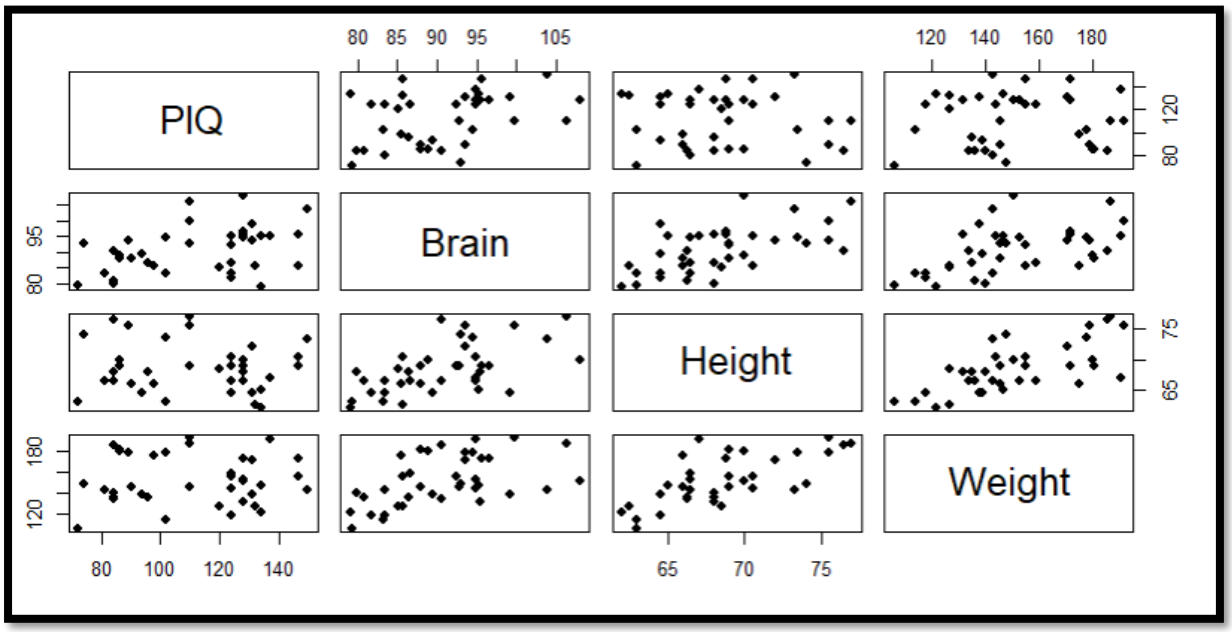
[Q1]

[Ans a]

```
head(iq)
```

```
  PIQ  Brain Height Weight
1 124  81.69   64.5    118
2 150 103.84   73.3    143
3 128  96.54   68.8    172
4 134  95.15   65.0    147
5 110  92.88   69.0    146
6 131  99.13   64.5    138
```

```
pairs(iq[,1:4], pch = 19)          # pch to increase data points size
```



The scatter plot doesn't seem to show linear relationship between PIQ and the variables Height and Weight. We observe the PIQ and Brain variables showing some positive association as data points are scattered in an uphill pattern even-though there are outliers. Since there are no curved patterns variables will not show non-linear relationships. However, we cannot confirm this by looking at the scatter plot alone.

Finding correlation coefficient value

To understand the quantified linear relationship between variables (only for analysis).

```
cor(iq, method = 'pearson')
```

```
                PIQ     Brain      Height      Weight
PIQ     1.000000000 0.3778155 -0.09315559 0.002512154
Brain   0.377815463 1.0000000  0.58836684 0.513486971
Height -0.093155590 0.5883668  1.00000000 0.699614004
Weight  0.002512154 0.5134870  0.69961400 1.000000000
```

The Pearson correlation coefficient value shows almost no linear relationship between PIQ and Height as well as between PIQ and Weight. But the association between PIQ and Height is greater than between PIQ and Weight. As the value of Height increases the PIQ rate will decrease. The linear relationship between PIQ and Brain is not strong. There is a positive association between the two variables so as values of Brain increases the PIQ would also increase.

Correlation test

Before performing hypothesis test we need to check the assumptions of the linearity and normality of the variables.

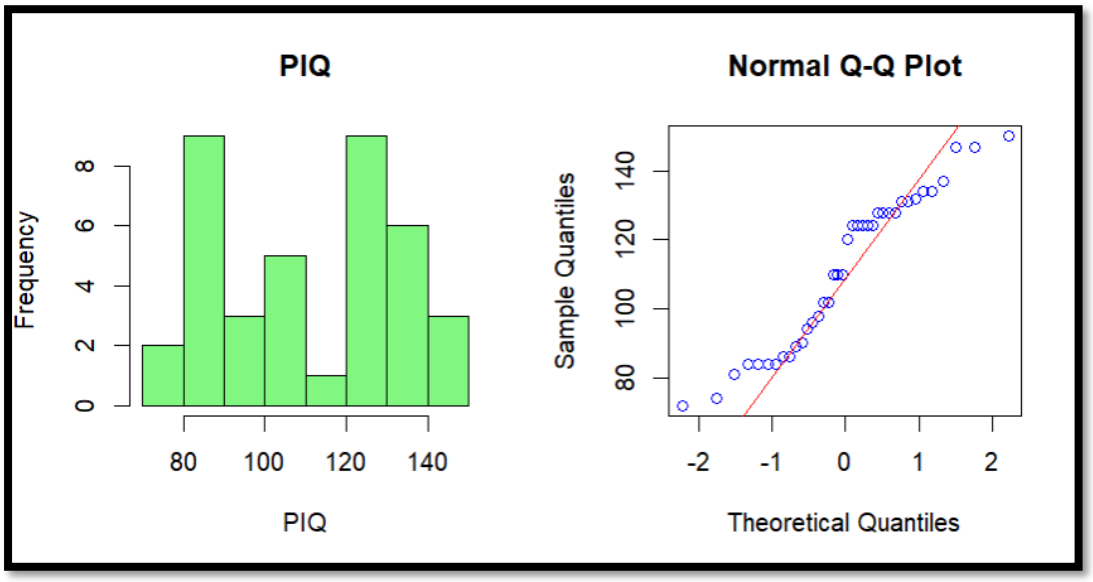
Linearity

From the scatter plot we observe the PIQ and Brain variables showing some positive association as data points are scattered in an uphill pattern even-though there are outliers.

