Draft Project

11/7/2017

Joining the two tables on the name of the college

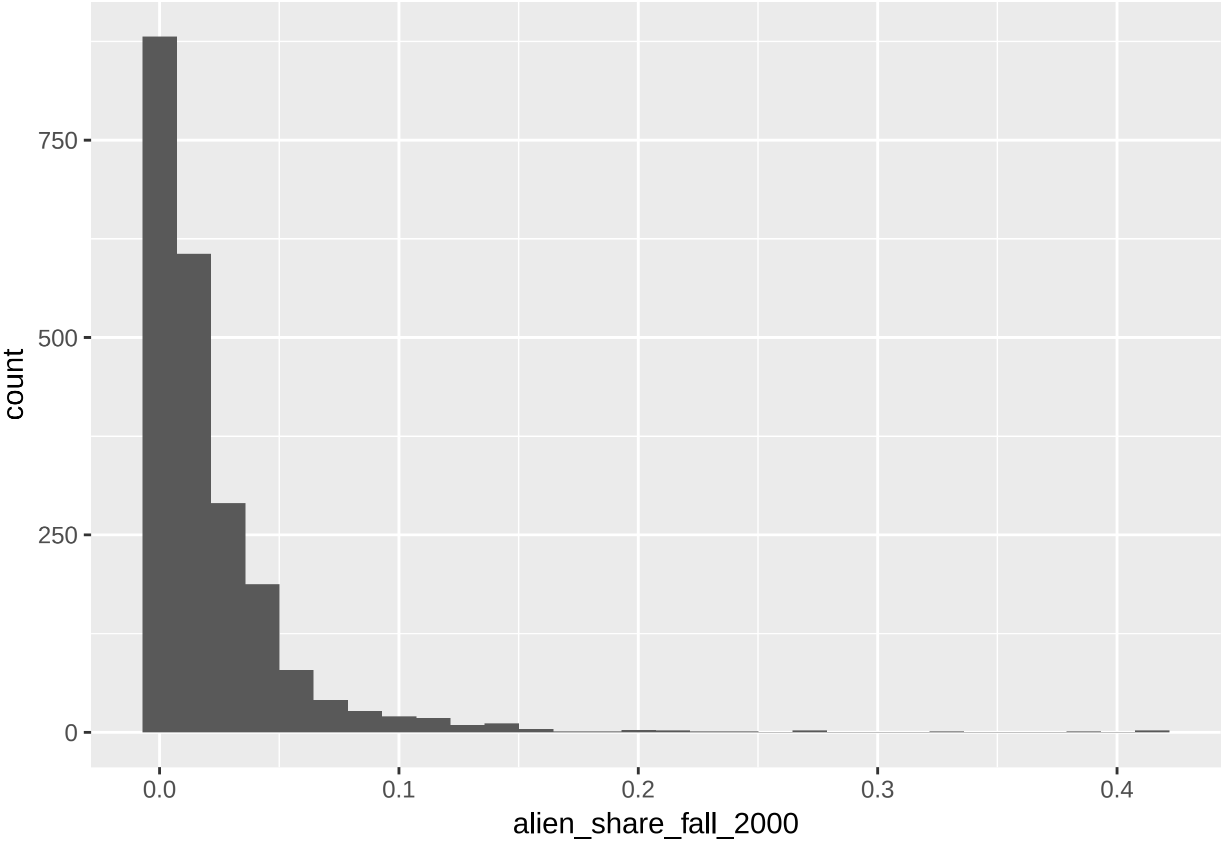
```
newData = inner_join(collegeChardata, collegedata, by="name")
```

Transforming the alien share explanatory variable

```
# before transformation
c <- ggplot(newData, aes(alien_share_fall_2000))
c+geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

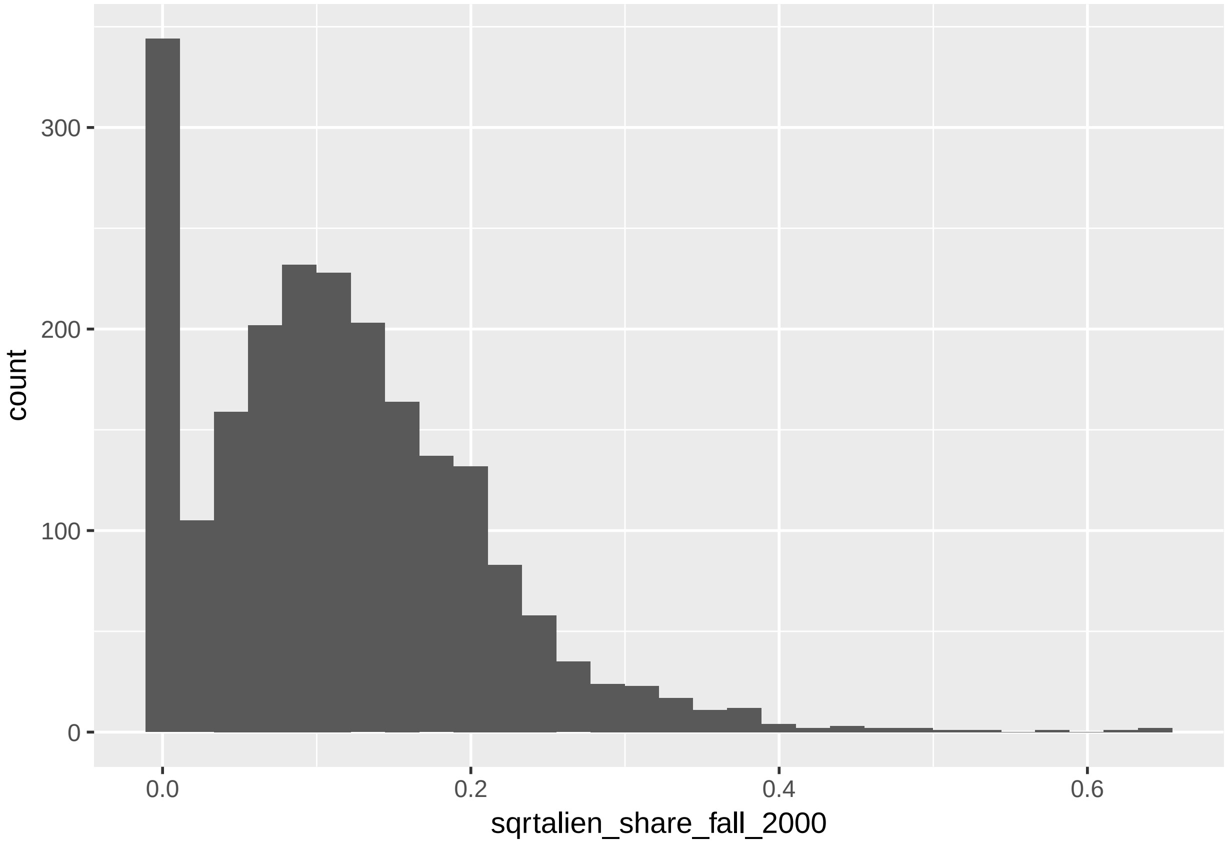
```
## Warning: Removed 11 rows containing non-finite values (stat_bin).
```



```
newData = newData %>%
  mutate(sqrtalien_share_fall_2000 = sqrt(alien_share_fall_2000))

