STAT495 (Advanced Data Analysis): takehome problem

Tam Tran The

November 30, 2016

Description of the College Scorecard Data

The College Scorecard Data, compiled by the College Board, covers a wide range of topics associated with academic institutions in the US, including types of academics offered, admission statistics, demographic of the student body, cost, financial aid, repayment, college completion, and student earnings. The data set proves to be helpful to both students who are choosing colleges and policy makers who are trying to improve college quality.

```
load("takehome.Rda")
summary(train)
```

```
##     INSTNM            AVGFACSAL        UG25ABV           COSTT4_A
##  Length:1200        Min.   : 1476   Min.   :0.00000   Min.   : 7715
##  Class :character   1st Qu.: 5960   1st Qu.:0.06695   1st Qu.:21539
##  Mode  :character   Median : 7131   Median :0.15095   Median :31124
##                     Mean   : 7354   Mean   :0.20429   Mean   :32688
##                     3rd Qu.: 8370   3rd Qu.:0.27625   3rd Qu.:41953
##                     Max.   :17861   Max.   :0.87310   Max.   :64233
##
##          REGION        PFTFAC       GRAD_DEBT_MDN      RET_FT4
##  Southeast  :317   Min.   :0.0249   Min.   : 2000   Min.   :0.0000
##  Mid East   :242   1st Qu.:0.4849   1st Qu.:21148   1st Qu.:0.6736
##  Great Lakes:185   Median :0.6982   Median :24696   Median :0.7552
##  Far West   :126   Mean   :0.6782   Mean   :23746   Mean   :0.7445
##  Plains     :123   3rd Qu.:0.9222   3rd Qu.:27000   3rd Qu.:0.8322
##  New England: 95   Max.   :1.0000   Max.   :44500   Max.   :1.0000
##  (Other)    :112
##     ADM_RATE          FEMALE
##  Min.   :0.0000   Min.   :0.09899
##  1st Qu.:0.5411   1st Qu.:0.51891
##  Median :0.6728   Median :0.57823
##  Mean   :0.6507   Mean   :0.58099
##  3rd Qu.:0.7823   3rd Qu.:0.64152
##  Max.   :1.0000   Max.   :0.97781
##
```

```
colnames(train) <- c("schoolname", "avgfacsal", "age25", "annualcost",
                     "region", "fulltimefac", "debt", "retentionrate",
                     "admissionrate", "female")
colnames(test) <- c("schoolname", "avgfacsal", "age25", "annualcost",
                    "region", "fulltimefac", "debt", "retentionrate",
                    "admissionrate", "female")
```

