Mathew Witek
05/02/2019
Statistics II
Final Project

Shoe Size and Height

Introduction

We are first presented with a data set containing both the shoe sizes and heights of both females and males. Our goal in this project is to split the data based on sexes, and then analyze the two data sets to see if shoe size is a useful predictor for height, along with other calculations and predictions to show that we can properly analyze and interpret the data.

Procedure

The tools used to complete this project were SAS statistical software and excel in order to compute and calculate the required values needed to adequately analyze and interpret the given data set(s). Although all code, output, and results will be attached to this writeup, screen shots of the tools and technologies used to find the answers will be provided throughout this writeup in order for the reader to better grasp where the answers came from.

Analysis

A.) Here we are asked to simply separate the data into two data sets, splitting the original data set based on sexes; male and female. This was done with a two simple 'if' statements in the SAS statistical software, resulting in…

Mens Data

Obs	Size	Height	Sex
1	10.5	70.0	M
2	13.0	72.0	M
3	10.5	74.5	M
4	12.0	71.0	M
5	10.5	71.0	M
6	13.0	77.0	M
7	11.5	72.0	M
8	10.0	72.0	M
9	8.5	67.0	M
10	10.5	73.0	M
11	10.5	72.0	M
12	11.0	70.0	M
13	9.0	69.0	M
14	13.0	70.0	M

Females Data

Obs	Size	Height	Sex
1	6.5	66.0	F
2	9.0	68.0	F
3	8.5	64.5	F
4	8.5	65.0	F
5	7.0	64.0	F
6	9.5	70.0	F
7	9.0	71.0	F
8	7.5	64.0	F
9	8.5	67.0	F
10	8.5	59.0	F
11	5.0	62.0	F
12	6.5	66.0	F
13	7.5	64.0	F
14	8.5	69.0	F

The following B. through J. are all analysis of the male data set.

B.) Moving on, we are asked to determine the sample regression equation with shoe size as the predictor variable for height. In SAS, there is a simple way to output a diagram and model a regression line based on a data set by using the proc reg data method and then

using model Height = Size because we know that size is being used as the predictor variable for height. This is what I did in SAS below…

```
/* B. Determine the sample regression equation with shoe size as the predictor variable for height.*/
Title 'Mens Regression Line';
ods graphics on;
   proc reg data=shoesizeMen plots=residualbypredicted;
      model Height = Size / r clm cli;
   run;
```

This then output a lot of results of which one must carefully read through to interpret and find the necessary values to construct the sample regression equation. The output is shown below and highlighted to show the y-intercept (in green) and the slope (in yellow).

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	61.67176	4.69243	13.14	<.0001
Size	1	0.89313	0.42474	2.10	0.0573

With the help of SAS statistical software, we get that the sample regression equation for the male data set is, $\hat{y} = 61.67176 + 0.89313x$.

C.) In order to find the standard error (S_e), we must first find the error sum of squares (*SSE*). The SSE is given in the output from the provided SAS code above.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	20.52604	20.52604	4.42	0.0573
Error	12	55.70611	4.64218		
Corrected Total	13	76.23214			

Now that we have SSE = 55.70611, we can use this value in the equation $S_e = \sqrt{\frac{SSE}{n-2}}$. We know that n = 14 and plugging that all in gives us $S_e = 2.155$. This means that the predicted height of a male in the sample differs, on average, from the observed height by 2.155 inches.

D.) To determine whether shoe size is useful for predicting height, we must complete a hypothesis test, where Ho: $\beta_1 = 0$, and Ha: $\beta_1 \neq 0$. The previously shown SAS codes output already does the work for us and provides us with the needed test statistic and P-value to interpret.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	61.67176	4.69243	13.14	<.0001
Size	1	0.89313	0.42474	2.10	0.0573

Here we see that the test statistic has a value of 2.10, and the P-value is 0.0573 when $\alpha = 0.05$, which means that the P-value is greater than α. That means that we cannot reject the null hypothesis of Ho: $\beta_1 = 0$. This means that the data does not provide sufficient evidence to conclude that shoe size is useful for predicting height.

E.) A point estimate for the mean height of all males who wear a size 10.5 shoe is given by simply plugging in 10.5 into the sample regression equation. This gives us

$$\hat{y} = 61.67176 + 0.89313(10.5)$$
$$\hat{y} = 71.05 \text{ inches.}$$

F.) Here we are asked to obtain a 95% confidence interval for the mean height of all males who wear a size 10.5 shoe. The SAS code shown in part B once again provides the output that we need in order to answer this question. The useful output that the code provides us with is…

