Universidad del Valle de Guatemala
10 de febrero de 2021
Rudik R. Rompich - Carné: 19857

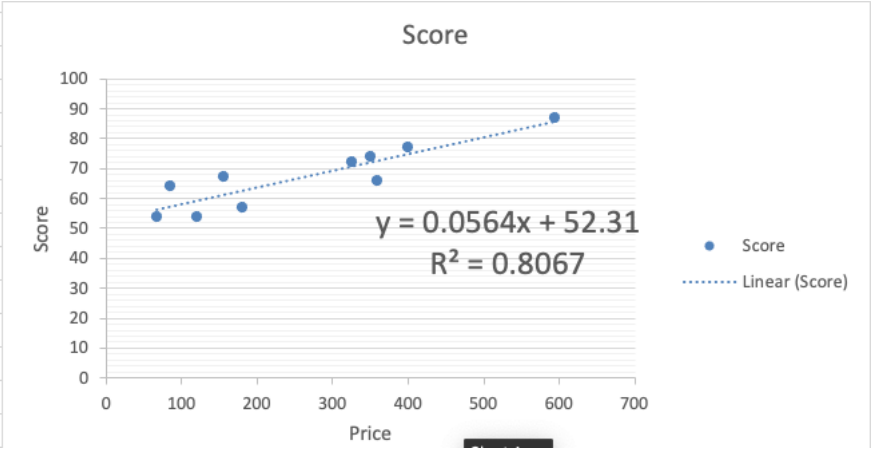
Estadística 2 - Eugenio Aristondo

Tarea 1

1. Sección 14

1.1. Ejercicio 9

- Trace un diagrama de dispersión utilizando el precio como la variable independiente.



- ¿Qué indica el diagrama de dispersión del inciso a) acerca de la relación entre las dos variables?

  Solución. Indica que tiene una relación positiva. ■

- Use el método de mínimos cuadrados para desarrollar la ecuación de regresión estimada.

	x	y	xy	xx
	325	72	23400	105625
	350	74	25900	122500
	67	54	3618	4489
	120	54	6480	14400
	85	64	5440	7225
	180	57	10260	32400
	360	66	23760	129600
	156	67	10452	24336
	595	87	51765	354025
	400	77	30800	160000
Suma	2638	672	191875	954600

$$\begin{aligned} intercepto &= \frac{\sum Y \cdot \sum X^2 - \sum X \cdot \sum XY}{n \cdot \sum X^2 - (\sum X)^2} = \frac{672 \cdot 954600 - 2638 \cdot 191875}{10 \cdot 954600 - 2638^2} \approx 52{,}31 \\ pendiente &= \frac{n \cdot \sum XY - \sum X \cdot \sum Y}{n \cdot \sum X^2 - (\sum X)^2} = \frac{10 \cdot 191875 - 2638 \cdot 672}{10 \cdot 954600 - (2638)^2} \approx 0{,}056 \end{aligned} \tag{1}$$

$$\implies y = mx + b$$
$$\implies y = 0{,}056x + 52{,}31$$

- Proporcione una interpretación para la pendiente de la ecuación de regresión estimada.

  Solución. Indica que por cada dolar hay un incremento de 0.056. ■

- La maleta de la marca Eagle Creek Hovercraft tiene un precio de $225. Usando la ecuación de regresión estimada desarrollada en el inciso c), prediga la puntuación para esta maleta.

  Solución.

$$\implies y = 0{,}056x + 52{,}31$$
$$\implies y = 0{,}056(225) + 52{,}31$$
$$\implies y = 64{,}91$$

■

1.2. Ejercicio 19

- Calculelas SCE, STC y SCR.

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	2272	2272	106.918	6.60903E-06
Residual	8	170	21.25		
Total	9	2442			

Solución.
- SCE = 170
- SCR = 2272

- STC = 2442

■

- Calcule el coeficiente de determinación r^2 . Haga un comentario sobre la bondad del ajuste.

