TOPIC 3

November-17-22 1:30 PM

TOPIC3: Model selection

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

in "data science" applications there are often scores of variables

THIS IS JUST ONE WORKFLOW

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. → "FULL" MODEL
2. Specify a strategy for selecting a model (backwards, forwards, stepwise, all-subsets)
3. Evaluate the reliability of the model chosen.

} depending on your full model + strategy, you will get a different outcome.

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.

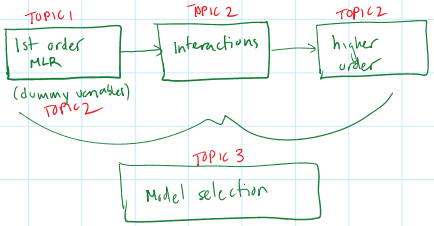
recommend starting with no interactions + no higher order

Step 1: Specifying the Maximum Model = FULL 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



↗ data dredging.

Take a "data mining" approach.

↳ one of the reasons to do this is because data scientists care most about predicting not about why

approach here → maximizing our ability to predict

BACKWARD FORWARD STEPWISE REGRESSION

You have a bunch of variables to screen + you want to figure out which ones are useful for predicting your response

} The problem statement of this section.

THERE IS NO SINGLE IDEAL MODEL

↓

It's just what's best for our purpose

HOW DOES A STEPWISE REGRESSION WORK?

VARIABLES $x_1, x_2, \cdots x_{10}$

1 Creates all the simple linear regression models

$\hat{y}_i = \hat{B}_0 + \hat{B}_1 x_1$ compare all models

unpack the stepwise black box

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, $x_2, \ldots, x_p$, 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: $\beta_1 = 0$

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.

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

```
##
## Attaching package: 'olsrr'
```

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

```
salary=read.csv("EXECSAL2.csv", header = TRUE)
model1<-lm(Y~X1, data = salary)
summary(model1)
```

```
##
## Call:
## lm(formula = Y ~ X1, data = salary)
```


Let's do this manually to show you → (There is an automated function) but for now let's do it step-by-step.

black box

1 Creates all the simple linear regression models

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_2$
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_{10}$

compare all models + find the one with highest t-score (highest $|t_{calc}|$) ∴ lowest p-value

2 Let's say x_1 had highest $|t_{calc}|$. Now use it to build models with 2 predictors

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_3$
etc

which should I add to x_1? based on $|t_{calc}|$

3 Assume $x_1 + x_3$ are now selected
- Check if x_1 is still significant
- If not remove x_1 + keep only x_3
- If it is, keep $x_1 + x_3$

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.51010 -0.08148  0.01533  0.09007  0.34663
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.090887   0.033055  335.52   <2e-16 ***
## X1           0.027839   0.002206   12.62   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1612 on 98 degrees of freedom
## Multiple R-squared:  0.619,  Adjusted R-squared:  0.6151
## F-statistic: 159.2 on 1 and 98 DF,  p-value: < 2.2e-16
```

this is the highest/tcalc/ of all simple regressions

```
model2<-lm(Y~X2, data = salary)
summary(model2)
```

```
##
## Call:
## lm(formula = Y ~ X2, data = salary)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.69058 -0.17417  0.01475  0.14929  0.60722
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.05594    0.17971  61.520   <2e-16 ***
## X2           0.02491    0.01110   2.243   0.0271 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2547 on 98 degrees of freedom
## Multiple R-squared:  0.04884,    Adjusted R-squared:  0.03914
## F-statistic: 5.032 on 1 and 98 DF,  p-value: 0.02713
```

```
model3<-lm(Y~X3, data = salary)
summary(model3)
```

```
##
## Call:
## lm(formula = Y ~ X3, data = salary)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.64801 -0.17344  0.02863  0.18306  0.53486
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.31231    0.04112 275.116  < 2e-16 ***
## X3yes        0.21623    0.05061   4.272 4.49e-05 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2398 on 98 degrees of freedom
## Multiple R-squared:  0.157,  Adjusted R-squared:  0.1484
## F-statistic: 18.25 on 1 and 98 DF,  p-value: 4.487e-05
```

```
model4<-lm(Y~X4, data = salary)
summary(model4)
```

```
##
## Call:
## lm(formula = Y ~ X4, data = salary)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.79069 -0.16613 -0.01677  0.18069  0.53399
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.134e+01  5.813e-02 195.157   <2e-16 ***
## X4          3.236e-04  1.535e-04   2.107   0.0376 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2554 on 98 degrees of freedom
## Multiple R-squared:  0.04335,    Adjusted R-squared:  0.03359
## F-statistic: 4.441 on 1 and 98 DF,  p-value: 0.03763
```

