CSC 423 Homework 5

Akhil Kumar Ramasagaram

May 6, 2016

6.6 Clerical staff work hours.

a)

```
library(leaps)
library(MASS)
library(ggplot2)
library(gridExtra)
library(MPV)
library(utils)
library(nnet)
load("rdata/CLERICAL.Rdata")
lm_model <- lm(Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7, data = CLERICAL)
model <- step(lm_model, direction = "both")
```

```
## Start:  AIC=256.6
## Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7
##
##        Df Sum of Sq    RSS    AIC
## - X7    1     92.38 5406.9 255.50
## - X3    1    108.46 5423.0 255.65
## <none>              5314.5 256.60
## - X1    1    261.73 5576.2 257.10
## - X6    1    311.75 5626.3 257.57
## - X2    1    395.05 5709.6 258.33
## - X4    1    734.70 6049.2 261.33
## - X5    1   1842.25 7156.8 270.08
##
## Step:  AIC=255.5
## Y ~ X1 + X2 + X3 + X4 + X5 + X6
##
##        Df Sum of Sq    RSS    AIC
## - X1    1    195.61 5602.5 255.35
## <none>              5406.9 255.50
## - X3    1    214.46 5621.3 255.52
## - X6    1    329.23 5736.1 256.57
## + X7    1     92.38 5314.5 256.60
## - X2    1    408.74 5815.6 257.29
## - X4    1    775.08 6182.0 260.46
## - X5    1   2307.92 7714.8 271.98
##
## Step:  AIC=255.35
## Y ~ X2 + X3 + X4 + X5 + X6
##
##        Df Sum of Sq    RSS    AIC
## <none>              5602.5 255.35
```

```
## + X1    1    195.61 5406.9 255.50
## - X3    1    329.66 5932.2 256.32
## - X2    1    361.53 5964.0 256.60
## - X6    1    377.54 5980.0 256.74
## + X7    1     26.26 5576.2 257.10
## - X4    1    761.70 6364.2 259.97
## - X5    1   2124.83 7727.3 270.07
```

```
summary(model)
```

```
##
## Call:
## lm(formula = Y ~ X2 + X3 + X4 + X5 + X6, data = CLERICAL)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.249  -7.439  -1.807   7.619  28.406
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 68.274431   7.701975   8.865 1.63e-11 ***
## X2           0.083086   0.048224   1.723  0.09162 .
## X3           0.013864   0.008427   1.645  0.10674
## X4          -0.043445   0.017372  -2.501  0.01602 *
## X5           0.044711   0.010705   4.177  0.00013 ***
## X6           0.229095   0.130120   1.761  0.08495 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.04 on 46 degrees of freedom
## Multiple R-squared:  0.545,  Adjusted R-squared:  0.4955
## F-statistic: 11.02 on 5 and 46 DF,  p-value: 5.19e-07
```

b) out of 7 initial varialbe after stepwise selection, we got 5 variable as significant variables.
Our intercept $\beta_0 = 68.27$ which can be interpreted as, when everything is zero we can expect that a staff worked 68.27 hr.
$\beta_1 = 0.08$ which can be interpreted as when a unit increase in x_2 result in 0.08 hr incraese in hours worked.
$\beta_2 = 0.01$ which can be interpreted as when a unit increase in x_3 result in 0.01 hr incraese in hours worked.
$\beta_3 = -0.04$ which can be interpreted as when a unit increase in x_4 result in 0.04 hr decrease in hours worked.
$\beta_4 = 0.045$ which can be interpreted as when a unit increase in x_5 result in 0.045 hr decrease in hours worked.
$\beta_5 = 0.23$ which can be interpreted as when a unit increase in x_6 result in 0.23 hr increase in hours worked.

c)

(i) Sometimes in might produce a model which the computer thinks best, but it may not be as good we though. So even after stepwise selection one should use domain expertise for making judgement on whether to use that model or not.

(ii) If the number of variables that you select for testing is large compared to the number of observations in your data set or if there is high multicollinearity among the variables, then we will end up using almost all variable.

