HW2 Solution

Taebin Kim

Data analysis on `rock` dataset

First, load the dataset and perform an initial data analysis.

```r
data(rock)
str(rock)
```

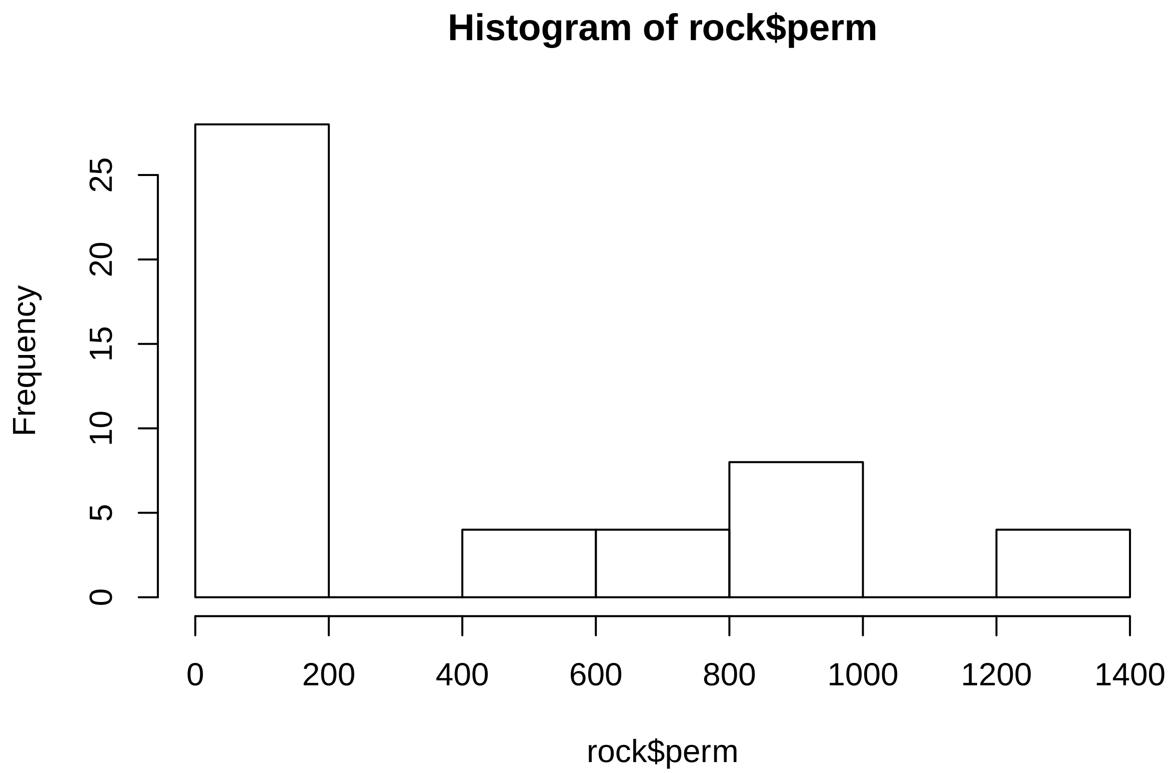
```
## 'data.frame':    48 obs. of  4 variables:
##  $ area : int  4990 7002 7558 7352 7943 7979 9333 8209 8393 6425 ...
##  $ peri : num  2792 3893 3931 3869 3949 ...
##  $ shape: num  0.0903 0.1486 0.1833 0.1171 0.1224 ...
##  $ perm : num  6.3 6.3 6.3 6.3 17.1 17.1 17.1 17.1 119 119 ...
```

```r
summary(rock)
```

```
##       area            peri            shape              perm
##  Min.   : 1016   Min.   : 308.6   Min.   :0.09033   Min.   :   6.30
##  1st Qu.: 5305   1st Qu.:1414.9   1st Qu.:0.16226   1st Qu.:  76.45
##  Median : 7487   Median :2536.2   Median :0.19886   Median : 130.50
##  Mean   : 7188   Mean   :2682.2   Mean   :0.21811   Mean   : 415.45
##  3rd Qu.: 8870   3rd Qu.:3989.5   3rd Qu.:0.26267   3rd Qu.: 777.50
##  Max.   :12212   Max.   :4864.2   Max.   :0.46413   Max.   :1300.00
```

