Exercise 1

for Advanced Methods for Regression and Classification

Dzhamilia Kulikieva

23.10.2024

1. Loading and Preprocessing

1a. Loading the College data with ISLR package and investigating the structure and headings:

```r
data(College, package = 'ISLR')
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 ...
```

```r
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
```

The College dataset provides information about 777 colleges, describing their main characteristics. It includes details on college type (private or public), the number of applications, admissions, and enrollments, financial information (tuition, room and board, books), as well as quality indicators such as faculty qualifications and graduation rates.

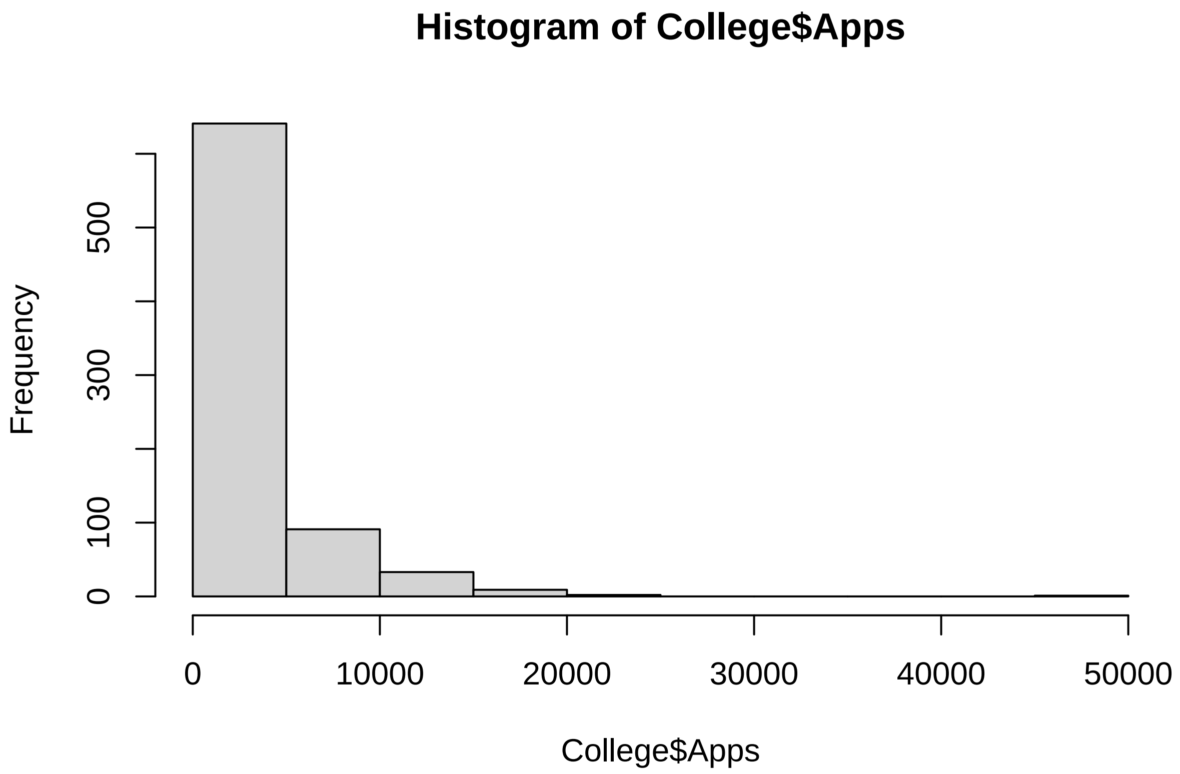
Our goal is to find a linear regression model which allows to predict the variable Apps, i.e. the number of applications received, using the remaining variables except of the variables Accept and Enroll.

1b. Checking if there are missing values in the table and looking at the distribution of the Apps data:

```
sum(is.na(College))
```

```
## [1] 0
```

```
hist(College$Apps)
```

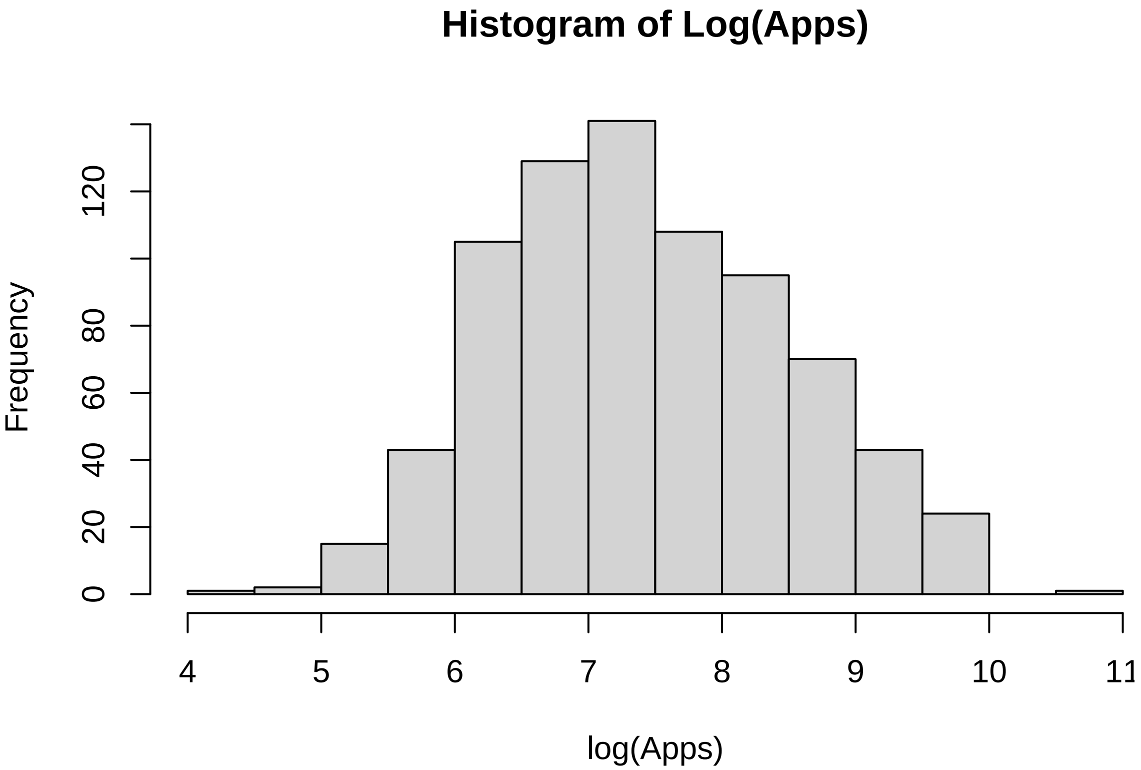


We see that the table has no missing values but the histogram shows that the distribution of the variable College$Apps is highly positively skewed (with a long tail to the right). Most colleges receive a relatively small number of applications, while a few colleges have a much higher number of applications, resulting in this "long tail" effect.

Such skewness can create issues for LRM, as they generally assume normally distributed errors and a balanced influence of all observations. To deal with this skewness and make the data more symmetric, a log transformation is often a suitable approach.

1c. Making a logarithmic transformation of the variable Apps:

```
College$log_Apps <- log(College$Apps + 1)
College$Apps <- NULL
hist(College$log_Apps, main = "Histogram of Log(Apps)", xlab = 'log(Apps)')
```



After the log transformation, the distribution looks much more symmetric, and it is now closer to a normal distribution.

