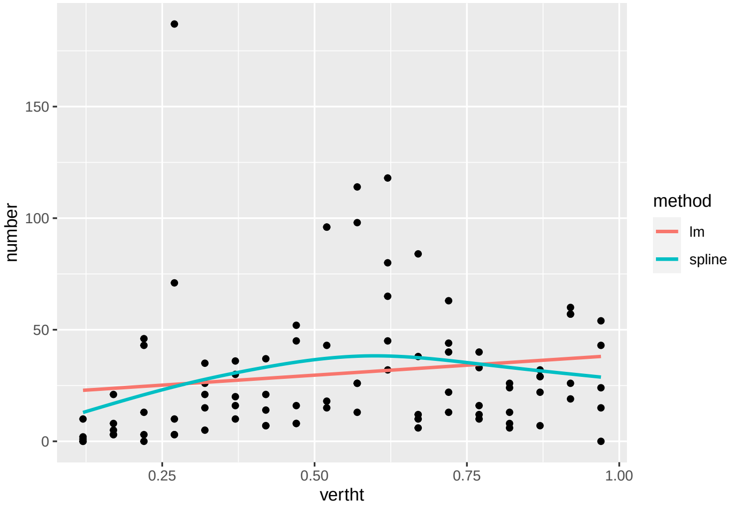
Assignment 2

Bilal Qureshi 46119043

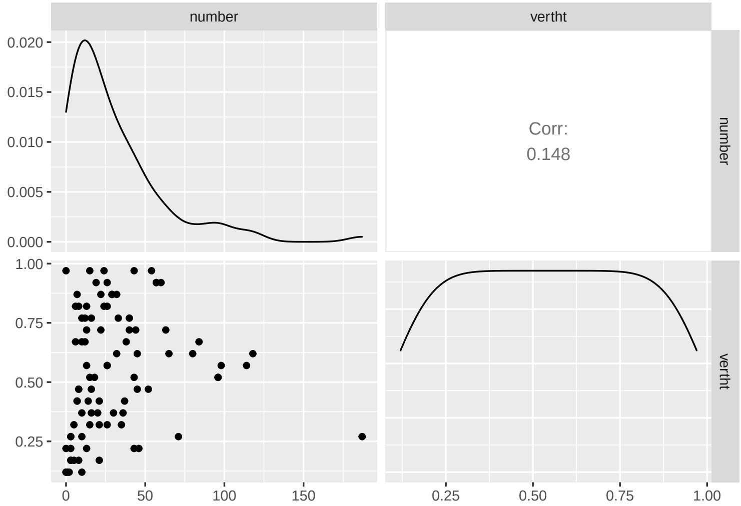
08/10/2021

Part a:

```
ggplot(tidal, aes(x = vertht, y = number)) +
  geom_point()+
geom_smooth(method = "lm", se = FALSE, aes(colour = "lm")) +
geom_smooth(method = "gam", se = FALSE, aes(colour = "spline")) +
scale_colour_discrete(name = "method")
```



```
ggpairs(tidal)
```



## R Markdown *Scatter plot* No evidence of linearity is found in the graph hence we can see that Poision or NB distribution is applicable here *Density curve* of vertht shows it is uniformly distributed and density curve of number shows occurences of organisms, inferring poisson distribution. *Correlation:* 0.148 coorelation found which tends to be weak positive relation

Therefore, vertht does not need any transformation for further analysis.

Including Plots

1b

You can also embed plots, for example:

```
model4 = glm(number ~ vertht, family = poisson(link = identity), data = tidal)
```

```
model5 = glm(number ~ vertht, family = poisson(link = log), data = tidal)
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

```
model6 = glm(number ~ vertht, family = poisson(link = sqrt), data = tidal)
```

```
broom::glance(model4) %>%add_case(broom::glance(model5)) %>%add_case(broom::glance(model6))%>%add_colum
```

```
## # A tibble: 3 x 9
##   link     null.deviance df.null logLik   AIC   BIC deviance df.residual  nobs
##   <chr>            <dbl>   <int>  <dbl> <dbl> <dbl>    <dbl>       <int> <int>
## 1 identity         2385.      89 -1358. 2721. 2726.    2299.          88    90
## 2 log              2385.      89 -1370. 2743. 2748.    2321.          88    90
## 3 sqrt             2385.      89 -1365. 2734. 2739.    2312.          88    90
```

For this dataset we will apply Poisson and NB Distribution because the variables number is a count and vertht is a countinous variable. Further for all 3 links we can over despersion. For all 3 (2298.961/88=26.12) for identity,(2321.307/88=26.37) for log, (2312.008/88=26.27) for sqrt. All dispression have values much larger then 2 hence we will omit Poisson distribution and go with Negative Binomial Distribution. This tells that these models have over dispursion.

```
modelnb2 = glm.nb(number ~ vertht, data = tidal,link = "log")
modelnb1 = glm.nb(number ~ vertht, data = tidal,link = "identity")
modelnb3 = glm.nb(number ~ vertht, data = tidal,link = "sqrt")
modelnbs3 = glm.nb(number ~ poly(vertht,2), data = tidal,link = "sqrt")
modelnbs2 = glm.nb(number ~ poly(vertht,2), data = tidal,link = "log")
```

```
 broom::glance(modelnb1)  %>%add_case(broom::glance(modelnb2))%>%add_case(broom::glance(modelnb3))%>%ad
```

```
## # A tibble: 3 x 9
##   link     null.deviance df.null logLik   AIC   BIC deviance df.residual  nobs
##   <chr>            <dbl>   <int>  <dbl> <dbl> <dbl>    <dbl>       <int> <int>
## 1 identity          109.      89  -396.  799.  806.     104.          88    90
## 2 log               107.      89  -397.  801.  808.     104.          88    90
## 3 sqrt              108.      89  -397.  800.  808.     104.          88    90
```

```
broom::glance(modelnbs2)  %>%add_case(broom::glance(modelnbs3))%>%add_column(link=c( "log","sqrt")) %>%
```

```
## # A tibble: 2 x 9
##   link  null.deviance df.null logLik   AIC   BIC deviance df.residual  nobs
##   <chr>         <dbl>   <int>  <dbl> <dbl> <dbl>    <dbl>       <int> <int>
## 1 log            119.      89  -392.  792.  802.     104.          87    90
## 2 sqrt           122.      89  -390.  788.  798.     103.          87    90
```

```
modelnbc2 = glm.nb(number ~ poly(vertht,3), data = tidal,link = "log")
modelnbc3 = glm.nb(number ~ poly(vertht,3), data = tidal,link = "sqrt")
broom::glance(modelnbc2)  %>%add_case(broom::glance(modelnbc3))%>%add_column(link=c( "log","sqrt")) %>%
```

```
## # A tibble: 2 x 9
##   link  null.deviance df.null logLik   AIC   BIC deviance df.residual  nobs
##   <chr>         <dbl>   <int>  <dbl> <dbl> <dbl>    <dbl>       <int> <int>
## 1 log            126.      89  -388.  787.  799.     103.          86    90
## 2 sqrt           129.      89  -387.  784.  796.     102.          86    90
```

We can see for above the lowest AIC models are for qubic model. Dispursion for all 3 models are close 1 and difference is as follows: $\Delta D = D_0 - D_1 \implies \Delta D < 1$ AIC for linear model(log and sqrt) is 800.9, Square equation: 788.4, cubic equation(sqrt): 783.8945 for . The lowest AIC is of cubic equation hene this will be selected based on model selection.