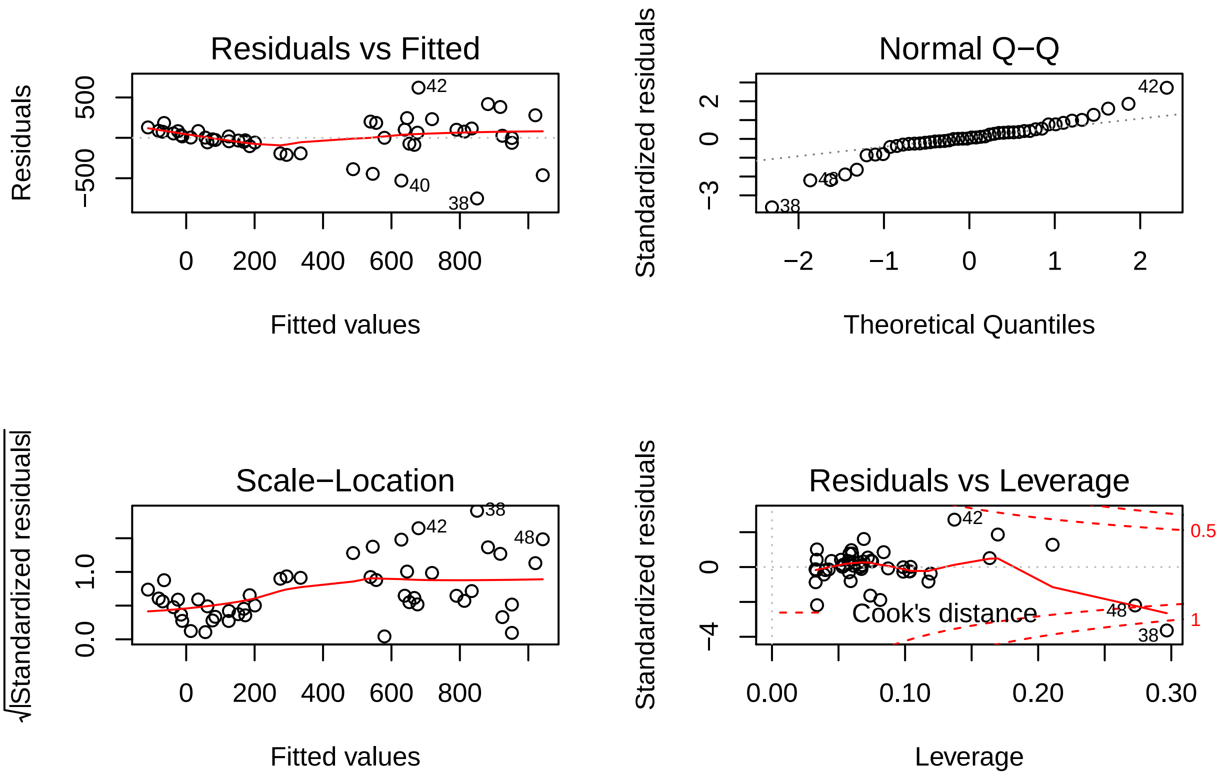
We can see that `perm` is right-skewed substantially. Hence draw a histogram of the variable.

```r
hist(rock$perm)
```



The histogram confirms the skewness of `perm`, and we should consider applying transformation methods to the variable. Fit a basic linear model to `rock` data beforehand.

```
lmod <- lm(perm ~ ., data = rock)
par(mfrow = c(2,2))
plot(lmod)
```

