

ADDITIONAL HOMEWORK PROBLEMS: ECON 371
Spring 2008

Chapter 2

Note: Unlike later additional questions, these two questions use data sets provided with Anderson, Sweeny, and Williams.

- 1) On page 58 of Anderson, Sweeny, and Williams, read problem 40 and examine the data set. Using Minitab, open the data set for this problem, “Golf,” and then perform the following operations in Minitab:
 - a) Tally the data to get counts for each possible response.
 - b) Using the Chart command, make a bar chart for the frequency and percent frequencies for each possible response.
 - c) Make a pie chart of the data, labeling each slice with percent.

- 2) On page 61 of Anderson, Sweeny, and Williams, read problem 46 and examine the data set. Using Minitab, open the data set “High-Low” and then perform the following operations in Minitab:
 - a) Make dotplots of high and low temperatures, with both dotplots overlaid in the same diagram.
 - b) Make boxplots of both high and low temperatures, with both boxplots overlaid in the same diagram.
 - c) Make a frequency histogram of the high temperatures.
 - d) Make a percent frequency histogram of the low temperatures.
 - e) Make a cumulative frequency histogram of the high temperatures.
 - f) Make a stem-and-leaf display of the low temperatures.
 - g) Plot the high temperatures against the low temperatures, putting the high temperatures on the Y-axis.

Chapter 4

1) Given $P[A] = .5$ and $P[A \cup B] = .6$, find $P[B]$ if:

- a) A and B are mutually exclusive.
- b) A and B are independent
- c) $P[A|B] = .4$.

2) A certain hapless individual can never remember whether the dishes in the dishwasher are clean or dirty. Eighty per cent of the time, he runs the dishwasher immediately after loading it, so eighty per cent of the time the dishes are clean. But twenty per cent of the time, he forgets to start the dishwasher, leaving the dishes dirty. Finding the dishwasher full, and unable to remember if the dishes are clean, he reaches into the dishwasher, pulls out a glass, and holds it up to the light. Unfortunately, this isn't a perfect test. Ten per cent of his glasses would look clean even without being run through the dishwasher. Five per cent would look dirty, even if they had been run through the dishwasher. If the glass looks clean, what is the probability that the dishwasher has been run? If the glass looks dirty, what is the probability that the dishwasher has been run?

3) Of all of Michael Jordan's fans:

- 25% wear Nikes
- 10% wear Nikes and eat Wheaties
- 30% eat Wheaties
- 8% wear Nikes and drink Gatorade
- 15% drink Gatorade
- 5% drink Gatorade and eat Wheaties
- 2% wear Nikes, eat Wheaties, and drink Gatorade

What is the probability that a randomly selected Michael Jordan fan:

- a) Uses one or more of these products?
- b) Drinks Gatorade or eats Wheaties, or does both?
- c) Wears Nikes given that he consumes either Gatorade or Wheaties?
- d) Drinks Gatorade, and doesn't wear Nikes?
- e) Drinks Gatorade, given that he doesn't wear Nikes?

4) A mail-order firm considers three possible foul-ups in filling an order:

- A: The wrong item is sent.
- B: The item is lost in transit.
- C: The item is damaged in transit.

Assume that event A is independent of both B and C and that events B and C are mutually exclusive. The individual event probabilities are $P(A) = .02$, $P(B) = .01$, and $P(C) = .04$. Find the probability that at least one of these foul-ups occurs for a randomly chosen order.

Chapter 5

1) A widow's only asset is a closely held family business whose current market value is unknown. However, if the business is sold at auction, it is known that the selling price (measured in hundreds of thousands of dollars) will be a random variable with the following distribution:

x	p(x)
1	.3
2	.1
3	.2
4	.1
5	.3

- a. Find the expected value and standard deviation of X,
- b. A family friend offers to pay \$250,000 (2.5 units of X) for the business with the intention of running the risk of the auction himself. would it be wise for the widow to take this deal if she is interested only in maximizing expected wealth? Explain.
- c. Suppose the widow has a utility function relating wealth to utility given by:

$$U(x) = 10x - x^2$$

- Would it be wise to take this deal if the widow is interested in maximizing expected utility? (Answer by explicitly calculating expected utility under each alternative)
- d. Do you think that the person with the utility function in part c is risk neutral or risk averse? Why? (Remember that one way we could think of measuring risk is by the variance of wealth)

2) Suppose the probability distribution for the number of quarterback sacks Reggie White (Defensive end) achieves in a game is given by:

$$p(x) = \frac{4-x}{10}, \text{ where } x = 0, 1, 2, \text{ or } 3$$

- a) Find the mean and variance of the number of sacks achieved per game.
 - b) Suppose White's contract includes an incentive clause, so that his base pay is \$100,000 per game with a \$7500 bonus for each quarterback sack achieved. What are the mean and variance of his compensation per game?
- 3) You have been given a binomial problem to solve, in which $n = 18$ and $p = .42$. The question asks you to find the probability of exactly 6 successes. Sitting at your computer, you generate the following Minitab output. What is the (numerical) answer?

MTB > cdf;
SUBC > binomial N=18 p=.42.

