L24 Backward elimination and stepwise

1. Backward elimination

```
proc reg;
     model y=x1 x2 x3/selection=Backward SLSTAY=0.1;
     run;
```

(1) The process
The process produces at most 4 models: one with all 3 predictors, one with 2 predictors, one with 1 predictor and one without any predictors.
It starts with the model with all 3 predictors. SAS displays its R^2, Cp, ANOVA table and Parameter table with p-values for testing the zero values of each coefficients of predictors in the model. The variable with the largest p-value¿0.1 is eliminated.
SAS then displays R^2, Cp, ANOVA table, and parameter table for the model with 2 predictors. Again, the predictor with the largest p-value¿0.1 for testing zero value of its coefficient is eliminated. ....
The process stops if all predictors are eliminated or if all predictors in the model produced p-value¡0.1.

(2) Abbreviation and defaul
Backward, SLSTAY can be abbreviated as B and SLS.
The default value for SLSTAY is 0.1.

(3) Order of elimination
At one time there are k predictors in the model, to eliminate a predictor k tests will be performed in the same full model with k different reduced models.

$$\begin{aligned} & \text{The largest p-value} \Longleftrightarrow \text{The largest } P(F(1,\, n-p) > F_{ob}) \\ \Longleftrightarrow\ & \text{The smallest } F_{ob} = \tfrac{SSE_r - SSE}{MSE} = \tfrac{SSM - SSM_r}{SSTO - SSM}\,(n-p) = \tfrac{R^2 - R_r^2}{1-R^2}\,(n-p) \\ \Longleftrightarrow\ & \text{The largest } R_r^2 \end{aligned}$$

Thus in each elimination step the largest R^2 criterion is used in a specified candidate model pool.

Comment: In forward selection, at one time there are k predictors remaining in the predictor pool. To select a new predictor k tests will be performed in k different full models with a common reduced model. The predictor with smallest p-value<SLENTRY is selected.

$$\begin{aligned} & \text{The smallest p-value} \Longleftrightarrow \text{The smallest } P(F(1,\, n-p) > F_{ob}) \\ \Longleftrightarrow\ & \text{The largest } F_{ob} = \tfrac{SSE_r - SSE}{MSE} = \tfrac{SSM - SSM_r}{SSTO - SSE}\,(n-p) = \tfrac{R^2 - R_r^2}{1-R^2}\,(n-p) \\ \Longleftrightarrow\ & \text{The largest } R^2 \end{aligned}$$

Thus in each selection step the largest R^2 criterion is used in a specified candidate model pool.

2. Stepwise

```
proc reg;
     model y=x1 x2 x3/selection=Stepwise SLENTRY=0.10 SLSTSAY=0.10;
     run;
```

(1) The process The main scheme is forward selection. But after each selection, the process shifts to backward elimination mode. Regardless the result of the elimination the process shifts back to forward selection.
The process stops if no predictors in the model can be eliminated, and no predictors still in the pool can be selected.

(2) An example

```
proc reg;
     model y=x1 x2 x3/selection=Stepwise SLENTRY=0.10 SLSTSAY=0.10;
     run;
```

```
                              Stepwise Selection: Step 1
                    Variable x2 Entered: R-Square = 0.9433 and C(p) = 18.6298

                                   ........

-----------------------------------------------------------------------------------------
                              Stepwise Selection: Step 2
                    Variable x3 Entered: R-Square = 0.9821 and C(p) = 5.1564

                                   ........

-----------------------------------------------------------------------------------------


          All variables left in the model are significant at the 0.1000 level.
      No other variable met the 0.1000 significance level for entry into the model.

                                Summary of Stepwise Selection

      Variable    Variable    Number    Partial     Model
Step  Entered     Removed     Vars In   R-Square    R-Square    C(p)     F Value   Pr > F

  1   x2                         1      0.9433      0.9433    18.6298     99.83   <.0001
  2   x3                         2      0.0388      0.9821     5.1564     10.81   0.0218
```

3. Other operations on variables

(1) Creating more predictors
With x_1 and x_2 one can create $x_3 = x_1^2$, $x_4 = x_2^2$, $x_5 = x_1 * x_2$,....

(2) Transformation of response variable
In stead of using original y as response, one can use $\sqrt{y}$, $\frac{1}{y}$, $\ln y$,....

(3) Check model adequacy
Note that $y = X\beta + \epsilon \Longrightarrow y - X\beta = \epsilon \sim (0, \sigma^2 I)$.
Thus the residual $e = y - X\widehat{\beta} = y - \widehat{y}$ should behave similar to $N(0, \sigma^2 I_n)$.
If the histogram of the residuals is bell-shaped and symmetric, then it is the indication of adequacy of the model.

L25 Influential observations: outliers

1. Influential observations: outlers

   An observation located far away from where it should be is called an outlier. Outlier has greater influence on the outcome of statistical inference. So we need to identify outliers, Investigate the conditions under which the observation was taken with the intention of deleting the outlier to cancel the influence of the observations observed under abnormal conditions.

2. Residuals and studentized residuals

   (1) Residual comparison method
   Residual vector $\widehat{e} = y - \widehat{y} = y - X\widehat{\beta}$ should behave similar to error vector $\epsilon = y - E(y) = y - X\beta \sim N(0, \sigma^2 I_n)$.
   If $|\widehat{e}_i|$ is far greater than others, then it is an indication that the ith observation is an outlier.

