Pruebas de hipótesis para el modelo de regresión lineal múltiple

En esta sección suponemos que $\boldsymbol{y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I})$ con $\boldsymbol{X}$ $n \times (k+1)$ dimensional de rango $k+1 < n$.

Prueba de regresión general

Empezamos intereándonos en la hipótesis de que ninguno de los covariables x considerados predicen a la variable de interés y en el modelo de regresión lineal múltiple con errores normales. En términos matemáticos esto significa que $\boldsymbol{\beta}_1 = (\beta_1, \beta_2, \dots, \beta_k)' = \boldsymbol{0}$. Obtenemos la prueba de hipótesis:

$$H_0 \ : \ \boldsymbol{\beta}_1 = \boldsymbol{0} \qquad \text{v.s.} \qquad H_1 \ : \ \boldsymbol{\beta}_1 \neq \boldsymbol{0}.$$

En las notas correspondientes a las distribuciones no centrales se demostró que para $\lambda_1 = \frac{\boldsymbol{\beta}_1' \boldsymbol{X}_c' \boldsymbol{X}_c \boldsymbol{\beta}_1}{2\sigma^2}$

$$F = \frac{SSR/\left(\sigma^2 k\right)}{SSE/\left(\sigma^2(n-k-1)\right)} = \frac{SSR/k}{SSE/(n-k-1)} \sim F(k, n-k-1, \lambda_1)$$

Teorema

Si H_0 es cierta, es decir $\boldsymbol{\beta}_1 = \boldsymbol{0}$, entonces $\lambda_1 = 0$ y $F \sim F(k, n-k-1)$; y si H_0 es falsa, es decir $\boldsymbol{\beta}_1 \neq \boldsymbol{0}$, entonces $\lambda_1 = \frac{\boldsymbol{\beta}_1' \boldsymbol{X}_c' \boldsymbol{X}_c \boldsymbol{\beta}_1}{2\sigma^2}$ y $F \sim F(k, n-k-1, \lambda_1)$.

Observe que $\lambda_1 = 0$ si y sólo si $\boldsymbol{\beta}_1 = \boldsymbol{0}$ dado que $\boldsymbol{X}_c' \boldsymbol{X}_c$ es positio definida. Podemos usar F como cantida pivotal para realizar la prueba de hipótesis como sigue:

Si $F > F_{\alpha,k,n-k-1}$, con $F_{\alpha,k,n-k-1}$ tal que $\mathbb{P}\left[F > F_{\alpha,k,n-k-1} | \boldsymbol{\beta}_1 = \boldsymbol{0}\right] = \alpha$, entonces rechazamos H_0.

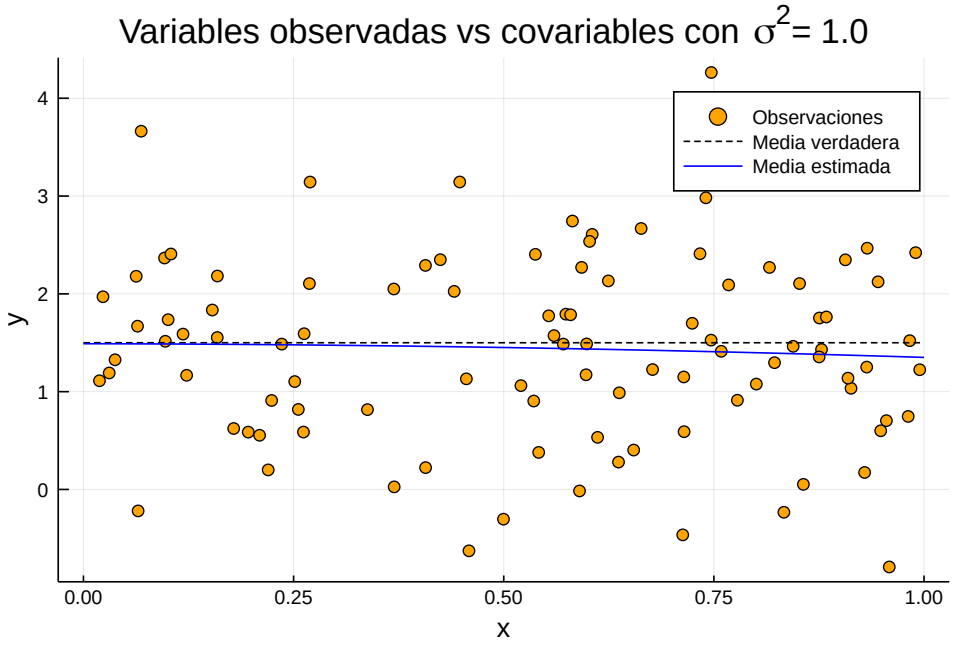
A continuación consideramos la prueba anterior para el caso i) en que $\boldsymbol{\beta}_1 = \boldsymbol{0}$ y ajustamos un modelo con $\boldsymbol{\beta}_1 \neq \boldsymbol{0}$; y para el caso ii) en que $\boldsymbol{\beta}_1 \neq \boldsymbol{0}$ y ajustamos un modelo con $\boldsymbol{\beta}_1 \neq \boldsymbol{0}$.

