



Verificación de cumplimiento de los supuestos del modelo

ANÁLISIS DE REGRESIÓN

Supuestos del modelo

Las principales premisas al estudiar análisis de regresión son:

- ▶ La relación entre la respuesta y el (los) regresor(es) es lineal, al menos en forma aproximada
- ▶ El término del error tiene **media cero**
- ▶ El término del error tiene **varianza constante**
- ▶ Los errores no están correlacionados
- ▶ Los errores tienen **distribución normal**

Media cero

Media cero

La falta de cumplimiento de este supuesto conlleva que los estimadores del modelo pierdan su naturaleza insesgada, lo cual indica la importancia del mismo. Y precisamente el carácter tan básico del supuesto hace que la verificación del cumplimiento del mismo pueda ser en extremo complicada. De hecho, en el análisis de regresión tradicional se suele considerar suficiente admitir que si el resto de los supuestos verificables se cumple, éste también se cumple.

Sin embargo, cabe mencionar que al momento de revisar el cumplimiento de normal en los residuales se hace bajo la consideración de una distribución normal estándar, una distribución normal con medio cero. Por lo general se incluye en el análisis el valor de la media de los residuales generados bajo el modelo de regresión.

Varianza Constante

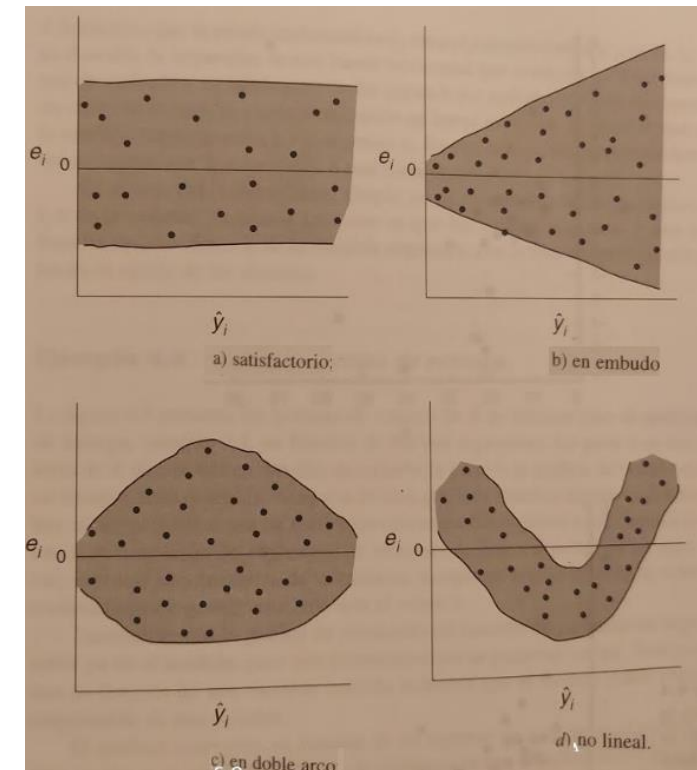
GRÁFICA Y ESTIMADA VS ERRORES

Gráfica y estimada vs errores

Se realiza la gráfica de respuesta estimada vs residuales, en eje horizontal y vertical, respectivamente

Lo que se desea es que no existan patrones visibles en la dispersión de los residuales mostrados en la gráfica

a) No hay defectos obvios en el modelo, b) la varianza de los residuales es creciente (si el embudo va de mas a menos implicaría varianza decreciente), c) indica que la varianza de los errores no es constante (con frecuencia se presenta cuando y es una proporción), d) podría indicar que se necesitan otras variables regresoras en el modelo (incluso modelos polinómicos)



Incorrelación

PRUEBA DURBIN-WATSON

Prueba Durbin-Watson

La gráfica de residuales en secuencia temporal puede indicar que los errores en un periodo se correlacionan con los de otros periodos. La correlación entre los errores del modelo en distintos periodos se llama autocorrelación.

Para detectar la presencia de autocorrelación de forma analítica se puede aplicar la prueba Durbin-Watson ya que se basa en la hipótesis de que los errores del modelo de regresión se generan en un proceso autorregresivo de primer orden, que se observa a intervalos de tiempo igualmente espaciados.

