Exercise 1- Advanced Methods for Regression and Classification

12433732 - Stefan Merdian

2024-10-18

Get the data

```
if(!require(ISLR)) install.packages("ISLR",repos = "http://cran.us.r-project.org")
```

```
## Loading required package: ISLR
```

```
data(College,package="ISLR")
```

Look into the dataset

```
 head(College)
```

```
##                              Private Apps Accept Enroll Top10perc Top25perc
## Abilene Christian University     Yes 1660   1232    721        23        52
## Adelphi University               Yes 2186   1924    512        16        29
## Adrian College                   Yes 1428   1097    336        22        50
## Agnes Scott College              Yes  417    349    137        60        89
## Alaska Pacific University        Yes  193    146     55        16        44
## Albertson College                Yes  587    479    158        38        62
##                              F.Undergrad P.Undergrad Outstate Room.Board Books
## Abilene Christian University        2885         537     7440       3300   450
## Adelphi University                  2683        1227    12280       6450   750
## Adrian College                      1036          99    11250       3750   400
## Agnes Scott College                  510          63    12960       5450   450
## Alaska Pacific University            249         869     7560       4120   800
## Albertson College                    678          41    13500       3335   500
##                              Personal PhD Terminal S.F.Ratio perc.alumni Expend
## Abilene Christian University     2200  70       78      18.1          12   7041
## Adelphi University               1500  29       30      12.2          16  10527
## Adrian College                   1165  53       66      12.9          30   8735
## Agnes Scott College               875  92       97       7.7          37  19016
## Alaska Pacific University        1500  76       72      11.9           2  10922
## Albertson College                 675  67       73       9.4          11   9727
##                              Grad.Rate
## Abilene Christian University        60
## Adelphi University                  56
## Adrian College                      54
## Agnes Scott College                 59
## Alaska Pacific University           15
## Albertson College                   55
```

```
 str(College)
```

```
## 'data.frame':    777 obs. of  18 variables:
##  $ Private    : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Apps       : num  1660 2186 1428 417 193 ...
```

```
##  $ Accept     : num  1232 1924 1097 349 146 ...
##  $ Enroll     : num  721 512 336 137 55 158 103 489 227 172 ...
##  $ Top10perc  : num  23 16 22 60 16 38 17 37 30 21 ...
##  $ Top25perc  : num  52 29 50 89 44 62 45 68 63 44 ...
##  $ F.Undergrad: num  2885 2683 1036 510 249 ...
##  $ P.Undergrad: num  537 1227 99 63 869 ...
##  $ Outstate   : num  7440 12280 11250 12960 7560 ...
##  $ Room.Board : num  3300 6450 3750 5450 4120 ...
##  $ Books      : num  450 750 400 450 800 500 500 450 300 660 ...
##  $ Personal   : num  2200 1500 1165 875 1500 ...
##  $ PhD        : num  70 29 53 92 76 67 90 89 79 40 ...
##  $ Terminal   : num  78 30 66 97 72 73 93 100 84 41 ...
##  $ S.F.Ratio  : num  18.1 12.2 12.9 7.7 11.9 9.4 11.5 13.7 11.3 11.5 ...
##  $ perc.alumni: num  12 16 30 37 2 11 26 37 23 15 ...
##  $ Expend     : num  7041 10527 8735 19016 10922 ...
##  $ Grad.Rate  : num  60 56 54 59 15 55 63 73 80 52 ...
```

```
dim(College)
```

```
## [1] 777  18
```

##Data preprocessing

If there is na values, it will be removed

```
if (sum(colSums(is.na(College))) > 0) {
  College <- na.omit(College)
}
```

Since our Goal is find a linear regression model which allows to predict the variable **Apps**, based on remaining variables except of the variables **Accept** and **Enroll**. First we will remove those rows, as well the **Apps** coloumn in a separated var as the prediction variable.

```
df <- data.frame(College)
head(College)
```

```
##                              Private Apps Accept Enroll Top10perc Top25perc
## Abilene Christian University     Yes 1660   1232    721        23        52
## Adelphi University               Yes 2186   1924    512        16        29
## Adrian College                   Yes 1428   1097    336        22        50
## Agnes Scott College              Yes  417    349    137        60        89
## Alaska Pacific University        Yes  193    146     55        16        44
## Albertson College                Yes  587    479    158        38        62
##                              F.Undergrad P.Undergrad Outstate Room.Board Books
## Abilene Christian University        2885         537     7440       3300   450
## Adelphi University                  2683        1227    12280       6450   750
## Adrian College                      1036          99    11250       3750   400
## Agnes Scott College                  510          63    12960       5450   450
## Alaska Pacific University            249         869     7560       4120   800
## Albertson College                    678          41    13500       3335   500
##                              Personal PhD Terminal S.F.Ratio perc.alumni Expend
## Abilene Christian University     2200  70       78      18.1          12   7041
## Adelphi University               1500  29       30      12.2          16  10527
## Adrian College                   1165  53       66      12.9          30   8735
## Agnes Scott College               875  92       97       7.7          37  19016
## Alaska Pacific University        1500  76       72      11.9           2  10922
## Albertson College                 675  67       73       9.4          11   9727
##                              Grad.Rate
## Abilene Christian University        60
## Adelphi University                  56
## Adrian College                      54
## Agnes Scott College                 59
## Alaska Pacific University           15
## Albertson College                   55
```

```r
predict_value <- df$Apps

df$Apps <- NULL
df$Accept <- NULL
df$Enroll <- NULL

dim(df)
```

```
## [1] 777  15
```

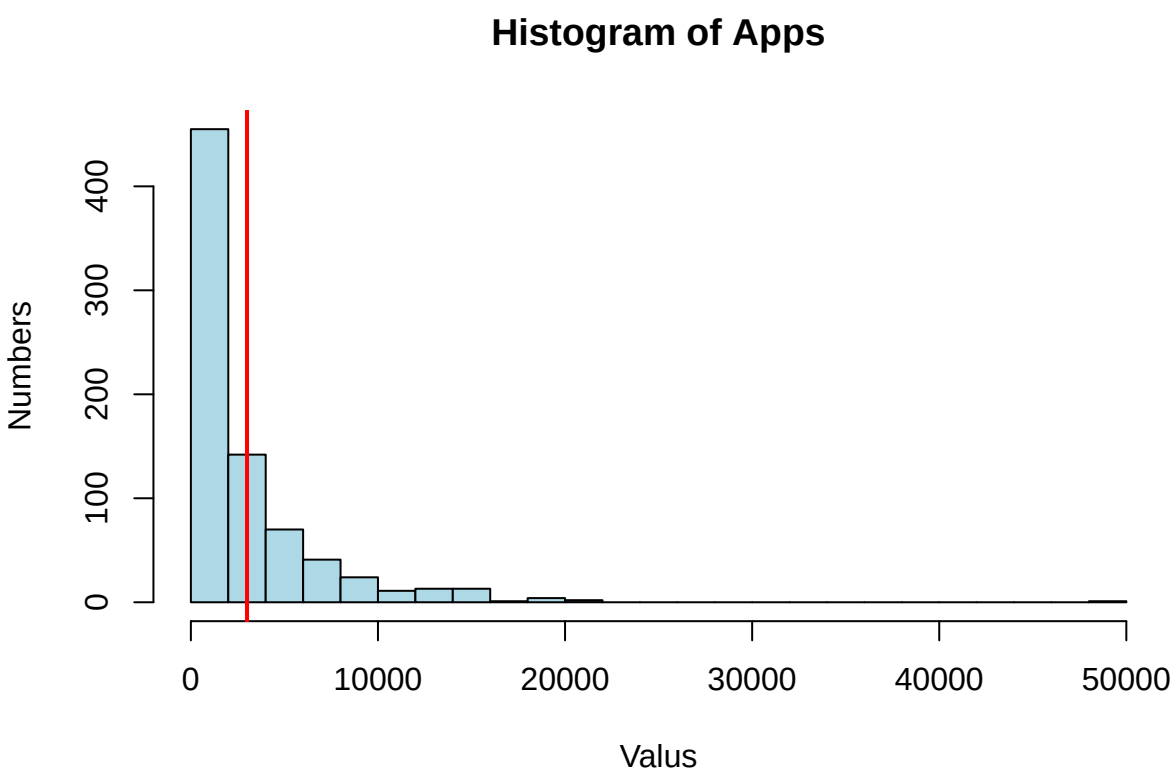
#Task 1

In linear regression, it's assumed that the residuals are normally distributed. If this assumption is violated, as shown by the Shapiro-Wilk test, it can lead to biased estimates and unreliable inference, like incorrect hypothesis testing. Heteroscedasticity (non-constant variance) or skewness in the data can result in inefficient coefficient estimates, making predictions less accurate. Additionally, if the data is heavily skewed or contains outliers, the model might struggle to generalize to new data, leading to poor performance.

