Regresión Avanzada

Regresión Lineal Múltiple

Regresión lineal múltiple

Buscamos extender las ideas expresadas anteriormente al caso de más de un regresor.
En nuestro ejemplo teníamos información sobre la inversión publicitaria en Radio y Diario.
Si corriéramos regresiones simples en este caso, tendríamos que ...

Regresión lineal múltiple

	Coeficientes	Error Estándar	Estadístico	p-valor
Intercept	9.312	0.563	16.54	< 0.0001
Radio	0.203	0.020	9.92	< 0.0001

Por cada $1000 invertidos en publicidad en radio las ventas se incrementan en promedio en 203 unidades.

	Coeficientes	Error Estándar	Estadístico	p-valor
Intercept	12.351	0.621	19.88	< 0.0001
Diario	0.055	0.017	3.30	< 0.0001

Por cada $1000 invertidos en publicidad en diario las ventas se incrementan en promedio en 55 unidades

Regresión lineal múltiple

Cómo hacer una sola predicción dados los datos de las tres inversiones?
Cada regresión ignora lo que pasa en los otros medios, si estuvieran vinculados esta información se perdería, se podrían llegar a conclusiones erróneas haciendo análisis unidimensionales.

Regresión lineal múltiple

Sean $X_1, \ldots, X_{p-1}$ variables regresoras. El modelo de regresión lineal múltiple está dado por

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1} + \epsilon.$$

- X_j es la j-ésima variable regresora.
- β_j es el parámetro que cuantifica la asociación entre la j-ésima variable regresora y la variable de respuesta. Es el efecto promedio de Y cuando X_j aumenta en una unidad.

Regresión lineal múltiple

En nuestro caso el modelo de regresión queda

$$ventas = \beta_0 + \beta_1 TV + \beta_2 Radio + \beta_3 Diario + \epsilon$$

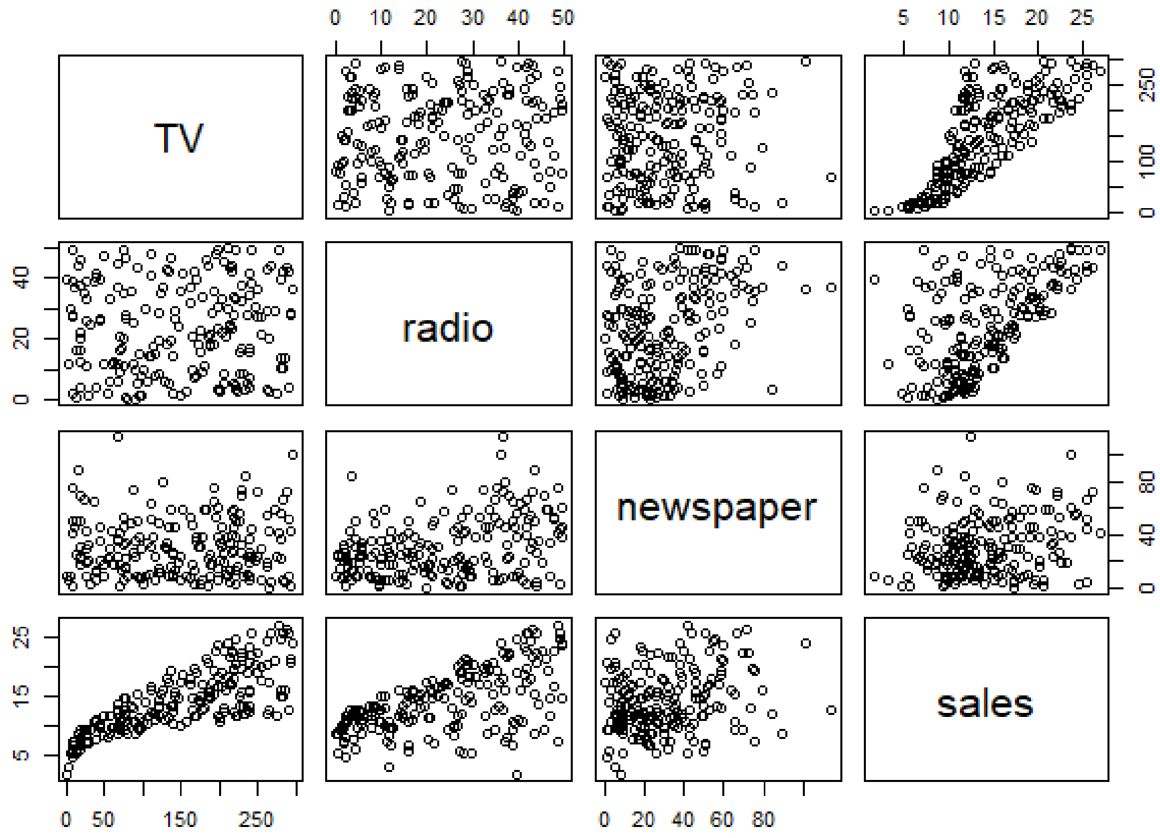
Sobre los errores hacemos los mismos supuestos que en el caso de regresión lineal

- Son independientes entre si e independientes de las variables regresoras.
- Tienen media cero, $E(\epsilon) = 0$, son homocedásticos, es decir que tienen todos igual varianza, $var(\epsilon) = \sigma^2$.
- Están normalmente distribuidos, $\epsilon \sim N(0, \sigma^2)$.

Regresión lineal múltiple

Scatter plots de todas las variables

pairs(advertising[,-1])



Regresión lineal múltiple

Realizamos un experimento n veces y tenemos los siguientes datos

Respuesta y	Variables Regresoras $(X_1, \ldots, X_{(p-1)})$
y_1	$x_{11}, x_{12}, \ldots, x_{1(p-1)}$
y_2	$x_{21}, x_{22}, \ldots, x_{2(p-1)}$
$\vdots$	$\vdots$
y_n	$x_{n1}, x_{n2}, \ldots, x_{n(p-1)}$

Regresión lineal múltiple

Asumimos que el modelo es

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{(p-1)} X_{(p-1)} + \epsilon$$

Luego las n-uplas siguen el siguiente modelo

$$\begin{gathered}
y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_{(p-1)} x_{1(p-1)} + \epsilon_1 \\
y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_{(p-1)} x_{2(p-1)} + \epsilon_2 \\
\vdots \\
y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_{(p-1)} x_{n(p-1)} + \epsilon_n
\end{gathered}$$

Regresión lineal múltiple

En forma matricial

$$\underbrace{\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}}_{Y} = \underbrace{\begin{pmatrix} 1 & x_{11} & \dots & x_{1(p-1)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{n(p-1)} \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix}}_{\beta} + \underbrace{\begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}}_{\epsilon}$$

$$n \times 1 \qquad n \times p \qquad p \times 1 \quad + \quad n \times 1$$

Regresión lineal múltiple

Supuestos del modelo lineal

- $E(\epsilon) = 0$.
- $E(\epsilon'\epsilon) = \sigma^2 I_n$. Es decir que son independientes entre si y tienen igual varianza.
- $rg(X) = p$.
- X no es una matriz estocástica. Asumimos esto durante el curso salvo que se indique lo contrario.
- $\epsilon \sim N(\mathbf{0}, \sigma^2 I_n)$.

Estos supuestos son necesarios para hacer inferencia. Para probar la consistencia tenemos que hacer un supuesto adicional.

- $\lim_{n\to\infty} \frac{X'X}{n} = \Delta$, donde Δ es una matriz no singular.

Observación: A' es la matriz traspuesta de A.

