Assignment 2

1. Do males and females differ significantly on their glycosylated hemoglobin values?

- Males and Females do not differ significantly on their glycosylated values, as the difference in the means of the baseline gender i.e., male and female is just 0.05634. The probability of getting a t-value 0.125 is more than 0.05 [p(t) = 0.901].
- The F-statistic of this model is 0.01568 on 1 and 128 DF, and the probability of getting this value is more than 0.05 [p(F) = 0.9006] Also, the amount of variance accounted for by the predictor glycosylated hemoglobin is -0.7689 % [Adjusted R- squared = -0.007689]. Hence, the model has a not a good fit, and the variables used as predictors do not have a strong-enough relationship with the outcome.

```
Call:
lm(formula = glyhb ~ gender, data = diabetes)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.1474 -1.5601 -0.9892  0.5464 10.1689 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   5.94105    0.33724  17.617   <2e-16 ***
genderfemale  0.05634    0.45003   0.125    0.901    
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.546 on 128 degrees of freedom
Multiple R-squared:  0.0001224, Adjusted R-squared:  -0.007689 
F-statistic: 0.01568 on 1 and 128 DF,  p-value: 0.9006
```

2. Are there significant differences in glycosylated hemoglobin across locations, after taking gender into account?

```
Call:
lm(formula = glyhb ~ gender + location, data = diabetes)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.9212 -1.5683 -0.9087  0.5151 10.4771 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)      6.5112     0.4489  14.504   <2e-16 ***
genderfemale     0.0517     0.4455   0.116   0.9078    
locationLouisa  -0.8783     0.4623  -1.900   0.0597 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.521 on 127 degrees of freedom
Multiple R-squared:  0.02775,	Adjusted R-squared:  0.01244 
F-statistic: 1.813 on 2 and 127 DF,  p-value: 0.1674
```

- There is a slight significance in glycosylated hemoglobin across the locations Louisa, and Buckingham when the later is considered as the baseline location, considering the slight difference in means of the two locations [coefficient = -0.8783]. The negative value implies that the mean of the location is lower than the baseline location.
- The F-statistic of this model is 1.813 on 2 and 127 DF, and the probability of getting this value is more than 0.05 [p(F) = 0.1674] Also, the amount of variance accounted for by the predictor glycosylated hemoglobin is 1.2 % [Adjusted R- squared = 0.01244]. Hence, the model has a not a good fit, although the model is better than the previous one.

3. Are cholesterol, stabilized glucose, HDL, cholesterol/HDL ratio, age, and weight/height ratio significant predictors of glycosylated hemoglobin?

- The null Hypothesis in this case is that there is no association between the cholesterol, stabilized glucose, HDL, cholesterol/HDL ratio, age, and weight/height ratio, and the alternate hypothesis is that there is an association between the variables. In this model, the stabilized glucose [p(t) = < 2e-16], and age [p(t) = 0.00678] are significant predictors of glycosylated hemoglobin as the probability of t-statistic is less than 0.05.
- The F-statistic of this model is 57.69 on 6 and 123 DF, and the probability of getting this value is less than 0.05 [p(F) = < 2.2e-16] Also, the amount of variance accounted for by the predictor glycosylated hemoglobin is 72.5 % [Adjusted R- squared = 0.725]. Hence, the model has a good fit, and the model is better than all the above models.

```
Call:
lm(formula = glyhb ~ chol + stab.glu + hdl + ratio + age + weight.height,

    data = diabetes)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.5256 -0.8281 -0.0771  0.5823  3.7783 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -1.961339   1.052699  -1.863  0.06483 .  
chol           0.002864   0.005424   0.528  0.59839    
stab.glu       0.030216   0.002273  13.292  < 2e-16 ***
hdl            0.011945   0.015208   0.785  0.43370    
ratio          0.261135   0.155303   1.681  0.09521 .  
age            0.023717   0.008612   2.754  0.00678 ** 
weight.height  0.276487   0.205981   1.342  0.18197    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.33 on 123 degrees of freedom
Multiple R-squared:  0.7378,	Adjusted R-squared:  0.725 
F-statistic: 57.69 on 6 and 123 DF,  p-value: < 2.2e-16
```

4. After taking into account all of the predictors listed above, does the effect of weight/height ratio vary significantly between the two genders?

