Trabajo Final - Modelos Lineales

Lucca Frachelle , Valentina Solidni , Cecilia Waksman

2024-06-28

Introducción

Este trabajo consiste en un análisis de la incidencia de ciertos factores sobre la presión sanguínea en personas hipertensas. Los factores que se tomarán en cuenta para ello son: edad, peso, superficie corporal, duración desde que a la persona le diagnosticaron hipertensión, el pulso en estado basal y un índice de estrés.

Para ello se utilizarán df de un estudio realizado en una policlínica universitaria y se realizará un modelo de regresión lineal múltiple, dónde la variable de respuesta, Y, es la presión arterial (**presion_art**) y las variables explicativas serán seleccionadas de las nombradas anteriormente (las cuales formarán la matriz X).

El modelo se podrá escribir como $Y = X\beta + \epsilon$, donde β es un vector de parámetros a estimar.

Supuestos a cumplir

Para poder obtener conclusiones confiables, el modelo debe cumplir con determinados supuestos

- **No multicolinealidad**: exacta ni aproximada, para asegurar que la matriz X sea de rango completo (conformable),
- **Linealidad**: la relación entre variables expicativas y la respuesta debe ser aproximadamente lineal,
- **Homoscedasticidad**: 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 si mismo, es recomendable identificar observaciones atípicas e influyentes al modelo.

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

Los datos consisten en mediciones de los siguientes parámetros:

- **Edad**: Varía entre 44 y 56 años, con una mediana de 48 años.
- **Peso**: Oscila entre 65.3 kg y 84.4 kg, con una media de 75.72 kg.
- **Superficie Corporal**: Se encuentra entre 1.69 m^2 y 2.33 m^2, con una media de 2.019 m^2.
- **Duración de la Hipertensión**: Registrada entre 2.5 y 12.8 años, con una mediana de 6.3 años.
- **Pulso** : Varía de 61 a 80 pulsaciones por minuto, con una media de 69.86 pulsaciones por minuto.
- **Índice de Estrés**: Mide entre 0 y 100, con una media de 53.05.
- **Presión Arterial** : Oscila entre 101 mmHg y 129 mmHg, con una media de 114.4 mmHg.

Análisis de Distribución y Correlación

1. **Distribución de Variables**
    - **Edad**: La distribución de la edad muestra una concentración alrededor de los 48-50 años.