Normality

- **PIQ variable**

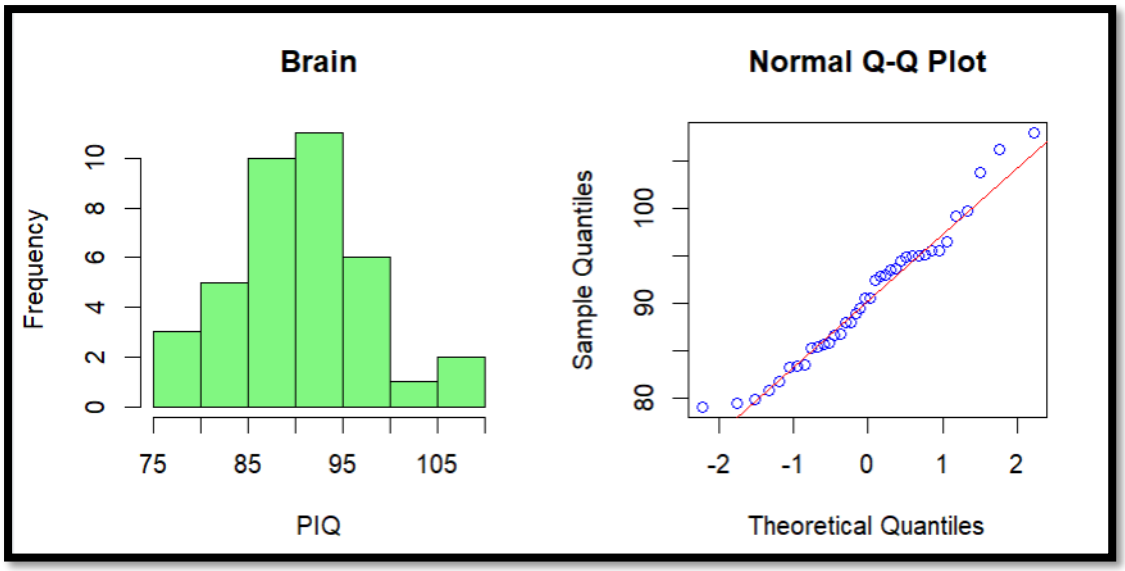
```
hist(iq$PIQ,breaks =10, main="PIQ", xlab="PIQ", col="#81F781")

qqnorm(iq$PIQ, col="blue")

qqline(iq$PIQ, col="red")

par(mfrow=c(1,2))
```



The histogram doesn’t look symmetric or have a bell shape. A proportion of data points do not lie on the straight line in the normal Q-Q plot so the data doesn't seem to follow a normal distribution.

- **Brain variable**

```
hist(iq$Brain,breaks =10, main="Brain", xlab="PIQ", col="#81F781")

qqnorm(iq$Brain, col="blue")

qqline(iq$Brain, col="red")

par(mfrow=c(1,2))
```



The histogram looks to have a bell shape. Majority of the data points lie on the straight line in the normal Q-Q plot so the data seem to follow a normal distribution.

Performing hypothesis testing to check the normality of the data

We formulate our hypothesis as follows:

Ho: sample drawn from normally distributed population

Ha: sample not drawn from normally distributed population

Using shapiro-wilk normality test to check our hypothesis

```
with(iq,shapiro.test(PIQ))
```

```
with(iq,shapiro.test(Brain))
```

```
> with(iq,shapiro.test(PIQ))

        Shapiro-Wilk normality test

data:  PIQ
W = 0.93157, p-value = 0.02249

> with(iq,shapiro.test(Brain))

        Shapiro-Wilk normality test

data:  Brain
W = 0.96536, p-value = 0.2822
```

The p-value for PIQ is significant ($p<0.05$) so we reject the null hypothesis and accept the alternate hypothesis. Thus, there is evidence that PIQ variable is not normally distributed.

Thus, we checked our assumptions and now check the correlation between PIQ and Brain

Hypothesis testing to check the relationship between PIQ and Brain

We formulate our hypothesis as follows:

Ho: no statistically significant linear relationship between PIQ and Brain

Ha: there is statistically significant linear relationship between PIQ and Brain

Since the normality of the data is not met we use spearman correlation method

```
with(iq, cor.test(Brain,PIQ, method = "spearman", exact=FALSE))
```

```
        Spearman's rank correlation rho

data:  Brain and PIQ
S = 5367.9, p-value = 0.01004
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.4126407
```

The Spearman correlation is 0.4126.

The p-value is significant($p<0.05$) and so we reject the null hypothesis and accept the alternate hypothesis. Thus, there is evidence that there is linear relationship between PIQ and Brain.

Conclusion

We conclude that there is an increasing linear association between PIQ and Brain. That is, as the Brain value increases, the PIQ rate increases.

[Ans b]