BINOMIAL WITH N = 18 P = 0.420000	
K	P(X LESS OR = K)
0	0.0001
1	0.0008
2	0.0052
3	0.0223
4	0.0687
5	0.1628
6	0.3105
7	0.4938
8	0.6764
9	0.8232
10	0.9189
11	0.9693
12	0.9906
13	0.9978
14	0.9996
15	0.9999
16	1.0000

- 4) a. A very large shipment of parts contains 10% defectives. Two parts are chosen at random from the shipment and checked. Let the random variable X denote the number of defectives found. Find the probability function of this random variable. {Hint: If the population size is very large compared to the sample size, the sample draws are nearly statistically independent, so that you can use a distribution based on independence with no appreciable loss of accuracy.}
- b. A shipment of twenty parts contains two defectives. Two parts are chosen at random from the shipment and checked. Let the random variable Y denote the number of defectives found. Find the probability function of this random variable. Explain why your answer is different from that part (a).
- 5) Assume that typographic errors in the page proofs of a book occur with a frequency that obeys the Poisson distribution. If the average number of typographic errors per page is 3.7, what is the probability that two pages will contain a total of 2 or fewer errors?
- 6) In Virginia's *Cash 5* lottery game run in the summer of 1993, gamblers had to pick 5 numbers (without replacement) from the numbers 1 to 34. If you matched 3 of the 5 numbers, you won \$5. If you matched 4 of the 5 numbers, you won \$100, and if you matched all 5 numbers, you won \$100,000.
- What was the probability of winning \$5?
 - What was the probability of winning \$100?

- c) What was the probability of winning \$100,000?
- d) Each ticket cost \$1. On average, how much did players lose on each ticket they purchased? {Cost of ticket minus expected value of the ticket}

7) The population of Hawaii is 1% black. If a random sample of 50 people is drawn, what is the chance it contains no blacks? At least one black? Calculate your answer in two ways: using the binomial distribution, and using a poisson approximation to the binomial.

8) The joint probability distribution of the returns of two stocks (in percentage terms) are given by the following table:

		Return of X		
		10	20	30
Return of Y	6	.30	.10	.10
	8	.10	.20	.20

- a) Find σ_{xy} and ρ_{xy} .
 - b) What is the probability $Y = 6$, given that $X = 10$?
 - c) If you held a portfolio in which 4 tenths of your wealth was invested in stock Y and 6 tenths was invested in stock X, what would be the expected return and the *standard deviation* of the return on this portfolio?
- 9) In the game of football, field goals are worth 3 points and touchdowns are worth 7 points. (No safeties, no two point conversions, and no missed extra points in this problem.) Suppose the mean number of touchdowns the Redskins score in a game is 1.7, and the standard deviation of the number of touchdowns is 1.0. The mean number of field goals is 2.1, while the standard deviation of the number of field goals is 1.2. The correlation between the number of touchdowns and number of field goals is .30. What are the mean and standard deviation of points scored per game?

Chapter 6

1) Suppose X is a continuous random variable with the following density function:

$$f(x) = \begin{cases} 2x & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Find:

- a. $\Pr(0 < X < 1)$
 - b. $\Pr(X < .5)$
 - c. $\Pr(X > .8)$
 - d. $\Pr(.5 < X < .8)$
- 2) Let X , the time it takes for a new 3/8" chain to break under 20,000 pounds of pressure, be normally distributed with a mean of 45 minutes and a standard deviation of ten minutes. State whether each of the following statements is true or false and why.
- a. It is equally probable that the chain will break after 45 minutes as that it will break before 45 minutes.
 - b. It is more probable that the chain will break between 45 minutes and an hour's wear than between one-half hour and 45 minutes' wear.
 - c. It is equally probable that the chain will break before the first half hour's wear as it is that it will break after the first hour's wear.
 - d. A chain will break before the first hour and five minutes of wear 95.5% of the time.
- 3) The kilowatt demand at any given time on the Amgar Power Plant is normally distributed with a mean of 120,000 and a standard deviation of 10,000. If the plant can generate at most 150,000 kilowatts, what is the probability that at any given time there will be an overload?
- 4) The scores on an achievement test given to 231,126 high school seniors are normally distributed about a mean of 500. The distribution has a standard deviation of 90.
- a. What is the probability that an achievement score is less than 500?
 - b. What is the probability that an achievement score is between 320 and 680?
 - c. The probability is 0.85 that a score is more than what value?
- 5) During registration, students consult advisors with questions about course selection. A particular advisor noted that during registration, an average of 2 students per hour came to ask questions, although the exact arrival times of the students were random and unpredictable.
- a. What is the probability of having exactly one student arrive during a particular hour?
 - b. What is the probability the *next* student will appear sometime between one and two hours from now?

Chapter 7

1) A library has 97.5 feet of empty shelf space. They have on order 600 new books. It is known from long experience that a book's width is a random variable with $\mu = 2$ inches and $\sigma = 1$ inch. Assuming the new books constitute a random sample:

- a. What is the probability that the books will be too wide to fit on the shelf?
- b. Is the Central Limit theorem relevant to this problem? Why or why not?

2) The concession manager of a 15,000 seat sports arena knows that the number of ounces of Coca-Cola purchased by a randomly selected fan is a random variable with the following distribution:

x	p(x)
0	.3
8	.1
12	.3
16	.2
24	.1

- a) If an event is sold out, how many ounces of Coca-Cola must be on hand to guarantee that there is only a 1% chance of running out?
- b) If an event is sold out, what is the probability that at least 69% of the fans will purchase Coca-Cola?

3) A gun is fired at a target. The probability that a round hits the target is .7.

- a. If 1000 rounds are fired, estimate the probability that the proportion of hits will be between .67 and .72 inclusive.
- b. How many shots need to be fired in order to be 90% sure that at least 64% are hits?

4) Do you agree or disagree with the following statements? Explain your answer.

- a. The probability that there are ten defectives in a sample of 200 is the same as that 95% are not defective in a sample of 200.
- b. In tossing a coin, the probability that the proportion of heads equals one-half goes to one as the number of tosses goes to infinity means that one could be reasonably certain that if one flipped a coin 10,000 billion times, one would get exactly 5000 billion heads.
- c. If the mean and variance of a variable are known, according to the Central Limit Theorem, we can use the normal distribution to approximate the probability that the variable will exceed some number.
- d. If X is normally distributed, the only information we need know about X to answer probability statements about it is its mean and standard deviation.
- e. The mean of a sample is always exactly normally distributed.