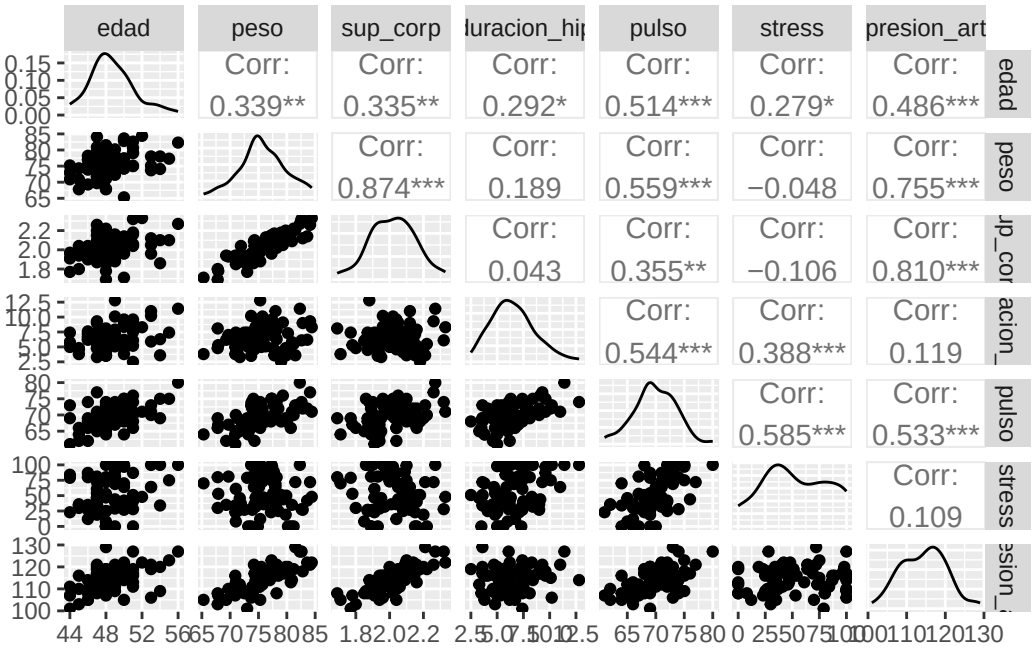

Figure 1: Distribución y Correlación

- **Peso**: La distribución del peso es relativamente normal, con una media alrededor de 75 kg.
- **Superficie Corporal**: La superficie corporal también sigue una distribución normal, centrada cerca de 2.0 m^2.
- **Duración de la Hipertensión**: La duración de la hipertensión muestra una mayor dispersión, pero con una tendencia hacia valores más bajos.
- **Pulso**: El pulso muestra una distribución concentrada entre 65 y 75 pulsaciones por minuto.
- **Estrés**: El índice de estrés varía considerablemente, pero con una mayor concentración en valores bajos.
- **Presión Arterial**: La presión arterial tiene una distribución aproximadamente normal, centrada alrededor de 115 mmHg.

2. **Correlaciones Significativas**

- **Edad y Peso**: Existe una correlación positiva significativa entre la edad y el peso (corr = 0.339).

- **Peso y Superficie Corporal**: La correlación entre el peso y la superficie corporal es muy alta (corr = 0.874), lo cual es esperable ya que ambos parámetros están relacionados físicamente.
- **Superficie Corporal y Peso**: Esta correlación refuerza la relación entre estas dos variables (corr = 0.874).
- **Duración de la Hipertensión y Edad**: Hay una correlación positiva moderada entre la duración de la hipertensión y la edad (corr = 0.544).
- **Pulso y Peso**: La correlación entre el pulso y el peso es positiva y significativa (corr = 0.533).
- **Estrés y Peso**: Existe una correlación moderada entre el estrés y el peso (corr = 0.533).
- **Presión Arterial y Edad**: La presión arterial está positivamente correlacionada con la edad (corr = 0.486).
- **Presión Arterial y Peso**: Hay una correlación significativa entre la presión arterial y el peso (corr = 0.755).
- **Presión Arterial y Superficie Corporal**: Existe una correlación positiva entre la presión arterial y la superficie corporal (corr = 0.810).
- **Presión Arterial y Pulso**: La correlación entre la presión arterial y el pulso es significativa (corr = 0.533).

Diagnóstico del modelo

En primera instancia, se evalúa un modelo en el que participan todas las variables del dataset.

Multicolinealidad

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

Se observa que al calcular el VIF (Variance Inflation Factor) del las variables del modelo, la variable *peso* tiene un valor mayor a 5, indicando una multicolinealidad aproximada alta, por esta razón es eliminada del modelo. Como *peso* está fuertementemente correlacionada con la variable *superfice corporal*, por lo que tiene sentido que estas dos cuenten con un VIF alto, ya que de cierta forma compiten por explicar la misma variablidad.

```
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
```

Observando el *summary*, es visible que el modelo explica un porcentaje relativamente alto de la variabilidad de los datos, contando con un $R^2 = 74,76$. Existen dos variables, las cuales no son significativas al 5%, estas son: *duracion_hip* y *stress*.

```
     edad     sup_corp duracion_hip        pulso       stress
 1.418656     1.480007     1.480623     2.782521     1.841093
```

Calculando nuevamente el VIF se obtiene que todos los valores cuentan con diferencias menores que 5, por lo que no hay multicolinealidad. Se puede continuar con el análisis.

