STAT3032S19_HW5

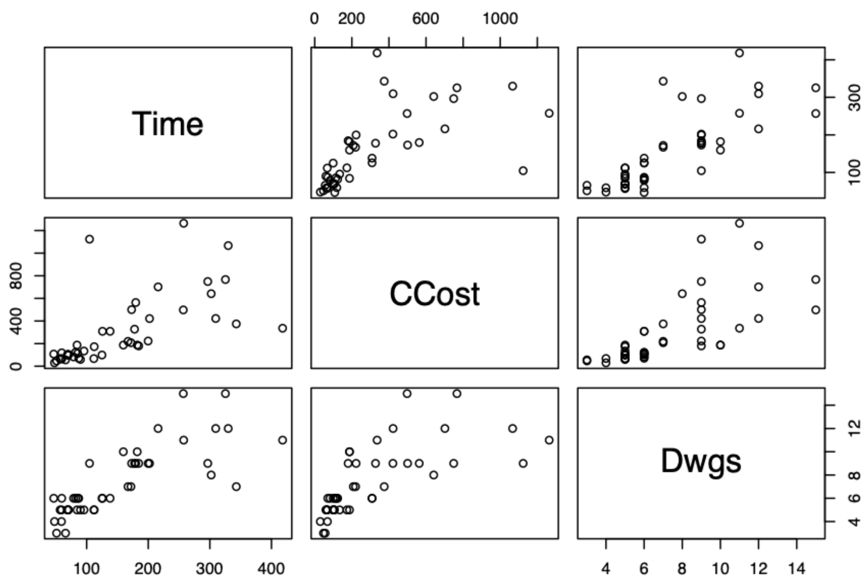
Mingming Xu

2019/4/15

Problem 1

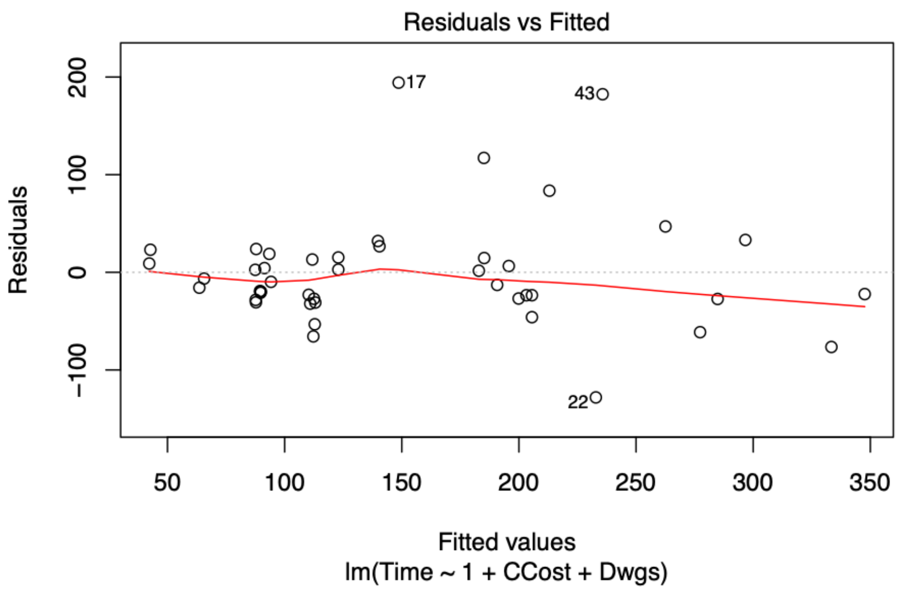
(a)

