STAT 4310 Final Project: Beauty data

Group F: Elizabeth Bekele, Neha Gyawali, Alison Cheek

Introduction

The name of our data is beauty and it comes from the wooldridge package. The data has 1260 observations/rows and 17 variables. In summary the variables include qualities of an employee. Some of the variables include: if they are female, black, married, living in the city, years of experience and education, the hourly wage, and a rating on their appearance. For an in-depth list of the variables look below:

Variable Name	Meaning
wage	hourly wage
lwage	log(wage)
belavg	1 if looks <= 2
abvavg	1 if looks >= 4
exper	years of work experience
looks	from 1-5
union	1 if union member
goodhlth	1 if good health
black	1 if black
female	1 if female
married	1 if married
south	1 if live in south
bigcity	1 if lives in big city
smllcity	1 if lives in small city
service	1 if in service industry
expersq	exper^2
educ	years of schooling

Data Exploration

```
##    wage    lwage belavg abvavg exper looks union goodhlth black female married
## 1  5.73 1.745715      0      1    30     4     0        1     0      1       1
## 2  4.28 1.453953      0      0    28     3     0        1     0      1       1
## 3  7.96 2.074429      0      1    35     4     0        1     0      1       0
## 4 11.57 2.448416      0      0    38     3     0        1     0      0       1
## 5 11.42 2.435366      0      0    27     3     0        1     0      0       1
## 6  3.91 1.363537      0      0    20     3     0        0     0      1       1
##   south bigcity smllcity service expersq educ
## 1     0       0        1       1     900   14
## 2     1       0        1       0     784   12
## 3     0       0        1       0    1225   10
## 4     0       1        0       1    1444   16
## 5     0       0        1       0     729   16
```

```
## 6     0       1         0       0     400   12
```

```r
#Check if there are NA's
sum(is.na(beauty))
```

```
## [1] 0
```

```r
#Check data type of each variable
str(beauty)
```

```
## 'data.frame':    1260 obs. of  17 variables:
##  $ wage    : num  5.73 4.28 7.96 11.57 11.42 ...
##  $ lwage   : num  1.75 1.45 2.07 2.45 2.44 ...
##  $ belavg  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ abvavg  : int  1 0 1 0 0 0 0 1 0 0 ...
##  $ exper   : int  30 28 35 38 27 20 12 5 5 12 ...
##  $ looks   : int  4 3 4 3 3 3 3 4 3 3 ...
##  $ union   : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ goodhlth: int  1 1 1 1 1 0 1 1 1 1 ...
##  $ black   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ female  : int  1 1 1 0 0 1 0 0 1 1 ...
##  $ married : int  1 1 0 1 1 1 1 0 0 0 ...
##  $ south   : int  0 1 0 0 0 0 0 0 0 0 ...
##  $ bigcity : int  0 0 0 1 0 1 1 0 0 0 ...
##  $ smllcity: int  1 1 1 0 1 0 0 1 0 1 ...
##  $ service : int  1 0 0 1 0 0 0 0 0 0 ...
##  $ expersq : int  900 784 1225 1444 729 400 144 25 25 144 ...
##  $ educ    : int  14 12 10 16 16 12 16 16 16 12 ...
##  - attr(*, "time.stamp")= chr "25 Jun 2011 23:03"
```

Factors that are currently integers:

- belavg (1 if looks <= 2)
- abavag (1 if looks >= 4)
- union
- goodhlth
- black
- female
- married
- south
- bigcity
- smllcity
- service
- looks (from 1 - 5)

```
##    looks_1 looks_2 looks_3 looks_4 looks_5
## 1:       0       0       0       1       0
## 2:       0       0       1       0       0
## 3:       0       0       0       1       0
## 4:       0       0       1       0       0
## 5:       0       0       1       0       0
## 6:       0       0       1       0       0
```

```
## Classes 'data.table' and 'data.frame':   1260 obs. of  21 variables:
##  $ wage    : num  5.73 4.28 7.96 11.57 11.42 ...
##  $ lwage   : num  1.75 1.45 2.07 2.45 2.44 ...
##  $ belavg  : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ abvavg  : Factor w/ 2 levels "0","1": 2 1 2 1 1 1 1 2 1 1 ...
##  $ exper   : int  30 28 35 38 27 20 12 5 5 12 ...
##  $ looks_1 : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ looks_2 : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ looks_3 : Factor w/ 2 levels "0","1": 1 2 1 2 2 2 2 1 2 2 ...
##  $ looks_4 : Factor w/ 2 levels "0","1": 2 1 2 1 1 1 1 2 1 1 ...
##  $ looks_5 : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ union   : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 2 1 1 ...
##  $ goodhlth: Factor w/ 2 levels "0","1": 2 2 2 2 2 1 2 2 2 2 ...
##  $ black   : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ female  : Factor w/ 2 levels "0","1": 2 2 2 1 1 2 1 1 2 2 ...
##  $ married : Factor w/ 2 levels "0","1": 2 2 1 2 2 2 2 1 1 1 ...
##  $ south   : Factor w/ 2 levels "0","1": 1 2 1 1 1 1 1 1 1 1 ...
##  $ bigcity : Factor w/ 2 levels "0","1": 1 1 1 2 1 2 2 1 1 1 ...
##  $ smllcity: Factor w/ 2 levels "0","1": 2 2 2 1 2 1 1 2 1 2 ...
##  $ service : Factor w/ 2 levels "0","1": 2 1 1 2 1 1 1 1 1 1 ...
##  $ expersq : int  900 784 1225 1444 729 400 144 25 25 144 ...
##  $ educ    : int  14 12 10 16 16 12 16 16 16 12 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

Simple Linear Regression Model

Model 1

To start off, we will make a linear regression model excluding predictors that are calculated based on other predictors. We are removing expersq, lwage, and the 5 columns of looks because we are keeping exper, wage, belavg, and abvag.

The response we are interested in is wage. We want to know if certain predictors can determine an employee's hourly wage.

