# Assignment 2: Multiple Regression Analysis

*Keith G. Williams - 800690755*

*Friday, May 29, 2015*

Completed as part of DSBA 6201, SUM I 2015.

## The Data Set

The "Boston Housing" dataset recorded properties of 506 housing zones in the Greater Boston area. Typically, one is interested in predicting `MEDV` (median home value) based on other attributes.

Here is a list of attribute information: 1. `CRIM`: per capita crime rate by town 2. `ZN`: proportion of residential land zoned for lots over 25,000 $ft^2$ 3. `INDUS`: proportion of non-retail business acres per town 4. `CHAS`: Charles River dummy variable (=1 if tract bounds river; 0 otherwise) 5. `NOX`: nitric oxides concentraion (parts per 10 million) 6. `RM`: average number of rooms per dwelling 7. `AGE`: proportion of owner-occupied units built prior to 1940 8. `DIS`: weighted distances to five Boston employment centers 9. `RAD`: index of accessibility to radial highways 10. `TAX`: full-value property-tax rate per $10,000 11. `PTRATIO`: pupil-teacher ratio by town 12. `LSTAT`: % lower status of the population 13. `MEDV`: median value of owneer-occupied homes in $1000's

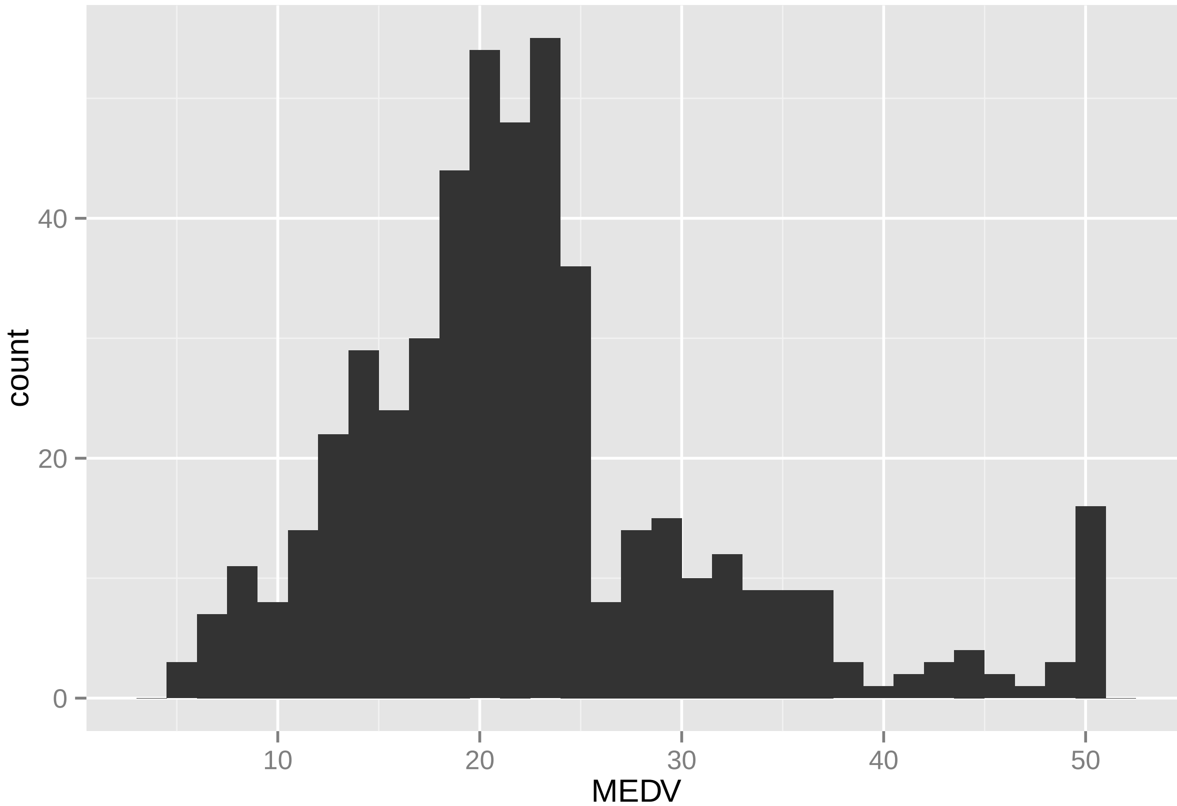
```
file <- "~/DSBA 6201/Boston Housing.csv"
bos <- read.csv(file)
```

## Pre-processing

`MEDV` has somewhat longish tail and is not so normally distributed, so we will take the log transform, and then predict `LMEDV` instead.

