Tertiary Enrollment Model

By: Nathan Smith, Hafid Pradipta, and Eric Baron

Research Question

What societal factors can predict the gross enrollment rate in tertiary education in a given country?

Response Variable

- For this research, we want to find out to what extent the size of a population living in urban areas, gender inequality, media influence, and social globalization are good predictors of variance in the **total gross enrollment in tertiary education**.
- Additionally, we want to know whether democratic countries will demonstrate higher enrollment than non-democratic.

Predictor Variables

➔ *Urban Population*
- *Percent of the population living in urban areas*

➔ *Gender Inequality*
- *Transformed from .0-1 ratio to percent to reflect degree of inequality*

➔ *Media Influence*
- *Combination of daily newspapers, tv sets, and internet access per 1,000 people*

➔ *Social Globalization*

➔ *Democratic*
- *A dichotomous variable that reads 0 = non democratic and 1 = democratic*

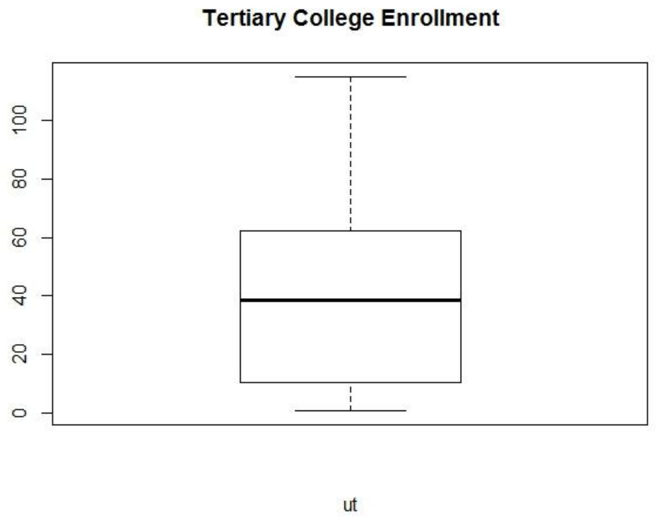
Hypothesis

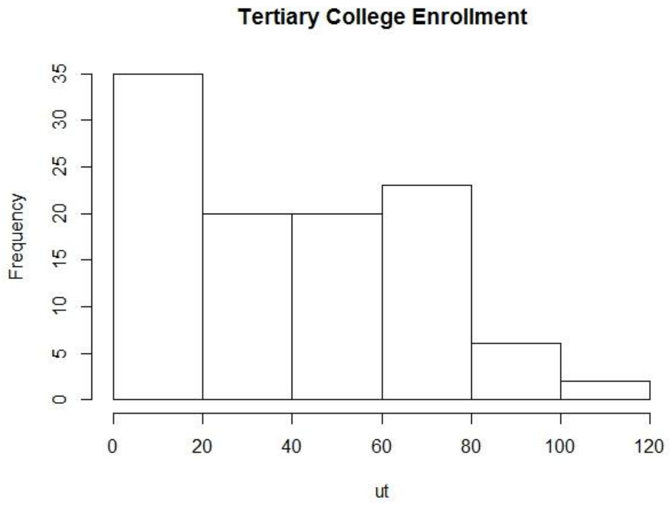
H_o: There exists no relationship between tertiary education enrollment and urban population, gender inequality, media influence, social globalization, or whether or not a country is democratic. These societal factors will not be able to predict a rise or fall in the rate of tertiary enrollment.

H_a: There is at least one inequality in this model.

