Pontificia Universidad Católica del Perú

Maestría en Estadística

Trabajo de Aplicación: Predicción de gasto reducido por seguro médico

---

Curso: Modelos lineales 1

Autor:

Lucio Enrique Cornejo Ramírez

Profesores:

Camiz Sergio

Alex de la Cruz Huayanay

Julio 2025

Tabla de Contenidos

Introducción **2**
- Marco del problema . . . 2
- Plan de modelamiento . . . 2

Metodología **3**
- Datos/Observaciones . . . 3
- Itinerario metodológico de la modelización . . . 4

Presentación y análisis de resultados **4**
- Carga de datos . . . 4
- Filtro de variables . . . 5
- Modelamiento . . . 6
- Modelo tras remover puntos aberrantes . . . 7
- Selección reducida de covariables . . . 8
- Modelo final . . . 8

Conclusiones **10**

Introducción

Marco del problema

Entre los costos que más desastibilizan económicamente a las personas, se encuentra el pago por procedimientos médicos. Estos precios pueden variar en gran medida dependiendo de características del paciente, como detallaremos más adelante.

En ese sentido, resulta de gran valor predecir adecuadamente el costo que un seguro médico cubrirá, respecto a un procedimiento médico. Para un paciente, aquella predicción puede servir para que planifique qué tanto sería desestabilizado económicamente debido a algún tipo de procedimiento particular. Por otro lado, también para las aseguradoras resulta útil aquellas predicciones, ya que pueden anticipar qué tanto dinero estarían perdiendo por el monto a cubrir de la operación; además, con ese conocimiento pueden monitorear mejor qué pedidos de cobertura resultan anómalos, potencialmente fraudulentos.

Plan de modelamiento

Anteriormente, hemos planteado como variable por predecir a la cantidad monetaria que la aseguradora de un paciente cubrirá debido a un procedimiento médico. Note que aquel monto está muy relacionado al precio que paga el paciente luego que el seguro descuenta parte del costo del procedimiento médico … ese monto, que denominaremos `charges`, se intentará predecir.

Asimismo, vale recalcar la influencia de los gastos del hospital debido al procedimiento médico, variable que denotaremos `Hospital_expenditure`, sobre `charges`.

Para el modelamiento, se considerará además las siguientes características del paciente:

- Sexo (`age`)
- Si fuma o no (`smoker`)
- Región de la que provee (`region`)
- Edad (`age`)
- Índice de masa corporal (`bmi`)
- Cantidad de hijos e hijas (`children`)
- Costo médico que pagaría en caso no se aplicase seguro médico (`Claim_Amount`)
- Número de procedimientos pasados (`past_consultations`)
- Número de pasos que realizó en cierto día (`num_of_steps`)
- Número de veces que ha sido hospitalizado (`Number_of_past_hospitalizations`)
- Salario anual (`Anual_Salary`)

Metodología

Datos/Observaciones

Las observaciones que consideraremos para este proyecto fueron descargadas de este sitio web.

Estos datos son de distintos pacientes que recibieron algún tipo de tratamiento médico, de los cuales se tienen variables recopiladas, como edad, sexo, si fuma o no, etc. Así, descartamos que los datos consistan de una serie de tiempo.

No obstante, aquella página web no provee información más específica sobre el origen de los datos. Por ejemplo, si han sido recopilados en un único hospital, o en diversos hospitales, pero de qué país, etc.

Aún así, en esta investigación, no solo se considera la predicción de la variable mencionada, sino también cómo es que influyen las variables que emplearemos como regresores, en la predicción final. Por ejemplo, si su relación es directa o inversamente proporcional.

