INFORME FINAL

Modelos Lineales

Daniel Czarnievicz
CI: 4.744.781-9

Lucía Coudet
CI: 4.545.399-9

Montevideo, julio de 2017

Índice

1. Introducción

El objetivo de este trabajo es implementar las técnicas de análisis estudiadas en el curso de modelos lineales, en particular, el modelo de regresión lineal múltiple.

En una primera instancia, se procederá a hacer un análisis descriptivo de los datos. Posteriormente, se estimará el modelo con el total de las variables explicativas, y se evaluará significación individual y significación global del mismo.

En la etapa de diagnóstico del modelo, mediante herramental estadístico y análisis visual de gráficos auxiliares se testeará normalidad de los errores, así como también la homoscedasticidad de los residuos. En función de los resultados obtenidos en esta etapa, se estimarán algunos modeos alternativos. No se ensayaron pruebas de restricciones lineales sobre los parámetros por no contar con un marco teórico que justificara dichas pruebas.

Por último, se utilizarán técnicas de *Cross Validation* para obtener estimaciones del Error Cuadrático Medio de los distintos modelos. Para ellos se perseguirán dos enfoques: *LOOCV*, y *k-fold CV* con k=10.

A lo largo de este trabajo se utilizará un nivel de significación del 5 % ($\alpha = 0{,}05$). Se utilizará la letra k para designar a la cantidad de regresores de los modelos, y se utilizará la letra n para designar la cantidad de observaciones en el modelo ($n = 1\,115$).

2. Descripción de la base

Para este informe se trabajó con una base de nacimientos. La misma consta de 1115 observaciones de nacimientos y con las siguiente variables:

1. *educ*: variable categórica que toma valores entre los enteros 0 y 17 según los años de educación completados por la madre.
2. *fuma*: variable dicotómica que toma valores 1 si la madre fuma y 0 si la madre no fuma.
3. *gest*: variable categórica que mide el tiempo de gestación del recién nacido (en semanas), y toma valores entre los enteros 20 y 40.

Es importante destacar que una de las limitaciones con la que se cuenta es no tener información sobre cuáles de las observaciones refieren a nacidos vivos y cuáles no, lo cual sería de gran utilidad en particular al momento de analizar heteroscedasticidad.

A efectos de tener un primer acercamiento con la estructura subyacente en la base de datos, se obtienen algunas estadísticas descriptivas.

Tabla 1: estadísticos descriptivos de las variables.

	Peso del nacido	**Educación de la madre**	**Tiempo de gestación**
Mín.	284	-	20,00
1st Q	2.900	11,00	38,00
Median	3.267	12,00	39,00
Mean	3.220	12,27	38,84
3rd Q	3.630	13,00	40,00
Max.	4.830	17,00	43,00

En la **Tabla 1** interesa notar que:

- La variable *peso* varía entre un mínimo de 284 y un máximo de 4.830 gramos, con una media de 3.220.
- La variable *educ* toma valores entre 0 y 17 años, con una media de aproximadamente 12 años.
- La variable *gest* toma valores entre 20 y 43 semanas, con una media de aproximadamente 39.

Por su parte, la proporción de madres fumadoras en el total es de 0,759 aproximadamente.

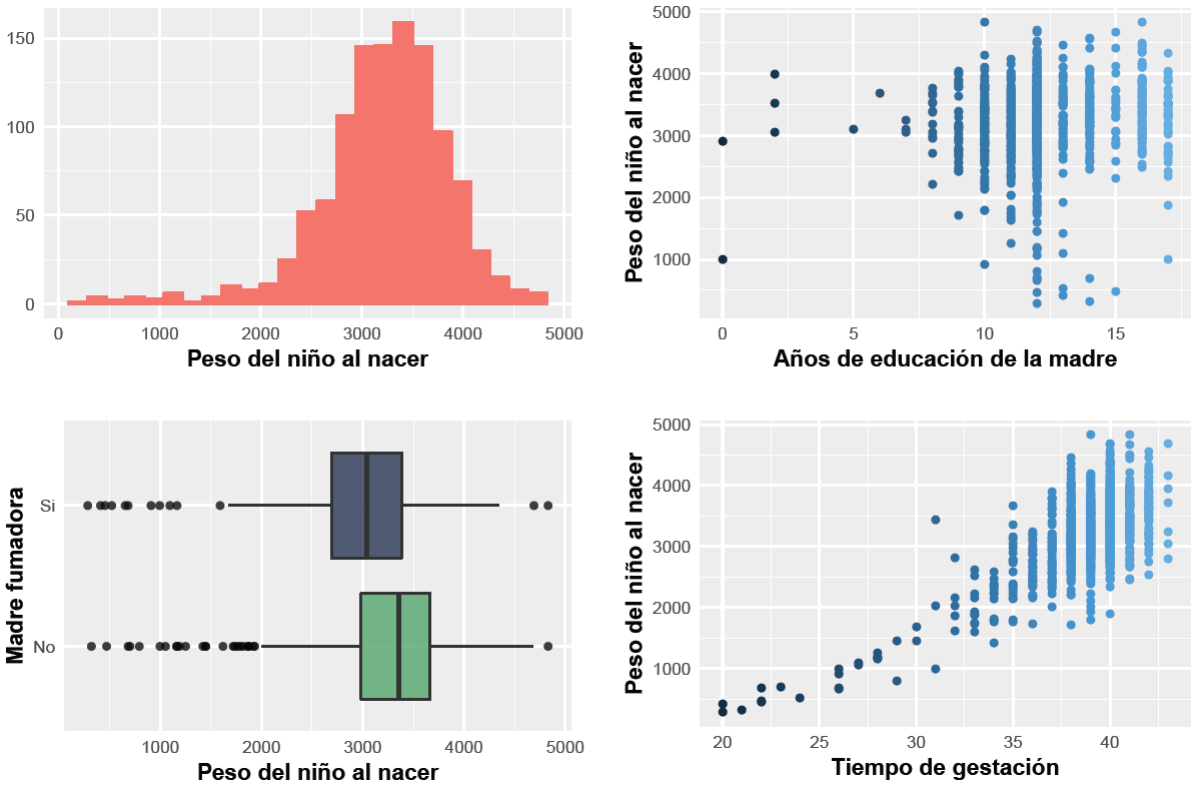
Tabla 2: proporción de madres fumadoras.

	Total	**Proporción**
Fuma	846	0,759
No fuma	269	0,241
Total	1.115	1,00

En el **Gráfico 1** se presenta:

- Histograma de la variable *peso*.
- Scatterplot de las variables *peso* y *educ*.
- Boxplot de la variable *peso* según la variable *fuma*.
- Scatterplot de las variables *peso* y *gest*.

Gráfico 1: descripción de las variables utilizadas.



La base parece presentar heterocedasticidad generada por la variable *gest*. Los residuos de los distintos modelos fueron sometidos al test de heterocedasticidad de White, donde se pudo descartar esta situación. Es probable que la hipótesis de heterocedasticidad se rechaze por la baja cantidad de observaciones en la cola del gráfico.

