CLAY HEATON – SAS Assessment 26 July 2013

/\* Question 1:

a: Null Hypothesis: The average production levels of the three bottling assembly

lines is the same.

Alternative Hypothesis: At least one of the bottling assembly lines has a

different level of production than the others.

\*/

/\* Question 1b: \*/

TITLE '1b. Box Plot of Units of Bottle Production by Line';

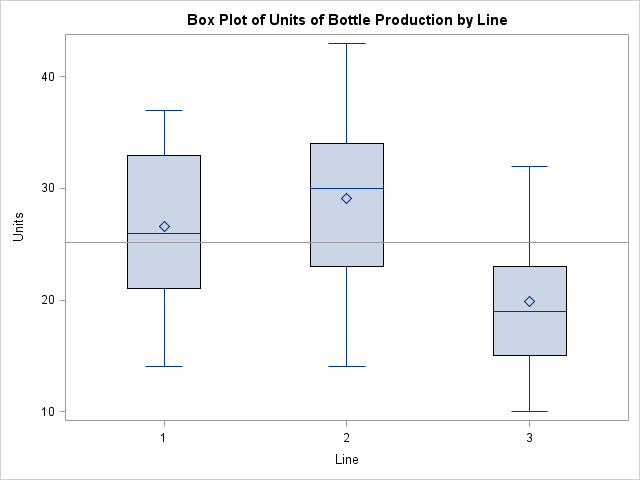
**PROC** **SGPLOT** DATA = sasuser.bottle;

VBOX units / CATEGORY = Line;

**RUN**;

TITLE;

The box plot below suggest that Line 3 has a mean bottle production that is significantly different from the average of all of the lines.



/\* Question 1c \*/

TITLE '1c. Overall Mean and Standard Deviation of Units of Production Across All Assembly Lines';

**PROC** **MEANS** DATA = sasuser.bottle MEAN STD;

VAR Units;

**RUN**;

TITLE;

/\* OUTPUT

Overall Mean and Standard Deviation of Units of Production Across All Assembly Lines 2

12:58 Friday, July 26, 2013

The MEANS Procedure

Analysis Variable : Units

Mean Std Dev

ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ

25.1884058 7.5230535

ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ

\*/

TITLE '1c. Mean and Standard Deviation of Units of Production for Each Individual Assembly Line';

**PROC** **MEANS** DATA = sasuser.bottle MEAN STD;

CLASS Line;

VAR Units;

**RUN**;

TITLE;

/\* OUTPUT

Mean and Standard Deviation of Units of Production for Each Individual Assembly Line 3

12:58 Friday, July 26, 2013

The MEANS Procedure

Analysis Variable : Units

N

Line Obs Mean Std Dev

ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ

1 23 26.5652174 6.1409659

2 23 29.0869565 7.2420053

3 23 19.9130435 6.1490062

ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ

\*/

/\* Question 1d \*/

**PROC** **GLM** DATA = sasuser.bottle PLOTS(ONLY) = DIAGNOSTICS;

CLASS Line;

MODEL Units = Line;

MEANS Line / HOVTEST;

**RUN**;

**QUIT**;

/\* Part 1: The predicted value vs. residuals plot appears more or less random, which

would indicate that there is equal variance of errors. The HOVTEST option

invokes Levene's Test for equality of variances. The results are:

The GLM Procedure

Levene's Test for Homogeneity of Units Variance

ANOVA of Squared Deviations from Group Means

Sum of Mean

Source DF Squares Square F Value Pr > F

Line 2 3025.8 1512.9 0.57 0.5678

Error 66 174914 2650.2

The null hypothesis is that the variances are equal for all of the Lines.

We fail to reject the null here, concluding that we do not have evidence

to show that there is variation in the errors.

Part 2: Normality of Errors: The histogram of residuals looks more or less normal,

perhaps with a slight skew to the right. Probably good enough to call normal.

The QQ plot also seems to indicate normality of errors.

