Modelos Lineales

Inferencia

Fernando Massa; Bruno Bellagamba

jueves 11 de abril 2024

Objetivos

- Continuaremos con los casos particulares de las pruebas de hipótesis lineales.
- Significación de una variable.
- Significación de un subconjunto de variables.
- Significación de una combinación lineal.

Repaso

La clase pasada:

1. Planteamos los estimadores máximo verosímiles (MV).

Repaso

La clase pasada:

1. Planteamos los estimadores máximo verosímiles (MV).
2. Vimos la distribución de $\hat{\beta}_{MV}$.

La clase pasada:

1. Planteamos los estimadores máximo verosímiles (MV).
2. Vimos la distribución de $\hat{\beta}_{MV}$.
3. Vimos la distribución de $\hat{\sigma}^2$.

Repaso

La clase pasada:

1. Planteamos los estimadores máximo verosímiles (MV).
2. Vimos la distribución de $\hat{\beta}_{MV}$.
3. Vimos la distribución de $\hat{\sigma}^2$.
4. Reconocimos su independencia.

La clase pasada:

1. Planteamos los estimadores máximo verosímiles (MV).
2. Vimos la distribución de $\hat{\beta}_{MV}$.
3. Vimos la distribución de $\hat{\sigma}^2$.
4. Reconocimos su independencia.
5. Planteamos hipótesis lineales.

Repaso

La clase pasada:

1. Planteamos los estimadores máximo verosímiles (MV).
2. Vimos la distribución de $\hat{\beta}_{MV}$.
3. Vimos la distribución de $\hat{\sigma}^2$.
4. Reconocimos su independencia.
5. Planteamos hipótesis lineales.
6. A partir de todo lo anterior, planteamos el estadístico F.

La clase pasada:

1. Planteamos los estimadores máximo verosímiles (MV).
2. Vimos la distribución de $\hat{\beta}_{MV}$.
3. Vimos la distribución de $\hat{\sigma}^2$.
4. Reconocimos su independencia.
5. Planteamos hipótesis lineales.
6. A partir de todo lo anterior, planteamos el estadístico F.
7. Vimos el primer caso particular, la significación global del modelo.

Repaso

La clase pasada:

1. Planteamos los estimadores máximo verosímiles (MV).
2. Vimos la distribución de $\hat{\beta}_{MV}$.
3. Vimos la distribución de $\hat{\sigma}^2$.
4. Reconocimos su independencia.
5. Planteamos hipótesis lineales.
6. A partir de todo lo anterior, planteamos el estadístico F.
7. Vimos el primer caso particular, la significación global del modelo.

Hoy continuaremos con los siguientes casos particulares.

Repaso

Las pruebas de hipótesis con las que trabajaremos tienen la forma general:

$$\begin{array}{lccc} H_0) & C\beta & = & b \\ H_1) & C\beta & \neq & b \end{array}$$

Las pruebas de hipótesis con las que trabajaremos tienen la forma general:

$$\begin{array}{lccc} H_0) & C\beta & = & b \\ H_1) & C\beta & \neq & b \end{array}$$

Y el estadístico de prueba tenía la forma:

$$F = \frac{\left(C\hat{\beta}_{MV} - b\right)^{'}\left[C\left(X^{'}X\right)^{-1}C^{'}\right]^{-1}\left(C\hat{\beta}_{MV} - b\right)/q}{\hat{\sigma}^2}$$

Donde el tipo de prueba viene dado por la elección de las matrices C y b.

La idea de hoy será transitar a lo largo de estas pruebas con un modelo de *RLM* correspondiente a un cierto caso.

La idea de hoy será transitar a lo largo de estas pruebas con un modelo de *RLM* correspondiente a un cierto caso.

La idea de hoy será transitar a lo largo de estas pruebas con un modelo de *RLM* correspondiente a un cierto caso.

Los datos corresponden a características de un conjunto de automóviles de 1974.