```r
attach(beauty)

mod.all <- lm(wage ~. -expersq -lwage -looks_1 -looks_2 -looks_3 -looks_4 -looks_5, beauty)
summary(mod.all)
```

```
##
## Call:
## lm(formula = wage ~ . - expersq - lwage - looks_1 - looks_2 -
##     looks_3 - looks_4 - looks_5, data = beauty)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.541 -2.133 -0.541  1.186 71.907
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.94468    0.85395  -1.106  0.26883
```

```
## belavg1     -0.77351    0.36973  -2.092  0.03663 *
## abvavg1      0.17226    0.26768   0.644  0.52000
## exper        0.07765    0.01068   7.271 6.30e-13 ***
## union1       0.58565    0.26792   2.186  0.02901 *
## goodhlth1   -0.02245    0.47593  -0.047  0.96238
## black1      -0.13452    0.46191  -0.291  0.77093
## female1     -2.12282    0.27652  -7.677 3.28e-14 ***
## married1     0.80987    0.27454   2.950  0.00324 **
## south1       0.37575    0.31211   1.204  0.22886
## bigcity1     1.70265    0.33668   5.057 4.89e-07 ***
## smllcity1    0.55932    0.27445   2.038  0.04176 *
## service1    -0.47562    0.28837  -1.649  0.09933 .
## educ         0.42641    0.05007   8.516  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.138 on 1246 degrees of freedom
## Multiple R-squared:  0.2199, Adjusted R-squared:  0.2118
## F-statistic: 27.02 on 13 and 1246 DF,  p-value: < 2.2e-16
```

The starting R^2 is 21.99% and the residual standard error, s, 4.138.

The low R^2 is due to extra predictors that are not significant in explaining the hourly wage of an employee. Predictors we want to remove include: abvavg, goodhlth, black, south, and service. These predictors have p-values that are above 0.05 and do not contribute significantly to explaining the model.

Residual Analysis

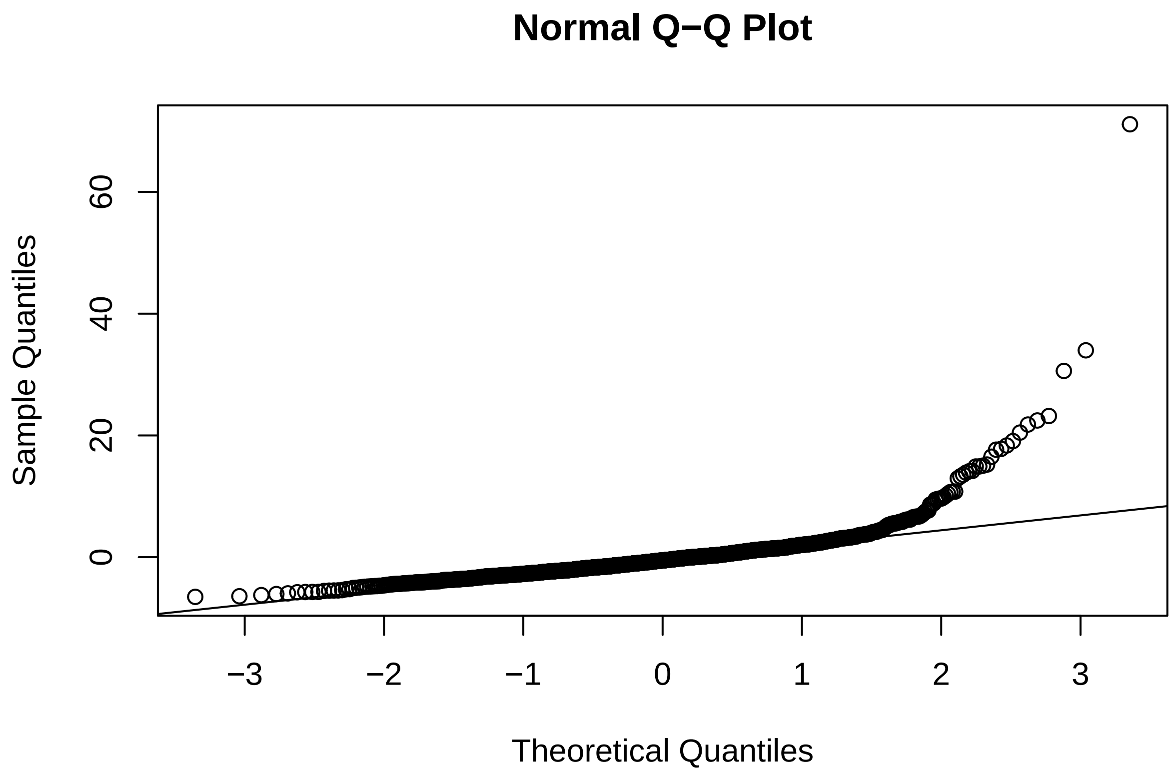
First we calculate the Hat matrix, and calculate the model residuals using Hat matrix. Then we adjust the residual using the Hat matrix to account for influential points.

Residual Assumptions

1) Normally Distributed

We check for normality