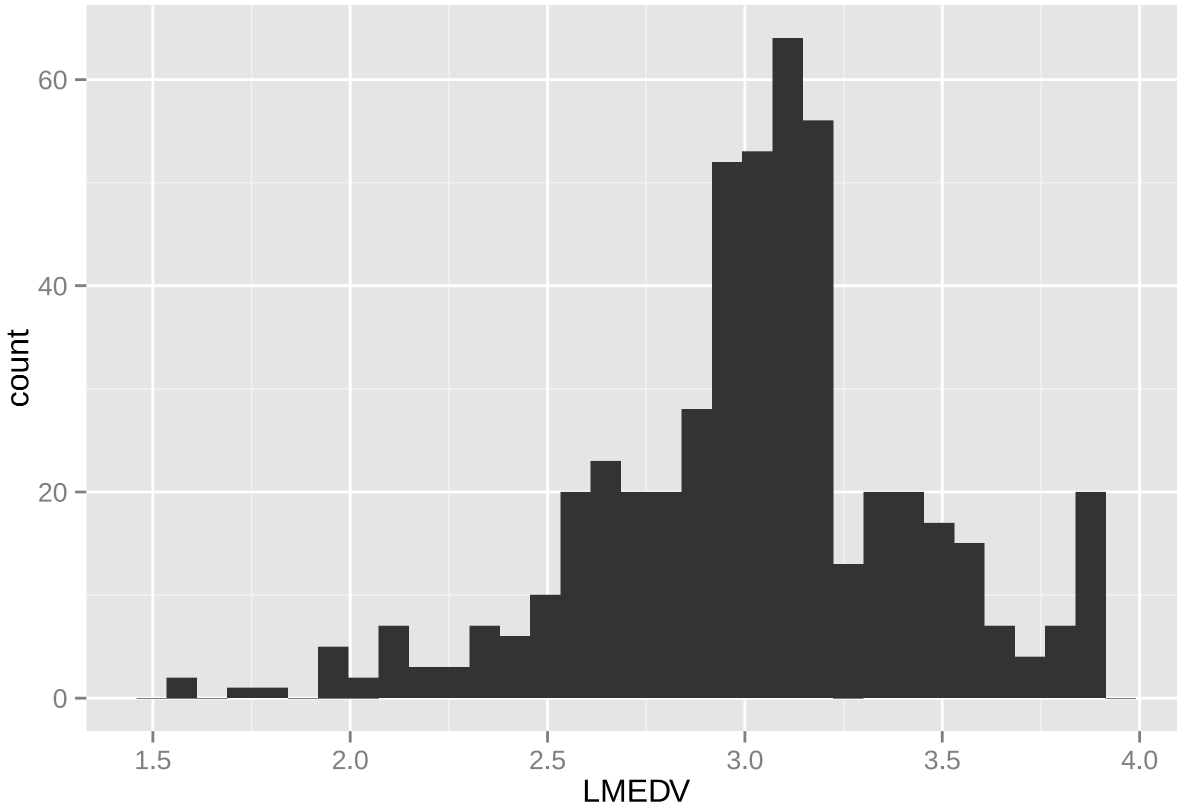
```
library(ggplot2)

# plot histogram of median value
h <- ggplot(bos, aes(MEDV)) + geom_histogram()
h
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```




```
bos$LMEDV <- log(bos$MEDV)
l <- ggplot(bos, aes(LMEDV)) + geom_histogram()
l
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```




### Questions

1. Please perform the multicollinearity diagnosis based on the VIF calculation results. Do we need to drop any variables that might have multicollinearity concerns?

```
library(car)
fit <- lm(LMEDV ~ . - MEDV, data = bos)
vif(fit)
```

```
##     CRIM       ZN    INDUS     CHAS      NOX       RM      AGE      DIS
## 1.767486 2.298459 3.987181 1.071168 4.369093 1.912532 3.088232 3.954037
##      RAD      TAX  PTRATIO    LSTAT
## 7.445301 9.002158 1.797060 2.870777
```

