Part 2

Kahvi Patel

Data Import and Cleaning

```
library(scales)

#read in .csv file
df <- read.csv('Data/stat123_regression.csv')
```

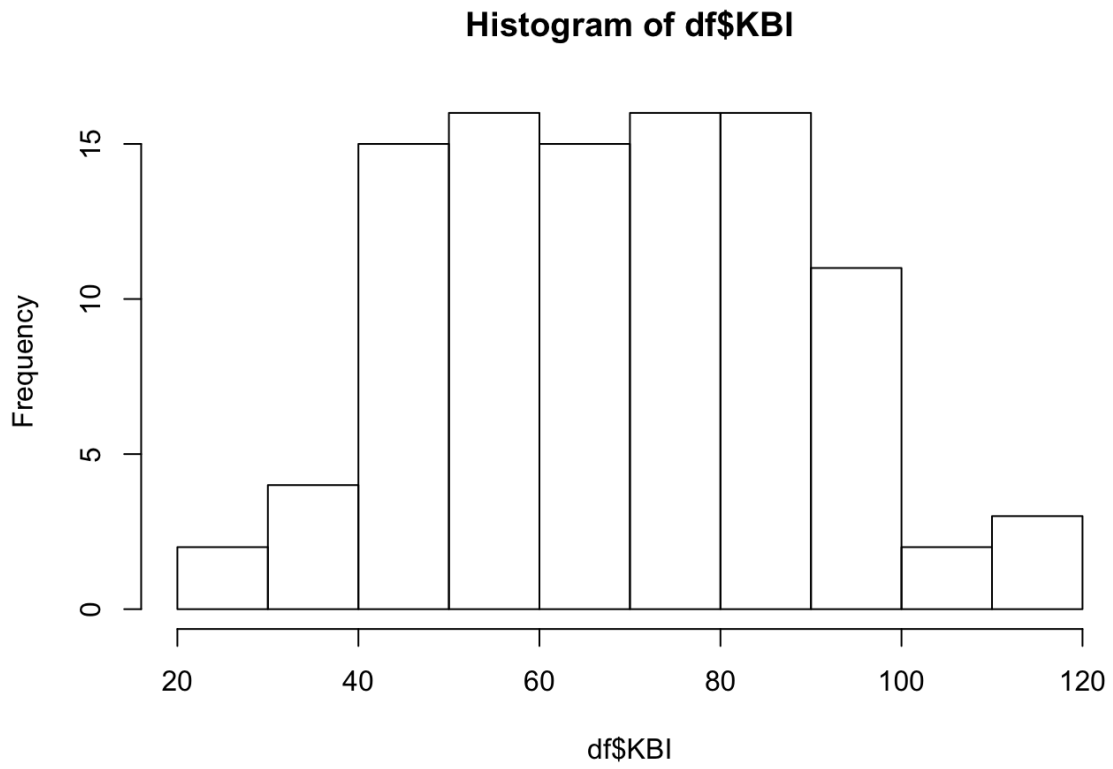
1. After reading the data into R using the read.csv function, provide summary of the data. Comment on your results and especially on any unusual features in the data.

The variables are spread in different ways. The KBI has values from 28 to 115 while the COG has values from 0 to 27. Therefore, it seems like none of the variable were scaled in the same way, or else they would have similar maxes and mins. The histograms of each variable show that: - The distribution of the KBI variable is roughly symmetric and unimodal. There doesn't appear to be any outliers. - The distribution of the ADL variable is not symmetric. - The distribution of the COG variable is not symmetric and bimodal. The distribution is also weakly skewed to the left. - The distribution of the MEM variable is unimodal and skewed to the left.

