

## PROBLEM 15

Clerical staff work hours. In any production process in which one or more workers are engaged in a variety of tasks, the total time spent in production varies as a function of the size of the work pool and the level of output of the various activities.

For example, in a large metropolitan department store, the number of hours worked (Y) per day by the clerical staff may depend on the following

variables:

X1 = Number of pieces of mail processed (open, sort, etc.)

X2 = Number of money orders and gift certificates sold,

X3 = Number of window payments (customer charge accounts) transacted ,

X4 = Number of change order transactions processed ,

X5 = Number of checks cashed ,

X6 =Number of pieces of miscellaneous mail processed on an "as available" basis , and

X7 =Number of bus tickets sold

The data are provided in **CLERICAL.csv** file count for these activities on each of 52 working days. Conduct a Stepwise Regression Procedure and All-Possible-Regressions procedure of the data using R software package.

### APPROACH ONE: TRY TO BUILD A "BEST MODEL" MANUALLY

1. Fit the model with all terms thought to be important as below in `firstordermodel`.
2. Remove terms that are not significant with individual t-tests, this becomes `model`.
3. Conduct a partial F test using `anova()` to determine if this was justified.

```
workhours=read.csv("CLERICAL.csv",header = TRUE)
firstordermodel<-lm(Y~X1+X2+X3+X4+X5+X6+X7,data=workhours)
summary(firstordermodel)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7, data = workhours)
##
## Residuals:
##      Min    1Q   Median    3Q   Max 
## -18.537 -7.038 -1.224  6.168 28.012 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 60.5537920  9.4952130  6.377  9.4e-08 ***
## X1          0.0013496  0.0009168  1.472  0.14813  
## X2          0.0972715  0.0482561  1.809  0.07736 .  
##
```

```

## X3      0.0086879  0.0091681   0.948  0.34850
## X4     -0.0427781  0.0173449  -2.466  0.01762 *
## X5      0.0467992  0.0119898   3.905  0.00032 ***
## X6      0.2092130  0.1302236   1.607  0.11530
## X7      0.0048192  0.0055105   0.875  0.38657
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.99 on 44 degrees of freedom
## Multiple R-squared:  0.5684, Adjusted R-squared:  0.4997
## F-statistic: 8.277 on 7 and 44 DF,  p-value: 2.053e-06

```

```

model<-lm(Y~X2+X4+X5,data=workhours)
summary(model)

```

```

##
## Call:
## lm(formula = Y ~ X2 + X4 + X5, data = workhours)
##
## Residuals:
##    Min     1Q Median     3Q    Max 
## -21.259 -9.075 -1.938  6.882 29.303 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 77.725640   6.910199 11.248 4.69e-15 ***
## X2          0.136264   0.045413  3.001 0.00426 **  
## X4         -0.034689   0.017140  -2.024 0.04857 *   
## X5          0.058268   0.009714   5.998 2.52e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.54 on 48 degrees of freedom
## Multiple R-squared:  0.4806, Adjusted R-squared:  0.4481
## F-statistic: 14.8 on 3 and 48 DF,  p-value: 5.91e-07

```

```

anova(model,firstordermodel)

```

```

## Analysis of Variance Table
##
## Model 1: Y ~ X2 + X4 + X5
## Model 2: Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7
##   Res.Df   RSS Df Sum of Sq    F Pr(>F)    
## 1     48 6395.3
## 2     44 5314.5  4   1080.8 2.2371 0.08035 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

4. The Partial  $F$  indicates removal of those terms was justifiable (strictly using  $\alpha = 0.05$ ). So determine whether it is justifiable to add two-way interactions across all main effects found in `firstordermodel`.

