Regresión: Tarea por Grupos

Cuervo Fernández, Esther García Prado, Sergio

7 de diciembre de 2017

Parte I
Ejercicio Kutner: Mantenimiento de Copiadoras

Para la realización del ejercicio de *Mantenimiento de Copiadoras* mediante SAS, lo primero que se ha llevado a cabo es la importación del conjunto de datos a partir del fichero. Para facilitar dicha tarea, se ha realizado una fase previa de preprocesado para convertir el fichero a formato `csv` denotando por Y la primera columna y X la segunda (tal y como indica el enunciado). Una vez hecho esto, se ha utilizado el fragmento de la figura 14 para importar el conjunto de datos en *SAS*. Por tanto, ya se puede comenzar a realizar el ejercicio.

1.20. Copier maintenance. The Tri-City Office Equipment Corporation sells an imported copier on a franchise basis and performs preventive maintenance and repair service on this copier. The data below have been collected from 45 recent calls on users to perform routine preventive maintenance service; for each call, X is the number of copiers serviced and Y is the total number of minutes spent by the service person. Assume that first-order regression model (1) is appropriate.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{1}$$

Analizaremos el modelo de la ecuación 1 mediante un estudio de regresión lineal simple, con Y_i como variable dependiente, X_i como variable independiente, β_0 como término independiente, y β_1 como término dependiente o pendiente de la recta de regresión. ε_i es el error aleatorio que supondremos normal con media 0 y varianza σ^2, es decir $\varepsilon_i \sim N(0, \sigma^2)$.

a. Obtain the estimated regression function.

Para obtener la generación del modelo de regresión a mediante *SAS* se ha utilizado el fragmento de códido *SAS* de la figura 15, que calcula los estimadores del modelo de regresión lineal simple que se ilustra en la ecuación (1). A partir de dicha sentencia se han obtenido distintas salidas que después han sido utilizadas en otros apartados pedidos por el enunciado del problema.

Para la resolución de este apartado, ha sido suficiente con consultar el "resumen", que se muestra en la figura 1, a partir de la cual se han obtenido los valores de β_0 y β_1, que se muestran en las ecuaciones (2) y (3) respectivamente.

Por tanto, la estimación de la función de regresión simple obtenida se muestra en la ecuación (4).

The REG Procedure
Model: MODEL1
Dependent Variable: Y

Number of Observations Read	45
Number of Observations Used	45

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	76960	76960	968.66	<.0001
Error	43	3416.37702	79.45063		
Corrected Total	44	80377			

Root MSE	8.91351	R-Square	0.9575
Dependent Mean	76.26667	Adj R-Sq	0.9565
Coeff Var	11.68729		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	90% Confidence Limits	
Intercept	1	-0.58016	2.80394	-0.21	0.8371	-5.29378	4.13347
X	1	15.03525	0.48309	31.12	<.0001	14.22314	15.84735

Figura 1: *Salida SAS*: Mantenimiento de Copiadoras - Resumen de Regresión Lineal Simple

$$\widehat{\beta}_0 = -0.58016 \tag{2}$$

$$\widehat{\beta}_1 = 15.03525 \tag{3}$$

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \varepsilon_i = -0.58016 + 15.03525 X_i + \varepsilon_i \tag{4}$$

b. Plot the estimated regression function and the data. How well does the estimated regression function fit the data?

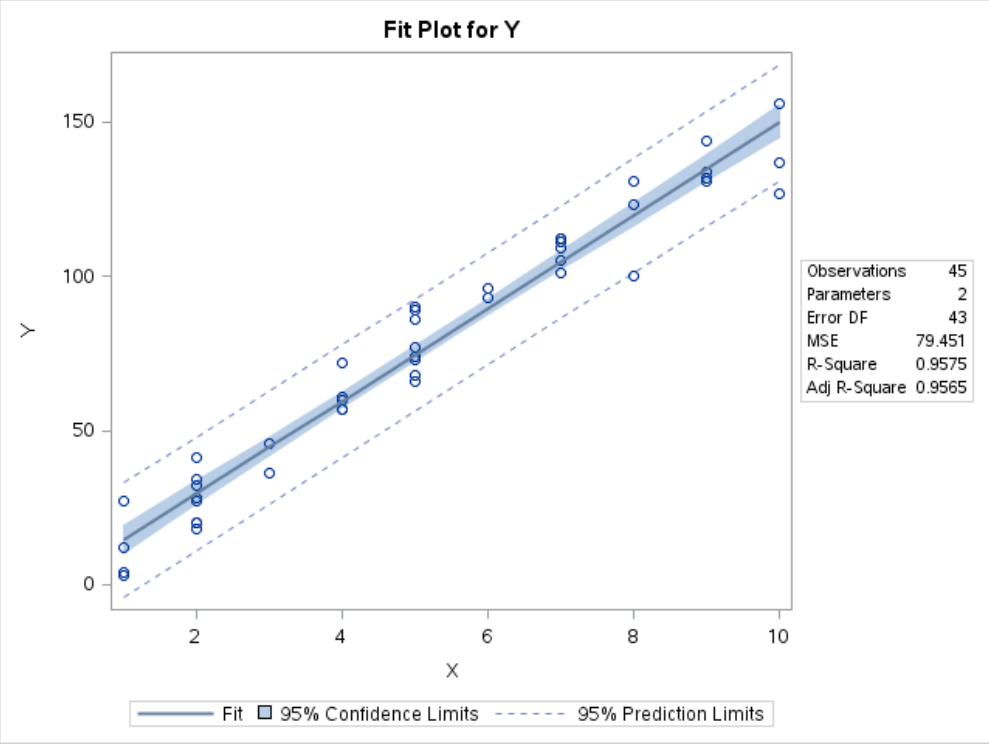

Figura 2: *Salida SAS*: Mantenimiento de Copiadoras - Gráfico de Regresión Lineal Simple

En este apartado se pide representar la recta de regresión obtenida en el apartado anterior. Afortunadamente, en este caso no ha sido necesaria la utilización de otro fragmento de código, sino que con el utilizado en el apartado anterior (figura 15) ya se generaba dicho gráfico.