Variables of interest

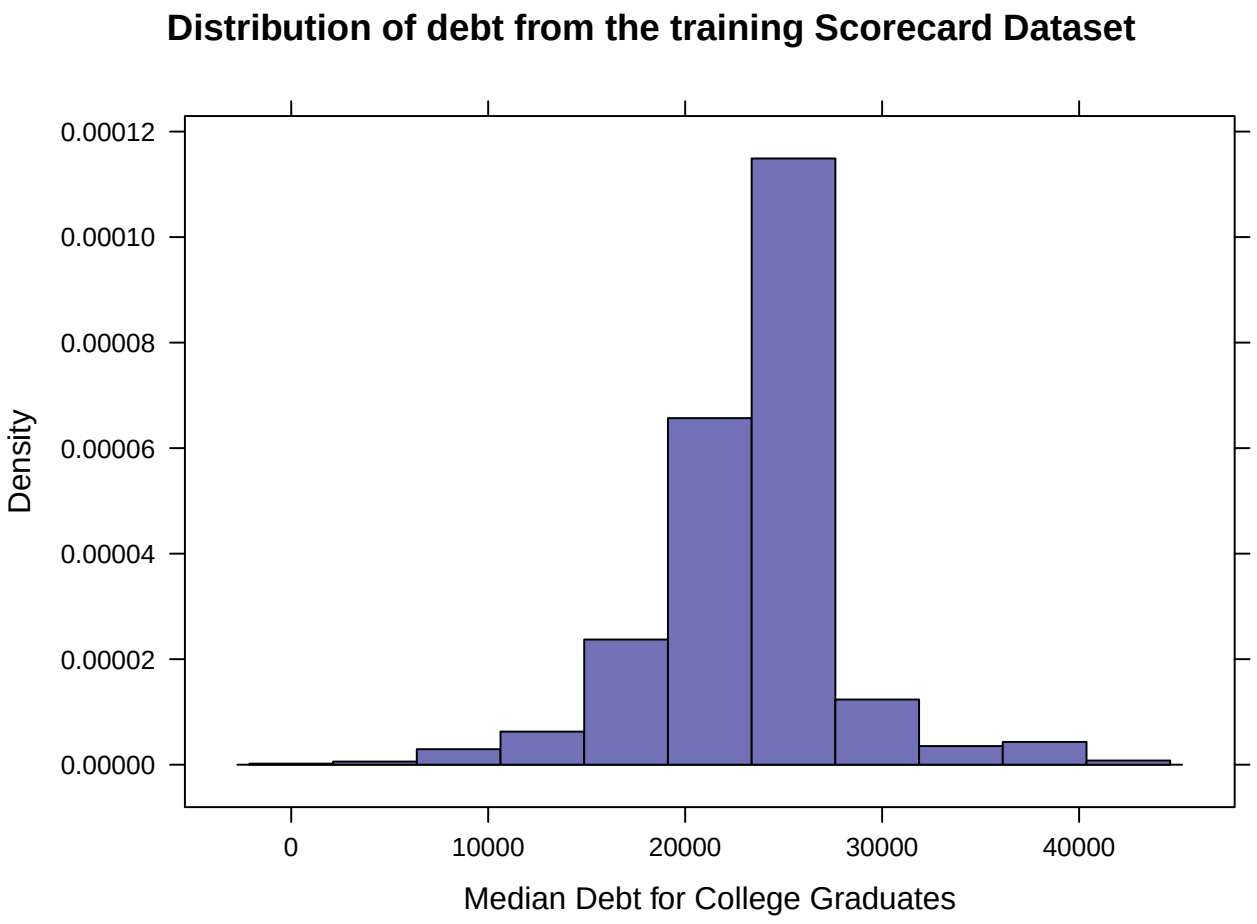
Response variable

Definition

`debt` (in $US) is defined as the median debt for students who have completed college.

Distribution

```
histogram(~debt, data=train,
          xlab = "Median Debt for College Graduates",
          ylab = "Density",
          main= "Distribution of debt from the training Scorecard Dataset")
```



`debt` has a unimodal distribution.

Predictors

Definitions

- `avgfacsal`: average faculty salary per month in $US
- `age25`: the proportion of undergraduates who are aged 25 or older
- `annualcost`: the average annual total cost in $US
- `region`: location of school
- `fulltimefac`: the proportion of full-time faculty
- `retentionrate`: the retention rate of first-time, full-time students at four-year institutions
- `admissionrate`: the admission rate
- `female`: the proportion of female students

Distributions

```
sal <- favstats(~avgfacsal, data=train)[c("min", "median", "mean", "max", "n")]
age25 <- favstats(~age25, data=train)[c("min", "median", "mean", "max", "n")]
cost <- favstats(~annualcost, data=train)[c("min", "median", "mean", "max", "n")]
reg <- favstats(~region, data=train)[c("min", "median", "mean", "max", "n")]
fac <- favstats(~fulltimefac, data=train)[c("min", "median", "mean", "max", "n")]
ret <- favstats(~retentionrate, data=train)[c("min", "median", "mean", "max", "n")]
ad <- favstats(~admissionrate, data=train)[c("min", "median", "mean", "max", "n")]
fem <- favstats(~female, data=train)[c("min", "median", "mean", "max", "n")]
summ <- rbind(sal, age25, cost, reg, fac, ret, ad, fem)
rownames(summ) <- c("avgfacsal", "age25", "annualcost", "region",
                    "fulltimefac", "retentionrate", "admissionrate", "female")
options(xtable.comment = FALSE)
xtable(summ)
```

	min	median	mean	max	n
avgfacsal	1476.00	7131.00	7353.86	17861.00	1200
age25	0.00	0.15	0.20	0.87	1200
annualcost	7715.00	31124.50	32687.94	64233.00	1200
region	1.00	4.00	4.10	8.00	1200
fulltimefac	0.02	0.70	0.68	1.00	1200
retentionrate	0.00	0.76	0.74	1.00	1200
admissionrate	0.00	0.67	0.65	1.00	1200
female	0.10	0.58	0.58	0.98	1200