  Solución. Dado:
  $$r^2 = \frac{SCR}{STC} * 100 = \frac{2272}{2442} * 100 = 93{,}04\,\%$$
  Eso quiere decir que el modelo explica el 93.04 de la variabilidad de x y y. ■

- ¿Cuál es el valor del coeficiente de correlación muestral?

  Solución.
  $$r = \sqrt{0{,}9304} = 0{,}9646$$

■

1.3. Ejercicio 27

- Use estos datos para desarrollar la ecuación de regresión estimada a efecto de estimar el precio de las mochilas y las botas para excursionismo con base en el soporte superior.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.893393
R Square	0.798151
Adjusted R	0.77292
Standard E	17.634
Observatio	10

ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regression	1	9836.737	9836.737	31.63366	0.000496
Residual	8	2487.663	310.9579		
Total	9	12324.4			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	49.93069	21.27398	2.347031	0.046898	0.872802	98.98858	0.872802	98.98858
Support	31.20792	5.548686	5.624381	0.000496	18.41263	44.00321	18.41263	44.00321

$$y = 31,20792x + 49,93069$$

- Empleando un nivel de significancia de 0.05, determine si hay relación entre soporte superior y precio.

  Solución. Basándose en el Valor-P, la H_0 se rechaza ya que el Valor-P $\leq 0{,}05$. ■

- ¿Confiaría en usar la ecuación de regresión estimada desarrollada en el inciso a) para estimar el precio de las mochilas y las botas con base en la evaluación del soporte superior?

  Solución. No, ya que el Valor-P de la pendiente no es significante. ■

- Estime el precio de una mochila que tiene 4 como evaluación del soporte superior.

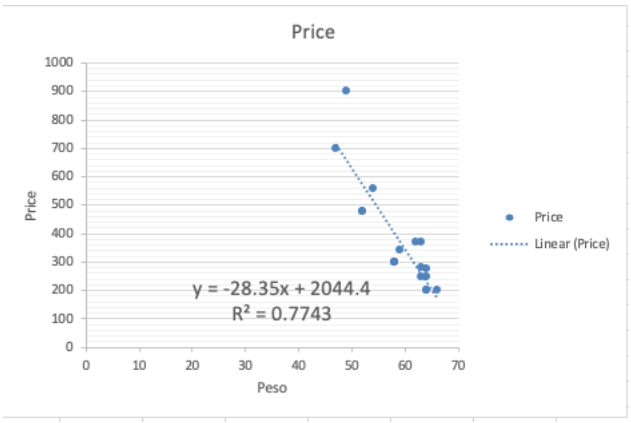
  Solución.

$$y = 31,20792x + 49,93069$$
$$y = 31,20792(4) + 49,93069$$
$$y = 165,4$$

■

1.4. Ejercicio 44

- Trace un diagrama de dispersión usando el peso como variable independiente.



- ¿Parece haber alguna relación entre las dos variables?

  Solución. Sí, una relación negativa. ■

- c) Obtenga la ecuación de regresión estimada que pueda utilizarse para predecir el precio de acuerdo con el peso.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.879962
R Square	0.774333
Adjusted R Square	0.760228
Standard Error	91.8098
Observations	18

ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regression	1	462761.1	462761.1	54.90082	1.48E-06
Residual	16	134864.6	8429.04		
Total	17	597625.8			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	2044.381	226.3543	9.031775	1.11E-07	1564.531	2524.231	1564.531	2524.231
Weight	-28.3499	3.826147	-7.40951	1.48E-06	-36.4609	-20.2388	-36.4609	-20.2388

$$y = -28,3499x + 2044,381$$

- Pruebe la significancia de la relación en un nivel de significancia de 0.05.

  Solución. Basándose en el Valor-P, la H_0 se rechaza ya que el Valor-P $\leq 0{,}05$. ■

- ¿La ecuación de regresión estimada proporciona un buen ajuste? Explique.

