Exercise2-1

Mehrab Atighi

4/14/2021

Introducing data to R

```
#after installing the msme packages we should library that`
library("msme")

## Loading required package: MASS

## Loading required package: lattice

data("medpar")
#we want to see small sample of our data:
head(medpar[,1:6])

##   los hmo white died age80 type
## 1   4   0     1    0     0    1
## 2   9   1     1    0     0    1
## 3   3   1     1    1     1    1
## 4   9   0     1    0     0    1
## 5   1   0     1    1     1    1
## 6   4   0     1    1     0    1
```

We know that the los column indicates the length of nights the person has been hospitalized.

hmo column indicates whether the person was covered by insurance or not (yes=1 ,No=0)

white column indicates whether the person is white or not (yes=1 , No=0)

died column indicates whether the person died within 48 hours of hospitalization or not (yes=1 ,No=0)

age80 column ididactes wheter the person age is more equal 80 or not (yes=1 ,No=0)

type coulmn idicates the person's kind of hospitalization (Optional=1 ,Instant=2 ,Emergency=3)

Solve:

chek correlation between the variables:

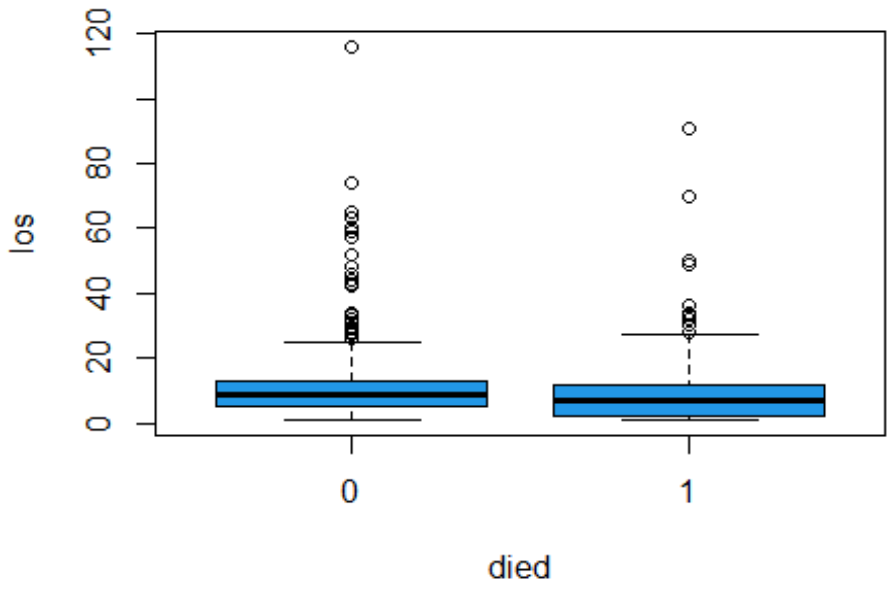
```
cor(medpar[,1:6])
```

```
##                los           hmo       white          died       age80
## los    1.00000000 -5.832123e-02 -0.06779545 -1.037458e-01 -0.03303782
## hmo   -0.05832123  1.000000e+00  0.05435482 -4.371603e-05 -0.03853239
## white -0.06779545  5.435482e-02  1.00000000  3.830089e-02  0.04647059
## died  -0.10374584 -4.371603e-05  0.03830089  1.000000e+00  0.13167978
## age80 -0.03303782 -3.853239e-02  0.04647059  1.316798e-01  1.00000000
## type   0.25511584 -1.127590e-01 -0.07471925  8.975658e-02 -0.03005332
##              type
## los    0.25511584
## hmo   -0.11275902
## white -0.07471925
## died   0.08975658
## age80 -0.03005332
## type   1.00000000
```

According to this matrix we can say that we dont have any significant dependence and correlation between variables.

