



MA34B – Estadística

Modelos Lineales

Prof. Rodrigo Abt B.

rabt@dim.uchile.cl

Introducción

- En muchas oportunidades a un investigador le interesa saber cómo el comportamiento de un conjunto de variables impacta sobre otra, no solo desde el punto de la influencia o grado de asociatividad, sino que describir la posible relación funcional entre las mismas.
- Muchas leyes de ciencias experimentales como la Física y la Biología intentan construir modelos basados en experimentos en que interviene el azar, lo que hace impracticable una formulación determinista.
- Es decir dado un conjunto de variables explicativas $X^{(1)}, X^{(2)}, \dots, X^{(p)}$ y una variable de interés Y , se intenta determinar la relación que existe entre ellas a través de una forma funcional f :

$$Y = f(X^{(1)}, X^{(2)}, \dots, X^{(p)})$$

- Un tipo de forma funcional de interés es la forma lineal, es decir, aquella en que Y depende linealmente de las “ p ” variables explicativas:

$$Y = \beta_0 + \beta_1 X^{(1)} + \beta_2 X^{(2)} + \dots + \beta_p X^{(p)}$$

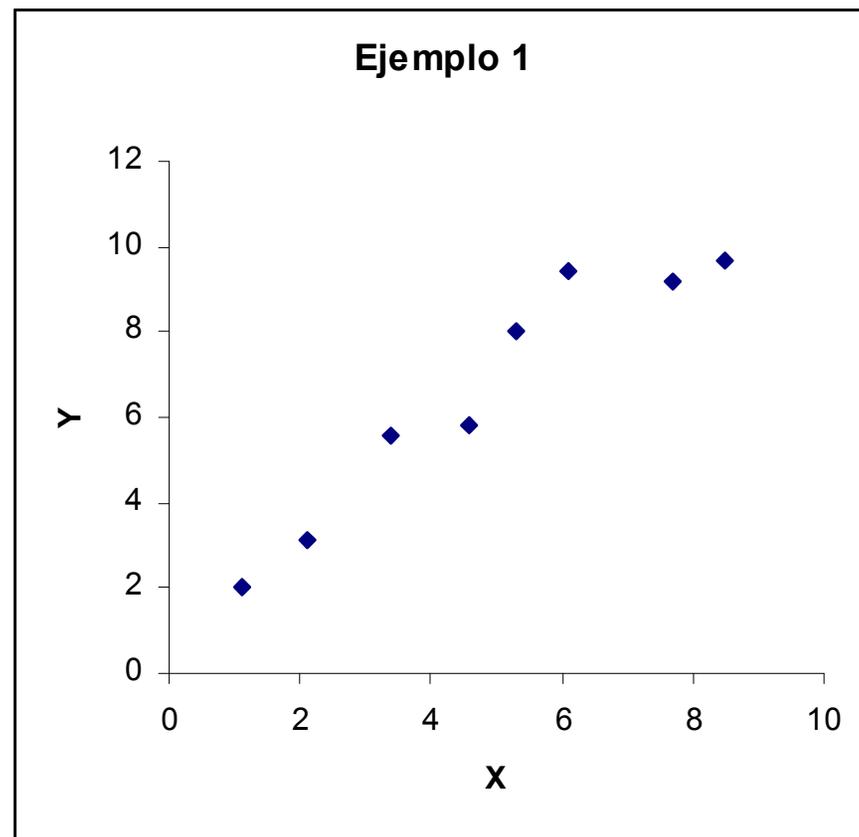
Caso: 1 variable explicativa (1)

- Supongamos que hemos recopilado la siguiente tabla con observaciones:

obs	X	Y
1	1,1	2
2	2,1	3,1
3	3,4	5,6
4	4,6	5,8
5	5,3	8
6	6,1	9,4
7	7,7	9,2
9	8,5	9,7

Caso: 1 variable explicativa (2)

