Ejercicio: Peso de recién nacidos

Alfonso Novales
Departamento de Economía Cuantitativa
Universidad Complutense*

10 de mayo de 2004

1.

*Versión muy preliminar. No citar sin permiso del autor. Las sucesivas versiones de este trabajo irán estando disponibles en http://www.ucm.es/info/ecocuan/anc, o pueden solicitarse en anovales@ccee.ucm.es

Peso de recién nacidos[1]

Analizamos en este ejemplo datos tomados de Wooldridge, Introducción a la Econometría: un enfoque moderno, primera edición. Partiendo de un modelo de regresión estimado en dicho texto,

- discutimos el modo de llevar a cabo un análisis descriptivo, tanto de tipo gráfico como de tipo estadístico, acerca de la capacidad explicativa que un conjunto de variables tiene sobre una determinada variable dependiente, y
- describimos cómo el habitual uso mecánico de los estadísticos tipo t de Student y F puede conducir a conclusiones erróneas sobre la capacidad explicativa de una variable o de un conjunto de ellas.

1.1. Descripción del ejemplo

Consideramos en este ejemplo la especificación de un modelo de regresión para tratar de caracterizar factores que pueden afectar al peso de bebés al nacer. La base de datos[2], tomada de Wooldridge (2001), contiene información sobre el peso de los bebés, recogido en 1.388 nacimientos, la renta de la familia en la que se produce el nacimiento ($renta_i$), el número medio de cigarrillos fumados diariamente por la madre durante el embarazo ($cigarrillos_i$), el número medio de cajetillas de tabaco fumados diariamente por la madre durante el embarazo, el número de orden que ocupa el recién nacido dentro de los hijos de la familia ($ordenac_i$), los años de educación del padre ($educp_i$) y de la madre ($educm_i$), el sexo del bebé y si éste es blanco o de otra raza. Estas dos últimas variables son ficticias, y aparecen en el archivo como variables dicotómicas, es decir, tomando dos valores únicamente. La variable sexo ha sido definida mediante $Sexo = 1$ si el recién nacido es varón, $Sexo = 0$ en caso contrario, mientras que la variable que recoge el grupo étnico se ha definido $Raza = 1$ si el bebé es de raza blanza, $Raza = 0$ en caso contrario. Falta información acerca del nivel educativo del padre del recié nacido en 196 nacimientos, faltando información acerca del nivel educativo de la madre en un caso más, por lo que las regresiones que incluyen estas variables como explicativas utilizan un máximo de 1191 observaciones.

En Wooldridge (2001) se estima el modelo de regresión,

[1] Fichero de trabajo de EVIEWS: Bwght.wf1

[2] El archivo Bwght.des contiene la descripción de las variables incluidas en el archivo Bwght.raw, algunas de las cuales se han utilizado en el ejemplo.

$$
\begin{aligned}
Peso_i &= \underset{(3,73)}{114,52} - \underset{(0,110)}{0,596}\, cigarillos_i + \underset{(0,037)}{0,056}\, renta_i + \underset{(0,659)}{1,788}\, ordenac_i + \underset{(0,283)}{0,472}\, educp_i - \underset{(0,320)}{0,370}\, e \\
\bar{R}^2 &= 0,035,\ \hat{\sigma}_u = 19,789
\end{aligned}
$$

donde se muestran entre paréntesis los estadísticos tipo t. El autor contrasta la significación conjunta de los niveles educativos de ambos padres mediante el estadístico F, en la forma del R^2, no rechazando la hipótesis nula de ausencia de capacidad explicativa de ambas variables, conjuntamente consideradas. Por tanto, el nivel educativo de los padres no parece ser un condicionante significativo del peso de los bebés al nacer.

La discusión que llevamos a cabo en la Sección XX ya sugiere que, en cualquier caso, la interpretación de este resultado no debe hacerse con carácter absoluto. El investigador debería decir que "una vez incluidas considerados como posibles factores explicativas del peso del recién nacido la renta de la familia, el número de cigarrillos fumados por la madre durante el embarazo y el número de orden del recién nacido entre sus hermanos, los indicadores educativos de los padres no aportan *información adicional* relevante".

El segundo matiz que hemos de hacer es que podría darse el caso de que los indicadores educativos contengan información relevante incluso una vez que ya se ha tenido en cuenta la información aportada por las variables mencionadas pero, por alguna razón, la información muestral disponible no permite medir con suficiente precisión el impacto que sobre el peso del bebé tiene el nivel educativo de los padres. Una reducida precisión podría conducir a un estadístico-t reducido y, con ello, a no rechazar la hipótesis nula de ausencia de relación entre nivel educativo de los padres y peso del bebé.

La tercera consideración a efectuar es que el contraste tipo F efectuado para analizar conjuntamente la información proporcionada por las dos variables educativas descansa sobre el supuesto de Normalidad del término de error del modelo de regresión, cuestión que habríamos de analizar.

