Exercise 1

Nikolaus Czernin

```r
# install.packages("ISLR")
library("ISLR")
library("tidyverse")
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library("knitr")

data(College,package="ISLR")
saveRDS(College, file = "College.rds")
```

```r
?College
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
summary(College)
```

```
##  Private        Apps           Accept          Enroll       Top10perc
##  No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00
##  Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00
##            Median : 1558   Median : 1110   Median : 434   Median :23.00
##            Mean   : 3002   Mean   : 2019   Mean   : 780   Mean   :27.56
##            3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:35.00
##            Max.   :48094   Max.   :26330   Max.   :6392   Max.   :96.00
##    Top25perc      F.Undergrad     P.Undergrad         Outstate
##  Min.   :  9.0   Min.   :  139   Min.   :    1.0   Min.   : 2340
##  1st Qu.: 41.0   1st Qu.:  992   1st Qu.:   95.0   1st Qu.: 7320
##  Median : 54.0   Median : 1707   Median :  353.0   Median : 9990
##  Mean   : 55.8   Mean   : 3700   Mean   :  855.3   Mean   :10441
##  3rd Qu.: 69.0   3rd Qu.: 4005   3rd Qu.:  967.0   3rd Qu.:12925
##  Max.   :100.0   Max.   :31643   Max.   :21836.0   Max.   :21700
##    Room.Board       Books           Personal         PhD
##  Min.   :1780   Min.   :  96.0   Min.   : 250   Min.   :  8.00
##  1st Qu.:3597   1st Qu.: 470.0   1st Qu.: 850   1st Qu.: 62.00
##  Median :4200   Median : 500.0   Median :1200   Median : 75.00
##  Mean   :4358   Mean   : 549.4   Mean   :1341   Mean   : 72.66
##  3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.:1700   3rd Qu.: 85.00
##  Max.   :8124   Max.   :2340.0   Max.   :6800   Max.   :103.00
##     Terminal       S.F.Ratio      perc.alumni        Expend
##  Min.   : 24.0   Min.   : 2.50   Min.   : 0.00   Min.   : 3186
##  1st Qu.: 71.0   1st Qu.:11.50   1st Qu.:13.00   1st Qu.: 6751
##  Median : 82.0   Median :13.60   Median :21.00   Median : 8377
##  Mean   : 79.7   Mean   :14.09   Mean   :22.74   Mean   : 9660
##  3rd Qu.: 92.0   3rd Qu.:16.50   3rd Qu.:31.00   3rd Qu.:10830
##  Max.   :100.0   Max.   :39.80   Max.   :64.00   Max.   :56233
##    Grad.Rate
##  Min.   : 10.00
##  1st Qu.: 53.00
##  Median : 65.00
##  Mean   : 65.46
##  3rd Qu.: 78.00
##  Max.   :118.00
```

Regarding `Apps`, the mean is twice as high as the median, suggesting the data is left-skewed. By log-transforming the data, we reduce the effect of high numbers and hopefully make the model more robust.

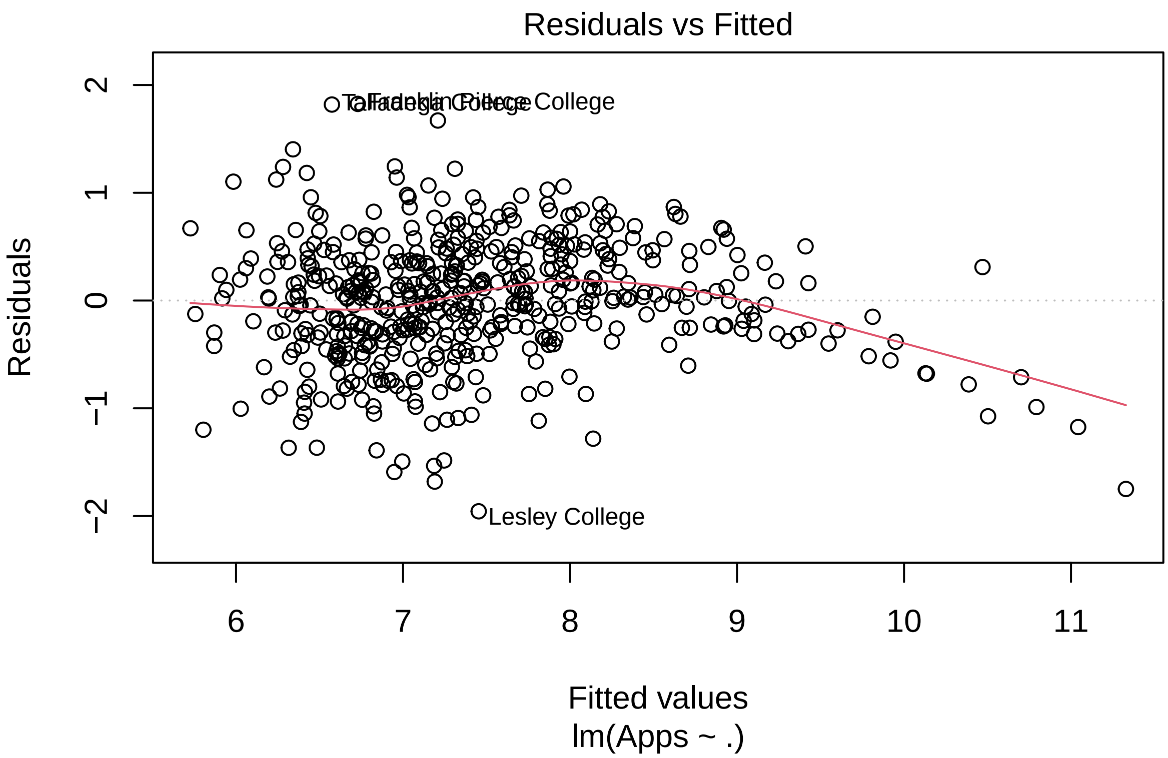
```r
College_processed <- College %>%
  mutate(Apps = log(Apps)) %>%
  select(-Accept, -Enroll)

n <- nrow(College_processed)
idx <- sample(1:n, n%/%3*2)
train = College_processed[idx,]
test =  College_processed[-idx,]
```

