Data 603:Statistical Modelling with Data

Multiple Linear Regression

Part III: Model Selection

Model Selection

One of the biggest problem in building a model to describe a response variable (Y) is choosing the important independent variables to be included. The list of potentially important independent variables is extremely long and we need some objective methods of screening out those which are not important. The problem of deciding which of a large set of independent variables to include in a model is a common one.

For example: Independent Variables in the Executive Salary

Independent Variable and Description

x_1: Experience (years)-quantitative

x_2: Education (years)-quantitative

x_3: Bonus eligibility (1 if yes, 0 if no)-qualitative

x_4: Number of employees supervised-quantitative

x_5: Corporate assets (millions of dollars)-quantitative

x_6: Board member (1 if yes, 0 if no)-qualitative

x_7: Age (years)-quantitative

x_8: Company profits (past 12 months, millions of dollars)-quantitative

x_9: Has international responsibility (1 if yes, 0 if no)-qualitative

x_{10}: Company's total sales (past 12 months, millions of dollars)-quantitative

Steps in Selecting the Best Regression Equation

To select the best regresson equation, carry out the following steps

1. Specify the maximum model to be considered.

2. Specify a strategy for selecting a model

3. Evaluate the reliability of the model chosen.

By following theses steps, you can convert the fuzzy idea of finding the best predictors of Y into simple, concrete action. Each step helps to ensure reliability and to reduce the work required.

Step 1: Specifying the Maximum Model

The maximum model is defined to be the largest model (the one having the most predictor variables) considered at any point in the process of model selection. A model created by deleting predictors from the maximum model is called *a restriction of the maximum model*.

Step 2: Specify a strategy for selecting a model

A systematic approach to building a restriction model from a large number of independent variables is difficult because the interpretation of multivariable interactions is complicated. We therefore turn to a screening procedure, available in most statistical software packages, objectively determine which independent variables in the list are the most important predictors of Y and which are the least important predictors. The most widely used method is **stepwise regression**, while another popular method, **backward** and **forward regression**, also are provided in this section.

Stepwise Regression Procedure

The user first identifies the response y and the set of potentially important independent variables x_1, x_2,...,x_p, where p is generally large. However, we often **include only the main effects** of both quantitative variables (first-order terms) and qualitative variables (dummy variables). The response and independent variables are then entered into the computer software, and the stepwise procedure begins.

Step 1 The software program fits all possible one-variable models of the form

$$E(Y) = \beta_0 + \beta_1 X_i$$

to the data, where X_i is the ith independent variable, $i = 1,2,\ldots,p$. For each model, the t-test for a single β_1 parameter is conducted to test the null hypothesis

H_0: β_1 = 0

against the alternative hypothesis

H_a: $\beta_1 \neq 0$

The independent variable that produces the largest (absolute) t -value is then declared the best one-variable predictor of Y. Call this independent variable X_1.

Step 2 The stepwise program now begins to search through the remaining $(p-1)$ independent variables for the best two-variable model of the form

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_i$$

This is done by fitting all two-variable models containing X_1 and each of the other $(p-1)$ options for the second variable X_i. The t-values for the test $H_0: \beta_2 = 0$ are computed for each of the $p-1$ models (corresponding to the remaining independent variables, $X_i, i = 2,3,\ldots,p-1$), and the variable having the largest t is retained. Call this variable X_2.

Before proceeding to Step 3, the stepwise routine will go back and check the t-value of $\widehat{\beta_1}$ after $\widehat{\beta_2}X_2$ has been added to the model. If the t-value has become nonsignificant at some specified α level (say $\alpha = 0.3$), the variable X_1 is removed and a search is made for the independent variable with a β parameter that will yield the most significant t-value in the presence of $\widehat{\beta_2}X_2$.

The reason the t-value for X_1 may change from step 1 to step 2 is that the meaning of the coefficient $\widehat{\beta_1}$ changes. In step 2, we are approximating a complex response surface in two variables with a plane. The best-fitting plane may yield a different value for $\widehat{\beta_1}$ than that obtained in step 1. Thus, both the value of $\widehat{\beta_1}$ and its significance usually changes from step 1 to step 2. For this reason, stepwise procedures that recheck the t-values at each step are preferred.

Step 3 The stepwise regression procedure now checks for a third independent variable to include in the model with X_1 and X_2. That is, we seek the best model of the form

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_i$$

To do this, the computer fits all the $(p-2)$ models using X_1, X_2, and each of the $(p-2)$ remaining variables, X_i , as a possible X_3. The criterion is again to include the independent variable with the largest (significant) t-value. Call this best third variable X_3. The better programs now recheck the t-values corresponding to the X_1 and X_2 coefficients, replacing the variables that yield nonsignificant t-values.

This procedure is continued until no further independent variables can be found that yield significant t-values (at the specified α level) in the presence of the variables already in the model.

