HW04 - Regression

Stat 131A, Fall 2018

Due Sep-24

General Instructions

- This assignment assumes that you have read the tutorials `11-regression-line.pdf` and `12-regression-lm.pdf` available in the course's github repository.
- Write your narrative and code in an `Rmd` (R markdown) file.
- Name this file as `hw04-first-last.Rmd`, where `first` and `last` are your first and last names (e.g. `hw04-gaston-sanchez.Rmd`).
- Please do not use code chunk options such as: `echo = FALSE`, `eval = FALSE`, `results = 'hide'`. All chunks must be visible and evaluated.
- Submit your Rmd and html files to bCourses.

---

1) The Centers for Disease Prevention 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).

Here's the code to import the data set in R (do NOT include this in your `.Rmd`). You will need to save the `.csv` file in the same directory (i.e. folder) containing the `.Rmd` file of this assignment.

```
# get a copy of the data file but do NOT include this code in your Rmd
repo = 'https://raw.githubusercontent.com/ucb-introstat/introstat-fall-2018/'
datafile = 'master/data/vegetables-smoking.csv'
url = paste0(repo, datafile)

# download csv to your working directory
download.file(url = url, destfile = 'vegetables-smoking.csv')
```

Assuming that the file `vegetables-smoking.csv` is in the same location of your `.Rmd` file, you can include the following command—inside a code chunk—in order to read in the data set:

```
# (include this in your Rmd)
# read in data set
dat = read.csv(file = 'vegetables-smoking.csv')
```

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 results of parts b), c), d), and e) to obtain the y-intercept of the regression line. What is the interpretation of this value?

g. 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 parts e) and f).

h. 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.

i. Using the regression line $\hat{y} = ax + b$, calculate a vector `e` of residuals $e = y - \hat{y}$; and display its `summary()` values.

j. Create a plot of residuals. This is a scatterplot of points (x_i, e_i) pairs where x_i are the vegetable values, and e_i are the residuals. Does the plot show homoscedasticity?

k. 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.)

HW continues on the next page

2) 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.

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

```
# get a copy of the data file but do NOT include this code in your Rmd
repo = 'https://raw.githubusercontent.com/ucb-introstat/introstat-fall-2018/'
datafile = 'master/data/stock-earnings-prices.csv'
url = paste0(repo, datafile)

# download csv to your working directory
download.file(url = url, destfile = 'stock-earnings-prices.csv')
```

Assuming that the file `stock-earnings-prices.csv` is in the same location of your `.Rmd` files, you can include the following command—inside a code chunk—in order to read in the data set:

```
# read in data set
dat = read.csv(file = 'stock-earnings-prices.csv')
```

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

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

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

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

e. What is the interpretation of the slope value?

f. 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?

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. Using the regression line $\hat{y} = ax + b$, calculate a vector `e` of residuals $e = y - \hat{y}$; and display its `summary()` values.

i. Create a plot of residuals. This is a scatterplot of points (x_i, e_i) pairs where x_i are the earnings values, and e_i are the residuals. Does the plot show homoscedasticity?

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

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

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