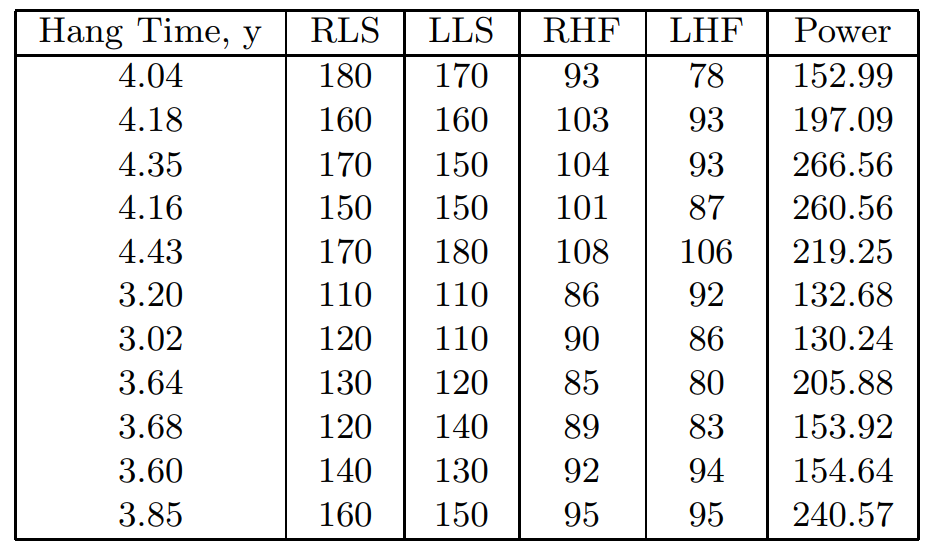
**Assignment #4**

**Stats 157 Winter 2018**

Sarah Ruckman

SID: 7194

Leg strength is a necessary ingredient of a successful punter in football1 . One measure of the quality of a good punt is the “hang time”. This is the time that the ball hangs in the air before being caught by the punt returner. To determine what leg strength factors influence hang time and to develop a model for predicting hang time, a study was conducted using 11 punters. Each of the punters punted a football ten times and the average hang time, along with various strength measures, were recorded:



where RLS = right leg strength (lbs), LLS = left leg strength, RHF = right hamstring muscle flexibility (degrees), LHF = left hamstring muscle flexibility (degrees), and Power = overall leg strength (foot lbs).

1. Read in and print out your data. (3 pts)

**SAS Code:**

/\*Set up options and turn off extra graphics\*/

options nocenter nodate nonumber ps=**55** ls=**78**;

ods graphics off;

/\*goptions formats the plot

cback color of the plot background

colors colors to use

ftitle font of plot title

htitle height of the title

htext height of the text on the plot \*/

goptions reset = all colors=(blue,red,green,purple) ftitle = swissb ftext=swissb htitle=**3**;

/\*Set up titles\*/

title1 'Statistics 157 Winter 2018';

title2 'Assignment #4';

title3 'Sarah Ruckman';

title4 'Question 1';

/\*Create new SAS temporary dataset\*/

**data** foot;

/\*Use an infile statement to read in the data \*/

infile 'C:\Users\sarah\Downloads\footballw18.dat' firstobs = **2**;

/\*Input the variables\*/

input HangTime RLS LLS RHF LHF Power @@;

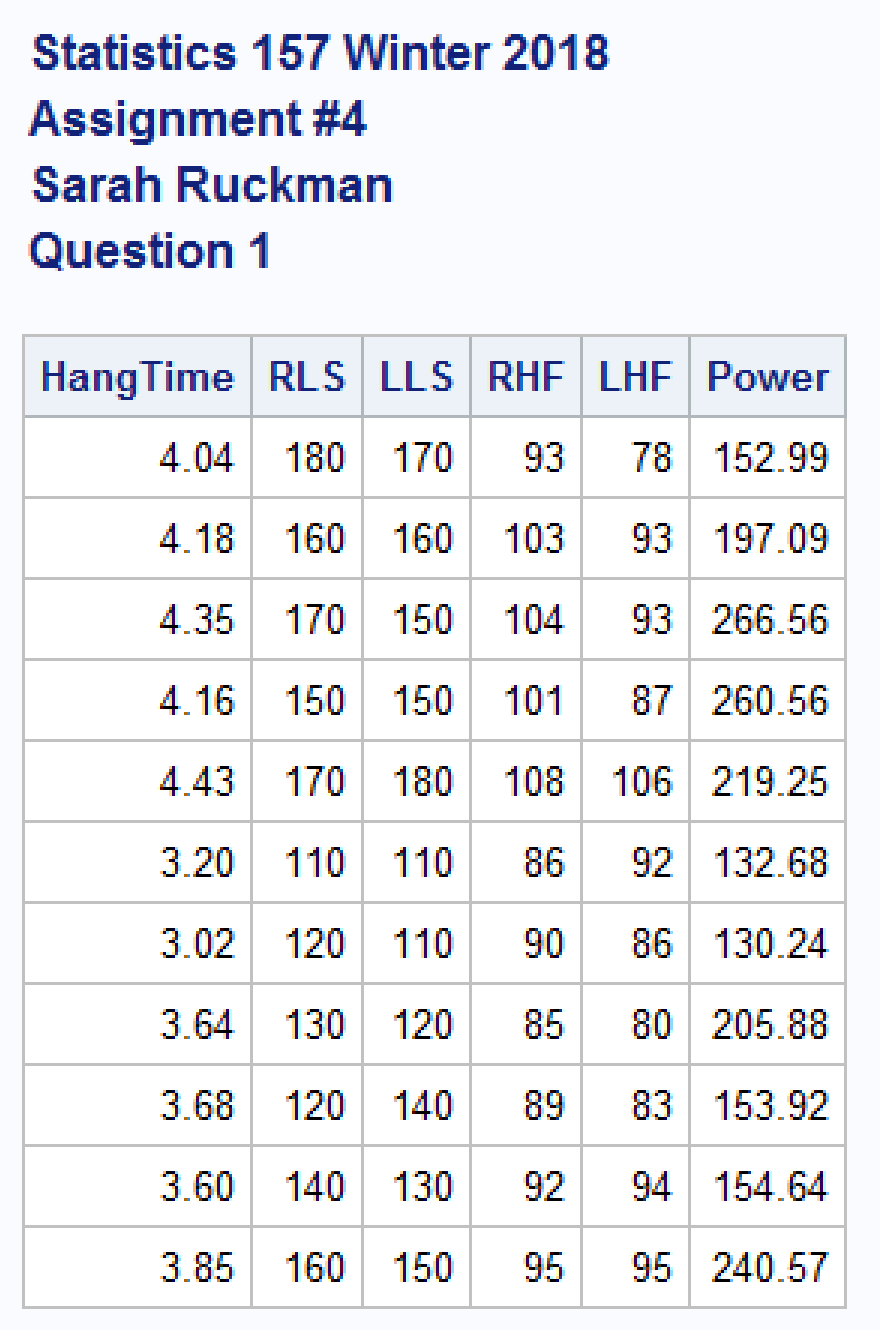
/\*Print as a check\*/

**proc** **print** noobs;

**run**;

**quit**;

**Output:**



1. Find the correlation between all pairs of variables. (3 pts)

**SAS Code:**

/\*Use proc corr to find the correlation between the variables\*/

**proc** **corr** nosimple noprob;

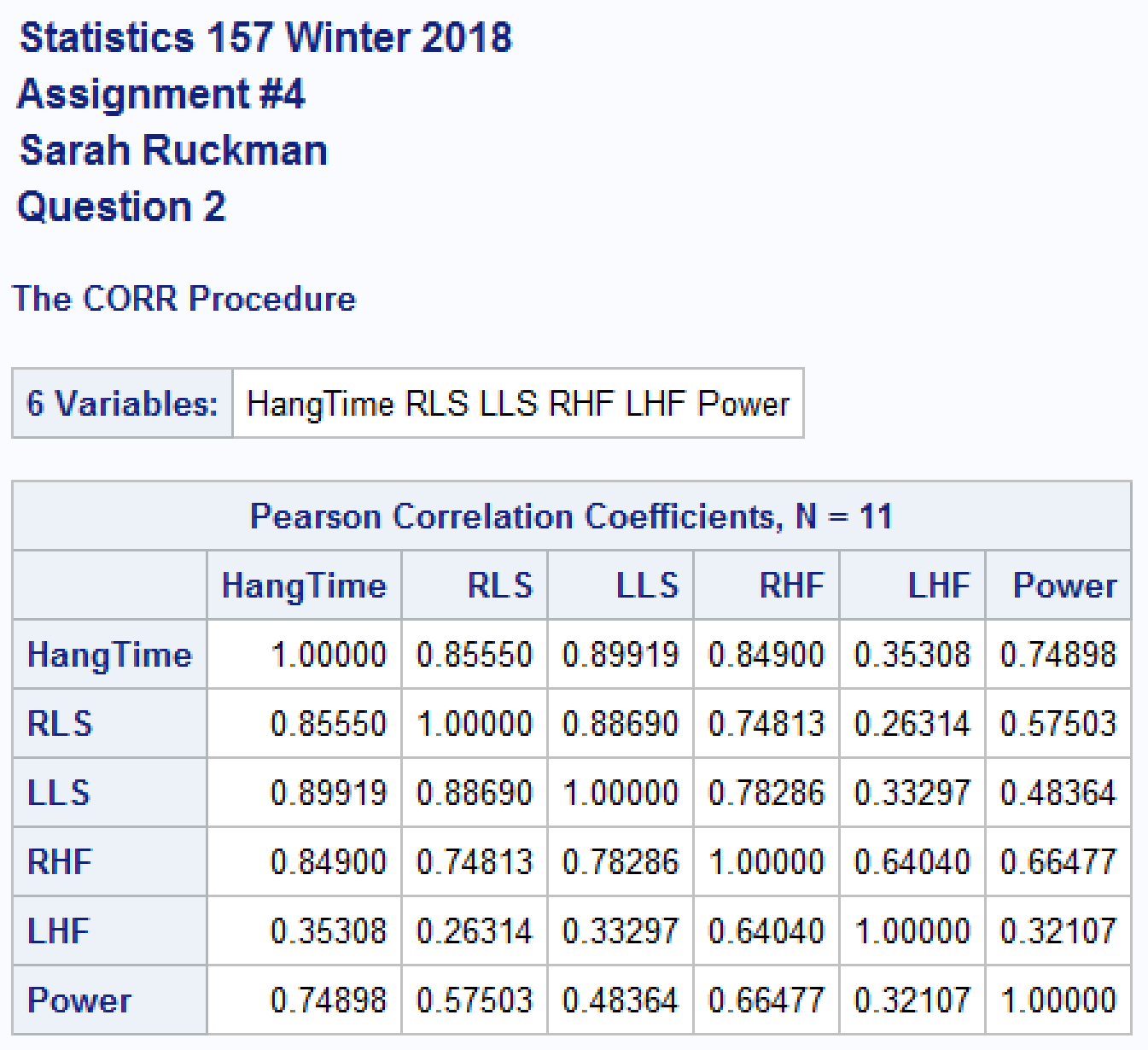
/\*Revise title4\*/

title4 'Question 2';

**run**;

**quit**;

**Output:**



* 1. Identify the two variables (including independent variables and dependent variable) that are the most highly correlated. (Be sure to give the value of the correlation coefficient!) (1 pt)

**The dependent variable, hang time, and independent variable, left leg strength, are most highly correlated with a correlation coefficient value of 0.89919.**

* 1. Identify the two variables (including independent variables and dependent variable) that are the least correlated. (Be sure to give the value of the correlation coefficient!) (1 pt)

