UNIVERSIDAD SIMÓN BOLÍVAR
DEPARTAMENTO DE CÓMPUTO CIENTÍFICO Y ESTADÍSTICA
CÁTEDRA: ESTADISTICA PARA INGENIEROS (CO3321)

Laboratorio de Regresión Lineal Múltiple.

Las técnicas de regresión lineal múltiple busca establecer una relación entre una variable de respuesta o variable dependiente y, y las variables explicativas, predictorias o independientes x_1 , x_2 , ..., x_p

La ecuación de regresión lineal múltiple tiene la forma:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$$

Hipótesis del problema:

1. $\varepsilon_i \sim N(0, \sigma^2)\,;\ i = 1, 2, .., n$
2. El número de datos n es mayor $p + 1$, es decir, que se debe tener suficientes datos para estimar los $p + 1$ parámetros.
3. Los regresores son linealmente independientes, es decir, que ninguno de ellos está exactamente determinado por otros.
4. $E(y|x_1, x_2, ..., x_p) = \sigma^2$
5. $y \sim N(\mu, \sigma^2)$ y su componentes son independientes.
6. Las y_i no están correlacionadas entre si, $i = 1, 2, ..., n$.
7. Las variables $x_1, x_2, ..., x_n$ son determinsticas.

Los coeficientes de regresión se estiman por el método de mínimos cuadrados, donde el modelo se puede escribir matricialmente como:

$$Y = X\hat{\beta} + \varepsilon$$

de aquí:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

donde:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{pmatrix}$$

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & ... & x_{1p} \\ 1 & x_{21} & x_{22} & ... & x_{2p} \\ . & . & . & ... & . \\ . & . & . & ... & . \\ . & . & . & ... & . \\ 1 & x_{n1} & x_{n2} & ... & x_{np} \end{pmatrix}$$

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ . \\ . \\ . \\ \hat{\beta}_p \end{pmatrix}$$

$$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{pmatrix}$$

Coeficiente de determinación múltiple:

$$R^2 = 1 - \frac{SSE}{SS_{yy}}$$

donde $SSE = Y^TY - \hat{\beta}^T X^T Y$ y $SS_{yy} = \sum_{i=1}^{n}(y_i - \hat{y})^2$

Inferencia respecto a los parámetros:

1. **Intervalos de confianza para** $\beta_i\,;\ i = 0, 1, 2, ..., p$

   El estadístico de prueba es:

$$T = \frac{\hat{\beta}_i - \beta_i}{s\sqrt{c_{ii}}} \sim t_{n-1-p}$$

   donde los C_{ii} se obtienen a partir de:

$$(X^TX)^{-1} = \begin{pmatrix} c_{00} & c_{01} & ... & c_{0p} \\ c_{10} & c_{11} & ... & c_{1p} \\ . & . & ... & . \\ . & . & ... & . \\ . & . & ... & . \\ c_{p0} & c_{p1} & ... & c_{pp} \end{pmatrix}$$

   y $S^2 = \frac{SSE}{n-p-1}$

   El intervalo de confianza es:

$$I = (\hat{\beta}_i - t_{n-p-1;\frac{\alpha}{2}} S\sqrt{c_{ii}}\ ,\ \hat{\beta}_i + t_{n-p-1;\frac{\alpha}{2}} S\sqrt{c_{ii}})$$

2. **Pruebas de hipótesis para** $\beta_i\,;\ i = 0, 1, 2, ..., p$

   El estadístico de prueba bajo H_0 es:

$$T = \frac{\hat{\beta}_i - \beta_{i0}}{S\sqrt{c_{ii}}} \sim t_{n-p-1}$$

   - Si la prueba es bilateral:
     $H_0 : \beta_i = \beta_{i0}$ contra $H_1 : \beta_i \neq \beta_{i0}$
     La región de rechazo es

$$RR = (-\infty\,,\ -t_{n-p-1;\,\frac{\alpha}{2}}) \cup (t_{n-p-1;\,\frac{\alpha}{2}}, \infty)$$

   - Si la prueba es unilateral derecha:
     $H_0 : \beta_i \leq \beta_{i0}$ contra $H_1 : \beta_i > \beta_{i0}$
     La región de rechazo es

$$RR = (t_{n-p-1;\,\alpha}, \infty)$$

   - Si la prueba es unilateral izquierda:
     $H_0 : \beta_i \geq \beta_{i0}$ contra $H_1 : \beta_i < \beta_{i0}$
     La región de rechazo es

$$RR = (-\infty\,,\ -t_{n-p-1;\,\alpha})$$

3. **Análisis de varianza en la regresión lineal múltiple:**

   $H_0 : \beta_1 = \beta_2 = ... = \beta_p$, es decir, las variables independientes no están relacionadas linealmente con la variable dependiente, contra

   $H_1 : \beta_i \neq 0$ para algún $i = 1, 2, ..., p$, es decir, existe al menos una variable independientes que está linealmente relacionada con la variable dependiente.