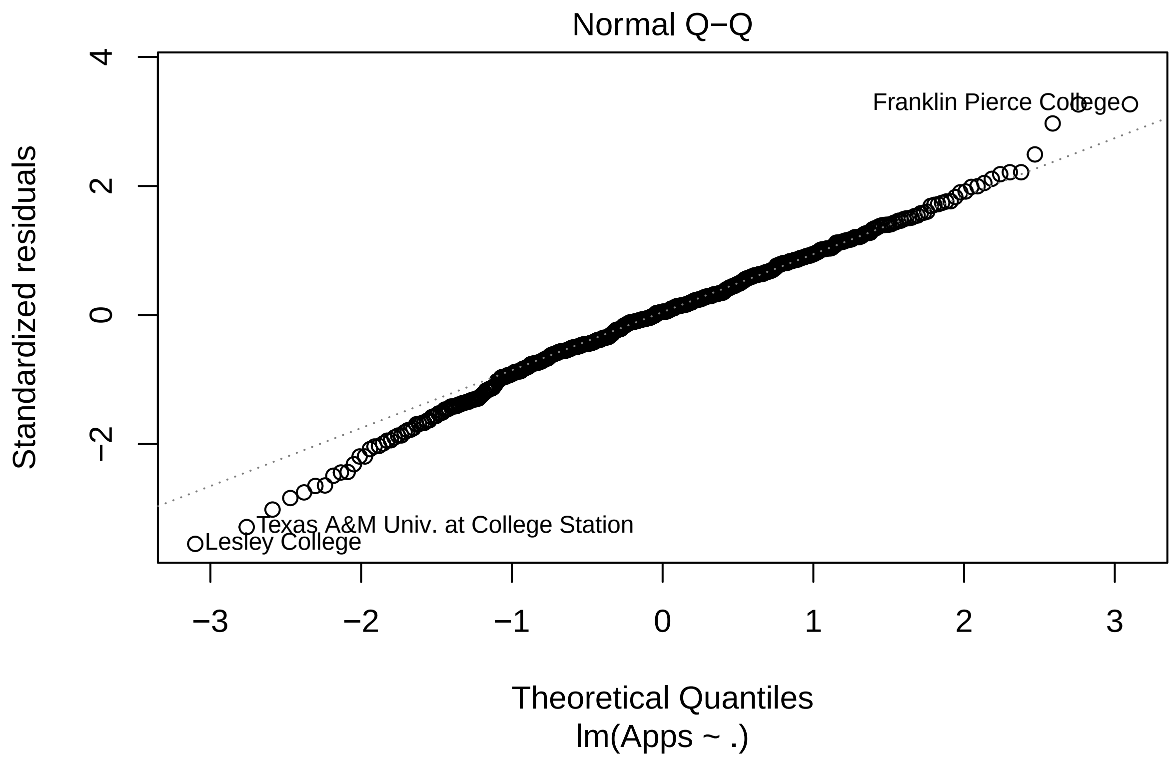
2. Full Model

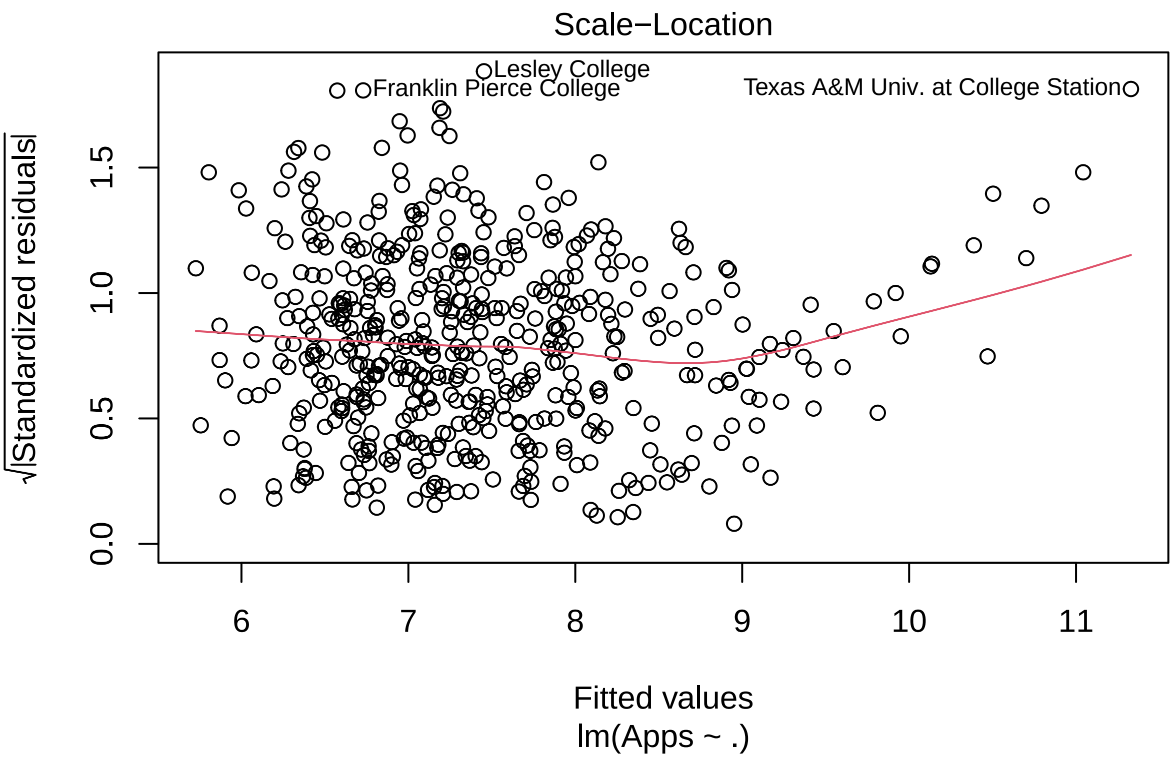
```r
full.lm <- lm(Apps ~ ., data = train)
full.lm %>% summary()
```

