HW05 - More Regression

Stat 20 & 131A, Spring 2017, Prof. Sanchez

Due Feb-23

1) For the first-year students at a certain university, the correlation between SAT scores and first-year GPA was 0.60. The scatter diagram is football-shaped. Please show your work (No work, No credit) to predict the percentile rank on the first-year GPA for a student whose percentile rank on the SAT was: $_{1pt}$

a. 90%

b. 30%

c. 50%

d. unknown

2) Pearson and Lee obtained the following results in a study of about 1,000 families: $_{1pt}$

- average height of husband $\approx$ 68 inches, SD $\approx$ 2.7 inches
- average height of wife $\approx$ 63 inches, SD $\approx$ 2.5 inches
- $r \approx 0.25$

Please show your work (No work, No credit). Predict the height of a wife when the height of her husband is:

a. 72 inches.

b. 64 inches.

c. 68 inches.

d. unknown.

3) Refer to the previous exercise. Again, please show your work (No work, No credit). $_{0.5pts}$

- average height of husband $\approx$ 68 inches, SD $\approx$ 2.7 inches
- average height of wife $\approx$ 63 inches, SD $\approx$ 2.5 inches
- $r \approx 0.25$

a. What percentage of the women were over 5 feet 8 inches?

b. Of the women who were married to men of height 6 feet, what percentage were over 5 feet 8 inches?

4) In one study, the correlation between the educational level of husbands and wives in a certain town was about 0.50; both averaged 12 years of schooling completed, with an SD of 3 years. $_{0.7pts}$

a. Predict the educational level of a woman whose husband has completed 18 years of schooling.

b. Predict the educational level of a man whose wife has completed 15 years of schooling.

c. Apparently, well-educated men mary women who are less well educated than themselves. But the women mary men with even less education. How is this possible?

5) A doctor is in the habit of measuring blood pressure twice. She notices that patients who are unusually high on the first reading tend to have somewhat lower second readings. She concludes that patients are more relaxed on the second reading. A colleague disagrees, pointing out that the patients who are unusually lon on the first reading tend to have somewhat higher second readings suggesing they get more nervous. Which doctor is right? Or perhaps both are wrong? Explain briefly. $_{0.4pts}$

6) For men age 18-24 in the HANES5 sample, the regression equation for predicting height from weight is:

predicted height = (0.0267 inches per pound) x (weight) + 62.5 inches

(Height is measured in inches and weight in pounds.) If someone puts on 20 pounds, will he get taller by:

20 pounds $\times$ 0.0267 inches per pound $\approx$ 0.5 inches?

If not, what does the slope mean? $_{0.4pts}$

7) A study is made of working couples. The regression equation for predicting wife's income from husband's income is:

wife's income = 0.1667 $\times$ husband's income + \$24,000

Another investigator solves this equation for husband's income, and gets:

husband's income = 6 $\times$ wife's income - \$144,000

True or false, and explain: the second investigator has found the regerssion equation for predicting husband's income from wife's income. If you want to compute anything, here's some data: $_{0.4pts}$

- husband's average income = \$54,000. SD = \$39,000
- wife's average income = \$33,000. SD = \$26,000
- r = 0.25

8) A law school finds the following relationship between LSAT scores and first-year scores: $_{0.5pts}$

- average LSAT score = 165, SD = 5
- average first-year score = 65, SD = 10
- $r \approx 0.6$

a. The admissions officer uses the regression line to predict first-year scores from LSAT scores. What is the r.m.s. error of the line?

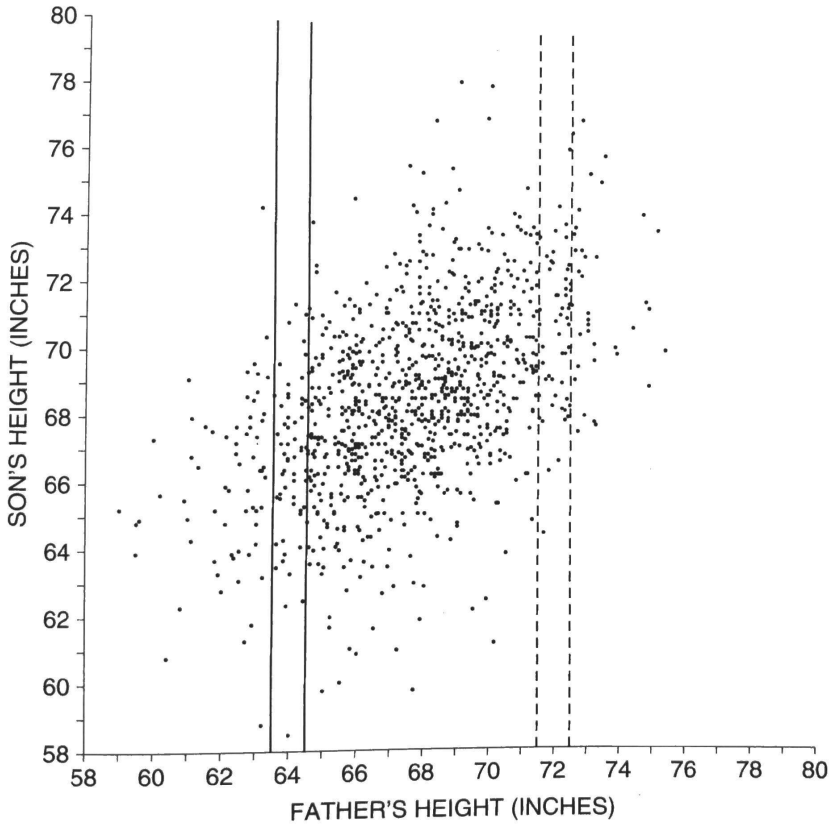
b. One of the students is chosen at random; you have to guess his first-year score, without being told his LSAT score. How would you do this?

c. Your r.m.s. error would be _____.

d. Repeat parts b) and c), if you are allowed to use his LSAT score.