In [2]:
```
using Distributions # Paquete con distribuciones de probabilidad
using Plots # Paquete para producir imágenes
using LaTeXStrings # Paquete para usar latex en strings
```

In [4]:
```
n = 100 # Consideramos 100 observaciones
x = rand(Uniform(),n) # n puntos aleatorios uniformes en (0,1), esta sería la pri
X = zeros(n,3) # Matriz de ceros para construir X
for i in 1:n
    X[i,:] = [ 1, x[i], x[i]^2 ] # iteramos para construir renglones de X
end
```

In [6]:
```
ϵ = rand(Normal(0,1.0),n) # Vector de errores normales con media μ=0 y varianza σ
y = 1.5 .+ ϵ # Observaciones provenientes del modelo con varianza 0.1
β_ml = ( X' * X)^(-1) * X' * y
f(x) = 1.5
f_ml(x) = β_ml[1] + β_ml[2]*x + β_ml[3]*x^2.0
mesh = collect(0.0:1.0/100.0:1.0)
scatter(x,y,color="orange",label="Observaciones")
plot!(mesh,f.(mesh), color = :black, linestyle=:dash ,label="Media verdadera")
plot!(mesh,f_ml.(mesh), color = :blue, label="Media estimada")
ylabel!("y")
xlabel!("x")
title!("Variables observadas vs covariables con \\sigma^{2}  = 1.0")
```

Out[6]:



Si calculamos el estadístico de prueba F

In [11]:
```
SSR = β_ml' * X' * y - n*mean(y)^2.0
SSE = y'*y - β_ml' * X' * y
F = (n-2-1)*SSR/( 2* SSE )
Fα = cquantile( FDist( 2, n-2-1), 0.05 );
```

Out[11]:
```
3.0901866751548672
```

Vemos que

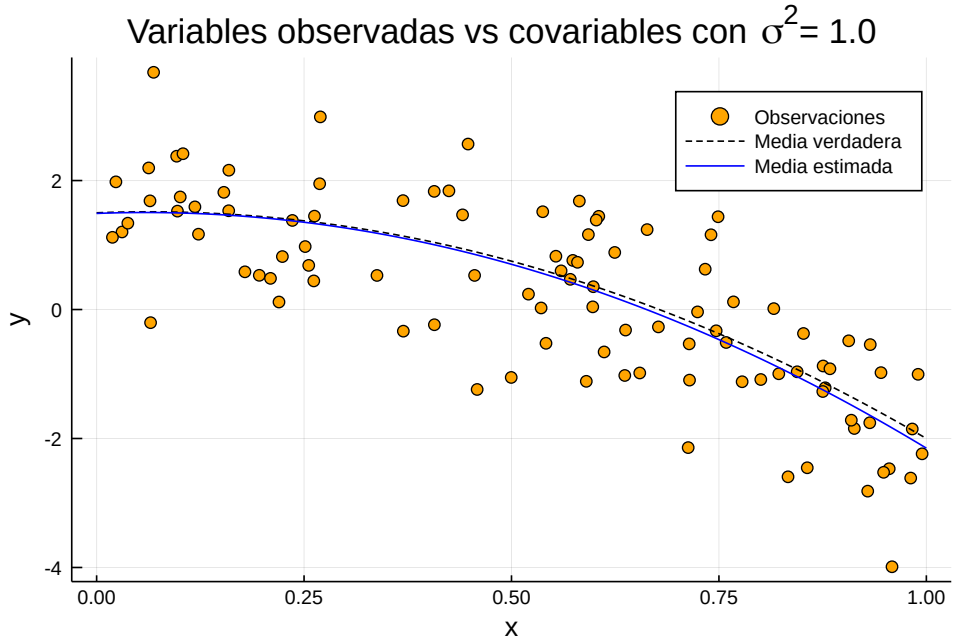
In [14]:
```
F >= Fα
```

Out[14]:
```
false
```

Por lo que no rechazamos $H_0 : \boldsymbol{\beta}_1 = \mathbf{0}$.