```
model5<-lm(Y~X5, data = salary)
summary(model5)
```

```
##
## Call:
## lm(formula = Y ~ X5, data = salary)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.70447 -0.17997 -0.00744  0.17354  0.57667
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.853365   0.293139   37.02   <2e-16 ***
## X5           0.003436   0.001668    2.06    0.042 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2557 on 98 degrees of freedom
## Multiple R-squared:  0.04152,    Adjusted R-squared:  0.03174
## F-statistic: 4.245 on 1 and 98 DF,  p-value: 0.04202
```

```
model6<-lm(Y-X6, data = salary)
summary(model6)
```

```
##
## Call:
## lm(formula = Y - X6, data = salary)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.77744 -0.18580 -0.00297  0.15596  0.59563
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.46777    0.03652 314.017   <2e-16 ***
## X6yes       -0.02603    0.05217  -0.499    0.619
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2608 on 98 degrees of freedom
## Multiple R-squared:  0.002533,   Adjusted R-squared:  -0.007645
## F-statistic: 0.2489 on 1 and 98 DF,  p-value: 0.619
```

```
model7<-lm(Y-X7, data = salary)
summary(model7)
```

```
##
## Call:
## lm(formula = Y - X7, data = salary)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.52333 -0.13687  0.02306  0.13711  0.49733
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.682669   0.098393 108.571  < 2e-16 ***
## X7           0.018029   0.002247   8.022 2.28e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2029 on 98 degrees of freedom
## Multiple R-squared:  0.3964, Adjusted R-squared:  0.3902
## F-statistic: 64.35 on 1 and 98 DF,  p-value: 2.27e-12
```

```
model8<-lm(Y-X8, data = salary)
summary(model8)
```

```
##
## Call:
## lm(formula = Y - X8, data = salary)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.78463 -0.17565  0.00108  0.14772  0.62316
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.388078   0.132479  85.961   <2e-16 ***
## X9           0.008693   0.016868   0.515    0.607
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2608 on 98 degrees of freedom
## Multiple R-squared:  0.002703,   Adjusted R-squared:  -0.007474
## F-statistic: 0.2656 on 1 and 98 DF,  p-value: 0.6075
```

```
model9<-lm(Y~X9, data = salary)
summary(model9)
```

```
##
## Call:
## lm(formula = Y ~ X9, data = salary)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.79002 -0.17332  0.00838  0.15368  0.60908
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.45432    0.02884 397.211   <2e-16 ***
## X9yes        0.00386    0.06797   0.057    0.955
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2611 on 98 degrees of freedom
## Multiple R-squared:  3.291e-05,  Adjusted R-squared:  -0.01017
## F-statistic: 0.003225 on 1 and 98 DF,  p-value: 0.9548
```

```
model10<-lm(Y~X10, data = salary)
summary(model10)
```

```
##
## Call:
## lm(formula = Y ~ X10, data = salary)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7916 -0.1661  0.0035  0.1677  0.5867
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.325878   0.238765  47.435   <2e-16 ***
## X10          0.005201   0.009558   0.544    0.588
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2607 on 98 degrees of freedom
## Multiple R-squared:  0.003012,   Adjusted R-squared:  -0.007161
## F-statistic: 0.2961 on 1 and 98 DF,  p-value: 0.5876
```

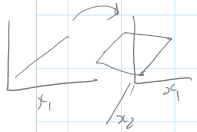
Step 2 The stepwise program now begins to search through the remaining $(p-1)$ independent variables for the best two-variable model of the form

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

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, $X_i, i = 2, 3, ..., p-1$), 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 $\hat{\beta}_1$ after $\hat{\beta}_2 X_2$ 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 $\hat{\beta}_2 X_2$.

The reason the t-value for X_1 may change from step 1 to step 2 is that the meaning of the coefficient $\hat{\beta}_1$ 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 $\hat{\beta}_1$ than that obtained in step 1. Thus, both the value of $\hat{\beta}_1$ 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.




```
library(olsrr)#need to install the package olsrr
salary=read.csv("EXECSAL2.csv", header = TRUE)
model1<-lm(Y~X1+X2, data = salary)
summary(model1)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2, data = salary)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41018 -0.08883 -0.00270  0.08998  0.35311
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.692577   0.110148  97.075  < 2e-16 ***
## X1           0.027835   0.002071  13.439  < 2e-16 ***
## X2           0.024866   0.006598   3.769 0.000282 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1513 on 97 degrees of freedom
## Multiple R-squared:  0.6676, Adjusted R-squared:  0.6608
## F-statistic: 97.43 on 2 and 97 DF,  p-value: < 2.2e-16
```