También se realizó una prueba de igualdad de medias para la variable peso, separando entre las madres fumadoras de las no fumadoras, descartandose que las medias fueran iguales. El análisis de ANOVA sobre la variable *fuma* confirmó estos resultados.

3. Modelo 1

Como ya se mencionó, el primer modelo estimado consiste en la regresión de la variable *peso*, contra las demás variables de la base. Por lo tanto, el modelo presenta la siguiente especificación:

$$peso_i = \beta_0 + \beta_1\ educ_i + \beta_2\ fuma_i + \beta_3\ gest_i + \varepsilon_i$$

Tabla 3: estimación del modelo 1

```
Residuals:
     Min       1Q   Median       3Q      Max
-1370.56  -298.35   -10.66   289.35  1486.90

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3200.582    210.838 -15.180  < 2e-16 ***
educ           14.956      6.506   2.299   0.0217 *
fumasi       -174.368     32.152  -5.423 7.18e-08 ***
gest          161.651      5.027  32.156  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 443.4 on 1111 degrees of freedom
Multiple R-squared:  0.5115,    Adjusted R-squared:  0.5102
F-statistic: 387.7 on 3 and 1111 DF,  p-value: < 2.2e-16
```

3.1. Significación individual de los regresores

Para cada uno de los regresores se relaiza la siguiente prueba de hipótesis:

$$H_0)\ \ \beta_i = 0 \quad Vs. \quad H_1)\beta_i \neq 0$$

Con región crítica: $RC = \left\{ \left(\mathbf{Xy} \right) \Big/ |e| \geq t_{n-k-1}(1 - {}^{\alpha}/_{2}) \right\}$

Para la cual se utiliza el estadístico: $e = \dfrac{\hat{\beta}_i}{\hat{sd}(\hat{\beta}_i)} \overset{H_0}{\sim} t_{n-k-1}$

Siguiendo el criterio del p-valor, la evidencia empírica sugiere que las variables *educ*, *fuma*, y *gest* son individualmente significativas para explicar a *peso*. Con lo cual se rechaza la hipótesis nula de no significación de los regresores, a un nivel $\alpha = 0{,}05$.

3.2. Significación global del modelo

Para cada uno de los regresores se realizó la siguiente prueba de hipótesis:

$$H_0) \;\; \beta_i = 0 \, \forall i = 1; \ldots; k+1 \quad Vs. \quad H_1) \exists i / \beta_i \neq 0 \; i = 1; \ldots; k+1$$

Con región crítica: $RC = \left\{ \left(\mathbf{Xy} \right) \Big/ F \geq F_{k-1;n-k-1}(1-\alpha) \right\}$

Para la cual se utiliza el estadístico:

$$F = \frac{(\mathbf{R}\hat{\beta} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^1(\mathbf{R}\hat{\beta} - \mathbf{r})/(k-1)}{\hat{\varepsilon}'\hat{\varepsilon}/(n-k-1)} \overset{H_0}{\sim} F_{k-1;n-k-1}$$

donde $\mathbf{R} = \mathbf{I}_k$ y $\mathbf{r} = \mathbf{0}_k$.

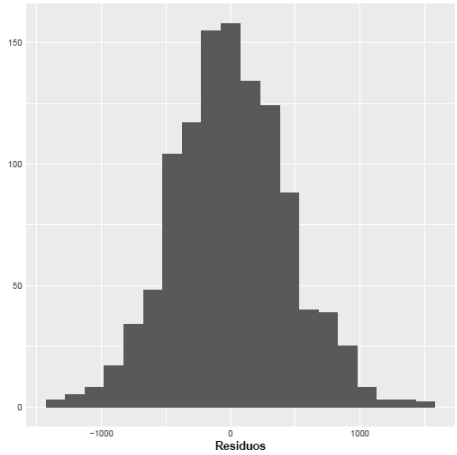
Siguiendo el criterio del p-valor, a un nivel del 5 %, la evidencia empírica sugiere que el modelo es globalmente significativo. Esto implica que, dada la evidencia empíica con la que se cuenta, no es posible rechazar la hipótesis de que los regresores utilizadas no contribuyen a explicar el regresando (*peso*).

3.3. Diagnóstico del modelo

3.3.1. Normalidad de los residuos

El histograma de los residuos estandarizados sugiere que podría llegar a ser razonable suponer distribución normal de los mismos.

Gráfico 2: histograma de los residuos estandarizados del modelo 1.



Por su parte, los test de normalidad Shapiro-Wilk y Jarque-Bera, según el criterio del p-valor y para un nivel de significación del 5 %, no rechazan la hipótesis nula de normalidad de los residuos. A continuación, las salidas de R correspondientes:

Tabla 4: tests de normalidad de residuos del modelo 1

```
        Shapiro-Wilk normality test

data:  m1$residuals
W = 0.99829, p-value = 0.334
```

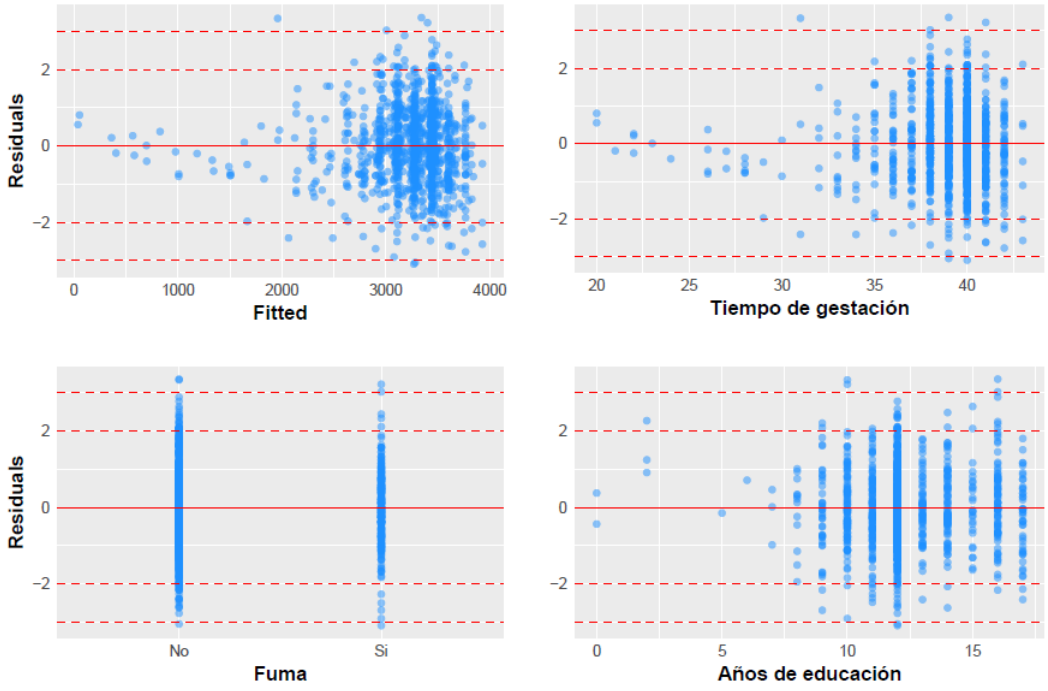
```
        Jarque Bera Test

data:  m1$residuals
X-squared = 3.2236, df = 2, p-value = 0.1995
```

3.3.2. Homoscedasticidad

En una primera instancia, se optó por recurrir a un análisis visual de los residuos estandarizados del modelo especificado. A continuación vemos los gráficos de los residuos en el eje de las ordenadas, y los valores ajustados y las regresoras en el eje de las abscisas, los cuales a primera vista parecerían sugerir la existencia de heteroscedasticidad.