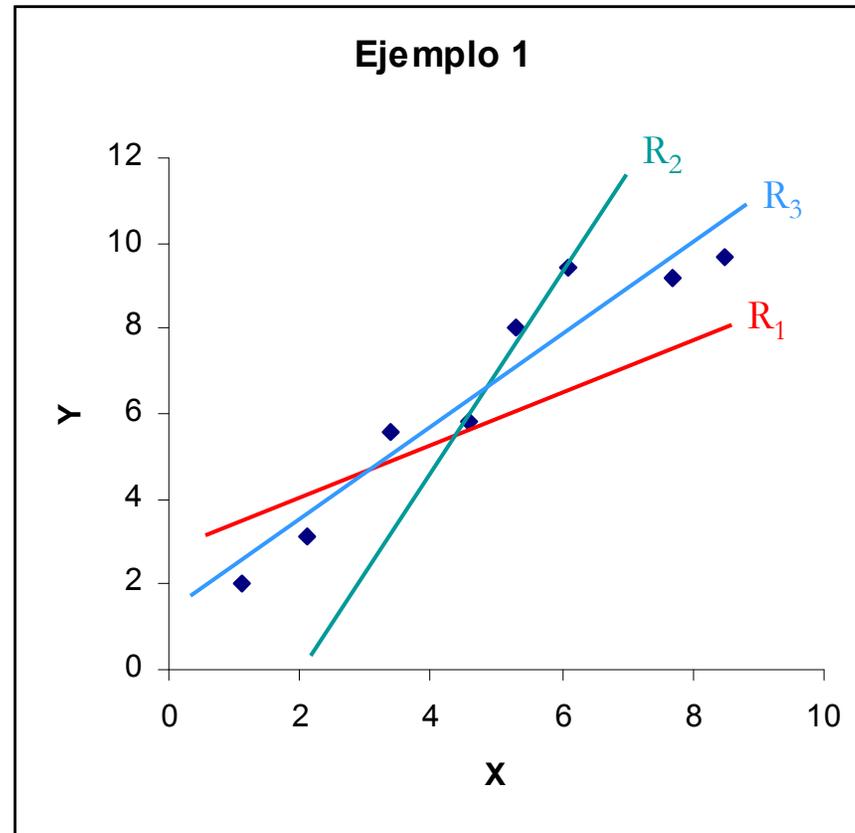
- Si graficamos los valores:



Caso: 1 variable explicativa (3)

- Se observa un tendencia lineal positiva (lo cual puede ser corroborado con el coeficiente de correlación lineal).
- Podríamos suponer entonces que Y es una función lineal de X , es decir $Y = a + bX$.
- El problema es, ¿cómo determino los valores de a y b ?
- Podríamos probar al ojo primero:

Caso: 1 variable explicativa (4)



Caso: 1 variable explicativa (5)

- De las 3 rectas representadas, R_1 , R_2 y R_3 , ¿Cuál es la que mejor representa la relación? ¿Qué criterio utilizaría para determinar la mejor recta?
- La intuición nos dice que aquella recta que se encuentre “más cerca” de los puntos será mejor.
- Una manera de medir “cercanía” es a través de distancias.
- Sea $\hat{Y} = a + bX$ la recta buscada.
- Si definimos $\varepsilon_i = Y_i - \hat{Y}_i$ como la diferencia entre lo observado y lo predicho, esperaremos que estas diferencias sean pequeñas.

Criterio de los Mínimos Cuadrados

- Sea

$$Q = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - (a + bx_i))^2$$

- Entonces buscamos valores de a y b tales que Q sea mínimo.
- Basta entonces con derivar e igualar a 0:

$$\frac{\partial Q}{\partial a} = 0 \quad \frac{\partial Q}{\partial b} = 0$$

- Y obtenemos un sistema de ecuaciones que al resolver nos da:

$$\hat{b} = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sum x_i^2 - n \bar{x}^2} \quad \hat{a} = \bar{y} - \hat{b} \bar{x}$$

Observaciones

- Este método se denomina Mínimos Cuadrados Ordinarios (MCO)
- Este es un método NUMÉRICO que nos proporciona el mejor ajuste de coeficientes para un conjunto de datos.
- Como no hacemos suposiciones estadísticas respecto del problema, no es posible llevar a cabo estimaciones ni tests de hipótesis.

