Outline

Purposes of Model Selection

- Recall a multiple linear regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i, \quad \varepsilon_i \sim \text{iid } N(0, \sigma^2),$$

  can any of the $p-1$ explanatory variables be dropped to simplify the model?
- If the purpose is description/explanation/understanding, then
  - Parsimony is a key idea.
  - **Occam's razor**: All things being equal, the simplest solution tends to be the right one.
- If the purpose is prediction, then
  - Models are evaluated by predictive accuracy/power.

Approaches to Model Selection

- Goal: Find the subset of explanatory variables that gives the “best" model or identify the subset of explanatory variables for further study.
- This is an “unsolved” problem in statistics: there are no magic procedures to get you the “best model.”
- Why not compute

$$t_k^* = \frac{\hat{\beta}_k}{\widehat{\text{SD}}(\hat{\beta}_k)}, \quad k = 1, \ldots, p-1$$

and drop the explanatory variables with large p-values?
Important explanatory variables may be dropped due to multicollinearity.
A good search algorithm should be able to address this issue.

Model Selection: General

- In some sense, model selection is "data mining."
- Data miners/machine learners often work with very many predictors
- To "implement" this, we need
  - A criterion or benchmark to compare two models.
  - A search strategy
- With a limit number of predictors, it is possible to search all possible models.
- The total number of possible regression models is 2^{p-1} for $p-1$ explanatory variables.

Outline

All Possible Subsets Methods

- Main idea:
  - Choose a model selection criterion.
  - Fit all possible regression models and compute the criterion for each model.
  - Pick the “best" model or a few “good" models according to the criterion.
  - Perform model diagnostics and take remedial measures if needed before determining the final model(s).
- Model selection criterion:
  - Coefficient of multiple determination R_p^2 criterion.
  - Adjusted Coefficient of multiple determination $R_{a,p}^2$ criterion.
  - Mallow’s C_p criterion.
  - AIC_p and BIC_p criteria.

R_p^2 Criterion

- Recall coefficient of multiple determination

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

  where p is the number of parameters (1 intercept and $p-1$ slopes) and $p = 1, 2, \ldots$.
- Not a good criterion. Always increase with model size $\rightarrow$ "optimum" is to take the biggest model

$R^2_{a,p}$ Criterion

- Recall adjusted coefficient of multiple determination

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

  where p is the number of parameters (1 intercept and $p-1$ slopes).
- When p increases, $R^2_{a,p}$ can increase or decrease.
- Possible procedures:
  - Select the models with the maximum $R^2_{a,p}$ or close to the maximum.
  - For fixed p, find the model with the largest $R^2_{a,p}$.
- The $R^2_{a,p}$ criterion is equivalent to the MSE_p criterion.

Mallow's C_p Criterion

Mallow's C_p

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

- $\hat{\sigma}^2 = SSE(F)/df_f$ is the "best" estimate of σ^2 we have (use the fullest model).
- $\text{SSE}(\mathcal{M}) = \left\| Y - \hat{Y}_{\mathcal{M}} \right\|^2$ is the SSE of the model $\mathcal{M}$.
- $p(\mathcal{M})$ is the number of predictors in $\mathcal{M}$, or the degrees of freedom used up by the model.
- Based on an estimate of:

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} \mathbb{E}(Y_i - \mathbb{E}\hat{Y}_i)^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} \mathbb{E}(Y_i - \hat{Y}_i)^2 + \text{Var}(\hat{Y}_i)$$

- Select the model that has small C_p value and $C_p \approx p$.

AIC_p and BIC_p Criteria

- Mallow's C_p is (almost) a special case of

Akaike Information Criterion (AIC)

$$\text{AIC}(\mathcal{M}) = -2 \log L(\mathcal{M}) + 2 \times p(\mathcal{M})$$

Schwarz's Bayeisan Information Criterion (BIC)

$$\text{BIC}(\mathcal{M}) = -2 \log L(\mathcal{M}) + \log n \times p(\mathcal{M})$$

- $L(\mathcal{M})$ is the likelihood function of the parameters in model $\mathcal{M}$ evaluated at the MLE.