- After considering all the predictors along with the interaction of weight/ratio and gender variable, the weight/height ratio does not vary significantly between the genders.

- There is no significance in glycosylated hemoglobin considering the effect of weight/height ratio vary significantly between the two genders, and the weight. height: gender female is considered as the baseline interaction term. The difference in means of the two locations is given by the coefficient [coefficient = 0.281066], and probability of t value is more than 0.05[p(t) = 0.5070]. In this model, the stabilized glucose [p(t) = < 2e-16], and age [p(t) = 0.0103] are significant predictors of glycosylated hemoglobin as the probability of t-statistic is less than 0.05.
- The F-statistic of this model is 42.93 on 8 and 121 DF, and the probability of getting this value is less than 0.05 [p(F) = < 2.2e-16] Also, the amount of variance accounted for by the predictor glycosylated hemoglobin is 72.22 % [Adjusted R- squared = 0.7222]. Hence, the model has a good fit, but the model is not better than the previous one due to less adjusted r- squared value and F-statistic value.

```
Call:
lm(formula = glyhb ~ chol + stab.glu + hdl + ratio + age + weight.height *
    gender, data = diabetes)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6358 -0.8456 -0.0422  0.5569  3.7620

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              -1.370779   1.330305  -1.030   0.3049
chol                      0.002230   0.005529   0.403   0.6875
stab.glu                  0.030513   0.002312  13.196   <2e-16 ***
hdl                       0.012244   0.015295   0.800   0.4250
ratio                     0.272832   0.157342   1.734   0.0855 .
age                       0.022761   0.008728   2.608   0.0103 *
weight.height             0.061220   0.355229   0.172   0.8635
genderfemale             -0.636765   1.197297  -0.532   0.5958
weight.height:genderfemale  0.281066   0.422381   0.665   0.5070
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.337 on 121 degrees of freedom
Multiple R-squared:  0.7395,    Adjusted R-squared:  0.7222
F-statistic: 42.93 on 8 and 121 DF,  p-value: < 2.2e-16
```

5. To model 3, add waist/hip ratio as a predictor and explain whether this is a significant predictor of glycosylated hemoglobin, after accounting for the variance accounted by all of the predictors in model 3.

- After adding waist/hip ratio as a predictor, the waist/hip ratio is not a significant predictor of glycosylated hemoglobin as the probability of t value is more than 0.05 [p(t) =0.75394]. In this model, the stabilized glucose [p(t) = < 2e-16], and age [p(t) = 0.00911] are significant predictors of glycosylated hemoglobin as the probability of t-statistic is less than 0.05.
- The F-statistic of this model is 49.1 on 7 and 122 DF, and the probability of getting this value is less than 0.05 [p(F) = < 2.2e-16] Also, the amount of variance accounted for by the predictor glycosylated hemoglobin is 72.3% [Adjusted R- squared = 0.723]. Hence, the model has a good fit, but the model is slightly better than the previous one.

```
Call:
lm(formula = glyhb ~ chol + stab.glu + hdl + ratio + age + weight.height
 +
    waist.hip, data = diabetes)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5443 -0.8196 -0.0851  0.5836  3.8253

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -2.388584   1.722198  -1.387  0.16799
chol           0.002847   0.005444   0.523  0.60191
stab.glu       0.030179   0.002285  13.210  < 2e-16 ***
hdl            0.012217   0.015289   0.799  0.42579
ratio          0.259636   0.155948   1.665  0.09850 .
age            0.023247   0.008772   2.650  0.00911 **
weight.height  0.276817   0.206742   1.339  0.18308
waist.hip      0.507493   1.615449   0.314  0.75394
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.335 on 122 degrees of freedom
Multiple R-squared:  0.738,     Adjusted R-squared:  0.723
F-statistic:  49.1 on 7 and 122 DF,  p-value: < 2.2e-16
```

