

MODELACION DEL COSTO MEDIO ANUAL DEL SERVICIO MEDICO, PARA LA
POBLACION AFILIADA A UNA EPS DEL VALLE DEL CAUCA EN EL AÑO 2011
COMO ESTRATEGIA DE EVALUACION DE LA SUFICIENCIA DE LA UPC

WILLIAM ALFONSO GALVIS RAMIREZ

UNIVERSIDAD DEL VALLE
FACULTAD DE INGENIERIA
ESCUELA DE ESTADISTICA
SANTIAGO DE CALI
2014

MODELACION DEL COSTO MEDIO ANUAL DEL SERVICIO MEDICO, PARA LA
POBLACION AFILIADA A UNA EPS DEL VALLE DEL CAUCA EN EL AÑO 2011
COMO ESTRATEGIA DE EVALUACION DE LA SUFICIENCIA DE LA UPC

WILLIAM ALFONSO GALVIS RAMIREZ

Proyecto de Grado para optar al Título de
ESTADISTICO



(Director)
JAIME MOSQUERA RESTREPO

UNIVERSIDAD DEL VALLE
FACULTAD DE INGENIERIA
ESCUELA DE ESTADISTICA
SANTIAGO DE CALI
2014

RESUMEN

En el desarrollo del presente estudio; se pretende aplicar la técnica estadística para la modelación del Costo Medio Anual de los servicios médicos requeridos por la población Afiliada a la EPS del Valle del Cauca en el año 2011. Las EPS's Empresas Administradora de Planes de Beneficio o Promotoras de Salud son las aseguradoras de la población respecto a los riesgos de salud a través del Plan Obligatorio de Salud (POS) en el Régimen Contributivo y/o Régimen Subsidiado entre otros. Teniendo en cuenta el propósito de este trabajo la técnica estadística usada en primer lugar modela la probabilidad de uso de los servicios médicos de acuerdo a los factores de riesgo de la población afiliada a través del Modelo Logístico y en segundo lugar para la modelación de los costos esperados se usa el Modelo Lineal Generalizado, una vez se tienen estos modelos se combinan en producto dando paso a lo que se define como el modelo de dos partes, que permite hallar el valor del Costo Medio Anual por servicios médicos, con la aplicación de este modelo en el presente trabajo se espera que la EPS cuente con más elementos que le permitan proponer al estado ajustes coherentes y eficientes en los incrementos del valor de la UPC. Para la realización y desarrollo de este estudio se usó la base datos que la EPS reporto al Ministerio de la Protección Social para el estudio de Suficiencia de la UPC en año 2011 por los servicios médicos dados en el plan Obligatorio de Salud en el año de la referencia; Los programas usados fueron: SQL Server, R-Project, Microsoft Excel y Powerpivot.

Palabras Claves: Unidad de Pago por Capitación (UPC), Modelo Logístico, Modelo Lineal Generalizado, Modelo de dos Partes, Sistema General de Seguridad Social en Salud (SGSSS).

ABSTRACT

In the development of this study; is intended to apply the statistical technique for modeling the Mean Annual Cost of medical services for the population EPS Affiliated to the Cauca Valley in the year 2011. The Manager Companies EPS's Benefit Plans or Health Coaches are insurers of the population regarding health risks through the Mandatory Health Plan (POS) in the contributive regime and / or Subsidized Regime among others. Given the purpose of this paper the statistical technique used first modeled the probability of use of health services according to risk factors for the population covered by Logistic Model and secondly for modeling costs expected the Generalized Linear Model is used, once these models are combined in the product giving rise to what is defined as the two-part model, which allows us to find the value of the Mean Annual Cost for medical services with the application of this model in this paper is expected to EPS has more elements to the state will propose coherent and efficient adjustments increases the value of the UPC. To carry out this study and development of the database that the EPS reported to the Ministry of Social Protection to study Sufficiency of UPC in 2011 for medical services provided in the Mandatory Health Plan in the year used the reference; The programs used were: SQL Server, R-Project, Microsoft Excel and PowerPivot.

TABLA DE CONTENIDO

1. PLANTEAMIENTO DEL PROBLEMA	12
1.1 Objetivos.....	14
1.1.1 Objetivo General.....	14
1.1.2 Objetivos Específicos.....	14
1.2 Justificación	15
2. REVISIÓN DE ANTECEDENTES.....	17
3. EL SISTEMA DE SEGURIDAD SOCIAL EN COLOMBIA	21
4. MARCO TEÓRICO ESTADÍSTICO	26
4.1 Modelo de dos partes.....	26
4.2 Modelo de Regresión Logística.....	28
4.2.1 Estimación e Interpretación de los Coeficientes Modelo de Regresión Logística.....	30
4.2.2 Estimación por Intervalos e Interpretación para β_i	31
4.2.3 Pruebas de Bondad de Ajuste y Significancia de los Parámetros	33
4.2.4 Supuestos del Modelo Logístico.....	37
4.2.5 Errores del Modelo	37
4.2.6 Selección de Variables	38
4.3 Modelo Lineal Generalizado	38
4.3.1 Funciones de probabilidad para los modelos lineales generalizados	39
4.3.2 Componente sistemático (predictor lineal)	41
Variables dummy	43
Términos mixtos.....	44
Estimación de los parámetros por medio de la máxima verosimilitud según (Dobson & Barnett, 2008).....	44
4.4 La Distribución gamma.....	48
Selección de las covariables.....	52
Prueba de Bondad de Ajuste	54
Inspección de los Residuales	54
5. METODOLOGÍA.....	56
5.1 Fuentes de Información	56

5.2	Análisis Exploratorio de los Datos.....	59
5.3	Modelación del Costo de Servicio	60
5.3.1	Modelación de la Probabilidad de Uso	60
5.3.2	Modelación Condicional del Costo del Servicio	62
5.4	Construcción del Modelo de dos Partes	63
5.5	Evaluación de la Suficiencia.....	63
6.	RESULTADOS	65
6.1	Análisis Exploratorio de Datos	65
6.2	Ajuste del Modelo de Regresión Logística para Probabilidad de Uso	80
6.3	Ajuste del Modelo de Lineal Generalizado Costo del Servicio.....	83
6.4	Construcción del Modelo de dos Partes	87
6.5	Evaluación de la Suficiencia de la UPC año 2011	88
7.	CONCLUSIONES	93
8.	REFERENCIAS BIBLIOGRAFICAS.....	98

INDICE DE TABLAS

Tabla No. 4.1: Factores de Riesgo utilizados por Países para Estimar del Gasto Esperado en Salud.....	28
Tabla No. 4.2 Valores regresión logística cuando la variable Independiente es dicotómica	31
Tabla No. 4.3 Distribuciones más importantes de la familia Exponencial	41
Tabla No. 5.3 Análisis de Sensibilidad y Especificidad capacidad Predictiva del Modelo.....	61
Tabla No. 6.1 % según Genero, Zona, Tipo Afiliado	66
Tabla No. 6.2 Primeros Tres Departamento de Residencia de acuerdo a Cantidad de Afiliados	69
Tabla No. 6.3 Resultados Test de Fisher.....	70
Tabla No. 6.4 Agrupación de diagnósticos.....	72
Tabla No. 6.5 Agrupación de Prestaciones	72
Tabla No. 6.6 Resumen de Estadísticas Costo	73
Tabla No. 6.7 Resultados del Modelo Completo (Seleccionado)	80
Tabla No. 6.8 Resultados Sensibilidad y Especificidad del Modelo Completo (Seleccionado)	82
Tabla No. 6.9 Resultados del Modelo Lineal Generalizado Completo.....	84
Tabla No. 6.10 Resultados del Modelo Lineal Generalizado Completo	84
Tabla No. 6.11 Estadísticas Residuales	86
Tabla No. 6.12 Resultados de Modelación Costo Medio Anual x Grupo de Edad.....	89
Tabla No. 6.13 Resultados de Modelación Costo Medio Anual x Genero	89
Tabla No. 6.14 Resultados de Modelación Costo Medio Anual x Zona Geográfica	90
Tabla No. 6.15 Resultados de Modelación Costo Medio Anual x Tipo Afiliado.....	90
Tabla No. 6.16 Resultados de Modelación Costo Medio Anual x Departamento de Residencia	90
Tabla No. 6.16 Resultados de Modelación Costo Medio Anual x Departamento de Residencia	91
Tabla No. 6.17 % de Siniestralidad Costo Ejecutado y Costo Modelado por Grupo de Edad.....	92
Tabla No. 6.18 % de Siniestralidad Costo Ejecutado y Costo Modelado por Genero	92
Tabla 5.1 variables población del estudio de suficiencia POS – UPC	95
Tabla 5.2 variables de prestación del estudio de suficiencia POS – UPC	97

INDICE DE FIGURAS Y GRAFICAS

Gráfica No. 4.2 Ejemplo de Curva ROC.....	35
Gráfica No. 5.1: Gestión de información del Estudio de suficiencia POS – UPC. Colombia, año 2008	56
Gráfica No. 6.1 Histograma Distribución Edad de los Afiliados	65
Gráfica No. 6.2. Pirámide poblacional de afiliados a EPS del régimen contributivo, año 2011.....	66
Gráfica No. 6.3. Afiliados por Departamento de Residencia.....	67
Gráfica No. 6.4 Distribución Etarea por Departamento de Residencia.....	68
Gráfica No. 6.5 Distribución Etarea por Tipo Afiliado	68
Gráfica No. 6.6 Distribución Etarea por Zona de Residencia	69
Gráfica No. 6.7 Porcentaje de Uso del Servicio Población Total.....	70
Gráfica No. 6.8 % de Uso de los Servicios	71
Gráfica 6.9 % de Uso de los Servicios por Grupo de Edad	71
Gráfica 6.10 Costos Promedio por Género según Grupo Edad	73
Gráfica No. 6.11 % Relación Costos de Servicios vs Edad.....	74
Gráfica No. 6.12 % Relación Costos de Servicios vs Edad según Genero.....	74
Gráfica No. 6.13 Comportamiento de los Costos según el Grupo de Edad	75
Gráfica No. 6.14 Comportamiento de los Costos según el Grupo de Edad y Genero	76
Gráfica No. 6.15 % Histogramas de los Costos de Servicios, Generales, según Genero, Zona	78
Gráfica No. 6.16 Costos Promedio por Departamentos	79
Gráfica No. 6.17 Curva ROC Probabilidad de Uso.....	82
Gráfica No. 6.18 Y estimado vs Residuales.....	85

DEDICATORIAS

Dedico este gran logro a mi amado DIOS que sin él hubiese sido imposible, a mis padres por forjar en mí la voluntad y disciplina para obtener cada triunfo en la vida a pulso y con esfuerzo; a no sucumbir ante las dificultades, a mi esposa Sandra Patricia y a mi hijo Miguel Ángel por estar siempre presentes y por ser el motor de mi vida; por sembrar en mí la pasión de hacer las cosas y hacerlas bien, a mis hermanos por estar pendientes siempre de mí, a mis queridas tías que están en el cielo, quienes permitieron que este sueño se iniciara algún día en esta ciudad, a las personas que de un modo u otro contribuyeron para que este propósito de vida se hiciera realidad.

AGRADECIMIENTOS

Infinitas gracias doy a mi Universidad del Valle, por haberme recibido y formarme como profesional, como persona, a la escuela de Estadística y su excelente grupo de profesores los cuales siempre están dispuestos apoyar con paciencia, con amor y generosamente a compartir su gran conocimiento. Para mi director de tesis quien creyó en mí, que supo animarme para que lograra el objetivo, que con sus consejos y orientaciones me supo llevar a la meta como cual niño es llevado de la mano, Gracias profesor Jaime Mosquera que DIOS le bendiga esa maravillosa labor de formar profesionales y sobre todo seres humanos. A la EPS del Valle de Cauca donde se pudo gestar este hermoso Proyecto, en especial a la Doctora María Fernanda Ochoa por permitirme el acceso a la información y uso de herramientas corporativas para el desarrollo de lo propuesto, por su orientación y convocatoria para llegar con buen paso a la meta.

INTRODUCCION

La Ley 100 de 1993, creó el Sistema General de Seguridad Social en Salud (SGSSS) y con él las entidades para administrar y prestar los servicios de salud tales como; las Empresas Promotoras de Salud (EPS, las Administradoras del Régimen Subsidiado (ARS, hoy EPS-S) y las Instituciones Prestadoras de Servicios de Salud (IPS). Las EPS y EPS-S son las aseguradoras de la población del Régimen Contributivo y el Régimen Subsidiado respectivamente. Estas entidades contratan directamente a las IPS (hospitales o centros de salud, incluyendo a las Empresas Sociales del Estado, ESE) para la prestación de los servicios de sus afiliados a través del Plan Obligatorio de Salud (POS). Así, las EPS y las ARS se comportan como compañías aseguradoras que administran riesgos y por lo cual reciben una remuneración.

El estado con el fin de garantizar el plan obligatorio de Salud implemento el estudio de la suficiencia y de los mecanismos de ajustes de riesgos de la Unidad de Pago por Capitación(UPC), para lo cual las EPS's deben consolidar y reportar la información de los servicios prestados a la población en un año, de esta manera el estado monitorea y evalúa la real cobertura del Plan Obligatorio de Salud (POS) y se determina el ajuste de la UPC a pagar a las EPS por los afiliados, de igual manera se estudia la inclusión de nuevas actividades, procedimientos e intervenciones.

Los ajustes e incrementos realizados por el estado al valor de la UPC, han hecho que las EPS's se pronuncien al respecto; argumentan que los ajustes son insuficientes; dada la gran demanda de servicios por parte de la población afiliada, es así que en muchos de los casos se ha deteriorado la calidad de la prestación de los servicios; colocando en riesgo la salud y la vida de las personas; por la escases de los recursos que informan las EPS's, teniendo en cuenta lo anterior muchas de la EPS's afectadas por los ajustes del gobierno han implementado sus propios estudios a fin de evaluar una mejor metodología para el cálculo del valor a pagar por la UPC.

Para la EPS del valle de Cauca donde se realiza el presente estudio, los ajustes por parte del estado al valor de la UPC han sido insuficientes, lo cual hace que la operación se vuelva compleja y constantemente deban idearse estrategias para que la atención a los afiliados sea con la debida calidad y prioridad requerida, a fin de evaluar nuevas metodologías para el cálculo de la UPC, en el presente trabajo se describen los aspectos y pasos metodológicos usados para analizar la información insumo del estudio, que comprende las variables, los procesos de calidad y método del cálculo para el logro de los objetivos, es así que a través de 7 capítulos se desarrollan las actividades del presente trabajo, para lo cual en la parte inicial se relaciona la problemática y relevancia de realizar este trabajo, luego se presentan estudios relacionados por otros autores los cuales fueron pertinentes para conocer las técnicas utilizadas en la modelación de los Costos Medios en Salud. Con el fin de alcanzar los objetivos propuestos; a través del marco teórico se revisan los métodos estadísticos utilizados en la metodología propuesta. Con el fin de tener una primera aproximación al comportamiento de los Costos por servicios médicos y frecuencias de uso del servicio, se realiza análisis descriptivo, relacionado en el capítulo 5, posteriormente se modela la probabilidad de uso del servicio a través del Modelo Logístico y

se Modelan los Costos Esperados por medio del Modelo Lineal Generalizado, usando el Modelo de dos partes se calcula el costo medio anual del servicio médico para la población afiliada a la EPS. Finalmente se formulan conclusiones y recomendaciones, que sirvan de base y discusión a estudios futuros.

CAPITULO 1

PLANTEAMIENTO DEL PROBLEMA

En Colombia, la ley 100 de 1.993 establece que todo ciudadano debe participar del Sistema General de Seguridad Social en Salud (SGSSS), el cual debe ser garantizado por el estado en todos los niveles de complejidad, a través de lo que se ha denominado como el Plan Obligatorio de Salud (POS). En este orden de ideas, el estado delega la responsabilidad de este derecho fundamental en empresas intermediarias (EPS) y reconoce por este servicio un pago por cada afiliado (Cotizante o Beneficiario), para que estas a su vez garanticen la prestación de los servicios de salud incluidos en el POS. Este pago se denomina Unidad de Pago por Capitación (UPC). (Fedesarrollo, 2012)

La UPC, informalmente, se puede definir como la suma mensual que el SGSSS reconoce a las EPS's por cada afiliado cotizante o beneficiario cubierto, para la organización y garantía de la prestación de los servicios incluidos en el POS; dicho valor, debe ser actualizado anualmente por el Consejo Nacional de Seguridad Social en Salud (CNSSS), quien mediante estudios de carácter técnico determina la cuantía en la cual debe ser actualizado dicho pago, teniendo como referente las variables fundamentales que afectan el gasto en salud del sistema, como son: Edad del afiliado, perfil epidemiológico de la población, tasas de crecimiento de la misma por género, zona de procedencia, entre otras.

De acuerdo con el estudio de Fedesarrollo en 2012, lo que se observa es que los incrementos en la UPC no reflejan de manera adecuada el aumento en los costos operativos que experimentan las EPS's. A pesar de que las EPS's demuestran en los estudios¹ realizados que existen una gran cantidad de prestaciones que no hacen parte del plan POS (NO POS) y que estas prestaciones solo se pueden recobrar en forma parcial al Fondo de Solidaridad y Garantía FOSYGA². Según ello las EPS's han estado asumiendo parte de estos costos, lo que ha generado una grave crisis y en general vulnerado el sistema, a punto tal de cuestionar su sostenibilidad. Pareciera que el marco de prestaciones del Plan POS se hubiese congelado a pesar que la demanda de servicios sigue creciendo. También la situación que ha hecho que lo No POS siga creciendo son las enfermedades raras o huérfanas³, cuyo costos son sumamente altos. Según las EPS's existen evidencias para pensar que el valor de UPC debe ser ajustado

¹ Estudios de la suficiencia y de los mecanismos de ajuste de riesgo para el cálculo de la Unidad de Pago por Capitación para garantizar el Plan Obligatorio de Salud en años 2006 a 2010 (Ministerio de la Protección Social Colombia)

² El Fosyga es una cuenta adscrita al Ministerio de Salud y Protección Social manejada por encargo fiduciario, sin personería jurídica ni planta de personal propia, cuyos recursos se destinan a la inversión en salud tomado de : www.fosyga.gov.co

³ Son aquellas crónicamente debilitantes, graves, que amenazan la vida y con una prevalencia menor de 1 por cada 2.000 personas (Ley 1392 de 2010. Artículo 2)

a la menor brevedad posible. En la actualidad se presenta alta preocupación, ya que el FOSYGA no cuenta con recursos garantizados para el sostenimiento del POS en el Régimen Contributivo debido al crecimiento de lo No POS. (Universidad Nacional de Colombia, 2010)

La EPS en la que se realizará este estudio; la situación anteriormente mencionada no es una problemática ajena o desconocida, al igual que el resto de EPS's del sector; se está viendo altamente impactada, según (Clavijo, Torrente, Santamaria, & Garcia, 2008), los ingresos por UPC son inferiores al gasto que las EPS's deben asumir para mantener la población afiliada, técnicamente el Egreso de gasto médico es superior al Ingreso, lo que pone en riesgo la operación de la entidad aseguradora.

De esta manera, con el objetivo de evaluar su riesgo operativo/financiero, es necesario evaluar periódicamente la suficiencia de la UPC, entendida como la capacidad de cobertura de los egresos. Para ello es necesario abordar el problema a través del análisis de su población afiliada, su perfil epidemiológico, características de Morbilidad y descripción de su estructura de egresos. En este sentido el presente trabajo de grado tiene como propuesta aplicar un modelo estadístico que incluya las variables críticas poblacionales, de consumo y Morbilidad para el valor de la UPC de acuerdo a la información particular de la EPS y de esta manera estimar, evaluar la suficiencia del valor de la prima cancelada. Como caso de estudio se tomaran los registros del año 2011.

1.1 Objetivos

1.1.1 Objetivo General

- Construir un Modelo estadístico para la estimación del costo medio del servicio médico del conjunto de afiliados a una EPS del Valle del Cauca en el año 2011.

1.1.2 Objetivos Específicos

- Caracterizar la población afiliada a la EPS en el año 2011, en cuanto a sus condiciones socioeconómicas y su perfil de consumo.
- Ajustar un Modelo Estadístico para la estimación del Costo Medio anual, según características de la Población afiliada a la EPS.
- Evaluar la suficiencia del valor actual de la UPC reconocido por el Ministerio de Protección Social frente a las condiciones de costo observadas y estimaciones de Costos obtenidas en el Modelo.

1.2 Justificación

Las EPS's se crearon en el SGSSS para garantizar la afiliación y prestación de los servicios de salud a la población Colombiana, servicio por el cual económicamente el estado paga un valor llamado prima o UPC por cada uno de sus afiliados. Desde 1993, momento que surgió la Ley 100 y todo el marco técnico legal para la implementación del SGSSS, según el informe del Ministerio de la Protección Social a la Comisión de Regulación en Salud en año 2010 se analiza que este valor no está siendo suficiente, lo cual hace riesgosa la operación de dichas entidades, toda vez que desequilibra el Sistema de Salud. Teniendo en cuenta esta situación; el estado colombiano se ha encaminado a realizar estudios a partir de la información que las EPS's consolidan por todos los servicios prestados en un año, llamado "Estudio de Suficiencia de la UPC", y es aplicable tanto a EPS's del Régimen Contributivo⁴ como del Régimen Subsidiado⁵, este estudio se viene realizando cada año y es utilizado para evaluar el incremento del valor de la UPC del año siguiente, es decir la información del año 2010 es utilizada para realizar el análisis y cálculo de la UPC del año 2011. En estudios contratados por el estado Colombiano como el de la Universidad Nacional en el año 2010 se ha encontrado que no es la mejor manera de hacerlo; se ha evidenciado que hay entidades que envían la información sin la calidad requerida, debido a la falta de recursos para contar con medios tecnológicos que permitan manejar la información con la calidad solicitada en el estudio. También se da la situación que de un periodo a otro cambia la forma de reporte; ej.: en un periodo enviaron datos detallados, al siguiente periodo envían la información agrupada lo cual impedirá realizar análisis comparativos o aplicar la técnica de análisis de manera eficaz.

En general se puede mencionar que la literatura referida al tema; a nivel Nacional se ha abordado primordialmente desde una perspectiva macro, toda vez que viene respaldada por la evidencia que proporciona para el Gobierno Nacional, contar con la totalidad de observaciones válidas que constituyen el universo poblacional de prestaciones médicas, a partir de las cuales se obtienen los valores esperados de la Unidad de Pago por Capitación (UPC) que debe pagar el Gobierno por cada persona afiliada al sistema de acuerdo a sus variables fundamentales (edad, género y ubicación geográfica). Desde luego que esta discusión reviste gran importancia, toda vez que el valor de la prima decretada constituye la piedra angular que permite garantizar, en mayor o menor medida, la sostenibilidad financiera del sistema general de seguridad social en salud (SGSSS) y más en las actuales circunstancias donde el debate gira en torno a la conveniencia de continuar o no con la intermediación por parte de las Entidades Promotoras de Salud (mismas que argumentan la conveniencia de ajustar al alza los valores de dicha prima, como una alternativa para aliviar la crisis del sistema), en un modelo que, dicho sea de paso, empieza a mostrar claros síntomas de agotamiento.

⁴ Estarán afiliados la clase laboral por empresa o independientes, pensionados o jubilados y las personas con capacidad de pago más los beneficiarios de sus respectivos grupos familiares: Tomado de Nueva Historia de Economía en Colombia Kalmanovitz 2010

⁵ Estarán vinculados aquellas personas sin capacidad de pago, entendiéndose aquella población vulnerable que definitivamente debe ser subsidiada. Tomado de Nueva Historia de Economía en Colombia Kalmanovitz 2010

Teniendo en cuenta lo anteriormente mencionado, se realizara la estimación del valor de UPC 2011 con la información de la EPS en mención, haciendo uso de métodos estadísticos y econométricos a fin contrastar de manera objetiva, si el valor de la prima determinada por el Gobierno para cada uno de los grupos de riesgo, cubren de forma suficiente las necesidades reales de funcionamiento de la EPS en la cual se realiza el presente estudio.

CAPÍTULO 2

REVISIÓN DE ANTECEDENTES

En este capítulo se relacionan algunos estudios e investigaciones que tienen relación con el manejo de la información en un sistema de salud para afiliados que incurren en un gasto; como aquellos afiliados; que no generan ningún gasto, estos estudios han sido realizados en diferentes partes del mundo de manera exitosa, siendo referentes de gran relevancia para el presente estudio.

“Modelado del riesgo usando los modelos lineales generalizados”: (Blough, Madden, & Hornbrook., 1999) realizaron un estudio en el estado de Washington (USA) a partir de la base de datos de los empleados estatales; inscritos entre 1991 y 1993, excluyeron del estudio a los jubilados, excepto los menores de 65 años, igualmente las personas mayores de 75 años fueron omitidos debido al pequeño número de estas personas; resultando una muestra de 126.393 individuos de la cual se tomo una muestra aleatoria del 1%, las 1.284 observaciones resultantes se destinaron a formar una base de datos de prueba para la validación del modelo. Las 125.109 observaciones restantes conformaron el conjunto de datos de entrenamiento que se utiliza para el desarrollo del modelo. En este artículo se presenta un nuevo enfoque para el modelado de la segunda etapa de los modelos de dos partes, los cuales permiten en una primera etapa a través de un modelo logístico o probit calcular la probabilidad de que una persona incurra en un gasto, en la segunda etapa; para aquellas personas de la población que tengan un gasto asociado; calcular la distribución de los costos. Una vez se tengan estas dos estimaciones se combinan obteniendo el valor esperado dado los factores de riesgo. Los autores utilizan extensiones del modelo lineal generalizado para la predicción de los gastos médicos de una persona, dadas ciertas características demográficas. En la segunda parte del modelo, se uso un modelo lineal generalizado con el fin de modelar la distribución del gasto de aquellas personas que los presentan. Este enfoque tiene pocos supuestos y evita la necesidad de transformar los datos, al tiempo que representa una reparametrización del modelo que permite conservar la escala original (en este caso dólares), de la variable de respuesta. Los datos de gastos médicos normalmente presentan una serie de características que deben ser tenidas en cuenta con el fin de desarrollar un modelo válido, el modelo de dos partes ha encontrado una gran aceptación en la metodología de evaluación de riesgos. La parte 1 de dicho modelo trata con la masa de datos en cero. La variable de respuesta de cada individuo en la parte 1 del modelo es dicotómica, ya sea que el individuo presente gastos médicos (1), o no los tenga (0) durante el periodo. La parte 2 del modelo sólo se refiere a las personas que presentan gastos (sumatoria del valor de los servicios en el periodo por cada individuo). Para estimar los parámetros del modelo en la parte 2 se hace uso del modelo de Regresión ordinario (Normalmente es una transformación logarítmica), de esta forma la regresión ordinaria se aplica sobre los datos transformados para obtener una predicción general de gastos para individuos, una vez se han obtenido los resultados de cada modelo se multiplican las dos predicciones de cada parte del modelo (de dos partes). Este enfoque de dos partes es de uso general para las distribuciones mixtas, lo que Posibilita modelar de manera efectiva la masa de datos en

cero por separado. La Especificación de la parte 1 y parte 2 de las distribuciones de probabilidad correspondientes; permiten obtener estimaciones de máxima verosimilitud de los parámetros del modelo. Se encontró que el uso de modelos lineales generalizados, la cuasi-verosimilitud y Cuasi-verosimilitud extendida ofrecen un enfoque flexible para el modelado de la segunda parte de los modelos de dos partes con respecto a los gastos médicos, lo que confirma que es la técnica apropiada para ser usada en el modelado de los datos de gasto médico de la EPS donde se realiza el trabajo de grado. También se encontró que este tipo de modelos usados en la segunda parte del modelo no requieren ninguna hipótesis de distribución y la elección de enlace y las funciones de varianza se pueden probar formalmente mediante la incorporación del modelo en clases paramétricas de cada uno, las variables en el modelo para estimar el gasto medico son el género, la edad de interacción con el tratamiento tomando cada año como una categoría separada, la composición de la familia, situación o estado y Grupos de Atención Ambulatoria ACG.

Mullahy (1998) realiza estudio a partir de los datos de la Encuesta Nacional de Salud de 1992 en Estados Unidos, con una muestra de $N = 36.111$ observaciones en personas con edades de 25 a 64 años, la variable dependiente es el número de visitas al médico en los últimos doce meses anteriores a la encuesta. Para esta medida de la variable dependiente, $y = 0$ contó con una muestra $N_0 = 8.513$ (23.6%) casos, con respecto a $y > 0$ la muestra fue $N_+ = 27.598$ casos (76,4%), la finalidad básica del estudio es evidenciar las circunstancias en las que el modelo estándar de dos partes con retransformación homocedástica dejará de comunicar conclusiones consistes sobre importantes parámetros. También el artículo ilustra algunos enfoques alternativos que puedan ser de utilidad en este tipo de aplicaciones. Considerando que el objetivo del presente trabajo de grado es modelar los datos de gasto medico en el año 2011 en una EPS, este artículo se considera relevante y útil para conocer en que circunstancias los modelos de dos partes, no son eficientes, adicionalmente el articulo dá a conocer acerca de otras metodologías aplicadas a fin de dar solución a un mismo problema. En economía de la salud muchos de los resultados que son hallados de manera empírica presentan dos propiedades estadísticas de gran importancia la primera que $y > 0$ y la segunda que $y = 0$ para lo cual se hace énfasis que estos resultados de valores en cero no se pueden ignorar, la econometría típicamente se ha basado en tres estrategias bien conocidas en la que se encuentran estas estructuras de datos como son: El modelo de dos partes (2PM), El modelo de selección de la muestra (SSM) y el método de regresión como es el enfoque de Heckman, (Heckman 1979).

