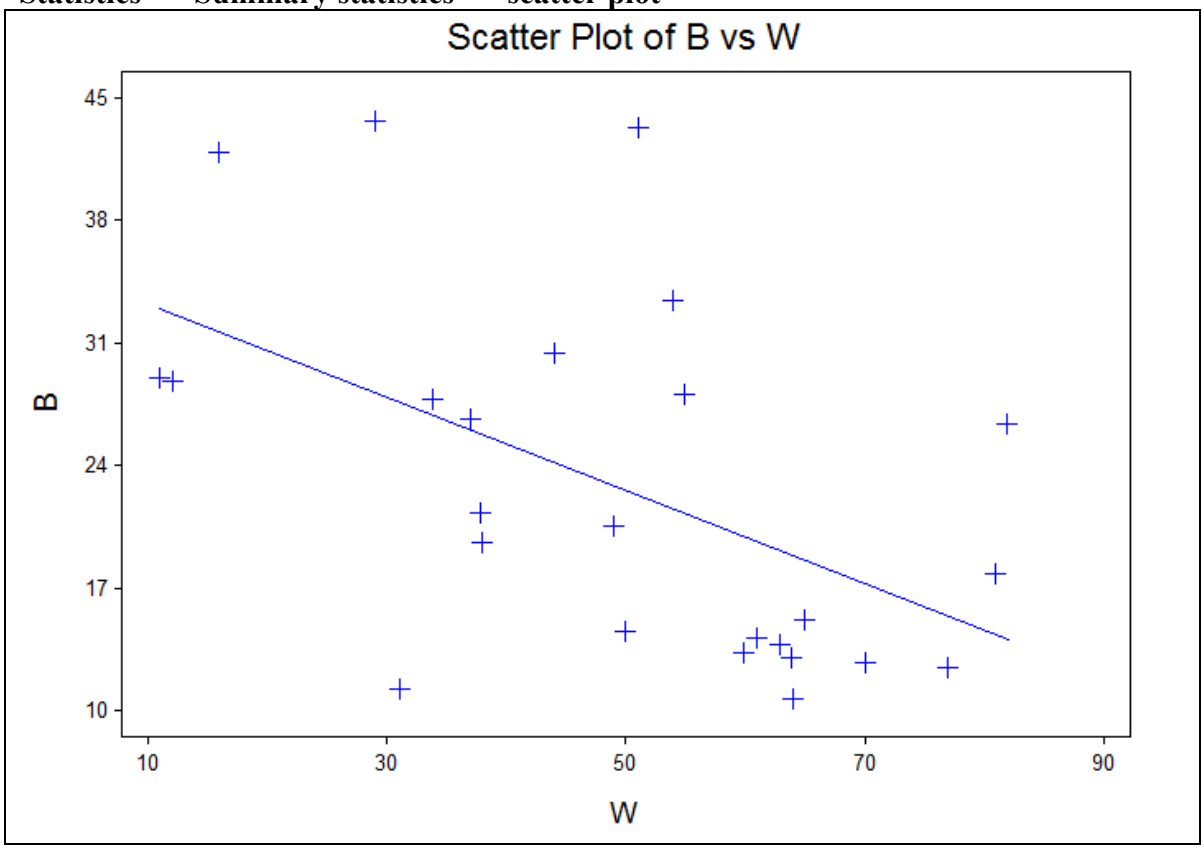
ESTADISTICA (Q) - Ejercicio de regresión - Junio 2010

El archivo **tasa_nacimiento.sx** contiene datos de tasas de nacimiento y otras variables para 26 naciones. Fuente: Statistical Abstract of the United States, 1995 and Human Development Report, 1995, Oxford University Press.

