Lista de ejercicios 3 - Análisis de regresión

Ciclo nivelación 2025-2

Mg. Sc. J. Eduardo Gamboa U.

Una empresa de análisis inmobiliario desea modelar el precio de venta de departamentos en una ciudad usando información estructural y comercial. Se construyó una base de datos de 100 propiedades vendidas recientemente.

Variable respuesta: Precio de venta (miles de USD)

Variables explicativas:

- x1: Área construida, en metros cuadrados
- x2: Índice de calidad constructiva (z-score), es un puntaje técnico de materiales y acabados.
- x3: Intensidad de publicidad digital (z-score), es una medida de exposición en portales y anuncios.
- x4: Tiene cochera (0: No, 1: Sí)
- x5: Zona de la ciudad (A = Zona periférica, B = Zona intermedia, C = Zona comercial, D = Zona premium)

Objetivos:

- Identificar leverages, residuales y valores influenciales
- Seleccionar variables
- Reportar el modelo resultante e interpretar sus coeficientes estimados.
- Presentar un primer alcance sobre el uso de testing set.

1. División del conjunto de datos en entrenamiento y prueba

```
library(readxl)
datos = read_xlsx('Lista3_datos.xlsx')
```

Divide los datos en 2 particiones: entrenamiento (75%) y prueba (25%), usando la semilla 12369874. Como resultado, obtenemos 375 observaciones en el (sub)conjunto de entrenamiento y 125 en el de prueba.

```
library(rsample)
set.seed(12369874)
split_obj = initial_split(datos, prop = 0.75)
train     = training(split_obj)
test      = testing(split_obj)
train |> nrow()
```

```
[1] 375
```

```
test |> nrow()
```

```
[1] 125
```

2. Creación del modelo inicial de regresión lineal

El modelo se construye con los datos de entrenamiento.

```
modelo = lm(y ~ x1 + x2 + x3 + x4 + x5, data = train)
modelo |> summary()
```

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-11.7807  -1.0193  -0.1207   0.8529  13.9607

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.55013    0.42123  13.176  < 2e-16 ***
x1           1.35969    0.04177  32.553  < 2e-16 ***
x2          -0.55100    0.06354  -8.672  < 2e-16 ***
x3           0.02829    0.10227   0.277 0.782246
x41          0.98436    0.20580   4.783 2.51e-06 ***
x5B          0.77653    0.27400   2.834 0.004851 **
x5C         -1.02727    0.26210  -3.919 0.000106 ***
x5D          1.18452    0.32541   3.640 0.000312 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.976 on 367 degrees of freedom
Multiple R-squared:  0.9966,    Adjusted R-squared:  0.9965
F-statistic: 1.532e+04 on 7 and 367 DF,  p-value: < 2.2e-16
```

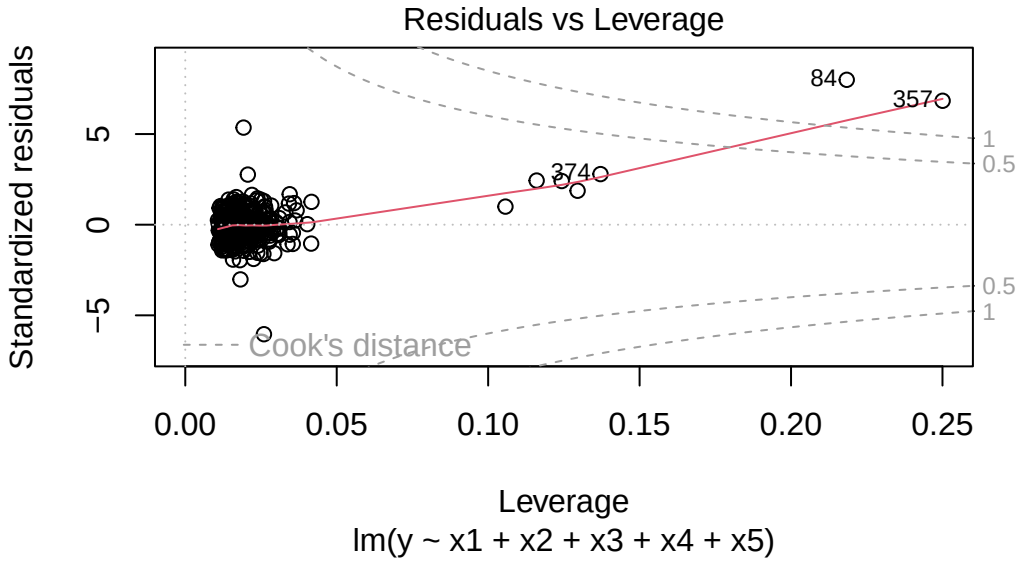
3. Detección de leverages

Se crea un vector con las observaciones que son leverage en el modelo, usando el criterio $h_{ii} > \frac{2k}{n}$

```
modelo |> model.matrix() -> X
X %*% solve(t(X) %*% X) %*% t(X) -> H
train |> nrow() -> n
modelo |> coef() |> length() -> k
as.vector(which(diag(H) > 2*k/n)) -> leverages
```

Obtenemos una gráfica donde se aprecian los residuales vs leverage, donde se aprecia que las observaciones 357, 84 y 374 son algunas de las que presentan leverage alto.

```
library(olsrr)
modelo |> plot(which=5)
```



Se identificaron 7 leverage, y efectivamente las observaciones 374, 357 y 84 presentan leverage alto.

```
leverages
```

```
[1]  72  84 319 327 357 362 374
```

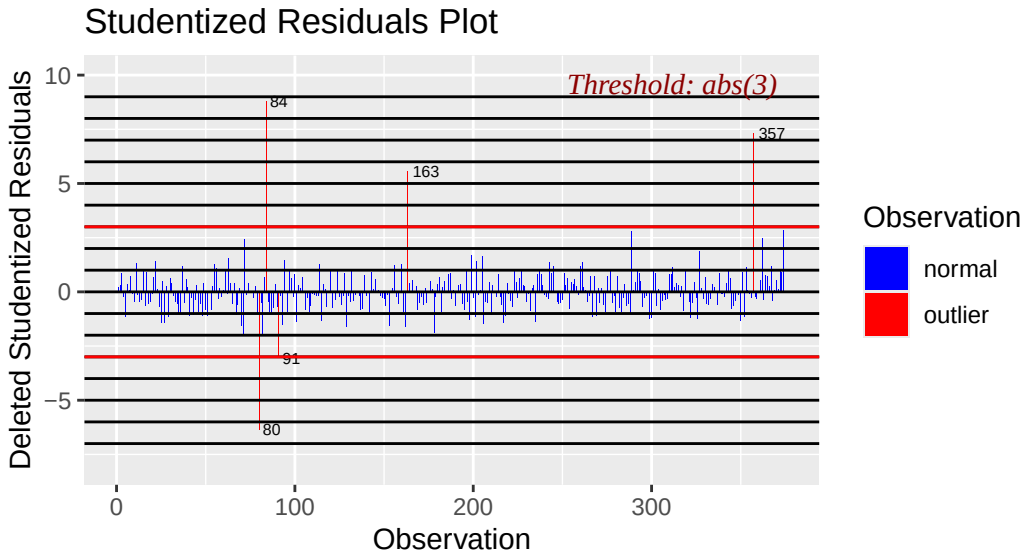
4. Detección de valores atípicos

Calcula los residuales estudentizados y verifica aquellos cuyo valor absoluto es mayor a 2. De ser así, esa observación sería outlier.

