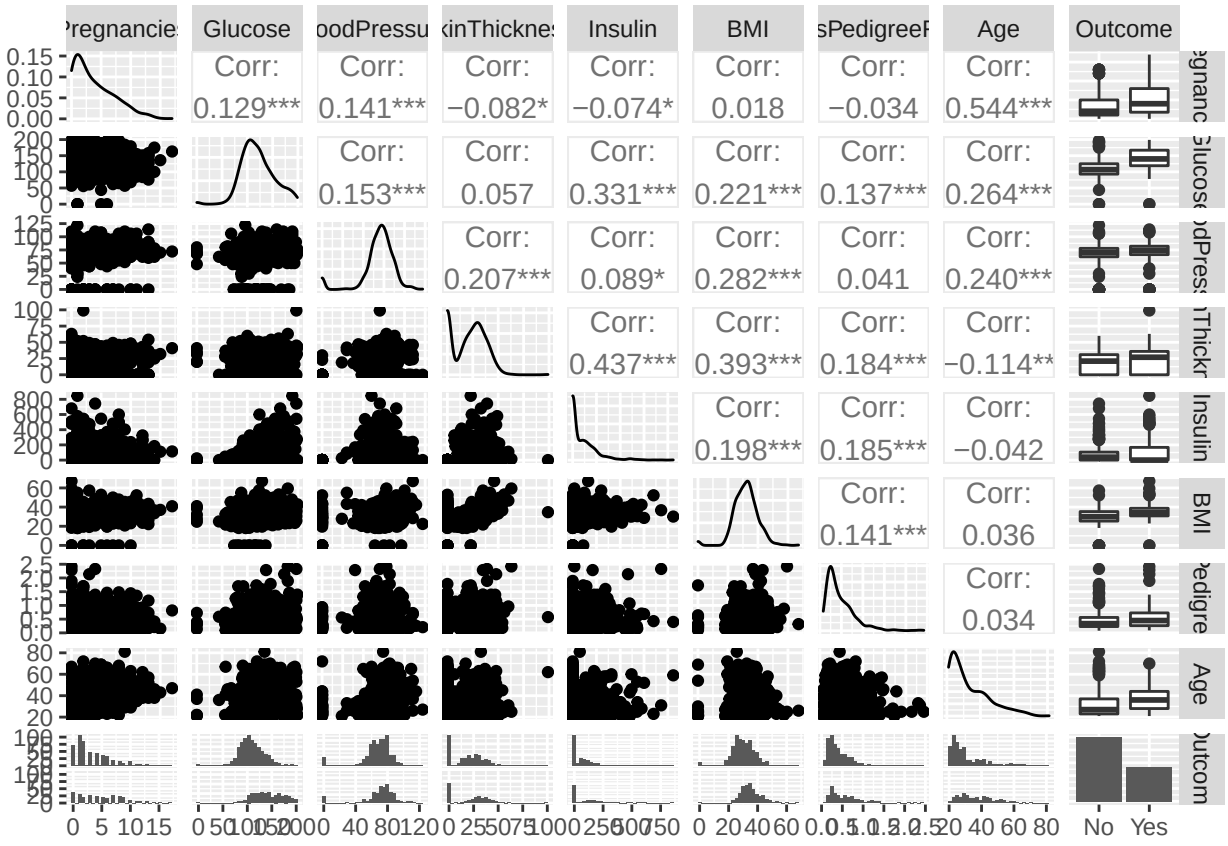
# Multiple Linear Regression Runs

Uyen Nguyen

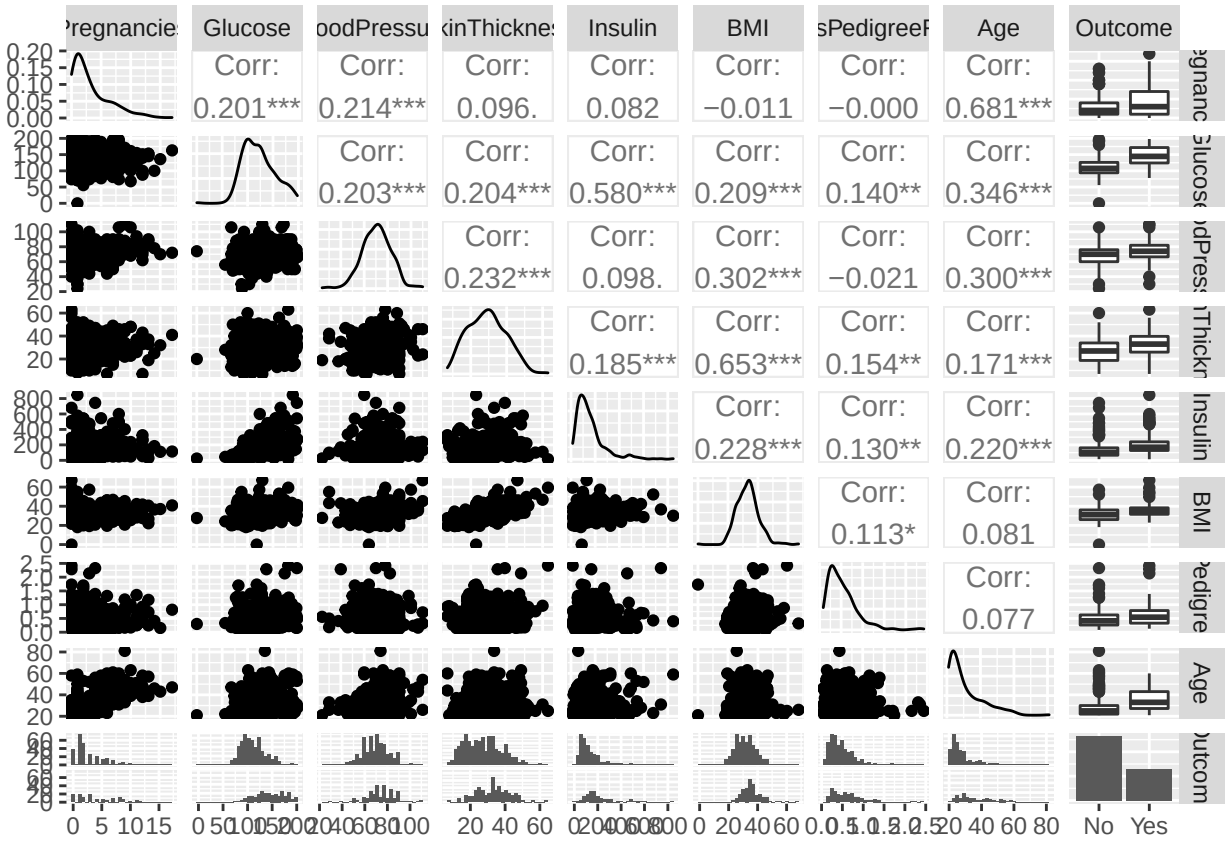
The data cleaning will ensure ggplot will color correctly for variable Outcome

```
# Scatter plot using GGpairs
GGally::ggpairs(diabetes)
```



