**Problema de investigación**

Interesa determinar si la variable peso del recién nacido (peso), medida en gramos, puede ser explicada por las siguientes variables:

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

Para ello, se cuenta con una base de datos de un total de 1115 observaciones.

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.

**Estructuración del trabajo**

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 modelo.

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 homoscedasticidad del modelo y se hará un análisis teórico de no autocorrelación de los errores.

En función de los resultados obtenidos en la etapa anterior, se optará por estimar modelos alternativos o no.

En última instancia, se hará predicción a partir de una partición de la base de datos en muestra de entrenamiento y muestra de prueba, y se calculará el error cuadrático medio resultante como medida de error, comparando entre los distintos modelos especificados si fuera ese el caso.

Asimismo, se obtendrán estimaciones del error de cada modelo usando Cross Validation Leave One Out (LOOCV) y K folds dejando 10 observaciones fuera.

**Análisis descriptivo**

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

Tabla 1: Estadísticas descriptivas



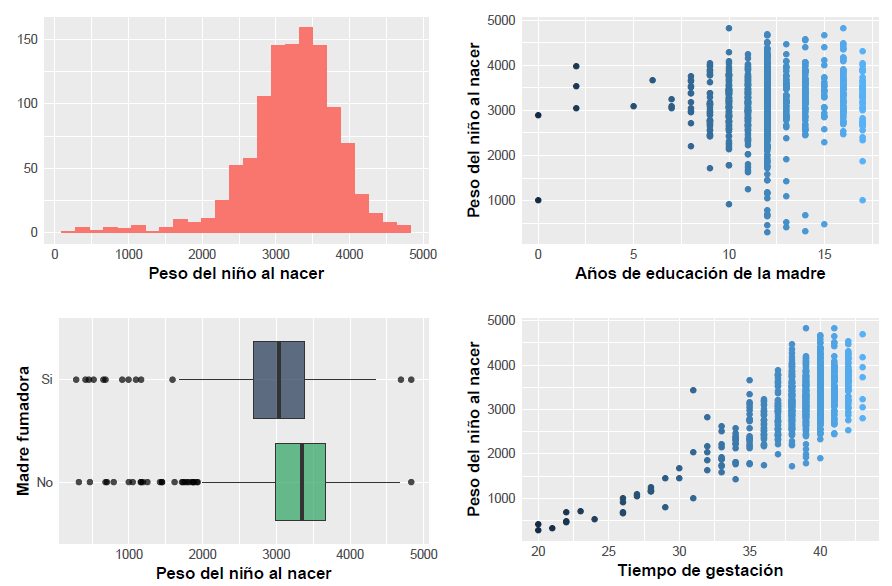
En la tabla 1 interesa notar que

* 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



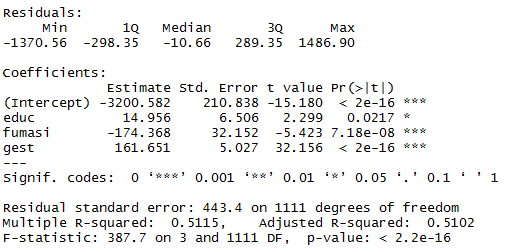


**Especificación del modelo**

Observaciones

1. En lo que sigue del modelo, se tomará en todas las pruebas un nivel de significación
2. Se tomará el valor K como la cantidad de regresores del modelo
3. Se tomará el valor n para simbolizar la cantidad de observaciones, que en nuestro caso son 1115.

**Modelo 1**: Se estima el siguiente modelo



**Significación individual**

**Estadístico de la prueba y distribución bajo H0**

**Región critica de la prueba**

Según 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 parámetros, a un nivel .