```
summary(modelnbc3)
```

```
##
## Call:
## glm.nb(formula = number ~ poly(vertht, 3), data = tidal, link = "sqrt",
##     init.theta = 1.313040942)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9053  -1.0259  -0.3404   0.2393   3.1617
##
## Coefficients:
```

```
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)          5.4075     0.2599  20.810  < 2e-16 ***
## poly(vertht, 3)1     5.1853     2.1335   2.430  0.01508 *
## poly(vertht, 3)2    -9.0032     2.2078  -4.078 4.54e-05 ***
## poly(vertht, 3)3     5.4787     2.1084   2.599  0.00936 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.313) family taken to be 1)
##
##     Null deviance: 129.06  on 89  degrees of freedom
## Residual deviance: 102.46  on 86  degrees of freedom
## AIC: 783.96
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.313
##           Std. Err.:  0.198
##
##  2 x log-likelihood:  -773.964
```

1c $NB(Y) \sim (\mu, \sigma)$

$\sqrt(Y_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3$

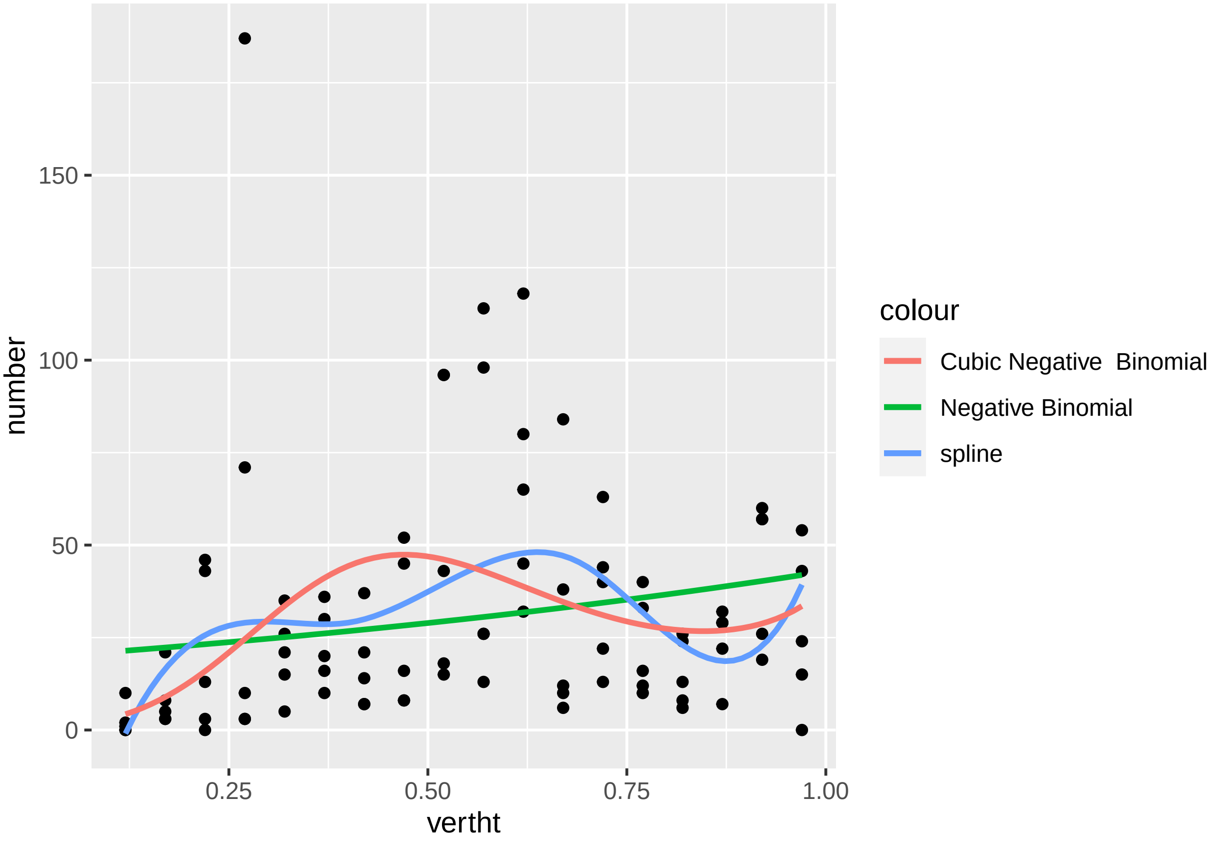
$Y_i = 2^{\beta_0} + 2^{\beta_1 x_i} + 2^{\beta_2 x_i^2} + 2^{\beta_3 x_i^3}$

$Y_i = 2^{5.4075} + 2^{5.1853 x_i} + 2^{-9.0032 x_i^2} + 2^{5.4787 x_i^3}$

$where, \quad Y_i = number, \quad x_i = vertht, \quad for\ i = for\ each\ observation\ 1, 2, 3...$

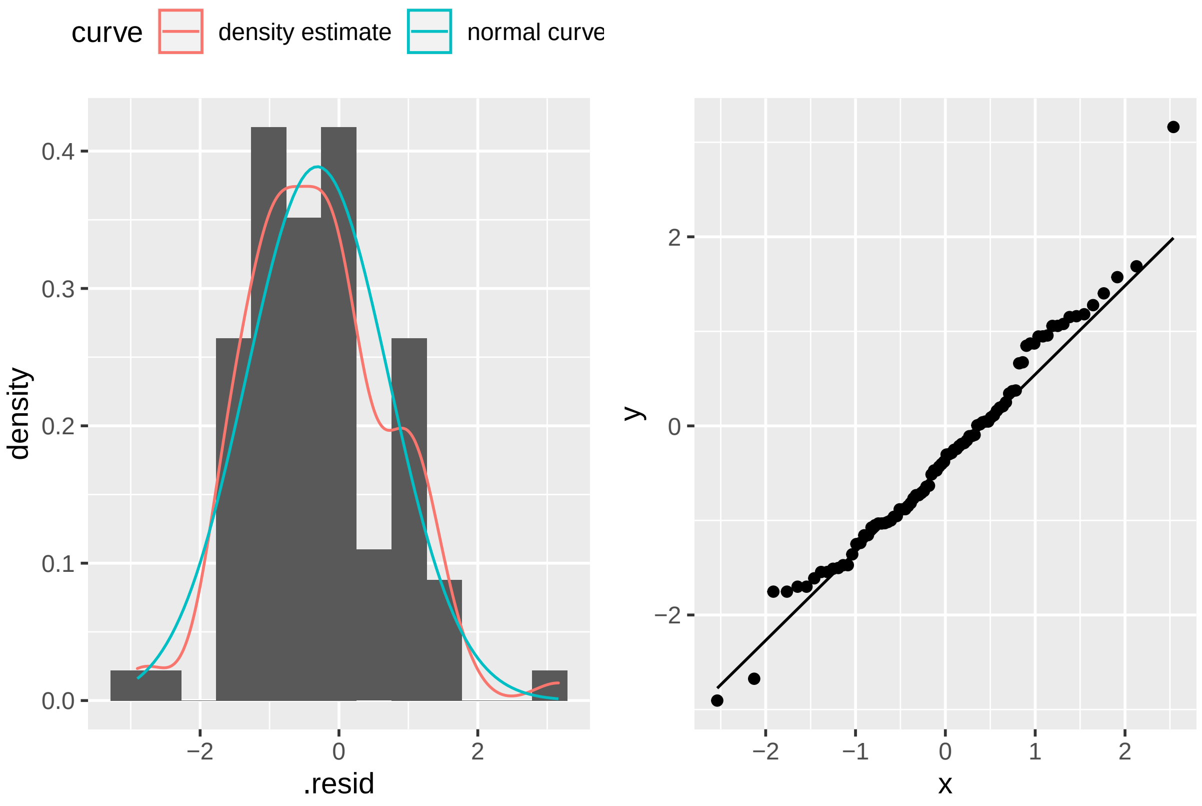
1d