```
qqnorm(residuals)
qqline(residuals)
```



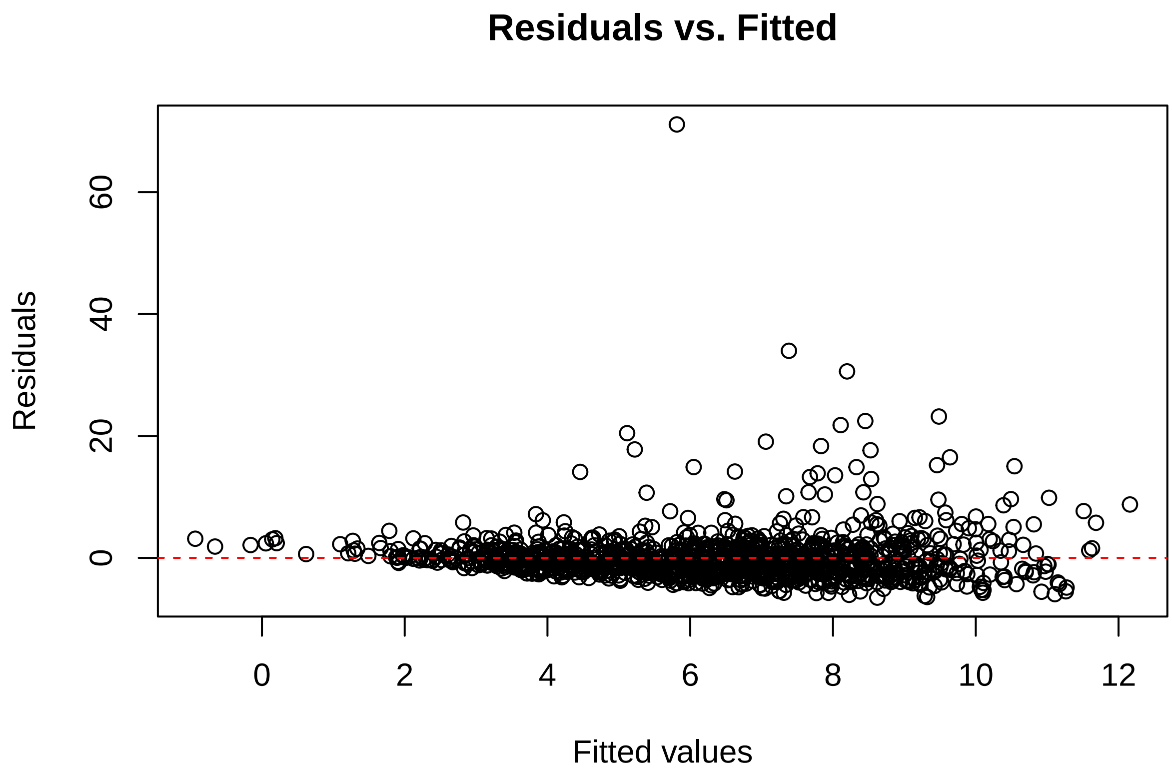

```
shapiro.test(residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals
## W = 0.64748, p-value < 2.2e-16
```

Using both tests, we find that the data is not normally distributed.

2) Checking for Zero Mean and Constant Variance

```
plot(mod.all$fitted.values, residuals,
xlab = "Fitted values",
ylab = "Residuals",
main = "Residuals vs. Fitted")
abline(h = 0, lty = 2, col = "red")
```



BP-Test for heteroskedasticity

```
library(lmtest)
bptest(mod.all)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  mod.all
## BP = 22.746, df = 13, p-value = 0.04482
```

In this case, the p-value of the test is 0.04482, which is less than the significance level of 0.05. There is heteroscedasticity in the residuals of the model. This means that the variance of the errors is not constant across the range of the predicted values, violating one of the assumptions of linear regression.

3) Independence