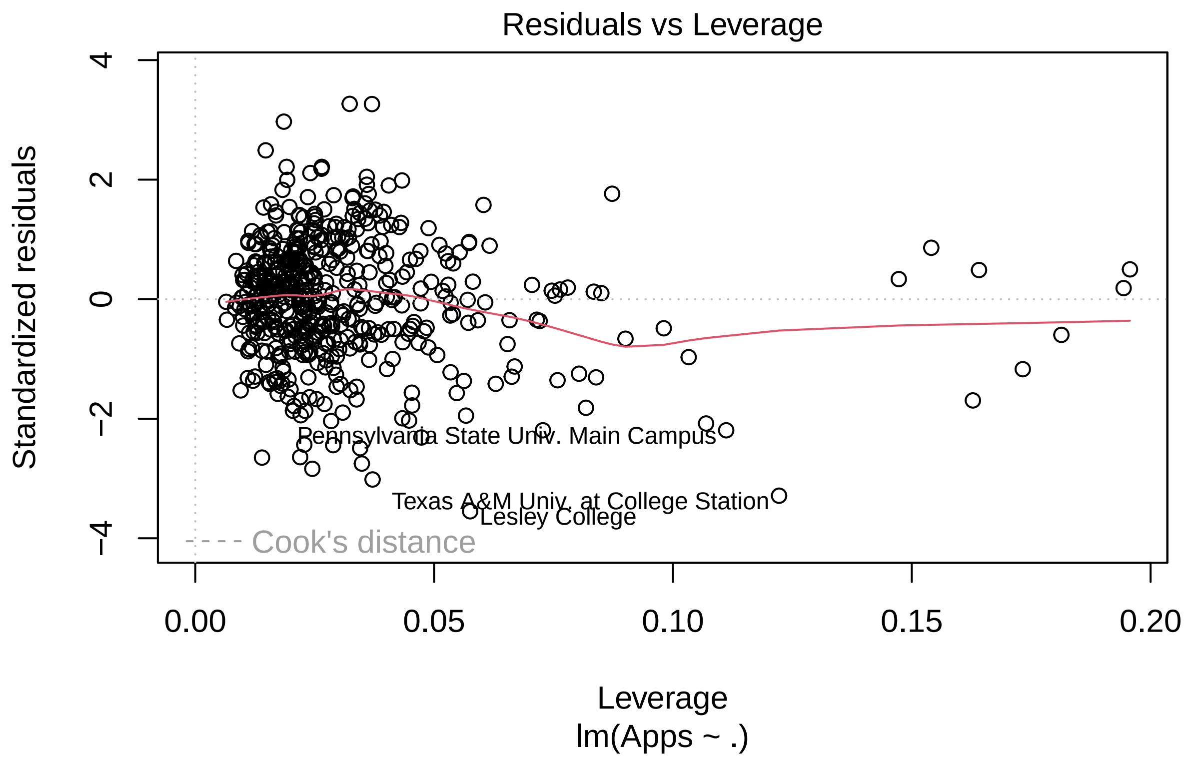
```
##
## Call:
## lm(formula = Apps ~ ., data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.95576 -0.31598  0.02826  0.36506  1.82463
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.254e+00  2.671e-01  15.926  < 2e-16 ***
## PrivateYes  -5.657e-01  9.278e-02  -6.098 2.15e-09 ***
## Top10perc   -1.163e-04  3.792e-03  -0.031  0.97554
## Top25perc    4.527e-03  2.978e-03   1.520  0.12917
## F.Undergrad  1.121e-04  8.801e-06  12.737  < 2e-16 ***
## P.Undergrad  1.359e-05  2.862e-05   0.475  0.63521
## Outstate     4.107e-05  1.316e-05   3.120  0.00191 **
## Room.Board   5.706e-05  3.388e-05   1.684  0.09279 .
## Books        3.705e-04  1.708e-04   2.169  0.03055 *
## Personal     5.639e-06  4.055e-05   0.139  0.88945
## PhD          7.496e-03  3.173e-03   2.363  0.01851 *
## Terminal     1.926e-03  3.393e-03   0.568  0.57057
## S.F.Ratio    3.895e-02  8.608e-03   4.524 7.57e-06 ***
## perc.alumni -8.444e-03  2.843e-03  -2.970  0.00312 **
## Expend       2.721e-05  1.152e-05   2.361  0.01859 *
## Grad.Rate    1.078e-02  1.985e-03   5.430 8.79e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5677 on 502 degrees of freedom
## Multiple R-squared:  0.7253, Adjusted R-squared:  0.7171
## F-statistic: 88.38 on 15 and 502 DF,  p-value: < 2.2e-16
```

```r
full.lm %>% plot()
```


Residuals vs Fitted
Talladega College
Franklin Pierce College
Lesley College
Residuals
Fitted values
lm(Apps ~ .)
−2
−1
0
1
2
6
7
8
9
10
11


Normal Q−Q
Standardized residuals
4
2
0
−2
Franklin Pierce College
Texas A&M Univ. at College Station
Lesley College
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(Apps ~ .)


Scale−Location
Lesley College
Franklin Pierce College
Texas A&M Univ. at College Station
√|Standardized residuals|
0.0
0.5
1.0
1.5
6
7
8
9
10
11
Fitted values
lm(Apps ~ .)



The Residuals vs Fitted plot shows how poorly the model is performing. The red line not being straight is a sign, that the variance is not constant, which is also observableby the residual points being more spread out on the left hand side than the right end.
The QQ-Plot shows that the residuals are somewhat normally distributed.

