Math 3424

Chapter 4. Criteria for choice of best model for linear regression

1. Sequential Variable Selection Procedures

   (a) Forward selection

   - the initial model contains only a constant term
   - enter the variable that produces the largest R^2 of any single regressor
   - add to the model that predictor that meets three equivalent criteria

     i. it has the highest sample partial correlation in absolute value with the response, adjusting for the predictors in the equation already

     ii. adding the variable will increase R^2 more than any other single variable

     iii. the variable added would have the largest t or F statistic of any of the variables that are not already in the model

   The above process continues until, at some stage, the candidate regressor for entry does not exceed a preselected F_{IN}.

   (b) Backward elimination

   The backward elimination begins with all regressors and eliminates one at a time. The variable to be removed is the one that has the smallest t or F value of all the variables in the equation. This is equivalent to removing the variable that causes the smallest change in R^2 or has the smallest absolute partial correlation with Y adjusting for all the other variables left in the model. The procedure is continued until the candidate regressor for removal experiences a partial F value which exceeds the preselected F_{OUT}.

   (c) Stepwise regression

   Stepwise regression provides an important modification of forward selection. At each stage, a regressor can be entered, and another many be eliminated.

2. Best Subset Selection Method

   (a) Cp statistic

   Impact of underfitting

$$\underset{\sim}{y} = \underset{\sim}{X}_1\underset{\sim}{\beta}_1 + \underset{\sim}{\varepsilon} \qquad (p' \text{ parameters})$$

true model:

$$\underset{\sim}{y} = \underset{\sim}{X}_1\underset{\sim}{\beta}_1 + \underset{\sim}{X}_2\underset{\sim}{\beta}_2 + \underset{\sim}{\varepsilon}^* \qquad (m' \text{ parameters}),\ m' > p'$$

$$\begin{aligned}
\text{residual mean square} &= \hat{\sigma}_p^2 \\
&= \frac{1}{n-p'}\underset{\sim}{y}^T[\underset{\sim}{I} - \underset{\sim}{X}_1(\underset{\sim}{X}_1{}^T\underset{\sim}{X}_1)^{-1}\underset{\sim}{X}_1]\underset{\sim}{y} \\
\text{E}(\underset{\sim}{y}) &= \underset{\sim}{X}_1\underset{\sim}{\beta}_1 + \underset{\sim}{X}_2\underset{\sim}{\beta}_2 \\
\text{Var}(\underset{\sim}{y}) &= \sigma^2 I
\end{aligned}$$

$$\begin{aligned}
\text{E}(\hat{\sigma}_p^2) &= \frac{1}{n-p'}[\sigma^2\text{trace}(\underset{\sim}{I} - \underset{\sim}{X}_1(\underset{\sim}{X}_1{}^T\underset{\sim}{X}_1)^{-1}\underset{\sim}{X}_1{}^T) + \\
&\quad (\underset{\sim}{X}_1\underset{\sim}{\beta}_1 + \underset{\sim}{X}_2\underset{\sim}{\beta}_2)^T(\underset{\sim}{I} - \underset{\sim}{X}_1(\underset{\sim}{X}_1{}^T\underset{\sim}{X}_1)^{-1}\underset{\sim}{X}_1{}^T)(\underset{\sim}{X}_1\underset{\sim}{\beta}_1 + \underset{\sim}{X}_2\underset{\sim}{\beta}_2)] \\
&= \sigma^2 + \frac{1}{n-p'}\underset{\sim}{\beta}_2{}^T[\underset{\sim}{X}_2{}^T\underset{\sim}{X}_2 - \underset{\sim}{X}_2{}^T\underset{\sim}{X}_1(\underset{\sim}{X}_1{}^T\underset{\sim}{X}_1)^{-1}\underset{\sim}{X}_1{}^T\underset{\sim}{X}_2]\underset{\sim}{\beta}_2 \\
&\neq \sigma^2
\end{aligned}$$

