SEMINARIO DE DOCTORADO

TÉCNICAS AVANZADAS DE INVESTIGACIÓN SOCIAL

Módulo 3B:
MODELOS DE REGRESIÓN
MODELOS NO LINEALES
INTEPRETACIÓN DE INTERACCIONES

Agustín Salvia
Santiago Poy

Modelos de Regresión Lineal Múltiple

Problemas de determinación

- **El investigador suele tener razones teóricas o prácticas para creer que los valores de una variable dependen del comportamiento de una o más variables distintas.**
- **Si hay suficientes observaciones empíricas sobre estas variables, el análisis de regresión es un método apropiado para predecir el comportamiento de la variable dependiente y describir la estructura, fuerza y sentido exacto de la relación.**

CORRELACIÓN ENTRE VARIABLES CUANTITATIVAS

- **El ajuste mide el nivel en que los pares de observaciones quedan representados en una línea en donde a cada valor de A le corresponde un valor en B. Si la nube de observaciones es estrecha y alargada, una línea recta representará a la nube de puntos y a la relación y por tanto ésta será fuerte.**

- **El sentido de la relación se refiere a cómo varían los valores de *B* con respecto a *A*. Si al crecer los valores de la variable *A* lo hacen los de *B*, será una relación positiva o directa. Si al aumentar *A*, disminuye *B*, será una relación negativa o inversa.**

- **La forma establece el tipo de línea a emplear para definir el mejor ajuste. Se pueden emplear tres tipos de líneas: una línea recta, una curva monotónica o una curva no monotónica.**

- **La fuerza es la pendiente de la recta. En cuantas unidades aumenta, o disminuye, la variable A al aumentar en una unidad la variable B.**

Modelos de Regresión Lineal

Problemas de Causalidad

Una pregunta importante que se plantea en el análisis de regresión es la siguiente: ¿Qué parte de la variación total en *Y* se debe a la variación en *X*? ¿Cuánto de la variación de Y no se explica por X?

- **El modelo permite diferenciar variables explicativas, independientes o predictivas (métricas), variables a explicar o dependientes, y variables control (métricas o transformadas en variables categoriales *dummy*).**
- **La distinción entre variables dependientes e independientes debe efectuarse con arreglo a fundamentos teóricos, por conocimiento o experiencia y estudios anteriores.**

Métodos de tipo: $Y : f(X, \epsilon) / Y = a + bX + e$

Construcción de variables dummy

- **Se utilizan para ingresar variables cualitativas con un rol de predictoras en un modelo de regresión.**
- **Consiste en la generación de tantas variables dicotómicas como categorías menos uno tenga la variable original**

Ejemplo: variable original: Sexo

- **Cantidad de categorías: 2 (Varón / Mujer)**
- **Necesidad de generar una variable dummy (N-1 categorías)**
- **Valores que asume la dummy: Si es varón = 0 / Si es mujer = 1.**

Modelos de Regresión Lineal

Respuestas Metodológicas

- **Estima el grado de ajuste o de bondad explicativa del modelo teórico independientemente del nivel de covarianza entre las variables introducidas**
- **Predice el valor medio que puede asumir la variable Y dado un valor de X (regresión a la media) bajo un intervalo de confianza**
- **Estima el efecto neto / fuerza / sentido de cada una de las variables predictoras de la variable dependiente (control sobre los demás efectos suponiendo independencia entre ellas).**

Modelos de Regresión Lineal

Salidas Estadísticas del Método

- **Se evalúa la bondad de ajuste del modelo teórico a a través del coeficiente de determinación R^2**
- **La capacidad explicativa del modelo se hace a partir del método de mínimos cuadrados (ANOVA), cuyo resultado es testeado a través de F de Fisher**
- **Predice los valores de la variable dependiente a partir de estimar el valor del coeficiente (B), el error estándar (S) y el coeficiente R parcial (BETA) de cada una de las variables y de la Constante**
- **Mide la fuerza, sentido y significancia estadística de las variables del modelo sobre la variable dependiente a través de la prueba t de Student**

Modelos de Regresión Lineal

Control de Supuestos

- **LINEALIDAD: a través de gráficas de dispersión simple y parciales. Transformación de las variables hasta lograr el mejor ajuste (mayor bondad de ajuste R2).**
- **NORMALIDAD DE LOS RESIDUOS: a través de un gráfico de de distribución de los residuos tipificados. Solución: eliminación de datos outliers.**
- **HETEROSCEDASTICIDAD: a través de gráficos de residuos ϵ para cada valor de $\hat{y}$. Solución: Eliminación de casos outliers, transformación de las variables independientes y/o estandarización de la variable dependiente *Y*.**
- **AUTOCORRELACIÓN DE ERRORES: a través de la prueba Durbin-Watson / el valor 2 indica no autocorrelación. Solución: Corrección de observaciones o eliminación de datos.**
- **MULTICOLINEALIDAD: a través de matrices de correlación simple entre las variables independientes, análisis de Tolerancia y Diagnóstico de Colinealidad. Solución: Seleccionar variables independiente con baja correlación entre sí y/o transformar en variables dummy no colineales.**

Linealidad

- Se obtiene del plot de los *valores predichos* versus la *variable independiente*. Si la relación no es lineal, la dispersión (scatter) de los puntos mostrará una desviación sistemática de la línea de regresión.
- Con el modelo de la regresión múltiple es mejor generar un gráfico simple (plot) de los *valores observados* versus los *valores predichos*. Teóricamente, en un gráfico de *observados* vs. *predichos* los puntos deberían moverse entre torno a la *línea recta diagonal*.
- El gráfico de valores residuales vs. valores predichos es esencialmente el mismo que el anterior, a excepción de que la línea de referencia es horizontal más que de 45 grados.

Linealidad

Caso donde se cumple el supuesto:

Figura: Gráfico de y vs x

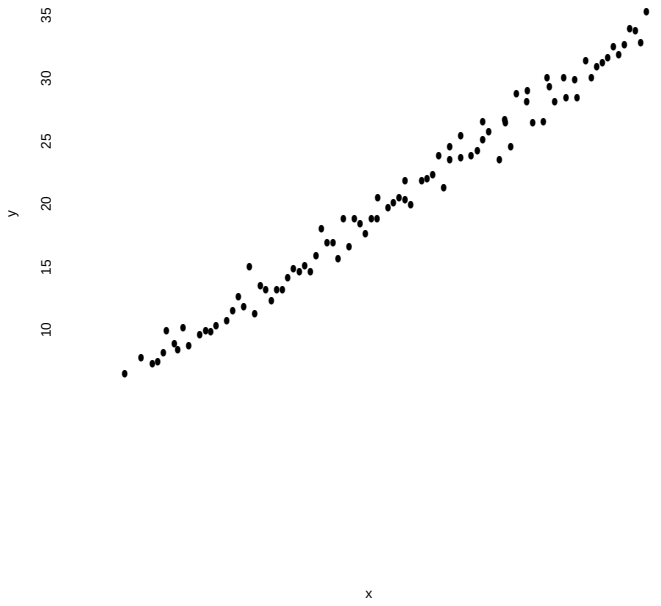
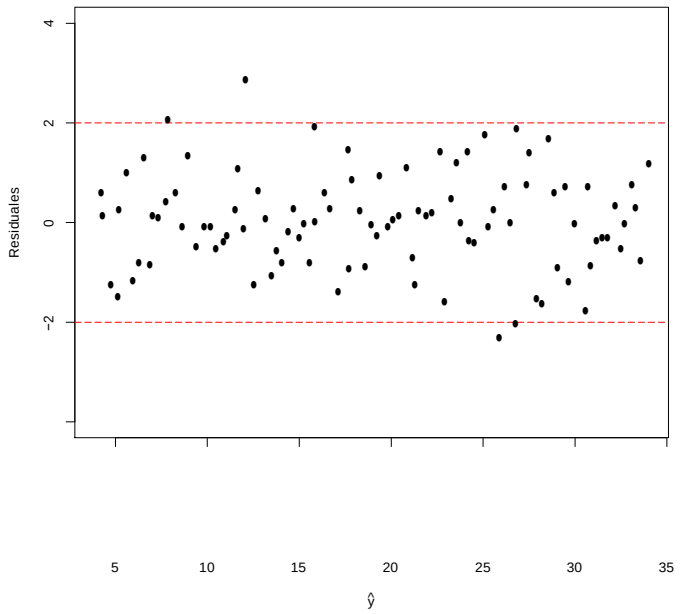