2. Full Model: Estimation the full regression model and interpretion the results

Let's split the data randomly into training and test data (2/3 and 1/3):

```
set.seed(2024)
sample_index <- sample(1:nrow(College), size = floor(2/3 * nrow(College)))
train_data <- College[sample_index, ]
test_data <- College[-sample_index, ]
```

2a. Bilding a complete regression model by using loc_Apps as the dependent variable and investigating the results:

```
model <- lm(log_Apps ~ ., data = train_data)
summary(model)
```

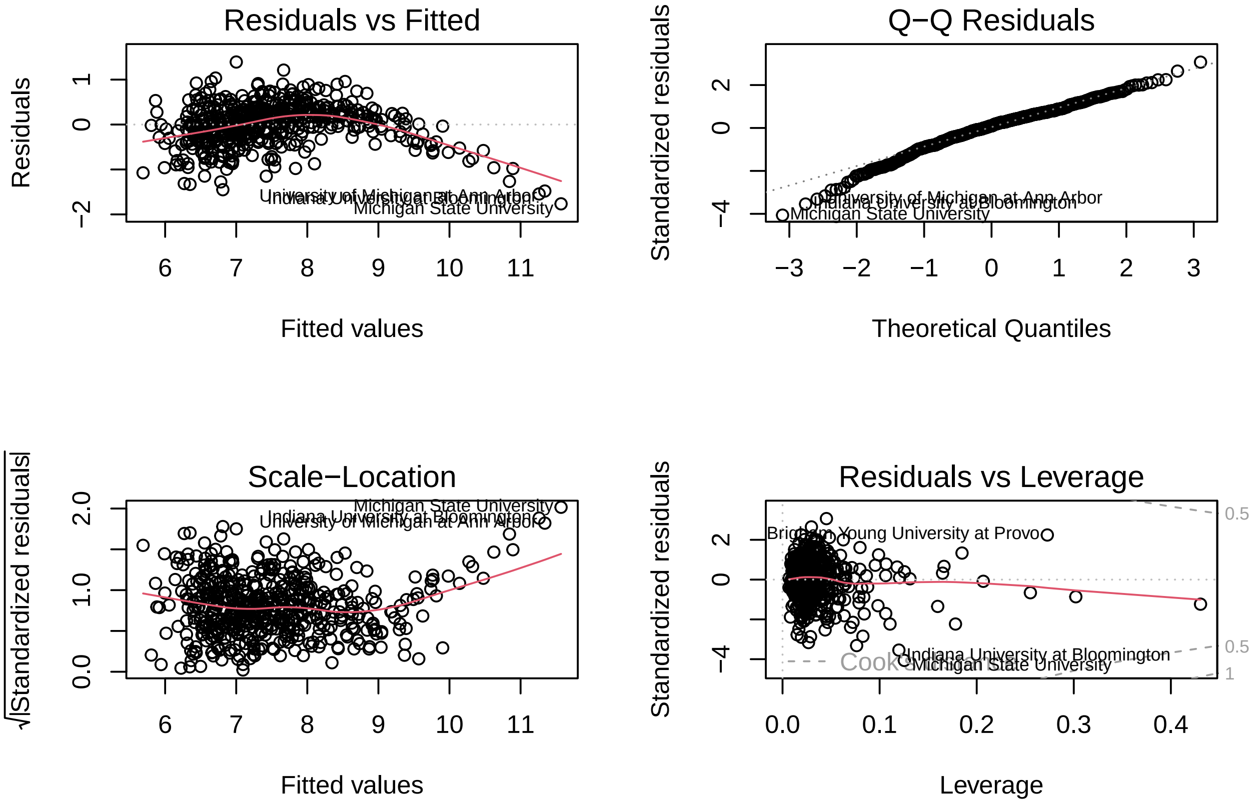
```
##
## Call:
## lm(formula = log_Apps ~ ., data = train_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.76507 -0.25989  0.04369  0.29780  1.39205
##
## Coefficients:
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.373e+00  2.324e-01  18.819  < 2e-16 ***
## PrivateYes  -5.010e-01  7.584e-02  -6.606 1.01e-10 ***
## Accept       2.794e-04  2.922e-05   9.563  < 2e-16 ***
## Enroll       1.408e-04  1.018e-04   1.383 0.167318
## Top10perc    7.771e-03  3.059e-03   2.541 0.011366 *
## Top25perc   -1.095e-03  2.365e-03  -0.463 0.643441
## F.Undergrad -4.089e-05  1.616e-05  -2.530 0.011697 *
## P.Undergrad  3.433e-05  1.607e-05   2.136 0.033162 *
## Outstate     6.158e-06  1.036e-05   0.594 0.552647
## Room.Board   5.775e-05  2.713e-05   2.129 0.033749 *
## Books        4.506e-04  1.340e-04   3.363 0.000829 ***
## Personal     4.564e-05  3.428e-05   1.332 0.183580
## PhD          7.696e-03  2.445e-03   3.147 0.001747 **
## Terminal     2.269e-03  2.805e-03   0.809 0.418832
## S.F.Ratio    4.916e-02  7.268e-03   6.763 3.77e-11 ***
## perc.alumni -2.694e-03  2.318e-03  -1.162 0.245671
## Expend       2.649e-05  6.827e-06   3.880 0.000118 ***
## Grad.Rate    7.463e-03  1.696e-03   4.399 1.33e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4646 on 500 degrees of freedom
## Multiple R-squared:  0.8143, Adjusted R-squared:  0.8079
## F-statistic: 128.9 on 17 and 500 DF,  p-value: < 2.2e-16
```

This linear regression model shows that several variables have a significant impact on the dependent variable log_Apps. The R^2 value (81.43%) indicates that the model describes the data well, and a number of predictors (such as PrivateYes, Accept, PhD, Grad.Rate etc) have a significant impact on the model, which is confirmed by their low p-values. However, some predictors do not have a significant influence, which could be considered when simplifying the model.

Let's embed plots for the model:

```r
par(mfrow=c(2, 2))
plot(model)
```



Conclusions:

- Residuals vs Fitted Plot: We see that the red line is curved, indicating a nonlinear pattern. The points have a smaller spread at the beginning and in the middle of the range of predicted values, the spread increases for larger predicted values, indicating a heteroscedasticity. This suggests that the assumption of linearity and homoscedasticity is violated.
- Q-Q Residuals: Most of the points are close to the diagonal line, indicating that the residuals are approximately normally distributed. However, there are some deviations at the ends, which suggest the presence of outliers or non-normality in the extreme values.
- Scale-Location Plot: The red line shows an increasing trend, and the spread of points also increases as the fitted values increase. This indicates heteroscedasticity—the variance of residuals increases with the fitted values.The assumption of homoscedasticity is not satisfied.
- Residuals vs Leverage Plot: There are several observations with high leverage, such as Brigham Young University at Provo. These observations may have a significant influence on the model.These influential points in the model violate the assumption that the model should not depend heavily on a few observations.

In general, the linear regression model has issues with nonlinearity, heteroscedasticity, influential observations, and minor normality deviations in residuals.

2b. Manually estimating the coefficients:

```
X <- model.matrix(log_Apps ~ ., data = train_data)
y <- train_data$log_Apps
beta_hat <- solve(t(X) %*% X) %*% t(X) %*% y
beta_hat
```

```
##                         [,1]
## (Intercept)  4.373418e+00
## PrivateYes  -5.010121e-01
## Accept       2.794038e-04
## Enroll       1.408171e-04
## Top10perc    7.771179e-03
## Top25perc   -1.095468e-03
## F.Undergrad -4.088829e-05
## P.Undergrad  3.432874e-05
## Outstate     6.157822e-06
## Room.Board   5.774935e-05
## Books        4.506036e-04
## Personal     4.564332e-05
## PhD          7.695916e-03
## Terminal     2.269431e-03
## S.F.Ratio    4.915613e-02
## perc.alumni -2.694389e-03
## Expend       2.649155e-05
## Grad.Rate    7.462697e-03
```

R handles binary variables by creating indicator variable, such as PrivateYes:

1 if the college is private (Private = "Yes"). 0 if the college is public (Private = "No")

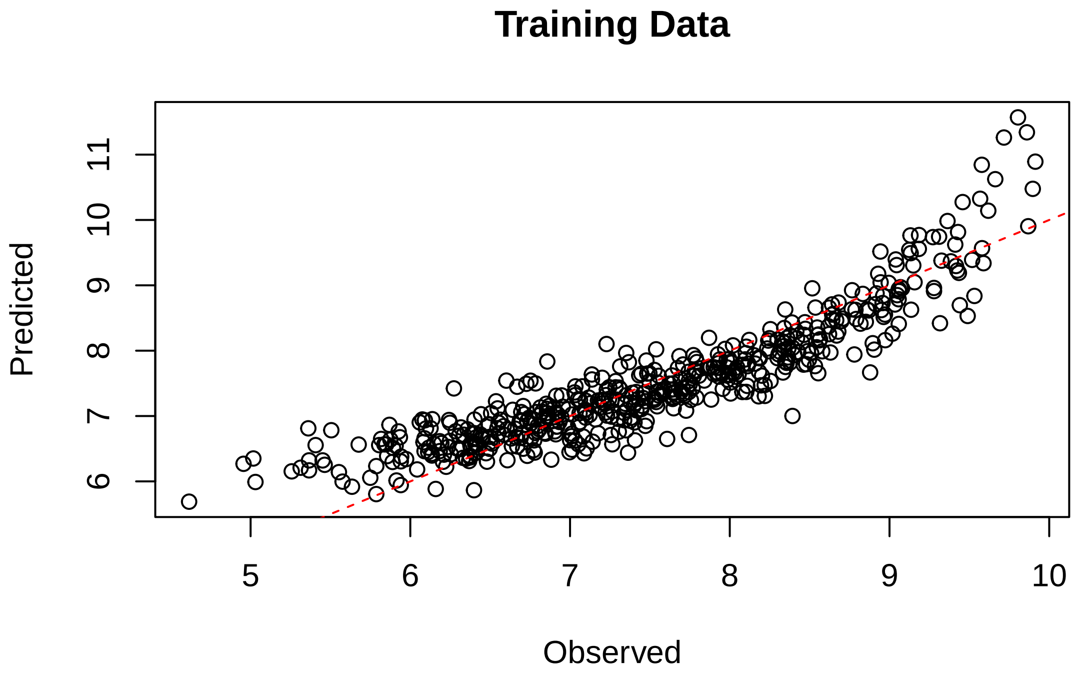
If a college is private (PrivateYes = 1), the expected value of log_Apps decreases by 0.501 (-5.010121e-01) units compared to a public college (PrivateYes = 0), assuming all other variables are held constant.