```
dwtest(mod.all, alternative = "two.sided")
```

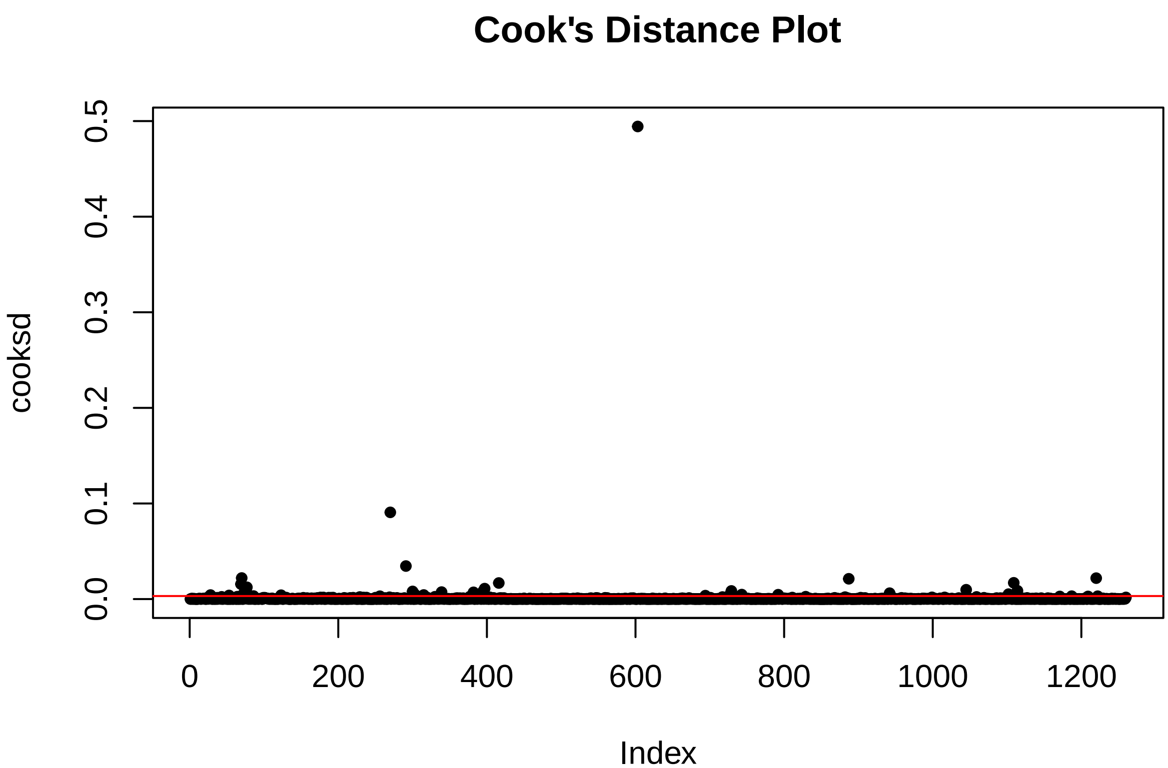
```
##
##  Durbin-Watson test
##
## data:  mod.all
## DW = 1.9044, p-value = 0.08147
## alternative hypothesis: true autocorrelation is not 0
```

The test statistic is the Durbin-Watson (DW) statistic, which ranges from 0 to 4. A value of 2 indicates no autocorrelation, values below 2 indicate positive autocorrelation, and values above 2 indicate negative autocorrelation.

In this case, the DW statistic is 1.9044, which is less than 2 but greater than 0, indicating that there may be some positive autocorrelation present in the residuals. However, the p-value of 0.08147 is greater than the typical significance level of 0.05, so we do not have enough evidence to reject the null hypothesis of no autocorrelation. Therefore, we cannot conclude that there is significant autocorrelation in the residuals.

We now caclulate cooks distance and take out any outliers.

```
cooksd <- cooks.distance(mod.all)
plot(cooksd, pch = 20, main = "Cook's Distance Plot")
abline(h = 4/length(cooksd), col = "red")
```




```
outliers <- order(cooksd, decreasing = TRUE)[1:3]
outliers
```

```
## [1] 603 270 291
```

Residual Analysis

```
beauty_clean <- beauty[-outliers,]
res_clean <- lm(wage ~. -expersq -lwage -looks_1 -looks_2 -looks_3 -looks_4 -looks_5, data = beauty_clea
summary(res_clean)
```