Lineanidad

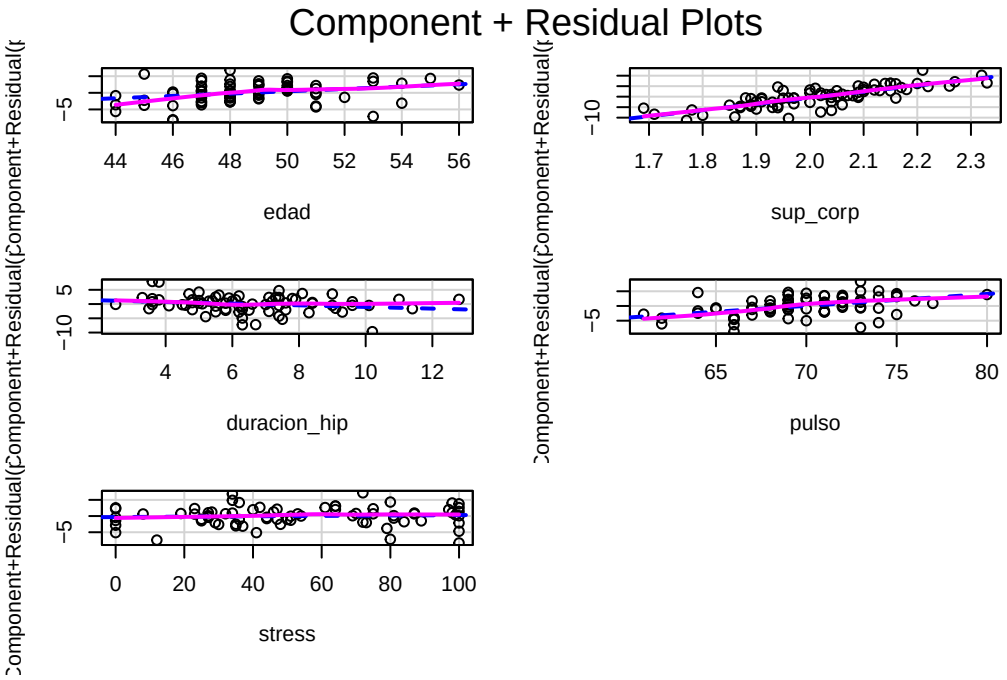

Figure 2: Linealidad

Parece no haber problemas de linealidad en el modelo. Ya que los residuos no presentan un patrón claro , como dispersión o curvatura a lo largo de los valores de x. Por lo que se puede seguir con el análisis.

Homoscedasticidad

ncvTest

- **Hipótesis**:
  - **Hipótesis Nula (H0)**: La varianza de los residuos es constante (homoscedasticidad).
  - **Hipótesis Alternativa (H1)**: La varianza de los residuos no es constante (heterocedasticidad).
- **Procedimiento**:
  - El *ncvTest* examina la relación entre los valores ajustados (predicciones del modelo) y la varianza de los residuos.

  - Se ajusta un modelo de regresión para predecir los residuos en función de los valores ajustados.
  - Se calcula un estadístico de prueba basado en esta relación.
  - El estadístico de prueba sigue una distribución chi-cuadrado.

  Un p-valor alto sugiere que no hay evidencia suficiente para rechazar la hipótesis nula, indicando homoscedasticidad.

  - Un p-valor bajo indica heteroscedasticidad.

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

Breusch-Pagan Test

- **Hipótesis**:
  - **Hipótesis Nula (H0)**: La varianza de los residuos es constante y no depende de las variables independientes (homoscedasticidad).
  - **Hipótesis Alternativa (H1)**: La varianza de los residuos depende linealmente de las variables independientes (heteroscedasticidad).
- **Procedimiento**:
  - El test de *Breusch-Pagan* examina si la varianza de los residuos depende linealmente de las variables independientes del modelo original.
  - Se realiza una regresión auxiliar de los residuos al cuadrado contra las variables independientes originales.
  - Se calcula un estadístico de prueba basado en la regresión auxiliar.
  - El estadístico de prueba sigue una distribución chi-cuadrado.

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

Ambos p-valores son suficientemente altos, por lo que no hay problemas de heteroestaticidad.

