Notas de clase: Modelo lineal general I

Alvaro J. Flórez

2021-08-12

Contents

Introducción

Estas son las notas de clase del curso Modelo Lineal General I. Las temáticas que se tratarán son:

1. Modelo lineal simple
2. Modelo lineal múltiple
3. Evaluación de los supuestos del modelo lineal
4. Transformaciones y mínimos cuadrados ponderados
5. Evaluación de puntos influyentes y atípicos

Tenga en cuenta que el propósito de estas notas de clase no es reemplazar los textos guías. Para el estudio más detallado de los temas que se tratan, se recomiendan las siguientes lecturas:

- *Introduction to Linear Regression Analysis*, Fifth Ed., 2012, by Montgomery, D. C., Peck, E. A. and Vining, G. G. **(Texto guía)**
- *Applied Regression Analysis*, Third Ed., 1998, by Draper, N. R. and Smith, H., Wiley.
- *Theory and Applications of the Linear Models*, 2000, by Graybill, F. A., Duxbury.
- *Applied Linear Statistical Models*, Fifth Ed., 2005, by Kutner, M. H, Nachtsheim, C. J., Neter, J. and Li, W., McGraw-Hill.
- *Análisis de Regresión. Introducción Teórica y Práctica basada en R*, 2011, by F. Tusell.
- *Applied Linear Regression*, Fourth Ed., 2014, by S. Weisberg.
- *Applied Regression Analysis & Generalized Linear Models*, 2016, by J. Fox.

Chapter 1

Modelo lineal simple

Datos de peso al nacer

Los datos `birthweight` (disponible en el campus virtual) contienen el peso y la edad gestacional de 42 recién nacidos. El objetivo del estudio es investigar cómo la edad gestational del feto influyen en el peso al nacer durante las últimas semanas del embarazo. Aunque la base de datos contiene otras variables, por ahora solo consideramos el peso y la edad gestacional.

La Figura 1.1 muestra la relación entre el peso (en kilogramos) y la edad gestacional (en semanas) del recién nacido. Por medio de este gráfico vemos que hay una relación aproximadamente lineal positiva (la correlación es igual a 0.73). Es decir que cuando la edad gestacional aumenta, el peso del recién nacido también lo hace. Por lo tanto, sería razonable describir el valor esperado del peso al nacer como una función lineal de la edad gestacional:

$$E(\text{weight}|\text{age} = x) = \beta_0 + \beta_1 x.$$

La Figura 1.1 se puede hacer con el siguiente código:

```
birthweight = read.csv("birthweight.csv",header = T)

plot(weight~age,data=birthweight,pch=20,xlab="Edad gestacional(semanas)",
     ylab='Peso(kilogramos)')
```

Con este conjunto de datos podemos plantear las siguientes preguntas:

- ¿Cómo afecta la edad gestional al peso del neonato?
- Si la edad gestacional aumenta en una unidad ¿en cuanto aumenta el peso del recién nacido? ¿ese aumento puede considerarse significativo?
- ¿Se puede predecir el peso al nacer por medio de la edad gestacional?

Estas preguntas se pueden resolver a partir de un análisis de regresión lineal.

1.1 Regresion lineal simple

En una análisis de regresión simple estamos interesados en modelar la relación entre una variable de entrada (regresor, variable independiente o covariable) X y una variable de salida (respuesta o variable dependiente) Y.

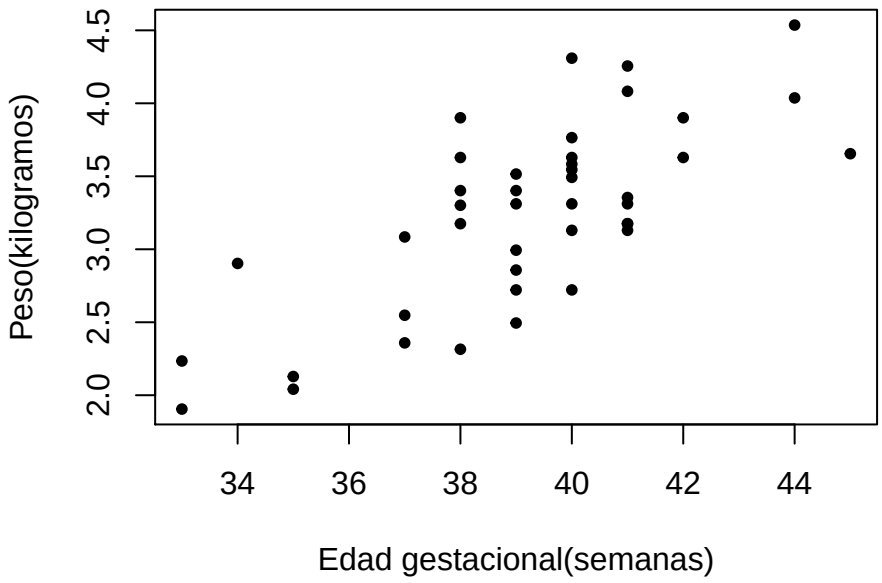

Figura 1.1: Gráfico de dispersion del peso del recien nacido y la edad gestacional.

El modelo de regresión nos permite:

- Evaluar cuanto cambia el valor esperado de Y debido a cambios en X,
- Predecir Y (o su valor esperado) en función de X.

1.2 Modelo lineal simple

Sea (y_i, x_i) la i-ésima observación de la variable respuesta (y) y la covariable (x), para $i = 1, \ldots, n$, con n igual al número total de observaciones. **El modelo lineal simple** se puede expresar de la forma:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

donde β_0 es el intercepto, β_1 es la pendiente y ε_i es el componente de error. Generalmente, los supuestos que acompañan este modelo son:

a. $E(\varepsilon_i) = 0$,
b. $V(\varepsilon_i) = \sigma^2$,
c. $cov(\varepsilon_i, \varepsilon_j) = 0$, para todo $i \neq j$,
d. ε_i se distribuye normalmente.

Por lo tanto, $\varepsilon_i \sim N(0, \sigma^2)$.

A partir de (a) se tiene que:

$$E(Y|X = x_i) = \beta_0 + \beta_1 x_i.$$

Entonces, para $x = 0$, el valor esperado de Y es igual a β_0. Cuando X incrementa en una unidad (de x a $x + 1$), el valor esperado de Y incrementa en:

$$E(Y|X = x + 1) - E(Y|X = x) = \beta_0 + \beta_1(x + 1) - (\beta_0 + \beta_1 x) = \beta_1.$$

Lo que indica que β_1 representa el cambio en el valor esperado de Y por un cambio unitario en X.

Dado (b), tenemos que $V(Y|X = x_i) = \sigma^2$. Es decir que para cualquier valor de x, la varianza de Y es la misma (homocedasticidad). Puesto que los errores están incorrelacionados (c), entonces las observaciones de Y también lo están.

Debido a (d), tenemos que la variable respuesta se distribuye de forma normal. Específicamente, tenemos que: $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$.

El proceso de generación de datos del modelo de regresión lineal se ilustra en la Figura 1.2. El valor esperado de $Y|X$ está representado por la linea negra. Tenemos que, cada observación de y (puntos negros) es una realización de la distribución de $Y|X = x_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ (curvas rojas). En esté ejemplo, suponemos que $\beta_0 = 0$, $\beta_1 = 1$, y $\sigma = 0.1$.

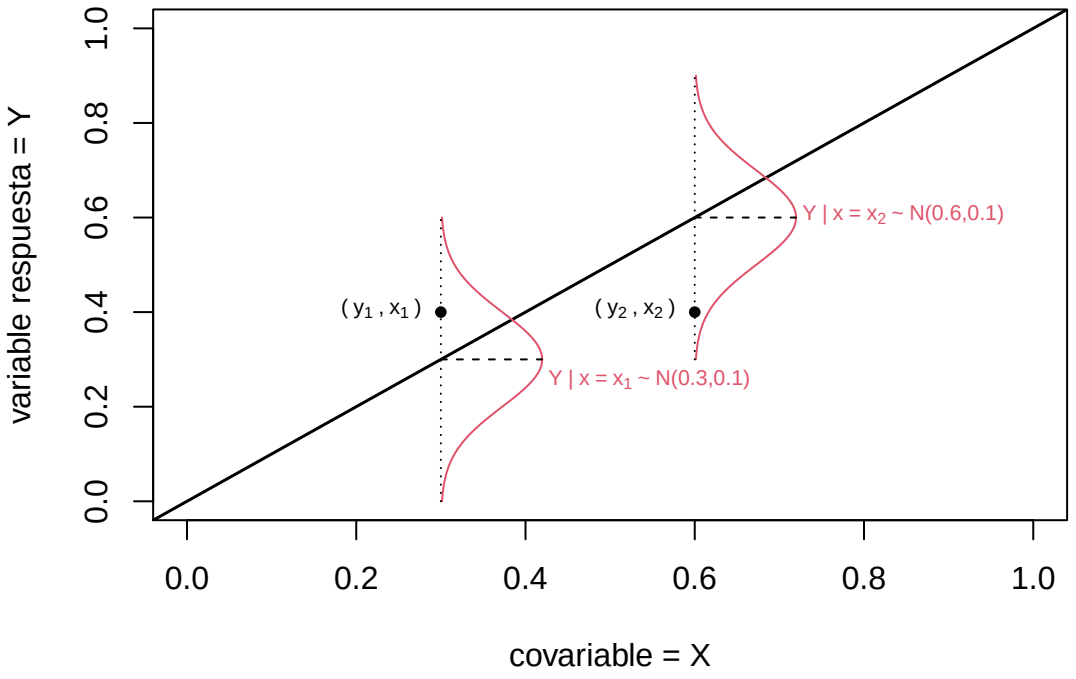

Figura 1.2: Proceso generador datos del modelo lineal simple.

1.3 Estimación de los parámetros

Los parámetros β_0 y β_1 son desconocidos y deben estimarse a partir de los datos. Para esto utilizamos el método de **mínimos cuadrados ordinarios (MCO)**.

La función objetivo es la siguiente:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = \sum_{i=1}^{n} e_i^2. \tag{1.1}$$

Entonces, tenemos que encontrar la combinación de β_0 y β_1 que minimizan (1.1). Para esto, primero debemos derivar $S(\beta_0, \beta_1)$ con respecto a β_0 y β_1, e igualar estas ecuación a cero. De esta forma obtenemos las **ecuaciones normales**:

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = -\frac{2}{n} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i), \tag{1.2}$$

y

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = -\frac{2}{n} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i. \tag{1.3}$$

Los estimadores por MCO ($\hat{\beta}_0$ y $\hat{\beta}_1$) se obtienen resolviendo el sistema de ecuaciones (1.2) y (1.3):

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

y

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} = cor(X, Y)\frac{s_Y}{s_X},$$

donde s_X y s_Y son las desviaciones estándar muestrales de X y Y.

La diferencia entre el valor observado (y_i) y el valor ajustado correspondiente ($\hat{y}_i$) es llamado **residuo** (o residual):

$$e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) = y_i - \hat{y}_i, \text{ para } i = 1, \ldots, n.$$

La Figura 1.3 presenta los residuos de forma gráfica. Estos juegan un papel importante para la evaluar la bondad del ajuste del modelo (detectar posibles desviaciones a los supuestos asumidos).

1.4 Estimación de σ^2

La estimación de σ^2 se hace a partir de la suma de cuadrados de los residuos:

$$SS_{res} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

El valor esperado de SS_{res} es $E(SS_{res}) = (n-2)\sigma^2$ (Para el caso de regresión múltiple, ver Sección C.3 de Montgomery et al. (2012)). Por lo tanto, un estimador insesgado de σ^2 es:

$$\hat{\sigma}_i = \frac{SS_{res}}{n-2} = MS_{res}.$$

La cantidad MS_{res} es llamada **cuadrado medio de los residuos**.

Datos de peso al nacer. Modelo y estimación de parametros

Para los datos de peso al nacer se propone el siguiente modelo:

$$\text{weight}_i = \beta_0 + \beta_1 \text{age}_i + \varepsilon_i, \tag{1.4}$$

donde $\varepsilon_i \sim N(0, \sigma^2)$.

En R, la estimación por MCO se realiza a través de la función `lm`:

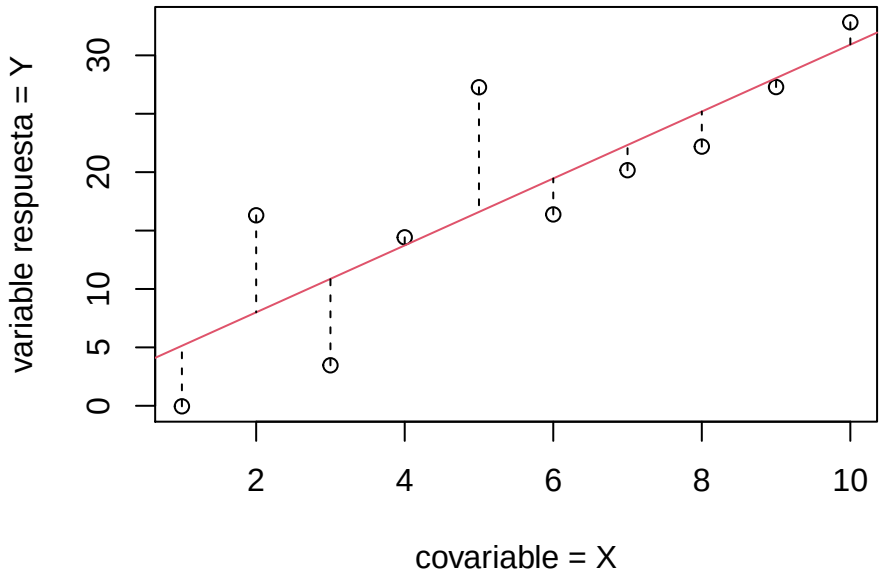

Figura 1.3: Diagrama gráfico de un ajuste por MCO. La recta representa la estimación por MCO, y las lineas discontinuas verticales entre los puntos observados y la recta estimada son los residuos.

```
mod = lm(weight~age, data=birthweight)
mod
```

```
##
## Call:
## lm(formula = weight ~ age, data = birthweight)
##
## Coefficients:
## (Intercept)          age
##     -3.6312       0.1752
```

De aquí obtenemos que $\hat{\beta}_0 = -3.63$ y $\hat{\beta}_1 = 0.18$. Note que la estimación del intercepto es negativa, lo que es físicamente imposible. Además tampoco tiene sentido una edad gestacional igual a cero. Por lo cual, este parámetro no tiene interpretación en este caso. β_0 solo tiene interpretación cuando las observaciones de x están alrededor de cero.

A partir de $\hat{\beta}_1$, podemos concluir que la edad gestacional tiene un efecto positivo sobre el peso del recién nacido (el coeficiente estimado es positivo). Por cada incremento de una semana en la edad gestacional, el valor esperado del peso del recién nacido aumenta 0.18 kilogramos.

La representación gráfica del modelo estimado se presenta en la Figura 1.4.

```
plot(weight~age,data=birthweight,pch=20,xlab="edad gestacional(semanas)",
     ylab='peso(kilogramos)')
abline(mod,lwd=2)
```

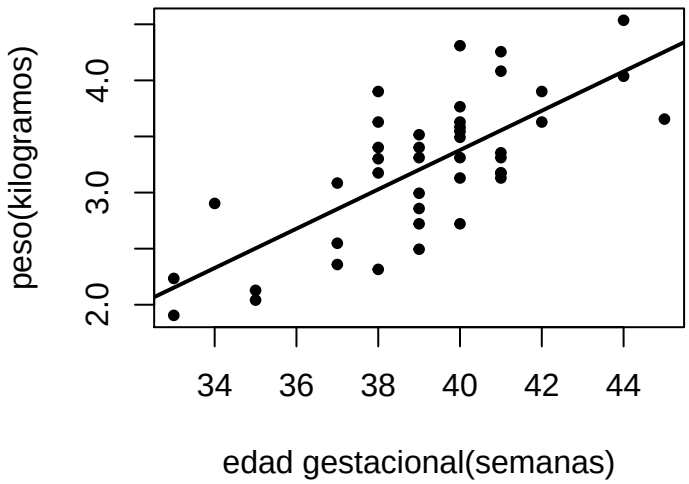

Figura 1.4: Gráfico de dispersion del peso del recien nacido y la edad gestacional. La linea representa la estimación por MCO.

La estimación de σ es:

```
sqrt(sum(mod$residuals^2)/22)
```

```
## [1] 0.5901017
```

1.5 Propiedades de los estimadores por MCO

Los estimadores de $\hat{\beta}_0$ y $\hat{\beta}_1$ son una combinación lineal de las observaciones:

$$\hat{\beta}_i = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} y_i = \sum_{i=1}^{n} c_i y_i. \tag{1.5}$$

y

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \sum_{i=1}^{n} \left(\frac{1}{n} - c_i \bar{x} \right) y_i = \sum_{i=1}^{n} d_i y_i.$$

Además, se tiene que $\sum_{i=1}^{n} c_i = 0$ y $\sum_{i=1}^{n} c_i x_i = 1$. Los valores ajustados también son combinaciones lineales de los datos:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = \sum_{i=1}^{n} (d_i + c_i x_i) y_i.$$

Puesto que los estimadores de β_0 y β_1 dependen de los errores, estos también son variables aleatorias. Por lo tanto debemos calcular el valor esperado y varianza de $\hat{\beta}_0$ y $\hat{\beta}_1$.

Si los supuestos del modelo se cumplen, tenemos que el valor esperado de $\hat{\beta}_1$ es:

$$\begin{aligned} E(\hat{\beta}_1) &= E\left(\sum_{i=1}^{n} c_i y_i \right) = \sum_{i=1}^{n} c_i E(y_i) = \sum_{i=1}^{n} c_i (\beta_0 + \beta_1 x_i + \varepsilon_i) \\ &= \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i = \beta_1. \end{aligned}$$

El valor esperado de $\hat{\beta}_0$ es:

$$\begin{aligned} E(\hat{\beta}_0) &= E\left(\bar{y} - \hat{\beta}_1 \bar{x} \right) = \frac{1}{n} \sum_{i=1}^{n} E(y_i) - \beta_1 \bar{x} \\ &= \frac{1}{n} \sum_{i=1}^{n} (\beta_0 + \beta_1 x_i) - \beta_1 \bar{x} = \beta_0. \end{aligned}$$

Es decir que $\hat{\beta}_0$ y $\hat{\beta}_1$ son estimadores insesgado de β_0 y β_1,respectivamente.

La varianza de $\hat{\beta}_1$ y $\hat{\beta}_0$ son:

$$V(\hat{\beta}_1) = V\left(\sum_{i=1}^{n} c_i y_i \right) = \sum_{i=1}^{n} c_i^2 V(y_i) = \sigma^2 \sum_{i=1}^{n} c_i^2 = \frac{\sigma^2}{S_{xx}},$$

y

$$\begin{aligned} V(\hat{\beta}_0) &= V\left(\bar{y} - \hat{\beta}_1 \bar{x} \right) = V(\bar{y}) + \bar{x}^2 V(\hat{\beta}_1) - 2\bar{x} Cov(\bar{y}, \hat{\beta}_1) \\ &= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}. \right), \end{aligned}$$

respectivamente. Finalmente, la covarianza entre $\hat{\beta}_0$ y $\hat{\beta}_1$ es:

$$\begin{aligned} Cov(\hat{\beta}_0, \hat{\beta}_1) &= Cov\left(\bar{y} - \hat{\beta}_1\bar{x}, \hat{\beta}_1\right) = Cov\left(\bar{y}, \hat{\beta}_1\right) - \bar{x}V\left(\hat{\beta}_1\right) \\ &= -\sigma^2\frac{\bar{x}}{S_{xx}}. \end{aligned}$$

Si se cumple que $E(\varepsilon_i) = 0$, $V(\varepsilon_i) = \sigma^2$ y $Cov(\varepsilon_i, \varepsilon_j) = 0$, se puede probar que los estimadores por MCO son insesgado y de varianza mínima (**teorema de Gauss-Markov**). Para la demostración en el caso de regresión múltiple, ver Sección C4 de Montgomery et al. (2012). Esto quiere decir que, comparado con todos los posibles estimadores insesgados que son combinación lineal de las observaciones, $\hat{\beta}_0$ y $\hat{\beta}_1$ tienen las varianzas más pequeñas. Por esto los estimadores por MCO son considerados los **mejores estimadores lineales insesgados**.

1.6 Inferencia

También podemos hacer pruebas de hipótesis e intervalos de confianza para los parámetros del modelo y/o pronósticos.

Por ejemplo, en los datos del peso al nacer podemos estar interesados en evaluar si la edad gestacional tiene un efecto positivo sobre el peso al nacer. Por lo tanto, debemos probar si $\beta_1 > 0$. También podríamos estar interesados en el valor esperado de un recién nacido para cierto valor especifico de edad gestacional, por ejemplo 38 semanas. Entonces, podemos calcular un intervalo de confianza para $E(Y|x = 38)$.

1.6.1 Pruebas de hipótesis

Suponga la siguiente hipótesis:

$$H_0 : \beta_1 = \beta_{10} \; H_1 : \beta_1 \neq \beta_{10}. \tag{1.6}$$

Dado que $\hat{\beta}_1$ es una combinación lineal de y_i (1.5), podemos concluir que:

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right).$$

Además,

$$t_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\frac{MS_{res}}{S_{xx}}}} \sim t_{n-2}. \tag{1.7}$$

Por lo tanto, t_0 es el estadístico de prueba para las hipótesis (1.6). Entonces, rechazamos H_0 si $|t_0| \geq t_{1-\alpha/2,n-p}$ (o por medio del valor-p asociado).

De igual forma, para evaluar:

$$H_0 : \beta_0 = \beta_{00} \; H_1 : \beta_0 \neq \beta_{00},$$

el estadístico de prueba es:

$$t_0 = \frac{\hat{\beta}_0 - \beta_{00}}{\sqrt{MS_{res}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \sim t_{n-2}. \tag{1.8}$$

1.6.2 Análisis de varianza

El análisis de varianza se basa en la partición de la variabilidad total de la variable respuesta y en dos componentes, uno debido al modelo ajustado y otro al error. Primero, empecemos con la siguiente

identidad:

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}). \tag{1.9}$$

La Figura 1.5 muestra la partición (1.9) en el punto $i = 3$. Aquí vemos que una parte de la diferencia entre y_3 y $\bar{y}$ es explicada por el modelo (línea discontinua roja).

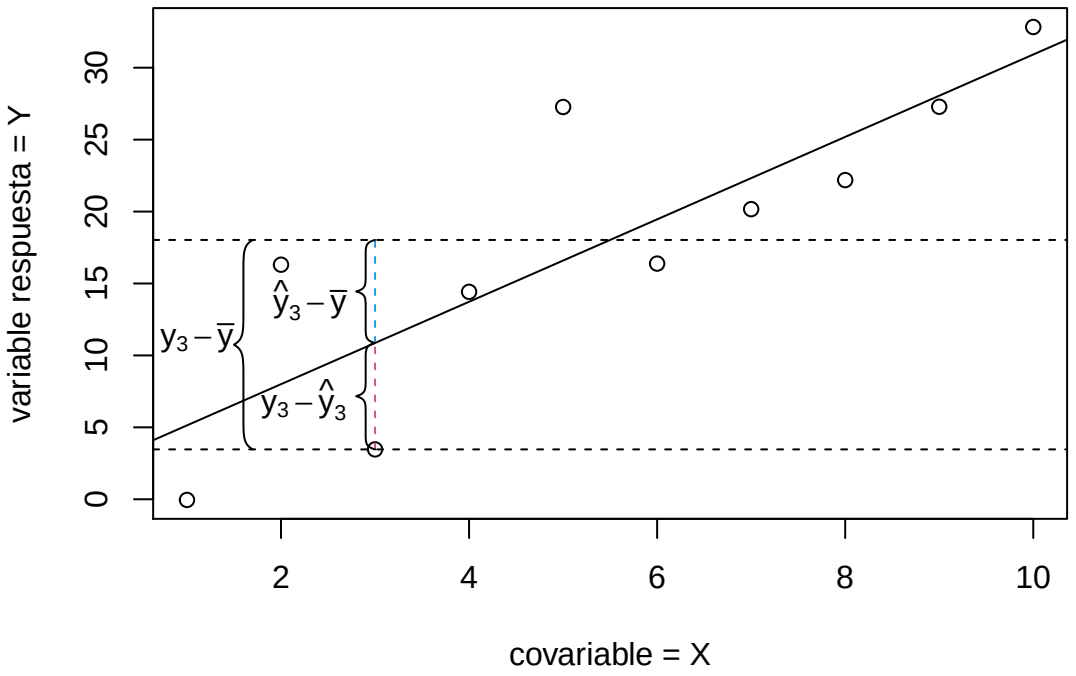

Figura 1.5: Representación gráfica de la partición
eqrefeq:decomposion.

Ahora evelavamos al cuadrado (1.9) y sumamos todos los componentes:

$$\begin{aligned} \sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + 2\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) \\ &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \\ SST &= SS_R + SS_{res}, \end{aligned} \tag{1.10}$$

donde SST es llamada la **suma de cuadrados totales** (con $n-1$ grados de libertad), SS_R es la **suma de cuadrados de la regresión** (con 1 grados de libertad), y SS_{res} es la **suma de cuadrados residual o del error** (con $n-2$ grados de libertad).

El análisis de varianza nos permite evaluar la siguiente hipótesis:

$$H_0 : \beta_1 = 0 \; H_1 : \beta_1 \neq 0. \tag{1.11}$$

Se puede demostrar que si H_0 es cierta:

$$SS_{res} = \frac{(n-2)MS_{res}}{\sigma^2} \sim \chi^2_{n-2}, \quad \frac{SS_R}{\sigma^2} \sim \chi^2_1,$$

y que SS_{res} y SS_R son independientes. Por lo tanto:

$$F_0 = \frac{SS_R/1}{SS_{res}/(n-2)} = \frac{MS_R}{MS_{res}} \sim F_{(1,n-2)}.$$

Además, $E(MS_{res}) = \sigma^2$ y $E(MS_{res}) = \sigma^2 + \beta_1^2 S_{xx}$.

Entonces, podemos utilizar F_0 como estadístico de prueba de (1.11). Rechazamos H_0 si $F_0 > F_{\alpha,1,n-2}$.

Si H_0 es falsa, F_0 sigue una distribución F no central con 1 y $n-2$ grados de libertad, y parámetro de no centralidad igual a $\lambda = (\beta_1^2 S_{xx})/\sigma^2$.

Estos resultados se pueden resumir en la Tabla 1.1.

Tabla 1.1: Tabla de ANOVA

Fuente de variación	g.l.	SS	MS	F
regresión	1	SS_R	MS_R	F_0
residuos	n-2	SS_{res}	MS_{res}	
Total	n-1	SST		

La cantidad:

$$R^2 = \frac{SS_R}{SS_{res}} = 1 - \frac{SS_{res}}{SST},$$

es llamada **coeficiente de determinación**, y cuantifica la cantidad de variabilidad de y que es explicada por x. Dado que $0 \leq SS_{res} \leq SST$, se tiene que $0 \leq R^2 \leq 1$. Por lo tanto, valores cercanos a 1 implican que el modelo explica gran parte de la variabilidad de y.

Datos de peso al nacer. Pruebas de hipótesis y ANOVA

Se quiere probar que la edad gestacional tiene influencia sobre el peso al nacer del recién nacido. Esto es:

$$H_0 : \beta_1 = 0 \qquad H_1 : \beta_1 \neq 0.$$

Tanto la prueba de hipótesis basada en t_0, el valor F_0 del ANOVA, y el R^2 se pueden observar usando la función `summary`:

```
summary(mod)
```

```
##
## Call:
## lm(formula = weight ~ age, data = birthweight)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.71314 -0.36708  0.01982  0.26741  0.93034
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.63119    1.01556  -3.576 0.000932 ***
## age          0.17525    0.02586   6.778 3.83e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4376 on 40 degrees of freedom
## Multiple R-squared:  0.5346, Adjusted R-squared:  0.5229
## F-statistic: 45.94 on 1 and 40 DF,  p-value: 3.828e-08
```

Del resultado anterior, tenemos que:

- $t_0 = 6.778$ con un valor-p asociado de 0. Por lo tanto, rechazamos H_0 y concluimos que la edad gestacional tiene un efecto significativo sobre el peso al nacer.
- La función `summary` no arroja como resultado una tabla ANOVA. Pero podemos observar el valor $F_0 = 45.94$ con un valor p asociado de 0.
- $R^2 = 0.535$, lo que indica que el 53.5% de la variabilidad del peso al nacer es explicada por la edad gestacional.

1.6.3 Intervalos de confianza

Intervalos de confianza para β_0, β_1 y σ^2

Los intervalos de confianza para β_0 y β_1 se construyen a partir de las distribuciones de probabilidad de $\widehat{\beta}_0$ y $\widehat{\beta}_1$. Esto es, (1.8) y (1.7), respectivamente.

Por lo tanto, el intervalo del $100(1-\alpha)\%$ de confianza para β_j es:

$$\widehat{\beta}_j \pm t_{1-\alpha/2,n-2}\sqrt{V(\widehat{\beta}_j)}, \text{ para } j = 0, 1.$$

Se puede demostrar que $(n-2)MS_{res}/\sigma^2 \sim \chi^2_{n-2}$. Por lo tanto:

$$P\left\{\chi^2_{\alpha/2,n-2} \leq (n-2)MS_{res}/\sigma^2 \leq \chi^2_{1-\alpha/2,n-2}\right\} = 1-\alpha.$$

Entonces, el intervalo del $100(1-\alpha)\%$ de confianza para σ^2 es:

$$\left\{\frac{(n-2)MS_{res}}{\chi^2_{1-\alpha/2,n-2}}; \frac{(n-2)MS_{res}}{\chi^2_{\alpha/2,n-2}}\right\}.$$

Intervalos de confianza para $E(y_i)$ y una predicción futura

Cuando el objetivo de ajustar un modelo de regresión es hacer predicciones, es posible hacer intervalos de confianza para la respuesta media, esto es $E(Y|x_0) = \mu_{Y|x_0}$, donde x_0 es un valor de la covariable dentro del rango de valores observados de x en los datos.

Una estimación insesgada de $\mu_{Y|x_0}$ es:

$$\widehat{\mu}_{Y|x_0} = \widehat{\beta}_0 + \widehat{\beta}_1 x_0.$$

La varianza de $\hat{\mu}_{Y|x_0}$ es:

$$\begin{aligned} V(\hat{\mu}_{Y|x_0}) &= V(\hat{\beta}_0 + \hat{\beta}_1 x_0) = V(\hat{\beta}_0) + x_0^2 V(\hat{\beta}_1) + 2x_0 Cov(\hat{\beta}_0, \hat{\beta}_1) \\ &= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) + \sigma^2 x_0^2 \frac{1}{S_{xx}} - 2\sigma^2 x_0 \frac{\bar{x}}{S_{xx}} \\ &= \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]. \end{aligned}$$

El intervalo de confianza se construye a partir de la siguiente distribución muestral:

$$\frac{\hat{\mu}_{Y|x_0} - E(Y|x_0)}{\sqrt{MS_{res}[1/n + (x_0 - \bar{x})^2/S_{xx}]}} \sim t_{n-2}.$$

Por lo tanto, el intervalo del $100(1-\alpha)\%$ de confianza para $\mu_{Y|x_0}$ es:

$$\hat{\mu}_{Y|x_0} \pm t_{1-\alpha/2,n-2} \sqrt{MS_{res}[1/n + (x_0 - \bar{x})^2/S_{xx}]}. \tag{1.12}$$

Note que la longitud del intervalo de confianza de $\hat{\mu}_{Y|x_0}$ depende del punto x_0. La menor longitud se obtiene en el punto $x_0 = \bar{x}$, y el intervalo es cada vez mas ancho a medida que nos alejamos de ese punto.

