Facultad de Ciencias
UNAM

Estadística II

Tarea 1B

Regresión Lineal Simple

Enríquez Hernández Leobardo
Huitrón Zambrano Victor Manuel
Suárez López David

28 de marzo de 2024

Índice

1. Regresión a través del origen.

2. Regresión lineal simple.

Considere el modelo de regresión $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, donde $E(\epsilon_i) = 0$, $V(\epsilon_i) = \sigma^2$ y $Cov(\epsilon_i, \epsilon_j) = 0$, $\forall i \neq j$; $i, j = 1, ..., n$.

Calcular $V(e_i)$, donde $e_i = y_i - \hat{y}_i$ y $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, con $\hat{\beta}_0$ y $\hat{\beta}_1$ los estimadores de los parámetros del modelo.

Hint: Se puede usar que $V(A - B) = V(A) + V(B) - 2Cov(A, B)$ y que $\hat{y}_i$ se puede escribir como una combinación lineal de las $y_{i's}$.

SOLUCIÓN

Como $V(y_i) = V(\beta_0 + \beta_1 x_i + \epsilon_i) = V(\epsilon_i) = \sigma^2$ por ser β_0, β_1 y x_i constantes.

Como $V(\hat{y}_i) = V(\hat{\beta}_0 + \hat{\beta}_1 x_i) = V(\beta_0) + V(\beta_1 x_i) + 2Cov(\beta_0, \beta_1 x_i) = V(\hat{\beta}_0) + x_i^2 V(\hat{\beta}_1) + 2x_i Cov(\hat{\beta}_0, \hat{\beta}_1) = \sigma^2(\frac{1}{n} + \frac{\bar{X}^2}{SSx}) + x_i^2(\frac{\sigma^2}{SSx}) + 2x_i(\frac{-\bar{X}\sigma^2}{SSx}) = \sigma^2(\frac{SSx + n\bar{X}^2}{nSSx} + \frac{x_i^2}{SSx} - \frac{2x_i\bar{X}}{SSx}) = \sigma^2(\frac{1}{n} + \frac{(x_i - \bar{X})^2}{SSx})$, con $SS_x = \sum_{i=1}^n (x_i - \bar{X})^2$.

Como $Cov(y_i, \hat{y}_i) = Cov(y_i, \bar{Y} + \hat{\beta}_1 x_i - \hat{\beta}_1 \bar{X}) = Cov(y_i, \bar{Y}) + Cov(y_i, \hat{\beta}_1 x_i) + Cov(y_i, -\hat{\beta}_1 \bar{X}) = Cov(y_i, \hat{\beta}_0 + \hat{\beta}_1 \bar{X}) + x_i Cov(y_i, \hat{\beta}_1) - \bar{X} Cov(y_i, \hat{\beta}_1) = Cov(y_i, \hat{\beta}_0) + \bar{X} Cov(y_i, \hat{\beta}_1) + x_i Cov(y_i, \hat{\beta}_1) - \bar{X} Cov(y_i, \hat{\beta}_1) = Cov(y_i, \hat{\beta}_0) + x_i Cov(y_i, \hat{\beta}_1) = (\frac{1}{n} - \frac{\bar{X}(x_i - \bar{X})}{SSx})\sigma^2 + x_i(\frac{x_i - \bar{X}}{SSx})\sigma^2 = \sigma^2(\frac{1}{n} - \frac{\bar{X}(x_i - \bar{X})}{SSx} + x_i(\frac{x_i - \bar{X}}{SSx})$.