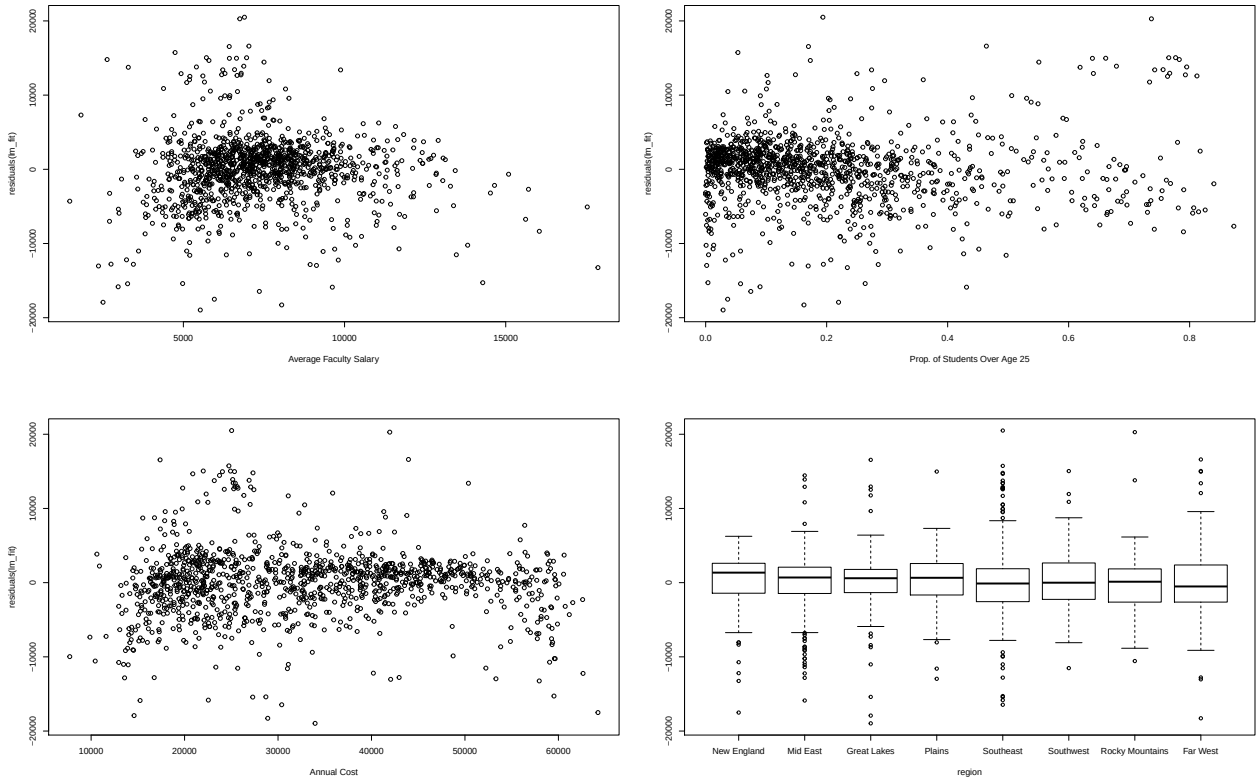
Analysis

The LINE assumptions

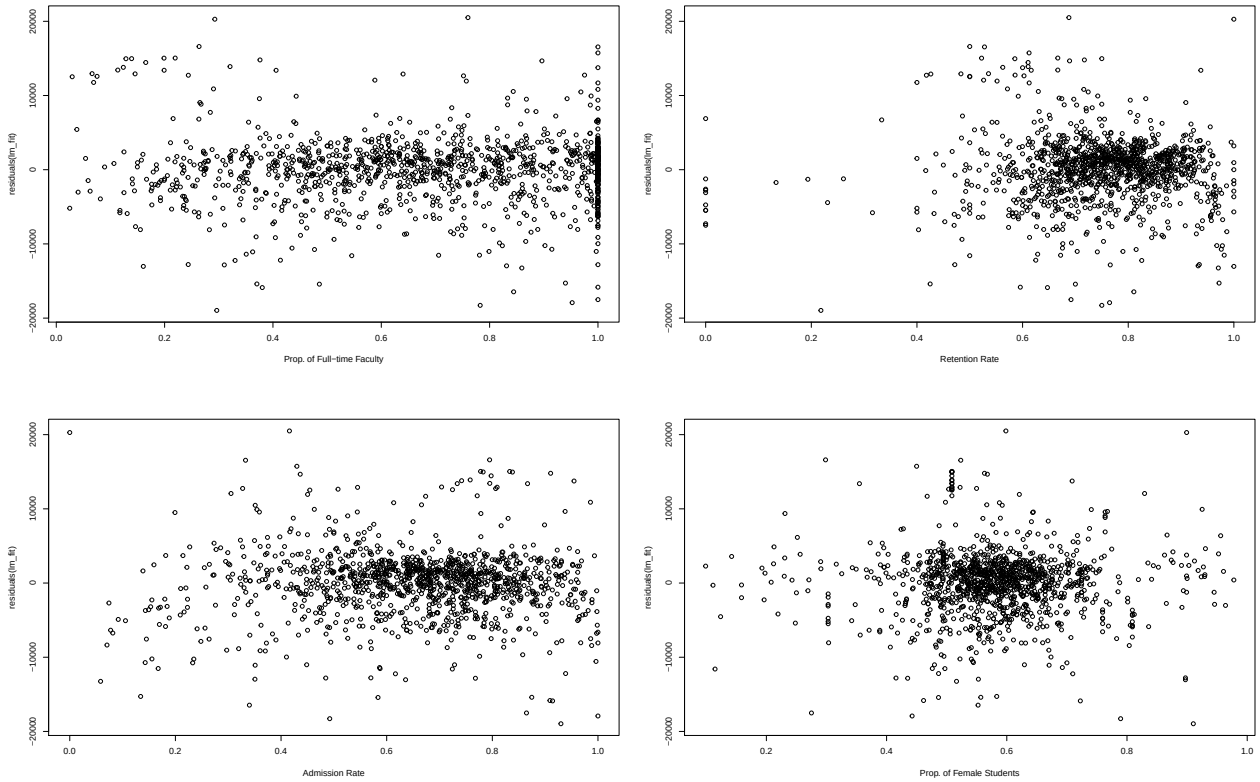
Linearity

```
lm_fit <- lm(debt ~ avgfacsal + age25 + annualcost + region +
               fulltimefac + retentionrate + admissionrate + female, data = train)
```

```
par(mfrow=c(2,2))
plot(train$avgfacsal, residuals(lm_fit),xlab="Average Faculty Salary")
plot(train$age25, residuals(lm_fit),xlab="Prop. of Students Over Age 25")
plot(train$annualcost, residuals(lm_fit),xlab="Annual Cost")
plot(train$region, residuals(lm_fit),xlab="region")
```