5) Below is some Minitab computer output, which simulates 300 samples of size 7 drawn from a uniform distribution on the interval [0, 1]. (that is, a number picked at random between 0 and 1.) These 300 samples are used to judge the relative efficiency of the sample mean and sample median as estimators of the population mean of a uniform distribution. Begin by looking over this output.

```
MTB > random 300 c1-c7
SUBC> uniform a=0 b=1
MTB > rmean c1-c6 c8
MTB > rmedian c1-c6 c9
MTB > name c8='means' c9='medians'
MTB > histogram c8 c9
SUBC> same.
```

Histogram of means N=300
Each * represents 2 obs.

Midpoint	Count	
0.1	0	
0.2	4	**
0.3	23	*****
0.4	74	*****
0.5	96	*****
0.6	67	*****
0.7	32	*****
0.8	4	**
0.9	0	

histogram of medians N=300
Each * represents 2 obs.

Midpoint	Count	
0.1	1	*
0.2	15	*****
0.3	36	*****
0.4	61	*****
0.5	61	*****
0.6	49	*****
0.7	45	*****
0.8	29	*****
0.9	3	**

MTB > describe c8 c9

	N	MEAN	MEDIAN	TRMEAN	STDEV	SEMEAN
means	300	0.50511	0.50246	0.50512	0.11502	0.00664
medians	300	0.51595	0.50764	0.51632	0.16844	0.00973

	MIN	MAX	Q1	Q3
means	0.22097	0.78792	0.42198	0.59002
medians	0.09089	0.91068	0.38482	0.65394

- In a couple of sentences, explain what these histograms and statistics suggest about the bias and efficiency of these two estimators.
- Use these numbers to estimate the relative efficiency of the mean compared to the median in this application.
- Suppose you show these pictures to a friend, and he or she, while pointing to the histogram of c8 ('means'), says "I'm confused. I thought I remembered the uniform distribution from class. It was the flat one. But this isn't flat - it looks like a hill. What's going on?" Explain what is going on.

Unless announced otherwise in class, consider the following question optional.

- Suppose you have taken a sample of size n *without replacement* from a finite population of size N . You are interested in estimating the population variance, σ^2 . Consider the estimator of the form:

$$\hat{\theta} = k \left[\left(\sum_{i=1}^n X_i^2 \right) - n\bar{X}^2 \right]$$

Evaluate the expected value of this estimator, and find a value of k for which it is unbiased. Hint: Remember that if one samples without replacement from a finite population, the standard deviation of the sample mean is

$$\sqrt{\frac{N-n}{N-1}} \frac{\sigma}{\sqrt{n}}$$

Chapter 8

1) Minitab calculated a confidence interval for the mean difference in abrasion resistance between treated and untreated half-pieces of rubber. The printout is below:

```
MTB> set C5
DATA> 2.6 3.1 -.2 1.7 .6 1.2 2.2 1.1 -.2 .6
DATA> end
MTB> tint * C5
```

	N	MEAN	STDEV	SEMEAN	*PERCENT C.I.
C5	10	1.27	1.126	.356	(.2657 2.2743)

As you can see, the level of confidence ($1-\alpha$) has been omitted (see the asterisks). What is the level of confidence?

2) Someone presents you with a 95% confidence interval calculated from sample data: $\mu = 8 \pm 4.2$. In explaining what this means, the person says that "The probability that μ lies in the interval 3.8 to 12.2. is .95" Explain what is wrong with this interpretation, and give a correct interpretation.

3) Of the 14,714 families in West Falls, a random sample of 217 families was taken in order to determine the mean family income in this depressed area. A 95% confidence interval (\$3812 to \$4116) was established on the basis of the sample results.

Using only the above information, which of the following statements are valid?

- Of all possible samples of size 217 drawn from this population, 95% of the sample means will fall in the interval.
- Of all possible samples of size 217 drawn from this population, 95% of the population means will fall in the interval.
- Of all possible samples of size 217 drawn from this population, 95% of the confidence intervals established by the above method will contain the population mean.
- 95% of families in West Falls have means between \$3812 and \$4116.
- Using the above method, exactly 95% of the intervals so established will contain the sample mean \bar{X} .
- We do not know whether the population mean is in the interval \$3812 to \$4116.

Chapter 9

- 1) The security department of a factory wants to know whether the true average time required by the night watchman to make his rounds is 30 minutes. If, in a random sample of 32 rounds, the night watchman averaged 30.8 minutes with $\sigma = 1.5$ minutes, determine at the 1% level of significance whether this is sufficient evidence to reject the null hypothesis, $\mu \leq 30$ in favor of the alternative, $\mu > 30$. What is the power of the test if the true mean time is 31 minutes?

- 2) A bowl contains seven marbles of which θ are red while the others are blue. In order to test the null hypothesis, $\theta = 2$, against the alternative, $\theta = 4$, two of the marbles are randomly drawn without replacement and the null hypothesis is rejected if and only if both are red. Find the probabilities of committing type I and type II errors with this criterion.

