MATH 1312 - Regression Analysis

Assignment 3

Anirudhda Pardhi - s3807109

Importing required libraries and reading the data

```
library(car)
```

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

```
library(Hmisc)
```

```
## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

## Loading required package: ggplot2

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
library(plyr)
```

```
##
## Attaching package: 'plyr'

## The following objects are masked from 'package:Hmisc':
##
##     is.discrete, summarize
```

```
library(tidyr)
library(magrittr)
```

```
##
## Attaching package: 'magrittr'
```

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

## The following objects are masked from 'package:Hmisc':
##
##     src, summarize

## The following object is masked from 'package:car':
##
##     recode

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(QuantPsyc)
```

```
## Loading required package: boot

##
## Attaching package: 'boot'

## The following object is masked from 'package:survival':
##
##     aml

## The following object is masked from 'package:lattice':
##
##     melanoma

## The following object is masked from 'package:car':
##
##     logit
```

```
## Loading required package: MASS

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select

##
## Attaching package: 'QuantPsyc'

## The following object is masked from 'package:base':
##
##     norm
```

```
library(TSA)
```

```
##
## Attaching package: 'TSA'

## The following objects are masked from 'package:stats':
##
##     acf, arima

## The following object is masked from 'package:utils':
##
##     tar
```

Q1

(a)

```
data2 <- read.csv("/Users/ADMIN/Desktop/Sem 3/Regression analysis/Asg 3/asphalt.csv",header=TRUE)
lm.fit1<-lm(y ~ visc + surf + base + fines + voids + run , data= data2)
summary(lm.fit1)
```

```
##
## Call:
## lm(formula = y ~ visc + surf + base + fines + voids + run, data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6781 -1.8309  0.1751  1.4858 11.1262
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -62.970450  36.118989  -1.743   0.0941 .
```

```
## visc          0.003071   0.008161   0.376   0.7100
## surf          7.498028   3.967155   1.890   0.0709 .
## base          6.225817   4.812723   1.294   0.2081
## fines         0.522211   1.174673   0.445   0.6606
## voids        -0.241275   1.684963  -0.143   0.8873
## run          -5.386297   0.985384  -5.466 1.28e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.94 on 24 degrees of freedom
## Multiple R-squared:  0.7274, Adjusted R-squared:  0.6592
## F-statistic: 10.67 on 6 and 24 DF,  p-value: 8.588e-06
```

Equation of best fit line is :

y = -62.97+ 0.003* b1+ 7.498* b2+ 6.225* b3+ 0.522* b4 - 0.241* b5 - 5.38* b6

where b1,b2,b3,b4,b5,b6 are slopes for predictors variables visc,surf,base,fines,voids,run

p-value of equation line is 8.588e-06 which is very less than 0.05 and this suggest that regression is statistically significant at 5% level of significance.

Since the sample size is $>$ 30 we can assume normal distribution for 95% confidence. Degree of freedom = 24. As per the description given in summary we can observe that only p-value for run variable is less than 0.05 hence we can say that it is significant predictor variable and all the other variables have p-value greater than 0.05 hence we can say that they are insignificant predictors.

anova table

```
anova(lm.fit1)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## visc       1 424.93  424.93 27.3767 2.310e-05 ***
## surf       1  16.13   16.13  1.0393    0.3182
## base       1  28.40   28.40  1.8295    0.1888
## fines      1  19.73   19.73  1.2711    0.2707
## voids      1  41.05   41.05  2.6450    0.1169
## run        1 463.77  463.77 29.8792 1.283e-05 ***
## Residuals 24 372.51   15.52
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As per ANOVA table p-value of slope for variable visc = 2.310e-05 and run = 1.283e-05 which are less than 0.05 which suggest that they are statistically significant at 5% level of significance. However slope of p-value for surf = 0.3182, base= 0.1888, fines= 0.2707, voids= 0.1169 are greater than 0.05 which suggest that these variables are statistically insignificant.

