Universidad de Chile
Facultad de Ciencias Físicas y Matemáticas
Departamento de Ingenieria industrial
IN-3242 – Estadistica

Pauta Regresión lineal

introducción y conceptos base

Profesor: Raimundo Undurraga
Auxiliares: Brandon Galarza, Camila Jáuregui Leonardo Meneses, Francisca Monetta, Matias Reyes, Bastian Urzua, Antonia Villegas

Comentes.-

1. Diga si los siguientes modelos son lineales:
   a) $y = \beta_0 + \beta_1 X^2 + \epsilon$
   b) $y = \frac{1}{\beta_0 + \beta_1 X} + \epsilon$
2. La recta resultante de una regresión lineal logra que los residuos sean cero.
3. El MCO asegura que se puede explicar completamente una variable a través de otra, pues el error que se puede cometer es mínimo.

P1.- Demostraciones básicas

Es sabido que la Regresión Lineal Simple supone que el fenómeno en estudio sigue una tendencia lineal de la forma:

$$y = \hat{\beta}_0 + \hat{\beta}_1 X$$

donde "y" es la variable dependiente, "X" la variable independiente y $\hat{\beta}_1, \hat{\beta}_0$ son parámetros a determinar. Considerando que $y_i = \hat{\beta}_0 + \hat{\beta}_1 Xi + \epsilon_i$ (la demanda observada $y_i, i = 1, ..., n$ comete un error ϵ_i), estos parámetros son escogidos de forma que la suma del cuadrado de los errores ϵ_i, provenientes de la observación i sea mínima. Demuestre que bajo lo anterior se tiene que:

$$\hat{\beta}_1 = \frac{\frac{1}{n-1}\sum_{i=1}^{n} X_i y_i - \bar{y}\bar{X}N}{\frac{1}{n-1}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 N} = \frac{Cov(X_i, y_i)}{Var(X_i)} \qquad \text{y} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{X}$$

Con $\bar{X}$ e $\bar{y}$ promedios muestrales de X e y respectivamente.

Respuesta: Minimizamos el cuadrado de la suma de los residuos, dados por $\epsilon_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)$, siguiendo:

$$(\hat{\beta}_0, \hat{\beta}_1) = argmin \sum_{i=1}^{N} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2$$

Trabajaremos un poco la expresión de $\epsilon_i^2 = (y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2$

$$((y_i - \hat{\beta}_1 X_i) - \hat{\beta}_0)^2$$
$$(y_i - \hat{\beta}_1 X_i)^2 - 2\hat{\beta}_0 (y_i - \hat{\beta}_1 X_i) + \hat{\beta}_0{}^2$$
$$y_i^2 - 2y_i \hat{\beta}_1 X_i + \hat{\beta}_1{}^2 X_i^2 - 2\hat{\beta}_0 y_i + 2\hat{\beta}_0 \hat{\beta}_1 X_i + \hat{\beta}_0{}^2$$

Luego con la sumatoria

$$\sum_{i=1}^{N} \epsilon_i^2 = \sum_{i=1}^{N} \left[y_i^2 - 2y_i \hat{\beta}_1 X_i + \hat{\beta}_1{}^2 X_i^2 - 2\hat{\beta}_0 y_i + 2\hat{\beta}_0 \hat{\beta}_1 X_i + \hat{\beta}_0{}^2 \right]$$

$$\sum_{i=1}^{N} y_i^2 - 2\hat{\beta}_1 \sum_{i=1}^{N} X_i y_i + \hat{\beta}_1{}^2 \sum_{i=1}^{N} X_i^2 - 2\hat{\beta}_0 \sum_{i=1}^{N} y_i + 2\hat{\beta}_0 \hat{\beta}_1 \sum_{i=1}^{N} X_i + \hat{\beta}_0{}^2 \cdot N \tag{1}$$