Manual computation of the coefficients

```
X <- model.matrix(Apps ~ ., data = train)
# get the manual estimator
full.estimator <- solve(t(X) %*% X) %*% (t(X) %*% train$Apps)

# bind it to the coefficients of the lm function
summary(full.lm) %>% .$coefficients %>% .[,1] %>% cbind(full.estimator)
```

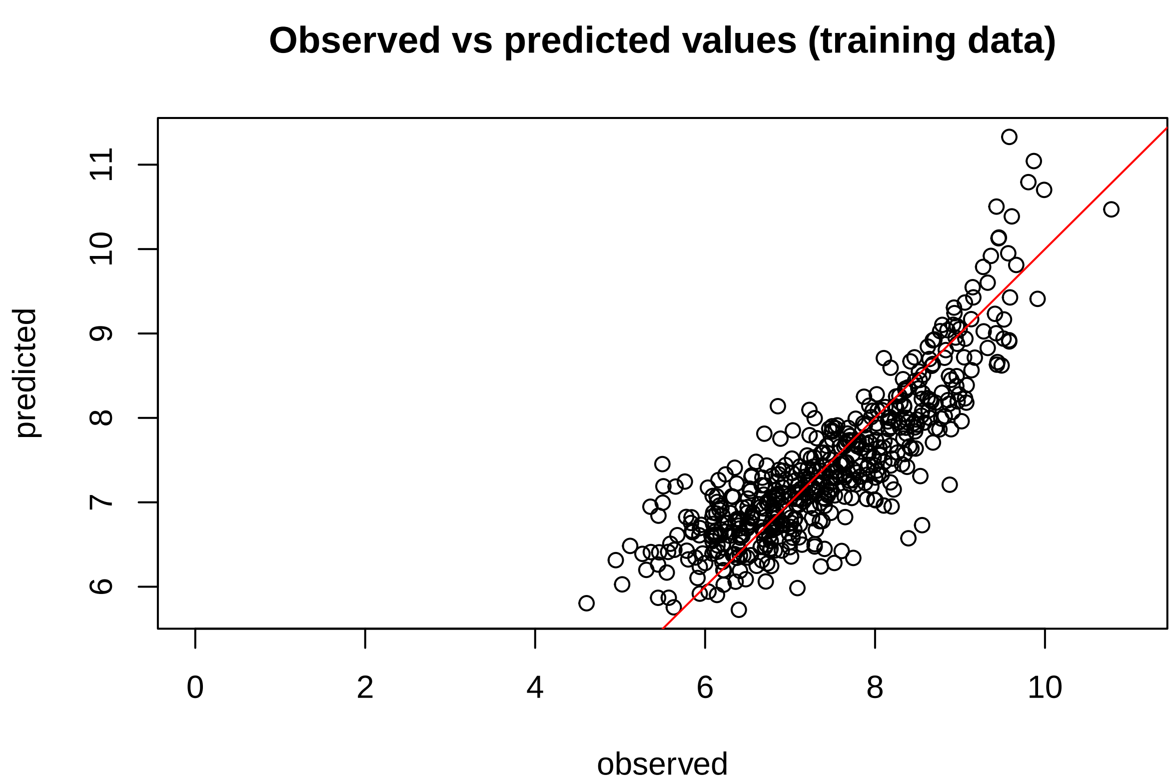
```
##                           .
## (Intercept)  4.253587e+00  4.253587e+00
## PrivateYes  -5.657271e-01 -5.657271e-01
## Top10perc   -1.163063e-04 -1.163063e-04
## Top25perc    4.526543e-03  4.526543e-03
## F.Undergrad  1.120963e-04  1.120963e-04
## P.Undergrad  1.358643e-05  1.358643e-05
## Outstate     4.106487e-05  4.106487e-05
## Room.Board   5.705838e-05  5.705838e-05
## Books        3.704522e-04  3.704522e-04
```

```
## Personal      5.639464e-06  5.639464e-06
## PhD           7.496296e-03  7.496296e-03
## Terminal      1.925552e-03  1.925552e-03
## S.F.Ratio     3.894692e-02  3.894692e-02
## perc.alumni -8.444168e-03 -8.444168e-03
## Expend        2.721288e-05  2.721288e-05
## Grad.Rate     1.078001e-02  1.078001e-02
```

The coefficients of the lm() function and the manual estimation are equal.
`PrivateYes` is a variable with highly significant coefficient of ~-0.5, meaning that a value of "Yes" negatively influences the response.

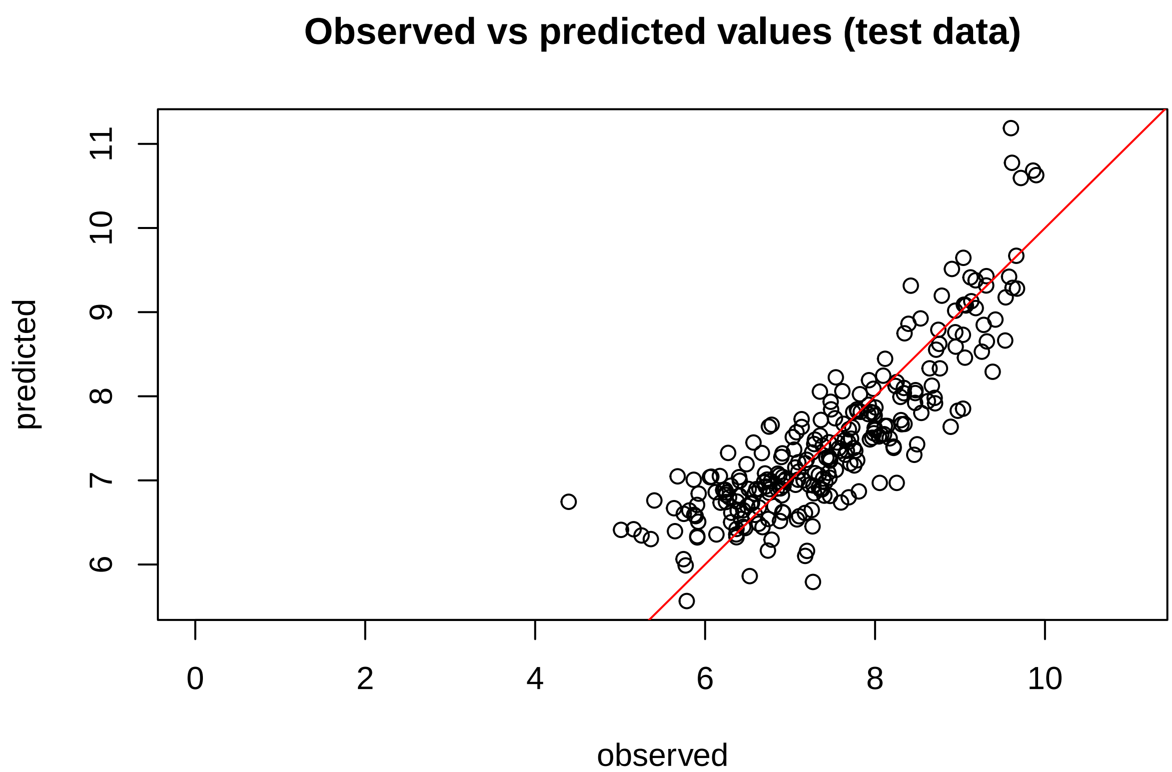
Predicting values

```
plot(train$Apps, full.lm %>% predict(train) , xlim = c(0, 11),
     main="Observed vs predicted values (training data)", xlab="observed", ylab="predicted",
     # ylim=c(0, 11)
     )
abline(coef = c(0,1), col="red")
```




```
plot(test$Apps, full.lm %>% predict(test) , xlim = c(0, 11),
     main="Observed vs predicted values (test data)", xlab="observed", ylab="predicted",
     # ylim=c(0, 11)
```

```
    )
abline(coef = c(0,1), col="red")
```