(iii) Sometimes a subset of variables that are treated as a group. Stepwise regression may throw some of them out, in which case one should have to manually put them back in later. ### 6.7

a) For 1, 2, 3 & 4 variable we get 4, 6, 4 & 1 possible models.
when n = 1; x_1, x_2, x_3 & x_4
when n = 2; x_1x_2, x_2x_3, x_3x_4, x_1x_3, x_1x_4 & x_2x_4
when n = 3; $x_1x_2x_3$, $x_2x_3x_4$, $x_1x_3x_4$ & $x_1x_2x_4$
when n = 4; $x_1x_2x_3x_4$

b)

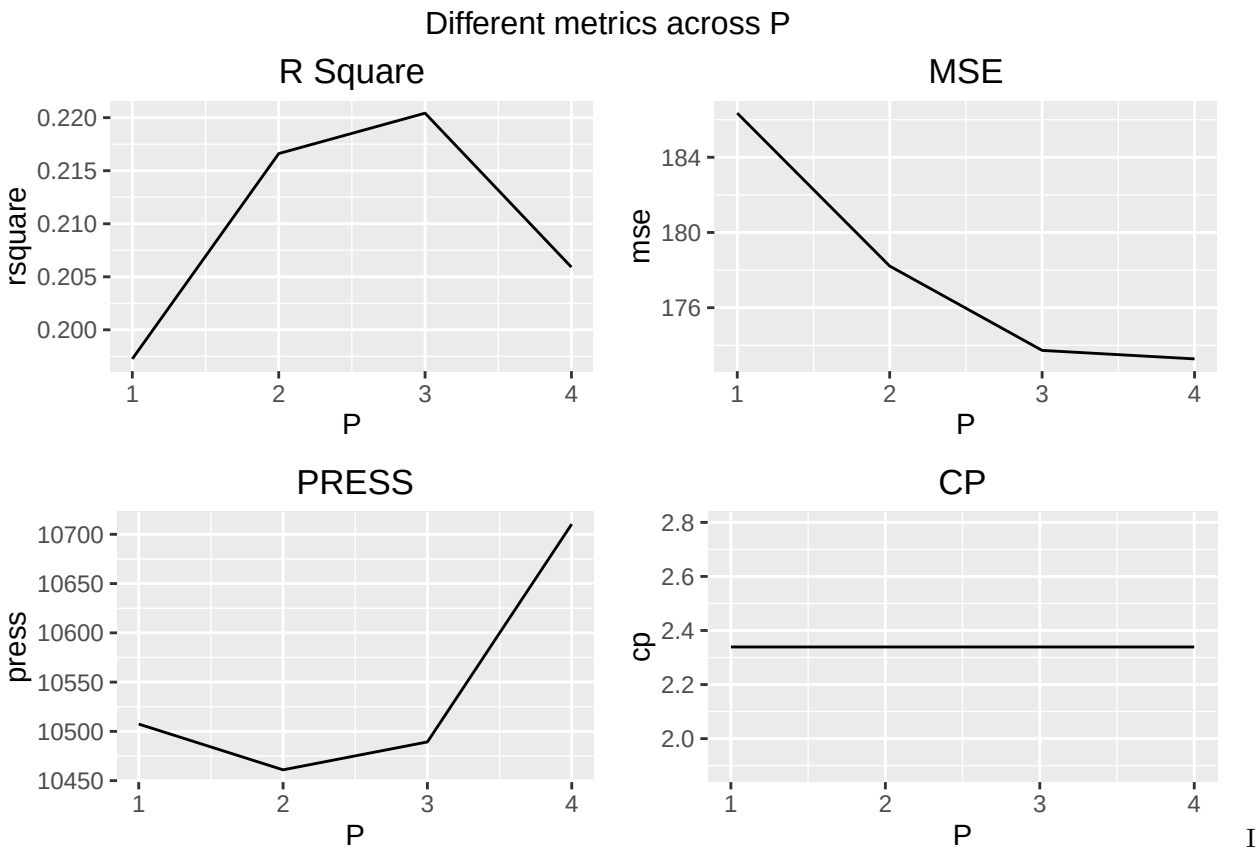
```
different_comb <- function(plot = F){
  final_result <- data.frame("P" = NULL, "rsquare" = NULL, "mse" = NULL, "press" = NULL, "cp" = NULL)
  for (p in seq(4)){
    xvars = c("X1", "X2", "X3", "X4")
    yvar <- "Y"
    rsquare <- NULL
    mse <- NULL
    press <- NULL
    cp <- NULL
    comb <- combn(xvars, p)
    for(i in seq(ncol(comb))){
      form <- as.formula(paste("Y ~ ", paste(comb[,i], collapse = " + "), sep = ""))
      lm_model <- lm(form, data = CLERICAL)
      sm <- summary(lm_model)
      rsquare[i] <- sm$adj.r.squared
      mse[i] <- mean(sm$residuals^2)
      press[i] <- PRESS(lm_model)
    }
    model=leaps( x=CLERICAL[,xvars], y=CLERICAL[,yvar], names=xvars, nbest=p, method="Cp")
    cp <- model$Cp
    final_result <- rbind(final_result, data.frame("P" = p, "rsquare" = max(rsquare),
                                                   "mse" = min(mse), "press" = min(press), "cp" = min(cp
  }
  if(plot == T){
    p1 <- ggplot(final_result, aes(y = rsquare, x = P)) + geom_line() + ggtitle("R Square")
    p2 <- ggplot(final_result, aes(y = mse, x = P)) + geom_line() + ggtitle("MSE")
    p3 <- ggplot(final_result, aes(y = press, x = P)) + geom_line() + ggtitle("PRESS")
    p4 <- ggplot(final_result, aes(y = cp, x = P)) + geom_line() + ggtitle("CP")
    grid.arrange(p1,p2,p3,p4, ncol = 2, top = "Different metrics across P")
  }
  else{
    return(final_result)
  }
}

different_comb(F)
```

```
##   P   rsquare      mse    press       cp
## 1 1 0.1972599 186.3452 10507.43 2.339122
```

```
## 2 2 0.2166171 178.2146 10460.99 2.339122
## 3 3 0.2204242 173.7292 10489.28 2.339122
## 4 4 0.2059069 173.2776 10710.36 2.339122
```