(b)

Equation for y with run indicator 1 and -1 as factor

```
fit1<-lm(y ~ visc + surf + base + fines + voids + factor(run) , data= data2)
summary(fit1)
```

```
##
## Call:
## lm(formula = y ~ visc + surf + base + fines + voids + factor(run),
##     data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6781 -1.8309  0.1751  1.4858 11.1262
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -57.584153  35.896498  -1.604   0.1218
## visc           0.003071   0.008161   0.376   0.7100
## surf           7.498028   3.967155   1.890   0.0709 .
## base           6.225817   4.812723   1.294   0.2081
## fines          0.522211   1.174673   0.445   0.6606
## voids         -0.241275   1.684963  -0.143   0.8873
## factor(run)1 -10.772593   1.970769  -5.466 1.28e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.94 on 24 degrees of freedom
## Multiple R-squared:  0.7274, Adjusted R-squared:  0.6592
## F-statistic: 10.67 on 6 and 24 DF,  p-value: 8.588e-06
```

```
# for run=1, y = -68.35+ 0.003*visc+ 7.498*surf+ 6.225*base+ 0.522*fines - -0.241*voids
# for run=-1 y = -57.58+ 0.003*visc+ 7.498*surf+ 6.225*base+ 0.522*fines - -0.241*voids
```

bo= -57.58 (intercept) b1= slope for visc , x1 = visc variable b2= slope for surf , x2 = surf variable b3= slope for base , x3 = base variable b4= slope for fines , x4 = fines variable b5= slope for voids , x5 = voids variable b6= slope for run , x6 = run indicator variable

When value of run indicator is 1 then in that case the intercepts values get added up and this intercept value will be different than in case of run indicator is -1.

for run indicator = 1 the fitted equation is :

y= (bo+b6) + b1*x1 + b2*x2 + b3*x3 + b4*x4 + b5*x5 y = -68.35+ 0.003*x1 + 7.498*x2 + 6.225*x3 + 0.522*x4 - 0.241*x5

for run indicator = -1 the fitted equation is :

y= (bo) + b1*x1 + b2*x2 + b3*x3 + b4*x4 + b5*x5 y = -57.58 + 0.003*x1 + 7.498*x2 + 6.225*x3 + 0.522*x4 - 0.241*x5

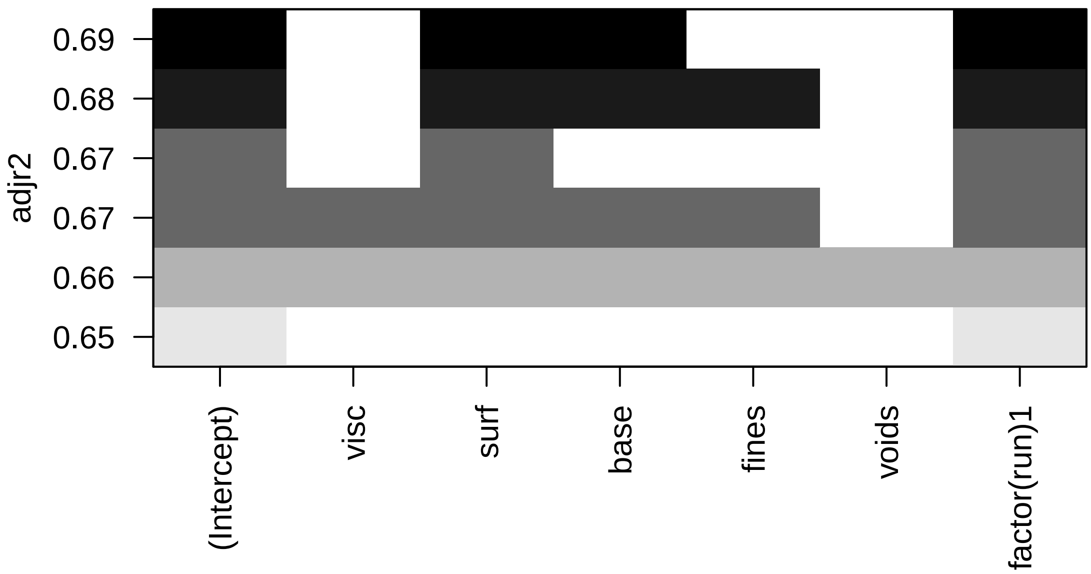
(c)