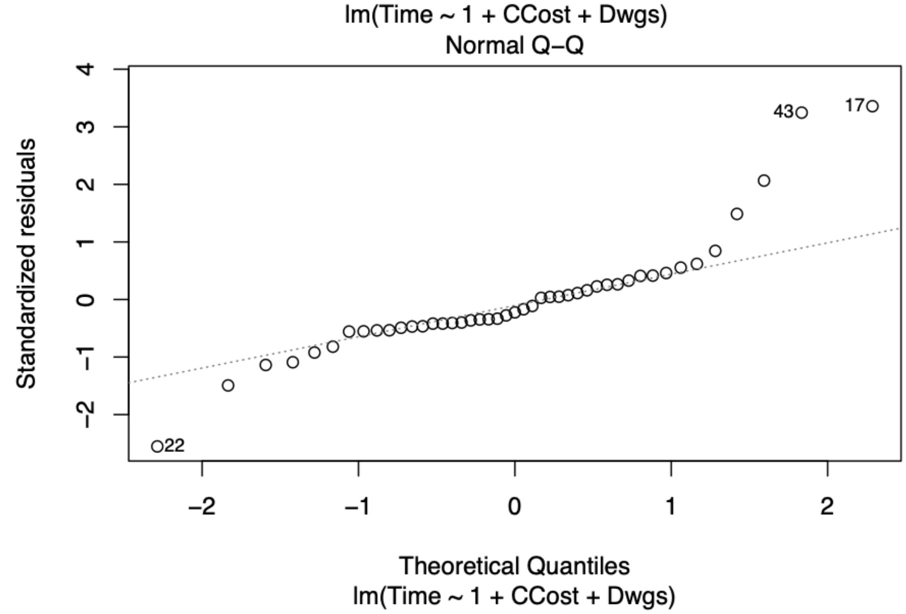
```
bridge=read.table("http://gattonweb.uky.edu/sheather/book/docs/datasets/bridge.txt",header = TRUE)
pairs(Time~CCost+Dwgs,data=bridge)
```

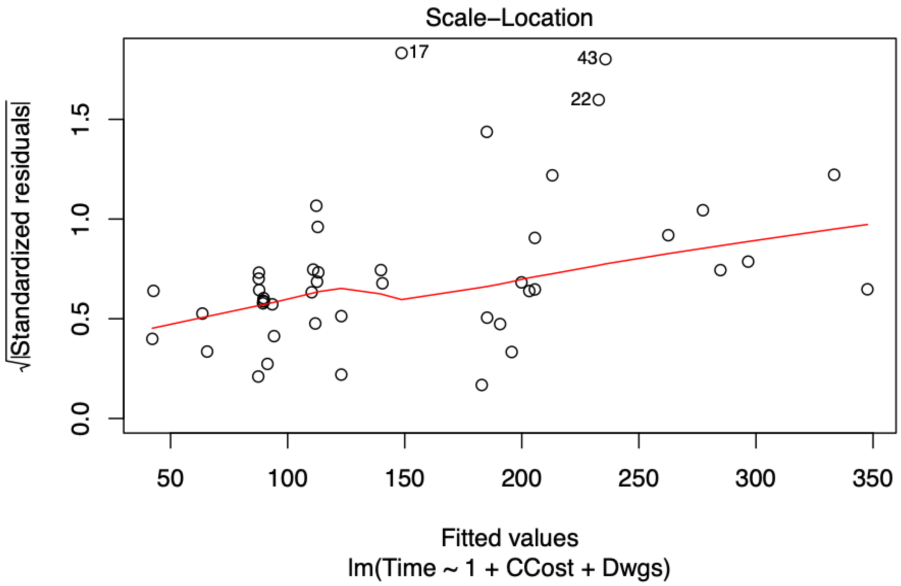


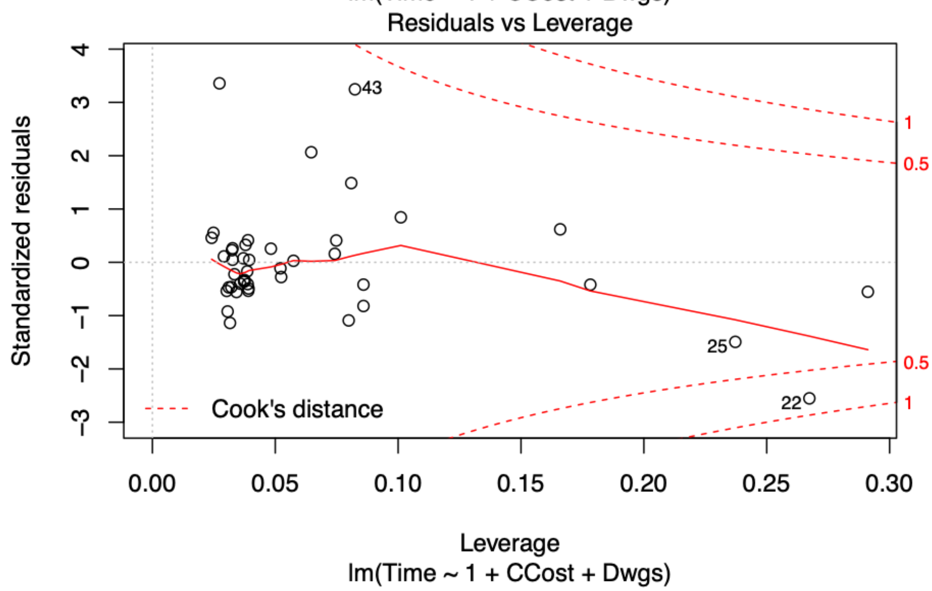
(b)

```
mod1_1=lm(Time~1+CCost+Dwgs,data=bridge)
plot(mod1_1)
```


