Trabajo Final
Modelos Lineales

Valentina Soldini, Lucca Frachelle, Cecilia Waksman

Datos

Los datos fueron obtenidos a partir de un estudio llevado a cabo en una policlínica universitaria.

Se cuenta con 73 observaciones, las cuales representan a personas que padecen hipertensión, y 8 variables, las mismas son: presión arterial, edad, peso, superficie corporal, duración desde que a la persona le diagnosticaron hipertensión, el pulso en estado basal y un índice de estrés.

Objetivo

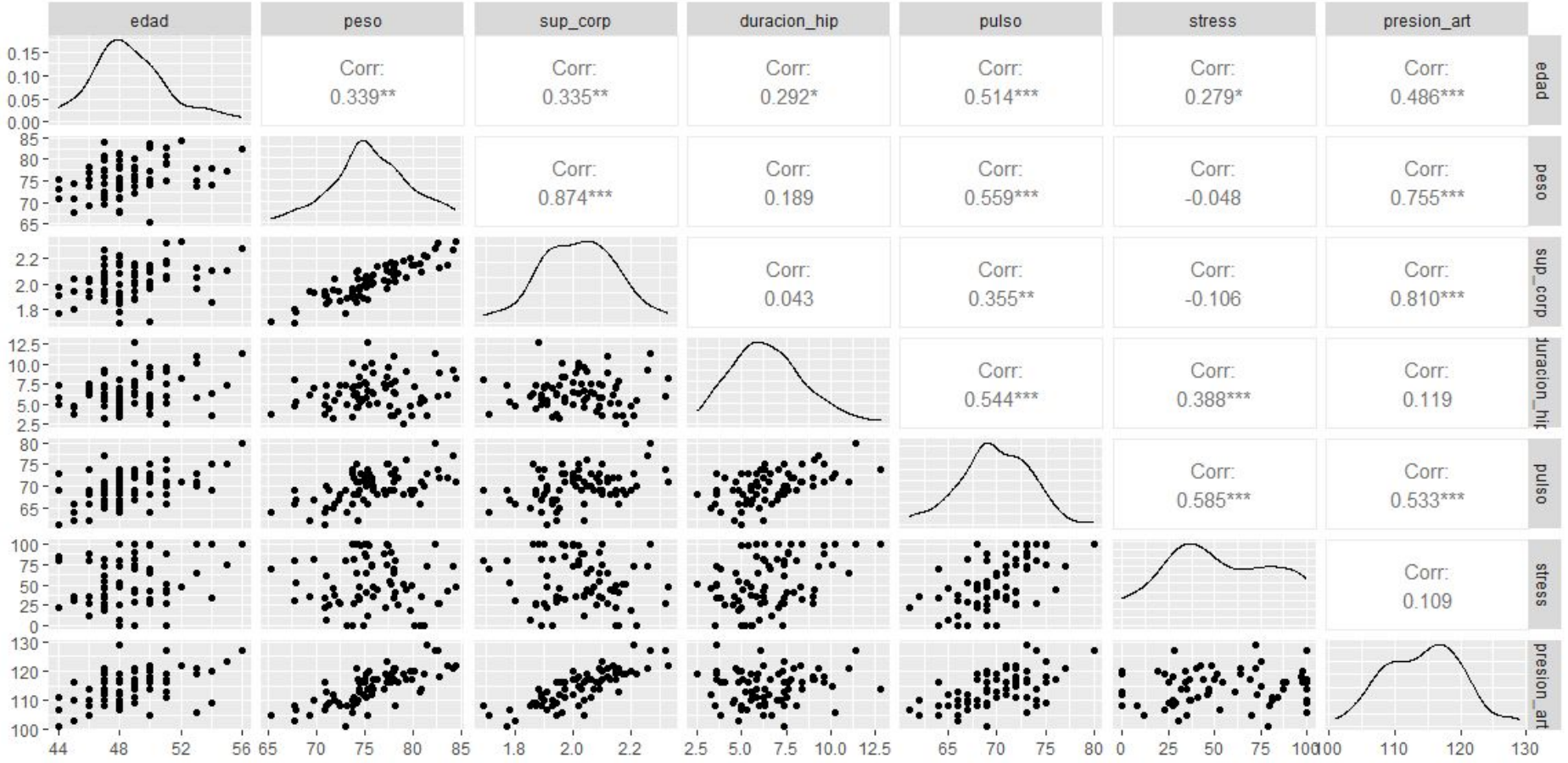
Se busca determinar qué factores inciden en la presión sanguínea en personas hipertensas. Para ello se llevará a cabo un análisis exploratorio que permita explicar y/o predecir el valor de la variable ***presión arterial*** a partir de un modelo lineal que esté basado en las demás variables mencionadas.