============================================================================================

The GLM Procedure

Dependent Variable: Units

Sum of

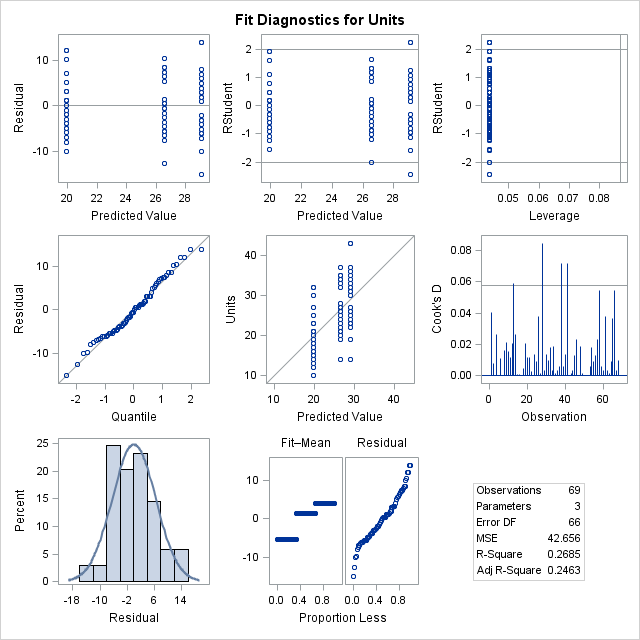
Source DF Squares Mean Square F Value Pr > F

Model 2 1033.246377 516.623188 12.11 <.0001

Error 66 2815.304348 42.656126

Corrected Total 68 3848.550725

============================================================================================



/\* Question 1e

Overall: The null hypothesis is that the mean production of units is the same across

all lines. The results (below) from SAS show a p-value of < 0.0001. We reject

our null hypothesis and conclude that there are differences in the means of

units across different lines in bottle production.

\*/

/\* Question 1f \*/

TITLE '1f. Line Comparisons for Bottle Production using Tukey';

**PROC** **GLM** DATA = sasuser.bottle PLOTS(ONLY) = DIFFPLOT;

CLASS Line;

MODEL Units = Line;

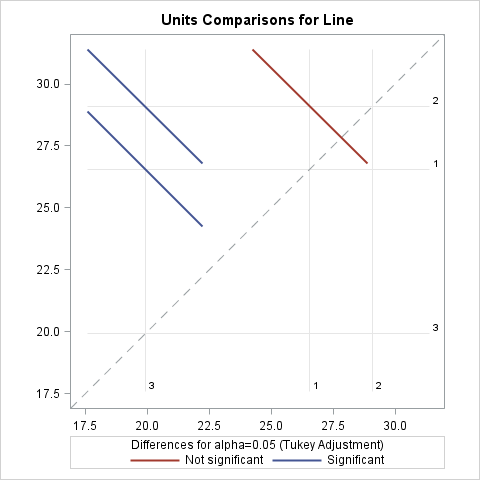
LSMEANS Line / PDIFF = ALL ADJUST = TUKEY;

**RUN**;

**QUIT**;

TITLE;

The diffogram shows that there is a difference between line 3 and line 1 and a difference between line 3 and line 2:



/\* Question 1g

Mr. Manager, the way that we collected our data may interfere with our analysis.