   El estadístico de prueba bajo H_0 es:

$$F = \frac{R^2(n-p-1)}{p(1-R^2)} \sim F_{p\,,\,n-p-1}$$

La región de rechazo es

$$R = (f_{p,n-p-1\,;\,\alpha}\,,\,\infty)$$

y el p-valor es $1 - P(F \leq f_{obs})$

Ejemplo:

Dados los siguientes datos:

y	x_1	x_2	x_3
3.015	0.0	-1.417	-0.927
2.454	0.0	-1.417	-0.927
1.232	0.5	0.958	0.0386
1.328	0.5	0.958	0.0386
1.102	0.5	0.958	0.0386
0.217	-0.5	0.208	1.313
-0.635	0.5	-1.042	3.436
-0.093	-0.5	1.208	-0.386
3.473	0.0	-0.917	-1.776
-1.396	-0.5	-0.292	2.162
2.631	0.0	0.083	-3.475
-0.394	-0.5	0.708	0.463

$> y < -c(3.015, 2.454, 1.232, 1.328, 1.102, 0.217, -0.635, -0.093, 3.473, -1.396, 2.631, -0.394)$

$> x1 < -c(0.0, 0.0, 0.5, 0.5, 0.5, -0.5, 0.5, -0.5, 0.0, -0.5, 0.0, -0.5)$

$> x2 < -c(-1.417, -1.417, 0.958, 0.958, 0.958, 0.208, -1.042, 1.208, -0.917, -0.292, 0.083, 0.708)$

$> x3 < -c(-0.927, -0.927, 0.0386, 0.0386, 0.0386, 1.313, 3.436, -0.386, -1.776, 2.162, -3.475, 0.463)$

$> ajuste < -lm(y \sim x1 + x2 + x3)$

$> summary(ajuste)$

Esto nos da como resultado:

Los coeficientes de regresión son:
$\hat{\beta}_0 = 1.07756;\ \hat{\beta}_1 = 1.17321;\ \ \hat{\beta}_2 = -0.60041;\ \hat{\beta}_3 = -0.71222$

El coeficiente de determinación es $R^2 = 0.9062$

El error estndar, t_{obs} y el p-valor para las pruebas de hipótesis de los coeficientes de regresión son las siguientes:
Para $\hat{\beta}_0$: $e.e(\hat{\beta}_0) = 0.16303$, $t_{obs} = 6.609$ y $p - valor = 0.000168 * **$
Para $\hat{\beta}_1$: $e.e(\hat{\beta}_1) = 0.39935$, $t_{obs} = 2.938$ y $p - valor = 0.018772*$
Para $\hat{\beta}_2$: $e.e(\hat{\beta}_2) = 0.17192$, $t_{obs} = -3.492$ y $p - valor = 0.008169 * *$
Para $\hat{\beta}_3$: $e.e(\hat{\beta}_3) = 0.09476$, $t_{obs} = -7.516$ y $p - valor = 6.82e - 05 * **$

Para un nivel de significación del 5%, se rechaza la hipótesis nula para todos los β_i, pero la variable más significativa, por tener un p-valor más pequeño es x_3, seguida de x_2 y finalmente x_1.

Análisis de varianza:
$f_{obs} = 25.77$ y $p - valor = 0.0001829$

Por ejemplo, para un nivel de significación del 5%, se rechaza la hipótesis nula. Por lo tanto, existe al menos una variable que es significativa en el modelo de regresión lineal.
En consecuencia, el modelo de regresión lineal es:

$$y = 1.07756 + 1.17321x_1 - 0.60041x_2 - 0.71222x_3$$

Matriz de correlación:
$M = cbind(y, x1, x2, x3)$
$> cor(M)$
Obtenemos la matriz de correlación:

$$\begin{pmatrix} & y & x1 & x2 & x3 \\ y & 1 & 3.180581e-01 & -3.781388e-01 & -8.137158e-01 \\ x1 & 0.3180581 & 1 & -8.949863e-18 & -1.186437e-05 \\ x2 & -0.3781388 & -8.949863e-18 & 1 & 6.256791e-05 \\ x3 & -0.8137158 & -1.186437e-05 & 6.256791e-05 & 1 \end{pmatrix}$$

Coeficiente de determinación:

$> n < -length(y)$
$> SSE < -sum(ajuste\$residuals \wedge 2)$
$> ybarra < -mean(y)$
$> y.barra < -rep(ybarra, n)$

$> SSyy < -sum((y - y.barra) \wedge 2)$
$> R.cuadrado < -1 - SSE/SSyy$
$> R.cuadrado$
[1]0.9062387

Intervalo de confianza para β_1:

$> p < -3$
$> ones < -rep(1, n)$
$> x < -cbind(ones, x1, x2, x3)$
$> cii < -solve(t(x)\% * \%x)$
$> c11 < -cii[2, 2]$
$> S < -sqrt(SSE/(n - p - 1))$
$> alpha < -0.05$
$> t.alphamedio < -qt(1 - alpha/2, n - p - 1)$
$> coef < -ajuste\$coefficients$
$> B1 < -coef[2]$
$> lim.inf < -B1 - t.alphamedio * S * sqrt(c11)$
$> lim.sup < -B1 + t.alphamedio * S * sqrt(c11)$
$> Intervalo < -c(lim.inf, lim.sup)$
$> Intervalo$

El intervalo de confianza para β_1 es (0.2523178 , 2.0941110)

Intervalo de confianza para β_2:

$> p < -3$
$> ones < -rep(1, n)$
$> x < -cbind(ones, x1, x2, x3)$
$> cii < -solve(t(x)\% * \%x)$
$> c22 < -cii[3, 3]$
$> S < -sqrt(SSE/(n - p - 1))$
$> alpha < -0.05$
$> t.alphamedio < -qt(1 - alpha/2, n - p - 1)$
$> coeficientes < -ajuste\$coefficients$
$> B2 < -coef[3]$

$> lim.inf < -B2 - t.alphamedio * S * sqrt(c22)$
$> lim.sup < -B2 + t.alphamedio * S * sqrt(c22)$
$> Intervalo < -c(lim.inf, lim.sup)$
$> Intervalo$

El intervalo de confianza para β_2 es $(-0.9968533\,,\ -0.2039657)$

Predicción:

Predecir el valor de y para $x_1 = 0.25$, $x_2 = -1.71$ y $x_3 = 3.11$

Solución:

$> B0 < -coef[1]$
$> B1 < -coef[2]$
$> B2 < -coef[3]$
$> B3 < -coef[4]$
$> B4 < -coef[5]$
$> B5 < -coef[6]$
$> x1 < -0.25$
$> x2 < -1.71$
$> x3 < 3.11$
$> y < -B0 + B1 * x1 + B2 * x2 + B3 * x3$
$> y$
$[1] - 1.870854$

Diagrama de dispersión

$> x11()$
$> pairs(M)$

Histograma de residuales:

$> x11()$
$> hist(resid(ajuste), main =$ 'Histograma de residuales')

El modelo es correcto si la distribución de los residuales es normal con media 0.

Gráfica de normalidad de los residuos:

$> x11()$
$> qqnorm(resid(ajuste), main =$ 'Gráfica de normalidad de los residuos')
$> qqline(resid(ajuste))$

Para que el modelo sea correcto, los puntos del diagrama de dispersión deben estar muy cercanos a la recta de regresión.

Gráfica de independencia de los residuos:

$> x11()$
$> plot(ajuste\$fitted.values, resid(ajuste), main =$ 'Gráfica de independencia')
$> abline(h = 0)$

Los puntos del diagrama de dispersión tienen que estar distribuidos por encima y por debajo de la recta $h = 0$. De ser así, se cumple la hipótesis de independencia de los residuos.

Laboratorio de Análisis de Varianza de un factor.

El Análisis de varianza (ANOVA) se basa en la comparación de medias de poblaciones normales.

Nuestro estudio consiste en la comparación de medias de k poblaciones, a partir de muestras independientes de tamaño $n_1, n_2, ..., n_k$.

Es decir, $H_0 : \mu_1 = \mu_2 = ... = \mu_k$ contra $H_1 : \mu_i \neq \mu_j$ para algún $i \neq j$
donde $Y_{i,j} \sim N(\mu_i, \sigma^2)$; $\quad i = 1, 2, ..., k; \quad j = 1, 2, ..., n_i$

El estadístico de prueba bajo H_0 y suponiendo igualdad de varianzas es:

$$F\frac{\tilde{S}^2}{\tilde{S}_k^2} \sim F_{k-1,n-k}$$

donde:
$\tilde{S}_k^2 = \frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$ representa la variabilidad interna de los grupos.

