

Stat 500 - Midterm 2 - Off campus, Week of 14-18 November 2005

Please put your name on the back of your answer book. Do NOT put it on the front.
Thanks.

- The exam is closed book, closed notes. Use only the formula sheet and tables I provide today. You may use a calculator.
- Write your answers in your blue book. Ask if you need a second (or third) blue book.
- You have 2 1/4 hours (135 minutes) to complete the exam.
- Tables, a formula sheet, and scrap paper are provided.
- Points are indicated for each question. The total is 120.
- Important reminders:
 - budget your time. Some parts of each question should be easy; others may be hard. Make sure you do all parts you can.
 - notice that some parts do not require any computations.
 - show your work neatly so you can receive partial credit.
- Good luck!

1. 35 points. Safety of a potential vaccine. Vaccines, like all potential medicines, undergo a rigorous review before they are approved for human or animal use. One of the first evaluations is safety: does the vaccine make healthy individuals sick? Hopefully, the answer is no. The following (made up) data came from a safety study of a potential vaccine. One of the safety concerns with this vaccine is whether it induces a fever after it is administered.

Two treatments were used in this study: VACCINE and PLACEBO. Individuals were given a dose of the vaccine (VACCINE group) or of a placebo (PLACEBO group). After 24 hours, if the individuals body temperature was higher than normal, the individual was recorded as 'FEVER'. If the temperature was normal or low, the individual was recorded as 'NORMAL'.

You are given the following summary of the data:

Treatment	Normal	Fever	Total
Placebo	310	140	450
Vaccine	277	173	450
Total	587	313	900

- Calculate the Chi-square statistic to test whether $P[\text{fever in the PLACEBO group}] = P[\text{fever in the VACCINE group}]$.
- Calculate the d.f. for the test in 1a. Approximate the p-value.
- Compute the odds ratio. Express the odds ratio so that if the vaccine increases the probability of fever, the odds ratio is larger than 1.
- Compute a 95% confidence interval for the odds ratio.
- Summarize your results in a short conclusion about whether this vaccine affects the probability of having a fever, and if so, by how much.

After you share your conclusions with the study coordinator, you discover that the data in the above table came from two studies.

- Most of the data came from a study of 40 individuals. 20 were randomly assigned to receive the VACCINE; the other 20 were randomly assigned to receive the PLACEBO. The presence of fever was recorded for 20 days for each person. You are given the following summary of the data from this study:

Treatment	Normal	Fever	Total
Placebo	263	137	400
Vaccine	238	162	400
Total	501	299	800

Is the Chi-square test an appropriate test of equal probability of fever for these data? Explain why or why not. Note: **No computations required.**

- The other study looked at 50 individuals. Each individual was given the VACCINE and the presence or absence of fever recorded. After a waiting period of two weeks, each individual was given the PLACEBO, and the presence or absence of fever recorded. You are given the following summary of the data from this study:

Treatment	Normal	Fever	Total
Placebo	47	3	50
Vaccine	39	11	50
Total	86	14	100

Is the Chi-square test an appropriate test of equal probability of fever for these data? Explain why or why not. Again: **No computations required.**

2. 45 points. Health of factory workers. The following data were collected in a study of the health of paint sprayers in an auto assembly plant. Two of the variables that were measured on each of the 103 workers in the study were H, the haemoglobin concentration, and L, the lymphocyte count. These are measures of two different components of the blood.

The following quantities may help you answer the questions:

The observed intercept and slope in the regression $H_i = \beta_0 + \beta_1 L_i + \epsilon_i$ are $b_0 = -59.4$, $b_1 = 2.4$

The estimated s.d. of observations around the line is $s_e = 2.1$

The error SS for the regression $H_i = \beta_0 + \beta_1 L_i + \epsilon_i$ is 445.41

The error SS for the regression $H_i = \beta_0 + \beta_1 L_i + \beta_2 L_i^2 + \epsilon_i$ is 442.75

The error SS for the regression $H_i = \beta_0 + \beta_1 L_i + \beta_2 L_i^2 + \beta_3 L_i^3 + \epsilon_i$ is 438.20

The sum-of-squares of lymphocyte counts, $\sum(x_i - \bar{x})^2$, = 1736.

The mean lymphocyte count is 31.

The correlation coefficient between H and L is 0.90.

- (a) What statistic is the most appropriate to describe the association between haemoglobin concentration and lymphocyte count? You may answer with one of the values I've provided, or some other statistic.

No matter how you answered the previous question, the investigators want you to fit the regression: $H_i = \beta_0 + \beta_1 L_i + \epsilon_i$.

- (b) Test $H_0: \beta_1 = 0$. Report your test statistic and two-sided p-value.
- (c) The investigators use the fitted regression to predict average haemoglobin concentration at three possible lymphocyte counts: $L_i = 26$, $L_i = 31$, and $L_i = 35$. Which prediction is the most precise?
- (d) Here are a residual plot and a normal quantile-quantile plot for the fitted regression. List the assumptions made in the regression, then assess each using the information in the plots.

- (e) The investigators want to know whether the relationship between haemoglobin concentration and lymphocyte count follows a straight line. Test this, if possible from the available information. If not, say what additional information you need.

- (f) There are a few additional workers for whom the lymphocyte count was not measured. One of those workers had a haemoglobin concentration of 5.4.
- i. If it is possible given the available data, estimate the lymphocyte count for that individual.
 - ii. If possible from the available data, Calculate the s.e. of your prediction.