The coefficients obtained from lm() match the manually calculated coefficients using model.matrix()

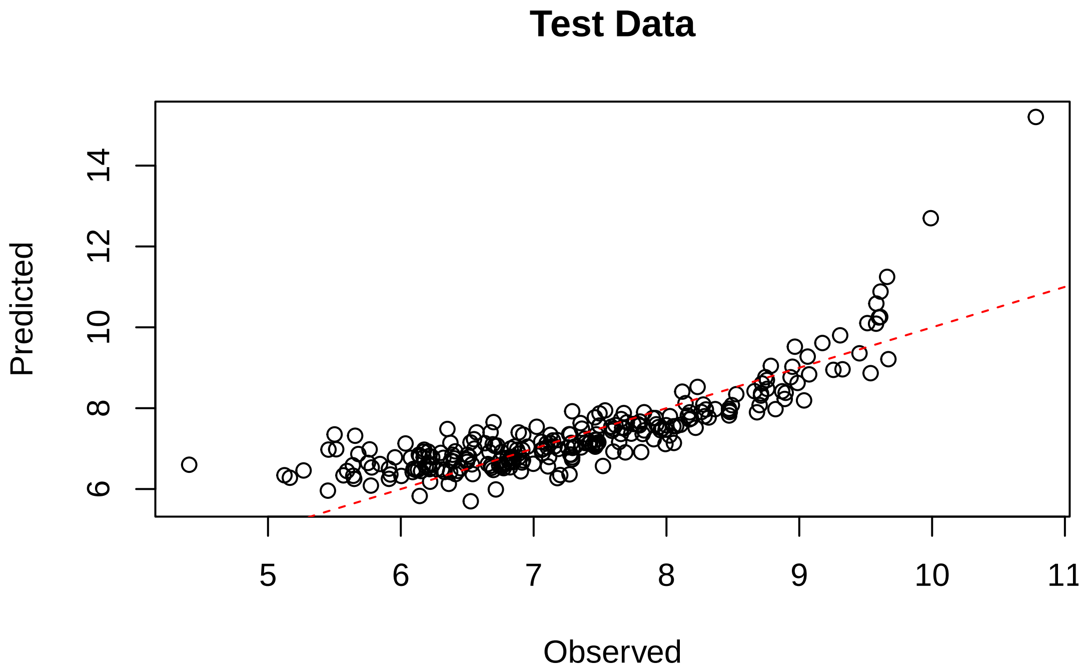
2c. Comparing graphically the observed and the predicted values of log_Apps for the train_data and for test_data:

```r
predicted_train <- predict(model, train_data)
predicted_test <- predict(model, test_data)
```

```r
plot(train_data$log_Apps, predicted_train, xlab = 'Observed', ylab = 'Predicted',
main = 'Training Data')
abline(0,1, col = 'red', lty = 2)
```



```r
plot(test_data$log_Apps, predicted_test, xlab = 'Observed', ylab = 'Predicted',
main = 'Test Data')
abline(0,1, col = 'red', lty = 2)
```



For the train data:

- Points are mostly clustered along the red line, indicating that the model predicts values reasonably well on the training data. However, there are some points that are far from the line, which indicates model errors for those observations.
- Points are relatively tightly distributed along the line in the central region, meaning the model predicts adequately for most observations. However, there are some distant points, indicating outliers or cases where the model struggles to predict well for certain values of the variables.

For the test data:

- Unlike the plot for the training data, here the points are less tightly clustered along the red line, and there are more outliers, especially in the area with high predicted values (above 10). This indicates that the model has higher errors on the test data compared to the training data, which may suggest overfitting.
- In the central part of the plot, the points are still grouped along the ideal line, indicating that the model is able to predict values adequately for most observations. However, there are outliers, such as the point above y = 14, which indicates difficulties for the model in predicting certain observations.

2d. Computing the RMSE separately for training and test data:

```
rmse_train <- sqrt(mean((train_data$log_Apps - predicted_train)^2))
rmse_test <- sqrt(mean((test_data$log_Apps - predicted_test)^2))

rmse_train
```

```
## [1] 0.4564835
```

```
rmse_test
```

```
## [1] 0.6340315
```

Conclusions:

The RMSE value for the test data is higher than for the training data (0.6340 vs 0.4565). This indicates that the model performs better on the data it was trained on than on new data. The difference in RMSE suggests that the model may exhibit signs of overfitting — meaning it does not generalize as well to new data.

On the other hand, a low RMSE value for the train data indicates that the model errors are small, and the model makes good predictions. The RMSE for the test data (0.6340) is also not too high, but it is higher than for the training data, which may be expected since the test data was not used for training.

3. Reduced model: Exclution not significant variables and computing the LS-estimator

```
reduced_model <- lm(log_Apps ~ Private + Accept + Top10perc + F.Undergrad + P.Undergrad
                    + Room.Board + Books + PhD + S.F.Ratio + Expend + Grad.Rate,
                    data = train_data)
summary(reduced_model)
```

```
##
## Call:
## lm(formula = log_Apps ~ Private + Accept + Top10perc + F.Undergrad +
##     P.Undergrad + Room.Board + Books + PhD + S.F.Ratio + Expend +
##     Grad.Rate, data = train_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.71989 -0.24804  0.05017  0.31016  1.35247
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.506e+00  2.082e-01  21.638  < 2e-16 ***
## PrivateYes  -5.150e-01  7.020e-02  -7.336 8.84e-13 ***
## Accept       3.052e-04  2.232e-05  13.675  < 2e-16 ***
## Top10perc    6.624e-03  1.872e-03   3.539 0.000438 ***
## F.Undergrad -2.508e-05  1.185e-05  -2.116 0.034866 *
## P.Undergrad  3.830e-05  1.588e-05   2.411 0.016255 *
## Room.Board   5.764e-05  2.460e-05   2.343 0.019512 *
## Books        5.046e-04  1.303e-04   3.872 0.000122 ***
## PhD          9.033e-03  1.659e-03   5.445 8.08e-08 ***
## S.F.Ratio    4.800e-02  7.178e-03   6.686 6.07e-11 ***
## Expend       2.853e-05  6.462e-06   4.415 1.23e-05 ***
## Grad.Rate    6.858e-03  1.622e-03   4.227 2.81e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4645 on 506 degrees of freedom
## Multiple R-squared:  0.8121, Adjusted R-squared:  0.808
## F-statistic: 198.8 on 11 and 506 DF,  p-value: < 2.2e-16
```

3a. Comparing results: full model vs redused model:

In the reduced model, all input variables are now significant at the 0.05 level.

