6. Model Selection

[MS 1] Why?

[MS 1] Why?

1. Model Interpretability.
   - Often not all predictors are associated with the response.

   $$Y = \beta_0 + \beta_A X_A + \beta_I X_I + \epsilon \quad (\beta_A \neq 0,\ \beta_I = 0)$$

     - Discover associated predictors X_A.
     - Remove irrelevant predictors X_I; reduce unnecessary complexity.
   - Occam's razor: Prefer models easier to interpret.
2. Prediction Accuracy.
   - Models are evaluated by prediction accuracy.

[MS 1] Why?

body fat ← measurements

1. Model Interpretability.
   - Often not all predictors are associated with the response.

$$Y = \beta_0 + \beta_A X_A + \beta_I X_I + \epsilon \quad (\beta_A \neq 0,\ \beta_I = 0)$$

     - Discover associated predictors X_A. ✓
     - Remove irrelevant predictors X_I; reduce unnecessary complexity.
   - Occam's razor: Prefer models easier to interpret.
2. Prediction Accuracy.
   - Models are evaluated by prediction accuracy.

[MS 1] Why?

1. Model Interpretability.
   - Often not all predictors are associated with the response.

   $$Y = \beta_0 + \beta_A X_A + \beta_I X_I + \epsilon \quad (\beta_A \neq 0,\ \beta_I = 0)$$

     - Discover associated predictors X_A.
     - Remove irrelevant predictors X_I; reduce unnecessary complexity.
   - Occam's razor: Prefer models easier to interpret.
2. Prediction Accuracy.
   - Models are evaluated by prediction accuracy.

[MS 1] Why?

1. Model Interpretability.
   - Often not all predictors are associated with the response.

$$Y = \beta_0 + \beta_A X_A + \beta_I X_I + \epsilon \quad (\beta_A \neq 0,\ \beta_I = 0)$$

     - Discover associated predictors X_A.
     - Remove irrelevant predictors X_I; reduce unnecessary complexity.
   - Occam's razor: Prefer models easier to interpret.
2. Prediction Accuracy.
   - Models are evaluated by prediction accuracy.

[MS 2] Criteria

[Criterion 1] Coefficient of multiple determination R^2

goodness-of-fit

corr(Y, $\hat{Y}$)

Given a particular model $\mathcal{M}$,

$$R^2(\mathcal{M}) = 1 - \frac{\text{SSE}(\mathcal{M})}{\text{SST}}$$

- $\text{SSE}(\mathcal{M}) = \|\underline{Y} - \underline{\hat{Y}_{\mathcal{M}}}\|^2$: SSE of the model $\mathcal{M}$
- $\text{SST} = \|\underline{Y - \bar{Y}}\|^2$: sum of squares total

- Always increases as the model size increases.
  - Tends to prefer a larger model.
- Can be used to compare 2 models of the same size.

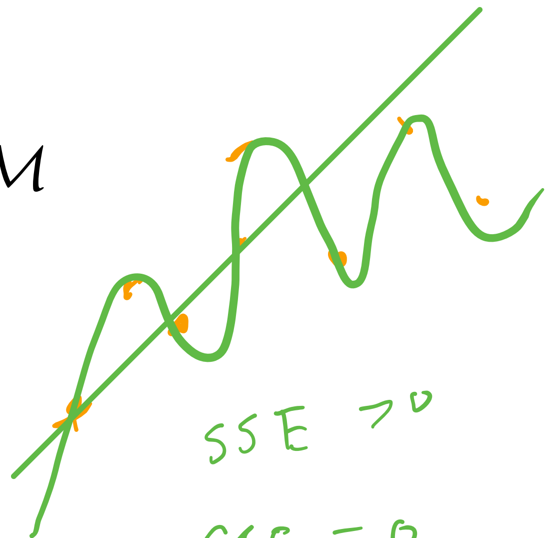
[Criterion 1] Coefficient of multiple determination R^2

Given a particular model $\mathcal{M}$,

$$R^2(\mathcal{M}) = 1 - \frac{\text{SSE}(\mathcal{M})}{\text{SST}}$$

- $\text{SSE}(\mathcal{M}) = \|Y - \hat{Y}_{\mathcal{M}}\|^2$: SSE of the model $\mathcal{M}$
- $\text{SST} = \|Y - \bar{Y}\|^2$: sum of squares total

SSE $\downarrow$



SSE > 0

SSE $= 0$

- Always increases as the model size increases.
  - Tends to prefer a larger model.
- Can be used to compare 2 models of the same size.