Normalidad

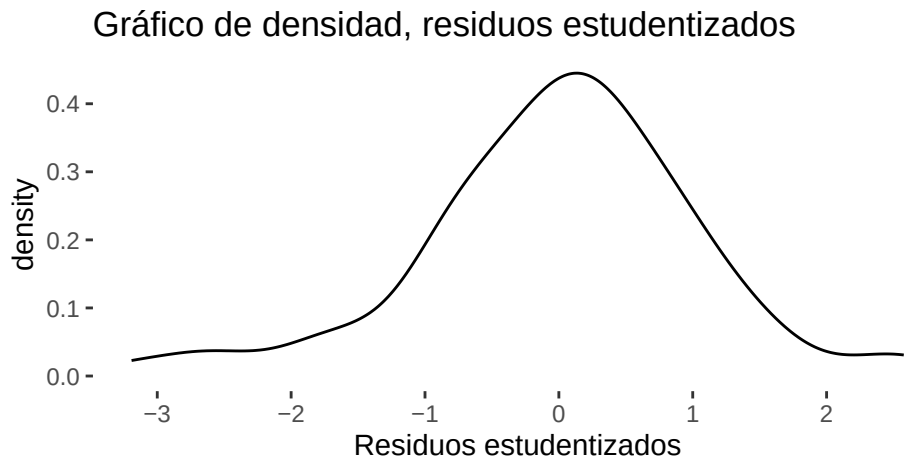

Figure 3: Normalidad de los residuos

A priori, la densidad de los residuos estudentizados parece comportarse normal. Sin embargo, las colas no son totalmente simétricas. Esto puede dar indicios de que los resiudos no se distriubuyen normal.

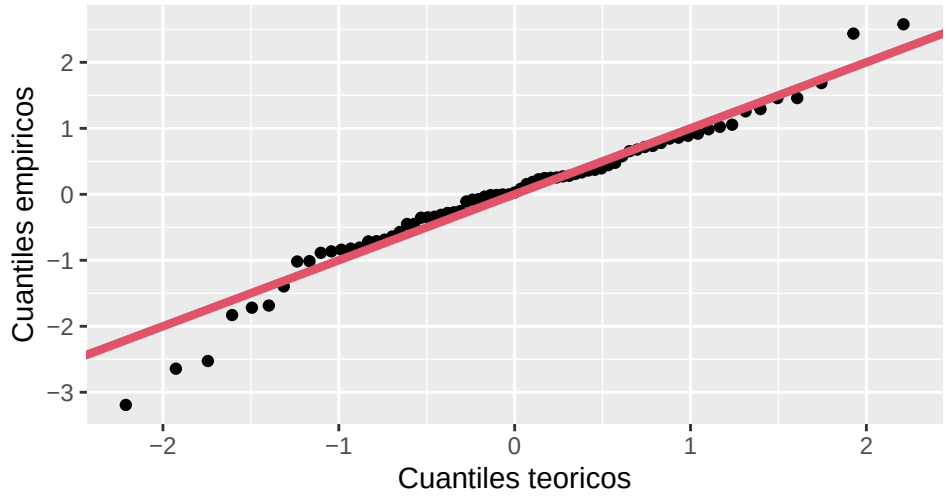

Figure 4: Comparación de los cuantiles teóricos y empíricos

Nuevamente, el gráfico refuerza la idea de que existe una asimetría en los extremos de la distribución. Se ve como la parte inferior de la cola es la que peor se ajusta a la recta, por lo que pareciera no ser del todo normal.

- **Shaphiro-Wilk**: se basa en la comparación de los cuantiles empíricos y teóricos bajo el supuesto de normalidad.
- **Jarque-Bera**: se basa en la comparación de los estadísticos de asimetría y kurtosis bajo el supuesto de normalidad.
- **Kolmogorov-Smirnov**: se basa en la máxima discrepacnoa entre la función de distribución empírica y la teórica bajo el supuesto 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 Jarque-Bera tiene un p-valor menor que 0.05, por lo que se rechaza la hipótesis nula,los residuos se distribuyen normal. Sin embargo, el test de Shapiro-Wilk y el test de Kolmogorov-Smirnov no rechazan la hipótesis nula. Por lo que no se puede afirmar que los residuos no son normales.

