

# Diagnosticos OLS y Errores Robustos

## Validando Nuestros Modelos de Salarios

EC3003B - Economia Laboral Aplicada

Tecnologico de Monterrey, Campus Puebla

Jueves 12 de febrero, 2025 | 3-5pm

# Contenido de la Sesión

- 1 Introducción
- 2 Heterocedasticidad
- 3 Multicolinealidad
- 4 Observaciones Influyentes
- 5 Especificación del Modelo
- 6 Resumen de Diagnosticos
- 7 Resumen

## Hasta ahora:

- M01: Ecuacion de Mincer, interpretacion de coeficientes
- M02: Variables categoricas e interacciones

## Hasta ahora:

- M01: Ecuacion de Mincer, interpretacion de coeficientes
- M02: Variables categoricas e interacciones

## Pero... ¿podemos confiar en nuestros resultados?

- ¿Los supuestos de OLS se cumplen?
- ¿Los errores estandar son correctos?
- ¿Hay observaciones influyentes distorsionando los resultados?

## Hasta ahora:

- M01: Ecuacion de Mincer, interpretacion de coeficientes
- M02: Variables categoricas e interacciones

## Pero... ¿podemos confiar en nuestros resultados?

- ¿Los supuestos de OLS se cumplen?
- ¿Los errores estandar son correctos?
- ¿Hay observaciones influyentes distorsionando los resultados?

## Hoy aprenderemos

Diagnosticos para validar modelos OLS y correcciones cuando hay problemas.

Al finalizar esta sesion, podras:

- ➊ Detectar heterocedasticidad y aplicar correcciones
- ➋ Identificar multicolinealidad y sus consecuencias
- ➌ Detectar observaciones influyentes
- ➍ Evaluar especificacion del modelo
- ➎ Elegir el tipo correcto de errores estandar

# ¿Que es Heterocedasticidad?

**Homocedasticidad** (supuesto OLS):

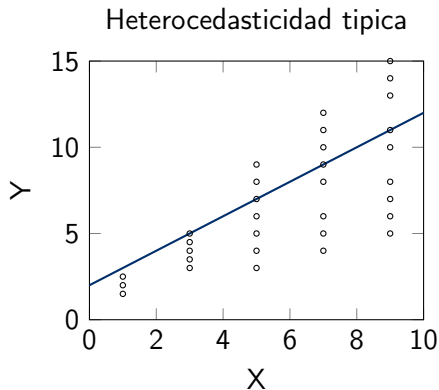
$$\text{Var}(\varepsilon_i|X) = \sigma^2 \quad \forall i$$

Varianza constante para todas las observaciones.

**Heterocedasticidad** (violacion):

$$\text{Var}(\varepsilon_i|X) = \sigma_i^2$$

Varianza cambia con  $X$ .



# ¿Por que es Comun en Datos Salariales?

## Razon economica:

- Personas con alta educacion tienen opciones mas diversas
- Mas variabilidad en salarios de profesionistas que de obreros
- Sectores de alta paga tienen mas dispersion



# ¿Por que es Comun en Datos Salariales?

## Razon economica:

- Personas con alta educacion tienen opciones mas diversas
- Mas variabilidad en salarios de profesionistas que de obreros
- Sectores de alta paga tienen mas dispersion

## Consecuencias de ignorar heterocedasticidad

- $\hat{\beta}$  sigue siendo **insesgado y consistente**
- Pero  $\text{Var}(\hat{\beta})$  esta **mal calculada**
- Errores estandar incorrectos  $\rightarrow$  inferencia invalida
- Tests t y F son **invalidos**

## 1. Grafico de residuos vs valores ajustados:

```
reg ln_salario escolaridad experiencia experiencia2  
rvfplot, yline(0)
```

## 1. Grafico de residuos vs valores ajustados:

```
reg ln_salario escolaridad experiencia experiencia2  
rvfplot, yline(0)
```

## 2. Test de Breusch-Pagan:

```
estat hettest  
* H0: Varianza constante (homocedasticidad)  
* Rechazar H0 = evidencia de heterocedasticidad
```

# Detectar Heterocedasticidad

## 1. Grafico de residuos vs valores ajustados:

```
reg ln_salario escolaridad experiencia experiencia2  
rvfplot, yline(0)
```