```
library(MASS)
as.vector(which((modelo |> studres() |> abs()) > 2)) -> outliers
```

La gráfica nos permite ver que las observaciones 80, 84, 91, 163 y 357 son valores atípicos, porque están fuera de los límites (margen rojo)

```
modelo |> ols_plot_resid_stud()
```

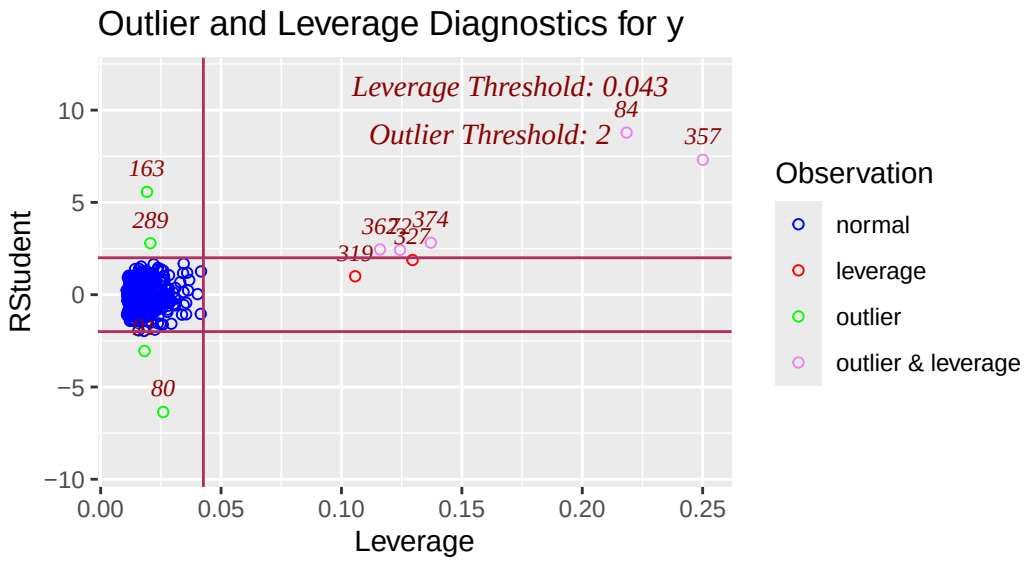


Luego, esta gráfica permite ver las observaciones con alto leverage y también las atípicas.

Por ejemplo:

- 289: es un valor atípico pero no leverage
- 357: es outlier y leverage
- 319: es leverage pero no atípico

```
modelo |> ols_plot_resid_lev()
```



Se identificaron 9 valores atípicos:

```
outliers
```

```
[1]  72  80  84  91 163 289 357 362 374
```

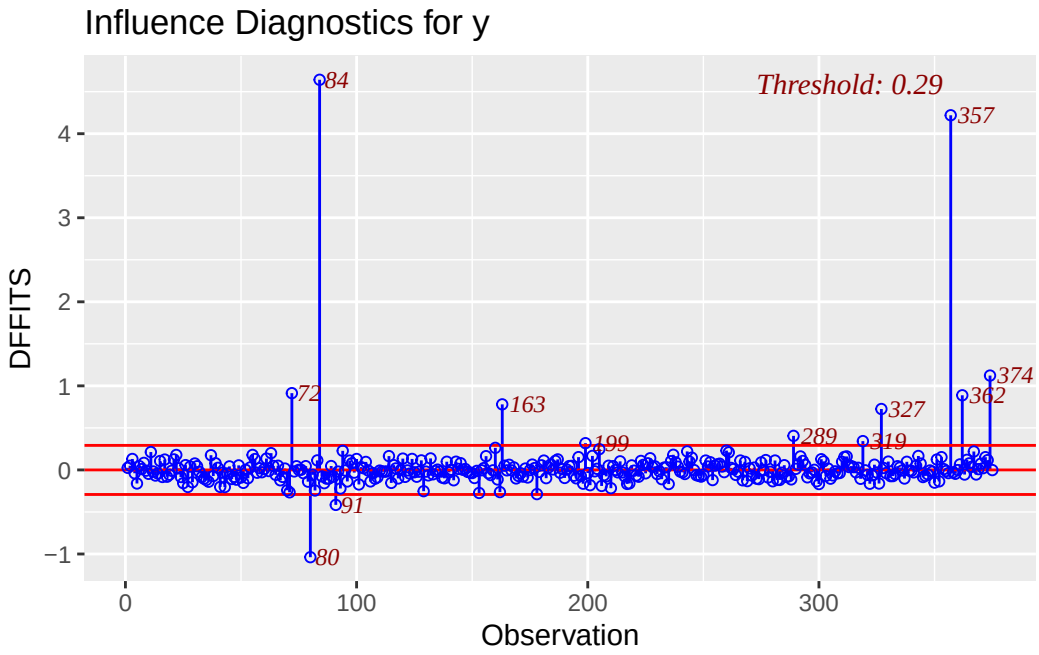
5. Detección de DFFITS

Evaluamos los DFFITS del modelo, es decir aquellas observaciones que al ser retiradas ocasionan un gran cambio en los $\hat{y}$

```
modelo |> coef() |> length() -> k
train |> nrow() -> n
(abs(modelo |> dffits())) -> dffits
as.vector(which(dffits >= 2*sqrt(k/n))) -> inf_dffits
```

Gráficamente, apreciamos las observaciones con mayor DFFITS, siendo la 357 y 84 las que más influyen.