Hipótesis	$H_0: \rho = 0$ equivalente a la incorrelación de los datos analizados		$H_1: \rho > 0$
Estadístico de prueba	$d = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2}$		
Criterio de decisión	Rechazo H_0 si $d < d_L$	No Rechazo H_0 si $d > d_U$	<i>Prueba no concluyente si $d_L < d < d_U$</i>

Valores críticos del estadístico Durbin-Watson

Table A.6 Critical values of the Durbin-Watson statistic

Sample Size	Probability in Lower Tail (Significance Level = α)	k = Number of Regressors (Excluding the Intercept)									
		1		2		3		4		5	
		d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	.01	.81	1.07	.70	1.25	.59	1.46	.49	1.70	.39	1.96
	.025	.95	1.23	.83	1.40	.71	1.61	.59	1.84	.48	2.09
	.05	1.08	1.36	.95	1.54	.82	1.75	.69	1.97	.56	2.21
20	.01	.95	1.15	.86	1.27	.77	1.41	.63	1.57	.60	1.74
	.025	1.08	1.28	.99	1.41	.89	1.55	.79	1.70	.70	1.87
	.05	1.20	1.41	1.10	1.54	1.00	1.68	.90	1.83	.79	1.99
25	.01	1.05	1.21	.98	1.30	.90	1.41	.83	1.52	.75	1.65
	.025	1.13	1.34	1.10	1.43	1.02	1.54	.94	1.65	.86	1.77
	.05	1.20	1.45	1.21	1.55	1.12	1.66	1.04	1.77	.95	1.89
30	.01	1.13	1.26	1.07	1.34	1.01	1.42	.94	1.51	.88	1.61
	.025	1.25	1.38	1.18	1.46	1.12	1.54	1.05	1.63	.98	1.73
	.05	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83

40	.01	1.25	1.34	1.20	1.40	1.15	1.46	1.10	1.52	1.05	1.58
	.025	1.35	1.45	1.30	1.51	1.25	1.57	1.20	1.63	1.15	1.69
	.05	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72	1.23	1.79
50	.01	1.32	1.40	1.28	1.45	1.24	1.49	1.20	1.54	1.16	1.59
	.025	1.42	1.50	1.38	1.54	1.34	1.59	1.30	1.64	1.26	1.69
	.05	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77
60	.01	1.38	1.45	1.35	1.48	1.32	1.52	1.28	1.56	1.25	1.60
	.025	1.47	1.54	1.44	1.57	1.40	1.61	1.37	1.65	1.33	1.69
	.05	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77
80	.01	1.47	1.52	1.44	1.54	1.42	1.57	1.39	1.60	1.36	1.62
	.025	1.54	1.59	1.52	1.62	1.49	1.65	1.47	1.67	1.44	1.70
	.05	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77
100	.01	1.52	1.56	1.50	1.58	1.48	1.60	1.45	1.63	1.44	1.65
	.025	1.59	1.63	1.57	1.65	1.55	1.67	1.53	1.70	1.51	1.72
	.05	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78

Source: Adapted from "Testing for Serial Correlation in Least Squares Regression II," by J. Durbin and G. S. Watson, *Biometrika*, Vol. 38, 1951, with permission of the Royal Society.

Ejemplo

Considere los datos “Gastar para vender” de la tabla anexa en la cual se desea analizar una posible relación lineal entre la inversión mensual (en miles de pesos) en un pequeño negocio y el rendimiento en ventas (en miles de pesos) del mismo.

Después de realizar una regresión lineal simple la cual queda significativa, se hace el cálculo de los residuales correspondientes.

x	y
gasto mensual	rendimiento ventas
25	34
16	14
42	48
34	32
10	26
21	29
19	20

Ejemplo

Haciendo uso de la tabla mostrada a la derecha, se tiene que $d=2.9245$

De la tabla de valores críticos se tiene que para una significancia de 0.05 no se rechaza H_0 : los residuales no muestran correlación.

Dada la cantidad de datos no es de forma inmediata la conclusión, note que si el valor de d_u crece conforme crece el tamaño muestral y $d_u=1.36$ es menor que el estadístico de prueba, un valor más pequeño de d_u mantiene la desigualdad.