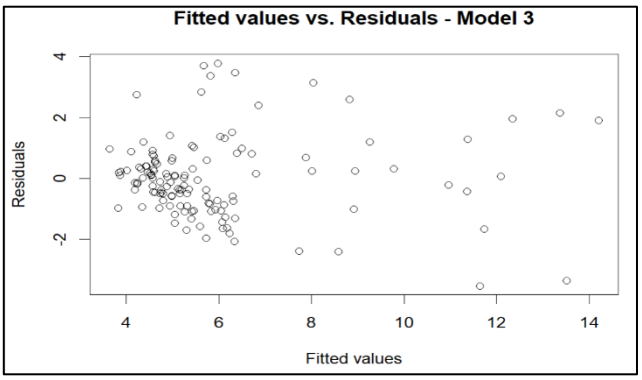
6. Which model among the 5 fit above yields best fit? Explain using appropriate out as evidence.

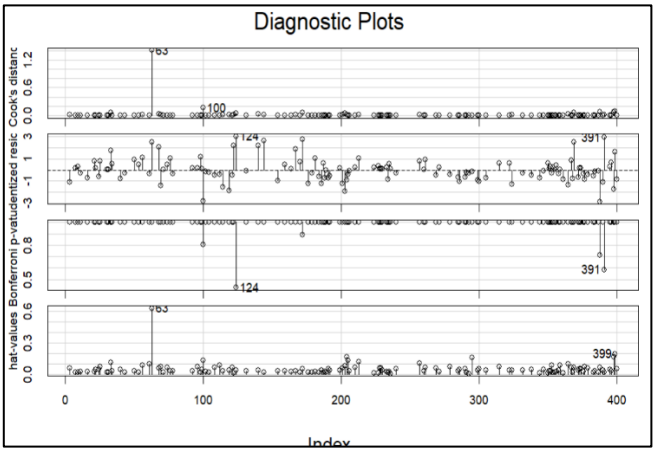
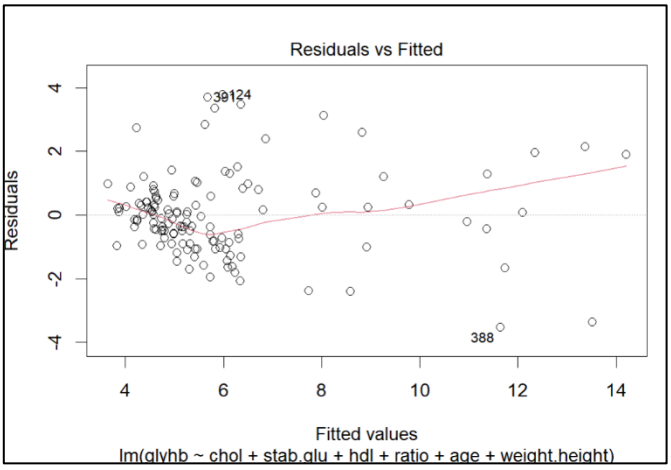
```
Analysis of Variance Table

Model 1: glyhb ~ gender
Model 2: glyhb ~ gender + location
Model 3: glyhb ~ chol + stab.glu + hdl + ratio + age + weight.height
Model 4: glyhb ~ chol + stab.glu + hdl + ratio + age + weight.height +
    weight.height * gender
Model 5: glyhb ~ chol + stab.glu + hdl + ratio + age + weight.height +
    waist.hip
  Res.Df    RSS Df Sum of Sq       F    Pr(>F)
1    128 829.76
2    127 806.83  1     22.93 12.8325 0.0004917 ***
3    123 217.57  4    589.26 82.4398 < 2.2e-16 ***
4    121 216.22  2      1.35  0.3787 0.6855920
5    122 217.40 -1     -1.18  0.6589 0.4185457
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Of all the models, model 2 and model 3 are statistically significant as the probability of F-statistic is less than 0.05 and p(F) are respectively, 0.0004917, and < 2.2e-16 respectively. And among model 2 and model 3, model 3 is more significant, considering the F-statistic and the degrees of freedom. The F-statistic of model 3 is 589.26, which is very large compared to model 2 and the degree of freedom is 123, which is less than model 2.
- Keeping this significance in the mind, it can be said that in this model, the stabilized glucose [p(t) = < 2e-16], and age [p(t) = 0.00678] are significant predictors of glycosylated hemoglobin. And the amount of variance accounted for by the predictor glycosylated hemoglobin is 72.5 %

[Adjusted R- squared = 0.725]. Hence, the model has a good fit, and the model is better than all the models.

