UVG UNIVERSIDAD DEL VALLE DE GUATEMALA

Regresión lineal: Diagnóstico y Pruebas de Hipótesis

Alan Reyes-Figueroa

Introducción a la Ciencia de Datos (Aula 33) 17.mayo.2021

Métodos de Diagnóstico

¿Qué podemos hacer para verificar los supuestos?

1. Verificar normalidad de los residuos $\widehat{\varepsilon}_i \sim \mathcal{N}(0, \sigma^2)$.
2. Verificar homoscedasticidad (varianza σ^2 constante).
3. Verificar que no haya autocorrelación de los residuos $\widehat{\varepsilon}_i$.
4. Verificar no colinealidad de las X_i.
5. Identificar influencia de las observaciones (*outliers*).

Para cada una de estas tenemos varias alternativas:

- Pruebas o herramientas gráficas,
- Pruebas de hipótesis.

Métodos de Diagnóstico

OLS Regression Results

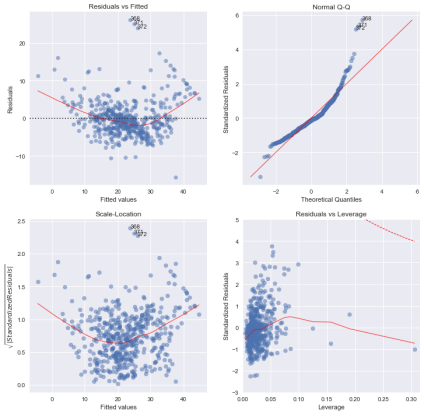
Dep. Variable:	y	R-squared:	0.988
Model:	OLS	Adj. R-squared:	0.988
Method:	Least Squares	F-statistic:	3959.
Date:	Thu, 13 May 2021	Prob (F-statistic):	1.05e-93
Time:	15:31:54	Log-Likelihood:	-142.71
No. Observations:	100	AIC:	291.4
Df Residuals:	97	BIC:	299.2
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.6302	0.301	5.415	0.000	1.033	2.228
x1	-0.1378	0.139	-0.991	0.324	-0.414	0.138
x2	3.1519	0.135	23.409	0.000	2.885	3.419

Omnibus:	1.426	Durbin-Watson:	2.235
Prob(Omnibus):	0.490	Jarque-Bera (JB):	1.075
Skew:	0.249	Prob(JB):	0.584
Kurtosis:	3.099	Cond. No.	24.6

Métodos de Diagnóstico

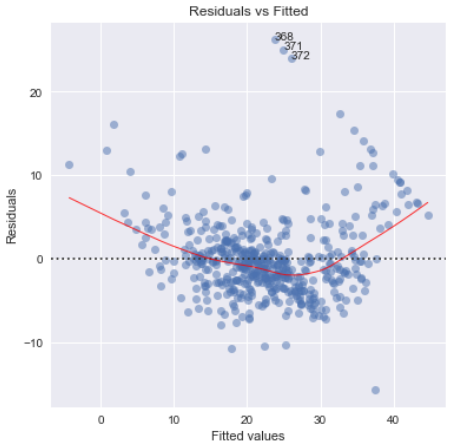
Luego de realizar el cálculo del modelo de regresión, típicamente hacemos varios plots de diagnóstico.



UVG UNIVERSIDAD DEL VALLE DE GUATEMALA

Métodos de Diagnóstico

Gráfico de Residuales (Residuals vs. Fitted)



Métodos de Diagnóstico

- Muestra si los residuales $\widehat{\varepsilon}_i = y_i - \widehat{y}_i$ tienen patrones no lineales.
  Podría haber una relación no lineal entre las variables predictoras y una variable de resultado y el patrón podría aparecer en este gráfico si el modelo no captura la relación no lineal.
- Si el modelo no cumple con el supuesto del modelo lineal, esperaríamos ver residuos que son muy grandes (gran valor positivo o gran valor negativo).
  Usualmente, valores estandarizados menores que -2 ó mayores que 2 se consideran problemáticos.
- También es útil para verificar el supuesto de linealidad y homocedasticidad.
  Buscamos asegurarnos de que no haya un patrón en los residuos y que estén igualmente distribuidos alrededor de $y = 0$.

