Name:

STAT 2112: Computer Assignment #2

1 Linear Model

1. For several years, I worked as a volunteer for a soup kitchen in Milwaukee where I supervised the Tuesday breakfast program. After opening the kitchen at 5:30 a.m., I would estimate, based on Monday's attendance, the number of meals which we would expect to serve that morning. Naturally, I used a statistical method namely the method of least squares to compute β_0 and β_1 of a linear model. In this model, the response variable, y, represented Tuesday's attendance, and the independent or explanatory variable, x, represented Monday's attendance. I used 15 weeks of data to compute estimates of the model's parameters:

Monday	255	295	355	350	240	295	305	345	255	295	305	345	265	240	300
Tuesday	225	285	295	280	225	240	295	295	245	250	280	290	205	205	285

$\sum_{i=1}^{15} x_i = 4445$ $\sum_{i=1}^{15} x_i^2 = 1339175$ $\sum_{i=1}^{15} x_i y_i = 1171415$ $\sum_{i=1}^{15} y_i = 3900$. After the line was fitted to the data, the sum of squared errors was computed to be: SSE=4596.

   (a) (8 pts) Write the linear model:

   (b) (8 pts) Make picture of the data.

   Suddenly, due to an emergency, you became the Tuesday breakfast co-ordinator. Monday's attendance was 280. Estimate Tuesday's expected attendance. To that end, compute:

   (c) (8 pts) Calculate by hand, SS_{xx}=

   (d) (8 pts) Calculate by hand, SS_{xy}=

   (e) (8 pts) Calculate by hand, $\widehat{\beta_0}$, and write the value which SPSS produced.

   (f) (8 pts) Calculate by hand, $\widehat{\beta_1}$, and write the value which SPSS produced.

   (g) (8 pts) When $x_p = 280$, calculate $\widehat{E[y_p]}$=

   (h) (8 pts) What are the 95% confidence intervals for β_0 and β_1 which SPSS produced? (Remember to check the *Confidence Interval* box in the *Statistics* sub-menu).

   Almost all of our patrons were destitute, and although we tried to prepare enough food, sometimes to my great embarrassment we did run out of hot meals, but that

happened rarely because I used the upper limit of the 95% confidence interval of $E[\widehat{E[y]}]$. Do the same. Find:

(i) (8 pts) $t_{n-2,\frac{\alpha}{2}}=$

(j) (8 pts) Given that $s^2 = \frac{SSE}{n-2}$, calculate s. (Note that SPSS produces s^2 in the ANOVA Table under the column *Mean Square* on row *Residual*).

(k) (8 pts) Given that $x_p = 280$, find the upper limit of the 95% CI about $E[\widehat{E[y]}]$. Use the appropriate formula found on page 642 of the McClave textbook.

(l) (8 pts) Produce a Q-Q plot.

(m) (8 pts) Plot residuals versus predicted values.

(n) (8 pts) Produce an ANOVA table. Test the hypothesis using the F test statistic

(o) (8 pts) Is the model a good model? Explain your answer.

2 Multiple Means

2. Testing hypotheses and confidence intervals are equivalent. The question of whether or not the difference between the two population means equals a specified value, d_0 as in:

$$H_0 : \mu_1 - \mu_2 = d_0 \quad vs \quad H_1 : \mu_1 - \mu_2 \neq d_0$$

at a level of significance α can be answered by observing if the confidence intervals overlap. The same conclusion can be obtained by comparing the test statistic to the appropriate quantile.

The equivalence would occur in higher dimensions when more than two means are compared as in:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \quad vs \quad H_1 : otherwise \tag{1}$$

if it were possible to draw four or higher dimensional confidence regions. It is, of course, impossible to draw pictures of things in four or higher dimensions. Nonetheless, the method of least squares makes it possible to test hypothesis 1. In this computer assignment, the F test statistic which SPSS will calculate and present in an ANOVA table will offer the opportunity to decide whether the null hypothesis that the means are the same can be rejected in favor of the alternative hypothesis which says that at least one mean is different from the rest.

