1.

In two to three paragraphs briefly summarize the paper and blog post relating to challenges with how we try to use race and ethnicity as variables in models. You are also welcome to provide any personal thoughts, experiences, or sources that you wish to incorporate.

Algorithms are nothing but tools that physicians are using to make decisions around health care. Understanding Race is a major challenge in clinical algorithms as it has multiple implications based on what context, Race has been used. There is an ongoing conflict between the population genetics and clinical implication of Race. In the introduction of the paper, they have given the example of hydralazine and isosorbide dinitrate which reduced the morality due to heart failure in between the patients who identified Black. But here Black in terms of colour or origin has not been specified clearly.

Study published in June 17, 2020 New England Journal of Medicine, looked at 13 algorithms for medical decisions. Example- some physicians may use Americans Heart Association's "Heart Failure Risk Score" for determining a heart patient's future treatment. Now here the algorithm embeds race as it include "**race-correction**" mechanism where a **Black** is automatically assumed to be lower risk that a **nonblack** patient which results given 3 fewer points under the given algorithm that *"may raise the threshold for using clinical resources for black patients*". The authors are not suggesting to exclude the race completely rather they propose3 questions for doctors developing algorithms to ask themselves, whether to include a " race – correction" mechanism in the algorithms i.e.

1. Is the need for race correction based on robust evidence and statistical analyses (like: with the consideration of internal & external validity and bias)?

2. Is there a plausible causal mechanism for racial differences that justifies this bias?

3. Would implicating this race correction relieve or exacerbate health inequalities?"

In this paper authors has concluded,

"Their understanding of race has changed over 2 decades. But still its equally important is the project of making medicine more antiracist field. So One change can be considered on this process is to ensure that clinical practices do not perpetuate the very inequities they aim to repair"

One of my Computer vision project, while finding the criminal identification, "the people who are of non-white colour treated as most threatening people compare to white people". I personally feel this bias should not be considered as its not evident to predict such misconception.

2.a

```
amniotic <- data.frame(
cells = c(1.13,1.20,1.00,0.91,1.05,1.75,1.45,1.55,1.64,1.60,
2.30,2.15,2.25,2.40,2.49,3.18,3.10,3.28,3.35,3.12),
temp = c(rep(40,5), rep(60,5), rep(80,5), rep(100,5)))
amniotic$ln_cells <- log(amniotic$cells) # calculate log(cells)
```

```
model <- glm( log(cells) ~ temp, data=amniotic)
```

```
summary(model)
```

```
Call:
glm(formula = log(cells) ~ temp, data = amniotic)

Deviance Residuals:
      Min         1Q     Median         3Q        Max
-0.168328  -0.049781  -0.002362   0.048546   0.114505

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.6681674  0.0527461  -12.67  2.1e-10 ***
temp         0.0185546  0.0007178   25.85  1.1e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.005152141)

    Null deviance: 3.535481  on 19  degrees of freedom
Residual deviance: 0.092739  on 18  degrees of freedom
AIC: -44.717

Number of Fisher Scoring iterations: 2
```

Interpretation of the slope on the original scale: If we exponentiate the slope our interpretation changes to the percent increase (or multiplicative change): $e^{0.019}=1.019$e0.019=1.019. On average, a one-unit increase in temperature results in a cell count that is 1.9% higher (1.019 times higher).

95% CI:

On log(cells),

0.0186 ± 1.96 * (0.0007178) = (0.0172, 0.02001)

```
confint(model,'temp',level=0.95)
```

```
Waiting for profiling to be done...
```

2.5 %	0.0171477975752772
97.5 %	0.0199614598149313

Interpretation of 95% CI:

The 95% confidence interval for the slope is (0.0172, 0.0199)

95% confident that the true mean geometric multiplicative increase in cells counts are in between 1.7%, 2.0% for each 1 unit increase in temperature.

2b.