# after transformation
d <- ggplot(newData, aes(sqrtalien_share_fall_2000))
d+geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 11 rows containing non-finite values (stat_bin).
```



Running stepwise regression

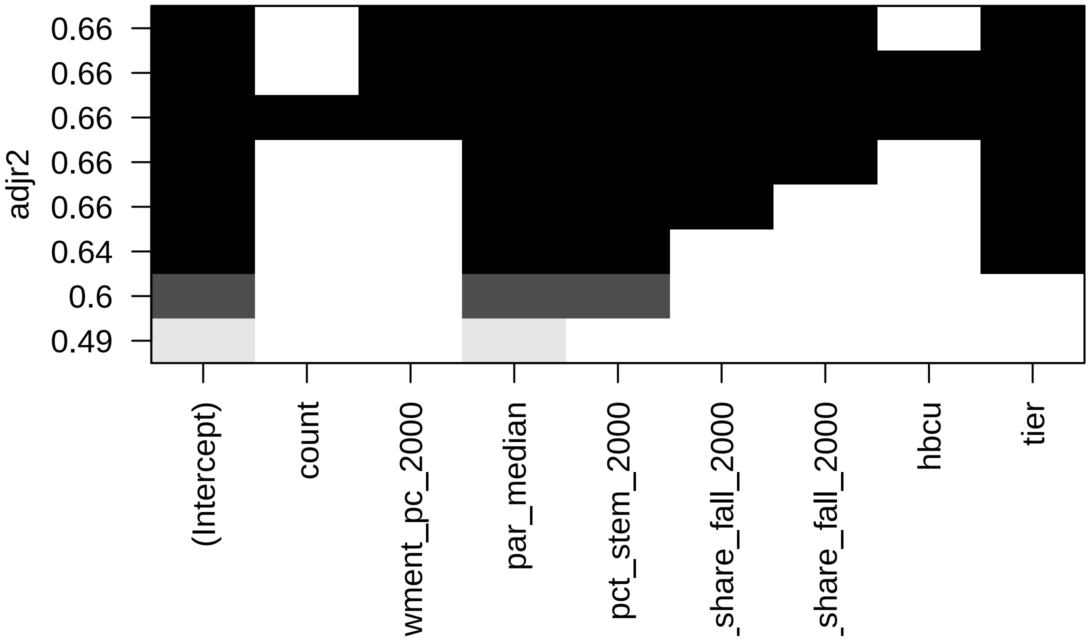
```
#just to remove missing values of variables
newData = newData%>%
  filter(!is.na(tier),!is.na(hbcu),!is.na(black_share_fall_2000),!is.na(sqrtalien_share_fall_2000),!is.n
         !is.na(par_median),!is.na(endowment_pc_2000),!is.na(count),!is.na(kq5_cond_parq1))

nullmodel <- lm(kq5_cond_parq1~1, data = newData)
fullmodel <- lm(kq5_cond_parq1~tier + hbcu + black_share_fall_2000 + sqrtalien_share_fall_2000 + pct_ste

# With 'scale=fullMSE', the AIC term can be interpreted as Cp
fullMSE <- (summary(fullmodel)$sigma)^2

step(nullmodel, scope = list(upper = fullmodel),
scale = fullMSE, direction = "both")
```

```
## Start:  AIC=1392.17
## kq5_cond_parq1 ~ 1
##
##                              Df Sum of Sq   RSS      Cp
## + par_median                  1     41592 42327  349.32
## + tier                        1     34753 49166  521.12
## + pct_stem_2000              1     23664 60256  799.70
## + sqrtalien_share_fall_2000  1     16822 67098  971.59
## + count                      1      6131 77789 1240.15
## + black_share_fall_2000      1      3657 80263 1302.32
## + hbcu                       1      1055 82865 1367.68
## + endowment_pc_2000          1       412 83508 1383.83
## <none>                                   83920 1392.17
##
## Step:  AIC=349.32
## kq5_cond_parq1 ~ par_median
##
##                             Df Sum of Sq   RSS      Cp
## + pct_stem_2000              1      8488 33839  138.08
## + tier                       1      7735 34592  157.00
## + sqrtalien_share_fall_2000  1      3232 39095  270.12
## + hbcu                       1       688 41639  334.03
## + black_share_fall_2000      1       454 41874  339.93
## <none>                                   42327  349.32
## + endowment_pc_2000          1        60 42267  349.81
## + count                      1        15 42312  350.94
## - par_median                 1     41592 83920 1392.17
##
## Step:  AIC=138.08
## kq5_cond_parq1 ~ par_median + pct_stem_2000
##
##                             Df Sum of Sq   RSS      Cp
## + tier                       1    4094.4 29745  37.223
## + sqrtalien_share_fall_2000  1    2186.1 31653  85.162
## <none>                                   33839 138.080
## + hbcu                       1      68.0 33771 138.373
## + black_share_fall_2000      1      59.6 33779 138.583
## + endowment_pc_2000          1      54.3 33785 138.715
## + count                      1      36.8 33802 139.155
## - pct_stem_2000              1    8488.4 42327 349.318
## - par_median                 1   26416.5 60256 799.696
##
## Step:  AIC=37.22
## kq5_cond_parq1 ~ par_median + pct_stem_2000 + tier
##
##                             Df Sum of Sq   RSS      Cp
## + sqrtalien_share_fall_2000  1    1043.9 28701  12.999
## + black_share_fall_2000      1     303.1 29442  31.609
## + hbcu                       1     253.9 29491  32.845
## <none>                                   29745  37.223
## + endowment_pc_2000          1      60.9 29684  37.692
## + count                      1       1.6 29743  39.183
## - tier                       1    4094.4 33839 138.080
## - pct_stem_2000              1    4847.4 34592 156.996
## - par_median                 1   12212.0 41957 342.003
##
## Step:  AIC=13
## kq5_cond_parq1 ~ par_median + pct_stem_2000 + tier + sqrtalien_share_fall_2000
##
##                             Df Sum of Sq   RSS       Cp
## + black_share_fall_2000      1     330.4 28370   6.6981
## + hbcu                       1     275.1 28426   8.0877
## + endowment_pc_2000          1     109.3 28591  12.2520
## <none>                                   28701  12.9988
## + count                      1      18.3 28682  14.5384
## - sqrtalien_share_fall_2000  1    1043.9 29745  37.2230
## - tier                       1    2952.2 31653  85.1619
## - pct_stem_2000              1    4661.1 33362 128.0911
## - par_median                 1   10469.8 39171 274.0138
##
## Step:  AIC=6.7
## kq5_cond_parq1 ~ par_median + pct_stem_2000 + tier + sqrtalien_share_fall_2000 +
##     black_share_fall_2000
##
##                             Df Sum of Sq   RSS       Cp
## + endowment_pc_2000          1      93.3 28277   6.3532
## <none>                                   28370   6.6981
## + hbcu                       1      33.8 28336   7.8489
## + count                      1      17.0 28353   8.2698
## - black_share_fall_2000      1     330.4 28701  12.9988
## - sqrtalien_share_fall_2000  1    1071.3 29442  31.6094
## - tier                       1    3276.2 31646  86.9993
## - pct_stem_2000              1    4747.1 33117 123.9514
## - par_median                 1    5993.6 34364 155.2648
##
## Step:  AIC=6.35
## kq5_cond_parq1 ~ par_median + pct_stem_2000 + tier + sqrtalien_share_fall_2000 +
##     black_share_fall_2000 + endowment_pc_2000
##
##                             Df Sum of Sq   RSS       Cp
## <none>                                   28277   6.3532
## - endowment_pc_2000          1      93.3 28370   6.6981
## + hbcu                       1      35.3 28242   7.4676
## + count                      1      16.6 28260   7.9370
## - black_share_fall_2000      1     314.4 28591  12.2520
## - sqrtalien_share_fall_2000  1    1115.4 29392  32.3738
## - tier                       1    3242.8 31520  85.8157
## - pct_stem_2000              1    4732.2 33009 123.2331
## - par_median                 1    6084.4 34361 157.2000
```