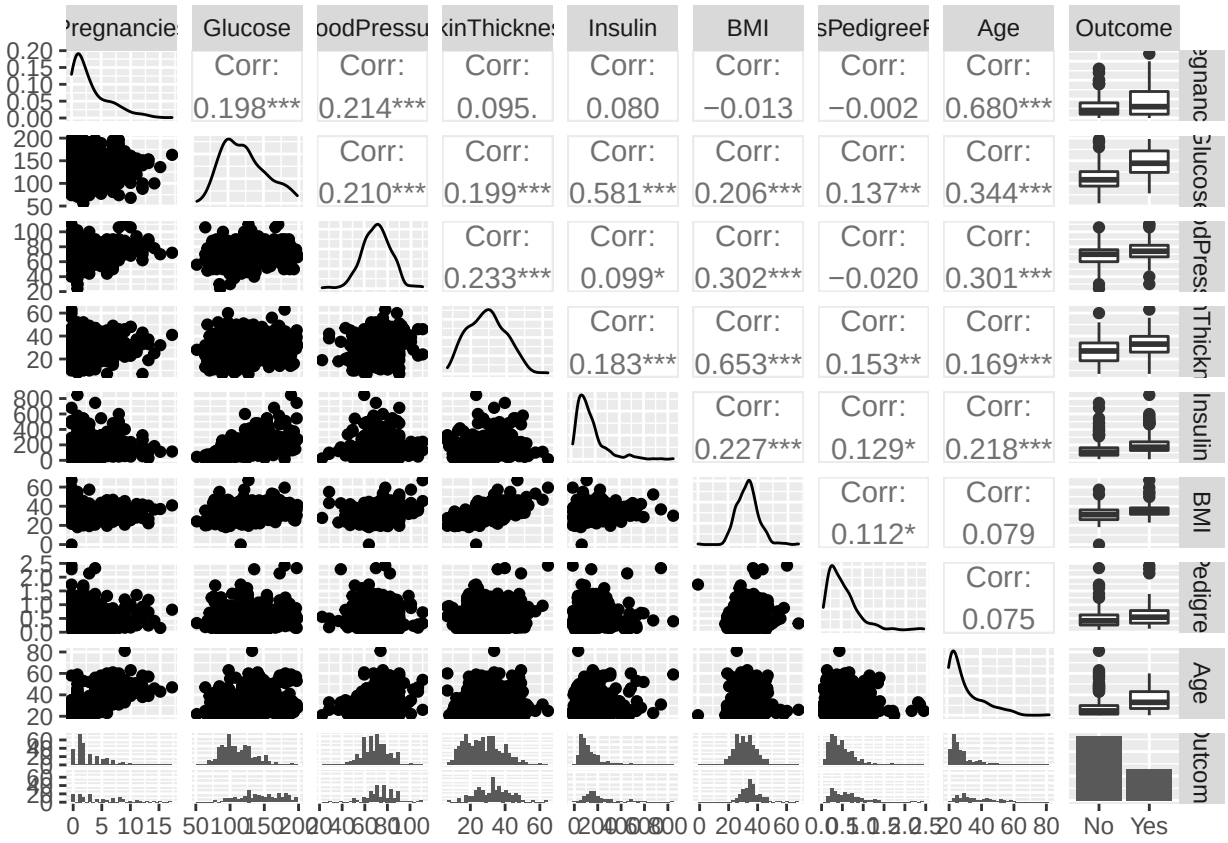

There were a lot of 0s in the plots and they might hurt the correlation so I removed them by Insulin then Glucose.

```
# Filter out 0 values in Insulin and plot pairs
noNullIns <- diabetes %>% filter(Insulin != 0)
GGally::ggpairs(noNullIns)
```



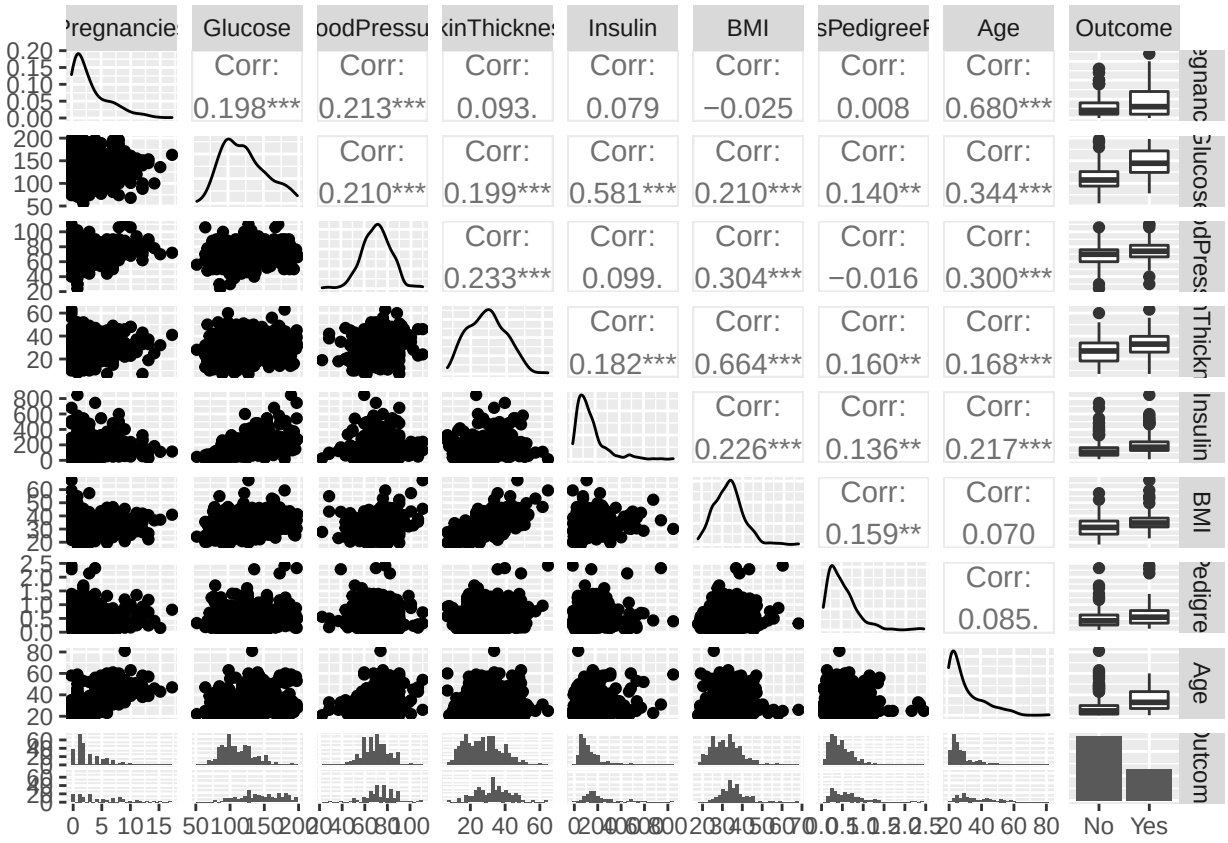

Correlation and graphs look better after taking out 0s from Insulin!

```
# Filter out 0 value in Glucose and plot pairs
noNullInsGlu <- noNullIns %>% filter(Glucose != 0)
GGally::ggpairs(noNullInsGlu)
```