Vayamos al R

Significación individual

Una vez determinada la significación global del modelo, suele ser pertinente examinar la **significación individual** de cada una de las variables explicativas.

Una vez determinada la significación global del modelo, suele ser pertinente examinar la **significación individual** de cada una de las variables explicativas.

En términos del modelo, esto implica llevar a cabo pruebas del estilo.

$$\begin{array}{lccc} H_0) & \beta_j & = & 0 \\ H_1) & \beta_j & \neq & 0 \end{array}$$

En el entendido de que, si la hipótesis nula es cierta, el coeficiente de la variable X_j es 0. De esta manera, el NO rechazo de $H_0)$ implicaría que no existe evidencia suficiente para corroborar que la variable explicativa en cuestión tenga un aporte *significativo* al modelo.

Las matrices correspondientes a las hipótesis lineales:

$$C = (0, 0, \ldots, 0, 1, 0, \ldots, 0) \text{ y } b = 0$$

Siendo la matriz C un vector de ceros $(q = 1)$ con un 1 en el j-ésimo lugar y el vector b un escalar.

Las matrices correspondientes a las hipótesis lineales:

$$C = (0, 0, \ldots, 0, 1, 0, \ldots, 0) \text{ y } b = 0$$

Siendo la matriz C un vector de ceros $(q = 1)$ con un 1 en el j-ésimo lugar y el vector b un escalar.

En este caso, el estadístico de prueba pasa a ser:

$$F = \frac{\left(\hat{\beta}_j - b\right)^2}{g_{jj}\hat{\sigma}^2}$$

Siendo g_{jj} el j-ésimo elemento de diagonal de la matriz $\left(X'X\right)^{-1}$

Sustituyendo el valor de $\hat{\sigma}^2$ en la expresión anterior:

$$F = \frac{\left(\hat{\beta}_j - b\right)^2 / 1}{g_{jj} \frac{SCRes}{n-k-1}}$$

Sustituyendo el valor de $\hat{\sigma}^2$ en la expresión anterior:

$$F = \frac{\left(\hat{\beta}_j - b\right)^2 / 1}{g_{jj} \frac{SCRes}{n-k-1}}$$

De esta manera se observa que la distribución del estadístico es $F_{1,n-k-1}$

Significación individual

Sustituyendo el valor de $\hat{\sigma}^2$ en la expresión anterior:

$$F = \frac{\left(\hat{\beta}_j - b\right)^2 / 1}{g_{jj} \frac{SCRes}{n-k-1}}$$

De esta manera se observa que la distribución del estadístico es $F_{1,n-k-1}$

Relación entre F y t

Sea X una variable aleatoria con distribución $F_{1,\nu}$, entonces $\sqrt{X}$ tiene distribución t_ν

Sustituyendo el valor de $\hat{\sigma}^2$ en la expresión anterior:

$$F = \frac{\left(\hat{\beta}_j - b\right)^2 / 1}{g_{jj}\frac{SCRes}{n-k-1}}$$

De esta manera se observa que la distribución del estadístico es $F_{1,n-k-1}$

Relación entre F y t

Sea X una variable aleatoria con distribución $F_{1,\nu}$, entonces $\sqrt{X}$ tiene distribución t_ν

Distribución t_ν

Sea X una variable aleatoria con distribución $N(0,1)$ y Y otra variable aleatoria **independiente** de X con distribución χ^2_ν entonces:

$$\frac{X}{\sqrt{Y/\nu}} \sim t_\nu$$

En nuestro caso es fácil comprobar este resultado si se multiplican el numerador y denominador por σ^2.

$$\sqrt{F} = \frac{\frac{\left(\hat{\beta}_j - b\right)}{\sqrt{g_{jj}}\sigma}}{\sqrt{\frac{SCRes}{\sigma^2(n-k-1)}}} = \frac{N(0,1)}{\sqrt{\chi^2_{n-k-1}/(n-k-1)}}$$