Figura: Gráfico de residuales vs y ajustados



Linealidad

Caso donde no se cumple el supuesto:

Figura: Gráfico de y vs x

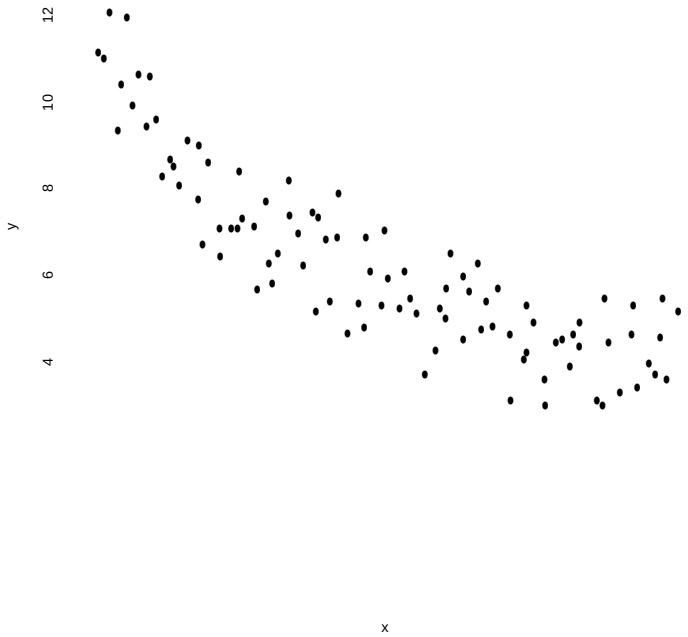
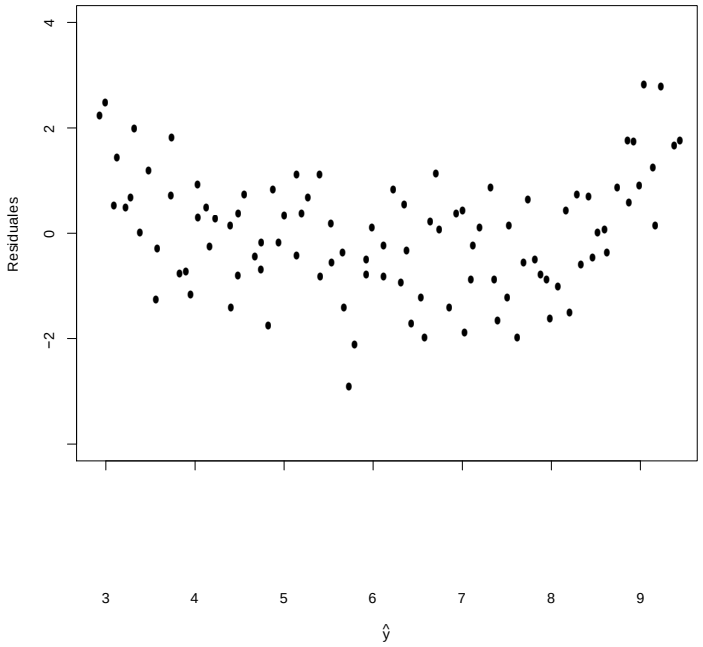


Figura: Gráfico de residuales vs y ajustados



AJUSTE DE VARIABLES A FUNCIONES NO LINEALES

- **Identificada dicha función, substituir los valores de una variable con sus valores cuadrados, logarítmicos o con alguna otra modificación, y hacer de nuevo la matriz de correlación.**

- **Proceder a análisis segmentados de la población explorando los diferentes comportamientos (p.e. Regresión de educación sobre ingresos en los jóvenes y Regresión de educación sobre ingresos en los mayores.**

Prueba de normalidad

- Mediante el histograma de los residuos tipificados. La curva se construye con media 0 y un desviación típica de 1.
- Mediante el gráfico de probabilidad normal (en el eje de las abscisas se representa la probabilidad acumulada de cada residuo y en de las ordenadas la probabilidad acumulada pronosticada)

Caso donde se cumple el supuesto:

Figura: Histograma de los residuales

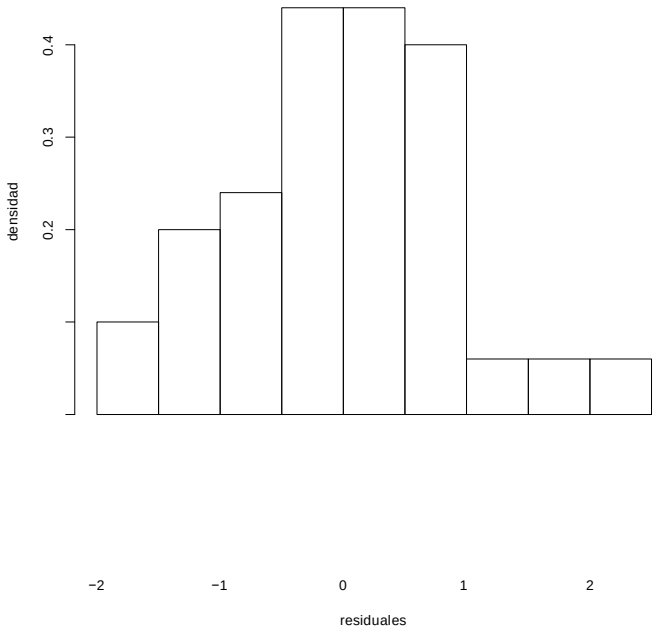
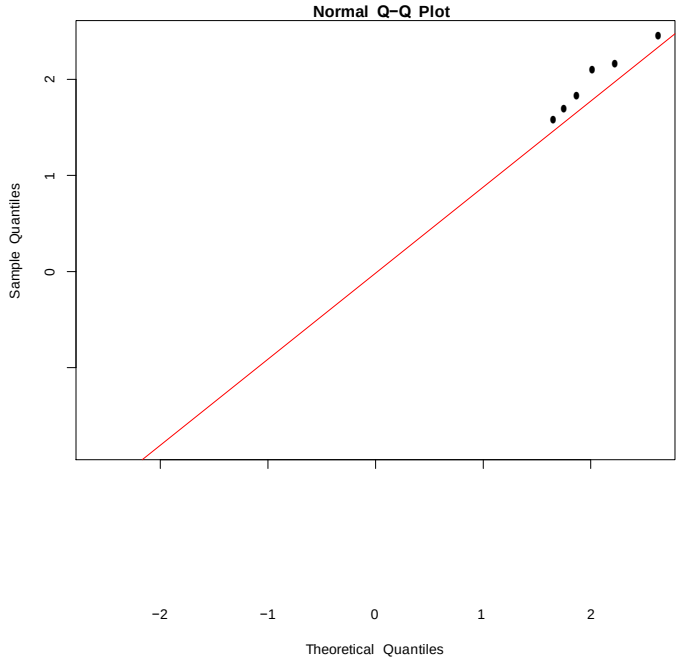


Figura: qq-plot de los residuales



Caso donde no se cumple el supuesto:

Figura: Histograma de los residuales

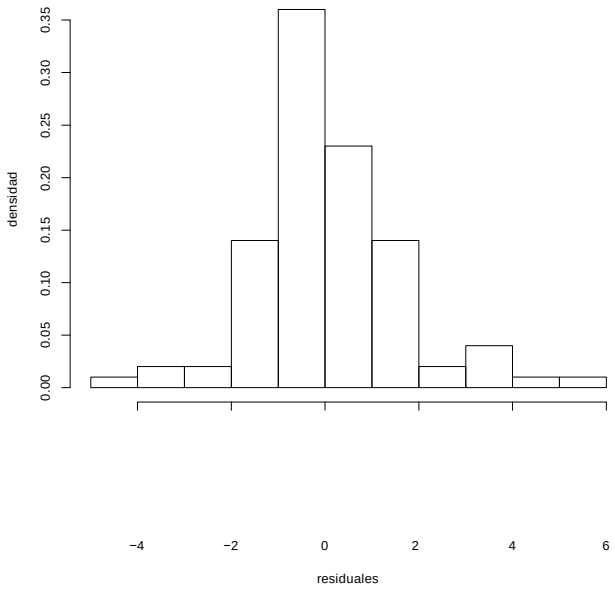
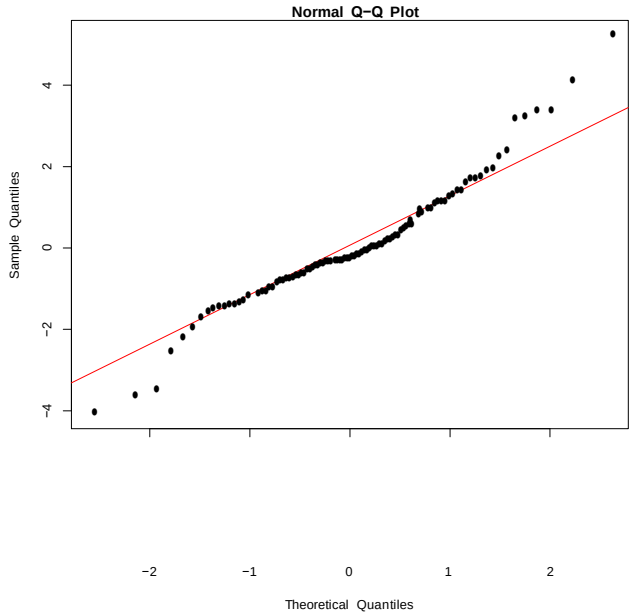


Figura: qq-plot de los residuales



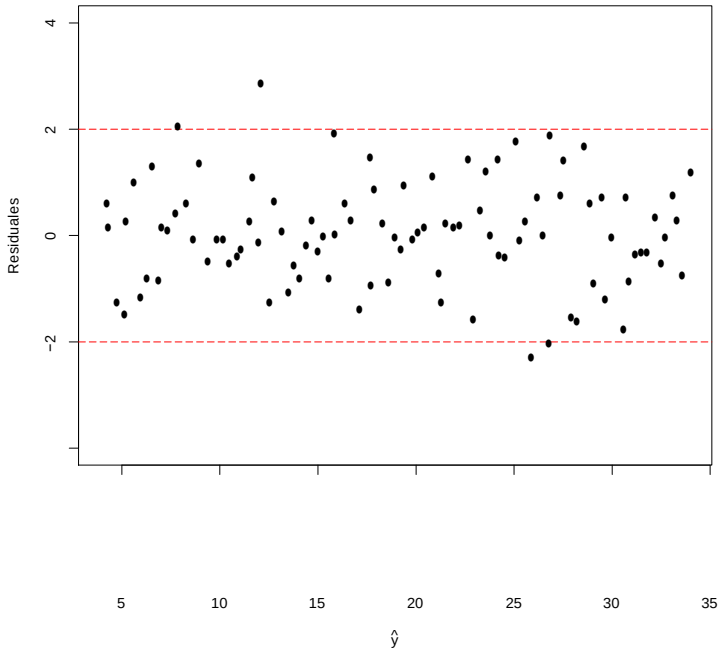
Homoscedasticidad

- En el cuadro de diálogo de gráficos se obtienen una serie de variables listadas para obtener diferentes gráficos de dispersión, por ejemplo: Los valores ZRESID se trasladan al eje Y y los valores ZPRED al eje X.
- La variación de los residuos debe ser uniforme en todo el rango de valores pronosticados; es decir, el tamaño de los residuos es independiente del tamaño de los pronósticos.

Homoscedasticidad

- El gráfico de dispersión no debe mostrar ninguna pauta de asociación entre los residuos tipificados y los valores pronosticados

Figura: Gráfico de residuales vs *y* ajustados



x

Homoscedasticidad

Caso donde no se cumple el supuesto:

Figura: Gráfico de y vs x

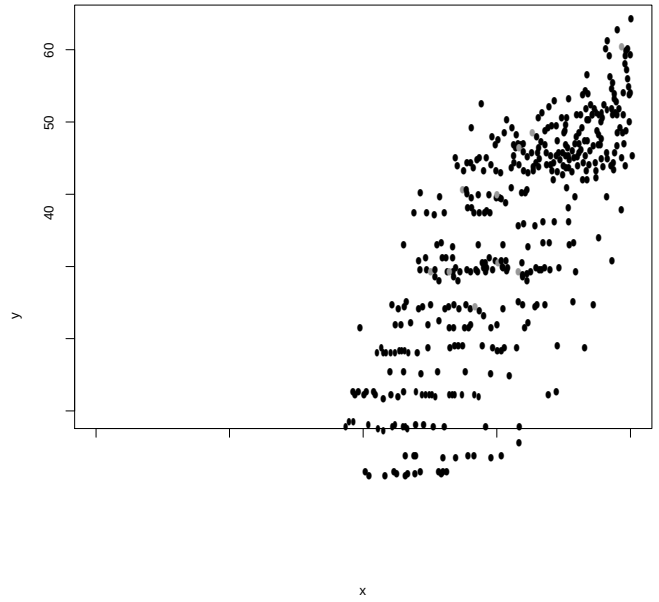
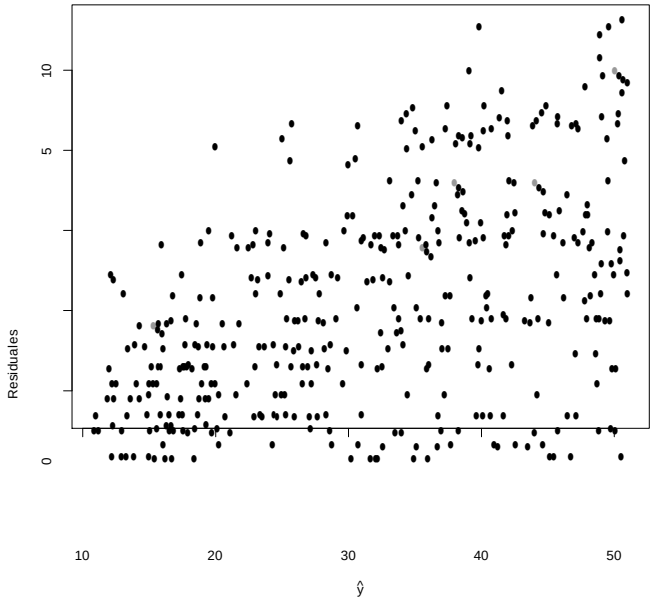


Figura: Gráfico de residuales vs y ajustados



Independencia de los residuos

- Uno de los supuestos básicos del MRL (modelos de la regresión lineal) es la independencia entre los residuos. El estadístico de *Durbin-Watson* aporta información sobre el grado de independencia existente entre ellos.
- El estadístico de *Durbin-Watson* (DW) proporciona información sobre el grado de independencia entre los residuales. El estadístico DW varía entre 0 y 4, y toma el valor 2 cuando los residuales son independientes.
- Valores menores que 2 indica autocorrelación positiva. Podemos asumir independencia entre los residuales cuando DW toma valores entre 1.5 y 2.5

Colinealidad

Estadísticos de colinealidad

Tolerancia y VIF (variancia inflation factors)

- Tolerancia: Una primera medida para para probar la colinealidad o no dependencia lineal entre los regresores ($T_p = 1 – R_p^{2)}$.
- Cuando tiene un valor máximo de 1, la variable no tiene ningún grado de colinealidad con las restantes, Un valor 0 indica que la variable es una combinación lineal perfecta de otros regresores. Es deseable que, en general, sea mayor a .40