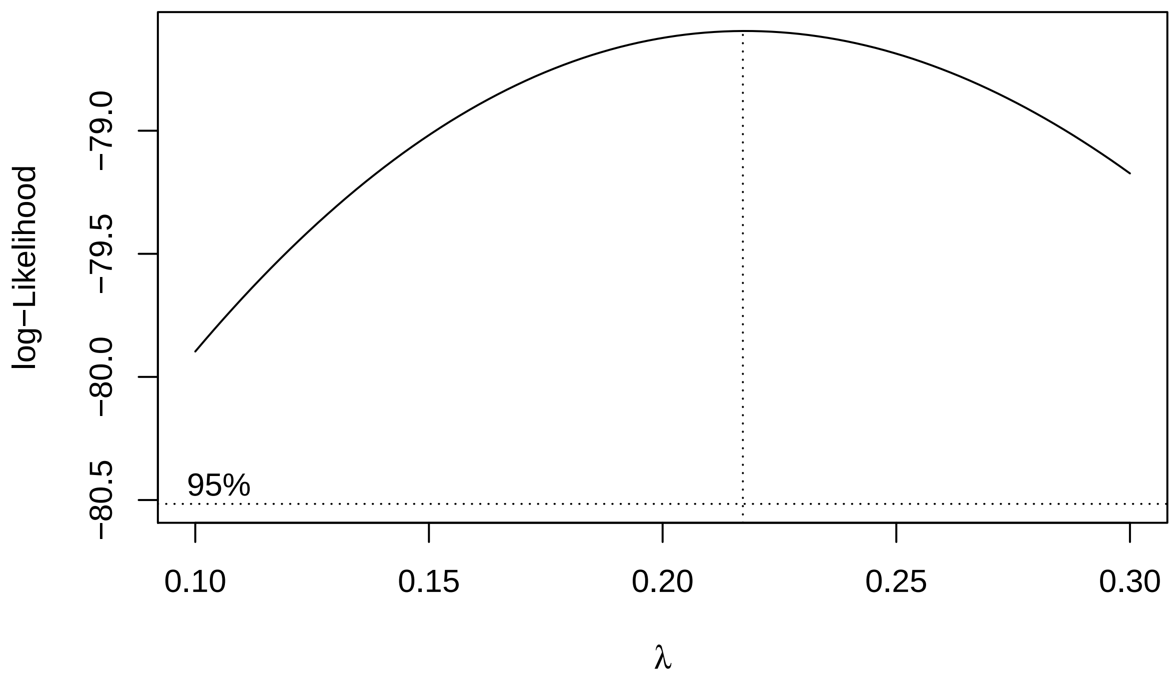



```r
summary(lmod)
```

```
##
## Call:
## lm(formula = perm ~ ., data = rock)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -750.26  -59.57   10.66  100.25  620.91
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 485.61797  158.40826   3.066 0.003705 **
## area          0.09133    0.02499   3.654 0.000684 ***
## peri         -0.34402    0.05111  -6.731 2.84e-08 ***
## shape       899.06926  506.95098   1.773 0.083070 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 246 on 44 degrees of freedom
## Multiple R-squared:  0.7044, Adjusted R-squared:  0.6843
## F-statistic: 34.95 on 3 and 44 DF,  p-value: 1.033e-11
```

Every plot shows that the residuals are not normal. The QQ plot is nonlinear and there are some outliers including observations 38, 42, and 48. Apply boxcox transformation to the data.