Model Without Log Tranformation

```
model_withoutlog <- glm( cells ~ temp, data=amniotic)
```

```
summary(model_withoutlog)
```

```
Call:
glm(formula = cells ~ temp, data = amniotic)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.2532  -0.0909  -0.0014   0.0814   0.2304

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.462400   0.104811  -4.412 0.000337 ***
temp         0.035820   0.001426  25.114 1.83e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.02034311)

    Null deviance: 13.19690  on 19  degrees of freedom
Residual deviance:  0.36618  on 18  degrees of freedom
AIC: -17.25

Number of Fisher Scoring iterations: 2
```

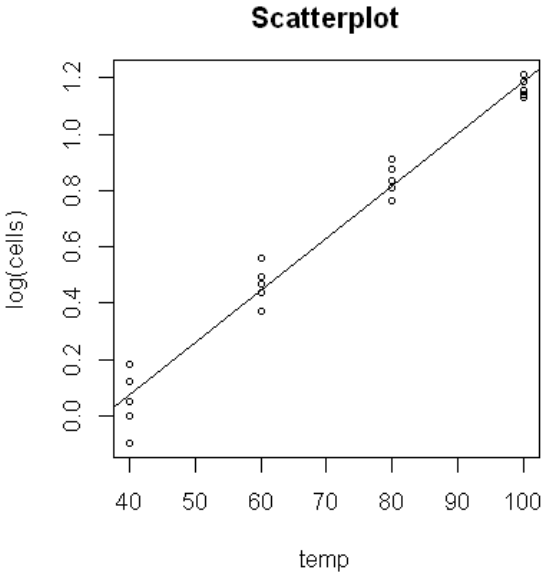
```
summary(model_withoutlog)$coefficients
```

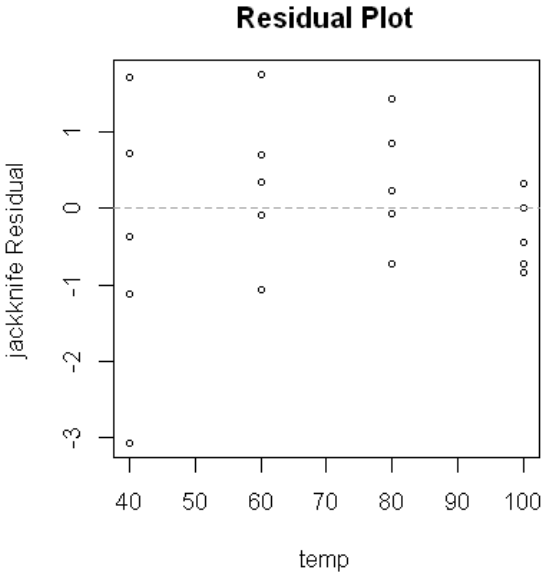
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.46240	0.104810686	-4.411764	3.365925e-04
temp	0.03582	0.001426293	25.114058	1.831172e-15

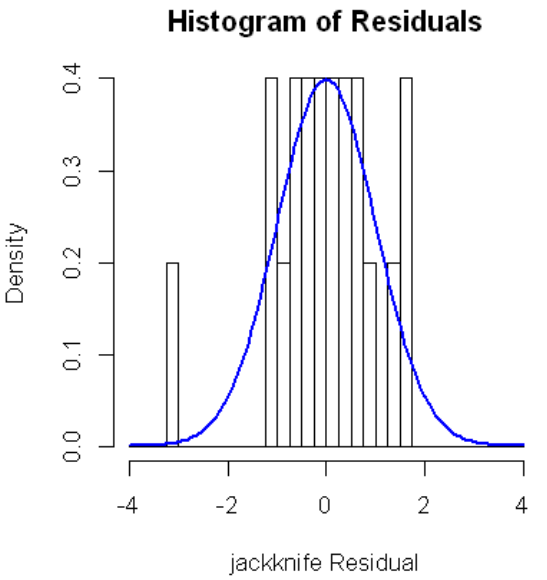
```
par(mfrow=c(2,2), mar=c(4.1,4.1,3.1,2.1))
plot(x=amniotic$temp, y=log(amniotic$cells), xlab='temp', ylab='log(cells)',
main='Scatterplot', cex=0.7); abline( model_withoutlog )

plot(x=amniotic$temp, y=rstudent(model_withoutlog), xlab='temp', ylab='jackknife Residual',
main='Residual Plot', cex=0.7); abline(h=0, lty=2, col='gray65')

hist(rstudent(model_withoutlog), xlab='jackknife Residual',
main='Histogram of Residuals', freq=F, breaks=seq(-4,4,0.25));


curve( dnorm(x,mean=0,sd=1), lwd=2, col='blue', add=T)
plot( ppoints(length(rstudent(model_withoutlog))), sort(pnorm(rstudent(model_withoutlog))),
xlab='Observed Cumulative Probability',
ylab='Expected Cumulative Probability',
main='Normal Probability Plot', cex=2, pch='.');
abline(a=0,b=1, col='gray65', lwd=1)
```