In [18]:
```
y = 1.5 .+ 0.5.*x .- 4.0.*x.^2.0 .+ ϵ # Observaciones provenientes del modelo con
β_ml = ( X' * X)^(-1) * X' * y
f(x) = 1.5 + 0.5*x - 4.0.*x^2.0
f_ml(x) = β_ml[1] + β_ml[2]*x + β_ml[3]*x^2.0
mesh = collect(0.0:1.0/100.0:1.0)
scatter(x,y,color="orange",label="Observaciones")
plot!(mesh,f.(mesh), color = :black, linestyle=:dash ,label="Media verdadera")
plot!(mesh,f_ml.(mesh), color = :blue, label="Media estimada")
ylabel!("y")
xlabel!("x")
title!("Variables observadas vs covariables con \\sigma^{2}  = 1.0")
```

Out[18]:



In [19]:
```
SSR = β_ml' * X' * y - n*mean(y)^2.0
SSE = y'*y - β_ml' * X' * y
F = (n-2-1)*SSR/( 2* SSE )
Fα = cquantile( FDist( 2, n-2-1), 0.05 );
```

Vemos que

In [20]:
```
F >= Fα
```

Out[20]:
```
true
```

Por lo que rechazamos $H_0 : \boldsymbol{\beta}_1 = \mathbf{0}$.

Prueba para $H_0 : C\beta = t$.

Consideramos la prueba de hipótesis:

$$H_0 : \boldsymbol{C\beta} = \boldsymbol{t} \quad \text{v.s.} \quad H_1 : \boldsymbol{C\beta} \neq \boldsymbol{t}.$$

Está prueba al ser muy general nos será de gran utilidad para hacer pruebas de hipótesis en el modelo de regresión lineal múltiple. Por ejemplo, esta prueba nos permite hacer hipótesis de la forma $H_0 : \beta_1 = \beta_2 + 5$. En lo siguiente suponemos que el sistema de equaciones $\boldsymbol{C\beta} = \boldsymbol{t}$ tiene por lo menos una solución, en β. Para plantear una prueba de hipótesis como en la sección anterior, ocupamos del siguiente teorema.

Teorema

Si $\boldsymbol{y} \sim N_n(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$ y $\boldsymbol{C}$ es $q \times (k+1)$ dimensional de rango $q \leq k+1$, entonces

1. $\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{t} \sim N_q(\boldsymbol{C\beta} - \boldsymbol{t}, \sigma^2\boldsymbol{C}\left(\boldsymbol{X'X}\right)^{-1}\boldsymbol{C'})$.
2. $\frac{SSH}{\sigma^2} = \frac{(\boldsymbol{C}\hat{\boldsymbol{\beta}}-\boldsymbol{t})'\left(\boldsymbol{C}(\boldsymbol{X'X})^{-1}\boldsymbol{C'}\right)^{-1}(\boldsymbol{C}\hat{\boldsymbol{\beta}}-\boldsymbol{t})}{\sigma^2} \sim \chi^2(q, \lambda)$ con $\lambda = \frac{(\boldsymbol{C\beta}-\boldsymbol{t})'\left(\boldsymbol{C}(\boldsymbol{X'X})^{-1}\boldsymbol{C'}\right)^{-1}(\boldsymbol{C\beta}-\boldsymbol{t})}{2\sigma^2}$
3. SSH y $SSE = \boldsymbol{y}'\left(\boldsymbol{I} - \boldsymbol{X}\left(\boldsymbol{X'X}\right)^{-1}\boldsymbol{X'}\right)\boldsymbol{y}$ son indepednientes.

dem

1. Sabemos que en este caso $\hat{\boldsymbol{\beta}} \sim N_{k+1}(\boldsymbol{\beta}, \sigma^2\left(\boldsymbol{X'X}\right)^{-1})$ por lo que, usando la distribución de $\boldsymbol{Ay}$ dado $\boldsymbol{y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ vista en clase, $\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{t} \sim N_q(\boldsymbol{C\beta} - \boldsymbol{t}, \sigma^2\boldsymbol{C}\left(\boldsymbol{X'X}\right)^{-1}\boldsymbol{C'})$.
2. Por un teorema visto en clase, vea notebook sobre distribuciones no centrales, basta ver que $\boldsymbol{B} = \text{Cov}\left(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{t}\right)\frac{1}{\sigma^2}\left(\boldsymbol{C}\left(\boldsymbol{X'X}\right)^{-1}\boldsymbol{C'}\right)^{-1}$ es idempotente de rango q; pero vemos que $\boldsymbol{B} = \sigma^2\boldsymbol{C}\left(\boldsymbol{X'X}\right)^{-1}\boldsymbol{C'})\frac{1}{\sigma^2}\left(\boldsymbol{C}\left(\boldsymbol{X'X}\right)^{-1}\boldsymbol{C'}\right)^{-1} = \boldsymbol{I}_q$ es idempotente de rango q.
3. Sabemos que $\frac{SSE}{\sigma^2} \sim \chi^2(n-k-1)$ es independiente de $\hat{\boldsymbol{\beta}}$, de lo que se sigue que $\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{t}$ es independiente de SSE y por lo tanto SSH es indepedniente de SSE.