Caso Multivariado

- Consideremos un modelo lineal general con un conjunto de “k” variables independientes o “regresores”:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \varepsilon_i \quad \forall i = 1, \dots, n$$

- La expresión anterior se puede escribir matricialmente como:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \dots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$y = X\beta + \varepsilon$

Solución MCO (1)

- Para encontrar la recta de mejor ajusta, podemos utilizar el mismo criterio que en el caso bivariado, es decir buscar $\beta_1, \beta_2, \dots, \beta_k$ tales que $Q = \sum \varepsilon_i^2$ sea mínimo, es decir, se tiene que resolver:

$$\text{Min } Q = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - (\beta_1 x_{i1} + \beta_2 x_{i2} + \dots, \beta_k x_{ik}))^2$$

- Derivando con respecto a cada coeficiente β , e igualando a 0, se tiene un sistema de “k” ecuaciones.

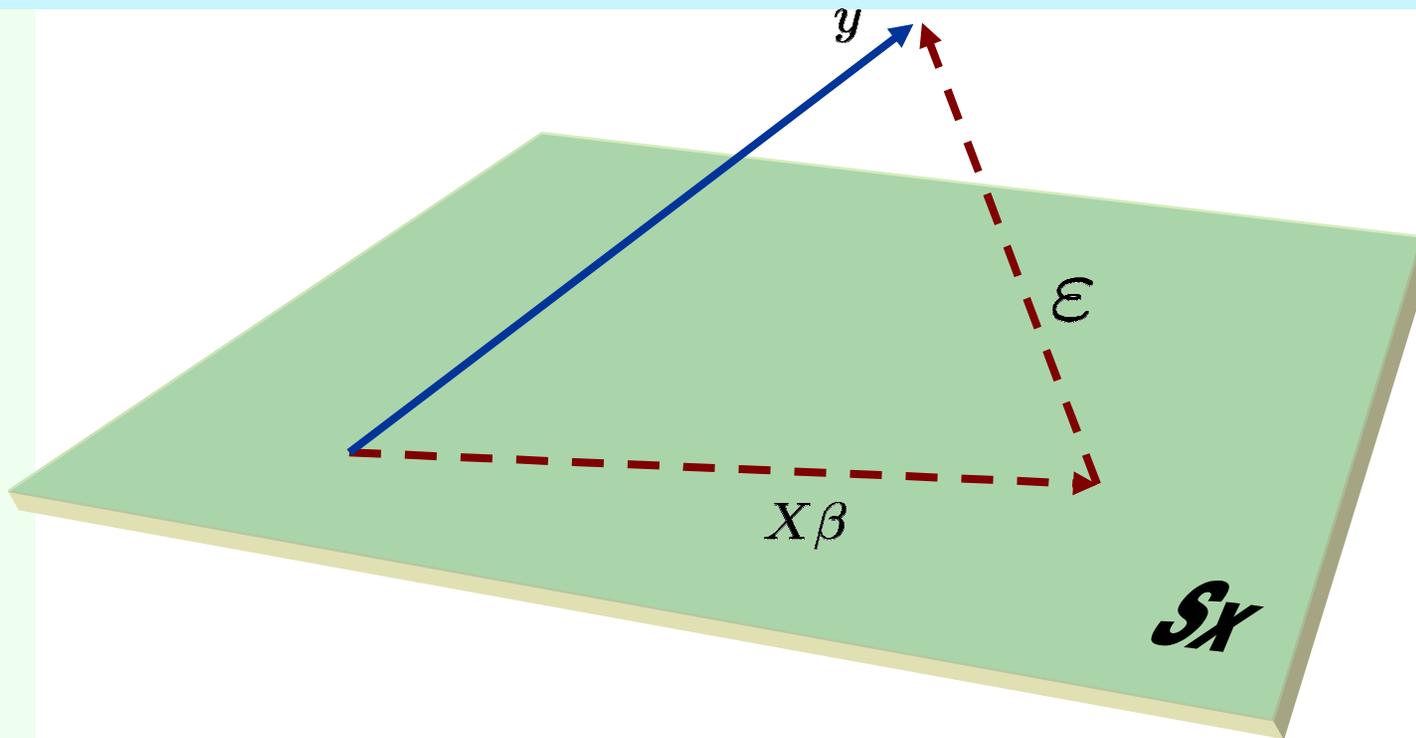
Solución MCO (2)