AIC_p and BIC_p Criteria

- In multiple linear regression model with i.i.d. error assumptions

AIC and BIC for linear model

$$AIC_p = n\log(SSE_p) - n\log(n) + 2p$$
$$BIC_p = n\log(SSE_p) - n\log(n) + p\log(n)$$

- Select the model with the lowest AIC_p (or BIC_p) value.
- The terms $2p$ and $p\log(n)$ are penalty that penalizes complex models.

AIC_p and BIC_p Criteria

- AIC is a relative measure. Only AIC values from the same data should be compared.
- A rule of thumb: Good models are those that are within 2 AIC units of the lowest AIC value. Models with more than 10 AIC units above the lowest AIC value are generally not considered.
- AIC can result in overfitting.
- AIC and most other criteria do not have measure of variation and thus the rules are generally approximations.

Outline

Automatic Search Procedures

- Automatic search procedures are useful when the number of possible predictors is large and it becomes difficult to compute criteria for all possible models.
- Two approaches:
  - "Best subsets": search all possible models and take the one with highest R_a^2 or lowest C_p.
  - Stepwise (forward, backward, or both): useful when the number of predictors is large. Choose an initial model and be "greedy".
    - "Greedy" means always take the biggest jump (up or down) in your selected criterion.

Forward Stepwise Selection

1. Fit simple linear regression model for each of the $p-1$ explanatory variables. Compute

$$t_k^* = \frac{\hat{\beta}_k}{\widehat{\text{SD}}(\hat{\beta}_k)}$$

   for each model and keep the model with the largest $|t_k^*|$ provided that it is significant at a pre-specified α level.
2. Add the next “best" explanatory variable using the same criterion.
3. Test whether any current explanatory variables can be deleted.
4. Repeat steps 2 and 3 in the procedure until no further explanatory variables can be deleted or added.

Forward Selection and Backward Elimination

- Two special cases:
  - **Forward selection**: Repeat step 2 but not step 3.
  - **Backward elimination**: Start with the full model and delete explanatory variables. Once deleted, an explanatory variable can not be added back to the model.
- The criterion $|t_k^*|$ can be replaced with R^2, R_a^2, and AIC, BIC, etc.
- Drawbacks of stepwise selection are:
  - Too automatic: No input from the experts.
  - Too random: Small changes in the data can give drastically different models.
  - Different approaches result in different “best" models.

Implementations in R

- "Best subset": use the function `leaps`.
  Works only for multiple linear regression models.
- Stepwise: use the function `step`.
  Works for any model with AIC. In multiple linear regression, AIC is (almost) a linear function of C_p.

Model Validation

- **Model validation** refers to checking a selected model against independent data.
- There are several possible approaches.
- Collect new data as validation data set and check
  - stability of regression coefficient estimation
  - accuracy of prediction
- Compare results with theory or simulations.
- Split data into training and validation set. Check
  - stability of regression coefficient estimation
  - accuracy of prediction

Mean Squared Prediction Error

- Define **mean squared prediction error (MSPE)**:

$$\text{MSPE} = \frac{\sum_{i=1}^{n^*}(Y_i - \hat{Y}_i)^2}{n^*}$$

  where
  - n^* is the sample size of the validation data set.
  - Y_i is the ith observed response in the validation data set.
  - $\hat{Y}_i$ is the ith predicted response in the validation data set.
- If MSPE $\approx$ MSE, then the model is probably adequate.
- If MSPE $>>$ MSE, then the model may not be very useful for general use.

K-Fold Cross Validation

- Split data randomly into K roughly equal parts.
- For $k = 1, \ldots, K$, fit the model using all but the kth part of the data and compute the prediction error sum of squares

$$\mathrm{CV}_k = \sum_{i=1}^{n_k} (Y_{ki} - \hat{Y}_{ki})^2$$

- Compute a K-fold cross-validation estimate

$$\mathrm{CV} = \sum_{k=1}^{K} \mathrm{CV}_k$$