NATION
B = tasa de nacimiento cruda (número de nacimientos por cada 1000 personas),
W = actividad económica femenina (mujeres en la fuerza laboral como porcentaje de varones en la fuerza laboral)
GNP = producto bruto per capita (en miles de millones de dólares)

Considere la tasa de nacimiento (B) como la variable respuesta y la actividad económica de las mujeres (W) como variable regresora.

a) Diagrama de dispersión de B vs W. Proponer una relación lineal entre ambas parece apropiado?
Statistics => Summary statistics => scatter plot



No parece inapropiado proponer una relación lineal. La nube de puntos es muy dispersa y no parece mostrar tendencia distinta de la lineal.

b) Ajuste un modelo lineal. Obtenga la ecuación de regresión ajustada o estimada e interprete el valor estimado para la pendiente y la ordenada al origen.

statistics =>linear regression

```
STATISTIX 7.0

UNWEIGHTED LEAST SQUARES LINEAR REGRESSION OF B

PREDICTOR
VARIABLES    COEFFICIENT    STD ERROR     STUDENT'S T       P
---------    -----------    ---------     -----------    ------
CONSTANT        35.8785       4.82167         7.44       0.0000
W              -0.26583       0.09052        -2.94       0.0074

R-SQUARED           0.2727      RESID. MEAN SQUARE (MSE)    80.5268
ADJUSTED R-SQUARED  0.2411      STANDARD DEVIATION          8.97368

SOURCE        DF       SS           MS          F        P
----------   ---   ----------   ----------   -----   ------
REGRESSION     1     694.500      694.500     8.62   0.0074
RESIDUAL      23     1852.12      80.5268
TOTAL         24     2546.62

CASES INCLUDED 25   MISSING CASES 1
```

Ecuación $E(B|W) = \beta_0 + \beta_1 W$

Ecuación estimada *Tasa nacim (B) = 35.9 - 0.266 Act. Econ. fem. (W)*

Interpretación pendiente

La tasa de nacimientos disminuye en 2.66 niños por cada 1000 habitantes cuando la tasa de actividad femenina aumenta en 10 unidades.

Interpretación de la ordenada al origen

Para países con actividad femenina = 0, esperamos encontrar una tasa de nacimientos de 36 niños cada 1000 habitantes.

c) Interesa hacer un test para saber si la tasa de nacimientos es independiente de la actividad económica femenina. Haga el test. Escriba las hipótesis y las conclusiones.

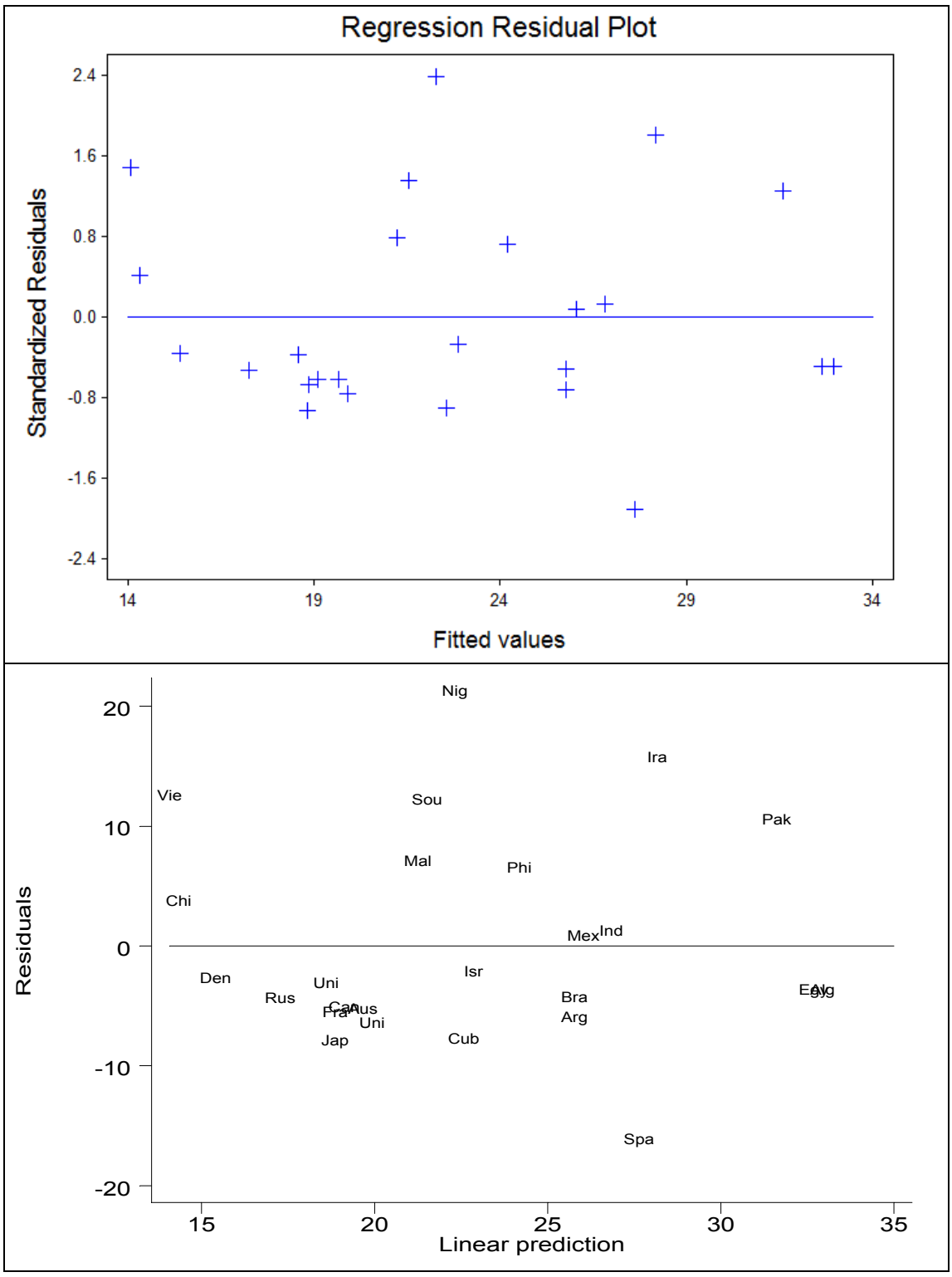
H_o: $\beta_1 = 0$ *vs* H_1: $\beta_1 \neq 0$