Residuals vs Fitted
Residuals
200
100
0
−100
50
100
150
200
250
300
350
17
43
22
Fitted values
lm(Time ~ 1 + CCost + Dwgs)


Normal Q−Q
Standardized residuals
4
3
2
1
0
−1
−2
−2
−1
0
1
2
43
17
22
Theoretical Quantiles
lm(Time ~ 1 + CCost + Dwgs)





Violation: Linearity.Form the Residuals vs Fitted plot and from the marginal scatterplotsin part(a), we could not see there is no clear linearity in Time and CCost and in Time and Dwgs.

(c)

```
library(car)
```

```
## Loading required package: carData
```

```r
powerTransform(bridge[,4:5])
```

```
## Estimated transformation parameters
##      CCost       Dwgs
## -0.1895545 -0.1781791
```

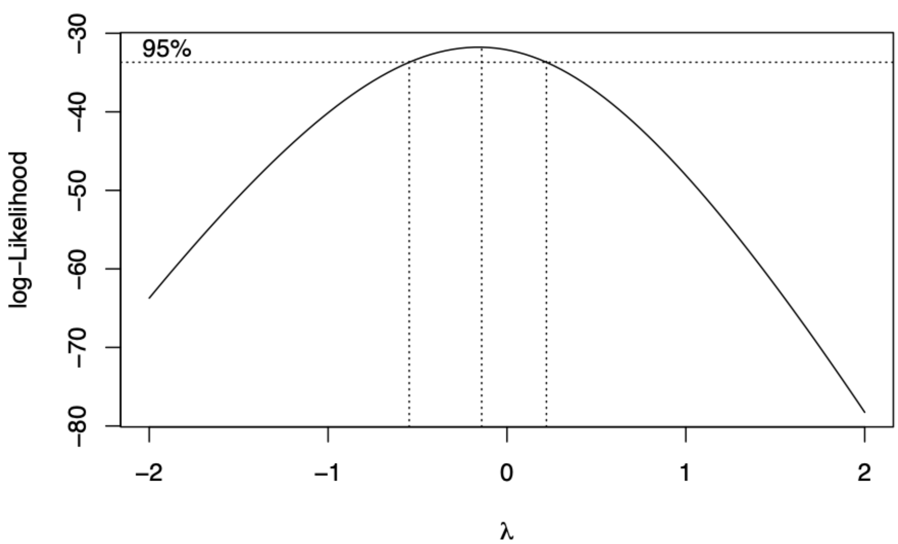
```r
summary(powerTransform(bridge[,4:5]))
```

```
## bcPower Transformations to Multinormality
##       Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## CCost   -0.1896           0      -0.4351       0.0560
## Dwgs    -0.1782           0      -0.6890       0.3327
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                              LRT df    pval
## LR test, lambda = (0 0) 2.375838  2 0.30486
##
## Likelihood ratio test that no transformations are needed
##                              LRT df       pval
## LR test, lambda = (1 1) 77.82844  2 < 2.22e-16
```

Based on output comes, estimated transformation parameters of CCost and Dwgs are -0.1896 and -0.1782,we consider log transformation which .