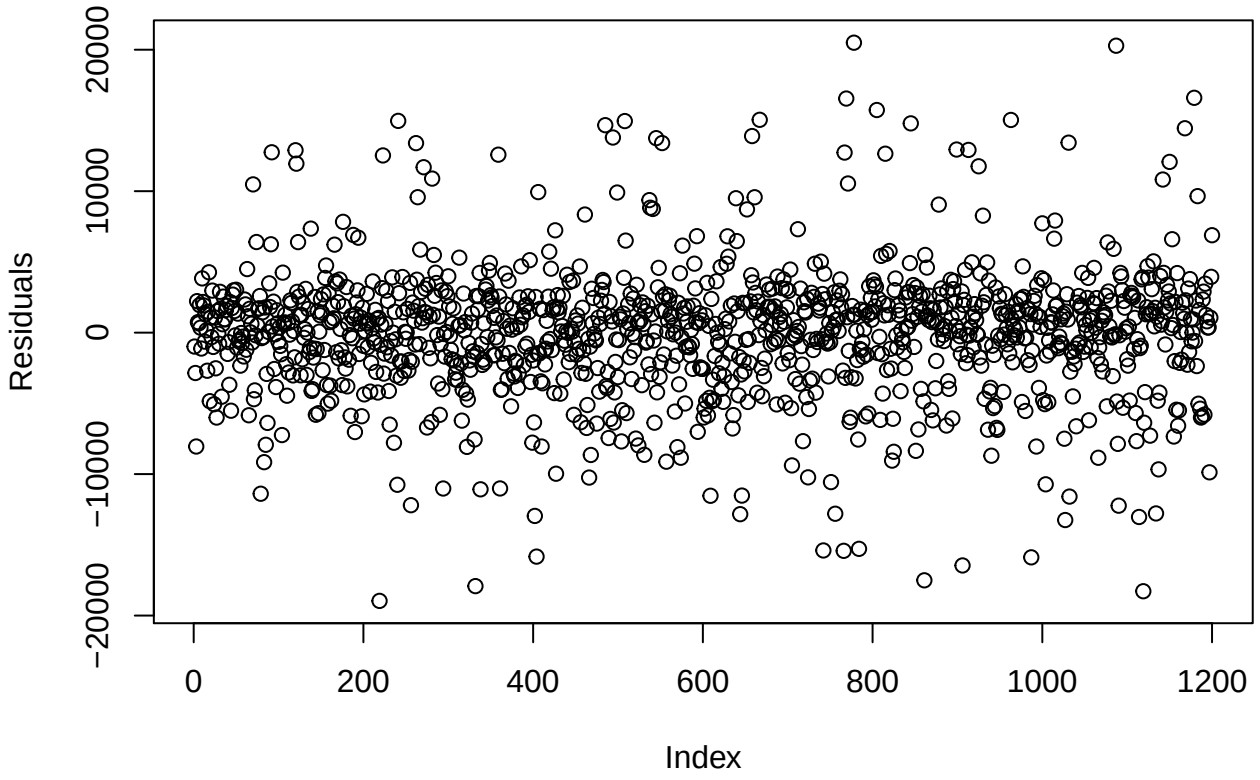

```
plot(train$fulltimefac, residuals(lm_fit),xlab="Prop. of Full-time Faculty")
plot(train$retentionrate, residuals(lm_fit),xlab="Retention Rate")
plot(train$admissionrate, residuals(lm_fit),xlab="Admission Rate")
plot(train$female, residuals(lm_fit),xlab="Prop. of Female Students")
```



The scatterplot between residuals and predictors checks for the linear relation. Since almost all of these plots show a pattern, the linear relationship assumption may not be met.

Independence

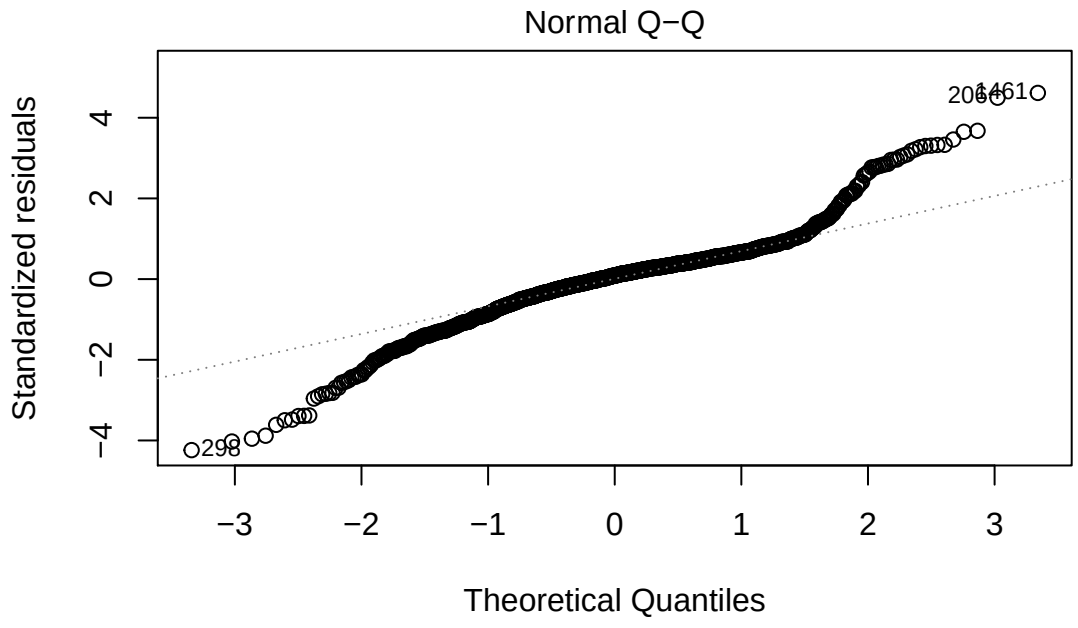
```
plot(residuals(lm_fit), ylab="Residuals")
```



There is no structure in this plot, which suggests that data is random and the residuals are independent.