The REG Procedure
Model: MODEL1
Dependent Variable: Height

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual	Std Error Residual	Student Residual	Cook's D
1	70.0	71.0496	0.6087	69.7235	72.3758	66.1715	75.9277	-1.0496	2.067	-0.508	0.011
2	72.0	73.2824	1.0388	71.0190	75.5459	68.0709	78.4940	-1.2824	1.888	-0.679	0.070
3	74.5	71.0496	0.6087	69.7235	72.3758	66.1715	75.9277	3.4504	2.067	1.669	0.121
4	71.0	72.3893	0.7246	70.8105	73.9682	67.4365	77.3421	-1.3893	2.029	-0.685	0.030
5	71.0	71.0496	0.6087	69.7235	72.3758	66.1715	75.9277	-0.0496	2.067	-0.024	0.000
6	77.0	73.2824	1.0388	71.0190	75.5459	68.0709	78.4940	3.7176	1.888	1.969	0.587
7	72.0	71.9427	0.6192	70.5937	73.2918	67.0584	76.8271	0.0573	2.064	0.028	0.000
8	72.0	70.6031	0.7066	69.0634	72.1427	65.6626	75.5435	1.3969	2.035	0.686	0.028
9	67.0	69.2634	1.1946	66.6605	71.8662	63.8956	74.6311	-2.2634	1.793	-1.262	0.354
10	73.0	71.0496	0.6087	69.7235	72.3758	66.1715	75.9277	1.9504	2.067	0.944	0.039
11	72.0	71.0496	0.6087	69.7235	72.3758	66.1715	75.9277	0.9504	2.067	0.460	0.009
12	70.0	71.4962	0.5760	70.2411	72.7513	66.6369	76.3555	-1.4962	2.076	-0.721	0.020
13	69.0	69.7099	1.0137	67.5012	71.9187	64.5219	74.8980	-0.7099	1.901	-0.373	0.020
14	70.0	73.2824	1.0388	71.0190	75.5459	68.0709	78.4940	-3.2824	1.888	-1.739	0.458

Looking at the table above, we can see that the observations which correspond to a shoe size of 10.5 are observations 1, 3, 5, 10, and 11. All of which have a 95% confidence interval of (69.72, 72.38). We can be 95% confident that the height of men who wear a size 10.5 shoe lies between 69.72 and 72.38 inches.

G.) The predicted height of a male who wears a size 10.5 shoe can be found by plugging in 10.5 into the sample regression equation.

$$\hat{y} = 61.67176 + 0.89313(10.5)$$
$$\hat{y} = 71.05 \text{ inches.}$$

H.) Looking at the same output used to find the 95% confidence interval in part f. we can see that next to the 95% confidence interval column, there is 95% confidence interval predict

column where we will find our answer. The 95% prediction interval for the height of a male who wears a size 10.5 shoe is (66.17, 75.93). We can be 95% confident that the height of males who wear a size 10.5 shoe lies between 66.17 and 75.93 inches.

I.) To find if the data provides sufficient evidence to conclude that shoe size and height are positively linearly correlated, we used the proc corr method in SAS to calculate the correlation value between Size and Height, and the P-value. The code is provided below along with the output it generated…

```
/*I.Correlation*/
Title 'Correlation between Height and Size';
PROC CORR DATA= shoesizeMen fisher;
    VAR Size Height;
RUN;
```

Pearson Correlation Coefficients, N = 14
Prob > |r| under H0: Rho=0

	Size	Height
Size	1.00000	0.51890 0.0573
Height	0.51890 0.0573	1.00000

As we can see in the boxed in region, we are provided with the correlation value of 0.519, and a P-value of 0.0573. The P-value of the right-tailed test is 0.0285. This gives us…

$$\text{P-value } (0.0285) < \alpha\ (0.05)$$

This means that we reject the Ho: p = 0, and we can conclude that at the 5% significance level, the data does provide sufficient evidence to conclude that shoe size and height are positively linearly correlated.

Upon completing I, J. states to repeat B-I for the female data set of shoe size and height.

B.) We are asked to determine the sample regression equation with shoe size as the predictor variable for height. The following code is used which provides a lot of useful output.

```
/*Repeating b-j for Female Data Set.*/

Title 'Womens Regression Line';
ods graphics on;
   proc reg data=shoesizeFemale plots=residualbypredicted;
      model Height = Size / r clm cli;
   run;
```

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	55.72527	5.13817	10.85	<.0001
Size	1	1.26678	0.64643	1.96	0.0737

The boxed in data gives us our sample regression equation for shoe size and height which is $\hat{y} = 55.725 + 1.267x$.

C.) In order to find the standard error (S_e), we must first find the error sum of squares

(*SSE*). The SSE is given in the output from the provided SAS code in part B.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	32.43873	32.43873	3.84	0.0737
Error	12	101.36484	8.44707		
Corrected Total	13	133.80357			

Now that we have SSE = 101.365, we can use this value in the equation $S_e = \sqrt{\frac{SSE}{n-2}}$. We know that n = 14 and plugging that all in gives us $S_e = 2.906.$ This means that the predicted height of a female in the sample differs, on average, from the observed height by 2.906 inches.

D.) In order to determine whether shoe size is useful for predicting height, we must complete a hypothesis test, where Ho: $\beta_1 = 0$, and Ha: $\beta_1 \neq 0$. The previously shown SAS codes output already does the work for us and provides us with the needed test statistic and P-value to interpret.