```
library(ggplot2)
ggplot(tidal, aes(y = number, x = vertht)) +
  geom_point() +
  geom_smooth(aes(colour = "Negative Binomial"), method = "glm.nb",se=FALSE) +
  geom_smooth(aes(colour = "spline"),method = "lm", formula = y ~ splines::bs(x, 5), se = FALSE)+ geom_s
              formula = y~poly(x,3),se=FALSE)
```



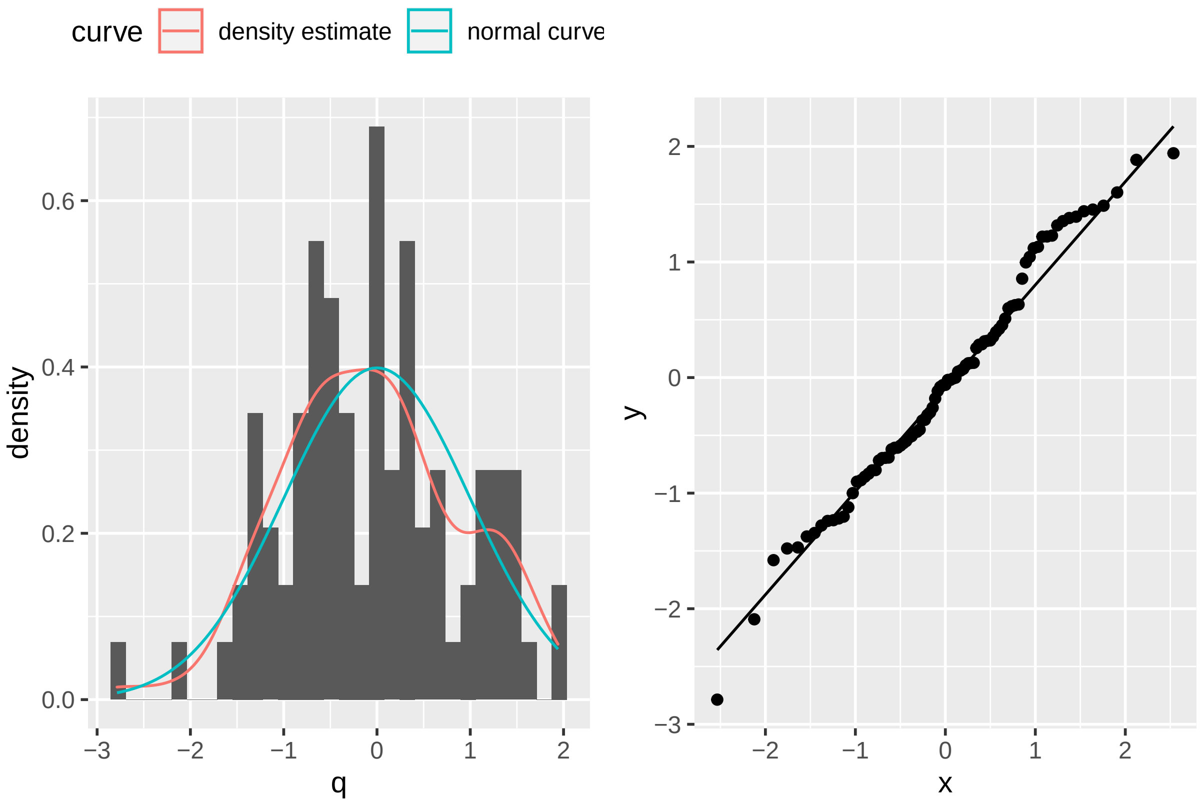
For the above model we can see that the Cubic Negative Binomial model is closely fitting with the spline model. This gives an indication, this shows that the data has more coverage while comparing it to the green line representing linear Negative Binomial. This tells the cubic model we selected is fitting properly.

```r
m11_diag <- broom::augment(modelnbc3)
p1 <- ggplot(m11_diag, aes(x = .resid)) +
  geom_histogram(bins = 13, aes(y = ..density..)) +
  geom_density(aes(colour = "density estimate")) +
  geom_function(aes(colour = "normal curve"),
                fun = dnorm,
                args = list(
                  mean = mean(m11_diag$.resid),
                  sd = sd(m11_diag$.resid)
                )
  ) +
  scale_colour_discrete(name = "curve") +
  theme(legend.position = "top")
p2 <- ggplot(m11_diag, aes(sample = .resid)) +
  geom_qq() +
  geom_qq_line()