Datos Atípicos

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

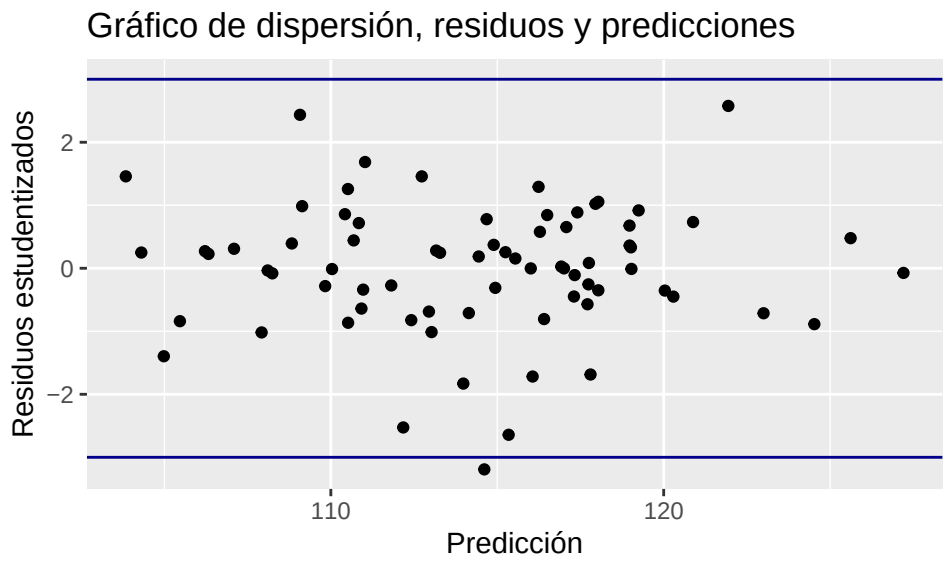

Figure 5: Presencia de Atípicos

Datos Influyentes

Para detectar la presencia de influyentes se utilizarán dos métodos: Leverage y Distancia de Cook.

Leverage: Este indicador es útil para determinar la influencia de cada observación sobre los valores ajustados, se presenta en la diagonal de la matriz H. Los puntos con un alto índice de Leverage tienen un potencial mayor para influir en la estimación de los coeficientes del modelo de regresión. Un punto de Leverage mayor que $\frac{2(k+1)}{n}$, donde k es el número de predictores y n es el número de observaciones, se considera un punto alto de Leverage.

Distancia de Cook: Este indicador cuantifica el cambio en el vector de estimaciones luego de remover la i-ésima observación. D_i cuantifica el cambio en el vector de valores ajustados. Para determinar si alguna observación tiene una influencia significativa $D_i > \frac{4}{n}$.

Leverage:

En base al indicador de Leverage, se puede observar la existencia de tres datos influyentes. Estos puntos podrían tener un impacto significativo en los resultados del análisis. En etapas posteriores, se evaluará como manejarlos, ya sea mediante su eliminación o su intervención.

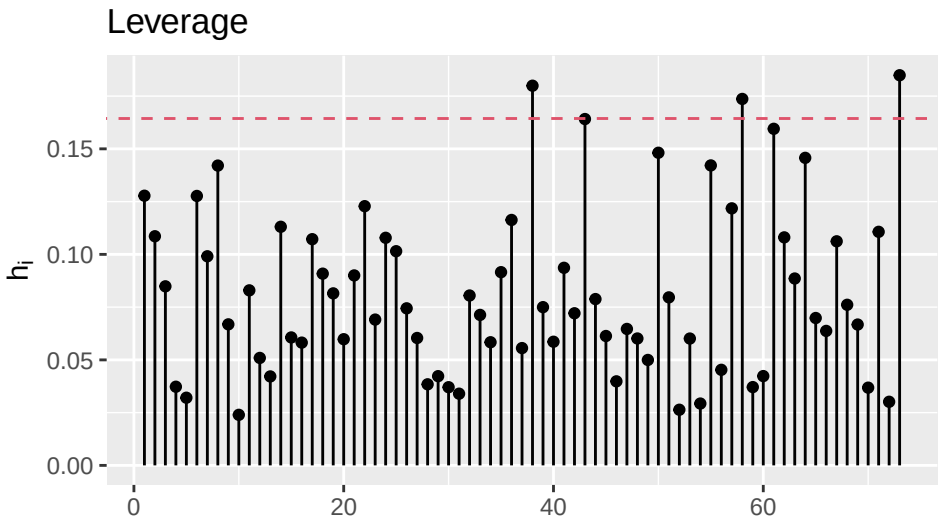

Figure 6: Leverage

Distancia de Cook:

En este caso, existen más datos influyentes que en el caso de Leverage. Sin embargo, se tratará de seguir el mismo enfoque, considerando la eliminación de estos datos o su intervención para poder continuar con el análisis.