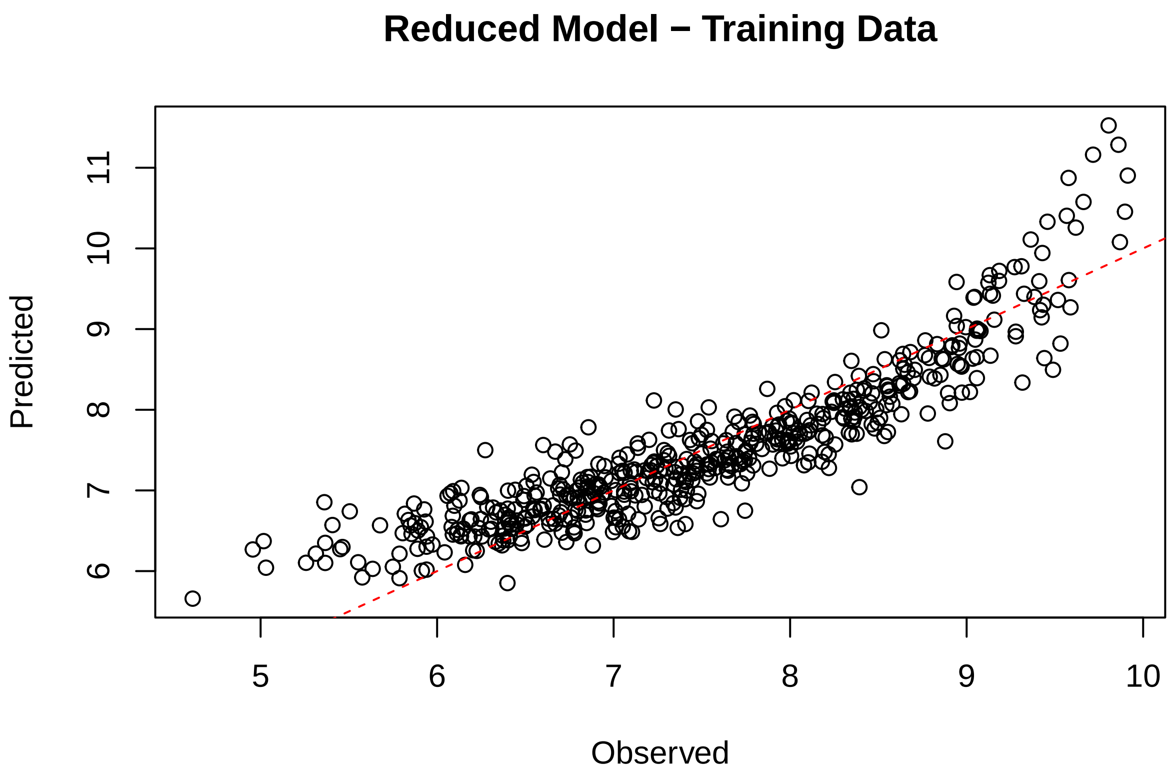
However, it's not always expected that all input variables will be significant in a reduced model. Variables can act as confounders, and removing them can change the relationships of others, possibly making them non-significant. Additionally, depending on how the data is split into training and test sets, or how samples are collected, different predictors may appear significant or non-significant.

Furthermore, due to interdependence between variables, when certain variables are removed, it can impact the statistical significance of the remaining variables.

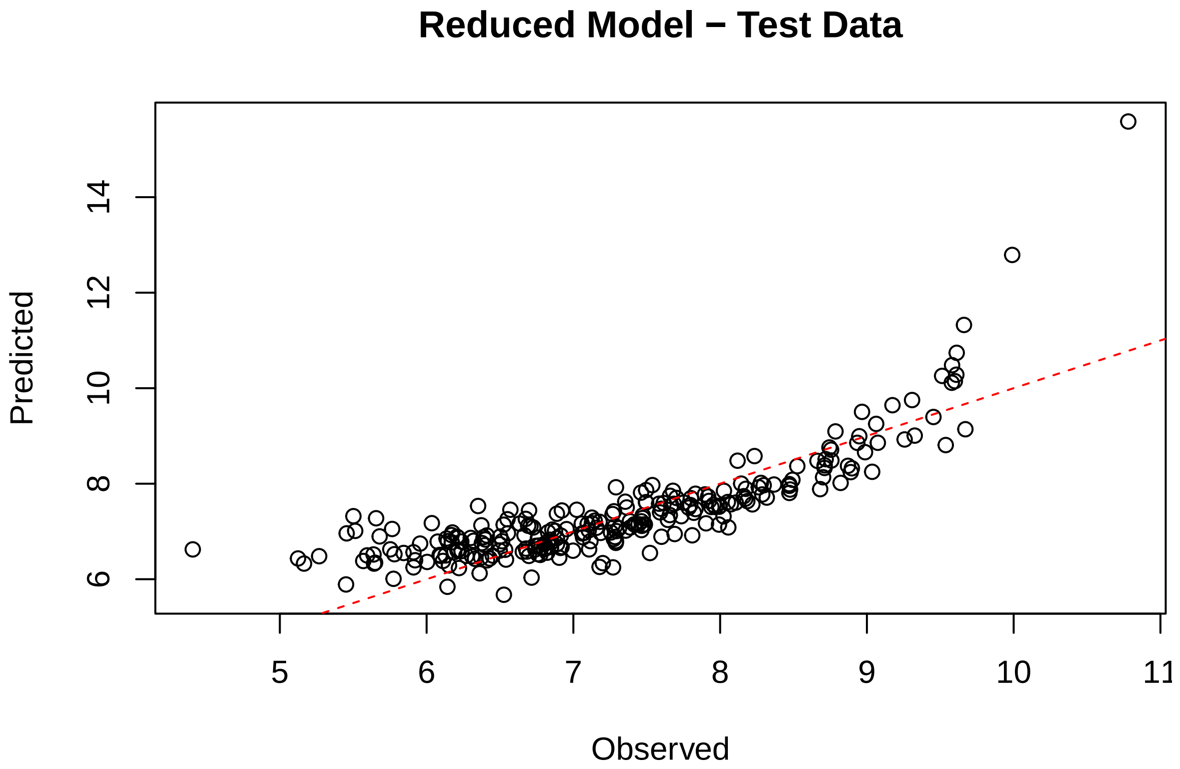
3b. Visualising the fit and the prediction from a new model:

```r
predicted_train_reduced <- predict(reduced_model, train_data)
predicted_test_reduced <- predict(reduced_model, test_data)

plot(train_data$log_Apps, predicted_train_reduced, xlab = "Observed", ylab = "Predicted",
     main = "Reduced Model - Training Data")
abline(0, 1, col = "red", lty = 2)
```



```r
plot(test_data$log_Apps, predicted_test_reduced, xlab = "Observed", ylab = "Predicted",
     main = "Reduced Model - Test Data")
abline(0, 1, col = "red", lty = 2)
```



3c. Computing the RMSE for the new model:

```
rmse_train_reduced <- sqrt(mean((train_data$log_Apps - predicted_train_reduced)^2))
rmse_test_reduced <- sqrt(mean((test_data$log_Apps - predicted_test_reduced)^2))

rmse_train_reduced
```

```
## [1] 0.4591263
```

```
rmse_test_reduced
```

```
## [1] 0.6479654
```

For the reduced model, we would expect similar or slightly higher RMSE, because removing less significant variables could lead to a small increase in error but reduce model complexity, making it easier to interpret.

In our case, the RMSE for the reduced model is indeed slightly higher compared to the full model(0.4564835, 0.6340315), confirming this expectation.

3d. Comparing two models with anova:

```
anova(model, reduced_model)
```

```
## Analysis of Variance Table
##
## Model 1: log_Apps ~ Private + Accept + Enroll + Top10perc + Top25perc +
##     F.Undergrad + P.Undergrad + Outstate + Room.Board + Books +
##     Personal + PhD + Terminal + S.F.Ratio + perc.alumni + Expend +
##     Grad.Rate
## Model 2: log_Apps ~ Private + Accept + Top10perc + F.Undergrad + P.Undergrad +
##     Room.Board + Books + PhD + S.F.Ratio + Expend + Grad.Rate
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    500 107.94
## 2    506 109.19 -6   -1.2535 0.9677 0.4464
```

The reduced model has more degrees of freedom (506 vs. 500 in the full model), making it simpler with fewer parameters to estimate and less prone to overfitting.

The reduced model's RSS is 109.19, slightly higher than the full model's 107.94, indicating a small increase in model error after removing some variables — a trade-off between complexity and accuracy.

The p-value of 0.4464 is much greater than the typical significance level (e.g., 0.05), and with F = 0.9677 (not large), it means that the reduced model is not significantly worse than the full model. Therefore, the reduced model is preferred for its simpler structure without a significant loss in quality.

