Regression Modeling with Actuarial and Financial Applications
Exercise 1.5

Note: See the last two pages of this document for code that is usable in a script file within R

Download 'AutoBI' data from Jed Frees' website
Follow the 'Data' link to find 'AutoBI'

Choose AutoBI.csv from its saved location

```
AutoBI <- read.table(choose.files(), header=TRUE,
     sep=",")
```

Notes

- The assignment operator **<-** assigns the chosen file to the variable `AutoBI`
- **`read.table`** reads a file in table format and creates a data frame from it, with cases corresponding to lines and variables to fields in the file.
- **`choose.files`** uses a Windows file dialog to choose a list of zero or more files interactively (this allows for more universal code, rather than using a specific location)
- **`header=TRUE`** tells `read.table` that the data has headers as its first row
- **`sep="," `** tells `read.table` the table is in comma separated format (.csv)
- For more information on R's syntax or functions, use the **`help`** function. The **`help`** function can be used to find more detailed explanations of other functions in R

```
help(Syntax)

help(help)
```

Get a summary of the data

```
summary(AutoBI)
```

a) Compute descriptive statistics only for `LOSS`

```
summary(AutoBI$LOSS)
```

Output:

```
Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
   0.005    0.640    2.331    5.953    3.995 1068.000
```

Don't forget to use `help` for more detailed explanations of functions!

Notes

- **`summary`** is a generic function used to produce result summaries of the results of various model fitting functions. The results will vary depending on the class of first argument
- In this case, summary's output will be descriptive statistics because the data are numerical
- **`$`** in `AutoBI$LOSS` extracts the component `LOSS` from `AutoBI`

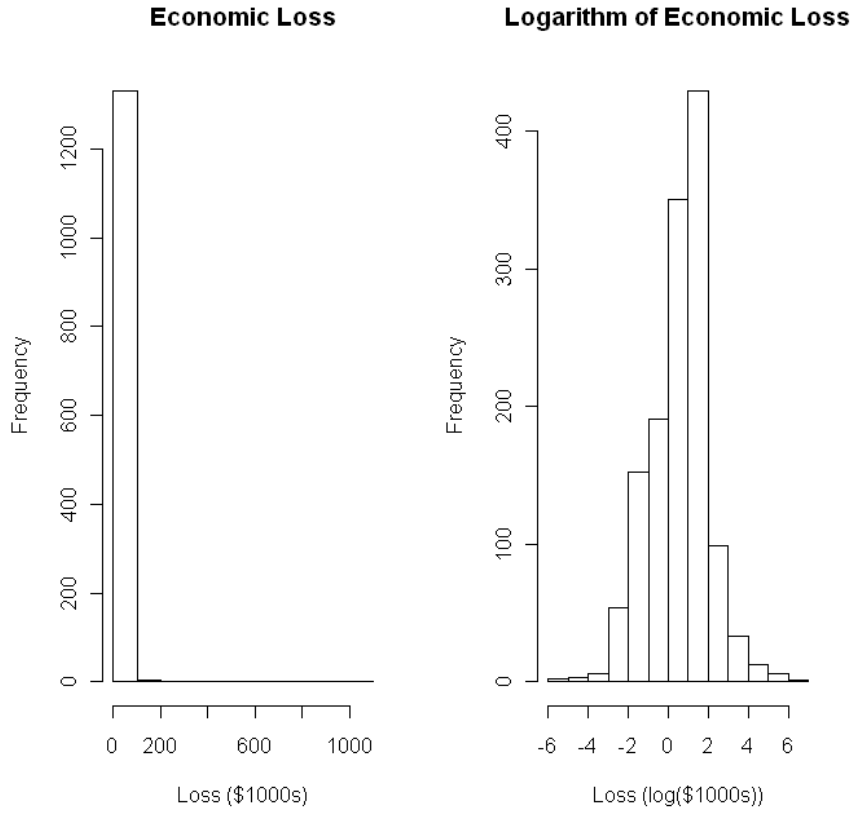
b) Compute a histogram and (normal) QQ plot for `LOSS`