Estimación de los parámetros $(\beta_0, \beta_1, \ldots, \beta_{(p-1)})$. Consideramos todos los posibles valores de $\beta \in \mathbb{R}^p$.
Objetivo: es hallar el valor $(\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_{(p-1)})$ que minimice la suma de los cuadrados de los residuos.

$$
\begin{aligned}
RSS &= \sum_{i=1}^{n} \epsilon_i^2 = \epsilon' \epsilon = (y - X\beta)'(y - X\beta) \\
&= y'y - y'X\beta - \beta' X' y + \beta' X' X \beta = \\
&= y'y - 2\beta' X' y + \beta' X' X \beta.
\end{aligned}
$$

Diferenciando respecto de β

$$
\frac{\partial RSS}{\partial \beta} = -2X'y + 2X'X\beta.
$$

Regresión lineal múltiple

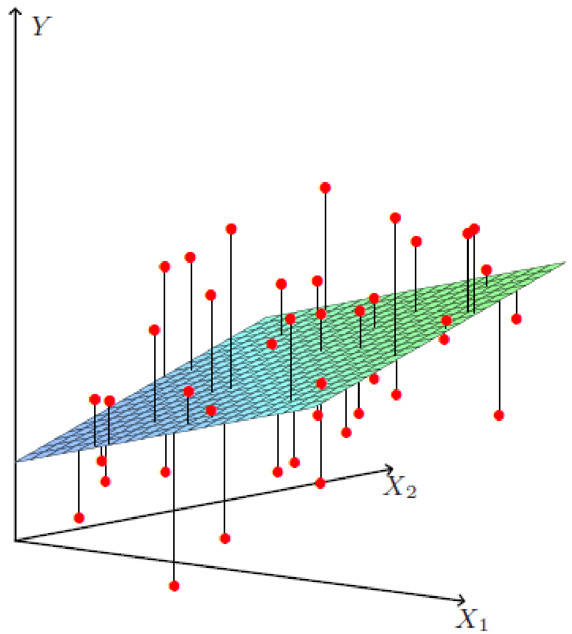
Las ecuaciones normales quedan

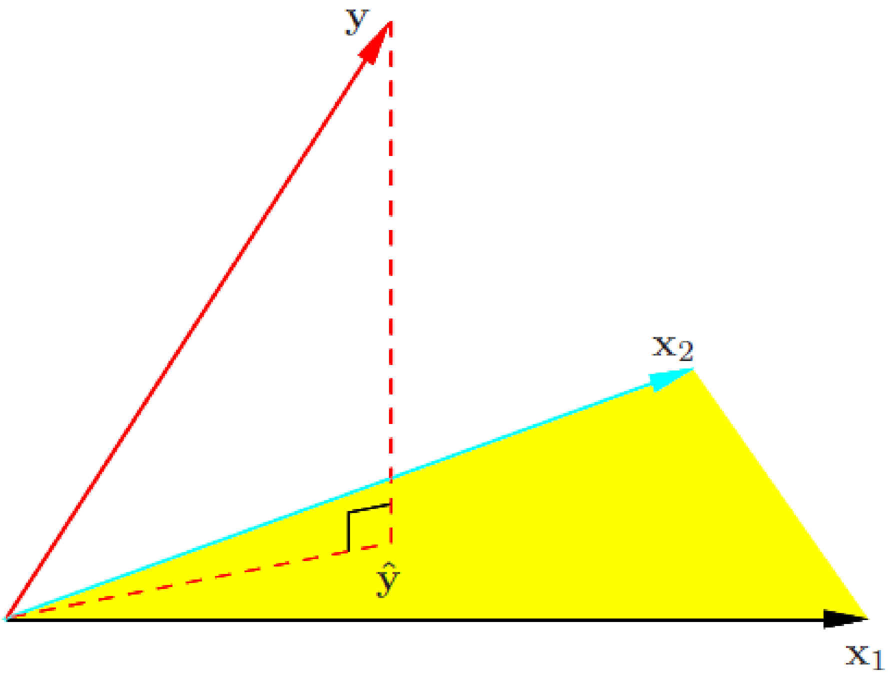
$$\begin{aligned} \left.\frac{\partial RSS}{\partial \beta}\right|_{\widehat{\beta}} &= -2X'y + 2X'X\widehat{\beta} = 0 \\ 2X'y &= 2X'X\widehat{\beta}. \end{aligned}$$

$$\boxed{\widehat{\beta} = (X'X)^{-1}X'y.}$$

Estamos usando que X es una matriz de rango completo. Cómo RSS es al menos definida no negativa entonces es un mínimo.

Regresión lineal múltiple





Representación de ajuste por mínimos cuadrados en con dos variables regresoras .

Representación geométrica del estimador de mínimos cuadrados. El valor predicho es la proyección ortogonal de *y* en hiperplano generado por X_1 y X_2.