```
##
## Call:
## lm(formula = kq5_cond_parq1 ~ par_median + pct_stem_2000 + tier +
##     sqrtalien_share_fall_2000 + black_share_fall_2000 + endowment_pc_2000,
##     data = newData)
##
## Coefficients:
##               (Intercept)                 par_median
##                 7.808e+00                  2.030e-04
##             pct_stem_2000                       tier
##                 2.323e-01                 -1.443e+00
## sqrtalien_share_fall_2000      black_share_fall_2000
##                 2.070e+01                 -4.345e+00
##         endowment_pc_2000
##                 -1.533e-05
```

All subsets approach

```
allsubsets<-regsubsets(kq5_cond_parq1~ count + endowment_pc_2000+par_median+pct_stem_2000+sqrtalien_sha
plot(allsubsets, scale= "adjr2")
```



Both gave us the same model!

We will also look at the correlation matrix to see if some variables are collinear

```
newData %>%
  select(c(sqrtalien_share_fall_2000,endowment_pc_2000, par_median, tier, black_share_fall_2000, pct_st
  cor()
```

```
##                           sqrtalien_share_fall_2000 endowment_pc_2000
## sqrtalien_share_fall_2000                1.00000000        0.12447010
## endowment_pc_2000                        0.12447010        1.00000000
## par_median                               0.37784467        0.13708928
## tier                                    -0.41122796       -0.08056512
## black_share_fall_2000                   -0.04115203       -0.01359417
## pct_stem_2000                            0.22244904        0.04110641
##                           par_median        tier black_share_fall_2000
## sqrtalien_share_fall_2000  0.3778447 -0.41122796         -0.0411520267
## endowment_pc_2000          0.1370893 -0.08056512         -0.0135941671
## par_median                 1.0000000 -0.55550003         -0.3925410156
## tier                      -0.5555000  1.00000000         -0.0848679726
## black_share_fall_2000     -0.3925410 -0.08486797          1.0000000000
## pct_stem_2000              0.3274408 -0.40968216          0.0003224705
##                           pct_stem_2000
## sqrtalien_share_fall_2000  0.2224490398
## endowment_pc_2000          0.0411064114
## par_median                 0.3274407968
## tier                      -0.4096821623
## black_share_fall_2000      0.0003224705
## pct_stem_2000              1.0000000000
```

Nothing is collinear!

Now lets fit this model

```
Lm1<-lm(kq5_cond_parq1~endowment_pc_2000+par_median+pct_stem_2000+sqrtalien_share_fall_2000+black_share_
summary(Lm1)
```