Histograms

```
layout(matrix(1:2, nrow = 1))
hist(AutoBI$LOSS)
LOGLOSS <- log(AutoBI$LOSS)
hist(LOGLOSS)
```

Histograms with custom labels

```
hist(AutoBI$LOSS, main = "Economic Loss",
     xlab = "Loss ($1000s)")
hist(LOGLOSS, main = "Logarithm of Economic Loss",
     xlab = "Loss (log($1000s))")
```



Don't forget to use `help` for more detailed explanations of functions!

Interpretation

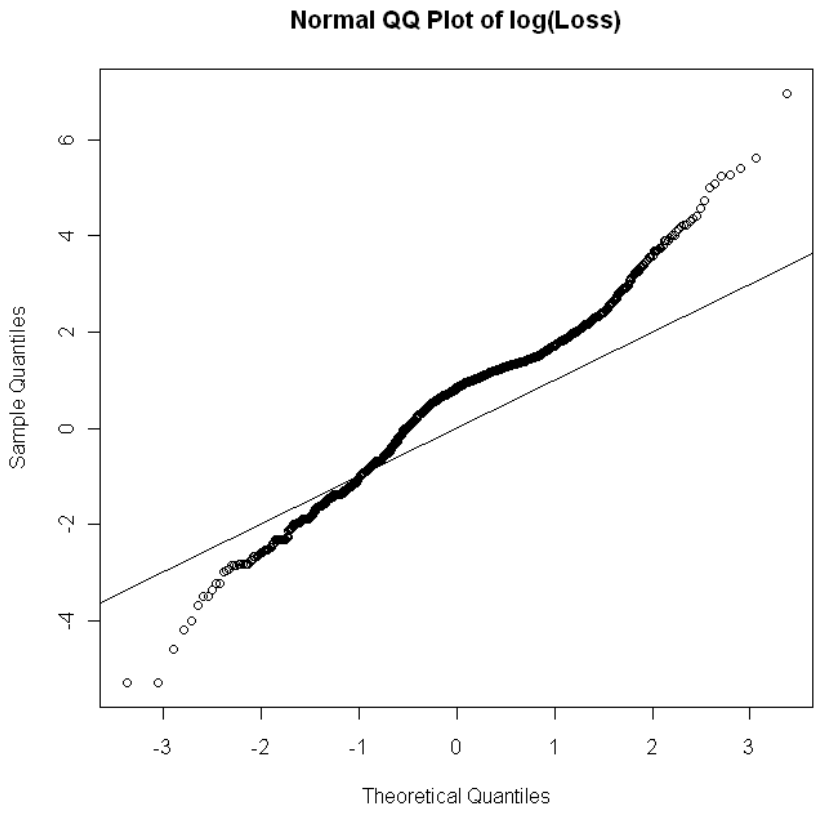
- Once a log transform is applied to the data and the histogram is plotted, the histogram seems to show a distribution that is skewed to the right.
- Histograms can sometimes be deceiving depending on the width and number of rectangles used to generate the graph.

Notes

- **`layout`** divides the device up into as many rows and columns as there are in matrix `mat`, with the column-widths and the row-heights specified in the respective arguments
- **`matrix`** creates a matrix from the given set of values
- In this case, **`nrow`** specifies the desired number of rows for the matrix
- **`hist`** computes a histogram of the given data values
- **`log`** computes logarithms, by default natural logarithms
- For details on **`main`**, **`xlab`**, and **`ylab`**, use `help(hist)`; these arguments can be used in many other graphs to create custom labels

Normal QQ Plot

```
dev.off()
qqnorm(LOGLOSS, main = "Normal QQ Plot of log(Loss)")
abline(0,1)
```



Don't forget to use `help` for more detailed explanations of functions!

