Figura 8

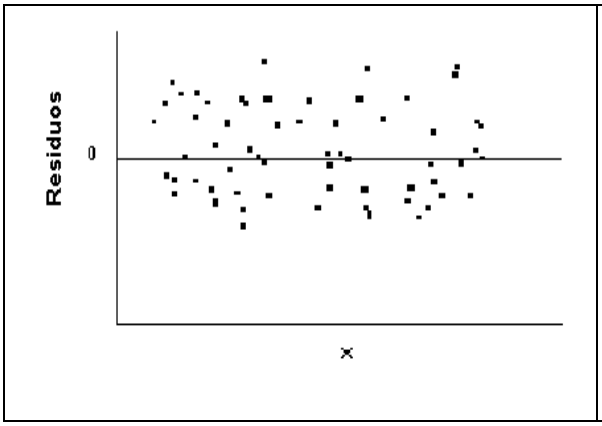
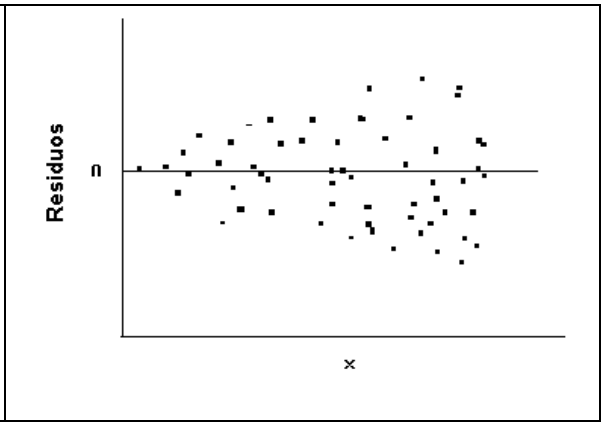
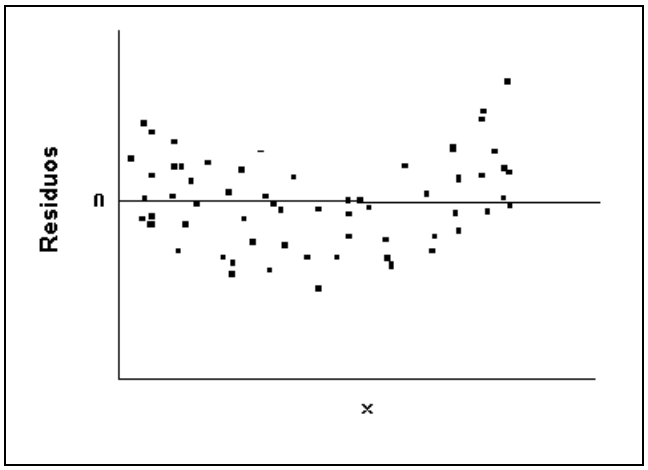


Figura 9. No se satisface el supuesto de homoscedasticidad



Si graficáramos los residuos contra los valores de *X* los puntos deberían estar distribuidos en forma de nube alrededor del valor 0 del residuo, para todos los valores de *X*, como se muestra en la figura 8. En la figura 9 los datos no satisfacen el supuesto c, ya que los residuos tienen variabilidad creciente a medida que *X* crece. En la figura 10 se aprecia una estructura curva en los residuos invalidando la linealidad.

Figura 10. No se satisface el supuesto de linealidad.



Cuando los supuestos a, b, c y d se satisfacen, los errores no están correlacionados y tienen una distribución Normal con media 0 y varianza constante.

Residuos Estandarizados

En el método de cuadrados mínimos los valores de la variable explicativa alejados de su media tienden a acercar la recta hacia ellos, esto es llamado “efecto palanca”. Como consecuencia, los residuos tienen una tendencia a ser menores para valores de x extremos, es decir que **si x_i está lejos de su promedio**, la varianza de los residuos será chica y el valor ajustado ($\hat{y}_i$) estará cerca del valor observado por efecto palanca.

$$\mathrm{Var}(e_i) = \sigma^2 (1 - h_{ii})$$

Para eliminar la tendencia de los residuos a ser menores para valores de x alejados de su media, consideramos el **residuo estandarizado** definido por:

$$r_{si} = \frac{y_i - \hat{y}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}} = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_{ii}}} \qquad (9)$$

que es el residuo dividido por su error estándar, donde

$$\hat{\sigma} = \sqrt{\frac{RSS}{n-2}} = \sqrt{\frac{\sum (y_i - \hat{y})^2}{n-2}} \qquad (10)$$

es un estimador de σ, el desvío estándar de los errores, y la cantidad

aumenta a medida que x_i se aleja del promedio $\bar{x}$.