```
##
## Call:
## lm(formula = wage ~ . - expersq - lwage - looks_1 - looks_2 -
##     looks_3 - looks_4 - looks_5, data = beauty_clean)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.614 -1.942 -0.465  1.230 30.549
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.160152   0.702347  -1.652  0.09882 .
## belavg1     -0.802836   0.303331  -2.647  0.00823 **
## abvavg1     -0.123514   0.220067  -0.561  0.57472
## exper        0.082477   0.008766   9.408  < 2e-16 ***
## union1       0.419666   0.220141   1.906  0.05684 .
## goodhlth1    0.668170   0.395196   1.691  0.09114 .
## black1      -0.867813   0.380570  -2.280  0.02276 *
## female1     -2.184820   0.227001  -9.625  < 2e-16 ***
## married1     0.521271   0.225529   2.311  0.02098 *
## south1       0.515376   0.256126   2.012  0.04442 *
## bigcity1     1.452681   0.276402   5.256 1.73e-07 ***
## smllcity1    0.413818   0.225407   1.836  0.06662 .
## service1    -0.636935   0.236785  -2.690  0.00724 **
## educ         0.421210   0.041083  10.253  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.394 on 1243 degrees of freedom
## Multiple R-squared:  0.2969, Adjusted R-squared:  0.2895
## F-statistic: 40.38 on 13 and 1243 DF,  p-value: < 2.2e-16
```

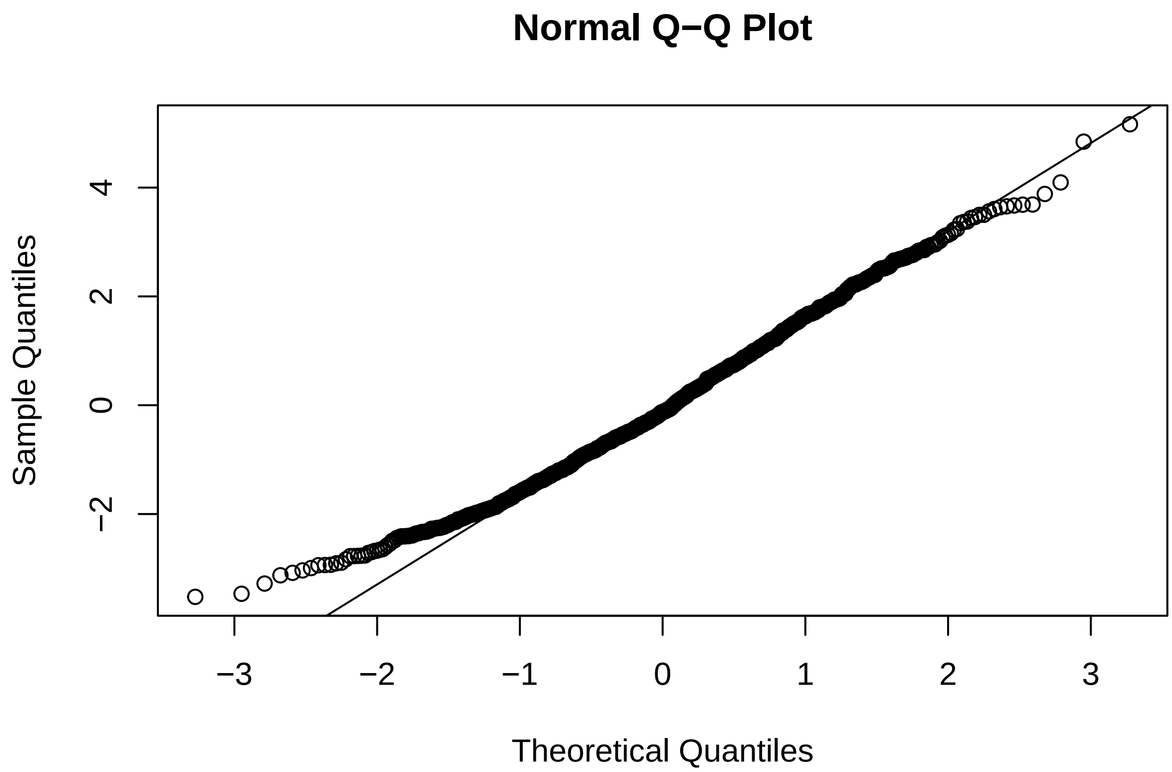
```
summary(cooksd)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0000000 0.0000257 0.0001185 0.0009785 0.0003762 0.4943687
```

Factoring in our findings from our cooks distance analysis, we create a new model.

```
beauty_clean <- beauty[cooksd < 0.0003762, ]
model.new <- lm(wage ~. -expersq -lwage -looks_1 -looks_2 -looks_3 -looks_4 -looks_5, data = beauty_clea

qqnorm(model.new$residuals)
qqline(model.new$residuals)
```




```
summary(mod.all)
```

```
##
## Call:
## lm(formula = wage ~ . - expersq - lwage - looks_1 - looks_2 -
##     looks_3 - looks_4 - looks_5, data = beauty)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.541 -2.133 -0.541  1.186 71.907
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.94468    0.85395  -1.106  0.26883
## belavg1     -0.77351    0.36973  -2.092  0.03663 *
## abvavg1      0.17226    0.26768   0.644  0.52000
## exper        0.07765    0.01068   7.271 6.30e-13 ***
## union1       0.58565    0.26792   2.186  0.02901 *
## goodhlth1   -0.02245    0.47593  -0.047  0.96238
## black1      -0.13452    0.46191  -0.291  0.77093
## female1     -2.12282    0.27652  -7.677 3.28e-14 ***
## married1     0.80987    0.27454   2.950  0.00324 **
## south1       0.37575    0.31211   1.204  0.22886
## bigcity1     1.70265    0.33668   5.057 4.89e-07 ***
## smllcity1    0.55932    0.27445   2.038  0.04176 *
```

```
## service1    -0.47562    0.28837  -1.649  0.09933 .
## educ         0.42641    0.05007   8.516  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.138 on 1246 degrees of freedom
## Multiple R-squared:  0.2199, Adjusted R-squared:  0.2118
## F-statistic: 27.02 on 13 and 1246 DF,  p-value: < 2.2e-16
```