Interpretation

- This QQ plot compares the distribution of the sample data (represented by the points) to the normal distribution (represented by the straight line).
- In this case, the QQ plot shows the sample data not following the normal distribution at all.

Notes

- **`dev.off`** shuts down the specified (by default the current) device; in this case, it also resets the layout of the graphical device
- **`qqnorm`** is a generic function, the default method of which produces a normal *qq* plot of the values in y
- **`abline`** adds one or more straight lines through the current plot. `abline(0,1)` produces a line with y-intercept = 0 and slope = 1.

c) Partition the dataset into two subsamples, one corresponding to those claims involving an attorney, and the other to those in which an attorney was not involved

```
Attorney1 <- subset(AutoBI, ATTORNEY==1)
Attorney2 <- subset(AutoBI, ATTORNEY==2)
```

Check to make sure data is partitioned correctly; the results should show that `Attorney1` has all 1's for `ATTORNEY`, and that `Attorney2` has all 2's for `ATTORNEY`

```
summary(Attorney1$ATTORNEY)
summary(Attorney2$ATTORNEY)
```

Output:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      1       1       1       1       1       1

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      2       2       2       2       2       2
```

i) For each subsample, compute the typical loss

```
summary(Attorney1$LOSS)
summary(Attorney2$LOSS)
```

Output:

```
 Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.052    2.162    3.417    9.863    5.831 1068.000

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0050  0.3195  0.9860  1.8650  2.4250 82.0000
```

Don't forget to use `help` for more detailed explanations of functions!

Notes

- **`subset`** returns subsets of vectors, matrices or data frames which meet certain conditions; in this case, when `ATTORNEY` equals `1` and `2` in `AutoBI`

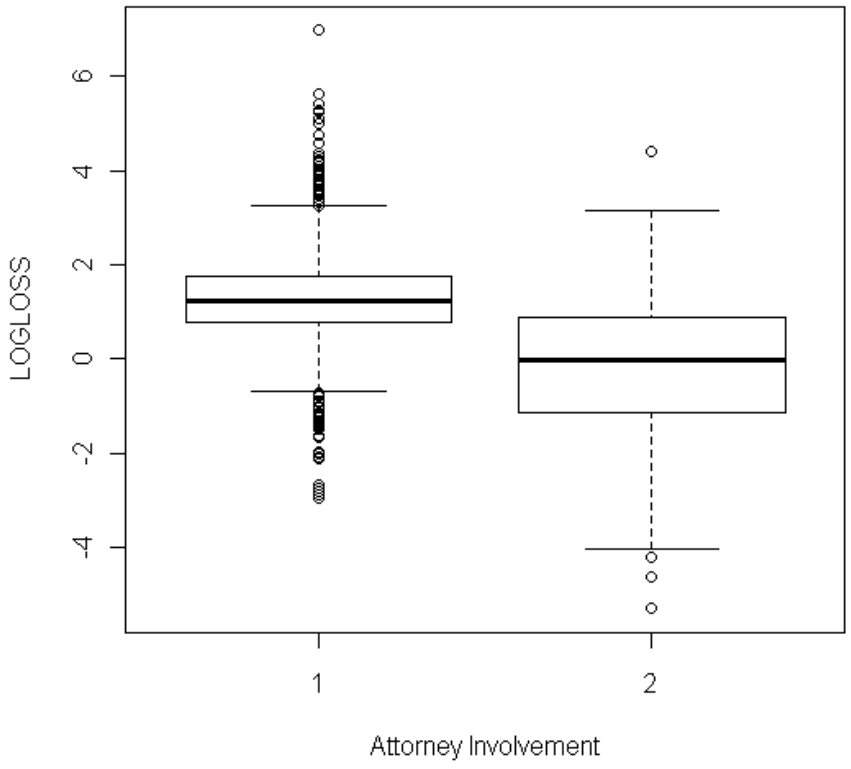
  ii) To compare distributions, compute a box plot by level of attorney involvement

  Box plot