**The independent variables right leg strength and left hamstring flexibility are least correlated with a correlation coefficient value of 0.26314.**

* 1. Identify the independent variable that is the most highly correlated with the dependent variable. (Be sure to give the value of the correlation coefficient! (1 pt)

**The independent variable, left leg strength, is most highly correlated with the dependent variable, hang time, with a correlation coefficient value of 0.89919**

1. Find the estimated regression equation which includes each of the variables in a linear fashion. (4 pts)

**SAS Code:**

/\*Use proc reg to find the regression equation for all of the variables\*/

**proc** **reg**;

model HangTime = RLS LLS RHF LHF Power / P R;

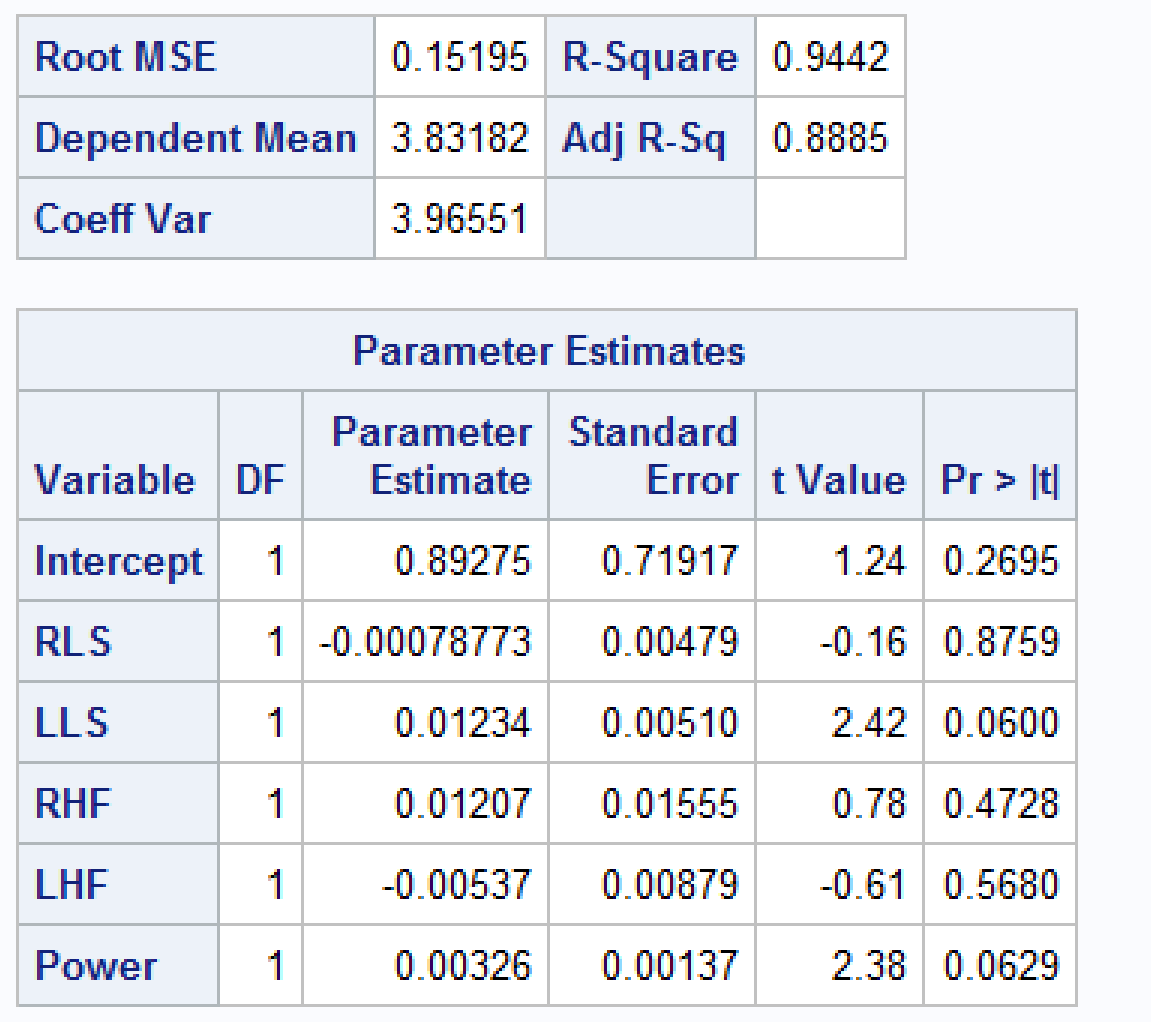
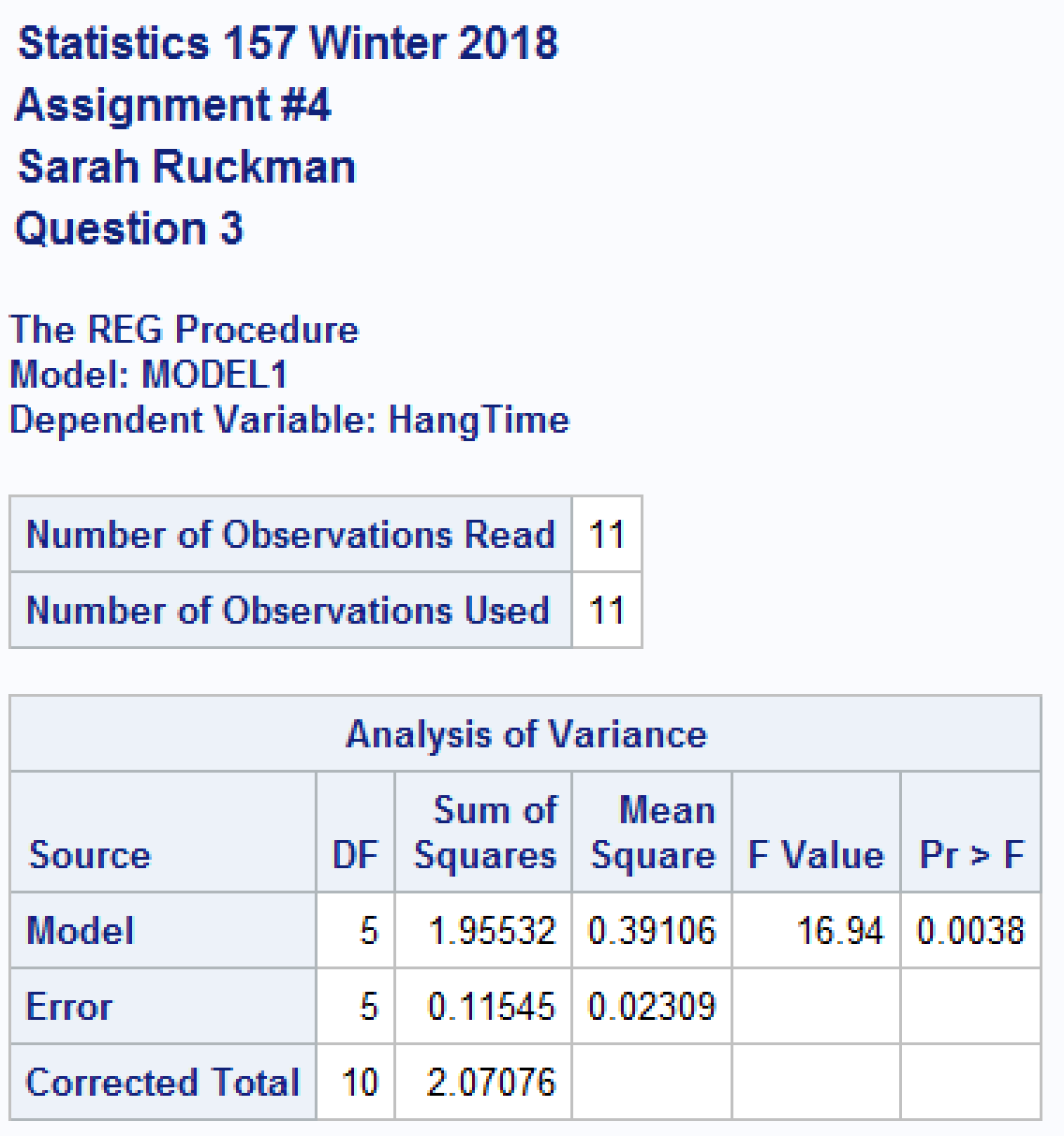
/\*Revise title 4\*/

title4 'Question 3';

**run**;

**quit**;

**Output:**



**The regression equation is:**

**ŷ = 0.89275 - 0.00078773(RLS) + 0.01234(LLS) + 0.01207(RHF) – 0.00537(LHF) + 0.00326(Power)**

* 1. Find and interpret the coefficient of determination. (3 pts)

**SAS Code and Output:**

**See above**

**The coefficient of determination is 0.9442 indicating that 94.42% of the variability in hang time can be explained by the variability in right leg strength, left leg strength, right hamstring flexibility, left hamstring flexibility, and power according to the model.**

* 1. Which variable, if any, would you consider removing from the model first? Justify your answer. (2 pts)

**As some of the values have p-values above α = 0.05, we choose to remove the variable with the highest p-value above α, which is right leg strength with a p-value = 0.8759.**

1. Use best all subsets regression to find the best model with best = 2. Be sure to generate the values of adjrsq, Cp, and the MSE. Indicate your choice for a final model (i.e., the variables that are included in the model). Be sure to justify your selection! (5 pts)

**SAS Code:**

/\*Use proc reg with the options selection = rsquare, adjrsq, mse, cp, and best = 2 to create a new reg equation\*/

**proc** **reg**;

model HangTime = RLS LLS RHF LHF Power / selection = rsquare adjrsq mse cp best=**2**;