No obstante, es importante señalar que si dejamos de lado aquellas pocas observaciones que se separan de la nube de puntos principal, como lo es en el gráfico Tiempo de gestación Vs. Residuos (observaciones con menos de 30 semanas de gestación), resulta razonable suponer que el modelo es homoscedastico.

Gráfico 3: análisis de los residuos estandarizados del modelo 1.



Por lo tanto, se propone acudir a un contraste de heteroscedasticidad. El contraste de Goldfeld y Quandt no resulta apropiado dado que se entiende que, con esta estructura de datos, se estaría forzando un rechazo de homoscedasticidad al separar a las observaciones en dos grupos, uno de los cuales quedaría conformado con los considerados como potenciales "outliers". En virtud de ello, se entiende que una posible alternativa es realizar el contraste propuesto por White.

Se sometieron los residuos a la siguiente prueba de hipótesis:

$$H_0) \text{ Homoscedasticidad} \quad Vs. \quad H_1) \text{ Heteroscedasticidad}$$

Para contrastar dicha hipótesis nula, el contraste de White propone los siguientes pasos:

- Estimar el modelo por MCO y guardar los residuos $\hat{\varepsilon}$.
- Estimar el siguiente model: $\hat{\varepsilon}_i^2 = \delta_0 + \sum_{j=1}^{k} \gamma_j X_{ij} + \sum_{j=1}^{k} \sum_{l=1}^{k} \delta_{jl} X_{ij} X_{il} + \mu_i$

  Esto implica regresar el cuadrado de los residuos de la regresión original, contra una constante, todos los regresores originalmente utilizados y sus interacciones (incluyendo con sigo mismos, es decir, cuadrados[1]).
- Contrastar: $H_0)\ \delta_{jl} = \gamma_j = 0 \quad Vs. \quad H_1)\ \exists\ \delta_{jl}$ o γ_{jl} distinto de cero.

  Para esto se utiliza el estadístico $\lambda = nR^2_{aux}$, el cual converge en distribución (bajo H_0) a una χ^2_p donde n es el número de observaciones, R^2_{aux} es el coeficiente de determinación de la regresión auxiliar, y p es el número de regresores de la regresión auxiliar menos uno (por la constante).

  El test tiene la siguiente región crítica: $RC = \left\{ \left(\mathbf{Xy} \right) / \lambda \geq \chi^2_p(1-\alpha) \right\}$

Esta prueba presenta dos potenciales problemas: no permite determinar qué variable es la que genera heteroscedasticidad, y puede recoger problemas de especificación en el modelo original (dado que trabaja con las interacciones de los regresores utilizados en el modelo original).

A partir de la implementación de la prueba, con un $R^2_{aux} = 0{,}009112$, no se rechaza la hipótesis nula de homoscedasticidad para el nivel de significación especificado.

[1] Deben excluirse los cuadrados de las variables dicotómicas, dado que generarían problemas de multicolinealidad.

Tabla 5: test de White del modelo 1

```
Residuals:
    Min      1Q  Median      3Q     Max
-279345 -173045 -104048   44594 1996275

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -49433.7  1065050.2  -0.046    0.963
educ         -14431.6    61408.7  -0.235    0.814
gest          21406.3    43053.5   0.497    0.619
fumasi      -546178.3   319864.6  -1.708    0.088 .
I(educ^2)      1653.5     1028.0   1.609    0.108
I(gest^2)      -201.4      488.2  -0.413    0.680
educ:gest      -732.2     1593.6  -0.459    0.646
educ:fumasi   12317.5    12446.0   0.990    0.323
gest:fumasi    9455.9     7323.3   1.291    0.197
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 290800 on 1106 degrees of freedom
Multiple R-squared:  0.009112,  Adjusted R-squared:  0.001945
F-statistic: 1.271 on 8 and 1106 DF,  p-value: 0.2545
```

4. Modelo 2

Como segundo modelo se estimó una regresión polinómica donde se incluyó el cubo de la variable *gest*.

$$peso_i = \beta_0 + \beta_1\ educ_i + \beta_2\ fuma_i + \beta_3\ gest_i + \beta_4\ gest_i^3 + \varepsilon_i$$

Previamente se ensayaron regresiones con el cuadrado de la variable *gest*, y con el cuadrado y el cubo de dicha variable. Cuando solo se incluyó el cuadrado de la variable, el mismo resultó no significativo. Por otra parte, cuando se incluyeron tanto cubos como cuadrados solo los primeros resultaron significativos. Por principio de parsimonia, se opró por conservar únicamente el cubo de dicha variable.

las pruebas de significación indivicual de los regresores, asi como la de significación global del modelo se realizaron siguiendo las mismas pruebas de hipótesis especificadas en las secciones 3.1 y 3.2. No se reproducen aquí las especificaciones de las mismas. No obstante, de la salida de R podemos ver que todos los regresores y el modelo resultaron significativos al 5 %.

Tabla 6: estimación del modelo 2

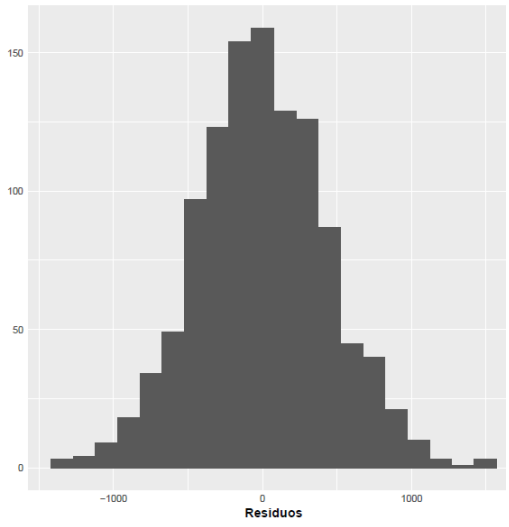
```
Residuals:
     Min       1Q   Median       3Q      Max
-1364.87  -294.06   -10.15   291.46  1476.04

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.375e+03  6.272e+02  -6.976 5.22e-12 ***
educ         1.547e+01  6.503e+00   2.378   0.0176 *
fumasi      -1.746e+02  3.211e+01  -5.439 6.59e-08 ***
gest         2.132e+02  2.642e+01   8.071 1.80e-15 ***
I(gest^3)   -1.406e-02  7.070e-03  -1.988   0.0470 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 442.8 on 1110 degrees of freedom
Multiple R-squared:  0.5132,    Adjusted R-squared:  0.5115
F-statistic: 292.6 on 4 and 1110 DF,  p-value: < 2.2e-16
```

4.1. Diagnóstico del modelo

4.1.1. Normalidad de los residuos

El histograma de los residuos estandarizados sugiere que podría llegar a ser razonable suponer distribución normal de los mismos.