[Criterion 2] Adjusted R^2

df. of $\mathcal{M}$

$$R^2_{adj}(\mathcal{M}) = 1 - \frac{\text{SSE}(\mathcal{M})/(n - p_{\mathcal{M}})}{\text{SST}\,/(n-1)} = 1 - \frac{\hat{\sigma}^2_{\mathcal{M}}}{\hat{\sigma}^2_{\text{null}}}$$

df. of null model (no covariates)

where $p_{\mathcal{M}}$ is the number of parameters in the model $\mathcal{M}$

- When $p_{\mathcal{M}}$ increases,
  - SSE($\mathcal{M}$) always decreases,
  - $\hat{\sigma}^2_{\mathcal{M}}$ could increase or decrease.
  - $R^2_{adj}(\mathcal{M})$ could increase or decrease.
- Prefer models with larger $R^2_{adj}(\mathcal{M})$
  - A goodness-of-fit measure.

[Criterion 2] Adjusted R^2

$$R^2_{adj}(\mathcal{M}) = 1 - \frac{\mathrm{SSE}(\mathcal{M})/(n - p_{\mathcal{M}})}{\mathrm{SST}\,/(n-1)} = 1 - \frac{\hat{\sigma}^2_{\mathcal{M}}}{\hat{\sigma}^2_{\text{null}}}$$

where $p_{\mathcal{M}}$ is the number of parameters in the model $\mathcal{M}$

- When $p_{\mathcal{M}}$ increases,
  - $\mathrm{SSE}(\mathcal{M})$ always decreases,
  - $\hat{\sigma}^2_{\mathcal{M}}$ could increase or decrease.
  - $R^2_{adj}(\mathcal{M})$ could increase or decrease.
- Prefer models with larger $R^2_{adj}(\mathcal{M})$
  - A goodness-of-fit measure.

[Criterion 2] Adjusted R^2

$$R^2_{adj}(\mathcal{M}) = 1 - \frac{\text{SSE}(\mathcal{M})/(n - p_{\mathcal{M}})}{\text{SST}/(n-1)} = 1 - \frac{\hat{\sigma}^2_{\mathcal{M}}}{\hat{\sigma}^2_{\text{null}}}$$

where $p_{\mathcal{M}}$ is the number of parameters in the model $\mathcal{M}$

- When $p_{\mathcal{M}}$ increases,
  - SSE($\mathcal{M}$) always decreases,
  - $\hat{\sigma}^2_{\mathcal{M}}$ could increase or decrease.
  - $R^2_{adj}(\mathcal{M})$ could increase or decrease.
- Prefer models with larger $R^2_{adj}(\mathcal{M})$
  - A goodness-of-fit measure.

[Criterion 2] Adjusted R^2

$$R^2_{adj}(\mathcal{M}) = 1 - \frac{\text{SSE}(\mathcal{M})/(n - p_{\mathcal{M}})}{\text{SST}/(n-1)} = 1 - \frac{\hat{\sigma}^2_{\mathcal{M}}}{\hat{\sigma}^2_{\text{null}}}$$

where $p_{\mathcal{M}}$ is the number of parameters in the model $\mathcal{M}$

$$1 - \frac{n-1}{n-p_{\mathcal{M}}} \times \frac{SSE}{SST}$$

$\Downarrow$

correction

penalty on model complexity

- When $p_{\mathcal{M}}$ increases,
  - $\text{SSE}(\mathcal{M})$ always decreases,
  - $\hat{\sigma}^2_{\mathcal{M}}$ could increase or decrease.
  - $R^2_{adj}(\mathcal{M})$ could increase or decrease.
- Prefer models with larger $R^2_{adj}(\mathcal{M})$
  - A goodness-of-fit measure.

$p_{\mathcal{M}} \uparrow$

[Criterion 2] Adjusted R^2

$$R^2_{adj}(\mathcal{M}) = 1 - \frac{\text{SSE}(\mathcal{M})/(n - p_{\mathcal{M}})}{\text{SST}\,/(n-1)} = 1 - \frac{\hat{\sigma}^2_{\mathcal{M}}}{\hat{\sigma}^2_{\text{null}}}$$

where $p_{\mathcal{M}}$ is the number of parameters in the model $\mathcal{M}$

- When $p_{\mathcal{M}}$ increases,
  - $\text{SSE}(\mathcal{M})$ always decreases,
  - $\hat{\sigma}^2_{\mathcal{M}}$ could increase or decrease.
  - $R^2_{adj}(\mathcal{M})$ could increase or decrease.
- Prefer models with larger $R^2_{adj}(\mathcal{M})$
  - A goodness-of-fit measure.