Análisis Exploratorio

Table 1: Resúmen estadístico

edad	peso	sup_corp	duracion_hippulso		stress	presion_art
Min. :44.00	Min. :65.30	Min. :1.690	Min. : 2.500	Min. :61.00	Min. : 0.00	Min. :101.0
1st Qu.:47.00	1st Qu.:73.70	1st Qu.:1.930	1st Qu.: 5.000	1st Qu.:68.00	1st Qu.: 30.00	1st Qu.:110.0
Median :48.00	Median :75.30	Median :2.020	Median : 6.300	Median :70.00	Median : 48.00	Median :115.0
Mean :48.68	Mean :75.72	Mean :2.019	Mean : 6.522	Mean :69.86	Mean : 53.05	Mean :114.4
3rd Qu.:50.00	3rd Qu.:78.00	3rd Qu.:2.100	3rd Qu.: 7.600	3rd Qu.:72.00	3rd Qu.: 80.00	3rd Qu.:119.0
Max. :56.00	Max. :84.40	Max. :2.330	Max. :12.800	Max. :80.00	Max. :100.00	Max. :129.0

Relación entre Variables



Supuestos a Cumplir

- **No multicolinealidad**: exacta o aproximada, para asegurar que la matriz X sea de rango completo (conformable),
- **Linealidad**: la relación entre variables explicativas y la respuesta debe ser aproximadamente lineal,
- **Homocedasticidad**: la varianza de los errores no depende de ninguna de las variables explicativas,
- **Normalidad**: los errores del modelo deben presentar una distribución normal,
- **Atípicos / Influyentes**: si bien no es un supuesto en sí mismo, es recomendable identificar observaciones atípicas e influyentes al modelo

Multicolinealidad

Primer modelo:

En primera instancia, decidimos crear un modelo que comprendiera todas las variables del dataset.
Testeamos multicolinealidad calculando el VIF (Variance Inflation Rate) de cada variable y obtuvimos los siguientes resultados:

edad	peso	sup_corp	duracion_hip	pulso	stress
1.472313	7.751650	5.319215	1.481739	4.600766	2.280361

Existen dos variables cuyo VIF es mayor que 5, *peso* y *superficie corporal*. Eliminamos la que cuenta con el mayor valor y repetimos el procedimiento

