**L22 Model selection**

1. $R^2$-criterion for selecting $k$ predictors

   (1) Selecting $k$ predictors
   Variables $y$ and $x_1, ..., x_m$ are available. We need to select $k$ (specified) predictors in the pool of $x_1, ..., x_n$ to form a "best" regression model with/without intercept (specified). Consider all models of $k$ predictors. There are $C_n^k = \frac{n!}{k!(n-k)!}$ such models. Since $SSM$ is the variation in $y$ explained by the model, by intuition we select one with largest SSM. We call this criterion largest SSM criterion.

   (2) Equivalent criteria
   Suppose $A$ and $B$ are two models with $k$ predictors. Then

$$\boxed{\begin{array}{ccc} SSM_A \geq SSM_B & \Leftrightarrow & MSM_A \geq MSM_B \\ \Updownarrow & & \Updownarrow \\ SSE_A \leq SSE_B & \Leftrightarrow & MSE_A \leq MSE_B \end{array}} \Rightarrow \boxed{F_A \geq F_B \quad \Leftrightarrow \quad p_A \leq p_B}$$

$$\Updownarrow$$

$$\boxed{\begin{array}{ccc} (\sqrt{MSE})_A \leq (\sqrt{MSE})_B & \Leftrightarrow & R_A^2 \geq R_B^2 \\ \Updownarrow & & \Updownarrow \\ \frac{(\sqrt{MSE})_A}{\overline{y}} \times 100 \leq \frac{(\sqrt{MSE})_B}{\overline{y}} \times 100 & & (adj - R^2)_A \geq (adj - R^2)_B \end{array}}$$

   Thus the largest SSM criterion, the largest MSM criterion, the smallest SSE criterion, the smallest MSE criterion, the largest F criterion, the smallest p-value criterion, the smallest Root of MSE criterion, the smallest coefficient of variation criterion, the largest $R^2$ criterion and the largest Adj-$R^2$ criterion are all equivalent.

   (3) Tool
   `proc reg; model y=x1 x2 x3/selection=RSQUARE; run;`
   produces $R^2$ for all $2^3 - 1 = 7$ models with intercept.
   `proc reg; model y=x1 x2 x3/noint selection=RSQUARE; run;`
   produces $R^2$ for all $2^3 - 1 = 7$ models without intercept.

**Ex1:** Implementation with SAS
Suppose we need two predictors to form a regression model for $y$ in a pool of 6 predictors.

```
proc reg;
     model y=x1 x2 x3 x4 x5 x6/selection=RSQUARE;
     run;
```

```
Number in
  Model      R-Square    Variables in Model

      1        0.9987    x6
      1        0.9987    x2
      1        0.0200    x3
      1        0.0182    x1
      1        0.0143    x5
      1        0.0112    x4
-------------------------------------------
                .....
-------------------------------------------
      6        0.9998    x1 x2 x3 x4 x5 x6
```

So the best model is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_6 + \epsilon$.

2. Adj-$R^2$ criterion for selecting a best group of predictors

   (1) Selecting a group of predictors
   Variables $y$ and $x_1, ..., x_m$ are available. We need to select a group of predictors to form a "best" regression model with/without intercept (specified).
   Because adding new predictors into the model increases SSM, the largest SSM criterion will always ends up with selecting all available predictors. Also because the models with different number of predictors are compared, So the largest SSM criterion and the largest MSM criterion are no longer equivalent.

   (2) MSE and equivalent criteria
   In a model MSE is the unbiased estimator for $\sigma^2$, the variance of random error. We may select the model with smallest MSE. For two models $A$ and $B$,

$$\begin{array}{rcl} MSE_A \leq MSE_B & \iff & \text{Adj-}R^2_A \geq \text{Adj-}R^2_B \\ \Updownarrow & & \\ (\sqrt{MSE})_A \leq (\sqrt{MSE})_B & & \\ \Updownarrow & & \\ \frac{(\sqrt{MSE})_A}{\overline{y}} \times 100 \leq \frac{(\sqrt{MSE})_B}{\overline{y}} \times 100 & & \end{array}$$

   Thus the smallest MSE criterion, the smallest Root of MSE criterion, the smallest coefficient of variation criterion and the largest Adj-$R^2$ criterion are all equivalent.

   (3) Tool
   proc reg; model y=x1 x2 x3/selection=ADJRSQ; run;
   produces Adj-$R^2$ for all $2^3 - 1 = 7$ models with intercept.
   proc reg; model y=x1 x2 x3/noint selection=ADFRSQ; run;
   produces ADJ-$R^2$ for all $2^3 - 1 = 7$ models without intercept.

