STAT 350 Post Exam 2 Material Review Problems Key

1. **True of False Questions (explain your answer):**

a) ANOVA stands for analysis of variance; therefore, it is a statistical inference procedure to test if the population variances are different.

FALSE
ANOVA is a statistical inference procedure that tests for any differences among the population by decomposing and comparing different sources of variation.

b) One-way ANOVA can be used only when there is a single level of a factor, we use k-way ANVOA when there are k levels of a factor.

FALSE
One-way ANOVA does not indicate the number of levels of a factor; it simply implies that there is one factor under consideration. The term 'one-way' refers to the number of factors being analyzed, not the number of levels within that factor. In contrast, two-way ANOVA is a statistical inference procedure used to study the effects of two factors on a single quantitative variable. So, the distinction lies in the number of factors (one vs. two) and not in the number of levels within a factor.

c) One-way ANOVA can be used only when there are two means to be compared.

FALSE
One-way ANOVA can be used when there are **three or more means** with one factor. Though it can be used when there are two means, a 2-sample independent t method is preferred.

d) In ANOVA, the null hypothesis is that the sample means are all equal.

FALSE
ANOVA tests the null hypothesis that the **population** means are all equal.

e) In rejecting the null hypothesis in ANOVA, one can conclude that all the means are different from one another.

FALSE
In rejecting the null hypothesis, one can conclude that **at least two means are different from each other**.

2. For each of the following situations, identify the response variable, the populations to be compared, and give *k*, *n* and (i) Degrees of freedom for group, for error, and for the total (ii) Null and alternative hypotheses (iii) Numerator and denominator degrees of freedom for the *F* statistic.

a) A poultry farmer is interested in reducing the cholesterol level in his marketable eggs. He wants to compare two different cholesterol-lowering drugs added to the hens' standard diet as well as an all-vegetarian diet. He assigns 25 of his hens to each of the three treatments.

Response: egg cholesterol level
Population: chickens with different diets or drugs
k = 3, n = 75, n_1=n_2=n_3=25

i)
dfa = k – 1 = 3 – 1 = 2
dfe = n – I = 75 – 3 = 72

dft -= n – 1 = dfa + dfe = 74

ii)
H_0: $\mu_1 = \mu_2 = \mu_3$,
H_a: at least two population means are different from each other.

iii)
df1 (numerator) = dfa = 2
df2 (denominator) = dfe = 72

b) A researcher is interested in students' opinions regarding an additional annual fee to support non-income-producing varsity sports. Students were asked to rate their acceptance of this fee on a seven-point scale. She received 94 responses, of which 31 were from students who attend varsity football or basketball games only, 18 were from students who also attend other varsity competitions, and 45 were from students who did not attend any varsity games

Response: rating on 7 points scale
Population: students from three different groups
k = 3, n = 94, n_1=31, n_2=18, n_3=45

i)
dfa = k – 1 = 3 – 1 = 2
dfe = n – I = 94 – 3 = 91
dft -= n – 1 = dfa + dfe = 93

ii)
H_0: $\mu_1 = \mu_2 = \mu_3$,
H_a: at least two population means are different from each other.

iii)
df1 (numerator) = dfa = 2
df2 (denominator) = dfe = 91

c) A professor wants to evaluate the effectiveness of her teaching assistants. In one class period, the 42 students were randomly divided into three equal-sized groups, and each group was taught power calculations from one of the assistants. At the beginning of the next class, each student took a quiz on power calculations, and these scores were compared.

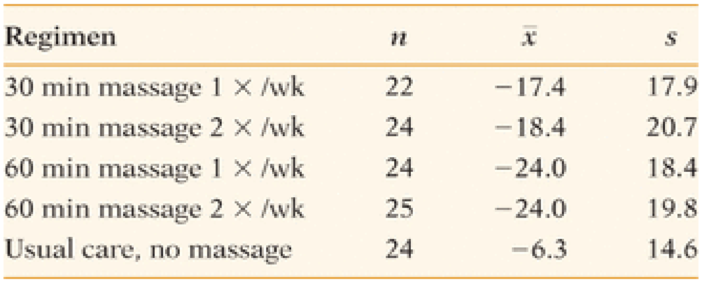
Response: quiz score
Population: students in each TA group
k = 3, n = 42, n_1=n_2=n_3=14

i)
dfa = k – 1 = 3 – 1 = 2
dfe = n – I = 42 – 3 = 39
dft -= n – 1 = dfa + dfe = 41

ii)
H_0: $\mu_1 = \mu_2 = \mu_3$,
H_a: at least two population means are different from each other.