El gráfico en cuestión se muestra en la figura 2, a partir del cual se puede observar la relación lineal existente entre la variable independiente X y la variable dependiente Y. Además, se puede apreciar como el modelo de regresión lineal simple se ajusta de manera apropiada a los datos.

En los siguientes apartados se analizará la dispersión de los datos así como los intervalos de confianza para la media y de predicción. Sim embargo, a simple vista y debido al contexto y unidades de medida de los datos, parece coherente el nivel de dispersión de los datos (variaciones entorno a 10 minutos entre instalaciones).

c. Interpret $\widehat{\beta}_0$ in your estimated regression function. Does $\widehat{\beta}_0$ provide any relevant information here? Explain.

El valor estimado de la ordenada en el origen (o "intercept") se muestra en la ecuación (2), el cual está muy próximo al valor 0. La interpretación del término independiente en este caso podría ser *el número de minutos que se emplean en realizar* 0 *mantenimientos*, el cual tiene sentido que esté próximo a 0 minutos.

Esto podría deberse a que no se contabiliza el tiempo cuando no hay mantenimiento por hacer, lo cual tiene sentido ya que el estudio trata de analizar la duración del proceso de *mantenimiento de copiadoras*. Algo a destacar es que, posiblemente debido a la muestra patrón utilizada para la generación del modelo, el término independiente ha sido ajustado con valor negativo. Sin embargo, a partir de este valor no se puede obtener ninguna conclusión adicional ya que su valor se ha ajustado de manera que el ajuste global (en términos de mínimos cuadrados) sea mínimo.

Tal y como se verá en el apartado e del ejercicio 2.5, no existen evidencias significativas para asumir que sea distinto de cero. Pero tal y como se discute en dicho apartado, no es apropiado eliminarlo ya que este haría que parte del error explicado por el modelo se convirtiese en error aleatorio, lo cual es algo negativo para el ajuste.

d. Obtain a point estimate of the mean service time when $X = 5$ copiers are serviced.

Para la obtención de una estimación puntual para $X = 5$, se ha decidido realizar el proceso de añadir una nueva observación al conjunto de datos (de manera que tan solo contenga el valor de la variable independiente X). Para ello, se ha utilizado el fragmento de código *SAS* de la figura 16. Nótese que esto podría haberse llevado a cabo mediante la simple sustitución del valor X en la función de regresión de la ecuación (4), pero esto no nos habría ofrecido una estimación del la varianza.

Por tanto, la salida obtenida a través de *SAS* se muestra en la figura 3, la cual indica el valor predicho, así como una estimador de la desviación típica respecto de la media en ese punto. Estos valore se muestran en las ecuaciones (5) y (6).

La interpretación que se le puede dar a dichos resultados es que para el mantenimiento de 5 copiadoras se dedica en torno a 74.6 minutos con una variación media de 1.15 minutos.

The REG Procedure
Model: MODEL1
Dependent Variable: Y

Output Statistics

Obs	X	Dependent Variable	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual	Cook's D
1	2	20	29.4903	2.0061	-9.4903	8.685	-1.093	0.032
2	4	60	59.5608	1.4331	0.4392	8.798	0.050	0.000
45	5	77	74.5961	1.3298	2.4039	8.814	0.273	0.001
46	5	.	74.5961	1.3298	.	.	.	.

Figura 3: *Salida SAS*: Mantenimiento de Copiadoras - Prediccion de Regresión Lineal Simple para $X = 5$

$$E\left[\widehat{Y} \mid X = 5\right] = E\left[\widehat{\beta}_0 + \widehat{\beta}_1 X + \varepsilon \mid X = 5\right] = -0.58016 + 15.03525 * 5 + 0 = 74.5961 \tag{5}$$

$$Var\left[\widehat{Y} \mid X = 5\right] = \sigma^2\left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right) = 1.1531^2 = 1.3298 \tag{6}$$

1.24. Refer to Copier maintenance Problem 1.20.

a. Obtain the residuals e_i and the sum of the squared residuals $\sum_i e_i^2$. What is the relation between the sum of the squared residuals here and the quantity Q in (7)?