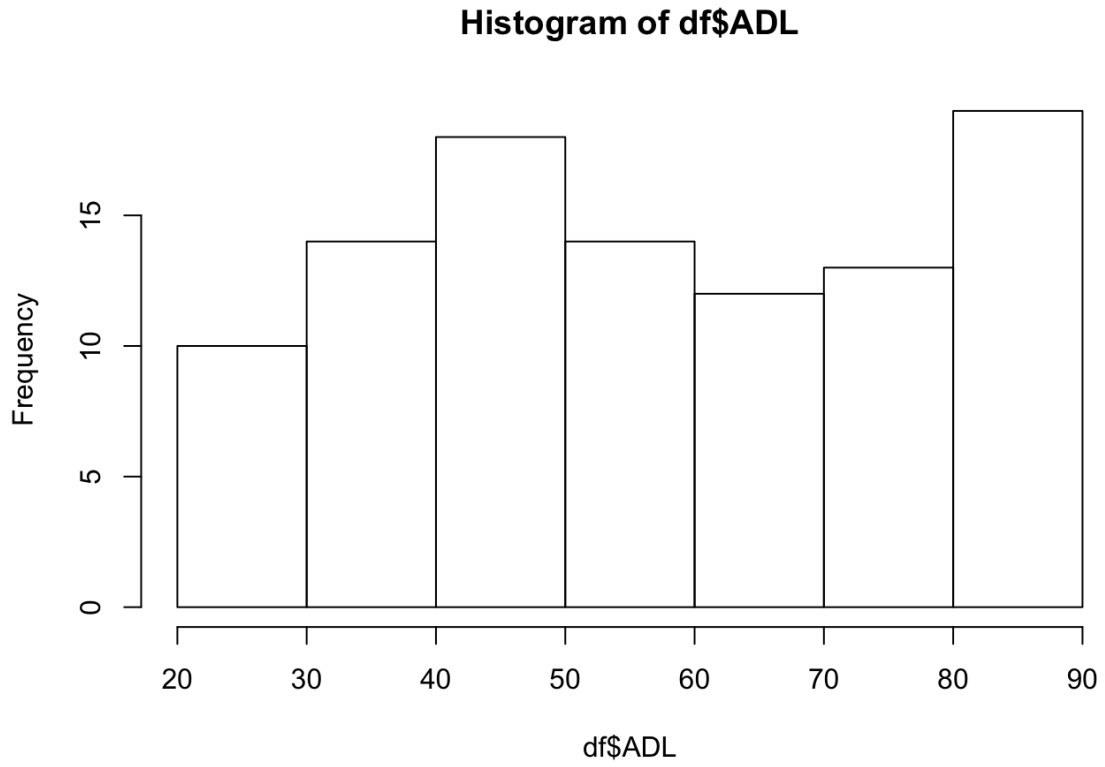
```
summary(df)
```

```
##       KBI              ADL              MEM             COG        
##  Min.   : 28.00   Min.   :22.00   Min.   : 3.0   Min.   : 0.00  
##  1st Qu.: 52.75   1st Qu.:42.00   1st Qu.:14.0   1st Qu.: 7.00  
##  Median : 69.50   Median :56.00   Median :24.0   Median :15.00  
##  Mean   : 69.24   Mean   :57.85   Mean   :26.3   Mean   :13.69  
##  3rd Qu.: 85.50   3rd Qu.:77.50   3rd Qu.:34.0   3rd Qu.:19.00  
##  Max.   :115.00   Max.   :90.00   Max.   :66.0   Max.   :27.00
```