The `TAX` variable has a high variance inflation factor, and can be seen to be highly correlated with `RAD`, so it will be removed from the model.

```
cor(bos$TAX, bos$RAD)
```

```
## [1] 0.9102282
```

```
fit2 <- update(fit, . ~ . - TAX)
vif(fit2)
```

```
##     CRIM       ZN    INDUS     CHAS      NOX       RM      AGE      DIS
## 1.767349 2.184172 3.217951 1.055023 4.343300 1.902642 3.085756 3.952445
##      RAD  PTRATIO    LSTAT
## 2.772208 1.787049 2.870408
```

2. Please run the linear regrssion analyses.

- Use the stepwise model selection approach to determine the final model. Drop variables based on thier significance.
- Report summary output for *each step*, including ANOVA, $R^2$, and parameter estimates.

```r
step(fit2, direction = "backward")
```

```
## Start:  AIC=-1638.39
## LMEDV ~ CRIM + ZN + INDUS + CHAS + NOX + RM + AGE + DIS + RAD +
##     PTRATIO + LSTAT
##
##           Df Sum of Sq    RSS     AIC
## - AGE      1    0.0099 18.948 -1640.1
## - INDUS    1    0.0444 18.983 -1639.2
## - ZN       1    0.0572 18.996 -1638.9
## <none>                 18.938 -1638.4
## - RAD      1    0.2905 19.229 -1632.7
## - CHAS     1    0.4880 19.426 -1627.5
## - RM       1    1.0424 19.981 -1613.3
## - NOX      1    1.1881 20.126 -1609.6
## - DIS      1    1.4416 20.380 -1603.3
## - PTRATIO  1    2.0415 20.980 -1588.6
## - CRIM     1    2.4735 21.412 -1578.3
## - LSTAT    1    8.1551 27.093 -1459.2
##
## Step:  AIC=-1640.12
## LMEDV ~ CRIM + ZN + INDUS + CHAS + NOX + RM + DIS + RAD + PTRATIO +
##     LSTAT
##
##           Df Sum of Sq    RSS     AIC
## - INDUS    1    0.0439 18.992 -1641.0
## - ZN       1    0.0525 19.001 -1640.7
## <none>                 18.948 -1640.1
## - RAD      1    0.2829 19.231 -1634.6
## - CHAS     1    0.4965 19.445 -1629.0
## - RM       1    1.1287 20.077 -1612.8
## - NOX      1    1.2166 20.165 -1610.6
## - DIS      1    1.6545 20.603 -1599.8
## - PTRATIO  1    2.0320 20.980 -1590.6
## - CRIM     1    2.4750 21.423 -1580.0
## - LSTAT    1    8.9776 27.926 -1445.9
##
## Step:  AIC=-1640.95
## LMEDV ~ CRIM + ZN + CHAS + NOX + RM + DIS + RAD + PTRATIO + LSTAT
##
```

```
##           Df Sum of Sq    RSS     AIC
## - ZN       1    0.0536 19.046 -1641.5
## <none>                 18.992 -1641.0
## - RAD      1    0.2591 19.251 -1636.1
## - CHAS     1    0.4820 19.474 -1630.3
## - RM       1    1.2115 20.204 -1611.7
## - NOX      1    1.5808 20.573 -1602.5
## - DIS      1    1.6207 20.613 -1601.5
## - PTRATIO  1    2.2159 21.208 -1587.1
## - CRIM     1    2.4519 21.444 -1581.5
## - LSTAT    1    9.1833 28.175 -1443.4
##
## Step:  AIC=-1641.52
## LMEDV ~ CRIM + CHAS + NOX + RM + DIS + RAD + PTRATIO + LSTAT
##
##           Df Sum of Sq    RSS     AIC
## <none>                 19.046 -1641.5
## - RAD      1    0.2955 19.341 -1635.7
## - CHAS     1    0.4760 19.522 -1631.0
## - RM       1    1.2972 20.343 -1610.2
## - NOX      1    1.6229 20.669 -1602.2
## - DIS      1    1.7742 20.820 -1598.5
## - CRIM     1    2.4070 21.453 -1583.3
## - PTRATIO  1    2.7296 21.776 -1575.8
## - LSTAT    1    9.1687 28.215 -1444.7
```

```
##
## Call:
## lm(formula = LMEDV ~ CRIM + CHAS + NOX + RM + DIS + RAD + PTRATIO +
##     LSTAT, data = bos)
##
## Coefficients:
## (Intercept)         CRIM         CHAS          NOX           RM
##    4.255994    -0.010619     0.123795    -0.918473     0.095826
##         DIS          RAD      PTRATIO        LSTAT
##   -0.045207     0.004528    -0.042095    -0.029971
```