```
summary(model.new)
```

```
##
## Call:
## lm(formula = wage ~ . - expersq - lwage - looks_1 - looks_2 -
##     looks_3 - looks_4 - looks_5, data = beauty_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5221 -1.1430 -0.1272  1.0457  5.1643
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.405644   0.405761  -3.464 0.000556 ***
## belavg1     -0.600605   0.157362  -3.817 0.000144 ***
## abvavg1      0.146868   0.115066   1.276 0.202140
## exper        0.083569   0.004865  17.177  < 2e-16 ***
## union1       0.939246   0.115290   8.147 1.20e-15 ***
## goodhlth1    0.386291   0.252469   1.530 0.126343
## black1      -0.028699   0.235632  -0.122 0.903086
## female1     -1.861660   0.117074 -15.902  < 2e-16 ***
## married1     0.387701   0.116498   3.328 0.000909 ***
## south1       0.217283   0.140874   1.542 0.123320
## bigcity1     1.532425   0.150127  10.208  < 2e-16 ***
## smllcity1    0.614330   0.115637   5.313 1.35e-07 ***
## service1    -0.399684   0.126266  -3.165 0.001599 **
## educ         0.396386   0.022763  17.414  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.53 on 930 degrees of freedom
## Multiple R-squared:  0.624,  Adjusted R-squared:  0.6188
## F-statistic: 118.7 on 13 and 930 DF,  p-value: < 2.2e-16
```

We find that the new model is a better fit for the data.

The original model has a larger residual standard error (4.138) and a lower Multiple R-squared (0.2199) compared to the new model with a smaller residual standard error (1.53) and a higher Multiple R-squared (0.624).

The F-statistic in the new model (118.7) is higher than the F-statistic in the original model (27.02), suggesting that the new model is a better fit than the original model

Model 2

Can we predict the hourly wage of an employee based on the following significant predictors:

- belavg
- exper
- union
- female
- married
- bigcity
- smllcity
- service
- educ

```
mod2 <- lm(wage ~ exper + female + bigcity + educ +
             married + belavg + union + smllcity, beauty_clean)
summary(mod2)
```

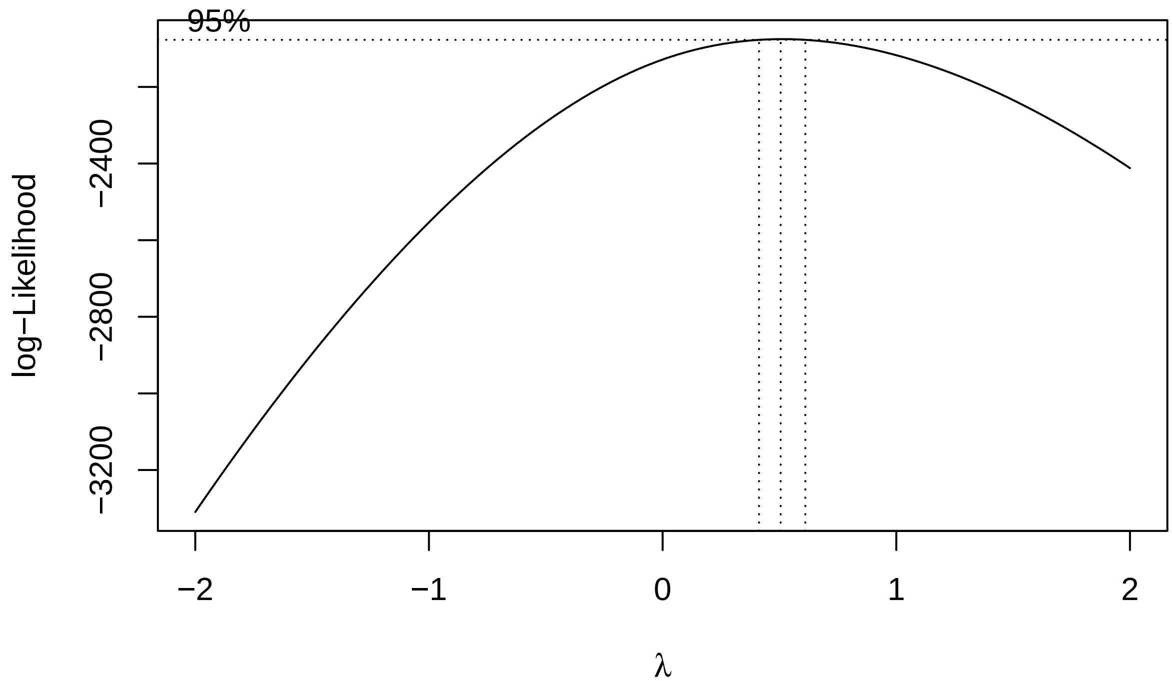
```
##
## Call:
## lm(formula = wage ~ exper + female + bigcity + educ + married +
##     belavg + union + smllcity, data = beauty_clean)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.532 -1.148 -0.139  1.060  5.134
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.848148   0.325369  -2.607  0.00929 **
## exper        0.081544   0.004834  16.868  < 2e-16 ***
## female1     -1.988797   0.111150 -17.893  < 2e-16 ***
## bigcity1     1.562387   0.149257  10.468  < 2e-16 ***
## educ         0.385126   0.021782  17.681  < 2e-16 ***
## married1     0.377231   0.116279   3.244  0.00122 **
## belavg1     -0.657821   0.153074  -4.297 1.91e-05 ***
## union1       0.951316   0.114925   8.278 4.31e-16 ***
## smllcity1    0.645260   0.114754   5.623 2.48e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.539 on 935 degrees of freedom
## Multiple R-squared:  0.6175, Adjusted R-squared:  0.6142
## F-statistic: 188.7 on 8 and 935 DF,  p-value: < 2.2e-16
```

- R^2 slightly decreased to 61.75%
- s slightly increased to 1.539

Model Transformation - BoxCox

Let's try and increase our R^2 and decrease our R_{SE}