```
modelo |> ols_plot_dffits()
```



12 valores influyentes según DFFITS, es decir aquellos que superaron el umbral de 0.29.

```
inf_dffits
```

```
 [1]  72  80  84  91 163 199 289 319 327 357 362 374
```

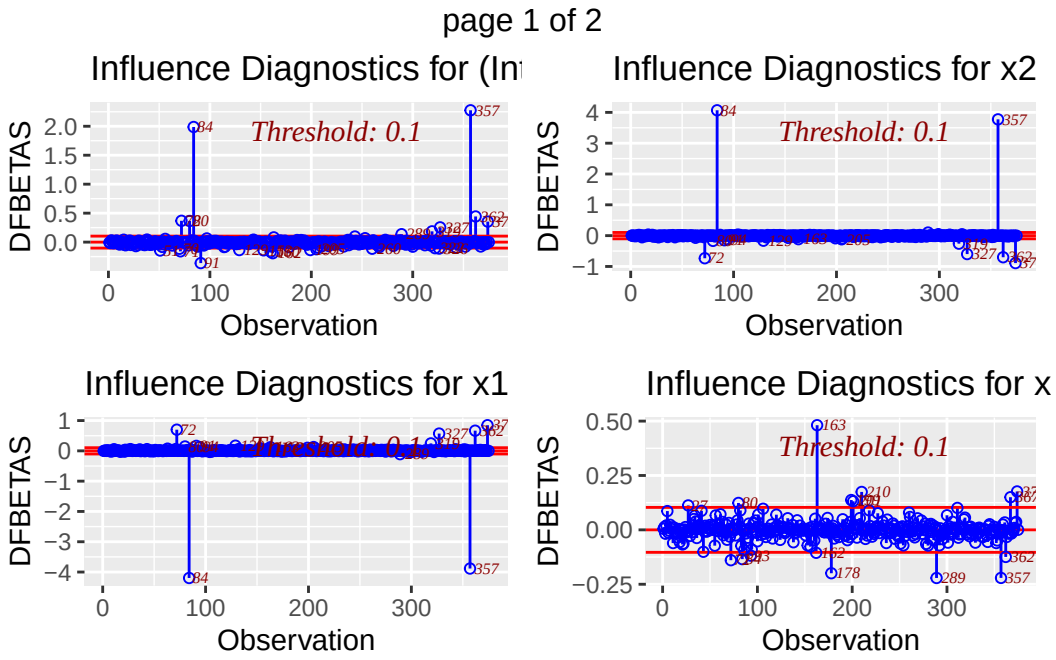
6. Detección de DFBETAS

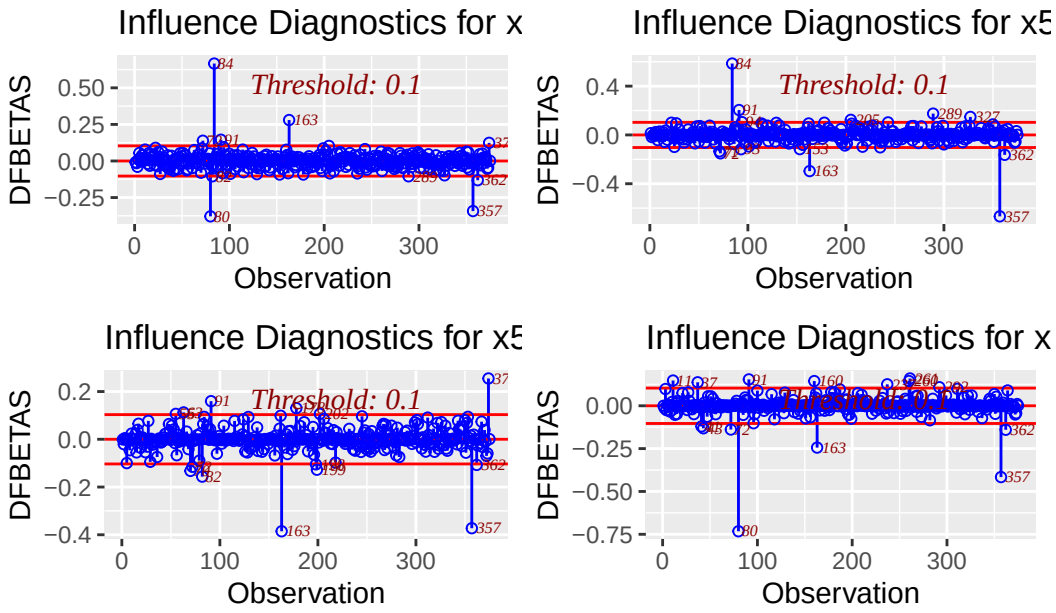
Los DFBETAS permiten evaluar el cambio que tiene cada β al eliminar cada observación.

```
(abs(modelo |> dfbetas())) -> dfbetas
as.vector(which(dfbetas[,1] >= 2/sqrt(n))) -> dfbeta0

as.vector(which(dfbetas[,2] >= 2/sqrt(n))) -> dfbeta1

as.vector(which(dfbetas[,3] >= 2/sqrt(n))) -> dfbeta2

as.vector(which(dfbetas[,4] >= 2/sqrt(n))) -> dfbeta3

as.vector(which(dfbetas[,5] >= 2/sqrt(n))) -> dfbeta4

as.vector(which(dfbetas[,6] >= 2/sqrt(n))) -> dfbeta5

as.vector(which(dfbetas[,7] >= 2/sqrt(n))) -> dfbeta6

as.vector(which(dfbetas[,8] >= 2/sqrt(n))) -> dfbeta7
```

Se aprecia los dfbetas para cada coeficiente / variable.

```
modelo |> ols_plot_dfbetas()
```





Valores influyentes según DFBETAS: hay observaciones que, al ser retiradas, influyen en varios coeficientes, como por ejemplo 72, 84, 91, 374, 80, 163.