Se tiene que la primera parte del modelo de dos partes asume que $\Pr(y > 0 | x)$ y se rige por un modelo de probabilidad binaria paramétrica, tales como el modelo logit o modelo probit. Para la parte dos del modelo; conceptualiza que $E[\ln(y)|y > 0, x]$ es una función lineal de x , e.g. donde $E[\ln(y)|y > 0, x] = x\beta$, para esta segunda parte se hace uso de un modelo lineal generalizado. Los resultados presentados en este artículo sugieren que se debe utilizar el enfoque de la norma Homocedástica del modelo de dos partes (2PM) con mucha cautela en aplicaciones micro econométricas para los centros de interés con $E[y|x]$ y efectos parciales asociados. La suposición básica de identificación de β en ese

modelo, a saber $E[\varepsilon|y > 0, x] = 0$, no es lo suficientemente potente como para identificar otros parámetros de interés como $E[\varepsilon|y], \delta(x)$.

Lin (2008) realizó estudio en las tres principales regiones geográficas de Taiwán como son: el área de Taiwán, las zonas montañosas y las islas del litoral, el objetivo propuesto examinar la relación entre los factores de riesgo controlables de la salud y gastos médicos reales de los individuos en Taiwán, para alcanzar dicho objetivo se combinó información de la Encuesta Nacional de Salud 2001 (ENS) y los datos de la Base de Datos Nacional de Investigación en Salud (NHIRD) del mismo año. Para el caso de estudio solo se tuvieron en cuenta los datos de la misma región de Taiwán, el número de observaciones en esta muestra fue de 18.144 personas; la fuente de información contiene amplia información sobre factores de riesgo controlables de la salud relacionados con el consumo de: tabaco, alcohol, la nuez de betel de mascar, incluso si las personas desayunan regularmente y si los encuestados hacen o no ejercicio.

Para realizar los análisis estadísticos el autor se apoyó en un modelo de regresión de dos partes, basándose en la teoría de los autores Duan, Manning, Morrisb, & Newhousea (1983), se logró explorar y conocer la asociación entre los factores de riesgo modificables de la salud y los gastos médicos en Taiwán, se resalta la ventaja de utilizar un modelo de dos partes, de acuerdo con lo expuesto en este estudio, este modelo proporciona información detallada sobre el proceso de la utilización de asistencia sanitaria, permite dar solución a dos problemas comunes; relacionados con los datos de los gastos de la salud y con el gran número de usuarios que no utilizan los servicios médicos, el modelo consiste en dos etapas de estimación. Entre los resultados se encontró que variables del estilo de vida están fuertemente correlacionadas con la probabilidad de uso de los servicios ambulatorios, hospitalarios y los costos relacionados en la mayoría de los modelos las personas con hábitos de ejercicio son menos propensos a utilizar los servicios de atención hospitalaria incurren en gastos de hospitalización más bajos. Como consecuencia de ello, las políticas de salud promueven la actividad física y no fumar para frenar el aumento de los gastos de salud en Taiwán. Este estudio queda como una propuesta abierta a futuras investigaciones que requieran evaluar la causalidad, el uso de modelos estadísticos más avanzados y mejores conjuntos de datos.

Deb, Munkin, & Trivedi, (2006) aplicaron un modelo bayesiano de dos partes a una muestra de 20.460 personas, obtenida de la encuesta de panel de gastos médicos. En este artículo se plantea estudiar el impacto de la atención administrada en los gastos totales de atención médica ambulatoria y hospitalaria. La variable de los gastos médicos, el resultado de interés, tiene una importante proporción de ceros que se manejan con el modelo de dos partes, se modela tanto los gastos de hospitalización, incluyendo todos los tratamientos hospitalarios, como los gastos ambulatorios, que incluyen el resto de los gastos totales de los tratamientos médicos; como son las consultas médicas, visitas ambulatorias, visitas a urgencias, y el gasto en medicamentos recetados. Para lograr el objetivo propuesto utilizan un modelo de dos partes, esta técnica introduce flexibilidad de modelado al permitir que los valores en cero y los

valores positivos de los gastos en salud se generen por medio de dos procesos separados, adicionalmente integran el modelo de dos partes y el modelo de selección Multinomial en un único marco a lo cual llaman el modelo de dos partes extendido o endógeno (ETPM) teniendo en cuenta lo anterior los autores implementaron una estrategia de estimación paramétrica que permite desarrollar, un marco de estimación bayesiano basado en un modelo de dos partes prolongado (ETPM) el cual respeta la endogeneidad y la naturaleza Multinomial de la elección del seguro. Se introduce la heterogeneidad no observada por medio de variables latentes, correlacionadas a través de las opciones de seguros. Se comenta que las características sobresalientes de los datos de gastos de salud particular y los datos de utilización más general, incluyen, además de no negatividad de los resultados, una fracción significativa de valores ceros. Los resultados están caracterizados por la asimetría positiva y el exceso de Curtosis. Se realizaron comparaciones de los resultados con estudios anteriores, comprobando que los planes de Organización y Mantenimiento de la Salud de acuerdo con sus siglas en inglés (HMO) tienen menores tasas de ingreso al hospital y duración de la estancia, igualmente hay un menor uso de pruebas y procedimientos costosos según la referencia de (Glied, 2000). El estudio también expone que el rápido crecimiento de la población de atención administrada durante la década de 1990, así como la consolidación y la organización de proveedores preferidos (PPO), con los individuos saludables y no saludables que entran en la HMO puede haber cambiado la composición de los afiliados, haciendo que haya aumento en la penetración de los planes de atención administrada, lo cual puede afectar a los mercados locales de salud y el uso del hospital, alentándolos a generar restricciones de control de costos en el uso del hospital para seguir siendo competitivos.

CAPITULO 3

EL SISTEMA DE SEGURIDAD SOCIAL EN COLOMBIA

A continuación se comentará parte de la evolución del Sistema de Seguridad Social en Salud SGSSS, importante referente para la comprensión de este estudio.

Las primeras instituciones que prestaron atención médica fueron las llamadas Fundaciones, que surgieron de la iniciativa privada y clerical desde tiempos coloniales y cuya financiación dependió de los aportes particulares o de la “Caridad Cristiana”, situación que se mantuvo durante la fase de la Republica Conservadora (1880 – 1930), En los años 40’s imperaba todavía el concepto de “Beneficencias”, financiada con recursos de las loterías y algunos impuestos reservados, que excluía la moderna responsabilidad de un estado de proveer “Bienestar” a todos los ciudadanos.

El estado fue incrementando su participación en la financiación de entidades de salud, concentrándose principalmente en el control de vectores de enfermedad en puertos y en las principales ciudades, que podían amenazar la economía y contagiar a otros países. Se inician campañas contra la malaria y se realizan jornadas de vacunación contra las enfermedades endémicas, apoyadas por la Fundación Rockefeller de 1917 en adelante; cuyos técnicos y médicos hicieron para la salud pública lo que Kemmerer⁶ había logrado para las instituciones monetarias y fiscales durante los años 20’s. Por lo demás había presiones sobre la calidad de la salubridad en los puertos de Buenaventura, Cartagena y Barranquilla, ejercidas por las autoridades norteamericanas de salud del canal de Panamá para prevenir epidemias.

En el periodo 1945 – 1946, inspirado en el modelo alemán Bismarckiano⁷; surgen los Sistemas de Salud Mexicano y Colombiano. En Colombia se crearon la Caja Nacional de Previsión (CAJANAL) y el Instituto Colombiano de los Seguros Sociales (ICSS), entidades que constituyeron al primer Sistema de Seguridad Social del país. Ellos surgieron como respuesta a la necesidad de un sistema público nacional que compensara la falencia de las empresas privadas que no garantizaba a los trabajadores el cabal cumplimiento de las llamadas *prestaciones patronales* (Pensiones, Cesantías, Salud y Riesgos profesionales), asociados todavía con la mentalidad caritativa católica.

⁶ **Edwin Walter Kemmerer** (Scranton, 29 de junio de 1875 - Princeton, 16 de diciembre de 1945), economista estadounidense, conocido como Money doctor (literalmente, «doctor dinero») o asesor económico de gobiernos de países de todo el mundo, especialmente latinoamericanos, tomado de www.wikipedia.org

⁷ El Canciller Alemán **Otto Von Bismarck** (el Canciller de Hierro) refrenda tres leyes sociales, que representan hasta hoy, la base del Sistema de Seguridad Social Universal: Seguro contra Enfermedad. 1883, Seguro contra Accidentes de Trabajo. 1884 Seguro contra la Invalidez y la Vejez. 1889, tomado de www.monografias.com

Paralelamente a Cajanal y al ICSS, fue constituyéndose una serie de entidades encargadas del manejo de la seguridad social de los empleados públicos de los niveles territorial y nacional. Con el paso del tiempo este sistema evidenció grandes deficiencias en términos de la calidad y oportunidad, lo que favoreció la creación de entidades, consultorios y sistemas de salud en las empresas, así como las instituciones de medicina prepaga y seguros de salud de naturaleza privada, cuyo mercado objetivo era la población con capacidad de pago, que demandaba mejores niveles de calidad en la atención y mayores comodidades hospitalarias. En el periodo de 1975 -1993 operó el llamado Sistema Nacional de Salud (SNS), conformado por tres subsistemas: i) el de Seguridad Social (entidades adscritas al ISS [antes ICSS], que le permitió al país realizar importantes avances en cobertura y cajas de previsión), ii) el privado (medicina prepagada y el gasto de bolsillo) y iii) el oficial (entidades públicas). En el SNS, el Ministerio de Salud era el responsable de la toma de decisiones financieras, del diseño de políticas de atención, del desarrollo de programas de salud pública y la administración del subsistema oficial, por su parte los entes territoriales eran los responsables de la administración y control de los hospitales de segundo y tercer nivel y los servicios seccionales de salud de cada departamento.

El SNS clasificó las atenciones en salud según el nivel y grado de complejidad y, con el fin de garantizar que cada persona recibiera la atención apropiada, las articuló entre sí mediante un sistema de referencia y contra referencia⁸ de pacientes. Con este esquema se esperaba que en los puestos de salud y hospitales locales (primer nivel de atención) se atendiera al 80% de los casos; en los hospitales regionales (segundo nivel) al 15%, en los hospitales universitarios (tercer nivel) y en las entidades especializadas (cuarto nivel), al restante 5% de los casos. En la práctica se presentó una subutilización de los centros de primer nivel y a la vez una sobredemanda en los hospitales de segundo y tercer nivel, hechos explicados por la deficiente calidad de la atención en el primer nivel y por el desconocimiento de la lógica del sistema por parte de la población, que prefería acudir a las entidades de mayor prestigio. Lo anterior se tradujo en un desempeño deficiente del SNS en materia de oportunidad, cobertura, calidad y eficiencia. En efecto la asignación centralizada de recursos, dependiente de los costos reportados por los hospitales y basada en los presupuestos históricos (subsidios a la oferta), impidió que surgieran los incentivos para que los diferentes actores del sistema público de salud buscaran mejorar la eficiencia. Adicionalmente, las autoridades territoriales y los directivos de hospitales no tenían potestad para modificar el funcionamiento de los hospitales y solucionar sus principales problemas.

A estas dificultades se sumaron problemas de equidad respecto a la población que se beneficiaba de la prestación de los servicios financiados con subsidios a la oferta. La Encuesta Nacional de Hogares de septiembre de 1992 reveló que cerca del 45% de la población urbana y el 80% de la población rural dependía de los servicios prestados por las entidades públicas o del sistema privado de salud o por el sector “informal” que ofrecía pocas garantías. Igualmente, el 12% de las hospitalizaciones y el 20% de las cirugías adelantadas en la red pública de hospitales, que debían dirigirse principalmente a la población más pobre, fueron recibidas por pacientes pertenecientes al 20% más rico de la población. Se

⁸ Mecanismo a través del cual el Sistema General de Seguridad Social en Salud, define estrategias que permitan garantizar a la población en general el acceso a los servicios de salud, tomado de (www.cruevalle.org)

evidenció, que el esquema de subsidios a la oferta no garantizó el acceso de buena parte de la población pobre a los servicios de salud. La constitución política de 1991 señaló la necesidad de reformar el sistema al establecerse que los servicios de salud debían organizarse con participación de la comunidad, de manera descentralizada y según niveles de atención.”

A principios de la década de los años 90 se decide realizar una reforma al sector salud, el principio para esta reforma indiscutiblemente va de la mano con lo propuesto en el “Pluralismo Estructurado”, en general planteaba la universalidad en la cobertura para el acceso de la población a los servicios de salud, tratando que el enfoque fuese una reforma equilibrada tanto para la población como para las instituciones. Esta reforma incluyó una gran cantidad de debates; la idea no gozaba de una gran aceptación; el gobierno se dio a la gran tarea de garantizar los beneficios de la implementación del nuevo sistema. Con la puesta en marcha de la ley 100 de 1993 se implementa el Sistema General de Seguridad Social en Salud (SGSSS). Esta nueva versión de la salud que ofrecía cobertura universal, mostraba la deficiencia del sistema anterior el país tan solo contaba con una cobertura del 33%. Para cumplir con esta cobertura universal el estado realizó ingentes esfuerzos a fin de contar con los recursos necesarios para la financiación de este nuevo sistema, el cual cuenta con actores tales como las entidades aseguradoras de la población denominadas Entidades Promotoras de Salud (EPS), las Instituciones Prestadoras de Servicio (IPS) como Clínicas, Hospitales, Laboratorios, etc., a la vez se cuenta con un Fondo que administra los recursos del sistema denominado FOSYGA.

Teniendo en cuenta los objetivos planteados con este nuevo sistema aún persistía la falta de equidad, razón por la cual y con el ánimo de aumentar la cobertura el gobierno establece que el sistema debe contar con unos regímenes: i) Régimen contributivo, ii) Régimen Subsidiado y iii) Régimen Especial (dado por convenciones colectivas de algunos gremios o empresas del estado). Al régimen contributivo estarían afiliados la clase laboral por empresa o independientes, pensionados o jubilados y las personas con capacidad de pago más los beneficiarios de su respectivos grupos familiar, en cuanto al régimen Subsidiado estarían vinculados aquellas personas sin capacidad de pago, entendiéndose aquella población vulnerable que definitivamente debe ser subsidiada. Para el régimen Especial, que es donde están algunas entidades del gobierno tales como Fuerzas Militares, Banco de la Republica, Ecopetrol, Magisterio, Universidades públicas, etc. Por último las Entidades Adaptadas al Sistema (EAS) que hace las veces de EPS, cuya condición para su funcionamiento es que no pueden afiliar personal, que no sean familiares o empleados de entidades como el caso de las empresas publicas de Medellín, Fondo Pasivo de Ferrocarriles Nacionales, Universidad del Valle, entre otras (Kalmanovitz, 2010).

Las EPS's, pueden ser de naturaleza Pública, Privada o mixta; estas entidades son las responsables de Afiliar y llevar los registros de los afiliados y tienen asignadas dos tipos de función: Gestión del aseguramiento o Protección de la Salud. En este sistema las EPS's reciben un valor por cada afiliado denominado UPC que es lo que se conoce en el mercado de las aseguradoras como Prima de Seguros. Entre las funciones que las EPS's deben realizar; está la de enviar al FOSYGA el valor diferencia entre los ingresos por Cotización y UPC en unas fechas determinadas.

Para regular a las EPS's el estado creó la Comisión de Regulación en Salud CRES la cual define el valor de UPC para cada régimen; este incremento es anual, en caso que este ente no realice dicho incremento se realizará con el valor de la inflación causada del año inmediatamente anterior. La CRES debe soportar las decisiones al respecto de la UPC, para tal fin se realizó un concurso por méritos y la institución que resultó ganadora fue la universidad Nacional de Colombia, la cual desarrolló la metodología y estudios para actualizar el POS según el acuerdo 008 de 2009 para los regímenes Contributivo y subsidiado; deben también garantizar la actuación de todos los actores y ciudadanos. La fuente de información son los datos de las EPS's relacionado por servicios prestados a los afiliados, identificándolos por la modalidad de plan de atención si corresponden a la cobertura del POS o están por fuera de dicha cobertura identificándolos como NO POS (Tutelas y Comité Técnico Científico denominado CTC).

Una de las problemáticas en el SGSSS es la selección adversa la cual se da cuando los consumidores de servicios de salud conocen el riesgo de salud y el asegurador no; lo que implica que el valor de la prima con el que se llega o desea cubrir las demandas de servicios es insuficiente. También se presenta la Integración vertical, ocurre cuando se opera en el sistema sin separación de funciones es decir es asegurador y prestador de servicios a la vez, lo que implica que la eficiencia se verá comprometida, esta figura de ser comprador de servicios y prestador a la vez se da con el fin de reducir costos.

La UPC: incluye el valor de la administración de dicho seguro para el régimen subsidiado, esto no se aplicó al régimen contributivo para dar la libertad y facilidad en la operación del manejo financiero y oportunidad de ganancia. Las EPS's cuentan con otros Ingresos como son los copagos, cuotas moderadoras, Incentivos por actividades de Promoción y Prevención (Eventos de PyP), Recobros al sistema de Riesgos profesionales, tránsito y NO POS a FOSYGA. (Universidad Nacional de Colombia, 2011).

La UPC para el régimen Contributivo es una UPC diferencial y se estableció para 56 grupos de riesgos según la edad, género, zona geográfica. Esta UPC diferencial se denomina "Ajuste de la UPC y Ponderación del Riesgo", de acuerdo con esta ponderación el valor de la UPC se ha beneficiado ya que ha tenido un incremento del 9% en los últimos dos años. No se puede considerar una UPC promedio y realizar un incremento sin tener en cuenta estas variables o agrupaciones. Esta forma de calcular una UPC diferencial teniendo en cuenta las tres variables; ha sido diseñado por el ente regulador para evitar la selección adversa⁹ por parte de las aseguradoras. Dado que el enfoque de este modelo no es comercial el valor de la prima no lo define el asegurador si no el ente regulador, tampoco es una aproximación actuarial, si fuese así entonces el asegurador establecería el valor a cobrar según las características del Riesgo. Existe el Riesgo moral¹⁰ lo cual incentiva los recobros al FOSYGA. El valor de la UPC se

⁹ Selección Adversa: "De manera resumida lo podemos describir como la exclusión, por parte del sistema, de aquellos "clientes" que eventualmente representan una mayor probabilidad de gastos y cuyas retribuciones nunca serán -en el aspecto monetario- las suficientes para el pago de lo que el sistema habría "invertido" en ellos" tomado de www.encolombia.com/heraldomed22-22700acerca.htm.

¹⁰ Los individuos tienen una mayor propensión a utilizar los servicios, así este uso no sea necesario. Esta tendencia a consumir más de lo necesario debido al aseguramiento corresponde a lo que, en la literatura, se denomina como riesgo moral. Tomado del Artículo "El Sector Salud en Colombia: Riesgo Moral y Selección Adversa en el Sistema General de Seguridad Social En Salud (SGSSS)" Por Mauricio Santa María S.1, Fabián García A. y Tatiana Vásquez B.

puede estimar por métodos actuariales o también haciendo uso de técnicas de modelización. El ajuste realizado para cálculo de la UPC se realizó teniendo en cuenta variables de Edad, Género, Zona, el sistema de información contiene variables importantes como prestaciones codificadas a través de Códigos Únicos de Prestación de Servicios CUPS y los códigos únicos de Medicamentos CUMS, estas variables permiten establecer enfoque epidemiológicos y así tomar decisiones acertadas o encaminar otro tipos de estudio.

CAPITULO 4

MARCO TEÓRICO ESTADÍSTICO

En este capítulo se relaciona la revisión teórica sobre el modelo estadístico empleado para la obtención de los resultados en el cálculo del valor de la UPC en entidades aseguradoras. El modelo empleado es el modelo de dos partes (o etapas), cuyo funcionamiento básico es el siguiente: En la primera etapa se usa un modelo logístico binario y en la segunda etapa se hace uso de un modelo lineal generalizado, sus estimadores son robustos en comparación con los estimadores obtenidos a partir de los mínimos cuadrados ordinarios (MCO), es compatible con una amplia variedad de distribuciones de probabilidad, que no necesariamente están supeditadas a la distribución de probabilidad Normal entre las cuales se encuentra la familia de distribución Exponencial.

4.1 Modelo de dos partes

La idea general es calcular en primera instancia la probabilidad de que un individuo solicite un servicio dados los factores de riesgo. El segundo paso es calcular el valor esperado del servicio dado que se realiza una reclamación según sus factores de riesgo. Cuando se combinan estos dos modelos en multiplicación se tiene el valor esperado del servicio dado los factores de riesgo. De esta manera es posible calcular el valor esperado de la reclamación para cada factor de riesgo. Los modelos de mixtura o dos partes tienen en cuenta la asimetría en la distribución de costos, la gran proporción de ceros (Afiliados que no incurren en ningún costo) y las colas largas, lo cual va en línea con las recomendaciones dadas en (Mihaylova, Briggs, & O' Hagan, 2010) y de (Manning & Mullahy, 2001) quienes muestran que los estimadores de los modelos lineales generalizados son robustos. La forma general de un modelo de dos partes se puede escribir de la siguiente manera:

$$E[y | \vec{x}] = \Pr(y > 0) \times E[y | y > 0, \vec{x}]$$

Donde y es la variable dependiente de interés (Costo del evento) y $\vec{x}$ es un vector de covariables relacionadas con el evento y/o su costo.

- $\Pr(y > 0)$: Es la probabilidad de incurrir en un gasto dado los factores de riesgo.
- $E[y | y > 0, \vec{x}]$: Es el valor esperado del costo dado los factores de riesgo.

La idea es modelar $\Pr(y > 0, \vec{x})$ y $E[y | y > 0, \vec{x}]$ por aparte. Siendo posible realizar una estimación de los parámetros de ambos modelos en conjunto utilizando el algoritmo EM (expectation–maximization) desarrollado por (Hastie, Tibshirani, & Friedman, 2001); sin embargo estos tres autores sostienen que

resulta conveniente estimar los modelos por aparte dado que hay una ganancia considerable en tiempo y recursos computacionales sin comprometer la exactitud de las estimaciones.

La primera parte $\Pr(y > 0 | \vec{x})$ puede ser estimada mediante un modelo logit o logístico. Para estimar $E[y | y > 0, \vec{x}]$ se puede emplear un modelo lineal generalizado, el cual es una extensión del modelo de mínimos cuadrados ordinarios que dá campo a la existencia de heteroscedasticidad y condición de no normalidad en la respuesta. En estos modelos se asume que $E[y_i | y_i > 0, x_i] = g(x_i)$. La función g se denomina link function (función de enlace)

Para estimar el modelo de dos partes es necesario determinar los factores de riesgo ($\vec{x}$) o (covariables) que afectan el gasto. En la tabla 4.1, que es una adaptación de (Rice & Smith, 2001), se hace explicito cada uno de los factores de riesgo utilizados en 20 países estudiados. Según los autores la morbilidad es la característica Individual que más afecta los costos estimados; sin embargo, puede ser utilizada estratégicamente por las entidades de salud y aseguradoras. En esta tabla se tiene factores a nivel individual como: edad, género, etnia y estado de discapacidad, en cuanto a los factores agregados por ejemplo se realizan por un área geográfica específica, estado socioeconómico entre otros, todo estos factores varían según el país en cuestión.

País	Factores Individuales	Factores Agregados	Factores Adicionales
Australia	Edad, género, grupo etnicidad, vivienda		Flujos a través de fronteras, variación en costos
Canadá	Edad, género, etnicidad, estado de bienestar	Distancia	Flujos a través de fronteras, variación en costos
Dinamarca	Edad	Edad de hijos de padre único	Impuestos locales
Inglaterra	Edad	Mortalidad, morbilidad, desempleo, comparte vivienda (personas tercera edad), etnicidad, estatus socioeconómico	Variabilidad en costos
Finlandia	Edad	Distancia	Impuestos
Francia	Discapacidad		
Alemania	Edad, género		Ingreso
Israel	Edad		Remoción de 5 categorías de enfermedad graves
Italia	Edad, género	Mortalidad	
Japón	Edad		
Holanda	Edad, género, bienestar/discapacidad	Urbanización	Ingreso
Nueva Zelandia	Edad, género, bienestar, etnicidad	Rural	
Irlanda del Norte	Edad, género	Mortalidad, comparte vivienda (personas de tercera edad), bienestar, peso al nacer	Costos rurales
País	Factores Individuales	Factores Agregados	Factores Adicionales

Noruega	Edad, género	Mortalidad	Impuestos
Escocia	Edad, género	Mortalidad	Costos rurales
España	.		Flujos a través de fronteras, población.
Suecia	Edad, comparte vivienda, empleo, Vivienda, diagnostico de salud anteriores		
Suiza	Edad, género, región		Ingreso
Gales	Edad, género	Mortalidad	Ajuste por variabilidad en costos

Tabla No. 4.1: Factores de Riesgo utilizados por Países para Estimar del Gasto Esperado en Salud.

Tomado de: Consultoría realizada por la Fundación Santa Fe de Bogotá, según contrato número 378-2008 de 2008 suscrito con el Instituto Colombiano para el Desarrollo de la Ciencia y la Tecnología-COLCIENCIAS

4.2 Modelo de Regresión Logística

El modelo de Regresión Logística es un caso particular del Modelo Lineal Generalizado (GLM), para el cual la variable respuesta Y es dicotómica con distribución binomial siendo del tipo (Éxito/Fracaso). Este modelo permite estimar la relación existente entre la variable respuesta dicotómica y un conjunto de k predictoras independientes $X' = x_1, x_2, \dots, x_k$ de cualquier naturaleza (Hosmer & Lemeshow, 1989).

Los Objetivos al ajustar este modelo de regresión son:

- Obtener una ecuación que estime la probabilidad de ocurrencia de $P(Y = 1|\vec{x})$ una vez conocidos los valores de $x_1, x_2, \dots, x_k$, siendo un modelo predictivo.
- Cuantificar la relación entre la variable (dependiente o respuesta) Y y el conjunto de variables independientes (predictoras o explicativas) $x_1, x_2, \dots, x_k$.

Para que el modelo proporcione directamente la probabilidad de pertenecer a cada uno de los grupos posibles, se realiza la transformación de la variable respuesta Y con el fin de garantizar que la respuesta a estimar se encuentre contenida en el intervalo $[0 ; 1]$.

Considerando un conjunto de k variables independientes denotados por el vector $\vec{X}' = x_1, x_2, \dots, x_k$, la probabilidad condicional de que el resultado está presente se denota por $P(Y = 1|\vec{x}) = \pi(\vec{x})$ donde $\pi(x)$ se usa con el fin de simplificar la notación, esta cantidad $\pi(\vec{x}) = E(Y|\vec{x})$ representa la media condicional de Y dado $\vec{X}$, el logit del modelo de regresión logística múltiple se da por la ecuación:

$$g(\vec{x}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_k x_k$$

$g(\vec{x})$ resulta de transformar a $\pi(\vec{x})$; es llamada la transformación logit. Esta transformación es fundamental para el modelo de regresión logística ya que tendrá muchas de las características deseables en un modelo de regresión lineal. En cuyo caso el modelo de regresión logística es:

$$\pi(\vec{x}) = \frac{e^{g(\vec{x})}}{1 + e^{g(\vec{x})}}$$

Entonces $\pi(\vec{x}) = P(Y = 1|\vec{X})$ estima la probabilidad de que un individuo con características definidas por el conjunto de $\vec{X}$ pertenezca a la población correspondiente a $Y = 1$. Dicho de otra manera expresa la probabilidad de que ocurra un éxito.

La forma específica del modelo de regresión logística queda expresado como:

$$\pi(\vec{x}) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_k x_k}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_k x_k}}$$

Los coeficientes estimados para las variables independientes representan la pendiente (iesima Velocidad de cambio) de una función de la variable dependiente por unidad de cambio en la variable independiente. Por lo tanto, la interpretación implica determinar la relación funcional entre la variable dependiente y la variable independiente, de esta manera se define adecuadamente la unidad de cambio de la variable independiente.

Se debe determinar qué función de la variable dependiente permite obtener una función lineal de las variables independientes. Esta se llama la función de enlace [ver (McCullagh & Nelder, 1989) o Dobson (1990)] en el caso de un modelo de regresión lineal, es la función Identidad desde la variable dependiente, por definición, es lineal en los parámetros. En el modelo de regresión logística la función de enlace es la transformación logit

$$g(\vec{x}) = \ln \left\{ \frac{\pi(\vec{x})}{[1 - \pi(\vec{x})]} \right\} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_k x_k$$

4.2.1 Estimación e Interpretación de los Coeficientes Modelo de Regresión Logística

Los coeficientes β 's del modelo son estimados a través de procedimientos de máxima verosimilitud, utilizando el método numérico de Mínimos Cuadrados Iterativamente Reponderados (IRLS Iteratively Reweighted Least Squares).

El método máxima verosimilitud enfocado en la estimación de los parámetros desconocidos del modelo de regresión logística, produce valores para los parámetros desconocidos que maximizan la probabilidad de obtener el conjunto de datos observados. Para aplicar este método, primero se debe construir la función de verosimilitud. Esta función expresa la probabilidad de los datos observados como una función de los parámetros desconocidos. Los estimadores de máxima verosimilitud de estos parámetros son elegidos por ser los valores que maximizan esta función. Por lo tanto, los estimadores resultantes son los que tendrán una relación más estrecha con los datos observados.