- El resultado del sistema, escrito de manera matricial es:

$$\hat{\beta} = (X^t X)^{-1} X^t y$$

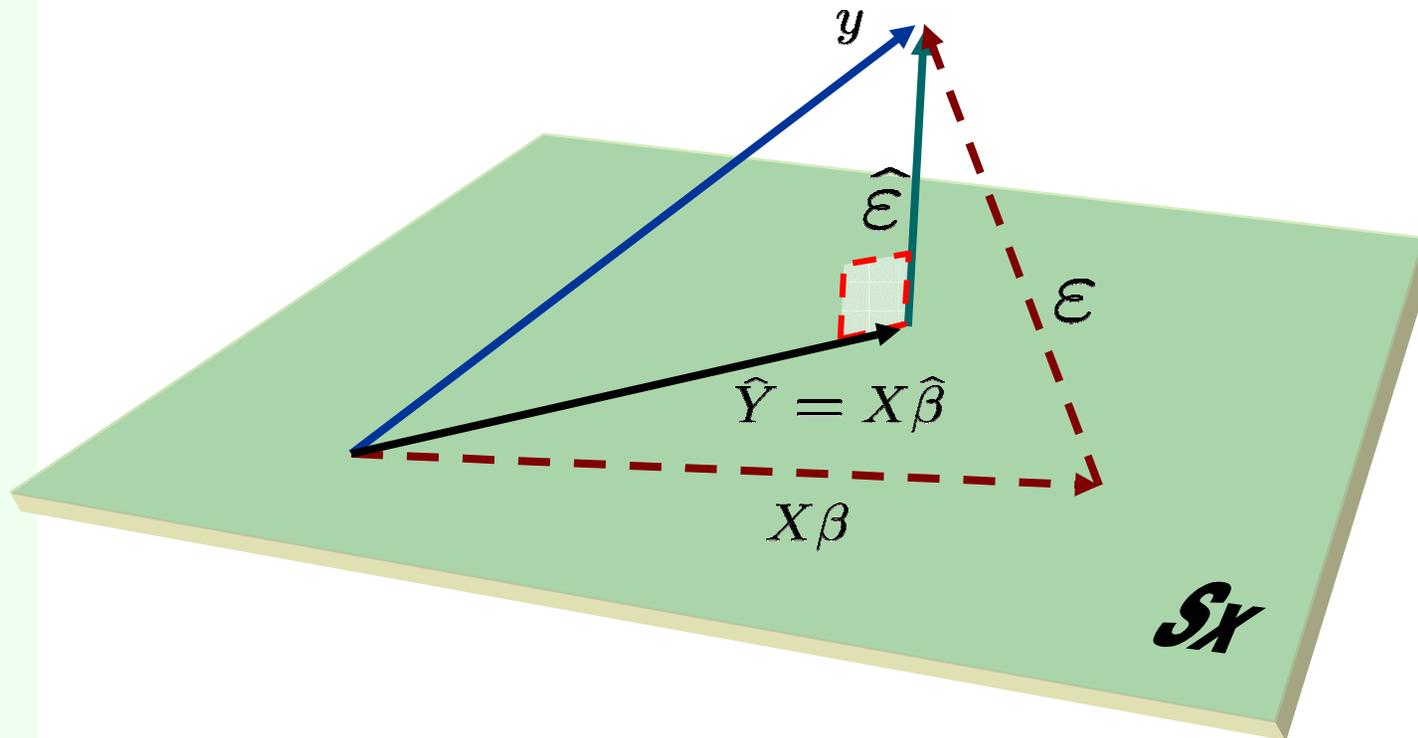
- Este resultado requiere que la matriz $X^t X$ sea invertible, es decir, que las columnas de X sean l.i.
- NOTA: Este resultado se puede obtener directamente aplicando derivadas a la expresión matricial.

Interpretación Geométrica (1)



- Se busca aproximar y que es de dimensión “n” por un vector $\hat{y} = \mathbf{X}\beta$ (de dimensión “k”) contenido en el subespacio generado por las columnas de \mathbf{X} (S_X).
- Para constituir una “buena” aproximación, el vector \hat{y} debe encontrarse “cerca” de y dentro del subespacio S_X .
- El vector requerido se obtiene proyectando ortogonalmente y sobre el subespacio S_X .