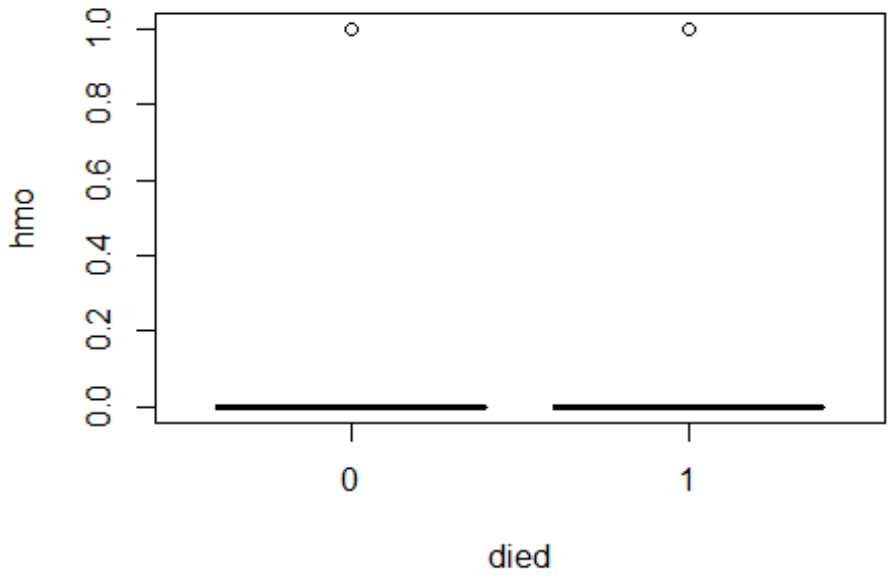
chek the relation beween response and variables with Boxplots:

```
#at the first we should attach the data
attach(medpar)
#now we want to see the Box plot of each variable with our response:
boxplot(los~died ,col=4)
```



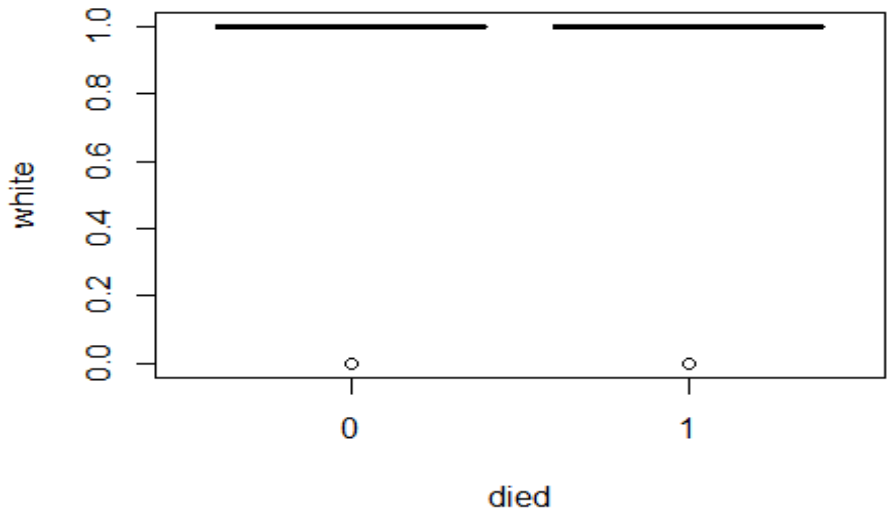
According to this Boxplot we can say that median of los for death and live persons are equal but for more los value we have more live persons.

```
boxplot(hmo~died ,col=3)
```



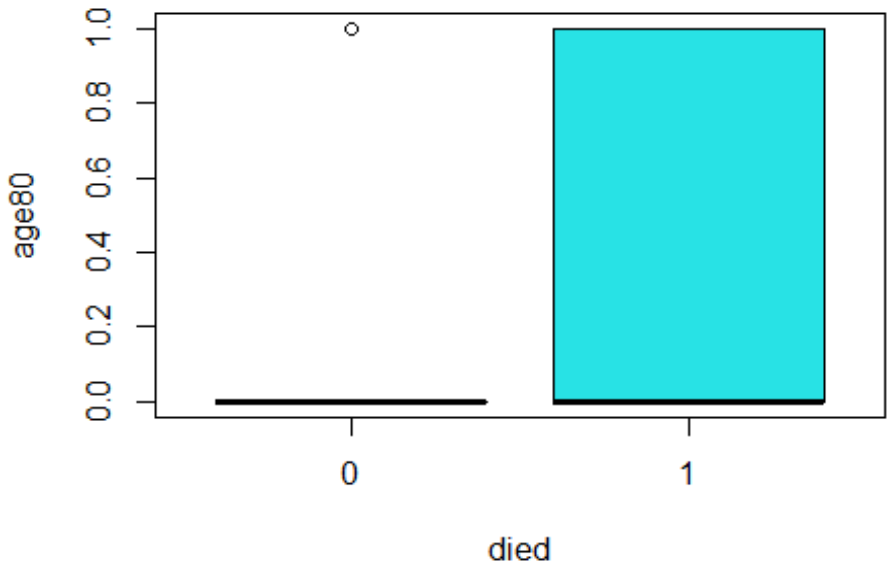
According to top Boxplot we can say that the majority of those admitted did not have insurance coverage.

```
boxplot(white~died ,col=2)
```