- 3) Imagine you are chief of security for an Atlantic City casino. One of your men drags in a red-faced gambler, who he thinks has switched dice in the crap game. The suspicious pair of dice are presented to you. Knowing that a pair of fair dice will roll 7 one sixth of the time, and the usual trick of dishonest gamblers is to weigh the dice so as to make 7 come up more frequently, you decide to roll the dice 15 times. If you get 6 or more 7's you will call the police, whereas if you get fewer than 6 7's you will release the gambler with a warning. This, of course, is a hypothesis test.
 - a) State formally the hypothesis being tested, and the alternate hypothesis.
 - b) Give an expression for the exact probability of a type I error. Do not evaluate. (Hint: Do not use a normal approximation.)
 - c) Use a normal approximation to approximate the probability of type I error in part
 - d) If the dice are dishonest and actually come up 7 thirty percent of the time, give an expression for the exact probability of a type II error. Do not evaluate. (Hint: Do not use a normal approximation.)
 - e) Fill in the blanks with either "cheating" or "honest" as appropriate.
 - f) Saying the gambler is _____ when actually he is _____ is a Type I error. Saying the gambler is _____ when actually he is _____ is a Type II error.

- 4) In the following someone has used Minitab to perform a one tailed test of the null hypothesis that $\mu \geq 10$. Calculate the p-value for the two-tailed alternative that μ is not equal to 10. Show how you obtained your answer.

```
MTB > ttest mu=10 c2;  
SUBC> alternative = -1.
```

TEST OF MU = 10.000 VS MU L.T. 10.000

	N	MEAN	STDEV	SE MEAN	T	P VALUE
C2	31	10.233	1.059	.190	1.23	0.89

5) To test whether medical intervention to lower blood cholesterol can lower the risk of heart attack, a major clinical trial was done. Called the CPPT, or Coronary Primary Prevention Trial, it was a double blind study. A group of men with elevated cholesterol levels was randomly divided into control and treatment groups of equal size. The treatment group was given a drug called cholestyramine that lowers blood cholesterol. The control group was given a placebo. The study lasted several years. By the end of the study, a total of 356 heart attacks had occurred: 160 heart attacks in the treatment group, and 196 in the control group.

- a) What fraction of heart attacks would be expected to come from the treatment group if the treatment were ineffective? Formally write out the null and alternate hypotheses for an appropriate one-sided test.
- b) Calculate a p-value for the null and alternate in part a, and test the hypothesis at a 5% significance level. Say whether you accept or reject the null hypothesis. If you believe in 5% tests, would this be evidence for or against the treatment?
- c) Clinical trials of drug therapy in the past have often discovered unwanted side effects that actually cause increases in disease rates. Therefore it might be argued that a 2-tailed test is appropriate. What is the two-sided p-value for this data? In this case, would you accept or reject the null. If you believe in 5% tests, would this be evidence for or against the treatment?

6) State null and alternative hypotheses in the following situations.

- a. Consumer Reports wishes to test a cereal manufacturer's claim that its 32 ounce packages contain at least 32 ounces of cereal.
- b. Jones Political Advisory Service claims the percentage of people who favor the Republican presidential candidate in New York State is the same as that in New Jersey.
- c. A safety engineer doubts a glove manufacturer's claim that the average width of a certain welding glove is 2".
- d. The proportion of defective items produced by a certain process is reported to be at most 10%. A prospective buyer wishes to test this claim, using a low probability of rejecting the claim erroneously.

7) M.S. Kanarek and associates studied the relationship between cancer rates and levels of asbestos in the drinking water, in 722 Census tracts around San Francisco Bay. After adjusting for age and various demographic variables, but not smoking, they found a "strong relationship" between the rate of lung cancer

among white males and the concentration of asbestos fibers in the drinking water: $p\text{-value} < .001$.

Multiplying the concentration of asbestos by a factor of 100 was associated with an increase in the level of lung cancer by a factor of about 1.05, on average. (If tract B has 100 times the concentration of asbestos fibers in the water as tract A, and the lung cancer rate for white males in tract A is 1 per 1000 persons per year, a rate of 1.05 per 1000 persons per year is predicted in tract B.)

The investigators tested over 200 relationships – different types of cancer, different demographic groups, different ways of adjusting for possible confounding variables; the p -value for lung cancer in white males was by far the smallest one they got.

Does asbestos in the drinking water cause cancer? Is the effect a strong one? Discuss briefly.¹

¹ From Freedman, Pisani, Purves, and Adhikari, *Statistics, Second edition* The published article referred to is M.S. Kanarek et. al. "Asbestos in drinking water and cancer incidence in the San Francisco Bay area," *American Journal of Epidemiology*, vol. 112 (1980), pp. 54-72.

Chapter 10

1) Suppose one were interested in knowing whether higher education improved one's IQ. Imagine that you have been able to find 5 pairs of identical twins, where in each case, one of the twins has had a college education, while the other has not. You administer a standard IQ test to each of the 10 individuals, with the following result:

	twin without college	twin with college
Smith twins	112	118
Jones twins	98	105
Brown twins	120	135
Hall twins	111	99
Davis twins	104	113

Test the hypothesis that college does not improve IQ using $\alpha=.10$. Do you accept or reject the null hypothesis?

2) The *fog index* is used to measure the reading difficulty of a written text. The index is calculated through the following steps:

- i. Find the average number of words per sentence.
- ii. Find the percentage of words with three or more syllables.
- iii. The fog index is 40% of the sum of (i) and (ii).

A random sample of six advertisements taken from *Scientific American* has the following fog indices:

15.75 11.55 11.16 9.92 9.23 8.20

An independent random sample of six advertisements from *Sports Illustrated* had the following fog indices:

9.17 8.44 6.10 5.78 5.58 5.36

Stating any assumptions you need to make, test at the 5% level the null hypothesis that the population mean fog indices are the same against the alternative that the true mean is higher for *Scientific American* than for *Sports Illustrated*.

Chapter 14, Part 1

1) The following data on the Boston Marathon were collected by the US Army Research Institute and were published in the *Boston Globe* on April 20, 1992.