Interpretación Geométrica (2)



- Si \mathbf{P} es el operador de proyección de \mathbf{y} sobre $S_{\mathbf{X}}$, entonces $\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{P}\mathbf{y}$
- El operador \mathbf{P} se puede encontrar en función de \mathbf{X} , dado que el vector $\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{P}\mathbf{y}$ es ortogonal a cada columna de \mathbf{X} .

Interpretación Geométrica (3)

- Se cumple entonces que:

$$\langle y - Py, X^{(j)} \rangle = \langle y - X\beta, X^{(j)} \rangle = 0 \quad \forall j = 1, \dots, k$$

- De manera matricial:

$$X^t(y - X\beta) = 0 \Rightarrow X^t y - X^t X \beta = 0$$

- Esta última expresión da origen al sistema de “ecuaciones normales”.
- Si $X^t X$ es invertible, entonces se obtiene:

$$\hat{\beta} = (X^t X)^{-1} X^t y$$

Propiedades

- Con lo anterior podemos deducir el operador de proyección \mathbf{P} como:

$$P = X(X^t X)^{-1} X^t$$

- El operador así determinado cumple lo siguiente:

- El \mathbf{P} es simétrico, es decir, $\mathbf{P}^t = \mathbf{P}$
- \mathbf{P} es idempotente de orden 2, es decir, $\mathbf{P}^2 = \mathbf{P}$
- \mathbf{P} es de dimensión k
- El operador $\mathbf{H} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t = \mathbf{I} - \mathbf{P}$ es tal que
 - $\mathbf{H} \perp \mathbf{P}$
 - \mathbf{H} es simétrico e idempotente de orden 2.
 - \mathbf{H} es dimensión $n-k$
 - $\mathbf{H}\mathbf{X} = 0$
 - $\hat{\varepsilon} = \mathbf{H}\mathbf{y} = \mathbf{H} \varepsilon$

Supuestos del Modelo Lineal (1)

- Los resultados anteriores son el resultado de condiciones geométrico-algebraicas, por lo que no tienen validez para el tratamiento de estimación y contraste de hipótesis.
- Desde el punto de vista estadístico se hacen los siguientes supuestos básicos para el modelo lineal:
 - El modelo es estocástico
 - El modelo es lineal en los coeficientes β .
 - La matriz \mathbf{X} es determinista y de rango completo
 - La media condicional de los errores respecto de los regresores es 0
 - Los errores son independientes y de varianza constante (esfericidad)
 - Los errores son normales

Supuestos del Modelo Lineal (2)

- Los 3 primeros supuestos indican que la variable y se considera como aleatoria, y que se modela a través de variables consideradas como “controlables” por el investigador (deterministas). Se supone además que el número de observaciones es superior al número de coeficientes a determinar ($n \gg k$). Esto es para evitar problemas de ajuste calzado o sistemas de ecuaciones indeterminados (problema de identificación).

Supuestos del Modelo Lineal (3)

- El cuarto supuesto se puede expresar como

$$E(\varepsilon_i) = 0 \quad \forall i = 1, \dots, n$$

- Esto indica que no existe un aporte **sistemático** de los errores en el modelo.

- El quinto supuesto se puede expresar como:

$$Cov(\varepsilon_i, \varepsilon_j) = 0 \quad \forall i \neq j$$

$$Var(\varepsilon_i) = \sigma^2 \quad \forall i = 1, \dots, n$$

- La matriz de varianza-covarianzas del error es:

$$Var(\varepsilon) = \sigma^2 I_n$$

Supuestos del Modelo Lineal (4)

- Estas son las condiciones de ausencia de correlación, y homocedasticidad en los errores.
- Finalmente, el último supuesto se expresa como:

$$\varepsilon_i \sim N(0, \sigma^2) \quad \forall i = 1, \dots, n$$

- Vectorialmente:

$$\varepsilon \sim N_n(0, \sigma^2 I_n)$$

- **NOTA:** El último supuesto es necesario solamente para llevar a cabo tests de hipótesis y procedimientos similares. Sin embargo, se pueden estudiar las propiedades del estimador MCO de los coeficientes sin necesidad de recurrir al mismo

