# [Criterion 3] Mallow's $C_p$

RSS

$$C_p(\mathcal{M}) = \frac{\mathrm{SSE}(\mathcal{M})}{\hat{\sigma}^2} - n + 2 \times p_{\mathcal{M}}$$

- $\hat{\sigma}^2 = \mathrm{SSE}(\mathcal{F})/df_{\mathcal{F}}$
  - $\mathcal{F}$ denotes the fullest model
  - best estimate of $\sigma^2$

- The criterion is motivated from the Model Error (ME).
  - $\mathrm{ME} = \| \mathrm{E}(Y) - \hat{Y} \|^2$
  - $\mathrm{E}(\mathrm{ME}) = \mathrm{E}(\mathrm{SSE}) + \sigma^2(-n + 2p)$
  - See the derivation on Notes.

# [Criterion 3] Mallow's $C_p$

$$C_p(\mathcal{M}) = \frac{\text{SSE}(\mathcal{M})}{\hat{\sigma}^2} - n + 2 \times p_{\mathcal{M}}$$

- $\hat{\sigma}^2 = \text{SSE}(\mathcal{F})/df_{\mathcal{F}}$
  - $\mathcal{F}$ denotes the fullest model
  - best estimate of $\sigma^2$

- The criterion is motivated from the Model Error (ME).
  - $\text{ME} = \| \underline{\text{E}(Y) - \hat{Y}} \|^2$
  - $\text{E(ME)} = \text{E(SSE)} + \sigma^2(-n + 2p)$
  - See the derivation on Notes.

# [Criterion 3] Mallow's $C_p$

$$C_p(\mathcal{M}) = \frac{\text{SSE}(\mathcal{M})}{\hat{\sigma}^2} - n + 2 \times p_{\mathcal{M}} \quad \checkmark$$

- $\hat{\sigma}^2 = \text{SSE}(\mathcal{F})/df_{\mathcal{F}}$
  - $\mathcal{F}$ denotes the fullest model
  - best estimate of $\sigma^2$

- The criterion is motivated from the Model Error (ME).
  - $\text{ME} = \| \text{E}(Y) - \hat{Y} \|^2$
  - $\text{E(ME)} = \underline{\text{E(SSE)} + \sigma^2(-n + 2p)}$
  - See the derivation on Notes.

- Let $a = E(Y) - \hat{Y}$.

$$E\|a\|^2 = \|E(a)\|^2 + \operatorname{tr}\{\operatorname{var}(a)\}$$

- Thus,

$$\begin{aligned} E(ME) &= \|E(Y) - E(\hat{Y})\|^2 + \operatorname{tr}\{\operatorname{var}(\hat{Y})\} \\ &= \text{bias}^2 + \text{variance} \end{aligned}$$

$$E\|a\|^2 = E\|a - E(a) + E(a)\|^2$$

$$= E\{\|a - E(a)\|^2 + \|E(a)\|^2 + 2\times\{a - E(a)\}^T E(a)\}$$

- Let $a = \underbrace{E(Y)}_{n\times 1} - \underbrace{\hat{Y}}_{n\times 1}$.

$$E\|a\|^2 = \|E(a)\|^2 + \text{tr}\{\text{var}(a)\}$$

E(a) is constant

- Thus,

$$E(ME) = \|E(Y) - E(\hat{Y})\|^2 + \text{tr}\{\text{var}(\hat{Y})\}$$

$$= \text{bias}^2 + \text{variance}$$

① $= E\|a - E(a)\|^2 = E\,\text{tr}\{\{a - E(a)\}\{a - E(a)\}^T\} = \text{tr}\{\text{var}(a)\}$

② $= \|E(a)\|^2$

③ $= 2\times E\{(a - E(a))^T E(a)\} = 2\times \underbrace{E\{a - E(a)\}^T}_{=0} E(a)$

$$E\{a - E(a)\} = E(a) - E(a) = 0$$

- Let $a = \mathrm{E}(Y) - \hat{Y}$.

