
**Universidad Nacional de Colombia**
**Facultad de Ciencias**

**Departamento de Estadística**

# Análisis de Regresión para la predicción del precio de un automóvil

**Docente:**
Mario Enrique Arrieta Prieto

**Grupo 7**

**Autores:**

| | | |
|---|---|---|
| Andrés Mauricio Rico Parada | Estadística | aricop@unal.edu.co |
| Ander Steven Cristancho Sanchez | Estadística | acristanchos@unal.edu.co |
| Juan David Carrascal Ibañez | Matemáticas | jdcarrascali@unal.edu.co |

**Diciembre 2022**

# 1. Descripción

Este trabajo, de carácter académico, busca afianzar los conocimientos adquiridos en el curso Análisis de Regresión, aplicando algunas de las temáticas vistas en clase en un caso práctico. Para esta tercera y última entrega, se realizará una selección y comparación de diferentes modelos, teniendo en cuenta criterios de información y enfoque de habilidad predictiva para predecir el precio de un automóvil en el mercado estadounidense. Al final, se construirá un modelo de regresión logística con una nueva base de datos sobre datos clínicos.

# 2. Fase de selección automática de variables

a) Inicialmente decidimos utilizar el criterio BIC, pues este critetio penaliza más fuertemente un modelo que tenga muchas variables[1] y consideramos que reducir esta cantidad nos puede permitir trabajar con un modelo más sencillo.

b) Hicimos la estimación de los modelos usando el set de datos completo, descartando únicamente las variables "$CarName$" y "Car_ID", pues estas corresponden a etiquetas de cada observación.Para el método del mejor subconjunto decidimos establecer un tamaño máximo de 9 variables para elegir el mejor modelo. Decidimos esto pues haciendo uso de la función "system.time" encontramos que R tarda 532.47 segundos en hacer este proceso. Para tamaños más grandes de subconjuntos, el algoritmo se ejecutó durante un período mucho más largo de tiempo y no logró terminar la orden. Para las estimaciones requeridas del modelo 4 simplemente usamos las funciones AIC y BIC en el script de la entrega 2.

c)

| Modelo | Método | Número de variables | $R^2$ ajust. | AIC | BIC |
|---|---|---|---|---|---|
| 1 | Forward | 22 | 0.9500111 | 45.70355 | -491.7414 |
| 2 | Backward | 21 | 0.9471112 | 55.57861 | -485.5043 |
| 3 | Mejor subconjunto | 9 | 0.9294914 | 103.7312 | -490.4341 |
| 4 | Entrega 2 | 27 | 0.9401 | 3745.693 | 3842.06 |

- Según la tabla, nos quedaríamos con el modelo por regresión forward, ya que observamos que el modelo de la entrega 2 tiene muchas más variables y un BIC más alto con respecto a los otros modelos. Además, es el único que tiene incluida una transformación no lineal en alguna de sus variables, lo cuál producía dificultades con su interpretabilidad. En términos del criterio AIC tenemos una situación análoga a lo descrito anteriormente con BIC. Respecto a al $R^2$ ajustado, observamos que todos tienen valores semejantes, lo cuál nos da más evidencia para descartar el uso de este valor como bondad de ajuste en los modelos que hemos trabajado.
- Optamos por continuar con el criterio BIC pues AIC produce resultados muy similares y como argumentamos anteriormente, el otro criterio queda completamente descartado.

d)

- Presentamos las estimaciones del mejor "modelo", en nuestro caso, el modelo 1:

| | |
|---|---|
| (Intercept) | -84363.944290 |
| enginelocation_rear | 17278.106766 |
| carlength | -56.139084 |
| carwidth | 1261.551408 |
| curbweight | 9.975033 |
| enginetype_l | -6555.084514 |
| enginetype_ohcv | 622.149230 |
| enginetype_rotor | 1197.568597 |
| cylindernumber_five | -1083.063727 |
| cylindernumber_three | 12578.089294 |
| fuelsystem_2bbl | -84.252810 |
| compressionratio | -123.628123 |
| highwaympg | -4.358241 |
| Marca_Chevrolet | 1988.472676 |
| Marca_Dodge | 140.281895 |
| Marca_Isuzu | 1583.851927 |
| Marca_Jaguar | 2638.213862 |
| Marca_Plymouth | -369.109903 |
| Marca_Porsche | 2740.820232 |
| Marca_Saab | 39.920175 |
| Marca_Volkswagen | -520.919585 |
| fuelsystem_idi | 1269.386612 |
| Marca_Peugeot | 0.001212 |