Year	Temperature	Injury Rate (%)
1984	46	4.7
1985	72	12.3
1986	59	6.5
1987	56	5.9
1988	54	6.6
1989	70	10.3
1990	68	8.4
1991	48	4.0

- Consider the simple linear regression of injury rates on temperatures from the data given above. Calculate the estimated slope and intercept of the regression line.
- Estimate the variance of the error terms in the population regression equation.
- Calculate r^2 for this regression.
- For a marathon run at 65 degrees, find a 99% prediction interval for the injury rate.

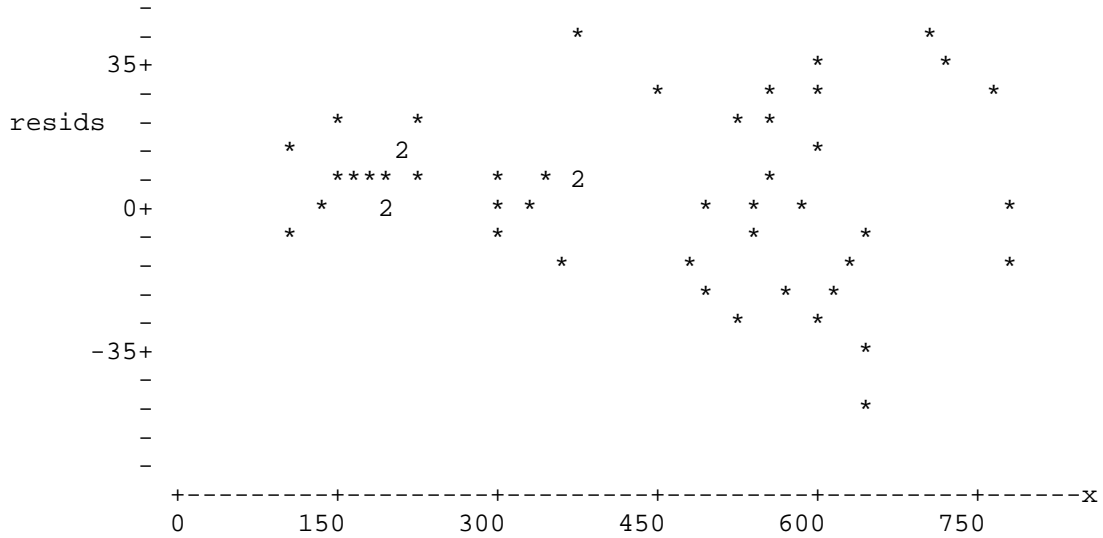
2) A 95% confidence interval for a regression slope was calculated on the basis of 1000 observations: $\beta = .38 \pm .27$. Calculate the p-value for the following hypothesis:

$$H_0 : \beta \leq 0$$

$$H_A : \beta > 0$$

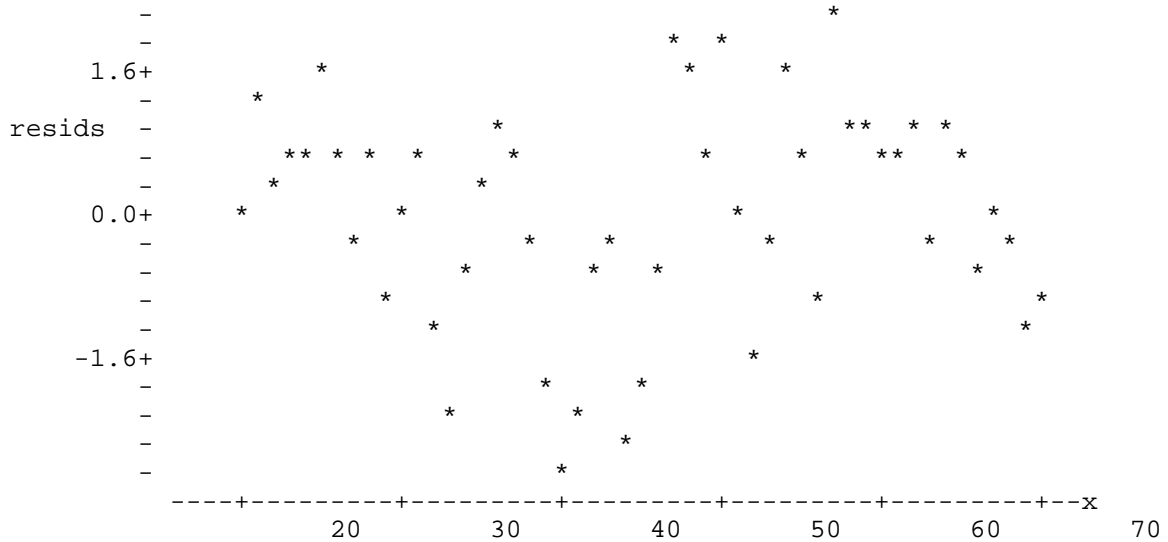
3) a) In the following residual plot, state which of the standard assumptions of the regression model appears to be violated, and explain what aspect of the plot led you to your conclusion.