Propiedades Estimador MCO

- Calculemos esperanza y varianza del estimador MCO de β .

$$\begin{aligned} E(\hat{\beta}) &= E[(X^t X)^{-1} X^t y] = (X^t X)^{-1} X^t E[y] \\ &= (X^t X)^{-1} X^t (X\beta + \varepsilon) = \beta \end{aligned}$$

- Luego el estimador MCO es insegado.
- La varianza del estimador MCO es:

$$Var(\hat{\beta}) = Var[(X^t X)^{-1} X^t y] = \sigma^2 (X^t X)^{-1}$$

- El error estimado se denomina **residuo**, y se obtiene como:

$$\hat{\varepsilon} = y - \hat{y} = y - X\hat{\beta} = Hy = H\varepsilon$$

Teorema de Gauss Markov

- Otra de las gracias del estimador MCO, es que dentro de todos los estimadores insesgados lineales en y , el estimador MCO es el de menos varianza.
- Teorema de Gauss Markov:

Si $E(\varepsilon) = 0$, y $Var(\varepsilon) = E(\varepsilon\varepsilon^t) = \sigma^2 I_n$, entonces la combinación lineal $a^t \hat{\beta}$ es de varianza mínima entre todos los estimadores insesgados lineales en y para $a^t \beta$

Estimación Máximo Verosímil

- Si suponemos normalidad de los errores, es decir, $\varepsilon_i \sim N(0, \sigma^2)$, se tiene que el estimador MCO coincide con el de máxima verosimilitud, esto es:

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

- Sin embargo el EMV para σ^2 es sesgado. Para esto se propone el siguiente estimador corregido:

$$\tilde{\sigma}^2 = \frac{\hat{\varepsilon}^t \hat{\varepsilon}}{n - k}$$

Hipótesis sobre coeficientes del modelo (1)

- Dado que los estimadores MCO son insesgados y lineales en \mathbf{y} , eso significa que son normales, y por ende los residuos también:

$$\hat{\beta} \sim N_n(\beta, \sigma^2(X^t X)^{-1})$$

$$\hat{\varepsilon} \sim N_n(0, \sigma^2 H \cdot I_n)$$

- Los tests de hipótesis sobre coeficientes se llevan a cabo con los estadísticos muestrales conocidos: t-Student, Chi-cuadrado y F-Fisher. En este caso se tiene que:

$$\frac{\hat{\beta}_i - \beta_i^{(0)}}{\sigma \sqrt{(X^t X)^{-1}_{ii}}} \sim N(0, 1)$$

$$\frac{(n - k)\tilde{\sigma}^2}{\sigma^2} \sim \chi_{n-k}^2$$

Hipótesis sobre coeficientes del modelo (2)

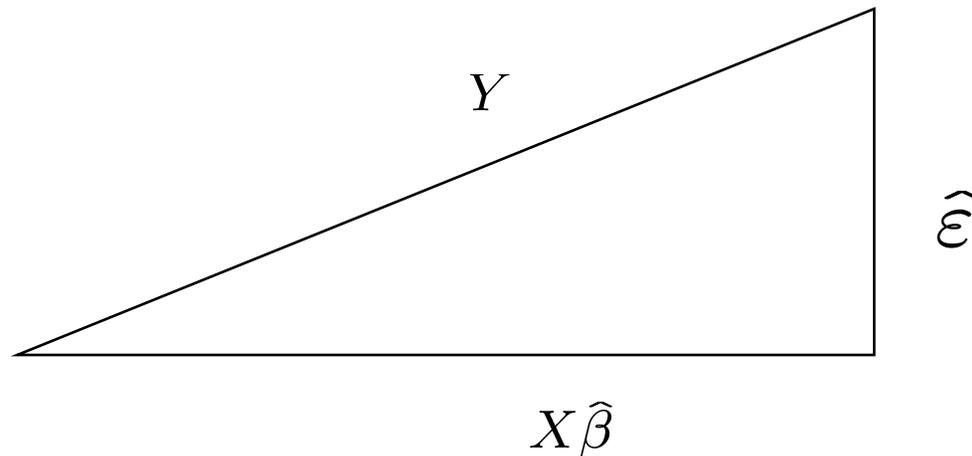
- Combinando ambas expresiones se tiene que:

$$\frac{\hat{\beta}_i - \beta_i^{(0)}}{\tilde{\sigma} \sqrt{(X^t X)^{-1}_{ii}}} \sim t_{n-k}$$