$\tilde{S}^2 = \frac{1}{k-1}\sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2$ representa la variabilidad entre grupos.
k es el número de poblaciones o grupos.
$\bar{y}_i = \frac{1}{n_i}\sum_{i=1}^{n_i} y_{ij}$ es la media de cada grupo.
$\bar{y} = \frac{1}{n}\sum_{i=1}^{k} n_i\bar{y}_i$ es la media global.
n_i es el número de individuos del i-ésimo grupo; $i = 1, 2, ..., k$
$n = \sum_{i=1}^{k} n_i$ es el número de individuos de la población.
y_{ij} es la observación para el j-ésimo individuo del i-ésimo grupo;
$i = 1, 2, ..., k; \quad j = 1, 2, ..., n_i$

Ejemplo:

En la siguiente tabla se presenta la altura de los árboles para 3 bosques distintos. haga un análisis de varianza al nivel $\alpha = 0.01$ para ver si existe una diferencia significativa en la altura media de los árboles de cada bosque. Suponga poblaciones normales con varianzas iguales.

Bosque 1	**Bosque 2**	**Bosque 3**
23.4	22.5	18.9
22.4	22.9	21.1
24.6	24.6	21.2
24.9	23.7	22.1
25.0	24.0	22.5
26.2	24.5	23.5
26.3	25.3	24.5
26.8	26.0	24.6
26.8	26.2	26.2
26.9	26.4	26.7
27.0	26.7	$---$
27.6	26.9	$---$
27.7	27.4	$---$

$> altura < -c(23.4, 22.4, 24.6, 24.9, 25.0, 26.2, 26.3, 26.8, 26.8, 26.9, 27.0, 27.6, 27.7, 22.5, 22.9, 24.6, 23.7, 24.0, 24.5, 25.3, 26.0, 26.2, 26.4, 26.7, 26.9, 27.4, 18.9, 21.1, 21.2, 22.1, 22.5, 23.5, 24.5, 24.6, 26.2, 26.7)$
$> bosque < -factor(rep(LETTERS[1:3], c(13, 13, 10)))$
$> altura.df < -data.frame(bosque, altura)$

$> altura.df$
$> aov.altura < -aov(altura \sim bosque, altura.df)$
$> summary(aov.altura)$

Esto nos da como resultado:

$k - 1 = 2$
$n - k = 33$
$\sum_{i=1}^{k} n_i(y_i - \bar{y})^2 = 42.95$
$\sum_{i=1}^{k} \sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 = 115.83$
$\tilde{S}^2 = 21.48$
$\tilde{S}_k^2 = 3.51$
$fobs = 6.118$
$p - valor = 0.00549$

Como $p - valor \leq \alpha$ los datos presentan suficientes evidencias para rechazar H_0 al nivel $\alpha = 0.01$. Por lo tanto, existe una diferencia significativa en la altura media de los árboles de cada bosque.

Tablas:

1. **Tabla ANOVA de Regresión Múltiple:**

	Sum of Squares	**df**	**Mean Square**	**F**	$Pr(>F)$
Regression	SSR	p	$MSR = \frac{SSR}{p}$	$\frac{MSR}{MSE}$	p-valor
Residual	SSE	$n-p-1$	$MSE = \frac{SSE}{n-p-1}$		
Total	$SSR + SSE$	$n-1$	$MSR + MSE$		

2. **Tabla de Coeficientes de Regresión Múltiple:**

	Beta estimate	**Std. error**	**t**	$Pr(>T)$
Intercept	$\hat{\beta}_0$	$s\sqrt{c_{00}}$	$\frac{\hat{\beta}_0}{S\sqrt{c_{00}}}$	p-valor
x_1	$\hat{\beta}_1$	$s\sqrt{c_{11}}$	$\frac{\hat{\beta}_1}{S\sqrt{c_{11}}}$	p-valor
x_2	$\hat{\beta}_2$	$s\sqrt{c_{22}}$	$\frac{\hat{\beta}_2}{S\sqrt{c_{22}}}$	p-valor
...	...	...	...	...
x_p	$\hat{\beta}_p$	$s\sqrt{c_{pp}}$	$\frac{\hat{\beta}_p}{S\sqrt{c_{pp}}}$	p-valor

3. **Tabla ANOVA de un factor:**

	Sum of Squares	**df**	**Mean Square**	**F**	$Pr(>F)$
Method	$(k-1)\tilde{S}^2$	$k-1$	$\tilde{S}^2$	$\frac{\tilde{S}^2}{\tilde{S}_k^2}$	p-valor
Residuals	$(n-k)\tilde{S}_k^2$	$n-k$	$\tilde{S}_k^2$		
Total	$(k-1)\tilde{S}^2 + (n-k)\tilde{S}_k^2$	$n-1$	$\tilde{S}^2 + \tilde{S}_k^2$		