Ahora consideremos hacer una predicción de una observación futura de y para cierto valor de x. Si queremos hacer la predicción para $x = x_0$, entonces la predicción de la nueva observación es:

$$\tilde{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0.$$

Asumiendo que el modelo es correcta, el verdadero valor de y_0 es:

$$\tilde{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 + \varepsilon_0.$$

y su varianza es:

$$V(\tilde{y}_0) = \sigma^2 + \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]. \tag{1.13}$$

El primer término a la derecha de (1.13) corresponde a la variabilidad de ε_0 y el segundo al error de estimación de los coeficientes β_0 y β_1. A partir de estos resultados, el intervalo del $100(1-\alpha)\%$ de predicción de una observación futura en $x = x_0$ es:

$$\hat{\mu}_{Y|x_0} \pm t_{1-\alpha/2,n-2} \sqrt{MS_{res}[1 + 1/n + (x_0 - \bar{x})^2/S_{xx}]}.$$

Datos de peso al nacer. Intervalos de confianza

El intervalo del 95% de confianza para los parámetros del modelo (1.4) son:

```
confint(mod)
```

```
##                  2.5 %     97.5 %
## (Intercept) -5.6837190 -1.5786621
## age          0.1229923  0.2275067
```

Para β_1, con un nivel de confianza del 95% podemos decir que cuando la edad gestacional aumenta en una unidad, el peso medio del recién nacido aumenta entre 123 y 228 gramos. Como mencionamos antes, β_0 no tiene interpretación en este modelo.

El intervalo del 95% de confianza para σ se calcula "a pie" de la siguiente forma:

```
var.limInf = sum(mod$residuals^2)/qchisq(0.975,df=mod$df.residual)
var.limSup = sum(mod$residuals^2)/qchisq(0.025,df=mod$df.residual)
sqrt(c(var.limInf,var.limSup))
```

```
## [1] 0.3593008 0.5599503
```

Se quiere predecir el peso medio de los recién nacidos en la semana gestacional 36. La estimación puntual es:

$$\hat{\mu}_{Y|x_0=36} = \hat{\beta}_0 + \hat{\beta}_1(36) = -3.6312 + 0.1751 * 36 = 2.672.$$

El intervalo del 95% de confianza para $\hat{\mu}_{Y|x_0=36}$ se puede calcular de la siguiente forma:

```
x.nuevo = data.frame(age=36)
pred.media = predict(mod,x.nuevo,interval = 'confidence')
pred.media
```

```
##        fit     lwr      upr
## 1 2.677792 2.46233 2.893254
```

Esto quiere decir que, con un nivel de confianza del 95%, el peso medio de los recién nacidos en la semana gestacional 36 está entre 2.46 y 2.89 kilogramos.

Ahora, se quiere predecir el peso de un recién nacido en la semana gestacional 38, para esto calculamos un intervalo de predicción del 95%:

```
x.nuevo = data.frame(age=38)
pred.nuevaObs = predict(mod,x.nuevo,interval = 'prediction')
pred.nuevaObs
```

```
##        fit      lwr      upr
## 1 3.028291 2.131178 3.925404
```

Por lo tanto, con un nivel de confianza del 95% el peso de un recién nacido en la semana gestacional 38 está entre 2.13 y 3.93 kilogramos.

Gráficamente, podemos ver los intervalos de confianza y predicción de la siguiente forma:

```
plot(weight~age,data=birthweight,pch=20,xlab="Edad gestacional(semanas)",
     ylab='Peso(kilogramos)')
abline(mod)
lines(x.nuevo$age,pred.media[,2],lty=2)
lines(x.nuevo$age,pred.media[,3],lty=2)
lines(x.nuevo$age,pred.nuevaObs[,2],lty=3)
lines(x.nuevo$age,pred.nuevaObs[,3],lty=3)
```

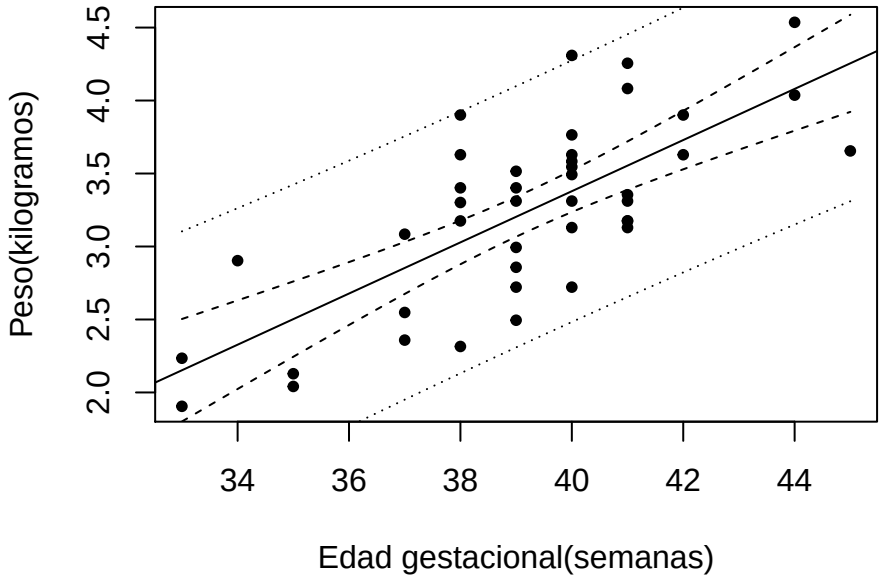

Figura 1.6: Intervalos del 95% de confianza (línea discontinua) y predicción (línea punteada) para el peso del recien nacido en función de la edad gestacional.

1.7 Estimador por máxima verosimilitud

Si consideramos que $\varepsilon_i \sim N(0, \sigma^2)$, entonces las observación también se distribuyen de forma normal $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$. Si además asumimos que las observaciones (y_i, x_i) son independientes, entonces la función de verosimilitud es:

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} \left(\sqrt{2\pi\sigma^2}\right)^{-1/2} \exp\left[-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2\right] \\ &= \left(\sqrt{2\pi\sigma^2}\right)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right], \end{aligned}$$

para $\theta = (\beta_0, \beta_1, \sigma^2)'$. La log-verosimilitud es:

$$\ell(\theta) = -\left(\frac{n}{2}\right)[\log(2\pi) - \log\sigma^2] - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2.$$

El estimador de máxima verosimilitud de θ debe satisfacer:

$$\left.\frac{\partial \ell(\theta)}{\partial \beta_0}\right|_{\tilde{\theta}} = \frac{1}{\tilde{\sigma}^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0,$$

$$\left.\frac{\partial \ell(\theta)}{\partial \beta_1}\right|_{\tilde{\theta}} = \frac{1}{\tilde{\sigma}^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i = 0,$$

y

$$\left.\frac{\partial \ell(\theta)}{\partial \sigma^2}\right|_{\tilde{\theta}} = -\frac{n}{2\tilde{\sigma}^2} + \frac{n}{2\tilde{\sigma}^4}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 = 0,$$

Luego de solucionar las ecuaciones anteriores se obtienen los estimadores por máxima verosimilitud:

$$\tilde{\beta}_0 = \bar{y} - \tilde{\beta}_1 \bar{x},$$

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{n} y_i(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

y

$$\tilde{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2}{n}.$$

Aquí observamos que los estimadores por máxima verosimilitud para β_0 y β_1 son equivalente a los estimadores por MCO. El estimador de σ^2 es sesgado, sin embargo el sesgo disminuye a medida que n crece.

Por lo general, los estimadores por máxima verosimilitud tienen mejores propiedades que los estimadores por MCO. Son asintoticamente insesgados, consistentes y asintoticamente de mínima varianza. Sin embargo, estos requieren de supuestos distribucionales completos. Recordemos que los estimadores por MCO solo requieren de una correcta especificación de los dos primeros momentos (valor esperado, varianza y covarianza).

1.8 Algunas consideraciones finales

- Las conclusiones sobre los modelos de regresión se hacen sobre el rango de valores observados de las covariables (interpolación). Por ejemplo, en los datos de los recién nacidos, se pueden hacer inferencias sobre el peso al nacer para bebés que nacen en entre las semanas 33 y 45. Cuando hacemos predicciones fuera de este rango estaríamos extrapolando.

Por extrapolación nos referimos a hacer predicciones fuera del rango observado de x. La Figura 1.7 muestra el problema que se puede cometer cuando extrapolamos. Si tenemos datos en el rango $x_{min} \leq x \leq x_{max}$, un modelo lineal es una buena aproximación de $E(Y|x)$. Pero, esa aproximación no es buena para $x > x_{max}$. Por lo tanto, se estaría cometiendo errores graves cuando hacemos predicciones de Y para valores de x mayores a x_{max} (por ejemplo en el punto x_0).

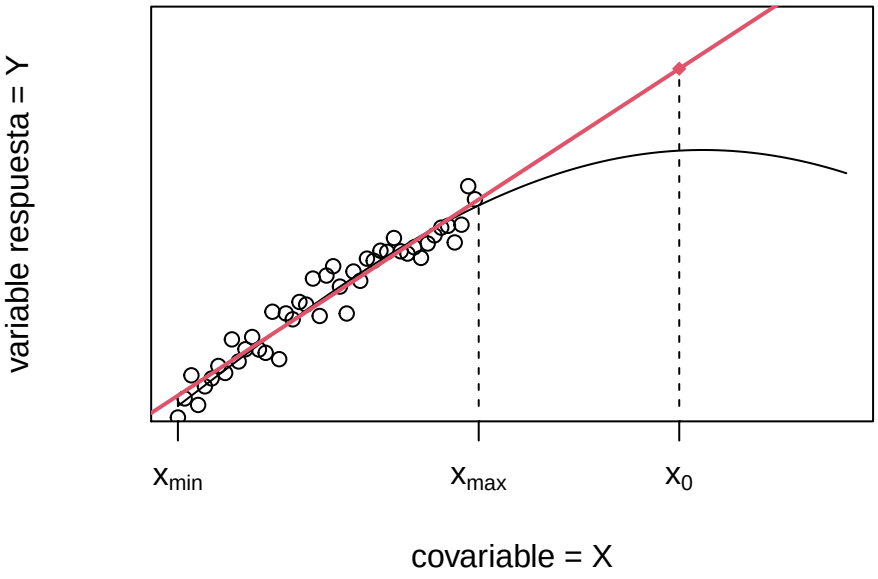

Figura 1.7: Peligro de extrapolar. La curva negra representa $E(Y|x)$ y la línea roja es el ajuste del modelo lineal con los datos observados de x. La predicción en el punto x_0 es bastante sesgada.

- La posición de los valores de x tienen una influencia sobre el ajuste por MCO. Particularmente, la estimación de β_1 está fuertemente influenciada por los valores alejados de x y y. A estos puntos se les denomina **puntos influyentes**. En la Figura 1.8(a) podemos ver como un solo punto tiene una influencia alta en la estimación de los parámetros del modelo.
- Los **valores atípicos** son observaciones que difieren considerablemente del resto de los datos (generalmente en y). Un punto atípico puede afectar la estimación de β_0 (ver Figura 1.8(b) y la estimación de σ^2.

Los métodos para la detección de puntos atípicos e influyentes se presentarán en un capítulo posterior.

- Una fuerte relación entre dos variables no necesariamente implica que la relación entre las variables es de causa-efecto. Un modelo de regresión nos permite modelar variables que estén correlacionadas, pero no se puede concluir que, necesariamente, es una relación causal.

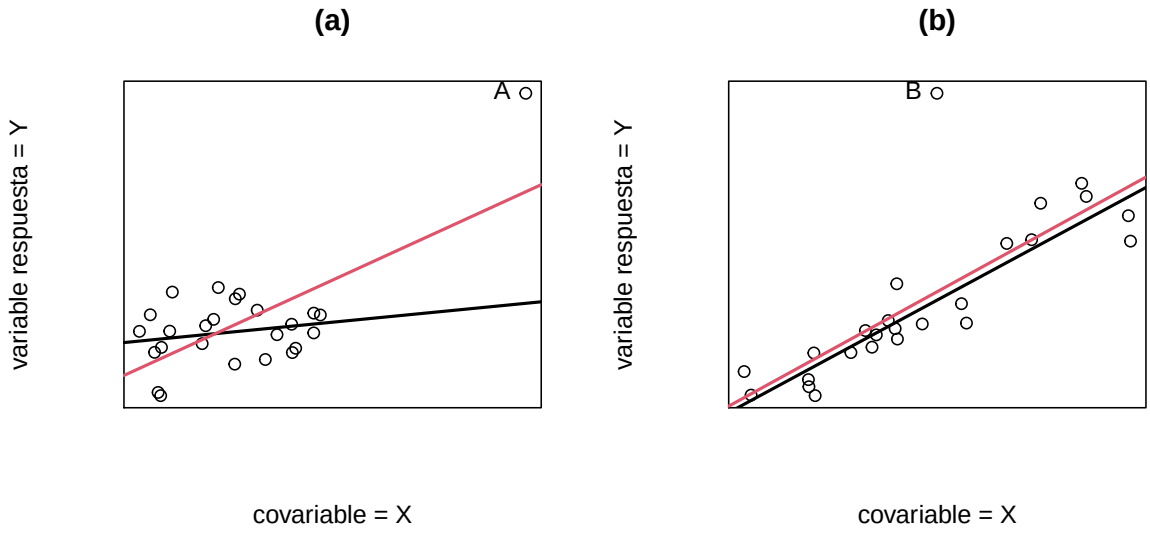

Figura 1.8: (a) Efecto de un punto influyente, la línea negra es la estimación sin el punto A y la linea roja es la estimación incluyento el punto A. (b) Efecto de un punto atípico, la línea negra es la estimación sin el punto B y la linea roja es la estimación incluyento el punto B.

Chapter 2

Modelo lineal múltiple

Bajo peso al nacer

Retomemos la base de datos de bajo peso al nacer (disponible en el campus virtual). Aparte de la edad gestacional, el peso del recién nacido puede estar explicado con otros factores. Por ejemplo, el peso de los padres, salud de la madre, entre otros. A parte de la edad gestacional y el peso del recién nacido, vamos a observar también la variable peso de la madre antes del embarazo.

La Figura 2.1 muestra la relación entre las variables de estudio. Aquí podemos observar una relación lineal positiva fuerte entre el peso al nacer y la edad gestacional (correlación igual a 0.73). La relación entre el peso al nacer y el peso de la madre es lineal positiva, aunque no tan fuerte como la anterior (correlación igual a 0.3).

La Figura 2.1 y la matriz de correlación se pueden hacer con los siguientes códigos:

```
birthweight = read.csv("birthweight.csv",header = T)
pairs(birthweight[,c(3,4,8)])
```

```
cor(birthweight[,c(3,4,8)])
```

```
##           weight       age     mppwt
## weight 1.0000000 0.7311334 0.3048027
## age    0.7311334 1.0000000 0.2505155
## mppwt  0.3048027 0.2505155 1.0000000
```

Por lo tanto, junto con la edad gestacional, vamos a incluir peso de la madre antes del embarazo (en kgs, `mppwt`) como covariable. Por lo tanto, el modelo propuesto es:

$$\text{weight}_i = \beta_0 + \beta_1 \text{weight}_i + \beta_2 \text{mppwt}_i + \varepsilon_i, \text{ para } i = 1, \dots, 42,$$

con $\varepsilon_i \sim N(0, \sigma^2)$, y $cov(\varepsilon_j, \varepsilon_k) = 0$ para todo $j \neq k$.

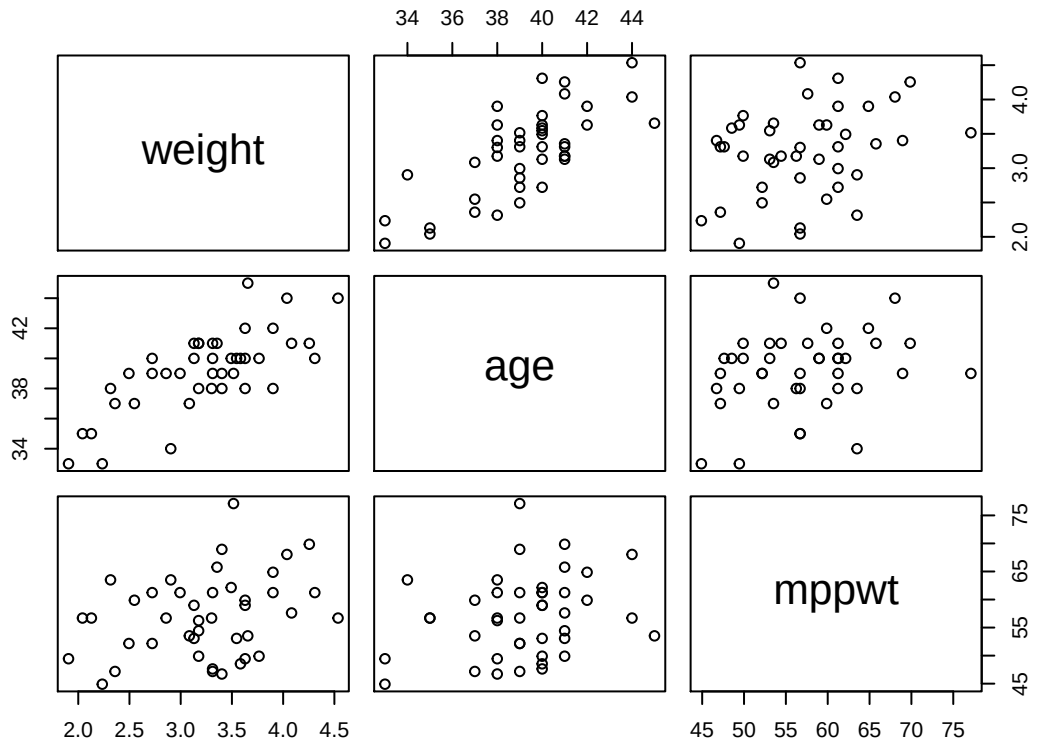

Figura 2.1: Gráfico de dispersion del peso del recien nacido y la edad gestacional.

2.1 Modelo lineal múltiple

En general, se puede relacionar la variable respuesta (y), con k covariables o variables predictoras. El modelo lineal múltiple se expresa de la siguiente forma:

$$\begin{aligned} y_i =& \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_{p-1} x_{i,p-1} + \varepsilon_i \\ =& x_i'\beta + \varepsilon_i, \qquad i = 1, ..., n, \end{aligned} \tag{2.1}$$

donde $x_i = (1, x_{i1}, x_{i2}, ..., x_{i,p-1})'$ es el vector de dimensión p de covariables del individuo i y $\beta = (\beta_0, \beta_1, ..., \beta_{p-1})'$ es el vector de dimensión p de coeficientes de regresión.

Los supuestos del modelo son los mismos que se plantearon en el capítulo anterior. Estos es: $\varepsilon_i \sim N(0, \sigma^2)$ y $cov(\varepsilon_j, \varepsilon_k) = 0$, para todo $j \neq k$.

Dado que $E(\varepsilon_i) = 0$, el valor esperado de Y es:

$$E(Y|x_{i1}, x_{i2}, ..., x_{i,p-1}) = E(Y|x_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_{p-1} x_{i,p-1} = x_i'\beta. \tag{2.2}$$

El intercepto β_0 es el valor esperado de Y cuando $x_i = (1, 0, 0, ..., 0)'$, es decir cuando todos las covariables toman el valor 0.

El parámetro de pendiente β_j indica el cambio en el valor esperado de Y debido a un aumento unitario en la covariable x_j cuando todas las demás variables predictoras se mantienen constantes. Sean $x_{i,j} = (1, x_{i1}, ..., x_{ij}, ..., x_{i,p-1})$ y $x_{i,j+1} = (1, x_{i1}, ..., x_{ij} + 1, ..., x_{i,p-1})$. A partir de (2.2), tenemos:

$$E(Y|x_{i,j}) = \beta_0 + \beta_1 x_{i1} + ... + \beta_j x_{ij} + ... + \beta_{p-1} x_{i,p-1},$$

y

$$E(Y|x_{i,j+1}) = \beta_0 + \beta_1 x_{i1} + ... + \beta_j (x_{ij} + 1) + ... + \beta_{p-1} x_{i,p-1}.$$

De aquí tenemos que:

$$E(Y|x_{i,j+1}) - E(Y|x_{i,j}) = \beta_j.$$

Es conveniente escribir el modelo de regresión múltiple (2.1) de forma matricial:

$$y = X\beta + \varepsilon,$$

donde:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} & x_{12} & ... & x_{1,p-1} \\ 1 & x_{21} & x_{22} & ... & x_{2,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & ... & x_{n,p-1} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Además, los supuestos sobre los errores se pueden expresar como $\varepsilon \sim N(0, \sigma^2 I)$, donde 0 es un vector con todas las entradas iguales a cero, y I es la matriz identidad.

2.2 Estimación de los parámetros de regresión

La estimación de β se hace a través del método de mínimos cuadrados ordinarios. Por lo tanto, debemos encontrar el vector $\widehat{\beta}$ que minimice:

$$S(\beta) = \sum_{i=1}^{n} (y_i - x_i'\beta)^2 = \sum_{i=1}^{n} \epsilon_i^2.$$

En forma matricial, tenemos:

$$\begin{aligned} S(\beta) &= \varepsilon'\varepsilon = (y - X\beta)'(y - X\beta) \\ &= y'y - \beta' X'y - y'X\beta + \beta' X'X\beta \\ &= y'y - 2\beta' X'y + \beta' X'X\beta. \end{aligned}$$

Por lo tanto, $\hat{\beta}$ debe satisfacer:

$$\left.\frac{\partial S}{\partial \beta}\right|_{\hat{\beta}} = -2X'y + 2X'X\hat{\beta} = 0.$$

A partir de aquí obtenemos las **ecuaciones normales**:

$$X'X\hat{\beta} = X'y.$$

En más detalle:

$$\begin{pmatrix} n & \sum x_{i1} & \sum_{i=1}^{n} x_{i2} & \dots & \sum_{i=1}^{n} x_{i,p-1} \\ \sum x_{i1} & \sum x_{i1}^2 & \sum x_{i1}x_{i2} & \dots & \sum x_{i1}x_{i,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum x_{i,p-1} & \sum x_{i1}x_{i,p-1} & \sum x_{i2}x_{i,p-1} & \dots & \sum x_{i,p-1}^2 \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_{p-1} \end{pmatrix} = \begin{pmatrix} \sum y_i \\ \sum x_{i1}y_i \\ \vdots \\ \sum x_{i,p-1}y_i \end{pmatrix}$$

Por lo cual, el estimador por mínimos cuadrados es:

$$\hat{\beta} = (X'X)^{-1}X'y.$$

Note que es necesario que X sea de rango completo, rango$(X) = p \leq n$. Esta restricción es necesaria para asegurar que $X'X$ sea no singular. Si $X'X$ es singular, implica que existe una combinación lineal entre las columnas de X, o que rango$(X) < p$.

El valor ajustado de y para el vector de covariables x_i es $\hat{y}_i = x_i'\hat{\beta}$. Definiendo $\hat{y} = (\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n)$, tenemos que:

$$\hat{y} = X\hat{\beta} = X(X'X)^{-1}X'y = Hy.$$

La matriz $(n \times n)$ $H = X(X'X)^{-1}X'$ es llamada **matriz hat** (sombrero) y desempeña un papel importante en el análisis de regresión.

Los residuos del modelo ($e_i = y_i - \hat{y}_i$) también se pueden expresar en forma matricial:

$$e = y - X'\hat{\beta} = y - X(X'X)^{-1}X'y = y - Hy = (I_n - H)y.$$

2.2.1 Estimación de σ^2

Al igual que en la regresión simple, el estimador de σ^2 es el cuadrado medio del error, definido como:

$$MS_{res} = \frac{SS_{res}}{n - p},$$

donde:

$$\begin{aligned} SS_{res} &= \sum_{i=1}^{n} e_i^2 = e'e = (y - X\hat{\beta})'(y - X\hat{\beta}) \\ &= (y - Hy)'(y - Hy) = y'(I_n - H)'(I_n - H)y = y'(I_n - H)y. \end{aligned}$$

Se puede demostrar que MS_{res} es un estimador insesgado de σ^2, es decir $E(MS_{res}) = \sigma^2$. Para esto debemos calcular el valor esperado de SS_{res}.

Sabemos que $E(y) = X\beta$ y $V(y) = \sigma^2 I_n$, entonces:

$$E(SS_{res}) = E\left[y'(I - H)y\right] = \sigma^2 \text{tr}\,(I - H) + \beta' X'(I - H)X\beta = (n - p)\sigma^2.$$

Por lo tanto, $E(MS_{res}) = E(SS_{res})/(n - p) = \sigma^2$.

2.2.2 Bajo peso al nacer - estimación de parámetros

Para ajustar el modelo:

$$\text{weight}_i = \beta_0 + \beta_1 \text{weight}_i + \beta_2 \text{mppwt}_i + \varepsilon_i, \text{ para } i = 1, \ldots, 42,$$

con $\varepsilon_i \sim N(0, \sigma^2)$, y $cov(\varepsilon_j, \varepsilon_k) = 0$ para todo $j \neq k$, usamos la función `lm` de R:

```
mod = lm(weight ~ age + mppwt, data=birthweight)
mod
```

```
##
## Call:
## lm(formula = weight ~ age + mppwt, data = birthweight)
##
## Coefficients:
## (Intercept)          age        mppwt
##    -3.97750      0.16746      0.01142
```

De aquí temenos que:

$$E(\text{weight}|\text{age}, \text{mppwt}) = -6.33824 + 0.16443\text{age} + 0.01914\text{mheight}.$$

Es decir que ambas covariables tienen un efecto positivo sobre el peso del bebé al nacer. Especificamente, tenemos que:

- Si la edad gestacional aumenta en una semana y el peso de la madre se mantiene constante, el valor esperado del peso al nacer crece 167 gramos.
- Por cada incremento de un kilogramo en el peso de la madre y manteniendo la edad gestacional constante, el peso al nacer medio aumenta 11 gramos.

Además, la estimación de σ^2 es:

```
sqrt(sum(mod$residuals^2)/22)
```

```
## [1] 0.5800073
```

Note que al adicionar la covariable `mppwt` se redujo el MS_{res}.

2.2.3 Propiedades de los estimadores por MCO

El valor esperado de $\hat{\beta}$ es:

$$\begin{aligned} E(\hat{\beta}) =& E\left[(X'X)^{-1}X'y\right] = E\left[X'X)^{-1}X'(X\beta + \varepsilon)\right] \\ =& E\left[(X'X)^{-1}X'X\beta + (X'X)^{-1}X'\varepsilon\right] = \beta \end{aligned}$$

Por lo tanto, $\hat{\beta}$ es un estimador insesgado de β (si el modelo está bien espeficado).

La matriz de varianzas-covarianzas de $\hat{\beta}$ es:

$$\begin{aligned} V(\hat{\beta}) &= V\left[(X'X)^{-1}X'y\right] = (X'X)^{-1}X'V(y)X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}X'X(X'X)^{-1} = \sigma^2(X'X)^{-1}. \end{aligned}$$

Si $C = (X'X)^{-1}$, entonces $V(\hat{\beta}_j) = \sigma^2 c_{jj}$ y $Cov(\hat{\beta}_j, \hat{\beta}_k) = \sigma^2 c_{jk}$, donde c_{jk} es la entrada (j, k) de la matriz C.

Teorema de Gauss-Markov

Si, $E(\varepsilon) = 0$ y $V(\varepsilon) = \sigma^2 I_n$, el estimador por MCO, $\hat{\beta} = (X'X)X'y$, es el mejor estimador lineal insesgado de β. Esto quiere decir que es el estimador con menor varianza entre la clase de estimador insesgados que son combinaciones lineales de y. Para la demostración, ver Sección C4 de Montgomery et al. (2012).

Además, si $\varepsilon \sim N(0, \sigma^2 I_n)$, el estimador por MCO coincide con el estimador por máxima verosimilitud.

2.3 Pruebas de hipótesis

Después de estimar el modelo podemos preguntarnos:

- ¿el modelo hace un buen ajuste de los datos?
- ¿cuales regresores específicos parecen importantes?

Para resolver estas preguntas podemos realizar pruebas de hipótesis. Generalmente, estos test requieren que $\varepsilon \sim N(0, \sigma^2 I_n)$.

2.3.1 Análisis de varianza

Para probar la significancia del modelo (determinar si que la relación entre y y algunas de las covariables es lineal) se plantean las siguientes hipótesis:

$$\begin{aligned} &H_0 : \beta_1 = \beta_2 = ... = \beta_{p-1} = 0 \\ &H_1 : \beta_j \neq 0 \text{ para al menos un } j. \end{aligned} \tag{2.3}$$

El rechazo de esta hipótesis nula implica que al menos uno de los regresores $x_1, x_2, ..., x_{p-1}$ contribuye significativamente al modelo.

Igual que en la regresión simple, el estadístico de prueba se encuentra a partir de la partición de la suma de cuadrados totales:

$$SS_T = SS_R + SS_{res},$$

donde:

- $SS_T = \sum_{i=1}^{n}(y_i - \bar{y})^2 = (y - \frac{1}{n}1'y)'(y - \frac{1}{n}1'y)$,
- $SS_R = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = (Hy - \frac{1}{n}1'y)'(Hy - \frac{1}{n}1'y)$,
- $SS_{res} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = (y - Hy)'(y - Hy)$,

y 1 es un vector cuyas entradas son iguales a 1.

Si H_0 es cierta, tenemos que:

$$\frac{SS_{res}}{\sigma^2} \sim \chi^2_{n-p} \text{ y } \frac{SS_R}{\sigma^2} \sim \chi^2_{p-1},$$

además, SS_{res} y SS_R son independientes. Por lo tanto,

$$F_0 = \frac{SS_R/(p-1)}{SS_{res}/(n-p)} = \frac{MS_R}{MS_{res}} \sim F_{p-1,n-p}.$$

También se puede probar que:

$$E(MS_{res}) = \sigma^2 \text{ y } E(MS_R) = \sigma^2 + \frac{\beta^{*'} X_c' X_c \beta^*}{(p-1)\sigma^2},$$

donde $\beta^* = (\beta_1, \beta_2, \dots, \beta_{p-1})'$ y

$$X_c = \begin{pmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \dots & x_{1,p-1} - \bar{x}_{p-1} \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \dots & x_{2,p-1} - \bar{x}_{p-1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{i1} - \bar{x}_1 & x_{i2} - \bar{x}_2 & \dots & x_{i,p-1} - \bar{x}_{p-1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \dots & x_{n,p-1} - \bar{x}_{p-1} \end{pmatrix}.$$

Si H_0 no es cierta, tenemos que F_0 sigue una distribución F no-central con $p-1$ y $n-p$ grados de libertad y parámetro de no centralidad:

$$\lambda = \frac{\beta' X_c' X_c \beta}{\sigma^2}.$$

Estos nos resultados nos indican que si el valor F_0 es grande, entonces al menos un β_j es diferente de cero.

Por lo tanto, para probar las hipótesis (2.3) calculamos el estadístico de prueba $F_0 = \frac{MS_R}{MS_{res}}$, y rechazamos H_0 si $F_0 > F_{1-\alpha,p-1,n-p}$.

A partir de las sumas de cuadrados podemos calcular el coeficiente de determinación:

$$R^2 = 1 - \frac{SS_{res}}{SS_T}.$$

A medida que agregamos mas covariables al modelo el R^2 aumenta (o permanece igual), sin importar si la covariable agregada tiene una contribución importante en el ajuste. Esto hace que sea difícil determinar si el incremento en el R^2 al agregar una covariable sea relevante. Por esta razón, también podemos usar el coeficiente de determinación ajustado:

$$R^2_{adj} = 1 - \frac{SS_{res}/(n-p)}{SS_T/(n-1)} = 1 - \frac{MS_{res}}{SS_T/(n-1)}.$$

Aunque no tiene interpretación, el R^2_{adj} puede usarse para comparar modelos al agregar covariables. Dado que $SS_T/(n-1)$ es constante, el R^2_{adj} solo aumentará al agregar una covariable nueva al modelo si la adición de la covariable reduce el MS_{res}.

2.3.2 Pruebas individuales sobre los coeficientes

Al rechazar H_0 de la prueba de hipótesis (2.3) concluimos que al menos un coeficiente es diferente de cero. Por lo tanto, una o más covariables tienen un aporte significativo en el modelo. El paso que sigue es identificar estas covariables.

Para esto podemos plantear las siguientes hipótesis individuales sobre los coeficientes del modelo:

$$H_0 : \beta_j = 0 \qquad H_1 : \beta_j \neq 0.$$

El estadística de prueba es:

$$t_0 = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{\sqrt{MS_{res} c_{jj}}},$$

donde c_{jj} es la entrada (j,j) de la matriz $C = (X'X)^{-1}$. Rechazamos H_0 si $|t_0| > t_{1-\alpha/2,n-p}$.

Este es una prueba parcial, puesto que estamos evaluando la significancia de x_j cuando las demás covariables x_k, para $k \neq j$, ya están incluidas en el modelo. Por lo tanto, si no rechazamos H_0, podemos concluir que, cuando los demás regresores están en el modelo, la covariable x_j no tiene un aporte significativo. Por lo tanto, podríamos retirarla del modelo.