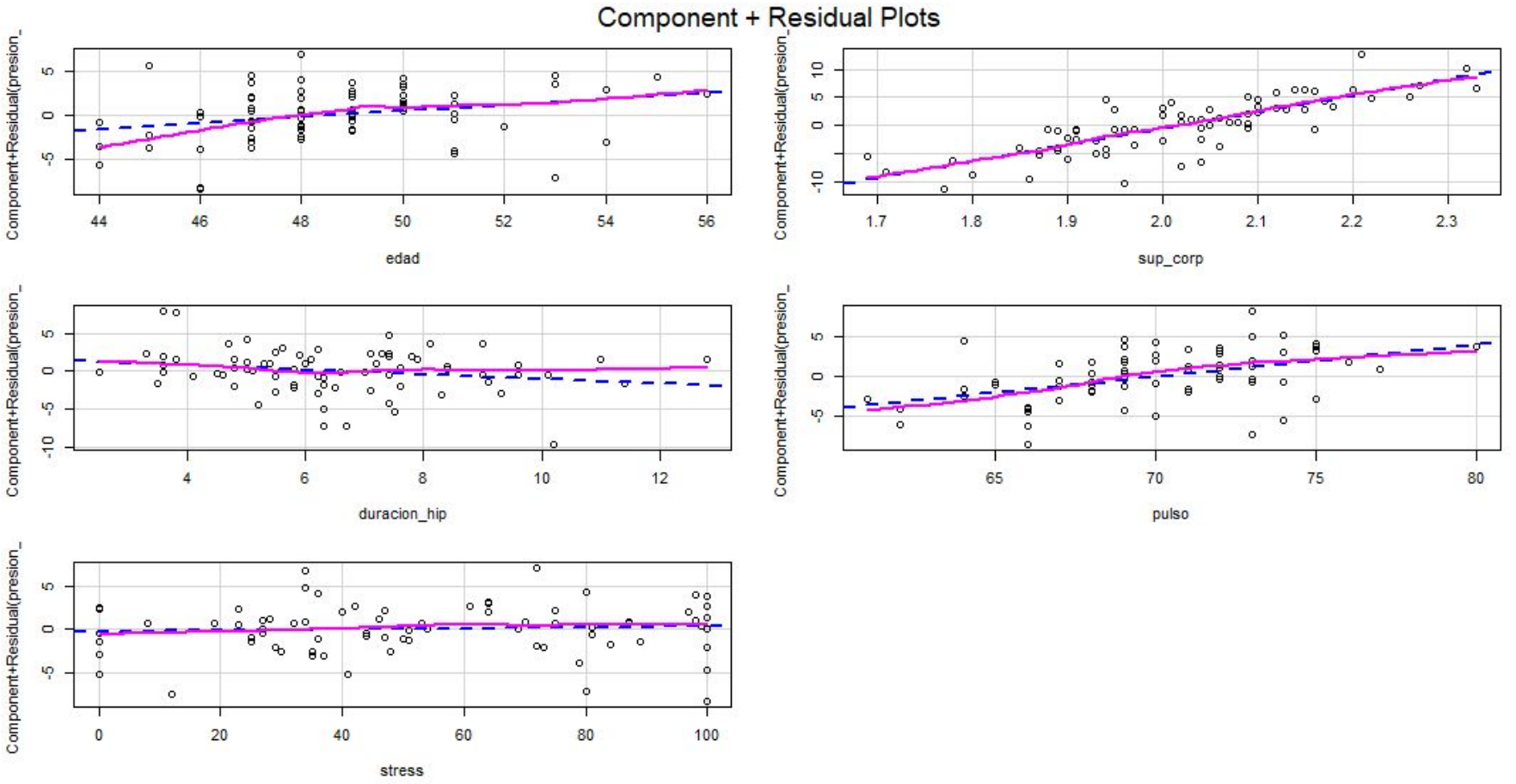
Modelo sin variable *peso*:

edad	sup_corp	duracion_hip	pulso	stress
1.418656	1.480007	1.480623	2.782521	1.841093

Summary: Pruebas de Significación Global e Individual

```
Call:
lm(formula = presion_art ~ edad + sup_corp + duracion_hip + pulso +
    stress, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-8.6105 -1.7111  0.0763  1.9243  7.0588

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.655551   9.331521   1.249   0.2160
edad          0.349633   0.170307   2.053   0.0440 *
sup_corp     29.268817   3.224683   9.076 2.79e-13 ***
duracion_hip -0.292461   0.210008  -1.393   0.1683
pulso         0.403266   0.163247   2.470   0.0161 *
stress        0.006022   0.015906   0.379   0.7062
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.046 on 67 degrees of freedom
Multiple R-squared:  0.7476,    Adjusted R-squared:  0.7288
F-statistic:  39.7 on 5 and 67 DF,  p-value: < 2.2e-16
```

