Review Problems for the Final

1. There are about 2,700 institutions of higher learning in the United States (including junior colleges and community colleges). In 1976, as part of a continuing study of higher education, the Carnegie Commission took a simple random sample of 256 of these institutions. The mean enrollment in the 256 sample schools was 3,500, with an standard deviation of 4,000. A histogram for the enrollments was plotted and did not follow the normal curve.

Say whether each of the following statements is *true* or *false*, and explain why.

(a) An approximate 95%-confidence interval for the average enrollment of all 2,700 institutions runs from 3,000 to 4,000.

(b) It is estimated that 95% of the institutions of higher learning in the United States enroll between $3,000$ and $4,000$ students.

(c) There is an approximately 95% chance that the average enrollment at all institutions of higher learning in the United States is between 3000 and 4000 students.

(d) The normal curve can't be used to figure confidence levels here at all, because the data doesn't follow the normal curve.

2. Match each of the statements below to the appropriate hypothesis test from the list that follows.

Statements:

(a) The average cost of a pack of cigarettes is higher in Canada than in the US.

(b) The probability that a biology major is male is higher than the probability that an economics major is a male.

(c) The proportions of red, green, blue, yellow, orange and brown M&Ms in a randomly selected 6oz bag matches the proportions claimed by the Mars Corporation.

(d) The more fertilizer used, the taller a pine tree grows in two years.

(e) Kids, their parents and their grandparents are equally likely to use Facebook.

(f) The mean cost of repairing a car after a crash is the same for compacts, midsize cars and full-size cars.

(g) The average amount of cash in a restaurant cash register is $1500.

(h) The probability that a randomly chosen college graduate majored in Biology is 32%.

(i) The weight of a *Quarter Pounder* is the same, on average, as the weight of a *Whopper*.

(j) The length of a bull elephants left tusk is the same as the length of a bull elephants right tusk.

Tests

i. One sample proportion test

ii. One sample t test

iii. Two sample proportion test

iv. Two sample t test

v. Matched pair test

vi. Regression slope test

vii. Goodness-of-fit test

viii. Test of homogeneity.

ix. Analysis of variance

3. Researchers surveyed a simple random sample of 459 drivers and collected data on seat belt usage and cigarette smoking, which yielded the data in the table below. Test the claim that seat belt usage is *independent* of smoking — use all of the usual steps in your hypothesis test.

	Don't smoke	Smoke
Wear seat belts	175	68
Don't wear seat belts	149	67

They were studying the theory that people who smoke more are less likely to use seat belts (because they are less concerned about health and safety). Is this theory supported by the sample data? Explain.

4. Investigators studying the relationship between cigarette smoking and blood pressure in adult men collected data from 625 U.S. men aged 20 - 40, and generated the following statistics:

$$\overline{X} = 24 \quad SD_X = 5.5$$
$$\overline{Y} = 135 \quad SD_Y = 9 \quad r = 0.7$$

where X = number of cigarettes per day, and Y = systolic blood pressure, measured in mmHG.

(a) Is there *significant* linear correlation between systolic blood pressure and number of cigarettes per day? Frame your answer in terms of the appropriate hypothesis test (include all of the usual steps).

(b) What is the predicted systolic blood pressure of a 28-year old man who smokes 30 cigarettes per day? Is this prediction reliable? *Show your work.*

(c) Can you use this data to predict the blood pressure of a 50-year old man who smokes 20 cigarettes per day? *Explain your answer.*

5. John Smith is running for office. One week before the election, his campaign manager hires a polling firm to survey likely voters. The firm surveyed a simple random sample of 2700 likely voters and found that 51% favor Smith. They also found that of the 1250 women in the survey, 54% favor Smith.

a. What percentage of the men in the survey favor Smith?

b. Compute 95% confidence intervals for the percentage of women who favor Smith, the percentage of men who favor Smith and the percentage of likely voters who favor Smith.

c. If you were running Smith's campaign, how would you spend your advertising money in the last few days before the election? Explain.

6. A simple random sample of 25 college students were each weighed at the beginning of their freshman year and then again at the end of their freshman year. The mean difference in the weights of these students was 12 lbs and the standard deviation of the differences was 8 lbs. Test the claim that students gain at least 15 lbs on average during their freshman year. Use a 0.05 significance level. You may assume that the differences in weight follow an approximately normal distribution (is this assumption necessary?).

7. There are about 25,000 high schools in the United States and each high school has a principal. These 25,000 high schools also employ a total of about one million teachers. As part of a national survey of education, a simple random sample of 625 high schools is chosen.