```
model2<-lm(Y~X1+X3, data = salary)
summary(model2)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X3, data = salary)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.38585 -0.08612  0.00136  0.09114  0.27781
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.968372   0.032010 342.659  < 2e-16 ***
## X1           0.027258   0.001801  15.134  < 2e-16 ***
## X3           0.197135   0.027777   7.097  2.1e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1314 on 97 degrees of freedom
## Multiple R-squared:  0.7492, Adjusted R-squared:  0.744
## F-statistic: 144.9 on 2 and 97 DF,  p-value: < 2.2e-16
```

thus is the highest t-calc given x_1

$x_1 + x_3$ ✓

check if thus is also still significant it is → keep it in.

3

```
model3<-lm(Y~X1+X4, data = salary)
summary(model3)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X4, data = salary)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.50928 -0.08646  0.01543  0.10257  0.29984
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.098e+01  4.414e-02 248.645  < 2e-16 ***
## X1          2.792e-02  2.077e-03  13.439  < 2e-16 ***
## X4          3.361e-04  9.123e-05   3.684 0.000378 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1517 on 97 degrees of freedom
## Multiple R-squared:  0.6657, Adjusted R-squared:  0.6589
## F-statistic:  96.6 on 2 and 97 DF,  p-value: < 2.2e-16
```

```
model4<-lm(Y~X1+X5, data = salary)
summary(model4)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X5, data = salary)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.42563 -0.08619  0.00795  0.08695  0.30373
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.050e+01  1.777e-01  59.077  < 2e-16 ***
## X1          2.781e-02  2.098e-03  13.256  < 2e-16 ***
## X5          3.378e-03  9.997e-04   3.379  0.00105 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1533 on 97 degrees of freedom
## Multiple R-squared:  0.6591, Adjusted R-squared:  0.6521
## F-statistic: 93.77 on 2 and 97 DF,  p-value: < 2.2e-16
```

```
model5<-lm(Y~X1+X6, data = salary)
summary(model5)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X6, data = salary)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.48700 -0.08907  0.00752  0.09016  0.34976
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.110322   0.036001 308.609   <2e-16 ***
## X1           0.027960   0.002199  12.712   <2e-16 ***
## X6Yes       -0.042898   0.032144  -1.335    0.185
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1606 on 97 degrees of freedom
## Multiple R-squared:  0.6258, Adjusted R-squared:  0.6181
## F-statistic: 81.13 on 2 and 97 DF,  p-value: < 2.2e-16
```

```
model6<-lm(Y~X1+X7, data = salary)
summary(model6)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X7, data = salary)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.49375 -0.08082  0.02015  0.08661  0.33368
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.042344   0.091739 120.367  < 2e-16 ***
## X1           0.026315   0.003480   7.562 2.26e-11 ***
## X7           0.001598   0.002816   0.568    0.572
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1617 on 97 degrees of freedom
## Multiple R-squared:  0.6202, Adjusted R-squared:  0.6124
## F-statistic: 79.21 on 2 and 97 DF,  p-value: < 2.2e-16
```

```
model7<-lm(Y~X1+X8, data = salary)
summary(model7)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X8, data = salary)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.50473 -0.08487  0.02426  0.08717  0.36774
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.030724   0.086844 127.018   <2e-16 ***
## X1           0.027828   0.002211  12.584   <2e-16 ***
## X8           0.007832   0.010450   0.749    0.455
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1616 on 97 degrees of freedom
## Multiple R-squared:  0.6212, Adjusted R-squared:  0.6134
## F-statistic: 79.53 on 2 and 97 DF,  p-value: < 2.2e-16
```

```
model8<-lm(Y~X1+X9, data = salary)
summary(model8)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X9, data = salary)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.51268 -0.07828  0.01660  0.09221  0.34383
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.093365   0.033831 327.909   <2e-16 ***
## X1           0.027871   0.002218  12.568   <2e-16 ***
## X9yes       -0.016080   0.042170  -0.381    0.704
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1619 on 97 degrees of freedom
## Multiple R-squared:  0.6195, Adjusted R-squared:  0.6117
## F-statistic: 78.98 on 2 and 97 DF,  p-value: < 2.2e-16
```

```
model9<-lm(Y~X1+X10, data = salary)
summary(model9)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X10, data = salary)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.50403 -0.06077  0.00449  0.08415  0.35693
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.272380   0.147215  76.571   <2e-16 ***
## X1           0.028314   0.002231  12.689   <2e-16 ***
## X10         -0.007560   0.005976  -1.265    0.209
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1607 on 97 degrees of freedom
## Multiple R-squared:  0.6252, Adjusted R-squared:  0.6174
## F-statistic: 80.89 on 2 and 97 DF,  p-value: < 2.2e-16
```

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.

