General form of the function which belong to exponential family

$$f(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}$$

Given density function =

$$f(y;p) = (y+1)p^2(1-p)^y \qquad y = 0,1,2,3\ldots$$

$$= \exp\left[\log\left\{(y+1)p^2(1-p)^y\right\}\right]$$

$$= \exp\left[\log(y+1) + \log p^2 + \log(1-p)^y\right]$$

$$= \exp\left[\log(y+1) + 2\log p + y\log(1-p)\right]$$

$$\Rightarrow \exp\left[y\log(1-p) + \log(y+1) + 2\log p\right]$$

$\theta = \log(1-p)$ [On comparing to general form]

$b(\theta) = -2\log(p)$

$\phi = 1$ as $c(y,\phi) = \log(y+1)$

Hence we can ~~sho~~ confirm that the above function belong to the exponential family.

$$E(y) = b'\theta$$

$$= \frac{d}{d\theta}(-2\log p)$$

$$= -2 \frac{d}{d\theta} (\log p)$$

$$= -2 \frac{d}{d\theta} (\log (1 - e^{\theta}))$$

$$= -2 \times \frac{1}{1 - e^{\theta}} \times -e^{\theta}$$

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

$\therefore \theta = \log(1-p)$

$e^{\theta} = 1 - p$

$p = 1 - e^{\theta}$

$$\text{Var}(y) = \phi \, b''(\theta)$$

$$= 1 \times \frac{d}{d\theta} b'(\theta)$$

$$\Rightarrow \frac{d}{d\theta} \left(\frac{2(1-p)}{p} \right)$$

$$\Rightarrow 2 \frac{d}{d\theta} \left(\frac{1-p}{p} \right)$$

$$\Rightarrow 2 \frac{d}{d\theta} \left(\frac{1}{p} - 1 \right)$$

$$\Rightarrow 2 \left[\frac{d}{d\theta} \left(\frac{1}{p} \right) - \frac{d}{d\theta} (1) \right]$$

$$= 2 \frac{d}{d\theta} \left[\frac{1}{(1 - e^{\theta})} \right]$$

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

811 Take Home

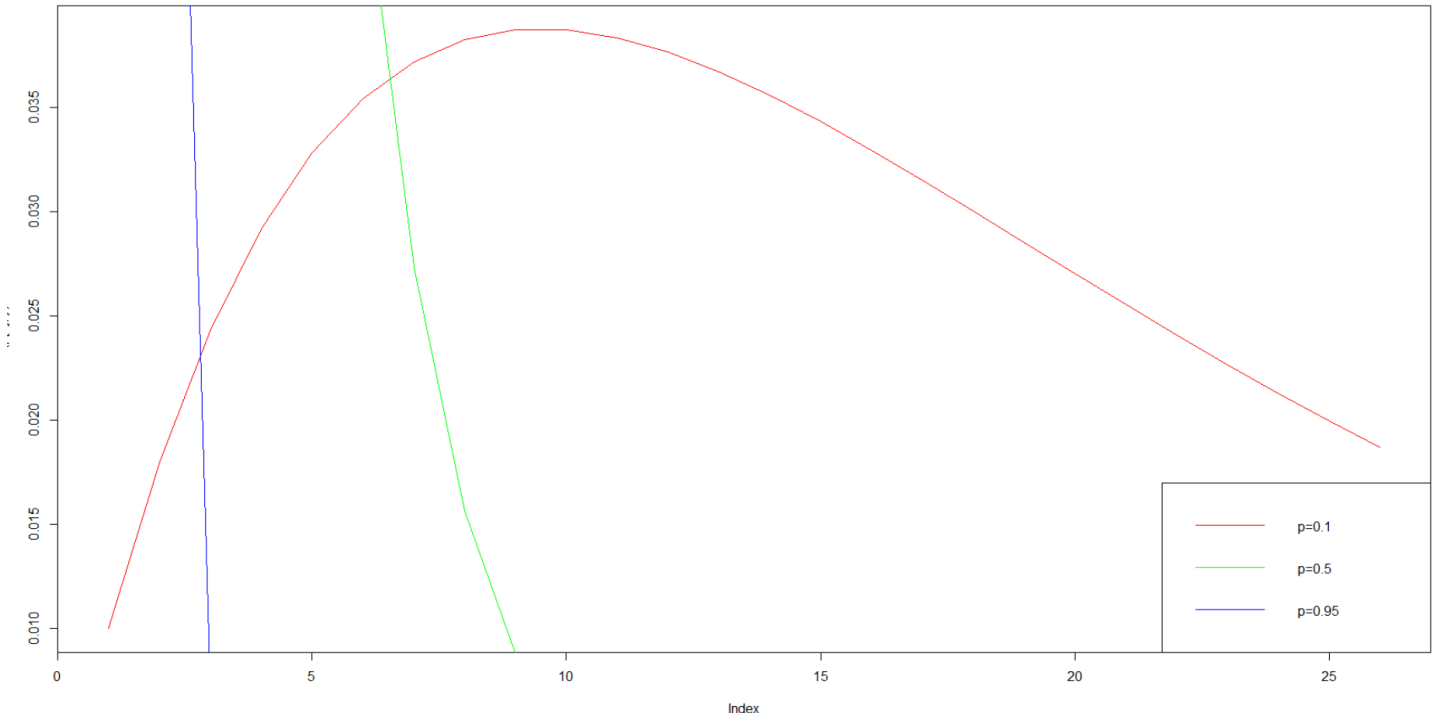
Sukhdeep Singh (44442467)

November 16, 2018

Question 1

c)

```
f = function(p,y) {(y+1)*p^2*(1-p)^y}
y=0:25
p=c(0.1,0.5,0.95)

plot(f(p[1],y) ,type='l' , col = 'red')
lines(f(p[2],y), col='green', type= 'l')
lines(f(p[3],y), col='blue' ,type ='l')
legend(x = "bottomright" , legend = c("0.1", "0.5", "0.95"), col = c("red", "green",
"blue"), lty= 1:1)
```



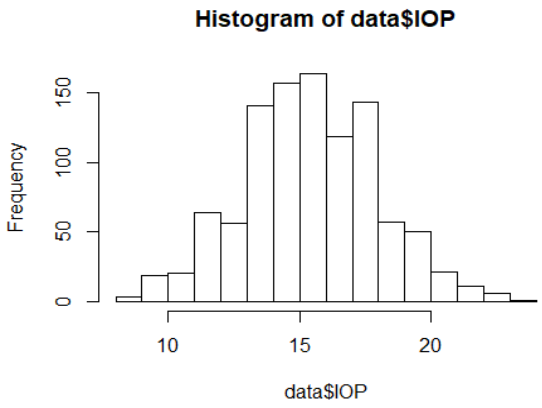
d) Negative binomial distribution

Question 2:

