Tema 3: Modelos generales de regresión

TEMA 1: MODELO DE REGRESIÓN LINEAL SIMPLE

Problema: Analizar la relación entre una variable numérica independiente (predictora), X, y una variable numérica de respuesta Y y expresar esta relación mediante una función matemática.

Ojetivos:

- Describir la relación entre las dos variables
- Predecir una respuesta Y a partir de los valores de X
- Cuantificar la contribución o efecto de la variable predictora sobre la variable de respuesta

Condiciones del modelo:

- Modelo lineal

$$Y = \alpha + \beta X + \mathcal{E}, \text{ donde } \mathcal{E} \text{ es la variable error o residual}$$

- Normalidad

$$\mathcal{E} \sim N(0, \sigma^2) \qquad \text{es decir} \quad Y \,|\, X = x_i \sim N(\alpha + \beta x_i, \sigma^2)$$

- Parámetros a estimar: α, β y σ^2

I. Descripción de la relación entre ambas variables

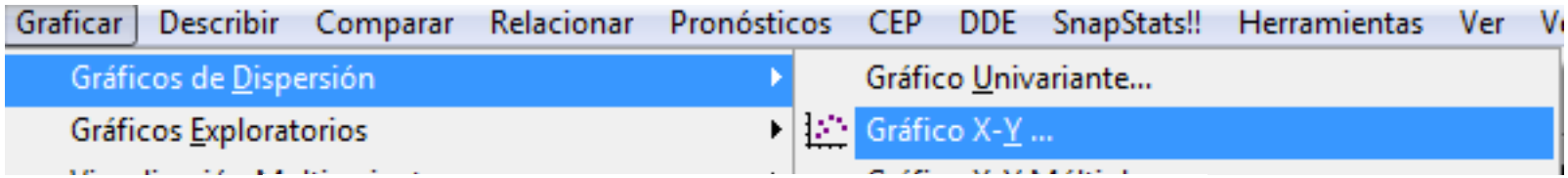
1. **Introducir los datos en dos columnas separadas en StatGraphics.**
2. **Representar gráficamente la nube de puntos o diagrama de dispersión**

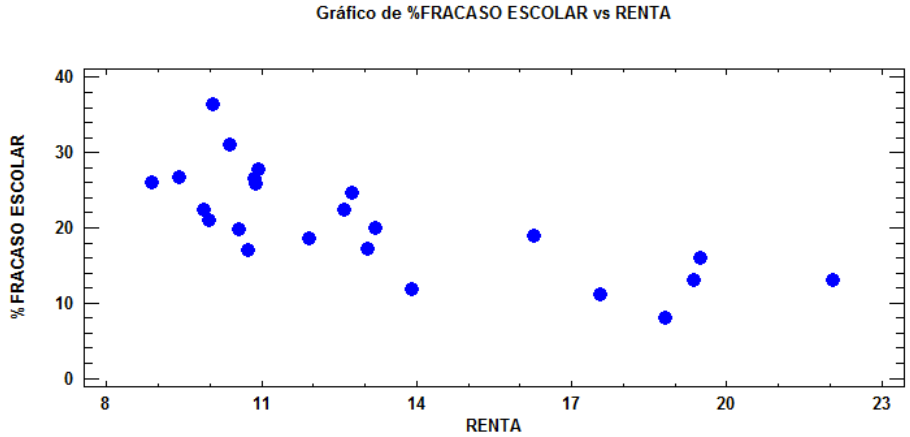
   Graficar$\longrightarrow$ Gráficos de dispersión$\longrightarrow$ Gráfico $X-Y$

   Introducir la variable dependiente en la columna marcada con Y y la variable independiente en la columna marcada con X
3. **Representar y obtener la ecuación de la recta de regresión**

   Relacionar$\longrightarrow$ Un factor $\longrightarrow$ Regresión Simple

   Introducir la variable dependiente en la columna marcada con Y y la variable independiente en la columna marcada con X