Preliminary Analysis Full Model

Box Plot and Histogram of Tertiary College Enrolment

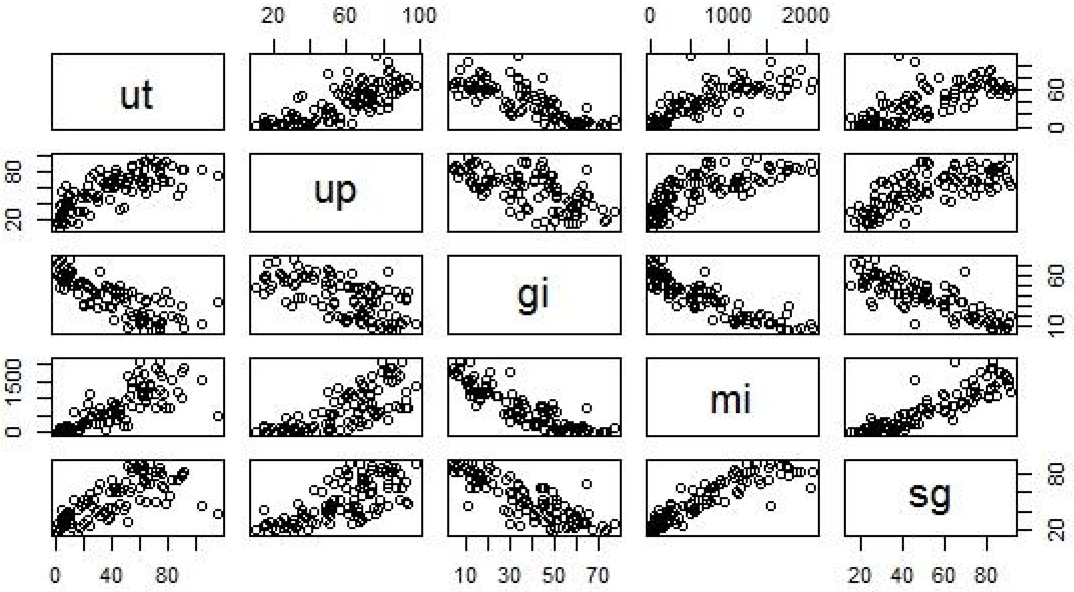




Correlation among Variables

	ut	up	gi	mi	sg	dc
ut	1.0000000	0.7578263	-0.8020527	0.7975787	0.7480792	0.5627572
up	0.7578263	1.0000000	-0.6100478	0.7313970	0.7036780	0.5170954
gi	-0.8020527	-0.6100478	1.0000000	-0.8801039	-0.8343895	-0.5891803
mi	0.7975787	0.7313970	-0.8801039	1.0000000	0.8908791	0.6257261
sg	0.7480792	0.7036780	-0.8343895	0.8908791	1.0000000	0.6060181
dc	0.5627572	0.5170954	-0.5891803	0.6257261	0.6060181	1.0000000

Preliminary Analysis Initial Model

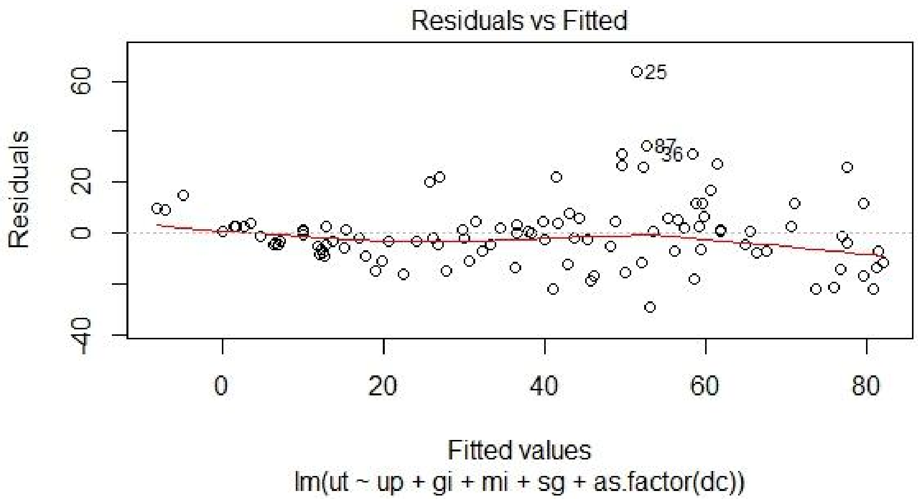


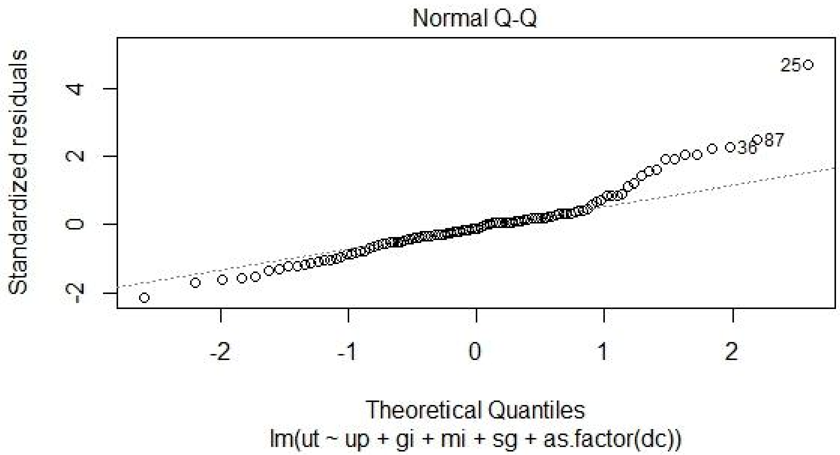
Coefficient of Initial Model

```
> summary(dataprojreg)

Call:
lm(formula = ut ~ up + gi + mi + sg + as.factor(dc), data = dataproject)

Residuals:
    Min      1Q  Median      3Q     Max 
-29.518  -7.094  -1.583   4.379  63.432 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)    39.140301  11.096380   3.527 0.000636 ***
up              0.519257   0.092999   5.583 2.03e-07 ***
gi             -0.763810   0.160749  -4.752 6.77e-06 ***
mi              0.003468   0.006735   0.515 0.607798    
sg             -0.071184   0.144443  -0.493 0.623220    
as.factor(dc)1  2.040006   3.593202   0.568 0.571484    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.15 on 100 degrees of freedom
Multiple R-squared:  0.7598,	Adjusted R-squared:  0.7478 
F-statistic: 63.27 on 5 and 100 DF,  p-value: < 2.2e-16
```

Residual of Initial Model

Residual plot and QQ plot of Initial Model





Residual of Initial Model

Shapiro Wilk Test for the normality of residual

Shapiro-Wilk normality test
data: datareg$residuals
W = 0.91718, p-value = 5.751e-06

Ho: Data is normally distributed
Ha: Data is not normally distributed
Interpretation: p-value of Shapiro Wilk test is less than our significance which is 0.05. It means that there is no significance evidence to support that Data is normally distributed. We reject the null hypothesis and accept alternate hypothesis.

Residual of Initial Model

Breusch Pagan Test for Constant Variance

Breusch Pagan test for full model:

```
        studentized Breusch-Pagan test
data:  datareg
BP = 10.778, df = 5, p-value = 0.05597
```

Hypothesis:
Ho: The variance of data is constant
Ha: The variance of data is not constant

Since the p-value is larger than 0.05, we fail to reject null hypothesis hence the variance of residual is still considered constant with 5% significance value.

Residual of Initial Model

Studentized Residuals

Critical Value for studentized Residual with Bonferonni method of n =105.

Individual significance = (0.05/2)/105 = 0.0002

Critical value based on t distribution = t(0.9998, 105) = 3.657524

Studentized residual for case 25th is considered as an outlier because the value is 5.260586442 hence this data is omitted from our model

Variable Selection

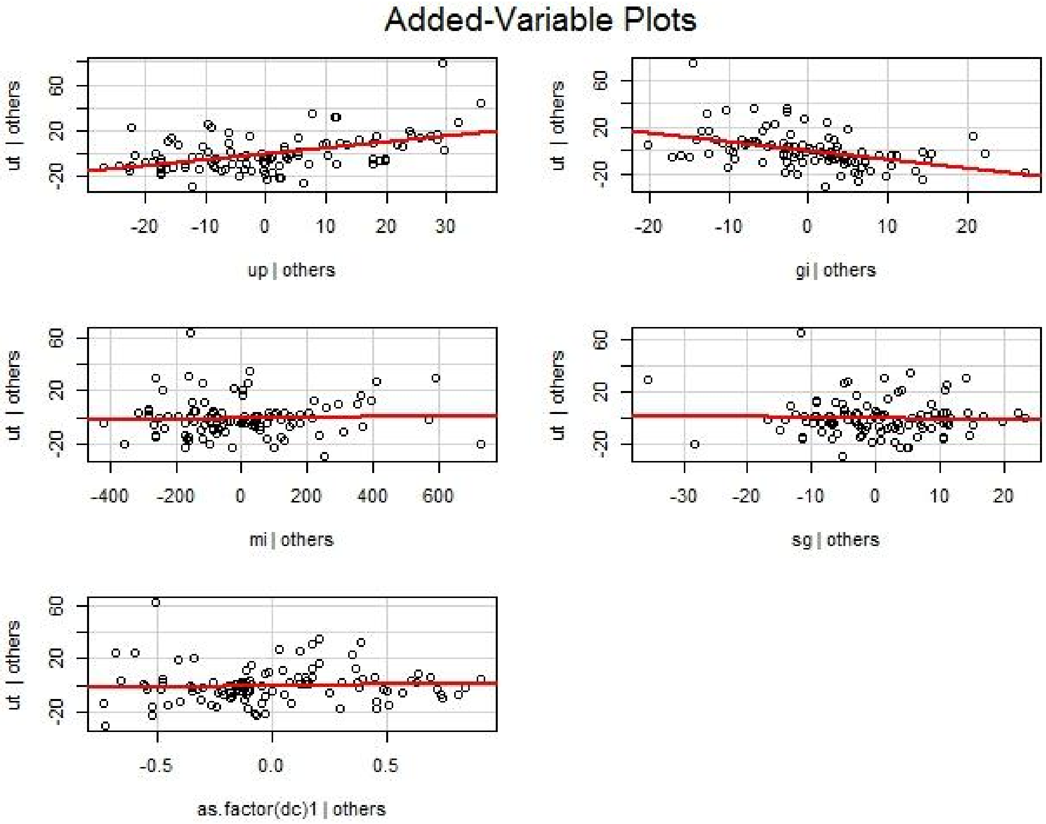
Stepwise Method

Lowest AIC is obtained based only two predictors which are urban population (up) and gender inequality(gi)

```
Step:  AIC=562.35
ut ~ up + gi

       Df Sum of Sq   RSS    AIC
<none>              20172 562.35
- up    1    9578.8 29750 601.54
- gi    1   15331.0 35503 620.28

Call:
lm(formula = ut ~ up + gi, data = dataproject)

Coefficients:
(Intercept)           up           gi
    39.6000       0.5366      -0.8069
```

Variable Selection

Added Variable Plot



Variable Selection

Subset of the model

Two predictors: Urban population and Gender Inequality

However we want to check with 3 predictors by testing democratic (dc) in our model

```
> bestsub
Subset selection object
Call: regsubsets.formula(ut ~ up + gi + mi + sg + as.factor(dc), data = dataproject)
5 Variables  (and intercept)
               Forced in Forced out
up                 FALSE      FALSE
gi                 FALSE      FALSE
mi                 FALSE      FALSE
sg                 FALSE      FALSE
as.factor(dc)1     FALSE      FALSE
1 subsets of each size up to 5
Selection Algorithm: exhaustive
         up  gi  mi  sg  as.factor(dc)1
1  ( 1 ) " " "*" " " " " " "
2  ( 1 ) "*" "*" " " " " " "
3  ( 1 ) "*" "*" " " " " "*"
4  ( 1 ) "*" "*" "*" " " "*"
5  ( 1 ) "*" "*" "*" "*" "*"
```

Model 2

Democracy as qualitative predictor is not significant in our second model, therefore we decided not to continue with this model and move to model 3

```
> anova(dataprojreg3,dataprojreg2)
Analysis of Variance Table

Model 1: ut ~ up + gi
Model 2: ut ~ up + gi + as.factor(dc)
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1    102 16242
2    101 15866  1    376.46 2.3965 0.1247
```

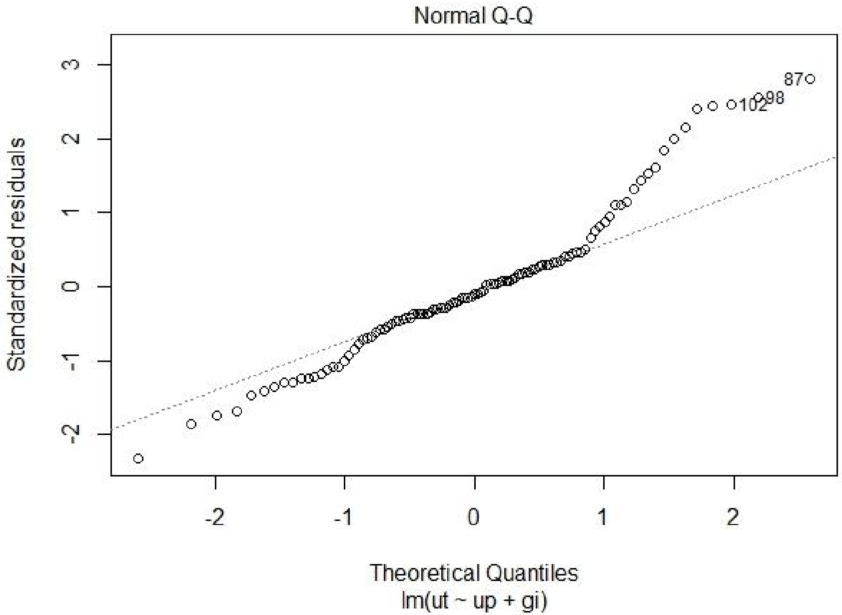
Mallow CP

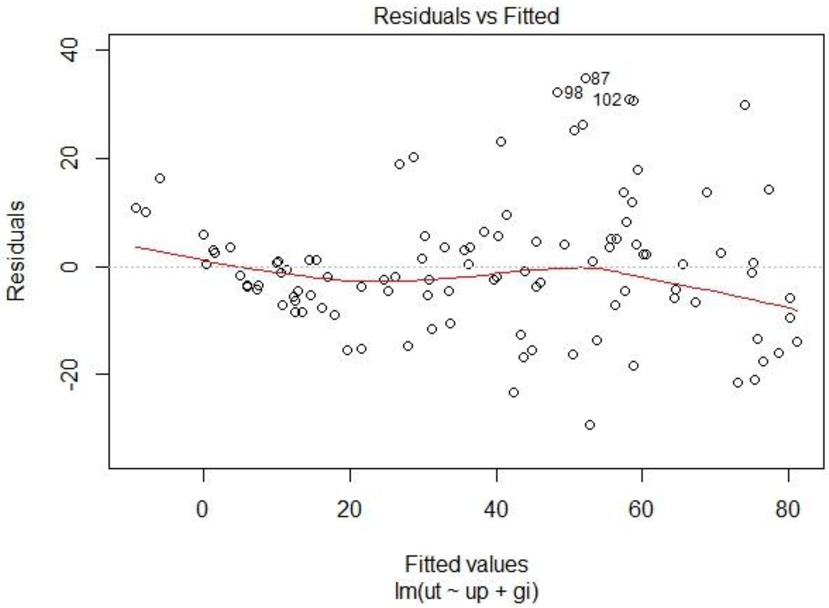
```
> (15866/200)-(105-2*3)
[1] -19.67
```

Model 3 Reduced Model

```
> summary(dataprojreg3)

Call:
lm(formula = ut ~ up + gi, data = dataproject2)

Residuals:
    Min      1Q  Median      3Q     Max 
-29.191  -6.680  -1.218   4.456  34.736 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 41.02502    6.53795   6.275 8.58e-09 ***
up           0.50930    0.06940   7.338 5.37e-11 ***
gi          -0.81871    0.08227  -9.951  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.62 on 102 degrees of freedom
Multiple R-squared:  0.7908,    Adjusted R-squared:  0.7867 
F-statistic: 192.8 on 2 and 102 DF,  p-value: < 2.2e-16
```

Residual of Reduced Model





Residual of Model 3 (Reduced model)

Shapiro-Wilk normality test
W = 0.95606, p-value = 0.001553

Ho: Data is normally distributed
Ha: Data is not normally distributed

P-value of model 3 (reduced model) for Shapiro Wilk test is less than 0.05 which means we reject the null hypothesis. The residual of this model is not normally distributed.

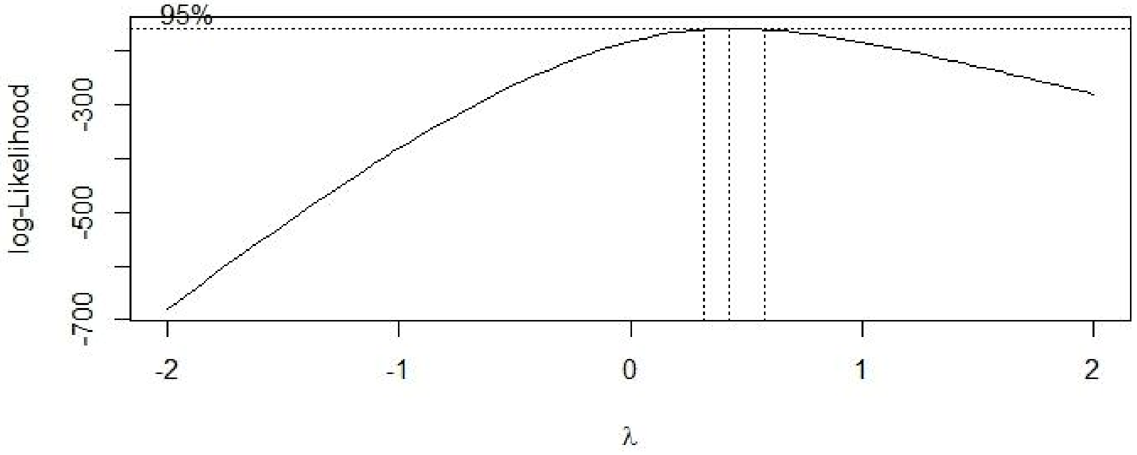
Breusch Pagan Test for constant variance
BP = 9.6214, df = 2, p-value = 0.008142
Hypothesis:
Ho: The variance of data is constant
Ha: The variance of data is not constant

Since the p-value is larger than 0.05, we fail to reject null hypothesis hence the variance of residual is still considered constant with 5% significance value.

Remedial Measure

Transformation

Since response variable is not normally distributed, we would like to suggest the transformation in response variable. Based on log-likelihood of box cox transformation, we would try to try with ⅄ =0.5



Model 4 (Reduced and Remedial)

```
> summary(dataprojreg4)

Call:
lm(formula = dataproject2$squt ~ up + gi, data = dataproject2)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.1917 -0.7792  0.0757  0.5306  2.4753 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  5.533668   0.537362  10.298  < 2e-16 ***
up           0.051554   0.005704   9.038 1.09e-14 ***
gi          -0.072353   0.006762 -10.700  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.037 on 102 degrees of freedom
Multiple R-squared:  0.8307,	Adjusted R-squared:  0.8274 
F-statistic: 250.3 on 2 and 102 DF,  p-value: < 2.2e-16
```

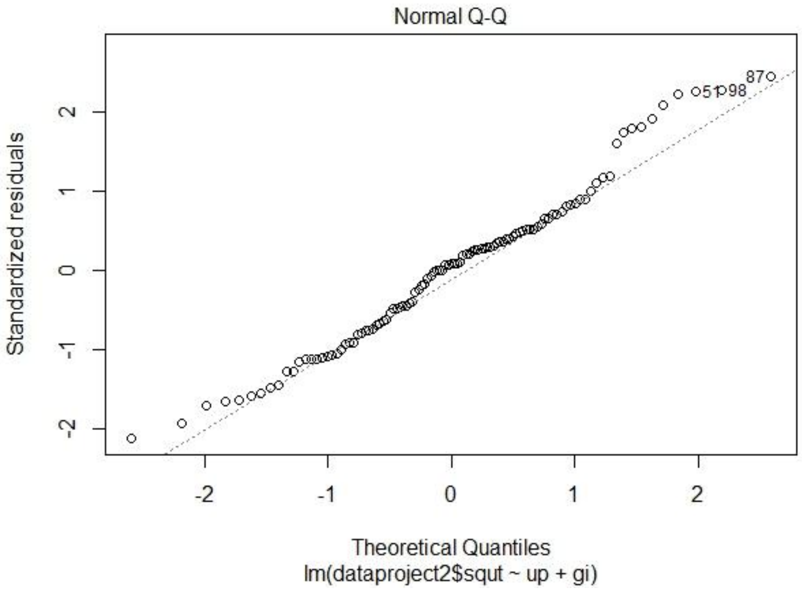
Model 4 (Reduced and Remedial)

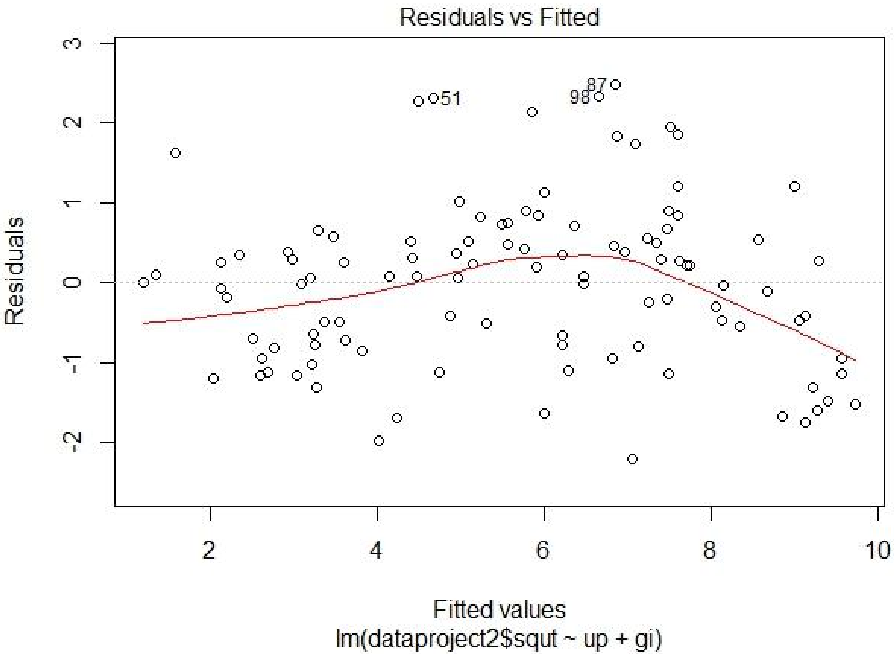
Regression Equation Model 4

$\sqrt{Yhat} = 5.533668 + 0.051554x1 - 0.072353x2$

$Yhat = (5.533668 + 0.051554x1 - 0.072353x2)^2$

Residual of Model 4





Residual of Model 4 (Reduced and Remedial)

Shapiro-Wilk normality test
data: dataprojreg4$residuals
W = 0.97928, p-value = 0.09942

Ho: Data is normally distributed
Ha: Data is not normally distributed

P-value of model 4 (reduced and model) for Shapiro Wilk test is more than 0.05 which means we fail to reject the null hypothesis. The residual of this model is considered normally distributed

Breusch Pagan Test for constant variance
BP = 2.6661, df = 2, p-value = 0.2637
Hypothesis:
Ho: The variance of data is constant
Ha: The variance of data is not constant

Since the p-value is larger than 0.05, we fail to reject null hypothesis hence the variance of residual is still considered constant with 5% significance value.

Model Validation

```
> selcri(dataprojreg)
           rsq   adj.rsq      aic      bic   press
[1,] 0.7598031 0.7477932 567.6126 583.5932 23069.8
> selcri(dataprojreg2)
           rsq   adj.rsq      aic      bic    press
[1,] 0.7956282 0.7895578 534.8865 545.5023 17173.24
> selcri(dataprojreg3)
          rsq   adj.rsq      aic      bic    press
[1,] 0.790779 0.7866766 535.3488 543.3106 17176.36
> selcri(dataprojreg4)
           rsq   adj.rsq      aic      bic    press
[1,] 0.8307107 0.8273913 10.62026 18.58215 116.2082
```

Analysis of the Final Model

- After adjusting for the effects of gender inequality,for every 1 percent increase in urban population, gross tertiary enrollment increases by 0.052 percent
- After adjusting for the effects of urban population, for every 1 percent increase in gender inequality, gross tertiary enrollment will decrease by 0.073 percent
- R-squared indicates that our model accounts for 83 percent of the variance in gross tertiary enrollment across 106 countries

Conclusion

- With p-values of less than 0.01 for both of our predictor variables, we reject our null hypothesis and accept that both the degree of gender inequality and the percentage of a total urban population is significantly associated with the variance in total tertiary enrollment in a given country.