$E(a) = E(Y) - E(\hat{Y})$

$$\mathrm{E}\,\|a\|^2 = \|\,\mathrm{E}(a)\|^2 + \mathrm{tr}\{\mathrm{var}(a)\}$$

- Thus,

$$\mathrm{E(ME)} = \|\,\mathrm{E}(Y) - \mathrm{E}(\hat{Y})\|^2 + \mathrm{tr}\{\mathrm{var}(\hat{Y})\}$$
$$= \mathrm{bias}^2 + \mathrm{variance}$$

$\mathrm{var}(a) = \mathrm{var}\{E(Y) - \hat{Y}\} = \mathrm{var}(-\hat{Y}) = \mathrm{var}(\hat{Y})$

Want both bias and variance to be small.

# [Criterion 4] AIC: Akaike Information Criterion

$Y = X\beta + \epsilon$

- Definition: For a general model $\mathcal{M}$ with parameter $\theta$,

$$\mathrm{AIC}(\mathcal{M}) = -2\log L_{\mathcal{M}}(\hat{\theta}) + 2 \times p_{\mathcal{M}} \Rightarrow \text{penalty}$$

where $L_{\mathcal{M}}(\hat{\theta})$ denotes the likelihood function of the parameters in the model $\mathcal{M}$ evaluated at the MLE. $\hat{\theta}$

negative $L \Rightarrow$ small

complexity $p \Rightarrow$ small

- Motivation: Kullback-Leibler discrepancy

$$KL(f, g) = \int \log \frac{f(\boldsymbol{y})}{g(\boldsymbol{y}; \theta)} f(\boldsymbol{y}) d\boldsymbol{y}$$

  - A measure of difference between a true fixed $f$ and various competing models $g$ depending on parameter $\theta$.
    - non-symmetric $KL(f, g) \neq KL(g, f)$.
    - $KL(f, g) \geqslant KL(f, f) = 0$.
    - $KL(f, g) = -\int \log g(\boldsymbol{y}; \theta) f(\boldsymbol{y}) d\boldsymbol{y} + \text{constant}$

# [Criterion 4] AIC: Akaike Information Criterion

- Definition: For a general model $\mathcal{M}$ with parameter $\theta$,

$$\mathrm{AIC}(\mathcal{M}) = -2\log L_{\mathcal{M}}(\hat{\theta}) + 2 \times p_{\mathcal{M}}$$

  where $L_{\mathcal{M}}(\hat{\theta})$ denotes the likelihood function of the parameters in the model $\mathcal{M}$ evaluated at the MLE.

- Motivation: Kullback-Leibler discrepancy

KL divergence

$$KL(f, g) = \int \log \frac{f(\boldsymbol{y})}{g(\boldsymbol{y}; \theta)} f(\boldsymbol{y}) d\boldsymbol{y}$$

f: Y

g: Xβ

parameters

  - A measure of difference between a true fixed $f$ and various competing models $g$ depending on parameter $\theta$.
    - non-symmetric $KL(f, g) \neq KL(g, f)$.
    - $KL(f, g) \geqslant KL(f, f) = 0$.
    - $KL(f, g) = -\int \log g(\boldsymbol{y}; \theta) f(\boldsymbol{y}) d\boldsymbol{y} + \text{constant}$

# [Criterion 4] AIC: Akaike Information Criterion

- Definition: For a general model $\mathcal{M}$ with parameter $\theta$,

$$\text{AIC}(\mathcal{M}) = -2 \log L_{\mathcal{M}}(\hat{\theta}) + 2 \times p_{\mathcal{M}}$$

where $L_{\mathcal{M}}(\hat{\theta})$ denotes the likelihood function of the parameters in the model $\mathcal{M}$ evaluated at the MLE.