```

#Model1
interactmodel1<-lm(Y~(X2+X4+X5)^2, data=workhours)
summary(interactmodel1)

##
## Call:
## lm(formula = Y ~ (X2 + X4 + X5)^2, data = workhours)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -20.734 -8.232 -1.018  7.021 28.770 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 6.598e+01 2.199e+01  3.000  0.00439 ** 
## X2          2.040e-01 1.840e-01  1.108  0.27363    
## X4         -2.027e-02 6.819e-02 -0.297  0.76759    
## X5          8.380e-02 4.244e-02  1.975  0.05444 .  
## X2:X4       1.241e-04 4.965e-04  0.250  0.80382    
## X2:X5      -1.712e-04 3.313e-04 -0.517  0.60789    
## X4:X5      -4.323e-05 9.246e-05 -0.468  0.64238    
## ---      
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
##
## Residual standard error: 11.87 on 45 degrees of freedom
## Multiple R-squared:  0.485, Adjusted R-squared:  0.4163 
## F-statistic: 7.063 on 6 and 45 DF, p-value: 2.43e-05

```

5. Interactions don't appear to improve anything, so let's go back to the model with only first-order effects among the variables identified as significant on an individual  $t$ -test.

```

bestmodel1<-lm(Y~X2+X4+X5, data=workhours)
summary(bestmodel1)

##
## Call:
## lm(formula = Y ~ X2 + X4 + X5, data = workhours)
##
## Residuals:
##     - - -
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 77.725640   6.910199 11.248 4.69e-15 *** 
## X2          0.136264   0.045413  3.001  0.00426 **  
## X4         -0.034689   0.017140 -2.024  0.04857 *  
## X5          0.058268   0.009714  5.998 2.52e-07 *** 
## ---      
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
##
## Residual standard error: 11.54 on 48 degrees of freedom

```

- borderline suggestion  
Naah!

no really good interactions  
keep reduced model

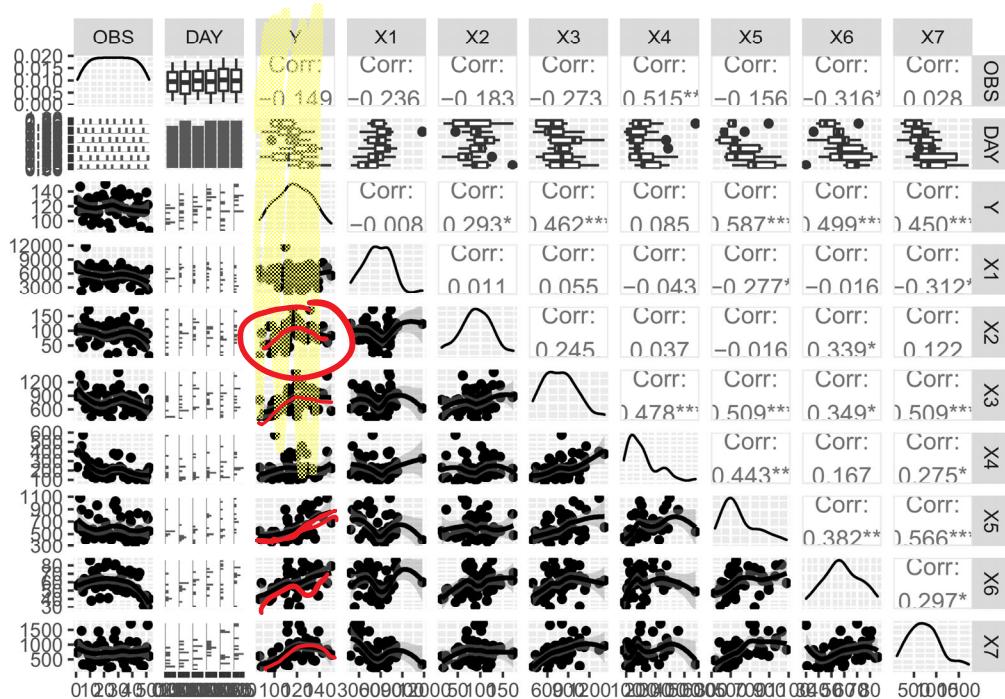


```
## Multiple R-squared:  0.4806, Adjusted R-squared:  0.4481
## F-statistic:  14.8 on 3 and 48 DF,  p-value: 5.91e-07
```

6. Now let's check if we can improve the model with *higher-order terms*. We should look to see which terms we might target optimally for potential curvilinearity (i.e., plots with  $Y$  against each of  $X1...X7$$ ). It looks as if  $X2$  and  $X5$  are worth trying. So let's do them separately as `bestmodel11` and `bestmodel12`