I propose that we begin to use what we call "blocking" to collect our data, because

it might remove nusiance factors from our analysis. Since we have 4 different suppliers,

I suggest that we create 3 blocks -- one for each Line in our plant -- and then randomly

process bottles from each of the 4 suppliers on those 3 lines. That will allow us to

determine whether the problem with line 3 is due, in fact, to the quality of the

products from our suppliers. \*/

/\*

============================================================================================

========================== QUESTION 2 ======================================================

============================================================================================

\*/

/\* Question 2a \*/

TITLE '2a. Correlation Analysis for Salary Data';

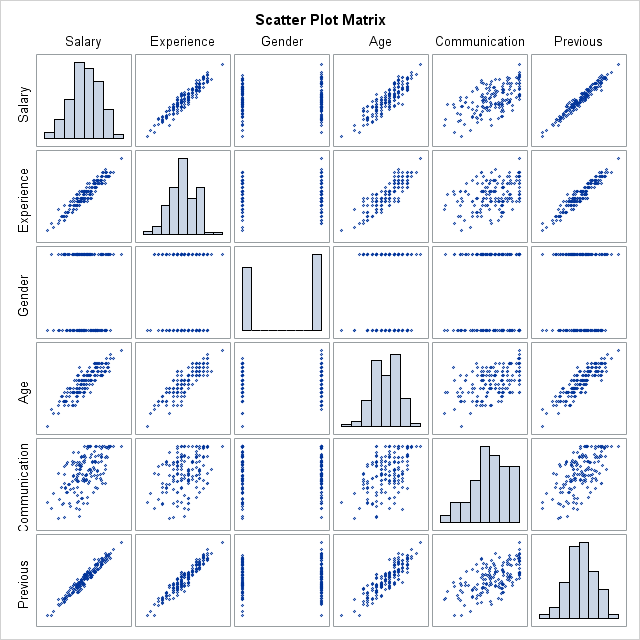
**PROC** **CORR** DATA = sasuser.salary PLOTS = MATRIX(NVAR = ALL HISTOGRAM);

VAR Salary Experience Gender Age Communication Previous;

**RUN**;

TITLE;

The correlation coefficient for salary and experience is 0.95. This indicates a very strong linear relationship between salary and experience.



Visually, it looks like there are many noteworthy linear associations here, everywhere there is a diagonal pattern going from the lower left to the upper-right. A quick visual glance would suggest that most of the variables are correlated.

/\* Question 2b \*/

/\*

When trying to predict salary using all other variables, I'm going to need to keep

potential multicollinearity in mind.

Experience is correlated with Age and Previous.

Communication appears to be correlated (more weakly) with Experience, Age, and Previous.

Based on the plot alone, it is difficult to tell if gender is correlated with the other

variables.

\*/

/\* Question 2c \*/

TITLE '2c. Predicting Salary with Experience, Gender, Age, Communication, and Previous';

**PROC** **REG** DATA = sasuser.salary;

All\_Variables:

MODEL Salary = Experience Gender Age Communication Previous;

OUTPUT OUT = out R = residuals;

**RUN**;

**QUIT**;

TITLE;

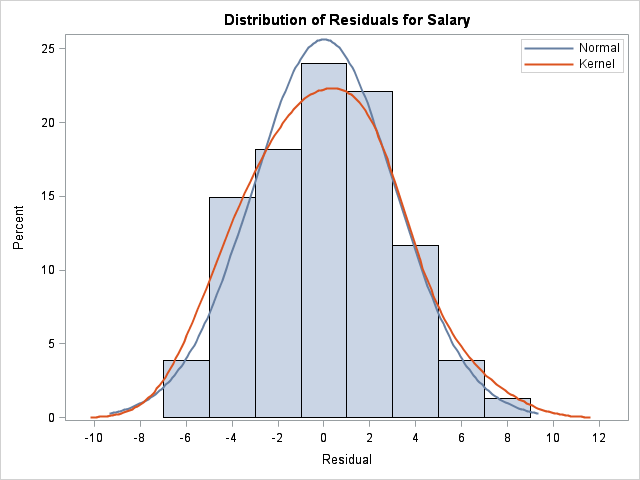
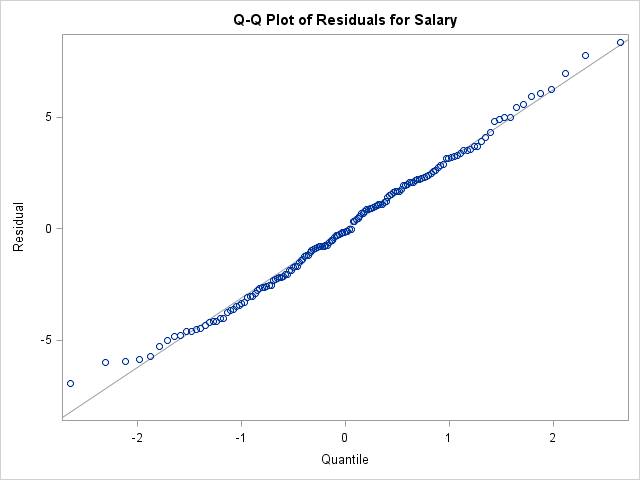
/\* Question 2c

