HW7

Manuel

10/30/2019

```
setwd('/home/noble_mannu/Documents/PhD/First/STAT_2131_Applied_Statistical_Methods_I/HW7')
Data <- data.frame(read.table(file = "Boston.txt", header = T, sep = "\t", stringsAsFactors = F))
Data <- Data[,!(colnames(Data)%in%c("LSTAT","b"))]
```

Exercise 3 Homework 7

Make the multilinear regression model

```
linearMod <- lm(mvalue ~ ., data=Data)
```

Display summary of our model

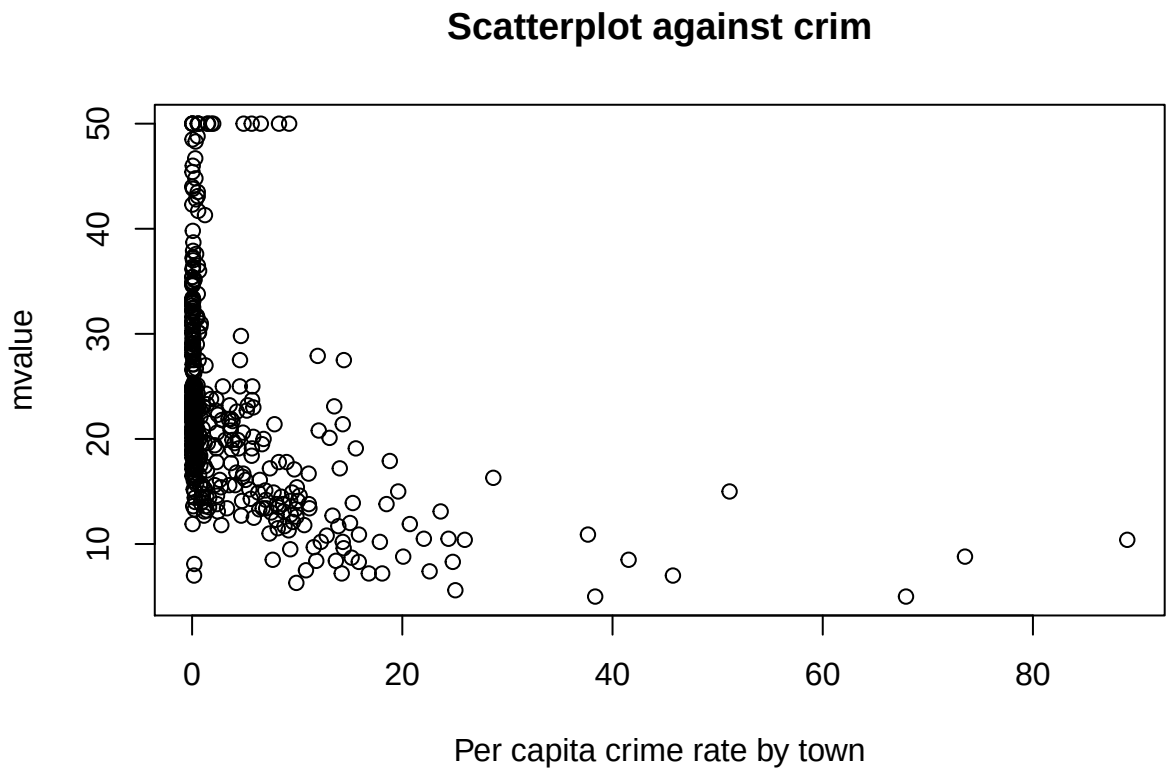
```
summary(linearMod)
```

```
##
## Call:
## lm(formula = mvalue ~ ., data = Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.846  -2.749  -0.624   1.994  37.089
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  27.152368   5.290506   5.132 4.12e-07 ***
## crim         -0.184032   0.036162  -5.089 5.12e-07 ***
## zn            0.039100   0.015424   2.535 0.011551 *
## indus        -0.042324   0.068920  -0.614 0.539425
## chas          3.487528   0.965890   3.611 0.000337 ***
## nox         -22.182110   4.271529  -5.193 3.03e-07 ***
## rooms         6.075744   0.397168  15.298  < 2e-16 ***
## age          -0.045188   0.013971  -3.234 0.001300 **
## distance     -1.583852   0.224166  -7.066 5.47e-12 ***
## radial        0.254722   0.074371   3.425 0.000666 ***
## tax          -0.012213   0.004229  -2.887 0.004053 **
## pt           -0.996206   0.146998  -6.777 3.50e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.339 on 494 degrees of freedom
## Multiple R-squared:  0.6703, Adjusted R-squared:  0.663
## F-statistic: 91.31 on 11 and 494 DF,  p-value: < 2.2e-16
```

Next we'll plot the scatterplot covariates vs mvalue and the residuals plots

PLot the response against X1

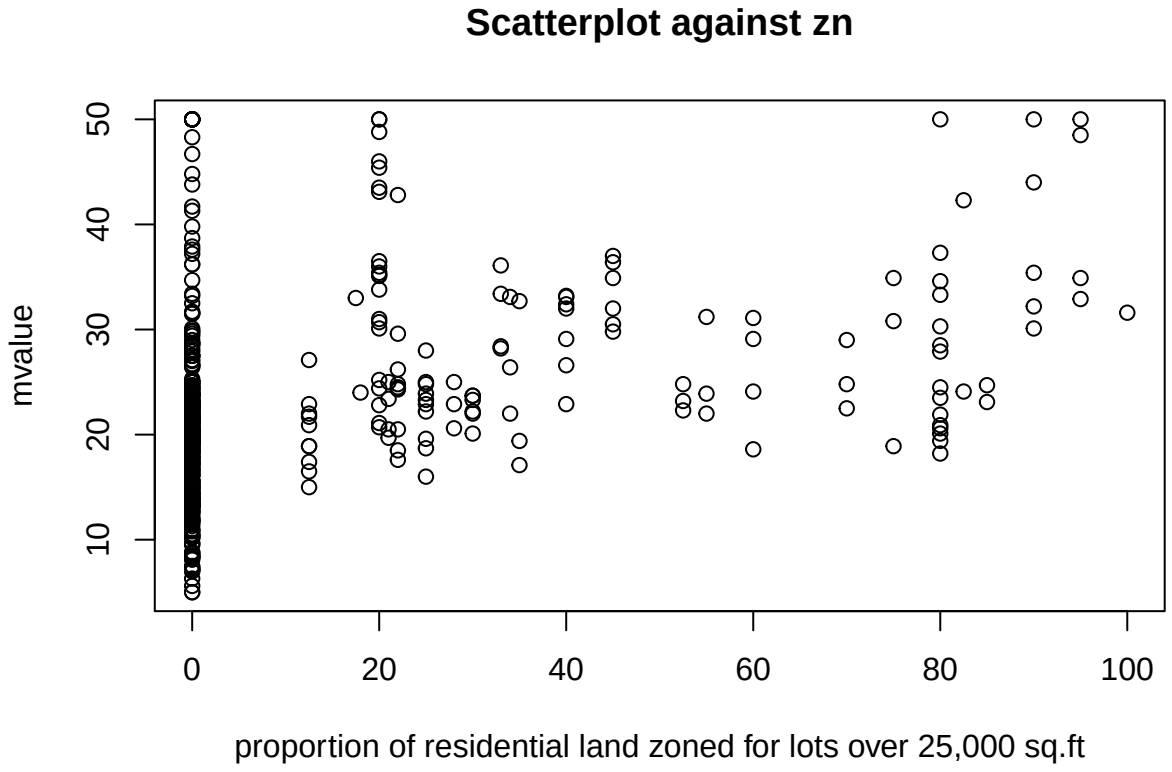
```
plot(Data$crim, Data$mvalue, xlab = "Per capita crime rate by town", ylab = "mvalue",
     main = 'Scatterplot against crim')
```



#

PLot the response against X2

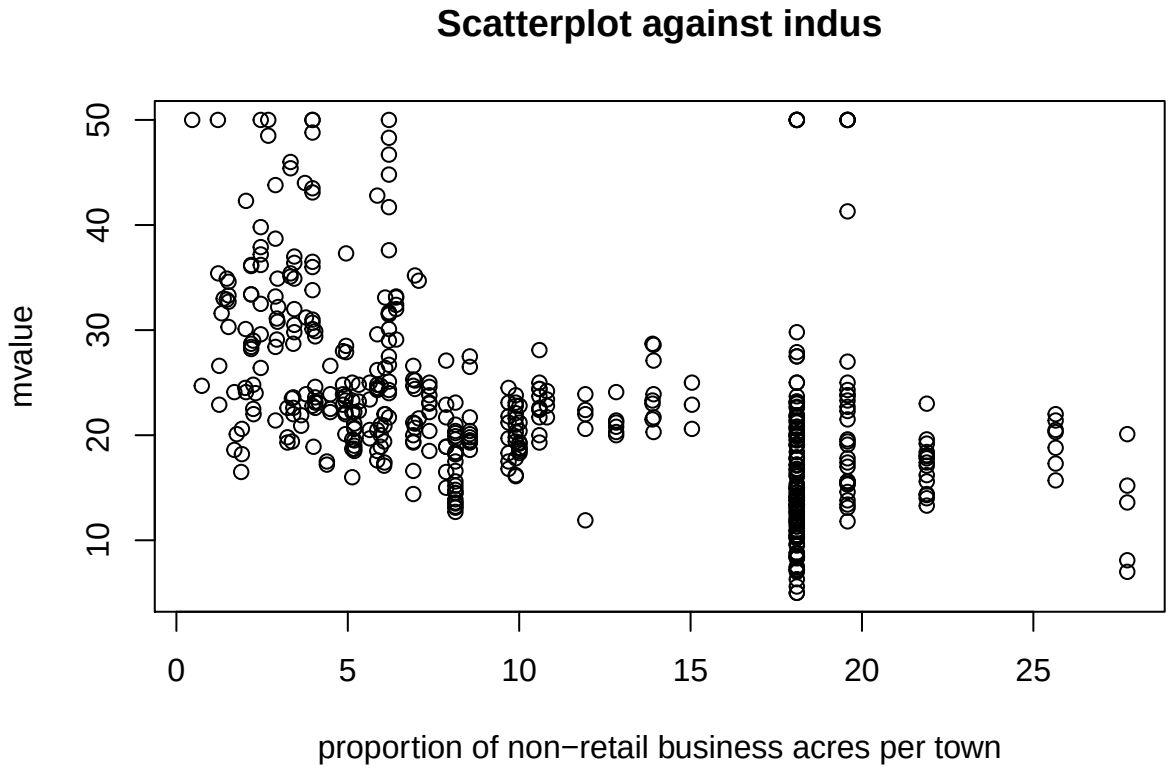
```
plot(Data$zn, Data$mvalue, xlab = "proportion of residential land zoned for lots over 25,000 sq.ft", yla
     main = 'Scatterplot against zn')
```