Para encontrar ese conjunto de estimadores se procede de la siguiente manera:

Para los pares (x_i, y_i) , donde $y_i = 1$, la contribución a la función de verosimilitud es $\pi(x)$, y para aquellos pares donde $y_i = 0$, la contribución a la función de probabilidad es $1 - \pi(x)$. Una forma conveniente de expresar la contribución a la función de probabilidad para el par (x_i, y_i) es a través de la expresión:

$$\pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i}$$

Dado que las n observaciones se supone que son independientes, la función de probabilidad se obtiene como el producto de los términos definidos en la expresión $\pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i}$ de la siguiente manera:

$$l(\beta) = \prod_{i=1}^n \pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i}$$

En este caso, es más fácil matemáticamente trabajar por medio de la expresión del logaritmo de verosimilitud, que se define como:

$$L(\beta) = \ln[l(\beta)] = \sum_{i=1}^n \{y_i \ln[\pi(x_i)] + (1 - y_i) \ln[1 - \pi(x_i)]\}$$

Para encontrar el conjunto de valores de β que maximiza $L(\beta)$ diferenciamos $L(\beta)$ con respecto a β_0 y β_i , las expresiones resultantes se igualan a cero.

$$\frac{\partial \ln(L(\beta))}{\beta_i} = 0$$

Estas ecuaciones, conocidas como las ecuaciones de probabilidad no son lineales en los β , requiere de métodos especiales para su solución. Estos métodos son de naturaleza iterativa y mediante el uso de software programado se obtienen los resultados requeridos.

4.2.2 Estimación por Intervalos e Interpretación para β_i

Al igual que en el caso del modelo lineal general, las estimaciones β pueden ser expresadas en términos de intervalos de confianza $100 * (1 - \alpha)\%$, para el coeficiente β_i es: IC para β_i

$$\langle \beta_i \rangle = \beta_i \pm Z_{\frac{\alpha}{2}} \sqrt{VAR(\hat{\beta}_i)}$$

Expresión en la cual $VAR(\hat{\beta}) = - \left[E \left(\frac{\delta^2 l(\beta; y; X)}{\delta \beta \delta \beta'} \right) \right]^{-1}$ Asumiendo normalidad asintótica de los estimadores Máximo -Verosímiles, la $VAR(\hat{\beta}_i)$.

La regresión utiliza como mecanismo de interpretación la razón de disparidad, el odds ratio, denotado como OR, el cual se define como la relación o razón de las probabilidades de que un suceso ocurra para ($x = 1$) como que no ocurra ($x = 0$) es decir, un número que expresa cuanto más probable es que se produzca frente a que no se produzca el hecho en cuestión, el OR está asociado a los parámetros del modelo. De esta manera es posible lograr una estación de OR siguiendo la expresión:

$$\widehat{OR} = \frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]}$$

Sustituyendo las expresiones para el modelo de regresión logística que se muestran en la tabla 4.2, para el caso de una variable predictora dicotómica.

Variable Resultado Y	Variable independiente (X)	
	$x = 1$	$x = 0$
$y = 1$	$\pi_1 = \frac{e^{(\beta_0 + \beta_i)}}{1 + \exp(\beta_0 + \beta_i)}$	$\pi_0 = \frac{e^{(\beta_0)}}{1 + e^{(\beta_0)}}$
$y = 0$	$1 - \pi_1 = \frac{1}{1 + e^{(\beta_0 + \beta_i)}}$	$1 - \pi_0 = \frac{1}{1 + e^{(\beta_0)}}$

Tabla No. 4.2 Valores regresión logística cuando la variable Independiente es dicotómica

Se obtiene la siguiente relación de probabilidades:

$$\widehat{OR} = \frac{\left(\frac{e^{\beta_0+\beta_i}}{1+e^{\beta_0+\beta_i}}\right) / \left(\frac{1}{1+e^{\beta_0+\beta_i}}\right)}{\left(\frac{e^{\beta_0}}{1+e^{\beta_0}}\right) / \left(\frac{1}{1+e^{\beta_0}}\right)} = \frac{e^{\beta_0+\beta_i}}{e^{\beta_0}} = e^{(\beta_0+\beta_i)-\beta_0} = e^{\beta_i}$$

De esta manera, cuando el Coeficiente β_i de una variable es positivo, se obtiene un valor de $OR > 1$. Por tanto esta variable representa un factor potenciador de la aparición del evento de interés. Análogamente, un coeficiente β_i negativo proporciona un valor de $OR < 1$. Siendo esta variable un factor protector para la ocurrencia del evento.

De aquí se obtiene la relación con los parámetros del modelo para el caso de un modelo de regresión logística simple con la variable independiente dicotómica:

$$\widehat{OR}_i = e^{(\beta_i)}$$

Es claro entonces, que un cambio unitario en el valor de la variable predictora está asociado con un aumento en OddsRatio.

Finalmente al aplicar el logaritmo natural se llega a la estimación del parámetro β_i mediante $\widehat{OR}_i$:

$$\ln(\widehat{OR}_i) = \hat{\beta}_i$$

Cuando un modelo de regresión logística contiene una variable independiente continua, la interpretación del coeficiente estimado depende de la forma en que se introduce en el modelo y las unidades especiales de la variable.

Dependiendo del tipo de estudio que se está realizando; el cambio en unidades para los datos de estudio puede considerarse relevante, muy a menudo el valor de "1" no es clínicamente interesante. Por ejemplo, un aumento de 1 año de edad o un 1 mm Hg de aumento de la presión arterial sistólica puede ser demasiado pequeño para ser considerado importante. Un cambio de 10 años o 10 mm de Hg podría ser considerado más útil. Por otra parte, si el rango de x es de cero a 1, entonces un cambio de 1 es demasiado grande y un cambio de 0.01 podría ser más realista. Por lo tanto, para proporcionar una interpretación útil de las covariables de escala continua se requiere desarrollar un método de estimación puntual y de intervalo para un cambio arbitrario de " c " unidades en la covariable. El log odds ratio para un cambio de c unidades en x se obtiene a partir de la diferencia logit $g(x+c) - g(x) = c\beta_1$ y la razón de probabilidad asociada se obtiene por $OR(c) = OR(x+c, x) = \exp(c\beta_1)$. Una estimación se puede obtener mediante la sustitución de β_1 a través de la máxima verosimilitud estimada de β_1 . Una estimación del error estándar necesario para la estimación del intervalo de confianza se obtiene al multiplicar el error estándar estimado de β_1 por c . De ahí que los criterios de valoración de la estimación $100(1 - \alpha)\%CI$ 100 son:

$$\exp[c\hat{\beta}_1 \pm Z_{1-\alpha/2}c \widehat{SE}(\hat{\beta}_1)] = \exp^{c[\beta_1 \pm Z_{1-\alpha/2}\widehat{SE}(\hat{\beta}_1)]}$$

Dado que tanto la estimación puntual y límites del intervalo de confianza dependen de la elección de c , el valor particular de c debe estar claramente especificado sobretodo presentar significado de interpretabilidad. La naturaleza más bien arbitraria de la elección de c puede ser problemática en algunos casos. Por ejemplo, ¿por qué utilizar un cambio de 10 años, cuando 5 o 15 o incluso 20 años pueden ser igual de buenos? por supuesto, se podría usar cualquier valor razonable; pero el objetivo debe mantenerse en mente: Conocer algunas formas de análisis con una indicación clara de cómo el riesgo del resultado estará presente en los cambios con la variable en cuestión.

(Hosmer & Lemeshow, 1989) Existe un dilema inevitable cuando covariables continuas se modelan de forma lineal en el logit. Cuando se cree que el modelo logit no es lineal en la covariable, la agrupación y el uso de variables dummy se debe considerar.

4.2.3 Pruebas de Bondad de Ajuste y Significancia de los Parámetros

Las siguientes son algunas de las pruebas de Bondad de ajuste que se usan para evaluar el modelo de Regresión Logística.

Devianza D

Según (Hosmer & Lemeshow, 1989), la devianza compara el logaritmo de la verosimilitud del modelo ajustado con el logaritmo de la verosimilitud del modelo saturado es decir el modelo que contiene todas las variables de interés que queramos evaluar y todas las interacciones posibles, para lograr el contraste de las siguientes hipótesis:

H_0 : El modelo ajustado no difiere del modelo saturado vs

H_1 : El modelo ajustado difiere del modelo saturado para mejorarlo

El estadístico de prueba está representado por:

$$D = -2 \sum_{i=1}^n \left[y_i \ln \left(\frac{\hat{\pi}_i}{y_i} \right) + (1 - y_i) \ln \left(\frac{1 - \hat{\pi}_i}{1 - y_i} \right) \right]$$

Donde $\hat{\pi}_i = \hat{\pi}(x_i)$

El Modelo Saturado es el modelo que contiene la mayor cantidad de parámetros como número de "observaciones" en el conjunto de datos, es decir, el tamaño de la muestra.

El Modelo Ajustado comúnmente se refiere al modelo donde se ha realizado una evaluación de la importancia de las variables en el modelo, esto normalmente implica la formulación y prueba de una

hipótesis estadística para determinar si las variables independientes en el modelo son "significativas" en relación con la variable de resultado.

La regla de decisión se construye apoyada en una distribución $X^2_{(n-k)}$ donde k corresponde al número de variables independientes de esta manera se obtiene:

- Si $D \leq X^2_{\alpha, n-k}$ El modelo ajustado no difiere del modelo saturado.
- Si $D > X^2_{\alpha, n-k}$ El modelo ajustado difiere del modelo saturado.

Prueba de Hosmer y Lemeshow

La prueba X^2 de Hosmer y Lemeshow consiste en calcular para cada observación del conjunto de datos, las probabilidades de la variable dependiente, agrupándolas, generalmente en deciles y calculando a partir de las frecuencias esperadas y las compara con las observadas, a través del siguiente estadístico de prueba:

$$\hat{C} = \sum_{j=1}^g \frac{(O_k - n'_j \bar{\pi}_k)^2}{n'_j \bar{\pi}_k (1 - \bar{\pi}_k)}$$

Donde n'_j es el número total de sujetos en el grupo de orden j *esimo*, c_k denota el número de patrones en las covariables en el decil k *esimo*, g corresponde a los grupos.

$$O_k = \sum_{i=1}^k y_i$$

Es el número de respuestas entre los patrones en las covariables c_k , y

$$\bar{\pi}_k = \sum_{j=1}^k \frac{m_j \hat{\pi}_j}{n'_j}$$

Es el promedio de la probabilidad estimada.

Con el uso de un amplio conjunto de simulaciones, Hosmer y Lemeshow (1980) demostraron que, cuando $J = n$, y el modelo de regresión logística ajustada es el modelo correcto, la distribución del estadístico $\hat{C}$ es bien aproximada por la distribución chi-cuadrado con $g - 2$ grados de libertad, $\chi^2(g - 2)$. Si bien no se ha examinado específicamente, es probable que $\chi^2(g - 2)$ también se aproxime a la distribución cuando $J = n$.

Los valores menores al estadístico calculado indican buen ajuste y valores superiores indican falta de ajuste.

Criterio de Información de Akaike AIC

Otra forma de valorar la Bondad de ajuste es sancionar la complejidad de los modelos por falta de parsimonia o sobreparametrización, buscando el modelo que mejor explique los datos con un mínimo de parámetros. Para ello el estadístico de prueba será:

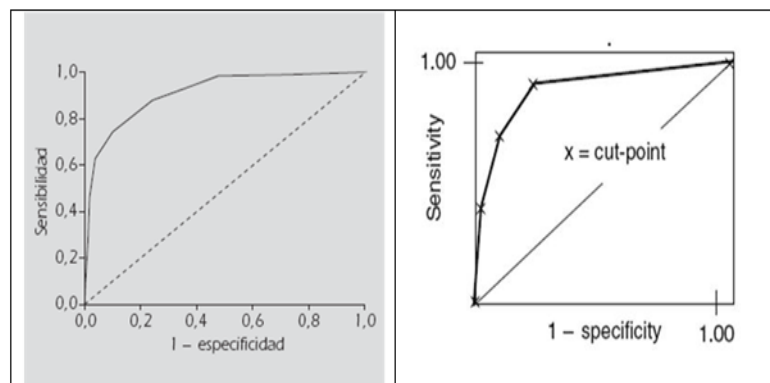
$$AIC = 2(P) - 2\ln(L)$$

Donde P es el número de parámetros estimados incluyendo el intercepto y L es el logaritmo de la verosimilitud alcanzado por el modelo estimado. El criterio de información de Akaike es una medida de la calidad relativa de un modelo estadístico, como tal AIC proporciona un medio para la selección del modelo. El criterio es elegir entre un grupo de modelos aquel que tenga menor AIC, este estadístico sirve para analizar la capacidad explicativa del modelo.

Curvas ROC

En el caso de los modelos predictivos; se cuenta con una alternativa para verificar la adecuación del modelo, el concepto curva ROC (Receiving Operating Characteristic) o (curvas de características operativas para el receptor), nacido de las telecomunicaciones y los problemas asociados a la recepción de señales electrónicas, se aplica frecuentemente en la evaluación de pruebas diagnósticas o de procedimientos de pronósticos. ROC es una representación gráfica de la sensibilidad (Se) vs 1 - especificidad (1 - Sp) derivado de varios puntos de corte para el valor predicho, en el eje de ordenadas se sitúa la sensibilidad (proporción de verdaderos positivos) y en el eje de abscisas el complementario de la especificidad (1 - especificidad o proporción de falsos positivos).

La sensibilidad como tal es la probabilidad de clasificar correctamente un individuo (éxito) es decir; la capacidad del test para detectar éxitos, mientras que la especificidad es la probabilidad de clasificar correctamente un individuo (fracaso) es decir; la capacidad del test para detectar los fracasos; toman valores de 0 a 1 (0 a 100%)



Gráfica No. 4.2 Ejemplo de Curva ROC

Cuando la prueba no tiene ningún poder de discriminación, es decir, se observan los mismos resultados en los individuos (éxito) que en los que no presentan (fracaso), la curva ROC está representada por la diagonal principal del gráfico (área bajo la curva igual a 0,5). El mejor punto de corte es el más próximo al ángulo superior izquierdo del gráfico. Como norma general, si el área bajo la curva es mayor de 0,9 se considera que la prueba es muy exacta, mientras que valores comprendidos entre 0,7 y 0,9 indican una exactitud moderada. Los valores comprendidos entre 0,5 y 0,7 se corresponden con una exactitud baja. En definitiva, cuanto más próxima es una curva ROC a la esquina superior izquierda, más alta es la exactitud global de la prueba.

Test de Wald (Significancia de los parámetros)

Cuando se utiliza la Estimación Máxima Verosímil, se puede llevar a cabo la prueba de hipótesis de significancia de los parámetros mediante el uso de uno de dos procedimientos, la prueba de razón de verosimilitud ó la prueba de Wald.

La significancia de las variables en el modelo de regresión logístico se evalúa con el siguiente contraste de hipótesis:

$$H_0: \beta_i = 0 \text{ vs } H_a: \beta_i \neq 0$$

$$H_0: OR = 1 \text{ vs } H_a: OR \neq 1$$

El estadístico Wald corresponde al cociente entre el valor de la estimación máximo verosímil $\hat{\beta}$ y la estimación del error estándar correspondiente (Hosmer & Lemeshow, 1989). La distribución de un estimador de máxima verosimilitud es aproximadamente normal, además de ser asintóticamente insesgados con varianza igual a:

$$VAR(\hat{\beta}) = - \left[E \left(\frac{\delta^2 l(\beta; y; X)}{\delta \beta \delta \beta'} \right) \right]^{-1}$$

El estadístico de prueba está dado por:

$$z_0 = \frac{\hat{\beta}_i}{\sqrt{VAR(\hat{\beta}_i)}} \sim N(0,1)$$

Si $|z_0| \leq z_{1-\frac{\alpha}{2}}$ Existe suficiente evidencia para rechazar H_0 .

Si $|z_0| > z_{1-\frac{\alpha}{2}}$ No Existe evidencia suficiente para rechazar H_0 .

Razón de Verosimilitud G

Consiste en comparar cada modelo cuando se adicionan o salen variables en el proceso de selección de las mismas para definir el modelo ajustado, este evalúa si la covariable debe ser incluida (removida) o no el modelo. Mide hasta qué punto el modelo se ajusta bien a los datos $H_0 = \beta_i = 0$ vs $H_a = \beta_i \neq 0$. El estadístico de prueba está representado por:

$$G = -2 \ln \left[\frac{\text{Verosimilitud del Modelo sin } X_i}{\text{Verosimilitud del Modelo con } X_i} \right]$$

- Si $G \leq \chi^2_{1-\alpha,1}$ No existe suficiente evidencia para rechazar H_0 .
- Si $G > \chi^2_{1-\alpha,1}$ Existe evidencia suficiente para rechazar H_0 .

Es el método de estimación y evaluación de la significancia del modelo vía máxima verosimilitud más adecuado en términos prácticos.

4.2.4 Supuestos del Modelo Logístico

Algunos supuestos en el modelo de regresión logístico son:

- Se requiere la existencia de una relación lineal entre $\text{logit}(x)$ y los predictores $x_1, x_2, \dots, x_k$.
- Al igual que con otras formas de regresión, la multicolinealidad entre los predictores puede llevar a estimaciones sesgadas y a errores típicos inflados.
- Para que la regresión logística tenga un sentido claro, las variables independientes deben tener una relación monótona con la probabilidad del evento de interés (Silva A., 2008).

4.2.5 Errores del Modelo

Los errores del modelo no son normales debido a la naturaleza de la variable de respuesta que es binaria, por tanto estos serán de la forma:

- Si $Y = 1$, los errores serán $\varepsilon = 1 - \pi(x)$
- Si $Y = 0$, los errores serán $\varepsilon = -\pi(x)$

En la distribución de los errores se tiene que $E(\varepsilon) = 0$ y $Var(\varepsilon) = \pi(x)(1 - \pi(x))$

4.2.6 Selección de Variables

Un punto que despierta natural interés es el concerniente a la “selección de variables en el modelo”. Si inicialmente se incluyen k variables, es posible que no todas sean relevantes para el problema. En tal caso, resulta conveniente eliminar las que no lo sean y ajustar un modelo más simple (o más “parsimonioso”), la ventaja de esta táctica es que cuanto menor sea el número de parámetros ($p + 1$) para estimar, menores serán los errores de muestreo inherentes al proceso de estimación, existen varios procedimientos para eliminar las variables superfluas. El más usado es la regresión paso a paso (step wise regresión) que, a grandes rasgos, consiste en construir sucesivos modelos de manera que cada uno difiera del precedente en una sola variable e ir comparando los resultados de cada versión con los de la anterior a través del indicador de la razón de verosimilitud.

Cualquier procedimiento por etapas para la selección o supresión de variables de un modelo se basa en un algoritmo estadístico que comprueba la "importancia" de las variables, y de esta manera las incluye o excluye del modelo, sobre la base de una regla de decisión fija.

La "importancia" de una variable se define en términos de una medida de la significación estadística del coeficiente de la variable. La estadística que se usa depende de los supuestos del modelo. En etapas de regresión lineal se utiliza una prueba F, puesto que se supone que los errores que se distribuye normalmente. En la regresión logística se supone que los errores que siguen una distribución Binomial, y la importancia se evalúa a través de la prueba de chi-cuadrado de razón de verosimilitud.

Para ver en detalle cada uno de los pasos para la selección de variables en el modelo de regresión logística ver (Hosmer y Lemeshow pág. 116)

4.3 Modelo Lineal Generalizado

Los modelos lineales generalizados pueden ser vistos como una extensión de los modelos lineales clásicos. Un modelo lineal clásico se puede resumir de la siguiente forma: Los componentes de Y son variables normales independientes con varianza constante σ^2 y

$$E(Y) = X\beta + \varepsilon \quad (4.3)$$

Donde X corresponde a la matriz de variables explicativas de Y , β es el vector de parámetros que relaciona Y con X y ε son los errores aleatorios normales de media cero y homocedásticos. El modelo lineal clásico ha demostrado su gran utilidad, sin embargo existen situaciones en las cuales la normalidad de los errores no puede ser garantizada. Para modelar este tipo de situaciones se presenta una generalización de este modelo, la cual fue propuesta por (Nelder & Wedderburn, 1972), cuya característica principal es la posibilidad de modelar variables con distribuciones asimétricas.

Para realizar la generalización se reorganiza ligeramente la ecuación (4.3) y se produce la siguiente especificación compuesta de tres partes:

1. **El componente aleatorio:** los componentes de Y tienen distribuciones normales independientes con $E(Y) = \mu$ y varianza constante σ^2 .

2. **El componente sistemático:** Las covariables $x_1, x_2, \dots, x_k$ producen un predictor lineal η dado por

$$\eta = \sum_{j=1}^k x_j \beta_j;$$

3. **El enlace entre los componentes aleatorios y sistemáticos es:**

$$\mu = \eta$$

Esta generalización introduce un nuevo símbolo η para el predictor lineal y la tercera componente entonces, especifica que μ y η son, de hecho, idénticas. Si escribimos

$$\eta = g(\mu)$$

Entonces $g(\cdot)$ será llamada la función de enlace. En esta formulación, los modelos lineales clásicos tienen una distribución normal (o Gaussiana) en el componente 1 y la función de la identidad para el vínculo en el componente 3.

Los Modelos Lineales Generalizados permiten dos extensiones; primero que la distribución en el componente 1 puede provenir de la familia exponencial, y en segundo lugar la función de enlace en el componente 3 puede convertirse en cualquier función diferenciable monótona.

4.3.1 Funciones de probabilidad para los modelos lineales generalizados

En este modelo se supone que cada componente de Y tiene una distribución que pertenece a la familia exponencial, lo cual significa que su función de densidad puede ser expresada como:

$$f(y; \theta, \phi) = \exp\{(y\theta - b(\theta))/a(\theta) + c(y, \phi)\} \quad (4.4)$$

Para algunas funciones específicas $a(\cdot)$, $b(\cdot)$ y $c(\cdot)$. Si se conoce ϕ , este es un modelo de la familia exponencial con el parámetro canónico θ . Puede o no puede ser de dos parámetros de la familia exponencial si ϕ es desconocido.

Por ejemplo, para la distribución Normal

$$f(y; \theta, \phi) = \frac{1}{\sqrt{(2\pi\sigma^2)}} \exp\{-(y - \mu)^2/2\sigma^2\}$$

$$= \exp\left\{(y\mu - \mu^2/2)/\sigma^2 - \frac{1}{2}(y^2/\sigma^2 + \log(2\pi\sigma^2))\right\}$$

haciendo $\theta = \mu$, $\phi = \sigma^2$, entonces puede expresarse

$$a(\phi) = \phi, \quad b(\theta) = \frac{\theta^2}{2}, \quad c(y, \phi) = -\frac{1}{2}\{y^2/\sigma^2 + \log(2\pi\sigma^2)\}$$

Se escribe $l(\theta, \phi; y) = \log f_Y(y; \theta, \phi)$ la función de log-verosimilitud considerada como una función de θ ; ϕ , y . La media y la varianza de Y se pueden derivar de las relaciones conocidas

$$E\left(\frac{\partial l}{\partial \theta}\right) = 0 \quad (4.5)$$

$$y \quad E\left(\frac{\partial^2 l}{\partial \theta^2}\right) + E\left(\frac{\partial l}{\partial \theta}\right) = 0 \quad (4.6)$$

de la ecuación (4.4) se tiene que

$$l(\theta; y) = \{y\theta - b(\theta)\}/a(\phi) + c(y, \phi)$$

de donde

$$\left(\frac{\partial l}{\partial \theta}\right) = \{y - b'(\theta)\}/a(\phi) \quad (4.7)$$

y

$$\left(\frac{\partial^2 l}{\partial \theta^2}\right) = -b''(\theta)/a(\phi) \quad (4.8)$$

Donde '' (primas) denotan la diferenciación con respecto a θ , de las ecuaciones (4.5) y (4.7) se tiene

$$0 = E\left(\frac{\partial l}{\partial \theta}\right) = \{\mu - b'(\theta)\}/a(\phi)$$

de modo que

$$E(Y) = \mu = b'(\theta)$$

Similarmente de la ecuación (4.6) y (4.8) se tiene

$$0 = \frac{b''(\theta)}{a(\phi)} + \frac{\text{var}(Y)}{a^2(\phi)}$$

de modo que

$$\text{var}(Y) = b''(\theta)a(\phi)$$

Así, la varianza de Y es el producto de dos funciones; uno, $b''(\theta)$, depende del parámetro canónico (y por lo tanto en la media) y sólo se llama la función de la varianza, mientras que el otro es independiente de θ y depende sólo de ϕ .

Para distribuciones diferentes a la normal, el modelo lineal generalizado puede ser ajustado mediante la especificación correcta de la función de enlace. Un referente se presenta en la tabla 4.3

	Normal	Poisson	Binomial	Gamma	Inversa de Gauss
Notación	$N(\mu, \sigma^2)$	$P(\mu)$	$B(m, \pi)/m$	$G(\mu, \nu)$	$IG(\mu, \sigma^2)$
Rango de y	$-\infty, \infty$	$0(1)\infty$	$\frac{0(1)m}{m}$	$(0, \infty)$	$(0, \infty)$
Parámetro de Dispersión ϕ	$\phi = \sigma^2$	1	$1/m$	$\phi = \nu^{-1}$	$\phi = \sigma^2$
Función Cumulant	$\theta^2/2$	$\exp(\theta)$	$\log(1 + e^\theta)$	$-\log(-\theta)$	$-(2\theta)^{1/2}$
$c(y; \phi)$	$-\frac{1}{2} \left(\frac{y^2}{2} + \log(2\pi\phi) \right)$	$-\log y!$	$\log \binom{m}{my}$	$\nu \log(\nu y) - \log y - \log \Gamma(\nu)$	$-\frac{1}{2} \left(\log(2\pi\phi y^3) + \frac{1}{\phi y} \right)$
$\mu(\theta) = E(Y; \theta)$	θ	$\exp(\theta)$	$e^\theta / (1 + e^\theta)$	$-1/\theta$	$(-2\theta)^{-1/2}$
Función de enlace	$\mu = \eta = X\beta$	$\eta = \text{Log}(\mu) = \text{Log}(X\beta)$	$\frac{e^{\beta_0 + \beta_1 x + \dots + \beta_k k}}{1 + e^{\beta_0 + \beta_1 x + \dots + \beta_k k}}$	$1/\mu$	$1/\mu^2$
Función de varianza $V(\mu)$	1	μ	$\mu(1 - \mu)$	μ^2	μ^3

Tabla No. 4.3 Distribuciones más importantes de la familia Exponencial

Fuente: Generalized linear Models (P. McCullagh and J.A Nelder FRS) Pag 30

- El parámetro de valor medio se denota por μ , o por π para la distribución Binomial.
- La parametrización de la distribución gamma es tal que su varianza es μ^2/ν para $u > 0, \nu > 0$

La función de enlace se refiere al predictor lineal η para el valor esperado μ de un punto de referencia y . En los modelos lineales clásicos la media y el predictor lineal son idénticos, y el enlace de la identidad es posible que tanto η y μ pueden tomar cualquier valor en la recta real.

4.3.2 Componente sistemático (predictor lineal)

Entre los aspectos del predictor lineal que se produce en todos los modelos lineales generalizados están las covariables, $x_1, \dots, x_k$ que pueden ser mediciones continuas, vectores de incidencia de factores cualitativos, o vectores de incidencia de las interacciones entre éstos. Para la especificación y ajuste de los modelos lineales generalizados se debe realizar una descripción concisa y construcción automática de tales vectores.

En cuanto a las covariables cuantitativas comprenden covariables tales como la masa, temperatura, tiempo, cantidad de fertilizante o de drogas, la concentración de un soluto y así sucesivamente, que puede tomar los valores en una escala continua. Los modelos que contienen sólo los términos con covariables continuas a menudo se llaman los modelos de regresión, se deben contrastar con análisis los modelos de la varianza, que sólo tienen términos que implican factores cualitativos.

La linealidad en el presente contexto; significa linealidad del predictor lineal η en los parámetros. En consecuencia una covariable x continua en un término del modelo puede ser sustituida por una función arbitraria $g(x)$. En particular, se puede utilizar $x^2, x^3, \dots$, Además de x para construir un polinomio en x , sin destruir la linealidad. Del mismo modo, el modelo lineal, $\beta_0 + \beta_1 x_1 + \beta_2 x_2$, puede ser ampliado para incluir el término producto $\beta_{12} x_1 x_2$, produciendo una relación bilineal. Si los términos se reorganizan en forma

$$(\beta_0 + \beta_2 x_2) + (\beta_1 + \beta_{12} x_2) x_1$$

La cual muestra una relación lineal en x_1 , que tanto en la pendiente como en la intersección son funciones lineales de x_2 . El reordenamiento alternativo expresa la bilinealidad en forma complementaria

$$(\beta_0 + \beta_1 x_1) + (\beta_2 + \beta_{12} x_1) x_2$$

Una función como $\exp(\gamma x)$, produce un modelo no lineal a menos que γ se conozca a priori. Si γ es desconocida, el modelo no es lineal, y se requiere alguna técnica de optimización no lineal para minimizar la función de discrepancia.

Con respecto a las covariables cualitativas es el conjunto de observaciones frecuentemente indexadas por uno o más factores de clasificación, cada factor tiene un índice asociado, cuyos datos están divididos en grupos disjuntos o clases.

Un factor puede tener sólo un conjunto limitado de valores posibles, que se llamará niveles. Los h niveles siempre pueden ser codificados utilizando los enteros $1, 2, \dots, h$, aunque la codificación $0, 1, \dots, h-1$ es a veces más conveniente. Tal codificación define los niveles formales de un factor.

El término más simple en un predictor lineal generado por un factor es un componente de la intersección. Al Considerar la posibilidad de un modelo con una covariable x y predictor lineal

$$\eta = \alpha + \beta x$$

Si A es un factor con índice i , entonces el predictor lineal extendido podría convertirse

$$\eta = \alpha_i + \beta x$$

lo que implica una intercepción por separado para cada nivel de A , pero una pendiente común β , que es asumido constante en los niveles del factor. Si un factor tiene niveles numéricos, también podría

tratarse como una covariable cuantitativa que tiene sólo unos pocos valores distintos. Si se trata como un factor, encajamos un efecto distinto para cada nivel de forma no estructurada, mientras que si se trata como una variable cuantitativa, se impone una forma lineal de la respuesta. Otra forma quizás la más recomendable es utilizar polinomios en los niveles reales para detectar desviaciones de la linealidad.