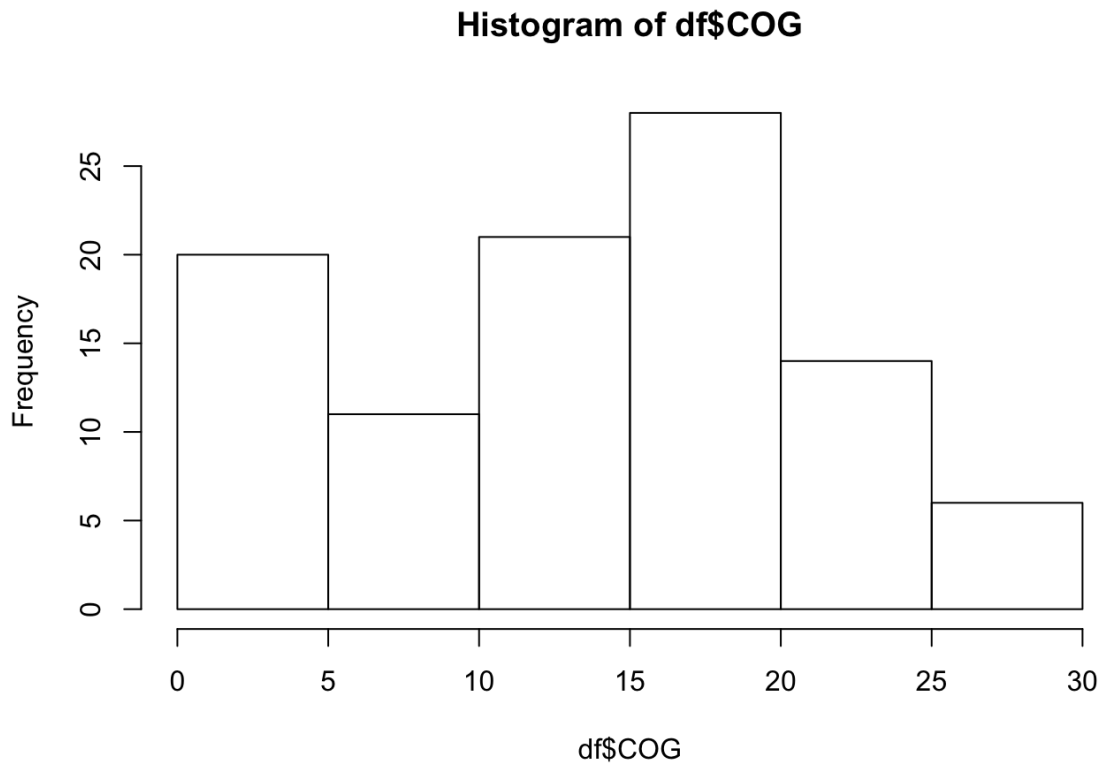
```
hist(df$KBI)
```


Histogram of df$KBI
Frequency
0
5
10
15
20
40
60
80
100
120
df$KBI

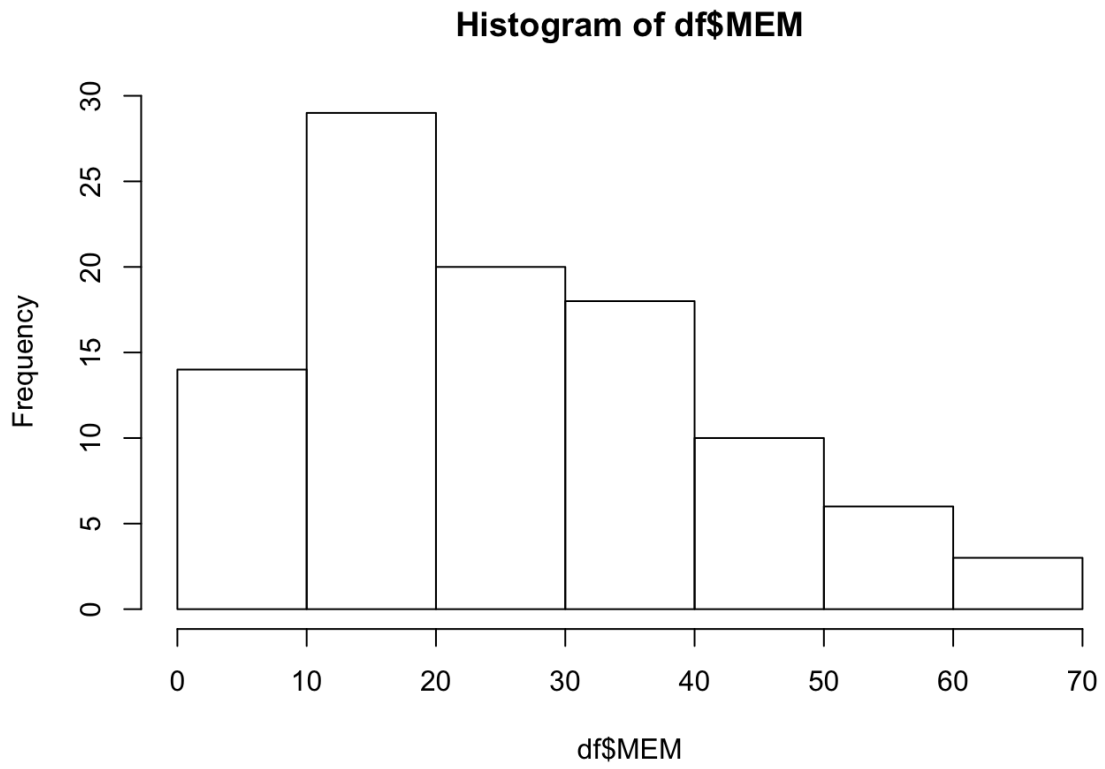
```
hist(df$ADL)
```


Histogram of df$ADL
Frequency
0
5
10
15
20
30
40
50
60
70
80
90
df$ADL

```
hist(df$COG)
```