#

PLot the response against X3

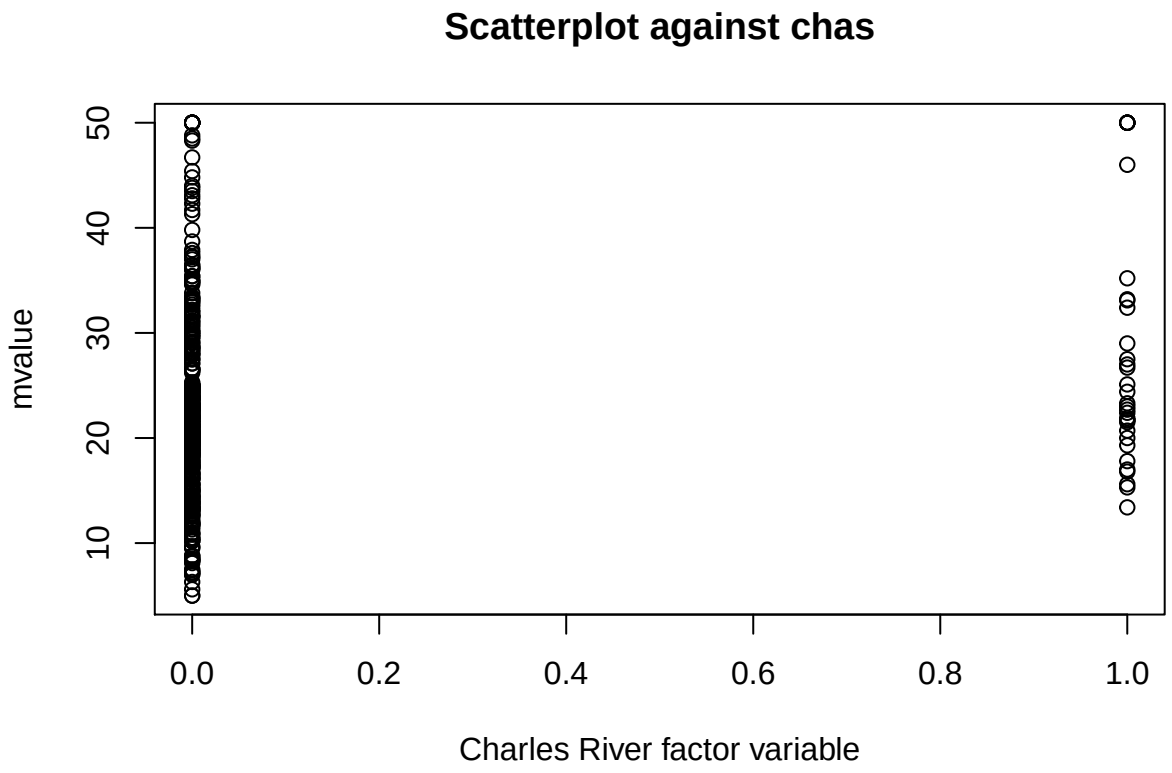
```
plot(Data$indus, Data$mvalue, xlab = "proportion of non-retail business acres per town", ylab = "mvalue
     main = 'Scatterplot against indus')
```



#

PLot the response against X4

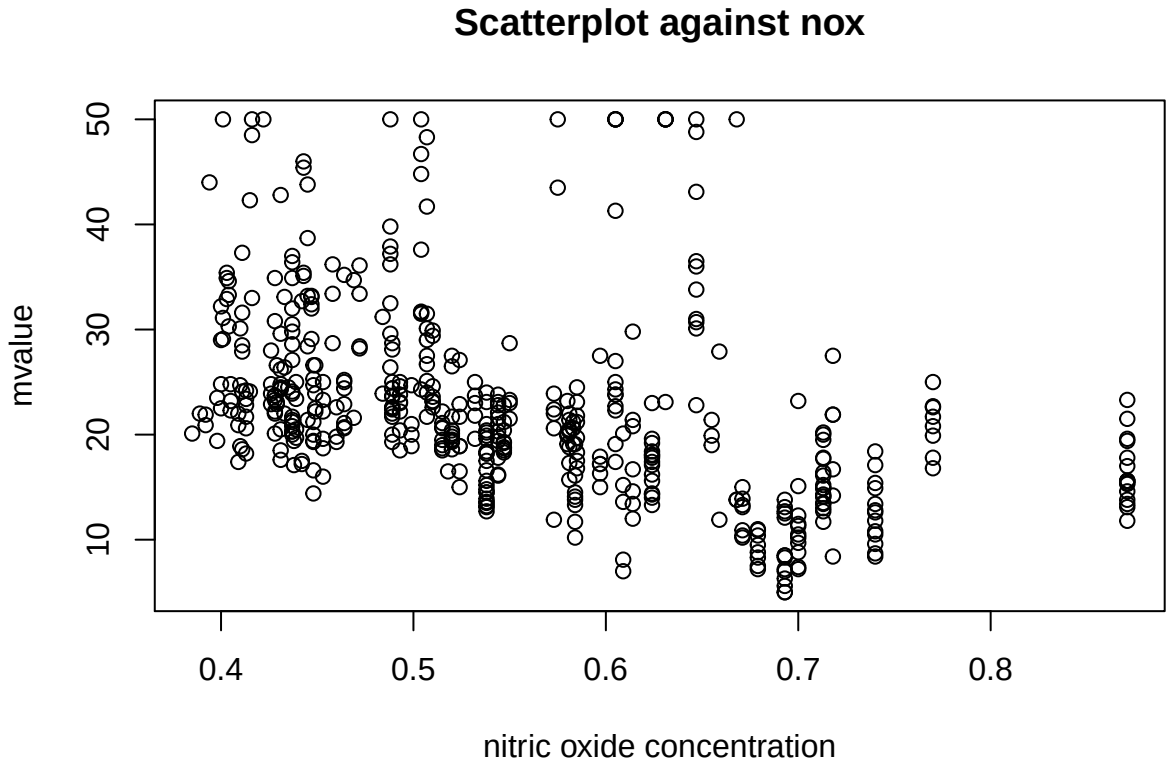
```
plot(Data$chas, Data$mvalue, xlab = "Charles River factor variable", ylab = "mvalue",
     main = 'Scatterplot against chas')
```



#

PLot the response against X5

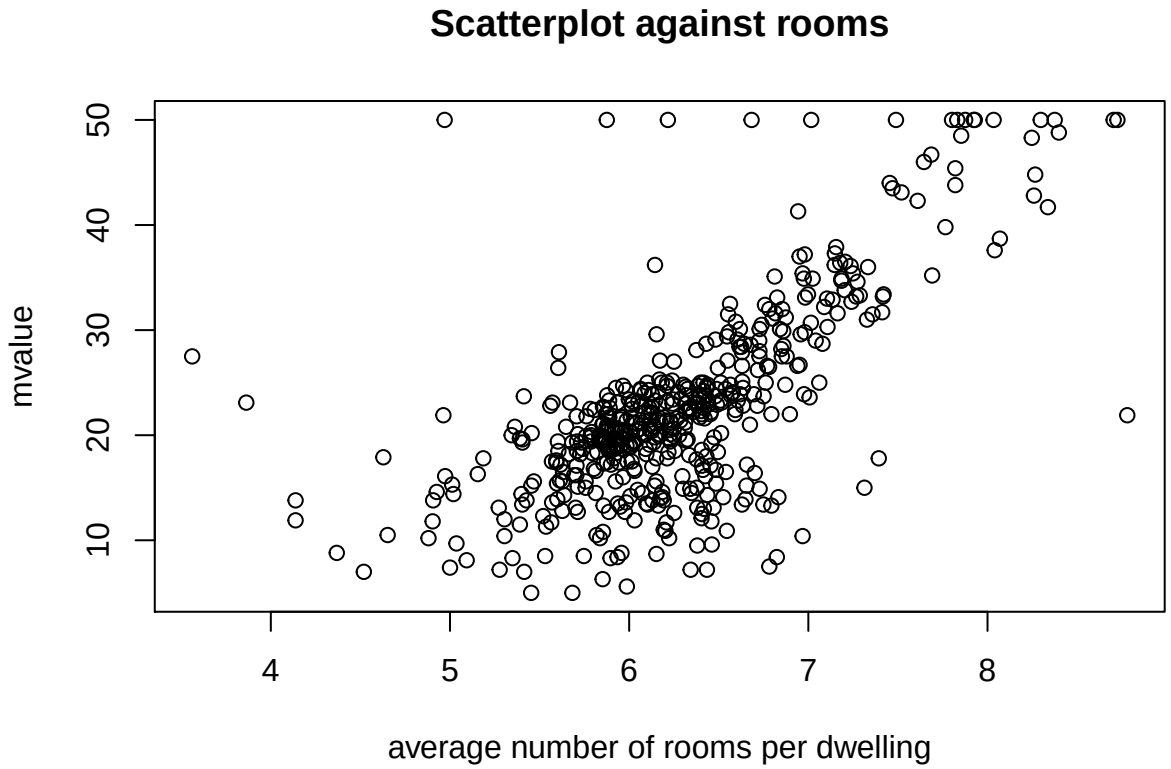
```
plot(Data$nox, Data$mvalue, xlab = "nitric oxide concentration", ylab = "mvalue",
     main = 'Scatterplot against nox')
```