$$\hat{\sigma}_p'^2 \quad \text{— biased estimator for } \sigma^2$$

$$\begin{aligned}
\mathrm{E}(\hat{\underset{\sim}{\beta}}_1) &= (\underset{\sim}{X}_1{}^T\underset{\sim}{X}_1)^{-1}\underset{\sim}{X}_1{}^T\mathrm{E}(\underset{\sim}{y}) \\
&= (\underset{\sim}{X}_1{}^T\underset{\sim}{X}_1)^{-1}\underset{\sim}{X}_1{}^T(\underset{\sim}{X}_1\underset{\sim}{\beta}_1 + \underset{\sim}{X}_2\underset{\sim}{\beta}_2) \\
&= \underset{\sim}{\beta}_1 + (\underset{\sim}{X}_1{}^T\underset{\sim}{X}_1)^{-1}\underset{\sim}{X}_1{}^T\underset{\sim}{X}_2\underset{\sim}{\beta}_2 \\
&\neq \underset{\sim}{\beta}_1
\end{aligned}$$

$$\hat{\underset{\sim}{\beta}}_1 = (\underset{\sim}{X}_1{}^T\underset{\sim}{X}_1)^{-1}\underset{\sim}{X}_1{}^T\underset{\sim}{y} \quad \text{— biased estimator for} \hat{\underset{\sim}{\beta}}_1$$

$$\hat{y}_{x_0} = \underset{\sim}{x}_{10}^T\hat{\underset{\sim}{\beta}}_1$$

$$\underset{\sim}{x}_0 = \begin{pmatrix} 1 \\ x_{01} \\ x_{02} \\ \vdots \\ x_{0p} \\ x_{0p+1} \\ \vdots \\ x_{0m} \end{pmatrix}$$

So,

$$\begin{aligned}
\mathrm{E}(\hat{y}_{x_0}) &= \mathrm{E}(\underset{\sim}{x}_{10}^T\hat{\underset{\sim}{\beta}}_1) \\
&= \underset{\sim}{x}_{10}^T\mathrm{E}(\hat{\underset{\sim}{\beta}}_1) \\
&= \underset{\sim}{x}_{10}^T\underset{\sim}{\beta}_1 + (\underset{\sim}{X}_1^T\underset{\sim}{X}_1)^{-1}\underset{\sim}{X}_1^T\underset{\sim}{X}_2\underset{\sim}{\beta}_2 \\
&= \underset{\sim}{x}_{10}^T\underset{\sim}{\beta}_1 + \underset{\sim}{x}_{10}^T(\underset{\sim}{X}_1^T\underset{\sim}{X}_1)^{-1}\underset{\sim}{X}_1^T\underset{\sim}{X}_2\underset{\sim}{\beta}_2 \qquad \hat{y}_{x_0} \text{ — biased estimator}
\end{aligned}$$

At x_0, under the true model $\underset{\sim}{y} = \underset{\sim}{X}_1\underset{\sim}{\beta}_1 + \underset{\sim}{X}_2\underset{\sim}{\beta}_2 + \underset{\sim}{\varepsilon}^*$, mean response at $x_0 = \underset{\sim}{x}_{10}^T\underset{\sim}{\beta}_1 + \underset{\sim}{x}_{20}^T\underset{\sim}{\beta}_2$

Impact of overfitting

Assumed model:

$$\underset{\sim}{y} = \underset{\sim}{X}_1\underset{\sim}{\beta}_1 + \underset{\sim}{X}_2\underset{\sim}{\beta}_2 + \underset{\sim}{\varepsilon}^* \qquad (m' \text{ parameters})$$

True model:

$$\underset{\sim}{y} = \underset{\sim}{X}_1\underset{\sim}{\beta}_1 + \underset{\sim}{\varepsilon} \qquad (p' \text{ parameters})$$

