Categorical Analysis Project:
Purposeful Variable Selection for the PBC Data

Jaehwan Han

Part 1

```
width=c(22.5,23.5,24.5,25.5,26.5,27.5,28.5,30)
freq=c(3,4,10,18,20,21,15,20)
x1=rep(width,freq)
freq2=c(5,11,13,15,9,5,4,1)
x2=rep(width,freq2)
x=c(x1,x2)

d1=rep(1,length(x1))
d2=rep(0,length(x2))
d=c(d1,d2)

fit.glm <- glm(d~x, family=binomial(logit))
summary(fit.glm)
```

```
library(broom)
model.data <- augment(fit.glm)
model.data
```

```
## # A tibble: 174 x 9
##        d     x .fitted .se.fit .resid   .hat .sigma .cooksd .std.resid
##  * <dbl> <dbl>   <dbl>   <dbl>  <dbl>  <dbl>  <dbl>   <dbl>      <dbl>
##  1     1  22.5  -1.21    0.375   1.71 0.0250   1.06 0.0439        1.74
## # ... with 164 more rows
```

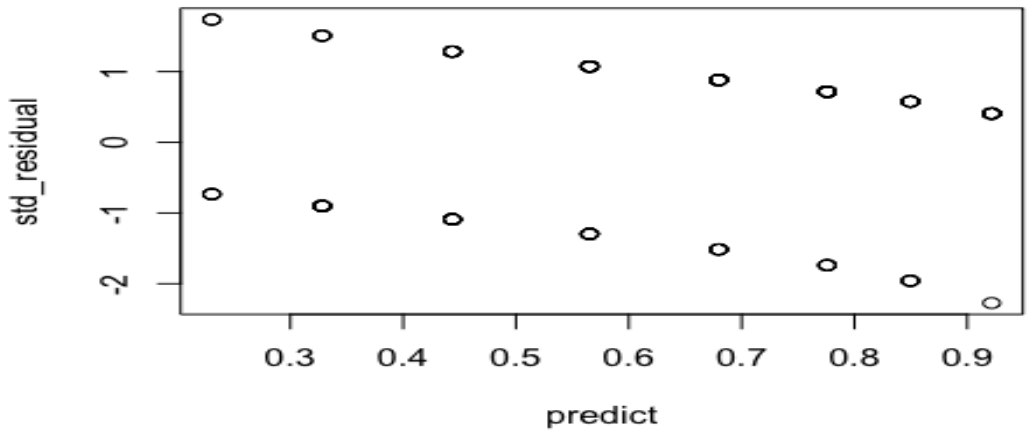
```
#Standarized Pearson residuals
std_residual=model.data$.std.resid
std_residual
```

```
##   [1]  1.7351659  1.7351659  1.7351659  1.5069575  1.5069575  1.5069575
```

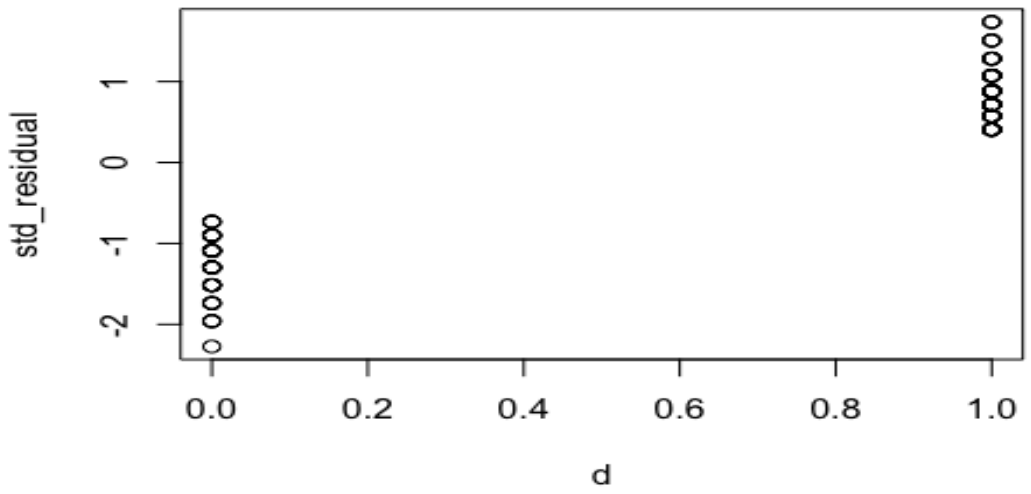
```
#Predicted Probabilities
predict <- predict(fit.glm, type = "response")
predict
```

```
##         1         2         3         4         5         6         7
## 0.2304300 0.2304300 0.2304300 0.3281900 0.3281900 0.3281900 0.3281900
```

```
#Standarized person residuals vs predicted probability
plot(predict,std_residual)
```