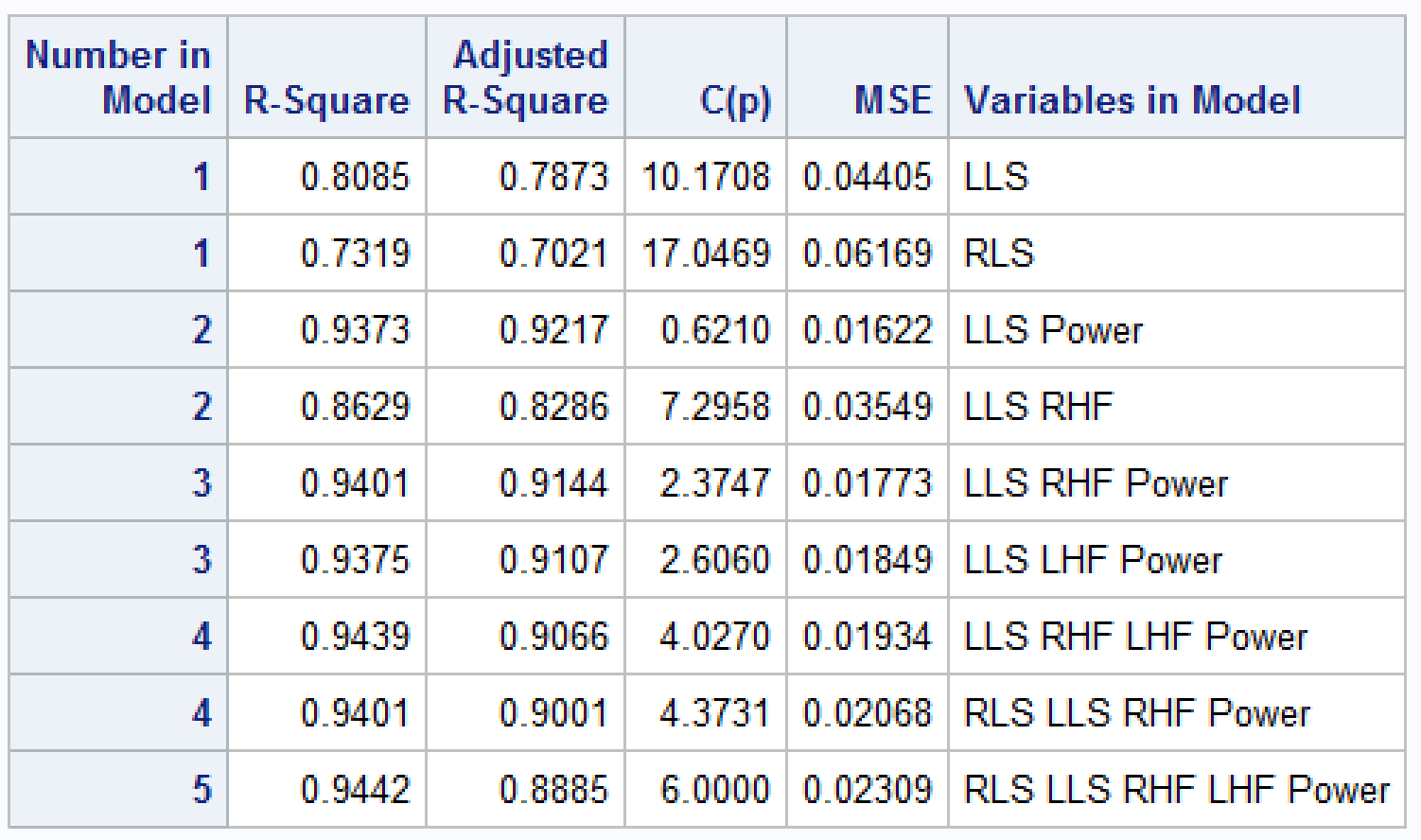
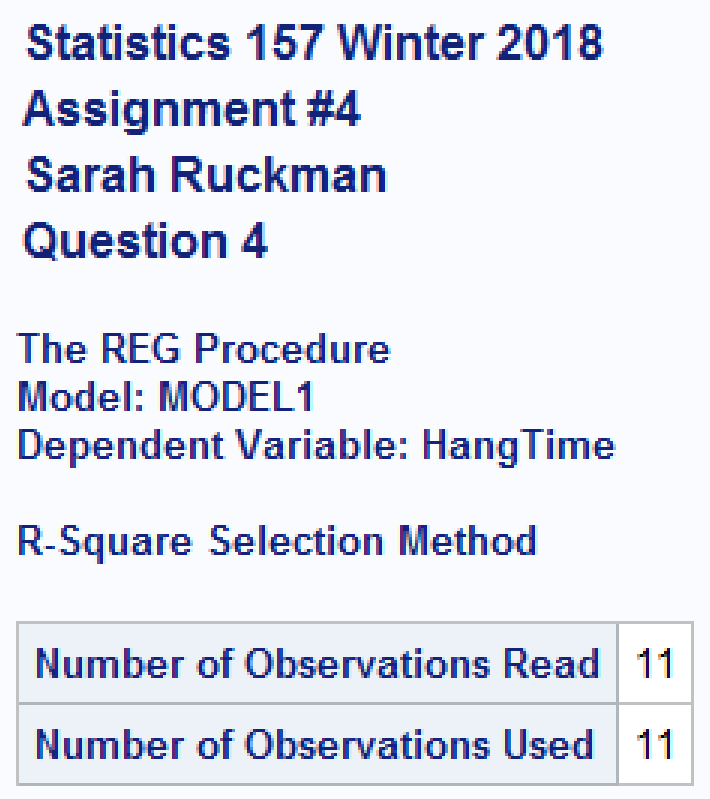
/\*Revise title 4\*/

title4 'Question 4';

**run**;

**quit**;

**Output:**



**The best model according to this selection method is the model with LLS and Power as its variables. This is due to the model having the highest R-square adjusted value and lowest MSE value. It also has a C(p) value less than 3 indicating that this is a good model.**

1. Find the best-fitting model using stepwise regression with alpha to enter = 0.05 and alpha to remove = 0.05. Using this model, complete the following:

**SAS Code:**

/\*Use proc reg with the options selection = stepwise sle = 0.05 entering less than, and leaving alpha sls = 0.05\*/

**proc** **reg**;

model HangTime = RLS LLS RHF LHF Power / P R selection = stepwise sle = **0.05** sls = **0.05**;

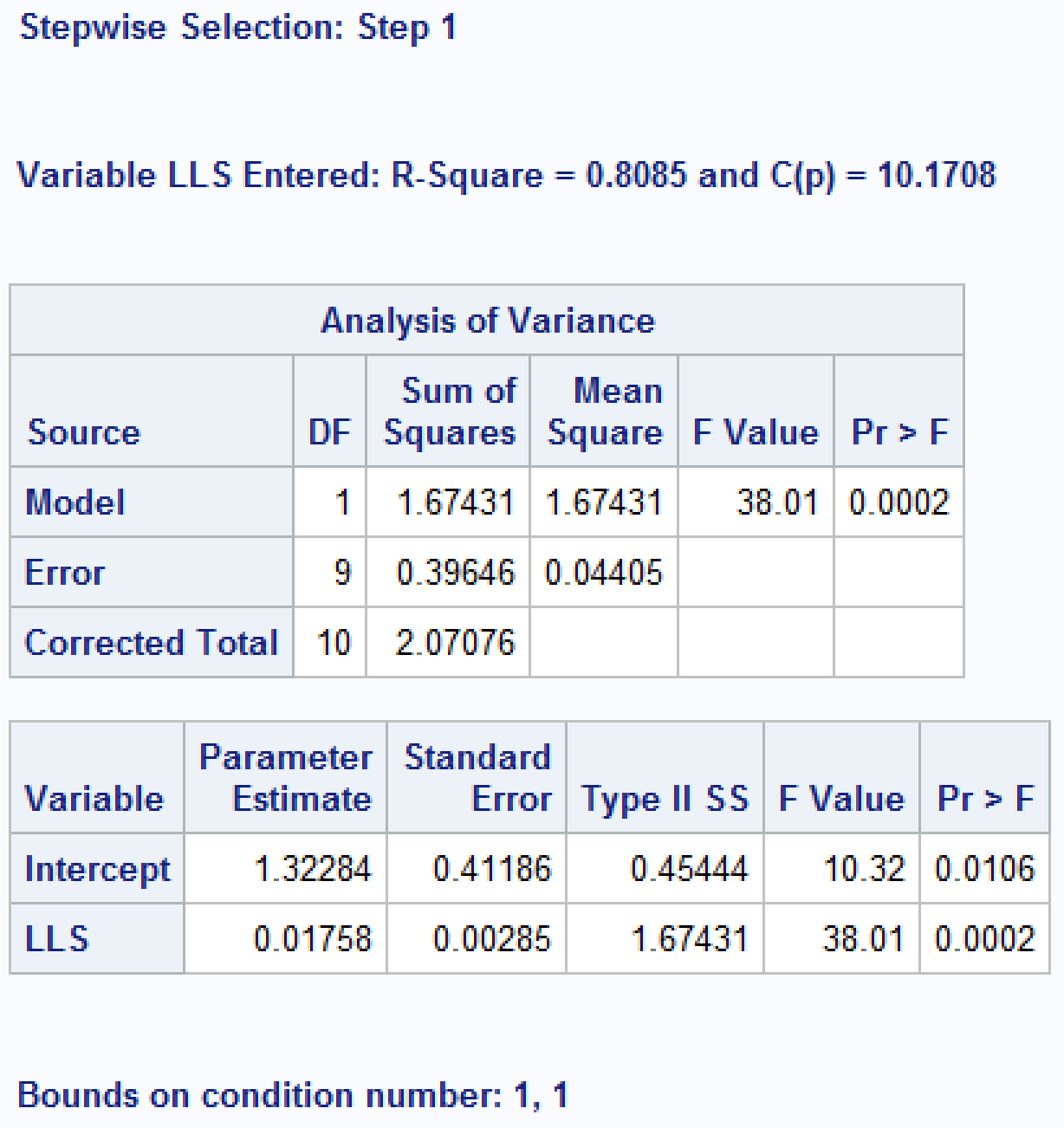
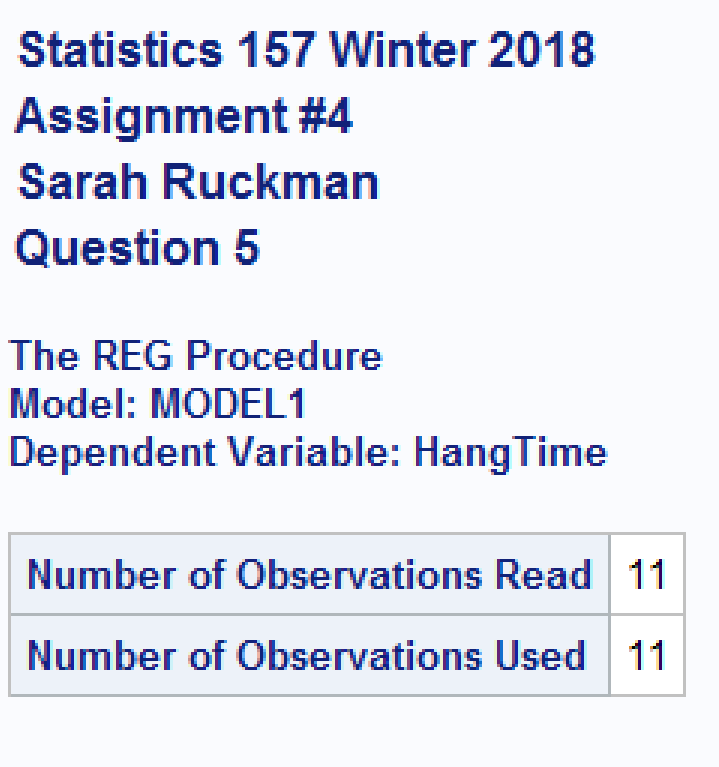
/\*Revise title 4\*/

