15.8 Leyendo la salida de un programa estadístico

Cada programa estadístico presenta los resultados de la regresión en forma diferente, pero la mayoría provee la misma información básica. La tabla 2 muestra la salida del Statistix para la regresión de la presión de transición sobre la temperatura.

Tabla 2

```
UNWEIGHTED LEAST SQUARES LINEAR REGRESSION OF PRESION

PREDICTOR
VARIABLES     COEFFICIENT     STD ERROR         STUDENT'S T        P
---------     -----------     ---------         -----------     ------
CONSTANT (a)     26079.9      11.9034            2190.97        0.0000
TEMP     (b)    -39.8935      0.35843 ⇐ s(b)    -111.30        0.0000

R-SQUARED           0.9985    RESID. MEAN SQUARE (MSE)  298.854 ⇐ σ̂²
ADJUSTED R-SQUARED  0.9985    STANDARD DEVIATION        17.2874 ⇐ σ̂

SOURCE         DF        SS           MS          F         P
----------    ---    ----------   ----------   -----    ------
REGRESSION      1      3702187      3702187   12387.96   0.0000
RESIDUAL       18      5379.36      298.854 ⇐ σ̂²

TOTAL          19      3707567

CASES INCLUDED 20   MISSING CASES 0
```

La columna encabezada **COEFFICIENT** presenta la ordenada al origen (CONSTANT = ***a*** = 26079.9) y la pendiente (***b*** = -39.8935).

La columna encabezada **STD ERROR** presenta los correspondientes errores estándar.

El cociente entre el valor del coeficiente y su error estándar aparece en la columna **STUDENT'S T**. Sus p-valores, presentados en la columna encabezada **P** , permiten testear si los coeficientes son significativamente distintos de cero. En este caso ambos p-valores (0.0000 y 0.0000) son menores que 0.01 y decimos que los coeficientes *a* y *b* son altamente significativos.

El coeficiente de la variable TEMP, es estadísticamente significativo distinto de cero, entonces decimos que **la variable** temperatura **explica la respuesta**, presión de transición

Otros estadísticos incluidos en la tabla 2 son:

R-SQUARED: Coeficiente de determinación $R^2 = 1 - \frac{5379.36}{3707567} = 0.9985$.

RESID. MEAN SQUARE ERROR (MSE), $\hat{\sigma}^2$: Cociente entre la suma de cuadrados de los residuos y sus grados de libertad, RSS / DF = 5379.36 / 18 = 298.854

STANDARD DEVIATION es la raíz cuadrada del MSE. Es un estimador del desvío σ. En nuestro ejemplo $\hat{\sigma}$ = $\sqrt{298.854}$ = 17.2874.

La salida incluye también una tabla, denominada tabla de análisis de la varianza correspondiente al ajuste, la cuál contiene las sumas de cuadrados, bajo el encabezado "**SS**", los grados de libertad de cada suma de cuadrados ("**DF**"), el cociente entre la suma de cuadrados y sus correspondientes grados de libertad ("**MS**") y finalmente el estadístico **F** y su **p**-valor asociado.

TOTAL SS Es la suma de cuadrados total, TSS = 3707567 (g.l. = 19)

RESIDUAL SS Es la suma de cuadrados residual, RSS = 5379.36 (g.l. = 18)

REGRESSION SS Es la suma de cuadrados explicada por el modelo, TSS - RSS = 3702187

REGRESSION MS (TSS - RSS) / 1 =3702187 / 1 = 3702187
RESIDUAL MS: RSS / 18 = 5379.36 / 18 = 298.854 $\Leftarrow \hat{\sigma}^2$

ESTADISTICO F Es una medida global de la bondad de la regresión y se calcula como el cociente de los dos últimos valores definidos, es decir

$$F = \frac{(TSS - RSS)/1}{RSS/18} = \frac{3702187}{298.854} = 12387.96$$

El correspondiente valor-p = 0.0000 y por lo tanto decimos que la regresión es altamente significativa.

Observación : En la **regresión lineal simple únicamente**, el p-valor del estadístico F coincide con el del test para decidir si la pendiente es estadísticamente distinta de cero. Esto no ocurre cuando se incluyen más variables explicativas al modelo.

