Diagnósticos Regresión Lineal

Abril 10, 2024

Prof. Sergio Béjar

Departamento de Estudios Políticos, CIDE

Objetivo(s) para Hoy

1. Entender los supuestos de la regresión lineal simple (OLS).
2. Familiarizarnos con las pruebas básicas de diagnóstico que debemos hacer en "todos" los modelos OLS.
3. Entender que debemos hacer cuando nos falle alguna prueba de diagnóstico.

Supuestos de OLS

OLS tiene muchos supuestos.

Es complejo determinar cuáles supuestos son los más importantes.

- Punto de vista maximalista: cualquier violación a los supuestos es suficiente para rechazar el modelo.
- Pero: la violación de algunos supuestos tiene ciertas *implicaciones* importantes.

LINI

Pensemos en la siguinete abreviación de los supuestos de OLS: LINI

- `L`: la variable dependiente y es una función *l*ineal de las variables independientes y de control.
- `I`: los errores (residuales) son *i*ndependendientes (i.e. no hay autocorrelación).
- `N`: la distribución de los errores es *n*ormal
- `I`: la varianza de los errores is *i*gual/constante (i.e. no heteroskedasticidad).

En realiad el orden no importa tanto.
Y estos supuestos tampoco nos dicen nada sobre la validez de los datos o que tan representativa es la muestra.

Primero, una Advertencia

Los libros generalmehte mencionarán como problemas lo siguiente:

1. Multicolinearidad
2. Especificación (“toidas las variables relevantes”)

Mi advertencia: son importantes, pero constituyen un problema trivial.

Multicolinearidad

Multicolinearidad existe cuando dos o más variables explicativas están altamente correlacionadas. Por lo tanto, nuestra regresión (OLS) no nos da efectos parciales confiables.

- Colinearidad *Perfecta*: el modelo no se puede "identificar".
- Alta colinearidad: tus errores estándar son MUY grandes.

Diagnóstico: matriz de correlación, factor inflación-varianza

- Matriz de correlación: valor absoluto por encima de .8 indica un problema.
- Factor inflación-varianza: un valor arriba de 5 indica un problema.

Solución(es):

1. No inlcuir una de las variables.
2. Análisis de componente principal (i.e. crear una medida latente)

Problemas de Especificación

Los problemas de especificación de un modelo generalmente se presentan "incluyendo todas las variebles" que predicen y. Advertencia:

- No existe una prueba formal para esto.

Los problemas de especificación son críticos únicamente cuando hacemos ajustes en las variables de control. Veamos a los siguientes escenarios:

1. `X` y `Z` ambas explican variación en `Y`, pero `X` y `Z` no están correlacionadas.
2. `X` y `Z` ambas explican variación en `Y`, y `X` y `Z` están correlacionadas.
3. `X` (pero no `Z`) explican variación en `Y`, y `X` y `Z` están correlacionadas.

Escenarios de problemas de especificación

Primer escenario: omitir `Z`no tiene influencia en el “verdadero” efecto de `X` en `Y`.

- Omitir `Z` reducirá la R^2, lo que no significa un problema real para identificación causal.
- *Nos tenemos que preguntar cuál es el objetivo de nuestro modelo*.

Segundo escenario: omitir `Z` sesga la relación entre `X` y `Y`.

- Incluir `Z` arregla esto, pero genera un problema de multicolinearidad.

Tercer escenario: esto es conocido como el problema de la variable instrumental (un tema avanzado).

Linearidad

OLS asume que y es una función lineal de variables que la predicen.

- Esto implica que el modelo es *aditivo*.
- Sin este supuesto el modelo deja de ser lineal.

Diagnóstico: fundamentalmente visual (plot residuos ajustados). Pero también:

- Prueba Utts (1982) del "Arcoiris"
- Prueba Harvey-Collier
- *Nota:. Ninguna de estas pruebas es muy buena; lo más recomendable es mirar a los datos/modelo*.

Soluciones:

- ¿Nuevo Modelo?
- Transformación logarítmica
  - e.g. si $y = abc$, entonces $log(y) = log(a) + log(b) + log(c)$
- ¿Interacciones/elevar variables al cuadrado?

Regresamos al Ejemplo de Turnout y Educación

```
## cargamos datos
datos <- rio::import("https://raw.githubusercontent.com/Sergio-Bejar/MC
```

