Unidad 3: Medidas de bondad de ajuste

Módulo 3: Modelamiento estadístico

Iris Ashimine

Medidas de bondad de ajuste

Las **medidas de bondad de ajuste** evalúan qué tan bien un modelo de regresión explica la variabilidad de la variable dependiente. En otras palabras, nos dicen qué tan bien los valores estimados del modelo representan los valores reales observados en los datos.

Si un modelo tiene una buena bondad de ajuste, significa que la variabilidad de la variable dependiente se explica en gran parte por las variables independientes incluidas en la regresión. Sin embargo, un buen ajuste **no garantiza causalidad ni capacidad de predicción**.

Coeficiente de determinación

R^2, el coeficiente de determinación, es la fracción de la varianza muestral de la variable dependiente y_i que es explicada por el (los) regresor(es) x_i.

Matemáticamente se puede escribir como el ratio entre la *Suma Explicada de Cuadrados* (SEC) y la *Suma Total de Cuadrados* (STC).

STC es la suma cuadrada de las desviaciones de y_i y su media $\bar{y}$. **SEC** es la suma cuadrada de las deviaciones de los valores *predichos* de $\hat{y}_i$ y del promedio $\bar{y}$.

$$SEC = \sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2 STC = \sum_{i=1}^{N} (y_i - \bar{y})^2 R^2 = \frac{SEC}{STC}$$

La *Suma de Residuos Cuadrados* (SRC) se define como:

$$SRC = \sum_{i=1}^{N} (\hat{u}_i)^2$$

Como $STC = SRC + SCE$ también se puede calcular como:

$$R^2 = 1 - \frac{SRC}{STC}$$

El R^2 se encuentra entre 0 y 1, y es fácil ver que, de no exisitr errores ($SRC = 0$), esto implicaría que $R^2 = 1$.

- Si $R^2 = 1 \rightarrow$ el modelo explica el 100% de la variabilidad de los datos.
- Si $R^2 = 0 \rightarrow$ el modelo no explica nada, y es igual a predecir con la media de y.

Ejemplo Si tenemos un $R^2 = 0.8$ en un modelo que preedice salarios en función de la educación y la experiencia, significa que el 80% de la variación de los salarios es explicada por estas variables, mientras que el 20% restante sedebe a otros factores no incluidos en el modelo.

Discusión

En las ciencias sociales, los coeficientes de detemrinación bajos son comunes, debido a la variabilidad inherente del comportamiento humano. Por tanto, en estas ciencas, los investigadores se enfocan más en el tamaño del efecto. Adicionalmente, R^2 captura la correlación, no la **causalidad**. En estudios de inferencia *causal* los investigadores priorizan efectos insesgados muy por encima de la capacidad de explicar la varianza que tienen estos.

Coeficiente de determinación Ajustado

R^2 ajustado es una medida de bondad de ajuste que adiferencia que el R^2 que incrementa siempre que una nueva variable es añadida al modelo (no penaliza incrementar variables irrelevantes), corrige esto, penalizando por cada regresor k del modelo.

$$\bar{R^2} = 1 - \frac{N-1}{N-1-k}\frac{SRC}{STC}$$

Entonces ahora, añadir un regresor provoca: - SSR disminuya - (N-1)/(N-1-k) incremente Una medida que contrarresta el mejor ajuste automático del modelo a una variable adicional.

Por construcción, el $R^2 \geq \bar{R^2}$.

Estadístico F

Es una medida que prueba la significancia general de un modelo (no de una sola variable).

- H_0: Todos los coeficientes (exceptuando el intercepto) son cero.
- H_1: Al menos un coeficiente no es cero.

Se calcula como:

$$F = \frac{\frac{SEC}{k}}{\frac{SRC}{N-k-1}}$$

Donde SEC es la Suma Explicada de Cuadrados y SRC es la Suma de Residuos Cuadrados. Los grados de libertadad son $df : N - k - 1$.

Error estándar de la regresión

El error estándar de la regresión (SER) es una medida de dispersión de la distribución de y alrededor de la línea de regresión. Se mide como la raíz cuadrada de la Suma de Residuos Cuadráticos (SRC) con un ajuste por grados de libertad:

$$SER = \sqrt{\frac{SRC}{N-k-1}}$$

$N - k - 1$ ajusta el sesgo hacia abajo introducido al estimar $k + 1$ coeficientes (k regresores e intercepto)

EL **SER** indica cuánto error comete el modelo al hacer predicciones. La regla general es: *mientra menor sea el SER, mejor es el ajuste del modelo.*

Ejemplo

```r
# Datos de ejemplo
horas_estudiadas <- c(2, 4, 6)
notas_actuales <- c(80, 90, 95)
notas_predichas <- c(78, 92, 94)

# Residuos
residuo <- notas_actuales - notas_predichas
# Suma cuadrada de los residuos
src <- sum(residuo^2)

# SER
n <- length(horas_estudiadas) #tres observaciones
k <- 1  #numero de predictores (más el intercepto)
ser <- sqrt(src/ (n - k - 1)) #1

cat("SSR:" ,src, "\n" ,"SER:", ser)
```

```
## SSR: 9
##  SER: 3
```

Esto nos indica que, en promedio, las notas predichas se desvían en 3 puntos de las notas actuales.