Selection of best model based on R square value

```
library(leaps)
r<-leaps::regsubsets(y ~ visc + surf + base + fines + voids + factor(run) , data= data2)
summary(r)
```

```
## Subset selection object
## Call: regsubsets.formula(y ~ visc + surf + base + fines + voids + factor(run),
##     data = data2)
## 6 Variables  (and intercept)
##              Forced in Forced out
## visc             FALSE      FALSE
## surf             FALSE      FALSE
## base             FALSE      FALSE
## fines            FALSE      FALSE
## voids            FALSE      FALSE
## factor(run)1     FALSE      FALSE
## 1 subsets of each size up to 6
## Selection Algorithm: exhaustive
##          visc surf base fines voids factor(run)1
## 1  ( 1 ) " "  " "  " "  " "   " "   "*"
## 2  ( 1 ) " "  "*"  " "  " "   " "   "*"
## 3  ( 1 ) " "  "*"  "*"  " "   " "   "*"
## 4  ( 1 ) " "  "*"  "*"  "*"   " "   "*"
## 5  ( 1 ) "*"  "*"  "*"  "*"   " "   "*"
## 6  ( 1 ) "*"  "*"  "*"  "*"   "*"   "*"
```

```
par(mfrow=c(1,1))
plot(r, scale="adjr2")
```



Based on the adjusted R square we can see that predictor variable surf, base and factor(run) are most significant one so the equation of fitted line will be obtained as per below calculation.

```
bestfit1<-lm(y ~ surf + base + factor(run) , data= data2)
summary(bestfit1)
```

```
##
## Call:
## lm(formula = y ~ surf + base + factor(run), data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9414 -1.7181 -0.1026  1.4959 11.0532
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -57.332     28.107  -2.040   0.0513 .
## surf            7.427      3.251   2.285   0.0304 *
## base            6.828      4.039   1.691   0.1024
## factor(run)1  -10.537      1.353  -7.786 2.26e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.738 on 27 degrees of freedom
## Multiple R-squared:  0.7239, Adjusted R-squared:  0.6932
## F-statistic:  23.6 on 3 and 27 DF,  p-value: 1.044e-07
```

Equation of line :

Bo = 57.332 B1 = 7.427 slope for surf variable , x1 = surf variable B2 = 6.828 slope for base variable , x2 = base variable b3 = -10.537 slope for run variable , x3 = run variable