```
##
## Call:
## lm(formula = kq5_cond_parq1 ~ endowment_pc_2000 + par_median +
##     pct_stem_2000 + sqrtalien_share_fall_2000 + black_share_fall_2000 +
##     tier, data = newData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.054  -3.839  -0.917   3.091  48.023
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                7.808e+00  2.204e+00   3.543 0.000421 ***
## endowment_pc_2000         -1.533e-05  1.001e-05  -1.532 0.125967
## par_median                 2.030e-04  1.642e-05  12.369  < 2e-16 ***
## pct_stem_2000              2.323e-01  2.130e-02  10.908  < 2e-16 ***
## sqrtalien_share_fall_2000  2.070e+01  3.909e+00   5.296 1.58e-07 ***
## black_share_fall_2000     -4.345e+00  1.545e+00  -2.812 0.005063 **
## tier                      -1.443e+00  1.598e-01  -9.030  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.306 on 711 degrees of freedom
## Multiple R-squared:  0.663,  Adjusted R-squared:  0.6602
## F-statistic: 233.2 on 6 and 711 DF,  p-value: < 2.2e-16
```

Variance Inflaction Factor

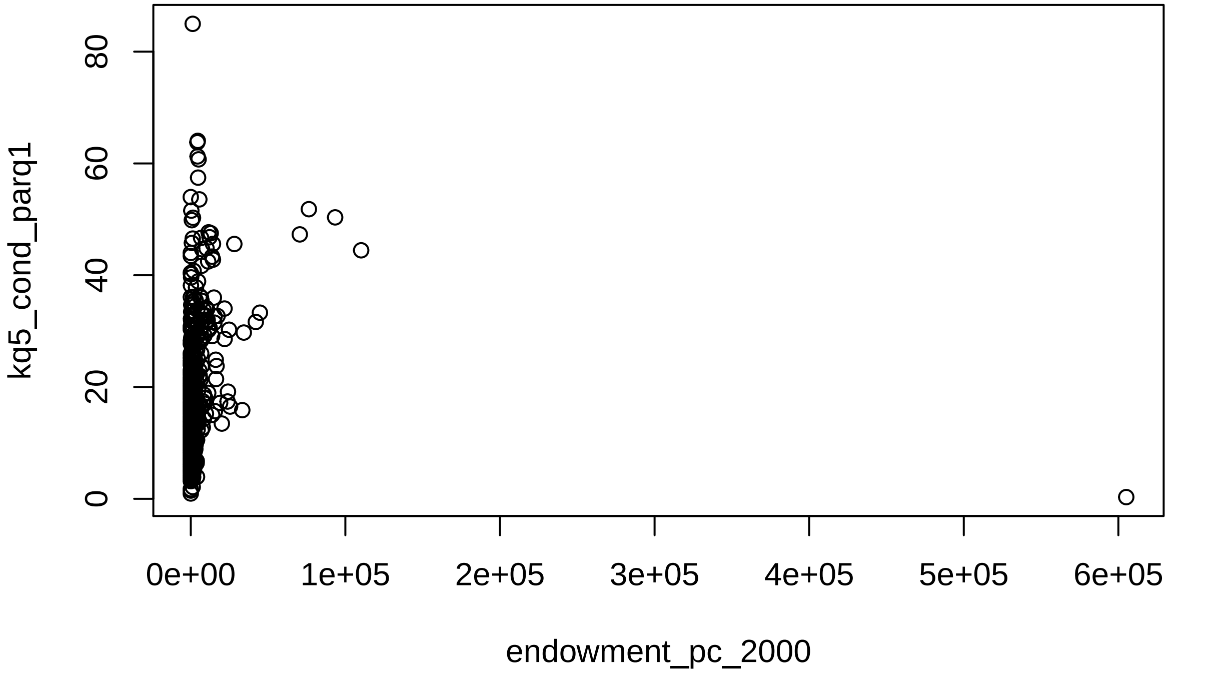
```
vif(Lm1)
```

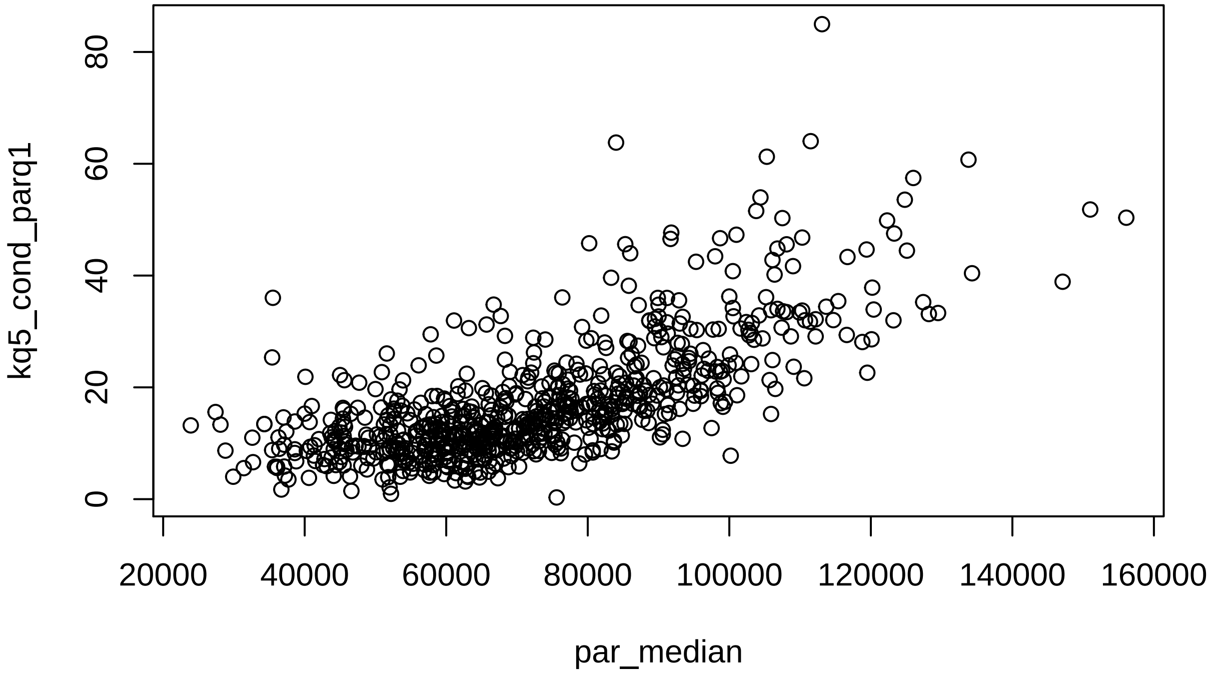
```
##         endowment_pc_2000                par_median
##                  1.027954                  2.135997
##             pct_stem_2000 sqrtalien_share_fall_2000
##                  1.226684                  1.262974
##     black_share_fall_2000                      tier
##                  1.407719                  1.925631
```

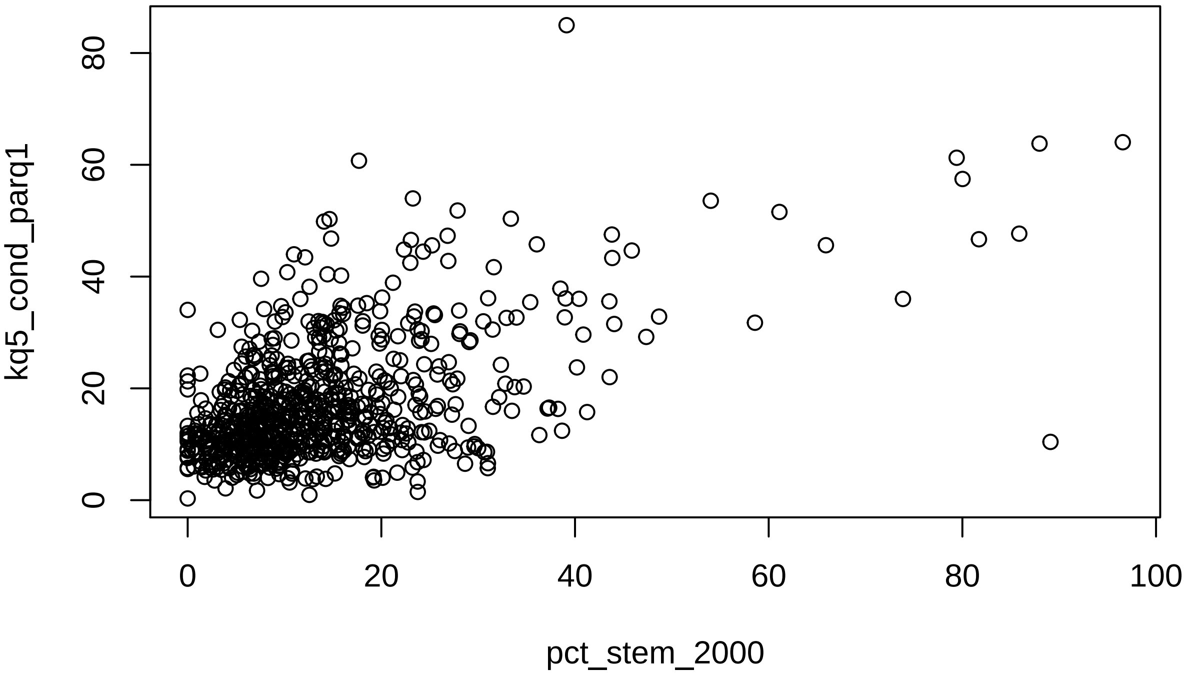
Everything looks good! Other *exploratory* analysis suggested that we may have some collinearity between par_median and tier so we will add an interaction term to explain this.

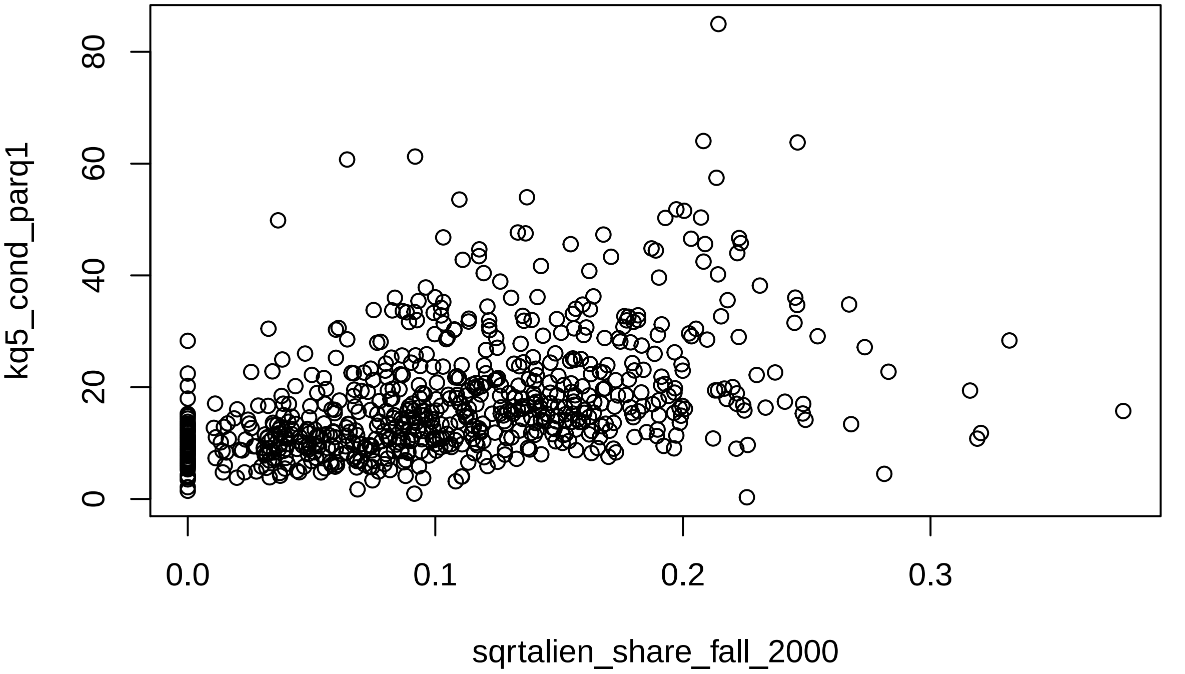
Model including the interaction term

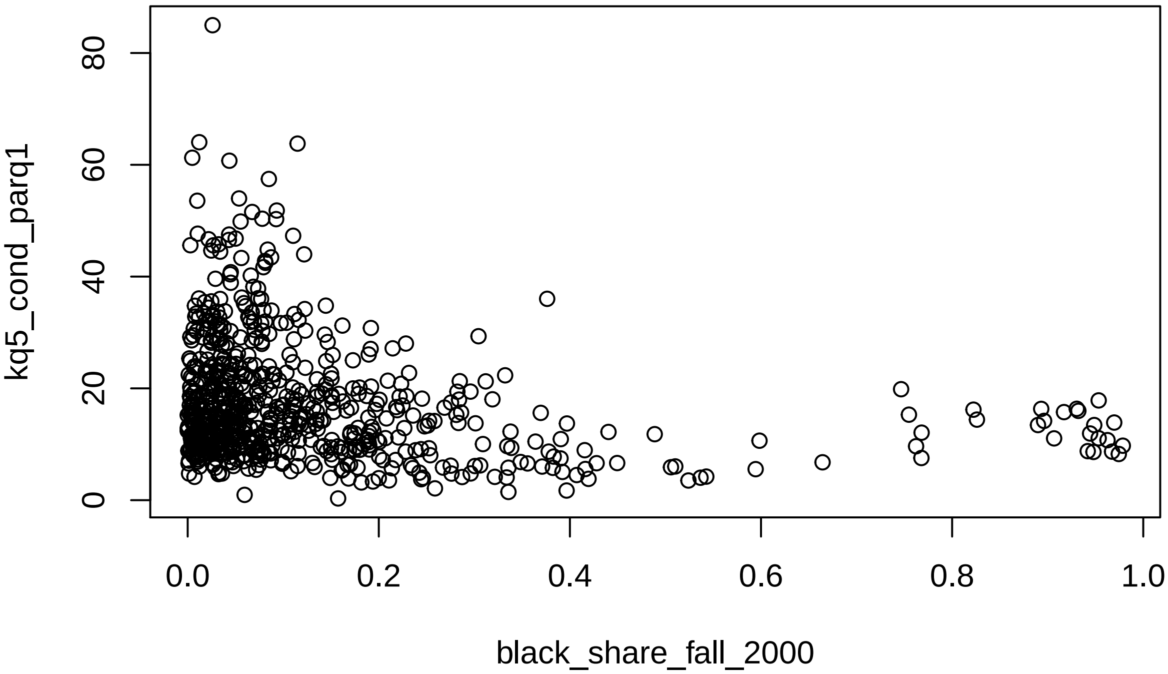
```
m1<-plot(kq5_cond_parq1~endowment_pc_2000+par_median+pct_stem_2000+sqrtalien_share_fall_2000+black_shar
```

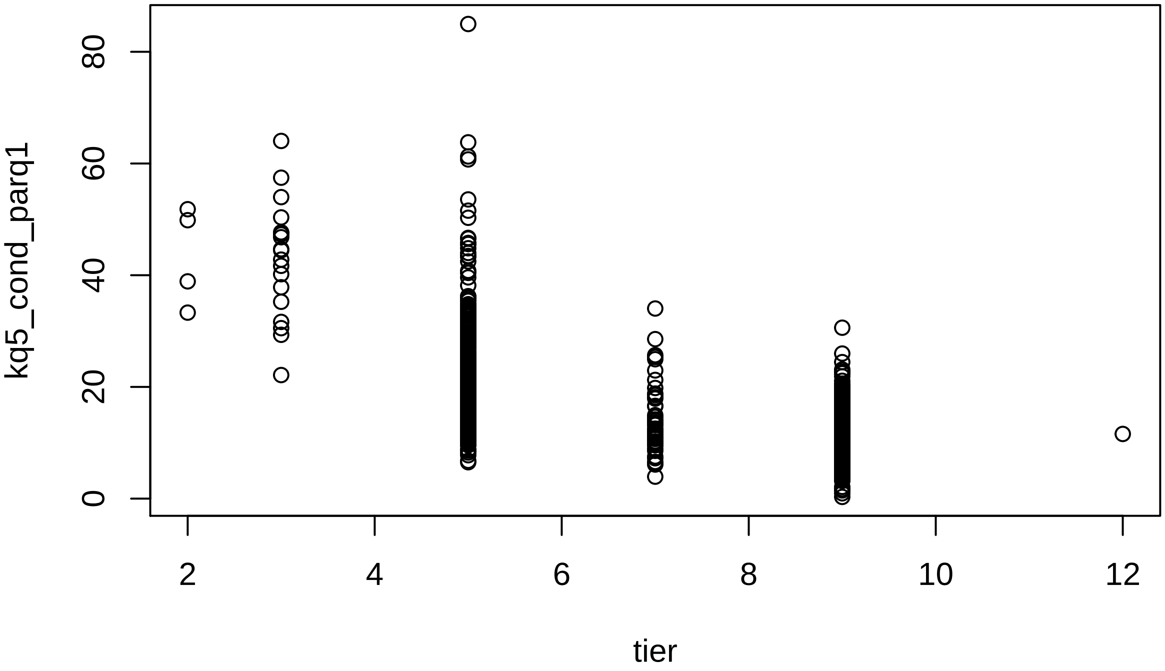





kq5_cond_parq1
0
20
40
60
80
0
20
40
60
80
100
pct_stem_2000


kq5_cond_parq1
0
20
40
60
80
0.0
0.1
0.2
0.3
sqrtalien_share_fall_2000





