Hw06ST430Yu

Haozhe (Jerry) Yu

2023-11-11

Question 1

A researcher studied the effects of the charge rate and temperature on the life of a new type of power cell in a preliminary small-scale experiment. The charge rate (Xl) was controlled at three levels (0.6, 1.0, and 1.4 amperes) and the ambient temperature (X2) was controlled at three levels (l0, 20, 30°C). Factors pertaining to the discharge of the power cell were held at fixed levels. The life of the power cell (Y) was measured in terms of the number of discharge-charge cycles that a power cell underwent before it failed.

The researcher was not sure about the nature of the response function in the range of the factors studied. Hence, the researcher decided to fit the second-order polynomial regression model

```
data <- read.table("Datasets/battery.txt", header=FALSE)
names(data) <- c("cycles","rate","temp")
attach(data)
```

##a. Find the correlation matrix and report any high correlation between predictor variables.

```
cor(data)
```

```
##            cycles       rate      temp
## cycles  1.0000000 -0.5555349 0.7512159
## rate   -0.5555349  1.0000000 0.0000000
## temp    0.7512159  0.0000000 1.0000000
```

The correlation between cycles and temp is 0.7512159. This is high and could be a sign of multicollinearity.

##b. Fit a full model (Shown above) and report the overall F value and individual t-values. Do you suspect any multicollinearity problem?

```
mod1<-lm(cycles~rate+temp+I(rate^2)+I(temp^2)+ I(rate*temp))
summary(mod1)
```

```
##
## Call:
## lm(formula = cycles ~ rate + temp + I(rate^2) + I(temp^2) + I(rate *
##     temp))
##
## Residuals:
##       1       2       3       4       5       6       7       8       9      10
## -21.465   9.263  12.202  41.930  -5.842 -31.842  21.158 -25.404 -20.465   7.263
```

```
##      11
##  13.202
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     337.7215   149.9616   2.252   0.0741 .
## rate           -539.5175   268.8603  -2.007   0.1011
## temp              8.9171     9.1825   0.971   0.3761
## I(rate^2)       171.2171   127.1255   1.347   0.2359
## I(temp^2)        -0.1061     0.2034  -0.521   0.6244
## I(rate * temp)    2.8750     4.0468   0.710   0.5092
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.37 on 5 degrees of freedom
## Multiple R-squared:  0.9135, Adjusted R-squared:  0.8271
## F-statistic: 10.57 on 5 and 5 DF,  p-value: 0.01086
```

Yes I do. The overall p value for the ANOVA is < 0.05, but each of the individual regression coefficient's p values are more than 0.05. This is a sign of multicollinearity. Additionaly, this is a polynomial regression that has not been centered so by definition it will have structural multicollinearity.

c. We can remove the high correlation between explanatory variables and their powers by centering.

```r
rate.code <- (rate-mean(rate))/0.4
temp.code <- (temp-mean(temp))/10
cor(cbind(rate.code,temp.code,rate.code^2,temp.code^2))
```

```
##               rate.code temp.code
## rate.code  1.000000e+00         0 -4.042173e-16 -1.994753e-17
## temp.code  0.000000e+00         1  0.000000e+00  0.000000e+00
##           -4.042173e-16         0  1.000000e+00  2.666667e-01
##           -1.994753e-17         0  2.666667e-01  1.000000e+00
```

In this new correlation matrix I do not observe any high correlations and therefore signs of multicollinearity.

d.Fit a new full model with the scaled new predictor variables and report the estimated regression function

```r
mod2<-lm(cycles~rate.code+temp.code+I(rate.code^2)+I(temp.code^2)+I(rate.code*temp.code))
summary(mod2)
```

```
##
## Call:
## lm(formula = cycles ~ rate.code + temp.code + I(rate.code^2) +
##     I(temp.code^2) + I(rate.code * temp.code))
##
```

```
## Residuals:
##       1       2       3       4       5       6       7       8       9      10 
## -21.465   9.263  12.202  41.930  -5.842 -31.842  21.158 -25.404 -20.465   7.263 
##      11 
##  13.202 
## 
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)    
## (Intercept)               162.84      16.61   9.805 0.000188 ***
## rate.code                 -55.83      13.22  -4.224 0.008292 ** 
## temp.code                  75.50      13.22   5.712 0.002297 ** 
## I(rate.code^2)             27.39      20.34   1.347 0.235856    
## I(temp.code^2)            -10.61      20.34  -0.521 0.624352    
## I(rate.code * temp.code)   11.50      16.19   0.710 0.509184    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 32.37 on 5 degrees of freedom
## Multiple R-squared:  0.9135, Adjusted R-squared:  0.8271 
## F-statistic: 10.57 on 5 and 5 DF,  p-value: 0.01086
```

```r
summary(mod2)$coeff[1,1]
```

```
## [1] 162.8421
```