Linealidad


Component + Residual Plots
Component+Residual(presion_
edad
sup_corp
duracion_hip
pulso
stress

Homocedasticidad

Non-Constant Error Variance Test:

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.1392125, Df = 1, p = 0.70907
```

Breusch-Pagan Test:

statistic	p.value	parameter	method	alternative
1.017454	0.9611458	5	Koenker (studentised)	greater

Normalidad

Gráfico de densidad:

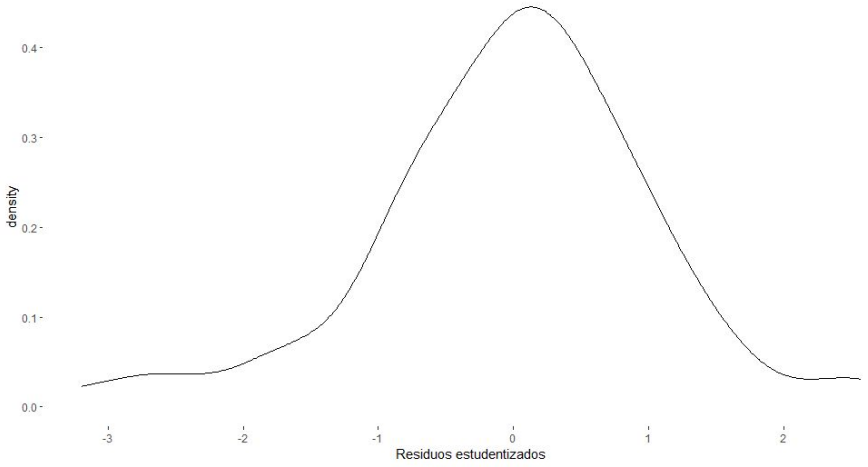
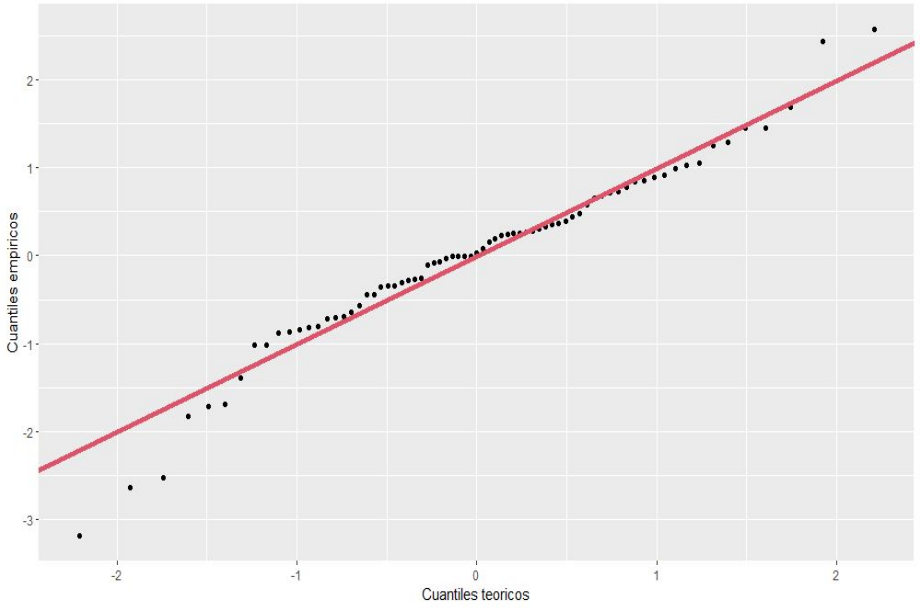


Gráfico de dispersión:



3 Tests de Normalidad

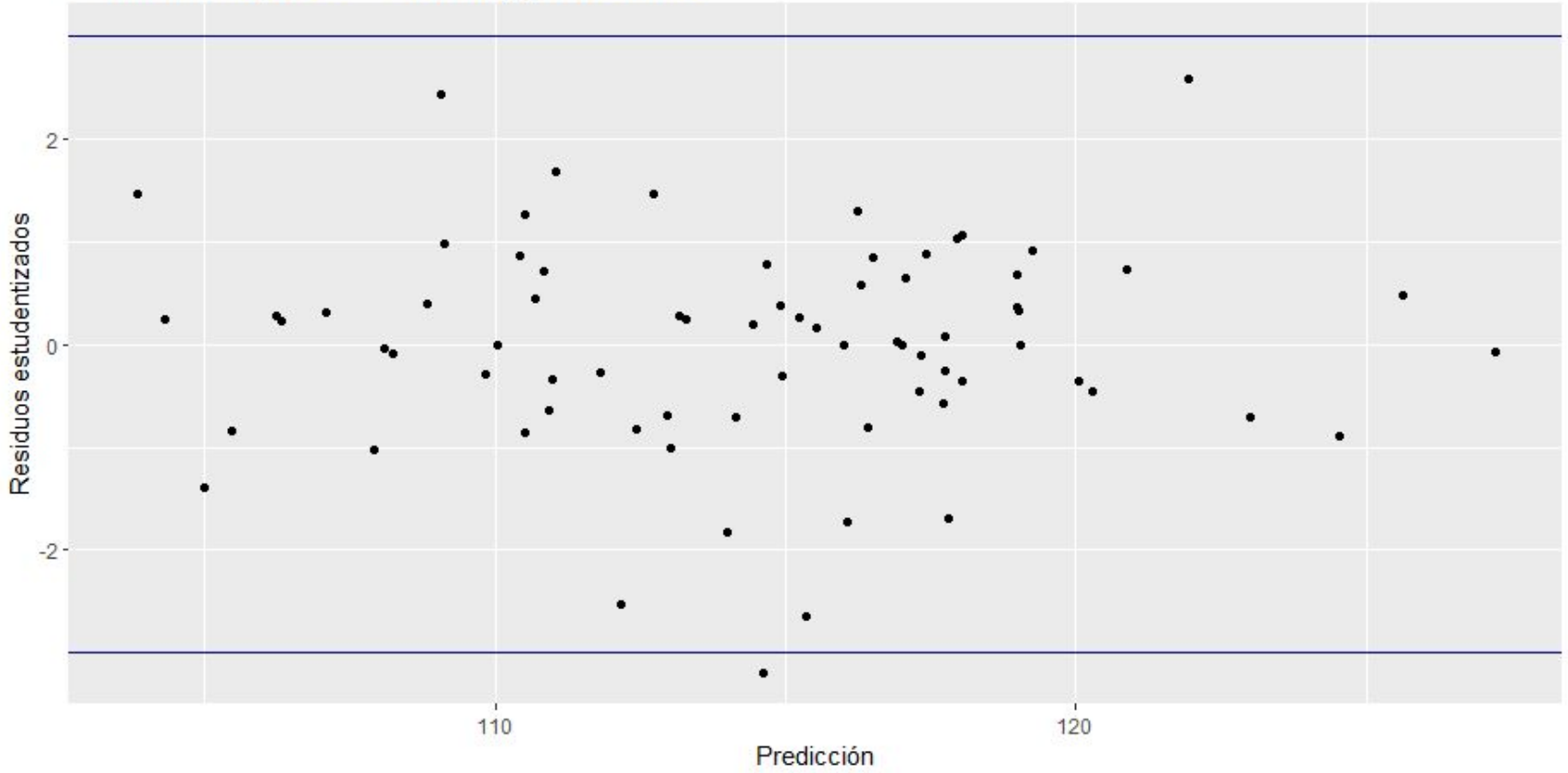
```
        Shapiro-Wilk normality test