Gráfico 4: histograma de los residuos estandarizados del modelo 2.



Por su parte, los test de normalidad Shapiro-Wilk y Jarque-Bera, según el criterio del p-valor y para un nivel de significación del 5 %, no rechazan la hipótesis nula de normalidad de los residuos. A continuación, las salidas de R correspondientes:

Tabla 7: tests de normalidad de residuos del modelo 2

```
        Shapiro-Wilk normality test

data:  resi
W = 0.99849, p-value = 0.4502
```

```
        Jarque Bera Test

data:  resi
X-squared = 2.7324, df = 2, p-value = 0.2551
```

4.1.2. Homoscedasticidad

El análisis visual de los residuos lleva a conclusiones similares respecto a las del modelo 1, por lo que se procede a implementar el test de homosedasticidad de White. De la salida se desprende que no se rechaza la hipótesis nula de homoscedasticidad al 5 %.

Gráfico 5: análisis de los residuos estandarizados del modelo 2.

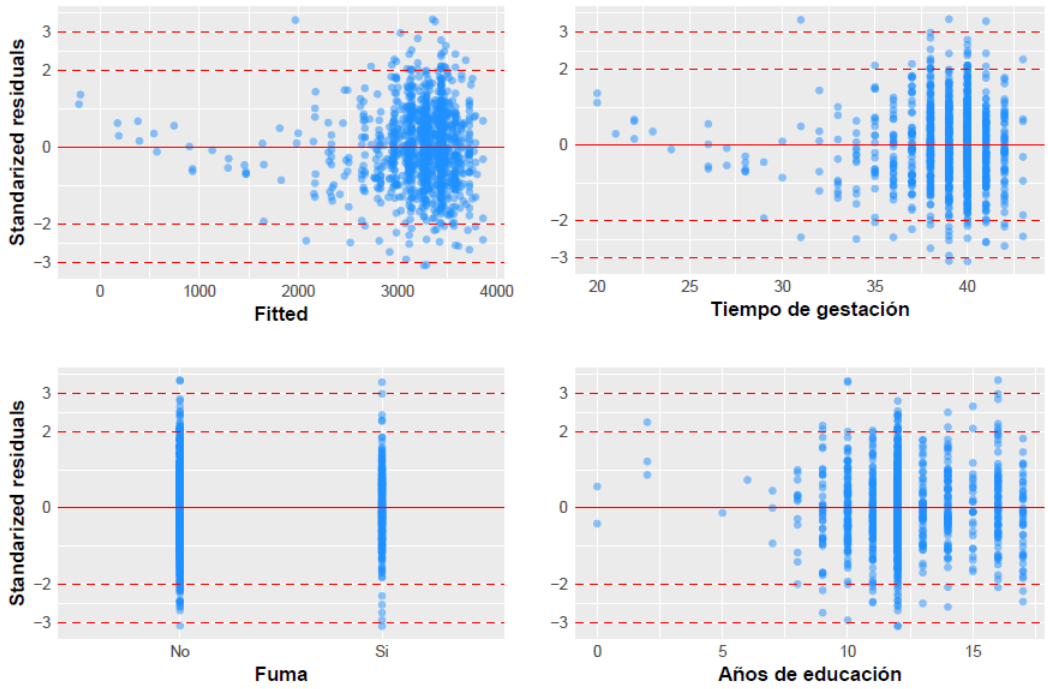


Tabla 8: test de White del modelo 2

```
Residuals:
    Min      1Q  Median      3Q     Max
-307729 -169082 -105830   48057 1973036

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       9.581e+07  6.704e+07   1.429   0.1532
educ             -1.066e+05  2.262e+05  -0.471   0.6377
gest             -1.430e+07  1.038e+07  -1.378   0.1684
fumasi            2.201e+05  9.936e+05   0.222   0.8247
I(gest^3)        -2.171e+04  1.729e+04  -1.256   0.2095
I(educ^2)         1.723e+03  1.026e+03   1.679   0.0935 .
I(gest^2)         8.183e+05  6.196e+05   1.321   0.1869
I(I(gest^3)^2)   -1.277e-02  1.238e-02  -1.032   0.3024
educ:gest         3.414e+03  9.806e+03   0.348   0.7278
educ:fumasi       1.104e+04  1.242e+04   0.889   0.3742
educ:I(gest^3)   -1.177e+00  2.649e+00  -0.444   0.6569
gest:fumasi      -2.544e+04  4.140e+04  -0.615   0.5390
gest:I(gest^3)    2.380e+02  2.008e+02   1.185   0.2362
fumasi:I(gest^3)  1.022e+01  1.110e+01   0.921   0.3574
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 287900 on 1101 degrees of freedom
Multiple R-squared:  0.01669,   Adjusted R-squared:  0.005075
F-statistic: 1.437 on 13 and 1101 DF,  p-value: 0.1354
```

5. Modelo 3

Por último se busca entender si existe (es estadísticamente significativa) una interacción entre las variables *fuma* y *gest*. Esto se debe a que, al menos desde un punto de vista teórico, parecería lógico que dicha interacción sí existiera. Para esto se estimo el siguiente modelo:

$$peso_i = \beta_0 + \beta_1\ educ_i + \beta_2\ gest_i + \beta_3\ gest_i fuma_i + \varepsilon_i$$

Al igual que con el caso polinómico, aquí también se estimaron distintos modelos que buscaban contemplar la posible interacción entre las variables *gest* y *fuma*. En el anexo de este trabajo se presentan todos los modelos que fueron descartados.

A continuación se presenta la salida correspondiente al modelo 3.

Tabla 9: estimación del modelo 3

```
Residuals:
     Min       1Q   Median       3Q      Max
-1366.30  -296.27   -11.13   288.89  1493.33

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3275.0198   207.9171 -15.752  < 2e-16 ***
educ           15.0179     6.5044   2.309   0.0211 *
gest          163.5324     4.9923  32.757  < 2e-16 ***
gest:fumasi    -4.5048     0.8316  -5.417 7.43e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 443.4 on 1111 degrees of freedom
Multiple R-squared:  0.5114,    Adjusted R-squared:  0.5101
F-statistic: 387.7 on 3 and 1111 DF,  p-value: < 2.2e-16
```

Una vez más, para las pruebas de significación individual de los regresores y global del modelo se siguió la especificación detallada en las secciones 3.1 y 3.2 respectivamente. Tal como se aprecia en la salida de la Tabla 9, todos las variables incluidas y el modelo resultaron significativos al 5 %.