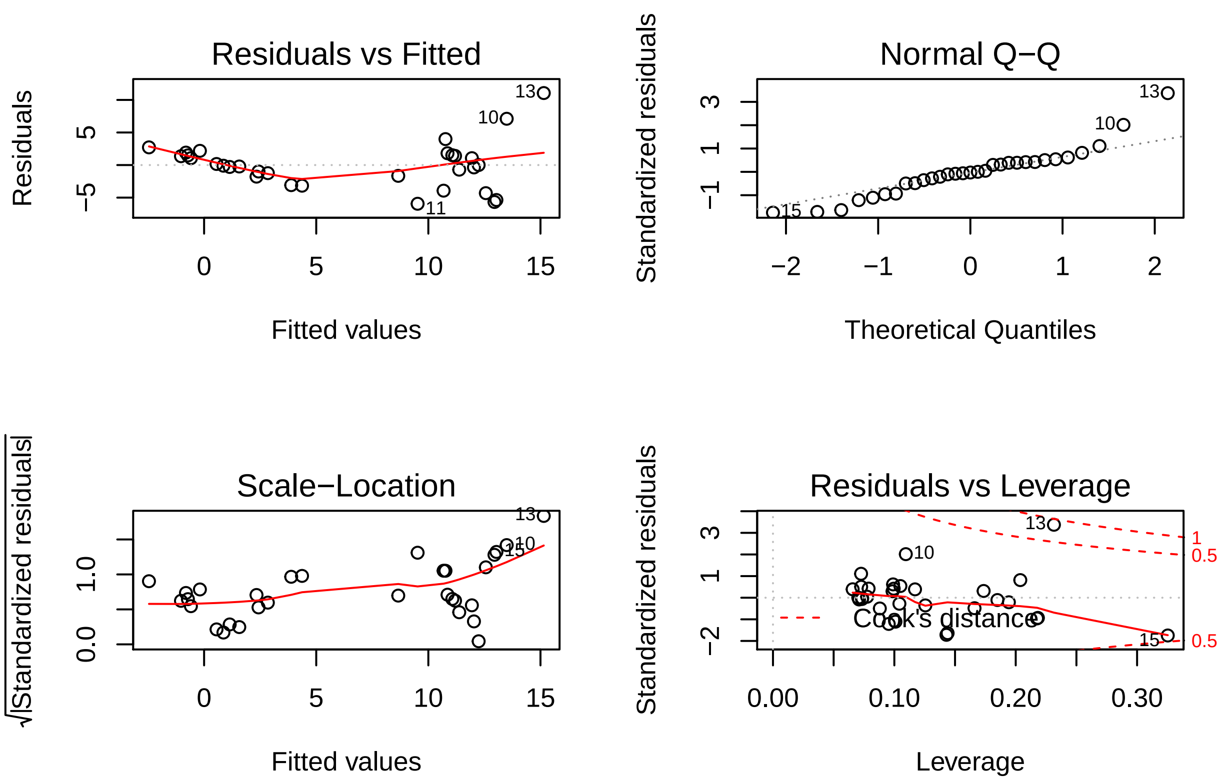
y = -57.332 + B1*x1 + B2*x2 + B3*x3

p-value of equation line is 1.044e-07 which is very less than 0.05 and this suggest that regression is statistically significant at 5% level of significance.Adjusted R-squared value is 69.32% which suggest goodness of fit is significant.

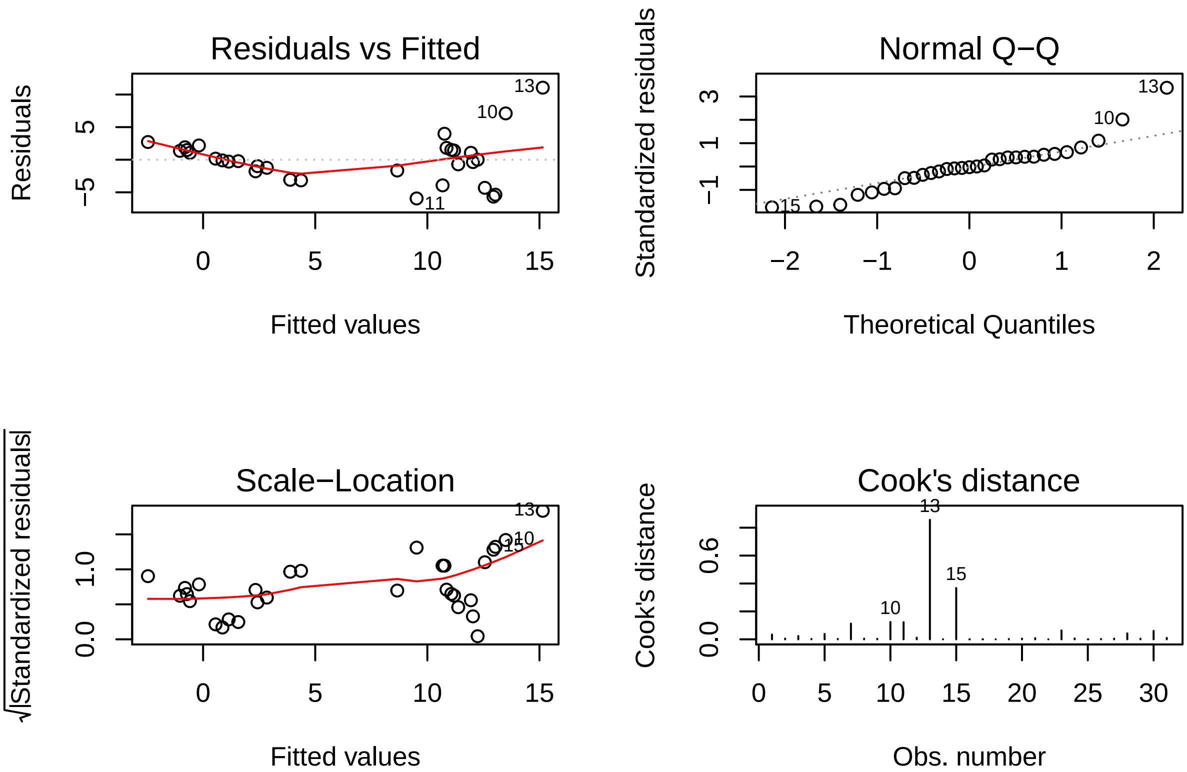
Since the sample size is > 30 we can assume normal distribution for 95% confidence. Degree of freedom = 27. As per the description given in summary we can observe that p-value for run and surf variable are less than 0.05 hence we can say that it is significant predictor variable and base variable have p-value greater than 0.05 hence we can say that this is insignificant predictor.