#

PLot the response against X6

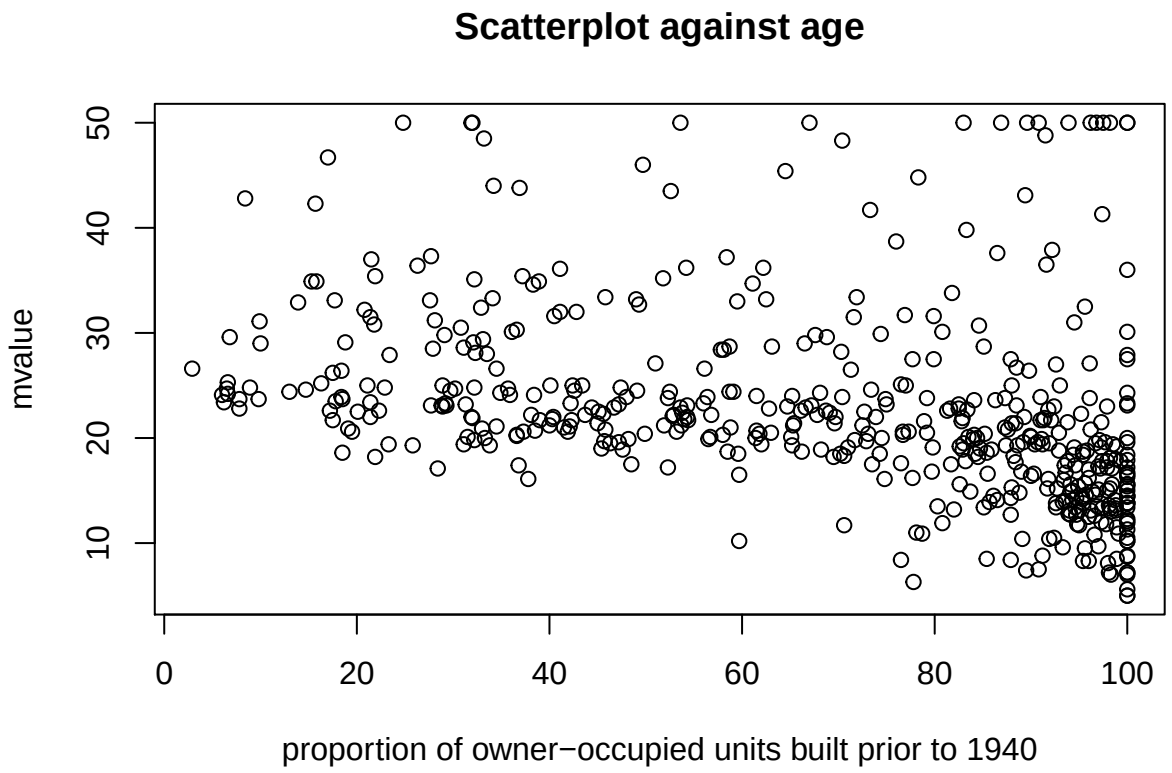
```
plot(Data$rooms, Data$mvalue, xlab = "average number of rooms per dwelling", ylab = "mvalue",
     main = 'Scatterplot against rooms')
```



#

PLot the response against X7

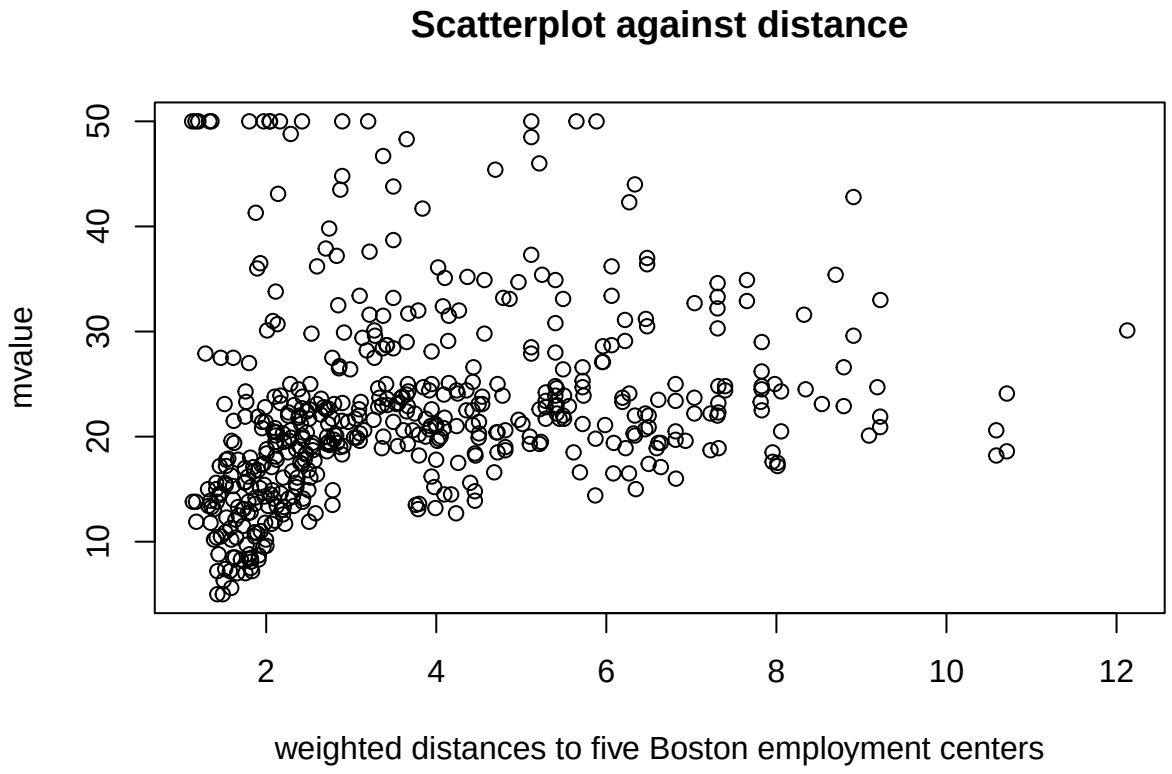
```
plot(Data$age, Data$mvalue, xlab = "proportion of owner-occupied units built prior to 1940", ylab = "mva
     main = 'Scatterplot against age')
```



#

PLot the response against X8

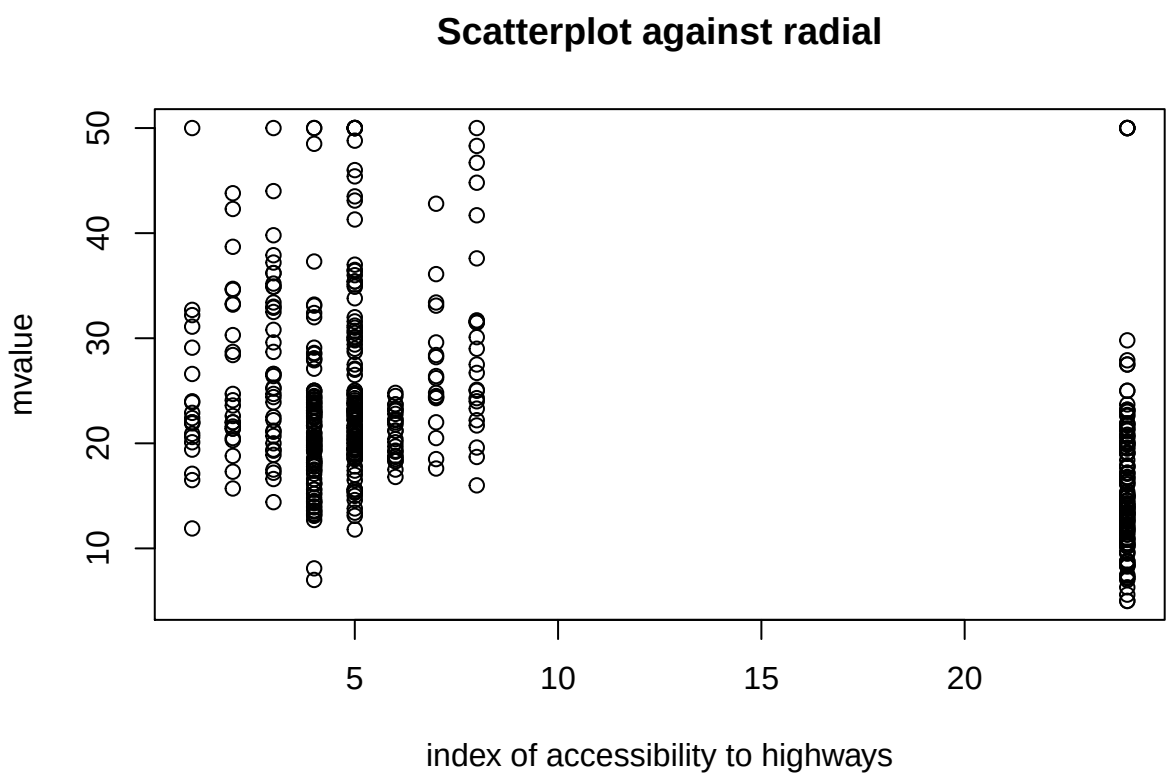
```
plot(Data$distance, Data$mvalue, xlab = "weighted distances to five Boston employment centers", ylab =
     main = 'Scatterplot against distance')
```