p1+p2
```

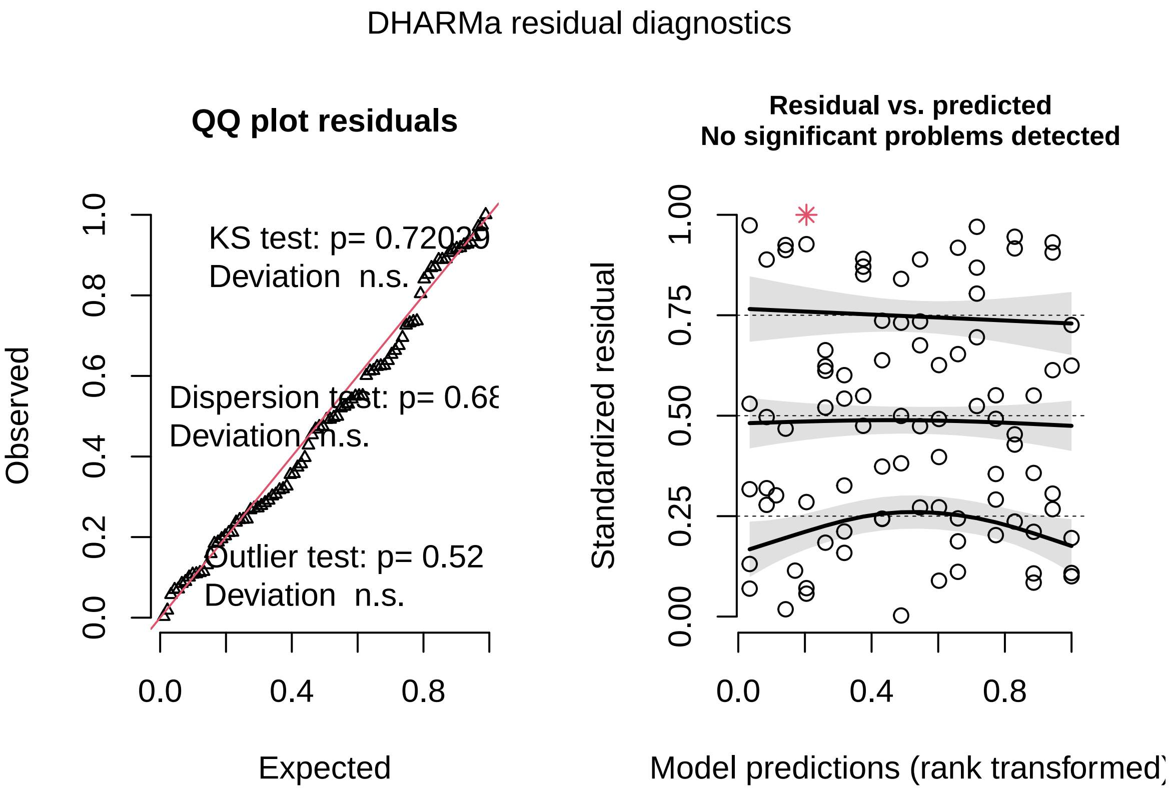


```
#Residual deviance: 102.46  on 86  degrees of freedom
```

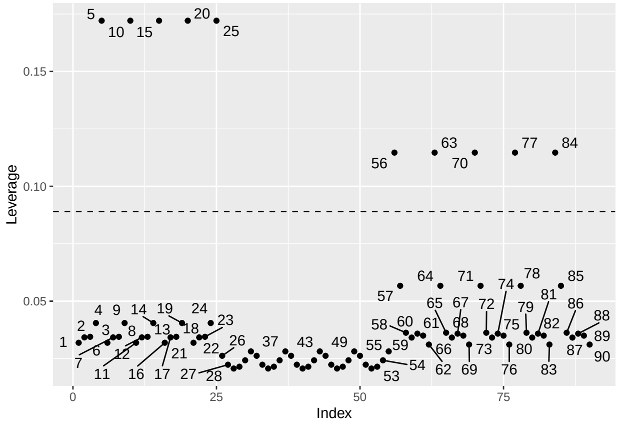
```
library(DHARMa)
# Quantile Residuals
q_unif <- simulateResiduals(modelnbc3, n = 500)
q_norm <- residuals(q_unif, quantileFunction = qnorm)

p3 <- ggplot(tibble(q = q_norm), aes(x = q)) +
  geom_histogram(aes(y = ..density..)) +
  geom_density(aes(colour = "density estimate")) +
  geom_function(aes(colour = "normal curve"),
              fun = dnorm
  ) +
  scale_colour_discrete(name = "curve") +
  theme(legend.position = "top")
p4 <- ggplot(tibble(q = q_norm), aes(sample = q)) +
  geom_qq()+
  geom_qq_line()

p3+p4
```



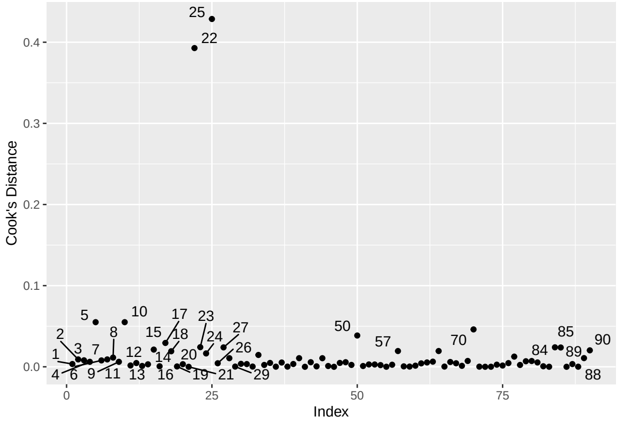
```
# Test in Quantile residuals
plot(q_unif)
```



```
library("ggrepel")
diag_m11=broom::augment(modelnbc3)
ggplot(diag_m11, aes(y = .hat, x = seq_along(.hat))) +
  geom_point() +
  xlab("Index") +
  ylab("Leverage") +
  geom_hline(yintercept = 0.089, linetype = 2) +
  ggrepel::geom_text_repel(aes(label = seq_along(.hat)))
```