data:  rstudent(mod)
W = 0.96959, p-value = 0.07481
```

```
        Jarque Bera Test

data:  rstudent(mod)
X-squared = 6.6665, df = 2, p-value = 0.03568
```

```
        Exact one-sample Kolmogorov-Smirnov test

data:  rstudent(mod)
D = 0.073951, p-value = 0.7919
alternative hypothesis: two-sided
```

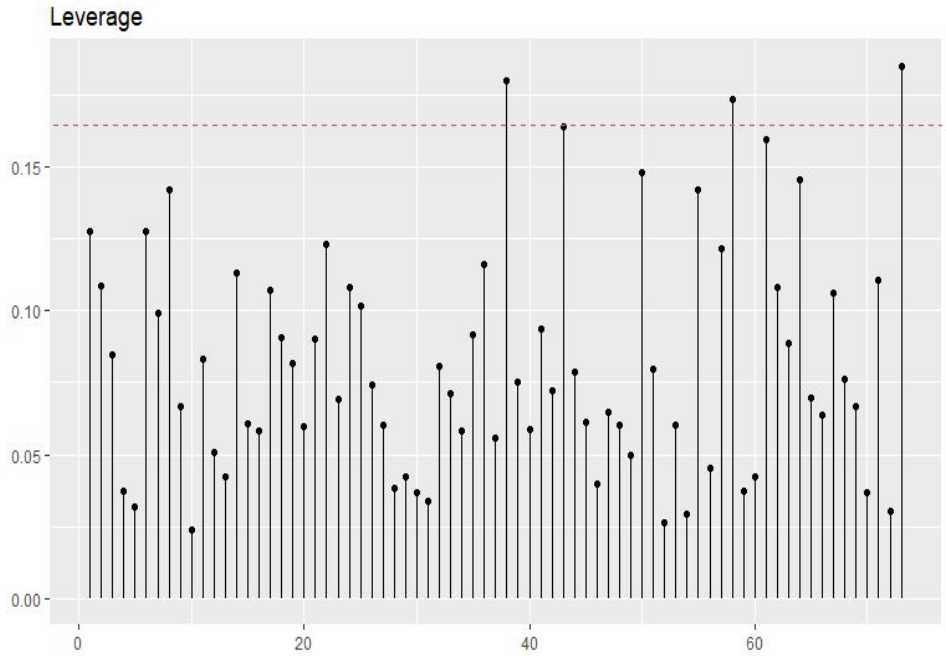
El test de Jarque Bera rechaza el supuesto de normalidad.

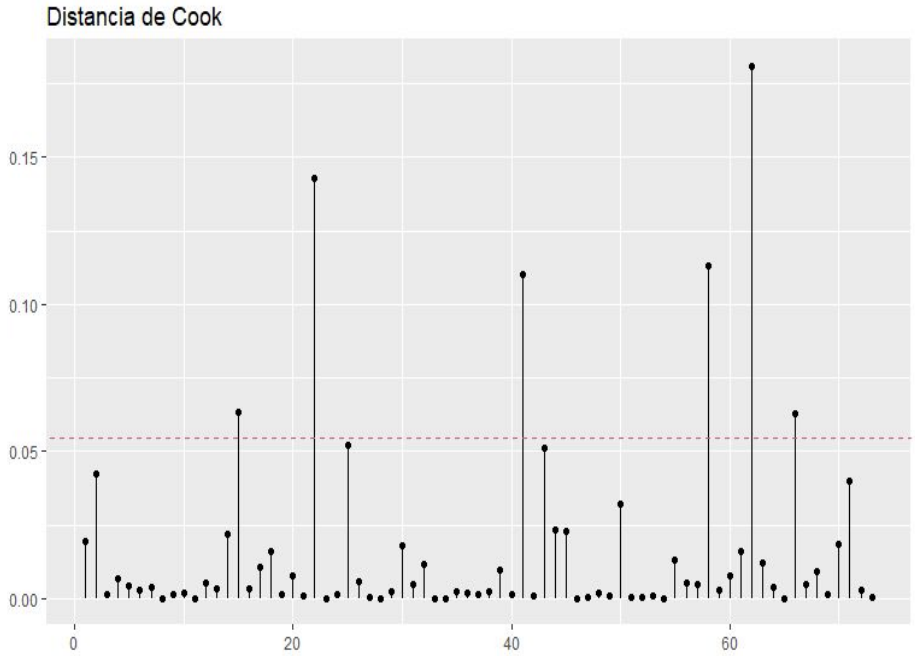
Atípicos



Existe una única observación que podría tomarse como dato atípico, esta es la observación 62.

Influyentes





Según el índice de Leverage, las observaciones más influyentes son: 73, 38, 58, en ese orden.

Según el índice de Distancia de Cook, las observaciones más influyentes son: 62, 22, 58, 41, 15, 66, respectivamente.

Leverage