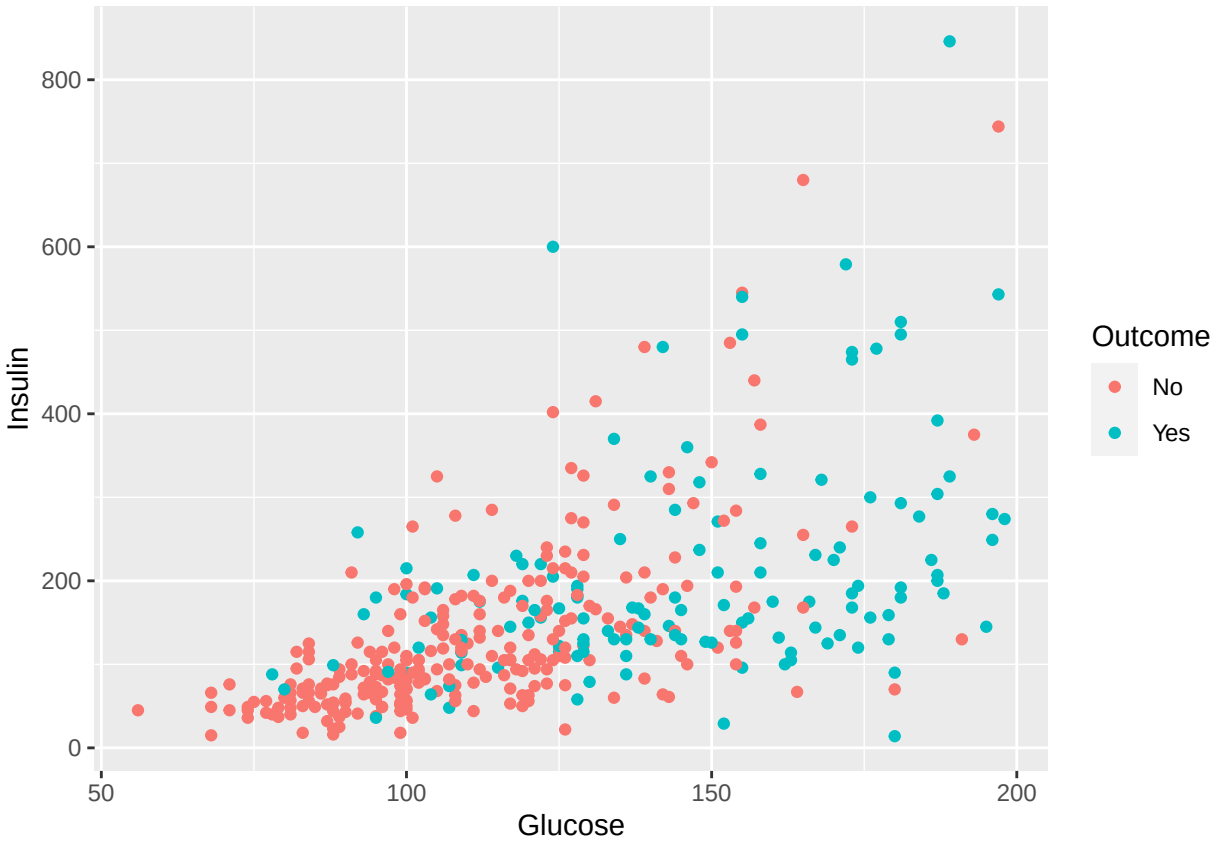

Some correlations dropped after taking the 0 from Glucose, but graphs look better.

```
# Filter out 0 value in Glucose and plot pairs
noNullInsGluBMI <- noNullInsGlu %>% filter(BMI != 0)
GGally::ggpairs(noNullInsGluBMI)
```




I noticed some 0s in BMI so I removed them as well.

```
# Scatterplot of Glucose and Insulin with Outcome
ggplot(noNullInsGluBMI, aes(Glucose, Insulin, color = Outcome)) +
  geom_point()
```




The variance was non-constant so Idk.

## First run for MLR

```
# Full model and summary
result <- lm(Glucose ~ Pregnancies + BloodPressure + SkinThickness + Insulin + BMI + DiabetesPedigreeFu
summary(result)
```

```
##
## Call:
## lm(formula = Glucose ~ Pregnancies + BloodPressure + SkinThickness +
##     Insulin + BMI + DiabetesPedigreeFunction + Age, data = noNullInsGluBMI)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -71.185 -15.558  -3.087  11.847  74.331
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              60.03002    8.44119   7.112 5.65e-12 ***
## Pregnancies               0.07383    0.52315   0.141 0.887848
## BloodPressure             0.21341    0.10769   1.982 0.048219 *
## SkinThickness             0.07433    0.15769   0.471 0.637628
```

```
## Insulin                   0.13321    0.01084  12.293  < 2e-16 ***
## BMI                       0.13038    0.24389   0.535 0.593239
## DiabetesPedigreeFunction  4.17855    3.62412   1.153 0.249635
## Age                       0.57734    0.17182   3.360 0.000857 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.1 on 384 degrees of freedom
## Multiple R-squared:  0.4012, Adjusted R-squared:  0.3903
## F-statistic: 36.76 on 7 and 384 DF,  p-value: < 2.2e-16
```

Pregnancies, skin thickness, BMI, and diabetes pedigree function didn't look significant based on p-values in the presence of the other predictors. Partial F test will be conducted to see if we can drop these variables.

```
# Checking for multicollinearity
faraway::vif(result)
```

```
##              Pregnancies            BloodPressure            SkinThickness
##                 1.900621                 1.219344                 1.851701
##                  Insulin                      BMI DiabetesPedigreeFunction
##                 1.116662                 1.978124                 1.055661
##                      Age
##                 2.068613
```

Checking for multicollinearity signs in this model and it looks fine. Everything is definitely under 5 so we're in the clear.

```
reduced <- lm(Glucose ~ BloodPressure + Insulin + Age, data=noNullInsGluBMI)
summary(reduced)
```

```
##
## Call:
## lm(formula = Glucose ~ BloodPressure + Insulin + Age, data = noNullInsGluBMI)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -73.576 -15.015  -3.763  12.144  78.621
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    65.4659     7.2187   9.069  < 2e-16 ***
## BloodPressure   0.2423     0.1022   2.371   0.0182 *
## Insulin         0.1372     0.0105  13.063  < 2e-16 ***
## Age             0.6036     0.1276   4.730 3.15e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.07 on 388 degrees of freedom
## Multiple R-squared:  0.3962, Adjusted R-squared:  0.3915
## F-statistic: 84.87 on 3 and 388 DF,  p-value: < 2.2e-16
```

Everything is significant in the reduced model so we'll run with this.

```
# Checking for multicollinearity
faraway::vif(reduced)
```

```
## BloodPressure       Insulin           Age
##      1.100343      1.050805      1.143554
```

No signs of multicollinearity here either.

```
# Conducting partial F test
anova(reduced, result)
```