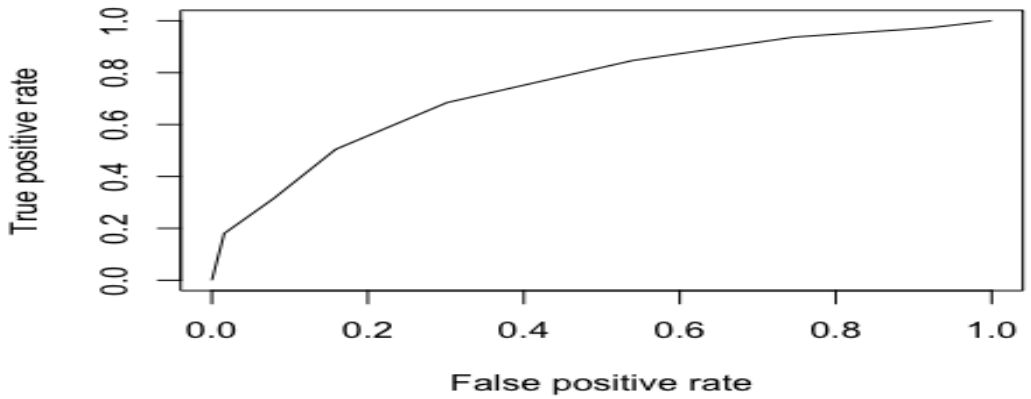



```
#Standarized person residuals vs index of covariate pattern
plot(d,std_residual)
```




```
ROC curve (I just want to see ROC value of this model which is not required.)
library(ROCR)
pr <- prediction(predict, d)
```

```
prf <- performance(pr, measure = "tpr", x.measure = "fpr")
plot(prf)
auc <- performance(pr, measure = "auc")
auc <- auc@y.values[[1]]
auc #0.7460
```




```
## [1] 0.7460317
```

Question D

```
#Saturated model
a <- factor(1:length(d))
fit <- glm(d~a,family=binomial("logit"), maxit=100)
summary(fit)

##
## Call:
## glm(formula = d ~ a, family = binomial("logit"), maxit = 100)
##
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.757e+01  5.871e+05       0        1
## a2          -9.277e-06  8.303e+05       0        1
## a3          -4.597e-07  8.304e+05       0        1
## a4          -4.358e-07  8.304e+05       0        1
##
##     Null deviance: 2.2780e+02  on 173  degrees of freedom
## Residual deviance: 3.7137e-10  on   0  degrees of freedom
## AIC: 348
##
## Number of Fisher Scoring iterations: 26
```

```r
#Log likelihood of the saturated model
logLik(fit)
```

```
## 'log Lik.' -1.85679e-10 (df=174)
```

Part I, Probelm 2

Question A

```r
# #Device(Present/Absent)->x,Years->x1,Number of accidents->y
x=c(0,0,0,0,0,0,0,0,1,1,1,1,1,1)
x1=c(8,8,7,7,8,7,8,7,2,3,2,2,2,3)
y=c(13,6,30,25,10,15,7,13,2,4,0,6,1,2)
df=data.frame(y,x,x1)

#Fit of x as a covariate, a poisson loglinear model
model=glm(y~1+x,family=poisson(link=log),data=df)
summary(model)
```

```
##
## Call:
## glm(formula = y ~ 1 + x, family = poisson(link = log), data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6182  -1.2786  -0.4124   0.6619   3.4411
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.69968    0.09167  29.450  < 2e-16 ***
## x           -1.78339    0.27399  -6.509 7.57e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 107.213  on 13  degrees of freedom
## Residual deviance:  42.554  on 12  degrees of freedom
## AIC: 96.023
##
## Number of Fisher Scoring iterations: 5
```

```r
#Define where device is not there
d1=which(x==0)
#Define where device is there
d2=which(x==1)

#Dataset where device is not there
d11=df[d1,]
#Dataset where device is there
d12=df[d2,]
```

```r
## Ratio of accidents where device is there and device is not there
Ratio=sum(exp(predict(model,newdata=d12)))/sum(exp(predict(model,newdata=d11)
))
Ratio
```

```
## [1] 0.1260504
```

```r
#Fit of x as a covariate,and log(years) as offset for a poisson loglinear mod
el
model1=glm(y~1+x+offset(log(x1)),family=poisson(link=log),data=df)
summary(model1)
```

```
## glm(formula = y ~ 1 + x + offset(log(x1)), family = poisson(link = log),
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8397  -1.4042  -0.4841   0.3905   3.7413
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.68478    0.09167   7.470 8.02e-14 ***
## x           -0.61579    0.27399  -2.247   0.0246 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 55.034  on 13  degrees of freedom
## Residual deviance: 49.119  on 12  degrees of freedom
## AIC: 102.59
##
```

```r
Ratio1=sum(exp(predict(model1,newdata=d12)))/sum(exp(predict(model1,newdata=d
11)))
Ratio1
```

```
## [1] 0.1260504
```

```r
#Pseudo R square
R2<- 1-(model$deviance/model$null.deviance)
R2_1<- 1-(model1$deviance/model1$null.deviance)
R2
```

```
## [1] 0.6030922
```

```r
R2_1
```

```
## [1] 0.1074886
```

Question B)

```r
#Wald Test for coefficient of the model where we have taken only the x as cov
ariate
summary(model)
```

```
##
## Call:
## glm(formula = y ~ 1 + x, family = poisson(link = log), data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6182  -1.2786  -0.4124   0.6619   3.4411
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.69968    0.09167  29.450  < 2e-16 ***
## x           -1.78339    0.27399  -6.509 7.57e-11 ***
```

```r
#Fitting of negative binomial using glm.nb function.
require(MASS)
```

```
## Loading required package: MASS
```

```r
model2=glm.nb(y~x, data = df)
summary(model2)
```

```
##
## Call:
## glm.nb(formula = y ~ x, data = df, init.theta = 4.819377101,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0069  -0.8605  -0.2596   0.5191   1.5517
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.6997     0.1853  14.568  < 2e-16 ***
## x            -1.7834     0.3682  -4.843 1.28e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(4.8194) family taken to be 1)
##
##     Null deviance: 40.530  on 13  degrees of freedom
## Residual deviance: 15.354  on 12  degrees of freedom
## AIC: 84.036
```

```r
#Pseudo R square
R2_2<- 1-(model2$deviance/model2$null.deviance)
R2_2
```

```
## [1] 0.6211726
```

```r
#Check AIC of two models
summary(model)
```

```
##
##     Null deviance: 107.213  on 13  degrees of freedom
## Residual deviance:  42.554  on 12  degrees of freedom
## AIC: 96.023
##
## Number of Fisher Scoring iterations: 5
```

```r
summary(model2)
```

```
##     Null deviance: 40.530  on 13  degrees of freedom
## Residual deviance: 15.354  on 12  degrees of freedom
## AIC: 84.036
##
```