iii)
df1 (numerator) = dfa = 2
df2 (denominator) = dfe = 41

3. Various studies have shown the benefits of massage to manage pain. In one study, 125 adults suffering from osteoarthritis of the knees were randomly assigned to one of five 8-week regimens. The primary outcome was the change in the Western Ontario and McMaster Universities Arthritis Index (WOMAC-Global). This index is used extensively to assess pain and functioning in those suffering from arthritis. Negative values indicate improvement. The following table summarizes the results of those completing the study.



Regimen	n	$\bar{x}$	s
30 min massage 1 × /wk	22	−17.4	17.9
30 min massage 2 × /wk	24	−18.4	20.7
60 min massage 1 × /wk	24	−24.0	18.4
60 min massage 2 × /wk	25	−24.0	19.8
Usual care, no massage	24	−6.3	14.6

Since you do not have the data to run, I have provided you with some of the R output below. (write down the code that was used to generate the output). Be sure to also include code and value of the critical value in the Tukey multiple comparison and any other value explicitly asked for in the questions.

```
            Df  Sum Sq   Mean Sq F value   Pr(>F)
Massage      4  5060.3   1265.09   3.728 0.00688
Residuals  114 38682.5    399.32

  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = pain ~ Massage)

$Massage
           diff      lwr      upr
30m1-30m2     1  -13.733   15.773
30m1-60m1   6.6   -8.173   21.373
30m1-60m2   6.6   -8.173   21.373
30m1-noc  -11.1  -25.873    3.673
30m2-60m1   5.6   -9.173   20.373
30m2-60m2   5.6   -9.173   20.373
30m2-noc  -12.1  -26.873    2.673
60m1-60m2     0  -14.773   14.733
60m1-noc  -17.7  -32.473   -2.927
60m2-noc  -17.7  -32.473   -2.927
```

a) What proportion of adults dropped out of the study before completion?

n = 22+24+24+25+24 = 119; dropout rate = (125 – 119)/125 = 4.8%

b) Is it reasonable to use the assumption of equal standard deviations when we analyze these data? Please explain your answer.

Yes, the ratio of the maximum to minimum standard deviations is $\frac{20.7}{14.6}$ = 1.42 < 2

c) Perform a hypothesis test at a 5% significance level for these data. Be sure to include the degrees of freedom(s) for the test statistic. I would suggest that you create the ANOVA table before you start. What code was used to generate the data shown above?

code:

```
fit <- aov(pain ~ Massage, data = PainStudy)
summary(fit)
```

The ANOVA table is provided from the R code except for dft and SST. If you are doing the problem by hand given SSA and MSE, the work is provided below:

ANOVA Table:

Source	df	SS	MS	F
Factor A	k – 1 = 5 – 1 = 4	5060.346	1265.09	3.73
Error	n – k = 119 – 5 = 114	38,682.48	339.32	
Total	dfa + dfe = 118	43,742.83		

$$MSA = \frac{SSA}{dfa} = \frac{5060.346}{4} = 1265.0865$$

SSE = MSE (dfe) = 339.32 (114) = 38,682.48

SST = SSA + SSE = 5060.346 + 38,682.48 = 43,742.83

$$F = \frac{MSA}{MSE} = \frac{1265.09}{339.32} = 3.73$$

Step 1:

μ_1 is the population mean index for 30 min massage 1 times per week

μ_2 is the population mean index for 30 min massage 2 times per week

μ_3 is the population mean index for 60 min massage 1 times per week

μ_4 is the population mean index for 60 min massage 2 times per week

μ_5 is the population mean index for usual care with no massage

Step 2:

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

H_a: at least two population means are different from each other.