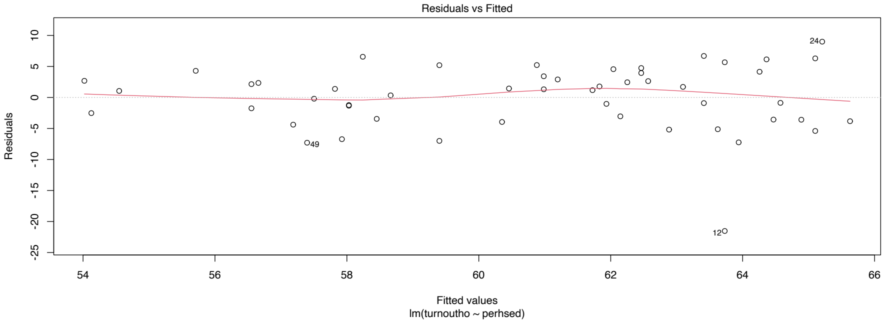
Estimamos el modelo (M1)

```
summary(M1 <- lm(turnoutho ~ perhsed, data=datos))
#>
#> Call:
#> lm(formula = turnoutho ~ perhsed, data = datos)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -21.529  -3.510   1.176   3.676   8.994
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept) -32.3027    21.3948  -1.510    0.138
#> perhsed       1.0553     0.2423   4.355 6.77e-05 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 5.247 on 49 degrees of freedom
#> Multiple R-squared:  0.2791, Adjusted R-squared:  0.2644
#> F-statistic: 18.97 on 1 and 49 DF,  p-value: 6.765e-05
```

Plot Valores Esperados vs. Residuales

Para identificar problemas de linealidad podemos usar la siguiente función que grafica los valores esperados de y y los residuales.

```
plot(M1, which = 1)
```



Si vemos una relación clara lineal entonces hay un problema de linealidad.

Errores Independientes (o: No Autocorrelación)

Otro supuesto grande: OLS asume que los datos son obtenidos aleatoriamente de una población. Es decir, no hay patrones de dependencia espacial, temporal o multinivel.

- La inclusión de una observaión no debe tener efecto en la inclusión de otra observación.
- El valor residual de una observación no puede depender en los residuales de otras observaciones.
- Si esto ocurre, OLS pierde su valor inferencial.

Diagnóstico: Hay 3 situaciones comunes.

1. Series de tiempo (i.e. y depende de valores pasados de y)
2. Modelos "Multinivel"/Jerárquicos
3. Sesgo de variables omitidas

Soluciones: Generalmente resolvemos este problema estimando un modelo muy diferente al que tenemos.

Checando Independencia de Errores (Series de Tiempo)

Usamos la función `dwtest` de la librería `lmtest` para checar aurocorrelación.

```
dwtest(M1)
#> 
#>  Durbin-Watson test
#> 
#> data:  M1
#> DW = 1.8147, p-value = 0.2548
#> alternative hypothesis: true autocorrelation is greater than 0
```

Un p-value por encima de 0.05 es consistente con la hipótesis nula de no autocorrelación.

Checando Independencia de Errores (Series de Tiempo)

Otra prueba que podemos usar para detectar autocorrelación es la Breusch-Godfrey. Usamos la función `dwtest` de la librería `lmtest` para checar aurocorrelación.

```
bgtest(M1)
#> 
#>  Breusch-Godfrey test for serial correlation of order up to 1
#> 
#> data:  M1
#> LM test = 0.335, df = 1, p-value = 0.5627
```

Si el valor de p está por debajo de .05 entonces tenemos un problema.

Soluciones Para Problemas de Autocorrelación

- Incluir tendencia de tiempo (t^2, t^3, etc.).
- Usar primeras diferencias.
- Usar efectos restardados (lagged effects).

Normalidad de Errores

OLS asume que la distribución de residuales es normal con una media de 0 y cierta varianza. Advertencias:

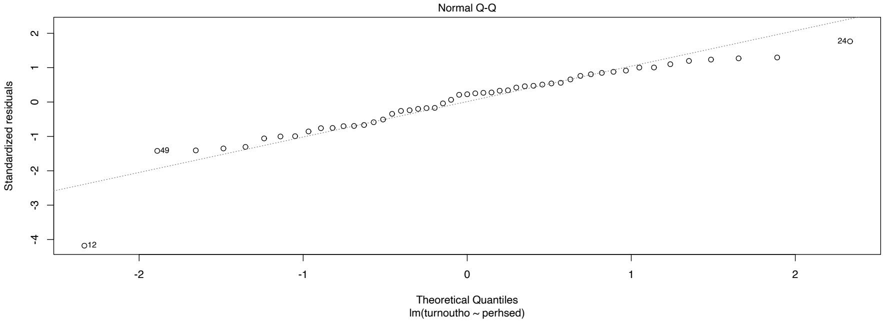
- Esto no significa que la variable dependiente es normal.
- No dice nada sobre las variables independientes o de control.
- Las pruebas de diagnóstico no son muy buenas porque son muy sensibles al tipo de VD que tengamos.
- La implicación de esta violación es sobre los errores, no sobre la regresión en sí.

Diagnósticos: Q-Q plot, y otras pruebas de normalidad que no son muy buenas.

Soluciones: usar otro tipo de modelos (diferentes a OLS.. una opción es GLS). Generalmente este problema sucede cuando estamos forzando OLS en casos cuando tenemos una VD con un conjunto finito de valores.