4. Performing variable selection based on stepwise regression, using the function step()

```
forward_model <- step(lm(log_Apps ~ 1, data = train_data), direction = "forward",
                      scope = formula(model))
```

```
## Start:  AIC=61.55
## log_Apps ~ 1
##
##               Df Sum of Sq    RSS     AIC
## + Accept       1    395.36 185.74 -527.28
## + Enroll       1    351.01 230.10 -416.34
## + F.Undergrad  1    315.65 265.45 -342.30
## + PhD          1    171.96 409.14 -118.20
## + Terminal     1    160.92 420.18 -104.41
## + Private      1    135.77 445.33  -74.30
## + P.Undergrad  1     92.98 488.12  -26.77
## + Top25perc    1     80.73 500.37  -13.94
## + Top10perc    1     75.66 505.44   -8.71
## + Expend       1     48.53 532.57   18.37
## + Room.Board   1     27.26 553.84   38.66
## + Grad.Rate    1     17.88 563.22   47.36
## + Personal     1     14.60 566.50   50.36
## + Books        1     13.64 567.46   51.24
## + S.F.Ratio    1     11.68 569.43   53.03
## + Outstate     1      8.56 572.55   55.86
## + perc.alumni  1      4.51 576.59   59.51
## <none>                     581.10   61.55
##
## Step:  AIC=-527.28
## log_Apps ~ Accept
##
##               Df Sum of Sq    RSS     AIC
## + PhD          1    36.789 148.95 -639.61
## + Terminal     1    31.712 154.03 -622.25
## + Top10perc    1    25.363 160.38 -601.33
## + Top25perc    1    21.051 164.69 -587.58
## + Expend       1    16.815 168.93 -574.43
## + Grad.Rate    1    11.897 173.84 -559.56
## + Room.Board   1    10.420 175.32 -555.18
## + Outstate     1     8.100 177.64 -548.37
## + Private      1     4.564 181.18 -538.16
## + Books        1     2.835 182.91 -533.24
## + perc.alumni  1     2.569 183.17 -532.49
## <none>                     185.74 -527.28
## + P.Undergrad  1     0.562 185.18 -526.84
## + Enroll       1     0.328 185.41 -526.19
## + S.F.Ratio    1     0.283 185.46 -526.06
## + F.Undergrad  1     0.030 185.71 -525.36
## + Personal     1     0.000 185.74 -525.28
##
## Step:  AIC=-639.61
## log_Apps ~ Accept + PhD
##
##               Df Sum of Sq    RSS     AIC
## + Private      1    5.4548 143.50 -656.94
## + Top10perc    1    4.9720 143.98 -655.20
## + Books        1    4.4439 144.51 -653.30
## + S.F.Ratio    1    3.3529 145.60 -649.41
## + Expend       1    3.2673 145.68 -649.10
## + Top25perc    1    2.9054 146.05 -647.82
## + Grad.Rate    1    2.5334 146.42 -646.50
## + Room.Board   1    1.7422 147.21 -643.71
## + Terminal     1    1.4512 147.50 -642.68
## + P.Undergrad  1    0.8865 148.06 -640.70
## <none>                     148.95 -639.61
## + Enroll       1    0.5231 148.43 -639.43
## + Personal     1    0.3371 148.61 -638.79
## + F.Undergrad  1    0.1673 148.78 -638.19
## + perc.alumni  1    0.1431 148.81 -638.11
## + Outstate     1    0.1415 148.81 -638.10
##
## Step:  AIC=-656.94
## log_Apps ~ Accept + PhD + Private
##
##               Df Sum of Sq    RSS     AIC
## + Top10perc    1   11.2073 132.29 -697.06
## + Expend       1    9.6609 133.84 -691.04
## + Grad.Rate    1    8.9032 134.59 -688.12
## + Room.Board   1    7.8689 135.63 -684.15
## + Outstate     1    7.0344 136.46 -680.98
## + Top25perc    1    6.4300 137.07 -678.69
## + Books        1    5.1324 138.36 -673.81
## + Terminal     1    1.8122 141.68 -661.52
## + F.Undergrad  1    0.6287 142.87 -657.21
## + S.F.Ratio    1    0.6233 142.87 -657.19
## <none>                     143.50 -656.94
## + perc.alumni  1    0.4735 143.02 -656.65
## + P.Undergrad  1    0.0705 143.43 -655.19
## + Enroll       1    0.0388 143.46 -655.08
## + Personal     1    0.0001 143.50 -654.94
##
## Step:  AIC=-697.06
## log_Apps ~ Accept + PhD + Private + Top10perc
##
##               Df Sum of Sq    RSS     AIC
## + Room.Board   1    5.5805 126.71 -717.39
## + Grad.Rate    1    4.1281 128.16 -711.48
## + S.F.Ratio    1    3.7890 128.50 -710.12
## + Books        1    3.3530 128.94 -708.36
## + Expend       1    2.2938 130.00 -704.12
## + Outstate     1    2.2713 130.02 -704.03
## + F.Undergrad  1    1.2856 131.00 -700.12
## + Terminal     1    0.9202 131.37 -698.68
## + P.Undergrad  1    0.6063 131.68 -697.44
## <none>                     132.29 -697.06
## + Enroll       1    0.5035 131.79 -697.04
## + perc.alumni  1    0.1598 132.13 -695.69
## + Top25perc    1    0.1323 132.16 -695.58
## + Personal     1    0.0138 132.28 -695.12
##
## Step:  AIC=-717.39
## log_Apps ~ Accept + PhD + Private + Top10perc + Room.Board
##
##               Df Sum of Sq    RSS     AIC
## + S.F.Ratio    1    4.9149 121.79 -735.88
## + Grad.Rate    1    2.5058 124.20 -725.73
## + Books        1    2.4299 124.28 -725.42
## + Expend       1    0.8991 125.81 -719.08
## <none>                     126.71 -717.39
## + F.Undergrad  1    0.4443 126.26 -717.21
## + P.Undergrad  1    0.3838 126.33 -716.96
## + Outstate     1    0.2543 126.45 -716.43
## + Personal     1    0.2327 126.48 -716.34
## + Terminal     1    0.2265 126.48 -716.31
## + perc.alumni  1    0.0938 126.61 -715.77
## + Top25perc    1    0.0910 126.62 -715.76
## + Enroll       1    0.0033 126.70 -715.40
## 
## Step:  AIC=-735.88
## log_Apps ~ Accept + PhD + Private + Top10perc + Room.Board + 
##     S.F.Ratio
## 
##               Df Sum of Sq    RSS     AIC
## + Expend       1    4.7275 117.07 -754.39
## + Grad.Rate    1    2.7677 119.03 -745.79
## + Books        1    2.5979 119.20 -745.05
## + Outstate     1    1.1067 120.69 -738.61
## + F.Undergrad  1    0.9137 120.88 -737.78
## + Personal     1    0.4736 121.32 -735.90
## <none>                     121.79 -735.88
## + Terminal     1    0.3618 121.43 -735.42
## + P.Undergrad  1    0.3547 121.44 -735.39
## + Top25perc    1    0.2540 121.54 -734.96
## + Enroll       1    0.0315 121.76 -734.01
## + perc.alumni  1    0.0066 121.79 -733.91
## 
## Step:  AIC=-754.39
## log_Apps ~ Accept + PhD + Private + Top10perc + Room.Board + 
##     S.F.Ratio + Expend
## 
##               Df Sum of Sq    RSS     AIC
## + Grad.Rate    1   3.01365 114.05 -765.90
## + Books        1   2.28511 114.78 -762.60
## + F.Undergrad  1   0.72854 116.34 -755.62
## <none>                     117.07 -754.39
## + Personal     1   0.37006 116.70 -754.03
## + Outstate     1   0.33739 116.73 -753.88
## + P.Undergrad  1   0.26581 116.80 -753.57
## + Terminal     1   0.20416 116.86 -753.29
## + perc.alumni  1   0.02549 117.04 -752.50
## + Enroll       1   0.02169 117.05 -752.48
## + Top25perc    1   0.00184 117.06 -752.40
## 
## Step:  AIC=-765.9
## log_Apps ~ Accept + PhD + Private + Top10perc + Room.Board + 
##     S.F.Ratio + Expend + Grad.Rate
## 
##               Df Sum of Sq    RSS     AIC
## + Books        1    3.1718 110.88 -778.51
## + Personal     1    0.9290 113.12 -768.13
## + P.Undergrad  1    0.9145 113.14 -768.07
## <none>                     114.05 -765.90
## + perc.alumni  1    0.4028 113.65 -765.73
## + Terminal     1    0.2474 113.81 -765.02
## + F.Undergrad  1    0.2368 113.82 -764.97
## + Outstate     1    0.0303 114.02 -764.03
## + Enroll       1    0.0226 114.03 -764.00
## + Top25perc    1    0.0138 114.04 -763.96
##
## Step:  AIC=-778.51
## log_Apps ~ Accept + PhD + Private + Top10perc + Room.Board +
##     S.F.Ratio + Expend + Grad.Rate + Books
##
##               Df Sum of Sq    RSS     AIC
## + P.Undergrad  1   0.72214 110.16 -779.89
## + Personal     1   0.43478 110.45 -778.54
## + F.Undergrad  1   0.43338 110.45 -778.54
## <none>                     110.88 -778.51
## + perc.alumni  1   0.27671 110.60 -777.80
## + Terminal     1   0.06220 110.82 -776.80
## + Top25perc    1   0.05986 110.82 -776.79
## + Outstate     1   0.05985 110.82 -776.79
## + Enroll       1   0.00030 110.88 -776.51
##
## Step:  AIC=-779.89
## log_Apps ~ Accept + PhD + Private + Top10perc + Room.Board +
##     S.F.Ratio + Expend + Grad.Rate + Books + P.Undergrad
##
##               Df Sum of Sq    RSS     AIC
## + F.Undergrad  1   0.96586 109.19 -782.45
## <none>                     110.16 -779.89
## + perc.alumni  1   0.26914 109.89 -779.16
## + Personal     1   0.25406 109.91 -779.09
## + Outstate     1   0.08616 110.07 -778.30
## + Top25perc    1   0.07249 110.09 -778.23
## + Enroll       1   0.05061 110.11 -778.13
## + Terminal     1   0.04980 110.11 -778.13
##
## Step:  AIC=-782.45
## log_Apps ~ Accept + PhD + Private + Top10perc + Room.Board +
##     S.F.Ratio + Expend + Grad.Rate + Books + P.Undergrad + F.Undergrad
##
##               Df Sum of Sq    RSS     AIC
## <none>                     109.19 -782.45
## + Personal     1   0.41421 108.78 -782.42
## + Enroll       1   0.33786 108.86 -782.06
## + perc.alumni  1   0.23732 108.96 -781.58
## + Terminal     1   0.07758 109.11 -780.82
## + Top25perc    1   0.06860 109.12 -780.78
## + Outstate     1   0.01331 109.18 -780.52
```

```r
summary(forward_model)
```

```
##
## Call:
## lm(formula = log_Apps ~ Accept + PhD + Private + Top10perc +
##     Room.Board + S.F.Ratio + Expend + Grad.Rate + Books + P.Undergrad +
##     F.Undergrad, data = train_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.71989 -0.24804  0.05017  0.31016  1.35247
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.506e+00  2.082e-01  21.638  < 2e-16 ***
## Accept       3.052e-04  2.232e-05  13.675  < 2e-16 ***
## PhD          9.033e-03  1.659e-03   5.445 8.08e-08 ***
## PrivateYes  -5.150e-01  7.020e-02  -7.336 8.84e-13 ***
## Top10perc    6.624e-03  1.872e-03   3.539 0.000438 ***
## Room.Board   5.764e-05  2.460e-05   2.343 0.019512 *
## S.F.Ratio    4.800e-02  7.178e-03   6.686 6.07e-11 ***
## Expend       2.853e-05  6.462e-06   4.415 1.23e-05 ***
## Grad.Rate    6.858e-03  1.622e-03   4.227 2.81e-05 ***
## Books        5.046e-04  1.303e-04   3.872 0.000122 ***
## P.Undergrad  3.830e-05  1.588e-05   2.411 0.016255 *
## F.Undergrad -2.508e-05  1.185e-05  -2.116 0.034866 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4645 on 506 degrees of freedom
## Multiple R-squared:  0.8121, Adjusted R-squared:  0.808
## F-statistic: 198.8 on 11 and 506 DF,  p-value: < 2.2e-16
```