Histogram of df$COG
Frequency
0
5
10
15
20
25
0
5
10
15
20
25
30
df$COG

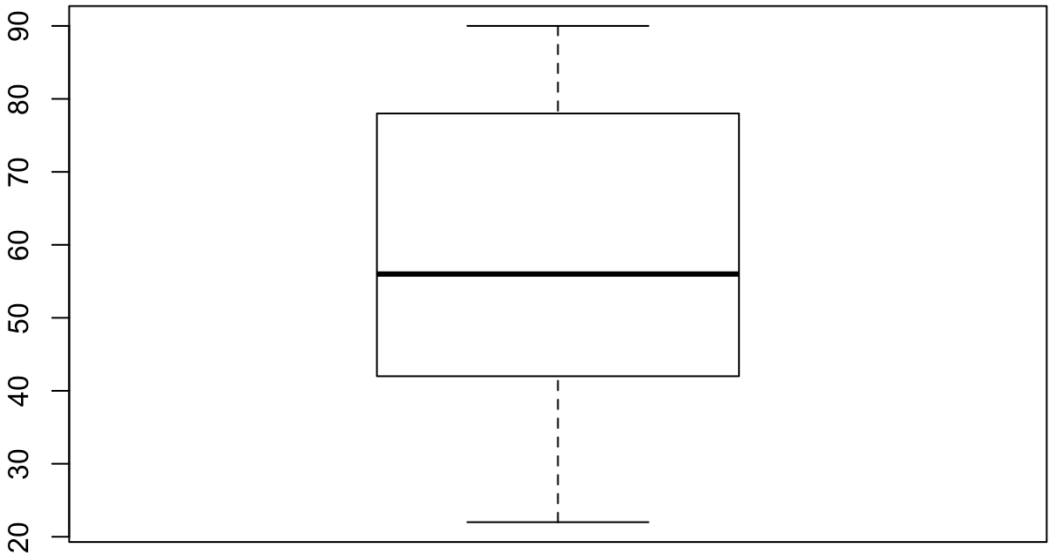
```
hist(df$MEM)
```



2. Produce the boxplot of ADL. Comment. Are there any outliers?

No, there doesn't appear to be any outliers. In fact, the boxplot looks normally spread and not skewed.