9) The data in the figure below can be summarized as: $_{0.5pts}$

- average height of fathers $\approx$ 68 inches, SD $\approx$ 2.7 inches
- average height of sons $\approx$ 69 inches, SD $\approx$ 2.7 inches
- $r \approx 0.5$



a. Find the r.m.s. error of the regression line for predicting son's height from father's height.

b. If a father is 72 inches tall, predict his son's height.

c. This prediction is likely to be off by _____ inches or so. If more information is needed, say what is, and why.

d. Repeat parts b) and c), if the father is 66 inches tall.

10) A statistical analysis was made of the midterm and final scores in a large course, with the following results:

- average midterm score $\approx$ 50, SD $\approx$ 25
- average final score $\approx$ 55, SD $\approx$ 15
- $r \approx 0.60$

The scatter diagram was football-shaped. For each student, the final score was predicted from the midterm score using the regression line. $_{0.6pts}$

a. For about 1/3 of the students, the prediction for the final score was off by more than _____.

b. Predict the final score for a student whose midterm score was 80.

c. By how many points this prediction is likely to be off?.

11) The Centers for Disease Preventaion and Control (CDC) Behavioral Risk Factor Surveillance System (BRFSS) collects data related to health conditions and risk behaviors. Aggregated data by state are in the file `vegetables-smoking.csv` (in the github repository). *2.8pts*

Here's the code to import the data set in R:

```
# assembling the URL of the CSV file
# (otherwise it won't fit within the margins of this document)
repo = 'https://raw.githubusercontent.com/ucb-introstat/introstat-spring-2017/'
datafile = 'master/data/vegetables-smoking.csv'
url = paste0(repo, datafile)

# read in data set
dat = read.csv(url)
```

The data set contains two variables. `vegetables` is the percent of adults in the state who report eating at least five servings of fruits and vegetables per day; `smoking` is the percent who smoke every day.

a. Plot a scatter diagram of `vegetables` (in the x-axis) and `smoking` (in the y-axis).

b. Calculate the average and SD for `vegetables`.

c. Calculate the average and SD for `smoking`.

d. Use the `cor()` function to find the correlation coefficient r between the two variables `vegetables` and `smoking`.

e. Use the results of parts b), c), and d) to obtain the slope of the regression line using the formula: $slope = r \times (SD_y/SD_x)$. What is the interpretation of this value?

f. Use the `lm()` function to obtain the coefficients of the regression line (regressing `smoking` onto `vegetables`). Compare the value of the slope provided by `lm()` with your computed value in part e).

g. Create a new scatterplot like the one in part a), but now add the regression line obtained with `lm()`, via the `abline()` function, to the graph.

h. Create a plot of residuals. One way to do this is by passing the `"lm"` object to the `plot()` function, and specifying the argument `which = 1`. Does the plot show homoscedasticity?

i. Use the regression method to predict the percentage of adults who smoke every day when the percentage of adults who consume fruits and vegetables in a state is 18%. (Don't use the `predict()` function.)

12) In the long run, the price of a company's stock ought to parallel changes in the company's earnings. The following table gives data on the annual growth rates in the earnings and in stock prices (both in percent) for major industry groups. $_{1.2pts}$

industry	earnings	price
auto	3.30	2.90
banks	8.60	6.50
chemicals	6.60	3.10
computers	10.20	5.30
drugs	11.30	10.00
electrical equipment	8.50	8.20
food	7.60	6.50
household products	9.70	10.10
machinery	5.10	4.70
oil domestic	7.40	7.30
oil international	7.70	7.70
oil equipment	10.10	10.80
railroad	6.60	6.60
retail food	6.90	6.90
department stores	10.10	9.50
soft drinks	12.70	12.00
steel	-1.00	-1.60
tobacco	12.30	11.70
utilities electric	2.80	1.40
utilities gas	5.20	6.20

Here's the code to import the data set in R:

```
# assembling the URL of the CSV file
# (otherwise it won't fit within the margins of this document)
repo = 'https://raw.githubusercontent.com/ucb-introstat/introstat-spring-2017/'
datafile = 'master/data/stock-earnings-prices.csv'
url = paste0(repo, datafile)

# read in data set
dat = read.csv(url)
```

a. Make a scatter diagram showing how earnings growth (`earnings`) explains growth in stock price (`prices`). Does it a ppear to be true that (on the average in the long run) stock price growth parallels earnings growth?

b. Use `lm()` to run a regression analysis (`price` explained by `earnings`). What is the obtained regression line?

c. Apply `summary()` on the `"lm"` object. What percent of the variation in stock price growth among industry groups can be explained by the linear relationship with earnings growth?

d. Use the `residuals` from the `"lm"` object to calculate the r.m.s. error for the regression line. How do you interpret this value?

e. What is the correlation between earnings growth and price growth?

f. If we had data on all of the individual companies in these 20 industries, would the correlation be higher or lower? Why?