Thats why we will take a look in our response data.

```r
hist(predict_value,
     main = "Histogram of Apps",
     xlab = "Valus",
     ylab = "Numbers",
     col = "lightblue",
     breaks = 20)


abline(v = mean(predict_value), col = "red", lwd = 2)
```




```r
# Shapiro-Wilk-Test durchführen
shapiro.test(predict_value)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  predict_value
## W = 0.65408, p-value < 2.2e-16
```

From these results, we can clearly see that the data deviates significantly from a normal distribution. The W-value of 0.65408 is quite far from 1, indicating a poor fit to the normal distribution, and the extremely small p-value confirms that this deviation is statistically significant.

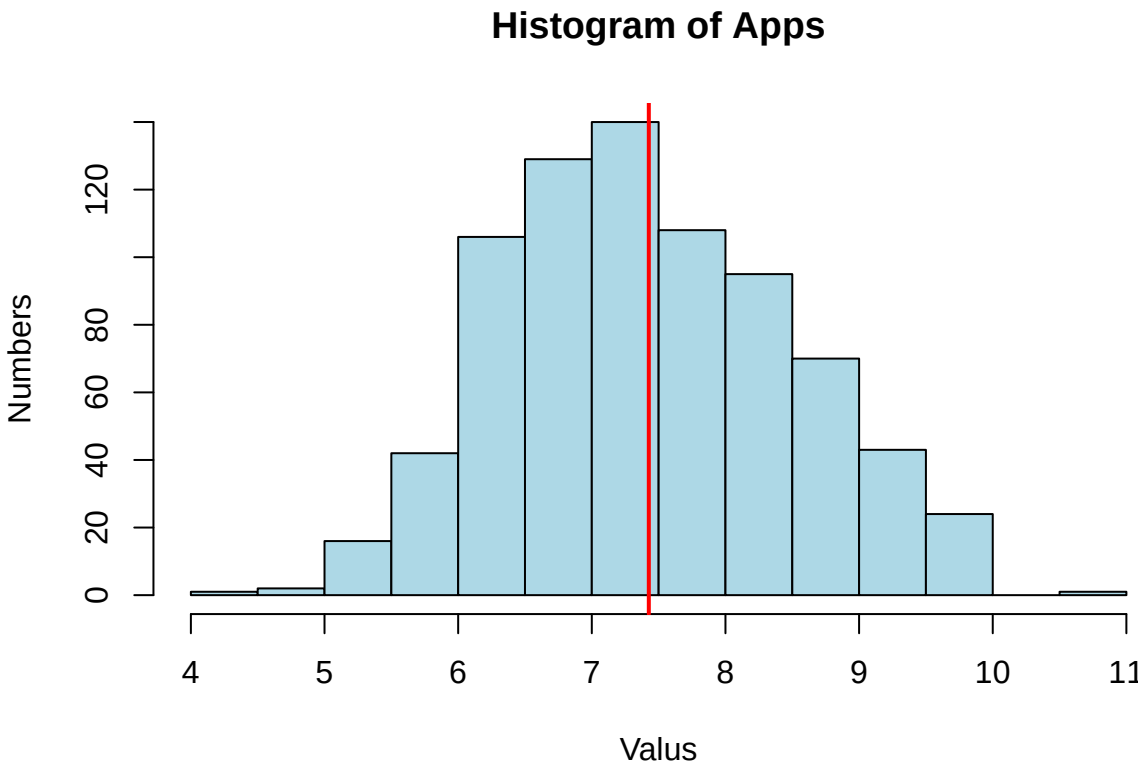
In the histogram, the data is not well spread, showing possible skewness or outliers, which is also confirmed by the Shapiro-Wilk test.

So we will do a log-transformation on the response data to adress the issues.

```
predict_value <- log(predict_value)
```

```
hist(predict_value,
     main = "Histogram of Apps",
     xlab = "Valus",
     ylab = "Numbers",
     col = "lightblue",
     breaks = 20)


abline(v = mean(predict_value), col = "red", lwd = 2)
```




```
shapiro.test(predict_value)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  predict_value
## W = 0.99085, p-value = 9.506e-05
```

After applying the transformation, several key aspects of the data have improved:

- The Shapiro-Wilk test now shows a much higher W-value, indicating that the residuals are much closer to a normal distribution.
- The transformation has helped to reduce skewness in the data

Now We will split our data into a training and test set. -> about 2/3 training and 1/3 test

```r
set.seed(187)

sample <- sample(c(TRUE, FALSE), nrow(df), replace=TRUE, prob=c(0.64,0.36))
train_x  <- df[sample, ]
test_x   <- df[!sample, ]

train_y  <- predict_value[sample ]
test_y <- predict_value[!sample ]

dim(train_x)
```

```
## [1] 490  15
```

```r
length(train_y)
```

```
## [1] 490
```

```r
dim(test_x)
```

```
## [1] 287  15
```

```r
length(test_y)
```

```
## [1] 287
```

Task 2

a) Function lm()

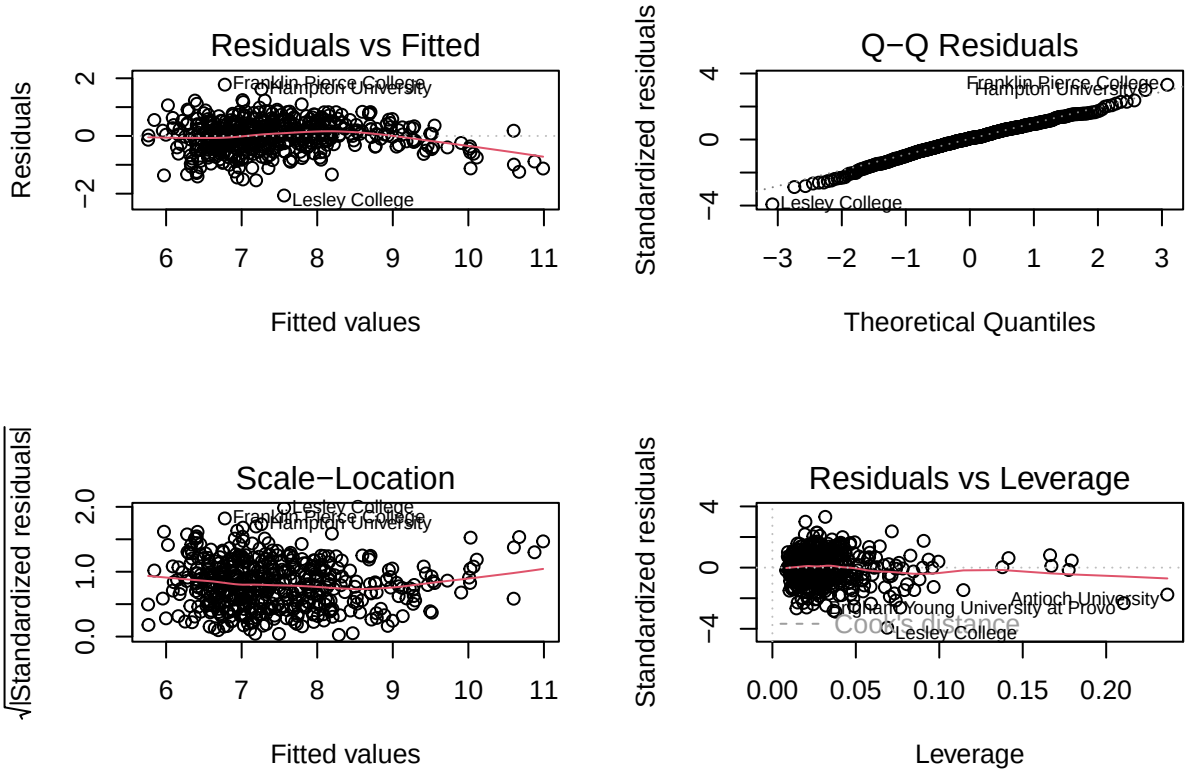
```r
res <- lm(train_y ~ ., data = train_x)  # Fit the linear model
summary(res)
```

```
##
## Call:
## lm(formula = train_y ~ ., data = train_x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0645 -0.3226  0.0425  0.3681  1.7772
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.250e+00  2.872e-01  14.801  < 2e-16 ***
## PrivateYes  -5.606e-01  9.658e-02  -5.805 1.18e-08 ***
## Top10perc    2.856e-04  3.532e-03   0.081  0.93559
## Top25perc    5.775e-03  2.865e-03   2.016  0.04439 *
## F.Undergrad  1.319e-04  8.474e-06  15.564  < 2e-16 ***
## P.Undergrad -6.100e-05  2.752e-05  -2.217  0.02712 *
## Outstate     4.877e-05  1.205e-05   4.048 6.04e-05 ***
## Room.Board   8.288e-05  3.179e-05   2.607  0.00941 **
## Books        3.325e-04  1.864e-04   1.784  0.07505 .
## Personal     3.614e-05  4.105e-05   0.880  0.37911
## PhD          4.601e-03  3.216e-03   1.431  0.15323
## Terminal     9.345e-04  3.492e-03   0.268  0.78910
## S.F.Ratio    4.777e-02  9.065e-03   5.270 2.07e-07 ***
## perc.alumni -7.825e-03  2.616e-03  -2.991  0.00293 **
## Expend       2.506e-05  8.243e-06   3.040  0.00249 **
## Grad.Rate    8.845e-03  1.872e-03   4.726 3.02e-06 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5446 on 474 degrees of freedom
## Multiple R-squared:  0.7423, Adjusted R-squared:  0.7341
## F-statistic:    91 on 15 and 474 DF,  p-value: < 2.2e-16
```

Seems like the significant predictors are 'PrivateYes', 'F.Undergrad', 'Outstate', 'Room.Board', 'perc.alumni', 'Expend', and 'Grad.Rate'. Some variables like 'Top10perc', 'Books', and 'PhD' do not seem to have a significant impact and could potentially be removed to simplify the model.

```r
par(mfrow=c(2,2))  # 2 rows, 1 column
plot(res)
```



Are the model assumptions fulfilled?

- Residuals vs Fitted: There's a slight curve in the red line, suggesting some non-linearity might be present.
- Q-Q Residuals: Most Residuals are on the line, idicates normal distribution, but with some extremes at the tail
- Scale-Location: The slight curved red line, indicate there is some heteroscedasticity.
- Residuals vs Leverag: Since there is no points outside the distance, meaning there is no significant outliers.

b) Manually compute the LS coefficients

```r
X <- model.matrix(train_y ~ ., data = train_x)
head(X)
```

```
##                             (Intercept) PrivateYes Top10perc Top25perc
## Abilene Christian University           1          1        23        52
## Adelphi University                     1          1        16        29
## Agnes Scott College                    1          1        60        89
## Alaska Pacific University              1          1        16        44
## Albertus Magnus College                1          1        17        45
## Albright College                       1          1        30        63
##                             F.Undergrad P.Undergrad Outstate Room.Board Books
```

```
## Abilene Christian University        2885          537     7440       3300   450
## Adelphi University                  2683         1227    12280       6450   750
## Agnes Scott College                  510           63    12960       5450   450
## Alaska Pacific University            249          869     7560       4120   800
## Albertus Magnus College              416          230    13290       5720   500
## Albright College                     973          306    15595       4400   300
##                                Personal PhD Terminal S.F.Ratio perc.alumni Expend
## Abilene Christian University     2200  70       78      18.1           12   7041
## Adelphi University               1500  29       30      12.2           16  10527
## Agnes Scott College               875  92       97       7.7           37  19016
## Alaska Pacific University        1500  76       72      11.9            2  10922
## Albertus Magnus College          1500  90       93      11.5           26   8861
## Albright College                  500  79       84      11.3           23  11644
##                                Grad.Rate
## Abilene Christian University          60
## Adelphi University                    56
## Agnes Scott College                   59
## Alaska Pacific University             15
## Albertus Magnus College               63
## Albright College                      80
```

```
beta_hat <- solve(t(X) %*% X) %*% t(X) %*% train_y
beta_hat
```

```
##                       [,1]
## (Intercept)  4.250309e+00
## PrivateYes  -5.606264e-01
## Top10perc    2.855535e-04
## Top25perc    5.774626e-03
## F.Undergrad  1.318950e-04
## P.Undergrad -6.100419e-05
## Outstate     4.877365e-05
## Room.Board   8.287991e-05
## Books        3.325251e-04
## Personal     3.614314e-05
## PhD          4.601077e-03
## Terminal     9.345377e-04
## S.F.Ratio    4.777464e-02
## perc.alumni -7.825236e-03
## Expend       2.506146e-05
## Grad.Rate    8.844586e-03
```

How is R handling binary variables, and how can you interpret the corresponding regression coefficient?

A: R handles binary variables by automatically converting them into dummy variables. For a binary variable like Private, which has values "Yes" and "No", R will convert this into a variable with values 0 (for "No") and 1 (for "Yes").

How can you interpret the corresponding regression coefficient?:

A: A negative coefficient means that, if the corresponding predictor variable increases (`PrivateYes`), the response variable (`Apps`) will decrease.

Since `PrivateYes` has a negative coefficient, it means that private institutions tend to have lower acceptance rates than non-private institution.

Comparing the coefficients of both models

```
lm_coef <- coef(res)

cbind(Manual = beta_hat, lm = lm_coef)
```

```
##                                    lm
## (Intercept)  4.250309e+00  4.250309e+00
```

```
## PrivateYes  -5.606264e-01 -5.606264e-01
## Top10perc    2.855535e-04  2.855535e-04
## Top25perc    5.774626e-03  5.774626e-03
## F.Undergrad  1.318950e-04  1.318950e-04
## P.Undergrad -6.100419e-05 -6.100419e-05
## Outstate     4.877365e-05  4.877365e-05
## Room.Board   8.287991e-05  8.287991e-05
## Books        3.325251e-04  3.325251e-04
## Personal     3.614314e-05  3.614314e-05
## PhD          4.601077e-03  4.601077e-03
## Terminal     9.345377e-04  9.345377e-04
## S.F.Ratio    4.777464e-02  4.777464e-02
## perc.alumni -7.825236e-03 -7.825236e-03
## Expend       2.506146e-05  2.506146e-05
## Grad.Rate    8.844586e-03  8.844586e-03
```

We can see both coefficients are same.

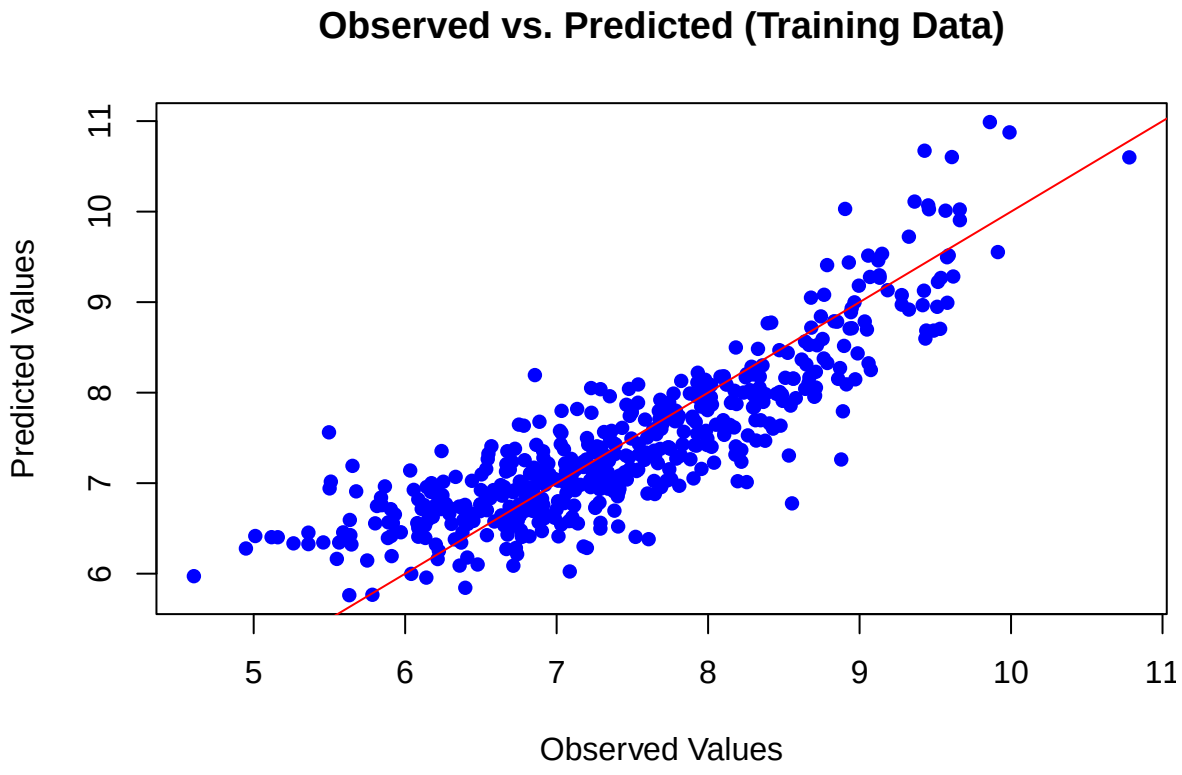
c) Compare graphically

Get Predictions

```
train_pred <- predict(res, newdata = train_x)
test_pred <- predict(res, newdata = test_x)
```

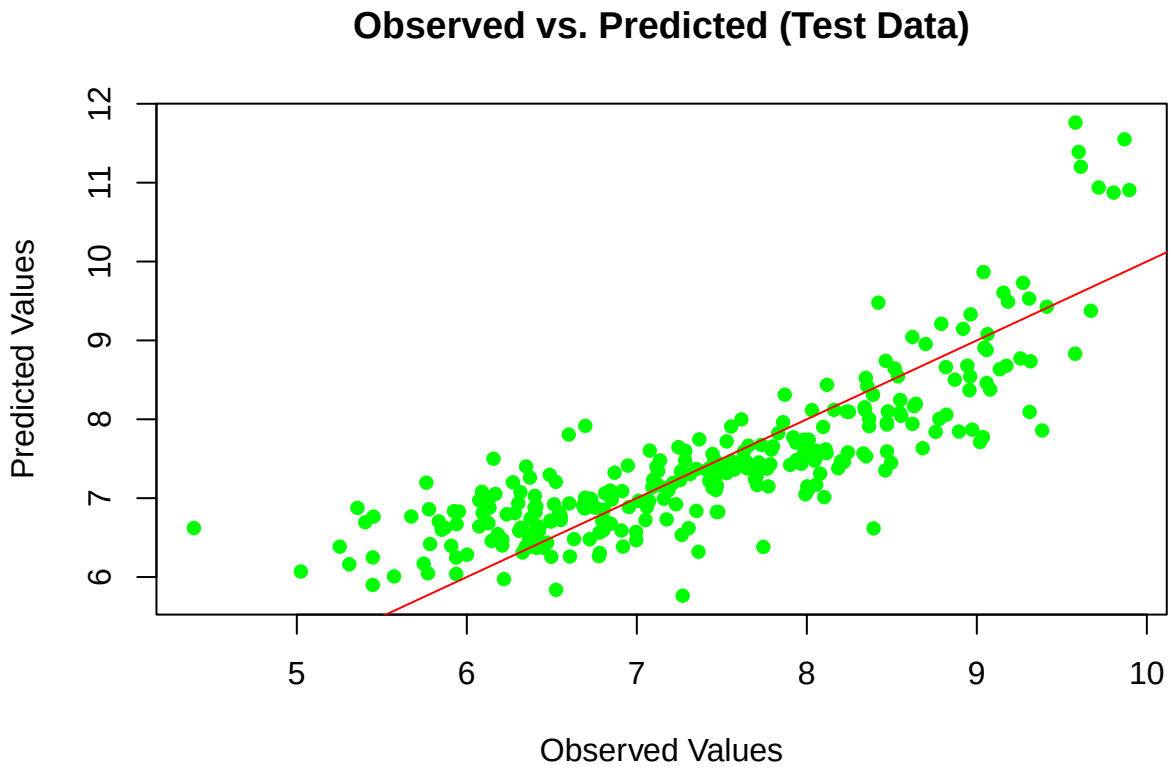
Graphically Compare Observed vs. Predicted Values:

```
plot(train_y, train_pred,
     main = "Observed vs. Predicted (Training Data)",
     xlab = "Observed Values",
     ylab = "Predicted Values",
     col = "blue", pch = 16)
abline(0, 1, col = "red")
```



Overall, the model performs well on the training data, with predictions closely matching the observed values for a large portion of the data. There are some deviations especially for lower and higher observed values, indicating potential areas for improvement, but the model generally captures the relationship well.

```r
plot(test_y, test_pred,
     main = "Observed vs. Predicted (Test Data)",
     xlab = "Observed Values",
     ylab = "Predicted Values",
     col = "green", pch = 16)
abline(0, 1, col = "red")
```



The model is showing kinda similar performance on the test data, but it is less accurate for higher observed values espacally for the values over 9, where it tends to underestimate.

d) RMSE

```r
rmse <- function(observed, predicted) {
  sqrt(mean((observed - predicted)^2))
}


train_rmse <- rmse(train_y, train_pred)
cat("RMSE for Training Data:", train_rmse, "\n")
```

```
## RMSE for Training Data: 0.5356183
```

```r
test_rmse <- rmse(test_y, test_pred)
cat("RMSE for Test Data:", test_rmse, "\n")
```

```
## RMSE for Test Data: 0.629906
```

Both values have smaller value, which is good, beacuse a lower RMSE indicates better model performance. The Training RMSE (0.5356) is lower than the Test RMSE (0.6299). This is expected, as models usually perform better on the data they were trained on, but the difference between the two RMSE values is relatively small, means the model is not significantly overfitting and is generalizing well

Task 3 Reduced model

We will exclude all input variables from the model which were not significant in 2(a), and compute the LS-estimator.

```
summary(res)
```

```
##
## Call:
## lm(formula = train_y ~ ., data = train_x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0645 -0.3226  0.0425  0.3681  1.7772
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.250e+00  2.872e-01  14.801  < 2e-16 ***
## PrivateYes  -5.606e-01  9.658e-02  -5.805 1.18e-08 ***
## Top10perc    2.856e-04  3.532e-03   0.081  0.93559
## Top25perc    5.775e-03  2.865e-03   2.016  0.04439 *
## F.Undergrad  1.319e-04  8.474e-06  15.564  < 2e-16 ***
## P.Undergrad -6.100e-05  2.752e-05  -2.217  0.02712 *
## Outstate     4.877e-05  1.205e-05   4.048 6.04e-05 ***
## Room.Board   8.288e-05  3.179e-05   2.607  0.00941 **
## Books        3.325e-04  1.864e-04   1.784  0.07505 .
## Personal     3.614e-05  4.105e-05   0.880  0.37911
## PhD          4.601e-03  3.216e-03   1.431  0.15323
## Terminal     9.345e-04  3.492e-03   0.268  0.78910
## S.F.Ratio    4.777e-02  9.065e-03   5.270 2.07e-07 ***
## perc.alumni -7.825e-03  2.616e-03  -2.991  0.00293 **
## Expend       2.506e-05  8.243e-06   3.040  0.00249 **
## Grad.Rate    8.845e-03  1.872e-03   4.726 3.02e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5446 on 474 degrees of freedom
## Multiple R-squared:  0.7423, Adjusted R-squared:  0.7341
## F-statistic:    91 on 15 and 474 DF,  p-value: < 2.2e-16
```

a) : Exclude all input

We will exlude `Top10perc`, `Books`, `Personal`, `PhD`,`Terminal`.

```
reduced_train <- train_x
reduced_train$Top10perc <- NULL
reduced_train$Books <- NULL
reduced_train$Personal <- NULL
reduced_train$PhD <- NULL
reduced_train$Terminal <- NULL

reduced_test <- test_x
reduced_test$Top10perc <- NULL
reduced_test$Books <- NULL
reduced_test$Personal <- NULL
reduced_test$PhD <- NULL
reduced_test$Terminal <- NULL

reduced_model <- lm(train_y ~ .,data = reduced_train)


summary(reduced_model)
```

```
##
## Call:
## lm(formula = train_y ~ ., data = reduced_train)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.03016 -0.33651  0.03738  0.37599  1.69957
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.705e+00  2.224e-01  21.151  < 2e-16 ***
## PrivateYes  -6.279e-01  9.221e-02  -6.810 2.94e-11 ***
## Top25perc    7.572e-03  1.660e-03   4.560 6.49e-06 ***
## F.Undergrad  1.356e-04  8.391e-06  16.166  < 2e-16 ***
## P.Undergrad -5.462e-05  2.719e-05  -2.009 0.045095 *
## Outstate     5.215e-05  1.172e-05   4.449 1.07e-05 ***
## Room.Board   1.013e-04  3.117e-05   3.252 0.001228 **
## S.F.Ratio    4.732e-02  9.008e-03   5.253 2.26e-07 ***
## perc.alumni -7.745e-03  2.583e-03  -2.998 0.002860 **
## Expend       2.834e-05  7.624e-06   3.718 0.000225 ***
## Grad.Rate    8.542e-03  1.868e-03   4.573 6.12e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5477 on 479 degrees of freedom
## Multiple R-squared:  0.7365, Adjusted R-squared:  0.731
## F-statistic: 133.9 on 10 and 479 DF,  p-value: < 2.2e-16
```

Are now all input variables significant in the model?

There are indeed all significant, even some values become more significant for example `Top25perc` become even more significant.

Why is this not to be expected in general?

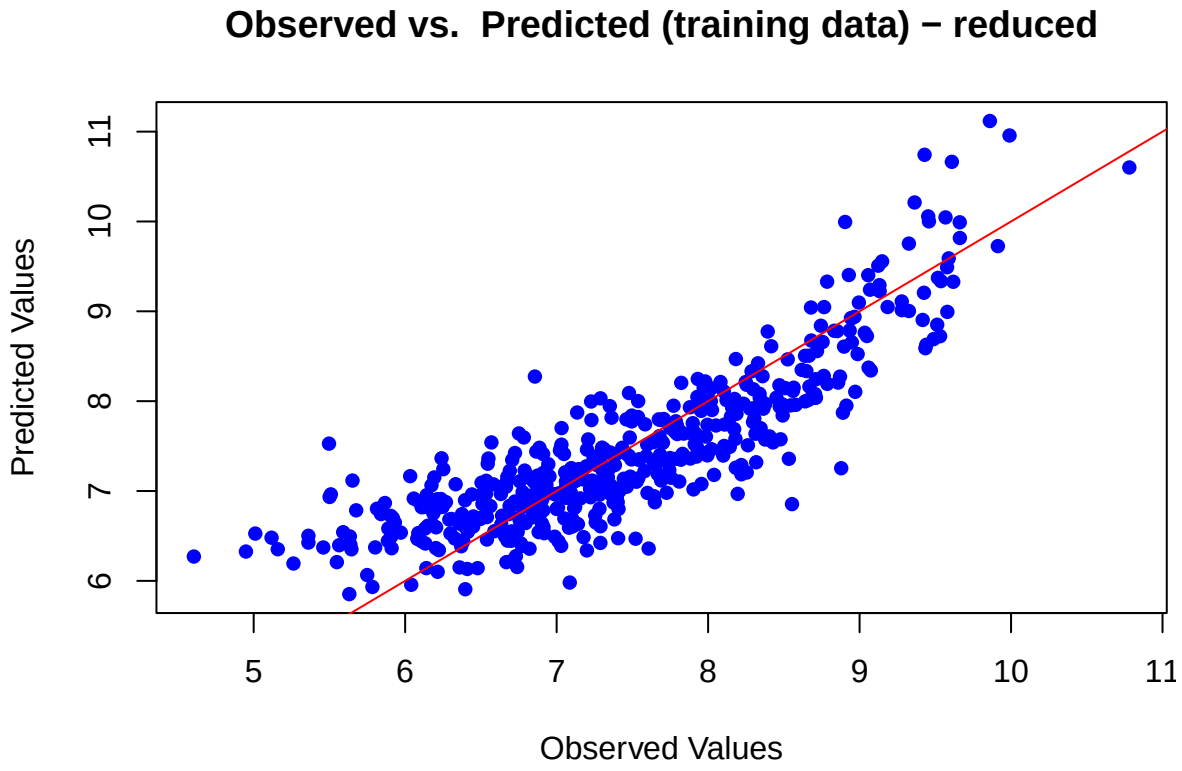
Some may become less significant or even insignificant because: - When predictors are correlated with each other (multicollinearity), their individual significance can fluctuate when other variables are added or removed from the model.

- By choosing only the significant variables from the full model, we introduce a selection bias that can artificially influence the significance of the remaining variables.

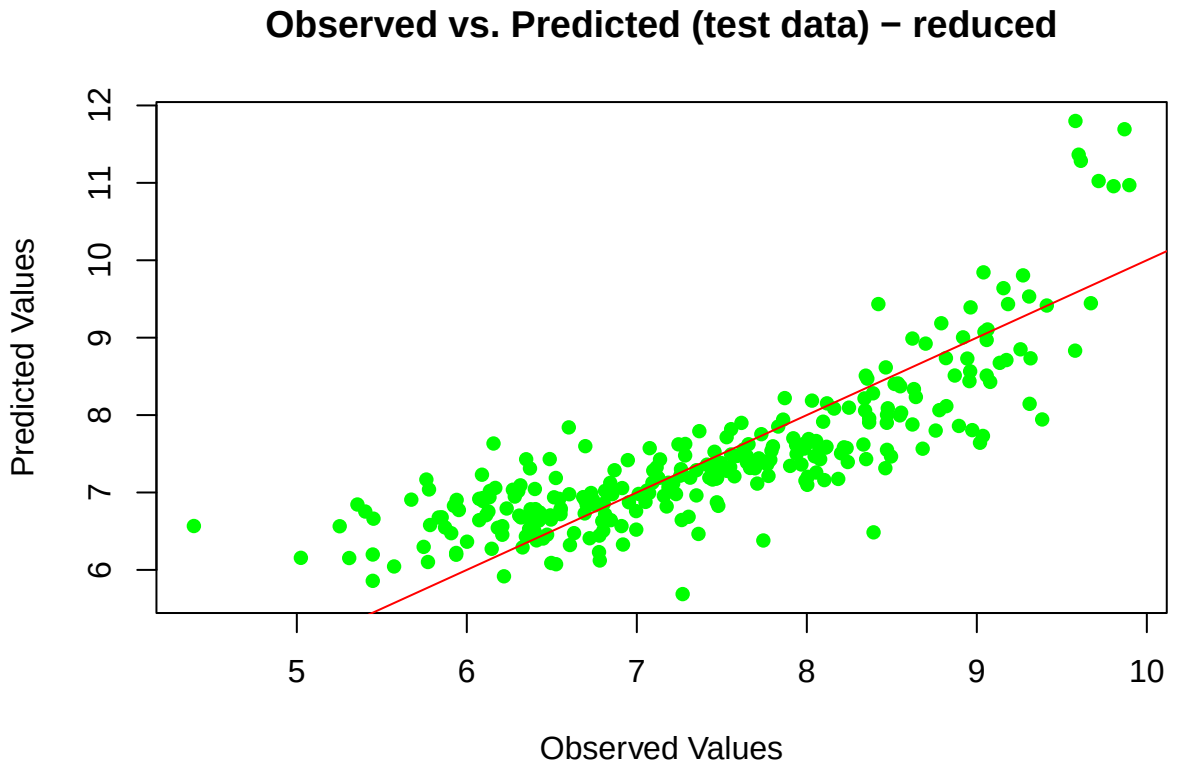
b)

```r
reduced_train_pred <- predict(reduced_model, newdata = reduced_train)
reduced_test_pred <- predict(reduced_model, newdata = reduced_test)
```

```r
plot(train_y, reduced_train_pred,
     main = "Observed vs.  Predicted (training data) - reduced",
     xlab = "Observed Values",
     ylab = "Predicted Values",
     col = "blue", pch = 16)
abline(0, 1, col = "red")
```




```
plot(test_y, reduced_test_pred,
     main = "Observed vs. Predicted (test data) - reduced",
     xlab = "Observed Values",
     ylab = "Predicted Values",
     col = "green", pch = 16)
abline(0, 1, col = "red")
```



For both reduced data sets, it didn't change much compared to the full model. ## c)

```
reduced_train_rmse <- rmse(train_y, reduced_train_pred)
cat("RMSE for Training Data:", train_rmse, "\n")
```

```
## RMSE for Training Data: 0.5356183
```

```r
reduced_test_rmse <- rmse(test_y, reduced_test_pred)
cat("RMSE for Test Data:", test_rmse, "\n")
```

```
## RMSE for Test Data: 0.629906
```

Also the RMSE values are similar to the full model.

d)

```r
anova(res, reduced_model)
```

```
## Analysis of Variance Table
##
## Model 1: train_y ~ Private + Top10perc + Top25perc + F.Undergrad + P.Undergrad +
##     Outstate + Room.Board + Books + Personal + PhD + Terminal +
##     S.F.Ratio + perc.alumni + Expend + Grad.Rate
## Model 2: train_y ~ Private + Top25perc + F.Undergrad + P.Undergrad + Outstate +
##     Room.Board + S.F.Ratio + perc.alumni + Expend + Grad.Rate
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    474 140.57
## 2    479 143.70 -5   -3.1252 2.1076 0.06331 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The RSS indicates a better fit for the full model, but its not really significant. Also a small p-value (typically < 0.05) indicates that the full model provides a significantly better fit to the data than the reduced model. Since the p-value is **0.06331**, indicates that the full model is not significant better than the reduced one. There is no strong evidence to suggest that the full model is significantly better.

Task 4 Variable selection

```r
full_model <- lm(train_y ~ ., data = train_x)

empty_model <- lm(train_y ~ 1, data = train_x)
```

```r
forward_model <- step(empty_model, scope = formula(full_model), direction = "forward")
```

```
## Start:  AIC=54.49
## train_y ~ 1
##
##                Df Sum of Sq    RSS     AIC
## + F.Undergrad   1   300.234 245.17 -335.30
## + PhD           1   140.902 404.50  -89.96
## + Private       1   128.293 417.11  -74.92
## + Terminal      1   127.397 418.01  -73.87
## + Top25perc     1    98.627 446.78  -41.25
## + P.Undergrad   1    84.303 461.10  -25.79
## + Top10perc     1    83.182 462.22  -24.60
## + Expend        1    42.536 502.87   16.70
## + Books         1    24.311 521.09   34.15
## + Grad.Rate     1    20.456 524.95   37.76
## + Room.Board    1    20.395 525.01   37.81
## + Personal      1    12.229 533.17   45.38
## + S.F.Ratio     1    11.241 534.16   46.28
## + Outstate      1     5.177 540.23   51.82
## <none>                      545.40   54.49
## + perc.alumni   1     1.360 544.04   55.27
```

```
##
## Step:  AIC=-335.3
## train_y ~ F.Undergrad
##
##               Df Sum of Sq    RSS     AIC
## + PhD          1    45.758 199.41 -434.53
## + Top25perc    1    45.341 199.83 -433.50
## + Top10perc    1    45.024 200.15 -432.73
## + Outstate     1    42.589 202.58 -426.80
## + Terminal     1    40.657 204.51 -422.15
## + Grad.Rate    1    39.360 205.81 -419.05
## + Expend       1    36.972 208.20 -413.40
## + Room.Board   1    34.838 210.33 -408.40
## + perc.alumni  1     7.524 237.65 -348.58
## + P.Undergrad  1     4.846 240.32 -343.09
## + Books        1     3.785 241.38 -340.93
## + Personal     1     1.702 243.47 -336.72
## + S.F.Ratio    1     1.470 243.70 -336.25
## <none>                     245.17 -335.30
## + Private      1     0.179 244.99 -333.66
##
## Step:  AIC=-434.53
## train_y ~ F.Undergrad + PhD
##
##               Df Sum of Sq    RSS     AIC
## + Grad.Rate    1   17.5853 181.82 -477.77
## + Outstate     1   13.7086 185.70 -467.43
## + Top25perc    1   13.5358 185.88 -466.97
## + Top10perc    1   13.4078 186.00 -466.63
## + Room.Board   1   12.1353 187.28 -463.29
## + Expend       1   11.4482 187.96 -461.50
## + Books        1    3.1649 196.25 -440.37
## + P.Undergrad  1    3.0595 196.35 -440.10
## + Terminal     1    1.3530 198.06 -435.86
## <none>                     199.41 -434.53
## + Private      1    0.5356 198.88 -433.85
## + Personal     1    0.2384 199.17 -433.11
## + S.F.Ratio    1    0.2354 199.18 -433.11
## + perc.alumni  1    0.1664 199.24 -432.94
##
## Step:  AIC=-477.77
## train_y ~ F.Undergrad + PhD + Grad.Rate
##
##               Df Sum of Sq    RSS     AIC
## + Private      1    5.5361 176.29 -490.92
## + Expend       1    5.2542 176.57 -490.13
## + Top25perc    1    5.0506 176.78 -489.57
## + Room.Board   1    4.9383 176.89 -489.26
## + Top10perc    1    4.6318 177.19 -488.41
## + Outstate     1    3.8587 177.97 -486.28
## + Books        1    3.1189 178.71 -484.24
## + S.F.Ratio    1    2.3086 179.52 -482.03
## + perc.alumni  1    2.1701 179.66 -481.65
## + Terminal     1    0.9754 180.85 -478.40
## <none>                     181.82 -477.77
## + P.Undergrad  1    0.4174 181.41 -476.89
## + Personal     1    0.1760 181.65 -476.24
##
## Step:  AIC=-490.92
## train_y ~ F.Undergrad + PhD + Grad.Rate + Private
##
##               Df Sum of Sq    RSS     AIC
## + Outstate     1   14.3365 161.95 -530.48
## + Expend       1   10.1231 166.17 -517.89
```

```
## + Room.Board   1    9.7654 166.52 -516.84
## + Top10perc    1    7.6272 168.66 -510.59
## + Top25perc    1    6.6267 169.66 -507.69
## + Books        1    3.6404 172.65 -499.14
## + Terminal     1    1.0027 175.29 -491.71
## + P.Undergrad  1    0.9625 175.33 -491.60
## <none>                     176.29 -490.92
## + perc.alumni  1    0.7063 175.58 -490.88
## + S.F.Ratio    1    0.4086 175.88 -490.05
## + Personal     1    0.0567 176.23 -489.07
##
## Step:  AIC=-530.48
## train_y ~ F.Undergrad + PhD + Grad.Rate + Private + Outstate
##
##               Df Sum of Sq    RSS     AIC
## + S.F.Ratio    1    3.7746 158.18 -540.03
## + Top25perc    1    3.4170 158.54 -538.93
## + Top10perc    1    3.1083 158.84 -537.97
## + Room.Board   1    2.7840 159.17 -536.98
## + Books        1    2.7382 159.22 -536.83
## + perc.alumni  1    2.7162 159.24 -536.77
## + Expend       1    2.1234 159.83 -534.95
## + P.Undergrad  1    0.9016 161.05 -531.21
## <none>                     161.95 -530.48
## + Personal     1    0.3214 161.63 -529.45
## + Terminal     1    0.0189 161.93 -528.54
##
## Step:  AIC=-540.03
## train_y ~ F.Undergrad + PhD + Grad.Rate + Private + Outstate +
##     S.F.Ratio
##
##               Df Sum of Sq    RSS     AIC
## + Expend       1    6.3908 151.79 -558.24
## + Top10perc    1    4.7988 153.38 -553.13
## + Top25perc    1    4.5871 153.59 -552.45
## + Books        1    3.2645 154.91 -548.25
## + Room.Board   1    3.0480 155.13 -547.57
## + perc.alumni  1    2.1952 155.98 -544.88
## + P.Undergrad  1    1.0388 157.14 -541.26
## + Personal     1    0.8402 157.34 -540.64
## <none>                     158.18 -540.03
## + Terminal     1    0.0518 158.13 -538.19
##
## Step:  AIC=-558.24
## train_y ~ F.Undergrad + PhD + Grad.Rate + Private + Outstate +
##     S.F.Ratio + Expend
##
##               Df Sum of Sq    RSS     AIC
## + Top25perc    1   2.79770 148.99 -565.36
## + perc.alumni  1   2.47982 149.31 -564.31
## + Room.Board   1   2.44026 149.35 -564.18
## + Books        1   2.43550 149.35 -564.17
## + Top10perc    1   1.76584 150.02 -561.98
## + P.Undergrad  1   0.68321 151.10 -558.45
## <none>                     151.79 -558.24
## + Personal     1   0.59262 151.19 -558.16
## + Terminal     1   0.02838 151.76 -556.33
##
## Step:  AIC=-565.36
## train_y ~ F.Undergrad + PhD + Grad.Rate + Private + Outstate +
##     S.F.Ratio + Expend + Top25perc
##
##               Df Sum of Sq    RSS     AIC
## + perc.alumni  1    3.5167 145.47 -575.06
```

```
## + Room.Board   1    2.9801 146.01 -573.26
## + Books        1    2.1256 146.86 -570.40
## <none>                     148.99 -565.36
## + Personal     1    0.5817 148.41 -565.28
## + P.Undergrad  1    0.4537 148.54 -564.85
## + Terminal     1    0.0124 148.98 -563.40
## + Top10perc    1    0.0002 148.99 -563.36
##
## Step:  AIC=-575.06
## train_y ~ F.Undergrad + PhD + Grad.Rate + Private + Outstate +
##     S.F.Ratio + Expend + Top25perc + perc.alumni
##
##               Df Sum of Sq    RSS     AIC
## + Room.Board   1   2.08536 143.39 -580.14
## + Books        1   1.66453 143.81 -578.70
## + P.Undergrad  1   0.93999 144.53 -576.24
## <none>                     145.47 -575.06
## + Personal     1   0.28226 145.19 -574.01
## + Terminal     1   0.06301 145.41 -573.27
## + Top10perc    1   0.01375 145.46 -573.11
##
## Step:  AIC=-580.14
## train_y ~ F.Undergrad + PhD + Grad.Rate + Private + Outstate +
##     S.F.Ratio + Expend + Top25perc + perc.alumni + Room.Board
##
##               Df Sum of Sq    RSS     AIC
## + P.Undergrad  1   1.36734 142.02 -582.83
## + Books        1   1.20803 142.18 -582.28
## <none>                     143.39 -580.14
## + Personal     1   0.28851 143.10 -579.12
## + Top10perc    1   0.04511 143.34 -578.29
## + Terminal     1   0.00859 143.38 -578.17
##
## Step:  AIC=-582.83
## train_y ~ F.Undergrad + PhD + Grad.Rate + Private + Outstate +
##     S.F.Ratio + Expend + Top25perc + perc.alumni + Room.Board +
##     P.Undergrad
##
##             Df Sum of Sq    RSS     AIC
## + Books      1   1.19874 140.82 -584.99
## <none>                   142.02 -582.83
## + Personal   1   0.45693 141.56 -582.41
## + Terminal   1   0.03064 141.99 -580.94
## + Top10perc  1   0.00498 142.02 -580.85
##
## Step:  AIC=-584.99
## train_y ~ F.Undergrad + PhD + Grad.Rate + Private + Outstate +
##     S.F.Ratio + Expend + Top25perc + perc.alumni + Room.Board +
##     P.Undergrad + Books
##
##             Df Sum of Sq    RSS     AIC
## <none>                   140.82 -584.99
## + Personal   1  0.225211 140.60 -583.77
## + Terminal   1  0.014452 140.81 -583.04
## + Top10perc  1  0.001380 140.82 -582.99
```

```
summary(forward_model)
```

```
##
## Call:
## lm(formula = train_y ~ F.Undergrad + PhD + Grad.Rate + Private +
##     Outstate + S.F.Ratio + Expend + Top25perc + perc.alumni +
##     Room.Board + P.Undergrad + Books, data = train_x)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.05462 -0.32725  0.03534  0.37862  1.76045
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.337e+00  2.510e-01  17.283  < 2e-16 ***
## F.Undergrad  1.324e-04  8.394e-06  15.775  < 2e-16 ***
## PhD          5.249e-03  2.096e-03   2.504 0.012621 *
## Grad.Rate    8.693e-03  1.856e-03   4.685 3.66e-06 ***
## PrivateYes  -5.656e-01  9.532e-02  -5.934 5.70e-09 ***
## Outstate     4.793e-05  1.186e-05   4.041 6.20e-05 ***
## S.F.Ratio    4.664e-02  8.952e-03   5.210 2.82e-07 ***
## Expend       2.568e-05  7.611e-06   3.374 0.000801 ***
## Top25perc    5.995e-03  1.738e-03   3.449 0.000612 ***
## perc.alumni -7.968e-03  2.581e-03  -3.088 0.002135 **
## Room.Board   8.224e-05  3.152e-05   2.609 0.009370 **
## P.Undergrad -5.793e-05  2.701e-05  -2.145 0.032475 *
## Books        3.669e-04  1.821e-04   2.015 0.044459 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5433 on 477 degrees of freedom
## Multiple R-squared:  0.7418, Adjusted R-squared:  0.7353
## F-statistic: 114.2 on 12 and 477 DF,  p-value: < 2.2e-16
```

Highest AIC is -584.99

```
backward_model <- step(full_model, direction = "backward")
```

```
## Start:  AIC=-579.85
## train_y ~ Private + Top10perc + Top25perc + F.Undergrad + P.Undergrad +
##     Outstate + Room.Board + Books + Personal + PhD + Terminal +
##     S.F.Ratio + perc.alumni + Expend + Grad.Rate
##
##               Df Sum of Sq    RSS     AIC
## - Top10perc    1     0.002 140.58 -581.84
## - Terminal     1     0.021 140.60 -581.77
## - Personal     1     0.230 140.80 -581.05
## <none>                     140.57 -579.85
## - PhD          1     0.607 141.18 -579.74
## - Books        1     0.944 141.52 -578.57
## - Top25perc    1     1.205 141.78 -577.66
## - P.Undergrad  1     1.457 142.03 -576.79
## - Room.Board   1     2.016 142.59 -574.87
## - perc.alumni  1     2.653 143.23 -572.69
## - Expend       1     2.742 143.32 -572.38
## - Outstate     1     4.859 145.43 -565.20
## - Grad.Rate    1     6.624 147.20 -559.29
## - S.F.Ratio    1     8.237 148.81 -553.94
## - Private      1     9.993 150.57 -548.20
## - F.Undergrad  1    71.841 212.42 -379.57
##
## Step:  AIC=-581.84
## train_y ~ Private + Top25perc + F.Undergrad + P.Undergrad + Outstate +
##     Room.Board + Books + Personal + PhD + Terminal + S.F.Ratio +
##     perc.alumni + Expend + Grad.Rate
##
##               Df Sum of Sq    RSS     AIC
## - Terminal     1     0.020 140.60 -583.77
## - Personal     1     0.231 140.81 -583.04
## <none>                     140.58 -581.84
## - PhD          1     0.628 141.21 -581.66
## - Books        1     0.948 141.52 -580.55
```

```
## - P.Undergrad  1     1.489 142.07 -578.68
## - Room.Board   1     2.015 142.59 -576.87
## - perc.alumni  1     2.654 143.23 -574.68
## - Expend       1     3.256 143.83 -572.62
## - Top25perc    1     3.466 144.04 -571.90
## - Outstate     1     4.870 145.45 -567.15
## - Grad.Rate    1     6.646 147.22 -561.21
## - S.F.Ratio    1     8.248 148.82 -555.90
## - Private      1    10.007 150.58 -550.15
## - F.Undergrad  1    72.715 213.29 -379.56
##
## Step:  AIC=-583.77
## train_y ~ Private + Top25perc + F.Undergrad + P.Undergrad + Outstate +
##     Room.Board + Books + Personal + PhD + S.F.Ratio + perc.alumni +
##     Expend + Grad.Rate
##
##               Df Sum of Sq    RSS     AIC
## - Personal     1     0.225 140.82 -584.99
## <none>                     140.60 -583.77
## - Books        1     0.967 141.56 -582.41
## - P.Undergrad  1     1.472 142.07 -580.67
## - PhD          1     1.857 142.45 -579.34
## - Room.Board   1     2.068 142.66 -578.62
## - perc.alumni  1     2.634 143.23 -576.67
## - Expend       1     3.261 143.86 -574.54
## - Top25perc    1     3.485 144.08 -573.77
## - Outstate     1     5.001 145.60 -568.65
## - Grad.Rate    1     6.627 147.22 -563.20
## - S.F.Ratio    1     8.232 148.83 -557.89
## - Private      1    10.316 150.91 -551.08
## - F.Undergrad  1    72.726 213.32 -381.48
##
## Step:  AIC=-584.99
## train_y ~ Private + Top25perc + F.Undergrad + P.Undergrad + Outstate +
##     Room.Board + Books + PhD + S.F.Ratio + perc.alumni + Expend +
##     Grad.Rate
##
##               Df Sum of Sq    RSS     AIC
## <none>                     140.82 -584.99
## - Books        1     1.199 142.02 -582.83
## - P.Undergrad  1     1.358 142.18 -582.28
## - PhD          1     1.851 142.67 -580.59
## - Room.Board   1     2.009 142.83 -580.04
## - perc.alumni  1     2.814 143.64 -577.29
## - Expend       1     3.361 144.18 -575.43
## - Top25perc    1     3.513 144.33 -574.91
## - Outstate     1     4.821 145.64 -570.49
## - Grad.Rate    1     6.479 147.30 -564.94
## - S.F.Ratio    1     8.013 148.84 -559.87
## - Private      1    10.395 151.22 -552.09
## - F.Undergrad  1    73.465 214.29 -381.27
```

```
summary(backward_model)
```

```
##
## Call:
## lm(formula = train_y ~ Private + Top25perc + F.Undergrad + P.Undergrad +
##     Outstate + Room.Board + Books + PhD + S.F.Ratio + perc.alumni +
##     Expend + Grad.Rate, data = train_x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.05462 -0.32725  0.03534  0.37862  1.76045
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.337e+00  2.510e-01  17.283  < 2e-16 ***
## PrivateYes  -5.656e-01  9.532e-02  -5.934 5.70e-09 ***
## Top25perc    5.995e-03  1.738e-03   3.449 0.000612 ***
## F.Undergrad  1.324e-04  8.394e-06  15.775  < 2e-16 ***
## P.Undergrad -5.793e-05  2.701e-05  -2.145 0.032475 *
## Outstate     4.793e-05  1.186e-05   4.041 6.20e-05 ***
## Room.Board   8.224e-05  3.152e-05   2.609 0.009370 **
## Books        3.669e-04  1.821e-04   2.015 0.044459 *
## PhD          5.249e-03  2.096e-03   2.504 0.012621 *
## S.F.Ratio    4.664e-02  8.952e-03   5.210 2.82e-07 ***
## perc.alumni -7.968e-03  2.581e-03  -3.088 0.002135 **
## Expend       2.568e-05  7.611e-06   3.374 0.000801 ***
## Grad.Rate    8.693e-03  1.856e-03   4.685 3.66e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5433 on 477 degrees of freedom
## Multiple R-squared:  0.7418, Adjusted R-squared:  0.7353
## F-statistic: 114.2 on 12 and 477 DF,  p-value: < 2.2e-16
```

Also highest AIC is -584.99

```r
rmse <- function(observed, predicted) {
  sqrt(mean((observed - predicted)^2))
}

forward_train_pred <- predict(forward_model, newdata = train_x)
forward_test_pred <- predict(forward_model, newdata = test_x)

forward_train_rmse <- rmse(train_y, forward_train_pred)
forward_test_rmse <- rmse(test_y, forward_test_pred)

cat("Forward Model RMSE (Training):", forward_train_rmse, "\n")
```

```
## Forward Model RMSE (Training): 0.5360889
```

```r
cat("Forward Model RMSE (Test):", forward_test_rmse, "\n")
```

```
## Forward Model RMSE (Test): 0.6285708
```

```r
backward_train_pred <- predict(backward_model, newdata = train_x)
backward_test_pred <- predict(backward_model, newdata = test_x)

backward_train_rmse <- rmse(train_y, backward_train_pred)
backward_test_rmse <- rmse(test_y, backward_test_pred)

cat("Backward Model RMSE (Training):", backward_train_rmse, "\n")
```

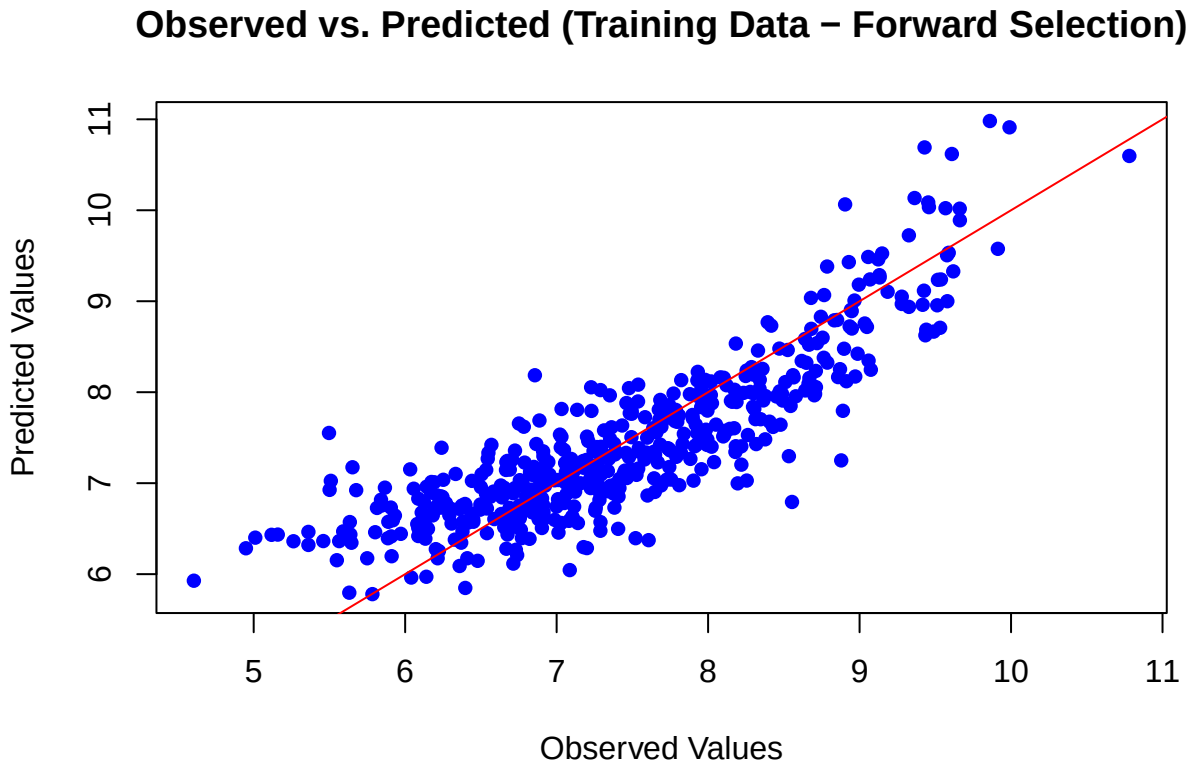
```
## Backward Model RMSE (Training): 0.5360889
```

```r
cat("Backward Model RMSE (Test):", backward_test_rmse, "\n")
```

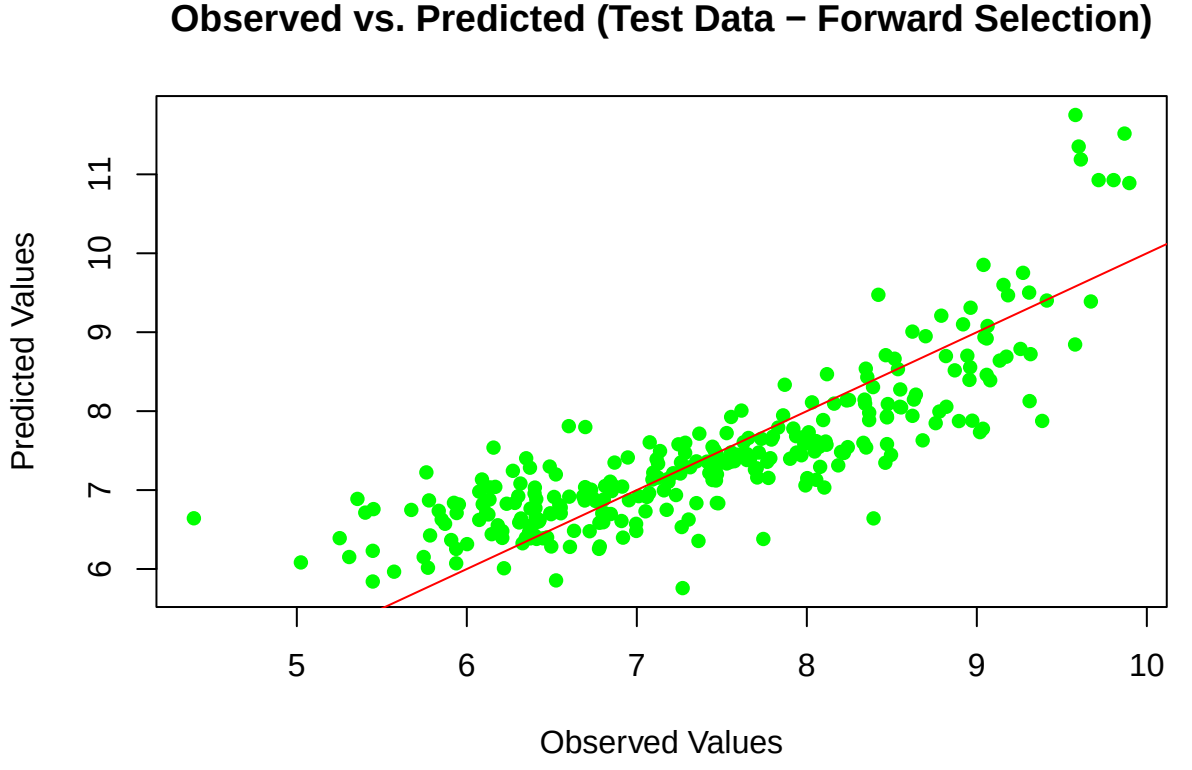
```
## Backward Model RMSE (Test): 0.6285708
```

So we have also similar RMSE values.

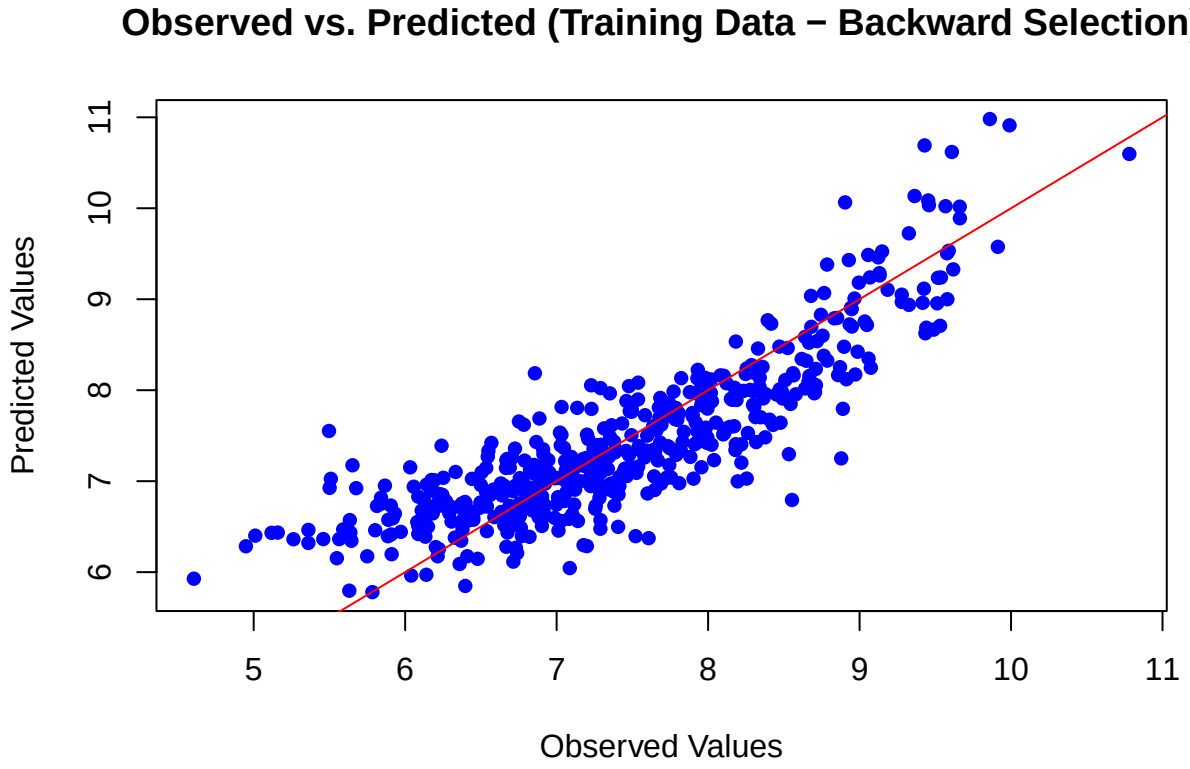
```r
plot(train_y, forward_train_pred,
     main = "Observed vs. Predicted (Training Data - Forward Selection)",
     xlab = "Observed Values",
     ylab = "Predicted Values",
     col = "blue", pch = 16)
abline(0, 1, col = "red")
```




```
plot(test_y, forward_test_pred,
     main = "Observed vs. Predicted (Test Data - Forward Selection)",
     xlab = "Observed Values",
     ylab = "Predicted Values",
     col = "green", pch = 16)
abline(0, 1, col = "red")
```




```
plot(train_y, backward_train_pred,
     main = "Observed vs. Predicted (Training Data - Backward Selection)",
     xlab = "Observed Values",
     ylab = "Predicted Values",
     col = "blue", pch = 16)
abline(0, 1, col = "red")
```




```
plot(test_y, backward_test_pred,
     main = "Observed vs. Predicted (Test Data - Backward Selection)",
     xlab = "Observed Values",
     ylab = "Predicted Values",
     col = "green", pch = 16)
abline(0, 1, col = "red")
```