$$\begin{aligned}
\text{Res. M.S.} &= \frac{1}{n-m'}\underset{\sim}{y}(\underset{\sim}{I} - \underset{\sim}{X}(\underset{\sim}{X}^T\underset{\sim}{X})^{-1}\underset{\sim}{X}^T)\underset{\sim}{y} \\
\mathrm{E}(\hat{\sigma}_{m'}^2) &= \frac{1}{n-m'}\{\sigma^2\text{trace}(\underset{\sim}{I} - \underset{\sim}{X}(\underset{\sim}{X}^T\underset{\sim}{X})^{-1}\underset{\sim}{X}^T) + (\underset{\sim}{X}_1\underset{\sim}{\beta}_1)^T(\underset{\sim}{I} - \underset{\sim}{X}(\underset{\sim}{X}^T\underset{\sim}{X})^{-1}\underset{\sim}{X}^T)(\underset{\sim}{X}_1\underset{\sim}{\beta}_1)\} \\
&= \sigma^2 + \frac{1}{n-m'}\underset{\sim}{\beta}_1^T\underset{\sim}{X}_1^T(\underset{\sim}{I} - \underset{\sim}{X}(\underset{\sim}{X}^T)\underset{\sim}{X})^{-1}\underset{\sim}{X}^T)\underset{\sim}{X}_1\underset{\sim}{\beta}_1 \\
&= \sigma^2 + \frac{1}{n-m'}\underset{\sim}{\beta}_1^T\{\underset{\sim}{X}_1^T\underset{\sim}{X}_1 - \underset{\sim}{X}_1{}^T\underset{\sim}{X}(\underset{\sim}{X}^T\underset{\sim}{X})^{-1}\underset{\sim}{X}^T\underset{\sim}{X}_1\}\underset{\sim}{\beta}_1 \\
&= \sigma^2 + \frac{1}{n-m'}\underset{\sim}{\beta}_1^T\underset{\sim}{0}\underset{\sim}{\beta}_1 \\
&= \sigma^2 \quad \Rightarrow \quad \hat{\sigma}_{m'}^2 \text{ — unbiased estimator for } \sigma^2
\end{aligned}$$

$$\frac{\hat{\sigma}_{m'}^2(n-m')}{\sigma^2} \sim \chi^2_{(n-m')}$$

$$\begin{aligned}\mathrm{Var}(\frac{\hat{\sigma}^2_{m'}(n-m')}{\sigma^2}) &= 2(n-m') \\ \Rightarrow \mathrm{Var}(\hat{\sigma}^2_{m'}) &= \frac{2\sigma^4}{n-m'} > \frac{2\sigma^4}{n-p'}\end{aligned}$$

$$\hat{\underset{\sim}{\beta}}^{*}\text{— unbiased}$$

$$\mathrm{Var}(\hat{\beta}_i^{\;*}) \;>\; \mathrm{Var}(\hat{\beta}_i)$$

$$\text{At } x_0,\ \hat{y}^{*}_{x_0}\text{— unbiased}$$

$$\mathrm{Var}(\hat{y}^{*}_{x_0}) \;>\; \mathrm{Var}(\hat{y}_{x_0})$$

Cp — Consider the mean square error of $\hat{y}(\underset{\sim}{x}_i)$ (the estimate of $\mathrm{E}(y(\underset{\sim}{x}_i))$)

$$\begin{aligned}\sum_{i=1}^{n}\frac{\mathrm{MSE}(\hat{y}(\underset{\sim}{x}_i))}{\sigma^2} &= \sum_{i=1}^{n}\frac{\mathrm{Var}(\hat{y}(\underset{\sim}{x}_i)) + [\mathrm{bias}(\hat{y}(\underset{\sim}{x}_i))]^2}{\sigma^2} \\ &= \sum_{i=1}^{n}\frac{\mathrm{Var}(\hat{y}(\underset{\sim}{x}_i))}{\sigma^2} + \sum_{i=1}^{n}\frac{[\mathrm{bias}(\hat{y}(\underset{\sim}{x}_i))]^2}{\sigma^2}\end{aligned}$$

Assumed model: p' – parameters

$$\underset{\sim}{y} = \underset{\sim}{X}_1\underset{\sim}{\beta}_1 + \underset{\sim}{\varepsilon}$$

True model: m' – parameters

$$\underset{\sim}{y} = \underset{\sim}{X}_1\underset{\sim}{\beta}_1 + \underset{\sim}{X}_2\underset{\sim}{\beta}_2 + \underset{\sim}{\varepsilon}^{*}$$

$$\begin{aligned}\hat{y}(\underset{\sim}{x}_i) &= \underset{\sim}{x}_{1i}^T\hat{\underset{\sim}{\beta}}_1 \\ \hat{\underset{\sim}{\beta}}_1 &= (\underset{\sim}{X}_1^T\underset{\sim}{X}_1)^{-1}\underset{\sim}{X}_1^T\underset{\sim}{y} \\ \mathrm{Var}(\hat{\underset{\sim}{\beta}}_1) &= (\underset{\sim}{X}_1^T\underset{\sim}{X}_1)^{-1}\sigma^2\end{aligned}$$