Estimaciones modelo 1

- Observamos que en el modelo 1 se incluyen categorías de variables que no se consideraron en el modelo de la entrega 2 como: "$carlength$", "$fuelsystem$", "$compressionratio$". Así como menos categorías de la variable "$Marca$".

| Estimación | Modelo 1 | Entrega 2 |
|---|---|---|
| (Intercept) | -84363.944290 | -63434.16 |
| carwidth | 1261.551408 | 1115.80 |
| MarcaJaguar | 2638.213862 | 4486.62 |
| MarcaSaab | 39.920175 | 4132.15 |
| MarcaIsuzu | 1583.851927 | 1827.25 |
| MarcaPorsche | 2740.820232 | 6898.55 |
| MarcaDodge | 140.281895 | 124.00 |
| MarcaPeugeot | 0.001212 | 1437.98 |
| MarcaVolkswagen | -520.919585 | -109.74 |

Comparación estimaciones

- Las estimaciones de los parámetros difiere bastante entre los dos modelos. Sin embargo, para cada variable, los signos son iguales, entonces la interpretación puede ser similar, pero va a cambiar mucho las unidades de cambio para cada variable.

# 3. Fase de evaluación de la habilidad predictiva.

a) El criterio de habilidad predictiva que se eligió es MSE ya que es el criterio de habilidad predictiva con el que más familiaridad se tiene, además, este criterio maximiza los errores grandes, haciendo que si el modelo es muy malo, se vea reflejado en su MSE, y por tanto, permitiendo que sea juzgado de una mejor manera. Se tomará el modelo con menos MSE. Además de esto, se escogió un mecanismo de validación 70-30, una partición de los datos de 70 % para entrenamiento y 30 % para el testeo, debido a que los modelos escogidos tienen hiperparámetros involucrados, por lo tanto, resulta más sencillo realizar su calibración, que se realizó con validación cruzada a 10-etapas con la porción de entrenamiento. Después de calibrar el hiperparámetro, se estimó el modelo con los datos de entrenamiento y se evaluó el error cuadrático medio en la porción de testeo. Sin embargo, cabe resaltar que con este mecanismo la estimación del error en el test puede cambiar de acuerdo a la semilla con la que se realice la partición.

b) La semilla predeterminada que se utilizará para este punto será la semilla 7, que resultará en la partición usada en los siguientes incisos. La partición es presentada en el código adjunto referente a este punto.

c) Se estimaron modelos utilizando selección del mejor subconjunto, el modelo de la entrega dos, el modelo con regresión ridge, el modelo con regresión LASSO y el modelo aditivo general, para este último, después de algunas pruebas se tomó el que mejor dio resultados, además, se realizó un modelo lineal generalizado (Bonus).

- **Modelo del mejor subconjunto:** Las estimaciones de los parámetros para este modelo, estimados sobre la porción de entrenamiento, son las siguientes:

| **Best Subset** | Estimate | Std. Error | t value | Pr($>\|t\|$) |
|---|---|---|---|---|
| (Intercept) | -6.402e+04 | 1.122e+04 | -5.707 | 7.05e-08 *** |
| horsepower | 4.419e+01 | 8.599e+00 | 5.139 | 9.57e-07 *** |
| carwidth | 9.532e+02 | 1.964e+02 | 4.852 | 3.34e-06 *** |
| curbweight | 3.297e+00 | 9.764e-01 | 3.377 | 0.000959 *** |
| Marca_Bmw | 1.165e+04 | 8.864e+02 | 13.139 | $<$ 2e-16 *** |
| Marca_Buick | 8.509e+03 | 1.782e+03 | 4.774 | 4.66e-06 *** |
| Marca_Jaguar | 8.745e+03 | 2.283e+03 | 3.831 | 0.000196 *** |
| cylindernumber_eight | 5.652e+03 | 1.841e+03 | 3.071 | 0.002587 ** |
| enginelocation_rear | 2.071e+04 | 2.305e+03 | 8.985 | 2.11e-15 *** |

