Variable (Model) Selection

Thus far. . .

- predictors identified in advance

Reality. . .

- Many predictors
- Several candidate models
  - all may pass the usual diagnostics and tests
- How do we pick the best model?

What is variable (model) selection?

- the process of choosing a “best” subset of available predictors.
- there might not be a single “best” subset.
- We do want a model we can interpret or justify with respect to the questions of interest.

Questions of Variable Selection

1. Which variables to include? $(X_1, X_2, \ldots)$.
2. What form should these variables take?
   - X_1^2, $\log(X_2)$, $1/X_3$, BMI = Weight/(Height)^2 ....
3. Should we include interaction terms, like `x1*x2`, $X_1/(X_1 + X_2)$?

- Ideally, we answer these questions simultaneously.
- Instead, we will focus on the first question.
- We can then use variable transformations as needed.
- It would be impossible to compare all possible forms of all possible variables.
  - With n observations & p available predictors
    there are p predictors + $\binom{p}{2} = \frac{p(p-1)}{2}$ possible interactions +
    numerous possible transformations
    Impossible to consider all of them

What Happens if We Miss Necessary Predictors or Include Unnecessary Predictors

Mean Squared Error

The **Mean Squared Error (MSE)** of $\hat{\beta}$ is defined to be

$$\text{MSE}(\hat{\beta}) = \text{E}\left[(\hat{\beta} - \beta)^2\right]$$

Warning: This MSE is different from the MSE = SSE/dfE of a MLR model.

One can show that $\boxed{\text{MSE} = \text{Variance} + (\text{Bias})^2}$

$$\begin{aligned}
\text{E}\left[(\hat{\beta} - \beta)^2\right] &= \text{E}\left[\left(\hat{\beta} - \text{E}[\hat{\beta}] + \text{E}[\hat{\beta}] - \beta\right)^2\right] \\
&= \text{E}\left[(\hat{\beta} - \text{E}[\hat{\beta}])^2 + 2(\hat{\beta} - \text{E}[\hat{\beta}])(\text{E}[\hat{\beta}] - \beta) + (\text{E}[\hat{\beta}] - \beta)^2\right] \\
&= \underbrace{\text{E}\left[(\hat{\beta} - \text{E}[\hat{\beta}])^2\right]}_{\text{Variance}} + 2\underbrace{\text{E}\left[\hat{\beta} - \text{E}[\hat{\beta}]\right]}_{=0}\left(\text{E}[\hat{\beta}] - \beta\right) + \underbrace{(\text{E}[\hat{\beta}] - \beta)^2}_{(\text{Bias})^2} \\
&= \left(\text{Variance of } \hat{\beta}\right) + \left(\text{Bias of } \hat{\beta}\right)^2
\end{aligned}$$

where (Bias of $\hat{\beta}$) is defined as $\text{E}[\hat{\beta}] - \beta$, which might not be 0 if $\text{E}[\hat{\beta}] \neq \beta$

Notations

Suppose the "correct" model contains q predictors

$$y_i = \overbrace{\underbrace{\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}}_{\text{retained}} + \underbrace{\beta_{p+1} x_{i,p+1} + \ldots + \beta_q x_{iq}}_{\text{omitted}}}^{\text{correct model}} + \varepsilon_i$$

- Let $\hat{\beta}_j^*$ be the estimated coefficients and $\hat{y}_i^*$ be the predicted values for the correct (big) model
- Let $\hat{\beta}_j$ be the estimated coefficients and $\hat{y}_i$ be the predicted values for the smaller model that retains only the first p predictors ($p < q$)

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i$$

What Happens if We Miss Necessary Predictors?

Gain: Smaller Variance

- $\text{Var}(\hat{\beta}_j^*) \geq \text{Var}(\hat{\beta}_j)$, for $j = 0, 1, \ldots, p$.
- Deleting variables cannot increase the variance.