(d)

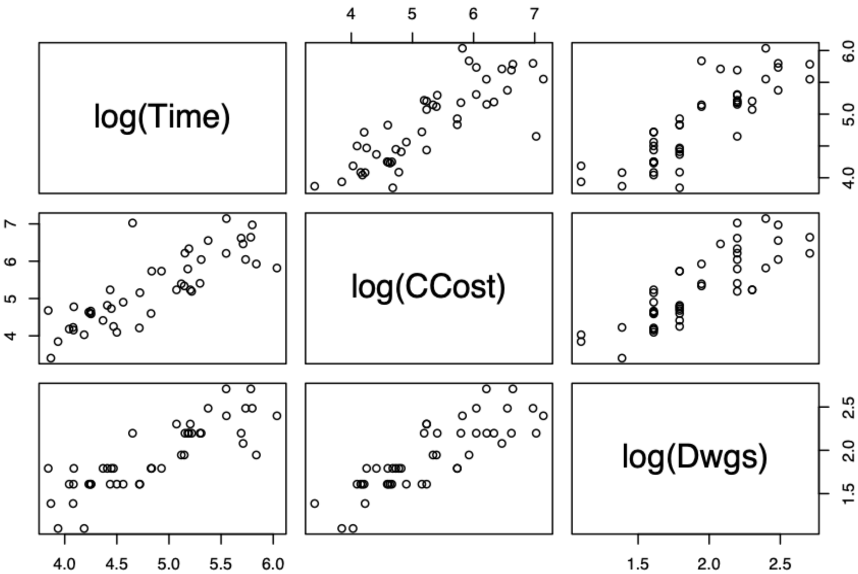
```r
library(MASS)
boxcox(Time~1+log(CCost)+log(Dwgs),data=bridge)
```



Based on the output, we could find that the value of transformation parameter of Time is close to 0 which maximizes log-likelihood. Thus, we consider log transformation for reponse variable.

(e)

```r
pairs(log(Time)~1+log(CCost)+log(Dwgs),data=bridge)
```



(f)

```r
mod1_2=lm(log(Time)~1+log(CCost)+log(Dwgs),data=bridge)
summary(mod1_2)
```

```
##
## Call:
## lm(formula = log(Time) ~ 1 + log(CCost) + log(Dwgs), data = bridge)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.87030 -0.16542 -0.02123  0.19536  0.80694
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.89890    0.26796   7.087 1.09e-08 ***
## log(CCost)   0.25931    0.08955   2.896 0.005981 **
## log(Dwgs)    0.81970    0.21869   3.748 0.000538 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3149 on 42 degrees of freedom
## Multiple R-squared:  0.7574, Adjusted R-squared:  0.7459
## F-statistic: 65.57 on 2 and 42 DF,  p-value: 1.206e-13
```

Interpret the slope coefficient of log(CCost):

Holding other constant variables,when log(CCost) increases one unit,log(Time) on average will increase 0.25931.

Yes.

(g)