Regresión lineal múltiple

```
adv.lm=lm(sales~TV+radio+newspaper)
summary(adv.lm)

Call:
lm(formula = sales ~ TV + radio + newspaper)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8277 -0.8908  0.2418  1.1893  2.8292

Coefficients:
              Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)   2.938889    0.311908     9.422    <2e-16 ***
TV            0.045765    0.001395    32.809    <2e-16 ***
radio         0.188530    0.008611    21.893    <2e-16 ***
newspaper    -0.001037    0.005871    -0.177      0.86
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared:  0.8972,        Adjusted R-squared:  0.8956
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

B_0 estimado

B_{TV} estimado

B_{Radio} estimado

B_{Diario} estimado

El modelo queda

$$ventas \approx 2.94 + 0.046\,TV + 0.189\,Radio - 0.001\,Diario.$$

Interpretación:

- Fijando la inversión en TV y diario, por cada \$ invertido en radio se venderán 189 unidades más.
- Fijando la inversión en radio y diario, por cada \$ invertido en TV se venderán 46 unidades más.
- Fijando la inversión en TV y diario, por cada \$ invertido en TV se venderá 1 unidades menos. Ojo: el p-valor es muy alto. Esto nos está diciendo que no es significativamente distinto de cero.

Acá se ve claro que la regresión simple y múltiple pueden ser muy distintas.

Regresión lineal múltiple

El predictor de y, $\widehat{y}$, está dado por:

$$\widehat{y} = X\widehat{\beta} = X(X'X)^{-1}X'y.$$

Llamamos $H = X(X'X)^{-1}X'$ a la matriz sombrero. Propiedades:

1. H es simétrica.
2. H es idempotente, $HH = H$.
3. $traza(H) = traza(X(X'X)^{-1}X') = traza(X'X(X'X)^{-1}) = traza(I_p) = p$.

Regresión lineal múltiple

Los residuos son la diferencia entre el valor observado, y_i y el valor predicho, $\widehat{y}_i$.

$$r = y - \widehat{y} = y - X\widehat{\beta} = y - Hy = (I - H)y.$$

La matriz $(I - H)$ también tiene propiedades interesantes.
Propiedades:

1. $I - H$ es simétrica.
2. $I - H$ es idempotente, $(I - H)(I - H) = (I - H)$.
3. $traza(I - H) = traza(I_n) - traza(H) = n - p$.

Regresión lineal múltiple

Propiedades de los estimadores de mínimos cuadrados.

(i) Error de estimación.

$$\begin{aligned}
\widehat{\beta}-\beta &= (X'X)^{-1}X'y-\beta \\
&= (X'X)^{-1}X'(X\beta+\epsilon)-\beta \\
&= (X'X)^{-1}X'X\beta+(X'X)^{-1}X'\epsilon-\beta \\
&= \beta+(X'X)^{-1}X'\epsilon-\beta=(X'X)^{-1}X'\epsilon.
\end{aligned}$$

(ii) Sesgo

$$E(\widehat{\beta}-\beta)=(X'X)^{-1}X'\underbrace{E(\epsilon)}_{0}=0,$$

el estimador es insesgado.

Regresión lineal múltiple

(iii) Matriz de covarianza

$$
\begin{aligned}
Var(\widehat{\beta}) &= E((\widehat{\beta}-\beta)(\widehat{\beta}-\beta)') \\
&= E(((X'X)^{-1}X'\epsilon)((X'X)^{-1}X'\epsilon)') \\
&= E((X'X)^{-1}X'\epsilon\epsilon'X(X'X)^{-1}) \\
&= (X'X)^{-1}X'E(\epsilon\epsilon')X(X'X)^{-1} \\
&= (X'X)^{-1}X'\sigma^2 I_n X(X'X)^{-1} \\
&= \sigma(X'X)^{-1}X'X(X'X)^{-1} = \sigma^2(X'X)^{-1}
\end{aligned}
$$

Regresión lineal múltiple

Estimación de σ^2

El criterio de mínimos cuadrados no brinda información para estimar σ^2, ya que no aparecen en la RSS.

Para estimarlo pensemos en los residuos $E(r_i^2) = \sigma^2$.

$$r = y - \widehat{y} = y - \underbrace{X(X'X)^{-1}X'}_{H} y = (I - H)y$$

La suma de los cuadrados de los residuos está dada por

$$\begin{aligned} RSS &= \sum_{i=1}^{n} r_i^2 = r'r \\ &= ((I-H)y)'((I-H)y) = y'(I-H)'(I-H)y = y'(I-H)y \end{aligned}$$

Reemplazando $y = X\beta + \epsilon$, queda

$$RSS = y'(I - H)y = \epsilon'(I - H)\epsilon.$$

Regresión lineal múltiple

Como $\epsilon \sim N(0, \sigma^2 I)$ entonces $y \sim N(X\beta, \sigma^2 I)$, por lo tanto

$$\frac{y'(I-H)y}{\sigma^2} = \frac{\epsilon'(I-H)\epsilon}{\sigma^2} \sim \chi^2_{n-p},$$

Luego,

$$E\left(\frac{y'(I-H)y}{\sigma^2}\right) = \left(\frac{\epsilon'(I-H)\epsilon}{\sigma^2}\right) = n-p$$

En forma equivalente,

$$E\left(\frac{y'(I-H)y}{n-p}\right) = \sigma^2.$$

De aquí que la *media* de la suma de los cuadrados de los residuos es un estimador de σ^2,

p = cant de parametros beta a estimar

$$\widehat{\sigma}^2 = \frac{RSS}{n-p}.$$

Regresión lineal múltiple

Theorem

Bajo las condiciones sobre los errores y las variables predictoras establecidas anteriormente. El estimador de mínimos cuadrados de β *es el estimador lineal insesgado de mínima varianza uniforme de* β.

Proof.

Observemos que el estimador de mínimos cuadrados es lineal en y.
$\widehat{\beta} = (X'X)^{-1}X'y$.
Consideramos una función lineal en y arbitraria como posible candidato a ser el estimador lineal insesgado de mínima varianza de $v'\beta$. Sea $\beta^* = a'y$. Calculamos el sesgo,
$E(\beta^*) = E(a'y) = a'X\beta$. El estimador es insesgado sii $v' = a'X$.
Calculamos la varianza de β^*.

$$var(a'y) = a'var(y)a = \sigma^2 a'a.$$

Por otro lado

$$
\begin{aligned}
var(v'\widehat{\beta}) &= v'var(\widehat{\beta})v = v'\sigma^2(X'X)^{-1}v \underbrace{=}_{v'=a'X} \\
&= \sigma^2 a'X(X'X)^{-1}X'a = \sigma^2 a'Ha.
\end{aligned}
$$

a' y $\qquad\qquad$ $\widehat{v`B}$

Luego, $var(\cancel{v'\widehat{\beta}}) - var(\cancel{a'y}) = \sigma a'(I-H)a \geq 0$, porque $I - H$ es semidefinida positiva.

Bajo los supuestos que venimos asumiendo en los errores $\epsilon \sim N(0, \sigma^2 I)$ el estimador de máxima verosimilitud de β coincide con el estimador de mínimos cuadrados de β.
Para ver esto alcanza con plantear la función de verosimilitud de los errores.

$$\begin{aligned}
\mathcal{L}ik(\beta, \sigma^2) &= \prod_{i=1}^{n} f_{\epsilon_i}(\epsilon_i, \beta, \sigma^2) \\
&= \frac{1}{(2\pi\sigma^2)^{n/2}} exp\left(-\frac{\sum_{i=1}^{n} \epsilon_i^2}{2\sigma^2}\right) \\
&= \frac{1}{(2\pi\sigma^2)^{n/2}} exp\left(-\frac{(y - X\beta)'(y - X\beta)}{2\sigma^2}\right).
\end{aligned}$$

En el caso del estimador de β, es claro que maximizar la $\mathcal{L}ik(\beta, \sigma^2)$ es equivalente a minimizar $(y - X\beta)'(y - X\beta)$ por lo tanto ambos estimadores coinciden.
Por otra parte, al buscar el estimador de σ^2 obtenemos $\tilde{\sigma}^2 = \frac{RSS}{n}$.

Regresión lineal múltiple

Consistencia del estimador de mínimos cuadrados de β.
Para mostrar la consistencia, alcanza con mostrar que la varianza tiende a cero, ya que sabemos que es insesgado. Recordemos que uno de los supuestos que teníamos era que $\lim_{n\to\infty} \frac{X'X}{n} = \Delta$

$$\begin{aligned}
\lim_{n\to\infty} var(\widehat{\beta}) &= \lim_{n\to\infty} \sigma^2 (X'X)^{-1} \\
&= \sigma^2 \lim_{n\to\infty} \frac{1}{n} \left(\frac{(X'X)}{n} \right)^{-1} = 0.
\end{aligned}$$

Regresión lineal múltiple

Estandarización de los coeficientes de regresión. En ocaciones es dificil comparar lo estimadores de los coeficientes de regresión ya que los mismos reflejan en las unidades en que fueron medidas las variables X_j. Ej: $\widehat{y} = 5 + X_1 + 1000X_2$ donde,

- y y X_2 están medidas en litros.
- X_1 está medida en mililitros.

Si bien $\widehat{\beta}_1 = 1$ y $\widehat{\beta}_2 = 1000$, es decir $\widehat{\beta}_2 \gg \widehat{\beta}_1$, el efecto de ambas variables explicativas es el mismo. Cada litro que se adiciona en cada una de las variables provoca el mismo efecto en la variable de respuesta.

Regresión lineal múltiple

Supongamos que tenemos el modelo

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

Si reescalamos una variable regresora tenemos que $(c \times X_1)$.

$$y = \beta_0 + \underbrace{\frac{\beta_1}{c}}_{\widetilde{\beta}_1} (c \times X_1) + \beta_2 X_2 + \epsilon.$$

Si reescalamos la variable de respuesta tenemos que $(c \times y)$.

$$c \times y = \underbrace{c \times \beta_0}_{\widetilde{\beta}_0} + \underbrace{\beta_1 \times c}_{\widetilde{\beta}_1} X_1 + \underbrace{\beta_2 \times c}_{\widetilde{\beta}_2} X_2 + \underbrace{\epsilon \times c}_{\widetilde{\epsilon}}$$

donde $\widetilde{\epsilon} \sim N(0, \sigma^2 \times c^2)$.

Regresión lineal múltiple

Test de hipótesis relacionados con los coeficientes del modelo.

- Cuál es la validez general del modelo?
- Cuáles son las variables más importantes?

Planteo general

$$H_0 : R\beta = r \text{ vs } H_A : R\beta \neq r.$$

donde $R \in \mathbb{R}^{Jxp}$ donde $rg(R) = J$ y $r \in \mathbb{R}^{Jx1}$.

Regresión lineal múltiple

Algunos casos importantes

- $H_0 : \beta_j = 0$ en este caso
  $R = (0, \ldots, 0, \underbrace{1}_{j-\text{ ésima coordenada}}, 0, \ldots, 0)$, $r = 0$ y $J = 1$.
- $H_0 : \beta_3 = \beta_4$ en forma equivalente $H_0 : \beta_3 - \beta_4 = 0$. Aquí tenemos $R = (0, 0, 1, -1, 0, \ldots, 0)$, $r = 0$ y $J = 1$.
- $H_0 : \beta_3 = \beta_4 = \beta_5$ en forma equivalente

$$\begin{aligned} H_0 \quad : \quad & \beta_3 - \beta_4 = 0 \\ & \beta_4 - \beta_5 = 0. \end{aligned}$$

En este caso $J = 2$, $r = (0, 0)'$ y

$$R = \begin{bmatrix} 0 & 0 & 1 & -1 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & \ldots & 0 \end{bmatrix}$$

Regresión lineal múltiple

Algunos casos importantes

- $H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$ en forma equivalente $H_0 : \beta_j = 0$ para $j = 1, \ldots, p-1$. En este caso $J = p - 1$, $r = (0, \cdots, 0)'$ y

$$R = \begin{bmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \end{bmatrix}$$

Este último caso es el *test de bondad de ajuste de una regresión*. Busca determinar si algún β_j es significativo en el modelo. La hipótesis nula señala que ninguna variable es significativa para este modelo. La hipótesis alternativa es $H_A : \beta_j \neq 0$ para algún $j = 1, \ldots, p-1$.

Regresión lineal múltiple

Para ver la región de rechazo pensemos nuevamanente en la descomposición de la varianza.

$$\begin{aligned} y_i &= \widehat{y}_i + r_i \\ y_i - \overline{y} &= \widehat{y}_i - \overline{y} + r_i \\ \sum_{i=1}^{n} (y_i - \overline{y})^2 &= \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2 + \sum_{i=1}^{n} r_i^2. \\ TSS &= ESS + RSS, \end{aligned}$$

TSS mide la variabilidad total (tiene $n-1$ grados de libertad).
ESS mide la variabilidad explicada por el modelo (tiene $p-1$ grados de libertad).
RSS mide la variabilidad no explicada o residual (tiene $n-p$ grados de libertad).

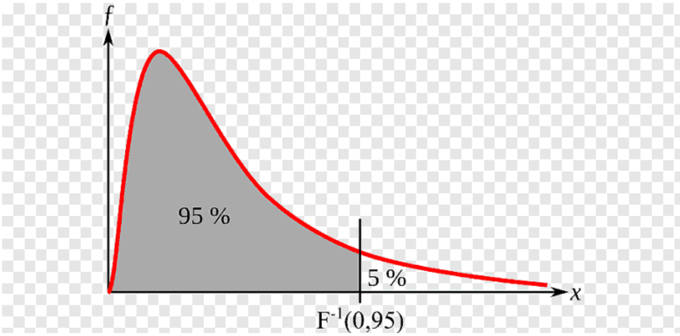
Regresión lineal múltiple

Comparamos la variabilidad explicada con la no explicada mediante el estadístico F:

$$F = \frac{ESS/(p-1)}{RSS/n-p}$$

Bajo H_0 el estadístico F sigue una distribución $F_{p-1,n-p}$. La región de rechazo de H_0 a nivel de significación α esta dada por

$$F > f_{p-1,n-p;1-\alpha}$$



Regresión lineal múltiple

Tabla de anova

Fuente de variación	SS	gl	Cuad. medios	Estad. *I*
Explicada (ESS)	$\sum_{i=n}^{n}(\widehat{y}_i - \overline{y})^2$	$p-1$	$\frac{\sum_{i=n}^{n}(\widehat{y}_i - \overline{y})^2}{p-1}$	F
Residual (RSS)	$\sum_{i=n}^{n} r_i^2$	$n-p$	$S_R^2 = \frac{\sum_{i=n}^{n} r_i^2}{n-p}$	
Total (TSS)	$\sum_{i=n}^{n}(y_i - \overline{y})^2$	$n-1$		

Regresión lineal múltiple