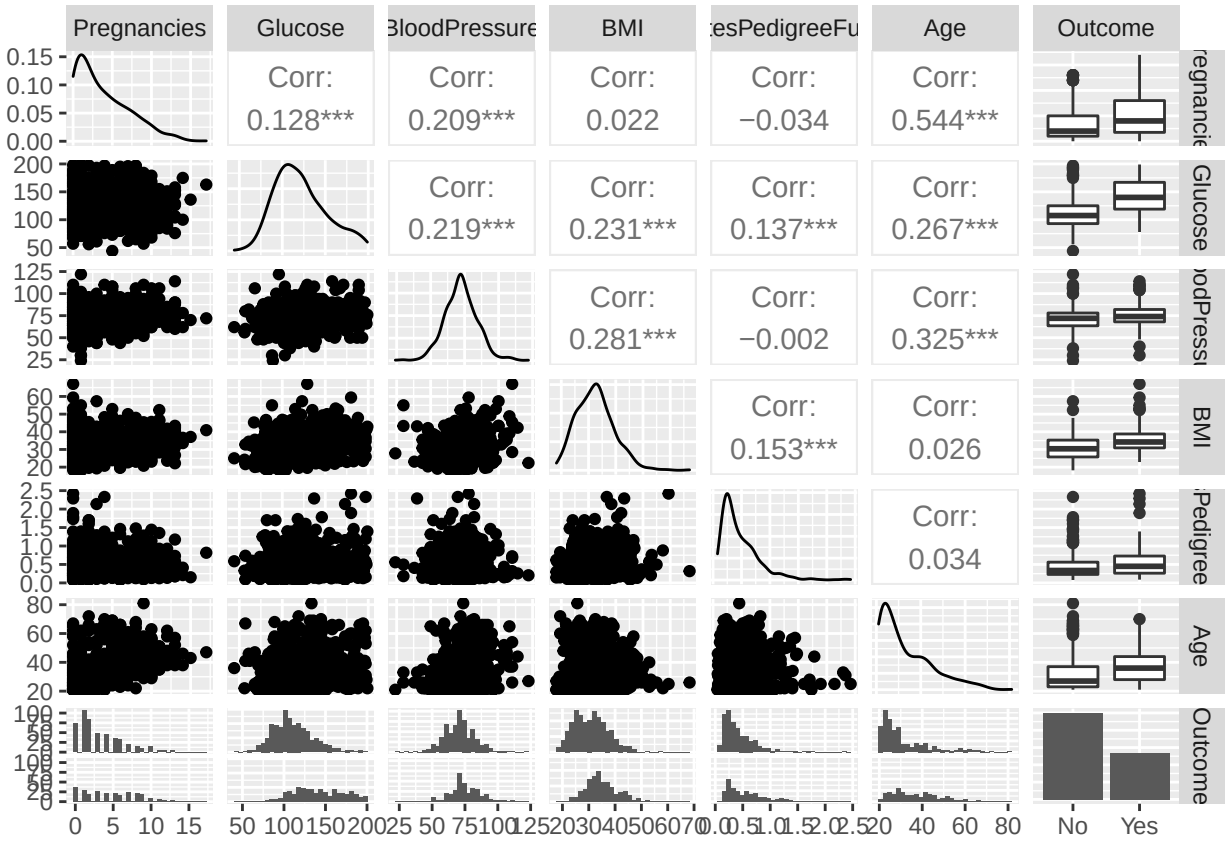
```
## Analysis of Variance Table
##
## Model 1: Glucose ~ BloodPressure + Insulin + Age
## Model 2: Glucose ~ Pregnancies + BloodPressure + SkinThickness + Insulin +
##     BMI + DiabetesPedigreeFunction + Age
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    388 224845
## 2    384 222975  4      1870 0.8051 0.5224
```

Insignificant p-value so we failed to reject the null and favor the reduced model.

# Second run for MLR

## Scatterplots

```
GGally::ggpairs(diabetes2)
```




## Full MLR with Pregnancies, Blood Pressure, BMI, DiabetesPedigeeFunction, and Age

```
result2 <- lm(Glucose ~ Pregnancies + BloodPressure + BMI + DiabetesPedigreeFunction + Age, data = diab
summary(result2)
```

```
##
## Call:
## lm(formula = Glucose ~ Pregnancies + BloodPressure + BMI + DiabetesPedigreeFunction +
##     Age, data = diabetes2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -77.261 -19.334  -2.609  16.460  85.447
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               53.86954    7.19234   7.490 1.91e-13 ***
## Pregnancies               -0.20870    0.36314  -0.575  0.56566
## BloodPressure              0.23689    0.09369   2.528  0.01166 *
## BMI                        0.81230    0.15770   5.151 3.30e-07 ***
## DiabetesPedigreeFunction  9.23650    3.13996   2.942  0.00336 **
## Age                        0.62331    0.10788   5.778 1.10e-08 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.35 on 762 degrees of freedom
## Multiple R-squared:  0.1381, Adjusted R-squared:  0.1325
## F-statistic: 24.42 on 5 and 762 DF,  p-value: < 2.2e-16
```

Pregnancies was insignificant in the presence of other variables so it's dropped.

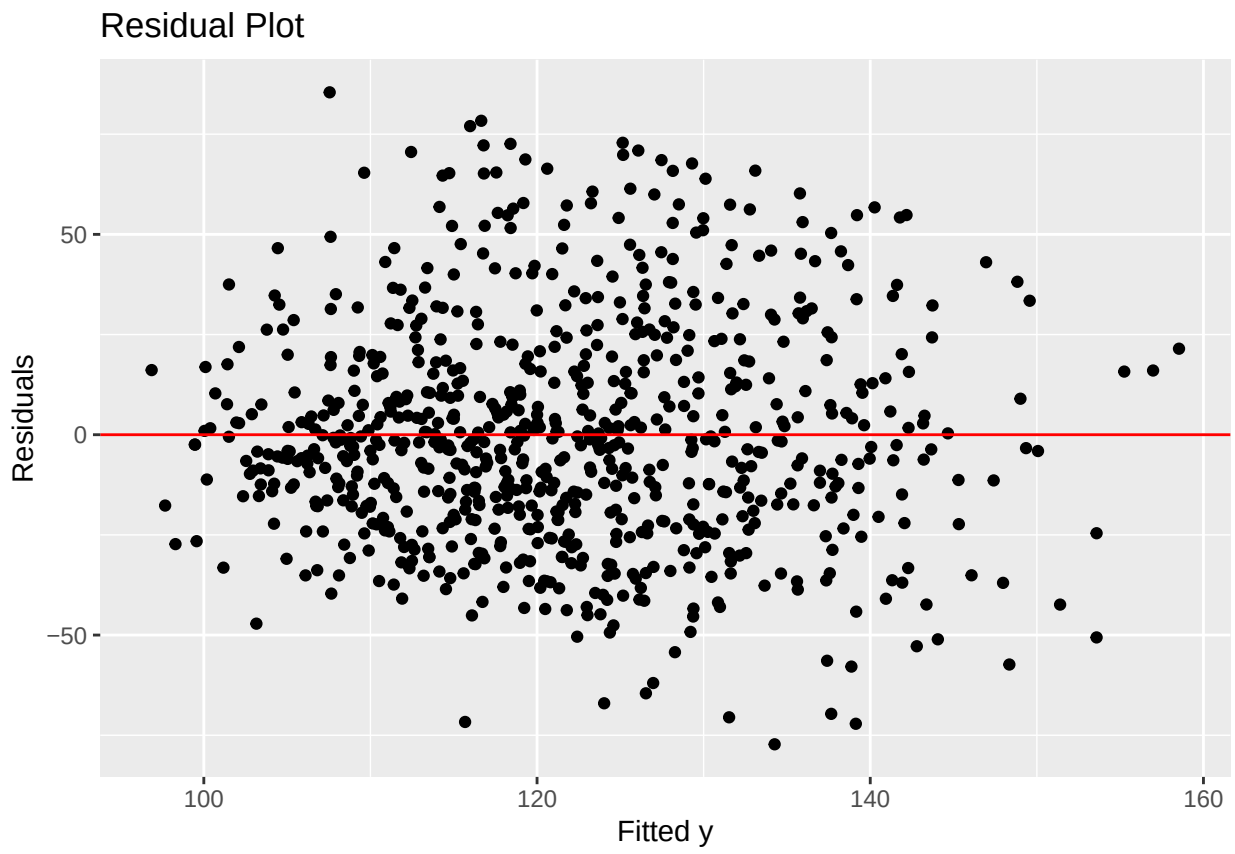
## Reduced MLR with Pregnancies dropped

```
reduced2 <- lm(Glucose ~ BloodPressure + BMI + DiabetesPedigreeFunction + Age, data = diabetes2)
summary(reduced2)
```

```
##
## Call:
## lm(formula = Glucose ~ BloodPressure + BMI + DiabetesPedigreeFunction +
##     Age, data = diabetes2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -75.438 -19.312  -2.569  16.997  85.677
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               54.23838    7.16050   7.575 1.04e-13 ***
## BloodPressure              0.23499    0.09360   2.511  0.01225 *
## BMI                        0.81161    0.15763   5.149 3.34e-07 ***
## DiabetesPedigreeFunction  9.34780    3.13260   2.984  0.00294 **
## Age                        0.59130    0.09235   6.403 2.67e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.34 on 763 degrees of freedom
## Multiple R-squared:  0.1378, Adjusted R-squared:  0.1332
## F-statistic: 30.47 on 4 and 763 DF,  p-value: < 2.2e-16
```