2.3.3 Pruebas sobre subconjuntos de coeficientes

Para probar la significancia de un subconjunto de coeficientes del modelo hacemos uso de la **suma de cuadrados extra**. Primero, consideremos el siguiente modelo de regresión:

$$y = X\beta + \varepsilon,$$

donde X es una matrix $n \times p$ y β es el vector de coeficientes de longitud p. Queremos probar si un subconjunto $r < p$ de covariables tienen un aporte significativo en el modelo. Para esto hacemos la siguiente partición del vector β:

$$\beta = \left(\begin{array}{c} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-r-1} \\ \hline \beta_{p-r} \\ \beta_{p-r+1} \\ \vdots \\ \beta_p \end{array}\right) = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix},$$

donde β_1 y β_2 son vector de dimensión $(p-r)$ y (r), respectivamente. Por lo tanto, queremos realizar la siguiente prueba de hipótesis:

$$H_0 : \beta_2 = 0 \qquad H_1 : \beta_2 \neq 0. \tag{2.4}$$

El modelo anterior se puede re-escribir de la siguiente forma:

$$y = X\beta + \varepsilon = X_1\beta_1 + X_2\beta_2 + \varepsilon,$$

donde X_1 es la matriz $n \times (p-r)$ que contiene las columnas de X asociadas con β_1, y X_2 es la matriz $n \times r$ que contiene las columnas de X asociadas con β_2. Este es llamado el **modelo completo**.

Para el modelo completo tenemos:

- Estimador de β:
$$\hat{\beta} = (X'X)^{-1}X'y.$$
- Suma de cuadrados del modelo:
$$SS_R(\beta) = \hat{\beta}'X'y \text{ (con } p \text{ grados de libertad).}$$
- Cuadrado medio del error:
$$MS_{res} = \frac{y'y - \hat{\beta}'X'y}{n-p}.$$

Para evaluar la contribución de los regresores asociados a β_2, ajustamos el modelo asumiendo que H_0 es cierta. De esta forma tenemos el **modelo reducido**:

$$y = X_1\beta_1 + \varepsilon.$$

Para el modelo reducido tenemos:

- Estimador de β_1:
$$\hat{\beta}_1 = (X_1'X_1)^{-1}X_1'y.$$
- Suma de cuadrados del modelo:
$$SS_R(\beta_1) = \hat{\beta}_1'X_1'y \text{ (con } p-r \text{ grados de libertad).}$$

Entonces, la suma de cuadrados debido a β_2 dado que β_1 ya está en el modelo es:

$$SS_R(\beta_2|\beta_1) = SS_R(\beta) - SS_R(\beta_1),$$

con $p-(p-r)=r$ grados de libertad. Esta suma de cuadrados es llamada la suma de cuadrados extra debido a β puesto que mide el incremento en la suma de cuadrados de la regresión como resultado de adicionar los regresores X_2 en el modelo que ya contiene X_1.

Dado que $SS_R(\beta_2|\beta_1)$ y MS_{res} son independientes, podemos utilizar el siguiente estadístico de prueba:

$$F_0 = \frac{SS_R(\beta_2|\beta_1)/r}{MS_{res}}.$$

Si H_0 es cierta entonces $F_0 \sim F_{r,n-p}$. Si H_0 no es cierta, entonces F_0 sigue una distribución F no-central con parámetro de no centralidad igual a:

$$\lambda = \frac{1}{\sigma^2}\beta_2' X_2' \left[I_n - X_1(X_1'X_1)^{-1}X_1'\right] X_2\beta_2.$$

Note que si hay una relación casi colineal entre X_1 y X_2 (multicolinealidad), λ es cercano a cero pesar que β_2 sea marcadamente distinto de cero. Es decir, que la prueba tiene poca capacidad de indicar diferencias (poco poder) en presencia de multicolinealidad. Caso contrario, el máximo poder se alcanza cuando X_1 y X_2 son ortogonales (es decir $X_2'X_1 = 0$).

Entonces, si $F_0 > F_{1-\alpha,r,n-p}$ rechazamos H_0 y concluimos que al menos un coeficiente en β_2 es diferente de cero. Consecuentemente, al menos una de las covariables en X_2 tiene un aporte significativo dentro del modelo.

Ejemplo

Considere el modelo:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i.$$

La suma de cuadrados del modelo se puede descomponer de la siguiente forma:

$$SS_R = SS_R(\beta_1,\beta_2,\beta_3|\beta_0) = SS_R(\beta_1|\beta_0) + SS_R(\beta_2|\beta_0,\beta_1) + SS_R(\beta_3|\beta_0,\beta_1,\beta_2),$$

donde cada suma de cuadrados en el lado derecho tiene un grado de libertad. Además, el order de los regresores en estos componentes marginales es arbitrario. Por lo que la siguiente descomposición alternativa es también válida:

$$SS_R(\beta_1,\beta_2,\beta_3|\beta_0) = SS_R(\beta_2|\beta_0) + SS_R(\beta_3|\beta_0,\beta_2) + SS_R(\beta_1|\beta_0,\beta_2,\beta_3).$$

Sin embargo, la siguiente partición de la suma de cuadrados de la regrsión es generalmente inválida:

$$SS_R(\beta_1,\beta_2,\beta_3|\beta_0) \neq SS_R(\beta_1|\beta_0,\beta_2,\beta_3) + SS_R(\beta_2|\beta_0,\beta_1,\beta_3) + SS_R(\beta_3|\beta_0,\beta_1,\beta_2).$$

2.4 Prueba de hipótesis lineal general

Suponga que estamos interesados en las siguientes hipótesis:

$$H_0 : T\beta = 0 \qquad H_1 : T\beta \neq 0, \tag{2.5}$$

donde T es una matriz $m \times p$ de constantes, tal que r de las m ecuaciones de $T\beta = 0$ son independientes.

El **modelo completo (FM)** es:

$$y = X\beta + \varepsilon,$$

El estimador de β es $\hat{\beta} = (X'X)^{-1}X'y$, y la suma de cuadrados de los residuos es $SS_{res}(FM)$ (con $n-p$ grados de libertad).

El **modelo reducido (RM)** se obtiene al resolver las r ecuaciones independientes de $T\beta = 0$ para los r coeficientes en el modelo completo en términos de los $p-r$ coeficientes restantes. Esto lleva al siguiente RM:

$$y = Z\gamma + \varepsilon,$$

donde Z es una matriz $n \times (p-r)$ y γ es un vector de dimiensión $(p-r)$ de coeficientes de regresión. La suma de cuadrados de los residuos de este modelo es $SS_{res}(RM)$ (con $n-p+r$ grados de libertad).

Dado que el modelo reducido tiene menos parámetros que el modelo completo, $SS_{res}(RM) \geq SS_{res}(FM)$. Para probar (2.5) usamos la diferencia entre las sumas de cuadrados de los residuos:

$$SS_H = SS_{res}(RM) - SS_{res}(FM),$$

con r grados de libertad. SS_H es llamado la suma de cuadrados debido a $H_0 : T\beta = c$. El estadístico de prueba es:

$$F_0 = \frac{SS_H/r}{SS_{res}(FM)/(n-p)} = \frac{\hat{\beta}'T[T(X'X)T']^{-1}T\hat{\beta}/r}{SS_{res}(FM)/(n-p)}.$$

Rechazamos H_0 si $F_0 > F_{1-\alpha,r,n-p}$.

La hipótesis anterior se puede generalizar de la siguiente forma:

$$H_0 : T\beta = c \qquad H_1 : T\beta \neq c, \tag{2.6}$$

Para este caso, el estadístico de prueba es:

$$F_0 = \frac{(T\hat{\beta} - c)'[T(X'X)T']^{-1}(T\hat{\beta} - c)/r}{SS_{res}(FM)/(n-p)}.$$

Si H_0 es cierta, $F_0 \sim F_{r,n-p}$. Por lo tanto, rechazamos H_0 si $F_0 > F_{1-\alpha,r,n-p}$.

ejemplo

Considere el modelo:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_3 x_{i3} + \varepsilon_i,$$

y queremos probar las siguientes hipótesis:

$$\begin{aligned} H_0 :& \beta_1 = 0 \\ & 2\beta_2 - \beta_3 = 3 \end{aligned} \qquad\qquad \begin{aligned} H_1 :& \beta_1 \neq 0 \\ & 2\beta_2 - \beta_3 \neq 3 \end{aligned}$$

De aquí tenemos que:

$$T = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & -1 & 0 \end{pmatrix} \text{ y } c = \begin{pmatrix} 0 \\ 3 \end{pmatrix}.$$

Si no rechazamos H_0, podríamos estimar β sujo a las restricciones impuestar por la hipótesis nula (usando mínimos cuadrados restringidos).

Bajo peso al nacer - pruebas de hipótesis

Los resultados de las pruebas de hipótesis individuales sobre los coeficientes, análisis de varianza y coeficientes de determiandción se obtiene a partir del resumen del modelo:

```
summary(mod)
```

```
##
## Call:
## lm(formula = weight ~ age + mppwt, data = birthweight)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.79566 -0.34083  0.05415  0.26504  0.88930
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.977501   1.053336  -3.776 0.000531 ***
## age          0.167456   0.026585   6.299 1.99e-07 ***
## mppwt        0.011416   0.009756   1.170 0.249032
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4356 on 39 degrees of freedom
## Multiple R-squared:  0.5503, Adjusted R-squared:  0.5273
## F-statistic: 23.87 on 2 and 39 DF,  p-value: 1.703e-07
```

A partir de estos resutados tenemos que $F_0 = 23.8665$ con un valor-p asociado de 0.000, es decir que al menos uno de los coeficientes de regresión es diferente de cero. Además, el 55% de la variabilidad del peso al nacer es explicada por la edad gestacional y el peso de la madre antes del embarazo ($R^2 = 0.5503$). Note que hubo un incremento leve en el R^2 respecto al modelo que solo incluye la edad gestacional como covariable ($R^2 = 0.535$).

A partir de las pruebas de hipótesis individuales, podemos decir que el peso de la madre antes del embarazo no tiene un aporte significativo cuando el modelo ya incluye la covariable edad gestacional ($t_0 = 1.17$ con un valor-p asociado de 0.249). Por otro lado, el efecto de la edad gestacional si es significativo ($t_0 = 6.3$ con un valor-p asociado de 0).

Ahora consideremos un modelo ingresando dos covariables más:

$$y_i = \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{mppwt}_i + \beta_3 \text{motherage}_i + \beta_2 \text{mnocig}_i + \varepsilon_i,$$

donde motherage_i y mnocig_i es la edad (en años) y el número medio de cigarrillos fumados por mes de la i-ésima madre, respectivamente. El resumen del modelo ajustado es:

```
mod.completo = lm(weight~age + mppwt + motherage + mnocig,data=birthweight)
summary(mod.completo)
```

```
##
## Call:
## lm(formula = weight ~ age + mppwt + motherage + mnocig, data = birthweight)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -0.78765 -0.35948  0.09209  0.35024  0.75018
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.104029   1.011723  -4.056 0.000247 ***
## age          0.168027   0.024916   6.744 6.24e-08 ***
## mppwt        0.014838   0.009530   1.557 0.127966
## motherage    0.001751   0.012335   0.142 0.887900
## mnocig      -0.014417   0.005421  -2.660 0.011493 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4073 on 37 degrees of freedom
## Multiple R-squared:  0.627,  Adjusted R-squared:  0.5867
## F-statistic: 15.55 on 4 and 37 DF,  p-value: 1.5e-07
```

Respecto al modelo anterior, hay un aumento del R^2 y el R^2_{adj}. Por lo cuál podemos concluir que al ingresar estas covariables el ajuste mejoró. Aunque, los efectos del peso y la edad de la madre no son significativos a partir de las pruebas individuales.

Esto no necesariamente quiere decir que podemos eliminar estas dos covariables del modelo, recordemos que las pruebas t son individuales (se evalúa el efecto de la covariable cuando el modelo ya incluye las restantes). Para determinar si podemos eliminar `mppwt` y `motherage` del modelo, realizamos la siguiente prueba de hipótesis:

$$H_0 : \beta_2 = \beta_3 = 0 \qquad H_0 : \beta_j \neq 0 \text{ para algún } j = 2, 3.$$

En R podemos hacer esto a través de la función `anova`:

```
anova(mod,mod.completo)
```

```
## Analysis of Variance Table
##
## Model 1: weight ~ age + mppwt
## Model 2: weight ~ age + mppwt + motherage + mnocig
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     39 7.4010
## 2     37 6.1386  2    1.2624 3.8043 0.03144 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Aquí vemos que $F_0 = 3.804$ con un valor-p asociado de 0.0314. Por lo tanto, no tenemos evidencia suficiente para rechazar H_0 y podemos retirar las dos covariables del modelo.

2.5 Intervalos de confianza

Al igual que en caso del modelo lineal simple, también podemos hacer estimaciones por intervalos de confianza para los coeficientes del modelo, valor esperado de Y y observaciones futuras.

Para que los intervalos de confianza sean válidos se requiere que se cumplan todos los supuestos del modelo, esto es $\varepsilon \sim N(0, \sigma^2 I_n)$.

2.5.1 Intervalos de confianza para β_j

El intervalo de confianza de β_j parte de:

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{V(\hat{\beta}_j)}} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{MS_{res}c_{jj}}} \sim t_{n-p},$$

donde c_{jj} es la entrada (j, j) de la matriz $C = (X'X)^{-1}$. Entonces, el intervalo del $(1-\alpha)100\%$ de confianza para $\hat{\beta}_j$ es:

$$\hat{\beta}_j \pm t_{1-\alpha/2,n-p}\sqrt{MS_{res}C_{jj}}.$$

2.5.2 Intervalos de confianza para el valor esperado de Y y una observación futura

Ahora queremos construir un intervalo de confianza para la respuesta media de Y para un punto particular $x_0 = (1, x_{01}, x_{02}, ..., x_{0,p-1})$. La estimación puntual en x_0 es:

$$\hat{\mu}_{Y|x_0} = x_0'\hat{\beta}.$$

Además, tenemos que $\hat{\mu}_{Y|x_0} \sim N[x_0\beta, V(\hat{\mu}_{Y|x_0})]$ con:

$$V(\hat{\mu}_{Y|x_0}) = \sigma^2 x_0'(X'X)^{-1}x_0.$$

Por lo tanto, el intervalo del $(1-\alpha)100\%$ de confianza para $E(Y|x_0)$ es:

$$\hat{\mu}_{Y|x_0} \pm t_{1-\alpha/2,n-p}\sqrt{MS_{res}x_0'(X'X)^{-1}x_0}.$$

De igual forma, el intervalo del $(1-\alpha)100\%$ de confianza para una observación futura en x_0 es:

$$\hat{\mu}_{Y|x_0} \pm t_{1-\alpha/2,n-p}\sqrt{MS_{res}\left[1 + x_0'(X'X)^{-1}x_0\right]}.$$

Bajo peso al nacer - intervalos de confianza

Siguiendo con el modelo inicial $y_i = \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{mppwt}_i + \varepsilon_i$, los intervalos del 95% de confianza para los coeficientes son:

```
confint(mod)
```

```
##                    2.5 %     97.5 %
## (Intercept) -6.108074493 -1.8469278
## age          0.113682140  0.2212300
## mppwt       -0.008317269  0.0311501
```

Si peso de la madre permance constante, por cada aumento de una semana en la edad gestacional, el peso medio del recién nacido incrementa entre 114 y 221 gramos con un nivel de confianza del 95%. Note que el intervalo de confianza para el coeficiente asociado al peso de la madre contiene el valor 0 (recordemos que no rechazamos la hipótesis nula de la prueba t sobre β_2).

Queremos determinar el peso medio de los recién nacidos en la semana gestacional 36 y de madres que pesan 50 kilogramos. Es decir $E(Y|\text{age} = 36, \text{mppwt} = 50)$. Para esto podemos contruir un intervalo del 95% de confianza:

```
x0 = data.frame(age=36,mppwt=50)
predict(mod,x0,interval='confidence')
```

```
##        fit      lwr      upr
## 1 2.621739 2.386238 2.857239
```

Por lo tanto, el peso medio de los recién nacidos en la semana gestacional 36 y de madres que pesan 50 kilogramos está entre 2.39 y 2.86 kilogramos con un nivel de confianza del 95%.

Ahora, estamos interesados en predecir el peso de un recién nacido en la semana 38 y cuya madre peso 65 kilogramos. Por lo cuál construimos un intervalo del 95% de predicción:

```
x0pred = data.frame(age=38,mppwt=65)
predict(mod,x0pred,interval='prediction')
```

```
##        fit     lwr      upr
## 1 3.127897 2.21775 4.038044
```

El peso del recién nacido con estas caracteristica está entre 2.22 y 4.04 kilogramos con un nivel de confianza del 95%.

2.6 Extrapolación oculta en regresión múltiple

Al igual que en regresión simple, al pronosticar una nueva respuesta en un punto dado x_0 se debe tener cuidado de no extrapolar fuera de la región de los datos originales. En regresión múltiple es fácil extrapolar inadvertidamente, puesto que la región que contiene los datos está definida de forma conjunta por los valores que toman las covariables y no por el rango individual de cada covariable.

La Figura 2.2 muestra un ejemplo de extrapolación en el caso de un modelo de regresión con dos covariables. Se quiere hacer una predicción en el punto (x_{01}, x_{02}) que está dentro del rango de ambos regresores, pero que fuera de la región conjunta de los datos (región roja en la figura). Por lo tanto, al realizar la predicción en este punto estaríamos extrapolando.

Determinar la región conjunta de los datos en regresión múltiple no es fácil, lo que hace díficil saber si se está extrapolando a la hora de hacer de una predicción. Por lo tanto, se ha propuesto determinar la región conjunta de los datos a partir del conjunto convexo mínimo que contiene todos los n datos originales, $(x_{i1}; x_{i2}, \dots, x_{i,p-1})$, para $i = 1, 2, \dots, n$, como la envolvente de las covariables (RVH). Entonces, si un punto $(x_{01}, x_{02}, \dots, \dots, x_{0,p-1})$ está dentro o en la frontera de la RVH, una predicción o una estimación implica interpolación, mientras que si está fuera de la RVH, se está extrapolando.

Una aproximación de la RVH es a través de la matriz H. El conjunto de puntos x que satisfacen, $x'(X'X)^{-1}x \leq \max(h_{ii})$, prudcen un elipsoide que encierra todos los puntos dentro de la RVH. Entonces, un punto de predicción x_0 está fuera de la RVH si $h_{00} > \max h_{ii}$, donde:

$$h_{00} = x_0'(X'X)^{-1}x_0.$$

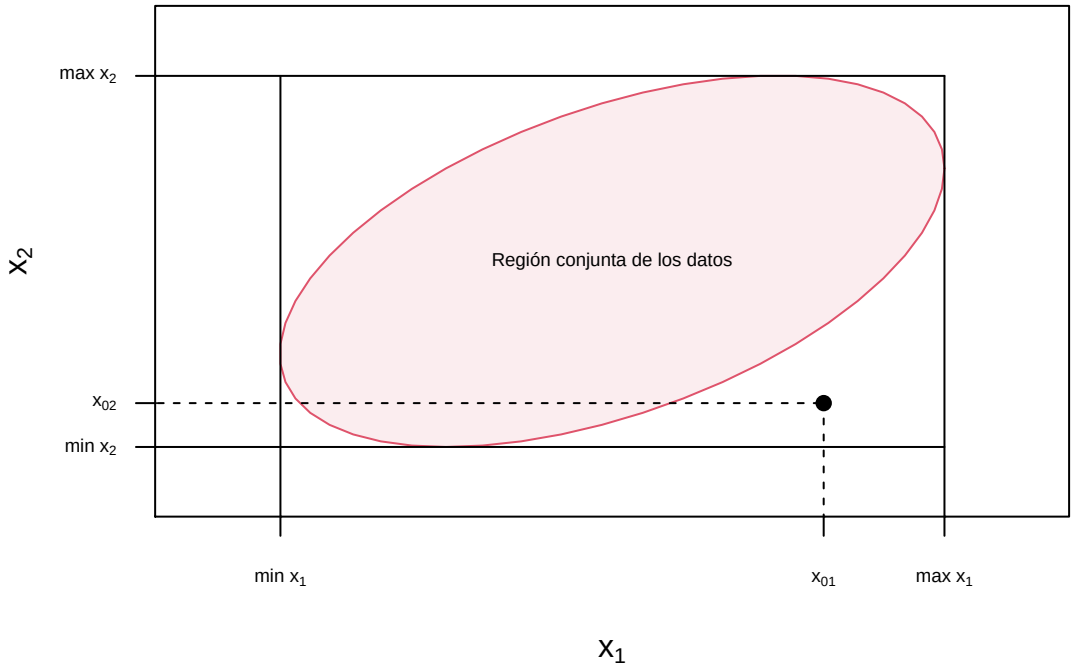

Figura 2.2: Ejemplo de extrapolación en regresión múltiple

Tabla 2.1: Bajo peso al nacer. Punto de predicción.

	1	2	3	4
Edad gestacional (semanas)	34	36	38	46
Peso de la madre (kg)	75	50	60	55

Bajo peso al nacer - interpolación

Suponga que se quiere hacer una predicción para recién nacidos con las características que muestra la Tabla @ref(tab:puntosPrediccion}.

La Figura 2.3 muestra el gráfico de dispersión de las covariables, donde los puntos rojos indican los valores donde se quieren hacer predicciones. Aquí vemos que en los puntos x_{02} y x_{03} no estaríamos extrapolando. Pero, es difícil de determinar para los puntos x_{01} y x_{04}. Para esto vamos a calcular las aproximaciones de la RVH y verificar si en estos puntos estaríamos extrapolando.

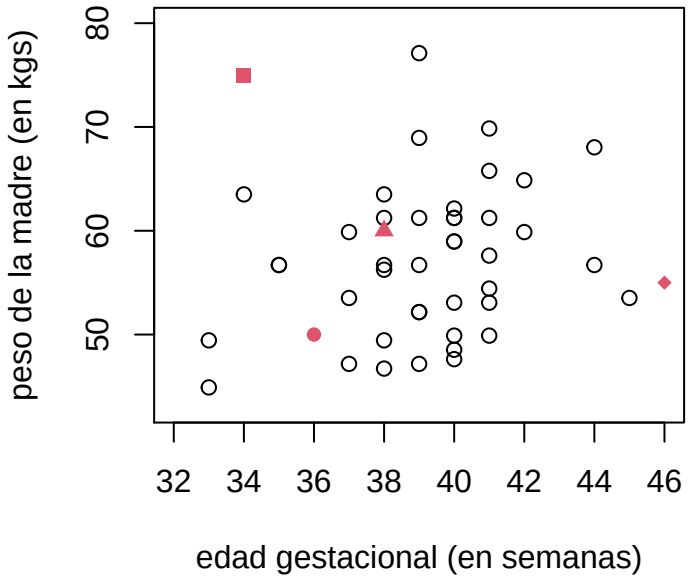

Figura 2.3: Bajo peso al nacer. Gráfico de dispersion de la edad gestacional y el peso de la madre antes del embarazo. Los puntos donde se quiere hacer predicción están en rojo.

```
newPoints = cbind(x0=rep(1,4),x1=c(34,36,38,46),x2=c(75,50,60,55))
X = model.matrix(mod)
XtX.inv = solve(t(X)%*%X)
h.values = hatvalues(mod)
hmax = max(h.values)
h0 = apply(newPoints,1,function(x){t(x)%*%XtX.inv%*%x})
h0 >hmax
```

```
## [1]  TRUE FALSE FALSE FALSE
```

Para la predicción en el punto x_{01}, tenemos que $h_0 = (1, 34, 77)(X'X)^{-1}(1, 34, 77)' = 0.3487 > h_{max} = 0.2276$. Por lo tanto, aquí se estaría extrapolando. Para el resto de punto no hay problemas de extrapolación.

2.6.1 Coeficientes normalizados de regresión

Los coeficientes de regresión están influenciados por las unidades de medida de las covariables. Exactamente las unidades de medida de β_j es:

$$\frac{\text{la unidad de medida de } y}{\text{la unidad de medida de } x_j}.$$

Dado que, por lo general, las covariables están medidas en unidades diferenes, la comparación de los coeficientes es complicada. En el ejemplo de los datos de los recién nacidos, la edad gestacional está en semanas y el peso de la madre en kilogramos.

Por esta razón, en algunas ocasiones es útil escalar los valores de las covariables y la respuesta para calcular los coeficientes de regresión adimensionales. Hay varias formas de hacer este escalamiento, aquí nos centraremos en el escalamiento de longitud unitaria.

2.6.1.1 Escalamiento de longitud unitaria

Una opción es hacer un **escalamiento de longitud unitaria** a las covariables:

$$z_{ij} = \frac{x_{ij} - \bar{x}}{\sqrt{S_{jj}}}, i = 1, 2, ..., n \quad j = 1, 2, ..., k,$$

y la variable respuesta:

$$y_i^* = \frac{y_i - \bar{y}}{\sqrt{SS_T}},$$

donde:

$$S_{jj} = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2.$$

Con estas variables transformadas, se puede ajustar el modelo:

$$y_i^* = b_1 z_{i1} + b_2 z_{i2} + ... + b_{p-1} z_{i,p-1} = z_i' b + \varepsilon.$$

El estimador por MCO es:

$$\hat{b} = (Z'Z)^{-1} Z' y^*.$$

Note que con este escalamiento, la matriz $(Z'Z)$ es igual a la matriz de correlación de las covariables. Esto es:

$$(Z'Z) = R = \begin{pmatrix} 1 & r_{12} & r_{13} & ... & r_{1,p-1} \\ r_{12} & 1 & r_{23} & ... & r_{2,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{1,p-1} & 1 & r_{2,p-1} & ... & 1 \end{pmatrix},$$

donde r_{jk} es la correlación entre las covariables x_j y x_k. Además, la matriz $Z'y^*$ es el vector de correlación entre la variable respuesta y cada covariable. Esto es:

$$Z'y^* = (r_{1y}, r_{2y}, r_{3y}, ..., r_{p-1,y})',$$

donde r_{jy} es la correlación entre la variable respuesta y la covariable x_j.

Bajo peso al nacer - coeficientes de regresión con variables escaladas

Para estimar los coeficientes de regresión escalados, primero debemos escalar las variables:

```
y = birthweight$weight
Z = apply(X[,-1],2,function(x){(x-mean(x))/sqrt(sum((x-mean(x))^2))})
ys = (y-mean(y))/sqrt(sum((y-mean(y))^2))
```

Ahora procedemos a estimar el modelo con las variables escaladas:

```
mod.std = lm(ys~Z-1)
summary(mod.std)
```

```
##
## Call:
## lm(formula = ys ~ Z - 1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.19612 -0.08401  0.01335  0.06533  0.21920
##
## Coefficients:
##        Estimate Std. Error t value Pr(>|t|)
## Zage     0.6986     0.1095   6.379 1.39e-07 ***
## Zmppwt   0.1298     0.1095   1.185    0.243
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.106 on 40 degrees of freedom
## Multiple R-squared:  0.5503, Adjusted R-squared:  0.5279
## F-statistic: 24.48 on 2 and 40 DF,  p-value: 1.142e-07
```

Las estimaciones de los coeficientes son ahora adimensionales y podemos comparar sus magnitudes. Por lo tanto, parece que la covariable edad gestacional es más importante para determinar el peso al nacer que la covariable peso de la madre. Note que, al escalar las variables, los resultados de las pruebas de hipótesis, estimación de σ^2, y los coeficientes de determinación no se ven alterados.

2.7 Multicolinealidad

Un problema que puede afectar enormente el ajuste de un modelo de regresión es la multicolinealidad. Este se presenta cuando hay una dependencia casi lineal entre las covariables.

Recordemos que el estimador por MCO es $\hat{\beta} = (X'X)^{-1}X'y$. Por lo tanto es necesario que la matriz $X'X$ sea no singular. En caso contrario, no es posible encontrar la inversa y las ecuaciones normales no tendrán una única solución. Cuando sucede esto se debe a que hay al menos una columna de X linealmente dependiente.

En regresión se utiliza las palabras multicolinealidad cuando hay una dependencia aproximada en las columnas de X. Es decir que al menos una covariable puede representarse, de forma aproximada, como una relación lineal de las otras:

$$x_{ij} \approx c_0 + c_1 x_{i1} + ... + c_{j-1}x_{i,j-1} + c_{j+1}x_{i,j+1} + ... + +c_{p-1}x_{i,p-1},$$

para $i = 1, \ldots, n$.

Hay que aclarar que la falta de ortogonalidad no es necesariamente un inconveniente, el problema es cuando la relación lineal entre los regresores es casi perfecta, lo que provoca problemas en las inferencias que se hagan. Uno de estos problemas se ilustra a continuación con un ejemplo.

ejemplo

Considere el siguiente modelo de regresión:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i, \text{ con } \varepsilon_i \sim N(0, \sigma^2),$$

y se plantean dos posibles matrices de diseño:

$$X_1 = \begin{pmatrix} 1 & 1 \\ 1 & 5 \\ 2 & 1 \\ 2 & 5 \end{pmatrix} \text{ y } X_2 = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 2 & 4 \\ 2 & 5 \end{pmatrix}.$$

Haciendo el escalamiento de longitud unitaria a las covariables tenemos que:

$$Z_1'Z_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \text{ y } Z_1'Z_1 = \begin{pmatrix} 1 & 0.95 \\ 0.95 & 1 \end{pmatrix}.$$

Por lo tanto, la varianza de $\hat{b}$ para ambos casos es:

$$V(\hat{b}_1) = \sigma^2 (Z_1'Z_1)^{-1} = \sigma^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}^{-1} = \sigma^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

y

$$V(\hat{b}_2) = \sigma_0^2 (Z_2'Z_2)^{-1} = \sigma^2 \begin{pmatrix} 1 & 0.95 \\ 0.95 & 1 \end{pmatrix}^{-1} = \sigma^2 \begin{pmatrix} 10 & 9.49 \\ 9.49 & 10 \end{pmatrix}.$$

Aquí podemos ver que la varianza de $\hat{b}_2$ está inflada debido a la alta correlación entre las columnas de X_2. Es 10 veces mayor que la varianza de $\hat{b}_1$ (las columnas de X_1 son independientes).

En el ejemplo anterior vemos que los valores de la diagonal de la matriz $(Z'Z)^{-1}$ nos indican en cuanto aumenta la varianza de las estimaciones de los coeficientes debido a la multicolinealidad. Por esta razón, estos valores toman el nombre de **factores de inflación de varianza (VIFs)** y son uno de los indicadores para el diagnostico de este problema.

Se puede demostrar que el VIF de β_j se puede calcular como:

$$\text{VIF}_j = \frac{1}{1 - R_j^2},$$

donde R_j^2 es el coeficiente de determinación obtenido ajustado una regresión de x_j sobre las demás covariables. Si x_j es casi linealmente dependiente de algunos de los otros regresores, entonces R_j^2 será cercano a uno y el VIF_j será muy alto. Generalmente, un VIF mayor de 10 indica problemas graves de multicolinealidad.

En un capítulo posterior ahondaremos más en este problema.

2.7.1 Bajo peso al nacer - factores de inflación de varianza

En el caso del peso de los recién nacidos, tenemos que los VIFs son:

```
library(car)
vif(mod)
```

```
##     age   mppwt
## 1.06696 1.06696
```

Lo que nos indica que la varianza de las estimaciones de los coeficientes no se inflan debido a multicolinealidad. Recordemos que la correlación entre las dos covariables no es alta (0.2505155).

Chapter 3

Evaluación de los supuestos del modelo

Ejemplo 1. Datos de peso al nacer

Retomemos la base de datos de bajo peso al nacer (disponible en el campus virtual), y consideremos el siguiente modelo:

$$\text{weight}_i = \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{motherage}_i + \beta_3 \text{mnocig}_i + \beta_4 \text{mppwt}_i + \varepsilon_i,$$

con $\varepsilon_i \sim N(0, \sigma^2)$ y $cov(\varepsilon_j, \varepsilon_k) = 0$, para todo $j \neq k$.

El ajuste del modelo es:

```
Birthweight = read.csv('birthweight.csv')
mod.birthweight = lm(weight ~ age + motherage + mnocig + mppwt, data=Birthweight)
summary(mod.birthweight)
```

```
##
## Call:
## lm(formula = weight ~ age + motherage + mnocig + mppwt, data = Birthweight)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.78765 -0.35948  0.09209  0.35024  0.75018
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.104029   1.011723  -4.056 0.000247 ***
## age          0.168027   0.024916   6.744 6.24e-08 ***
## motherage    0.001751   0.012335   0.142 0.887900
## mnocig      -0.014417   0.005421  -2.660 0.011493 *
## mppwt        0.014838   0.009530   1.557 0.127966
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.4073 on 37 degrees of freedom
## Multiple R-squared:  0.627,  Adjusted R-squared:  0.5867
## F-statistic: 15.55 on 4 and 37 DF,  p-value: 1.5e-07
```

La edad gestacional y el número de cigarrillos consumidos por la madre tienen efectos signficativos sobre el peso del recién nacido. El primero es un factor de protección, a mayor edad gestacional mayor será el peso del bebé. Mientras que, el consumo de cigarrillos es un factor de riesgo. A mayor consumo, menor peso tendrá el recién nacido. La edad y el peso de la madre, aunque tienen efectos positivos, no son covariables significativas cuando las otras dos covariables ya están incluidas en el modelo.

Ejemplo 2. Ventas de helados

La base de datos `icecream` (del paquete `orcutt` de R) recopila la siguiente información tomada cada cuatro semanas durante dos años (marzo 1951 a julio 1953):

- **price**: precio promedio del helado (dolares por bote)
- **cons**: consumo medio de helado (botes por persona)
- **temp**: temperatura promedio (en Fahrenheit)

La Figura 3.1 muestra la relación entre las variables. Aquí podemos observar que a mayor temperatura, el consumo de helado se incrementa. Por otro lado, la relación con el precio no es tan fuerte.