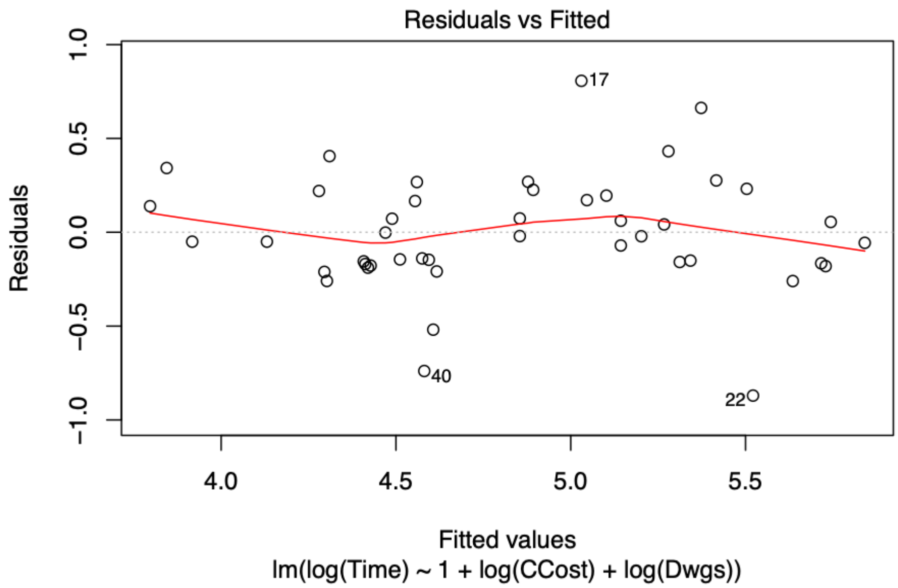
```
vif(mod1_2)
```

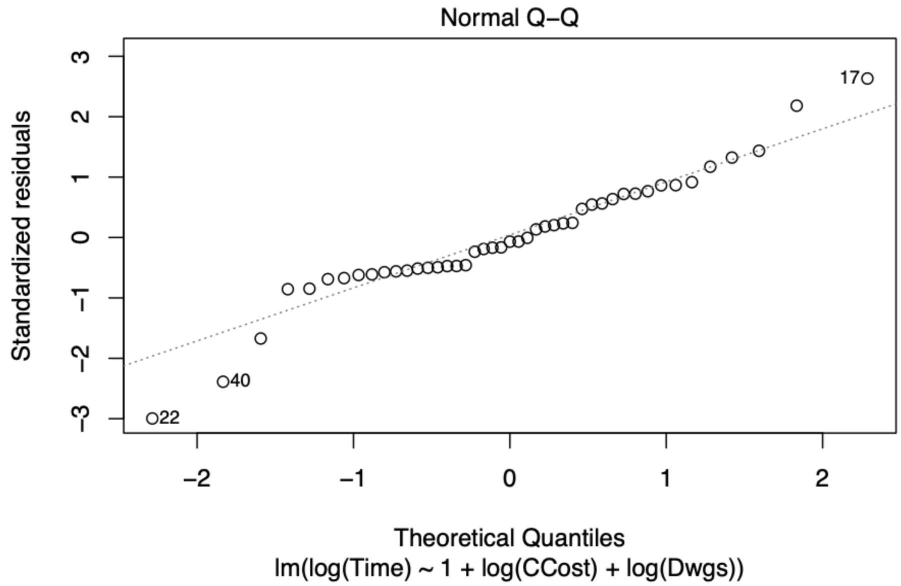
```
## log(CCost)  log(Dwgs)
##    3.23961    3.23961
```

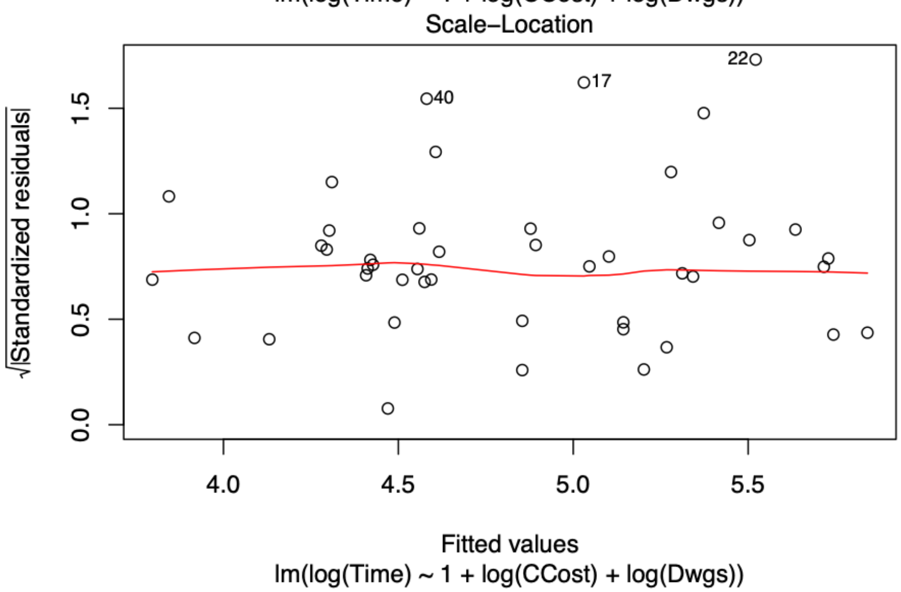
The VIF valu of log(CCost) and log(Dwgs) are 3.23961 and 3.23961 .Hence,using 5 as the threshold, the model does not have a multicollinearity issue.