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	55.72527	5.13817	10.85	<.0001
Size	1	1.26678	0.64643	1.96	0.0737

Here we see that the test statistic has a value of 1.96, and the P-value is 0.0737 when $\alpha = 0.05$, which means that the P-value is greater than α. That means that we cannot reject the null hypothesis of Ho: $\beta_1 = 0$. This means that the data does not provide sufficient evidence to conclude that shoe size is useful for predicting height.

E.) A point estimate for the mean height of all females who wear a size 8 shoe is given by simply plugging in 8 into the sample regression equation. This gives us

$$\hat{y} = 55.725 + 1.267(8)$$
$$\hat{y} = 65.861 \text{ inches.}$$

F.) Here we are asked to obtain a 95% confidence interval for the mean height of all females who wear a size 8 shoe. The SAS code shown in part B once again provides the output that we need in order to answer this question. The useful output that the code provides us with is… (next page due to sizing)

Womens Regression Line

The REG Procedure
Model: MODEL1
Dependent Variable: Height

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual	Std Error Residual	Student Residual	Cook's D
1	66.0	63.9594	1.1718	61.4063	66.5124	57.1316	70.7871	2.0406	2.660	0.767	0.057
2	68.0	67.1263	1.0720	64.7907	69.4620	60.3768	73.8758	0.8737	2.701	0.323	0.008
3	64.5	66.4929	0.8809	64.5735	68.4123	59.8760	73.1099	-1.9929	2.770	-0.720	0.026
4	65.0	66.4929	0.8809	64.5735	68.4123	59.8760	73.1099	-1.4929	2.770	-0.539	0.015
5	64.0	64.5928	0.9541	62.5139	66.6716	57.9278	71.2577	-0.5928	2.745	-0.216	0.003
6	70.0	67.7597	1.3158	64.8929	70.6265	60.8086	74.7109	2.2403	2.591	0.864	0.096
7	71.0	67.1263	1.0720	64.7907	69.4620	60.3768	73.8758	3.8737	2.701	1.434	0.162
8	64.0	65.2261	0.8103	63.4606	66.9917	58.6521	71.8001	-1.2261	2.791	-0.439	0.008
9	67.0	66.4929	0.8809	64.5735	68.4123	59.8760	73.1099	0.5071	2.770	0.183	0.002
10	59.0	66.4929	0.8809	64.5735	68.4123	59.8760	73.1099	-7.4929	2.770	-2.705	0.370
11	62.0	62.0592	2.0036	57.6936	66.4248	54.3677	69.7506	-0.0592	2.105	-0.028	0.000
12	66.0	63.9594	1.1718	61.4063	66.5124	57.1316	70.7871	2.0406	2.660	0.767	0.057
13	64.0	65.2261	0.8103	63.4606	66.9917	58.6521	71.8001	-1.2261	2.791	-0.439	0.008
14	69.0	66.4929	0.8809	64.5735	68.4123	59.8760	73.1099	2.5071	2.770	0.905	0.041

where we can see that the 95% confidence interval for the mean height of all females who wear a size 8 shoe is (64.16, 67.56). This means that we can be 95% confident that the height of all females who wear size 8 shoes lies between 64.16 and 67.56 inches.

G.) The predicted height of a female who wears a size 8 shoe can be found by plugging in 8 into the sample regression equation.

$$\hat{y} = 55.725 + 1.267(8)$$
$$\hat{y} = 65.861 \text{ inches.}$$

H.) Looking at the same output used to find the 95% confidence interval in part f. we can see that 95% prediction interval for the height of a female who wears a size 8 shoe is (59.3, 72.42). We can be 95% confident that the height of females who wear a size 8 shoe lies between 59.3 and 72.42 inches.

I.) To find if the data provides sufficient evidence to conclude that shoe size and height are positively linearly correlated, we used the proc corr method in SAS to calculate the correlation value between Size and Height, and the P-value. The code is provided below along with the output it generated…

```
/*I.Correlation*/
Title 'Correlation between Height and Size';
PROC CORR DATA= shoesizeFemale fisher;
    VAR Size Height;
RUN;
```

Pearson Correlation Coefficients, N = 14
Prob > |r| under H0: Rho=0

	Size	Height
Size	1.00000	0.49238 0.0737
Height	0.49238 0.0737	1.00000

As we can see in the boxed in region, we are provided with the correlation value of 0.49238, and a P-value of 0.0737. The P-value of the right-tailed test is 0.03685. This gives us…

$$\text{P-value } (0.03685) < \alpha\ (0.05)$$

This means that we reject the Ho: $p = 0$, and we can conclude that at the 5% significance level, the data does provide sufficient evidence to conclude that shoe size and height are positively linearly correlated.

Conclusion

We have found that in both data sets, shoe size and height are positively linearly correlated, and that shoe size is not a good predictor variable for height. Both sample regression equations were simply computed, as were the confidence intervals. All code, data, and spreadsheets are attached to this write up. This project has helped me gain a better understanding of chapters 14 and 15, as well as gave me the chance to brush up on my SAS programming.