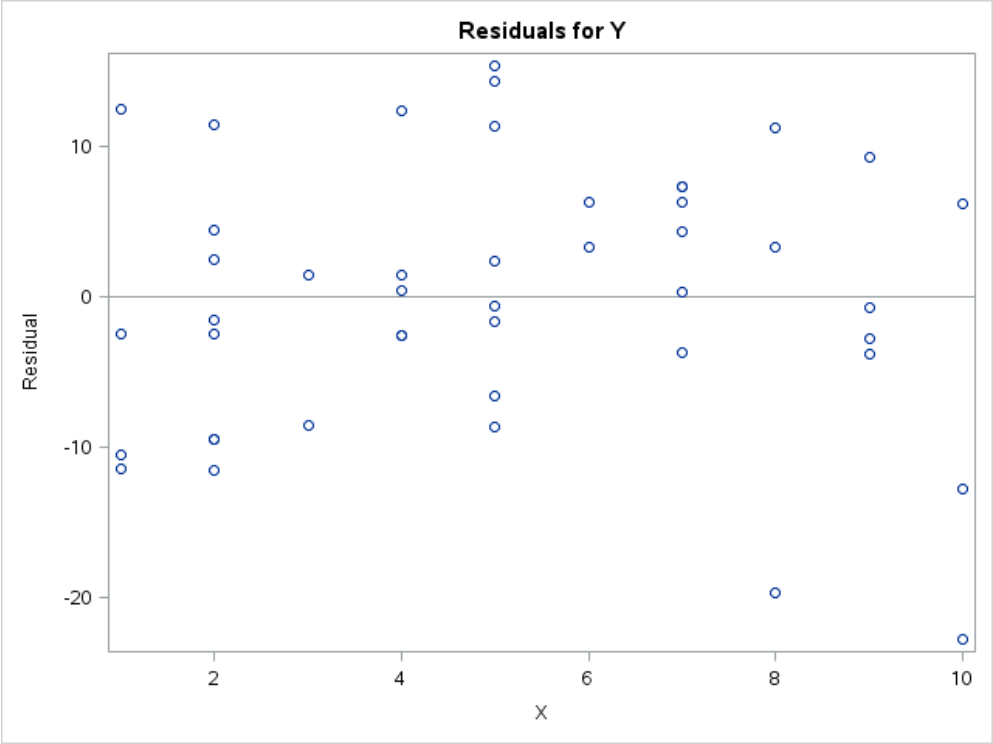

Figura 4: *Salida SAS*: Mantenimiento de Copiadoras - Gráfico de Residuos de Regresión Lineal Simple

En este apartado se pide obtener los residuos e_i. Para ello, se ha creido conveniente la representación de un gráfico de resiudos, el cual ha sido obtenido mediante el fragmento de código de la figura 15 utilizado en

apartados anteriores.

Este gráfico de residuos se muestra en la figura 4 y a partir de él se puede apreciar una distribución más o menos uniforme de los errores (tiene cierta forma de parábola, pero esta característica es sutil). También se puede apreciar la dispersión de los datos en torno al valor 10, tal y como se indicó en partaados anteriores.

$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 = 3416.37702 \tag{7}$$

En el título del apartado, también se pide el valor de la suma de errores al cuadrado $\sum_i e_i^2$, así como su relación con el valor Q, que se muestra en la ecuación (7). El valor Q es aquel que se trata de minimizar bajo el criterio de mínimos cuadrados en el modelo de regresión simple. Este ha sido extraido del resumen generado por *SAS*, que se muestra en la figura 1 utilizada en otros apartados.

Por tanto, es la medida del error en términos de mínimos cuadrados de la recta de regresión obtenida con respecto del conjunto de datos. sto puede escribirse como $Q = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2$, y dado que los errores e_i se definen como $Y_i - \widehat{Y}_i$ la equivalencia es directa, es decir, $Q = \sum_i e_i^2$.

b. Obtain point estimates of σ^2 and σ. In what units is σ expressed?

En este apartado se han pedido obtener estimaciones acerca de la dispersión de los errores, es decir, del valor σ. Al igual que en anteriores apartados, este valor se ha obtenido a partir del código *SAS* de la figura 15, que ha generado los resultados obtenidos en la figura 1.

La varianza y la desviación típica de los errores se muestran en las ecuaciones (8) y (9) respectivamente. En el título del apartado se indica además que se explique cuáles son las unidades de medida de la desviación típica de los errores σ. Es sencillo entender que las unidades de esta medida serán en *minutos*, ya que esta mide la dispersión de los datos en torno a la predicción obtenida por la recta de regresión, la cual se refiere a la variable dependiente Y, que representa el número de minutos en realizar el mantenimiento.

$$\widehat{\sigma}^2 = MSE = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2}{n-2} = \frac{3416.37702}{45-2} = 79.45063 \tag{8}$$

$$\widehat{\sigma} = \sqrt{\widehat{\sigma^2}} = \sqrt{79.45063} = 8.91351 \tag{9}$$

2.5. Refer to Copier maintenance Problem 1.20.

a. Estimate the change in the mean service time when the number of copiers serviced increases by one. Use a 90 % confidence interval. Interpret your confidence interval.

El enunciado de este apartado pide que se obtenga un intervalo de confianza del 90 % para la estimación de la pendiente de la regresión β_1. Para ello se ha utilizado la salida obtenida por el fragmento de código de la figura 15, que se muestra en la figura 1.

$$t_{n-2;1-\frac{\alpha}{2}} = t_{43;0.95} = 1.681071 \tag{10}$$

$$Var\left[\widehat{\beta}_1\right] = 0.23337 \tag{11}$$

En esta se aparece el intervalo de confianza al 90 % para el valor β_1, el cual se muestra en la ecuación (12). Este se construye a partir de la estimación $\widehat{\beta}_1 = 15.03525$, cuyo valor se relaja a partir de un determinado error de estimación $t_{n-2;1-\frac{\alpha}{2}} \sqrt{Var\left[\widehat{\beta}_1\right]} = 1.681071 * \sqrt{0.23337} = 0.8121$.

$$\text{I.Conf.} = \left[\widehat{\beta}_1 \pm t_{n-2;1-\frac{\alpha}{2}} \sqrt{Var\left[\widehat{\beta}_1\right]}\right] = [15.03525 \pm 0.8121] = [14.2232, 15.8474] \tag{12}$$

La interpretación de dicho intervalo es la siguiente: Con una seguridad del 90 % el tiempo medio de mantenimiento de cada copiadora se encuentra entre 14.2232 y 15.8474 minutos.

b. Conduct a *t-test* to determine whether or not there is a linear association between X and Y here; control the α risk at 0.1. State the alternatives, decision rule, and conclusion. What is the *P-value* of your test?

