Lab 8

No homework questions.

1. Derive least squares estimators for A and B when

$$Y_i = A + BX_i + e_i, \qquad \mathbb{E}[e_i] = 0.$$

   i.e. find A and B that minimize the sum of squares

$$S = \sum_i (Y_i - A - BX_i)^2.$$

2. In R, run the code in the `anscombe` dataset help section (run `?anscombe` in R): The four graphs are all fit by the same linear model. Is regression an appropriate tool for analyzing the four datasets?

3. Show that if X, Y are independent N(0,1) random variables, and $\rho \in [-1, 1]$ then X and $\rho X + \sqrt{1-\rho^2}Y$ are both $N(0,1)$ with correlation ρ. This is called the bivariate normal distribution. Plot a sample from the bivariate normal distribution in R.

4. Regression in R: Work through the following commands and see what happens.

```
x=rnorm(1000)                    #produce some data
y=x*2+4+rnorm(length(x),0,1)    #where y depends on x
(l=lm(y~x)) #fit a linear model Y=A+BX+e_i
plot(x,y)   #plot the data
abline(l)   #line of best fit
summary(l)  #t=tests for coefficients
plot(x,l$residuals)   #plot the residuals against x
plot(1:length(x),l$residuals[order(x)]) #plot the residuals
                                        #accoring to the order of x
qqnorm(l$residuals);qqline(l$residuals) #QQ plot

(l=lm(y~0+x))
summary(l)
plot(x,y)
abline(l)
#Now plot the residuals

(l=lm(y~1))
summary(l)
plot(x,y)
abline(l)
#Now plot the residuals

(l=lm(y~offset(2*x))) #fix the coefficient B to have value 2
summary(l)
plot(x,y)
abline(l$coefficients[1],2)

y=x^2 + 3*x+ 4=rnorm(length(x))  #generate some new data
plot(x,y)

(l=lm(y~x))  #Fit a linear model.
#Is it a good fit?
#Do a QQ-plot for the residuals, etc

(l=lm(y~x+I(x^2)))  #Fit a quadratic model.
#Is it a good fit?
```

5. The term regression comes from the phenomenon of "regression towards the mean". Here is an example:

```
#Install and load the UsingR package
install.packages("UsingR")
library(UsingR)
# Description of the dataset:
?father.son
```

Plot the father.son dataset. Fit a linear model for the sons' heights in terms of the fathers' heights. Plot the line of best fit. Do a t-test to decide which coefficients to include in your linear model.

6. Repeat exam results. Every student in a number theory class has a "skill level" which reflects their (i) affinity for number theory, (ii) there previous educational background, etc. We will assume that your "skill" is fixed over relatively short time periods. The students take a midterm exam, and an end of term exam.

```
skill=rnorm(100,60,10) # generate "skill" data
exam1=skill+rnorm(100,0,10)
exam2=skill+rnorm(100,0,10)
```

Try to predict the second set of exam results in terms of the first. Try to make the model as simple as possible (i.e. if you can set the constant term to zero, do so. If you can set the slope to a precise value, do so.)