## 2. Test de Breusch-Pagan:

```
estat hettest  
* H0: Varianza constante (homocedasticidad)  
* Rechazar H0 = evidencia de heterocedasticidad
```

## 3. Test de White:

```
estat imtest, white  
* Mas general, no asume forma funcional especifica
```

## Errores robustos de Huber-White (HC):

\* Forma clasica (incorrecta si hay heterocedasticidad)

```
reg ln_salario escolaridad experiencia experiencia2
```

\* Forma robusta (valida con heterocedasticidad)

```
reg ln_salario escolaridad experiencia experiencia2, robust
```

# Solucion: Errores Estandar Robustos

## Errores robustos de Huber-White (HC):

\* Forma clasica (incorrecta si hay heterocedasticidad)

```
reg ln_salario escolaridad experiencia experiencia2
```

\* Forma robusta (valida con heterocedasticidad)

```
reg ln_salario escolaridad experiencia experiencia2, robust
```

## ¿Que hace robust?

- No cambia  $\hat{\beta}$  (mismos coeficientes)
- Cambia  $\widehat{\text{Var}}(\hat{\beta})$
- Errores estandar validos **incluso con heterocedasticidad**
- Inferencia (tests t, IC) ahora es valida

## Regla practica

En datos salariales, **siempre** usar `. robust.`

# Variantes de Errores Robustos

Tipo	Comando Stata	Uso
HC1 (default)	<code>, robust</code>	Heterocedasticidad general
Cluster	<code>, cluster(var)</code>	Correlacion dentro de grupos
HAC	<code>newey</code>	Datos de series de tiempo
Bootstrap	<code>, vce(bootstrap)</code>	Muestras pequenas

# Variantes de Errores Robustos

Tipo	Comando Stata	Uso
HC1 (default)	, robust	Heterocedasticidad general
Cluster	, cluster(var)	Correlacion dentro de grupos
HAC	newey	Datos de series de tiempo
Bootstrap	, vce(bootstrap)	Muestras pequenas

## ¿Cuándo usar cluster?

Si los errores estan correlacionados dentro de grupos:

- Empleados de la misma empresa
- Trabajadores del mismo estado
- Observaciones del mismo individuo en panel

```
reg ln_salario escolaridad experiencia, cluster(estado)
```



# ¿Que es Multicolinealidad?

## Definicion

Existe **multicolinealidad** cuando las variables independientes estan altamente correlacionadas entre si.

# ¿Que es Multicolinealidad?

## Definicion

Existe **multicolinealidad** cuando las variables independientes estan altamente correlacionadas entre si.

## Ejemplos en ecuacion de Mincer:

- Edad y experiencia:  $Exp = Edad - 5 - 6$
- Escolaridad y nivel educativo (dummies)
- Ingreso familiar e ingreso individual

# ¿Que es Multicolinealidad?

## Definicion

Existe **multicolinealidad** cuando las variables independientes estan altamente correlacionadas entre si.

## Ejemplos en ecuacion de Mincer:

- Edad y experiencia:  $Exp = Edad - 5 - 6$
- Escolaridad y nivel educativo (dummies)
- Ingreso familiar e ingreso individual

## Consecuencias

- $\hat{\beta}$  sigue siendo insesgado
- Pero  $Var(\hat{\beta})$  es **grande**
- Coeficientes individuales imprecisos
- Signos pueden ser “incorrectos”

# Detectar Multicolinealidad: VIF

**Factor de Inflacion de Varianza (VIF):**

$$\text{VIF}_j = \frac{1}{1 - R_j^2}$$

Donde  $R_j^2$  es el  $R^2$  de regresar  $X_j$  sobre las demas  $X$ .