```
dfbeta0
```

```
 [1]  51  70  71  72  80  84  91 129 153 160 162 199 205 260 289 319 322 326 327
[20] 357 362 374
```

```
dfbeta1
```

```
 [1]  72  80  84  91  94 129 163 205 289 319 327 357 362 374
```

```
dfbeta2
```

```
 [1]  72  80  84  91  94 129 163 205 319 327 357 362 374
```

```
dfbeta3
```

```
 [1]  27  72  80  84  93 162 163 178 199 201 210 289 357 362 367 374
```

```
dfbeta4
```

```
 [1]  72  80  82  84  91 163 289 357 362 374
```

```
dfbeta5
```

```
 [1]  55  63  70  72  82  91 163 178 198 199 202 357 362 374
```

```
dfbeta6
```

```
 [1]  71  72  84  91  93  94 153 163 205 289 327 357 362
```

```
dfbeta7
```

```
 [1]  11  37  41  43  72  80  91 160 163 237 260 261 292 357 362
```

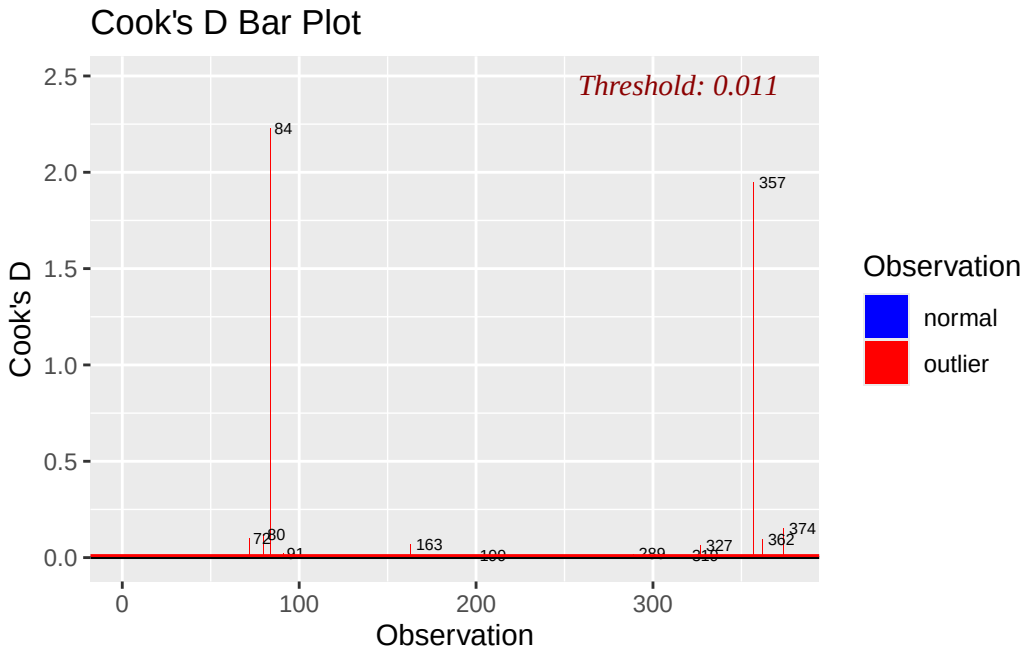
7. Detección de valores influyentes según distancia de Cook

Se halla la distancia de Cook para cada observación del modelo, y luego se verifica cuáles superan el umbral 4/n.

```
modelo |> cooks.distance() -> cookd
as.vector(which(cookd > 4/n)) -> inf_cook
```

Gráficamente, observamos que 84 y 357 (que los identificamos como leverage y outliers), presentan las mayores distancias de Cook, es decir son aquellas que, al ser retiradas, cambia más el vector de coeficientes β. Además hay otras observaciones con distancia de Cook algo altas (72, 80, 163, 362, 374)

```
modelo |> ols_plot_cooksd_bar()
```



12 valores influyentes según Distancia de Cook:

```
inf_cook
```

```
 [1]  72  80  84  91 163 199 289 319 327 357 362 374
```

8. Detección de valores influyentes según COVRATIO

Se halla el COVRATIO de cada observación del modelo y luego se evalúa, convirtiendo en dos vectores aquellos con mayor ($> 1 + 3k/n$) y menor ($< 1-3k/n$) COVRATIO. Finalmente, todo se junta en un solo vector.

```
modelo |> covratio() |> abs() -> covra
as.vector(which(covra > 1+3*k/n)) -> inf_covra1
as.vector(which(covra < 1-3*k/n)) -> inf_covra2
c(inf_covra1, inf_covra2) -> inf_covra
```

9 valores influyentes según COVRATIO:

```
inf_covra
```

```
[1] 185 319 327  80  84  91 163 289 357
```

En resumen, la observación 357 es aquella que es atípica, tiene leverage alto y es influencial en todos los criterios.

```
lista <- list(leverages,
              outliers,
              inf_dffits,
              dfbeta0,dfbeta1,dfbeta2,dfbeta3,dfbeta4,dfbeta5,dfbeta6,dfbeta7,
              inf_cook,
              inf_covra)

inter <- Reduce(intersect, lista)
inter
```

```
[1] 357
```

9. Remoción de los valores leverage, atípicos e influyentes, es decir solo la observación 357.

```
train2 <- train[-c(inter),]
```

10. Creación de un nuevo modelo con el conjunto de datos de entrenamiento "limpio".

```
modelo2 = lm(y ~ x1 + x2 + x3 + x4 + x5, data = train2)
```

11. Comparación de los reportes del modelo original y el nuevo

```
modelo |> summary()
```

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-11.7807  -1.0193  -0.1207   0.8529  13.9607

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.55013    0.42123  13.176  < 2e-16 ***
x1           1.35969    0.04177  32.553  < 2e-16 ***
x2          -0.55100    0.06354  -8.672  < 2e-16 ***
x3           0.02829    0.10227   0.277 0.782246
x41          0.98436    0.20580   4.783 2.51e-06 ***
x5B          0.77653    0.27400   2.834 0.004851 **
x5C         -1.02727    0.26210  -3.919 0.000106 ***
x5D          1.18452    0.32541   3.640 0.000312 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.976 on 367 degrees of freedom
Multiple R-squared:  0.9966,    Adjusted R-squared:  0.9965
F-statistic: 1.532e+04 on 7 and 367 DF,  p-value: < 2.2e-16
```

```
modelo2 |> summary()
```

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5, data = train2)

Residuals:
     Min       1Q   Median       3Q      Max
-11.7913  -0.9513  -0.1185   0.8672  17.3718

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.65340    0.41271  11.275  < 2e-16 ***
x1           1.51145    0.04425  34.158  < 2e-16 ***
x2          -0.77528    0.06690 -11.589  < 2e-16 ***
x3           0.04942    0.09571   0.516 0.605908
x41          1.05033    0.19273   5.450 9.29e-08 ***
x5B          0.87202    0.25665   3.398 0.000754 ***
x5C         -0.86361    0.24620  -3.508 0.000508 ***
x5D          1.31120    0.30490   4.300 2.19e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.848 on 366 degrees of freedom
Multiple R-squared:  0.9969,    Adjusted R-squared:  0.9969
F-statistic: 1.688e+04 on 7 and 366 DF,  p-value: < 2.2e-16
```

Tendió a aumentar la significancia de las variables, al disminuir la mayoría de los p-valores. Además, cambian los valores de los coeficientes estimados.

12. Selección de variables según R^2 predictivo

Se obtuvo un reporte de indicadores para todos los posibles modelos ($2^5 = 32$), de los cuales aquí se analiza el de R^2 predictivo.

Se observan los modelos con el mayor R^2 predictivo, todos los valores son altos, se elige uno simple y con alto R^2 pred.