Como p = 0.007 concluimos que la pendiente es significativamente distinta de cero, es decir existe asociación entre B y W. Además mirando el valor del coeficiente asociado a W vemos que la asociación es negativa a mayor actividad femenina menor tasa de nacimientos.

d) Evalúe los supuestos del modelo lineal. Para ello haga:

d1) el gráfico de residuos vs valores predichos

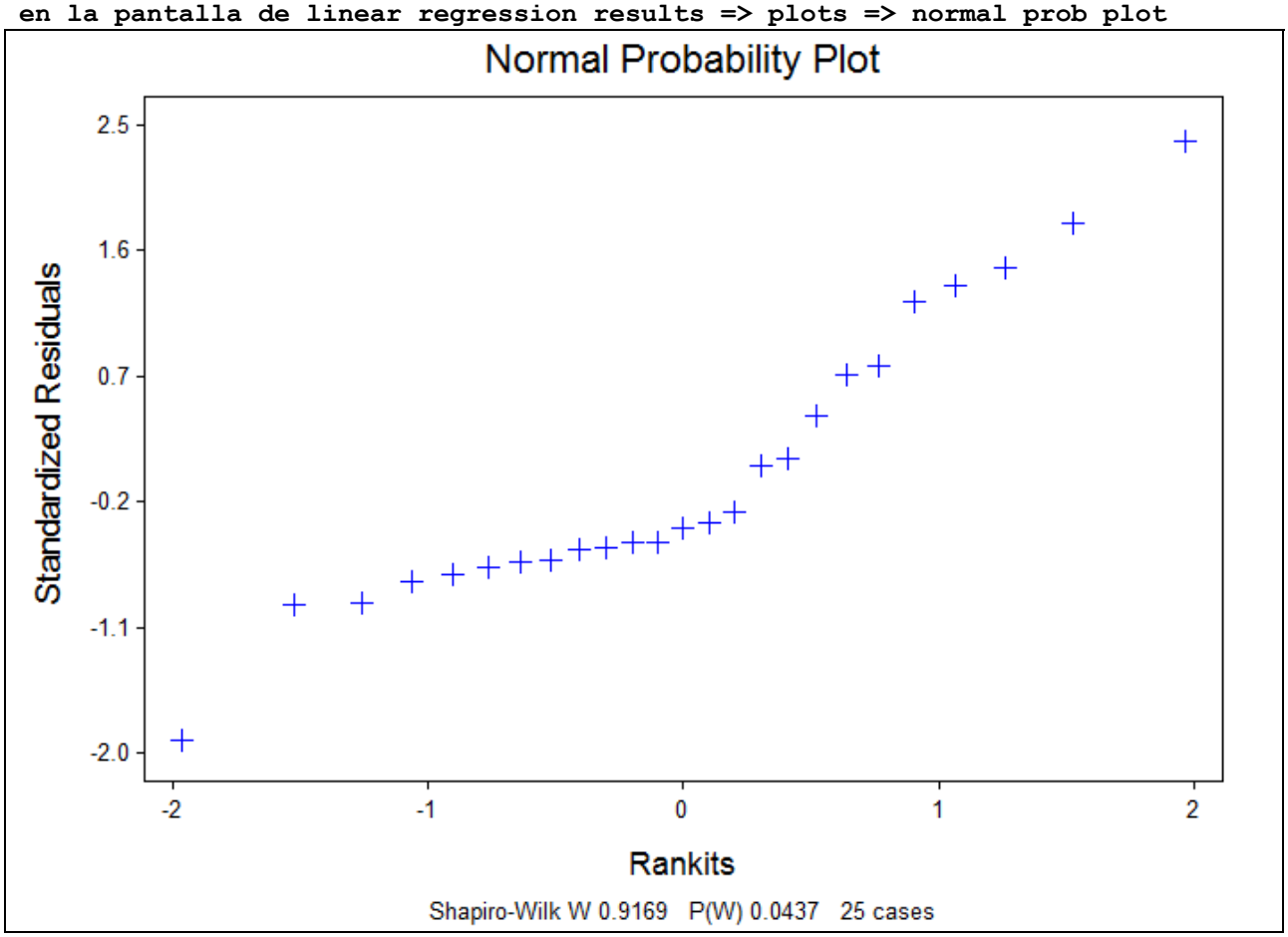
en la pantalla de linear regression results => plots =>st residuals vs fitted values