  Solución.
  $$r^2 = \frac{SCR}{STC} * 100 = \frac{462761}{134864} * 100 = 77,43\,\%$$
  Por lo que la regresión proporciona un buen ajuste. ■

1.5. Ejercicio 49

- Obtenga una ecuación de regresión estimada que pueda utilizarse para pronosticar los precios de venta dada la extensión en pies cuadrados.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.754682
R Square	0.569546
Adjusted F	0.545631
Standard E	19166
Observatic	20

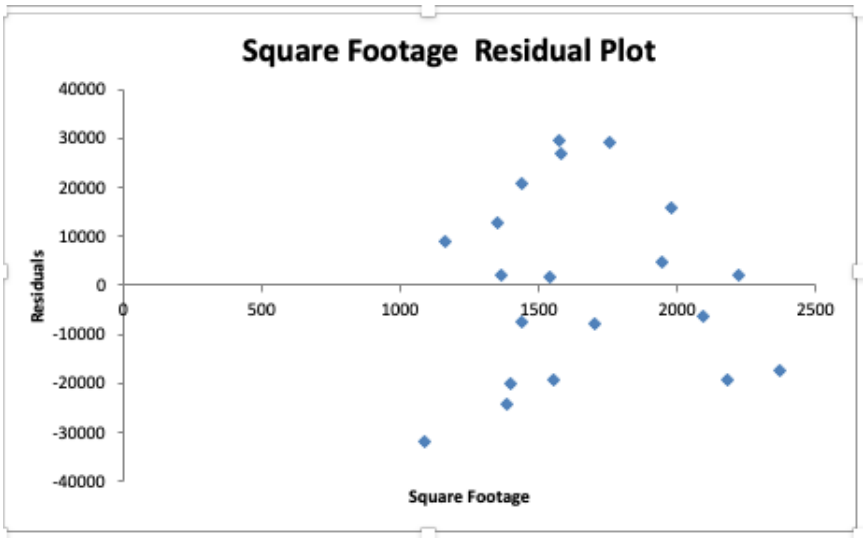
ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regressior	1	8.75E+09	8.75E+09	23.81627	0.00012
Residual	18	6.61E+09	3.67E+08		
Total	19	1.54E+10			

	Coefficients	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Jpper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	22635.95	20460.07	1.106348	0.283144	-20349.1	65620.96	-20349.1	65620.96
Square Fo	58.95954	12.0814	4.880192	0.00012	33.57746	84.34162	33.57746	84.34162

$$y = 58,95954x + 22635,95$$

- Construya una gráfica de residuales estandarizados contra la variable independiente.



- A la luz de la gráfica, ¿los supuestos acerca de los términos del error y de la forma del modelo parecen razonables?

  Solución. No, pareciera que la gráfica no tiene una forma definida. ■

2. Sección 15

2.1. Ejercicio 10

- Desarrolle una ecuación de regresión estimada para predecir la proporción de juegos ganados, dada la proporción de anotaciones de campo del equipo.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.448158
R Square	0.200846
Adjusted R	0.171247
Standard E	0.126636
Observatio	29

ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regression	1	0.10882	0.10882	6.785719	0.014763
Residual	27	0.43299	0.016037		
Total	28	0.54181			

	Coefficients	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	-1.22071	0.661717	-1.84477	0.076069	-2.57845	0.137017	-2.57845	0.137017
FG%	3.957595	1.519265	2.604941	0.014763	0.840321	7.074869	0.840321	7.074869

$$y = 3,957595x - 1,22071$$

- Interprete la pendiente de la ecuación de regresión estimada obtenida con el inciso a).

  Solución. Quiere decir que por cada proporción de anotaciones de campo del equipo, hay un incremento de 3,957595 en la proporción de juegos ganados. ■

- Obtenga una ecuación de regresión estimada para predecir la proporción de juegos ganados dada la proporción de anotaciones de campo del equipo, el porcentaje de tiros de tres puntos del equipo contrario y el número de pérdidas de balón del equipo adversario.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.750846
R Square	0.563769
Adjusted R	0.511422
Standard E	0.097233
Observatio	29

ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regression	3	0.3054558	0.10182	10.76971	9.94E-05
Residual	25	0.236354	0.00945		
Total	28	0.5418099			

	Coefficients	*tandard Erre*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	-1.23459	0.6002511	-2.0568	0.050285	-2.47083	0.001655	-2.47083	0.001655
FG%	4.816565	1.1830431	4.07134	0.000412	2.380043	7.253088	2.380043	7.253088
Opp 3 Pt%	-2.58947	0.7041013	-3.6777	0.001128	-4.03959	-1.13934	-4.03959	-1.13934
Opp TO	0.034425	0.0125325	2.74686	0.010994	0.008614	0.060236	0.008614	0.060236

$$y = 4,816565x_1 - 2,58947x_2 + 0,034425x_3 - 1,23459$$

- Analice las implicaciones prácticas de la ecuación obtenida en el inciso c).