Normality

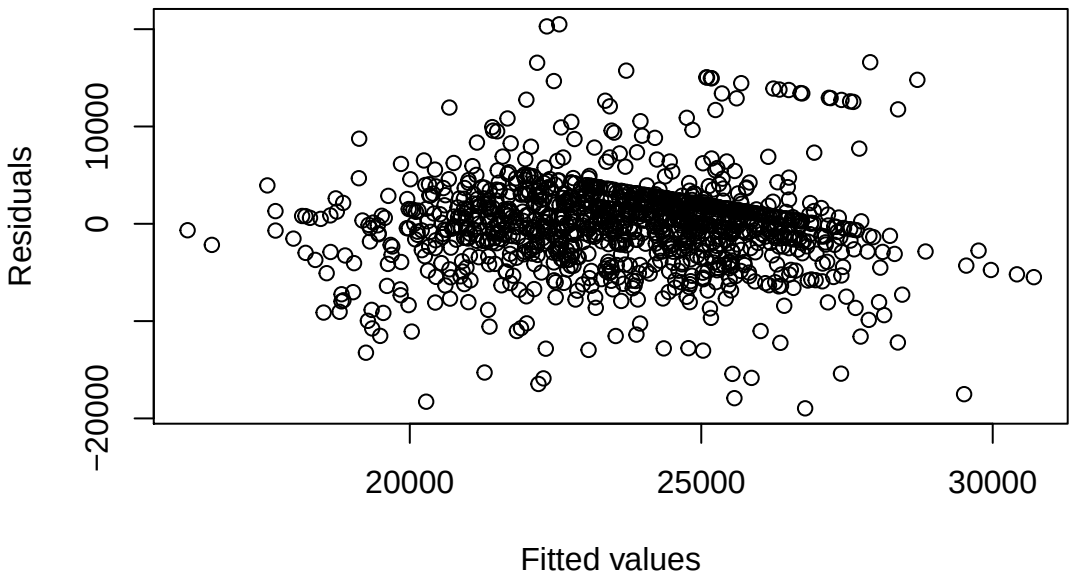
```
plot(lm_fit, which = 2, sub = "")
```



The qqplot checks for the normal distribution of the residuals. Since the residual points on the left and right ends do not fall on the theoretical line, normality is not guaranteed.

Equal variance

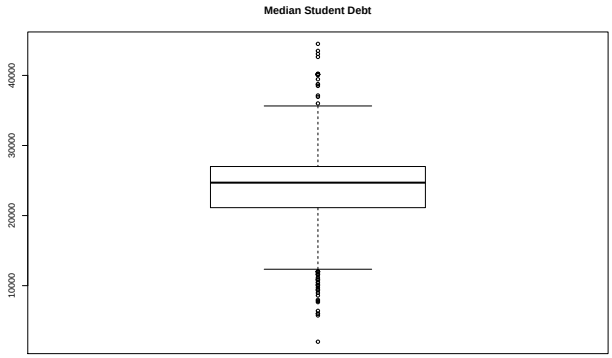
```
plot(fitted(lm_fit), residuals(lm_fit), xlab="Fitted values", ylab="Residuals")
```

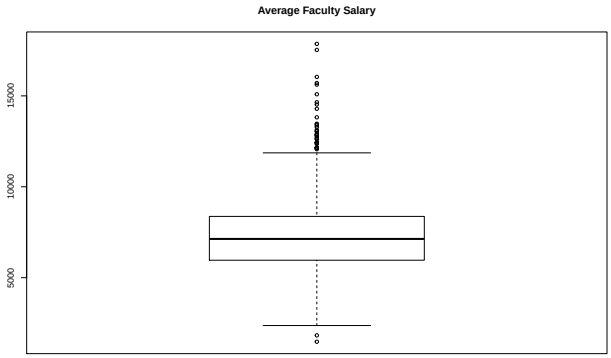


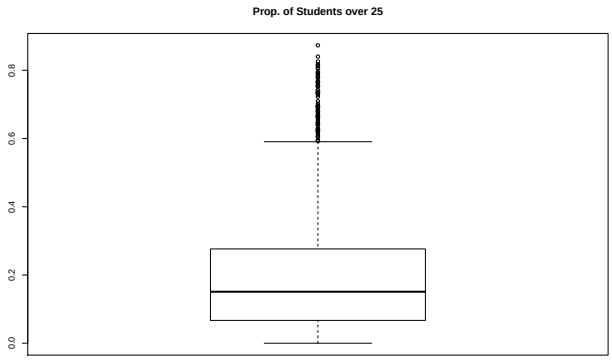
The scatterplot between residuals and fitted values checks for the homogeneity of the variance. Since there is no discernable pattern in the plot, the equal variance assumption is met.

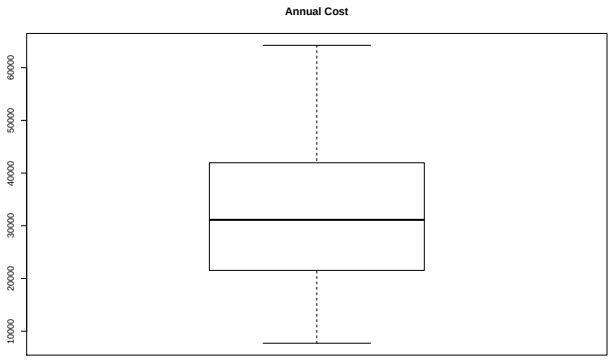
Outliers

```
par(mfrow=c(2,2))
boxplot(train$debt, main="Median Student Debt")
boxplot(train$avgfacsal, main="Average Faculty Salary")
boxplot(train$age25, main="Prop. of Students over 25")
boxplot(train$annualcost, main="Annual Cost")
```