c) & d)

```
different_comb(T)
```



I would go with p = 3 as it has good R square and MSE and 2nd best press value.

6.8 Collusive bidding in road construction.

a)

```
load("rdata/FLAG2.Rdata")
FLAG2$SUBCONT[which(is.na(FLAG2$SUBCONT))] <- 0
districts <- data.frame(class.ind(FLAG2$DISTRICT))[1:4]
names(districts) <- c("district1", "district2", "district3", "district4")
FLAG2$DISTRICT <- NULL
FLAG2 <- cbind(FLAG2, districts)
lm_model <- lm(LOWBID ~ ., data = FLAG2)
model <- step(lm_model, direction = "both", trace = F)
summary(model)
```

```
##
```

```
## Call:
## lm(formula = LOWBID ~ DOTEST + LBERATIO + RDLNGTH + PCTASPH +
##     PCTBASE + PCTEXCAV + district4, data = FLAG2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1971750   -63233     7861    70976  1388637
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.617e+05  9.942e+04  -7.662 3.26e-13 ***
## DOTEST       9.363e-01  8.467e-03 110.582  < 2e-16 ***
## LBERATIO     8.345e+05  9.061e+04   9.209  < 2e-16 ***
## RDLNGTH      6.217e+03  4.296e+03   1.447   0.1490
## PCTASPH     -1.265e+05  6.714e+04  -1.884   0.0606 .
## PCTBASE      2.348e+05  1.526e+05   1.539   0.1249
## PCTEXCAV    -2.632e+05  1.361e+05  -1.935   0.0541 .
## district4   -2.823e+05  1.183e+05  -2.387   0.0177 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 249300 on 271 degrees of freedom
## Multiple R-squared:  0.9818, Adjusted R-squared:  0.9813
## F-statistic:  2087 on 7 and 271 DF,  p-value: < 2.2e-16
```

So you mode suitable variables for our model are DOTEST, LBERATIO, RDLNGTH, PCTASPH, PCTBASE, PCTEXCAV, district4

b) $\beta_0 = -761725$, when everything is 0 and district is 5 our lowest bid is -761725, which has no meaning in this context. $\beta_1 = 0.9314$ for a unit increase in dotest, our lowbid decreases by 0.93 $\beta_2 = 834464$ for a unit increase in lberatio, our lowbid increases by 834464 $\beta_3 = 6217$ for a unit increase in rdlength, our lowbid increases by 6217 $\beta_4 = -126491$ for a unit increase in pctasph, our lowbid decreases by 126491 $\beta_5 = 234831$ for a unit increase in pctbase, our lowbid increases by 234831 $\beta_6 = -263220$ for a unit increase in pctexcav, our lowbid decreases by 263220 $\beta_7 = -282346$ if the district is 4, our lowbid decreases by 282346