   (2) Studentized residual method
   $\widehat{e} = y - \widehat{y} = y - X\widehat{\beta} = y - Hy = (I - H)Y \sim N(0, (I - H)\sigma^2)$. So

   $$\widehat{e}_i \sim N(0, (1 - h_{ii})\sigma^2)$$

   where $\sigma^2_{\widehat{e}_i} = (1 - h_{ii})\sigma^2$ is estimated by $s^2_{\widehat{e}_i} = (1 - h_{ii})\, MSE$. It can be shown that

   $$t_i = \frac{\widehat{e}_i}{s_{\widehat{e}_i}} = \frac{\widehat{e}_i}{\sqrt{(1 - h_{ii})\, MSE}} \sim t(n - p).$$

   This t_i is the studentized residual and h_{ii} is the leverage for the ith observation.
   Based on the probability $P(t(n - p) > |t_i|)$, we know how rare is for t_i at its location. Conventionally the ith observation is an outlier if $|t_i| > 2.5$.

Ex1: SAS can create and store P ($\widehat{y}$), R ($\widehat{e}$), Student (Studentized residual) and H (leverage); display them and do plot.

```
data a;
  infile "D:\Example.txt";
  input y x1 x2;
proc reg;
  model y=x1 x2;
  output out=a P=yhat R=Residual STUDENT=Studebt H=leverage;
  run;
proc print;
  run;
proc plot;
  plot Student*y;
  plot Studebt*yhat;
  run;
```

3. Deleted residuals and studentized deleted resuduals

(1) Deleted residuals
In $\widehat{e}_i = y_i - \widehat{y}_i$, $\widehat{y}_i$ is obtained with the presence of the ith observation. To get the ith residual without the influence of the ith observation. We estimate $E(y_i)$ with data $y_{(i)}$ and $X_{(i)}$, the original y and X with the ith row deleted. The resulted estimated $E(y_i)$ is denoted as $\widehat{y}_{i(i)}$ and the residual $y_i - \widehat{y}_{i(i)}$ is called the ith deleted residual and is denoted as $\widehat{e}_{i(i)}$.

(2) Formula and usage
It can be shown that

$$\widehat{e}_{i(i)} = \frac{\widehat{e}_i}{1 - h_{ii}}$$

If $|\widehat{e}_{i(i)}|$ is far greater than others, then the ith observation is an outlier.

(3) Studentized deleted residual
$\widehat{e}_i \sim N(0,\, (1-h_{ii})\sigma^2) \Longrightarrow \widehat{e}_{i(i)} = \frac{\widehat{e}_i}{1-h_{ii}} \sim N\left(0,\, \frac{\sigma^2}{1-h_{ii}}\right)$.
Here the variance of $\widehat{e}_{i(i)}$, $\sigma^2_{\widehat{e}_{i(i)}} = \frac{\sigma^2}{1-h_{ii}}$ is estimated by its unbiased estimator $s^2_{\widehat{e}_{i(i)}} = \frac{MSE_{(i)}}{1-h_{ii}}$ where $\text{MSE}_{(i)}$ is the MSE in the model without the original ith observation. Then

$$t_{(i)} = \frac{\widehat{e}_{i(i)}}{s_{\widehat{e}_{i(i)}}} = \frac{\widehat{e}_i}{\sqrt{(1-h_{ii})\, MSE_{(i)}}} \sim t(n-1-p)$$

is the ith studentized deleted residual.
Conventionally if $|t_{(i)}| > 2.5$, then the ith observation is an outlier.

Ex2: In SAS deleted residual is called PRESS and studentized deleted residul is called RSTUDENT.

```
proc reg;
   model y=x;
   output out=a R=Resi STUDENT=StuResi H=H PRESS=DelResi RSTUDENT=StuDelRes;
   run;
proc print; run;
```

data c contains 14 observations, data a contains the first 13 observations.

Obs	x	y	Resi	StuResi	H	DelResi	StuDelRes
1	0	0.8	0.10893	0.16432	0.23214	0.14186	0.1575
2	1	1.2	0.00357	0.00510	0.14286	0.00417	0.0049
3	2	1.6	-0.10179	-0.14099	0.08929	-0.11176	-0.1351
4	3	2.0	-0.20714	-0.28415	0.07143	-0.22308	-0.2730
5	4	2.2	-0.51250	-0.70989	0.08929	-0.56275	-0.6944
6	5	2.7	-0.51786	-0.73938	0.14286	-0.60417	-0.7246
7	5	2.6	-0.61786	-0.88216	0.14286	-0.72083	-0.8734
8	4	2.4	-0.31250	-0.43286	0.08929	-0.34314	-0.4177
9	3	2.1	-0.10714	-0.14697	0.07143	-0.11538	-0.1408
10	2	1.7	-0.00179	-0.00247	0.08929	-0.00196	-0.0024
11	1	1.4	0.20357	0.29065	0.14286	0.23750	0.2793
12	0	1.1	0.40893	0.61687	0.23214	0.53256	0.6002
13	6	3.1	-0.62321	-0.94012	0.23214	-0.81163	-0.9352
14	6	6.0	2.27679	3.43453	0.23214	2.96512	25.2209

Based on $t_{14} = 3.43453$ and $t_{(14)} = 25.2209$, the 14th observation is an outlier.