A continuación justificamos el posible uso de los caracteres presentes en los datos, como covariables:

- **Sexo**: Debido al riesgo y costos distintos entre hombres y mujeres, para ciertos tipos de operaciones; por ejemplo, parto.
- **Si fuma o no**: Pues fumar aumenta la probabilidad de desarrollar complicaciones médicas
- **Región de la que provee**: Ya que el costo de un procedimiento médico puede variar mucho por región, así que también varía cuánto cubriría una aseguradora.
- **Edad**: Puesto que pacientes mayores suelen requerir más cuidados.
- **Índice de masa corporal**: En base a que un IMC elevado está asociado a mayores riesgos durante cirugías.
- **Cantidad de hijos e hijas**: Esto puede influir en el tipo de cobertura familiar (de seguro) que tiene el paciente.
- **Costo médico que pagaría en caso no se aplicase seguro médico**: Importante incluirlo, pues incluso se espera que presente una fuerte correlación positiva con la variable por predecir.
- **Número de procedimientos pasados**: Puede resultar útil en base a que pacientes con muchos procedimientos suelen tener enfermedades crónicas, por lo que se esperaría una mayor cobertura.
- **Número de pasos que realizó en cierto día**: Esta variable tampoco se explica en la fuente, pero la podemos considerar como una medida de la condición física de una persona, qué tan activa es.
- **Número de veces que ha sido hospitalizado**: Pues más hospitalizaciones implican mayor riesgo en la operación, aumentando posiblemente así los costos que cubre la aseguradora.
- **Salario anual**: Como indicador de nivel socioeconómico, se espera que pacientes con ingresos altos cuenten con aseguradoras que cubren mayor

parte el costo por intervención médica.

Itinerario metodológico de la modelización

A continuación, describrimos los pasos a seguir para la construcción de diferentes modelos de predicción:

1. Descarte de observaciones que presenten algún valor faltante para cualquier variable.
2. Gráficos de dispersión para pares de variables
3. Debido al máximo establecido en este proyecto, respecto al número de covariables, calculamos las correlaciones múltiples
4. Filtro de observaciones al azar, debido a máximo establecido en este proyecto.
5. Limpieza de datos
6. Construcción del modelo OLS, empleando todas las covariables
7. Gráficos de valores observados y residuos contra valores estimados. Interpretar R^2
8. Emplear el test de Levine y Shapiro para averiguar la homocedasticidad y la normalidad.
9. En caso positivo, evaluar por medio de ANOVA si el modelo tiene sentido. En caso positivo, determinar qué variables explicativas tienen sentido.
10. Si se encuentran puntos aberrantes, recorrer el modelo sin aquellos y repetir los pasos mencionados.
11. Emplear selección por delante, para atrás y stepwise.
12. Ejecutar los tests de tipo ANOVA
13. En caso el test ANOVA relevante resulte positivo, incluir los tests post-hoc.

Presentación y análisis de resultados

Carga de datos

A continuación, mostramos los datos descargados del sitio web mencionado en la sección previa.

Descartamos las observaciones con alguna variable faltante.

Filtro de variables

Graficos de dispersión

Nótese que la variable por predecir, `charges`, parece presentar una relación lineal con la covariable `Annual_Salary`. Asimismo, parece haber indicios de que resulta posible transformar las variables `num_of_steps` y `Hospital_expenditure` por funciones logaritmo y exponcial, respectivamente, con fin que se tenga una fuerte relación lineal entre el predictor creado y la variable por predecir.

Correlaciones parciales entre covariables

Note que cuatro covariables presentan correlación parcial mayor a 0.8, en orden descendente `Anual_Salary`, `Hospital_expenditure`, `num_of_steps` y `Number_of_past_hospitalizations`. Aquellas variables son muy explicadas por las demás (posible multicolinearidad).

Inspeccionemos ahora, de manera particular, las correlaciones entre covariables

Observamos una alta correlación entre `Anual_Salary` y `Hospital_expenditure`, con un valor de 0.9692177 Asimismo, como la variable de salario anual es más sencilla de recopilar (por ejemplo, en una encuesta) que la de gasto de hospital, descartamos la variable cuantitativa `Hospital_expenditure`.

Variable cuantitativa `num_of_steps`

Inicialmente se consideró descartar la variable referente al número de pasos que realizó el paciente en cierto día. Esto pues, a primera vista, no se esperaría que tal información resulte relevante para el costo final por el procedimiento médico.

Graficamos tal posible regreso contra la variable respuesta:

En base a que la relación parece asemejarse a una exponencial, graficamos la variable `num_of_steps` contra el logaritmo de la variable respuesta:

En base a que aquella relación parece ser *aproximadamente* lineal, optamos por no descartar la variable cuantitativa `num_of_steps`.

