**Assignments for Lesson 10**

1. Refer to the GRADES data. Suppose that the instructor of the class wants to see if students performed at consistent levels during the semester. There would be a problem with the grading procedure if, for example, students who earned high grades at the beginning of the semester tended to have lower grades toward the end of the semester, or if students who performed well in one week performed poorly the next week. One way to evaluate this consistency numerically is to use a split-half reliability coefficient. Choose one way to divide the 13 homework grades into two groups: one with seven assignments, one with six assignments. For example, you may choose to divide the assignment into early and late assignments, odd-number and even-number assignments, or a randomly chosen group of seven and the remaining six assignments. Then calculate the correlation between the total of the first group of assignments and the total of the second group of assignments. Typically, for a grading procedure to be considered ‘reliable’, this correlation should be 0.7 or higher. Would you conclude that the grading policy is reliable from your calculations?

CODE

**data** grades;

infile '\\Client\C$\Users\Leland\Desktop\ISM6930 SAS\Data\grades.txt' dlm='/';

input student **1**-**4** Week1 **6** Week2 **7** Week3 **8** Week4 **9** Week5 **10** Week6 **11** Week7 **12** Week8 **13** Week9 **14** Week10 **15** Week11 **16** Week12 **17** Week13 **18**;

array grades (**13**) Week1-Week13;

TotalEarly= sum(of week1-week7);

TotalLate= sum(of week8-week13);

**run**;

**proc** **corr** data=grades outp=gradescorr;

var totalearly ;

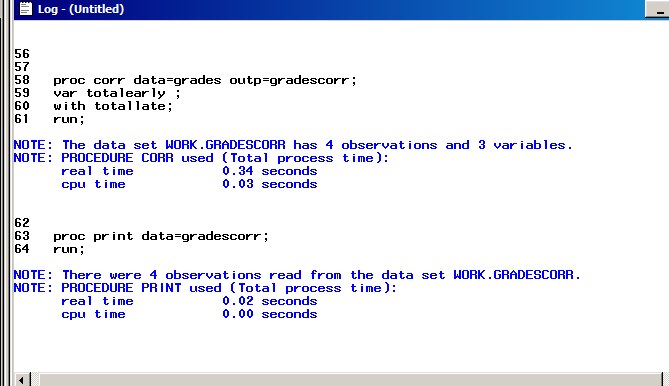
with totallate;

**run**;

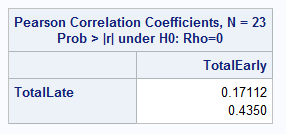
**proc** **print** data=gradescorr;

**run**;

LOG



OUTPUT



1. Refer to the IRIS data. Find all of the pairwise correlations among sepal length, sepal width, petal length and petal width, using all of the observations. Now, calculate the averages of sepal length, sepal width, petal length, and petal width separately for each of the three species. You can do this within PROC UNIVARIATE by listing all four variables in the VAR statement, then use an OUTPUT statement similar to this:

OUTPUT OUT=IRISAVGS MEAN=AVGSL AVGSW AVGPL AVGPW;

Next, find the correlations among the four averages. For example, one of these is the correlation between the average sepal length and the average sepal width, calculated from three data points: one from *setosa* averages, one from *versicolor* averages, and one from *virginica* averages. The first correlations are called total correlations, while the correlations between averages are called between-groups correlations. These two types of correlations are used in Fisher’s linear discriminant analysis to find mathematical expressions which describe the differences among the three species.

**CODE**

**data** iris;

infile '\\Client\C$\Users\Leland\Desktop\ISM6930 SAS\Lesson 4\Homework\iris.txt' firstobs=**2**;

input class $ SL SW PL PW;

label SL = 'Sepal Length'

SW = 'Sepal Width'

PL = 'Pupil Length'

PW = 'Pupil Width';

**run**;

\*descending goes before the var;

**proc** **sort** data=iris;

by class;

**run**;

**proc** **corr** data= iris outp=irisco;

var sl sw pl pw;

**run**;

**proc** **univariate** data=iris noprint;

title 'iris univariate';

var SL SW PL PW;

by class;

OUTPUT OUT=IRISAVGS MEAN=AVGSL AVGSW AVGPL AVGPW;

**run**;

**proc** **corr** data=irisavgs outp=iriscorr noprint;

var AVGSL AVGSW AVGPL AVGPW;

**run**;

**proc** **print** data=irisco ;

title 'Iris Correlation of Class';