```
library(olsrr)#need to install the package olsrr
salary=read.csv("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:
```


continue until no more variables to add or remove.

Let's do STEPWISE REGRESSION AUTOMATICALLY

"AUTOMAGICALLY?"
No, Paul we are scientists!

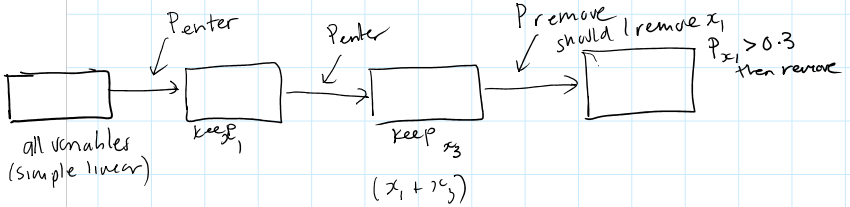
made sure that these qualitative are specified

start by specifying the "FULL MODEL" + run it with lm( )

```
##       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

stepmod=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
## ------------------------------------------------------------------------------------
```

P < Penter

P < Penter

```
##   (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.
## ----------------------------------------------------------------
```

```
##  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
##
##  - X1 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
##
```

Also checking X1 against Premove

P < Premove.

```
##                                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
```

Nothing is greater than Premier so keep everything in.

```
## ----------------------------------------------------------------------------
##
##
##
##                        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
```

No more variables meet the Penter criteria

So STOP

```
##  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
## ----------------------------------------------------------------------------------------
```

output of stepwise regression

```
summary(stepmod$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
```

final model everything significant.

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 10

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 $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$ 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.

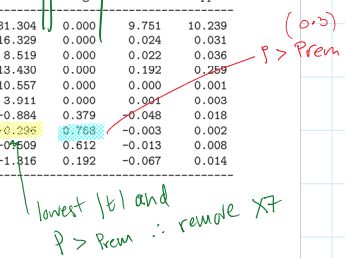
```
library(olsrr) #need to install the package olsrr
salary=read.csv("EXECSAL2.csv", header = TRUE)
fullmodel<-lm(Y~X1+X2+factor(X3)+X4+X5+factor(X6)+X7+X8+factor(X9)+X10, data = salary)
backmodel=ols_step_backward_p(fullmodel, prem = 0.3, details=TRUE)
```

```
## 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
##
```


BACKWARDS REGRESSION SELECTION

1. Start with the full model (best additive)
2. Find the variable with the smallest $|t|$ and remove it, provided it is not significant (specify that with Premove)
3. Repeat until no further non-significant variables can be found.

```
##                         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
## ----------------------------------------------------------------------------------------
```


p is not > Prem
so STOP!

```
##
##
##
## 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
##
##                                ANUVA
## --------------------------------------------------------------------
##                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
```

this was kept in because Prem = 0.3 our next step might be to rerun this model without this variable do this manually with lm(-)

R function: 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 11

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.

CREDIT.CSV

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.

no step backward just forwards.

```
library(olsrr) #need to install the package olsrr
salary=read.csv("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)
```

Much the same as stepwise

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


LIMITATIONS OF BACKWARDS IN PRACTICE

① Need to fit a model with everything in it at STEP 1
- if there are lots of variables thus could be time prohibitive with large data sets.

↓

in such a case we might prefer FORWARD or STEPWISE

```
##
## 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)
##
```

variable with highest |t|

```
##                       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
## ----------------------------------------------------------------
```

No more variables to add based on $P < P_{enter}$

```
## 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
## ----------------------------------------------------------------------------------------
```

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

— we can get the model output this way

```
##
## 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
```

check that all terms are non-zero in indiv t-tests

STEP, FW, BW } all methods are limited in one way or another → they may miss certain combinations of variables

```
## 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 12

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 aware 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 $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$ will yield the largest R^2.

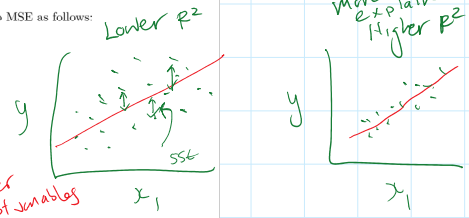
2. Adjusted R^2 or RMSE Criterion

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:

$$R^2_{adj} = 1 - \frac{\frac{SSE}{n-p-1}}{\frac{SST}{n-1}}$$

$$R^2_{adj} = 1 - (n-1)\frac{MSE}{SST}$$

$$s = RMSE = \sqrt{\frac{1}{n-p-1} SSE}$$



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.

Another measure of model fit

3. Mallow's Cp Criterion

$C_p = \frac{SSE}{s^2} - n + 2p$

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. smaller is better

(2) a value of Cp near $p + 1$, a property that indicates that slight (or no) bias exists in the subset regression model. keeps as close to this limit to reduce bias

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.