Cycles = 162.8421053 + -55.8333333rate.code + 75.5temp.code + 27.3947368rate.code^2 + -10.6052632temp.code^2 + 11.5[rate.code * temp.code]

(Goodness of fit) To test whether the second order polynomial regression function is good fit or not? Report the p-value and conclusion.

```r
mod.full <- lm(cycles~0+factor(rate.code)+factor(temp.code)+
                 factor(rate.code)*factor(temp.code))
anova(mod2, mod.full)
```

```
## Analysis of Variance Table
## 
## Model 1: cycles ~ rate.code + temp.code + I(rate.code^2) + I(temp.code^2) + 
##     I(rate.code * temp.code)
## Model 2: cycles ~ 0 + factor(rate.code) + factor(temp.code) + factor(rate.code) * 
##     factor(temp.code)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1      5 5240.4                           
## 2      2 1404.7  3    3835.8 1.8205 0.3738
```

p: 0.3738

Conclusion: We fail to reject the null hypothesis and conclude that there is no lack of fit for this model.

(Test higher order terms) The researcher wants to know whether a first-order model would be sufficient or not? Write the null and alternate hypothesis, p-value and conclusion.

```
mod.linear <- lm(cycles~rate.code+temp.code)
anova(mod.linear,mod2)
```

```
## Analysis of Variance Table
##
## Model 1: cycles ~ rate.code + temp.code
## Model 2: cycles ~ rate.code + temp.code + I(rate.code^2) + I(temp.code^2) +
##     I(rate.code * temp.code)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1      8 7700.3
## 2      5 5240.4  3    2459.9 0.7823 0.5527
```

H0: $\beta_3$rate.code^2 = $\beta_4$temp.code^2 = $\beta_5$rate.code * temp.code = 0, aka cycles = β_{01} + β_{11}rate.code + β_{21}temp.code = β_{02} + β_{12}rate.code + β_{22}temp.code + β_3(rate.code^2) + β_4($temp.code^2$) + β_5I(rate.code * temp.code)

HA: at least 1 of β_3,β_4, or $\beta_5 \neq 0$, so tghe 2 regression equations are not the same

p: 0.5527

Conclusion: At $\alpha = 0.05$, p $> \alpha$ so we fail to reject the null hypothesis and conclude that there is no evidence to conclude that the linear model and the polynomial model are any different, and thus as the linear model is more straightforward, it should be used instead.

Converting back to the original scale.

```
cf <- coefficients(mod.linear)
cf.rate <- cf["rate.code"]/0.4
cf.rate
```

```
## rate.code
## -139.5833
```

```
cf.temp <- cf["temp.code"]/10
cf.temp
```

```
## temp.code
##      7.55
```

```
const <- cf[1] - cf[2]/0.4 - cf[3]*20/10
const
```

```
## (Intercept)
##    160.5833
```

90% Bonferroni's Confidence interval for the estimate of the linear effects of the two predictor variables of the first order model

```
ci <- confint(mod.linear, level = 0.95)
ci["rate.code",] / 0.4
```

```
##      2.5 %     97.5 %
## -212.60206  -66.56461
```

```
ci["temp.code",] / 10
```

```
##     2.5 %    97.5 %
##  4.629251 10.470749
```

Question 2

A study obtained mortgage yields in n = 18 U.S. metropolitan areas in the 1960s. The researcher obtained the following variables and fit a linear regression model to see which factors (variables) were associated with yield (each variable was obtained for each metro area):

• Y = Mortgage Yield (Interest Rate as a %) • X1 = Average Loan/Mortgage Ratio (High Values • Low Down Payments/Higher Risk) • X2 = Distance from Boston (in miles) – (Most of population was in Northeast in the 1960s) • X3 = Savings per unit built (Measure of Available capital versus building rate) • X4 = Savings per capita • X5 = Population increase from 1950 to 1960 (%) • X6 = Percent of first mortgage from inter-regional banks (Measures flow of money from outside SMSA)

```
city <- as_tibble(read.table("Datasets/city.txt", header=TRUE))
```

Fit the Full Model

```
citym_f <- lm(Y~X1+X2+X3+X4+X5+X6,data=city)
```

###i. i. Test whether any of the independent variables are associated with mortgage yield. What proportion of variation in Y is "explained" by the independent variables?