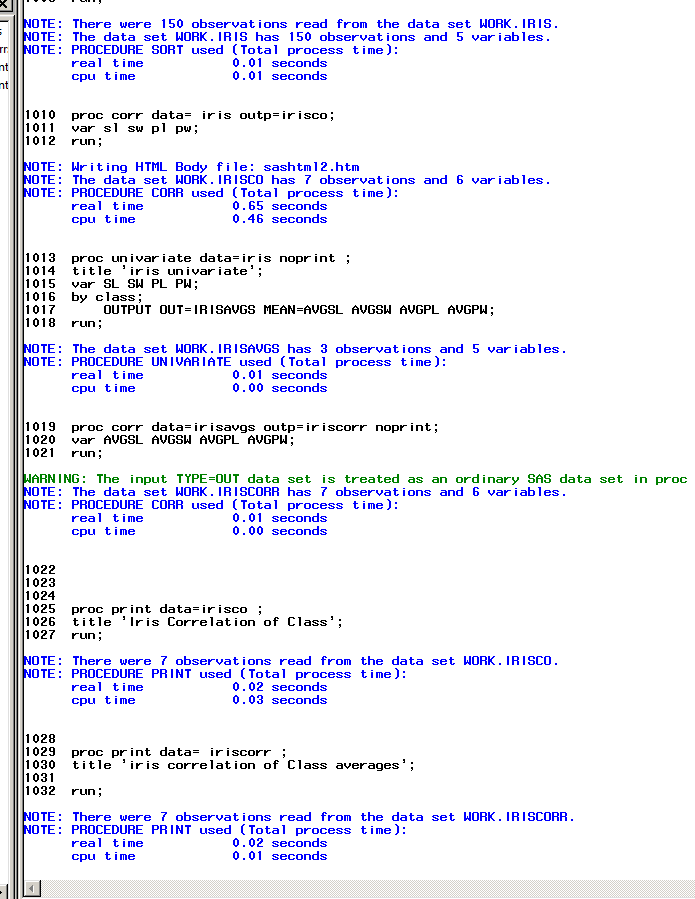
**run**;

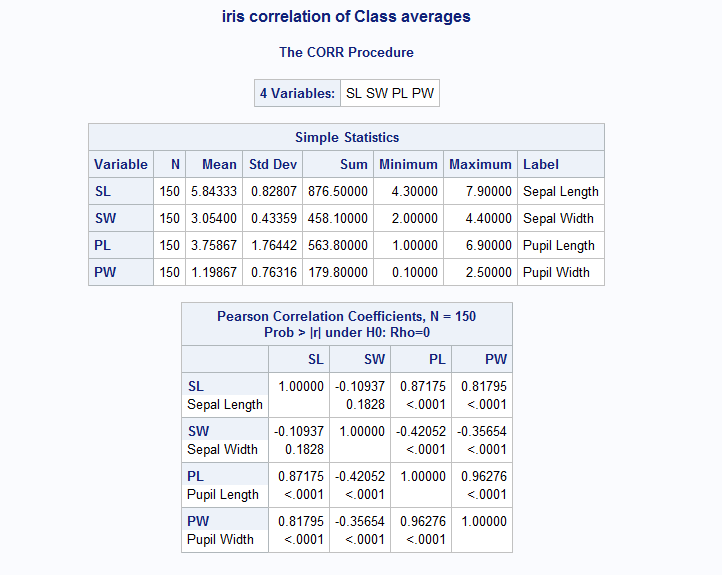
**proc** **print** data= iriscorr ;

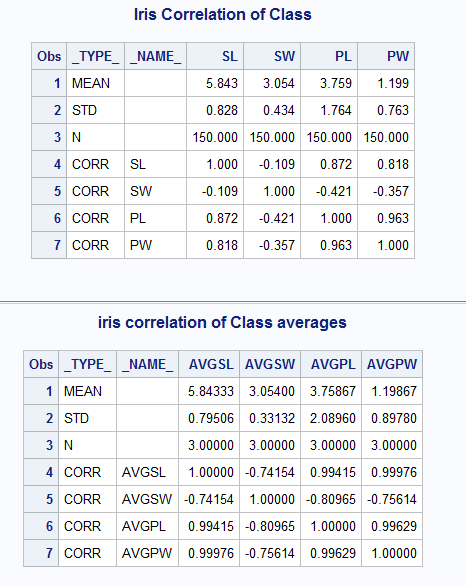
title 'iris correlation of Class averages';

**run**;

**LOG**

 **OUTPUT**





1. Sanford Weisberg listed the following example in his book *Applied Linear Regression, Second Edition* (New York, Wiley, 1985). Consider the following data.