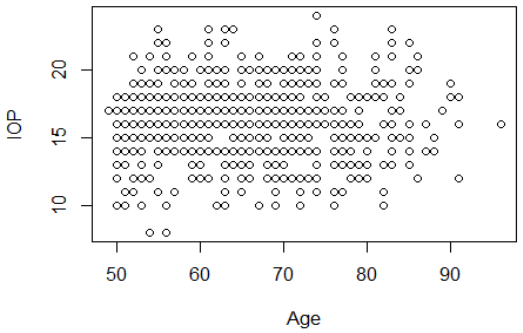
```
library('nlme')
library('dplyr')

data = read.csv(file = 'eye_takehome_2018.csv', header = TRUE)

#checking NA's
summary(data)

##       age             sex           hearimp         alcohol
##                                  NA's   :6       NA's   :50
```

```
##      systbp         diastbp

##  NA's   :8       NA's   :8
```

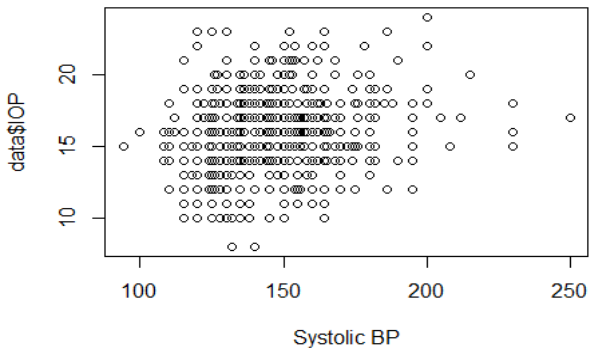
```
#removing Na's as we want to fit models on same observation.
data = na.omit(data)

hist(data$IOP, breaks = 20)
```



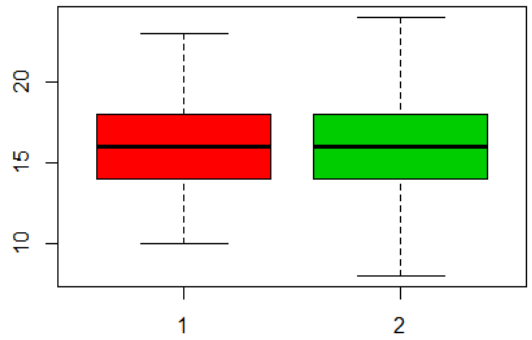

```
plot(data$age,data$IOP, xlab = "Age", ylab = "IOP")
```



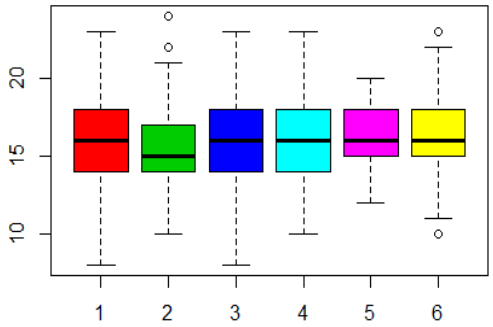

```
plot(data$systbp, data$IOP, xlab = "Systolic BP")
```



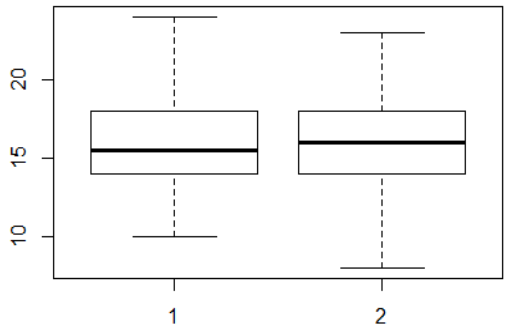

```
boxplot(data$IOP ~ factor(data$sex), col = 2:3)
```




```
boxplot(data$IOP ~ factor(data$alcohol) ,col = 2:7)
```




```
boxplot(data$IOP ~ factor(data$hearimp))
```



Initial screening:

- We observe some of the NA values in the covariates, which are omitted.
- Distribution of IOP looks fairly normal.
- No observed pattern is visible b/w response and covariates
- Sex as a factor is balanced

- Alcohol as a factor has 6 levels not equally balanced
- Hearimp as a factor has two levels one of which is sparsely populated.
- Covariate selection: All continuous covariates pass the screening. No categorical varirate passed

Let's try to compare the models:

```
model1 <- lme(IOP~age, random=~1|IdNum, data=data, correlation = corAR1(form=~1|IdNum))

model2 <- lme(IOP~systbp, random=~1|IdNum, data=data, correlation=corSymm(form=~1|IdNum))

model3<-  lme(IOP~diastbp+age, random=~1|IdNum, data=data)

AIC( model1, model2, model3)
```

```
##        df      AIC
## model1  5 4116.563
## model2  5 4097.664
## model3  5 4110.368
```

We see model 2 is favored by AIC and has the best explanatory power.

Final model equation:

Y_ij|u_i = (12.42) + 0.023907(X_ij1) + u_i, Correlation(Y_i1, Y_i2) = 0

```
summary(model2)
```

