Modelos de Regresión y Series de Tiempo (MRST)
2025 - 02

Clase 9 – Propiedades de los parámetros estimados, estimación de σ^2, ANOVA y prueba de utilidad

Docente: Natalia Jaramillo Quiceno
Escuela de Ingenierías
natalia.jaramilloq@upb.edu.co

Regresión lineal múltiple
Propiedades de los estimadores de mínimos cuadrados

- El vector $\hat{\beta}$ es un estimador insesgado de β. Lo cual se puede demostrar al evaluar la esperanza de $\hat{\beta}$ así:

$$E\left(\hat{\boldsymbol{\beta}}\right) = E\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\right] = E\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})\right]$$

$$= E\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\right] = \boldsymbol{\beta}$$

Considerando que $E(\boldsymbol{\varepsilon}) = \mathbf{0}$

y $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} = \mathbf{I}$

- La $Var(\hat{\beta}_i) < Var(\beta_i^*)$ para cualquier β_i^* diferente al obtenido por mínimos cuadrados (mínima varianza)

- La matriz de varianzas-covarianzas de $\hat{\beta}$ está data por: $\mathrm{Cov}\left(\hat{\boldsymbol{\beta}}\right) = \mathrm{Var}\left(\hat{\boldsymbol{\beta}}\right) = \mathrm{Var}\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\right]$

$$= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

Así, tendríamos que…

$$\widehat{\boldsymbol{\beta}} \sim \boldsymbol{N}(\boldsymbol{\beta}, \boldsymbol{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$$

Regresión lineal múltiple

Estimación de σ^2

Así como en la RLS, se puede desarrollar un estimador de σ^2 a partir de la suma de cuadrados de los residuales:

$$SS_{\text{Res}} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 = \mathbf{e}'\mathbf{e}$$

Al sustituir $\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ se obtiene lo siguiente:

$$SS_{\text{Res}} = \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)'\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right) = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \mathbf{y}'\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y}'\mathbf{y} - 2\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}$$

Ya que $\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$, entonces

$$\boxed{SS_{\text{Res}} = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}}$$

Suma de cuadrados de los residuales con $\boldsymbol{n-p}$ grados de libertad.

Siendo $\boldsymbol{p = k + 1}$

Así, un estimador insesgado de σ^2 es

$$\hat{\sigma}^2 = MS_{\text{Res}} = \frac{SS_{\text{Res}}}{n-p}$$

Varianza de los errores

Ruido no explicado por el MRLM

Debe minimizarse

Regresión lineal múltiple
Ejemplo – RStudio

Se tiene la siguiente información de una localidad turística sobre los turistas extranjeros llegados de 5 países de procedencia:

Observación (i)	**Y** Número de turistas	X_1 Ingresos medios anuales (miles de euros)	X_2 Distancia (cientos de km)
1	18	5	17
2	25	10	15
3	7	2	32
4	12	4	25
5	19	6	20

Estime la varianza de los errores σ^2 (RStudio)

Regresión lineal múltiple

Ejemplo en RStudio

```
## Resumen del modelo ajustado
summary(mod)
```

```
Call:
lm(formula = turistas ~ ingreso + distancia)

Residuals:
      1       2       3       4       5 
-0.1927 -0.2867  0.1869 -0.8510  1.1435 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)  20.9060     4.9563   4.218   0.0519 .
ingreso       1.2123     0.3434   3.530   0.0717 .
distancia    -0.5162     0.1491  -3.462   0.0742 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.045 on 2 degrees of freedom
Multiple R-squared:  0.9885,	Adjusted R-squared:  0.9771 
F-statistic: 86.28 on 2 and 2 DF,  p-value: 0.01146
```

Residuales para cada valor de y

Parámetros estimados

Estimación error estándar de los β_j

Estimación del error estándar de los residuales

$\sqrt{\sigma^2}$ con **n – p** grados de libertad

Regresión lineal múltiple

Actividad 1 en RStudio

Volviendo al caso estudio de riesgo de infección en los hospitales…

A partir del muestreo de 113 hospitales en Estados Unidos, se evaluó la incidencia de diferentes factores en la probabilidad de que un paciente adquiera una infección mientras está hospitalizado. Las variables analizadas fueron:

Variable Respuesta (y): riesgo de infección en porcentaje **(riesgo)**

Potencial predictor (x_1): tiempo de hospitalización promedio de los pacientes (**tiempo**)

Potencial predictor (x_2): edad promedio de los pacientes (**edad**)

Potencial predictor (x_3): índice de rayos X realizados (**tasarayosx**)

Estime la varianza de los errores σ^2 (RStudio)

Regresión lineal múltiple

ANOVA

Recordemos la identidad fundamental del análisis de varianza ANOVA:

SST = SSR +SSE

SST	SSR	SSE
Suma de cuadrados totales	Suma de cuadrados debida a la regresión	Suma de cuadrados de residuales
Desviación total	**Desviación explicada por el modelo**	**Desviación debida al error del ajuste**