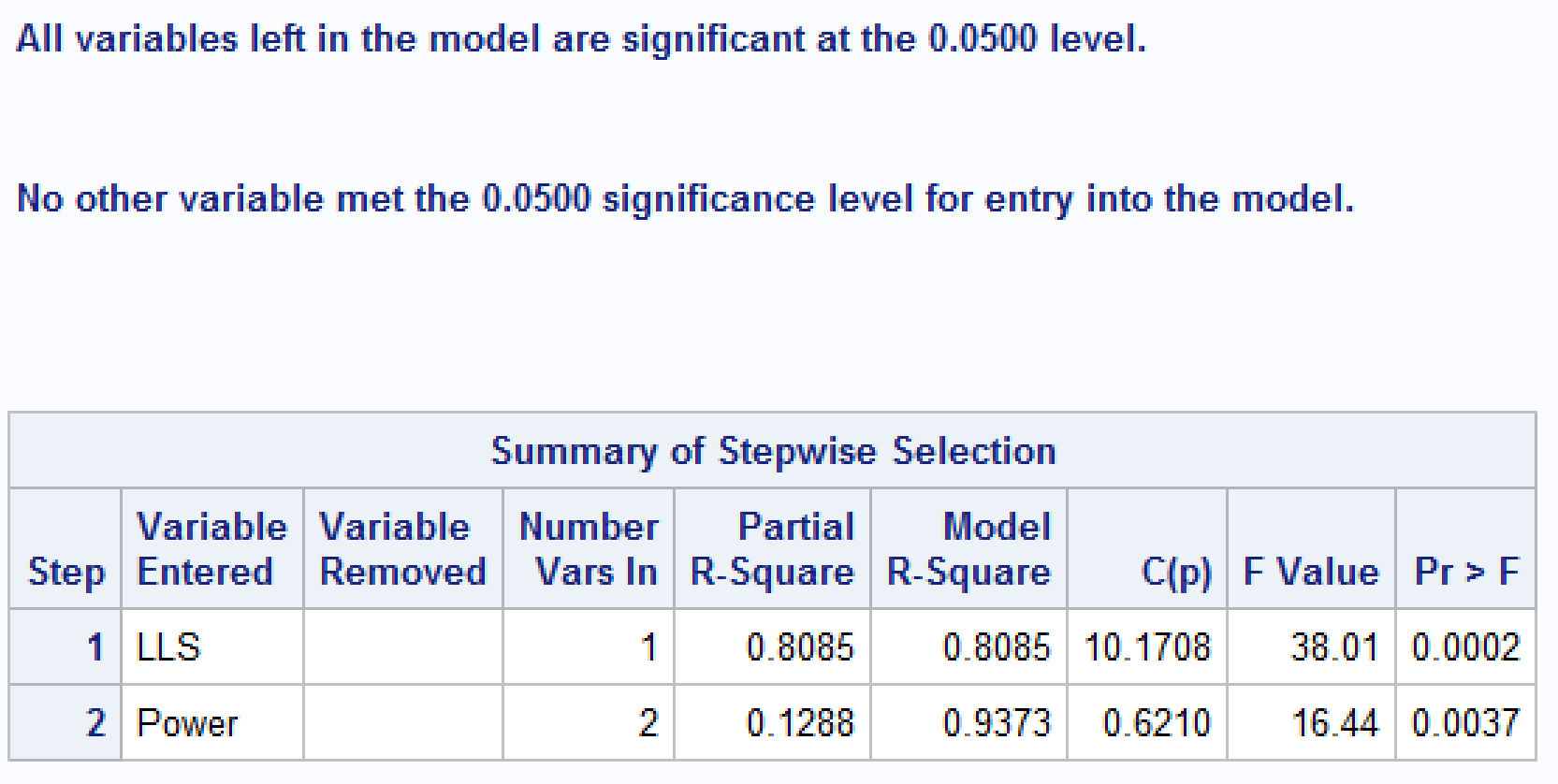
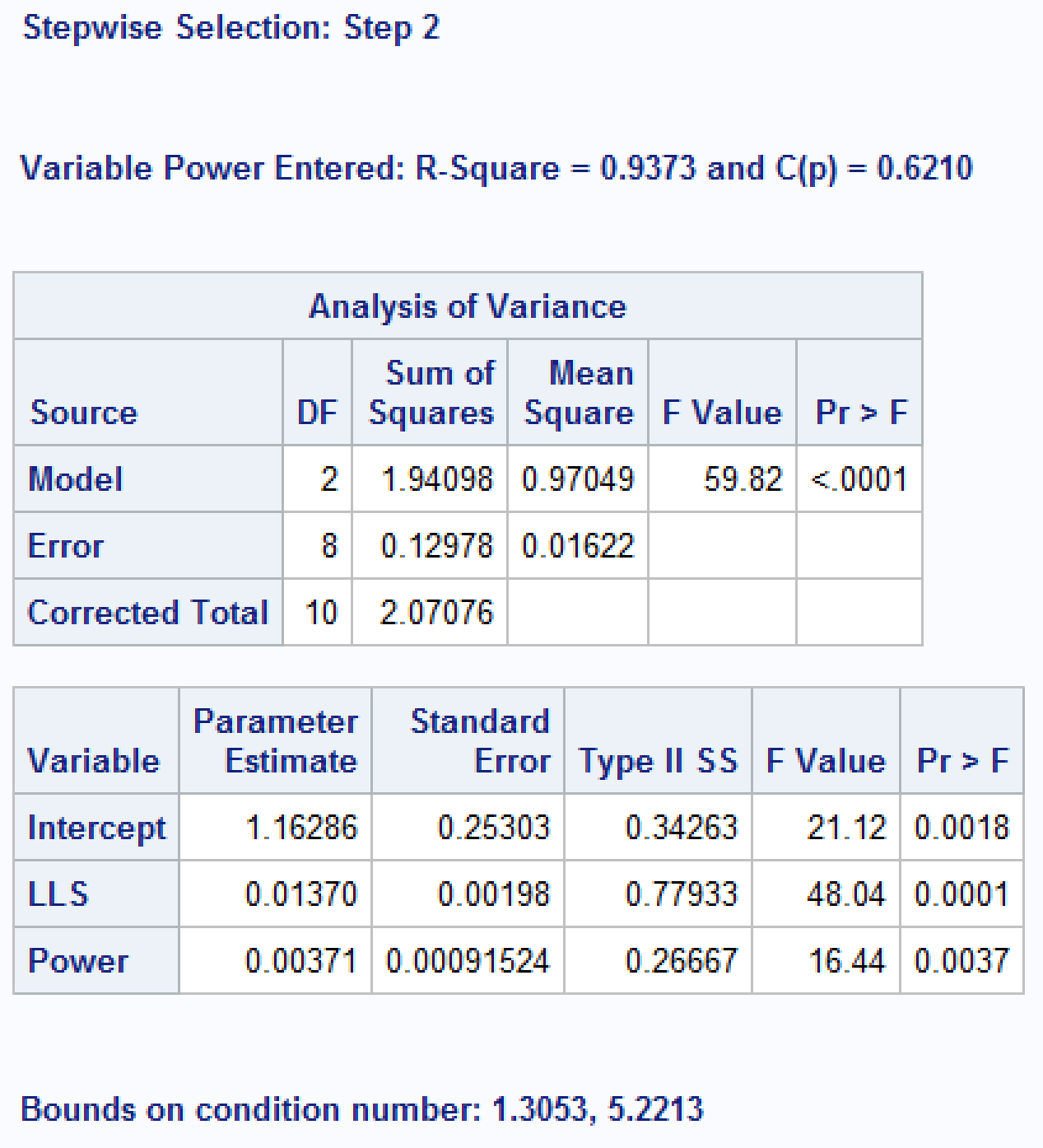
title4 'Question 5';

**run**;

**quit**;

**Output:**





**The best model has the variables LLS and Power according to stepwise selection.**

* 1. Which term enters the model, first? (1 pt)

**LLS, left leg strength enters the model first.**

* 1. Which terms are included in the final model? (2 pts)

**LLS, left leg strength, and Power are included in the final model.**

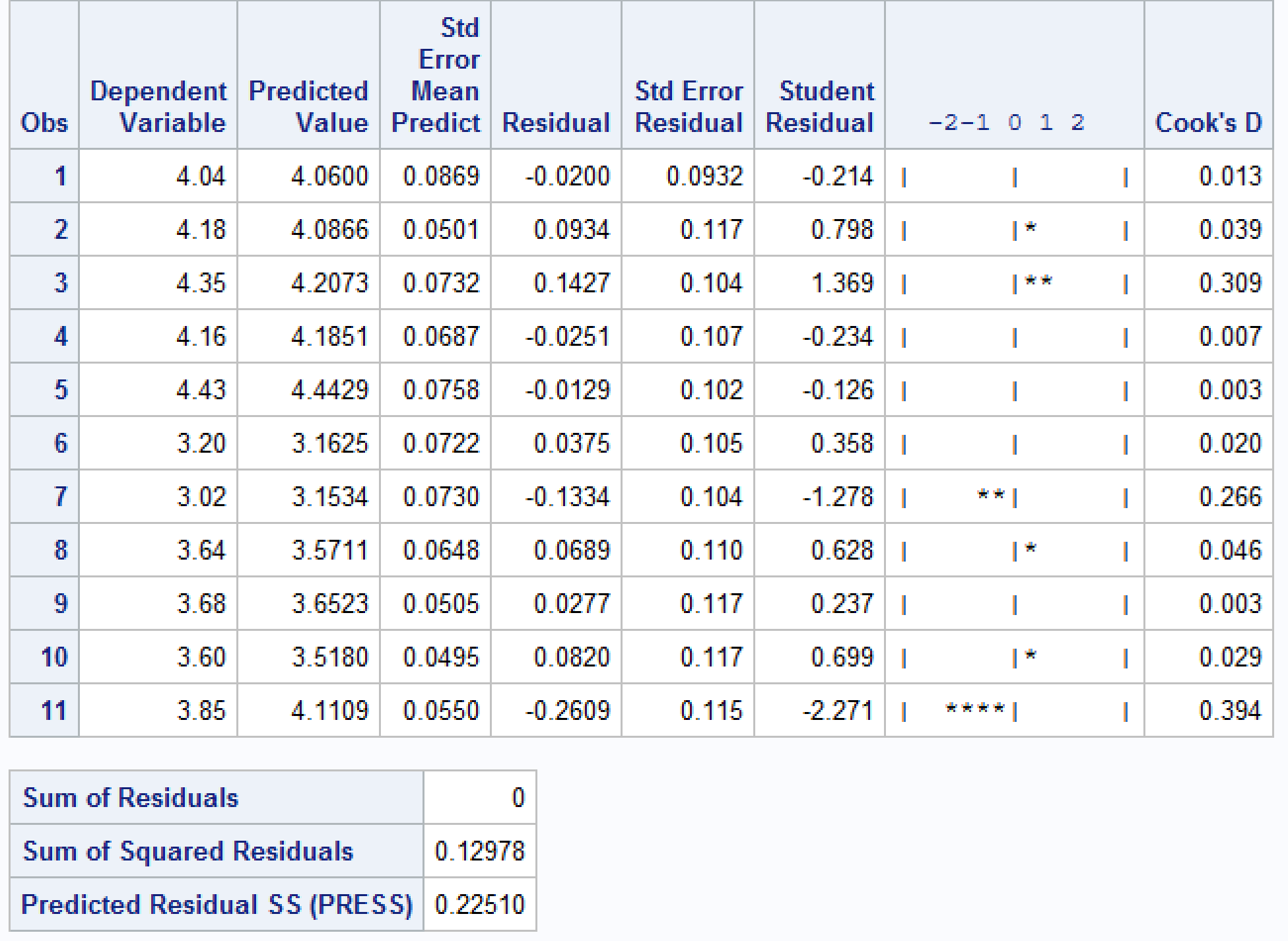
* 1. Find the estimated regression equation (with independent variables in Part (ii)).