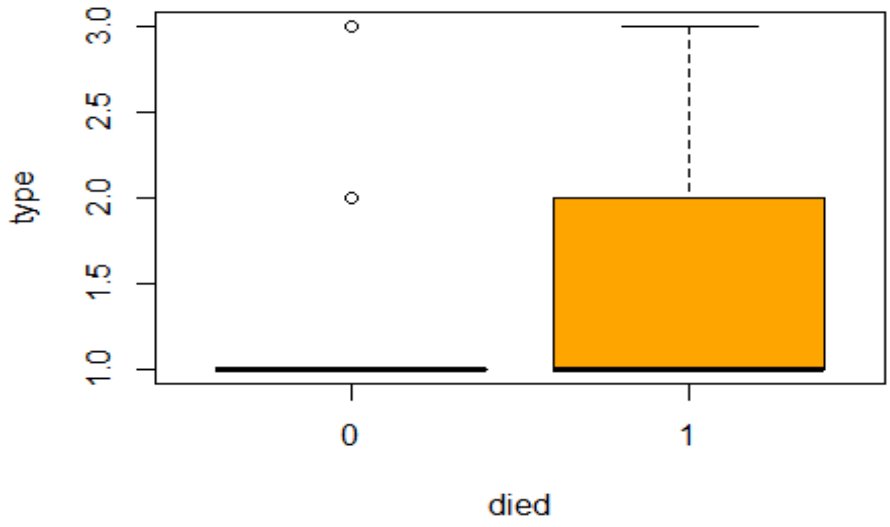
According to top Boxplot we can say that the majority of those admitted were whith white skin.

```
boxplot(age80~died ,col=85)
```



According to top Boxplot we can say that the majority of those admitted that death, are more and equal 80 years old.

```
boxplot(type~died, col="orange")
```



According to top Boxplot we can say that the majority of those admitted that their hospitalization were instant and emergency includ the majority of death.

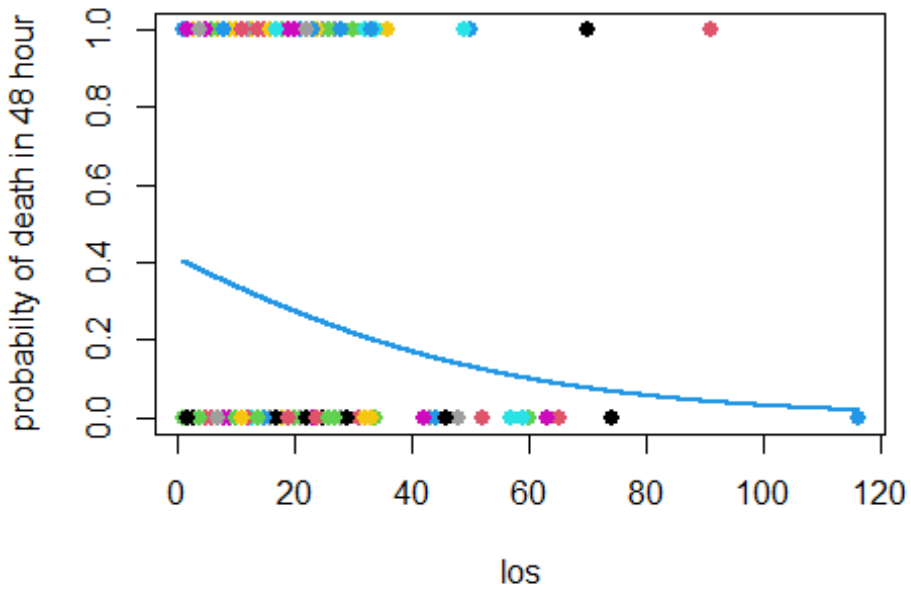
Logstics Regression with severan variables and univariables