Refer to the Executive Salary Example. A preliminary step in the construction of this model is the determination of the most important independent variables. For one firm, 10 potential independent variables (seven quantitative and three qualitative) were measured in a sample of 100 executives. The data are saved in the **EXECSAL2.CSV** file. Since it would be very difficult to construct a complete first-order model with all of the 10 independent variables, use stepwise regression to decide which of the 10 variables should be included in the building of the final model.

```r
library(olsrr)#need to install the package olsrr

##
## Attaching package: 'olsrr'

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

salary=read.csv("c:/Users/thuntida.ngamkham/OneDrive - University of
Calgary/dataset603/EXECSAL2.csv", header = TRUE)
fullmodel<-lm(Y~X1+X2+factor(X3)+X4+X5+factor(X6)+X7+X8+factor(X9)+X10, data
= salary)
summary(fullmodel)

##
## Call:
## lm(formula = Y ~ X1 + X2 + factor(X3) + X4 + X5 + factor(X6) +
##     X7 + X8 + factor(X9) + X10, data = salary)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.201770 -0.050464  0.004435  0.046826  0.185952
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.002e+01  1.481e-01  67.692  < 2e-16 ***
## X1             2.792e-02  1.773e-03  15.745  < 2e-16 ***
## X2             2.903e-02  3.426e-03   8.475 4.57e-13 ***
## factor(X3)yes  2.243e-01  1.708e-02  13.135  < 2e-16 ***
## X4             5.140e-04  4.922e-05  10.443  < 2e-16 ***
## X5             2.048e-03  5.250e-04   3.901 0.000186 ***
## factor(X6)yes -1.538e-02  1.686e-02  -0.912 0.364124
## X7            -5.097e-04  1.438e-03  -0.355 0.723795
## X8            -2.633e-03  5.128e-03  -0.513 0.608896
## factor(X9)yes -2.656e-02  2.037e-02  -1.304 0.195613
## X10           -9.774e-04  2.959e-03  -0.330 0.741955
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07608 on 89 degrees of freedom
## Multiple R-squared:  0.9229, Adjusted R-squared:  0.9142
## F-statistic: 106.5 on 10 and 89 DF,  p-value: < 2.2e-16

stepw=ols_step_both_p(fullmodel,pent = 0.1, prem = 0.3, details=TRUE)

## Stepwise Selection Method
## ---------------------------
##
## Candidate Terms:
##
```

```
## 1. X1
## 2. X2
## 3. factor(X3)
## 4. X4
## 5. X5
## 6. factor(X6)
## 7. X7
## 8. X8
## 9. factor(X9)
## 10. X10
##
## We are selecting variables based on p value...
##
##
## Stepwise Selection: Step 1
##
## - X1 added
##
##                         Model Summary
## -------------------------------------------------------------
## R                         0.787       RMSE                0.161
## R-Squared                 0.619       Coef. Var           1.407
## Adj. R-Squared            0.615       MSE                 0.026
## Pred R-Squared            0.601       MAE                 0.122
## -------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
## --------------------------------------------------------------------
##                 Sum of
##                 Squares        DF    Mean Square       F         Sig.
## --------------------------------------------------------------------
## Regression      4.136          1           4.136    159.204    0.0000
## Residual        2.546         98           0.026
## Total           6.683         99
## --------------------------------------------------------------------
##
##                                    Parameter Estimates
## --------------------------------------------------------------------------
---------------
##       model      Beta    Std. Error    Std. Beta       t        Sig
lower     upper
## --------------------------------------------------------------------------
---------------
## (Intercept)    11.091         0.033                 335.524    0.000
11.025    11.156
##          X1     0.028         0.002        0.787     12.618    0.000
0.023     0.032
```

```
## -----------------------------------------------------------------------------
---------------
##
##
##
## Stepwise Selection: Step 2
##
## - factor(X3) added
##
##                         Model Summary
## -------------------------------------------------------------
## R                       0.866       RMSE              0.131
## R-Squared               0.749       Coef. Var         1.147
## Adj. R-Squared          0.744       MSE               0.017
## Pred R-Squared          0.732       MAE               0.104
## -------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
## --------------------------------------------------------------------
##                Sum of
##               Squares        DF    Mean Square       F         Sig.
## --------------------------------------------------------------------
## Regression      5.007         2          2.503    144.887    0.0000
## Residual        1.676        97          0.017
## Total           6.683        99
## --------------------------------------------------------------------
##
##                                    Parameter Estimates
## -----------------------------------------------------------------------------
-----------------
##         model      Beta    Std. Error    Std. Beta       t        Sig
lower     upper
## -----------------------------------------------------------------------------
-----------------
##   (Intercept)    10.968         0.032                 342.659    0.000
10.905    11.032
##            X1     0.027         0.002        0.770     15.134    0.000
0.024     0.031
## factor(X3)yes     0.197         0.028        0.361      7.097    0.000
0.142     0.252
## -----------------------------------------------------------------------------
-----------------
##
##
##
##                         Model Summary
## -------------------------------------------------------------
```

```
## R                          0.866       RMSE                0.131
## R-Squared                  0.749       Coef. Var           1.147
## Adj. R-Squared             0.744       MSE                 0.017
## Pred R-Squared             0.732       MAE                 0.104
## -----------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
## ------------------------------------------------------------------
##               Sum of
##              Squares        DF    Mean Square       F         Sig.
## ------------------------------------------------------------------
## Regression      5.007         2          2.503    144.887    0.0000
## Residual        1.676        97          0.017
## Total           6.683        99
## ------------------------------------------------------------------
##
##                                     Parameter Estimates
## ----------------------------------------------------------------------------
-----------------
##         model      Beta    Std. Error    Std. Beta       t        Sig
lower     upper
## ----------------------------------------------------------------------------
-----------------
##   (Intercept)    10.968         0.032                342.659    0.000
10.905    11.032
##            X1     0.027         0.002        0.770     15.134    0.000
0.024     0.031
## factor(X3)yes     0.197         0.028        0.361      7.097    0.000
0.142     0.252
## ----------------------------------------------------------------------------
-----------------
##
##
##
## Stepwise Selection: Step 3
##
## - X4 added
##
##                          Model Summary
## -----------------------------------------------------------
## R                          0.916       RMSE                0.106
## R-Squared                  0.839       Coef. Var           0.924
## Adj. R-Squared             0.834       MSE                 0.011
## Pred R-Squared             0.825       MAE                 0.082
## -----------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
```

```
##  MAE: Mean Absolute Error
##
##                                ANOVA
## --------------------------------------------------------------------
##                Sum of
##               Squares        DF    Mean Square       F         Sig.
## --------------------------------------------------------------------
## Regression      5.607         3          1.869    166.873    0.0000
## Residual        1.075        96          0.011
## Total           6.683        99
## --------------------------------------------------------------------
##
##                                      Parameter Estimates
## ------------------------------------------------------------------------------
-----------------
##         model      Beta    Std. Error    Std. Beta       t        Sig
lower     upper
## ------------------------------------------------------------------------------
-----------------
##   (Intercept)    10.783         0.036                298.170    0.000
10.711    10.854
##            X1     0.027         0.001        0.771     18.801    0.000
0.024     0.030
## factor(X3)yes     0.233         0.023        0.427     10.170    0.000
0.187     0.278
##            X4     0.000         0.000        0.307      7.323    0.000
0.000     0.001
## ------------------------------------------------------------------------------
-----------------
##
##
##
##                         Model Summary
## -------------------------------------------------------------
## R                         0.916       RMSE              0.106
## R-Squared                 0.839       Coef. Var         0.924
## Adj. R-Squared            0.834       MSE               0.011
## Pred R-Squared            0.825       MAE               0.082
## -------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
## --------------------------------------------------------------------
##                Sum of
##               Squares        DF    Mean Square       F         Sig.
## --------------------------------------------------------------------
## Regression      5.607         3          1.869    166.873    0.0000
## Residual        1.075        96          0.011
```

```
## Total           6.683        99
## ----------------------------------------------------------------------
##
##                                     Parameter Estimates
## ----------------------------------------------------------------------------
-----------------
##          model      Beta    Std. Error    Std. Beta       t        Sig
lower     upper
## ----------------------------------------------------------------------------
-----------------
##   (Intercept)    10.783         0.036                 298.170    0.000
10.711    10.854
##            X1     0.027         0.001        0.771     18.801    0.000
0.024     0.030