$$\begin{aligned}\sum_{i=1}^{n}\frac{\mathrm{Var}(\hat{y}(\underset{\sim}{x}_i))}{\sigma^2} &= \sum_{i=1}^{n}\frac{\mathrm{Var}(\underset{\sim}{x}_i^T\hat{\underset{\sim}{\beta}}_1)}{\sigma^2} \\ &= \sum_{i=1}^{n}\frac{\underset{\sim}{x}_{1i}^T(\underset{\sim}{X}_1^T\underset{\sim}{X}_1)^{-1}\sigma^2\underset{\sim}{x}_{1i}}{\sigma^2} \\ &= \sum_{i=1}^{n}\underset{\sim}{x}_{1i}^T(\underset{\sim}{X}_1^T\underset{\sim}{X}_1)^{-1}\underset{\sim}{x}_{1i} \\ &= \sum_{i=1}^{n}\mathrm{trace}(\underset{\sim}{x}_{1i}^T(\underset{\sim}{X}_1^T\underset{\sim}{X}_1)^{-1}\underset{\sim}{x}_{1i}) \\ &= \sum_{i=1}^{n}\mathrm{trace}(\underset{\sim}{x}_{1i}\underset{\sim}{x}_{1i}^T(\underset{\sim}{X}_1^T\underset{\sim}{X}_1)^{-1}) \\ &= \mathrm{trace}\left(\sum_{i=1}^{n}\underset{\sim}{x}_{1i}\underset{\sim}{x}_{1i}^T(\underset{\sim}{X}_1^T\underset{\sim}{X}_1)^{-1}\right) \\ &= \mathrm{trace}(\underset{\sim}{I}_{p'}) \\ &= p'\end{aligned}$$

$$\begin{aligned}
\mathrm{E}(\hat{y}(\underset{\sim}{x}_i)) &= \underset{\sim}{x}_{1i}^T\left[\underset{\sim}{\beta}_1 + (\underset{\sim}{X}_1^T\underset{\sim}{X}_1)^{-1}\underset{\sim}{X}_1^T\underset{\sim}{X}_2\underset{\sim}{\beta}_2\right] \\
&= \underset{\sim}{x}_{1i}^T\underset{\sim}{\beta}_1 + \underset{\sim}{x}_{1i}^T\underset{\sim}{A}\underset{\sim}{\beta}_2 \qquad \text{where } \underset{\sim}{A} = (\underset{\sim}{X}_1^T\underset{\sim}{X}_1)^{-1}\underset{\sim}{X}_1^T\underset{\sim}{X}_2
\end{aligned}$$

$$\mathrm{E}(y(\underset{\sim}{x}_i)) = \underset{\sim}{x}_{1i}^T\underset{\sim}{\beta}_1 + \underset{\sim}{x}_{2i}^T\underset{\sim}{\beta}_2$$

$$\begin{aligned}
\mathrm{bias}(\hat{y}(\underset{\sim}{x}_i)) &= \mathrm{E}(\hat{y}(\underset{\sim}{x}_i) - \mathrm{E}(y(\underset{\sim}{x}_i)) \\
&= (\underset{\sim}{x}_{1i}^T\underset{\sim}{A} - \underset{\sim}{x}_{2i}^T)\underset{\sim}{\beta}_2
\end{aligned}$$

$$\begin{aligned}
\Big[\mathrm{bias}(\hat{y}(\underset{\sim}{x}_i))\Big]^2 &= \underset{\sim}{\beta}_2^T(\underset{\sim}{x}_{1i}^T\underset{\sim}{A} - \underset{\sim}{x}_{2i}^T)^T(\underset{\sim}{x}_{10}^T\underset{\sim}{A} - \underset{\sim}{x}_{2i}^T)\underset{\sim}{\beta}_2 \\
&= \underset{\sim}{\beta}_2^T(\underset{\sim}{A}^T\underset{\sim}{x}_{1i} - \underset{\sim}{x}_{2i})(\underset{\sim}{x}_{1i}^T\underset{\sim}{A} - \underset{\sim}{x}_{2i}^T)\underset{\sim}{\beta}_2 \\
&= \underset{\sim}{\beta}_2^T(\underset{\sim}{x}_{2i} - \underset{\sim}{A}^T\underset{\sim}{x}_{1i})(\underset{\sim}{x}_{2i}^T - \underset{\sim}{x}_{1i}^T\underset{\sim}{A})\underset{\sim}{\beta}_2 \\
\sum_{i=1}^{n}\frac{[\mathrm{bias}(\hat{y}(\underset{\sim}{x}_i))]^2}{\sigma^2} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}\underset{\sim}{\beta}_2^T(\underset{\sim}{x}_{2i} - \underset{\sim}{A}^T\underset{\sim}{x}_{1i})(\underset{\sim}{x}_{2i}^T - \underset{\sim}{x}_{1i}^T\underset{\sim}{A})\underset{\sim}{\beta}_2 \\
&= \frac{1}{\sigma^2}\underset{\sim}{\beta}_2^T\sum_{i=1}^{n}(\underset{\sim}{x}_{2i} - \underset{\sim}{A}^T\underset{\sim}{x}_{1i})(\underset{\sim}{x}_{2i}^T - \underset{\sim}{x}_{1i}^T\underset{\sim}{A})\underset{\sim}{\beta}_2 \\
&= \frac{1}{\sigma^2}\underset{\sim}{\beta_2}^T[\underset{\sim}{X}_2{}^T\underset{\sim}{X}_2 - \underset{\sim}{X}_2{}^T\underset{\sim}{X}_1(\underset{\sim}{X}_1{}^T\underset{\sim}{X}_1)^{-1}\underset{\sim}{X}_1{}^T\underset{\sim}{X}_2]\underset{\sim}{\beta}_2
\end{aligned}$$