**The regression equation is: ŷ = 1.16286 + 0.01370(LLS) + 0.00371(Power)**

* 1. Using the model in part (ii), do there appear to be any suspect or extreme outliers? Justify your answer! (If there are outliers present, identify them!) (2 pts)

**Output:**





**There is one suspect outlier in observation 11 with a student residual value of -2.271. This makes observation 11 a suspect outlier because the residual value lies between -2 and -3.**

* 1. Generate a residual plot (using either the residual or standardized residuals). What, if any, conclusions can be drawn from the plot? (3 pts)

**SAS Code:**

/\*Output the data\*/

output out = hangtime2 P = pred R = resid Student = stdres;

/\*Create a residual plot using proc gplot\*/

**proc** **gplot** data = hangtime2;

/\*Revise title 4 and 5\*/

title4 'Question 5 Part V';

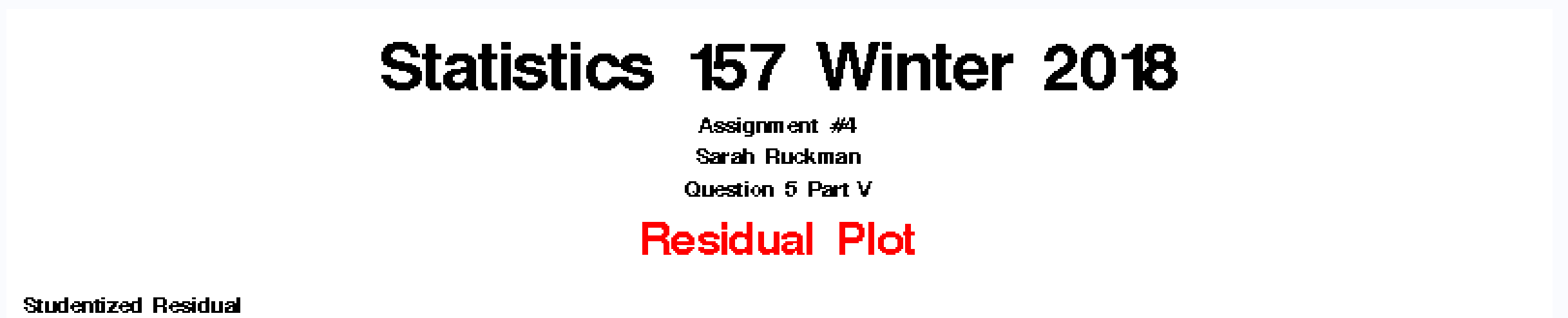
title5 height = **2** color = red 'Residual Plot';

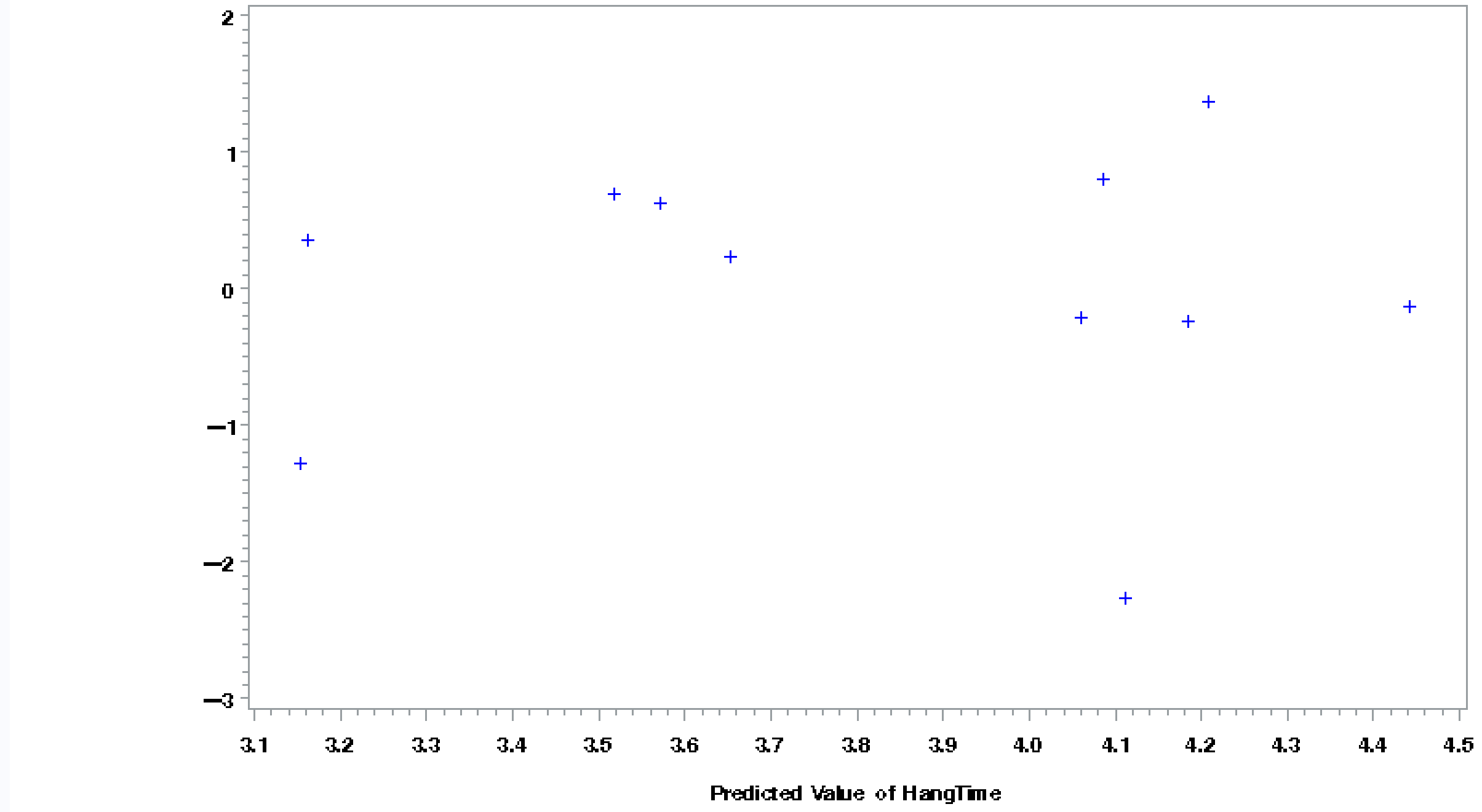
plot stdres\*pred;

**run**;

**quit**;

**Output:**





**The residual plot looks random, but there is a suspect outlier circled above. It is a suspect outlier because it has a studentized residual value between -2 and -3.**

* 1. Test for normality of the errors. (2 pts)

**SAS Code:**

/\*Use proc univariate with the normal option to generate the test for normality information\*/

**proc** **univariate** data = hangtime2 normal;

ods select TestsForNormality;

var stdres;

/\*Revise title 4 and 5\*/

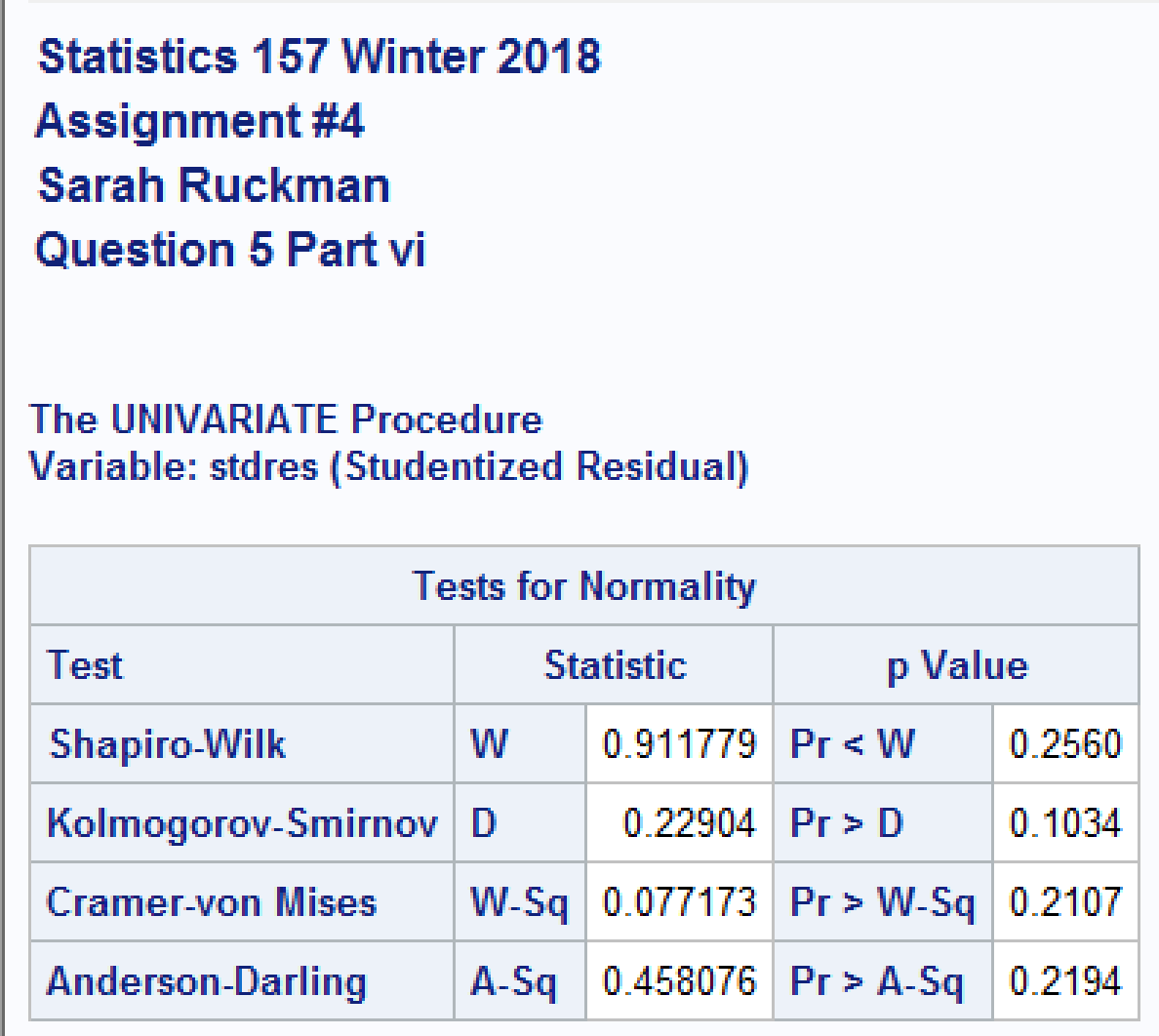
title4 'Question 5 Part vi';

title5 ' ';

**run**;

**quit**;

**Output:**



**H0: Errors are normally distributed.**

**Ha: Errors are not normally distributed.**

**TS = Shapiro-Wilks = 0.911779 with p-value = 0.2560**

**RR: Reject H0 if p-value < α = 0.05**

**Conclusion:**

**Since the p-value = 0.2560 is greater than α = 0.05, We do not reject H0 and it is reasonable to assume the errors are normally distributed.**

1. Refer to Question 5 and the stepwise model. Did you observe any outliers? If so, remove the data point(s) representing the outlier(s), without altering the data file, and generate the new stepwise regression model. Be sure to state the new model! Is this new regression model an improvement over the previous model? [Note: If there were no outliers present, just state that.] (10 pts)

NOTE: Be sure you address the following in Question 6: Coefficient of determination, residuals and possible outliers, residual plot, and normality of the errors

**Yes, the last point was a suspect outlier and was removed and the statistics was rerun:**

**SAS Code:**

/\*Create a new SAS temporary dataset to remove the last data point\*/

**data** football2;

/\*Use an infile statement with firstobs and obs options to remove the last data point\*/

infile 'C:\Users\sarah\Downloads\footballw18.dat' firstobs = **2** obs = **10**;

/\*input the variables\*/

input HangTime RLS LLS RHF LHF Power @@;

/\*Print as check\*/

**proc** **print** noobs;

/\*Revise title 4\*/

title4 'Question 6';

/\*Use proc reg with the options selection = stepwise sle = 0.05 entering less than, and leaving alpha sls = 0.05\*/

**proc** **reg**;

model HangTime = RLS LLS RHF LHF Power / P R selection = stepwise sle = **0.05** sls = **0.05**;

/\*Revise title 5\*/

title5 'Regression Model 2';

/\*Output the data\*/

output out = hangtime3 P = pred1 R = resid1 Student = stdres1;

/\*Create a residual plot using proc gplot\*/

**proc** **gplot** data = hangtime3;

/\*Revise title 5\*/

title5 height = **2** color = red 'Residual Plot of Model 2';

plot stdres1\*pred1;

/\*Use proc univariate with the normal option to generate the test for normality information\*/

**proc** **univariate** data = hangtime3 normal;

ods select TestsForNormality;

var stdres1;

