1.7 Exercise (Basic stats)

CJF-4) Summarize the BondYield data in the AER library. What is the kurtosis? What is the skewness? Does the skewness value make sense given the values of mean and the median?

```
library(AER)
data(BondYields)
library(fBasics)
basicStats(BondYield)
```

R produces the following output:

```
             BondYield
nobs         60.000000
NAs           0.000000
Minimum       6.660000
Maximum       9.720000
1. Quartile   7.902500
3. Quartile   8.945000
Mean          8.290833
Median        8.300000
Sum         497.450000
SE Mean       0.104351
LCL Mean      8.082027
UCL Mean      8.499640
Variance      0.653354
Stdev         0.808303
Skewness     -0.234511
Kurtosis     -0.823370
```

Since the mean is very close to the median, it may seem that skewness should be very close to 0. However, the official measure is not that close to zero. The point is that the simple comparison of mean and median to indicate skewness is merely a crude approximation.

1.8 Exercise (Model comparison)

CJF-5) Using the Guns data in the AER library, regress the following 2 models:

Violent = b0 + b1(prisoners) + b2(income) + b3(density) + ϵ
Violent = b0 + b1(prisoners) + b2(income) + b3(density) + b4(law) + ϵ
Compare the 2 models. Which model is better? How can you tell?

```
rm(list=ls())
library(AER)
data(Guns)
reg1=lm(violent~prisoners+income+density,data=Guns)
reg2=lm(violent~prisoners+income+density+law,data=Guns)
summary(reg1);summary(reg2)
```

R produces the following output:

```
Call:
lm(formula = violent ~ prisoners + income + density, data = Guns)

Residuals:
     Min       1Q   Median       3Q      Max
-1173.95  -145.04   -35.04   119.28   671.42

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.605e+02  3.539e+01   4.536 6.33e-06 ***
prisoners   8.525e-01  4.420e-02  19.288  < 2e-16 ***
income      8.435e-03  2.732e-03   3.088  0.00206 **
density     9.554e+01  5.510e+00  17.341  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 210.5 on 1169 degrees of freedom
Multiple R-squared: 0.6044,     Adjusted R-squared: 0.6034
F-statistic: 595.3 on 3 and 1169 DF,  p-value: < 2.2e-16

>summary(reg2)

OUTPUT by R
Call:
lm(formula = violent ~ prisoners + income + density + law, data = Guns)
```

```
Residuals:
     Min       1Q   Median       3Q      Max 
-1225.51  -137.98   -39.65   118.48   777.04 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.882e+02  3.393e+01   5.546 3.61e-08 ***
prisoners    9.069e-01  4.257e-02  21.303  < 2e-16 ***
income       8.350e-03  2.612e-03   3.197  0.00142 ** 
density      8.637e+01  5.339e+00  16.176  < 2e-16 ***
lawyes      -1.465e+02  1.390e+01 -10.535  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 201.3 on 1168 degrees of freedom
Multiple R-squared: 0.6387,     Adjusted R-squared: 0.6375 
F-statistic: 516.3 on 4 and 1168 DF,  p-value: < 2.2e-16 

anova(reg1, reg2)
Analysis of Variance Table

Model 1: violent ~ prisoners + income + density
Model 2: violent ~ prisoners + income + density + law
  Res.Df      RSS  Df Sum of Sq      F    Pr(>F)    
1   1169 51807994                                   
2   1168 47312125   1   4495869 110.99 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
```

The analysis of variance shows that the second model is better. It includes the factor of whether the state has 'law' in effect for each year. The anova summary indicates that law is significant at any reasonable level.
Using the Guns data, run a regression of income on murder and plot the results and include a regression line in the plot.

```
plot(murder~income, data=Guns)
reg1=lm(murder~income, data=Guns)
abline(reg1)
```

Use the regression above, but this time include a quadratic term. Does the quadratic term help the explanation? Which model is better?

```
reg2=lm(murder~income+I(income^2), data=Guns)
anova(reg1,reg2)
```

R produces the following output:

```
Analysis of Variance Table

Model 1: murder ~ income
Model 2: murder ~ income + I(income^2)
  Res.Df   RSS   Df Sum of Sq      F    Pr(>F)
1   1171 63099
2   1170 54868    1      8231 175.52 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Including the square term gives a better model as indicated by the statistical significance of the F-test.

1.9 Exercise (Regression data plots)

CJF-6) using the data from USMacroB from the AER package, plot gnp and mbase (include a legend) and regress mbase and lagged gnp on gnp. What is the residual sum of squares (RSS)? Run a Durbin-Watson test to test for autocorrelation. Comment on the results

```
rm(list=ls())
library(AER)
data(USMacroB) #cannot access by name
#attach(data.frame(USMacroB))
plot(USMacroB[,c("gnp", "mbase")], lty=c(3,1),
plot.type="single", ylab="", lwd=1.5)
legend("topleft", legend = c("gnp", "money base"),
lty = c(3,1), bty="n")
library(dynlm)
reg1=dynlm(gnp~mbase + L(gnp), data=USMacroB)
summary(reg1)
deviance(reg1)
dwtest(reg1)
```