**Ex2:** Implementation with SAS

```
proc reg;
     model y=x1 x2 x3 x4 x5 x6/selection=ADJRSQ;
     run;
```

| Number in Model | Adjusted R-Square | R-Square | Variables in Model |
|---|---|---|---|
| 3 | 0.9942 | 0.9943 | x1 x2 x6 |
| 4 | 0.9941 | 0.9944 | x1 x2 x4 x6 |
| 4 | 0.9941 | 0.9943 | x1 x2 x3 x6 |
| 4 | 0.9941 | 0.9943 | x1 x2 x5 x6 |
| 5 | 0.9941 | 0.9944 | x1 x2 x3 x4 x6 |
| | .... | | |
| 1 | -.0005 | 0.0096 | x3 |

The best model is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_6 + \epsilon$.

**Caution:** Adj-$R^2$ could be a negative number.

3. Cp-criterion

(1) Mallows $C_p$-criterion
Statistician Colin Mallows proposed Cp-criterion. In implementation Cp is calculated for each model and the model with smallest Cp is selected.

**Ex3:** Implementation with SAS.

```
proc reg;
     model y=x1 x2 x3 x4 x5 x6/selection=Cp;
     run;
```

| Number in Model | C(p) | R-Square | Variables in Model |
|---|---|---|---|
| 2 | 0.3948 | 1.0000 | x1 x6 |
| 3 | 1.7666 | 1.0000 | x1 x3 x6 |
| 3 | 2.0336 | 1.0000 | x1 x4 x6 |
| 3 | 2.0431 | 1.0000 | x1 x5 x6 |
| 3 | 2.2593 | 1.0000 | x1 x2 x6 |
| | ... | | |
| 1 | 4456522 | 0.0003 | x5 |

So the model selected by $C_p$ criterion is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_6 + \epsilon$.

(2) Formula

$$C_p = 2p - n + \frac{SSE_{(p)}}{MSE}.$$

Here $MSE$ is from the model with all $m$ predictors. $SSE_{(p)}$ is the model with some predictors such that the number of components in $\beta$ is $p$. Clearly C in Cp is for Criterion and $p$ is for the number of components in $\beta$.

(3) How Cp is defined
$y \sim N(E(y),\, \sigma^2 I_n)$.
With selected predictions, $\widehat{y}_* = H_* y = X_*(X_*'X_*)^{-1}X_*'y$ where $X_* \in R^{n\times p}$.
The error of this model is measured by parameter $\Gamma_p = \frac{E\|E(y)-\widehat{y}_*\|^2}{\sigma^2}$.
To find the expression for $\Gamma_p$ recall: $Z \sim N(\mu,\, \Sigma) \Longrightarrow E\|Z\|^2 = E(Z'Z) = \|\mu\|^2 + \mathrm{tr}(\Sigma)$.
With $-\widehat{y}_* = -H_* y \sim N(-H_* E(y),\, \sigma^2 H_*)$,
$E(y) - \widehat{y}_* \sim N\left((I - H_*)E(y),\, \sigma^2 H_*\right) \Longrightarrow E\|E(y) - \widehat{y}_*\|^2 = \|(I - H_*)E(y)\|^2 + \sigma^2 p$.
$y - \widehat{y}_* \sim N\left((I - H_*)E(y),\, \sigma^2(I - H_*)\right) \Longrightarrow E(SSE_*) = \|(I - H_*)E(y)\|^2 + \sigma^2(n - p)$.
So $E\|E(y) - \widehat{y}_*\|^2 = \sigma^2(2p - n) + E(SSE_*)$ and $\Gamma_p = 2p - n + \frac{E(SSE_*)}{\sigma^2}$.
Replace $\sigma^2$ by its UE MSE and $E(SSE_*)$ by its UE $SSE_*$ now denoted as $SSE_{(p)}$.
We obtain

$$Cp = 2p - n + \frac{SSE_{(p)}}{MSE}.$$

**L23 Information criteria and Forward selection**

1. AIC, BIC and SBC

(1) The largest maximized likelihood criterion
Over all models with/without intercept (specified),
The largest maximized lielihood criterion

$$\begin{aligned}
&= \text{The lagest } \left(\tfrac{1}{2\pi e}\right)^{n/2} \left(\tfrac{SSE}{n}\right)^{-n/2} \text{ criterion} \\
&= \text{The largest } \left(\tfrac{SSE}{n}\right)^{-n/2} \text{ criterion} \\
&= \text{The smallest } -2\ln\left(\tfrac{SSE}{n}\right)^{-n/2} \text{ criterion} \\
&= \text{The smallest } n\ln\tfrac{SSE}{n} \text{ criterion} \\
&= \text{The smallest } SSE \text{ criterion}
\end{aligned}$$

.

The largest likelihood criterion is equivalent to the smallest SSE criterion. But adding new predictors into model always decreases SSE. This criterion will always ends up with selecting all predictors and thus can not be used for the comparison for models with different number of predictors.

(2) AIC, BIC and SBC
To create a criterion for comparing models with different number of predictors, consider

$$r(p) + n\ln\frac{SSE_{(p)}}{n}$$

where the increasing function $r(p)$ is the penalty for the trend of decreasing $SSE_{(p)}$ from a model with $p$ components in $\beta$.