Métodos de Diagnóstico

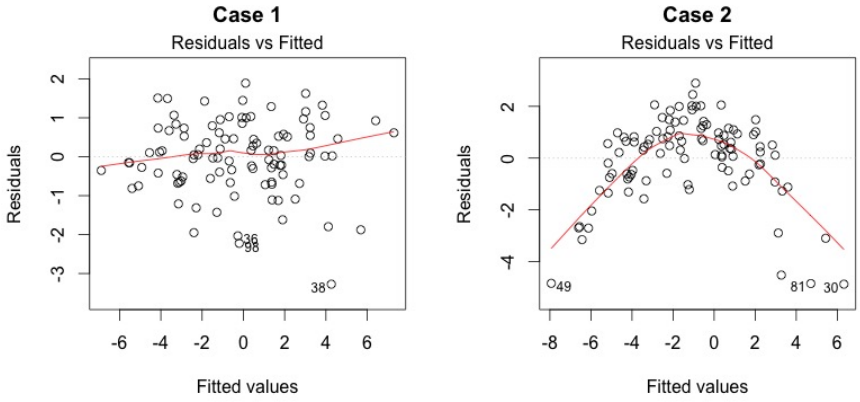
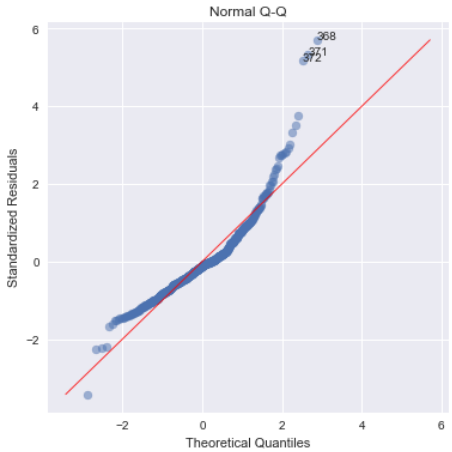


Gráfico de residuales: (a) buen ajuste; (b) mal ajuste.

Métodos de Diagnóstico

Gráficos Cuantil-Cuantil (*QQ plot*)

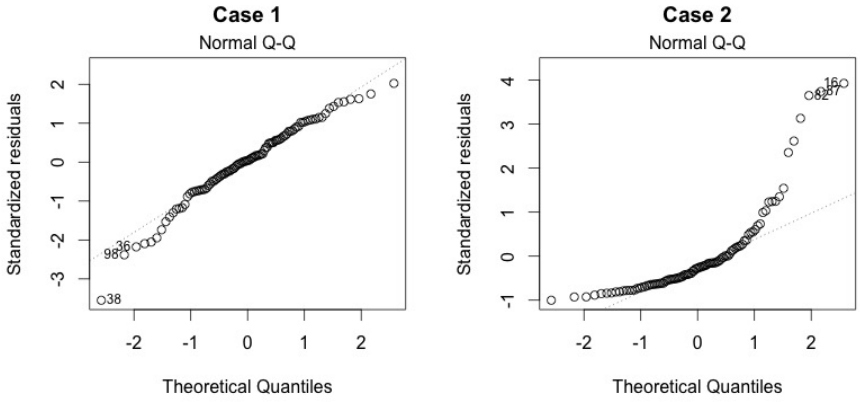


Métodos de Diagnóstico

La gráfica cuantil-cuantil o QQ es una prueba de comparación entre la distribución de los residuales $\widehat{\varepsilon}_i$ y una distribución normal estándar.

- El supuesto de normalidad de los residuos se puede evaluar comparando éstos con las observaciones normales “ideales” (recta identidad).
- Entre más alejadas las observaciones de la diagonal se considera que los residuales se alejan de una distribución normal.
- Similarmente, encontrar estructuras o comportamientos particulares, dan evidencia en contra de la normalidad.

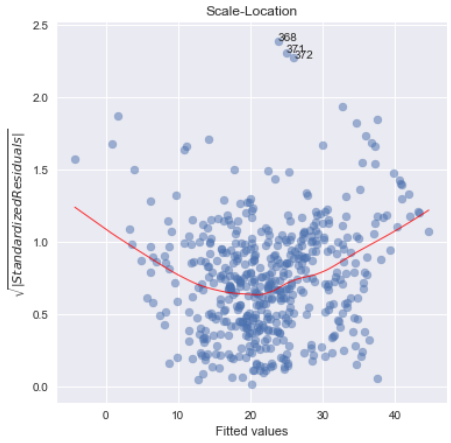
Métodos de Diagnóstico



Gráfica cuantil-cuantil: (a) buen ajuste; (b) maj ajuste.

Métodos de Diagnóstico

Residuos estandarizados (*Scale-location / Spread-location*)



Métodos de Diagnóstico

- Las gráficas de ubicación a escala (residuo estandarizado de raíz cuadrada vs. valor predicho) eson útiles para verificar el supuesto de homoscedasticidad.
- Muestra si los residuos se distribuyen por igual a lo largo de los rangos de predictores.
- También es útil para determinar aleatoriedad y normalidad.

Observación Importante!

- Los supuestos de una muestra aleatoria y observaciones independientes no se puede probar con gráficos de diagnóstico. Es una suposición que puede probar examinando el diseño del estudio.
- Tampoco es un criterio para decidir sí/no el modelo es bueno. Se usa más para entender los datos.

