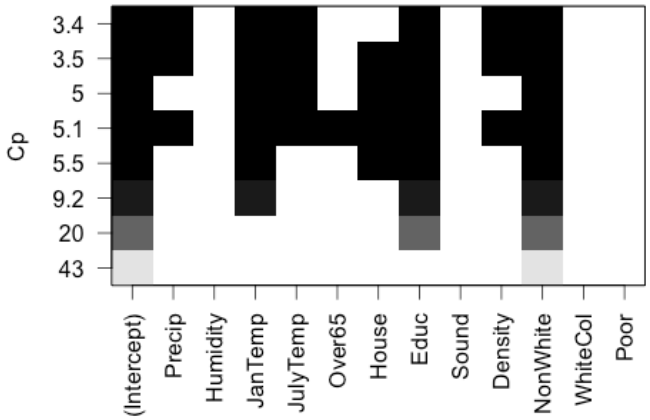
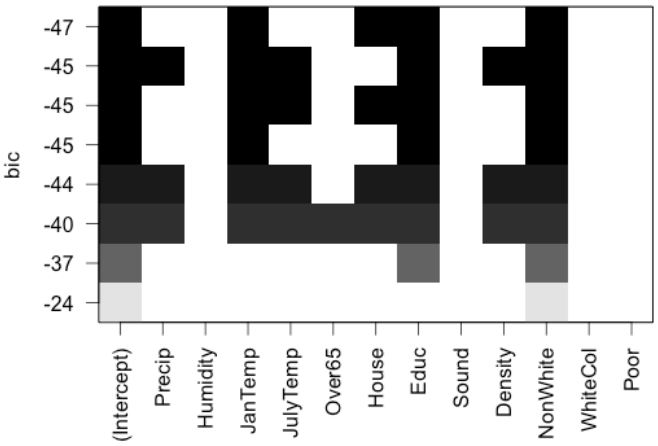
Name: Parth Pareek
UNI: PP2547
Date: 4/6/2016
Assignment: HW8

1.

a.



Smallest Cp = 3.443
Model variables: Precip, JanTemp, JulyTemp, Educ, Density, NonWhite



Smallest BIC = 453
Model variables: JanTemp, House, Educ, NonWhite

After adding pollution variables F = ((66518-52712)/3)/(52712/50) = 4.365 (p-value = .0083)

b.
Stepwise regression leads to same model

F= ((74,651-63018)3)/(63018/52) = 3.2 (p-value = .03)

There is a difference in p-values however, the result doesn't change.

2. The data set is separated into Native and Non-Native. After taking log transformation, we use best subset methodology. It is noticed that the best subset has only 1 variable, *Area*.

```
Call:
lm(formula = Native ~ ., data = dat.native)

Residuals:
     Min       1Q   Median       3Q      Max
-2.11603 -0.29983  0.09883  0.36696  1.00491

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.06730    1.49763   2.048  0.05164 .
Area         0.34170    0.09330   3.662  0.00123 **
Elev        -0.13572    0.29019  -0.468  0.64423
DistNear    -0.05866    0.10371  -0.566  0.57688
DistSc      -0.01637    0.11973  -0.137  0.89237
AreaNear    -0.03634    0.04141  -0.878  0.38881
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7022 on 24 degrees of freedom
Multiple R-squared:  0.7127,    Adjusted R-squared:  0.6529
F-statistic: 11.91 on 5 and 24 DF,  p-value: 7.223e-06
```

A model with all parameter also suggests that only Area is of significance (all other parameters have p-value > 0.005)

```
Subset selection object
Call: regsubsets.formula(Native ~ ., data = dat.native, method = "forward")
5 Variables  (and intercept)
         Forced in Forced out
Area          FALSE      FALSE
Elev          FALSE      FALSE
DistNear      FALSE      FALSE
DistSc        FALSE      FALSE
AreaNear      FALSE      FALSE
1 subsets of each size up to 5
Selection Algorithm: forward
         Area Elev DistNear DistSc AreaNear
1  ( 1 ) "*"  " "  " "      " "    " "
2  ( 1 ) "*"  " "  " "      " "    "*"
3  ( 1 ) "*"  " "  "*"      " "    "*"
4  ( 1 ) "*"  "*"  "*"      " "    "*"
5  ( 1 ) "*"  "*"  "*"      "*"    "*"
```

3.

a.

```
Call:
glm(formula = Failure ~ ., family = "binomial", data = dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2125  -0.8253  -0.4706   0.5907   2.0512

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 10.87535    5.70291   1.907   0.0565 .
Temperature -0.17132    0.08344  -2.053   0.0400 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 28.975  on 23  degrees of freedom
Residual deviance: 23.030  on 22  degrees of freedom
AIC: 27.03

Number of Fisher Scoring iterations: 4
```

b. One sided p-value = 0.04/2 = 0.02

c. Drop in deviance = 28.975 – 23.030 = 5.94

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: Failure

Terms added sequentially (first to last)


            Df Deviance Resid. Df Resid. Dev
NULL                           23     28.975
Temperature  1   5.9441        22     23.030
```

d. CI = -0.3348 to -0.0078

e. Logit = 5.5644; Probability = 0.9962

f. Can’t say for sure since data available is not in range of explanatory variable is not in range of the dataset used for regression.

4.
k = 1, p-value = 0.00048
k = 2, p-value = 0.01163
k = 3, p-value = 0.00003
k = 4, p-value = 0.00462
k = 5, p-value = 0.00014
k = 6, p-value = 0.00026
k = 7, p-value = 0.24819

b.