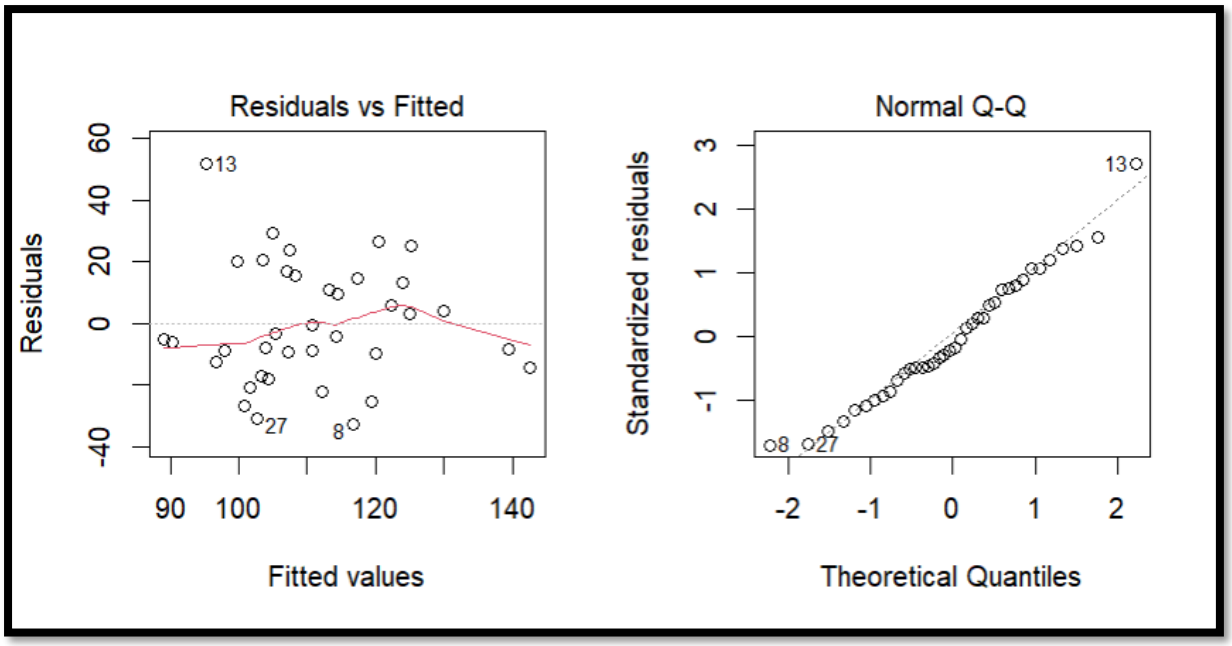
Multiple Regression is the appropriate model as it uses several explanatory variables to predict the outcome of response variable (PIQ)

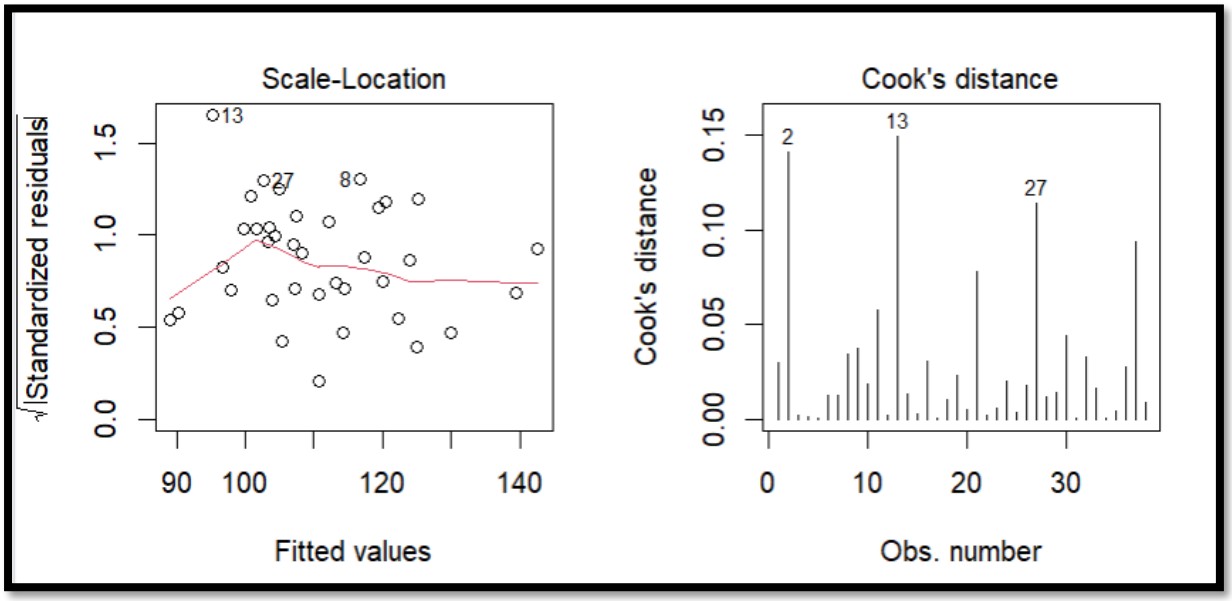
Assumptions:

1. Linearity. The mean of the response variable is linear in the unknown parameters and the explanatory variables. The linearity is not required between the response variable and the explanatory variables as the latter are treated as fixed values.
2. The variation of the random error is constant which does not depend on the corresponding observed values of dependent variables.
3. The random error is normally distributed.
4. The observations are independent

Checking assumptions for the full model

```
m.full <- with(iq, lm(PIQ ~ Brain + Height + Weight))
```





Residual plot: (Residuals vs Fitted)

The model is good as we observe a random scatter around the dotted black horizontal line, correlation between residuals and fitted values close to 0 and the assumption of constant variance is met by this plot.

Normal Q-Q plot:

The assumption of normality of random errors is met as almost majority of data points lie on the normal line.

Standardised residual plot: (Scale-Location)

Very few data points are above the threshold value of 1.4. This plot again suggests that the model is good.

Linearity is checked by the scatter plot from part-a.

Thus, assumptions are satisfied and we can use multiple regression model.

Full Model

Hypothesis for F-test

Ho: none of the explanatory variables affects the outcome (PIQ)

Ha: at-least one of the explanatory variable affects the outcome (PIQ)

Hypothesis for t-test

Ho: $\beta i = 0$

Ha: $\beta i \neq 0$

Multiple Linear Regression model

```
m.full <- with(iq, lm(PIQ ~ Brain + Height + Weight))
summary(m.full)
```

```
Call:
lm(formula = PIQ ~ Brain + Height + Weight)

Residuals:
   Min     1Q Median     3Q    Max
-32.74 -12.09  -3.84  14.17  51.69

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.114e+02  6.297e+01   1.768 0.085979 .
Brain        2.060e+00  5.634e-01   3.657 0.000856 ***
Height      -2.732e+00  1.229e+00  -2.222 0.033034 *
Weight       5.599e-04  1.971e-01   0.003 0.997750
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.79 on 34 degrees of freedom
Multiple R-squared:  0.2949,    Adjusted R-squared:  0.2327
F-statistic: 4.741 on 3 and 34 DF,  p-value: 0.007215
```

We reject the null hypothesis as p-value of the F-statistic is significant (p-value<0.05). Thus, there is evidence that at-least one of the explanatory variable affects the outcome.

p-values for Brain and Height is significant so these variables should remain in the final model. Thus, we reject the null hypothesis for t-test.

p-value for Weight is non-significant so this variable could be removed by further model selection.