Como es un problema de minimización, tomamos las primeras derivadas para cada beta, obtenidos en la ecuación 1 e imponemos las CPO:

$$\frac{\partial(1)}{\partial \hat{\beta}_1} = -2 \sum_{i=1}^{N} X_i y_i + 2\hat{\beta}_1 \sum_{i=1}^{N} X_i^2 + 2\hat{\beta}_0 \sum_{i=1}^{N} X_i \stackrel{!}{=} 0 \tag{2}$$

$$\frac{\partial(1)}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^{N} y_i + 2\hat{\beta}_1 \sum_{i=1}^{N} X_i + 2\hat{\beta}_0 N \stackrel{!}{=} 0 \tag{3}$$

De la ecuación (3) despejamos $\hat{\beta}_0$

$$2\hat{\beta}_0 N = 2 \sum_{i=1}^{N} y_i - 2\hat{\beta}_1 \sum_{i=1}^{N} X_i$$
$$\hat{\beta}_0 N = \sum_{i=1}^{N} y_i - \hat{\beta}_1 \sum_{i=1}^{N} X_i$$

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{N} y_i}{n} - \hat{\beta}_1 \frac{\sum_{i=1}^{N} X_i}{n}$$

$$\boxed{\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{X}}$$

De la ecuación (2) despejamos $\hat{\beta}_1$, Remplazando $\hat{\beta}_0$

$$-2\sum_{i=1}^{N} X_i y_i + 2\hat{\beta}_1 \sum_{i=1}^{N} X_i^2 + 2\hat{\beta}_0 \sum_{i=1}^{N} X_i \stackrel{!}{=} 0$$

$$-\sum_{i=1}^{N} X_i y_i + \hat{\beta}_1 \sum_{i=1}^{N} X_i^2 + \hat{\beta}_0 \sum_{i=1}^{N} X_i \stackrel{!}{=} 0$$

$$-\sum_{i=1}^{N} X_i y_i + \hat{\beta}_1 \sum_{i=1}^{N} X_i^2 + (\bar{y} - \hat{\beta}_1 \bar{X}) \sum_{i=1}^{N} X_i \stackrel{!}{=} 0$$

$$-\sum_{i=1}^{N} X_i y_i + \hat{\beta}_1 \sum_{i=1}^{N} X_i^2 + \bar{y} \sum_{i=1}^{N} X_i - \hat{\beta}_1 \bar{X} \sum_{i=1}^{N} X_i \stackrel{!}{=} 0$$

$$\hat{\beta}_1 \left(\sum_{i=1}^{N} X_i^2 - \bar{X} \sum_{i=1}^{N} X_i \right) = \sum_{i=1}^{N} X_i y_i - \bar{y} \sum_{i=1}^{N} X_i$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} X_i y_i - \bar{y} \sum_{i=1}^{N} X_i}{\sum_{i=1}^{N} X_i^2 - \bar{X} \sum_{i=1}^{N} X_i}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} X_i y_i - \bar{y} \bar{X} N}{\sum_{i=1}^{N} X_i^2 - \bar{X} \bar{X} N}$$

Aplicamos nikita nipone:

$$\hat{\beta}_1 = \frac{\frac{1}{n-1} \sum_{i=1}^{N} X_i y_i - \bar{y} \bar{X} N}{\frac{1}{n-1} \sum_{i=1}^{N} X_i^2 - \bar{X}^2 N}$$

$$\boxed{\hat{\beta}_1 = \frac{Cov(X_i, y_i)}{Var(X_i)}}$$

P2.- Aplicación

El ministerio de salud desea realizar un análisis de productividad de los médicos que atienden en consultorios. Se desea estudiar como la experiencia de los doctores (medida

en años transcurridos desde que comenzaron a trabajar en los consultorios), X - afecta la productividad, medido en número de pacientes atendidos por hora, y. Para esto, se recopiló información de 200 médicos de distintas especialidades de distintas zonas del país. Para cada médico, se calculó su experiencia (exp) y la productividad durante el último año (prod). La siguiente tabla muestra algunos estadísticos obtenidos para los 200 doctores. La correlación entre las dos variables es 0.64. Asuma que la varianza de los residuos es 1,65.