  Solución. Es un modelo con un R^2 bastante malo. No predice perfectamente la relación lineal. ■

- Estime la proporción de juegos ganados por un equipo para el que los valores de las tres variables independientes son: FG % = 0.45; Opp 3 Pt %= 0.34, y Opp TO =17.

$$y = 4,816565x_1 - 2,58947x_2 + 0,034425x_3 - 1,23459$$

$$y = 4,816565(0,45) - 2,58947(0,34) + 0,034425(17) - 1,23459$$

$$y = 0,638$$

2.2. Ejercicio 18

- a) En el inciso c) del ejercicio 10 se obtuvo una ecuación de regresión estimada que arrojó la proporción de juegos ganados dado el porcentaje de anotaciones de campo del equipo, la proporción de tiros de tres puntos del conjunto contrario y la cantidad de recuperaciones de balón del equipo adversario. ¿Cuáles son los valores de R^2 y R_a^2 ?

  Solución.

$$r^2 = \frac{SCR}{STC} = \frac{0,3054558}{0,5418099} = 0,564$$

$$R_a^2 = 1 - \left(1 - R^2\right)\frac{n-1}{n-p-1} = 1 - (1 - 0,564)\frac{29-1}{29-2-1} = 0,564$$

■

- b) ¿Esta ecuación de regresión estimada proporciona un buen ajuste a los datos? Explique.

  Solución. No, ya que R_a^2 es un valor muy alejado a 1. Por lo que no es una buena ecuación. ■

2.3. Ejercicio 24

- a) Desarrolle la ecuación de regresión estimada para predecir el sueldo del entrenador dados los ingresos generados por el programa y el porcentaje de victorias.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.564555
R Square	0.318723
Adjusted F	0.280874
Standard E	0.328622
Observatic	39

ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regressior	2	1.818792	0.909396	8.420952	0.001
Residual	36	3.887715	0.107992		
Total	38	5.706508			

	Coefficients	*ndard Err*	*t Stat*	*P-value*	*Lower 95%*	*pper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	-0.68204	0.504361	-1.35228	0.184719	-1.70493	0.340853	-1.70493	0.340853
Revenue	0.049828	0.013451	3.704238	0.000708	0.022547	0.077108	0.022547	0.077108
%Wins	0.014683	0.006291	2.333842	0.025303	0.001924	0.027442	0.001924	0.027442

$$y = 0,049828x_1 + 0,014683x_2 - 0,68204$$

- Use la prueba F para determinar la significancia global de la relación. ¿Cuál es su conclusión empleando 0.05 como nivel de significancia?

  Solución. La hipótesis nula se rechaza por el Valor-P, aunque es evidencia que la relación es significante. ■

- Utilice la prueba t para determinar la significancia de cada una de las variables indepen- dientes. ¿Cuál es su conclusión con un nivel de significancia de 0.05?

  Solución. Dados los valores-P de las dos variables independientes y menores a 0.05, se concluye que son significantes. ■

2.4. Ejercicio 51

- Calcule las entradas que faltan en esta pantalla.

```
The regression equation is
Y = 8.103 + 7.602 X1 + 3.111 X2

Predictor            Coef        SE Coef           T
Constant            8,103          2.667      3,0382
X1                  7,602          2.105      3,6114
X2                  3,111          0.613      5,0750

S = 3.335      R-sq = 92.3%     R-sq(adj) = 91,02 %

Analysis of Variance

SOURCE              DF          SS          MS          F
Regression           2        1612         806    71,9221
Residual Error      12    134,4789    11,2066
Total               14   1746,4789
```

- Use la prueba F y $\alpha = 0{,}05$ para identificar si existe una relación significativa.