```
#Improving model Individual T test
library(GGally)

ggpairs(workhours,lower = list(continuous = "smooth_loess", combo =
  "facethist", discrete = "facetbar", na = "na"))
```



```
bestmodel11<-lm(Y~X2+I(X2^2)+X4+X5, data=workhours)
summary(bestmodel11)
```

```
##
## Call:
## lm(formula = Y ~ X2 + I(X2^2) + X4 + X5, data = workhours)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -24.315   -6.480    1.185    5.320   26.482 
## 
## Coefficients:
```

```

##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 61.0933183  8.7297596  6.998 8.22e-09 ***
## X2          0.5762076  0.1611431   3.576 0.000821 ***
## I(X2^2)     -0.0024326  0.0008596  -2.830 0.006827 **
## X4          -0.0326852  0.0160268  -2.039 0.047054 *
## X5          0.0571700  0.0090822   6.295 9.62e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.78 on 47 degrees of freedom
## Multiple R-squared:  0.5562, Adjusted R-squared:  0.5184
## F-statistic: 14.73 on 4 and 47 DF,  p-value: 7.196e-08

bestmodel12<-lm(Y~X2+I(X2^2)+X4+X5+I(X5^2), data=workhours)
summary(bestmodel12)

```

```

## 
## Call:
## lm(formula = Y ~ X2 + I(X2^2) + X4 + X5 + I(X5^2), data = workhours)
##
## Residuals:
##    Min      1Q      Median      3Q      Max
## -23.5578 -6.9145  0.8808  5.8568 25.3756
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.097e+01  2.059e+01  2.475 0.017051 *
## X2          5.892e-01  1.625e-01  3.570 0.000849 ***
## I(X2^2)     -2.479e-03  8.703e-04 -2.848 0.006648 **
## X4          -3.358e-02  1.623e-02 -2.069 0.044214 *
## X5          9.044e-02  6.187e-02  1.462 0.150604
## I(X5^2)     -2.425e-05  4.459e-05 -0.544 0.589264
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.86 on 46 degrees of freedom
## Multiple R-squared:  0.559, Adjusted R-squared:  0.5111
## F-statistic: 11.66 on 5 and 46 DF,  p-value: 2.612e-07

```

## APPROACH 2: Use Stepwise regression

- Run a stepwise regression on the additive model with all potential terms. Note, we have `details=FALSE` here, and the output of the stepwise only is shown.

```

# Stepwise Method_
library(olsrr) #need to install the package olsrr

```

```

## Warning: package 'olsrr' was built under R version 4.2.2
##
## Attaching package: 'olsrr'

```

✓ Good decision  
to add  
X<sub>2</sub>

X probably  
"best" DATA 63  
didn't help.  
mode by hand

```

## The following object is masked from 'package:datasets':
##
##      rivers

library(leaps) #need to install the package leaps for best.subset() function

## Warning: package 'leaps' was built under R version 4.2.2

workhours=read.csv("CLERICAL.csv",header = TRUE)

#Using Stepwise Regression Procedure for data selection
firstordermodel<-lm(Y~X1+X2+X3+X4+X5+X6+X7,data=workhours)
step <- ols_step_both_p(firstordermodel,pent = 0.1, prem = 0.3, details=FALSE)
summary(step$model)

## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##      data = 1)
##
## Residuals:
##      Min        1Q    Median        3Q       Max
## -22.7666  -8.3861  -0.4456   8.5525  25.9007
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 70.44910   7.72424   9.121 5.73e-12 ***
## X5          0.05075   0.01024   4.957 9.73e-06 ***
## X2          0.10212   0.04766   2.143   0.0373 *
## X4         -0.03398   0.01669  -2.036   0.0474 *
## X6          0.25226   0.13168   1.916   0.0615 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.23 on 47 degrees of freedom
## Multiple R-squared:  0.5182, Adjusted R-squared:  0.4772
## F-statistic: 12.64 on 4 and 47 DF,  p-value: 4.647e-07

```

$X_2, X_4$  if  
 - potentially  
 consider  
 (marginal)

- Take the output of stepwise regression and see if two-way interaction terms are justifiable. The first attempt `interactmodel2` shows that there *may be a chance that  $X_2, X_6$  might remain significant when added to the first-order model with just main effects.* So let's try just that in `bestmodel21`, but we find that it doesn't work that well as the  $P > \alpha$  (i.e. 0.05), so let's go back to just the main effects in `bestmodel22`.