```
library(orcutt)
data("icecream")
pairs(icecream[,c(2,1,4)])
```

El objetivo del estudio es explicar el consumo de helado en función del precio y la temperatura. Para esto se propone el siguiente modelo:

$$\text{cons}_i = \beta_0 + \beta_1 \text{price}_i + \beta_2 \text{temp}_i + \varepsilon_i,$$

con $\varepsilon_i \sim N\left(0, \sigma^2\right)$ y $cov(\varepsilon_j, \varepsilon_k) = 0$, para todo $j \neq k$.

El resumen del modelo es el siguiente:

```
  mod.icecream = lm(cons~price+temp,data=icecream)
summary(mod.icecream)
```

```
##
## Call:
## lm(formula = cons ~ price + temp, data = icecream)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.08226 -0.02051  0.00184  0.02272  0.10076
##
## Coefficients:
```

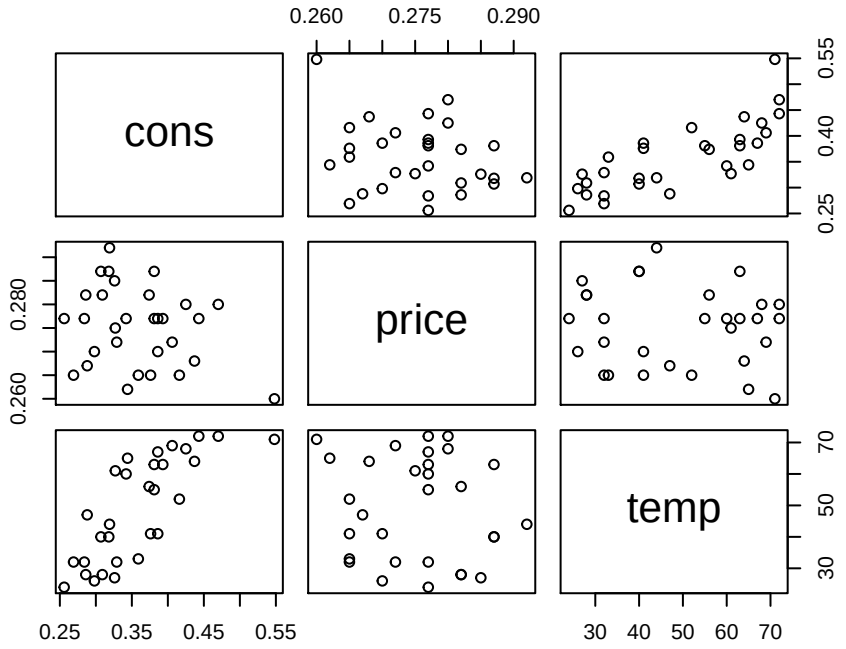

Figura 3.1: Relación entre las variables de los datos de ventas de helados.

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.59655    0.25831   2.309   0.0288 *
## price       -1.40176    0.92509  -1.515   0.1413
## temp         0.00303    0.00047   6.448 6.56e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04132 on 27 degrees of freedom
## Multiple R-squared:  0.6328, Adjusted R-squared:  0.6056
## F-statistic: 23.27 on 2 and 27 DF,  p-value: 1.336e-06
```

De aquí podemos concluir que alrededor del 70% de la variabilidad del consumo de helado está explicado por el modelo propuesto. Además, el efecto de la temperatura sobre el consumo de helado es signficativamente positivo. Aunque la relación con el precio es negativa, esta no es significativa.

Ejemplo 3. Longitud del pez lobina boca chica

La base de datos `wblake` (de la librería `alr4`) contiene la edad (en años) y longitud (en mm) de 439 peces lobina boca chica del lago West Bearskin en el nordeste de Minnesota en 1999. El objetivo del estudio es determinar los patrones de crecimiento de este tipo de pez.

```
library(alr4)
data("wblake")
plot(Length~Age,data=wblake,xlab='edad (años)',ylab='longitud (mm)')
```

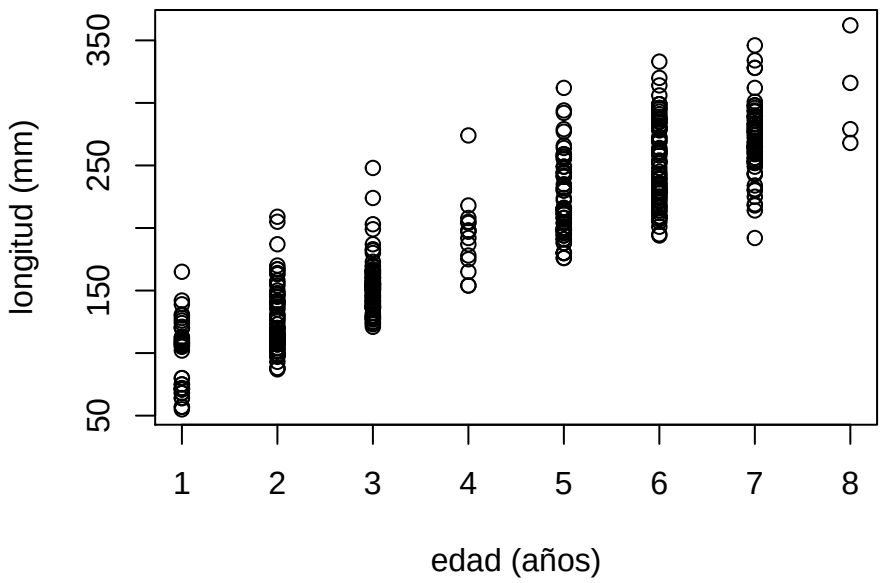

Figura 3.2: Relación entre la edad y la longitud de los peces lobina boca chica.

La Figura 3.2 muestra que la relación entre estas dos variables se puede aproximar a una recta, por lo cuál se propone el siguiente:

$$\text{length}_i = \beta_0 + \beta_1 \text{age}_i + \varepsilon_i$$

con $\varepsilon_i \sim N(0, \sigma^2)$ y $cov(\varepsilon_j, \varepsilon_k) = 0$, para todo $j \neq k$.

El resumen del ajuste es el siguiente:

```
mod.bass = lm(Length~Age,data=wblake)
summary(mod.bass)
```

```
##
## Call:
## lm(formula = Length ~ Age, data = wblake)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -85.794 -19.499  -4.499  16.177  94.853
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  65.5272     3.1974   20.49   <2e-16 ***
## Age          30.3239     0.6877   44.09   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.65 on 437 degrees of freedom
## Multiple R-squared:  0.8165, Adjusted R-squared:  0.8161
## F-statistic:  1944 on 1 and 437 DF,  p-value: < 2.2e-16
```

Estos resultados muestran que la edad del pez tiene un efecto significativamente positivo. Por cada año del pez, la longitud aumenta 30 milimetros en promedio. Adicionalmente, esta covariable explica el 81% de la variabilidad de la longitud.

La validez de las conclusiones hechas en estos ejemplos descansa en el cumplimiento de los supuestos sobre los errores. Por esta razón, es de gran importancia que tengamos herramientas para evaluar si los datos analizados no muestran ningún alejamiento de los supuestos asumidos.

3.1 Supuestos del modelo linea múltiple

En el modelo de regresión lineal múltiple:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_{p-1} x_{i,p-1} + \varepsilon_i,$$

asumimos que la relación entre la variable respuesta y las covariables es lineal, al menos de forma aproximada. Además,

a. $E(\varepsilon_i) = 0$, para $i = 1, ..., n$,
b. $Var(\varepsilon_j) = \sigma^2$. Homogeneidad de varianza en los errores,
c. $Cov(\varepsilon_i, \varepsilon_j) = 0$ para todo $i \neq j$. Los errores están incorrelacionados,
d. $\varepsilon_j \sim Normal(0, \sigma^2)$. Los errores se distribuyen de forma normal.

La importancia de realizar procedimientos para validar los supuestos, radica en que ellos inciden en las cualidades de los estimadores por MCO. En caso de no cumplirse se pueden perder propiedades importantes. Si no se cumple el supuesto (a) se obtienen estimaciones sesgadas. Si no se cumplen (b) y (c) los estimadores MCO pierden la condición de optimalidad. Si no se cumple (d) se pierde eficiencia e imposibilita la aplicación de inferencias basadas en normalidad.

En general, no se puede detectar una violación a los supuestos a partir de estadísticos del ajuste del modelo (R^2, F_0, valores-t, etc). El diagnostico se puede hacer por métodos gráficos y pruebas formales (pruebas de hipótesis). Ambos métodos son complementarios, los gráficos sugieren formas particulares de incumplimiento del supuesto, mientras las pruebas formales evalúan su importancia (Behar, 2002).

3.2 Efectos del incumplimiento de los supuestos

3.2.1 Sesgo por omisión de variables relevantes

Si $E(\varepsilon) = 0$, entonces $E(y|X) = X\beta$, y el estimador por MCO es insesgado. Sin embargo, si omitimos variables relevantes dentro del modelo las estimaciones serán sesgadas. Para ver esto, supongamos que el modelo generador de los datos es:

$$y = X\beta + \varepsilon = X_1\beta_1 + X_2\beta_2 + \varepsilon,$$

donde las columnas de la matriz $n \times p$ de covariables X está divida en dos submatrices X_1 y X_2 de dimensiones $n \times (p-r)$ y $n \times r$, respectivamente. Además, asumimos que $\varepsilon \sim N(0, \sigma^2 I_n)$.

Ahora, consideramos estimar el siguiente modelo:

$$y = X_1\beta_1 + \varepsilon^*,$$

es decir estamos omitiendo las covariables contenidas en X_2. El estimador de β_1 es:

$$\hat{\beta}_1 = (X_1'X_1)^{-1}X_1'y,$$

y el estimador de σ^2 es:

$$\hat{\sigma}_1^2 = \frac{y'(I_n - H_1)y}{n-(p-r)}, \text{ donde } H_1 = X_1(X_1'X_1)^{-1}X_1'.$$

El valor esperado de $\hat{\beta}_1$ es:

$$\begin{aligned} E(\hat{\beta}_1) &= E\left[(X_1'X_1)^{-1}X_1'y\right] = (X_1'X_1)^{-1}X_1'E(y) \\ &= (X_1'X_1)^{-1}X_1'E(X_1\beta_1 + X_2\beta_2 + \varepsilon) = \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2. \end{aligned}$$

Evidentemente, el sesgo de $\hat{\beta}_1$ depende de la magnitud de β_2. Entre más importante sean los efectos asociados a las covariables omitidas (β_2), mayor será el sesgo. Si las columnas de X_1 son ortogonales de las columnas de X_2, tenemos que $X_1'X_2 = 0$. Así que, en este caso particular, $\hat{\beta}_1$ es insesgado (así omitamos las covariables en X_2).

El valor esperado de $\hat{\sigma}_1^2$ es:

$$E(\hat{\sigma}_1^2) = \sigma^2 + \frac{\beta_2'X_2'(I_n - H_1)X_2\beta_2}{n-p_1}.$$

Dado que $(I - H_1)$ es idempotente y, por lo tanto, positiva semi-definida, entonces $E(\hat{\sigma}_1^2) > \sigma^2$. Esto quiere decir que $\hat{\sigma}_1^2$ es un estimador sesgado de σ^2

Ahora veamos el efecto de omitir covariables relevantes sobre las predicciones de y en el punto $x_0 = (x_{01}', x_{02}')'$. Tenemos que:

$$\hat{y}_0 = x_{01}'\hat{\beta}_1 = x_{01}'(X_1'X_1)^{-1}X_1'y.$$

El valor esperado de $\hat{y}_0$ es:

$$E(\hat{y}_0) = x_{01}'E(\hat{\beta}_1) = x_{01}'\left[\beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2\right].$$

Por lo tanto las predicciones también son sesgadas, $E(\hat{y}_0) \neq x_{01}\beta_1 + x_{02}\beta_2$.

3.2.2 Incorrecta matriz de varianzas de los errores

Cuando $V(\varepsilon) = \sigma^2 V$, pero asumimos erroneamente que $V(\varepsilon) = \sigma^2 I_n$, el estimador $\hat{\beta}$ sigue siendo insesgado. Pero, tenemos que:

$$\begin{aligned} V(\hat{\beta}) &= V\left[(X'X)^{-1}X'y\right] = (X'X)^{-1}X'V(y)X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}X'VX(X'X)^{-1}, \end{aligned}$$

es, generalmente, diferente de $\sigma^2(X'X)^{-1}$ (la varianza que asumimos como cierta). Igualmente, el estimador por MCO pierde su condición de optimalidad. Es decir, deja de ser el mejor estimador lineal insesgado.

El estimador de σ^2 es sesgado:

$$E(\hat{\sigma}^2) = \frac{\sigma^2}{n-p} E\left[y'(I_n - H)y\right] = \frac{\sigma^2}{n-p} \mathrm{tr}\left[V(I_n - H)\right].$$

Las predicciones son insesgadas, pero:

$$V(\hat{y}_0) = V(x_0'\hat{\beta}) = \sigma^2 x_0'(X'X)^{-1}X'VX(X'X)^{-1}x_0,$$

que es diferente de $\sigma^2 x_0'(X'X)^{-1}x_0$ (la varianza que asumimos como cierta).

Cuando V es una matriz diagonal:

$$V = \begin{pmatrix} v_{11} & 0 & 0 & ... & 0 \\ 0 & v_{22} & 0 & ... & 0 \\ 0 & 0 & v_{33} & ... & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & ... & v_{nn} \end{pmatrix},$$

tenemos que hay **heterocedasticidad**. Cada error tiene una varianza diferente $V(\varepsilon_j) = \sigma^2 v_{jj}$, pero están incorrelacionados. Si los valores fuera de la diagonal de V son diferentes de cero, entonces los errores están correlacionados.

La correlación de los errores puede esperarse en algunas situaciones. Por ejemplo, si las observaciones son tomadas en el tiempo puede presentarse correlación temporal. En situaciones en las que se pueda garantizar que las observaciones $(y_1, y_2, ..., y_n)$ constituyen una muestra aleatoria, no existirá correlación entre los errores, es decir, que es posible controlar este aspecto, algunas ocasiones, controlando el procedimiento de selección de la muestra.

3.2.3 Distribución no normal de los errores

La normalidad de los errores permite la estimación por intervalos de confianza no sólo para los coeficientes de regresión, sino también para la predicción. Igualmente, permite el planteamiento de pruebas de hipótesis sobre los parámetros del modelo. Cuando los errores no son normales, estas inferencias no son exactas y pueden llegar a ser inválidas.

Sin embargo, el teorema central del límite asegura que, bajo ciertas condiciones muy amplias, la inferencias basadas en el estimador de mínimos cuadrados son aproximadamente válidas si el tamaño de muestra es suficientemente grande. Esto significa que los niveles de las pruebas y cobertura de los intervalos de confianza son aproximadamente correctos.

De la misma forma, los efectos negativos de la no normalidad dependen de que tan alejados estamos de la normalidad. Si la distribución de los errores es parecida a la normal (por ejemplo, t-Student), los efectos negativos no son considerables.

3.3 Residuos del modelo

Los residuos están definidos como:

$$e_i = y_i - \hat{y}_i, \text{ en forma matricial } e = y - \hat{y} = (I_n - H)y.$$

Los residuos representan las desviaciones entre las observaciones y el ajuste. Además, estos son combinaciones lineales de los errores:

$$e = (I_n - H)(X\beta + \varepsilon) = (I_n - H)\varepsilon.$$

Por lo tanto toda desviación de las premisas de los errores se debe reflejar en los residuales. Si $\varepsilon \sim N(0, \sigma^2 I_n)$, entonces:

$$e \sim N\left[0, \sigma^2(I_n - H)\right].$$

De aquí tenemos que $V(e_i) = (1 - h_{ii})\sigma^2$ y $Cov(e_i, e_j) = -h_{ij}\sigma^2$, para todo $i \neq j$. Lo que indica que, aún cuando los errores sean homogéneos en varianza e incorrelacionados, no implica que los residuos lo sean también. Note que los residuos asociados a puntos alejados del centro de los datos tienen menor varianza. Lo que hace difícil detectar violaciones.

Cuando n es grande comparado con el número de parámetros en el modelo, los residuos si reflejan a los errores en cuanto al comportamiento de su varianza y correlación. Esto es porque $|h_{ij}| \leq 1$, $\sum_{i=1}^{n} h_{ii} = n - p$, y $\sum_{i=1}^{n} h_{ij} = \sum_{j=1}^{n} h_{ij} = 1$. Por lo tanto, cuando $n \to \infty$, $V(e_i) = \sigma^2$ y $Cov(e_i, e_j) = 0$.

3.3.1 Residuos estudentizados

Para evitar el inconveniente de la varianza no constante de los residuos, es preferible utilizar los **residuos estudentizados**:

$$r_i = \frac{e_i}{\sqrt{\hat{\sigma}^2(1 - h_{ii})}}, \qquad i = 1, 2, \ldots, n.$$

Entonces, r_i tiene varianza constante ($V(r_i) = 1$) independiente del lugar de x_i.

3.3.2 Residuos PRESS y R-student

Como veremos más adelante, los residuos estudentizados se pueden utilizar detectar puntos atípicos. El problema es que si la i-ésima observación es bastante inusual, el ajuste del modelo puede estar muy influenciado por esta observación. Lo que puede producir un residuo pequeño. Por esta razón, también se pueden calcular los residuos de predicción (PRESS). Estos se calculan de la siguiente forma:

$$e_{(i)} = y_i - \hat{y}_{(i)}, \text{ para } i = 1, \ldots, n,$$

donde $\hat{y}_{(i)}$ es el valor ajustado para la i-ésima observación usando todas las observaciones excepto la i-ésima. Esto implicaría que para calcular los residuos PRESS es necesario ajustar n veces el modelo. Sin embargo, esto no es así, ya que se puede demostrar que:

$$e_{(i)} = \frac{e_i}{1 - h_{ii}}.$$

La varianza de los residuos PRESS es:

$$V(e_{(i)}) = V\left(\frac{e_i}{1 - h_{ii}}\right) = \frac{1}{(1 - h_{ii})^2} V(e_i) = \frac{1}{(1 - h_{ii})^2}[\sigma^2(1 - h_{ii})] = \frac{\sigma^2}{(1 - h_{ii})}.$$

Si estudentizamos los residuos PRESS obtenemos los **residuos R-Student**:

$$t_i = \frac{e_i}{\sqrt{\hat{\sigma}^2_{(i)}(1 - h_{ii})}},$$

donde $\hat{\sigma}^2_{(i)}$ es la estimación de σ usando todas las observaciones excepto la i-ésima. Se puede demostrar que:

$$\hat{\sigma}^2_{(i)} = \frac{(n - p)\hat{\sigma}^2 - \epsilon_i^2/(1 - h_{ii})}{n - p - 1}.$$

3.4 Evaluación del cumplimiento de los supuestos

En esta sección mostramos la evaluación de los supuestos a través del análisis de los residuos del ajuste (ya sean los residuos estudentizados o los R-Student) usando gráficos y pruebas de hipótesis.

3.4.1 Gráficos de residuos

Un gráfico de los residuos es una forma efectiva de investigar posibles alejamientos de los supuestos. Generalmente, se grafican los residuos estudentizados (r_i) contra los valores ajustados $\hat{y}_i$ (o contra alguna de las covariables x_{ij}). Este tipo de gráfico es de gran ayuda para detectar la correcta especificación del modelo y homocedasticidad. Algunos patrones de residuos se pueden observar en la Figura 3.3.

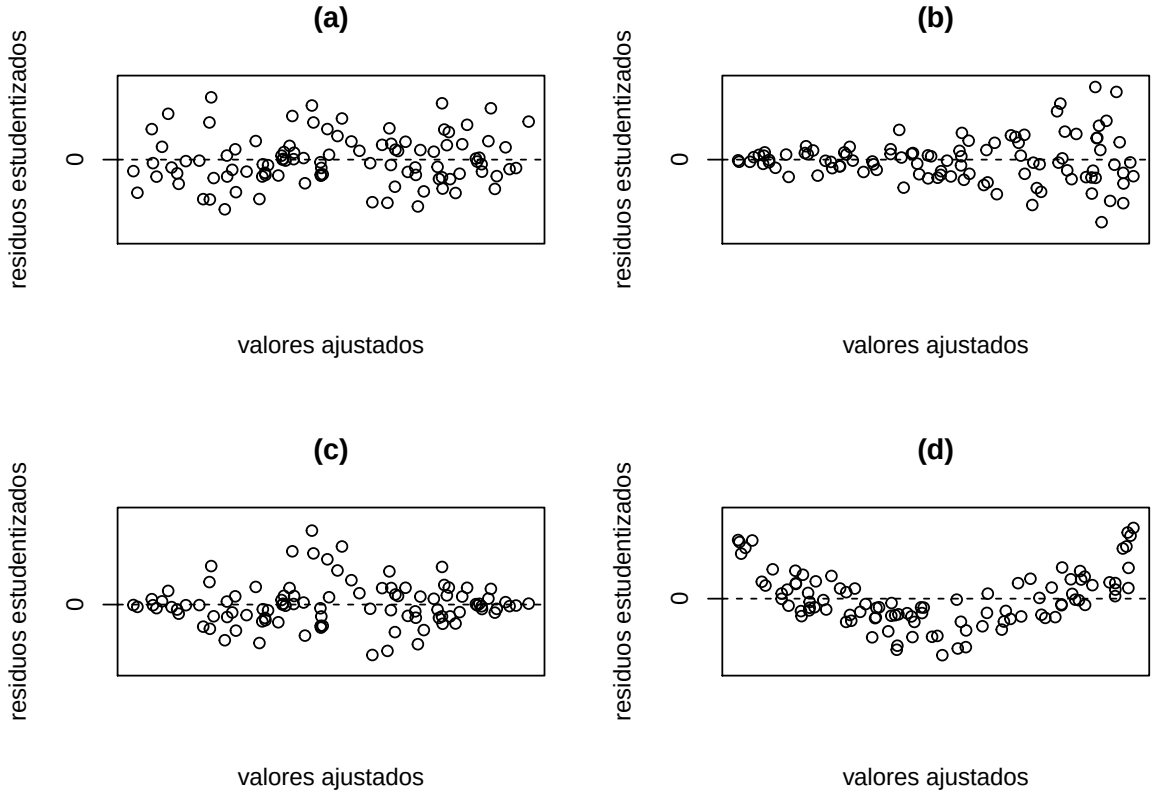

Figura 3.3: Ejemplos de posibles patrones de residuos

La Figura 3.3(a) muestra que los residuos se encuentran alrededor de cero y no se observa ningún patrón claro. Esto es un indicio que el modelo está bien especificado y hay homocedasticidad. En la Figura 3.3(b) vemos que los residuos están alrededor de cero pero la variabilidad crece a medida que los valores ajustados aumenta. Esto es un indicador de heterocedasticidad. La Figura 3.3(c) también muestra un patrón de heterocedasticidad, la variabilidad aumenta hasta cierto punto y luego decrece. En la Figura 3.3(d) observamos que los residuos no fluctuan alrededor de cero, sino que siguen una curva. Esto nos que la relación entre la variable respuesta y las covariables no es lineal.

Adicionalmente, para detectar más facilmente heterocedasticidad, se pueden graficar el valor absoluto de los residuos estudentizados (o al cuadrado) contra los valores ajustados (o las covariables). La Figura 3.4 muestra los mismos patrones pero graficando los residuos en valor absoluto. En las Figuras 3.4(b-c) se evidencia claramente la heterocedasticidad.

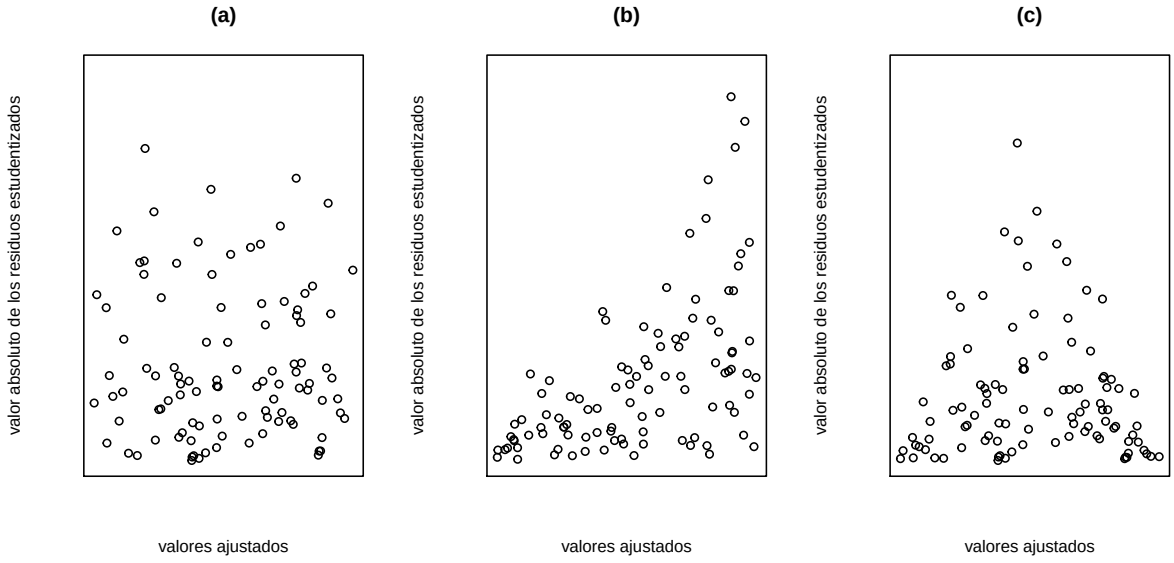

Figura 3.4: Ejemplos de posibles patrones de residuos (graficando el valor absoluto de los residuos).

Bajo peso al nacer - gráfico de residuos

Para el modelo ajustado para los datos de peso al nacer, el gráfico de los residuos contra los valores ajustados se obtienen de la siguiente forma:

```
library(MASS)
res.stud.birthweight = studres(mod.birthweight)
mod.fit.birthweight = mod.birthweight$fitted.values
par(mfrow=c(1,2))
plot(mod.fit.birthweight,res.stud.birthweight, ylab='residuos estudentizados',
     xlab='valores ajustados',main='(a)')
abline(h=0,lty=2)
lines(lowess(res.stud.birthweight~mod.fit.birthweight), col = 2)
plot(mod.fit.birthweight,abs(res.stud.birthweight),
     ylab='valor absoluto de los residuos estudentizados',
     xlab='valores ajustados',main='(b)')
lines(lowess(abs(res.stud.birthweight)~mod.fit.birthweight), col = 2)
```

La Figura @ref(fig:residuosBWdata})(a) muestra que los residuos están alrededor de cero sin mostrar ningún patrón. Note que este gráfico es similar a la Figura 3.3(a). La línea roja es una suaviación LOWESS (Locally weighted scatterplot smoothing, más detalle ver Apéndice A.5 de Weisberg (2014)). Estas suavizaciones permiten ver fácilmente patrones de comportamiento. Aquí vemos que la suavización está cerca de la recta en cero sin mostrar ninguna tendencia o curvatura muy marcada. Por lo tanto, podemos afirmar que la relación entre el peso del recién nacido y las covariables propuesta es lineal. Además, no se observa un problema notorio de heterocedasticidad en ninguno de los dos gráficos de residuos.

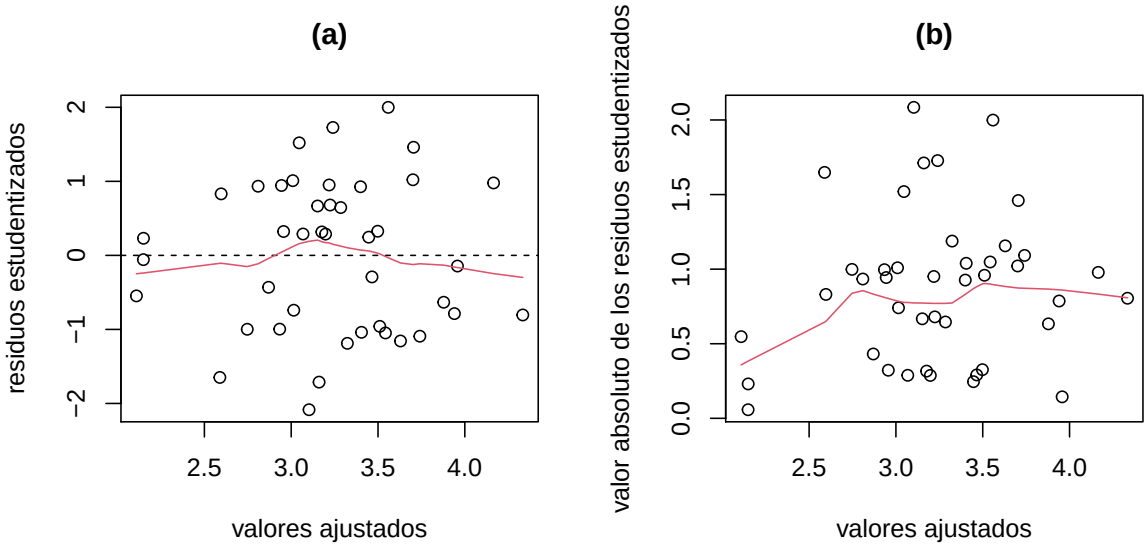

Figura 3.5: Datos de peso al nacer. Gráfico de los residuos estudentizados contra los valores ajustados.

3.4.2 Gráficos de residuos parciales

Los gráficos de residuos parciales permiten estudiar el efecto marginal de una covariable sobre la respuesta condicionado a que los demás regresores ya están en el modelo. En caso que el gráfico de los residuos muestre posibles curvaturas (por ejemplo, Figura 3.3(d)), los residuos parciales permiten verificar si estas se presentan debido a una covariable especifica.

Considere el modelo:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_{p-1} x_{i,p-1} + \varepsilon_i.$$

Para calcular los residuos parciales, primero estimamos los parámetros $(\hat{\beta}_1, ..., \hat{\beta}_{p-1})$ y los residuos ordinarios $(e_1, ..., e_n)$. Luego, los residuos parciales para la covariable x_j se obtienen de la siguiente forma:

$$e_i^*(y|x_j) = e_i - \hat{\beta}_j x_{ij}, \text{ para } i = 1, ..., n.$$

El gráfico de residuos parciales para la covariable x_j se obtiene graficando $e_i^*(y|x_j)$ contra x_j. Si la covariable x_j entra al modelo linealmente, entonces el gráfico de residuos parciales debe mostrar una tendencia lineal. Por el contrario, si se observa una curva, x_j no entra al modelo de forma lineal.