#

PLot the response against X9

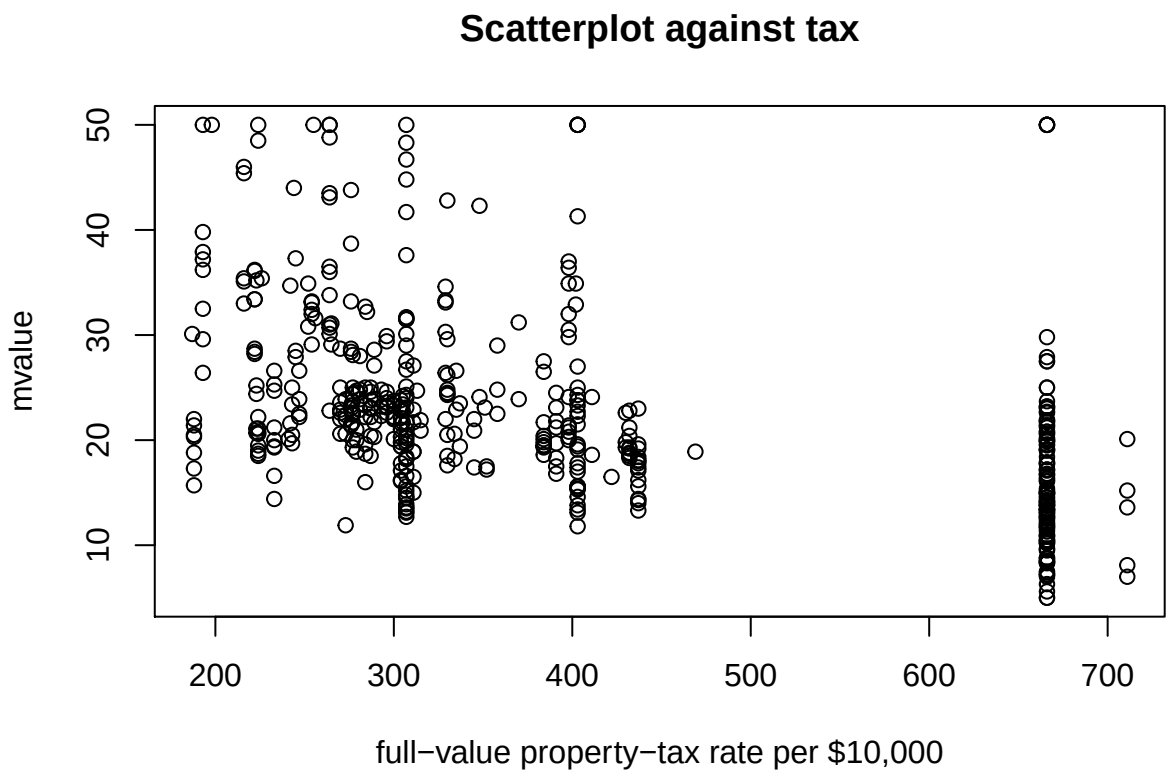
```
plot(Data$radial, Data$mvalue, xlab = "index of accessibility to highways", ylab = "mvalue",      main =
```



#

PLot the response against X10

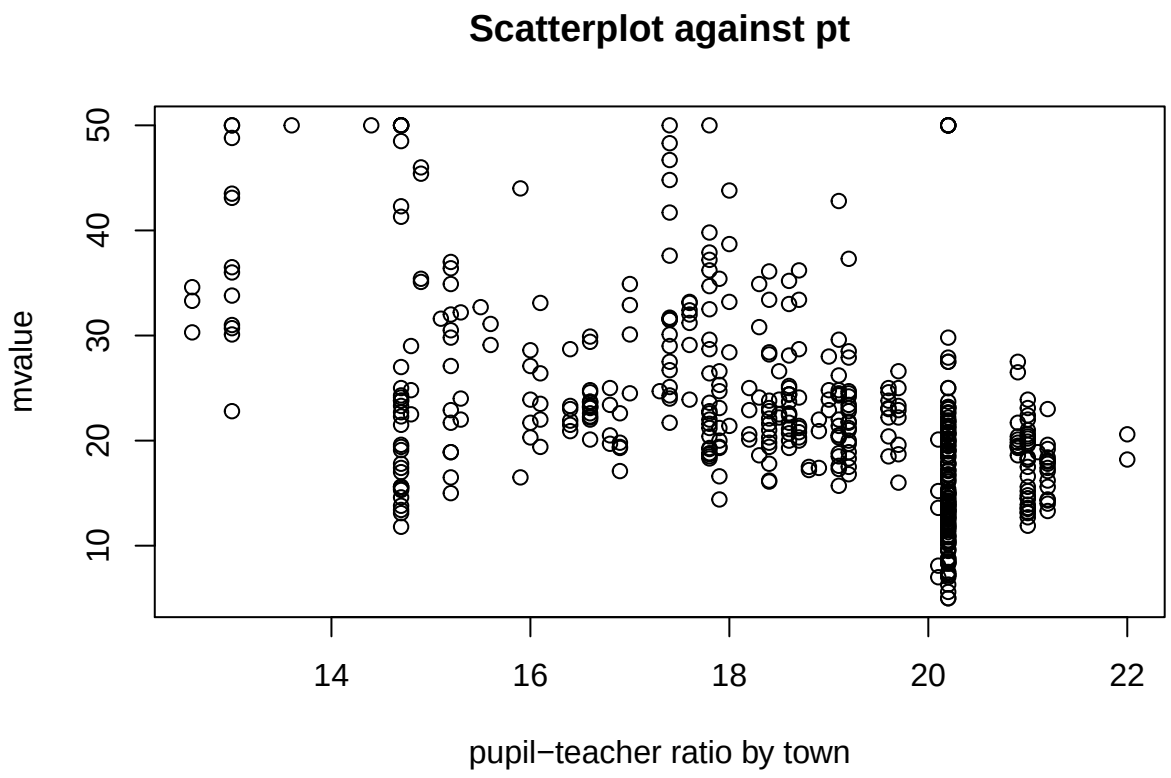
```
plot(Data$tax, Data$mvalue, xlab = "full-value property-tax rate per $10,000", ylab = "mvalue", main =
```



#

PLot the response against X11

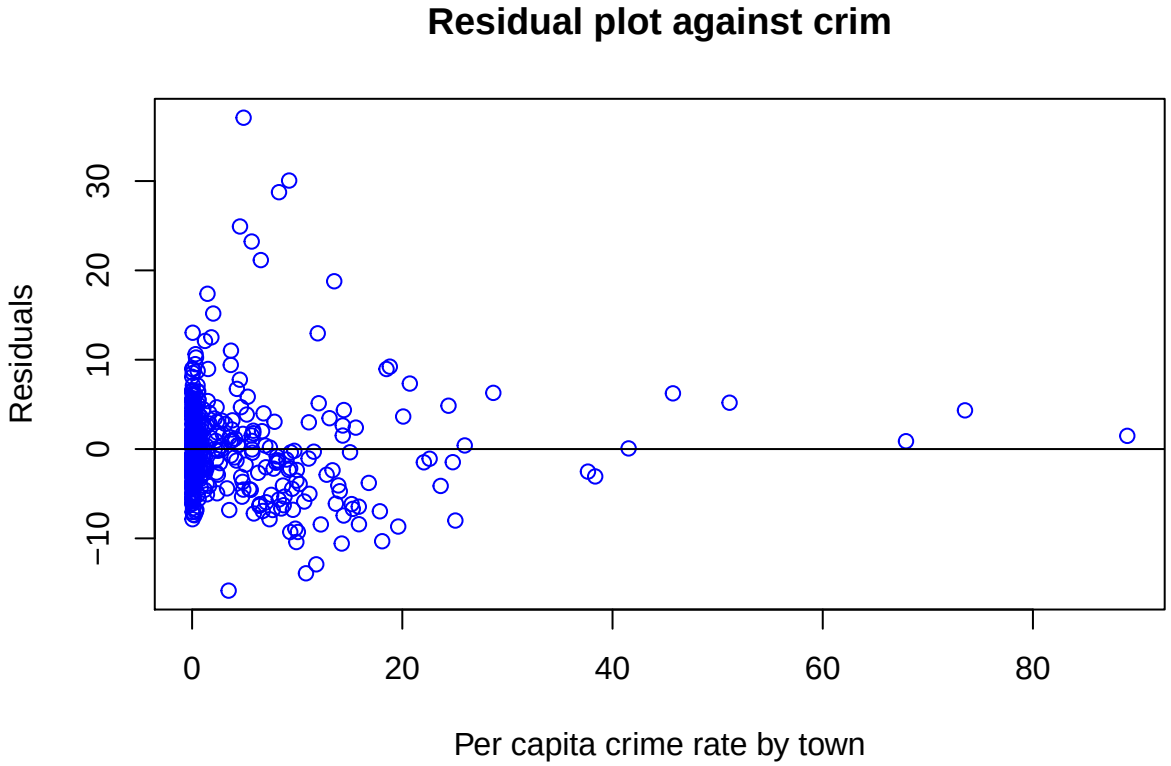
```
plot(Data$pt, Data$mvalue, xlab = "pupil-teacher ratio by town", ylab = "mvalue",
     main = 'Scatterplot against pt')
```