7. For the best model, identify the pertinent assumptions underlying the OLS multiple regression approach and using appropriate evidence, determine the degree to which each assumption was met. Present your results in a summarized tabular format – a table with three columns: (1) the name of the assumption, (2) evidence (refer to appropriate graphical/numeric information) to indicate support/no support, and (3) conclusion on the degree to which the assumption is met.

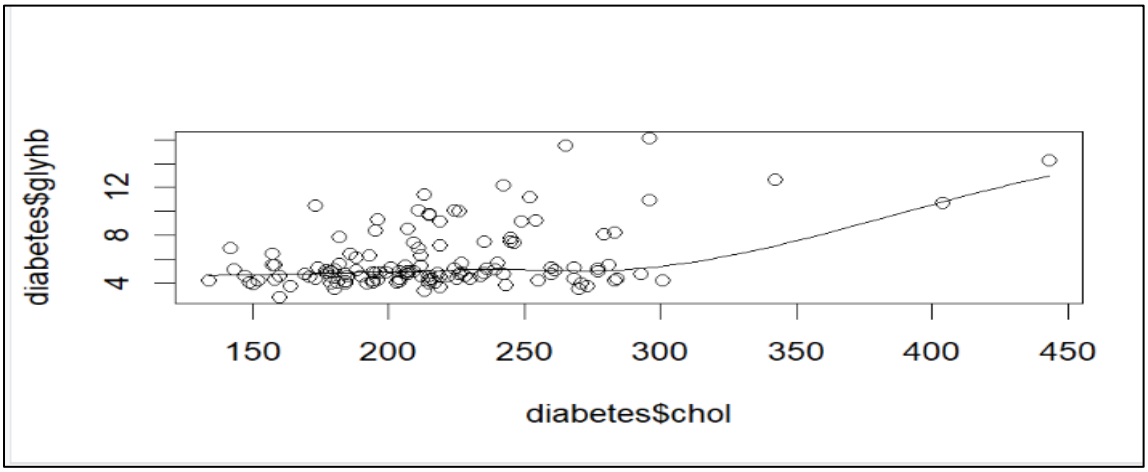
ASSUMPTION	EVIDENDENCE (graphical or numerical)	CONCLUSION
Multi-collinearity	chol 4.831737 stab.glu 1.297869 hdl 5.554972 ratio 7.539853 age 1.225463 weight.height 1.129863	As no variable has inflation factor greater than 10, this assumption is met.
Independent errors	lag Autocorrelation D-W Statistic p-value 1 0.08984088 1.805719 0.302 Alternative hypothesis: rho != 0	As the value of D-W Statistic is close to 2, condition is most likely to have been met.
Homoscedasticity		As the residuals are not increasing with predicted values, this assumption is most likely to be met.
Normality of residuals	Shapiro-Wilk normality test data: lm.model.3$residuals W = 0.96097, p-value = 0.0008745	The distribution of the model residuals is significantly different from normal distributions. So, this assumption is not met.

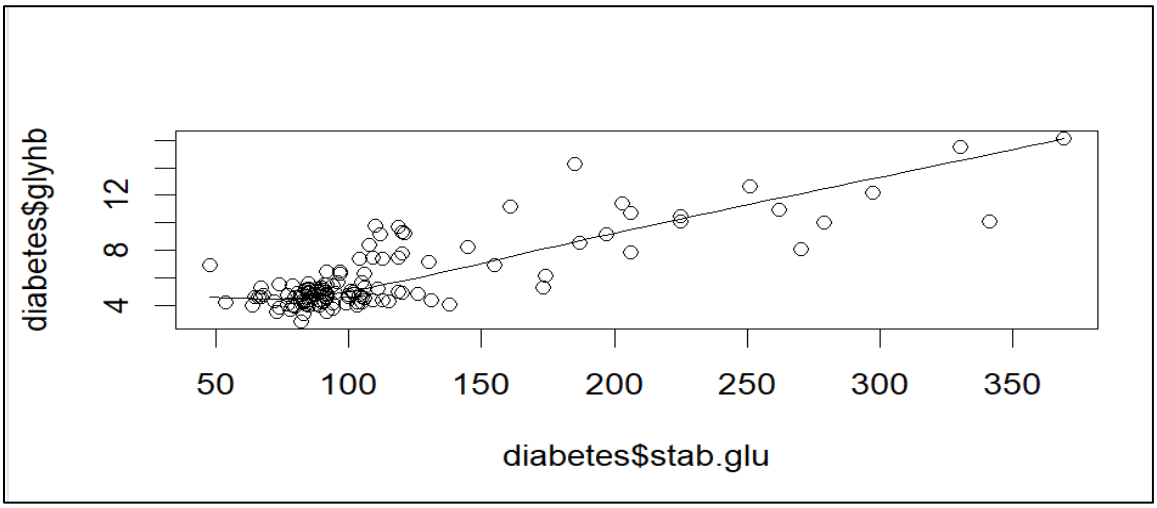
Presence of outliers		There are only 2 observations whose cook's value is away from 1. This implies that the assumption is almost met
Linearity		The linearity assumption is not met.
Non-zero variance in the outcome variable	The variance values is 6.433062	Since the value is not close to zero, this assumption is met.