Bajo peso al nacer - gráfico de residuos parciales

El gráfico de residuos parciales se obtiene así:

```
library(car)
par(mfrow=c(2,2))
crPlots(mod.birthweight,'age',xlab='edad gestacional')
crPlots(mod.birthweight,'motherage',xlab='edad de la madre')
crPlots(mod.birthweight,'mnocig',xlab='número de cigarrillos por mes')
crPlots(mod.birthweight,'mppwt',xlab='peso de la madre')
```

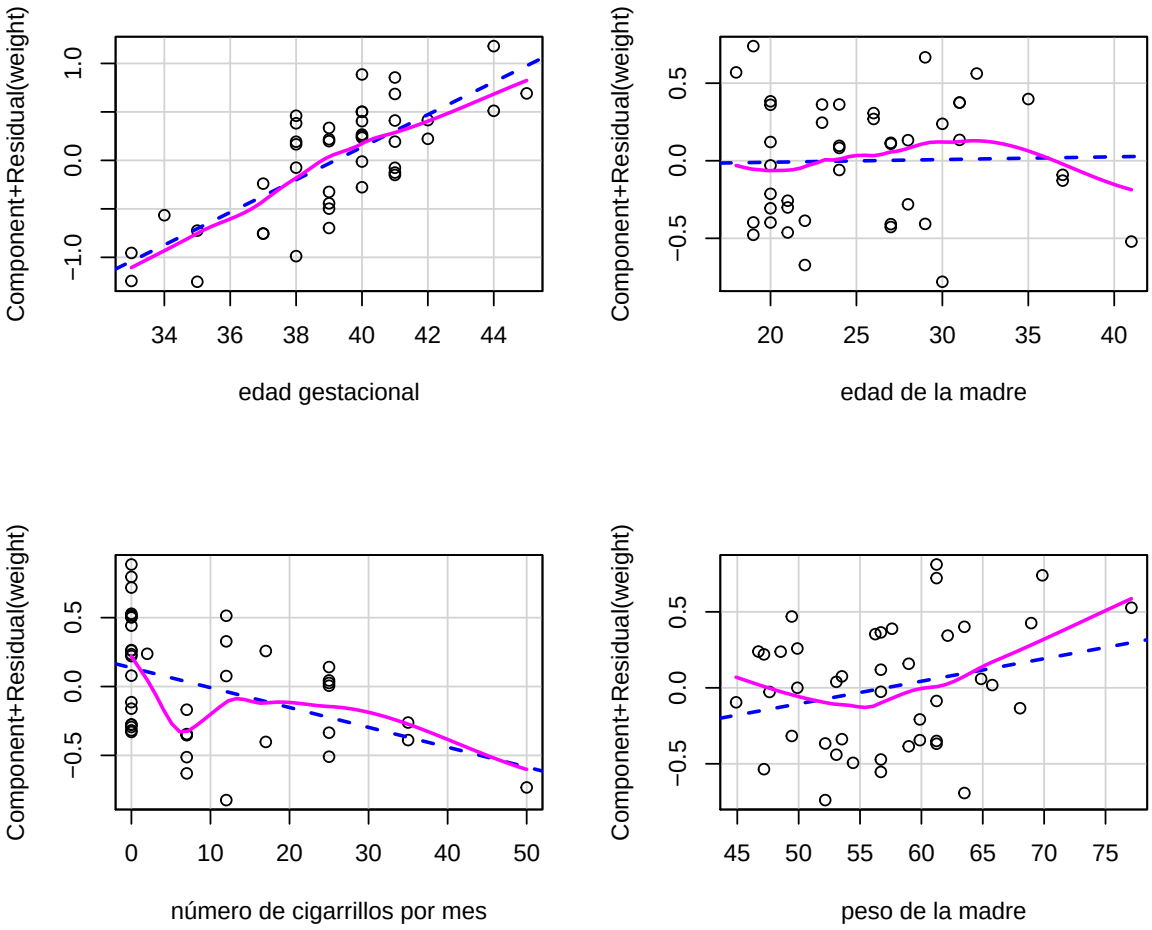

Figura 3.6: Datos de peso al nacer. Gráfico de los residuos parciales para cada covariable.

Igual que en el gráfico de residuos, los residuos parciales no muestran tendencias no lineales muy marcada. Por lo que se puede asumir que la relación entre el peso al nacer y las covariables es lineal.

3.4.3 Gráficos de normalidad

Los gráficos cuantil-cuantil (qqplot) comparan los cuantiles muestrales contra los cuantiles que se esperarían con la distribución de probabilidad asumida para los datos (cuantiles teóricos). En el caso de regresión lineal, estamos asumiendo que los errores del modelo siguen una distribución normal. Por lo tanto, debemos comparar los cuartiles muestrales de los residuos con los cuartiles teóricos que se esperarían bajo una distribución normal.

Sea $(x_1, x_2, \dots, x_n)$ una muestra aleatoría de la variable X con función de distribución desconocida $F_X(x)$, y sean $(x_{[1]}, x_{[2]}, \dots, x_{[n]})$ los estadísticos de orden (observaciones ordenadas de forma creciente). La función empirica de distribución es:

$$S_n(x_{[i]}) = \frac{i}{n} = \frac{\#\text{ de observaciones } \leq x_{[i]}}{n}.$$

Si asumimos que $X \sim N(0,1)$, entonces los puntos $(x_{[i]}, \Phi^{-1}\left\{S_n(x_{[i]})\right\})$, donde $\Phi^{-1}()$ es la inversa de la función acumulativa de una normal estándar, deben seguir aproximadamente una línea recta.

La Figura 3.7 muestra diferentes patrones de gráficos de normalidad para datos generados a partir de tres distribuciones diferentes: normal estándar (derecha), exponencial con $\lambda = 1$ (centro), y t-Student con 2 grados de libertad (derecha). Aquí vemos que para los datos normales, los cuantiles muestrales y teóricos siguen aproximadamente la linea recta de referencia. Mientras que en los otros dos casos, los puntos se alejan en los extremos. En el caso de los datos exponenciales, es al lado izquierdo, mostrando que los datos presentan asímetrica. Mientras que con los datos t-Student, es a ambos lados, indicando que hay muchos valores en las colas (más de los esperados bajo normalidad).

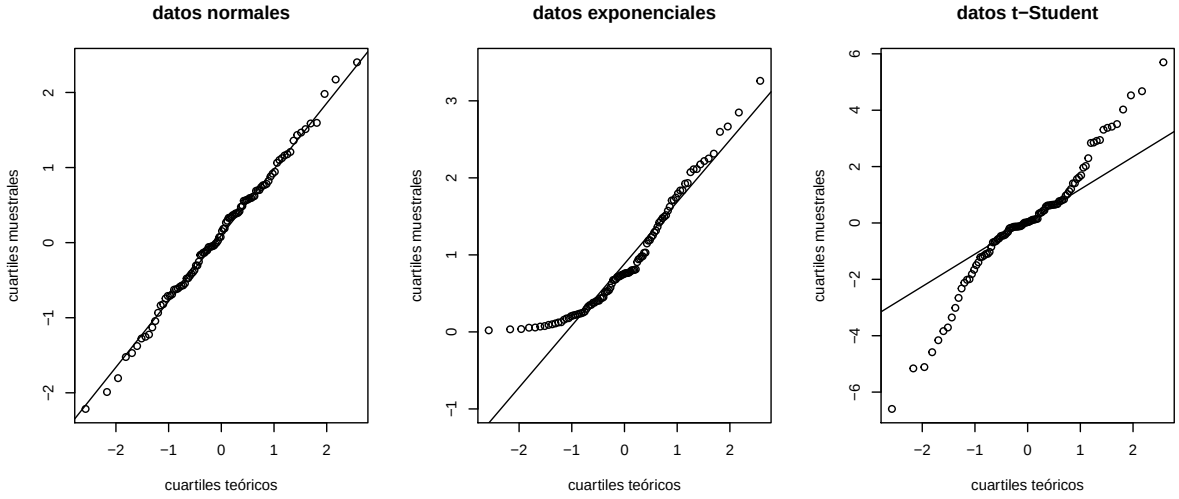

Figura 3.7: Gráficos de normalidad para datos aleatorios generados a partir de una distribución normal estándar (izquierda), exponencial (centro), y t-Student con 2 grados de libertad (derecha).

Bajo peso al nacer - gráfico de normalidad

El gráfico cuartil-cuartil de los residuos estudentizados se obtiene así:

```
car::qqPlot(mod.birthweight,xlab='cuantiles teóricos',ylab='residuos estudentizados',
       distribution = 'norm')
```

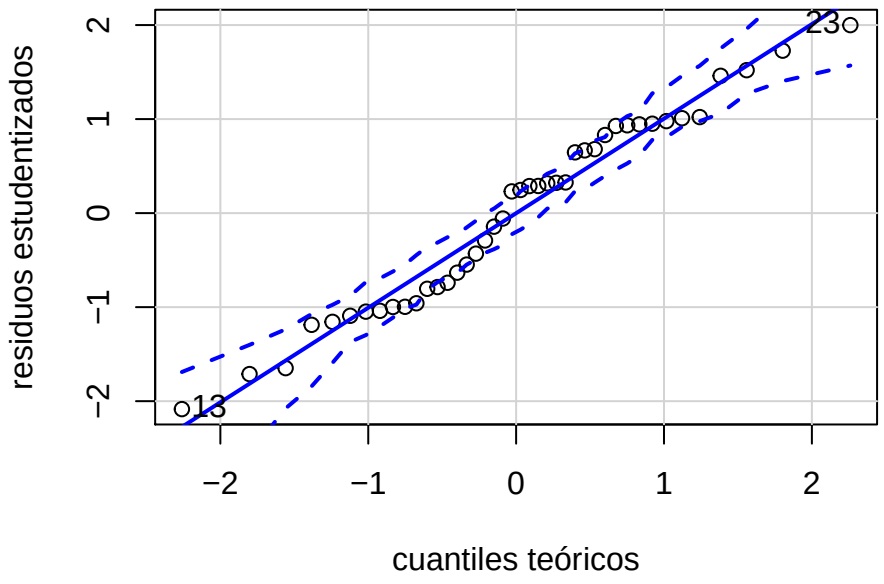

Figura 3.8: Datos de peso al nacer. Gráfico cuantil-cuantil para normalidad.

```
## [1] 13 23
```

Note que la función `qqPlot` incluye intervalos del 95% de confianza para los estadísticos de orden. Para más detalle de como se calculan, ver Sección 12.1.1 de Fox (2016).

La Figura 3.8 muestra que los puntos siguen de forma aproximada una línea recta. Además, la mayoría de los puntos están dentro de las bandas de confianza. Por lo tanto, se estaría cumpliendo el supuesto de normalidad.

3.4.4 Gráfico de residuos frente a el tiempo

Cómo se mencionó anteriormente, si las observaciones fueron tomadas de forma independiente, entonces se puede garantizar que los errores también lo sean. Para los datos del peso de los recién nacidos, tendríamos que asumir que los bebés fueron seleccionados de forma totalmente aleatoria, y así garantizar que los errores no están correlacionados. Por el contrario, los datos del consumo de helado fueron tomados

a lo largo del tiempo (cada dos semanas), por lo que se puede presentar **correlación temporal**. Es decir, observaciones tomadas en tiempo cercanos se espera que estén altamente correlacionadas.

En caso de posible correlación temporal, se puede hacer un gráfico de los residuos con respecto al tiempo. Alternativamente, se puede hacer un gráfico de los residuos rezagados. Es decir, los residuos en el tiempo t (e_t) contra los residuos en el tiempo inmediatamente anterior (e_{t-1}).

La Figura 3.9 muestra diferentes patrones de comportamiento para residuos correlacionados temporalmente. En la columna (a) vemos el comportamiento de residuos incorrelacionados. Estos fluctúan alrededor de cero sin mostrar ningún patrón claro, además el gráfico de residuos rezagados no muestra niguna tendencia. Por el contrario, en las columnas (b) y (c) vemos patrones de comportamiento de residuos correlacionados de forma positiva y negativa, respectivamente. Cuando hay correlación positiva, los valores de los residuos que están cercanos en el tiempo tienden a ser muy similares, además el gráfico de los residuos rezagados muestra una relación positiva entre e_t y e_{t-1}. Cuando la correlación es negativa, vemos que los residuos en el tiempo cambian de signo constantemente, además el gráfico de residuos rezagados muestra una tendencia negativa.

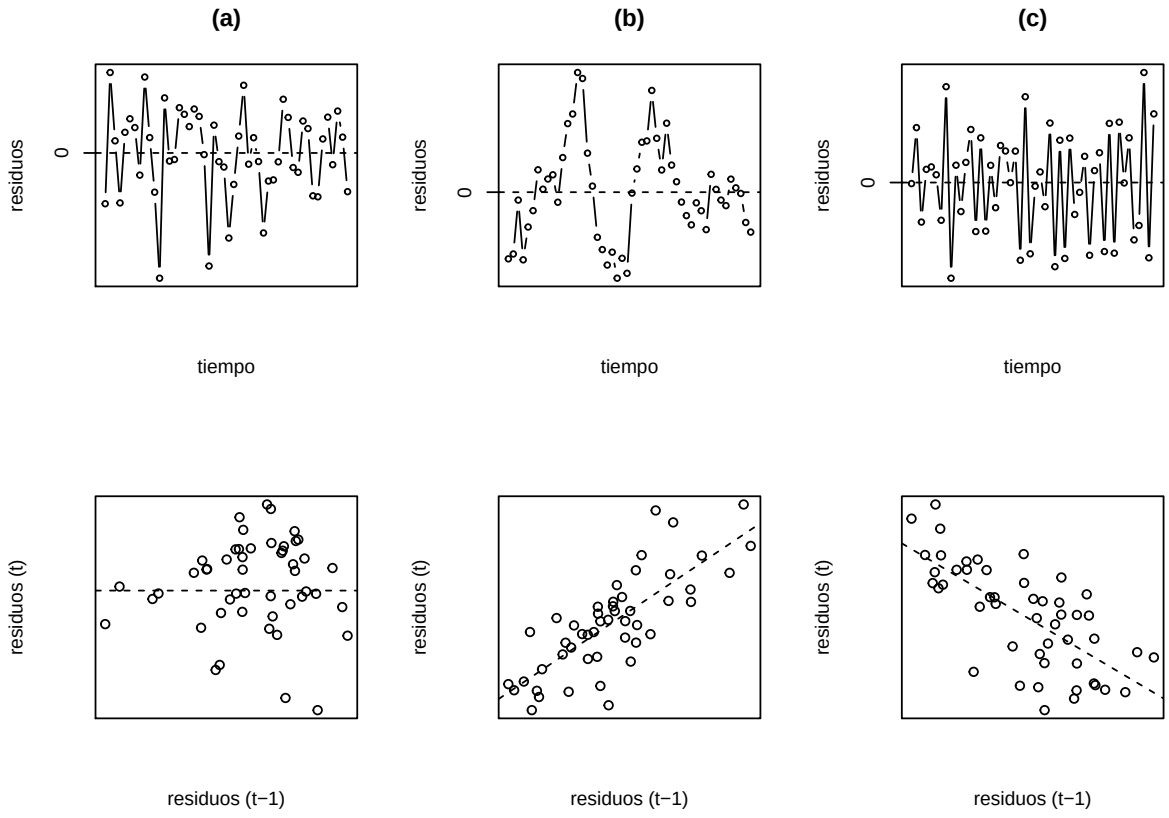

Figura 3.9: Patrones de correlación temporal de los residuos. En la columna (a) los residuos están incorrelacionados, (b) los residuos tienen correlación positiva, y en (c) los residuos tienen correlación negativa.

Consumo de helado - gráfico de los residuos contra el tiempo

Los gráficos de los residuos para el modelo ajustado a los datos de consumo de helado se observan en la Figura 3.10. Aquí vemos que la relación entre la variable respuesta y las covariables es aproximadamente lineal, no hay problemas de heterocedasticidad, y que los residuos siguen una distribución normal. Aunque se ve la presencia de un punto atípico.

```
res.stud.icecream = studres(mod.icecream)
mod.fit.icecream = mod.icecream$fitted.values
par(mfrow=c(1,2))
plot(mod.fit.icecream,res.stud.icecream, ylab='residuos estudentizados',
     xlab='valores ajustados')
abline(h=0,lty=2)
lines(lowess(res.stud.icecream~mod.fit.icecream), col = 2)
car::qqPlot(mod.icecream,xlab='cuantiles teóricos',ylab='residuos estudentizados',
       distribution = 'norm')
```

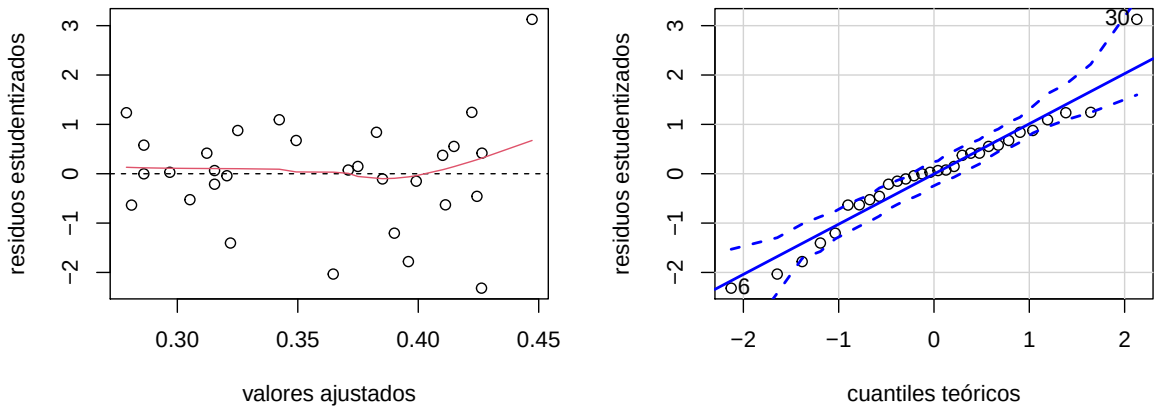

Figura 3.10: Datos de consumo de helado. Gráfico de los residuos estudentizados contra los valores ajustados (izquierda) y gráfico de normalidad (derecha).

```
## [1]  6 30
```

Dado que las observaciones del consumo de helado tienen un orden temporal (fueron tomadas cada 4 semanas), se pueden graficar los residuos contra el tiempo y de los residuos rezagados. Estos se pueden observar en la Figura 3.11. Aquí podemos ver que hay una correlación temporal positiva. Además, se tiene que $cor(e_t, e_{t-1}) = 0.605$.

```
par(mfrow=c(1,2))
plot(res.stud.icecream, ylab='residuos estudentizados',
     xlab='tiempo',type='b')
abline(h=0,lty=2)
plot(res.stud.icecream[-30],res.stud.icecream[-1], ylab='residuos estudentizados (t)',
     xlab='residuos estudentizados (t-1)')
abline(lm(res.stud.icecream[-1] ~ res.stud.icecream[-30]),lty=2)
```

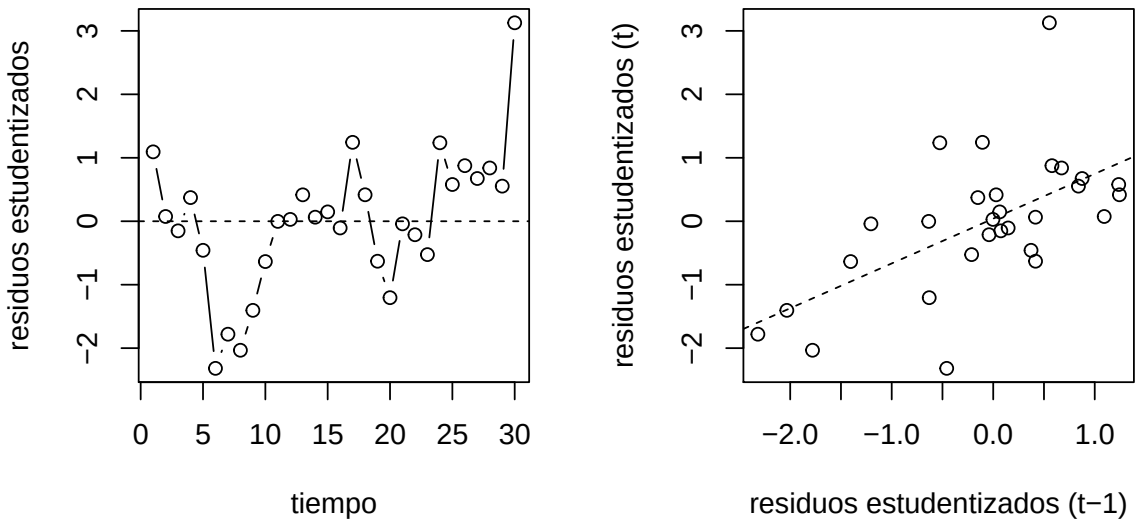

Figura 3.11: Datos de consumo de helado. Gráfico de los residuos estudentizados contra el tiempo (izquierda) y gráfico de residuos rezagados (derecha).

3.5 Pruebas de hipótesis para evaluar los supuestos

3.5.1 Prueba de falta de ajuste

El objetivo de la prueba de falta de ajuste es determinar si la relación entre la variable respuesta y las covariables puede asumirse cómo lineal. Hay que tener en cuenta que esta prueba es sensible a ajamientos de los supuestos de normalidad, varianza constante e independencia de los errores. Además, requiere que se tengan mútiple observaciones de y para diferentes niveles de x. Por ejemplo, el test puede implementarse en los datos de la longitud de los peces puesto que tenemos varios individuos con las mismas edades. En los otros casos, no es posible, por lo menos de forma exacta. El objetivo de estas observaciones replicadas es tener una estimación independiente de σ^2.

En regresión lineal simple, suponga que se tienen n_i observaciones de la variable respuesta para el i-ésimo nivel de x_i, para $i = 1, \ldots, m$, y denotemos y_{ij} como la j-ésima observación de la respuesta en x_i, para $j = 1, \ldots, n_i$. Por lo tanto, tenemos $n = \sum_{i=1}^{m} n_i$ observaciones en total. Esto permite que se puede descomponer la suma de cuadrados de los residuos en dos:

$$SS_{res} = SS_{PE} + SS_{LOF},$$

donde SS_{PE} y SS_{LOF} son la suma de cuadrados del error puro y falta de ajuste, respectivamente.

Para esto, primero observemos que los residuos se pueden descomponer así:

$$y_{ij} - \widehat{y}_i = (y_{ij} - \bar{y}_i) + (\bar{y}_i - \widehat{y}_i), \tag{3.1}$$

donde $\bar{y}_i$ es el promedio de las n_i observaciones en x_i. Ahora, elevando al cuadrado ambos lados de (3.1), y sumando para todo i y j, tenemos:

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-\hat{y}_i)^2 = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2 + \sum_{i=1}^{m} n_i(\bar{y}_i-\hat{y}_i)^2,$$
$$SS_{res} = SS_{PE} + SS_{LOF}.$$

Note que la suma de cuadrados de la falta de ajuste es una suma ponderada de las diferencias entre el promedio de las observaciones en cada nivel de x y el correspondiente valor ajustado. Por lo tanto, si la relación entre las variables es aproximadamente lineal, entonces se espera que SS_{LOF} sea cercana a cero.

Los grados de libertad de SS_{PE} y SS_{LOF} son $\sum_{i=1}^{m}(n_i-1) = n-m$ y $m-2$, respectivamente. De aquí se define el cuadrado medio del error puro y el cuadrado medio de la falta de ajuste:

$$MS_{PE} = \frac{SS_{PE}}{n-m} \text{ y } MS_{LOF} = \frac{SS_{LOF}}{m-2},$$

respectivamente. Si se cumple el supuesto de homocedasticidad, el valor esperado de MS_{PE} es:

$$E(MS_{PE}) = \frac{1}{n-m}E\left[\sum_{j=1}^{n_i}(y_{ij}-\hat{y}_i)^2\right] = \frac{1}{n-m}\sum_{j=1}^{n_i}E\left[(y_{ij}-\hat{y}_i)^2\right] = \frac{\sigma^2}{n-m}\sum_{j=1}^{n_i}(n_i-1) = \sigma^2.$$

Es decir que el cuadrado medio del error puro es una estimación de σ^2 independiente del ajuste del modelo.

Además, el valor esperado del cuadrado medio de la falta de ajuste es:

$$E(MS_{LOF}) = \sigma^2 + \frac{\sum_{i=1}^{m} n_i\left[E(y_i|x_i)-\beta_0-\beta_1 x_i\right]^2}{m-2}.$$

Por lo tanto, si la función de la media es lineal, entonces $E(y_i|x_i) = \beta_0+\beta_1 x_i$, y $E(MS_{LOF}) = \sigma^2$. Si la función media no es lineal, entonces $E(MS_{LOF}) > \sigma^2$

La prueba de falta de ajuste plantea las siguiente hipótesis:

H_0 : el lineal modelo proporciona buen ajuste $\quad$ H_1 : el modelo lineal no proporciona buen ajuste,

El estadístico de prueba es:

$$F_0 = \frac{SS_{LOF}/(m-2)}{SS_{PE}/(n-m)} = \frac{MS_{LOF}}{MS_{PE}}.$$

Si H_0 es cierta, F_0 sigue una distribución $F_{m-2,n-m}$. Por lo tanto, se rechaza H_0 (es decir, la función de regresión no es lineal) si $F_0 > F_{1-\alpha,m-2,n-m}$.

Como se mencionó antes, la prueba exige que se tenga múltiples observaciones para cada nivel de x. Lo cuál es un gran desventaja. En caso que esto no ocurra, se puede hacer agrupaciones de los valores de las covariables (vecinos más cercanos), y considerarlas como repeticiones; de esta manera la prueba es aproximada.

3.5.1.1 Base de datos de la longitud de los peces - prueba de falta de ajuste

La Figura 3.12 muestra los gráficos de los residuos para el modelo ajustado a los datos de la longitud de los peces.

```
res.stud.bass = studres(mod.bass)
mod.fit.bass = mod.bass$fitted.values
plot(mod.fit.bass,res.stud.bass, ylab='residuos estudentizados',
     xlab='valores ajustados')
abline(h=0,lty=2)
lines(lowess(res.stud.bass~mod.fit.bass), col = 2)
```

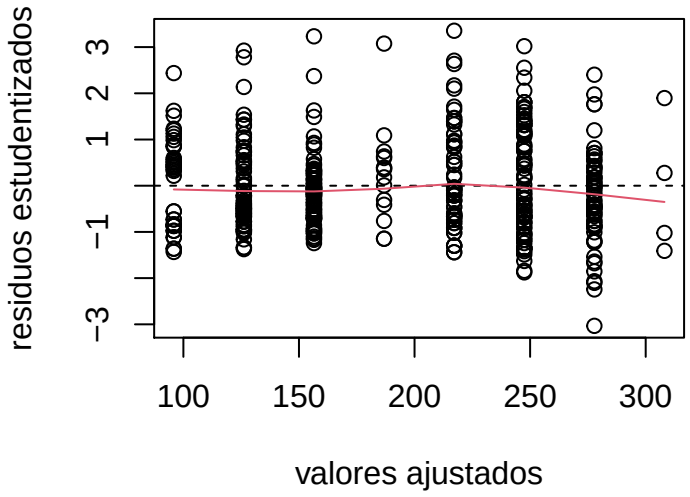

Figura 3.12: Datos de la longitud de los peces. Gráfico de los residuos estudentizados contra los valores ajustados.

La prueba de falta se ajuste se puede hacer usando la función `anovaPE` de la librería `EnvStats`:

```
library(EnvStats)
anovaPE(mod.bass)
```

```
##                 Df  Sum Sq Mean Sq   F value Pr(>F)
## Age              1 1595359 1595359 1973.0736 <2e-16 ***
## Lack of Fit      6   10104    1684    2.0827 0.0541 .
## Pure Error     431  348492     809
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A partir del valor-p podemos concluir que no se rechaza H_0, por lo tanto no hay suficiente evidencia para dudar del ajuste lineal propuesto.

3.5.2 Prueba de heterocedasticidad

Algunas pruebas de heterocedasticidad asumen que la varianza de los errores se compone de una parte constante y otra que varía según unas variables (z):

$$\sigma_i^2 = f(\sigma^2, z_i),$$

donde σ^2 es la parte fija de la varianza, z_i el conjunto de variables cuyos valores se asocian con los cambios en la varianza de los errores. Por lo general se asume que la función de varianza depende de algunas de las covariables del modelo, es decir que $z_i = x_i$.

La **prueba de Breusch-Pagan** asume que la varianza es una función aditiva de las covariables:

$$\sigma_i^2 = E(\varepsilon_i^2) = \gamma_0 + \gamma_1 x_{i1} + \gamma_1 x_{i2} + ... + \gamma_{p-1} x_{i,p-1}.$$

Por lo tanto, se pueden plantear las siguientes hipótesis:

$$\begin{aligned} H_0 &: \gamma_1 = \gamma_2 = ... = \gamma_{p-1} = 0 \qquad \text{(homocedasticidad)}, \\ H_1 &: \gamma_j \neq 0 \text{ para algún } j = 1, ..., p-1 \text{ (heterocedasticidad)}. \end{aligned}$$

Dado que los errores no son observables, el estadístico de prueba se construye a partir de los residuos del modelo estimado. Primero, se ajusta el siguiente modelo de regresión:

$$e_i^2 = \gamma_0 + \gamma_1 x_{i1} + \gamma_1 x_{i2} + ... + \gamma_{p-1} x_{i,p-1} + \upsilon_i,$$

donde se asume que $\upsilon_i \sim N(0, \sigma_\upsilon^2)$, y se obtiene el coeficiente de determinación R_e^2.

si H_0 es cierta, se tiene que $W = nR_e^2 \sim \chi_{p-1}^2$. Entonces, si $W > \chi_{1-\alpha,p-1}^2$ se rechaza H_0. Por lo tanto, hay heterocedasticidad.

El test de Breusch-Pagan sólo detecta formas lineales de heterocedasticidad. La **prueba de White** propone que la relación entre la varianza y las covariables es cuadrática:

$$\begin{aligned} \sigma_i^2 &= \left(\gamma_0^* + \sum_{j=1}^{p-1} \gamma_j^* x_{ij} \right)^2 \\ &= \gamma_0 + \sum_{j=1}^{p-1} \gamma_j x_{ij} + \sum_{j=1}^{p-1} \gamma_{jj} x_{ij}^2 + \sum_{j=1}^{p-1} \sum_{k \neq j} \gamma_{jk} x_{ij} x_{ik}. \end{aligned}$$

Dado que el número de parámetros del modelo se incrementa rápidamente a medida que tengamos más covariables, se pueden omitir las interacciones entre las covariables.

Por ejemplo si se tiene un modelo con tres covariables, se plantea la siguiente función para la varianza:

$$\sigma_i^2 = \gamma_0 + \gamma_1 x_{i1} + \gamma_1 x_{i2} + \gamma_1 x_{i3} + \gamma_4 x_{i1}^2 + \gamma_5 x_{i2}^2 + \gamma_6 x_{i3}^2 + \gamma_7 x_{i1} x_{i2} + \gamma_8 x_{i1} x_{i3} + \gamma_9 x_{i2} x_{i3}$$

y las hipótesis son:

$$\begin{aligned} H_0 &: \gamma_1 = \gamma_2 = ... = \gamma_9 = 0 \qquad \text{(homocedasticidad)}, \\ H_1 &: \gamma_j \neq 0 \text{ para algún } j = 1, ..., 9 \text{ (heterocedasticidad)}. \end{aligned}$$

Para calcular el estadístico de prueba, primero ajustamos el siguiente modelo auxiliar:

$$e_i^2 = \gamma_0 + \gamma_1 x_{i1} + \gamma_1 x_{i2} + \gamma_1 x_{i3} + \gamma_4 x_{i1}^2 + \gamma_5 x_{i2}^2 + \gamma_6 x_{i3}^2 + \gamma_7 x_{i1} x_{i2} + \gamma_8 x_{i1} x_{i3} + \gamma_9 x_{i2} x_{i3} + \upsilon_i,$$

y el coeficiente de determinación asociado R_e^2. El estadístico de prueba es $W = nR_e^2$, y rechazamos H_0 si $W > \chi_{1-\alpha,8}^2$.

Bajo peso al nacer - prueba de heterocedasticidad

En la prueba de White se asume que:

$$\sigma_i^2 = \gamma_0 + \gamma_1 \text{age}_i + \gamma_2 \text{motherage}_i + \gamma_3 \text{mnocig}_i + \gamma_4 \text{mppwt}_i + \gamma_5 \text{age}_i^2 + \gamma_6 \text{motherage}_i^2 + \gamma_7 \text{mnocig}_i^2 + \gamma_8 \text{mppwt}_i^2.$$

Dado que se tienen varias covariables, se omitieron las interacciones entre las covariables. Se plantean las siguientes hipótesis:

$$H_0 : \gamma_1 = \gamma_2 = ... = \gamma_8 = 0 \qquad \text{(homocedasticidad)},$$
$$H_1 : \gamma_j \neq 0 \text{ para algún } j = 1, ..., 8 \text{ (heterocedasticidad)}.$$

El ajuste del modelo auxiliar es:

```
res.stud.birthweight=mod.birthweight$residuals
mod.res.birthweight = lm(res.stud.birthweight^2 ~ age + motherage + mnocig + mppwt +
                           I(age^2) + I(motherage^2) + I(mnocig^2) + I(mppwt^2),
                         data=birthweight)
summary(mod.res.birthweight)
```

```
##
## Call:
## lm(formula = res.stud.birthweight^2 ~ age + motherage + mnocig +
##     mppwt + I(age^2) + I(motherage^2) + I(mnocig^2) + I(mppwt^2),
##     data = birthweight)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.20691 -0.08384 -0.01338  0.02286  0.41575
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -6.1348535  3.4447222  -1.781   0.0841 .
## age             0.2877513  0.1701394   1.691   0.1002
## motherage      -0.0406689  0.0359059  -1.133   0.2655
## mnocig          0.0024339  0.0051823   0.470   0.6417
## mppwt           0.0433585  0.0435969   0.995   0.3272
## I(age^2)       -0.0037481  0.0021924  -1.710   0.0967 .
## I(motherage^2)  0.0007444  0.0006445   1.155   0.2563
## I(mnocig^2)    -0.0001672  0.0001335  -1.253   0.2192
## I(mppwt^2)     -0.0003450  0.0003688  -0.935   0.3564
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1457 on 33 degrees of freedom
## Multiple R-squared:  0.2397, Adjusted R-squared:  0.05544
## F-statistic: 1.301 on 8 and 33 DF,  p-value: 0.2772
```

Note que el estadístico F_0 es pequeño y que las pruebas individuales sobre los coeficientes no son significativas. Esto es un indicio que H_0 es cierta. El coeficiente de determinación es $R_e^2 = 0.24$. Entonces, $W = 10.069$, con un valor-p asociado de 0.2602. Por lo tanto, no tenemos evidencia suficiente para determinar que hay heterocedasticidad.

La prueba se puede implementar directamente usando la función `bptest` de la librería `lmtest` llegando a los mismos resultados:

```
bptest(mod.birthweight, ~ age + motherage + mnocig + mppwt +
                          I(age^2) + I(motherage^2) + I(mnocig^2) + I(mppwt^2),
                        data=birthweight)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  mod.birthweight
## BP = 10.069, df = 8, p-value = 0.2602
```

Otras pruebas de heterocedasticidad que se pueden utilizar son:

- Prueba de Goldfeld–Quandt.
- Prueba de Barlett.
- Prueba de Cochran.
- Prueba de Hartley.

Las últimas tres pruebas requieren que se tengan múltiples observaciones para cada nivel de x. En caso que no se tengan repeticiones, es posible agruparlas por los vecinos más cercanos e implementar la prueba de forma aproximada.

3.5.3 Prueba de normalidad

Para probar normalidad podemos utilizar la prueba de Shapiro-Wilks.

Suponga que se tiene una muestra aleatoria $x_1, \dots, x_n$ que se asumen sigue una distribución normal. Por lo cuál, se plantean las siguientes hipótesis:

$$H_0 : \text{la distribución de } X \text{ es normal} \qquad H_0 : \text{la distribución de } X \text{ no es normal}$$

El estadístico de prueba propuestos por Shapiro y Wilks es:

$$W = \frac{\sum_{i=1}^{[n/2]} a_{in} \left(x_{[n-i+1]} - x_{[i]} \right)}{\sum_{i=1}^{n} (x_i - \bar{x})^2},$$

donde $(x_{[1]}, x_{[2]}, \dots, x_{[n]})$ son los estadísticos de orden y los valores a_{in}, así como los valores críticos, están dados en tablas tabuladas por los autores.

Bajo peso al nacer - prueba de normalidad

La prueba de Shapiro-Wilks para ajuste de los datos de peso al nacer es: `lmtest`:

```
shapiro.test(res.stud.birthweight)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res.stud.birthweight
## W = 0.96593, p-value = 0.2395
```

A partir de este resultado, no tenemos evidencia suficiente para rechazar que los errores se distribuyen normal. En esta función de R, el valor-p es calculado usando una aproximación propuesta por Royston (1995).

Otras pruebas de normalidad que se pueden utilizar son:

- Modificaciones de la prueba Shapiro-Wilks, como son: D'agostino o Shapiro-Francia.
- Pruebas de bondad de ajuste generales, como son: Kolmogorov-Smirnov, Cramer-Von Mises, Anderson-Darling.

3.5.4 Prueba de correlación temporal de los errores

Cuando la correlación es debido a que las observaciones fueron tomadas en el tiempo, se puede asumir que hay **autocorrelación**. Aquí se asume que los errores que están separados t unidades de tiempo siempre tienen la misma correlación lineal. Además que la correlación disminuye a medida que las observaciones se separan en el tiempo.

El modelo de regresión, con errores autoregresivos de orden uno, es el siguiente:

$$y_t = x_t'\beta + \varepsilon_t, \text{ con } \varepsilon_t = \phi\varepsilon_{t-1} + a_t,$$

donde y_t y x_t son la variable respuesta observada y el conjunto de covariables observadas en el tiempo t, respectivamente, y ϕ es el parámetro de autocorrelación ($|\phi| < 1$). Además, se asume que $a_t \sim N(0, \sigma_a^2)$ y $cov(a_j, a_k) = 0$, para todo $j \neq k$. A partir de estos resultados, se tiene que:

$$E(\varepsilon_t) = 0, V(\varepsilon_t) = \sigma_a^2\left(\frac{1}{1-\phi^2}\right), \text{ y } cov(\varepsilon_t, \varepsilon_{t\pm k}) = \phi^k\sigma_a^2\left(\frac{1}{1-\phi^2}\right).$$

Por lo tanto, la correlación entre dos errores separados en k periodos de tiempo es $cor(\varepsilon_t, \varepsilon_{t\pm k}) = \phi^k$. Si $\phi > 0$, los errores están correlacionados positivamente, pero la magnitud de la correlación disminuye a medida que los errores se separan más. Por otro lado, si $\phi = 0$ los errores están incorrelacionados.

La prueba de Durbin-Watson plantea las siguientes hipótesis:

$$H_0 : \phi = 0 \text{ (independencia)} \qquad H_1 : \phi \neq 0 \text{ (autocorrelación)}$$

El estadístico de prueba es:

$$d = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}.$$

La distribución de probabilidad de d, bajo H_0, depende de la estructura de X y es difícil de determinar. Por lo tanto, los valores críticos están tabulados para diferentes valores de significancia, tamaño de muestra y número de parámetros. Otra alternativa, usada por los paquetes estadísticos, es calcular la significancia a través de métodos de remuestreo y aproximaciones del estadístico de prueba a la distribución normal.

Ventas de helado - prueba de correlación temporal

La prueba de Durbin-Watson para ajuste de los datos de ventas de helado se puede hacer a través de la función `durbinWatsonTest` de la librería `car`:

```
durbinWatsonTest(mod.icecream,method='resample',reps=1000)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.5412411      0.655856       0
##  Alternative hypothesis: rho != 0
```

Estos resultados muestran que hay correlación serial en los datos. Note que con esta función, el valor-p es calculado a partir de técnicas de remuestreo usando 1000 repeticiones.

3.6 Comentarios finales

Algunas consideraciones, muchas de ellas tomadas de Behar (2002):

- La efectividad de las pruebas formales depende del tamaño de la muestra. Si n es pequeño, la potencia es baja. Por lo tanto, es difícil detectar alejamientos de la hipótesis nula. Si por el contrario, la muestra es grande, la potencia es alta. Entonces, se rechaza la hipótesis nula ante cualquier alejamiento ligero.
- El incumplimiento de un supuesto puede reflejarse como el incumplimiento de otros. Por ejemplo, la falta de ajuste del modelo puede reflejarse como heterogeneidad de los errores y/o como correlación de los mismos.
- Hay que tener en cuenta que algunas pruebas de hipótesis suponen cierto alejamiento particular del supuesto que se quiere probar. Por ejemplo, el test de White asume que la varianza es una función cuadrática de las covariables. Por lo tanto si rechazamos esta prueba, no necesariamente podemos asegurar con total certeza que no hay heterocedasticidad. Es posible que la función de varianza tome otra forma.
- Adicionalmente, varias pruebas son muy sensibles al alejamiento de la suposición de normalidad. Es decir que, si los errores no son normalmente distribuidos, el nivel real de significancia puede ser muy diferente del especificado. Sin embargo, el rechazo de la hipótesis nula podría sugerir que al menos uno de los dos supuestos no se cumple.
- A la hora de validar el supuesto de normalidad de los errores se está interesado en saber si el alejamiento de ese modelo normal, es aceptable desde el punto de vista de la conservación de las propiedades y ventajas que se heredan de la normalidad. Las estimaciones de $\widehat{\beta}$ son generalmente robustas a desviaciones de la normalidad (Teorema del límite central).

Chapter 4

Transformaciones y mínimos cuadrados ponderados

Ejemplo 1. Datos de la ONU

La base de datos `UN11` de la librería `alr4` contiene las siguientes estadísticas de varios miembros de las naciones unidas (y otras regiones independientes) durante los años 2009-2011:

- **fertility**: Número esperado de nacidos vivos por mujer.
- **ppgdp**: producto nacional bruto per cápita (PNB, en dólares).
- **Purban**: el porcentaje de la población que vive en un área urbana.
- **lifeExpF**: esperanza de vida femenina (años).

El objetivo del estudio es ver la relación entre la fertilidad con las otras variables. Por ahora, empecemos con un modelo de la fertilidad en función del producto nacional bruto y el porcentaje de población urbana.

La Figura 4.1 muestra la relación entre las variables. Aquí vemos que ambas covariables tienen una relación negativa con la fertilidad. Note, además, que la relación con el producto nacional bruto no es lineal. Esto último podría traer problemas a la hora de ajustar un modelo lineal.

```
library(alr4)
data("UN11")
pairs(UN11[,-c(1:2,5)])
```

Por ahora consideremos solamente el PNB y el % de población en área urbana como covariables. Por lo tanto, el modelo propuesto es el siguiente:

$$\text{fertility}_i = \beta_0 + \beta_1 \text{ppgdp}_i + \beta_2 \text{pctUrban}_i + \varepsilon_i,$$

donde $\varepsilon_i \sim N(0, \sigma^2)$ y $cov(\varepsilon_j, \varepsilon_k) = 0$.

Luego de ajustar el modelo se procede a hacer un análisis de residuos. El gráfico de los residuos estudentizados (Figura 4.2 muestra que el ajuste presenta problemas de no linealidad y heterocedasticidad. En la Figura 4.3 de los residuos parciales podemos observar que estos problemas se deben a la covariable PNB.

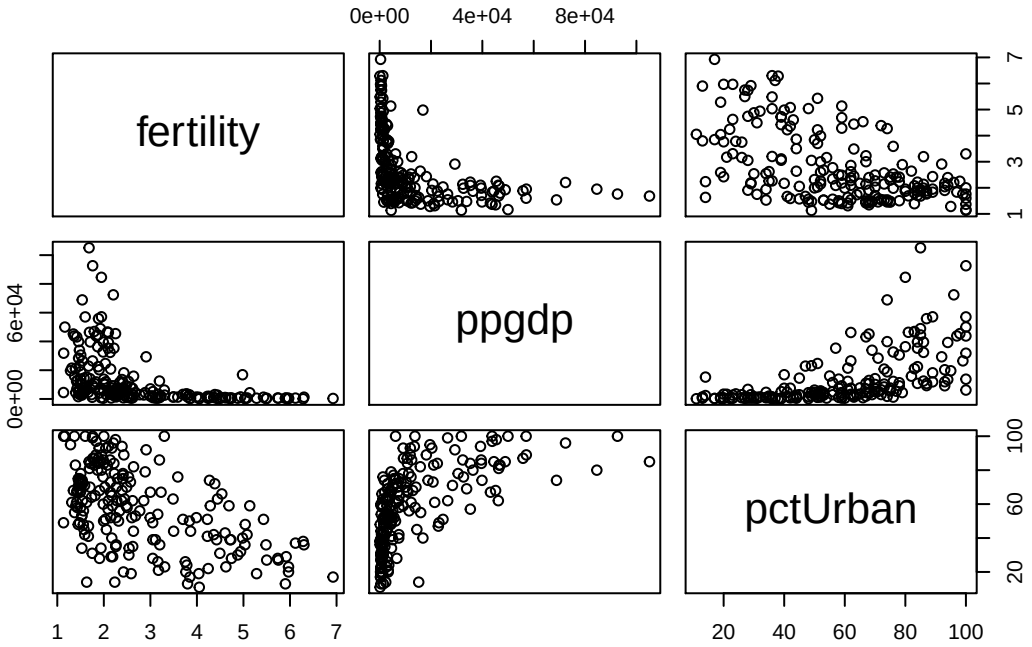

Figura 4.1: Datos de la ONU. Relación entre las variables.

```
mod.UN11 = lm(fertility~ppgdp+pctUrban,data=UN11)
library(MASS)
res.UN11 = studres(mod.UN11)
plot(mod.UN11$fitted.values,res.UN11,
     xlab='valores ajustados',ylab='residuos estudentizados')
lines(lowess(res.UN11~mod.UN11$fitted.values),col=2)
abline(h=0,lty=2)
```

```
library(car)
crPlots(mod.UN11,main='')
```

Ejemplo 2. Datos de educación

La base de datos `education` de la librería `robustbase` contiene información sobre gastos en educación de 50 estados de los EEUU en el año 1975. Las variables observadas son:

- **Y**: gasto per cápita en educación pública (dólares, proyectado para 1975).
- **X1**: número de residentes en áreas urbanas en 1970 (en miles).
- **X2**: ingreso per cápita en 1973 (en miles dolares).
- **X3**: número de residentes menores de 18 años en 1974 (en miles)

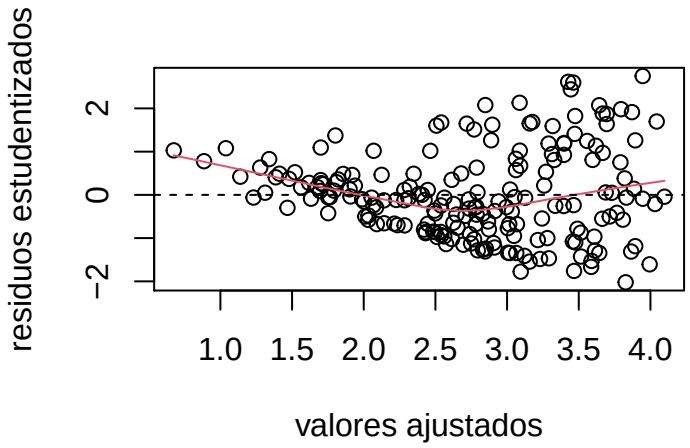

Figura 4.2: Datos de la ONU. Gráfico de los residuos estudentizados.

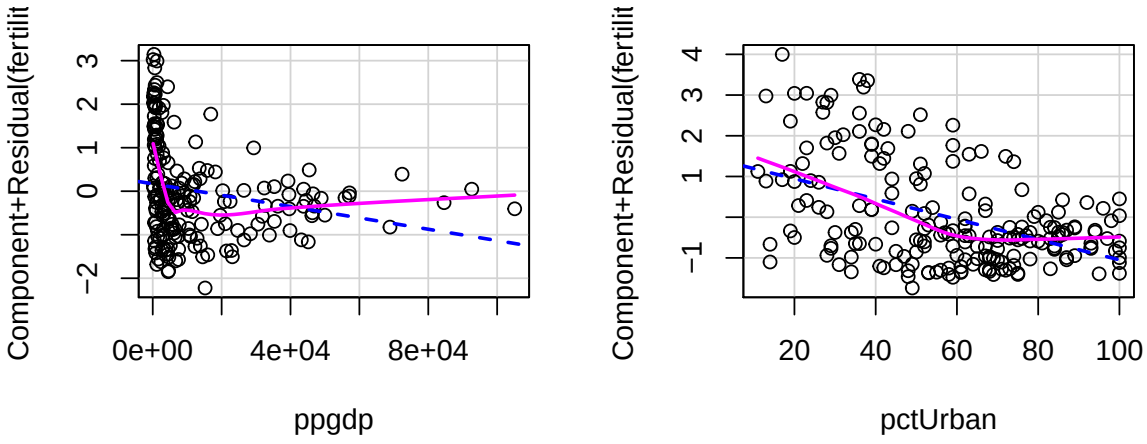

Figura 4.3: Datos de la ONU. Gráfico de los residuos parciales.

La relación entre las variables se observa en la Figura 4.4. Se observa una relación positiva aproximadamente lineal entre la variable respuesta y las covariables, aunque no es tan fuerte con la covariable número de residentes menores de 18 años. Además, vemos que hay por lo menos un posible valor atípico.

```
library(robustbase)
data("education")
education$X2 = education$X2/1000 # cambio de unidad de medida (miles de dolares)
pairs(education[,c(6,3:5)])
```

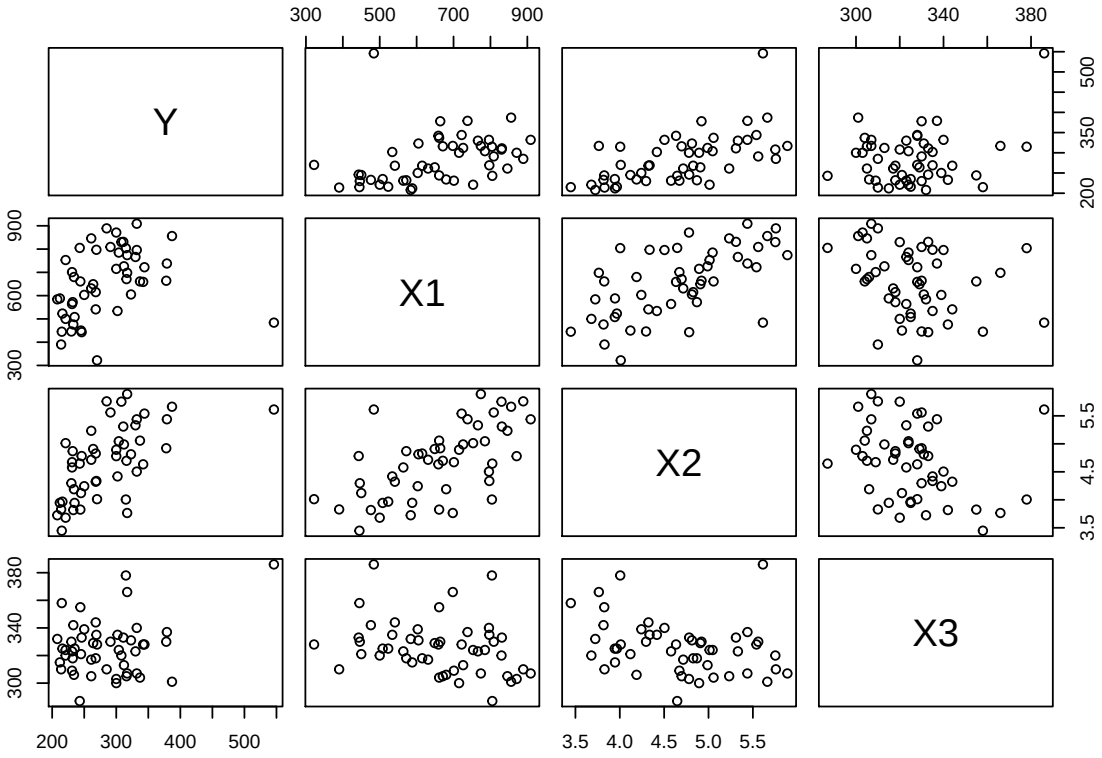

Figura 4.4: Datos de educación. Relación entre las variables.

El objetivo es ajustar un modelo de regresión para el gasto per cápita en educación pública en función de las demás variables:

$$Y_i = \beta_0 + \beta_1 \text{X1}_i + \beta_2 \text{X2}_i + \beta_3 \text{X3}_i + \varepsilon_i,$$

donde $\varepsilon_i \sim N(0, \sigma^2)$ y $cov(\varepsilon_j, \varepsilon_k) = 0$.

Antes de hacer inferencias sobre el modelo hacemos un análisis de los residuos. La Figura 4.5 exhibe el gráfico de los residuos estudentizados. Aunque la relación entre la variable respuesta y covariables es aproximadamente lineal, hay presencia de heterocedasticidad. La variabilidad de los residuos aumenta con los valores ajustados.

```
mod.educ = lm(Y~X1+X2+X3,data=education)
library(MASS)
res.educ = studres(mod.educ)
```

```
par(mfrow=c(1,2))
plot(mod.educ$fitted.values,res.educ,
     xlab='valores ajustados',ylab='residuos estudentizados')
lines(lowess(res.educ~mod.educ$fitted.values),col=2)
abline(h=0,lty=2)
plot(mod.educ$fitted.values,abs(res.educ),
     xlab='valores ajustados',ylab='| residuos estudentizados |')
lines(lowess(abs(res.educ)~mod.educ$fitted.values),col=2)
```

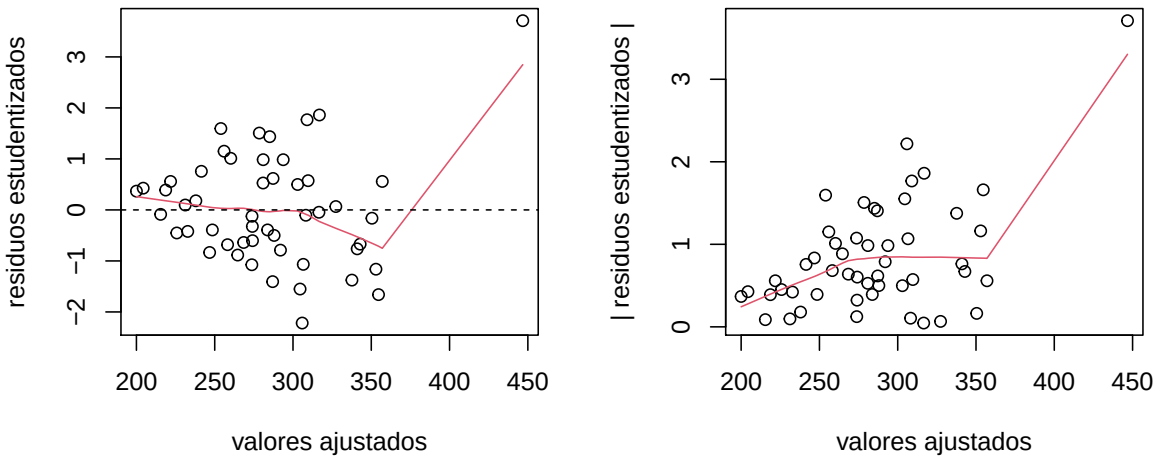

Figura 4.5: Datos de educación. Gráficos de los residuos estudentizados.

Dado que los modelos ajustados presentan desviaciones considerables de los supuestos asumidos, las inferencias que se hagan pueden ser invalidas. Por lo tanto, en este capítulo presentaremos dos herramientas para la corrección de estos problemas: (1) transformación de variables (incluyendo la transformación de Box-Cox) y (2) mínimos cuadrados ponderados.

4.1 Transformación de los datos

Los objetivos de realizar transformaciones sobre los datos son:

- linealizar la relación de las variables,
- estabilizar la varianza,
- y corregir la normalidad.

Las transformaciones pueden hacerse sobre la variable respuesta, las covariables, o ambas.

La desventaja de hacer transformaciones es que la interpretación del modelo estimado, así como las inferencias, se hacen sobre las variables transformadas, y no en su escala original.

Tabla 4.1: Funciones linealizables

	Función	Transformación	Forma lineal
(a)	$y = \beta_0 x^{\beta_1}$	$y^* = \log(y), x^* = \log(x)$	$y^* = \log(\beta_0) + \beta_1 x^*$
(b)	$y = \beta_0 e^{\beta_1 x}$	$y^* = \log(y)$	$y^* = \log(\beta_0) + \beta_1 x$
(c)	$y = \beta_0 + \beta_1 ln(x)$	$x^* = \log(x)$	$y^* = \beta_0 + \beta_1 x^*$
(d)	$y = \frac{x}{\beta_0 x + \beta_1}$	$x^* = \frac{1}{x}, y^* = \frac{1}{y}$	$y^* = \beta_0 - \beta_1 x^*$

Tabla 4.2: Algunas transformaciones para estabilizar la varianza

Relación entre σ^2 y $E(Y)$	Transformación
$\sigma^2 \propto C$	$y^* = y$
$\sigma^2 \propto E(y)$	$y^* = \sqrt{y}$
$\sigma^2 \propto E(y)[1 - E(y)]$	$y^* = \sin^{-1}\sqrt{y} (0 \leq y_i \leq 1)$
$\sigma^2 \propto E(y)^2$	$y^* = \log y$ o también $y^* = \log(y+1)$ (si $y \geq 0$)
$\sigma^2 \propto E(y)^3$	$y^* = \frac{1}{\sqrt{y}}$
$\sigma^2 \propto E(y)^4$	$y^* = \frac{1}{y}$

4.1.1 Transformaciones para linealizar el modelo

Recordemos que el modelo lineal asume que la relación entre la media y las covariables es aproximadamente lineal. En algunos casos, dada la naturaleza de los datos, este supuesto puede ser violado. Por lo tanto, para seguir utilizando la metodología de los modelos lineales, es posible linealizar funciones no-lineales por medio de transformaciones.

Algunas de estas funciones linealizables y su representación gráfica, se muestran en la Tabla 4.1 y Figura 4.6, respectivamente. Por ejemplo, considere que el modelo generador de los datos es:

$$y_i = \beta_0 \exp(\beta_1 x_{i1})\, \varepsilon_i \text{ [ver figura 4.6(b)].}$$

Esta relación no lineal se puede linealizar aplicando una transformación logaritmica a ambos lados:

$$\log y_i = y_i^* = \log[\beta_0 \exp(\beta_1 x_i)\, \varepsilon_i] = \log \beta_0 + \beta_1 x_i + \log \varepsilon_i.$$

Note que estaríamos asumiendo que $\log \varepsilon_i$ está normalmente distribuido. Para que esto sea cierto, ε_i debe seguir una distribución log-normal.

4.1.2 Transformaciones para estabilizar la varianza

Un caso frecuente es que la variable respuesta sigue una distribución de probabilidad en la que la varianza se relaciona en forma funcional con la media:

$$V(Y|X = x) = \sigma^2 g[E(Y|X = x)].$$

Por ejemplo, en la distribución Poisson, la varianza es igual a la media. Algunas transformaciones comunes para estabilizar varianza se muestran en la Tabla 4.2 (Behar, 2002).

Algunas consideraciones cuando se hacen transformaciones sobre las variables:

- Transformaciones pueden ser sugeridas por experiencia (o teoría). En otros casos, la selección se hace empíricamente.

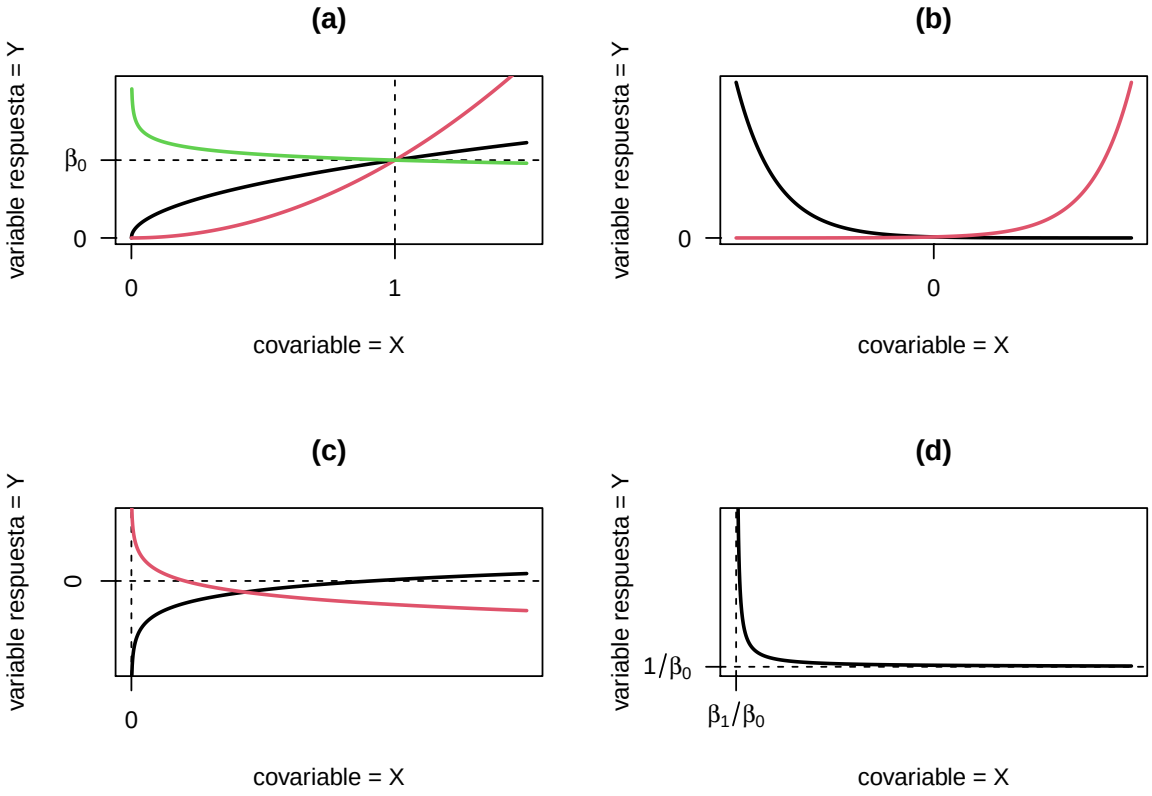

Figura 4.6: Diferentes patrones linealizables.

- Luego de realizar las transformaciones se debe verificar si el modelo transformado cumple los supuestos.
- El estimador de MCO tiene propiedades de mínimos cuadrados con respecto a los datos transformados.
- Las predicciones son sobre las respuestas transformadas, no las originales. Devolverse a la variable respuesta original no es fácil. Recordemos que

$$E[g(y)] \neq g[E(y)].$$

Al aplicar la transformación inversa a las predicciones de la respuesta transformada estamos estimando la mediana, y no la media. Por otro lado, a las estimaciones por intervalos de confianza si se les puede aplicar la transformación inversa. Esto porque los percentiles no se ven afectados por transformaciones.

Datos de la ONU. Transformación para linealizar los datos

Al realizar un análisis de residuos del ajuste del modelo para los datos de la ONU, vimos que hay una relación no-lineal entre la fertilidad y las dos covariables propuestas. Particularmente, esto se debe a la covariable producto nacional bruto.

Por lo tanto, podemos aplicar una transformación logarítmica tanto a la variable respuesta, así como la covariable producto nacional bruto. En la Figura 4.7 vemos como al aplicar esta transformación se linealiza la relación.

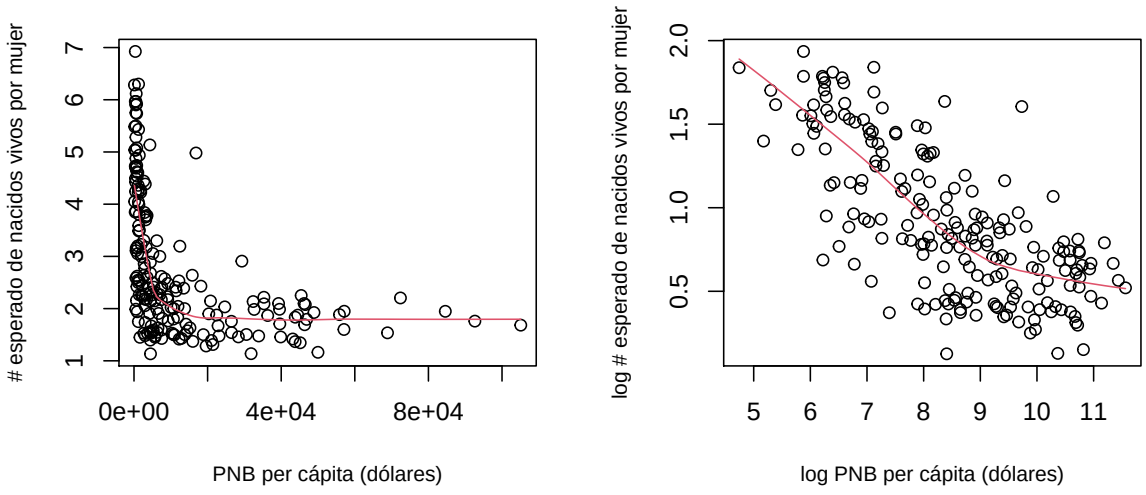

Figura 4.7: Datos de la ONU. Relación entre las variables.

Producto de esta transformación, se propone el siguiente modelo:

$$\log \text{fertility}_i = \beta_0 + \beta_1 \log \text{ppgdp}_i + \beta_2 \text{pctUrban}_i + \varepsilon_i,$$

donde $\varepsilon_i \sim N(0, \sigma^2)$ y $cov(\varepsilon_j, \varepsilon_k) = 0$.

La Figura 4.8 muestra el gráfico de los residuos para este ajuste. Aquí vemos que los problemas de no linealidad y heterocedasticidad se corrigieron al realizar la transformación logarítmica.