## factor(X3)yes     0.233         0.023        0.427     10.170    0.000
0.187     0.278
##            X4     0.000         0.000        0.307      7.323    0.000
0.000     0.001
## ----------------------------------------------------------------------------
-----------------
##
##
##
## Stepwise Selection: Step 4
##
## - X2 added
##
##                         Model Summary
## ---------------------------------------------------------------
## R                       0.953       RMSE               0.081
## R-Squared               0.907       Coef. Var          0.704
## Adj. R-Squared          0.904       MSE                0.007
## Pred R-Squared          0.896       MAE                0.062
## ---------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                              ANOVA
## ----------------------------------------------------------------------
##                Sum of
##               Squares        DF    Mean Square       F         Sig.
## ----------------------------------------------------------------------
## Regression      6.064         4          1.516    232.936    0.0000
## Residual        0.618        95          0.007
## Total           6.683        99
## ----------------------------------------------------------------------
##
##                                     Parameter Estimates
## ----------------------------------------------------------------------------
```

```
-----------------
##         model      Beta    Std. Error    Std. Beta       t        Sig
lower     upper
## ---------------------------------------------------------------------------
-----------------
##   (Intercept)    10.278         0.066                155.154    0.000
10.146    10.409
##            X1     0.027         0.001        0.771     24.677    0.000
0.025     0.029
## factor(X3)yes     0.232         0.017        0.425     13.297    0.000
0.197     0.267
##            X4     0.001         0.000        0.354     10.920    0.000
0.000     0.001
##            X2     0.030         0.004        0.266      8.379    0.000
0.023     0.037
## ---------------------------------------------------------------------------
-----------------
##
##
##
##                        Model Summary
## ---------------------------------------------------------------
## R                       0.953       RMSE                0.081
## R-Squared               0.907       Coef. Var           0.704
## Adj. R-Squared          0.904       MSE                 0.007
## Pred R-Squared          0.896       MAE                 0.062
## ---------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                               ANOVA
## ---------------------------------------------------------------------
##                Sum of
##               Squares        DF    Mean Square       F         Sig.
## ---------------------------------------------------------------------
## Regression      6.064         4          1.516    232.936    0.0000
## Residual        0.618        95          0.007
## Total           6.683        99
## ---------------------------------------------------------------------
##
##                                    Parameter Estimates
## ---------------------------------------------------------------------------
-----------------
##         model      Beta    Std. Error    Std. Beta       t        Sig
lower     upper
## ---------------------------------------------------------------------------
-----------------
##   (Intercept)    10.278         0.066                155.154    0.000
10.146    10.409
```

```
##            X1     0.027         0.001        0.771     24.677    0.000
0.025     0.029
## factor(X3)yes     0.232         0.017        0.425     13.297    0.000
0.197     0.267
##            X4     0.001         0.000        0.354     10.920    0.000
0.000     0.001
##            X2     0.030         0.004        0.266      8.379    0.000
0.023     0.037
## ----------------------------------------------------------------------------
-----------------
##
##
##
## Stepwise Selection: Step 5
##
## - X5 added
##
##                         Model Summary
## --------------------------------------------------------------
## R                       0.959       RMSE               0.075
## R-Squared               0.921       Coef. Var          0.656
## Adj. R-Squared          0.916       MSE                0.006
## Pred R-Squared          0.909       MAE                0.059
## --------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
## ---------------------------------------------------------------------
##                Sum of
##               Squares        DF    Mean Square       F         Sig.
## ---------------------------------------------------------------------
## Regression      6.152         5          1.230    218.061    0.0000
## Residual        0.530        94          0.006
## Total           6.683        99
## ---------------------------------------------------------------------
##
##                                   Parameter Estimates
## ----------------------------------------------------------------------------
--------------
##         model     Beta    Std. Error    Std. Beta      t        Sig
lower     upper
## ----------------------------------------------------------------------------
--------------
##   (Intercept)    9.962         0.101                 98.578    0.000
9.761    10.163
##            X1    0.027         0.001        0.771    26.501    0.000
0.025     0.029
## factor(X3)yes    0.225         0.016        0.412    13.742    0.000
```

```
0.192     0.257
##            X4    0.001         0.000        0.337    11.064    0.000
0.000     0.001
##            X2    0.029         0.003        0.258     8.719    0.000
0.022     0.036
##            X5    0.002         0.000        0.116     3.947    0.000
0.001     0.003
## ----------------------------------------------------------------------------
--------------
##
##
##
##                           Model Summary
## ---------------------------------------------------------------
## R                         0.959       RMSE                0.075
## R-Squared                 0.921       Coef. Var           0.656
## Adj. R-Squared            0.916       MSE                 0.006
## Pred R-Squared            0.909       MAE                 0.059
## ---------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                  ANOVA
## ----------------------------------------------------------------------
##                Sum of
##               Squares        DF    Mean Square       F         Sig.
## ----------------------------------------------------------------------
## Regression      6.152         5          1.230    218.061    0.0000
## Residual        0.530        94          0.006
## Total           6.683        99
## ----------------------------------------------------------------------
##
##                                     Parameter Estimates
## ----------------------------------------------------------------------------
--------------
##         model     Beta    Std. Error    Std. Beta      t        Sig
lower     upper
## ----------------------------------------------------------------------------
--------------
##   (Intercept)    9.962         0.101                 98.578    0.000
9.761    10.163
##            X1    0.027         0.001        0.771    26.501    0.000
0.025     0.029
## factor(X3)yes    0.225         0.016        0.412    13.742    0.000
0.192     0.257
##            X4    0.001         0.000        0.337    11.064    0.000
0.000     0.001
##            X2    0.029         0.003        0.258     8.719    0.000
0.022     0.036
```

```
##            X5    0.002         0.000        0.116     3.947    0.000
0.001     0.003
## ----------------------------------------------------------------------------
--------------
##
##
##
## No more variables to be added/removed.
##
##
## Final Model Output
## ------------------
##
##                         Model Summary
## -------------------------------------------------------------
## R                       0.959       RMSE              0.075
## R-Squared               0.921       Coef. Var         0.656
## Adj. R-Squared          0.916       MSE               0.006
## Pred R-Squared          0.909       MAE               0.059
## -------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
## --------------------------------------------------------------------
##                Sum of
##               Squares        DF    Mean Square       F         Sig.
## --------------------------------------------------------------------
## Regression      6.152         5          1.230    218.061    0.0000
## Residual        0.530        94          0.006
## Total           6.683        99
## --------------------------------------------------------------------
##
##                                   Parameter Estimates
## ----------------------------------------------------------------------------
--------------
##         model     Beta    Std. Error    Std. Beta      t        Sig
lower     upper
## ----------------------------------------------------------------------------
--------------
##   (Intercept)    9.962         0.101                  98.578    0.000
9.761    10.163
##            X1    0.027         0.001        0.771    26.501    0.000
0.025     0.029
## factor(X3)yes    0.225         0.016        0.412    13.742    0.000
0.192     0.257
##            X4    0.001         0.000        0.337    11.064    0.000
0.000     0.001
##            X2    0.029         0.003        0.258     8.719    0.000
```

```
0.022     0.036
##            X5    0.002         0.000        0.116     3.947    0.000
0.001     0.003
## ----------------------------------------------------------------------------
--------------
```

```r
summary(stepw$model)
```

```
##
## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##     data = l)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.201219 -0.056016 -0.003581  0.053656  0.187251
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.9619345  0.1010567  98.578  < 2e-16 ***
## X1            0.0272762  0.0010293  26.501  < 2e-16 ***
## factor(X3)yes 0.2246932  0.0163503  13.742  < 2e-16 ***
## X4            0.0005244  0.0000474  11.064  < 2e-16 ***
## X2            0.0290921  0.0033367   8.719 9.71e-14 ***
## X5            0.0019623  0.0004972   3.947 0.000153 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07512 on 94 degrees of freedom
## Multiple R-squared:  0.9206, Adjusted R-squared:  0.9164
## F-statistic: 218.1 on 5 and 94 DF,  p-value: < 2.2e-16
```