- *Cierta curvatura? Los residuos tienden a ser positivos luego negativos, luego positivos, aunque no parece muy marcado.*
- *Heterogeneidad de varianzas? Si me paro en distintos valores de yhat los residuos tienen distinta dispersión?*
- *Datos atípicos o alejados? Nigeria? Spain?*

- *Datos influyentes?*

d2) El normal probability plot de los residuos.
en la pantalla de linear regression results => plots => normal prob plot

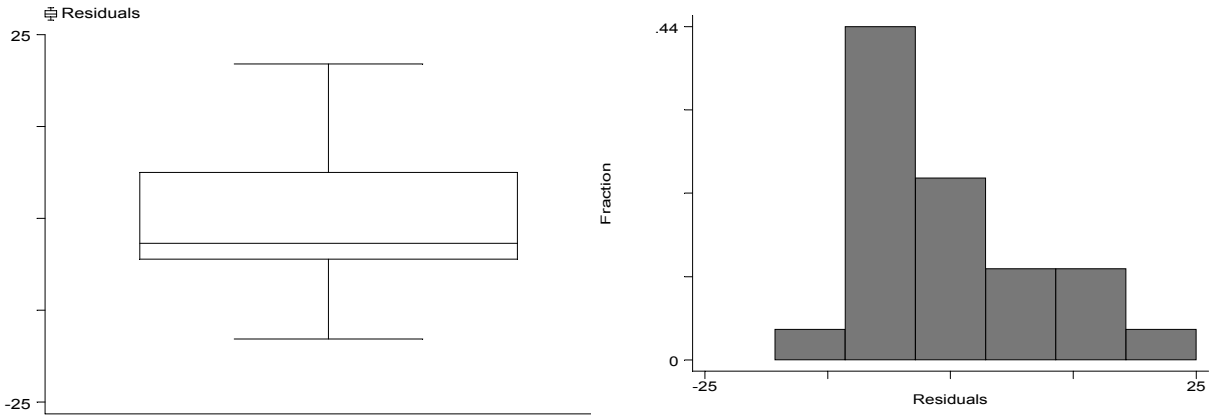


- *Apartamiento de la distribución normal? Asimetría derecha?*

d3) El test de Shapiro-Wilk sobre los residuos

El supuesto de distribución normal de los errores no parece apropiado para los datos. El p-valor es pequeño (<0.20). A pesar de no tener gran potencia nuestros datos permiten rechazar la hipótesis nula de que los errores son normales.

Veamos algunos gráficos para entender el tipo de apartamiento de la normal.



Asimetría derecha. Esto en general se logra corregir con una transformación de la variable dependiente.

d4)¿Cuál es su conclusión final sobre la validez de los supuestos del análisis de regresión?