Null Model

```
m.null <- with(iq, lm(PIQ ~ 1))
```

Final model by stepwise AIC model selection

By forward selection we start with null model and keep on adding explanatory variables one at a time

```
stepAIC(m.null,scope=list(lower=m.null,upper=m.full),direction="forward",
k=2)
```

```
Start:  AIC=237.94
PIQ ~ 1

         Df Sum of Sq   RSS    AIC
+ Brain   1   2697.09 16198 234.09
<none>                18895 237.94
+ Height  1    163.97 18731 239.61
+ Weight  1      0.12 18894 239.94

Step:  AIC=234.09
PIQ ~ Brain

         Df Sum of Sq   RSS    AIC
+ Height  1   2875.65 13322 228.66
+ Weight  1    940.94 15256 233.82
<none>                16198 234.09

Step:  AIC=228.66
PIQ ~ Brain + Height

         Df Sum of Sq   RSS    AIC
<none>                13322 228.66
+ Weight  1 0.0031633 13322 230.66

Call:
lm(formula = PIQ ~ Brain + Height)

Coefficients:
(Intercept)        Brain       Height
    111.276        2.061       -2.730
```

The final model selected is:

```
m.final <- with(iq, lm(PIQ ~ Brain + Height))
```

```
m.final
```

Thus, the final model includes Brain and Height as the independent variables.

Interpretation

```
summary(m.final)
```

```
Call:
lm(formula = PIQ ~ Brain + Height)

Residuals:
    Min      1Q  Median      3Q     Max
-32.750 -12.090  -3.841  14.174  51.690

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 111.2757    55.8673   1.992 0.054243 .
Brain         2.0606     0.5466   3.770 0.000604 ***
Height       -2.7299     0.9932  -2.749 0.009399 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.51 on 35 degrees of freedom
Multiple R-squared:  0.2949,    Adjusted R-squared:  0.2546
F-statistic: 7.321 on 2 and 35 DF,  p-value: 0.002208
```

The multiple regression model equation can be represented as:

PIQ = 111.276 + 2.06*Brain + -2.730*Height + e

From the estimate column we understand that for every unit increase in Brain there is an associated 2.06 unit increase in PIQ, and that for every unit increase in Height there is an associated -2.73 unit decrease in PIQ.

Hypothesis for F-test

Ho: none of the explanatory variables affects the outcome (PIQ)

Ha: at-least one of the explanatory variable affects the outcome (PIQ)

Hypothesis for t-test

Ho: $\beta i = 0$

Ha: $\beta i \neq 0$

The p-value of the F-statistic is significant($p\text{-value}<0.05$) hence we reject the null hypothesis and accept the alternate hypothesis. Thus, there is evidence that at-least one the variable affects PIQ. The significant p-values of the t-statistics

of the estimated coefficients again leads us to reject null hypothesis and conclude that both Brain and Height likely influences PIQ.

The final selected model explains relatively low percent (29.5%) of variation, which is not good. This indicates that the regression model fits the observations poorly which means that Brain and Height are not explaining much in the variation of PIQ.

[Ans c]