Visually, the variance of the predicted vs observed data points look similar in the plots of the training and the test data.

```
get_rmse <- function(y, yhat){
  sqrt(mean((y-yhat)^2))
}

paste(
  "RMSE of training set:",
  get_rmse(train$Apps, full.lm %>% predict(train)) %>% round(4),
  " ---- RMSE of test set:",
  get_rmse(test$Apps, full.lm %>% predict(test)) %>% round(4)
)
```

```
## [1] "RMSE of training set: 0.5589  ---- RMSE of test set: 0.574"
```

The RMSE of the training set being lower than that of the test set alsp checks out.

3. Slim model

Manually removing all insignificant variables from the full model, we are left with:
- An Intercept that is not zero

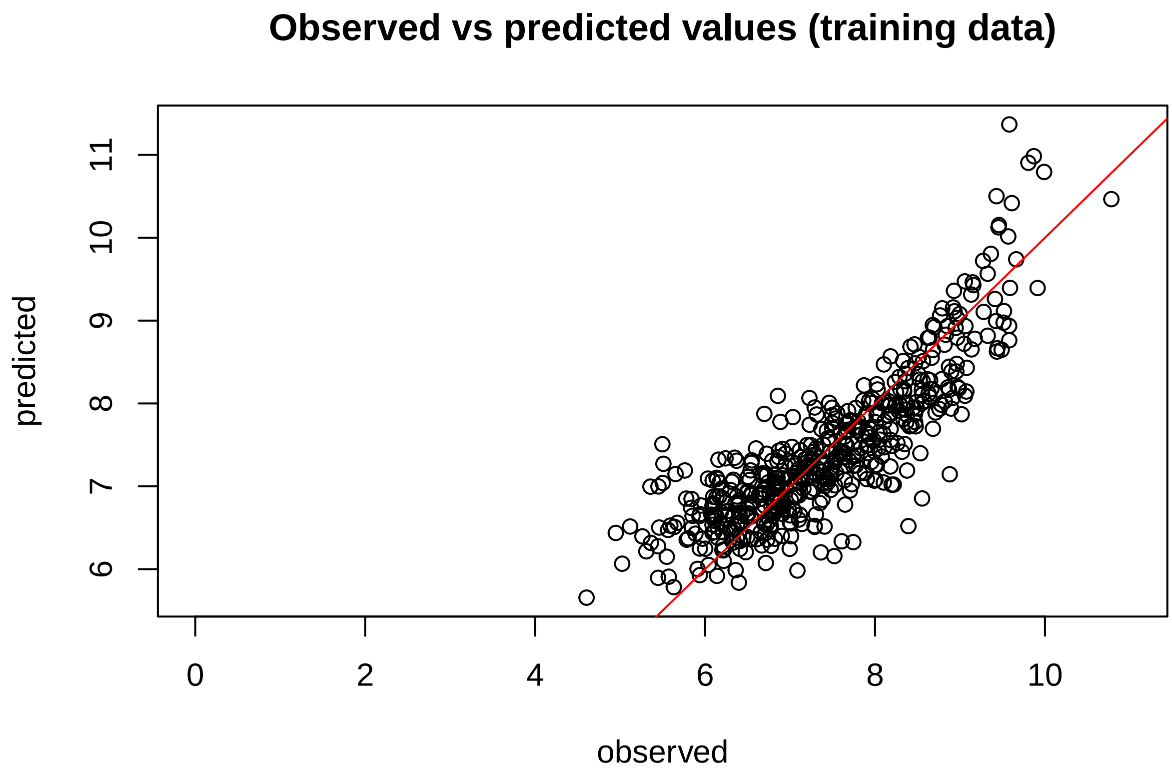
```
- Private
- F.Undergrad
- Outstate
- Room.Board
- Books
- PhD
- S.F.Ratio
- perc.alumni
- Expend
- Grad.Rate
```

```
slim.lm <- lm(Apps ~ Private + F.Undergrad + Outstate + Room.Board + Books + PhD + S.F.Ratio + perc.alu
slim.lm %>% summary()
```