- **Modelo Entrega 2:** Las estimaciones de los parámetros sobre la porción de entrenamiento son las siguientes:

| **Entrega 2** | Estimate | Std. Error | t value | Pr(> \|t\|) |
|---|---|---|---|---|
| (Intercept) | -7.271e+04 | 1.383e+04 | -5.257 | 6.84e-07 *** |
| I(carwidth) | 1.266e+03 | 2.146e+02 | 5.900 | 3.73e-08 *** |
| Marca_Audi | -6.527e+02 | 1.890e+03 | -0.345 | 0.73040 |
| Marca_Bmw | 1.077e+04 | 1.116e+03 | 9.656 | < 2e-16 *** |
| Marca_Buick | 8.070e+02 | 2.785e+03 | 0.290 | 0.77253 |
| Marca_Honda | 4.230e+02 | 7.564e+02 | 0.559 | 0.57707 |
| Marca_Mazda | 3.786e+02 | 9.047e+02 | 0.419 | 0.67632 |
| Marca_Mitsubishi | -1.072e+03 | 8.401e+02 | -1.276 | 0.20446 |
| Marca_Jaguar | 3.691e+03 | 2.691e+03 | 1.372 | 0.17277 |
| Marca_Nissan | -4.112e+02 | 7.200e+02 | -0.571 | 0.56905 |
| Marca_Peugeot | 4.166e+01 | 1.128e+03 | 0.037 | 0.97060 |
| Marca_Plymouth | -2.669e+02 | 9.811e+02 | -0.272 | 0.78605 |
| Marca_Porsche | 9.713e+02 | 3.822e+03 | 0.254 | 0.79986 |
| Marca_Saab | 3.662e+03 | 1.106e+03 | 3.311 | 0.00124 ** |
| Marca_Subaru | 9.454e+01 | 8.548e+02 | 0.111 | 0.91213 |
| Marca_Dodge | 5.085e+01 | 1.172e+03 | 0.043 | 0.96548 |
| Marca_Volkswagen | -6.989e+02 | 9.621e+02 | -0.726 | 0.46908 |
| Marca_Volvo | 3.498e+03 | 1.095e+03 | 3.196 | 0.00180 ** |
| I(enginesize$^2$) | 2.325e-01 | 4.292e-02 | 5.417 | 3.37e-07 *** |
| Marca_Isuzu | 1.957e+03 | 1.654e+03 | 1.183 | 0.23916 |
| carbody_hatchback | 3.173e+02 | 1.133e+03 | 0.280 | 0.78001 |
| carbody_wagon | 4.222e+02 | 1.218e+03 | 0.347 | 0.72956 |
| carbody_convertible | 4.100e+03 | 1.930e+03 | 2.124 | 0.03582 * |
| carbody_sedan | 6.607e+02 | 1.122e+03 | 0.589 | 0.55712 |
| cylindernumber_eight | 1.700e+03 | 2.937e+03 | 0.579 | 0.56374 |
| cylindernumber_four | -3.816e+03 | 1.279e+03 | -2.983 | 0.00349 ** |
| cylindernumber_six | -2.990e+03 | 1.637e+03 | -1.826 | 0.07039 . |
| enginelocation_rear | 1.661e+04 | 5.149e+03 | 3.226 | 0.00164 ** |

- **Modelo con regresión ridge:** Para este modelo, las estimaciones, al ser demasiadas, se prefiere mostrar junto a las del modelo con regresión LASSO, para comodidad del lector.
- **Modelo con regresión LASSO:** Mismo caso que con regresión ridge, debido a que como estos modelos no se ven afectados por la multicolinealidad, ajusta un modelo sobre todas las 52 variables disponibles.

| **Ridge** | Estimate |
|---|---|
| (Intercept) | -64798.65 |
| enginetype_dohcv | -9646.31 |
| enginetype_ohcv | -2566.13 |
| Marca_Mitsubishi | -1643.07 |
| Marca_Renault | -1220.46 |
| Marca_Plymouth | -1146.16 |
| enginetype_dohc | -1137.12 |
| Marca_Mercury | -1087.49 |
| Marca_Dodge | -995.72 |
| aspiration_std | -906.50 |
| Marca_Chevrolet | -650.94 |
| Marca_Subaru | -575.23 |
| drivewheel_fwd | -570.92 |
| Marca_Honda | -525.66 |
| Marca_Peugeot | -499.81 |
| carbody_wagon | -373.49 |
| carbody_hatchback | -324.02 |
| cylindernumber_four | -313.68 |
| drivewheel_rwd | -158.71 |
| Marca_Volkswagen | -155.87 |
| carlength | -50.05 |
| enginetype_l | -26.91 |
| peakrpm | 1.83 |
| curbweight | 3.50 |
| horsepower | 21.02 |
| enginesize | 44.92 |
| highwaympg | 88.77 |
| Marca_Nissan | 94.37 |
| carbody_sedan | 130.07 |
| wheelbase | 144.10 |
| carwidth | 652.66 |
| Marca_Volvo | 672.92 |
| enginetype_ohc | 731.18 |
| aspiration_turbo | 785.98 |
| Marca_Mazda | 990.69 |
| Marca_Isuzu | 1008.47 |
| enginetype_ohcf | 1348.10 |
| cylindernumber_two | 1465.23 |
| cylindernumber_twelve | 1480.86 |
| enginetype_rotor | 1562.32 |
| Marca_Audi | 1977.99 |
| cylindernumber_six | 2044.79 |
| carbody_convertible | 2865.07 |
| Marca_Saab | 2966.97 |
| cylindernumber_three | 4349.24 |
| Marca_AlfaRomeo | 4387.26 |
| Marca_Buick | 5800.89 |
| Marca_Porsche | 5861.04 |
| Marca_Jaguar | 6983.98 |
| enginelocation_rear | 7447.14 |
| Marca_Bmw | 8353.31 |
| cylindernumber_eight | 8960.41 |