R functions ols_step_both_p(): Build regression model from a set of candidate predictor variables by entering and removing predictors based on p values

Note!

pent: variables with p value less than pent will enter into the model.

prem: variables with p value more than prem will be removed from the model.

details: print the regression result at each step.

From the output, the regression model is $Y = X_1 + X_2 + X_3 + X_4 + X_5 + \epsilon$. Is this model the best fit for predicting executive salary?

Inclass Practice Problem From the credit example in MLR Modelling Part 2, use **Stepwise Regression Procedure** to find the potentially important independent variables for predicting credit card balance.

Backward Elimination Procedure

The Backward procedure initially fits a model containing terms for all potential independent variables. That is, for p independent variables, the model $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$ is fit in step 1. The variable with the smallest t (or F) statistic for testing $H_0: \beta_i = 0$ is identified and dropped from the model if the t-value is less than some specified critical value or p-value more than a cut-off. The model with the remaining $(p - 1)$ independent variables is fit in step 2, and again, the variable associated with the smallest nonsignificant t-value is dropped. This process is repeated until no further nonsignificant independent variables can be found.

```r
library(olsrr) #need to install the package olsrr
salary=read.csv("c:/Users/thuntida.ngamkham/OneDrive - University of Calgary/dataset603/EXECSAL2.csv", header = TRUE)
fullmodel<-lm(Y~X1+X2+factor(X3)+X4+X5+factor(X6)+X7+X8+factor(X9)+X10, data = salary)
backmodel=ols_step_backward_p(fullmodel, prem = 0.3, details=TRUE)
```

```
## Backward Elimination Method
## ---------------------------
##
## Candidate Terms:
##
## 1 . X1
## 2 . X2
## 3 . factor(X3)
## 4 . X4
## 5 . X5
## 6 . factor(X6)
## 7 . X7
## 8 . X8
## 9 . factor(X9)
## 10 . X10
##
## We are eliminating variables based on p value...
##
## - X10
##
## Backward Elimination: Step 1
##
##  Variable X10 Removed
##
##                         Model Summary
## --------------------------------------------------------------
## R                       0.961       RMSE                0.076
## R-Squared               0.923       Coef. Var           0.661
## Adj. R-Squared          0.915       MSE                 0.006
## Pred R-Squared          0.904       MAE                 0.058
## --------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
```

```
##  MAE: Mean Absolute Error
##
##                                 ANOVA
## ---------------------------------------------------------------------
##                Sum of
##               Squares        DF    Mean Square       F         Sig.
## ---------------------------------------------------------------------
## Regression      6.167         9          0.685    119.551    0.0000
## Residual        0.516        90          0.006
## Total           6.683        99
## ---------------------------------------------------------------------
##
##                                     Parameter Estimates
## ---------------------------------------------------------------------------
----------------
##         model      Beta    Std. Error    Std. Beta      t        Sig
lower     upper
## ---------------------------------------------------------------------------
----------------
##   (Intercept)     9.995         0.123                81.304    0.000
9.751    10.239
##            X1     0.028         0.002        0.785    16.329    0.000
0.024     0.031
##            X2     0.029         0.003        0.258     8.519    0.000
0.022     0.036
## factor(X3)yes     0.225         0.017        0.413    13.430    0.000
0.192     0.259
##            X4     0.001         0.000        0.332    10.557    0.000
0.000     0.001
##            X5     0.002         0.001        0.121     3.911    0.000
0.001     0.003
## factor(X6)yes    -0.015         0.017       -0.028    -0.884    0.379    -
0.048     0.018
##            X7     0.000         0.001       -0.014    -0.296    0.768    -
0.003     0.002
##            X8    -0.003         0.005       -0.016    -0.509    0.612    -
0.013     0.008
## factor(X9)yes    -0.027         0.020       -0.040    -1.316    0.192    -
0.067     0.014
## ---------------------------------------------------------------------------
----------------
##
##
## - X7
##
## Backward Elimination: Step 2
##
##  Variable X7 Removed
##
##                           Model Summary
```

```
## -----------------------------------------------------------------
## R                          0.961       RMSE              0.075
## R-Squared                  0.923       Coef. Var         0.658
## Adj. R-Squared             0.916       MSE               0.006
## Pred R-Squared             0.906       MAE               0.058
## -----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
## --------------------------------------------------------------------
##               Sum of
##              Squares        DF    Mean Square       F         Sig.
## --------------------------------------------------------------------
## Regression      6.166         8          0.771    135.846    0.0000
## Residual        0.516        91          0.006
## Total           6.683        99
## --------------------------------------------------------------------
##
##                                    Parameter Estimates
## ----------------------------------------------------------------------------
----------------
##         model      Beta    Std. Error    Std. Beta      t        Sig
lower     upper
## ----------------------------------------------------------------------------
----------------
##   (Intercept)     9.978         0.108                 92.466    0.000
9.764    10.192
##            X1     0.027         0.001        0.773    26.473    0.000
0.025     0.029
##            X2     0.029         0.003        0.259     8.648    0.000
0.022     0.036
## factor(X3)yes     0.225         0.017        0.411    13.605    0.000
0.192     0.257
##            X4     0.001         0.000        0.331    10.607    0.000
0.000     0.001
##            X5     0.002         0.001        0.122     3.978    0.000
0.001     0.003
## factor(X6)yes    -0.013         0.016       -0.026    -0.839    0.404    -
0.045     0.018
##            X8    -0.003         0.005       -0.015    -0.509    0.612    -
0.013     0.007
## factor(X9)yes    -0.026         0.020       -0.039    -1.302    0.196    -
0.066     0.014
## ----------------------------------------------------------------------------
----------------
##
##
## - X8
```

```
##
## Backward Elimination: Step 3
##
##  Variable X8 Removed
##
##                         Model Summary
## ---------------------------------------------------------------
## R                         0.960       RMSE               0.075
## R-Squared                 0.923       Coef. Var          0.655
## Adj. R-Squared            0.917       MSE                0.006
## Pred R-Squared            0.907       MAE                0.058
## ---------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
## ----------------------------------------------------------------------
##                 Sum of
##                Squares        DF    Mean Square       F         Sig.
## ----------------------------------------------------------------------
## Regression      6.165         7          0.881    156.475    0.0000
## Residual        0.518        92          0.006
## Total           6.683        99
## ----------------------------------------------------------------------
##
##                                    Parameter Estimates
## ----------------------------------------------------------------------------
----------------
##         model      Beta    Std. Error    Std. Beta      t        Sig
lower     upper
## ----------------------------------------------------------------------------
----------------
##   (Intercept)     9.966         0.105                  94.885    0.000
9.758    10.175
##            X1     0.027         0.001        0.773    26.575    0.000
0.025     0.029
##            X2     0.029         0.003        0.258     8.669    0.000
0.022     0.036
## factor(X3)yes     0.224         0.016        0.411    13.652    0.000
0.192     0.257
##            X4     0.001         0.000        0.332    10.680    0.000
0.000     0.001
##            X5     0.002         0.001        0.119     3.966    0.000
0.001     0.003
## factor(X6)yes    -0.012         0.016       -0.023    -0.768    0.444    -
0.043     0.019
## factor(X9)yes    -0.025         0.020       -0.037    -1.254    0.213    -
0.064     0.015
## ----------------------------------------------------------------------------
```

```
----------------
##
##
## - factor(X6)
##
## Backward Elimination: Step 4
##
##  Variable factor(X6) Removed
##
##                           Model Summary
## --------------------------------------------------------------
## R                         0.960       RMSE               0.075
## R-Squared                 0.922       Coef. Var          0.653
## Adj. R-Squared            0.917       MSE                0.006
## Pred R-Squared            0.909       MAE                0.058
## --------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                  ANOVA
## ---------------------------------------------------------------------
##                 Sum of
##                Squares        DF    Mean Square       F         Sig.
## ---------------------------------------------------------------------
## Regression      6.162          6          1.027    183.264    0.0000
## Residual        0.521         93          0.006
## Total           6.683         99
## ---------------------------------------------------------------------
##
##                                      Parameter Estimates
## ---------------------------------------------------------------------------
----------------
##         model      Beta    Std. Error    Std. Beta      t        Sig
lower     upper
## ---------------------------------------------------------------------------
----------------
##   (Intercept)     9.946         0.101                  98.028    0.000
9.745    10.147
##            X1     0.027         0.001        0.772    26.623    0.000
0.025     0.029
##            X2     0.029         0.003        0.260     8.807    0.000
0.023     0.036
## factor(X3)yes     0.223         0.016        0.409    13.667    0.000
0.191     0.256
##            X4     0.001         0.000        0.337    11.071    0.000
0.000     0.001
##            X5     0.002         0.001        0.122     4.112    0.000
0.001     0.003
## factor(X9)yes    -0.025         0.020       -0.038    -1.287    0.201    -
```

```
0.065     0.014
## ---------------------------------------------------------------------------
----------------
##
##
##
## No more variables satisfy the condition of p value = 0.3
##
##
## Variables Removed:
##
## - X10
## - X7
## - X8
## - factor(X6)
##
##
## Final Model Output
## ------------------
##
##                         Model Summary
## -------------------------------------------------------------
## R                       0.960       RMSE                0.075
## R-Squared               0.922       Coef. Var           0.653
## Adj. R-Squared          0.917       MSE                 0.006
## Pred R-Squared          0.909       MAE                 0.058
## -------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
## --------------------------------------------------------------------
##                Sum of
##               Squares        DF    Mean Square       F          Sig.
## --------------------------------------------------------------------
## Regression      6.162         6          1.027    183.264    0.0000
## Residual        0.521        93          0.006
## Total           6.683        99
## --------------------------------------------------------------------
##
##                                    Parameter Estimates
## ---------------------------------------------------------------------------
----------------
##         model      Beta    Std. Error    Std. Beta      t        Sig
lower     upper
## ---------------------------------------------------------------------------
----------------
##   (Intercept)     9.946         0.101                 98.028    0.000
9.745    10.147
```

```
##             X1     0.027          0.001        0.772    26.623    0.000
0.025     0.029
##             X2     0.029          0.003        0.260     8.807    0.000
0.023     0.036
## factor(X3)yes     0.223          0.016        0.409    13.667    0.000
0.191     0.256
##             X4     0.001          0.000        0.337    11.071    0.000
0.000     0.001
##             X5     0.002          0.001        0.122     4.112    0.000
0.001     0.003
## factor(X9)yes    -0.025          0.020       -0.038    -1.287    0.201    -
0.065     0.014
## ------------------------------------------------------------------------------
----------------
```

```
summary(backmodel$model)
```

```
##
## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##     data = l)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.20278 -0.05332 -0.00050  0.05115  0.18286
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.946e+00  1.015e-01  98.028  < 2e-16 ***
## X1             2.733e-02  1.027e-03  26.623  < 2e-16 ***
## X2             2.933e-02  3.330e-03   8.807 6.82e-14 ***
## factor(X3)yes  2.232e-01  1.633e-02  13.667  < 2e-16 ***
## X4             5.230e-04  4.724e-05  11.071  < 2e-16 ***
## X5             2.062e-03  5.014e-04   4.112 8.46e-05 ***
## factor(X9)yes -2.549e-02  1.980e-02  -1.287    0.201
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07486 on 93 degrees of freedom
## Multiple R-squared:  0.922,  Adjusted R-squared:  0.917
## F-statistic: 183.3 on 6 and 93 DF,  p-value: < 2.2e-16
```

R functions ols_step_backward_p():Build regression model from a set of candidate predictor variables by removing predictors based on p values

From the output, the first order regression model by using Backward Regression Procedure is $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_2 X_3 + \beta_3 X_4 + \beta_4 X_5 + \beta_5 X_9 + \epsilon$. Consider the predictor X9 has tcal=-1.287 with the p-value= 0.201, this predictor should be dropped out from the output. Therefore, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_2 X_3 + \beta_3 X_4 + \beta_4 X_5 + \epsilon$ is the first order model to predict salary by using For Backward Regression Procedure.