Part 1: Normally Distributed Errors. The errors appear to be normally distributed, based

on the histogram in the lower-left of the fit-diagnostics matrix and the QQ plot

of the residuals. Let's do a formal analysis:

\*/

**PROC** **UNIVARIATE** DATA = out NORMAL;

VAR residuals;

**RUN**;

/\* The null hypothesis for the Anderson-Darling test is normality. We want to fail to

reject the null. Here are the results:

============================================================================================

Tests for Normality

Test --Statistic--- -----p Value------

Shapiro-Wilk W 0.992616 Pr < W 0.6154

Kolmogorov-Smirnov D 0.034985 Pr > D >0.1500

Cramer-von Mises W-Sq 0.031239 Pr > W-Sq >0.2500

Anderson-Darling A-Sq 0.224322 Pr > A-Sq >0.2500

============================================================================================

The p-value for Anderson-Darling is > 0.25, which means we fail to reject the null hypothesis

and conclude that we don't have the evidence to prove that the residuals are anything other

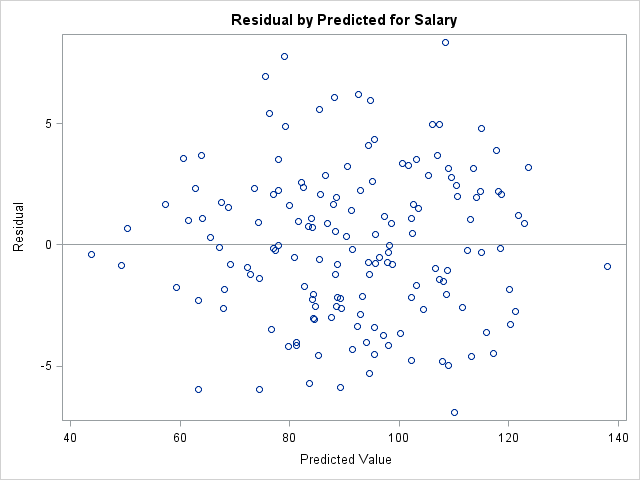
than normal. (aka, they are normal)

\*/

/\* Question 2c Part 2

The errors appear mostly to have equal variance, which perhaps a slight bit of a fan shape at

the lower predicted values. \*/



/\* Question 2c Part 3

The plot of predicted salary by value suggests that there is a linear relationship in the model

\*/



/\* Question 2d

Here are the parameter estimates and p-values for the variables in the model:

============================================================================================

Parameter Estimates

Parameter Standard

Variable DF Estimate Error t Value Pr > |t|

Intercept 1 11.95713 4.12437 2.90 0.0043

Experience 1 3.03172 0.34515 8.78 <.0001

Gender 1 -0.36363 0.51287 -0.71 0.4794

Age 1 0.24591 0.17123 1.44 0.1531

Communication 1 0.25062 0.03354 7.47 <.0001

Previous 1 0.21623 0.09174 2.36 0.0197

============================================================================================