```
Lm1<-lm(kq5_cond_parq1~endowment_pc_2000+par_median+pct_stem_2000+sqrtalien_share_fall_2000+black_share_

summary(Lm1)
```

```
##
## Call:
## lm(formula = kq5_cond_parq1 ~ endowment_pc_2000 + par_median +
##     pct_stem_2000 + sqrtalien_share_fall_2000 + black_share_fall_2000 +
##     tier + par_median * tier, data = newData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.585  -3.656  -0.424   2.828  47.250
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -1.118e+01  3.388e+00  -3.302  0.00101 **
## endowment_pc_2000         -2.229e-05  9.717e-06  -2.294  0.02209 *
## par_median                 4.459e-04  3.724e-05  11.973  < 2e-16 ***
## pct_stem_2000              2.167e-01  2.068e-02  10.478  < 2e-16 ***
## sqrtalien_share_fall_2000  2.584e+01  3.843e+00   6.726 3.59e-11 ***
## black_share_fall_2000     -3.361e+00  1.499e+00  -2.242  0.02524 *
## tier                       1.494e+00  4.358e-01   3.429  0.00064 ***
## par_median:tier           -4.050e-05  5.619e-06  -7.207 1.46e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.092 on 710 degrees of freedom
## Multiple R-squared:  0.686,  Adjusted R-squared:  0.6829
## F-statistic: 221.6 on 7 and 710 DF,  p-value: < 2.2e-16
```

It explains 68% of the variability in the response.

Nested F- tests

Nested F-test to check endowment