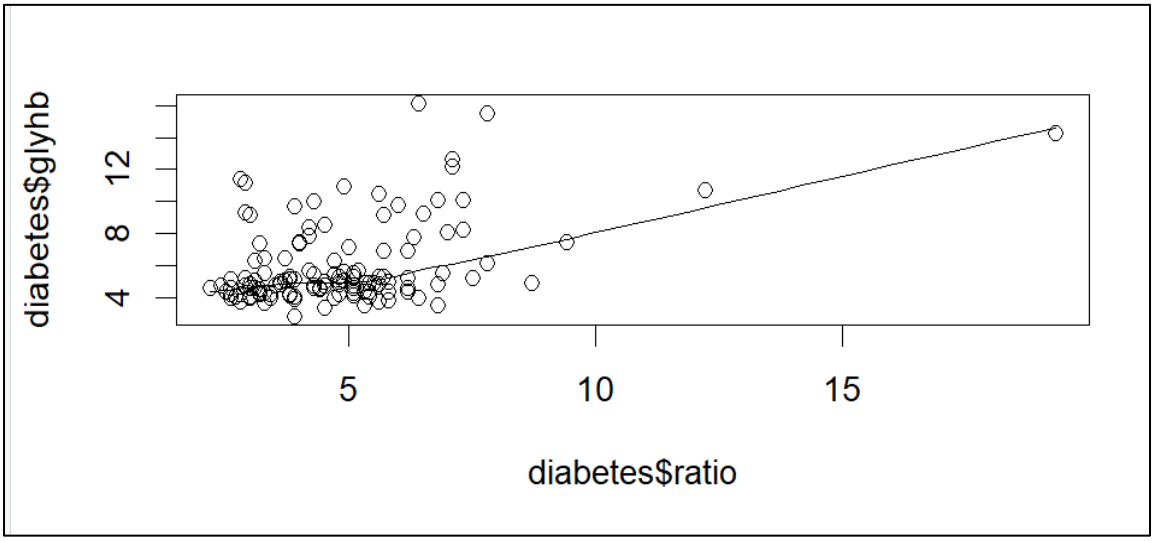
8. Based on the validity of assumptions, are any remedial steps needed to allow the model assumptions to be valid? Identify and perform the needed remedial steps, and comment on the degree to which these steps improved the satisfaction of model assumptions.

- The linearity assumption and normality of residuals assumption are not met. To allow the model assumptions to be valid, the linearity of every variable is checked, and log and square root transformations are performed for variables not obeying the linearity. The cholesterol, stabilized glucose, HDL, cholesterol/HDL ratio are found to be nonlinear.
- The transformations did not allow the model assumptions to be valid.

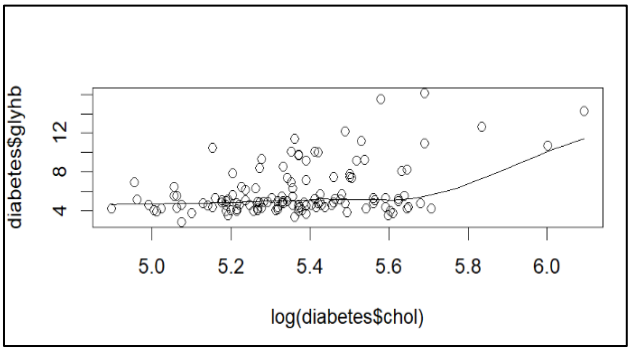
Before Transformations:

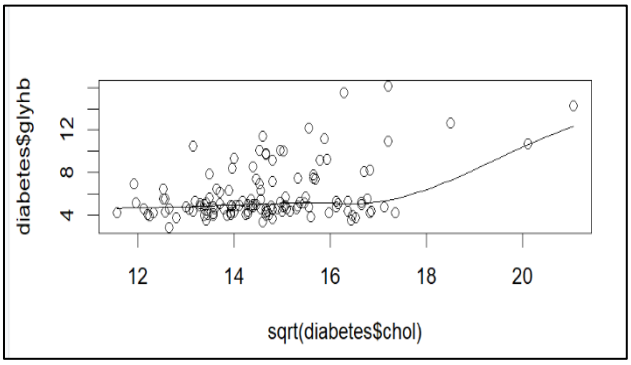


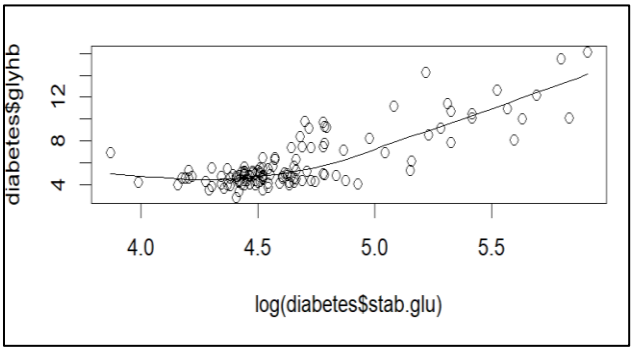


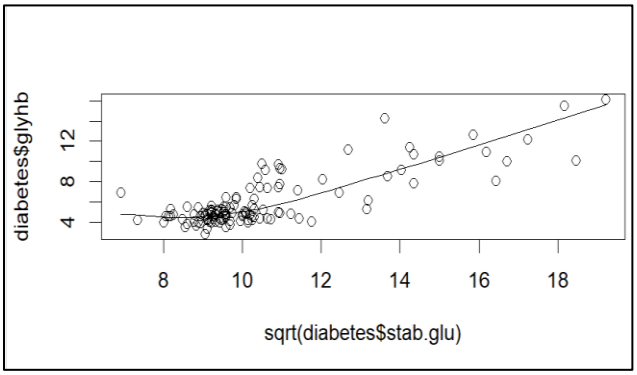


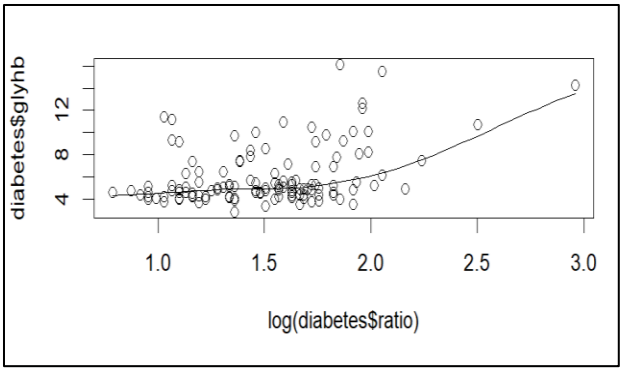
After Transformation:

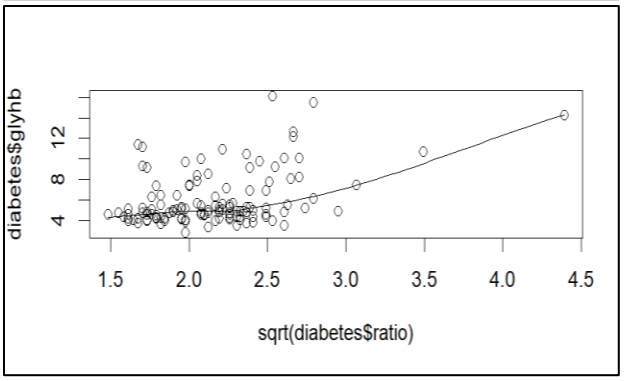












9. Are any further remedial measures needed? Explain via suitable evidence.

- The weighted regression analysis is done to improve the model. The only significant predictor is stabilised glucose with the probability of t value less than 0.05 [p(t) = <2e-16]. The F-statistic of this model is 41.55 on 7 and 122 DF, and the probability of getting this value is less than 0.05 [p(F) = < 2.2e-16] Also, the amount of variance accounted for by the predictor glycosylated hemoglobin is 68.75% [Adjusted R- squared = 0.6875]. Hence, the model has a good fit, but the model is not better than the original one.