```
library(olsrr)
library(dplyr)
ols_step_all_possible(modelo2)$result |> data.frame() -> resultados
resultados |>
  select(n,predrsq,predictors) |>
  arrange(-predrsq) |>
  head()
```

```
  n   predrsq     predictors
1 4 0.9960979    x1 x2 x4 x5
2 5 0.9960822 x1 x2 x3 x4 x5
3 3 0.9958170       x1 x2 x5
4 4 0.9958046    x1 x2 x3 x5
5 3 0.9956196       x1 x2 x4
6 4 0.9956054    x1 x2 x3 x4
```

Modelo resultante: Y = f(x1, x2, x5)

13. Selección de variables según AIC

Del reporte de todos los modelos con sus indicadores, se elige los 6 con menor AIC (penaliza el número de variables).

El modelo con x1, x2, x4 y x5 es el que tiene menor AIC, y no hay un modelo más simple que ese con un AIC tan bajo o con una diferencia de solo 2 unidades.

```
resultados |>
  select(n,aic,predictors) |>
  arrange(aic) |>
  head()
```

```
  n      aic     predictors
1 4 1529.078    x1 x2 x4 x5
2 5 1530.806 x1 x2 x3 x4 x5
3 3 1556.697       x1 x2 x5
4 4 1557.986    x1 x2 x3 x5
5 3 1586.814       x1 x2 x4
6 4 1587.844    x1 x2 x3 x4
```

Modelo resultante: Y = f(x1, x2, x4, x5)

14. Selección de variables según SBC

Del reporte de todos los modelos con sus indicadores, se elige los 6 con menor SBC (penaliza el número de variables).

Coloca en primer lugar al mismo modelo que el AIC, y del mismo modo no compite con ningún otro.

```
resultados |>
  select(n,sbc,predictors) |>
  arrange(sbc) |>
  head()
```

```
  n      sbc     predictors
1 4 1560.472    x1 x2 x4 x5
2 5 1566.124 x1 x2 x3 x4 x5
3 3 1584.167       x1 x2 x5
4 4 1589.380    x1 x2 x3 x5
5 3 1606.435       x1 x2 x4
6 4 1611.390    x1 x2 x3 x4
```

Modelo resultante: Y = f(x1, x2, x4, x5)

15. Selección de variables según SBIC

Del reporte de todos los modelos con sus indicadores, se elige los 6 con menor SBIC (penaliza el número de variables).

Coloca en primer lugar al mismo modelo que el AIC, y del mismo modo no compite con ningún otro, ya que el segundo es más complejo y los demás tienen SBIC muy alto.

```
resultados |>
  select(n,sbic,predictors) |>
  arrange(sbic) |>
  head()
```

```
  n     sbic     predictors
1 4 463.9225    x1 x2 x4 x5
2 5 465.7012 x1 x2 x3 x4 x5
3 3 490.8719       x1 x2 x5
4 4 492.0656    x1 x2 x3 x5
5 3 524.2984       x1 x2 x4
6 4 525.0742    x1 x2 x3 x4
```

Modelo resultante: Y = f(x1, x2, x4, x5)

16. Selección de variables según Cp de Mallows

Se reporta los 6 modelos con mejor criterio de Cp de Mallows, es decir aquellos cuyo valor se acerque al número de variables.

El mejor modelo incluye las variables x1, x2, x4, x5 igual que los criterios previos, y la diferencia absoluta cp - variables es solo 2.266637

```
resultados |>
  select(n,cp,predictors) |>
  mutate(dif = n-cp,.after=cp) |>
  arrange(abs(dif)) |>
  head()
```

```
  n        cp        dif     predictors
1 4  6.266637  -2.266637    x1 x2 x4 x5
2 5  8.000000  -3.000000 x1 x2 x3 x4 x5
3 3 34.452883 -31.452883       x1 x2 x5
4 4 35.699554 -31.699554    x1 x2 x3 x5
5 4 69.193442 -65.193442    x1 x2 x3 x4
6 3 68.318519 -65.318519       x1 x2 x4
```

Modelo resultante: Y = f(X1, X2, X4, X5)

17. Selección del modelo según mejores subconjuntos

Se reportan los mejores modelos de 1 variable, de 2 variables, de 3 variables, etc.

Luego, se elige el modelo 4 (Y ~ x1 + x2 + x4 + x5), porque tiene los mejores indicadores:

- R^2 predictivo más alto (aunque empatado con el 5)
- C_p de Mallows más cercano al número de variables (6.2666 - 4 = 2.2666)
- AIC más bajo (1529.0782), aunque compite con el modelo 5 (1530.8059) porque difieren en menos de 2 unidades, pero el modelo 4 es más simple
- SBIC más bajo (463.9225), aunque compite con el modelo 5 (465.7012) porque difieren en menos de 2 unidades, pero el modelo 4 es más simple
- SBC más bajo (1560.4723) que todos los demás.