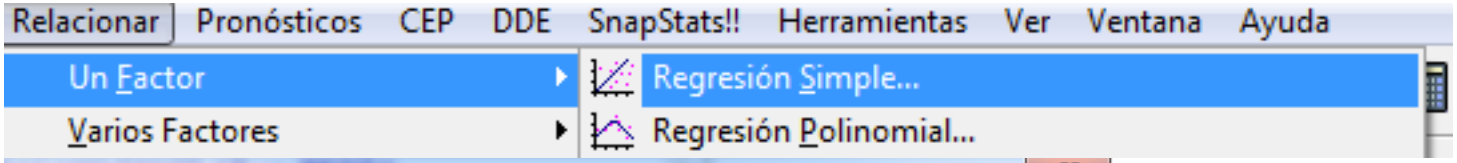




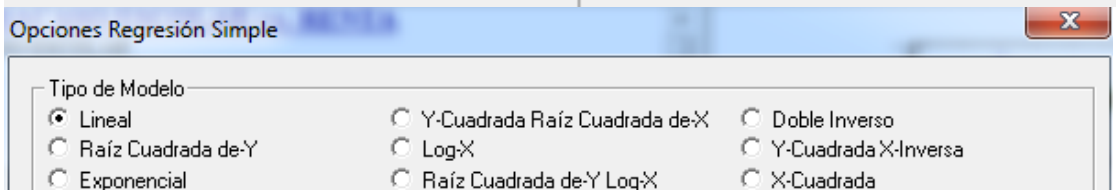
4. Estimación puntual de los parámetros.

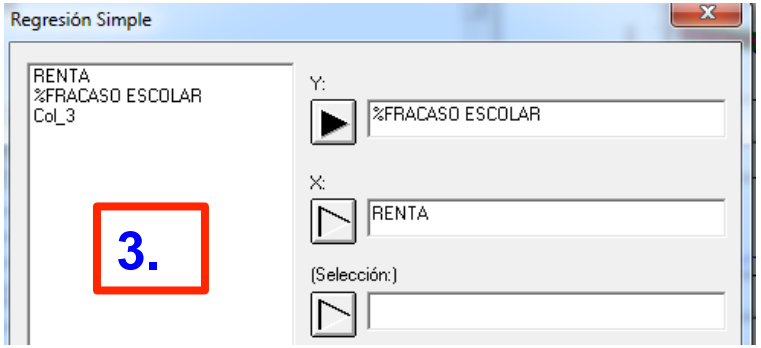
El resultado de realizar el apartado anterior es una pantalla dividida en tablas (izquierda) y gráficos(derecha). En la primera tabla, Regresión simple, aparecen las estimaciones puntuales para el parámetro α = Intercepto y para el parámetro β = Pendiente. También aparece el valor del coeficiente de determinación que explica la proporción de la variación de la variable Y que se puede explicar a través del modelo ajustado. Un valor próximo a uno nos indicará que el modelo ajustado representa de forma adecuada la relación entre los datos.
La tabla de Análisis de la Varianza recoge la información relativa al contraste de regresión con el P-valor para dicho contraste.

1.

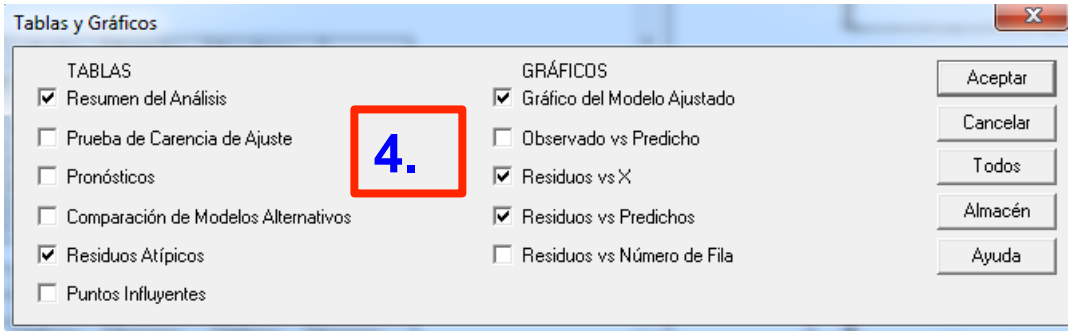

Relacionar
Pronósticos
CEP
DDE
SnapStats!!
Herramientas
Ver
Ventana
Ayuda
Un Factor
Varios Factores
Regresión Simple...
Regresión Polinomial...

