Regresión: Tarea por Grupos

Cuervo Fernández, Esther García Prado, Sergio Martín Villares, Pablo

27 de noviembre 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 6 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.

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

- Y_i:
- β_0:
- β_1:
- X_i:
- ϵ_i:

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 7, 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.

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

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

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

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

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

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

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 7) 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 regresió 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

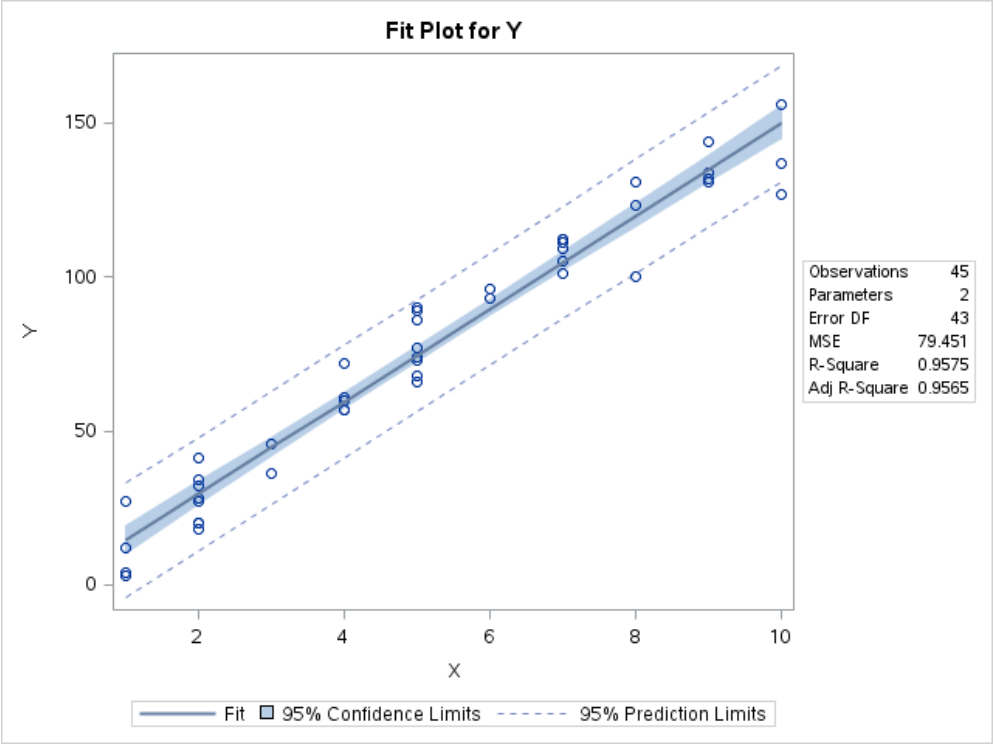

Figura 2: *Salida SAS*: Mantenimiento de Copiadoras - Gráfico de Regresión Lineal Simple

término independiente ha sido ajustado con valor negativo. Esto no puede tener ninguna interpretación, 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 8. 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.

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

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

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 $X = 5$

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)?

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 7 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.

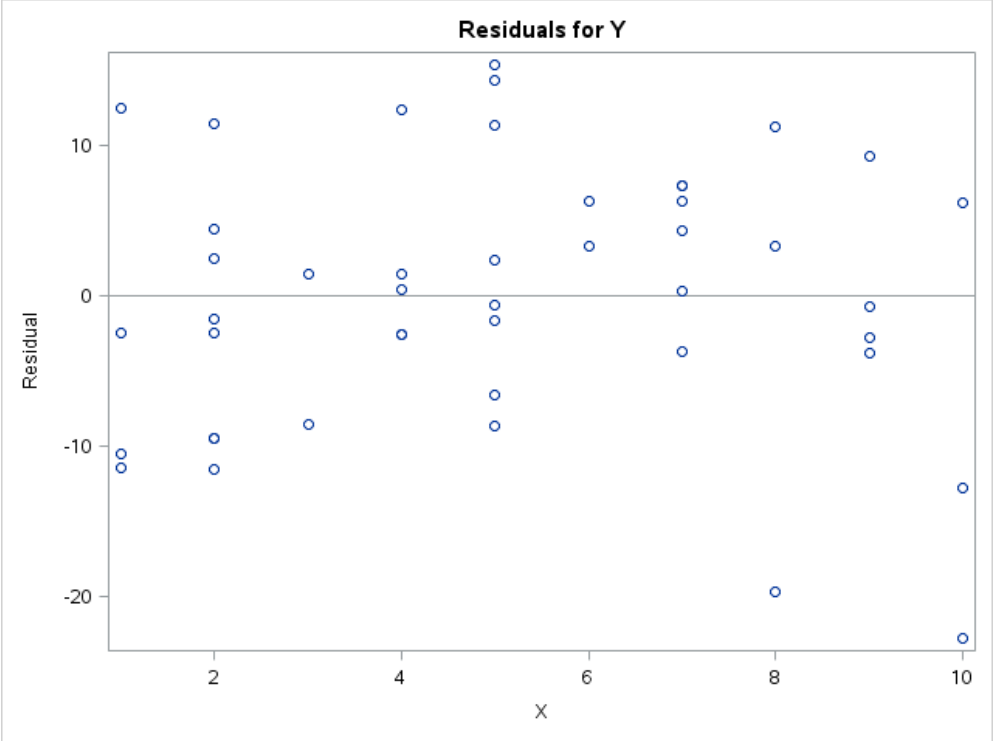

Figura 4: *Salida SAS*: Mantenimiento de Copiadoras - Gráfico de Residuos de Regresión Lineal Simple

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

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 7, 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.

[TODO]

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

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

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

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 0.10. State the alternatives, decision rule, and conclusion. What is the *P-value* of your test?

[TODO]

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

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

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

[TODO]

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?

[TODO]

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

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

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?

[TODO]

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

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

2.14. Refer to Copier maintenance Problem 1.20.

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

[TODO]

$$\widehat{Y}_h = \widehat{\beta}_0 + \widehat{\beta}_1 6 + \epsilon_h = 89.6313 \tag{19}$$

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

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

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

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 $X = 6$

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?

[TODO]

$$\widehat{Y}_{pred} = \widehat{\beta}_0 + \widehat{\beta}_1 6 + \epsilon_i = 89.6313 \tag{23}$$

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

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

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

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*

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

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 7 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.

[TODO]

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

Parte II
Ejercicios Montgomery:

[TODO]

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

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

Figura 7: *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 8: *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 9: *Código SAS:* Mantenimiento de Copiadoras - Predicción para $X = 6$.

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/.