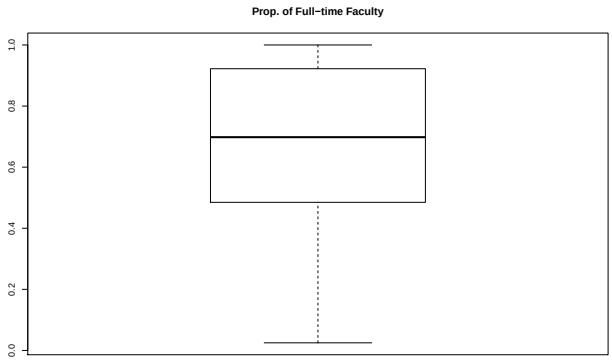


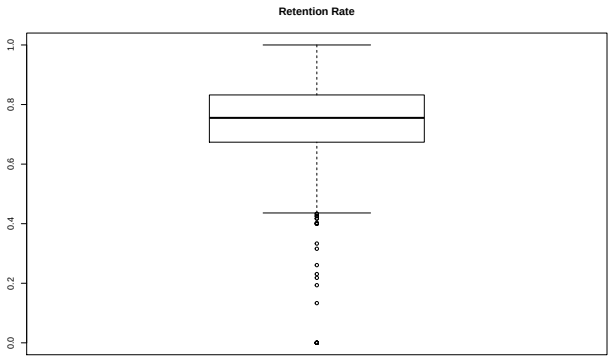


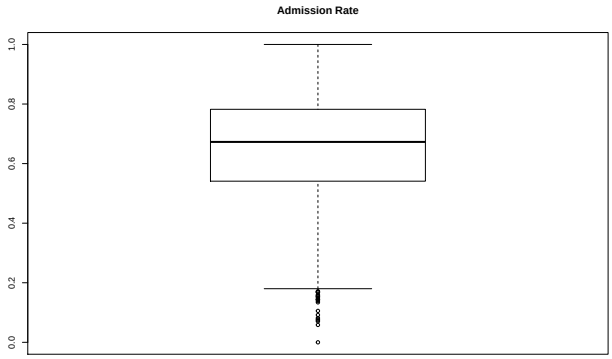


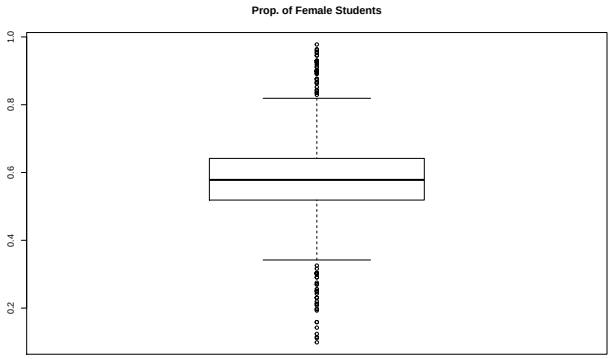



```r
boxplot(train$fulltimefac, main="Prop. of Full-time Faculty")
boxplot(train$retentionrate, main="Retention Rate")
boxplot(train$admissionrate, main="Admission Rate")
boxplot(train$female, main="Prop. of Female Students")
```









Linear regression model

Model fitting

```
options(xtable.comment = FALSE)
xtable(summary(lm_fit))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	25110.6910	1529.5054	16.42	0.0000
avgfacsal	-0.4963	0.0837	-5.93	0.0000
age25	2820.6896	937.9231	3.01	0.0027
annualcost	0.1398	0.0122	11.50	0.0000
regionMid East	917.5288	559.9358	1.64	0.1016
regionGreat Lakes	1131.4211	588.2312	1.92	0.0547
regionPlains	-125.7528	643.7070	-0.20	0.8451
regionSoutheast	811.6463	564.5486	1.44	0.1508
regionSouthwest	-1039.9600	706.5506	-1.47	0.1413
regionRocky Mountains	-626.9499	999.3604	-0.63	0.5305
regionFar West	-661.7374	627.8009	-1.05	0.2921
fulltimefac	-1265.2534	567.8485	-2.23	0.0261
retentionrate	-3030.5822	1265.7228	-2.39	0.0168
admissionrate	3069.4720	766.2030	4.01	0.0001
female	-3695.9613	1169.9962	-3.16	0.0016

Interpretation

According to this linear regression model, `avgfacsal`, `age25`, `annualcost`, `fulltimefac`, `retentionrate`, `admissionrate`, and `female` are statistically significant at an α level of 0.05. Specifically, `age25`, `annualcost` and `admissionrate` are positive predictors whereas `avgfacsal`, `fulltimefac`, `retentionrate`, and `female` are negative ones. This model only explains 18.73% of the variation in the response variable `debt`.

Holding other factors fixed:

- An additional dollar in average faculty salary per month results in a decrease of 50 cents for median student debt.
- A unit increase in the proportion of students over 25 predicts an increase of 2820 dollars in median student debt.
- An additional dollar in annual cost for college is expected to increase the median student debt by 14 cents.
- A unit increase in the proportion of full-time faculty results in a decrease of 1265 dollars for median student debt.
- A unit increase in the student retention rate results in a decrease of 3030 dollars for median student debt.
- A unit increase in the admission rate yields an increase of 3069 dollars for student debt.
- A unit increase in the proportion of female students results in a decrease of 3695 dollars for student debt.

Generally, schools with higher average faculty salary, proportion of full-time faculty, retention rate, and proportion of female students might be better for students who want to graduate with less debt.

Training and test error

```
train_error1 <- sqrt(sum((predict(lm_fit) - train$debt)^2)/nrow(train))
train_error1
```