As F = 12.3349032, and p is 2.5233194×10^{-4} and $p < \alpha$ when $\alpha = 0.05$, we reject the null hypothesis and conclude that at least one of the independent variables is associated with mortgage yield. The r^2 is 0.8706026 so 87.0602574 percent of the variation in mortgage yield is associated with the independent variables, but as we have many predictors it would be better to use the Adjusted R^2 which is 0.8000222 which would indicate that 80.002216 percent of the variation in mortgage yield is associated with the independent variables.

###ii.) Obtain the parameter estimates and t-tests for the individual partial regression coefficient and test individually for each variable (controlling for all others).

```
sumt <- summary(citym_f)$coeff %>% as_tibble()

sumtfinal <- select(sumt,-"Std. Error") %>% add_column(
  ifelse(
```

```
    sumt$`Pr(>|t|)` < 0.05,
    "Reject H0, controlling others this var is related to Y",
    "Fail to Reject H0, controlling others this var is not related to Y"
  ),
  "Parameter" = c("Intercept", "X1", "X2", "X3", "X4", "X5", "X6")
) %>% select("Parameter", everything())
sumtfinal
```

```
## # A tibble: 7 x 5
##   Parameter   Estimate `t value` `Pr(>|t|)` `ifelse(...)`
##   <chr>          <dbl>     <dbl>      <dbl> <chr>
## 1 Intercept  4.29          6.41   0.0000499 Reject H0, controlling others this ~
## 2 X1         0.0203        2.18   0.0515    Fail to Reject H0, controlling othe~
## 3 X2         0.0000136     0.290  0.778     Fail to Reject H0, controlling othe~
## 4 X3        -0.00158      -2.10   0.0593    Fail to Reject H0, controlling othe~
## 5 X4         0.000202      1.79   0.100     Fail to Reject H0, controlling othe~
## 6 X5         0.00128       0.727  0.483     Fail to Reject H0, controlling othe~
## 7 X6         0.000236      0.102  0.920     Fail to Reject H0, controlling othe~
```

iii. Obtain the partial sum of squares for each independent variable, and conduct the F-tests for individually for each variable (controlling for all others). Show that this is equivalent to the t-tests in the previous part.

```
l1 <- as_tibble(t(ftest(citym_f,matrix(c(1,0,0,0,0,0,0),nrow=1))))
l2 <- as_tibble(t(ftest(citym_f,matrix(c(0,0,1,0,0,0,0),nrow=1))))
l3 <- as_tibble(t(ftest(citym_f,matrix(c(0,0,0,1,0,0,0),nrow=1))))
l4 <- as_tibble(t(ftest(citym_f,matrix(c(0,0,0,0,1,0,0),nrow=1))))
l5 <- as_tibble(t(ftest(citym_f,matrix(c(0,0,0,0,0,1,0),nrow=1))))
l6 <- as_tibble(t(ftest(citym_f,matrix(c(0,0,0,0,0,0,1),nrow=1))))
smuf <- tibble("Parameter"=c("X1","X2","X3","X4","X5","X6")) %>%  bind_cols(bind_rows(l1,l2,l3,l4,l5,l6
smuf
```

```
## # A tibble: 6 x 5
##   Parameter       F   df1   df2 `p-value`
##   <chr>       <dbl> <dbl> <dbl>     <dbl>
## 1 X1        41.1        1    11 0.0000499
## 2 X2         0.0839     1    11 0.778
## 3 X3         4.42       1    11 0.0593
## 4 X4         3.22       1    11 0.100
## 5 X5         0.528      1    11 0.483
## 6 X6         0.0105     1    11 0.920
```

The p values for both rows match.

b) Test whether X2 (Distance from Boston), X5 (Population increase from 1950 to 1960), and X6 (Percent of first mortgage from inter-regional banks) are associated with mortgage yield, after controlling for X1, X3, and X4.

```r
tm3 <- matrix(c(0,0,1,0,0,0,0,
         0,0,0,0,0,1,0,
         0,0,0,0,0,0,1),nrow=3,ncol=7,byrow=TRUE)
ftest(citym_f,tm3)
```

```
##          F        df1        df2    p-value
##  0.2044452  3.0000000 11.0000000  0.8911770
```