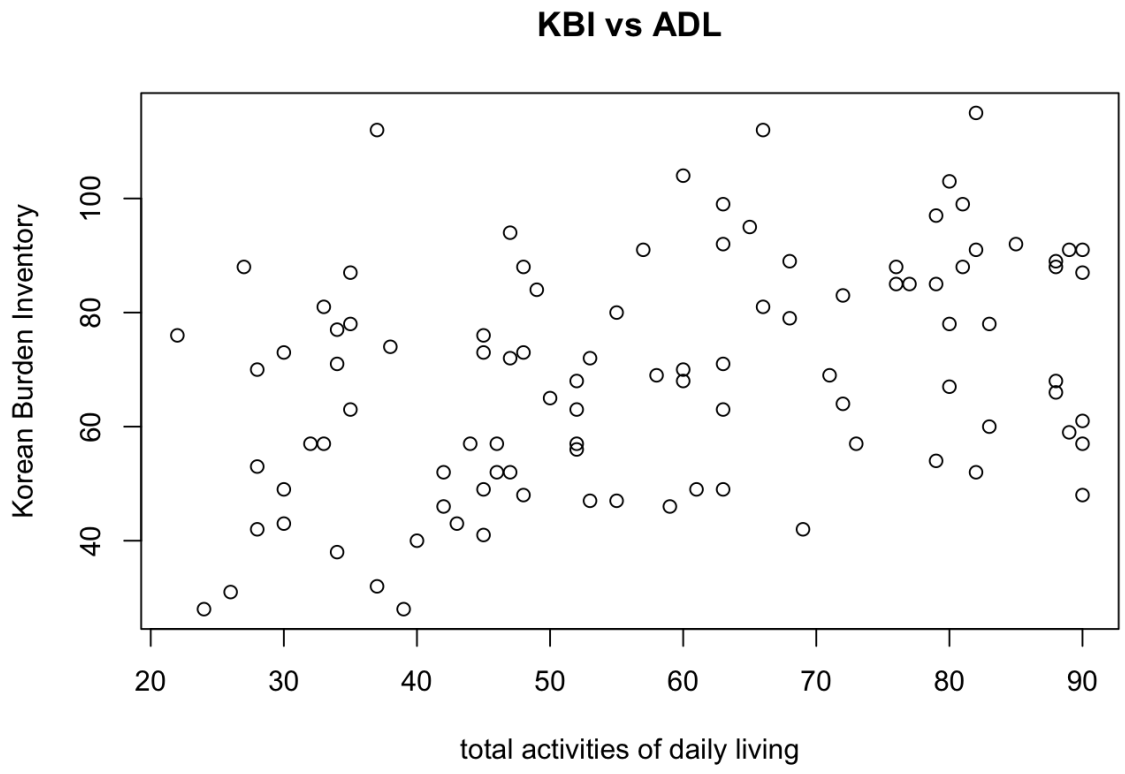
```
boxplot(df$ADL)
```



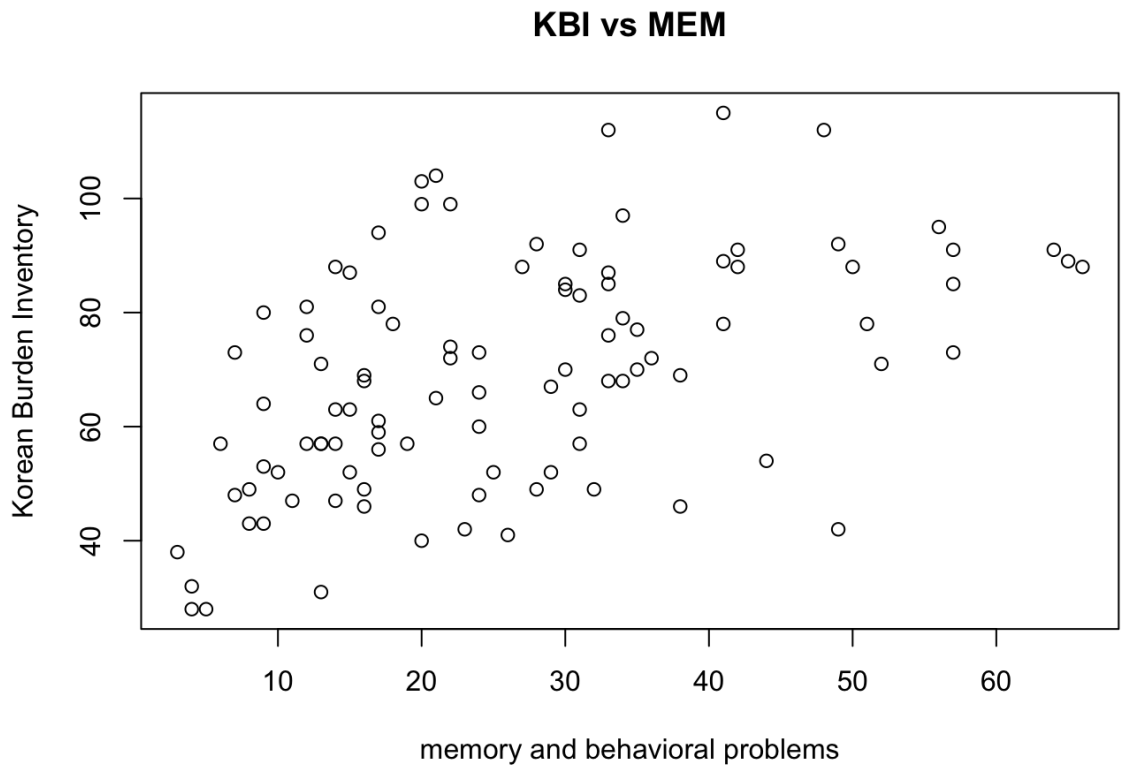
3. Produce the scatterplots of Y and the X's. Comment. Is a linear model appropriate for this data? Why or why not? Are the X's correlated amongst themselves.

A linear model may be appropriate for the regression variables ADL, MEM and COG. The relationship would be scattered for all three variables since the data doesn't look explicitely grouped together. The relation for ADL and MEM would be positive.

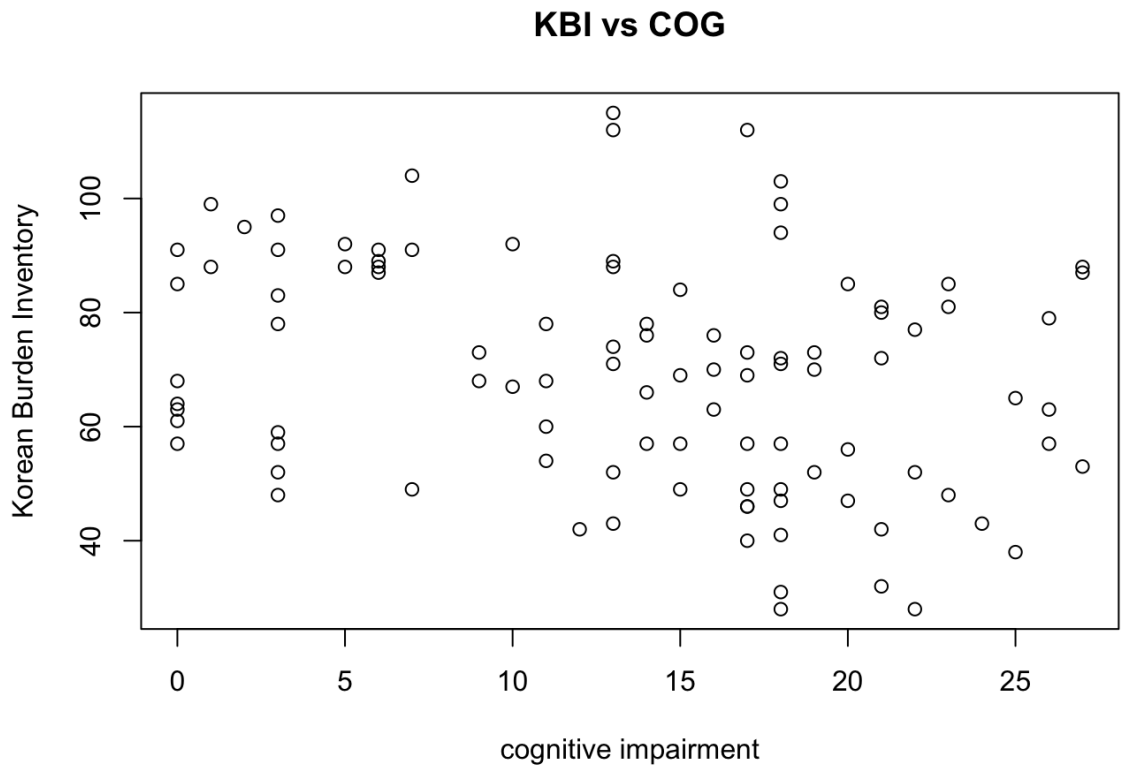
```
plot(df$ADL, df$KBI,
     main='KBI vs ADL',
     xlab="total activities of daily living", ylab="Korean Burden Invento
ry")
```




```
plot(df$MEM, df$KBI,
     main='KBI vs MEM',
     xlab="memory and behavioral problems", ylab="Korean Burden Inventor
y")
```




```
plot(df$COG, df$KBI,
     main='KBI vs COG',
     xlab="cognitive impairment", ylab="Korean Burden Inventory")
```