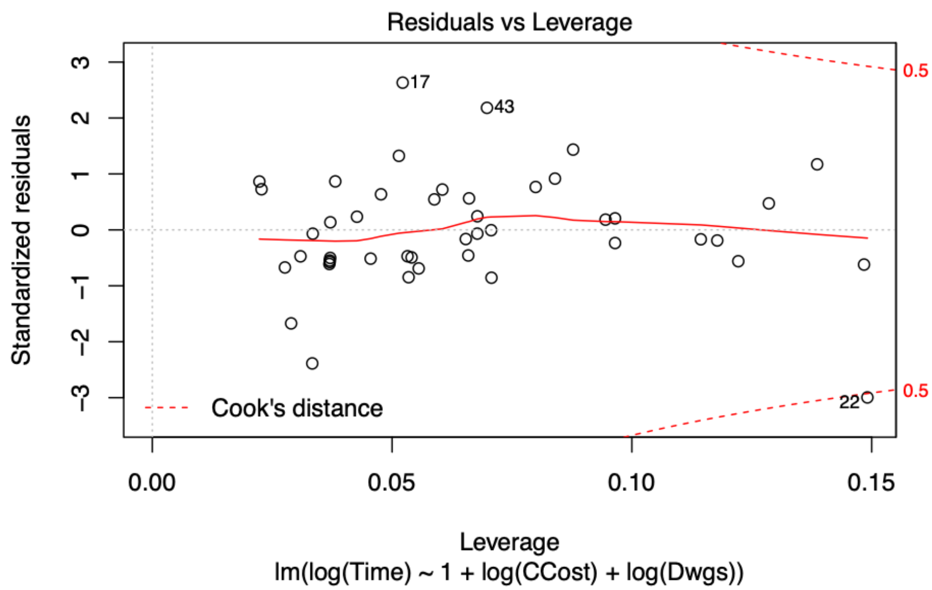
(h)

```
plot(mod1_2)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(log(Time) ~ 1 + log(CCost) + log(Dwgs))
17
40
22
−3
−2
−1
0
1
2
3


Scale−Location
√|Standardized residuals|
Fitted values
lm(log(Time) ~ 1 + log(CCost) + log(Dwgs))
22
17
40
0.0
0.5
1.0
1.5
4.0
4.5
5.0
5.5



Compare and contrast with the diagnostic plots in (b),I could not see any violation of the linear regression assumptions.

(i)

```r
new.dat=data.frame(CCost=300,Dwgs=6)
predict(mod1_2,newdata =new.dat)
```

```
##        1
## 4.846626
```

```r
predict(mod1_2,newdata =new.dat,interval ='predict',level=0.95)
```

```
##        fit      lwr     upr
## 1 4.846626 4.190921 5.50233
```

Hence, the estimated design time is 4.846626 and the 95% prediction interval for the estimated time is (4.190921, 5.50233).

Problem 2

Consider Y is a binary random variable, taking values 0 or 1 depending on a set of predictor variables, x1,x2,. . . ,xp.

The logit link function, where $\theta_x \in (0,1)$,$logit(\theta_x) = log(\theta_x/(1-\theta_x)) = \beta_0 + \beta_1 * x_1 + ... + \beta_p * x_p$ (a)

If$Y \sim Ber(\theta)$,$E(Y|\theta) = P(Y=1|\theta)*1 + P(Y=0|\theta)*0 = P(Y=1|\theta) = \theta$, $Var(Y|\theta) = E(Y^2|\theta) - (E(Y|\theta)^2) = E(Y^2|\theta) - \theta^2 = P(Y=1|\theta)*1^2 + P(Y=0|\theta)*0^2 - \theta^2 = \theta - \theta^2 = \theta(1-\theta)$.

(b) $\theta_x = e^{\beta_0+\beta_1*x_1+...+\beta_p*x_p}/(1+e^{\beta_0+\beta_1*x_1+...+\beta_p*x_p})$,

The denominator and numerator are divided by $e^{\beta_0+\beta_1*x_1+...+\beta_p*x_p}$, $\theta_x = 1/(1+e^{-\beta_0+\beta_1*x_1+...+\beta_p*x_p})$.

(c)

Let $Y_1, Y_2, ..., Y_{10}$ has $Ber(\theta = 0.7)$. If $W = \sum_{i=1}^{10} Y_i$,W has Binomial(10,0.7).

$P(W <= 3) = P(W = 0) + P(W = 1) + P(W = 2) + P(W = 3) = \binom{10}{0} * 0.7^0 * (1 - 0.7)^{10} + \binom{10}{1} * 0.7^1 * (1 - 0.7)^9 + \binom{10}{2} * 0.7^2 * (1 - 0.7)^8 + \binom{10}{3} * 0.7^3 * (1 - 0.7)^7$=0.010592078.

Problem3