Variables dummy

Las variables dummy o ficticias son variables cualitativas, también conocidas como indicativas, binarias, categóricas y dicotómicas. Sólo pueden asumir los valores 0 y 1, indicando respectivamente ausencia o presencia de una cualidad o atributo. Las variables dummy o indicatoras sirven para identificar categorías o clase a las que pertenecen las observaciones. Si i es el índice para los niveles de factor cualitativo A con h niveles, el término A_i puede ser escrito en notación vectorial como:

$$\alpha_1 U_1 + \alpha_2 U_2 + \cdots + \alpha_h U_h$$

Donde los U_j son variables aleatorias ficticias cuyos componentes tomarán el valor 1 si la unidad tiene el factor A en el nivel j , y cero en caso contrario. También se utilizan la incidencia vector de términos y el indicador del vector. Por lo tanto si $h = 3$ y los niveles formales para cinco observaciones son 1, 2, 2, 3, 3, las variables aleatorias ficticias (U_1, U_2, U_3) toman valores de la siguiente manera:

Unidad	A	U_1	U_2	U_3
1	1	1	0	0
2	2	0	1	0
3	2	0	1	0
4	3	0	0	1
5	3	0	0	1

Teniendo en cuenta que

$$U_1 + U_2 + U_3 \equiv 1$$

Independientemente de la asignación de los niveles a las unidades, la constante vectorial 1, es la variable aleatoria ficticia (Dummy) correspondiente al término de intersección, a menudo escrito como μ , en el predictor lineal.

Un término compuesto tal como $(\alpha\beta)_{ij}$ tiene variables aleatorias ficticias, $(uv)_{ij}$, cuyos valores son los productos de componentes de interfaz de u_i y v_j , las variables aleatorias ficticias para A y B como términos de factor único. Se deduce entonces que

$$\sum_i (uv)_{ij} = v_j \quad y \quad \sum_j (uv)_{ij} = u_i$$

Independientemente de la asignación de los niveles a las unidades de los factores, los efectos principales están intrínsecamente relacionados con las interacciones en las que se incluyen.

Términos mixtos

Son aquellos en el que la intersección varía con el nivel de factor, pero donde la pendiente es constante en los niveles. A veces la pendiente puede también cambiar con el nivel de factor, lo que requiere el término βx para ser sustituido por $\beta_i x$. Condiciones en el predictor lineal en la que una pendiente o coeficiente de regresión cambia con el nivel de uno o más factores se denominan mixtas, debido a que incluyen aspectos de ambas covariables continuas y cualitativas.

Las variables dummy para términos mixtos adoptan la misma forma que los factores, excepto que los 1s se sustituyen por los correspondientes valores de x. por escrito como (u1, u2 , u3), toman valores de la siguiente manera:

Unidad	A	X	u ₁	u ₂	u ₃
1	1	1	1	0	0
2	2	3	0	3	0
3	2	5	0	5	0
4	3	7	0	0	7
5	3	9	0	0	9

Aquí $U_1 + U_2 + U_3 = x$

Estimación de los parámetros por medio de la máxima verosimilitud según (Dobson & Barnett, 2008).

Teniendo en cuenta lo mencionado en el punto 4.3.1 y dado que la notación cambia en relación a los autores; entonces

$$l = l_i \quad y$$

$$f(y; \theta; \phi) = f(y_i; \theta).$$

Para estimar los parámetros β que están relacionados con Y_i a través de $E(Y_i) = \mu_i$ y $g(\mu_i) = x_i^T \beta$. Para cada Y_i , la función de log-verosimilitud es:

$$l_i = y_i b(\theta_i) + c(\theta_i) + d(y_i)$$

Donde las funciones de b, c y d se definen por medio de las funciones de la familia de distribución exponencial, para lo cual se debe considerar la posibilidad de una sola variable Y aleatoria cuya distribución de probabilidad depende de un único parámetro θ . La distribución pertenece a la familia exponencial si se puede escribir en la forma:

$$f(y_i; \theta) = s(y)t(\theta)e^{a(y)b(\theta)}$$

Donde a, b, s y t son funciones conocidas. Observe la simetría entre Y y θ . La ecuación anterior se valida si se reescribe como: para lo cual $a(y) = a(\phi)$

$$f(y; \theta) = \exp[a(y)b(\theta) + c(\theta) + d(y)]$$

Donde $s(y) = \exp d(y)$ y $t(\theta) = \exp c(\theta)$

Si $a(y) = y$, la distribución se dice que está en forma canónica (es decir, normal) y $b(\theta)$ a veces se llama el parámetro natural de la distribución.

Si hay otros parámetros, además del parámetro de interés θ , se consideran como parámetros de perturbación formando partes de las funciones de a, b, c y d , y se tratan como si se conociesen. Muchas distribuciones conocidas pertenecen a la familia exponencial.

Teniendo en cuenta lo anteriormente mencionado donde las funciones de b, c y d se definen por medio $f(y; \theta) = \exp[a(y)b(\theta) + c(\theta) + d(y)]$, También

$$E(Y_i) = \mu_i = c'(\theta_i)/b'(\theta_i)$$

$$\text{var}(Y_i) = [b''(\theta_i)c'(\theta_i) - c''(\theta_i)b'(\theta_i)/b'(\theta_i)]^3 \text{ y}$$

$$g(\mu_i) = x_i^T \beta = \eta_i$$

Donde x_i es un vector con elementos $x_{ij}, j = 1, \dots, k$. La función de log-verosimilitud para todo los Y_i 's es

$$l = \sum_{i=1}^n l_i = \sum_{i=1}^n y_i b(\theta_i) + \sum_{i=1}^n c(\theta_i) + \sum_{i=1}^n d(y_i)$$

Para obtener el estimador de máxima verosimilitud para el parámetro β_j se necesita

$$\frac{\partial l}{\partial \beta_j} = U_j = \sum_{i=1}^n \left[\frac{\partial l}{\partial \beta_j} \right] = \sum_{i=1}^n \left[\frac{\partial l_i}{\partial \theta_i} \cdot \frac{\partial \theta_i}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \beta_j} \right] = \vec{0}$$

Utilizando la regla de la cadena para la diferenciación. Se considera cada término en el lado derecho de la ecuación anterior por separado. Primero

$$\frac{\partial l_i}{\partial \theta_i} = y_i b'(\theta_i) + c'(\theta_i) = b'(\theta_i)(y_i - \mu_i)$$

Mediante la diferenciación de $l_i = y_i b(\theta_i) + c(\theta_i) + d(y_i)$ y sustituyendo $E(Y_i) = \mu_i = c'(\theta_i)/b'(\theta_i)$. Luego

$$\frac{\partial \theta_i}{\partial \mu_i} = 1 / \frac{\partial \mu_i}{\partial \theta_i}$$

La diferenciación de $E(Y_i) = \mu_i = c'(\theta_i)/b'(\theta_i)$ da

$$\frac{\partial \mu_i}{\partial \theta_i} = \frac{-c''(\theta_i)}{b'(\theta_i)} + \frac{c'(\theta_i)b''(\theta_i)}{[b'(\theta_i)]^2} = b'(\theta_i)var(y_i)$$

A partir de $var(Y_i) = [b''(\theta_i)c'(\theta_i) - c''(\theta_i)b'(\theta_i)/b'(\theta_i)]^3$. Por último, a partir de $g(\mu_i) = x_i^T \beta = \eta_i$ se tiene:

$$\frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i} x_{ij}$$

Por lo tanto el resultado, dado en

$$\frac{\partial l}{\partial \beta_j} = U_j = \sum_{i=1}^n \left[\frac{\partial l}{\partial \beta_j} \right] = \sum_{i=1}^n \left[\frac{\partial l_i}{\partial \theta_i} \cdot \frac{\partial \theta_i}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \beta_j} \right], \text{ es}$$

$$U_j = \sum_{i=1}^n \left[\frac{(y_i - \mu_i)}{var(Y_i)} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \right]$$

La matriz de varianza-covarianza de los U_j 's tiene términos:

$$\mathfrak{I}_{jk} = E[U_j, U_k]$$

Que constituyen la matriz de información $\mathfrak{I}$. de U_j

$$\begin{aligned}\mathfrak{J}_{jk} &= E \left\{ \sum_{i=1}^n \left[\frac{(y_i - \mu_i)}{\text{var}(Y_i)} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \right], \sum_{l=1}^K \left[\frac{(y_l - \mu_l)}{\text{var}(Y_l)} x_{lk} \left(\frac{\partial \mu_l}{\partial \eta_l} \right) \right] \right\} \\ &= \sum_{i=1}^n \frac{[(y_i - \mu_i)^2] x_{ij} x_{ik}}{[\text{var}(Y_i)]^2} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2\end{aligned}$$

Dado $E[(y_i - \mu_i)(y_l - \mu_l)] = 0$ for $i \neq l$ siendo los Y_i 's independientes y usando

$E[(y_i - \mu_i)^2] = \text{var}(Y_i)$, por lo cual la expresión $\sum_{i=1}^n \frac{[(y_i - \mu_i)^2] x_{ij} x_{ik}}{[\text{var}(Y_i)]^2} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2$ se puede simplificar a

$$\mathfrak{J}_{jk} = \sum_{i=1}^n \frac{x_{ij} x_{ik}}{\text{var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2$$

La estimación de la ecuación $\theta^{(m)} = \theta^{(m-1)} + \frac{U^{(m-1)}}{\mathfrak{J}^{(m-1)}}$ por el método de puntuación se generaliza a:

$$b^{(m)} = b^{(m-1)} + [\mathfrak{J}^{(m-1)}]^{-1} U^{(m-1)}$$

Donde $b^{(m)}$ es el vector de estimaciones de los parámetros $\beta_1, \dots, \beta_p$ en la iteración m th.

En la ecuación $b^{(m)} = b^{(m-1)} + [\mathfrak{J}^{(m-1)}]^{-1} U^{(m-1)}$, $[\mathfrak{J}^{(m-1)}]^{-1}$ es la inversa de la

Matriz de información con elementos $\mathfrak{J}_{jk}$ dado por $\mathfrak{J}_{jk} = \sum_{i=1}^n \frac{x_{ij} x_{ik}}{\text{var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2$ y

$U^{(m-1)}$ Es el vector de los elementos dados por $b^{(m)} = b^{(m-1)} + [\mathfrak{J}^{(m-1)}]^{-1} U^{(m-1)}$ todos evaluado en $b^{(m-1)}$. Si ambos lados de la ecuación $b^{(m)} = b^{(m-1)} + [\mathfrak{J}^{(m-1)}]^{-1} U^{(m-1)}$ se multiplican por $[\mathfrak{J}^{(m-1)}]^{-1}$ se obtiene

$$\mathfrak{J}^{(m-1)} b^{(m)} = \mathfrak{J}^{(m-1)} b^{(m-1)} + U^{(m-1)}$$

De la ecuación que se puede escribir como $\mathfrak{J}_{jk} = \sum_{i=1}^n \frac{x_{ij} x_{ik}}{\text{var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2$ se obtiene

$$\mathfrak{J} = X^T W X$$

donde W es la matriz diagonal $N \times N$ con elementos

$$w_{ii} = \frac{1}{var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2$$

La expresión en el lado derecho de $\mathfrak{J}^{(m-1)} \mathbf{b}^m = \mathfrak{J}^{(m-1)} \mathbf{b}^{(m-1)} + U^{(m-1)}$ es el vector con elementos

$$\sum_{k=1}^p \sum_{i=1}^n \frac{x_{ij} x_{ik}}{var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 b_k^{(m-1)} + \sum_{i=1}^n \frac{(y_i - \mu_i) x_{ij}}{var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)$$

La cual es Evaluada en $\mathbf{b}^{(m-1)}$; de lo que se genera las ecuaciones:

$$\mathfrak{J}_{jk} = \sum_{i=1}^N \frac{x_{ij} x_{ik}}{var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \text{ y } U_j = \sum_{i=1}^N \left[\frac{(y_i - \mu_i)}{var(Y_i)} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \right]. \text{ Así, el lado derecho de la}$$

Ecuación $\mathfrak{J}^{(m-1)} \mathbf{b}^m = \mathfrak{J}^{(m-1)} \mathbf{b}^{(m-1)} + U^{(m-1)}$ se puede escribir como:

$$\mathbf{X}^T \mathbf{W} \mathbf{z}$$

donde $\mathbf{z}$ tiene elementos

$$z_i = \sum_{k=1}^p x_{ik} b_k^{(m-1)} + (y_i - \mu_i) \left(\frac{\partial \eta_i}{\partial \mu_i} \right)$$

Con μ_i y $\partial \eta_i / \partial \mu_i$ evaluado en $\mathbf{b}^{(m-1)}$

Por lo tanto la ecuación iterativa $\mathfrak{J}^{(m-1)} \mathbf{b}^m = \mathfrak{J}^{(m-1)} \mathbf{b}^{(m-1)} + U^{(m-1)}$, se puede escribir como:

$$\mathbf{X}^T \mathbf{W} \mathbf{X} \mathbf{b}^{(m)} = \mathbf{X}^T \mathbf{W} \mathbf{z}$$

4.4 La Distribución gamma

La distribución Gamma es una distribución adecuada para modelizar el comportamiento de variables aleatorias continuas positivas con asimetría positiva; es decir, variables que presentan una mayor densidad de sucesos a la izquierda de la media que a la derecha. En su expresión se encuentran dos parámetros, siempre positivos, α y β de los que depende su forma y alcance por la derecha, y también la función Gamma $\Gamma(\alpha)$ responsable de la convergencia de la distribución.

En cuanto a los parámetros el primer parámetro α sitúa la máxima intensidad de probabilidad y por este motivo en algunas fuentes se denomina “la forma” de la distribución: cuando se toman valores próximos a cero aparece entonces un dibujo muy similar al de la distribución exponencial. Cuando se

toman valores más grandes de α , el centro de la distribución se desplaza a la derecha y va apareciendo la forma de una campana de Gauss con asimetría positiva. Es el segundo parámetro β el que determina la forma o alcance de esta asimetría positiva desplazando la densidad de probabilidad en la cola de la derecha. Para valores elevados de β la distribución acumula más densidad de probabilidad en el extremo derecho de la cola, alargando mucho su dibujo y dispersando la probabilidad a lo largo del plano. Al dispersar la probabilidad la altura máxima de densidad de probabilidad se va reduciendo; de aquí que se le denomine “escala”. Valores más pequeños de β conducen a una Distribución más simétrica y concentrada, con un pico de densidad de probabilidad más elevado.

Los Parámetros de la distribución Gamma pueden expresarse según su escala y su forma:

$$\gamma(v, \lambda): f(y) = \frac{1}{\Gamma(v)} \left(\frac{1}{\lambda}\right)^v y^{v-1} e^{-\frac{y}{\lambda}} \quad \text{donde } y > 0; v > 0 \text{ y } \lambda > 0$$

$$\mu = E(y) = v\lambda \quad ; \quad var(y) = v\lambda^2 = \mu\lambda$$

λ : Parámetro de escala ; v : Parámetro de forma

Otra parametrización mediante el parámetro de forma y la media es:

$$f(y) = \frac{1}{\Gamma(v)} \left(\frac{v}{\mu}\right)^v y^{v-1} e^{-\frac{v}{\mu}y} \quad \text{donde } y > 0; v > 0 \text{ y } \mu > 0$$

$$E(y) = \mu \quad ; \quad var(y) = \frac{\mu^2}{v}$$

La distribución gamma es de tipo exponencial:

$$\begin{aligned} \exp\left\{-\log\Gamma(v) + v\log(v) - v\log(\mu) + (v-1)\log(y) - \frac{v}{\mu}y\right\} = \\ \exp\left\{v\left[y\left(-\frac{1}{\mu}\right) - \log(\mu)\right] + (v-1)\log(y) + v\log(v) - \log\Gamma(v)\right\} \end{aligned}$$

Los componentes de la ecuación son iguales a:

$$v = \frac{1}{\theta} \quad ; \quad \left(-\frac{1}{\mu}\right) = \theta \quad ; \quad \log(\mu) = b(\theta) \text{ por último}$$

$$(v-1)\log(y) + v\log(v) - \log\Gamma(v) = c(y, \phi)$$

Entonces

- $\theta = -\frac{1}{\mu}$ por lo tanto $\mu = -\frac{1}{\theta}$
- $b(\theta) = -\log(-\theta)$
- $a(\phi) = \phi = \frac{1}{v}$

- $V(\mu) = b''(\theta) = \left(\frac{1}{\theta}\right)^2 = \mu^2$
- $Var(y) = a(\phi)b''(\theta) = \frac{1}{v}\mu^2$

Un modelo lineal generalizado gamma estará determinado por el predictor lineal

$$\eta_i = g(\mu_i) = x_i^T \beta$$

Dado que en la distribución gamma $\theta_i = -\frac{1}{\mu_i}$, el enlace canónico es la función recíproca:

$$g(\mu_i) = \frac{1}{\mu_i} = x_i^T \beta$$

Otro enlace habitual considerado en el ajuste de modelos gamma es $\log(\mu_i) = x_i^T \beta$ para lo cual se debe considerar las diferencias entre transformar la respuesta o transformar la media. Suponiendo que se tiene un estimador de un parámetro que tiene la varianza dependiente del mismo por decir $Var_\theta(T) = v(\theta)$ se requerirá de una transformación monótona que estabilice la varianza lo cual hace que la varianza del estimador transformado se constante, una aproximación la proporciona el método delta:

$$Var[h(T)] \approx [h'(\theta)]^2 v(\theta) \propto 1$$

$$h(\theta) \propto \int \frac{1}{\sqrt{v(\theta)}} d\theta$$

Como en el presente modelo la respuesta que se tiene es gamma entonces:

$$y \rightarrow \gamma(v, \mu) \quad ; \quad E(y) = \mu \quad ; \quad Var(y) = \frac{\mu^2}{v}$$

$$h(\mu) \propto \int \frac{1}{\sqrt{\mu}} d\mu = \log(\mu)$$

La densidad gamma se escribe de la forma:

$$\frac{1}{\Gamma} \left(\frac{vy}{\mu}\right)^v \exp\left(-\frac{vy}{\mu}\right) d(\log y); \quad y \geq 0, v > 0, \mu > 0$$

Por razones de brevedad se escribe $Y \sim G(\mu, v)$. Desde su función generadora cumulante, $-v \log(1 - \mu t/v)$, los cuatro primeros cumulantes se encuentran como:

$$\kappa_1 = E(Y) = \mu$$

$$\begin{aligned}\kappa_2 &= \text{var}(Y) = \mu^2/\nu \\ \kappa_3 &= E(Y - \mu)^3 = 2\mu^3/\nu^2 \\ \kappa_4 &= \text{var}(Y) = 6\mu^4/\nu^3\end{aligned}$$

En general, $k_r = (r-1)! \mu^r / \nu^{r-1}$. El valor de ν determina la forma de la distribución. Si $0 < \nu < 1$ la densidad tiene un polo en el origen y disminuye monótonamente a medida $y \rightarrow \infty$. El caso especial de $\nu = 1$ corresponde a la distribución exponencial.

En la parametrización de la distribución gamma la función de la varianza es cuadrática. Este resultado se puede obtener directamente escribiendo el logaritmo de verosimilitud como una función tanto ν y μ en la forma estándar:

$$\nu(-y/\mu - \log \mu) + \nu \log y + \nu \log \nu - \log \Gamma(\nu)$$

De ello se desprende, en términos de la parametrización que $\theta = -1/\mu$ es el parámetro canónico, y $b(\theta) = -\log(-\theta)$ (B) es la función cumulante. De éstos la media $b'(\theta) = \mu$ y la varianza función $b''(\theta) = \mu^2$ la cual se puede derivar.

Para comparar el logaritmo de la verosimilitud del modelo ajustado con el logaritmo de la verosimilitud del modelo saturado se hace uso de la Deviance

Se podría suponer que se dispone de r subpoblaciones independientes, cada una con distribución gamma y con una dispersión común, pero medias posiblemente distintas. De la subpoblación i -ésima se extraen observaciones i.i.d, entonces tomando ν como una constante conocida, el logaritmo de verosimilitud se puede escribir como

$$\sum_{i=1}^r \nu(-y_i/\mu_i - \log \mu_i)$$

para las observaciones independientes. Si el índice no es constante, sino que es proporcional a los pesos conocidos, $V_i = VWI$, el logaritmo de verosimilitud es igual

$$\sum_{i=1}^r w(-y_i/\mu_i - \log \mu_i)$$

La máxima verosimilitud alcanzable ocurre en $\mu = y$, y el valor obtenido es $\sum_i w_i(1 - \log y_i)$ que es finito, a menos $y_i = 0$ para algún i . La desviación, que es proporcional a dos veces la diferencia entre el logaritmo de verosimilitud alcanzado bajo el modelo y el valor máximo alcanzable, es

$$D(y; \hat{\mu}) = -2 \sum_{i=1}^r w_i \{\log(y_i / \hat{\mu}_i)\} - (y_i - \hat{\mu}_i) / \hat{\mu}_i$$

Esta estadística sólo se define si todas las observaciones son estrictamente positivas. En términos más generales, si algunos componentes de y son cero se puede reemplazar $D(y; \mu)$ por

$$D^+(y; \hat{\mu}) = 2C(y) + 2 \sum_{i=1}^r w_i \log \hat{\mu}_i + 2 \sum_{i=1}^r w_i y_i / \hat{\mu}_i$$

Donde $C(y)$ es una función acotada arbitraria de y . La única ventaja de $D(y; \hat{\mu})$ sobre $D^+(y; \hat{\mu})$ es que la primera función es siempre positiva y se comporta como una suma residual de cuadrados. Pero hay que tener en cuenta, que la estimación de máxima probabilidad de ν es una función de $D(y; \hat{\mu})$ y no de $D^+(y; \hat{\mu})$. Además, si cualquiera de los componentes de y es cero, entonces $\hat{\nu} = 0$.

El enlace canónico

La función de enlace canónico produce estadísticas suficientes que son funciones lineales de los datos y se da por:

$$\eta = \log(\mu_i) = x_i^T \beta$$

A diferencia de los enlaces canónicos para la Poisson y distribuciones Binomiales, la transformación recíproca, que a menudo es interpretable como la tasa de un proceso, no asigna el rango de μ en toda la recta real. Por lo tanto la exigencia de que $\eta > 0$ implica restricciones a los β en cualquier modelo lineal. Se debe tener precauciones en el cálculo de β , de modo que se eviten los valores negativos de μ .

Selección de las covariables

Además de la elección de la función de enlace y distribución de error, el problema de la modelización se reduce a la búsqueda de uno o más conjuntos parsimoniosos de covariables apropiadas, correspondientes a una matriz X de orden $n \times p$.

Como en otros lugares, es importante que el modelo final o modelos deben tener sentido físico: como mínimo, esto generalmente significa que las interacciones no deben incluirse sin efectos principales ni los términos de polinomios de grado superior sin sus parientes de grado menor.

La selección de un útil conjunto de covariables de un gran conjunto de posibles covariables como para formar un modelo parsimonioso es entonces un ejercicio no trivial. Hay dos problemas estadísticos e informáticos, esta última derivada de la "explosión combinatoria" que se produce cuando todos los posibles subconjuntos de covariables deben ser probados para su inclusión en el modelo

En el lado estadístico, el problema es el de definir el equilibrio necesario entre dos efectos opuestos de la inclusión de un nuevo término en el modelo. El buen efecto puede ser una reducción en la discrepancia entre los datos y los valores ajustados. El efecto negativo es que, a menos que exista un buen conocimiento previo de que la covarianza tiene una influencia no despreciable en la respuesta, la inclusión de la covarianza generalmente complica el modelo y las declaraciones de las conclusiones derivadas de ella. En un extremo, si la adición de una sola covariable reduce la media residual cuadrado, digamos, a un tercio de su valor original no dudamos en incluir en el modelo, sobre todo si el número de grados de libertad del residuo es grande. En el otro extremo, si dicha adición no provoca ninguna reducción, por el principio de la navaja de Occam, la parsimonia gana y nos excluye. En los casos intermedios que causan problemas. Por ejemplo, si hay un gran número de covarianza irrelevante, a continuación, accidentes estadísticos se producen unos pocos falsos positivos que parecen influir en la respuesta.

El F-estadística habitual para la reducción de la desviación o la suma de los cuadrados es la base de la mayoría de los criterios para la selección de covariables. Con el fin de excluir términos irrelevantes el nivel de significación para la aceptación se ha fijado en un nivel bajo, pero no debe ser tan bajo que términos importantes están excluidos de esta manera. Otro enfoque se basa en la idea de proporcionar la mejor predicción de los valores de respuesta sobre un conjunto de valores de covarianza, y otra utiliza un criterio basado en una medida de la información. (Atkinson, 1981), señala que todos estos procedimientos se puede representar (en nuestra notación), a continuación se relacionan algunas formas como casos especiales de minimizar la expresión:

- Selección hacia adelante, con lo que en cada etapa se añade una covariable, la mejor covariable seleccionada debe satisfacer el criterio de selección hasta que no hay más candidatos.
- Eliminación hacia atrás, que se inicia con la serie completa y elimina la peor de las covariables proceso que se realiza una por una hasta que todos los covariables restantes son las necesarias.
- Regresión por pasos según los (Efroymson, 1960), que combina los dos procedimientos anteriores, a raíz de la eliminación hacia atrás por la selección hacia adelante hasta tanto no logran cambiar el modelo.

(Wolstenholme, O'Brien, & Nelder, 1988), Describen una interfaz basada en el conocimiento para el paquete de estadísticas GLIM 3.77, desarrollado con métodos y herramientas de programación lógica. Utiliza una estrategia de selección de modelo que resulta en un árbol de modelos candidatos, con el nodo extremo de cada rama de la formación de un posible modelo parsimonioso. El paso básico en el algoritmo tiene como entrada un núcleo, que contiene términos ya aceptados como sea necesario, y un conjunto de términos libres, cuyo estado es incierto.

También puede suceder que algunas covariables son mucho más complejas de medir que otras, y esto no está permitido en un criterio basado en consideraciones puramente estadísticos.

Prueba de Bondad de Ajuste

Con respecto a l prueba de Bonda de Ajuste se tiene :

La Devianza D Según (Hosmer & Lemeshow, 1989), la devianza compara el logaritmo de la verosimilitud del modelo ajustado con el logaritmo de la verosimilitud del modelo saturado es decir el modelo que contiene todas las variables de interés que queramos evaluar y todas las interacciones posibles, para lograr el contraste de las siguientes hipótesis:

H_0 : El modelo ajustado no difiere del modelo saturado vs

H_1 : El modelo ajustado difiere del modelo saturado para mejorarlo

El estadístico de prueba está representado por:

$$D = -2 \sum_{i=1}^n \left[y_i \ln \left(\frac{\hat{\pi}_i}{y_i} \right) + (1 - y_i) \ln \left(\frac{1 - \hat{\pi}_i}{1 - y_i} \right) \right]$$

Donde $\hat{\pi}_i = \hat{\pi}(x_i)$

El Modelo Saturado es el modelo que contiene la mayor cantidad de parámetros como número de "observaciones" en el conjunto de datos, es decir, el tamaño de la muestra.

El Modelo Ajustado comúnmente se refiere al modelo donde se ha realizado una evaluación de la importancia de las variables en el modelo, esto normalmente implica la formulación y prueba de una hipótesis estadística para determinar si las variables independientes en el modelo son "significativas" en relación con la variable de resultado.

La regla de decisión se construye apoyada en una distribución $X^2_{(n-k)}$ donde k corresponde al número de variables independientes de esta manera se obtiene:

- Si $D \leq X^2_{\alpha, n-k}$ El modelo ajustado no difiere del modelo saturado.
- Si $D > X^2_{\alpha, n-k}$ El modelo ajustado difiere del modelo saturado.

Inspección de los Residuales

Si la desviación se utiliza como una medida de la discrepancia de un modelo generalizado, a continuación, cada unidad contribuye una cantidad d_i , de manera que $\sum d_i = D$ lo cual se define como

$$r_D = \text{signo}(y - \mu) \sqrt{d_i}$$

se pueden definir distintos tipos de residuales para un Modelo Lineal Generalizado entre estos está:

Analisis de Residuales de Pearson

$$r_p = \frac{y - \hat{\mu}}{\sqrt{\widehat{var}(\hat{\mu})}}$$

CAPITULO 5

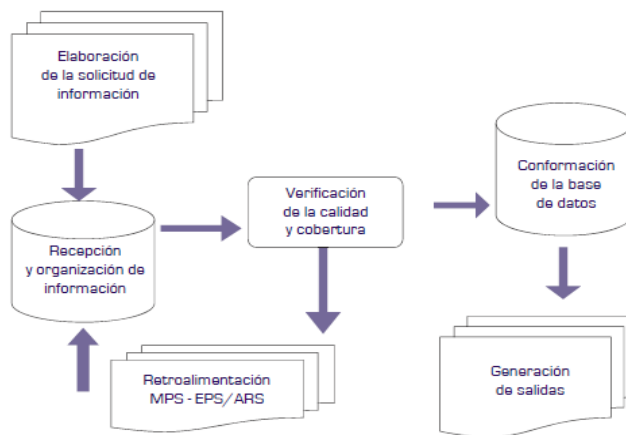
METODOLOGÍA

La metodología a seguir para el alcance de los objetivos planteados está dividida en 2 etapas, la primera enfocada al análisis exploratorio de los datos y la segunda al planteamiento de un Modelo Estadístico para estimar el costo anual del servicio, con la intención de evaluar los factores de influencia en el costo y posteriormente la suficiencia de la UPC para el año 2011.

5.1 Fuentes de Información

La fuente de información corresponde a la base de datos de los afiliados y los servicios de salud prestados para el año 2011. Con el fin de garantizar que se relacione el 100% de los servicios del periodo requerido por parte de las EPS's; el Ministerio de la Protección Social solicita a las EPS's extraer servicios Año 2011 a Facturas Radicadas entre Enero 1 de 2011 a Mayo 31 de 2012. En esta base de datos se registran, una a una, las atenciones solicitadas y su facturación asociada (cada afiliado puede aparecer varias veces). Esta base de datos es un consolidado de múltiples fuentes propias de la EPS, entre ellas la información del Registro Individual de Prestación de Servicios (RIPS), el sistema de facturación, las autorizaciones, los recobros (Tutelas y Comité Técnico Científico).

