Regresión: Tarea por Grupos

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

27 de noviembre de 2017

Parte I

Ejercicios Montgomery:

2.1. 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).

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.

$$y_i = \beta_0 + \beta_1 x_{8i} + \varepsilon_i \tag{1}$$

Analizaremos el modelo de la ecuación 1 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 procedimiento *REG* permite estimar los valores del intercepto y la pendiente, y analizar las hipótesis nulas:

$$H_0 : \beta_0 = 0 \tag{2}$$
$$H_0' : \beta_1 = 0 \tag{3}$$

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 1: 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 $95\,\%$. El *p-valor* para el intercepto también permite rechazar la hipótesis nula.

Por tanto, los estimadores son:

$$\hat{\beta}_0 = 21.78825 \tag{4}$$
$$\hat{\beta}_1 = -0.00703 \tag{5}$$

Con la recta de regresión:

$$y_i = 21.78825 - 0.00703x_{8i} + \varepsilon_i \tag{6}$$

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.

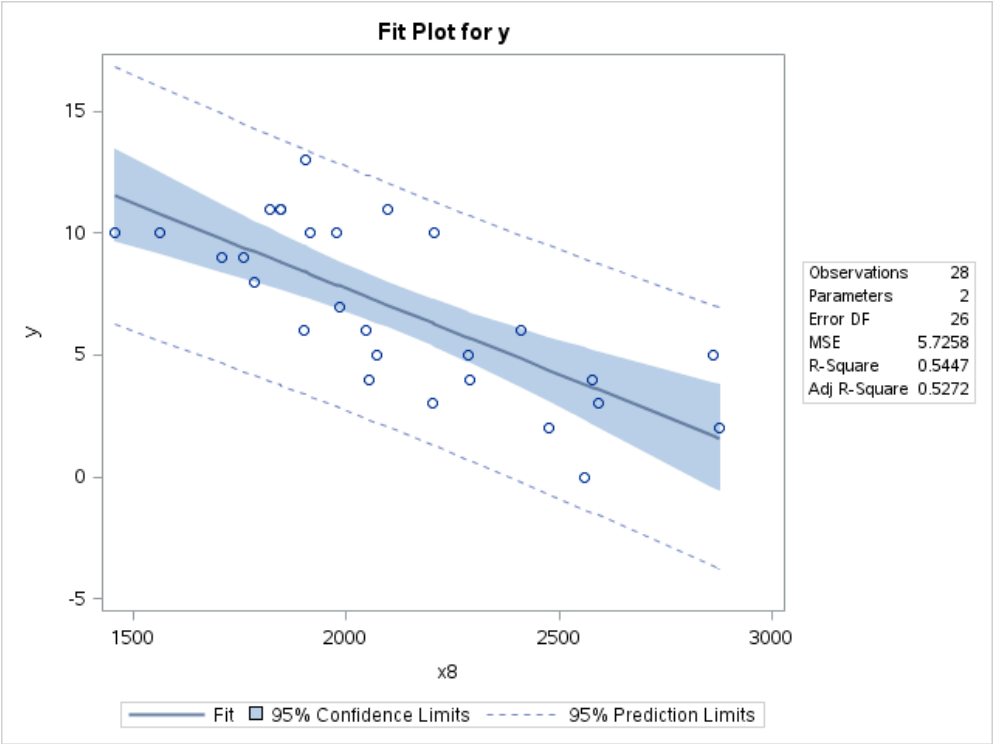

Figura 2: Fit plot del modelo de regresión lineal simple

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 procedimiento *REG*.

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 3: 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 obtenida por el error aleatorio ε. Si se cumple la hipótesis $H_0 : \beta_1 = 0$, *Mean Square* para el *Model* también es un estimador de la varianza σ^2. Por tanto

$$F = \frac{MSM}{MSE} \tag{7}$$

es un estadístico que permite realizar el test de hipótesis sobre la hipótesis nula H_0. Esta tabla también nos proporciona el *p-valor* para este contraste, que es menor que $0,05$, 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 procedimiento *REG*, añadiendo 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 4: Estimadores para los parámetros con intervalos de confianza del 95 %

El intervalo de confianza de la pendiente es el correspondiente al parámetro x_8, por lo que la pendiente se situa entre -0.00961 y -0.00444.

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 5: 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.

$$MST = \frac{SST}{N-1} = \frac{326,96429}{27} = 12.109788519 \tag{8}$$

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:

$$\frac{SS_{Modelo}}{SS_{Total}} \tag{9}$$

Que es un coeficiente conocido como R^2, que también podemos obtener directamente con la siguiente tabla obtenida del procedimiento *REG*:

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

Figura 6: 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. *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, añadimos la opción *CLI* al modelo del procedimiento *REG*.

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 7: 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.

2.2. 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 % 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, 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 8: 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 II
Código Fuente

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

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

Figura 10: *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 11: *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 12: *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 13: *Código SAS:* Ejercicio 2.2 del Montgomery - Obtención de valores predichos con intervalos de confianza del 90 %.