```
#load dataset already stored in R
data(esoph)
#gives details about dataset
help(esoph)
#make sure the age group and alcohol consumption variables
# are read as basic categorical
esoph$agegp = factor(esoph$agegp, ordered=FALSE)
esoph$alcgp = factor(esoph$alcgp, ordered=FALSE)
esoph$tobgp = factor(esoph$tobgp, ordered=FALSE)
```

(a)

```
mod3_1=glm(cbind(ncases,ncontrols)~agegp+alcgp+tobgp,family="binomial",data=esoph)
summary(mod3_1)
```

```
##
## Call:
## glm(formula = cbind(ncases, ncontrols) ~ agegp + alcgp + tobgp,
##     family = "binomial", data = esoph)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6891  -0.5618  -0.2168   0.2314   2.0642
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -5.9108     1.0302  -5.737 9.61e-09 ***
## agegp35-44    1.6095     1.0676   1.508 0.131652
## agegp45-54    2.9752     1.0242   2.905 0.003675 **
## agegp55-64    3.3584     1.0198   3.293 0.000991 ***
## agegp65-74    3.7270     1.0253   3.635 0.000278 ***
## agegp75+      3.6818     1.0645   3.459 0.000543 ***
## alcgp40-79    1.1216     0.2384   4.704 2.55e-06 ***
## alcgp80-119   1.4471     0.2628   5.506 3.68e-08 ***
## alcgp120+     2.1154     0.2876   7.356 1.90e-13 ***
## tobgp10-19    0.3407     0.2054   1.659 0.097159 .
## tobgp20-29    0.3962     0.2456   1.613 0.106708
## tobgp30+      0.8677     0.2765   3.138 0.001701 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 227.241  on 87  degrees of freedom
## Residual deviance:  53.973  on 76  degrees of freedom
## AIC: 225.45
##
## Number of Fisher Scoring iterations: 6
```

(b)

```r
n=nrow(esoph)
backBIC = step(mod3_1, direction="backward", data=diamonds,k=log(n))
```

```
## Start:  AIC=255.18
## cbind(ncases, ncontrols) ~ agegp + alcgp + tobgp
##
##         Df Deviance    AIC
## - tobgp  3   64.572 252.35
## <none>       53.973 255.18
## - alcgp  3  120.028 307.80
## - agegp  5  131.484 310.31
##
## Step:  AIC=252.35
## cbind(ncases, ncontrols) ~ agegp + alcgp
##
##         Df Deviance    AIC
## <none>       64.572 252.35
## - agegp  5  138.789 304.18
## - alcgp  3  139.112 313.46
```

```r
summary(backBIC)
```

```
##
## Call:
## glm(formula = cbind(ncases, ncontrols) ~ agegp + alcgp, family = "binomial",
##     data = esoph)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8979  -0.5592  -0.1995   0.5029   2.6250
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -5.6180     1.0217  -5.499 3.82e-08 ***
## agegp35-44    1.5376     1.0646   1.444 0.148669
## agegp45-54    2.9470     1.0217   2.884 0.003922 **
## agegp55-64    3.3116     1.0172   3.255 0.001132 **
## agegp65-74    3.5774     1.0209   3.504 0.000458 ***
## agegp75+      3.5858     1.0620   3.377 0.000734 ***
## alcgp40-79    1.1392     0.2367   4.814 1.48e-06 ***
## alcgp80-119   1.4951     0.2600   5.749 8.97e-09 ***
## alcgp120+     2.2228     0.2843   7.820 5.29e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 227.241  on 87  degrees of freedom
## Residual deviance:  64.572  on 79  degrees of freedom
## AIC: 230.05
##
## Number of Fisher Scoring iterations: 6
```

The best fitted model choosen is cbind(ncases, ncontrols) ~ agegp + alcgp.

(c)

Probility: $\hat{\theta_x} = 1/(1+e^{-(-5.6180+1.5376*agegp35-44+2.9470*agegp45-54+3.3116*agegp55-64+3.5774*agegp65-74+3.5858*agegp75+1.1392*alc}$

log odds:$\hat{\eta} = log(\hat{\theta_x}/(1-\hat{\theta_x})) = -5.6180+1.5376*agegp35-44+2.9470*agegp45-54+3.3116*agegp55-64+3.5774*agegp65-74+3.5858*agegp75+1.1392*alcgp40-79+1.4951*alcgp80-119+2.2228*alcgp120+$

odds:$\hat{\theta_x}/(1-\hat{\theta_x}) = e^{-5.6180+1.5376*agegp35-44+2.9470*agegp45-54+3.3116*agegp55-64+3.5774*agegp65-74+3.5858*agegp75+1.1392*alcgp40-}$

(d)

Holding other variables constant, when the age is in the group 75+,on average,the predicted log odds that gets esophageal cancer will increase 3.5858 units than the age is in the group 25-34.

(e)

Holding other variables constant , when the alcohol consumption is 40-79g/day,on average ,the predicted odds that gets esophageal cancer will increase by a factor of e^1.1392 units than the alcohol consumption is 0-39g/day.

(f)

