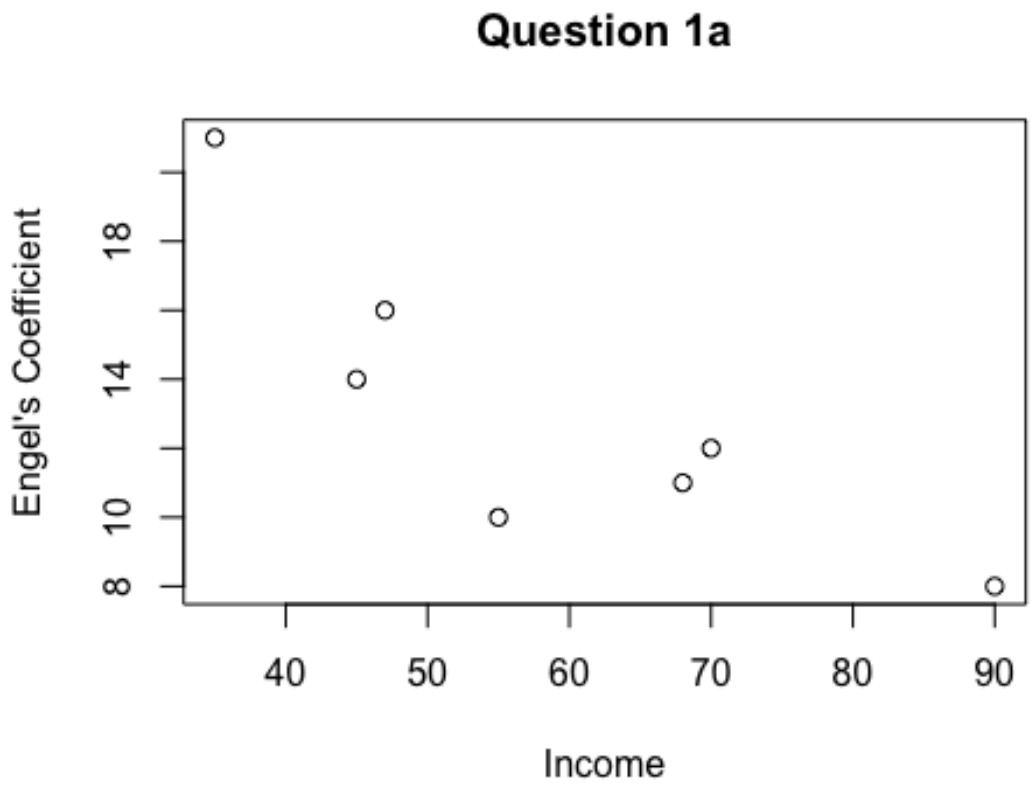
# Assignment_4

1. A researcher surveyed household income (in thousand dollars) and Engle's coefficient (percentage of income used for food) of nine housesholds in the city A

(a) Make a scatter plot (X: Income, Y : Engel's coefficient).

```
X <- c(35,90,47,45,68,70,55)
Y <- c(21,8,16,14,11,12,10)

plot(X,Y,main="Question 1a",xlab="Income",ylab="Engel's Coefficient")
```




(b) Regress Engel's coefficient Y on Income X with a linear model Y = a + bX + ε by 'lm' function.

```
Model <- lm(Y~X)
```

(c) Do the same estimation as (b) by matrix algebra in R (without using the lm function).

```
X <- matrix(c(1,1,1,1,1,1,1,35,90,47,45,68,70,55), ncol=2)
Y <- matrix(c(21,8,16,14,11,12,10), ncol=1)

Z <- solve((t(X) %*% X),diag(1,nrow=2))
```

```
print(Z %*% (t(X) %*% Y))
```

```
##           [,1]
## [1,] 24.71247
## [2,] -0.19753
```

(d) Do the same estimation as (b) by numerically minimizing sum of squared errors. Tips: The optim function finds the minimizer of a function.

```
# XY <- data.frame(x=X,y=Y)
#
# func <- function(XY)
# {
#   Z <- solve((t(XY$x.2) %*% XY$x.2),diag(1,nrow=1))
#   return(Z %*% (t(XY$x.2) %*% XY$y))
# }
#
# func(XY)
#
#
# optim(par = XY, fn=func, method="BFGS")
```

2. (Exercise 2 in the Chapter 6 slides.) Using the iris data

(a) regress Sepal.Length on Sepal.Width for the first 50 observations (i.e., setosa species) by simple linear regression.