#

PLot the residuals against X1

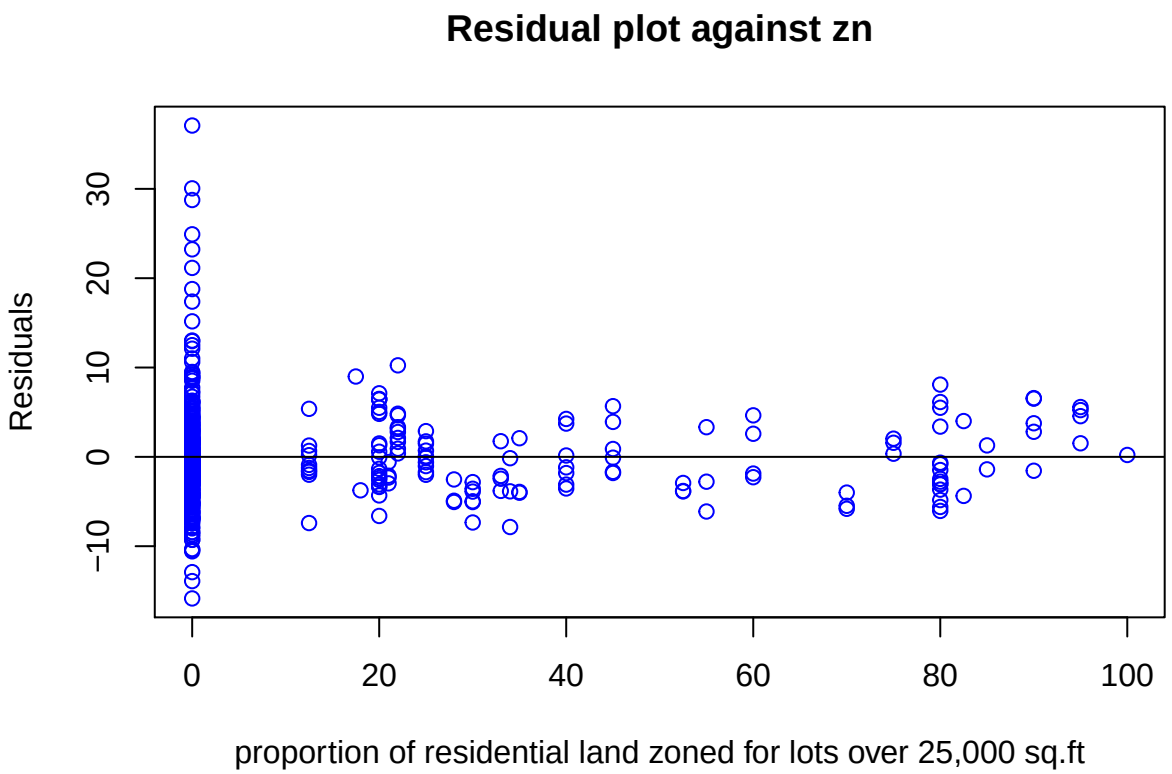
```
plot(Data$crim, resid(linearMod), xlab = "Per capita crime rate by town", ylab = "Residuals", main = 'R
abline(a=0, b=0)
```



#

PLot the residuals against X2

```
plot(Data$zn, resid(linearMod), xlab = "proportion of residential land zoned for lots over 25,000 sq.ft
abline(a=0, b=0)
```



#

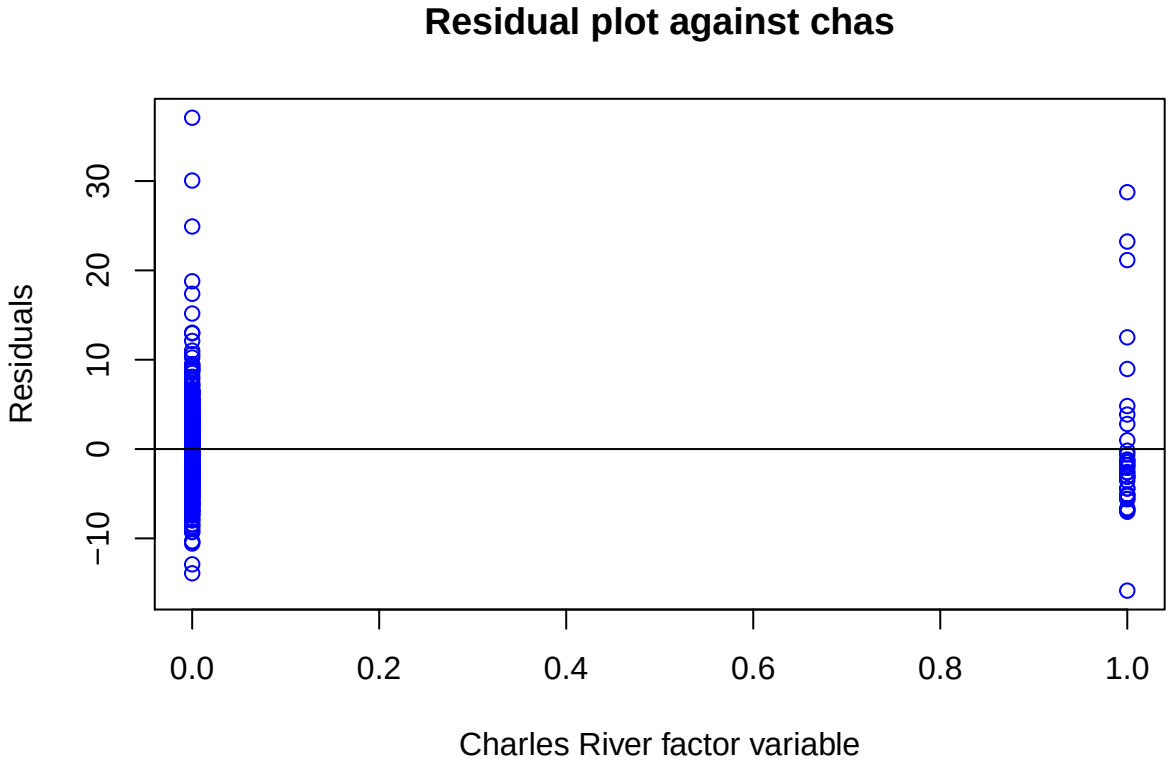
PLot the residuals against X3

```
plot(Data$indus, resid(linearMod), xlab = "proportion of non-retail business acres per town", ylab = "R
     main = 'Residual plot against indus', col = 'blue')
abline(a=0, b=0)
```

#

PLot the residuals against X4

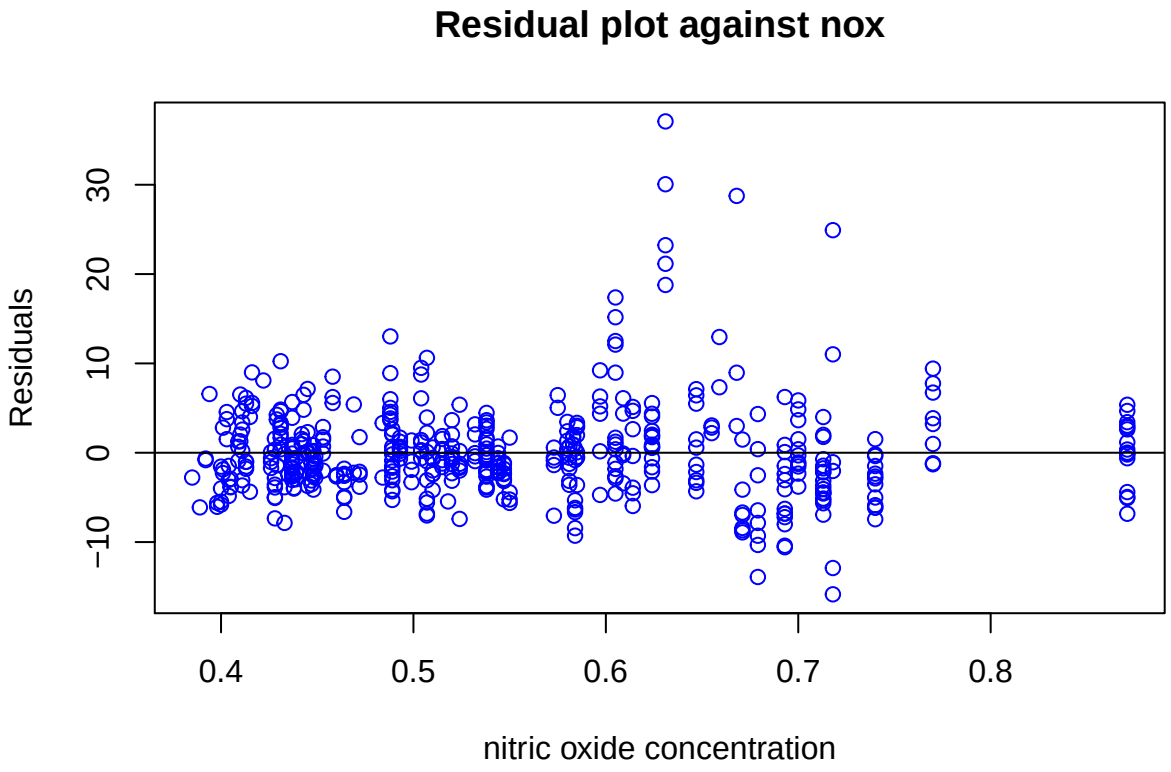
```
plot(Data$chas, resid(linearMod), xlab = "Charles River factor variable", ylab = "Residuals", main = 'R
abline(a=0, b=0)
```