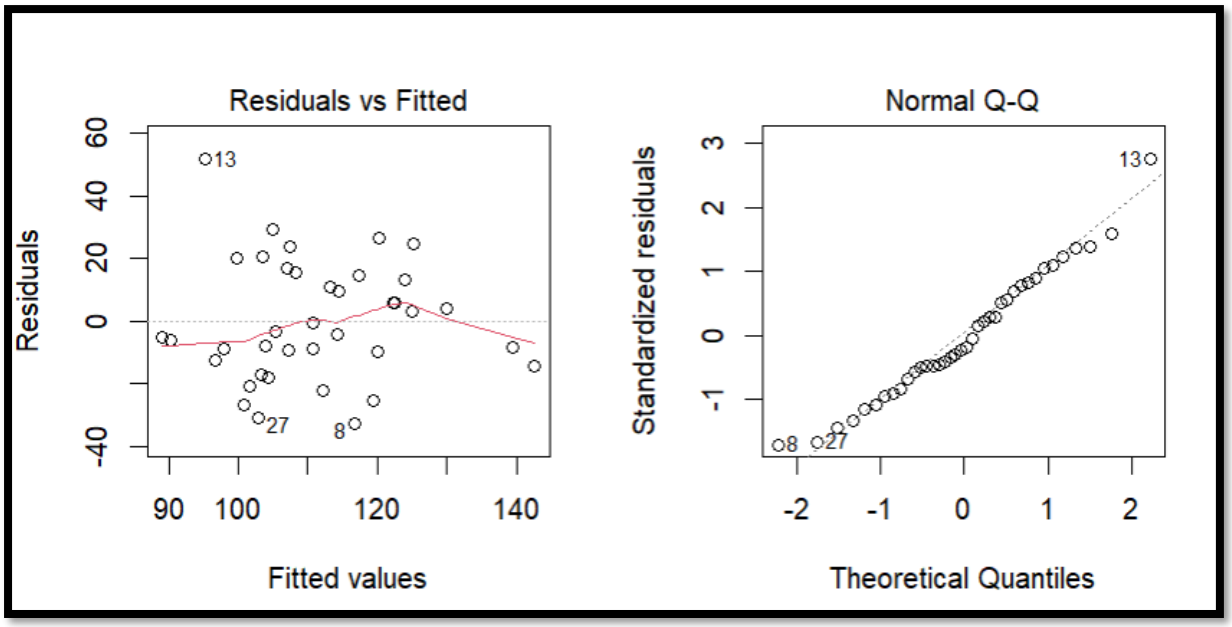
Diagnostics

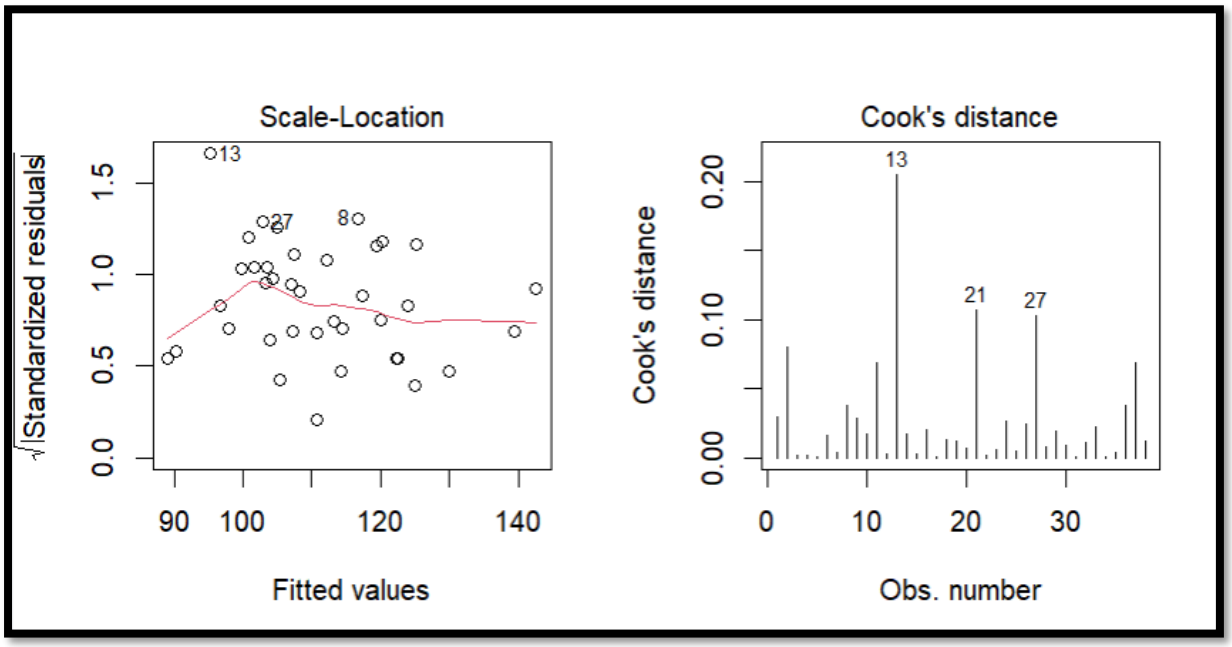
Assumptions:

1. Linearity. The mean of the response variable is linear in the unknown parameters and the explanatory variables. The linearity is not required between the response variable and the explanatory variables as the latter are treated as fixed values.
2. The variation of the random error is constant which does not depend on the corresponding observed values of dependent variables.
3. The random error is normally distributed.
4. The observations are independent

Checking assumptions for the final model

```
plot(m.final, which = 1:4)
```





Residual plot: (Residuals vs Fitted)

The model is good as we observe a random scatter around the dotted black horizontal line, correlation between residuals and fitted values close to 0 and the assumption of constant variance is met by this plot.

Normal Q-Q plot:

The assumption of normality of random errors is met as almost majority of data points lie on the normal line.

Standardised residual plot: (Scale-Location)

Very few data points are above the threshold value of 1.4. This plot again suggests that the model is good.

Linearity is checked by the scatter plot from part-a

Cooks Distance

From this plot it seems observation 2,13 and 27 are influential.

Thus, assumptions are satisfied.

[Q2]

[Ans a]

[PC: Principal Component]

[PCA: Principal Component Analysis]

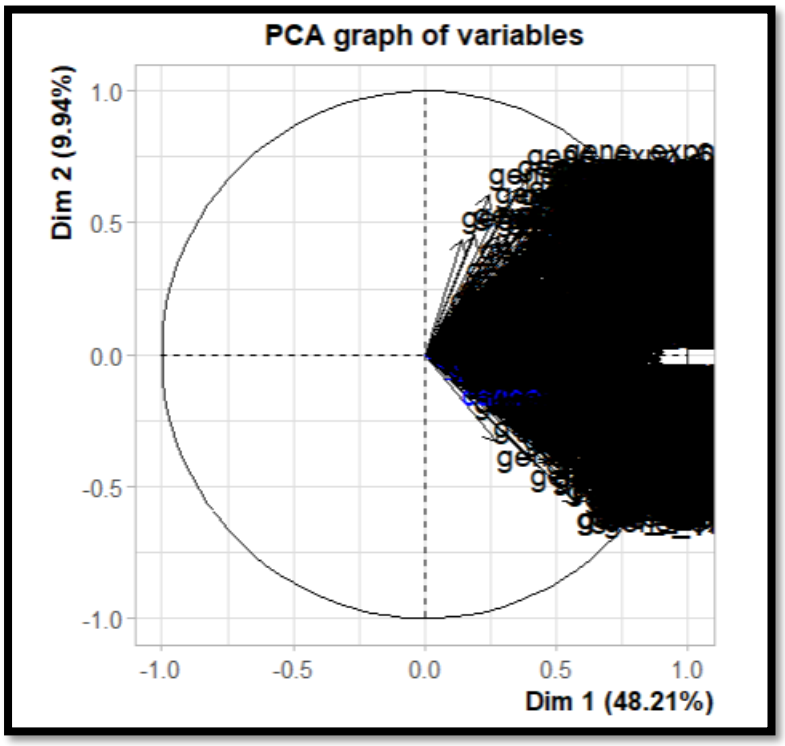
```
head(d_cancer)
```

```
# A tibble: 6 x 855
  cancer gene_expn_1 gene_expn_2 gene_expn_3 gene_expn_4 gene_expn_5 gene_expn_6 gene_expn_7
   <dbl>       <dbl>       <dbl>       <dbl>       <dbl>       <dbl>       <dbl>       <dbl>
1      1        259.        282.        242.        161.         99.9        74.1        144.
2      0        497.        339.        739.        399.        118.        246.        181.
3      1        262.        160.        242.        133.        231.        177.        349.
4      0        164.        163.         75.3       100.        271.        261.        246.
5      1        263.        465.        370.        868.        189.        206.        325.
6      0        399.        335.        138.        245.        154.        343.        287.
```