```
ggplot(diag_m11, aes(y = .cooksd, x = seq_along(.hat))) +
  geom_point() +
  xlab("Index") +
  ylab("Cook's Distance") +
  ggrepel::geom_text_repel(aes(label = seq_along(.hat)))
```

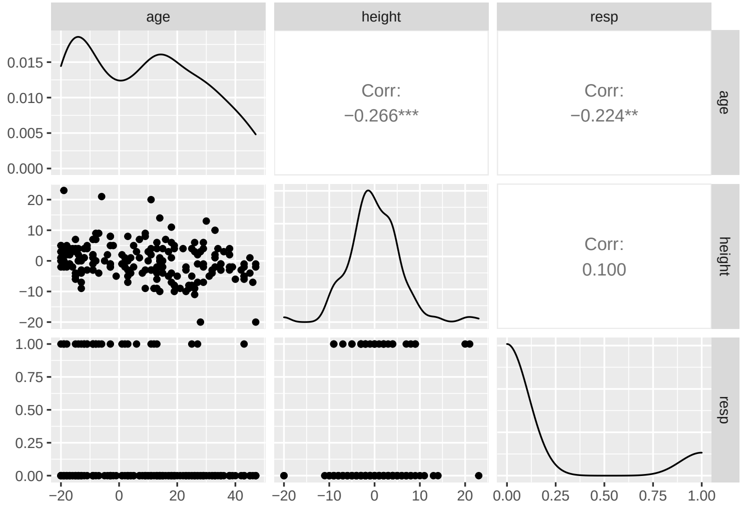


1e The first graph shows histogram with density estimates and normal curve. And its shows the densitity curve is closer to the normal curve indicating that the residual of the model coverage is good. No indication of it diviating away from the normal curve is seen. Few observations are observed to be bit diviated from the normal curve but the model is not effected by them Seen above we can see Outlier test to be p=0.54, Dipserion test to be p=0.684, KS test to be p=0.72079. Showing that all test are not significant which tells that our model fits well. In Model prediction we can clearly observe that no problems are dedicted inside the model. In the first graph of cooks distance we can see that points 56,63,70,77 and 84 are well off from other points, there are greater then H(i) value but since they are greater then the cutoff of 0.08 hence they can be potential outliers. We can also observe for points 5,10,15,20 and 25 that thepoints are observed to be far off from other points but since they are above cut off leverage of 0.08 we can see them as potential outliers. For second graph we can see the values 22 and 25 to be far off from other points but since they have cooks distance lesser than 1 hence they can't be be considered as outliers.

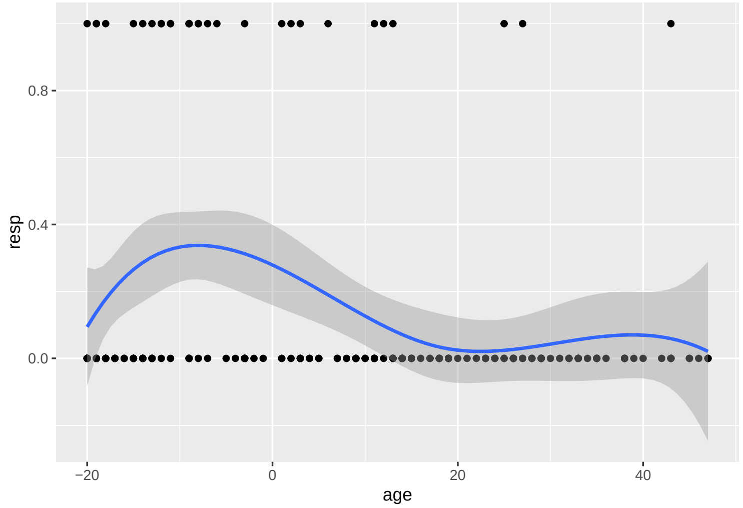
##Question 2 a

```
respiratory = read.table("respiratory.csv",head = TRUE, sep=",")
```

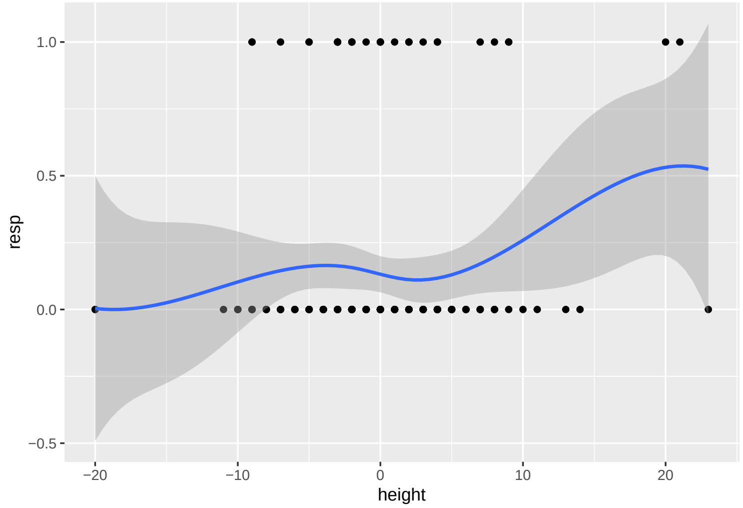
```
library(ggplot2)
library(GGally)

ggpairs(respiratory[,c("age","height","resp")])
```



```r
ggplot(respiratory, aes(x = age, y = resp)) +
  geom_point() +
  geom_smooth(method = "lm", formula = y ~ splines::bs(x, 5))
```



```r
ggplot(respiratory, aes(x = height, y = resp)) +
  geom_point() +
  geom_smooth(method = "lm", formula = y ~ splines::bs(x, 5))
```



```
# Categorical variables
library(janitor)
(tab1 <- respiratory %>%
    janitor::tabyl(resp,female))
```

```
##  resp  0  1
##     0 98 68
##     1 20  9
```

```
#11.68% chance of getting infection if its a female, 16.95% for male

janitor::chisq.test(tab1)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tab1
## X-squared = 0.64545, df = 1, p-value = 0.4217
```

```
# pvalue from the chi 2 is very high meaning that there is no link between resp and female

(tab2 <- respiratory %>%
    janitor::tabyl(resp,stunted))
```

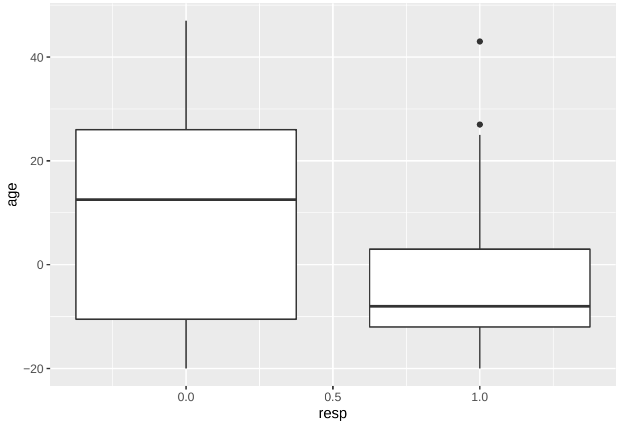
```
##  resp   0  1
##     0 142 24
##     1  26  3
```

```
#11.1 % chance of infection if stunted, 15.48% if not stunted

# pvalue from the chi 2 is very high meaning that there is no link between resp and stunted
janitor::chisq.test(tab2)
```

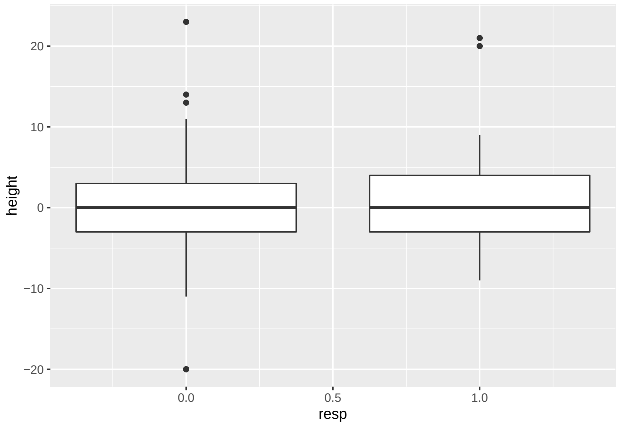
```
##
```

```
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tab2
## X-squared = 0.090196, df = 1, p-value = 0.7639
```

```
par(mfrow=c(1,2))
ggplot(respiratory, aes(x=resp, y=age, group=resp)) +
    geom_boxplot()
```



```
ggplot(respiratory, aes(x=resp, y=height, group=resp)) +
    geom_boxplot()
```



Seeing the coorelations Respiratory has less coorelation with Height ie 0.1 corr. Hence indicating that Height should be included. For Age and Height have negative coorelation with -0.26 and is not very strong but the graph proves it since the relationship between the two is not linear.

As per graphs Age does not seem to be normally distributed and is right skewed. On the other hand height seems to have normal distribution.

Scatter plot suggest inverse relationship between age and resp since when age increases respiratory infection decreases. Relationship between height and resp is directly propotional since resp increases so does height.

Box plots: Resp vs Height The mean of the box plot seem to be same for non infection and infection. Number of observations are slightly the same but outliers are also detected, but are insignificant.

Resp vs age The mean age of non infection is higher then infection. The number of observations for non infections are also greater than infection. No signficant outliers are detected in both box plots

Categorical Variable

Chances of getting infection for males is observed as 16.9% where as for female to be 11.7%. But they have no relationship between them since the chi sq p value is observed to be greater than 0.05.

Chances of getting infection for non stunted is observed as 15.5% where as for female to be 11.1%. But they have no relationship between them since the chi sq p value is observed to be greater than 0.05.