MTB > plot c3 c2



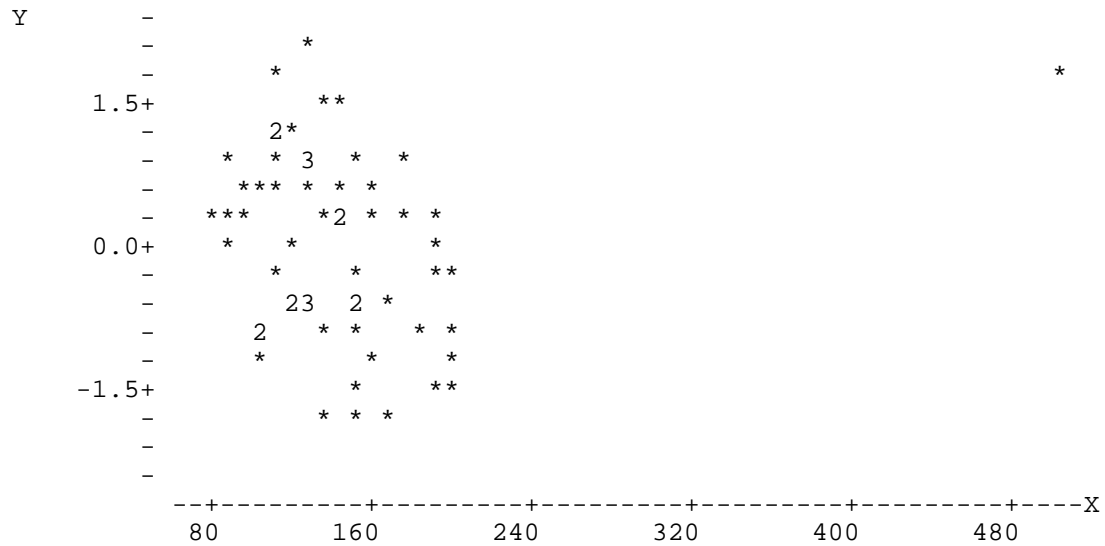
b) In the following residual plot, state which of the standard assumptions of the regression model appears to be violated, and explain what aspect of the plot led you to your conclusion.

MTB > plot c3 c1



c) Suppose created a scatter plot of Y versus X and got the following plot. What would your concern be? If you were fitting a regression of Y versus X, what diagnostic statistics would you ask the computer to calculate that would bear on your concern? {Give the formulas that the computer would be using, and say what the use of each statistic is.} What else might you consider doing?

MTB > plot c4 c2



Chapter 14, part 2

1) The following problem uses data on requests and appropriations for research and development from 1953 to 1973 for the US Army. US Air Force data is also provided, but is not used in this problem. The data set is stored on your floppy disk using filename "approp." The variables included in the data set are the year, the value of the GNP deflator (a useful price index), the Army's requested level of funding (in millions of dollars) and the Army's appropriated level of funding (also in millions of dollars).

We want to relate appropriations to requests. However, from an economist's point of view, it is probably a mistake to relate *nominal* appropriations to *nominal* requests. Inflation will create an upward trend in both these variables, introducing a spurious correlation. Therefore, it is a good idea to convert nominal measures to constant dollar, or real, terms by dividing by a price index. Use the GNP deflator data to convert Army appropriations and requests to 1972 dollars. {Hint: use Minitab's LET command.}

a) Make a plot of real appropriations versus real requests. Use the NAME command so that there are some labels on the plot. What type of function (straight line, parabola, etc.) will best fit the data? Are there any obvious outliers?

b) Fit a regression relating real appropriations to real requests. In this regression:

i) Test the following hypothesis about the intercept:

$$H_0: \beta_0 = 0$$

$$H_A: \beta_0 \neq 0$$

using a 5% significance level.

ii) Interpret the null and alternate hypotheses.

iii) Compute a 98% Confidence Interval for β_1 , the slope coefficient.

c) Use the PREDICT subcommand to calculate:

i) A 95% Confidence Interval for the mean value of real appropriations when requests are equal to 1587 million dollars and the GNP deflator is at its 1973 value. What is the implied confidence interval for the mean value of *nominal* appropriations?

ii) A 95% Prediction interval for a new observation of real appropriations when requests are equal to 1587 million dollars and the GNP deflator is at its 1973 value. What is the implied prediction interval for *nominal* appropriations?

d) {Thought question.} You are a lieutenant colonel in Army research and development, and you want a secretary and a bigger office. You calculate, on the basis of your regression, that if the department requests an additional \$75,000, enough additional money will be appropriated to pay for the perks you desire. Will this scheme work? Is the regression evidence that it will work?

2) This is a continuation of the NAEP computer case in the textbook addressing the effect of school spending on academic achievement. Our point of departure from the text is the observation that some states have a much higher cost of living than others, so spending \$7000 per student means something different in New York than it does in Alabama. We are going to adjust the nominal expenditure to reflect the cost of living. Use the data set NAEP2 downloaded from the web page. As in problem 1, we get an index of real expenditure by dividing nominal expenditure by the cost of living index. Then regress the dependent variable on this measure of real spending. Does it seem to work better than the regression on nominal expenditure recommended by the book? Explain why or why not.

Chapter 15

1) Consider the following computer output, which we will assume was generated by Minitab:

Predictor	Coef	Stdev	t-ratio	p
Constant	450.3	748.5	0.60	0.551
X1	635.4	682.3	0.93	0.357
X2	129.1	67.86	1.90	0.064
X3	-5.85	5.13	-1.14	0.261
X4	-9.37	5.96	-1.57	0.123

Where the model is taken to be:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

For each of the following null and alternate hypotheses, give the correct p-value.

- a. $H_0: \beta_1 \leq 0$
 $H_A: \beta_1 > 0$
- b. $H_0: \beta_2 = 0$
 $H_A: \beta_2 \neq 0$
- c. $H_0: \beta_3 \leq 0$
 $H_A: \beta_3 > 0$
- d. $H_0: \beta_4 \geq 0$
 $H_A: \beta_4 < 0$
- e. In parts b, c, and d, explain exactly what the p-value means by completing the following sentence: "This p-value is the probability that . . ."

