Universidad de la República

Facultad de Ciencias Económicas y de Administración

Licenciatura en Estadística

Caracterización de los jóvenes uruguayos que no asisten al Sistema Educativo.

Informe final de Pasantía para la obtención del título de Licenciado en Estadística.

Natalia Caballero

Gerardo Jadra

Tutora:

Ec. Laura Nalbarte

Octubre 2013

Agradecimientos.

Queremos agradecer a todas las personas que nos apoyaron y ayudaron brindándonos su tiempo para cumplir con este proyecto. En primer lugar, agradecemos a la Profesora Laura Nalbarte, nuestra tutora, por el apoyo brindado y las recomendaciones brindadas en todo momento. En segundo lugar queremos agradecer a nuestras familias por el apoyo y por ser sostén en todo momento brindándonos su aliento para concretar este proyecto. Por último un agradecimiento a nuestros amigos y compañeros de trabajo que también nos han acompañado en este camino. A todos, muchas gracias.

Resumen Ejecutivo.

La temática que aborda el presente trabajo es determinar las características que impactan en la asistencia a un centro educativo de los jóvenes uruguayos de 14 a 17 años de edad. Se toma como fuente de datos los provenientes de la ECH2011 que elabora el INE.

Se trabaja con una muestra de 8.868 jóvenes, de los cuales 7.207 asisten a un centro educativo (81,2 %) y 1.661 no asisten (18,8 %). Los datos fueron ponderados de forma de expandirlos a toda la población de interés, es decir los jóvenes de 14 a 17 años de edad. A su vez se adoptó como estrategia particionar dicha muestra en una muestra de entrenamiento, la cual se utilizó para aplicar las técnicas, y una muestra de prueba en la que se evaluó su poder predictivo.

Se realizaron distintos análisis: estadística univariada para la caracterización general de los jóvenes que asisten y de los que no asisten, técnicas paramétricas como ser Modelos Lineales Generalizados y no paramétricas como por ejemplo Análisis Factorial y Árboles de decisión. El Análisis Factorial, en particular Análisis de Correspondencia Múltiple se empleó para la construcción de un índice sintético que resuma características del hogar (Índice de Confort) y acceso a tecnologías de la información (Índice de TIC's), para su posterior incorporación al modelo predictivo de la asistencia. Posteriormente se compara la eficiencia de incluir estos indicadores en el modelo vs. incluir las variables que los componen en forma desagregada.

Como complemento a los Modelos Lineales Generalizados se emplea la técnica de Árboles de decisión, en particular se construyen árboles de clasificación para la variable asistencia. Se analiza su aporte a la descripción de la problemática de la asistencia y se compara su performance con la de los Modelos Lineales Generalizados. Los resultados obtenidos para los distintos modelos reflejan qué características como la edad, el ser activo, tener hijos a cargo y ser jefe del hogar tienen un impacto negativo en la asistencia, lo cual (disminuye la probabilidad de asistencia del joven). En cambio características como los años de educación y el clima educativo tienen un impacto positivo. Es decir que cuanto mayor sean los años de educación acumulados por el joven y el clima educativo del hogar más probable es que asista. También surge un comportamiento

diferenciado en cuanto al género encontrándose una mayor probabilidad de asistencia en las mujeres respecto a los varones.

En lo que refiere a las características del hogar se constata una mayor probabilidad de asistencia en hogares donde la calidad de la vivienda es buena frente a aquellos donde es deficitaria. La condición de hacinamiento tiene un impacto negativo, disminuyendo la probabilidad de asistir.

Respecto al lugar de residencia, el hecho de residir en la capital aumenta la probabilidad de asistir frente a los que residen en el interior del país.

Al comparar los modelos que incluyen los índices con los que no los incluyen, surge que ambos presentan una buena performance en cuanto a la predicción. Esto permite la aplicación de ambos según se quiera analizar el impacto de la variable agregada (índice sintético) o analizar el impacto de determinada variable en forma individual (variables individuales).

En lo que respecta a los árboles de clasificación y de su comparación con los modelos anteriores, se aprecia que dicha técnica aporta a la descripción de la variable asistencia y a determinar cuáles son las variables que discriminan entre los que asisten y los que no a un centro educativo. Por otra parte su rendimiento en cuanto a la predicción no es del todo satisfactorio presentando tasas de error altas para predecir los que no asisten.

Como recomendaciones a considerar a futuro, se plantea indagar sobre los motivos de la no asistencia a un centro educativo, donde a modo de ejemplo surge que un 63 % de los jóvenes se desvincula mostrando como principal motivo el desinterés. Siguiendo la línea de investigación, se plantea el estudio del grupo de jóvenes que se desvinculan por desinterés.

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción.

En los últimos años el debate público ha puesto el foco en la educación, en particular ha preocupado a las autoridades el elevado número de jóvenes que no asisten a un centro educativo, que se desafilian del sistema sin finalizar sus estudios secundarios. Se han implementado programas de relativo éxito apuntando a revertir este fenómeno y que un mayor número de jóvenes permanezca dentro del sistema educativo. La temática que se aborda en éste trabajo es el fenómeno de la no asistencia a centros educativos de los jóvenes en Uruguay.

En particular el relevamiento que lleva adelante el Instituto Nacional de Estadística (en adelante INE) en su Encuesta Continua de Hogares (en adelante ECH), muestra que para los años 2009 a 2011 la no asistencia de los jóvenes entre 14 y 17 años de edad a un centro educativo se ha mantenido en el entorno de un 20 %, mostrando una persistencia del fenómeno y la incapacidad de las políticas educativas para revertirlo. Por otra parte la nueva Ley de Educación amplía la cobertura estableciendo la obligatoriedad para el ciclo de Enseñanza Media Superior (EMS) y si se considera a los jóvenes de entre 14 a 17 años considerados por edades simples los datos revelan un aumento progresivo de las tasas de no asistencia conforme aumenta la edad del joven.

Este trabajo pretende analizar cuáles son las características que impactan en la asistencia y cuáles en la no asistencia a un centro educativo de los jóvenes uruguayos de entre 14 y 17 años de edad.

El presente estudio se realizó con los datos de la ECH del año 2011 relevada por el INE. La temática de la asistencia es un fenómeno que posee varias aristas, es multidimensional, pasando por factores que abarcan desde lo micro-social a lo macro-social. Éste trabajo se centró en lo que la teoría denomina aspectos micro-sociales es decir en aquellos que toman como centro al

individuo y su entorno inmediato. Por lo tanto se hace especial hincapié en las características de los jóvenes que no asisten a un centro educativo y la situación del hogar en los que los mismos habitan. Pero es pertinente destacar que en el fenómeno de la no asistencia también influyen otros aspectos que no son menos relevantes como ser características concernientes a los centros educativos, como su tamaño, el clima organizacional, las políticas pedagógicas llevadas adelante por los centros, etc. O aspectos de carácter macro-social como ser el sistema educativo en su conjunto, la política educativa, el gasto público en educación, etc..

En éste informe se tomó como eje a los jóvenes como protagonistas centrales del proceso educativo y las características que comparten aquellos que declaran estar asistiendo y los que declaran no hacerlo.

El análisis de las características compartidas por los jóvenes de entre 14 y 17 años que no asisten a un centro educativo, así como la de los hogares en los que habitan se considera que pueden ser el punto de partida en el diseño y ejecución de estrategias de intervención que tengan como finalidad desincentivar la no asistencia y aumentar las tasas de retención por parte de los centros educativos. Más aún, si tomamos en cuenta el papel que juega la educación como generadora de oportunidades, protección social y motor de desarrollo de un país. Como bien señala Fernández en su artículo "Desafiliación educativa y desprotección social"(2010), al decir que la desafiliación es una trayectoria de transición al mundo adulto que deja a quien la sigue en un estado de vulnerabilidad social.

1.1. Objetivos y principales hipótesis.

El objetivo principal de éste trabajo es modelizar y caracterizar la asistencia al sistema educativo por parte de los jóvenes uruguayos de entre 14 y 17 años de edad.

Asímismo se plantean los siguientes objetivos específicos:

• Construir mediante Análisis de Correspondencia Múltiple tres indicadores que resuman características de la vivienda en cuanto a salubridad, nivel de confort y acceso a teconologías de la información y comunicaciones.
• Analizar qué factores son determinantes a la hora de optar por no asistir a un centro educativo y entender en que sentido operan los mismos. Para ello se estima un modelo lineal generalizado

que permita evaluar el impacto de un conjunto de características del joven en la probabilidad de asistencia a un centro educativo.

• Construir una regla de clasificación (árbol de clasificación) como complemento al modelo lineal generalizado que permita el estudio de diferentes "trayectorias" por las cuales un joven puede llegar a tomar la decisión de no asistir.

• Realizar una comparación del poder predictivo de ambas técnicas (modelo lineal generalizado-árbol de clasificación).

• Aportar evidencia de carácter nacional a la problemática de la asistencia y no asistencia al sistema educativo.

Las hipótesis se centran en características de los individuos y de los hogares donde los mismos residen, así como el sentido en el que operan las mismas, es decir favoreciendo o no la asistencia.

A la luz de los antecedentes analizados se espera encontrar diferencias en cuanto al sexo de los jóvenes, es decir una mayor propensión a la asistencia en las mujeres que en los varones. Por otra parte y teniendo en cuenta el tramo etario que se analiza en el presente trabajo, muchos jóvenes comienzan a participar del mercado laboral dándose una competencia en el uso del tiempo que destinan a estudiar y el que destinan a trabajar, por lo que se espera que la condición de ser activo opere en forma negativa en la asistencia. A su vez, tal como señalan estudios anteriores (Casacuberta y Buchelli, 2010), se espera encontrar una caída en los niveles de asistencia conforme aumenta la edad del joven. Por otro lado, el tener hijos a cargo también compite con el tiempo destinado a estudiar por lo que se espera un impacto negativo en la asistencia. Cuanto mayor sea el número de hijos a cargo mayor el impacto en la no asistencia. Respecto a los años de educación acumulados por el joven, la literatura señala un impacto positivo cuanto mayor son los años de educación acumulados por el joven, o más cercano se encuentra a los años acumulados esperados dada la edad del joven, es decir, cuando no existe repetición (Ferrari, Martínez y Saavedra, 2010). En cuanto al lugar de residencia de los jóvenes se espera encontrar una mayor proporción de jóvenes que no asisten en el interior del país, esto tal vez relacionado con la distancia a la que se encuentra el centro educativo más próximo (dificultad de acceso) o con decisiones más tempranas de incorporación al mercado de trabajo en relación con los jóvenes que residen en la capital.

En cuanto a las características de los hogares donde los jóvenes habitan se espera que cuanto mejor sean las condiciones estructurales de la vivienda, mejores sean las condiciones sanitarias, es decir el acceso al agua potable, la condición de no hacinamiento, etc. más probable es que el joven asista. Por último en relación al clima educativo del hogar se espera que hogares con mejor clima educativo, es decir mayor promedio de años de educación de los adultos, más probable sea la asistencia por parte de los jóvenes. Se espera que en estos hogares los padres tengan mayores habilidades para acompañar los procesos de aprendizaje de sus hijos.

El presente trabajo se organiza del siguiente modo: Un primer apartado donde se lleva a cabo una revisión de los antecedentes a nivel nacional e internacional de la problemática de la no asistencia a un centro educativo. Un segundo apartado donde se presenta la dimensión del fenómeno bajo estudio y los distintos enfoques para abordar la temática de la desafiliación educativa. Luego un capítulo dedicado a exponer la metodología utilizada para cumplir con los objetivos. Un capítulo donde se realiza una descripición exhaustiva de los datos de la ECH2011 y sus estadísticas descriptivas. Posteriormente se presentan los resultados obtenidos. Por último se exponen las conclusiones extraídas y recomendaciones para futuros estudios.

Capítulo 2

Antecedentes.

A continuación, se hace referencia a investigaciones previas a nivel nacional e internacional que han abordado el tema de la asistencia a un centro educativo y los factores que determinan que el joven tome la decisión de no asistir.

2.1. Antecedentes a nivel nacional.

Dentro de las investigaciones a nivel nacional cabe destacar en primer término el trabajo, "Asistencia a instituciones educativas y actividad laboral de los adolescentes en Uruguay, 1986-2008", de Casacuberta y Bucheli (2010). En segundo término se encuentra el trabajo, "Estrategia nacional para la infancia y la adolescencia" de Ferrari, Martínez y Saavedra (2010). Por otra parte se cita también el trabajo, "Factores escolares y desafiliación en la Enseñanza Media Superior(2003-2007)", de Fernández (2010).

• Casacuberta y Bucheli analizan la asistencia a centros de educación y desempeño en el mercado de trabajo (actividad laboral, empleo y desempleo) de jóvenes entre 14 y 17 años de edad en las cohortes nacidas entre 1972 y 1991 en Uruguay. En una primera parte del documento los autores estudian la evolución de la asistencia y la participación laboral tomando como fuente de información la construcción de un cuasi panel conformado por las ECH del INE entre 1986 y 2008. Allí se constata un crecimiento en la proporción de jóvenes que asisten al sistema educativo para el grupo de 14 a 17 años, pero se advierte también el hecho de que este aumento no sería gradual si no que tuvo un escalón a fines de los noventa. Los autores atribuyen este fenómeno a los efectos que tuvieron las medidas en políticas educativas que se tomaron a mediados de los noventa. En cuanto a la participación laboral se encuentra un comportamiento opuesto con una caída en la proporción de activos. Esta caída no es gradual sino que también presenta un escalón

al igual que el encontrado para la asistencia. Se analizan a su vez, las características de los adolescentes, de sus hogares y las condiciones generales del mercado de trabajo que influyen en la asistencia/participación en la población económicamente activa (PEA). Se utiliza como fuente de información los datos aportados por la ECH 2008 y se analiza un modelo probit-bivariado. A partir de dicho modelo se analiza cuál es el impacto de las variables en la probabilidad de asistir y trabajar. Entre los principales resultados de dicha investigación se encuentra que la edad, el haber tenido hijos, la extraedad y el hecho de que el hogar se encuentre en un contexto socioeconómico bajo impactan en forma negativa en la probabilidad de asistir. También aparece como factor determinante el clima educativo del hogar, donde allí se destaca que la presencia de adultos con más de 9 años de educación en promedio, impacta en forma positiva en la asistencia.

• Ferrari, Martínez y Saavedra en su trabajo se plantean como objetivo general proponer alternativas de política pública que apunten a la permanencia o retorno al sistema educativo de los adolescentes entre 14 y 17 años económicamente activos y de aquellos que no estudian, no trabajan y no buscan trabajo. Para ello realizan una caracterización de los jóvenes que trabajan o buscan empleo y sus hogares, analizándose en primer lugar la situación de los adolescentes en relación a la participación en el sistema educativo, en segundo lugar el perfil de los hogares de los adolescentes que trabajan o buscan empleo, el nivel de remuneraciones de los adolescentes y su importancia en los ingresos del hogar; tomando como referencia los datos que surgen de la ECH 2008.

En dicho artículo los autores describen que de los adolescentes de 14 a 17 años de edad, el 79 % asistía a algún centro educativo, siendo mayor la asistencia de mujeres (82 %) que la de varones (75 %). Señalan por otra parte que los niveles más significativos de desvinculación comienzan entre los 13 y los 14 años, y se incrementan con la edad. Muestran que el primer momento en la trayectoria educativa donde se produce un alto porcentaje de abandono es en primer año de educación media, hallando que un 13 % de los que culminan primaria no culminan el 1er. año del liceo.

Entre los factores encontrados por los autores más vinculados con el fenómeno de la desafiliación y la inserción temprana en el mercado de trabajo se encuentra el clima educativo del hogar. Plantean que cuanto menor es el nivel educativo alcanzado por el jefe de hogar y su cónyuge aumenta en gran medida la probabilidad de los adolescentes de desafiliarse y de ser económicamente activos. En el mismo sentido operan los ingresos, siendo mayor la probabilidad de desafiliación y de trabajar cuanto menor es el nivel de ingresos del hogar. Por otra parte, establecen que la repetición en primaria impacta negativamente sobre la asistencia, sobre todo en los varones;

al tiempo que impacta positivamente sobre la probabilidad de trabajar, y especialmente en el trabajo no remunerado de las mujeres.

• Fernández comienza su investigación planteándose la pregunta de si las instituciones educativas (liceos) de Educación Media Superior inciden en la desafiliación de sus estudiantes. El fenómeno de la desafiliación y la no asistencia es multidimensional y posee varias aristas, el autor se propone abordar la problemática desde la perspectiva del centro educativo, y evaluar diversos factores como si la misma es pública o privada, la estructura académica (general o técnica), así como el tamaño y el clima organizacional. Para llevar adelante dicha investigación el autor trabaja con una muestra de datos panel de estudiantes evaluados por el Proyecto Programa de Evaluación Internacional de Estudiantes 2003 (PISA, por su sigla en inglés) de la Organización para la Cooperación y el Desarrollo Económico (OCDE), y se aplica un modelo lineal generalizado de tipo jerárquico, en la cual se analizan los distintos efectos vinculados al centro educativo y su impacto en la desafiliación. En cuanto a la pregunta que fue eje de su investigación el autor concluye de forma afirmativa, encontrando que las escuelas sí hacen diferencia en la desafiliación y no asistencia de sus estudiantes. Entre los principales resultados de dicha investigación se destaca que la probabilidad de desafiliación es mayor para: un varón que una mujer, los que asisten a un centro en una localidad con menos de 5 mil habitantes, los que cursan Bachillerato Diversificado en comparación con los que cursan Tecnológico, y los que asisten a un liceo cuyo nivel de matriculación es superior a los 1200 alumnos.

2.2. Antecedentes a nivel internacional.

A nivel internacional es variada la bibliografía sobre la temática de la desafiliación, entre otros se destacan por compartir el mismo objeto de estudio: "Factores individuales, familiares e institucionales relacionados con la deserción en una escuela preparatoria de Yucatán"(2010) de Pilar Angelina Canto Bonilla, "Caracterización de la deserción estudiantil en la Universidad Nacional de Colombia"(2006) de Darío Alberto Rico Huguita. Es de especial mención, no solo por compartir el mismo objeto de estudio sino también por estar enfocada a la misma población objetivo, la investigación "Deserción escolar y trabajo infantil en Costa Rica"(2003) de Ana María Cerdas, llevado adelante por el Instituto de Economía de la Pontificia Universidad Católica de Chile.

• Cerdas en su trabajo analiza los determinantes de la deserción escolar en conjunto con el trabajo infantil para adolescentes costarricenses entre 12 y 17 años. Para ello utiliza un modelo de asignación de tiempo estimado mediante un modelo probit bivariado. Para su implementación

se utilizó la Encuesta de Hogares de Propósitos Múltiples del año 2000 para Costa Rica, que es realizada por el Instituto Nacional de Estadísticas y Censos de dicho país. Cerdas obtiene como resultados que la edad tiene un impacto negativo en la asistencia, conforme la edad del niño aumenta se verifica una disminución en su probabilidad de estudiar. El hecho de residir en una zona urbana tiene un impacto negativo en la asistencia, mientras que no halla una diferencia en cuanto al sexo.

Cerdas señala que las características del hogar ejercen distintos efectos en la probabilidad de estudiar, encontrando que el tamaño del hogar y la presencia de sólo un padre (hogares monoparentales) no parece tener un peso significativo en la decisión de asistir. En cambio la educación promedio de los padres si lo tiene y este es positivo. También se destaca que ni el ingreso familiar y ninguna de las variables referidas a las características del sistema educativo resultan significativas a la hora de asistir a un centro educativo. En cuanto a la incorporación al mercado de trabajo, Cerdas encuentra que cuanto mayor es la edad del joven mayor es su probabilidad de incorporarse al mercado de trabajo, y esta es mayor en comparación si son varones. En cuanto al impacto de las características del hogar se plantea que el tamaño del hogar tiene un efecto positivo y que cuanto mayor es el promedio de años de educación de los padres menor la probabilidad de trabajar. Por último el ingreso del hogar tiene un impacto negativo sobre la probabilidad de trabajar, es decir que cuanto mayor es el ingreso del hogar menor la probabilidad de trabajar del joven.

Capítulo 3

Marco Teórico.

3.1. Algunos conceptos previos: Desafiliación, ausentismo, abandono y no asistencia.

Para el desarrollo de la siguiente sección se siguió fundamentalmente el artículo "Desafiliación educativa y desprotección social" de Fernández et al.(2010).

Al analizar la bibliografía existente sobre la temática de la desafiliación, se advierte el hecho de que en varias ocasiones se utilizan diferentes términos como sinónimos para referirse al tema de la desvinculación educativa. Fernández et al. insisten en diferenciar unos conceptos de otros, y en hacer un esfuerzo para dar una definición de cada uno de ellos, a los efectos de determinar cuál es la problemática que se pretende abordar y también la metodología para medir cada uno de los fenómenos. A su vez señalan que son diversos los eventos que se presentan en el curso de vida de los adolescentes y jóvenes, y que son cubiertos por el mismo término: "deserción" como ser, el ausentismo, el abandono de un curso y la decisión de dejar de estudiar. El Sistema de Información de Tendencias Educativas en América Latina (SITEAL) que analiza los vínculos entre sociedad y educación, define la deserción como la no concurrencia actual al centro de estudios que se infiere resultante de la no inscripción o matriculación este año; ya sea esta una decisión tomada por la familia o por el mismo joven.

El análisis debe permitir ubicar a los adolescentes en algún punto del espectro de integración/ desvinculación al centro educativo que tiene por extremos los estados de: plena afiliación y desafiliación. A continuación se definen los conceptos de ausentismo, abandono, desafiliación y no asistencia, como forma de delimitar unos fenómenos de otros y encuadrar el objeto de estudio del presente trabajo que atañe a la no asistencia a un centro educativo.

3.1.1. Ausentismo.

Fernández et al. definen el ausentismo como un atributo de un estudiante que caracteriza su comportamiento regular de estudiante durante un año lectivo o ciclo escolar. El mismo surge de contabilizar las inasistencias que se dieron durante el año y de definir también un umbral, un "límite de faltas" el cual si se supera trae aparejado la pérdida del ciclo.

3.1.2. Abandono.

Siguiendo a Fernández et al. por abandono se alude a la ausencia definitiva y sin causa justificada del centro escolar por parte de un alumno sin haber finalizado la etapa educativa que esté cursando. Es decir que se trata de un evento que tiene como consecuencia la reprobación del curso y la baja de registros del centro de estudio. Este fenómeno se explica por mecanismos que son diferentes a la salida definitiva del centro educativo, es decir que básicamente se trata de una elección errónea por parte del joven respecto a un curso o una escuela que se genera por información incompleta como puede ser señales de mercado de trabajo, costo-oportunidad de estudiar, así como características de la escuela, su clima, sus docentes, así como también entran en juego características subjetivas respecto a la autovaloración que hacen los jóvenes de si mismos y a las probabilidades subjetivas de finalizar con éxito el nivel que han comenzado. Cabe destacar en relación al evento abandono que el mismo es reversible, dado que un estudiante puede reinscribirse al año siguiente en el mismo curso, puede ir a otra escuela o incluso puede cambiar de orientación en sus estudios.

3.1.3. Desafiliación.

El término desafiliación refiere a un estado que se caracteriza por la desvinculación del joven del sistema educativo y el truncamiento de su trayectoria educativa. El Centro Nacional de Estadísticas de Educación (NCES por su sigla en inglés) establece que para configurarse el estado de desafiliación se deben registrar tres eventos: dos decisiones en años consecutivos de no matricularse y que no haya acreditado el nivel educativo a través de programas alternativos. De esta definición se desprende que para hablar de desvinculación es necesario contar con datos de tipo longitudinal sobre el estudiante y no solamente con datos contables de matriculación.

Fernández et al. definen a la desafiliación como una trayectoria de transición al mundo adulto que deja a quien la sigue en un estado de vulnerabilidad social. Se caracteriza por el truncamiento (o falta de acreditación) de la trayectoria académica en el ciclo medio, la pérdida de expectativas respecto al bienestar futuro que podría derivarse de la educación, y por el relegamiento a una

posición social vulnerable o directamente excluida de la protección social asociada a la asistencia al centro de estudios. Nuevamente a raíz de esta definición es claro cómo para hablar de desvinculación se hace necesario el estudio de las trayectorias educativas de los jóvenes y por esto la necesidad de contar con información de carácter longitudinal que abarque un determinado período en el tiempo de ese joven. Asimismo se hace hincapié en resaltar el estado de vulnerabilidad en que quedan aquellos jóvenes que ven truncada su trayectoria educativa, entendido en relación a que una mayor cantidad de años de educación posibilita un mejor acceso al mercado de trabajo y mayores ingresos asociados, dado que cuanto mayor sea la calificación mayor es la posibilidad de desempeñarse en procesos de trabajo más complejos. Tal es así, que muchos autores en sus enfoques multidimensionales de pobreza clasifican a aquellos jóvenes que no han acreditado la educación básica obligatoria como pobres.

3.1.4. Desafiliación y no asistencia a un centro educativo.

Fernández advierte que es común en muchos casos encontrar que se confunda la desafiliación con la no asistencia a un centro educativo. Se constata que en diversas mediciones se toma la no asistencia en conjunto con la falta de acreditación del ciclo o nivel y se lo asimila a deserción. Es decir, que si ante la pregunta de asistencia éste responde que no asiste y además que no ha completado el nivel o ciclo cursado, se lo registra como "desertor". Surge de forma clara que esta forma de medición constituye una aproximación al fenómeno, pues detecta un proceso solo a través de un único evento como es la no asistencia. De este modo no se puede inferir que la persona se haya desafiliado sin cometer errores. Además la no asistencia puede deberse a la no inscripción o al abandono, es decir que dicho estado puede ser transitorio o deberse a una situación de carácter coyuntural.

La no asistencia a un centro educativo es relevada por la ECH que elabora el INE, dado que en el presente trabajo se trabaja con la ECH2011 y teniendo en cuenta a su vez todos los conceptos que fueron vertidos anteriormente, es relevante hacer hincapié en que sólo se puede hablar de asistencia y no de deserción.

3.2. Diferentes enfoques para abordar la desafiliación educativa.

A continuación se hará una breve reseña de los diferentes abordajes que se presentan en la bibliografía para tratar la temática de la desafiliación educativa. Para el desarrollo de la presente

sección se siguió la exposición que hace Fernández en su artículo "Enfoques para explicar la desafiliación"(2010).

Los distintos enfoques se diferencian unos de otros por el énfasis que cada uno de ellos hace para explicar el fenómeno de la desafiliación, algunos ponen mayor atención en aspectos individuales representados por el individuo y su familia, otros en características exclusivamente institucionales, como ser el centro educativo, su organización, su estrategia pedagógica; y por otro lado aquellos que ponen mayor relevancia en aspectos de la sociedad en su conjunto o del sistema educativo en su totalidad, etc. También suelen diferenciarse por el grado en que los diferentes factores pueden ser manipulables por la voluntad, ya sea de profesores o de los hacedores de políticas (administradores).

3.2.1. Enfoques que toman como centro al individuo y su entorno.

En dichas teorías se pone mayor énfasis en la situación del hogar al que pertenece el individuo, se establece una relación significativa entre los ingresos del hogar, el clima educativo y la desafiliación. Es decir que sería un factor determinante la disponibilidad y nivel de recursos que un hogar tiene para mantener a los jóvenes estudiando a tiempo completo sin que deban ingresar al mercado de trabajo, tener que aportar un ingreso adicional y configurarse el estado de desafiliación. En estos enfoques se establece que existe una competencia entre el tiempo dedicado a estudiar y el tiempo dedicado a trabajar. Se desprende de dichos enfoques que la causa más persistente de la desafiliación sería la pobreza, los jóvenes provenientes de hogares pobres tienen mayor probabilidad de un ingreso temprano al mercado de trabajo y por consiguiente configurar un estado de desafiliación.

Por otra parte existen un conjunto de teorías que ponderan la racionalidad que los jóvenes individualmente ponen en sus acciones y decisiones, las de sus familias y su influencia en los grupos de pares. Se encuentran aquí teorías que postulan que los jóvenes evalúan, comparan, razonan y toman una decisión. Se supone que los estudiantes al inicio de cada año realizan una evaluación del costo/beneficio de adicionar un año más de educación a sus vidas. En segundo término la teoría también supone la existencia de un umbral en la utilidad del individuo, y que todo individuo acumulará años de educación hasta llegar a un nivel óptimo. En tercer lugar se establece que en las decisiones influyen expectativas sobre el futuro, en este caso serían tasas de retorno de la educación. En el marco de dichos enfoques se desprende que todos los individuos tendrían una propensión a la desafiliación educativa, y que la desafiliación observada sería el resultado de que el óptimo del individuo se encontraría por debajo del óptimo social en la acu-

mulación educativa.

3.2.2. Enfoques que tienen como centro la comunidad y las organizaciones.

Bajo este ítem se encuentra un conjunto de teorías que ponen de relieve aspectos que tienen que ver con la localidad o el barrio del centro educativo, y con características organizacionales de los centros educativos. Estos enfoques parten del supuesto de que la decisión de desafiliación no es el resultado exclusivo de factores individuales sino que también entran en juego las características de los territorios así como de las organizaciones a las que pertenecen los jóvenes. Una de las causas más relevantes de desafiliación que se plantean es la marginación socioterritorial que caracteriza a ciertos barrios con escasez de recursos materiales y servicios públicos, y también de localidades con alta concentración de minorías étnicas o raciales. Por otro lado en lo que respecta a las organizaciones se pone énfasis en los entornos normativos que dichos centros generan para sus docentes y alumnos. Se entiende que aquellos centros en donde se establece un clima organizacional estimulante, centrado en el aprendizaje, y existe un buen relacionamiento entre docentes y alumnos terminan operando con una mejor retención de sus alumnos por parte del centro y con ello una caída en la probabilidad de desafiliación.

3.2.3. Enfoques que toman como centro aspectos de la sociedad nacional y global.

En estos enfoques se toma como eje de la discusión ya no aspectos individuales, como la clase social, el género o condición de pobreza, ni tampoco las características del centro educativo o barrio. La probabilidad de desafiliación estaría vinculada con aspectos de la sociedad toda, como ser su economía, su estado, sus instituciones o su historia. Uno de los supuestos más manejados es que la desafiliación tiene una relación inversa con el nivel de bienestar que, en promedio, cada país garantiza a sus habitantes. Entra en juego también el porcentaje del PIB que se destina a la educación. Incluso se encuentran estudios que vinculan y hacen ordenamientos de diferentes países evaluando sus tasas de abandono escolar con su PIB per cápita.

Si bien los diversos enfoques de teoría presentados anteriormente atañen exclusivamente al fenómeno de la desafiliación, los autores consideramos que son perfectamente extrapolables al fenómeno de la no asistencia entendida esta como un factor de riesgo para configurarse el posterior estado de desafiliación.

Capítulo 4

Metodología.

A continuación se describe la metodología utilizada para el cumplimiento de los objetivos del presente trabajo. En la primera sección se presenta la teoría utilizada para la construcción de distintos indicadores de interés, en las secciones siguientes, se presentan los instrumentos empleados para la obtención del mejor modelo que represente la asistencia de los jóvenes al sistema educativo y se expone la teoría utilizada para la selección del modelo así como distintos test empleados tanto para la significación de las variables como para la predicción del modelo. Para la construcción de los indicadores antes mencionados se utilizaron distintos instrumentos de los cuales se entiende que es relevante incluir los aspectos teóricos asociados a los mismos: Análisis factorial, en particular: Análisis de Correspondencia Múltiple (ACM) y la teoría para la construcción de un Índice de Calidad de la Vivienda. En la búsqueda del mejor modelo que explique la asistencia de los jóvenes a un centro de educación se utilizó una combinación de técnicas paramétricas y no paramétricas: Modelos Probit, Modelos Logit y Árboles de Clasificación. Se detallan a continuación las metodologías mencionadas.

4.1. Teorías para la construcción de indicadores

4.1.1. Análisis de correspondencia múltiple.

El ACM es una técnica de análisis multivariado enmarcada dentro de las técnicas de Análisis Factorial. El Análisis Factorial es una técnica cuyo propósito, entre otros, consiste en buscar el número mínimo de dimensiones o factores capaces de explicar el máximo de información contenida en los datos. El nombre factorial se debe a que la descomposición que se pretende de la matriz de datos se realiza en factores o también llamados ejes de inercia sobre los cuales se realiza la proyección de la nube de datos. Esta técnica transforma las proximidades entre los datos en distancias euclídeas, lo que permite una fácil interpretación.

Los objetivos son:

• Reducir la dimensionalidad de la matriz de datos con el fin de eliminar información redundante y manifestar las relaciones existentes entre las variables. Se obtienen nuevas variables o factores que resumen la información esencial que surge de la nube de puntos.
• Con respecto a los individuos, dos individuos serán más parecidos cuanto más cercanos sean sus valores en el conjunto de las variables.
• En referencia a las variables, dos variables serán más cercanas cuantos más individuos compartan.

