1 Model Selection

1.1 Introduction

- It won't always be obvious **which predictors** should be **included** in a multiple regression model and **which** should be **omitted**.

  When the predictors are *uncorrelated*, it's safe to simply drop from the model those whose coefficients aren't statistically significant according to the ***t* tests**.

  More often, though, there will be some degree of multicollinearity among the predictors, and in this case special ***model selection*** procedures should be used.

  The goal is to find a model that accomplishes *both* of two *competing* objectives:

  1. The model should **fit** the data **well**.
  2. The model should be **parsimonious** (contain only a small number of predictors).

  The challenge is that there's a **tradeoff** – the more parsimonious the model, the less well it fits the data.

- We'll look at several **model selection criteria** for comparing models:

  1. R_p^2 and SSE_p
  2. $R_{a,p}^2$ and MSE_p
  3. AIC_p and BIC_p (or SBC_p)
  4. PRESS_p

- We'll use the following notation:

  $\boldsymbol{P-1}$ = The total number of predictors **available** for inclusion in a model.
  $\boldsymbol{p-1}$ = The number of predictors **in** the model **currently being considered** (so $p-1 \le P-1$).

1.2 R_p^2 and SSE_p

- $\boldsymbol{R_p^2}$ and $\textbf{SSE}_{\boldsymbol{p}}$ are just the usual **coefficient of multiple determination** and **error sum of squares** (Class Notes 11), i.e.

$$\text{SSE}_p = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

and

$$R_p^2 \;=\; 1 - \frac{\text{SSE}_p}{\text{SSTO}}\,,$$

except now, because they'll be used to compare models with **different** numbers of parameters, the number of parameters in the model (p) is explicitly represented in the notation.

- Recall that a **small** SSE_p and **large** R_p^2 indicate that the model **fits** the data **well**.
- But recall also that SSE_p *always increases* and R_p^2 *decreases* when a predictor is dropped from a model. So **neither** of these is useful for comparing two models that have **different** numbers of predictors.

1.3 $R^2_{a,p}$ and MSE_p

- $\boldsymbol{R^2_{a,p}}$ and $\textbf{MSE}_{\boldsymbol{p}}$ are just the **adjusted coefficient of multiple determination** and usual **mean squared error** (Class Notes 11), i.e.

$$\text{MSE}_p \;=\; \frac{\text{SSE}_p}{n-p}$$

and

$$R^2_{a,p} \;=\; 1 - \frac{\text{SSE}_p/(n-p)}{\text{SSTO}/(n-1)}\,,$$

except now the number of parameters in the model (p) is explicitly represented in the notation.

- Recall that a **small** MSE_p and **large** $R^2_{a,p}$ indicate that the model **fits** the data **well**.
- These criteria take into account the number of predictors in the model, so that they're **useful** for comparing two models that have **different** numbers of predictors.

Using these criteria, the model that has **larger** $R^2_{a,p}$ or, equivalently, **smaller** MSE_p is **preferred**.

1.4 AIC_p and BIC_p

- ***Akaike's Information Criterion***, denoted $\textbf{AIC}_{\boldsymbol{p}}$, is defined as

Akaike's Information Criterion:

$$\text{AIC}_p \;=\; n\log\text{SSE}_p \;-\; n\log n \;+\; 2p$$

- The ***Bayesian Information Criterion***, denoted $\mathbf{BIC}_p$ is defined as

> **Bayesian Information Criterion**:
>
> $$\mathrm{BIC}_p \;=\; n \log \mathrm{SSE}_p - n \log n + (\log n)\, p$$

- For both AIC_p and BIC_p:

  1. The first term $\boldsymbol{n \log \mathrm{SSE}_p}$ will be **small** if the model **fits** the data **well**.
  2. The second term $\boldsymbol{n \log n}$ is **constant** for fixed n (i.e. it doesn't depend on the how many predictors are in the model or on how well the model fits the data).
  3. The last term $\boldsymbol{2p}$ or $\boldsymbol{(\log n)\, p}$ will be **small** if the model is **parsimonious** (i.e. if the number of predictors in the model, $\boldsymbol{p-1}$, is small).

  Using these criteria, the model that has **smaller** AIC_p (or BIC_p) is **preferred**, and accomplishes better the *two* competing objectives of Subsection 1.1.

  Note that the term $\boldsymbol{2p}$ in AIC_p (and $\boldsymbol{(\log n)\, p}$ in BIC_p) acts as a ***penalty*** for including too many predictors in the model.

1.5 $\mathbf{PRESS}_p$

- The idea behind $\mathbf{PRESS}_p$ is to successively **delete one observation** (row) at a time from the data set, **fit** a given **model** to the **remaining $\boldsymbol{n-1}$ observations**, and for each fitted model calculate the ***delete-one prediction error***

$$\text{Delete-one Prediction Error} \;=\; Y_i - \hat{Y}_{i(i)}\,,$$

  where $\hat{Y}_{i(i)}$ is the predicted value for the deleted Y_i based on the model fitted to the other $n-1$ observations.

  $\mathbf{PRESS}_p$ is the **sum** of **squared delete-one prediction errors**:

> $\mathbf{PRESS}_p$:
>
> $$\mathrm{PRESS}_p \;=\; \sum_{i=1}^{n} (Y_i - \hat{Y}_{i(i)})^2$$

  Using this criteria, the model that has **smaller** $\mathbf{PRESS}_p$ is **preferred**.