La Gráfica 5.1 muestra la manera como se gestiona la información del Estudio de Suficiencia POS – UPC Colombia por parte del Ministerio de la Protección Social. Este mismo esquema es adoptado por las EPS's y se hace la gestión hacia los prestadores de servicios de salud como son las IPS, médicos y profesionales de la salud que prestaron los servicios a la población afiliada a la EPS en el periodo correspondiente.



Gráfica No. 5.1: Gestión de información del Estudio de suficiencia POS – UPC. Colombia, año 2008

Fuente: Tomado del estudio de la suficiencia de UPC Año 2009 Minprotección Social.

La población objeto de estudio corresponde entonces a los registros de costos para la fase de Modelación del Costo y la población de Afiliados para la Modelación de la Probabilidad de uso del Servicio, para lo cual se relaciona las estructuras con las variables de población y variables de prestación de servicios del plan de beneficio del Estudio de suficiencia POS –UPC para la generación de la información.

La estructura y variables relacionadas en las fuentes de datos para el estudio de la Suficiencia POS-UPC corresponde a estructura diseñada por el Ministerio de la Protección Social para realizar la solicitud a las EPS's, estas fuentes de información son la bases de datos de los afiliados y base de datos de los servicios de salud prestados en el año 2011 por la EPS. Cada tipo de información requiere un conjunto de variables que se detallan en las Tabla 5.1 y Tabla 5.2 del diccionario de datos del Capítulo 7; de acuerdo con el uso, el indicador o el análisis con el que se relacionan y tipo de información que generan.

Para la afiliada, la base de datos contiene las siguientes variables: Tipo de identificación, identificación, fecha de nacimiento, género, código del departamento, código del municipio, zona geográfica, días compensados, afiliados compensados equivalentes, tipo de afiliado y fecha de afiliación al SGSSS.

Para la base de datos de prestación de servicios del plan de beneficios se dispone de las siguientes variables: tipo de identificación, identificación, fecha de nacimiento, género, código del departamento, código del municipio, zona geográfica, código del diagnóstico, fecha de prestación del servicio, código de actividad, intervención o procedimiento y medicamentos; ámbito de prestación de la actividad, intervención o procedimiento y medicamento; forma de reconocimiento y pago de la actividad, intervención o procedimiento y medicamento; número de días estancia normal; valor actividad, intervención o procedimiento y medicamento y valor asumido por el usuario. La información detallada de la prestación de servicios de salud permitirá obtener las frecuencias y los costos por actividad, intervención o procedimiento del plan de beneficios.

La vigencia de la información corresponde a las atenciones realizadas durante el año 2011, para lo cual el ministerio solicita a las EPS realizar un recorrido en las bases de datos de servicios, desde enero 01 año requerido hasta Mayo o Junio 30 del año siguiente al periodo solicitado, con el fin de asegurar la mayoría de las atenciones dadas por la EPS. Esta forma de solicitar la información se justifica por que los prestadores de Servicios o IPS no alcanzan a facturar a las EPS el 100% de los servicios al cierre del año en referencia.

La información que las EPS's deben reportar para el estudio de suficiencia de la UPC al ministerio de la Protección Social debe estar previamente depurada; sin errores de estructura y contenido, para lo cual el ministerio de la Protección Social establece 7 criterios de calidad que las EPS's deben implementar. A continuación se relacionan los criterios de calidad requeridos por el ente de control.

Calidad 1

Verificación de la estructura de los archivos de reporte que incluye la completitud de los campos y el cumplimiento del formato en la solicitud de información

Calidad 2 y 3

Verificación de la consistencia interna de la información, se relaciona con la revisión de las relaciones entre edad y género; y actividad, intervenciones y procedimientos con género. Verificación cruzada de la información buscando la correspondencia entre las actividades, intervenciones y procedimientos con su ámbito y días de estancia; y en el ámbito ambulatorio alertas en cuanto a valores reconocidos máximos y mínimos.

En la consistencia interna de la información se realiza validaciones cruzadas entre dos o más variables de la solicitud, utilizando dos tablas una para diagnósticos y otra para Actividad, intervención y procedimiento.

- Para la validación del diagnóstico se tiene en cuenta el género, edad mínima y edad máxima, se hacen validaciones cruzadas de la siguiente manera:

Género (M) Masculino, (F) Femenino, (A) Indistinto

Edad: El valor de la edad se calcula en año con la fecha de nacimiento que aparece en el reporte a 31 de diciembre del año de servicio solicitado.

- Actividad, intervención y procedimiento

Para la validación de actividad, intervención y procedimiento se utilizará la tabla CUPS, la cual contiene especificaciones por género, ámbito, estancia.

Calidad 4

Verificación de derechos de los usuarios, cruza los usuarios a los que se les prestaron servicios contra la totalidad de la base de datos de población de la EPS.

Calidad 5

Verificación de atenciones en salud únicas, se debe aplicar a fin de detectar registros iguales por año y día con las variables de identificación y fecha de servicios.

Calidad 6

Verificación de frecuencias y valor de afiliados con frecuencias mayores a 100 actividades anuales y valores de prestación de servicios mayores a \$100.000.000 de pesos

Calidad 7

Verificación de afiliados con el mismo tipo de identificación y número de identificación que aparecen con diferente género o fecha de nacimiento.

Una vez la base de datos cuenta con las respectivas validaciones de calidad, se realiza el reporte al Ministerio de la Protección Social y si el contenido del reporte enviado es considerado satisfactorio será incluido en el cálculo de la Unidad de Pago por Capitación para el país.

5.2 Análisis Exploratorio de los Datos

Con el fin de comprender la estructura de los datos y resaltar algunas características de este conjunto de Cotizantes y Beneficiarios, el primer acercamiento se hará a través del análisis exploratorio, el cual permitirá caracterizar la población afiliada a la EPS en el año 2011 en cuanto a sus condiciones socioeconómicas, perfil de consumo (cantidad y costos de los servicios), así mismo construir perfiles de riesgos epidemiológicos a partir de los diagnósticos, edad, género y zonas geográficas¹¹. Las agrupaciones en cuanto a la atención se realizarán teniendo en cuenta para los diagnósticos la tabla de clasificación internacional de enfermedades CIE 10 y para las prestaciones como son los procedimientos y/o medicamentos según tablas de códigos únicos de prestaciones CUPS y/o código único de medicamentos CUMS, fecha de atención, cantidad de servicios, cantidad de usuarios y costo de las atenciones.

De manera más amplia la intención del análisis exploratorio es detallar las bases de datos y conocer el comportamiento de los afiliados en este sistema de salud. La base de datos de Afiliados consta de 1.107.233 registros que corresponde igualmente al número de individuos únicos afiliados en el rango de tiempo de año 2011. En cuanto a la base de datos de Costos de Servicios el número de registros es de 9.360.283. En este análisis se pretende evaluar diferencias de aquellos individuos que generan atenciones en salud frente a individuos que no tuvieron ninguna atención. En cuanto al análisis demográfico, la caracterización de la base de datos se realizará por género de los afiliados, por ciudad de residencia, grupo de edad. La razón de dicha agrupación es coincidir con la metodología que viene usando el Ministerio de la Protección Social para cubrir los costos de atención en salud y de esta manera realizar las comparaciones correspondientes.

Dentro de este análisis exploratorio también se realizará una caracterización de la morbilidad de la población de Afiliados, a fin de conocer de las enfermedades más frecuentes de la población que consultó los servicios médicos en el año 2011, esto permitirá realizar contrastes entre géneros teniendo en cuenta variables antes mencionadas como la edad o la ciudad, o ámbito de prestación del servicio que dará a conocer que modalidad de atención fue la de mayor demanda, entre estos ámbitos están servicios de Urgencias, atención Ambulatorios y Hospitalaria.

Finalmente se realiza una exploración a la plana de datos respecto a los costos causados por cada atención, para ello las agrupaciones que se pueden proponer son los costos por enfermedad, usuarios más costosos, por ciudades entre otras.

¹¹ Las zonas geográficas son : Normal, Especial y Conurbados, para ver el detalle que ciudades corresponden a cada zona ver Resolución 4480 de 2012

Las herramientas de análisis propuestas serán gráficas de columnas, histogramas así como el uso de tablas resumen que consoliden información por diferentes factores de agrupación, ciudad, por diagnósticos y prestaciones y así de esta manera dar a conocer en un primer enfoque exploratorio la composición de los datos a modelar.

Vale la pena resaltar que para la fase de modelación no resulta conveniente incluir variables relacionadas con la morbilidad del paciente, ya que si bien este es un factor asociado con el costo, no puede convertirse en un discriminante a priori de tarificación, dado que la enfermedad se adquiere durante la exposición.

5.3 Modelación del Costo de Servicio

Para la fase de Modelación se propone el ajuste de un modelo de dos partes (o etapas), cuyo funcionamiento básico postula una primera etapa en la que se usa un modelo logístico binario para la estimación de la Probabilidad de Uso del Servicio, esto debido a que en esta EPS, para el periodo de estudio, un 38% de Afiliados no presentan costo alguno, y en una segunda etapa se hace uso de un modelo lineal generalizado para estimar el valor esperado del costo, según características del Afiliado. La estimación final del costo del servicio se obtiene como el producto de las dos componentes ajustadas previamente.

5.3.1 Modelación de la Probabilidad de Uso

Para Modelar la probabilidad de uso del servicio se tendrán en cuenta todos los afiliados a la EPS en el año 2011, discriminando aquellos que generan costo ($Y > 0$) de aquellos que no ($Y = 0$) o atención médica en el periodo en mención. La técnica estadística a usar en esta componente será el ajuste de un Modelo de Regresión Logística. En el modelo se considerará como predictoras del uso las variables género, zona geográfica y edad, que son las que actualmente contempla el ministerio para la asignación de precios en la UPC. Adicionalmente se consideraran potenciales factores asociados al uso del servicio que se piensan pueden llegar a afectar los costos o la frecuencia del uso tales como Tipo de afiliado, Periodos Compensados.

El Modelo a Ajustar presenta la siguiente estructura:

$$P(Y = 1) = \pi(\vec{x}) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k}}$$

$\pi(\vec{x})$: Representa la probabilidad de que un individuo con perfil $\vec{x}$ haga uso de los servicios médicos durante el año.

En cuanto a la selección de variables existen varios procedimientos para eliminar las variables superfluas. El método a utilizar es la eliminación hacia atrás, que se inicia con la serie completa y elimina la peor de las covariables, proceso que se realiza una por una hasta que todas las covariables restantes son las necesarias. Para evaluar la pertinencia del modelo en cada iteración se realizara el test de razón de verosimilitud, teniendo en cuenta la tabla No 5.3

Evaluación de la capacidad Predictiva del Modelo

	Positivo	Negativo	Total
Positivo	(a)	(b)	(a + b)
Negativo	(c.)	(d)	(c + d)
Total	(a + c)	(b + d)	(a + b + c + d)
Sensibilidad :	a/(a+c).	Espcificidad :	d/ (b+d).

Tabla No. 5.3 Análisis de Sensibilidad y Especificidad capacidad Predictiva del Modelo

La bondad de ajuste del modelo se evaluara a través del Criterio de Información de Akaike AIC, que busca el modelo que mejor explique los datos con un mínimo de parámetros. El estadístico de prueba es:

$$AIC = 2(P) - 2\ln(L)$$

Donde P es el número de parámetros estimados incluyendo el intercepto y L es el logaritmo de la verosimilitud alcanzado por el modelo estimado. El criterio de información de Akaike es una medida de la calidad relativa de un modelo estadístico, como tal AIC proporciona un medio para la selección del modelo. El criterio es elegir entre un grupo de modelos aquel que tenga menor AIC, este estadístico sirve para analizar la capacidad explicativa del modelo.

Como criterio complementario de validación se utilizará la evaluación de la curva ROC asociada al modelo (Receiving Operating Characteristic) o (curvas de características operativas para el receptor) es una prueba más de Bondad de Ajuste, a partir de una representación gráfica de la sensibilidad (Se) vs. 1 - especificidad (1 - Sp), es decir en el eje de ordenadas se sitúa la sensibilidad (proporción de verdaderos positivos) y en el eje de abscisas el complementario de la especificidad (1 – especificidad o proporción de falsos positivos), La sensibilidad como tal es la probabilidad de clasificar correctamente un individuo enfermo es decir; la capacidad del test para detectar la enfermedad, mientras que la especificidad es la probabilidad de clasificar correctamente un individuo sano es decir; la capacidad del test para detectar a los sanos; toman valores de 0 a 1 (0 a 100%)

5.3.2 Modelación Condicional del Costo del Servicio

Para la modelación de los costos de los servicios se utilizará un modelo lineal generalizado. En este caso se filtran de la base de datos global, a los individuos que no presentan costos y se totaliza el costo por año para aquellos que si lo presentan. Este enfoque tiene pocos supuestos y evita la necesidad de transformar los datos, al tiempo que representa una reparametrización del modelo que permite conservar la escala original (en este caso pesos), de la variable de respuesta.

Dado que los datos de Costos son variables aleatorias continuas con asimetría positiva, la función de distribución que mejor modela este tipo de datos es la función de distribución Gamma; por tanto será considerada en la modelación.

El modelo lineal generalizado gamma estará determinado por el predictor lineal

$$\ln(\mu) = \eta_i = g(\mu_i) = x_i^T \beta$$

Con respecto a la selección de las variables del modelo, en el presente modelo se empleara la eliminación hacia atrás, en la cual se inicia con un modelo completo y se elimina la peor de las covariables, proceso que se realiza paso paso hasta que todas las covariables restantes se consideran necesarias.

El ajuste del modelo al conjunto de datos se hará usando de la prueba de bondad de ajuste. La bondad de ajuste del modelo se evaluara a través del Criterio de Información de Akaike AIC, que busca el modelo que mejor explique los datos con un mínimo de parámetros. El estadístico de prueba es:

$$AIC = 2(P) - 2\ln(L)$$

Donde P es el número de parámetros estimados incluyendo el intercepto y L es el logaritmo de la verosimilitud alcanzado por el modelo estimado. El criterio de información de Akaike es una medida de la calidad relativa de un modelo estadístico, como tal AIC proporciona un medio para la selección del modelo. El criterio es elegir entre un grupo de modelos aquel que tenga menor AIC, este estadístico sirve para analizar la capacidad explicativa del modelo.

Verificación de los supuestos $E(e) = 0$ y $\sigma(e) = 0$. Complementariamente se evaluara el Modelo, principalmente frente a cumplimiento del supuesto de media cero y homogeneidad de varianzas en sus errores. Para lo cual se emplearan métodos gráficos.

5.4 Construcción del Modelo de dos Partes

Finalmente, los resultados obtenidos en el modelo logístico y en el lineal generalizado se multiplican para obtener una cuantificación del costo medio causado por cada individuo; esto apoyado en la siguiente expresión:

$$E[y | \vec{x}] = \Pr(y > 0) \times E[y | y > 0, \vec{x}]$$

Donde y es la variable dependiente de interés y $\vec{x}$ es un conjunto de covariables relacionadas con la ocasión y/o monto del gasto medico.

Los resultados obtenidos a través del Modelo serán contrastados contra el valor de capitación reconocido por el estado, evaluando la suficiencia en cada caso.

5.5 Evaluación de la Suficiencia

Teniendo en cuenta el impacto económico desfavorable que se viene dando en las EPS's, por el bajo valor de la Unidad de Pago por Capitación UPC; que el gobierno paga a estas entidades, se tiene como propuesta realizar la Modelación del Costo Esperado en la EPS del Valle del Cauca año 2011, la propuesta es través de un modelo de dos partes, dado que las metodologías usadas por el estado tales como, el modelo de celdas, modelo de Regresión presentan un bajo rendimiento; lo anterior genera a las EPS's daños y prejuicios por falta de capacidad financiera. Esta modelación le dará a la EPS otra alternativa para debatirle al estado que la metodología usada es ineficiente para el cálculo de dicho valor, teniendo en cuenta lo anterior el desarrolló del presente trabajo se hizo a partir de la base de datos de la población Afiliada y de la base de datos de los servicios prestados a los afiliados, para lo cual se realizó la Modelación de la probabilidad de uso y los Costos esperados de los servicios médicos, los Modelos usados fueron el Modelo de Regresión Logística y el Modelo Lineal Generalizado, los cuales se combinaron teniendo en cuenta lo realizado por otros autores y de esta manera establecer el Modelo de dos Partes, este último Modelo permitió Calcular el Costo Medio Anual de los Servicios que Requirió la Población en el 2011; con los resultados del Modelo de dos Partes se realizará la comparación de los montos que actualmente destina el gobierno para suplir los servicios de los Afiliados frente al costo global causado, estas comparaciones se realizaran por Grupos de Edad, Genero, Zona Geográfica de Residencia. Se establecerán algunos indicadores que permitan verificar la suficiencia de dicho valor, entre estos indicadores se calculara la Siniestralidad, la cual consiste en establecer la proporción del Costo ejecutado sobre el valor del Ingreso o UPC pagado por el estado, así; de esta manera si el valor es mayor al 90%, indica que los costos han sobrepasado el límite donde la EPS pierde rentabilidad frente al negocio, este desbordamiento de los Costos sobre los Ingresos hacen que la maniobrabilidad por parte de la EPS frente a la administración de los recursos financieros se un tanto complicada; dado que ante todo debe garantizarle a la población afiliada el derecho a la salud y cumplir con las obligaciones adquiridas en el sector, tales como los pagos de los prestadores, acatando las normas establecidas por los entes de Control para evitar Sanciones jurídicas y económicas. Otra manera como

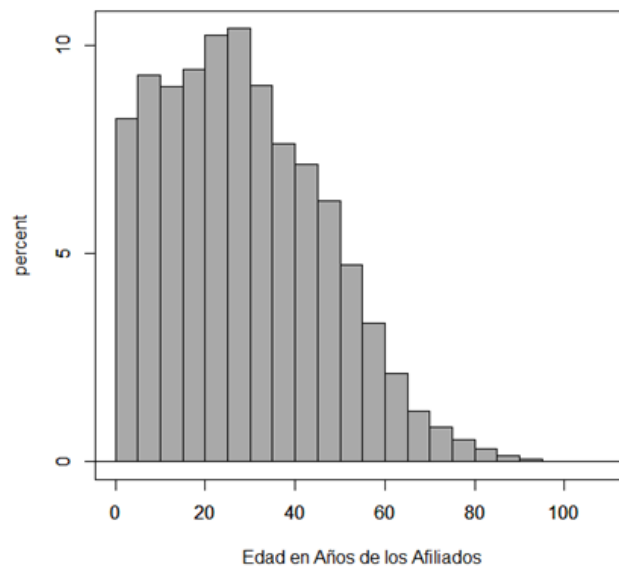
se evaluara el resultado del Modelo consiste en medir de forma proporcional el Costo Modelado vs el Costo Observado esta forma de comparación es relevante para la EPS dado que se viene presentando una insuficiencia con los recursos que actualmente reciben las EPS y de esta manera podrían definir si la asignación de los recursos debe ser usando los resultados del Modelo en vez del Valor de UPC como criterio de asignación.

CAPITULO 6

RESULTADOS

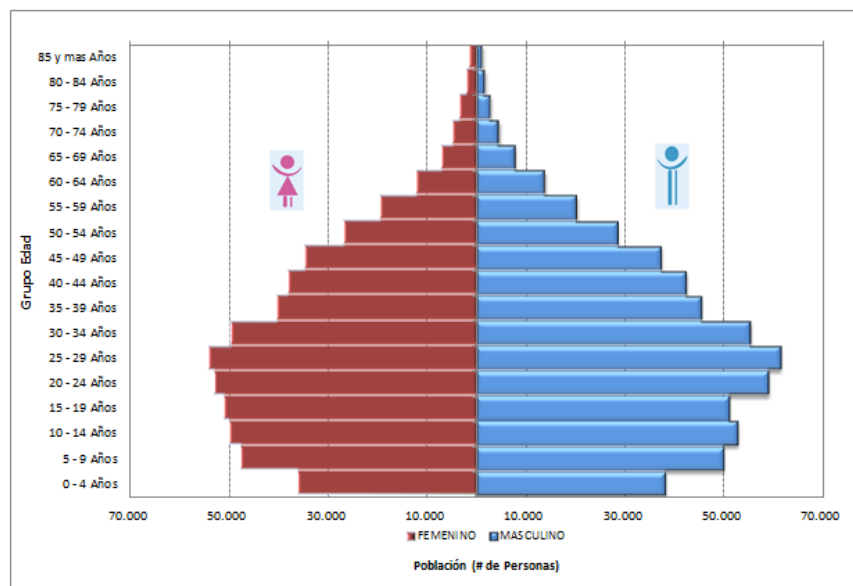
6.1 Análisis Exploratorio de Datos

En este capítulo se presentara una primera aproximación al comportamiento de los individuos respecto a la distribución por Edad, Género, Tipo Afiliado, Zona de residencia del Afiliado, concentración de individuos según enfermedades relacionadas y prestaciones consumidas, los resultados y gráficas se realizaran a través del análisis exploratorio. Las siguientes son las gráficas y tablas de resultado generadas en el análisis exploratorio propuesto:



Gráfica No. 6.1 Histograma Distribución Edad de los Afiliados

La gráfica No. 6.1 muestra la distribución de la edad para cada uno de los Afiliados, de forma clara se observa que los rangos de edad que están entre 20 y 35 años de edad concentran la mayor cantidad de Afiliados el 37.78% y los afiliados de 70 años en adelante es el grupo de menos Afiliados 2.08%.



Gráfica No. 6.2. Pirámide poblacional de afiliados a EPS del régimen contributivo, año 2011

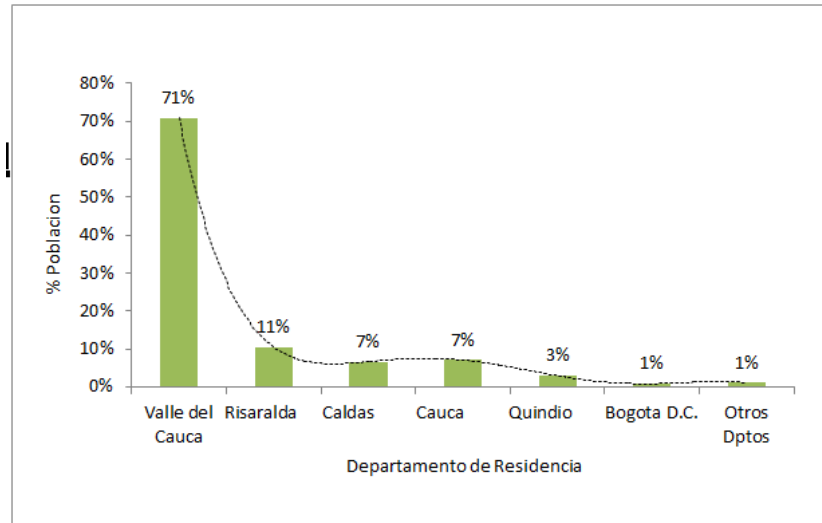
Con el fin de conocer la distribución de los afiliados según el Género; se elaboró la pirámide poblacional de los Afiliados a la EPS del régimen contributivo en el Suroccidente de Colombia, en esta pirámide se muestra que gran parte de la población son personas mayores de edad, el rango de edad de 20 a 34 años para ambos géneros presentan el mayor número de Afiliados; Mujeres 29.34%, Hombres 30.69%, mientras los Afiliados mayores a 75 años representan el grupo más pequeño de los Afiliados en ambos géneros, es importante resaltar que la población afiliada a esta EPS se considera una población joven lo cual se puede considerar conveniente para una Aseguradora en salud; pues se esperaría que los más jóvenes sean los que menos consumo de servicios demanden. El total de afiliados para el año 2011 es de 1.107.233 de los cuales las mujeres representan el 48.31% del total de afiliados y el 51.69% restantes los hombres.

Genero	Femenino	48,3%
	Masculino	51,7%
Zona Geografica	Conurbada	66,3%
	Normal	33,7%
Tipo Afiliado	Cotizantes	45,8%
	Beneficiarios	54,2%

Tabla No. 6.1 % según Genero, Zona, Tipo Afiliado

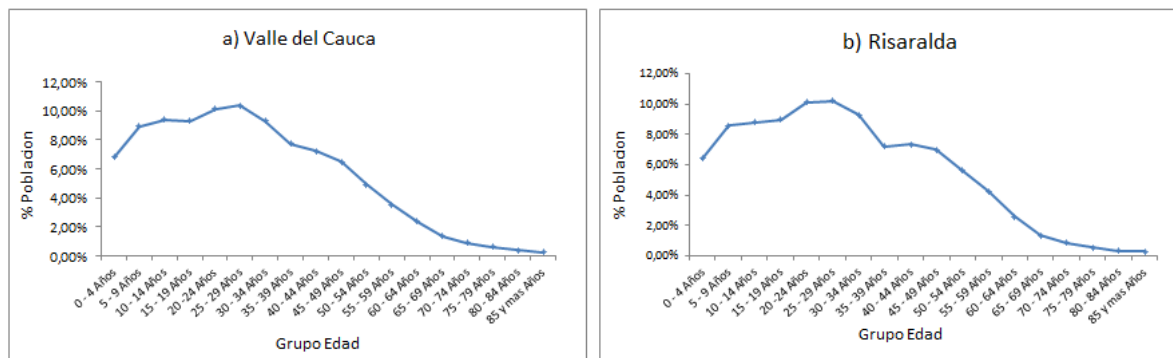
La distribución porcentual de la población Afiliada a la EPS en año 2011 corresponde; el 51,7% al género Masculino mientras el 48,3% restante son de género Femenino, se puede confirmar que el

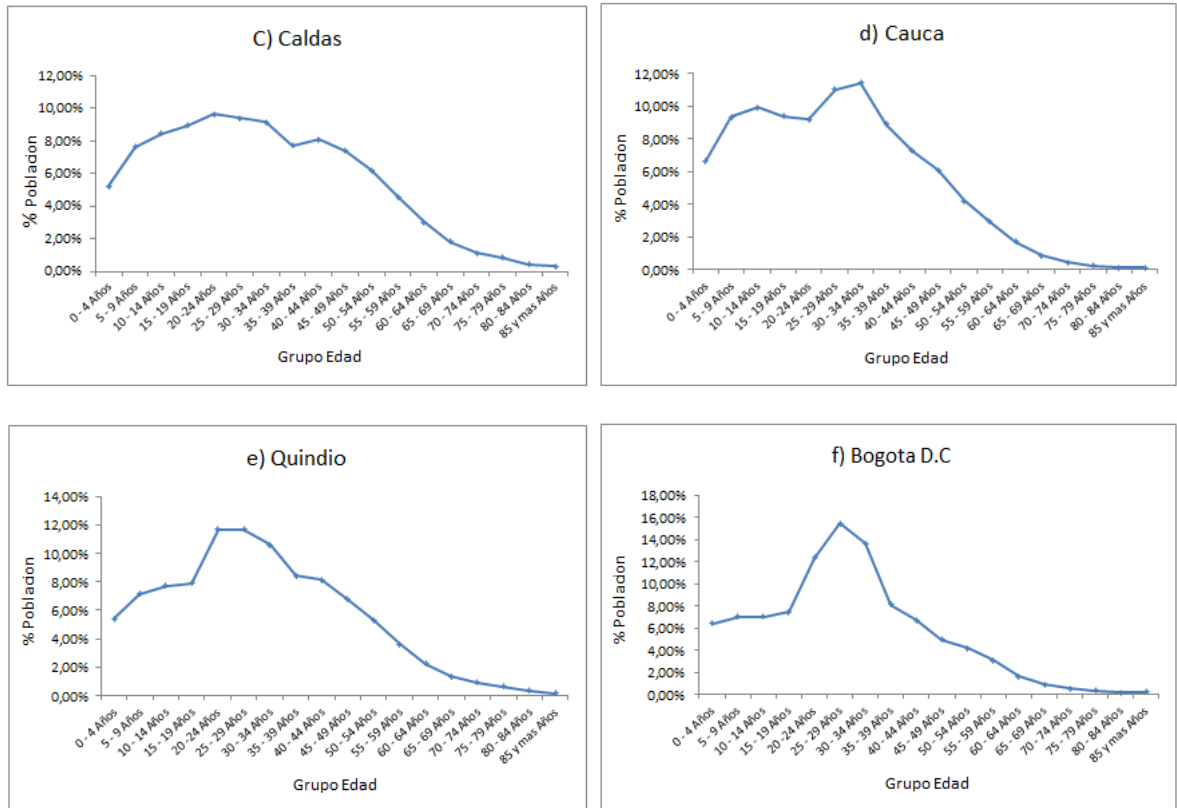
66,3% de la población residen en ciudades que se agrupan como zona geográfica Conurbada; los Afiliados que residen en la zona geográfica Normal consolidan el 33,7%, estas agrupaciones hace que se tenga un pago diferencial por parte del estado en el valor de la UPC, respecto al tipo de afiliación es decir si se vincularon a la EPS como cotizantes o beneficiarios, el 46% corresponde a personas que cotizan al Sistema de Seguridad Social en Salud y el 54% son los beneficiarios que han sido afiliados a través de los Cotizantes ver Tabla No 6.1.