```
adv.lm=lm(sales~TV+radio+newspaper)
summary(adv.lm)

Call:
lm(formula = sales ~ TV + radio + newspaper)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8277 -0.8908  0.2418  1.1893  2.8292

Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  2.938889  0.311908     9.422   <2e-16 ***
TV           0.045765  0.001395    32.809   <2e-16 ***
radio        0.188530  0.008611    21.893   <2e-16 ***
newspaper   -0.001037  0.005871    -0.177     0.86
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared:  0.8972,     Adjusted R-squared:  0.8956
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

Estadístico F
Grados de libertad
p=4 entonces p-1=3
n=200 entonces n-p=196.

p-valor=P(F>570.3) donde
$F \sim f_{3,196}$

Regresión lineal múltiple

Si queremos testear

$$H_0 : \beta_j = 0 \text{ vs } H_A : \beta_j \neq 0.$$

Calculamos la suma de los residuos parciales bajo H_0

$$\begin{aligned} RSS_0 &= \sum_{i=1}^{n} (y_i - \widehat{\beta_0} - \widehat{\beta_1} x_{i1} - \cdots - \widehat{\beta_{j-1}} x_{i(j-1)} - \\ &- \widehat{\beta_{j+1}} x_{i(j+1)} - \cdots - \widehat{\beta_{p-1}} x_{i(p-1)})^2. \end{aligned}$$

Y se calcula el estadístico F

$$F = \frac{(RSS - RSS_0)/1}{RSS/n-p} \text{bajo } H_0 \text{ tiene distribución } f_{1,n-p}.$$

Se puede ver que $f_{1,n-p} = (t_{n-p})^2$.

Regresión lineal múltiple

Distribución de los estimadores de los coeficientes: Todos los estimadores $\widehat{\beta}_j$ verifican:

$$\frac{\widehat{\beta}_j - \beta_j}{\widehat{se}(\widehat{\beta}_j)} \sim t_{n-p}$$

donde $\widehat{se}(\widehat{\beta}_j)$ es el error estándar de $\widehat{\beta}_j$.
Luego se rechaza $H_0 : \beta_j = 0$ cuando

$$\left| \frac{\widehat{\beta}_j}{\widehat{se}(\widehat{\beta}_j)} \right| > t_{n-p;1-\alpha/2}.$$

Regresión lineal múltiple

Para cualquiera de los coeficientes del modelo podemos construir intervalos de nivel $1-\alpha$ de confianza:

$$IC_{1-\alpha}(\beta_j) = \left(\widehat{\beta}_j \pm t_{n-p;1-\alpha/2}\widehat{se}(\widehat{\beta}_j)\right).$$

Regresión lineal múltiple

```
adv.lm=lm(sales~TV+radio+newspaper)
summary(adv.lm)
```

```
Call:
lm(formula = sales ~ TV + radio + newspaper)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8277 -0.8908  0.2418  1.1893  2.8292

Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  2.938889  0.311908     9.422   <2e-16 ***
TV           0.045765  0.001395    32.809   <2e-16 ***
radio        0.188530  0.008611    21.893   <2e-16 ***
newspaper   -0.001037  0.005871    -0.177     0.86
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared:  0.8972,     Adjusted R-squared:  0.8956
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

Errores estándar estimados
Para los parámetros β.

Estadísticos correspondientes a los test individuales para cada B. Estos tests tienen distribución t_{196}

P-valores correspondientes a los test indivuales.

Regresión lineal múltiple

En nuestro caso

$$H_0 : \beta_0 = 0 \text{ vs } H_A : \beta_0 \neq 0.$$

Tenemos que $t_{obs} = 2.9388/0.3119 = 9.422$ luego p-valor$= 2 \times 10^{-16}$. El intercept es significativo.

$$H_0 : \beta_{TV} = 0 \text{ vs } H_A : \beta_{TV} \neq 0.$$

Tenemos que $t_{obs} = 0.045765/0.001395 = 32.809$ luego p-valor$= 2 \times 10^{-16}$. La inversión publicitaria en TV es significativa.

Regresión lineal múltiple

$$H_0 : \beta_{Radio} = 0 \text{ vs } H_A : \beta_{Radio} \neq 0.$$

Tenemos que $t_{obs} = 0.188530/0.008611 = 21.893$ luego p-valor$= 2 \times 10^{-16}$. La inversión publicitaria en radio es significativa.

$$H_0 : \beta_{Diario} = 0 \text{ vs } H_A : \beta_{Diario} \neq 0.$$

Tenemos que $t_{obs} = -0.001037/0.005871 = -0.177$ luego p-valor$= 0.86$. La inversión publicitaria en diario no es significativa.

Regresión lineal múltiple