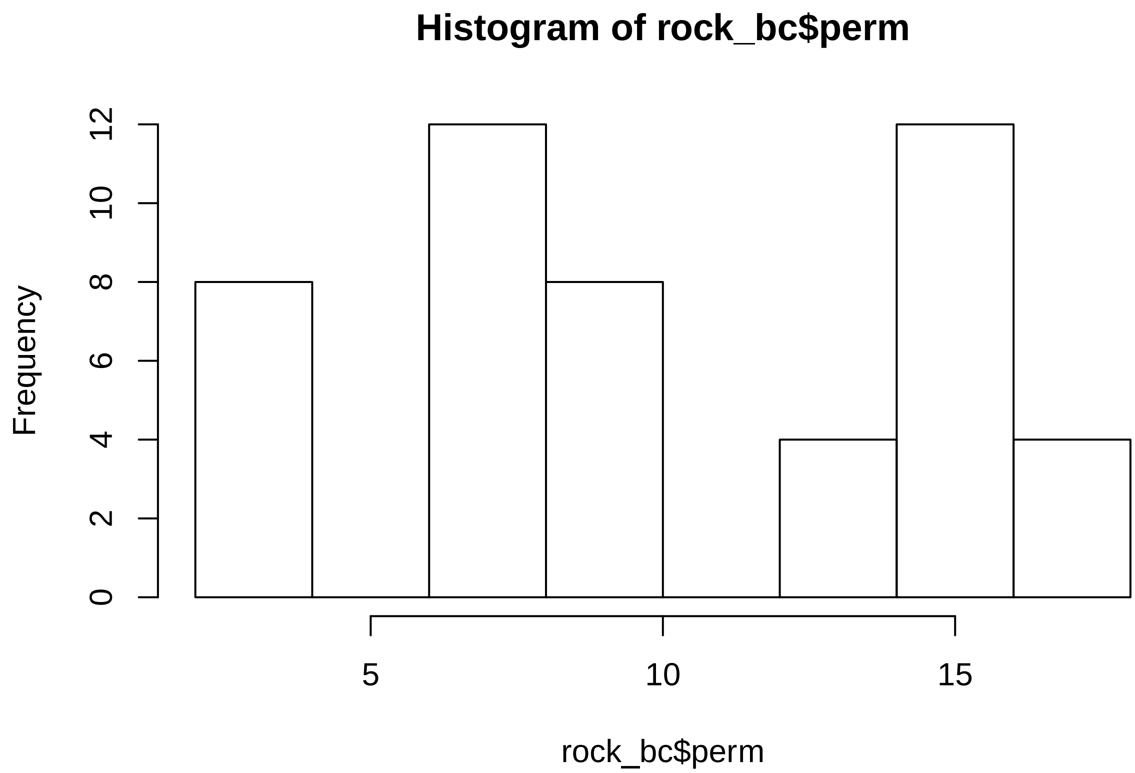
```r
library(MASS)
boxcox(lmod, lambda=seq(0.1, 0.3,by=0.01))
```




```r
lamb <- 0.22
rock_bc <- rock
rock_bc$perm <- (rock$perm^lamb - 1) / lamb
```

Transform the data with the optimal $\lambda = 0.22$. Then perform forward or backward variable selection.