Residual analysis

```
par(mfrow=c(2,2))
plot(bestfit1)
```




```
plot(bestfit1, which = 1:4)
```




```
ncvTest(bestfit1)
```

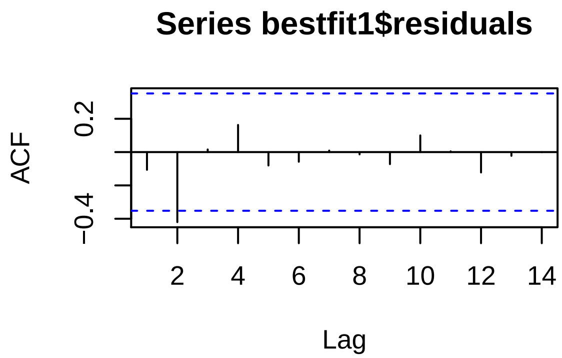
```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 13.25413, Df = 1, p = 0.00027198
```

```
shapiro.test(bestfit1$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  bestfit1$residuals
## W = 0.93002, p-value = 0.04391
```

```
acf(bestfit1$residuals)
durbinWatsonTest(bestfit1)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      -0.1063792      2.159149   0.798
##  Alternative hypothesis: rho != 0
```



NCV test- In residual vs fitted graph we can see that the red line is curved so there may be heteroscedasticity exists.So we do "NCV test" and p = 0.0027 which is less than significance level 0.05 we fail to reject the null hypothesis that the variance of the residuals is constant and can say that Heteroscedasticity is present.

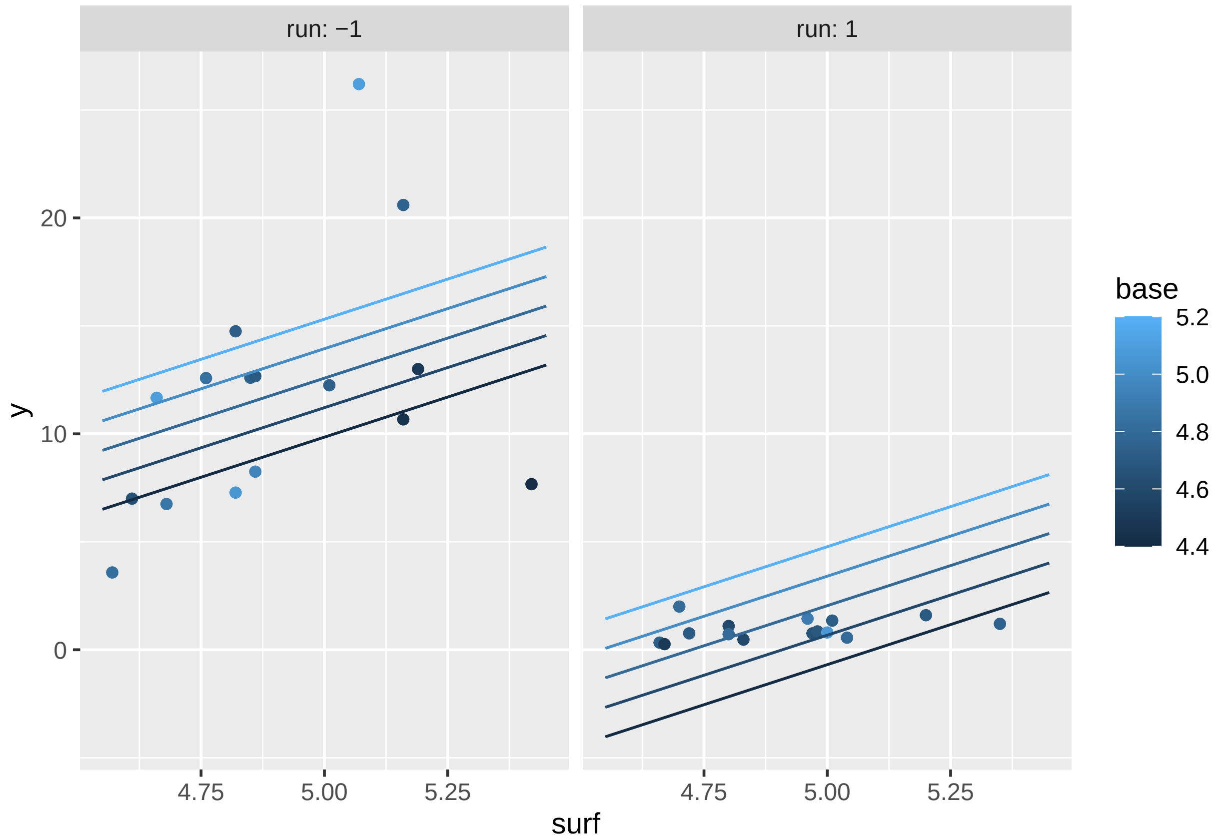
To test the Normality we can see the QQ plot and can say that there is not any gross deviations from normality.But since the number of observations are more than 30 according to central limit theorem we can assume normality.To confirm this we do "Shapiro test".Obtained p value = 0.043 which is less than significance level 0.05 which is implying that the distribution of the data are significantly different from normal distribution and allowing us to assume the normality as per this test.