Table 1:

Laboratory ID	1	2	3	4
	0.01434	0.00290	0.10417	0.21726
	0.10511	0.00739	0.42503	0.36502
	0.27031	0.35320	0.00713	0.42656
	0.45811	0.00317	0.00189	0.00259
	0.45734	0.00543	0.00515	0.00397
	0.45710	0.00065	0.00486	0.42211
	0.45698	0.00776	0.00636	0.00490
	0.45576	0.01149		0.41975
	0.45466			0.00190
	0.44371			0.00188
	0.45342			
	0.44324			

All motorcycle helmets which are sold in the United States must meet USDOT safety standards. Before a lot of helmets is shipped from a manufacturer to the stores, a random sample of helmets is selected and they are subjected to a destructive test. The helmet is clamped to a device which slides vertically downward by gravity when it is released so that the helmet strikes a flat steel anvil. Inside the helmet, there is a form which imitates the human head and inside the head is a device which measures the g force when the helmet strikes the anvil. The deformation of the helmet upon impact is measured. Suppose four independent laboratories which perform these tests are given helmets from the same lot and same manufacturer, in order to ascertain the quality of the laboratories. Theoretically, under these circumstances, the four laboratories should show no significant difference. The measurements from four laboratories are given in Table 1. Test hypothesis 1 at $\alpha = .05$.

To that end, enter measurements into SPSS such that you have two columns: the first column will contain the laboratory ID and the second column will contain the corresponding measurements as illustrated by Table 2.

Choose Analyze → Compare Means → One-Way ANOVA. Put **laboratory** into box **Factor** and put **measurement** into box **Dependent**. Under options, select *Means Plot*.

Inspect the ANOVA.

(a) (8 pts) What is the F test statistic?

(b) (8 pts) The F quantile is: $F_{3,33;.05} = 2.891564$

(c) (8 pts) Test the hypothesis.

(d) (8 pts) Shown in the SPSS ANOVA is **Sig.**. It is the p-value. What is it?

(e) (8 pts) Use this SPSS p-value to test the hypothesis.

Table 2:

laboratory	measurement
1	0.01434
1	0.10511
1	0.27031
.	.
.	.
.	.
2	0.00290
2	0.00739
2	0.35320
.	.
.	.
.	.

(f) (8 pts) Is the quality the same for these four laboratories?

3 Contingency Table

3. The results of a survey made to determine whether the age of a driver 21 years of age and older has any effect on the number of automobile accidents in which he is involved (including all minor accidents) are shown in Table 3. At a significance level of .05, test the hypothesis that the number of accidents is independent of the age of the driver.

Table 3: Contingency Table

		Age of Driver				
		20-30	31-40	41-50	51-60	61-70
Number of Accidents	0	748	821	786	720	672
	1	74	60	51	66	50
	2	31	25	22	16	15
	> 2	9	10	6	5	7

Find the following:

(a) (8 pts) X^2=

(b) (8 pts) Degrees of freedom=

(c) (8 pts) The appropriate X^2 quantile=

(d) (8 pts) Give your decision with justification.

Three steps are required to use SPSS in this problem:

(a) Enter the data into SPSS as follows:

Table 4:

Accident	Age	Freq
0	20-30	748
0	31-40	821
0	41-50	786
0	51-60	720
0	61-70	672
1	20-30	74
1	31-40	60
1	41-50	51
1	51-60	66
1	61-70	50
2	20-30	31
2	31-40	25
2	41-50	22
2	51-60	16
2	61-70	15
> 2	20-30	9
> 2	31-40	10
> 2	41-50	6
> 2	51-60	5
> 2	61-70	7

Go to the Variable tab of the spreadsheet; change var00001 to Accident and change `numeric` to `string` Likewise, use `string` instead of `numeric` for Age. When the cells are in `string` mode, letters can be entered into them or what looks like a number is treated as letter. For example, `20-30` is not subtraction but the name `20-30`.

(b) Choose Data → Weight Cases → Freq

(c) Choose Analyze → Descriptive → Crosstabs . Put **Accident** into box **Rows** and put **Age** into box **Columns**. Under Statistics, select *Chi-square*.

Instructions:

You may collaborate with a classmate or a friend to help you understand how to use SPSS and to do the computations. I want a report which contains the SPSS output and an answer sheet which contains the answers to the problems. The plots remain with the SPSS output. However, I want the report written in your own words showing your work for the various manual computations.