Step 3:

F = 3.73 with df1 = 4, df2 = 114 (the work is above)

code if data is not provided:

```
> pf(3.73,4,114,lower.tail = FALSE)
[1] 0.006853962
```

p = P(F > 3.73) = 0.0069

Step 4:

Reject H_0 because 0.0069 < 0.05

The data shows support (p = 0.0069) to the claim that the population mean change in WOMAC-Global scores is different for at least one of the different regimens.

d) In this study, there are ten pairs of means to compare. If you do this by hand, assume that the number of observations for each case is 24. Determine the critical value for the Tukey multiple-comparisons method at a 5% significance level. Which pairs of means are found to be significantly different? You may use either the hand calculations or the data provided from the output. Please include a graphical representation of your result. Write a short summary of your analysis including what you would recommend to reduce pain. What code was used to generate the output above?

code if data is not provided:

```
> Q <- qtukey(0.95,5,114)
> Q
[1] 3.920091
> crit <- Q/sqrt(2)
> crit
[1] 2.771923
> MSE <- 339.32
> ni = nj = 24
> SE <- sqrt(MSE*(1/ni+1/nj))
> SE
[1] 5.317581
> crit*SE
[1] 14.73992
```

code if data is provided:

```
TukeyHSD(fit, conf.level = 0.95)
```

Note that the information below was calculated use a Margin of error of 14.773 which is the value that was generated using the table. This will change the confidence intervals, but not whether it was significant or not.

If the difference between the two means is greater than 14.74, then they are greater than the margin of error so they are different. If the difference between the two means is less than 14.964, then they are within the margin of error so they are the same.

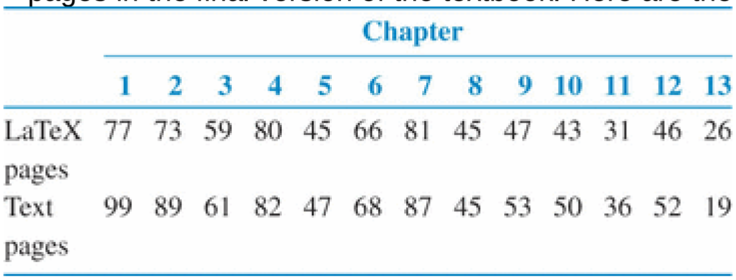
first mean	second mean	difference	Confidence Interval	significant
1	2	1	(-13.733,15.773)	no
1	3	6.6	(-8.173,21.373)	no
1	4	6.6	(-8.173, 21.373)	no
1	5	-11.1	(-25.873,3.673)	no
2	3	5.6	(-9.173,20.373)	no
2	4	5.6	(-9.173,20.373)	no
2	5	-12.1	(-26.873,2.673)	no
3	4	0	(-14.773,14.773)	no
3	5	-17.7	(-32.473,-2.927)	yes
4	5	-17.7	(-32.473,-2.927)	yes

Therefore, everything is the same except for 60 min massage 1 or 2 times a week is different from no massage. This can be graphical displayed as:

3	4	2	1	5
-24	-24	-18.4	-17.4	-6.3

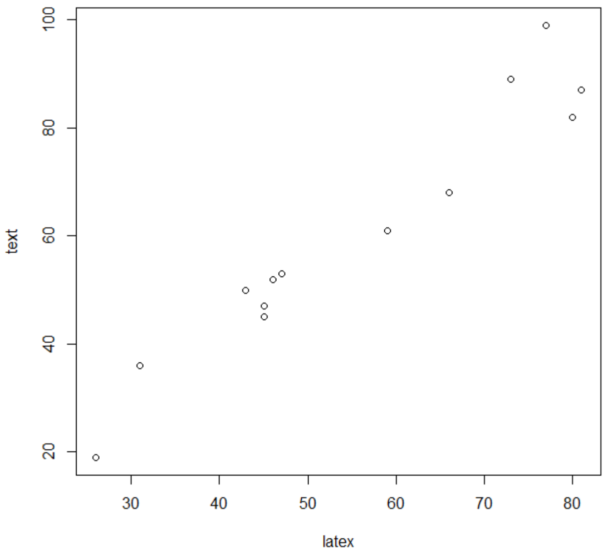
If I was going to improve pain management the most, I would have a 60 minute massage 1 time per week. There was no improvement for two times per week and 2 times per week is more expensive than 1 time per week. Note that both of the 30-minute massages were statistically the same as the two 60-minute massages. However, they are also the same as the usual care.