ACF test- In ACF we check for early lags.Before lag = 5 we can observe thato only 1 correlation value is crossing significant confidence boundaries hence we can comprehend that stochastic component of data may be a white noise.As per durbinWatsonTest result we can see that p value is 0.75 which is greater than 0.05 and suggests we fail to reject null hypothesis i.e First-order autocorrelation does not exist.

Cook's distance shows the presence of influential points or possible outliers.In our case we have spotted one such point at 13th observation.

Model fitting as per run indicator

```
library(ggiraphExtra)
fit3<-lm(y ~  surf + base +  run , data= data2)
ggPredict(fit3)
```



When value of run indicator is 1 then in that case the intercepts values get added up and this intercept value will be different than in case of run indicator is -1.

Q2

(a)

```
data1 <- read.csv("/Users/ADMIN/Desktop/Sem 3/Regression analysis/Asg 3/byssinosis.csv",header=TRUE)
class(data1)
```

```
## [1] "data.frame"
```

```
model1<-glm(cbind(BysYes,BysNo)~Dust+Race+Sex+Smoke+Employ , family=binomial,data1)
summary.glm(model1)
```

```
##
## Call:
## glm(formula = cbind(BysYes, BysNo) ~ Dust + Race + Sex + Smoke +
##     Employ, family = binomial, data = data1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4126  -0.7573  -0.2421   0.3688   1.9804
```

```
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.4852     0.6060  -0.801  0.42331
## Dust         -1.3751     0.1155 -11.901  < 2e-16 ***
## Race          0.2463     0.2061   1.195  0.23203
## Sex          -0.2590     0.2116  -1.224  0.22095
## Smoke        -0.6292     0.1931  -3.259  0.00112 **
## Employ        0.3856     0.1069   3.607  0.00031 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 322.527  on 64  degrees of freedom
## Residual deviance:  69.509  on 59  degrees of freedom
## AIC: 188.19
##
## Number of Fisher Scoring iterations: 5
```

z value is analogous to t-statistics in multiple regression output. z value $>$ 2 implies the corresponding variable is significant.According to this Race, smoke and employee are significant variables.

p value determines the probability of significance of predictor variables. With 95% confidence level, a variable having $p < 0.05$ is considered an important predictor. The same can be inferred by observing stars against p value.According to this Race, smoke and employee are significant variables.

Significant predictors

```
confint.default(model1)
```

```
##                  2.5 %     97.5 %
## (Intercept) -1.6729540  0.7025201
## Dust        -1.6015997 -1.1486543
## Race        -0.1576071  0.6501886
## Sex         -0.6737381  0.1557304
## Smoke       -1.0075930 -0.2507521
## Employ       0.1760540  0.5951812
```

As per confidence interval those predictors are significant which does not include 0 so Dust,employee and Smoke are significant.

check for multicollinearity

```
vif(model1)
```

```
##     Dust     Race      Sex    Smoke   Employ
## 1.239990 1.546530 1.215603 1.047165 1.460025
```

We are checking multicollinearity between the predictors using the vif value. All the VIF values are below 5 which suggests multicollinearity problem does not exist.

(b)

model adequacy

```
deviance(model1)
```

```
## [1] 69.50926
```

```
pchisq(model1$deviance, df=model1$df.residual, lower.tail = FALSE)
```

```
## [1] 0.1645594
```

The chi-square test statistic of 69.50926 with 59 degree of freedom gives a p-value of 0.1645647, indicating that the null hypothesis is plausible, and we can conclude that logistic model is adequate.

(c)

This question has been performed manually on excel and this file has been submitted along with pdf. As per calculation done in excel using formula below are the probabilities for a person suffering from byssinosis :

Answer (i) - P(x) = 0.042967226

Answer (ii) - P(x) = 0.204549276

Answer (iii) - P(x) = 0.023309701

Answer (iv) - P(x) = 0.003156909