```
Lm1<-lm(kq5_cond_parq1~endowment_pc_2000+par_median+pct_stem_2000+sqrtalien_share_fall_2000+black_share_

nested1 <- lm(kq5_cond_parq1~par_median+pct_stem_2000+sqrtalien_share_fall_2000+black_share_fall_2000+ t
anova(nested1, Lm1)
```

```
## Analysis of Variance Table
##
## Model 1: kq5_cond_parq1 ~ par_median + pct_stem_2000 + sqrtalien_share_fall_2000 +
##     black_share_fall_2000 + tier + par_median * tier
## Model 2: kq5_cond_parq1 ~ endowment_pc_2000 + par_median + pct_stem_2000 +
##     sqrtalien_share_fall_2000 + black_share_fall_2000 + tier +
##     par_median * tier
##   Res.Df   RSS Df Sum of Sq      F  Pr(>F)
## 1    711 26544
## 2    710 26349  1    195.26 5.2614 0.02209 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The nested model does not have endowment. Since the p-value < 0.05 we should use the full model and keep endowment.

Nested F-test to check par_median

```
Lm<-lm(kq5_cond_parq1~endowment_pc_2000+par_median+pct_stem_2000+sqrtalien_share_fall_2000+black_share_
nested2 <- lm(kq5_cond_parq1~endowment_pc_2000 + pct_stem_2000+sqrtalien_share_fall_2000+ black_share_fa
anova(nested2, Lm)
```

```
## Analysis of Variance Table
##
## Model 1: kq5_cond_parq1 ~ endowment_pc_2000 + pct_stem_2000 + sqrtalien_share_fall_2000 +
##     black_share_fall_2000 + tier
## Model 2: kq5_cond_parq1 ~ endowment_pc_2000 + par_median + pct_stem_2000 +
##     sqrtalien_share_fall_2000 + black_share_fall_2000 + tier +
##     par_median * tier
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    712 34361
## 2    710 26349  2      8012 107.94 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The nested model does not have par_median Since the p-value < 0.05 we should use the full model and keep par_median.

Nested F-test to check tier

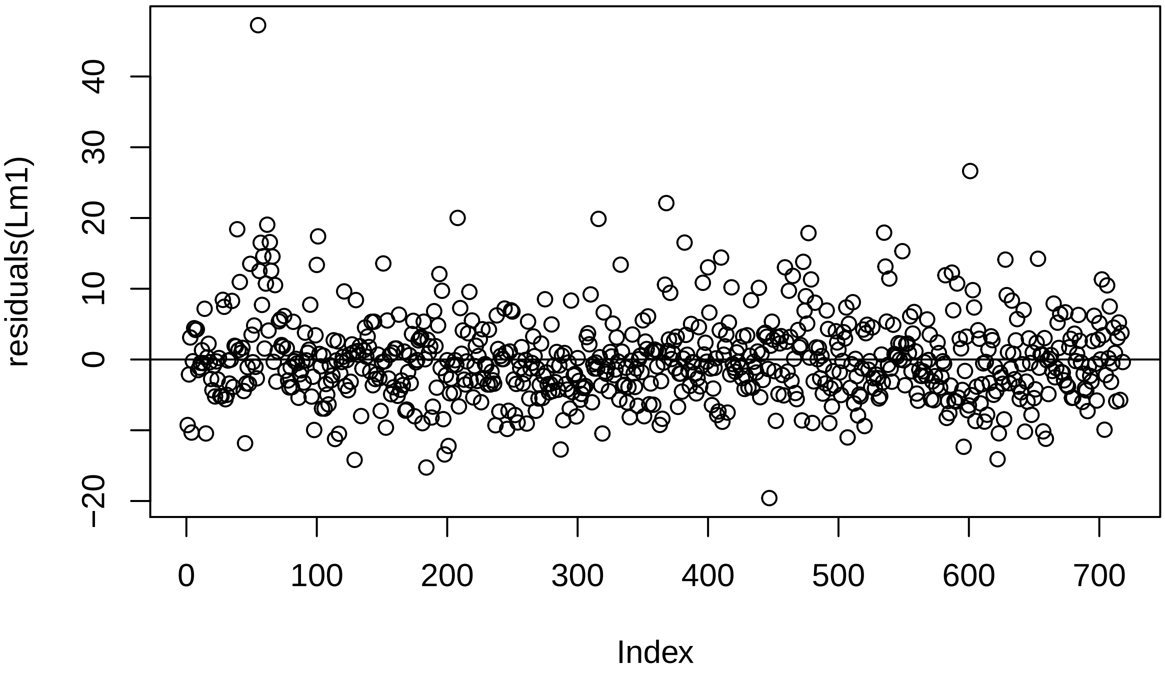
```
Lm<-lm(kq5_cond_parq1~endowment_pc_2000+par_median+pct_stem_2000+sqrtalien_share_fall_2000+black_share_

nested2 <- lm(kq5_cond_parq1~ endowment_pc_2000 + pct_stem_2000+sqrtalien_share_fall_2000+ black_share_
anova(nested2, Lm)
```

```
## Analysis of Variance Table
##
## Model 1: kq5_cond_parq1 ~ endowment_pc_2000 + pct_stem_2000 + sqrtalien_share_fall_2000 +
##     black_share_fall_2000 + par_median
## Model 2: kq5_cond_parq1 ~ endowment_pc_2000 + par_median + pct_stem_2000 +
##     sqrtalien_share_fall_2000 + black_share_fall_2000 + tier +
##     par_median * tier
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    712 31520
## 2    710 26349  2    5170.4 69.661 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