Logstics regression for each variable and response:

```
fit1<-glm(died~los,family = binomial)
coef(fit1)

## (Intercept)         los
## -0.36170695 -0.03048316

summary(fit1)

##
## Call:
## glm(formula = died ~ los, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0160  -0.9449  -0.8767   1.3614   2.5212
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.361707   0.088436  -4.090 4.31e-05 ***
## los         -0.030483   0.007691  -3.964 7.38e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1922.9  on 1494  degrees of freedom
## Residual deviance: 1904.6  on 1493  degrees of freedom
## AIC: 1908.6
##
## Number of Fisher Scoring iterations: 4
```

According to the summary and coef function outputs we can say that we have nagative relationship(betha los = -0.03048316) and our p-value for signifacting H0 (Betha0 = Betha1 = 0)is equal to 7.38e-05 and its less than alpha(0.05) so we say the H0 reject and the regression is signifact.

```
seq1<-data.frame(los=seq(min(los),max(los),len=10^4))
seq1$died=predict(fit1,newdata= seq1,type="response")
plot(los,died,col= c(1:length(los)),pch=19,cex=1.1,ylab="probabilty of death
in 48 hour")
lines(died~los , seq1 ,col=4,lwd=2)
```



According to top plot we can see the probabilty of death in 48 hour for los value between (0, 35) is value between (0.2 , 0.4) and for los more equal than 60 day is less than 0.1 .

```r
fit2<-glm(died~hmo,family = binomial)
coef(fit2)
```

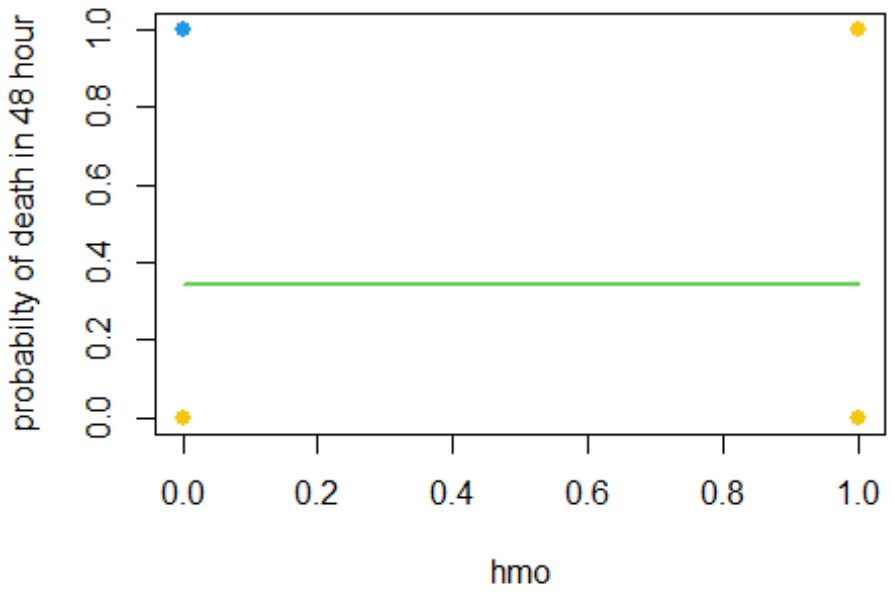
```
##   (Intercept)           hmo
## -0.6492752962 -0.0002512619
```

```r
summary(fit2)
```

```
##
## Call:
## glm(formula = died ~ hmo, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9169  -0.9169  -0.9169   1.4626   1.4627
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.6492753  0.0594332 -10.924   <2e-16 ***
## hmo         -0.0002513  0.1486501  -0.002    0.999
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1922.9  on 1494  degrees of freedom
## Residual deviance: 1922.9  on 1493  degrees of freedom
## AIC: 1926.9
##
## Number of Fisher Scoring iterations: 4
```