Question C

```r
model3=glm(y~1+x,family=quasipoisson(link=log),data=df)
summary(model3)
```

```
##
## Call:
## glm(formula = y ~ 1 + x, family = quasipoisson(link = log), data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6182  -1.2786  -0.4124   0.6619   3.4411
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.6997     0.1739   15.52 2.63e-09 ***
## x            -1.7834     0.5199   -3.43  0.00498 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 3.600561)
##
##     Null deviance: 107.213  on 13  degrees of freedom
## Residual deviance:  42.554  on 12  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

```r
#Value of scale parameter
deviance(model)
```

```
## [1] 42.55368
```

```r
pr <- residuals(model,"pearson")
sum(pr^2)
```

```
## [1] 43.20672
```

```r
phi <- sum(pr^2)/df.residual(model)
round(c(phi,sqrt(phi)),4)
```

```
## [1] 3.6006 1.8975
```

Part II : Data Analysis Project

```r
library(foreign)
library(caTools)
getwd()
```

```
## [1] "/Users/jaehwanhan/Desktop"
```

```r
#Load the data and check for missing values
cirrhosisData = read.dta("PBC.dta")
head(cirrhosisData)
```

```
##        age sex edema bili albumin mort3yr hstage
## 1 58.76523   1     1 14.5    2.60       1      3
## 2 56.44627   1     0  1.1    4.14       0      2
## 3 70.07255   0     1  1.4    3.48       1      3
```

```r
sum(is.na(cirrhosisData))
```

```
## [1] 0
```

```r
summary(cirrhosisData)
```

```r
#Create training and validaiton set
(Used to provide an unbiased evaluation of model fit)
# (Data set is used to minimize overfitting)
set.seed(4)
split = sample.split(cirrhosisData$mort3yr, SplitRatio = 0.7)
dt = subset(cirrhosisData,split == TRUE)
dv = subset(cirrhosisData, split == FALSE)

#train basic model
model = glm(mort3yr~., data = dt, family = binomial)
nothing = glm(mort3yr~1, data = dt, family = binomial)
summary(model)
```