```
Call:
glm(formula = Site ~ ., family = "binomial", data = dat)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.39021  -0.74745  -0.01854   0.72308   1.90308

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  9.80304    3.38934   2.892  0.00382 **
PctRing1    -0.05708    0.03713  -1.537  0.12422
PctRing2     0.11730    0.04990   2.351  0.01873 *
PctRing3    -0.12181    0.05199  -2.343  0.01913 *
PctRing4     0.01694    0.04277   0.396  0.69201
PctRing5    -0.03296    0.03905  -0.844  0.39875
PctRing6    -0.10891    0.06631  -1.642  0.10051
PctRing7     0.05157    0.03619   1.425  0.15415
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 83.178  on 59  degrees of freedom
Residual deviance: 52.107  on 52  degrees of freedom
AIC: 68.107

Number of Fisher Scoring iterations: 6
```

Simple logistic regression suggests that Ring 2 and 3 are critical covariates for regression. However, 1 cannot be ignored.

Taking hint into account, ring percentages is converted to circle percentages by cumulative addition of rings.

Adding a ring at a time, we can conclude that ring 1,2,3 and 5 (between 1.6 and 1.78 km) are the most relevant.

CODES

```
#Question 1
library(Sleuth3)
library(leaps)
attach(ex1217)
dat <- ex1217
dat1 <- dat[,-(15:17)]

subset1 <- regsubsets(Mortality~.,dat1[,-1])
plot(subset1,scale = "Cp")
plot(subset1,scale = "bic")
min(leaps(x=dat1[,-(1:2)],y=dat1[,2])$Cp)

model11 <- lm(Mortality~Precip + JanTemp + JulyTemp + Educ + Density +
                NonWhite, data = dat)
model12 <- lm(Mortality~Precip + JanTemp + JulyTemp + Educ + Density +
                NonWhite + HC + NOX + SO2, data = dat)

anova(model11,model12)

subset2 <- regsubsets(Mortality~.,dat1[,-1])
bestmodel <- which.max(summary(subset2)$adjr2)
plot(subset2,scale = "Cp")
plot(subset2,scale = "bic")
summary(subset2)


model21 <- lm(Mortality~JanTemp + House + Educ + NonWhite, data = dat)
model22 <- lm(Mortality~JanTemp + House + Educ + NonWhite
              + HC + NOX + SO2, data = dat)

anova(model21,model22)


#Question 2
library(Sleuth3)
attach(ex1220)
dat <- ex1220
dat$DistSc <- dat$DistSc + 2
dat.native <- dat[,-(1:2)]
dat.native <- log(dat.native)
dat.nonnat <- cbind("NonNative" = (dat[,2] - dat[,3]),dat[,-(1:3)])
dat.nonnat <- log(dat.nonnat)

subset.native <- regsubsets(Native~., data = dat.native, method =
"forward")
best.sub.native <- which.max(summary(subset.native)$adjr2)

model1 <- lm(Native~.,data = dat.native)
summary(model1)

model2 <- lm(NonNative~.,data = dat.nonnat)
summary(model2)
```

```
#Question 3
library(MASS)
library(Sleuth3)
attach(ex2011)
dat <- ex2011

model1 <- glm(Failure~., data = dat, family = "binomial")
summary(model1) #z-value is -2.054; 2 sided p-value is 0.4
                #therefore, one-sided would approximately be 0.2

model2 <- glm(Failure~.-1, data = dat, family = "binomial")
anova(model2,model1)

predict(model1, list(Temperature = 31))
(exp(5.564414)-1)/exp(5.564414)


#Question 4
library(MASS)
library(leaps)
library(Sleuth3)
attach(ex2015)
dat <- ex2015

rr1 <- dat$PctRing1[which(dat$Site == "Random")]
rr2 <- dat$PctRing2[which(dat$Site == "Random")]
rr3 <- dat$PctRing3[which(dat$Site == "Random")]
rr4 <- dat$PctRing4[which(dat$Site == "Random")]
rr5 <- dat$PctRing5[which(dat$Site == "Random")]
rr6 <- dat$PctRing6[which(dat$Site == "Random")]
rr7 <- dat$PctRing7[which(dat$Site == "Random")]
rn1 <- dat$PctRing1[which(dat$Site == "Nest")]
rn2 <- dat$PctRing2[which(dat$Site == "Nest")]
rn3 <- dat$PctRing3[which(dat$Site == "Nest")]
rn4 <- dat$PctRing4[which(dat$Site == "Nest")]
rn5 <- dat$PctRing5[which(dat$Site == "Nest")]
rn6 <- dat$PctRing6[which(dat$Site == "Nest")]
rn7 <- dat$PctRing7[which(dat$Site == "Nest")]

t.test(rr1,rn1, alternative = "less")$p.value
t.test(rr2,rn2, alternative = "less")$p.value
t.test(rr3,rn3, alternative = "less")$p.value
t.test(rr4,rn4, alternative = "less")$p.value
t.test(rr5,rn5, alternative = "less")$p.value
t.test(rr6,rn6, alternative = "less")$p.value
t.test(rr7,rn7, alternative = "less")$p.value

model <- glm(Site~., data = dat, family = "binomial")
model1 <- glm(Site~PctRing1, data = dat, family = "binomial")

R1 <- PctRing1
R2 <- R1+PctRing2
```

```
R3 <- R2+PctRing3
R4 <- R3+PctRing4
R5 <- R4+PctRing5
R6 <- R5+PctRing6
R7 <- R6+PctRing7

sub <- regsubsets(dat$Site~R1+R2+R3+R4+R5+R6+R7, data = dat)
summary(sub)


#Question 6
library(MASS)
library(leaps)
X <- c(-2,-1,1,2)
Y <- c(0,0,1,1)
dat <- data.frame(X,Y)
model <- glm(Y~X-1,data = dat, family = binomial)
summary(model)
```