- Y lo que usualmente se prueba es si el coeficiente en cuestión es 0 o no (significancia individual), y con esto se puede determinar si la variable correspondiente tiene peso o no en el modelo.
- Este estadístico además se puede utilizar para encontrar intervalos de confianza para un coeficiente en particular.

Análisis De Varianza En El Modelo Lineal

- Recordemos la situación geométrica del estimador MCO:



- Por Pitágoras, tenemos que:

$$\|Y\|^2 = \|X\hat{\beta}\|^2 + \|\hat{\varepsilon}\|^2$$

$$Y^t Y = (X\hat{\beta})^t (X\hat{\beta}) + \hat{\varepsilon}^t \hat{\varepsilon}$$

Test De Significación Global (1)

- Sean $SSE = \hat{\beta}^t X^t X \hat{\beta}$ y $SSR = \hat{\varepsilon}^t \hat{\varepsilon}$
- Si $\beta = 0$, entonces: $\frac{\hat{\beta}^t X^t X \hat{\beta}}{k} = \frac{SSE}{k}$ es insesgado para σ^2
- Independiente de β , se tiene que: $\frac{\hat{\varepsilon}^t \hat{\varepsilon}}{n - k} = \frac{SSR}{n - k}$ es insesgado para σ^2
- Luego es de esperar que el estadístico:
$$F = \frac{\frac{SSE}{k}}{\frac{SSR}{n - k}}$$

se encuentre cerca de 1 si suponemos $H_0: \beta = 0$, pero mayor que 1 si la hipótesis es falsa.

Test De Significación Global (2)

- Bajos los supuestos de normalidad de los errores, entonces se puede probar que el estadístico F sigue una distribución F-Fisher.
- En general los modelos se suponen con constante. Supongamos un modelo con “k” coeficientes (incluida la constante):

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_{k-1} x_{k-1,1} + \varepsilon_i$$

- Sean: $\tilde{y}_i = y_i - \bar{y}$ y $\tilde{x}_{ij} = x_{ij} - \bar{x}_j$

- Se puede escribir el modelo equivalente:

$$\tilde{y}_i = \beta_1 \tilde{x}_{i1} + \beta_2 \tilde{x}_{i2} + \dots + \beta_{k-1} \tilde{x}_{k-1,1} + \varepsilon_i$$

- Y se tiene que

$$SST = \sum (y_i - \bar{y})^2 \quad SSE = \sum (\hat{y}_i - \bar{y})^2 \quad SSR = \sum (y_i - \hat{y}_i)^2$$

donde $SST = SSE + SSR$

- Y por ende, el estadístico F es:

$$F = \frac{\frac{SSE}{k-1}}{\frac{SSR}{n-k}} = \frac{\frac{\sum (\hat{y}_i - \bar{y})^2}{k-1}}{\frac{\sum (y_i - \hat{y}_i)^2}{n-k}} \sim F_{k-1, n-k}$$

Test De Significación Global (3)

- Luego el estadístico F anterior sirve para testear las hipótesis:

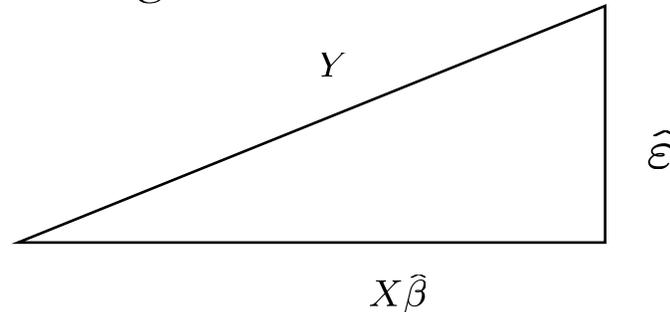
$$H_0 : \beta_1 = \beta_2 = \dots = \beta_{k-1} = 0$$

$$H_1 : \text{Alguno de los } \beta_i \neq 0$$

- A esto se le denomina test de validez global del modelo.
- **IMPORTANTE:** Este test solo tiene validez cuando existe constante en el modelo.