∴ if there are 4 variables a $C_p \approx 5$ is least biased

too few variables to adequately predict.

select lowest C_p closest to $C_p = p+1$

4. AIC (Akaike's information criterion) — another metric of model fit

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

smaller is better

AIC → R²

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

don't worry about formula

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

Where

n : the number of observations in the dataset

p : the number of parameters in the model

5. BIC (Bayesian information criteria)

Bayesian information criterion (BIC) is another criterion for model selection. It is based, in part, on the likelihood function, and it is closely related to Akaike information criterion (AIC). The models can be tested using corresponding BIC values. Lower BIC value indicates a better model.[2]

smaller is better

$$BIC = n\ln(\frac{SSE}{n}) + (p)\ln(n)$$

used equivalently to AIC

Note!

n : the number of observations in the dataset

p : the number of parameters in the model

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

```
# Option 1
library(olsrr)
salary=read.csv("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 ols_best_subset(firstordermodel, details=TRUE)
# for the output interpretation
rsquare<-c(ks$rsquare)
AdjustedR<-c(ks$adjr)
```

full model

ks ← ols_step(———)

CRITERIA TO ASSESS MODEL FIT

- R^2_{adj} — higher is better
- C_p — lower is better, near p+1
- AIC / BIC — lower is better
- RMSE — lower is better

```
cp<-c(ks$cp)
AIC<-c(ks$aic)
cbind(rsquare,AdjustedR,cp,AIC)
```

Take all metrics stick em in a dataframe

a different model

```
##          rsquare AdjustedR        cp       AIC
## [1,] 0.8189795 0.6150915 343.856582  -77.26778
## [2,] 0.7492075 0.7440365 195.519164 -117.09051
## [3,] 0.8390930 0.8340647  93.753768 -159.47046
## [4,] 0.9074746 0.9035788  16.812839 -212.80484
## [5,] 0.9206284 0.9164065   3.627915 -226.13906
## [6,] 0.9220182 0.9169871   4.023513 -225.90557
## [7,] 0.9255151 0.9166195   5.449923 -224.54476
## [8,] 0.9273354 0.9159429   7.195556 -222.82953
## [9,] 0.9288103 0.9150913   9.109093 -220.92652
## [10,] 0.9229048 0.9142424  11.000000 -219.04902
```

this represents the best model that has only (1) predicts

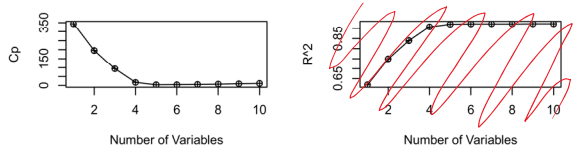
this represents the best model that has (6) predictors

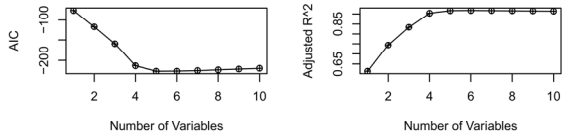
Depending on the criterion we might say the best model has 5 or 6 predictors

By examining this table we are deciding how many variables are in the best model

you can plot this table to see patterns

```
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$aic,type = "o",pch=10, xlab="Number of Variables",ylab= "AIC")
plot(ks$adjr,type = "o",pch=10, xlab="Number of Variables",ylab= "Adjusted R^2")
```





R functions

ols_step_best_subset Best subsets regression, select the subset of predictors that do the best at meeting some

one approach to doing all subsets regression

well-defined objective criterion, such as having the largest adj R^2 value or the smallest MSE, Mallow's Cp or AIC.[1] BIC values are not provided

```
#  Option 2
library(olsrr)
salary=read.csv("EXECSAL2.csv", header = TRUE)
firstordermodel<-lm(Y~X1+X2+X3+X4+X5+X6+X7+X8+X9+X10, data= salary)

library(leaps) #need to install the package leaps for regsubsets() function
```

The "leaps" package does the same thing but is a bit easier to interpret.

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

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)
```

only show the best 10 models

```
## 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 ) "*" "*" "*"   "*" "*" "*"   "*" "*" "*"   "*"
```

because of this I recommend using regsubsets(→) in the leaps package

starting with the same full model

let's say we decide these are the one's we are going to defend!