```

library(olsrr) #need to install the package olsrr
library(leaps) #need to install the package leaps for best.subset() function
workhours=read.csv("CLERICAL.csv",header = TRUE)

#Model2
mod2=lm(Y~X2+X4+X5+X6, data=workhours)
interactmodel2<-lm(Y~(X2+X4+X5+X6)^2, data=workhours)
summary(interactmodel2)

```

```

## 
## Call:
## lm(formula = Y ~ (X2 + X4 + X5 + X6)^2, data = workhours)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -25.1071  -7.0977  -0.5452  6.9982 23.7102 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 2.861e+00 3.921e+01  0.073  0.9422    
## X2          5.175e-01 2.785e-01  1.858  0.0704 *  
## X4         -5.762e-02 9.360e-02 -0.616  0.5416    
## X5          1.071e-01 6.585e-02  1.626  0.1116    
## X6          1.571e+00 7.092e-01  2.216  0.0323 *  
## X2:X4     -9.526e-05 5.644e-04 -0.169  0.8668    
## X2:X5      1.425e-04 3.411e-04  0.418  0.6783    
## X2:X6     -8.931e-03 4.503e-03 -1.983  0.0541 .  
## X4:X5     -2.535e-05 9.734e-05 -0.260  0.7958    
## X4:X6      7.972e-04 1.984e-03  0.402  0.6899    
## X5:X6     -1.100e-03 8.586e-04 -1.282  0.2071    
## ---      
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## Residual standard error: 11.41 on 41 degrees of freedom
## Multiple R-squared:  0.5661, Adjusted R-squared:  0.4603 
## F-statistic: 5.349 on 10 and 41 DF,  p-value: 5.085e-05

bestmodel21<-lm(Y~X2+X4+X5+X6+X2:X6, data=workhours)
summary(bestmodel21)

```

- maybe we should try  $Y^2 + 6$

```

## 
## Call:
## lm(formula = Y ~ X2 + X4 + X5 + X6 + X2 * X6, data = workhours)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -24.364  -7.618  -0.616   7.252  24.350 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 40.296332 19.291735  2.089  0.0423 *  
## X2          0.474925  0.224291  2.117  0.0397 *  
## X4         -0.033759  0.016362 -2.063  0.0448 *  
## X5          0.047836  0.010183  4.698  2.4e-05 *** 
## X6          0.841726  0.370116  2.274  0.0277 *  
## X2:X6     -0.096664  0.003922 -1.699  0.0960 .  
## ---      
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## Residual standard error: 11.02 on 46 degrees of freedom
## Multiple R-squared:  0.5467, Adjusted R-squared:  0.4974 
## F-statistic: 11.09 on 5 and 46 DF,  p-value: 4.785e-07

```

- more marginally significant

```

bestmodel22<-lm(Y~X2+X4+X5+X6, data=workhours)
summary(bestmodel22)