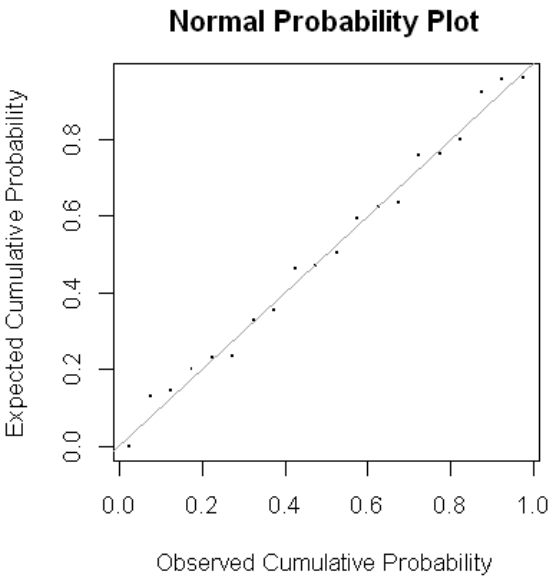
Log Transformation

```
par(mfrow=c(2,2), mar=c(4.1,4.1,3.1,2.1))
plot(x=amniotic$temp, y=log(amniotic$cells), xlab='temp', ylab='log(cells)',
main='Scatterplot', cex=0.7); abline( model )

plot(x=amniotic$temp, y=rstudent(model), xlab='temp', ylab='jackknife Residual',
main='Residual Plot', cex=0.7); abline(h=0, lty=2, col='gray65')

hist(rstudent(model), xlab='jackknife Residual',
main='Histogram of Residuals', freq=F, breaks=seq(-4,4,0.25));


curve( dnorm(x,mean=0,sd=1), lwd=2, col='blue', add=T)
plot( ppoints(length(rstudent(model))), sort(pnorm(rstudent(model))),
xlab='Observed Cumulative Probability',
ylab='Expected Cumulative Probability',
main='Normal Probability Plot', cex=2, pch='.');
abline(a=0,b=1, col='gray65', lwd=1)
```









```
coef(model)
```

(Intercept)	-0.668167432254476
temp	0.0185546286951042

This results in E[log(FEV)] = -0.66816 + 0.01855 × temp

With respect to log(Y),

we have our usual interpretations:

The intercept of -0.66816 is the mean log(FEV) for someone who is having temp 0.

The slope term indicates that for each 1 cell increase, on average, log(FEV) increases by 0.01855.

Our Predicted regression equation is Y = 0.01855 × (X) - 0.66816

Here Y is cells and X = Temp

The estimates are defined by β0 = -0.66816, β1 = 0.01855, which depicts that for every 1 unit increase of temperature the natural log of cells is -0.668.

95% CI:

E[log(FEV)] = -0.66816 + 0.01855 × temp

= exp(-0.66816 + 0.01855 × temp)

Exp(E| log(cells)] = exp[(-0.66816) + exp(0.01855 × temp)]

= 0.513 * (1.019)^temp

```
summary(model)$coefficients
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.66816743	0.0527461488	-12.66761	2.100675e-10
temp	0.01855463	0.0007177842	25.84987	1.103493e-15

For every 1 unit increase in temperature, the number of cells increase 0.01855 and the expected geometric mean number of cells at temperature with 0 degrees of freedom is 0.513 * 10^6

Question: 3

```
Heart_study <- read.csv('D:/3rd_Semester/6611_biostatisticalmethod/hw8/frmgham2_baseline_subset.csv')
```

```
summary(Heart_study)
```

```
 Qu.:144.0   3rd Qu.: 90.00   3rd Qu.:1.0000   3rd Qu.:20.000
.   :295.0   Max.   :142.50   Max.   :1.0000   Max.   :70.000
                                                NA's   :32
   BMI           DIABETES           BPMEDS           HEARTRTE
.   :15.54   Min.   :0.00000   Min.   :0.00000   Min.   : 44.00
: Qu.:23.09   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.: 68.00
dian :25.45   Median :0.00000   Median :0.00000   Median : 75.00
an   :25.85   Mean   :0.02729   Mean   :0.03293   Mean   : 75.89
 Qu.:28.09   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.: 83.00
.   :56.80   Max.   :1.00000   Max.   :1.00000   Max.   :143.00
's   :19                       NA's   :61        NA's   :1
 GLUCOSE            educ          PREVCHD           PREVAP
.   : 40.00   Min.   :1.000   Min.   :0.00000   Min.   :0.00000
: Qu.: 72.00   1st Qu.:1.000   1st Qu.:0.00000   1st Qu.:0.00000
dian : 78.00   Median :2.000   Median :0.00000   Median :0.00000
an   : 82.19   Mean   :1.976   Mean   :0.04375   Mean   :0.03315
 Qu.: 87.00   3rd Qu.:3.000   3rd Qu.:0.00000   3rd Qu.:0.00000
.   :394.00   Max.   :4.000   Max.   :1.00000   Max.   :1.00000
```