According to the summary and coef function outputs we can say that we have very Weak nagative relationship(betha hmo= -0.0002512619) and our p-value for signifacting H0 (Betha0 = Betha1 = 0)is equal to 0.999 and its not less than alpha(0.05) so we say the H0 accept and the regression is not signifact. i think its not good variable for our regression.

```
seq2<-data.frame(hmo=seq(min(hmo),max(hmo),len=10^4))
seq2$died=predict(fit2,newdata= seq2,type="response")
plot(hmo,died,col= c(1:length(hmo)),pch=19,cex=1.1,ylab="probabilty of death
in 48 hour")
lines(died~hmo , seq2 ,col=3,lwd=2)
```



According to top plot we can see the fix line about probabilty of death is equal to 0.34 and its fix when the hmo variable change, so we can say its not good variable for our regression.

```r
fit3<-glm(died~white,family = binomial)
coef(fit3)
```

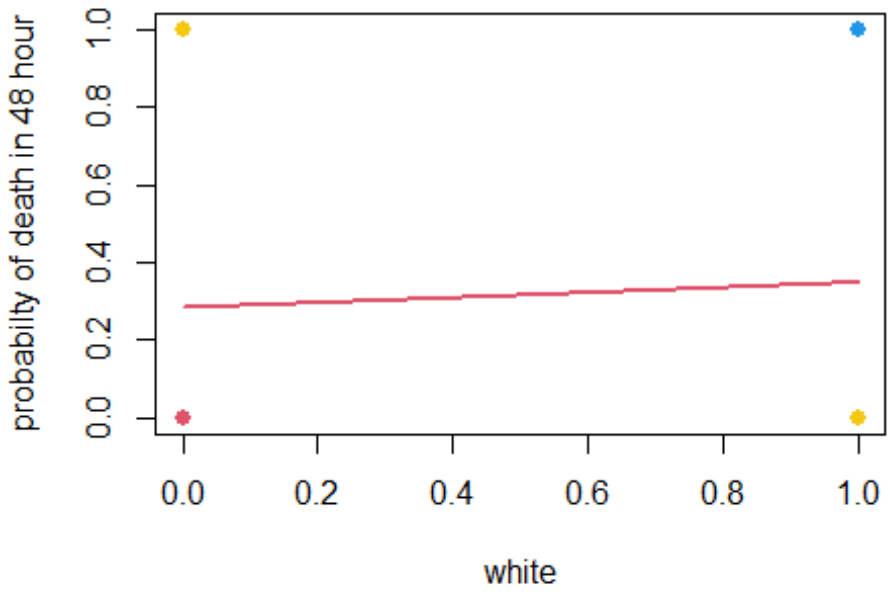
```
## (Intercept)       white
##  -0.9273406   0.3025126
```

```r
summary(fit3)
```

```
##
## Call:
## glm(formula = died ~ white, family = binomial)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -0.926  -0.926  -0.926   1.452   1.588
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.9273     0.1969  -4.710 2.48e-06 ***
## white         0.3025     0.2049   1.476     0.14
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1922.9  on 1494  degrees of freedom
## Residual deviance: 1920.6  on 1493  degrees of freedom
## AIC: 1924.6
##
## Number of Fisher Scoring iterations: 4
```

According to the summary and coef function outputs we can say that we have weak positive relationship(betha white= 0.3025126) But our p-value for signifacting H0 (Betha0 = Betha1 = 0)is equal to 0.14 and its not less than alpha(0.05) so we say the H0 accept and the regression isn't signifact, so we should remove it.

```
seq3<-data.frame(white=seq(min(white),max(white),len=10^4))
seq3$died=predict(fit3,newdata= seq3,type="response")
plot(white,died,col= c(1:length(white)),pch=19,cex=1.1,ylab="probabilty of de
ath in 48 hour")
lines(died~white , seq3 ,col=2,lwd=2)
```