Gráfica No. 6.3. Afiliados por Departamento de Residencia

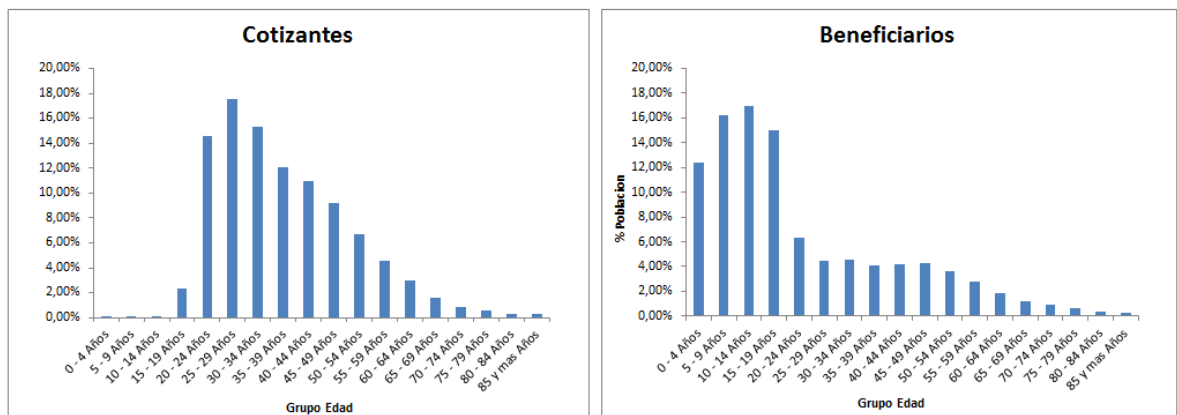
El departamento de residencia que mayor cantidad de población afiliada concentra es el valle del Cauca con el 71%; esto se debe a que en este departamento encuentra la sede principal de la EPS, después del Valle del Cauca los departamentos con mayor cantidad de Afiliados son Risaralda, Caldas y Cauca 11%,7.0% y 7.0% respectivamente, estos resultados según la Gráfica No. 6.3.





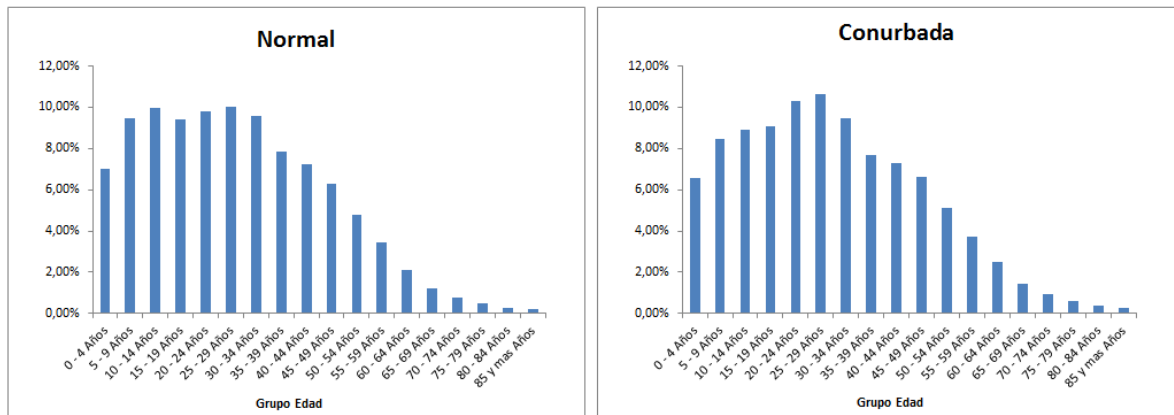
Gráfica No. 6.4 Distribución Etarea por Departamento de Residencia

Con el fin de observar la distribución por grupos de edad para los 6 primeros departamentos que concentran la mayoría de la población de la EPS, las gráficas 6.4 ; muestran que los grupos que van de 20 a 35 años de edad tiene una mayor participación frente al resto de grupo de edad, este mismo comportamiento se observó en las gráficas No. 6.1 y 6.2, las tendencias por grupos de edad según el departamento de residencia son muy similares.



Gráfica No. 6.5 Distribución Etarea por Tipo Afiliado

Al comparar los grupos de edad de Cotizantes y Beneficiarios se puede notar en la gráfica 6.5 que la mayoría de los Beneficiarios son menores a 19 años aproximadamente el 61%, de los Cotizantes muy pocos son menores de edad el 2.52%, los menores de edad pueden ser cotizantes de acuerdo con lo pronunciado por la Corte Suprema de Justicia en la sentencia T-760/08. Como es de esperarse la gran mayoría de los Cotizantes son mayores de edad el 79.64% tiene edades entre 20 y 49 años de edad, el 17.83% es mayor a 49 años.



Gráfica No. 6.6 Distribución Etarea por Zona de Residencia

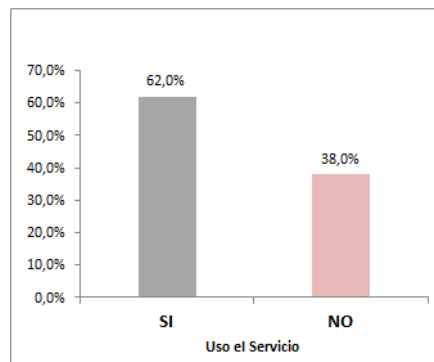
Los Afiliados que residen en zonas Conurbadas y Normal muestran una tendencia muy parecida, 38.04% y 29.39% de los afiliados están en los grupos de edad 20 a 34 años para cada una de la zonas, se puede ver que los afiliados 0 a 14 años en zona Normal según gráfica 6.6 tienen una mayor participación que los residentes en zona conurbada, 26.44% y 23.95% respectivamente.

Departamento Residencia	Ciudad de Residencia	% Mujeres	% Hombres	% Total general
Valle del Cauca	Santiago de Cali	38,1%	37,4%	37,8%
	Palmira	5,2%	5,0%	5,0%
	Buenaventura	3,6%	4,1%	3,9%
Total Valle del Cauca		46,9%	46,5%	46,7%
Risaralda	Pereira	7,3%	6,4%	6,9%
	Dosquebradas	2,7%	2,4%	2,5%
	Santa Rosa de Cabal	0,5%	0,4%	0,5%
Total Risaralda		10,5%	9,3%	9,9%
Caldas	Manizales	4,6%	4,4%	4,5%
	Chinchina	0,3%	0,3%	0,3%
	Supia	0,3%	0,3%	0,3%
Total Caldas		5,2%	5,0%	5,1%
Total Primeros 3		62,6%	60,8%	61,6%
Total General		100,0%	100,0%	100,0%

Tabla No. 6.2 Primeros Tres Departamento de Residencia de acuerdo a Cantidad de Afiliados

En la tabla No 6.2 Se relaciona la participación porcentual de los tres primeros departamentos con sus respectivas tres primeras ciudades, el 61.6% de los Afiliados viven en las ciudades principales de los departamentos en referencia, lo cual corresponde con los puntos de atención con que cuenta la EPS en cada una de estas ciudades.

Ya en cuanto al uso de servicios que inducen algún costo, el interés se centra en identificar potenciales factores que puedan influenciar:



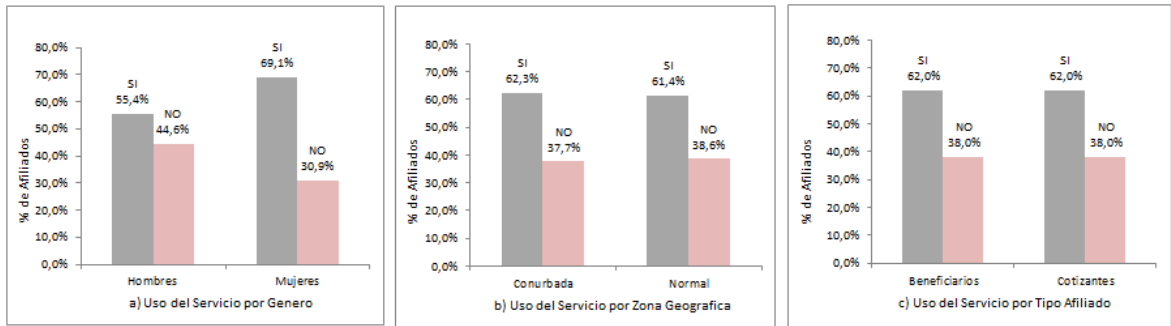
Gráfica No. 6.7 Porcentaje de Uso del Servicio Población Total

Según la gráfica 6.7. El 62% de los afiliados hizo uso de algún tipo de servicio en el periodo de referencia, el 38% restante no se tendrá en cuenta para el ajuste del Modelo de costos de servicios (2^{da} parte)

Test de Fisher	Valor-p	OR	IC[OR]
Género (Hombres)	2.2e-16	0.55	[0.550 ; 0.558]
Zona (Normal)	2.2e-16	0.96	[0.955 ; 0.971]
Tipo Afiliado (Cotizante)	0.4446	0,99	[0.98 ; 1.00]

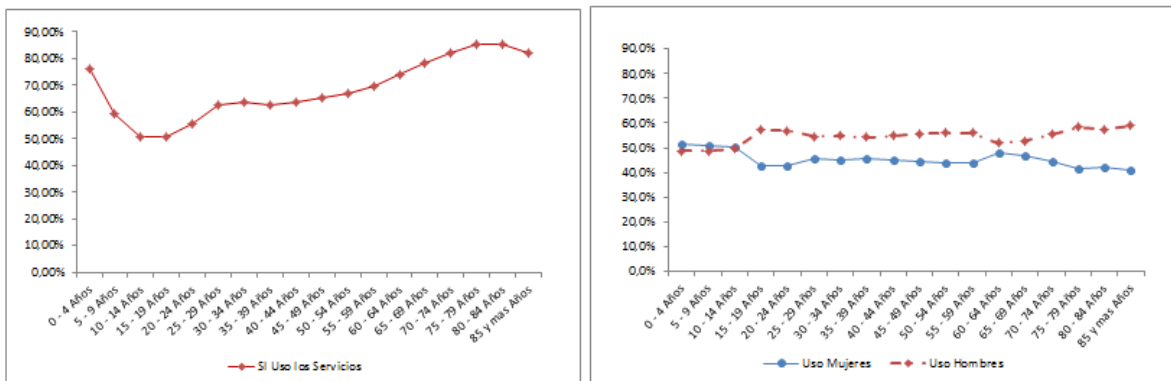
Tabla No. 6.3 Resultados Test de Fisher

Al observar la tabla No 6.2, la probabilidad de hacer del uso de servicio aumentara cuando el afiliado sea de género Femenino y disminuirá, cuando el Afiliado es de género Masculino, por lo cual ser de género Masculino se considera factor protector para un aumento de la probabilidad del uso de los servicios médicos; según los resultados del OR. Es decir residir en zona geográfica Normal, ser un afiliado Cotizante; de género Masculino minimizara la probabilidad de hacer uso de los servicios médicos.



Gráfica No. 6.8 % de Uso de los Servicios

Al observar el comparativo del uso de servicios de los afiliados a la EPS se tiene que la demanda de servicios de salud en las mujeres es mayor en comparación con los hombres, el 69.1% de las Mujeres tienen asociado el uso de los servicios de Salud; mientras que en los hombres es del 55.4%, El nivel del uso de servicios no presentan menores diferencias entre zonas Conurbadas y Normales, adicionalmente no se observan diferencias entre los Afiliados cotizantes y Afiliados beneficiarios ver gráfica 6.8 sobre el uso de los servicios, en ambas categorías la participación porcentual es del 62%



Gráfica 6.9 % de Uso de los Servicios por Grupo de Edad

El 76% de los Afiliados entre 0 a 4 años hicieron uso de los servicios, de manera muy similar los adultos mayores a 60 años de edad presentan uso del servicio por encima del 74%, el menor uso de servicios se relaciona en el grupo de 10 a 14 años el 50.44%, tal como se observa en la gráfica 6.9, se puede observar que los hombres usan en mayor proporción los servicios que las mujeres, en general la población tiene un alto uso de los servicios,

GRUPO DIAGNOSTICO EPS	% Part
FACTORES QUE INFLUYEN EN EL ESTADO DE SALUD Y CONTACTO CON LOS SERVICIOS DE SALUD	35,0%
SINTOMAS, SIGNOS Y HALLAZGOS ANORMALES CLINICOS Y DE LABORATORIO, NO CLASIFICADOS EN OTRA PARTE	17,3%
ENFERMEDADES DEL SISTEMA DIGESTIVO	8,2%
ENFERMEDADES DEL SISTEMA CIRCULATORIO	6,6%
ENFERMEDADES DEL SISTEMA RESPIRATORIO	4,9%
ENFERMEDADES DEL SISTEMA OSTEOMUSCULAR Y DEL TEJIDO CONJUNTIVO	4,4%
ENFERMEDADES ENDOCRINAS, NUTRICIONALES Y METABOLICAS	4,0%
ENFERMEDADES DEL SISTEMA GENITOURINARIO	3,6%
CIERTAS ENFERMEDADES INFECCIOSAS Y PARASITARIAS	3,4%
TRAUMATISMOS, ENVENENAMIENTOS Y ALGUNAS OTRAS CONSECUENCIAS DE CAUSAS EXTERNAS	2,7%
ENFERMEDADES DEL OJO Y SUS ANEXOS	1,9%
ENFERMEDADES DE LA PIEL Y DEL TEJIDO SUBCUTANEO	1,7%
ENFERMEDADES DEL SISTEMA NERVIOSO	1,3%
ENFERMEDADES DEL OIDO Y DE LA APOFISIS MASTOIDES	1,2%
TRASTORNOS MENTALES Y DEL COMPORTAMIENTO	1,1%
TUMORES [NEOPLASIAS]	1,0%
OTRAS PATOLOGIAS	1,7%

Tabla No. 6.4 Agrupación de diagnósticos

Se identificaron los diagnósticos de las enfermedades por los cuales los afiliados usaron los servicios médicos, debido a que las distintas enfermedades que se presentaron eran demasiadas, se utilizó una agrupación propia de la EPS para clasificar cada uno de los diagnósticos del CIE-10 relacionados en la base de datos de servicios, de esta manera se resumió la información en la tabla 6.4 encontrando que el 35.0% de los afiliados que usaron el servicio presentaron patologías por factores que influyen en el estado de salud, así como Síntomas, Signos y hallazgos Clínicos de laboratorios 17.3%, el resto de atenciones se dio por enfermedades en sistema Digestivo, Circulatorio y Respiratorio entre otros...

AGRUPACION DE PRESTACIONES EPS	% PARTICIPACION
MEDICAMENTOS	28,6%
LABORATORIO CLINICO	18,9%
ATENCION MEDICA GENERAL	16,8%
PROMOCION Y PREVENCION	6,1%
ATENCION ESPECIALIZADA	5,9%
CIRUGIAS	3,7%
ATENCION DE URGENCIAS	3,7%
TERAPIAS	3,6%
AYUDAS DIAGNOSTICAS ESPECIALIZADAS	3,4%
INSUMOS Y DISPOSITIVOS	3,2%
IMAGENOLOGIA	2,6%
ATENCION ODONTOLOGICA	2,0%
PROCEDIMIENTOS NO QUIRURGICOS	0,6%
HOSPITALIZACION	0,3%
ATENCION DOMICILIARIA	0,2%
PARTOS-CESAREAS	0,1%
IMAGENOLOGIA ESPECIALIZADA	0,1%
HOSPITALIZACION UCI	0,1%
ALTO COSTO	0,1%

Tabla No. 6.5 Agrupación de Prestaciones

La ruta de inicio para la atención medica de los afiliados la mayoría de veces inicia con una consulta médica general razón por la cual se observa su participación del 18.8% en el grupo de las tres primeras

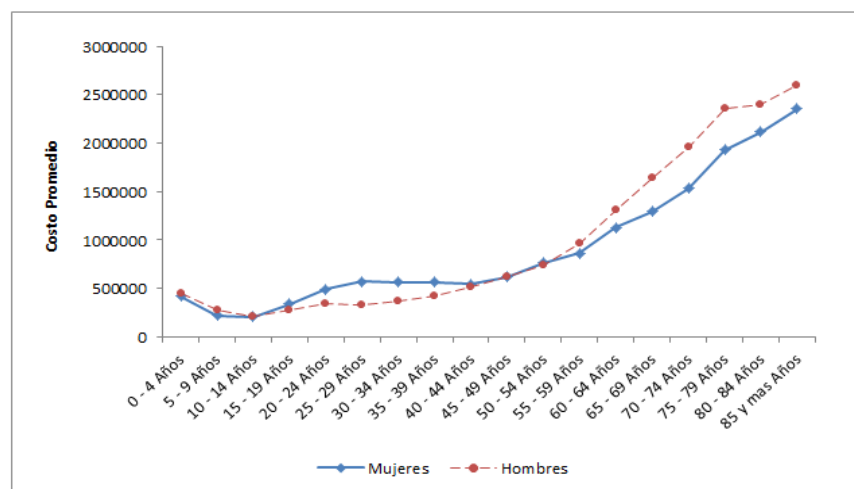
prestaciones, una vez el medico ha valorado al paciente ordena exámenes médicos y ayudas diagnosticas casi siempre laboratorios, esto se confirma con los resultados de la tabla No 6.5 donde los medicamentos representan el 28.6% y los laboratorios clínicos el 18.9% .

En cuanto al comportamiento de los costos se tiene:

Variable	Categoria	Costos ≥ 0			Costos > 0		
		mean	sd	cv	mean	sd	cv
Género	Mujer	394.483,9	2.235.484	5,67	570.763,8	2.670.191	4,68
	Hombre	285.218,4	2.923.889	10,25	515.138,1	3.914.372	7,60
Zona Geográfica	Conurbada	349.376,8	2.722.473	7,79	560.763,4	3.431.879	6,12
	Normal	315.653,4	2.388.497	7,57	513.855,2	3.030.718	5,90
Tipo Afiliado	Cotizante	375.223,7	2.636.404	7,03	604.741,5	3.326.172	5,50
	Beneficiario	306.566,3	2.595.705	8,47	494.652,4	3.283.047	6,64

Tabla No. 6.6 Resumen de Estadísticas Costo

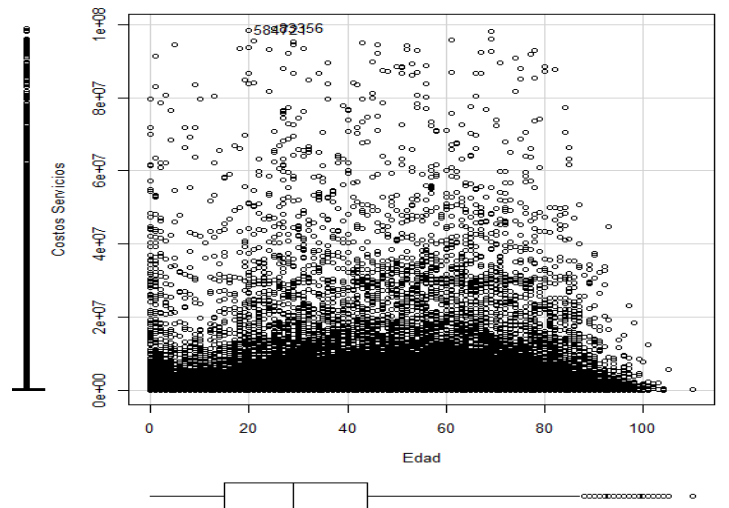
La tabla No. 6.6. Relaciona un resumen de algunas medidas estadísticas, los valores relacionados son mayores e iguales a ceros y mayores que cero. Respecto al coeficiente de variación se observa que hay una mayor heterogeneidad en los costos cuando están incluidas las colas de valores cero, la mayor heterogeneidad entre las variables evaluadas está en la variable género hombres; en ambos casos, el costo promedio de los servicios médicos según el género es más alto en las mujeres para ambos escenarios. En zona geográfica el costo promedio más alto esta para zona geográfica Conurbada al igual que el Coeficiente de variación, los Cotizantes tienen un Costo promedio mayor que lo Beneficiarios.



Gráfica 6.10 Costos Promedio por Género según Grupo Edad

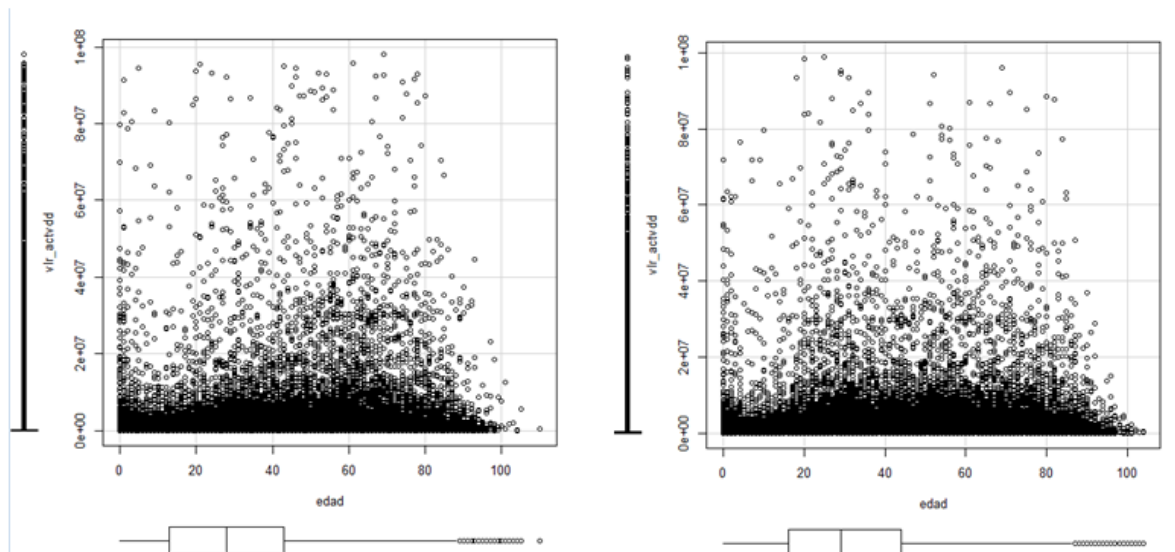
En cuanto al costo promedio por Genero se puede apreciar en la gráfica 6.14 (grafico no contiene colas de ceros) que las Mujeres entre 10 y 44 años tienen un costo promedio mayor respecto al de los Hombres, pero esta tendencia se invierte cuando los Hombres tienen 55 o más años de edad, para

ambos géneros los mayores costos promedios se encuentran a partir de los 60 años de edad siendo el máximo para hombres \$2.592.036 y para las mujeres \$2.354.188.



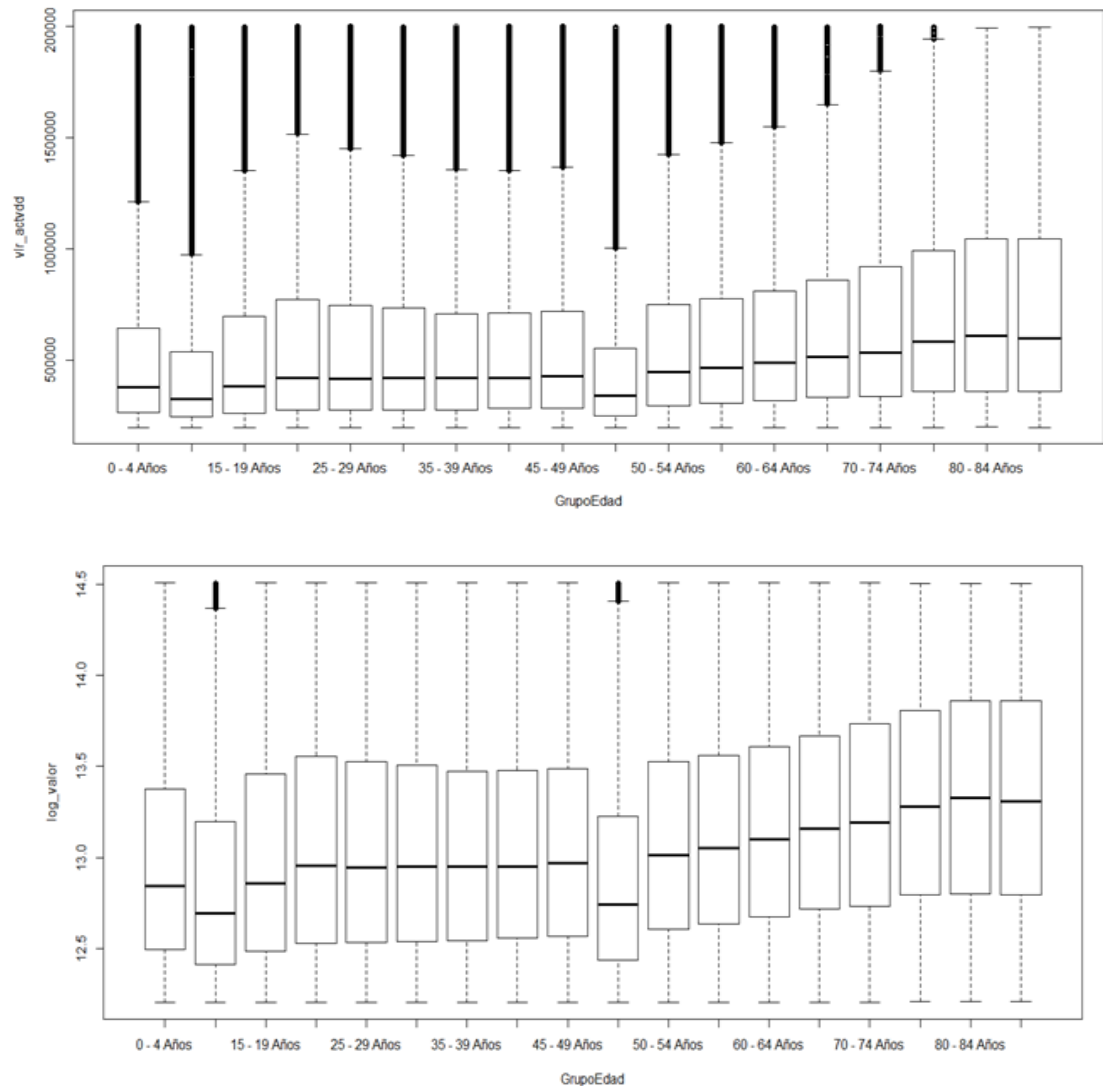
Gráfica No. 6.11 % Relación Costos de Servicios vs Edad

La relación de los Costos respecto a la Edad de los afiliados muestra que la mayoría de los Costos están por debajo de los 20 millones; de igual manera se observa que los valores se concentran en las edades que van de 20 a 70 años y que en casi todas las edades se tienen afiliados con atenciones que superan los 40 millones, se debe tener en cuenta que para la elaboración de la gráfica 6.10 se realizó acotación de los valores a relacionar tomando los menores a 100 millones de pesos esto con el fin de observar de manera más clara la relación planteada.



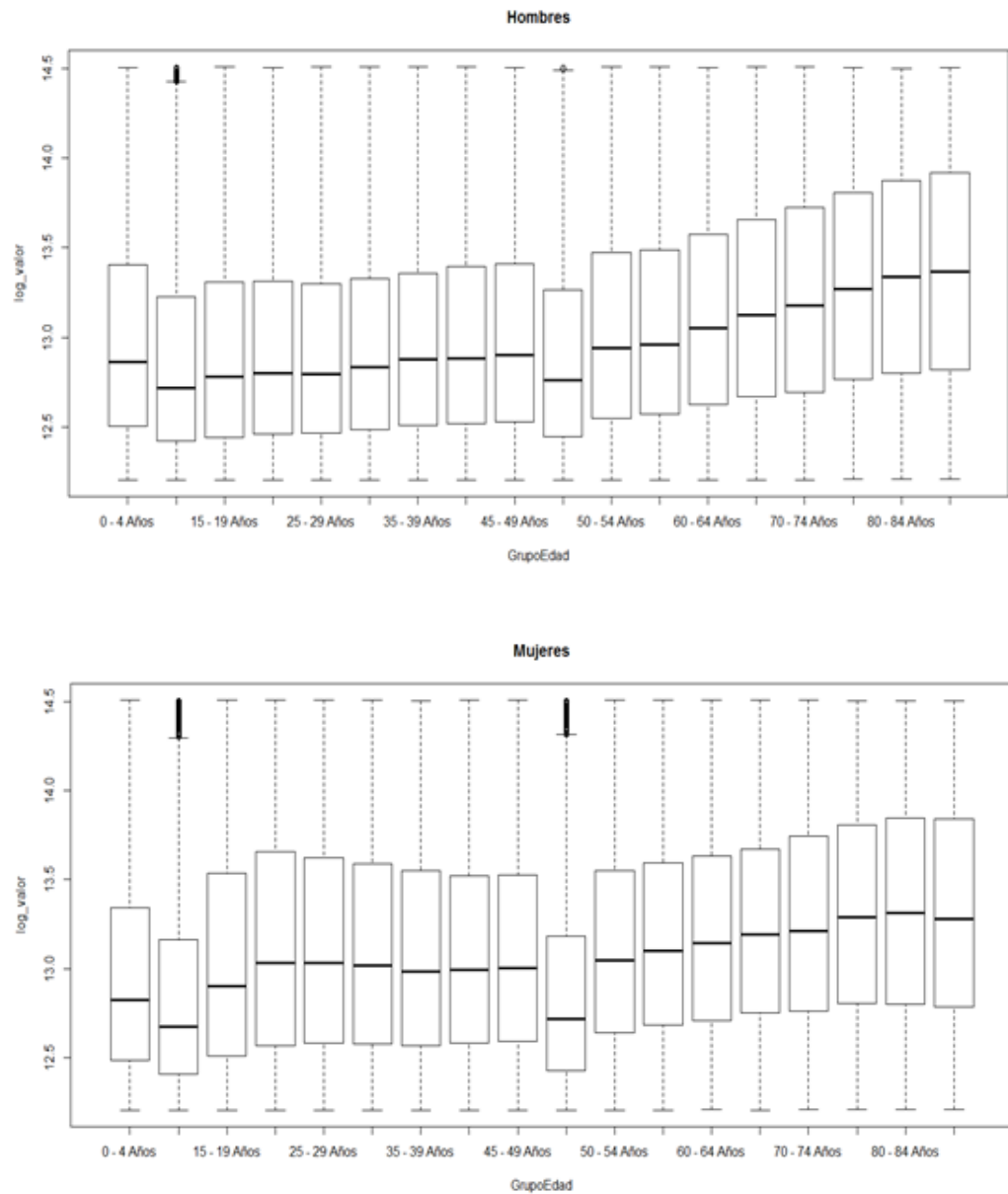
Gráfica No. 6.12 % Relación Costos de Servicios vs Edad según Género

La comparación de la relación de Costos según el Género de los Afiliados, muestran un comportamiento muy parecido, aunque se puede apreciar que los Costos de servicios de los Hombres son un poco más bajos que los de las Mujeres.



