Assignment 3
Due **Tuesday, February 14, 2017**

Total 75 points.

Please submit a well-commented R script, documenting all code used in this problem set, along with a write up answering all questions below. Use figures as appropriate to support your answers, and when required by the problem. Some questions modified from E. Spiro.

1) For each of the following situations, state whether the parameter of interest is a mean or a proportion. It may be helpful to examine whether individual responses are numerical or categorical.

> a) (1 pt) In a survey, one hundred college students are asked how many hours per week they spend on the Internet.
>
> b) (1 pt) In a survey, one hundred college students are asked: “What percentage of the time you spend on the Internet is part of your course work?”
>
> c) (1 pt) In a survey, one hundred college students are asked whether or not they cited information from Wikipedia in their papers.
>
> d) (1 pt) In a survey, one hundred college students are asked what percentage of their total weekly spending is on alcoholic beverages.
>
> e) (1 pt) In a sample of one hundred recent college graduates, it is found that 85 percent expect to get a job within one year of their graduation date.

2) (5 pt) In 2013, the Pew Research Foundation reported that “45% of U.S. adults report that they live with one or more chronic conditions”. However, this value was based on a sample, so it may not be a perfect estimate for the population parameter of interest on its own. The study reported a standard error of about 1.2%, and a normal model may reasonably be used in this setting. Create a 95% confidence interval for the proportion of U.S. adults who live with one or more chronic conditions. Also interpret the confidence interval in the context of the study.

3) The nutrition label on a bag of potato chips says that a one ounce (28 gram) serving of potato chips has 130 calories and contains ten grams of fat, with three grams of saturated fat. A random sample of 35 bags yielded a sample mean of 136 calories with a standard deviation of 17 calories.

> a) (4 pt) Write down the null and alternative hypotheses for a two-sided test of whether the nutrition label is lying.
>
> b) (4 pt) Calculate the test statistic and find the p value.
>
> c) (2 pt) If you were the potato chip company would you rather have your alpha = 0.05 or 0.025 in this case? Why?

4) Regression was originally used by Francis Galton to study the relationship between parents and children. He wondered if he could predict a man's height based on the height of his father? This is the question we will explore in this problem. You can obtain data similar to that used by Galton as follows:

```
library(UsingR)
height <- get("father.son")
```

a) (5 pt) Perform an exploratory analysis of the father and son heights. What does the relationship look like? Would a linear model be appropriate here?

b) (5 pt) Use the Im function in R to fit a simple linear regression model to predict son's height as a function of father's height. Write down the model,

$$y_{sheight} = \beta_0 + \beta_1 \text{ x fheight}$$

filling in estimated coefficient values and interpret the coefficient estimates.

c) (5 pt) Find the 95% confidence intervals for the estimates. You may find the `confint()` command useful.

d) (5 pt) Produce a visualization of the data and the least squares regression line.

e) (5 pt) Produce a visualization of the residuals versus the fitted values. (You can inspect the elements of the linear model object in R using `names()`). Discuss what you see. Do you have any concerns about the linear model?

f) (5 pt) Using the model you fit in part (b) predict the height was 5 males whose father are 50, 55, 70, 75, and 90 inches respectively. You may find the `predict()` function helpful.

g) (5 pt) What do the estimates of the slope and height mean? Are the results statistically significant? Are they practically significant?

5) An investigator is interested in understanding the relationship, if any, between the analytical skills of young gifted children and the father's IQ, the mother's IQ, and hours of educational TV. The data are here:

```
library(openintro)
data(gifted)
```

a) (5 pt) Run two regressions: one with the child's analytical skills test score ("score") and the father's IQ ("fatheriq") and the child's score and the mother's IQ score ("motheriq").

b) (5 pt) What are the estimates of the slopes for father and mother's IQ score with their 95% confidence intervals? (Note, estimates and confidence intervals are usually reported: Estimate (95% CI: CI_{lower}, CI_{upper})

c) (5 pt) How are these interpreted?

d) (5 pt) What conclusions can you draw about the association between the child's score and the mother and father's IQ?