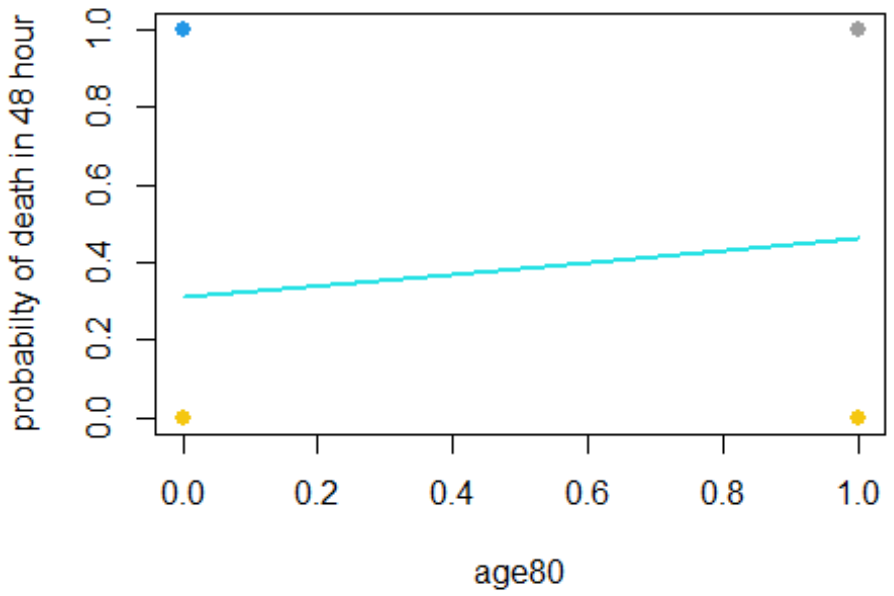


According to top plot we can see the line probabilty of death is between(0.283,0.36) and its apprixmimately fix when the white variable change, so we can say its not good variable for our regression.

```
fit4<-glm(died~age80,family = binomial)
coef(fit4)

## (Intercept)       age80
##  -0.8007213   0.6428183

summary(fit4)

##
## Call:
## glm(formula = died ~ age80, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1111  -0.8612  -0.8612   1.2452   1.5308
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.80072    0.06336 -12.639  < 2e-16 ***
## age80        0.64282    0.12732   5.049 4.45e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1922.9  on 1494  degrees of freedom
## Residual deviance: 1897.7  on 1493  degrees of freedom
## AIC: 1901.7
##
## Number of Fisher Scoring iterations: 4
```

According to the summary and coef function outputs we can say that we have positive relationship(betha age80= 0.64282) and our p-value for signifacting H0 (Betha0 = Betha1 = 0)is equal to 4.45e-07 and its less than alpha(0.05) so we say the H0 reject and the regression is signifact.

```
seq4<-data.frame(age80=seq(min(age80),max(age80),len=10^4))
seq4$died=predict(fit4,newdata= seq4,type="response")
plot(age80,died,col= c(1:length(age80)),pch=19,cex=1.1,ylab="probabilty of de
ath in 48 hour")
lines(died~age80 , seq4 ,col=85,lwd=2)
```



According to top plot we can see the line with positive slope about probabilty of death is between (0.309 ,0.46) and its increasing when the age80 variable increase.

```r
fit5<-glm(died~type,family = binomial)
coef(fit5)
```

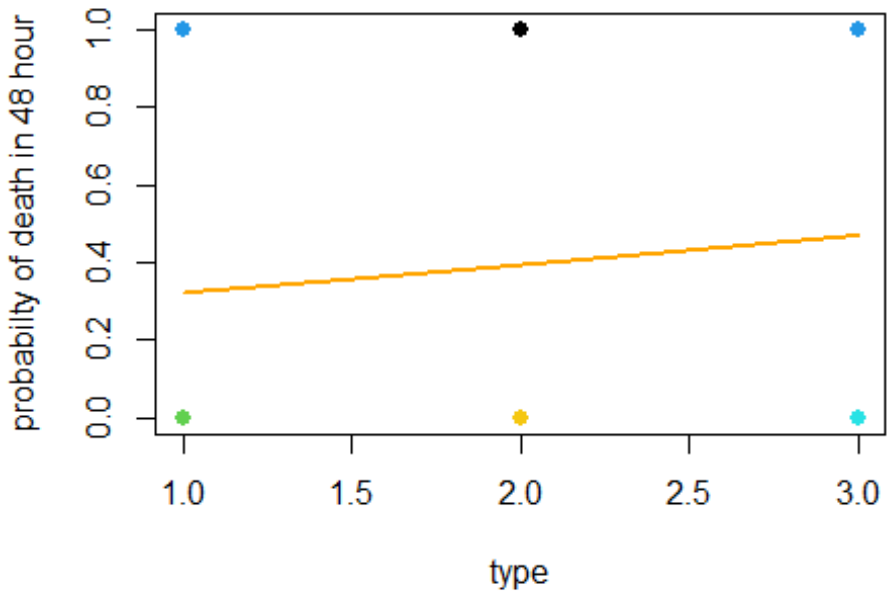
```
## (Intercept)        type
##  -1.0613241   0.3121013
```

```r
summary(fit5)
```

```
##
## Call:
## glm(formula = died ~ type, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1248  -0.8799  -0.8799   1.3678   1.5075
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.06132    0.13245  -8.013 1.12e-15 ***
## type         0.31210    0.09055   3.447 0.000568 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1922.9  on 1494  degrees of freedom
## Residual deviance: 1911.1  on 1493  degrees of freedom
## AIC: 1915.1
##
## Number of Fisher Scoring iterations: 4
```