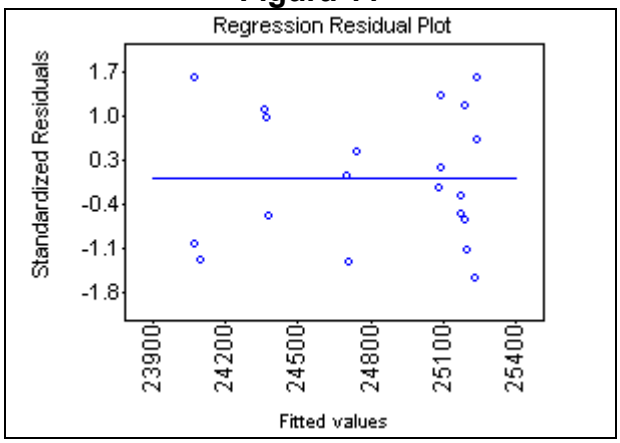
$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{k=1}^{n} (x_k - \bar{x})^2}$$

Suma fija

(11)

es una medida del alejamiento del valor x_i respecto de su media llamada **palanca,** o **"leverage".**

Figura 11



La figura 11 muestra el diagrama de dispersión (scatter plot) de los residuos estandarizados correspondientes al ejemplo de la presión de transición, graficados en función de los valores ajustados. El gráfico tiene un aspecto satisfactorio.

Recordemos: un gráfico de probabilidad Normal de un conjunto de datos indica que estos tienen una distribución aproximadamente normal cuando los puntos caen aproximadamente sobre una recta. La figura 12 muestra el gráfico de probabilidad Normal de los residuos correspondientes al mismo ejemplo. Vemos que es razonablemente lineal aunque presenta colas un poco más livianas que lo esperable bajo Normalidad. El valor del estadístico de Shapiro-Wilks es 0.9423 y su valor-p= 0.2652 > 0.20 (cuanto más alto es el p-valor mayor es la evidencia a

favor de la hip. nula de Normalidad de los errores). No se rechaza el supuesto.

Figura 12

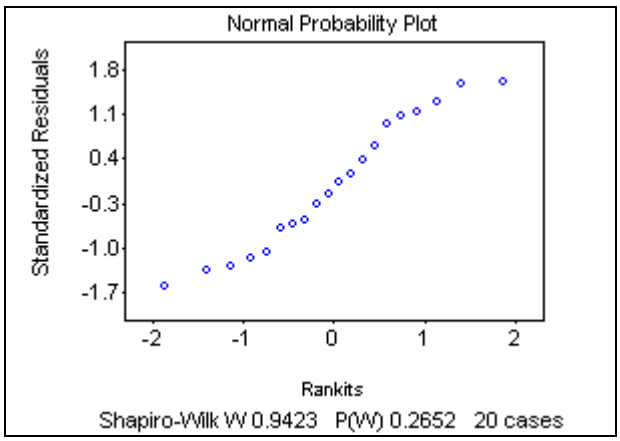


Figura 13

Con respecto al supuesto d, es útil realizar un diagrama de dispersión de los residuos en función del orden en que fueron realizadas las mediciones, figura 13. Los tests para estudiar la auto-correlación de los errores (ε) se basan en examinar los residuos (*e*). Usualmente esto es realizado mediante el test de Durbin-Watson:

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n}(e_i)^2} . \qquad (12)$$

```
DURBIN-WATSON TEST FOR AUTOCORRELATION

DURBIN-WATSON STATISTIC  2.3061

P-VALUES, USING DURBIN-WATSON'S BETA APPROXIMATION:
  P (POSITIVE CORR) = 0.6750,  P (NEGATIVE CORR) = 0.3250

EXPECTED VALUE OF DURBIN-WATSON STATISTIC     2.1065
EXACT VARIANCE OF DURBIN-WATSON STATISTIC    0.17682
```