5.1. Diagnóstico del modelo

5.1.1. Normalidad de los residuos

El histograma de los residuos estandarizados sugiere que podría llegar a ser razonable suponer distribución normal de los mismos.

Por su parte, los test de normalidad Shapiro-Wilk y Jarque-Bera, según el criterio del p-valor y para un nivel de significación del 5 %, no rechazan la hipótesis nula de normalidad de los residuos. A continuación, las salidas de R correspondientes.

Gráfico 6: histograma de los residuos estandarizados del modelo 3.

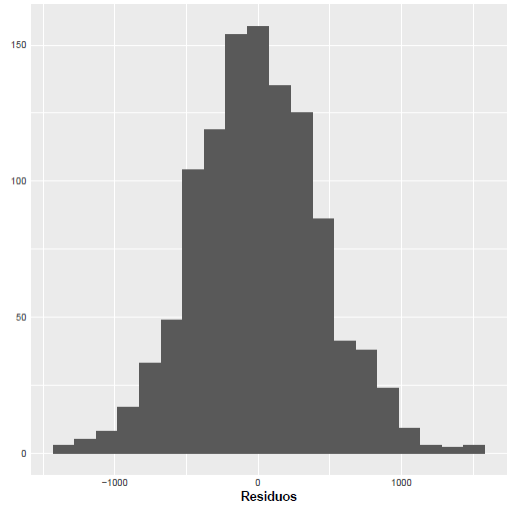


Tabla 10: tests de normalidad de residuos del modelo 3

```
        Shapiro-Wilk normality test

data:  resi
W = 0.99825, p-value = 0.3125
```

```
        Jarque Bera Test

data:  resi
X-squared = 3.4461, df = 2, p-value = 0.1785
```

5.1.2. Homoscedasticidad

El análisis visual de los residuos lleva a conclusiones similares respecto a las del modelo 1, por lo que se procede a implementar el test de homosedasticidad de White. De la salida correspondiente a la Tabla 11 se desprende que no se rechaza la hipótesis nula de homoscedasticidad al 5 %.

Gráfico 7: análisis de los residuos estandarizados del modelo 3.

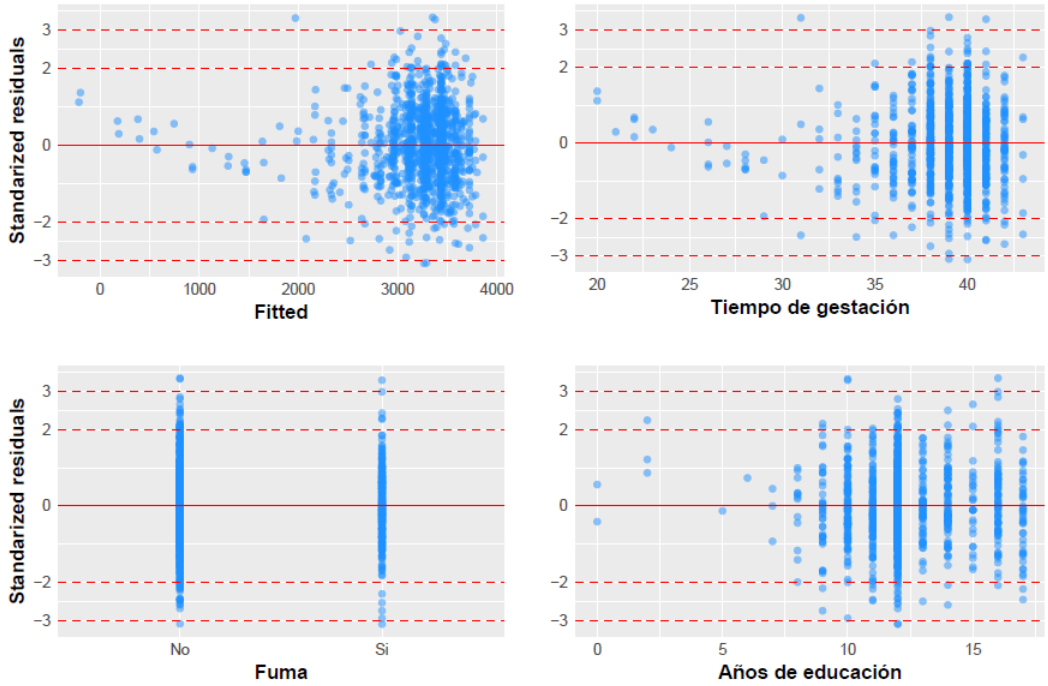


Tabla 11: test de White del modelo 3

```
Residuals:
    Min      1Q  Median      3Q     Max
-330648 -173487 -103834   43009 2027534

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      -434155.8   984329.2  -0.441    0.659
educ                7152.4    57215.3   0.125    0.901
gest               34499.1    40197.0   0.858    0.391
I(educ^2)           1167.6      912.2   1.280    0.201
I(gest^2)           -348.4      469.2  -0.742    0.458
educ:gest           -923.5     1575.4  -0.586    0.558
gest:fumasi       -14541.7     9317.7  -1.561    0.119
I(gest^2):fumasi     349.6      238.8   1.464    0.144

Residual standard error: 291200 on 1107 degrees of freedom
Multiple R-squared:  0.008486,  Adjusted R-squared:  0.002216
F-statistic: 1.353 on 7 and 1107 DF,  p-value: 0.2216
```

6. Selección del modelo

Existen varios métodos para seleccionar entre distintos modelos. En la siguient tabla se reportan el coeficiente de determinación R^2 asi como también su versión ajustada R_a^2. Por otra parte, se reportan tambien los criterios de información AIC y BIC para cada modelo.

Tabla 12: selección de modelos.

	$\mathbf{R^2}$	$\mathbf{R_a^2}$	**AIC**	**BIC**
M1	0,5115	0,5102	16.760,85	16.785,94
M2	0,5132	0,5115	16.758,89	16.788,99
M3	0,5114	0,5101	16.760,92	16.786,00

Salvo por el críterio BIC, todos los demás criterios señalan al modelo 2 como el que mejor explica la varianza en el regresando (*peso*).

Por otra parte podrían también implementarse métodos basados en los críterios de información, como ser el método *Forward* o el método *Backward*. Estos métodos se vuelven muy útiles cuando se desconoce la distribución de los residuos, y por tanto, se vuelve compleja la relaización de inferencia sobre los parámetros[2]. Pero en nuestro caso, para todos lo modelos ensayados, la normalidad no fue un problema, por lo que no se justifica su uso (dado que podemos recurrir a pruebas de hipótesis).

6.1. Cross-Validation

No obstante todo lo anterior, tal vez la técnica más poderosa de selección de modelos sea la validación cruzada. Para este trabajo se utilizaron dos enforques de la misma, $LOOCV$[3] y *k-fold CV*. En ambos casos el objetivos es estimar el *test error* del modelo.

El procedimiento de *Cross-Validation* es el siguiente. Primero se separa la base de datos en dos, una de testeo, y una de entrenamiento. Luego se utiliza la base de entrenamiento para estimar el modelo. Acto seguido se utiliza el modelo para predecir el regresando en la base de testeo. Por último se calcula el error cuadrático medio de dichas predicciones.