4. The editor of a statistics textbook would like to plan for the next edition. A key variable is the number of pages that will be in the final version. Text files are prepared by the authors using LaTeX, and separate files contain figures and tables. For the previous edition of the textbook, the number of pages in the LaTeX files can easily be determined, as well as the number of pages in the final version of the textbook. Here are the data:



	Chapter												
	1	2	3	4	5	6	7	8	9	10	11	12	13
LaTeX pages	77	73	59	80	45	66	81	45	47	43	31	46	26
Text pages	99	89	61	82	47	68	87	45	53	50	36	52	19

Scatter Plot



```
              Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  -6.20176     5.71233   -1.086    0.301
latex         1.20810     0.09828     xxx    8.95e-8
Multiple R-squared:  0.9321,    Adjusted R-squared:  0.926
```

a) What is the explanatory variable? Response variable?

Explanatory variable: LaTex Page
Response variable: Text Pages

b) Describe the form, direction and strength of the scatter plot above. Are there any outliers?

form: Linear, direction: positive, strength: strong?. I do not see any outliers.

c) Find the equation of the least-squares regression line based on the software output.

TextPages = - 6.20 + 1.21 LaTeXPages

d) Interpret the slope of the regression line.

This is the number of text pages that would be produced if you increased the number of LaTex pages by 1.

e) What is the meaning of the y-intercept? Does this variable make sense in this situation? Please explain your response.

The y-intercept is the number of pages in the text version when there are no pages in the Latex version. This does not make any sense because all chapters have to have some pages.

This does not mean that we don't have to include this value in the line; this just means that the number has no practical significance.

f) Find the predicted number of pages for the next edition if the number of LaTeX pages for a chapter is 62.

TextPages = - 6.20 + 1.21 (62) = 68.82 => 69
The number of pages has to be an integer so I rounded up.

g) What proportion of the variation in Text pages is explained by LaTex pages?

Multiple R-squared: 0.9321
The proportion of the variation in Text pages is explained by LaTex pages is 93.2%

Note: No words or work is required because this is from the output.

h) What is the value of the correlation coefficient r?

$r = +\sqrt{0.9321} = 0.965$
It is positive because the slope is positive.

i) Using the t test, test whether there is an association between LaTeX pages and text pages at a 10% significance level.

We are not given enough information to perform the F test.

Step 1: β_1 is the population slope between TextPages and LaTexPages

Step 2: H_0: $\beta_1 = 0$ H_a: $\beta_1 \neq 0$

Step 3:

$$t = \frac{b_1 - \beta_{10}}{SE_{b_1}} = \frac{1.21 - 0}{0.09828} = 12.31$$

df = 13 – 2 =11
p = 8.95e-8

Step 4:
reject H_0 because 8.95e-8 < 0.1
The data provides support (p = 8.95e-8) to the claim that there is a linear association between Text pages and LaTex pages.

j) Find and interpret a 90% confidence interval for the slope β_1

df = 13 – 2 = 11.

```
> qt(0.1/2,11,lower.tail = FALSE)
[1] 1.795885
```

$1.2081 \pm 1.7959 \cdot 0.09828 = 1.2081 \pm 0.1765 \Rightarrow (1.0316, 1.385)$

Interpretation: We are 90% confident that the population slope of pages of text versus pages of LaTex is covered by the interval from 1.0316 to 1.385.

k) Suppose there are 80 LaTex pages in a chapter for the new edition. Raj obtained a 95% confidence interval for the mean pages and a 95% prediction interval for the pages of the final version but did not label them. Which of the following two is the prediction interval? Please explain your answer.

Interval A	(75, 106)
Interval B	(84, 97)

Interval A is the prediction interval; Interval B is the confidence interval.
This is because Interval A is wider than interval B.

k) Which interval should the editor use to plan for the next edition of the statistics textbook?

Prediction Interval because we want to predict a value. The confidence interval is only for the average value.

l) One new chapter has 200 LaText pages. Discuss if it is appropriate use the regression line obtained in part (c) to predict the final textbook pages.

Not really. We need to be careful about extrapolation since the range of values is between 30 and 90.

5. Can a pretest in mathematics skills predict success in a statistics course? The 62 students in an introductory statistics class took a pretest at the beginning of the semester. The least-squares regression line for predicting the score, y, on the final exam from the pretest score, x, was y = 13.8 + 0.81x. The standard error of b_1 was 0.43.

a) Perform a hypothesis test to determine if there is a linear association between the pretest score and the score on the final exam at a 5% significance level.

This is a t-test because that is the only information that is provided.