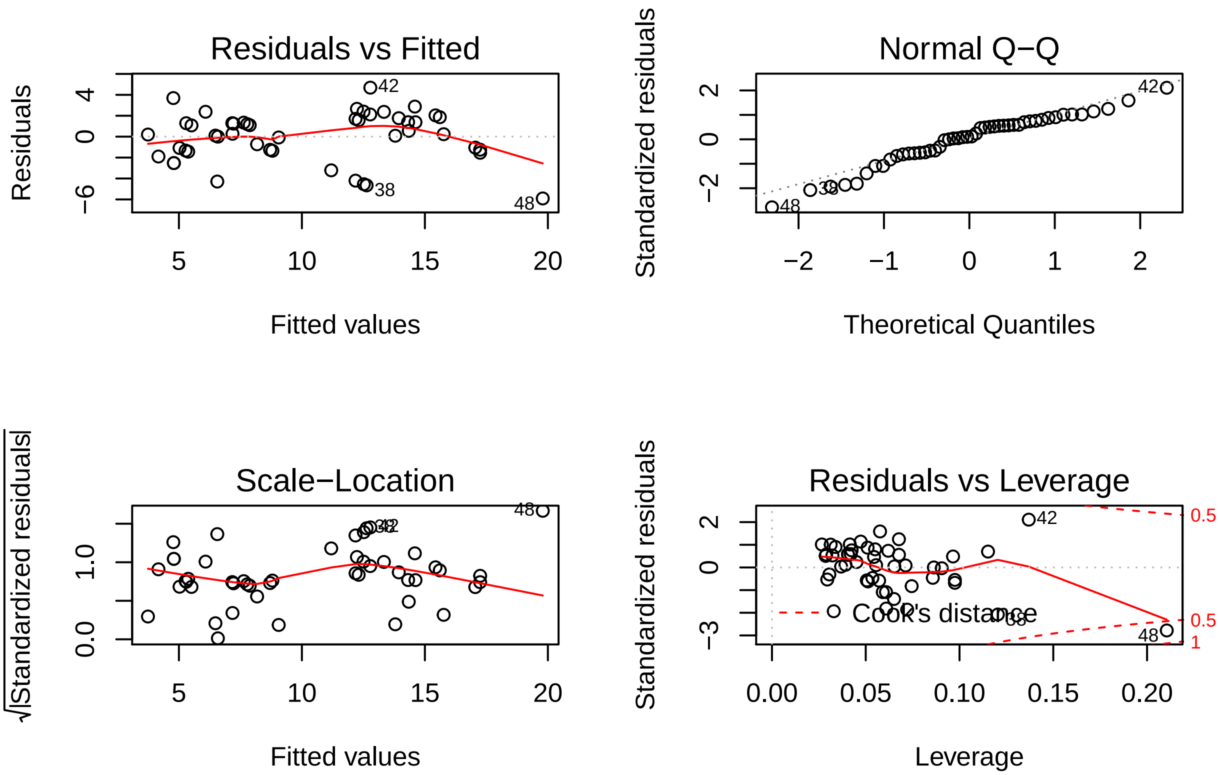
```r
hist(rock_bc$perm)
```




```r
lmod_bc <- lm(perm ~ ., data = rock_bc)
step(lmod_bc, trace = F)
```

```
##
## Call:
## lm(formula = perm ~ area + peri, data = rock_bc)
##
## Coefficients:
## (Intercept)         area         peri
##   12.792722     0.001461    -0.004843
```

```r
lmod_step <- lm(perm ~ area + peri, data = rock_bc)
par(mfrow = c(2,2))
plot(lmod_step)
```



The result shows that the residuals are more normal and the QQ plot is more linear. Observation 42 is still an outlier. Thus, we predict the `perm` of the observation from the linear model obtained after omitting it.

```
rock_omit <- rock_bc[-42,]
lmod_omit <- lm(perm ~ area + peri, data = rock_omit)
predict(lmod_omit, newdata = rock_bc[42,], interval = "prediction", level = 0.95)
```

```
##         fit      lwr      upr
## 42 12.03884 7.058357 17.01932
```

```
rock_bc[42,]$perm
```

```
## [1] 17.4658
```

We can observe that `perm` of observation 42 is outside the 95% prediction interval from the omitted linear model. We conclude that the observation is an outlier. The overall result of the linear model is not satisfactory. This may be because the structure of the `rock` dataset is very strange. It consists of twelve different specimens repeated four times. This fact could be verified by the following code.

```
help(rock)
```

Data analysis on `prostate` dataset

First, load the dataset and perform an initial data analysis.

```r
library(faraway)
data(prostate)
str(prostate)
```