2.


Opciones Regresión Simple
Tipo de Modelo
Lineal
Raíz Cuadrada de-Y
Exponencial
Y-Cuadrada Raíz Cuadrada de-X
Log-X
Raíz Cuadrada de-Y Log-X
Doble Inverso
Y-Cuadrada X-Inversa
X-Cuadrada


Regresión Simple
RENTA
%FRACASO ESCOLAR
Col_3
Y:
%FRACASO ESCOLAR
X:
RENTA
(Selección:)

3.


Tablas y Gráficos
TABLAS
Resumen del Análisis
Prueba de Carencia de Ajuste
Pronósticos
Comparación de Modelos Alternativos
Residuos Atípicos
Puntos Influyentes
GRÁFICOS
Gráfico del Modelo Ajustado
Observado vs Predicho
Residuos vs X
Residuos vs Predichos
Residuos vs Número de Fila
Aceptar
Cancelar
Todos
Almacén
Ayuda

4.

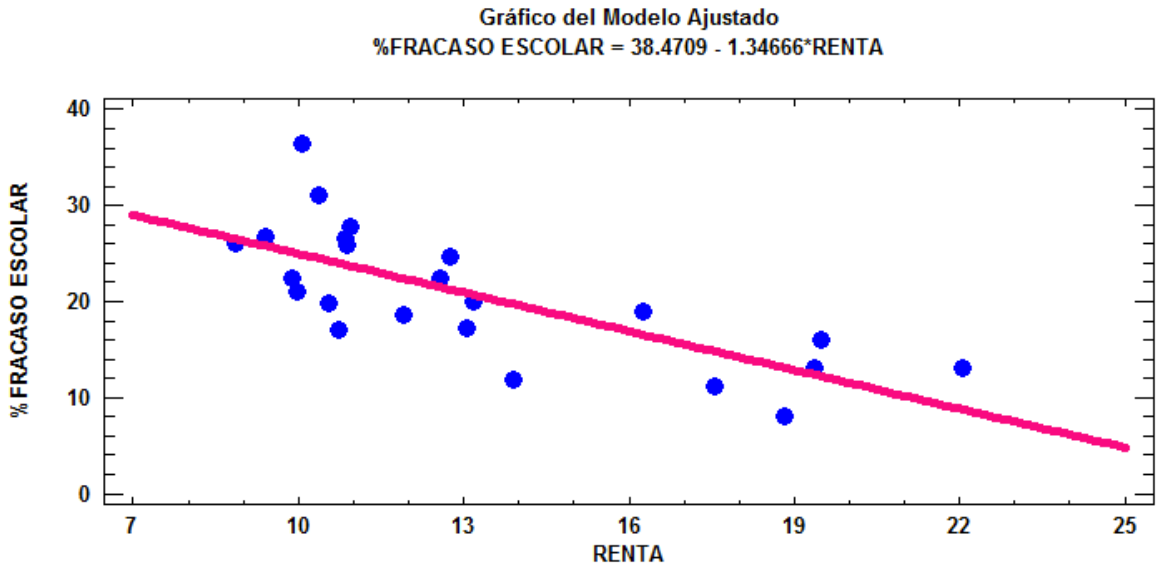
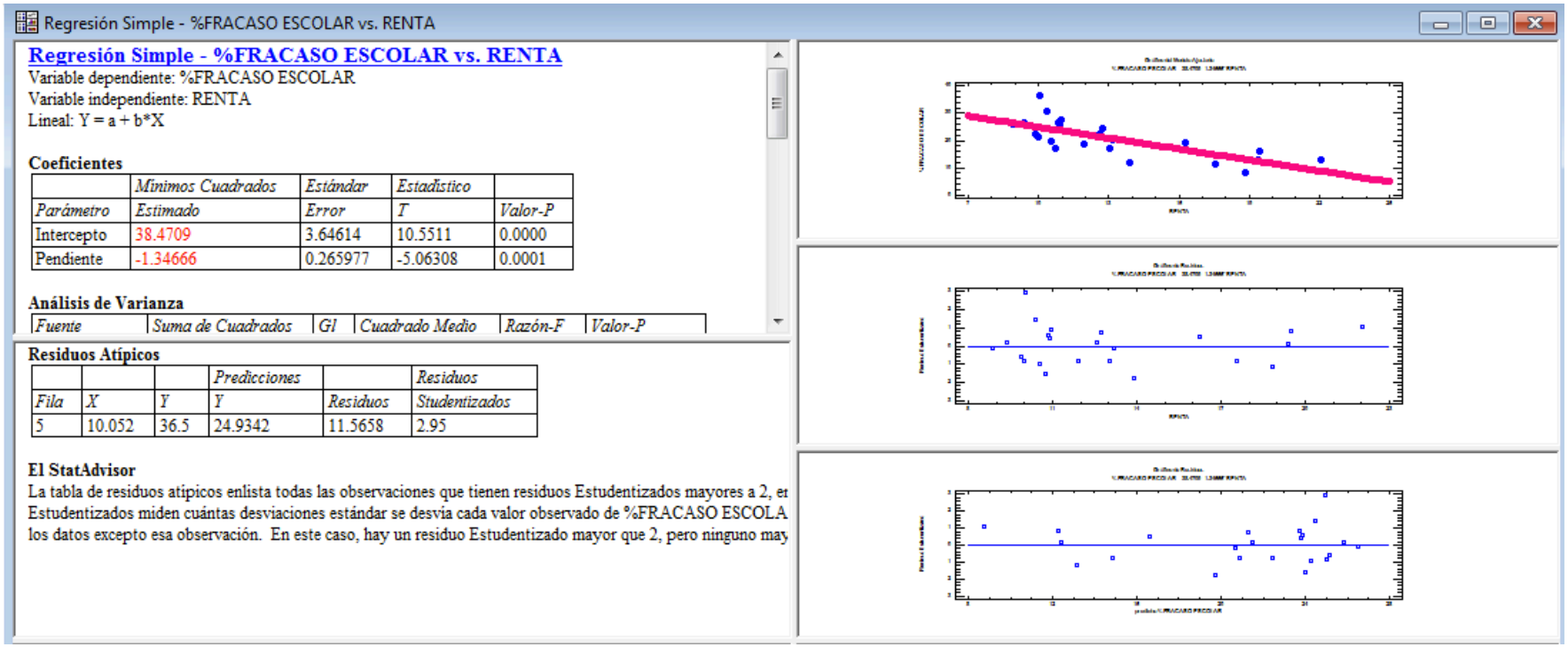