16 Intervalo de confianza para la pendiente

Podemos calcular un intervalo de confianza para la pendiente de la recta ajustada y también realizar un test para decidir si es significativamente distinta de cero. Una pendiente cero querría decir que no hay relación lineal entre *Y* y *X*.

Recordemos que la pendiente de la recta ajustada es:

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

Luego Var (b) = $\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ (considerando que las x's son fijas y sin error)

y el **desvío estándar estimado de la pendiente** es:

$$s(b) = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \qquad (19)$$

donde $\hat{\sigma}^2$ es un estimador de σ^2, calculado como suma de los cuadrados de los residuos dividida por *n*-2.

Rara vez será necesario utilizar (19) para el cálculo ya que s(b) es un valor que muestra el Statistix automáticamente al realizar un ajuste por cuadrados mínimos. En este caso, *s(b)* = 0.35843 (Vea la tabla 2, **STD ERROR**). Sin embargo es interesante detenerse en su expresión ; vemos que:

- intervienen todas las observaciones,
- a medida que aumenta el número de observaciones (*n*), *s*(b) se hace cada vez más pequeño,
- para un mismo número de observaciones, cuanto más dispersos estén los valores x_i tanto más pequeño será *s*(b).

Los extremos de un intervalo de confianza de nivel (1-α) 100% para la pendiente son:

$$b \pm t_{n-2,\,\alpha/2} * s(b)\,, \qquad (20)$$

Conclusión: si queremos que la estimación de la pendiente de la recta sea lo más precisa posible, debemos elegir un tamaño de la muestra grande y los valores de la variable explicativa lo más espaciados que se pueda dentro del rango de interés.

Tomando α = 0.05, en nuestro ejemplo *n* -2 = 18, b = -39.8935, *s(b)* = 0.35843 y $t_{18,\,0.025}$ = 2.10, resulta el intervalo del 95% de confianza (-40.646, -39.141) de las pendientes compatibles con los datos.
Como el cero no pertenece al intervalo de confianza obtenido, se rechaza la hipótesis de pendiente nula al nivel 0.05.

17. Intervalos de confianza e intervalos de predicción.

Hemos dicho (sección 15.3) que la recta ajustada puede utilizarse de dos maneras distintas

a) para *estimar* de la media poblacional de Y para cada x fijo.
b) para *predecir* un valor futuro de Y para un valor fijo de x.

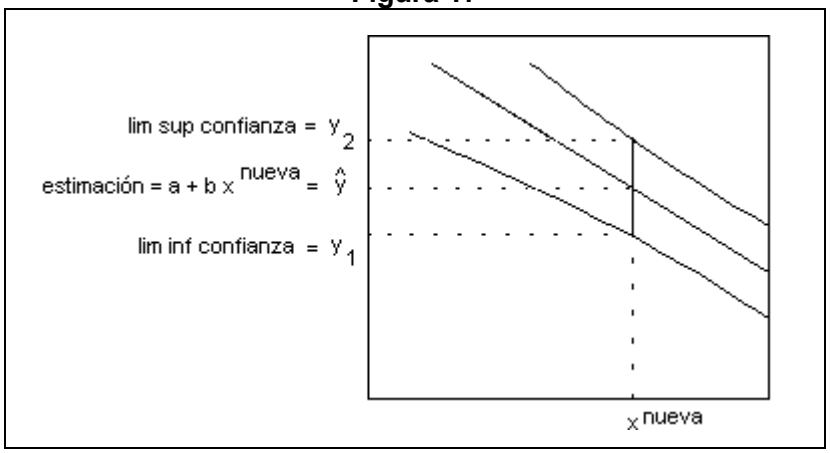
Agregaremos algunas más

c) para *estimar* de la media poblacional de Y para varios valores de x diferentes.
d) para *predecir* varios valores futuros de Y para cada uno con un valor fijo de x diferente.
e) realizar predicciones del valor de X que dio lugar al nuevo valor observado de Y. Esto se llama *predicción inversa.*

Los intervalos que resultan de a) - d) están todos centrados en $\hat{y}$, difieren únicamente en su amplitud debido a la diferencia en las varianzas.