```
df = summary(adv.lm)$coef
print(cbind(df[, 1:2], confint(adv.lm, level = 0.95)))
```

	Estimate	Std. Error	2.5 %	97.5 %
(Intercept)	2.938889369	0.311908236	**2.32376228**	**3.55401646**
TV	0.045764645	0.001394897	**0.04301371**	**0.04851558**
radio	0.188530017	0.008611234	**0.17154745**	**0.20551259**
newspaper	-0.001037493	0.005871010	**-0.01261595**	**0.01054097**

Límite inferior de los intervalos de confianza de nivel 0.95

Límite superior de los intervalos de confianza de nivel 0.95

Análisis: dejando fijos todos las otras variables por cada $1000 de inversión publicitaria en TV se venderán entre 43 y 48 unidades más.

Retomamos el planteo general

$$H_0 : R\beta = r \text{ vs } H_A : R\beta \neq r.$$

donde $R \in \mathbb{R}^{Jxp}$ donde $rg(R) = J$ y $r \in \mathbb{R}^{Jx1}$.
Bajo H_0 tenemos que

$$\begin{aligned} E(R\widehat{\beta}) &= RE(\widehat{\beta}) = R\beta = r. \\ Var(R\widehat{\beta}) &= RVar(\widehat{\beta})R' = \sigma^2 R(X'X)^{-1}R'. \end{aligned}$$

Luego,

$$\begin{aligned} R\widehat{\beta} &\sim N(R\beta, \sigma^2 R(X'X)^{-1}R') \\ R\widehat{\beta} - r &\sim N(0, \sigma^2 R(X'X)^{-1}R'). \end{aligned}$$

Si calculamos los errores queda,

$$\frac{(R\widehat{\beta} - r)' \left[R(X'X)^{-1}R'\right]^{-1} (R\widehat{\beta} - r)}{\sigma^2} \sim \chi_J$$

siempre que esta inversa exista.

Regresión lineal múltiple

Bajo $H_0 \cup H_A$, los errores están dados por,

$$\frac{(y - X\widehat{\beta})'(y - X\widehat{\beta})}{\sigma^2} = \frac{(n-p)\widehat{\sigma}^2}{\sigma^2} \sim \chi_{n-p}.$$

Luego al compararlos tenemos que,

$$\frac{(R\widehat{\beta} - r)'\left[R(X'X)^{-1}R'\right]^{-1}(R\widehat{\beta} - r)}{J\widehat{\sigma}^2} \sim F_{J,n-p}.$$

Luego, rechazamos H_0 si

$$\frac{(R\widehat{\beta} - r)'\left[R(X'X)^{-1}R'\right]^{-1}(R\widehat{\beta} - r)}{J\widehat{\sigma}^2} > f_{J,n-p,1-\alpha}.$$

Quiero testear si el impacto de invertir en TV y radio es igual

$$H_0: B_{tv}=\beta_{radio} \text{ vs } H_A: B_{tv}\neq\beta_{radio}$$

```
library(car)
linearHypothesis(adv.lm,hypothesis.matrix = c(0,1,-1,0),rhs=0)
```

```
Linear hypothesis test
Hypothesis:
TV - radio = 0

Model 1: restricted model
Model 2: sales ~ TV + radio + newspaper

  Res.Df    RSS      Df   Sum of Sq     F        Pr(>F)
1   197    1308.89
2   196     556.83    1    752.07     264.72    < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Rechazamos la Hip. Nula, las Inversiones en los Dos medios tienen Diferente impacto

Podemos usar las tablas de ANOVA para comparar modelos

#Comparacion del modelo completo y el modelo solamente usando tv

anova(adv.tv.lm,adv.lm)

Analysis of Variance Table

Model 1: sales ~ TV

Model 2: sales ~ TV + radio + newspaper

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	198	2102.53				
2	196	556.83	2	1545.7	272.04	**< 2.2e-16 *****

Agregar las variables radio y Diario mejora el modelo significativa mente

---

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

#Comparacion del modelo completo y modelo sin newspaper

adv.lm.nonews<-lm(sales~TV+radio)

anova(adv.lm.nonews,adv.lm)

Analysis of Variance Table

Model 1: sales ~ TV + radio

Model 2: sales ~ TV + radio + newspaper

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	197	556.91				
2	196	556.83	1	0.088717	0.0312	**0.8599**

Quitar la variable Diario no empeora el modelo.

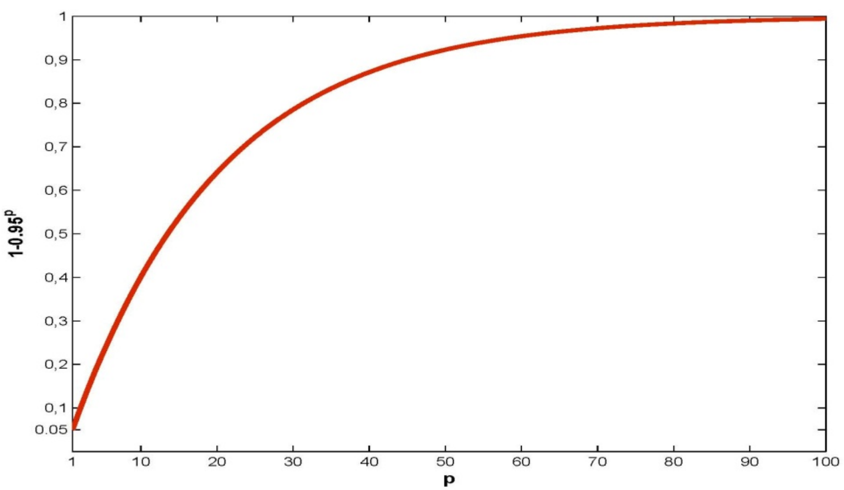
Regresión lineal múltiple

Porqué es necesario mirar el test F? no alcanza con mirar los test individuales?

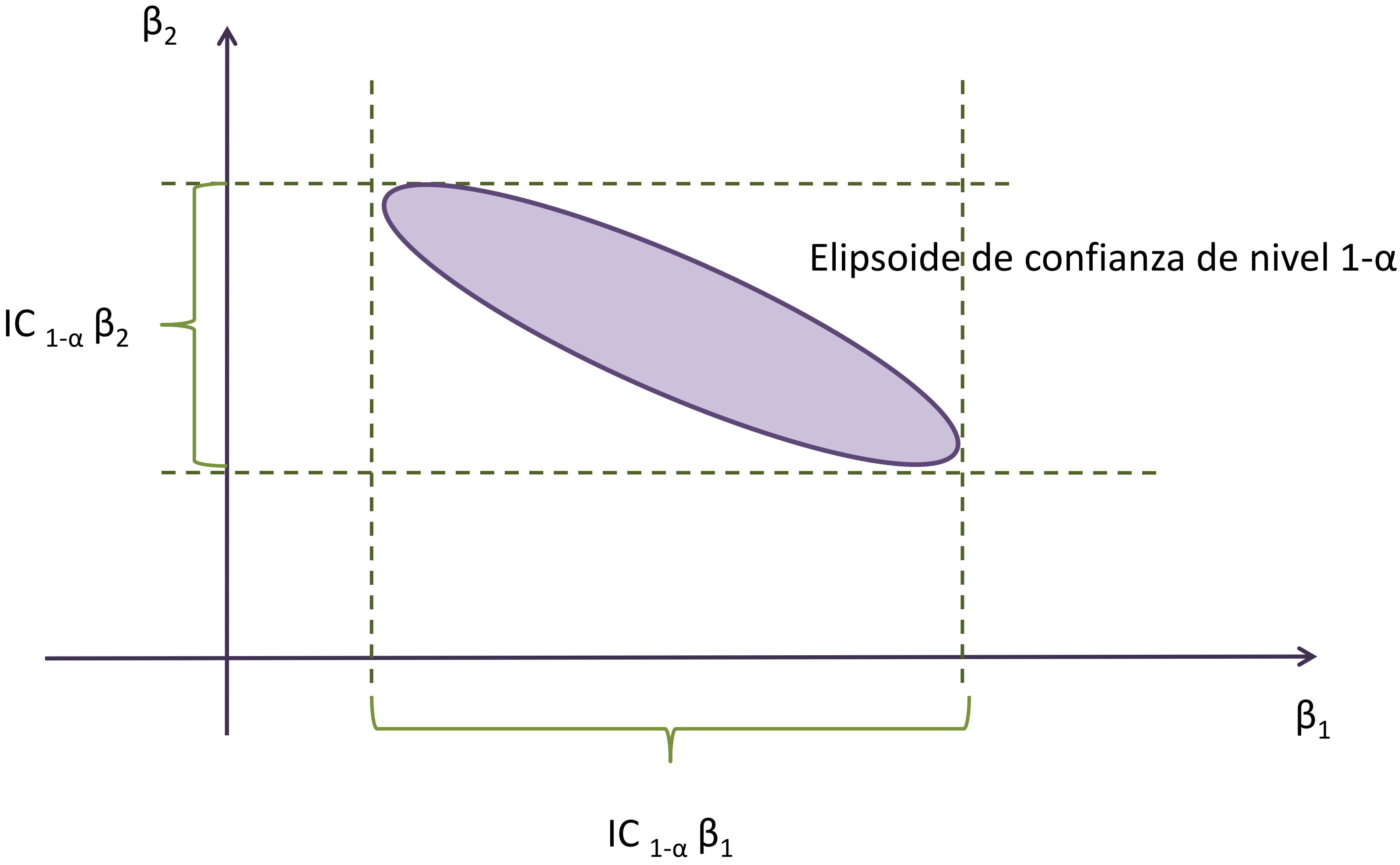
Supongamos que realizamos $H_{01}, \ldots, H_{0p}$ test de nivel α.
Definimos $A_i =$ rechazar el i-ésimo test. Sabemos que $P(A_i) = \alpha$.
Entonces