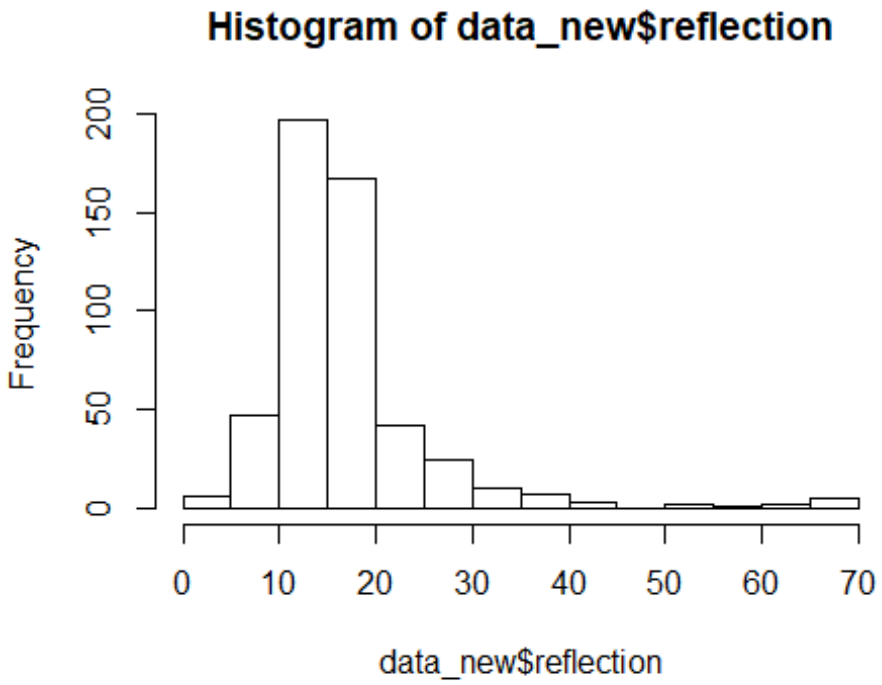
```
## Linear mixed-effects model fit by REML
##  Data: data
##        AIC      BIC    logLik
##   4097.664 4122.351 -2043.832
##
## Random effects:
##  Formula: ~1 | IdNum
##         (Intercept)  Residual
## StdDev:    2.387566 0.8711182
##
## Correlation Structure: General
##  Formula: ~1 | IdNum
##  Parameter estimate(s):
##  Correlation:
##   1
## 2 0
## Fixed effects: IOP ~ systbp
##                 Value Std.Error  DF   t-value p-value
## (Intercept) 12.423005 0.7539263 516 16.477745       0
## systbp       0.023907 0.0051332 516  4.657375       0
##  Correlation:
##        (Intr)
## systbp -0.99
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -5.13060381 -0.24574715 -0.02462687  0.19547278  5.20094448
##
## Number of Observations: 1032
## Number of Groups: 518
```

We can confirm there no significant relationship between IOP~Age after correcting for other parameters

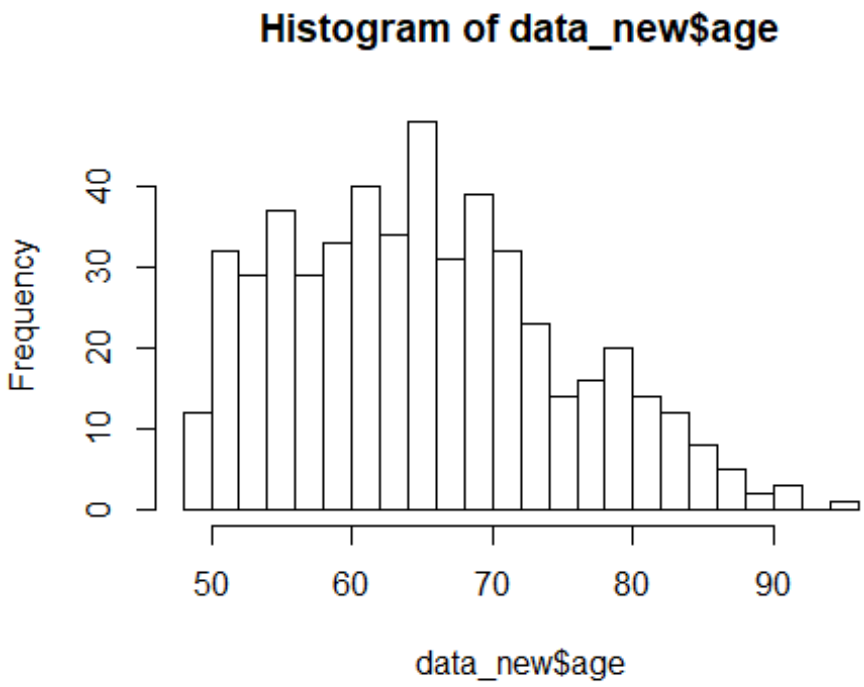
Question 2

Graphical data exploration:

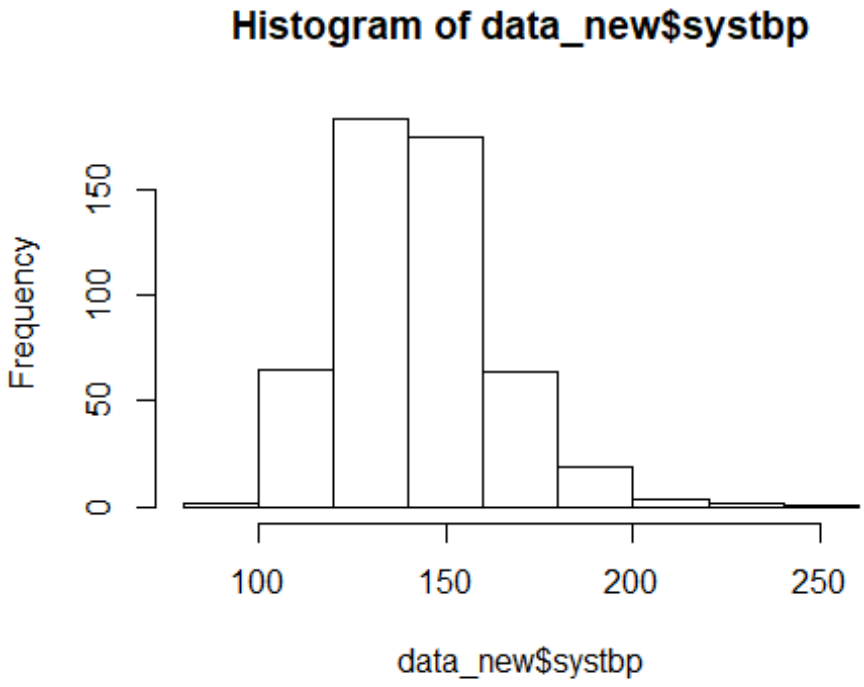
```
#data tranformation
data$reflection = 70-data$VA

#subsetting
data_new = subset(data, eye == 1, select = IdNum:reflection)

#Distribution of response
hist(data_new$reflection, breaks = 20)
```