4. Fit univariable linear models, Y versus Xi for each of the three X regressor variables.

```
lmADL = lm(KBI~ADL, data = df)
lmMEM = lm(KBI~MEM, data = df)
lmCOG = lm(KBI~COG, data = df)
```

a. What are the estimated regression models?

```
summary(lmADL)
```

```
##
## Call:
## lm(formula = KBI ~ ADL, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -34.12 -15.78  -0.12  12.10  50.64
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 47.37433    5.69237   8.322 5.18e-13 ***
## ADL          0.37797    0.09297   4.066 9.68e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.65 on 98 degrees of freedom
## Multiple R-squared:  0.1443, Adjusted R-squared:  0.1356
## F-statistic: 16.53 on 1 and 98 DF,  p-value: 9.683e-05
```

```
summary(lmMEM)
```

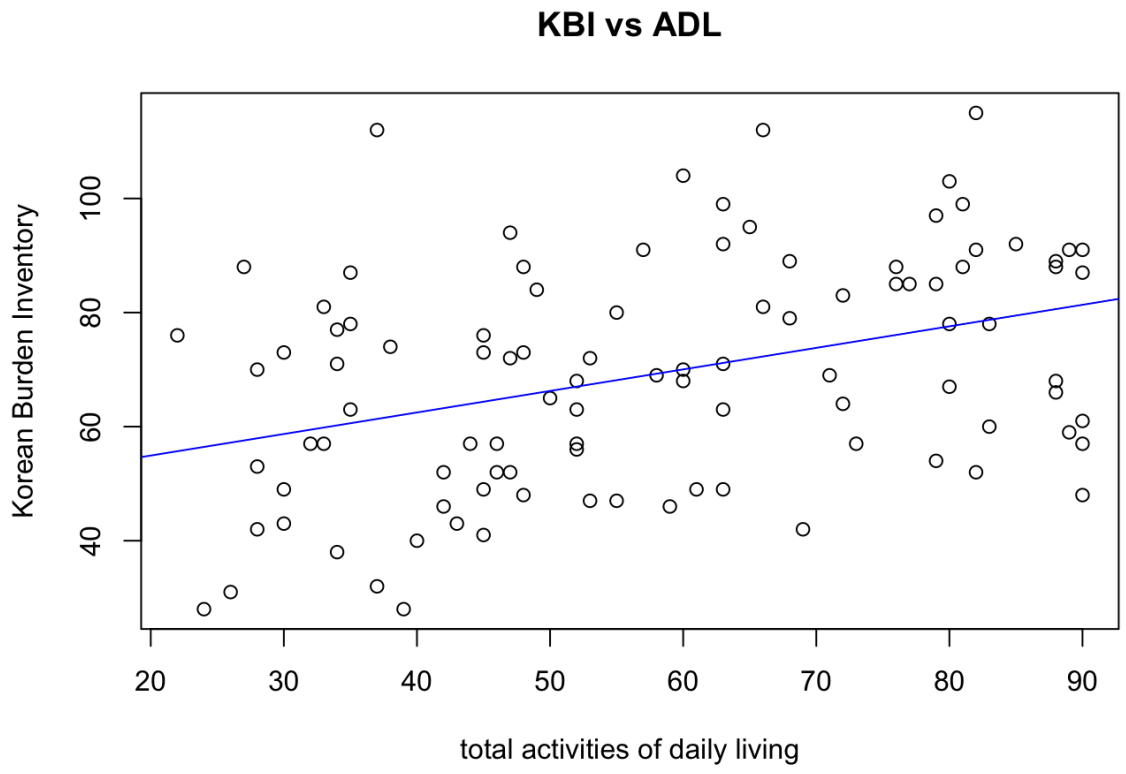
```
##
## Call:
## lm(formula = KBI ~ MEM, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.182 -10.394  -2.473  10.837  38.350
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  51.9287     3.4810  14.918  < 2e-16 ***
## MEM           0.6582     0.1146   5.746 1.03e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.44 on 98 degrees of freedom
## Multiple R-squared:  0.252,  Adjusted R-squared:  0.2444
## F-statistic: 33.02 on 1 and 98 DF,  p-value: 1.032e-07
```

```
summary(lmCOG)
```