2) A multiple regression of lung capacity Y (in milliliters), as a function of age (in years), height (in inches), amount of smoking (in packs per day) and sex was computed for a sample of 25 randomly selected individuals from a large population of workers.

Suppose your computer printout gave you the following results:

Variable	Estimated Coefficient	Standard Error	t-ratio
Age	-39		-4.29
Height	98.4		3.08
Smoking	-180		-1.38
Sex	23		1.78

- a. Fill in the missing column in the table.
- b. If sex is a dummy variable equal to one for males, test the hypothesis that the sexes have equal lung capacity against the alternative that males have greater lung

capacity, holding age, height, and smoking constant, at the 10% level of significance.

- c. As far as lung capacity is concerned, holding other variables constant, the effect of smoking 1 pack per day is equivalent to aging how many years?
- d. Why is your answer to part c only an approximate answer?

3) The data for this problem is on your data diskettes, stored in a file named Coleman. The data is from a sociological study that explored the relationship between student performance in school and various measures of the socioeconomic status of the school district as well as teacher characteristics. The variables in the data set are:

y verbal mean test score (all sixth graders),

x_1 staff salaries per pupil,

x_2 6th grade percent white-collar fathers

x_3 socioeconomic status composite deviation: 6th grade means for, family size, family intactness, father's education, mother's education, per cent white-collar fathers, and home items,

x_4 mean teacher's verbal test score,

x_5 6th grade mean mother's education level (1 unit = 2 school years).

- a) Fit a regression explaining y with x_2 and x_4 , that is:

$$y = \beta_0 + \beta_1 x_2 + \beta_2 x_4$$

- i) How would you interpret β_1 ? (Be precise!)
- ii) Test the following hypothesis using $\alpha = .05$:

$$H_0: \beta_1 \leq 0$$

$$H_A: \beta_1 > 0$$

Do you reject the null hypothesis? What do you conclude?

- b) Fit a regression explaining y with x_4 and x_5 , that is:

$$y = \beta_0 + \beta_1 x_4 + \beta_2 x_5$$

- i) How would you interpret β_2 ? (Be precise!)
- ii) Test the following hypothesis using $\alpha = .05$:

$$H_0: \beta_2 \leq 0$$

$$H_A: \beta_2 > 0$$

Do you reject the null hypothesis? What do you conclude?

- c) Fit a regression explaining y with x_2 , x_4 and x_5 , that is:

$$y = \beta_0 + \beta_1 x_2 + \beta_2 x_4 + \beta_3 x_5$$

- i) Compare the coefficients of like variables with earlier regressions. Are they the same?
- ii) Test the following hypothesis using $\alpha = .05$:

$$H_0: \beta_1 \leq 0$$

$$H_A: \beta_1 > 0$$

Do you reject the null hypothesis? What do you conclude?

iii) Test the following hypothesis using $\alpha = .05$:

$$H_0: \beta_3 \leq 0$$

$$H_A: \beta_3 > 0$$

Do you reject the null hypothesis? What do you conclude?

iv) Are your answers in c, parts (ii) and (iii) consistent with the answers in parts a and b? Compare the standard error of the estimated coefficient of x_2 before and after adding x_5 . Compare the standard error of the estimated coefficient of x_5 before and after adding x_2 . Does adding the other variable make these bigger or smaller? Why? What is going on?

d) Fit a regression explaining y with x_3 and x_4 , that is:

$$y = \beta_0 + \beta_1 x_3 + \beta_2 x_4$$

i) Construct a 99% confidence interval for β_1 .

ii) Construct a 99% confidence interval for β_2 .

e) Now fit a regression explaining y with x_1 , x_3 and x_4 , that is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_3 + \beta_3 x_4$$

Should the variable x_1 stay in the regression? Why or why not?

- 4) Using election data, investigators made a study of the various factors influencing voting behavior. They estimate that the issue of inflation contributed about 7 percentage points to the Republican vote in a certain election. However, the standard error for this estimate is about 5 percentage points. Therefore, the increase is not statistically significant. The investigators conclude that “in fact, and contrary to widely held views, inflation has no impact on voting behavior.” Does the conclusion follow from the statistical test? Answer yes or no, and explain briefly.²

² Taken from Freedman, Pisani, Purves, and Adhikari, *Statistics, second edition*. The question is based on Arcelus and Meltzer, "The Effect of aggregate Economic variables on congressional voting," *American Political Science Review*, vol 69 (1965), pp. 1232-69.

Chapter 16

1) Using the Coleman report data set, perform the following regression:

$$Y = \beta_0 + \beta_1 X_2 + \beta_2 X_4 + \beta_3 X_5$$

and do F tests to test the following hypotheses, using $\alpha = .05$.

- a) $H_0: \beta_1 = \beta_2 = \beta_3 = 0$
 $H_A: \text{At least one not zero}$
- b) $H_0: \beta_1 = \beta_3 = 0$
 $H_A: \text{At least one not zero}$

(Hint: part b will require regressing Y on X_4 .)

c) What is surprising about your finding in part (b) ?

(Hint: relate your finding in part b to part c of the previous question.)

2) The data set resgas (located on your floppy disk) is data on the residential use of natural gas by year and quarter of the year, and it can be used to illustrate the use of dummy variables for seasonal adjustment.

a) Retrieve the data set and make a scatter plot of gas use against number of customers. Then make a scatter plot of the log of gas use against the log of the number of customers. Which looks like it will give a better fit when we define seasonal dummies? Why?