Performing principal component analysis on gene expression data

Excluding cancer column (Dependent variable) as we are interested to study the reason for cancer according to the genome types so we perform PCA on the dataset containing only genome types.

```
pca = PCA(d_cancer, quanti.sup = 1)
```



From the above plot we see all the variables contribute to PC1. A total of 58.15% [Dim.1(42.21%) + Dim.2(9.94%)] of variance is captured in the entire dataset by PC1 and PC2.

```
summary(pca)
```

- Variables

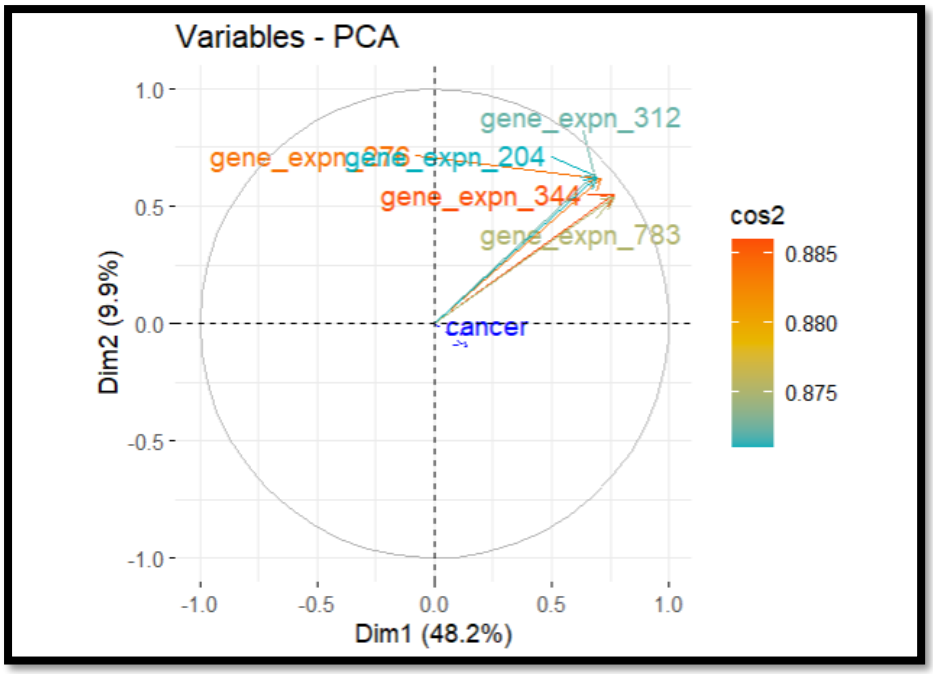
```
Variables (the 10 first)
               Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2
gene_expn_1   |  0.539  0.071  0.290 |  0.309  0.112  0.095 |  0.325  0.179  0.105 |
gene_expn_2   |  0.890  0.193  0.793 | -0.066  0.005  0.004 | -0.101  0.017  0.010 |
gene_expn_3   |  0.632  0.097  0.400 |  0.455  0.244  0.207 |  0.005  0.000  0.000 |
gene_expn_4   |  0.521  0.066  0.272 |  0.480  0.272  0.231 |  0.031  0.002  0.001 |
gene_expn_5   |  0.713  0.123  0.508 | -0.089  0.009  0.008 |  0.132  0.029  0.017 |
gene_expn_6   |  0.666  0.108  0.443 | -0.334  0.131  0.112 |  0.292  0.145  0.085 |
gene_expn_7   |  0.796  0.154  0.633 | -0.346  0.141  0.119 |  0.283  0.136  0.080 |
gene_expn_8   |  0.837  0.170  0.701 | -0.191  0.043  0.036 | -0.122  0.025  0.015 |
gene_expn_9   |  0.859  0.179  0.738 | -0.235  0.065  0.055 |  0.116  0.023  0.013 |
gene_expn_10  |  0.663  0.107  0.440 |  0.594  0.416  0.353 |  0.086  0.012  0.007 |

Supplementary continuous variable
               Dim.1   cos2    Dim.2   cos2    Dim.3   cos2
cancer        |  0.138  0.019 | -0.097  0.009 | -0.223  0.050 |
```

Dim column values are basically the coefficients in the linear combination between the variables and the specific principal component.

A high ‘cos2’ value indicates a good representation of the variable on the principal component. From the above information we understand that majority of the variables are very well represented PC1.

Correlation Circle

```
fviz_pca_var(pca, col.var = "cos2", select.var= list(cos2 = 5),gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"), repel = TRUE)
```



The above plot indicates top 5 variables with high 'cos2' values. The longer lines indicate variables are positively correlated and are well represented by the first two dimensions.

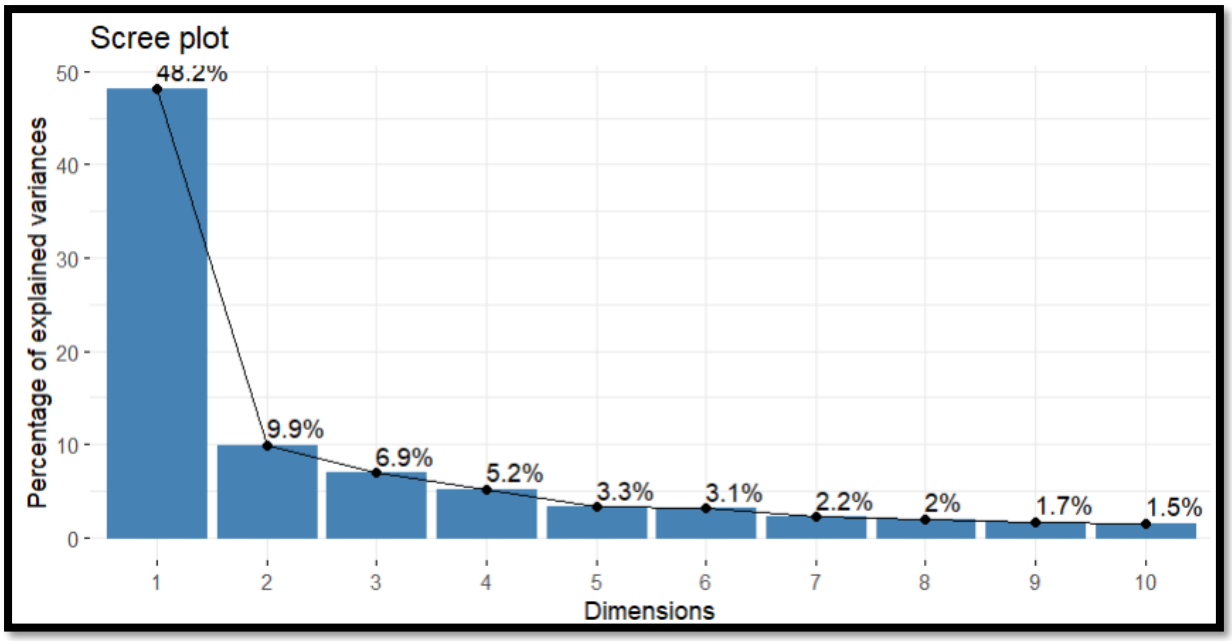
The 'ctr' column tells us the importance of each variable in constructing the PC. For PC1 'gene_expn_2' variable has high importance.