| **LASSO** | Estimate |
|---|---|
| (Intercept) | -51618.41 |
| enginetype_dohcv | -2502.72 |
| cylindernumber_four | -1195.56 |
| Marca_Mitsubishi | -1027.18 |
| Marca_Renault | -1019.62 |
| aspiration_std | -1015.68 |
| enginetype_ohcv | -641.93 |
| drivewheel_fwd | -340.26 |
| Marca_Plymouth | -266.15 |
| carbody_wagon | -220.30 |
| Marca_Peugeot | -111.58 |
| Marca_Dodge | -31.38 |
| Marca_Nissan | -22.76 |
| Marca_Subaru | -18.76 |
| highwaympg | 0.00 |
| carlength | 0.00 |
| Marca_Chevrolet | 0.00 |
| Marca_Honda | 0.00 |
| Marca_Mercury | 0.00 |
| Marca_Volkswagen | 0.00 |
| carbody_hatchback | 0.00 |
| cylindernumber_six | 0.00 |
| cylindernumber_twelve | 0.00 |
| enginetype_dohc | 0.00 |
| enginetype_l | 0.00 |
| enginetype_ohc | 0.00 |
| enginetype_ohcf | 0.00 |
| drivewheel_rwd | 0.00 |
| aspiration_turbo | 0.00 |
| peakrpm | 0.84 |
| curbweight | 2.49 |
| horsepower | 13.68 |
| wheelbase | 35.11 |
| enginesize | 44.10 |
| enginetype_rotor | 142.99 |
| carbody_sedan | 283.67 |
| Marca_Mazda | 530.33 |
| carwidth | 671.68 |
| Marca_Isuzu | 703.20 |
| cylindernumber_two | 1189.77 |
| Marca_Audi | 1253.51 |
| Marca_Volvo | 1360.24 |
| Marca_AlfaRomeo | 1578.70 |
| Marca_Saab | 1914.47 |
| cylindernumber_three | 2036.64 |
| carbody_convertible | 2713.67 |
| cylindernumber_eight | 4763.90 |
| Marca_Porsche | 4776.94 |
| Marca_Buick | 6275.54 |
| Marca_Jaguar | 6426.32 |
| Marca_Bmw | 9280.38 |
| enginelocation_rear | 10773.21 |

- **Regresión GAM:** Después de calibrar con varios for los grados de libertad para las variables del modelo del mejor subconjunto para habilidad predictiva, y agregando la variable Marca_Volvo ya que nos dio buenos resultados, el modelo GAM que mejor MSE dió fue el siguiente.

| **GAM** | Estimate |
|---|---|
| (Intercept) | 3429.85 |
| ns(curbweight, 2)1 | 12014.49 |
| ns(curbweight, 2)2 | 8152.67 |
| ns(carwidth, 2)1 | 7785.75 |
| ns(carwidth, 2)2 | 9721.77 |
| ns(horsepower, 13)1 | 814.37 |
| ns(horsepower, 13)2 | 1808.99 |
| ns(horsepower, 13)3 | 3201.81 |
| ns(horsepower, 13)4 | -2641.17 |
| ns(horsepower, 13)5 | 2189.42 |
| ns(horsepower, 13)6 | -3190.82 |
| ns(horsepower, 13)7 | 6848.10 |
| ns(horsepower, 13)8 | -931.76 |
| ns(horsepower, 13)9 | 6631.81 |
| ns(horsepower, 13)10 | 53.37 |
| ns(horsepower, 13)11 | 14168.30 |
| ns(horsepower, 13)12 | 8117.32 |
| ns(horsepower, 13)13 | 653.38 |
| Marca_Bmw | 9729.25 |
| Marca_Buick | 3175.23 |
| Marca_Jaguar | 5254.90 |
| Marca_Volvo | 1428.00 |
| cylindernumber_eight | 8649.96 |
| enginelocation_rear | 17401.82 |