If you answered in part a that the logs will work best, use the log of gas and the log of customers to do the rest of the problem. If you answered in part a that the raw variables would work best, use gas and customers without a transformation to do the rest of the problem. Begin by defining dummy variables for the quarter of the year. Then regress your gas use variable on your number of customers variable and the dummy variables. (Hint: done correctly this will require three dummy variables.)

b) If G is your gas use variable and C is your number of customers variable, you should have fit something like:

$$G = \beta_0 + \beta_1 C + \beta_2 Q_1 + \beta_3 Q_2 + \beta_4 Q_3.$$

Use an F test to test the statistical significance of seasonality, i.e.,

$$H_0: \beta_2 = \beta_3 = \beta_4 = 0$$

$$H_A: \text{At least one not zero}$$

- c) Interpret the coefficients of the dummy variables. What are they estimates of?
 d) Interpret the coefficient of customers. What is it an estimate of?
 e) Regress (if you have not already)

$$G = \beta_0^* + \beta_1^* C.$$

- i) Compute a 95% Confidence Interval for β_1^*
 ii) Compute a 95% Confidence Interval for β_1 in part b.

iii) Compare the two. Has including seasonality substantially improved the precision of your estimate of β_1^* obtained from the simple regression?

3) Suppose you were interested in trying to predict attendance at major league ball games, and you gathered the following data: citypop is estimated population of the city in 1978 (in thousands), metpop is estimated population of the metropolitan area in 1978 (in thousands), stadcap is the seating capacity of the baseball stadium, 78win% is the percentage of games won in 1978 (eg., if the team wins half its games, this variable is .500), 77win% is the percentage of games won in 1977, attend is the 1978 attendance, compete is whether or not there is another major league team in the same city {1=Yes, 0=No}, and division is the league and division of each team {0=American League West, 1=National League East, 2=National League West, 3=American League East}. This data is stored in a data set named baseball on your floppy disk.

Use this data to find the best model for predicting attendance, as if you were acting as a consultant to a baseball team. I especially want you to think about which variables belong in the regression and which don't, and why. You should also define dummy variables, do appropriate t and F-tests, and - once you've settled on a specification - interpret the magnitudes of the estimated coefficients.

4) In the *Minitab Handbook*, there is a continuing problem involving estimating the volume of trees from measurements of their heights and diameters, using the data set known as trees. Using the "trees" data set, estimate the relationship in logs; that is, regress the log of the volume of the tree on both the log of its height and the log of its diameter.

a) The most natural null hypotheses to test for this problem are the null that $\beta_{ht} = 1$, versus the alternative that $\beta_{ht} \neq 1$, and the null that $\beta_{diam} = 2$, versus the alternative that $\beta_{diam} \neq 2$. Explain the theory that would lead to these null and alternate hypotheses. {This requires some general knowledge from high school geometry.}

b) Test these two hypotheses using $\alpha = .10$. Be sure to state the calculated and critical values of the test statistic, and whether you accept or reject the null hypothesis.

c) Now fit a simple regression explaining log volume with log diameter. Compare these results with those obtained in the multiple regression. Note that the coefficient of log diameter has changed. What is going on? *A correct explanation will also explain why the coefficient in the simple regression is LARGER than in the multiple regression.*

5) A group of 4 physicians hired a management consultant to see whether he could reduce the long waiting times of their patients. He randomly selected 200 patients, and found their waiting times had an average of 32 minutes, and a standard deviation of 15 minutes. To determine the factors that influence waiting time, he ran a multiple regression:

$$\text{WAIT} = 22 + .09 \text{ DRLATE} - .24 \text{ PALATE} + 2.61 \text{ SHORT}$$

Where WAIT = waiting time, in minutes, DRLATE = the lateness of the doctors in arriving that morning (sum of their times, in minutes.), PALATE = the lateness of the patient in arriving for his appointment (in minutes), and SHORT = 1 if the clinic was short staffed, and some of the appointments had to be rebooked; 0 if fully staffed with all 4 physicians.

Variable	Coef	StDev	$R^2 = .72$
Constant	22.0	14.0	
DRLATE	.090	.010	
PALATE	-.240	.050	
SHORT	2.61	.820	

Analysis of Variance	Df.	SS
Regression	3	32238
Error	196	12537
Total	199	44775

Answer True or False; if false, correct it:

- a) Since the coefficient of SHORT is biggest, it is the most important factor in accounting for the variation in WAIT.
- b) If two of the doctors were late that morning (by 20 minutes and 40 minutes), the expected increase in waiting time for a patient that day would be 2.7 minutes.
- c) If the consultant had studied all the patients, he would have found, with 95% confidence, that:
 - i) Their average waiting time would be somewhere between 2 and 62 minutes.
 - ii) The regression coefficient of PALATE would be somewhere between -.12 and -.36.
- d) Since patients who are late are likely to wait longer, the office staff is providing a strong incentive for patients to arrive on time.
- e) If he included another factor in the multiple regression, R^2 would necessarily be larger, as would the adjusted R^2 .
- f) Suppose the researcher wants to include the identity of the patient's doctor as an explanatory variable in the regression. There are 4 doctors in the clinic - lets call them Fred, Wilma, Betty, and Barney. {4 names chosen entirely at random.} Each patient's data includes a physician name. Suppose the first 4 patients in the sample belong to Betty, Barney, Wilma, and Fred respectively.
 - i) Explain how you would input "physician identity" as an explanatory variable: Illustrate by defining appropriate variables and showing how, for these first four patients, you'd input their data to Minitab.

ii) Suppose, after including the "physician identity" variable, you repeated the regression, leaving all the other variables in the model as well as the physician identity variables. As a result, you get the following analysis of variance table. Test the hypothesis that the physician's identity is a determinant of waiting time, against the null that it has no effect, using $\alpha = .01$.

Analysis of Variance	Df.	SS
Regression	6	38033
Error	193	6742
Total	199	44775