```
reg.summary<-summary(best.subset)

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

If we choose Model 7 it has:
X1
X2
X3
X4
X5
X6
X9

produce a similar table to the one produced above

```
AdjustedR<-c(reg.summary$adjr2)
RMSE<-c(reg.summary$rss)
BIC<-c(reg.summary$bic)
cbind(rsquare,cp,BIC,RMSE,AdjustedR)
```

how many variables are in the model

Different model fit criteria

```
##          rsquare        cp       BIC     RMSE AdjustedR
##  [1,] 0.6189795 343.856582  -87.27986 2.5462337 0.6150915
##  [2,] 0.7492075 195.519164 -124.49741 1.6759632 0.7440365
##  [3,] 0.8390930  93.753768 -164.27219 1.0752880 0.8340647
##  [4,] 0.9074746  16.812839 -215.00141 0.6183162 0.9035788
##  [5,] 0.9200284   3.627915 -225.73645 0.5304140 0.9169871
##  [6,] 0.9220182   4.023513 -222.89179 0.5211265 0.9187871
##  [7,] 0.9225151   5.449923 -218.92582 0.5178061 0.9180335
##  [8,] 0.9227354   7.195594 -214.60542 0.5163336 0.9159429
##  [9,] 0.9228103   9.109093 -210.09723 0.5158331 0.9150913
## [10,] 0.9229048  11.000000 -205.61456 0.5152916 0.9142424
```

best model with 1 predictor

we looked at various fit criteria to find optima but they tell conflicting stories

10 predictors

best model with

When looking at Mallow's Cp in this table find lowest

```
par(mfrow=c(3,2)) # split the plotting panel into a 3 x 2 grid
plot(reg.summary$cp,type = "o",pch=10, xlab="Number of Variables",ylab= "Cp")
plot(reg.summary$bic,type = "o",pch=10, xlab="Number of Variables",ylab= "BIC")
plot(reg.summary$rsq,type = "o",pch=10, xlab="Number of Variables",ylab= "R^2")
```

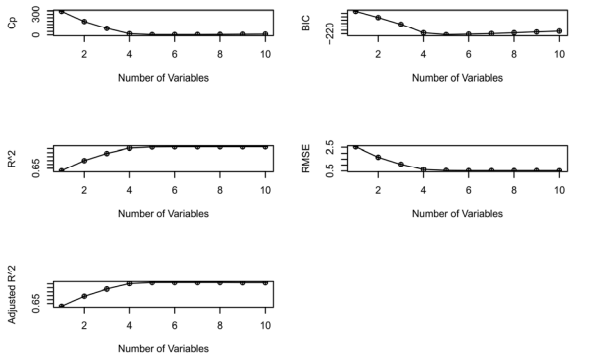
OPTION 1 Choose your favourite criterion + use it

OPTION 2 Choose a balance of criteria. + look for consensus

OPTION 3 Make "forward" several

```
## [10,] 0.9229048  11.000000 -205.61456 0.5152916 0.9142424
```

```
par(mfrow=c(3,2)) # split the plotting panel into a 3 x 2 grid
plot(reg.summary$cp,type = "o",pch=10, xlab="Number of Variables",ylab= "Cp")
plot(reg.summary$bic,type = "o",pch=10, xlab="Number of Variables",ylab= "BIC")
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

When looking at Mallow's Cp in this table find lowest value that is closest to the number of variables in model.

OPTION 3

Consensus
Move "forward" several competing models.

WARNING *DANGER*
don't do all subsets with too many variables or it might take longer than the universe has existed.

regsubsets() :performs best sub- set selection by identifying the best model that contains a given number of predictors. No AIC values are provided

---

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 13

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.

checking if interactions + higher-order terms are justified.

this is our favourite additive model let's see if we can make it better

```
salary=read.csv("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
```

interaction model

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

↑ all pairwise interactions


```
##
## 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                  -1.439e-04  2.269e-04  -0.633  0.52853
```

only one significant interaction

These approaches to model selection.

1. All are automated.
2. End results may differ
3. Best subsets gives all possible models a try and suggests which model is the best
4. Best subsets more complex to use & potentially time consuming

A summary

Alt risk the chance that you will make a mistake because you are testing so many times.

Recommended work order

1. 1st order (additive)
2. Interaction model
3. Higher-order model

- automated procedure
- do it by hand
- do it using first principles

DATA603 MLR MASTER RECIPE

```
## 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  2.345e-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 **
```


- only one significant interaction at $\alpha = 0.05$
- rerun it with just that one significant interaction
- cleaner model with everything significant
- ADDITIVE INTERACTIONS