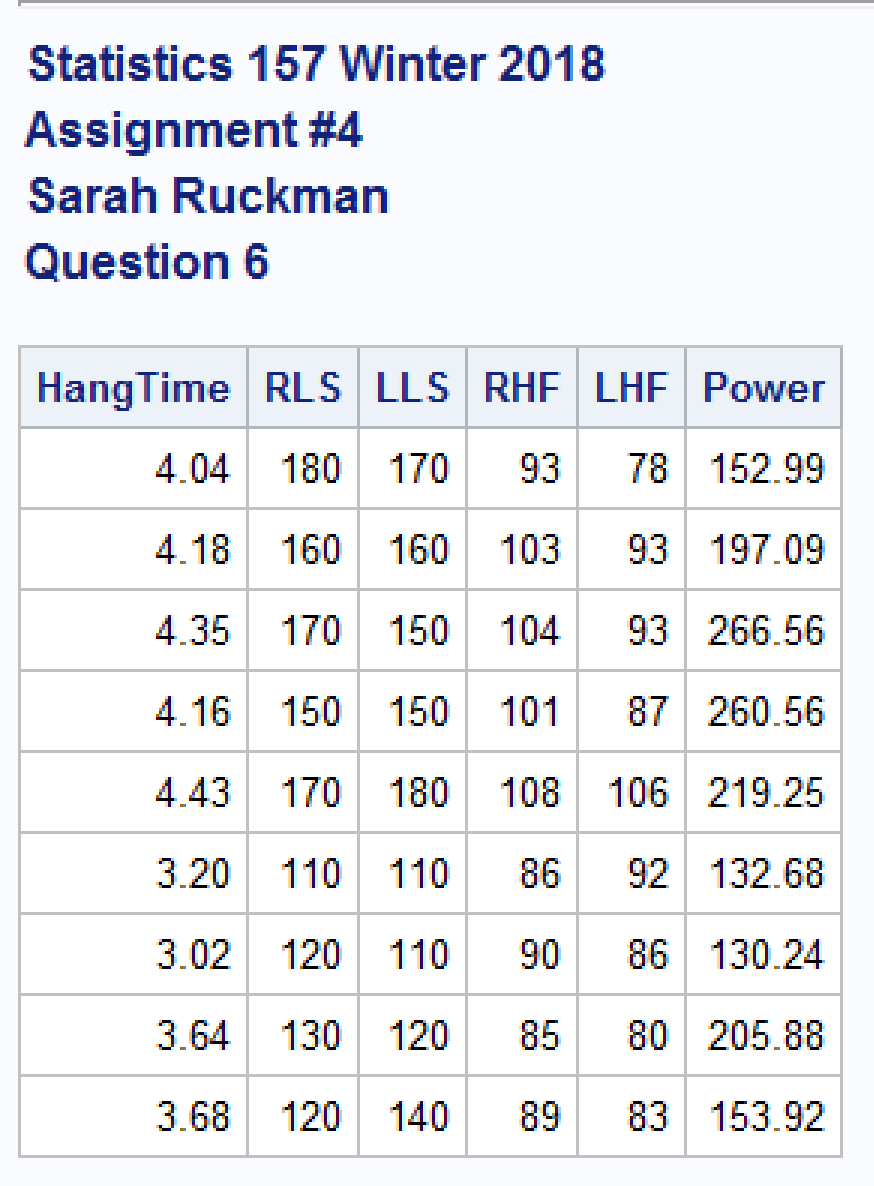
/\*Revise title 5\*/

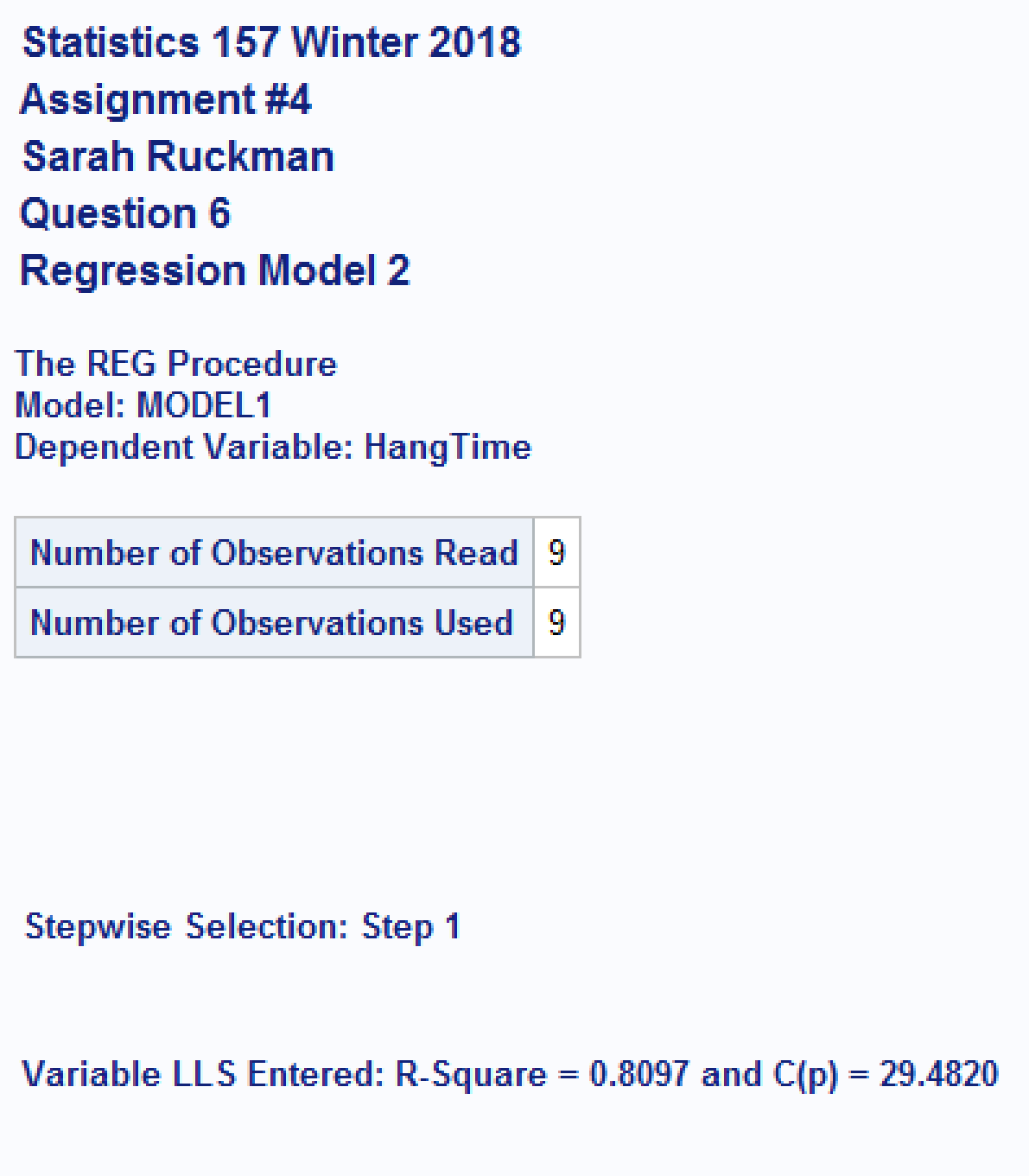
title5 'Test for Normality';

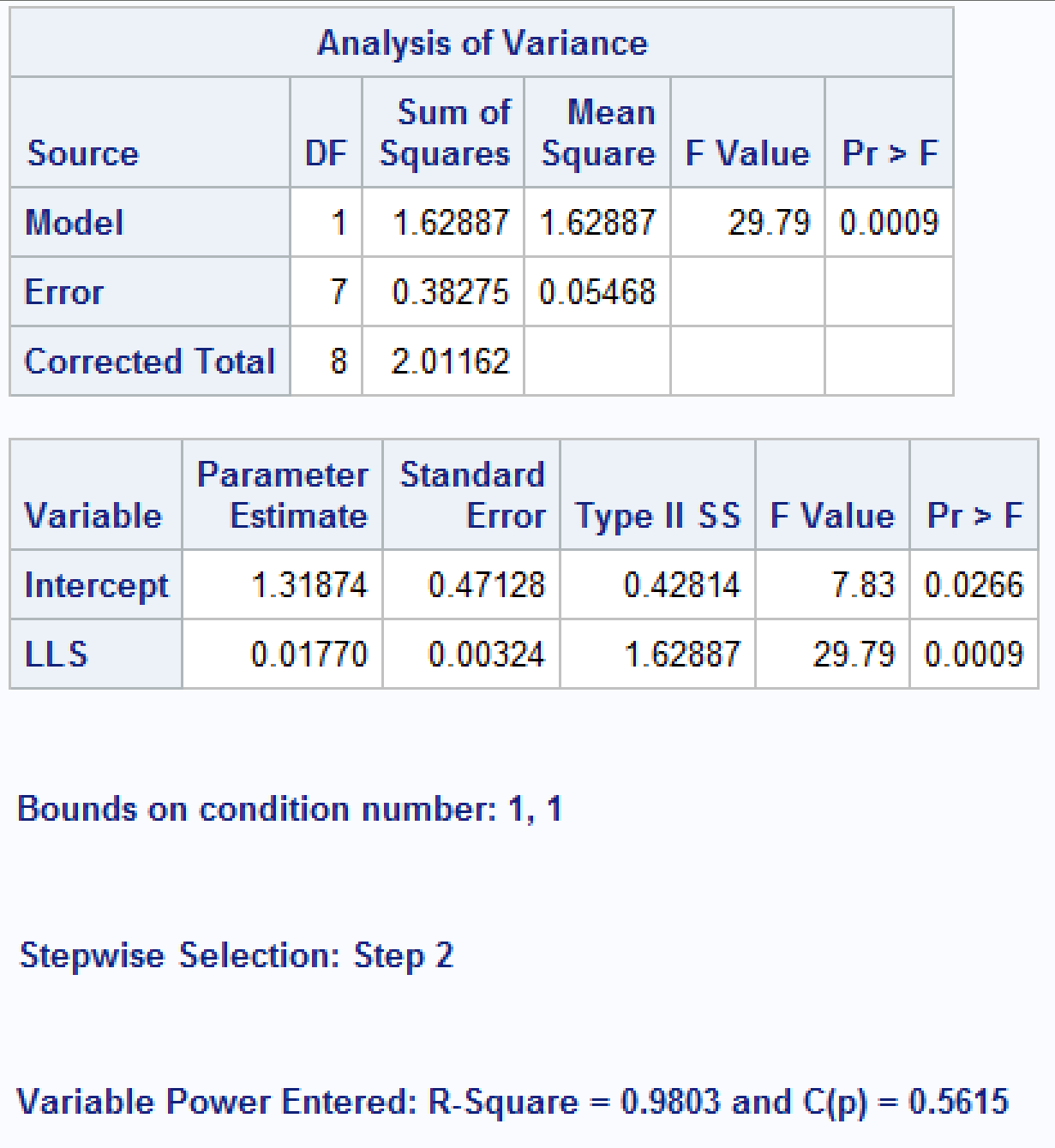
**run**;