```
# b part
#(part 1)

m1=glm(resp~age,data=respiratory,family = binomial)
summary(m1)
```

```
##
## Call:
## glm(formula = resp ~ age, family = binomial, data = respiratory)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8224  -0.6438  -0.4649  -0.3288   2.5376
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.63048    0.20729  -7.866 3.67e-15 ***
## age         -0.03601    0.01197  -3.007  0.00263 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 163.99  on 194  degrees of freedom
## Residual deviance: 153.45  on 193  degrees of freedom
## AIC: 157.45
##
## Number of Fisher Scoring iterations: 5
```

```
m2=glm(resp~poly(age,2),data=respiratory,family = binomial)
m3=glm(resp~poly(age,3),data=respiratory,family = binomial)
summary(m3)
```

```
##
## Call:
## glm(formula = resp ~ poly(age, 3), family = binomial, data = respiratory)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8958  -0.6704  -0.4258  -0.2320   2.6759
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -2.0716     0.2685  -7.715 1.21e-14 ***
## poly(age, 3)1 -11.5259     3.8194  -3.018  0.00255 **
## poly(age, 3)2  -0.8205     3.3765  -0.243  0.80801
## poly(age, 3)3   8.1441     3.3363   2.441  0.01465 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 163.99  on 194  degrees of freedom
## Residual deviance: 146.19  on 191  degrees of freedom
## AIC: 154.19
##
## Number of Fisher Scoring iterations: 5
```

```
m4=glm(resp~poly(age,4),data=respiratory,family = binomial)
# The linear age and the cubic age are significant terms based on the p values <0.25

m5=glm(resp~height,data=respiratory,family = binomial)
summary(m5)
```

```
##
## Call:
## glm(formula = resp ~ height, family = binomial, data = respiratory)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9035  -0.5994  -0.5496  -0.4870   2.1430
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.76681    0.20515  -8.612   <2e-16 ***
## height       0.04703    0.03377   1.393    0.164
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 163.99  on 194  degrees of freedom
## Residual deviance: 162.07  on 193  degrees of freedom
## AIC: 166.07
##
## Number of Fisher Scoring iterations: 4
```

```
m6=glm(resp~poly(height,2),data=respiratory,family = binomial)
m7=glm(resp~poly(height,3),data=respiratory,family = binomial)
m8=glm(resp~poly(height,4),data=respiratory,family = binomial)
# Only linear term of height is significant with p value < 0.25

m9 = glm(resp ~ factor(female), data = respiratory, family = binomial)
summary(m9)
```

```
##
## Call:
## glm(formula = resp ~ factor(female), family = binomial, data = respiratory)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6095  -0.6095  -0.6095  -0.4986   2.0720
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.5892     0.2454  -6.477 9.36e-11 ***
## factor(female)1  -0.4330     0.4313  -1.004    0.315
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 163.99  on 194  degrees of freedom
## Residual deviance: 162.94  on 193  degrees of freedom
## AIC: 166.94
##
## Number of Fisher Scoring iterations: 4
```

```r
m10 = glm(resp ~ factor(stunted), data = respiratory, family = binomial)
summary(m10)
```

```
##
## Call:
## glm(formula = resp ~ factor(stunted), family = binomial, data = respiratory)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5799  -0.5799  -0.5799  -0.4854   2.0963
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -1.6977     0.2133  -7.959 1.74e-15 ***
## factor(stunted)1  -0.3817     0.6484  -0.589    0.556
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 163.99  on 194  degrees of freedom
## Residual deviance: 163.61  on 193  degrees of freedom
## AIC: 167.61
##
## Number of Fisher Scoring iterations: 4
```

2 b i Note: Some model Summaries are hidden since showing most of them will cause the page limit to exceed. From m1 to m4 variables, models have been produced of age from power 1 to power 4. Upon seeing the summaries linear and cubic terms have p<0.25. p value is 0.00263<0.25. For height we can see the linear term signficant lesser then <0.25. Both stunted and female equations have non sig values >0.25. 0.315 and 0.556 respectively

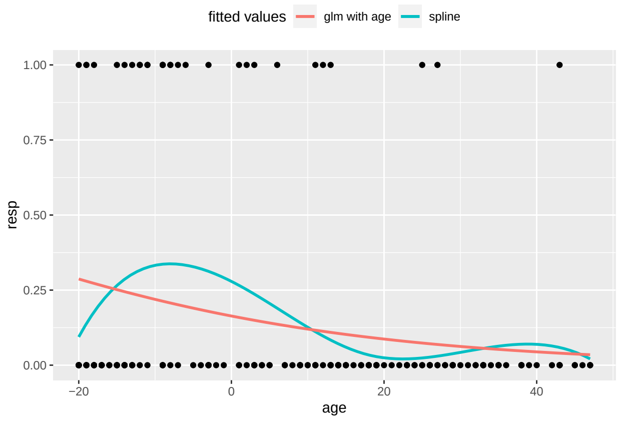
ii Equation for height: $ln(Y_i) = \epsilon + 0.04703X_i + -1.76681 \implies where\ \epsilon = error,\ Y_i = Respiratory\ infection,\ X_i = height$

Equation for Female: $ln(Yi) = \epsilon - 0.4330X_i - 1.5892 \implies where\ infection,\ \epsilon = error,\ X_i = 1\ for\ female,\ = 0\ for\ male,\ Y_i = Respiratory$

iii Height: For height we can see positive directly propotional relationship as for every decrearse on increase in height by 4.82 percent $1 - e^{0.04703}$ Positive or negative increase depends upon the sign of Beta variable.

Female: For the odds of female getting infection is decreased by percentage of 35 in comparison with females. Where as odds of infection for male 20.4 percent.

