Homework policies reminder: Late homework is not accepted, with one exception: you may hand in an assignment late (by the beginning of the next class meeting) *once* during the semester. If you are planning to hand in a homework late, email me by the beginning of class on the original due date.

Your assignments should be typed, except for equations, symbols, and graphics, which may be handwritten. Please integrate graphics into the flow of your text; do not put all your graphics and tables in an appendix at the end of your written material.

I encourage you to discuss the homework with other students and with me. Your final answer should be written in your own words. On your paper, list the students with whom you had substantial collaborations.

SHORT-ANSWER QUESTIONS (3 POINTS EACH)

(1) In class, we said that if the standard regression model assumptions are satisfied, then the least-squares line passes through the conditional means. Sketch a picture of a dataset in which not all of the standard regression model assumptions are satisfied, and the least-squares line does not pass through all the conditional means of the Y-values. Explain briefly or indicate on your picture where the LS line misses the conditional mean.

(2) In a survey of 988 men aged 18–24, the regression equation for predicting height from weight was

$$\text{height in inches} = 62.4 + (0.047)(\text{weight in pounds})$$

Is the following statement a correct interpretation of the regression line: “If someone puts on 10 pounds, he will get taller by (0.047)(10) = 0.47 inches”? If not, explain what the slope means.

(3) In the dataset above, suppose the conditional SD of Y is $s = 2.2$ inches. What percentage of all 200-pound men are taller than 74 inches? (Assume the regression model assumptions are met.)

(4) Suppose we have two datasets each having n observations: dataset 1 with variables X_1 and Y_1, and dataset 2 with variables X_2 and Y_2. (X_1 and X_2 are the explanatory variables, and Y_1 and Y_2 are the response variables.) There is a positive linear relationship between X_1 and Y_1, and a positive linear relationship between X_2 and Y_2. Suppose Y_1 has marginal standard deviation 10 and conditional standard deviation 8, and Y_2 has marginal standard deviation 10 and conditional standard deviation 2. For each of Y_1 and Y_2, assume the conditional standard deviation is constant. Which one of the following statements must be true? Explain briefly or draw a picture.

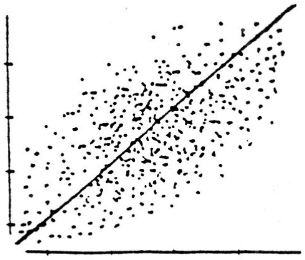
(a) The correlation between X_1 and Y_1 equals the correlation between X_2 and Y_2.

(b) The correlation between X_1 and Y_1 is greater than the correlation between X_2 and Y_2.

(c) The correlation between X_1 and Y_1 is less than the correlation between X_2 and Y_2.

(d) None of the above statements can be determined from the information given.

(5) Is the line drawn in the scatterplot at right the regression line? Why or why not? Explain briefly. You may assume that the standard regression model assumptions are satisfied.



Skyscrapers (18 points)

How does the height (*Y*) of a skyscraper depend on the number of stories it has (*x*)? Sixty buildings were selected at random from a table of US tall buildings given in the *World Almanac*, and their heights and number of stories were recorded. A few other selected skyscrapers were added to the dataset. The data are on Moodle. You can read in the data using the following commands:

```
# read in dataset
setwd("~/Documents/Folder1/Folder2/Fill in your pathname here")
data <- read.csv("skyscrapers.csv")

# define variables
height <- data[,"height"]
stories <- data[,"stories"]
year <- data[,"year"]
building <- data[,"building"]
```

(6a) How does the height of a tall building depend on the number of stories it has? To explore this question, make a scatterplot of height vs stories using the `plot` command. (Height should be on the Y axis.) Briefly describe the relationship between height and stories: Does it appear to be linear? Is the relationship a strong one? Copy the scatterplot and hand it in (paste it into your write-up).

(6b) Calculate the correlation of height and stories using the `cor` command. What is the value of the correlation coefficient. Would you say this is a strong correlation, medium, or low?

(6c) Calculate the regression of height and stories using the following commands:

```
fit1 <- lm(height ~ stories)
summary(fit1)
```

Copy the output and hand it in. What are the equation of the regression line, the value of the conditional SD of height, and the value of R-squared?

(6d) Calculate a 95% confidence interval for ß$_1$. How would you explain the meaning of this confidence interval, in the specific context of this dataset, to an architect who has not taken statistics?

(6e) Using the output from **(c)**, test the hypothesis that ß$_1$ = 0. State your null and alternative hypotheses and report the test statistic and p-value; circle or highlight the relevant part of the output from **(c)**. Is your conclusion surprising?

(6f) Make a residual plot for the regression of height and stories using the following commands:

```
plot(fitted(fit1), resid(fit1))
abline(h=0)
```

Copy the residual plot and hand it in. Are there any apparent violations of the regression model assumptions? Explain briefly.

(6g) Make a Normal probability plot of the residuals using the following commands:

```
qqnorm(resid(fit1))
qqline(resid(fit1))
```

Copy the NPP and hand it in. Are there any apparent violations of the regression model assumptions? Explain briefly.

(6h) Have there been any trends in the height/story relationship over time? Make a scatterplot of the residuals vs year and hand it in. Are there any patterns over time or other notable features?

(6i) Are there any buildings that seem to be unusually high for their number of stories? If so, which building(s)? You can identify points on the residual plot using the following commands:

```
plot(year, resid(fit1))
abline(h=0)
identify(year, resid(fit1), labels=building, cex=.6)
```

Each time you click on a point in the plot, the name of the building should appear. To break out of "point identification mode" and return to normal usage of R, command-click or control-click on the plot, or type the ESC key (depending on what operating system you are using). What factors might account for the unusual height of this building(s)? Hint: Think about how exactly the "height" of a building might be measured — particularly a building that does not have a flat roof. If the architect puts a TV antenna on top of a building, does that increase its "height"? Some background research may be helpful. For instance, see http://www.skyscraperpage.com/diagrams/ for pictures of many of these buildings. Another possibly useful site is http://www.emporis.com/statistics/history-of-worlds-tallest-buildings.