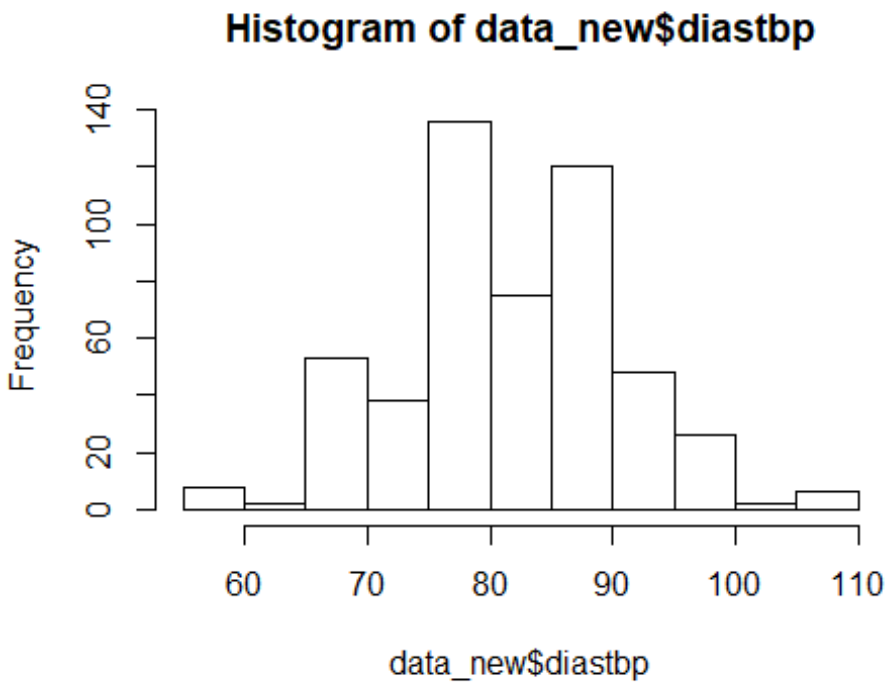

```
hist(data_new$age, breaks = 20)
```



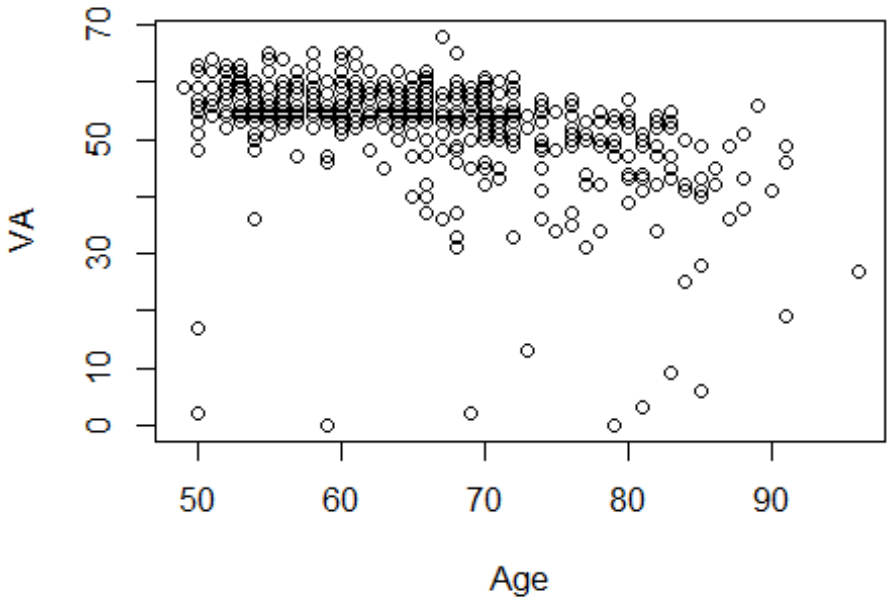

```r
hist(data_new$systbp)
```



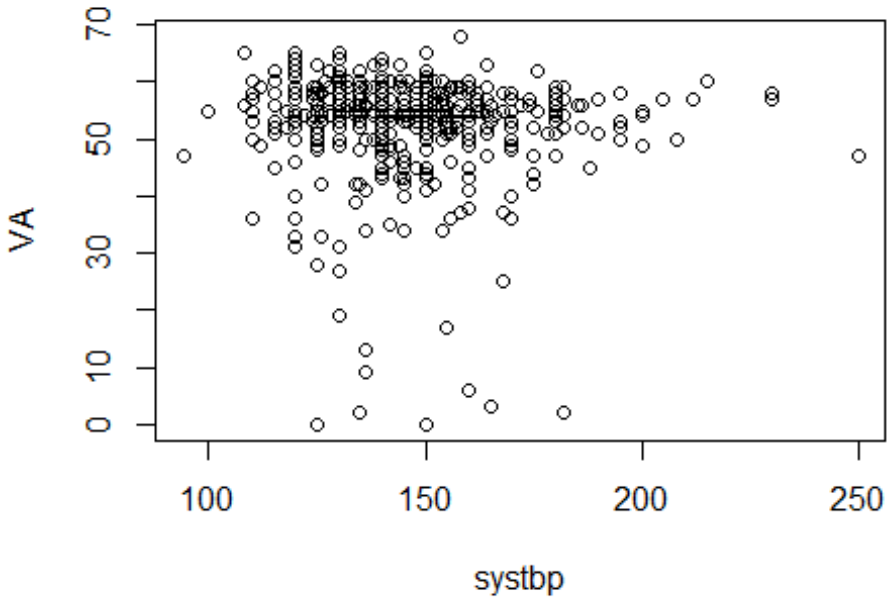

```r
hist(data_new$diastbp)
```



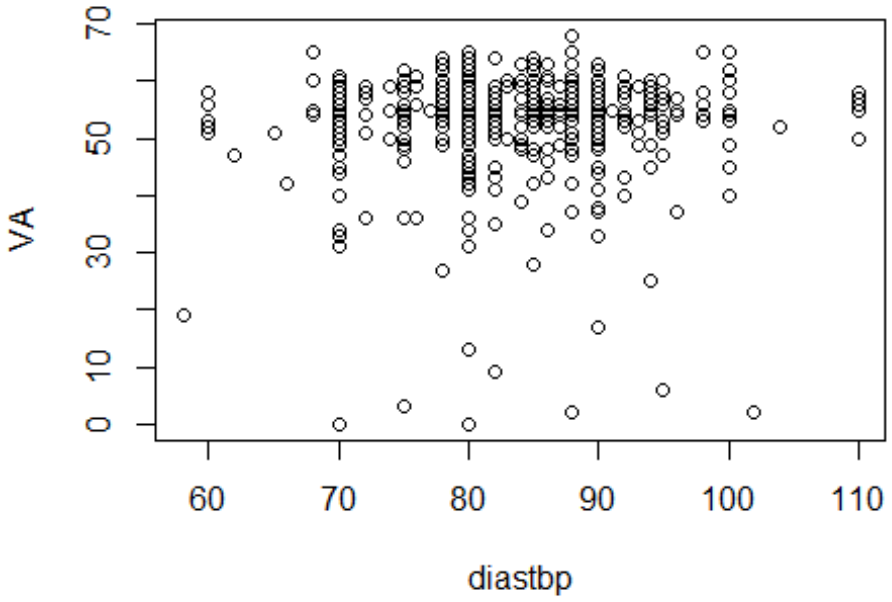

```r
#Scatterplot analysis
plot(data_new$age, data_new$VA, xlab = "Age", ylab= "VA")
```