Variable cuantitativa `age`

Variable cuantitativa `bmi`

Variable cuantitativa `children`

Variable cuantitativa `past_consultations`

Descartamos la variable cuantitativa `children`, pues, en base a este simple análisis inicial, no parece indicar algún de tipo de relación lineal con la variable por predecir. Es más, su gráfico de dispersión parece sugerir que consideremos a la variable `children` como cualitativa.

Variables categóricas

Para el filtro de variables categóricas, descartaremos aquella para la cual las distribuciones de la variable respuesta, respecto a los valores de aquella variable categórica sean relativamente similares.

Variable cualitativa `region`

Inspeccionamos la distribución de la variable respuesta, respecto a los valores de la variable categórica `region`.

En base a que aquellas funciones densidad no presentan una difencia resaltante, descartaremos la variable `region`. De esa manera, las variables cualitativas que emplearemos para esta investigación son solo `sex` y `smoker`.

Variables finales

- Variables cualitativas:
  - `sex`
  - `smoker`
- Variables cuantitativas:
  - `age`
  - `bmi`
  - `Claim_Amount`
  - `past_consultations`
  - `num_of_steps`
  - `Number_of_past_hospitalizations`
  - `Anual_Salary`
  - `charges` (**variable respuesta**)

Para limitarnos a 500 filas, según la restricción de este proyecto, realizaremos un muestreo:

Modelamiento

Note que los tipos de datos de las covariables son adecuadas. En particular, las funciones por emplear se encargarán de la conversión numérica a las covariables categóricas `age` y `sex`.

Comencemos definiendo algunas funciones auxliares en el análisis y modelamiento.

Modelo completo

En base al valor del R^2, note que este modelo explica alrededor del **99.1%** de la varianza de `charges`.

Tests

Prueba de homocedasticidad

En base al p-valor menor a 0.05, se tiene suficiente evidencia de que la **varianza de los residuos no es constante**, es decir, no se cumple la homocedasticidad.

Prueba de normalidad de los errores

En base al p-valor menor a 0.05, contamos con suficiente evidencia para afirmar que los **residuos no siguen una distribución normal**.

Al ejecutar el test para los residuos estandarizados, se llega a la misma conclusión respecto a la normalidad de los residuos.

Como se ha fallado en ambos tests, no estamos en condición formal de aplicar ANOVA. Sin embargo, inspeccionemos su resultado de todas maneras, según la tabla de análisis de la varianza.

Como el p-valor es mucho menor que 0.05, concluimos que este modelo tiene sentido. En particular, existe alguna covariable que explica significativamente la varianza asociada a `charges`.

Por otro lado, según este otro test, solo se puede afirmar para las variables `age`, `num_of_steps`, `Number_of_past_hospitalizations` y `Anual_Salary` que tienen sentido en el modelo.

Puntos aberrantes

Puntos más lejos de 3 promedios

Según aquel criterio, se tienen 38 puntos aberrantes … una cantidad significativa.

Puntos más lejos de cuatro veces los regresores

Según este otro criterio, se tienen 20 puntos aberrantes, también una cantidad significativa.

Mayores apalancamientos: Resaltamos en rojo los 10 puntos con mayor apalancamiento.

Modelo tras remover puntos aberrantes

Note que el porcentaje de varianza explicada aumentó de 99.1% a **99.4%**.

Asimismo, resaltamos que el **residuo promedio es menor** en el modelo tras remover puntos aberrantes.

El p-valor asociado al test de homocedasticidad prácticamente no ha cambiado.

Por otro lado, el p-valor asociado al Test de Shapiro **aumentó significativamente**, de $1.01 * 10^{-5}$ a 0.0379. Este último valor es cercano a 0.05, aunque aún menor, por lo cual se sigue evidenciando la **no normalidad** de los residuos tras haber removido aquellos puntos aberrantes.

En base a estas comparaciones, el modelo tras haber removido los puntos aberrantes resulta **mejor** que el modelo inicialmente construido.

Selección reducida de covariables

Calculamos las correlaciones parciales para estas observaciones.