**Significación global del modelo**

**Estadístico de la prueba y distribución bajo H0**

Con

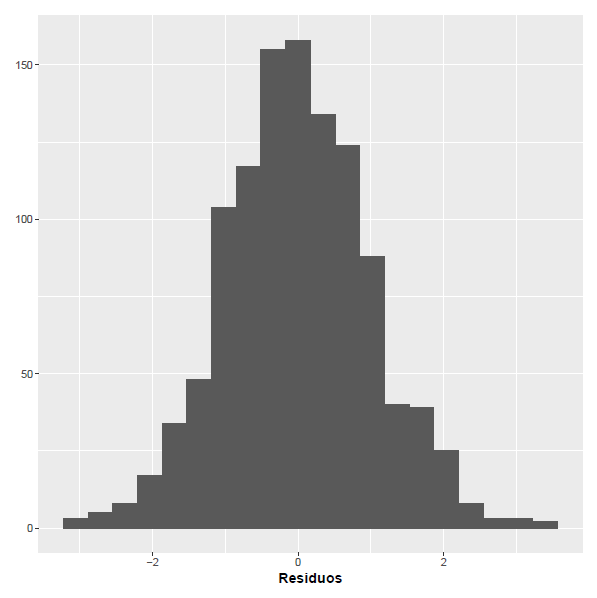
**Región crítica de la prueba**

Según el criterio del p-valor, a un nivel , la evidencia empírica sugiere que el modelo es globalmente significativo.

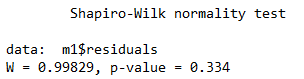
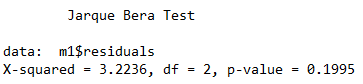
**Diagnóstico del modelo 1**

**Normalidad de los residuos**

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

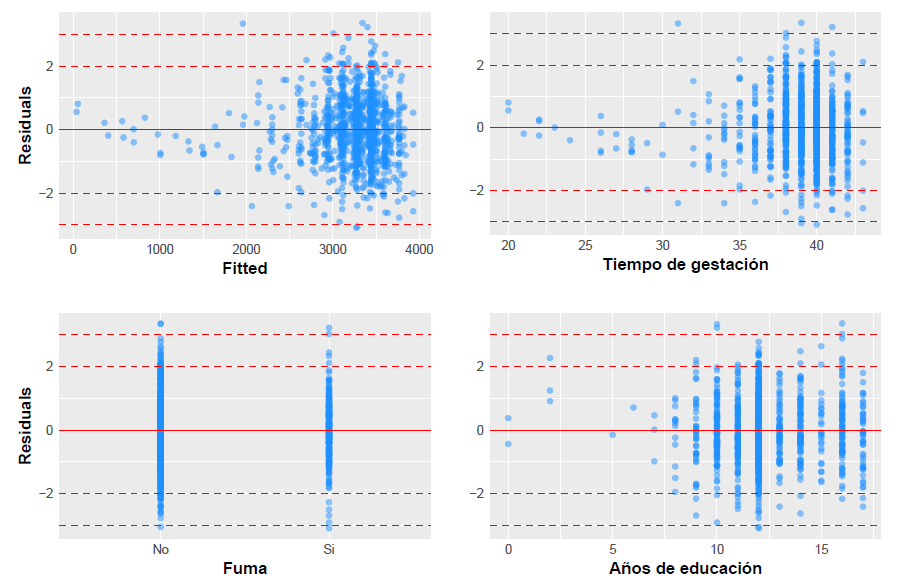


Por su parte, los test de normalidad Shapiro-Wilk y Jarque-Bera, según el criterio del p-valor y para un nivel de significación , ambos no rechazan la hipótesis nula de normalidad de los residuos. A continuación, las salidas de R correspondientes:

**Homoscedasticidad**

En una primera instancia, se opta 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 homoscedasticidad en el modelo.

Por lo tanto, se propone acudir a un contraste de heteroscedasticidad. El contraste de Gold-Feld y Quandt no resulta apropiado dado que se en tiende 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 outliers.

En virtud de ello, se entiende que una posible alternativa es realizar el contraste propuesto por White.

**Prueba de heteroscedasticidad de White**

**Metodología**

Dado el modelo

Se especifica la regresión auxiliar

Es decir, la regresión de los residuos al cuadrado del modelo original sobre las variables del modelo, los cuadrados de las variables y los productos cruzados. Obsérvese que se omitió el cuadrado de la variable por ser una variable dummy lo cual introduciría multicolinealidad.

Se estima la regresión auxiliar por el método MCO, y se obtiene el coeficiente de determinación .