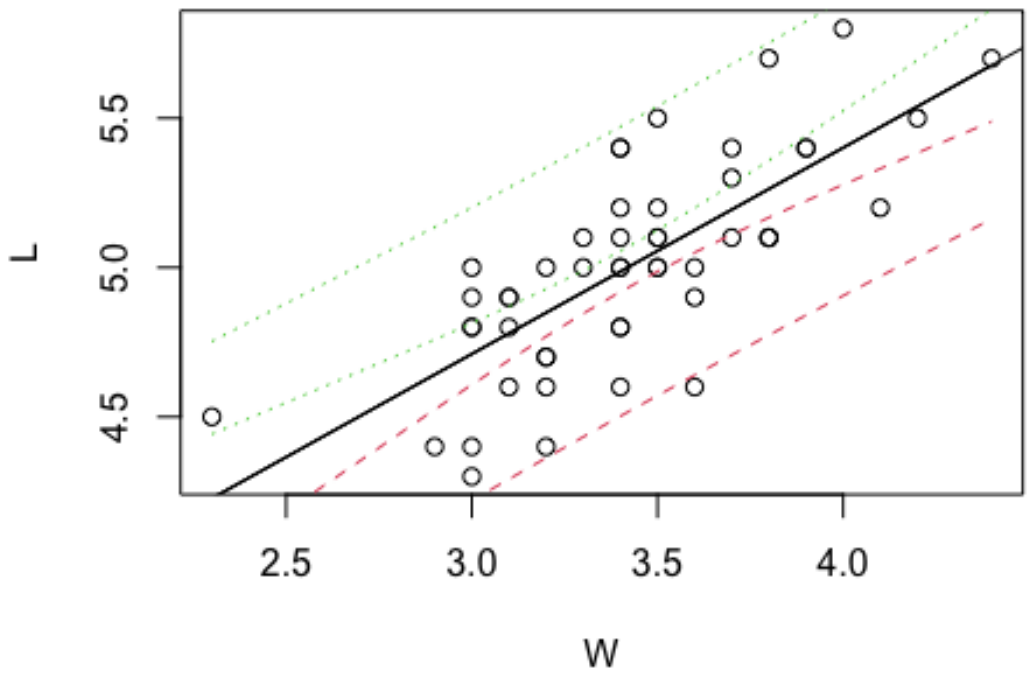
```
L <- iris$Sepal.Length[1:50]
W <- iris$Sepal.Width[1:50]

SetosaModel <- lm(L~W)
```

(b) create a scatter plot with the regression line, confidence and prediction bands

```
CI <- predict(SetosaModel, int="c")
CI <- CI[order(W),]
PI <- predict(SetosaModel, int="p")
```

```
## Warning in predict.lm(SetosaModel, int = "p"): predictions on current data
refer to _future_ responses
```

```
PI <- PI[order(W),]

plot(L~W)
abline(SetosaModel)

matlines(sort(W),CI)
matlines(sort(W),PI)
```