```
model3 <- lm(TOTCHOL ~ CURSMOKE + BMI + PREVMI +PREVCHD + PREVSTRK+ PREVHYP, data = Heart_study)
summary(model3)
```

```
Call:
lm(formula = TOTCHOL ~ CURSMOKE + BMI + PREVMI + PREVCHD + PREVSTRK +
    PREVHYP, data = Heart_study)

Residuals:
    Min      1Q  Median      3Q     Max
-128.41  -30.03   -3.41   26.37  452.40

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 211.9363     4.5331  46.753  < 2e-16 ***
CURSMOKE     -2.1263     1.3536  -1.571    0.116
BMI           0.8466     0.1721   4.920 8.97e-07 ***
PREVMI        1.3375     6.3899   0.209    0.834
PREVCHD       1.6353     4.3548   0.376    0.707
PREVSTRK     -7.0327     8.1004  -0.868    0.385
PREVHYP      13.1050     1.5071   8.696  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 43.93 on 4357 degrees of freedom
  (70 observations deleted due to missingness)
Multiple R-squared:  0.03336,   Adjusted R-squared:  0.03203
F-statistic: 25.06 on 6 and 4357 DF,  p-value: < 2.2e-16
```

Regression equation:

$$Y = 211.9363 - 2.12(X1) + 0.8466(X2) + 1.3375(X3) + 1.6353(X4) - 7.0327(X5) + 13.1050(X6)$$

Interpretation:

Here we got intercept as 211.93 which is the expected cholesterol level by assuming all the predictors = 0

for all X1, X2, X3, X4,X5 and X6 the interpretations are as follows:

- Cholesterol level decreases by 2.1263 if any individual is a current smoker (X1)
- for every 1 unit increase in individual (X2)BMI, cholesterol decrease by 0.8466

- For chronic heart disease(X3), their cholesterol increase on an average by 1.3375

- If anyone had a heart choke in past (X4), then their cholesterol increases by 1.6353

- Any individual who had a stroke previously (X5), their cholesterol decreases by 7.0327

- Who had hypertension (X6) previously, cholesterol increases on an average by 13.1050

3b

```
filtered_data <- Heart_study %>%
        select('TOTCHOL', 'CURSMOKE', 'BMI', 'PREVMI', 'PREVCHD', 'PREVSTRK', 'PREVHYP') %>%
        filter(complete.cases(.))
```

```
dim(filtered_data)
```

```
4364  7
```

```
dim(Heart_study)
```

```
4434  39
```

Here Dimesion has filtered by 70 rows as mentioned

3c

Evaluate if the entire set of six independent variability contribute significantly to the prediction of Y . Write out the null and alternative hypothesis being tested and your conclusion.

Null Hypothesis: There is no significant by any predictor variables on the outcome totchol

Alternative Hypothesis: At least one predictor variable has significant impact on our outcome variable totchol

```r
model4 <- glm(TOTCHOL ~ ., data = filtered_data)
summary(model4)
```

```
Call:
glm(formula = TOTCHOL ~ ., data = filtered_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-128.41   -30.03    -3.41    26.37   452.40

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 211.9363     4.5331  46.753  < 2e-16 ***
CURSMOKE     -2.1263     1.3536  -1.571    0.116
BMI           0.8466     0.1721   4.920 8.97e-07 ***
PREVMI        1.3375     6.3899   0.209    0.834
PREVCHD       1.6353     4.3548   0.376    0.707
PREVSTRK     -7.0327     8.1004  -0.868    0.385
PREVHYP      13.1050     1.5071   8.696  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 1930.274)

    Null deviance: 8700437  on 4363  degrees of freedom
Residual deviance: 8410206  on 4357  degrees of freedom
AIC: 45409