Gráfico del Modelo Ajustado
%FRACASO ESCOLAR = 38.4709 - 1.34666*RENTA
% FRACASO ESCOLAR
RENTA
0
10
20
30
40
7
10
13
16
19
22
25

4. Estimación puntual de los parámetros.

El resultado de realizar el apartado anterior es una pantalla dividida en tablas (izquierda) y gráficos(derecha). En la primera tabla, Regresión simple, aparecen las estimaciones puntuales para el parámetro α = Intercepto y para el parámetro β = Pendiente. También aparece el valor del coeficiente de determinación que explica la proporción de la variación de la variable Y que se puede explicar a través del modelo ajustado. Un valor próximo a uno nos indicará que el modelo ajustado representa de forma adecuada la relación entre los datos.
La tabla de Análisis de la Varianza recoge la información relativa al contraste de regresión con el P-valor para dicho contraste.



Contraste de regresión: Consiste en comprobar si el modelo representa la variabilidad de los datos. Se plantean dos hipótesis

$$H_0 : \beta = 0, \quad \text{no existe relación lineal entre las variables } X \text{ e } Y$$
$$H_1 : \beta \neq 0, \quad \text{sí existe relación lineal entre las variables } X \text{ e } Y$$

Este contraste se raliza con un nivel de significación fijado de antemano. Los niveles de significación usuales son 0.05 (StatGraphics por defecto), 0.01, 0.1
El Valor-P (llamado P-valor) nos indica si se debe aceptar o rechazar la hipótesis H_0. El criterio es el siguiente:

Si P-valor $>$ nivel de significación, se **ACEPTA** H_0
Si P-valor $<$ nivel de significación, se **RECHAZA** H_0

Si rechazamos la hipótesis nula, estaremos concluyendo que sí existe una relación entre las variables X e Y y que es mediante la recta que nos aparece estimada, $y = \alpha + \beta X$. Lo que no se determina con este contraste es si la recta es el mejor modelo o cuánto de bueno es. Para ver si la recta es un buen modelo o no comprobamos el valor de R^2 o del coeficiente de correlación, ρ, que son las medidas de bondad de ajuste.