## 
## Call:
## lm(formula = Y ~ X2 + X4 + X5 + X6, data = workhours)
## 
## Residuals:
##      Min    1Q   Median    3Q    Max 
## -22.7666 -8.3861 -0.4456  8.5525 25.9007 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 70.44910   7.72424   9.121 5.73e-12 ***
## X2          0.10212   0.04766   2.143  0.0373 *  
## X4         -0.03398   0.01669  -2.036  0.0474 *  
## X5          0.05075   0.01024   4.957 9.73e-06 *** 
## X6          0.25226   0.13168   1.916   0.0615 .  
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## Residual standard error: 11.23 on 47 degrees of freedom 
## Multiple R-squared:  0.5182, Adjusted R-squared:  0.4772 
## F-statistic: 12.64 on 4 and 47 DF,  p-value: 4.647e-07

```

- Move forward without  
+  
X2+X6

3. Now let's see if it is justifiable to add any second-order terms. We'll let you look back about at the `ggpairs()` plot. We think (this time) that  $X_2$  and  $X_6$  might be good candidates to be non-linear terms, given that we found that  $X_5$  wasn't so great last time. But this is really just a guess. And this time we find that  $X_2$  was justified, but not  $X_6$  as a quadratic (i.e., as  $\$X_2^2\$$ ). But we'd probably want to re-run with  $X_6$  `bestmodel23`.

```

#Improving model from Stepwise method
library(olsrr) #need to install the package olsrr
library(GGally) # need toinstall the GGally package for ggpairs function

workhours=read.csv("CLERICAL.csv",
                    header = TRUE)

#ggpairs(workhours, lower = list(continuous = "smooth_loess", combo =
# "facethist", discrete = "facetbar", na = "na"))

bestmodel21<-lm(Y~X2+I(X2^2)+X4+X5+X6*I(X6^2), data=workhours)
summary(bestmodel21)

```

```

## 
## Call:
## lm(formula = Y ~ X2 + I(X2^2) + X4 + X5 + X6 + I(X6^2), data = workhours)
## 
## Residuals:
##      Min    1Q   Median    3Q    Max 
## -25.980 -6.680   1.176   6.473  23.328

```

```

## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 45.9256582 24.9048015   1.844 0.07177 .  
## X2          0.5160220  0.1643950   3.139 0.00299 ** 
## I(X2^2)    -0.0022744  0.0008637  -2.633 0.01154 *  
## X4          -0.0337748  0.0162257  -2.082 0.04310 *  
## X5          0.0517135  0.0098304   5.261 3.85e-06 *** 
## X6          0.5867189  0.8456851   0.694 0.49139    
## I(X6^2)    -0.0032293  0.0071060  -0.454 0.65169    
## ---      
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

## Residual standard error: 10.68 on 45 degrees of freedom
## Multiple R-squared:  0.5829, Adjusted R-squared:  0.5273 
## F-statistic: 10.48 on 6 and 45 DF,  p-value: 2.967e-07

```

```

bestmodel22<-lm(Y~X2+I(X2^2)+X4+X5+X6, data=workhours)
summary(bestmodel22)

```

```

## 
## Call:
## lm(formula = Y ~ X2 + I(X2^2) + X4 + X5 + X6, data = workhours)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.3063 -7.4222 -0.0186  6.2710 23.4825
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 56.4807725 9.0226268  6.258 1.19e-07 *** 
## X2          0.5130400  0.1628411   3.151 0.00286 ** 
## I(X2^2)    -0.0022381  0.0008525  -2.625 0.01171 *  
## X4          -0.0322631  0.0157435  -2.049 0.04616 *  
## X5          0.0510943  0.0096512   5.294 3.26e-06 *** 
## X6          0.2067173  0.1253390   1.649 0.10591    
## ---      
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

## Residual standard error: 10.59 on 46 degrees of freedom
## Multiple R-squared:  0.581, Adjusted R-squared:  0.5354 
## F-statistic: 12.76 on 5 and 46 DF,  p-value: 8.518e-08

```

```

bestmodel23<-lm(Y~X2+I(X2^2)+X4+X5, data=workhours)
summary(bestmodel23)

```

```

## 
## Call:
## lm(formula = Y ~ X2 + I(X2^2) + X4 + X5, data = workhours)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24.315 -6.480   1.185   5.320  26.482

```

*Wain't  
worth  
adding!*

*- Final  
model*

```

## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 61.0933183  8.7297596  6.998 8.22e-09 ***
## X2          0.5762076  0.1611431   3.576 0.000821 ***
## I(X2^2)     -0.0024326  0.0008596  -2.830 0.006827 **
## X4          -0.0326852  0.0160268  -2.039 0.047054 *
## X5          0.0571700  0.0090822   6.295 9.62e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.78 on 47 degrees of freedom
## Multiple R-squared:  0.5562, Adjusted R-squared:  0.5184
## F-statistic: 14.73 on 4 and 47 DF,  p-value: 7.196e-08

```

### APPROACH THREE: Use all-best-subsets regression.

- Run an all-best subsets regression using `regsubsets()`, and then combine the salient model selection diagnostics into a single table to display using `cbind()`.