Intervención de influyentes según 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)
```

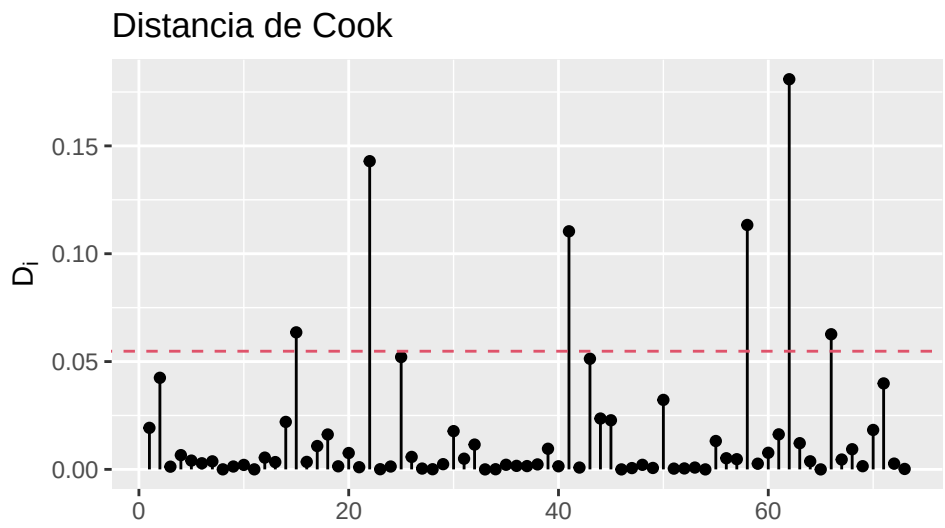

Figure 7: Distancia de Cook

```
D = 0.072055, p-value = 0.8288
alternative hypothesis: two-sided
```

Para este caso, se intervinieron las observaciones 38 y 73, que fueron identificadas como influyentes en el gráfico de este indicador.

Se testean los supuestos en el modelo intervenido según Leverage:

- Test de Multicolinealidad: Los valores de VIF fueron todos inferiores a 5, indicando ausencia de multicolinealidad significativa.
- Test de Linealidad: Todas las variables siguen una relación lineal adecuada.
- Test de Homocedasticidad: No se rechazó la hipótesis nula, con un p-valor de 0.704, indicando que no hay un problema de heterocedasticidad.
- Test de Normalidad: Se utilizaron los métodos de Shapiro-Wilk, Jarque-Bera y Kolmogorov-Smirnoc para evaluar normalidad en los residuos. Los tres métodos dan un p-valor mayor a 5% por lo tanto se acepta la normalidad.

Interviención de influyentes según 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
```

Se realizó el mismo análisis, pero según la Distancia de Cook. Para este caso, se intervinieron varios datos y finalmente se llegó a la conclusión que únicamente interviniendo el dato 62 se cumple con todos los supuestos.

- Test de Multicolinealidad: Los valores de VIF fueron todos inferiores a 5, indicando ausencia de multicolinealidad significativa.
- Test de Linealidad: Todas las variables siguen una relación lineal adecuada.
- Test de Homocedasticidad: No se rechaza la hipótesis nula, con un p-valor de 0.82783, indicando que no hay un problema de heterocedasticidad.
- Test de Normalidad: Se utilizaron los métodos de Shapiro-Wilk, Jarque-Bera y Kolmogorov-Smirnoc para evaluar normalidad en los residuos. Los tres métodos dan un p-valor mayor a 5% por lo tanto se asepta la normalidad.

Como solo es necesario intervenir una observación, la cual a su vez anteriormente se identificó como dato atípico, preferimos este modelo al creado interviniendo según Leverage.

```
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 consecuancia, se procederá a repetir todo el análisis excluyendo ambas.

Modelo sin variables no significativas:

Se procedió a eliminar las variables no significativas mencionadas anteriormente, *duracion_hip* y *stress*, del modelo que creamos anteriormente. Luego, se verificó que cumpla con los supuestos.

```
    Shapiro-Wilk normality test

data:  rstudent(mod_II2)
W = 0.96128, p-value = 0.02611
```

```
    Jarque Bera Test

data:  na.omit(rstudent(mod_II2))
X-squared = 7.2341, df = 2, p-value = 0.02686
```

```
    Exact one-sample Kolmogorov-Smirnov test
```

```
data:  rstudent(mod_II2)
D = 0.090553, p-value = 0.5652
alternative hypothesis: two-sided
```

En las pruebas de normalidad los resultados indicaron que el p-valor de los métodos Shapiro-Wilk y Jarque-Bera fue inferior al 5%, lo cual sugiere que no se cumple el supuesto de normalidad para los residuos del modelo.

Modelo Final

Al no lograr cumplir con el supuesto de normalidad, incluso después de eliminar las variables no significativas, se procede a crear otro modelo excluyendo esas variables del análisis, el cual será validado de forma tal que no necesite de dicho supuesto.