$$P(\cup A_i) = 1 - P(\cap A_i^c) = 1 - \prod_{i=1}^{p}(1-\alpha) = 1 - (1-\alpha)^p.$$

Luego, si los A_i fueran independientes el nivel global del test es $1-(1-\alpha)^p$ entonces tenemos que hacer tests simultáneos para los parámetros. En este caso al ser no independientes los A_i es más difícil determinar la probabilidad.



Regresión lineal múltiple



Regresión lineal múltiple

Definimos la región de confianza de nivel $1-\alpha$ para el vector β
Sabemos que

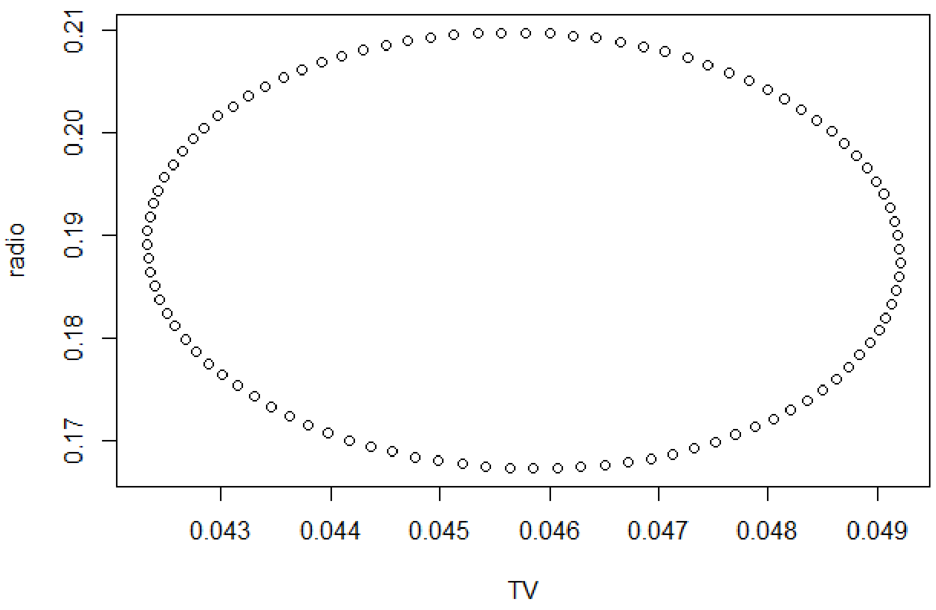
$$\frac{(\widehat{\beta}-\beta)'(X'X)(\widehat{\beta}-\beta)}{pRSS/(n-p)} \sim F_{p,n-p}$$

Luego, la región de confianza de nivel $1-\alpha$ para β esta dada por

$$\frac{(\widehat{\beta}-\beta)'(X'X)(\widehat{\beta}-\beta)}{pRSS/(n-p)} \leq f_{p,n-p;1-\alpha}.$$

###Elipses de confianza, solo se pueden visualizar en dimension 2

```
library(ellipse)
#estoy haciendo la elipse de 95% de confianza para TV y Radio.
plot(ellipse(adv.lm, which=c(2,3), level=0.95))
```



Regresión lineal múltiple

El elipsoide de confianza tiene en cuenta la correlación entre los estimadores de los parámetros.
Si la correlación es alta, entonces el elipsoide resultará *alargado* sino será más redondeado.
Los intervalos de confianza son fáciles de interpretar mientras que el elipsoide no, por ende hay que llegar a un equilibrio analizando la correlación entre los estimadores.

El coeficiente de determinación de un modelo es una medida de bondad de ajuste del mismo y está dado por

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}.$$

Propiedades:

- $0 \leq R^2 \leq 1$.
- Cuando $R^2 = 1$ existe una relación exacta entre la respuesta y las $p - 1$ variables regresoras.
- Cuando $R^2 = 0$ $\widehat{\beta}_0 = \overline{y}$,además $\beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$. No existe relación lineal entre y y las X_i.
- Podemos interpretar R^2 o como un coeficiente de correlación múltiple entre y y las $p - 1$ variables regresoras.
- Se verifica que
$$F = \frac{R^2}{1 - R^2} \frac{n - p}{p - 1}.$$

Regresión lineal múltiple

El coeficiente de determinación para comparar distintos modelos de regresión entre sítiene el siguiente inconveniente: Siempre que se añade una nueva variable regresora al modelo, R^2 aumenta, aunque el efecto de la variable regresora sobre la respuesta no sea significativo. Por ello se define el coeficiente de determinación ajustado o corregido por grados de libertad

$$R^2_{adj} = 1 - \frac{RSS/(n-p)}{TSS/(n-1)} = 1 - \left(\frac{n-1}{n-p}\right)(1-R^2).$$

sólo disminuye al introducir una nueva variable en el modelo si la varianza residual disminuye.

Regresión lineal múltiple

- El R^2_{adj} puede ser negativo. Si $n = 10$, $p = 3$ y $R^2 = 0.16$, entonces $R^2_{adj} = 1 - \left(\frac{9}{7}\right)(1 - 0.16) = -0.08 < 0$. Esto no tiene interpretación.
- En modelos sin intercept no se puede definir el R^2.
- R^2 es sensible a valores extremos, no es robusto.
- R^2 aumenta siempre al aumentar el numero de variables, aunque las mismas sean irrelevantes.

- No es sencillo comparar modelos.
  Supongamos que tenemos, Modelo 1

$$y_i = \beta X_i + \epsilon_i,$$

  y el Modelo 2

$$\log(y_i) = \gamma X_i + \nu_i.$$

  Los respectivos coeficientes de determinación serían

$$R_1^2 = 1 - \frac{\sum_{i=1}^n (y_i - \widehat{y}_i)^2}{\sum_{i=1}^n (y_i - \overline{y})^2}$$

  y

$$R_1^2 = 1 - \frac{\sum_{i=1}^n (\log(y_i) - \widehat{\log(y_i)})^2}{\sum_{i=1}^n (\log(y_i) - \overline{\log(y)})^2}$$

  respectivamente. Estas cantidades no son comparables. Para que fuesen comparables habría que transformar $\log(y)$ en el segundo modelo. Es decir definir,

$$R_3^2 = 1 - \frac{\sum_{i=1}^n (y_i - exp(\widehat{\log(y_i)}))^2}{\sum_{i=1}^n (y_i - \overline{y})^2}.$$