X Y1 Y2 Y3

10 8.04 9.14 7.46

8 6.95 8.14 6.77

13 7.58 8.74 12.74

9 8.81 8.77 7.11

11 8.33 9.26 7.81

14 9.96 8.10 8.84

6 7.24 6.13 6.08

4 4.26 3.10 5.39

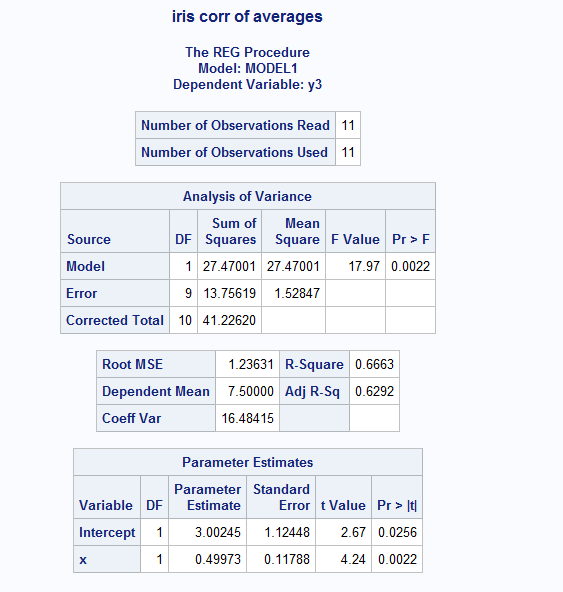
12 10.84 9.13 8.15

7 4.82 17.26 6.42

5 5.66 4.74 5.73

On three separate scatterplots, plot Y1, Y2, and Y3 on the vertical axis versus X on the horizontal axis. Then, calculate the linear regression line to model Y1 using X as a predictor. Do the same for Y2 using X and for Y3 using X. Write down your estimates from the regression equations in the following table: **How do you determine the slope?**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Response Variable | Intercept | Slope | R2 | Mean Square Error |
| Y1 | 3 | .5 | .67 | 1.53 |
| Y2 | 5.5 | .32 | .09 | 12.91 |
| Y3 | 3 | .5 | .67 | 1.53 |



CODE

**data** num;

input x y1 y2 y3;

datalines;

10 8.04 9.14 7.46

8 6.95 8.14 6.77

13 7.58 8.74 12.74

9 8.81 8.77 7.11

11 8.33 9.26 7.81

14 9.96 8.10 8.84

6 7.24 6.13 6.08

4 4.26 3.10 5.39

12 10.84 9.13 8.15

7 4.82 17.26 6.42

5 5.66 4.74 5.73

;

**run**;

**proc** **gplot** data=num uniform;

plot y1\*x y2\*x y3\*x;

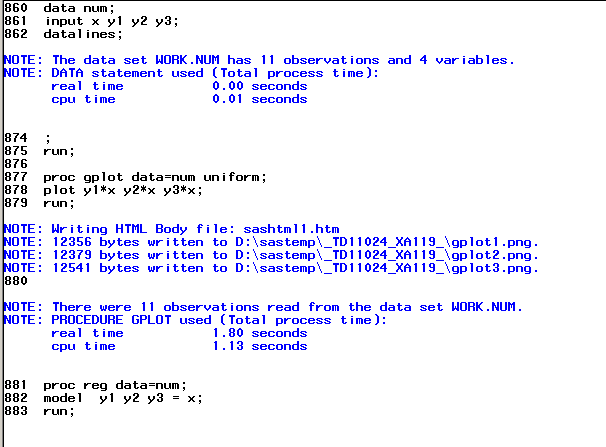
**run**;

**proc** **reg** data=num;

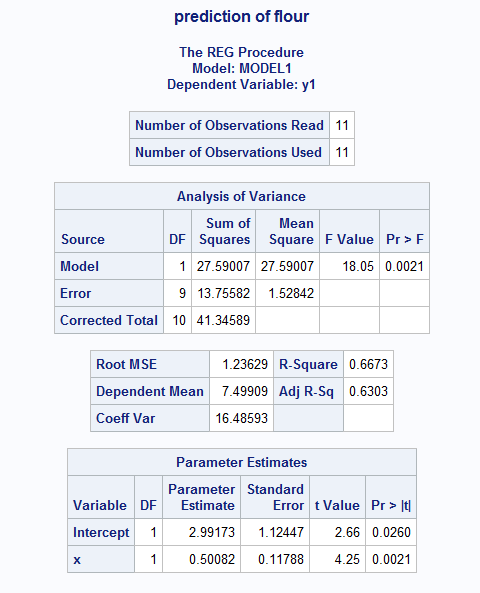
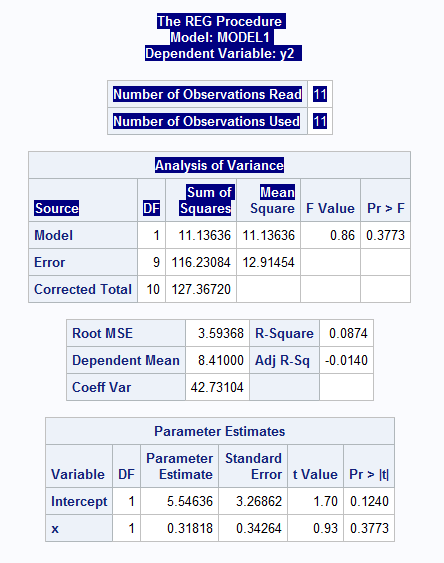
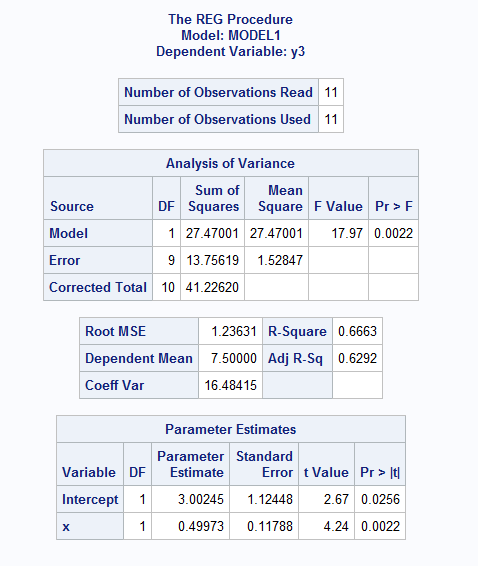
model y1 y2 y3 = x;