En la práctica, es de interés, obtener un primer eje factorial en el cual su dirección haga máxima la inercia respecto al baricentro. Una vez encontrado el primer eje, al segundo se le impone la condición de ser ortogonal al primero y así sucesivamente se van encontrando el resto de los ejes. La inercia de la nube proyectada sobre un eje u_s es igual a: $\sum_i p_i[F_s(i)]^2$
Siendo p_i el peso de la fila i y $F_s(i)$ la inercia del elemento i. Matricialmente se puede escribir en función de la matriz diagonal de los pesos D y del vector de las coordenadas de las proyecciones sobre u_s, F_s, como $F_s'DF_s$. Como $F_s = XMu_s$, resulta: Inercia $= u_s'MX'DXMu_s$
Siendo X la matriz de datos, M la métrica y u_s la dirección, eje de inercia o eje factorial.
Cuando las variables son cualitativas se puede aplicar el ACM para obtener indicadores sintéticos. Cada variable tiene un conjunto de categorías o modalidades excluyentes entre las cuales el individuo es clasificado. En el caso de que alguna de las variables sea cuantitativa, esta se puede transformar en cualitativa, dividiendo su rango en intervalos. Cada uno de estos intervalos será una categoría de la nueva variable cualitativa. Se puede desarrollar en dos tipos de tablas: la tabla disjunta completa, que es una tabla de individuos por modalidades, y la tabla de Burt, que cruza modalidades por modalidades. A continuación se presenta la técnica ACM mediante la tabla disjunta completa, para ello se sigue fundamentalmente el texto de Blanco, J.(2006)

La tabla disjunta completa está compuesta únicamente por ceros y unos. K_j es el número de modalidades de la variable j, tal que $\sum_{j=1}^{J} K_j = K$ y $K = \sum_j [1]^j = JK_j$, en este caso K es el número total de modalidades de la tabla. Por otro lado, x_{ih} vale 1 cuando el individuo i presenta la categoría h de la variable j y 0 en caso contrario, siendo $\sum_{h=1}^{K_j} x_{ih} = 1$ y $\sum_{h=1}^{K} x_{ih} = J$. El número de filas corresponde a la cantidad de individuos considerados I, por lo que $\sum_{i=1}^{n} x_{ih} = I_h$, siendo I_h, la cantidad de individuos que tienen cierta modalidad h.

Con respecto a los individuos, dos individuos serán "iguales" si presentan las mismas modalidades. La distancia χ^2 entre dos individuos será:

$d^2(i,l) = \sum_k \frac{IJ}{I_k}(\frac{x_{ik}}{J} - \frac{x_{lk}}{J})^2 = \frac{1}{J}\sum_k \frac{I}{I_k}(x_{ik} - x_{lk})^2$
Con respecto a las modalidades, la distancia entre dos modalidades k y h será:

$d^2(k,h) = \sum_i I(\frac{x_{ik}}{I_k} - \frac{x_{ih}}{I_h})^2 = I(\frac{A_k}{I_k I_h} + \frac{A_h}{I_k I_h})$
Siendo A_k el número de individuos que tienen la modalidad k y no la h y A_h el de individuos que poseen la modalidad h y no la k.

Como los valores en esta tabla son ceros o unos, se puede observar claramente que cuando aumentan las diferencias de una misma modalidad, entre dos individuos, esta distancia se hace más grande. También sucede, que una modalidad "rara" estará siempre lejos del centro de la nube ya que el baricentro de las modalidades es constante y el perfil de la columna será cero o $1/I_k$.
La inercia global respecto al baricentro es:

$\sum_{k=1}^{K}$(Inercia de k en relación a G) $= \sum_{k=1}^{K} \frac{1}{J}(1 - \frac{I_k}{I}) = \sum_{k=1}^{K} \frac{1}{J} - \sum_{k=1}^{K} \frac{I_k}{JI} = (\frac{K}{J} - 1)$
La inercia será siempre igual al cociente entre dos modalidades menos uno que es igual a la suma de los valores propios. Por lo que no depende, en otras palabras, de la estructura de los datos. El valor propio λ_s asociado a un factor s es igual a la media de las correlaciones entre el factor y cada variable. Esta técnica permite relacionar las proyecciones de la nube de las filas con las proyecciones de la nube de las columnas. Esto se realiza mediante las formulas de transición:

$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_{k \in K} \frac{x_{ik}}{J} G_s(k)$,$G_s(k) = \frac{1}{\sqrt{\lambda_s}} \sum_{i \in I} \frac{x_{ik}}{I_k} F_s(i)$
Se aprecia que la proyección de un individuo i sobre un eje factorial s corresponde con el baricentro de las modalidades que fueron observadas sobre el i-ésimo individuo con un coeficiente de expansión $1/\sqrt{\lambda_s}$. La segunda ecuación muestra que la i-ésima modalidad es representada sobre un eje factorial s por el baricentro de los individuos observados que presentan la modalidad i (con un coeficiente de expansión $1/\sqrt{\lambda_s}$).

La proximidad sobre un eje factorial entre los individuos observados indica gráficamente la similitud entre individuos. La proximidad entre variables refleja diferentes conceptos: la proximidad entre las variables indicadoras mide la asociación entre ellas, mientras que la proximidad de medias de clases de individuos deja en evidencia su relación con las demás modalidades. Dos clases de individuos estarán tanto más cerca entre si cuanto posean más características similares en el

conjunto.

La información de cada factor se puede medir con la inercia explicada, esta depende del tamaño de la tabla, por lo que subestima la importancia de los ejes no resultando una buena medida de información. Para corregir la subestimación, algunos autores proponen distintas funciones:

Índice de Benzecri

$p(\lambda_s) = [\frac{J}{J-1}]^2[\lambda_s - \frac{1}{J}]^2$, siendo λ_s el autovalor de orden s.
Índice de Greenacre

$p(\lambda_s) = [\frac{J}{J-1}]^2[\sqrt{\lambda_s} - \frac{1}{J}]^2$, siendo λ_s el autovalor de orden s.

4.1.2. Índice de calidad de la vivienda

Se describe a continuación la metodología utilizada para la construcción del Índice de calidad de la vivienda y servicios . Este Índice fue elaborado en forma conjunta por el Ministerio de Desarrollo Social (MIDES) y la Oficina de Planeamiento y Presupuesto (OPP) en su documento "Principales Características del Uruguay Social" para el Reporte Social 2009. Para la construcción de este indicador se consideran las siguientes dimensiones: materiales de la vivienda, acceso al agua potable, acceso al saneamiento y acceso a la energía eléctrica.

4.1.2.1. Materiales de la vivienda

Esta dimensión se basa en la condición estructural de la vivienda, se define según la combinación de tres variables: materiales de construcción en paredes, piso y techos. Asumiendo las siguientes categorías: deficitaria, regular-recuperable, aceptable y buena.

Cuadro 4.1: Categorías según condición estructural de la vivienda.

Material en techos	Material en piso	Material en paredes externas	Categoría
Planchada de hormigón con protección (tejas y otros)	Cerámica, parqué, moqueta, linóleo o baldosas calcáreas	Ladrillos, ticholos o bloques terminados	Buena
Planchada de hormigón sin protección, o cubierta liviana con cielorraso, o cubierta liviana sin cielorraso	Cerámica, parqué, moqueta, linóleo o baldosas calcáreas	Todo menos materiales de desecho	Aceptable
Planchada de hormigón con protección	Cerámica, parqué, moqueta, linóleo o baldosas calcáreas	Todo menos ladrillo, ticholos o bloques terminados o materiales de desecho	Aceptable
Planchada de hormigón sin protección, o planchada de hormigón sin protección, o cubierta liviana con cielorraso o cubierta liviana sin cielorraso	Alisado de hormigón	Todos menos materiales de desecho	Aceptable
Todo menos materiales de desecho	Solo contrapiso, sin piso	Todo menos materiales de desecho	Regular - recuperable
Quinchado	Cerámica, parqué, moqueta, linóleo o baldosas calcáreas o alisado de hormigón	Todo menos materiales de desecho	Regular - recuperable
Techo con materiales de desecho	Todos	Todos	Deficitaria
Todos	Tierra sin piso ni contrapiso	Todos	Deficitaria
Todos	Todos	Materiales de desecho	Deficitaria

Fuente: Principales características del Uruguay social. MIDES-OPP.

4.1.2.2. Acceso al agua potable en la vivienda

Esta dimensión se define según la combinación de dos variables: el origen del agua utilizada por el hogar para beber y cocinar, y la forma en que llega el agua a la vivienda, tomando entonces las siguientes categorías: buen acceso, regular o malo.

Cuadro 4.2: Categorías de acceso al agua potable en la vivienda.

Categorías	Descripción
Buen acceso	Origen en red general o en pozo surgente protegido y acceso por cañería dentro de la vivienda.
Regular	Origen en pozo surgente no protegido, aljibe o arroyo y acceso por cañería dentro de la vivienda; o todos los orígenes excepto otros y acceso por cañería fuera de la vivienda.
Malo	Origen en canilla pública u otros y/o acceso por otros medios.

Fuente: Principales características del Uruguay social. MIDES-OPP.

4.1.2.3. Acceso al saneamiento

Para el cálculo de esta dimensión se combinan dos variables: existencia de baño (con descarga y sin descarga) y forma de evacuación del servicio sanitario. Las categorías resultantes son: adecuado y no adecuado (regular o malo).

Cuadro 4.3: Categorías según acceso al saneamiento.

Categorías	Descripción
Adecuado	Servicio con descarga y evacuación por red general, o fosa séptica o pozo negro.
Regular	Servicio con descarga y evacuación por entubado hacia un arroyo.
Malo	Servicio sin descarga y/o evacuación por otros medios, o no existencia de baño.

Fuente: Principales características del Uruguay social. MIDES-OPP.

4.1.2.4. Acceso a la energía eléctrica

Por último se calcula la dimensión acceso a la energía eléctrica, para esto se estudian los hogares según acceso a la energía eléctrica para iluminación de la vivienda a través del servicio de energía eléctrica de UTE, grupo electrógeno o cargador de batería.

4.1.2.5. ICV (Índice de calidad de la vivienda y servicios)

Luego del estudio de cada uno de estas dimensiones, se calcula el Índice de calidad de la vivienda y servicios. Este indicador se obtiene combinando el índice de calidad de los materiales de la vivienda, el acceso al agua potable, el acceso al saneamiento y el acceso a la energía eléctrica.

La metodología elaborada en el Reporte Social 2009, calcula el índice como:

Cuadro 4.4: Categorías del ICV según reporte social.

Categorías	Descripción
Buena calidad de materiales y servicios	Calidad de materiales buena, acceso al agua potable bueno, acceso al saneamiento adecuado y acceso a la energía eléctrica para iluminación.
Situación intermedia materiales y servicios	Se construye como categoría residual de las otras dos categorías.
Déficit de materiales y/o servicios	Calidad de materiales deficitaria y/o acceso al agua potable malo y/o acceso al saneamiento malo y/o no acceso a la energía eléctrica para iluminación.

Fuente: Principales características del Uruguay social. MIDES-OPP.

En el presente estudio se realizó una transformación de dicho índice fusionando las primeras dos categorías y asumiendo únicamente dos categorías finales:

Cuadro 4.5: Categorías del ICV reformulado.

Categorías	Descripción
Déficit de materiales y/o servicios	Calidad de materiales deficitaria y/o acceso al agua potable malo y/o acceso al saneamiento malo y/o no acceso a la energía eléctrica para iluminación
Calidad de materiales y servicios regular o buena	Se construye como categoría residual de la categoría anterior

Fuente: Elaboración propia.

4.2. Modelos lineales generalizados.

Para el desarrollo del siguiente apartado se siguió fundamentalmente el texto de Agresti, A.(2007) En un principio se planteó como objetivo buscar los factores determinantes que inciden en la decisión de que un joven de entre 14 a 17 años asista o no a un centro educativo. Para determinar estos factores, se modeliza esta decisión mediante un modelo lineal generalizado.

4.2.1. Componentes de un modelo lineal generalizado.

Los modelos lineales generalizados tienen tres componentes. Una componente aleatoria que identifica la variable de respuesta y su distribución de probabilidad. El componente sistemático que identifica el conjunto de variables explicativas. Un tercer componente que es la función de enlace (link), la cual especifica una función del valor esperado de Y_i (variable de respuesta del modelo); este componente vincula el componente aleatorio con el sistemático.

4.2.1.1. Componente aleatoria.

La componente aleatoria consiste en una variable aleatoria Y_i con observaciones independientes $(y_1, \ldots, y_n)$. También se especifica cuál es su distribución. En muchas aplicaciones, las observaciones de Y son binarias, y se denotan como éxito o fracaso. En forma más general, Y_i podría ser el número de éxitos en una serie de n pruebas, en este caso la variable de respuesta seguiría una distribución binomial. En otros casos también pueden asumirse distribuciones Poisson o binomiales negativas, cuando cada observación es por ejemplo un recuento.

4.2.1.2. Componente sistemática.

El componente sistemático especifica el conjunto de variables explicativas (predictores) que se incluyen en el modelo. Estas variables ingresan al modelo en forma de combinación lineal en la siguiente forma: $\alpha + \beta_1 X_1 + ... + \beta_k X_k$.

4.2.1.3. Función de enlace (link).

La función link especifica una función $g(.)$ que vincula la esperanza de Y ($E(Y) = p$) con el predictor lineal de la siguiente forma: $g(p) = \alpha + \beta_1 X_1 + ... + \beta_k X_k$.
La función link más sencilla es la función identidad, $g(p) = p$, que da lugar al modelo de regresión lineal clásico. Otras funciones link permiten una relación no lineal entre la esperanza de la variable de respuesta y el conjunto de predictores, estas son entre otras la logística, la función de distribución acumulada de una normal típica, etc. Cuando la función link es la logística se habla de modelos logísticos o logit, en el caso de la función de distribución de una normal típica se los denomina modelos probit.

Modelos logit.

Estos modelos se basan en la función de distribución acumulativa logística. En estos modelos la función link es de la forma:

$g(p) = \log[\frac{p}{1-p}]$, dicha función se conoce como link logit.

Los modelos que adoptan dicha función se conocen como modelos de regresión logísticos.

Modelos probit.

En los modelos probit la función link $g(p)$ es la función de distribución acumulada de una variable aleatoria normal típica.

4.3. Modelos generalizados para datos binarios.

En este caso tenemos una variable de respuesta que toma sólo dos valores:

$Y_i = 1$(Asiste a un centro educativo),
$Y_i = 0$ (No asiste a un centro educativo);

con probabilidad:

$$Y_i \begin{cases} 1 \text{ con probablidad p} \\ 0 \text{ con probablidad 1-p} \end{cases}$$

El valor 1 denota que el individuo ha tomado una decisión (en nuestro caso asistir a un centro educativo).

• La variable Y sigue una distribución Bernuolli con función de cuantía:

$f(y) = P(Y = y) = p^y(1-p)^{1-y}, y = 0, 1$

• $E(Y) = p$

• $V(Y) = p(1-p)$

En este estudio interesa analizar cual es la probabilidad de que el individuo asista a un centro educativo dadas ciertas características del joven, es decir la $Prob(Y = 1/\underline{X})$.

Para el caso de un modelo de regresión logística dicha probabilidad adopta la siguiente forma: $Prob(Y = 1/\underline{X}) = \frac{e^{(\alpha+\beta_1 X_1+\ldots+\beta_k X_k)}}{1+e^{(\alpha+\beta_1 X_1+\ldots+\beta_k X_k)}} = p$ esta función se corresponde con la función de distribución logística.

En un modelo de regresión probit el modelo a ajustar es: $Prob(Y = 1/\underline{X}) = F(\alpha + \beta_1 X_1 + \ldots + \beta_k X_k)) = p$, donde F es la función de distribución acumulada de una normal estándar.
Por lo que tenemos que: $F(x) = \Phi(x) = \int_{-\infty}^{x} \phi(x)\, dx$, donde $\phi(x)$ es la función de densidad de una normal típica. (Ver anexo metodológico para estimación del modelo y pruebas de significación, págs. 89-90).

4.3.1. Interpretación de coeficientes.

La interpretación de los coeficientes del modelo dependerá tanto de su signo como de la naturaleza de la variable a la que está asociado, es decir si la misma es continua o categórica. Si el signo asociado al coeficiente $\hat{\beta}$ es positivo aumentará la $P(Y = 1/\underline{X} = \underline{x})$, si es negativo disminuirá dicha probabilidad. En el caso de una variable continua para analizar el impacto de incrementar en una unidad dicha variable debemos analizar la magnitud de $e^{\hat{\beta}}$, dejando las demás variables constantes. La magnitud de $e^{\hat{\beta}}$ refleja el impacto en el cociente de odds, es decir en el cociente entre la probabilidad de asistir y de no asistir de incrementar en una unidad dicha variable, respecto del cociente entre la probabilidad de asistir/no asistir sin dicho incremento, dejando las demás variables fijas. En el caso de una variable categórica la magnitud $e^{\hat{\beta}}$ representa el impacto en el cociente de odds de poseer dicha categoría frente a una categoría que se toma como referencia y que no es incluida en el modelo.
En los modelos logísticos, la función link es el $\log[\frac{p}{1-p}]$, que no es más que el log del cociente entre la probabilidad de éxito y la probabilidad de fracaso. A modo de ejemplo un modelo donde se incluye una variable indicatriz para indicar si el individuo es hombre o mujer, podemos ver reflejado el impacto en el cociente de odds de uno frente a otro, es decir cuanto más probable es la asistencia de los varones respecto a la de las mujeres. Es decir se puede calcular el siguiente cociente: $\frac{\frac{\hat{p}}{1-\hat{p}}_{hombre}}{\frac{\hat{p}}{1-\hat{p}}_{mujer}} = e^{\hat{\beta}_{hombre}}$. Si el cociente es mayor que uno se puede ver cuanto más probable es la asistencia en los varones respecto a las mujeres.

4.3.2. Poder predictivo.

Para evaluar la bondad de ajuste de un MLG se utiliza como indicador el poder predictivo, allí se evalua la performance del modelo, comparando lo observado con lo que es predicho por el modelo. Dos formas útiles de resumir dicho poder predictivo es a través de las tablas de clasificación y las curvas ROC. Predecir el valor de la variable de respuesta en función de ciertos valores de las variables explicativas implica transformar las probabilidades estimadas por el modelo en respuestas binarias (0 o 1), para ello es necesario definir un valor crítico (punto de corte) a partir del cual las estimaciones del valor esperado implican un valor de 1 para la variable de respuesta. Por lo general se establece dicho punto en un valor de 0,5. Entonces tenemos que el valor predicho de un individuo será 1 si su probabilidad predicha supera el valor de 0,5 ($Y_i = 1$), y será cero si es menor o igual a 0,5. Dicha estrategia es válida si es igualmente probable que ocurra 0 o 1, y si los costos de predecir en forma incorrecta 0 o 1 son prácticamente los mismos. Aunque muchas veces ese punto de corte se toma como 0,5, es más apropiado tomar como punto de corte la proporción de valores $Y = 1$ en la muestra, e incluso más aconsejable es probar distintos puntos de corte y quedarnos con el que maximice la tasa de clasificaciones correctas.

Al estimar los errores con toda la muestra se produce un sesgo en las estimaciones, subestimando las mismas. Para corregir el sesgo se pueden adoptar diferentes estrategias como partición de la muestra, emplear validación cruzada o técnicas de remuestreo.

4.3.2.1. Partición de la muestra.

Una de las formas de corregir el sesgo en el cálculo del error del modelo consiste en particionar la muestra en dos submuestras. Una de ellas es la que se emplea para el ajuste o estimación del modelo que se denomina muestra de entrenamiento. Otra se emplea para evaluar la performance del modelo, es decir para hacer predicción, se la denomina muestra de prueba. Posteriormente se calculan los errores de clasificación en ambas y se comparan dichos errores.

4.3.2.2. Validación cruzada.

La validación cruzada es otra de las técnicas empleadas para validar modelos. Consiste en repetir y calcular los errores de clasificación para diferentes particiones de entrenamiento y prueba. Esta técnica es comúmente empleada cuando se trabaja con pocos datos. Existen diferentes tipos de validación cruzada como ser validación cruzada con K subconjuntos (K-folders) y leaving one out.

Validación cruzada con K subconjuntos.
En esta técnica los datos se dividen en K subconjuntos. Uno de los conjuntos se emplea como datos de prueba y el resto como datos de entrenamiento. Posteriormente el proceso de validación cruzada es repetido en los k grupos. En cada etapa se calculan los errores de clasificación y luego se obtienen tasas promedio de error. La ventaja de esta técnica es que se evalúa la performance del modelo a partir de K combinaciones de prueba y entrenamiento, la desventaja es su lentitud desde el punto de vista computacional.

Validación cruzada quitando una observación ("leaving one out").
En este procedimiento en cada subconjunto se deja una de las observaciones fuera la cual será la muestra de prueba, el resto de las observaciones constituye la muestra de entrenamiento. La evaluación del modelo se obtiene por medio de evaluar el error en cada grupo y luego obteniendo una tasa de error promedio. La desventaja de este procedimiento es su alto costo computacional dado que tenemos tantas subconjuntos o grupos como observaciones tenga la muestra.

4.3.2.3. Técnicas de remuestreo.

Otra de las posibilidades es emplear el remuestreo o bootstrap, el cual consiste en obtener varias muestras con reemplazo de la muestra original. Se pueden obtener tantas muestras como se desee y analizarlas en forma independiente calculando la regla de clasificación y las correspondientes tasas de error en cada una de ellas. Luego se puede calcular el error global, errores promedios o ver cual es el error más frecuente, se construyen de esta forma medidas de error que son más robustas.

A continuación se presenta dos posibles formas de resumir o presentar el poder predictivo del modelo.

Tabla de clasificación.

A efectos de visualizar la performance del modelo se puede construir la siguiente tabla:

Cuadro 4.6: Tabla de observado vs. predicho por el modelo.

predicho / observado	$\hat{Y} = 0$	$\hat{Y} = 1$	Total
$Y = 0$	n_1	n_2	$n_1 + n_2$
$Y = 1$	n_3	n_4	$n_3 + n_4$
Total	$n_1 + n_3$	$n_2 + n_4$	N

A partir de dicha tabla se puede evaluar qué tan bueno es el modelo a la hora de predecir, se puede comparar los valores predichos por el modelo y los observados en la muestra y construir medidas como el porcentaje de aciertos o la tasa de error. Dos medidas que resumen el poder predictivo son la sensibilidad y la especificidad. Se define a la sensibilidad como la $P(\hat{Y} = 1/Y = 1)$ y a la especificidad como la $P(\hat{Y} = 0/Y = 0)$. A través de la tabla de clasificación se puede tener una aproximación a dichas medidas:

$$\text{Sensibilidad} = n_4/(n_3 + n_4), \text{ Especificidad} = n_1/(n_1 + n_2)$$

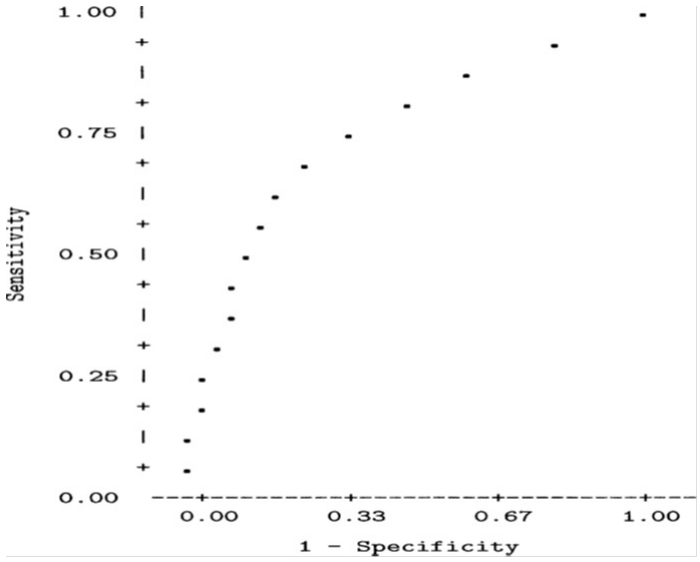
Estas dos medidas reflejan las tasas de acierto del modelo, es decir el porcentaje de coincidencias entre lo observado y predicho por el modelo. En forma análoga también podríamos plantearnos el porcentaje de error del modelo, analizando el porcentaje de observaciones que el modelo clasificó como Y=1 y cuyo valor observado fue Y=0 y al revés.

Curvas ROC.

Una curva ROC (Receiver Operating Characteristic) es un gráfico de la sensibilidad como función de (1 - especificidad) para los posibles puntos de corte π_0. De este modo una curva ROC es más informativa que una tabla de clasificación, dado que resume el poder predictivo del modelo para todos los posibles valores de π_0 . El área bajo la curva ROC es una medida del poder predictivo llamada índice de concordancia, proporciona una medida de la habilidad del modelo para discriminar entre las observaciones que presentan el suceso de interés versus aquellos que no. Se establece como regla que:

• Si el área es 0,5 indica que no existe discriminación.
• Si el área es mayor o igual a 0,7 y menor a 0,8 indica una discriminación y capacidad predictiva aceptable.
• Si dicho valor de área es mayor o igual a 0,8 y menor a 0,9 decimos que el modelo posee muy buena discriminación y capacidad predictiva.

Figura 4.1: Curva ROC.



Fuente: Agresti, A.(2007)

4.4. Árboles de decisión.

Para el desarrollo de la siguiente sección se siguió el libro de Breiman et al.(1984)
Los árboles de decisión son técnicas multivariadas no paramétricas que se utilizan para encontrar reglas de clasificación (predicción), asi como para realizar análisis descriptivos de los datos. El

problema se puede presentar como dado un conjunto de datos D=($\underline{X}$, Y), donde Y es la variable a explicar y $X = (X_1, \ldots, X_k)$ es un conjunto de k características que se le miden a los individuos, el objetivo es predecir el valor de Y a partir de los valores observados de las variables X. Es decir, dado D se construye un predictor $\varphi(\underline{x}, D)$ que permita obtener estimaciones para valores desconocidos de Y.
Los árboles de decisión se pueden clasificar tomando en cuenta la naturaleza de la variable dependiente o a explicar en:

• Árboles de clasificación: la variable dependiente es de naturaleza cualitativa. Se tiene que $Y \in (1, 2, ..., J)$ y lo que se busca es clasificar a los individuos en alguno de los J grupos predeterminados usando k características $(X_1, ..., X_k)$

• Árboles de regresión: la variable dependiente es de naturaleza cuantitativa. Es decir se tiene que $Y \in \mathbf{R}$ y el objetivo es, al igual que en un modelo de regresión, obtener una estimación del valor de $E(Y)$.

Esta técnica, también puede ser utilizada para seleccionar variables, es decir determinar cuáles características son las que mejor definen a las J clases. La relación entre Y y las X consiste en una función constante por conjuntos, los valores predichos de Y pueden expresarse como:

$$E(Y/X = x) = \sum_{s=1}^{S} c_s I_{N_s}(x)$$

Donde N_s representa una partición del espacio de las variables explicativas, I_{N_s} es la indicatriz que vale 1 cuando $x \in N_s$ y 0 en otro caso. La partición es tal que $N_s \cap N_i = \emptyset$, por lo que c_s es la predicción de Y cuando $x \in N_s$. La determinación de la predicción de Y difiere según si el árbol es de clasificación o de regresión.

$$\text{Árbol de clasificación } c_s = max_l \left\{ \frac{card(Y=l)}{card(N_s)} \right\},$$
$$\text{Árbol de regresión } c_s = \frac{1}{card(N_s)} \sum_{i/X_i \in N_s} Y_i$$

En particular los árboles de decisión que se utilizaron en este trabajo son del tipo CART (Classification and Regression Trees, por sus siglas en inglés), que son técnicas de clasificación y regresión del tipo binarias. Estos árboles van generando una partición recursiva del espacio de representación a partir de un conjunto de reglas de decisión. Se parte de un nodo inicial que contiene a todos los datos, luego el nodo se particiona en dos nodos hijos de acuerdo a una regla de decisión, lo que se pretende es que los dos grupos resultantes sean lo más homogéneos posibles en su interior. Esta regla de decisión está basada en una única variable, y la misma se escoge de modo que la partición se haga en dos conjuntos lo más homogéneos posible.

4.4.1. Elementos básicos necesarios en el proceso de construcción del árbol.

El proceso de construcción del árbol es recursivo y se comienza con todo el conjunto de datos de entrenamiento $D = (Y, \underline{X})$. Luego se consideran un conjunto de particiones s, representadas por preguntas binarias del tipo ¿ $x \in Q$?, donde Q es un subconjunto de la muestra y se crea de acuerdo a una regla que toma en consideración una única variable. Luego se utiliza un criterio de bondad de ajuste para evaluar y determinar la mejor partición s, para ello es necesario contar con una medida de la impureza del nodo resultante. Una vez elegida la mejor partición, el conjunto de los datos es dividido en dos subconjuntos, luego en cada uno de los subconjuntos resultante se repite el proceso. El proceso continúa de este modo hasta que se verifica determinada regla de detención previamente definida o hasta que se obtienen nodos puros. A los nodos finales, resultantes de todo el proceso se los denomina terminales. Esta breve descripción da cuenta que en el proceso de partición recursiva se deben tener presente los siguientes elementos:

1. Conjunto de preguntas con respuesta binaria.
2. Criterio de bondad de ajuste de la partición que evalúa en cada nodo t la bondad de la partición s.
3. Regla de detención.
4. Regla para asignar cada nodo terminal a una clase.

4.4.1.1. Conjunto de preguntas con respuesta binaria.

Para cada nodo t se tiene un conjunto de reglas de decisión s. Las particiones se seleccionan según si la variable interviniente es continua o categórica. Si la variable es cuantitativa las reglas son del tipo s: $X_i \leq m$ con m $\in$ R. Existen N posibles divisiones, que consisten en igualar m con cada uno de los valores observados de X_i. Si X_i es de tipo cualitativa con modalidades $c_1, ..., c_L$ las preguntas serán del estilo: $X_i \in C$, siendo C subconjunto de $C = \{c_1, c_2, \ldots, c_l\}$. Con L categorías se definen $2^{L-1} - 1$ reglas para particionar el nodo t. Es decir que para el caso de que la variable interviniente en la decisión sea continua, se verifica si el valor de dicha variable es mayor que cierto valor específico. Si es mayor se sigue el camino de la derecha y si es menor el de la izquierda. Luego este procedimiento se vuelve a repetir en cada nodo, es decir se selecciona una variable y un punto de corte para dividir la muestra en dos partes más homogéneas.

4.4.1.2. Criterio de bondad de ajuste de la partición que evalúa en cada nodo t la bondad de la partición s.

En cada nodo se evalúa la bondad de la partición y se selecciona la mejor de ellas, para ello es necesario tener un criterio que permita evaluar que tan buena es la partición que se genera, este criterio está basado en la medida de impureza del nodo t (I(t)). La impureza de un nodo está asociado a la heterogeneidad presente en la variable dependiente en dicho nodo. Las formas de medir la impureza variará según el tipo de árbol con el que se trabaje, es decir si el mismo es de clasificación o de regresión. Para cada regla s se calcula la caída que se produce en la impureza al utilizar la regla s para dividir t, es decir se considera $\phi(s,t) = I(t) - I(t_i) - I(t_d)$, donde t_i e t_d representan los nodos izquierdo y derecho resultantes de la partición del nodo t. La regla elegida es aquella que maximice $\phi(s,t)$. A modo de ejemplo se presenta una de las formas de definir la impureza de un nodo cuando el árbol es de clasificación:

$$I(t) = -\sum_{g=1}^{G} p(g/t) \log p(g/t) \text{ donde } p(g/t) = \frac{N_j(t)}{N(t)}$$

En este caso $p(g/t)$ representa la probabilidad de que las observaciones que llegan al nodo t pertenezcan a cada una de las clases. Esta es máxima cuando $p(g/t) = \frac{1}{G}$. Por lo tanto la variable a ser seleccionada será la que minimice la heterogeneidad o impureza que resulta de la división del nodo.

4.4.1.3. Regla de detención.

Las reglas de detención son las que permiten declarar a un nodo como terminal y que el proceso de partición se detenga. Por lo general el analista adopta un criterio para decidir cuándo un nodo es lo suficientemente homogéneo y que el proceso se detenga. Si dicho criterio se verifica el nodo será un nodo terminal, en caso contrario será un nodo intermedio. Es habitual el utilizar como criterio de parada el grado de impureza del nodo, y establecer que el proceso se detenga cuando el decrecimiento en el nivel de impureza alcanza determinado umbral, establecido como crítico por el analista. También pueden adoptarse criterios relacionados con la cantidad de elementos o el grado de impureza del nodo.

4.4.1.4. Regla para asignar cada nodo terminal a una clase.

Para asignar cada nodo terminal a una clase nos fijaremos en la frecuencia de las observaciones contenidas en dicho nodo. Es decir que asignaremos todas las observaciones al grupo más probable en dicho nodo, lo que equivale a decir, el grupo con máxima p(g/t).

4.4.2. Proceso de construcción del árbol óptimo.

La construcción del árbol se realiza en dos pasos, primero se construye un "árbol maximal" y luego se procede a la poda para obtener el árbol óptimo. Este procedimiento es válido tanto para árboles de regresión como árboles de clasificación. Si el proceso de partición se continúa hasta el final se obtienen nodos puros. El inconveniente es que el árbol resultante puede ser muy complejo, difícil su interpretación y no tener una buena performance en lo que a predicción se refiere ya que el mismo está muy ajustado a los datos de entrenamiento. Por tal motivo el árbol resultante es podado, cortando sucesivas ramas o nodos terminales hasta encontrar el tamaño adecuado del árbol. Breiman sugiere algunas ideas como forma de determinar el árbol óptimo. Una forma es considerar una secuencia de árboles anidados con el maximal de tamaño decreciente, luego estos son comparados para determinar el óptimo. Esta comparación está basada en una función de costo-complejidad, $R_\alpha(T)$.

Para cada árbol T, la función de costo-complejidad se define como:

$$R_\alpha(T) = R(T) + \alpha T$$

donde R(T) puede ser la tasa de error global o la suma de cuadrados residuales total dependiendo del tipo de árbol, T es la complejidad del árbol, entendida como el número de nodos del subárbol y α es el parámetro de complejidad. De la secuencia de árboles anidados es necesario seleccionar el árbol óptimo para ello se evalúa tanto la complejidad como el poder predictivo del árbol resultante. Para evaluar el poder predictivo de la secuencia de árboles anidados se emplea un procedimiento de partición de la muestra o de validación cruzada. En el mismo se emplea una parte de la muestra para construir la regla de clasificación y la otra parte se emplea para predecir la variable de respuesta. Luego se obtienen los errores de predicción y se selecciona el árbol con el menor error de predicción.