  Solución. Se rechaza la H_0 y se determina que es significante. ■

- Utilice la prueba t y $\alpha = 0{,}05$ para demostrar $H_0 : \beta_1 1 = 0$ y $H_0 : \beta_2 = 0$

  Solución. Las variables X_1 y X_2 son significativas debibo a su valor-P mayor a 0.05. ■

- Calcule R_a^2.

$$R_a^2 = 1 - \left(1 - R^2\right) \frac{n-1}{n-p-1} = 1 - (1 - 0{,}923) \frac{15-1}{15-2-1} \approx 0{,}9102 = 91{,}02\,\%$$

2.5. Ejercicio 52

,

- Calcule las entradas que faltan en esta pantalla.

```
The regression equation is
Y = -1.41 + .0235 X1 + .00486 X2

Predictor            Coef        SE Coef           T
Constant          -1.4053         0.4848    -2,9084
X1               0.023467       0.008666      2,7117
X2                 0,0486       0.001077      4,5125

S = 0.1298      R-sq = 93,73%    R-sq(adj) = 91.74%

Analysis of Variance

SOURCE              DF          SS          MS          F
Regression           2     1.76209      0,8810    52,3053
Residual Error       7     0,11791      0,0168
Total                9     1.88000
```

- Use la prueba F y 0.05 como nivel de significancia para saber si existe una relación significativa.

  Solución. Basándose en la prueba F, hay suficiente evidencia para afirmar que la relación es significante. ■

- Utilice la prueba t y $\alpha = 0{,}05$ para probar $H_0 : \beta_1 = 0$ y $H_a : \beta_2 = 0$.

  Solución. Hay evidencia para afirmar que X_2 es significante, mientras que X_1 no lo es. ■

- ¿La ecuación de regresión estimada proporciona un buen ajuste a los datos? Explique.

  Solución. Basándose en R^2 y R_a^2 se puede afirmar que la ecuación de la regresión estimada tiene un buen ajuste. ■

3. Sección 16

3.1. Ejercicio 11

En un análisis de regresión con 30 observaciones se obtuvo la siguiente ecuación de regresión estimada.

$$\hat{y} = 17{,}6 + 3{,}8x_1 - 2{,}3x_2 + 7{,}6x_3 + 2{,}7x_4$$

Para esta ecuación de regresión estimada, STC = 1805ySCR = 1760.

- Con $\alpha = 0{,}05$, pruebe la significancia de la relación entre las variables. Suponga que las variables x_1 y x_4 se retiran del modelo y se obtiene la siguiente ecuación de regresión estimada.

  $$\hat{y} = 11{,}1 - 3{,}6x_2 + 8{,}1x_3$$

  Para este modelo, STC = 1805 y SCR = 1705.

  Solución.

  $$\begin{aligned} MSE &= \frac{SSE}{n-p-1} = \frac{SST - SSR}{n-p-1} = \frac{1805 - 1760}{30-4-1} = 1{,}8 \\ MSR &= \frac{SSR}{p} = \frac{1760}{4} = 440 \end{aligned} \tag{2}$$

  $$F = \frac{MSR}{MSE} = \frac{440}{1{,}8} = 244{,}4444$$

  Se rechaza la H_0 y se concluye que la relación es significativa. ■

- Calcule SCE (x_1, x_2, x_3, x_4)

  Solución.

  $$SSE = SST - SSR = 45$$

  ■

- Calcule SCE (x_2, x_3)

  Solución.

$$SSE = SST - SSR = 100$$

■

- Utilice la prueba F y 0.05 como nivel de significancia para determinar si x_1 y x_2 contribuyen significativamente al modelo.