Comenzamos nuestro análisis indagando la información que cada una de las potenciales variables explicativas contiene sobre el peso del recién nacido. Al hacerlo individualmente, estamos ignorando el hecho de que distintas variables pueden contener información común; debemos interpretar que se trata de un análisis que trata de detectar la ausencia de capacidad explicativa en alguna variable. Si, como es habitual, nos limitásemos al análisis de los estadísticos tipo t, diríamos que, entre las variables consideradas, el número de cigarrillos fumados por la madre

afecta al peso del recién nacido, habiendo asimismo un efecto estadísticamente significativo en relación con el número de orden que el recién nacido ocupa entre los hijos de la familia. Los indicadores de educación no parecen aportar información relevante, al igual que tampoco parece haber relación con la renta de la familia en la que se produce el nacimiento.

1.2. Características muestrales de las variables (archivo bwght.wf1)

Los histogramas[3] de las variables revelan características interesantes (ver *HIS_nombre variable en el fichero* bwght.wf1): la variable dependiente *peso* es una variable continua, cuyo exceso de curtosis genera un comportamiento no Normal en la muestra, rechazándose claramente dicha hipótesis mediante el test de Jarque-Bera. Este resultado despierta dudas acerca del uso de las distribuciones habituales tipo t de Student y F de Fisher-Sendecor para los estadísticos utilizados en la contrastación de hipótesis.

Las variables *cigarrillos* y *paquetes* tienen una correlación exactamente igual a 1,0. Esto significa que se han construido una a partir de la otra, pues si se hubiera encuestado sobre ambas existiría una relación algo menos que perfecta entre ellas. Examinando sus valores, vemos que la primera es igual a 20 veces el valor numérico de la segunda en todos los casos, por lo que utilizaremos únicamente la variable *cigarrillos*. Esta es una variable discreta, con un valor mínimo de 0 y un valor máximo de 50; la mediana es 0, reflejando el hecho de que en casi un 85% de los 1.388 nacimientos recogidos en la muestra, la madre declaró no haber fumado durante el embarazo[4]. Sólo en 212 casos, la madre del recién nacido declaró haber fumado un número medio de cigarrillos por día mayor que cero. Esto sugiere que disponemos de una información relativamente reducida para estimar la contribución al peso del bebé de un cigarrillo adicional, lo que podría hacer que dicha estimación se obtenga con una precisión no muy alta, salvo si la diferencia entre el peso de los bebés de madres fumadoras y no fumadoras es muy sistemática.

La educación de la madre toma valores entre 2 y 18 años, con una mediana de 12 años; ésta es también la moda, recogiendo el 40,5% de las observaciones muestrales. La educación del padre toma valores entre 1 y 18 años, también con una mediana y moda igual a 12 años; valor que aparece en un 37,2% de los nacimientos. El elevado número de observaciones en el nivel educativo correspondiente a

[3]Los nombres en cursivas, entre paréntesis, denotan elementos del archivo de trabajo Bwght.wf1.

[4]Por tanto, la moda de esta variable es cero.

12 años segmenta la muestra de padres y madres entre los que alcanzan el grado medio y los que continúan con estudios superiores.

La información numérica sobre la renta familiar, en miles de dólares, tiene el aspecto de haber sido redondeada, apareciendo únicamente valores numéricos entre 0,5 y 19,5, además de 22,5, 27,5, 32,5, 37,5, 42,5, 47,5, 65,0. Por tanto, la variable renta tiene naturaleza discreta, tomando un número relativamente alto de valores igualmente espaciados en el primer rango mencionado, para pasar a tomar valores más dispersos posteriormente. Un 38% de las observaciones están en el rango $(0,5; 19,5)$ de renta, estando el 62% restante en niveles de renta superiores, por lo que el proceso de redondeo afecta a un alto número de observaciones. Si hubiera una relación continua entre la renta de la familia y el peso del recién nacido, tal proceso de simplificación numérica podría dificultar notablemente su estimación. Aunque ignoramos el modo en que la concentración de valores numéricos ha sido hecha, imaginemos que se ha asignado un dato de renta de 65,0 a las familias con renta en $(56,75, 65,0)$, asignando renta de 47,5 a aquellas familias con renta en $(47,5; 56,75)$. El peso del recén nacido podría crecer suavemente con la renta, pero ésta se ha colapsado en los dos extremos del intervalo, generando una importante cantidad de errores en cualquier relación lineal entre peso y renta. Por tanto, tenderíamos apensar que dicha relación no existe.

La variable $ordenaci_i$, que recoge el orden del recién nacido entre los hijos de la familia, toma valores entre 1 y 6, siendo la moda igual a 1, con una frecuencia relativa de 57,3%. Por tanto, la mediana de esta variable es asimismo igual a 1.