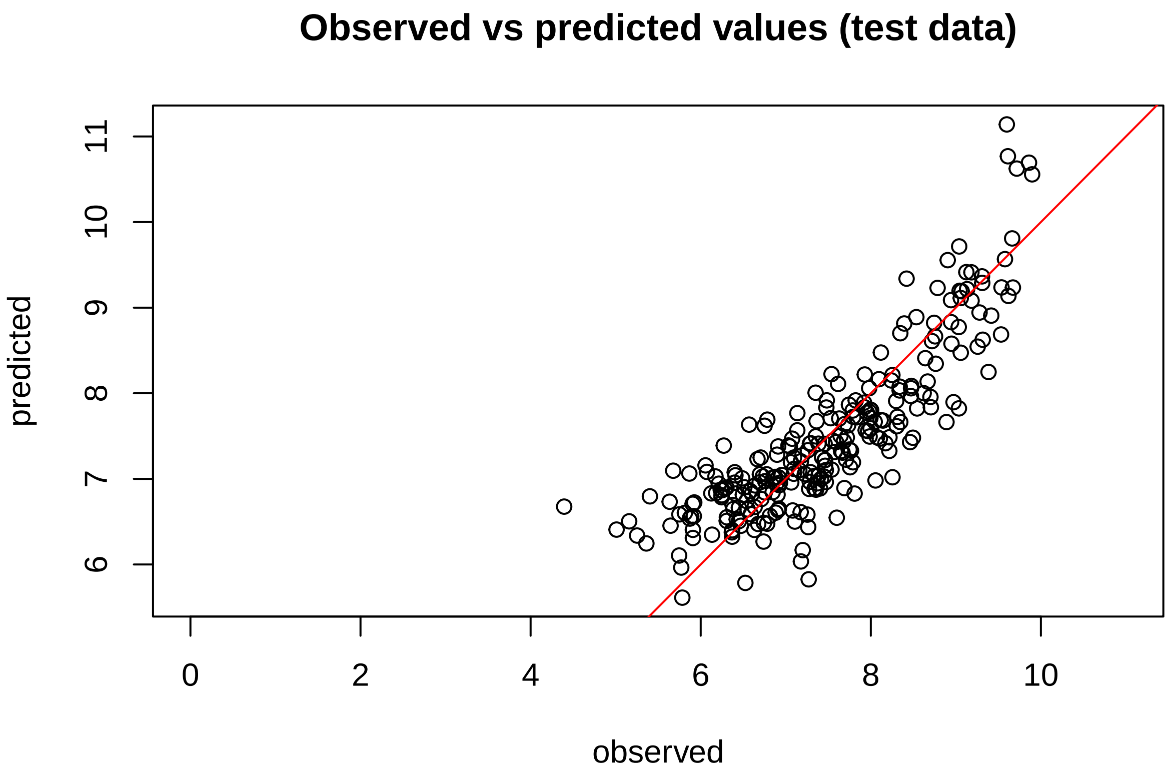
```
##
## Call:
## lm(formula = Apps ~ Private + F.Undergrad + Outstate + Room.Board +
##     Books + PhD + S.F.Ratio + perc.alumni + Expend + Grad.Rate,
##     data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.01115 -0.31224  0.01305  0.36900  1.87276
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.292e+00  2.344e-01  18.313  < 2e-16 ***
## PrivateYes  -5.736e-01  9.242e-02  -6.206 1.13e-09 ***
## F.Undergrad  1.182e-04  7.406e-06  15.965  < 2e-16 ***
## Outstate     4.291e-05  1.309e-05   3.279  0.00111 **
## Room.Board   5.605e-05  3.301e-05   1.698  0.09014 .
## Books        4.496e-04  1.635e-04   2.750  0.00618 **
## PhD          1.012e-02  2.138e-03   4.733 2.88e-06 ***
## S.F.Ratio    3.862e-02  8.554e-03   4.515 7.88e-06 ***
## perc.alumni -6.985e-03  2.749e-03  -2.541  0.01134 *
## Expend       3.161e-05  1.005e-05   3.144  0.00177 **
## Grad.Rate    1.152e-02  1.906e-03   6.041 2.97e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5688 on 507 degrees of freedom
## Multiple R-squared:  0.7216, Adjusted R-squared:  0.7161
## F-statistic: 131.4 on 10 and 507 DF,  p-value: < 2.2e-16
```

After pruning the variables that were not significant in the full model, all remaining variables' coefficients are significant in the pruned model.
Generally, this is not always the case, as highly correlated variables that are significant may not be significant anymore if you remove one.

```
plot(train$Apps, slim.lm %>% predict(train) , xlim = c(0, 11),
     main="Observed vs predicted values (training data)", xlab="observed", ylab="predicted",
     # ylim=c(0, 11)
     )
abline(coef = c(0,1), col="red")
```




```r
plot(test$Apps, slim.lm %>% predict(test) , xlim = c(0, 11),
     main="Observed vs predicted values (test data)", xlab="observed", ylab="predicted",
     # ylim=c(0, 11)
     )
abline(coef = c(0,1), col="red")
```



Visually, I don't see an obvious improvement of the slim model's performance to the full model.

```
paste(
  "RMSE of training set:",
  get_rmse(train$Apps, slim.lm %>% predict(train)) %>% round(4),
  " ---- RMSE of test set:",
  get_rmse(test$Apps, slim.lm %>% predict(test)) %>% round(4)
)
```

```
## [1] "RMSE of training set: 0.5627  ---- RMSE of test set: 0.5753"
```

On the test set, the RMSE has gotten marginally worse, which is to be expected when pruning predictors. On the other hand, also unsurprisingly, the RMSE on the test set has improved, though also by a bizmal amount.

```
anova(full.lm, slim.lm)
```

```
## 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    502 161.79
## 2    507 164.01 -5    -2.217 1.3757 0.2319
```

The p-value is >6%, so I would conservatively rule this slim model not to be significantly different from the full model.

4. Stepwise variable selection

```
step.fw.lm <- step(full.lm, direction = "forward")
```

```
## Start:  AIC=-570.78
## Apps ~ Private + Top10perc + Top25perc + F.Undergrad + P.Undergrad +
##     Outstate + Room.Board + Books + Personal + PhD + Terminal +
##     S.F.Ratio + perc.alumni + Expend + Grad.Rate
```

```
step.bw.lm <- step(full.lm, direction = "backward")
```

```
## Start:  AIC=-570.78
## 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
## - Top10perc     1     0.000 161.79 -572.78
## - Personal      1     0.006 161.80 -572.76
## - P.Undergrad   1     0.073 161.86 -572.55
## - Terminal      1     0.104 161.89 -572.45
## <none>                      161.79 -570.78
## - Top25perc     1     0.744 162.53 -570.40
## - Room.Board    1     0.914 162.71 -569.86
## - Books         1     1.516 163.31 -567.95
## - Expend        1     1.797 163.59 -567.06
## - PhD           1     1.799 163.59 -567.05
## - perc.alumni   1     2.844 164.63 -563.76
## - Outstate      1     3.138 164.93 -562.83
## - S.F.Ratio     1     6.598 168.39 -552.08
## - Grad.Rate     1     9.503 171.29 -543.22
## - Private       1    11.984 173.77 -535.77
## - F.Undergrad   1    52.289 214.08 -427.72
##
## Step:  AIC=-572.78
## Apps ~ 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
## - Personal      1     0.006 161.80 -574.76
## - P.Undergrad   1     0.075 161.87 -574.54
## - Terminal      1     0.108 161.90 -574.44
## <none>                      161.79 -572.78
## - Room.Board    1     0.920 162.71 -571.84
## - Books         1     1.518 163.31 -569.94
## - PhD           1     1.817 163.61 -569.00
```