```r
#Implement forward stepwise procedure
forward = step(nothing, scope = list(lower=formula(nothing),upper=formula(mod
el)), direction = c("forward"))
```

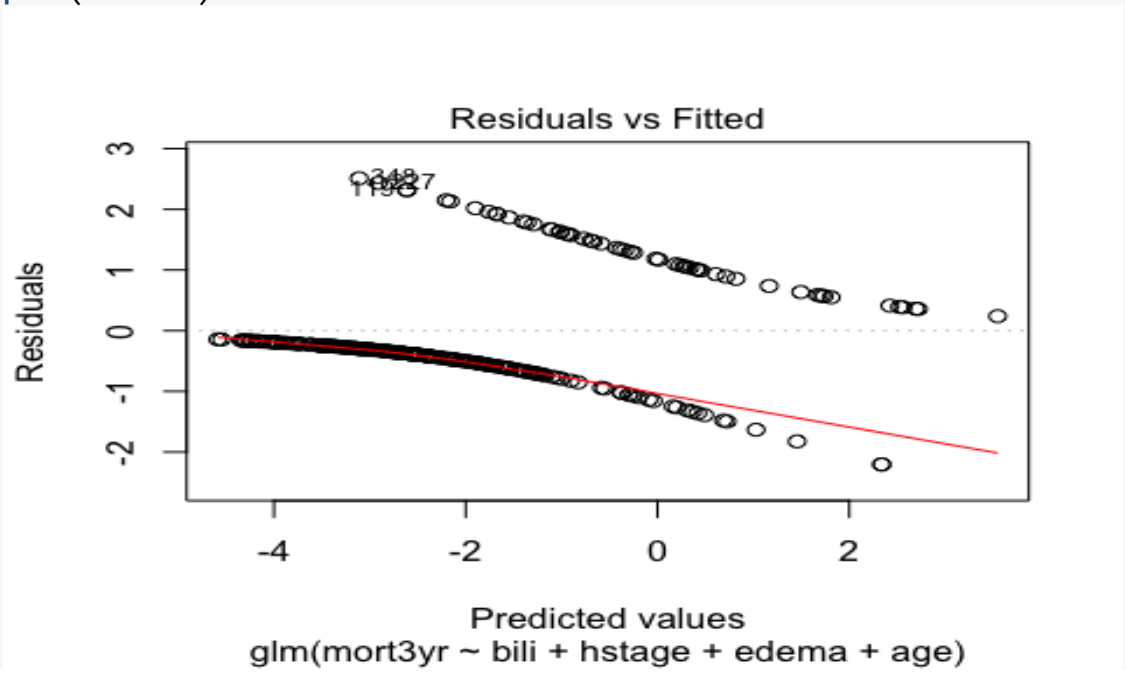
```
## Start:  AIC=282.45
## mort3yr ~ 1
##
##           Df Deviance    AIC
## + bili     1   235.25 239.25
## + edema    1   243.21 247.21
## + albumin  1   249.62 253.62
## + hstage   1   251.07 255.07
## + age      1   262.28 266.28
## + sex      1   277.76 281.76
## <none>         280.45 282.45
##
## Step:  AIC=239.25
## mort3yr ~ bili
##
##           Df Deviance    AIC
## + hstage   1   213.11 219.11
## + edema    1   213.77 219.77
## + age      1   214.12 220.12
## + albumin  1   222.74 228.74
## + sex      1   231.27 237.27
## <none>         235.25 239.25
##
## Step:  AIC=219.11
## mort3yr ~ bili + hstage
##
##           Df Deviance    AIC
## + edema    1   198.77 206.77
## + age      1   200.28 208.28
## + albumin  1   208.03 216.03
## + sex      1   210.27 218.27
## <none>         213.11 219.11
##
## Step:  AIC=206.77
## mort3yr ~ bili + hstage + edema
##
##           Df Deviance    AIC
## + age      1   191.90 201.90
## + sex      1   196.21 206.21
## <none>         198.77 206.77
## + albumin  1   197.22 207.22
##
## Step:  AIC=201.9
## mort3yr ~ bili + hstage + edema + age
##
##           Df Deviance    AIC
```

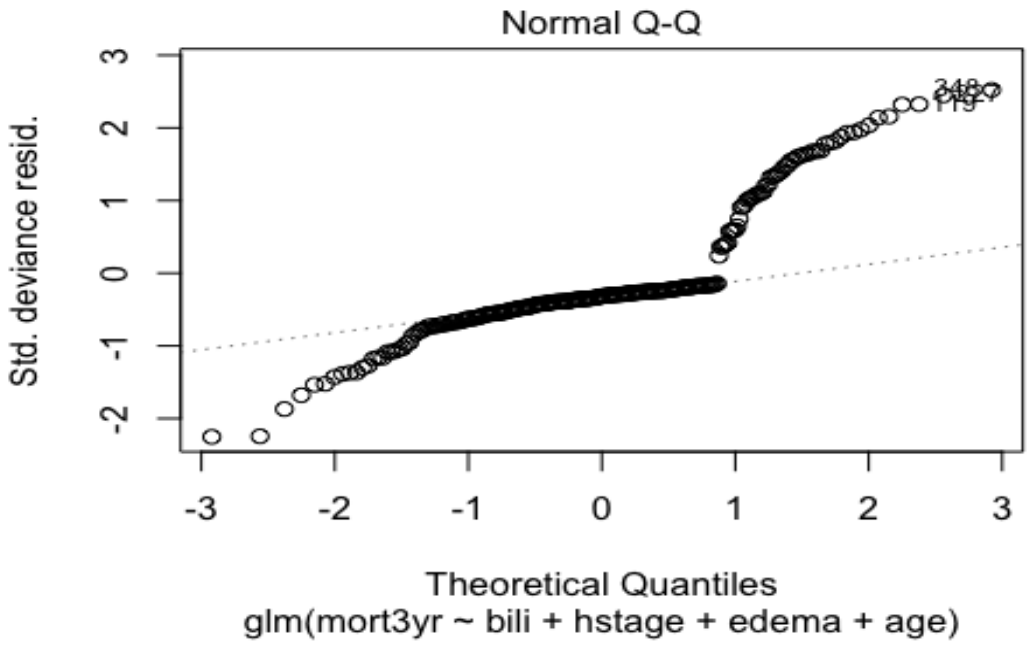
```
## <none>          191.90 201.90
## + albumin  1   190.25 202.25
## + sex      1   191.20 203.20
```

```
summary(forward) ## Using 0.05 significance level, bili, hstage, edma and age
is selected.
```

```
## glm(formula = mort3yr ~ bili + hstage + edema + age, family = binomial,
##     data = dt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2057  -0.5041  -0.3163  -0.1892   2.5117
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.99898    1.16771  -5.994 2.05e-09 ***