Regresión Simple - %FRACASO ESCOLAR vs. RENTA

Regresión Simple - %FRACASO ESCOLAR vs. RENTA

Variable dependiente: %FRACASO ESCOLAR
Variable independiente: RENTA
Lineal: Y = a + b*X

Coeficientes

	Mínimos Cuadrados	*Estándar*	*Estadístico*	
Parámetro	*Estimado*	*Error*	*T*	*Valor-P*
Intercepto	38.4709	3.64614	10.5511	0.0000
Pendiente	-1.34666	0.265977	-5.06308	0.0001

Análisis de Varianza

Fuente	*Suma de Cuadrados*	*Gl*	*Cuadrado Medio*	*Razón-F*	*Valor-P*
Modelo	580.513	1	580.513	25.63	0.0001
Residuo	475.556	21	22.6455		
Total (Corr.)	1056.07	22			

Coeficiente de Correlación = -0.741412
R-cuadrada = 54.9692 porciento
R-cuadrado (ajustado para g.l.) = 52.8249 porciento
Error estándar del est. = 4.75873
Error absoluto medio = 3.75676
Estadístico Durbin-Watson = 1.88337 (P=0.3023)
Autocorrelación de residuos en retraso 1 = 0.0392624

Como el P-valor es menor que el nivel de significación α=0.05, se rechaza la hipótesis de que β=0. Por lo que sí existe relación lineal entre las variables

%FRACASO=34.4709 - 1.34666*RENTA

II. Predicciones e intervalos

Una vez obtenida la recta podemos hacer predicciones para distintos valores de la variable X. Estas predicción puede ser puntual o por intervalos de confianza.

1. Predicciones puntuales: Sobre la pantalla donde aparece la recta de regresión y el contraste de ANOVA, señalamos el icono de Tablas y Gráficas, marcamos Pronósticos y Aceptamos. El cuadro que aparece nos da la predicción puntual para un valor concreto de X, las bandas de predicción y las bandas de confianza. Podemos cambiar o hacer varias predicciones mediante Opciones de Ventana.

2. Predicciones por intervalos: En el mismo cuadro donde aparecen las predicciones puntuales nos encontramos los intervalos de predicción y los de confianza. El intervalo de predicción nos de mayor amplitud que el intervalo de confianza porque tiene en cuenta todas las predicciones para todos los valores de X y el intervalo de confianza se va calculando para cada valor específico de X