En este apartado se pide realizar un test de hipótesis acerca de la existencia de asociación entre X e Y. Es decir, si la pendiente de nuestra recta de regresión puede ser considerada nula. Para ello, se puede modelar el test de hipótesis tal y como se indica en la ecuación (13).

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_1 &: \beta_1 \neq 0 \end{aligned} \tag{13}$$

En este caso se pide realizar el test de hipótesis mediante el procedimiento de *test t* (ya que en la sección b del ejercicio 2.24 se pide realizar el mismo test, pero esta vez a partir de la distribución F, para comprobar su equivalencia). El *p-valor* de este test se obtiene tal y como se indica en la ecuación (14). Por ser un test de *2 vías* se toma el valor absoluto y se multiplica por 2 el valor de probabilidad ($\alpha/2$ a cada lado). Para ello se ha utilizado la salida obtenida por el fragmento de código de la figura 15, que se muestra en la figura 1.

$$\text{p-value} = 2Pr\left[t_{n-2} > \left|\frac{\widehat{\beta}_1 - 0}{\sqrt{Var\left[\widehat{\beta}_1\right]}}\right|\right] = 2Pr\left[t_{43} > \frac{15.03525}{0.48309}\right] = 2Pr\left[t_{43} > 31.12\right] < 0.0001 \tag{14}$$

El *p-valor* obtenido es muy próximo a 0, por lo que a nivel $\alpha = 0.1$ nos vemos obligados a rechazar la hipótesis de que β_1 es nulo. Esto quiere decir por tanto que asumiremos que existe relación lineal entre el X e Y.

c. Are your results in parts (a) and (b) consistent? Explain.

En este caso se nos pide analizar los resultados obtenidos en los dos últimos apartados para comprobar si son coherentes entre sí, ya que ambos se refieren a procesos de inferencia sobre la estimación de $\widehat{\beta}_1$.

Los resultados si que son coherentes entre sí, ya que en el primero de ellos se ha calculado un intervalo de confianza (o aceptación) para el valor de la pendiente β_1 de la regresión, mientras que en el siguiente caso se ha realizado un test sobre el valor de dicha pendiente (si se puede tomar como nula). Puesto que en el primer caso el límite izquierdo del intervalo se encuentra en torno al valor 14 con un nivel de confianza del 90 %, es razonable que sobre el mismo nivel de confianza se tenga que rechazar la hipótesis de que el valor 0 está incluido en dicho intervalo de aceptación.

d. The manufacturer has suggested that the mean required time should not increase by more than 14 minutes for each additional copier that is serviced on a service call. Conduct a test to decide whether this standard is being satisfied by Tri-City. Control the risk of a Type I error at 0.05. State the alternatives, decision rule, and conclusion. What is the *P-value* of the test?

En este caso se pide realizar un test de hipótesis acerca de un intervalo, referido una vez más al valor de la pendiente de la relación β_1. Un fabricante quiere comprobar si es razonable asumir que el tiempo de mantenimiento de una copiadora es menor que 14 minutos con un nivel de confianza del 95 %. Por tanto. el test de hipótesis se define tal y como se indica en la ecuación (15).

$$\begin{aligned} H_0 &: \beta_1 \leq 14 \\ H_1 &: \beta_1 > 14 \end{aligned} \tag{15}$$

En este caso para la realización del *test t*, ya no hay que selecciona $\alpha/2$ a cada lado, ya que que el test es de *1 vía*. El *p-valor* del test se ha obtenido tal y como se indica en la ecuación (16).

$$\text{p-value} = Pr\left[t_{n-2} > \frac{\widehat{\beta}_1 - 14}{\sqrt{Var\left[\widehat{\beta}_1\right]}}\right] = Pr\left[t_{43} > \frac{1.03525}{0.48309}\right] = Pr\left[t_{43} > 2.14297\right] = 0.01890824 \tag{16}$$

El *p-valor* en este caso, no es tan próximo a 0 como en anteriores tests. Sin embargo, con un nivel de confianza del 95 % nos vemos obligados a rechazar la hipótesis de que β_1 sea menor que 14. Nótese que esto no se puede rechazar con un nivel de confianza del 99 %.

e. Does $\widehat{\beta}_0$ give any relevant information here about the start-up time on calls-i.e., about the time required before service work is begun on the copiers at a customer location?

En este apartado se pide realizar un test de hipótesis que permita comprobar si el valor β_0 es nulo. Esto puede interpretarse como que el tiempo para realizar 0 mantenimientos de copiadoras también es 0. El test se plantea por tanto como se indica en la ecuación (17).

$$\begin{aligned} H_0 &: \beta_0 = 0 \\ H_1 &: \beta_0 \neq 0 \end{aligned} \tag{17}$$

Para la realización de dicho test de hipótesis se ha calculado el *p-valor* tal y como se indica en la ecuación (18), que tal y como ocurría en el análogo de β_1, se plantea como un test de *2 vías* por lo que se toma el valor absoluto y se multiplica por 2 (debido a la simetría de la distribución t). Para ello se ha utilizado la salida obtenida por el fragmento de código de la figura 15, que se muestra en la figura 1.

$$\text{p-value} = 2Pr\left[t_{n-2} > \left|\frac{\widehat{\beta}_0 - 0}{\sqrt{Var\left[\widehat{\beta}_0\right]}}\right|\right] = 2Pr\left[t_{43} > \left|\frac{-0.58016}{2.80394}\right|\right] = 2Pr\left[t_{43} > 0.21\right] = 0.8371 \tag{18}$$

Debido al elevado valor obtenido en el *p-valor* estamos en condiciones suficientes como para aceptar la hipótesis nula de la ecuación (17), que nos indica que el valor de la ordenada en el origen es 0. Esto puede interpretarse como que cuando no se realiza ningún mantenimiento, tampoco se dedica ningún minuto.