El 48% de los recién nacidos (665) son mujeres y el 52% (723) varones, por lo que la muestra está bastante equilibrada en este sentido; por el contrario, el 78% son de raza blanca y el 22% restante de otras razas. Los posibles efectos del sexo y la raza del recién nacido sobre su peso no han sido considerados en la regresión anterior, pero los consideraremos más adelante. Es asimismo interesante observar que de las 212 madres que declararon haber fumado durante el embarazo, 165 eran de raza blanca, mientras que de las 1089 madres que declararon no haber fumado durante el embarazo, 924 eran de raza blanca.[5] Como se muestra en Bwght.xls, haber fumado durante el embarazo es independiente de la raza de la madre.

[5]Esto se muestra en Bwght.xls, multiplicando las columnas de variables dicotómicas {0,1} "Fuma" y "Blanco", y hallando la suma de dicho producto, y repitiendo el cálculo con "Blanco" y 1-"Fuma". Suponemos aquí que la raza de la madre y del recién nacido son las mismas. De modo análogo, puede verse que de las 212 madres que declararon haber fumado, 100 tuvieron un hijo varón. Esta división aproximada entre hijos varones y mujeres es, por supuesto, muy razonable.

1.3. Asociación con la variable dependiente, peso del recién nacido.

Los coeficientes de correlación habituales son reducidos (Tabla *correlaciones*), siendo el más elevado numéricamente (-0,16) el del número de cigarrillos fumados, que es de signo negativo, como esperaríamos. Recuérdese que una desviación típica aproximada del coeficiente de correlación es el inverso de la raíz cuadrada del tamaño muestral, que estaría en torno a 0,027. Ello haría que la correlación mencionada, aun siendo reducida, fuese estadísticamente significativa. Sin embargo, el resto de las correlaciones recogidas en la tabla sugiere que la búsqueda de capacidad explicativa del peso del recién nacido en las variables disponibles puede resultar poco fructífera. Entre las variables explicativas, la renta de la familia tiene coeficientes de correlación superiores a 0,40 con los niveles educativos del padre y la madre que, a su vez, muestran una correlación de 0,64 entre ellos.

Sin embargo, las variables explicativas tienen naturaleza discreta, por lo que los coeficientes de correlación habituales no están plenamente justificados. Esto mismo hace que las nubes de puntos con la variable dependiente no sean tan informativas como en otros casos; como muestra, recogemos en el fichero de trabajo la nube de puntos entre el peso y el orden que el recién nacido ocupa entre los hijos. Un efecto negativo, por ejemplo, vendría dado por una reducción del peso al aumentar el valor de la variable $ordenac_i$. La nube de puntos nos da un intervalo de pesos observados entre los recién nacidos que comparten un mismo valor de la variable $ordenac_i$, y se trataría de ver si el valor representativo de cada intervalo de pesos es decreciente al aumentar $ordenac_i$.

Esto nos dirige a estimar la asociación entre variables mediante tablas de clasificación de sus valores, así como contrastando la igualdad de medias y medianas entre clases. Por ejemplo, para analizar la posible asociación entre el peso del bebé y la educación de la madre, calculamos la mediana del peso de los bebés para cada uno de los posibles niveles educativos de la madre, contrastando la igualdad de dichos valores mediana. Si estas dos variables no estuvieran relacionadas, las medidas de posición central (mediana o media) de la variable *peso* serían similares para los distintos niveles educativos; si existe una asociación positiva entre ambas variables, esperaríamos que la media o mediana de *peso* fuese creciente con el nivel educativo, y lo contrario ocurriría si existiera una relación negativa entre ambas. En ambos casos se rechazaría la hipótesis nula de igualdad de medias así como la de igualdad de medianas. Para ello, debe calcularse la media o mediana de la variable dependiente para cada uno de los distintos rangos de valores numéricos de la variable explicativa que se considera. Nos centramos en las medianas y no en las medias debido a la fuerte desviación que muestran las distribuciones de

estas variables respecto de la Normalidad, tanto por razón de la muy elevada frecuencia observada en el valor modal, como de su asimetría. El lector interesado puede reproducir nuestro análisis contrastando la igualdad de medias muestrales del peso para los distintos niveles educativos de la madre o el padre.