```
boxplot(LOGLOSS ~ ATTORNEY, AutoBI)
```

  Box plot with custom labels

```
boxplot(LOGLOSS ~ ATTORNEY, AutoBI,
     xlab = "Attorney Involvement",
     ylab = "LOGLOSS")
```



Interpretation

- Comparison of the two box plots reveals that losses when an attorney was involved (ATTORNEY = 1) were higher than losses when no attorney was involved (ATTORNEY = 2).

Don't forget to use `help` for more detailed explanations of functions!

- The large number of outliers associated with ATTORNEY = 1 shows greater variability than the few outliers associated with ATTORNEY = 2. It also suggests that losses related to attorney involvement do not follow a normal distribution.

Notes

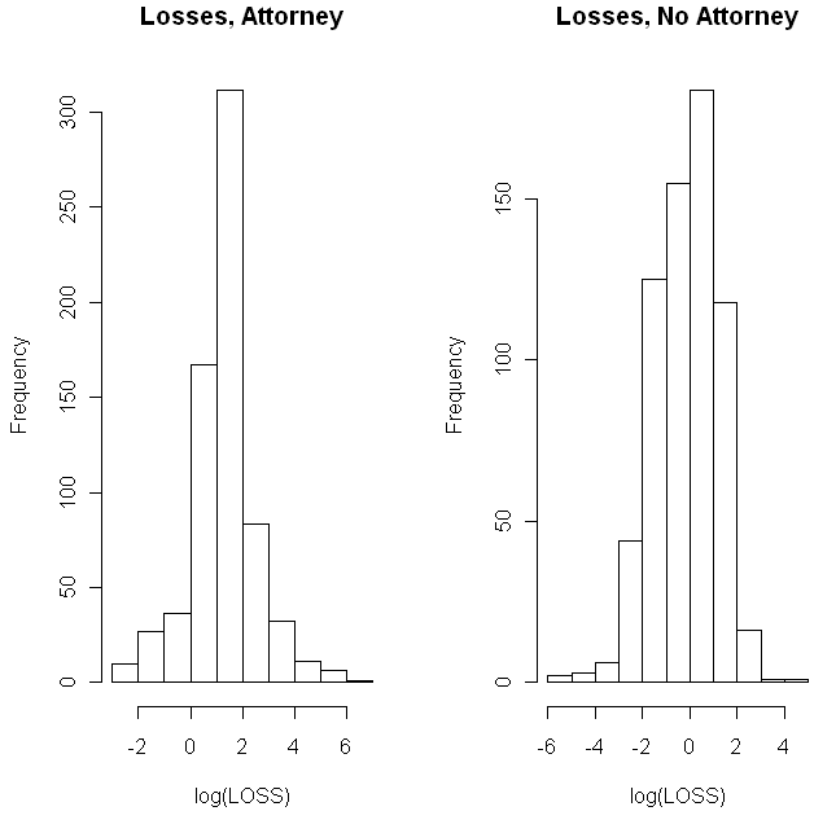
- **`boxplot`** produces box-and-whisker plot(s) of the given (grouped) values
- Use `help(boxplot)` to find out more about the arguments involved in `boxplot`
- **~** denotes a forumla

iii) For each subsample, compute a histogram and *qq* plot

```
LOGLOSS_A1 <- log(Attorney1$LOSS)
LOGLOSS_A2 <- log(Attorney2$LOSS)
```

Histogram Comparison