Regresión lineal múltiple

```
adv.lm=lm(sales~TV+radio+newspaper)
summary(adv.lm)

Call:
lm(formula = sales ~ TV + radio + newspaper)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8277 -0.8908  0.2418  1.1893  2.8292

Coefficients:
              Estimate   Std. Error  t value   Pr(>|t|)
(Intercept)   2.938889   0.311908     9.422    <2e-16 ***
TV            0.045765   0.001395    32.809    <2e-16 ***
radio         0.188530   0.008611    21.893    <2e-16 ***
newspaper    -0.001037   0.005871    -0.177      0.86
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared:  0.8972,     Adjusted R-squared:  0.8956
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

RSE

R^2

R^2 ajustado

Regresión lineal múltiple

Analicemos la matriz de correlación

	TV	radio	diario	Ventas
TV	**1**	**0,0548**	**0,0567**	**0,7822**
Radio		**1**	**0,3541**	**0,5762**
Diario			**1**	**0,2286**
Ventas				**1**

Regresión lineal múltiple

Se tiende a invertir más plata en publicidad en diarios en mercados donde se invirtió más plata en publicidad en radios.

Al hacer la regresión simple la inversión publicitaria en diarios saca créditos de la inversión publicitaria en radio. En la múltiple esto se pone en evidencia.

Hay más ataques de tiburones en las playas donde se venden más helados.

Regresión lineal múltiple

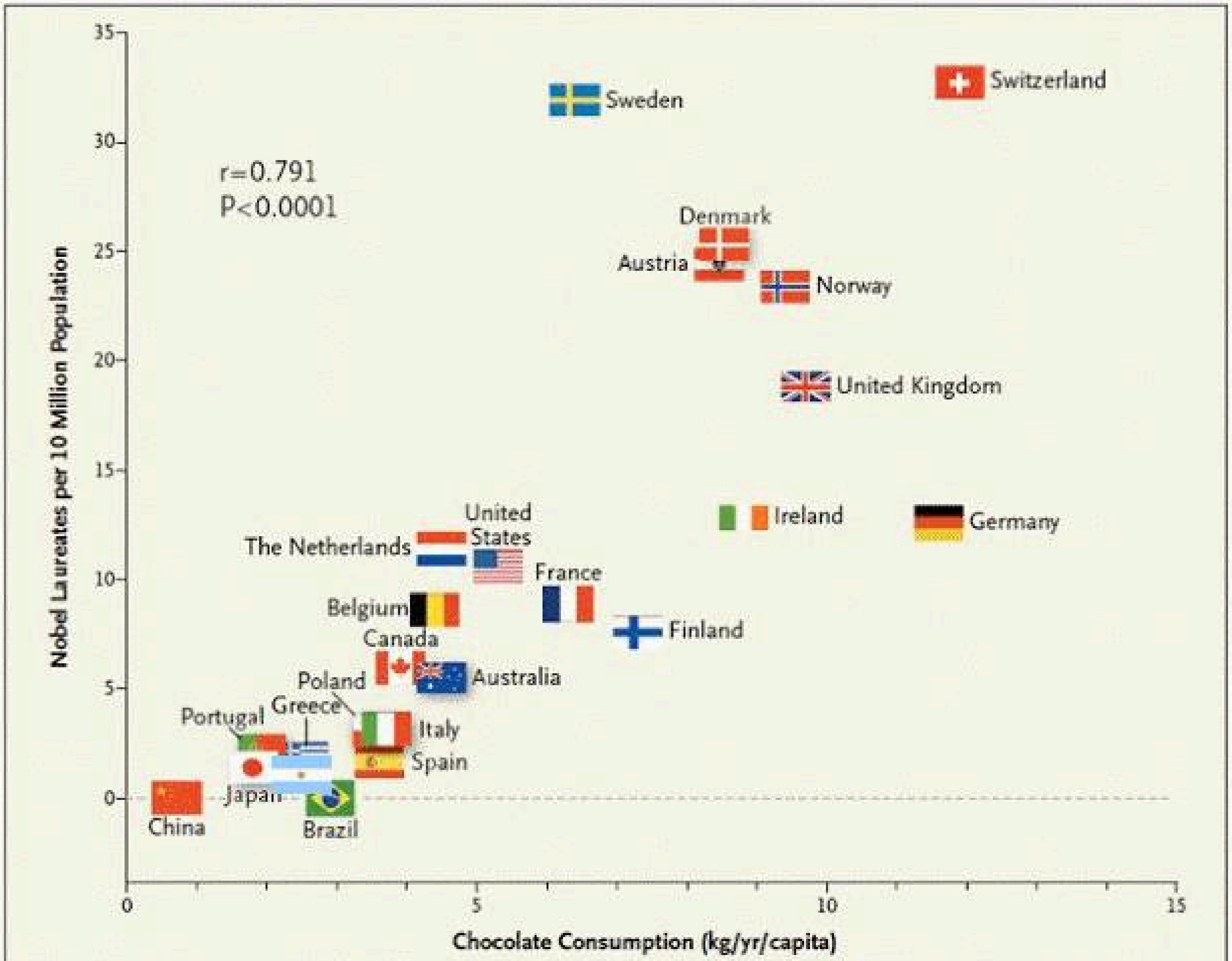


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Regresión lineal múltiple

Todo sugiere que la variables diario no aporta nada al modelo. Consideremos entonces un nuevo modelo sin ella.

$$ventas_i = \beta_0 + \beta_1 TV_i + \beta_2 radio_i + \epsilon_i \text{ para } i = 1, \ldots, 200.$$

Regresión lineal múltiple

```
adv.lm.2=lm(sales~TV+radio)
summary(adv.lm.2)

Call:
lm(formula = sales ~ TV + radio)

Residuals:
    Min      1Q  Median      3Q     Max
-8.7977 -0.8752  0.2422  1.1708  2.8328