**run**;

LOG



OUTPUT

1. I was trying out a new bread recipe the other day. I spilled something on the recipe booklet, and I can’t read how much flour I’m supposed to use in the recipe. I do know that I need to use 1 cup of water, 2 tablespoons of oil, 2 tablespoons of sugar, 1 ½ teaspoons of salt, and 2 ¼ teaspoons of yeast.

Help me out. Refer to BREAD data. Find the least-squares regression equation to predict flour amounts from water, oil, sugar, salt, and yeast, and use that equation to estimate how much flour I need in my recipe. Make sure that SAS prints the estimated amount of flour needed.

Would this be a proc glm anova?

CODE

**data** bread;

infile '\\Client\C$\Users\Leland\Desktop\ISM6930 SAS\data\bread.txt' dlm = ','firstobs = **3**;

input Dough $ Water Oil Sugar Salt DryMilk Flour Yeast Wheat Oregano Eggs;

**run**;

**proc** **print** data=bread;

**run**;

**data** bread2;

input Dough $ Water Oil Sugar Salt DryMilk Flour Yeast Wheat Oregano Eggs;

datalines;

new\_bread 1 2 2 1.5 . . 2.15 . . .

;

**run**;

**data** bread3;

set bread bread2;

**run**;

**proc** **reg** data=bread3;

model flour= water oil sugar salt yeast;

output out= new predicted= pred;

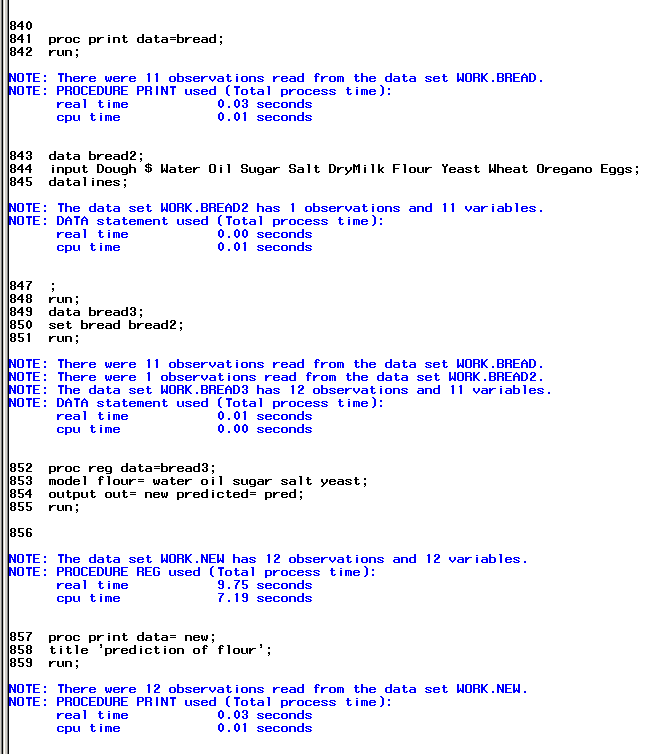
**run**;

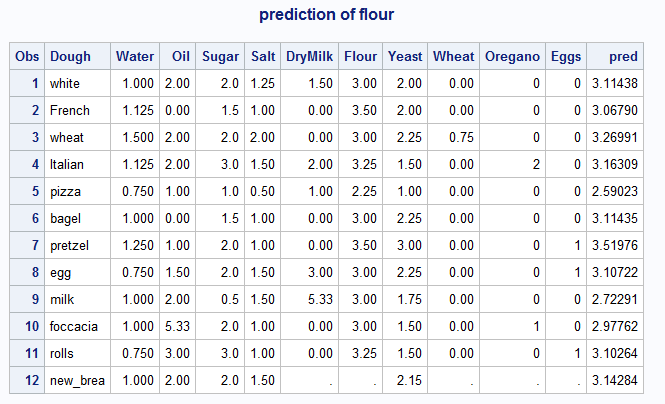
**proc** **print** data= new;

title 'prediction of flour';