Step 1: β_1 is the population slope between final exam score and pretest score.

Step 2: H_0: $\beta_1 = 0$ H_a: $\beta_1 \neq 0$

$$Step\ 3: t = \frac{b_1 - 0}{SE_{b_1}} = \frac{0.81}{0.43} = 1.88$$

df = 62 – 2 = 60

```
> 2*pt(0.81/0.43,60,lower.tail = FALSE)
[1] 0.06445124
```

p = 0.0644

Step 4: fail to reject H_0 because 0.0645 > 0.05
The data does not provide support (p = 0.0645) to the claim that there is a linear relationship between pretest score and score on the final exam.

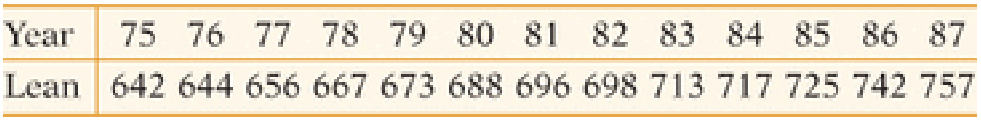
b) What would the decision be for the test to determine if the slope is positive at the same significance level? Explain your answer. Use the results of a) to answer this question. No hypothesis test is required.

For the one-sided alternative, we would divide the p-value by 2, so

$$p = \frac{0.0644}{2} = 0.0322$$

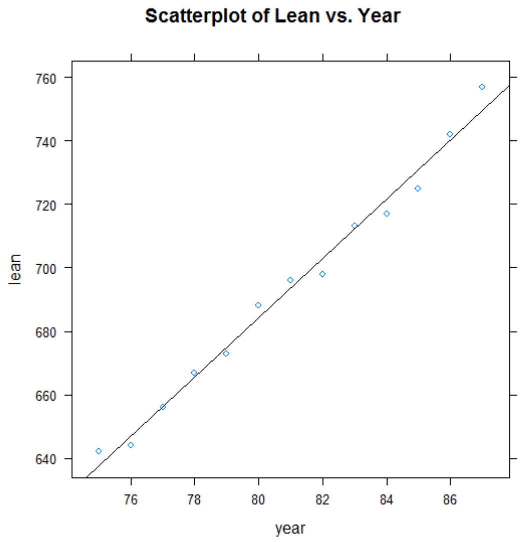
Therefore, we would reject H_0; however, I would consider this a 'maybe' situation.

6. The Leaning Tower of Pisa is an architectural wonder. Engineers concerned about the tower's stability have done extensive studies of its increasing tilt. Measurements of the lean of the tower over time provide much useful information. The following table gives measurements for the years 1975 to 1987. The variable "lean" represents the difference between where a point on the tower would be if the tower were straight and where it actually is. The data are coded as tenths of a millimeter in excess of 2.9 meters, so that the 1975 lean, which was 2.9642 meters, appears in the table as 642. Only the last two digits of the year were entered into the computer.

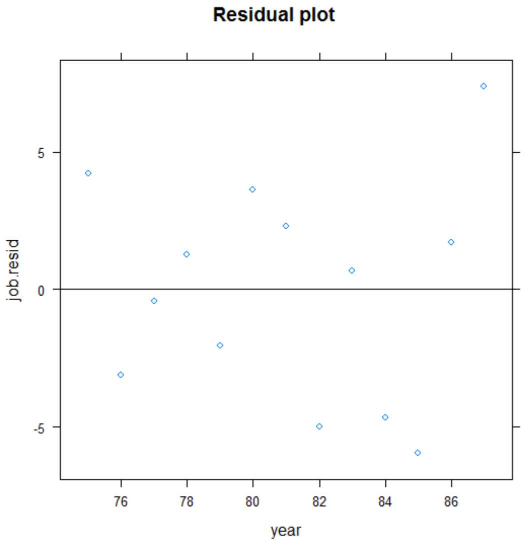


Year	75	76	77	78	79	80	81	82	83	84	85	86	87
Lean	642	644	656	667	673	688	696	698	713	717	725	742	757

Please use the following information in your analysis.