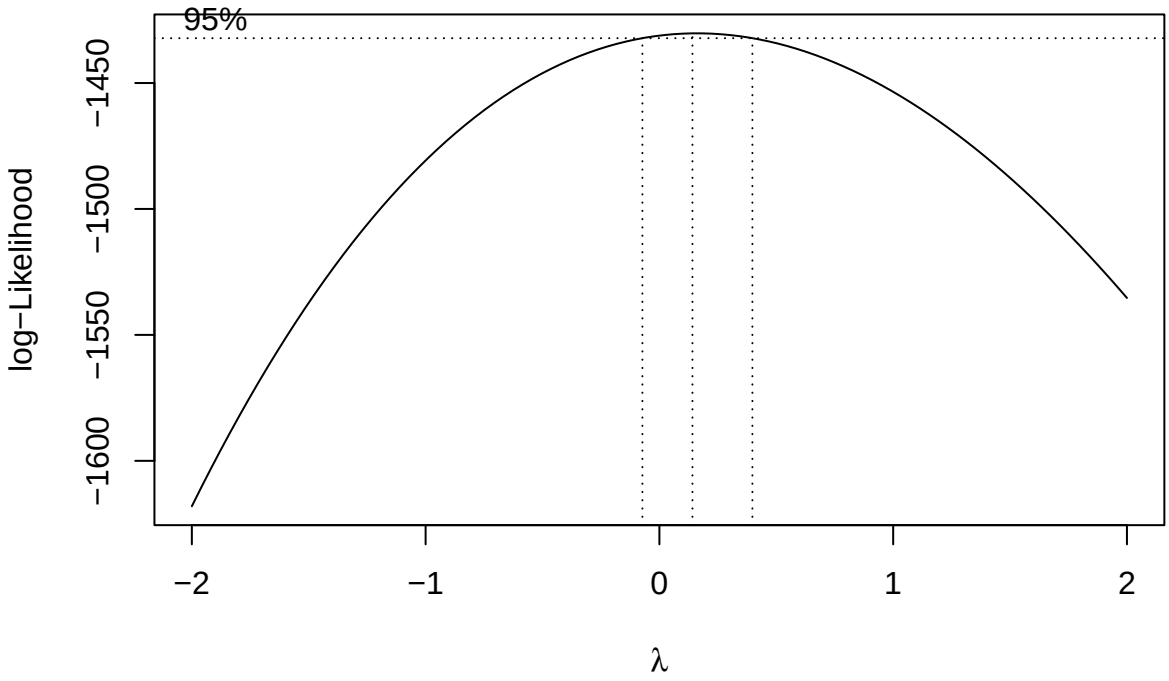
## Residual plot of reduced model

```
yhat <- result2$fitted.values
res <- result2$residuals
diabetes2 <- data.frame(diabetes2, yhat, res)

ggplot(diabetes2, aes(x = yhat,y = res))+
  geom_point()+
  geom_hline(yintercept=0, color="red")+
  labs(x="Fitted y",
       y="Residuals",
       title="Residual Plot")
```