# Detectar Multicolinealidad: VIF

**Factor de Inflacion de Varianza (VIF):**

$$VIF_j = \frac{1}{1 - R_j^2}$$

Donde  $R_j^2$  es el  $R^2$  de regresar  $X_j$  sobre las demas  $X$ .

```
reg ln_salario escolaridad experiencia experiencia2 edad  
vif
```

VIF	Interpretacion
1	Sin multicolinealidad
1-5	Moderada (aceptable)
5-10	Alta (preocupante)
>10	Severa (problematica)

## ① Eliminar variables redundantes

- No incluir edad Y experiencia (están relacionadas)
- No incluir escolaridad Y dummies de nivel

## ① Eliminar variables redundantes

- No incluir edad Y experiencia (están relacionadas)
- No incluir escolaridad Y dummies de nivel

## ② Combinar variables

- Crear índices compuestos
- Usar componentes principales

## ① Eliminar variables redundantes

- No incluir edad Y experiencia (están relacionadas)
- No incluir escolaridad Y dummies de nivel

## ② Combinar variables

- Crear índices compuestos
- Usar componentes principales

## ③ Aumentar la muestra

- Mas datos = mejor identificación



## ① Eliminar variables redundantes

- No incluir edad Y experiencia (están relacionadas)
- No incluir escolaridad Y dummies de nivel

## ② Combinar variables

- Crear índices compuestos
- Usar componentes principales

## ③ Aumentar la muestra

- Mas datos = mejor identificación

## ④ Aceptar y reportar

- Si la predicción conjunta es correcta
- Reportar correlaciones y VIF

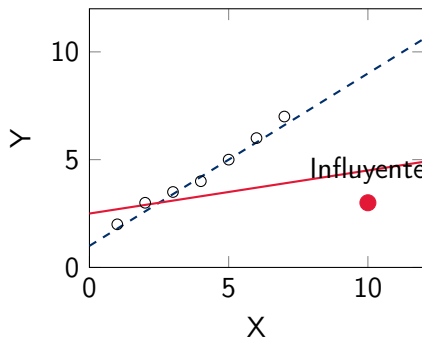
# Outliers vs Observaciones Influyentes

## Outlier:

- Valor inusual de  $Y$  o  $X$
- Puede o no afectar la regresion

## Observacion influyente:

- Afecta sustancialmente  $\hat{\beta}$
- Removerla cambia los resultados



## 1. Leverage (apalancamiento):

```
predict leverage, leverage  
summarize leverage, detail
```

Mide que tan extremo es  $X_i$ . Regla:  $h_i > 2(k + 1)/n$  es alto.

## 1. Leverage (apalancamiento):

```
predict leverage, leverage  
summarize leverage, detail
```

Mide que tan extremo es  $X_i$ . Regla:  $h_i > 2(k + 1)/n$  es alto.

## 2. Residuos estudentizados:

```
predict rstudent, rstudent  
list if abs(rstudent) > 2
```

Mide que tan extremo es  $Y_i$  dado  $X_i$ .

## 1. Leverage (apalancamiento):

```
predict leverage, leverage  
summarize leverage, detail
```

Mide que tan extremo es  $X_i$ . Regla:  $h_i > 2(k+1)/n$  es alto.

## 2. Residuos estudentizados:

```
predict rstudent, rstudent  
list if abs(rstudent) > 2
```

Mide que tan extremo es  $Y_i$  dado  $X_i$ .

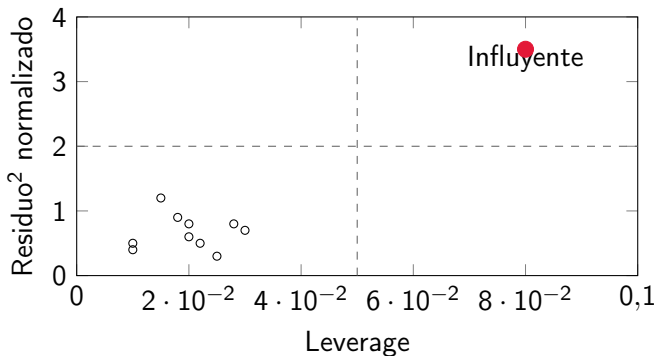
## 3. Distancia de Cook:

```
predict cooksd, cooksd  
list if cooksd > 4/e(N)
```

Combina leverage y residuo. Mide impacto en  $\hat{\beta}$ .