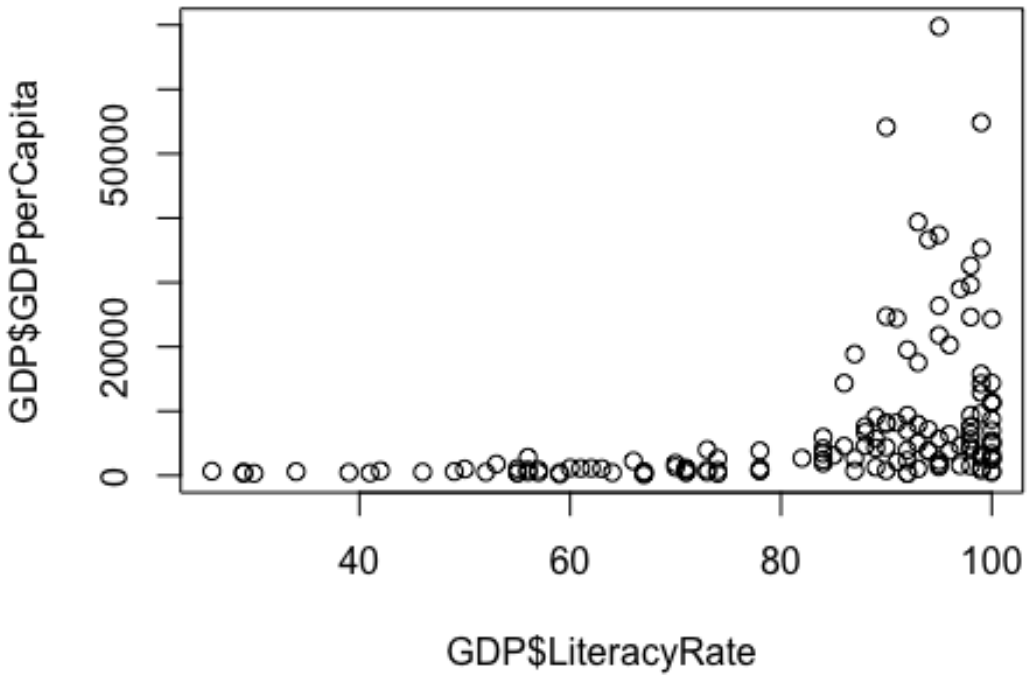



3. The attached ”GDPLiteracy.csv” file includes GDP per capita (in 2009) and literacy rate (in 2009 or latest) of 143 countries and regions

(a) Make a scatter plot for literacy rate (x-axis) and GDP per capita (y-axis). What problems do you see to apply simple linear regression?

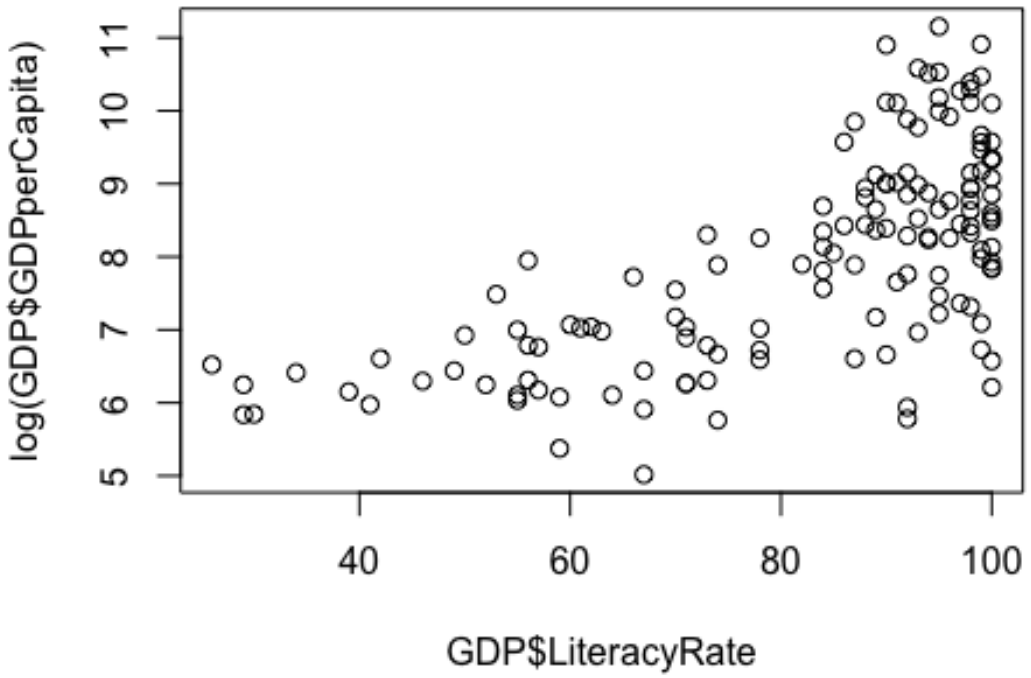
The points have a non-linear relation to each other.

```
GDP <- read.csv(file = "GDPLiteracy.csv")
plot(GDP$GDPperCapita~GDP$LiteracyRate)
```




(b) Transform GDP per capita by natural logarithm, and make a scatter plot again.

```
plot(log(GDP$GDPperCapita)~GDP$LiteracyRate)
```




(c) Regress the logarithm of GDP per capita on literacy rate. Report the intercept, slope, and their standard deviations and t-statistics. Is the slope significantly different from zero?