```
# (c)
# Part 1
ggplot(respiratory, aes(x = age, y = resp)) +
  geom_point() +
  geom_smooth(aes(colour="spline"),method = "lm", formula = y ~ splines::bs(x, 5),se=FALSE)+
  geom_smooth(aes(colour = "glm with age"),
              method = "glm", method.args = list(family = "binomial"),
              se = FALSE) +
  scale_colour_discrete(name="fitted values")+
  theme(legend.position = "top")
```

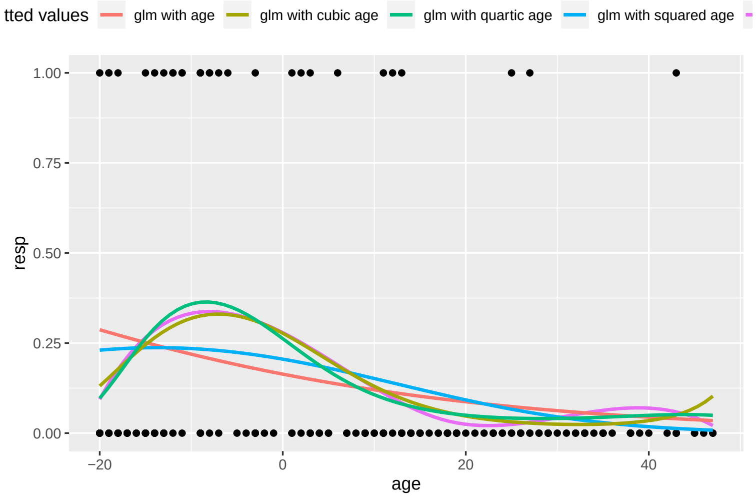


```
#Part 2
AIC(m1,m2,m3,m4)
```

```
##    df      AIC
## m1  2 157.4504
## m2  3 157.9307
## m3  4 154.1937
## m4  5 155.2269
```

c i For the above graph glm with age is not nearer to/coinciding the fitted values therefore this model is not sufficient. ii Using AIC values cube is the best model since it has the least AIC. Hence we will use the cubic AIC.

```
# Part 3

ggplot(respiratory, aes(x = age, y = resp)) +
  geom_point() +
  geom_smooth(aes(colour="spline"),method = "lm", formula = y ~ splines::bs(x, 5),se=FALSE)+
  geom_smooth(aes(colour = "glm with age"),
              method = "glm", method.args = list(family = "binomial"),
              se = FALSE) +
  geom_smooth(aes(colour = "glm with squared age"),
              method = "glm",formula = y~x + I(x^2), method.args = list(family = "binomial"),
              se = FALSE) +
  geom_smooth(aes(colour = "glm with cubic age"),
              method = "glm",formula = y~ x + I(x^2) + I(x^3), method.args = list(family = "binomial"),
              se = FALSE) +
  geom_smooth(aes(colour = "glm with quartic age"),
              method = "glm",formula = y~ x + I(x^2) + I(x^3)+ I(x^4), method.args = list(family = "bin
              se = FALSE) +
  scale_colour_discrete(name="fitted values")+
  theme(legend.position = "top")
```



iii The best model is the cubic model in this graph. Since its most closest to the spline line. Hence it is the best fit for the model.

```
# (d)
m11=glm(resp~poly(age,3)+height,data=respiratory,family = binomial)

AIC(m11,m3)
```

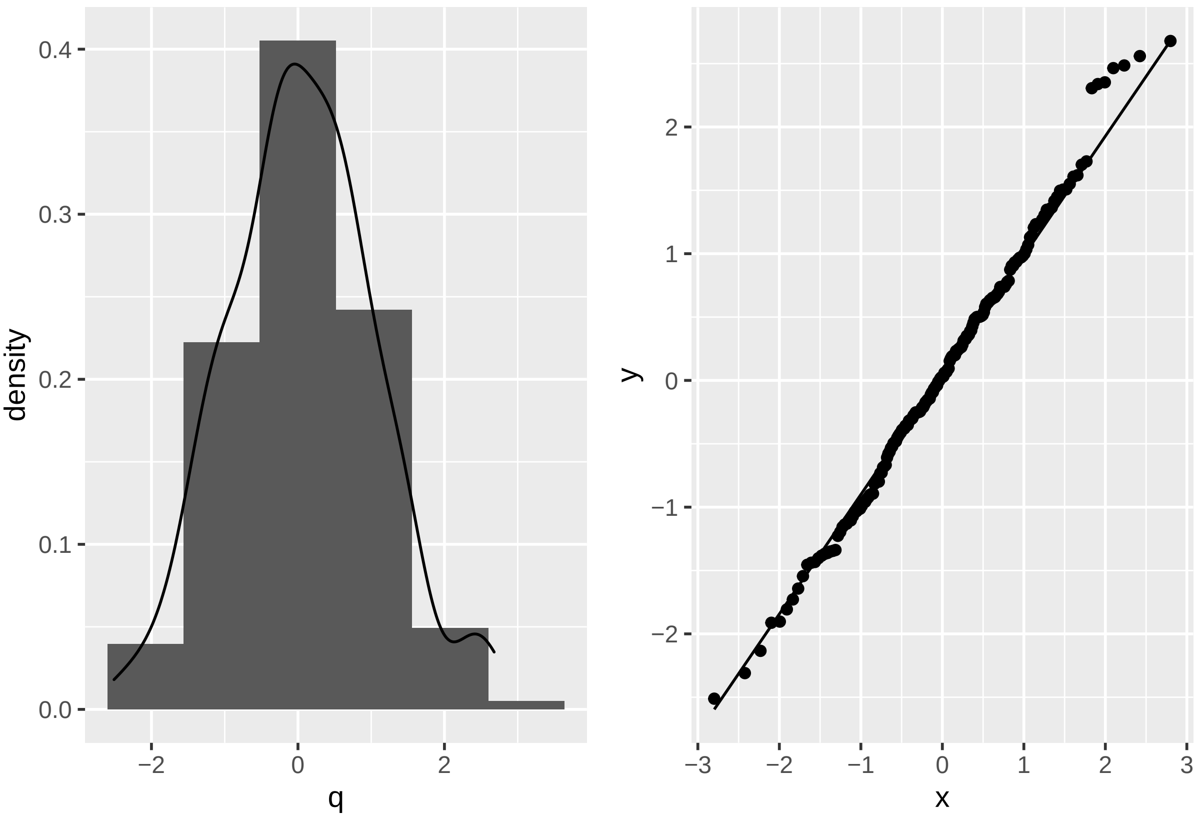
```
##     df      AIC
## m11  5 155.8367
## m3   4 154.1937
```