En nuestro caso es fácil comprobar este resultado si se multiplican el numerador y denominador por σ^2.

$$\sqrt{F} = \frac{\frac{\left(\hat{\beta}_j - b\right)}{\sqrt{g_{jj}}\sigma}}{\sqrt{\frac{SCRes}{\sigma^2(n-k-1)}}} = \frac{N(0,1)}{\sqrt{\chi^2_{n-k-1}/(n-k-1)}}$$

Esta expresión cumple las 3 condiciones del teorema anterior:

- Bajo H_0 cierta, el numerador de esta expresión tiene distribución $N(0,1)$.
- El denominador es la raíz cuadrada de un estadístico con distribución χ^2_{n-k-1} dividido entre sus grados de libertad.
- Adicionalmente, numerador y denominador son independientes debido a que $\hat{\beta}_{MV}$ y $SCRes$ son independientes.

En nuestro caso es fácil comprobar este resultado si se multiplican el numerador y denominador por σ^2.

$$\sqrt{F} = \frac{\frac{\left(\hat{\beta}_j - b\right)}{\sqrt{g_{jj}}\sigma}}{\sqrt{\frac{SCRes}{\sigma^2(n-k-1)}}} = \frac{N(0,1)}{\sqrt{\chi^2_{n-k-1}/(n-k-1)}}$$

Esta expresión cumple las 3 condiciones del teorema anterior:

- Bajo H_0 cierta, el numerador de esta expresión tiene distribución $N(0,1)$.
- El denominador es la raíz cuadrada de un estadístico con distribución χ^2_{n-k-1} dividido entre sus grados de libertad.
- Adicionalmente, numerador y denominador son independientes debido a que $\hat{\beta}_{MV}$ y $SCRes$ son independientes.

Entonces se comprueba que el estadístico tiene distribución t_{n-k-1}.

En resumen, a la hora de llevar a cabo la prueba

$$\begin{array}{llcl} H_0) & \beta_j & = & b \\ H_1) & \beta_j & \neq & b \end{array}$$

Da lo mismo utilizar el estadístico:

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

que usar el estadístico

$$t = \frac{\left(\hat{\beta}_j - b\right)}{\sqrt{g_{jj}\hat{\sigma}^2}} \sim t_{n-k-1}$$

Ya que (a través de distribuciones diferentes) ambos llevan al mismo p-valor.

Ejemplo

A partir de lo anterior, estamos en condiciones de leer e interpretar gran parte del *summary* del modelo.

A partir de lo anterior, estamos en condiciones de leer e interpretar gran parte del *summary* del modelo.

Volvamos a los datos de los autos en R.

Con los temas que hemos visto hasta aquí en el curso, podemos identificar varios elementos en la salida de la función *summary* del R.

Medidas de resumen de los residuos

```
Call:
lm(formula = mpg ~ wt + qsec + am + carb, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1184 -1.5414 -0.1392  1.2917  4.3604

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.8972     7.4725   1.726 0.095784 .
wt           -3.4343     0.8200  -4.188 0.000269 ***
qsec          1.0191     0.3378   3.017 0.005507 **
am            3.5114     1.4875   2.361 0.025721 *
carb         -0.4886     0.4212  -1.160 0.256212
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.444 on 27 degrees of freedom
Multiple R-squared:  0.8568,	Adjusted R-squared:  0.8356
F-statistic: 40.39 on 4 and 27 DF,  p-value: 5.064e-11
```

Algunas estadísticas descriptivas de los residuos del modelo.