17.1 Intervalos de confianza para la respuesta media

Figura 17



Debe tenerse en cuenta la incerteza de la recta ajustada. Para ello se construye una *banda* alrededor de la recta de regresión ajustada, tal que para cada valor fijo de x (x^{nueva}), el intervalo determinado por la banda y una recta vertical a la abscisa en x^{nueva}, sea un intervalo de confianza del (1-α) 100%:

$$a + b\, x^{nueva} \pm t_{n-2,\alpha/2}\, \sigma \sqrt{\frac{1}{n} + \frac{(x^{nueva} - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Si llamamos $s(a + b\,x^{nueva})$ a $\sigma\sqrt{\frac{1}{n} + \frac{(x^{nueva} - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$ tenemos que los límites de un intervalo de confianza para la media de la variable Y dado el valor x^{nueva} son

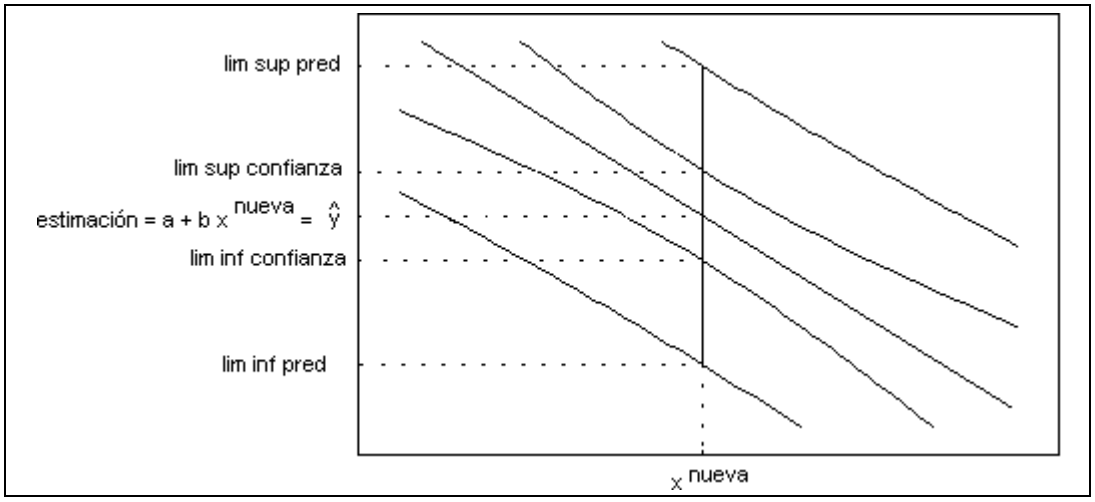
$$a + b\,x^{nueva} \pm t_{n-2,\alpha/2}\; s(a + b\,x^{nueva}) \qquad (21)$$

En particular si α = 0.05, el 95% de confianza significa que el intervalo es uno de una familia de intervalos, tal que 95 de cada 100 contienen la verdadera media de *Y* para ese valor fijo de x (x^{nueva}) ; 5 no. Podemos confiar en que el que tenemos es uno de esos 95.

17.2 Intervalos de predicción para una observación futura

Para un mismo valor (*x*) de la variable explicativa, un intervalo de predicción refleja, además de la variabilidad debida a que la recta estimada no representa exactamente la media verdadera de la variable respuesta para ese valor de X, la **variabilidad** individual de la variable respuesta alrededor de la media verdadera y es por esa razón es de mayor amplitud que el intervalo de confianza.

Figura 18. Intervalos de confianza junto con los intervalos de predicción para una observación futura



La expresión general de los límites de predicción del (1-α) 100 % para una observación futura (y^{nueva}) para el valor x^{nueva} de la variable explicativa es:

$$a + b\,x^{\text{nueva}} \pm t_{n-2,\alpha/2}\,\hat{\sigma}\,\sqrt{1+\frac{1}{n}+\frac{(x^{\text{nueva}}-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}\,, \qquad (22)$$

$$a + b\,x^{\text{nueva}} \pm t_{n-2,\alpha/2}\;\; s(\text{pred } Y)$$

La única diferencia entre el intervalo de confianza (21) y el de predicción (22) es que aparece un 1 dentro de la raíz. Esta diferencia hace que la longitud de los intervalos de confianza pueda hacerse tan pequeña como se quiera, con tal de tomar suficientes observaciones, mientras que la longitud de los intervalos de predicción nunca pueda ser menor que $2\,t_{n-2,\,\alpha/2}\;\hat{\sigma}$

Si la cantidad de observaciones es grande la raíz que aparece en la expresión (22) es aproximadamente igual a 1 y la longitud del intervalo de predicción de nivel 0.95, resulta cerca de 4*s.* Por lo tanto, si estamos interesados en predicción, $4\hat{\sigma}$ es un excelente indicio de la calidad del ajuste, y como consecuencia, de la incerteza de las predicciones.