**run**;

LOG

  
OUTPUT



1. Refer to the USEDCARS data. Calculate the regression line to predict the price of a used car based on the year in which it was manufactured. Obtain the residuals from this regression model, and use the PROC UNIVARIATE to examine their distribution. In the PROC UNIVARIATE output, identify the largest five and smallest five residuals by the name of the used car dealer. Do the residuals appear to be normally distributed, as we assume when conducting the t- and F- tests?

CODE

**data** usedcars;

infile '\\Client\C$\Users\Leland\Desktop\ISM6930 SAS\Lesson 4\Homework\usedcars.txt'

firstobs=**2** obs=**51** dsd missover;

input Year **1**-**2**

@**9** Manufacturer :$10.

@**24** Model $11.

@**38** Miles comma7.0

@**49** Price dollar10.0

@**61** Dealer $23. ;

label Miles = 'Milage'

Price = 'Cost';

**run**;

**proc** **sort** data=usedcars;

by price;

**run**;

**proc** **reg** data=usedcars;

model price=year;

output out= usedreg

residual= rescar;

**run**;

**proc** **sort** data=usedreg;

by dealer;

**run**;

ods select extremeobs;

**proc** **univariate** data=usedreg;

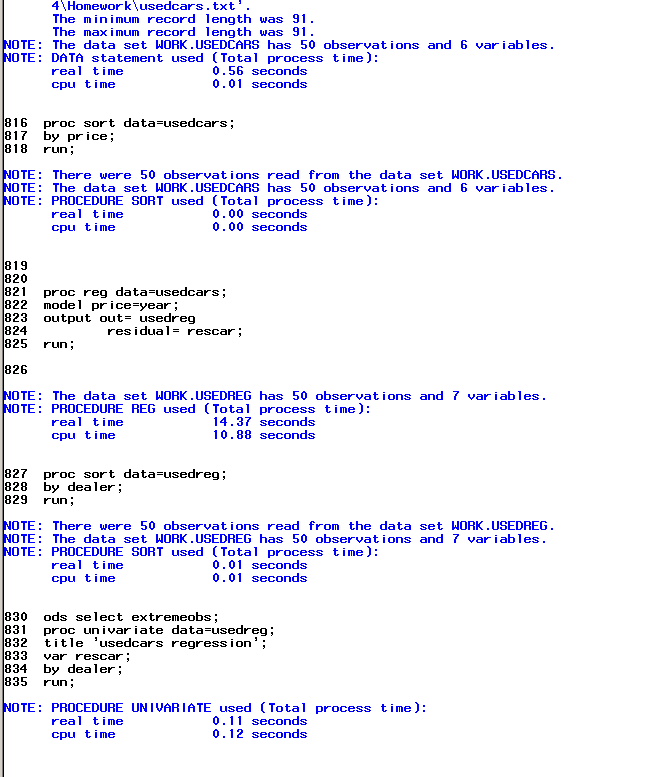
title 'usedcars regression';

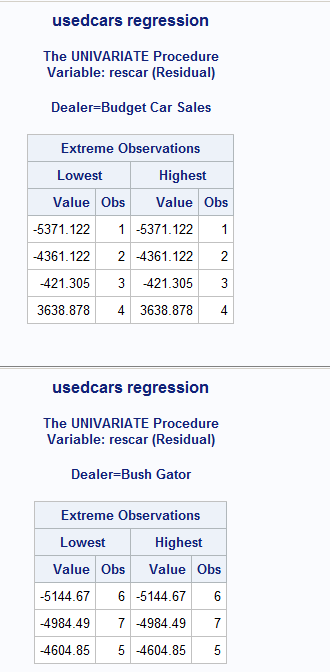
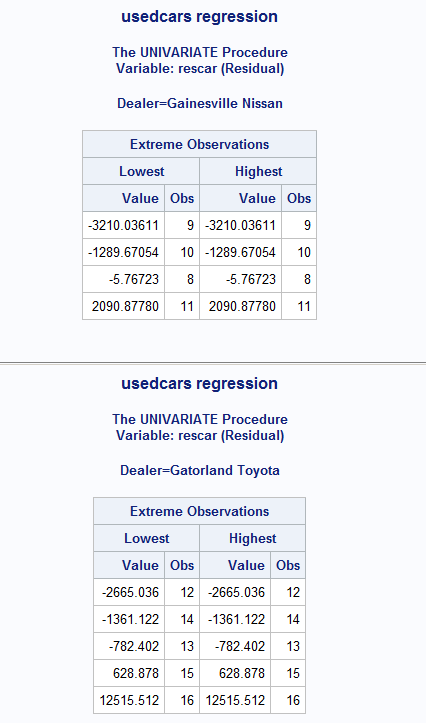
var rescar;