$\int \log\left(\frac{g}{f}\right) \times g \; d$

- Motivation: Kullback-Leibler discrepancy

$$KL(f, g) = \int \log \frac{f(\boldsymbol{y})}{g(\boldsymbol{y}; \theta)} f(\boldsymbol{y}) d\boldsymbol{y}$$

  - A measure of difference between a true fixed $f$ and various competing models $g$ depending on parameter $\theta$.
    - non-symmetric $KL(f, g) \neq KL(g, f)$.
    - $KL(f, g) \geqslant KL(f, f) = 0$.
    - $KL(f, g) = -\int \log g(\boldsymbol{y}; \theta) f(\boldsymbol{y}) d\boldsymbol{y} + \text{constant}$

# [Criterion 4] AIC: Akaike Information Criterion

- Definition: For a general model $\mathcal{M}$ with parameter $\theta$,

$$\mathrm{AIC}(\mathcal{M}) = -2\log L_{\mathcal{M}}(\hat{\theta}) + 2 \times p_{\mathcal{M}}$$

  where $L_{\mathcal{M}}(\hat{\theta})$ denotes the likelihood function of the parameters in the model $\mathcal{M}$ evaluated at the MLE.

$KL(f,f)$
$= \int \log \frac{f(y)}{f(y)} f(y)\,dy$
$= \int \log(1) \times f(y)\,dy$
$= 0$

- Motivation: Kullback-Leibler discrepancy

$$KL(f,g) = \int \log \frac{f(\boldsymbol{y})}{g(\boldsymbol{y};\theta)} f(\boldsymbol{y})d\boldsymbol{y}$$

  - A measure of difference between a true fixed $f$ and various competing models $g$ depending on parameter $\theta$.
    - non-symmetric $KL(f,g) \neq KL(g,f)$.
    - $KL(f,g) \geqslant KL(f,f) = 0$.
    - $KL(f,g) = -\int \log g(\boldsymbol{y};\theta) f(\boldsymbol{y})d\boldsymbol{y} + \text{constant}$

# [Criterion 4] AIC: Akaike Information Criterion

- Definition: For a general model $\mathcal{M}$ with parameter $\theta$,

$$\text{AIC}(\mathcal{M}) = -2 \log L_{\mathcal{M}}(\hat{\theta}) + 2 \times p_{\mathcal{M}}$$

  where $L_{\mathcal{M}}(\hat{\theta})$ denotes the likelihood function of the parameters in the model $\mathcal{M}$ evaluated at the MLE.

- Motivation: Kullback-Leibler discrepancy

$$KL(f, g) = \int \log \frac{f(\boldsymbol{y})}{g(\boldsymbol{y}; \theta)} f(\boldsymbol{y}) d\boldsymbol{y}$$

  - A measure of difference between a true fixed $f$ and various competing models $g$ depending on parameter $\theta$.
    - non-symmetric $KL(f, g) \neq KL(g, f)$.
    - $KL(f, g) \geqslant KL(f, f) = 0$.
    - $KL(f, g) = -\int \log g(\boldsymbol{y}; \theta) f(\boldsymbol{y}) d\boldsymbol{y} + \text{constant}$

# Modification from KL

① Because $\theta$ is unknow plug in MLE $\hat{\theta}(Y)$ from the observed data $Y$ $\Rightarrow$

$$\Delta := -\int \log g(z; \hat{\theta}(Y)) f(z)\, dz$$

$$= -E_Z \log g(Z; \hat{\theta}(Y)) \quad (Z \text{ independent with } Y)$$

can be used to approximate KL-divergence

② In the MLR with $\sigma^2$ known, next prove

$$E_Y(AIC) = E_Y(2 \times \Delta) + \text{constant}$$

## Proof:

### Step 1. Form of AIC

(1.1) Likelihood of $Y$ under the model $E(Y) = X\beta$ is

$$L_M(\beta) = -\frac{1}{2\sigma^2}\|Y - X\beta\|^2 - \underline{\frac{n}{2}\log(2\pi\sigma^2)} \quad \text{known}$$

(1.2) MLE $\hat{\beta} = (X^TX)^{-1}X^TY$

(1.3)
$$\begin{aligned} AIC &= -2L_M(\hat{\beta}) + 2 \times p \\ &= \|Y - X\hat{\beta}\|^2 + \text{constant} + 2 \times p \\ &= RSS_M + \text{constant} + 2 \times p \end{aligned}$$