```
layout(matrix(1:2, nrow = 1))
hist(LOGLOSS_A1,
     main = "Losses, Attorney",
     xlab = "log(LOSS)")
hist(LOGLOSS_A2,
     main = "Losses, No Attorney",
     xlab = "log(LOSS)")
```



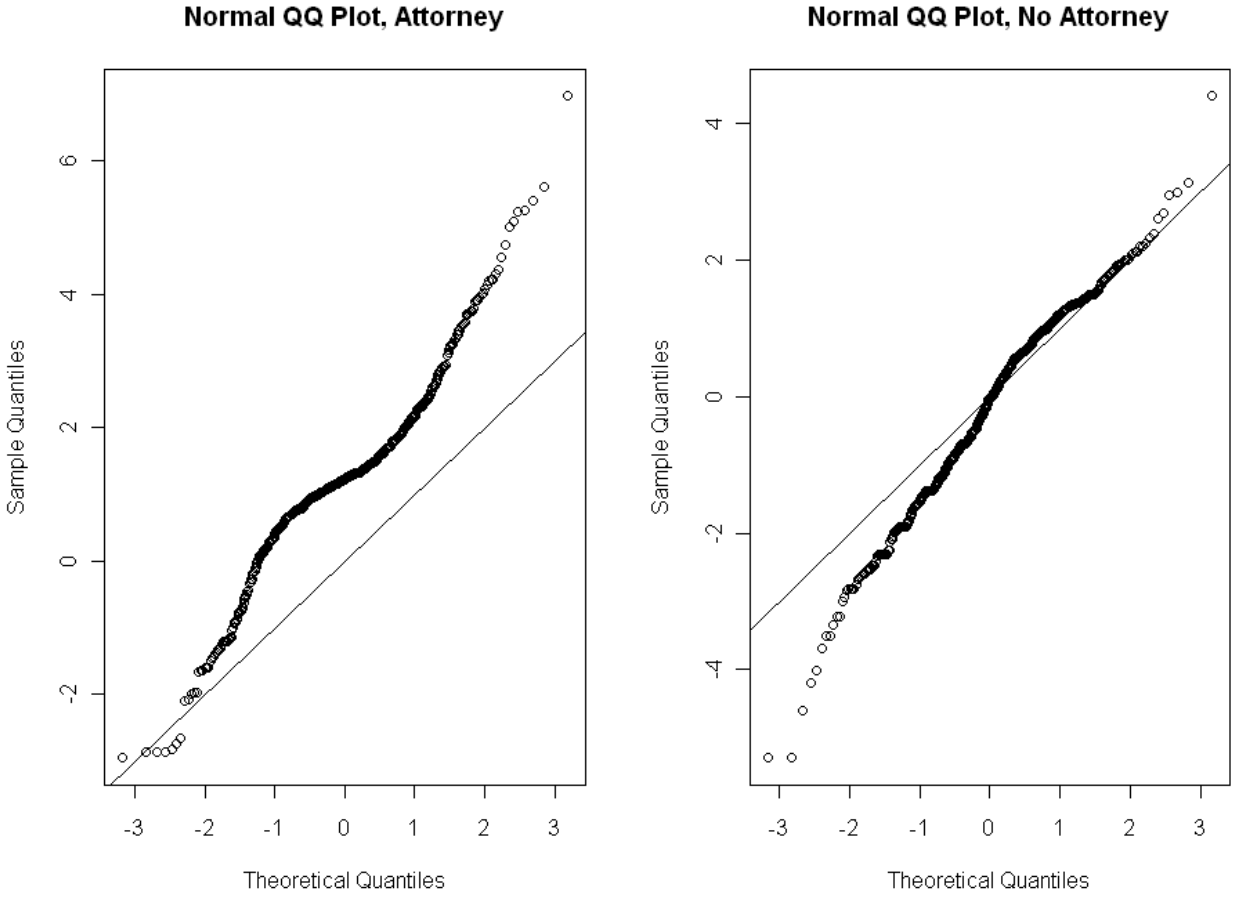
Don't forget to use `help` for more detailed explanations of functions!

Interpretation

- Losses associated with attorney involvement seem to be right skewed, whereas losses associated with no attorney involvement seem to be more normally distributed.

QQ Plot Comparison