Capítulo 5

Estadística Univariada.

5.1. Datos utilizados.

Para el desarrollo del presente trabajo se utilizaron como datos los de la ECH2011, elaborada por el INE. Para el año 2011 esta encuesta de cobertura nacional reúne información de 46.669 viviendas seleccionadas mediante un muestreo probabilístico estratificado bietápico.

La muestra se selecciona en dos etapas: zona censal y vivienda particular, y es independiente mes a mes. En cada departamento, las unidades primarias de muestreo son las zonas censales (manzanas o territorio identificable), seleccionadas con probabilidad proporcional al tamaño medido en número de viviendas particulares. Las unidades secundarias de muestreo son las viviendas particulares dentro de cada zona.

Se presenta información de 130.804 personas que hacen referencia a 46.669 hogares. Esta información describe características socio-demográficas de las personas (edad, sexo, región en que habita, nivel de instrucción, etc.), y a su vez se miden características de la vivienda, ingresos percibidos e ingresos del hogar, participación económica y actividad laboral entre otros.

El objetivo central del presente proyecto es el análisis del fenómeno de la asistencia al Sistema Educativo, en particular de los jóvenes uruguayos de 14 a 17 años. Del análisis de los datos de ECH previas como la del 2009 y 2010, se aprecia que la no asistencia a centros educativos por parte de los jóvenes se ha mantenido en el entorno del 20 %, lo cual marca una persistencia de dicho fenómeno.

5.2. Los jóvenes uruguayos y la asistencia al sistema educativo.

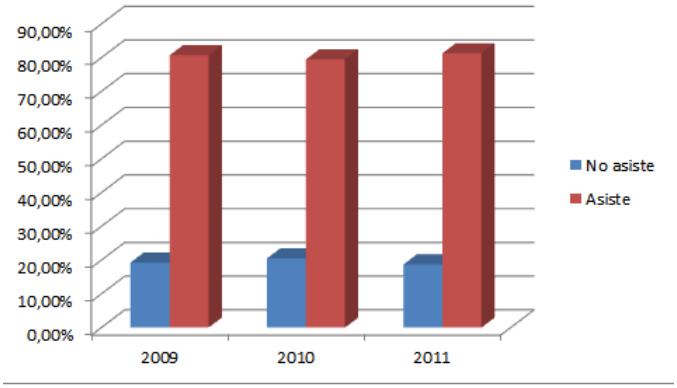
A continuación se presenta un conjunto de estadísticas univariadas que reflejan el comportamiento de la variable asistencia para determinadas características de los jóvenes y de los hogares donde residen.
Los datos fueron expandidos a toda la población de jóvenes de 14 a 17 años de edad.
En el siguiente cuadro y figura se aprecia la evolución que ha tenido la asistencia para los jóvenes de entre 14 y 17 años de edad en los años 2009, 2010 y 2011 respectivamente.

Cuadro 5.1: Asistencia y no asistencia a la educación de jóvenes de 14 a 17 por año.

Asistencia	2009		2010		2011	
	Frecuencia	Porcentaje	Frecuencia	Porcentaje	Frecuencia	Porcentaje
No Asiste	39.169	19,3 %	49.559	20,5 %	42.773	18,7 %
Asiste	164.034	80,7 %	192.553	79,5 %	182.206	81,3 %
Total	203.203	100 %	242.112	100 %	224.979	100 %

Fuente: ECH2009-2010-2011, elaboración propia

Figura 5.1: Asistencia y no asistencia a la educación de jóvenes de 14 a 17 años por año.



Fuente: ECH2009-2010-2011, elaboración propia

5.2.1. La asistencia según sexo y edades simples.

Si se analiza la asistencia según edades simples para jóvenes de 14 a 17 años, se aprecia un aumento progresivo en el porcentaje de no asistencia conforme aumenta la edad del joven, pasando

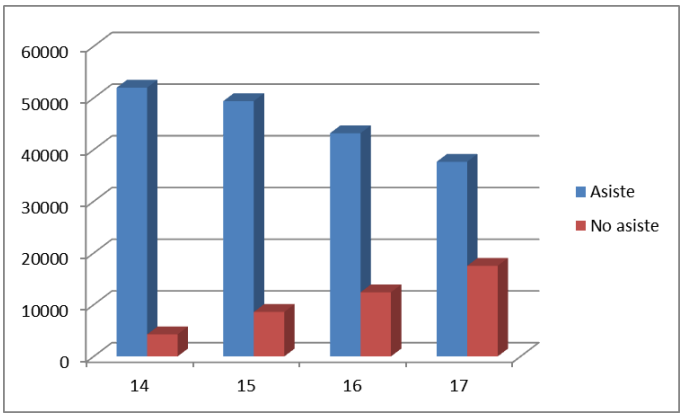
de un valor de 7,6 % en jóvenes de 14 años a un porcentaje de 31,8 % para los que poseen 17 años de edad. Este resultado se encuentra en línea con los hallazgos realizados por Casacuberta y Buchelli (2008) que señalan un aumento progresivo en los niveles de no asistencia para el tramo de edad 14-17 años conforme aumenta la edad del joven.

Cuadro 5.2: Asistencia a la educación de jóvenes de 14 a 17 según edades simples.

Edad	Asistencia				
	Asiste	Porc.	No asiste	Porc.	Total
14	51.986	92,4 %	4.253	7,6 %	56.239
15	49.391	85,2 %	8.598	14,8 %	57.989
16	43.174	77,7 %	12.399	22,3 %	55.573
17	37.655	68.2 %	17.523	31.8 %	55.178
Total	182.206	80,1 %	42.773	19,9 %	224.979

Fuente: ECH2011, elaboración propia

Figura 5.2: Asistencia a la educación de jóvenes de 14 a 17 según edades simples.



Fuente: ECH2011, elaboración propia

5.2.2. Asistencia según el sexo.

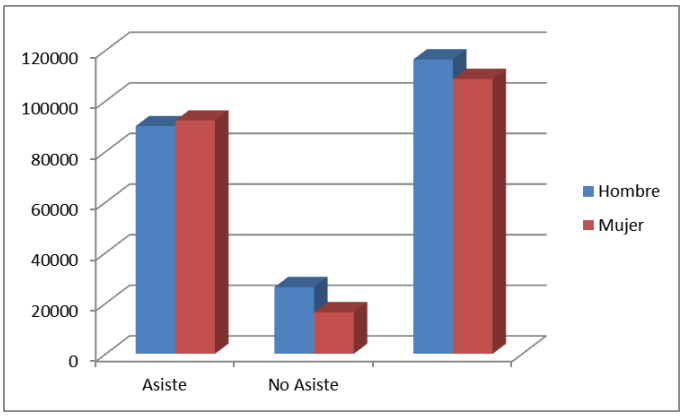
Si se analiza la asistencia al sistema educativo según el sexo del joven, encontramos un comportamiento diferenciado. Se aprecia un mayor porcentaje de no asistencia para los varones (22, 7 %) que para las mujeres, donde la no asistencia es de 15, 2 %. Este resultado se encuentra en línea con la hipótesis de partida del presente trabajo que señalaba un comportamiento diferencial en cuanto al sexo y una mayor propensión de las mujeres a asistir al sistema educativo.

Cuadro 5.3: Asistencia a la educación de jóvenes de 14 a 17 por sexo.

Asistencia	Sexo			
	Hombre	Porc.	Mujer	Porc.
Asiste	89.989	77,3 %	92.217	84,8 %
No asiste	26.347	22,7 %	16.426	15,2 %
Total	116.336	100 %	108.643	100 %

Fuente: ECH2011, elaboración propia

Figura 5.3: Asistencia a la educación de jóvenes de 14 a 17 por sexo.



Fuente: ECH2011, elaboración propia

5.2.3. Asistencia según región.

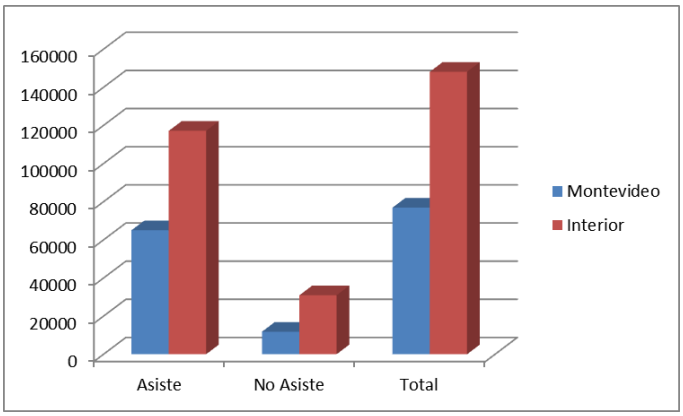
Analizando la asistencia por región se observan también diferencias en cuanto al lugar de residencia de los jóvenes uruguayos. El porcentaje de los jóvenes de entre 14 a 17 años que no asisten es mayor para aquellos que residen en el Interior del país (20, 9 %) que para los que residen en la capital, el cual asciende a un 15, 4 % del total. Esto se podría deber principalmente a la proximidad del centro educativo más cercano (dificultades de acceso), la mayor oferta educativa existente en la capital, así como una vinculación más temprana con el mercado de trabajo.

Cuadro 5.4: Asistencia a la educación de jóvenes de 14 a 17 por región.

Asistencia	Región			
	Montevideo	Porc.	Interior	Porc.
Asiste	65.037	84,6 %	117.169	79,1 %
No Asiste	11.853	15,4 %	30.920	20,9 %
Total	76.890	100 %	148.089	100 %

Fuente: ECH2011, elaboración propia

Figura 5.4: Asistencia a la educación de jóvenes de 14 a 17 por región.



Fuente: ECH2011, elaboración propia

5.2.4. Los años acumulados de educación.

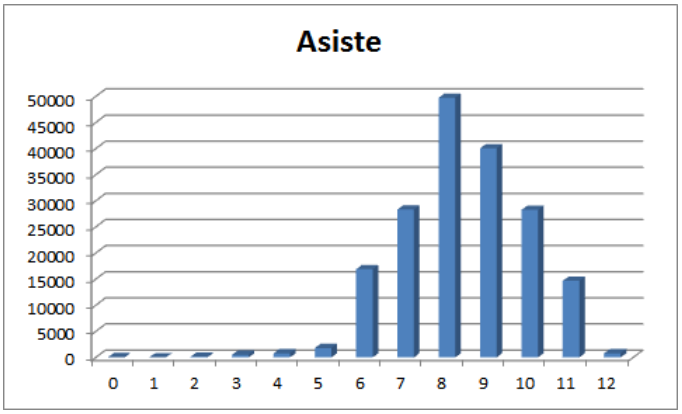
Tomando en cuenta los años de educación acumulados por el joven, en los siguientes cuadros se puede apreciar como conforme aumentan los años de educación acumulados aumenta la proporción de adolescentes que asisten al sistema educativo. Esto se verifica hasta alcanzar los ocho años de educación donde comienza a descender. Lo anterior se condice con estudios que señalan un descenso de los niveles de asistencia en los jóvenes que se encuentran cursando bachillerato, es decir aquellos que tienen entre 14 y 17 años de edad y han completado el ciclo básico (9 años de educación). Del mismo modo se destaca como a partir de los 7 años de educación acumulados por el joven, la proporción de los que asisten es mayor que la de los que no asisten.

Cuadro 5.5: Asistencia de jóvenes de 14 a 17 a centros educativos por años de educación acumulados.

Asistencia	Años de educación acumulados												
	0	1	2	3	4	5	6	7	8	9	10	11	12
Asiste	53	35	160	527	801	1.804	16.903	28.343	49.768	40.076	28.261	14.679	796
No Asiste	209	83	229	362	1390	888	20.027	8.097	6.234	3.622	895	225	472
Total	262	118	389	889	2.191	2.692	36.930	36.440	56.002	43.698	29.156	14.904	1.268

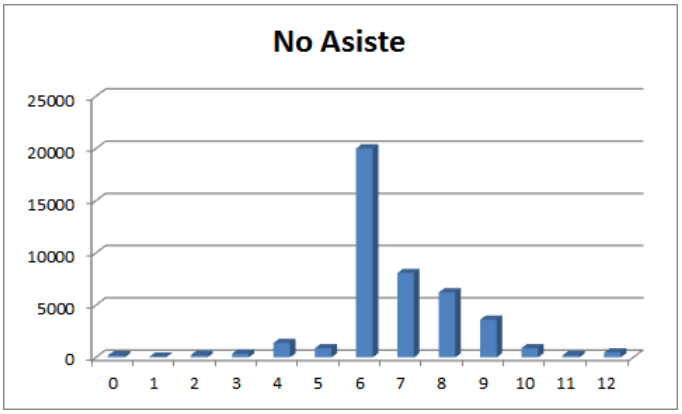
Fuente: ECH2011, elaboración propia

Figura 5.5: Asistencia de jóvenes de 14 a 17 a centros educativos por años de educación acumulados.



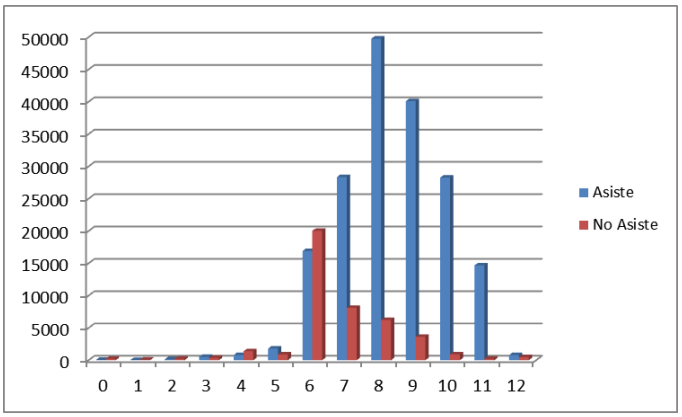
Fuente: ECH2011, elaboración propia

Figura 5.6: No Asistencia de jóvenes de 14 a 17 a centros educativos por años de educación acumulados.



Fuente: ECH2011, elaboración propia

Figura 5.7: Asistencia y No Asistencia de jóvenes de 14 a 17 a centros educativos por años de educación acumulados.



Fuente: ECH2011, elaboración propia

5.2.5. Asistencia según la inserción en el mercado de trabajo.

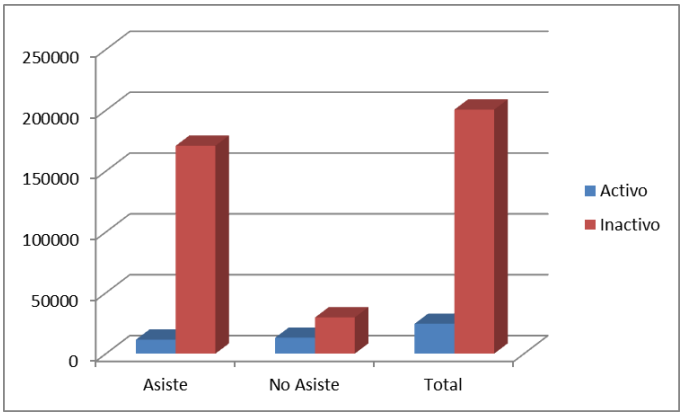
Una de las hipótesis de la presente investigación se relaciona con el hecho de que si el joven es activo tiene menos probabilidad de asistir al sistema educativo, respecto de aquellos que no trabajan. Los datos analizados operan en el sentido esperado, se aprecia que el porcentaje de jóvenes que no asisten (53 %) es mayor entre aquellos que son activos que entre los inactivos (14,82 %).

Cuadro 5.6: Asistencia de jóvenes de 14 a 17 a centros educativos por condición de actividad.

Asistencia	Condición de actividad del joven			
	Activo	Porc.	No activo	Porc.
Asiste	11.450	46,7 %	170.756	85,18 %
No asiste	13.054	53,3 %	29.719	14,82 %
Total	24.504	100 %	200.475	100 %

Fuente: ECH2011, elaboración propia

Figura 5.8: Asistencia de jóvenes de 14 a 17 a centros educativos por condición de actividad.



Fuente: ECH2011, elaboración propia

5.2.6. Región, asistencia y vinculación con el mercado de trabajo.

Cuadro 5.7: Asistencia de jóvenes de 14 a 17 a centros educativos en Montevideo por condición de actividad.

Asistencia	Condición de actividad del joven			
	Activo	Porc.	No activo	Porc.
Asiste	2.986	50,6 %	62.051	87,5 %
No Asiste	2.911	49,4 %	8.942	12,5 %
Total	5.897	100 %	70.993	100 %

Fuente: ECH2011, elaboración propia

Cuadro 5.8: Asistencia de jóvenes de 14 a 17 a centros educativos en el interior del país por condición de actividad.

Asistencia	Condición de actividad del joven			
	Activo	Porc.	No activo	Porc.
Asiste	8.464	45,5 %	108.705	84 %
No Asiste	10.143	54,5 %	20.777	16 %
Total	18.607	100 %	129.482	100 %

Fuente: ECH2011, elaboración propia

Si se analiza los datos por región y se los controla por la condición de actividad del joven se encuentra un mayor porcentaje de jóvenes activos en el interior del país que en Montevideo, esto se podría relacionar con estudios que hablan de una vinculación más temprana con el mercado laboral en los jóvenes que residen en el interior del país. En lo que respecta a la asistencia, si se mira a los jóvenes activos, se aprecia un mayor porcentaje de no asistencia en el interior (54,5 %) respecto a un 49,4 % para los que residen en la capital. La doble condición de residir en el interior y ser activo tendría un mayor impacto al momento de decidir no asistir si se compara con un joven que reside en la capital y que es activo.

5.2.7. Asistencia según tenencia de hijos a cargo.

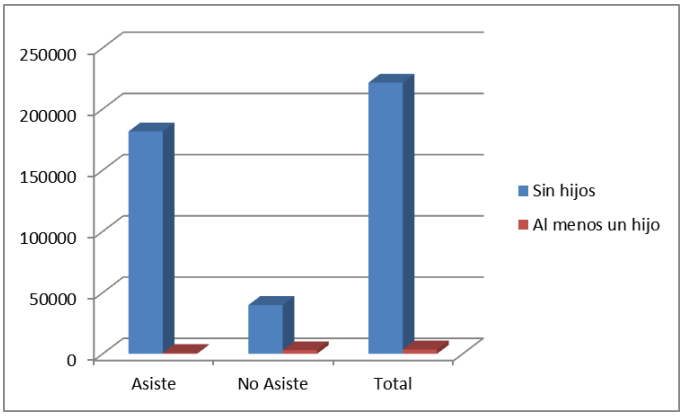
Otra de la hipótesis del presente trabajo vincula la asistencia con la tenencia de hijos a cargo. La hipótesis de partida establecía un vínculo negativo entre ambas, es decir la presencia de hijos a cargo tiene un impacto negativo en la asistencia de los jóvenes. En los datos analizados se observa un mayor porcentaje de no asistencia entre los jóvenes que poseen al menos un hijo a su cargo (83 %) de aquellos que no (18 %).

Cuadro 5.9: Asistencia de jóvenes de 14 a 17 a centros educativos por la tenencia de hijos a cargo.

Asistencia	Hijos a cargo			
	Sin hijos	Porc.	Al menos un hijo a cargo	Porc.
Asiste	181.606	82 %	600	17 %
No asiste	39.875	18 %	2.898	83 %
Total	221.481	100 %	3.498	100 %

Fuente: ECH2011, elaboración propia

Figura 5.9: Asistencia de jóvenes de 14 a 17 a centros educativos por la tenencia de hijos a cargo.



Fuente: ECH2011, elaboración propia

5.2.8. Asistencia según ascendencia de los jóvenes.

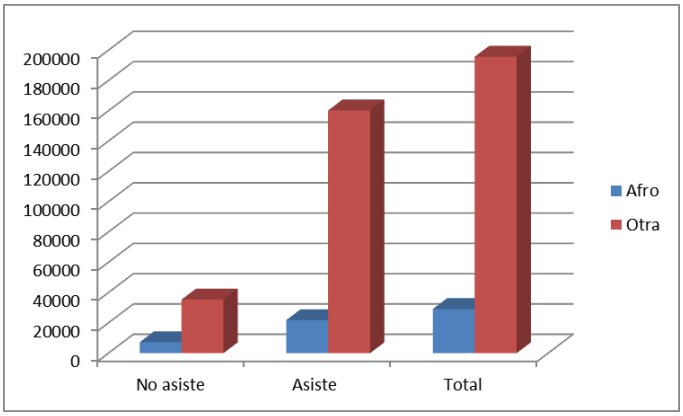
Analizando los niveles de no asistencia según la ascendencia de los jóvenes, se encuentra que existe un comportamiento que es disímil. Los datos reflejan que para el tramo etario bajo estudio, el porcentaje de jóvenes afrodescendientes que no asisten al sistema educativo (25 %) es más elevado que el porcentaje hallado para los jóvenes cuya ascendencia cae en la categoría otra, el cual es de 18 % (esta última incluye a aquellos jóvenes cuya ascendencia es blanca). Sería de esperar entonces, un signo negativo para la variable afro en el modelo de regresión logística que intenta explicar la probabilidad de asistencia a un centro educativo.

Cuadro 5.10: Asistencia de jóvenes de 14 a 17 a centros educativos por ascendencia.

Asistencia	Ascendencia			
	Afro	Porc.	Otra	Porc.
No asiste	7.322	25 %	35.451	18 %
Asiste	21.882	75 %	160.324	82 %
Total	29.204	100 %	195.775	100 %

Fuente: ECH2011, elaboración propia

Figura 5.10: Asistencia de jóvenes de 14 a 17 a centros educativos por ascendencia.



Fuente: ECH2011, elaboración propia

5.3. La asistencia a centros educativos y las características de los hogares.

5.3.1. El clima educativo del hogar.

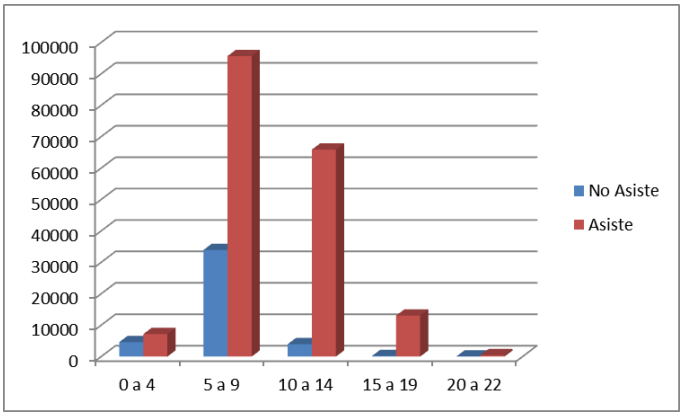
En lo que respecta al clima educativo del hogar, entendido como el promedio de años de educación de los adultos del hogar (en su defecto el jefe del hogar), la literatura establece un impacto positivo en la asistencia. Se señala que en estos hogares los padres cuentan con mejores herramientas como para seguir e intervenir en el proceso educativo de sus hijos. La hipótesis inicial respecto al impacto que tiene el clima educativo en los niveles de asistencia se encuentra en línea con este pensamiento. Del análisis de los datos, surge que para todos los tramos en los cuales se agrupó la variable clima educativo, se encuentra que es mayor el porcentaje de asistencia que de no asistencia. Es decir que los datos analizados estarían operando en el sentido señalado por la teoría que señala que el clima educativo del hogar donde habitan los jóvenes tiene un impacto positivo en los niveles de asistencia. El mayor porcentaje de no asistencia se verifica para aquellos hogares que poseen entre 0 a 9 años en promedio de educación. Del mismo modo se verifica que a medida que aumenta el clima educativo del hogar disminuye la no asistencia de los jóvenes.

Cuadro 5.11: Asistencia de jóvenes de 14 a 17 a centros educativos por promedio de años de educación de los adultos del hogar.

Asistencia	Promedio de años de educación de los adultos del hogar									
	0 a 4	Porc.	5 a 9	Porc.	10 a 14	Porc.	15 a 19	Porc.	20 a 22	Porc.
No Asiste	4.618	39,2 %	33.933	26,19 %	3.989	5,7 %	208	1,6 %	25	5,15 %
Asiste	7.160	60,8 %	95.655	73,81 %	65.881	94,3 %	13.050	98,4 %	460	94,85 %
Total	11.778	100 %	129.588	100 %	69.870	100 %	13.258	100 %	485	100 %

Fuente: ECH2011, elaboración propia

Figura 5.11: Asistencia de jóvenes de 14 a 17 a centros educativos por promedio de años de educación de los adultos del hogar.



Fuente: ECH2011, elaboración propia

5.3.2. El impacto del hacinamiento.

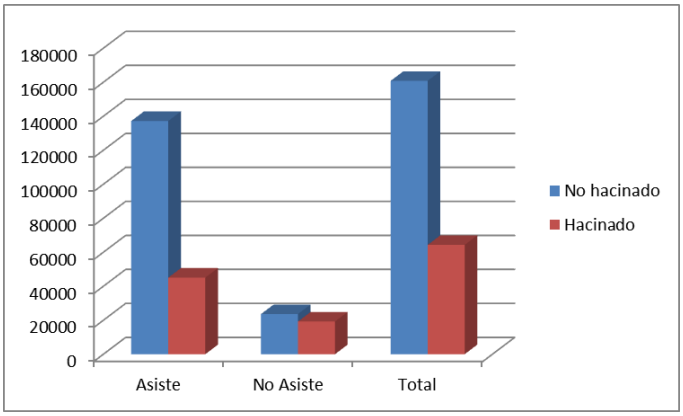
Otra de las hipótesis de partida del presente trabajo fue considerar un vínculo negativo entre el hacinamiento y la asistencia. Aquellos jóvenes que habitan en hogares en condiciones de hacinamiento presentarían mayores niveles de no asistencia que los que no se encuentran hacinados.

Cuadro 5.12: Asistencia de jóvenes de 14 a 17 a centros educativos por situación de hacinamiento en el hogar.

Asistencia	Hacinamiento			
	No hacinado	Porc.	Hacinado	Porc.
No Asiste	23.573	14,7 %	19.200	30 %
Asiste	137.153	85,3 %	45.053	70 %
Total	160.726	100 %	64.253	100 %

Fuente: ECH2011, elaboración propia

Si se analiza el cuadro anterior se aprecia que los datos aquí analizados operan en el sentido esperado, es decir, el porcentaje de no asistencia es mayor para aquellos jóvenes que habitan en hogares en condiciones de hacinamiento (30 %) que aquellos que no lo están, donde el porcentaje de no asistencia es de 14.7 %.



Fuente: ECH2011, elaboración propia

5.3.3. Asistencia, hacinamiento y región de residencia.

Cuadro 5.13: Asistencia de jóvenes de 14 a 17 en el interior del país por situación de hacinamiento del hogar.

Asistencia	Hacinamiento	
	No hacinado	Hacinado
No Asiste	54 %	46 %
Asiste	73 %	27 %
Total	69 %	31 %

Fuente: ECH2011, elaboración propia

Cuadro 5.14: Asistencia de jóvenes de 14 a 17 en Montevideo por situación de hacinamiento del hogar.

Asistencia	Hacinamiento	
	No hacinado	Hacinado
No Asiste	58 %	42 %
Asiste	79 %	21 %
Total	76 %	24 %

Fuente: ECH2011, elaboración propia

Si se controlan los datos por región de residencia se encuentra que el hacinamiento en jóvenes de entre 14 a 17 años de edad resulta mayor en el interior del país que en la capital. Por otra parte el porcentaje de no asistencia es mayor en el Interior (30 %) que en Montevideo (27 %) si

se considera solamente a los jóvenes que habitan en hogares con condiciones de hacinamiento. Es decir que habría un doble impacto negativo en este caso por el hecho de residir en el interior del país y vivir en condiciones de hacinamiento.

5.3.4. La asistencia y el indicador de calidad de la vivienda.

El indicador de calidad de vivienda (ICV), es un indicador que resume determinadas características de la vivienda en cuanto a la calidad de la misma y el acceso a determinados servicios. En el mismo se toman en cuenta los materiales de construcción de la vivienda presentes en pisos, paredes y techos, así como el acceso al agua potable, saneamiento, y energía eléctrica. Los datos reflejan que cuánto más precaria son las condiciones estructurales de la vivienda o más deficitario sea el acceso al agua potable, energía eléctrica y saneamiento, mayores son los niveles de no asistencia de los jóvenes que habitan en viviendas con estas características. Al analizar el cuadro que se presenta a continuación se aprecia que aquellos jóvenes que habitan en viviendas cuya calidad es deficitaria, presentan un porcentaje de no asistencia que asciende al 59.6 % del total. En comparación, los que habitan en viviendas cuya calidad es regular, presentan un porcentaje de no asistencia del 18.8 %. Se aprecia un fuerte impacto de la calidad de la vivienda en la decisión de no asistir por parte de los jóvenes con una diferencia de 40.8 % en los niveles de no asistencia.

Cuadro 5.15: Asistencia de jóvenes de 14 a 17 a centros educativos por calidad de la vivienda.

Asistencia	Buena o regular	Porc.	Déficit	Porc.
No asiste	41.964	18,8 %	809	59,6 %
Asiste	181.657	81,2 %	549	40,4 %
Total	223.621	100 %	1.358	100 %

Fuente: ECH2011, elaboración propia

Capítulo 6

Resultados.

6.1. Discusión.

Para el desarrollo de la presente sección se tomó como fundamento la exposición de Damonte et al. "Análisis de datos provenientes de diseños muestrales complejos: Aplicaciones a la Encuesta de Hogares y Empleo de la Provincia de Bs. As."

Las encuestas de hogares que llevan adelante las Oficinas de Estadísticas Oficiales, presentan por lo general diseños muestrales que se denominan complejos. Al contar con información respecto al diseño se plantea si es necesario incorporar la misma al análisis o no.

En Estadística es frecuente que se presente una discusión en cuanto a tres enfoques de inferencia: a) el enfoque de inferencia basado en el diseño de la muestra; b) el basado en modelos superpoblacionales y c) la inferencia asistida por modelos.

En el enfoque de inferencia basado en modelos superpoblacionales, los valores de la población finita $(y_1, ..., y_n)$, se consideran realizaciones de un vector aleatorio $(Y_1, ..., Y_n)$, donde Y_i es una variable aleatoria que sigue determinada distribución. Los valores de la población se consideran provenientes de una superpoblación con distribución $p(Y/\theta)$, con parámetros fijos θ. La estimación se realiza en base a la frecuencia de distintas realizaciones del vector aleatorio y la misma está condicionada a una sola muestra, que es la realizada, y no a otras muestras posibles.

En el enfoque de inferencia basado en el diseño, los valores $(y_1, ..., y_n)$ de la característica Y que se evalua para cada miembro de la población de tamaño N, son considerados valores fijos y no aleatorios. La aleatorieadad proviene de considerar todas las muestras posibles de tamaño

n, y una distribución de probabilidad que se define sobre este conjunto de muestras posibles. Entran en juego bajo este enfoque las probabilidades de inclusión de los diferentes individuos en la población.

Existen autores que plantean que en la estimación de modelos no sería necesaria la incorporación de la información contenida en el diseño muestral (forma en que se extrajo la muestra, ponderaciones, etc.). Por otra parte, otros plantean la necesidad de incorporar la información respecto al diseño si es que se cuenta con dicha información. También existen otros que señalan que se debería llevar adelante ambos enfoques con y sin información del diseño y efectuar una comparación, si se arriba a modelos similares estaría indicando que no sería necesario incorporar tal información.

Existen varios artículos que hablan sobre la importancia de no ignorar dicha información cuando se dispone de la misma, al respecto Damonte et al. señalan lo siguiente: "el costo de ignorar la información de diseño puede conducir a errores que pueden ser graves, en términos de modificar las conclusiones de inferencia, y en este sentido se recomienda contemplar dicha información, aún cuando implique un costo en términos de eficiencia de los estimadores por el aumento en el estimador de la varianza de diseño respecto al estimador basado en un modelo."

Dado que en el presente trabajo se trabajan con datos provenientes de la ECH2011, la cual posee un diseño muestral complejo, y que se dispone de dicha información, se optó por incorporarla a la estimación del modelo de la asistencia para los jóvenes uruguayos de 14 a 17 años de edad. Se utiliza la función svyglm contenida en el paquete 'survey' del software R. Esta función fue desarrollada por Thomas Lumley, quien recomienda su uso bajo diseños muestrales complejos.

Por otra parte se debe mencionar que en la mayoría de los antecedentes analizados sobre la temática de la desafiliación (Casacuberta y Buchelli, 2010; Ferrari, Martínez y Saavedra, 2010) los autores han optado por modelizar la asistencia a un centro educativo mediante un modelo probit-bivariado. Se adopta esta estrategia como forma de analizar o modelizar en forma conjunta la asistencia y la participación en el mercado de trabajo por parte de los jóvenes. La mayoría de los autores revisados parten del supuesto de la existencia de una estrecha relación entre no asistir a un centro educativo y participar en el mercado de trabajo por tal razón es que estiman dos ecuaciones en forma simultánea, una para la asistencia y otra para la participación del joven en el mercado de trabajo. En este trabajo se busca determinar cuáles características son las que determinan la asistencia de un joven a un centro educativo y no su vinculación con la participación en el mercado de trabajo, por ello se optó por trabajar con modelos uniecuacionales.