d)

| Modelo | Método | N. de variables | $R^2$ ajust. | AIC | BIC | MSE |
|---|---|---|---|---|---|---|
| 1 | Mejor subconjunto | 8 | 0.9297432 | 25.75642 | -343.3772 | 5356094 |
| 2 | Entrega 2 | 27 | 0.9287 | 2219.014 | 2712.753 | 6298335 |
| 3 | Ridge | 51 | 0.9606541 | -12507370325 | -12507370155 | 1634678 |
| 4 | LASSO | 51 | 0.9530696 | -12408622048 | -12408621925 | 2113738 |
| 5 | GAM | 9 | 0.9499116* | 2573.32 | 2647.391 | 6175728 |
| 6 | Bonus1 | 9 | 0.9244404* | 2536.005 | 2568.597 | 5305869 |
| 7 | Bonus2 | 9 | 0.9259998* | 2533.014 | 2565.605 | 260636298 |

Se puede ver que el modelo de la entrega 2 no es tan bueno en habilidad predictiva que tiene de los MSE más altos y en término de los otros criterios tampoco sobresale respecto a los otros modelos, es preferible usar otro modelo. Los mejores en habilidad predictiva son la regresión Ridge y Lasso, que solucionan los problemas multicolinealidad, y con lambdas óptimos, permiten que aunque sean sesgados los estimadores, sean mejores en términos de MSE que los de una regresión lineal simple. También, se puede observar que el GLM con función de enlace identidad supera en términos de habilidad predictiva a la mayoría de los otros modelos, menos los métodos de regularización Ridge y Lasso.
El modelo lineal generalizado con función de enlace log es mejor en los criterios de información que la función de enlace identidad, pero sorpresivamente, en MSE este modelo es el más alto de todos.
Se puso en asterisco el $R^2$ ajustado de los modelos GAM, Bonus 1 y Bonus 2, debido a que para estos, sólo se pudo calcular el Pseudo-$R^2$ ajustado.

**BONUS: Modelo lineal generalizado:** La variable precio es continua, mayor a cero y tiene sesgo a la derecha, por lo tanto, se hará uso de la familia Gamma para el modelo lineal generalizado. Se estimarán dos modelos con las funciones de enlace identidad y logaritmo. Las variables seleccionadas para estos modelos serán las seleccionadas en el modelo mejor subconjunto para habilidad predictiva, y agregando la variable Marca_Volvo pues dio buenos resultados al hacer pruebas.

| **Identity** | Estimate | Std. Error | z value | Pr(> \|z\|) |
|---|---|---|---|---|
| (Intercept) | -4.350e+04 | 8.905e+03 | -4.884 | 2.93e-06 *** |
| horsepower | 3.765e+01 | 8.139e+00 | 4.626 | 8.74e-06 *** |
| carwidth | 4.007e+00 | 7.444e-01 | 5.383 | 3.21e-07 *** |
| curbweight | 6.216e+02 | 1.536e+02 | 4.046 | 8.78e-05 *** |
| Marca_Bmw | 1.090e+04 | 1.500e+03 | 7.264 | 2.84e-11 *** |
| Marca_Buick | -7.985e+03 | 3.402e+03 | 2.347 | 0.02040 * |
| Marca_Jaguar | 9.564e+03 | 4.936e+03 | 1.938 | 0.05478 . |
| Marca_Volvo | 2.425e+03 | 1.181e+03 | 2.054 | 0.04191 * |
| cylindernumber_eight | 7.718e+03 | 3.650e+03 | 2.115 | 0.03631 * |
| enginelocation_rear | 2.111e+04 | 5.669e+03 | 3.723 | 0.00029 *** |

| **Log** | Estimate | Std. Error | z value | Pr(> \|z\|) |
|---|---|---|---|---|
| (Intercept) | 4.4782651 | 0.7929609 | 5.648 | 9.45e-08 *** |
| horsepower | 0.0028551 | 0.0006076 | 4.699 | 6.44e-06 *** |
| carwidth | 0.0004278 | 0.0000698 | 6.129 | 9.38e-09 *** |
| curbweight | 0.0521339 | 0.0138804 | 3.756 | 0.000257 *** |
| Marca_Bmw | 0.5266401 | 0.0632357 | 8.328 | 8.81e-14 *** |
| Marca_Buick | 0.2725324 | 0.1263575 | 2.157 | 0.032817 * |
| Marca_Jaguar | 0.0325696 | 0.1625323 | 0.200 | 0.841483 |
| Marca_Volvo | 0.1156395 | 0.0664632 | 1.740 | 0.084189 . |
| cylindernumber_eight | -0.0329829 | 0.1302755 | -0.253 | 0.800521 |
| enginelocation_rear | 0.8636211 | 0.1629968 | 5.298 | 4.72e-07 |