```
modelo2 |> ols_step_best_subset()
```

```
   Best Subsets Regression
----------------------------
Model Index    Predictors
----------------------------
     1         x1
     2         x1 x2
     3         x1 x2 x5
     4         x1 x2 x4 x5
     5         x1 x2 x3 x4 x5
----------------------------
```

Subsets Regression Summary

Model	R-Square	Adj. R-Square	Pred R-Square	C(p)	AIC	SBIC	SBC	MSEP	FPE	HSP
1	0.9949	0.9948	0.9947	239.6075	1709.6139	646.4032	1721.3867	2094.0560	5.6290	0.0151
2	0.9960	0.9960	0.9953	102.0304	1614.3672	551.6919	1630.0642	1618.9601	4.3635	0.0117
3	0.9967	0.9966	0.9958	34.4529	1556.6974	490.8719	1584.1672	1369.2324	3.7203	0.0100
4	0.9969	0.9969	0.9961	6.2666	1529.0782	463.9225	1560.4723	1268.4153	3.4555	0.0093
5	0.9969	0.9969	0.9961	8.0000	1530.8059	465.7012	1566.1242	1270.9456	3.4715	0.0093

```
AIC: Akaike Information Criteria
 SBIC: Sawa's Bayesian Information Criteria
 SBC: Schwarz Bayesian Criteria
 MSEP: Estimated error of prediction, assuming multivariate normality
 FPE: Final Prediction Error
 HSP: Hocking's Sp
 APC: Amemiya Prediction Criteria
```

Modelo resultante: Y = f(x1, x2, x4, x5)

18. Selección del mejor modelo según forward selection

La selección via forward va introduciendo variables al modelo. La primera a ser introducida fue X_1 por tener el menor p-valor entre todas las variables, lo que le otorgó la mayor significancia.

Luego, estando ya X_1 en el modelo, la segunda variable en entrar fue X_2, luego X_4 y finalmente X_5. Como consecuencia natural y lógica, el AIC fue reduciéndose.

```
modelo2 |> ols_step_forward_p()
```

```
                              Stepwise Summary
-------------------------------------------------------------------------
Step    Variable        AIC         SBC         SBIC        R2       Adj. R2
-------------------------------------------------------------------------
 0      Base Model    3678.527    3686.375    2613.180    0.00000    0.00000
 1      x1            1709.614    1721.387     646.403    0.99486    0.99484
 2      x2            1614.367    1630.064     551.692    0.99603    0.99601
 3      x4            1586.814    1606.435     524.298    0.99633    0.99631
 4      x5            1529.078    1560.472     463.922    0.99691    0.99686
-------------------------------------------------------------------------

Final Model Output
------------------

                         Model Summary
---------------------------------------------------------------
R                       0.998       RMSE                 1.829
R-Squared               0.997       MSE                  3.346
Adj. R-Squared          0.997       Coef. Var            1.456
Pred R-Squared          0.996       AIC                1529.078
MAE                     1.178       SBC                1560.472
---------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
 AIC: Akaike Information Criteria
 SBC: Schwarz Bayesian Criteria

                                ANOVA
--------------------------------------------------------------------------
                Sum of
               Squares         DF    Mean Square        F          Sig.
--------------------------------------------------------------------------
Regression    403616.821        6      67269.470    19727.742    0.0000
Residual        1251.430      367          3.410
Total         404868.252      373
--------------------------------------------------------------------------

                                Parameter Estimates
-----------------------------------------------------------------------------------------
      model      Beta    Std. Error    Std. Beta       t        Sig      lower     upper
-----------------------------------------------------------------------------------------
(Intercept)     4.646         0.412                11.275    0.000     3.835     5.456
         x1     1.511         0.044        1.506     34.192    0.000     1.425     1.598
         x2    -0.775         0.067       -0.511    -11.600    0.000    -0.907    -0.644
        x41     1.057         0.192        0.016      5.500    0.000     0.679     1.435
        x5B     0.883         0.256        0.012      3.454    0.001     0.380     1.385
        x5C    -0.872         0.245       -0.012     -3.554    0.000    -1.355    -0.390
```

```
        x5D     1.300         0.304        0.014       4.279    0.000     0.703     1.898
-------------------------------------------------------------------------------------------
```

La variable X_1 fue la primera en entrar porque redujo en mayor medida el AIC (más que las otras variables), pasando de 3678 a 1709. La segunda en entrar fue X_2, reduciendo el AIC de 1709 a 1614, y así sucesivamente entraron X_4 y X_5.

```
modelo2 |> ols_step_forward_aic()
```

```
                              Stepwise Summary
-------------------------------------------------------------------------
Step    Variable        AIC         SBC         SBIC        R2       Adj. R2
-------------------------------------------------------------------------
 0      Base Model    3678.527    3686.375    2613.180    0.00000    0.00000
 1      x1            1709.614    1721.387     646.403    0.99486    0.99484
 2      x2            1614.367    1630.064     551.692    0.99603    0.99601
 3      x5            1556.697    1584.167     490.872    0.99665    0.99661
 4      x4            1529.078    1560.472     463.922    0.99691    0.99686
-------------------------------------------------------------------------

Final Model Output
------------------

                         Model Summary
---------------------------------------------------------------
R                       0.998       RMSE                 1.829
R-Squared               0.997       MSE                  3.346
Adj. R-Squared          0.997       Coef. Var            1.456
Pred R-Squared          0.996       AIC               1529.078
MAE                     1.178       SBC               1560.472
---------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
 AIC: Akaike Information Criteria
 SBC: Schwarz Bayesian Criteria

                                ANOVA
-----------------------------------------------------------------------
                Sum of
               Squares         DF    Mean Square        F          Sig.
-----------------------------------------------------------------------
Regression    403616.821        6      67269.470    19727.742    0.0000
Residual        1251.430      367          3.410
Total         404868.252      373
-----------------------------------------------------------------------

                                Parameter Estimates
-------------------------------------------------------------------------------------------
      model      Beta    Std. Error    Std. Beta       t        Sig      lower     upper
-------------------------------------------------------------------------------------------
(Intercept)     4.646         0.412                11.275    0.000     3.835     5.456
         x1     1.511         0.044        1.506     34.192    0.000     1.425     1.598
         x2    -0.775         0.067       -0.511    -11.600    0.000    -0.907    -0.644
        x5B     0.883         0.256        0.012      3.454    0.001     0.380     1.385
        x5C    -0.872         0.245       -0.012     -3.554    0.000    -1.355    -0.390
        x5D     1.300         0.304        0.014      4.279    0.000     0.703     1.898
        x41     1.057         0.192        0.016      5.500    0.000     0.679     1.435
-------------------------------------------------------------------------------------------
```

Modelo resultante: Y = f(x1, x2, x4, x5)

19. Selección del mejor modelo según backward selection

Comienza con el modelo completo y retira X_3 porque fue la variable con el mayor pvalor. No retira más variables porque las demás no tienen un pvalor alto (solo x3 tiene pvalor>0.05).