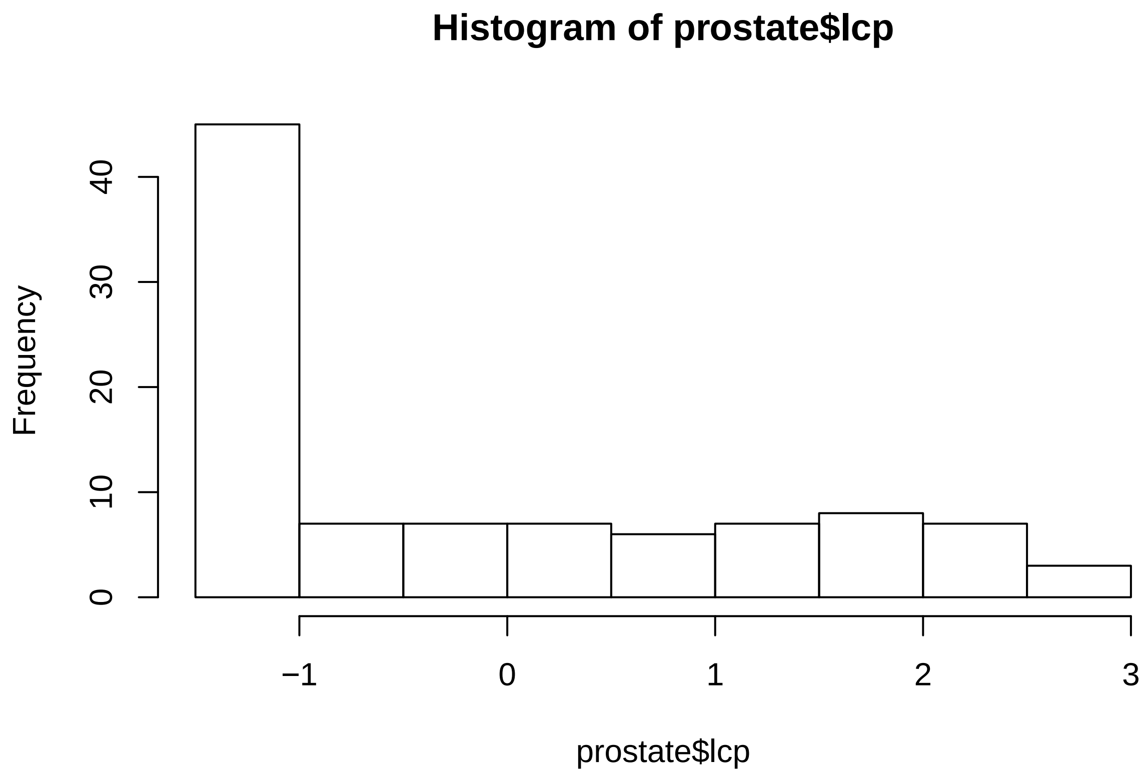
```
## 'data.frame':    97 obs. of  9 variables:
##  $ lcavol : num  -0.58 -0.994 -0.511 -1.204 0.751 ...
##  $ lweight: num  2.77 3.32 2.69 3.28 3.43 ...
##  $ age    : int  50 58 74 58 62 50 64 58 47 63 ...
##  $ lbph   : num  -1.39 -1.39 -1.39 -1.39 -1.39 ...
##  $ svi    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ lcp    : num  -1.39 -1.39 -1.39 -1.39 -1.39 ...
##  $ gleason: int  6 6 7 6 6 6 6 6 6 6 ...
##  $ pgg45  : int  0 0 20 0 0 0 0 0 0 0 ...
##  $ lpsa   : num  -0.431 -0.163 -0.163 -0.163 0.372 ...
```

```r
summary(prostate)
```