## bili         0.20261    0.04022   5.037 4.72e-07 ***
## hstage       0.85627    0.27040   3.167  0.00154 **
## edema        1.25969    0.43016   2.928  0.00341 **
## age          0.04693    0.01854   2.531  0.01137 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 280.45  on 286  degrees of freedom
## Residual deviance: 191.90  on 282  degrees of freedom
## AIC: 201.9
```

```
plot(forward)
```






```
# Check if albumin is significant by adding albumin variable
m1<- glm(formula = mort3yr ~ bili + hstage + age + edema + albumin,
    family = binomial, data = dt)
m2<- glm(formula = mort3yr ~ bili + hstage + age + edema,
    family = binomial, data = dt)
summary(m1)

##
## Call:
## glm(formula = mort3yr ~ bili + hstage + age + edema + albumin,
##     family = binomial, data = dt)
##     Null deviance: 280.45  on 286  degrees of freedom
## Residual deviance: 190.25  on 281  degrees of freedom
## AIC: 202.25

summary(m2)

##
## Call:
## glm(formula = mort3yr ~ bili + hstage + age + edema, family = binomial,
##     data = dt)
##     Null deviance: 280.45  on 286  degrees of freedom
## Residual deviance: 191.90  on 282  degrees of freedom
## AIC: 201.9
```

Compare AIC level of M1, M2, you could see M2 AIC level is lower than M1. (Wi thout Albumin is better)

```
library(rsq)
rsq(m1, adj=TRUE)
```

```
## [1] 0.3164038
```

```
rsq(m2, adj=TRUE)
```

```
## [1] 0.3179628
```

Without albumin adj R squared model is greater than the model with albumin model. Check AIC, R Squared value between two model with albuim and without albumin. albumin variable should not be included.