e) El mejor modelo en términos de MSE resulto ser el modelo que utiliza regresión ridge, por lo que se estimó sobre todo el dataset, las estimaciones son las siguientes:

| **Ridge** | Estimate |
|---|---|
| (Intercept) | -64798.65 |
| enginesize | 44.92 |
| wheelbase | 144.10 |
| horsepower | 21.02 |
| carwidth | 652.66 |
| curbweight | 3.50 |
| highwaympg | 88.77 |
| peakrpm | 1.83 |
| carlength | -50.05 |
| Marca_AlfaRomeo | 4387.26 |
| Marca_Audi | 1977.99 |
| Marca_Bmw | 8353.31 |
| Marca_Buick | 5800.89 |
| Marca_Chevrolet | -650.94 |
| Marca_Dodge | -995.72 |
| Marca_Honda | -525.66 |
| Marca_Isuzu | 1008.47 |
| Marca_Jaguar | 6983.98 |
| Marca_Mazda | 990.69 |
| Marca_Mercury | -1087.49 |
| Marca_Mitsubishi | -1643.07 |
| Marca_Nissan | 94.37 |
| Marca_Peugeot | -499.81 |
| Marca_Plymouth | -1146.16 |
| Marca_Porsche | 5861.04 |
| Marca_Renault | -1220.46 |

| **Ridge** | Estimate |
|---|---|
| Marca_Saab | 2966.97 |
| Marca_Subaru | -575.23 |
| Marca_Volkswagen | -155.87 |
| Marca_Volvo | 672.92 |
| carbody_convertible | 2865.07 |
| carbody_hatchback | -324.02 |
| carbody_sedan | 130.07 |
| carbody_wagon | -373.49 |
| cylindernumber_eight | 8960.41 |
| cylindernumber_four | -313.68 |
| cylindernumber_six | 2044.79 |
| cylindernumber_three | 4349.24 |
| cylindernumber_twelve | 1480.86 |
| cylindernumber_two | 1465.23 |
| aspiration_std | -906.50 |
| aspiration_turbo | 785.98 |
| enginetype_dohc | -1137.12 |
| enginetype_dohcv | -9646.31 |
| enginetype_l | -26.91 |
| enginetype_ohc | 731.18 |
| enginetype_ohcf | 1348.10 |
| enginetype_ohcv | -2566.13 |
| enginetype_rotor | 1562.32 |
| drivewheel_fwd | -570.92 |
| drivewheel_rwd | -158.71 |
| enginelocation_rear | 7447.14 |

| **Entrega2** | Estimate |
|---|---|
| (Intercept) | -63434.16 |
| I(carwidth) | 1115.80 |
| Marca_Audi | 1801.26 |
| Marca_Bmw | 10181.33 |
| Marca_Buick | 6777.28 |
| Marca_Honda | 379.05 |
| Marca_Mazda | 745.85 |
| Marca_Mitsubishi | -394.80 |
| Marca_Jaguar | 4486.62 |
| Marca_Nissan | -596.06 |
| Marca_Peugeot | 1437.98 |
| Marca_Plymouth | -56.82 |
| Marca_Porsche | 6898.55 |
| Marca_Saab | 4132.15 |
| Marca_Subaru | -219.14 |
| Marca_Dodge | 124.00 |
| Marca_Volkswagen | -109.74 |
| Marca_Volvo | 3690.80 |
| I(enginesize^2) | 0.19 |
| Marca_Isuzu | 1827.25 |
| carbody_hatchback | -367.26 |
| carbody_wagon | -271.41 |
| carbody_convertible | 4594.69 |
| carbody_sedan | -130.02 |
| cylindernumber_eight | 199.66 |
| cylindernumber_four | -2264.92 |
| cylindernumber_six | 265.87 |
| enginelocation_rear | 9443.70 |

Las diferencias más notorias entre el modelo con regresión ridge y el modelo de la entrega 2, es la cantidad de variables, pues el modelo con regresión ridge, al no sufrir de problemas de multicolinealidad, estima sobre todas las variables. Por otra parte, en la entrega dos esto no fue posible, ya que el dataset tenía serios problemas de multicolinealidad, haciendo que uno de los mejores modelos que se pudieron hallar tenga 28 variables y no las 52. Cabe resaltar que en el caso de la regresión Ridge, no hay ninguna transformación en alguna variable, cosa que sí pasa en el modelo de la entrega 2, donde engine size está al cuadrado, lo que simplifica nuestro modelo.