Entonces:

$$V(e_i) = V(y_i - \hat{y}_i) = V(y_i) + V(\hat{y}_i) - 2Cov(y_i, \hat{y}_i) = \sigma^2 + \sigma^2(\frac{1}{n} + \frac{(x_i - \bar{X})^2}{SS_x}) - 2\sigma^2(\frac{1}{n} - \frac{\bar{X}(x_i - \bar{X})}{SSx} + \frac{x_i(x_i - \bar{X})}{SSx})$$

$$= \sigma^2 + \frac{\sigma^2}{n} + \frac{\sigma^2(x_i - \bar{X})^2}{SSx} - \frac{-2\sigma^2}{n} + \frac{2\sigma^2\bar{X}(x_i - \bar{X})}{SSx} - \frac{2\sigma^2 x_i(x_i - \bar{X})}{SSx} = \sigma^2 + \frac{\sigma^2}{n} + (\frac{-\sigma^2 x_i^2 - \sigma^2\bar{X}^2 + 2\sigma^2\bar{X}x_i}{SSx})$$

$$= \sigma^2 + \frac{\sigma^2}{n} - \frac{\sigma^2}{SSx}(x_i - \bar{X})^2 = \sigma^2 + \frac{\sigma^2}{n} - \frac{\sigma^2}{n} = \sigma^2$$

Se usaron los siguientes resultados:

$\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}$

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = (\bar{Y} - \bar{X}\hat{\beta}_1) + \hat{\beta}_1 x_i = \bar{Y} + \hat{\beta}_1 x_i - \hat{\beta}_1 \bar{X}$

$V(\hat{\beta}_0) = Cov(\hat{\beta}_0, \hat{\beta}_0) = Cov(\sum_{i=1}^n k_{i_0} y_i, \sum_{j=1}^n k_{j_0} y_j) = \sigma^2 \sum_{i=1}^n k_{i_0}^2 = \sigma^2 \sum_{i=1}^n (\frac{1}{n} - \frac{\bar{X}(x_i - \bar{X})}{SSx})^2 = \sigma^2(\frac{1}{n} + \frac{\bar{X}^2}{SSx})$,

$V(\hat{\beta}_1) = Cov(\hat{\beta}_1, \hat{\beta}_1) = Cov(\sum_{i=1}^n k_{i_1} y_i, \sum_{j=1}^n k_{j_1} y_j) = \sigma^2 \sum_{i=1}^n k_{i_1}^2 = \sigma^2 \sum_{i=1}^n (\frac{x_i - \bar{X}}{SSx})^2 = \frac{\sigma^2}{(SSx)^2} \sum_{i=1}^n (x_i - \bar{X})^2 = \frac{\sigma^2}{SSx}$,

$Cov(\hat{\beta}_0, \hat{\beta}_0) = Cov(\sum_{i=1}^n k_{i_0} y_i, \sum_{j=1}^n k_{j_1} y_j) = \sigma^2 \sum_{i=1}^n k_{i_0} k_{i_1} = \sigma^2 \sum_{i=1}^n (\frac{1}{n} - \frac{\bar{X}(x_i - \bar{X})}{SSx})(\frac{x_i - \bar{X}}{SSx}) = -\frac{\bar{X}\sigma^2}{SSx}$,

$Cov(y_i, \hat{\beta}_0) = Cov(y_i, \sum_{i=1}^n k_{i_0} y_i) = k_{i_0} Cov(y_i, y_i) = k_{i_0} V(y_i) = k_{i_0} \sigma^2 = (\frac{1}{n} - \frac{\bar{X}(x_i - \bar{X})}{SSx})\sigma^2$,

$Cov(y_i, \hat{\beta}_1) = Cov(y_i, \sum_{i=1}^n k_{i_1} y_i) = k_{i_1} Cov(y_i, y_i) = k_{i_1} V(y_i) = k_{i_1} \sigma^2 = (\frac{x_i - \bar{X}}{SSx})\sigma^2$,

$\frac{SSx}{(x_i - \bar{X})^2} = \frac{\sum_{i=1}^n (x_i - \bar{X})^2}{(x_i - \bar{X})^2} = \sum_{i=1}^n 1 = n$