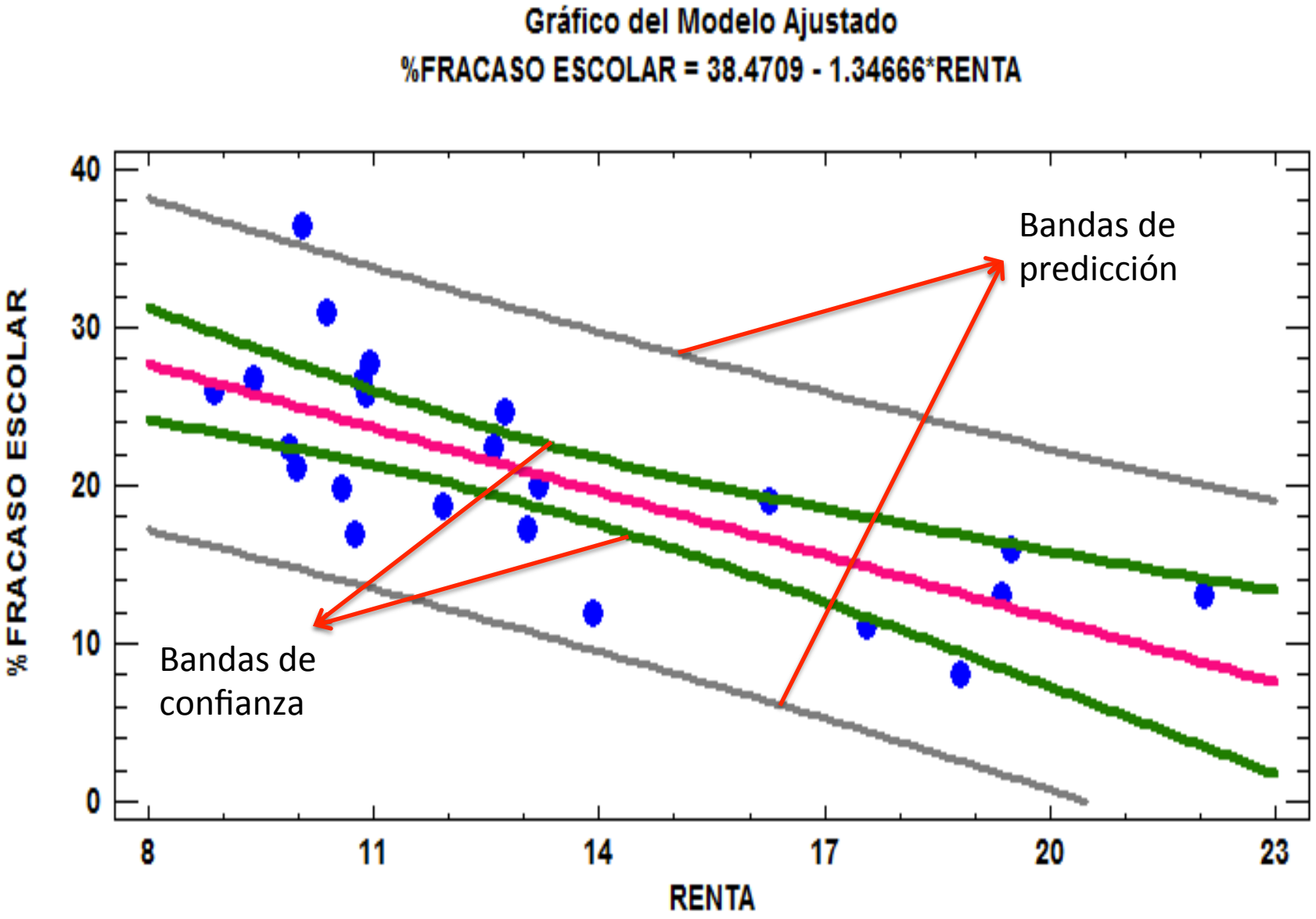


Gráfico del Modelo Ajustado
%FRACASO ESCOLAR = 38.4709 - 1.34666*RENTA
% FRACASO ESCOLAR
40
30
20
10
0
8
11
14
17
20
23
RENTA
Bandas de predicción
Bandas de confianza

Regresión Simple - %FRACASO ESCOLAR vs. RENTA

Valores Predichos

		95.00%		*95.00%*	
	Predicciones	*Límite*	*Predicción*	*Límite*	*Confianza*
X	*Y*	*Inferior*	*Superior*	*Inferior*	*Superior*
8.864	26.5341	16.1454	36.9227	23.3739	29.6943
22.05	8.77699	-2.4572	20.0112	3.46008	14.0939