Experience parameter est: 3.03, p-value: < 0.0001

Gender parameter est:-0.36, p-value: 0.4794

Age parameter est: 0.25, p-value: 0.1531

Commun. parameter est: 0.25, p-value: < 0.0001

Previous parameter est: 0.22, p-value: 0,0197

\*/

/\* Question 2e \*/

TITLE '2e. Predicting Salary with Experience, Gender, Age, Communication, and Previous';

**PROC** **REG** DATA = sasuser.salary PLOTS(ONLY) = DIAGNOSTICS (UNPACK);

All\_Variables:

MODEL Salary = Experience Gender Age Communication Previous / VIF;

OUTPUT OUT = out R = residuals;

**RUN**;

**QUIT**;

TITLE;

/\* RESULTS for VIF

============================================================================================

Parameter Estimates

Parameter Standard Variance

Variable DF Estimate Error t Value Pr > |t| Inflation

Intercept 1 11.95713 4.12437 2.90 0.0043 0

Experience 1 3.03172 0.34515 8.78 <.0001 25.64787

Gender 1 -0.36363 0.51287 -0.71 0.4794 1.00335

Age 1 0.24591 0.17123 1.44 0.1531 4.75831

Communication 1 0.25062 0.03354 7.47 <.0001 3.58657

Previous 1 0.21623 0.09174 2.36 0.0197 34.41884

============================================================================================

VIF is the Variance Inflation Factor. It tells us the severity of multicollinearity in an

ordinary least squares regression model. Any value above 10 is too high.

The values here indicate that there is a high degree of multicollinearity affecting both

Experience and Previous. Since we never remove more than 1 variable at a time from a

multiple linear regression model, the next step would be to remove Previous (highest VIF)

and run the model again.

\*/

/\* Question 2f \*/

TITLE '2f. Predicting Salary with Experience, Gender, Age, and Communication';

**PROC** **REG** DATA = sasuser.salary;

Better\_Model:

MODEL Salary = Experience Gender Age Communication

/ SELECTION = RSQUARE ADJRSQ;

**RUN**;

**QUIT**;

TITLE;

/\* RESULTS

============================================================================================

2f. Predicting Salary with Experience, Gender, Age, and Communication

The REG Procedure

Model: Better\_Model

Dependent Variable: Salary

R-Square Selection Method

Number of Observations Read 154

Number of Observations Used 154

Number in Adjusted

Model R-Square R-Square Variables in Model

1 0.9061 0.9055 Experience

1 0.7757 0.7742 Age

1 0.3504 0.3461 Communication

1 0.0016 -.0049 Gender

-------------------------------------------------------------------------

2 0.9669 0.9665 Experience Communication

2 0.9183 0.9172 Experience Age

2 0.9063 0.9050 Experience Gender

2 0.8105 0.8079 Age Communication

2 0.7757 0.7727 Gender Age

2 0.3516 0.3430 Gender Communication

-------------------------------------------------------------------------

3 0.9677 0.9671 Experience Age Communication

3 0.9671 0.9664 Experience Gender Communication

3 0.9183 0.9167 Experience Gender Age

3 0.8105 0.8067 Gender Age Communication

-------------------------------------------------------------------------

4 0.9679 0.9670 Experience Gender Age Communication

============================================================================================

The best models are:

Number in Adjusted

Model R-Square R-Square Variables in Model

1 0.9061 0.9055 Experience

2 0.9669 0.9665 Experience Communication

3 0.9677 0.9671 Experience Age Communication

4 0.9679 0.9670 Experience Gender Age Communication

The R-square value always increases with the addition of variables to a model. That is why I

included the Adj-R-square value here, too.