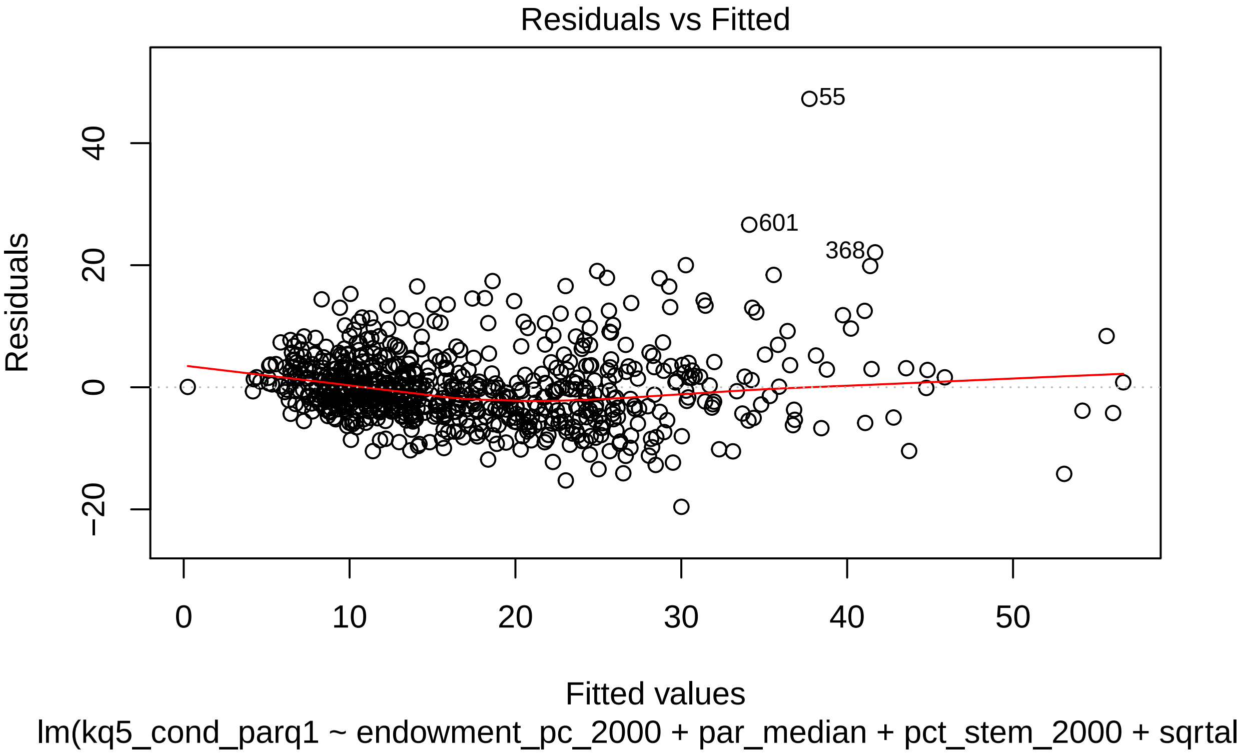
The nested model does not have tier Since the p-value < 0.05 we should use the full model and keep tier

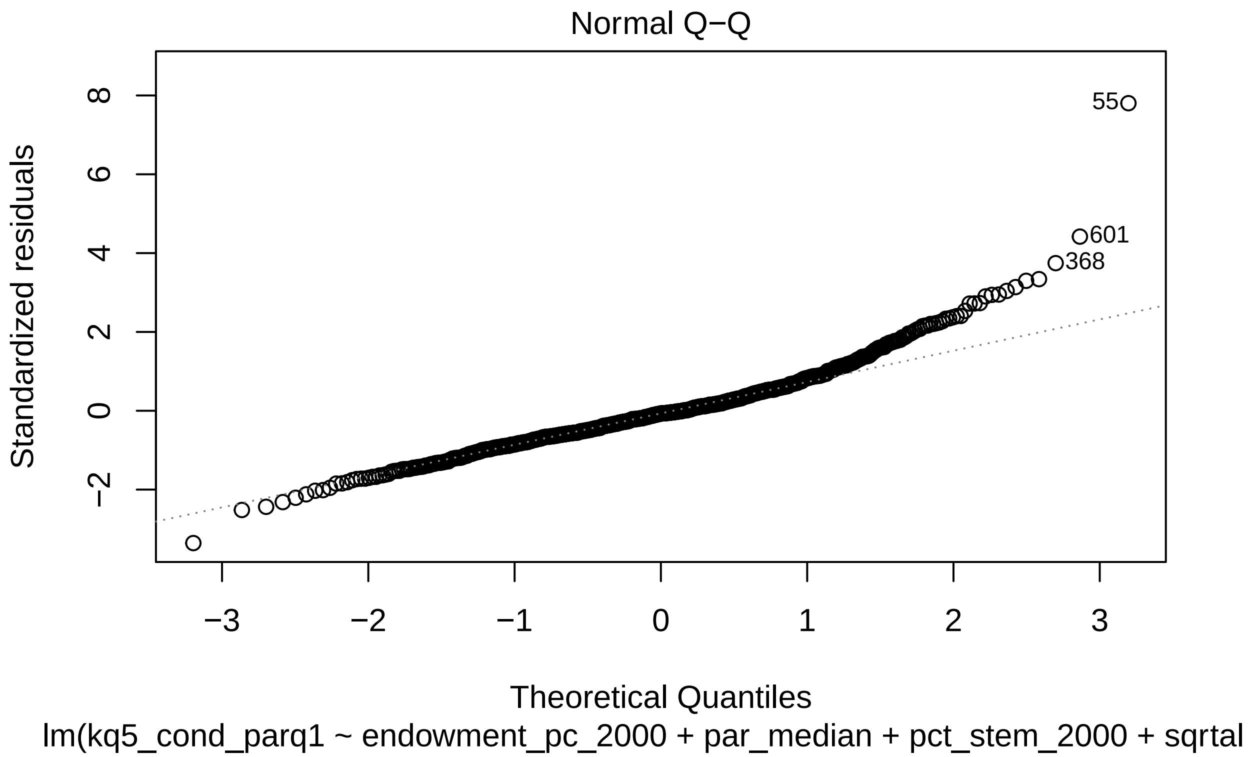
Analysis of residuals

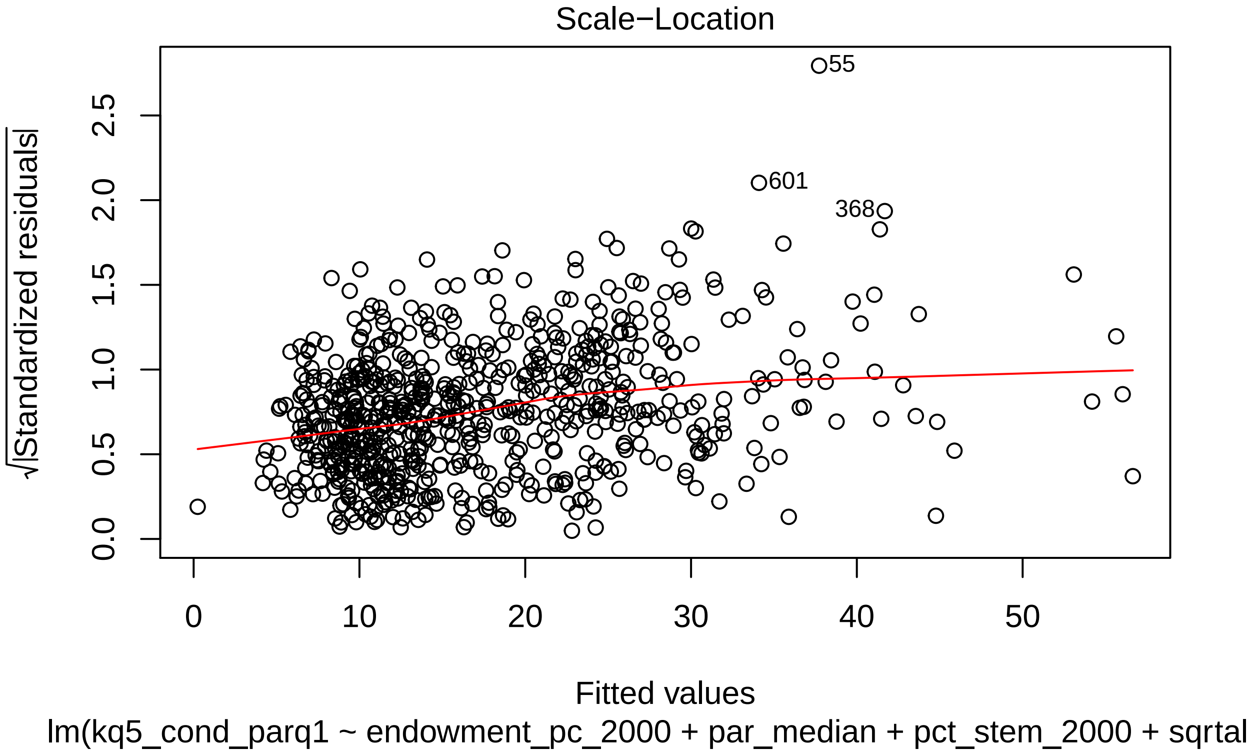
```
plot(residuals(Lm1))
abline(0,0)
```