La diferencia entre $LOOCV$ y *k-fold CV* pasa por la selección de la partición de la base a utilizar. Mientras en $LOOCV$ se entrenan n modelos con $n-1$ observaciones cada uno, en *k-fold CV* se toman k particiones aproximadamente iguales de la base. Luego, en cada etapa, una de dichas particiones es utilizada como base de testeo, y las demás se utilizan para estimar el modelo. De nuevo, luego de realizado esto se calcula el promedio de los errores cuadráticos medios de cada submuestra.

[2]En estos casos también podrían obtenerse regiones críticas mediante *Bootstrap*

[3]Leave-one-out cross-validation

	LOOCV	*k-fold CV*
Error Cuadrático Medio estimado	$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} ECM_i$	$CV_{(n)} = \frac{1}{k}\sum_{i=1}^{k} ECM_i$

A continuación se presentan los resultados de los métodos de cross-validation implementados para los tres modelos.

Tabla 13: cross-validation.

	LOOCV	*k-fold CV*
M1	197.246,7	197.272,6
M2	196.950,2	197.110,8
M3	197.254,7	196.842,3

Por métodos de *LOOCV* el modelo a elegir sería el modelo 2, pero por *k-fold CV* deberíamos quedarnos con el modelo 3. Dado que los demás criterios de selección brindaron evidencia respecto de la capacidad del modelo 2, este fue el seleccionado por el equipo.

7. Observaciones influyentes

En el presenta apartado presentamos los resultados encontrados en el análisis de observaciones influyentes.

Se comenzó utilizando el método de clustering agregativo por single-linkage. Este método es muy sencible a la presencia de observaciones atípicas, por lo que tiene una muy buena performance a la hora de detectar observaciones influyentes en un modelo lineal, en tanto y en cuanto no se utilice con todas las variables, sino con los regresores únicamente. Tal como se observa en el gráfico 8, únicamente las observaciones 2, 4, y 6 toman valores atípicos en los regresores. Por ser solo tres, no parece que su influencia en el modelo pueda ser muy grande. No obstante se realizó un estudio de *leverage* y de distancias de Cook para corroborar esta afirmación. Los resultados de dicho análisis se presentan en los gráficos 9, 10, y 11. Estos sugieren que son otras las observaciones influyentes, pero nuevamente, la capacidad de influir de estas es baja.

Gráfico 8: clustering de regresores (dingle-linkage).

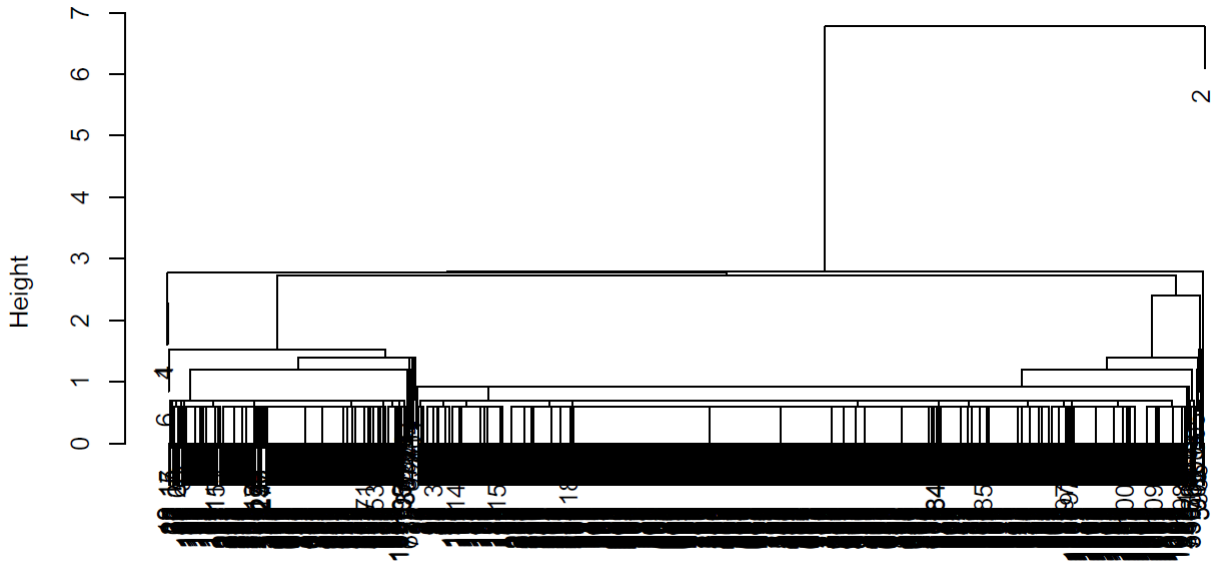


Gráfico 9: distancia de Cook para cada observación.

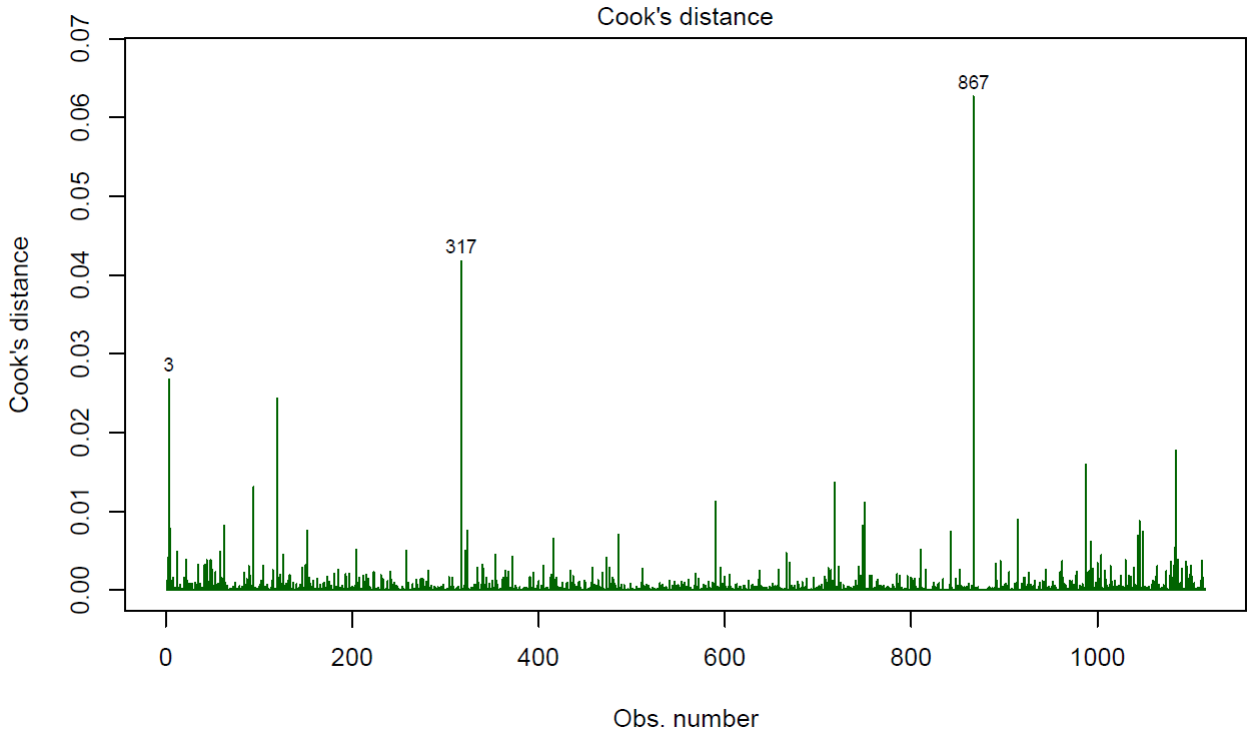


Gráfico 10: distancias de Cook Vs. leverage.