Con los temas que hemos visto hasta aquí en el curso, podemos identificar varios elementos en la salida de la función *summary* del R.

```
Call:
lm(formula = mpg ~ wt + qsec + am + carb, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1184 -1.5414 -0.1392  1.2917  4.3604

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.8972     7.4725   1.726 0.095784 .
wt           -3.4343     0.8200  -4.188 0.000269 ***
qsec          1.0191     0.3378   3.017 0.005507 **
am            3.5114     1.4875   2.361 0.025721 *
carb         -0.4886     0.4212  -1.160 0.256212
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.444 on 27 degrees of freedom
Multiple R-squared:  0.8568,    Adjusted R-squared:  0.8356
F-statistic: 40.39 on 4 and 27 DF,  p-value: 5.064e-11
```

$n-k-1$

k

El número de observaciones (n) y variables (k) que conforman los grados de libertad de los estadísticos de prueba.

Con los temas que hemos visto hasta aquí en el curso, podemos identificar varios elementos en la salida de la función *summary* del R.

R^2

```
Call:
lm(formula = mpg ~ wt + qsec + am + carb, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.1184 -1.5414 -0.1392  1.2917  4.3604 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  12.8972     7.4725   1.726 0.095784 .  
wt           -3.4343     0.8200  -4.188 0.000269 ***
qsec          1.0191     0.3378   3.017 0.005507 ** 
am            3.5114     1.4875   2.361 0.025721 *  
carb         -0.4886     0.4212  -1.160 0.256212    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.444 on 27 degrees of freedom
Multiple R-squared:  0.8568,	Adjusted R-squared:  0.8356 
F-statistic: 40.39 on 4 and 27 DF,  p-value: 5.064e-11
```

El valor del R^2.

Con los temas que hemos visto hasta aquí en el curso, podemos identificar varios elementos en la salida de la función *summary* del R.

$\sqrt{\hat{\sigma}^2}$

```
Call:
lm(formula = mpg ~ wt + qsec + am + carb, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1184 -1.5414 -0.1392  1.2917  4.3604

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.8972     7.4725   1.726 0.095784 .
wt           -3.4343     0.8200  -4.188 0.000269 ***
qsec          1.0191     0.3378   3.017 0.005507 **
am            3.5114     1.4875   2.361 0.025721 *
carb         -0.4886     0.4212  -1.160 0.256212
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.444 on 27 degrees of freedom
Multiple R-squared:  0.8568,    Adjusted R-squared:  0.8356
F-statistic: 40.39 on 4 and 27 DF,  p-value: 5.064e-11
```

La estimación del desvío standard de los residuos del modelo (cuidado, el R presenta a $\hat{\sigma}$ en lugar de $\hat{\sigma}^2$).

summary DEL R

Con los temas que hemos visto hasta aquí en el curso, podemos identificar varios elementos en la salida de la función *summary* del R.

```
Call:
lm(formula = mpg ~ wt + qsec + am + carb, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1184 -1.5414 -0.1392  1.2917  4.3604

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.8972     7.4725   1.726 0.095784 .
wt           -3.4343     0.8200  -4.188 0.000269 ***
qsec          1.0191     0.3378   3.017 0.005507 **
am            3.5114     1.4875   2.361 0.025721 *
carb         -0.4886     0.4212  -1.160 0.256212
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.444 on 27 degrees of freedom
Multiple R-squared:  0.8568,    Adjusted R-squared:  0.8356
F-statistic: 40.39 on 4 and 27 DF,  p-value: 5.064e-11
```

significación global

El valor del estadístico de la prueba de significación global del modelo así como sus grados de libertad y el p-valor de la prueba.

Con los temas que hemos visto hasta aquí en el curso, podemos identificar varios elementos en la salida de la función *summary* del R.

$\hat{\beta}_{MCO} = \hat{\beta}_{MV}$

```
Call:
lm(formula = mpg ~ wt + qsec + am + carb, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1184 -1.5414 -0.1392  1.2917  4.3604

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.8972     7.4725   1.726 0.095784 .
wt           -3.4343     0.8200  -4.188 0.000269 ***
qsec          1.0191     0.3378   3.017 0.005507 **
am            3.5114     1.4875   2.361 0.025721 *
carb         -0.4886     0.4212  -1.160 0.256212
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.444 on 27 degrees of freedom
Multiple R-squared:  0.8568,    Adjusted R-squared:  0.8356
F-statistic: 40.39 on 4 and 27 DF,  p-value: 5.064e-11
```

Las estimaciones de los coeficientes del modelo.