  Solución.

$$\begin{aligned} F &= \frac{\frac{\text{SSE}(x_2,x_3)-\text{SSE}(x_1,x_2,x_3,x_4)}{nmerodetrminosextra}}{\frac{\text{SSE}(x_1,x_2,x_3,x_4)}{n-p-1}} \\ &= \frac{\frac{100-45}{2}}{\frac{45}{30-4-1}} \\ &\approx 15,2778 \end{aligned} \tag{3}$$

  Se concluye que las variables son significativas. ■

3.2. Ejercicio 12

- Desarrolle una ecuación de regresión estimada para pronosticar la Scoring Avg. de todos los eventos dado el número promedio de putts en los golpes dados en Green in Reg.

SUMMARY OUTPUT

Regression Statistics	
Multiple F	0.62189
R Square	0.38675
Adjusted	0.36485
Standard	0.5106
Observati	30

ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regressio	1	4.60363	4.60363	17.6582	0.00024
Residual	28	7.29984	0.26071		
Total	29	11.9035			

	Coefficient	*ndard Er*	*t Stat*	*P-value*	*ower 95%*	*pper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	46.2774	6.02602	7.6796	2.3E-08	33.9337	58.6211	33.9337	58.6211
Putting A	14.1028	3.35609	4.20216	0.00024	7.22819	20.9775	7.22819	20.9775

$$y = 14,1028x + 46,2774$$

- Desarrolle una ecuación de regresión estimada para pronosticar la Scoring Avg. de todos los eventos dado el tiempo promedio en que una jugadora es capaz de golpear el Green in Reg, y el promedio de veces en que consigue "subir y bajar" una vez que se encuentra en la trampa de arena.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.79796
R Square	0.63675
Adjusted	0.59483
Standard	0.40781
Observati	30

ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regressio	3	7.57948	2.52649	15.1917	6.5E-06
Residual	26	4.32399	0.16631		
Total	29	11.9035			

	Coefficient	*ndard Er*	*t Stat*	*P-value*	*ower 95%*	*pper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	59.0219	5.77423	10.2216	1.3E-10	47.1528	70.891	47.1528	70.891
Greens in	-10.281	2.87661	-3.5741	0.0014	-16.194	-4.3683	-16.194	-4.3683
Putting A	11.4132	2.75984	4.13547	0.00033	5.74029	17.0861	5.74029	17.0861
Sand Save	-1.813	0.92103	-1.9685	0.05976	-3.7062	0.08018	-3.7062	0.08018

$$y = -10,281x_1 + 11,4132x_2 - 1,813x_3 + 59,0219$$

- Con un el nivel de significancia de 0.05, pruebe si las dos variables independientes agregadas en el inciso el porcentaje de veces en que una jugadora consigue llegar al green en regulación y el promedio de veces en que es capaz de "subir y bajar" una vez que se encuentra en la trampa de arena al lado del green, contribuyen significativamente el desarrollo de la ecuación de regresión en el inciso a). Explique.

  Solución.

$$\begin{aligned} F &= \frac{\frac{\text{SSE}(x_2) - \text{SSE}(x_1,x_2,x_3)}{\text{número de términos extra}}}{\frac{\text{SSE}(x_1,x_2,x_3)}{n-p-1}} \\ &= \frac{\frac{7,2998-4,3240}{2}}{\frac{4,3240}{30-3-1}} \\ &= 8,9468 \end{aligned} \tag{4}$$

  Por lo que se concluye, que es una regresión significativa. ■

3.3. Ejercicio 14

- Desarrolle una ecuación de regresión estimada para predecir el riesgo de fumar dada la edad y el nivel de presión sanguínea.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.89801
R Square	0.80641
Adjusted I	0.78364
Standard I	6.90826
Observatic	20

ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regressio	2	3379.64	1689.82	35.4081	8.7E-07
Residual	17	811.31	47.7241		
Total	19	4190.95			

	Coefficients	*ndard Err*	*t Stat*	*P-value*	*Lower 95%*	*Jpper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	-110.942	16.4699	-6.73607	3.5E-06	-145.691	-76.1939	-145.691	-76.1939
Age	1.315	0.17329	7.58847	7.4E-07	0.9494	1.68061	0.9494	1.68061
Blood Pressure	0.2964	0.05107	5.80401	2.1E-05	0.18866	0.40415	0.18866	0.40415

$$y = 1,315x_1 + 0,2964x_2 - 110,942$$

- Considere la adición de dos variables independientes al modelo desarrollado en el inciso a): una para la interacción entre la edad y el nivel de presión arterial y otra que indique si la persona es fumadora. Desarrolle una ecuación de regresión estimada utilizando estas cuatro variables independientes.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.93606
R Square	0.8762
Adjusted I	0.84319
Standard I	5.88127
Observatic	20

ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regressio	4	3672.11	918.027	26.5407	1.2E-06
Residual	15	518.841	34.5894		
Total	19	4190.95			

	Coefficients	*ndard Err*	*t Stat*	*P-value*	*Lower 95%*	*Jpper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	-123.165	56.9424	-2.16298	0.0471	-244.535	-1.79518	-244.535	-1.79518
Age	1.51298	0.77956	1.94083	0.07131	-0.1486	3.17457	-0.1486	3.17457
Blood Pressure	0.44825	0.34573	1.29654	0.21438	-0.28866	1.18516	-0.28866	1.18516
Smoker	8.86555	3.07365	2.88438	0.01135	2.31423	15.4169	2.31423	15.4169
A-B	-0.00276	0.00481	-0.57333	0.57492	-0.013	0.00749	-0.013	0.00749

$$y = -0,00276x_4 + 8,86555x_3 + 0,44825x_2 + 1,51298x_1 - 123,165$$

- Con un nivel de 0.05 de significancia, lleve a cabo una prueba para determinar si la adición del término interacción y la variable fumador contribuyen significativamente a la ecuación de regresión estimada desarrollada en el inciso a).

Solución.

$$\begin{aligned} F &= \frac{\frac{\text{SSE}(x_1,x_2)-\text{SSE}(x_1,x_2,x_3,x_4)}{\text{nmbero de términos extra}}}{\frac{\text{SSE}(x_1,x_2,x_3,x_4)}{n-p-1}} \\ &= \frac{\frac{811{,}3097-518{,}8406}{2}}{\frac{518{,}8406}{20-4-1}} \\ &= 4,2277 \end{aligned} \tag{5}$$

Por lo tanto, es significativo. ■

3.4. Ejercicio 17

The Ladies Professional Golfers Association (LPGA) lleva estadísticas sobre el desempeño y las ganancias de los miembros del LPGA Tour. Las estadísticas de fin de año sobre el papel de las 30 jugadoras que obtuvieron las mejores ganancias totales en la LPGA Tour de 2005 apArecen en el archivo titulado LPGATour2 (sitio web de LPGATour, 2006). Earnings (ganancias) constituyen el resultado total en miles de dólares en todos los eventos de la gira; Scoring Avg. es la puntuación promedio para todos los eventos; Drive Average es la distancia promedio en yardas alcanzada en el drive por la jugadora; Greens in Reg. es el porcentaje de veces que la golfista llega al green en regulación; Putting Avg. es el promedio de putts en el green en regulación, y Sand Saves es el porcentaje de veces que una jugadora es capaz de logra "subir y bajar" (up and down) cuando se encuentra en la trampa de arena al lado del green. Éste se considera un golpe en la regulación si alguna parte de la bola toca la superficie del putting y la diferencia entre el valor del par de hoyos y el número de golpes que lleva a golpear el green es por lo menos de 2. DriveGreens denota una nueva variable independiente que representa la interacción entre la distancia media alcanzada en el drive por la jugadora y el porcentaje de veces que es capaz de alcanzar el green en regulación. Utilice los métodos de esta sección a efecto de desarrollar la mejor ecuación de regresión múltiple estimada para calcular el Scoring Avg. de una jugadora en todos los eventos.

