UNIDAD 7. Regresión Lineal Simple

(1) Procedimiento para realizar una regresión lineal simple

1. Identificar las variables: Realizar un diagrama de dispersión para mostrar cómo la variable dependiente (Y) y la variable independiente (X) se relacionan entre sí.

Statgraphics: *Gráficos / Dos dimensiones / Gráfico X-Y.*

2. Estimar β_0 y β_1 a partir de b_0 y b_1: Se aplica el método de los mínimos cuadrados para obtener la recta ajustada preliminar ($\hat{y} = b_0 + b_1 \cdot x$).

Statgraphics: *Relacionar / Modelos Reg. Simple / Regresión Simple. Seleccionar el modelo "lineal" y en la primera corrida marcar la casilla "incluir constante". Seleccionar las casillas de tablas "Resumen del Análisis", "Pronósticos", "Comparación Modelos Alternativos", "Residuos Atípicos" y de gráficos "Gráfico del Modelo Ajustado" y "Residuos vs Predichos".*

3. Determinar si β_0 y β_1 son significativos: Se determina si $\beta_0 \neq 0$ y $\beta_1 \neq 0$ a través de los siguientes métodos.

3.1. ANOVA: Consiste en realizar una prueba de hipótesis para β_1 utilizando la D. Fisher con $H_0: \beta_1 = 0$ y $H_1: \beta_1 \neq 0$. Se rechaza H_0 si $f_0 > f_{1,n-2,\alpha}$ o si $Valor\ P < \alpha$, $Valor\ P = P(F_{1,n-2} > f_0)$. Si la hipótesis nula no es rechazada, el modelo no es válido para predecir.

Fuente	Suma de Cuadrados	Gl	Cuadrado Medio	Razón-F	Valor-P
Modelo	SSR	1	$MS_R = \frac{SSR}{1}$	$f_0 = \frac{SSR/1}{SSE/(n-2)} = \frac{SSR}{s_e^2}$	$P(F_{1,n-2} > f_0)$
Residuo	SSE	$n-2$	$MS_E = s_e^2 = \frac{SSE}{n-2}$		
Total (Corr.)	SST	$n-1$			

Caso Especial: *Cuando β_0 no es significativo (es igual a cero), el ANOVA presenta modificaciones. Se rechaza H_0 si $f_0 > f_{1,n-1,\alpha}$ o si $Valor\ P < \alpha$, $Valor\ P = P(F_{1,n-1} > f_0)$.*

Fuente	Suma de Cuadrados	Gl	Cuadrado Medio	Razón-F	Valor-P
Modelo	SSR	1	$MS_R = \frac{SSR}{1}$	$f_0 = \frac{SSR/1}{SSE/(n-1)} = \frac{SSR}{s_e^2}$	$P(F_{1,n-1} > f_0)$
Residuo	SSE	$n-1$	$MS_E = s_e^2 = \frac{SSE}{n-1}$		
Total (Corr.)	SST	n			

Statgraphics: *Revisar la tabla "Análisis de Varianza" de la ventana "Regresión simple". Por defecto el software utiliza un nivel de significancia de 5%, si se desea modificar hacer clic derecho a "Opciones tabulares" y modificar en "Nivel alpha de valor P" colocando el valor deseado en porcentaje.*

3.2. Pruebas de hipótesis utilizando la D. t de Student: Se realizan para β_0 ($H_0: \beta_0 = B_0 = 0$ y $H_1: \beta_0 \neq B_0 \neq 0$) y β_1 ($H_0: \beta_1 = B_1 = 0$ y $H_1: \beta_1 \neq B_1 \neq 0$), se calculan los grados de libertad ($n-2$) y el estadístico de prueba (t_0).

- Pruebas de hipótesis para β_0:

$$t_0 = \frac{b_o - B_0}{\sqrt{s_e^2 \cdot \left(\frac{1}{n} + \frac{\bar{x}^2}{SS_x}\right)}}$$

- Pruebas de hipótesis para β_1:

$$t_0 = \frac{b_1 - B_1}{\sqrt{\frac{s_e^2}{SS_x}}}$$

Si la hipótesis nula $H_0: \beta_0 = 0$ no es rechazada, el modelo no tiene intercepto. Si la hipótesis nula $H_0: \beta_1 = 0$ no es rechazada, el modelo no es válido para predecir.

