ANALISIS DE REGRESION

ANALISIS DE RESIDUALES

El análisis de los residuos permite validar algunos de los supuestos del modelo de regresión lineal múltiple como: linealidad, varianza constante, independencia, normalidad, etc.; es un método efectivo para detectar deficiencias en el modelo, utilizando diversos tipos de gráficos. El análisis de residuales además permite comprobar la adecuación del modelo.

Definición de residuales

Los residuos mínimo-cuadráticos vienen dados por:

$$e_i = y_i - \hat{y}_i \qquad i=1,\ldots,n$$

o en forma matricial

$$\vec{\mathbf{e}} = \vec{\mathbf{Y}} - \hat{\mathbf{Y}}.$$

Tienen media cero y varianza estimada igual al CMRes

1° El residuo Ordinario está definido como :

$$e_i = y_i - \hat{y}_i .$$

Algunos residuos transformados son:

Métodos para transformar o escalar Residuales

Para eliminar el efecto que la escala de medición de las variables regresoras, ejerce sobre la variable de respuesta, es conveniente transformar o escalar los residuos.

2º Residuo estandarizado:

Es la razón entre los residuos y su desviación estándar.
Estos residuos tienen promedio cero y varianza uno.

$$d_i = \frac{e_i}{\hat{\sigma}}$$

di > 3 indica que la "i-ésima observación" es un potencial dato discordante o atípico.

3º Residuo estudentizado:

Estos residuales mejoran la aproximación de la varianza al utilizar la desviación estándar exacta de i-ésimo residual, Recordando que se halló el vector de los residuales como:

$$e = (I - H)y$$

Donde H= $X(X'X)^{-1}X'$ es la matriz de sombrero, recodemos que esta matriz posee propiedades útiles:
es simétrica $(H' = H)$,
es idempotente $(HH = H)$.
La matriz $I - H$ es simétrica e idempotente.
Si utilizamos $y = X\beta + \varepsilon$

$$e = (I - H)y = (I - H)(X\beta + \varepsilon) = X\beta - HX\beta + (I - H)\varepsilon$$
$$= X\beta - X(X'X)^{-1}X'X\beta + (I - H)\varepsilon = (I - H)\varepsilon$$
$$e = (I - H)\varepsilon$$

Los residuales son una transformación lineal de las observaciones y y los errores ε.

La matriz de covarianza de los residuales es

$$Var(e) = Var\left[(I - H)\varepsilon\right] = (I - H)Var(\varepsilon)(I - H)' = \sigma^2(I - H)$$

La varianza del i-ésimo residual es

$$Var(e_i) = \sigma^2(1 - h_{ii})$$
$$Cov(e_i, e_j) = -\sigma^2 h_{ij}$$

Sabemos que hii es una medida del lugar o ubicación del i-ésimo punto en el espacio de x.

Luego es necesario examinar los residuales estudentizados que se definen como:

$$r_i = \frac{e_i}{\sqrt{MS_{Res}(1-h_{ii})}} \qquad \text{para todo } i = 1,2,3,\ldots,n$$

O También:

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}},$$

Sin embargo, ya que cualquier punto con un residual grande y también una h_{ii} grande tiene una influencia potencial muy grande sobre el ajuste por mínimos cuadrados, se recomienda por lo general examinar los residuales estudentizados.

Residuales PRESS

En este caso la i-ésima observación es eliminada del análisis, y luego con el resto de los datos; es decir, a las n-1 observaciones restantes, se procede a estimar los parámetros del modelo, calcular los valores de la variable ajustada y posteriormente los residuos correspondientes. Permite identificar observaciones donde el modelo no se ajusta bien a los datos o que produzca inadecuadas predicciones.

Esto es: $e_{(i)} = y_i - \hat{y}_{(i)}$ llamado **error de predicción** correspondiente a la observación eliminada.

Si esto se repite n veces, esto es para cada observación, a esos errores de predicción se le llama **residuales PRESS** y se define como

$$e_{(i)} = y_i - \hat{y}_{(i)} = \frac{e_i}{1-h_{ii}} \quad i=1\ldots n$$

Y la varianza del i-ésimo residual PRESS es

$$Var(e_{(i)}) = \sigma^2 / (1- h_{ii})$$

por lo que un ***residual PRESS estandarizado*** es

$$d_{(i)} = \frac{e_i}{\sigma\sqrt{1-h_{ii}}}$$

Si se usa el CMRes para estimar σ^2,
resulta ser el ***residual PRESS estudentizado*** que se describió anteriormente.

$$r_{(i)} = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}},$$

En general, una gran diferencia entre el residual ordinario y el residual PRESS indica un punto donde el modelo se ajusta bien a los datos, pero un modelo formado sin ese punto hace malas predicciones

ESTADÍSTICA PRESS

Podemos decir que la Estadística PRESS, es una medida que indica lo adecuado del modelo para realizar predicciones de nuevas observaciones, es decir, se considera como una medida de la calidad del modelo. La estadística PRESS se define como:

$$PRESS = \sum_{i=1}^{n}\left(\frac{e_i}{1-h_{ii}}\right)^2$$

Un modelo adecuado es aquel que tenga PRESS$\rightarrow$0.

R^2 para la predicción

Se puede calcular el coeficiente de determinación para la predicción a partir de la estadística PRESS y se define como:

$$R^2_{prediccion} = 1 - \frac{PRESS}{SCT}$$

Este estadístico da cierta indicación de la capacidad predictiva del modelo de regresión.

Una aplicación muy importante de la estadística PRESS es comparar modelos de regresión.
En general, un modelo con pequeño valor de PRESS es preferible a uno con PRESS grande.

R de Student.

El residual estudentizado r_i utiliza la varianza estimada $\hat{\sigma}^2$=MSRes que es generado con las n observaciones.
Otra forma es, eliminar la observación atípica y volver a estimar σ^2
Al estimador de σ^2, obtenido de esa forma se le representa por $S^2_{(i)}$, que se define como:

$$S^2_{(i)} = \frac{(n-p)MS_{Res} - \frac{e_i^2}{(1-h_{ii})}}{n-p-1}$$

Esto permitirá obtener residuos estudentizados externamente y se llama R de Student o R-Student y se define como:

$$t_i = \frac{e_i}{\sqrt{S^2_{(i)}(1-h_{ii})}}$$ para todo $i = 1,2,3,\ldots,n$ con distribución $t_{(n-p-1)}$