[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 σ^2

$X_1 \cdots X_{10}$

- The criterion is motivated from Model Error (ME)
  - $\text{ME} = \| E(Y \mid X) - \hat{Y} \|^2$

[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 σ^2

$E(Y|X) = X\beta$

$\hat{Y} = X\hat{\beta}$

- The criterion is motivated from Model Error (ME)
  - $\text{ME} = \| \text{E}(Y \mid X) - \hat{Y} \|^2$

prediction to mean value of the design matrix X

Step 1: ME form

$$ME = \| \mu - P\mu + P(Y-\mu) \|^2 \qquad P = X(X^T X)^{-1} X^T$$

$$\star = \|(I-P)\mu\|^2 + \|P(Y-\mu)\|^2 + 0$$

$\Rightarrow$ cross term $=0$ by $P(I-P)=0$

$$= \|(I-P)\mu\|^2 + \|P\epsilon\|^2$$

(by $Y = \mu + \epsilon$)

$$= \mu^T (I-P)^T (I-P)\mu + \epsilon^T P^T P \epsilon$$

$$= \mu^T (I-P)\mu + \epsilon^T P \epsilon$$

Step 2: Calculate expectation of ME

By $E(\epsilon^T P \epsilon) = E\{ tr(P\epsilon\epsilon^T)\} = tr\{P E(\epsilon\epsilon^T)\}$

$$= tr\{P \times \sigma^2 I_n\} = \sigma^2 tr(P) = p$$

(By $tr(P) = p$ Notes Sep 26)

$$E(ME) = \mu^T(I-P)\mu + E(\epsilon^T P \epsilon)$$

$$= \mu^T (I-P)\mu + p\sigma^2$$

Step 3: Expectation of SSE

If we fit a model with p parameters, let

$P_p = X_p (X_p^T X_p)^{-1} X_p$ corresponding hat matrix.

RSS_p denote the results of residual sum of squares

$$E[RSS_p] = E[Y^T(I-P_p)Y]$$

$$= E[(\mu+\epsilon)^T(I-P_p)(\mu+\epsilon)]$$

$$= \mu^T(I-P_p)\mu + E\cancel{(\epsilon^T(I-P_p)\mu)}$$

$$+ E\cancel{\{\mu^T(I-P_p)\mu\}} + E\{\epsilon^T(I-P_p)\epsilon\}$$

$$= \mu^T(I-P_p)\mu + tr(I-P_p)\times\sigma^2$$

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

Then $\dfrac{E(ME)}{\sigma^2} = \dfrac{E(RSS_p)}{\sigma^2} + 2p-n$

Plug in $\hat{\sigma}^2$ obtaned from fullest model

$$\Rightarrow \quad C_p = \frac{RSS_p}{\hat{\sigma}^2} + 2p - n$$

Comment 1 on C_p:

By $$E(C_p) \approx \frac{E(RSS_p)}{\sigma^2} + 2p - n$$

$$= \underbrace{\frac{\mu^T(I-P_p)\mu}{\sigma^2}}_{\text{If} \to 0} + n - p + 2p - n$$

$$\approx 0 + n - p + 2p - n = p$$

Thus select model with small C_p and $C_p \approx p$.

Commen 2 on ME:

Note $\mu = E(Y)$ & $P\mu = PE(Y) = E(PY) = E(\hat{Y})$

In ☆, we obtain

$$
\begin{aligned}
ME &= \|\mu - P\mu\|^2 + \|PY - P\mu\|^2 \\
&= \underbrace{\|E(Y) - E(\hat{Y})\|^2}_{① \text{ (Fixed)}} + \underbrace{\|\hat{Y} - E(\hat{Y})\|^2}_{② \text{ (Random)}}
\end{aligned}
$$

Let $a = \hat{Y} - E(\hat{Y})$ satisfying $E(a) = 0$

$E(②) = E(a^T a) = E\{tr(aa^T)\} = tr\{E(aa^T)\} = tr\{Var(a)\}$

$E(ME) = ① + E(②) = \text{Total Bias} + \text{Total Variance}$