Si los sucesivos residuos son no correlacionados entonces $d \cong 2$. Valores de d alejados de 2 indican la presencia de cierto tipo de auto-correlación, la denominada auto-correlación a lag 1. Un p-valor alto asociado a este estadístico, por ejemplo mayor que 0.10, indica que la diferencia de d con 2 no es estadísticamente significativa, que puede atribuirse al azar, y por lo tanto no se rechaza la hipótesis nula de residuos no correlacionados. La aplicación del test de Durbin-Watson sólo tiene sentido cuando los datos están ordenados en el tiempo o en el espacio, o sea por ejemplo, cuando el índice i representa el tiempo. Por otra parte, y debido a que el test fue diseñado para detectar cierto tipo especial de auto-correlación, puede ocurrir que con este estadístico no se detecten auto-correlaciones cuando en realidad el supuesto no se satisface.

Concluimos que se satisfacen razonablemente los supuestos del Método de Cuadrados Mínimos.

15.7.1 Evaluación de la asociación: Asociación positiva, asociación negativa

Dos grupos de datos (que corresponden a valores de dos variables) están asociados en forma **positiva** cuando los valores que están por **encima del promedio** en uno de ellos tienden a ocurrir mayoritariamente con valores por **encima del promedio** del otro. Dos variables están asociadas en forma **negativa** cuando valores por **encima del promedio** de una suelen estar acompañados por valores por **debajo del promedio** de la otra y viceversa.

El signo de la pendiente, b, de la recta ajustada nos indica si la asociación es positiva o negativa. Sin embargo, la pendiente no mide directamente la fuerza (o grado) de la asociación. Esto se debe a que el valor absoluto de la pendiente está intrínsecamente vinculado a las unidades en las que se han expresado las mediciones. Podemos obtener valores tan grandes o tan chicos como queramos con sólo elegir las unidades adecuadamente.

Las medidas de asociación que consideramos a continuación no varían con cambios en las unidades de medición.

15.7.2 Correlación de Pearson

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}. \qquad (13)$$

El coeficiente de correlación de Pearson mide el grado de asociación *lineal* de un conjunto de datos $(x_1, y_1), \ldots, (x_n, y_n)$, de tamaño n, correspondiente a observaciones de dos variables continuas X e Y

Como la pendiente de la recta ajustada por cuadrados mínimos está dada por

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

resulta que el coeficiente de correlación se puede expresar de la siguiente manera

$$r = \frac{s_x}{s_y} b$$

donde s_x y s_y son las dispersiones muestrales de *X* e *Y* respectivamente. Cuando las dispersiones muestrales son iguales (s_x = s_y), la correlación es igual a la pendiente. Esto ocurre cuando las variables se han estandarizado, en cuyo caso ambas variables tienen desvío 1. Es por esta razón que el coeficiente de correlación de Pearson es también llamado *coeficiente de regresión estandarizado*.

15.7.3. Propiedades del Coeficiente de correlación

- La correlación, a diferencia de la pendiente b, trata a los valores x´s e y´s en forma simétrica. El valor del coeficiente de correlación muestral r , no depende de las unidades en que se miden las variables y su valor está siempre entre -1 y 1.
- A mayor valor absoluto de r, mayor el grado de asociación lineal.
- r tiene el mismo signo que b.
- Cuando r = 0 también b = 0, no hay una tendencia lineal creciente o decreciente en la relación entre los valores x´s e y´s.
- Los valores extremos, *r* = 1 y *r* = -1, ocurren únicamente cuando los puntos en un diagrama de dispersión caen exactamente **sobre** una recta. Esto corresponde a asociaciones positivas ó negativas perfectas. En este caso, el error de predicción es cero al utilizar la recta ajustada $\hat{y}$ = a + b x, para predecir el valor de Y.
- Valores de *r* cercanos a 1 ó -1 indican que los puntos yacen **cerca** de una recta.
- Valores de r positivos indican que la mayoría de los desvíos $x_i - \bar{x}$ e $y_i - \bar{y}$ tendrán el mismo signo, es decir, hay una asociación positiva entre las variables.
- Valores de r negativos indican que la mayoría de los desvíos $x_i - \bar{x}$ e $y_i - \bar{y}$ tendrán signos opuestos, es decir, hay una asociación negativa entre las variables.