For all-regression approaches, it is possible also to use Mallow's Cp as a deciding factor

for choosing a best model. Otherwise, one can use automated forward selection, backward

elimination, or stepwise selection. Based on the work done here, I would select the

aforemented model with 3 variables as the best. \*/

/\*

============================================================================================

========================== QUESTION 3 ======================================================

============================================================================================

\*/

/\* Question 3a \*/

TITLE '3a. PROC FREQ to look at Frequency and Crosstabulation Tables';

**PROC** **FREQ** DATA = sasuser.pain;

TABLES Pain Treatment Pain\*Treatment;

**RUN**;

TITLE;

/\* RESULTS

============================================================================================

3a. PROC FREQ to look at Frequency and Crosstabulation Tables 28

12:58 Friday, July 26, 2013

The FREQ Procedure

Cumulative Cumulative

Pain Frequency Percent Frequency Percent

ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ

No 33 55.00 33 55.00

Yes 27 45.00 60 100.00

Cumulative Cumulative

Treatment Frequency Percent Frequency Percent

ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ

A 20 33.33 20 33.33

B 20 33.33 40 66.67

P 20 33.33 60 100.00

Table of Pain by Treatment

Pain Treatment

Frequency‚

Percent ‚

Row Pct ‚

Col Pct ‚A ‚B ‚P ‚ Total

ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ

No ‚ 17 ‚ 13 ‚ 3 ‚ 33

‚ 28.33 ‚ 21.67 ‚ 5.00 ‚ 55.00

‚ 51.52 ‚ 39.39 ‚ 9.09 ‚

‚ 85.00 ‚ 65.00 ‚ 15.00 ‚

ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ

Yes ‚ 3 ‚ 7 ‚ 17 ‚ 27

‚ 5.00 ‚ 11.67 ‚ 28.33 ‚ 45.00

‚ 11.11 ‚ 25.93 ‚ 62.96 ‚

‚ 15.00 ‚ 35.00 ‚ 85.00 ‚

ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ

Total 20 20 20 60

33.33 33.33 33.33 100.00

============================================================================================

Based on the crosstabulation table, it appears that there is a relationship between

treatment and an outcome of pain. In particular, at quick glance, I would say that

patients are least likely to experience pain when on treatment A and most likely

to experience pain while using a placebo. \*/

/\* Question 3b \*/

TITLE '3b. Crosstabulation of Pain and Treatment with Expected Cell Counts and Other Measures';

**PROC** **FREQ** DATA = sasuser.pain;

TABLES Pain\*Treatment / CHISQ EXPECTED CELLCHI2 RELRISK;

**RUN**;

TITLE;

/\* RESULTS

============================================================================================

3b. Crosstabulation of Pain and Treatment with Expected Cell Counts and Other Measures 31

12:58 Friday, July 26, 2013

The FREQ Procedure

Table of Pain by Treatment

Pain Treatment

Frequency ‚

Expected ‚

Cell Chi-Square‚

Percent ‚

Row Pct ‚

Col Pct ‚A ‚B ‚P ‚ Total

ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ

No ‚ 17 ‚ 13 ‚ 3 ‚ 33

‚ 11 ‚ 11 ‚ 11 ‚

‚ 3.2727 ‚ 0.3636 ‚ 5.8182 ‚

‚ 28.33 ‚ 21.67 ‚ 5.00 ‚ 55.00

‚ 51.52 ‚ 39.39 ‚ 9.09 ‚

‚ 85.00 ‚ 65.00 ‚ 15.00 ‚

ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ

Yes ‚ 3 ‚ 7 ‚ 17 ‚ 27

‚ 9 ‚ 9 ‚ 9 ‚

‚ 4 ‚ 0.4444 ‚ 7.1111 ‚

‚ 5.00 ‚ 11.67 ‚ 28.33 ‚ 45.00

‚ 11.11 ‚ 25.93 ‚ 62.96 ‚

‚ 15.00 ‚ 35.00 ‚ 85.00 ‚

ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ

Total 20 20 20 60

33.33 33.33 33.33 100.00

Statistics for Table of Pain by Treatment

Statistic DF Value Prob

ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ

Chi-Square 2 21.0101 <.0001

Likelihood Ratio Chi-Square 2 22.8621 <.0001

Mantel-Haenszel Chi-Square 1 19.4680 <.0001

Phi Coefficient 0.5918

Contingency Coefficient 0.5093

Cramer's V 0.5918

Sample Size = 60

============================================================================================

The expected cell counts all are above 5, so we do not need to use Fisher's Exact Test.