Diagnóstico de Colinealidad

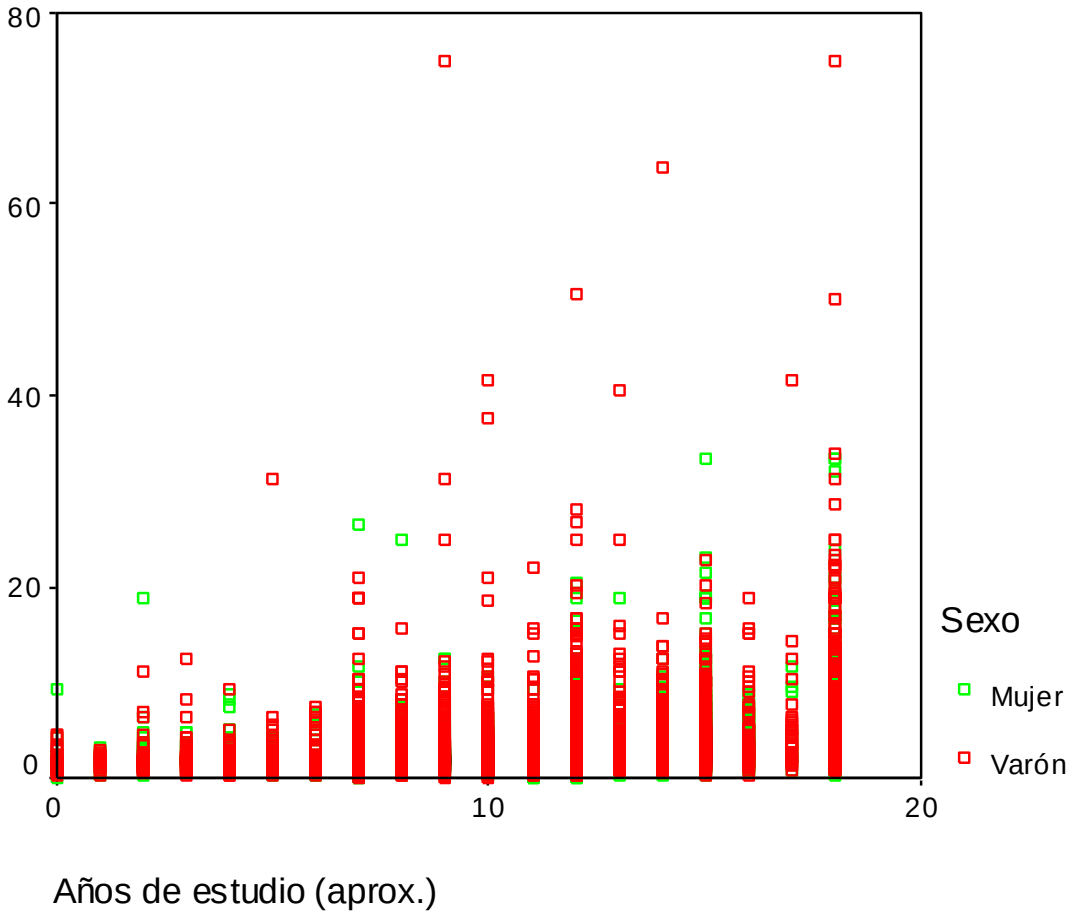
- Dimensiones: factores diferentes que se hallan en el conjunto de variables independientes.
- Autovalores: los valores próximos a 0 indican colinealidad.
- Índices de condición: raíz cuadrada (autovalormayor/autovalor). Valores por encima de 15 indican posibles problemas de colinealidad
- Proporciones de variancia: proporción de la variancia de cada coeficiente de la regresión parcial b_j que está explicada por cada factor.
- Proporciones de variancia: Hay problema de colinealidad si una dimensión (de índice de condición alto) explica gran cantidad de la varianza de dos o más variables.

- VIF (variance inflation factor): a medida que es mayor la multicolinealidad, en un de los regresores, la variancia de su coeficiente comienza a crecer. La multicolinealidad infla la variancia del coeficiente ($VIF_p = 1/(1-R_{xp}^2)$).
- La VIF tomará un valor mínimo de 1 cuando no hay colinealidad y no tendrá límite superior en el caso de multicolinealidad.
- En presencia de multicolinealidad, una solución lógica consiste en eliminar del modelo aquellas variables con más alto VIF (o más baja tolerancia).

Modelos de Regresión Lineal

ANÁLISIS DE UN EJEMPLO

El ingreso horario de los ocupados (entre 25 y 45 años) no se ve afectados por el sexo sino que depende de la cantidad de años de instrucción