Since

$$\begin{aligned}
\mathrm{E}(\hat{\sigma}_p^2) &= \sigma^2 + \frac{1}{n-p'}\underset{\sim}{\beta_2}^T[\underset{\sim}{X}_2{}^T\underset{\sim}{X}_2 - \underset{\sim}{X}_2{}^T\underset{\sim}{X}_1(\underset{\sim}{X}_1{}^T\underset{\sim}{X}_1)^{-1}\underset{\sim}{X}_1{}^T\underset{\sim}{X}_2]\underset{\sim}{\beta}_2 \\
\Rightarrow &\quad \frac{1}{n-p'}\sum_{i=1}^{n}[\mathrm{bias}(\hat{y}(\underset{\sim}{x}_i))]^2 = \mathrm{E}(\hat{\sigma}_{p'}^2) - \sigma^2
\end{aligned}$$

$$\begin{aligned}
\sum_{i=1}^{n}\frac{\mathrm{MSE}(\hat{y}(\underset{\sim}{x}_i))}{\sigma^2} &= \sum_{i=1}^{n}\frac{\mathrm{Var}(\hat{y}(\underset{\sim}{x}_i))}{\sigma^2} + \sum_{i=1}^{n}\frac{[\mathrm{bias}(\hat{y}(\underset{\sim}{x}_i))]^2}{\sigma^2} \\
&= p' + \frac{1}{\sigma^2}(n-p')(\mathrm{E}(\hat{\sigma}_{p'}^2) - \sigma^2) \\
&= p' + \frac{(\mathrm{E}(\hat{\sigma}_{p'}^2) - \sigma^2)(n-p')}{\sigma^2}
\end{aligned}$$

Define

$$\begin{aligned}
Cp &= p' + \frac{(\hat{\sigma}_{p'}^2 - \hat{\sigma}_{\text{full model}}^2)(n-p')}{\hat{\sigma}_{\text{full model}}^2} \\
&= p' + \frac{\hat{\sigma}_{p'}^2(n-p')}{\hat{\sigma}_{\text{full model}}^2} - (n-p') \\
&= 2p' - n + \frac{RSS_{p'}}{\hat{\sigma}_{\text{full model}}^2}
\end{aligned}$$

(b) <u>Cross validation for model selection</u>

$$n_1\left\{\begin{array}{cccc}
y_1\ x_{11} & x_{12}\ldots & x_{1p} \\
\vdots\ \vdots & \vdots & \vdots \\
y_{n1}\ x_{n1,1} & x_{n1,2}\ldots & x_{n1,p}
\end{array}\right\}\text{fitting} \Rightarrow \text{estimate regression coeff.}\, y_j = \hat{\beta}_0 + \hat{\beta}_1 x_{j1} + \ldots + \hat{\beta}_p x_{jp}$$

$$n_2\left\{\begin{array}{cccc}
y_{n_1+1}\ x_{n_1+1,1} & x_{n_1+1,2}\ldots & x_{n_1+1,p} \\
\vdots\ \vdots & \vdots & \vdots \\
y_{n_1+n_2}\ x_{n_1+n_2,1} & x_{n_1+n_2,2}\ldots & x_{n_1+n_2,p}
\end{array}\right\}\text{validation} \quad \widetilde{y_{n_1+1}} = \hat{\beta}_0 + \hat{\beta}_1 x_{n_1+1,1} + \ldots + \hat{\beta}_p x_{n_1+1,p}$$