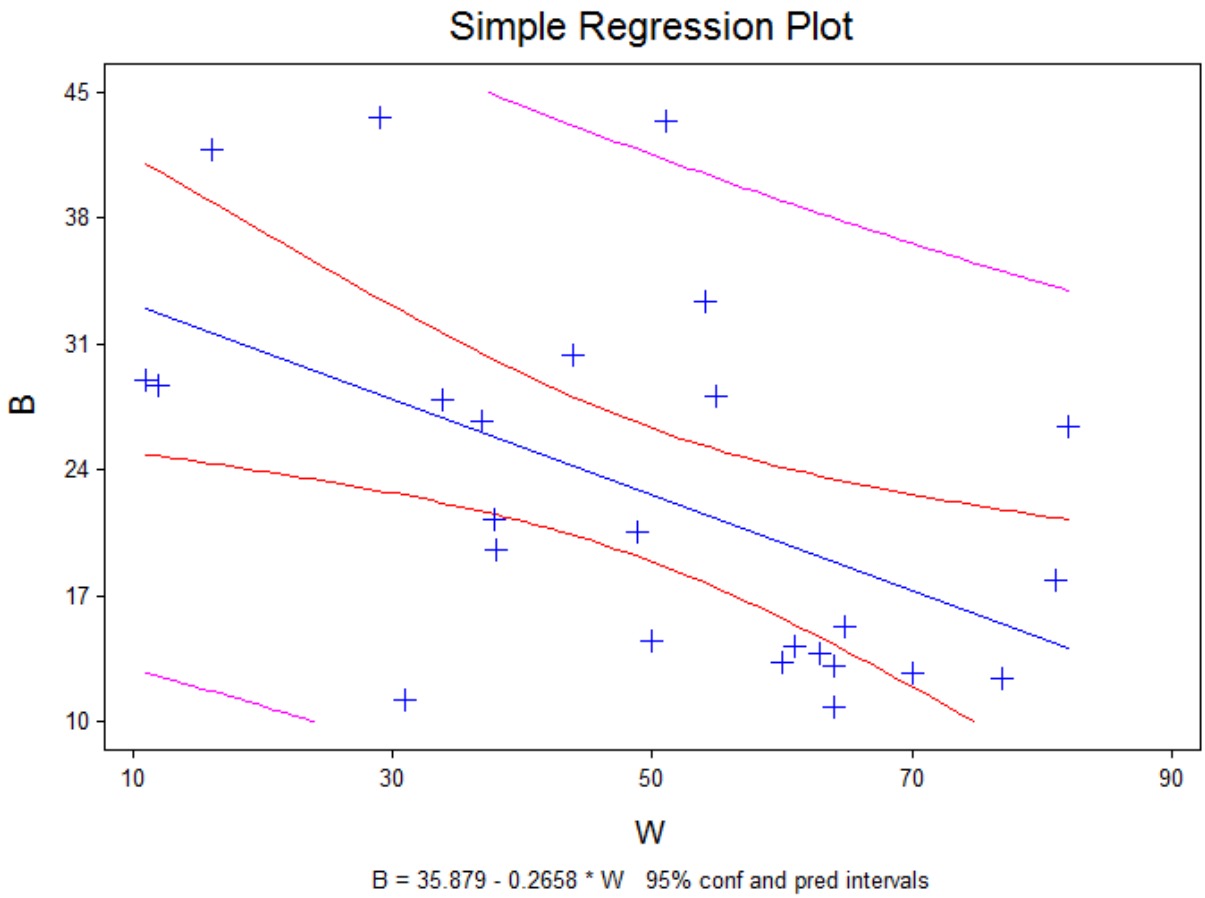
1) *El modelo lineal parece una buena primera aproximación a la relación entre B y W*
2) *El supuesto de normalidad no sería válido, en consecuencia los test e IC basados en la distribución normal pueden no ser válidos.*
3) *Condicional a W las tasas de nacimiento (B) tienen gran varianza, pero el supuesto de homogeneidad de varianzas parece razonable*

e) Calcule el coeficiente de correlación y el coeficiente de determinación. Interprete.