La figura 14 muestra cómo los valores de r se acercan a cero, i.e. se alejan del 1 ó -1, a medida que decrece el grado de asociación lineal entre las variables.

Figura 14. Comportamiento del coeficiente de correlación a medida que decrece el grado de asociación lineal.

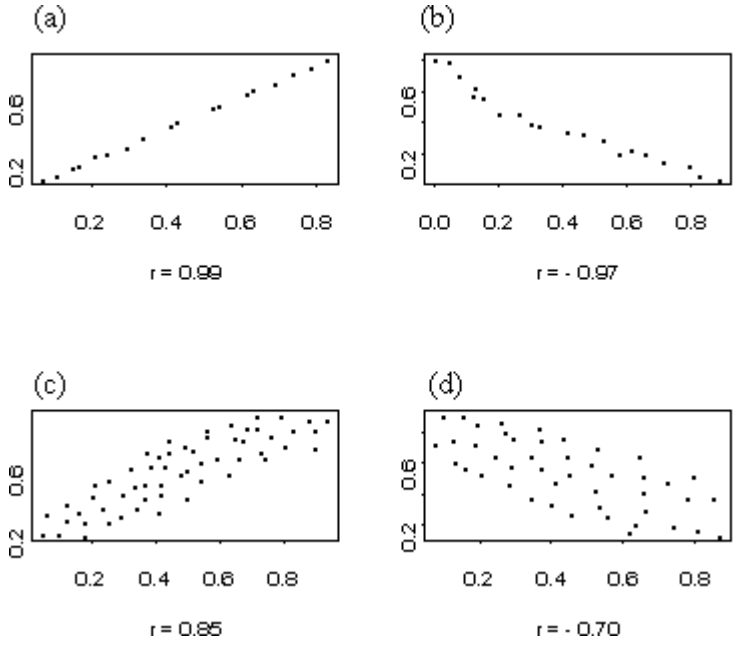
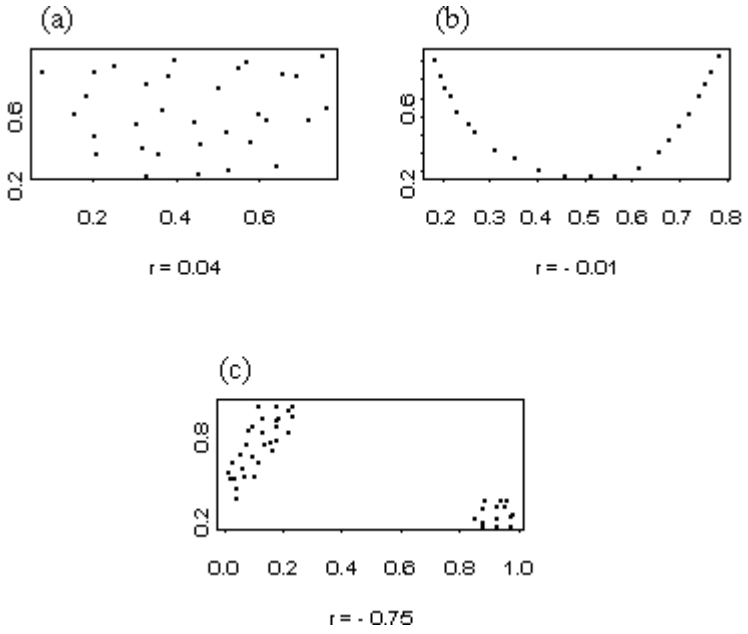


Figura 15. Comportamiento del coeficiente de correlación ante diferentes tipos de asociaciones no lineales.



La figura 15. (a) muestra una falta total de asociación entre las dos variables y un coeficiente de correlación cercano a cero, en cambio la (b), que también tiene un r cercano a cero, muestra una clara relación funcional entre las variables y la (c) muestra dos grupos uno con asociación positiva y otro con asociación nula, sin

embargo el coeficiente de correlación, -0.75, indica una asociación negativa.

Podemos escribir a la pendiente de la recta ajustada como:

$$b = r \frac{s_Y}{s_X}. \tag{14}$$

Si ambas variables (X e Y) están estandarizadas, de manera que sus medias sean cero y su desvíos estándar 1, entonces la recta de regresión tiene pendiente r y pasa por el origen.