## Step 2: $E(AIC)$

By analysis of $C_p$ (Step 3)

we have calculated

$$E(RSS_M) = \mu^T(I - P_p)\mu + (n-p)\sigma^2$$

with $\mu = E(Y)$

$P_p = X_p(X_p^T X_p)^{-1} X_p$ using $p$ covariates

Thus $E(AIC)$

$$= \frac{1}{\sigma^2} E(RSS_M) + 2p + \text{constant}$$

$$= \frac{1}{\sigma^2} \mu^T(I - P_p)\mu + p + n$$

## Step 3: Calculate $\Delta$

- $Z$ denotes a random matrix independent with $Y$ but follows the same distribution as $Y$
- So we have $Z = \mu + \epsilon_Z$

$$\Delta = -E_Z \log g(Z; \hat{\theta}(Y))$$

$$= \frac{1}{2\sigma^2} E_Z \| Z - X \cdot \hat{\beta}(Y) \|^2 + \text{constant}$$

$$= \frac{1}{2\sigma^2} E_Z \| \mu + \epsilon_Z - X \hat{\beta}(Y) \|^2 + \text{constant}$$

$$= \frac{1}{2\sigma^2} \left\{ \| \mu - X\hat{\beta}(Y) \|^2 + E(\epsilon_Z^T \epsilon_Z) \right\}$$

$$= \frac{1}{2\sigma^2} \left( \| \mu - X\hat{\beta}(Y) \|^2 + n\sigma^2 \right)$$

Step 4: Calculate $E_Y(2\Delta)$

$$E_Y(2\Delta) = \frac{1}{\sigma^2} E_Y\left[\|\mu - X\hat{\beta}(Y)\|^2\right] + n$$

$$= \frac{1}{\sigma^2} E_Y\left[\|\mu - PY\|^2\right] + n$$

(same as ME in $C_p$)

$$= \frac{1}{\sigma^2} \mu^T (I - P)\mu + p\sigma^2$$

In summary,

$$E_Y(2\Delta) = \frac{1}{\sigma^2}\mu^T(I-P)\mu + p$$

$$E_Y(AIC) = \frac{\mu^T(I-P)\mu}{\sigma^2} + (n-p) + 2p$$

$n$ is fixed, does not influence model selection

## AIC under multiple linear models

- if $\sigma^2$ is known,

$$\text{AIC} = \frac{\text{SSE}_{\mathcal{M}}}{\sigma^2} + 2p_{\mathcal{M}}.$$

  - Similar to $C_p$ if replace $\sigma^2$ by $\hat{\sigma}^2$ (only differ by $-n$)

- if $\sigma^2$ is unknown,

$$\text{AIC} = n \log(\text{SSE}_{\mathcal{M}} / n) + 2p_{\mathcal{M}}$$

## AIC under multiple linear models

- if $\sigma^2$ is known,

$$\text{AIC} = \frac{\text{SSE}_{\mathcal{M}}}{\sigma^2} + 2p_{\mathcal{M}}.$$

  - Similar to $C_p$ if replace $\sigma^2$ by $\hat{\sigma}^2$ (only differ by $-n$)

- if $\sigma^2$ is unknown,

$$\text{AIC} = n \log(\text{SSE}_{\mathcal{M}} / n) + 2p_{\mathcal{M}}$$

# [Criterion 5] BIC: Bayesian Information Criterion

For a general model $\mathcal{M}$ with parameter $\theta$,

$$\text{BIC}(\mathcal{M}) = -2 \log L_{\mathcal{M}}(\hat{\theta}) + \log(n) \times p_{\mathcal{M}}$$

where $L_{\mathcal{M}}(\hat{\theta})$ denotes the likelihood function of the parameters in the model $\mathcal{M}$ evaluated at the MLE.

- BIC penalizes larger models more heavily and so will tend to prefer smaller models in comparison to AIC.

# [Criterion 5] BIC: Bayesian Information Criterion

For a general model $\mathcal{M}$ with parameter $\theta$,

$$\mathrm{BIC}(\mathcal{M}) = -2\log L_{\mathcal{M}}(\hat{\theta}) + \log(n) \times p_{\mathcal{M}}$$

where $L_{\mathcal{M}}(\hat{\theta})$ denotes the likelihood function of the parameters in the model $\mathcal{M}$ evaluated at the MLE.

- BIC penalizes larger models more heavily and so will tend to prefer smaller models in comparison to AIC.