```
## [1] 4540.957
```

```
test_error1 <- sqrt(sum((predict(lm_fit, newdata=test) - test$debt)^2)/
                      nrow(test))
test_error1
```

```
## [1] 4443.278
```

The test error for this model is smaller than the training error, which is unexpected.

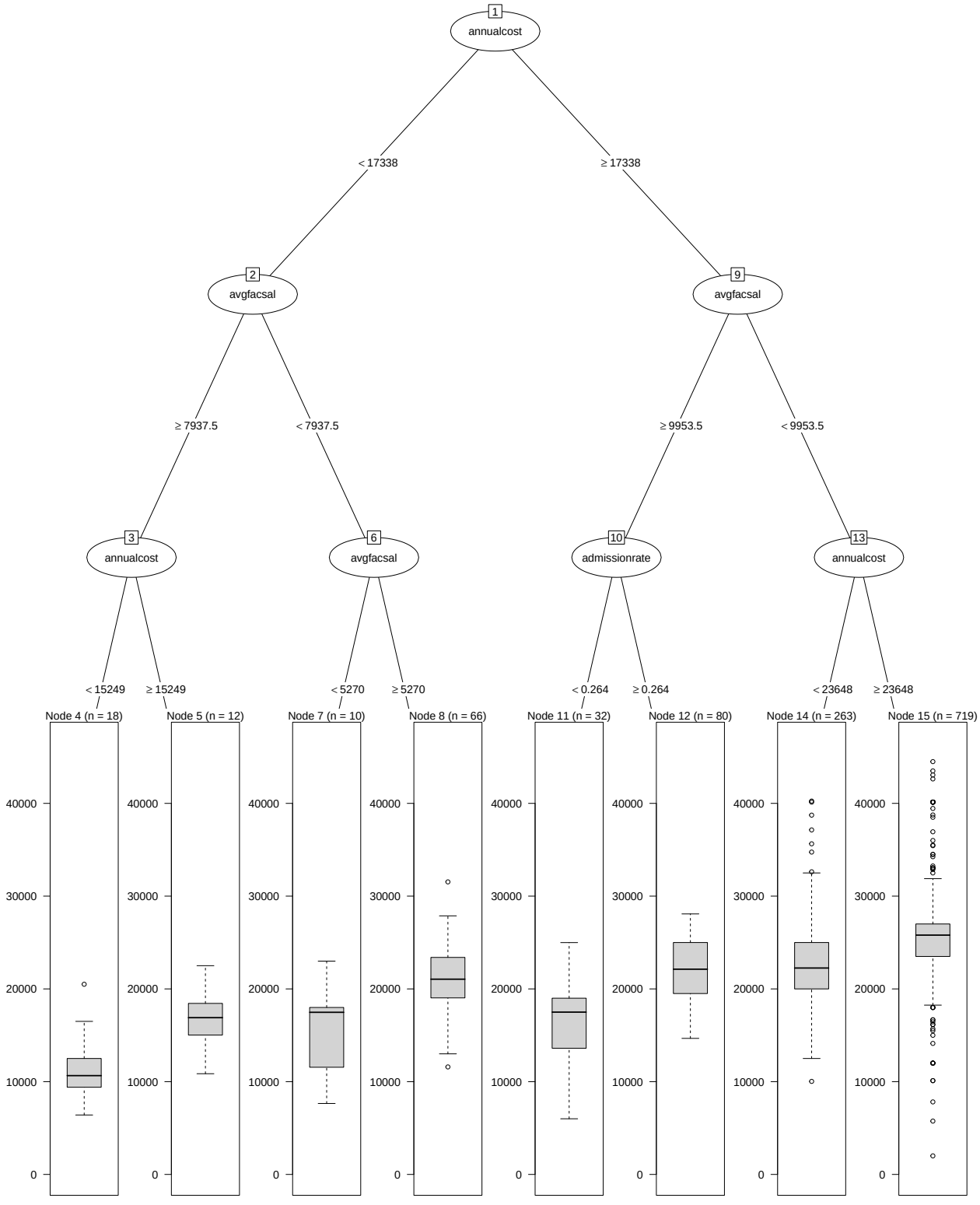
Decision tree

Model fitting

```
tree_fit <-rpart(debt ~ avgfacsal + age25 + annualcost + region +
                fulltimefac + retentionrate + admissionrate + female, data = train,
                  control=rpart.control(cp=0.005, maxdepth=3))
tree_fit
```

```
## n= 1200
##
## node), split, n, deviance, yval
##       * denotes terminal node
##
##  1) root 1200 30446090000 23745.66
##    2) annualcost< 17338 106  2794748000 18436.22
##      4) avgfacsal>=7937.5 30   474375900 13540.92
##        8) annualcost< 15249 18   183947600 11488.17 *
##        9) annualcost>=15249 12   100808100 16620.04 *
##      5) avgfacsal< 7937.5 76  1317668000 20368.58
##       10) avgfacsal< 5270 10   189204200 15728.10 *
##       11) avgfacsal>=5270 66   880495800 21071.68 *
##    3) annualcost>=17338 1094 24373670000 24260.10
##      6) avgfacsal>=9953.5 112  2372428000 20498.57
##       12) admissionrate< 0.26405 32   564019500 16022.33 *
##       13) admissionrate>=0.26405 80   910762400 22289.07 *
##      7) avgfacsal< 9953.5 982 20235800000 24689.11
##       14) annualcost< 23648 263  5063456000 22685.94 *
##       15) annualcost>=23648 719 13730970000 25421.84 *
```