Solución. Matriz de correlación:

	Scoring Avg.	*nings ($10*	*Drive Average*	*Greens in Reg.*	*Putting Avg.*	*Sand Saves*	*DriveGreens*
Scoring Avg.	1						
Earnings ($1000)	-0.75393899	1					
Drive Average	-0.303704195	0.267689	1				
Greens in Reg.	-0.574877327	0.565748	0.356147201	1			
Putting Avg.	0.621889954	-0.50675	-0.259380878	-0.232217067	1		
Sand Saves	-0.312977045	0.499926	-0.020079244	0.095130449	-0.0744091	1	
DriveGreens	-0.560227807	0.532741	0.784664141	0.858181299	-0.3017199	0.06280341	1

Por otra parte, la regresión lineal:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.866000555
R Square	0.74995696
Adjusted R Sc	0.684728341
Standard Erro	0.359732932
Observations	30

ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regression	6	8.92708768	1.487848	11.49736	5.98E-06
Residual	23	2.976378987	0.129408		
Total	29	11.90346667			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	-49.92520686	48.04436538	-1.03915	0.309537	-149.313	49.46214	-149.313	49.46214
Earnings ($10	-0.000388424	0.000278775	-1.39332	0.176842	-0.00097	0.000188	-0.00097	0.000188
Drive Average	0.446670187	0.190347095	2.346609	0.027928	0.052907	0.840433	0.052907	0.840433
Greens in Reg	158.6231405	70.02189001	2.265336	0.03322	13.77182	303.4745	13.77182	303.4745
Putting Avg.	7.858943246	2.890027288	2.719332	0.01223	1.880466	13.83742	1.880466	13.83742
Sand Saves	-0.090426931	1.02409938	-0.0883	0.930403	-2.20894	2.028084	-2.20894	2.028084
DriveGreens	-0.656568379	0.278722405	-2.35564	0.02739	-1.23315	-0.07999	-1.23315	-0.07999

Se determinó que el valor-P de la variable Sand Saves es demasiado elevado, por lo que se decidió eliminarlo.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.865951615
R Square	0.749872199
Adjusted R Square	0.69776224
Standard Error	0.352218442
Observations	30

ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regression	5	8.926079	1.785216	14.39019	1.47E-06
Residual	24	2.977388	0.124058		
Total	29	11.90347			

	Coefficients	*ndard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	-50.6641456	46.32165	-1.09375	0.284923	-146.267	44.93904	-146.267	44.93904
Earnings ($1000)	-0.000400696	0.000237	-1.69346	0.103309	-0.00089	8.76E-05	-0.00089	8.76E-05
Drive Average	0.449773961	0.183166	2.45556	0.02169	0.071739	0.827809	0.071739	0.827809
Greens in Reg.	159.8149892	67.27335	2.375606	0.025851	20.96963	298.6604	20.96963	298.6604
Putting Avg.	7.793844065	2.736034	2.848592	0.008867	2.146948	13.44074	2.146948	13.44074
DriveGreens	-0.661026378	0.268386	-2.46297	0.021338	-1.21495	-0.10711	-1.21495	-0.10711

Por otra parte, también se determinó que la variable Earnings, tiene un Valor-P superior a 0.05, por lo que se decidió eliminarlo.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.848518577
R Square	0.719983775
Adjusted R Square	0.675181179
Standard Error	0.365139086
Observations	30

ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regression	4	8.570303	2.142576	16.07014	1.23E-06
Residual	25	3.333164	0.133327		
Total	29	11.90347			

	Coefficients	*ndard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	-88.09849682	42.20023	-2.08763	0.04718	-175.011	-1.18549	-175.011	-1.18549
Drive Average	0.590714727	0.169151	3.492231	0.0018	0.242342	0.939088	0.242342	0.939088
Greens in Reg.	209.1869123	62.8518	3.328256	0.002709	79.7412	338.6326	79.7412	338.6326
Putting Avg.	9.735789766	2.575255	3.780515	0.000869	4.431953	15.03963	4.431953	15.03963
DriveGreens	-0.867656398	0.247815	-3.50123	0.00176	-1.37804	-0.35727	-1.37804	-0.35727

Finalmente, se llegó a un modelo de 4 variables, considerando:

- Drive Average $= x_1$

- Greens in Reg $= x_2$
- Putting Avf $= x_3$
- DriveGreen $= x_4$

$$\implies y = 0,5907x_1 + 209,1869_2 + 9,7378x_3 - 0,8676x_4 - 88,0984$$

■