**REGRESIÓN LINEAL**

**Supuestos principales:** para que la RL sea válida

1. Relación lineal entre X e Y
2. Los errores (e) se distribuyen normalmente.
   1. ei ​= yi ​− y​ˆi. ​Valor real – valor predicho. Los errores deben seguir una distribución normal con media cero; esto garantiza que las estimaciones de los coeficientes (betas) y sus pruebas de hipótesis (p-valores) sean validas. Media cero, en promedio los residuos deben ser = 0, de lo contrario el modelo tiene un sesgo sistemático.

p-valor: es la probabilidad de obtener un resultado tan extremo como el observado de tus datos. Mide que tan compatible son los datos con la H0. No mide H0, directamente si es cierta o falsa, si no que tan improbable sería mi resultado si H0 fuera cierta.

La normalidad permite aplicar la teoría de los mínimos cuadrados ordinarios (OLS). Los estimadores de los coeficientes (beta) son BLUE.

La normalidad se evalúa por medio de gráficas

Histograma de residuos: debe parecerse a una campa de gauss (simétrico, unimodal)

Q-Q Plot: compara los cuantiles de los residuos contra una distribución normal teórica, si los puntos se alinean en la diagonal -< normalidad aceptada

Por medio de pruebas estadísticas:

Shapiro-Will test: H0: los datos provienen de una distribución normal; el p-valor > 0.05 -> no rechazo H0 -> normalidad aceptada

Smirnov-Kolmogorov test: similar al anterior, compara con una distribución normal

Jarque-Bera test: evalúa la curtosis y asimetría de los residuos. Ideal para grandes muestras.

Si los datos no son normales se debe

Transformar las variables: raíz cuadrada a Y

Usar modelos robustos (RANSAC, HuberRegressor)

Aumentar la complejidad de los modelos: polinomios o modelos no lineales

Mas datos.

1. Homocedasticidad (varianza constante de los errores)
   1. Los errores deben tener la misma varianza en todos lo niveles de X, si esto no se cumple -> heterocedasticidad (problema común en datos económicos). O sea, para todas las X’s el error deben ser =, de lo contrario el modelo sería heterocedastico. La homocedasticidad se evalúa por medio de

Gráficas de residuos vs predicciones

Pruebas estadísticas.

Si no hay heterocedasticidad se debe:

Trasformar la variable dependiente

Usar modelos robustos

1. Independencia de los errores
   1. Los residuos no deben estar correlacionados entre sí. En series temporales suele violarse este supuesto
2. No multicolinealidad (X’s no deben ser fuertemente correlacionada entre sí)
   1. Aunque no es un supuesto directo de los residuos, si las variables están muy correlacionadas -> los residuos reflejan problemas en la estimación de los coeficientes. Cuando hay multicolinealidad los <betas> pueden cambiar mucho con pequeños cambios en los datos

La multicolinealidad se puede detectar por medio de:

La matriz de correlación: pares con correlación >0.8 o < 0.8 sospecha de multicolinealidad

VIF (varianza Inflator Factor) ide cuanto aumenta la varianza de un coeficiente por culpa de la correlación con otras variables. VIF=1 -> no hay problema. VIF entre 5 y 10 -> cuidado. VIF > 10 multicolinealidad severa.

La multicolinealidad se pue de manejar de la siguiente manera:

Eliminar una de las variables correlacionadas

Cambiar variables

Regularización

Recoger más datos

Métricas de evaluación

1. R2: proporción de la varianza de Y explicada por el modelo (0-1):
   1. Mide que proporción de la variable dependiente (Y) puede ser explicada por las X’s, o sea que tan bien el modelo se ajusta a los datos. R2 = 1 (100%) -> el modelo explica toda la variabilidad R2 = 0 el modelo no explica nada y R2 < 0 el modelo es peor que predecir la media.

R2 = 80% significa que el 80% de la variación de Y es explicada por las X’s

1. MSE (Error cuadrático medio)
   1. Mide el promedio de los errores al cuadrado entre los valores reales y los valores predichos por el modelo. MSE bajo, mejor el modelo (menos error en promedio). Como los errores están al cuadrado los errores grandes se penalizan más fuertes que los pequeños. El MSE es sensible a los outlier (si hay un valor muy desviado, el MSE aumenta)
2. RMSE (Raíz del error cuadrático medio):
   1. Es la raíz del MSE, mide el MSE pero en las mismas unidades que Y, más intuitivo de interpretar
3. MAE (Error absoluto medio):
   1. Mide el promedio de los errores absolutos entre los valores reales y los valores predichos. Indica en promedio, cuando se equivoca el modelo, en la misma unidad de la variable dependiente.

Para saber si el MAE es grande o pequeño depende la Y, se pude calcular la relación entre el MAE/media de Y. También se puede comparar con la desviación estándar (DE) si el MAE<DE significa que el modelo predice mejor que simplemente usar la media