**Estadístico de la prueba y su distribución asintótica bajo H0**

**Región crítica de la prueba**

**Resultados obtenidos**

A partir de la implementación de la prueba, con un , no se rechaza la hipótesis nula de homoscedasticidad para el nivel de significación especificado.

**No autocorrelación**

En lo que sigue del trabajo y para todos los modelos a especificar, vamos a suponer la existencia de no autocorrelación en los residuos. El argumento es teórico y está dado ya que no sería razonable suponer que el peso y los determinantes del mismo en una de las observaciones afectan el de otras.

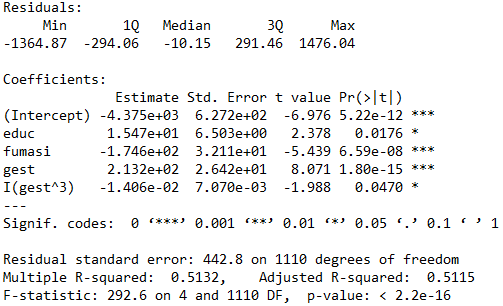
No obstante lo anterior, interesa señalar que como resultado de las estadísticas descriptivas, se pudo observar una especia de curvatura en el gráfico de la variable regresora gest contra la variable explicada peso, lo cual nos podría llevar a suponer algún tipo de relación polinómica entre ellas. De igual forma, a partir del gráfico de la variable regresora educ contra la variable explicada peso, no queda del todo claro la existencia de una relación lineal entre ellas y no polinómica.

Dado lo anterior, se procede a estimar modelos con cuadrados y cubos de ambas variables. Los modelos especificados con el cuadrado y el cuadrado y cubo de la variable educ, ambos fueron descartados dado que ninguna de ellas resultó significativa al nivel seleccionado.

Por otra parte, si bien el modelo especificado con el cuadrado de la variable gest resultó en una no significación de la misma, del modelo con el cuadrado y el cubo resultó en que el cuadrado de gest fue no significativa pero el cubo sí. Por lo tanto, se procedió a estimar un modelo con todas las variables regresoras más el cubo de la variable gest:

Modelo 2

La salida de R correspondiente

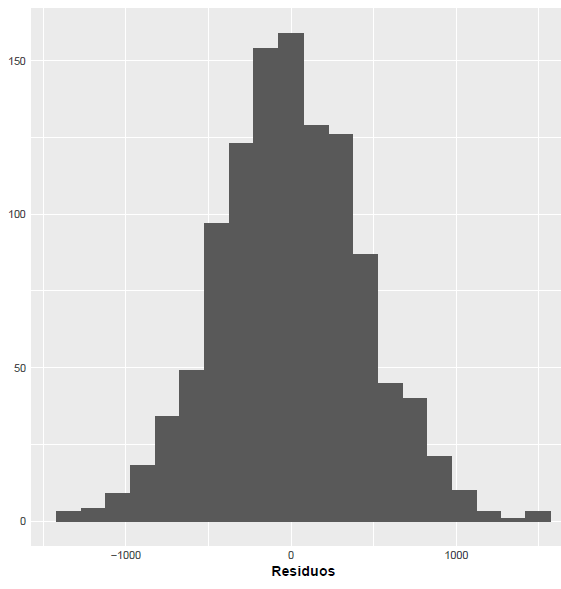


Como se puede observar en la salida, el cubo de la variable gest continúa siendo significativa por lo que, por un criterio de parsimonia, se opta por continuar el análisis con el modelo 2.

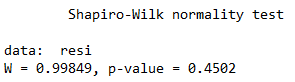
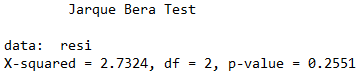
**Diagnóstico del modelo 2**

**Normalidad de los residuos**

Por su parte, el histograma de los residuos estandarizados sugiere que podría llegar a ser razonable suponer distribución normal de los mismos.