### BoxCox of reduced model

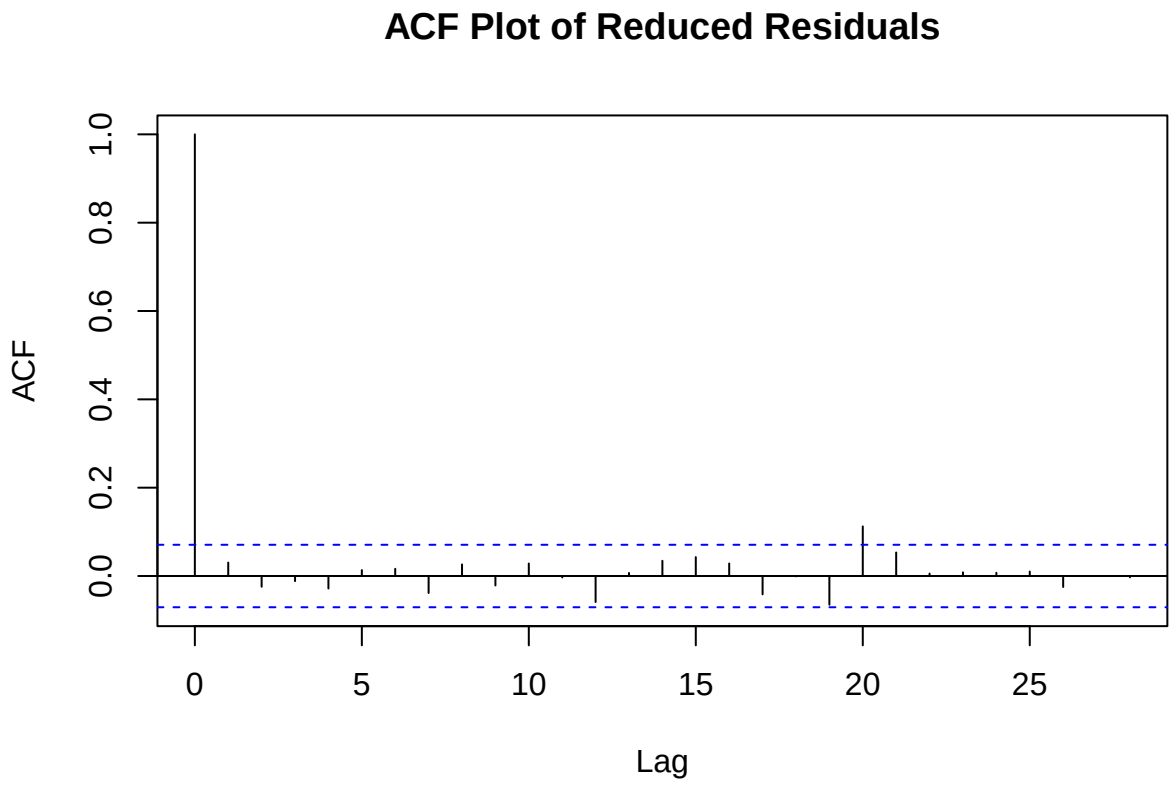
```
boxcox(reduced2)
```




Does not need any transformation on the response variable

### ACF of reduced model

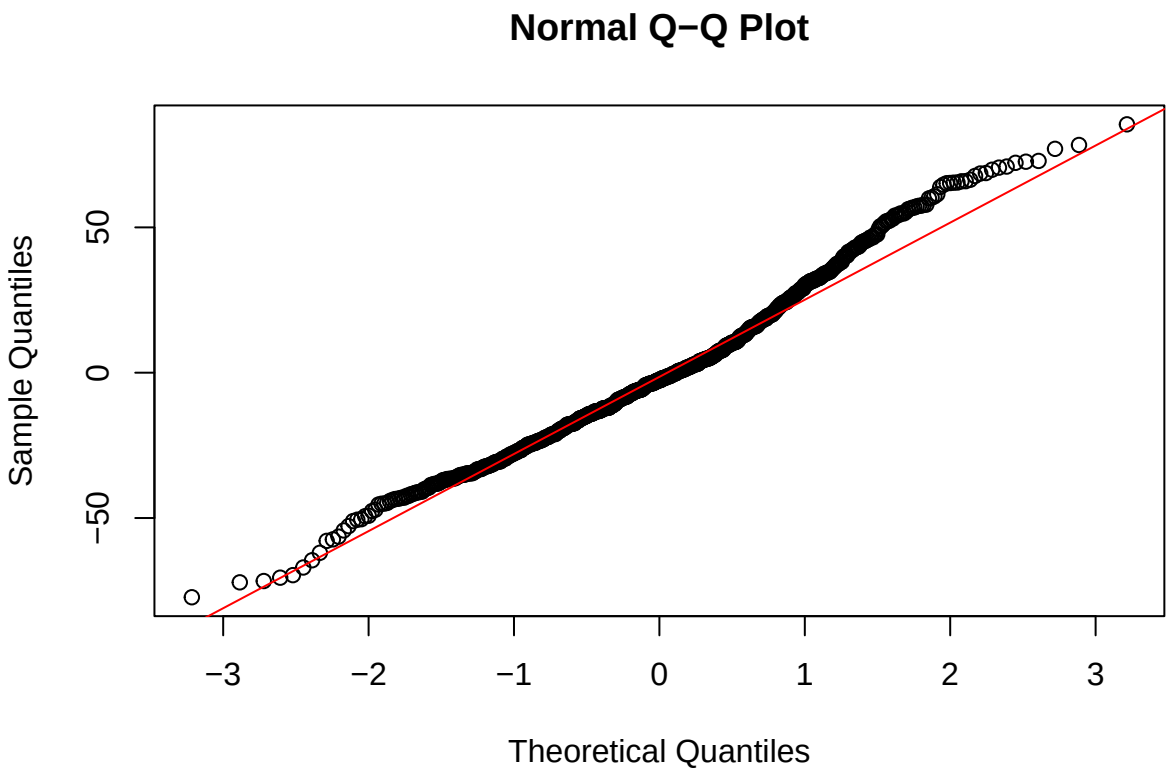
```
acf(res, main = "ACF Plot of Reduced Residuals")
```




Lag 19 and 20 look sus

### QQ plot of reduced

```
qqnorm(res)
qqline(res, col = "red")
```



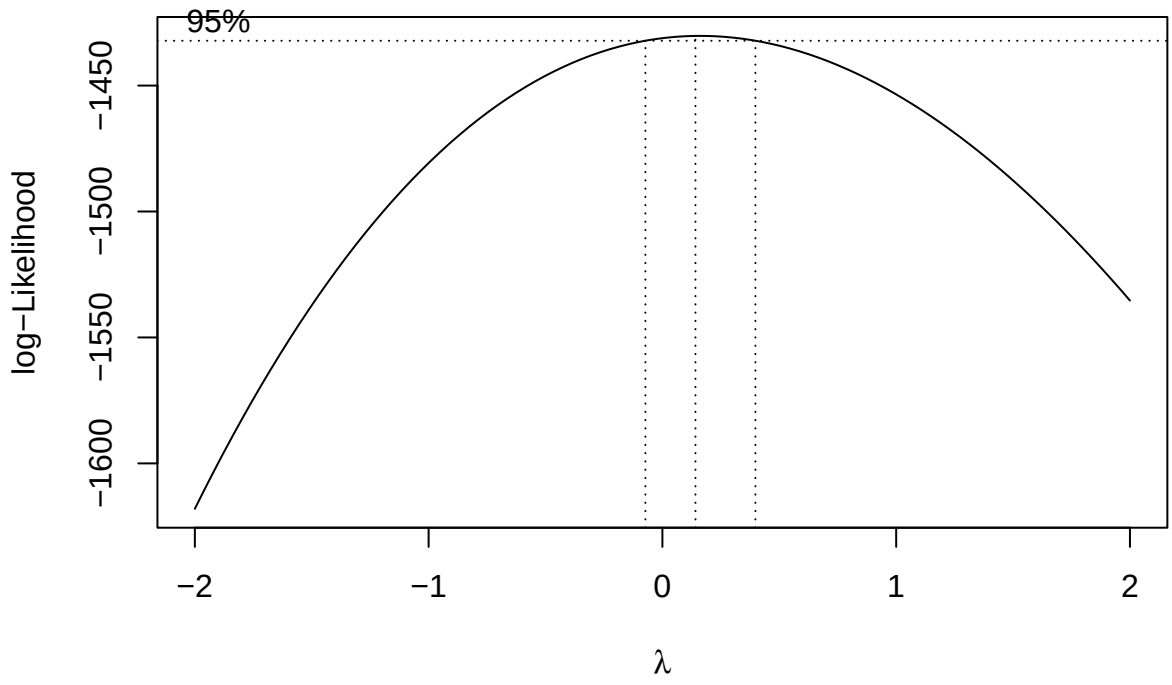

```
ystar <- log(diabetes2$Glucose)
diabetes2 <- data.frame(diabetes2, ystar)
```

```
reduced3 <- lm(Glucose ~ BloodPressure + BMI + DiabetesPedigreeFunction + Age, data = diabetes2)
summary(reduced3)
```

```
##
## Call:
## lm(formula = Glucose ~ BloodPressure + BMI + DiabetesPedigreeFunction +
##     Age, data = diabetes2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -75.438 -19.312  -2.569  16.997  85.677
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               54.23838    7.16050   7.575 1.04e-13 ***
## BloodPressure              0.23499    0.09360   2.511  0.01225 *
## BMI                        0.81161    0.15763   5.149 3.34e-07 ***
## DiabetesPedigreeFunction  9.34780    3.13260   2.984  0.00294 **
## Age                        0.59130    0.09235   6.403 2.67e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 28.34 on 763 degrees of freedom
## Multiple R-squared:  0.1378, Adjusted R-squared:  0.1332
## F-statistic: 30.47 on 4 and 763 DF,  p-value: < 2.2e-16
```

```
boxcox(reduced3)
```