Al comparar las variables *peso* y *educm*, los contrastes Kruskal-Wallis y van der Waerden de igualdad de medianas rechazan la igualdad de medianas, sugiriendo asociación entre ambas variables ($MEDN\_PESO\_EDUCM$). Repetimos el contraste llevando a cabo cierta agrupación de los niveles educativos, para eliminar el problema de que algunos niveles educativos recogen un número muy reducido de observaciones: para algunos niveles educativos hay una sóla observación muestral. La agrupación proporciona indicios aún más claros en contra de la igualdad de medianas. Los valores numéricos de las medianas por clases de niveles educativos[6] después de la agregación, recogidas en ($MEDN\_PESO\_EDUCM2$) *sugiere cierta asociación positiva* entre ambas variables, puesto que la mediana del peso parece ser creciente con el nivel educativo de la madre. Así lo sugieren asimismo los valores p de los contrastes de la chi-cuadrado, de Kruskal-Wallis y de van der Waerden que aparecen en la tabla. Tal asociación podría reflejarse en un gráfico de barras que mostrase los pesos medianas que aparecen debajo del rótulo *Category Statistics* en la tabla $MEDN\_PESO\_EDUCM2$ como función de los valores centrales de los intervalos que aparecen para la variable $educm_i$. Sin embargo, tal como muestra el gráfico de barras de *Med_peso_educm2*, la asociación, si existe, es débil.

También en la relación con el nivel educativo del padre, hemos efectuado dos veces el contraste de igualdad de medianas: una, sin agrupar los niveles educativos ($MEDN\_PESO\_EDUCP$), y otra, agrupándolos ($MEDN\_PESO\_EDUCP2$); la segunda es preferible, a pesar de que el nivel de agrupación es relativamente arbitrario. En casos como los que estamos analizando, 15 clases parece un número razonable, pues permite que aflore cierta disparidad entre medianas, a la vez que

[6]Para obtener una clasificación de la variable *Peso* utilizando como clasificador los niveles educativos de la madre, seleccionar *Peso* y entrar en *"Descriptive Statistics/Stats by Classification"* escribiendo EDUCM en la ventana *"Series/Group for Classify"*. Para contrastar la igualdad de medianas entre grupos a la vez que se lleva a cabo la clasificación, entrar en *"Tests for Descriptive Statistics/ Equality Tests by Classification"*, escribiendo EDUCM en *"Series/Group for Classify"*, y marcando *"Mediana"*, en vez de *"Media"* bajo *"Test Equality of"*. Para obtener una clasificación con agrupación de niveles educativos, a la derecha, donde aparece *"Group into Bins if "* marcar un número reducido (por ej., 10) en la ventana *"# of values"*, que se refiere al número de rangos de valores que se quieren utilizar para la variable que se utiliza como clasificador, en este caso, EDUCM.

permite recoger una mínima frecuencia dentro de cada clase. Si juzgamos por los valores p de los contrastes, la evidencia contraria a la hipótesis nula de igualdad de medianas, lo que sugeriría una posible asociación entre las variables *peso* y *educp*, es claramente menor que en el caso del nivel educativo de la madre, sugiriendo que el nivel educativo del padre podría no ser muy relevante para explicar el peso del bebé. Sin embargo, no hemos de olvidar que estamos comparando únicamente una medida de posición central de la variable *peso* para los distintos grupos definidos para *educm* o *educp;* no examinamos el conjunto de todos los valores de *peso* observados dentro de cada nivel educativo, lo que podría arrojar ciertas diferencias entre distintos niveles de $educm_i$. Por ejemplo, podríamos observar que los rangos observados para $peso_i$ se amplían o se estrechan al aumentar $educm_i$, sugiriendo que la varianza de la variable $peso_i$ es función del nivel educativo de la madre. Una evolución creciente de los pesos mínimo y máximo sugeriría asimismo una relación positiva, siendo negativa si si se observase la evolución contraria; esto podría ocurrir sin observar variaciones significativas en los valores mediana.

La evidencia a favor de asociación es bastante más clara en la comparación de *peso* y *renta* ($MEDN\_PESO\_RENTA$), y todavía más clara en el caso de *peso* y *cigarrillos* ($MEDN\_PESO\_CIGS2$). Un diagrama de barras de las medianas de *peso* por clases de *renta* sugiere una asociación positiva (MED_PESO_RENTA), mientras que un diagrama de medianas de *peso* por clases de valores de *cigarrillos* sugiere una asociación negativa (MED_PESO_CIGS2), si bien esta última clasificación está contaminada por el elevado porcentaje muestral con un valor cero de la variable *cigarrillos*. En el fichero de trabajo se incluye asimismo la variable $FUMA$, que hemos definido de modo que el valor 0 si la madre no fumó durante el embarazo, y el valor 1 si lo hizo. El valor mediana de los pesos de los bebés fue de 111 y 120 onzas, respectivamente, en cada caso, lo que sugiere cierta dependencia negativa entre el peso y el hábito de fumar. Los valores p de los contrastes en MED_PESO_FUMA son bastante concluyentes respecto a la existencia de tal dependencia.