```
Call:
lm(formula = glyhb ~ chol + stab.glu + hdl + ratio + age + weight.
height +
    waist.hip, data = diabetes, weights = 1/wts)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-6.9342 -1.0036 -0.2919  1.2420  7.7376

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.550241   1.538311  -0.358    0.721
chol           0.006059   0.005001   1.212    0.228
stab.glu       0.029241   0.002230  13.115   <2e-16 ***
hdl            0.005276   0.013849   0.381    0.704
ratio          0.152740   0.146176   1.045    0.298
age            0.011445   0.008027   1.426    0.157
weight.height  0.142911   0.177926   0.803    0.423
waist.hip     -0.069965   1.423157  -0.049    0.961
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.892 on 122 degrees of freedom
Multiple R-squared:  0.7045,    Adjusted R-squared:  0.6875
F-statistic: 41.55 on 7 and 122 DF,  p-value: < 2.2e-16
```

- So, the outcome variable is transformed and the model is bulit. The significant predictors are stabilised glucose [p(t) = <2e-16] and age [p(t) = 0.335] with the probability of t value less than 0.05. The F-statistic of this model is 33.2 on 7 and 122 DF, and the probability of getting this value is less than 0.05 [p(F) = < 2.2e-16] Also, the amount of variance accounted for by the predictor glycosylated hemoglobin is 63.6% [Adjusted R- squared = 0.636]. Hence, the model has a good fit, but the model is not better than the original one or the previous non-transformed weighted regression.