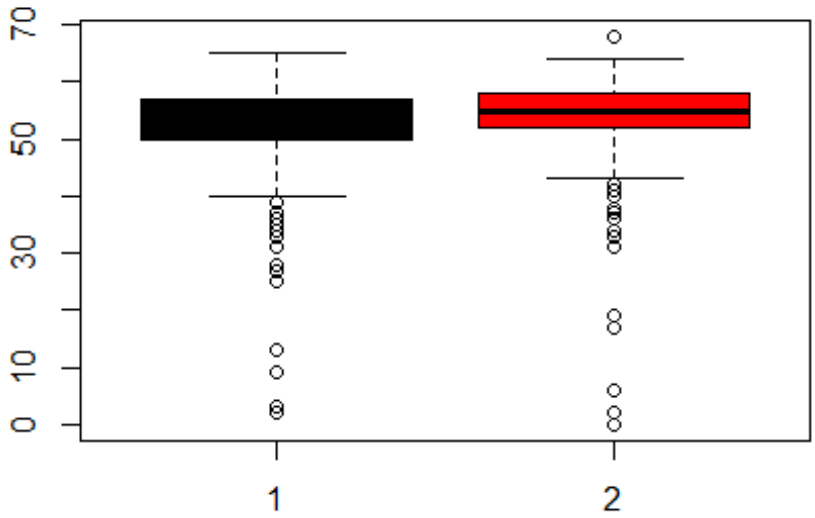

```r
plot(data_new$systbp, data_new$VA, xlab = "systbp", ylab= "VA")
```



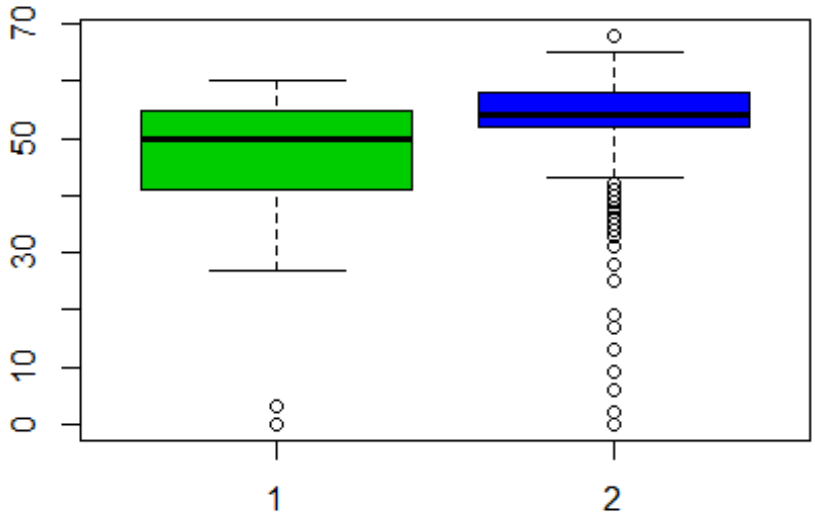

```r
plot(data_new$diastbp, data_new$VA, xlab = "diastbp", ylab= "VA")
```



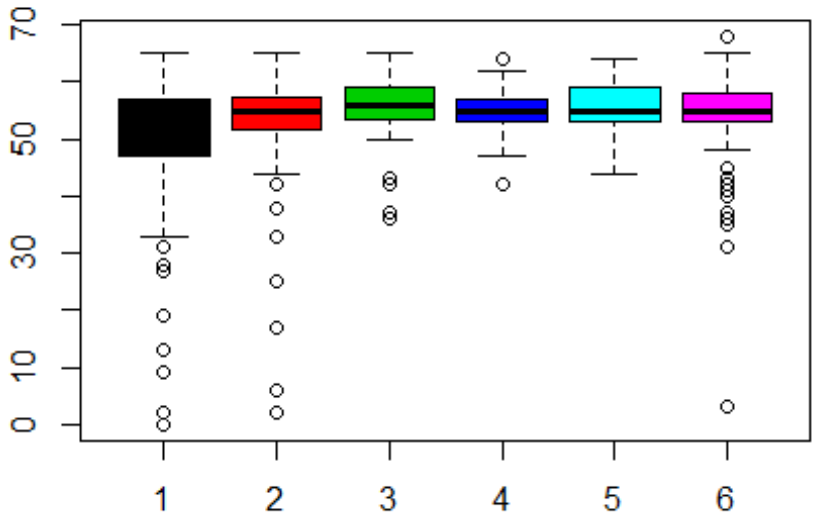

```r
boxplot(data_new$VA ~ factor(data_new$sex), col = 1:2)
```



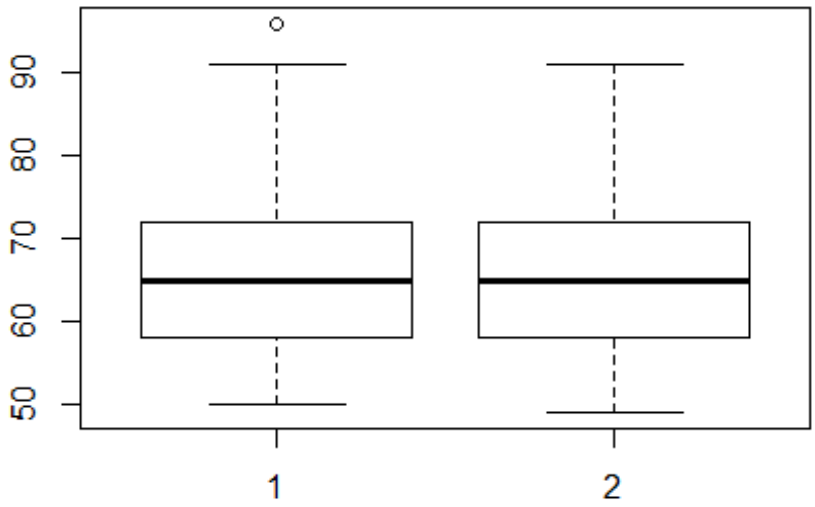

```
boxplot(data_new$VA ~ factor(data_new$hearimp), col= 3:4)
```