6.2. Resultados obtenidos.

En este capítulo se exponen los principales resultados del análisis empírico. El capítulo se estructura en tres secciones que siguen el orden de las etapas planteadas en la metodología, utilizando como herramienta de cálculo el paquete estadístico R (ver programa en anexo). En la primera sección se presentan los resultados del cálculo de indicadores pertinentes; el Índice de Vivienda y Salubridad (VIV), el Índice de Calidad de la Vivienda (ICV), el Índice de Confort, y por último un Índice de Tecnología de la Información y Comunicaciones (TICS). En la segunda sección se analizan las variables que discriminan la asistencia de la no asistencia, se analizan distintas metodologías de selección de variables como son "backward" y "test de razón de verosimilitud" para la estimación del mejor modelo que represente la asistencia de los jóvenes al sistema educativo. Por último, en la tercera sección se analiza el poder predictivo de los modelos seleccionados. Se construyen, como forma de complementar los modelos anteriores, árboles de clasificación para la variable asistencia y se compara los resultados con los modelos planteados anteriormente.

6.2.1. Indicadores pertinentes.

6.2.1.1. Índice de vivienda y salubridad (VIV) mediante ACM.

Con el fin de resumir información sobre las características de la vivienda donde vive el joven en una única dimensión y así poder observar como interviene esta dimensión en la asistencia, se construye el Índice de Vivienda y Salubridad (VIV). Para calcular este indicador se empleó Análisis de Correspondencia Múltiple (ACM) considerando las siguientes variables:

Cuadro 6.1: Variables para el indicador VIV mediante ACM.

Variables	Descripción
Agua en red general	vale 1 si tiene agua en red gral., 0 si no tiene
Tiene baño con cisterna	vale 1 si tiene baño con cisterna, 0 si no tiene
Evacuación en red general	vale 1 si tiene evacuación en red gral., 0 si no tiene
Techo	vale 0 si el techo es de quincha o material de desecho, 1 en otros casos
Piso	vale 0 si el piso es solo contrapiso sin piso o de tierra sin piso ni contrapiso, 1 en otros casos
Paredes	vale 0 si el piso es de material liviano sin revestimiento, de adobe o material de desecho, 1 en otros casos
Fuente para iluminar	vale 1 si es energía eléctrica, 0 en otros casos

Fuente: ECH2011, elaboración propia

El objetivo del presente trabajo es obtener un indicador que resuma la información de las dimensiones de vivienda y salubridad, debido a esto se va a considerar únicamente el primer eje factorial. Al observar los resultados de este análisis, se encuentra que existen dos modalidades "raras" ("Pared No" y "Fuente iluminar No") que se ubican lejos del resto de la nube de las modalidades. Estas modalidades "raras" presentan frecuencias extremas y se ubican por ese motivo lejos del resto de la nube. El 99 % de los hogares con al menos un joven de 14 a 17 años tiene como fuente para iluminar energía eléctrica, mientras que únicamente el 1 % de estos hogares no tiene como fuente para iluminar energía eléctrica. En referencia a la calidad de las paredes, el 99 % de los hogares donde residen estos jóvenes tienen paredes de calidad regular o buena, mientras que únicamente el 1 % de los hogares tiene materiales de mala calidad en las paredes.

Cuadro 6.2: Frecuencias de variables para el indicador VIV

Variables	No	Si	Total
Agua en red general	8 %	92 %	100 %
Tiene baño con cisterna	10 %	90 %	100 %
Evacuación en red general	47 %	53 %	100 %
Techo	14 %	86 %	100 %
Piso	25 %	75 %	100 %
Paredes	1 %	99 %	100 %
Fuente para iluminar	1 %	99 %	100 %

Fuente: ECH2011, elaboración propia

Estas frecuencias extremas conducen a que estas modalidades (Fuente para iluminar No y Pared No) interfieran en el análisis aumentando la amplitud del indicador innecesariamente. Por este motivo, se decide colocarlas como suplementarias y correr nuevamente el análisis.

Mediante la técnica factorial de ACM se obtuvieron los ejes de inercia, las proyecciones de las diferentes modalidades, las contribuciones de las diferentes modalidades a la inercia de cada factor y las respectivas masas y cosenos cuadrados.

Al tomar el total país, se aprecia que muchas de las modalidades son baricéntricas dificultando la interpretación de los resultados. Se adoptó como estrategia el realizar un análisis diferenciado para el interior y otro para la capital como forma de evaluar si los resultados obtenidos se podían deber a diferencias subyacentes en cuanto a la región. Como siguiente paso se elaboró un índice sólo para los hogares del interior y otro exclusivamente para Montevideo y se analizaron los resultados. De los resultados del ACM se aprecia que el Interior presenta un comportamiento muy similar a lo que es el total del país y por otra parte Montevideo un comportamiento diferenciado con una gran cantidad de modalidades que son baricéntricas.

Debido a la dificultad en la interpretación que significaba el trabajar con un Índice en donde la mayoría de las modalidades son baricéntricas, se optó por descartarlo y elaborar un Índice de Calidad de la Vivienda (ICV) utilizando la metodología del Reporte Social 2009. Las salidas y gráficos del ACM se encuentran adjuntas en el Anexo de resultados.

6.2.1.2. Índice de Calidad de la Vivienda (ICV), metodología del Reporte Social 2009.

Como se detalló en la metodología, se construye un Índice de Calidad de la Vivienda utilizando la metodología del Reporte Social 2009. En este reporte, se presenta el ICV con tres categorías realizando una transformación, fusionando las primeras dos categorías del mismo asumiendo únicamente dos valores finales; 1 si la vivienda presenta déficit de materiales y/o de servicios, y 0 si la calidad de los materiales y servicios de la vivienda son regulares o buenos. Obteniendo así un indicador de vivienda de los hogares donde vive al menos un joven de 14 a 17 años.

Este último indicador discrimina claramente los jóvenes que viven en viviendas de buena calidad o regular de los que viven en viviendas precarias. Estos últimos representan el 0,6 % del total de los hogares con al menos un joven de 14 a 17 años.

6.2.1.3. Índice de confort.

En referencia a los bienes de confort, el propósito consiste en buscar nuevamente un único factor capaz de explicar el máximo de información contenida en los datos con el fin de observar como interviene esta dimensión de confort del hogar en la asistencia por parte del joven. Para eso, se construye un Índice de Confort utilizando ACM con las siguientes variables explicativas:

Cuadro 6.3: Variables para el índice de confort.

Variables	Descripción
Calefón	1 si el hogar tiene calefón, 0 si no tiene
Refrigerador	1 si el hogar tiene refrigerador, 0 si no tiene
Lavarropas	1 si el hogar tiene lavarropas, 0 si no tiene
Secadora	1 si el hogar tiene secadora, 0 si no tiene
Lavavajillas	1 si el hogar tiene lavavajillas, 0 si no tiene
Microondas	1 si el hogar tiene microondas, 0 si no tiene
Aire	1 si el hogar tiene aire, 0 si no tiene
DVD	1 si el hogar tiene DVD, 0 si no tiene
Auto o moto	1 si el hogar tiene Auto o Moto, 0 si no tiene

Fuente: ECH2011, elaboración propia

En un principio se aplicó ACM a las ocho variables explicativas, como el objetivo es obtener un factor que resuma la máxima información posible sobre los bienes de confort del hogar en el que

vive el joven, se analizó únicamente el comportamiento de las modalidades sobre el primer eje factorial. El gráfico de las modalidades en el plano principal se puede observar en el anexo de resultados.

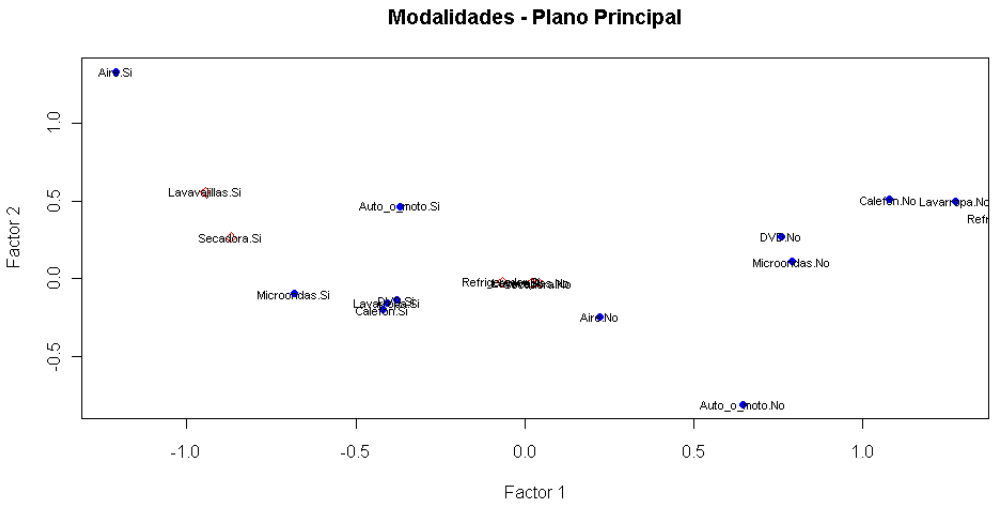
Cuadro 6.4: Tenencia de bienes de confort.

Tenencia de bienes de confort			
Variable	No	Si	Total
Calefón	28 %	72 %	100 %
Refrigerador	3 %	97 %	100 %
Lavarropas	24 %	76 %	100 %
Secadora	93 %	7 %	100 %
Lavavajillas	96 %	4 %	100 %
Microondas	46 %	54 %	100 %
Aire	85 %	15 %	100 %
DVD	33 %	67 %	100 %
Auto o moto	36 %	64 %	100 %

Fuente: ECH2011, elaboración propia

Se observa que existen tres modalidades "raras" ("Refrigerador No", "Secadora Si" y "Lavavajilla Si") que se ubican lejos del resto de la nube de las modalidades. Estas modalidades "raras" presentan frecuencias extremas en la muestra y por ese motivo se ubican lejos del resto de la nube. Por ejemplo "Refrigerador No", presenta una distribución del: 97 %-3 % indicando que la mayoría de los hogares tienen refrigerador mientras que muy pocos no tienen. De este modo, la variable "Refrigerador" puede estar distorsionando los resultados. Sucede algo similar con las variables "Secadora" y "Lavavajillas", se observa nuevamente frecuencias extremas, en este caso, existen pocos hogares que poseen estos bienes por lo que al igual que "Refrigerador", pueden estar interfiriendo en nuestro análisis. Se decidió entonces, correr un nuevo ACM considerando estas tres variables como suplementarias.

Figura 6.1: Índice de confort con variables suplementarias



Fuente: ECH2011, elaboración propia

Se obtuvo como resultado que el primer eje factorial separa a la derecha del cero las modalidades "No tiene" y a la izquierda "Si tiene". Siendo esta la distribución de las modalidades en el primer factor, observamos que cuanto más chico el indicador, el hogar se encuentra en una mejor situación en referencia a los bienes de confort que posee, y por el contrario, cuanto más grande, el hogar se encuentra en una peor situación. Este indicador tiene un mínimo de –0,93260 y un máximo de 1,28300.

Observando la descomposición de la inercia, se encuentra que con un factor la proporción de la inercia es muy baja, cercana al 29 %. Apelando a la transformación de Benzecri, que reponderá los valores propios, con sólo retener un valor se obtiene una importante proporción de inercia (98 %). Los resultados se pueden observar en el anexo de resultados.

6.2.1.4. Índice de Tecnologías de la Información y la Comunicación (TIC)

Para realizar este índice se utiliza el mismo procedimiento que para el Índice de Confort. En un principio se realiza un ACM con las siguientes variables:

Cuadro 6.5: Variables para el índice de TIC.

Variable	Descripción
TV	1 si el hogar tiene TV, 0 si no tiene
Cable	1 si el hogar tiene cable, 0 si no tiene
Internet	1 si el hogar tiene internet, 0 si no tiene
Teléfono	1 si el hogar tiene teléfono, 0 si no tiene
Computadora	1 si el hogar tiene computadora, 0 si no tiene

Fuente: ECH2011, elaboración propia.

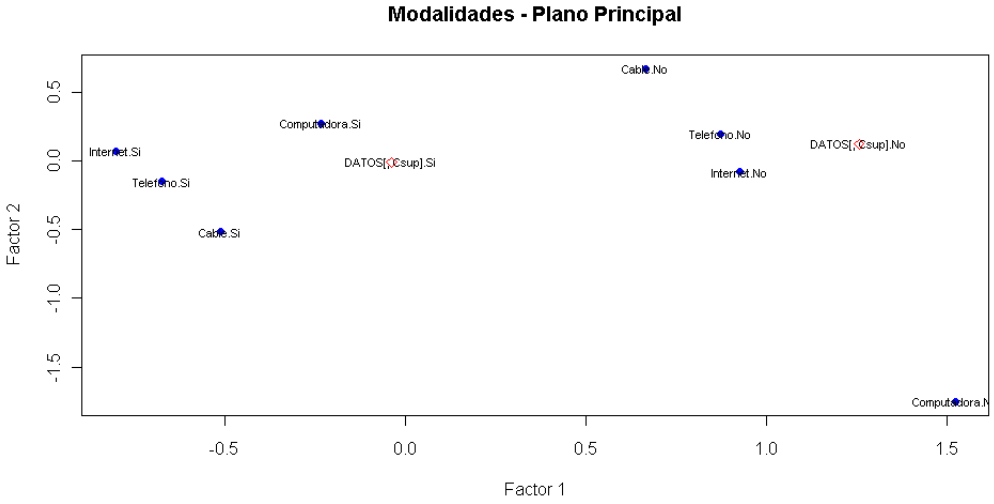
El gráfico del plano principal se encuentra en el anexo, en este se observa que la modalidad "TV No" se encuentra muy lejos de la nube de las modalidades debido a su poca frecuencia, por lo que esta modalidad es considerada como "rara". Al ubicar esta variable como suplementaria no encontramos variables que distorsionen el análisis. Claramente se observa que las amplitudes de ambos ejes disminuyen, disminuyendo así la dispersión de la nube. En el primer factor, se ubican nuevamente las modalidades "No tiene" a la derecha del "0" y las "Si tiene" a la izquierda. Por lo que cuanto más chico este Índice en una mejor situación se encuentra el hogar en referencia a las TIC. Se aprecia también la asociación existente entre "Teléfono" e "Internet". Para poder acceder a internet desde el hogar es necesaria una previa instalación telefónica o la utilización de ADSL móvil. La mayoría de los hogares en Uruguay se caracterizan por tener instalación telefónica para la utilización de internet explicando así la asociación éntre las variables "Teléfono" e "Internet".

Cuadro 6.6: Tenencia de TIC.

Variables	No	Si	Total
TV	2 %	98 %	100 %
Cable	44 %	56 %	100 %
Internet	46 %	54 %	100 %
Teléfono	44 %	56 %	100 %
Computadora	13 %	87 %	100 %

Fuente: ECH2011, elaboración propia.

Figura 6.2: Índice de TIC con variables suplementarias.



Fuente: ECH2011, elaboración propia.

Observando la descomposición de la inercia se encuentra que con un factor la proporción de la inercia es de casi 51 %. Apelando a la transformación de Benzecri, que repondera los valores propios, se encuentra que con sólo retener un valor se obtiene un porcentaje acumulado del 100 %.

En cuanto a la calidad de representación de las modalidades, se observa que las modalidades que refieren a las variables "Internet" y "Teléfono" se encuentran bien representados en el primer eje factorial presentando un coseno cuadrado mayor que 0,5. Las restantes poseen una baja calidad de representación con valores de coseno cuadrado cercano a 0,35. Los resultados se pueden apreciar en el anexo de resultados.

6.2.2. Selección de variables.

Siguiendo las líneas de los autores que han analizado la temática, se consideran las siguientes variables explicativas para determinar las decisiones de asistencia:

Cuadro 6.7: Descripción de variables.

Variable	Descripción
Variables Socio demográficas	
Afro	1 si es de raza afro, 0 es otra raza
Sexo	1 si es hombre, 0 si es mujer
E27	Años de edad
JovenActivo	1 si el joven es activo, 0 si no lo es
Hijos	1 si tiene al menos un hijo, 0 si no tiene hijos
Jefe	1 si el joven es el jefe de hogar, 0 si no lo es
LN YSVL Sin joven prom	Log del ingreso promedio del hogar sin valor locativo sin el ingreso del joven
Madre Ausente	1 si tiene madre ausente, 0 otros casos
Actividad del Jefe	1 si el jefe es activo, 0 en otros casos
Aniosed	Años de educación
Climaeducativo	Promedio de años de educ. de los adultos del hogar, en su defecto del jefe del hogar
Hacinamiento	1 si el hogar presenta hacinamiento, 0 en otro caso
Mdeo	1 si reside en Montevideo, 0 en otro caso
Indicadores	
Icv	Índice de Calidad de Vivienda (Reporte Social)
Viv	Índice de Vivienda y Salubridad (ACM)
Confort	Índice de Confort
Tic	Índice de Tecnología de la Información y Comunicación
Variables de Tecnología e Información	
Computadora	1 si tiene computadora, 0 si no tiene computadora
Teléfono	1 si tiene teléfono, 0 si no tiene teléfono
Internet	1 si tiene internet, 0 si no tiene internet
Cable	1 si tiene cable, 0 si no tiene cable
Variables de Confort	
Calefón	1 si tiene calefón, 0 si no tiene calefón
DVD	1 si tiene DVD, 0 si no tiene DVD
Microondas	1 si tiene microondas, 0 si no tiene microondas
Aire	1 si tiene aire acondicionado, 0 si no tiene aire acondicionado
Auto o moto	1 si tiene auto o moto, 0 sino tiene auto o moto
Auto	1 si tiene auto, 0 si no tiene auto
Moto	1 si tiene moto, 0 si no tiene moto
Secadora	1 si tiene secadora, 0 si no tiene secadora
Lavavajillas	1 si tiene lavavajillas, 0 otros casos
Refrigerador	1 si tiene refrigerador, 0 si no tiene refrigerador
TV	1 si tiene TV, 0 si no tiene TV

Fuente: ECH2011, elaboración propia.

6.2.2.1. Determinación de la muestra de entrenamiento y selección de variables mediante el método "backward" y "test de razón de verosimilitud".

Con el fin de determinar el mejor modelo que explique la asistencia de los jóvenes de 14 a 17 años al sistema educativo, se adoptó como estrategia particionar la muestra en una de entrenamiento y otra de prueba. Con la muestra de entrenamiento se estima el mejor modelo para luego evaluar

su capacidad predictiva con la muestra de prueba. En la base sin ponderar (ECH) existen 8.868 jóvenes de 14 a 17 años de edad, se toma el 80 % de dicha muestra para obtener la muestra de entrenamiento. Para ello se realiza entonces un muestreo aleatorio simple estratificado, respetando los estratos de asistencia antes presentados en el capítulo de estadística univariada (18,73 % No asiste y 81,27 % Asiste), obteniendo así una muestra de 7.094 jóvenes de 14 a 17 años. Los 1.774 jóvenes restantes representan la muestra de prueba con la cual se aprecia la capacidad predictiva del modelo. Del total de jóvenes pertenecientes a la muestra de entrenamiento, 5.765 (81 %) pertenecen al estrato 1 (Asisten), mientras que 1.329 (19 %) pertenecen al estrato 2 (No asisten).

Una vez obtenidos el Índice de Confort, ICV y TIC, y seleccionada la muestra de entrenamiento, se procede a estimar el modelo que mejor explique la asistencia por parte de los jóvenes a un centro educativo para luego poder apreciar su capacidad predictiva y así seleccionar el mejor modelo tanto en significación de las variables como en predicción.

Se consideraron distintos grupos de variables como modelos originales, con el fin de observar, luego de la selección de las variables, los cambios tanto en el nivel de significación de las variables como del modelo.

Escenarios:

1) **Modelo Original 1**: sin índice de confort ni índice de TIC. Para este primer modelo se utilizaron las variables socio-demográficas, las variables de Confort, las variables de Tecnología e Información y el indicador de vivienda.

2) **Modelo Original 2**: con índice de confort e índice de TIC. Para este segundo modelo se utilizaron la totalidad de las variables socio-demográficas, el Índice de Confort y el Índice de TIC.

El planteo de los dos escenarios se realiza con el fin de descubrir si el hecho de incorporar los indicadores antes calculados (confort, TIC, ICV) resulta más eficiente que utilizar las variables por separado para explicar la asistencia.

Se consideraron distintas formas de calcular las variables para observar los cambios en los niveles de significación de las mismas. Dentro de estas variables analizadas se encuentran: “auto o moto”, “auto” y “moto”, así como también distintas transformaciones en las variables de ingreso, descendencia y región. En referencia a las variables “auto o moto”, “auto” y “moto” cabe acla-

rar que tanto la combinación "auto o moto"-"auto" y "auto o moto"-"moto" no son conjuntos disjuntos por lo que estas combinaciones de variables no fueron incluidas en los modelos, únicamente se consideraron las variables únicas.

Selección de Variables.

Luego de la discusión planteada al comienzo del capítulo para determinar cuál función es la adecuada cuando se tiene diseño de muestra ("svyglm" o "glm"), se decide obtener los modelos finales utilizando "svyglm". De todas formas, una vez obtenidos los modelos finales y seleccionadas las variables se procedió a realizar un "glm" con las variables seleccionadas por el "svyglm" con el fin de observar el cambio en la significación de las variables cuando no se utiliza el diseño de muestra. Los resultados obtenidos se analizan a partir de la estimación de los modelos probit y logit para los distintos escenarios. Para realizar la estimación de estos modelos se utilizan las funciones antes mencionadas "svyglm" y "glm" que se encuentran dentro de los paquetes "survey" y "stats" respectivamente. En la función "svyglm" se incorpora la función "link" (probit o logit), los datos a usar (jóvenes de 14 a 17 años pertenecientes a la muestra de entrenamiento) y el diseño de muestra (en este caso el diseño de la ECH2011).

Una vez definidas las funciones a utilizar y determinados los grupos de variables se procede con la selección de las variables. Tanto el Test de Razón de Verosimilitud como el procedimiento de backward son instrumentos a utilizar para llevar a cabo el proceso de construcción del modelo. Cabe aclarar que en "R" la función "stepwise" utilizada usualmente para la selección de modelos en "glm" no es viable en este caso, ya que al utilizar la función "svyglm" se utiliza el estimador de máxima verosimilitud ponderado y no es posible calcular el AIC. El estimador de máxima verosimilitud ponderado es el que se obtiene de maximizar la función de verosimilitud ponderada, el adjetivo ponderada viene del hecho de adicionar a la función de verosimilitud las probabilidades de inclusión de los individuos en la muestra.Por este motivo, se decide utilizar como estrategias de selección de variables, el método "backward manual"[1] y por otro lado el Test de Razón de Verosimilitud detallados en el anexo.

La metodología "backward manual" parte de un modelo completo y elimina en una primera instancia la variable con mayor p-valor. Repite el mismo procedimiento hasta llegar a un modelo con todas las variables significativas. Partiendo del Modelo Original 1, e incorporando la varia-

[1]Método Backward: Es un procedimiento de selección de variables en el cual se parte del modelo incluyendo todas las variables explicativas y luego se van eliminando de a una según su capacidad explicativa. La primer variable que se elimina es aquella que presenta un menor coeficiente de correlación parcial con la variable dependiente y así sucesivamente.

ble “auto o moto” y utilizando la metodología “backward”, se selecciona el siguiente grupo de variables de salida en el orden determinado según su p-valor:

Cuadro 6.8: Posición de salida de variables utilizando la metodología backward partiendo del modelo original 1 e incorporando la variable “auto o moto”.

Posición de Salida	Variable
1	Internet
2	Madre Ausente
3	Calefón
4	Cable
5	Actividad del Jefe
6	Refrigerador
7	DVD
8	LN YSVL sin joven prom
9	Secadora
10	Mdeo
11	TV
12	Aire
13	Lavavajillas

Fuente: ECH2011, elaboración propia

Las salidas parciales se pueden encontrar en el Anexo de Resultados. Se presenta a continuación la salida final:

Figura 6.3: Salida de R mediante el método backward partiendo del modelo original 1

```
svyglm(formula = Asiste ~ afro + sexo + E27 + JovenActivo + hijos +
    jefe + aniosed + climaeducativo + icv2 + Computadora + Telefono +
    Microondas + Auto_o_moto + Hacinamiento, family = quasibinomial(link = "logit"),
    data = Per1417.con.muestra, design = diseño_personas_14_17,
    subset = (Stratum != 0))

Survey design:
subset(diseño_personas, (Personas.con.muestra$E27 >= 14 & Personas.con.muestra$E27 <=
    17))

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         9.66602    0.71326  13.552  < 2e-16 ***
afroafro            0.37079    0.12917   2.871  0.00411 **
sexohombre         -0.21176    0.09401  -2.253  0.02432 *
E27                -0.98787    0.04982 -19.828  < 2e-16 ***
JovenActivo1       -1.41775    0.11316 -12.529  < 2e-16 ***
hijosAl menos un hijo -2.42438    0.36283  -6.682 2.57e-11 ***
jefejefe           -1.64659    0.71163  -2.314  0.02071 *
aniosed             0.82275    0.04312  19.080  < 2e-16 ***
climaeducativo      0.13076    0.02146   6.092 1.18e-09 ***
icv2deficit        -1.68286    0.53583  -3.141  0.00169 **
ComputadoraSi       0.44532    0.12501   3.562  0.00037 ***
TelefonoSi          0.24085    0.10500   2.294  0.02184 *
MicroondasSi       0.25439    0.10466   2.431  0.01510 *
Auto_o_motoSi      -0.28746    0.09693  -2.966  0.00303 **
Hacinamiento1      -0.21054    0.10508  -2.004  0.04516 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

En este resultado se observa como la variable “afro” tiene un coeficiente positivo. En otras palabras esto significa que el hecho de que el joven sea de descendencia afro impacta positivamente en la asistencia. Esto, por un lado, contradice los datos brindados en el anexo que informan a priori que los jóvenes que tienen descendencia afro asisten menos que los que no tienen descendencia afro (75 % -82 %), además de distintos análisis realizados, cruzando variables (ver anexo) la variable “afro” puede llegar a presentar asociación con la variable “años de educación“ y con “clima educativo”, esta posible asociación puede estar interfiriendo en el análisis presentando en el resultado del modelo un signo positivo erroneo por parte de la variable “afro”. Por este motivo sería conveniente excluir a la variable “afro” del análisis.

Cuadro 6.9: Asistencia según ascendencia.

	No afro	Afro	Total
No asiste	18 %	25 %	19 %
Asiste	82 %	75 %	81 %
Total	100 %	100 %	100 %

Fuente: ECH2011, elaboración propia

Por otro lado, se observa que la variable “auto o moto” presenta signo negativo, esto significa, que el hecho de que el joven tenga auto o moto impacta negativamente en la asistencia. Este signo sugiere realizar un pequeño análisis descriptivo de esta variable y observar la construcción de la misma. Se decide por lo tanto estudiar como varía la asistencia de los jóvenes a un centro educativo según “moto”, “auto” y “auto o moto” en referencia a los que no tienen. Comparando estas tres variables según la asistencia se observa que los porcentajes de asistencia aumentan cuando el joven tiene “autoz cuando tiene “auto o moto” pero no cuando tiene “moto”. La asistencia de los jóvenes que tienen moto es menor que los que no tienen. En particular, sería incorrecto incorporar la variable “moto” al modelo original ya que si esta variable presenta signo negativo en el modelo final no significa que el hecho de que el joven tenga moto tenga un impacto negativo en la asistencia sino que los jóvenes que tienen moto presentan características diferentes a los que no tienen y esto es en en definitiva lo que determina la asistencia, se debería estudiar la posible asociación de dicha variable con otras presentes en el modelo. Este análisis conduce a optar por utilizar únicamente la variable “auto”.

Cuadro 6.10: Asistencia por tenencia de bienes.

	Auto o moto	Moto	Auto
No Tiene	78 %	82 %	76 %
Tiene	83 %	79 %	90 %

Fuente: ECH2011, elaboración propia

Luego de hacer estas observaciones se plantea un nuevo modelo original ahora incorporando la variable "auto", y sin las variables "afro" y "auto o moto". Al aplicar nuevamente la metodología "backward" obtenemos un modelo final similar al modelo obtenido anteriormente pero ahora sin la variable "Teléfono" y con la variable "Montevideo":

Modelo Final 1:

Figura 6.4: Salida de R mediante el método backward partiendo del modelo original 1 sin las variables afro ni auto o moto.

```
MF1=svyglm(formula = Asiste ~ sexo + E27 + JovenActivo + hijos +
    jefe + aniosed + climaeducativo + Mdeo + icv2 + Computadora +
    Microondas + Hacinamiento, family = quasibinomial(link = "logit"),
    data = Personas.con.muestra, design = diseño_personas_14_17,
    subset = (Stratum != 0))

Survey design:
subset(diseño_personas, (Personas.con.muestra$E27 >= 14 & Personas.con.muestra$E27 <=
    17))

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           9.59562    0.70541  13.603  < 2e-16 ***
sexohombre           -0.20392    0.09421  -2.165 0.030462 *
E27                  -0.98670    0.04960 -19.893  < 2e-16 ***
JovenActivo1         -1.43249    0.11353 -12.618  < 2e-16 ***
hijosAl menos un hijo -2.38013    0.35761  -6.656 3.06e-11 ***
jefejefe             -1.68127    0.72187  -2.329 0.019888 *
aniosed               0.82452    0.04303  19.163  < 2e-16 ***
climaeducativo        0.12589    0.02107   5.974 2.44e-09 ***
MdeoMontevideo        0.22188    0.10203   2.175 0.029700 *
icv2deficit          -1.61784    0.55196  -2.931 0.003390 **
ComputadoraSi         0.44441    0.12509   3.553 0.000384 ***
MicroondasSi          0.23178    0.10292   2.252 0.024363 *
Hacinamiento1        -0.21601    0.10430  -2.071 0.038394 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasibinomial family taken to be 0.9840903)

Number of Fisher Scoring iterations: 6
```

Partiendo del modelo original 2 (con índices) y utilizando la metodología "backward", se selecciona el siguiente grupo de variables de salida en el orden determinado:

Cuadro 6.11: Posición de salida de variables mediante el método backward partiendo del modelo original 2.

Posición de Salida	Variable
1	Madre ausente
2	Actividad del Jefe
3	Confort
4	Ingreso
5	Hacinamiento

Obteniendo así el **Modelo Final 2:**

Figura 6.5: Salida de R mediante el método backward partiendo del modelo original 2

```
MF2=svyglm(formula = Asiste ~ sexo + E27 + JovenActivo + hijos +
    jefe + aniosed + climaeducativo + TIC1_1417 + Mdeo + icv2,
    family = quasibinomial(link = "logit"), data = Personas.con.muestra,
    design = diseño_personas_14_17, subset = (Stratum != 0))

Survey design:
subset(diseño_personas, (Personas.con.muestra$E27 >= 14 & Personas.con.muestra$E27 <=
    17))

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         10.17677    0.68251  14.911  < 2e-16 ***
sexohombre          -0.19541    0.09367  -2.086 0.037010 *
E27                 -0.99666    0.04913 -20.285  < 2e-16 ***
JovenActivo1        -1.43062    0.11316 -12.642  < 2e-16 ***
hijosAl menos un hijo -2.40325    0.35428  -6.783 1.28e-11 ***
jefejefe            -1.74597    0.74503  -2.344 0.019135 *
aniosed              0.82348    0.04289  19.198  < 2e-16 ***
climaeducativo       0.12901    0.02134   6.044 1.59e-09 ***
TIC                 -0.29279    0.07729  -3.788 0.000153 ***
MdeoMontevideo       0.23899    0.10130   2.359 0.018338 *
icv2deficit         -1.61837    0.54696  -2.959 0.003099 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasibinomial family taken to be 1.00443)