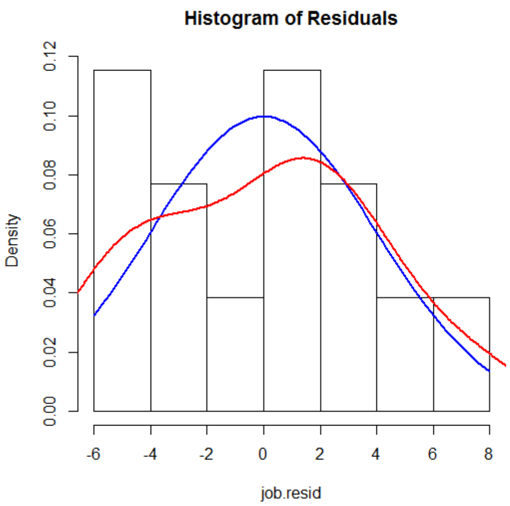


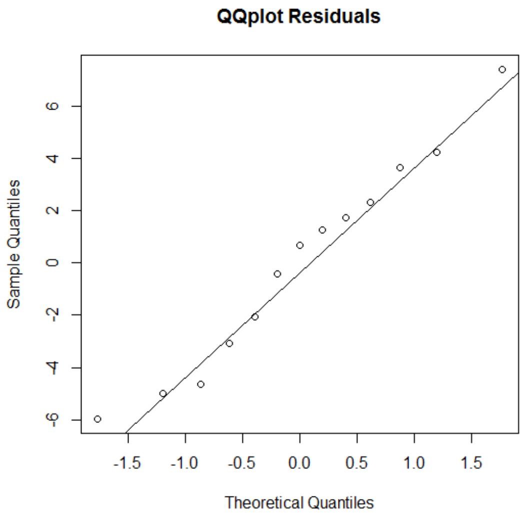
Scatterplot of Lean vs. Year



Residual plot

STAT 350 Post Exam 2 Material Review Problems Key





```
Coefficients:
                Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)       xxxx        xxx       xxxx     0.0333
year              xxxx        xxx       xxxx     6.5e-12
```

S_{XX} = 182, S_{YY} = 15996.77, S_{XY} = 1696, x̄ = 81, ȳ = 693.69

a) Does the trend in lean over time appear to be linear? Be sure to indicate which plot(s) you are using.

Yes. Both the scatterplot and the residual plot show that this is a linear relationship.

b) Are the assumptions met? Please explain your answer. Be sure to indicate all plot(s) you are using for each assumption.

Linear: yes (see part a)
Constant variance: This looks ok from the scatterplot and the residual plot.
Normality of the residuals: With only 13 points, we cannot use CLT to help us out. I would question this assumption. t
SRS/Independent: we have to assume this from the problem.

Please continue to do the problem no matter what you stated in parts a) and b).

c) What is the equation of the least-squares line? What is the correlation? What percent of the variation in lean is explained by this line?

$$b_1 = \frac{S_{XY}}{S_{XX}} = \frac{1696}{182} = 9.32$$

$b_0 = \bar{y} - b_1\bar{x}$ = 693.69 – (81)(9.32) = -61.12

Lean = -61.12 + 9.32 Year

$$r = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}} = \frac{1696}{\sqrt{(182)(15996)}} = 0.9940$$

In this case, we are not given the ANOVA table (that is asked in part d), so we have to say that $r^2 = (0.9940)^2 = 0.988$

d) Calculate the ANOVA table for linear regression. What percent of the variation in lean is explained by this line?

Source	df	SS	MS	F
Regression	1	15806.72	SSR = 15806.72	918.46
Error	n – 2 = 13 – 2 = 11	189.28	17.21	
Total	n – 1 = 13 - 1 = 12	S_{YY} = 15996.77		

$$SSR = b_1 S_{XY} = \frac{S_{XY}^2}{S_{XX}} = (9.32)(1696) = 15{,}806.72\ (or\ using\ the\ second\ formula\ 15804.48)$$

SSE = 15996 – 15806.72 = 189.28 (15996 – 15804.48 = 191.52)

$$MSE = \frac{SSE}{dfe} = \frac{189.28}{11} = 17.21\ \left(\frac{191.52}{11} = 17.41\right)$$

$$F = \frac{MSR}{MSE} = \frac{15806.72}{17.21} = 918.46\left(\frac{15806.72}{17.41} = 907.91\right)$$

$$R^2 = \frac{SSR}{SST} = \frac{15806.72}{15996.77} = 0.9881$$

Note that this value is the same as in part c).