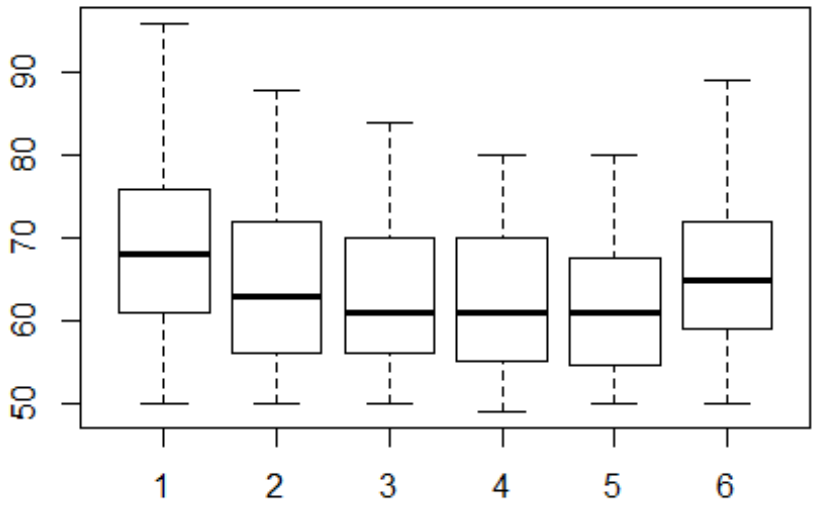

```
boxplot(data_new$VA ~ factor(data_new$alcohol), col = 1:6)
```



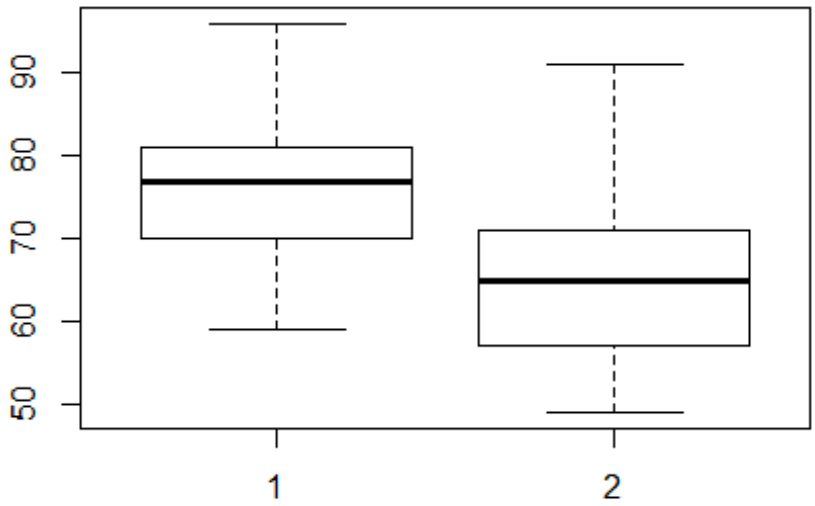

```r
#checking for relationship b/w covariates
boxplot(data_new$age ~ data_new$sex)
```




```r
boxplot(data_new$age ~ data_new$alcohol)
```




```
boxplot(data_new$age ~ data_new$hearimp)
```



- Graphical plots, reveal new transfomed variable is right skewed, Blood pressure is fairly normally distributed. Whereas age looks like slightly right skewed
- No obvious relationship is observable from the scatter plots or boxplots.
- No significant replationship b/w covariates

Model building

- As reflected VA is discreet score we consider poisson model
- Deviance for poisson indicates overdispersion
- Need to try negative binomial models
- All variables pass the initial screening

```
library('knitr')
table1 = read.table('2b1.txt',sep = '\t', header = TRUE)

kable(table1)
```

Model	Residual.deviance	df	AIC
Age	1435.4	512	3807.1
Age+systbp	1424.8	511	3798.5
Age+systbp+diastbp	1424.8	510	3800.5
Age+systbp+factor(sex)	1419.9	510	3795.6
Age+systbp+factor(alcohol)	1389.7	506	3773.5
Age+systbp+factor(hearimp)	1403.1	510	3778.9
Age+systbp+factor(alcohol)+factor(hearimp)	1365.8	505	3751.6

```
data_new<- within(data_new, hearimp <- relevel(factor(hearimp), ref = 2))

modelpois<-glm(reflection ~ age+systbp+factor(hearimp), family = poisson(link = "log"),
data= data_new)

summary(modelpois)

##
## Call:
## glm(formula = reflection ~ age + systbp + factor(hearimp), family = poisson(link =
"log"),
##     data = data_new)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.5940  -1.0425  -0.2687   0.4418  11.0617
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       1.6786193  0.0926244  18.123  < 2e-16 ***
## age               0.0216395  0.0011100  19.495  < 2e-16 ***
## systbp           -0.0019398  0.0005218  -3.718 0.000201 ***
## factor(hearimp)1  0.2051797  0.0430022   4.771 1.83e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1861.8  on 513  degrees of freedom
## Residual deviance: 1403.1  on 510  degrees of freedom
## AIC: 3778.9
```

```
##
## Number of Fisher Scoring iterations: 4
```

We choose to ignore the interaction terms to avoid model complexity.

Based on AIC we can pick up the model age+systbp+factor(hearimp) as its performance is close enough to the fully fitted model.

We can see the case of overdispersion here as residual deviance is much higher than the degrees of freedom. Let's look at the negative binomial model instead (using the same covariates).