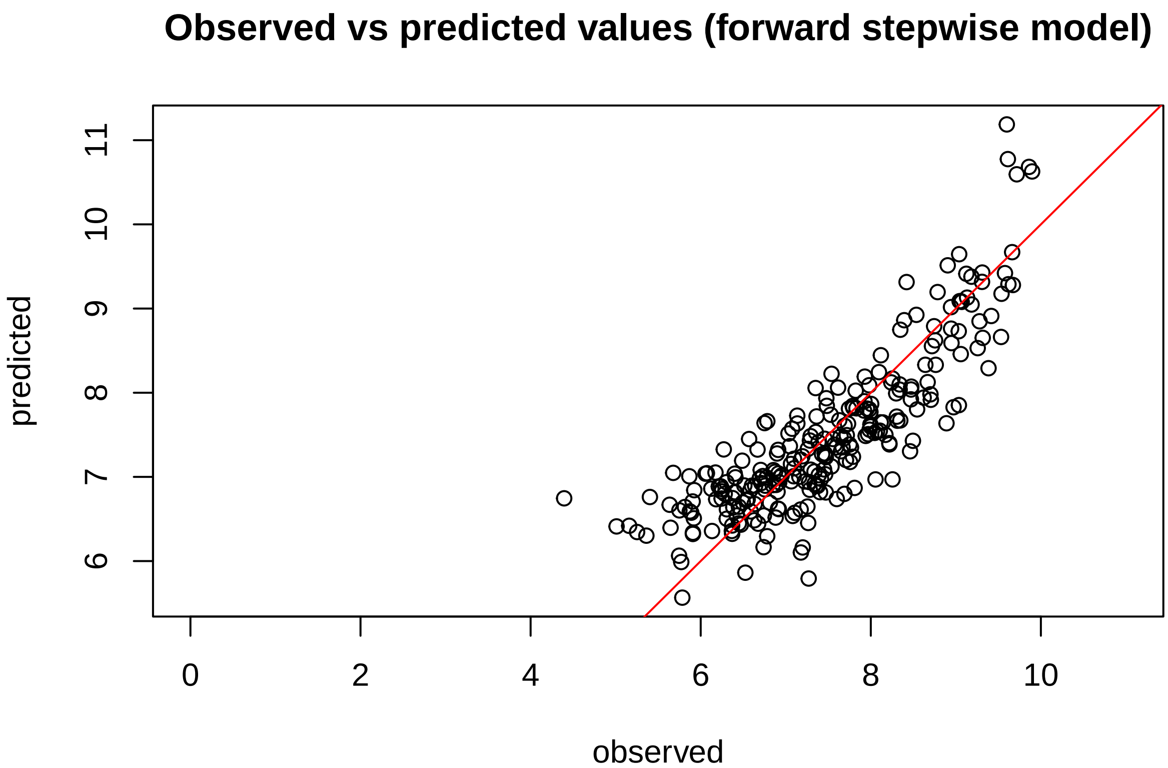
```
## - Top25perc    1     1.956 163.75 -568.56
## - Expend       1     2.223 164.01 -567.71
## - perc.alumni  1     2.849 164.64 -565.74
## - Outstate     1     3.148 164.94 -564.80
## - S.F.Ratio    1     6.619 168.41 -554.01
## - Grad.Rate    1     9.599 171.39 -544.92
## - Private      1    11.983 173.77 -537.77
## - F.Undergrad  1    52.305 214.10 -429.68
##
## Step:  AIC=-574.76
## Apps ~ Private + Top25perc + F.Undergrad + P.Undergrad + Outstate +
##     Room.Board + Books + PhD + Terminal + S.F.Ratio + perc.alumni +
##     Expend + Grad.Rate
##
##               Df Sum of Sq    RSS     AIC
## - P.Undergrad  1     0.079 161.88 -576.51
## - Terminal     1     0.109 161.91 -576.41
## <none>                     161.80 -574.76
## - Room.Board   1     0.914 162.71 -573.84
## - Books        1     1.650 163.45 -571.50
## - PhD          1     1.813 163.61 -570.99
## - Top25perc    1     1.952 163.75 -570.55
## - Expend       1     2.240 164.04 -569.64
## - perc.alumni  1     2.901 164.70 -567.56
## - Outstate     1     3.147 164.94 -566.78
## - S.F.Ratio    1     6.652 168.45 -555.89
## - Grad.Rate    1     9.624 171.42 -546.83
## - Private      1    12.019 173.82 -539.64
## - F.Undergrad  1    52.672 214.47 -430.78
##
## Step:  AIC=-576.51
## Apps ~ Private + Top25perc + F.Undergrad + Outstate + Room.Board +
##     Books + PhD + Terminal + S.F.Ratio + perc.alumni + Expend +
##     Grad.Rate
##
##               Df Sum of Sq    RSS     AIC
## - Terminal     1     0.116 161.99 -578.14
## <none>                     161.88 -576.51
## - Room.Board   1     1.009 162.88 -575.29
## - Books        1     1.668 163.54 -573.20
## - PhD          1     1.814 163.69 -572.74
## - Top25perc    1     1.927 163.80 -572.38
## - Expend       1     2.178 164.05 -571.59
## - perc.alumni  1     2.998 164.87 -569.00
## - Outstate     1     3.147 165.02 -568.54
## - S.F.Ratio    1     6.622 168.50 -557.74
## - Grad.Rate    1     9.576 171.45 -548.74
## - Private      1    12.200 174.08 -540.87
## - F.Undergrad  1    74.028 235.90 -383.43
##
## Step:  AIC=-578.14
## 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>                      161.99 -578.14
## - Room.Board   1     1.128 163.12 -576.54
## - Books        1     1.774 163.77 -574.50
## - Top25perc    1     2.016 164.01 -573.73
## - Expend       1     2.173 164.16 -573.24
## - perc.alumni  1     2.906 164.90 -570.93
## - Outstate     1     3.224 165.22 -569.93
## - PhD          1     5.218 167.21 -563.72
## - S.F.Ratio    1     6.596 168.59 -559.47
## - Grad.Rate    1     9.483 171.47 -550.67
## - Private      1    12.450 174.44 -541.78
## - F.Undergrad  1    75.018 237.01 -383.01
```