Inclass Practice Problem From the credit example in MLR Modelling Part 2, use **Backward Regression Procedure** to find the potentially important independent variables for predicting credit card balance.

Forward selection procedure This method is nearly identical to the stepwise procedure previously outlined. The only difference is that the forward selection technique provides no option for rechecking the t-values corresponding to the X's that have entered the model in an earlier step.

```
library(olsrr) #need to install the package olsrr
salary=read.csv("c:/Users/thuntida.ngamkham/OneDrive - University of
Calgary/dataset603/EXECSAL2.csv", header = TRUE)
fullmodel<-lm(Y~X1+X2+factor(X3)+X4+X5+factor(X6)+X7+X8+factor(X9)+X10, data
= salary)
formodel=ols_step_forward_p(fullmodel,penter = 0.1, details=TRUE)
```

```
## Forward Selection Method
## ---------------------------
##
## Candidate Terms:
##
## 1. X1
## 2. X2
## 3. factor(X3)
## 4. X4
## 5. X5
## 6. factor(X6)
## 7. X7
## 8. X8
## 9. factor(X9)
## 10. X10
##
## We are selecting variables based on p value...
##
```

```
##
## Forward Selection: Step 1
##
## - X1
##
##                           Model Summary
## ---------------------------------------------------------------
## R                         0.787       RMSE               0.161
## R-Squared                 0.619       Coef. Var          1.407
## Adj. R-Squared            0.615       MSE                0.026
## Pred R-Squared            0.601       MAE                0.122
## ---------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                  ANOVA
## ----------------------------------------------------------------------
##                 Sum of
##                Squares        DF    Mean Square       F         Sig.
## ----------------------------------------------------------------------
## Regression      4.136          1          4.136    159.204    0.0000
## Residual        2.546         98          0.026
## Total           6.683         99
## ----------------------------------------------------------------------
##
##                                      Parameter Estimates
## ----------------------------------------------------------------------------
---------------
##       model      Beta    Std. Error    Std. Beta       t        Sig
lower     upper
## ----------------------------------------------------------------------------
---------------
## (Intercept)    11.091         0.033                 335.524    0.000
11.025    11.156
##          X1     0.028         0.002        0.787     12.618    0.000
0.023     0.032
## ----------------------------------------------------------------------------
---------------
##
##
##
## Forward Selection: Step 2
##
## - factor(X3)
##
##                           Model Summary
## ---------------------------------------------------------------
## R                         0.866       RMSE               0.131
## R-Squared                 0.749       Coef. Var          1.147
```

```
## Adj. R-Squared          0.744       MSE                0.017
## Pred R-Squared          0.732       MAE                0.104
## -------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                               ANOVA
## --------------------------------------------------------------------
##                Sum of
##               Squares        DF    Mean Square       F         Sig.
## --------------------------------------------------------------------
## Regression      5.007         2          2.503    144.887    0.0000
## Residual        1.676        97          0.017
## Total           6.683        99
## --------------------------------------------------------------------
##
##                                    Parameter Estimates
## --------------------------------------------------------------------------
-----------------
##         model      Beta    Std. Error    Std. Beta       t        Sig
lower     upper
## --------------------------------------------------------------------------
-----------------
##   (Intercept)    10.968         0.032                 342.659    0.000
10.905    11.032
##            X1     0.027         0.002        0.770     15.134    0.000
0.024     0.031
## factor(X3)yes     0.197         0.028        0.361      7.097    0.000
0.142     0.252
## --------------------------------------------------------------------------
-----------------
##
##
##
## Forward Selection: Step 3
##
## - X4
##
##                         Model Summary
## -------------------------------------------------------------
## R                       0.916       RMSE               0.106
## R-Squared               0.839       Coef. Var          0.924
## Adj. R-Squared          0.834       MSE                0.011
## Pred R-Squared          0.825       MAE                0.082
## -------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
```

```
##                                ANOVA
## --------------------------------------------------------------------
##                Sum of
##               Squares        DF    Mean Square       F         Sig.
## --------------------------------------------------------------------
## Regression      5.607         3          1.869    166.873    0.0000
## Residual        1.075        96          0.011
## Total           6.683        99
## --------------------------------------------------------------------
##
##                                     Parameter Estimates
## ----------------------------------------------------------------------------
-----------------
##         model      Beta    Std. Error    Std. Beta       t        Sig
lower     upper
## ----------------------------------------------------------------------------
-----------------
##   (Intercept)    10.783         0.036                 298.170    0.000
10.711    10.854
##            X1     0.027         0.001        0.771     18.801    0.000
0.024     0.030
## factor(X3)yes     0.233         0.023        0.427     10.170    0.000
0.187     0.278
##            X4     0.000         0.000        0.307      7.323    0.000
0.000     0.001
## ----------------------------------------------------------------------------
-----------------
##
##
##
## Forward Selection: Step 4
##
## - X2
##
##                         Model Summary
## -------------------------------------------------------------
## R                       0.953       RMSE                0.081
## R-Squared               0.907       Coef. Var           0.704
## Adj. R-Squared          0.904       MSE                 0.007
## Pred R-Squared          0.896       MAE                 0.062
## -------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
## --------------------------------------------------------------------
##                Sum of
##               Squares        DF    Mean Square       F         Sig.
## --------------------------------------------------------------------
```

```
## Regression      6.064         4          1.516    232.936    0.0000
## Residual        0.618        95          0.007
## Total           6.683        99
## -------------------------------------------------------------------
##
##                                    Parameter Estimates
## -------------------------------------------------------------------------
-----------------
##         model      Beta    Std. Error    Std. Beta       t        Sig
lower     upper
## -------------------------------------------------------------------------
-----------------
##   (Intercept)    10.278         0.066                155.154    0.000
10.146    10.409
##            X1     0.027         0.001        0.771     24.677    0.000
0.025     0.029
## factor(X3)yes     0.232         0.017        0.425     13.297    0.000
0.197     0.267
##            X4     0.001         0.000        0.354     10.920    0.000
0.000     0.001
##            X2     0.030         0.004        0.266      8.379    0.000
0.023     0.037
## -------------------------------------------------------------------------
-----------------
##
##
##
## Forward Selection: Step 5
##
## - X5
##
##                         Model Summary
## --------------------------------------------------------------
## R                       0.959       RMSE              0.075
## R-Squared               0.921       Coef. Var         0.656
## Adj. R-Squared          0.916       MSE               0.006
## Pred R-Squared          0.909       MAE               0.059
## --------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
## -------------------------------------------------------------------
##               Sum of
##              Squares        DF    Mean Square       F         Sig.
## -------------------------------------------------------------------
## Regression      6.152         5          1.230    218.061    0.0000
## Residual        0.530        94          0.006
## Total           6.683        99
```

```
## ---------------------------------------------------------------------
##
##                                    Parameter Estimates
## -----------------------------------------------------------------------------
--------------
##          model     Beta    Std. Error    Std. Beta      t        Sig
lower     upper
## -----------------------------------------------------------------------------
--------------
##   (Intercept)    9.962         0.101                98.578    0.000
9.761    10.163
##            X1    0.027         0.001        0.771    26.501    0.000
0.025     0.029
## factor(X3)yes    0.225         0.016        0.412    13.742    0.000
0.192     0.257
##            X4    0.001         0.000        0.337    11.064    0.000
0.000     0.001
##            X2    0.029         0.003        0.258     8.719    0.000
0.022     0.036
##            X5    0.002         0.000        0.116     3.947    0.000
0.001     0.003
## -----------------------------------------------------------------------------
--------------
##
##
##
## No more variables to be added.
##
## Variables Entered:
##
## + X1
## + factor(X3)
## + X4
## + X2
## + X5
##
##
## Final Model Output
## ------------------
##
##                          Model Summary
## --------------------------------------------------------------
## R                       0.959       RMSE               0.075
## R-Squared               0.921       Coef. Var          0.656
## Adj. R-Squared          0.916       MSE                0.006
## Pred R-Squared          0.909       MAE                0.059
## --------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
```

```
##
##                                ANOVA
## --------------------------------------------------------------------
##                Sum of
##               Squares        DF    Mean Square       F          Sig.
## --------------------------------------------------------------------
## Regression      6.152         5          1.230    218.061    0.0000
## Residual        0.530        94          0.006
## Total           6.683        99
## --------------------------------------------------------------------
##
##                                   Parameter Estimates
## --------------------------------------------------------------------------
--------------
##          model     Beta    Std. Error    Std. Beta      t        Sig
lower     upper
## --------------------------------------------------------------------------
--------------
##    (Intercept)    9.962         0.101                98.578    0.000
9.761    10.163
##             X1    0.027         0.001        0.771    26.501    0.000
0.025     0.029
## factor(X3)yes    0.225         0.016        0.412    13.742    0.000
0.192     0.257
##             X4    0.001         0.000        0.337    11.064    0.000
0.000     0.001
##             X2    0.029         0.003        0.258     8.719    0.000
0.022     0.036
##             X5    0.002         0.000        0.116     3.947    0.000
0.001     0.003
## --------------------------------------------------------------------------
--------------
```

```r
summary(formodel$model)
```

```
##
## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##     data = l)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.201219 -0.056016 -0.003581  0.053656  0.187251
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.9619345  0.1010567  98.578  < 2e-16 ***
## X1             0.0272762  0.0010293  26.501  < 2e-16 ***
## factor(X3)yes 0.2246932  0.0163503  13.742  < 2e-16 ***
## X4             0.0005244  0.0000474  11.064  < 2e-16 ***
```

```
## X2           0.0290921  0.0033367   8.719 9.71e-14 ***
## X5           0.0019623  0.0004972   3.947 0.000153 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07512 on 94 degrees of freedom
## Multiple R-squared:  0.9206, Adjusted R-squared:  0.9164
## F-statistic: 218.1 on 5 and 94 DF,  p-value: < 2.2e-16
```