## Grafico: Leverage vs Residuo

```
reg ln_salario escolaridad experiencia experiencia2, robust  
lvr2plot, mlabel(id)
```



# ¿Que Hacer con Observaciones Influyentes?

## ① Verificar si son errores de datos

- Salarios de \$0 o \$999,999
- Edades imposibles

# ¿Que Hacer con Observaciones Influyentes?

## ① Verificar si son errores de datos

- Salarios de \$0 o \$999,999
- Edades imposibles

## ② Reportar sensibilidad

- Estimar con y sin la observacion
- Si los resultados cambian mucho, reportar ambos



# ¿Que Hacer con Observaciones Influyentes?

## ① Verificar si son errores de datos

- Salarios de \$0 o \$999,999
- Edades imposibles

## ② Reportar sensibilidad

- Estimar con y sin la observacion
- Si los resultados cambian mucho, reportar ambos

## ③ Usar metodos robustos

- Regresion robusta: `rreg`
- Regresion cuantilica (M05)

# ¿Que Hacer con Observaciones Influyentes?

## ① Verificar si son errores de datos

- Salarios de \$0 o \$999,999
- Edades imposibles

## ② Reportar sensibilidad

- Estimar con y sin la observacion
- Si los resultados cambian mucho, reportar ambos

## ③ Usar metodos robustos

- Regresion robusta: `rreg`
- Regresion cuantilica (M05)

## ④ Winsorizar o truncar

- Reemplazar valores extremos por percentiles

## Tipos de errores:

- **Variables omitidas:** Falta una variable relevante
- **Forma funcional incorrecta:** Deberia ser cuadratico, no lineal
- **Variables irrelevantes:** Incluir variables que no pertenecen

## Tipos de errores:

- **Variables omitidas:** Falta una variable relevante
- **Forma funcional incorrecta:** Deberia ser cuadratico, no lineal
- **Variables irrelevantes:** Incluir variables que no pertenecen

## Test RESET de Ramsey

Detecta errores de forma funcional.

- $H_0$ : Modelo correctamente especificado
- $H_1$ : Faltan terminos no lineales

# Test RESET en Stata

```
* Modelo sin termino cuadratico
reg ln_salario escolaridad experiencia, robust
estat ovtest

* Modelo con termino cuadratico
reg ln_salario escolaridad experiencia experiencia2, robust
estat ovtest
```

## Interpretacion:

- $p\text{-valor} < 0,05$ : Rechazar  $H_0$ , hay problema de especificacion
- $p\text{-valor} \geq 0,05$ : No rechazar, no hay evidencia de mala especificacion

# Test RESET en Stata

```
* Modelo sin termino cuadratico
reg ln_salario escolaridad experiencia, robust
estat ovtest

* Modelo con termino cuadratico
reg ln_salario escolaridad experiencia experiencia2, robust
estat ovtest
```

## Interpretacion:

- $p\text{-valor} < 0,05$ : Rechazar  $H_0$ , hay problema de especificacion
- $p\text{-valor} \geq 0,05$ : No rechazar, no hay evidencia de mala especificacion

## Nota

El test RESET no dice *cual* es el problema, solo que existe.

# Checklist de Diagnosticos

Problema	Test	Comando Stata	Solucion
Heterocedasticidad	Breusch-Pagan	<code>estat hetttest</code>	, robust
Heterocedasticidad	White	<code>estat imtest, white</code>	, robust
Multicolinealidad	VIF	<code>vif</code>	Eliminar/combinar
Influencia	Cook's D	<code>predict, cooksd</code>	Verificar/reportar
Especificacion	RESET	<code>estat ovtest</code>	Agregar terminos
Normalidad	Shapiro-Wilk	<code>swilk residuos</code>	N grande OK

# Flujo de Trabajo Recomendado

```
* 1. Estimar modelo
reg ln_salario escolaridad experiencia experiencia2, robust

* 2. Guardar residuos y predicciones
predict residuos, residuals
predict fitted, xb

* 3. Diagnosticos
estat hettest           // Heterocedasticidad
vif                     // Multicolinealidad
estat ovtest           // Especificacion
predict cooksd, cooksd  // Influencia
summarize cooksd, detail

* 4. Graficos
rvfplot, yline(0)       // Residuos vs fitted
lvr2plot                // Leverage vs residuo
```



## Problemas diagnosticados:

- Heterocedasticidad
- Multicolinealidad
- Observaciones influyentes
- Especificación incorrecta

## Mensaje principal:

En datos salariales, **siempre** usar errores robustos y verificar observaciones influyentes antes de reportar resultados.

# ¿Preguntas?

Proxima Sesion:

## **M04: Descomposicion Oaxaca-Blinder**

Lunes 16 de febrero, 3-5pm

**Recordatorio: E1 se entrega HOY 11:59pm**