```
Call:
lm(formula = log(glyhb) ~ chol + stab.glu + hdl + ratio + age +
    weight.height + waist.hip, data = diabetes, weights = 1/wts)

Weighted Residuals:
     Min       1Q   Median       3Q      Max
-0.84417 -0.16822 -0.06198  0.22802  1.11412

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.9024301  0.2160735   4.176 5.59e-05 ***
chol           0.0008516  0.0007024   1.212   0.2277
stab.glu       0.0035991  0.0003132  11.493  < 2e-16 ***
hdl           -0.0004136  0.0019453  -0.213   0.8320
ratio          0.0112718  0.0205320   0.549   0.5840
age            0.0024248  0.0011276   2.151   0.0335 *
weight.height  0.0304124  0.0249918   1.217   0.2260
waist.hip      0.0044407  0.1998988   0.022   0.9823
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4062 on 122 degrees of freedom
Multiple R-squared:  0.6558,    Adjusted R-squared:  0.636
F-statistic:  33.2 on 7 and 122 DF,  p-value: < 2.2e-16
```

10. What overall conclusions can you draw from the model results, taking into account the diagnostic information related to the validity of the modeling assumptions?

- The non-weighted linear regression model is better compared all the other models, but as the two assumptions failed for the model we can consider other algorithms to get best fit model. If we see the assumptions as less needed ones, we can proceed with that model to draw conclusions about the population dataset.