R functions ols_step_forward_p():Build regression model from a set of candidate predictor variables by entering predictors based on p values penter: p value; variables with p value less than penter will enter into the model. By default, penter=0.3

From the output, we specified our penter = 0.1 to follow the same procedure of Stepwise regression. Therefore, the regression model by using Forward Regression Procedure is $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon$.

Inclass Practice Problem From the credit example in MLR Modelling Part 2, use **Forward Regression Procedure** to find the potentially important independent variables for predicting credit card balance.

Note!

R also provides a function for selecting a subset of predictors from a larger set. You can use stepwise selection (backward,forward,both) by using the stepAIC() function from the MASS package. This function will select variable by extracting AIC (AIC value is explained in the next topic).

CAUTION!

Be wary of using the results of stepwise regression to make inferences about the relationship between $E(Y)$ and the independent variables in the first order model.

First, an extremely large number of t-tests have been conducted, leading to a high probability of making more Type I errors.

Second, stepwise regression should be used only when necessary- that is when you want to determine which of a large number of potentially important independent variables should be used in the model building process.

All-Possible-Regressions Selection Procedure

We presented stepwise regression as an objective screening procedure. Stepwise does not only provide the largest t-value, but also the techniques differ with respect to the criteria for selecting the ''best'' subset of variables. In this section, we describe four criteria widely used in practice,

1. R^2 **Criterion** the multiple coefficient of determination

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

will increase when independent variables are added to the model. Therefore, the model that includes all p independent variables $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$ will yield the largest R^2.

2. Adjusted R^2 **or RMSE Criterion**

We can use the adjusted R^2 instead of R^2. It is easy to show that R^2_{adj} is related to MSE as follows:

$$\begin{aligned} R^2_{adj} &= 1 - \frac{\frac{SSE}{n-p-1}}{\frac{SST}{n-1}} \\ R^2_{adj} &= 1 - (n-1)\frac{MSE}{SST} \\ s &= RMSE = \sqrt{\frac{1}{n-p-1}SSE} \end{aligned}$$

Note that R^2_{adj} increases only if RMSE decreases [since SST remains constant for all models]. Thus, an equivalent procedure is to search for the model with the minimum, or near minimum, RMSE.

3. Mallows's Cp Criterion

The Cp criterion, named for Colin Lingwood Mallow, selects as the best subset model with

(1) a small value of Cp (i.e., a small total mean square error), means that the model is relatively precise.

(2) a value of Cp near p + 1, a property that indicates that slight (or no) bias exists in the subset regression model.

Thus, the Cp criterion focuses on minimizing total mean square error and the regression bias. If we are mainly concerned with minimizing total mean square error, we will want to choose the model with the smallest Cp value, as long as the bias is not large. On the other hand, we may prefer a model that yields a Cp value slightly larger than the minimum but that has slight (or no) bias.

4. AIC (Akaike's information criterion)

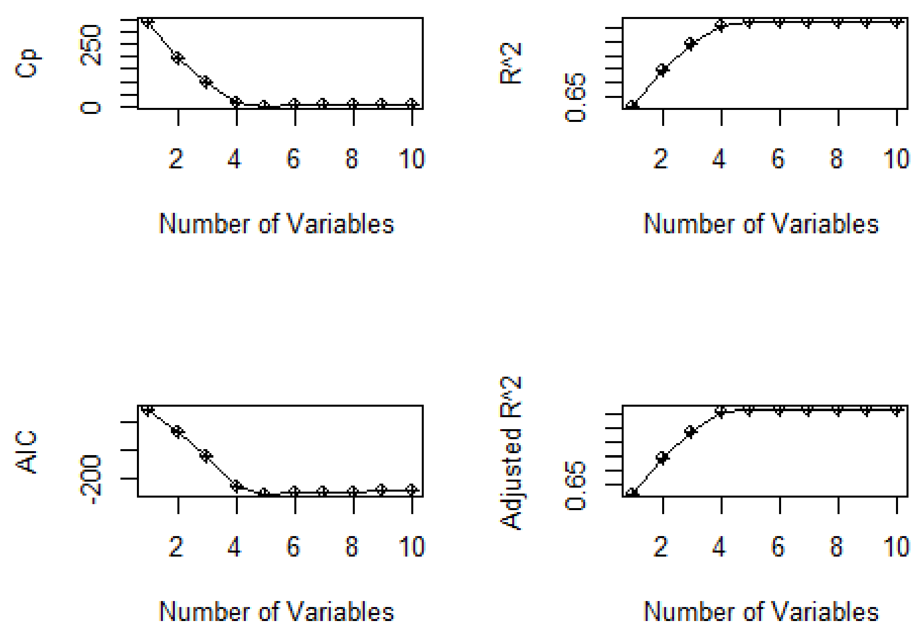
When using the model to predict Y, some information will be lost. Akaike's information criterion estimates the relative information lost by a given model. It is defined as

$$AIC = n\ln(\frac{SSE}{n}) + 2(p+1)$$

The formula is formulated by the statistician **Hirotugu Akaike**. Models with smaller values of AIC are preferred.

In this class, we are going to use R software package to calculate all values.

```
#  Option 1
library(olsrr)
salary=read.csv("c:/Users/thuntida.ngamkham/OneDrive - University of
Calgary/dataset603/EXECSAL2.csv", header = TRUE)
firstordermodel<-lm(Y~X1+X2+X3+X4+X5+X6+X7+X8+X9+X10, data= salary)
#Select the subset of predictors that do the best at meeting some well-
defined objective criterion, such as having the largest R2 value or the
smallest MSE, Mallow's Cp or AIC.
ks=ols_step_best_subset(firstordermodel, details=TRUE)
par(mfrow=c(2,2)) # split the plotting panel into a 2 x 2 grid
plot(ks$cp,type = "o",pch=10, xlab="Number of Variables",ylab= "Cp")
plot(ks$rsq,type = "o",pch=10, xlab="Number of Variables",ylab= "R^2")
#plot(ks$rss, xlab="Number of Variables",ylab= "RMSE")
plot(ks$aic,type = "o",pch=10, xlab="Number of Variables",ylab= "AIC")
plot(ks$adjr,type = "o",pch=10, xlab="Number of Variables",ylab= "Adjusted
R^2")
```