15.7.4 Coeficiente de determinación

El coeficiente de determinación es una medida de la proporción en que se reduce el error de predicción de una variable respuesta (Y) cuando se predice Y utilizando los valores de una variable X en la ecuación de predicción, $\hat{y}$ = a + b x, con respecto al error de predicción que se obtendría sin usarla.

El **coeficiente de determinación R^2** se define como

$$R^2 = (TSS - RSS) / TSS = 1 - RSS / TSS \tag{15}$$

- TSS, es la suma de cuadrados total. Es decir, la suma de los cuadrados de las desviaciones de cada respuesta observada a la media:

  $$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2. \tag{16}$$

  Mide el error de predicción si no se tiene en cuenta la variable explicativa y se predice la respuesta por $\bar{y}$, la media muestral de las respuestas observadas.
- RSS, es la suma de cuadrados de los residuos:

  $$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 . \tag{17}$$

  Es una medida del error de predicción que se comete cuando la variable respuesta se predice por $\hat{y}$ = a + b***x***, la teniendo en cuenta la variable explicativa.

Cuando hay una fuerte relación lineal entre X e Y, el modelo ajustado provee predictores $\hat{y}$ mucho mejores que $\bar{y}$, en el sentido que la suma de cuadrados de los errores de predicción es mucho menor.

Como el numerador de R^2 , TSS - RSS es igual a $\Sigma (\hat{Y}_i - \bar{Y})^2$, R^2 puede expresarse como

$$R^2 = \frac{\sum \left(\hat{Y}_i - \bar{Y}\right)^2}{\sum \left(Y_i - \bar{Y}\right)^2}.$$

El numerador se denomina suma de cuadrados explicada por el modelo (REGRESSION

sum of squares en el STATISTIX, y ESS en otros programas). Por lo tanto R^2 mide la proporción de la variación total explicada por la regresión.

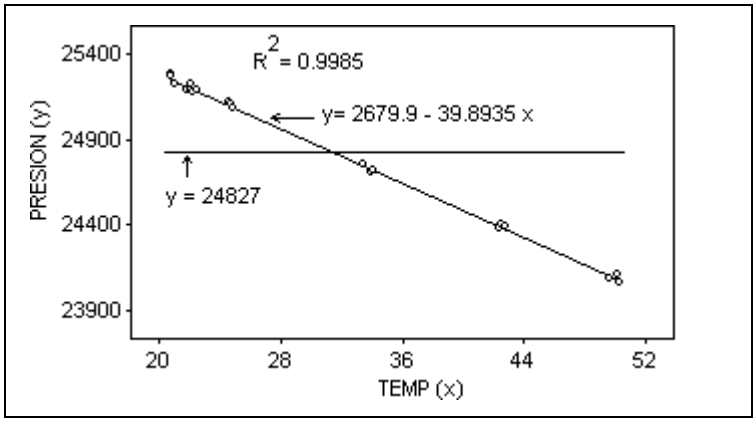
La correlación al cuadrado coincide con el coeficiente de determinación **únicamente en regresión lineal simple**:

$$r^2 = R^2 \qquad (18)$$

Consideremos el diagrama de dispersión del presión de transición versus la temperatura que aparece en la figura 16. La línea horizontal representa la media de los valores de PRESION, 24827, de las 20 observaciones. Los valores observados de Y varían, como lo indican las desviaciones verticales de los puntos a la línea horizontal.

La otra recta es la de mínimos cuadrados. Los desvíos verticales de los puntos a esta recta son, en muchísimo menores. Sobre esta recta, cuando *x* cambia *y* cambia, de manera que esta relación lineal explica una parte de la variación de *Y*. Como $R^2 = 0.9985$ decimos que la recta de regresión explica el 99% de la variación total observada en la presión de transición.

Figura 16



- El coeficiente de determinación no depende de las unidades en que se expresan los datos y toma valores entre cero y uno.
- Vale 0 cuando la regresión no explica nada; en ese caso, la suma de cuadrados total es igual a la suma de cuadrados de los residuos.
- Vale 1 cuando la variabilidad observada de la respuesta es explicada totalmente por la regresión; en ese caso, la suma de los cuadrados de los residuos es cero.