El Coeficiente De Determinación R^2 (1)

- Miremos nuevamente la geometría del modelo con MCO:



- Otra forma de medir que tan cerca la recta estimada del plano es mirando la relación angular entre ambos.

- Sea

$$R_u^2 = \cos^2 \theta = \frac{\|\hat{Y}\|^2}{\|Y\|^2}$$

- Si la recta está cerca del plano (buen ajuste), entonces R_u^2 tiende a 1, y si está lejos tenderá a 0. Sin embargo, bajo la presencia de constante en el modelo este indicador puede traer distorsiones.

El Coeficiente De Determinación R^2 (2)

- Para modelos con constante, se utiliza otra versión de R^2 :

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

y es este indicador el que reportan los paquetes estadísticos. Se denomina Coeficiente de Determinación o Coeficiente de Correlación Lineal Múltiple.

- A su vez R^2 y F se encuentran relacionados:

$$F = \frac{(n - k)R^2}{(k - 1)(1 - R^2)}$$

Hipótesis Sobre Combinación De Coeficientes

- Suponga que se quiere llevar a cabo un test sobre una combinación lineal de coeficientes:

- $H_0 : a^t \beta = a^t \beta_0$

- $H_1 : a^t \beta \neq a^t \beta_0$

- Se tiene que: $E(a^t \hat{\beta}) = a^t \beta$

$$Var(a^t \hat{\beta}) = a^t \sigma (X^t X)^{-1} a$$

- De donde:

$$t = \frac{\frac{a^t \hat{\beta} - a^t \beta}{\sqrt{a^t \sigma (X^t X)^{-1} a}}}{\sqrt{\frac{(n-k) \tilde{\sigma}^2}{\sigma^2} / (n-k)}} = \frac{a^t \hat{\beta} - a^t \beta_0}{\tilde{\sigma} \sqrt{a^t (X^t X)^{-1} a}} \sim t_{n-k}$$

Comparación Entre Modelos

- El test F Fisher también tiene aplicación para comparar modelos, cuando se tiene la sospecha que un CONJUNTO de variables no es significativa.
- Si un conjunto de variables no es significativa, debería esperarse que la variabilidad aportada por ambos modelos debe ser similar, o que las sumas residuales (variabilidad aportada por los residuos) sean similares.
- Se puede demostrar que:

$$\frac{\frac{SSR_r - SSR_c}{k_c - k_r}}{\frac{SSR_c}{n - k_c}} \sim F_{k_c - k_r, n - k_c}$$

- En que SSR_c es la suma residual del modelo sin restringir (completo), y SSR_r es la suma residual del modelo restringido (con variables eliminadas). Se tiene que k_c es la cantidad de coeficientes en el modelo completo, y k_r la cantidad de coeficientes en el modelo restringido.

Predicción

- Una predicción es un valor que toma el modelo bajo una nueva observación:

$$y_{n+1} = x_0^t \beta + \varepsilon_{n+1}$$

- Se puede definir entonces el error de predicción como:

$$\hat{\varepsilon}_{n+1} = y_{n+1} - \hat{y}_{n+1} \quad \hat{\varepsilon}_{n+1} \sim N(0, \sigma^2(1 + x_0^t(X^t X)^{-1}x_0))$$

- Y con ello se pueden construir intervalos de confianza para el error de predicción.

Indicaciones Generales

- Para analizar un modelo, se deben tener en cuenta algunas cosas:
 - En lo posible graficar la variable de respuesta en función de cada regresor para detectar posibles anomalías y puntos extremos.
 - Armar la matriz de correlaciones y estudiar la relación lineal entre cada variable independiente y la variable respuesta, para identificar redundancias y los posibles mejores predictores.
 - El análisis de los residuos permite detectar problemas en los supuestos del modelo lineal.
 - Por construcción, en un modelo con constante, la suma de los residuos es siempre igual a 0.