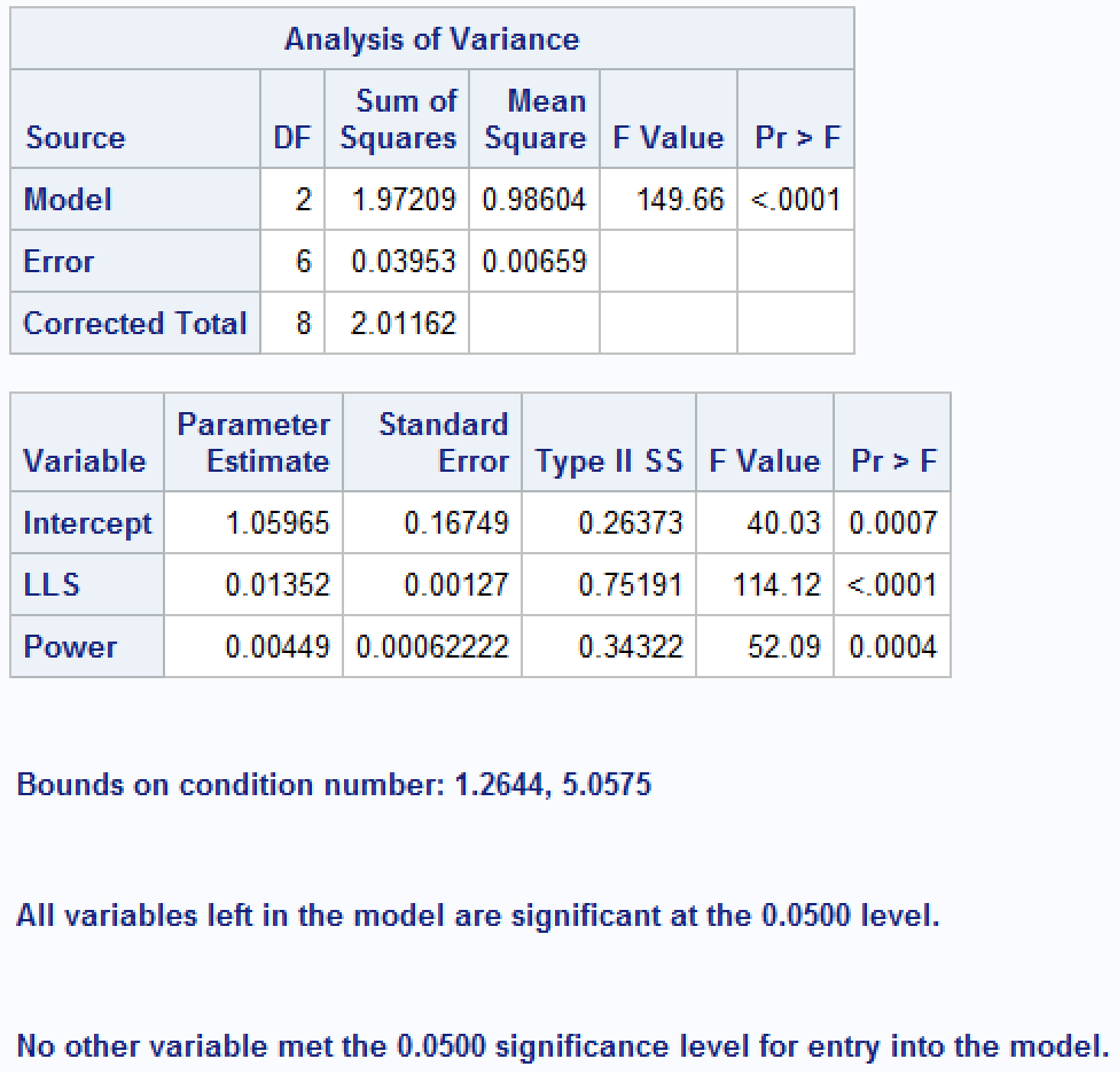
**quit**;