Negative binomial model

Model equation:

Y ~ NB(mu_i, k)

log(reflection) = 1.722713 + (0.020655 x age) -(0.001789 x systbp) + (0.198600 x factor(hearimp)1)

```
library('boot')
```

```
##
## Attaching package: 'boot'
```

```
## The following object is masked from 'package:gamlss.data':
##
##     aids
```

```
library('car','MASS')
```

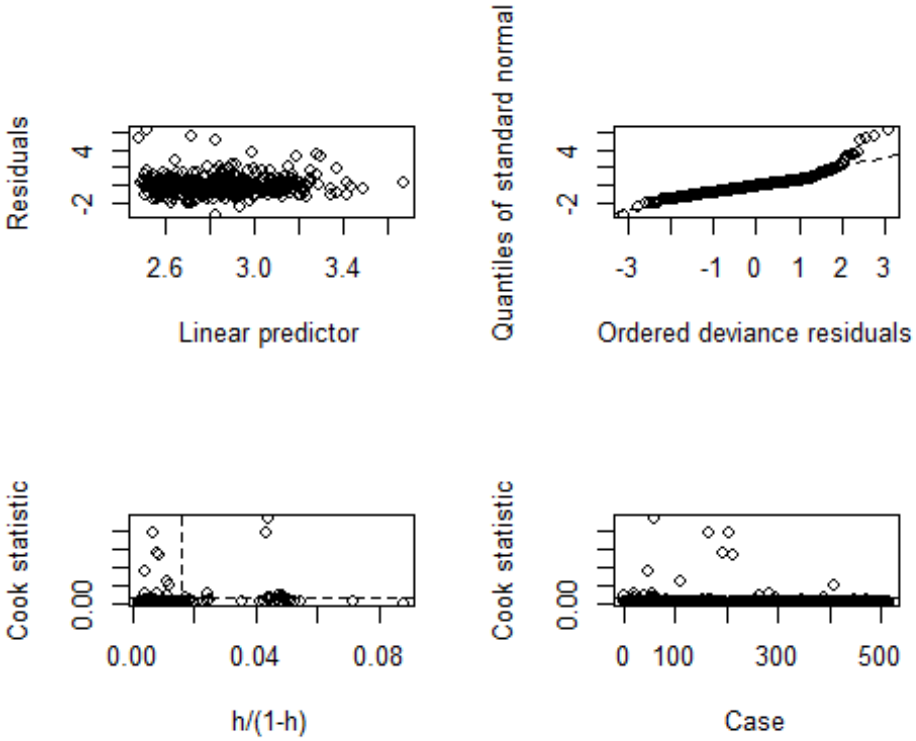
```
data_new<- within(data_new, hearimp <- relevel(factor(hearimp), ref = 2))
modelnb<-glm.nb(reflection ~ age+ systbp  +factor(hearimp), data= data_new)
```

```
summary(modelnb)
```

```
##
## Call:
## glm.nb(formula = reflection ~ age + systbp + factor(hearimp),
##     data = data_new, init.theta = 11.04466147, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3614  -0.6748  -0.1686   0.2692   6.2856
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       1.921313   0.181607  10.579   <2e-16 ***
## age               0.020655   0.001845  11.192   <2e-16 ***
## systbp           -0.001789   0.000849  -2.107   0.0351 *
## factor(hearimp)2 -0.198600   0.077314  -2.569   0.0102 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(11.0447) family taken to be 1)
##
##     Null deviance: 657.07  on 513  degrees of freedom
## Residual deviance: 491.78  on 510  degrees of freedom
## AIC: 3340.1
##
```

```
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  11.04
##           Std. Err.:  1.07
##
##  2 x log-likelihood:  -3330.059
```

```
glm.diag.plots(modelnb)
```



Model diagnostics:

1. Residuals are randomly scattered. 2 Ordered deviances are fairly normal 3.Cook's distance do not show a cause of concern as its well below the threshold for tolerance

Note that we can chosen hearimp level 2 as reference for our model as it is densely populated.

Model interpretation: Effect of age on the reflection of VA can be given by e^0.020655 = 1.02 which means a 2% increase in VA reflection (this means actual VA goes down by 2%). Effect of systbp on the reflection of VA can be given by e^-0.001789 = 0.99821 which means a 0.17% decrease in VA reflection (this means actual VA goes up by 0.17%)

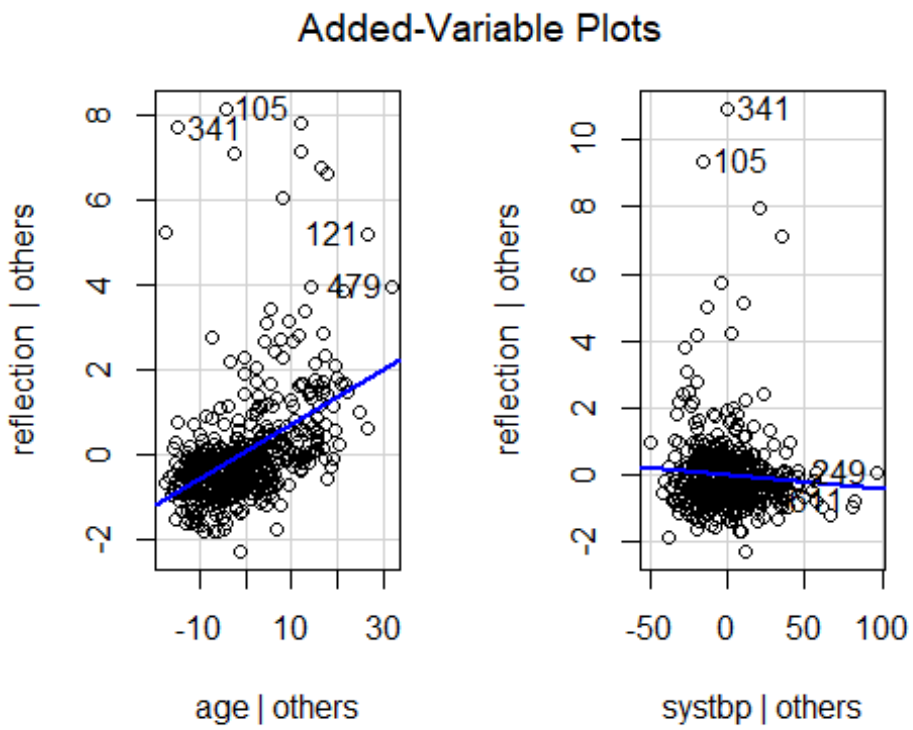
Checking the partial regression plots and link function for continuous predictors of the model0.

```
modelnb_cont<-glm.nb(reflection ~ age + systbp , data= data_new)

summary(modelnb_cont)
```