As $p > 0.05$, we fail to reject the null hypothesis and conclude that controlling for X1,X3,and X4, X2,X5, and X6 are not significantly associated w Y.

c) Fit a first order model with all predictor variables. Use the regression subsets function in leaps package for the variable selection methods to determine the "best: model based on

Adjusted R2 Mallows Cp BIC criteria

```r
cityrefgull <- regsubsets(Y~X1+X2+X3+X4+X5+X6,data=city)
rsum <- summary(cityrefgull)
names(rsum)
```

```
## [1] "which"  "rsq"    "rss"    "adjr2"  "cp"     "bic"    "outmat" "obj"
```

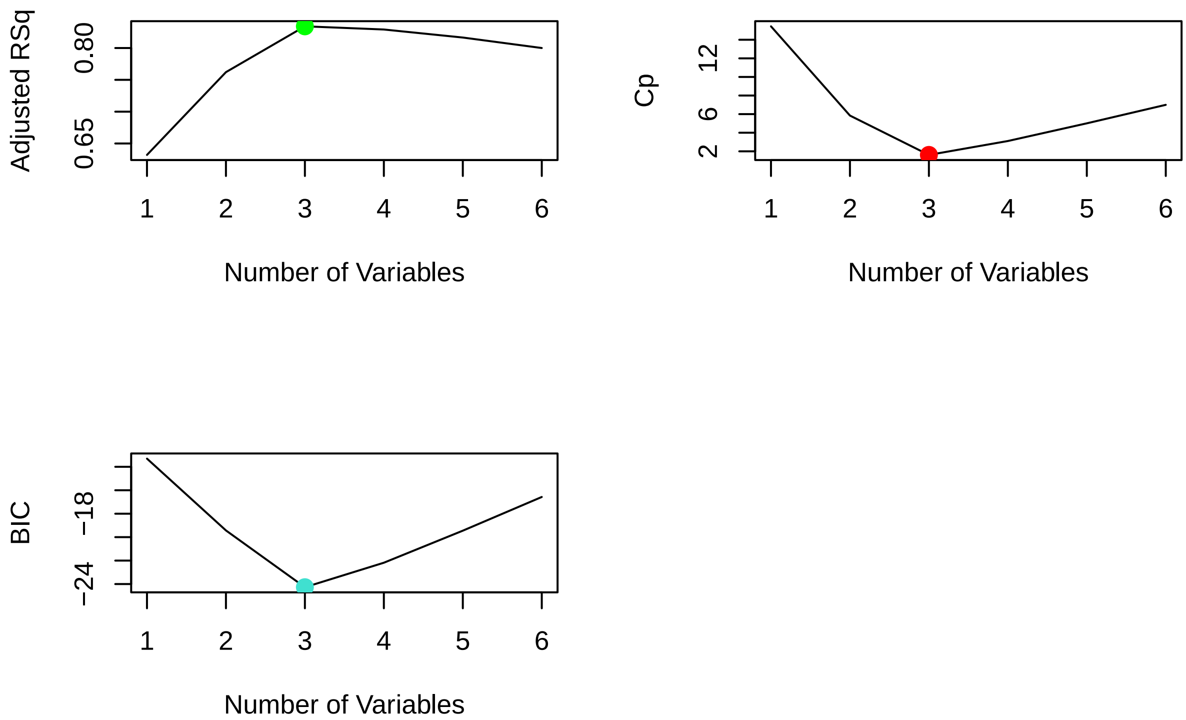
```r
rsum$which
```

```
##   (Intercept)   X1    X2    X3    X4    X5    X6
## 1        TRUE TRUE FALSE FALSE FALSE FALSE FALSE
## 2        TRUE TRUE  TRUE FALSE FALSE FALSE FALSE
## 3        TRUE TRUE FALSE  TRUE  TRUE FALSE FALSE
## 4        TRUE TRUE FALSE  TRUE  TRUE  TRUE FALSE
## 5        TRUE TRUE  TRUE  TRUE  TRUE  TRUE FALSE
## 6        TRUE TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
```

```r
par(mfrow = c(2,2))
plot(rsum$adjr2, xlab = "Number of Variables", ylab = "Adjusted RSq", type = "l")
adjr2_max<-which.max(rsum$adjr2)
points(adjr2_max, rsum$adjr2[adjr2_max],col="green",cex = 2, pch = 20)

plot(rsum$cp, xlab = "Number of Variables", ylab = "Cp", type = "l")
cp_min = which.min(rsum$cp) # 7
points(cp_min, rsum$cp[cp_min], col = "red", cex = 2, pch = 20)

plot(rsum$bic, xlab = "Number of Variables", ylab = "BIC", type = "l")
bic_min = which.min(rsum$bic) # 6
points(bic_min, rsum$bic[bic_min], col = "turquoise", cex = 2, pch = 20)
```