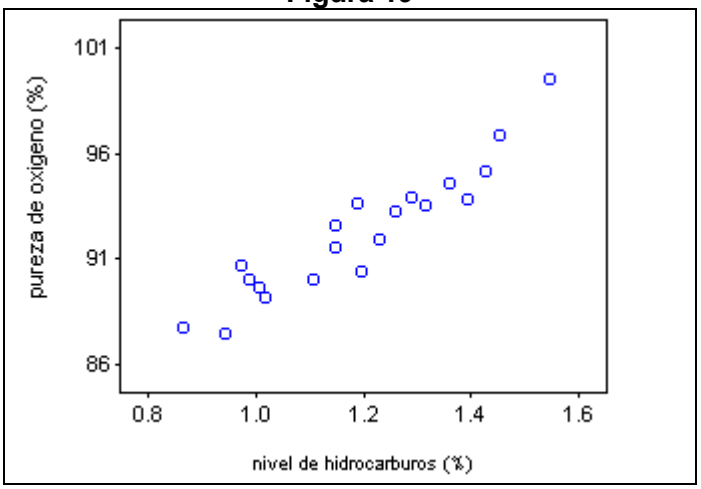
Ejemplo: Interesa estudiar la relación entre la pureza del oxigeno (Y) producido en un proceso de destilación y el porcentaje de hidrocarburos (X) presentes en el condensador principal de la unidad de destilación. No se conoce un modelo determinístico funcional que relacione la pureza del oxígeno con los niveles de hidrocarburo.

Tabla 3. Niveles de oxigeno e hidrocarburos

x(%)	y(%)	x(%)	y(%)	x(%)	y(%)	x(%)	y(%)
0.99	90.01	1.36	94.45	1.19	93.54	1.2	90.39
1.02	89.05	0.87	87.59	1.15	92.52	1.26	93.25
1.15	91.43	1.23	91.77	0.98	90.56	1.32	93.41
1.29	93.74	1.55	99.42	1.01	89.54	1.43	94.98
1.46	96.73	1.4	93.65	1.11	89.85	0.95	87.33

El diagrama de dispersión de la figura 19 muestra que a pesar de que ninguna curva simple pasará por todos los puntos hay una tendencia lineal creciente de manera que es razonable suponer que la media de la pureza de oxígeno esté relacionada linealmente con el nivel de hidrocarburos.

Figura 19



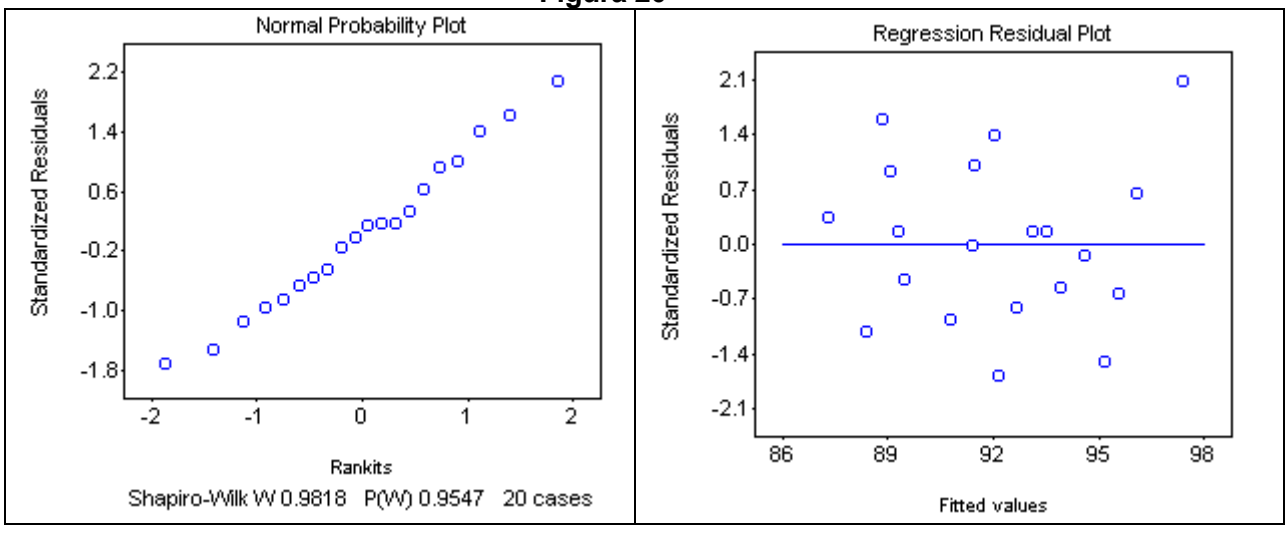
La tabla 4 los coeficientes y sus errores estandar, resultantes de un ajuste de cuadrados mínimos a los datos de la pureza de oxígeno. La variable X (% de hidrocarburos) es estadísticamente significativa.