Modelos de Regresión Lineal

ANÁLISIS DE UN EJEMPLO

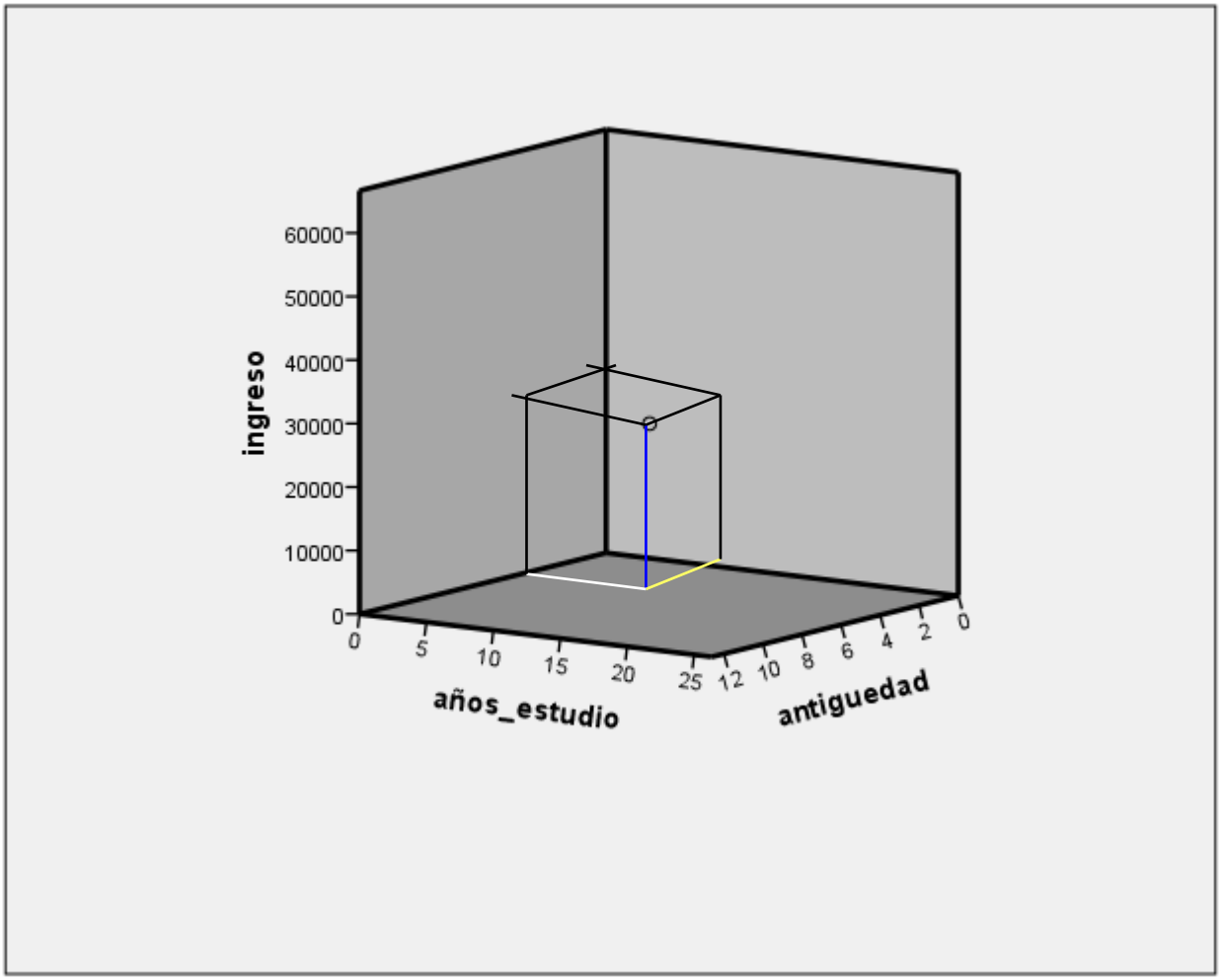
❑ CORRELACIÓN DE PEARSON

Correlations

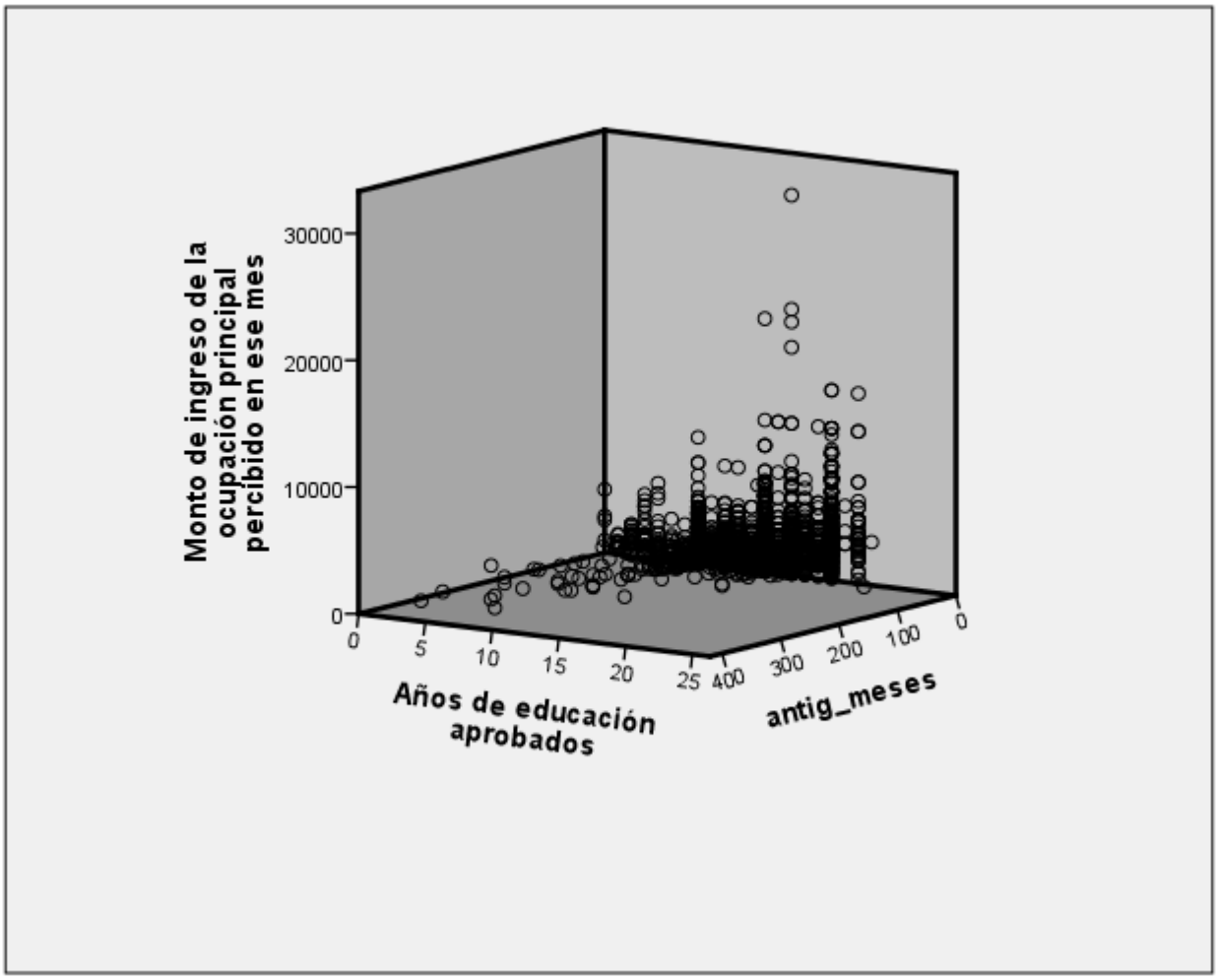
		Ingreso horario de la ocupación ppal	Sexo (dummy: 0=Varón)	Años de estudio (aprox.)
Ingreso horario de la ocupación ppal	Pearson Correlation	1,000	-,014	,354**
	Sig. (2-tailed)	,	,149	,000
	N	10339	10339	10338
Sexo (dummy: 0=Varón)	Pearson Correlation	-,014	1,000	,137**
	Sig. (2-tailed)	,149	,	,000
	N	10339	10339	10338
Años de estudio (aprox.)	Pearson Correlation	,354**	,137**	1,000
	Sig. (2-tailed)	,000	,000	,
	N	10338	10338	10338

**. Correlation is significant at the 0.01 level (2-tailed).

Identificación de un punto en el espacio según datos de tres variables



Identificación de nube de puntos en el espacio según datos de tres variables



Modelos de Regresión Lineal

ANÁLISIS DE UN EJEMPLO

❑ BONDAD DE AJUSTE DEL MODELO (R^2)

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	Sexo (dummy: 0=Varón)[a]	,	Enter
2	Años de estudio (aprox.)[a]	,	Enter

a. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,014[a]	,000	,000	3,3032
2	,359[b]	,129	,129	3,0832

a. Predictors: (Constant), Sexo (dummy: 0=Varón)

b. Predictors: (Constant), Sexo (dummy: 0=Varón), Años de estudio (aprox.)

Modelos de Regresión Lineal

ANÁLISIS DE UN EJEMPLO

- **ANÁLISIS DE VARIANZA DE LOS MODELOS**

ANOVA[c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	22,486	1	22,486	2,061	,151[a]
	Residual	112779,9	10336	10,911		
	Total	112802,4	10337			
2	Regression	14557,248	2	7278,624	765,683	,000[b]
	Residual	98245,112	10335	9,506		
	Total	112802,4	10337			

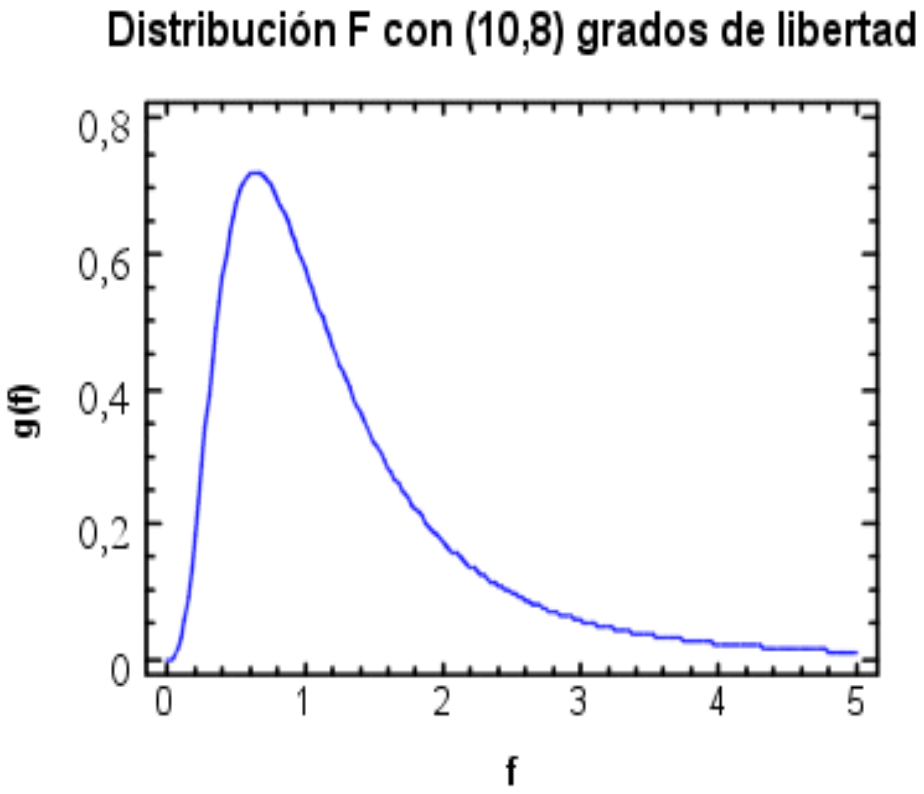
a. Predictors: (Constant), Sexo (dummy: 0=Varón)

b. Predictors: (Constant), Sexo (dummy: 0=Varón), Años de estudio (aprox.)