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

```
## Start:  AIC=-776.43
## log_Apps ~ Private + Accept + Enroll + 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
## - Top25perc    1    0.0463 107.99 -778.21
## - Outstate     1    0.0762 108.02 -778.07
## - Terminal     1    0.1413 108.08 -777.76
## - perc.alumni  1    0.2916 108.23 -777.04
## - Personal     1    0.3828 108.32 -776.60
## - Enroll       1    0.4128 108.35 -776.46
## <none>                    107.94 -776.43
## - Room.Board   1    0.9784 108.92 -773.76
## - P.Undergrad  1    0.9850 108.92 -773.73
## - F.Undergrad  1    1.3823 109.32 -771.84
## - Top10perc    1    1.3935 109.33 -771.79
## - PhD          1    2.1382 110.08 -768.27
## - Books        1    2.4422 110.38 -766.84
## - Expend       1    3.2504 111.19 -763.07
## - Grad.Rate    1    4.1773 112.12 -758.77
## - Private      1    9.4218 117.36 -735.08
## - S.F.Ratio    1    9.8754 117.81 -733.09
## - Accept       1   19.7433 127.68 -691.42
##
```

```
## Step:  AIC=-778.21
## log_Apps ~ Private + Accept + Enroll + Top10perc + 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
## - Outstate     1    0.0803 108.07 -779.83
## - Terminal     1    0.1168 108.10 -779.65
## - perc.alumni  1    0.3127 108.30 -778.71
## - Personal     1    0.3903 108.38 -778.34
## <none>                     107.99 -778.21
## - Enroll       1    0.4367 108.42 -778.12
## - P.Undergrad  1    0.9729 108.96 -775.57
## - Room.Board   1    0.9861 108.97 -775.50
## - F.Undergrad  1    1.4079 109.39 -773.50
## - PhD          1    2.1542 110.14 -769.98
## - Books        1    2.4127 110.40 -768.77
## - Top10perc    1    2.6136 110.60 -767.82
## - Expend       1    3.5946 111.58 -763.25
## - Grad.Rate    1    4.1500 112.14 -760.68
## - Private      1    9.4563 117.44 -736.73
## - S.F.Ratio    1    9.8917 117.88 -734.81
## - Accept       1   19.7160 127.70 -693.34
##
## Step:  AIC=-779.83
## log_Apps ~ Private + Accept + Enroll + Top10perc + F.Undergrad +
##     P.Undergrad + Room.Board + Books + Personal + PhD + Terminal +
##     S.F.Ratio + perc.alumni + Expend + Grad.Rate
##
##               Df Sum of Sq    RSS     AIC
## - Terminal     1    0.1404 108.21 -781.15
## - perc.alumni  1    0.2639 108.33 -780.56
## - Personal     1    0.3646 108.43 -780.08
## - Enroll       1    0.4119 108.48 -779.86
## <none>                     108.07 -779.83
## - P.Undergrad  1    0.9850 109.05 -777.13
## - Room.Board   1    1.2956 109.36 -775.65
## - F.Undergrad  1    1.4847 109.55 -774.76
## - PhD          1    2.1951 110.26 -771.41
## - Books        1    2.4056 110.47 -770.42
## - Top10perc    1    2.7556 110.82 -768.78
## - Expend       1    4.0459 112.11 -762.79
## - Grad.Rate    1    4.3809 112.45 -761.24
## - S.F.Ratio    1    9.8115 117.88 -736.81
## - Private      1    9.8656 117.93 -736.57
## - Accept       1   21.2825 129.35 -688.71
##
## Step:  AIC=-781.15
## log_Apps ~ Private + Accept + Enroll + Top10perc + F.Undergrad +
##     P.Undergrad + Room.Board + Books + Personal + PhD + S.F.Ratio +
##     perc.alumni + Expend + Grad.Rate
##
##               Df Sum of Sq    RSS     AIC
## - perc.alumni  1    0.2373 108.44 -782.02
## - Personal     1    0.3380 108.54 -781.54
## - Enroll       1    0.3984 108.61 -781.25
## <none>                     108.21 -781.15
## - P.Undergrad  1    1.0023 109.21 -778.38
## - F.Undergrad  1    1.4317 109.64 -776.35
## - Room.Board   1    1.4710 109.68 -776.16
## - Books        1    2.6392 110.85 -770.67
## - Top10perc    1    2.7502 110.96 -770.15
## - Expend       1    4.1623 112.37 -763.60
## - Grad.Rate    1    4.3561 112.56 -762.71
## - PhD          1    6.5781 114.78 -752.58
## - S.F.Ratio    1    9.7562 117.96 -738.44
## - Private      1   10.1209 118.33 -736.84
## - Accept       1   21.2793 129.49 -690.16
##
## Step:  AIC=-782.02
## log_Apps ~ Private + Accept + Enroll + Top10perc + F.Undergrad +
##     P.Undergrad + Room.Board + Books + Personal + PhD + S.F.Ratio +
##     Expend + Grad.Rate
##
##               Df Sum of Sq    RSS     AIC
## - Enroll       1    0.3350 108.78 -782.42
## - Personal     1    0.4113 108.86 -782.06
## <none>                     108.44 -782.02
## - P.Undergrad  1    1.0068 109.45 -779.23
## - F.Undergrad  1    1.3862 109.83 -777.44
## - Room.Board   1    1.5880 110.03 -776.49
## - Top10perc    1    2.5721 111.02 -771.88
## - Books        1    2.7249 111.17 -771.16
## - Expend       1    4.0699 112.51 -764.93
## - Grad.Rate    1    4.1189 112.56 -764.71
## - PhD          1    6.3753 114.82 -754.43
## - S.F.Ratio    1   10.0439 118.49 -738.14
## - Private      1   11.2554 119.70 -732.87
## - Accept       1   22.5729 131.02 -686.07
##
## Step:  AIC=-782.42
## log_Apps ~ Private + Accept + Top10perc + F.Undergrad + P.Undergrad +
##     Room.Board + Books + Personal + PhD + S.F.Ratio + Expend +
##     Grad.Rate
##
##               Df Sum of Sq    RSS     AIC
## - Personal     1     0.414 109.19 -782.45
## <none>                     108.78 -782.42
## - P.Undergrad  1     1.032 109.81 -779.53
## - F.Undergrad  1     1.126 109.91 -779.09
## - Room.Board   1     1.370 110.15 -777.94
## - Top10perc    1     2.782 111.56 -771.34
## - Books        1     2.801 111.58 -771.25
## - Grad.Rate    1     4.086 112.86 -765.32
## - Expend       1     4.127 112.91 -765.13
## - PhD          1     6.363 115.14 -754.98
## - S.F.Ratio    1     9.936 118.72 -739.15
## - Private      1    11.360 120.14 -732.97
## - Accept       1    40.763 149.54 -619.56
##
## Step:  AIC=-782.45
## log_Apps ~ Private + Accept + Top10perc + F.Undergrad + P.Undergrad +
##     Room.Board + Books + PhD + S.F.Ratio + Expend + Grad.Rate
##
##               Df Sum of Sq    RSS     AIC
## <none>                     109.19 -782.45
## - F.Undergrad  1     0.966 110.16 -779.89
## - Room.Board   1     1.185 110.38 -778.86
## - P.Undergrad  1     1.255 110.45 -778.54
## - Top10perc    1     2.703 111.90 -771.79
## - Books        1     3.236 112.43 -769.33
## - Grad.Rate    1     3.856 113.05 -766.48
## - Expend       1     4.207 113.40 -764.87
## - PhD          1     6.398 115.59 -754.96
## - S.F.Ratio    1     9.647 118.84 -740.60
## - Private      1    11.612 120.81 -732.10
## - Accept       1    40.356 149.55 -621.54
```

```
summary(backward_model)
```

```
##
## Call:
## lm(formula = log_Apps ~ Private + Accept + Top10perc + F.Undergrad +
##     P.Undergrad + Room.Board + Books + PhD + S.F.Ratio + Expend +
##     Grad.Rate, data = train_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.71989 -0.24804  0.05017  0.31016  1.35247
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.506e+00  2.082e-01  21.638  < 2e-16 ***
## PrivateYes  -5.150e-01  7.020e-02  -7.336 8.84e-13 ***
## Accept       3.052e-04  2.232e-05  13.675  < 2e-16 ***
## Top10perc    6.624e-03  1.872e-03   3.539 0.000438 ***
## F.Undergrad -2.508e-05  1.185e-05  -2.116 0.034866 *
## P.Undergrad  3.830e-05  1.588e-05   2.411 0.016255 *
## Room.Board   5.764e-05  2.460e-05   2.343 0.019512 *
## Books        5.046e-04  1.303e-04   3.872 0.000122 ***
## PhD          9.033e-03  1.659e-03   5.445 8.08e-08 ***
## S.F.Ratio    4.800e-02  7.178e-03   6.686 6.07e-11 ***
## Expend       2.853e-05  6.462e-06   4.415 1.23e-05 ***
## Grad.Rate    6.858e-03  1.622e-03   4.227 2.81e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4645 on 506 degrees of freedom
## Multiple R-squared:  0.8121, Adjusted R-squared:  0.808
## F-statistic: 198.8 on 11 and 506 DF,  p-value: < 2.2e-16
```