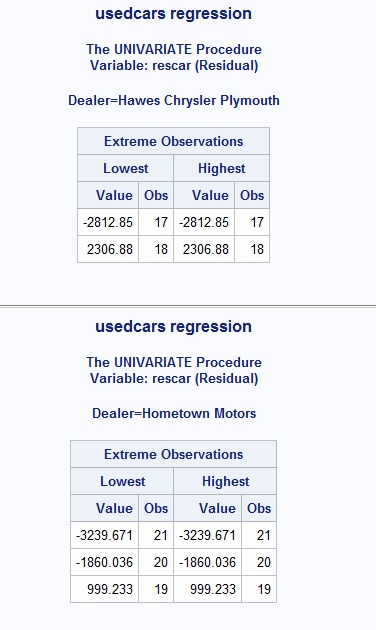
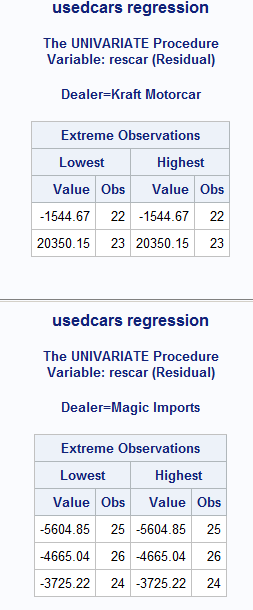
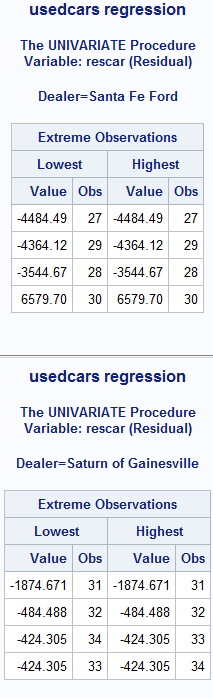
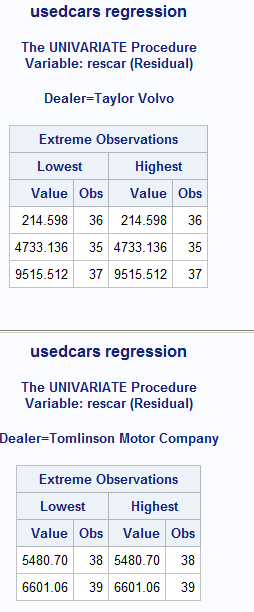
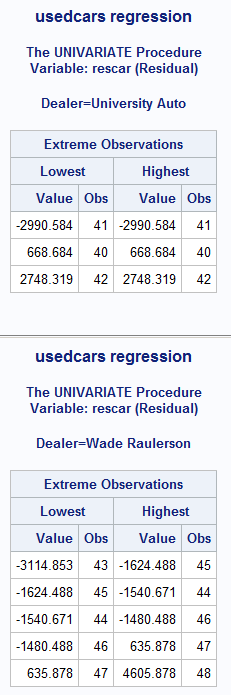
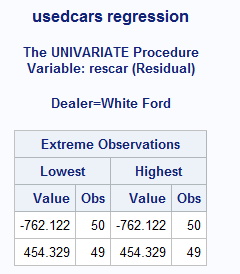
by dealer;

**run**;

LOG

  
OUTPUT

1. Refer to the HANKS data. During Tom Hanks’s career, he has played both humorous and dramatic roles. Over time, has he increasingly accepted serious roles over lighter ones? To see if this is true, find the correlations of the length of the movie, the drama rating, and the humor rating with year. (The ratings are ordinal, so the correlations of year with humor and drama must not be interpreted too rigidly. Instead, they give a rough indication of positive or negative trends with time.)

CODE

**data** hanks;

infile '\\Client\C$\Users\Leland\Desktop\ISM6930 SAS\Lesson 4\Homework\hanks.txt'

dlm='09'x missover firstobs=**2**;

input Title $ **1**-**25** Year **26**-**29** Length **34**-**36** MPAA $ **42**-**46** Action **50**-**51** Drama **58**-**59** Humor **66**-**67** Sex **74** Violence **82**-**83** Suspense **90** Offbeat **98** ;

**run**;

**proc** **format**;

value $rating 'G' = 'Suitable'

'PG' = 'Questionable'

'PG-13' = 'Questionable'

'R' = 'Not Suitable'

;

**run**;

**proc** **sort** data=hanks;

by year;