Sin embargo, no tiene ningún sentido eliminar el estimador $\widehat{\beta}_0$ del modelo, puesto que su valor (muy próximo a cero pero no cero) permite que la recta haya sido ajustada a los datos de manera óptima, y su eliminación tan solo perjudicaría la minimización cuadrática del error del modelo.

2.14. Refer to Copier maintenance Problem 1.20.

Output Statistics												
Obs	X	Dependent Variable	Predicted Value	Std Error Mean Predict	90% CL Mean		90% CL Predict		Residual	Std Error Residual	Student Residual	Cook's D
1	2	20	29.4903	2.0061	26.1180	32.8627	14.1313	44.8494	-9.4903	8.685	-1.093	0.032
2	4	60	59.5608	1.4331	57.1517	61.9699	44.3842	74.7375	0.4392	8.798	0.050	0.000
45	5	77	74.5961	1.3298	72.3605	76.8316	59.4460	89.7462	2.4039	8.814	0.273	0.001
46	6	.	89.6313	1.3964	87.2839	91.9788	74.4643	104.7983	.	.	.	.

Figura 5: *Salida SAS*: Mantenimiento de Copiadoras - Prediccion de Regresión Lineal Simple para $X = 6$

a. Obtain a 90 % confidence interval for the mean service time on calls in which six copiers are serviced. Interpret your confidence interval.

En este apartado, se pide obtener un intervalo de confianza del 90 % para la media de la variable dependiente Y cuando la variable dependiente toma el valor $X = 6$. Para ello, se ha utilizado el fragmento de código *SAS* incluido en la figura 17, a partil del cual se ha obtenido la salida de la figura 5.

$$E[\widehat{Y}_h] = E[\widehat{\beta}_0 + \widehat{\beta}_1 * 6 + \varepsilon_h] = 89.6313 \tag{19}$$

$$t_{n-2;1-\frac{\alpha}{2}} = t_{43;0.95} = 1.681071 \tag{20}$$

$$Var\left[\widehat{Y}_h\right] = 1.9499 \tag{21}$$

El resultado del intervalo de confianza para la media se muestra en la ecuación (22).

$$\text{I.Conf.} = \left[E[\widehat{Y}_h] \pm t_{n-2;1-\frac{\alpha}{2}} \sqrt{Var\left[\widehat{Y}_h\right]}\right] = [89.6313 \pm 2.3475] = [87.2839, 91.9788] \tag{22}$$

b. Obtain a 90 % prediction interval for the service time on the next call in which six copiers are serviced. Is your prediction interval wider than the corresponding confidence interval in part (a)? Should it be?

En este apartado, se pide obtener un intervalo de predicción del 90 % para la media de la variable dependiente Y cuando la variable dependiente toma el valor $X = 6$. Para ello, se ha utilizado el fragmento de código *SAS* incluido en la figura 17, a partil del cual se ha obtenido la salida de la figura 5.

$$E[\widehat{Y}_{pred}] = E[\widehat{\beta}_0 + \widehat{\beta}_1 6 + \varepsilon_i] = 89.6313 \tag{23}$$

$$t_{n-2;1-\frac{\alpha}{2}} = t_{43;0.95} = 1.681071 \tag{24}$$

$$Var\left[\widehat{Y}_{pred}\right] = \widehat{\sigma}^2 + Var\left[\widehat{Y}_h\right] = 79.45063 + 1.9499 = 81.40053 \tag{25}$$

El resultado del intervalo de predicción se muestra en la ecuación (26).

$$\text{I.Pred.} = \left[E[\widehat{Y}_{pred}] \pm z_{1-\frac{\alpha}{2}} \sqrt{Var\left[\widehat{Y}_{pred}\right]}\right] = [89.6313 \pm 15.167] = [74.4643, 104.7983] \tag{26}$$

Tal y como se puede apreciar, el intervalo de predicción es mucho más amplio que el intervalo de confianza para la media. Esto tiene sentido ya que el intervalo de predicción tiene en cuenta la dispersión generada en cada observación, es decir, su varianza tiene un $\widehat{\sigma}^2$ "mas" que el intervalo de confianza.

2.24. Refer to Copier maintenance Problem 1.20.

b. Conduct an *F-test* to determine whether or not there is a linear association between time spent and number of copiers serviced; use $\alpha = 0.1$. State the alternatives, decision rule, and conclusion.

En este caso, se pide realizar un test de la F para comprobar si existe relación de dependencia entre la variable dependiente Y y la variable dependiente X. Por tanto, esto se puede modelar de la misma manera que se hizo en el apartado b del ejercicio 2.5.

El test de hipótesis que se realizará se muestra en la ecuación (27), el cual es equivalente al del apartado previamente citado. Sin embargo, la manera de proceder para la realización del test, en este caso será diferente. Tal y como se indicó anteriormente, mediante el *test-t* se prueba que la variable de interés tome un determinado valor (lo cual lo hace más general que el que se ha realizado en esta sección).

En este caso, en lugar de realizar el test directamente sobre la estimación de β_1, se realiza sobre la variación recogida por el modelo. En concreto, se compara la relación entre la variación recogida por el modelo y la aleatoria con la distribución F con 1 y $n-2$ grados de libertad, para así obtener el *p-valor* del test.