```
## ---
## 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
```

```
## Warning: package 'GGally' was built under R version 4.2.2
```

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

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
#option 1: using function ggpairs()
salarydata <-data.frame(salary$Y,salary$X1,salary$X2,salary$X3,salary$X4,salary$X5)
salarydata
```

```
##     salary.Y salary.X1 salary.X2 salary.X3 salary.X4 salary.X5
## 1    11.6009        17        16        no       520       180
## 2    11.0837         2        17        no       590       190
## 3    11.2159         2        18        no       600       190
## 4    11.2810        13        12        no       390       170
## 5    11.3218        11        14        no       440       150
## 6    10.9819         4        18        no        70       150
## 7    11.3964        13        16        no       420       170
## 8    11.5973        25        19        no       150       200
## 9    11.1732         2        17        no       430       190
## 10   11.4648        13        13        no       570       180
## 11   10.8493         3        12        no       440       190
## 12   11.5991        22        17        no       370       200
## 13   11.1065         9        12        no       180       160
## 14   11.3278        10        18        no        90       180
## 15   11.4917        16        17        no       380       160
## 16   11.9621        24        12       yes       530       200
## 17   11.5703         9        13       yes       560       170
## 18   11.5768        14        18       yes       110       170
## 19   11.5750        18        13       yes       190       190
## 20   11.2567        10        14       yes       110       160
## 21   11.7707        21        13       yes       430       190
## 22   11.7448        26        15       yes       210       190
## 23   11.7110        22        18       yes       320       160
## 24   11.4742         3        16       yes       560       180
## 25   11.7668        17        18       yes       450       190
## 26   11.1872         2        16       yes       410       180
## 27   11.2810         8        17       yes        90       190
## 28   11.4731        13        15       yes       290       160
## 29   11.4606         3        18       yes       530       180
```


Higher-order ?

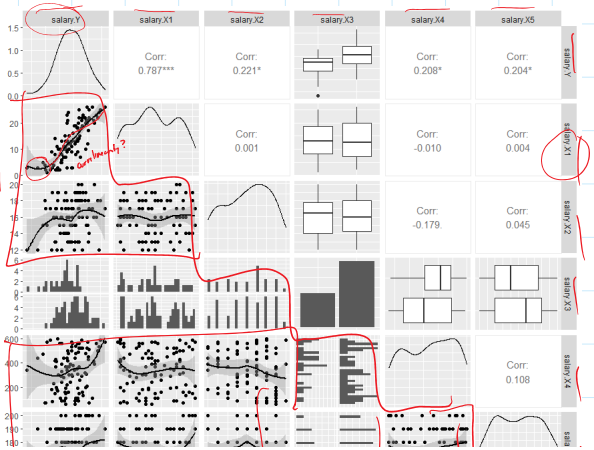
how do we decide if we need higher-order terms

```
## 30   11.4648      11      15      yes      500      190
## 31   12.0634      26      17      yes      570      190
## 32   11.5806      20      20      yes       90      150
## 33   11.5129      19      12      yes      340      160
## 34   11.5199      12      13      yes      440      170
## 35   11.9369      22      18      yes      500      160
## 36   11.2554       2      15      yes      560      190
## 37   11.6639      23      19      yes      130      150
## 38   11.5759      13      19      yes      310      150
## 39   11.6182       7      19      yes      520      200
## 40   11.9798      25      18      yes      590      160
## 41   11.7159      10      19      yes      480      200
## 42   11.1169       3      19      yes       80      160
## 43   11.3874      20      14       no      370      170
## 44   11.1619      14      13       no      420      160
## 45   11.2292      10      19       no      300      170
## 46   11.3794      23      14       no      220      170
## 47   11.4175      15      16       no      300      150
## 48   11.5560      18      19       no      350      160
## 49   11.3998      12      17       no      480      190
## 50   10.6643       3      12       no      340      150
## 51   11.5815      20      17       no      490      160
## 52   11.0186       1      15       no      570      180
## 53   11.3574      11      17       no      190      160
## 54   11.3953      21      13       no      500      160
## 55   11.4436      12      15      yes      240      170
## 56   11.7753      25      14      yes      510      160
## 57   11.2172       3      19      yes      170      170
## 58   11.6553      19      12      yes      520      150
## 59   11.6457      18      18      yes      290      170
## 60   11.1927       2      17      yes      200      180
## 61   11.5954      14      13      yes      560      180
## 62   11.1360       4      16      yes      230      160
## 63   11.8629      21      16      yes      410      180
## 64   11.4175      10      13      yes      370      190
## 65   11.2037      11      12      yes      180      170
## 66   11.5229      12      19      yes       60      200
## 67   11.3551      10      19      yes       60      180
## 68   11.8372      26      17      yes      110      200
## 69   11.3181       7      15      yes      280      190
## 70   11.3563       7      19      yes      110      180
## 71   11.7527      12      15      yes      570      200
## 72   11.2910       6      16      yes      240      180
## 73   11.6046      15      18      yes      260      170
## 74   11.1662       8      13      yes      150      160
## 75   11.1732       2      13      yes      370      190
## 76   11.3551      13      14      yes      150      160
## 77   11.7345      21      15      yes      310      180
## 78   11.7361      20      16      yes      520      160
## 79   11.7134      20      19      yes      200      170
## 80   10.9988       2      17      yes       70      160
## 81   11.4690       9      17      yes      300      160
## 82   11.8706      20      20      yes      390      170
## 83   11.3609      13      19       no      370      200
```



```
## 84   11.2910       8      17       no      560      170
## 85   11.6448      21      20       no      590      180
## 86   11.3771       9      18       no      440      180
## 87   11.5415      19      15       no      480      190
## 88   11.3457      15      14      yes      160      170
## 89   11.4360      12      13      yes      390      190
## 90   11.2823       5      17      yes      330      160
## 91   11.2709       5      16      yes      290      200
## 92   10.9526       5      15       no      470      150
## 93   11.4109      24      14       no      160      180
## 94   11.5327       8      18      yes      540      150
## 95   11.5268      19      15      yes       90      180
## 96   11.9144      23      16      yes      560      180
## 97   11.3783       3      16      yes      340      190
## 98   11.7830      22      17      yes       70      200
## 99   11.6579      22      16      yes      160      190
## 100  11.5405      13      18      yes      110      180
```

```
#ggpairs(salarydata)
#LOESS or LOWESS: LOcally WEighted Scatter-plot Smoother
#ggpairs(salarydata,lower = list(continuous = "smooth_loess", combo =
  #"facethist", discrete = "facetbar", na = "na"))
```