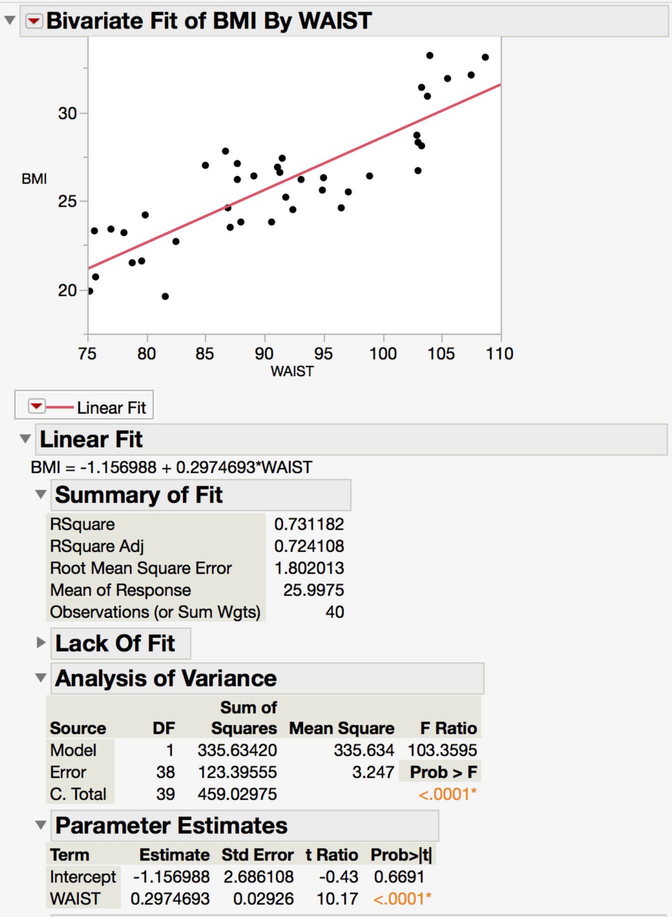
(a) In 510 of the sample high schools the principal has an advanced degree. Use the sample data to test the claim that over 80% of high school principals in the US have advanced degrees. Use a 5% significance level.

(b) The 625 sample high schools described above employ a total of 12,000 teachers, of whom 4,200 had advanced degrees and your friend uses this data to compute the 95%-confidence interval 35% ± 0.85% for the percentage of US high school teachers with advanced degrees. Is your friend right? Explain.

8. A person from the 'at-risk' population has a 10% chance of being HIV-positive. A preliminary screening test for HIV is correct 95% of the time (i.e., 5% false positives and 5% false negatives). If a randomly selected person from the 'at-risk' population tests positive for HIV in the initial screening, what is the probability that they are actually infected?

9. A researcher studies the relation between Male BMI (body mass index – kg/m^2) and waist circumference (cm). The sample linear correlation between BMI and Waist is $r = 0.855$. Test the claim that there is a significant linear relation between BMI and waist circumference at the 1% significance level.

10. Continuing from the previous problem... The JMP output of a linear regression predicting Male BMI by waist circumferences is displayed below, including the residual plot and a normal quantile plot of the residuals.



Bivariate Fit of BMI By WAIST

Linear Fit

Linear Fit

BMI = -1.156988 + 0.2974693*WAIST

Summary of Fit

RSquare	0.731182
RSquare Adj	0.724108
Root Mean Square Error	1.802013
Mean of Response	25.9975
Observations (or Sum Wgts)	40

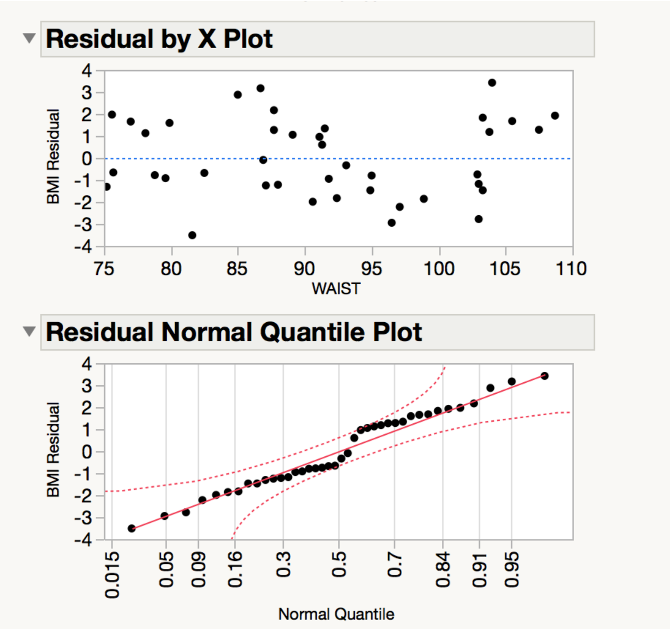
Lack Of Fit

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	335.63420	335.634	103.3595
Error	38	123.39555	3.247	Prob > F
C. Total	39	459.02975		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-1.156988	2.686108	-0.43	0.6691
WAIST	0.2974693	0.02926	10.17	<.0001*



Residual by X Plot

Residual Normal Quantile Plot

(a) Do the residuals appear to be *'healthy'* (i.e., do they appear to satisfy the requirements of the standard linear regression model)? Explain.