Se puede demostar que los resultados obtenidos sobre el test de existencia de correlación entre X e Y son equivalente entre el *test t* y el *test F*

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_1 &: \beta_1 \neq 0 \end{aligned} \tag{27}$$

En este caso, se ha obtenido un *p-valor* muy próximo a cero, tal y como se indica en la ecuación (28). Dado que el valor α se ha fijado en 0.1, nos vemos obligados a rechazar la hipótesis de que el valor de β_1 es igual a cero. Este resultado es equivalente a obtenido para el *test t*

$$\text{p-value} = Pr\left[F_{1;n-2} > \frac{MSM}{MSE}\right] = Pr\left[F_{1;43} > \frac{76960}{79.45063}\right] = Pr\left[F_{1;43} > 968.66\right] < 0.0001 \tag{28}$$

c. By how much, relatively, is the total variation in number of minutes spent on a call-reduced when the number of copiers serviced is introduced into the analysis? Is this a relatively small or large reduction? What is the name of this measure?

En este apartado se pide estudiar el grado de variación recogida por el modelo. Dicho grado de variación se puede estudiar a partir del coeficiente de determinación R^2, que representa el ratio de variación en términos de suma de cuadrados entre el modelo y el total del conjunto de datos. Por tanto, este siempre toma valores en el intervalo $[0, 1]$.

En este caso, el modelo recoge el 95.74 % de la variación del conjunto de datos. Por tanto, se cree que dicho ajuste ha sido acertado. La expresión del coeficiente de determinación se muestra en la ecuación (29). Esta ha sido extraida de la figura 1, generada a partir del fragmento de código de la figura 15 utilizado en otros apartados.

$$R^2 = \frac{SSM}{SST} = \frac{76960}{76960 + 3416.37702} = 0.9574 \tag{29}$$

d. Calculate r and attach the appropriate sign.

En este apartado, se pide obtener el coeficiente de correlación entre las variables. Esta medida sirve para cuantizar el grado de relación entre variables, así como si esta relación es positiva o negativa. El coeficiente de correlación toma valores en el interva $[-1, 1]$ indicando valores proximos al 1 una relación lineal fuerte positiva y -1 una relación lineal fuerte negativa, mientras que los valores próximos al 0 indican una relación débil entre las variables.

Este coeficiente se puede definir como $r = \frac{Cov[X,Y]}{\sqrt{Var[X]Var[Y]}}$, aunque en este caso se ha obtenido mediante la propiedad de la ecuación (30). Cuando se utiliza esta ecuación es necesario ajustar el signo de manera "manual", ya que el coeficiente de determinación R^2 tan solo toma valores positivos.

$$r = \pm\sqrt{R^2} = +0.9785 \tag{30}$$

En este caso, tras analizar los resultados de la ecuación (30) y aplicar el signo positivo (basta con consultar el gráfico de la regresión de la figura 2 para entender que la relación es positiva), podemos asegurar que existe una fuerte relación entra estas variables.

Este valor es coherente con la interpretación del conjunto de datos, que relaciona el tiempo de realización de un servicio (en este caso en minutos) con el número de unidades de trabajo que hay que desempeñar en dicho servicio (en este caso el número de copiadoras a las que realizar tareas de mantenimiento).

Parte II
Ejercicios Montgomery:

2.25. En la tabla B.1 del apéndice aparecen datos sobre el desempeño de los 26 equipos de la Liga Nacional de Fútbol en 1976. Se cree que la cantidad de yardas ganadas por tierra por los contrarios (x_8) tiene un efecto sobre la cantidad de juegos que gana un equipo (y).

Los datos se han importado del *.xls* disponible en la página web de la asignatura, tras lo cual se ha realizado un preprocesado, eliminando del dataset las columnas x_{1-7} y x_9, ya que estas no son necesarias para realizar el ejercicio. Esto se ha realizado mediante el fragmento de código 18.

a. Ajustar un modelo de regresión lineal simple que relacione los juegos ganados, y, con las yardas ganadas por tierra por los contrarios, x_8.

$$y_i = \beta_0 + \beta_1 x_{8i} + \varepsilon_i \tag{31}$$

Analizaremos el modelo de la ecuación 31 mediante un estudio de regresión lineal simple, con y como variable dependiente, x_8 como variable independiente, β_0 el intercepto, y β_1 la pendiente. ε es el error aleatorio.

El fragmento de código 19 permite estimar los valores del intercepto y la pendiente, y analizar las hipótesis nulas:

$$H_0 : \beta_0 = 0 \tag{32}$$
$$H_0' : \beta_1 = 0 \tag{33}$$

La hipótesis más importante en la regresión es la que incumbe a β_1, ya que si la pendiente de la recta es 0, no existe regresión, y se trata de una población simple sobre la que x_8 no tiene efecto.

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	21.78825	2.69623	8.08	<.0001
x8	x8	1	-0.00703	0.00126	-5.58	<.0001

Figura 6: Estimadores y p-valores para la regresión.

Como podemos ver, el *p-valor* para β_1 está por debajo de $0,05$, por lo que podemos rechazar la hipótesis nula con una confianza del $95\,\%$, y podemos afirmar que existe regresión. El *p-valor* para el intercepto también nos permite rechazar la hipótesis nula, que significaría que se puede ajustar un modelo sin intercepto.