Además de esto, no perdemos mucha interpretación, pero las estimaciones si cambian notoriamente en algunos casos, por ejemplo, en el caso de la estimación relacionada a cylindernumber_eight, que pasó de ser de 199,66 a 8960,41. Esto se da en varias variables, por lo que las estimaciones cambian notoriamente, lo cual es lógico por lo dicho anteriormente, pues la regresión toma en cuenta más variables.

# 4. Fase de construcción de un modelo con respuesta binaria

a) Para esta fase decidimos trabajar con una base de datos diferente a la propuesta originalmente. pues debido a su alta cantidad de variables y la información contenida en ellas, los modelos que estimamos presentaban bastantes problemas.

El nuevo data set contiene información sobre datos clínicos recolectada por el Instituto de diabetes y enfermedades digestivas y renales de pacientes en la India. Todos los pacientes son mujeres de al menos 21 años de edad. Las variables encontradas en este data set son las siguientes:

- **Pregnancies:** Número de embarazos del paciente. Cuantitativa.
- **Glucose:** Concentración de glucosa en la sangre después de dos horas de someterse a un test para

medir esta variable. Cuantitativa.

- **BloodPressure:** Presión diastolica medida en (mm Hg). Cuantitativa.
- **SkinThickness:** Grosor del pliegue cutáneo del tríceps. Cuantitativa.
- **Insulin:** Insulina en sangre medida cada dos horas. Cuantitativa.
- **BMI:** Indice de masa corporal. Cuantitativa.
- **DiabetesPedigreeFunction:** Valor númerico que mide el riesgo de contraer diabetes basado en la historia familiar del paciente. Cuantitativa.
- **Age:** Edad del paciente. Cuantitativa.
- **Outcome:** Variable binaria que determina si un paciente tiene diabetes (1) o no (0). Cualitativa

Elejimos la variable **Outcome**, pues en este caso nos interesa predecir si un paciente de este grupo poblacional padece diabetes o no, claramente esta variable codifca esta respuesta.
b)Después de estimar un modelo de respuesta Bernoulli, usando todo el set de datos y posteriormente usando un método de selección automática (backward). Obtenemos un modelo con la siguiente estructura:

$$\begin{cases} Y_k \sim Ber(\pi_k); \pi_k \in (0,1) \\ g(\pi_k) = \beta_0 + \beta_1 x_{k1} + \beta_2 x_{k2} + \beta_3 x_{k3} + \cdots + \beta_7 x_{k7} \\ Y_1, Y_2, ..., Y_n inid \end{cases}$$

Donde únicamente descartamos la variable **Skinthickness** de las que se encontraban originalmente en el set de datos.

c)

| Modelo | Método | P-$R^2$ ajust. | AIC | BIC |
|---|---|---|---|---|
| 1 | Logit | 0.26 | 741.45 | 783.24 |
| 2 | Probit | 0.27 | 739.45 | 776.60 |
| 3 | Cloglog | 0.25 | 751.77 | 788.92 |

d) Debido a que obtuvo un mejor desempeño en todos los criterios de información, optamos por el modelo con función de enlace "probit". Las estimaciones de sus paramétros son las siguientes:

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -6.1117 | 0.4864 | -12.57 | 0.0000 |
| Pregnancies | 0.0832 | 0.0216 | 3.85 | 0.0001 |
| Glucose | 0.0243 | 0.0024 | 10.30 | 0.0000 |
| BloodPressure | -0.0107 | 0.0035 | -3.07 | 0.0021 |
| Insulin | -0.0008 | 0.0005 | -1.50 | 0.1337 |
| BMI | 0.0654 | 0.0100 | 6.55 | 0.0000 |
| DiabetesPedigreeFunction | 0.3310 | 0.1928 | 1.72 | 0.0860 |
| Age | 0.0091 | 0.0066 | 1.38 | 0.1679 |

Procedemos a interpretar 5 de sus paramétros:

- $\hat{\beta}_1$(Pregnacies)= 0.0832: Por cada embarazo que experimente el paciente la probabilidad de padecer de diabetes aumenta .
- $\hat{\beta}_2$(Glucose)=0.0243: Por un aumento de 1 g/L (gramo por litro) en la concentración de glucosa medida en el test, la probabilidad de padecer de diabetes aumenta.