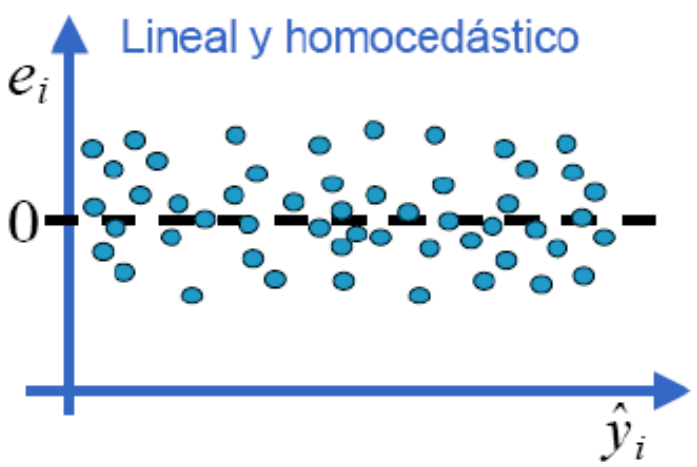
III. Diagnóstico del modelo

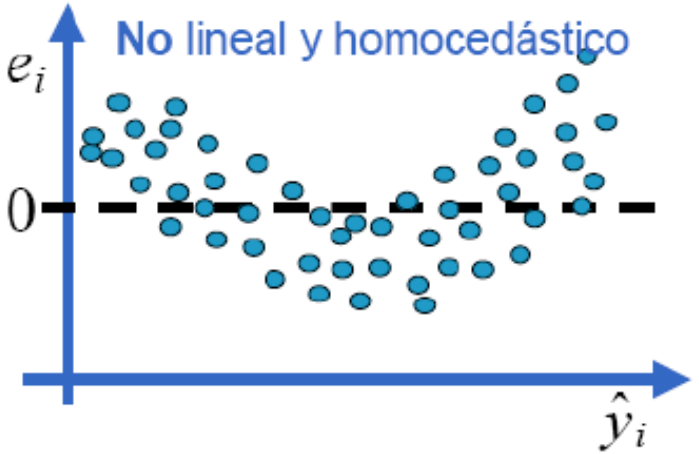
Es conveniente comprobar el cumplimiento de las hipótesis del modelo. Una manera de llevar esto a cabo es estudiar el diagrama de los residuos y verificar las propiedades. El gráfico que utilizaremos nos apareció cuando realizamos el cálculo de la recta de regresión.

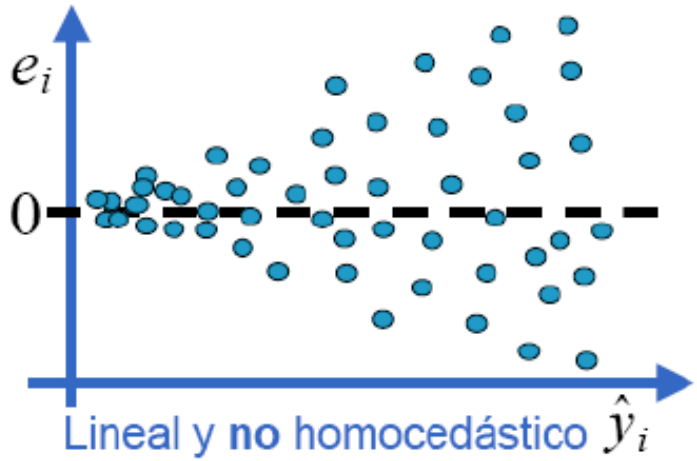
Si las hipótesis del modelo son ciertas, los residuos son Normales, con media cero, independientes, varianza constante y no hay residuos atípicos.

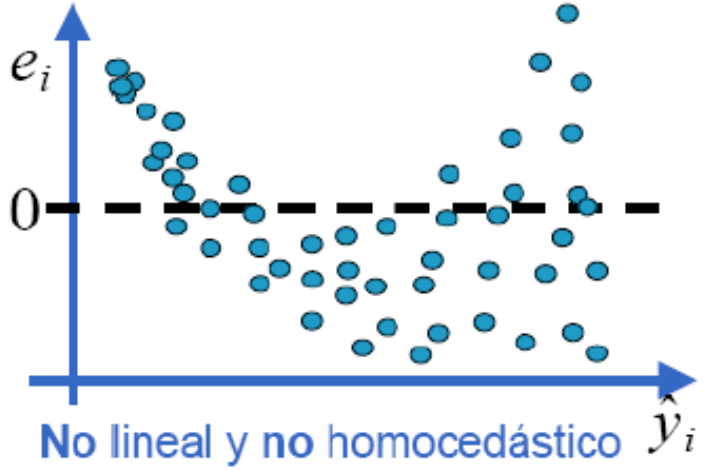
- Para comprobar esta hipótesis miraremos si en el gráfico, los residuos no muestran estructura determinada. En este caso los residuos corresponden a valores independientes obtenidos de una variable aleatoria normal. En caso contrario, si los residuos presentan alguna forma o estructura derterminada, no se cumplirán las hipótesis del modelo.