R-squared = 0.2727 r = -0.522

El 27% de la variabilidad total en la tasa de nacimiento es explicada por el índice de actividad económica femenina.

f) Haga un gráfico que incluya la recta de regresión estimada e intervalos de estimación y de predicción de nivel 95%
en la pantalla de linear regression results => plots => simple regression plots



g) Estime la media de la tasa de nacimiento para países con actividad económica femenina W=75 usando un intervalo del 95% de confianza.
en la pantalla de linear regression results => prediction => en la ventana "predictor values" escribimos "75"

```
PREDICTED/FITTED VALUES OF B

LOWER PREDICTED BOUND    -3.5853     LOWER FITTED BOUND      9.8841
PREDICTED VALUE           15.941     FITTED VALUE            15.941
UPPER PREDICTED BOUND     35.468     UPPER FITTED BOUND      21.999
SE (PREDICTED VALUE)      9.4393     SE (FITTED VALUE)       2.9281

UNUSUALNESS (LEVERAGE)    0.1065
PERCENT COVERAGE            95.0
CORRESPONDING T             2.07

PREDICTOR VALUES: W = 75.000
```

*IC 95% para E(B / W=75) = yhat ± $t_{23,\ 0.025}$ * SE(yhat) = 15.94 ± 2.07 * 2.928 = (9.88, 21.99)*

Interpretación: A partir de la regresión estimamos con un 95% de confianza que la media de la tasa de natalidad para países con actividad económica femenina igual a 75, es un valor entre 10 y 22.

h) ¿Cuál es el valor predicho y el residuo para Nigeria?
Para Nigeria Y = 43.3, Yhat = 22.32131 res = 20.97869

i) Excluya a Nigeria. Haga nuevamente la regresión. ¿Cambia el valor estimado para la pendiente? ¿Concluiría usted que Nigeria tiene mucha influencia en la estimación de la recta?
data => omit/select/restore => EN LA VENTANA ESCRIBO 'omit if nation="Nigeria"' [oculta del archivo el registro de Nigeria, se recupera en omit/select/restore clickeando el botón SELECT]

```
UNWEIGHTED LEAST SQUARES LINEAR REGRESSION OF B

PREDICTOR
VARIABLES    COEFFICIENT    STD ERROR     STUDENT'S T       P
---------    -----------    ---------     -----------    ------
CONSTANT        35.1757       4.28437         8.21       0.0000
W              -0.26930       0.08029        -3.35       0.0029

R-SQUARED           0.3383      RESID. MEAN SQUARE (MSE)    63.3435
ADJUSTED R-SQUARED  0.3083      STANDARD DEVIATION          7.95886

SOURCE         DF       SS           MS          F        P
----------   ---   ----------   ----------   -----   ------
REGRESSION     1     712.564      712.564    11.25   0.0029
RESIDUAL      22     1393.56      63.3435
TOTAL         23     2106.12

CASES INCLUDED 24   MISSING CASES 1
```

Sin Nigeria

```
PREDICTOR
VARIABLES    COEFFICIENT    STD ERROR     STUDENT'S T       P
---------    -----------    ---------     -----------    ------
CONSTANT        35.1757       4.28437         8.21       0.0000
W              -0.26930       0.08029        -3.35       0.0029
```

Con Nigeria

```
PREDICTOR
VARIABLES    COEFFICIENT    STD ERROR     STUDENT'S T       P
---------    -----------    ---------     -----------    ------
CONSTANT        35.8785       4.82167         7.44       0.0000
W              -0.26583       0.09052        -2.94       0.0074
```