**run**;

**proc** **corr** data=hanks;

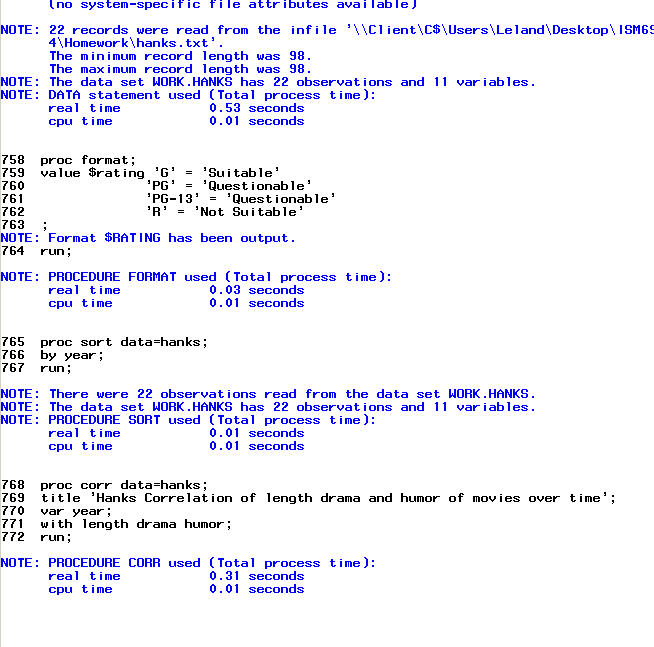
title 'Hanks Correlation of length drama and humor of movies over time';

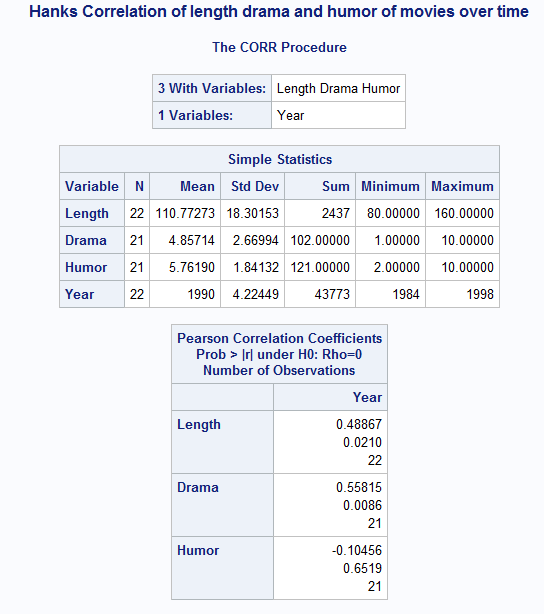
var year;

with length drama humor;

**run**;

LOG

  
OUTPUT



1. Refer to the LIMES data. Suppose that a citrus grower ships limes to be included in holiday gift baskets. To prevent bruising, the limes to be shipped are placed in plastic trays with individual compartments for each fruit. These compartments can be manufactured so that each lime is placed either on end or lengthwise in the compartment. The grower does not want the limes to roll around in the box, so the goal is to choose the tray orientation which would snugly hold more limes in place.

A statistician might solve this problem by determining whether the variance of the lime diameters is significantly larger or smaller than the variance of the lime lengths. Then, the orientation with lower variance would be the better orientation for packing. To test this hypothesis, we can use Pitman’s test. Calculate the sum of the fruit diameter and the fruit length, calculate the fruit diameter and the fruit length, calculate the fruit diameter minus the fruit length, and find the correlation between that sum and the difference. The implied results are shown below:

* Significant positive correlation implies that Var(diameter)>Var(length).
* Nonsignificant correlation implies that the two variances are not significantly different.
* Significant negative correlation implies that Var(diameter)<Var(length).

Based on the p-value for the correlation, what would you conclude? Write down your answer.

Based on the p-value of .0007 it is significant that the negative correlation results in diameter being less than length.

CODE

**data** limes;

infile '\\Client\C$\Users\Leland\Desktop\ISM6930 SAS\Lesson 3\Homework\limes.txt'

dlm=',' firstobs=**2** dsd;

input Date :MMDDYY10. FruitDia FruitLength FruitWt FruitVol JuiceVol JuiceWt PeelWt ;

format date MMDDYY8.;

label Fruitdia = "Fruit Diameter"

FruitLength= "Fruit Length"

FruitWt= "Fruit Weight"

FruitVol= "Fruit Volume"

JuiceVol= "Juice Volume"

JuiceWt= "Juice Weight"

PeelWt= "Peel Weight";

**run**;

**proc** **sort** data= newlimes;

by descending diameter;

**run**;

**data** limecalc;

set limes;

Sum= fruitdia + fruitlength;

Diff= fruitdia - fruitlength;