Por tanto, los estimadores son:

$$\hat{\beta}_0 = 21.78825 \tag{34}$$
$$\hat{\beta}_1 = -0.00703 \tag{35}$$

Con la recta de regresión:

$$\hat{y}_i = 21.78825 - 0.00703x_{8i} + \varepsilon_i \tag{36}$$

La interpretación de estos parámetros es que con 0 yardas ganadas por el contrario, un equipo gana de media $21,78825$ juegos, y que por cada yarda que gana el contrario, el equipo pierde de media 0.00703 juegos. Esta pendiente puede parecer muy cercana a 0, pero debemos considerar que las muestras tienen una magnitud de *miles de yardas*, por lo que la pendiente vale 7.03 juegos perdidos, si se midiera x_8 en la unidad de *miles de yardas*.

Figura 7: Fit plot del modelo de regresión lineal simple

El *fitplot* nos permite analizar cómo de bien se ajusta la muestra a la recta de regresión hallada. Como podemos ver, la distribución de las muestras tiene una forma clara de recta. Podemos ver que los valores tienen bastante variabilidad, bastante mayor en los extremos que en el centro de la recta.

b. Formar la tabla de análisis de varianza y probar el significado de la regresión.

La tabla de análisis de la varianza se obtiene también mediante el fragmento de código 19.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	178.09231	178.09231	31.10	<.0001
Error	26	148.87197	5.72585		
Corrected Total	27	326.96429			

Figura 8: Tabla de análisis de la varianza

Esta tabla nos permite obtener la suma de cuadrados del modelo y del error, lo cual nos da un modo alternativo de hacer el contraste sobre el coeficiente β_1. *Mean Square* para el *Error* es el estimador de la varianza σ^2 cuando la muestra se trata de una sola población $y = \beta_0 + \varepsilon$. Si se cumple la hipótesis $H_0 : \beta_1 = 0$, *Mean Square* para el *Model*, que estima la varianza con el modelo completo, también es un estimador de la varianza σ^2. Por tanto

$$F = \frac{MSM}{MSE} = 31.10 \tag{37}$$

es un estadístico que permite realizar el test de hipótesis sobre la hipótesis nula H_0. Si este estadístico F es mayor que la distribución F con 1 y 26 grados de libertad, entonces podemos considerar que las sumas de errores medias son diferentes, y por tanto rechazar la hipótesis nula.

Esta tabla también nos proporciona el *p-valor* para este contraste, es decir, $Pr(F(1;26) > 31.1)$ que es menor que $0,0001$, por lo que podemos rechazar la hipótesis, y afirmar que existe regresión con un 95 % de confianza.

c. Determinar un intervalo de confianza de 95 % para la pendiente.

La pendiente es el parámetro β_1, y podemos obtener los intervalos de confianza para los parámetros mediante el fragmento de código 19, habiendo añadido la opción *CLB* al modelo.

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	Intercept	1	21.78825	2.69623	8.08	<.0001	16.24606	27.33044
x8	x8	1	-0.00703	0.00126	-5.58	<.0001	-0.00961	-0.00444

Figura 9: Estimadores para los parámetros con intervalos de confianza del 95 %

Estos intervalos se calculan mediante la siguiente fórmula:

$$IC = \left[\hat{\beta}_1 \pm t_{1-\frac{\alpha}{2};n-2}\sqrt{Var[\hat{\beta}_1]}\right] \tag{38}$$

$$\sqrt{Var[\hat{\beta}_1]} = StandardError(x_8) = 0.00126 \tag{39}$$

$$t_{1-\frac{\alpha}{2};n-2} = t_{0.975;26} = -2.05552944 \tag{40}$$

$$IC = [-0.00703 \pm -2.05552944 * 0.001126] = [-0.009619967, -0.004440033] \tag{41}$$

Por tanto la pendiente de la recta de regresión está entre -0.00961 y -0.00444 con una confianza del 95 %.

d. ¿Qué porcentaje de variabilidad total da y, y explica este modelo?

Para obtener la medida de variabilidad total de y debemos volver a la tabla de análisis de varianza del apartado **b**:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	178.09231	178.09231	31.10	<.0001
Error	26	148.87197	5.72585		
Corrected Total	27	326.96429			

Figura 10: Tabla de análisis de la varianza

Esta tabla nos da la *Suma de cuadrados total*, que en este caso es igual a $326,96429$. Para obtener la variabilidad de la muestra, conocida como MST, se divide este valor por sus grados de libertad, que son $N-1$, siendo N el número de observaciones.

$$MST = \frac{SST}{N-1} = \frac{326,96429}{27} = 12.109788519 \tag{42}$$

Este valor MST es la variabilidad total de la muestra σ_T^2, parte de la cual intentamos explicar mediante el modelo de regresión lineal simple. Existirá otra parte explicada por el error aleatorio, σ^2.

Para calcular la cantidad de variabilidad que explica el modelo podemos calcular:

$$\frac{SS_{Modelo}}{SS_{Total}} \tag{43}$$

Que es un coeficiente conocido como R^2, que también podemos obtener directamente con la siguiente tabla obtenida del fragmento de código 19:

Root MSE	2.39287	R-Square	0.5447
Dependent Mean	6.96429	Adj R-Sq	0.5272
Coeff Var	34.35921		

Figura 11: Coeficientes de la Regresión Lineal Simple.

R-square indica que el modelo de regresión explica un 54.47 % de la variabilidad. También puede interpretarse como un indicador de cómo de bien se ajustan los datos a la línea de regresión ajustada.

e. Determinar un intervalo de confianza de 95 % para la cantidad promedio de juegos ganados, si la distancia ganada por tierra por los contrarios se limita a 2000 yardas.