Del teorema anterior se sigue que para $\lambda = (\boldsymbol{C\beta} - \boldsymbol{t})'\left(\boldsymbol{C}\left(\boldsymbol{X'X}\right)^{-1}\boldsymbol{C'}\right)^{-1}(\boldsymbol{C\beta} - \boldsymbol{t})/2\sigma^2$

$$F = \frac{SSH\,/\left(\sigma^2 q\right)}{SSE\,/\left(\sigma^2(n-k-1)\right)} = \frac{SSH\,/q}{SSE\,/(n-k-1)} \sim F(q, n-k-1, \lambda)$$

Teorema

Si H_0 es cierta, es decir $\boldsymbol{C\beta} = \boldsymbol{t}$, entonces $\lambda = 0$ y $F \sim F(q, n-k-1)$; y si H_0 es falsa, es decir $\boldsymbol{C\beta} \neq \boldsymbol{t}$, entonces $\lambda = \frac{(\boldsymbol{C\beta}-\boldsymbol{t})'\left(\boldsymbol{C}(\boldsymbol{X'X})^{-1}\boldsymbol{C'}\right)^{-1}(\boldsymbol{C\beta}-\boldsymbol{t})}{2\sigma^2}$ y $F \sim F(q, n-k-1, \lambda)$.

Podemos usar F como cantidad pivotal para realizar la prueba de hipótesis como sigue:

Si $F > F_{\alpha,q,n-k-1}$, con $F_{\alpha,q,n-k-1}$ tal que $\mathbb{P}\left[F > F_{\alpha,q,n-k-1} | H_0\right] = \alpha$, entonces rechazamos H_0.

Si queremos realizar una prueba de hipótesis del estilo

$$H_0 \ : \ \boldsymbol{a}'\boldsymbol{\beta} = 0 \qquad \text{v.s.} \qquad H_1 \ : \ \boldsymbol{a}'\boldsymbol{\beta} \neq 0$$

entonces tenemos el estadístico de prueba

$$F = \frac{\left(\boldsymbol{a}'\hat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{a}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{a}\right)^{-1}\boldsymbol{a}'\hat{\boldsymbol{\beta}}}{s^2} = \frac{\left(\boldsymbol{a}'\hat{\boldsymbol{\beta}}\right)^2}{s^2\boldsymbol{a}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{a}} \sim F\left(1, n-k-1, \boldsymbol{\beta}'\boldsymbol{a}\left(\boldsymbol{a}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{a}\right)^{-1}\boldsymbol{a}'$$

$$\sim F\left(1, n-k-1, \frac{\left(\boldsymbol{a}'\boldsymbol{\beta}\right)^2}{2\sigma^2\boldsymbol{a}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{a}}\right)$$

donde $s^2 = SSE/(n-k-1)$.

Para realizar la prueba $H_0 \ : \ \beta_j = 0 \qquad \text{v.s.} \qquad H_1 \ : \ \beta_j \neq 0$ tomamos $\boldsymbol{a}' = (0, \ldots, 0, 1, 0, \ldots, 0) = \boldsymbol{e}_j$ arriba para obtener el estadístico de prueba

$$F = \frac{\hat{\beta}_j^2}{s^2 g_{j,j}} \sim F\left(1, n-k-1, \frac{\beta_j^2}{2\sigma^2 g_{j,j}}\right)$$

con $g_{j,j}$ el j-ésimo elemento diagonal de $\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$. Rechazamos $H_0 \ : \ \beta_j = 0$ si $F \geq F_{\alpha,1,n-k-1}$.

Equivalentemente, al tener nuestro estadístico de prueba F con 1 y $n-k-1$ grados de libertad, podemos utilizar

$$t_j = \frac{\hat{\beta}_j}{s\sqrt{g_{j,j}}} \sim t\left(n-k-1, \frac{\beta_j}{2\sigma^2 g_{j,j}}\right),$$

donde hemos usado el teorema ante-anterior para estandarizar a $\hat{\beta}_j$.

Rechazamos $H_0 \ : \ \beta_j = 0$ si $|t_j| \geq t_{\alpha/2,n-k-1}$.

In [ ]: