Linear Regression

Victoria Okereke

```r
#importing libraries
library(faraway)
library(visdat)
library(olsrr)
```

```
##
## Attaching package: 'olsrr'

## The following object is masked from 'package:faraway':
##
##     hsb

## The following object is masked from 'package:datasets':
##
##     rivers
```

```r
library(lmtest)
```

```
## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
library(caret)
```

```
## Loading required package: ggplot2

## Loading required package: lattice

##
## Attaching package: 'lattice'

## The following object is masked from 'package:faraway':
##
##     melanoma
```

```r
library(kernlab)
```

```
##
## Attaching package: 'kernlab'

## The following object is masked from 'package:ggplot2':
##
##     alpha
```

```r
library(ipred)
#setting seed
set.seed(123)
```

Aim: To predict wage from the usawage dataset in the faraway library

Data Exploration

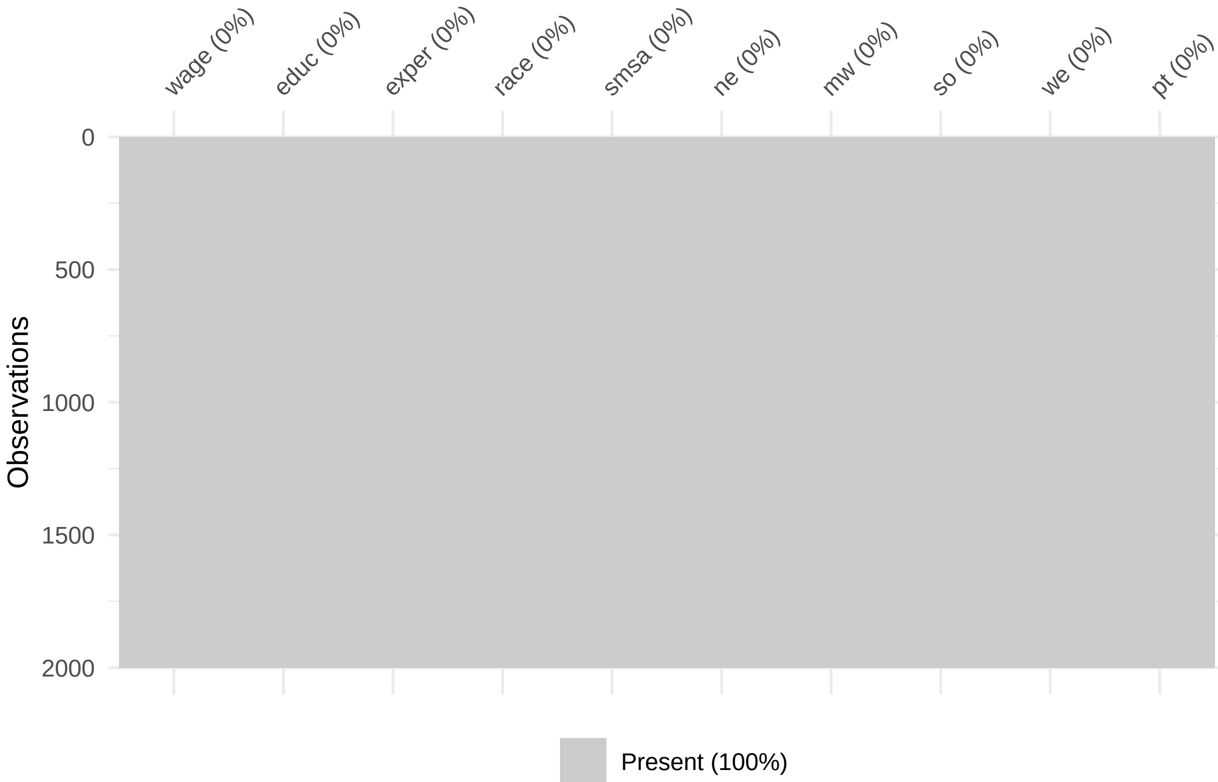
```r
#reading in dataset
data("uswages")
#viewing data structure
str(uswages)
```

```
## 'data.frame':    2000 obs. of  10 variables:
##  $ wage : num  772 617 958 617 902 ...
##  $ educ : int  18 15 16 12 14 12 16 16 12 12 ...
##  $ exper: int  18 20 9 24 12 33 42 0 36 37 ...
##  $ race : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ smsa : int  1 1 1 1 1 1 1 1 1 0 ...
##  $ ne   : int  1 0 0 1 0 0 0 0 0 0 ...
##  $ mw   : int  0 0 0 0 1 0 0 1 0 1 ...
##  $ so   : int  0 0 1 0 0 0 1 0 0 0 ...
##  $ we   : int  0 1 0 0 0 1 0 0 1 0 ...
##  $ pt   : int  0 0 0 0 0 0 1 1 1 0 ...
```

```r
#viewing first 6 rows of data
head(uswages)
```

```
##         wage educ exper race smsa ne mw so we pt
## 6085  771.60   18    18    0    1  1  0  0  0  0
## 23701 617.28   15    20    0    1  0  0  0  1  0
## 16208 957.83   16     9    0    1  0  0  1  0  0
## 2720  617.28   12    24    0    1  1  0  0  0  0
## 9723  902.18   14    12    0    1  0  1  0  0  0
## 22239 299.15   12    33    0    1  0  0  0  1  0
```

```r
#viewing the pattern of missingness
vis_miss(uswages)
```



No missing data so we do not need to worry about missingness.

A careful review of the data shows that columns ne, mw, so, and we seem to have been coded from the same categorical variable so we will drop one of them from the model

```r
#dropping the 'we' variable
uswages_reduced = uswages[-c(9)]
#fitting the linear regression model
uswages_reg = lm(wage ~., data = uswages_reduced)
#getting a summary statistics
summary(uswages_reg)
```

```
##
## Call:
## lm(formula = wage ~ ., data = uswages_reduced)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -875.7 -213.8  -53.3  128.5 7505.5
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -203.9184    53.6126  -3.804 0.000147 ***
## educ          48.8034     3.2489  15.022  < 2e-16 ***
## exper          9.1353     0.7262  12.579  < 2e-16 ***
## race        -119.1585    35.1922  -3.386 0.000723 ***
## smsa         115.6783    21.7386   5.321 1.15e-07 ***
## ne           -53.9265    27.9738  -1.928 0.054028 .
## mw           -60.1990    27.3487  -2.201 0.027839 *
## so           -50.4333    26.3703  -1.913 0.055955 .
```

```
## pt          -336.2156    31.9381 -10.527  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 412.1 on 1991 degrees of freedom
## Multiple R-squared:  0.2001, Adjusted R-squared:  0.1969
## F-statistic: 62.25 on 8 and 1991 DF,  p-value: < 2.2e-16
```

Some variables are not significant. Let's use a variable selection method to retain only significant variables in the model

```
#performing a stepwise both ways variable selection
kstep_both = ols_step_both_p(uswages_reg,pent=0.1,prem=0.05)#,details = TRUE)
kstep_both
```

```
##
##                                   Stepwise Selection Summary
## ---------------------------------------------------------------------------------------------
##                        Added/                   Adj.
## Step    Variable      Removed      R-Square    R-Square      C(p)          AIC           RMSE
## ---------------------------------------------------------------------------------------------
##    1      educ       addition        0.062       0.061     339.4730    30076.8956    445.5394
##    2      exper      addition        0.135       0.134     158.6640    29915.8785    427.8538
##    3       pt        addition        0.182       0.180      44.9450    29807.3688    416.2994
##    4      smsa       addition        0.192       0.191      19.9850    29782.7211    413.6389
##    5      race       addition        0.198       0.196       8.9290    29771.6878    412.3967
## ---------------------------------------------------------------------------------------------
```

```
#fitting the selected model
uswages_reg_final = lm(wage~ educ + exper + pt + smsa + race,data = uswages_reduced)
summary(uswages_reg_final)
```

```
##
## Call:
## lm(formula = wage ~ educ + exper + pt + smsa + race, data = uswages_reduced)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -885.8 -212.9  -56.8  128.9 7499.3
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -243.4879    50.8767  -4.786 1.83e-06 ***
## educ          48.6616     3.2478  14.983  < 2e-16 ***
## exper          9.0798     0.7259  12.509  < 2e-16 ***
## pt          -336.9503    31.9420 -10.549  < 2e-16 ***
## smsa         115.5466    21.5772   5.355 9.54e-08 ***
## race        -124.9292    34.6004  -3.611 0.000313 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 412.4 on 1994 degrees of freedom
## Multiple R-squared:  0.1977, Adjusted R-squared:  0.1957
## F-statistic: 98.26 on 5 and 1994 DF,  p-value: < 2.2e-16
```

From the summary statistics above, we see that all variables in the model are now significant. We also notice a low R-squared value of 0.1977 and Adjusted R-squared of 0.1957

Regression function:

yhat = -243.4879 + 48.6616(educ) + 9.0798(exper) - 336.9503(pt) + 115.5466(smsa) - 124.9292(race)

Now let's check to see if all the linear regression assumptions are met.

Checking for Multicollinearity

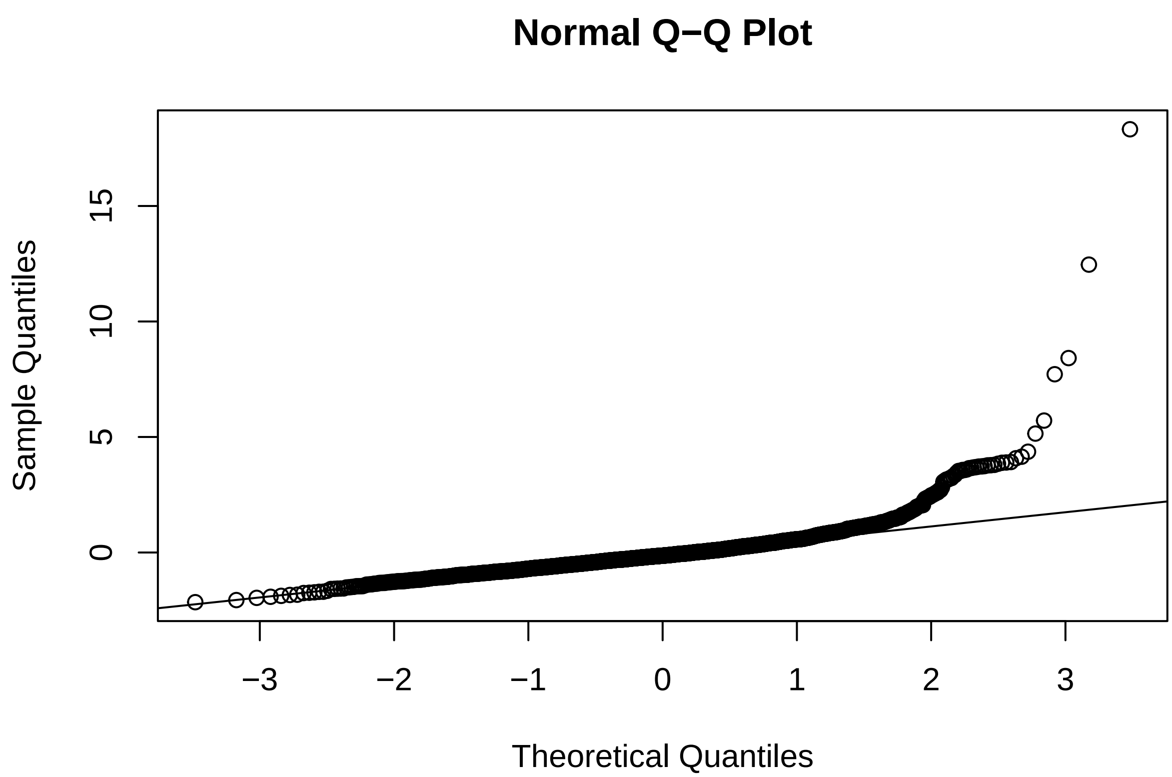
```
vif(uswages_reg_final)
```

```
##     educ    exper       pt     smsa     race
## 1.118935 1.107952 1.007190 1.009953 1.012483
```

All VIFs are below 10. There is no multicolinearity in the data

Checking for normality assumption

```
#Obtaining the standardized residuals
stdres = rstandard(uswages_reg_final)
#Normal probability plot of the standardized residuals
qqnorm(stdres)
qqline(stdres)
```



The QQ plot above shows a heavy upper tail. Which means that the model could be violating the normality assumption. Let's confirm through the Shapiro-Wilk test

```
shapiro.test(uswages_reg_final$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  uswages_reg_final$residuals
## W = 0.71014, p-value < 2.2e-16
```

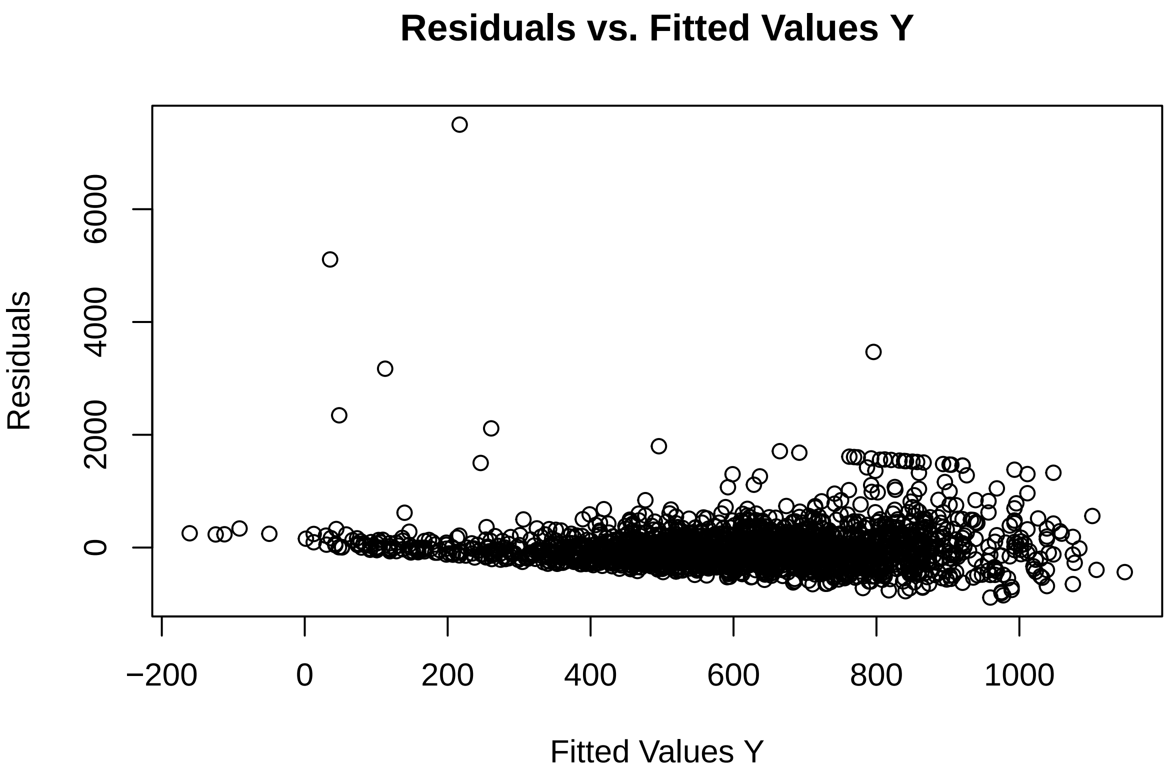
Ho: residuals are normally distributed

Ha: residuals are not normally distributed

The p-value < 2.2e-16, which signifies that we should reject the null hypothesis and conclude that the residuals are not normally distributed. This confirms that the model failed the normality assumption

Let's check for constant variance

```
#obtaining the residual
ei = uswages_reg_final$residuals
Y_hat = uswages_reg_final$fitted.values
#scatter plot of the residuals against fitted values Y
plot(Y_hat,ei,main = "Residuals vs. Fitted Values Y",
     xlab = "Fitted Values Y",ylab = "Residuals")
```



The plot above shows that the error term is not constant. We also notice some outliers. The plot also shows that the relationship is linear

```
#conducting Brausch-Pagan test to confirm
bptest(uswages_reg_final, studentize = FALSE)
```

```
##
##  Breusch-Pagan test
##
## data:  uswages_reg_final
## BP = 1010.4, df = 5, p-value < 2.2e-16
```

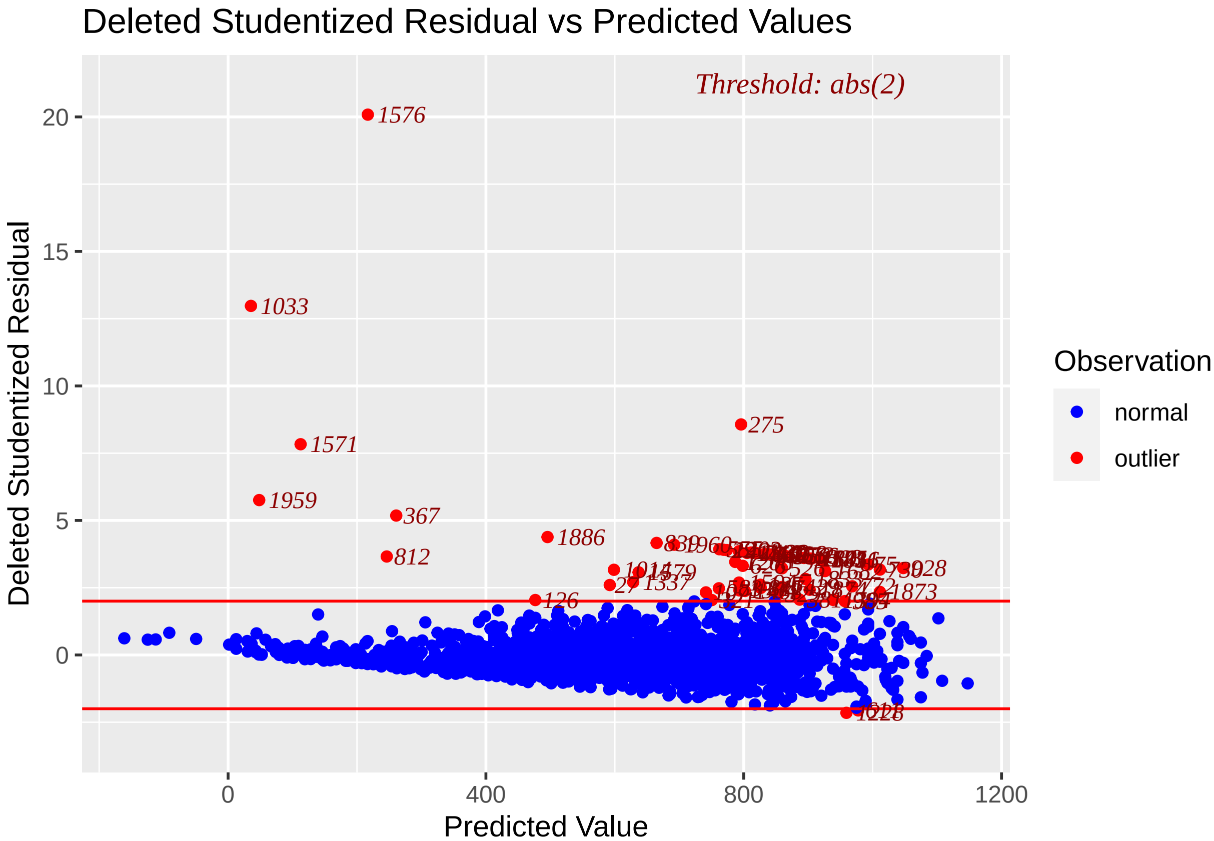
Ho: Error variance is constant

Ha: Error variance is not constant

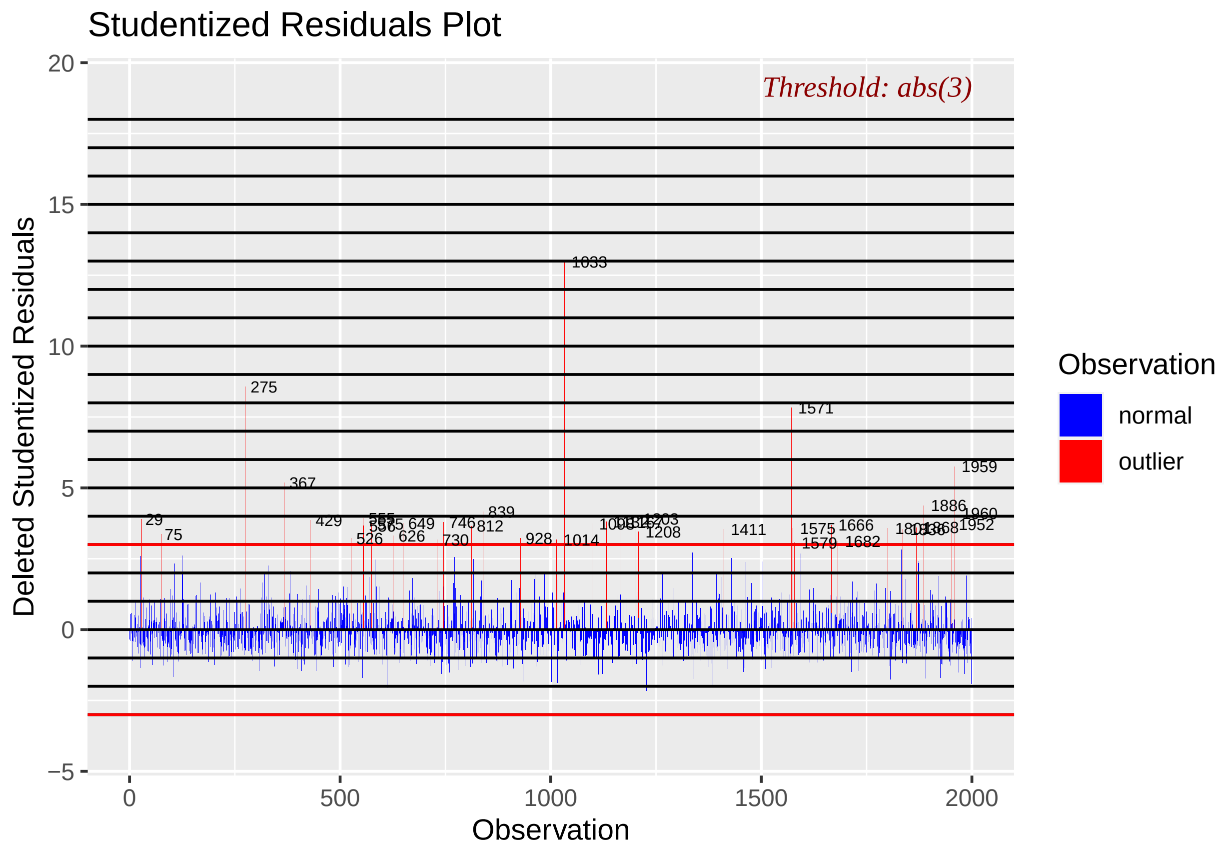
From the results above, we see that the p-value (< 2.2e-16) is significant (i.e. less than 0.05). So we reject the null hypothesis and conclude that error variance is not constant. Therefore the model also violates the constant variance assumption.

Let's investigate the outliers

```
#Checking for outliers
ols_plot_resid_stud_fit(uswages_reg_final)
```




```
ols_plot_resid_stud(uswages_reg_final)
```



From the plots above, we notice a lot of outlying observations.

Let's try to improve the R-square of our model by transforming the data. To determine type of transformatio needed, we use Box-Cox

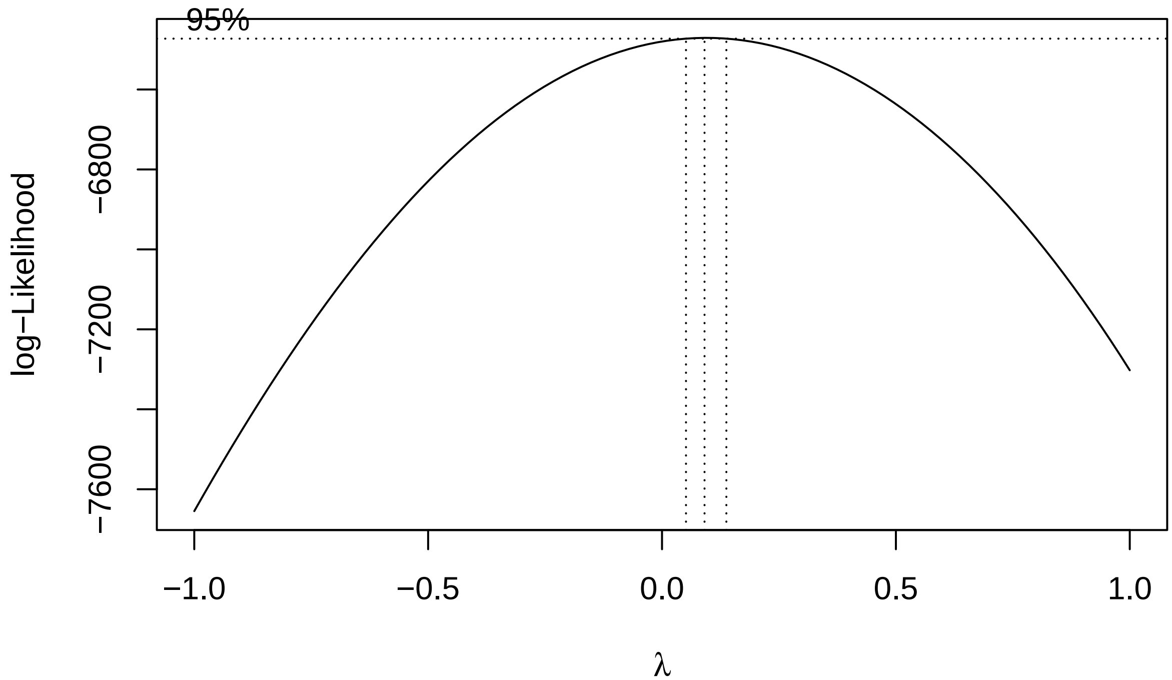
```
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:olsrr':
##
##     cement
```

```
par(mfrow=c(1,1))

boxcox(uswages_reg_final,lambda=seq(-1,1,by=.1))
```



The Box Cox suggest lambda close to zero, which means a log transformation of the outcome variable.

```
#fitting the model with a log-scale of the response variable
uswages_reg_trans = lm(log(wage) ~., data = uswages_reduced)
#performing a stepwise both ways variable selection
kstep_both_trans = ols_step_both_p(uswages_reg_trans,pent=0.1,prem=0.05)#,details = TRUE)
kstep_both_trans
```

```
##
##                                 Stepwise Selection Summary
## -----------------------------------------------------------------------------------------
##                       Added/                    Adj.
## Step    Variable     Removed      R-Square    R-Square      C(p)          AIC         RMSE
## -----------------------------------------------------------------------------------------
##    1        pt       addition        0.207       0.207    569.8570    3944.9102    0.6481
##    2       educ      addition        0.288       0.288    310.2140    3731.8044    0.6143
##    3      exper      addition        0.369       0.368     52.5070    3494.6388    0.5788
##    4       smsa      addition        0.378       0.377     22.8200    3465.3986    0.5745
##    5       race      addition        0.384       0.383      5.4230    3448.0318    0.5718
## -----------------------------------------------------------------------------------------
```

```
#refitting the selected model
uswages_reg_trans_final = lm(log(wage)~ educ + exper + pt + smsa + race,data = uswages_reduced)
summary(uswages_reg_trans_final)
```

```
##
## Call:
## lm(formula = log(wage) ~ educ + exper + pt + smsa + race, data = uswages_reduced)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5319 -0.3330  0.0495  0.3563  3.9435
##
## Coefficients:
```

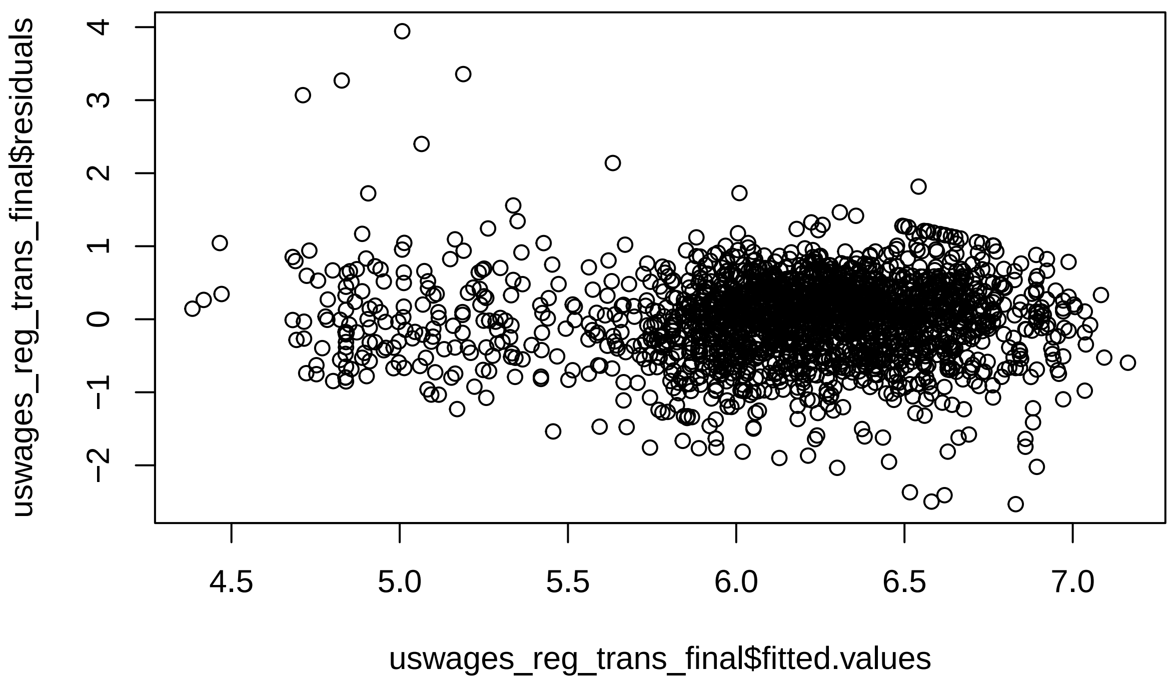
```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.725711   0.070545  66.989  < 2e-16 ***
## educ         0.086566   0.004503  19.223  < 2e-16 ***
## exper        0.016037   0.001006  15.934  < 2e-16 ***
## pt          -1.098583   0.044290 -24.804  < 2e-16 ***
## smsa         0.174543   0.029919   5.834 6.30e-09 ***
## race        -0.211327   0.047976  -4.405 1.11e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5718 on 1994 degrees of freedom
## Multiple R-squared:  0.3843, Adjusted R-squared:  0.3827
## F-statistic: 248.9 on 5 and 1994 DF,  p-value: < 2.2e-16
```

Regression function:

log(yhat) = 4.725711 + 0.086566(educ) + 0.016037(exper) - 1.098583(pt) + 0.174543(smsa) - 0.211327(race)

Comparing the output from the transformed and untransformed model, we see that after transforming the response variable, Adjusted R-squared increased greatly from 0.1957 to 0.3827. We also see from the plot of fitted values vs residuals below that the plot looks more random compared to the previous plot from the untransformed model.

```
plot(uswages_reg_trans_final$fitted.values,uswages_reg_trans_final$residuals)
```



Note that we could check to see if the outliers are still present, we could fit a robust regression model since it is robust to outliers (robust regression methods to consider are huber and bi-square regression)

Finally, from our regression function, we can make predictions. For instance, what would be the expected wage of an individual who has 15 educ, 30 exper, pt 0, smsa 1, and race 1.

From our regression function:

log(yhat) = 4.725711 + 0.086566(educ) + 0.016037(exper) - 1.098583(pt) + 0.174543(smsa) - 0.211327(race)

```r
yhat = exp(4.725711 + (0.086566*15) + (0.016037*30) - (1.098583*0) +  (0.174543*1) - (0.211327*1))
print(yhat)
```

```
## [1] 644.5336
```

Our model predicts a wage of 644.5336

Let's use the predict function in R

```r
#create a dataframe with the new observation
data = data.frame(educ=15,exper=30,pt=0,smsa=1,race=1)
#predict confidence interval
yh = predict(uswages_reg_trans_final,data,se.fit=TRUE, interval = "confidence",
             level = 0.95)
#taking the exp of the fit
fit_yh = exp(c(yh$fit[,1]))
#obtaining the lower limits
lower <- exp(c(yh$fit[,2]))
#obtaining the upper limits
upper <- exp(c(yh$fit[,3]))
print(fit_yh)
```

```
## [1] 644.5374
```

```r
print(lower)
```

```
## [1] 584.5296
```

```r
print(upper)
```

```
## [1] 710.7056
```

Our result is the same. Wage is predicted to be 644.5374 with 95% confidence interval of (584.5296,710.7056), which does not include zero, which means it is significant.