We see that there isn't a difference in terms of significant/non-significant predictors between the two models and they both ended up selecting only the statistically significant predictors that contributed meaningfully to the model.

Let's compare the resulting models with the RMSE, and with plots of response vs predicted values.

For the forward selection model:

```
predicted_forward_train <- predict(forward_model, train_data)
predicted_forward_test <- predict(forward_model, test_data)

rmse_train_forward <- sqrt(mean((train_data$log_Apps - predicted_forward_train)^2))
rmse_test_forward <- sqrt(mean((test_data$log_Apps - predicted_forward_test)^2))
rmse_train_forward
```

```
## [1] 0.4591263
```

```
rmse_test_forward
```

```
## [1] 0.6479654
```

For the backward selection model:

```
predicted_backward_train <- predict(backward_model, train_data)
predicted_backward_test <- predict(backward_model, test_data)

rmse_train_backward <- sqrt(mean((train_data$log_Apps - predicted_backward_train)^2))
rmse_test_backward <- sqrt(mean((test_data$log_Apps - predicted_backward_test)^2))

rmse_train_backward
```

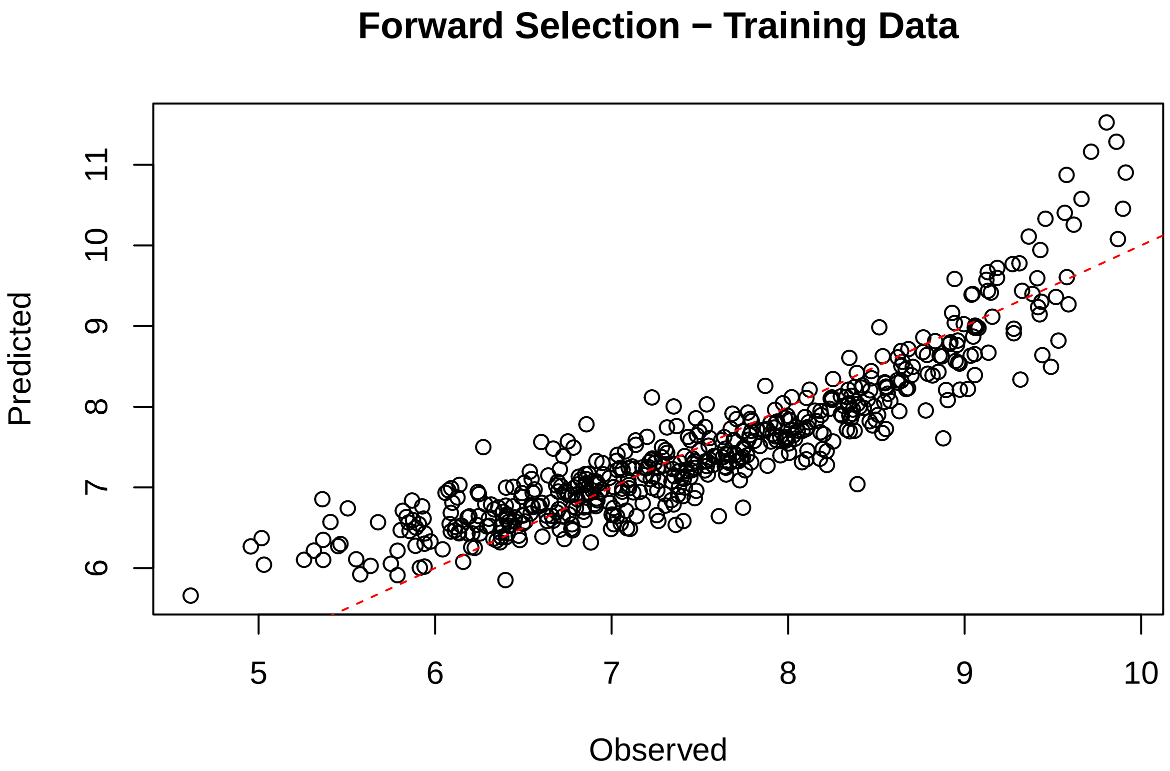
```
## [1] 0.4591263
```

```
rmse_test_backward
```