Number of Fisher Scoring iterations: 6
```

Hasta el momento se tienen dos posibles modelos finales; el Modelo Final 1(MF1) que se obtuvo partiendo del modelo original 1 (sin indices) y el Modelo Final 2(MF2) que se obtuvo del modelo original 2 (con índices) , estos difieren en la incorporación de las variables "Computadora", "Microondas", "Hacinamiento" y "TIC". En el cuadro que se presenta a continuación se resumen las variables que son significativas en ambos modelos.

Cuadro 6.12: Modelos finales MF1 y MF2.

Variable	MF1	MF2
Sexo	√	√
Edad	√	√
Actividad del Joven	√	√
Hijos	√	√
Jefe	√	√
Años de educación	√	√
Clima educativo del hogar	√	√
Montevideo	√	√
ICV	√	√
TIC		√
Computadora	√	
Microondas	√	
Hacinamiento	√	

Fuente: ECH2011, construcción propia.

Después de obtener los modelos finales con la estrategia de selección "backward", se procede a utilizar una nueva metodología de selección de variables: Test de Razón de Verosimilitudes sobre los modelos originales con el fin de analizar si al cambiar la estrategia de selección cambia el modelo final elegido.

Dicho test de razón de verosimilitud tiene como objetivo el comparar dos modelos de regresión logística, el denominado modelo completo, frente al que se conoce como modelo reducido. Este segundo modelo puede verse como un sub-modelo del modelo completo. La hipótesis nula sometida a prueba establece que los parámetros correspondientes a las variables que forman parte del modelo completo, pero no del modelo reducido, valen cero.

Se presenta a continuación un cuadro que resume los modelos obtenidos con la estrategia de selección razón de verosimilitud. Partiendo del modelo original 1 se obtuvo el modelo final 3 (MF3) y partiendo del modelo original 2 se obtuvo el modelo final 4 (MF4)

Cuadro 6.13: Modelos finales MF3 y MF4.

Variable	MF3	MF4
Sexo	√	√
Edad	√	√
Actividad del Joven	√	√
Hijos	√	√
Jefe	√	√
Años de educación	√	√
Clima educativo del hogar	√	√
ICV	√	√
TIC		√
Computadora	√	
Teléfono	√	
Hacinamiento	√	

Fuente: ECH2011, construcción propia.

Partiendo de los modelos originales 1 y 2 y utilizando las distintas estrategias de selección de variables se obtienen entonces cuatro modelos finales:

Cuadro 6.14: Modelos finales.

Variable	MF1	MF2	MF3	MF4
Sexo	√	√	√	√
Edad	√	√	√	√
Actividad del Joven	√	√	√	√
Hijos	√	√	√	√
Jefe	√	√	√	√
Años de educación	√	√	√	√
Clima educativo del hogar	√	√	√	√
Montevideo	√	√		
ICV	√	√	√	√
TIC		√		√
Computadora	√		√	
Microondas	√			
Hacinamiento	√		√	
Teléfono			√	

Fuente: ECH2011, construcción propia.

Logit vs Probit. Luego de obtener estos cuatro modelos finales, se procede a analizar qué tipo de modelo de elección discreta, "logit" o "probit", es más eficiente a la hora elegir el mejor modelo que explique la asistencia. Al analizar los cuatro modelos con el cambio en la función "link" no se encuentran diferencias en cuanto a la significación de las variables. Por este motivo el análisis se concentra en elegir la función "link" que genere el menor error predictivo.

"Svyglm" vs "glm". Por otro lado, como se menciona anteriormente, una vez obtenidos los cuatro modelos finales con la función "svyglm" se utiliza la función "glm" con las variables seleccionadas por el "svyglm" con el fin de estudiar los cambios en la significación de las mismas cuando no se utiliza un diseño de muestra.

Con este procedimiento se observa que para los cuatro modelos finales las variables "Jefe" y "Sexo" tienen un cambio en la significación. La variable "Jefe" no es significativa con el "glm" mientras que con el "svyglm" es significativa al 5 %. Lo mismo ocurre en el Modelo Final 1 con

la variable" Hacinamiento". No ocurre lo mismo con la variable "Sexo", la misma obtiene una "pérdida" de significación en los cuatro modelos, mientras que con el "glm" es significativa al 10 % con el "svyglm" lo es al 5 %. Por último, en el Modelo Final 3 dos nuevas variables sufren cambios en la significación estas son: "Microondas" y "Teléfono." Con el modelo "glm" la variable "Microondas" es significativa al 1 %, mientras que con el "svyglm" pasa a ser significativa al 5 %. En referencia a la variable "Teléfono" la misma pasa a ser significativa al 1 % en el "glm" a serlo al 5 % en el "svyglm".

La mayoría de las variables son elegidas por todos los modelos. Considerando una significación del 5 %, en los modelos finales 1 y 3 existen dos variables que no serían elegidas según "glm", mientras que en los modelo final 2 y 4 una única variable no sería elegida según "glm". Los respectivos p-valores se pueden apreciar en el anexo de resultados.

No se encuentran grandes diferencias en cuanto a la selección de variables cuando se utiliza la función "svyglm" o "glm". Dada la discusión planteada al comienzo del capítulo sobre la utilización o no de diseño de muestra a la hora de elegir el mejor modelo cuando se tiene un diseño muestral y observando los resultados obtenidos mediante "svyglm" y "glm" se opta por elegir los modelos "svyglm".

6.2.3. Predicción.

Para finalizar la elección del modelo se procede a evaluar el poder predictivo de estos cuatro modelos finales. Como se mencionó anteriormente al inicio del capítulo, se adoptó como estrategia particionar la muestra en una de entrenamiento (7.094 jóvenes) y otra de prueba (1.774 jóvenes), con la muestra de entrenamiento se contruyeron los distintos modelos y la muestra de prueba se utilizó para evaluar el poder predictivo de estos modelos.

Para evaluar éste poder de predicción se observa por un lado las tablas de clasificación donde se resumen los errores de clasificación y por otro las curvas ROC. El valor predicho de un individuo será 1 si su probabilidad predicha supera el valor del "punto de corte" ($Y_i = 1$), y será cero si es menor o igual al "punto de corte". Se evalúan los distintos puntos de corte y se observan las tablas de clasificación y las curvas ROC.

6.2.3.1. Tablas de Clasificación.

Observando las distintas tablas de clasificación para los diferentes modelos se encuentra que los modelos tienen una capacidad predictiva muy similar. Los cuatro modelos tienen una buena predicción llegando a tasas de acierto mayores al 80 %. Para encontrar el modelo con mejor capacidad predictiva se debe encontrar el modelo que minimice las tasas de error. En este análisis se considera más importante el error de clasificar a un joven como que asiste cuando en realidad no lo hace (n_2) que clasificar a un joven como que no asiste cuando en realidad asiste (n_3). Por este motivo se busca el punto de corte que minimice n_3 y n_2 , priorizando n_2. A partir de un punto de corte de 0,85 se observa cómo n_2 presenta valores menores al 20 % para los cuatro modelos.

Cuadro 6.15: Tabla de clasificación del modelo.

predicho / observado	$\hat{Y} = 0$	$\hat{Y} = 1$	Total
Y = 0	N_1	N_2	$N_1 + N_2$
Y = 1	N_3	N_4	$N_3 + N_4$
Total	$N_1 + N_3$	$N_2 + N_4$	N

Se definen las tasas de error del modelo tanto para el grupo de los que asisten como de los que no asisten:

$$n_3 = \frac{N_3}{N_3+N_4},\ n_2 = \frac{N_2}{N_1+N_2}$$

Las tasas de acierto del modelo son respectivamente:

$$n_4 = \frac{N_4}{N_3+N_4},\ n_1 = \frac{N_1}{N_1+N_2}$$

Cuadro 6.16: Errores y aciertos de clasificación según punto de corte.

Punto	MF1				MF2				MF3				MF4			
de Corte	n3	n2	n4	n1	n3	n2	n4	n1	n3	n2	n4	n1	n3	n2	n4	n1
0,79	0,1429	0,2651	0,8571	0,7349	0,1422	0,2741	0,8578	0,7259	0,1422	0,2741	0,8578	0,7259	0,1414	0,2801	0,8585	0,7198
0,8	0,1498	0,2651	0,8502	0,7349	0,1498	0,2620	0,8502	0,7380	0,1526	0,2620	0,8474	0,7380	0,1497	0,2680	0,8502	0,7319
0,81	0,1574	0,2530	0,8426	0,7470	0,1581	0,2560	0,8419	0,7440	0,1574	0,2470	0,8426	0,7530	0,1553	0,2530	0,8446	0,7469
0,82	0,1637	0,2349	0,8363	0,7651	0,1664	0,2440	0,8336	0,7560	0,1678	0,2410	0,8322	0,7590	0,1657	0,2379	0,8342	0,7620
0,83	0,1768	0,2349	0,8232	0,7651	0,1803	0,2259	0,8197	0,7741	0,1803	0,2259	0,8197	0,7741	0,1775	0,2228	0,8224	0,7771
0,84	0,1886	0,2259	0,8114	0,7741	0,1872	0,2169	0,8128	0,7831	0,1872	0,2169	0,8128	0,7831	0,1893	0,2138	0,8106	0,7861
0,845	0,1928	0,2048	0,8072	0,7952	0,1956	0,2078	0,8044	0,7922	0,1928	0,1988	0,8072	0,8012	0.1914	0,2138	0,8085	0,7861
0,85	0,1990	0,1958	0,8010	0,8042	0,2039	0,2018	0,7961	0,7982	0,1949	0,1928	0,8051	0,8072	0,1976	0,2018	0,8023	0,7981
0,86	0,2115	0,1867	0,7885	0,8133	0,2150	0,1898	0,7850	0,8102	0,209	0,1807	0,7906	0,8193	0,2122	0,1927	0,7877	0,8072
0,87	0,2288	0,1717	0,7712	0,8283	0,2254	0,1657	0,7746	0,8343	0,2240	0,1687	0,7760	0,8313	0,2226	0,1867	0,7773	0,8132

Fuente: ECH2011, elaboración propia.

Si se considera como criterio igualar n_4 con la tasa de asistencia inicial 81,27 % se eligiría un punto de corte entre 0,845 y 0,85. Con los resultados de esta tabla se puede apreciar la similitud y buena capacidad predictiva de los cuatro modelos pero es difícil elegir el más eficiente. Por lo que se decide utilizar como complemento la metodología de las curvas ROC.

La curva ROC es más informativa que una tabla de clasificación, dado que resume el poder predictivo del modelo para todos los posibles valores del punto de corte. El área bajo la curva ROC es una medida del poder predictivo llamada índice de concordancia, proporciona una medida de la habilidad del modelo para discriminar entre las observaciones que presentan el suceso de interés versus aquellos que no.

En la siguiente tabla se aprecia el valor del área debajo de la curva (o poder predictivo del modelo), para cada punto de corte:

Cuadro 6.17: Área bajo la curva ROC por modelo según punto de corte.

Punto de Corte	Area bajo la curva ROC			
	MF1	MF2	MF3	MF4
0,6	0,7445	0,7400	0,7457	0.7392
0,7	0,7754	0,7726	0,7769	0.7715
0,75	0,7849	0,7855	0,7807	0.7804
0,76	0,7858	0,7876	0,7803	0.7869
0,77	0,7894	0,7909	0,7832	0.7881
0,78	0,7901	0,7919	0,7905	0.786
0,79	0,7904	0,7911	0,7953	0.7907
0,8	0,7926	0,7937	0,8027	0.7921
0,81	0,7948	0,7929	0,7979	0.7959
0,82	0,8007	0,797	0,7927	0.7993
0,83	0,7989	0,7952	0,7905	0.7998
0,84	0,8001	0,7979	0,8048	0.7949
0,85	0,8026	0,7972	0,8073	0.8004
0,86	0,8009	0,7976	0,8106	0.7983
0,87	0,7997	0,8045	0,8076	0.8015
0,88	0,8001	0,7982	0,8068	0.8037
0,89	0,8019	0,8052	0,7983	0.8002
0,9	0,7927	0,7997	0,7949	0.7992

Fuente: ECH2011, elaboración propia.

Los modelos finales 1 y 3 fueron obtenidos desde el modelo orginal 1 (modelo sin índices) utilizando distinas estrategias de selección de variables, estos modelos difieren en la incorporación de las variables “Montevideo”, “Microondas” y “Teléfono”.Se observa que no exísten grandes diferencias entre ellos presentando valores similares bajo la curva ROC, pero a la hora de elegir el más eficiente el Modelo final 3 sería el más indicado ya que supera al Modelo final 1 en la predicción alcanzando un valor máximo bajo la curva (0,8106).

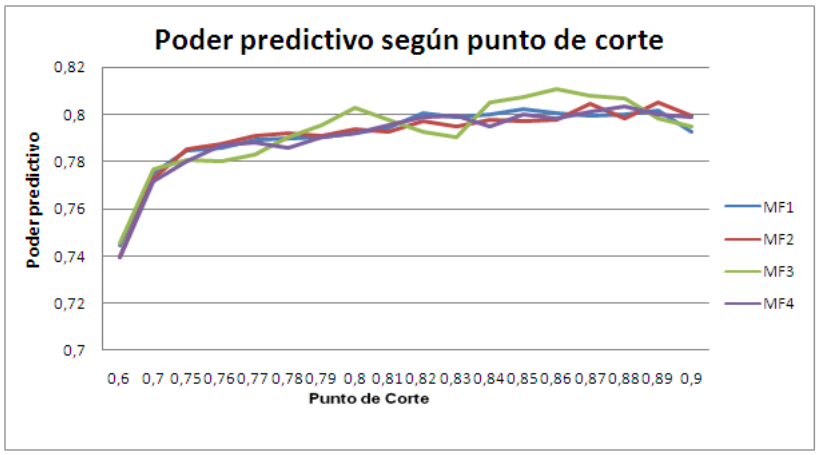
En referencia a los modelos finales 2 y 4, estos fueron obtenidos desde el modelo original 2 (modelo con índices) siendo su única diferencia la incorporación de la variable “Montevideo”, en estos modelos tampoco se observan grandes diferencias en los valores de predicción, pero a la hora de elegir el mas eficiente el modelo final 2 es el que tiene mejor predicción alcanzando el

valor máximo del área bajo la curva (0,8045).

Se observa entonces que los modelos finales 2 y 3 son los que tienen mejor desempeño, siendo el MF3 el que presenta el valor máximo bajo la curva el cual corresponde a un punto de corte igual a 0,86. Como este valor del área máximo es mayor a 0,8 y menor a 0,9 decimos que el modelo posee muy buena discriminación y capacidad predictiva.

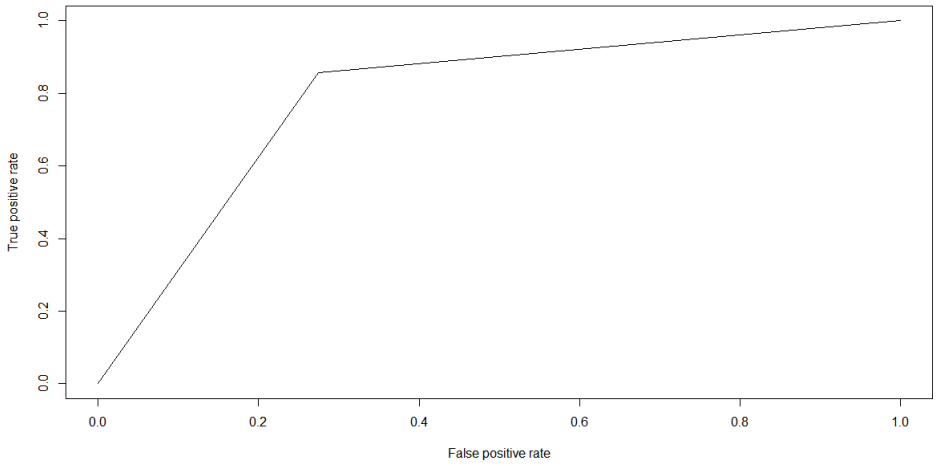
En el siguiente gráfico se aprecia la capacidad predictiva de cada modelo según los distintos puntos de corte. Se observa como el Modelo Final 3 tiene mejor predicción que los Modelos Finales 1, 2 y 4 presentando un pico máximo en el punto de corte 0,86 con un valor de área de 0,8106.

Figura 6.6: Poder predictivo de los modelos finales según punto de corte.



Fuente: ECH2011, elaboración propia.

Figura 6.7: Curva ROC del Modelo Final 3.



Fuente: ECH2011, elaboración propia.

Para evaluar si el buen desempeño del modelo MF3 es independiente de la muestra elegida, se presenta un cuadro que compara la tabla de clasificación obetenida mediante la muestra de prueba con la de entrenamiento. Se observa que no existen grandes diferencias entre estos valores, esto indica que el ajuste del modelo es adecuado y es independendiente de la muestra elegida.

Cuadro 6.18: Errores y aciertos de clasificación de MF3 con muestra de prueba y de entrenamiento según punto de corte.

Punto de	Muestra de prueba				Muestra de entrenamiento			
Corte	n3	n2	n4	n1	n3	n2	n4	n1
0,79	0,1422	0,2741	0,8578	0,7259	0,1537	0,1926	0,8463	0,8074
0,80	0,1526	0,2620	0,8474	0,7380	0,1622	0,1874	0,8378	0,8126
0,81	0,1574	0,2470	0,8426	0,7530	0,1717	0,1761	0,8283	0,8239
0,82	0,1678	0,2410	0,8322	0,7590	0,1768	0,1655	0,8232	0,8345
0,83	0,1803	0,2259	0,8197	0,7741	0,1899	0,1573	0,8101	0,8427
0,84	0,1872	0,2169	0,8128	0,7831	0,1995	0,1422	0,8005	0,8578
0,845	0,1928	0,1988	0,8072	0,8012	0,2045	0,1369	0,7955	0,8631
0,85	0,1949	0,1928	0,8051	0,8072	0,2116	0,1339	0,7884	0,8661
0,86	0,2090	0,1807	0,7906	0,8193	0,2281	0,1204	0,7719	0,8796
0,87	0,2240	0,1687	0,7760	0,8313	0,2373	0,1159	0,7627	0,8841

Fuente: ECH2011, elaboración propia.

Es interesante observar que mas allá de que los modelos finales 2 y 3 son los que tienen mejor desempeño, cualquiera de los cuatro modelos pueden ser utilizados a la hora de estudiar la asistencia, ya que los mismos no presentan grandes cambios en su predicción. El modelo final 3 resulta ser el más eficiente pero la elección entre los modelos 2 y 3 depende de la información que se posea y del tipo de análisis que se quiera realizar. Para cierta población puede ser de interes utilizar el modelo final 2 el cual incorpora el indicador de tecnologías de información (TIC) para así poder evaluar como impacta este indicador en la asistencia para cierta población.

Se presenta a continuación una breve descripción de la estimación de los coeficientes para el modelo final 3, la cual se detallará en el capítulo de conclusiones.

Características como el ser hombre, ser activo, tener hijos a cargo y la condición de ser jefe del hogar tienen un impacto negativo en la asistencia, lo mismo ocurre con la edad, la probabilidad de no asistencia se incrementa a medida que la misma aumenta, lo cual se ve reflejado en el signo negativo de los coeficientes estimados asociados a dichas variables. En referencia a las características de los hogares donde los jóvenes residen se puede concluir que cuanto mejor son las condiciones estructurales de la vivienda así como la ausencia de hacinamiento mayor será la probabilidad de asistencia del joven. Por otro lado las variables años de educación, clima educativo, tener computadora, tener teléfono y/o tener cable, tienen un impacto positivo en la asistencia. Cuanto mayor son los años de educación del joven , mayor es la probabilidad de asistencia a un centro educativo, ocurriendo lo mismo con el clima educativo del hogar.

Si se compara el impacto en cuanto al sexo, se tiene que $e^{\hat{\beta}_{hombre}} = 0,82$. De este modo se aprecia que el ser hombre respecto a ser mujer disminuye el cociente de probabilidad de asistencia/probabilidad de no asistencia en un 20 % dejando las demás variables constantes. En lo que respecta a la edad, el impacto de incrementar en 1 año la edad del joven, disminuye el cociente de probabilidad de asistencia/probabilidad de no asistencia en casi un 60 % $[e^{\hat{\beta}_{edad}} = 0,375]$ dejando las demás variables fijas. En lo que refiere a la condición de actividad del joven, se aprecia que el ser activo impacta disminuyendo en casi un 80 % dicho cociente de probabilidades dejando las demás variables constantes. En este caso el cociente de odds estimado es $e^{\hat{\beta}_{activo}} = 0,23$. Si se toma en cuenta como impacta el hecho de tener al menos un hijo y la condición de ser jefe del hogar, se tiene que el primero disminuye el cociente de probabilidad de asistencia/probabilidad de no asistencia en un 90 % y el segundo en casi un 20 % dejando en cada caso fijas el resto de las variables.
En lo que refiere a los años de educación acumulados por el joven, el impacto de incrementar en una unidad los años de educación ($e^{\hat{\beta}_{añosed}} = 2,26$), hace que aumente el cociente de probabilidad

de asistir/probabilidad de no asistir en dos veces o en otras palabras en un 100 %, dejando las demás constantes. Si se analiza el impacto de incrementar en 1 año el clima educativo del hogar, se debe calcular $e^{\hat{\beta}_{climaed}}$ que refleja el impacto en el cociente de odds de incrementar en 1 año el clima educativo del hogar. En este caso dicho valor es de 1,137, es decir que aumentar en un año el clima educativo del hogar tiene un impacto de incrementar en un 13 % en el cociente de probabilidad de asistir/probabilidad de no asistir, dejando las demás constantes.
Si se realiza el mismo ejercicio, ahora para determinar como impactan en el cociente de probabilidad de asistencia/probabilidad de no asistencia, la situación estructural de la vivienda y la tenencia de determinados bienes de confort en el hogar, se aprecia que el hecho de que el joven habite en un hogar donde la situación de la vivienda es deficitaria disminuye en un 20 % tal cociente, dejando las restantes constantes. Por otra parte la tenencia de computadora, teléfono y microondas impactan incrementando dicho cociente de probabilidades en un 47,6 % el primero, un 27,8 % el segundo, y un 26,5 % el tercero, dejando en todos los casos las demás variables constantes.

6.3. Árbol de clasificación para la variable Asistencia.

Como forma de complementar los modelos presentados anteriormente para modelizar la asistencia se empleó la técnica no paramétrica árbol de decisión. Para que los resultados de esta técnica sean comparables con los de los modelos lineales generalizados se empleó la misma estrategia de construcción. En lo que respecta a la muestra se trabajó con una muestra de entrenamiento con la cual se construyó el árbol y otra muestra (de prueba) en la que se evaluó su poder predictivo. También se partió del mismo conjunto de predictores respetando cada uno de los diferentes escenarios propuestos para la formulación de los modelos. En este caso se utilizó la muestra sin ponderar. Para la construcción de los árboles de clasificación se empleó la función "rpart".

Escenario1:

Al igual que se hizo para el Modelo 1 se consideró el siguiente conjunto de predictores: sexo, edad, condición de actividad del jóven, condición de jefe, clima educactivo, años de eduación, hijos, madre ausente, índice de calidad de vivienda, montevideo, logaritmo del ingreso del hogar sin el jóven en promedio, variables de tecnología e información (TV, computadora, internet, teléfono, cable) y variables de confort (calefón, dvd, microondas, aire, auto o moto, secadora, lavavajillas, refrigerador).

Para la construcción y posterior selección del árbol óptimo se siguió la estrategia sugerida por Breiman. La misma consiste en dejar crecer el árbol hasta que se obtiene el arbol maximal y luego se procede a su poda considerando la secuencia de árboles anidados que se generan. Para determinar árbol óptimo se considera en forma conjunta tanto la complejidad como el error global de clasificación. En la siguiente figura se aprecia un resumen del árbol maximal obtenido para la variable asistencia:

Figura 6.8: Árbol maximal para la variable asistencia.

```
Classification tree:
rpart(formula = Asiste ~ sexo + E27 + JovenActivo + Madre_ausente +
    hijos + jefe + aniosed + climaeducativo + icv2 + Computadora +
    Telefono + Microondas + Auto_o_moto + Cable + Internet +
    TV + Lavavajillas + Secadora + Refrigerador + Calefon + DVD +
    Aire + Mdeo + LN_YSVL_sin_joven_prom + Hacinamiento, data =
per_training,
    method = "class", control = rpart.control(cp = 0))

Variables actually used in tree construction:
 [1] aniosed                Auto_o_moto            Cable
Calefon
 [5] climaeducativo         Computadora            DVD
E27
 [9] Hacinamiento           hijos                  JovenActivo
LN_YSVL_sin_joven_prom
[13] Mdeo                   Microondas             Refrigerador
sexo
[17] Telefono               TV

Root node error: 1329/7094 = 0.18734

n= 7094

           CP nsplit rel error  xerror     xstd
1  0.12678706      0   1.00000 1.00000 0.024728
2  0.03235515      2   0.74643 0.74643 0.021980
3  0.01617758      3   0.71407 0.74041 0.021905
4  0.00376223      7   0.64936 0.65388 0.020778
5  0.00338600     10   0.63807 0.66892 0.020982
6  0.00263356     15   0.61550 0.64861 0.020706
7  0.00225734     17   0.61023 0.65237 0.020758
8  0.00150489     31   0.56885 0.65538 0.020799
9  0.00125408     45   0.54628 0.66065 0.020870
10 0.00112867     51   0.53875 0.67494 0.021063
11 0.00100326     61   0.52596 0.68548 0.021203
12 0.00075245     73   0.51392 0.68849 0.021242
13 0.00060196     91   0.50038 0.70805 0.021496
14 0.00037622     96   0.49737 0.72385 0.021698
15 0.00030098    108   0.49285 0.73890 0.021887
16 0.00018811    113   0.49135 0.74266 0.021933
17 0.00000000    117   0.49059 0.74944 0.022017
```

Fuente:ECH2011, elaboración propia.

La salida anterior presenta un breve resumen del ajuste del modelo, muestra desde el árbol minimal (un único nodo) al árbol máximo que presenta en este caso 118 nodos. Se lista el número de particiones y no el número de nodos. El número de nodos es siempre 1 + el número de particiones. También se aprecian las variables que fueron utilizadas para generar la partición

recursiva del modelo. La mayor parte de las variables incluidas en la partición coinciden con las seleccionadas por el Modelo 1, además se aprecia que se incluyen otras como ser Auto o moto, Cable, Calefón, DVD, Refrigerador, Teléfono, TV y logaritmo del ingreso. Del mismo modo la variable Jefe resultó significativa para el Modelo 1 pero no se incluye en el proceso de partición. Por otra parte se aprecia que las columnas de error han sido re-escaladas de forma tal que el primer nodo tiene un error de 1.

Posteriormente se considera la secuencia de árboles anidados podando el maximal para diferentes parámetros de complejidad. En la siguiente tabla se observa la secuencia de árboles anidados, las respectivas tasas de error y la complejidad del árbol resultante.

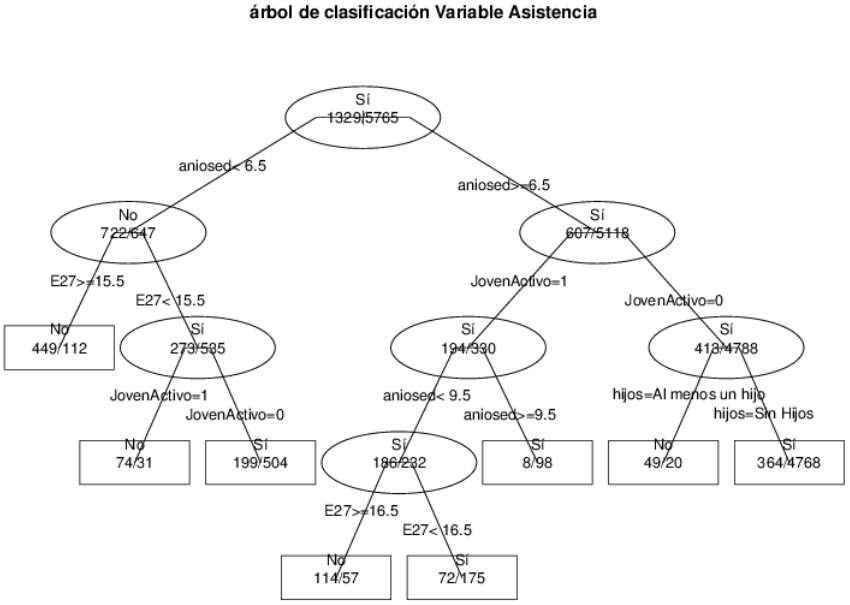
Cuadro 6.19: Secuencia de árboles anidados bajo escenario 1.

Parámetro de complejidad (α)	Tasa de error global	Complejidad del árbol (nodos terminales)
0,001	13,13 %	74
0,002	12,9 %	32
0,003	13 %	16
0,004	12,16 %	8
0,02	14 %	4

Fuente:ECH2011, elaboración propia.

El objetivo aquí es encontrar la proporción óptima entre la tasa de mala clasificación y la complejidad del árbol, siendo la tasa de mala clasificación el cociente entre las observaciones mal clasificadas y el número total de observaciones. El parámetro de complejidad penaliza la complejidad del árbol, entendida como el número de nodos terminales. Tomando en cuenta lo anterior se considera como árbol óptimo el que posee 8 nodos terminales, una tasa de error global de 12,16 % y se corresponde con un $cp = 0,004$. En la siguiente figura se aprecia la estructura que adopta el árbol de clasificación óptimo.

Figura 6.9: Árbol óptimo para la variable asistencia.



Fuente:ECH2011, elaboración propia.

Se aprecia que las variables que participan el proceso de partición son la edad, la condición de actividad, los años de educación, el clima educativo y el tener hijos a cargo. Estas son las variables que más discriminan a los jóvenes que asisten de los que no asisten a un centro educativo, esto se ve reflejado en el hecho de que las mismas son las primeras que ingresan en el algoritmo de partición recursiva. Si se considera un árbol de mayor complejidad se aprecia que el algoritmo recurre a las mismas variables anteriores y por ejemplo solo se agrega la variable Computadora (Ver anexo de resultados).

Como complemento al MLG el árbol se puede emplear para analizar diferentes "trayectorias" que pueden tener los jóvenes para ser clasificados como que no asisten a un centro educativo. A modo de ejemplo se puede apreciar que si el joven tiene menos de 6,5 años acumulados de educación y es mayor a 15, 5 años no asiste. Por otro lado si el joven tiene más de 6,5 años de educación, es inactivo y no tiene hijos a su cargo entonces asiste. Si a la "trayectoria" anterior agregamos el tener hijos a cargo entonces el joven no asiste a un centro educativo. Para evaluar el poder predictivo del árbol se seleccionó una muestra de prueba diferente de la que se usó para construir la regla de clasificación. Se tomó la misma muestra que se empleó para evaluar el po-

der predictivo del Modelo 1 como forma de asegurar resultados comparables. En las siguientes tablas se presenta una comparación entre lo observado y lo predicho por el modelo (regla de clasificación), tanto para la muestra de entrenamiento como la muestra de prueba.

Cuadro 6.20: Poder predictivo del árbol óptimo (datos de entrenamiento).

observado \ predicho	$\hat{Y} = 0$	$\hat{Y} = 1$	Total
Y = 0	686	643	1329
Y = 1	220	5545	5765
Total	906	6188	7094

Fuente: ECH2011, elaboración propia.

Cuadro 6.21: Poder predictivo del árbol óptimo (datos de prueba).

observado \ predicho	$\hat{Y} = 0$	$\hat{Y} = 1$	Total
Y = 0	150	182	332
Y = 1	42	1400	1442
Total	192	1582	1774

Fuente: ECH2011, elaboración propia.

Si se analiza el error del clasificador con los datos de entrenamiento, se aprecia que el mismo presenta una tasa de error global calculada del orden del 12,16 %. La tasa de error para el grupo de los que no asisten es de 48,4 % y para los que asisten es de 3,8 %. Se nota una elevada tasa de error para el grupo de los que no asisten. Si se analiza la performance del clasificador con los datos de prueba se aprecia que la tasa de error global es de 12,6 %. La tasa de error para los que no asisten es de 54,82 % y para los que asisten es de 2,9 %. Surge de este modo como la regla de clasificación comete menos error al predecir los que asisten con respecto a los que no asisten, no se produce una mejora de la performance con los datos de prueba al contrario el error en el grupo de los que no asisten es mayor aún, esto indicaría un ajuste del clasificador a los datos de entrenamiento. Si se compara con el poder predictivo del Modelo 1, éste tiene una mejor performance que el árbol. La regla de clasificación aporta a las características descriptivas del problema de la asistencia pero su rendimiento en cuanto a predicción no resulta del todo satisfactorio. Esto se

debe fundamentalmente al hecho de que los árboles de clasificación son muy sensibles a los datos con los cuales se trabaja. Como alternativa Breiman propone el uso de otras técnicas, como ser Bagging y Random Forest donde tiene un mayor peso la predicción y no el carácter descriptivo y el análisis de "trayectorias" mediante las cuales se puede tomar la decisión de asistir/no asistir.

Escenario 2.

Para este escenario se consideró el siguiente conjunto de predictores acorde al Modelo 2: sexo, edad, condición de actividad del jóven, condición de jefe del hogar, clima educativo, años de educación, hijos a cargo, madre ausente, índice de calidad de vivienda, residencia del joven, logaritmo del ingreso del hogar sin jóven en promedio, índice TIC, índice de confort, hacinamiento.

Para este escenario se adoptó la misma estrategia de obtener árbol maximal para posteriormente considerar la secuencia de árboles anidados y de ese modo obtener el óptimo. En la siguiente figura se aprecia un resumen del árbol maximal obtenido bajo el escenario 2.

Figura 6.10: Árbol maximal para la variable asistencia.