```
library(MASS)
bc <- boxcox(mod2, lambda = seq(-2.0, 2.0, 1), plotit = T)
```



Based on the BoxCox transformation our eigenvalue is 0.5 so, we will need to have a square root transformation of our response variable, wage.

Model 3: Using Transformation

```
mod3 <- lm(sqrt(wage) ~ exper + female + bigcity + educ +
             married + belavg + union + smllcity, beauty_clean)
summary(mod3)
```

```
##
## Call:
## lm(formula = sqrt(wage) ~ exper + female + bigcity + educ + married +
##     belavg + union + smllcity, data = beauty_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -0.94836 -0.23722 -0.01169  0.24090  0.95141
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.935239   0.069512  13.454  < 2e-16 ***
## exper        0.017166   0.001033  16.621  < 2e-16 ***
## female1     -0.454626   0.023746 -19.145  < 2e-16 ***
## bigcity1     0.324483   0.031887  10.176  < 2e-16 ***
## educ         0.083553   0.004653  17.955  < 2e-16 ***
## married1     0.069613   0.024842   2.802  0.00518 **
## belavg1     -0.140794   0.032703  -4.305 1.84e-05 ***
## union1       0.216177   0.024553   8.805  < 2e-16 ***
## smllcity1    0.133072   0.024516   5.428 7.27e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3288 on 935 degrees of freedom
## Multiple R-squared:  0.6272, Adjusted R-squared:  0.624
## F-statistic: 196.6 on 8 and 935 DF,  p-value: < 2.2e-16
```

- Our R^2 increased to 62.72%
- s decreased to 0.3288

Step(): Reduced Model

We will create a reduced model using the step() function on the transformed model.

```
#Reduced Model
mod.red <- step(mod3)
```

```
## Start:  AIC=-2090.85
## sqrt(wage) ~ exper + female + bigcity + educ + married + belavg +
##     union + smllcity
##
##            Df Sum of Sq    RSS     AIC
## <none>                  101.11 -2090.8
## - married   1     0.849 101.95 -2085.0
## - belavg    1     2.004 103.11 -2074.3
## - smllcity  1     3.186 104.29 -2063.6
## - union     1     8.383 109.49 -2017.7
## - bigcity   1    11.197 112.30 -1993.7
## - exper     1    29.873 130.98 -1848.5
## - educ      1    34.861 135.97 -1813.2
## - female    1    39.637 140.74 -1780.6
```

The predictors we have are currently the best that can predict the hourly wage. No, predictors were removed. The AIC is -2090.85.

Model 4: Using Reduced Model

