Regresión múltiple y otras técnicas multivariadas

Tarea 01

Rivera Torres Francisco de Jesús
Rodríguez Maya Jorge Daniel
Samayoa Donado Víctor Augusto
Trujillo Bariios Georgina

Febrero 06, 2019

Ejercicio 1

Validar la primera de las afirmaciones de Galton:

> *Los hijos de padres altos no son tan altos como sus padres.*

Utilizar el conjunto de datos **Galton** del paquete HistData de R. Suponer que los padres altos son aquellos que miden más de 176 cm. Utilizar $\alpha = 0.1$.

Primero, se procede a realizar una prueba de hipótesis para la igualdad de varianzas (ya que necesitamos saber si la comparación de las medias se realizará con varianzas desconocidas iguales o distintas),

El planteamiento de hipótesis para igualdad de varianzas está dado por:

$$H_0 : \frac{\sigma_y^2}{\sigma_x^2} = 1 \text{ vs. } H_1 : \frac{\sigma_y^2}{\sigma_x^2} \neq 1$$

donde la región de rechazo está dada por

$$C = \left\{ x \in X \left| \left(\frac{\frac{1}{n_1}\sum_{i=1}^{n_1}(y_i - \mu_y)^2}{\frac{1}{n_2}\sum_{i=1}^{n_2}(x_i - \mu_x)^2} \right) > F^{1-\alpha/2}_{(n_1-1,n_2-1)} \right. \right\}$$

donde $\alpha = 0.1$ para una confianza del 90% y n1 = n2 = 928.

Calculando el estadístico se tiene que .

```
alpha <- 0.1

datos <- Galton2 %>%
         filter(parent > 176)

n1 <- nrow(datos)
n2 <- nrow(datos)

est.f <- qf(1 - alpha, df1 = n1 - 1, df2 = n2 - 1)
```

obteniendo así un valor de $F^{1-\alpha/2}_{(n_1-1,n_2-1)} = 1.1553028$.

Realizando los cálculos para la región de rechazo, se obtiene que

```
x <- datos$parent
y <- datos$child

var_child <- sum((y - mean(y))^2)/n1
var_parent <- sum((x - mean(x))^2)/n2

f <- var_child/var_parent
```

$$\left(\frac{\frac{1}{n_1}\sum_{i=1}^{n_1}(y_i - \mu_y)^2}{\frac{1}{n_2}\sum_{i=1}^{n_2}(x_i - \mu_x)^2} \right) = 6.5187405$$

En este caso observamos que:

$$6.5187405 = \left(\frac{\frac{1}{n_1}\sum_{i=1}^{n_1}(y_i - \mu_y)^2}{\frac{1}{n_2}\sum_{i=1}^{n_2}(x_i - \mu_x)^2} \right) > F^{1-\alpha/2}_{(n_1-1,n_2-1)} = 1.1553028$$

por lo tanto se rechaza la hipótesis nula de que ambas poblaciones tienen varianza igual. Es decir, consideramos que las poblaciones (padres e hijos) tienen varianza distinta.

Con lo anterior, se procede a realizar una prueba de hipótesis unilateral (de una cola), para las medias de ambas poblaciones (padres e hijos).

```
t.test(x, y, alternative = "greater", paired = FALSE,
       conf.level = 0.9, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  x and y
## t = 5.8749, df = 410.72, p-value = 4.375e-09
## alternative hypothesis: true difference in means is greater than 0
## 90 percent confidence interval:
##  1.750836      Inf
## sample estimates:
## mean of x mean of y
##  178.3328  176.0925
```

Lo anterior nos indica que, con una cofianza del 90%, podemos afirmar que la media de los padres ($\mu_{\text{parent}} = 178.3328391$) es mayor que la media de los hijos ($\mu_{\text{child}} = 176.092511$).

Ejercicio 2

Repetir el ejercicio anterior para validar la segunda afirmación de Galton:

> *Los hijos de padres bajos no son tan bajos como sus padres.*

Suponer que los padres bajos son aquellos que miden menos de 172 cm.

Procedemos de una manera enteramente análoga al ejercicio anterior, pero ahora lo hacemos solamente para los datos asociados a los padres "bajos" (menores a 172 cm).

Calculando el estadístico se tiene que .

```
alpha <- 0.1

datos <- Galton2 %>%
         filter(parent < 172)

n1 <- nrow(datos)
n2 <- nrow(datos)

est.f <- qf(1 - alpha, df1 = n1 - 1, df2 = n2 - 1)
```

obteniendo así un valor de $F^{1-\alpha/2}_{(n_1-1,n_2-1)} = 1.1385438$.

Realizando los cálculos para la región de rechazo, se obtiene que

```
x <- datos$parent
y <- datos$child

var_child <- sum((y - mean(y))^2)/n1
var_parent <- sum((x - mean(x))^2)/n2

f <- var_child/var_parent
```

$$\left(\frac{\frac{1}{n_1}\sum_{i=1}^{n_1}(y_i - \mu_y)^2}{\frac{1}{n_2}\sum_{i=1}^{n_2}(x_i - \mu_x)^2} \right) = 4.424449$$

En este caso observamos que:

$$4.424449 = \left(\frac{\frac{1}{n_1}\sum_{i=1}^{n_1}(y_i - \mu_y)^2}{\frac{1}{n_2}\sum_{i=1}^{n_2}(x_i - \mu_x)^2} \right) > F^{1-\alpha/2}_{(n_1-1,n_2-1)} = 1.1385438$$

por lo tanto se rechaza la hipótesis nula de que ambas poblaciones tienen varianza igual. Es decir, consideramos que las poblaciones (padres e hijos) tienen varianza distinta.

Con lo anterior, se procede a realizar una prueba de hipótesis unilateral (de una cola), para las medias de ambas poblaciones (padres e hijos).

```
t.test(x, y, alternative = "less", paired = FALSE,
       conf.level = 0.9, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  x and y
## t = -3.6729, df = 559.16, p-value = 0.0001314
## alternative hypothesis: true difference in means is less than 0
## 90 percent confidence interval:
##        -Inf -0.7626884
## sample estimates:
## mean of x mean of y
##  169.3247  170.4969
```

Lo anterior nos indica que, con una cofianza del 90%, podemos afirmar que la media de los padres ($\mu_{\text{parent}} = 169.3246939$) es menor que la media de los hijos ($\mu_{\text{child}} = 170.496852$).

Ejercicio 3

Utilizar las expresiones obtenidas en clase para calcular las estimaciones de β_0, β_1 y σ^2 con el conjunto de datos de estaturas de Galton.

En la clase se obtuvieron las siguientes expresiones para los estimadores del módelo de regresión lineal simple:

$$\hat{\beta}_0 = \bar{y} - \frac{S_{xy}}{S_{xx}}\bar{x}, \qquad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \qquad \text{y} \qquad \hat{\sigma}^2_{\text{MCO}} = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{1}$$

Para el caso de los datos de estaturas de Galton, se utilizará como varaible no aleatoria (X) la estatura de los padres y como variable aleatoria continua (Y) la estatura de los hijos. Obteniendo así, los siguientes valores de los estimadores:

Tabla 1: Estimadores del modelo de regresión lineal simple

$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\sigma}^2_{\text{MCO}}$
60.81149	0.6462906	32.32957

Ejercicio 4

El modelo de regresión lineal simple (RLS) sin intercepto establece que

$$Y_i = \beta x_i + \varepsilon_i, \qquad i = 1, \ldots, n \tag{2}$$

donde los ε_i son errores aleatorios, con media 0 y varianza σ^2. Obtener el estimador de MCO de β.

Demostración. La estimación del valor esperado para el modelo de regresión lineal simple (RLS) sin intercepto está dada por:

$$\hat{y}_i = bx_i, \qquad i = 1, \ldots, n$$

Usando el método de mínimos cuadrados ordinarios (MCO), buscamos minimizar la función suma de cuadrados de los residuos en términos del coeficiente b, dado en la ecuación anterior.

$$Q(b) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - bx_i)^2$$

Derivando la función anterior se obtiene:

$$\begin{aligned} Q'(b) &= \sum_{i=1}^{n} 2\,(y_i - bx_i) \cdot (-x_i) \\ &= -2 \sum_{i=1}^{n} \left(y_i x_i - b x_i^2 \right) \\ &= -2 \sum_{i=1}^{n} y_i x_i + b \sum_{i=1}^{n} x_i^2 \end{aligned}$$

Para obtener los puntos críticos se iguala la derivada anterior a cero:

$$0 = Q'(b) = -2 \sum_{i=1}^{n} y_i x_i + b \sum_{i=1}^{n} x_i^2 \qquad \text{si y solo si} \qquad b \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

por ende

$$b = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

es un punto crítico para la función suma de cuadrados de los residuos $Q(b)$.

Verificamos que sea un mínimo usando el criterio de la segunda derivada:

$$Q''(b) = -2 \sum_{i=1}^{n} \left(-x_i^2 \right) = 2 \sum_{i=1}^{n} x_i^2$$

entonces se tiene que $Q'' \left(\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \right) > 0$, onteniendo así que es el valor que minimiza la función.

Por lo tanto, se tiene que el estimador MCO de β está dado por:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

□

Ejercicio 5

Utilizar el conjunto de datos **Galton** del paquete HistData de R para responder lo siguiente.

Inciso 5.a

Ajustar un modelo RLS sin intercepto. Reportar la estimación de β

Usando la expresión obtenida en el ejercicio 4, se procede a calcular el estimador de β.

Para el caso de los datos de estaturas de Galton, se utilizará como varaible no aleatoria (X) la estatura de los padres y como variable aleatoria continua (Y) la estatura de los hijos. Obteniendo así el siguiente valor para el estimador de β.

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} = 0.9965439$$

Inciso 5.b

Si utilizamos la suma de cuadrados de los residuos como criterio de comparación de modelos, ¿qué modelo *ajusta* mejor a los datos? ¿RLS con o sin intercepto?

Utilizando los coeficientes del modelo de regresión lineal simple con intercepto $\left(\hat{\beta}_0, \hat{\beta}_1\right)$ obtenidos en el ejercicio 3 y el coeficiente del modelo de regresión lineal simple sin intercepto $\left(\hat{\beta}\right)$ obtenido en el ejercicio 5.a, se procede a calcular la suma de los cuadrados de los residuos para cada modelo:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$$

Tabla 2: Suma de los cuadrados de los residuos $\left(\sum_{i=1}^{n} e_i^2\right)$ del modelo de regresión lineal simple

Con intercepto	Sin intercepto
29937.18	32282.6

De la tabla anterior, se aprecia que el modelo con intercepto tiene una menor suma de cuadrado de los residuos. Por lo tanto, podemos decir que el modelo RLS con intercepto *ajusta* mejor a los datos.

Ejercicio 6

Mostrar las siguientes igualdades

Inciso 6.a

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \tag{3}$$

Demostración. Usando la definición de S_{xx} se tiene que:

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

desarrollando el binomio cuadrado, se obtiene

$$= \sum_{i=1}^{n} \left(x_i^2 - 2\bar{x}x_i + \bar{x}^2 \right),$$

distribuyendo la suma sobre cada uno de los términos, se obtiene

$$= \sum_{i=1}^{n} x_i^2 - 2\bar{x} \sum_{i=1}^{n} x_i + \bar{x}^2 \sum_{i=1}^{n} 1,$$

multiplicando el segundo término por un 1 dn la forma $\frac{n}{n}$, se obtiene

$$= \sum_{i=1}^{n} x_i^2 - 2n\bar{x} \left(\frac{1}{n} \sum_{i=1}^{n} x_i \right) + n\bar{x}^2,$$

aplicando la definición de promedio, se obtiene

$$= \sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2,$$

$$= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

□

Inciso 6.b

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} \tag{4}$$

Demostración. Usando la definición de S_{xy} se tiene que:

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

desarrollando el producto de binomios, se obtiene

$$= \sum_{i=1}^{n} (x_i y_i - \bar{y}x_i - \bar{x}y_i + \bar{x}\bar{y}),$$

distribuyendo la suma sobre cada uno de los términos, se obtiene

$$= \sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i - \bar{x}\sum_{i=1}^{n} y_i + \bar{x}\bar{y}\sum_{i=1}^{n} 1,$$

multiplicando el segundo y tercer término un 1 de la forma $\frac{n}{n}$, se obtiene

$$= \sum_{i=1}^{n} x_i y_i - n\bar{y}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) - n\bar{x}\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) + n\bar{x}\bar{y},$$

aplicando la definición de promedio, se obtiene

$$= \sum_{i=1}^{n} x_i^2 - n\bar{y}\bar{x} - n\bar{x}\bar{y} + n\bar{x}\bar{y},$$

$$= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

□