3. Expresión alternativa para R^2

4. Problema Anova. Equivalencia con la estimación considerando dos poblaciones normales.

5. Problema ANOVA. Medicamentos.

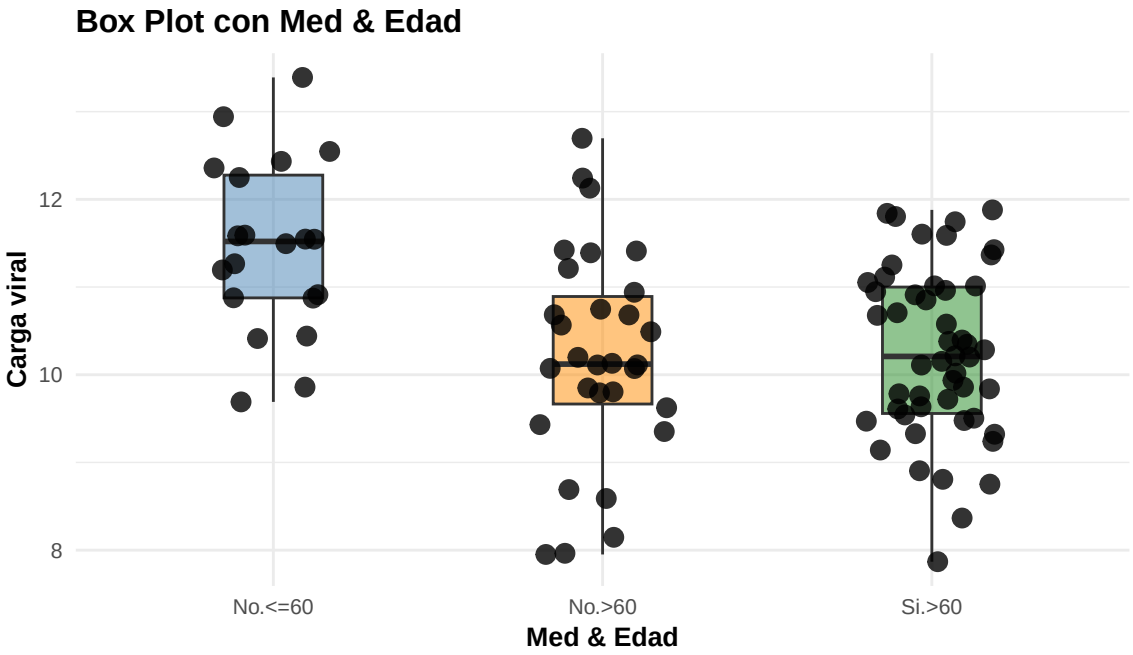
I. Análisis descriptivo y/o visualización de datos.

En la base de datos `Ejercicio5B` se tiene información del índice de carga viral (Y), si se aplicó o no el medicamento contra Covid (Med) y si los pacientes son mayores o menores (o iguales) a 60 años (Edad). Tenemos un total de 100 pacientes, de los cuales 80 tienen más de 60 años y a la mitad de todos los pacientes se le aplicó el medicamento (tratados).

En el siguiente Cuadro se muestra la estadística descriptiva del dato numérico, que es el índice de carga viral, es posible observar los pacientes presentaron un mínimo de 7.8681037 y un maximo de 13.3899626, con una media de 10.4800674. La desviación estándar de 1.1477774 es pequeña, por lo que parece que los datos no son tan dispersos.

Statistic	N	Mean	St. Dev.	Min	Max
datos$Y	100	10.480	1.148	7.868	13.390

A continuación podemos observar en la gráfica de caja y bigotes que a los pacientes menores de 60 años a quienes no se les aplicó la vacuna (no fueron tratadas) tienen una mayor carga viral, sin embargo no disponemos de datos para personas menores a 60 años a quienes se les aplicó el medicamento, para una mejor comparación. Por otro lado, para mayores de 60 años, parece no haber una diferencia clara entre los pacientes a los que se le aplicó la vacuna y a las personas a las que no se le aplicó, pues ambos grupos tienen cargas virales menores pero muy parecidas.