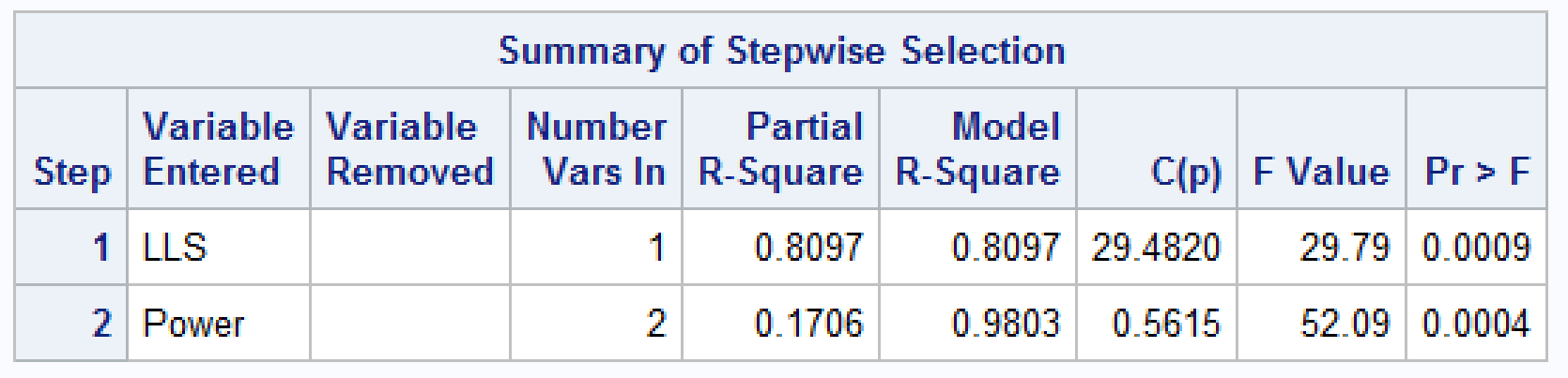
**Output:**



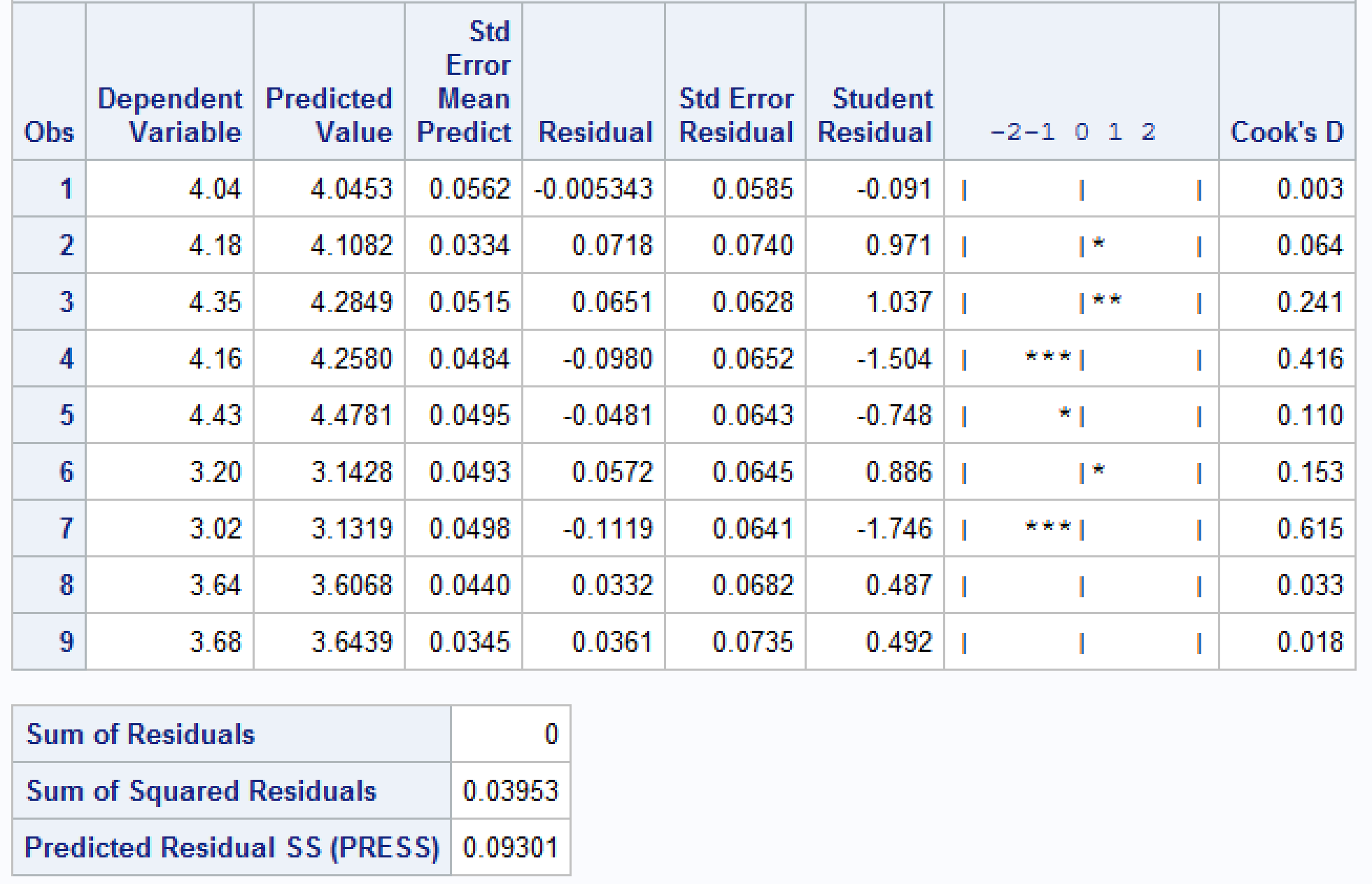


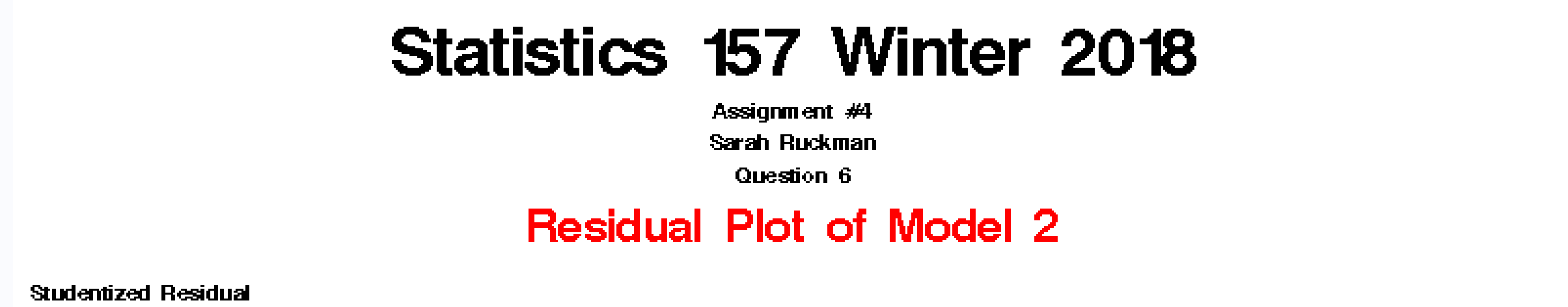


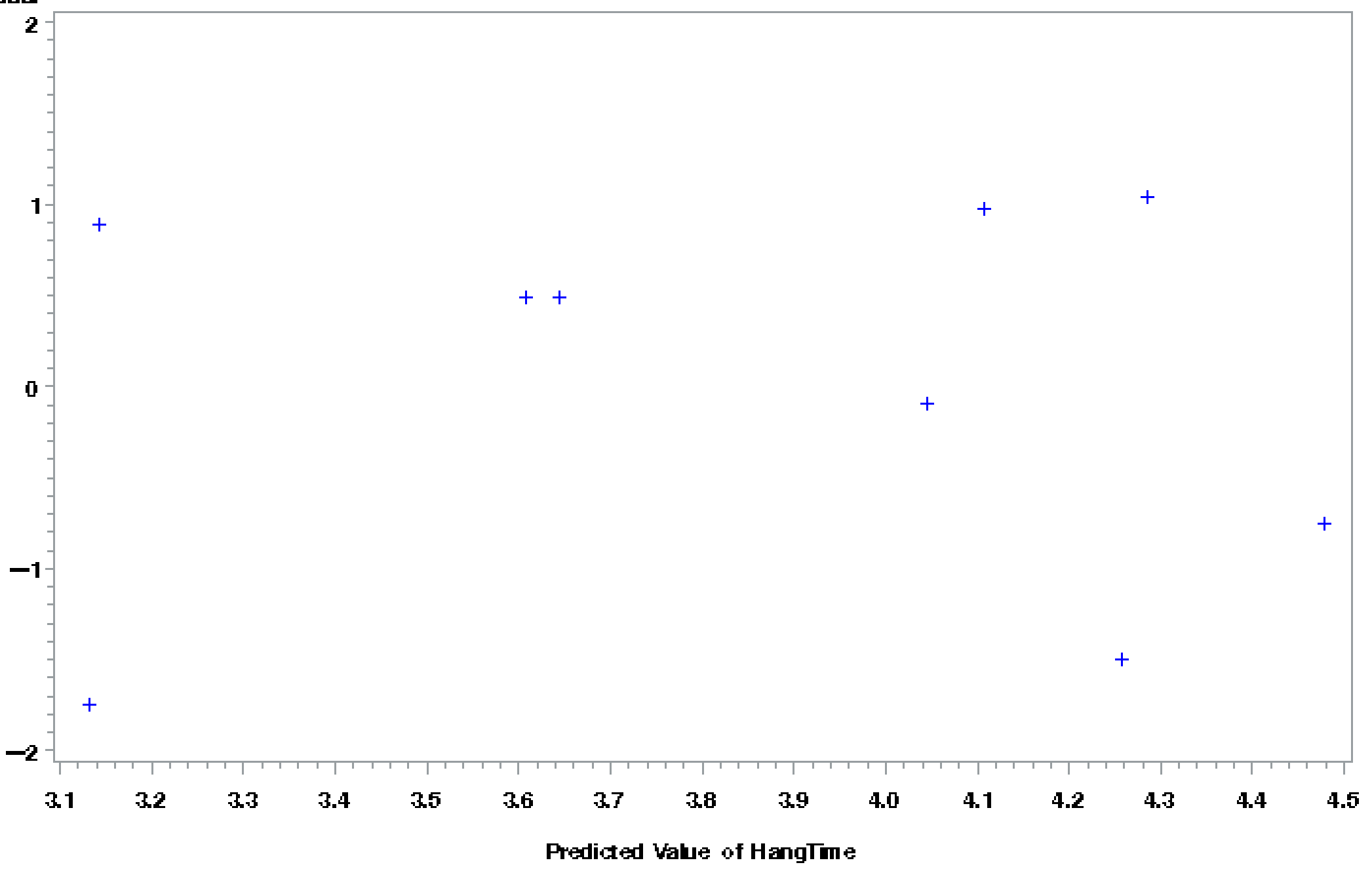


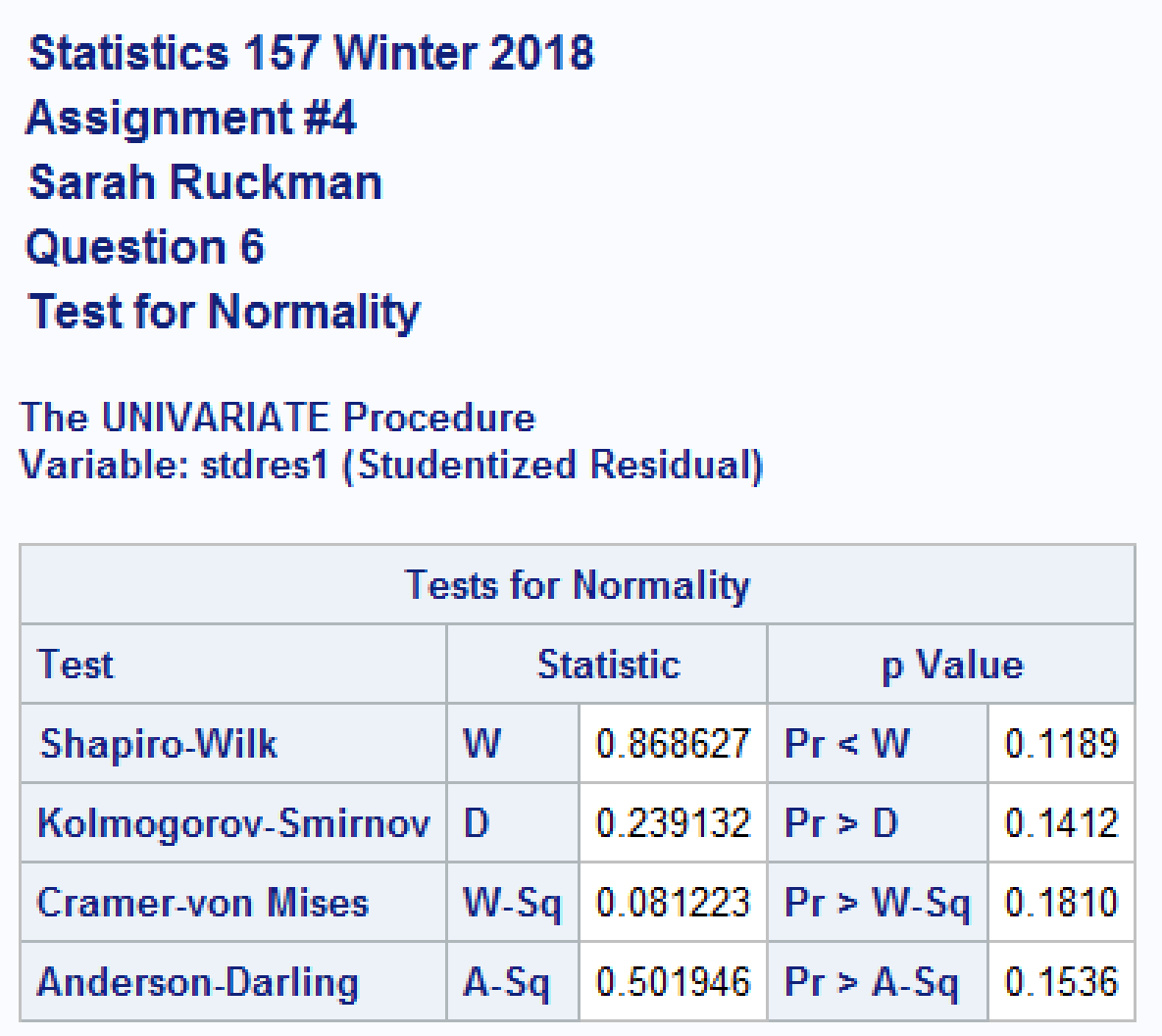












* **The regression equation is ŷ = 1.05965 + 0.01352(LLS) + 0.00449(Power)**
* **Left leg strength (LLS) and Power are the only variables in the model.**
* **The coefficient of determination is 0.9803 indicating that 98.03% of the variability in hang time can be explained by the variability in left leg strength and power according to the model.**
* **Looking at the residual values, there are no suspect or true outliers as all studentized residuals are between -2 and 2. If values were between 2 and 3 in absolute value they would be suspect outliers and if they were above 3 in absolute value they would be outliers.**
* **The residual plot above is random and has no suspect or true outliers as all points lie between 2 and -2.**
* **The test for normality of errors:**

**H0: Errors are normally distributed.**

**Ha: Errors are not normally distributed.**

**TS = Shapiro-Wilks = 0.868627 with p-value = 0.1189**

**RR: Reject H0 if p-value < α = 0.05**

**Conclusion:**

**Since the p-value = 0.1189 is greater than α = 0.05, We do not reject H0 and it is reasonable to assume the errors are normally distributed.**

**Complete SAS Code:**

/\*Set up options and turn off extra graphics\*/

options nocenter nodate nonumber ps=**55** ls=**78**;

ods graphics off;

/\*goptions formats the plot

cback color of the plot background

colors colors to use

ftitle font of plot title

htitle height of the title

htext height of the text on the plot \*/

goptions reset = all colors=(blue,red,green,purple) ftitle = swissb ftext=swissb htitle=**3**;

/\*Set up titles\*/

title1 'Statistics 157 Winter 2018';

title2 'Assignment #4';

title3 'Sarah Ruckman';

title4 'Question 1';

/\*Create new SAS temporary dataset\*/

**data** foot;

/\*Use an infile statement to read in the data \*/

infile 'C:\Users\sarah\Downloads\footballw18.dat' firstobs = **2**;

/\*Input the variables\*/

input HangTime RLS LLS RHF LHF Power @@;

/\*Print as a check\*/

**proc** **print** noobs;

/\*Use proc corr to find the correlation between the variables\*/

**proc** **corr** nosimple noprob;

/\*Revise title4\*/

title4 'Question 2';

/\*Use proc reg to find the regression equation for all of the variables\*/

**proc** **reg**;

model HangTime = RLS LLS RHF LHF Power / P R;

/\*Revise title 4\*/

title4 'Question 3';

/\*Use proc reg with the options selection = rsquare, adjrsq, mse, cp, and best = 2 to create a new reg equation\*/

**proc** **reg**;

model HangTime = RLS LLS RHF LHF Power / selection = rsquare adjrsq mse cp best=**2**;

/\*Revise title 4\*/

title4 'Question 4';

/\*Use proc reg with the options selection = stepwise sle = 0.05 entering less than, and leaving alpha sls = 0.05\*/

**proc** **reg**;

model HangTime = RLS LLS RHF LHF Power / P R selection = stepwise sle = **0.05** sls = **0.05**;

/\*Revise title 4\*/

title4 'Question 5';

/\*Output the data\*/

output out = hangtime2 P = pred R = resid Student = stdres;

/\*Create a residual plot using proc gplot\*/

**proc** **gplot** data = hangtime2;

/\*Revise title 4 and 5\*/

title4 'Question 5 Part V';

title5 height = **2** color = red 'Residual Plot';

plot stdres\*pred;

/\*Use proc univariate with the normal option to generate the test for normality information\*/

**proc** **univariate** data = hangtime2 normal;

ods select TestsForNormality;

var stdres;

/\*Revise title 4 and 5\*/

title4 'Question 5 Part vi';

title5 ' ';

/\*Create a new SAS temporary dataset to remove the last data point\*/

**data** football2;

/\*Use an infile statement with firstobs and obs options to remove the last data point\*/

infile 'C:\Users\sarah\Downloads\footballw18.dat' firstobs = **2** obs = **10**;

/\*input the variables\*/

input HangTime RLS LLS RHF LHF Power @@;

/\*Print as check\*/

**proc** **print** noobs;

/\*Revise title 4\*/

title4 'Question 6';

/\*Use proc reg with the options selection = stepwise sle = 0.05 entering less than, and leaving alpha sls = 0.05\*/

**proc** **reg**;

model HangTime = RLS LLS RHF LHF Power / P R selection = stepwise sle = **0.05** sls = **0.05**;

/\*Revise title 5\*/

title5 'Regression Model 2';

/\*Output the data\*/

output out = hangtime3 P = pred1 R = resid1 Student = stdres1;

/\*Create a residual plot using proc gplot\*/

**proc** **gplot** data = hangtime3;

/\*Revise title 5\*/

title5 height = **2** color = red 'Residual Plot of Model 2';

plot stdres1\*pred1;

/\*Use proc univariate with the normal option to generate the test for normality information\*/

**proc** **univariate** data = hangtime3 normal;

ods select TestsForNormality;

var stdres1;

/\*Revise title 5\*/

title5 'Test for Normality';

**run**;

**quit**;