Raíz del error cuadrático medio

$RMSE$ (Root Means Squared Error), es la raíz cuadrada del Error Cuadrático Medio (MSE), sin ajuste por grados de libertad. Es una medida dimilar al **SER**, pero más utilizada en modelos predictivos:

$$MSE = \frac{SRC}{N} RMSE = \sqrt{MSE}$$

Es una medida útil cuando se quiere evaluar la capacidad predictiva de un modelo (análisis de regresión, forecasting, machine learning).

Ejemplo

```r
# Datos
precio_actual <- c(200000, 240000, 300000)
precio_pred <- c(210000, 230000, 310000)

# RMSE
rmse <- sqrt(mean((precio_actual - precio_pred)^2))
cat("RMSE:",rmse)
```

```
## RMSE: 10000
```

Esto infica que, en promedio, los precios predicos difieren de los reales en $10.000.

Supuestos

1. Linealidad

2. Muestra i.i.d.
3. No colinealidad perfecta entre regresores
4. Exogeneidad
5. Condiciones de regularidad
6. Homocedasticidad[1]

Linealidad en los parámetros

Este supuesto establece que la relación entre la variable dependiente y y las variables independientes x es **lineal en parámetros**, aunque no necesariamente en las variables.

Los parámetros β_k deben tener una relación lineal con la variable dependiente, y el error entra de manera aditiva.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + e$$

Este supuesto nos dice que el modelo es una *combinación lineal de los coeficientes β, aunque las variables x pueden transformarse de forma no lineal. **Ejemplo**

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}^2 + e_i$$

El modelo incluye un **término cuadrático**, pero sigue siendo **lineal en los parámetros** $\beta_0, \beta_1, \beta_2$, por lo que no se está violando el supuesto de linealidad.

$$y_i = e^{\beta_0 + \beta_1 x_{i1}} + e_i$$

En cambio, en este segundo modelo, β_0 y β_1 aparecen dentro de una **función exponencial**, lo que viola este supuesto.

¿Qué hacer si el modelo no es lineal en parámetros?

En estos casos, podemos intentar aplicarle una transformación para lograr la linealidad en parámetros. En casos más complejos, se pueden aplicar métodos como **Máxima Verosimilitud (MLE)**.

Datos Independientes e Identicamente Distribuidos

Uno de los supuestos fundamentales en Mínimos Cuadrados Ordinarios es que los datos deben ser **independientes e idénticamente distribuidos** (i.i.d.). Este supuesto garantiza que las observaciones en la muestra no están correlacionadas entre sí y que provienen de una misma distribución.

1. **Las observaciones son independientes:** - Cada (y_i, x_i) es independiente de las demás observaciones (y_j, x_j), donde $i \neq j$. - Esto implica también que el error de una observación no debe estar relacionado con el error de otra: $E[e_i, e_j] = 0$ para $i \neq j$.
2. **Las observaciones tienen la misma distribución:** - Los errores siguen la misma distribución en toda la muestra

Ejemplo

Supongamos que estimamos un modelo de ventas mensuales y en función de los gastos en marketing x_1, y variables de característica de los empleados de cada empresa x_2 - Escenario 1: si utilizamos una serie de tiempo, las ventas de un mes dependen de las ventas del mes anterior → los residuos van a estar autocorrelacionados, violando el supuesto de independencia.

[1] Como se vio en la clase 1, la homocedasticidad es un supuesto que se suele asumir para simplificar demostraciones y bajo el cual funciona el Teorema de Gauss Markov, sin embargo, es la excepción en la práctica.

- Escenario 2: Utilizamos datos de corte transversal, sin embargo, las observaciones de las características de los empleados de cada empresa, estarán relacionadas entre sí, puesto a que provienen de una misma empresa → se viola el supuesto de independencia.

Soluciones

Utilizar errores estándares robustos o agrupados (cluestered standard errors), o cambiar de modelo.

Exogeneidad

La esperanza o media condicional del error es cero, esto quiere decir que, sin importar el valor de x, el término de error e no puede mostrar un patrón sistemático con los regresores.

$$E[e_i|x_{ik}] = 0$$

Este supuesto nos garantiza que la relación encontrada entre x y y es causal y no está sesgada por efectos no observados.

¿Qué hacer si el supuesto de exogeneidad no se cumple?

- Utilizar Variables Instrumentales: Encontrar una variable relacionada con x pero no con e, y utilizarla como "instrumento" para obtener una estimación no sesgada.
- Diseños experimentales: RCT o Randomized Controlled Trial, si se implementa correctamente se puede eliminar el problema de endogeneidad.