Note que aún existen covariables con correlación parcial elevada (`num_of_steps`, `Anual_Salary` y `Number_of_past_hospitalizations`), mayor que 0.8; pero ya no existe covariable con correlación parcial mayor a 0.9.

Sin embargo, los valores pequeños en valor absoluto para correlación parcial entre par de covariables implica que existe **poca multicolinealidad** entre las covariables.

Modelo final

Comparando las covariables finales de los tres últimos modelos creados, además de sus coeficientes respectivos, note que se trata de un único modelo.

Asimimo, aquel modelo presenta un R^2 elevado, similar al del previo mejor modelo, también con un valor aproximado a 99.4%.

No obstante, recalcamos que el nuevo modelo presenta un **residuo promedio menor** que el del mejor modelo que habíamos construído hasta ahora.

En ese sentido, el mejor modelo que planteamos es `modelo_both`. Este presenta como covariables a `age`, `Anual_Salary`, `bmi`, `smoker`, `num_of_steps` y `Number_of_past_hospitalizations`.

Prueba de homocedasticidad

Prueba de normalidad de los errores

Para este modelo también se concluye que no se cumple la homocedasticidad y que los residuos no presentan una distribución normal.

A partir de estos dos últimos tests, se concluye que este modelo tiene sentido, pero que la covariable `bmi` no es significativa para ese modelo.

En base a que residuos de este modelo no satisfacen la hipótesis de homocedasticidad ni de distribución normal, no presentaremos los análisis de ANOVA tipo I, II ni III, por tratarse aquellas hipótesis de condiciones necesarias.

Note que el último modelo presentado cuenta con el mayor R^2 ajustado, 99.39%, entre los modelos expuestos. Este criterio refuerza su selección como modelo final para este proyecto.

Entre lo positivo de este estudio, recalcamos que el modelo final presenta un alto valor de R^2. Sin embargo, el hecho que para todos los modelos que creamos se llegó a concluir no homocedasticidad y residuos con distribución no normal,

parece sugerir que el modelo de regresión lineal posiblemente no sea el adecuado para estos datos.

En todo caso, resulta posible que un modelo lineal generalizado resulte más apropiado para el uso con estos datos, en particular debido a la no homocedasticidad encontrada.

Respecto al mejor modelo que presentamos, analicemos la relevancia de las covariables que consideró para su definición:

- `age`: Predictor muy relevante con la variable respuesta, pues, como se mencionó en una sección previa, sucede que paciente de mayor edad suelen requerir más cuidados médicos, su salud está en mayor riesgo, por lo que se espera que la aseguradora cubra más el costo de un tratamiento médico. En efecto, tal es el caso, pues el coeficiente asociado a `age` en el modelo resulta negativo. Así, a mayor edad del paciente, se espera que pague menos (pues la aseguradora cubre mayor costo) por un tratamiento médico.
- `Anual_Salary`: En el caso de pacientes con alto ingreso anual, se espera que sus tratamientos médicos sean también de alto costo. Esto implica que la aseguradora cubra una **menor proporción** del costo médico, pues la cobertura ya resulta alta en base al precio total del procedimiento. Esta relación se hace evidente en el hecho que el coeficiente asociado a `Annual_Salary` es positivo; es decir, a mayor salario anual, menos costo cubre la aseguradora.
- `bmi`: La relación entre esta covariable y la variable por predecir es muy similar la relación de la edad y la variable por predecir. Por ello, el análisis el análogo, y, simplemente recalcamos que el modelo resalta lo esperado (en base al coeficiente negativo asociado a `bmi`), pues, a mayor índice de masa corporal (`bmi`), la aseguradora cubre más del costo, por lo cual el paciente paga menos.
- `smoker`: En el modelo, esta covariable categórica ha sido convertida en 0 y 1; considerando el caso 1 cuando el paciente es fumador. En ese sentido, el coeficiente positivo asociado a esta covariable indice que, si el paciente es fumador, entonces su precio a pagar por tratamiento médico es también mayor, pues la aseguradora cubre **menos** del costo del procedimiento médico.
- `num_of_steps`: Recordemos que esta covariable la interpretamos como un indicador del estado de salud del paciente. Es decir, un mayor valor de `num_of_steps` representa un mejor estado de salud, de actividad física, del paciente. En ese sentido, es coherente que aquella covariable presente un coeficiente positivo según el modelo. Esto pues, mientras más saludable sea una persona, su gasto por procedimiento médico será mayor; es decir, la aseguradora cubrirá una **menor** cantidad del costo del procedimiento.
- `Number_of_past_hospitalizations`: Esta covariable es posiblemente la que más relación se espera tenga con la variable por predecir. El coeficiente negativo asociada a esta covariable es coherente con el hecho que, a mayor

