Exercise 1

Konstantinos Vakalopoulos 12223236

2022-10-19

Premilinary work

First, the ISLR package was installed

```
install.packages("ISLR")
```

and then the data was loaded with

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

For addition information about the data, the following commands were used

```
?College
str(College)
```

The observations which contain missing values and the variables Accept and Enroll were removed from the original data

```
College <- na.omit(College) #Check for NA observations
College <- College[,-c(3,4)] #Remove the Accept and Enroll variables
```

Afterwards, the log transformation was applied to the Apps variable, because this variable is the one that needs to be predicted. The reason, why the log transformation was used, is shown to the below 2 histograms. In the histogram with log transformation, the data are normally distributed compared to the data with no transformation.

```
hist(College$Apps, breaks = 20, col="dodgerblue3",
     main = "Histogram of Apps before transformation",
     xlab = "Apps")
```

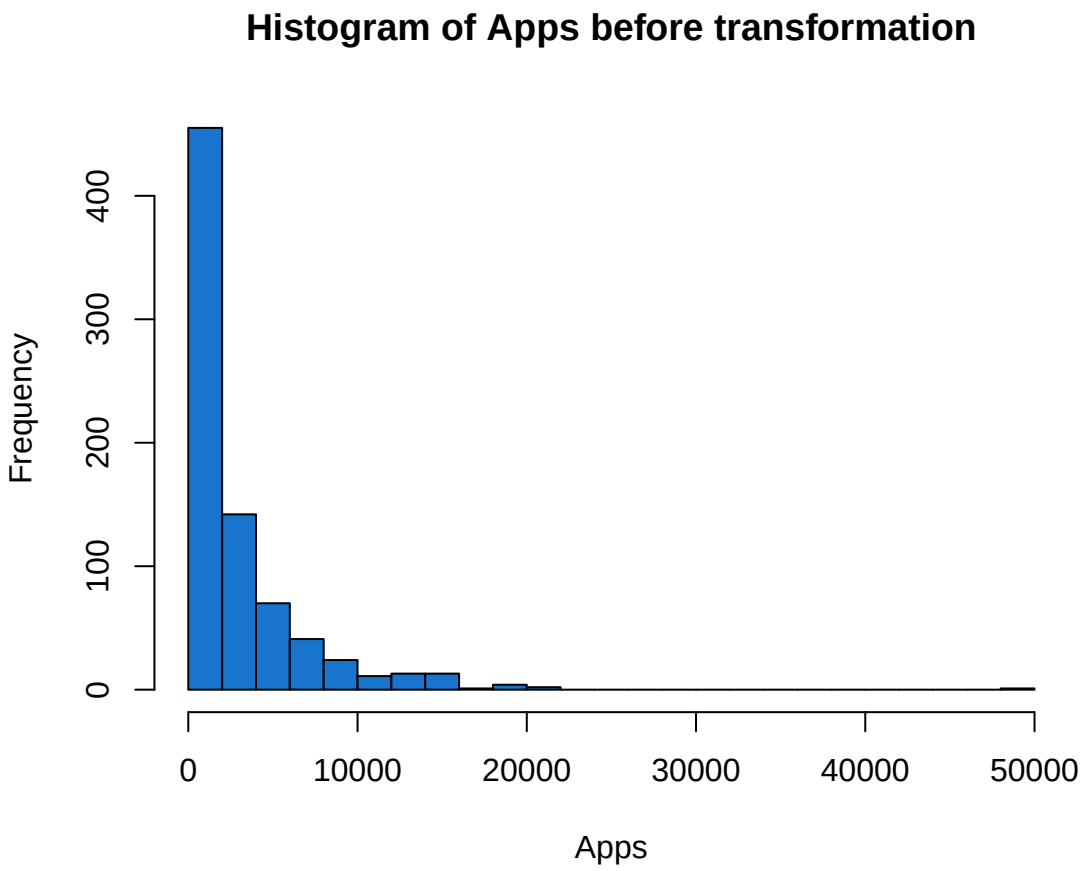

Figure 1: Before log transformation

```
hist(log(College$Apps), breaks = 20, col="dodgerblue3",
     main = "Histogram of Apps after log transformation",
     xlab = "log(Apps)")
```

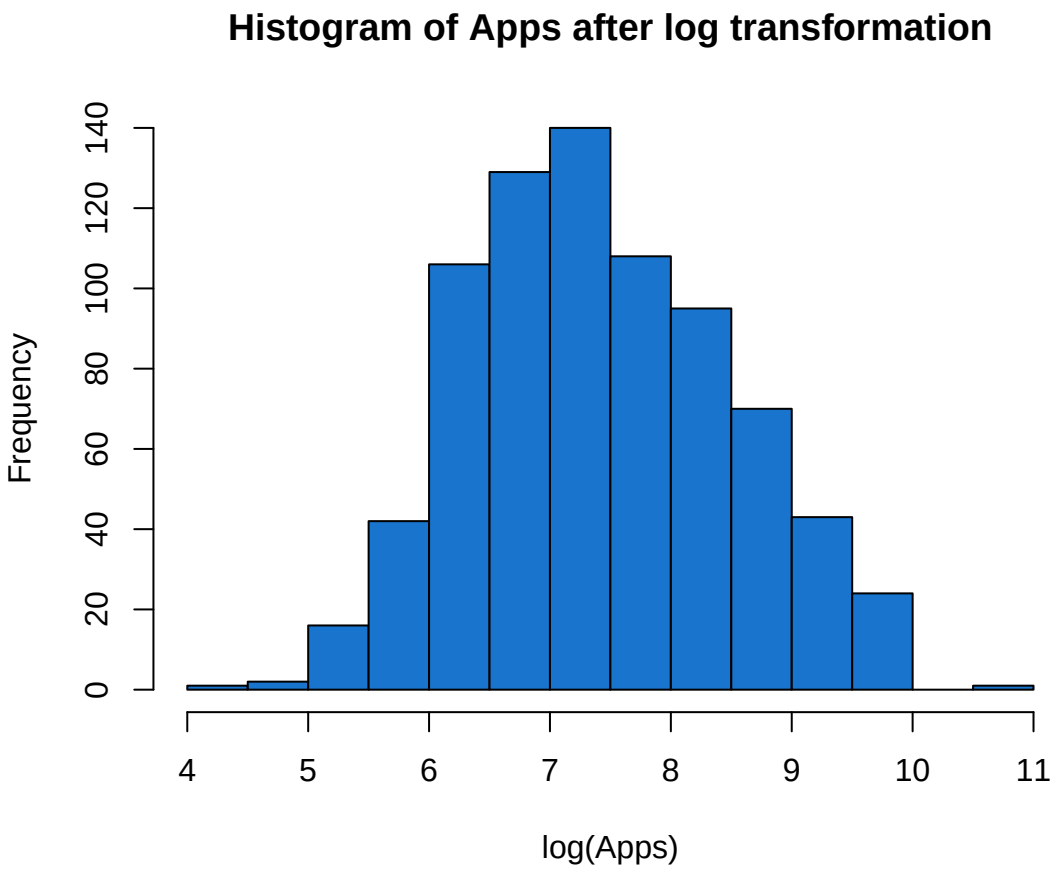

Figure 2: After log transformation

Thus, the values of the Apps variable has been changed to log transformed

```
College$Apps <- log(College$Apps) #log transformation
```

Finally, the data was splitted randomly into training and test data (about 2/3 and 1/3)

```
#Split the data into train and test
set.seed(12223236)
n <- nrow(College)
train <- sample(1:n,round(n*2/3))
test <- (1:n)[-train]
```

Question 1(a)

The function lm() was applied to compute the estimator. To interpret the outcome, the function summary() was used

```
res <- lm(Apps~., data = College, subset = train) #Apply the lm() function
summary(res) #Summary the results
```

```
##
## Call:
## lm(formula = Apps ~ ., data = College, subset = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.19876 -0.30015  0.03285  0.36024  1.74500
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.081e+00  2.550e-01  16.004  < 2e-16 ***
## PrivateYes  -6.215e-01  9.097e-02  -6.831 2.44e-11 ***
## Top10perc   -1.812e-03  3.576e-03  -0.507 0.612603
## Top25perc    5.010e-03  2.884e-03   1.737 0.082977 .
## F.Undergrad  1.078e-04  7.748e-06  13.915  < 2e-16 ***
## P.Undergrad  1.353e-05  1.830e-05   0.739 0.460063
## Outstate     4.542e-05  1.254e-05   3.622 0.000322 ***
## Room.Board   1.047e-04  3.103e-05   3.374 0.000799 ***
## Books        3.599e-04  1.472e-04   2.445 0.014817 *
## Personal     2.806e-05  3.925e-05   0.715 0.474941
## PhD          6.103e-03  2.810e-03   2.172 0.030322 *
## Terminal     1.159e-03  3.176e-03   0.365 0.715377
## S.F.Ratio    3.905e-02  8.445e-03   4.624 4.79e-06 ***
## perc.alumni -6.835e-03  2.576e-03  -2.653 0.008229 **
## Expend       2.564e-05  7.143e-06   3.589 0.000364 ***
## Grad.Rate    1.187e-02  1.939e-03   6.119 1.89e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5478 on 502 degrees of freedom
## Multiple R-squared:  0.7343, Adjusted R-squared:  0.7264
## F-statistic: 92.49 on 15 and 502 DF,  p-value: < 2.2e-16
```

The variables were chosen according to their p values, where they were less or equal than 0.05. In R, to find these variables, the below code was used

```
pvalues <- summary(res)$coefficients[,4]     #Get the p values from summary
pvalues <- pvalues[-c(1)]                    #Remove the intercept P value
variables <- names(which(pvalues<=0.05))     #Select the variables with P value<=0.05
```

The variables contribute to explaining the variable Apps are:

```
variables
```

```
##  [1] "PrivateYes"  "F.Undergrad" "Outstate"    "Room.Board"  "Books"
##  [6] "PhD"         "S.F.Ratio"   "perc.alumni" "Expend"      "Grad.Rate"
```

Finally, the diagnostics plots were plotted

```
par(mfrow = c(2, 2))
plot(res)
```

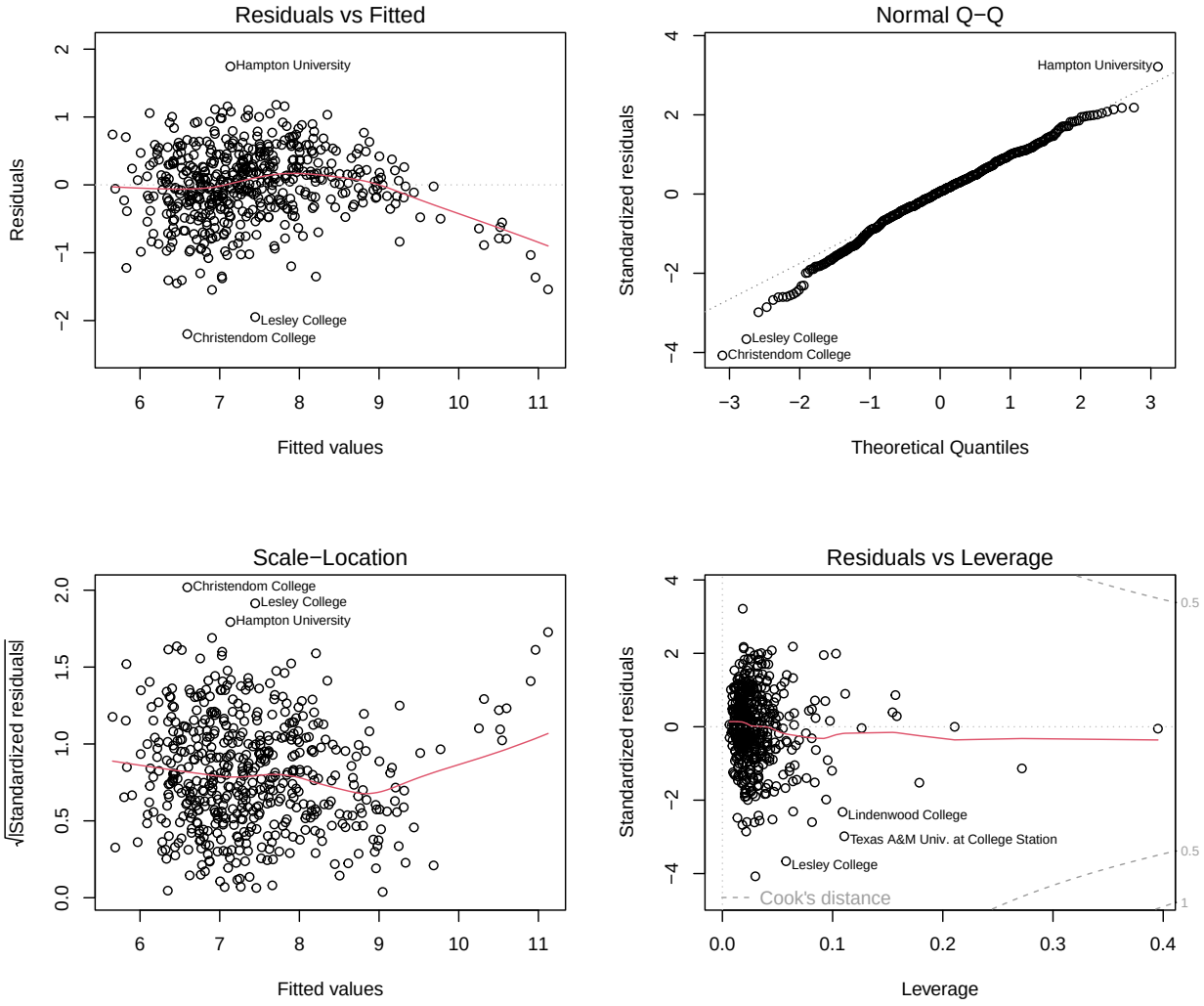

Figure 3: Model Assumptions

According to the plots, the model assumptions are valid. In residuals vs fitted plot, the line is horizontal, which indicates a good linear relationship. In normal Q-Q plot, the residuals points follow a straight dash line, which means that the residuals are normally distributed. The scale-Location plot shows if residuals are distributed equally across the prediction ranges. In our case, the red line is horizontal which is good and means that the points are spread equally. Finally, the Residuals vs Leverage plot all of the points, with some exceptions of course, are between the range of [-2,2], which is very good for the linear regression model.

Question 1(b)

In this part of the exercise, the LS coefficients were manually computed. The command lm() was replaced by model.matrix(), the matrix multiplication was done by %*% and the inverse matrix was computed with solve()

```
X <- model.matrix(Apps~., data = College)
LS_coefficients <- solve(t(X[train,])%*%X[train,])%*%t(X[train,])%*%College$Apps[train]
```

```
LS_coefficients
```

```
##                       [,1]
## (Intercept)  4.080714e+00
## PrivateYes  -6.214811e-01
## Top10perc   -1.811881e-03
## Top25perc    5.010155e-03
## F.Undergrad  1.078051e-04
## P.Undergrad  1.352558e-05
## Outstate     4.541600e-05
## Room.Board   1.046868e-04
## Books        3.598539e-04
## Personal     2.806283e-05
## PhD          6.102672e-03
## Terminal     1.158761e-03
## S.F.Ratio    3.904909e-02
## perc.alumni -6.835174e-03
## Expend       2.563585e-05
## Grad.Rate    1.186682e-02
```

In R, dummy variables are created to handle the binary variables. In this specific case, a PrivateYes dummy variable has been created that takes on a value of 1 when Private is Yes and 0 if Private is no. This can be proved by using the function contrasts()

```
contrasts(College$Private)
```

```
##     Yes
## No    0
## Yes   1
```

The corresponding regression coefficient is negative, which means that the value Yes is associated with the reduction of the value Apps.

Question 1(c)

The function predict() was used to predict the values of Apps for the training and test data

```
pred.train <- predict(res,newdata = College[train,])
pred.test <- predict(res,newdata = College[test,])
```

Graphically the observed and predicted values are shown to the figures 4 and 5

```
plot(College[train,"Apps"],pred.train,xlab="y measured train",
      ylab="y predicted train",cex.lab=1.3,
      xlim=c(4,13),ylim=c(4,13))
```

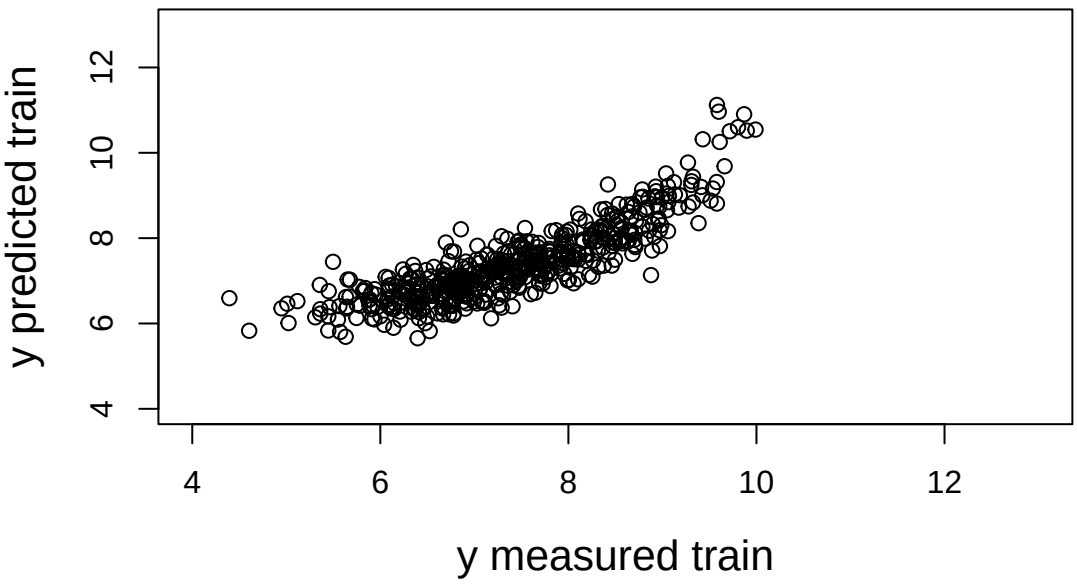

Figure 4: Prediction of the training data

```
plot(College[test,"Apps"],pred.test,xlab="y measured test"
     ,ylab="y predicted test",cex.lab=1.3,
     xlim=c(4,13),ylim=c(4,13))
```

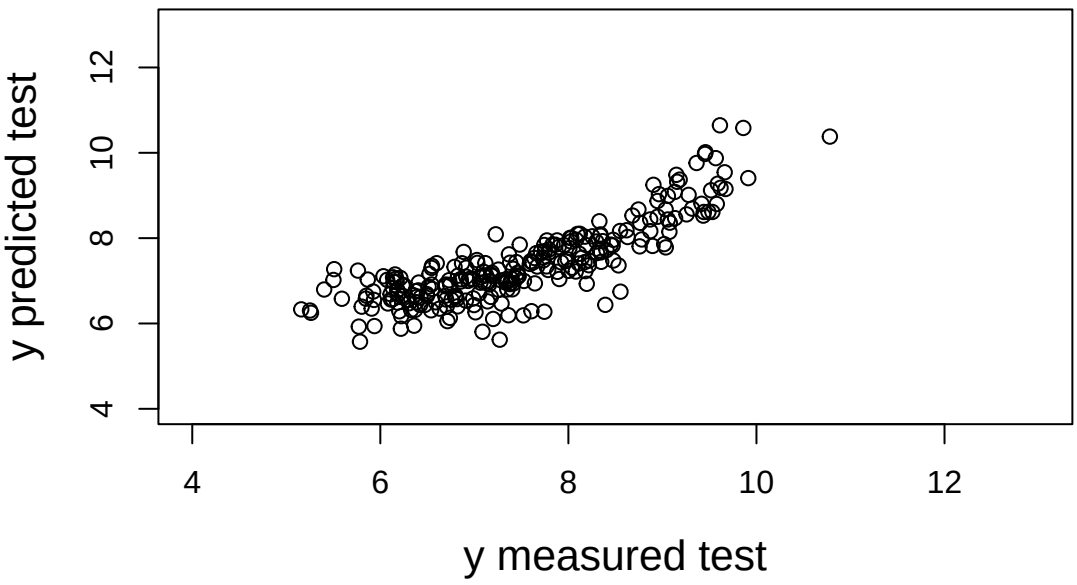

Figure 5: Prediction of the test data

According to the plots, the performance of the model is quite high because of the linearity of the actual and predicted values. In both plots, the linearity is distinctive while the actual values are roughly lower than 10. As the actual values increase, the model tend to become non-linear.

Question 1(d)

The RMSE for the training data is

```
sqrt(sum((College$Apps[train]-pred.train)^2)/length(train))
```

```
## [1] 0.5393014
```

and the RMSE for the test data is

```
sqrt(sum((College$Apps[test]-pred.test)^2)/length(test))
```

```
## [1] 0.6152971
```

First, the RMSE for the training data is lower than the one for the test data. This due to the fact that the model was created from the train data. Furthermore, both RMSE values are very low which indicates better fit and thus good performance of the model. Finally, the difference between the two values are not that high, as a result the model has the ability to predict not only the training data but unknown data quite effectively.

Question 2(a)

First, all input variables, from the model which were not significant in 1(a), were excluded. Afterwards, new data named new_College were created and the function lm() was applied

```
variables <- c(variables, "Apps")  #Add the Apps variable to the new data set
variables[1] <- "Private" #Change the name to Private from PrivateYes
new_College <- College[, (names(College) %in% variables)] #Create the new data
reduced_model <- lm(Apps~., data = new_College, subset = train) #Create the model
summary(reduced_model)
```

```
##
## Call:
## lm(formula = Apps ~ ., data = new_College, subset = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1479 -0.2918  0.0205  0.3661  1.8205
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.198e+00  2.249e-01  18.663  < 2e-16 ***
## PrivateYes  -6.322e-01  9.057e-02  -6.980 9.29e-12 ***
## F.Undergrad  1.140e-04  6.853e-06  16.638  < 2e-16 ***
## Outstate     4.664e-05  1.232e-05   3.786 0.000172 ***
## Room.Board   1.053e-04  3.063e-05   3.438 0.000634 ***
## Books        4.128e-04  1.430e-04   2.887 0.004058 **
## PhD          8.142e-03  1.877e-03   4.337 1.74e-05 ***
## S.F.Ratio    3.777e-02  8.405e-03   4.493 8.69e-06 ***
## perc.alumni -5.916e-03  2.516e-03  -2.351 0.019107 *
## Expend       2.650e-05  6.655e-06   3.982 7.83e-05 ***
## Grad.Rate    1.242e-02  1.821e-03   6.823 2.55e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5487 on 507 degrees of freedom
## Multiple R-squared:  0.7308, Adjusted R-squared:  0.7255
## F-statistic: 137.6 on 10 and 507 DF,  p-value: < 2.2e-16
```

and the LS-estimator was computed

```
X <- model.matrix(Apps~., data = new_College)
LS_coefficients_reduced = solve(t(X[train,])%*%X[train,]) %*%
               t(X[train,])%*%new_College$Apps[train]
LS_coefficients_reduced
```

```
##                      [,1]
## (Intercept)  4.197622e+00
## PrivateYes  -6.321780e-01
## F.Undergrad  1.140193e-04
## Outstate     4.663891e-05
## Room.Board   1.052988e-04
## Books        4.128348e-04
```

```
## PhD          8.141699e-03
## S.F.Ratio    3.776835e-02
## perc.alumni -5.915837e-03
## Expend       2.650206e-05
## Grad.Rate    1.242266e-02
```

According to their p values, the variables remain significant to the new model. However, for the variable perc.alumni, the p value has been increased up to 0.0191, which is a higher from the previous one, which was 0.00822. Thus, there are cases where the variables, after the variable selection, have worse p value than the p value before the variable selection. The reason why this is happening is that the p values are adjusted for the terms of the new model.

Question 2(b)

In figures 6 and 7 are visualized the fit and the prediction from the new/reduced model. The function predict() was used once again in order to find the predicted values for the training and test data

```
pred.train.reduced <- predict(reduced_model,newdata = new_College[train,])
pred.test.reduced <- predict(reduced_model,newdata = new_College[test,])
```

```
plot(College[train,"Apps"],pred.train.reduced,xlab="y measured train",
      ylab="y predicted train",cex.lab=1.3,
      xlim=c(4,13),ylim=c(4,13))
```

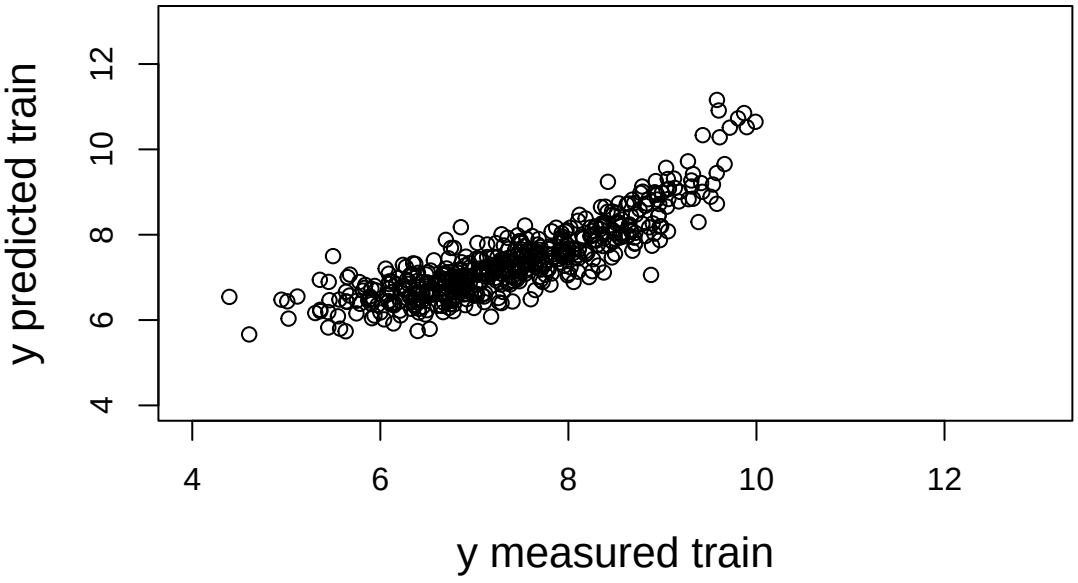

Figure 6: Prediction of the training data

```
plot(College[test,"Apps"],pred.test.reduced,xlab="y measured test"
     ,ylab="y predicted test",cex.lab=1.3,
     xlim=c(4,13),ylim=c(4,13))
```

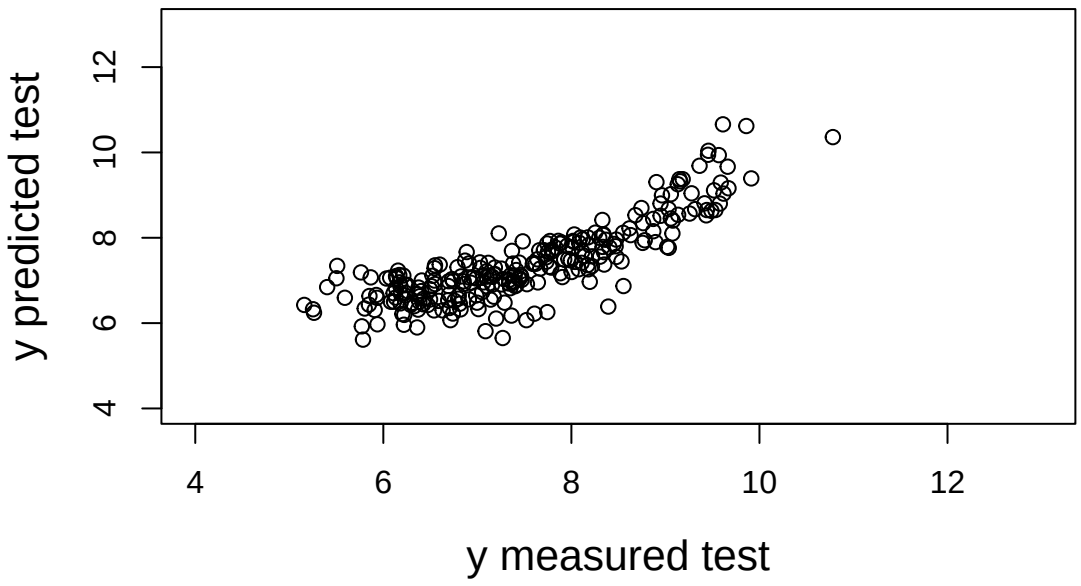

Figure 7: Prediction of the test data

Question 2(c)

The RMSE for the train and test data for the new/reduced model is

```
sqrt(sum((new_College$Apps[train]-pred.train.reduced)^2)/length(train))
```

```
## [1] 0.5428701
```

```
sqrt(sum((new_College$Apps[test]-pred.test.reduced)^2)/length(test))
```

```
## [1] 0.6171016
```

We would except that both values of RMSEs for the reduced model will be lower the RMSEs values for the full model. However, it is shown that the new RMSEs have slightly higher values and thus the variable selection leads to the formation of slightly worse model than the original.

Question 2(d)

The two models were compared with the function anova()

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

```
## Analysis of Variance Table
##
## Model 1: Apps ~ Private + Top10perc + Top25perc + F.Undergrad + P.Undergrad +
##     Outstate + Room.Board + Books + Personal + PhD + Terminal +
##     S.F.Ratio + perc.alumni + Expend + Grad.Rate
## Model 2: Apps ~ Private + F.Undergrad + Outstate + Room.Board + Books +
##     PhD + S.F.Ratio + perc.alumni + Expend + Grad.Rate
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    502 150.66
## 2    507 152.66 -5   -2.0005 1.3331 0.2487
```

The two models, after the application of the function anova(), are approximately the same. This is due to the fact that the difference between the RSS value is quite small. In case this difference was very large, the full model will explain the data significantly better than the reduced model.

Question 3

In this part of the exercise, variable selection was performed based on stepwise regression. The function step() was used for the forward and backward selection

```
# backward selection:
model.backward <- step(res,direction="backward")
```

```
## Start:  AIC=-607.71
## Apps ~ 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
## - Terminal     1     0.040 150.70 -609.57
## - Top10perc    1     0.077 150.74 -609.45
## - Personal     1     0.153 150.81 -609.18
## - P.Undergrad  1     0.164 150.82 -609.15
## <none>                     150.66 -607.71
## - Top25perc    1     0.906 151.56 -606.61
## - PhD          1     1.416 152.07 -604.86
## - Books        1     1.795 152.45 -603.58
## - perc.alumni  1     2.112 152.77 -602.50
## - Room.Board   1     3.416 154.07 -598.10
## - Expend       1     3.866 154.52 -596.58
## - Outstate     1     3.937 154.60 -596.35
## - S.F.Ratio    1     6.417 157.07 -588.10
## - Grad.Rate    1    11.238 161.90 -572.44
## - Private      1    14.006 164.66 -563.66
## - F.Undergrad  1    58.108 208.77 -440.74
##
## Step:  AIC=-609.57
## Apps ~ Private + Top10perc + 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
## - Top10perc    1     0.092 150.79 -611.26
## - Personal     1     0.148 150.85 -611.06
## - P.Undergrad  1     0.165 150.86 -611.01
## <none>                     150.70 -609.57
## - Top25perc    1     0.986 151.68 -608.20
## - Books        1     1.919 152.62 -605.02
## - perc.alumni  1     2.074 152.77 -604.49
## - Room.Board   1     3.525 154.22 -599.60
## - PhD          1     3.619 154.32 -599.28
## - Expend       1     3.908 154.61 -598.31
## - Outstate     1     4.086 154.78 -597.72
## - S.F.Ratio    1     6.483 157.18 -589.75
## - Grad.Rate    1    11.254 161.95 -574.27
## - Private      1    14.167 164.87 -565.03
## - F.Undergrad  1    58.782 209.48 -440.97
##
## Step:  AIC=-611.26
## Apps ~ 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     0.145 150.94 -612.76
## - P.Undergrad  1     0.182 150.97 -612.63
## <none>                     150.79 -611.26
## - Top25perc    1     1.594 152.38 -607.81
## - Books        1     1.881 152.67 -606.84
## - perc.alumni  1     2.098 152.89 -606.10
## - PhD          1     3.543 154.33 -601.23
## - Room.Board   1     3.595 154.38 -601.05
## - Expend       1     3.975 154.76 -599.78
## - Outstate     1     4.063 154.85 -599.48
## - S.F.Ratio    1     6.510 157.30 -591.36
## - Grad.Rate    1    11.210 162.00 -576.11
## - Private      1    14.335 165.13 -566.21
## - F.Undergrad  1    58.737 209.53 -442.85
##
## Step:  AIC=-612.76
## Apps ~ 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
## - P.Undergrad  1     0.228 151.16 -613.98
## <none>                     150.94 -612.76
## - Top25perc    1     1.603 152.54 -609.29
## - Books        1     2.104 153.04 -607.59
## - perc.alumni  1     2.198 153.13 -607.27
## - Room.Board   1     3.511 154.45 -602.85
## - PhD          1     3.622 154.56 -602.47
## - Outstate     1     3.934 154.87 -601.43
```

```
## - Expend       1     4.059 154.99 -601.01
## - S.F.Ratio    1     6.392 157.33 -593.28
## - Grad.Rate    1    11.070 162.00 -578.10
## - Private      1    14.227 165.16 -568.10
## - F.Undergrad  1    61.188 212.12 -438.47
##
## Step:  AIC=-613.98
## Apps ~ Private + Top25perc + F.Undergrad + Outstate + Room.Board +
##     Books + PhD + S.F.Ratio + perc.alumni + Expend + Grad.Rate
##
##               Df Sum of Sq    RSS     AIC
## <none>                     151.16 -613.98
## - Top25perc    1     1.496 152.66 -610.88
## - Books        1     2.132 153.29 -608.72
## - perc.alumni  1     2.208 153.37 -608.47
## - Room.Board   1     3.730 154.89 -603.35
## - PhD          1     3.776 154.94 -603.20
## - Outstate     1     3.872 155.03 -602.88
## - Expend       1     4.072 155.23 -602.21
## - S.F.Ratio    1     6.335 157.50 -594.71
## - Grad.Rate    1    10.857 162.02 -580.05
## - Private      1    14.382 165.54 -568.90
## - F.Undergrad  1    77.945 229.11 -400.58
```

```
# forward selection
model.empty <- lm(Apps~1,data=College, subset=train)
model.forward <- step(model.empty,scope=formula(res),direction="forward")
```

```
## Start:  AIC=48.84
## Apps ~ 1
##
##               Df Sum of Sq    RSS     AIC
## + F.Undergrad  1   299.820 267.21 -338.89
## + PhD          1   151.193 415.83 -109.80
## + Terminal     1   142.210 424.82  -98.73
## + Private      1   124.872 442.16  -78.01
## + Top25perc    1    85.724 481.30  -34.06
## + P.Undergrad  1    81.271 485.76  -29.29
## + Top10perc    1    65.054 501.97  -12.28
## + Expend       1    33.178 533.85   19.61
## + Grad.Rate    1    22.606 544.42   29.77
## + Room.Board   1    22.176 544.85   30.18
## + S.F.Ratio    1    21.770 545.26   30.56
## + Personal     1    18.996 548.03   33.19
## + Books        1    17.305 549.72   34.79
## + Outstate     1     5.825 561.20   45.49
## <none>                     567.03   48.84
## + perc.alumni  1     1.985 565.04   49.03
##
## Step:  AIC=-338.89
## Apps ~ F.Undergrad
##
##               Df Sum of Sq    RSS     AIC
## + PhD          1    50.426 216.78 -445.22
```

```
## + Terminal     1    45.407 221.80 -433.37
## + Outstate     1    43.806 223.40 -429.64
## + Room.Board   1    43.153 224.05 -428.13
## + Grad.Rate    1    42.660 224.55 -426.99
## + Top25perc    1    40.400 226.81 -421.80
## + Top10perc    1    38.831 228.38 -418.23
## + Expend       1    33.608 233.60 -406.52
## + perc.alumni  1     6.835 260.37 -350.31
## + Books        1     4.570 262.64 -345.83
## + Personal     1     2.628 264.58 -342.01
## <none>                     267.21 -338.89
## + P.Undergrad  1     0.476 266.73 -337.81
## + S.F.Ratio    1     0.389 266.82 -337.65
## + Private      1     0.325 266.88 -337.52
##
## Step:  AIC=-445.22
## Apps ~ F.Undergrad + PhD
##
##               Df Sum of Sq    RSS     AIC
## + Grad.Rate    1   21.3512 195.43 -496.93
## + Room.Board   1   19.6140 197.17 -492.35
## + Outstate     1   13.9821 202.80 -477.76
## + Top25perc    1   10.5770 206.20 -469.14
## + Expend       1   10.1345 206.65 -468.02
## + Top10perc    1    9.2014 207.58 -465.69
## + Books        1    5.9787 210.80 -457.71
## + Terminal     1    2.6612 214.12 -449.62
## <none>                     216.78 -445.22
## + Personal     1    0.8116 215.97 -445.17
## + Private      1    0.5459 216.24 -444.53
## + S.F.Ratio    1    0.5111 216.27 -444.45
## + P.Undergrad  1    0.2525 216.53 -443.83
## + perc.alumni  1    0.0352 216.75 -443.31
##
## Step:  AIC=-496.93
## Apps ~ F.Undergrad + PhD + Grad.Rate
##
##               Df Sum of Sq    RSS     AIC
## + Room.Board   1    9.3101 186.12 -520.22
## + Private      1    8.1975 187.23 -517.13
## + Books        1    5.1476 190.28 -508.76
## + Expend       1    4.4456 190.98 -506.85
## + Outstate     1    2.9204 192.51 -502.73
## + perc.alumni  1    2.9149 192.51 -502.72
## + S.F.Ratio    1    2.4957 192.93 -501.59
## + Terminal     1    2.2268 193.20 -500.87
## + Top25perc    1    2.0750 193.35 -500.46
## + Top10perc    1    1.5641 193.87 -499.10
## <none>                     195.43 -496.93
## + P.Undergrad  1    0.5608 194.87 -496.42
## + Personal     1    0.0368 195.39 -495.03
##
## Step:  AIC=-520.22
## Apps ~ F.Undergrad + PhD + Grad.Rate + Room.Board
```

```
##
##               Df Sum of Sq    RSS     AIC
## + Private      1   15.9079 170.21 -564.50
## + S.F.Ratio    1    5.3841 180.74 -533.42
## + Books        1    3.1523 182.97 -527.07
## + perc.alumni  1    2.7885 183.33 -526.04
## + Expend       1    1.4572 184.66 -522.29
## + Top25perc    1    1.4247 184.69 -522.20
## + Terminal     1    0.7758 185.34 -520.38
## + Top10perc    1    0.7606 185.36 -520.34
## <none>                     186.12 -520.22
## + P.Undergrad  1    0.2439 185.88 -518.90
## + Personal     1    0.2018 185.92 -518.78
## + Outstate     1    0.0113 186.11 -518.25
##
## Step:  AIC=-564.5
## Apps ~ F.Undergrad + PhD + Grad.Rate + Room.Board + Private
##
##               Df Sum of Sq    RSS     AIC
## + Outstate     1    5.0028 165.21 -577.95
## + Expend       1    4.3888 165.82 -576.03
## + Books        1    2.9664 167.25 -571.61
## + Top25perc    1    2.6474 167.56 -570.62
## + Top10perc    1    2.5489 167.66 -570.31
## + S.F.Ratio    1    1.1122 169.10 -565.90
## + Terminal     1    0.7909 169.42 -564.91
## <none>                     170.21 -564.50
## + perc.alumni  1    0.5020 169.71 -564.03
## + Personal     1    0.1951 170.02 -563.09
## + P.Undergrad  1    0.0534 170.16 -562.66
##
## Step:  AIC=-577.95
## Apps ~ F.Undergrad + PhD + Grad.Rate + Room.Board + Private +
##     Outstate
##
##               Df Sum of Sq    RSS     AIC
## + S.F.Ratio    1    3.1807 162.03 -586.02
## + Books        1    2.9037 162.31 -585.14
## + perc.alumni  1    2.0154 163.19 -582.31
## + Expend       1    1.6554 163.55 -581.17
## + Top25perc    1    1.3798 163.83 -580.30
## + Top10perc    1    1.0378 164.17 -579.22
## <none>                     165.21 -577.95
## + Personal     1    0.4946 164.71 -577.51
## + Terminal     1    0.2716 164.94 -576.80
## + P.Undergrad  1    0.1240 165.09 -576.34
##
## Step:  AIC=-586.02
## Apps ~ F.Undergrad + PhD + Grad.Rate + Room.Board + Private +
##     Outstate + S.F.Ratio
##
##               Df Sum of Sq    RSS     AIC
## + Expend       1    5.0608 156.97 -600.46
## + Books        1    3.2479 158.78 -594.51
```

```
## + Top25perc    1    2.0081 160.02 -590.48
## + Top10perc    1    1.9981 160.03 -590.45
## + perc.alumni  1    1.4618 160.57 -588.72
## + Personal     1    0.8418 161.19 -586.72
## <none>                     162.03 -586.02
## + Terminal     1    0.2278 161.80 -584.75
## + P.Undergrad  1    0.1467 161.88 -584.49
##
## Step:  AIC=-600.46
## Apps ~ F.Undergrad + PhD + Grad.Rate + Room.Board + Private +
##     Outstate + S.F.Ratio + Expend
##
##               Df Sum of Sq    RSS     AIC
## + Books        1   2.64446 154.32 -607.26
## + perc.alumni  1   1.79937 155.17 -604.43
## + Top25perc    1   1.24509 155.72 -602.58
## + Personal     1   0.62572 156.34 -600.53
## <none>                     156.97 -600.46
## + Top10perc    1   0.55438 156.41 -600.29
## + Terminal     1   0.16421 156.80 -599.00
## + P.Undergrad  1   0.15864 156.81 -598.98
##
## Step:  AIC=-607.26
## Apps ~ F.Undergrad + PhD + Grad.Rate + Room.Board + Private +
##     Outstate + S.F.Ratio + Expend + Books
##
##               Df Sum of Sq    RSS     AIC
## + perc.alumni  1   1.66419 152.66 -610.88
## + Top25perc    1   0.95275 153.37 -608.47
## <none>                     154.32 -607.26
## + Top10perc    1   0.33826 153.99 -606.40
## + Personal     1   0.28475 154.04 -606.22
## + P.Undergrad  1   0.14477 154.18 -605.75
## + Terminal     1   0.03016 154.29 -605.36
##
## Step:  AIC=-610.88
## Apps ~ F.Undergrad + PhD + Grad.Rate + Room.Board + Private +
##     Outstate + S.F.Ratio + Expend + Books + perc.alumni
##
##               Df Sum of Sq    RSS     AIC
## + Top25perc    1   1.49631 151.16 -613.98
## + Top10perc    1   0.63203 152.03 -611.03
## <none>                     152.66 -610.88
## + Personal     1   0.18640 152.47 -609.51
## + P.Undergrad  1   0.12093 152.54 -609.29
## + Terminal     1   0.11452 152.54 -609.27
##
## Step:  AIC=-613.98
## Apps ~ F.Undergrad + PhD + Grad.Rate + Room.Board + Private +
##     Outstate + S.F.Ratio + Expend + Books + perc.alumni + Top25perc
##
##               Df Sum of Sq    RSS     AIC
## <none>                     151.16 -613.98
## + P.Undergrad  1  0.227617 150.94 -612.76
```

```
## + Personal     1  0.190028 150.97 -612.63
## + Top10perc    1  0.106703 151.06 -612.35
## + Terminal     1  0.050781 151.11 -612.15
```

The RMSE for the test data was calculated for both backward and forward model as described in question 2(c) and 1(d). Also, the plots of response values and the prediction values were created

```
#Predict
pred.back <- predict(model.backward,College[test,])
pred.forward <- predict(model.forward,College[test,])

#Resulting models with RMSE
sqrt(sum((College$Apps[test]-pred.back)^2)/length(test))
```

```
## [1] 0.6128023
```

```
sqrt(sum((College$Apps[test]-pred.forward)^2)/length(test))
```

```
## [1] 0.6128023
```

```
plot(College[test,"Apps"],pred.back,xlab="y measured",
      ylab="y predicted backward",cex.lab=1.3,
      xlim=c(4,13),ylim=c(4,13))
```

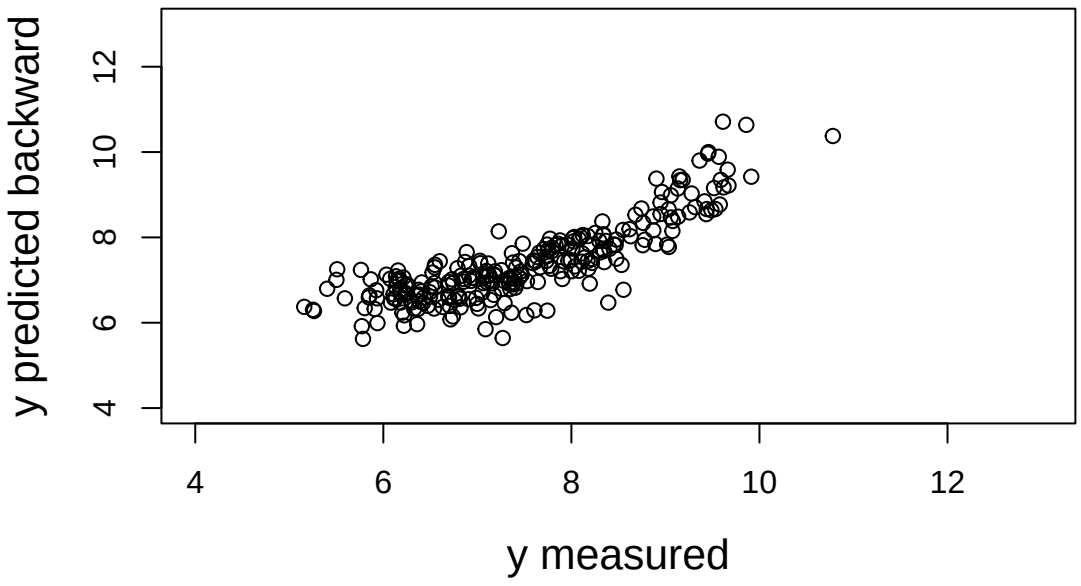

Figure 8: Prediction of the test data for the backward model

```
plot(College[test,"Apps"],pred.forward,xlab="y measured"
     ,ylab="y predicted forward",cex.lab=1.3,
     xlim=c(4,13),ylim=c(4,13))
```

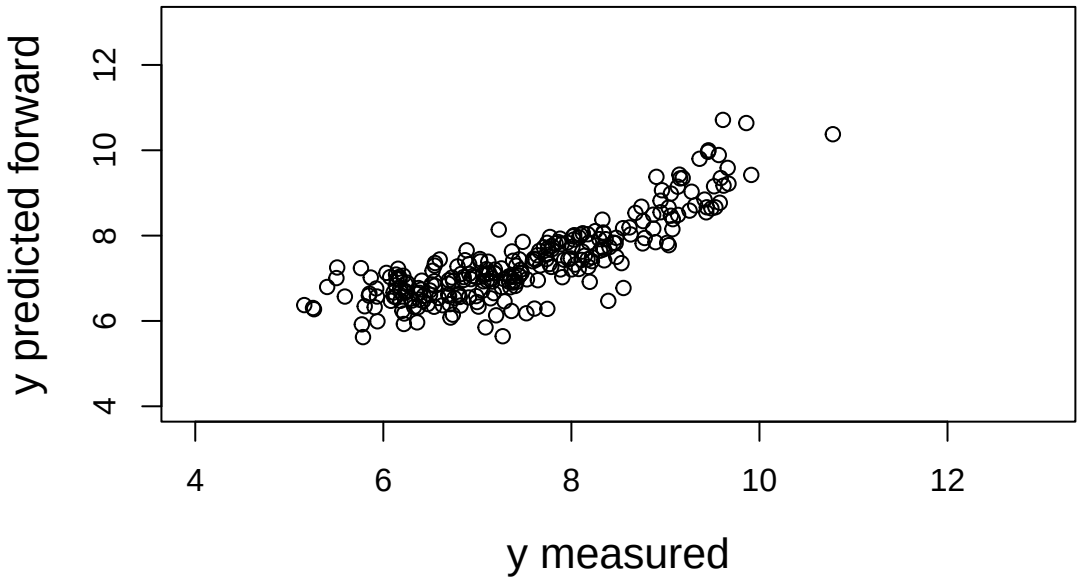

Figure 9: Prediction of the test data for the forward model

In conclusion, according to the RMSE values and the plot, both models' performance is equal. Despite the equality of the models, the RMSE values are significantly low which means that both models have the ability to fit unknown data quite well.