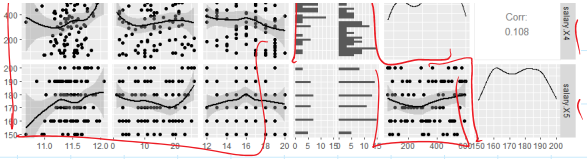
Evidence of Curvilinearity

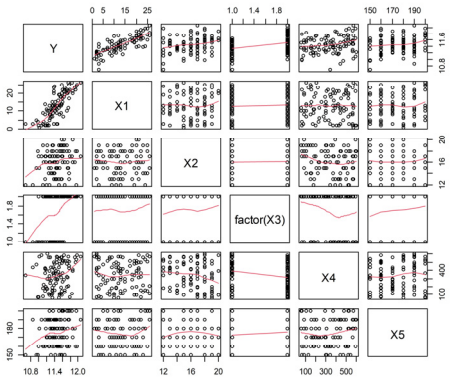
```
## 100  11.5405         13         18        yes        110        180
```

```
#ggpairs(salarydata)
#LOESS or LOWESS: LOcally WEighted Scatter-plot Smoother
#ggpairs(salarydata, lower = list(continuous = "smooth_loess", combo =
  # "facethist", discrete = "facetbar", na = "na"))
#option2: using function pairs()
#pairs(~Y+X1+X2+factor(X3)+X4+X5,data=salary,panel = panel.smooth)

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
```





```
bestmodel1<-lm(Y~X1+I(X1^2)+I(X1^3)+X2+factor(X3)+X4+X5+factor(X3)*X4,data=salary)
summary(bestmodel1)
```

```
##
## Call:
## lm(formula = Y ~ X1 + I(X1^2) + I(X1^3) + X2 + factor(X3) + X4 +
##     X5 + factor(X3) * X4, data = salary)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.163271 -0.048191 -0.000127  0.040151  0.122471
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       9.854e+00  9.980e-02  98.737  < 2e-16 ***
## X1                4.742e-02  1.057e-02   4.485 2.12e-05 ***
## I(X1^2)          -9.854e-04  9.277e-04  -1.062 0.290972
## I(X1^3)           8.853e-06  2.316e-05   0.382 0.703128
## X2                3.094e-02  2.964e-03  10.439  < 2e-16 ***
## factor(X3)yes     1.198e-01  3.805e-02   3.148 0.002222 **
## X4                3.352e-04  8.249e-05   4.063 0.000103 ***
## X5                2.367e-03  4.504e-04   5.256 9.64e-07 ***
## factor(X3)yes:X4  2.921e-04  9.633e-05   3.033 0.003158 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06627 on 91 degrees of freedom
## Multiple R-squared:  0.9402, Adjusted R-squared:  0.9349
## F-statistic: 178.8 on 8 and 91 DF,  p-value: < 2.2e-16
```

Push it further → do we have a cubic term?

→ it falls out → we pushed it too far go back to quadratic x2

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^2_{adj} 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 14

From the credit card example, when we investigate the scatter plots for all pairwise combinations between variables, find the best fitted model to predict balance. You may include interaction terms and higher order terms to improve the model.

Inclass Practice 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.

[1]https://www.rdocumentation.org/packages/olsrr/versions/0.5.3/topics/ols_step_best_subset

[2]https://medium.com/@analyttica/what-is-bayesian-information-criterion-bic-b3396a894be6

References

-Gareth James & Daniela Witten & Trevor Hastie Robert Tibshirani, An Introduction to Statistical Learning with Applications in R: Springer New York Heidelberg Dordrecht London.

-Wickham and Grolemund, R for Data Science: O'Reilly Media