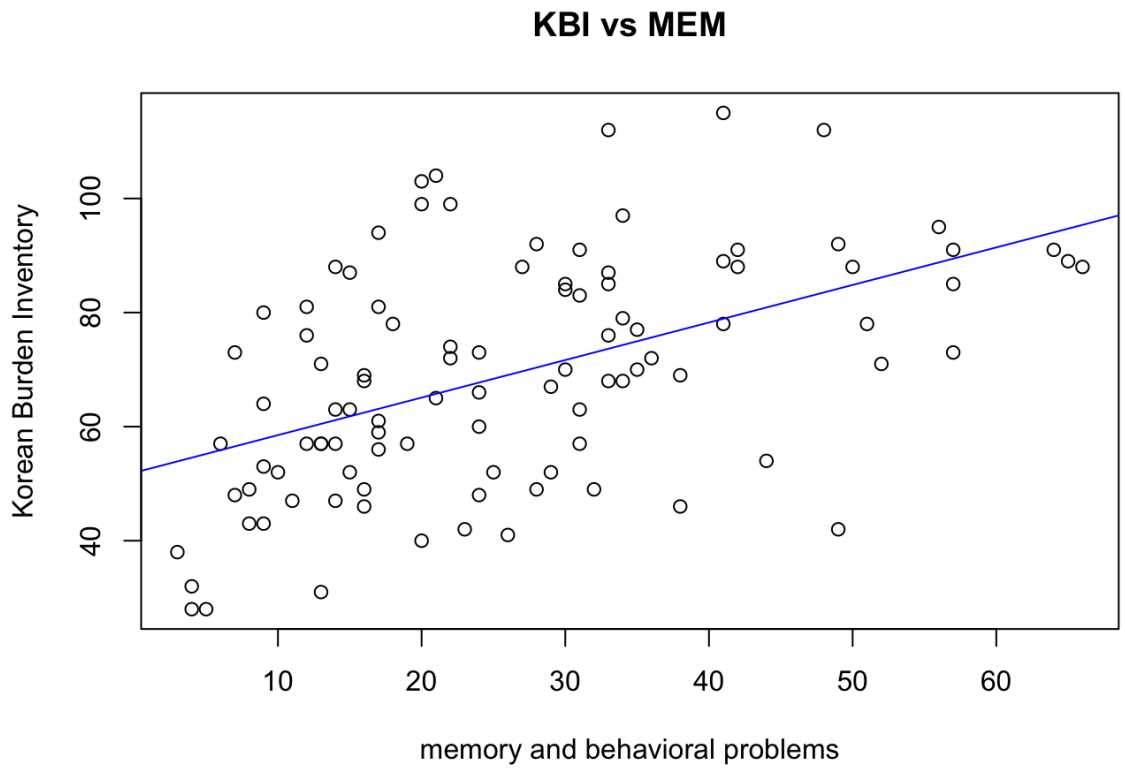
```
##
## Call:
## lm(formula = KBI ~ COG, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.156 -17.305   1.697  14.618  45.266
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  79.0354     3.9563   19.98  < 2e-16 ***
## COG          -0.7155     0.2520   -2.84  0.00549 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.38 on 98 degrees of freedom
## Multiple R-squared:  0.07604,    Adjusted R-squared:  0.06661
## F-statistic: 8.065 on 1 and 98 DF,  p-value: 0.005488
```

b. Compare the univariable models above.