```
        Shapiro-Wilk normality test

data:  rstudent(mod_I)
W = 0.97135, p-value = 0.1039
```

```
        Jarque Bera Test

data:  na.omit(rstudent(mod_I))
X-squared = 5.7869, df = 2, p-value = 0.05538
```

```
        Exact one-sample Kolmogorov-Smirnov test

data:  rstudent(mod_I)
D = 0.072055, p-value = 0.8288
alternative hypothesis: two-sided
```

Distancia de Cook

```
        Shapiro-Wilk normality test

data:  rstudent(mod_II)
W = 0.9711, p-value = 0.09601
```

```
        Jarque Bera Test

data:  na.omit(rstudent(mod_II))
X-squared = 4.1096, df = 2, p-value = 0.1281
```

```
        Exact one-sample Kolmogorov-Smirnov test

data:  rstudent(mod_II)
D = 0.078031, p-value = 0.7433
alternative hypothesis: two-sided
```

- Al intervenir los influyentes según Leverage, 73 y 38, se acepta normalidad.
- Al intervenir el dato atípico 62 e influyente según distancia de Cook, se acepta normalidad.

Problemas de significación individual

```
Call:
lm(formula = presion_art ~ . - peso - I38 - I73, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-7.1030 -1.2581  0.0243  1.4774  7.4521

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.84715    8.75439   1.239  0.21971
edad          0.44801    0.16265   2.754  0.00759 **
sup_corp     28.82542    3.02717   9.522 5.19e-14 ***
duracion_hip -0.18933    0.19957  -0.949  0.34624
pulso         0.34722    0.15409   2.253  0.02757 *
stress        0.01149    0.01501   0.765  0.44700
I62          -9.65361    3.02431  -3.192  0.00216 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.856 on 66 degrees of freedom
Multiple R-squared:  0.7814,    Adjusted R-squared:  0.7615
F-statistic: 39.32 on 6 and 66 DF,  p-value: < 2.2e-16
```

Las variables *duracion_hip* y *stress* continúan siendo no son significativas en el modelo. En consecuencia, se procederá a repetir todo el análisis excluyendo ambas.

Modelo Final

```
Call:
lm(formula = presion_art ~ . - peso - stress - duracion_hip,
    data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-9.1494 -1.3864  0.2515  1.8003  8.1535