The purpose of PCA is to reduce the dimensionality. We need to choose the best principal components that explains the maximum variance in the data.

Below are the steps to choose the best principal components

Screeplot:

```
fviz_screeplot(pca, addlabels = TRUE)
```



According to Scree plot we retain PCs to the left of the elbow point (at PC2). So, we have to retain PC1 as it explains 48% variance in the data.

Eigen values

pca$eig

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	411.6765662	48.20568691	48.20569
comp 2	84.8539308	9.93605747	58.14174
comp 3	59.0389711	6.91322847	65.05497
comp 4	44.5468350	5.21625703	70.27123
comp 5	27.8093861	3.25636840	73.52760
comp 6	26.6147098	3.11647655	76.64407
comp 7	18.9135913	2.21470624	78.85878
comp 8	17.1949944	2.01346539	80.87225
comp 9	14.4301611	1.68971441	82.56196
comp 10	13.1375683	1.53835695	84.10032
comp 11	11.3491145	1.32893613	85.42925
comp 12	9.6694947	1.13225933	86.56151
comp 13	8.4552952	0.99008140	87.55159
comp 14	7.4969450	0.87786241	88.42946
comp 15	6.9470718	0.81347445	89.24293
comp 16	6.4584301	0.75625645	89.99919
comp 17	6.3947371	0.74879826	90.74799
comp 18	5.3441741	0.62578151	91.37377

Eigenvalues measure the amount of variation by each Principal Component.

Principal Components from 1 to 17 explains 90% variation in the data (from last column) and their Eigenvalues are greater than 1, so we would like to retain them.

Thus, we performed PCA, that is by using PCA we have reduced the dimensionality and decided that 854 variables to be represented by only 17 principal components while losing about 10% variance.

I think 17 principal components would give an adequate representation of variation of the gene expression data as they are capable to capture about 90% variability in data so majority of the information is captured.

[Ans b]

We would like to predict the outcome (cancer) by using the selected principal components as our predictors.

Will use Logistic Regression model to predict the cancer status as the outcome variable(cancer) is dichotomous (output either 1 or 0).

Individuals co-ordinates

```
pca$ind$coord
```

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
1	-19.2933169	8.69185155	-2.50929105	-2.61897108	2.7130648
2	-12.2718507	13.78553411	1.26085415	1.48559112	4.8297294
3	-16.2895790	-3.65096743	-1.72810725	-0.15675790	-3.6039394
4	-22.8999965	-4.73198477	-0.82673468	-2.94736850	-2.5578427
5	-11.3946718	2.39578415	-8.10422029	0.47927588	3.3533886

Assumptions for using the model

1. Y1, ..., Yn are independent.
2. Yi not normally distributed, usually follows a distribution from the exponential family like binomial or Poisson.
3. The relationship between the transformed response by the link and the explanatory variables is linear.

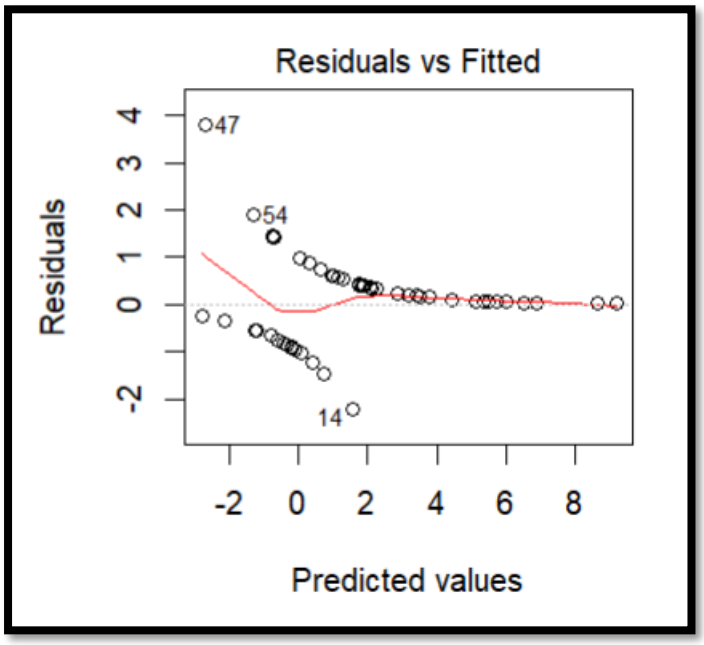
4. Random errors are independent but not normally distributed.
5. Uses MLE rather than OLS to estimate parameters, so relies on large sample approximations.

Logistic Regression

```
m1 = glm(d_cancer$cancer ~ pca$ind$coord, family = 'binomial')
```

Passing first 5 principal components as the predictor variables. Using family as binomial as dependent variable takes value either 0 or 1.

```
plot(m1)
```



The above plot shows the sigmoid function through the upper cluster but not through the lower cluster. However, our model is capable for classification but there would be incorrect predictions.