La igualdad de medianas no se rechaza cuando se clasifica la variable *peso* de acuerdo con los valores de la variable *ordenac,* sugiriendo que el orden del recién nacido entre sus hermanos podría no ser información relevante para explicar su peso. Este análisis descriptivo es preliminar, habiendo relacionado, alternativamente, cada una de las variables explicativas, con la variable dependiente. No hemos considerado, por tanto, la posible colinealidad entre variables explicativas, es decir, que éstas puedan proporcionar información común. A título preliminar, podríamos concluir con una ordenación de variables por niveles de capacidad ex-

plicativa, comenzando con el número de cigarrillos y la renta familiar, junto con una posible dependencia débil respecto del nivel educativo de la madre, mientras que el orden del recién nacido dentro de los hijos de la familia parece no aportar información relevante acerca de su peso. Esta evidencia es coherente con la obtenida en la regresión mostrada al inicio en lo relativo al efecto del número de cigarrillos fumados, pero no en cuanto a los posibles efectos de las variables $renta_i, ordenac_i$, o $educm_i$.

1.4. Análisis de regresión

Nuevamente hay que hacer notar que aunque esta sección debería comenzar presentando las nubes de puntos de las variables de la regresión pero, debido a la naturaleza de las variables explicativas, no lo hacemos. Si lo desea, el lector puede utilizar el fichero de trabajo para construir dichos gráficos. Estimamos regresiones individuales sobre las dos variables aparentemente más relevantes, *cigarrillos* y *renta*, obteniendo,

$$\begin{aligned} Peso_i &= \underset{\substack{(0{,}57)\\(209.3)}}{119.77} - \underset{\substack{(0{,}090)\\(-5{,}68)}}{0,514}\, cigarrillos_i + \hat{u}_i, && (1.1)\\ \bar{R}^2 &= 0,022,\ \hat{\sigma}_u = 20,13,\ Ratio = 0,011 && (1.2)\\ Peso_i &= \underset{\substack{(1{,}00)\\(115.0)}}{115,27} + \underset{\substack{(0{,}029)\\(4{,}08)}}{0,118}\, renta_i + \hat{u}_i, && (1.3)\\ \bar{R}^2 &= 0,011,\ \hat{\sigma}_u = 20,24,\ Ratio = 0,005 && (1.4) \end{aligned}$$

donde *Ratio* denota el cociente entre la desviación típica muestral de los residuos, y la de la variable peso, que es de 20,35.

Estos modelos de regresión simple puedan estar incorrectamente especificados por omitir algún efecto significativo. Si así fuese, el coeficiente estimado (la pendiente del modelo de regresión) en la primera estaría sesgado, en el sentido de no medir el efecto que sobre el peso tiene la única variable explicativa incluída en la regresión, *cigarrillos*; la estimación de dicho coeficiente estaría recogiendo asimismo los efectos de variables omitidas que no sean independientes de la variable incluida, por ejemplo, la renta de la familia, o la ordenación del recién nacido entre sus hermanos. Sabemos algo más: de acuerdo con la discusión teórica relativa al sesgo por variables omitidas, al omitir una variable explicativa negativamente correlacionada con *cigarrillos*, el coeficiente de ésta se subestimará,sobreestimándose

si la variable omitida tiene correlación positiva con *cigarrillos* pues, en ambos casos, asignaríamos a *cigarrillos* el efecto combinado de ambas variables. Esto es precisamente lo que diría nuestra intuición.

El primer paréntesis debajo de cada coeficiente estimado contiene la desviación típica de la estimación, mientras que el segundo contiene el estadístico tipo-t, cociente entre la estimación y su desviación típica. En muestras amplias de sección cruzada es habitual obtener un valor numérico muy reducido para el coeficiente de determinación, si bien desearíamos que fuese algo mayor del obtenido en estas regresiones individuales. En todo caso, los niveles obtenidos del $\bar{R}^2$ en absoluto indican ausencia de relación.

Este es un caso en el que el uso habitual de los estadísticos tipo-t sugeriría que ambas variables tienen capacidad explicativa relevante, siendo *estadísticamente significativas*; de acuerdo con tal criterio, nadie dudaría en incluirlas en un modelo de regresión. Sin embargo, las desviaciones típicas residuales, y los *Ratios* indican que la capacidad explicativa de cada una de estas variables por separado es, verdaderamente, muy reducida. El coeficiente estimado para *cigarrillos*, implica que, para el valor mediana de los cigarillos fumados durante el embarazo (cuando no son cero), que es de 10, la diferencia en peso de bebés de madres fumadoras y madres no fumadoras sería de 5 onzas, menor que la diferencia observada en la muestra, de 112 a 121 onzas, a que antes nos referimos.

Evidencia adicional acerca de la reducida información que *cigarrillos* y *renta* proporcionan sobre *peso* aparece en FIG_RES_CIGS y FIG_RES_RENTA, que representan los valores ajustados y los residuos de ambas regresiones. Este es un tipo de gráficos que siempre hemos de examinar, tras estimar un modelo de regresión. Estos gráficos son la evidencia más clara acerca de la reducidísima capacidad explicativa de las dos variables, ya que la mayor parte de la fluctuación en peso de unos bebés a otros permanece en los esiduos, npo habiendo sido explicada por las variables utilizadas como explicativas en la regresión.