```
modelo2 |> ols_step_backward_p()
```

```
                           Stepwise Summary
-------------------------------------------------------------------------
Step    Variable        AIC         SBC        SBIC        R2       Adj. R2
-------------------------------------------------------------------------
 0      Full Model    1530.806    1566.124    465.701    0.99691    0.99685
 1      x3            1529.078    1560.472    463.922    0.99691    0.99686
-------------------------------------------------------------------------

Final Model Output
------------------

                        Model Summary
---------------------------------------------------------------
R                       0.998       RMSE                 1.829
R-Squared               0.997       MSE                  3.346
Adj. R-Squared          0.997       Coef. Var            1.456
Pred R-Squared          0.996       AIC               1529.078
MAE                     1.178       SBC               1560.472
---------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
 AIC: Akaike Information Criteria
 SBC: Schwarz Bayesian Criteria

                                ANOVA
-------------------------------------------------------------------------
                Sum of
               Squares         DF    Mean Square        F          Sig.
-------------------------------------------------------------------------
Regression    403616.821        6      67269.470    19727.742    0.0000
Residual        1251.430      367          3.410
Total         404868.252      373
-------------------------------------------------------------------------

                                Parameter Estimates
-----------------------------------------------------------------------------------------
      model      Beta    Std. Error    Std. Beta       t        Sig      lower     upper
-----------------------------------------------------------------------------------------
(Intercept)     4.646         0.412                11.275    0.000     3.835     5.456
         x1     1.511         0.044        1.506     34.192    0.000     1.425     1.598
         x2    -0.775         0.067       -0.511    -11.600    0.000    -0.907    -0.644
        x41     1.057         0.192        0.016      5.500    0.000     0.679     1.435
        x5B     0.883         0.256        0.012      3.454    0.001     0.380     1.385
        x5C    -0.872         0.245       -0.012     -3.554    0.000    -1.355    -0.390
        x5D     1.300         0.304        0.014      4.279    0.000     0.703     1.898
-----------------------------------------------------------------------------------------
```

Comienza con el modelo completo y solo retira X_3 porque es la única, que al ser retirada, el AIC disminuye.

```
modelo2 |> ols_step_backward_aic()
```

```
                           Stepwise Summary
--------------------------------------------------------------------------
Step    Variable        AIC         SBC        SBIC        R2       Adj. R2
--------------------------------------------------------------------------
 0      Full Model    1530.806    1566.124    465.701    0.99691    0.99685
 1      x3            1529.078    1560.472    463.902    0.99691    0.99686
--------------------------------------------------------------------------

Final Model Output
------------------

                        Model Summary
---------------------------------------------------------------
R                       0.998       RMSE                 1.829
R-Squared               0.997       MSE                  3.346
Adj. R-Squared          0.997       Coef. Var            1.456
Pred R-Squared          0.996       AIC               1529.078
MAE                     1.178       SBC               1560.472
---------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
 AIC: Akaike Information Criteria
 SBC: Schwarz Bayesian Criteria

                                ANOVA
--------------------------------------------------------------------------
                Sum of
               Squares         DF    Mean Square        F          Sig.
--------------------------------------------------------------------------
Regression    403616.821        6      67269.470    19727.742    0.0000
Residual        1251.430      367          3.410
Total         404868.252      373
--------------------------------------------------------------------------

                                Parameter Estimates
-----------------------------------------------------------------------------------------
      model      Beta    Std. Error    Std. Beta       t        Sig      lower     upper
-----------------------------------------------------------------------------------------
(Intercept)     4.646         0.412                11.275    0.000     3.835     5.456
         x1     1.511         0.044        1.506   34.192    0.000     1.425     1.598
         x2    -0.775         0.067       -0.511  -11.600    0.000    -0.907    -0.644
        x41     1.057         0.192        0.016    5.500    0.000     0.679     1.435
        x5B     0.883         0.256        0.012    3.454    0.001     0.380     1.385
        x5C    -0.872         0.245       -0.012   -3.554    0.000    -1.355    -0.390
        x5D     1.300         0.304        0.014    4.279    0.000     0.703     1.898
-----------------------------------------------------------------------------------------
```

Modelo resultante: Y = f(x1, x2, x4, x5)

20. Selección del mejor modelo según stepwise selection

El método stepwise agrega y retira variables, pero en este caso solo agrega por lo que el resultado es equivalente al de forward selection.

```
modelo2 |> ols_step_both_p()
```

```
                              Stepwise Summary
-------------------------------------------------------------------------------
Step    Variable        AIC         SBC         SBIC        R2       Adj. R2
-------------------------------------------------------------------------------
 0      Base Model    3678.527    3686.375    2613.180    0.00000    0.00000
 1      x1 (+)        1709.614    1721.387     646.403    0.99486    0.99484
 2      x2 (+)        1614.367    1630.064     551.692    0.99603    0.99601
 3      x4 (+)        1586.814    1606.435     524.298    0.99633    0.99631
 4      x5 (+)        1529.078    1560.472     463.922    0.99691    0.99686
-------------------------------------------------------------------------------

Final Model Output
------------------

                         Model Summary
---------------------------------------------------------------
R                       0.998       RMSE                 1.829
R-Squared               0.997       MSE                  3.346
Adj. R-Squared          0.997       Coef. Var            1.456
Pred R-Squared          0.996       AIC                1529.078
MAE                     1.178       SBC                1560.472
---------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
 AIC: Akaike Information Criteria
 SBC: Schwarz Bayesian Criteria

                                 ANOVA
-------------------------------------------------------------------------
                Sum of
               Squares         DF    Mean Square        F          Sig.
-------------------------------------------------------------------------
Regression    403616.821        6      67269.470    19727.742    0.0000
Residual        1251.430      367          3.410
Total         404868.252      373
-------------------------------------------------------------------------

                                Parameter Estimates
-----------------------------------------------------------------------------------------
      model      Beta    Std. Error    Std. Beta       t        Sig      lower     upper
-----------------------------------------------------------------------------------------
(Intercept)     4.646         0.412                11.275    0.000     3.835     5.456
         x1     1.511         0.044        1.506   34.192    0.000     1.425     1.598
         x2    -0.775         0.067       -0.511  -11.600    0.000    -0.907    -0.644
        x41     1.057         0.192        0.016    5.500    0.000     0.679     1.435
        x5B     0.883         0.256        0.012    3.454    0.001     0.380     1.385
        x5C    -0.872         0.245       -0.012   -3.554    0.000    -1.355    -0.390
        x5D     1.300         0.304        0.014    4.279    0.000     0.703     1.898
-----------------------------------------------------------------------------------------
```

El método stepwise agrega y retira variables, pero en este caso solo agrega por lo que el resultado es equivalente al de forward selection.