Thus, our model is not efficient in predicting the cancer status with exact accuracy, this could be because our model is being trained using only 5 principal components so there is much loss of information. If we train our model with all selected 17 principal component's we would get better results.

summary(m1)

```
Call:
glm(formula = d_cancer$cancer ~ pca$ind$coord, family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8842  -0.7264   0.1085   0.5663   2.3394

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         1.62871    0.62777   2.594  0.00947 **
pca$ind$coordDim.1  0.06312    0.03391   1.861  0.06269 .
pca$ind$coordDim.2 -0.04753    0.06514  -0.730  0.46555
pca$ind$coordDim.3 -0.11151    0.07622  -1.463  0.14350
pca$ind$coordDim.4  0.33608    0.10726   3.133  0.00173 **
pca$ind$coordDim.5 -0.03182    0.08927  -0.356  0.72154
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 71.743  on 55  degrees of freedom
Residual deviance: 45.584  on 50  degrees of freedom
AIC: 57.584

Number of Fisher Scoring iterations: 6
```

Hypothesis:

Ho: none of the explanatory variables will affect the log odds of cancer.

Ha: at-least one of the explanatory variable will affect the log odds of cancer.

It appears that some p-values are non-significant.

The key point to observe is that ‘cancer’ is driven largely by PC4.

The p-value of the estimated coefficient of 'pcaindcoordDim.4 (PC4)’ is significant, hence it will affect the log odds of cancer. Here we reject the null hypothesis.

The estimated coefficients are in the terms of log odds

In other words, for a one-unit increase in pcaindcoordDim.4, the expected change in log odds is 0.336. The odds of cancer = exp(0.336) = 1.399. Odds

of cancer when pcaindcoordDim.4 = 1 is 1.399 times the odds of cancer when pcaindcoordDim.4 = 0.

Usefulness of PCA on this dataset:

1. PCA helped in overcoming data overfitting by decreasing the number of features from 854 to 17.
2. It would have been impossible to visualise the original dataset of 854 features. However, by using PCA we could capture about 58.142% variance in the dataset by just using 2 principal components and were able to visualise and interpret it successfully.
3. We were able to exclude the unnecessary predictors in the dataset which have no impact on cancer variable that would eventually be unhelpful for the prediction using regression model.
4. Standard regression assumes that the number of patients exceeds the number of predictors. In this dataset we would always need greater than 854 patients. When this is not the case, the model will fail to provide a solution. This situation is completely avoided by using PCA.
5. Since principal components are orthogonal to each other correlation between them are always zero, so it helps in avoiding multi-collinearity problem that occurs when many variables in the dataset are highly correlated.

Q3

[Ans a]

1.

T = Result of test being positive

$\overline{T}$ = Result of test being negative

$\overline{D}$ = Patient does not have the disease

D = Patient really has the disease

P[T] = probability of the test conducted on a random individual yields a positive result.

P[T] = P[T|D] × P[D] + P[T|$\overline{D}$] × P[$\overline{D}$]

Sensitivity = P[T|D] = 0.95

Specificity = P[$\overline{T}$|$\overline{D}$] = 0.89

P[T|$\overline{D}$] = 1 – P[$\overline{T}$|$\overline{D}$] = 1- 0.89 = 0.11

P[D] = 0.03

P[$\overline{D}$] = 1-0.03 = 0.97

P[T] = [0.95 × 0.03 + 0.11 × 0.97] = 0.1352

P[T] = 13.52%

Hence shown that probability of the test conducted on a random individual yields a positive result with probability 13.52%.

2.

By using Bayes Theorem

$P[D|T] = {P[T|D] \times P[D]} / {P[T]}$

$P[D|T] = 0.95 \times 0.03/0.1352$

P[D|T] = 0.2107

The probability that individual truly has the virus is 21.07%

[Ans b]

1.

Hypothesis testing:

sample size(n) = 1000;

true population proportion(p) = 0.135

np = 1000*0.135 = 135

n(1-p) = 1000(1-0.135) = 865

since both np and n(1-p) > 5 we can conduct one proportion test.

Since sample size > 30 according to Central Limit Theorem we assume the data given to be normally distributed.

Hypothesis testing for the proportion(p)

We formulate our hypothesis as follows:

Ho : p = 0.135

Ha : p != 0.135

As with sample means, we can apply the central limit theorem: the sampling distribution of a proportion is normal with mean equal to the true proportion p, and variance p(1-p)/p. Under the null hypothesis, therefore, the variance is by central limit theorem

 # Calculating variance

variance(v) = p(1-p)/n

v = 0.135*(1-0.135)/1000

v = 0.000116775

 # Calculating standard error

se = sqrt(v)

se = 0.01080625

 # Calculating the probability of the individuals who tested positive

p1 = 165/1000

p1 = 0.165

 # Calculating test-statistic

test_statistic = (p1-0.135)/se

test_statistic = 2.776172

 # Calculating p_value

P_value = 2*(1-pnorm(test_statistic))

P_value = 0.005500316

p-value is significant (p-value<0.05) hence we reject the null hypothesis. Thus, there is evidence that the population proportion is not equal to 0.135.

We can confirm our above results and check whether our hypothesis holds true or not by using prop.test:

```
prop.test(165, n= 1000, p= 0.135)
```

```
        1-sample proportions test with continuity correction

data:  165 out of 1000, null probability 0.135
X-squared = 7.4524, df = 1, p-value = 0.006335
alternative hypothesis: true p is not equal to 0.135
95 percent confidence interval:
 0.1428144 0.1898051
sample estimates:
    p
0.165
```

Here too p-value < 0.05 so we reject the null hypothesis.(R is making a 'continuity correction' so we are getting difference in P-value)

We estimate with 95% confidence that the population proportion is not in between confidence interval of the sample proportion i.e not in between 0.1428 and 0.1898

The probability of a test conducted on a random individual yields a positive is 0.135 (13.52%) which is calculated using a prevalence of 3%. Since we rejected the null hypothesis so we will also consider prevalence ≠3%.

2.

- According to me the 1000 individuals selected for testing would not be representing the actual population so this might hinder the results.
- The sample population could be too small compared to the actual true population. Small sample size could cause problems in generalising the actual results.
- There could have been biases while selecting the population like selection bias or the presence of confounders may have effected the result.