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.2043     8.2988   1.712  0.09146 .
edad          0.3399     0.1699   2.001  0.04938 *
sup_corp     29.5856     2.8897  10.238 1.75e-15 ***
pulso         0.3417     0.1172   2.916  0.00478 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.047 on 69 degrees of freedom
Multiple R-squared:  0.7399,    Adjusted R-squared:  0.7286
F-statistic: 65.42 on 3 and 69 DF,  p-value: < 2.2e-16
```

El modelo sin las variables que no son significativas no acepta normalidad, aún interviniendo observaciones influyentes y/o eliminando observaciones atípicas.

Debemos entonces testear su validez mediante otro método.

Randomization Test

El mismo consiste en extraer P muestras aleatorias con reposición de los datos. Como los mismos ya son una muestra, a estas “sub-muestras” las llamamos réplicas y son del mismo tamaño que los datos. Posteriormente se particiona el problema en modelos RLS, tantos como variables explicativas hayan, de tal forma que se obtenga una nueva versión de la variable explicativa de interés en cada modelo libre del efecto de las demás.

Summary:

	Estimate	Std. Error	t value	Pr(>|t|)	Pr(>|t|)_rand
(Intercept)	14.204258	8.298761	1.711612	0.091460	0.0914
edad	0.339870	0.169887	2.000561	0.049376	0.0544
sup_corp	29.585604	2.889726	10.238203	0.000000	0.0000
pulso	0.341703	0.117189	2.915830	0.004781	0.0044

Luego de generar 5000 réplicas, al comparar los p-valores de la prueba de significación individual y de la prueba de permutaciones, se obtiene que los resultados son relativamente similares, por lo que el modelo es válido aunque sus errores no se distribuyan de forma normal.

Desempeño Predictivo: LOOCV

Para evaluar la actuación del modelo, se debe exponer el mismo a un nuevo conjunto de datos. En este caso, al no contar con otros datos, más que los que son utilizados para crear el modelo, se acude a métodos de validación cruzada, en este caso se usa el método denominado Leave one out (LOOCV).

Table 3: R cuadrado

R2_conv	R2_LOO
0.739884	0.7042639

EL $R2$ se encuentra por encima del 70%, lo cual permite afirmar que él mismo explica suficientemente bien los datos de la muestra. Se observa que el $R2$ por el método LOOCV es un 3,5% más bajo que el $R2$ de la prueba de significación global.

Referencias

- Notas de la unidad curricular Modelos Lineales
- Farrar, Thomas J., and University of the Western Cape. 2024. Skedastic: Handling Heteroskedasticity in the Linear Regression Model. https://github.com/tjfarrar/skedastic.
- Fox, John, and Sanford Weisberg. 2019. An R Companion to Applied Regression. Third. Thousand Oaks CA: Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.
- Maechler, Martin, Peter Rousseeuw, Christophe Croux, Valentin Todorov, Andreas Ruckstuhl, Matias Salibian-Barrera, Tobias Verbeke, Manuel Koller, Eduardo L. T. Conceicao, and Maria Anna di Palma. 2024. Robustbase: Basic Robust Statistics. http://robustbase.r-forge.r-project.org/.
- Schloerke, Barret, Di Cook, Joseph Larmarange, Francois Briatte, Moritz Marbach, Edwin Thoen, Amos Elberg, and Jason Crowley. 2024. GGally: Extension to 'Ggplot2'. https://CRAN.R-project.org/package= GGally.
- Todorov, Valentin, and Peter Filzmoser. 2009. "An Object-Oriented Framework for Robust Multivariate Analysis." Journal of Statistical Software 32 (3): 1–47. https://www.jstatsoft.org/article/view/v032i03/.
- Trapletti, Adrian, and Kurt Hornik. 2024. Tseries: Time Series Analysis and Computational Finance. https://CRAN.R-project.org/package=tseries.
- Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." Journal of Open Source Software 4 (43): 1686. https://doi.org/10.21105/joss.01686.
- Wickham, Hadley, and Jennifer Bryan. 2023. Readxl: Read Excel Files. https://CRAN.R-project.org/ package=readxl.
- Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In Implementing Reproducible Computational Research, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC.
- ———. 2015. Dynamic Documents with R and Knitr. 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC. https://yihui.org/knitr/.
- ———. 2023. Knitr: A General-Purpose Package for Dynamic Report Generation in r. https://yihui.org/ knitr/.