```
modelo2 |> ols_step_both_aic()
```

```
                              Stepwise Summary
-------------------------------------------------------------------------
Step    Variable        AIC         SBC         SBIC        R2       Adj. R2
-------------------------------------------------------------------------
 0      Base Model    3678.527    3686.375    2613.180    0.00000    0.00000
 1      x1 (+)        1709.614    1721.387     646.403    0.99486    0.99484
 2      x2 (+)        1614.367    1630.064     551.692    0.99603    0.99601
 3      x5 (+)        1556.697    1584.167     490.872    0.99665    0.99661
 4      x4 (+)        1529.078    1560.472     463.922    0.99691    0.99686
-------------------------------------------------------------------------

Final Model Output
------------------

                         Model Summary
---------------------------------------------------------------
R                       0.998       RMSE                 1.829
R-Squared               0.997       MSE                  3.346
Adj. R-Squared          0.997       Coef. Var            1.456
Pred R-Squared          0.996       AIC                1529.078
MAE                     1.178       SBC                1560.472
---------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
 AIC: Akaike Information Criteria
 SBC: Schwarz Bayesian Criteria

                                ANOVA
-------------------------------------------------------------------------
                Sum of
               Squares         DF    Mean Square        F          Sig.
-------------------------------------------------------------------------
Regression    403616.821          6      67269.470    19727.742    0.0000
Residual        1251.430        367          3.410
Total         404868.252        373
-------------------------------------------------------------------------

                               Parameter Estimates
-----------------------------------------------------------------------------------------
      model      Beta    Std. Error    Std. Beta       t        Sig      lower     upper
-----------------------------------------------------------------------------------------
(Intercept)     4.646         0.412                11.275    0.000     3.835     5.456
         x1     1.511         0.044        1.506    34.192    0.000     1.425     1.598
         x2    -0.775         0.067       -0.511   -11.600    0.000    -0.907    -0.644
        x5B     0.883         0.256        0.012     3.454    0.001     0.380     1.385
        x5C    -0.872         0.245       -0.012    -3.554    0.000    -1.355    -0.390
        x5D     1.300         0.304        0.014     4.279    0.000     0.703     1.898
        x41     1.057         0.192        0.016     5.500    0.000     0.679     1.435
-----------------------------------------------------------------------------------------
```

Modelo resultante: Y = f(x1, x2, x4, x5)

21. Uso del testing

Hasta este punto, el modelo candidato con mayor fuerza es:

Y = f(x1, x2, x4, x5)

Lo vamos a poner a prueba.

Construimos el modelo3:

```
modelo3 = lm(y ~ x1 + x2 + x4 + x5, data = train2)
```

Obtenemos las predicciones del modelo3 en el training y el testing.

```
pred_train <- predict(modelo3, newdata = train2)
pred_test  <- predict(modelo3, newdata = test)
```

Defino métricas para evaluar las predicciones:

```
metricas <- function(y, yhat){
  rmse <- sqrt(mean((y - yhat)^2)) # root mean square error
  mae  <- mean(abs(y - yhat)) # mean absolute error
  r2   <- 1 - sum((y - yhat)^2)/sum((y - mean(y))^2) # R^2
  c(RMSE = rmse, MAE = mae, R2 = r2)
}
```

Calculo las métricas en el training y testing, esperando que no haya una discrepancia muy grande. De haber discrepancia, significaría que el modelo funciona bien en el training, pero no en el testing (sobreajuste).

```
m_train <- metricas(train2$y, pred_train)
m_test  <- metricas(test$y,  pred_test)
```

```
m_train
```

```
    RMSE      MAE       R2
1.829227 1.177658 0.996909
```

```
m_test
```

```
     RMSE       MAE        R2
2.6027405 1.3958899 0.9931682
```

Lo usual es que el RMSE y el MAE aumenten en el testing (un aprox más de 25% a 50%), y que el R2 baje.

En este caso, la variación es pequeña, por lo que el modelo estaría funcionando bien en el conjunto testing.

21. Interpretación de coeficientes del mejor modelo

```
modelo3 |> coef()
```

```
(Intercept)          x1          x2         x41         x5B         x5C
  4.6455416   1.5114489  -0.7752010   1.0567180   0.8827581  -0.8721613
        x5D
  1.3004364
```

$$\hat{Y} = 4.646 + 1.511x_1 - 0.7752x_2 + 1.0567x_4 + 0.883x_{5b} - 0.872x_{5c} + 1.3x_{5d}$$

Variable respuesta: Precio de venta (miles de USD)

Variables explicativas:

- x1: Área construida, en metros cuadrados
- x2: Índice de calidad constructiva (z-score), es un puntaje técnico de materiales y acabados.
- x3: Intensidad de publicidad digital (z-score), es una medida de exposición en portales y anuncios.
- x4: Tiene cochera (0: No, 1: Sí)
- x5: Zona de la ciudad (A = Zona periférica, B = Zona intermedia, C = Zona comercial, D = Zona premium)

$\hat{\beta}_0 = 4.646$ no tiene interpretación porque el área no puede ser de 0 m^2

$\hat{\beta}_1 = 1.511$: por cada metro cuadrado adicional de área construida, el precio promedio de venta aumenta en 1.511 miles de dólares, manteniendo constante las demás variables.

$\hat{\beta}_2 = -0.7752$: Un aumento de una unidad en el índice de calidad constructiva se asocia con una disminución promedio de 0.7752 miles de dólares en el precio de venta, manteniendo constantes las demás variables.

$\hat{\beta}_4 = 1.0567$: los inmuebles que tienen cochera presentan en promedio un precio de venta 1.0567 miles de dólares mayor que aquellos que no tienen cochera, manteniendo constantes las demás variables.

$\hat{\beta}_{5b} = 0.883$: los inmuebles ubicados en la zona intermedia tienen en promedio un precio de venta 0.883 miles de dólares mayor que los ubicados en la zona periférica, manteniendo constantes las demás variables.

$\hat{\beta}_{5c} = -0.872$: los inmuebles ubicados en la zona comercial tienen en promedio un precio de venta 0.872 miles de dólares menor que los ubicados en la zona periférica, manteniendo constantes las demás variables.

$\hat{\beta}_{5d} = 1.3$: los inmuebles ubicados en la zona premium tienen en promedio un precio de venta 1.3 miles de dólares mayor que los ubicados en la zona periférica, manteniendo constantes las demás variables.

¿Cuál es la ecuación estimada de regresión para un inmueble con cochera en la zona periférica?

$$\hat{Y} = 4.646 + 1.511x_1 - 0.7752x_2 + 1.0567$$
$$\hat{Y} = 5.7207 + 1.511x_1 - 0.7752x_2$$