## BIC is derived under the Bayesian perspective

- Consider the multiple linear model with $\sigma^2$ known.
- Suppose we fit a submodel with $\mathbf{X}_p\beta_p$
  - $p$ can be smaller than the total number of covariates
  - Assume $\beta$ has prior distribution $N_p(\mathbf{m}, \sigma^2 V)$
  - The log posterior distribution of $\beta_p$ is proportional to

$$\text{BIC} = \frac{\text{SSE}_{\mathcal{M}}}{\sigma^2} + \log(n)p_{\mathcal{M}}$$

- Detailed proof: read Linear Regression Analysis (Lee and Seber) Section 12.3.4

## BIC is derived under the Bayesian perspective

- Consider the multiple linear model with $\sigma^2$ known.
- Suppose we fit a submodel with $\mathbf{X}_p \beta_p$
  - $p$ can be smaller than the total number of covariates
  - Assume $\beta$ has prior distribution $N_p(\mathbf{m}, \sigma^2 V)$
  - The log posterior distribution of $\beta_p$ is proportional to

$$\text{BIC} = \frac{\text{SSE}_{\mathcal{M}}}{\sigma^2} + \log(n) p_{\mathcal{M}}$$

- Detailed proof: read Linear Regression Analysis (Lee and Seber) Section 12.3.4

## BIC is derived under the Bayesian perspective

- Consider the multiple linear model with $\sigma^2$ known.
- Suppose we fit a submodel with $\mathbf{X}_p\beta_p$
  - $p$ can be smaller than the total number of covariates
  - Assume $\beta$ has prior distribution $N_p(\mathbf{m}, \sigma^2 V)$
  - The log posterior distribution of $\beta_p$ is proportional to

$$\text{BIC} = \frac{\text{SSE}_{\mathcal{M}}}{\sigma^2} + \log(n)p_{\mathcal{M}}$$

- Detailed proof: read Linear Regression Analysis (Lee and Seber) Section 12.3.4

## BIC is derived under the Bayesian perspective

- Consider the multiple linear model with $\sigma^2$ known.
- Suppose we fit a submodel with $\mathbf{X}_p\boldsymbol{\beta}_p$
  - $p$ can be smaller than the total number of covariates
  - Assume $\boldsymbol{\beta}$ has prior distribution $N_p(\mathbf{m}, \sigma^2 V)$
  - The log posterior distribution of $\boldsymbol{\beta}_p$ is proportional to

$$\text{BIC} = \frac{\text{SSE}_{\mathcal{M}}}{\sigma^2} + \log(n)p_{\mathcal{M}}$$

- Detailed proof: read Linear Regression Analysis (Lee and Seber) Section 12.3.4

## BIC is derived under the Bayesian perspective

- Consider the multiple linear model with $\sigma^2$ known.
- Suppose we fit a submodel with $\mathbf{X}_p\beta_p$
  - $p$ can be smaller than the total number of covariates
  - Assume $\beta$ has prior distribution $N_p(\mathbf{m}, \sigma^2 V)$
  - The log posterior distribution of $\beta_p$ is proportional to

$$\text{BIC} = \frac{\text{SSE}_{\mathcal{M}}}{\sigma^2} + \log(n)p_{\mathcal{M}}$$

- Detailed proof: read Linear Regression Analysis (Lee and Seber) Section 12.3.4

# Summary

- $R^2$: motivated from $\text{corr}^2(\hat{Y}, Y)$ (prefer larger)
- Adjusted $R^2$: penalizes model complexity (prefer larger)
- $C_p$: motivated from $\| \mathrm{E}(Y) - \hat{Y} \|^2$ (prefer smaller)
- AIC and BIC: motivated from KL divergence (prefer smaller)

# Summary

- $R^2$: motivated from $\mathrm{corr}^2(\hat{Y}, Y)$ (prefer larger)
- Adjusted $R^2$: penalizes model complexity (prefer larger)
- $C_p$: motivated from $\| \mathrm{E}(Y) - \hat{Y} \|^2$ (prefer smaller)
- AIC and BIC: motivated from KL divergence (prefer smaller)