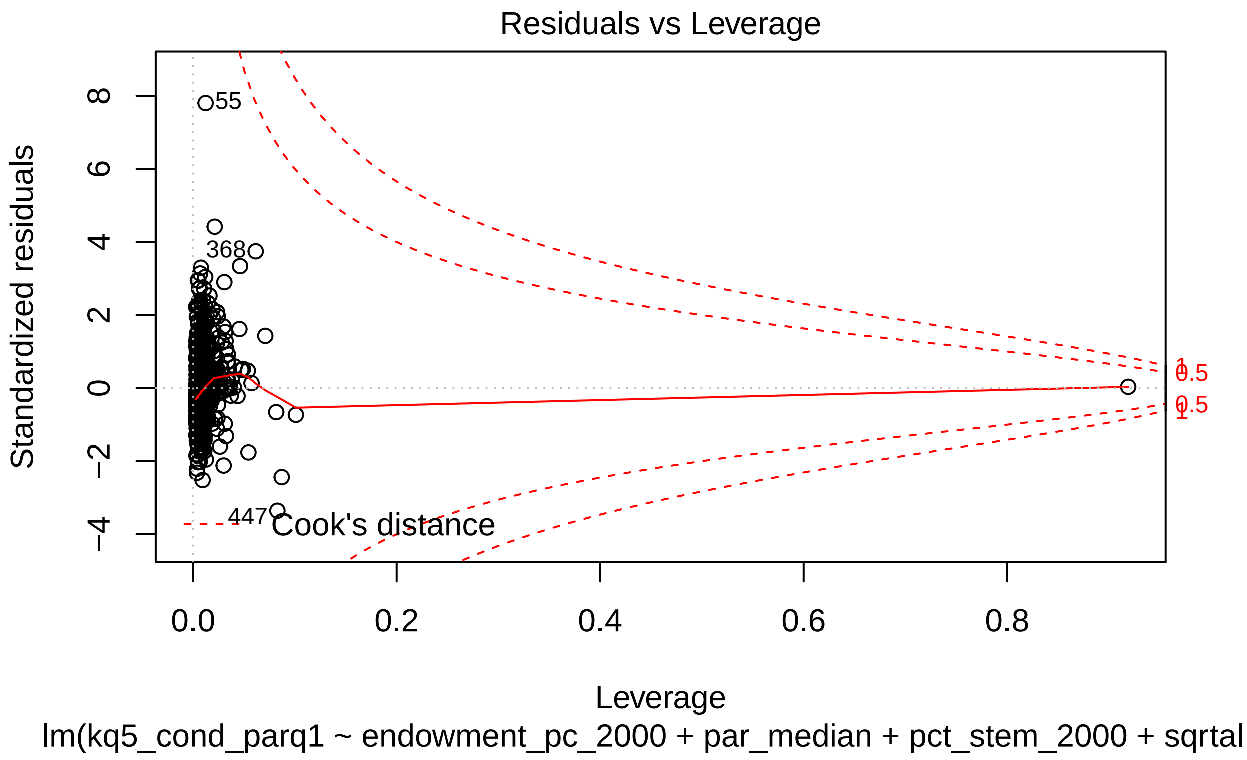

residuals(Lm1)
40
30
20
10
0
−20
0
100
200
300
400
500
600
700
Index

```
plot(Lm1)
```


Residuals vs Fitted
Residuals
40
20
0
−20
55
601
368
0
10
20
30
40
50
Fitted values
lm(kq5_cond_parq1 ~ endowment_pc_2000 + par_median + pct_stem_2000 + sqrtal


Normal Q−Q
Standardized residuals
−2
0
2
4
6
8
55
601
368
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(kq5_cond_parq1 ~ endowment_pc_2000 + par_median + pct_stem_2000 + sqrtal


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
55
601
368
0
10
20
30
40
50
Fitted values
lm(kq5_cond_parq1 ~ endowment_pc_2000 + par_median + pct_stem_2000 + sqrtal



```
summary(Lm1)
```

```
##
## Call:
## lm(formula = kq5_cond_parq1 ~ endowment_pc_2000 + par_median +
##     pct_stem_2000 + sqrtalien_share_fall_2000 + black_share_fall_2000 +
##     tier + par_median * tier, data = newData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.585  -3.656  -0.424   2.828  47.250
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -1.118e+01  3.388e+00  -3.302  0.00101 **
## endowment_pc_2000         -2.229e-05  9.717e-06  -2.294  0.02209 *
## par_median                 4.459e-04  3.724e-05  11.973  < 2e-16 ***
## pct_stem_2000              2.167e-01  2.068e-02  10.478  < 2e-16 ***
## sqrtalien_share_fall_2000  2.584e+01  3.843e+00   6.726 3.59e-11 ***
## black_share_fall_2000     -3.361e+00  1.499e+00  -2.242  0.02524 *
## tier                       1.494e+00  4.358e-01   3.429  0.00064 ***
## par_median:tier           -4.050e-05  5.619e-06  -7.207 1.46e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.092 on 710 degrees of freedom
## Multiple R-squared:  0.686,  Adjusted R-squared:  0.6829
## F-statistic: 221.6 on 7 and 710 DF,  p-value: < 2.2e-16
```

After having checked the importance of our variable, our fitted model is

$$\widehat{kq5\_cond\_parq1} = -11.18 - 0.00002229 endowment\_pc\_2000 + 0.0004459 par\_median + 0.2167 pct\_stem\_2000 + 25.84 sqrtalien\_share\_fall\_2000 - 3.361 black\_share\_fall\_2000 + 1.494 tier + 0.0000405 par\_median \cdot tier$$

68% of the variability in the percentage of children who reach the Top 20% of the income distribution among children with parents in the bottom 20% of the income distribution (kq5_cond_parq1) is explained by the model.

According to the model:

- When there are no variables, the kq5_cond_parq1 decreases by 11.18%, *Holding all else constant, a $1 increase in endowment results in a 0.00002229% decrease in kq5_cond_parq1,
- Holding all else constant, a $1 increase in parents' median income results in a 0.0004459% increase in kq5_cond_parq1,
- Holding all else constant, a percentage increase in stem students results in a 0.2167% increase in kq5_cond_parq1,
- Holding all else constant, if we add one more foreign student, we increase the percentage of kq5_cond_parq1 by 25.84%,
- Holding all else constant, if we add one more black student, kq5_cond_parq1 decreases by 3.361%, *Holding all else constant, a one unit increase in tier increases the kq5_cond_parq1 by 1.494%. We will need to make this variable a factor.* Holding all else constant, a $1 increase in parents median results in a 0.0000405% increase in kq5_cond_parq1 while moderating for tier.