Siempre debo buscar que:

$$SSR >> SSE$$

En el MRLM las sumas de cuadrados están definidos de forma matricial así:

$$SS_{\mathrm{T}} = \mathbf{y}'\mathbf{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

$$SS_{\mathrm{R}} = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

$$SS_{\mathrm{Res}} = SS_{\mathrm{T}} - SS_{\mathrm{R}}$$

$$SS_{\mathrm{Res}} = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}$$

Regresión lineal múltiple

ANOVA

Así, la tabla ANOVA toma la siguiente forma:

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	Fo	Valor p
Regresión	SSR	k	$MSR = \frac{SSR}{k}$	$F_0 = \frac{MSR}{MSE}$	
Residuales	SSE	$n - k - 1$	$MSE = \frac{SSE}{n-k-1}$		
Total	SST	$n - 1$			

Regresión lineal múltiple
Utilidad del MRLM

Esta prueba se utiliza para determinar si hay una relación lineal entre la respuesta y cualquiera de las variables regresoras $x_1, x_2, \dots, x_k$.

Las hipótesis pertinentes son:

$H_0\colon \beta_0 = \beta_1 = \cdots = \beta_k = 0$ → Las x no explican a y

$H_1\colon \beta_j \neq 0$ para algún j → Al menos una x sí explica a y

Estadístico de prueba:

$$F_0 = \frac{SS_R/k}{SS_{Res}/(n-k-1)} = \frac{MS_R}{MS_{Res}} \overset{\text{bajo } H_0}{\sim} F_{k,n-k-1}$$

Criterio de rechazo: $F_0 > F_{\alpha,\, k,\, n-p}$ o $Valor\, p < \alpha$

NOTA: Rechazar H_0 significa que existe **al menos una** variable regresora que **sí** explica el comportamiento de los datos, NO TODAS!!!

Regresión lineal múltiple

¿Qué tan útil es el MRLM?

Para evaluar **qué tan útil** es el MRLM su utilizan los estadísticos R^2 y R^2_{Adj} (R^2 ajustado)

R^2 o coeficiente de determinación

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SS_{Res}}{SST}$$

- Mide el porcentaje de la varianza de y que es explicada por las variables regresoras
- Preferible que $\boldsymbol{R^2}$ sea alto
- **Problemita:** aumenta cuando se agregan variables regresoras al modelo, independientemente del valor de la contribución de la variable agregada. Por tanto, es difícil juzgar si un incremento en $\boldsymbol{R^2}$ dice algo importante.

Regresión lineal múltiple

Utilidad del MRLM - Medidas de bondad de ajuste

Otras dos maneras de evaluar la utilidad general del MRLM son los estadísticos R^2 y R^2_{Adj} (R^2 ajustado)

R^2 ajustado (R^2_{Adj})

$$R^2_{\text{Adj}} = 1 - \frac{SS_{\text{Res}}/(n-p)}{SS_{\text{T}}/(n-1)}$$

- Penaliza la adición de variables regresoras. Sólo aumentará al agregar una variable, si esa adición reduce el MSE.
- También es preferible que $\boldsymbol{R^2_{Adj}}$ sea alto.

Regresión lineal múltiple

Actividad 2 en RStudio

Volviendo al caso estudio de riesgo de infección en los hospitales…

Considerando todas las variables regresoras para construir el MRLM:

- Desarrolle completamente la prueba de utilidad del modelo. Para esto utilice un nivel de significancia de 0.05

- Determiné qué tan útil es el modelo para describir la variabilidad del riesgo de infección.

Regresión lineal múltiple

Ejemplo en Rstudio – ANOVA y prueba de utilidad

```
## Resumen del modelo ajustado
summary(mod)
```

```
Call:
lm(formula = turistas ~ ingreso + distancia)

Residuals:
      1       2       3       4       5 
-0.1927 -0.2867  0.1869 -0.8510  1.1435 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)  20.9060     4.9563   4.218   0.0519 .
ingreso       1.2123     0.3434   3.530   0.0717 .
distancia    -0.5162     0.1491  -3.462   0.0742 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.045 on 2 degrees of freedom
Multiple R-squared:  0.9885,	Adjusted R-squared:  0.9771 
F-statistic: 86.28 on 2 and 2 DF,  p-value: 0.01146
```

Resultados prueba de utilidad:

- Estadístico F
- Grados de libertad: k/(n-p)
- $Valor\ p$

En este caso:

Dado que $Valor\ p < 0.05$

Se rechaza H_0

Por tanto, existe **al menos una** variable regresora que **sí** explica el comportamiento de los datos.

Regresión lineal múltiple

Ejemplo en Rstudio - R^2 y R^2_{Adj}

```
## Resumen del modelo ajustado
summary(mod)
```

```
Call:
lm(formula = turistas ~ ingreso + distancia)

Residuals:
      1       2       3       4       5 
-0.1927 -0.2867  0.1869 -0.8510  1.1435 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  20.9060     4.9563   4.218   0.0519 .
ingreso       1.2123     0.3434   3.530   0.0717 .
distancia    -0.5162     0.1491  -3.462   0.0742 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.045 on 2 degrees of freedom
Multiple R-squared:  0.9885,	Adjusted R-squared:  0.9771 
F-statistic: 86.28 on 2 and 2 DF,  p-value: 0.01146
```

R^2_{Adj}

R^2

MUCHAS GRACIAS

Natalia Jaramillo Quiceno

e-mail: natalia.jaramilloq@upb.edu.co