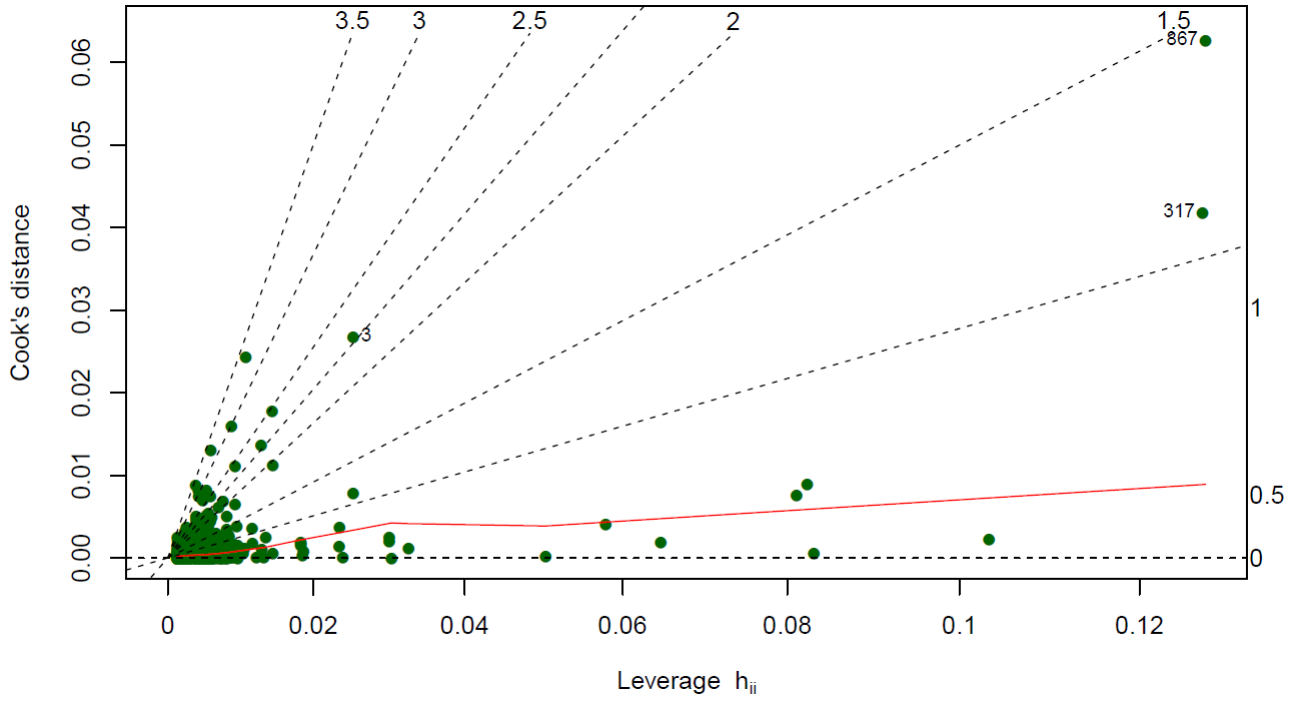
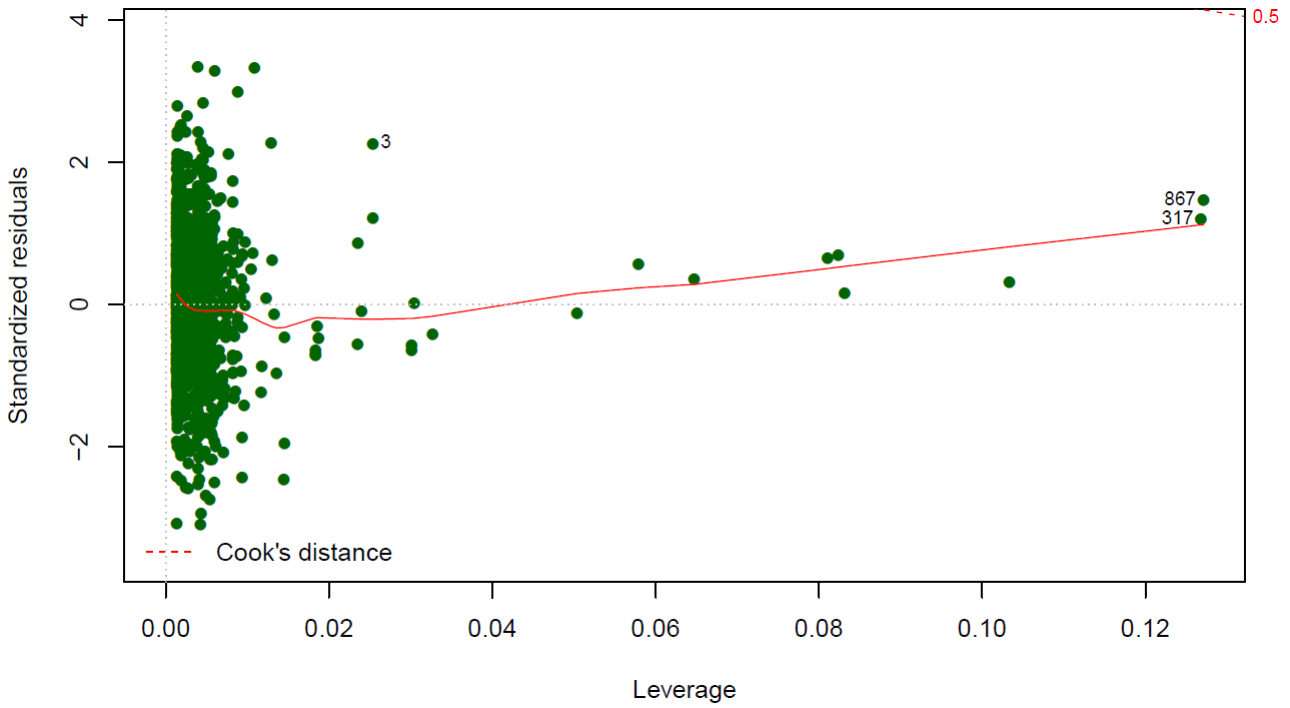


Gráfico 11: residuos estandarizados Vs. leverage.