c) discussed in 6.6 c

6.9

```
yvar <- "LOWBID"
xvars <- names(FLAG2)[-1]
model=leaps( x=FLAG2[,xvars], y=FLAG2[,yvar], names=xvars, nbest=3, method="Cp")
imp_var <- model$which[which.min(model$Cp),]
names(imp_var)[as.character(imp_var) == "TRUE"]
```

```
## [1] "DOTEST"    "LBERATIO"  "PCTBASE"   "PCTEXCAV"  "district4"
```

```
lm_model <- lm(LOWBID ~ DOTEST+LBERATIO+PCTBASE+PCTEXCAV+district4, FLAG2)
summary(lm_model)
```

```
##
## Call:
## lm(formula = LOWBID ~ DOTEST + LBERATIO + PCTBASE + PCTEXCAV +
##     district4, data = FLAG2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1971292   -58859    13866    62834  1406194
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.983e+05  9.464e+04  -8.435 1.95e-15 ***
## DOTEST       9.375e-01  8.455e-03 110.885  < 2e-16 ***
## LBERATIO     8.325e+05  9.060e+04   9.189  < 2e-16 ***
## PCTBASE      3.047e+05  1.480e+05   2.059   0.0404 *
## PCTEXCAV    -2.141e+05  1.327e+05  -1.613   0.1079
## district4   -2.839e+05  1.186e+05  -2.393   0.0174 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 250100 on 273 degrees of freedom
## Multiple R-squared:  0.9815, Adjusted R-squared:  0.9812
## F-statistic:  2901 on 5 and 273 DF,  p-value: < 2.2e-16
```

its almost close. we just dont have RDLNGTH & PCTASPH in this model now. The R square is also almost equal.

6.10 Cooling method for gas turbines.

```
load("rdata/GASTURBINE.Rdata")
GASTURBINE$ENGINE <- ifelse(GASTURBINE$ENGINE == "Traditional",0,1)
lm_model <- lm(HEATRATE ~ ., GASTURBINE)
model <- step(lm_model, direction = "both", trace = F)
summary(model)
```

```
##
## Call:
## lm(formula = HEATRATE ~ ENGINE + RPM + CPRATIO + INLETTEMP +
##     POWER + LHV, data = GASTURBINE)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -462.36 -100.64  -16.01   72.66  410.34
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.438e+04  4.659e+02  52.321  < 2e-16 ***
## ENGINE       2.281e+02  5.889e+01   3.873 0.000268 ***
## RPM          2.193e-02  5.818e-03   3.769 0.000376 ***
## CPRATIO      3.720e+01  8.779e+00   4.237 7.91e-05 ***
## INLETTEMP   -1.180e+00  3.451e-01  -3.418 0.001138 **
## POWER        2.810e-03  3.829e-04   7.339 6.60e-10 ***
```

```
## LHV         -3.922e+02  1.682e+01 -23.315  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 163.9 on 60 degrees of freedom
## Multiple R-squared:  0.9904, Adjusted R-squared:  0.9894
## F-statistic:  1032 on 6 and 60 DF,  p-value: < 2.2e-16
```

```
model <- step(lm_model, direction = "backward", trace = F)
summary(model)
```

```
##
## Call:
## lm(formula = HEATRATE ~ ENGINE + RPM + CPRATIO + INLETTEMP +
##     POWER + LHV, data = GASTURBINE)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -462.36 -100.64  -16.01   72.66  410.34
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.438e+04  4.659e+02  52.321  < 2e-16 ***
## ENGINE       2.281e+02  5.889e+01   3.873 0.000268 ***
## RPM          2.193e-02  5.818e-03   3.769 0.000376 ***
## CPRATIO      3.720e+01  8.779e+00   4.237 7.91e-05 ***
## INLETTEMP   -1.180e+00  3.451e-01  -3.418 0.001138 **
## POWER        2.810e-03  3.829e-04   7.339 6.60e-10 ***
## LHV         -3.922e+02  1.682e+01 -23.315  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 163.9 on 60 degrees of freedom
## Multiple R-squared:  0.9904, Adjusted R-squared:  0.9894
## F-statistic:  1032 on 6 and 60 DF,  p-value: < 2.2e-16
```

```
yvar <- "HEATRATE"
xvars <- names(GASTURBINE)[-9]
model=leaps( x=GASTURBINE[,xvars], y=GASTURBINE[,yvar], names=xvars, nbest=3, method="Cp")
imp_var <- model$which[which.min(model$Cp),]
names(imp_var)[as.character(imp_var) == "TRUE"]
```

```
## [1] "ENGINE"    "RPM"       "CPRATIO"   "INLETTEMP" "POWER"     "LHV"
```

All three methods exact same variables.