# Final Exam

Stats 120A
Winter 2005

Name:

Instructions:
This final is due Wednesday, March 23rd, 4pm. You may ask me any question you like, but you may not talk to anyone else about the exam. This includes emails, internet "chats", text messaging, or any form of communication with another human being. You are welcome, though, to make use of the UCLA library and its electronic archives. You are welcome to use any books or any notes from class. And you are welcome to ask me any question you like.

I will accept hardcopies only. No emailed exams. If you can't turn it in hardcopy form by March 23, then you need to turn it in earlier.

The first question is worth 20% each. The last is worth 80%.

The first two questions involve analyzing data from your fellow students. The directions are rather vague: "Analyze". This means you should provide:

1) The URL of the file.
Only the last two items in the pathname are unique, and you need only supply these. For example, do not write "http://www.stat.ucla.edu/~rgould/120w05/dataproject/smiith/experiment.txt" Instead write "smiith/experiment.txt".

2) The "research question" you will answer. This is a question about the real world and not about statistics. So "What is the value of the slope between height and weight?" is NOT a valid research question. But "To what extent are height and weight related?" is a valid question.

3) Your answer to the research question.

4) A short description of how you analyzed the data to achieve this answer, along with a clear statement of any assumptions you made.

5) A discussion of why you think those assumptions are valid. If you feel they are not valid, discuss how this affects your conclusion. (Be sure you chose a research question that can be answered with this data!)

CHOOSE ONE OF PROBLEMS #1 AND #2:
1. Choose a data set from the dataproject directory that is NOT your own and analyze it.

2. Choose either your own "controlled" or "observational" data set (the one that you submitted) and analyze it.

DO PROBLEM #3:
3. A recent NY Times article ("More Birthdays and Less Alcohol", March 1 2005) announced a UCLA study that confirmed the long-held belief that the amount of alcohol that people drink declines with age. The study was based on a random sample of about 14000 adults in the US who were surveyed at approximately 5 year intervals for 20 years. A small subset of the data is available at

**http://www.stat.ucla.edu/~rgould/120w05/datasets/alcohol.small.txt**

This is a text file that is tab-delimited and includes a header.

About the data:
qfi stands for "quantity/frequency index". Approximately, it is average the number of alcoholic drinks per day the subject has reported drinking, for the last month (at the time of the survey.) There are two such variables. qfi82 was assessed in 1982 and qfi92 was assessed in 1992. Keep in mind that these are the same people -- the first variable is how much they drank in 1982 and the second how much in 1992. The other variables are self-explanatory, except for X, which is simply the row number.

A note about log transforms:
You might possibly want to take the log of a variable that has negative values. This is impossible, of course, and so there's a little trick. Suppose that x is a vector and x has negative values:

```
smallest.x <- min(x[!is.na(x)])
log.x <- log(x - smallest.x + 1)
```

The basic idea is that we shift the distribution of x over so that the new smallest value is bigger than 0. To do this we need to know the minimum value, and that's what the first line does. (R assumes that the "missing" value is the smallest possible value, and so we have to exclude missing values from our consideration, which is what that "!is.na(x)" is all about.

Answer the following questions.

a) (2 pts for right choice, 3 pts for good explanation.) At one stage of their investigation, researchers fit a regression line to examine how drinking varies with age. This is their output:

$$\text{Log(qfi92)} = -1.455 - 0.011 * \text{age}$$

and here is the R output:

```
Call:
lm(formula = log(qfi92) ~ age)

Residuals:
    Min      1Q  Median      3Q     Max
-3.77012 -1.45388 -0.04917  1.63039  5.08444

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.455222   0.092719 -15.695  < 2e-16 ***
age         -0.010764   0.002173  -4.954 7.53e-07 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 1.88 on 4862 degrees of freedom
Multiple R-Squared: 0.005022,         Adjusted R-squared: 0.004817
F-statistic: 24.54 on 1 and 4862 DF,  p-value: 7.53e-07
```

Assuming the assumptions behind the linear model are satisfied, which of the following is a valid interpretation of this model? Choose A or B and explain (in one or two sentences) your choice.

Choice A: The negative slope shows that people tend to drink less as they age.
Choice B: In 1992, the older people tended to drink less than the younger.

b) (20 pts) **Do people drink less as they age?** Use the data provided to answer this question to the best of your ability. You should answer two research questions:

i) do people drink less as they age?
ii) does the answer to this question depend on gender and/or race?

Your analysis must

i) include an initial exploration of the data, including graphics and summary statistics. Do not simply printout all possible graphs and statistics, but instead describe what you learned from them, and include only important features.
ii) state all assumptions required to reach your conclusion. Show whether the data support these assumptions.
iii) if you base your answer on a statistical model, explain how you achieved that model. Describe the model. What are the relative merits and weaknesses of your model over other possibilities?

You do not need to demonstrate mastery of each and every technique covered in class. You will be graded based on whether you make choices appropriate to the challenge (determine whether people drink less as they age) and whether you make valid interpretations and discussions about your ability (and the data's ability) to make this inference. Please do not include R output unless it is needed to support your argument.

Note: the published study used a 20 years worth of data based on four surveys administered every 5 years. The extract that you have been provided with includes some of the responses from the second and fourth surveys.