#

PLot the residuals against X5

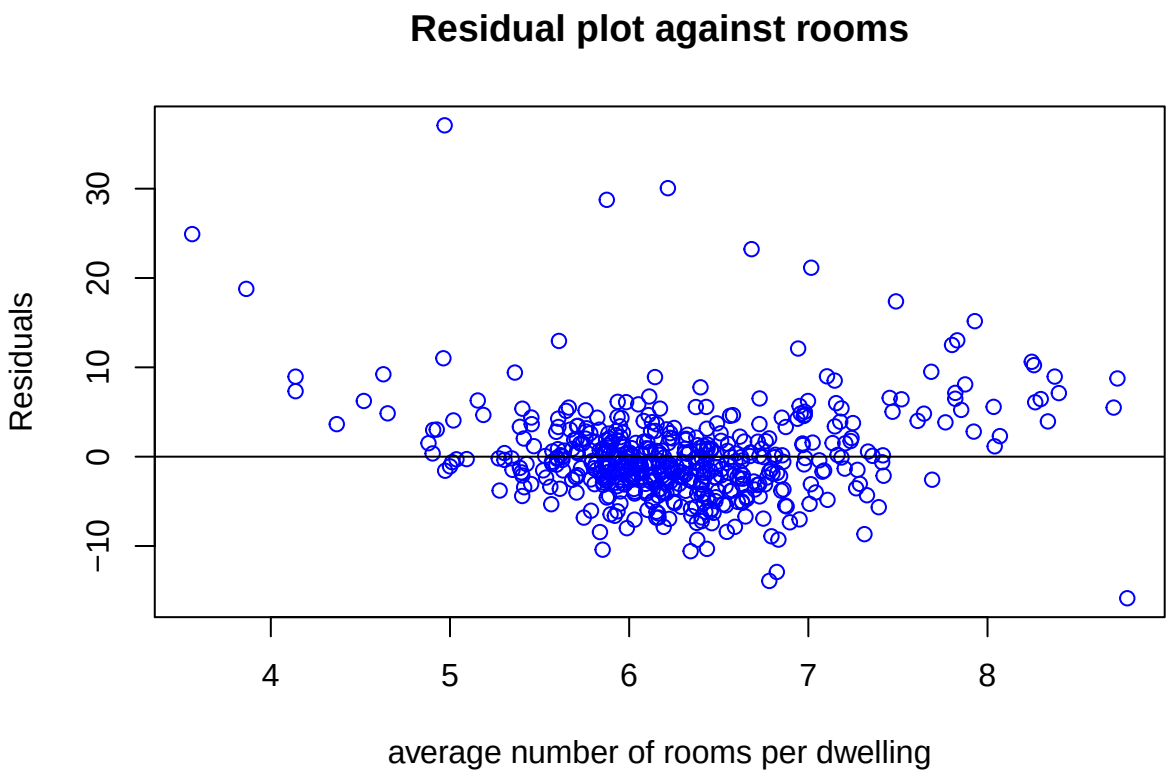
```
plot(Data$nox, resid(linearMod), xlab = "nitric oxide concentration", ylab = "Residuals",
     main = 'Residual plot against nox', col = 'blue')
abline(a=0, b=0)
```



#

PLot the residuals against X6

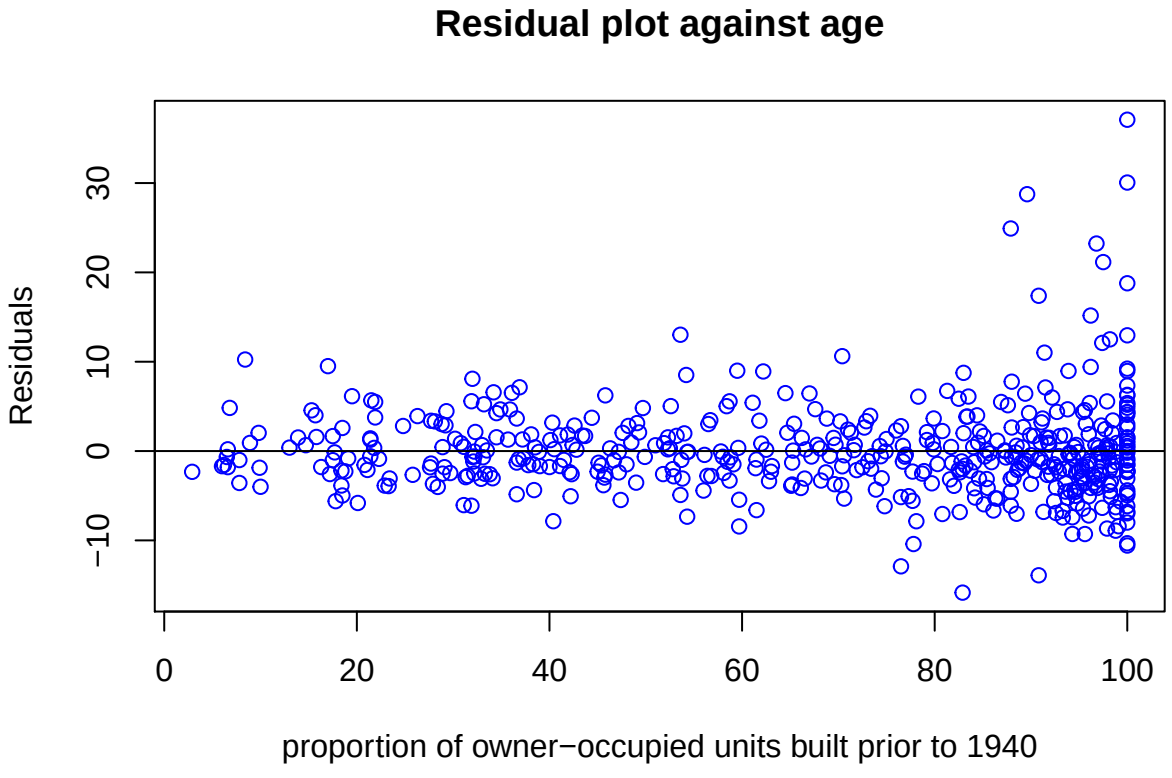
```
plot(Data$rooms, resid(linearMod), xlab = "average number of rooms per dwelling", ylab = "Residuals", ma
abline(a=0, b=0)
```



#

PLot the residuals against X7

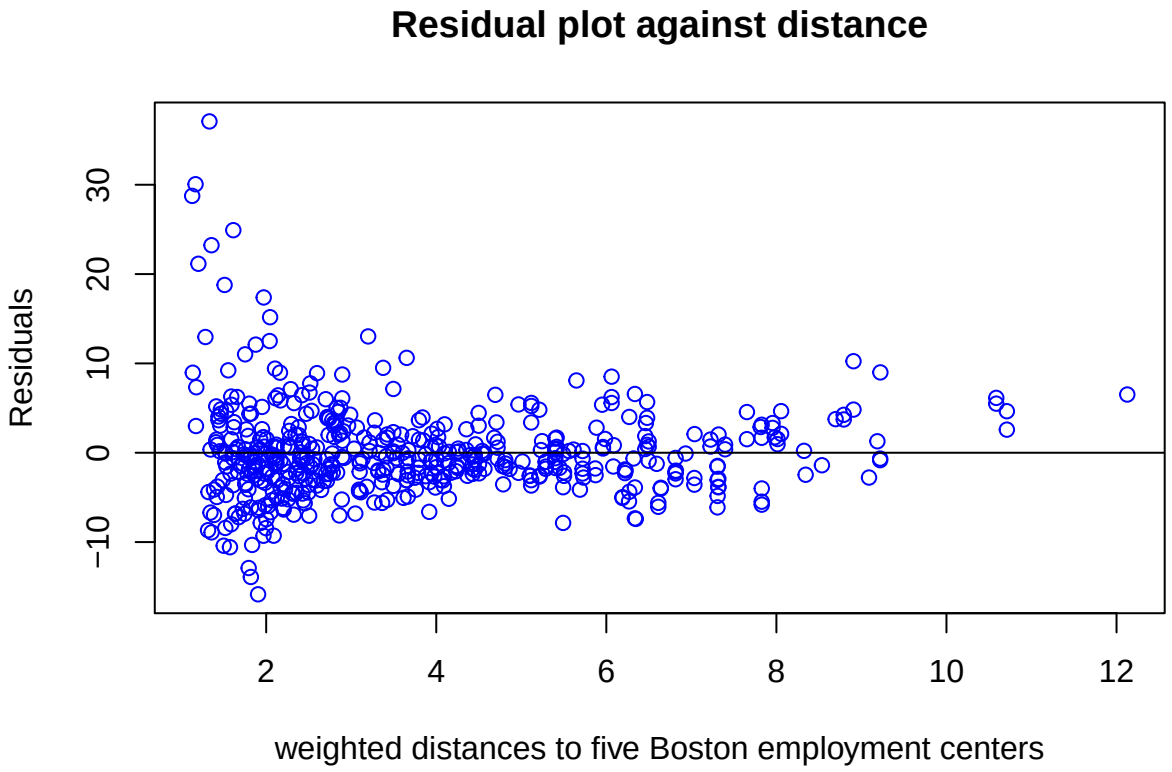
```
plot(Data$age, resid(linearMod), xlab = "proportion of owner-occupied units built prior to 1940", ylab =
abline(a=0, b=0)
```