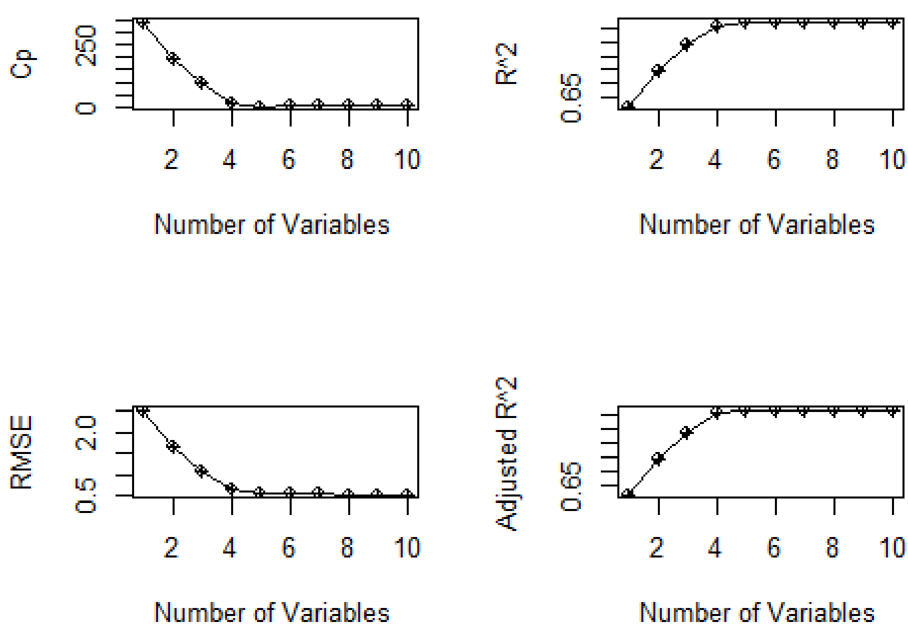


```r
# Option 2

library(leaps) #need to install the package leaps for best.subset() function
#by default, regsubsets() only report results up to the best 8-variable model
best.subset<-regsubsets(Y~X1+X2+X3+X4+X5+X6+X7+X8+X9+X10, data= salary, nv=10
)
#by default, regsubsets() only reports results up to the best 8-variable
model
#Model selection by exhaustive search, forward or backward stepwise, or
sequential replacement
#The summary() command outputs the best set of variables for each model size
using RMSE.
summary(best.subset)

## Subset selection object
## Call: regsubsets.formula(Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 +
##     X9 + X10, data = salary, nv = 10)
## 10 Variables  (and intercept)
##       Forced in Forced out
## X1        FALSE      FALSE
## X2        FALSE      FALSE
## X3yes     FALSE      FALSE
## X4        FALSE      FALSE
## X5        FALSE      FALSE
## X6yes     FALSE      FALSE
## X7        FALSE      FALSE
## X8        FALSE      FALSE
## X9yes     FALSE      FALSE
## X10       FALSE      FALSE
## 1 subsets of each size up to 10
## Selection Algorithm: exhaustive
##           X1  X2  X3yes X4  X5  X6yes X7  X8  X9yes X10
## 1  ( 1 )  "*" " " " "   " " " " " "   " " " " " "   " "
## 2  ( 1 )  "*" " " "*"   " " " " " "   " " " " " "   " "
## 3  ( 1 )  "*" " " "*"   "*" " " " "   " " " " " "   " "
## 4  ( 1 )  "*" "*" "*"   "*" " " " "   " " " " " "   " "
## 5  ( 1 )  "*" "*" "*"   "*" "*" " "   " " " " " "   " "
## 6  ( 1 )  "*" "*" "*"   "*" "*" " "   " " " " "*"   " "
## 7  ( 1 )  "*" "*" "*"   "*" "*" "*"   " " " " "*"   " "
## 8  ( 1 )  "*" "*" "*"   "*" "*" "*"   " " "*" "*"   " "
## 9  ( 1 )  "*" "*" "*"   "*" "*" "*"   "*" "*" "*"   " "
## 10  ( 1 ) "*" "*" "*"   "*" "*" "*"   "*" "*" "*"   "*"

reg.summary<-summary(best.subset)

# for the output interpretation
rsquare<-c(reg.summary$rsq)
cp<-c(reg.summary$cp)
AdjustedR<-c(reg.summary$adjr2)
```

```r
RMSE<-c(reg.summary$rss)
cbind(rsquare,cp,RMSE,AdjustedR)
```

```
##         rsquare         cp      RMSE AdjustedR
##  [1,] 0.6189795 343.856582 2.5462337 0.6150915
##  [2,] 0.7492075 195.519164 1.6759632 0.7440365
##  [3,] 0.8390930  93.753768 1.0752880 0.8340647
##  [4,] 0.9074746  16.812839 0.6183162 0.9035788
##  [5,] 0.9206284   3.627915 0.5304140 0.9164065
##  [6,] 0.9220182   4.023513 0.5211265 0.9169871
##  [7,] 0.9225151   5.449923 0.5178061 0.9166195
##  [8,] 0.9227354   7.195556 0.5163336 0.9159429
##  [9,] 0.9228103   9.109093 0.5158331 0.9150913
## [10,] 0.9229048  11.000000 0.5152016 0.9142424
```

```r
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")
```



R functions regsubsets() :performs best sub- set selection by identifying the best model that contains a given number of predictors. ols_step_best_subset() : perform best sub- set selection by identifying the best model that contains a given number of predictors

From the output, the first order regression model is $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon$. Is this model the best fitted model for predicting executive salary?

Inclass practice Problem

From the credit card example, using All Possible Regressions Selection Procedure to analyse which independent predictors should be used in the model.

3. Evaluate the reliability of the model chosen.

After using model selection by automatic methods or all possible regression methods, we might not have the best fit model yet, as we consider only main effects on independent variables. After eliminating some variables that are not important out of the model, we consider interaction terms and/or high order multiple regression model to improve the model.

```
salary=read.csv("c:/Users/thuntida.ngamkham/OneDrive - University of
Calgary/dataset603/EXECSAL2.csv", header = TRUE )

firstordermodel<-lm(Y~X1+X2+factor(X3)+X4+X5,data=salary)
summary(firstordermodel)

##
## Call:
## lm(formula = Y ~ X1 + X2 + factor(X3) + X4 + X5, data = salary)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.201219 -0.056016 -0.003581  0.053656  0.187251
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.9619345  0.1010567  98.578  < 2e-16 ***
## X1            0.0272762  0.0010293  26.501  < 2e-16 ***
## X2            0.0290921  0.0033367   8.719 9.71e-14 ***
## factor(X3)yes 0.2246932  0.0163503  13.742  < 2e-16 ***
## X4            0.0005244  0.0000474  11.064  < 2e-16 ***
## X5            0.0019623  0.0004972   3.947 0.000153 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07512 on 94 degrees of freedom
## Multiple R-squared:  0.9206, Adjusted R-squared:  0.9164
## F-statistic: 218.1 on 5 and 94 DF,  p-value: < 2.2e-16
```

```r
# building the best model with interation term
interacmodel<-lm(Y~(X1+X2+factor(X3)+X4+X5)^2,data = salary)
summary(interacmodel)

##
## Call:
## lm(formula = Y ~ (X1 + X2 + factor(X3) + X4 + X5)^2, data = salary)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.174954 -0.051664 -0.001672  0.047063  0.163348
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       9.467e+00  7.451e-01  12.705  < 2e-16 ***
## X1                4.238e-02  1.514e-02   2.798  0.00637 **
## X2                7.323e-02  3.893e-02   1.881  0.06344 .
## factor(X3)yes    -1.140e-01  2.029e-01  -0.562  0.57564
## X4                6.225e-04  6.279e-04   0.991  0.32436
## X5                3.466e-03  4.453e-03   0.778  0.43858
## X1:X2            -7.848e-04  4.976e-04  -1.577  0.11850
## X1:factor(X3)yes  7.695e-04  2.271e-03   0.339  0.73556
## X1:X4            -2.135e-07  6.283e-06  -0.034  0.97298
## X1:X5            -1.804e-05  6.987e-05  -0.258  0.79686
## X2:factor(X3)yes -5.825e-03  7.254e-03  -0.803  0.42424
## X2:X4            -8.966e-06  2.151e-05  -0.417  0.67785
## X2:X5            -1.430e-04  2.260e-04  -0.633  0.52853
## factor(X3)yes:X4  2.346e-04  1.076e-04   2.179  0.03211 *
## factor(X3)yes:X5  1.898e-03  1.096e-03   1.732  0.08703 .
## X4:X5            -6.789e-07  3.275e-06  -0.207  0.83627
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07333 on 84 degrees of freedom
## Multiple R-squared:  0.9324, Adjusted R-squared:  0.9203
## F-statistic: 77.25 on 15 and 84 DF,  p-value: < 2.2e-16

bestinteracmodel<-lm(Y~X1+X2+factor(X3)+X4+X5+factor(X3)*X4,data=salary)
summary(bestinteracmodel)

##
## Call:
## lm(formula = Y ~ X1 + X2 + factor(X3) + X4 + X5 + factor(X3) *
##     X4, data = salary)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.210078 -0.052939  0.003473  0.046302  0.155280
##
## Coefficients:
```

```
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.002e+01  1.001e-01 100.096  < 2e-16 ***
## X1                2.690e-02  1.006e-03  26.741  < 2e-16 ***
## X2                2.977e-02  3.240e-03   9.189 1.06e-14 ***
## factor(X3)yes     1.234e-01  4.071e-02   3.032 0.003150 **
## X4                3.263e-04  8.655e-05   3.770 0.000286 ***
## X5                2.043e-03  4.823e-04   4.236 5.34e-05 ***
## factor(X3)yes:X4 2.744e-04  1.016e-04   2.700 0.008249 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07273 on 93 degrees of freedom
## Multiple R-squared:  0.9264, Adjusted R-squared:  0.9216
## F-statistic: 195.1 on 6 and 93 DF,  p-value: < 2.2e-16
```