According to the backward step model selection, the final model should include `CRIM`, `CHAS`, `NOX`, `RM`, `DIS`, `RAD`, `PTRATIO`, and `LSTAT`.

```r
mod1 <- update(fit2, . ~ . - AGE)
mod2 <- update(mod1, . ~ . - INDUS)
mod3 <- update(mod2, . ~ . - ZN)

# analysis of variance
anova(fit2, mod1, mod2, mod3)
```

```
## Analysis of Variance Table
##
## Model 1: LMEDV ~ CRIM + ZN + INDUS + CHAS + NOX + RM + AGE + DIS + RAD +
##     PTRATIO + LSTAT
## Model 2: LMEDV ~ CRIM + ZN + INDUS + CHAS + NOX + RM + DIS + RAD + PTRATIO +
##     LSTAT
```

```
## Model 3: LMEDV ~ CRIM + ZN + CHAS + NOX + RM + DIS + RAD + PTRATIO + LSTAT
## Model 4: LMEDV ~ CRIM + CHAS + NOX + RM + DIS + RAD + PTRATIO + LSTAT
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    494 18.938
## 2    495 18.948 -1 -0.009918 0.2587 0.6112
## 3    496 18.992 -1 -0.043910 1.1454 0.2850
## 4    497 19.046 -1 -0.053634 1.3990 0.2375
```

```
fit3 <- lm(LMEDV ~ CRIM + CHAS + NOX + RM + DIS + RAD + PTRATIO + LSTAT, data = bos)
```

3. How do you interpret the final model?

```
summary(fit3)
```

```
##
## Call:
## lm(formula = LMEDV ~ CRIM + CHAS + NOX + RM + DIS + RAD + PTRATIO +
##     LSTAT, data = bos)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.68024 -0.10665 -0.01233  0.09954  0.81810
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.255994   0.199360  21.348  < 2e-16 ***
## CRIM        -0.010619   0.001340  -7.925 1.51e-14 ***
## CHAS         0.123795   0.035124   3.525 0.000463 ***
## NOX         -0.918473   0.141139  -6.508 1.87e-10 ***
## RM           0.095826   0.016470   5.818 1.07e-08 ***
## DIS         -0.045207   0.006644  -6.804 2.93e-11 ***
## RAD          0.004528   0.001630   2.777 0.005694 **
## PTRATIO     -0.042095   0.004988  -8.440 3.50e-16 ***
## LSTAT       -0.029971   0.001938 -15.468  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1958 on 497 degrees of freedom
## Multiple R-squared:  0.7743, Adjusted R-squared:  0.7706
## F-statistic: 213.1 on 8 and 497 DF,  p-value: < 2.2e-16
```

The natural log of median home value ($1000) changes by the estimated coefficient for each feature for each unit change in that feature, while all other features are held constant. To use `CRIM` as an example:

$log(MEDV)$ decreases by $0.010619 \pm 0.0013399$ for every 1 unit increase in the per capita crime rate for the same `CHAS`, `NOX`, `RM`, `DIS`, `RAD`, `PTRATIO`, and `LSTAT`.

The 95% confidence interval for each coefficient is calculated by obtaining the 0.975 t-quantile on 497 degrees of freedom multiplied by the standard error for that coefficient. This value is added and subtracted from the fitted value at each $X_i$ to get the upper and lower bounds.

Finally the adjusted $R^2$ value of 0.7706 means that 77% of the variance in the data are explained by the model.