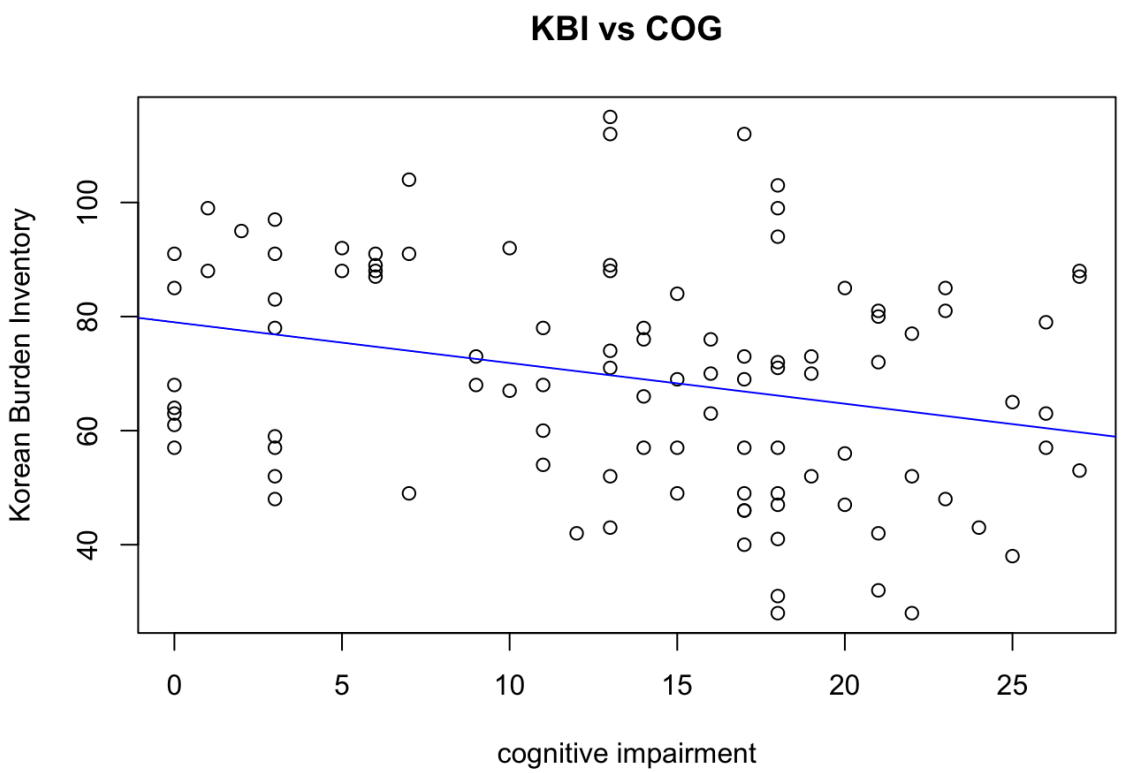
```
plot(df$ADL, df$KBI,
     main='KBI vs ADL',
     xlab="total activities of daily living", ylab="Korean Burden Invento
ry")
abline(lmADL, col="blue")
```




```
plot(df$MEM, df$KBI,
     main='KBI vs MEM',
     xlab="memory and behavioral problems", ylab="Korean Burden Inventor
y")
abline(lmMEM, col="blue")
```




```
plot(df$COG, df$KBI,
     main='KBI vs COG',
     xlab="cognitive impairment", ylab="Korean Burden Inventory")
abline(lmCOG, col="blue")
```



c. Check the fit of the models and comment.

The MEM linear model fits the best, as shown below by its r-squared value. That said, none of the regression models have an value over 50% which indicates that the data is scattered.

```
print(paste("ADL r-squared: ", percent(summary(lmADL)$r.squared)))
```

```
## [1] "ADL r-squared:  14.4%"
```

```
print(paste("MEM r-squared: ", percent(summary(lmMEM)$r.squared)))
```

```
## [1] "MEM r-squared:  25.2%"
```

```
print(paste("COG r-quared: ", percent(summary(lmCOG)$r.squared)))
```

```
## [1] "COG r-quared:  7.60%"
```

d. Explain each of the estimated regression parameters (except the intercept) in words.

Coefficients: - The slope coeficient is the value in the second row of the Estimates column. This value represents how much the Y value (KBI) increases for one step in the X value or regression variable.

Residual standard error: - Residuals are the difference between the actual observed regression values and the response values that the model predicted.

Multiple R-squared: - The r-squared value represents the measure of the linear relationship between our regression variable and our response variable.

F-statistic: - The F-test lets us quantify how well our data fits a model. The F-statistic = (Sum of squares for regression)/(Sums of squares for error). - The p-value indicates how likely our null hypothesis (in this case a slope of zero) is. The scientific consensus is that if our p-value is < 0.05, then the result is statistically significant and the null hypothesis is rejected.