Indicios adicionales acerca de la baja capacidad explicativa aparecen en $CORR\_PESO\_AJU$ que muestra coeficientes de correlación entre *peso* y los residuos de las dos regresiones, así como de la regresión que incluye ambas variables, *cigarrillos* y *renta*, como variables explicativas, y de otras regresiones que analizaremos posteriormente. Las variables mencionadas son las que han sido incluidas como explicativas en cada regresión. Todas las correlaciones son muy elevadas, lo que significa que la parte de la variable *Peso* que queda sin explicar por las variables *renta* y *cigarrillos* es muy similar a la propia variable *Peso*, es decir, que las regresiones apenas explican las diferencias en *peso* entre bebés. Es interesante que la correlación sea

algo menor cuando se utilizan ambas variables, lo que sugiere que la información que contienen no es exactamente común, si bien es reducida en ambos casos.

Correlaciones tan elevadas pueden interpretarse asimismo en el sentido de que, si utilizásemos las regresiones estimadas para predecir el peso de un recién nacido utilizando las variables *cigarrillos* y *renta* como predictores, la correlación entre la previsión resultante y el peso observado del bebé sería muy pequeña o, lo que es equivalente, la calidad de la predicción sería muy baja. Por ejemplo, para el nivel mediana de renta, 27,5, el modelo (1.3) predice un peso de 118,52 onzas. En la muestra se observa[7], para dicho nivel de renta, un rango de pesos entre 80 y 167 onzas; demasiada dispersión para poder prever con precisión, lo que explica el bajo ajuste del modelo.

El lector puede proceder a continuación a comprobar que ninguna de las dos variables de educación aporta capacidad explicativa adicional. Para ello, puede utilizar cada una de ellas por sí sólas como variable explicativa en una regresión y constatar: *a*) el reducido valor del R^2, y del *Ratio* de ajuste, así como *b*) la elevada correlación entre los residuos resultantes y la variable *Peso* original, tal como hemos hecho con *cigarrillos* y *renta; c)* alternativamente, puede añadir uno de los niveles educativos como variable explicativa a una cualquiera de las regresiones anteriores y comprobar que el coeficiente de correlación entre los residuos de ambas regresiones es muy elevado; algunas de estas correlaciones aparecen en $CORR\_PESO\_AJUSTE$, significando que la inclusión del nivel educativo de los padres no ha añadido capacidad explicativa significativa a las variables *cigarrillos* y *renta*. Las nubes de puntos $COMP\_RES\_1$ a $COMP\_RES\_4$ aportan una evidencia similar: $COMP\_RES\_1$ y $COMP\_RES\_2$ relacionan la variable dependiente, *Peso*, con los residuos de dos regresiones simples, la primera utilizando *cigarrillos* como unica variable explicativa, mientras la segunda utiliza *Renta* en su lugar. $COMP\_RES\_3$ compara los residuos de la regresión simple sobre *cigarrillos*, con los que se obtienen de una regresión que incluye *cigarrillos* y *renta* como variables explicativas; el hecho de que ambos residuos sean tan similares sugiere que la variable *renta* no aporta mucho a la variable *cigarrillos* para explicar el peso del recién nacido.[8] $COMP\_RES\_4$ compara los residuos de la regresión simple sobre *cigarrillos*, con los que se obtienen de una regresión que incluye *cigarrillos,*

[7] Ver Bwght.xls

[8] Nótese que esto no significa en modo alguno que cigarrillos tenga más o menos capacidad explicativa, sino tan sólo que la información proporcionada por la renta familiar no aporta nada a la contenida en el número de cigarillos fumado porla madre durante el embarazo, que podría ser relevante o no serlo.

renta, los niveles educativos del padre y de la madre, y el número de orden del recién nacido entre sus hermanos como variables explicativas; la interpretación es similar, y no parece que el resto de las variables aporte mucha información a la que pueda incorporar el número de cigarrillos.

Para profundizar en la información proporcionada por los niveles educativos, y dada la excesiva concentración de cada una de estas dos variables en el nivel 12 años, definimos una variable ficticia en el caso de las mujeres, *edm*, que es igual a 0 si *educm* es inferior a 12 años, es igual a 1 si *educm* es igual a 12 años, y es igual a 2 si *educm* toma cualquier valor numérico superior a 12 años. En ocasiones, es difícil medir con precisión el efecto de cambios unitarios en una variable como *educm*, pero se mide mejor el efecto que tiene sobre la variable dependiente el paso de un nivel de *educm* a otro. Aunque no incidimos aquí en los resultados, la variable ficticia así construida, que se incluye en el archivo de trabajo, no parece aportar capacidad explicativa significativa. Finalmente, concluimos que los niveles educativos no son relevantes para explicar el peso de los recién nacidos, una vez que se tiene en cuenta la información proporcionada por *cigarrillos*. Algo similar puede decirse del nivel educativo del padre.