No colinealidad perfecta entre regresores

En la muestra, ninguno de los regresores es una constante y no existe una relación lineal entre ellos. Este supuesto exige que **ninguna variable independiente sea combinación lineal exacta de otras**. Matemáticamente, esto significa que:

$$Det[X'X] \neq 0$$

Si el determinante es cero, la matriz no tiene inversa y no podemos calcular:

$$\hat{\beta} = (X'X)^{-}1(X'Y)$$

Intuitivamente, esto quiere decir que si tenemos una variable explicativa que está completamente explicada por una o más variables (es una combinación lineal exacta), esto hace imposible separar el efecto de cada variable sobre y, lo que hace que β no sea único.

- Este supuesto solo descarta la colinealidad *perfecta* entre variables explicativas; la **colinealidad imperfecta** está permitida.
- Si una variable independiente es resultado de la combinación lineal de otras varibles independientes, es superflua y debe ser eliminada.
- Variables constantes también deben ser descartadas (colinalidad con el intercepto).

Ejemplo de colinearidad perfecta Si se estima el siguiente modelo para medir el efecto de diferentes factos sobre la presión arterial de los pacientes:

$$presion\_art = \beta_0 + \beta_1 Sodio + \beta_2 Sal + \gamma_j x_j + e$$

El problema de este ejemplo yace en la relación exacta que existe entre los gramos de sal y su contenido en mg de sodio:

$$Sodio(mg) = 0.393 \times Sal$$

Esto significa que una de las variables es simplemente un múltiplo de la otra.

```r
# Crear datos con colinealidad perfecta
set.seed(123)
n <- 100
x1 <- rnorm(n)  # Variable independiente 1
x2 <- 2 * x1    # X2 es una combinación exacta de X1
y <- 5 + 3 * x1 + rnorm(n)

# Intentamos estimar el modelo
df <- data.frame(y, x1, x2)
modelo <- lm(y ~ x1 + x2, data = df)

# Verificamos la salida
summary(modelo)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9073 -0.6835 -0.0875  0.5806  3.2904
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.89720    0.09755   50.20   <2e-16 ***
## x1           2.94753    0.10688   27.58   <2e-16 ***
## x2                NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9707 on 98 degrees of freedom
## Multiple R-squared:  0.8859, Adjusted R-squared:  0.8847
## F-statistic: 760.6 on 1 and 98 DF,  p-value: < 2.2e-16
```

R detecta y elimina automáticamente una de las variables

Condiciones de regularidad

Las condiciones de regularidad que nos dicen que los datos atípicos (outliers) son poco probables, para ello la variable dependiente y las variables independientes deben tener una kurtosis finita: $E[y^4] < \infty; E[x_j^4] < \infty$.

Este supuesto implica que tanto y como x_j **no tienen colas extremadamente pesadas** en su distribución, es decir, no presentan valores extremos con alta probabilidad.

En términos prácticos, este supuesto ayuda a garantizar que el Teorema Cetral del Límite (TCL) se aplique correctamente, y que por tanto, los estimadores tengan una distribución asintótica normal.

Homocedasticidad

El supuesto de *homocedasticidad* establece que la varianza **condicional** de los errores e_i es una constante para todas las observaciones:

$$E[e_i^2|x_i] = \sigma^2 \ \forall i$$

En otras palabras, la dispersión de los errores no depende de los valores de x. - Cuando hay homocedasticidad, el ruido en los datos es el mismo sin importar los valores de las variables explicativas.

- En un modelo que intenta predecir los ingresos de las personas en función de la educación, en personas con menos años de educación, los ingresos suelen ser bastante homogéneos, pero a niveles más altos de educación, los ingresos varían mucho.

Cómo comprobar los supuestos en R

1. Linealided: usar el test `RESET` de Ramsey o gráficos de residuos para evaluar si la especificación lineal del modelo tiene un buen ajuste a los datos. Examina si los valores ajustados del modelo (o sus potencias) contienen información adicional que el modelo original no está capturando.
2. Independencia: usar la prueba de *Durbin-Watson* para evaluar la independencia de los errores. \

Normalidad de los errores: usar la prueba de *Shapiro-Wilk* o revisar visualmente mediante un **histograma de los residuos** o un `qq-plot`.

3. No colinealidad: Si existe colinealidad perfecta, el software no podrá encontrar los resultados, pero para probar colinealidad imperfecta se puede utilizar el *Factor de Inflación de la Varianza*.

$$VIF = \frac{1}{1 - R_k^2}$$

siendo R_k^2 el coeficiente de determinación de la regresión auxiliar de la variable x_k sobre el resto de las variables explicativas. El VIF para cada término del modelo mide el efecto combinado que tienen las dependencias entre regresores sobre la varianza de ese término. 4. Homocedasticidad: usar la prueba de *Breush-Pagan* que evalúa si la varianza de los errores depende de x, o la prueba de *White*, es similar al BP, pero no asume una relación lineal entre la varianza del error y x.