

# Diagnósticos OLS y Errores Robustos

## Validando Nuestros Modelos de Salarios

EC3003B - Economía Laboral Aplicada

Tecnológico 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 Diagnósticos
- 7 Resumen

## Hasta ahora:

- M01: Ecuación de Mincer, interpretación de coeficientes
- M02: Variables categoricas e interacciones

## Hasta ahora:

- M01: Ecuación de Mincer, interpretación de coeficientes
- M02: Variables categoricas e interacciones

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

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

## Hasta ahora:

- M01: Ecuación de Mincer, interpretación de coeficientes
- M02: Variables categoricas e interacciones

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

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

## Hoy aprenderemos

Diagnósticos para validar modelos OLS y correcciones cuando hay problemas.

Al finalizar esta sesión, podrás:

- 1 Detectar heterocedasticidad y aplicar correcciones
- 2 Identificar multicolinealidad y sus consecuencias
- 3 Detectar observaciones influyentes
- 4 Evaluar especificación del modelo
- 5 Elegir el tipo correcto de errores estándar

# ¿Qué es Heterocedasticidad?

**Homocedasticidad** (supuesto OLS):

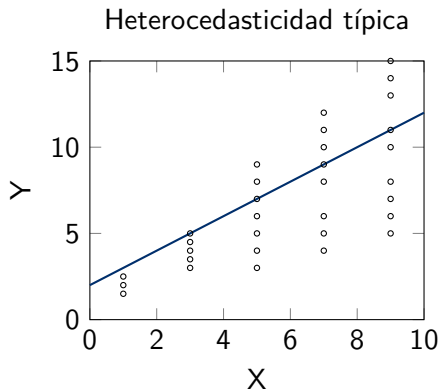
$$\text{Var}(\varepsilon_i|X) = \sigma^2 \quad \forall i$$

Varianza constante para todas las observaciones.

**Heterocedasticidad** (violación):

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

Varianza cambia con  $X$ .



# ¿Por qué es Común en Datos Salariales?

## **Razon económica:**

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



# ¿Por qué es Común en Datos Salariales?

## Razon económica:

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

## Consecuencias de ignorar heterocedasticidad

- $\hat{\beta}$  sigue siendo **insesgado** y **consistente**
- Pero  $\text{Var}(\hat{\beta})$  está **mal calculada**
- Errores estándar incorrectos  $\rightarrow$  inferencia inválida
- Tests t y F son **inválidos**

## 1. Gráfico de residuos vs valores ajustados:

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

## 1. Gráfico 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. Gráfico 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 específica
```

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

\* Forma clásica (incorrecta si hay heterocedasticidad)

```
reg ln_salario escolaridad experiencia experiencia2
```

\* Forma robusta (válida con heterocedasticidad)

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

# Solución: Errores Estandar Robustos

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

\* Forma clásica (incorrecta si hay heterocedasticidad)

```
reg ln_salario escolaridad experiencia experiencia2
```

\* Forma robusta (válida con heterocedasticidad)

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

## ¿Qué hace robust?

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

## Regla práctica

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>	Correlación dentro de grupos
HAC	<code>newey</code>	Datos de series de tiempo
Bootstrap	<code>, vce(bootstrap)</code>	Muestras pequeñas

# Variantes de Errores Robustos

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

## ¿Cuándo usar cluster?

Si los errores están 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)
```



# ¿Qué es Multicolinealidad?

## Definicion

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

# ¿Qué es Multicolinealidad?

## Definición

Existe **multicolinealidad** cuando las variables independientes están altamente correlacionadas entre sí.

## Ejemplos en ecuación de Mincer:

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

# ¿Qué es Multicolinealidad?

## Definición

Existe **multicolinealidad** cuando las variables independientes están altamente correlacionadas entre sí.

## Ejemplos en ecuación de Mincer:

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

## Consecuencias

- $\hat{\beta}$  sigue siendo insesgado
- Pero  $\text{Var}(\hat{\beta})$  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	Interpretación
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 identificacion



## 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

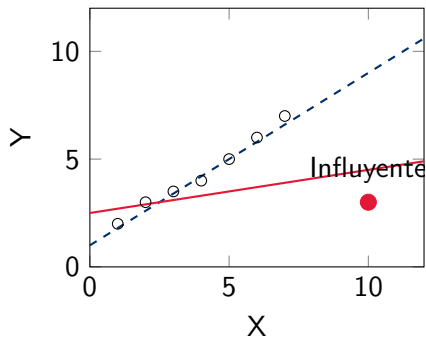
# Outliers vs Observaciones Influyentes

## Outlier:

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

## 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  $X_i$ .

# Diagnósticos de Influencia

## 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  $X_i$ .

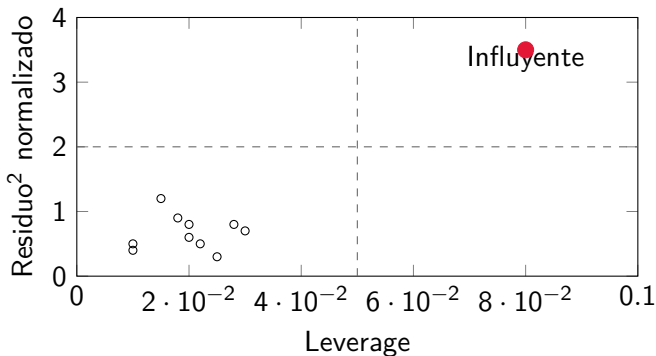
## 3. Distancia de Cook:

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

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

## Gráfico: Leverage vs Residuo

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



# ¿Qué Hacer con Observaciones Influyentes?

## ① Verificar si son errores de datos

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

# ¿Qué 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



# ¿Qué 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 métodos robustos

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

# ¿Qué 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 métodos robustos

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

## ④ Winsorizar o truncar

- Reemplazar valores extremos por percentiles

## Tipos de errores:

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

## Tipos de errores:

- **Variables omitidas:** Falta una variable relevante
- **Forma funcional incorrecta:** Debería ser cuadrático, 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 términos no lineales

# Test RESET en Stata

```
* Modelo sin término cuadrático
reg ln_salario escolaridad experiencia, robust
estat ovtest

* Modelo con término cuadrático
reg ln_salario escolaridad experiencia experiencia2, robust
estat ovtest
```

## Interpretación:

- $p\text{-valor} < 0.05$ : Rechazar  $H_0$ , hay problema de especificación
- $p\text{-valor} \geq 0.05$ : No rechazar, no hay evidencia de mala especificación

# Test RESET en Stata

```
* Modelo sin término cuadrático
reg ln_salario escolaridad experiencia, robust
estat ovtest

* Modelo con término cuadrático
reg ln_salario escolaridad experiencia experiencia2, robust
estat ovtest
```

## Interpretación:

- $p\text{-valor} < 0.05$ : Rechazar  $H_0$ , hay problema de especificación
- $p\text{-valor} \geq 0.05$ : No rechazar, no hay evidencia de mala especificación

## Nota

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

# Checklist de Diagnósticos

Problema	Test	Comando Stata	Solución
Heterocedasticidad	Breusch-Pagan	<code>estat hettest</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
Especificación	RESET	<code>estat ovtest</code>	Agregar términos
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. Diagnósticos
estat hettest           // Heterocedasticidad
vif                     // Multicolinealidad
estat ovtest            // Especificación
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?

Próxima Sesión:

## **M04: Descomposición Oaxaca-Blinder**

Lunes 16 de febrero, 3-5pm

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