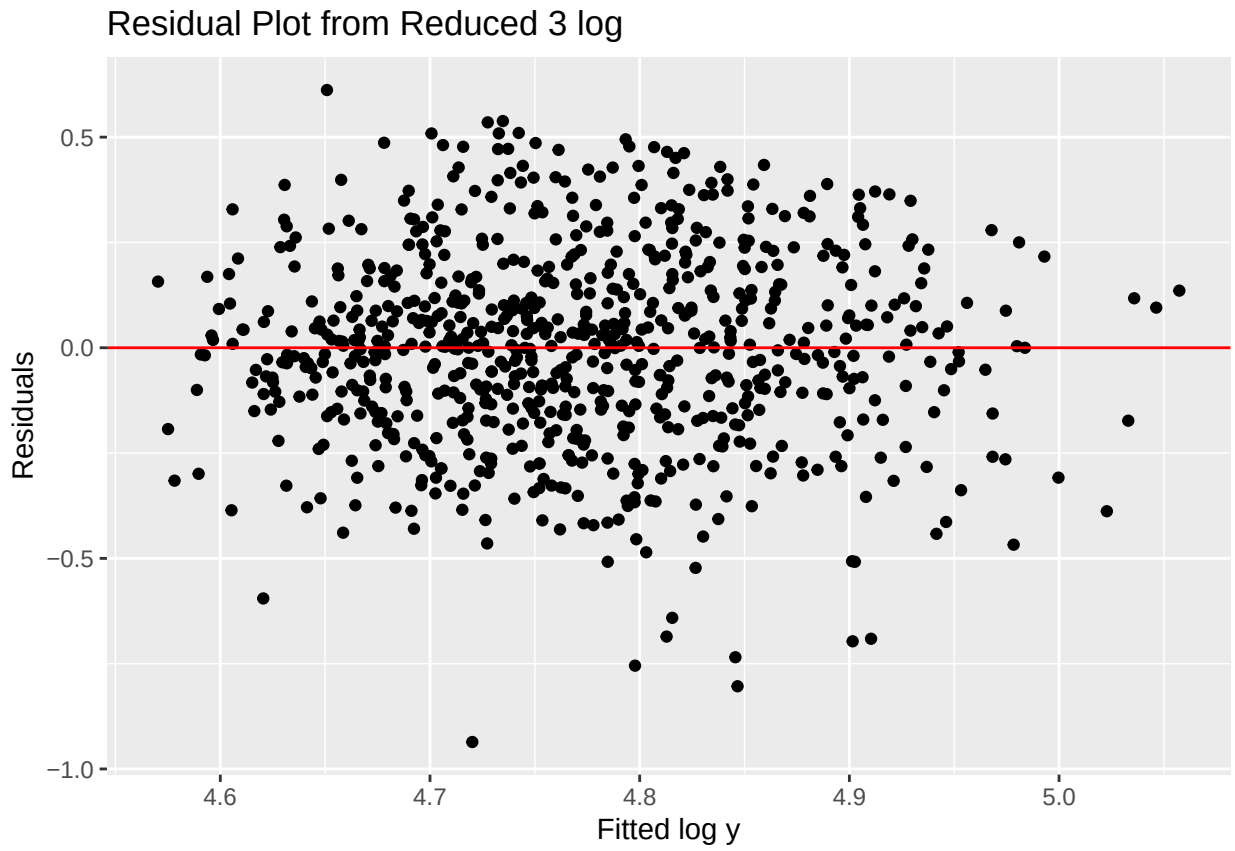



```
reduced3log <- lm(ystar ~ BloodPressure + BMI + DiabetesPedigreeFunction + Age, data = diabetes2)
summary(reduced3log)
```

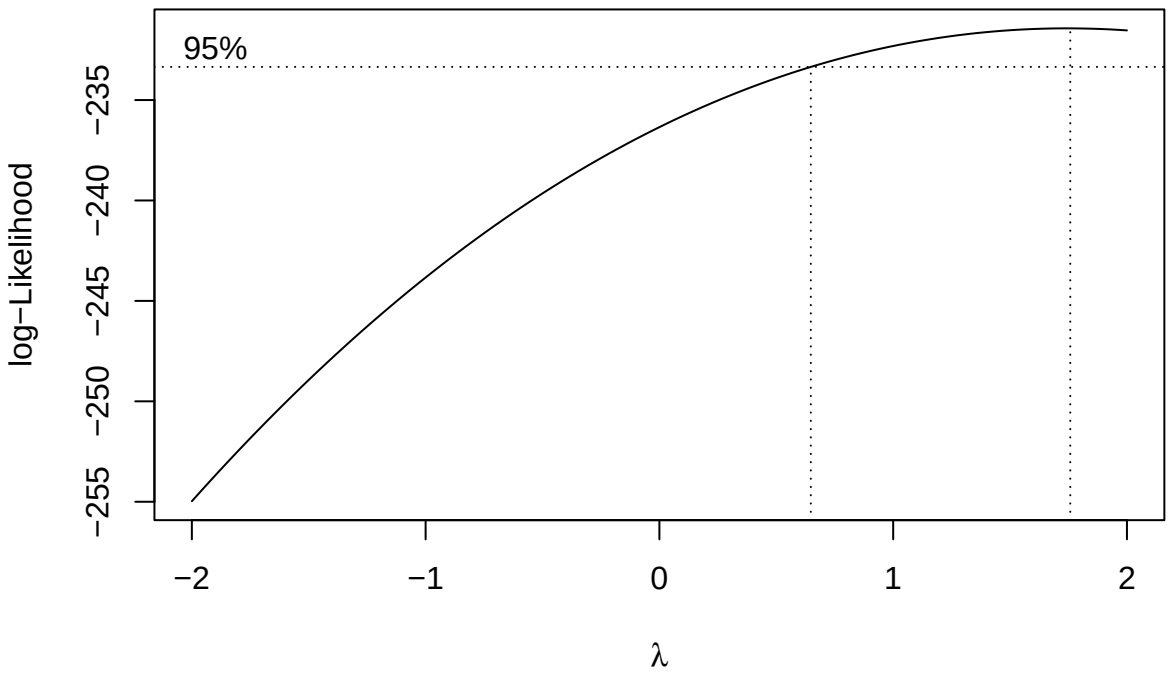
```
##
## Call:
## lm(formula = ystar ~ BloodPressure + BMI + DiabetesPedigreeFunction +
##     Age, data = diabetes2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.93600 -0.15099  0.00434  0.15470  0.61182
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               4.2205276  0.0589689  71.572  < 2e-16 ***
## BloodPressure             0.0020811  0.0007708   2.700  0.00709 **
## BMI                       0.0065896  0.0012981   5.076 4.84e-07 ***
## DiabetesPedigreeFunction 0.0696175  0.0257979   2.699  0.00712 **
## Age                       0.0045839  0.0007605   6.027 2.60e-09 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2334 on 763 degrees of freedom
## Multiple R-squared:  0.1311, Adjusted R-squared:  0.1265
## F-statistic: 28.78 on 4 and 763 DF,  p-value: < 2.2e-16
```