Evaluando Normalidad de Errores

```r
plot(M1, which = 2)
```



Evaluando Normalidad de Errores

También podemos usar las pruebas de Shapiro o Kolmogorov-Smirnoff, vía las funciones `shapiro.test()`. Un p-value menor a 0.05 rechaza la hipótesis nula de normalidad en la distribución de los residuos:

```
shapiro.test(resid(M1))
#>
#>  Shapiro-Wilk normality test
#>
#> data:  resid(M1)
#> W = 0.90809, p-value = 0.0007924
```

Homoesquedasticidad

OLS asume que la dispersión de los errores no depende en los valores esperados (homoesquedasticidad).

- Si esto sucede, la línea de regresión está bien pero los errores estándar NO.
- Esto tiene consecuencias muy importantes para pruebas de significancia de nuestros coeficientes.

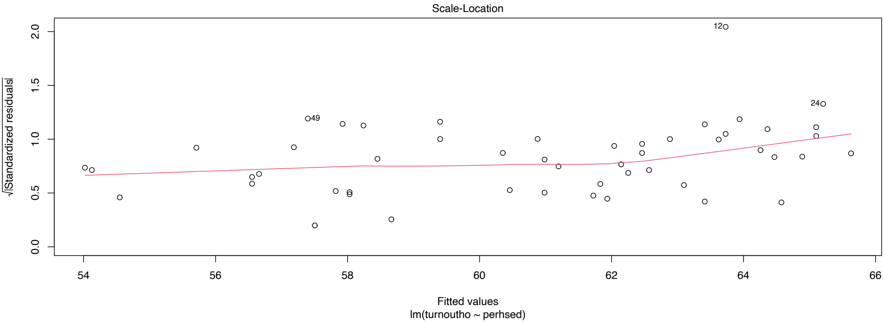
Diagnósticos: Plot de residuales esperados, prueba de Breusch-Pagan.

Solutions: más "pruebas de robustes" comparando OLS con otros estimadores.

- e.g. transformación de VD/VI, mínimos cuadrados ponderados (WLS), bootstrapping.

Evaluando Homoesquedasticidad

```r
plot(M1, which = 3)
```



Buscamos que: (a) la línea roja sea aproximadamente horizontal y (b) la dispersión de los puntos no cambie mucho en funció de los valores esperados.

Evaluando Homoesquedasticidad

La prueba de Breusch-Pagan nos ayudará a determinar con más precisión si tenemos un problema de heteroesquedasticidad.

```
bptest(M1)
#> 
#>  studentized Breusch-Pagan test
#> 
#> data:  M1
#> BP = 2.0467, df = 1, p-value = 0.1525
```

La hipótesis nula en esta prueba es homoesquedasticidad. Si la rechazamos con un p-value bajo (menor a 0.05), tenemos heteroesquedasticidad.

Solucionando Heteroesquedasticidad

En caso de tener heteroesquedasticidad podemos lidiar con este problema corrigiendo los errores estándar después de estimar el modelo y diagnosticarlo. Para ello, usamos la función `coeftest` del paquete `{sandwich}`.

```
coeftest(M1, vcov = vcovHC)
#>
#> t test of coefficients:
#>
#>              Estimate Std. Error t value  Pr(>|t|)
#> (Intercept) -32.30270   20.93430 -1.5431    0.1293
#> perhsed       1.05529    0.24134  4.3726 6.389e-05 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Estos resultados son obtenidos con errores estándar robustos. Esta es una práctica muy común en ciencias políticas y relaciones internacionales.

Solucionando Heteroesquedasticidad

También podemos estimar una regresión de mínimos cuadrados ponderados (WLS) para minimizar el problema de heteroesquedasticidad.

```r
# Calculamos el inverso de la varianza
# queremos dar menos peso a observaciones con varianza alta
# y más pedo a obervaciones con varianza alta
weights <- 1 / (datos$perhsed)^2

# Fit WLS model
m_wls <- lm(turnoutho ~ perhsed, weights = weights, data = datos)
```

Solucionando Heteroesquedasticidad

```
summary(m_wls)
#> 
#> Call:
#> lm(formula = turnoutho ~ perhsed, data = datos, weights = weights)
#> 
#> Weighted Residuals:
#>       Min        1Q    Median        3Q       Max 
#> -0.23663 -0.03900  0.01319  0.04122  0.09727 
#> 
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)    
#> (Intercept) -32.4347    20.6929  -1.567    0.123    
#> perhsed       1.0568     0.2349   4.499 4.21e-05 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusión

OLS tiene supuestos que debemos conocer.

- Hay variación en la importancia de los supuestos de OLS.
- Y esa variación en la importancia tiene efectos más o menos importantes en las inferencias que hacemos cuando usamos OLS.

Independientemente, *siempre* debemos asegurarnos que nuestras estimaciones sean lo más robustas posibles.

Table of Contents