número de hospitalizaciones pasadas, se espera que el seguro cubra una mayor parte del costo del procedimiento médico, por lo cual el gasto del paciente es **menor** por procedimiento médico.

Como no existe una interpretación física, realista, al caso `bmi = 0`, no interpretaremos el intercepto asociado al modelo.

Este modelo se limita a las hipótesis del método MCO. Ahora analizaremos si un modelo del tipo **lineal generalizado** genera mejoras significativas.

Comencemos por una inspección de la homocedasticidad y autocorrelación de los errores del modelo MCO final.

Vía el test de Breusch-Pagan, con un p-valor de $9.796 \cdot 10^{-6}$, existe suficiente evidencia para afirmar el modelo planteado no presenta homocedasticidad. Por otro lado, vía el test de Durbin-Watson, con un p-valor de 0.9838, no se cuenta con suficiente evidencia para rechazar que los errores del modelo son independientes entre sí.

En ese sentido, para los modelos de tipo lineal general por plantear no consideraremos el tipo **AR(1)**. De esa forma, plantearemos modelos del tipo con varianza ponderada, o varianza constante por categoría, o incluso una combinación de ambos modelos.

Asimismo, por medio del test de Breusch-Pagan, con un p-valor de $2.394 \cdot 10^{-12}$, existe suficiente evidencia para afirmar que el modelo completo (considerando todas las covariables) no presenta homocedasticidad. De esa forma, queda descartado el uso de técnicas de regularización como Lasso y Ridge, pues ambas suponen que los errores del modelo tienen varianza constante.

Conclusiones

1. **Validación de hipótesis sobre correlaciones**: Se confirmaron las hipótesis respecto al tipo de correlación (positiva o negativa) entre las covariables del modelo final y el precio por pagar por procedimiento médico.
2. **Capacidad predictiva del modelo**: El último modelo presentado logró explicar el **99.4%** de la varianza en el costo para el paciente por procedimiento médico, demostrando así una capacidad predictiva alta. Este nivel de precisión sugiere que las variables seleccionadas capturan efectivamente los factores determinantes en la cobertura de seguros médicos.
3. **Variables más influyentes identificadas**: El análisis reveló que las variables más significativas para predecir los costos cubiertos son: edad (`age`), salario anual (`Annual_Salary`), índice de masa corporal (`bmi`), hábito de fumar (`smoker`), número de pasos (`num_of_steps`), y número de hospitalizaciones previas (`Number_of_past_hospitalizations`). Estas variables representan factores demográficos, socioeconómicos y de estilo de vida que las aseguradoras consideran en sus decisiones de cobertura.

4. **Limitaciones metodológicas detectadas**: A pesar del alto poder predictivo, el modelo no satisface supuestos importante para la regresión lineal, la homocedasticidad y la normalidad de los residuos. Estas limitaciones sugieren que un modelo lineal simple podría no ser la aproximación más adecuada para estos datos.

5. **Recomendación de modelos alternativos**: Los hallazgos de heterocedasticidad y no normalidad de residuos indican que un **modelo lineal generalizado (GLM)** podría ser más apropiado para este tipo de datos. Estos enfoques alternativos podrían proporcionar estimaciones más confiables, además de intervalos de confianza más precisos.

6. **Necesidad de validación externa**: Aunque el modelo muestra alta precisión en los datos analizados, se recomienda validar estos resultados con datos de diferentes poblaciones, sistemas de salud y contextos geográficos para confirmar su generalización. Recalcamos que no se especifica en la fuente online de estos datos, la proveniencia de las observaciones analizadas.

7. Sin transformar los datos, se require el uso de un modelo que considere heterocedasticidad pero no autocorrelación.