8. Interpretación de los resultados y conclusiones

Habiendo seleccionado un modelo adecuado para explicar el peso de los recien nacidos en función de las variables *gest*, *fuma*, y *educ*, procedemos a interpretar los resultados del mismo.

- Cosntante: sin interpretación para el intercepto.
- Educación: antes un aumento de uno en los años de educación de la madre, se espera un aumento de 15,47 gramos en el peso del recien nacido, *ceteris paribus*.
- Fuma: se espera que el peso de los recien nacidos está 174,6 gramos por debajo de la media cuando la madre es fumadora, *ceteris paribus*.
- Ante un aumento de una semana de gestación se espera que el peso del recien nacido aumente en $\hat{\beta}_3 + 2\hat{\beta}_4 \ gest_i^2$, *ceteris paribus*. Nótese que el efecto de la cantidad de semanas de gestación sobre el peso esperado del recien nacido no es lineal y constante, sino que depende del punto de comparación.

De esto puede concluirse que mejorar el nivel de educación de las madres, así como implementar campañas contra el tabaquismo mejoraría los embarazos y ayudarían a prevenir el nacimiento de niños de bajo peso, lo cual pone en riesgo vida del recién nacido.

Anexo - Modelos descartados

Polinomios de educación

```
Call:
lm(formula = peso ~ gest + fuma + poly(educ, 2), data = naci)

Residuals:
     Min       1Q   Median       3Q      Max
-1365.33  -299.62    -9.75   287.09  1480.05

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -3026.474    197.069 -15.357  < 2e-16 ***
gest             161.885      5.031  32.177  < 2e-16 ***
fumasi          -173.283     32.164  -5.387 8.72e-08 ***
poly(educ, 2)1  1048.867    455.311   2.304   0.0214 *
poly(educ, 2)2   488.471    443.885   1.100   0.2714
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 443.3 on 1110 degrees of freedom
Multiple R-squared:  0.512,     Adjusted R-squared:  0.5103
F-statistic: 291.2 on 4 and 1110 DF,  p-value: < 2.2e-16


Call:
lm(formula = peso ~ gest + fuma + poly(educ, 3), data = naci)

Residuals:
     Min       1Q   Median       3Q      Max
-1373.34  -298.67    -6.34   286.78  1503.84

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -3041.784    197.264 -15.420  < 2e-16 ***
gest             162.279      5.036  32.223  < 2e-16 ***
fumasi          -173.197     32.149  -5.387 8.73e-08 ***
poly(educ, 3)1  1046.950    455.096   2.301   0.0216 *
poly(educ, 3)2   489.848    443.674   1.104   0.2698
poly(educ, 3)3  -636.918    443.803  -1.435   0.1515
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 443.1 on 1109 degrees of freedom
Multiple R-squared:  0.5129,    Adjusted R-squared:  0.5107
F-statistic: 233.6 on 5 and 1109 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = peso ~ gest + fuma + educ + I(educ^3), data = naci)

Residuals:
    Min      1Q  Median      3Q     Max
-1365.8  -299.1    -9.3   289.5  1477.3

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.071e+03  2.591e+02 -11.855  < 2e-16 ***
gest         1.618e+02  5.030e+00  32.163  < 2e-16 ***
fumasi      -1.735e+02  3.217e+01  -5.394  8.4e-08 ***
educ        -2.034e+00  2.078e+01  -0.098    0.922
I(educ^3)    3.643e-02  4.231e-02   0.861    0.389
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 443.4 on 1110 degrees of freedom
Multiple R-squared:  0.5118,    Adjusted R-squared:   0.51
F-statistic: 290.9 on 4 and 1110 DF,  p-value: < 2.2e-16
```

Polinomios de educación y gestación

```
Call:
lm(formula = peso ~ poly(educ, 2) + fuma + poly(gest, 2), data = naci)

Residuals:
     Min       1Q   Median       3Q      Max
-1360.10  -298.64   -13.31   288.23  1466.25

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     3261.41      15.36 212.270  < 2e-16 ***
poly(educ, 2)1  1082.66     455.30   2.378   0.0176 *
poly(educ, 2)2   482.85     443.49   1.089   0.2765
fumasi          -173.57      32.14  -5.401 8.09e-08 ***
poly(gest, 2)1 14435.63     448.32  32.200  < 2e-16 ***
poly(gest, 2)2  -774.09     443.40  -1.746   0.0811 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 442.9 on 1109 degrees of freedom
Multiple R-squared:  0.5133,    Adjusted R-squared:  0.5112
F-statistic:   234 on 5 and 1109 DF,  p-value: < 2.2e-16
```

Interacción entre fumadora y gestación

```
Call:
lm(formula = peso ~ educ + fuma + gest:fuma, data = naci)

Residuals:
     Min       1Q   Median       3Q      Max
-1368.94  -297.52   -11.17   289.14  1486.90

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3224.991    268.311 -12.020   <2e-16 ***
educ           14.965      6.510   2.299   0.0217 *
fumasi       -116.477    394.600  -0.295   0.7679
fumano:gest   162.273      6.569  24.705   <2e-16 ***
fumasi:gest   160.771      7.814  20.574   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 443.6 on 1110 degrees of freedom
Multiple R-squared:  0.5115,    Adjusted R-squared:  0.5097
F-statistic: 290.6 on 4 and 1110 DF,  p-value: < 2.2e-16


Call:
lm(formula = peso ~ educ + gest + fuma + gest:fuma, data = naci)

Residuals:
     Min       1Q   Median       3Q      Max
-1368.94  -297.52   -11.17   289.14  1486.90

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3224.991    268.311 -12.020   <2e-16 ***
educ           14.965      6.510   2.299   0.0217 *
gest          162.273      6.569  24.705   <2e-16 ***
fumasi       -116.477    394.600  -0.295   0.7679
gest:fumasi    -1.502     10.206  -0.147   0.8830
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 443.6 on 1110 degrees of freedom
Multiple R-squared:  0.5115,    Adjusted R-squared:  0.5097
F-statistic: 290.6 on 4 and 1110 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = peso ~ educ + gest:fuma, data = naci)

Residuals:
     Min       1Q   Median       3Q      Max
-1366.30  -296.27   -11.13   288.89  1493.33

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3275.020    207.917 -15.752   <2e-16 ***
educ           15.018      6.504   2.309   0.0211 *
gest:fumano   163.532      4.992  32.757   <2e-16 ***
gest:fumasi   159.028      5.115  31.088   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 443.4 on 1111 degrees of freedom
Multiple R-squared:  0.5114,    Adjusted R-squared:  0.5101
F-statistic: 387.7 on 3 and 1111 DF,  p-value: < 2.2e-16
```

Bibliogrfía

- Carmona (2003) - Modelos lineales
- Gallego Gómez - Apuntes de econometría
- Hastie, Tibshirani, et al. (2013) - An introduction to statistical learning with applications in R