**run**;

**proc** **corr** data=limecalc;

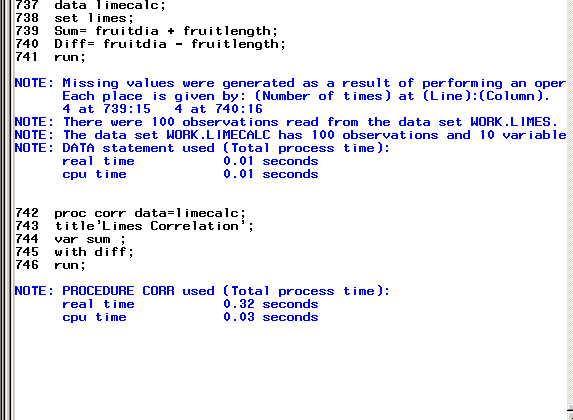
title'Limes Correlation';

var sum ;

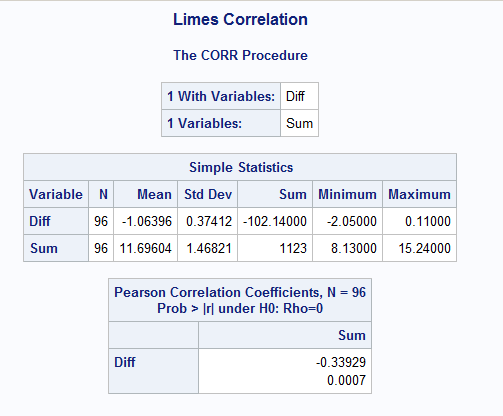
with diff;

**run**;

LOG



OUTPUT



1. Refer to the CLINTON data. Using all of the polls taken since 1993, calculate the number of days elapsed between successive polls and the number points that the approval rating changed in that time. For example, from January 24 to January 29, 1993, the elapsed time was 5 days and the change in approval rating was -4 points. Be sure to assign negative values to the “change” variable if the approval rating decreased and positive values if it increased. Then, find the correlation between the elapsed time in days and the change in approval rating. Also, calculate the correlation between the current approval rating and the number of days since the last poll was taken. Would you conclude that the Gallup organization has conducted its polls more frequently when President Clinton’s approval ratings have been unusually high, or low or when they have changed dramatically?

I read the data as they averaged polling every two weeks and their polling time reduced over the period while his approval dropped. I would read that as they are polling less as his approval falls.

CODE

**data** clinton;

infile '\\Client\C$\Users\Leland\Desktop\ISM6930 SAS\Lesson 5\Homework\Clinton.txt' firstobs=**3**;

input Date date18. Approval Disapproval No\_Opinion;

format date mmddyy10.;

**run**;

**proc** **sort** data=clinton;

by date;

**run**;

**data** clinton98;

set clinton;

retain App Dat;

elapsed=date-dat;

if elapsed=**.** then elapsed=**0**;

if app=**.** then change=**.**;

change= approval-app;

if change=**.** then change=**0**;

if app=**.** then change=**.**;

dat=date;

app=approval;

**run**;

**data** clintonrate;

set clinton98;

drop app dat;

**run**;

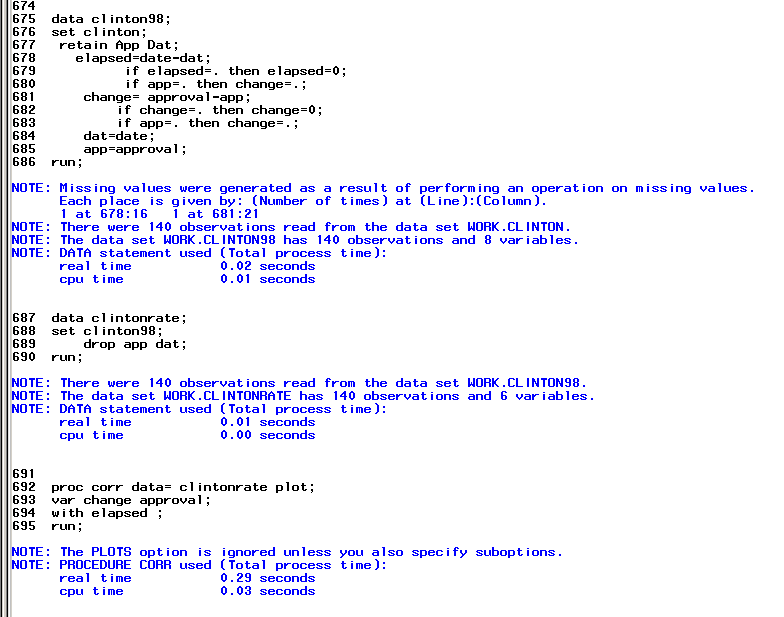
**proc** **corr** data= clintonrate plot;

var change approval;

with elapsed ;

**run**;

LOG



OUTPUT