II. Planteamiento del modelo y supuestos.

Para ver si la menor carga viral está asociada con la aplicación del medicamento planteamos un modelo de regresión donde la variable dependiente es la carga viral y_i y la variable independiente x_i se puede ver como categórica, donde $x_i = 1$ si el paciente es tratado y $x_i = 0$ si no.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Suponemos que el modelo presenta linealidad, homocedasticidad en los errores y no autocorrelación en los errores.

III. Prueba de hipótesis.

En el siguinte Cuadro se muestra el modelo ajustado por mínimos cuadrados ordinarios. El intercepto $\beta_0 =$ 10.7141904, el término asociado a la variable categórica de si se aplicó el medicamento es β_1 = -0.4682459, y ambos son estadísticamente significativos de manera individual con la prueba $t-student$, es decir, para ambos casos se rechaza la hipótesis nula de que $\beta_0 = 0$ y $\beta_1 = 0$ con el 5 % de nivel de significancia, respectivamente. La prueba de hipótesis global también rechaza la hipótesis nula de que $\beta_i = 0, \forall i = 1, .., n$, es decir, de acuerdo con la prueba F, al menos una de los coeficientes es distinto de cero. Cabe notar que como estamos en el caso de una solo variable explicativa, esta prueba coincide con la asociada a β_1.

Como puede observarse, el nivel de referencia es $MedNo$, es decir, que no se haya tratado o no se haya aplicado el medicamento. Entonces, de acuerdo con el modelo, el tratamiento disminuye la carga viral en -0.4682459 unidades.

IV. Consideración de la variable Edad.

Como se mencionó anteriormente, en la base de datos tenemos 20 observaciones de personas no tratadas menores de 60 años y no tenemos personas de ese grupo de edades que sean tratadas por el medicamento, lo que podría sesgar los resultados. Aunado a esto, tenemos que para el grupo de edades mayores de 60 años, no se puede notar gráficamente una disferencia clara entre tratados y no tratados. Por lo tanto, consideramos que los rsultados anteriores no son contundentes, por lo que convendría controlar por la variable de Edad.

	Dependent variable:
	Y
MedSi	-0.468^{**}
	s.e.(0.226)
	t-value: -2.074
	Pr(>|t|): 0.0407
Constant	10.714^{***}
	s.e.(0.160)
	t-value: 67.097
	Pr(>|t|): <2e-16
Observations	100
R^2	0.042
Adjusted R^2	0.032
Residual Std. Error	1.129 (df = 98)
F Statistic	4.299^{**} (df = 1; 98); p-value: 0.04075
Note:	*p<0.1; **p<0.05; ***p<0.01

V. Planteamiento del nuevo modelo y pruebas de hipótesis.

En este caso planteamos el mismo modelo, pero como no tenemos individuos tratados y no tratados menores de 60 años en la base de datos (solamente tenemos a los no tratados), tomaremos en cuenta a los grupos homogéneos de tratados y no tratados mayores o iguales a 60 años, por lo que nuestra base de datos se reduce de 100 a 80 observaciones.

A continuación se muestra el Cuadro correspondiente a la regresión lineal para los pacientes tratados y no tratados mayores a 60 años. La prueba global F muestra que no es posible rechazar la hipótesis nula de que los coeficientes asociados a las variables explicativas son cero, pues el p-value asociado es grande, incluso mayor a 0.1.

Al parecer, las conclusiones obtenidas al tomar toda la muestra estaban sesgadas por el grupo de edad, pues cuando quitamos a los no tratados menores de 60 años que tenían una carga viral bastante importante, los resultados cambiaron.

6. Uso del modelo de regresión lineal simple.

7. Regresión lineal simple con datos de "performance".

	Dependent variable:
	Y
MedSi	0.029
	s.e.(0.245)
	t-value: 0.119
	Pr(>|t|): 0.906
Constant	10.217***
	s.e.(0.194)
	t-value: 52.704
	Pr(>|t|): <2e-16
Observations	80
R^2	0.0002
Adjusted R^2	−0.013
Residual Std. Error	1.062 (df = 78)
F Statistic	0.014 (df = 1; 78); p-value: 0.9056
Note:	*p<0.1; **p<0.05; ***p<0.01