If answers to either question are not possible without additional information, say what additional information is needed.

- (g) The investigators are concerned about collecting redundant data. If the correlation between haemoglobin and lymphocyte exceeds 0.8, they may consider collecting only one variable, instead of two. Test whether the correlation is larger than 0.8, i.e. test $H_0: \rho \leq 0.8$ vs $H_a: \rho > 0.8$.
- (h) The investigators will repeat the study on a different population of industrial workers and come to you for a sample size recommendation. Both the variance of the lymphocyte counts and the variance of observations around the regression line in the new population are expected to be the same as in the current data set. If the desired s.e. for β_1 is 0.1, what sample size (# of observations) would you recommend?

3. 40 points. Estimating demand for gasoline. Econometricians are often interested in estimating demand curves. A demand curves describes the relationship between price and consumption of a specific product (demand). Estimating such curves correctly involves many issues well beyond what we have considered in this class. You are to use the regression tools we have considered in this class.

The following data are from a study to estimate the relationship between gasoline price and consumption. The data are the annual data on the US economy from 1960 to 1986. There are twenty seven (27) observations in the data set, one for each year. An economic detail: all variables are adjusted for the effects of inflation. The prices and income look small, but that is because they are expressed in 1967 dollars. The variables in the data set are:

Variable	explanation	units	Variable	explanation	units
gascons	total gasoline consumption	million barrels	gasprice	price index for gasoline	dollars
income	per capita disposable income	dollars	newcar	price index for new cars	dollars
usedcar	price index for used cars	dollars	bus	price index for public transportation	dollars
durable	price index for durable goods	dollars	nondur	price index for non-durable goods	dollars
service	price index for services	dollars	year2	year squared	
year	calendar year				

The goal of the study is to estimate the relationship between gasoline price and gasoline consumption. Is a higher gas price associated with a change in consumption? If so, how big is the effect.

Summary statistics for each of the 11 variables (10 X variables and *gascons*, the response) are:

Variable	N	Mean	Std Dev	Minimum	Maximum
year	27	73.0000000	7.9372539	60.0000000	86.0000000
gascons	27	207.0333333	43.7989287	129.7000000	269.4000000
gasprice	27	1.9021111	1.1679056	0.9140000	4.1090000
income	27	8513.52	1455.63	6036.00	10780.00
newcar	27	1.3813333	0.4279747	0.9910000	2.2400000
usedcare	27	1.7122963	0.9747435	0.8360000	3.7970000

bus	27	1.8623333	1.1083126	0.8100000	4.2640000
durable	27	0.6748889	0.2231702	0.4440000	1.0530000
nondur	27	0.6112593	0.2731639	0.3310000	1.0750000
service	27	0.6008148	0.3026135	0.3020000	1.2240000
year2	27	5389.67	1160.15	3600.00	7396.00

A scatterplot matrix of the data is on the next page.

The investigators considered 5 models. The error SS are included for each model

	SS_{error}	model equation
(1)	127.29	$gascons = \beta_0 + \beta_1 gasprice + \beta_2 year + \beta_4 income + \beta_5 newcar + \beta_6 usedcar + \beta_7 bus + \beta_8 durable + \beta_9 nondur + \beta_{10} service + \epsilon$
(2)	207.64	$gascons = \beta_0 + \beta_1 gasprice + \beta_2 year + \beta_3 year^2 + \beta_4 income + \beta_5 newcar + \beta_6 usedcar + \beta_7 bus + \beta_8 durable + \beta_9 nondur + \beta_{10} service + \epsilon$
(3)	800.96	$gascons = \beta_0 + \beta_1 gasprice + \beta_4 income + \beta_6 usedcar + \epsilon$
(4)	841.25	$gascons = \beta_0 + \beta_1 gasprice + \beta_4 income + \beta_7 bus + \epsilon$
(5)	607.42	$gascons = \beta_0 + \beta_1 gasprice + \beta_2 year + \beta_3 year^2 + \beta_{10} service + \epsilon$

Please use these results to answer the following:

- The estimate $\hat{\beta}_1$ in model 2 is -23.65. Please give a careful interpretation of this value. Include units, if possible.
- Construct a test of whether model 2 fits significantly better than model 3, if this is possible from the available information. If not, explain what other information is needed.
- Construct a test of whether model 5 fits significantly better than model 4, if this is possible from the available information. If not, explain what other information is needed.

The remaining four questions concern model 3. SAS output with additional information for model 3 is included on the last two pages.

- Use model 3 to predict gas consumption when $GASPRICE = 1.10$, $USED CAR = 0.90$, and $INCOME = 7000$.
- Estimate the standard deviation of a predicted **observation** at $GASPRICE = 1.10$, $USED CAR = 0.90$, and $INCOME = 7000$. The appropriate value of $x_i(X'X)^{-1}x_i'$ is 0.0830.
- The plot of residuals vs. predicted values (1st page of SAS output) suggests a problem with lack of fit. We have ignored violations of some assumptions because they have little effect on the desired answers. Do you have any concerns about the prediction and standard deviation you calculated in previous two parts?
- The primary goal of the study is to estimate β_1 , the partial regression coefficient for gas price. Do you have any concerns about the regression? Briefly explain of your concerns. If you don't have any, just say 'none'.