```
layout(matrix(1:2, nrow = 1))
qqnorm(LOGLOSS_A1,
     main = "Normal QQ Plot, Attorney")
abline(0,1)
qqnorm(LOGLOSS_A2,
     main = "Normal QQ Plot, No Attorney")
abline(0,1)
```



Interpretation

- Comparison of the *qq* plots shows that losses associated with attorney involvement do not follow the normal distribution at all.
- The Q-Q plot, showing losses associated with no attorney involvement, suggests a long tail on the lower end of the data.

Don't forget to use `help` for more detailed explanations of functions!

Follow these steps to copy the following into a script file in order to copy code into R more easily:

- Open R
- Select File -> New Script
- Copy and paste this into the new script
- R will ignore text following #, which allows notes to be made in script files
- Highlighting lines of code and right-clicking allows the selection to be run in R

```
# Regression Modeling with Financial and Actuarial Applications
# Exercise 1.5

# Download 'AutoBI' data from Jed Frees' website
# Follow the 'Data' hyperlink to find 'AutoBI'

# Choose AutoBI.csv from its saved location

AutoBI <- read.table(choose.files(), header=TRUE, sep=",")

help(Syntax)
help(help)

# Get a summary of the data

summary(AutoBI)

# a) Compute descriptive statistics only for LOSS

summary(AutoBI$LOSS)

# b) Compute a histogram and (normal) Q-Q plot for LOSS

# Histograms

layout(matrix(1:2, nrow = 1))
hist(AutoBI$LOSS)
LOGLOSS <- log(AutoBI$LOSS)
hist(LOGLOSS)

# Histograms with custom labels

hist(AutoBI$LOSS, main = "Economic Loss",
xlab = "Loss ($1000s)")
hist(LOGLOSS, main = "Logarithm of Economic Loss",
xlab = "Loss (log($1000s))")

# Normal Q-Q Plot

dev.off()
qqnorm(LOGLOSS, main = "Normal QQ Plot of log(Loss)")
abline(0,1)

# c) Partition the dataset into two subsamples,
# one corresponding to those claims involving an attorney,
# and the other to those in which an attorney was not involved
```

Don't forget to use `help` for more detailed explanations of functions!

```
Attorney1 <- subset(AutoBI, ATTORNEY==1)
Attorney2 <- subset(AutoBI, ATTORNEY==2)

# Check to make sure data is partitioned correctly by checking the
summary
# statstics of each partition's ATTORNEY data.

summary(Attorney1$ATTORNEY)
summary(Attorney2$ATTORNEY)

# i) For each subsample, compute the typical loss

summary(Attorney1$LOSS)
summary(Attorney2$LOSS)

# ii) To compare distributions, compute a box plot
# by level of attorney involvement

# Box plot

boxplot(LOGLOSS ~ ATTORNEY, AutoBI)

# Box plot with custom labels

boxplot(LOGLOSS ~ ATTORNEY, AutoBI,
xlab = "Attorney Involvement",
      ylab = "LOGLOSS")

# iii) For each subsample, compute a histogram and qq plot

LOGLOSS_A1 <- log(Attorney1$LOSS)
LOGLOSS_A2 <- log(Attorney2$LOSS)

# Histogram Comparison

layout(matrix(1:2, nrow = 1))
hist(LOGLOSS_A1,
main = "Losses, Attorney",
xlab = "log(LOSS)")
hist(LOGLOSS_A2,
main = "Losses, No Attorney",
xlab = "log(LOSS)")

# QQ Plot Comparison

layout(matrix(1:2, nrow = 1))
qqnorm(LOGLOSS_A1,
main = "Normal QQ Plot, Attorney")
abline(0,1)
qqnorm(LOGLOSS_A2,
main = "Normal QQ Plot, No Attorney")
abline(0,1)
```

Don’t forget to use `help` for more detailed explanations of functions!