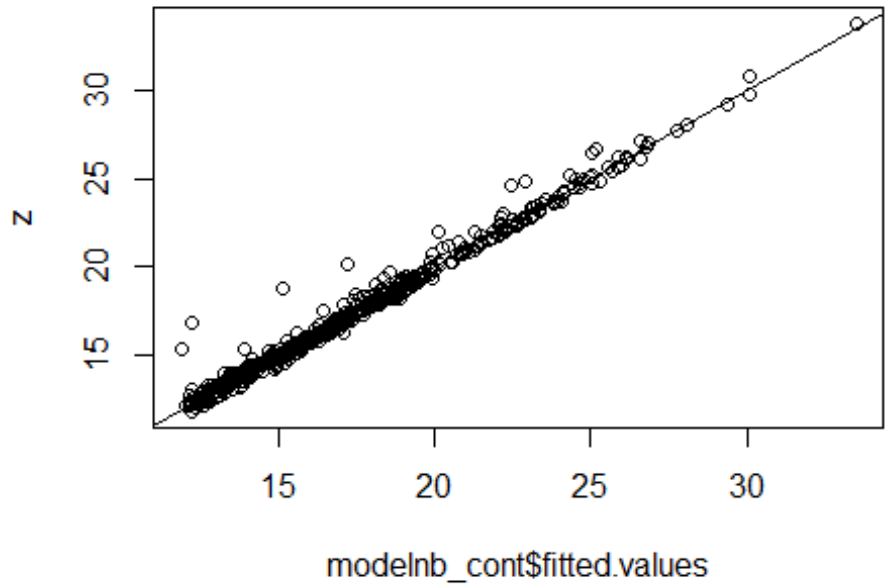
```
##
## Call:
## glm.nb(formula = reflection ~ age + systbp, data = data_new,
##     init.theta = 10.81117282, link = log)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -3.3708  -0.6606  -0.1812   0.2758   6.2755
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.6333110  0.1455399  11.222   <2e-16 ***
## age          0.0217012  0.0018187  11.932   <2e-16 ***
## systbp      -0.0015743  0.0008513  -1.849   0.0644 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(10.8112) family taken to be 1)
##
##     Null deviance: 648.29  on 513  degrees of freedom
## Residual deviance: 491.85  on 511  degrees of freedom
## AIC: 3344.7
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  10.81
##           Std. Err.:  1.04
##
##  2 x log-likelihood:  -3336.719
```

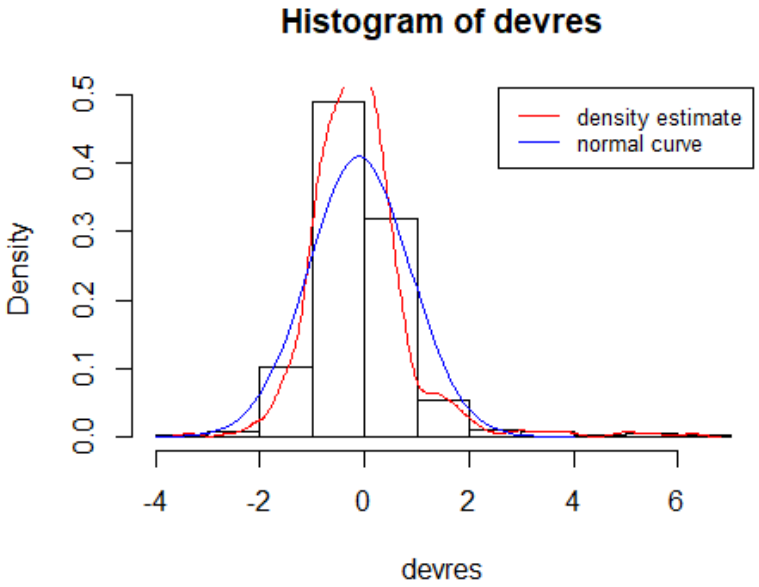
```
avPlots(modelnb_cont)
```




```
#checking the link function
z = modelnb_cont$fitted.values+
(data_new$reflection-modelnb_cont$fitted.values)/modelnb_cont$fitted.values
plot(modelnb_cont$fitted.values,z)
abline(lm(z~modelnb_cont$fitted.values))
```




```
#checking the normality of residuals
devres = resid(modelnb_cont)
hist(devres, freq=FALSE)
lines(density(devres), col="red")
lines(seq(-4,4, by=.1), dnorm(seq(-4,4, by=.1),
mean(devres), sd(devres)), col="blue")
legend("topright",legend=c("density estimate","normal curve"),
lty=1,col=c("red","blue"),cex=.8)
```



Looking at the above plots we can confirm the choice of link funciton is appropriate and residuals are fairly normal. AvPlots also seems to conform with our expectations.

Question 2b:

To check if the mean VA for both eyes is same or not we fit the below model:

Taking the same covariates as before and adding factor(eye) and random effects to account for the correlation.

```
data<- within(data, hearimp <- relevel(factor(hearimp), ref = 2))
modelnb_finale<-gamlss(reflection ~ age + systbp  +factor(hearimp)+ factor(eye) +
random(factor(IdNum)), family = NBI, data= data)

summary(modelnb_finale)
```

```
## ******************************************************************
## Family:  c("NBI", "Negative Binomial type I")
##
## Call:  gamlss(formula = reflection ~ age + systbp + factor(hearimp) +
##     factor(eye) + random(factor(IdNum)), family = NBI,      data = data)
##
## Fitting method: RS()
##
## ------------------------------------------------------------------
## Mu link function:  log
## Mu Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.6220216  0.0746675  21.723  < 2e-16 ***
## age               0.0217548  0.0008683  25.056  < 2e-16 ***
## systbp           -0.0017851  0.0004260  -4.190 3.14e-05 ***
## factor(hearimp)1  0.1574077  0.0349517   4.504 7.82e-06 ***
## factor(eye)2     -0.0144009  0.0167019  -0.862    0.389
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## Sigma link function:  log
## Sigma Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -4.3704     0.1839  -23.77   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## NOTE: Additive smoothing terms exist in the formulas:
##  i) Std. Error for smoothers are for the linear effect only.
## ii) Std. Error for the linear terms maybe are not accurate.
## ------------------------------------------------------------------
## No. of observations in the fit:  1032
## Degrees of Freedom for the fit:  328.8757
##       Residual Deg. of Freedom:  703.1243
##                       at cycle:  9
##
## Global Deviance:     5699.026
##             AIC:     6356.777
##             SBC:     7981.178
## ******************************************************************
```

We can see that the factor(eye) is not a significant term in this model so we can confirm the fact that mean VA is same for both the eyes.