```
logodds=-5.6180+3.5774+1.1392
```

(g)

```
pro=1/(1-exp(-(-5.6180+3.5774+1.1392)))
```

Problem 4

```
TitanicPartial=read.csv(file = "TitanicPartial.csv")
```

(a) Based on the scatterplots,Male2 has the lowest survival odds

(b)

```
mod4_1=glm(Survived~1 + as.factor(Pclass) + Age,family="binomial",data=TitanicPartial)
mod4_2=glm(Survived~1 + as.factor(Pclass) + Age+Sex,family="binomial",data=TitanicPartial)
summary(mod4_1)
```

```
##
## Call:
## glm(formula = Survived ~ 1 + as.factor(Pclass) + Age, family = "binomial",
##     data = TitanicPartial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1524  -0.8466  -0.6083   1.0031   2.3929
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         2.296012   0.317629   7.229 4.88e-13 ***
## as.factor(Pclass)2 -1.137533   0.237578  -4.788 1.68e-06 ***
## as.factor(Pclass)3 -2.469561   0.240182 -10.282  < 2e-16 ***
## Age                -0.041755   0.006736  -6.198 5.70e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 964.52  on 713  degrees of freedom
```

```
## Residual deviance: 827.16  on 710  degrees of freedom
## AIC: 835.16
##
## Number of Fisher Scoring iterations: 4
```

```r
summary(mod4_2)
```

```
##
## Call:
## glm(formula = Survived ~ 1 + as.factor(Pclass) + Age + Sex, family = "binomial",
##     data = TitanicPartial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7303  -0.6780  -0.3953   0.6485   2.4657
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         3.777013   0.401123   9.416  < 2e-16 ***
## as.factor(Pclass)2 -1.309799   0.278066  -4.710 2.47e-06 ***
## as.factor(Pclass)3 -2.580625   0.281442  -9.169  < 2e-16 ***
## Age                -0.036985   0.007656  -4.831 1.36e-06 ***
## Sexmale            -2.522781   0.207391 -12.164  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 964.52  on 713  degrees of freedom
## Residual deviance: 647.28  on 709  degrees of freedom
## AIC: 657.28
##
## Number of Fisher Scoring iterations: 5
```

(c)

$\hat{\theta_x} = 1/(1 + e^{-(3.777013 - 1.309799 * as.factor(Pclass)2 - 2.580625 * as.factor(Pclass)3 - 0.036985 * Age - 2.522781 * Sexmale)})$

//Holdong all other contant variables, when

(d)

```r
anova(mod4_1,mod4_2,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Survived ~ 1 + as.factor(Pclass) + Age
## Model 2: Survived ~ 1 + as.factor(Pclass) + Age + Sex
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       710     827.16
## 2       709     647.28  1   179.88 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Because the p-value is so small, we could reject he null hypothesis.Hence, the Modle 2 is better.

(e) Model 1: Survived ~ 1 + as.factor(Pclass) + Age Model 2: Survived ~ 1 + as.factor(Pclass) + Age + Sex

$H_0 $:Model1 better $ vs. H_A $: Model2 better

$G^2_{H_0} = 827.16$,$G^2_{H_A} = 647.28$,

Test statistic =$G^2_{H_0} - G^2_{H_A}$= 179.88

```r
827.16-647.28
```

```
## [1] 179.88
```

$df_{H_0} - df_{H_A} = 1$;

P-value

```r
pchisq(179.88, 1,lower.tail=FALSE)
```

```
## [1] 5.147795e-41
```

Because the p-value is so small, we could reject he null hypothesis.Hence, the Modle 2 is better.