```

#Using All possible Regression
library(olsrr) #need to install the package olsrr
library(leaps) #need to install the package leaps for best.subset() function

workhours=read.csv("CLERICAL.csv",
header = TRUE)

#option 2
best.subset<-regsubsets(Y~X1+X2+X3+X4+X5+X6+X7, data= workhours)
summary(best.subset)

## Subset selection object
## Call: regsubsets.formula(Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7, data = workhours)
## 7 Variables (and intercept)
## Forced in Forced out
## X1 FALSE FALSE
## X2 FALSE FALSE
## X3 FALSE FALSE
## X4 FALSE FALSE
## X5 FALSE FALSE
## X6 FALSE FALSE
## X7 FALSE FALSE
## 1 subsets of each size up to 7
## Selection Algorithm: exhaustive
##          X1 X2 X3 X4 X5 X6 X7
## 1 ( 1 ) " " " " " * " "
## 2 ( 1 ) " " " * " " " " * " "
## 3 ( 1 ) " " " * " " " * " " * " "
## 4 ( 1 ) " " " * " " " * " " * " "
## 5 ( 1 ) " " " * " " * " " * " " * " "
## 6 ( 1 ) " * " " * " " * " " * " " * " "
## 7 ( 1 ) " * " " * " " * " " * " " * " "

```

```

reg.summary<-summary(best.subset)
rsquare<-c(reg.summary$rsq)
cp<-c(reg.summary$cp)
AdjustedR<-c(reg.summary$adjr2)
RMSE<-c(reg.summary$rss)

## Display model selection diagnostics we just aggregated
cbind(rsquare, cp, RMSE, AdjustedR)

```

# of Variables in the best model of that size

using  $R^2$  alone we might select

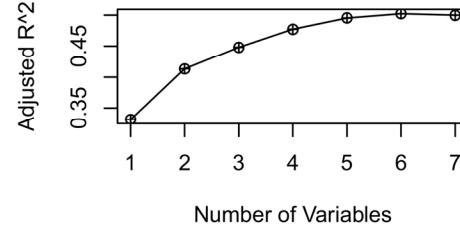
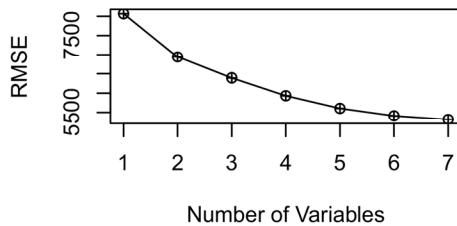
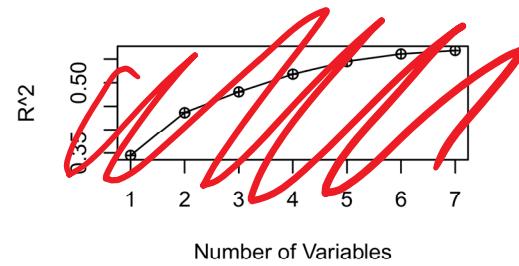
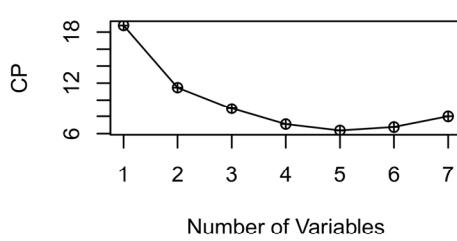
pretty good on CP!