c. Dependent Variable: Ingreso horario de la ocupación ppal

Modelos de Regresión Lineal

Distribución F de Fisher-Snedecor



- Nunca adopta valores menores de 0 y es asimétrica positiva. En el modelo de regresión representa la relación entre el total de la varianza de la variable dependiente y la parte explicada de dicha varianza.

- Es una familia de curvas, en función de los llamados "grados de libertad" del numerador y del denominador. La distribución F equivale a una razón entre dos chi-cuadrados (de ahí que se hable en el caso de F de grados de libertad en el numerador y en el denominador)

Modelos de Regresión Lineal

ANÁLISIS DE UN EJEMPLO

- **COEFICIENTES B Y PRUEBAS T DE SIGNIFICANCIA**

Coefficients[a]

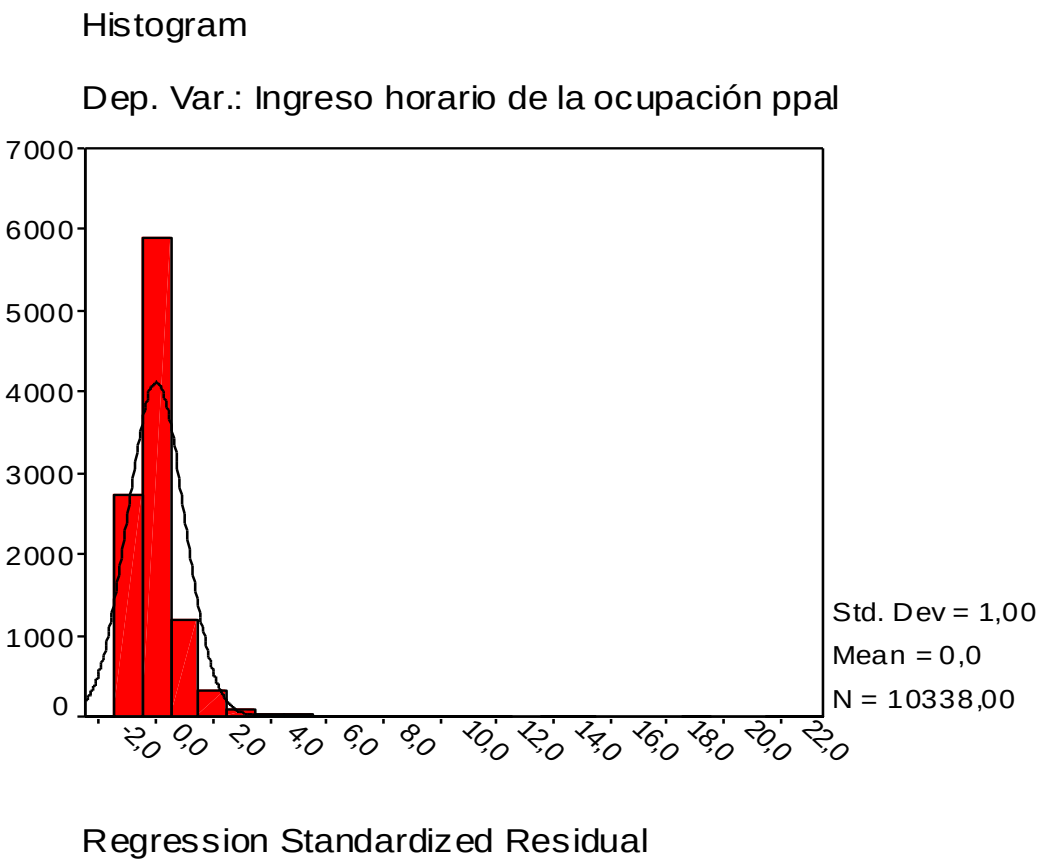
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3,476	,043		80,455	,000
	Sexo (dummy: 0=Varón)	-,0941	,066	-,014	-1,436	,151
2	(Constant)	,271	,091		2,964	,003
	Sexo (dummy: 0=Varón)	-,426	,062	-,064	-6,898	,000
	Años de estudio (aprox.)	,306	,008	,362	39,102	,000

a. Dependent Variable: Ingreso horario de la ocupación ppal

Modelos de Regresión Lineal

ANÁLISIS DE UN EJEMPLO

❑ PRUEBAS DE NORMALIDAD DE LOS RESIDUOS



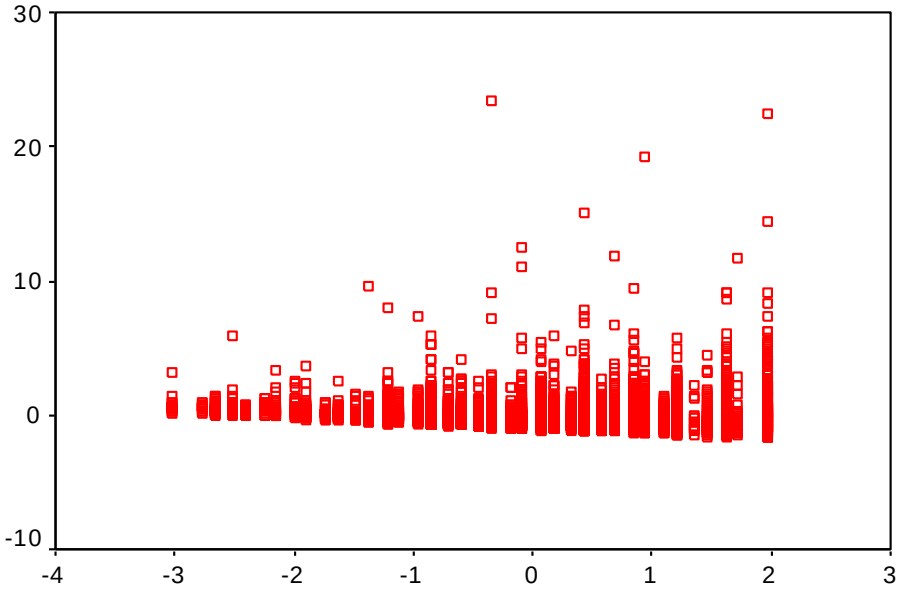
Modelos de Regresión Lineal

ANÁLISIS DE UN EJEMPLO

❑ PRUEBAS DE HETEROSCEDASTICIDAD

Scatterplot

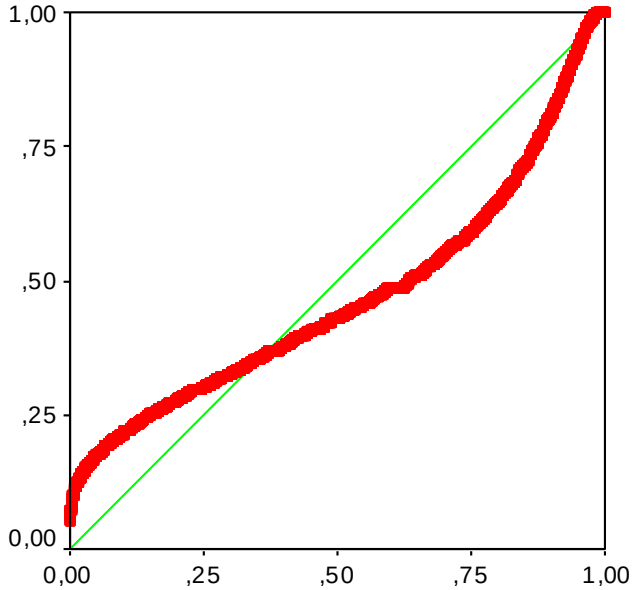
Dependent Variable: Ingreso horario de la ocupación ppal



Regression Standardized Predicted Value

Normal P-P Plot of Regression Standardized Res.