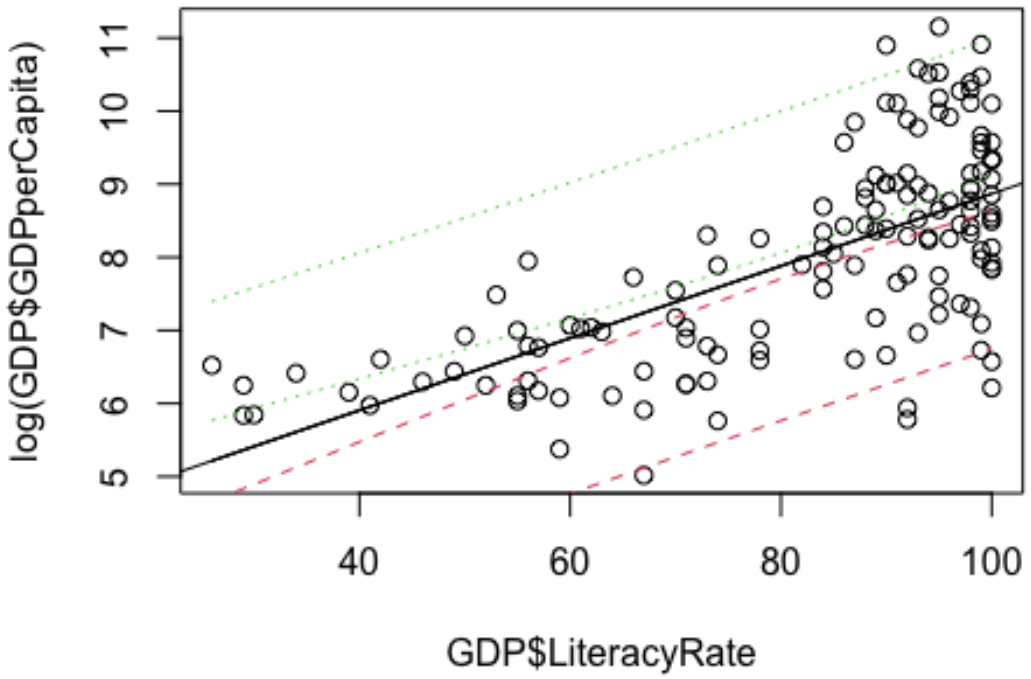
No slope is less than .05

```
GDPModel <- lm(log(GDP$GDPperCapita)~GDP$LiteracyRate)

GDPModel.Summary <- summary(GDPModel)

intercept <- GDPModel.Summary$coefficients[1,1]
print(paste("Intercept: ",intercept))

## [1] "Intercept:  3.93117072999145"

slope <- GDPModel.Summary$coefficients[2,1]
print(paste("Slope: ",slope))

## [1] "Slope:  0.0493771803186194"

SD <- GDPModel.Summary$coefficients[,2]
print(paste("Standard Deviation: ",SD))
```

```
## [1] "Standard Deviation:  0.39866758108186"
## [2] "Standard Deviation:  0.00471514749118862"

TS <- GDPModel.Summary$coefficients[,3]
print(paste("T-Statistic: ",TS))

## [1] "T-Statistic:  9.86077352796902" "T-Statistic:  10.472033040513"
```

(d) Overlay a prediction band on the scatter plot in (b).

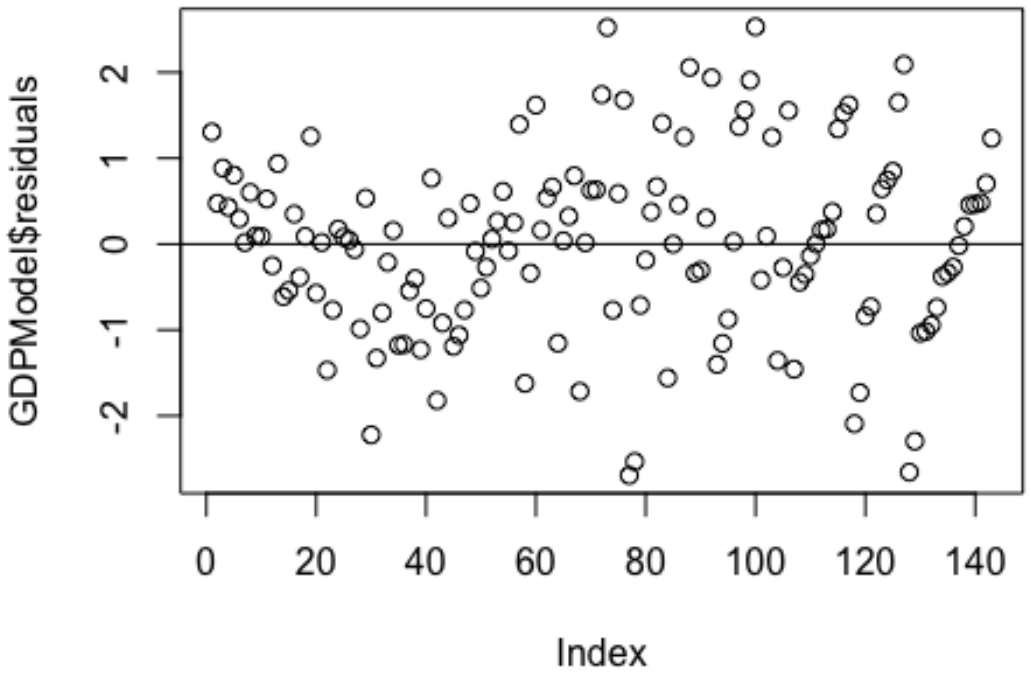
```
plot(log(GDP$GDPperCapita)~GDP$LiteracyRate)
abline(GDPModel)

CI <- predict(GDPModel, int="c")
CI <- CI[order(GDP$LiteracyRate),]
PI <- predict(GDPModel, int="p")

## Warning in predict.lm(GDPModel, int = "p"): predictions on current data
refer to _future_ responses

PI <- PI[order(GDP$LiteracyRate),]

matlines(sort(GDP$LiteracyRate),CI)
matlines(sort(GDP$LiteracyRate),PI)
```