Métodos de Diagnóstico

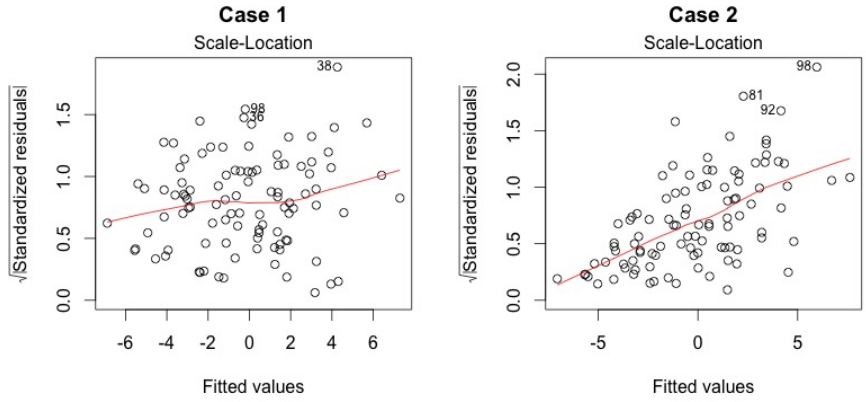


Gráfico de residuos estandarizados: (a) buen ajuste; (b) mal ajuste.

Puntos Influyentes

En el análisis de regresión, la **apalancamiento** (*leverage*) es una medida de qué tan lejos están los valores de las variables independientes de una observación de los de las otras observaciones.

Los puntos de alto apalancamiento son aquellas observaciones, (valores extremos o periféricos de las variables independientes), tales que la falta de observaciones vecinas significa que el modelo de regresión ajustado pasará cerca de esa observación.

En el modelo de regresión lineal ordinaria (OLS), el apalancamiento de la i-ésima observación $\mathbf{x}_i$ se define como

$$h_{ii} = (P)_{ii}, \quad \text{donde } P \text{ es la matrix de proyección } P = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T.$$

En particular, $h_{ii} = \mathbf{x}_i^T(\mathbb{X}^T\mathbb{X})^{-1}\mathbf{x}_i$.

Puntos Influyentes

h_{ii} también se llama la **auto-sensitividad** de $\mathbf{x}_i$, ya que $h_{ii} = \frac{\partial \widehat{y}_i}{\partial y_i}$.

El apacalancamiento h_{ii} está relacionado con la **influencia** de $\mathbf{x}_i$ sobre los coeficientes, ya que $\frac{\partial \beta}{\partial y_i} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbf{x}_i$. Además, con el grado de variación de los coeficientes, al remover la observación $(\mathbf{x}_i, y_i)$:

$$\widehat{\beta} - \widehat{\beta}^{(i)} = \frac{(\mathbb{X}^T\mathbb{X})^{-1}\mathbf{x}_i\,\widehat{\varepsilon}_i}{1 - h_{ii}}.$$

Distancia de Cook

Distancia de Cook: La distancia de Cook o **D de Cook** es una estimación de la influencia de un punto. Fue introducida R. Dennis Cook (1977).

Se usa para indicar puntos influyentes cuya validez merece la pena comprobar; o para indicar regiones del espacio de diseño donde sería bueno poder obtener más puntos de datos.

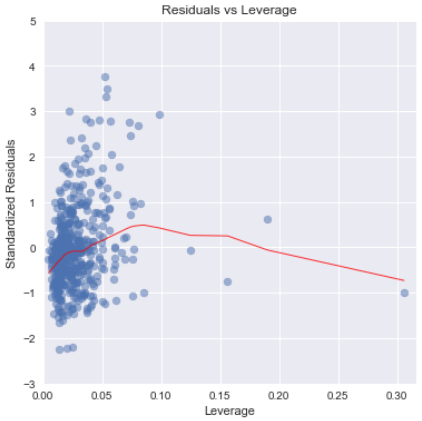
La distancia de Cook's D_i de la observación $\mathbf{x}_i$ se define como la suma de todos los cambios en el modelo de regresión cuando se elimina la observación $(\mathbf{x}_i, y_i)$:

$$D_i = \frac{\sum_{j=1}^{n}(\widehat{y}_j - \widehat{y}_{j(i)})}{ds^2},$$

donde $\widehat{y}_{j(i)}$ es la estimación de y_j al remover la boservación i, y $s^2 = \frac{1}{n-d}\widehat{\varepsilon}^T\widehat{\varepsilon}$ es el error cuadrático medio del modelo de regresión.