RESIDUOS – VALORES PRONOSTICADOS









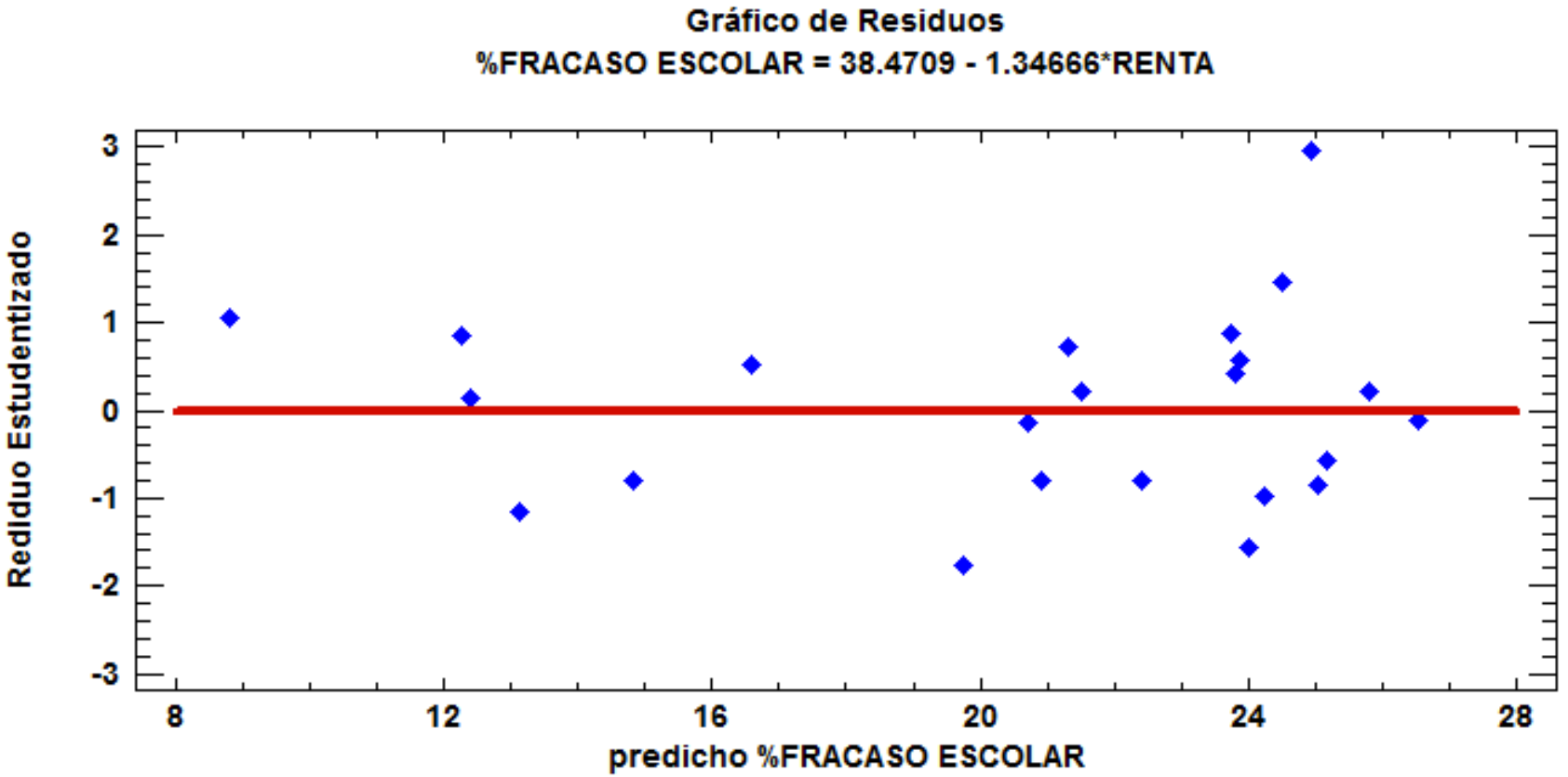


Gráfico de Residuos
%FRACASO ESCOLAR = 38.4709 - 1.34666*RENTA
Rediduo Estudentizado
3
2
1
0
-1
-2
-3
8
12
16
20
24
28
predicho %FRACASO ESCOLAR