Number of Fisher Scoring iterations: 2
```

In case of BMI p value is approximately 8.97, we reject our null hypothesis that the true intercept is 0.

It can be seen that p-value of the F-statistic is < 2.2e-16, which is highly significant. This means that, at least, one of the predictor variables is significantly related to the outcome variable.

TOTCHOL is statistically impacted by an individual BMI. For every 1 unit increase in BMI, totchol increases on an average of 0.81 units

3d

Null Hypothesis: Additional coefficients implicated in model will be equal to 0 or having no impact on our predictor variable totchol

Alternative Hypothesis: At least one of the 4 cardiovascular conditions coefficient is not 0

```r
model_c <- glm(TOTCHOL ~ BMI + CURSMOKE, data = filtered_data)
summary(model_c)
```

```
Call:
glm(formula = TOTCHOL ~ BMI + CURSMOKE, data = filtered_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-128.35   -30.59    -3.66    27.02   462.19

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 205.4060     4.5068  45.577  < 2e-16 ***
BMI           1.2780     0.1662   7.691 1.79e-14 ***
CURSMOKE     -2.8298     1.3617  -2.078   0.0378 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 1963.086)

    Null deviance: 8700437  on 4363  degrees of freedom
Residual deviance: 8561019  on 4361  degrees of freedom
AIC: 45479

Number of Fisher Scoring iterations: 2
```

Interpretation:

Here P value is nearly 6.021, so we are rejecting Null hypothesis

```r
all_d <- glm(TOTCHOL ~ ., data = filtered_data)
```

```r
anova (all_d, model_c, test = "F")
```

Resid. Df	Resid. Dev	Df	Deviance	F	Pr(>F)
4357	8410206	NA	NA	NA	NA
4361	8561019	-4	-150813.6	19.53266	6.021663e-16

$$F = \frac{[SS_{model}(full) - SS_{model}(reduced)]/k}{MS_{error}(full)} = \frac{[290231.2 - 139417.6]/4}{1930.27} = 19.53271$$

As null hypothesis is rejected so we can say that the 4 cardiovascular conditions significantly contributes to our predictor outcome **totchol**.

$p<0.001$ so we reject H_0 and conclude that at least one β_k is not 0. We should keep this set of four variables in the model since they contribute significantly to a model that already includes smoking status and BMI.

3e

Null Hypothesis: Current smoker coefficient will be 0 and will not have any effect on the predicting outcome variable totchol.

Alternative Hypothesis: current smoker coefficient can't be 0 and current smoker variable will have effect on predicting out come variable totchol.

```
smoke_m <- glm(TOTCHOL ~  BMI + PREVMI +PREVCHD + PREVSTRK+ PREVHYP, data = filtered_data)
anova(all_d, smoke_m, test = "F")
```

Resid. Df	Resid. Dev	Df	Deviance	F	Pr(>F)
4357	8410206	NA	NA	NA	NA
4358	8414969	-1	-4763.093	2.467573	0.1162899

round(summary(all_d)$coefficients, 6)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	211.936264	4.533141	46.752632	0.000000
CURSMOKE	-2.126312	1.353605	-1.570851	0.116290
BMI	0.846552	0.172061	4.920081	0.000001
PREVMI	1.337489	6.389908	0.209313	0.834214
PREVCHD	1.635278	4.354847	0.375508	0.707301
PREVSTRK	-7.032683	8.100367	-0.868193	0.385336
PREVHYP	13.104980	1.507081	8.695604	0.000000

Here CURSOKE, $t=-1.571$ and $p=0.116$. $p>0.05$, we fail to reject H_0 and we cannot conclude that it contributes significantly to our model above and beyond the other independent variables. We cannot conclude the slope is significantly different from 0.

By adding smoking status will not have any effect on the predicting outcome variable totchol.

3f.