```
# Check if sex is significant by adding sex variable
m3<- glm(formula = mort3yr ~ bili + hstage + age + edema + sex,
    family = binomial, data = dt)
m4<- glm(formula = mort3yr ~ bili + hstage + age + edema,
    family = binomial, data = dt)
summary(m3)
```

```
##
## Call:
## glm(formula = mort3yr ~ bili + hstage + age + edema + sex, family = binomi
al,
##     data = dt)
##     Null deviance: 280.45  on 286  degrees of freedom
## Residual deviance: 191.20  on 281  degrees of freedom
## AIC: 203.2
```

```
summary(m4)
```

```
##
## Call:
## glm(formula = mort3yr ~ bili + hstage + age + edema, family = binomial,
##     data = dt)
##
##     Null deviance: 280.45  on 286  degrees of freedom
## Residual deviance: 191.90  on 282  degrees of freedom
## AIC: 201.9
##
## Number of Fisher Scoring iterations: 5
```

Compare AIC level of M3, M4, you could see M4 AIC level is lower than M3.

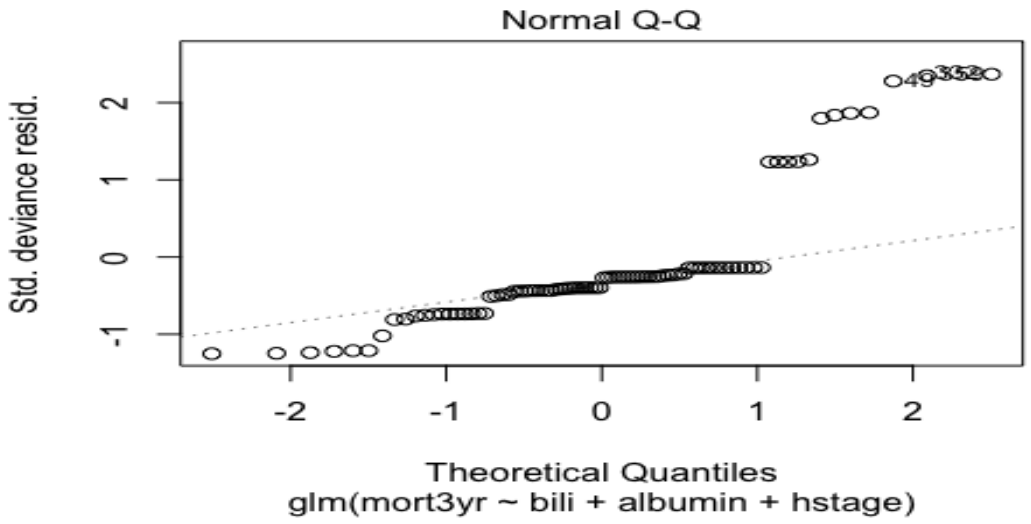
```
anova(m3,m4, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: mort3yr ~ bili + hstage + age + edema + sex
## Model 2: mort3yr ~ bili + hstage + age + edema
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
```

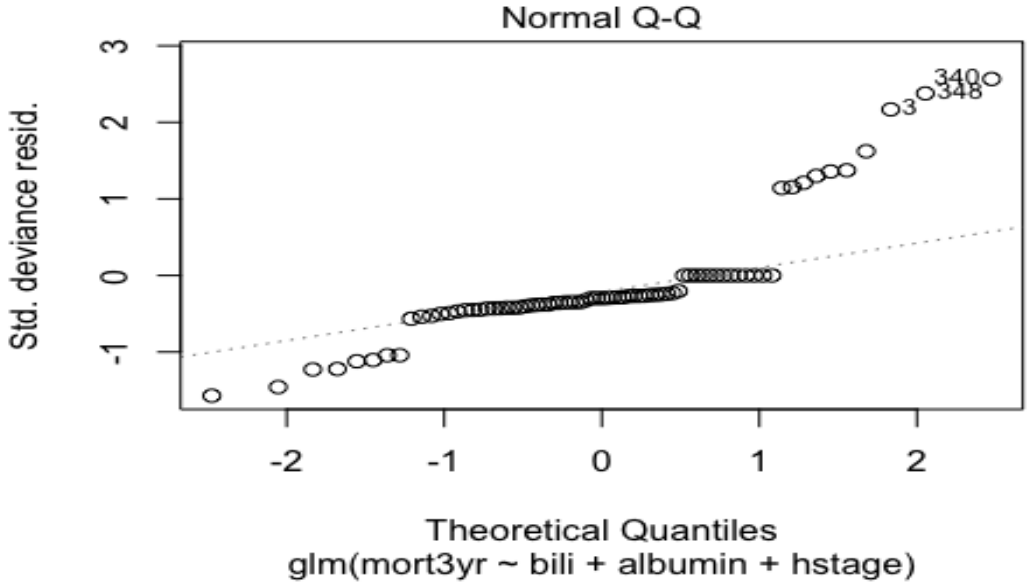
```
## 1       281      191.2
## 2       282      191.9 -1 -0.70205   0.4021
```

Using Anova test is also good way to decide variable selection. P-value is greater than 0.05, Fail to reject
Null hypothesis. We can conclude that Sex varialbe is not signficant. It should be removed from model.

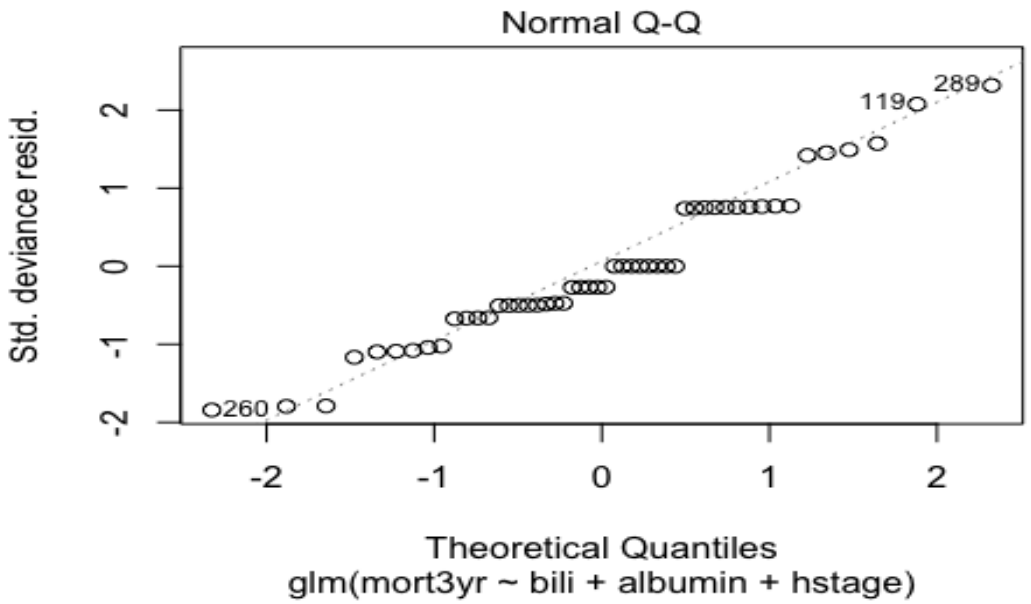
```
#Variable Categorisation (Continous variable Age, bili)
age <- as.factor(cut(dt$age, c(quantile(dt$age)[1] - 1,quantile(dt$age)[2],quantile(dt$age)[3],quantile(dt$age)[4],quantile(dt$age)[5] + 1)))
bili <- as.factor(cut(dt$bili, c(quantile(dt$bili)[1] - 1,quantile(dt$bili)[2],quantile(dt$bili)[3],quantile(dt$bili)[4],quantile(dt$bili)[5] + 1)))
dt1 = dt
dt1$age = age
dt1$bili = bili