Dep. Var.: Ingreso horario de la ocupación ppal



Observed Cum Prob

Modelos de Regresión Lineal

ANÁLISIS DE UN EJEMPLO

- **DURBIN WATSON: EVALUACIÓN DE AUTOCORRELACIÓN**

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,359[a]	,129	,129	3,0832	1,707

a. Predictors: (Constant), Sexo (dummy: 0=Varón), Años de estudio (aprox.)

b. Dependent Variable: Ingreso horario de la ocupación ppal

Modelos de Regresión Lineal

¿QUÉ HACER FRENTE A LOS SESGOS DE ESTIMACIÓN?

- **Eliminar casos OUTLIERS que afectan la distribución.**
- **Recodificación de las variables independientes y/o transformación LOGÍSTICA de la variable dependiente.**
- **Estratificación del análisis a partir de usar una variable independiente como CRITERIO PARA DIVIDIR a la población en grupos comparables.**

Modelos de Regresión Lineal

ANÁLISIS DE UN EJEMPLO

❑ CORRELACIÓN SIMPLE

Correlations

		L_INGHOR	Sexo (dummy 1-Varón)	Años de estudio (aprox.)
L_INGHOR	Pearson Correlation	1,000	-,021*	,421**
	Sig. (2-tailed)	,	,031	,000
	N	10339	10339	10338
Sexo (dummy 1-Varón)	Pearson Correlation	-,021*	1,000	-,137**
	Sig. (2-tailed)	,031	,	,000
	N	10339	10339	10338
Años de estudio (aprox.)	Pearson Correlation	,421**	-,137**	1,000
	Sig. (2-tailed)	,000	,000	,
	N	10338	10338	10338

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

Modelos de Regresión Lineal

ANÁLISIS DE EJEMPLOS

❑ BONDAD DE AJUSTE DE LOS MODELOS (R^2)

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,359[a]	,129	,129	3,0832	1,707

a. Predictors: (Constant), Años de estudio (aprox.), Sexo (dummy 1-Varón)

b. Dependent Variable: Ingreso horario de la ocupación ppal

Modelo Original

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,401[a]	,161	,160	2,5866	1,675

a. Predictors: (Constant), Años de estudio (aprox.), Sexo (dummy 1-Varón)

b. Dependent Variable: Ingreso horario de la ocupación ppal

Excluyendo desvíos mayores a 8z

Modelos de Regresión Lineal

ANÁLISIS DE UN EJEMPLO

- **BONDAD DE AJUSTE DEL MODELO (R^2)**

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,021[a]	,000	,000	,7307	
2	,422[b]	,178	,178	,6625	1,622

a. Predictors: (Constant), Sexo (dummy 1-Varón)

b. Predictors: (Constant), Sexo (dummy 1-Varón), Años de estudio (aprox.)

c. Dependent Variable: L_INGHOR

Variable dependiente logaritmo ing. horario

Modelos de Regresión Lineal

ANÁLISIS DE UN EJEMPLO

❑ ANÁLISIS DE VARIANZA DE LOS MODELOS

ANOVA[c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2,504	1	2,504	4,689	,030[a]
	Residual	5518,817	10336	,534		
	Total	5521,321	10337			
2	Regression	985,393	2	492,696	1122,596	,000[b]
	Residual	4535,928	10335	,439		
	Total	5521,321	10337			

a. Predictors: (Constant), Sexo (dummy 1-Varón)

b. Predictors: (Constant), Sexo (dummy 1-Varón), Años de estudio (aprox.)

c. Dependent Variable: L_INGHOR

Modelos de Regresión Lineal

ANÁLISIS DE UN EJEMPLO

- **COEFICIENTES B Y PRUEBAS T DE SIGNIFICANCIA**

Coefficients[a]

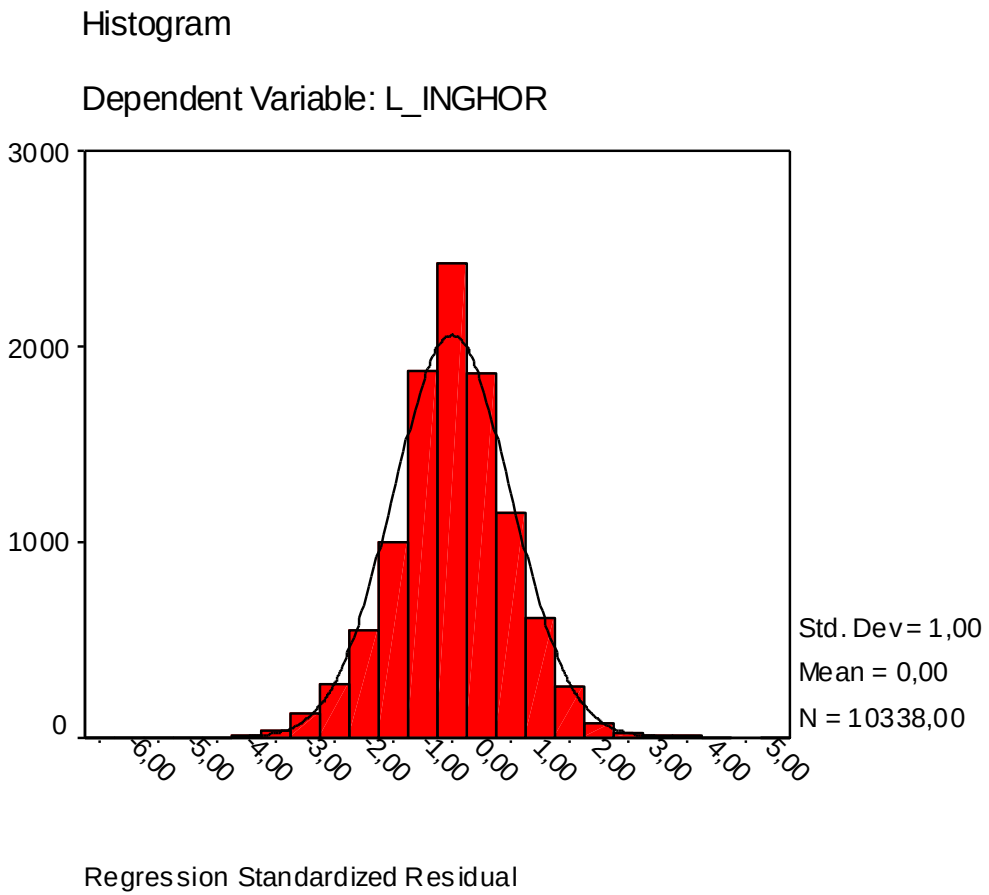
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	,976	,011		89,504	,000
	Sexo (dummy 1-Varón)	-,0314	,014	-,021	-2,165	,030
2	(Constant)	,0557	,022		2,553	,011
	Sexo (dummy 1-Varón)	,0549	,013	,037	4,139	,000
	Años de estudio (aprox.)	,0796	,002	,426	47,323	,000

a. Dependent Variable: L_INGHOR

Modelos de Regresión Lineal

ANÁLISIS DE UN EJEMPLO

- **GRAFICAS DE DISPERSIÓN DE RESIDUOS**



Modelos de Regresión Lineal

ANÁLISIS DE UN EJEMPLO

- **DURBIN WATSON: EVALUACIÓN DE AUTOCORRELACIÓN**

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,021[a]	,000	,000	,7307	
2	,422[b]	,178	,178	,6625	1,622

a. Predictors: (Constant), Sexo (dummy 1-Varón)

b. Predictors: (Constant), Sexo (dummy 1-Varón), Años de estudio (aprox.)