- $\hat{\beta}_3$(BloodPressure)=-0.0107: Por una aumento de 1 mm Hg (milímetro de mercurio) en la concentración de glucosa medida en el test, la probabilidad de padecer de diabetes disminuye..
- $\hat{\beta}_4$(Insulin)=-0.0008:Por un aumento de una microunidad por mililitro ($\mu U/ml$) de insulina en sangre, la probabilidad de padecer de diabetes disminuye.
- $\hat{\beta}_6$(DiabetesPedigreeFunction)=0.3310: Por un aumento de una unidad de riesgo en el puntaje dado por la función "DiabetesPedigreeFunction", la probabilidad de padecer de diabetes aumenta.

e) Únicamente haremos uso de los criterios "precisión" y "exhaustividad". No usaremos la tasa de error aparente pues algunas categorías tienen frecuencias desbalanceadas.
Para $\tau = 0,2$:

| | $y = 0$ (fracaso) | $y = 1$ (éxito) |
|---|---|---|
| $\hat{\pi}_k \leq \tau$ | 279 | 28 |
| $\hat{\pi}_k > \tau$ | 221 | 240 |

Matriz de confusión $\tau = 0,2$

| | |
|---|---|
| Precisión | 0,52 |
| Exhaustividad | 0,896 |

Indicadores bondad de ajuste

Para $\tau = 0,5$:

| | $y = 0$ (fracaso) | $y = 1$ (éxito) |
|---|---|---|
| $\hat{\pi}_k \leq \tau$ | 445 | 111 |
| $\hat{\pi}_k > \tau$ | 55 | 157 |

Matriz de confusión $\tau = 0,5$

| | |
|---|---|
| Precisión | 0,74 |
| Exhaustividad | 0,59 |

Indicadores bondad de ajuste

Para $\tau = 0,7$:

| | $y = 0$ (fracaso) | $y = 1$ (éxito) |
|---|---|---|
| $\hat{\pi}_k \leq \tau$ | 478 | 168 |
| $\hat{\pi}_k > \tau$ | 22 | 100 |

Matriz de confusión $\tau = 0,7$

| | |
|---|---|
| Precisión | 0,81 |
| Exhaustividad | 0,37 |

Indicadores bondad de ajuste

Si elegimos el indicador "precisión" el valor óptimo de $\tau$ será $\tau = 0,7$. En caso de elegir "exhaustividad" será $\tau = 0,2$.

**Bonus**

a) Observamos que 500 observaciones tienen outcome 0, es decir, 65,1 % de los pacientes de la muestra no

padecen de diabetes. De esta forma, consideramos que la muestra no parece estar balanceada. Por ello, obtamos por implementar una técnica de muestreo estratificado para mitigar este desbalance. Además, hacemos la partición del set de datos tomando un 70 % para datos de entrenamiento y el otro 30 % para el testeo.

b) Obtenemos los siguientes resultados al estimar el modelo obtenido en c) con los datos de entrenamiento.

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | -0.9408 | 0.1029 | -9.14 | 0.0000 |
| Pregnancies | 0.0195 | 0.0060 | 3.27 | 0.0012 |
| Glucose | 0.0053 | 0.0006 | 8.79 | 0.0000 |
| BloodPressure | -0.0022 | 0.0010 | -2.24 | 0.0257 |
| Insulin | -0.0001 | 0.0002 | -0.74 | 0.4575 |
| BMI | 0.0140 | 0.0023 | 6.05 | 0.0000 |
| DiabetesPedigreeFunction | 0.2266 | 0.0536 | 4.23 | 0.0000 |
| Age | 0.0055 | 0.0018 | 3.05 | 0.0024 |

c)Obtenemos la siguiente tabla para las estimaciones de la sucesión propuesta.

| tau | AER | precision | recall | F1 |
|---|---|---|---|---|
| 0.00 | 0.55 | 0.39 | 1.00 | 0.56 |
| 0.10 | 0.47 | 0.42 | 0.98 | 0.59 |
| 0.20 | 0.38 | 0.48 | 0.93 | 0.63 |
| 0.30 | 0.31 | 0.54 | 0.84 | 0.66 |
| 0.40 | 0.26 | 0.61 | 0.73 | 0.67 |
| 0.50 | 0.24 | 0.69 | 0.56 | 0.62 |
| 0.60 | 0.26 | 0.74 | 0.42 | 0.54 |
| 0.70 | 0.29 | 0.76 | 0.23 | 0.36 |
| 0.80 | 0.33 | 0.64 | 0.11 | 0.19 |
| 0.90 | 0.34 | 0.67 | 0.07 | 0.13 |
| 1.00 | 0.35 | 0.50 | 0.04 | 0.07 |

De donde, el valor obtenido de $\tau$ con mayor F1: $\tau = 0,4$ con F1=0.67