age <- as.factor(cut(dv$age, c(quantile(dt$age)[1] - 1,quantile(dt$age)[2],quantile(dt$age)[3],quantile(dt$age)[4],quantile(dt$age)[5] + 1)))
bili <- as.factor(cut(dv$bili, c(quantile(dt$bili)[1] - 1,quantile(dt$bili)[2],quantile(dt$bili)[3],quantile(dt$bili)[4],quantile(dt$bili)[5] + 1)))
dv1 = dv
dv1$age = age
dv1$bili = bili

#Create New Age Model
levels(dt1$age)

## [1] "(25.3,43]"   "(43,51.3]"   "(51.3,58.9]" "(58.9,79.4]"

ageModel1 = glm(mort3yr~bili+albumin+hstage, data = dt1[age=="(43,51.3]",], family = binomial)
ageModel2 = glm(mort3yr~bili+albumin+hstage, data = dt1[age=="(51.3,58.9]",], family = binomial)
ageModel3 = glm(mort3yr~bili+albumin+hstage, data = dt1[age=="(58.9,79.4]",], family = binomial)
summary(ageModel1)

summary(ageModel2)

summary(ageModel3)

plot(ageModel1,2)
```



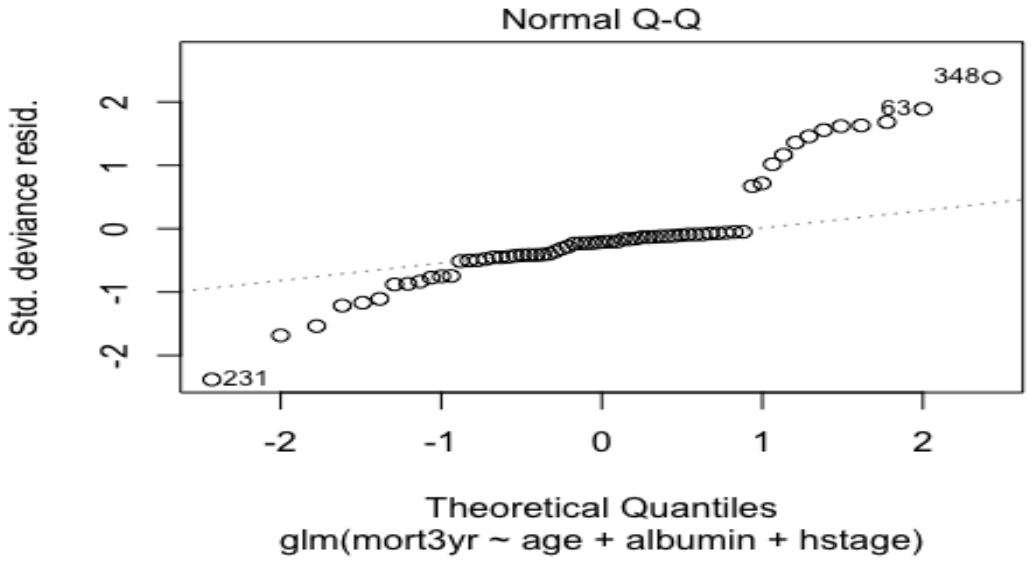

```
plot(ageModel2,2)
```



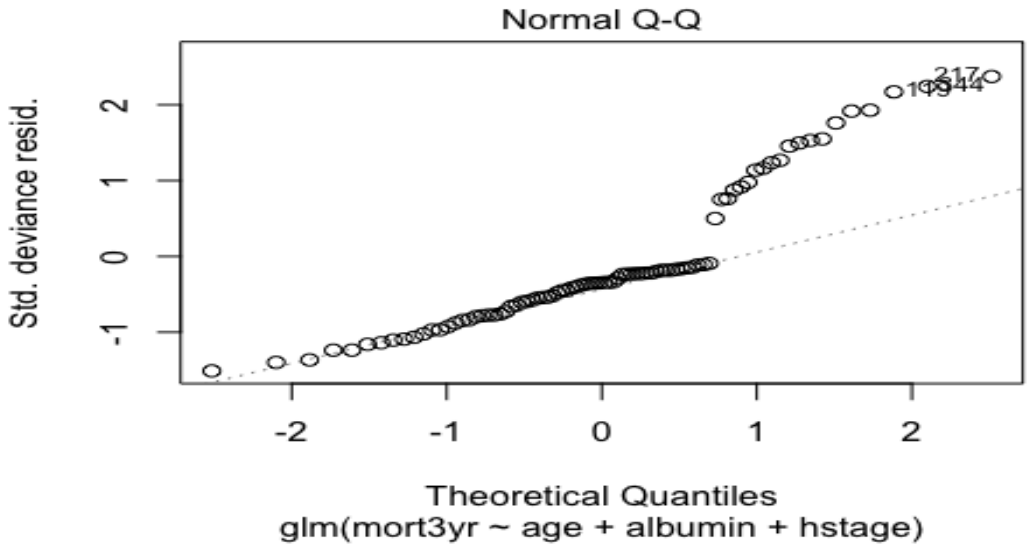

```
plot(ageModel3,2)
```



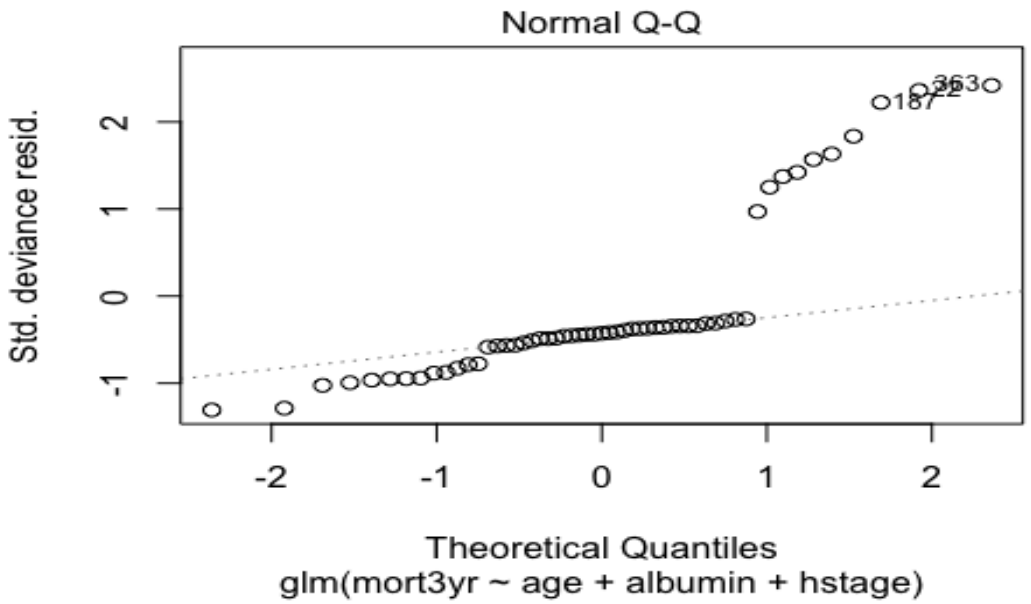

```r
#Create new bili models
levels(dt1$bili)

## [1] "(-0.7,0.8]" "(0.8,1.4]"  "(1.4,3.5]"  "(3.5,26.5]"

biliModel1 = glm(mort3yr~age+albumin+hstage, data = dt1[bili=="(0.8,1.4]" ,],
family = binomial)
biliModel2 = glm(mort3yr~age+albumin+hstage, data = dt1[bili=="(1.4,3.5]" ,],
family = binomial)
biliModel3 = glm(mort3yr~age+albumin+hstage, data = dt1[bili=="(3.5,26.5]" ,]
, family = binomial)
summary(biliModel1)

summary(biliModel2)

summary(biliModel3)