Este apartado nos pide hacer una predicción de y cuando $x_8 = 2000$. Este valor no se encuentra entre nuestras observaciones, pero si que está dentro del rango definido por el mínimo y el máximo de x_8 para las observaciones, por lo que podremos realizar esta predicción. Añadiremos manualmente a los datos una observación en la que solo rellenamos el valor de x_8 mediante el fragmento de código 20. *SAS* realizará la predicción sobre el valor de la variable dependiente y para este valor de x_8 dado el modelo de regresión conseguido mediante las observaciones completas.

Para mostrar los intervalos de confianza para las predicciones, utilizamos el fragmento de código 21, con la opción *CLI* en el modelo.

Output Statistics						
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Predict		Residual
⋮						
29	.	7.7381	0.4730	2.7242	12.7519	.

Figura 12: Intervalo de confianza del 95 % para la predicción de $x_8 = 2000$.

Como podemos ver, el valor predicho medio es $7,7381$ juegos ganados, y podemos afirmar con una confianza del 95 % que el valor real está entre $2,7242$ y $12,7519$ juegos ganados.

2.26. Supóngase que se quiere usar el modelo desarrollado en el problema 2.1 para pronosticar la cantidad de juegos que ganará un equipo si puede limitar los avances por tierra de sus contrarios a 1800 yardas. Determinar un estimado de punto de la cantidad de juegos ganados cuando $x_8 = 1800$. Determinar un intervalo de predicción de 90 por ciento para la cantidad de juegos ganados.

Como el apartado **2.1.e**, este ejercicio requiere hacer una predicción de y. El valor de x_8 pedido no está en las observaciones, pero se encuentra en el rango definido por el mínimo y el máximo de observaciones para x_8. Añadimos una observación con valor $x_8 = 1800$, al igual que hicimos para el anterior apartado, con el fragmento de código 21 para el cual *SAS* realizará una predicción de y dado el modelo de regresión simple. Para conseguir el intervalo del 90 % de confianza, definimos en las opciones del modelo el *ALPHA* como igual a 0.1.

Output Statistics						
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	90% CL Predict		Residual
30	.	9.1431	0.5976	4.9364	13.3497	.

Figura 13: Intervalo de confianza del 90 % para la predicción de $x_8 = 1800$.

El estimado de punto es $9,1431$ juegos ganados, y el intervalo nos permite afirmar con un 90 % de confianza que el valor real está entre $4,9364$ y $13,3497$ juegos ganados.

Parte III
Código Fuente

```
filename reffile '/folders/myshortcuts/sas/regression-group-task/data/CH01PR20.csv';

proc import datafile=reffile dbms=csv out=copiers;
  getnames=yes;
run;

proc print data=copiers;
run;
```

Figura 14: *Código SAS:* Mantenimiento de Copiadoras - Importación del conjunto de datos.

```
proc reg data=copiers;
  model y=x /clb alpha=0.1;
  id x;
run;
```

Figura 15: *Código SAS:* Mantenimiento de Copiadoras - Modelo de regresión simple.

```
data copiers_new_observation;
  x = 5;
run;

proc append base=copiers data=copiers_new_observation;
run;

proc reg data=copiers;
  model y = x /r clm;
  id x;
run;
```

Figura 16: *Código SAS:* Mantenimiento de Copiadoras - Predicción para $X = 5$.

```
data copiers_new_observation;
  x = 6;
run;

proc append base=copiers data=copiers_new_observation;
run;

proc reg data=copiers;
  model y = x /r cli clm alpha=0.1;
  id x;
run;
```

Figura 17: *Código SAS:* Mantenimiento de Copiadoras - Predicción para $X = 6$.

```
FILENAME REFFILE '/folders/myfolders/data-table-B1.XLS';

PROC IMPORT DATAFILE=REFFILE
  DBMS=XLS
  OUT=WORK.IMPORT;
  GETNAMES=YES;
RUN;

DATA FUTBOL;
  SET IMPORT;
  DROP X1 X2 X3 X4 X5 X6 X7 X9;
RUN;
```

Figura 18: *Código SAS:* Ejercicio 2.1 del Montgomery - Importación y preprocesado del conjunto de datos.

```
PROC REG DATA=FUTBOL;
  MODEL Y=X8/CLB;
RUN;
```

Figura 19: *Código SAS:* Ejercicio 2.1 del Montgomery - Programa de Regresión Lineal Simple con intervalo para los parámetros.

```
DATA AUX;
  INPUT X8;
  CARDS;
  2000
  1800
RUN;

PROC APPEND BASE=FUTBOL DATA=AUX; RUN;
```

Figura 20: *Código SAS:* Ejercicio 2.1 y 2.2 del Montgomery - Addición de valores para predicción.

```
PROC REG DATA=FUTBOL;
  MODEL Y=X8/CLI;
RUN;
```

Figura 21: *Código SAS:* Ejercicio 2.1 del Montgomery - Obtención de valores predichos con intervalos de confianza del 95 %.

```
PROC REG DATA=FUTBOL;
  MODEL Y=X8/CLI ALPHA=0.1;
RUN;
```

Figura 22: *Código SAS:* Ejercicio 2.2 del Montgomery - Obtención de valores predichos con intervalos de confianza del 90 %.

Referencias

[1] Barba Escribá, L. Regresión y ANOVA, 2017/18. Facultad de Ciencias: Departamento de Estadística.

[2] Montgomery, D. C., Peck, E. A., and Vining, G. G. *Introduction to linear regression analysis*, vol. 821. John Wiley & Sons, 2012.

[3] Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. *Applied linear statistical models*, vol. 4. Irwin Chicago, 1996.

[4] SAS® Software Institute. Sas. https://www.sas.com/.