(i) AIC (Akaike Information Criterion) by Akaike (1974)
$AIC = 2p + n\ln\frac{SSE_{(p)}}{n}$. Select the model with smallest AIC.

(ii) BIC (Bayesian Information Criterion) by Sawa (1978)
$BIC = 2(p+2)q - 2q^2 + n\ln\frac{SSE_{(p)}}{n}$ where $q = \frac{n\,MSE}{SSE_{(p)}}$
Select the model with smallest BIC.

(iii) SBC (Schwartz Bayesian Criterion) by Schwartz (1978)
$SBC = p\ln n + n\ln\frac{SSE_{(p)}}{n}$. Select the model with smallest SBC.

**Ex1:** Implementation with SAS

```
proc reg;
     model y=x1 x2 x3/selection=ADJRSQ AIC SBC BIC;
     run;
```

| Number in Model | Adjusted R-Square | R-Square | AIC | BIC | SBC | Variables in Model |
|---|---|---|---|---|---|---|
| 3 | 0.8821 | 0.9116 | -4.8012 | 0.3593 | -2.54139 | x1 x2 x3 |
| 1 | 0.8472 | 0.8599 | -2.8173 | -1.1866 | -1.68740 | x1 |
| 1 | 0.8450 | 0.8579 | -2.6307 | -1.0569 | -1.50075 | x2 |
| 2 | 0.8431 | 0.8693 | -1.7188 | 0.1424 | -0.02397 | x2 x3 |
| 2 | 0.8351 | 0.8626 | -1.0675 | 0.4983 | 0.62735 | x1 x2 |
| 2 | 0.8333 | 0.8611 | -0.9255 | 0.5766 | 0.76932 | x1 x3 |
| 1 | 0.8140 | 0.8295 | -0.2617 | 0.6078 | 0.86820 | x3 |

Model $y = \beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\epsilon$, the model with all available predictors, is selected by $R^2$-criterion, $adj-R^2$ criterion, AIC and SBC, but not by BIC.

2. Forward selection

```
proc reg;
     model y=x1 x2 x3 x4 x5 x6/selection=Forward SLENTRY=0.1;
     run;
```

There are 6 predictors in the pool. We consider models with intercept.

(1) The process
First predictor: In each of 6 models with 1 predictor, test the hypothesis that the coefficient of the predictor is zero. Select one with smallest p-value$< 0.1$.
Second predictor: With the first predictor in, in each of the 5 models with 2 predictors, test the hypothesis that the coefficient of the newly joint predictor is zero and select the one with smallest p-value$< 0.1$.
Third predictor: With the two predictors in, in each of the 4 models with 3 predictors, test the hypothesis that the coefficient of the newly joint predictor is zero and select the one with smallest p-value$< 0.1$
....
The process stops once all predictors are selected, or when none of the predictors remaining can produce ;-value$< 0.1$.

(2) Comments
The process does go through all models. For example, there are 15 models wihth 2 predictors, but only 5 models are investigated.
For selecting models without intercept use "noint".
Forward and SLENTRY can be abbreviated as F and SLE. The default value for SLENTRY is 0.5.

**Ex2:** Implementation with SAS

```
proc reg;
     model y=x1 x2 x3/selection=Forward SLENTRY=0.1;
     run;
```

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


                          Analysis of Variance

                                    Sum of           Mean
Source                    DF       Squares         Square    F Value    Pr > F

Model                      1       55.53721       55.53721      99.83    <.0001
Error                      6        3.33779        0.55630
Corrected Total            7       58.87500


                  Parameter     Standard
      Variable     Estimate        Error   Type II SS  F Value  Pr > F

      Intercept     0.76254      0.33799      2.83159     5.09  0.0649
      x2            0.54515      0.05456     55.53721    99.83  <.0001
-----------------------------------------------------------------------------------------

                          Forward Selection: Step 2
```

Variable x3 Entered: R-Square = 0.9821 and C(p) = 5.1564

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 2 | 57.81946 | 28.90973 | 136.94 | <.0001 |
| Error | 5 | 1.05554 | 0.21111 | | |
| Corrected Total | 7 | 58.87500 | | | |

| Variable | Parameter Estimate | Standard Error | Type II SS | F Value | Pr > F |
|---|---|---|---|---|---|
| Intercept | 0.25074 | 0.25996 | 0.19639 | 0.93 | 0.3791 |
| x2 | 0.85121 | 0.09897 | 15.61697 | 73.98 | 0.0004 |
| x3 | -0.07388 | 0.02247 | 2.28226 | 10.81 | 0.0218 |

---

No other variable met the 0.1000 significance level for entry into the model.

Summary of Forward Selection

| Step | Variable Entered | Number Vars In | Partial R-Square | Model 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 |