Cuando se considera la variable *ordenac*, la escasa contribución informativa es aún más evidente, como ya sugería el análisis descriptivo que antes hicimos, por lo que concluimos que esta variable no aporta información relevante a la ya proporcionada por las variables *renta* y *cigarrillos*. Esto ocurre a pesar de que esta variable aparece con un valor numérico del estadístico t de Student superior a 2 en la regresión que incluye todas las variables explicativas [REG_TODAS], propiedad que se mantiene si excluimos de dicha regresión todas las variables explicativas con estadístico t inferior a 2 en valor absoluto, y volvemos a estimar el modelo. Si siguiéramos este procedimiento, habitual en el análisis empírico, pero en absoluto recomendable, nos quedaríamos con una regresión que utiliza *cigarrillos* y *ordenac* como únicas variables explicativas [$REG\_CIGS\_ORDENAC$]. Sin embargo, la correlación entre los residuos de esta regresión [RES_CIGS_ORDEN] y los que utiliza únicamente la variable *cigarrillos* como explicativa [RES_CIGS] es superior a 0,997, indicando que *ordenac* apenas añade información a la que pueda incluir la variable *cigarrillos*.

Finalmente, si el investigador decidiera utilizar todas las variables simultáneamente, como hicimos en la regresión mostrada en primer lugar, obtendría unos residuos muy altamente correlacionados con los de las regresiones previas, así como con la variable *Peso* original [ver la correlación entre RES_TODAS (el residuo de la regresión con todas las variables explicativas) y RES_CIGS, RES_RENTA, en

la tabla *CORR_PESO_AJUSTE*, así como la nube de puntos *COMP_RES_4*]. Nuevamente, la interpretación es la misma, en términos de la reducida capacidad explicativa del conjunto de variables considerado, como perfectamente ilustra *FIG_RES_TODAS*.

En general, el ejemplo que estamos considerando ilustra la necesidad de huir de la aplicación mecánica de los estadísticos tipo t de Student. A pesar del elevado valor numérico de este estadístico, especialmente en las regresiones individuales, la única conclusión razonable en el análisis que hemos presentado, es que ninguna de las variables, tal como aparece recogida en la muestra, explica de manera importante el comportamiento del peso de los recién nacidos[9]. Por ejemplo, la regresión *REG_EDUCP_EDUCM*, que explica el peso del recién nacido utilizando únicamente por los niveles educativos de los padres únicamente, también genera un estadístico t superior a 2,0 en valor absoluto para la variable *EDUCP*, sin que de ello deba inferirse que esta variable aporta capacidad explicativa alguna, ni siquiera cuando se utiliza por sí sóla, como ya hemos discutido ampliamente.

También es interesante observar que el estadístico tipo F habitual para el contraste de significación global del modelo, es decir, para contrastar la hipótesis nula que afirma que las variables explicativas, consideradas conjuntamente, no aportan capacidad explicativa alguna, arroja un valor numérico de 9,55, con un valor-p igual a 0, por lo que una interpretación estricta del mismo conduciría a admitir la capacidad explicativa conjunta de las variables consideradas acerca del peso de los recién nacidos, contrariamente a las conclusiones que hemos obtenido.

Sin embargo, un investigador todavía debería pronunciarse acerca de la posible evidencia existente en la información muestral sobre la influencia que las distintas variables consideradas pueden tener sobre el peso del recién nacido. En este sentido, si consideramos los cigarillos fumados durante el embarazo, la diferencia entre las medianas que antes mencionamos para los pesos de los bebés nacidos de mujeres no fumadoras y de mujeres fumadoras es notable, siendo menor la mediana del peso para los hijos de mujeres fumadoras, lo que sugiere una relación negativa entre estas dos variables, como quizá cabría esperar. Esta es la única variable recogida en la muestra para la que detectamos un efecto significativo; los datos disponibles sugieren que el consumo de cigarrillos durante el embarazo tiende a disminuir significativamente el peso de los bebés al nacer, lo que ocurre es

[9]Por supuesto, la ilustración en Wooldridge (2001) acerca de la ausencia de capacidad explicativa de las dos variables de educación es cierta. Sin embargo, el resultado es aún más estricto, por cuanto que tampoco las variables *renta, cigarrillos, ordenac*, tienen verdaderamente una capacidad explicativa de gran significación.

que la información muestral no nos permite estimar con precisión las variaciones en peso producidas por cada incremento en el número de cigarrillos fumados por la madre durante el embarazo.