Variable	Obs	Mean	Std. Dev.	Min	Max
exp(x)	200	6,81	4,27	0,00	14,00
prod(y)	200	2,69	1,39	0,45	10,91

1.- Se desea estimar una regresión $y = \beta_0 + \beta_1 X + \epsilon$. Con los datos proporcionados, estime β_0 y β_1 mediante Mínimos Cuadrados Ordinarios

Respuesta:

Utilizando las fórmulas obtenidas en la pregunta anterior, se obtiene que

$$\hat{\beta}_1 = \frac{Cov(x,y)}{Var(x)} = \frac{\rho_{xy} \cdot \sigma_x \cdot \sigma_y}{\sigma_x^2} = \frac{\rho_{xy} \cdot \sigma_y}{\sigma_x}$$

Reemplazamos los datos

$$\hat{\beta}_1 = \frac{0.64 \cdot 1,39}{4,27} = 0,208$$

De forma análogo para obtener el intercepto

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{X}$$

Reemplazamos los datos

$$\hat{\beta}_0 = 2,69 - 0,208 \cdot 6,81 = 1,274$$

Por lo tanto la regresión lineal con los coeficientes calculado sería:

$$\boxed{y = 1,274 + 0,208 \cdot X}$$

2.-¿Qué se puede decir sobre la significancia de β_1? Realice un Test de Hipótesis.

Queremos testear la influencia que tiene el coeficiente β_1 que acompaña a la variable independiente "X"por sobre la variables dependiente ε". Asumimos que $\beta_1 \sim N\left(\hat{\beta}_1, \frac{S_R^2}{(N-1)\cdot\sigma^2}\right)$, en donde S_R^2 es la Varianza de los errores y σ^2 es la Varianza de X. Por lo tanto realizamos el test de hipótesis siguiendo el "Recetario" realizado en el Aux #04.

Paso 1.-

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Paso 2.- (No se da como dato, pero se debe establecer, intentar probar con otro α)

$$\alpha = 0.05$$

Paso 3.-

$$z_0 = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{S_R^2}{(N-1)\cdot\sigma^2}}} = \frac{\hat{\beta}_1 - 0}{\sqrt{\frac{S_R^2}{(N-1)\cdot\sigma^2}}}$$

Paso 4.- Haremos la construcción de un intervalo de confianza para evidenciar otras formas de testeo. Usando TCL (valor crítico de Normalidad) ya que tenemos 200 datos.

$$\left[\hat{\beta}_1 \pm Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{S_R^2}{(N-1)\cdot\sigma^2}}\right]$$

$$\left[0,208 \pm 1,96 \cdot \sqrt{\frac{1,65}{(200-1)\cdot 4,27^2}}\right]$$

$$[0,166; 0,249]$$

Paso 5.- Como en la hipótesis nula $\beta_1 = 0 \notin [0,166; 0,249]$. Se rechaza H_0. Entonces, existe evidencia estadística para rechazar la H_0, con una significancia de α= 0.05 (nivel de confianza del 95 %).

P3.- Interpretación de resultados

Se realizó una regresión lineal en R, para estudiar la incidencia del PIB per cápita en el índice de obesidad mundial, por cada país. Donde *obesidad_pais* es la variable dependiente y *pib_per_capita* la independiente.

1.- ¿Qué otras variables podrían ser útiles para mejorar la estimación de la variable obesidad, además de pib per cápita? Proponga otros 2 regresores y justifique.

Respuesta: Ademas del PIB hay muchas otras variables que podrían explicar obesidad. Pero como ejemplo podrían tenerse variables como la cantidad de horas promedio que se hace deporte, y tambien otro regresor puede ser el costo de la comida saudable en un pais.

2.- ¿Qué efecto se espera de pib per cápita y las otras dos variables propuestas, sobre la variable dependiente?