$$y_j - \widetilde{y}_j \qquad\qquad j = n_1+1, \ldots, n_1+n_2$$

$$\sum_{j=n_1+1}^{n_1+n_2} (y_j - \widetilde{y}_j)^2$$

$$\sum_{j=n_1+1}^{n_1+n_2} |y_j - \widetilde{y}_j|$$

PRESS statistic
Set $n_1 = n-1, \quad n_2 = 1$. We can find

$$y_i - \widetilde{y}_j,\ j = 1, 2, \ldots, n \quad \text{PRESS residual}$$

$$\begin{aligned} \text{PRESS} &= \sum_{i=1}^{n} (y_i - \widetilde{y}_i)^2 \\ &= \sum_{i=1}^{n} \frac{\hat{e}_i^2}{(1-h_{ii})^2} \end{aligned}$$

Choose the model with the smallest PRESS.

Simple Linear Regression:

$$y_i = \beta_0 + \beta_1 x_i + e_i \qquad\qquad i = 1, \ldots, n$$

$$\underset{\sim}{X}^T \underset{\sim}{X} = \begin{pmatrix} n & \sum_{j=1}^n x_j \\ \sum_{j=1}^n x_j & \sum_{j=1}^n {x_j}^2 \end{pmatrix} \qquad \underset{\sim}{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \qquad \underset{\sim}{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

$$\begin{aligned} (\underset{\sim}{X}^T \underset{\sim}{X})^{-1} &= \frac{1}{n\sum_{j=1}^n (x_j - \bar{x})^2} \begin{pmatrix} \sum_{j=1}^n {x_j}^2 & -\sum_{j=1}^n x_j \\ -\sum_{j=1}^n x_j & n \end{pmatrix} \\ h_{ii} &= \underset{\sim}{x_i}^T (\underset{\sim}{X}^T \underset{\sim}{X})^{-1} \underset{\sim}{x_i} \\ &= \begin{pmatrix} 1 & x_i \end{pmatrix} (\underset{\sim}{X}^T \underset{\sim}{X})^{-1} \begin{pmatrix} 1 \\ x_i \end{pmatrix} \\ &= \frac{1}{n\sum_{j=1}^n (x_j - \bar{x})^2} \begin{pmatrix} 1 & x_i \end{pmatrix} \begin{pmatrix} \sum_{j=1}^n {x_j}^2 & -\sum_{j=1}^n x_j \\ -\sum_{j=1}^n x_j & n \end{pmatrix} \begin{pmatrix} 1 \\ x_i \end{pmatrix} \\ &= \frac{1}{n\sum_{j=1}^n (x_j - \bar{x})^2} \begin{pmatrix} \sum_{j=1}^n {x_j}^2 - x_i n\bar{x} & -n\bar{x} + nx_i \end{pmatrix} \begin{pmatrix} 1 \\ x_i \end{pmatrix} \\ &= \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^n (x_j - \bar{x})^2} \qquad i = 1, 2, \ldots, n \\ \Rightarrow \sum_{i=1}^n h_{ii} &= 2 \quad \text{and} \quad h_{ii} > 1/n \end{aligned}$$

(c) Choose the model with the largest R^2.

(d) Choose the model with the largest R^2_{adj}, where $R^2_{adj} = 1 - \dfrac{MSE|_p}{Total\ S.S./(n-1)}$.

$\Rightarrow$ Choose the model with the smallest MSE.

Remark: R^2_{adj} can be negative.

(e) Choose the model with the smallest AIC, where $\text{AIC} = n\log(\frac{\text{Res.S.S.}}{n}) + 2p'$.

The first part n log(SSEp) measures the goodness of fit of the model, which is penalized by model complexity in the second part 2(p+1). The constant 2 in the penalty term is often referred to as complexity or penalty parameter. In model selection, we calculate AIC value for each model with the same data set, and the "best" model is the one with minimum AIC value.

(f) Choose the model with the smallest BIC, where $\text{BIC} = n\log(\frac{\text{Res.S.S.}}{n}) + p'\log n$.

It applies a larger penalty for overfitting.