Assuming we are trying to predict pain outcomes, we should use the Pearson Chi-Square test

because the predictor variable, Treatment, is not ordinal.

The Chi-Square value is <0.0001, indicating that there is an association between Treatment

and Pain outcome. \*/

/\* Problem 3c \*/

TITLE '3c. Logistic Model Pain = Treatment Age';

**PROC** **LOGISTIC** DATA = sasuser.pain PLOTS(ONLY) = (EFFECT ODDSRATIO);

CLASS Treatment(REF = 'P') / PARAM = REF;

MODEL Pain(EVENT = 'No') = Treatment Age / CLODDS = PL;

**RUN**;

TITLE;

/\* PARTIAL RESULTS

============================================================================================

Analysis of Maximum Likelihood Estimates

Standard Wald

Parameter DF Estimate Error Chi-Square Pr > ChiSq

Intercept 1 4.1398 3.9507 1.0981 0.2947

Treatment A 1 3.6276 0.9331 15.1146 0.0001

Treatment B 1 2.5562 0.8279 9.5335 0.0020

Age 1 -0.0841 0.0567 2.1947 0.1385

============================================================================================

The parameter estimates and p-values are:

- Treatment A parameter est 3.63, p-value 0.0001

- Treatment B parameter est 2.56, p-value 0.0020

- Age parameter est-0.08, p-value 0.1385

\*/

/\* Problem 3d \*/

/\* The p-value for Age is 0.13, which means that it is not a statistically significant

variable in the model.

The p-value for Treatment A is 0.0001, which means that there is a statistically

significant difference between Treatment A and the Placebo (reference) group.

The p-value for Treatment B is 0.0020, which means that there is a statistically

significant difference between Treatment B and the Placebo (reference) group.

\*/

/\* Problem 3e \*/

/\* The `CLODDS = PL` option on the MODEL statement in the code snippet in part 3c

is what uses the profile likelihoods instead of the Wald estimates.

The odds ratio results are:

============================================================================================

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals

Effect Unit Estimate 95% Confidence Limits

Treatment A vs P 1.0000 37.622 7.151 296.642

Treatment B vs P 1.0000 12.886 2.870 78.582

Age 1.0000 0.919 0.815 1.022

============================================================================================

The odds ratio for age is 0.919, but the confidence interval includes the value of 1, which

means that it is insignificant in the model. \*/

/\* Problem 3f \*/

TITLE '3f. Logistic Model Pain = Treatment Age with effects coding';

**PROC** **LOGISTIC** DATA = sasuser.pain PLOTS(ONLY) = (EFFECT ODDSRATIO);

CLASS Treatment(REF = 'P');

MODEL Pain(EVENT = 'No') = Treatment Age / CLODDS = PL;

**RUN**;

TITLE;

/\* PARTIAL RESULTS

============================================================================================

Analysis of Maximum Likelihood Estimates

Standard Wald

Parameter DF Estimate Error Chi-Square Pr > ChiSq

Intercept 1 6.2011 4.0663 2.3256 0.1273

Treatment A 1 1.5663 0.5090 9.4690 0.0021

Treatment B 1 0.4949 0.4442 1.2412 0.2652

Age 1 -0.0841 0.0567 2.1947 0.1385

============================================================================================

Treatment A now has a p-value of 0.0021, which is significant. Since we are using effects

coding, we are testing the treatments against the overall average effect of all treatments,

including the Placebo group. This means that there is a statistically significant difference

between Treatment A and the overall average of all treatments.

Treatment B now has a p-value of 0.2652, which is not significant. This means that we do not

have the evidence to show that Treatment B is statistically different from the overall

average effect of all treatments. \*/