Homework1

R functions used:

```
r2<-function(y_hat,y) {

  RSS<-sum((((y_hat))-(y))^2)

  TSS<-sum(((y)-(mean(y)))^2)

  return(1-RSS/TSS)}

rmse=function(y_hat,y)
{
  return(sqrt(mean((y-y_hat)^2)))
}
```

For questions 1 and 2, please run a linear regression on the data using `lm`

As output, include:

a) a scatterplot of the dependent and indepedent variable with a line added (using `curve`) representing best fit least squares model

b) Parameter estimates and 95% CI for slope and intercept parameters

c) A metric of model fit (calculate either R2 or RMSE using the functions above) reminder: `y_hat` is calculated using the equation for a line applied to the predictor variable

Question 1:

This question refers to the following dataset: math_scores.csv

```
math<-read.csv("math_scores.csv")
head(math)

##   LSD_concentration MATH_score
## 1              1.17      78.93
## 2              2.97      58.20
## 3              3.26      67.47
## 4              4.69      37.47
## 5              5.83      45.65
## 6              6.00      32.92
```

These data represent the average performance test scores of groups of human subjects with varying tissue concentrations of LSD (yes this is a real dataset–see citation below). Using least squares regression, address the following questions:

A. What level of LSD tissue concentration do you need to ensure a test score of >85%?

B. How well does LSD tissue concentration predict test performance?

C. Why might the normal distribution be inappropriate to model these data?

Source: Wagner, Agahajanian, and Bing (1968). Correlation of Performance Test Scores with Tissue Concentration of Lysergic Acid Diethylamide in Human Subjects. Clinical Pharmacology and Therapeutics, Vol.9 pp635-638.

Question 2:

These data refer to the following data frame: miracle_food.csv

```
miracle<-read.csv("miracle_food.csv")
head(miracle)
```

```
##   Weight_loss pomegranate
## 1       -0.89           2
## 2        6.31           2
## 3      -30.21           3
## 4       -6.28           7
## 5       11.38           4
## 6        1.67           2
```

These data are from an observational study that relates number of pomegranates eaten per day to weight gain/loss (in lbs) for 3 months. A pomegranate farmer's association has noted the significant p-value for effect of pomegranates on weight loss. The farmer's association is mounting a major ad campaign touting the miraculous weight loss effect of eating pomegranates.

With the least squares output in mind, do you agree or disagree with the miracle claim?

Question 3:

Mean Absolute Error (MAE) is an alternative to RMSE for assessing model fit. MAE uses the absolute difference between predicted and observed values, rather than the squared difference. Like RMSE, MAE represents prediction error on the scale of the original response variable. While RMSE gives higher weight to larger errors, MAE gives equal weight to all errors. Some have argued that MAE is easier to interpret.

A. Translate the mathematical equation for MAE into a function in R.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

B. Compare RMSE, R2, and MAE for the linear models in questions 1 and 2. How do these metrics of model fit differ?

Question 4:

Simulate linear data.

Step 1: Create a predictor variable using `runif` or `seq`

Step 2: Decide on a value for the intercept and slope

Step 3: Use `rnorm` to simulate draws from a normal distribution for your dataset

A. Plot your data

B. Estimate the slope and intercept parameters from the data using linear regression.

C. How do your estimates compare to the true values you came up with in step 2?

Question 5:

Unequal variance (heteroskedasticity) between subgroups of the response variable violates assumptions of least squares. Using `rnorm` and steps above in Question 5, simulate a normally-distributed random variable where the variance of the response variable increases as a function of a predictor variable.

A. Construct a plot showing your simulated data

B. What is a potential biological explanation for the data you have simulated?