```
mod.UN11.trans = lm(log(fertility)~log(ppgdp)+pctUrban,data = UN11)
res.UN11.trans = studres(mod.UN11.trans)
par(mfrow=c(1,2))
plot(mod.UN11.trans$fitted.values,res.UN11.trans,xlab='valores ajustados',
     ylab='residuos estudentizados')
lines(lowess(res.UN11.trans~mod.UN11.trans$fitted.values),col=2)
abline(h=0,lty=2)
plot(mod.UN11.trans$fitted.values,abs(res.UN11.trans),xlab='valores ajustados',
     ylab='| residuos estudentizados |')
lines(lowess(abs(res.UN11.trans)~mod.UN11.trans$fitted.values),col=2)
```

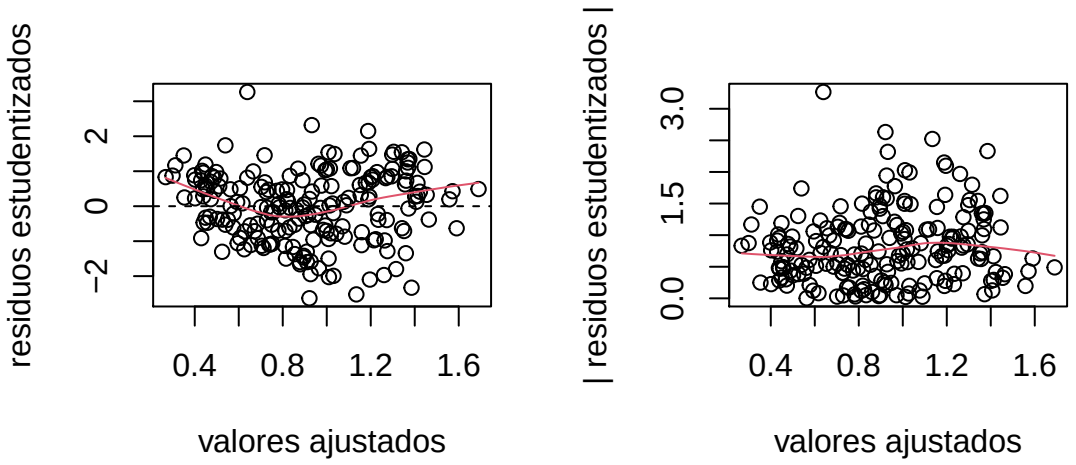

Figura 4.8: Datos de la ONU. Gráficos de los residuos para el modelo transformado.

4.2 Método de Box-Cox

Para la corrección del supuesto de normalidad y varianza constante, es posible implementar transformaciones en potencia para la variable respuesta. Esto es, $y^* = y^\lambda$. Dado que el valor de λ es desconocido, la idea del **método de Box-Cox** es estimar el modelo lineal para diferentes valores de λ y determinar el valor que proporciona el mejor ajuste. Sin embargo, aquí encontramos dos problemas.

Primero, la transformación en potencia tiene un problema de discontinuidad en $\lambda = 0$. Puesto que cuando λ tiende a cero, y^* se acerca a 1. Para resolver esto, se puede utilizar $y^* = (y^\lambda - 1)/\lambda$. De esta forma, cuando λ tiende a cero, y^* se acerca a $\log y$. Segundo, cuando λ cambia, los valores y^* varía drásticamente. Esto hace que los modelos ajustados no se puedan comparar fácilmente.

La transformación que permite que los modelos ajustados sean comparables es:

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda \dot{y}^{\lambda-1}}, & \text{si } \lambda \neq 0, \\ \dot{y} \log y & \text{si } \lambda = 0, \end{cases} \tag{4.1}$$

donde $\dot{y} = \log\left[1/n \sum_{i=1}^{n} \log y_i\right]$ es la media geométrica de la variable respuesta.

Entonces, el método de Box-Cox es el siguiente:

1. Determinar una secuencia de valores para λ, $(\lambda_1, \lambda_2, \dots, \lambda_K)$. Por lo general, se seleccionan valores en el intervalo $[-2, 2]$.
2. Ajustar el modelo:
$$y_{ij}^{(\lambda_k)} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_{p-1} x_{i,p-1} + \varepsilon_i,$$
para cada valor de λ_k, $k = 1, \dots, K$. Al utilizar la transformación (4.1), las sumas de cuadrados para modelos con diferentes valores de λ son comparables.
3. Seleccionar el modelo que minimiza la suma de cuadrados de los residuos, $SS_{res}(\lambda)$. Equivalentemente, el λ que maximiza la verosimilitud.
4. Luego de encontrar el valor de λ óptimo, se ajusta el modelo transformando la variable respuesta y^λ, si $\lambda \neq 0$, o $\log y$ si $\lambda = 0$. Es decir que (4.1) se utiliza solo en el paso de comparación de modelos ajustados.

Puesto que λ es una variable aleatoria, también se puede hacer una estimación por intervalos de confianza. Para esto primero consideremos la función de log-verosimilitud:

$$\ell(\beta, \sigma^2|\lambda) = -\frac{2}{n}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i^\lambda - x_i'\beta\right)^2 = -\frac{2}{n}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}SS_{res}(\lambda).$$

Aquí vemos que el valor de λ que maximiza la verosimilitud es el mismo que minimiza la suma de cuadrados de los residuos.

Un intervalo de confianza se puede construir a partir del siguiente estadístico de prueba:

$$G_0^2 = -2\left[\ell(\beta, \sigma^2|\lambda = 1) - \ell(\beta, \sigma^2|\lambda = \hat{\lambda})\right].$$

Si $\lambda = 1$, entonces asintóticamente $G_0^2 \sim \chi_1^2$. Por lo tanto, el intervalo del $(1-\alpha)\times 100\%$ de confianza para λ está definido por los valores de λ que cumplen con la condición:

$$\ell(\beta, \sigma^2|\lambda) \geq \ell(\beta, \sigma^2|\lambda = \hat{\lambda}) - \frac{1}{2}\chi_{1,1-\alpha}^2.$$

4.2.1 Datos de educación. Transformación de Box-Cox

Dado que el análisis de residuso para el ajuste del modelo para los datos de educación mostró que hay problemas de heterocedasticidad, vamos a encontrar una transformación que resuelva de problema usando el método de Box-Cox. Para esto usamos la función `boxcox` de la librería `MASS`:

```
boxcox.educ = MASS::boxcox(mod.educ,lambda=seq(-3,3,length.out = 1000),
                           ylab='log-verosimilitud')
```

```
boxcox.educ$x[boxcox.educ$y ==max(boxcox.educ$y)] # valor que maximiza la log-verosimilitud
```

```
## [1] -1.09009
```

Estos resultados nos indican que $\hat{\lambda} = -1.09$. Por lo cuál, podemos utilizar una transformación inversa ($\lambda = -1$). Entonces, el modelo propuesto es:

$$1/y_i = \beta_0 + \beta_1 \text{X1}_i + \beta_2 \text{X2}_i + \beta_3 \text{X3}_i + \varepsilon_i,$$

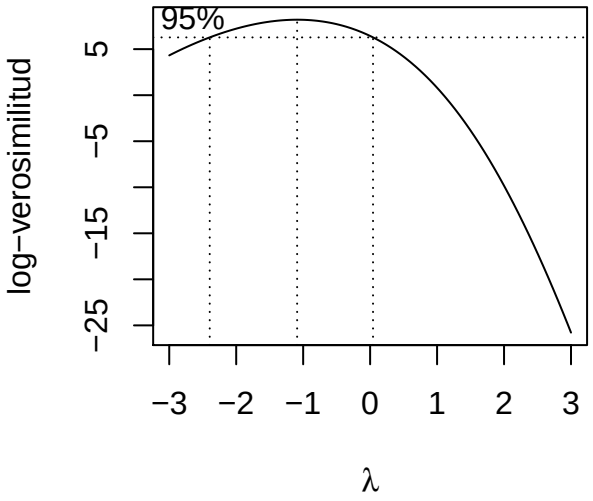

Figura 4.9: Datos de educación. Perfiles de verosimilitud para λ.

donde $\varepsilon_i \sim N(0, \sigma^2)$ y $cov(\varepsilon_j, \varepsilon_k) = 0$.

Ahora procedemos a hacer el análisis de los residuos del modelo transformado:

```
mod.educ.trans = lm(1/Y~X1+X2+X3,data=education)
res.educ.trans = studres(mod.educ.trans)
par(mfrow=c(1,2))
plot(mod.educ.trans$fitted.values,res.educ.trans,
     xlab='valores ajustados',ylab='residuos estudentizados')
lines(lowess(res.educ.trans~mod.educ.trans$fitted.values),col=2)
abline(h=0,lty=2)
plot(mod.educ.trans$fitted.values,abs(res.educ.trans),
     xlab='valores ajustados',ylab='| residuos estudentizados |')
lines(lowess(abs(res.educ.trans)~mod.educ.trans$fitted.values),col=2)
```

En la Figura 4.10 vemos que la transformación propuesta corrigió el problema de heterocedasticidad. Adicionalmente, por medio de la Figura 4.11 podemos verificar que el supuesto de normalidad se cumple.

```
car::qqPlot(mod.educ.trans,distribution = 'norm',ylab='residuos estudentizados')
```

```
## [1] 10 47
```

Puesto que se cumplen los supuestos del modelo transformado, ahora procedemos a interpretar los resultados.

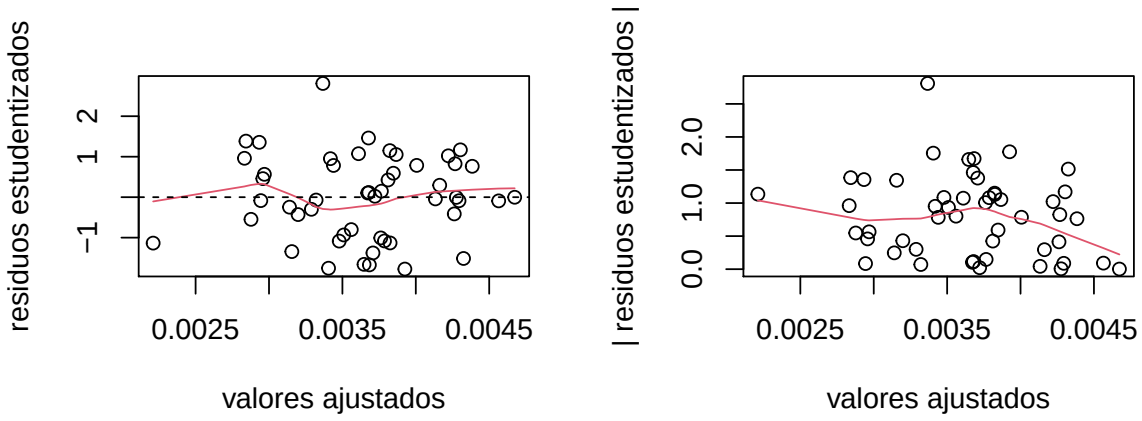

Figura 4.10: Datos de educación. Graficos de los residuos para el modelo transformado.

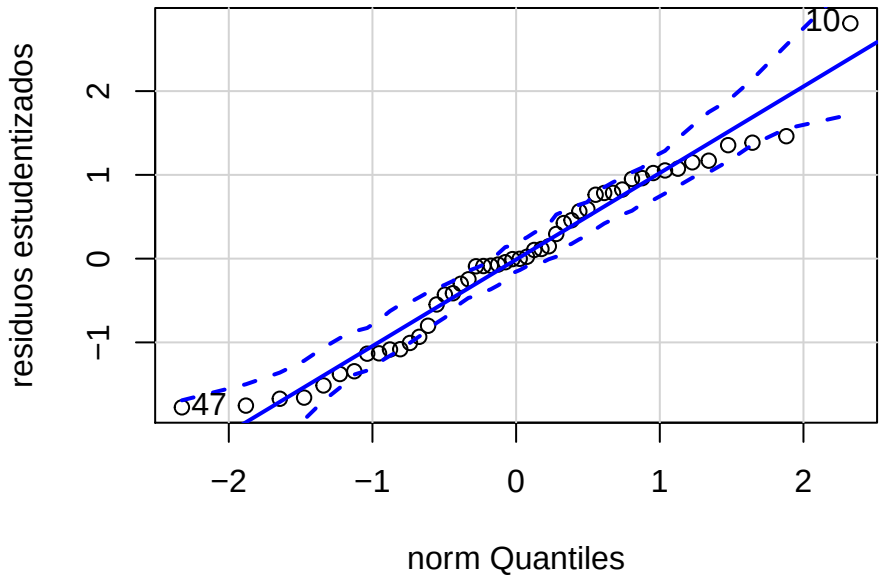

Figura 4.11: Datos de educación. Graficos de normalidad de los residuos para el modelo transformado.

```
summary(mod.educ.trans)
```

```
##
## Call:
## lm(formula = 1/Y ~ X1 + X2 + X3, data = education)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -7.609e-04 -3.732e-04 -1.640e-06  3.468e-04  1.158e-03
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.220e-02  1.366e-03   8.934 1.29e-11 ***
## X1          -6.613e-07  5.698e-07  -1.161 0.251816
## X2          -7.302e-04  1.286e-04  -5.676 8.83e-07 ***
## X3          -1.444e-05  3.489e-06  -4.139 0.000147 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0004487 on 46 degrees of freedom
## Multiple R-squared:  0.5897, Adjusted R-squared:  0.5629
## F-statistic: 22.04 on 3 and 46 DF,  p-value: 5.417e-09
```

A partir de estos resultados podemos concluir que la población en áreas urbanas no tiene un aporte signficativo en el modelo cuando las otras dos covariables ya están incluidas. Mientras que, el ingreso per cápita y la población menor de 18 años tienen un efecto positivo significativo sobre el gasto en educación pública (recordemos que se hizo una transformación inversa).

Puesto que hicimos transformaciones, la interpretación de los coeficientes se hace sobre la variable respuesta transformada.

Ahora, supogamos que queremos hacer la predicción del gasto medio en educación pública para los estados que tengan una población de X1 = 650, un ingreso per cápita de X2=4.5 y una población menor de 18 años de X3 = 320. A partir del modelo ajustado tenemos que:

```
x0.educ = data.frame(X1=650,X2=4.5,X3=320)
pred.educ.trans = predict(mod.educ.trans,x0.educ,interval='confidence')
1/pred.educ.trans
```

```
##        fit      lwr      upr
## 1 258.6228 268.5117 249.4364
```

Por lo que intervalo del 95% de confianza para el gasto medio en educación pública para los estados con las características expresadas anteriormente es (249.44, 268.51).

4.3 Mínimos cuadrados ponderados

El método de mínimos cuadrados ponderados (MCP) es una alternativa para estimar un modelo lineal en presencia de heterocedasticidad. La idea de esta técnica es calcular las desviaciones entre las observaciones (y_i) y los valores ajustados ($\hat{y}_i$) usando pesos (w_i) inversamente proporcionales a la varianza de y_i.

Asumamos que el modelo generador de los datos es:

$$y = X\beta + \varepsilon, \text{ donde } \varepsilon \sim N(0, \sigma^2 V), \tag{4.2}$$

donde V es una matriz diagonal $(n \times n)$:

$$V = \begin{pmatrix} v_{11} & 0 & 0 & ... & 0 \\ 0 & v_{22} & 0 & ... & 0 \\ 0 & 0 & v_{33} & ... & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & ... & v_{nn} \end{pmatrix}.$$

Esto quiere decir que tenemos presencia de hetorocedasticidad, donde $V(\varepsilon_i) = \sigma^2 v_{ii}$. Recordemos que si aplicamos el estimador por MCO, las estimaciones siguen siendo insesgadas pero son ineficientes. Además, las varianzas estimadas de $\widehat{\beta}$, y de las predicciones, están mal calculadas. Lo que puede afectar la cobertura de los intervalos de confianza y el nivel de signficancia de las pruebas de hipótesis.

La función de verosimilitud del modelo (4.2), asumiendo que V es conocida, es:

$$L(\beta, \sigma^2 | V) = \frac{1}{(2\pi)^{n/2} |\sigma^2 V|^{1/2}} \exp\left[-\frac{1}{2\sigma^2} (y - X'\beta)' V^{-1} (y - X'\beta) \right].$$

Por lo que para encontrar el estimador de β debemos minimizar la suma de cuadrados ponderada:

$$SS_{Wres} = (y - X'\beta)' W (y - X'\beta) = \sum_{i=1}^{n} w_{ii} (y_i - x_i'\beta)^2, \tag{4.3}$$

donde $W = V^{-1}$, es decir que $w_{ii} = \frac{1}{v_{ii}}$. Note que las observaciones con mayor varianza tienen menor peso en la estimación de β.

Al minimizar (4.3) se obtiene el estimador por MCP:

$$\widehat{\beta}_W = (X'WX)^{-1} X'Wy.$$

Además,

$$V(\widehat{\beta}_W) = \sigma^2 (X'WX)^{-1}. \tag{4.4}$$

Si V está correctamente especificada, se puede probar que $\widehat{\beta}_W$ es el mejor estimador insesgado de β.

Los residuos del ajuste del modelo son:

$$e_{Wi} = \sqrt{w_{ii}} (y_i - \widehat{y}_i), \text{ de forma matricial } e_W = W^{1/2} (y - X\widehat{\beta}_W),$$

donde $W = W^{1/2} W^{1/2}$. De este resultado obtenemos el estimador insesgado de σ^2:

$$\widehat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^{n} w_{ii} (y_i - x_i'\widehat{\beta})^2 = \frac{1}{n-p} (y - X\widehat{\beta})' W (y - X\widehat{\beta}).$$

Aqui estamos asumiendo que los pesos (w_{ii}) son conocidos. Lo que en la práctica es poco común. Estos pesos pueden ser determinados por conocimiento de los datos, experiencia, o información teórica del modelo. Alternativamente, estos se pueden calcular a partir de los residuos obtenidos por el estimador MCO.

Es común que σ_i^2 varíe de acuerdo a una o varias covariables, o con respecto $E(y_i|x_i)$. Por ejemplo, en los datos de educación parece que la varianza incrementa con $E(y_i|x_i)$. En este caso podemos aplicar el siguiente procedimiento:

1. Ajustar el modelo por MCO y analizar los residuos.
2. Ajustar un modelo para el valor absoluto de los residuos ($|e_i|$), o los residuso al cuadrado (e_i^2), en función de las covariables:

$$|e_i| = \gamma_0 + \gamma_1 x_{i1} + ... + \gamma_{p-1} x_{i,p-1} + \varepsilon_{ei}.$$

   De esta ajuste obtenemos una estimación para la desviación estándar (σ_i). Si ajustamos un modelo para los residuos al cuadrado, estamos estimando la varianza (σ^2).
3. Usar los valores ajustados el modelo anterior para estimar los pesos (w_{ii}).

Si $\hat{\beta}_W$ difiere mucho de $\hat{\beta}$, es posible repetir el proceso anterior una vez más (mínimos cuadrados iterativamente ponderados).

Dado que estamos estimando los pesos, la varianza de los coeficientes (4.4) es aproximada. Sin embargo, esta aproximación es muy buena si el tamaño de muestra no es muy pequeño.

Datos de educación. Minímos cuadrados ponderados

Retomemos el modelo para los datos de educación. En la Figura 4.5 vemos que hay heterocedasticidad. Particularmente vemos que la variabilidad crece a medida que aumentan los valores ajustados. Ya vimos que por medio de una transformación inversa sobre la variable respuesta se corrige el problema.

El problema de hacer transformaciones es que los coeficientes estimados pierden interpretación. Así que, alternativamente, vamos a implementar el método de minímos cuadrados ponderados. Dado que desconocemos los pesos, vamos a estimarlos usando el procedimiento que se explicó anteriormente.

Primero vamos a ajustar el siguiente modelo para el valor absoluto de los residuos:

$$|e_i| = \gamma_0 + \gamma_1 \text{X1}_i + \gamma_2 \text{X2}_i + \gamma_3 \text{X3}_i + \varepsilon_{ei}.$$

```
mod.stdev.educ = lm(abs(res.educ)~X1+X2+X3,data = education)
summary(mod.stdev.educ)
```

```
##
## Call:
## lm(formula = abs(res.educ) ~ X1 + X2 + X3, data = education)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.23416 -0.38164 -0.00938  0.32668  1.37892
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.6137734  1.7454477  -2.643   0.0112 *
## X1          -0.0014511  0.0007281  -1.993   0.0522 .
## X2           0.7274821  0.1643850   4.425 5.87e-05 ***
## X3           0.0092325  0.0044583   2.071   0.0440 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5734 on 46 degrees of freedom
## Multiple R-squared:  0.3199, Adjusted R-squared:  0.2756
## F-statistic: 7.214 on 3 and 46 DF,  p-value: 0.0004565
```

Por lo tanto, las estimaciones de las desviaciones estándar de los errores son:

$$\tilde{\sigma}_i = s_i = -4.614 - 0.001\text{X1}_i + 0.727\text{X2}_i + 0.009\text{X3}_i.$$

De este ajuste obtenemos los pesos $w_{ii} = \frac{1}{s_i^2}$, y luego, estimamos el modelo por MCP:

```
w = 1/mod.stdev.educ$fitted.values^2
mod.educ.mcp = lm(Y~X1+X2+X3,data=education,weights = w)
summary(mod.educ.mcp)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3, data = education, weights = w)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -93.670 -32.522  -9.149  31.582  86.742
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -404.85597   84.79548  -4.774 1.87e-05 ***
## X1             0.05062    0.03778   1.340    0.187
## X2            62.43586   10.29479   6.065 2.32e-07 ***
## X3             1.11736    0.20099   5.559 1.32e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 44.98 on 46 degrees of freedom
## Multiple R-squared:  0.6528, Adjusted R-squared:  0.6302
## F-statistic: 28.83 on 3 and 46 DF,  p-value: 1.219e-10
```

La Figura 4.12 muestra los gráficos de los residuos ponderados. Aquí podemos observar que los residuos ponderados no muestra heterocedasticidad.

```
res.educ.mcp = mod.educ.mcp$residuals*sqrt(w)
par(mfrow=c(1,2))
plot(mod.educ.mcp$fitted.values,res.educ.mcp,
     xlab='valores ajustados',ylab='residuos ponderados')
lines(lowess(res.educ.mcp~mod.educ.mcp$fitted.values),col=2)
abline(h=0,lty=2)
plot(mod.educ.mcp$fitted.values,abs(res.educ.mcp),
     xlab='valores ajustados',ylab='| residuos ponderados |')
lines(lowess(abs(res.educ.mcp)~mod.educ.mcp$fitted.values),col=2)
```

El ajuste por mínimos cuadrados ponderados es:

```
summary(mod.educ.mcp)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3, data = education, weights = w)
##
```

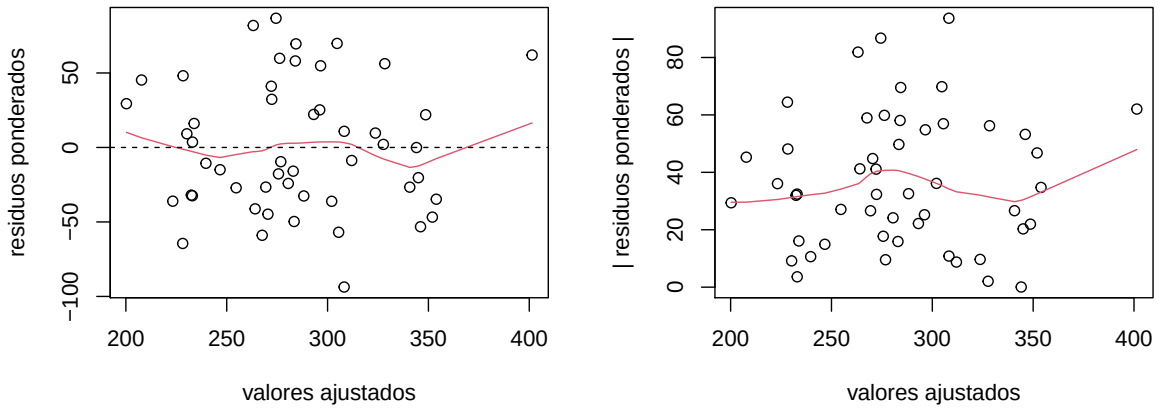

Figura 4.12: Datos de educación. Graficos de los residuos ponderados.

```
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -93.670 -32.522  -9.149  31.582  86.742
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -404.85597   84.79548  -4.774 1.87e-05 ***
## X1             0.05062    0.03778   1.340    0.187
## X2            62.43586   10.29479   6.065 2.32e-07 ***
## X3             1.11736    0.20099   5.559 1.32e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 44.98 on 46 degrees of freedom
## Multiple R-squared:  0.6528, Adjusted R-squared:  0.6302
## F-statistic: 28.83 on 3 and 46 DF,  p-value: 1.219e-10
```

Aquí llegamos a conclusiones similares a las obtenidas por medio del modelo transformado. El efecto de la población en áreas urbanas no es significativo. Las otras dos covariables si aportan significativamente al modelo. En este caso, dado que no aplicamos transformaciones, los coeficientes si tienen interpretación. Por ejemplo de la estimación de β_3 podemos concluir que: el gasto per cápita medio en educación pública aumenta en 1.117 USD por cada 1000 personas menores de 18 años.

Ahora, hagamos la predicción del gasto medio en educación pública para las características: X1 = 650, X2=4.5 y X3 = 320:

```
predict(mod.educ.mcp,x0.educ,interval='confidence')
```

```
##        fit      lwr      upr
## 1 266.5599 257.8442 275.2755
```

Por medio del estimador de MCP obtenemos el siguiente intervalo del 95% de confianza: (257.844,

275.275). Este intervalo es parecido al calculado usando el modelo transformado, aunque tiene una longitud un poco más pequeña.

---

nocite: Montgomery et al. (2012), Fox (2016)

Chapter 5

Evaluación de puntos influyentes y atípicos

5.1 Datos de la ONU

Retomemos los datos de la ONU (`UN11` de la librería `alr4`). El modelo propuesto es:

$$\log \text{fertility}_i = \beta_0 + \log \text{ppgdp}_i \beta_1 + \text{pctUrban}_i \beta_2 + \varepsilon_i, \tag{5.1}$$

donde $\varepsilon_i \sim N(0, \sigma^2)$ y $cov(\varepsilon_j, \varepsilon_k) = 0$.

El análisis de los residuos (ver Figura 5.1) muestra que el modelo está bien especificado y no hay problemas de heterocedasticidad. Sin embargo, podemos observar que algunos residuos presentan valores muy altos. La estimación de la log fertility para Guinea Ecuatorial es considerablemente baja en comparación con el valor observado. Caso contrario pasa con Corea del Norte, Bosnia y Herzegovina, y Moldavia.

5.2 Importancia de detectar valores influyentes y atípicos

En el análisis de datos pueden observarse algunos valores atípicos. Como atípicas nos referimos a las observaciones que no siguen el patrón de la mayoría de los datos. En un análisis de regresión, pueden presentarse valores atípicos sobre la variable respuesta y/o sobre algunas covariables. Por lo tanto, se podría identificar diferentes tipos de puntos "atípicos' '. Estos se pueden identificar en la Figura 5.2 para el caso de una regresión simple:

- A **es un punto atípico:** valor que no se ajusta bien en Y, pero regular en X.
- B **es un punto de balanceo:** observación que se ajusta bien en Y, pero es inusual en X.
- C **es un punto de influyente:** medición que no se ajusta bien en Y y es inusual en X.

Estos puntos inusuales pueden ser problemáticos a la hora de ajustar un modelo lineal por MCO, ya que pueden tener mucha influencia en los resultados, y su presencia puede ser una señal de que el modelo no captura caraterísticas importantes de los datos.

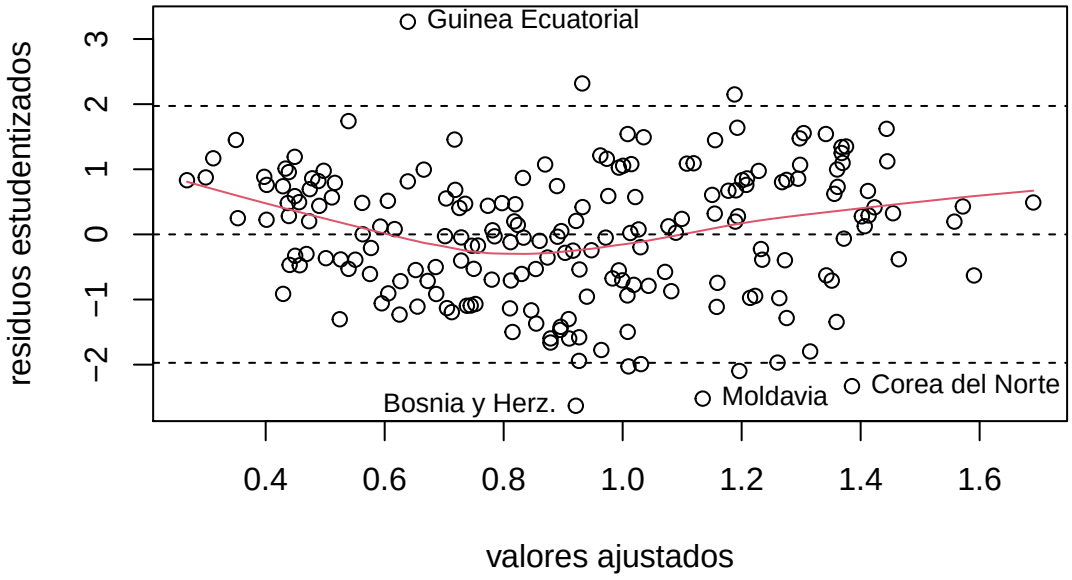

Figura 5.1: Datos de la ONU. Gráfico de los residuos estudentizados.

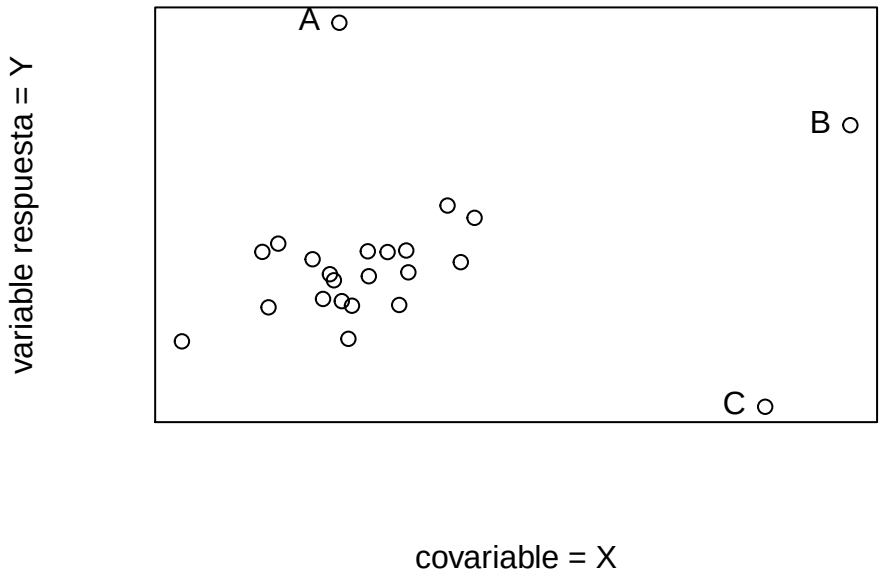

Figura 5.2: Diferentes tipos de datos. A: Punto atípico, B: punto de balanceo, C: punto influyente.

En la Figura 5.3(a) vemos que un punto atípico no tiene un efecto grande en la estimación de la recta de regresión. Sin embargo, dado que los puntos atípicos generan valores altos (en valor absoluto) para los residuos, estas observaciones inflan la varianza de las estimaciones y afectan las inferencias. En este caso, la estimación de la varianza con todos los datos es de $\hat{\sigma}^2 = 6.37$, y si se omite el dato A, tenemos que $\hat{\sigma}^2 = 1.55$. Como vemos en la Figura 5.3(b) los puntos de balanceo tampoco tienen mucha influencia sobre las estimaciones por MCO ya que estos están en línea con el resto de los datos.

Por el otro lado, en la Figura 5.3(c) vemos que los puntos influyentes afectan notablemente las estimaciones por MCO. Además, también inflan considerablemente la variabilidad. Con todos los datos tenemos que $\hat{\sigma}^2 = 3.09$. Mientras que $\hat{\sigma}^2 = 1.55$ al eliminar la observación C.

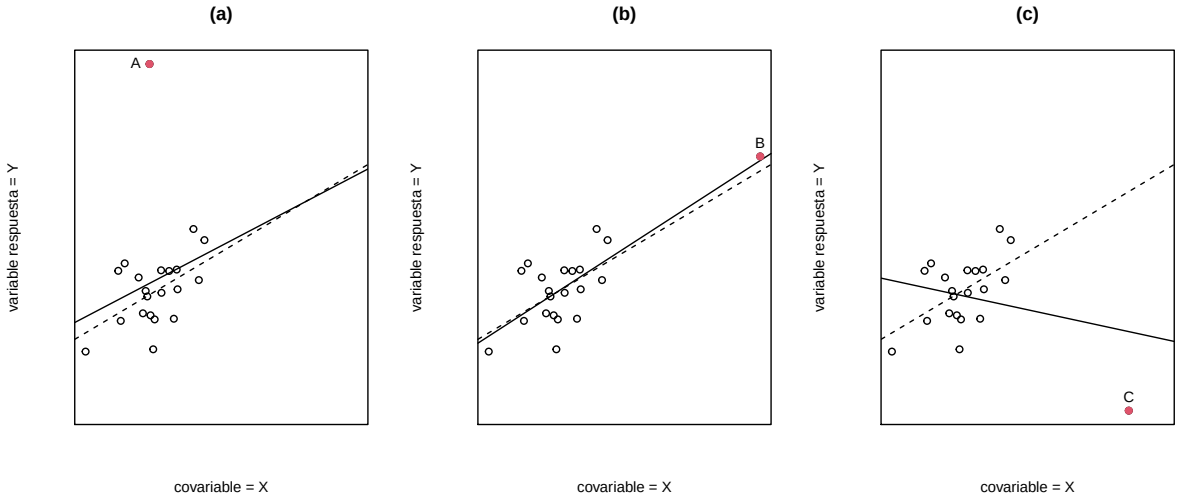

Figura 5.3: Efecto de los puntos inusuales. En cada gráfico, la línea continua es la estimación por MCO con todos los datos, mientras que la línea discontinua es la estimación por MCO omitiendo el punto inusual (circulo rojo). Izquierda: efecto de un punto atípico. Centro: efecto de un punto de balanceo. Derecha: efecto de un punto influyente.

5.3 Valores atípicos

Para identificar valores atípicos podemos hacer uso de los residuos del ajuste. Recordemos que, aunque los errores tenga varianza constante y sean incorrelacionados, los residuos no cumplen con estas propiedades. Por lo tanto, es recomendado usar los residuos estudentizados (o los residuos R-Student):

$$r_i = \frac{e_i}{\sqrt{\hat{\sigma}^2(1-h_{ii})}}, i = 1, \dots, n,$$

o los residuos R-Student:

$$t_i = \frac{e_i}{\sqrt{\hat{\sigma}^2_{(i)}(1-h_{ii})}} = r_i\sqrt{\frac{n-p-1}{n-p-r_i^2}},$$

donde $\hat{\sigma}^2_{(i)}$ es la estimación de σ usando todas las observaciones excepto la i-ésima.

Si se cumplen los supuestos del modelo, se puede demostrar que $r_i \sim t_{n-p-1}$. Los residuos estudentizados siguen también está distribución pero de forma aproximada. Por lo tanto, se pueden identificar posibles valores atípicos haciendo un gráfico de los R-Student (o residuos estudentizados) contra los valores ajustados y trazar líneas de referencia en los percentiles 0.025 y 0.975 de la distribución t con $n-p-1$ grados de libertad.

Esta verificación no es estrictamente una prueba de hipótesis. Puesto que estamos haciendo múltiplescomparaciones de los residuos R-Student con los valores criticos de la distribución t. Por lo que es necesario hacer una corrección utilizando el método de Bonferroni.

Datos de la ONU - valores atípicos

En la Figura 4.2 podemos observar que hay varios residuos que sobrepasan los puntos de corte. Particularmente, la observación de Guinea Ecuatorial presenta un residuo muy alto.

5.4 Puntos de balanceo

Recordemos que la estimación de la recta de regresión es un promedio ponderado de las observaciones:

$$\hat{y} = X\hat{\beta} = X(X'X)^{-1}X'y = Hy.$$

Para la observación y_i tenemos que $\hat{y}_i = h_{ii}y_i + \sum_{k\neq i} h_{ik}y_j$. Particularmente, el elemento h_{ij} pueden ser visto como la cantidad de balanceo o palanqueo ejercido por la j-ésima observación (y_i) sobre el i-ésimo valor ajustado ($\hat{y}_i$).

Entonces, para detectar valores influyentes vamos a centrarnos en la matriz *hat* (H). El elemento h_{ij} de esta matriz se calcula como:

$$h_{ij} = x_i'(X'X)x_j. \tag{5.2}$$

Algunas propiedades de la matriz H son:

- $\sum_{i=1}^{n} = p$.
- $\sum_{i=1}^{n} h_{ij} = \sum_{j=1}^{n} h_{ij} = 1$
- Cada valor h_{ii} está acotado entre $1/n$ y $1/r$ (r es el número de columnas de X iguales a x_i).

Además, la diagonal de la matriz H es una medida estandarizada de la distancia de las observaciones al centro (centroide) del espacio de x. Por lo tanto, valores altos en la diagonal de H pueden indicarnos observaciones que son potencialmente influyentes porque están alejadas en el espacio de las covariables. Viendo (5.2), si h_{ii} es muy cercano a uno, $\hat{y}_i$ estará muy cerca de y_i (dado que el peso de las demás observaciones será casi cero).

Dado que $\bar{h} = p/n$, observaciones con h_{ii} superiores a $2p/n$ son considerados **puntos de balanceo** (y posibles puntos influyentes). Note que, si $2p/n > 1$, el punto de corte no aplica.

Datos de la ONU - diagonal de la matrix hat

La Figura 5.4(a) muestra los valores de la diagonal de la matrix H. Aquí podemos observar que algunos países presentan valores más altos del punto de corte ($2\frac{p}{n} = 0.0302$). Comparado con los demás valores, el valor asociado a Trinidad y Tobago es muy alto. Por lo cuál este país lo podemos considerar como un punto de balanceo.

```
library(alr4)
data(UN11)
Names = rownames(UN11)
mod.UN11 = lm(log(fertility)~log(ppgdp)+pctUrban,data=UN11)
hii = hatvalues(mod.UN11)
par(mfrow=c(1,2))
plot(hii,type='h',xlab='índice',ylab='valores de la diagonal de la matrix hat')
abline(h=2*3/199,lty=2)
hii[hii>2*3/199]
```

```
##            Djibouti             Liberia               Nauru         North Korea             Somalia
##          0.03105290          0.03747901          0.03708513          0.03105839          0.04390710
##           Sri Lanka Trinidad and Tobago
##          0.03334886          0.07590435
```

```
plot(log(UN11$ppgdp),UN11$pctUrban,xlab='log del PNB per cápita',
     ylab='% de población urbanas')
points(log(UN11$ppgdp)[Names=='Trinidad and Tobago'],
       UN11$pctUrban[Names=='Trinidad and Tobago'],col=2,pch=19)
```

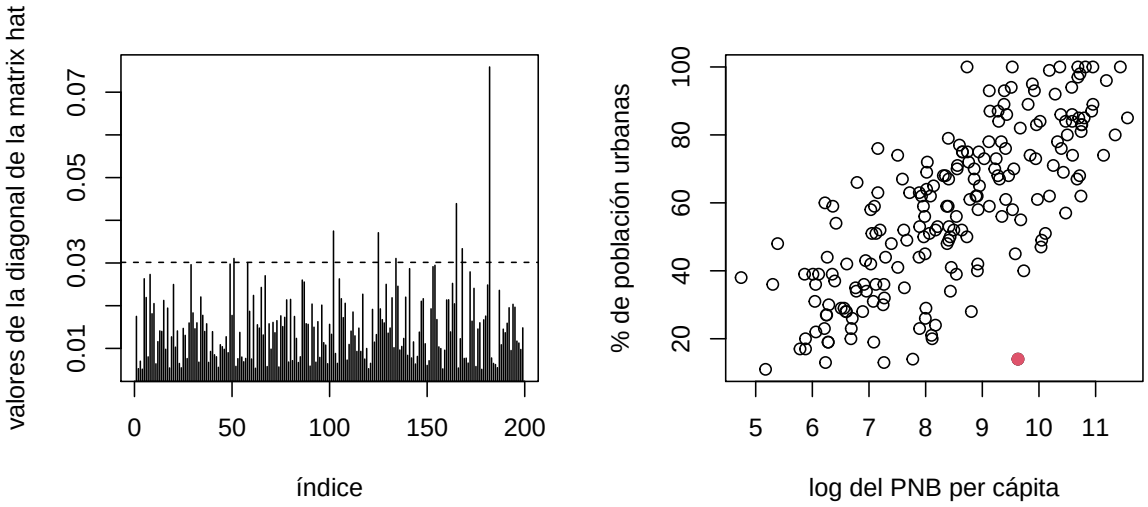

Figura 5.4: Datos de la ONU. (a) Valores de la diagonal de lam matriz hat. (b) Diagrama de dispersión de las covariables (derecha). El punto rojo indica a Trinidad y Tobago.

En laFigura 5.4(b) podemos observar que Trinidad y Tobago tiene un porcentaje muy pequeño de población en áreas urbanas y valor del PNB per cápita relativamente alto. Por esta razón el valor h_{ii} asociado es alto.

La Figura 5.5 muestra los valores de la diagonal de H contra los residuos estudentizados. Aquí vemos que aunque Trinidad y Tobago tiene un valor h_{ii} alto, el R-Student asociado es bajo. Por lo que no se puede considerar como un punto influyente. Por el contrario, Guinea Ecuatorial y Corea del Norte presentan ambos valores altos (residuos y h_{ii}), por lo que se pueden considerar como puntos influyentes. Las observaciones que tienen residuos altos pero valores h_{ii} bajos se consideran como atípicos.

```
rstud.UN11 = res.UN11*sqrt( (199-3-1)/(199-3-res.UN11^2) )
plot(hii,rstud.UN11,ylab='r Student',
     xlab='valores de la diagonal de la matrix hat')
abline(h=0,lty=2)
abline(h= c(-1,1)*qt(0.975,199-3-1),lty=2)
abline(v=2*3/199,lty=2)
text(hii[Names=="Trinidad and Tobago"],
     rstud.UN11[Names=='Trinidad and Tobago'],'Trinidad y Tobago',pos=2,cex=0.8)
text(hii[Names=="Equatorial Guinea"],
     rstud.UN11[Names=='Equatorial Guinea'],'Guinea Ecuatorial',pos=4,cex=0.8)
text(hii[Names=="North Korea"],
     rstud.UN11[Names=='North Korea'],'Corea del Norte',pos=4,cex=0.8)
text(hii[Names=="Somalia"],
     rstud.UN11[Names=='Somalia'],'Somalia',pos=4,cex=0.8)
text(hii[Names=="Bosnia and Herzegovina"],
     rstud.UN11[Names=='Bosnia and Herzegovina'],'Bosnia y Herz.',pos=4,cex=0.8)
```

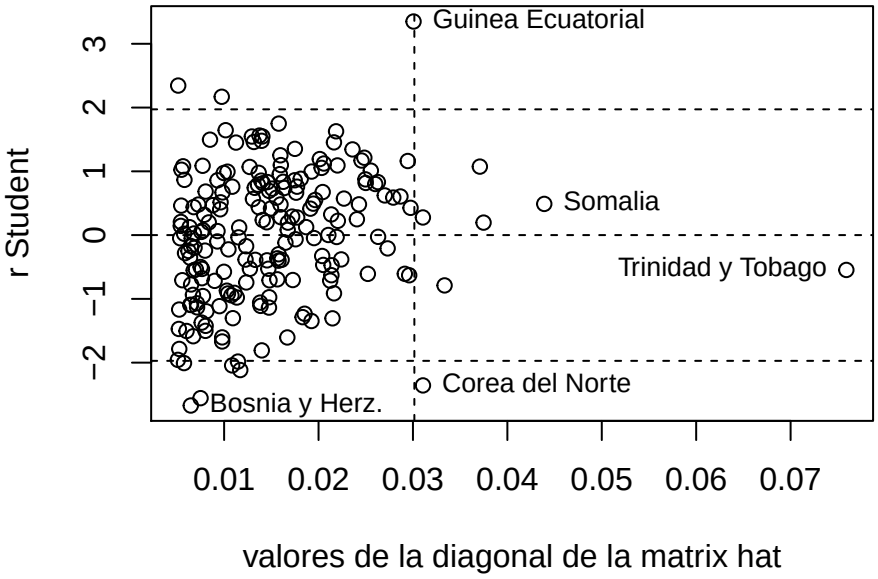

Figura 5.5: Datos de la ONU. Gráfico de valores influyentes

5.4.1 Medidas de influencia

El procedimiento para determinar si un punto es influyente se puede hacer evaluando los cambios que ocurren en el modelo ajustado cuando se elimina dicha observación.

Por ejemplo, la Figura 5.6 muestra cuanto cambian las estimaciones de β_1 y β_2 al eliminar un país a

la vez. Aquí vemos que los cambios mas grandes ocurren cuando se eliminan a Guinea Ecuatorial o a Corea del Norte. Mientras que los países que identificamos como puntos de balanceo (Trinidad y Tobago y Somalia) o atípicos (Bosnia y Herzegovina) no tienen mucha influencia sobre las estimaciones. ….

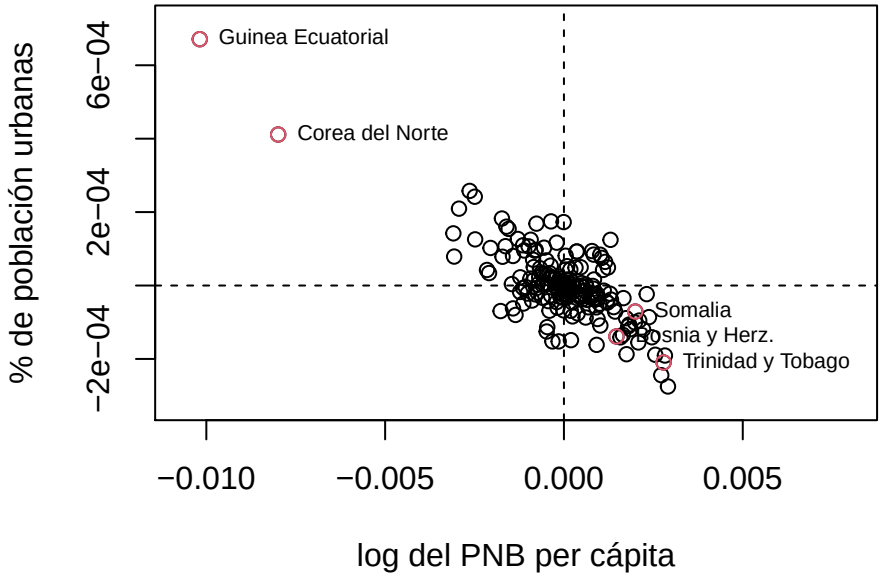

Figura 5.6: Datos de la ONU. Diferencias entre las estimaciones con todos los datos y las estimaciones obtenidas al eliminar una observación a la vez

De igual forma se podría evaluar cuanto cambian las estimaciones de $E(y|x_i)$ o las varianzas de los coeficientes $V(\beta_j)$ al eliminar observaciones una a una.

A continuación se presentan algunos indicadores estadísticos para detectar puntos influyentes:

5.4.1.1 Distancia de Cook

Esta es una medida de la distancia entre las estimaciones por MCO basado en los n puntos ($\hat{\beta}$), y el estimado obtenido eliminando el i-ésimo punto ($\hat{\beta}_{(i)}$). Es decir que es un indicador global de cuanto cambian todas las estimaciones de los coeficientes de regresión en conjunto. Esta medida se expresa de la siguiente forma:

$$\begin{aligned} D_i &= \frac{(\hat{\beta}_{(i)} - \hat{\beta})' X' X (\hat{\beta}_{(i)} - \hat{\beta})'}{p\hat{\sigma}^2} \\ &= \frac{(\hat{y}_{(i)} - \hat{y})'(\hat{y}_{(i)} - \hat{y})}{p\hat{\sigma}^2} = \frac{r_i^2}{p} \frac{h_{ii}}{1 - h_{ii}}, \qquad i = 1, 2, \ldots, n. \end{aligned}$$

Aquí podemos observar que D_i consta de dos componentes, uno asociado a el residuo (r_i) y otro a la distancia del vector x_i al centroide de la matriz de las covariables. Ambos (o alguno de ellos) puede contribuir a valores altos de este indicador.

Entonces, los puntos asociados a valores altos de D_i tienen gran influencia sobre la estimación de β por MCO. Se considera como un punto influyente si tiene asociado un $D_i > 4/n$ (algunos textos sugieren $D_i > 1$).

5.4.1.2 DFBETAS

Esta medida indica cuánto cambia el coeficiente de regresión $\hat{\beta}_j$, en unidades de desviaciones estándar, si se omitiera la i-ésima observación. Se calcula como:

$$DFBETAS_{(i,j)} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\hat{\sigma}^2_{(i)} C_{jj}} = \frac{r_{(j,i)}}{\sqrt{r'_j r_j}} \frac{t_i}{\sqrt{1-h_{ii}}},$$

donde C_{jj} es el j-ésimo valor de la diagonal de $(X'X)^2$. r_j es la j-ésima fila de $R = (X'X)^{-1}X$.

Un valor grande de $DFBETAS_{(j,i)}$ indica que la observación i tiene gran influencia sobre el j-ésimo coeficiente de regresión. Se sugiere que si $|DFBETAS_{(j,i)}| > 2/\sqrt{n}$ es necesario examinar la i-ésima observación.

5.4.1.3 DFFITS

Una medida que indica la influencia de la observación i-ésima sobre el valor ajustado ($\hat{y}_i$). Esta se calcula así:

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\hat{\sigma}^2_{(i)} h_{ii}} = \sqrt{\frac{h_{ii}}{1-h_{ii}}} t_i.$$

El $DFFITS_i$ puede ser grande si el dato es atípico (t_i grande) o si el dato tiene gran balanceo (h_{ii} grande). Se sugiere que si $|DFFITS_{(j,i)}| > 2\sqrt{p/n}$ es necesario examinar la i-ésima observación.

5.4.1.4 COVRATIO

Los diagnósticos anteriores permiten ver el efecto de las observaciones sobre $\hat{\beta}$ o $\hat{y}$. Pero no proporcionan información sobre la precisión general de la estimación. Una medida global de la precisión es la varianza generalizada:

$$GV(\hat{\beta}) = |Var(\hat{\beta})| = |\sigma^2 (X'X)^{-1}|.$$

Para determinar la influencia de la i-ésima observación en la precisión de la estimación se define la razón de covarianzas:

$$COVRATIO_i = \frac{|(X'_{(i)} X_{(i)})^{-1} \hat{\sigma}^2_{(i)}|}{|(X'X)^{-1}\hat{\sigma}^2|} = \left(\frac{S^2_{(i)}}{S^2}\right)^p \left(\frac{1}{1-h_{ii}}\right).$$

Si $COVRATIO_i > 1 + 3p/n$ o $COVRATIO_i > 1 - 3p/n$ se debería considerar a la i-ésima observación como influyente para la precisión de $\hat{\beta}$.

Datos de la ONU - medidas de influencia

En R las medidas de influencia se pueden calcular usando la función `influence.measures`. Aunque también se pueden calcular cada indicador de forma independiente.

La distancia de Cook se observa en la Figura 5.7. Aquí vemos que Guinea Ecuatorial y Corea del Norte son los países que más influyen en las estimaciones de los parámetros del modelo. Lo demás países presentan valores de la distancia de Cook mucho más bajos. Note que aunque observamos que Trinidad y Tobago es un punto de balanceo, este país no es influyente (valor de la distancia de Cook de 0.0082).

```
CD.UN11 = cooks.distance(mod.UN11)
OrderCD.UN11 = order(CD.UN11,decreasing = T)
plot(CD.UN11,type='h',ylab = 'distancia de Cook',xlab='índice')
text(OrderCD.UN11[1:2],CD.UN11[OrderCD.UN11[1:2]],Names[OrderCD.UN11[1:2]],pos=4,cex=0.6)
abline(h=4/199,lty=2)
```

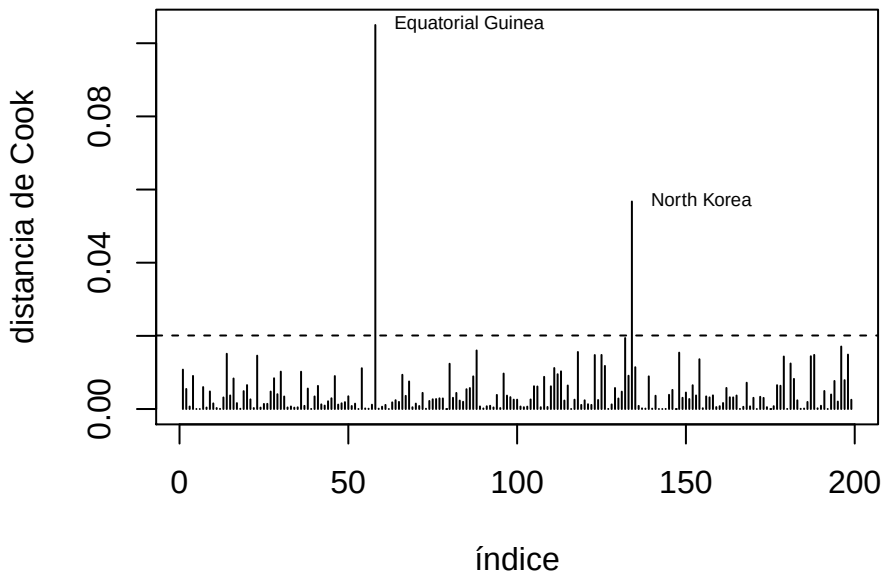

Figura 5.7: Datos de la ONU. Distancia de Cook.

Ya identificamos que Guinea Ecuatorial y Corea del Norte son influyentes y pueden afectar considerablemente las estimaciones de los coeficientes de regresión. Pero esta influencia puede ser solamente sobre algún o algunos parámetros. Para ver la influencia sobre cada parámetro, la Figura 5.8 muestra los DFBETAS para los coeficientes β_1 y β_2. En ambos gráficos podemos ver que estos dos países son influyentes para ambos parámetros. También se pude observar que otros pocos países tienen cierta influencia en la estimación de β_2, aunque los valores de los DFBETAS están muy cerca de los puntos de corte.

```
DFBetas.UN11 = dfbetas(mod.UN11)
OrderDB1.UN11 = order(abs(DFBetas.UN11[,2]),decreasing = T)
OrderDB2.UN11 = order(abs(DFBetas.UN11[,3]),decreasing = T)
```

```
par(mfrow=c(1,2))
plot(DFBetas.UN11[,2],ylab=quote('DFBETA'~(beta[1])),xlab='índice',main='(a)')
text(OrderDB1.UN11[1:2],DFBetas.UN11[OrderDB1.UN11[1:2],2],
     Names[OrderDB1.UN11[1:2]],pos=4,cex=0.6)
abline(h = c(-1,1)*2/sqrt(199),lty=2)
plot(DFBetas.UN11[,3],ylab = quote('DFBETA'~(beta[2])),xlab='índice',main='(b)')
text(OrderDB2.UN11[1:2],DFBetas.UN11[OrderDB2.UN11[1:2],3],
     Names[OrderDB2.UN11[1:2]],pos=4,cex=0.6)
abline(h = c(-1,1)*2/sqrt(199),lty=2)
```

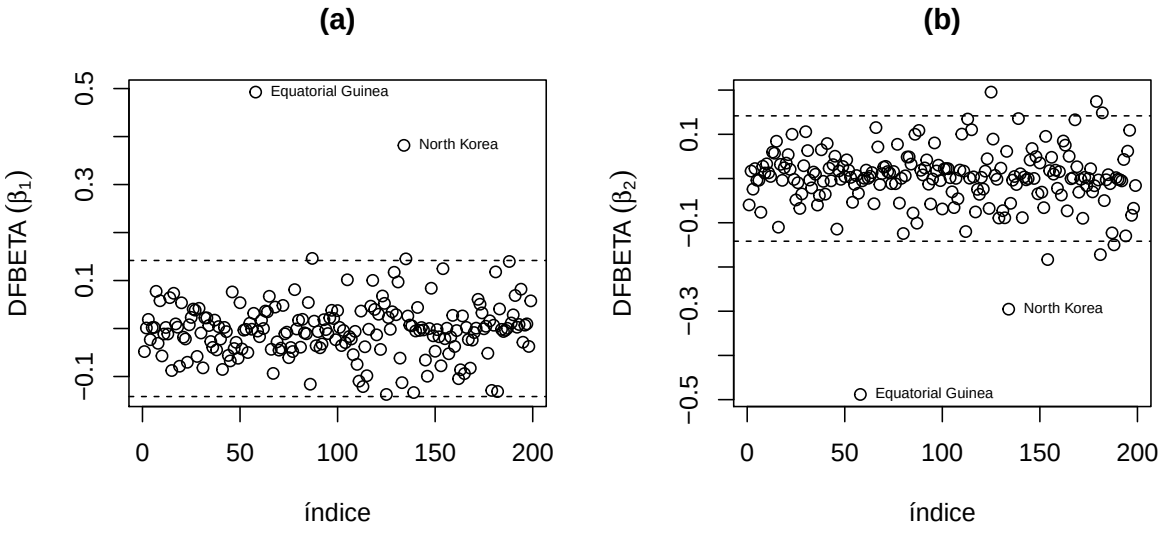

Figura 5.8: Datos de la ONU. DFBetas.

Los DFFITS y COVRATIO se observan en la Figura 5.9. A partir de los DFFITS se puede concluir que Guinea Ecuatorial y Corea del Norte también tienen gran influencia sobre las predicciones. A partir de los COVRATIO vemos que Trinidad y Tobago, Guinea Ecuatorial, Bosnia y Herzegovina, y Moldavia tienen gran influencia en la varianza de las estimaciones de los coeficientes de regresión. Note que aunque Corea del Norte fue influyente para las estimaciones, este país no influye en la varianza de estas.

```
DFfits.UN11 = dffits(mod.UN11)
OrderDFF.UN11 = order(abs(DFfits.UN11),decreasing = T)

Covratio.UN11 = covratio(mod.UN11)
OrderCR.UN11 = order(abs(1-Covratio.UN11),decreasing = T)
par(mfrow=c(1,2))
plot(DFfits.UN11,ylab='DFFITS',xlab='índice',main='(a)')
text(OrderDFF.UN11[1:2],DFfits.UN11[OrderDFF.UN11[1:2]],
     Names[OrderDFF.UN11[1:2]],pos=4,cex=0.6)
abline(h = c(-1,1)*2*sqrt(3/199),lty=2)
plot(Covratio.UN11,ylab = 'COVRATIO',xlab='índice',main='(b)')
text(OrderCR.UN11[1:4],Covratio.UN11[OrderCR.UN11[1:4]],
     Names[OrderCR.UN11[1:4]],pos=c(4,2,4,4),cex=0.6)
abline(h = 1+c(-1,1)*3*3/199,lty=2)
```

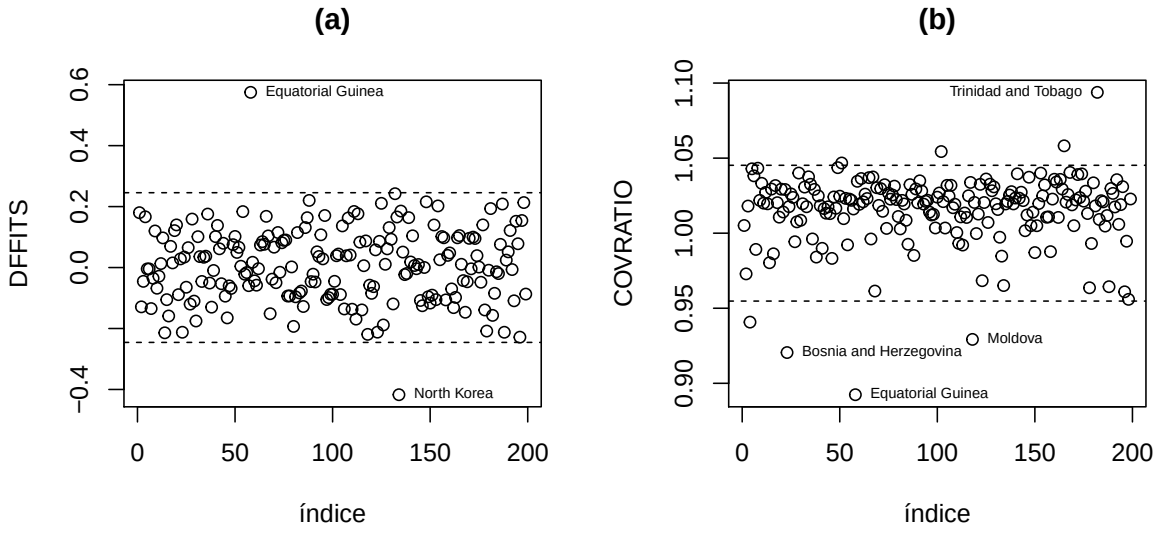

Figura 5.9: Datos de la ONU. DFFITS y COVRATIO.

A partir de estos indicadores encontramos que Guinea Ecuatorial y Corea del Norte son observaciones influyentes en la estimación del modelo (5.1). Particularmente, Guinea Ecuatorial tiene una tasa de fertilidad (4.98) muy superior a la estimada por el modelo. Esto se debe que los países con PNB y porcentaje de población urbana similares a este país tienen tasas de fertilidad más baja. Caso contrario pasa con Corea del Norte.

Para disminuir la influencia de estos países se pueden incorporar nuevas covariables dentro del modelo que ayuden a explicar estas discrepancias. Por ejemplo se podría ingresar una covariable asociada al continente.

5.5 Comentarios finales

- Dentro de la literatura hay muchos puntos de corte diferentes para los indicadores de observaciones influyentes. Esto es porque es difícil de determinar las distribuciones muestrales. Por lo que se recomienda verificar si hay algunas observaciones que tenga valores muy altos con respecto a los demás.
- Las observaciones influyentes o atípicas se deben descartar si estas corresponden a errores de medición o si son inválidas (por ejemplo, si pertenecen a otra población).
- Pero si las observaciones influyentes o atípicas son válidas, no hay justificación para eliminarla. Lo que se puede hacer es incluir nuevas covariables que puedan explicar mejor los datos y reducir la influencia de estas observaciones. En los datos de la ONU podríamos incluir covariables relacionadas con el continente.

Bibliography

Behar, R. (2002). *Validación de supuestos en el modelo de regresión*.

Fox, J. (2016). *Applied regression analysis and generalized linear models*. Third edition edition.

Montgomery, D. C., Peck, E. A., and Vining, G. G. (2012). *Introduction to linear regression analysis*. Wiley, 5th ed edition.

Weisberg, S. (2014). *Applied linear regression*. Fourth edition edition.