According to the summary and coef function outputs we can say that we have positive relationship(betha type= 0.3121013) and our p-value for signifacting H0 (Betha0 = Betha1 = 0)is equal to 0.000568 and its less than alpha(0.05) so we say the H0 reject and the regression is signifact.

```
seq5<-data.frame(type=seq(min(type),max(type),len=10^4))
seq5$died=predict(fit5,newdata= seq5,type="response")
plot(type,died,col= c(1:length(type)),pch=19,cex=1.1,ylab="probabilty of deat
h in 48 hour")
lines(died~type , seq5 ,col="Orange",lwd=2)
```



According to top plot we can see the line with positive slope about probabilty of death is between (0.320 ,0.468) and its increasing when the age80 variable increase.

Logstics regression for all variable and response(Multiple logstic regression)

```
full.fit<-glm(died~los+hmo+white+age80+type,family = binomial)
coef(full.fit)

## (Intercept)         los         hmo       white       age80        type
## -1.32099883 -0.03679831  0.06156626  0.25907749  0.65167023  0.46900157

summary(full.fit)

##
## Call:
## glm(formula = died ~ los + hmo + white + age80 + type, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5196  -0.8928  -0.7923   1.2790   2.3160
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.320999   0.253700  -5.207 1.92e-07 ***
## los         -0.036798   0.007891  -4.663 3.11e-06 ***
## hmo          0.061566   0.152732   0.403    0.687
## white        0.259077   0.210062   1.233    0.217
## age80        0.651670   0.129545   5.030 4.89e-07 ***
## type         0.469002   0.097186   4.826 1.39e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1922.9  on 1494  degrees of freedom
## Residual deviance: 1856.2  on 1489  degrees of freedom
## AIC: 1868.2
##
## Number of Fisher Scoring iterations: 4
```

According to this models we can say that just the (hmo,white) variables p-values is more than 0.05 and its not good for our model and we should remove it, the others variable have positive relationships with response expect los.

Reduce logstics model

```
reduce.fit<-glm(died~los+age80+type,family = binomial)
coef(reduce.fit)

## (Intercept)         los       age80        type
##  -1.0540990  -0.0373605   0.6566579   0.4576727

summary(reduce.fit)

##
## Call:
## glm(formula = died ~ los + age80 + type, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5085  -0.8845  -0.8018   1.2856   2.2566
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.054099   0.147929  -7.126 1.04e-12 ***
## los         -0.037360   0.007876  -4.743 2.10e-06 ***
## age80        0.656658   0.129178   5.083 3.71e-07 ***
## type         0.457673   0.096384   4.748 2.05e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1922.9  on 1494  degrees of freedom
## Residual deviance: 1858.0  on 1491  degrees of freedom
## AIC: 1866
##
## Number of Fisher Scoring iterations: 4
```

We can see that the reduce models outputs show that the (hmo,white) variablse wasn't important and dont have effects on response.

the Reduce model is best model without any bad variable and all of them are signifact and we have good predict for our response,

we can see the logstics predicton function here for reduce model:

y = died , x1 = los , x2 = age80 , x3 = type.

$$P(X) = P(y = 1|X) = \frac{\exp\{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3\}}{1 + \exp\{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3\}}$$

$$= \frac{\exp\{-1.054099 + (-0.037360 * x_1) + (0.656658 * x_2) + (0.457673 * x_3)}{1 + \exp\{-1.054099 + (-0.037360 * x_1) + (0.656658 * x_2) + (0.457673 * x_3)}$$

End.