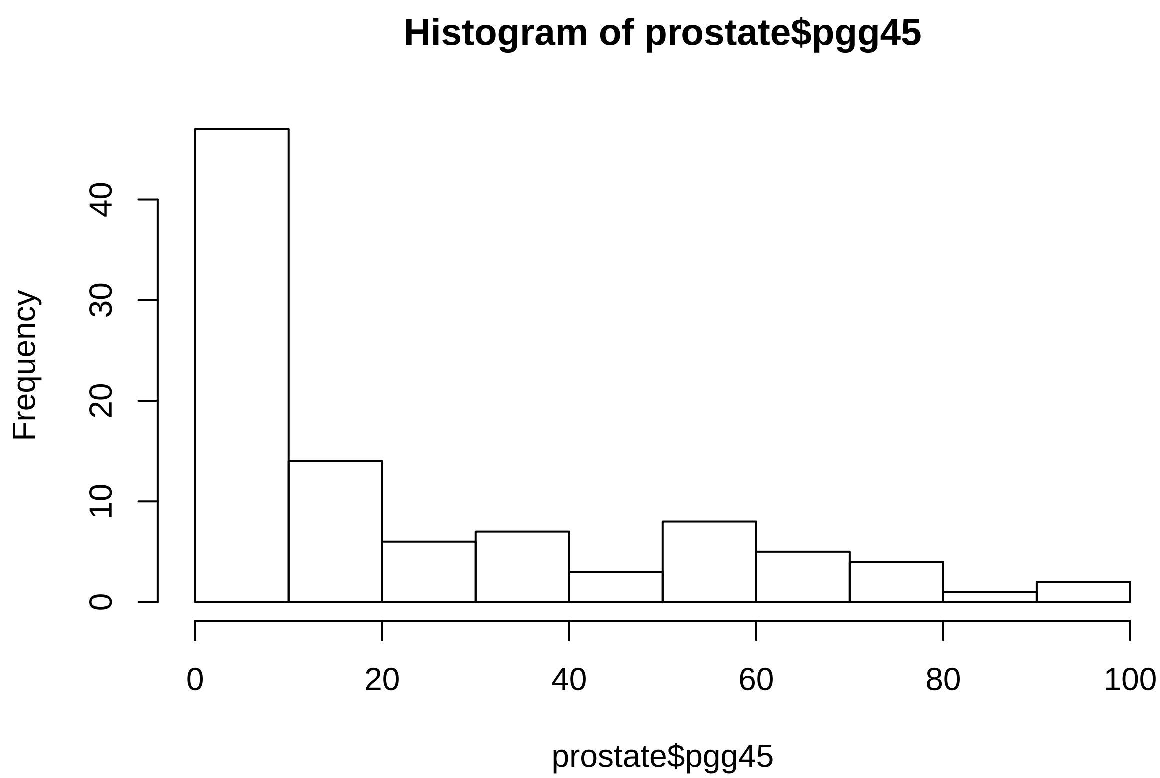
```
##      lcavol           lweight           age             lbph        
##  Min.   :-1.3471   Min.   :2.375   Min.   :41.00   Min.   :-1.3863  
##  1st Qu.: 0.5128   1st Qu.:3.376   1st Qu.:60.00   1st Qu.:-1.3863  
##  Median : 1.4469   Median :3.623   Median :65.00   Median : 0.3001  
##  Mean   : 1.3500   Mean   :3.653   Mean   :63.87   Mean   : 0.1004  
##  3rd Qu.: 2.1270   3rd Qu.:3.878   3rd Qu.:68.00   3rd Qu.: 1.5581  
##  Max.   : 3.8210   Max.   :6.108   Max.   :79.00   Max.   : 2.3263  
##       svi              lcp             gleason          pgg45       
##  Min.   :0.0000   Min.   :-1.3863   Min.   :6.000   Min.   :  0.00  
##  1st Qu.:0.0000   1st Qu.:-1.3863   1st Qu.:6.000   1st Qu.:  0.00  
##  Median :0.0000   Median :-0.7985   Median :7.000   Median : 15.00  
##  Mean   :0.2165   Mean   :-0.1794   Mean   :6.753   Mean   : 24.38  
##  3rd Qu.:0.0000   3rd Qu.: 1.1786   3rd Qu.:7.000   3rd Qu.: 40.00  
##  Max.   :1.0000   Max.   : 2.9042   Max.   :9.000   Max.   :100.00  
##       lpsa        
##  Min.   :-0.4308  
##  1st Qu.: 1.7317  
##  Median : 2.5915  
##  Mean   : 2.4784  
##  3rd Qu.: 3.0564  
##  Max.   : 5.5829  
```

We can see that `lcp` and `pgg45` are right-skewed. Hence draw histograms of the variables.

```r
hist(prostate$lcp)
```




```r
hist(prostate$pgg45)
```