TABLA AUXILIAR ESTADÍSTICO DURBIN-WATSON

gasto mensual	rendimiento ventas	<i>Residuos et</i>	et^2	diferencias ²
25	34	4.07154213	16.5774553	
16	14	-8.61685215	74.2501409	160.995349
42	48	4.26073132	18.1538314	165.832156
34	32	-5.24006359	27.4582665	90.265104
10	26	8.25755167	68.1871596	182.185618
21	29	2.32114467	5.3877126	35.240928
19	20	-5.05405405	25.5434624	54.3935563
		sumas	235.558029	688.912711

Normalidad

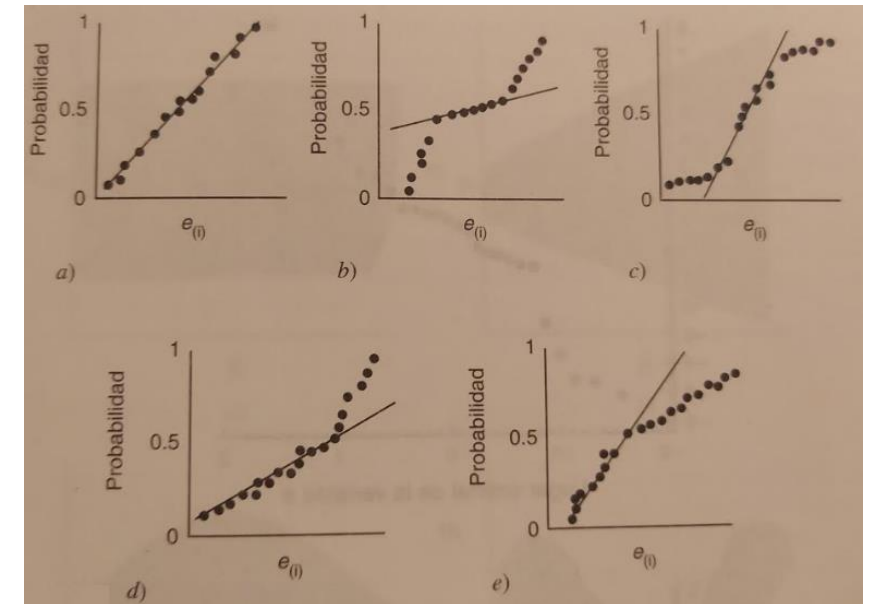
GRÁFICA DE PROBABILIDAD NORMAL QQPLOT

Gráfica qqplot

Se suele utilizar la grafica de probabilidad normal QQplot

La grafica idealizada es aquella en la que los puntos caen aproximadamente sobre la recta como en el grafico a), si las colas de la distribución son demasiado gruesas para poder considerar la distribución normal la grafica observada será de la forma b), si las colas son muy delgadas como para poder considerarla normal la grafica observada será de la forma c), si la distribución presenta asimetría positiva la grafica observada será de la forma d); mientras que una asimetría negativa la grafica observada será de la forma e)

Cabe mencionar que las pequeñas desviaciones respecto a esta hipótesis no afectan mucho al modelo, pero una no normalidad grande es potencialmente seria; dado a que las predicciones y los intervalos de confianza dependen de esta suposición.



¿Cómo realizar la gráfica?

1. Realice la regresión de interés para obtener de ésta los residuales correspondientes
2. Ordene los residuales de menor a mayor, considerando incluso el signo de éstos
3. Asigne el rango i correspondiente al valor del residual de 1 hasta n
4. Genere el valor de k correspondiente a los cuantiles, donde $k = \frac{i-0.375}{n+0.25}$
5. Obtenga el valor Z_k de tal manera que $P(Z < Z_k) = k$ para cada valor k generado; es decir, se busca el valor Z_k cuya probabilidad acumulada considerando una distribución normal estándar sea igual a k
6. Obtener los valores esperados multiplicando por \sqrt{MSE} , dicho valor es generado en la tabla ANOVA de la regresión de la cual se obtienen los residuales analizados
7. Grafique residuales ordenados vs los valores esperados calculados

Ejemplo

Considere los datos “Gastar para vender” de la tabla anexa en la cual se desea analizar una posible relación lineal entre la inversión mensual (en miles de pesos) en un pequeño negocio y el rendimiento en ventas (en miles de pesos) del mismo.

Después de realizar una regresión lineal simple la cual queda significativa, se hace el cálculo de los residuales correspondientes.