Hay otros aspectos, potencialmente relevantes, que no hemos considerado en la discusión previa: los residuos de la regresión más completa, REG_TODAS, tienen una media de -3,70 para los recien nacidos de raza no blanca (186 bebés), y de 0,68 para los de raza blanca (1.005 bebés). Esto está en consonancia con la posibilidad de que los bebés de raza blanca tengan más peso. Dichos residuos tienen media de -1,92 para las mujeres, y de 1,78 para los varones, sugiriendo asimismo que los varones puden tener un peso al nacer mayor que el de las mujeres. Ambos efectos son además acordes a la intuición, por lo que procede analizarlos en algún detalle.

Al incluir ambas variables ficticias junto con las cinco variables antes analizadas, el R^2 de la regresión aumenta apreciablemente, a 0,049, a la vez que la desviación típica residual se reduce a 19,65, y el Ratio de ajuste se eleva a 3,4%. Si restringimos el modelo a incluir las dos variables ficticias, de sexo y raza, junto con *cigarrillos*, la regresión apenas varía, con residuos muy altamente correlacionados con los obtenidos en todas las regresiones consideradas, un R^2 de 0,042, desviación típica de 19,92, y *Ratio* de ajuste de 2,1%.

Esta última [REG_CIG_FIC] es, sin embargo, quizá la regresión más razonable,

$$\begin{aligned} Peso_i &= \underset{(1,306)}{113,277} - \underset{(0,090)}{0,506}\, cigarillos_i + \underset{(1,071)}{3,052}\, male_i + \underset{(1,301)}{6,230}\, white_i + u_i \; i = 1, 2, ..., N \\ \bar{R}^2 &= 0,044,\ \hat{\sigma}_u = 19,925 \end{aligned}$$

cuyos residuos aparecen en $FIG\_RES\_CIG\_FIC$. El coeficiente estimado para la variable ficticia *WHITE* es de 6,23, siendo de 3,05 el coeficiente estimado para *MALE*. El coeficiente estimado para *WHITE* está en línea con la diferencia observada entre el promedio de los residuos correspondientes a bebés de cada grupo racial en la regresión que no incluía esta variable explicativa. De igual modo, el coeficiente estimado para *MALE* es muy similar a la diferencia entre los residuos de varones y mujeres en la regresión que no incluía esta variable explicativa.

Los valores numéricos mencionados sugieren que un bebé de raza blanca pesa, en promedio, 6,23 onzas más que un bebé de otra raza de iguales características en lo relativo a: número de cigarros fumados por la madre durante el embarazo, renta familiar, niveles educativos del padre y la madre, y orden del bebé dentro de los hijos de la familia. La diferencia de peso estimada entre un niño de raza

blanca al nacer, y una niña de raza negra, es de 9,28 onzas, comparable a la que obtuvimos antes entre los bebés de madres fumadoras y no fumadoras. En la hoja de cálculo Bwght.xls se muestra cómo la mediana del número de cigarrillos fumados por la madre durante el embarazo, entre el conjunto de las observaciones en que dicho número es no nulo, es de 10 cigarillos. Ello significa que, de acuerdo con la regresión anterior, la diferencia de peso entre distintos bebés sería,

Pesos estimados	
Características recién nacido	*Peso*
Madre fumadora, mujer, no blanca	108,217
Madre fumadora, varón, no blanco	111,269
Madre no fumadora, mujer, no blanca	113,277
Madre fumadora, mujer, blanca	114,447
Madre no fumadora, varón, no blanco	116,329
Madre fumadora, varón, blanco	117,499
Madre no fumadora, mujer, blanca	119,507
Madre no fumadora, varón, blanco	122,559

Como puede apreciarse, la diferencia entre los pesos estimados de dos recién nacidos que sólo difieren en que su madre declarase ser fumadora, es siempre de 5 onzas, procedente del producto de 10 cigarrillos, escogido como representativo del nivel de tabaco consumido diariamente, por el coeficiente estimado en la regresión. Esta diferencia estimada es inferior a la diferencia de 9 onzas entre las medianas de los pesos para ambos grupos de madres.

1.5. Ejercicios

1. Contraste mediante el estadístico F la significación conjunta de las variables *ordenac, renta* y *cigarrillos* en la regresión que explica *peso* incluyendo las variables mencionadas como variables explicativas, junto con los niveles educativos del padre y la madre. Utilice la expresión del estadístico en función de las Sumas Residuales de los modelos restringido y sin restringir, así como la expresión del mismo en función de los coeficientes de determinación de ambos modelos. Compruebe que ambas expresiones generan, efectivamente, el mismo valor numérico del estadístico, 13,06, aproximadamente, que conduce al rechazo de la hipótesis nula. De acuerdo con una interpretación mecánica de este resultado, ¿cuál sería la conclusión alcanzada?. Discuta, a la luz del análisis efectuado en esta sección el resultado del contraste, y proponga una interpretación del mismo.