plot(biliModel1,2)
```




```
plot(biliModel2,2)
```




```
plot(biliModel3,2)
```

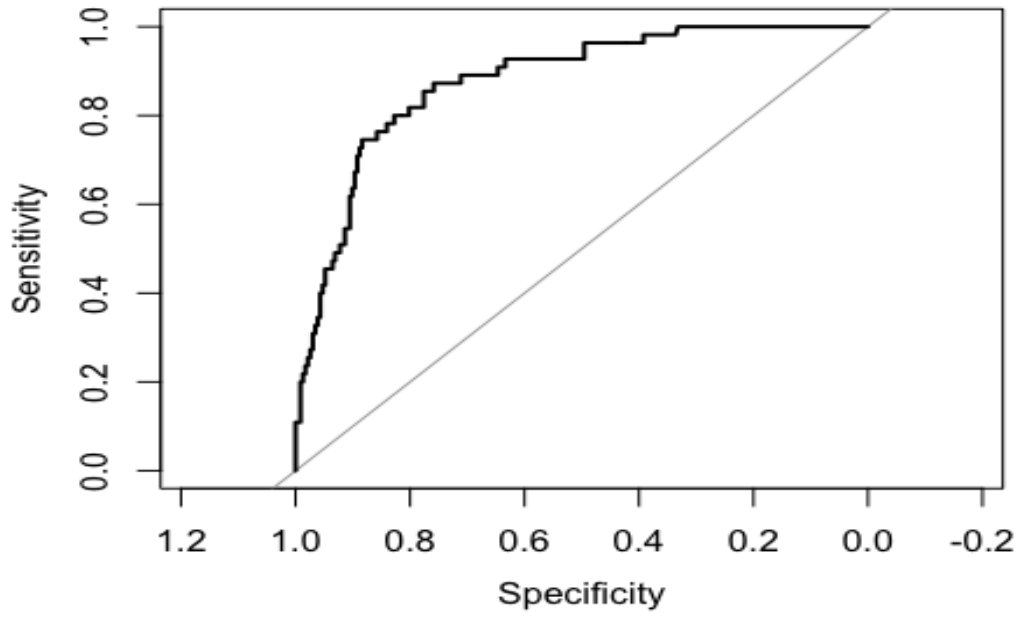



```
# Interaction
Inter = glm(mort3yr ~ bili*age*edema*hstage, data=dt1, family = "binomial")

summary(Inter)

##
## Call:
## glm(formula = mort3yr ~ bili * age * edema * hstage, family = "binomial",
##     data = dt1)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
##  -8.49    0.00    0.00    0.00    8.49
##
## Coefficients: (14 not defined because of singularities)
##                                             Estimate Std. Error
## (Intercept)                                -4.190e+13  3.480e+07
## bili(0.8,1.4]                               4.190e+13  5.885e+07
## bili(1.4,3.5]                              -2.659e+15  5.942e+07
## bili(3.5,26.5]                             -2.961e+15  6.320e+07
## age(43,51.3]                               -8.940e+15  7.333e+07
## age(51.3,58.9]                             -4.462e+15  4.834e+07
## age(58.9,79.4]                             -6.063e+15  5.403e+07
## edema                                      -2.597e+15  1.513e+08
## hstage                                      5.967e+12  1.981e+07
## bili(0.8,1.4]:age(43,51.3]                  8.940e+15  1.007e+08
## bili(1.4,3.5]:age(43,51.3]                  7.787e+15  1.040e+08
## bili(3.5,26.5]:age(43,51.3]                 7.117e+15  1.072e+08
```

```r
#Hoslem Test
library(ResourceSelection)

h1 = hoslem.test(model$y,fitted(model), g=10)
h1

##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  model$y, fitted(model)
## X-squared = 9.8591, df = 8, p-value = 0.2751
```

Greater than 0.05 Fail to reject Ho, there is no evidence of poor fit.

```r
#AUC
prob=predict(forward,type=c("response"))
dt$prob=prob
library(pROC)

g <- roc(mort3yr ~ prob, data = dt)
g

##
## Call:
## roc.formula(formula = mort3yr ~ prob, data = dt)
##
## Data: prob in 232 controls (mort3yr 0) < 55 cases (mort3yr 1).
## Area under the curve: 0.8782

plot(g)
```



```
summary(forward)
```

```
##
## Call:
## glm(formula = mort3yr ~ bili + hstage + edema + age, family = binomial,
##     data = dt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2057  -0.5041  -0.3163  -0.1892   2.5117
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.99898    1.16771  -5.994 2.05e-09 ***
## bili         0.20261    0.04022   5.037 4.72e-07 ***
## hstage       0.85627    0.27040   3.167  0.00154 **
## edema        1.25969    0.43016   2.928  0.00341 **
## age          0.04693    0.01854   2.531  0.01137 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 280.45  on 286  degrees of freedom
## Residual deviance: 191.90  on 282  degrees of freedom
## AIC: 201.9
##
## Number of Fisher Scoring iterations: 5
```

FINAL MODEL which include bili, hstage, edema, age variable.

Log(p/1-p) = -6.99898 + 0.20261 bili + 0.85627 hstage + 1.25969 edema + 0.04693 age

P-value is all close to 0. They are all highly significant.

(Using test statistics : Beta/SD(Beta), H0: Beta=0, H1: Beta is not 0)

Change in odds = e^0.20261 (bili)= 1.22459. It indicates that the risk of mortality within 3 years increase by a factor of 1.224 or22% for each bilirubin increase.

Change in odds = e^0.85627 (hstage) = 2.354

Change in odds = e^1.25969 (edema) = 3.243. It indicate that the risk of mortaility within 3 years Increase by a factor of 3.243 or presence of edema is 3 times greater than no edema.

Change in odds = e^0.04693 (age) = 1.0480. It indicate that the risk of mortality within 3 years Increase by a factor of 1.0480 or 4% for every 1 age increase.

Since they all positive coefficient, we can tell that People who are getting old, have high serum bilirubin r ate, high level of histologic stage of disease and having edema is more likely to die within 3 year compare d to the people who are younger and doesn't have high serum, high level of stage of disease, edema.