```
Classification tree:
rpart(formula = Asiste ~ sexo + E27 + JovenActivo + Madre_ausente +
    hijos + jefe + aniosed + climaeducativo + icv2 + Confort1 +
    TIC1 + Mdeo + LN_YSVL_sin_joven_prom + Hacinamiento, data =
per_training,
    method = "class", control = rpart.control(cp = 0))

Variables actually used in tree construction:
 [1] aniosed                climaeducativo         Confort1
E27
 [5] Hacinamiento           hijos                  JovenActivo
LN_YSVL_sin_joven_prom
 [9] Mdeo                   sexo                   TIC1

Root node error: 1329/7094 = 0.18734

n= 7094

           CP nsplit rel error  xerror     xstd
1  0.12678706      0   1.00000 1.00000 0.024728
2  0.03235515      2   0.74643 0.74643 0.021980
3  0.01617758      3   0.71407 0.73740 0.021868
4  0.00376223      7   0.64936 0.65914 0.020850
5  0.00357412     10   0.63807 0.65839 0.020840
6  0.00351141     15   0.61701 0.65914 0.020850
7  0.00263356     20   0.59518 0.66516 0.020932
8  0.00225734     23   0.58691 0.66065 0.020870
9  0.00150489     30   0.57111 0.66140 0.020881
10 0.00112867     46   0.54552 0.67570 0.021073
11 0.00100326     64   0.51994 0.68322 0.021173
12 0.00084650     68   0.51543 0.69601 0.021341
13 0.00075245     79   0.50489 0.70128 0.021409
14 0.00062704     85   0.50038 0.70579 0.021467
15 0.00050163     91   0.49661 0.71859 0.021631
16 0.00045147    112   0.48608 0.72837 0.021755
17 0.00037622    122   0.48157 0.74567 0.021970
18 0.00025082    140   0.47479 0.74492 0.021961
19 0.00000000    146   0.47329 0.75696 0.022109
```

Fuente: ECH2011, elaboración propia.

La salida anterior presenta un breve ajuste del modelo, se aprecia desde el árbol mínimo (un único nodo) al árbol maximal que en este caso presenta 147 nodos terminales. Las columnas de error han sido re-escaladas de forma que el primer nodo tenga un error igual a 1. La columna "xerror" representa el error de validación cruzada y "xstd" es el error estándar. También se brinda información respecto a la cantidad de variables que participaron en la partición y cuales fueron. En este caso de las 14 variables se tomaron 11 en el proceso. Se aprecia al mismo tiempo como las variables Confort y TIC son incluidas en la partición siendo esta la diferencia sustancial con el escenario 1 donde las mismas aparecían en forma desagregada y no como resultado de aplicar una técnica factorial.

Del mismo modo que se procedió para el caso 1 se considerará la secuencia de diferentes árboles anidados para distintos parámetros de complejidad (α), es decir que se considerarán diferentes árboles contenidos en el maximal que se obtienen podando sucesivas ramas para diferentes valores de α. En la tabla siguiente se aprecia dicha secuencia y la evaluación de la respectiva

complejidad y la tasa de error global del árbol resultante.

Cuadro 6.22: Secuencia de árboles anidados bajo escenario 2.

Parámetro de complejidad (α)	Tasa de error global	Complejidad del árbol (nodos terminales)
0,001	13,50 %	69
0,002	13,00 %	31
0,003	13,13 %	21
0,004	12,63 %	8
0,020	14,00 %	4

Fuente: ECH2011, elaboración propia.

Las tasas de error global se obtuvieron evaluando el poder predictivo del árbol en la respectiva muestra de prueba. Evaluando en forma conjunta complejidad y error global se concluye que el árbol óptimo es el que posee 8 nodos terminales y error global del orden de 12,63 % ($cp = 0,004$). Como se puede ver la regla de clasificación óptima coincide con la que se obtuvo bajo el escenario 1, no se producen cambios en lo que refiere a considerar las variables de confort y tic's en forma desagregada o bajo la forma de índice. Incluso las variables que participan en la partición son de hecho las mismas. De hecho para ver el ingreso de las variables TIC y Confort actuando en el proceso de partición se debe considerar un árbol de mayor complejidad que el óptimo (ver anexo de Resultados, sección árboles de clasificación).

En el siguiente cuadro se aprecia el poder predictivo de la regla de clasificación óptima, se comparan las categorías observadas para la variable Asistencia con las predichas por el modelo.

Cuadro 6.23: Poder predictivo del árbol óptimo (datos de entrenamiento).

predicho / observado	$\hat{Y} = 0$	$\hat{Y} = 1$	Total
Y = 0	686	643	1329
Y = 1	220	5545	5765
Total	906	6188	7094

Fuente: ECH2011, elaboración propia.

Cuadro 6.24: Poder predictivo del árbol óptimo (datos de prueba).

observado \ predicho	$\hat{Y} = 0$	$\hat{Y} = 1$	Total
Y = 0	150	182	332
Y = 1	42	1400	1442
Total	192	1582	1774

Fuente: ECH2011, elaboración propia.

Se aprecia que el poder predictivo de la regla de clasificación es idéntica a la que se obtuvo bajo el escenario 1. No existe una mejora en la predicción al incorporar los índices de TIC's y de Confort, el resultado es el mismo.

Capítulo 7

Conclusiones.

El objetivo general del presente trabajo consiste en la caracterización de los jóvenes uruguayos de entre 14 a 17 años de edad que no asisten a un centro educativo. Para llevar a cabo dicha caracterización se evalua un conjunto de variables que reflejaban aspectos del individuo, aspectos de los hogares donde los mismos residen, asi como sus condiciones de vida. Luego se plantea un modelo lineal generalizado para poder estimar la probabilidad de que un joven asista a un centro educativo dado un conjunto de características.

Se plantean dos escenarios, el primero de ellos con la totalidad de las variables socio-demográficas, de confort, de tecnología e información y el indicador de vivienda. El segundo de ellos con la totalidad de las variables socio-demográficas, el Índice de confort y el Índice de TIC.

Se utilizan distintas estrategias para la selección de variables, como ser la metodología de selección hacia atrás o "backward" y el "test de razón de verosimilitud".

Como resultado se obtienen cuatro modelos finales. Estos modelos difieren de la incorporación de seis variables: Montevideo, Computadora, Microondas, Hacinamiento, Teléfono, TIC, manteniendo el resto de las variables en los cuatro modelos.

Si se evalua la predicción, como se mencionó en el capítulo anterior los cuatro modelos tienen un poder predictivo similar, siendo el que no incorpora la región y el índice de TIC (modelo final 3) el que se desempeña mejor. Este modelo es el que presenta mayor área bajo la curva ROC, para un punto de corte de 0,86.

De todas formas resulta interesante observar que cualquiera de los cuatro modelos pueden ser

utilizados a la hora de estudiar la asistencia, ya que los mismos no presentan grandes cambios en su predicción. En otras palabras, para cierta población puede ser de interés utilizar el modelo final 2 el cual incorpora el indicador de tecnologías de infomración (TIC) para así poder evaluar como impacta este indicador en la asistencia para cierta población.

En este contexto se estudia la estimación de los coeficientes para los cuatro modelos finales. Como resultado de la estimación y comparando la asistencia según sexo, se observa que los hombres tienen mayor probabilidad de no asistir que las mujeres dadas todas las demás variables constantes. Se destaca por otra parte el efecto que tiene la edad, la cual incrementa la probabilidad de no asistencia a medida que la misma aumenta. Del mismo modo características como el ser activo, tener hijos a cargo y la condición de ser jefe del hogar tienen un impacto negativo en la asistencia, lo cual se ve reflejado en el signo negativo de los coeficientes estimados asociados a dichas variables.

En cuanto a los años de educación del joven, su impacto es positivo, conforme aumentan los años de educación mayor es la probabilidad de asistencia a un centro educativo. Si se considera el clima educativo del hogar podemos concluir que cuanto mayor sea el promedio de años de educación de los mayores de 18 años que residen en dicho hogar (o en su defecto del jefe del hogar) mayor también la probabilidad de asistencia del joven. Esto se puede relacionar con las mayores habilidades que poseen los padres que habitan en hogares con mayor clima educativo para acompañar los procesos de aprendizaje de sus hijos.

Otro aspecto que se puede destacar es el hecho de que el logaritmo del ingreso promedio sin valor locativo excluido el ingreso del joven no fue una variable significativa en ninguno de los cuatro modelos propuestos, esto muestra como el fenómeno de la no asistencia no es exclusivo de los jóvenes que habitan en hogares con menos recursos, sino que se relaciona con otras características del hogar. El ingreso no es determinante a la hora de asistir o no. Por otro lado, en referencia a las características de los hogares donde los jóvenes residen se puede concluir que cuanto mejor son las condiciones estructurales de la vivienda así como la ausencia de hacinamiento (únicamente en el MF1) mayor será la probabilidad de asistencia del joven. De estas dos últimas observaciones se destaca que por mas que el ingreso no sea un factor determinanate directo en la asistencia, puede ser un factor determinante indirecto en la misma, ya que el aumento del ingreso del hogar a largo plazo puede mejorar la situación estructural de la vivienda presentando entonces un impacto positivo en la asistencia.

Las ocho variables mencionadas anteriormente se comportan de forma similar en los cuatro

modelos. Ahora si se toma en consideración el lugar de residencia de los jóvenes (MF1 y MF2) se concluye una mayor probabilidad de asistencia en los jóvenes que residen en Montevideo con respecto a los que residen en el Interior del país. Esto se puede ver explicado por la mayor oferta educativa presente en la capital con respecto al Interior, así como también por las dificultades de acceso o por la cercanía del centro educativo más próximo.

Considerando los elementos de confort de los hogares donde el joven reside se destaca la presencia microondas con un efecto positivo en la asistencia tanto en el MF1 como en el MF4.

Tener computadora en el hogar es un factor que promueve una mayor asistencia entre los jóvenes. Esto se puede relacionar con el impacto cada vez más preponderante que tienen las tecnologías de la información y de las comunicaciones en la educación, esto se observa en los modelos MF1 y MF3, lo mismo sucede con el teléfono en el MF3.

De forma similar en los modelos con índices (MF2 y MF4) la variable TIC presenta un signo negativo en la salida lo que se traduce en un efecto positivo en la asistencia. Un bajo valor del índice significa que el hogar se encuentra en una buena situación en referencia a las TIC, el signo negativo en la salida del modelo representa entonces que las TIC's tienen un efecto positivo en la asistencia.

Como forma de complementar los resultados del modelo para la asistencia se realizó un árbol de clasificación para dicha variable. Los resultados se encuentran en coherencia con los que se obtuvieron para el modelo de asistencia. Se destaca cómo las variables que discriminan más entre los jóvenes que asisten de los que no son la edad del joven, los años acumulados de educación, su condición de actividad, el tener hijos a cargo y el clima educativo del hogar. Las que aportan en menor medida son el ingreso del hogar y algunos elementos de confort como ser el poseer teléfono, computadora, calefón, microondas y cable. Este resultado no se vio alterado por el hecho de incluir en la partición los índices de confort y tecnologías de la información. Es de destacar que todos los resultados a los que se arriba en el presente trabajo se encuentran en línea con las hipótesis planteadas, así como también con investigaciones previas sobre la asistencia en el Uruguay. En cuanto al poder predictivo del clasificador es bajo, se aprecia una elevada tasa de error en el grupo de los que no asisten, esto puede deberse al ajuste de la técnica a los datos de entrenamiento. Por lo tanto el clasificador sólo permite el estudio de diferentes "trayectorias". Si el objetivo es hacer predicción se deberían aplicar otras técnicas como ser Bagging o Random Forests, donde tiene un mayor peso la predicción de la variable de respuesta y no el estudio de "trayectorias".

Como ya se ha mencionado con anterioridad, el fenómeno de la asistencia comprende varios aspectos que van desde aspectos microsociales a macrosociales. En el presente trabajo se puso en relieve una de las aristas del fenómeno que tuvo como eje a los jóvenes como centro de la problemática. Sería recomendable en posteriores estudios indagar sobre los motivos de la no asistencia a un centro educativo, a modo de ejemplo se observan los motivos por los cuales no finalizan sus estudios los jóvenes de entre 14 a 17 años. El 63 % de los jóvenes no finalizan la educación media superior por falta de interés, el 9 % porque le resultada dificil, el 6 % por comenzar a trabajar, el 5 % por asuntos familiares, el 4 % por embarazo y el 13 % restante por otras razones. Siguiendo con la línea de investigación, se podría estudiar la población de jóvenes que se desvinculan del sistema mostrando como principal motivo el desinterés (63 %).

Por otro lado, los autores también quieren hacer énfasis en la necesidad de contar con una base de datos que estudie a los jóvenes en forma longitudinal (datos de panel), si lo que se pretende es tener una mejor aproximación al fenómeno de la deserción educativa en el Uruguay. Bien sabemos por todo lo expuesto que la asistencia es un factor de riesgo, pero que el fenómeno de la deserción tiene más que ver con las trayectorias educativas que adoptan los jóvenes, por lo tanto un proceso que se va desarrollando en el tiempo.

Bibliografía

[1] Agresti, A.(2007). *An Introduction to Categorical Data Analysis*. Second edition. Wiley.

[2] Blanco, J.(2006). *Introducción al Análisis Multivariado: Teoría y aplicaciones a la realidad latinoamericana*. Universidad de la República. Facultad de Ciencias Económicas y de Administración.

[3] Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J.(1984). *Classification and Regression Trees*. Chapman Hall.

[4] Canto Bonilla, P. A.(2010). *Factores individuales, familiares e institucionales relacionados con la deserción en una escuela preparatoria estatal de Yucatán*. Universidad Autónoma de Yucatán. Facultad de Educación.

[5] Casacuberta, C.(2006). *Situación de la Vivienda en Uruguay: Informe de divulgación*. Encuesta Nacional de Hogares Ampliada-Módulo Vivienda. INE.

[6] Casacuberta, C.; Buchelli, M.(2010). *Asistencia a instituciones educativas y actividad laboral de los adolescentes en Uruguay, 1986-2008*. Universidad de la República.CSIC.

[7] Cerdas, A. M.(2003). *Deserción escolar y trabajo infantil en Costa Rica*. Universidad Pontificia Católica de Chile. Instituto de Economía.

[8] Damonte, C.; Monteverde, M.; Pérez, V.; Sotelo, R.(2012). *Análisis de datos provenientes de diseños muestrales complejos: Aplicaciones a la Encuesta de Hogares y Empleo de la Provincia de Bs. As.*. Versión presentada para el V Congreso de la Asociación Latinoamericana de Población. Montevideo, Uruguay.

[9] De Los Campos, H.(2000). *El Índice de Necesidades Básicas Insatisfechas: Crítica de la definición oficial y propuesta de una metodología alternativa*. Universidad de la República. Facultad de Ciencias Sociales. Departamento de Trabajo Social.

[10] Fernández, T.; Cardozo, S.; Pereda, C.(2010). *Desafiliación y desprotección social*. Universidad de la República. CSIC.

[11] Fernández, T.(2010). *Enfoques para explicar la desafiliación*. Universidad de la República. CSIC.

[12] Fernández, T.(2010). *Factores escolares y desafiliación en la Enseñanza Media Superior(2003-2007)* Universidad de la República. CSIC.

[13] Ferrari, F.; Martínez, J. P.; Saavedra, E.(2010). *Identificación y dimensionamiento económico de alternativas para favorecer la permanencia o motivar el retorno a la educación media*. Universidad de la República. Facultad de Ciencias Económicas y de Administración. Instituto de Economía.

[14] Greenacre, J. M.(1984). *Theory and Applications of Correspondence Analysis*. Academic Press London.

[15] Greenacre, J. M.(1993). *Correspondence Analysis in Practice*. Academic Press London.

[16] Hosmer, D., Lemeshow, D.(2000). *Applied Logistic Regression*. Wiley & Sons.

[17] Llambí, C.(2009). *Propuesta de Relevamiento de Información sobre Educación para el VIII Censo de Población, IV de Hogares y VI de Viviendas (2010)*. INE.

[18] Long, J. S.; Freese, J.(2001). *Regression Models for Categorical Dependent Variables Using Stata*

[19] Lumley, T.(2010). *Complex Surveys, a guide to analysis using R*. Wiley.

[20] Nalbarte, L.; Castrillejo, A.; Debera, L.; Altmark, S.(2006) *Elaboración de pruebas diagnósticas al ingreso a la Facultad de Ciencias Económicas y de Administración*. Universidad de la República. Facultad de Ciencias Económicas y de Administración. Instituto de Estadística.

[21] Peña, D. (2002). *Análisis de Datos Multivariantes*. Alianza.

[22] Rencher, A.(2000). *Linear Models in Statistics*. John Wiley & Sons, Inc.

[23] Rencher, A.(1995). *Methods of Multivariate Analysis*. Wiley & Sons.

[24] Reporte Social 2011. *Principales Características del Uruguay Social*. OPP-MIDES.

[25] Rico, D. A.(2006). *Caracterización de la Deserción Estudiantil en la Universidad Nacional de Colombia Sede Medellín*. Universidad Nacional de Colombia Sede Medellín. Oficina de Planeación.

[26] Thernau, T. M.; Atkinson, E. J.(1997) *An Introduction to Recursive Partitioning Using the RPART Routines*. Division of Biostatistics, Mayo Foundation.

Capítulo 8

Anexo Metodológico.

A continuación se presenta la estimación de los parámetros en un modelo lineal generalizado y las pruebas de significación.

8.1. Estimación del modelo.

La derivación de estimadores muestrales en un MLG se realiza por el método de Máxima Verosimilitud. La lógica de este método consiste en dada una muestra, se elige como valor estimado aquel que maximiza la probabilidad (verosimilitud) de que precisamente esa muestra sea la observada.

Supongamos una muestra $(Y_1, Y_2, ..., Y_n)$ donde cada Y_i es una variable dicotómica que puede tomar el valor cero o uno y son mutuamente independientes.

Dada una muestra aleatoria, la probabilidad conjunta de observar a los n individuos será: $Pr(Y_1 = y_1, ..., Y_n = y_n) = \prod Pr((Y_1 = y_1, ..., Y_n = y_n)/ \underline{X}\underline{\beta}) = L(\underline{\beta}; y_1, ..., y_n) = \prod p^y(1-p)^{1-y}$

La probabilidad conjunta $L(\underline{\beta}; y_1, ..., y_n)$ como función de los parámetros es la función de verosimilitud de la muestra $(Y_1, Y_2, ..., Y_n)$. Dicha función para cada valor de los parámetros informa cuan verosimil (probable) resulta que se haya generado la muestra que observamos $(y_1, ..., y_n)$. En general resulta más conveniente trabajar con el logaritmo de la función de verosimilitud (log-verosimilitud), en este caso el logaritmo de la función queda

$\log L(\underline{\beta}; y_1, ..., y_n) = \sum_{i=1}^{i=n} y_i \log(p_i) + (1 - y_i) \log(1 - p_i)$

La estimación de los parámetros será aquella que haga máxima dicha función. Para la deri-

vación de los estimadores máximos verosímiles se emplean métodos numéricos, por lo general el método más empleado es el de Newton-Raphson.

8.2. Significación del modelo.

Posteriormente al ajuste del modelo y estimación de los parámetros debemos someterlo a determinadas pruebas de hipótesis entre ellas tenemos la prueba de significación del modelo y las de significación de parámetros. En la prueba de significación del modelo nos planteamos:

$H_0) \beta_1 = \beta_2 = ... = \beta_k = 0$

$H_1)$ Algún $\beta_k \neq 0$

El estadístico asociado a la prueba de hipótesis anterior es $Q = -2 \log \frac{L_0}{L_1}$ que sigue una distribución asintótica chi- cuadrado con q grados de libertad. En este caso L_0 corresponde a la verosimilitud bajo H_0 cierta y L_1 bajo H_1; los grados de libertad q surgen de la diferencia de parámetros estimados bajo H_1 menos la cantidad de parámetros estimados bajo H_0. Se rechaza la hipótesis nula cuando el p-valor sea menor al nivel de significación de la prueba o bien cuando el valor del estadístico observado (Q obs) sea mayor que el valor crítico (Q crítico).

8.3. Significación de los parámetros.

Asociada a la prueba de significación del modelo, que nos permite identificar si el modelo con el cual se va a trabajar es significativo en su globalidad, se encuentran las pruebas de significación de parámetros. En dichas pruebas se puede analizar en forma individual si una variable es significativa o no y por lo tanto tomar decisiones en cuanto a incluirlas en el modelo con el que se va a trabajar o por el contrario descartarlas.

La forma general de la prueba de significación de parámetros es la siguiente:

$H_0) \beta_k = 0$
$H_1) \beta_k \neq 0$

El estadístico asociado a dicha prueba es W (estadístico de Wald) que se define como $W = \frac{\hat{\beta}}{\sqrt{\hat{V}(\hat{\beta})}}$

que tiene una distribución asintótica $N(0,1)$ bajo H_0 cierta . Se rechaza la hipótesis nula cuando el p-valor sea menor al nivel de significación de la prueba (α).

Capítulo 9

Anexo de Resultados.

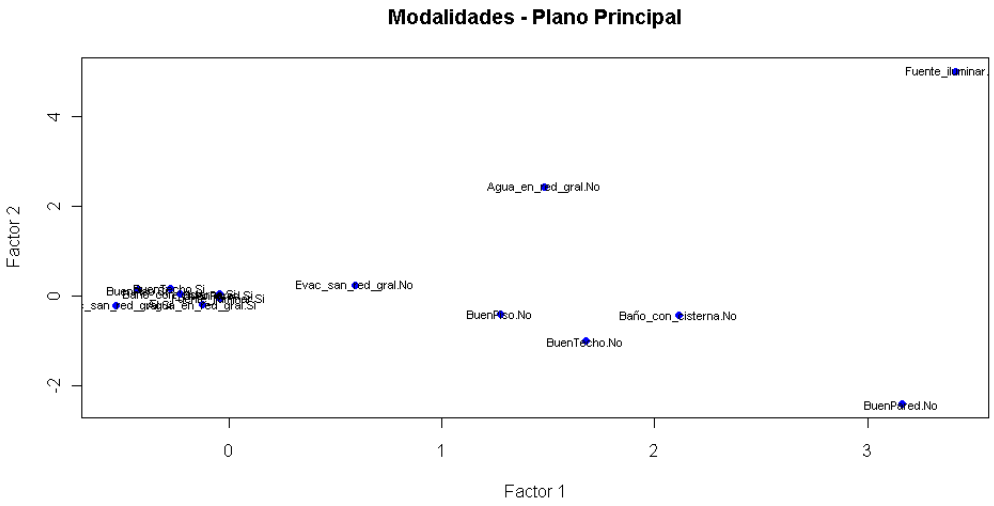
9.1. Cálculo del índice de vivienda y salubridad mediante ACM.

Para el i-ésimo hogar el valor del índice VIV será el resultado de:

$$VIV_i = 1/v\lambda^* \; [\beta_{Agua\_en\_red\_graIS} Agua\_en\_red\_graIS + \beta_{Agua\_en\_red\_graIN} Agua\_en\_red\_graIN + \beta_{bañosisS} BañosisS + \beta_{bañosisN} BañosisN + \beta_{evac\_red\_graIS} Evac\_red\_graIS + \beta_{evac\_red\_graIN} Evac\_red\_graIN + \beta_{BuenTechoS} BuenTechoS + \beta_{BuenTechoN} BuenTechoN + \beta_{BuenPisoS} BuenPisoS + \beta_{BuenPisoN} BuenPisoN + \beta_{BuenasParedesS} BuenasParedesS + \beta_{BuenasParedesN} BuenasParedesN + \beta_{FuenteIluminarS} FuenteIluminarS + \beta_{FuenteIluminarN} FuenteIlumniarN] / 7$$

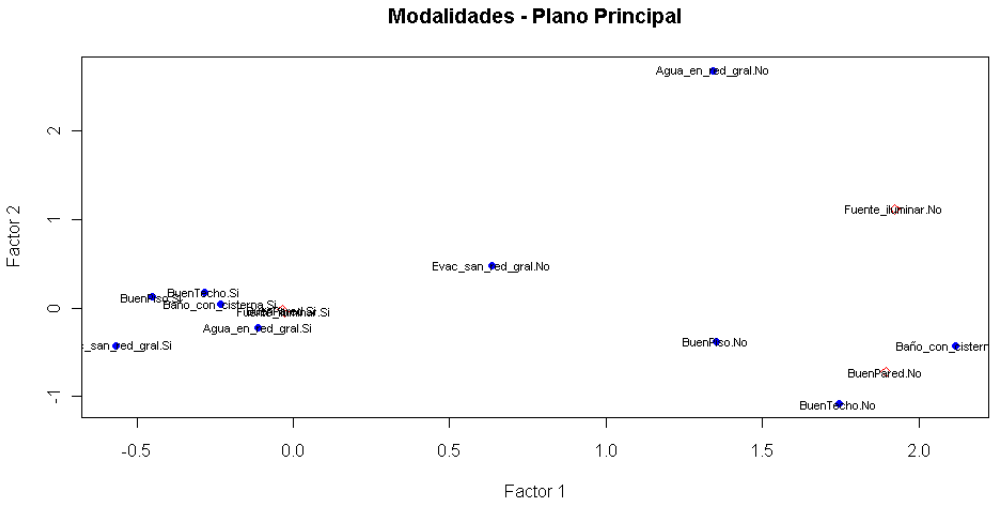
9.2. Resultados del ACM para el índice de vivienda y salubridad.

Figura 9.1: Índice de Vivienda y Salubridad.



Fuente: ECH2011, elaboración propia.

Figura 9.2: Índice de vivienda y salubridad variables suplementarias.



Fuente: ECH2011, elaboración propia.

```
        Componentes Disponibles

 [1]  "INERCIA e INERCIA ACUMULADA"
 [2]  "Masa"
 [3]  "Inercia de las Modalidades"
 [4]  "Contribuciones Parciales de las Modalidades"
 [5]  "Calidad de Representacion(cos2) de las Modalidades"
 [6]  "Coordenadas Modalidades"
 [7]  "Coordenadas Individuos"
 [8]  "Coordenadas de Modalidades Suplementarias"
 [9]  "Ajuste de la inercia: Benzecri"
 [10] "Ajuste de la inercia:Greenacre"
```

```
> acmviv [1]
[[1]]
[[1]]$TOT
     inertia       cum     ratio
1 0.42205040 0.4220504 0.4220504
2 0.21418887 0.6362393 0.6362393
3 0.14195063 0.7781899 0.7781899
4 0.12307072 0.9012606 0.9012606
5 0.09873937 1.0000000 1.0000000
```

```
> acmviv [2]
[[1]]
Agua_en_red_gral.No    0.01569284
Agua_en_red_gral.Si    0.18430716
Baño_con_cisterna.No   0.01967440
Baño_con_cisterna.Si   0.18032560
Evac_san_red_gral.No   0.09444324
Evac_san_red_gral.Si   0.10555676
BuenTecho.No           0.02808531
BuenTecho.Si           0.17191469
BuenPiso.No            0.05004198
BuenPiso.Si            0.14995802
```

```
> acmviv [3]
[[1]]
                    Inercia de las Modalidades
Agua_en_red_gral.No                  0.18719463
Agua_en_red_gral.Si                  0.01280537
Baño_con_cisterna.No                 0.18080537
Baño_con_cisterna.Si                 0.01919463
Evac_san_red_gral.No                 0.10410738
Evac_san_red_gral.Si                 0.09589262
BuenTecho.No                         0.17205369
BuenTecho.Si                         0.02794631
BuenPiso.No                          0.14955705
BuenPiso.Si                             0.05044295
```

```
> acmviv [4]
[[1]]
                          Contr_1     Contr_2
Agua_en_red_gral.No  0.066836359 0.524961402
Agua_en_red_gral.Si  0.005690784 0.044697855
Baño_con_cisterna.No 0.208888163 0.016530677
Baño_con_cisterna.Si 0.022790713 0.001803577
Evac_san_red_gral.No 0.090315121 0.100036273
Evac_san_red_gral.Si 0.080806316 0.089503979
BuenTecho.No         0.202815566 0.153076672
BuenTecho.Si         0.033133520 0.025007789
BuenPiso.No          0.216481990 0.033277016
BuenPiso.Si          0.072241469 0.011104760

> acmviv [5]
[[1]]
                          Cos2_1     Cos2_2
Agua_en_red_gral.No  0.1229630 0.49014149
Agua_en_red_gral.Si  0.1905002 0.75935116
Baño_con_cisterna.No 0.4757131 0.01910533
Baño_con_cisterna.Si 0.5024541 0.02017929
Evac_san_red_gral.No 0.3717556 0.20897155
Evac_san_red_gral.Si 0.3507680 0.19717396
BuenTecho.No         0.4950474 0.18962129
BuenTecho.Si         0.5007932 0.19182216
BuenPiso.No          0.6158079 0.04803971
BuenPiso.Si          0.6028199 0.04702651
```

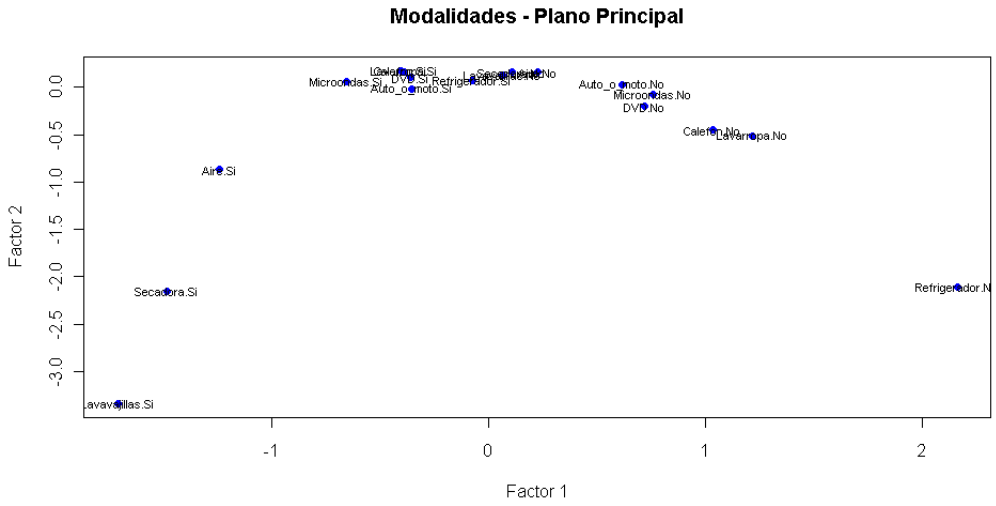
9.3. Cálculo del índice de confort.

Para el i-ésimo hogar el valor del Ìndice de confort será el resultado de computar:

$$IConf_i = 1/v\lambda * [(\beta_{microN}MicroN + \beta_{microS}MicroS + \beta_{calefN}CalefN + \beta_{calefS}CalefS + \beta_{dvdN}DvdN + \beta_{dvdS}DvdS + \beta_{lavarrN}LavarrN + \beta_{lavarrS}LavarrS + \beta_{aireN}AireN + \beta_{aireS}AireS + \beta_{auto\_o\_motoN}Auto\_o\_motoN + \beta_{autoS}Auto\_o\_motoS + \beta_{refrigN}RegrifN + \beta_{refrigS}RefrigS + \beta_{lavavajN}LavavajN + \beta_{lavavajS}LavavajS + \beta_{secadN}SecadN + \beta_{secadN}SecadS)] / 9.$$

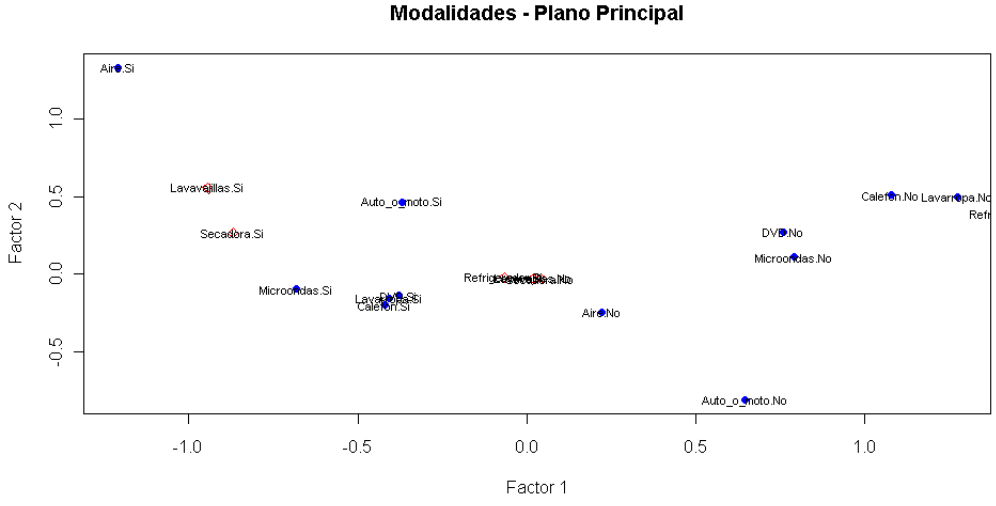
9.4. Resultados del ACM para el índice de confort.

Figura 9.3: Índice de Confort.



Fuente: ECH2011, elaboración propia.

Figura 9.4: Índice de confort con variables suplementarias.



Fuente: ECH2011, elaboración propia.

```
         Componentes Disponibles

 [1]  "INERCIA e INERCIA ACUMULADA"
 [2]  "Masa"
 [3]  "Inercia de las Modalidades"
 [4]  "Contribuciones Parciales de las Modalidades"
 [5]  "Calidad de Representacion(cos2) de las Modalidades"
 [6]  "Coordenadas Modalidades"
 [7]  "Coordenadas Individuos"
 [8]  "Coordenadas de Modalidades Suplementarias"
 [9]  "Ajuste de la inercia: Benzecri"
 [10] "Ajuste de la inercia:Greenacre"
```

```
> acmconfortsup[1]
[[1]]
[[1]]$TOT
     inertia       cum     ratio
1 0.38382360 0.3838236 0.3838236
2 0.15468771 0.5385113 0.5385113
3 0.13810306 0.6766144 0.6766144
4 0.13045653 0.8070709 0.8070709
5 0.09862980 0.9057007 0.9057007
6 0.09429929 1.0000000 1.0000000
```

```
> acmconfortsup [2]
[[1]]

Calefon.No      0.04662844
Calefon.Si      0.12003823
DVD.No          0.05545757
DVD.Si          0.11120910
Lavarropa.No    0.04040263
Lavarropa.Si    0.12626404
Microondas.No   0.07715637
Microondas.Si   0.08951030
Aire.No         0.14088950
Aire.Si         0.02577717
Auto_o_moto.No  0.06071879
Auto_o_moto.Si  0.10594787
```

```
> acmconfortsup [3]
[[1]]
              Inercia de las Modalidades
Calefon.No                     0.12241611
Calefon.Si                     0.04425056
DVD.No                         0.11136465
DVD.Si                         0.05530201
Lavarropa.No                   0.12686801
Lavarropa.Si                   0.03979866
Microondas.No                  0.09031320
Microondas.Si                  0.07635347
Aire.No                        0.02543624
Aire.Si                        0.14123043
Auto_o_moto.No                 0.10747204
Auto_o_moto.Si                 0.05919463
```

```
> acmconfortsup [4]
[[1]]
                  Contr_1     Contr_2
Calefon.No     0.14122365 0.078005039
Calefon.Si     0.05485784 0.030300787
DVD.No         0.08345959 0.026507774
DVD.Si         0.04161949 0.013218854
Lavarropa.No   0.17070326 0.064191846
Lavarropa.Si   0.05462252 0.020540442
Microondas.No  0.12541707 0.006312764
Microondas.Si  0.10810741 0.005441497
Aire.No        0.01801481 0.054162132
Aire.Si        0.09846300 0.296032323
Auto_o_moto.No 0.06580085 0.257635484
Auto_o_moto.Si 0.03771051 0.147651058
```

```
> acmconfortsup [5]
[[1]]
                  Cos2_1     Cos2_2
Calefon.No     0.4202120 0.09354225
Calefon.Si     0.4852556 0.10802143
DVD.No         0.2868406 0.03671654
DVD.Si         0.2892641 0.03702676
Lavarropa.No   0.5087216 0.07709787
Lavarropa.Si   0.5293068 0.08021759
Microondas.No  0.5274655 0.01069993
Microondas.Si  0.5483232 0.01112304
Aire.No        0.2724947 0.33017810
Aire.Si        0.2640549 0.31995167
Auto_o_moto.No 0.2291009 0.36151406
Auto_o_moto.Si 0.2480361 0.39139320
```

```
> acmconfortsup[6]
[[1]]
                   Comp1       Comp2
Calefon.No      1.0781871  0.50870243
Calefon.Si     -0.4188180 -0.19760370
DVD.No          0.7600174  0.27191559
DVD.Si         -0.3790043 -0.13559842
Lavarropa.No    1.2734502  0.49575035
Lavarropa.Si   -0.4074852 -0.15863279
Microondas.No   0.7898748  0.11249981
Microondas.Si  -0.6808588 -0.09697294
Aire.No         0.2215345 -0.24385770
Aire.Si        -1.2108344  1.33284562
Auto_o_moto.No  0.6449410 -0.81015697
Auto_o_moto.Si -0.3696161  0.46430147
```

```
> acmconfortsup[8]
[[1]]
                     Axis1       Axis2
Secadora.No     0.04330729 -0.02554736
Secadora.Si    -0.86970855  0.26711442
Lavavajillas.No 0.01864751 -0.02831760
Lavavajillas.Si -0.94492645  0.55959275
Refrigerador.No 1.43525881  0.38239612
Refrigerador.Si -0.06597130 -0.01858305
```

```
> acmconfortsup[9]
$Benzecri
      Inercia Inercia Ajustada      Porcentaje   Porc.Acumulado
   0.38382360       0.06790627      1.00000000      1.00000000
```

```
> acmconfortsup[10]
$Greenacre
      Inercia Inercia Ajustada      Porcentaje   Porc.Acumulado
    0.3838236        0.2953292       1.0000000       1.0000000
```

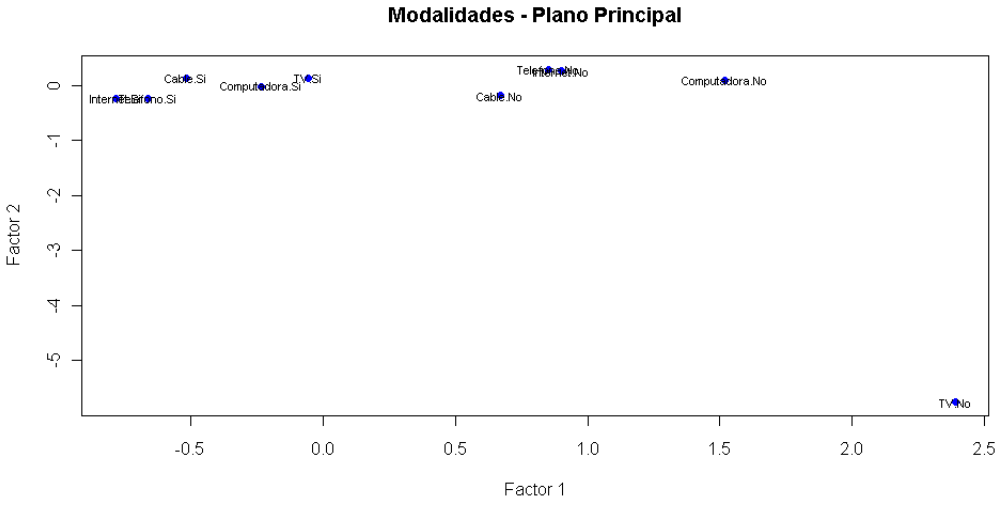
9.5. Cálculo del índice de tecnología de la información y comunicaciones (TIC's).

Para el i-ésimo hogar el valor del índice de TIC's (ACM) será el resultado de computar:

$$ITIC_i = 1/v\lambda * [(\beta_{ComputadraN}ComputadoraN + \beta_{ComputadoraS}ComputadoraS + \beta_{CableN}CableN + \beta_{CableS}CableS + \beta_{InternetN}InternetN + \beta_{InternetS}InternetS + \beta_{TelefonoN}TelefonoN + \beta_{TelefonoS}TelefonoS + \beta_{TVN}TVN + \beta_{TVS}TVS)] \quad / 5.$$

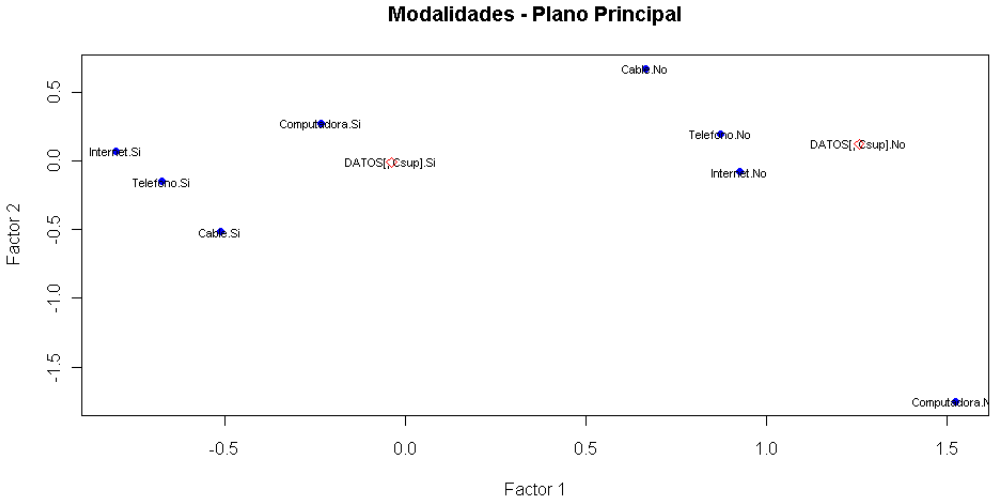
9.6. Resultados del ACM para el índice de TIC's.

Figura 9.5: Indice de TIC.



Fuente: ECH2011, elaboración propia.

Figura 9.6: Indice de TIC con variables suplementarias.



Fuente: ECH2011, elaboración propia.

```
         Componentes Disponibles

 [1]  "INERCIA e INERCIA ACUMULADA"
 [2]  "Masa"
 [3]  "Inercia de las Modalidades"
 [4]  "Contribuciones Parciales de las Modalidades"
 [5]  "Calidad de Representacion(cos2) de las Modalidades"
 [6]  "Coordenadas Modalidades"
 [7]  "Coordenadas Individuos"
 [8]  "Coordenadas de Modalidades Suplementarias"
 [9]  "Ajuste de la inercia: Benzecri"
 [10] "Ajuste de la inercia:Greenacre"
```

```
> acmticsup [1]
[[1]]
[[1]]$TOT
     inertia       cum     ratio
1 0.50786078 0.5078608 0.5078608
2 0.21417037 0.7220311 0.7220311
3 0.18406269 0.9060938 0.9060938
4 0.09390616 1.0000000 1.0000000
```

```
> acmticsup [2]
[[1]]

Cable.No       0.1088394
Cable.Si       0.1411606
Internet.No    0.1162058
Internet.Si    0.1337942
Telefono.No    0.1091946
Telefono.Si    0.1408054
Computadora.No 0.0334591
Computadora.Si 0.2165409
```

```
> acmticsup [3]
[[1]]
              Inercia de las Modalidades
Cable.No                       0.14429530
Cable.Si                       0.10570470
Internet.No                    0.13392617
Internet.Si                    0.11607383
Telefono.No                    0.14060403
Telefono.Si                    0.10939597
Computadora.No                 0.21748322
Computadora.Si                 0.03251678
```

```
> acmticsup [4]
[[1]]
                 Contr_1     Contr_2
Cable.No      0.09475110 0.228540976
Cable.Si      0.07305621 0.176212599
Internet.No   0.19575147 0.003464901
Internet.Si   0.17001838 0.003009413
Telefono.No   0.16348709 0.019575447
Telefono.Si   0.12678415 0.015180749
Computadora.No 0.15257610 0.479868421
Computadora.Si 0.02357550 0.074147494
```

```
> acmticsup [5]
[[1]]
                 Cos2_1      Cos2_2
Cable.No      0.3238805 0.329442245
Cable.Si      0.3587949 0.364956287
Internet.No   0.7414648 0.005534662
Internet.Si   0.7446198 0.005558213
Telefono.No   0.5916035 0.029872645
Telefono.Si   0.5877418 0.029677653
Computadora.No 0.3462571 0.459249896
Computadora.Si 0.3698144 0.490494529
```

```
> acmticsup [6]
[[1]]
                  Comp1       Comp2
Cable.No       0.6649230  0.67060779
Cable.Si      -0.5126774 -0.51706063
Internet.No    0.9249342 -0.07991184
Internet.Si   -0.8033443  0.06940680
Telefono.No    0.8719942  0.19594539
Telefono.Si   -0.6762310 -0.15195555
Computadora.No  1.5218031 -1.75260342
Computadora.Si -0.2351434  0.27080580
```

```
> acmticsup [8]
[[1]]
                      Axis1       Axis2
DATOS[, Csup].No  1.25545974  0.12001262
DATOS[, Csup].Si -0.03921157 -0.01098028
```

```
> acmticsup [9]
$Benzecri
        Inercia Inercia Ajustada       Porcentaje   Porc.Acumulado
      0.5078608        0.1182083        1.0000000        1.0000000
```

```
> acmticsup [10]
$Greenacre
        Inercia Inercia Ajustada       Porcentaje   Porc.Acumulado
      0.5078608        0.3805138        1.0000000        1.0000000
```

9.7. Estudio de asociación entre variables explicativas.

Se analizó la posible asociación entre la variable afro y las variables años de educación y clima educativo. Para ello se realizaron las siguientes pruebas de hipótesis:

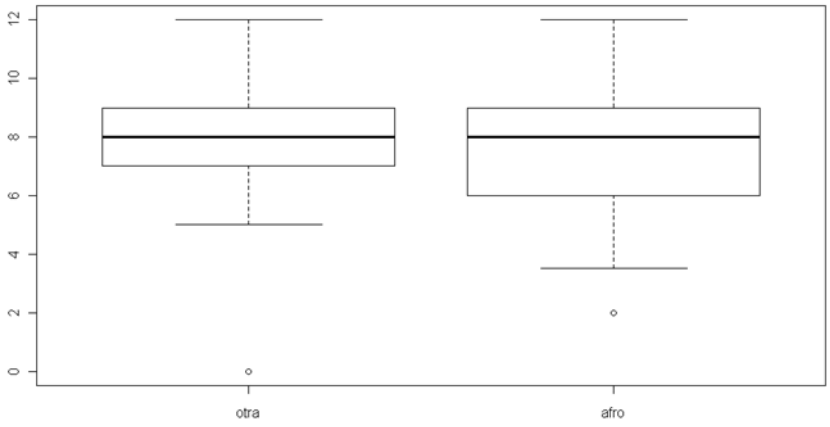
PH1: Afro y años de educación.

$H_0) \mu^{afro}_{añosed} = \mu^{otra}_{añosed}$

$H_1) \mu^{afro}_{añosed} \neq \mu^{otra}_{añosed}$

El resultado de la prueba muestra un p valor de $2,2 * e^{-16}$ menor que 0,05, es decir que a un nivel de 5 % de significación, se rechaza la hipótesis nula de que las medias son iguales para las variables años de educación y afro.

Figura 9.7: Diagrama de caja para las variables años de educación y afro con datos ponderados.



Fuente: ECH2011, elaboración propia

PH2: Afro y clima educativo.

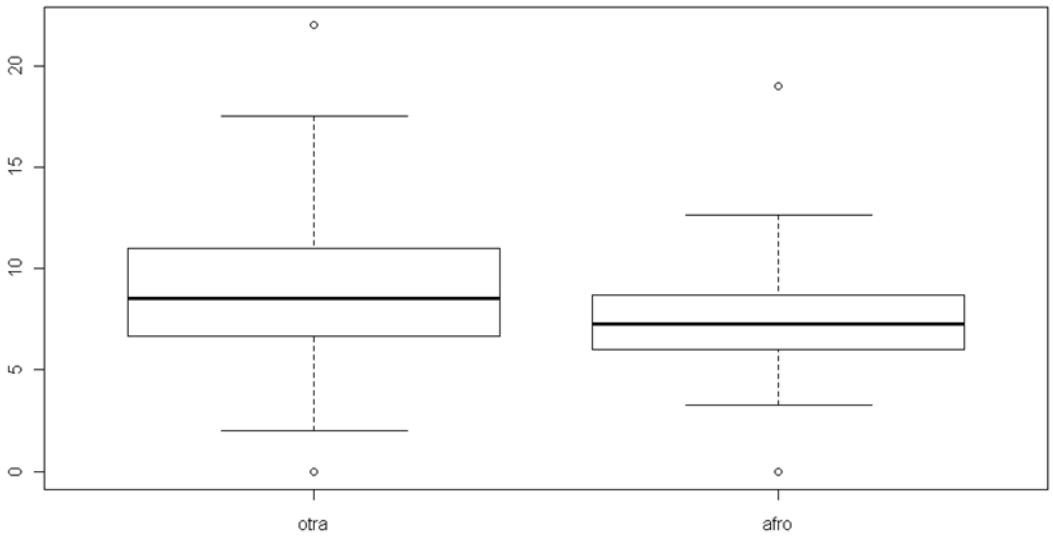
En referencia a las variables afro y clima educativo se plantea la siguiente prueba de hipótesis:

Ho) $\mu^{afro}_{climaeducativo} = \mu^{otra}_{climaeducativo}$

H1) $\mu^{afro}_{climaeducativo} \neq \mu^{otra}_{climaeducativo}$

El resultado de la prueba muestra un p valor de $2,2 * e^{-16}$ menor que 0,05, es decir que a un nivel de 5 % de significación, se rechaza la hipótesis nula de que las medias son iguales para las variables clima educativo y afro.

Figura 9.8: Diagrama de caja para las variables clima educativo y afro con datos ponderados.



Fuente: ECH2011, elaboración propia

Estos resultados podrían revelar a priori la existencia de cierto grado de asociación entre las variables, deberán tenerse en cuenta a la hora de incorporar las mismas al modelo de asistencia.

La posible existencia de asociación entre afro y años de educación y afro y clima educativo podría interferir en el modelo obteniéndose resultados erróneos. Por lo que se debe ser cauteloso a la hora de incoporar estas variables al modelo de asistencia.

9.8. Estrategia de selección razón de verosimilitud partiendo del modelo Original 1.

Partiendo del modelo original 1 se realiza un test de razón de verosimilitud, se presentan a continuación las pruebas de hipótesis realizadas:
En referencia las variables de confort se presenta la siguiente prueba de hipótesis:

H_0) $\beta_{Calefon} = \beta_{DVD} = \beta_{Aire} = \beta_{Secadora} = \beta_{Lavavajillas} = \beta_{Refrigerador} = 0$
H_1) Al menos un β distinto de cero

Al realizar el test entre el modelo completo (modelo original) y modelo reducido (modelo sin variables de H_0 se obtiene como resultado: p= 0,35751 que es mayor a 0,05, por lo que no se rechaza H_0 de que los coeficientes son iguales a 0. En otras palabras, es conveniente no considerar este grupo de variables en el modelo.

Una vez eliminadas estas variables se vuelve a repetir el test para distintos grupos de variables:

- Variables de TICs (Internet, Cable)
- LN YSVL sin joven prom, madre ausente, actividad del jefe
- Montevideo
- Hacinamiento

Se plantean las siguientes pruebas de hipótesis:

H_0) $\beta_{Internet} = \beta_{Cable} = \beta_{TV} = 0$
H_1) Al menos un β distinto de cero

H_0) $\beta_{Ingreso} = \beta_{MadreAusente} = \beta_{ActJefe} = 0$
H_1) Al menos un β distinto de cero

H_0) $\beta_{Mdeo} = 0$
H_1) $\beta_{Mdeo} \neq 0$

H_0) $\beta_{Hacinamiento} = 0$

H_1) $\beta_{Hacinamiento} \neq 0$

En todos los casos se obtuvieron p-valores mayores a 0,05 no rechazando de este modo H_0.

Figura 9.9: Salida de R mediante el test de razón de verosimilitud partiendo del modelo original 1.

```
MF3=svyglm(formula = Asiste ~ sexo + E27 + JovenActivo + hijos +
    jefe + aniosed + climaeducativo + icv2 + Computadora + Telefono +
    +Microondas, family = quasibinomial(link = "logit"), data = Per1417.con.muestra,
    design = diseño_personas_14_17, subset = (Stratum != 0))

Survey design:
subset(diseño_personas, (Personas.con.muestra$E27 >= 14 & Personas.con.muestra$E27 <=
    17))

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          9.48315    0.70243  13.500  < 2e-16 ***
sexohombre          -0.19784    0.09366  -2.112  0.03470 *
E27                 -0.98091    0.04944 -19.842  < 2e-16 ***
JovenActivo1        -1.43909    0.11339 -12.691  < 2e-16 ***
hijosAl menos un hijo -2.42633    0.35551  -6.825 9.64e-12 ***
jefejefe            -1.70031    0.71473  -2.379  0.01739 *
aniosed              0.81521    0.04206  19.380  < 2e-16 ***
climaeducativo       0.12873    0.02111   6.099 1.13e-09 ***
icv2deficit         -1.63558    0.53966  -3.031  0.00245 **
ComputadoraSi        0.38961    0.12232   3.185  0.00145 **
TelefonoSi           0.24531    0.10368   2.366  0.01801 *
MicroondasSi         0.23542    0.10296   2.286  0.02226 *
```

Fuente: ECH2011, elaboración propia

Figura 9.10: Salida de R mediante el test de razón de verosimilitud partiendo del modelo original 2.

```
> summary(MF4)

Call:
svyglm(formula = Asiste ~ sexo + E27 + JovenActivo + hijos +
    jefe + aniosed + climaeducativo + icv2 + TIC1, family = quasibinomial(link =
"logit"),
    data = Personas.con.muestra, design = diseño_personas_14_17,
    subset = (Stratum != 0))

Survey design:
subset(diseño_personas, (Personas.con.muestra$E27 >= 14 & Personas.con.muestra$E27 <=
    17))

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           10.18249    0.68231  14.924  < 2e-16 ***
sexohombre            -0.19427    0.09364  -2.075  0.03806 *
E27                   -0.99045    0.04893 -20.241  < 2e-16 ***
JovenActivo1          -1.44801    0.11284 -12.832  < 2e-16 ***
hijosAl menos un hijo -2.40848    0.34983  -6.885 6.36e-12 ***
jefejefe              -1.70565    0.72613  -2.349  0.01886 *
aniosed                0.81377    0.04235  19.216  < 2e-16 ***
climaeducativo         0.13559    0.02114   6.415 1.51e-10 ***
icv2deficit           -1.66566    0.54295  -3.068  0.00217 **
TIC1                  -0.30710    0.07684  -3.997 6.50e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasibinomial family taken to be 1.001339)

Number of Fisher Scoring iterations: 6
```

9.9. Modelos svyglm vs. glm.

En la siguiente tabla, se aprecia una comparación a través de los tres modelos trabajados de la significación de las variables bajo la metodología svyglm y glm.

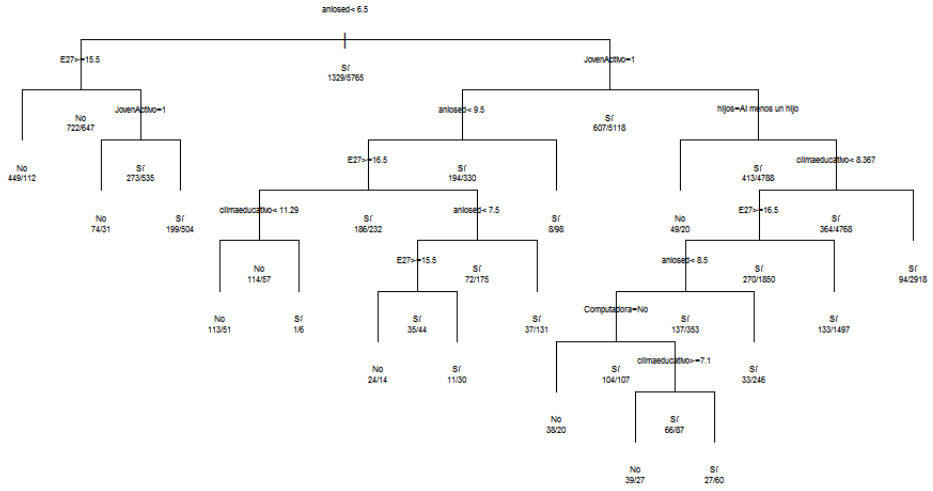
Variables	MF1 (svyglm)	MF1 (glm)	MF2 (svyglm)	MF2 (glm)	MF3 (svyglm)	MF3 (glm)
Constante	$2e^{-16}$	$2e^{-16}$	$2e^{-16}$	$2e^{-16}$	$2e^{-16}$	$2e^{-16}$
Sexo-Hombre	0,030462	0,00327	0, 037010	0,00288	0,03470	0,002920
E27	$2e^{-16}$	$2e^{-16}$	$2e^{-16}$	$2e^{-16}$	$2e^{-16}$	$2e^{-16}$
JovenActivo1	$2e^{-16}$	$2e^{-16}$	$2e^{-16}$	$2e^{-16}$	$2e^{-16}$	$2e^{-16}$
Hijos-Al menos un hijo	$3,06e^{-11}$	$3,66e^{-16}$	$1,28e^{-11}$	$2e^{-16}$	$9,64e^{-12}$	$2e^{-16}$
Jefe-jefe	0,019888	0,16807	0,019135	0,14622	0,01739	0,18809
Aniosed	$2e^{-16}$	$2e^{-16}$	$2e^{-16}$	$2e^{-16}$	$2e^{-16}$	$2e^{-16}$
Climaeducativo	$2,44e^{-9}$	$2,72e^{-13}$	$1,59e^{-9}$	$1,90e^{-12}$	$1,13e^{-9}$	$7,26e^{-13}$
ICV-déficit	0,003390	0,02241	0,003099	0,00141	0,00245	0,001072
Mdeo-Montevideo	0,029700	0,02241	0,018338	0,01640		
ComputadoraSi	0,000384	$5,55e^{-6}$			0,00145	$2,74e^{-5}$
MicroondasSi	0,024363	0,03662			0,02226	0,087078
Hacinamiento1	0,038394	0,13073				
TelefonoSi					0,01801	0,000673
TIC			0,000153	$7,34e^{-7}$		

Los modelos finales obtenidos son:

```
MF1=svyglm(formula = Asiste ~ sexo + E27 + JovenActivo + hijos + jefe + aniosed + climaeducativo +
    + icv2 + Computadora + Microondas + Hacinamiento, family = quasibinomial(link = "logit"),
    data = Personas.con.muestra, design = diseño_personas_14_17,subset = (Stratum != 0))

MF2=svyglm(formula = Asiste ~ sexo + E27 + JovenActivo + hijos + jefe + aniosed + climaeducativo +
    Mdeo + icv2 + TIC1, family = quasibinomial(link = "logit"), data = Personas.con.muestra,
    design = diseño_personas_14_17, subset = (Stratum != 0))

MF3=svyglm(formula = Asiste ~ sexo + E27 + JovenActivo + hijos + jefe + aniosed + climaeducativo +
    icv2 + Computadora + Telefono + Hacinamiento + Microondas, family = quasibinomial(link = "logit"),
    data = Personas.con.muestra, design = diseño_personas_14_17, subset = (Stratum != 0))

MF4=svyglm(formula = Asiste ~  sexo + E27 + JovenActivo + hijos +  jefe + aniosed + climaeducativo +
    icv2 + TIC1 , family = quasibinomial(link = "logit"), data = Personas.con.muestra,
    design = diseño_personas_14_17, subset = (Stratum != 0))
```

9.10. Árboles de clasificación para la variable Asistencia.

Figura 9.11: Árbol de clasificación con $cp = 0,003$ bajo escenario 1.



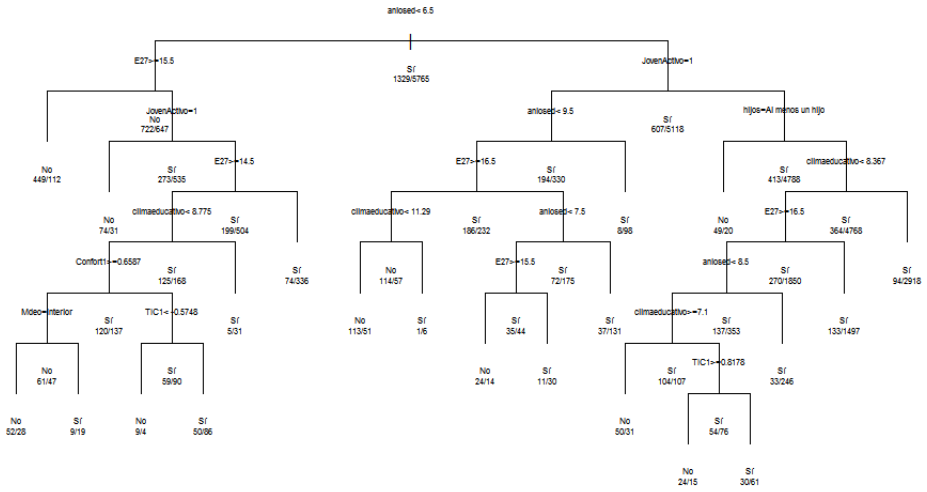
Fuente ECH2011, elaboración propia.

Figura 9.12: Árbol de clasificación óptimo.

```
n= 7094

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 7094 1329 Sí (0.18734142 0.81265858)
   2) aniosed< 6.5 1369  647 No (0.52739226 0.47260774)
     4) E27>=15.5 561  112 No (0.80035651 0.19964349) *
     5) E27< 15.5 808  273 Sí (0.33787129 0.66212871)
      10) JovenActivo=1 105   31 No (0.70476190 0.29523810) *
      11) JovenActivo=0 703  199 Sí (0.28307255 0.71692745) *
   3) aniosed>=6.5 5725  607 Sí (0.10602620 0.89397380)
     6) JovenActivo=1 524  194 Sí (0.37022901 0.62977099)
      12) aniosed< 9.5 418  186 Sí (0.44497608 0.55502392)
        24) E27>=16.5 171   57 No (0.66666667 0.33333333) *
        25) E27< 16.5 247   72 Sí (0.29149798 0.70850202) *
      13) aniosed>=9.5 106    8 Sí (0.07547170 0.92452830) *
     7) JovenActivo=0 5201  413 Sí (0.07940781 0.92059219)
      14) hijos=Al menos un hijo 69   20 No (0.71014493 0.28985507) *
      15) hijos=Sin Hijos 5132  364 Sí (0.07092751 0.92907249) *
```

Fuente ECH2011, elaboración propia.

Figura 9.13: Árbol de clasificación con $cp = 0,003$ bajo escenario 2.



Fuente ECH2011, elaboración propia.

9.11. Sintaxis R.

```
#### Sintaxis ##################################################################################################

Hogares=h2011
rm(h2011)
Personas=p2011
rm(p2011)

#######################
#### Asistencia ################################################################################################
#######################

Personas$Asiste= ifelse(Personas$E49==2 | ((Personas$E193==2 | Personas$E193==3 )
                                           & (Personas$E197==2 | Personas$E197==3)
                                           & (Personas$E201==2 | Personas$E201==3)
                                           & (Personas$E212==2 | Personas$E212==3)
                                           & (Personas$E215==2 | Personas$E215==3)
                                           & (Personas$E218==2 | Personas$E218==3)
                                           & (Personas$E221==2 | Personas$E221==3)), c(0) , c(1))

Personas$Asiste2=ifelse(Personas$E193==1 | Personas$E197==1|Personas$E201==1|Personas$E212==1|
                        Personas$E215==1|Personas$E218==1|Personas$E221==1, c(1), c(0))
Personas$Asiste=as.numeric(Personas$Asiste)

############################################
# Mujer, Afro, Hijos, Madre ausente y  Jefe ###################################################################
############################################

## Sexo
Personas$E26=as.numeric(Personas$E26)
Personas$sexo <- ifelse (Personas$E26==1,c(1),c(0))
Personas$sexo <- as.factor(Personas$sexo)
levels(Personas$sexo)= c("mujer","hombre")

## Raza ## no fue utilizada en el modelo
## Personas$raza <-ifelse(Personas$e29_6=="Afro o negra",c("Afro o negra"),
##ifelse((Personas$e29_6=="Asiática o amarilla"
## Personas$e29_6=="Indígena" |  Personas$E29_6=="Otra"), c("Otra"),
##ifelse(Personas$e29_6=="Blanca", c("Blanca"),c("99"))))

##Madre ausente
Personas$e31=as.numeric(Personas$E31)
Personas$Madre_ausente<- ifelse(Personas$E31== "99", c("1"), c("0"))
Personas$Madre_ausente<- as.factor(Personas$Madre_ausente)
levels(Personas$Madre_ausente)<- c("Presente", "Ausente")

##Hijos
Personas$E186_1=as.numeric(Personas$E186_1)
Personas$E31=as.numeric(Personas$E31)
Personas$hijos=ifelse(Personas$E186_1>0,c(1),c(0))
Personas$hijos =as.factor(Personas$hijos)
levels(Personas$hijos)= c("Sin Hijos","Al menos un hijo")

##Jefe
Personas$E30=as.numeric(Personas$E30)
Personas$jefe=ifelse(Personas$E30==1,c(1),c(0))
Personas$jefe <- as.factor(Personas$jefe)
levels(Personas$jefe)= c("otro","jefe")

############################################
## Condición de Actividad del Jefe del Hogar ###################################################################
############################################

Jefes <-subset(Personas ,E30==1,select=c("NUMERO","POBPCOAC"))
Hogares<- merge(Hogares,Jefes, by.x="numero", by.y="NUMERO", all.x=T)
Hogares$POBPCOAC=as.numeric(Hogares$POBPCOAC)
Hogares$Actividad_del_Jefe = ifelse(Hogares$POBPCOAC ==2, c(1),c(0))
Hogares$Actividad_del_Jefe =as.factor(Hogares$Actividad_del_Jefe)
levels(Hogares$Actividad_del_Jefe)= c("Jefe Inactivo o Desocupado","Jefe Activo")

#######################
## Actividad del Joven#########################################################################################
#######################

Personas$JovenActivo<-Personas$F66
Personas$JovenActivo<-ifelse(Personas$F66==1, c(1), c(0))
```

```
############################
## Clima Educativo del Hogar ##########################################################################
############################

## Años de educación
Personas$E51_2_a=ifelse(Personas$E51_2==9,0,Personas$E51_2)
Personas$E51_3_a=ifelse(Personas$E51_3==9,0,Personas$E51_3)
Personas$E51_4_a=ifelse(Personas$E51_4==9,0,Personas$E51_4)
Personas$E51_5_a=ifelse(Personas$E51_5==9,0,Personas$E51_5)
Personas$E51_6_a=ifelse(Personas$E51_6==9,0,Personas$E51_6)
Personas$E51_7_a=ifelse(Personas$E51_7==9,0,Personas$E51_7)
Personas$E51_8_a=ifelse(Personas$E51_8==9,0,Personas$E51_8)
Personas$E51_9_a=ifelse(Personas$E51_9==9,0,Personas$E51_9)
Personas$E51_10_a=ifelse(Personas$E51_10==9,0,Personas$E51_10)
Personas$E51_11_a=ifelse(Personas$E51_11==9,0,Personas$E51_11)

#creo una variable que indique años educ para cada valor de enseñanza técnica
Personas$E51_71=ifelse(Personas$E51_7_1==1,Personas$E51_7_a,0)
Personas$E51_72=ifelse(Personas$E51_7_1==2,Personas$E51_7_a,0)
Personas$E51_73=ifelse(Personas$E51_7_1==3,Personas$E51_7_a,0)
Personas$E51_74=ifelse(Personas$E51_7_1==4,Personas$E51_7_a,0)
#Personas$E51_72=ifelse(Personas$E51_72>3,3,Personas$E51_72)

attach(Personas)
Personas$aniosed = ifelse(E49==2,0,
                          ifelse(E51_11_a %in% 1:6, pmax(12+E51_9_a+E51_11_a, 12+E51_8_a+E51_11_a,
                                 12+E51_10_a+E51_11_a),
                                 ifelse(E51_9_a %in% 1:8 | E51_10_a %in% 1:8 | E51_8_a %in% 1:8 |
                                        E51_7_1==1 | (E51_7_1==2 & E51_72>3),
                                        pmax(12+E51_9_a, 12+E51_10_a, 12+E51_8_a, 12+E51_71, 9+E51_72),
                                        ifelse(E51_7_1==2 | E51_6_a %in% 1:3 | E51_5_a %in% 1:3,
                                               pmax(9+E51_72, 9+E51_6_a, 9+E51_5_a, 6+E51_73),
                                               ifelse(E51_4_a %in% 1:3 | E51_7_1==3, pmax(6+E51_4_a,
                                                      6+E51_73),
                                                      ifelse(E51_2_a %in% 1:6  | E51_7_1==4 |
                                                             E51_3_a %in% 1:4 , pmax(E51_2_a, E51_74,
                                                             E51_3_a),
                                                               ifelse(E193 %in% 1:2, 0, 0
                                                              )))))))
detach(Personas)

#trunco años ed en 22
Personas$aniosed = ifelse(Personas$aniosed<12 & (Personas$E51_9==9 | Personas$E51_8==9 |
                   Personas$E51_10==9 |(Personas$E51_7==9 & Personas$E51_7_1==1)), 12, Personas$aniosed)
#trunco años ed en 22
Personas$aniosed = ifelse(Personas$aniosed>22, 22,Personas$aniosed)

## Fin de años de educación

##Hogares con mayores
mayores = subset(Personas, E27>17)
aniosedh_media_may = tapply(mayores$aniosed,mayores$NUMERO,mean)
Hogares = merge(Hogares, aniosedh_media_may, by.x="numero", by.y="row.names", all.x=T)
names(Hogares) = c(names(Hogares)[1:length(names(Hogares))-1], "aniosedh_media_may")
##Hogares sin mayores
#Si no hay individuos mayores de 18 en el hogar, se imputan los años aprobados por el jefe del hogar
mayor_edadh = tapply(Personas$E27,Personas$NUMERO,max)
table(mayor_edadh)
Hogares = merge(Hogares, mayor_edadh, by.x="numero", by.y="row.names", all.x=T)
names(Hogares) = c(names(Hogares)[1:length(names(Hogares))-1], "mayor_edadh")
Hogares$hog_sin_mayores = ifelse(Hogares$mayor_edadh<18, 1, 0)
table(Hogares$hog_sin_mayores)

Personas$aniosed_jefe = ifelse(Personas$E30==1,Personas$aniosed,0)
table(Personas$aniosed_jefe)
aniosed_jefe = tapply(Personas$aniosed_jefe, Personas$NUMERO, max)
Hogares = merge(Hogares, aniosed_jefe, by.x="numero", by.y="row.names", all.x=T)
names(Hogares) = c(names(Hogares)[1:length(names(Hogares))-1], "aniosed_jefe")

Hogares$climaeducativo = ifelse(Hogares$hog_sin_mayores==1,Hogares$aniosed_jefe,Hogares$aniosedh_media_may)

### FIN CLIMA EDUCATIVO ##############################################################################
```

```
Personas$Asiste=as.factor(Personas$Asiste)
Personas$afro=as.factor(Personas$afro)
Personas$sexo=as.factor(Personas$sexo)
Personas$hijos=as.factor(Personas$hijos)
Personas$JovenActivo=as.factor(Personas$JovenActivo)
Hogares$climaeducativo=as.numeric(Hogares$climaeducativo)

###########
## Región ##################################################################################################
###########

Hogares$Mdeo=as.factor(ifelse(Hogares$region==1,c(1),c(0)))
#Hogares$Interior= ifelse(Hogares$región=="Interior localidades < 5000" |
#Hogares$región=="Interior localidades 5000 o más",c("1"),c("0"))
#Hogares$Rural= ifelse(Hogares$región==4,c("1"),c("0"))
Hogares$Mdeo= as.factor(Hogares$Mdeo)
#Hogares$Interior= as.factor(Hogares$Mdeo)
levels(Hogares$Mdeo) = c("Interior", "Montevideo")
#levels(Hogares$Interior) = c("otro", "Interior", )
#levels(Hogares$Rural) = c("otro", "Rural", )

############
## Ingreso ###################################################################################################
############

# YSVL= Ingreso del hogar sin valor locativo
# HT19 = Cantidad de personas sin servicio doméstico
# YSVL/HT19 = Ingreso per cápita del hogar

Hogares$YSVL=as.integer(Hogares$ysvl)
Hogares$HT19=as.integer(Hogares$ht19)
Hogares$YSVL_pc=(Hogares$YSVL/Hogares$ht19)
Hogares$numero=as.character(Hogares$numero)
Personas$NUMERO=as.character(Personas$NUMERO)

Ingreso <-subset(Hogares ,select=c("numero","YSVL","YSVL_pc"))
Ingreso$numero=as.character(Ingreso$numero)

Personas<- merge(Personas,Ingreso, by.x="NUMERO", by.y="numero", all.x=T)

rm(aniosed_jefe)
rm(aniosedh_media_may)
rm(mayor_edadh)
rm(mayores)
rm(Jefes)
rm(Ingreso)

Personas$YSVL_sin_joven= ifelse(Personas$E27>= 14 & Personas$E27 <= 17,c(Personas$YSVL-Personas$PT4),
                        c(Personas$YSVL))

YSVL_sin_joven=as.data.frame(subset(Personas, select=c(NUMERO,YSVL_sin_joven)))
YSVL_sin_joven_prom=aggregate(YSVL_sin_joven$YSVL_sin_joven, list(YSVL_sin_joven$NUMERO), FUN=mean)
colnames(YSVL_sin_joven_prom)=c("NUMERO","YSVL_sin_joven_prom")

Hogares=merge(Hogares,YSVL_sin_joven_prom, by.x="numero", by.y="NUMERO")

###################################
## Logartimo Neperiano del ingreso ##########################################################################
###################################

Hogares$LN_YSVL_sin_joven_prom=log(Hogares$YSVL_sin_joven_prom)

LN_YSVL_sin_joven_prom=subset(Hogares,select=c("numero","LN_YSVL_sin_joven_prom"))

Personas=merge(Personas,LN_YSVL_sin_joven_prom, by.x="NUMERO", by.y="numero")
rm(LN_YSVL_sin_joven_prom)
rm(YSVL_sin_joven)
rm(YSVL_sin_joven_prom)
rm(mayores)
```

```
#################################
#ÍNDICE DE CALIDAD DE LA VIVIENDA#################################################################################
#################################

#############
#Materiales#
############

Hogares$c3=as.numeric(Hogares$c3)

attach(Hogares)
buena=ifelse((c3==1 & (c4==1 | c4==2) & c2==1),1,0)
detach(Hogares)

attach(Hogares)
aceptable=ifelse((((c3==2 | c3==3 |c3==4) & (c4==1 | c4==2) & (c2<6)) |
                    ((c3==1) & (c4==1 | c4==2) & (c2>1 & c2<6)) | ((c3==1| c3==2 | c3==3 | c3==4 ) &
                    (c4==3) & (c2<6))),1,0)
detach(Hogares)

attach(Hogares)
regular=ifelse((((c3<6)  & (c4==4) & (c2<6)) | ((c3==5) & (c4==1 | c4==2 |  c4==3) & (c2<6))),1,0)
detach(Hogares)

attach(Hogares)
deficitaria=ifelse((c3==6 | c4==5 | c2==6),1,0)
detach(Hogares)

Hogares=cbind(Hogares,buena,aceptable,regular,deficitaria)

Hogares$materiales=ifelse(Hogares$buena==1,1,
                          ifelse(Hogares$aceptable==1,2,
                                 ifelse(Hogares$regular==1,3,
                                        ifelse(Hogares$deficitaria==1,4,"xxx"))))
########################
#ACCESO A AGUA POTABLE#
######################

attach(Hogares)
buenacceso=ifelse(((d11==1 | d11==2) & (d12==1)),1,0)
detach(Hogares)

attach(Hogares)
aguaregular=ifelse(((d11<6 ) & (d12==2|d12==3)),1,0)
detach(Hogares)

attach(Hogares)
aguamalo=ifelse(((d11==6 )  | (d12==4)),1,0)
detach(Hogares)

Hogares=cbind(Hogares,buenacceso,aguaregular,aguamalo)

Hogares$accesoagua=ifelse(Hogares$buenacceso==1,1,
                          ifelse(Hogares$aguaregular==1,2,
                                 ifelse(Hogares$aguamalo==1,3,"xxx")))

######################
#ACCESO A SANEAMIENTO#
####################

attach(Hogares)
sanadecuado=ifelse(((d13==1)& (d16==1 | d16==2)),1,0)
detach(Hogares)

attach(Hogares)
sanregular=ifelse(((d13==1)& (d16==1 | d16==2)),1,0)
detach(Hogares)
attach(Hogares)
sanmalo=ifelse(((d13==2 | d13==3) | (d16==4)),1,0)
detach(Hogares)

Hogares=cbind(Hogares,sanadecuado,sanregular,sanmalo)

Hogares$saneamiento=ifelse(Hogares$sanadecuado==1,1,
                           ifelse(Hogares$sanregular==1,2,
                                  ifelse(Hogares$sanmalo==1,3,"xxx")))
```

```
##################
#ACCESO A ENERGÍA#
#################

attach(Hogares)
energia=ifelse((d18==1 | d18==2),1,0)
detach(Hogares)

Hogares=cbind(Hogares,energia)

#################################
#ÍNDICE DE CALIDAD DE LA VIVIENDA#
#################################

attach(Hogares)
buenacalidad=ifelse((materiales==1 & accesoagua==1 & saneamiento==1 & energia==1),1,0)
detach(Hogares)
attach(Hogares)
deficit=ifelse(((materiales==4 | accesoagua==3 | saneamiento==3) & energia==0),1,0)
detach(Hogares)

Hogares=cbind(Hogares,buenacalidad,deficit)

Hogares$icv=ifelse(Hogares$buenacalidad==1,1,
                   ifelse(Hogares$deficit==1,3,2))

Hogares$icv=as.factor(Hogares$icv)
levels(Hogares$icv)=c("buenacalidad","regular","deficit")
rm(buenacceso)
rm(deficit)
rm(deficitaria)
rm(energia)
rm(regular)
rm(sanadecuado)
rm(sanmalo)
rm(sanregular)
rm(aceptable)
rm(aguamalo)
rm(aguaregular)
rm(buena)
rm(buenacalidad)

Hogares$icv2=ifelse(Hogares$icv=="deficit",c(2), c(1))
Hogares$icv2=as.factor(Hogares$icv2)
levels(Hogares$icv2)=c("buena o regular","deficit")

##################
#### Hacinamiento ##############################################################################################

##Hacinamiento = Total de personas en el hogar/ habitaciones para dormir
Hogares$Hacinamiento = ifelse (((Hogares$d25/Hogares$d10)>2), c(1), c(0))

########################
#### Variables para ACM ##########################################################################################

Hogares$Calefon = ifelse(Hogares$d21_1==1, c(1), c(0))
Hogares$Calefon = as.factor(Hogares$Calefon)
levels(Hogares$Calefon) = c("No", "Si")

Hogares$Refrigerador = ifelse(Hogares$d21_3==1, c(1), c(0))
Hogares$Refrigerador = as.factor(Hogares$Refrigerador)
levels(Hogares$Refrigerador) = c("No", "Si")

Hogares$TV = ifelse(Hogares$d21_4==1 | Hogares$d21_5==1, c(1), c(0))
Hogares$TV = as.factor(Hogares$TV)
levels(Hogares$TV) = c("No", "Si")

Hogares$Cable = ifelse(Hogares$d21_7==1 , c(1), c(0))
Hogares$Cable = as.factor(Hogares$Cable)
levels(Hogares$Cable) = c("No", "Si")

Hogares$DVD = ifelse(Hogares$d21_9==1 , c(1), c(0))
Hogares$DVD = as.factor(Hogares$DVD)
levels(Hogares$DVD) = c("No", "Si")

Hogares$Lavarropa = ifelse(Hogares$d21_10==1 , c(1), c(0))
Hogares$Lavarropa= as.factor(Hogares$Lavarropa)
levels(Hogares$Lavarropa) = c("No", "Si")
```

```r
Hogares$Secadora = ifelse(Hogares$d21_11==1 , c(1), c(0))
Hogares$Secadora= as.factor(Hogares$Secadora)
levels(Hogares$Secadora) = c("No", "Si")

Hogares$Lavavajillas = ifelse(Hogares$d21_12==1 , c(1), c(0))
Hogares$Lavavajillas= as.factor(Hogares$Lavavajillas)
levels(Hogares$Lavavajillas) = c("No", "Si")

Hogares$Microondas = ifelse(Hogares$d21_13==1 , c(1), c(0))
Hogares$Microondas= as.factor(Hogares$Microondas)
levels(Hogares$Microondas) = c("No", "Si")

Hogares$Aire = ifelse(Hogares$d21_14==1 , c(1), c(0))
Hogares$Aire= as.factor(Hogares$Aire)
levels(Hogares$Aire) = c("No", "Si")

Hogares$Computadora= ifelse(Hogares$d21_15==1 , c(1), c(0))
Hogares$Computadora= as.factor(Hogares$Computadora)
levels(Hogares$Computadora) = c("No", "Si")

Hogares$Internet= ifelse(Hogares$d21_16==1 , c(1), c(0))
Hogares$Internet=as.factor(Hogares$Internet)
levels(Hogares$Internet) = c("No", "Si")

Hogares$Telefono= ifelse(Hogares$d21_17==1 , c(1), c(0))
Hogares$Telefono=as.factor(Hogares$Telefono)
levels(Hogares$Telefono) = c("No", "Si")
Hogares$Auto_o_moto = ifelse(Hogares$d21_18==1 | Hogares$d21_19==1, c(1), c(0))
Hogares$Auto_o_moto = as.factor(Hogares$Auto_o_moto)
levels(Hogares$Auto_o_moto) = c("No", "Si")
Hogares$Auto = ifelse(Hogares$d21_18==1,c(1), c(0))
Hogares$Auto = as.factor(Hogares$Auto)
levels(Hogares$Auto) = c("No", "Si")

Hogares$Moto = ifelse(Hogares$d21_19==1,c(1), c(0))
Hogares$Moto = as.factor(Hogares$Moto)
levels(Hogares$Moto) = c("No", "Si")

Hogares$Agua_en_red_gral= ifelse(Hogares$d11==1 , c(1), c(0))
Hogares$Agua_en_red_gral = as.factor(Hogares$Agua_en_red_gral)
levels(Hogares7$Agua_en_red_gral) = c("No", "Si")

Hogares$Baño_con_cisterna= ifelse(Hogares$d13==1 , c(1), c(0))
Hogares$Baño_con_cisterna = as.factor(Hogares$Baño_con_cisterna)
levels(Hogares$Baño_con_cisterna) = c("No", "Si")

Hogares$Evac_san_red_gral= ifelse(Hogares$d16==1 , c(1), c(0))
Hogares$Evac_san_red_gral= as.factor(Hogares$Evac_san_red_gral)
levels(Hogares$Evac_san_red_gral) = c("No", "Si")

Hogares$Fuente_iluminar= ifelse(Hogares$d18==1 , c(1), c(0))
Hogares$Fuente_iluminar= as.factor(Hogares$Fuente_iluminar)
levels(Hogares$Fuente_iluminar) = c("No", "Si")

Hogares$BuenPared= ifelse(Hogares$c2==1| Hogares$c2==2 | Hogares$c2==3, c(1), c(0))
Hogares$BuenPared= as.factor(Hogares$BuenPared)
levels(HogaresBuenPared) = c("No", "Si")

Hogares$BuenTecho= ifelse(Hogares$c3==1| Hogares$c3==2 | Hogares$c3==3, c(1), c(0))
Hogares$BuenTecho = as.factor(Hogares$BuenTecho)
levels(Hogares.con.14.17$BuenTecho) = c("No", "Si")
Hogares$BuenPiso= ifelse(Hogares$c4==1| Hogares$c4==2, c(1), c(0))
Hogares$BuenPiso = as.factor(Hogares$BuenPiso)
levels(Hogares$BuenPiso) = c("No", "Si")

## Fin de variables para ACM ##################################################################################

############
## Joven ###################################################################################################
############

Personas$joven=ifelse((Personas$E27 >= 14 & Personas$E27 <= 17), 1, 0)
```

```
##############################################################################
## Clima Educativo, Actividad del Jefe, Mdeo, hacinamiento y ht19 en Personas ##########################
##############################################################################

var<-Hogares[,c("numero","climaeducativo","Actividad_del_Jefe","Hacinamiento","Mdeo","ht19","icv")]
Personas<-merge(Personas, var, by.x="NUMERO", by.y="numero")

rm(var)

############
## MUESTRA ##################################################################################################
############

## n total=8868 y n muestra= 7095
Personas$nper=seq(1,nrow(Personas))
Personas14_17=subset(Personas, (Personas$E27>=14 & Personas$E27<=17))
tabla.asistencia=prop.table(table(Personas14_17$Asiste))
nper=nrow(Personas14_17)
nmuestra=round(nper*80/100)
## 18,73 % No asiste y 81.27 % Asiste
asiste=as.numeric(tabla.asistencia[1])
noasiste=as.numeric(tabla.asistencia[2])

## No asiste 7095*18.73027/100= 1329
## Asiste    7095*81.26973/100= 5766
E.asiste=round(nmuestra*asiste)
E.noasiste=round(nmuestra*noasiste)

install.packages("sampling")
library( sampling )

## Muestreo aleatorio simple estratificado
estratos <- strata(Personas14_17, stratanames = c("Asiste"), size = c(E.noasiste,E.asiste),
                   method = "srswor")
Personas14_17.muestreado <- getdata(Personas14_17,estratos)

##Personas muestreadas
per14_17.muestreado2=subset(Personas14_17.muestreado, select=c("NUMERO","nper","Stratum"))

save(per14_17.muestreado2,file="per14_17.muestreado2.Rdata")
save(Personas14_17.muestreado,file="Personas14_17.muestreado.Rdata")
save(Personas14_17,file="Personas14_17.Rdata")
save(Personas,file="Personas")
save(Hogares,file="Hogares")

rm(E.asiste)
rm(E.noasiste)
rm(asiste)
rm(nmuestra)
rm(noasiste)
rm(nper)
rm(tabla.asistencia)
rm(estratos)
Personas.con.muestra = merge(Personas,per14_17.muestreado2, by.x=c("NUMERO","nper"),
                             by.y=c("NUMERO","nper"), all.x=TRUE)
Personas.con.muestra$Stratum=as.integer(Personas.con.muestra$Stratum)
Personas.con.muestra$Stratum[is.na(Personas.con.muestra$Stratum) ] <- 0
save(Personas.con.muestra,file="Personas.con.muestra.Rdata")

##Con esta base se va a realizar la predicción tomando el subset de 14 a 17.
Personas.no.muestra=subset(Personas.con.muestra, (Personas.con.muestra$Stratum==0))
save(Personas.no.muestra,file="Personas.no.muestra.Rdata")
Peronas14_17.no.muestra=subset(Personas14_17, (Personas14_17$Stratum==0))

#############
## INDICES ##################################################################################################
#############

library(foreign)
library("ade4")
source('acmPOND.R')

per1417 = tapply(Personas.con.muestra$joven,Personas.con.muestra$NUMERO,max)
per1417=as.data.frame(per1417)
names(per1417)="hog.con.joven"
```

```
Hogares.con.14.17 = merge(Hogares, per1417, by.x="numero", by.y="row.names", all.x=T)
Hogares.solo.14.17=subset(Hogares.con.14.17,Hogares.con.14.17$hog.con.joven==1)
Hogares.solo.14.17$seq=seq(1,nrow(Hogares.solo.14.17), by=1)

confort<- Hogares.solo.14.17[,c("Calefon","DVD","Lavarropa","Microondas","Aire", "Auto_o_moto",
                                "Secadora","Lavavajillas","Refrigerador")]
tic<- Hogares.solo.14.17[,c("Cable", "TV", "Internet", "Telefono","Computadora")]
ViviendaySalubridad =Hogares.solo.14.17[,c("Agua_en_red_gral","Baño_con_cisterna","Evac_san_red_gral",
                                           "BuenTecho","BuenPiso","BuenPared","Fuente_iluminar")]

##CONFORT
acmconfort<- acm(confort,NF=2,Csup=NULL,ByG=T,pesos=Hogares.solo.14.17$pesoano)
acmconfortsup<- acm(confort,NF=2,Csup=(7:9),ByG=T,pesos=Hogares.solo.14.17$pesoano)

##VIVIENDA
acmviv<- acm(ViviendaySalubridad,NF=2,Csup=NULL,ByG=T,pesos=Hogares.solo.14.17$pesoano)
acmvivsup<- acm(ViviendaySalubridad,NF=2,Csup=(6:7),ByG=T,pesos=Hogares.solo.14.17$pesoano)

##TIC
acmtic<- acm(tic,NF=2,Csup=NULL,ByG=T,pesos=Hogares.solo.14.17$pesoano)
acmticsup<- acm(tic,NF=2,Csup=(2),ByG=T,pesos=Hogares.solo.14.17$pesoano)

IConfort <- read.table("ACM_Individuos_Confort.txt", header=T, sep="\t", dec=",")
Confort<-c('Confort1_1417', 'Confort2_1417')
names(IConfort)<-Confort
table(is.na(IConfort$Confort1_1417))

ITIC <- read.table("ACM_Individuos_TIC.txt", header=T, sep="\t", dec=",")
TIC<-c('TIC1_1417','TIC2_1417')
names(ITIC)<-TIC
table(is.na(ITIC$TIC1_1417))
ITIC <- read.table("ACM_Individuos_TIC.txt", header=T, sep="\t", dec=",")
TIC<-c('TIC1_1417','TIC2_1417')
names(ITIC)<-TIC
table(is.na(ITIC$TIC1_1417))

IVIV <- read.table("C:/Documents and Settings/usuario/Escritorio/Tesis a entregar/Resultados
                   /ACM_Individuos_VIV.txt", header=T, sep="\t", dec=",")
VIV<-c('VIV1_1417', 'VIV2_1417')
names(IVIV)<-VIV
table(is.na(IVIV$VIV1_1417))

Hogares.solo.14.17 = merge(Hogares.solo.14.17, IConfort, by.x="seq", by.y="row.names", all.x=T)
Hogares.solo.14.17 = merge(Hogares.solo.14.17, ITIC, by.x="seq",  by="row.names", all.x=T)
Hogares.solo.14.17 = merge(Hogares.solo.14.17, IVIV, by.x="seq", by.y="row.names", all.x=T)
Indices= subset(Hogares.solo.14.17, select=c("numero","Confort1_1417","TIC1_1417","VIV1_1417"))
Indices= subset(Hogares.solo.14.17, select=c("numero","Confort1_1417","TIC1_1417"))

rm(ITIC)
rm(TIC)
rm(IConfort)
rm(Confort)
rm(TIC)
rm(IVIV)
rm(VIV)
rm(per14_17.muestreado2)

#Hogares
save(Hogares,file="Hogares.Rdata")
save(Hogares.con.14.17,file="Hogares.con.14.17.Rdata")
save(Hogares.solo.14.17,file="Hogares.solo.14.17.Rdata")

save(Indices,file="Indices.Rdata")

#Personas
save(Personas,file="Personas.Rdata")
save(Personas.con.muestra,file="Personas.con.muestra.Rdata")
save(Personas.no.muestra,file="Personas.no.muestra.Rdata")
save(Personas14_17,file="Personas14_17.Rdata")
save(Personas14_17.muestreado,file="Personas14_17.muestreado.Rdata")
save(Indices,file="Indices.Rdata")
save(Personas14_17.no.muestra,file="Personas14_17.no.muestra.Rdata")

rm(per1417)
rm(confort)
rm(tic)
```

```
##Pegamos Indices en Personas##

Personas.con.muestra<-merge(Personas.con.muestra,Indices, by.x="NUMERO", by.y="numero", all.x=T)
Personas14_17.muestreado=merge(Personas14_17.muestreado,Indices, by.x="NUMERO", by.y="numero", all.x=T)
Personas14_17=merge(Personas14_17,Indices, by.x="NUMERO", by.y="numero", all.x=T)
Personas.no.muestra=merge(Personas.no.muestra,Indices, by.x="NUMERO", by.y="numero", all.x=T)
Peronas14_17.no.muestra<- merge(Personas14_17.no.muestra,Indices, by.x="NUMERO", by.y="numero", all.x=T)


save(file="Personas.con.muestra.Rdata")
save(file="Personas.muestra.Rdata")
save(file="Personas14_17.Rdata")
save(file="Personas14_17.muestreado.Rdata")
save(file="Personas14_17.no.muestra.Rdata")

## Diseño ##########################################################################################

library("survey")

diseño_personas = svydesign(id=~NUMERO, strata=~ESTRATOGEO, data=Personas.con.muestra,
                            weights=Personas.con.muestra$PESOANO)
diseño_hogares = svydesign(id=~numero, strata=~estratogeo, data=Hogares.con.14.17,
                           weights=Hogares.con.14.17$pesoano)

diseño_personas_14_17 = subset(diseño_personas,(Personas.con.muestra$E27>=14 &
                                                  Personas.con.muestra$E27<=17))
diseño_hogares_14_17=subset(diseño_hogares, Hogares.con.14.17$hog.con.joven==1)
## MOdelos Finales ##################################################################################


MF1=svyglm(formula = Asiste ~ sexo + E27 + JovenActivo + hijos + jefe + aniosed + climaeducativo +
    + icv2 + Computadora + Microondas + Hacinamiento, family = quasibinomial(link = "logit"),
    data = Personas.con.muestra, design = diseño_personas_14_17,subset = (Stratum != 0))

MF2=svyglm(formula = Asiste ~ sexo + E27 + JovenActivo + hijos + jefe + aniosed + climaeducativo +
    Mdeo + icv2 + TIC1, family = quasibinomial(link = "logit"), data = Personas.con.muestra,
    design = diseño_personas_14_17, subset = (Stratum != 0))

MF3=svyglm(formula = Asiste ~ sexo + E27 + JovenActivo + hijos + jefe + aniosed + climaeducativo +
    icv2 + Computadora + Telefono + Hacinamiento + Microondas, family = quasibinomial(link = "logit"),
    data = Personas.con.muestra, design = diseño_personas_14_17, subset = (Stratum != 0))

MF4=svyglm(formula = Asiste ~  sexo + E27 + JovenActivo + hijos +  jefe + aniosed + climaeducativo +
    icv2 + TIC1 , family = quasibinomial(link = "logit"), data = Personas.con.muestra,
    design = diseño_personas_14_17, subset = (Stratum != 0))


## Predicción ######################################################################################

#Sintaxis para hacer predicción#
aa<-predict(MF3,newdata= Per.1417.no,muestra, type = c("response"))
aa1<- ifelse(aa<=0.79, c("no asiste"), c("asiste"))
aa1<- as.factor(aa1)
bb<-table(Per.1417.no.muestra$Asiste, aa1)

#Hacemos el cuadro observado vs. predicho#
MF3Pred<-prop.table(bb, m=1)
MF3Pred
aa2=as.numeric(aa1)

##Curva Roc
plot(roc(Per.1417.no.muestra$Asiste,aa2))
```

```
#############################################################################
#ARBOL DE CLASIFICACIÃN COHERENTE CON EL ESCENARIO 1#########################
#############################################################################
per$Asiste<-factor(per$Asiste, levels=0:1, labels = c("No", "Si"))
#Árbol por defecto con cp=0.01#
arbolmf1<-rpart(Asiste ~ sexo + E27 + JovenActivo + Madre_ausente + hijos + jefe +
                  aniosed + Madre_ausente + climaeducativo + icv2 + Computadora +
 Telefono  + Microondas + Auto_o_moto + Cable + Internet + TV + Lavavajillas +
Secadora + Refrigerador + Calefon + DVD + Aire + Mdeo + LN_YSVL_sin_joven_prom +
Hacinamiento,  data = per_training, method = 'class')
summary(arbolmf1)
print(arbolmf1)
plot(arbolmf1, uniform = TRUE, compress=T, margin=0.1)
text(arbolmf1, use.n = TRUE, all=T, pretty=0, cex = 0.5)
post(arbolmf1, title. = "Arbol de clasificaciÃ³n variable Asistencia", pretty = 0)
#Arbol maximal##############################################################################
arbolmax1<-rpart(Asiste ~ sexo + E27 + JovenActivo + Madre_ausente + hijos + jefe +
                   aniosed + climaeducativo + icv2 + Computadora + Telefono  +
                   Microondas + Auto_o_moto + Cable + Internet + TV + Lavavajillas +
Secadora + Refrigerador + Calefon + DVD + Aire + Mdeo + LN_YSVL_sin_joven_prom +
Hacinamiento,  data = per_training, method = 'class', control=rpart.control(cp=0))
print(arbolmax1)
printcp(arbolmax1)
summary(arbolmax1)
plot(arbolmax1, uniform = TRUE, compress=T, margin=0.1)
text(arbolmax1, use.n = TRUE, all=T, pretty=0, cex = 0.5)
pp<-predict(arbolmax1, newdata=per_test, type="class")
aa1<-table(per_test$Asiste, predict(arbolmax1, newdata=per_test, type="class"))
#Construcción de secuencia de Ã¡rboles anidados#####################################
fit1<-prune(arbolmax1, cp=0.001)
plot(fit1, uniform = TRUE, compress=T, margin=0.1)
text(fit1, use.n = TRUE, all=T, pretty=0, cex = 0.5)
pp<-predict(fit1, newdata=per_test, type="class")
aa1<-table(per_test$Asiste, predict(fit1, newdata=per_test, type="class"))
prop.table(aa1, m=1)
#############################################################################################
```

```
#############################################################################################
fit2<-prune(arbolmax1, cp=0.002)
plot(fit2, uniform = TRUE, compress=T, margin=0.1)
text(fit2, use.n = TRUE, all=T, pretty=0, cex = 0.5)
pp<-predict(fit2, newdata=per_test, type="class")
aa1<-table(per_test$Asiste, predict(fit2, newdata=per_test, type="class"))
prop.table(aa1, m=1)
#############################################################################################
fit3<-prune(arbolmax1, cp=0.003)
plot(fit3, uniform = TRUE, compress=T, margin=0.1)
text(fit3, use.n = TRUE, all=T, pretty=0, cex = 0.5)
pp<-predict(fit3, newdata=per_test, type="class")
aa1<-table(per_test$Asiste, predict(fit3, newdata=per_test, type="class"))
prop.table(aa1, m=1)
#############################################################################################
fit4<-prune(arbolmax1, cp=0.004)
plot(fit4, uniform = TRUE, compress=T, margin=0.1)
text(fit4, use.n = TRUE, all=T, pretty=0, cex = 0.5)
print(fit4)
pp<-predict(fit4, newdata=per_test, type="class")
aa1<-table(per_test$Asiste, predict(fit4, newdata=per_test, type="class"))
prop.table(aa1, m=1)
post(fit4, title. = "Ã¡rbol de clasificaciÃ³n Variable Asistencia", pretty = 0)
#############################################################################################
fit5<-prune(arbolmax1, cp=0.01)
plot(fit5, uniform = TRUE, compress=T, margin=0.1)
text(fit5, use.n = TRUE, all=T, pretty=0, cex = 0.5)
pp<-predict(fit5, newdata=per_test, type="class")
aa1<-table(per_test$Asiste, predict(fit5, newdata=per_test, type="class"))
prop.table(aa1, m=1)
```

```
#############################################################################
#ARBOL DE CLASIFICACIÓN COHERENTE CON EL ESCENARIO 2#########################
#############################################################################
#Arbol máximo#
arbolmax2<-rpart(Asiste ~ sexo + E27 + JovenActivo + Madre_ausente + hijos + jefe  +
                   aniosed + climaeducativo + icv2 + Confort1 + TIC1 + Mdeo +
LN_YSVL_sin_joven_prom + Hacinamiento, data = per_training, method = 'class',
control=rpart.control(cp=0))
print(arbolmax2)
printcp(arbolmax2)
summary(arbolmax2)
plot(arbolmax2, uniform = TRUE, compress=T, margin=0.1)
text(arbolmax2, use.n = TRUE, all=T, pretty=0, cex = 0.5)
pp<-predict(arbolmax2, newdata=per_test, type="class")
aa1<-table(per_test$Asiste, predict(arbolmax2, newdata=per_test, type="class"))
#ConstrucciÃ³n de secuencia de Ã¡rboles anidados#
fit1<-prune(arbolmax2, cp=0.001)
plot(fit1, uniform = TRUE, compress=T, margin=0.1)
text(fit1, use.n = TRUE, all=T, pretty=0, cex = 0.5)
printcp(fit1)
pp<-predict(fit1, newdata=per_test, type="class")
aa1<-table(per_test$Asiste, predict(fit1, newdata=per_test, type="class"))
prop.table(aa1, m=1)
#############################################################################
fit2<-prune(arbolmax2, cp=0.002)
plot(fit2, uniform = TRUE, compress=T, margin=0.1)
text(fit2, use.n = TRUE, all=T, pretty=0, cex = 0.5)
printcp(fit2)
pp<-predict(fit2, newdata=per_test, type="class")
aa1<-table(per_test$Asiste, predict(fit2, newdata=per_test, type="class"))
prop.table(aa1, m=1)
#############################################################################


#############################################################################
fit4<-prune(arbolmax2, cp=0.004)
plot(fit4, uniform = TRUE, compress=T, margin=0.1)
text(fit4, use.n = TRUE, all=T, pretty=0, cex = 0.5)
printcp(fit4)
pp<-predict(fit4, newdata=per_test, type="class")
aa1<-table(per_test$Asiste, predict(fit4, newdata=per_test, type="class"))
prop.table(aa1, m=1)
post(fit4, title. = "Ã¡rbol de clasificaciÃ³n Variable Asistencia", pretty = 0)
#############################################################################
fit5<-prune(arbolmax2, cp=0.01)
plot(fit5, uniform = TRUE, compress=T, margin=0.1)
text(fit5, use.n = TRUE, all=T, pretty=0, cex = 0.5)
printcp(fit5)
pp<-predict(fit5, newdata=per_test, type="class")
aa1<-table(per_test$Asiste, predict(fit5, newdata=per_test, type="class"))
prop.table(aa1, m=1)
#############################################################################
```