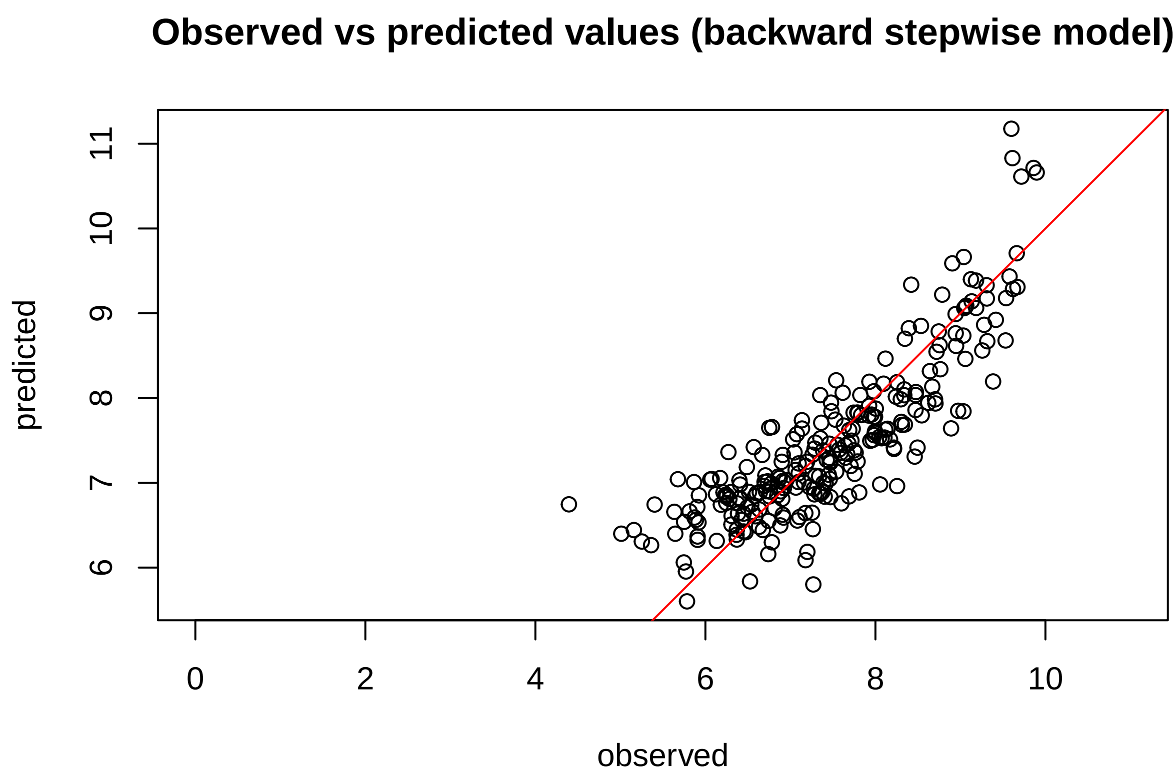
```r
data.frame(
  model = c("Full", "Slim", "Step.Forward", "Step.Backward"),
  rmse_train = c(
    get_rmse(train$Apps, predict(full.lm, train)),
    get_rmse(train$Apps, predict(slim.lm, train)),
    get_rmse(train$Apps, predict(step.fw.lm, train)),
    get_rmse(train$Apps, predict(step.bw.lm, train))
  ),
  rmse_test = c(
    get_rmse(test$Apps, predict(full.lm, test)),
    get_rmse(test$Apps, predict(slim.lm, test)),
    get_rmse(test$Apps, predict(step.fw.lm, test)),
    get_rmse(test$Apps, predict(step.bw.lm, test))
  )
) %>% kable()
```

model	rmse_train	rmse_test
Full	0.5588714	0.5739589
Slim	0.5626873	0.5753301
Step.Forward	0.5588714	0.5739589
Step.Backward	0.5592182	0.5735426

```r
plot(test$Apps, step.fw.lm %>% predict(test) , xlim = c(0, 11),
     main="Observed vs predicted values (forward stepwise model)", xlab="observed", ylab="predicted",
     # ylim=c(0, 11)
     )
abline(coef = c(0,1), col="red")
```




```r
plot(test$Apps, step.bw.lm %>% predict(test) , xlim = c(0, 11),
     main="Observed vs predicted values (backward stepwise model)", xlab="observed", ylab="predicted",
     # ylim=c(0, 11)
     )
abline(coef = c(0,1), col="red")
```




```r
anova(full.lm, step.fw.lm)
```

```
## 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 + Top10perc + Top25perc + F.Undergrad + P.Undergrad +
##     Outstate + Room.Board + Books + Personal + PhD + Terminal +
##     S.F.Ratio + perc.alumni + Expend + Grad.Rate
##   Res.Df    RSS Df Sum of Sq F Pr(>F)
## 1    502 161.79
## 2    502 161.79  0         0
```

```r
anova(full.lm, step.bw.lm)
```

```
## 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 + Top25perc + 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    502 161.79
## 2    506 161.99 -4  -0.20086 0.1558 0.9604
```

From looking at the resulting RMSE scores, the observed vs predicted plots and the results of the ANOVA tests, the stepwise models did not make a mentionable difference.