Tip: *H_0 y H_1 pueden plantearse también para otro tipo de pruebas. Por ejemplo para β_0 $H_1: \beta_0 \neq B_0$; $H_1: \beta_0 > B_0$; $H_1: \beta_0 < B_0$. Y por ejemplo para β_1 $H_1: \beta_1 \neq B_1$; $H_1: \beta_1 > B_1$; $H_1: \beta_1 < B_1$.*

Caso Especial: *Cuando β_0 no es significativo (es igual a cero), los grados de libertad para la D. t de Student en las pruebas de hipótesis para β_1 se calculan como $n-1$.*

Statgraphics: *Revisar la tabla "Coeficientes" de la ventana "Regresión simple".*

Parámetro	Mínimos Cuadrados Estimado	Estándar Error	Estadístico T	Valor-P
Intercepto	b_0	$\sqrt{s_e^2 \cdot \left(\frac{1}{n} + \frac{\bar{x}^2}{SS_x}\right)}$	t_0 para β_0	$Valor\ P$ para β_0
Pendiente	b_1	$\sqrt{\frac{s_e^2}{SS_x}}$	t_0 para β_1	$Valor\ P$ para β_1

3.3. Intervalos de confianza: Para β_0 y β_1 con un nivel de confianza de $(1-\alpha) \cdot 100\%$.

$$b_0 - t_{n-2,\alpha/2} \cdot \sqrt{s_e^2 \cdot \left(\frac{1}{n} + \frac{\bar{x}^2}{SS_x}\right)} \leq \beta_0 \leq b_0 + t_{n-2,\alpha/2} \cdot \sqrt{s_e^2 \cdot \left(\frac{1}{n} + \frac{\bar{x}^2}{SS_x}\right)}$$

$$b_1 - t_{n-2,\alpha/2} \cdot \sqrt{\frac{s_e^2}{SS_x}} \leq \beta_1 \leq b_1 + t_{n-2,\alpha/2} \cdot \sqrt{\frac{s_e^2}{SS_x}}$$

Caso Especial: *Cuando β_0 no es significativo (es igual a cero), el intervalo de confianza para β_1 presenta modificaciones.*

$$b_1 - t_{n-1,\alpha/2} \cdot \sqrt{\frac{s_e^2}{SS_x}} \leq \beta_1 \leq b_1 + t_{n-1,\alpha/2} \cdot \sqrt{\frac{s_e^2}{SS_x}}$$

Statgraphics: *No muestra los I.C. pero se pueden calcular manualmente a partir de la información que se muestra en la tabla "Coeficientes" de la ventana "Regresión simple".*

4. Escribir la recta ajustada definitiva: Según los resultados del paso 3 se originan tres casos.

- Si $\beta_1 = 0$ no hay relación entre X y Y (no hay recta).
- Si $\beta_0 = 0$ y $\beta_1 \neq 0$ la recta no tiene intercepto ($\hat{y} = b_1 \cdot x$).
- Si $\beta_0 \neq 0$ y $\beta_1 \neq 0$ la recta sería igual a la preliminar ($\hat{y} = b_0 + b_1 \cdot x$).

Statgraphics: *Si se obtiene que la recta no tiene intercepto, entonces se debe repetir el paso 2 y realizar nuevamente la regresión lineal simple indicando que el intercepto es cero (se desmarca la casilla "incluir constante"). Esto modificará el valor de la pendiente (b_1) y por eso se deben repetir los pasos 3.1, 3.2 y 3.3.*

UNIDAD 7. Regresión Lineal Simple

(2) Procedimiento para realizar una regresión lineal simple (continuación)

5. Verificar qué tan buena es la recta: Se calcula el coeficiente de correlación y el coeficiente de determinación.

5.1. Coeficiente de correlación: Mide el grado de asociación lineal entre las dos variables. Se interpreta según su valor entre -1 y 1.

- Si $\rho = 0$ no existe regresión lineal.
- Si $\rho \pm 1$ existe relación lineal perfecta entre las variables.
- Valores de ρ cercanos a la unidad implican una buena correlación entre las variables.
- Valores de ρ cercanos a cero implican poca o ninguna correlación entre las variables.

5.2. Coeficiente de determinación: Determina la calidad del modelo para replicar los resultados, y a su vez representa la proporción de variación de los resultados que puede explicarse por el modelo. Se interpreta según su valor entre 0 y 1.

- Si R^2 va de 0.9 a 1.0 el modelo es excelente.
- Si R^2 va de 0.8 a 0.9 el modelo es muy bueno.
- Si R^2 va de 0.6 a 0.8 el modelo es bueno.
- Si R^2 va de 0.5 a 0.6 el modelo es regular.
- Si R^2 es menor de 0.5 el modelo es malo.

Statgraphics: *Los valores de los coeficientes de interés se pueden leer en la ventana de "Regresión simple". Por ejemplo…*

Coeficiente de Correlación = 0.959903 $= \rho$
~~R-cuadrada = 92.1413 porciento~~
R-cuadrado (ajustado para g.l.) = 91.9776 porciento $= R^2$
Error estándar del est. = 5.851 $= s_e = \sqrt{SSE/(n-2)}$
~~Error absoluto medio = 3.75083~~
Estadístico Durbin-Watson = 2.02763 (P=0.4797) $= D_w$
~~Autocorrelación de residuos en retraso 1 = -0.0165066~~

6. Verificar supuestos sobre los errores: Se verifica la normalidad, independencia y homocedasticidad de los errores. Si uno o más supuestos no se pueden verificar entonces la recta no sería válida.