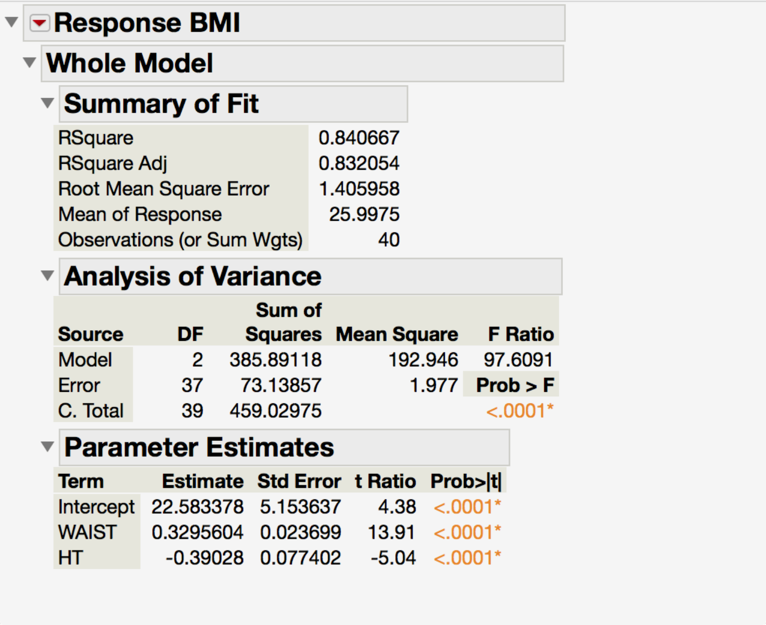
(b) Is the model significant? Give your answer in terms of an appropriate hypothesis test (all the usual steps).

(c) What percentage of variation in BMI is explained by Waist circumference? Explain.

(d) Give a 'practical' interpretation of the WAIST coefficient.

(e) Find a 95% confidence interval for the WAIST coefficient.

11. Continuing from the previous problem... Below is JMP output for a regression of BMI (kg/m^2) on Waist circumfrence (cm) and Height (inches).



Response BMI

Whole Model

Summary of Fit

RSquare	0.840667
RSquare Adj	0.832054
Root Mean Square Error	1.405958
Mean of Response	25.9975
Observations (or Sum Wgts)	40

Analysis of Variance

		Sum of		
Source	DF	Squares	Mean Square	F Ratio
Model	2	385.89118	192.946	97.6091
Error	37	73.13857	1.977	Prob > F
C. Total	39	459.02975		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	22.583378	5.153637	4.38	<.0001*
WAIST	0.3295604	0.023699	13.91	<.0001*
HT	-0.39028	0.077402	-5.04	<.0001*

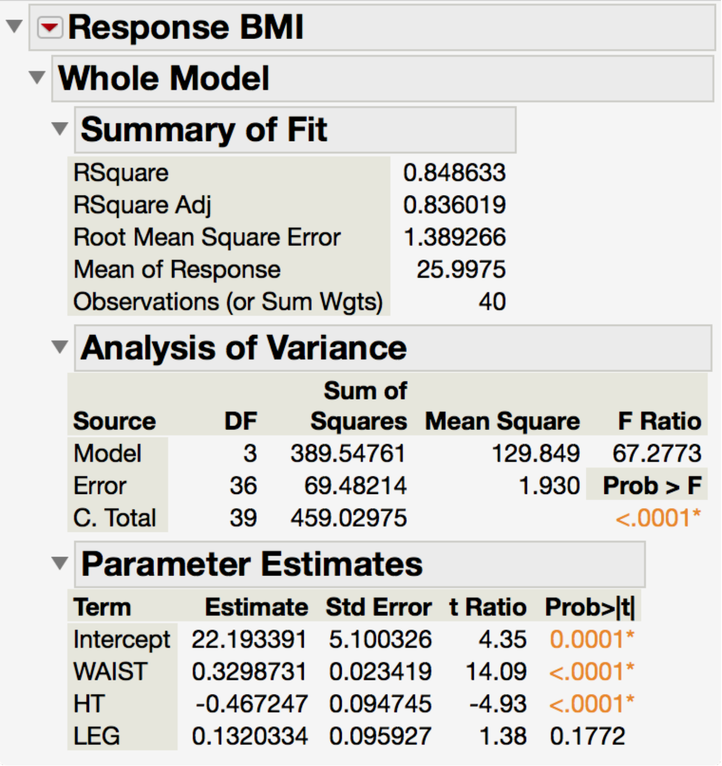
(a) According to this output, is the linear model significant at the 0.01 significance level? Explain.