The best model based on all 3 Model Selection Criteria is the one with 3 predictor variables, Y = b0 + b1X1 + b3X3 + b4X4. This is because it has the Highest Adj R^2, the Lowest Cp value, and the Lowest BIC value.

d) Fit a complete second-order model which contains all quadratic and cross-product terms. Use the regression subsets function in leaps package for the variable selection methods to determine the "best: model based on

1. Adjusted R2
2. Mallows Cp
3. BIC criteria

```
cityrefgull2 <- regsubsets(Y~X1+X2+X3+X4+X5+X6+I(X1^2)+I(X2^2)+I(X3^2)+I(X4^2)+I(X5^2)+I(X6^2)+
                             X1*X2 + X1*X3 + X1*X4 + X1*X5 + X1*X6 +
                                     X2*X3 + X2*X4 + X2*X5 + X2*X6 +
                                             X3*X4 + X3*X5 + X3*X6 +
                                                     X4*X5 + X4*X6 +
                                                             X5*X6
                            ,data=city)
```

```
## Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax, force.in =
## force.in, : 10 linear dependencies found
```

```r
rsum2 <- summary(cityrefgull2)
names(rsum2)
```

```
## [1] "which"  "rsq"    "rss"    "adjr2"  "cp"     "bic"    "outmat" "obj"
```