6.1. Normalidad: Se verifica a través de una prueba de bondad de ajuste Chi-cuadrado con H_0: Los errores distribuyen normal con $\mu = 0$ y $\sigma^2 = s_e^2$ y H_1: Los errores no distribuyen normal con $\mu = 0$ y $\sigma^2 = s_e^2$.

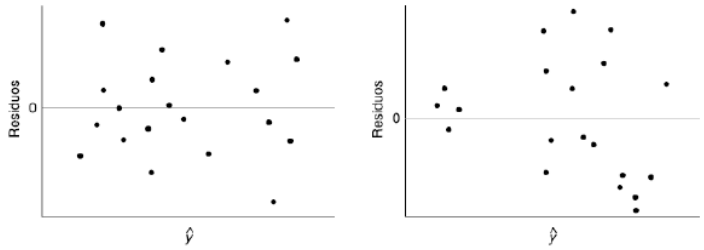
Statgraphics: *Análisis / Resultados marcar la casilla "Residuos". Luego Describir / Ajuste Distribución / Ajuste Datos No Censurados seleccionar los datos de nombre "Residuos". En las opciones de ajuste de distribuciones seleccionar "Normal". Seleccionar las casillas de tablas "Resumen del Análisis" y "Pruebas de Bondad-de-Ajuste". Los resultados aparecen en el menú "Ajuste de Distribuciones". Seleccionar la venta "Pruebas de Bondad-de-Ajuste" y una vez allí hacer clic derecho para "Opciones de ventana". En las opciones de "Pruebas de Bondad de Ajuste" desmarcar las casillas "Kolmogórov-Smirnov" y "Usar clases equiprobables", y marcar la casilla "Chi-cuadrada". Verificar si se recalcularon los resultados para la ventana "Pruebas de Bondad-de-Ajuste" utilizando una Distribución Chi-cuadrado. Copiar la tabla agrupada para realizar la prueba de bondad de ajuste con el procedimiento visto en el curso.*

6.2. Independencia: Se verifica a través de una prueba Durbin-Watson con H_0: Los errores son independientes y H_1: Los errores no son independientes. El estadístico Durbin-Watson se interpreta según su valor entre 0 y 4, y si este valor es cercano a 2 los residuos se asumen independientes. Para un nivel de significancia de 5% y $k^* = 1$ se buscan los valores de d_L y d_u para el n que corresponda en la Tabla Durbin-Watson.

- Si $0 \leq D_w \leq d_L$ la correlación es negativa
- Si $d_L \leq D_w \leq du$ es inconcluso
- Si $d_u \leq D_w \leq 4 - d_u$ los residuos son independientes
- Si $4 - d_u \leq D_w \leq 4 - d_L$ es inconcluso
- Si $4 - d_L \leq D_w \leq 4$ la correlación es positiva

Statgraphics: *El estadístico Durbin-Watson se puede leer en la ventana de "Regresión simple".*

6.3. Homocedasticidad: Indica la homogeneidad de la varianza, para verificarlo se realiza una gráfica de los errores contra las variables X y/o Y con H_0: Los residuos son homocedásticos y H_1: Los residuos son heterocedásticos. Si las fluctuaciones de los residuales parecen ser aleatorias alrededor de cero y no muestran ningún patrón, entonces, se dice que los residuos son homocedásticos.



Statgraphics: *Revisar la ventana "Gráfico de residuos" del menú "Regresión Simple".*

7. Inferencias para Y: Intervalo de confianza de $(1-\alpha)\cdot 100\%$ para la respuesta media $E(y|x_0)$ e Intervalos de predicción de $(1-\alpha)\cdot 100\%$ de confianza para y dado $x = x_0$.

$$\hat{y} - t_{n-2,\alpha/2}\cdot s_e\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{SS_x}} \leq E(y|x_0) \leq \hat{y} + t_{n-2,\alpha/2}\cdot s_e\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{SS_x}}$$

$$\hat{y} - t_{n-2,\alpha/2}\cdot s_e\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{SS_x}} \leq y|x_0 \leq \hat{y} + t_{n-2,\alpha/2}\cdot s_e\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{SS_x}}$$

Caso Especial: *Cuando β_0 no es significativo (es igual a cero), los intervalos de confianza y predicción presentan modificaciones.*

$$\hat{y} - t_{n-1,\alpha/2}\cdot s_e\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{SS_x}} \leq E(y|x_0) \leq \hat{y} + t_{n-1,\alpha/2}\cdot s_e\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{SS_x}}$$

$$\hat{y} - t_{n-1,\alpha/2}\cdot s_e\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{SS_x}} \leq y|x_0 \leq \hat{y} + t_{n-1,\alpha/2}\cdot s_e\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{SS_x}}$$

Statgraphics: *Revisar la ventana "Valores predichos" del menú "Regresión Simple". Para cambiar el tipo de intervalo (unilateral o bilateral) y los valores de la variable independiente (x) hacer clic derecho y seleccionar "Opciones de ventana".*