#

PLot the residuals against X8

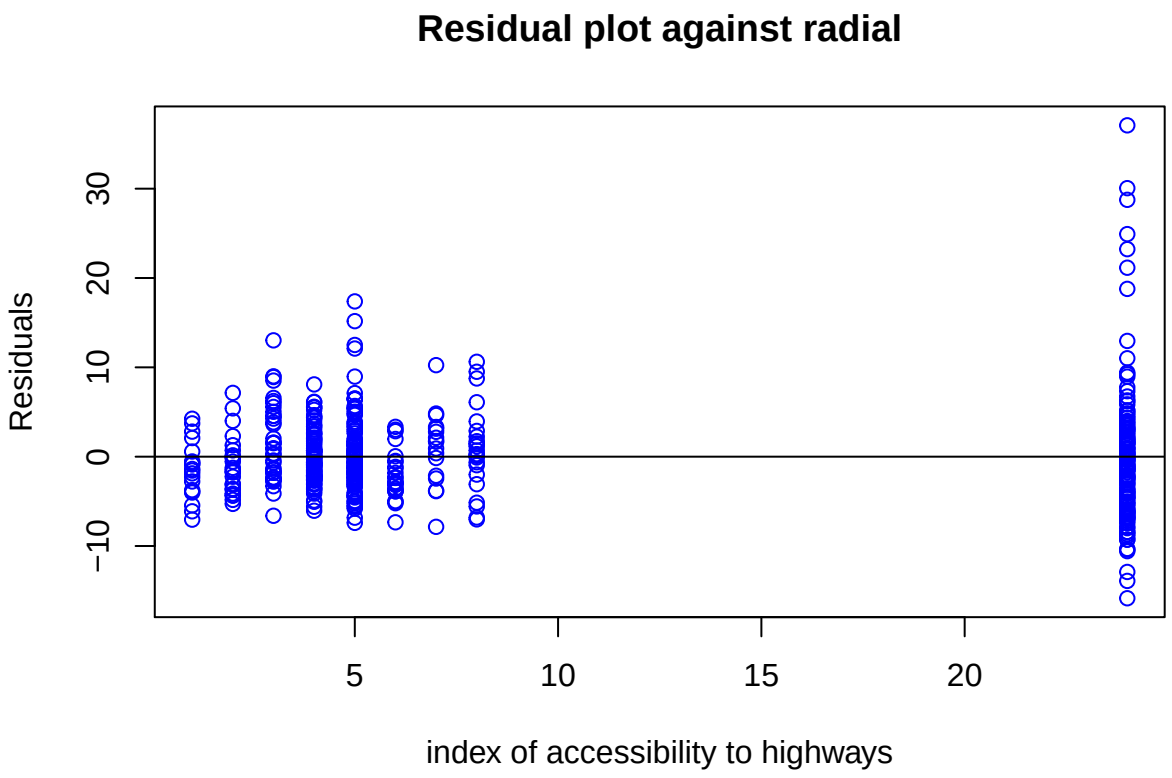
```
plot(Data$distance, resid(linearMod), xlab = "weighted distances to five Boston employment centers", yla
     main = 'Residual plot against distance', col = 'blue')
abline(a=0, b=0)
```



#

PLot the residuals against X9

```
plot(Data$radial, resid(linearMod), xlab = "index of accessibility to highways", ylab = "Residuals", ma
abline(a=0, b=0)
```



#

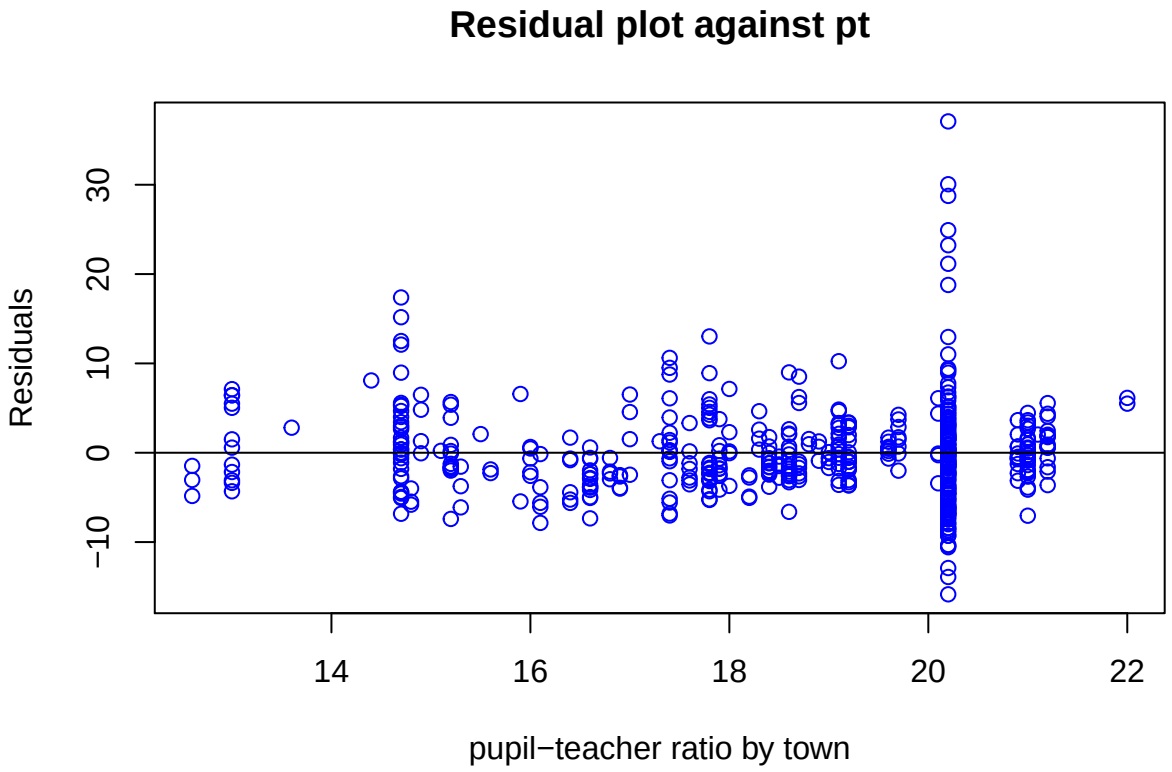
PLot the residuals against X10

```
plot(Data$tax, resid(linearMod), xlab = "full-value property-tax rate per $10,000", ylab = "Residuals",
abline(a=0, b=0)
```

#

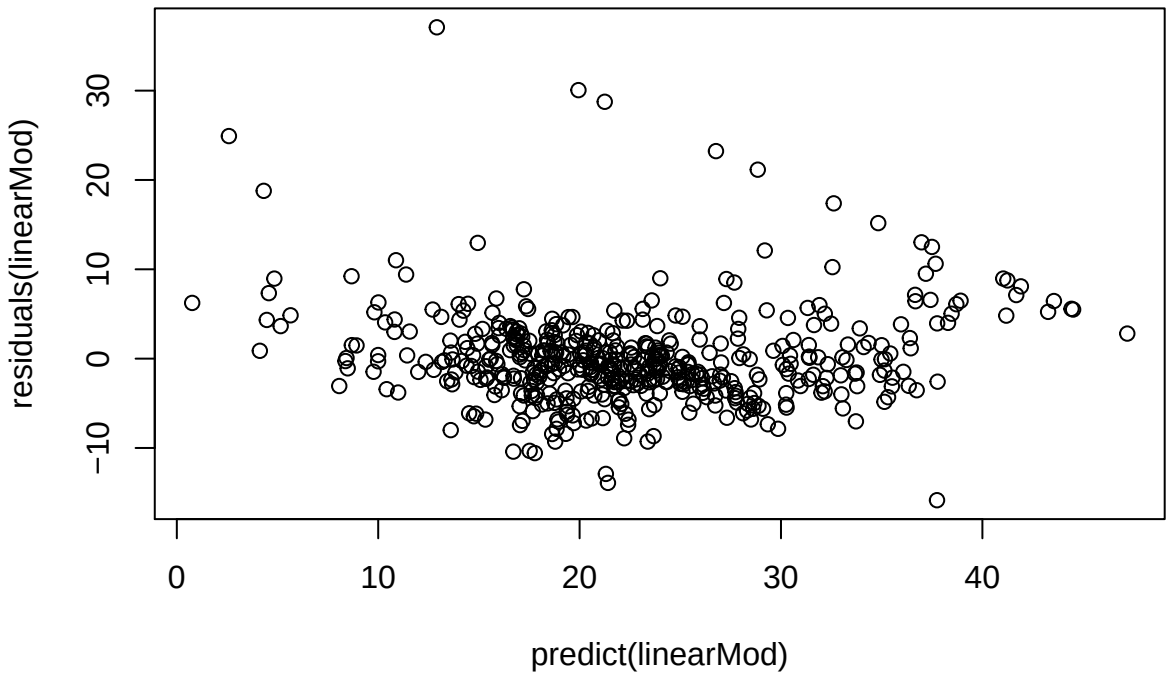
PLot the residuals against X11

```
plot(Data$pt, resid(linearMod), xlab = "pupil-teacher ratio by town", ylab = "Residuals",
     main = 'Residual plot against pt', col = 'blue')
abline(a=0, b=0)
```