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.92110    0.29449   9.919   <2e-16 ***
TV           0.04575    0.00139  32.909   <2e-16 ***
radio        0.18799    0.00804  23.382   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.681 on 197 degrees of freedom
Multiple R-squared:  0.8972,	Adjusted R-squared:  0.8962
F-statistic: 859.6 on 2 and 197 DF,  p-value: < 2.2e-16
```

Regresión lineal múltiple

Comparemos los dos modelos

	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	RSE	R^2	R^2_{adj}
Con diario	2.94	0.046	0.188	-0.001	1.686	0.8972	0.8956
Sin diario	2.92	0.046	0.188	—	1.681	0.8972	0.8962

- Prácticamente no hay diferencias entre los R^2 y los R^2_{adj}.
- Los $\widehat{\beta}$ casi no cambian.

Nos quedamos con el modelo que predice las ventas como función de la inversión en TV y en el diario, seleccionamos en forma heurística las variables. Más adelante en la materia abordaremos este tema en detalle.

Predicción

Sean $\widehat{\beta_0}, \ldots, \widehat{\beta_{p-1}}$ estimadores de $\beta_0, \ldots, \beta_{p-1}$, hay tres factores vinculados con la incertidumbre.

- **Error reducible** El hiperplano ajustado por mínimos cuadrados está dado por

$$\widehat{y} = \widehat{\beta_0} + \widehat{\beta_1} X_1 + \cdots + \widehat{\beta_{p-1}} X_{p-1},$$

  que es una estimación del modelo

$$f(X_1, \ldots, X_{p-1}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}.$$

- El modelo lineal es una aproximación a la realidad, hay un sesgo de selección de modelo. Podría haber modelos mejores.
- **Error irreducible** Aún conociendo explícitamente a $f(X)$ no se puede predecir con exactitud y, puesto que ϵ es desconocido.

Regresión lineal múltiple

Se pueden construir intervalos de confianza de nivel $1-\alpha$ para los valores predichos, $\widehat{y}$. Para tratar este problema podemos hacer inferencia sobre la variables de respuesta. Se tienen dos problemas estrechamente ligados.

- **Estimar el valor medio de** y para los individuos de la población donde $X = x_0$.
- **Predecir el valor individual que tomará la variable** y para una nueva observación $X = x_0$.

Regresión lineal múltiple

Predicción del valor medio de y

Necesitamos predecir el valor de y en un punto dado $x_0 = (x_{01}, \ldots, x_{0p})$.

El estimador puntual está dado por $\widehat{y}_0 = \widehat{\beta} x_0$.

Luego su esperanza es $E(\widehat{y}_0) = E(\widehat{\beta} x_0) = \beta x_0$, que es un estimador insesgado de $E(y)$.

Su varianza es $var(\widehat{y}_0) = var(\widehat{\beta} x_0) = \sigma^2 x_0'(X'X)^{-1} x_0 = \sigma^2 \nu_0$.

Se puede deducir un intervalo de confianza de nivel $1 - \alpha$ está para $E(y|x_0)$ está dado por

$$\widehat{\beta} x_0 \pm t_{n-p;1-\alpha/2} RSE \sqrt{\nu_0},$$

donde RSE es el estimador de σ^2 y está dado por $RSE = \sqrt{\frac{RSS}{n-p}}$.

Regresión lineal múltiple

Predicción de un valor individual de y

Necesitamos predecir el valor de y en un punto dado $x_0 = (x_{01}, \ldots, x_{0\cancel{p}})$. p-1

El estimador puntual está dado por $\widehat{y}_0 = \widehat{\beta} x_0 + \epsilon_0$. Luego su esperanza es $E(\widehat{\beta} x_0 + \epsilon_0) = \beta x_0$, coincide con el caso anterior. Su varianza es $var(\widehat{\beta} x_0) = \sigma^2 \left(1 + x_0'(X'X)^{-1} x_0\right) = \sigma^2(1 + \nu_0)$. Luego, se puede deducir un intervalo de predicción de nivel $1 - \alpha$ está para $E(y|x_0)$ está dado por

$$\widehat{\beta} x_0 \pm t_{n-p;1-\alpha/2} RSE \sqrt{1 + \nu_0}.$$

Regresión lineal múltiple

Qué problema es mas difícil de los dos?

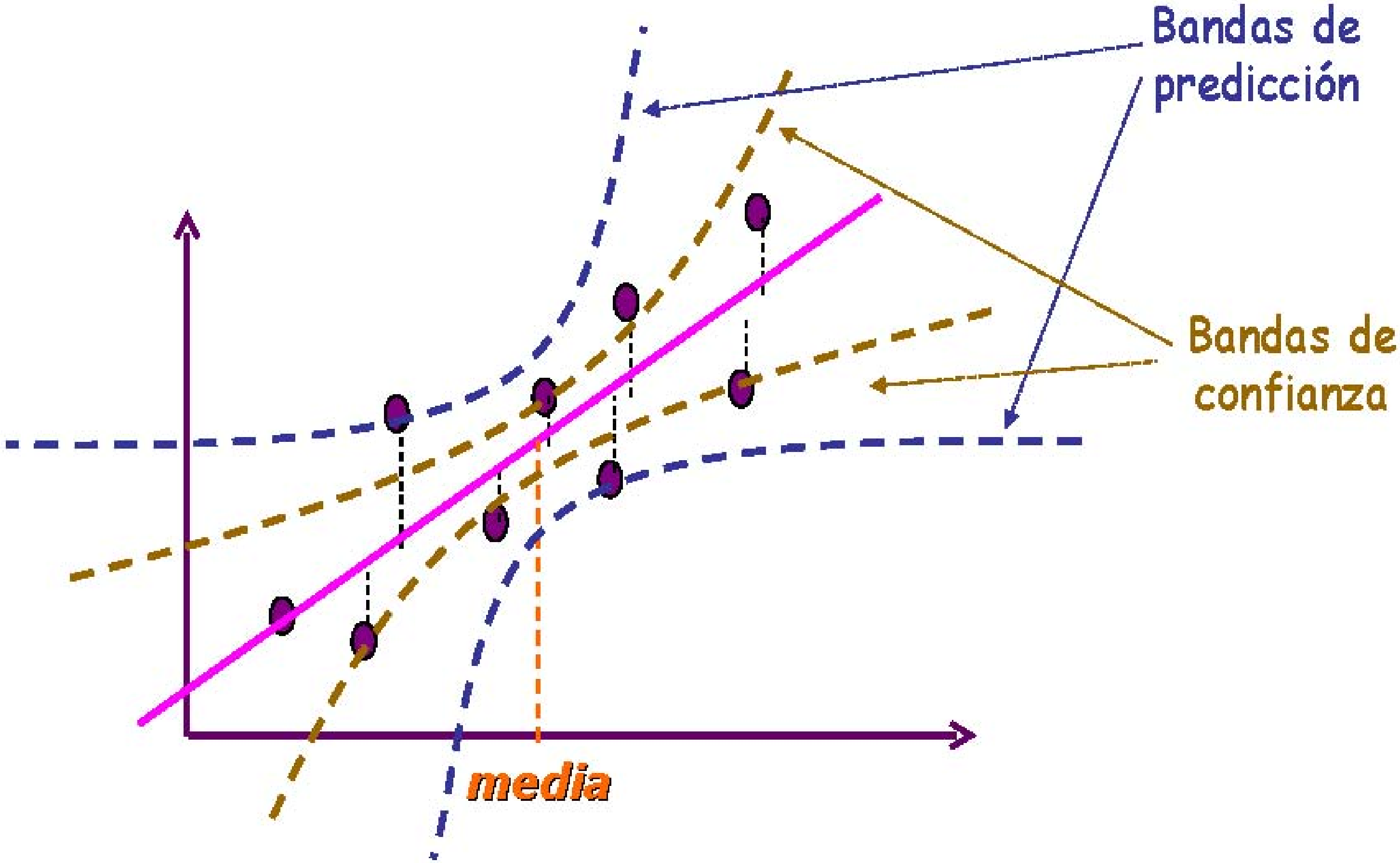
Qué estimador y que predicción resultan razonables para $\boldsymbol{\mu_0}$ y Y_0?

En ambos casos, el estimador (o predicción) puntual es:

$$\widehat{\mathrm{Y}}_0 = \boldsymbol{x_0}\widehat{\boldsymbol{\beta}}.$$

Sin embargo, el intervalo de confianza para $\boldsymbol{\mu_0}$ es diferente del intervalo de predicción para Y_0.

Regresión lineal múltiple



Regresión lineal múltiple

En nuestro ejemplo…

Calculamos el intervalo de confianza para cuantificar la venta media a lo largo de las ciudades. Si se invierten $100000 en publicidad en TV y $20000 en publicidad en radio en cada ciudad se espera que la venta promedio este entre [11985, 11528] unidades a nivel 0.95.

*El 95% de los intervalos de esta forma contienen el verdadero valor de f(**X**)*

Regresión lineal múltiple

El intervalo de predicción se utiliza para cuantificar a incertidumbre en una ciudad específica, con las mismas inversiones del ejemplo anterior el intervalo de predicción está dado por [7930, 14580] unidades.

El 95% de los intervalos de esta forma contienen el verdadero valor de las ventas Y para esa ciudad.

Regresión lineal múltiple

```
sales_fit=predict(adv.lm.2, newdata = data.frame(TV=100,radio=20),
    interval = c("confidence"), level = 0.95)
sales_fit
       fit        lwr       upr
1 11.25647  10.98525  11.52768
sales_pred=predict(adv.lm.2, newdata = data.frame(TV=100,radio=20),
    interval = c("predict"), level = 0.95)
sales_pred
       fit        lwr       upr
1 11.25647  7.929616  14.58332
```

Regresión lineal múltiple

Siempre ocurre que el intervalo de predicción es más largo que el intervalo de confianza, aunque están centrados en el mismo valor.

Esto ocurre porque la incertidumbre aumenta cuando hablamos de las ventas un mercado en particular y no de las ventas en promedio en muchos mercados.

Regresión lineal múltiple

Extrapolación

- Los modelos lineales tienen validez en el rango de valores estudiados y no fuera del mismo.
- En un contexto multivariado, hay que ser precavido ya que marginalmente la observación puede estar en el rango estimado pero no en forma conjunta y si se predice en ese caso se estaría incorrectamente extrapolando.

Regresión lineal múltiple

- An introduction to Statistical Learning, 7th ed. Gareth, J., Witten, D., Hastie, T., Tibshirani, R., (2013), Springer. Capitulos 2 y 3.
- Elements of Statistical Learning, Hastie 2nd Ed, T., Tibshirani, R., Friedman, J., (2009), Springer. Capítulo 3.1 y 3.2.
- Linear Models and Generalizations: Least Squares and Alternatives, 3rd ed. Radhakrishna Rao C., Shalabh H. T., Heunaman (2008), Springer. Capitulo 3.