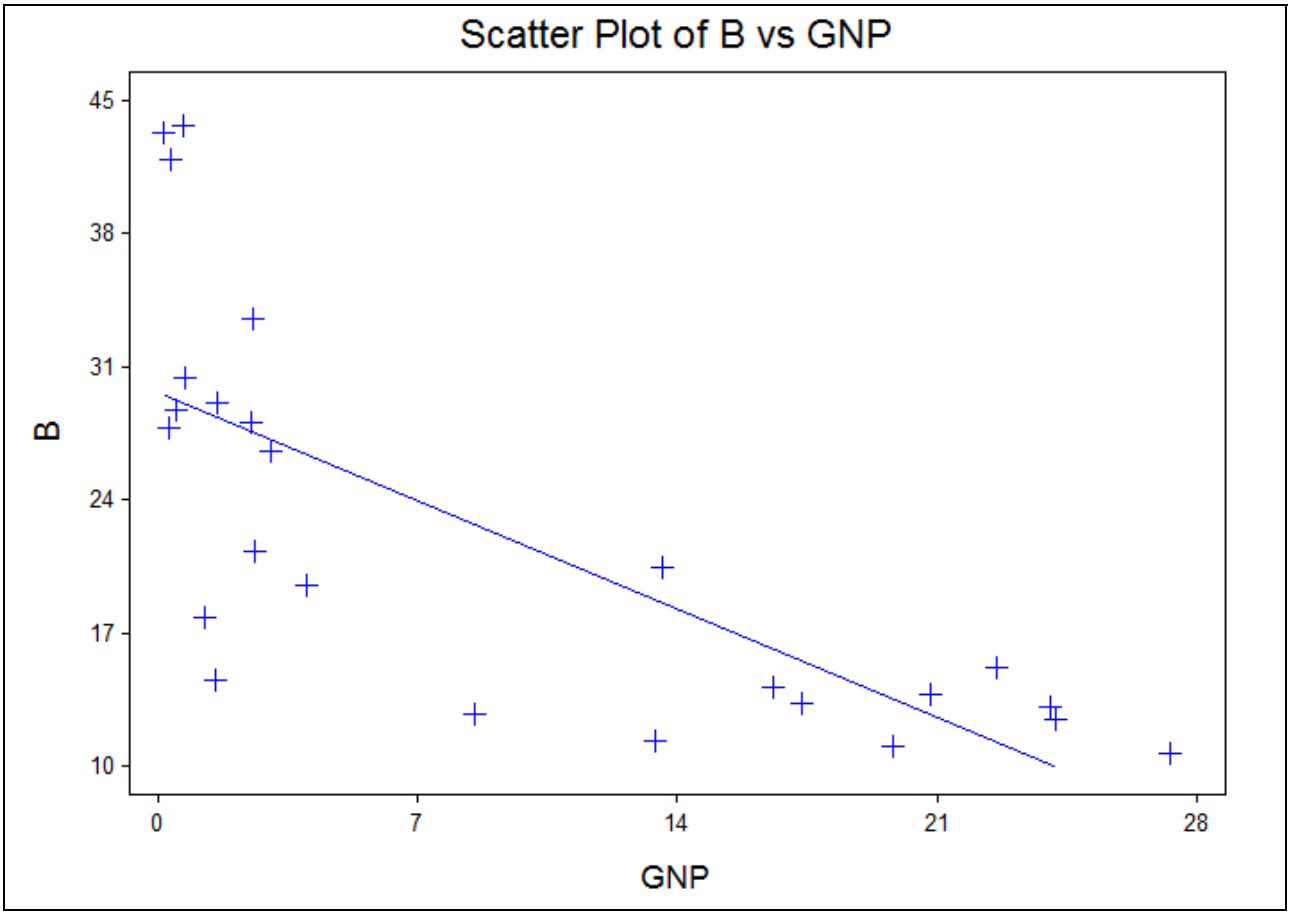
A pesar de que el residuo de Nigeria es muy grande, el sacar este dato prácticamente no modifica la estimación de los coeficientes (si la de las varianzas). En conclusión no es un dato que influya demasiado en el resultado de la estimación, el valor de W para Nigeria es muy cercano a la media de los W's de estos datos.