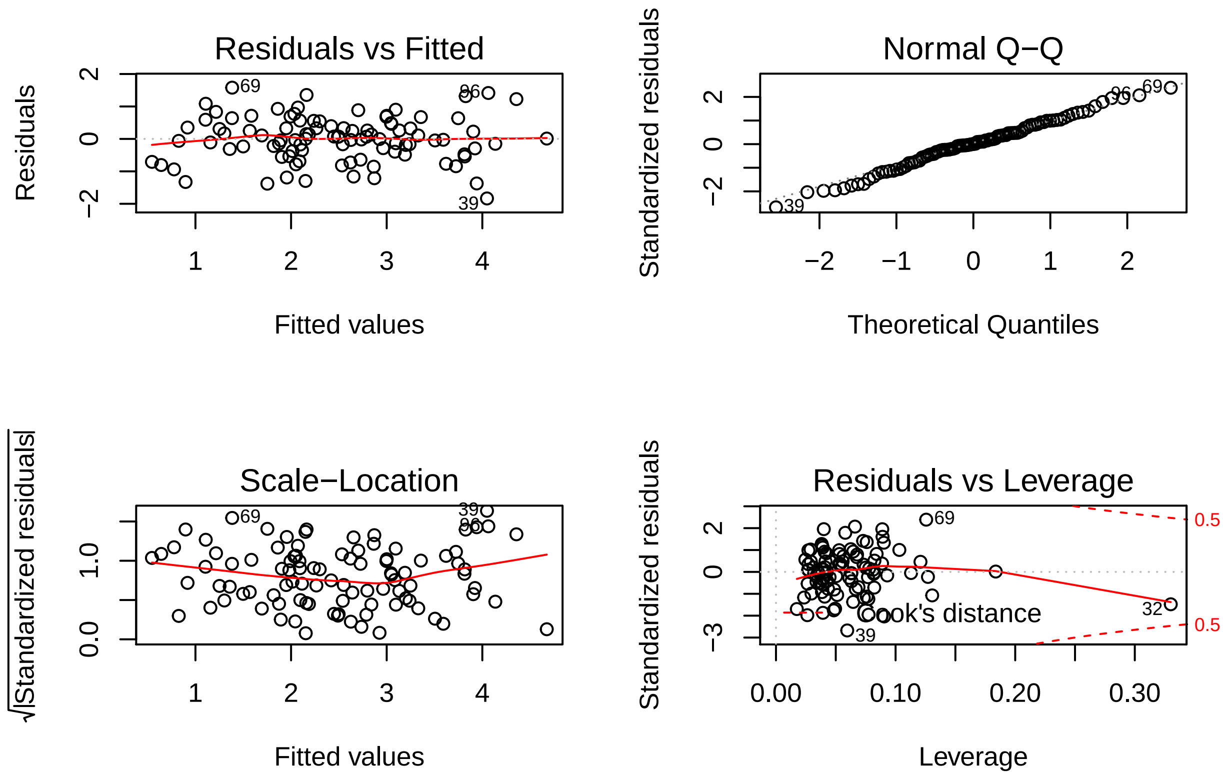
The histogram confirms the skewness of the variables, but it is not substantial. Perform forward or backward variable selection.

```
lmod <- lm(lpsa ~ ., data = prostate)
step(lmod, trace = F)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi, data = prostate)
##
## Coefficients:
## (Intercept)       lcavol      lweight          age         lbph
##     0.95100      0.56561      0.42369     -0.01489      0.11184
##         svi
##     0.72095
```

The best linear model is `lpsa` ~ `lcavol` + `weight` + `age` + `lbph` + `svi`.

```
lmod_step <- lm(lpsa ~ lcavol + lweight + age + lbph + svi, data = prostate)
par(mfrow = c(2,2))
plot(lmod_step)
```



The result shows that the residuals are normal and the QQ plot is linear. Since observation 69 is clearly an outlier, we calculate the prediction interval of the observation with the omitted linear model.

```
prostate_omit <- prostate[-69,]
lmod_omit <- lm(lpsa ~ lcavol + lweight + age + lbph + svi, data = prostate_omit)
predict(lmod_omit, newdata = prostate[69,], interval = "prediction", level = 0.95)
```

```
##         fit        lwr      upr
## 69 1.157023 -0.3059446 2.619991
```

```
prostate[69,]$lpsa
```

```
## [1] 2.96269
```

We can observe that `lpsa` of observation 69 is outside the 95% prediction interval from the omitted linear model. We conclude that the observation is an outlier.