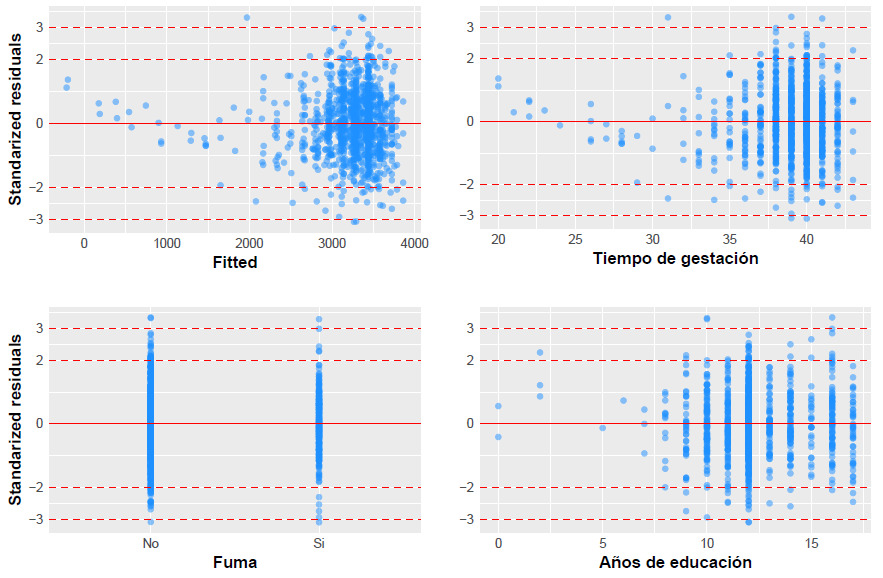


Por su parte, los test de normalidad Shapiro-Wilk y Jarque-Bera, según el criterio del p-valor y para un nivel de significación , ambos no rechazan la hipótesis nula de normalidad de los residuos. A continuación, las salidas de R correspondientes:

**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.



Se especifica la regresión auxiliar

Y se estima por MCO. El coeficiente de determinación de la regresión auxiliar no supera el valor de tabla de una distribución que acumula probabilidad 0.95 por lo que no se rechaza la hipótesis nula de homoscedasticidad.

**No autocorrelación**

El comentario es el mismo que para el modelo 1.

**Validación Cruzada**

Se implementó validación cruzada “leave one out” y dejando diez observaciones fuera, de lo cual se obtuvieron los siguientes resultados:

**Predicción**

Finalmente, se comparan los modelos seleccionados en función de su performance en materia de predicción. La medida de error seleccionada es el Error Cuadrático Medio, el cuál se define de la siguiente manera:

**Test de igualdad de medias**

**Estadístico de prueba**

**Región crítica de la prueba**

Ahora bien, resulta interesante testear la hipótesis de igualdad de medias por grupos conformados según la variable fuma.

Para la variable gest, aplicando una prueba T, resulta que la resta de las medias de los grupos formados por la variable fuma es significativamente distinta de cero, con lo cual se rechaza la hipótesis nula de igualdad de medias.

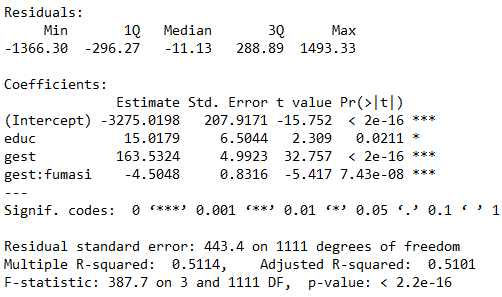
Lo mismo ocurre para la variable explicada peso.

Lo anterior es una justificación de agregar un tercer modelo, que contemple las interacciones entre la variable fuma y las variables gest y fuma; y gest y peso.

De las alternativas especificadas, se selecciona un tercer modelo que incluye la interacción entre la variable gest y la variable fuma.

**Modelo 3**

En la salida de R correspondiente indica que todas las variables incluidas en el modelo resultan individualmente significativas, y el modelo resulta también globalmente significativo.



No obstante, dado que el error cuadrático medio obtenido con las técnicas de LOOVC y K Fold CV no presenta una caída significativa, siendo incluso ambos mayores que en el modelo 1, el modelo 3 es descartado.