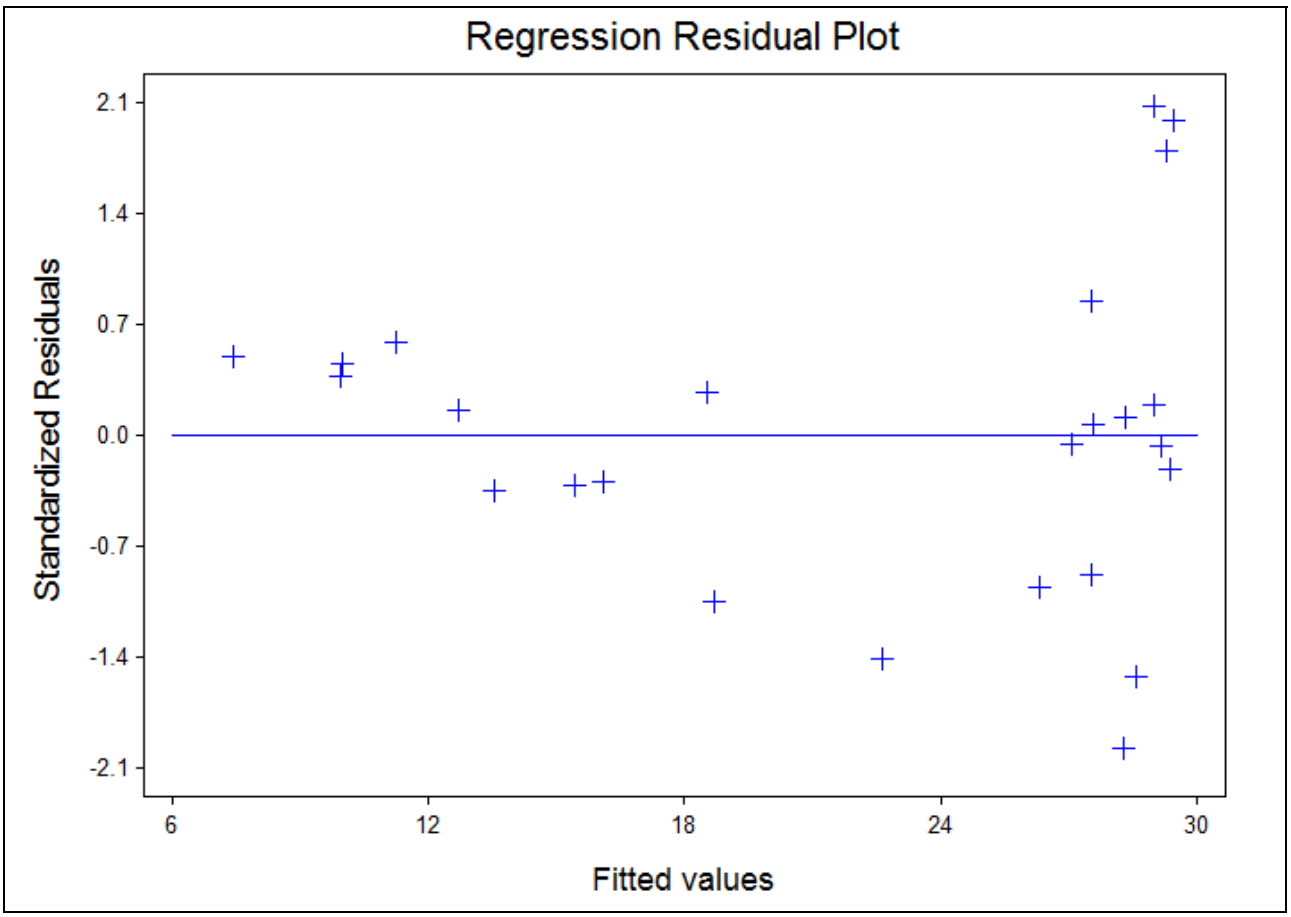
Consideramos ahora el producto bruto (GNP) como variable regresora.

a) Ajuste un modelo lineal para B y GNP.

```
UNWEIGHTED LEAST SQUARES LINEAR REGRESSION OF B

PREDICTOR
VARIABLES     COEFFICIENT     STD ERROR      STUDENT'S T        P
---------     -----------     ---------      -----------     ------
CONSTANT         29.6227        2.03742         14.54        0.0000
GNP             -0.81331        0.15503         -5.25        0.0000

R-SQUARED             0.5447      RESID. MEAN SQUARE (MSE)     52.6968
ADJUSTED R-SQUARED    0.5249      STANDARD DEVIATION           7.25925

SOURCE          DF        SS            MS          F         P
----------    ---    ----------    ----------    -----    ------
REGRESSION      1      1450.26       1450.26     27.52    0.0000
RESIDUAL       23      1212.03       52.6968
TOTAL          24      2662.29

CASES INCLUDED 25   MISSING CASES 1
```

b) Haga un scatter plot y un gráfico de residuos vs predichos y diga si el modelo lineal parece apropiado





El modelo lineal no parece apropiado!!! La recta no es un buen modelo para la relación entre GNP y B. Los residuos muestran estructura: a valores pequeños de yhat son positivos, luego negativos y finalmente positivos.