```
mod4 <- summary(mod.red)
mod4
```

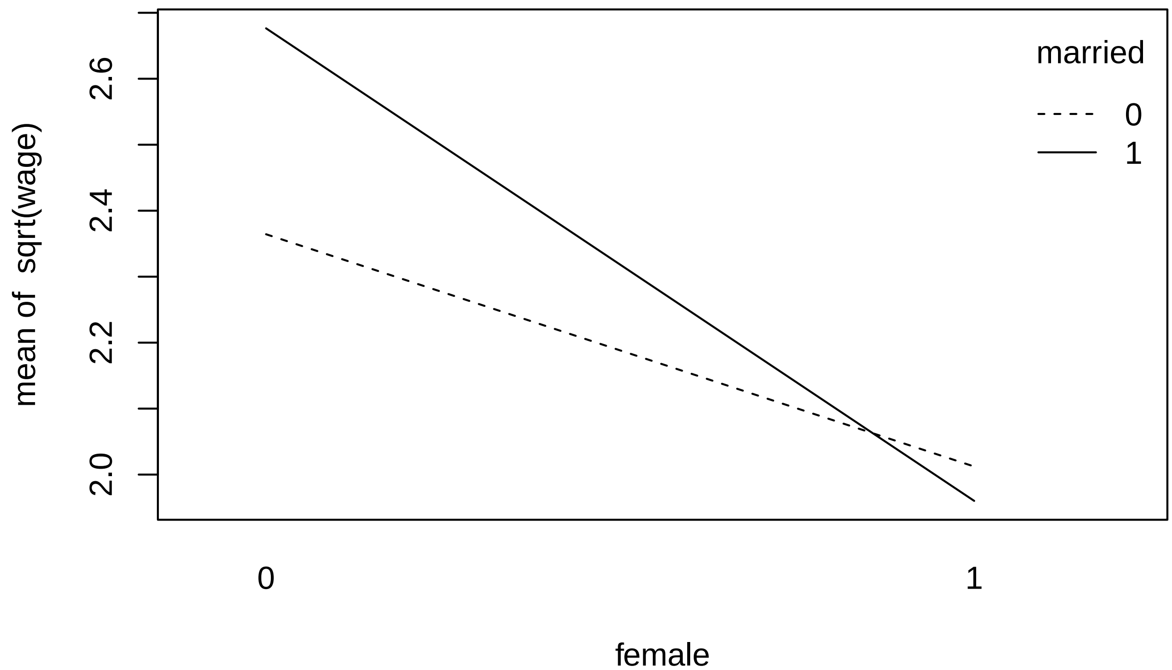
```
##
## Call:
## lm(formula = sqrt(wage) ~ exper + female + bigcity + educ + married +
##     belavg + union + smllcity, data = beauty_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.94836 -0.23722 -0.01169  0.24090  0.95141
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.935239   0.069512  13.454  < 2e-16 ***
## exper        0.017166   0.001033  16.621  < 2e-16 ***
## female1     -0.454626   0.023746 -19.145  < 2e-16 ***
## bigcity1     0.324483   0.031887  10.176  < 2e-16 ***
## educ         0.083553   0.004653  17.955  < 2e-16 ***
## married1     0.069613   0.024842   2.802  0.00518 **
## belavg1     -0.140794   0.032703  -4.305 1.84e-05 ***
## union1       0.216177   0.024553   8.805  < 2e-16 ***
## smllcity1    0.133072   0.024516   5.428 7.27e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3288 on 935 degrees of freedom
## Multiple R-squared:  0.6272, Adjusted R-squared:  0.624
## F-statistic: 196.6 on 8 and 935 DF,  p-value: < 2.2e-16
```

The summary of the reduced model is identical to the summary of the transformed model. Although, the R^2 is not above 70% it's the best it can do with the given predictors.

Model 5: Interaction

As a last attempt, to increase the R^2 and s we will try to include an interaction term between 2 predictors. We are looking at the interaction between the predictors female and married against the square root of hourly wage.

```
#Interaction Plot
interaction.plot(female, married, sqrt(wage))
```



Based on the interaction plot, we can see that the lines intersect. This indicates an interaction.

```
#Model with interaction
mod5 <- lm(sqrt(wage) ~ exper + female + bigcity + educ +
             married + belavg + union + smllcity + married*female, beauty_clean)
summary(mod5)
```

```
##
## Call:
## lm(formula = sqrt(wage) ~ exper + female + bigcity + educ + married +
##     belavg + union + smllcity + married * female, data = beauty_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.91671 -0.22844 -0.00635  0.24576  0.93761
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.899112   0.070727  12.712  < 2e-16 ***
## exper            0.016799   0.001040  16.158  < 2e-16 ***
## female1         -0.375247   0.039003  -9.621  < 2e-16 ***
## bigcity1         0.322317   0.031804  10.134  < 2e-16 ***
## educ             0.083045   0.004644  17.883  < 2e-16 ***
## married1         0.132288   0.034819   3.799 0.000154 ***
## belavg1         -0.136388   0.032651  -4.177 3.23e-05 ***
## union1           0.217194   0.024483   8.871  < 2e-16 ***
```

```
## smllcity1         0.131186   0.024454   5.364 1.02e-07 ***
## female1:married1 -0.125045   0.048826  -2.561 0.010592 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3279 on 934 degrees of freedom
## Multiple R-squared:  0.6298, Adjusted R-squared:  0.6262
## F-statistic: 176.5 on 9 and 934 DF,  p-value: < 2.2e-16
```

After including the interaction term, we achieve the highest R^2 at 62.98% and a low residual standard error of 0.3279.

```
anova(mod3, mod5)
```

```
## Analysis of Variance Table
##
## Model 1: sqrt(wage) ~ exper + female + bigcity + educ + married + belavg +
##     union + smllcity
## Model 2: sqrt(wage) ~ exper + female + bigcity + educ + married + belavg +
##     union + smllcity + married * female
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    935 101.11
## 2    934 100.40  1   0.70505 6.5589 0.01059 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the ANOVA model, we are able to see that the interaction between an employee's marital status and gender is significant in predicting the square root of their hourly wage. The P-value is 0.01059 which is below a significance level of 5%. The interaction term is needed in the linear regression model.

Conclusion

After removing outliers and conducting 5 different models, we were able to achieve the highest R^2 at 62.98% by including an interaction term between the predictors married and female. This model used the square root of wage as a response based on our findings in the boxcox transformation. The remaining 8 predictors used in the model include: experience, female, big city, education, married, below average, union, and small city. The predictors in this model are able to explain the response, sqrt(wage), more accurately with this interaction term included. We suggest using other factors that impact an employee's hourly wage in order to achieve a higher R^2.