Puntos Influyentes

Puntos influyentes (*Residuals vs. Leverage*)



Puntos Influyentes

- Ayuda a encontrar casos atípicos influyentes (los **puntos palanca**) si los hay. No todos los valores atípicos influyen en el análisis de regresión lineal. Son casos extremos contra la regresipon, de modo que su presencia altera los resultados del modelo si se excluyen del análisis. En otras palabras: no se llevan bien con la tendencia en la mayoría de los casos.

- A diferencia de las otras parcelas, esta vez los patrones no son relevantes. Vigilamos los valores atípicos en la esquina superior derecha o en la esquina inferior derecha Buscamos casos fuera de una línea discontinua, la distancia de Cook.

Puntos Influyentes

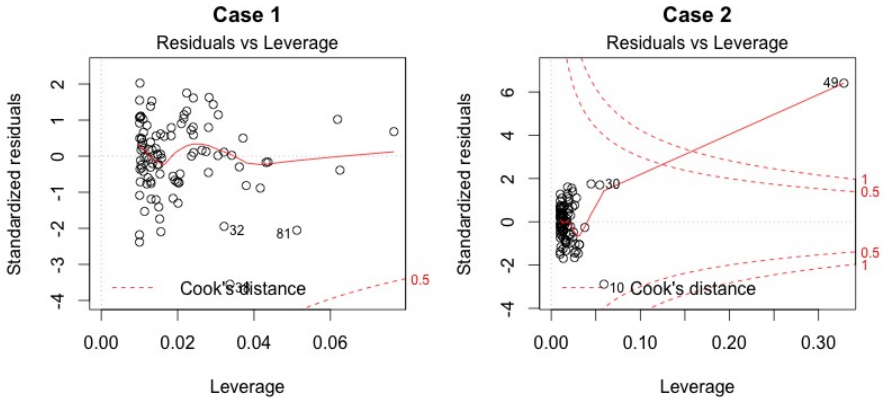


Gráfico de influencia y D de Cook: (a) buen ajuste; (b) mal ajuste.

Pruebas de Hipótesis

Existen también pruebas formales para verificar los supuestos.

Test de Normalidad: (para los $\widehat{\varepsilon_i}$)

- D'Agostino's K-squared test,
- **Jarque–Bera test**,
- Anderson–Darling test,
- Criterio de Cramer–von Mises,
- Lilliefors test (basado en Kolmogorov–Smirnov),
- **Shapiro–Wilk test**,
- Pearson's χ^2 test.
- Kolmogorov–Smirnov test, (sólo funciona si se conoce la media y varianza poblacional),

Pruebas de Hipótesis

Test de Jarque-Bera: (1980) Diseñado para evaluar la bondad de ajuste de una muestra dada la kurtosis y el coeficiente de asimetría. Define

$$JB = \frac{n}{6}\left(S^2 + \frac{1}{4}(K-3)^2\right),$$

donde n es el tamaño de la muestra, $S = \frac{\widehat{\mu}_3}{\widehat{\sigma}^3}$ el coeficiente de asimetría, y $K = \frac{\widehat{\mu}_4}{\widehat{\sigma}^4}$ la kurtosis.

Test de Shapiro-Wilk: (1965) Define el estadístico W como

$$W = \frac{\sum_{i=1}^{n} a_i \mathbf{x}_{(i)}}{\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})},$$

donde $(a_1, \ldots, a_n) = \frac{\mathbf{m}^T V^{-1}}{||V^{-1}\mathbf{m}||} = \frac{\mathbf{m}^T V^{-1}}{(\mathbf{m}^T V^{-1} V^{-1} \mathbf{m})^{1/2}}$, y el vector $\mathbf{m} = (m_1, \ldots, m_n)$ es el valor esperado de los estadísticos de orden de una normal estándar. V es la matriz de covarianza de esas estadísticas de orden.

Pruebas de Hipótesis

Pruebas de Homoscedasticidad:

- Bartlett's test,
- Levene's test,
- Brown–Forsythe test,
- **Box's *M*-test**,

Pruebas para Auto-correlación:

- **Durbin–Watson test**,
- Breusch–Godfrey test,
- Ljung–Box test.

Métodos de Diagnóstico

Entonces, ¿qué significa tener patrones en los residuos para su investigación? No es sólo una señal de *go/no-go*. Da información sobre el modelo y los datos. Es posible que el modelo actual no sea lo mejor.

En ese caso, es posible que uno quiera volver a la teoría e hipótesis iniciales. ¿Es realmente una relación (lineal) adecuada entre los predictores y el resultado?
Es posible que desee incluir otros términos adicionales, por ejemplo; o aplicar alguna transformación logarítmica puede representar mejor los fenómenos que le gustaría modelar.
¿O hay alguna variable importante que se dejó fuera del modelo?
O, tal vez, los datos fueron sesgados sistemáticamente al recopilarlos. Es posible que se desee rediseñar los métodos de recopilación de datos.