```
vif(all_d)
```

CURSMOKE	1.03526311567343
BMI	1.12330768133091
PREVMI	1.78330118729897
PREVCHD	1.80338122021788
PREVSTRK	1.01278156229332
PREVHYP	1.12138021885526

Here VIF > 10, there is collinearity in data

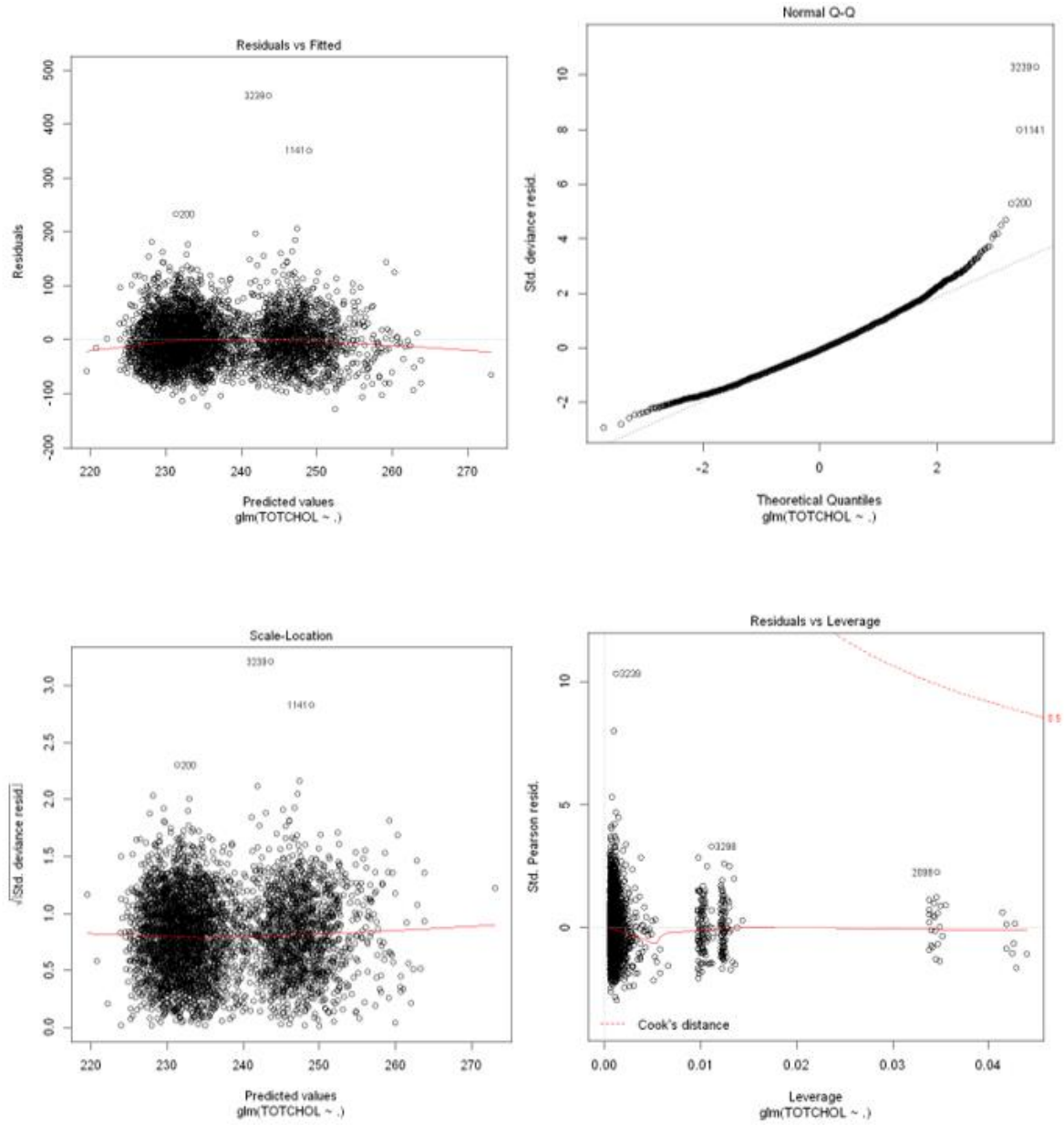
But here all the VIFs are varying from 1.01 to 1.8, hence no collinearity in data (variables are not correlated linearly with each other)

3g.

```
1 summary(all_d)
```

```
Call:
glm(formula = TOTCHOL ~ ., data = filtered_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-128.41   -30.03    -3.41    26.37   452.40

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 211.9363     4.5331  46.753  < 2e-16 ***
CURSMOKE     -2.1263     1.3536  -1.571    0.116
BMI           0.8466     0.1721   4.920 8.97e-07 ***
PREVMI        1.3375     6.3899   0.209    0.834
PREVCHD       1.6353     4.3548   0.376    0.707
PREVSTRK     -7.0327     8.1004  -0.868    0.385
PREVHYP      13.1050     1.5071   8.696  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 1930.274)

    Null deviance: 8700437  on 4363  degrees of freedom
Residual deviance: 8410206  on 4357  degrees of freedom
AIC: 45409

Number of Fisher Scoring iterations: 2
```

plot(all_d)



Here individual data points correlation has been shown