Respuesta: Se esperaría que a mayor PIB haya mas indice de obesidad, por lo cual se espera un efecto positivo, es decir $\beta_{PIB} > 0$.

Para las variables propuestas, se espera que a mayor horas de deporte haya menor obesidad, esperando un efecto negativo entonces de este regresor sobre la dependiente, con lo cual $\beta_{ejercicio} < 0$ y finalmente, para la variable costo de alimentación saludable, se espera que mientras mayor sea este costo, menos acceso a comida saludable tenga la poblacion y por tanto es esperable un mayor indice de obesidad, así $\beta_{comidasaludable} > 0$

Luego se obtienen los resultados de la regresión, siendo los siguientes:

Regresión p3

	Obesidad por pais
pib_per_capita	0.0002***
	(0.00004)
Constant	17.1735***
	(0.9750)
N	185
R^2	0.0912
Adjusted R^2	0.0863
Residual Std. Error	10.7483 (df = 183)
F Statistic	18.3684*** (df = 1; 183)
Notes:	***Significant at the 1 percent level.
	**Significant at the 5 percent level.
	*Significant at the 10 percent level.

Figura 1: Resultado de la regresión

3.- ¿Cuáles son los valores de beta 1 y del intercepto?¿Cuál es la desviación estándar del valor de beta 1?

Respuesta: El valor de beta 1 corresponde al valor junto al nombre de la variable, 0.0002. Mientras que el valor de la desviacion estandar asociado a este coeficiente es el parentesis justo debajo: (0.00004).

4.- ¿Se obtuvo el efecto esperado que se mencionó en 2? ¿Es significativo el resultado obtenido?

Respuesta: En efecto, dado que beta 1 es un valor positivo, quiere decir que a mayor PIB, se tendra un mayor valor de la variable dependiente obesidad. El valor obtenido es significativo, pese a que es pequeño, y esto es debido a que su pvalor es extremadamente pequeño incluso al nivel 1. Esto es un adelanto del proximo auxiliar pero en simples palabras significa que el coeficiente es suficientemente distinto de cero para poder establecer que si hay un efecto del PIB sobre obesidad.

5.- Con la información de la tabla responda: ¿basta con el regresor pib per cápita elegido para explicar el índice de obesidad?

Respuesta:Observando el valor de R cuadrado, que corresponde a 0.09, se puede decir que este modelo que intenta explicar el comportamiento de la variable obesidad mediante el PIB sólo explica el 9 % de la información. En base a ello, se puede decir que falta mucha información por ser explicada, y por tanto este regresor no es suficiente para explicar obesidad.

Resumen

Regresion lineal simple: Busca aproximar el comportamiento de una variable de forma lineal a través de otra variable, de la forma más exacta posible, es decir, minimizando los residuo. Se escribe de la forma:

$$y_i = \beta_0 + \beta_{1i} X_i + \epsilon_i$$

Donde y es la variable a explicar, o **variable dependiente**, X_i es la variable explicativa o **regresor** y ϵ es el **error**, que corresponde a un factor aleatorio no observable.

Mínimos cuadrados ordinarios: (MCO u OLS) Minimizar el residuo significa hacer que sea lo más pequeño posible. Si se hiciera directamente, sería algo del estilo: $\sum_{i=1}^{n} \epsilon_i \to -\infty$, lo cual resultaría una estimación extremadamente distinta de y, por lo que no es correcto minimizarlo de esta manera. De aquí nace el concepto de **mínimos cuadrados ordinarios (MCO)**, con lo cual se busca minimizar el la diferencia entre la estimación y la observación (residuo) al cuadrado, de forma que : $\sum_{i=1}^{n} \epsilon_i^2 \to 0$.