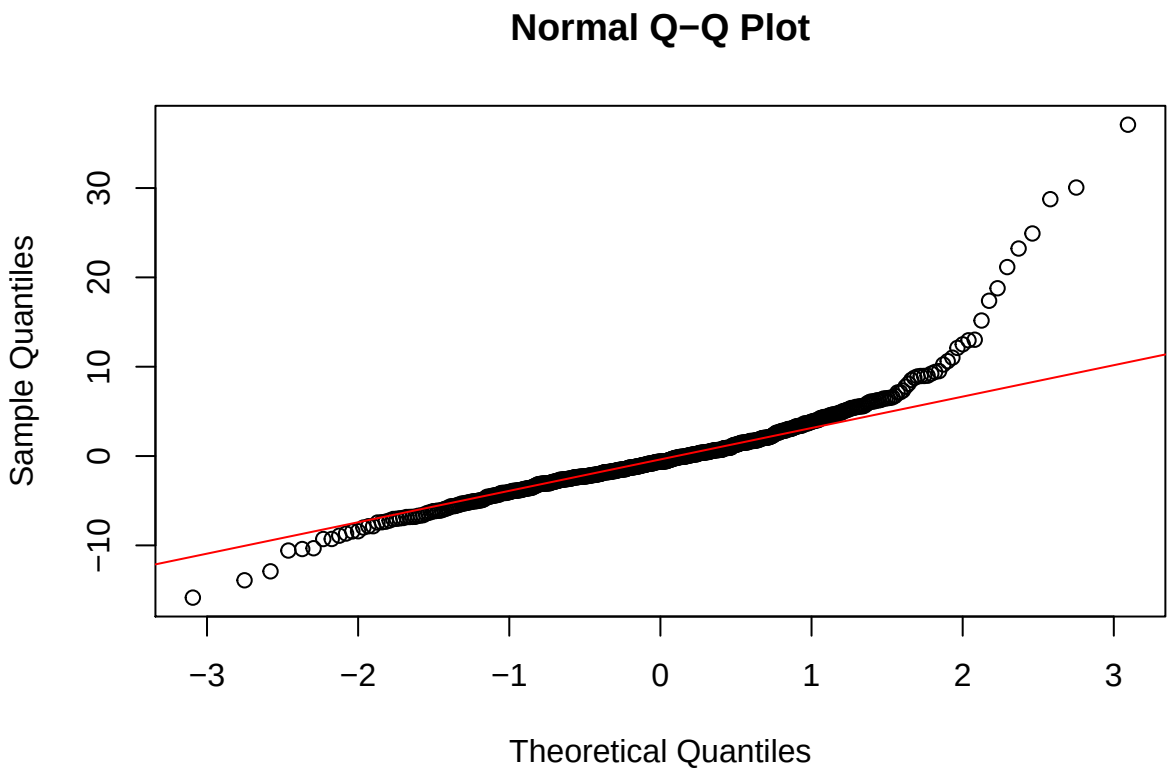
Predicted vs residuals

```
plot(predict(linearMod), residuals(linearMod))
```



Plotting the Q-Q Plot

```
qqnorm(residuals(linearMod)); qqline(residuals(linearMod), col="red")
```



##

Fitting the full model and a proposed model ## # Make the full model seen in class

```
linearMod <- lm(mvalue ~ ., data=Data)
```

Make the improved model seen in class

This is the improved model we saw in class, where we dropped predictor 'indus'

```
linearMod1 <- lm(mvalue ~ crim+zn+chas+nox+rooms+age+distance+radial+tax+pt, data=Data)
```

Make the proposed model (Based on residual plots and scatterplots)

This is my proposed model. For this one I added the square of indus, crim and distance

```
test <- lm(mvalue ~ .+I(indus^2)+I(crim^2)+I(distance^2), data=Data)
```

Now we'll run the best subset selection for the Original model and the Proposed model. We actually did this for the Original model in class and came up with the model were we dropped 'indus'. Nevertheless, I ran the code for this model again in addition to the one for my Proposed model.

Running best subset selection for both models

Run best subset selection for the original full model. This may take a minute or two...

```
# Run best subset selection for the original model. This takes about 2 mins.
Best.subset <- olsrr::ols_step_best_subset(linearMod)
# Run best subset selection for the proposed model. This took my computer about 20-30. mins...
Best.subset.test <- olsrr::ols_step_best_subset(test)
```

Comparing models Using R2 and adjusted R2

```
# Choosing the model based on R2 (Original model)
which.max(Best.subset$rsquare)
```

```
## [1] 11
```

```
# Returns row 11, this corresponds to the model with all predictors
# Prints the names of the predictors used in the best model with R2 criteria
Best.subset$predictors[which.max(Best.subset$rsquare)]
```

```
## [1] "crim zn indus chas nox rooms age distance radial tax pt"
```

```
# Choosing the model based on adjusted R2 (Original model)
which.max(Best.subset$adjr)
```

```
## [1] 10
```

```
# Returns row 10, this corresponds to the model with all predictors except indus
# Prints the names of the predictors used in the best model with adjusted R2 criteria
Best.subset$predictors[which.max(Best.subset$adjr)]
```

```
## [1] "crim zn chas nox rooms age distance radial tax pt"
```

```
# Choosing the model based on R2 (Proposed model)
which.max(Best.subset.test$rsquare)
```

```
## [1] 14
```

```
# Returns row 14, this corresponds to the model with all predictors
# Prints the names of the predictors used in the best model (Proposed model) with R2 criteria
Best.subset.test$predictors[which.max(Best.subset.test$rsquare)]
```

```
## [1] "crim zn indus chas nox rooms age distance radial tax pt I(indus^2) I(crim^2) I(distance^2)"
```

```
# Choosing the model based on adjusted R2 (Proposed model)
which.max(Best.subset.test$adjr)
```

```
## [1] 13
```

```
# Returns row 13, this corresponds to the model with all predictors except zn
# Prints the names of the predictors used in the best model (Proposed model) with adjusted R2 criteria
Best.subset.test$predictors[which.max(Best.subset.test$adjr)]
```

```
## [1] "crim indus chas nox rooms age distance radial tax pt I(indus^2) I(crim^2) I(distance^2)"
```

Comparing best subset of Original model vs best subset of Proposed model with R^2 criteria

```
r_origin <- Best.subset$rsquare[which.max(Best.subset$rsquare)]
r_prop <- Best.subset.test$rsquare[which.max(Best.subset.test$rsquare)]
r_prop > r_origin
```

```
## [1] TRUE
```

```
# This indicates the Proposed model is an improvement over the Original model using R^2 criteria
```

Comparing best subset of Original model vs best subset of Proposed model with adjusted R^2 criteria

```
r_adj_origin <- Best.subset$adjr[which.max(Best.subset$adjr)]
r_adj_prop <- Best.subset.test$adjr[which.max(Best.subset.test$adjr)]
r_adj_prop > r_adj_origin
```

```
## [1] TRUE
```

```
# This indicates the Proposed model is an improvement over the Original model using adjusted R^2 criter
```

Comparing models Using AIC and BIC

```
# Choosing the model based on AIC (Original model)
which.min(Best.subset$aic)
```

```
## [1] 10
```

```
# Returns row 10, this corresponds to the model with all predictors exceptc indus
# Prints the names of the predictors used in the best model with AIC criteria
Best.subset$predictors[which.min(Best.subset$aic)]
```

```
## [1] "crim zn chas nox rooms age distance radial tax pt"
```

```
# Choosing the model based on AIC (Proposed model)
which.min(Best.subset.test$aic)
```

```
## [1] 13
```

```
# Returns row 13, this corresponds to the model with all predictors exceptc zn
# Prints the names of the predictors used in the best model with AIC criteria
Best.subset.test$predictors[which.min(Best.subset.test$aic)]
```

```
## [1] "crim indus chas nox rooms age distance radial tax pt I(indus^2) I(crim^2) I(distance^2)"
```

Comparing best subset of Original model vs best subset of Proposed model with IAC criteria

```
aic_origin <- Best.subset$aic[which.min(Best.subset$aic)]
aic_prop <- Best.subset.test$aic[which.min(Best.subset.test$aic)]
aic_origin > aic_prop
```

```
## [1] TRUE
```

```
# This indicates the Proposed model is an improvement over the Original model using AIC criteria
```

```
# Choosing the model based on BIC (Original model)
which.min(Best.subset$sbc)
```

```
## [1] 10
```

```
# Returns row 10, this corresponds to the model with all predictors exceptc indus
# Prints the names of the predictors used in the best model with BIC criteria
Best.subset$predictors[which.min(Best.subset$sbc)]
```

```
## [1] "crim zn chas nox rooms age distance radial tax pt"
```

```
# Choosing the model based on BIC (Proposed model)
which.min(Best.subset.test$sbc)
```

```
## [1] 11
```

```
# Returns row 11, this corresponds to the model with the model that omits zn, indus and indus^2
# Prints the names of the predictors used in the best model with BIC criteria
Best.subset.test$predictors[which.min(Best.subset.test$sbc)]
```

```
## [1] "crim chas nox rooms age distance radial tax pt I(crim^2) I(distance^2)"
```

Comparing best subset of Original model vs best subset of Proposed model with BIC criteria

```
bic_origin <- Best.subset$sbc[which.min(Best.subset$sbc)]
bic_prop <- Best.subset.test$sbc[which.min(Best.subset.test$sbc)]
bic_origin > bic_prop
```

```
## [1] TRUE
```

```
# This indicates the Proposed model is an improvement over the Original model using BIC criteria
```