Gráfica No. 6.13 Comportamiento de los Costos según el Grupo de Edad

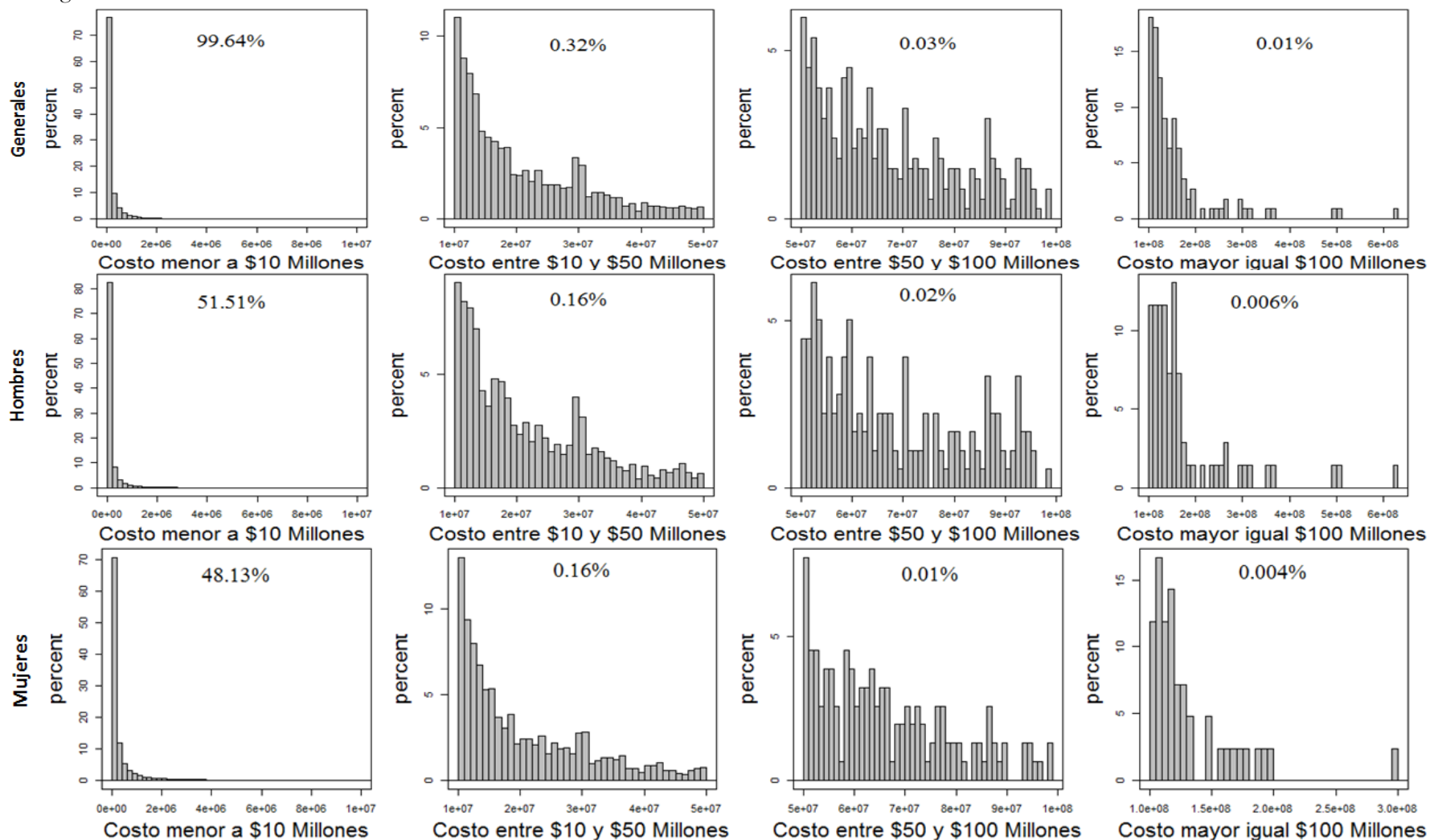
Con el fin de observar la distribución de los Costos se realizó filtro en la base de datos para los valores menores o iguales a 200 mil pesos, este filtro en razón a que es en este rango se concentran la mayoría de los valores, la gráfica 6.12 izquierda corresponde a los valores puros y la gráfica de la derecha son los valores en escala logarítmica, los grupos de edad mayores a 80 años presentan los costos más altos, mientras que las edades entre 5 a 10 años y 40 a 45 año presentan los costos más bajos.

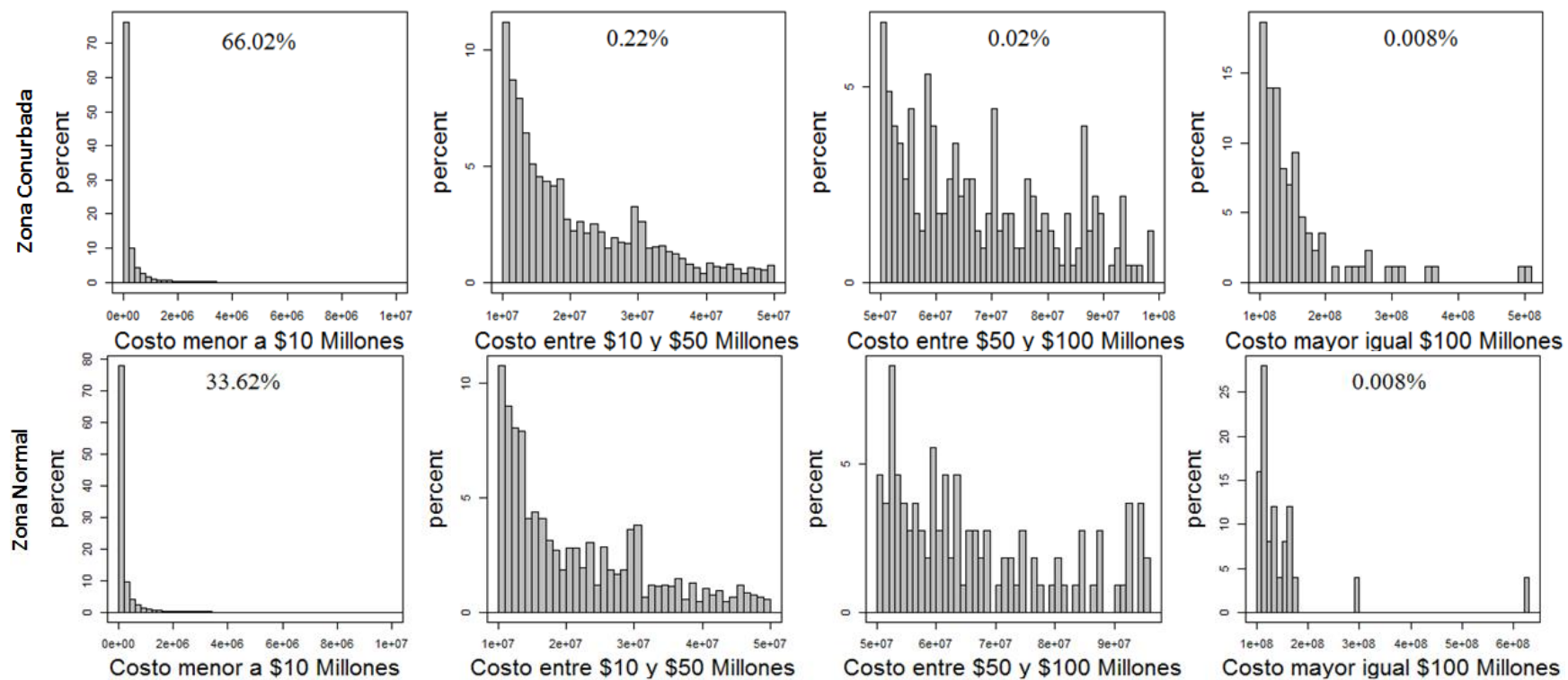


Gráfica No. 6.14 Comportamiento de los Costos según el Grupo de Edad y Genero

Al comparar la distribución de los Costos entre Hombres y Mujeres se observa que las mujeres en casi la mayoría de los grupos de edad son más costosas a excepción de los afiliados con edad entre 0 y 4 años y 45 a 49 años, los Hombres a medida que aumenta la edad su Costo muestra la tendencia a incrementarse.

Histogramas Costo de los Servicios Médicos en año 2011

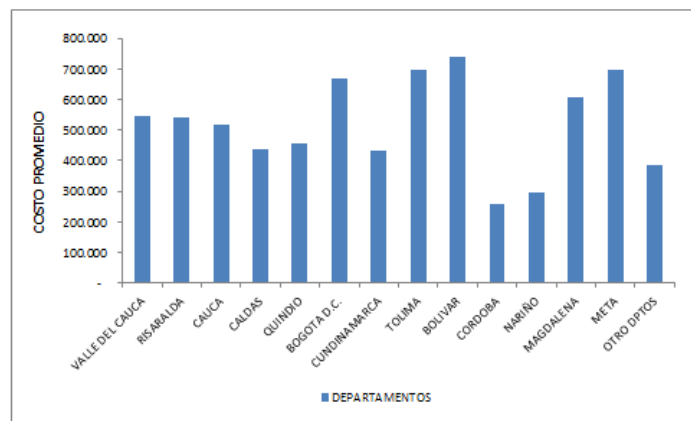




Gráfica No. 6.15 % Histogramas de los Costos de Servicios, Generales, según Genero, Zona

Debido a la gran cantidad de valores de servicios menores al costo promedio lo cual impide la correcta visualización en un solo histograma de frecuencias, como estrategia para subsanar esta dificultad y visualizar la distribución de los costos, la variable costos; se segmentó en cuatro grupos; el primero para los valores menores a \$10 millones 99.64%, seguido de valores entre \$10 y \$50 millones 0.32%, un tercer segmento para valores entre \$50 y \$100 millones 0.02% y por último los valores mayores a \$100 millones de pesos 0.008%, de igual manera se realizaron contrastes entre género y zona geográfica.

Se observa que la mayoría de los costos son menores o iguales a \$200 mil, la frecuencia de estos valores representan el 99.64% de los casos; los hombres concentran el 51.51%, mientras las mujeres el 48.13% de estos valores, el resto de los valores en ambos géneros presentan un comportamiento similar. La zona geográfica Conurbada es la de mayor demanda de servicios con costos menores \$10 millones 66.02%, el 33.62% restante de estos valores es para la zona Normal, para ambas zonas los valores menores a \$200 mil son los de mayor frecuencia, hay 7 Afiliados que presentan un costo de servicios mayor igual a \$300 millones, el paciente más costoso tiene un valor de \$624 millones. En general la gran mayoría de los servicios prestados los afiliados a la EPS tienen un costo inferior igual a los \$200 mil, mientras que los grandes valores se concentran en un grupo poblacional relativamente pequeño.



Gráfica No. 6.16 Costos Promedio por Departamentos

Se observa que los departamentos con mayor costo promedio son aquellos donde la cantidad de afiliados es menor, esta situación puede darse porque en ellos se encuentren afiliados con patologías de alta complejidad o que requirieron de una atención especializada en un momento dado del año 2011, el mayor costo promedio de los departamentos relacionados en la gráfica 6.16 corresponde a Bolívar con \$740.490 seguido por el departamento del Meta con \$696.172, el departamento del Valle del Cauca donde se encuentra la mayor cantidad de población afiliada tiene un costo promedio de \$546.066.

6.2 Ajuste del Modelo de Regresión Logística para Probabilidad de Uso

Se ajusta el modelo con todas las variables que se tienen disponibles en la base de afiliados y que se creen, son las variables que nos darán los resultados de la probabilidad de uso, la variable dependiente corresponde a variable dummy o recodificación de la variable valor actividad, la cual se codifico 1 si el afiliado presenta un costo asociado y cero (0) en caso que el afiliado no presente valor asociado por servicios de salud, como variables explicativas se tienen las variables Edad, Género, Zona Geográfica, Tipo Afiliado, Periodos Compensados más las interacciones de las variables Edad*Género y Edad*Tipo Afiliado; la variable Zona Geográfica agrupa las ciudades en dos categorías Normal, Conurbada (Especial y diferencial) de acuerdo con la norma del Ministerio de la Protección Social, Tipo Afiliado corresponde a las categorías C= cotizantes, B= Beneficiarios, dado que la variable Edad tiene presencia de un de factor cuadrático en el punto de 15 años y con el fin de minimizar la influencia de este factor; se crean dos variables; una para los mayores iguales a 15 años y otra para los menores a 15 años, a partir de la diferencia del valor de la Edad menos 15 se eleva al cuadrado y se multiplica por una variable dummy de la Edad(1 para ≥ 15 años, cero otros casos), las dos nuevas variables son $Edad^2 < 15$ y $Edad^2 > 15$.

La modelación inicia planteando un modelo completo:

$$P(Y = 1) =$$

$$\frac{e^{\beta_0 + \beta_1 Edad + \beta_2 Gnro + \beta_3 Zona + \beta_4 TAfido + \beta_5 Prdos Cmpnsds + \beta_6 Edad^2 < 15 + \beta_7 Edad^2 > 15 + \beta_8 Edad * Gnro + \beta_9 Edad * TAfido}}{1 + e^{\beta_0 + \beta_1 Edad + \beta_2 Gnro + \beta_3 Zona + \beta_4 TAfido + \beta_5 Prdos Cmpnsds + \beta_6 Edad^2 < 15 + \beta_7 Edad^2 > 15 + \beta_8 Edad * Gnro + \beta_9 Edad * TAfido}}$$

Cuyos resultados se presentan en la tabla No 6.7

Variables	Coeficientes	e^{β}	valor-p	
(Intercept)	-1.860e+00	0,15567	<2e-16	***
edad	1.423e-02	1,01433	<2e-16	***
Género[T.M]	-2.595e-01	0,77144	<2e-16	***
Zona[T.Normal]	1.013e-01	1,10661	<2e-16	***
tpo_afido[T.C]	7,265e-01	2,06783	<2e-16	***
periodos_compensados	2.156e-01	1,24061	<2e-16	***
$Edad^2 > 15$	1.688e-04	1,00017	<2e-16	***
$Edad^2 < 15$	8.885e-03	1,00892	<2e-16	***
edad:sexo[T.M]	-1.353e-02	0,98656	<2e-16	***
edad:tpo_afido[T.C]	-1.009e-02	0,98996	<2e-16	***
Devianza	1226800	AIC	1226820	

Tabla No. 6.7 Resultados del Modelo Completo (Seleccionado)

El modelo seleccionado muestra significancia en todas las variables a la explicación de la probabilidad de ocurrencia del evento $P(Y = 1)$. Ya en la explicación del modelo se obtiene que el incremento de la edad genera propensión al uso del servicio, el cual se incrementa cuando el individuo es cotizante y por cada mes Compensado la probabilidad de uso se incrementa en 1.234 veces. Por otra parte se presentan algunos factores protectores tales como el ser de género Masculino, con vivienda en zona Normal y ser Cotizante en el sistema. Esto implica una menor propensión al uso entre los Hombres ($OR = 0.82$), Cotizantes ($OR = 0.934$).

El modelo global permite realizar las estimaciones de probabilidad de uso requeridas para la estimación del Costo Medio. A manera de ejemplo se presenta el siguiente caso de estimación:

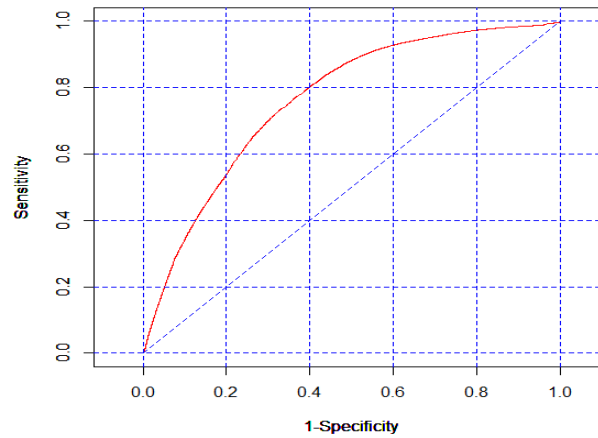
La variable edad se fijara por cada 10 años.

- Edad: 10;
- Género: Masculino (1);
- Zona Geográfica: Normal (1)
- Tipo Afiliado: Beneficiario (0)
- Periodos Compensados (12)
- Edad*Género: 10*1
- Edad* Tipo Afiliado Beneficiario 10*1

$$P(Y > 0) = \frac{e^{-1.292 + 0.0053Edad - 0.204Gnro + 0.093Zona - 0.068TAfido + 0.210PrdosCmpnsds - 0.0147Edad*Gnro + 0.009Edad*TAfido}}{1 + e^{\beta_0 + \beta_1Edad + \beta_2Gnro + \beta_3Zona + \beta_4TAfido + \beta_5PrdosCmpnsds + \beta_6Edad*Gnro + \beta_7Edad*TAfido}} = 0.74$$

La probabilidad de que un individuo con el perfil antes mencionado haga uso de un servicio en la EPS es del 74%

Para evaluar la bondad de ajuste del Modelo se dispone de su curva ROC, la cual nos indica si el área bajo la curva es mayor de 0,9 se considera que la prueba es muy exacta, mientras que valores comprendidos entre 0,7 y 0,9 indican una exactitud moderada. Los valores comprendidos entre 0,5 y 0,7 se corresponden con una exactitud baja. En definitiva, cuanto más próxima es una curva ROC a la esquina superior izquierda, más alta es la exactitud global de la prueba. Para el caso de la curva resultante en el modelo planteado se tiene una exactitud moderada 0.763.



Gráfica No. 6.17 Curva ROC Probabilidad de Uso

La evaluación de la capacidad Predictiva del Modelo a través de la Sensibilidad y Especificidad, para lo cual se validaron resultados en los puntos de corte 0.5, 0.6, 0.7 0.8 y 0.9, observando los siguientes resultados en la tabla No 6.8, según estos puntos de corte relacionados, se observar que el modelo pierde capacidad cuando el punto de corte es mayor al 0.7 y menor al 0.6, este mismo resultado se puede observar gráfica 6.17 de la Curva ROC.

	Puntos de Corte Evaluados				
	0,5	0,6	0,7	0,8	0,9
Sensibilidad	0,482	0,454	0,416	0,285	0,013
Especificidad	0,788	0,817	0,845	0,911	0,996
Falsos Usuarios	0,212	0,183	0,155	0,089	0,004
Falsos no usuarios	0,518	0,546	0,584	0,715	0,987

Tabla No. 6.8 Resultados Sensibilidad y Especificidad del Modelo Completo (Seleccionado)

6.3 Ajuste del Modelo de Lineal Generalizado Costo del Servicio

Para el cálculo del valor esperado se usa el modelo Lineal Generalizado con familia Exponencial Gamma y función de enlace Logarítmica, la metodología para selección de las variables en la modelación es el método de eliminación hacia atrás, el modelo completo evalúa como variable dependiente la variable Valor actividad que corresponde a los Costos por atenciones de servicios médicos a los afiliados en el año 2011, el criterio de filtro en la base de datos se restringe a casos cuyo valor es mayor a cero, como variables explicativas del modelo se evaluará: Edad, Género, Zona geográfica, Tipo afiliado, Periodos compensados, dado que la variable Edad tiene presencia de un factor cuadrático en el punto de 15 años y con el fin de minimizar la influencia de este factor; se crean dos variables; una para los mayores iguales a 15 años y otra para los menores a 15 años, a partir de la diferencia del valor de la Edad menos 15 se eleva al cuadrado y se multiplica por una variable dummy de la Edad (1 para ≥ 15 años, cero otros casos), las dos nuevas variables son $Edad^2 < 15$ y $Edad^2 > 15$, de igual manera se relacionaron las interacciones de $Edad^2 < 15$ por las variables mencionadas anteriormente excepto la edad, para el caso de la variable $Edad^2 > 15$ previamente se validó en el ajuste de un modelo inicial encontrando que no es significativa por lo cual se descarta su inclusión.

La modelación inicia planteando un modelo completo:

$$g(\mu) = \eta = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + \dots a_px_p$$

$$\ln(\mu) = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + \dots a_px_p$$

$$\ln(E(\text{costo})) = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + \dots a_px_p$$

$$E(\text{Costo}) = e^{a_0 + a_1\text{Edad} + a_2\text{Gnro} + a_3\text{Pcomp} + a_4\text{Edad}^2 < 15 + a_5\text{Zona} + a_6\text{TAflado} + a_7\text{Gnro} * \text{Edad}^2 < 15 + a_8\text{Pcomp} * \text{Edad}^2 < 15 + a_9\text{Zona} * \text{Edad}^2 < 15 + a_{10}\text{TAflado} * \text{Edad}^2 < 15}$$

Los resultados del Modelos completo se presentan en la tabla No 6.9.

Variables	Coefficientes	e^{β}	valor-p	
(Intercept)	1.166e+01	115.844,0	<2e-16	***
Edad	2.877e-02	1,029188	<2e-16	***
Género[T.M]	-1.796e-01	0,835604	<2e-16	***
Periodos_compensados	5.052e-02	1,051818	<2e-16	***
Edad ² < 15	5.121e-03	1,005134	<2e-16	***
Zona[T.Normal]	-6.920e-03	0,993104	0.73458	
Tpo_afldo[T.C]	-7.065e-04	0,999294	0.97236	
Género[T.M] : Edad ² < 15	1.946e-03	1,001948	5.48e-08	***
Periodos_Compensados : Edad ² < 15	-1.294e-04	0,999871	0.00389	**
Zona[T.Normal] : Edad ² < 15	-1.008e-04	0,999899	0.78824	
Tpo_afldo[T.C] : Edad ² < 15	-2.097e-03	0,997905	0.63719	
Devianza	1732382	AIC	19050305	

Tabla No. 6.9 Resultados del Modelo Lineal Generalizado Completo

Al evaluar los resultados del Modelo completo se observa en la tabla 6.9 que los coeficientes estimados del Modelo; muestran que las variables que tienen significancia son: Edad, Género, Periodos Compensados y la nueva variable Edad²<15 más las interacciones Género*Edad²<15, Periodos Compensados*Edad²<15, respecto a las variables que pierden significancia en este planteamiento corresponde a Zona geográfica, Tipo Afiliado y las interacciones de estas dos últimas variables con Edad²<15. Teniendo en cuenta los resultados de la tabla 6.9, se ajustara el Modelo con la exclusión de la covariables que menos aporta a la estimación del Costo esperado, el modelo ajustado es el siguiente:

$$E(Costo) = e^{a_0 + a_1 Edad + a_2 Gnro + a_3 PrdosCmps + a_4 Edad^2 < 15 + a_5 Gnro * Edad^2 < 15 + a_6 PrdosCmps * Edad^2 < 15}$$

Lo resultados del modelo ajustado se relaciona en la siguiente tabla

Variables que salen	# de Parametros	Devianza	AIC
edad:Género[T.M] Zona[T.Normal] Tpo_afldo[T.C] Zona[T.Normal] : Edad ² < 15 Tpo_afldo[T.C] : Edad ² < 15	6	1732382	19021613

Variables	Coefficientes	e^{β}	valor-p	
(Intercept)	1.165e+01	114.691,4	<2e-16	***
edad	2.878e-02	1,02920	<2e-16	***
Género[T.M]	-1.799e-01	0,83535	<2e-16	***
Periodos_compensados	5.061e-02	1,05191	<2e-16	***
Edad ² < 15	5.062e-03	1,00507	<2e-16	***
Género[T.M] : Edad ² < 15	1.948e-03	1,00195	4.63e-08	***
Periodos_Compensados : Edad ² < 15	-1.268e-04	0,99987	0.0043	**
Devianza	1732411	AIC	19050311	

Tabla No. 6.10 Resultados del Modelo Lineal Generalizado Completo

El resultado del Modelo ajustado muestra que todas las covariables son significativas, de igual manera se valida la significancia del modelo a través de la diferencia entre la desviación residual del Modelo anterior con el nuevo Modelo, este resultado se evalúa con la prueba chi-cuadrado y 6 grados de libertad; el resultado es menor a $\alpha = 0.05$, lo cual indica que sus coeficientes son diferentes de cero,

dado los resultados y significancia del Modelo ajustado, es el Modelo seleccionado para realizar la estimación del Costo esperado.

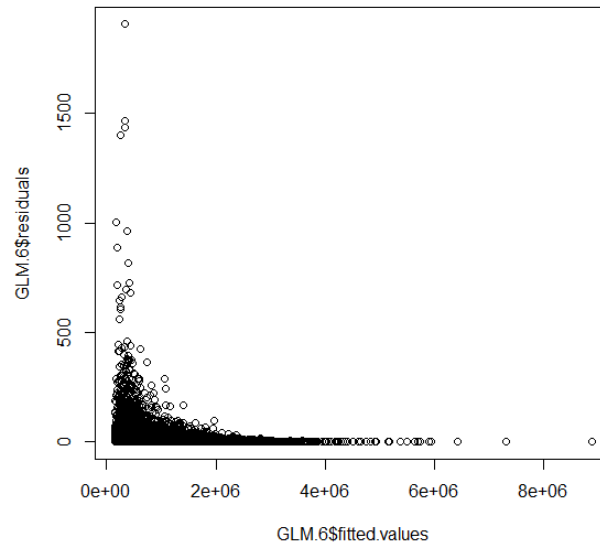
El modelo ajustado permitirá estimar el Costo esperado el cual se requiere para la estimación del Costo Medio. A continuación se presenta un ejemplo de estimación, caso particular:

- Edad: 36;
- Género: Masculino (1);
- Periodos Compensados (12)
- $\text{Edad}^2 < 15$: 441
- Género* $\text{Edad}^2 < 15$: $1 \cdot 441$
- Periodos Compensados* $\text{Edad}^2 < 15$: $12 \cdot 441$

$$E(\text{Costo}) = e^{11,65 + 0,029\text{Edad} + -0,18\text{Sexo} + 0,051\text{PrdosCmps} + 0,0051\text{Edad}^2 < 15 + 0,0019\text{Sexo} * \text{Edad}^2 < 15 + -0,00013\text{PrdosCmps} * \text{Edad}^2 < 15}$$

$$= \$5.575.455$$

La validación del ajuste del Modelo a los datos se realizará por medio del gráfico de los Residuales vs los valores Predichos.



Gráfica No. 6.18 Y estimado vs Residuales

Según lo observado en el gráfico 6.18 la forma de la distribución de los valores de $\hat{Y}$ vs lo Residuales corresponde a una disminución de la varianza ver (Behar Gutierrez, 2003), por lo cual no se cumple el supuesto de homogeneidad de la varianza, aunque el resultado de la desviación como tal es muy pequeña frente a los valores o escala real de los Costos de servicios, el supuesto de normalidad se probó

de forma gráfica obteniendo como resultado el cumplimiento del supuesto. A continuación se presentan los resultados del cálculo de la media y la desviación en la tabla No. 6.11

Media	Desvstand
0.000	7.3182

Tabla No. 6.11 Estadísticas Residuales

6.4 Construcción del Modelo de dos Partes

En la generación del modelo de dos Partes intervienen la Modelación de la Probabilidad de uso y el Costo Esperado de los servicios, al combinar estos dos Modelos se puede calcular el Costo Medio Anual por los servicios médicos que requieren los Afiliados. Para el caso del presente trabajo la variable edad está presentando un componente cuadrático cuyo punto de inflexión se encuentra en la Edad de 15 años, teniendo en cuenta lo anterior; a continuación se relaciona las ecuaciones de los Modelos que intervienen en el Modelo de dos Partes propuesto:

Modelación de la Probabilidad de uso

$$\pi(\vec{x}) = \frac{e^{g(\vec{x})}}{1 + e^{g(\vec{x})}}$$

Entonces $\pi(\vec{x}) = P(Y > 0 | \vec{X})$ estima la probabilidad de que un individuo con características definidas por el conjunto de $\vec{X}$ pertenezca a la población correspondiente a $Y = 1$.

$$g(\vec{x}) = \ln \left\{ \frac{\pi(\vec{x})}{[1 - \pi(\vec{x})]} \right\} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_k x_k$$

La forma específica del modelo de regresión logística queda expresado como:

$$P(Y > 0) = \pi(\vec{x}) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k}}$$

Modelo Costo Esperado

$$g(\mu) = \eta = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + \dots a_p x_p$$

$$\ln(\mu) = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + \dots a_p x_p$$

$$E[y | y > 0, \vec{X}] = \ln(E(\text{costo})) = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + \dots a_p x_p$$

$$E[y | y > 0, \vec{X}] = e^{a_0 + a_1 x_1 + a_2 x_2 + \dots + a_k x_k}$$

Finalmente, los resultados obtenidos en el modelo logístico y en el lineal generalizado se multiplican para obtener una cuantificación del costo medio causado por cada individuo; esto apoyado en la siguiente expresión:

$$E[y | \vec{x}] = P(Y > 0 | \vec{X}) \times E[y | y > 0, \vec{X}]$$

Donde y es la variable dependiente de interés y $\vec{x}$ es un conjunto de covariables relacionadas con la ocasión y/o monto del gasto médico.

Los resultados del Modelo de dos Partes correspondiente al Costo Medio Anual por los servicios médicos, se compararan con el valor de la UPC año 2011 que el gobierno le ha pagado a la EPS del valle del Cauca, así de esta manera se determinara si el valor pagado por el Estado corresponde con el valor que se esperaría la EPS gastara en la atención de sus afiliados en el término de un año, o si por el contrario con esta metodología se debería ajustar dicho valor.