```

## [1]   rsquare      cp      RMSE AdjustedR
## [2,] 0.436 18.775229 8065.390 0.3318425
## [3,] 0.46622 11.466378 6941.028 0.4132525
## [4,] 0.47043 8.948273 6395.312 0.4481208
## [5,] 0.47013 7.113662 5932.152 0.4771972
## [6,] 0.470 6.384302 5602.489 0.4955168
## [7,] 0.470 6.764836 5406.883 0.5023111
## [8,] 0.470 3.000000 5314.503 0.4996966

## Plot these (if it helps).
par(mfrow=c(2,2)) # split the plotting panel into a 2 x 2 grid
plot(reg.summary$cp,type = "o",pch=10, xlab="Number of Variables",ylab= "CP")
plot(reg.summary$rsq,type = "o",pch=10, xlab="Number of Variables",ylab= "R^2")
plot(reg.summary$rss,type = "o",pch=10, xlab="Number of Variables",ylab= "RMSE")
plot(reg.summary$adjr2,type = "o",pch=10, xlab="Number of Variables",ylab= "Adjusted R^2")

```



2. We decide, using our favourite diagnostic (or diagnostics) let's say, that the model with 6 terms (i.e., [6,] on the sixth row, is the best choice. So let's move it forward and see if any two-way interactions are justified. They are! So let's add just those two interactions (i.e., X1:X6 and X2:X6) into the original first-order model and see if they stay significant. They do. So let's move it forward.

```

workhours=read.csv("CLERICAL.csv", header = TRUE)
mod3=lm(Y~X1+X2+X3+X4+X5+X6, data=workhours)
interactmodel3<-lm(Y~(X1+X2+X3+X4+X5+X6)^2, data=workhours)
summary(interactmodel3)

##
## Call:
## lm(formula = Y ~ (X1 + X2 + X3 + X4 + X5 + X6)^2, data = workhours)
##
## Residuals:
##    Min      1Q   Median      3Q     Max 
## -19.1807 -5.3442  0.7369  4.1799 21.0237 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 1.109e+01  8.736e+01   0.127  0.8998    
## X1          -3.470e-03  8.140e-03  -0.426  0.6729    
## X2           1.185e+00  5.544e-01   2.137  0.0409 *  
## X3           7.461e-02  8.003e-02   0.932  0.3587    
## X4          -2.598e-01  1.822e-01  -1.425  0.1644    
## X5           3.511e-02  1.011e-01   0.347  0.7308    
## X6           6.822e-01  1.256e+00   0.543  0.5911    
## X1:X2        -7.714e-05  5.982e-05  -1.289  0.2071    
## X1:X3        -6.290e-06  8.452e-06  -0.744  0.4626    
## X1:X4        1.875e-05  1.762e-05  1.064  0.2959    
## X1:X5        2.033e-06  8.322e-06  0.244  0.8087    
## X1:X6        2.140e-04  1.069e-04  2.002  0.0544 .  
## X2:X3        3.022e-04  4.404e-04  0.686  0.4979    
## X2:X4        -1.179e-03  8.281e-04  -1.423  0.1649    
## X2:X5        5.483e-04  4.720e-04  1.162  0.2545    
## X2:X6        -1.791e-02  6.934e-03  -2.583  0.0149 *  
## X3:X4        8.826e-05  1.580e-04  0.559  0.5806    
## X3:X5        -1.071e-04  9.664e-05  1.108  0.2767    
## X3:X6        -2.485e-04  8.660e-04  -0.287  0.7761    
## X4:X5        1.460e-04  2.460e-04  0.594  0.5572    
## X4:X6        1.135e-03  2.471e-03  0.459  0.6492    
## X5:X6        1.667e-04  1.318e-03  0.126  0.9002    
## ---    
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
##
## Residual standard error: 11.16 on 30 degrees of freedom
## Multiple R-squared:  0.6967, Adjusted R-squared:  0.4844 
## F-statistic: 3.281 on 21 and 30 DF,  p-value: 0.001509

bestmodel3<-lm(Y~X1+X2+X3+X4+X5+X6+X2*X6+X1*X6, data=workhours)
summary(bestmodel3)