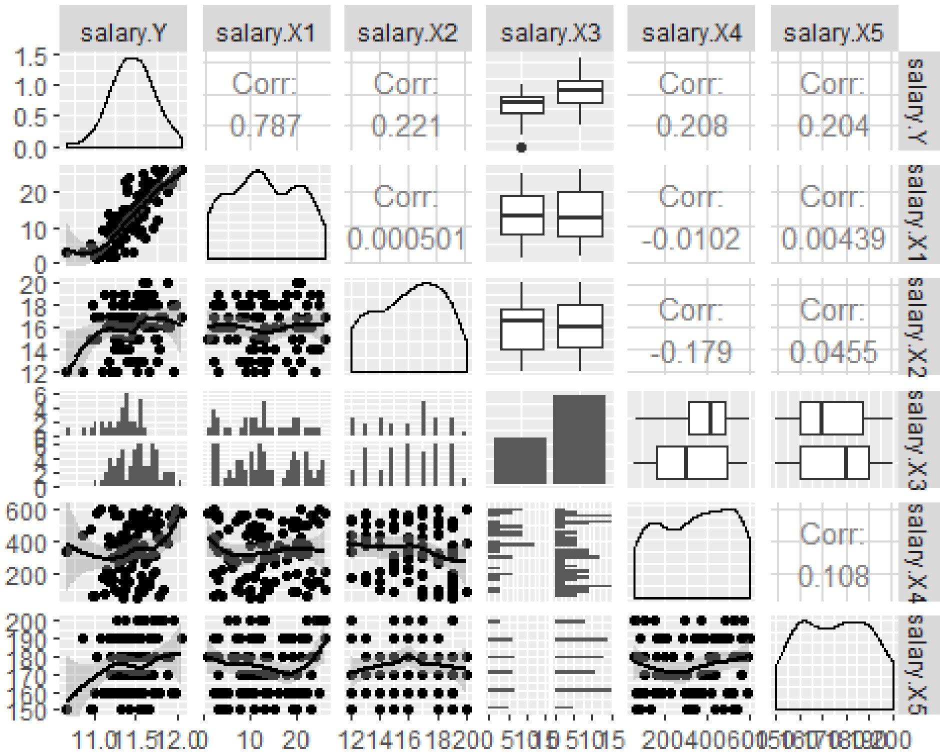
```
#considering high order model between Xs and Y to improve the model
library(GGally) # need to install the GGally package for ggpairs function
```

```
## Loading required package: ggplot2
```

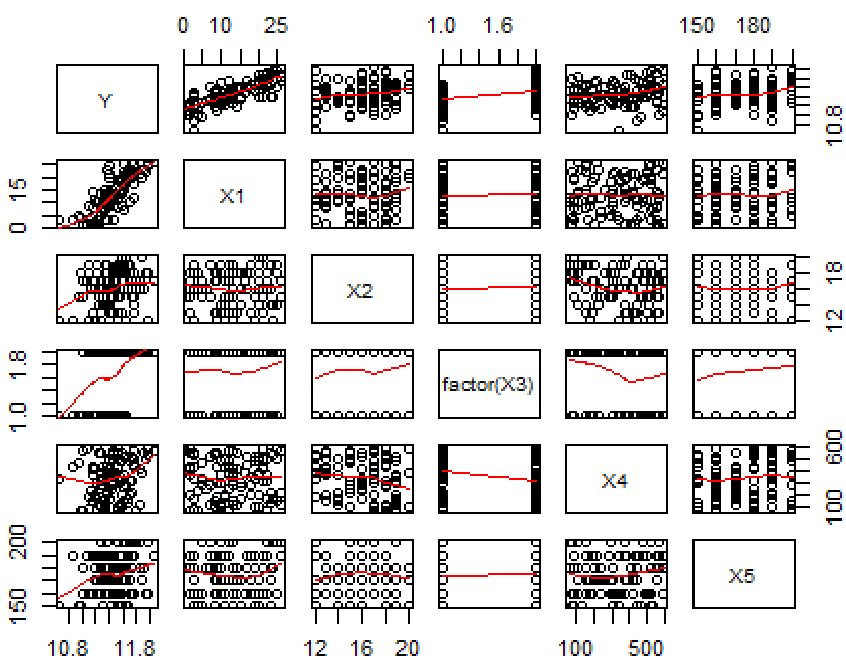
```
#option 1: using function ggpairs()
salarydata <-
data.frame(salary$Y,salary$X1,salary$X2,salary$X3,salary$X4,salary$X5)
#ggpairs(salarydata)
#LOESS or LOWESS: LOcally WEighted Scatter-plot Smoother
ggpairs(salarydata,lower = list(continuous = "smooth_loess", combo =
  "facethist", discrete = "facetbar", na = "na"))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```r
#option2: using function pairs()
pairs(~Y+X1+X2+factor(X3)+X4+X5,data=salary,panel = panel.smooth)
```



```r
bestmodel<-lm(Y~X1+I(X1^2)+X2+factor(X3)+X4+X5+factor(X3)*X4,data=salary)
summary(bestmodel)
```

```
##
## Call:
## lm(formula = Y ~ X1 + I(X1^2) + X2 + factor(X3) + X4 + X5 + factor(X3) *
##     X4, data = salary)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.163466 -0.048971 -0.001111  0.041345  0.124534
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        9.862e+00  9.703e-02 101.634  < 2e-16 ***
## X1                 4.364e-02  3.761e-03  11.604  < 2e-16 ***
## I(X1^2)           -6.347e-04  1.384e-04  -4.588 1.41e-05 ***
## X2                 3.094e-02  2.950e-03  10.487  < 2e-16 ***
## factor(X3)yes      1.166e-01  3.696e-02   3.155  0.00217 **
## X4                 3.259e-04  7.850e-05   4.152 7.36e-05 ***
## X5                 2.391e-03  4.439e-04   5.386 5.49e-07 ***
## factor(X3)yes:X4   3.020e-04  9.239e-05   3.269  0.00152 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06596 on 92 degrees of freedom
## Multiple R-squared:  0.9401, Adjusted R-squared:  0.9355
## F-statistic: 206.3 on 7 and 92 DF,  p-value: < 2.2e-16
```

R Functions ggpairs(): look at all pairwise combinations of continuous variables in scatterplots. pairs(): optional function for pairwise combinations panel.smooth: add a smooth loess curve on the scatters

From the output, you can see that after including an interaction term $(X_3 * X_4)$ and quadratic term X_1^2, they led to such a big improvement in the model as following,

1. all the p-values < 0.05, which means that all regression coefficients were significantly non-zero.

2. R_{adj}^2 increases from 0.9164 to 0.9355

3. Standard error of residuals (RMSE) decreases from 0.07512 to 0.06596

Therefore, it is clear that adding the additional terms really has led to a better fit to the data.

Inclass Practice Problem

From the credit card example, when we investigate the scatter plots for all pairwise combinations between variables, we found that Rating and Limit variable are correlated to each other (R^2 is very high)

Inclass Practice Problem

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.