e) Using the F test, test whether there is an association between the year and the amount of Lean at a 1% significance level

Step 1: none

Step 2: H_0: there is no linear association between year and amount of lean
H_a: there is a linear association between year and amount of lean

Step 3: F = 918.46 (907.91)
df1 = 1, df2 = 11

p = P(F > 918.46 or 907.91) = 6.0e-12 (6.4e-12)

```
pf(fts,1,11,lower.tail = FALSE)
```

Step 4: reject H_0 because 6e-12 < 0.01

The data provides strong support (p = 6e-12) to the claim that there is a linear association between year and amount of lean.

f) Calculate and interpret the 99% confidence interval for the average rate of change (tenths of a millimeter per year) of the lean.

Note that 'rate of change' means the slope.

df = 13 – 2 = 11

```
> qt(0.01/2,11,lower.tail = FALSE)
[1] 3.105807
```

$$b_1 \pm t_{0.005,11}\sqrt{\frac{MSE}{S_{XX}}} = 9.32 \pm 3.1058\sqrt{\frac{17.21}{182}} = 9.32 \pm 3.1058 \cdot 0.3075 = 9.32 \pm 0.955$$

$$\Longrightarrow (8.36, 10.28)$$

Interpretation: We are 99% confidence that the population slope of lean vs. year is covered by the interval from 8.36 to 10.28.

g) In 1918 the lean was 2.9071 meters. (The coded value for the year is 18, the coded value for the lean is 71.) Using the least-squares equation for the years 1975 to 1987, calculate a predicted value for the lean in 1918.

Lean$_{18}$ = -61.12 + 9.32(18) = 107 → a lean of 2.9107

h) Do you think that the predicted value in part g) is appropriate? Please explain your answer. Use numerical and graphical summaries to support your explanation.

No. 18 is beyond range of the data shown in the scatterplot. There is no reason to believe that the lean was linear outside of the range of values provided.

i) The engineers working on the Leaning Tower of Pisa were most interested in how much the tower would lean if no corrective action was taken. Use the least-squares equation to predict the tower's lean in the year 2013. (*Note:* The tower was renovated in 2001 to make sure it does not fall down.) The question implies that the engineers wanted to know what would happen if the linearity was valid in 2013.

Lean$_{2013}$ = -61.12 + 9.32(113) = 992 → a lean of 2.9992

j) Calculate and interpret the 99% confidence interval for the average lean in 2013. [In the exam, you will either be given the standard error of the mean at a point and the prediction interval or you will be given the output from both options.]

Though the SE is calculated here, this will not be required on the exam although you will need to know the formula for both standard errors.

$$lean_{2013} \pm t_{0.005,11}\sqrt{MSE\left(\frac{1}{n}+\frac{(x^*-\bar{x})^2}{S_{XX}}\right)} = 992 \pm 3.1058\sqrt{17.21\left(\frac{1}{13}+\frac{(113-81)^2}{182}\right)}$$
$$= 992 \pm 3.1058\cdot 9.9073 = 992 \pm 30.77 \Longrightarrow (961.23,1022.77)$$

interpretation: we are 99% confident that the average value of the lean in 2013 would have been covered by the interval from 2.996123 to 3.002277 meters if no correction was made.

k) Calculate and interpret the 99% prediction interval for the lean in 2013. [In the exam, you will either be given the standard error of the mean at a point and the prediction interval or you will be given the output from both options.]

Though the SE is calculated here, this will not be required on the exam although you will need to know the formula for both standard errors.

$$lean_{2013} \pm t_{0.005,11}\sqrt{MSE\left(1+\frac{1}{n}+\frac{(x^*-\bar{x})^2}{S_{XX}}\right)} = 992 \pm 3.1058\sqrt{17.21\left(1+\frac{1}{13}+\frac{(113-81)^2}{182}\right)}$$
$$= 992 \pm 3.1058\cdot 10.7408 = 992 \pm 33.36 \Longrightarrow (958.64,1025.36)$$

interpretation: we are 99% that the actual lean in 2013 would have been covered by the interval from 2.95861 to 3.002536 meters if no correction was made.

l) To give a margin of error for the lean in 2013, would you use a confidence interval for a mean response (j) or a prediction interval (k)? Explain your choice.

Prediction Interval since we are predicting for one specific year. We use the confidence interval if we were interested in an average value.