```
BIC(m11,m3)
```

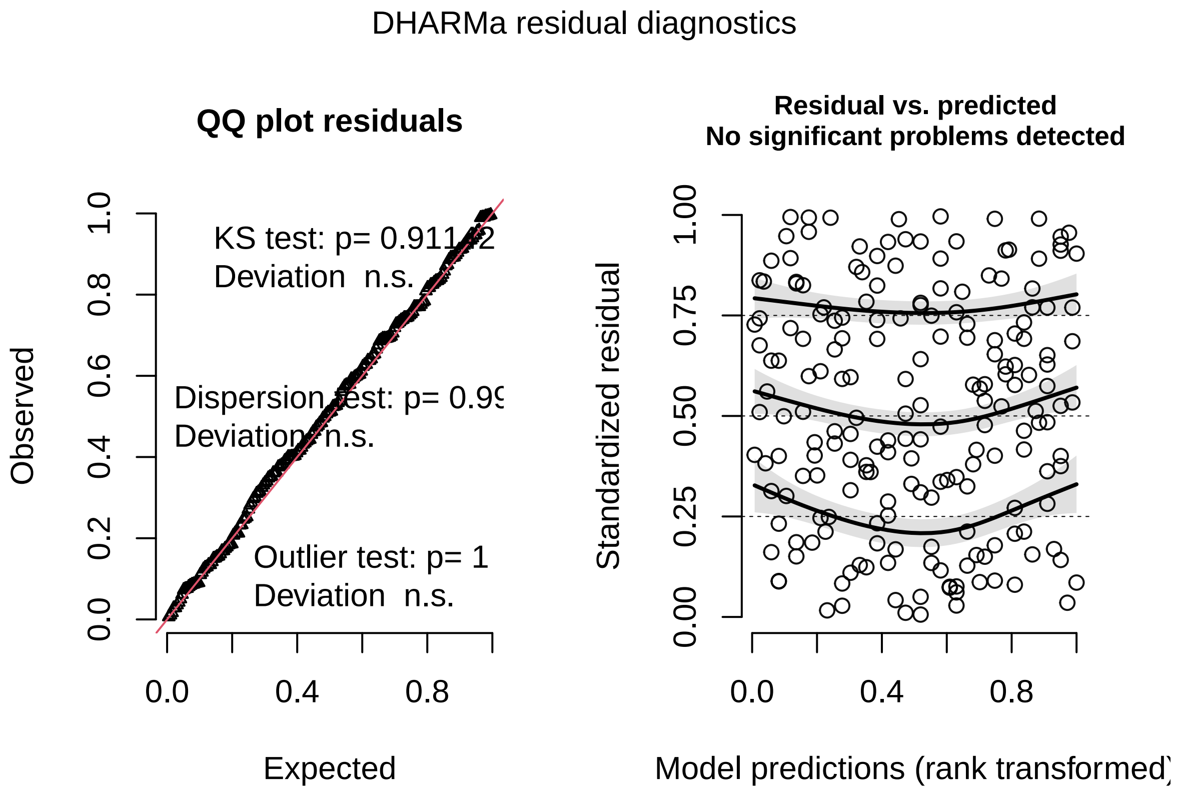
```
##     df      BIC
## m11  5 172.2017
## m3   4 167.2857
```

From the above AIC and BIC comparison, m3 which is cubic model of age is better since it has the least AIC value. Same goes for the BIC value. Hence our final will be cubic model of age it looks more adequate then the model with hieght.

```
# (e)
# Residuals
res_model <- simulateResiduals(fittedModel = m3)
q_model <- residuals(res_model, quantileFunction = qnorm)
p5 <- ggplot(tibble(q = q_model), aes(x = q)) +
  geom_histogram(aes(y = ..density..), bins = 6) +
  geom_density()
p6 <- ggplot(tibble(q = q_model), aes(sample = q)) +
  geom_qq() +
  geom_qq_line()
p5 + p6
```



```
plot(res_model)
```



```
table(respiratory$resp)
```

```
##
##   0   1
## 166  29
```

```
#select cut off from original prob (29/(29+166))

probs.quad <- fitted(m3)
table(respiratory$resp, probs.quad >= 0.25)
```

```
##
##     FALSE TRUE
##   0   136   30
##   1    14   15
```

The deviance plot has a normal distribution meaning the distribution deviance is normal through out. QQ plot is normal also we don't see any points deviating from the straight line.

Seeing the observed vs expected observations we can see the KS test, Dispresion test and outlier test all them are non significent. This tells that there is no dispersion that effected the model and neither such outlier present.
cut off probability is $\frac{29}{166+29} = 14.87\%$ classification indidcates that sensitivity $= \frac{15}{30+15} = 33\%$ specificity $= \frac{136}{14+136} = 90\%$

Question NO. 3 Part a:

$E(Y) = 0.(1-p) + 1.p \quad E(Y) = p \ E[Y] = \sum yP[Y = y]$

Part b:

$f(y) = (1-p)^{1-y}1.p^y = (1-p)(\frac{p}{1-p})^y$

$exp[log(1-p) + ylog(\frac{p}{1-p})]$

$\theta = log(\frac{p}{1-p})$

$e^{-\theta} = frac1 - pp, \quad e^{\theta} = \frac{p}{1-p}$

$e^{-\theta} = -1 + frac1p \implies e^{-\theta} + 1 = \frac{1}{p} \implies \frac{1}{e^{-\theta}+1} = p$

$\frac{e^\theta}{e^\theta+1} = p \quad , \ \frac{1}{1+e^\theta} = 1-p$

$f(y) = exp[-log(1+e^\theta) + y\theta] \implies exp[c(y,\phi) + \frac{-b(\theta)+y\theta}{a\phi}]$

$where \ \ a(\phi) = 1 \ \ c(y,\phi) = 0, \quad b(\theta) = log(a + e^\theta), \ \theta = log\frac{p}{1-p}$

part c:

$\mu = E(X) = b^{'}(\theta) = \frac{e^\theta}{e^\theta+1} = p$

$Var(X) = a(\phi).\frac{\delta^2}{\delta\theta^2}b(\theta)$

$\frac{e^\theta + e^{2\theta} - e^{2\theta}}{(1+e^\theta)^2}$

$\frac{e^\theta}{(1+e^\theta)^2} \implies \frac{e^\theta}{(1+e^\theta)}.\frac{1}{(1+e^\theta)}$

$p(1-p)$

Canonical Link:

$\frac{e^\theta}{1+e^\theta} = p = \mu$

$\theta = log\frac{p}{1-p} = log\frac{\mu}{1-\mu}$

Question 4

a

$F(\mu) = (-e^{-\mu})\frac{-1}{(e^{-\mu}+1)^2} = \frac{d}{d\mu}F_\epsilon(\mu)$

$-\infty < \mu < \infty, \frac{e^{-\mu}}{(e^{-\mu})^2+1}$

Part c:

$P(Y_i = 1|X_i = x_i, \beta) = P(\Psi_i \geq 0|X_i = x_i, \beta)P(x_i^T\beta + \epsilon_i \geq 0) = P(\epsilon \geq -x_i^T, \beta) = P(\epsilon \leq x_i^T, \beta)$

According to CDF, we have

$p_i = P(\epsilon \leq x_i^T, \beta) = F_{\epsilon_i}(x_i^T\beta) = \frac{1}{1+e^{-x_i^T\beta}}$

Part d:

Based on what is given, Generalized Linear Model can be defined as Logistic Regression Model with link function logit:

$log(\frac{p_i}{-p_i+1})$