Tabla 4

PREDICTOR VARIABLES	COEFFICIENT	STD ERROR	STUDENT'S T	P
CONSTANT	74.2833	1.59347	46.62	0.0000
X	14.9475	1.31676	11.35	0.0000

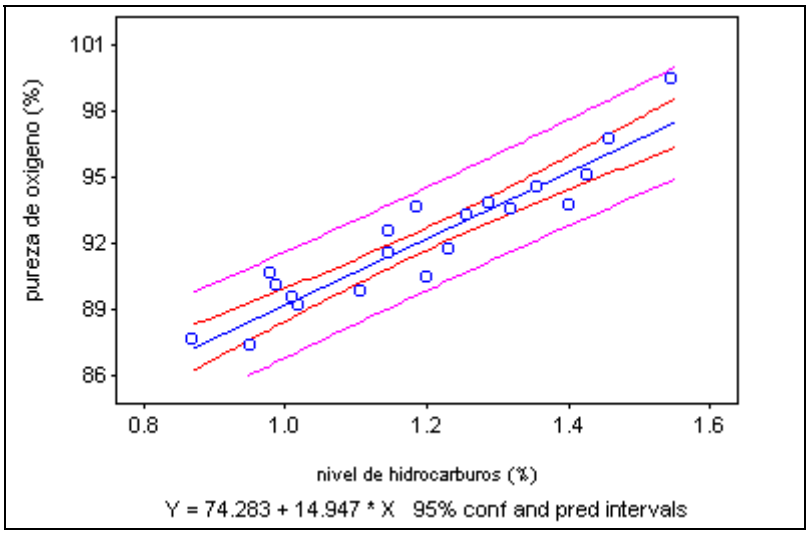
R-SQUARED	0.8774	RESID. MEAN SQUARE (MSE)	1.18055
ADJUSTED R-SQUARED	0.8706	STANDARD DEVIATION	1.08653

Figura 20



Los gráficos de la figura 20 nos permiten concluir que los datos no presentan alejamientos de los supuestos de Normalidad y homoscedasticidad.

Figura 21. Recta ajustada junto con las bandas de confianza y de predicción del 95%



La banda interna de la figura 21 es la banda de confianza (21). El intervalo más angosto, (91.650 , 92.671), se encuentra en el nivel promedio de hidrocarburos ($\bar{x}$ = 1.196 %). Los intervalos se van ensanchando a medida que aumenta la distancia a dicho valor promedio. Un alto porcentaje de valores observados cae fuera de la banda de confianza. Esto pone de manifiesto que dichas bandas están formadas por intervalos de confianza para la respuesta media, nada dicen respecto de los valores de la variable de interés.

La longitud de estos intervalos decrece con el aumento del tamaño de la muestra y/o de la dispersión de los valores de la variable independiente.

Siguiendo con el ejemplo, en el nivel promedio de hidrocarburos (1.196 %), el intervalo de predicción es (89.821 ; 94.500). Los intervalos de predicción (22) del 95% también se ensanchan con la distancia al nivel promedio de hidrocarburos, aunque esto no se ve fácilmente de la figura.