# Summary

- $R^2$: motivated from $\mathrm{corr}^2(\hat{Y}, Y)$ (prefer larger)
- Adjusted $R^2$: penalizes model complexity (prefer larger)
- $C_p$: motivated from $\| \mathrm{E}(Y) - \hat{Y} \|^2$ (prefer smaller)
- AIC and BIC: motivated from KL divergence (prefer smaller)

# Summary

- $R^2$: motivated from $\mathrm{corr}^2(\hat{Y}, Y)$ (prefer larger)
- Adjusted $R^2$: penalizes model complexity (prefer larger)
- $C_p$: motivated from $\| \mathrm{E}(Y) - \hat{Y} \|^2$ (prefer smaller)
- AIC and BIC: motivated from KL divergence (prefer smaller)

# [MS 3] Criterion-based Procedures

# 3.1 Best Subset Selection

1. Let $\mathcal{M}_0$ denote the null model, which contains no predictors. This model simply predicts the $\underbrace{\text{sample mean}}_{\bar{Y}}$ for each observation.

2. For $k = 1, 2, \ldots, p$:

   (a) Fit all $\binom{p}{k}$ models that contain exactly $k$ predictors.

   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here best is defined as having the smallest RSS = SSE, or equivalently largest $R^2$.

3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using $C_p$, AIC, BIC, or adjusted $R^2$.

# 3.1 Best Subset Selection

1. Let $\mathcal{M}_0$ denote the null model, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \ldots, p$:
   (a) Fit all $\binom{p}{k}$ models that contain exactly $k$ predictors.
   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here best is defined as having the smallest RSS = SSE, or equivalently largest $R^2$.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using $C_p$, AIC, BIC, or adjusted $R^2$.

# 3.1 Best Subset Selection

1. Let $\mathcal{M}_0$ denote the null model, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \ldots, p$:
   (a) Fit all $\binom{p}{k}$ models that contain exactly $k$ predictors.
   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here best is defined as having the smallest RSS $=$ SSE, or equivalently largest $R^2$.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using $C_p$, AIC, BIC, or adjusted $R^2$.

# 3.1 Best Subset Selection

1. Let $\mathcal{M}_0$ denote the null model, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \ldots, p$:
   (a) Fit all $\binom{p}{k}$ models that contain exactly $k$ predictors.
   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here best is defined as having the smallest $\text{RSS} = \text{SSE}$, or equivalently largest $R^2$.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using $C_p$, AIC, BIC, or adjusted $R^2$.

# 3.2 Stepwise Selection: Forward

1. Let $\mathcal{M}_0$ denote the null model, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 0, 1, \ldots, p-1$:
   (a) Consider all $p - k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
   (b) Choose the best among these $p - k$ models, and call it $\mathcal{M}_{k+1}$. Here best is defined as having smallest RSS or highest $R^2$.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using $C_p$, AIC, BIC, or adjusted $R^2$.

# 3.2 Stepwise Selection: Forward

1. Let $\mathcal{M}_0$ denote the null model, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 0, 1, \ldots, p-1$:
   (a) Consider all $p-k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
   (b) Choose the best among these $p-k$ models, and call it $\mathcal{M}_{k+1}$. Here best is defined as having smallest RSS or highest $R^2$.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using $C_p$, AIC, BIC, or adjusted $R^2$.

# 3.2 Stepwise Selection: Forward

1. Let $\mathcal{M}_0$ denote the null model, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 0, 1, \ldots, p - 1$:
   (a) Consider all $p - k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
   (b) Choose the best among these $p - k$ models, and call it $\mathcal{M}_{k+1}$. Here best is defined as having smallest RSS or highest $R^2$.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using $C_p$, AIC, BIC, or adjusted $R^2$.

# 3.2 Stepwise Selection: Forward

1. Let $\mathcal{M}_0$ denote the null model, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 0, 1, \ldots, p - 1$:
   (a) Consider all $p - k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
   (b) Choose the best among these $p - k$ models, and call it $\mathcal{M}_{k+1}$. Here best is defined as having smallest RSS or highest $R^2$.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using $C_p$, AIC, BIC, or adjusted $R^2$.