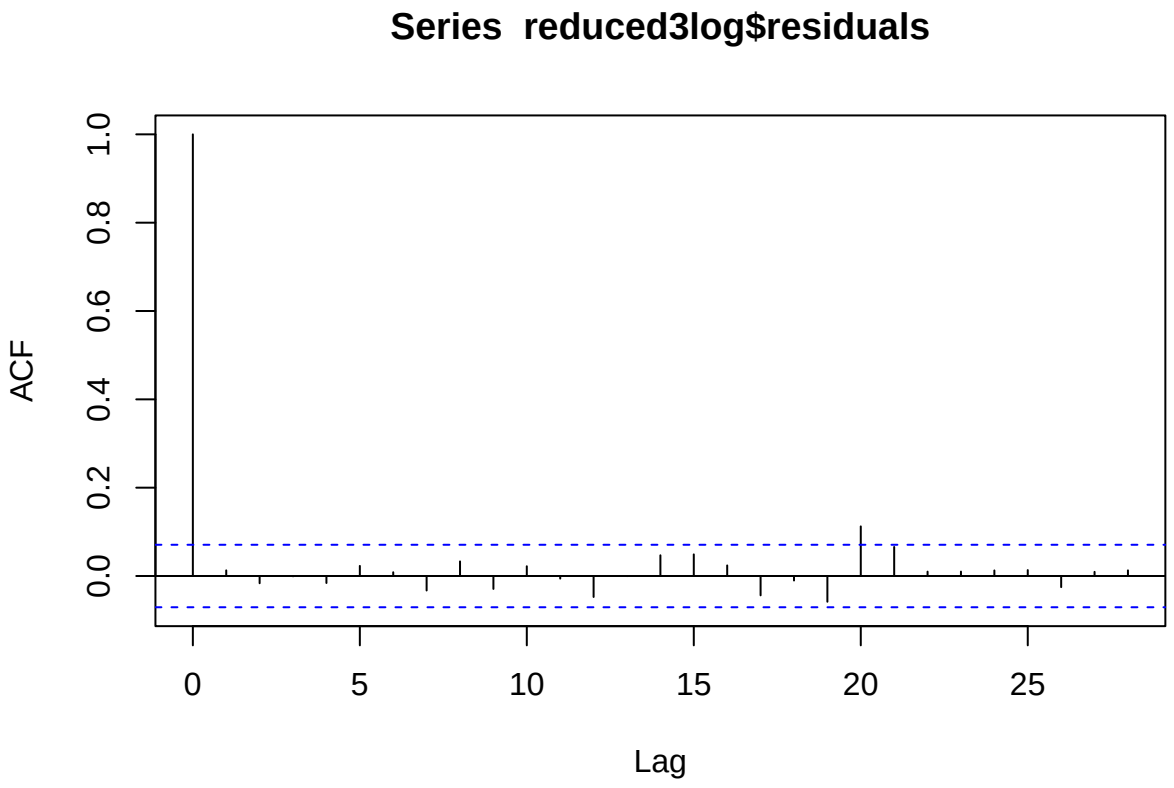
```
yhat3log <- reduced3log$fitted.values
res3 <- reduced3log$residuals
diabetes2 <- data.frame(diabetes2, yhat3log, res3)

ggplot(diabetes2, aes(x = yhat3log,y = res3))+
  geom_point()+
  geom_hline(yintercept=0, color="red")+
  labs(x="Fitted log y",
       y="Residuals",
       title="Residual Plot from Reduced 3 log")
```



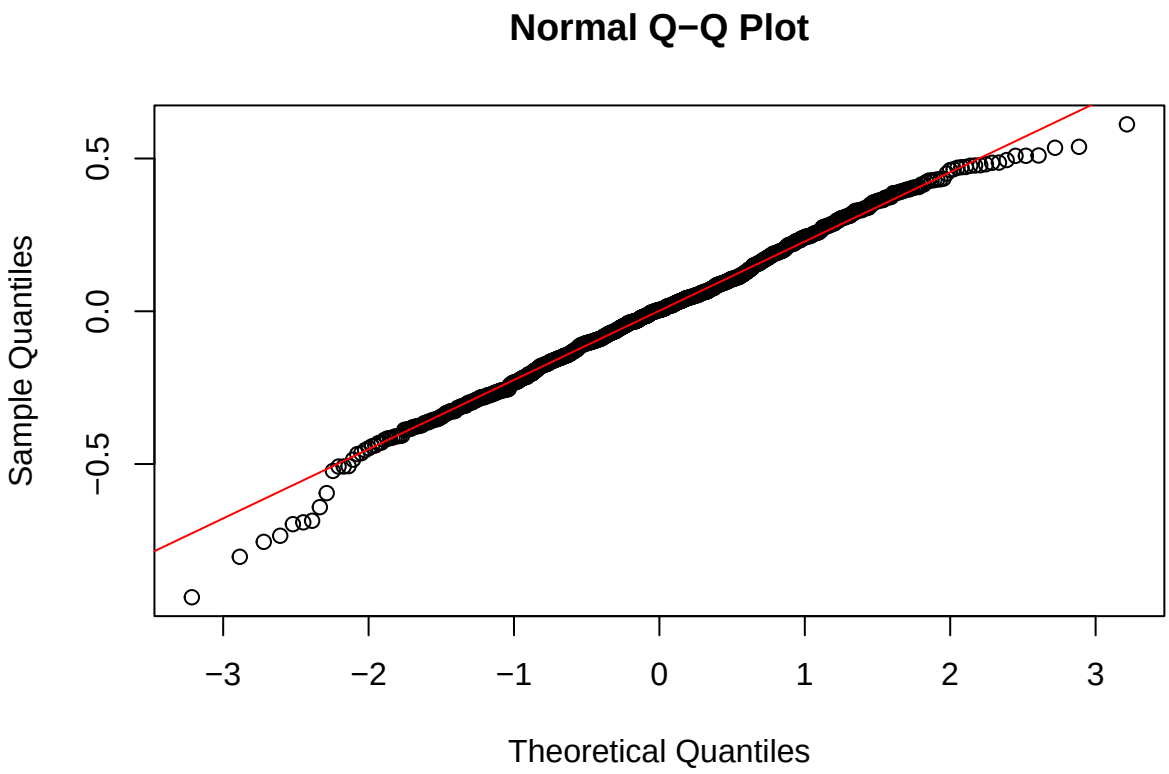

```
boxcox(reduced3log)
```




```
acf(reduced3log$residuals)
```




```
qqnorm(reduced3log$residuals)
qqline(reduced3log$residuals, col = "red")
```


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
0.5
0.0
−0.5
−3
−2
−1
0
1
2
3