```
plot(as.party(tree_fit))
```


1
annualcost
< 17338
≥ 17338
2
avgfacsal
9
avgfacsal
≥ 7937.5
< 7937.5
≥ 9953.5
< 9953.5
3
annualcost
6
avgfacsal
10
admissionrate
13
annualcost
< 15249
≥ 15249
< 5270
≥ 5270
< 0.264
≥ 0.264
< 23648
≥ 23648
Node 4 (n = 18)
Node 5 (n = 12)
Node 7 (n = 10)
Node 8 (n = 66)
Node 11 (n = 32)
Node 12 (n = 80)
Node 14 (n = 263)
Node 15 (n = 719)
0
10000
20000
30000
40000

Interpretation

Here, the first question is whether `annualcost` is above or below $17,338.

If `annualcost` is below $17,338, the next question is whether `avgfacsal` is above or below $7937.5.

- For those whose `avgfacsal` is above $7937.5, a threshold of `annualcost` of $15,249 is used to determine the median student debt. The student debt is expected to be around $11,488 if `annualcost` is below the threshold, and around $16,620 otherwise.
- For those whose `avgfacsal` is below $7937.5, we use a tighter threshold of `avgfacsal`, $5,270, to predict student debt. The student debt is around $15,728 if `avgfacsal` is below the threshold, and around $21,071 otherwise.

If `annualcost` is above $17,338, we again check `avgfacsal`, but this time the threshold is $9953.5.

- For those whose `avgfacsal` is above $9953.5, we ask about `addmissionrate` with the threshold of 0.264. The student debt is expected to be around $16,022 if `admissionrate` is below the threshold, and around $22,289 otherwise.
- For those whose `avgfacsal` is below $9953.5, we ask about `annualcost` with the threshold of $23,648. The student debt is around $22,685 if `annualcost` is below the threshold, and around $25,421 otherwise.

Training and test error

```
train_error2 <- sqrt(sum((predict(tree_fit) - train$debt)^2)/nrow(train))
train_error2
```

```
## [1] 4244.964
```

```
test_error2 <- sqrt(sum((predict(tree_fit, newdata=test) - train$debt)^2)/
                      nrow(test))
test_error2
```

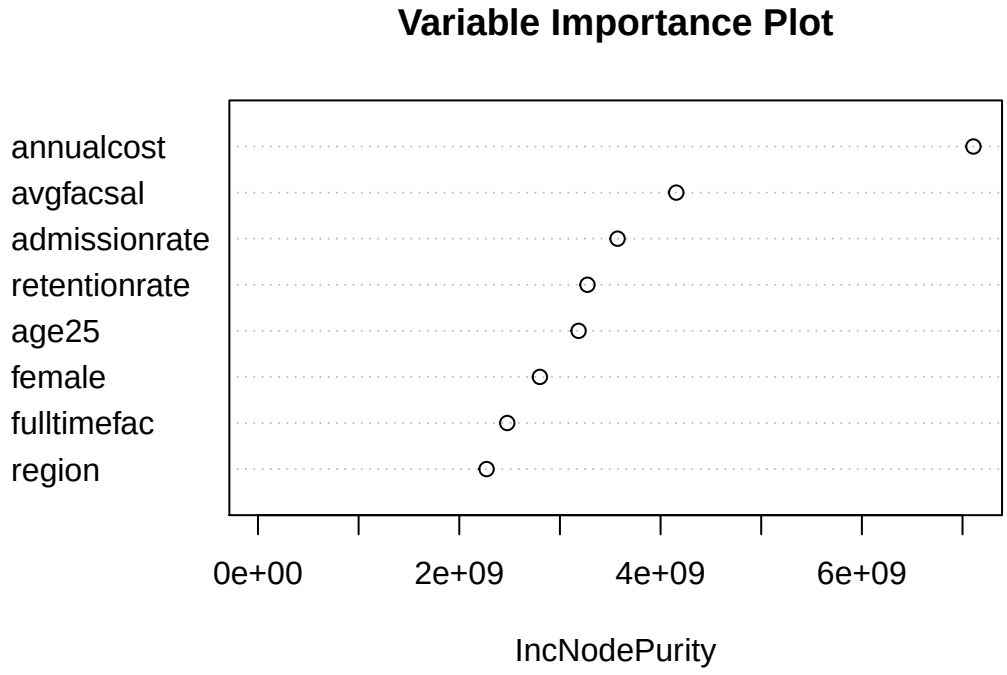
```
## [1] 10418.34
```

The test error is a lot greater than the train error, which suggests that this model is not a good fit and may suffer from overfitting.

Random forest

Model fitting

```
set.seed(1)
rf_fit <- randomForest(debt ~ avgfacsal + age25 + annualcost + region +
               fulltimefac + retentionrate + admissionrate + female,
               data = train, mtry=3, importance=TRUE)
varImpPlot(rf_fit, type=2, main="Variable Importance Plot")
```



Interpretation

From the variable importance plot, we see that annual cost is the most important predictor, followed by average faculty salary, admission rate. This is consistent with the decision tree's findings, since the decision tree also uses these variables to split the set of observations.

Training and test error

```
train_error3 <- sqrt(sum((predict(rf_fit) - train$debt)^2)/nrow(train))
train_error3
```

```
## [1] 3902.407
```

```
test_error3 <- sqrt(sum((predict(rf_fit, newdata=test) - train$debt)^2)/
                    nrow(test))
test_error3
```

```
## [1] 10650.64
```

The test error is a lot greater than the train error, which suggests that this model is not a good fit and may suffer from overfitting.