c. Dependent Variable: L_INGHOR

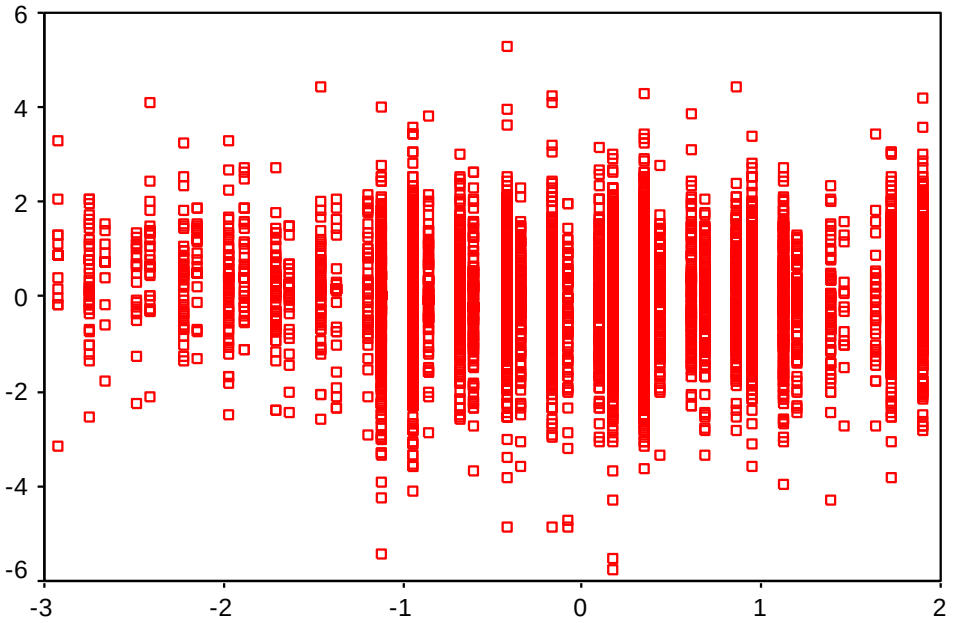
Modelos de Regresión Lineal

ANÁLISIS DE UN EJEMPLO

❑ PRUEBAS DE HETEROSCEDASTICIDAD

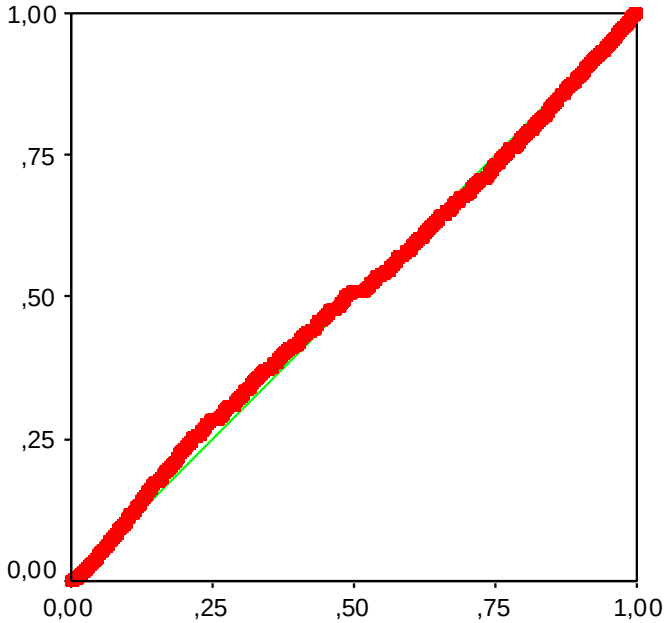
Scatterplot

Dependent Variable: L_INGHOR



Normal P-P Plot of Regression Standardized Residual

Dependent Variable: L_INGHOR



REGRESIONES CON EFECTOS DE INTERACCIÓN

SI ASUMIMOS A TRAVÉS DE LOS TEST DE TOLERANCIA Y ANÁLISIS DE CORRELACIÓN PARCIAL QUE LA RELACIÓN XY ESTÁ MODULADA/CONDICIONADA POR Z, EL EFECTO DE X SOBRE Y NO ES EL MISMO DEPENDIENDO DE LOS VALORES DE Z. EN LOS CASOS EN QUE EL EFECTO DE X-Y DEPENDE DEL VALOR O NIVEL DE Z, DEBE CONTROLARSE EL EFECTO INTERACCIÓN XZ

1) **CREAR LA VARIABLE INTERACCIÓN COMO PRODUCTO XZ (ASEGURARSE DE QUE LA VARIABLE ES MÉTRICA O DUMMY).**
2) **EVALUACIÓN EL PESO/SIGNIFICANCIA DE LOS CAMBIOS EN LOS COEFICIENTES PRINCIPALES Y EL COEFICIENTE INTERACCIÓN.**
3) **LECTURA / INTERPRETACIÓN DE LOS EFECTOS DEPENDIENDO DEL DISEÑO.**

En caso de COLINEALIDAD, debe evaluarse el efecto diferencial que la variable Z ejerce en la relación de X con Y. A este respecto, podemos calcular el efecto de X sobre Y para los distintos valores de Z. Para ello, reestructuramos la ecuación de regresión de la siguiente forma:

$$\mathbf{Y = (b0) + (b1x) + (b2z) + (b3zx)}$$

$$Y = (b0) + (b1*X) + (b2*Z) + (b1*X)\,(b2*Z)$$

DIFERENTES MODELOS:

X Y Z AMBAS SON DUMMY

X ES METRICA Y Z ES DUMMY

X Y Z AMBAS SON MÉTRICAS

REGRESIONES CON EFECTOS DE INTERACCIÓN

SI ASUMIMOS POR LOS TEST DE TOLERANCIA Y ANÁLISIS DE CORRELACIÓN PARCIAL QUE LA RELACIÓN XY ESTÁ MODULADA/CONDICIONADA POR Z, ESTO ES, EL EFECTO DE X SOBRE Y NO ES EL MISMO DEPENDIENDO DE LOS VALORES DE Z, SINO QUE TAL EFECTO DEPENDE DEL VALOR O NIVEL DE Z, DEBEMOS CONTROLAR EL EFECTO INTERACCIÓN XZ

1) **CREAR LA VARIABLE INTERACCIÓN COMO PRODUCTO XZ (ASEGURARSE DE QUE LA VARIABLE ES MÉTRICA O DUMMY).**
2) **EVALUAR EL CAMBIO EN LA BONDAD DE AJUSTE / ANOVA DE LOS MODELOS (R2) CON Y SIN EFECTO INTERACCIÓN.**
3) **EVALUACIÓN EL PESO/SIGNIFICANCIA DE LOS CAMBIOS EN LOS COEFICIENTES PRINCIPALES Y EL COEFICIENTE INTERACCIÓN.**
4) **LECTURA / INTERPRETACIÓN DE LOS EFECTOS DEPENDIENDO DEL DISEÑO.**

En caso de COLINEALIDAD, debe evaluarse el efecto diferencial que la variable Z ejerce en la relación de X con Y. A este respecto, podemos calcular el efecto de X sobre Y para los distintos valores de Z. Para ello, reestructuramos la ecuación de regresión de la siguiente forma:

$$\mathbf{Y = (b0) + (b1x) + (b2z) + (b3zx)}$$

$$Y = (b0) + (b1*X) + (b2*Z) + (b1*X)\,(b2*Z)$$

DIFERENTES MODELOS:

X Y Z AMBAS SON DUMMY

X ES METRICA Y Z ES DUMMY

X Y Z AMBAS SON MÉTRICAS