##

```

*Surely  
possible  
interactions*

*- marginal  
but let's  
try it.*

```

## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X2 * X6 + X1 *
##      X6, data = workhours)
##
## Residuals:
##   Min     1Q Median     3Q    Max 
## -20.232 -6.195 -0.924  4.847 25.003 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 7.560e+01 2.893e+01 2.613 0.012315 *  
## X1          -8.461e-03 4.249e-03 -1.992 0.052799 .  
## X2           5.103e-01 2.116e-01  2.411 0.020255 *  
## X3           1.676e-02 8.466e-03  1.980 0.054173 .  
## X4           -4.627e-02 1.598e-02 -2.896 0.005922 ** 
## X5           4.031e-02 1.109e-02  3.635 0.000738 *** 
## X6           1.604e-01 4.596e-01  0.349 0.728758    
## X2:X6       -8.283e-03 3.686e-03 -2.247 0.029832 *  
## X1:X6       1.578e-04 6.641e-05  2.376 0.022012 *  
## ---        
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

## Residual standard error: 10.11 on 43 degrees of freedom
## Multiple R-squared:  0.643, Adjusted R-squared:  0.5766 
## F-statistic: 9.681 on 8 and 43 DF,  p-value: 1.492e-07

```

3. And finally, let's see if we can justify any higher-order terms on these. We saw earlier, of course, that  $X2^2$  might be justified. So let's give it a go. Perhaps we ought to leave out  $X1$  and  $X6$ . Have a look at the  $R^2_{adj}$  and see if you can come to your own decision.

```

#Improving model from best subset function
library(GGally) # need to install the GGally package for ggpairs function

workhours=read.csv("CLERICAL.csv",header = TRUE)

#ggpairs(workhours)
#pairs(~Y+X1+X2+X3+X4+X5+X6+X7, data=workhours)

bestmodel31<-lm(Y~X1+X2+I(X2^2)+X3+X4+X5+X6+X1*X6, data=workhours)
summary(bestmodel31)

## 
## Call:
## lm(formula = Y ~ X1 + X2 + I(X2^2) + X3 + X4 + X5 + X6 + X1 *
##      X6, data = workhours)
##
## Residuals:
##   Min     1Q Median     3Q    Max 
## -20.559 -6.328 -0.694  5.042 25.934 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 7.560e+01 2.893e+01 2.613 0.012315 *  
## X1          -8.461e-03 4.249e-03 -1.992 0.052799 .  
## X2           5.103e-01 2.116e-01  2.411 0.020255 *  
## X3           1.676e-02 8.466e-03  1.980 0.054173 .  
## X4           -4.627e-02 1.598e-02 -2.896 0.005922 ** 
## X5           4.031e-02 1.109e-02  3.635 0.000738 *** 
## X6           1.604e-01 4.596e-01  0.349 0.728758    
## X2:X6       -8.283e-03 3.686e-03 -2.247 0.029832 *  
## X1:X6       1.578e-04 6.641e-05  2.376 0.022012 *  
## ---        
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

```

we get a  
 bit of a  
 different  
 model w/  
 all subsets  
 could be down  
 to which  
 we selected

```

## (Intercept) 7.927e+01 2.636e+01 3.007 0.004397 **
## X1          -4.913e-03 4.299e-03 -1.143 0.259355
## X2          4.779e-01 1.645e-01 2.906 0.005766 **
## I(X2^2)    -2.273e-03 8.318e-04 -2.733 0.009074 **
## X3          1.597e-02 8.276e-03 1.930 0.060187 .
## X4          -4.489e-02 1.561e-02 -2.877 0.006228 **
## X5          4.657e-02 1.089e-02 4.278 0.000103 ***
## X6          -3.455e-01 3.466e-01 -0.997 0.324476
## X1:X6     1.022e-04 6.652e-05 1.536 0.131967
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ',' 1
##
## Residual standard error: 9.865 on 43 degrees of freedom
## Multiple R-squared: 0.6601, Adjusted R-squared: 0.5969
## F-statistic: 10.44 on 8 and 43 DF, p-value: 5.581e-08

```