(e) Make a residual plot with a line y = 0 (use abline function. Do you see any patterns?

The residuals are concentrated towards the the line Y = 0

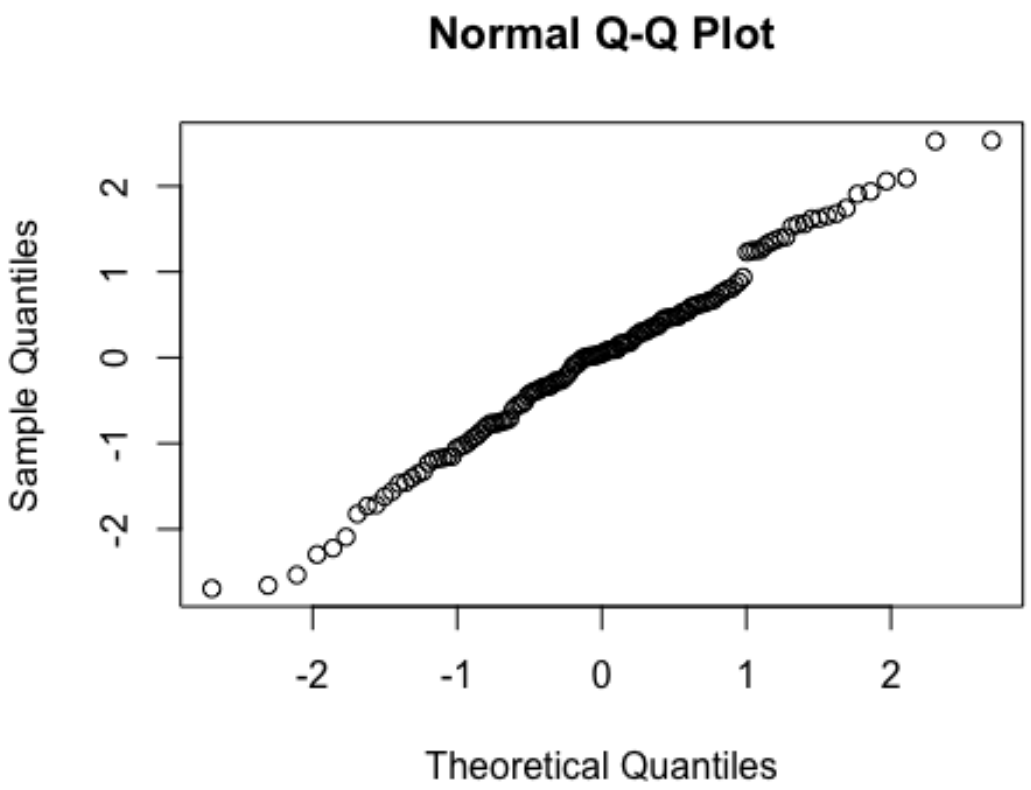
```
plot(GDPModel$residuals)
abline(a=0,b=0)
```




(f) Make a normal Q-Q plot for residuals. Residuals are approximately normal?

Yes

```
qqnorm(GDPModel$residuals)
```




(g) If we can add one more independent variable to predict GDP per capita, what variable will you add?

Country