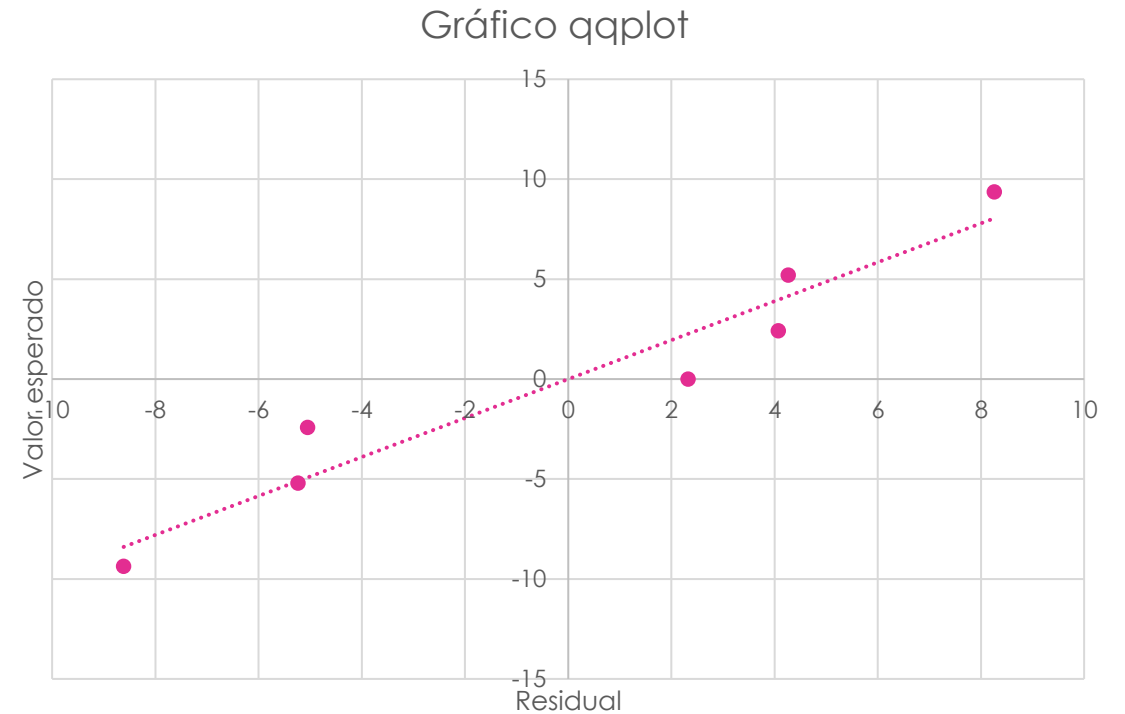
x	y
gasto mensual	rendimiento ventas
25	34
16	14
42	48
34	32
10	26
21	29
19	20

Ejemplo

TABLA AUXILIAR PARA QQPLOT				
residuo ordenado	rango i	k	Zk	valor esperado
-8.616852146	1	0.086206897	-1.36448875	-9.365563485
-5.240063593	2	0.224137931	-0.75829256	-5.204760459
-5.054054054	3	0.362068966	-0.35293399	-2.422464573
2.321144674	4	0.5	0	0
4.07154213	5	0.637931034	0.35293399	2.422464573
4.26073132	6	0.775862069	0.75829256	5.204760459
8.257551669	7	0.913793103	1.36448875	9.365563485

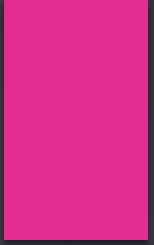
EJEMPLO

La gráfica qqplot generada es la mostrada a la derecha, de la cual se debe comentar sobre el ajuste de los pares a la línea recta



Comentarios Finales

- ▶ Ya que se ha generado una regresión significativa, incluso después de analizar diversos modelos y de seleccionar uno como candidato al modelo de los datos, se procede a la verificación del cumplimiento de los supuestos de los residuales.
- ▶ La verificación se hace para un modelo de regresión, ya sea simple o múltiple.
- ▶ En ocasiones es posible realizar las pruebas visuales, mediante graficas; sin embargo, siempre que se cuente con la prueba analítica ha de preferirse éste procedimiento ya que puede concluirse considerando porcentajes de confianza.
- ▶ Hasta aquí sólo verificamos de forma individual si se cumplen o no los supuestos; sin embargo, un tema a revisar son las correcciones o medidas a realizar en búsqueda del cumplimiento de dichos supuestos.



Las imágenes anexas a este contenido fueron tomadas del libro d indicado como referencia del curso.

Introducción al Análisis de Regresión Lineal

Douglas C. Montgomery

3ra Edición