Reemplazando lo anterior de la fórmula de regresión lineal entonces se tendrá:

$$\begin{aligned} 0 &= \sum_{i=1}^{n} \epsilon_i^2 \\ 0 &= \sum_{i=1}^{n} (y_i - \beta_0 - \beta_{1i} X_i)^2 \end{aligned} \tag{4}$$

Resolviendo la ecuación anterior, se encuentran los óptimos $\hat{\beta}_0$ y $\hat{\beta}_1$ tales que finalmente el resultado de la regresión lineal simple será lo que se conoce como **valores predichos**:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

Visualmente lo que busca la regresión lineal es la recta que relacione ambas variables tal que minimice la distancia de cada observación a sí misma de la **mejor** forma:

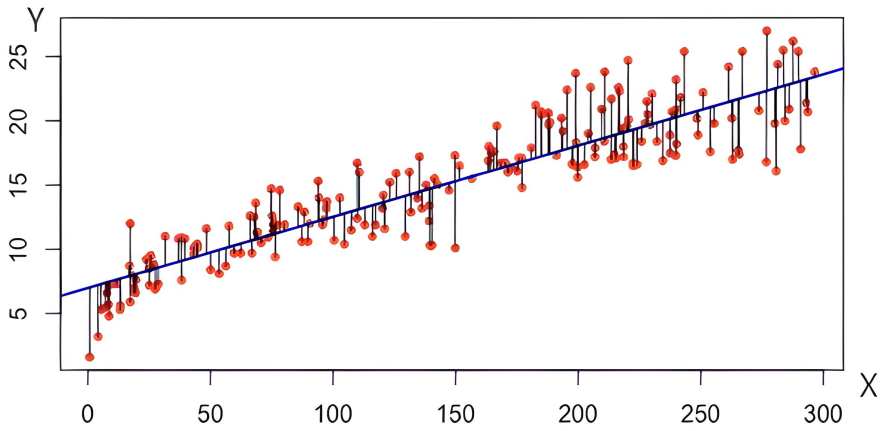

Figura 2: Regresión lineal en azul, observaciones en rojo

Importante destacar que el término residuo también se suele llamar más comunmente error, pero no confundir con error estándar pues no tienen nada que ver.

Resolucion del modelo con CPO: mediante el método de MCO se resuleve la ecuación (1), para lo cual se obtienen las ecuaciones:

- Para la pendiente: $\hat{\beta}_1 = \frac{Cov(X_i, y_i)}{Var(X_i)}$
- Para el intercepto: $\hat{\beta}_0 = \bar{y}_i - \hat{\beta}_1 X_i$

De lo anterior, $\hat{\beta}_1$ también puede escribirse como $\hat{\beta}_1 = \beta_1 + \frac{Cov(X, \varepsilon)}{Var(X)}$, entonces es directo

obtener que si $\frac{Cov(X,\varepsilon)}{Var(X)} = 0$, entonces $\hat{\beta}_1$ es insesgado. En específico:

- Si Cov(X, ε) = 0, entonces $\hat{\beta}_1$ es insesgado con respecto a β_1.
- Si $Cov(X, \varepsilon) \neq 0$, entonces $\hat{\beta}_1$ es sesgado con respecto a β_1.

Resultados del MCO: se tienen los siguientes elementos como resultado de todo lo anterior:

- Valores predichos: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
- Residuo: $e_i = y_i - \hat{y}_i$
- Suma de cuadrados:
  - sum of squares total: $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$
  - sum of squared regression: $SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
  - sum of square errors: $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

Propiedades del MCO

- $\frac{1}{n}\sum_{i=1}^{n}\epsilon_i = 0$
- $Cov(\epsilon_i, X_i) = 0$, es decir: $\sum_{i=1}^{n} X_i\epsilon_i = 0$
- SST = SSR + SSE

R cuadrado: este coeficiente se obtiene de las regresiones lineales, y se calcula como: $R^2 = \frac{SSR}{SST} \Leftrightarrow R^2 = 1 - \frac{SSE}{SST}$

Es un valor entre 0 y 1 y representa el porcentaje de los datos correspondientes a y, que son explicados por el modelo.