```
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 es significativo tanto global- como individualmente, con una variabilidad explicada (R^2) del 73,99%.

Multicolinealidad:

```
    edad sup_corp    pulso
1.410444 1.187482 1.432667
```

Observaciones influyentes:

```
[1] 62 41 22 58 66
```

```
[1] 73 61 38 58 55
```

Observaciones aípicas:

```
[1] 62
```

Debio a que se utilizaron diversas estrategias para cumplir con el supuesto de normalidad sin éxito, se optó por utilizar el método de *Randomization Test*, el cual se describirá en detalle a continuación.

Randomization Test

Una alternativa al método convencional para la validación de un modelo es el uso de métodos robustos ante el no cumplimiento de ciertos supuestos. En las etapas de diagnóstico anteriormente planteadas se ha rechazado el supuesto de normalidad, por lo que en este apartado se utilizará la **prueba basada en permutaciones** o **randomization test**, método el cual prescinde de dicho supuesto (o cualquier otro supuesto referente a la distribución de los datos).

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.

Para cada RLS y en cada réplica, se calcula la estimación del parámetro $\hat{\beta}_j$ correspondiente y su desvío estimado, y en función de los mismos se calcula el estadístico de interés, en este caso el mismo es el estadístico t, siendo este:

$$t = \frac{\hat{\beta}_j}{\widehat{desvo}}$$

A partir de esta colección de valores de t se aproximará la distribución del estadístico t y se calcula un p-valor como la proporción de valores de la colección de t que pertenecen al rango $(-|t_{obs}|, |t_{obs}|)$ (frente al total de los mismos), siendo t_{obs} el valor del estadístico t obtenido en las prueba de hipótesis individual del método convencional.

Si los p-valores obtenidos en la prueba de permutaciones son similares a los obtenidos en las pruebas de hipótesis de significación individual, el modelo es válido, aún sin necesidad de cumplir el supuesto de normalidad.

	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.

Evaluación del desempeño predictivo del modelo

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, mas 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).

El mismo consiste en quitar una observación del dataset y calcular los coeficientes del modelo con este nuevo subconjunto de datos. Esto se lleva a cabo de manera iterativa con cada una de las observaciones de nuestros datos. En cada iteración se predice además el valor de la variable de respuesta de la observación quitada en función de los valores que las variables explicativas toman en esta observación y los valores de las nuevas estimaciones de los parámetros del modelo.

Posteriormente se calcula el R^2 de los resultados obtenidos por este método, es decir:

$$R^2 = 1 - \frac{SCRes}{SCTotal} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{yi})^2}{\sum_{i=1}^{n}(yi - \bar{y})^2}$$

donde SCRes es la suma de los cuadrados de los residuos y SCTotal es la suma de los cuadrados Totales. En nuestros datos y, la variable de respuesta, es presion_ert.

Table 3: R cuadrado

R2_conv	R2_rand
0.739884	0.7042639

EL R^2 se encuentra por encima del 70%, lo cual permite afirmar que el mismo explica sificientemente bien los datos de la muestra.

Se observa que el R^2 por el método LOOCV es un 3,5% más bajo que el R^2 de la prueba de significación global. Esto se debe a que el modelo predice mejor los datos con los que fue creado en comparación con datos nuevos; a este efecto se le llama *overfitting*. En nuestro caso el overfitting es relativamente bajo. Podemos afirmar entonces que el modelo es competente en sus predicciones.

Conclusión:

Para finalizar se puede concluir que el modelo de regresión lineal multiple proporciona una buena representación de la relación entre las variables eplicativas y la variable de respuesta (presión arterial). Luego de pasar ciertos desafíos con la normalidad de los residuos, se ha confirmado que el modelo es fiable y tiene un buen desempeño predictivo. Por lo tanto, se puede decir que el mismo es adecuado para el propósito del análisis y puede ser utilizado para predecir la presión arterial basado en las variables edad, superficie corporal y pulso incluidas en el estudio.