(b) Is the coefficient for HEIGHT significant at the 0.01 significance level?

(c) What is the interpretation of the coefficient for HEIGHT? Does it make sense that it is *negative*?

(d) Which of the two models for predicting BMI is better — this one or the previous one? Explain your answer in detail.

12. BMI problems concluded... The JMP output for a regression of BMI on waist circumference, height and upper leg length (cm) is given below. Should the variable LEG be included in the model? Explain your answer in detail.



Response BMI

Whole Model

Summary of Fit

RSquare	0.848633
RSquare Adj	0.836019
Root Mean Square Error	1.389266
Mean of Response	25.9975
Observations (or Sum Wgts)	40

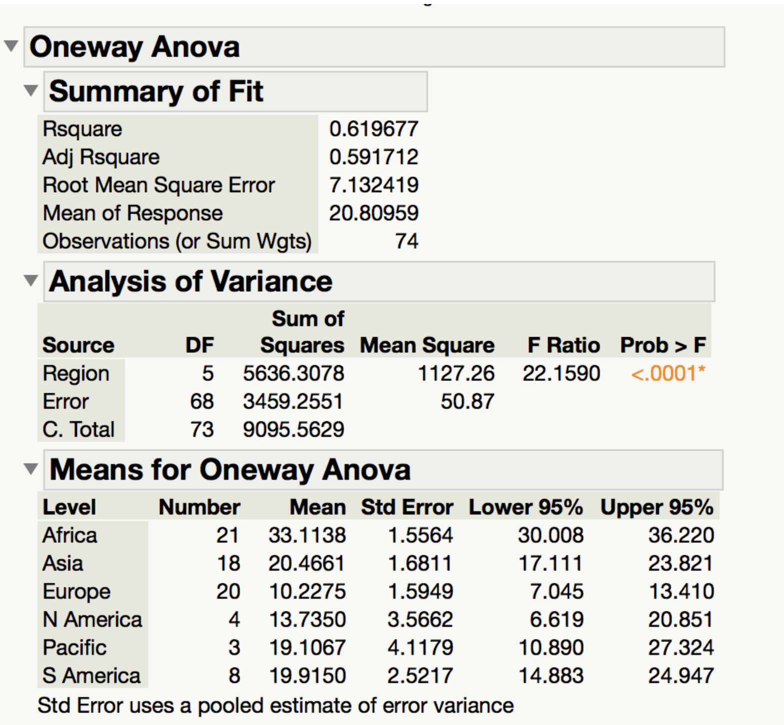
Analysis of Variance

		Sum of		
Source	DF	Squares	Mean Square	F Ratio
Model	3	389.54761	129.849	67.2773
Error	36	69.48214	1.930	Prob > F
C. Total	39	459.02975		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	22.193391	5.100326	4.35	0.0001*
WAIST	0.3298731	0.023419	14.09	<.0001*
HT	-0.467247	0.094745	-4.93	<.0001*
LEG	0.1320334	0.095927	1.38	0.1772

13. It is common to divide the world into the following six regions: Africa, Asia, Europe, N. America, Pacific and S. America. The JMP output of a study of birth-rate by country in each region is given below.



Oneway Anova

Summary of Fit

Rsquare	0.619677
Adj Rsquare	0.591712
Root Mean Square Error	7.132419
Mean of Response	20.80959
Observations (or Sum Wgts)	74

Analysis of Variance

		Sum of			
Source	DF	Squares	Mean Square	F Ratio	Prob > F
Region	5	5636.3078	1127.26	22.1590	<.0001*
Error	68	3459.2551	50.87		
C. Total	73	9095.5629			

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
Africa	21	33.1138	1.5564	30.008	36.220
Asia	18	20.4661	1.6811	17.111	23.821
Europe	20	10.2275	1.5949	7.045	13.410
N America	4	13.7350	3.5662	6.619	20.851
Pacific	3	19.1067	4.1179	10.890	27.324
S America	8	19.9150	2.5217	14.883	24.947

Std Error uses a pooled estimate of error variance

(a) Use the output from the table below to help you test the claim that the mean birth weights in all regions are equal. Follow all the usual steps, in particular identify the precise distribution of the test statistic.

(b) Is there a region that has a significantly larger mean birth rate? Justify your answer using appropriate statistics from the JMP output.