Con los temas que hemos visto hasta aquí en el curso, podemos identificar varios elementos en la salida de la función *summary* del R.

```
Call:
lm(formula = mpg ~ wt + qsec + am + carb, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1184 -1.5414 -0.1392  1.2917  4.3604

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.8972     7.4725   1.726 0.095784 .
wt           -3.4343     0.8200  -4.188 0.000269 ***
qsec          1.0191     0.3378   3.017 0.005507 **
am            3.5114     1.4875   2.361 0.025721 *
carb         -0.4886     0.4212  -1.160 0.256212
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.444 on 27 degrees of freedom
Multiple R-squared:  0.8568,	Adjusted R-squared:  0.8356
F-statistic: 40.39 on 4 and 27 DF,  p-value: 5.064e-11
```

Significación individual de $\forall x_j$

Los estadísticos de las pruebas de significación individual junto con sus respectivos grados de libertad. Tengan en cuenta que la columna de los estadísticos t surge de dividir la columna de los $\hat{\beta}_{MV}$ entre sus correspondientes desvíos standard ($s_{\hat{\beta}}$).

Significación de un subconjunto

La secuencia a seguir al analizar la contribución de las variables puede tener un paso más.

La secuencia a seguir al analizar la contribución de las variables puede tener un paso más.

1. **Significancia global**.

   Se lleva a cabo mediante el estadístico $F_{k,n-k-1}$. En caso de decidir que al menos una de las variables explicativas aporta a explicar/predecir a la Y, se llevan a cabo las pruebas individuales.

Significación de un subconjunto

La secuencia a seguir al analizar la contribución de las variables puede tener un paso más.

1. **Significancia global**.

   Se lleva a cabo mediante el estadístico $F_{k,n-k-1}$. En caso de decidir que al menos una de las variables explicativas aporta a explicar/predecir a la Y, se llevan a cabo las pruebas individuales.

2. **Significación individual**.

   Se lleva a cabo mediante k estadísticos t_{n-k-1}. Una a una, se analiza la contribución de cada variable ajustando por el efecto de las demás.

La secuencia a seguir al analizar la contribución de las variables puede tener un paso más.

1. **Significancia global**.

   Se lleva a cabo mediante el estadístico $F_{k,n-k-1}$. En caso de decidir que al menos una de las variables explicativas aporta a explicar/predecir a la Y, se llevan a cabo las pruebas individuales.

2. **Significación individual**.

   Se lleva a cabo mediante k estadísticos t_{n-k-1}. Una a una, se analiza la contribución de cada variable ajustando por el efecto de las demás.

3. **Significación de un subconjunto**.

   Se lleva a cabo mediante un estadístico $F_{q,n-k-1}$. En caso de detectar que un conjunto de variables explicativas no presentan un aporte significativo al modelo, es posible comprobar su aporte conjunto. En caso de no rechazar H_0, es posible simplificar el modelo quitando estas variables.

A la hora de construir el estadístico de prueba, la matriz C se compone a partir de las filas de la matriz empleada en la prueba de significación global reteniendo solamente las filas correspondientes a las variables que se desean testear.

A la hora de construir el estadístico de prueba, la matriz C se compone a partir de las filas de la matriz empleada en la prueba de significación global reteniendo solamente las filas correspondientes a las variables que se desean testear.