6.5 Evaluación de la Suficiencia de la UPC año 2011

Para evaluar la suficiencia del valor de la Unidad de Pago por Capitación UPC; que el gobierno pago a la EPS del Valle del Cauca en el año 2011, el desarrolló del presente trabajo se hizo a partir de la base de datos de la población Afiliada y de la base de datos de los servicios prestados a los afiliados, para lo cual se realizó la Modelación de la probabilidad de uso y los Costos esperados de los servicios médicos, los Modelos usados fueron el Modelo de Regresión Logística y el Modelo Lineal Generalizado, los cuales se combinaron teniendo en cuenta lo realizado por otros autores y de esta manera establecer el Modelo de dos Partes, este último Modelo permitió Calcular el Costo Medio Anual de los Servicios que Requirió la Población en el 2011; con los resultados del Modelo de dos Partes se realizará la comparación de los montos que actualmente destina el gobierno para suplir los servicios de los Afiliados frente al Costo global causado, estas comparaciones se realizaran por Grupos de Edad, Genero, Zona Geográfica de Residencia. Se establecerán algunos indicadores que permitan verificar la suficiencia de dicho valor, entre estos indicadores se calculará la Siniestralidad, la cual consiste en establecer la proporción del Costo ejecutado sobre el valor del Ingreso o UPC pagado por el estado, así; de esta manera si el valor es mayor al 90%, indica que los costos han sobrepasado el límite donde la EPS pierde rentabilidad frente al negocio; el 10% restante es para gastos de administración de la EPS, este desbordamiento de los Costos sobre los Ingresos hacen que la maniobrabilidad por parte de la EPS frente a la administración de los recursos financieros sea un tanto complicada; dado que ante todo debe garantizarle a la población afiliada el derecho a la salud y cumplir con las obligaciones adquiridas en el sector, tales como los pagos de los prestadores, acatando las normas establecidas por los entes de Control para evitar Sanciones jurídicas y económicas. Otra manera como se evaluara el resultado del Modelo consiste en medir de forma proporcional el Costo Modelado vs el Costo Observado esta forma de comparación es relevante para la EPS dado que se viene presentando una insuficiencia con los recursos que actualmente reciben las EPS y de esta manera podría definirse si la asignación de los recursos debe ser usando los resultados del Modelo en vez del Valor de UPC como criterio de asignación. A continuación se presentan los resultados del Costo Modelado para el cálculo de la suficiencia de la UPC

Grupo Edad	Costo Ejecutado	Costo Modelado	UPC Anual	% Part. Costo_Ejctdo/ Costo Mod
0 - 4 Años	\$24.708.504.700	\$23.242.974.396	\$31.598.404.529	94,1%
5 - 9 Años	\$14.588.813.800	\$17.252.138.571	\$13.054.582.066	118,3%
10 - 14 Años	\$10.947.602.600	\$14.374.631.637	\$11.936.869.862	131,3%
15 - 19 Años	\$16.254.318.900	\$15.333.191.232	\$14.799.740.345	94,3%
20 - 24 Años	\$26.383.040.600	\$19.781.957.666	\$25.791.874.632	75,0%
25 - 29 Años	\$33.589.415.400	\$26.912.805.891	\$30.114.254.803	80,1%
30 - 34 Años	\$31.866.511.700	\$30.041.041.680	\$29.511.243.205	94,3%
35 - 39 Años	\$26.833.249.100	\$29.379.844.079	\$24.927.878.901	109,5%
40 - 44 Años	\$27.109.128.900	\$32.862.440.034	\$24.059.178.511	121,2%
45 - 49 Años	\$29.084.266.000	\$34.895.743.750	\$27.171.309.704	120,0%
50 - 54 Años	\$28.043.360.400	\$31.789.184.784	\$27.014.287.727	113,4%
55 - 59 Años	\$25.431.631.500	\$27.417.167.875	\$24.412.306.368	107,8%
60 - 64 Años	\$23.511.912.500	\$21.509.905.816	\$21.370.529.965	91,5%
65 - 69 Años	\$17.168.123.000	\$14.879.865.153	\$15.811.914.071	86,7%
70 - 74 Años	\$13.744.872.200	\$11.629.593.996	\$12.543.467.145	84,6%
75 - 79 Años	\$11.629.058.800	\$9.326.492.351	\$10.604.299.452	80,2%
80 - 84 Años	\$7.573.603.700	\$6.720.515.837	\$6.739.206.913	88,7%
85 y mas Años	\$5.779.295.100	\$5.971.292.847	\$4.704.346.432	103,3%
Total general	\$374.246.708.900	\$373.320.787.594	\$356.165.694.629	99,8%

Tabla No. 6.12 Resultados de Modelación Costo Medio Anual x Grupo de Edad

Los resultados en la tabla 6.12 muestran los valores del Costo Ejecutado por servicios médicos en año 2011, Costo Modelado resultado del Modelo de dos Partes y el Valor de UPC pagado por el Estado en cada grupo de edad; al comparar el Costo ejecutado respecto a l Costo Modelado se observa que los valores son muy cercanos, la diferencia es de \$925 millones la cual corresponde al 0.25%, el Modelo es el 99.8% los valores del Costo real, con este resultado se ratifica la capacidad de la Modelación de dos Partes. Se observa que el Grupo de edad donde se presenta la menor estimación del Costo Medio Anual es para el Grupo de Edad 20 a 24 años 75%, y la máxima sobre estimación se da en el Grupo de Edad 10 a 14 años 31%, en general los valores estimados por la metodología estadística seleccionada, muestran unos resultados aceptables, inicialmente se había planteado la Modelación sin intervenir el factor cuadrático que relaciona la variable Edad, los resultados del Modelo mostraban una diferencia muy alta de \$23.000 millones el 93,69% razón por la que se ajustó el Modelo propuesto. En la tabla No 6.13 se relaciona la Modelación del Costo Medio respecto al género de los Afiliados

Género	Costo Ejecutado	Costo Modelado	UPC Anual	% Part. Costo_Ejctdo/ Costo Mod
Mujeres	\$ 210.999.944.400	\$ 218.889.481.305	\$ 203.752.077.979	103,7%
Hombres	\$ 163.246.764.500	\$ 154.431.306.289	\$ 152.413.616.651	94,6%
Total general	\$ 374.246.708.900	\$ 373.320.787.594	\$ 356.165.694.629	99,8%

Tabla No. 6.13 Resultados de Modelación Costo Medio Anual x Género

Como se observó en el Análisis exploratorio de los datos, las mujeres presentan un mayor Costo asociado por servicios médicos, de igual manera se observa en la Modelación del Costo; que el Modelo relaciona esta misma tendencia. De nuevo se confirma que la metodología del Modelo de dos Partes es adecuada para la modelación de datos de Costos de Salud, de acuerdo a la Probabilidad de uso y Costo

Esperado, continuando con el análisis de los resultados del Modelo; se relaciona los resultados por Zona Geográfica, ver tabla 6.14.

Zona Geografica	Costo Ejecutado	Costo Modelado	UPC Anual	% Part. Costo_Ejctdo/ Costo Mod
conurbada	\$ 256.358.050.900	\$ 253.056.829.094	\$ 245.108.725.206	98,7%
Normal	\$ 117.888.658.000	\$ 120.263.958.500	\$ 111.056.969.423	102,0%
Total general	\$ 374.246.708.900	\$ 373.320.787.594	\$ 356.165.694.629	99,8%

Tabla No. 6.14 Resultados de Modelación Costo Medio Anual x Zona Geográfica

Las diferencias del Costo Medio generado en la Modelación muestra son del 2% aproximadamente, realmente son esta diferencia se puede considerar mínima para ambas Zonas Geográficas con respecto al Costo Ejecutado, estos resultados generan la confianza a la EPS del Valle del Cauca sobre el uso de la Modelación de dos Partes.

Tipo Afiliado	Costo Ejecutado	Costo Modelado	UPC Anual	% Part. Costo_Ejctdo/ Costo Mod
Beneficiario	\$ 184.024.057.900	\$ 187.358.289.439	\$ 183.765.455.738	101,8%
Cotizante	\$ 190.222.651.000	\$ 185.962.498.155	\$ 172.400.238.891	97,8%
Total general	\$ 374.246.708.900	\$ 373.320.787.594	\$ 356.165.694.629	99,8%

Tabla No. 6.15 Resultados de Modelación Costo Medio Anual x Tipo Afiliado

Los resultados que se observan en la tabla No 6.15 muestran que la estimación del Costo Medio Anual presenta una subestimación en los cotizantes de aproximadamente un 2.2% \$4.260 millones de pesos, se considera que el Modelo cumple con lo esperado la agrupación generada por la Variable Tipo Afiliado sigue la tendencia del Costo ejecutado de manera muy parecida y cercana.

Departamento	Costo Ejecutado	Costo Modelado	UPC Anual	% Part. Costo_Ejctdo/ Costo Mod
VALLE DEL CAUCA	\$ 268.463.932.100	\$ 265.141.081.879	\$ 254.954.269.391	98,8%
RISARALDA	\$ 46.720.626.300	\$ 42.367.235.286	\$ 40.512.508.084	90,7%
CALDAS	\$ 19.697.471.900	\$ 26.987.921.407	\$ 25.376.799.539	137,0%
CAUCA	\$ 22.346.700.100	\$ 23.620.166.334	\$ 22.052.522.840	105,7%
QUINDIO	\$ 8.030.004.200	\$ 9.544.597.341	\$ 8.638.988.666	118,9%
BOGOTA D.C.	\$ 3.611.422.300	\$ 2.643.612.934	\$ 2.850.570.839	73,2%
OTROS DEPARTAMENTOS	\$ 374.246.708.900	\$ 373.320.787.594	\$ 356.165.694.629	99,8%
TOTAL GENERAL	\$ 374.246.708.900	\$ 373.320.787.594	\$ 356.165.694.629	99,8%

Tabla No. 6.16 Resultados de Modelación Costo Medio Anual x Departamento de Residencia

La estimación por departamentos de residencia del Afiliado en especial; para los primeros 5 departamentos muestra que los resultados son muy cercanos al verdadero Costo Ejecutado, el Valle del Cauca que concentra la mayor cantidad de Afiliados que concentra la mayoría de la población afiliada

está realmente cerca al verdadero valor la diferencia es del 1.8%, a pesar de que para el departamento de Caldas se observe una Sobreestimación del Costo del 37%

Agrupacion de Edad	Costo Ejecutado	Costo Modelado	UPC Anual	% Part. Costo_Ejctdo/ Costo Mod
Menores de 15	\$ 50.244.921.100	\$ 54.869.744.603	\$ 56.589.856.456	109,2%
Mayores Iguales a 15	\$ 324.001.787.800	\$ 318.451.042.991	\$ 299.575.838.174	98,3%
Total general	\$ 374.246.708.900	\$ 373.320.787.594	\$ 356.165.694.629	99,8%

Tabla No. 6.16 Resultados de Modelación Costo Medio Anual x Departamento de Residencia

Teniendo en cuenta que en la Modelación se encontró un factor cuadrático en la variable Edad y que se incluyeron variables para tratar de minimizar este efecto, se presenta la tabla 6.16, la agrupación para las Edades menores a 15 años y mayores o iguales a 15, los resultados continúan siendo cercanos al verdadero Costo, por lo cual se confirma que la metodología y Modelación cumplen con las expectativas de Modelar los Costos Medios en un Año los más fiel posible a la realidad.

Al inicio del presente capítulo se planteaba la manera de evaluar el Valor de UPC pagado por el estado, para lo cual se planteó la construcción de Indicadores que puedan informar si realmente el valor de la UPC es suficiente, entre los indicadores que se podrían realizar se encuentra la Siniestralidad término de la teoría del aseguramiento, esta medición consiste en la proporción del Costo Ejecutado vs el Valor de UPC, cuando se observa que el valor del Costo Ejecutado es mayor al valor del Ingreso o UPC el cual se afecta con el descuento del 10% por concepto de gastos administrativos que autoriza a descontarse el Ministerio de la Protección Social, indica que los Costos han sobrepasado los Ingresos, y están generando impacto que puede comprometer la operación de la EPS.

Se observa en las tablas relacionadas anteriormente que el Costo ejecutado es mayor al valor de UPC lo cual generar un indicador por encima del 90% como se mencionó esto coloca en riesgo la operación del negocio, del mismo modo que se evalúan los Costos Ejecutados se evaluara el Costo Medio Anual de la Modelación respecto la suficiencia del valor de la UPC, estos resultados le permitirán a la EPS un criterio para solicitar un cambio de metodología de cálculo para la asignación del valor de UPC que el estado debe pagar, seguidamente se presentan los resultados del Indicador de la Siniestralidad.

Grupo Edad	Costo Ejecutado	Costo Modelado	UPC Anual	% Siniestralidad Costo Ejecutado	% Siniestralidad Costo Modelado
0 - 4 Años	\$24.708.504.700	\$23.242.974.396	\$31.598.404.529	78,2%	73,6%
5 - 9 Años	\$14.588.813.800	\$17.252.138.571	\$13.054.582.066	112%	132,2%
10 - 14 Años	\$10.947.602.600	\$14.374.631.637	\$11.936.869.862	92%	120,4%
15 - 19 Años	\$16.254.318.900	\$15.333.191.232	\$14.799.740.345	110%	103,6%
20 - 24 Años	\$26.383.040.600	\$19.781.957.666	\$25.791.874.632	102%	76,7%
25 - 29 Años	\$33.589.415.400	\$26.912.805.891	\$30.114.254.803	112%	89,4%
30 - 34 Años	\$31.866.511.700	\$30.041.041.680	\$29.511.243.205	108%	101,8%
35 - 39 Años	\$26.833.249.100	\$29.379.844.079	\$24.927.878.901	108%	117,9%
40 - 44 Años	\$27.109.128.900	\$32.862.440.034	\$24.059.178.511	113%	136,6%
45 - 49 Años	\$29.084.266.000	\$34.895.743.750	\$27.171.309.704	107%	128,4%
50 - 54 Años	\$28.043.360.400	\$31.789.184.784	\$27.014.287.727	104%	117,7%
55 - 59 Años	\$25.431.631.500	\$27.417.167.875	\$24.412.306.368	104%	112,3%
60 - 64 Años	\$23.511.912.500	\$21.509.905.816	\$21.370.529.965	110%	100,7%
65 - 69 Años	\$17.168.123.000	\$14.879.865.153	\$15.811.914.071	109%	94,1%
70 - 74 Años	\$13.744.872.200	\$11.629.593.996	\$12.543.467.145	110%	92,7%
75 - 79 Años	\$11.629.058.800	\$9.326.492.351	\$10.604.299.452	110%	88,0%
80 - 84 Años	\$7.573.603.700	\$6.720.515.837	\$6.739.206.913	112%	99,7%
85 y mas Años	\$5.779.295.100	\$5.971.292.847	\$4.704.346.432	123%	126,9%
Total general	\$374.246.708.900	\$373.320.787.594	\$356.165.694.629	105%	104,8%

Tabla No. 6.17 % de Siniestralidad Costo Ejecutado y Costo Modelado por Grupo de Edad

La tabla No. 6.17 relaciona el valor de la siniestralidad respecto a los Costos Ejecutados y Costos Modelados, la totalidad del Costo Medico deja ver la insuficiencia del valor de UPC, en ambos cálculos los resultados están por encima del 90% lo cual ya demuestra la falta de capacidad de los ingresos para cubrir el gasto de la EPS en la atención de sus Afiliados en el Año 2011. Al observar los resultados en los Grupos de Edad entre 5 y 19 años muestran que la EPS presenta mayores Costos que Ingresos para estas edades, esta misma situación se presenta en las edades de 30 a 64 años y los mayores de 80 años de edad. Estos resultados contribuyen a que la EPS en algún momento contemple la posibilidad de solicitar al Ministerio de Protección Social un ajuste o incremento en el Valor de UPC para estos grupos de edad teniendo en cuenta el perfil del riesgo de cada Afiliado, con la demostrado en los cálculos del Modelo de dos Partes la Metodología usada en el presente trabajo sería la que debería adoptar el ente de Control, en general se observa en la tabla No 6.17 que la EPS en el periodo de Atención 2011 presento una relación de Costos vs Ingresos desfavorable y riesgosa, para mantenerse en el Mercado y de paso sea dicho de mantener la continuidad de los servicios médicos a sus Afiliados.

Género	Costo Ejecutado	Costo Modelado	UPC Anual	% Siniestralidad Costo Ejecutado	% Siniestralidad Costo Modelado
Mujeres	\$ 210.999.944.400	\$ 218.889.481.305	\$ 203.752.077.979	103,6%	107,4%
Hombres	\$ 163.246.764.500	\$ 154.431.306.289	\$ 152.413.616.651	107,1%	101,3%
Total general	\$ 374.246.708.900	\$ 373.320.787.594	\$ 356.165.694.629	105,1%	104,8%

Tabla No. 6.18 % de Siniestralidad Costo Ejecutado y Costo Modelado por Genero

El Indicador de la siniestralidad respecto al Genero de los Afiliados (Tabla 6.18) muestra que el valor asignado de Ingreso de UPC en los hombres no alcanza a cubrir lo gastado por estos, en ambos géneros sucede lo mismo el valor de la UPC no le es viable a la EPS para cubrir la demanda de Servicios de Salud.

CAPITULO 7

CONCLUSIONES

Los resultados obtenidos en el análisis exploratorio de los datos permitieron identificar los grupos con mayor concentración de población discriminada por Género, Ciudad, Tipo de afiliado, de igual manera la distribución de los costos, y morbilidades de los afiliados; éste primer acercamiento permitió conocer que para la EPS donde se realizó el estudio existen afiliados cuyo costo en el año, va de \$300 a \$600 millones, a pesar de que es un grupo de 7 afiliados, el valor es muy significativo; situaciones como estas hacen que los ingresos asignados por el estado a las EPS's pierdan la capacidad para garantizar y cubrir la demanda de servicios por parte de los afiliados, la razón del elevado costo se debe a patologías de alta complejidad, que hacen demandar el uso de moléculas y prestaciones muy costosas.

Respecto a los resultados en el Modelo Logístico, se muestra que la edad y el número de Periodos Compensados son factores que incrementan la probabilidad de uso de los servicios, mientras que ser afiliado de género Masculino y Cotizante se convierte en factores que minimizan la probabilidad de uso del servicio. Al observar la curva ROC se tiene que la capacidad de clasificación del modelo es aproximadamente del 76% lo que es considerada como una capacidad de clasificación moderada. La estimación del Costo a través del Modelo Lineal Generalizado muestra que el aporte de los coeficientes de las Covariables Edad, Periodos Compensados, Género y Tipo Afiliado concuerdan con los resultados obtenidos por medio del Modelo de Regresión Logística.

Los resultados del Modelo de dos Partes muestran mínimas diferencias entre los valores ajustados y los valores ejecutados; la metodología estadística del Modelo muestra que es adecuada para la Modelación de los Costos en Salud, tiene en cuenta la probabilidad de uso y la asimetría hacia la derecha que presentan los Costos; la cual es bien interpretada por el Modelo Lineal Generalizado a través de la función distribución Gamma perteneciente a la familia Exponencial. Dados los resultados en el Modelo de dos Partes, se confirma para la EPS, que la metodología y Modelación cumplen con las expectativas de Modelar los Costos Medios en un Año los más fiel posible a la realidad. Por lo cual, la EPS cuenta con evidencia que le puede servir para solicitar un cambio en la metodología para la asignación del Ingreso al Ministerio de la Protección Social. El método de cálculo de Celdas utilizado por el Ministerio de la Protección Social actualmente, está generando una Siniestralidad que impacta la operación de la EPS de manera negativa.

El Supuesto de normalidad se comprobó de manera gráfica obteniendo que el Modelo se comporta de manera adecuada frente a este supuesto. Respecto a el supuesto de varianza homogénea se da un comportamiento asimétrico correspondiente con lo descrito en (Behar Gutierrez, 2003), indicando que esta forma se corresponde con una reducción en la varianza, por lo que no es posible asumir el cumplimiento de este supuesto, resaltando nuevamente como posible fuente de esta situación los valores extremos en los costos de atención.

Como recomendación a futuros estudios la EPS podría considerar otras variables que puedan complementar el cálculo de Costo Medio Anual, asociado con el perfil del riesgo que tenga asociado el afiliado según el grupo de edad al que pertenece entre otros, de igual manera es necesario definir o estructurar la base de datos que contemple las categorías o valores dicotómicos, que deben relacionarse en las nuevas variables a fin evitar tiempos de procesamiento. La metodología implementada para modelar la información de los Costos Médicos de la EPS en el año 2011, debe ser considerada por parte de la EPS como necesaria e innovadora en este tipo de datos, y debe seguir considerándola para futuros proyectos de estimación del Costo de la UPC.

Diccionario de Datos

Variable	Descripción	Uso de la variable	indicador o Análisis con el que se relaciona	Tipo de información que genera
tipo de identificación	tipo de identificación del afiliado	llave principal, identificar persona Verificar derechos, identificación, registros, duplicados	Cobertura, intensidad	
identificación	número de identificación del afiliado según el tipo de identificación			
Fecha de nacimiento	fecha de nacimiento del afiliado	Cálculo de edad y grupos etario y análisis por grupo etario y género.	Análisis por grupo etario y género de: afiliación y estructura demográfica.	Características población, factor de ajuste
Género	Género del grupo de afiliados			
Código departamento	Código del departamento donde reside el afiliado	Análisis por zona geográfica	Análisis por zona geográfica de: afiliación y estructura demográfica	Características regionales, factor de ajuste
Código municipio	Código del municipio donde reside el afiliado			
zona geográfica	zona normal, especial o conurbada donde reside el afiliado			
Días compensados (RC)	Suma de días compensados por todos los afiliados en cada grupo etario y género en el periodo	Cálculo de equivalencia en afiliados	Cobertura % de rotación	Factor de ajuste
Afiliados compensados equivalentes (RC)	usuarios compensados equivalentes por grupo etario y género para el periodo	denominador de la población con derecho a la atención	Estructura población	Características población
tipo de afiliado	tipo de afiliado	Cálculos por tipo de afiliado	Análisis de afiliación	Características de aseguramiento
fecha de afiliación al SGSSS	fecha de afiliación al SGSSS	Cálculo de la antigüedad en el SGSSS	Análisis de afiliación	Características de aseguramiento

Tabla 5.1 variables población del estudio de suficiencia POS – UPC

Variable	Descripción	Uso de la variable	indicador o Análisis con el que se relaciona	Tipo de información que genera
tipo de identificación	tipo de identificación del afiliado	llave principal, identificar persona Verificar derechos, identificación, registros, duplicados	Cobertura, intensidad	
identificación	número de identificación del afiliado según el tipo de identificación			
fecha de nacimiento	fecha de nacimiento del afiliado	Cálculo de edad y grupos etario y análisis por grupo etario y género.	Análisis por grupo etario y género de: afiliación y estructura demográfica.	Características población, factor de ajuste
género usuario	género del grupo de afiliados			
Código departamento	Código del departamento donde reside el afiliado	Análisis por zona geográfica	Análisis por zona geográfica de: afiliación y estructura demográfica	Características regionales, factor de ajuste
Código municipio	Código del municipio donde reside el afiliado			
zona geográfica	zona normal, zona normal, especial o conurbada o especial donde reside el afiliado			
Código del diagnóstico	Código del diagnóstico principal	Perfil epidemiológico Coberturas, tendencias, Variable de ajuste	morbilidad incidencia Prevalencia indicadores de SP modelos de ajuste	Perfil epidemiológico, costo, factores de ajuste
fecha de prestación del servicio	fecha en que fue prestado el servicio	Validación de correspondencia al periodo; identificación registros duplicados; cálculos incidencia; periodicidad de las atenciones o diagnósticos	morbilidad incidencia Prevalencia	Perfil epidemiológico, utilización
Código de actividad, intervención o procedimiento, y medicamentos	Código de actividad, intervención o procedimiento, medicamentos de acuerdo con las tablas de CUPS, mapipos y SOAT y los Acuerdos 228, 236 y 282.	utilización de servicios tendencias	frecuencias de uso Coberturas intensidad de uso tendencias	utilización, costo, ingresos, factores de ajuste
Ámbito de prestación de la actividad, intervención, procedimiento o medicamento	identificador para determinar el ámbito de prestación según la ubicación funcional	utilización de servicios tendencias	frecuencias de uso Coberturas intensidad de uso tendencias	utilización, costo, ingresos, factores de ajuste
forma de reconocimiento y pago de la actividad, intervención o procedimiento, medicamento	Clasificar la actividad o el procedimiento según tipo de forma reconocimiento y pago	utilización de servicios tendencias	frecuencias de uso intensidad de uso tendencias Costo total por forma de reconocimiento y pago	utilización, costo

Variable	Descripción	Uso de la variable	Indicador o Análisis con el que se relaciona	Tipo de información que genera
número de días estancia normal	suma de los días de estancia facturados que no sean en UCI	relación entre procedimientos y estancia Estructura de costos	Costo promedio Valor per cápita	utilización
Valor actividad, intervención y procedimiento, medicamento	Valor reconocido por la Aseguradora al prestador por concepto de la atención prestada	Estructura de costos	Costo promedio Valor per cápita distribución de los costos	Costo
Valor asumido por el usuario (cuota y copago)	Valor asumido por el usuario por concepto de la atención	Estructura de costos	Costo promedio Valor per cápita	Costo
Valor a recobrar	Valor de la factura que corresponde a evento Pos irrecoobrable por concepto de Tutelas y CTC y no reconocido por el FOSYGA	Estructura de costos	Costo promedio, valor per cápita	Costo

Tabla 5.2 variables de prestación del estudio de suficiencia POS – UPC

REFERENCIAS BIBLIOGRAFICAS

Atkinson, A. (1981). Likelihood ratios, posterior odds and information criteria. Elsevier Journal of Econometrics, 16, 15 - 20.

Behar Gutierrez, R. (11 de 2003). Serie Monograficas. Validacion de Supuestos en el Modelo de Regresión. Santiago de Cali, Colombia: Universidad del Valle.

Behar, R. G. (27 de 02 de 2003). Index of /~robehar/Material de apoyo. Obtenido de <http://pino.univalle.edu.co/~robehar/Material%20de%20apoyo/>

Blough, D. K., Madden, C. W., & Hornbrook, M. C. (1999). Modeling risk using generalized linear models. Journal of Health Economics 18, 153–171.

Clavijo, S., Torrente, C., Santamaria, M., & Garcia, F. (2008). El deficit Fiscal de la Salud en Colombia, El Sistema de Salud Colombiano logros y retos despues de quince años de reforma. Carta financiera, 7 a 15, 24 a 30.

Deb, P., Munkin, M. K., & Trivedi, P. K. (2006). Bayesian Analysis Of The Two-Part Model With. Journal of Applied Econometrics 21, 1081–1099.

Dobson, A. J., & Barnett, A. G. (2008). An Introduction to Generalized Linear Modelr. Boca Raton: Chapman & CRC Press.

Domenech, J. M. (2002). Gráficos de Dispersión.

Duan, N., Manning, W. G., Morrisb, C. N., & Newhousea, J. P. (1983). A Comparison of Alternative Models for the Demand for Medical Care. Journal of Business & Economic Statistics, 115-126.

Efroymson, M. A. (1960). "Multiple regression analysis" . Wilf (eds). Mathematical Methods for Digital Computers, 1, 191 - 203.

Fedesarrollo. (2012). La Sostenibilidad Financiera del Sistema de Salud Colombiano - Dinámica del gasto y principales retos de cara al futuro. Bogota: La Imprenta Editores S.A.

Glied, S. (2000). Managed care. In Handbook of Health Economics. Elsevier, 707 - 745.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). The Elements of Statistical Learning; Data Mining, Inference, and Prediction. California: Springer.

Hosmer, D., & Lemeshow, S. (1989). Applied Logistic Regression. New York: Wiley.

Kalmanovitz, S. (2010). Nueva Historia de Economia en Colombia (Pag 257 a Pag 261). Bogota (Colombia): Taurus.

- Langrand, C., & Pinzon, L. M. (2009). *Análisis de datos Metodos y ejemplos*. Bogotá: Escuela Colombiana de Ingeniería.
- Lin, T.-F. (2008). Modifiable health risk factors and medical expenditures – The case of Taiwan. *ELSEVIER Social Science & Medicine*, 1727-1736.
- Manning, W., & Mullahy, J. (2001). Estimating log models: to transform or not to transform? *Health Economics*, 401- 494.
- Maria Elena Rodriguez, E. B. (2001). *Coefficientes de Asociación*.
- McCullagh, P., & Nelder, J. (1989). *Generalized Linear Models Second Edition*. Londres and Harpenden: Chapman and Hall.
- Mihaylova, B., Briggs, A., & O' Hagan, A. (2010). Review of statistical methods for analysing healthcare resources and costs. *Health Economics*, 101-109.
- Mullahy, J. (1998). Much ado about two: reconsidering retransformation and the two-part model in health econometrics. *Journal of Health Economics* 17. 247–281, 247–281.
- Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society*, 370 - 384.
- Rice, N., & Smith, P. C. (2001). Capitation and risk adjustment in health care financing: An international progress report. *The Milbank Quarterly*, 79, 1.
- Silva A., L. C. (2008). *Regresión Logística*. Madrid España: La Muralla.
- Universidad Nacional de Colombia. (2011). *Informe del Cálculo de la Unidad de Pago por Capitación Nota Técnica*. Bogotá: Universidad Nacional de Colombia.
- Universidad Nacional de Colombia, F. d. (2010). *Informe de Avance del Cálculo de la Unidad de Pago por Capitación*. Bogotá: Universidad Nacional de Colombia Facultad de Ciencias Economicas - Facultad de Medicina.
- Wolstenholme, D. E., O'Brien, C. M., & Nelder, J. A. (1988). *GLIMPSE: a knowledge-based front end for statistical analysis*. Elsevier, 173-178.