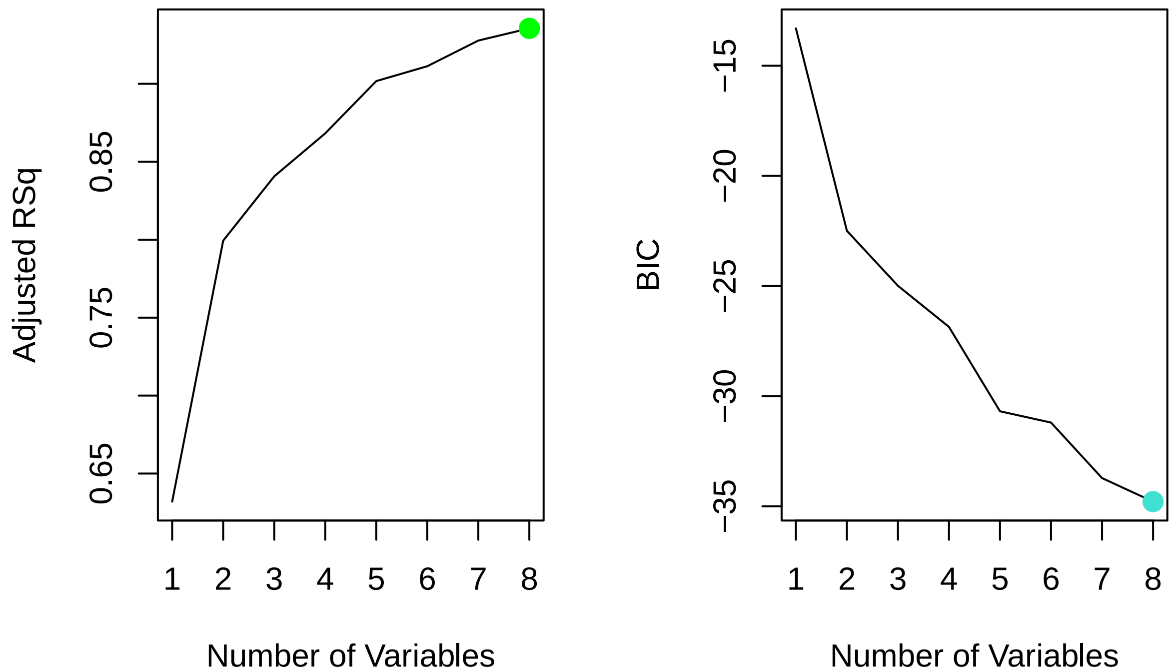
```r
rsum2$which
```

```
##   (Intercept)    X1    X2    X3    X4    X5    X6 I(X1^2) I(X2^2) I(X3^2)
## 1        TRUE  TRUE FALSE FALSE FALSE FALSE FALSE   FALSE   FALSE   FALSE
## 2        TRUE FALSE FALSE  TRUE FALSE FALSE FALSE   FALSE   FALSE   FALSE
## 3        TRUE FALSE FALSE FALSE  TRUE FALSE FALSE   FALSE   FALSE   FALSE
## 4        TRUE FALSE FALSE FALSE FALSE  TRUE FALSE    TRUE   FALSE   FALSE
## 5        TRUE FALSE FALSE FALSE  TRUE  TRUE FALSE   FALSE   FALSE   FALSE
## 6        TRUE FALSE FALSE FALSE  TRUE  TRUE FALSE   FALSE   FALSE   FALSE
## 7        TRUE FALSE FALSE FALSE FALSE FALSE FALSE   FALSE    TRUE   FALSE
## 8        TRUE  TRUE FALSE FALSE  TRUE FALSE FALSE    TRUE   FALSE   FALSE
##   I(X4^2) I(X5^2) I(X6^2) X1:X2 X1:X3 X1:X4 X1:X5 X1:X6 X2:X3 X2:X4 X2:X5 X2:X6
## 1   FALSE   FALSE   FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 2   FALSE   FALSE   FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
## 3   FALSE   FALSE   FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
## 4   FALSE   FALSE   FALSE FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE
## 5    TRUE    TRUE   FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE
## 6    TRUE    TRUE    TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 7    TRUE    TRUE   FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE  TRUE FALSE
## 8    TRUE   FALSE   FALSE FALSE  TRUE  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE
##   X3:X4 X3:X5 X3:X6 X4:X5 X4:X6 X5:X6
## 1 FALSE FALSE FALSE FALSE FALSE FALSE
## 2 FALSE FALSE FALSE FALSE FALSE FALSE
## 3 FALSE FALSE FALSE FALSE FALSE FALSE
## 4 FALSE FALSE FALSE FALSE FALSE FALSE
## 5 FALSE FALSE FALSE FALSE FALSE FALSE
## 6 FALSE FALSE FALSE FALSE FALSE  TRUE
## 7 FALSE  TRUE FALSE FALSE FALSE FALSE
## 8  TRUE FALSE FALSE FALSE FALSE FALSE
```

```r
par(mfrow = c(1,2))
plot(rsum2$adjr2, xlab = "Number of Variables", ylab = "Adjusted RSq", type = "l")
adjr2_max<-which.max(rsum2$adjr2)
points(adjr2_max, rsum2$adjr2[adjr2_max],col="green",cex = 2, pch = 20)

# plot(rsum2$cp, xlab = "Number of Variables", ylab = "Cp", type = "l")
# cp_min = which.min(rsum2$cp) # 7
# points(cp_min, rsum2$cp[cp_min], col = "red", cex = 2, pch = 20)

plot(rsum2$bic, xlab = "Number of Variables", ylab = "BIC", type = "l")
bic_min = which.min(rsum2$bic) # 6
points(bic_min, rsum2$bic[bic_min], col = "turquoise", cex = 2, pch = 20)
```



The best model based on Adjusted R^2 and BIC Criteria is the one with all 8 variables, with he regression equation Y= b0 + b1X1 + b2X4 + b3X1^2 + b4X4^2 + b5X1∗X3 + b6X1∗X4 + b7X2∗X3 +b8X3∗X4. This had the highest r^2 and the lowest BIC, while Mallow's Cp was negative infinity for all of them, so not useful.

e) Pick one best model from part c and part d and find press statistic to pick the final model

```
lm1 <- lm(Y~X1+X3+X4,data = city)
lm2 <- lm(Y~X1+X4+I(X1^2) + I(X4^2) + X1*X3 + X1*X4 + X3*X4 + X2*X3,data=city)

PRESS.statistic1 <- sum( (resid(lm1)/(1-hatvalues(lm1)))^2 )
print(paste("PRESS statistic= ", PRESS.statistic1))
```

```
## [1] "PRESS statistic=  0.199522369764292"
```

```
PRESS.statistic2 <- sum( (resid(lm2)/(1-hatvalues(lm2)))^2 )
print(paste("PRESS statistic= ", PRESS.statistic2))
```

```
## [1] "PRESS statistic=  0.23292658620828"
```

The best first order model has the lower PRESS statistic, so I would pick it as the final model.