A modo de ejemplo, considere el siguiente modelo de RLM.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

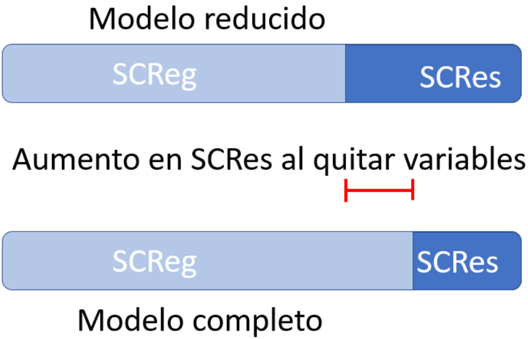
Si se desesa poner a prueba la hipótesis

$$H_0) \quad \beta_2 = \beta_4 = 0$$
$$H_1) \quad \text{al menos uno de ellos es } \neq 0$$

La matriz C adopta la siguiente forma:

$$C = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Una manera gráfica que puede resultar de ayuda para comprender esta prueba es la siguiente:



El objetivo de la prueba, es determinar si ese aumento en *SCRes* (o disminución en *SCReg*) es significativo o no.

Veamos si es posible remover del modelo a las variables *carb* y *am*.

Veamos si es posible remover del modelo a las variables *carb* y *am*.

Volvamos al R.

Este último caso particular del estadístico F será de particular utilidad cuando veamos los modelos de Análisis de Varianza. Pero veámoslo ahora en un contexto más general.

Este último caso particular del estadístico F será de particular utilidad cuando veamos los modelos de Análisis de Varianza. Pero veámoslo ahora en un contexto más general.

La idea fundamental es que en algunas ocasiones interesa contrastar cierta hipótesis sobre algunas relaciones entre los coeficientes del modelo. Supongamos la siguiente situación.

Este último caso particular del estadístico F será de particular utilidad cuando veamos los modelos de Análisis de Varianza. Pero veámoslo ahora en un contexto más general.

La idea fundamental es que en algunas ocasiones interesa contrastar cierta hipótesis sobre algunas relaciones entre los coeficientes del modelo. Supongamos la siguiente situación. A algún investigador podiría interesarle responder la siguiente pregunta:

¿Qué tiene más importancia sobre el salario, un año más de experiencia o un año más de educación

A fin de responder esta pregunta, se podría plantear el siguiente modelo de *RLM*.

$$Ingreso_i = \beta_0 + \beta_1 Educación_i + \beta_2 Experiencia_i + \varepsilon_i$$

La hipótesis a contrastar sería:

A fin de responder esta pregunta, se podría plantear el siguiente modelo de *RLM*.

$$Ingreso_i = \beta_0 + \beta_1 Educación_i + \beta_2 Experiencia_i + \varepsilon_i$$

La hipótesis a contrastar sería:

$$\begin{array}{lccc} H_0) & \beta_1 & = & \beta_2 \\ H_0) & \beta_1 & \neq & \beta_2 \end{array}$$

A fin de responder esta pregunta, se podría plantear el siguiente modelo de *RLM*.

$$Ingreso_i = \beta_0 + \beta_1 Educación_i + \beta_2 Experiencia_i + \varepsilon_i$$

La hipótesis a contrastar sería:

$$\begin{array}{lccc} H_0) & \beta_1 & = & \beta_2 \\ H_0) & \beta_1 & \neq & \beta_2 \end{array}$$

Que en el contexto de hipótesis lineales equivale a emplear la siguiente matriz C.

$$C = (0, 1, -1)$$

Para ilustrar este caso, emplearemos los datos contenidos en el archivo *ingresos.txt*.

Para ilustrar este caso, emplearemos los datos contenidos en el archivo *ingresos.txt*.

Volvamos al R.

En la próxima clase

La próxima hablaremos de:

- Intervalos de confianza:
- para los coeficientes.
- para las predicciones.
- para $\mathbb{E}(\hat{y})$.

Carmona, Francesc (2003). *Modelos Lineales (notas de curso).* Departament d'Estadística.

Farraway, Julian (2014). *Linear Models with R, second edition.* Chapman Hall/CRC.

Rencher, Alvin y Bruce Schaalje (2008). *Linear Models in Statistics, second edition.* John Wiley Sons, Inc.

¿Preguntas?

Muchas Gracias