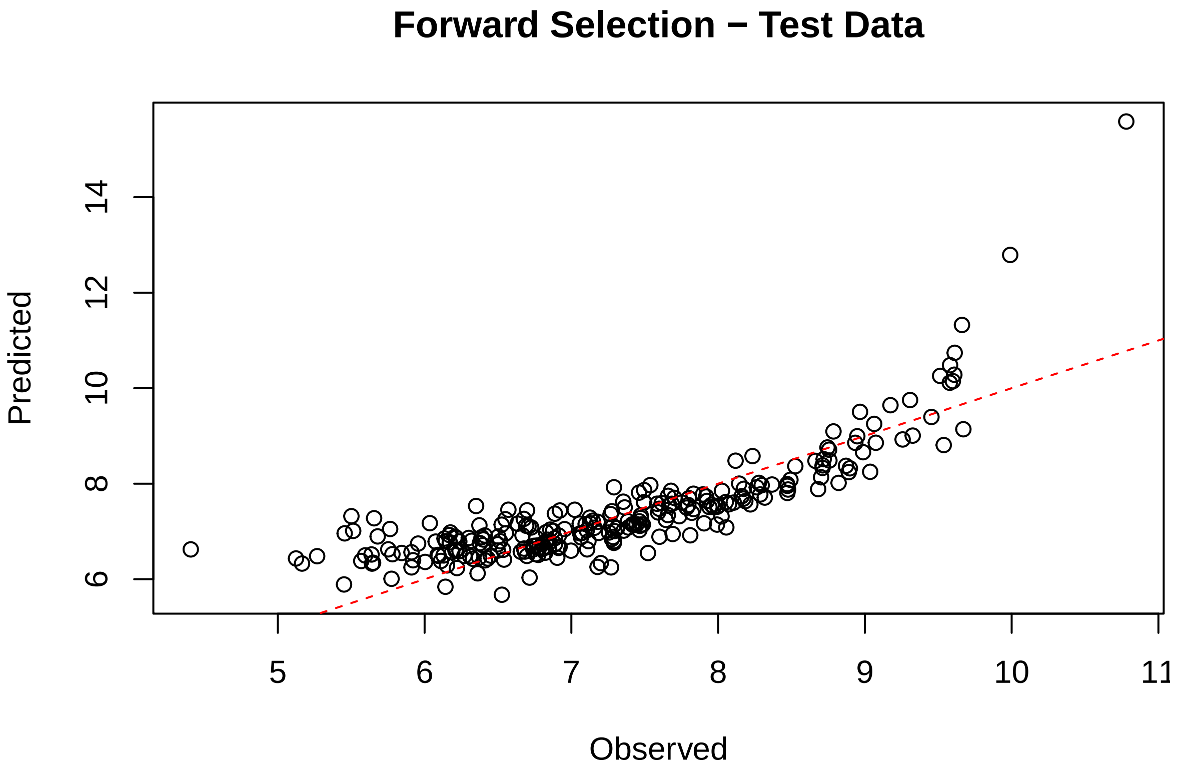
```
## [1] 0.6479654
```

The RMSE values indicate that both models have the same error for training and test data, suggesting similar predictive performance.

```
plot(train_data$log_Apps, predicted_forward_train, xlab = "Observed", ylab = "Predicted",
     main = "Forward Selection - Training Data")
abline(0, 1, col = "red", lty = 2)
```

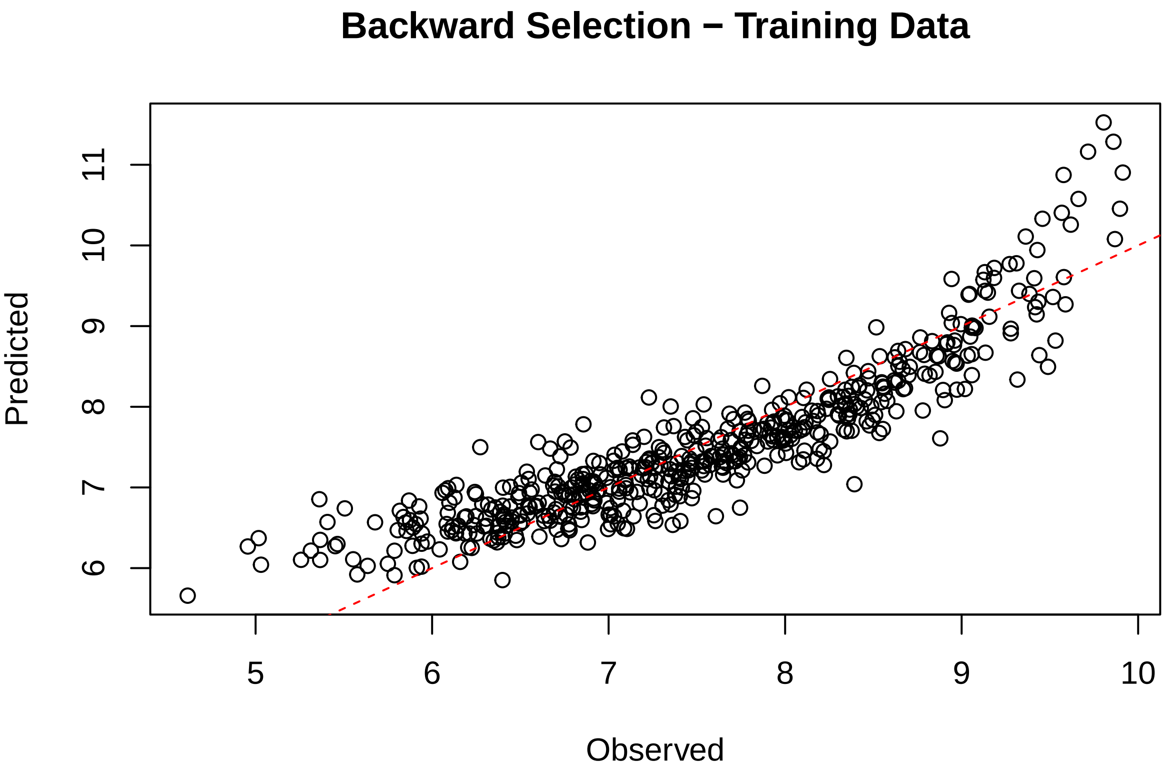


```
plot(test_data$log_Apps, predicted_forward_test, xlab = "Observed", ylab = "Predicted",
     main = "Forward Selection - Test Data")
abline(0, 1, col = "red", lty = 2)
```

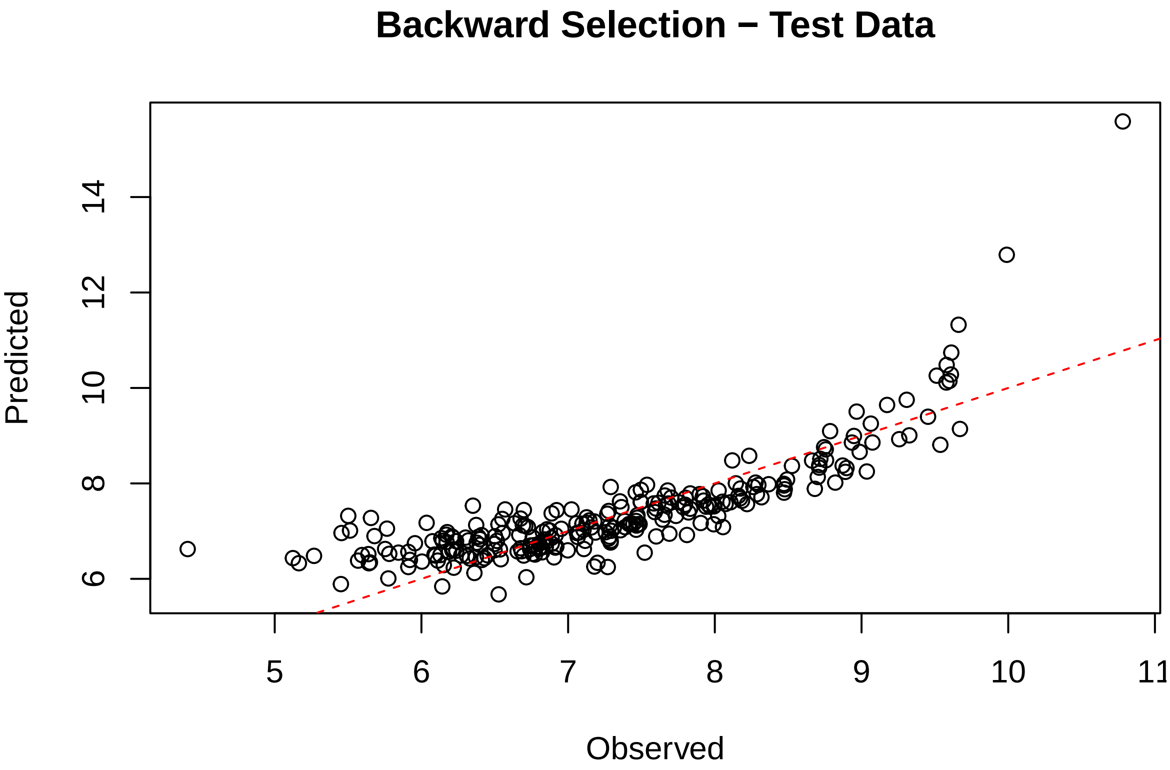


```
plot(train_data$log_Apps, predicted_backward_train, xlab = "Observed", ylab = "Predicted",
     main = "Backward Selection - Training Data")
```

```
abline(0, 1, col = "red", lty = 2)
```



```
plot(test_data$log_Apps, predicted_backward_test, xlab = "Observed", ylab = "Predicted",
     main = "Backward Selection - Test Data")
abline(0, 1, col = "red", lty = 2)
```



Based on the provided plots and RMSE results, we conclude that both models have identical predictors and yield similar outcomes. Therefore, there are no significant differences in their predictive ability, and both models have similar levels of error on both the training and test datasets.

In general, the results for the reduced model, created using the linear model (lm) with only the significant predictors, match the models obtained through both forward and backward selection. This indicates that all three approaches resulted in identical models, confirming consistency in the selection of significant variables across the different methods.