Loss: Biased Estimates

- Bias $= \text{E}[\hat{\beta}] - \beta$
- Bias $= 0$ if the β_j's of the omitted X_j's are all 0 or if the predictors are uncorrelated
- Bias is small if β_j of the omitted X_j's are small (relative to their SDs)

Variance-Bias Tradeoff: The smaller model might have smaller MSE if the increment in the $(\text{Bias})^2$ is less than the reduction in variance

What Happens if We Miss Necessary Predictors?

Effect of deleting predictors on the prediction of Y is similar

- the smaller model has smaller variance $\text{Var}(\hat{y}_i^*) \geq \text{Var}(\hat{y}_i)$ but greater bias
- MSE will decrease for the smaller model if the increment in the $(\text{Bias})^2$ is less than the reduction in variance
- The best models will keep the important variables —- those with high $\left|\beta_j\right|$.

What Happens if We Include Unnecessary Predictors?

$$y_i = \overbrace{\underbrace{\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}}_{\text{necessary}} + \underbrace{\beta_p x_{i,p+1} + \ldots + \beta_q x_{iq}}_{=0,\ \text{redudant}}}^{\text{bigger model}} + \varepsilon_i$$

If some X_j's have coefficients β_j's equal to 0, but we include them in the model,

- We gain nothing in the precision in estimating β's and predicting y
- The variance in estimation and prediction will increase

Uses of Regression Models

Model Selection Criteria (Model Evaluation)

The way we evaluate a model depends on what we hope to achieved with our model:

- Description
- Prediction

In many cases, these uses overlap.

There might not be a single best model

Goal 1: Description

- Goal: To describe a given process or understand and model the variation in a complex interacting system.
- Interpretability: Lots of thinking required about which variables are important
- Two conflicting requirements:
  1. Account for as much of the variation as possible;
  2. The principle of parsimony;
     Understanding and interpretation are easier with fewer variables.
- Strategy: Choose the smallest set of variables that accounts for the largest percentage of variation in the response.

Goal 2: Prediction

- Goal: To use the patterns in our current data set to estimate the mean of future response or to predict a future value.
- This is also called **forecasting**. The focus is on applying knowledge to observations not in the current data.
- Strategy: Minimize the MSE of the estimation or prediction

$$\text{MSE}(\hat{y}) = \text{E}\left[(\hat{y} - \text{E}(y|\mathbf{x}))^2\right] \quad \text{or} \quad \text{MSE}(\hat{y}) = \text{E}\left[(\hat{y} - y)^2\right]$$

Criteria for Evaluating Models

Summary of Model Comparison Methods

Nested models

- F-test (can provide P-values)

Any two models with the same response (no P-values)

- MSE (Mean Squared Error = SSE$/(n - p - 1)$)
- AIC (Akaike Information Criterion)
- BIC (Bayesian Information Criterion)
- Mallow's C_p (has fallen out of favor)

Any two models with the same response (but possibly differently transformed)

- adjusted R^2

Information Criteria

Both Akaike and Bayesian Information Criteria
reward small variance (SSE_p/n small) and penalize larger models (p large).

$$\text{AIC} = n\log_e(\text{SSE}_p/n) + 2p$$
$$\text{BIC} = n\log_e(\text{SSE}_p/n) + p\log_e(n)$$

- **CAUTION:** For AIC, BIC, and C_p,
  p = number of parameters (including the intercept).
  different from our usual meaning of the letter p (= # of predictors).
- Smaller AIC/BIC is better
- Models with AIC differ < 2 should be considered equally adequate.
- Similarly, models with BIC differ ≤ 2 are considered equally good
- BIC penalty for larger models is more severe
  - $p\log_e(n) > 2p$ (whenever $n > 8$)

Variable Selection Procedures

Searching Over All Possible Subsets (1)

The most direct and ideal approach is to examine **all possible subsets** of potential predictors.

What are "all possible subsets"?

- e.g., if there are 3 potential predictors: `X1`, `X2`, and `X3`, the models we could consider include

0 predictor	1 predictor	2 predictors	3 predictors
• $Y \sim 1$	• $Y \sim X1$ • $Y \sim X2$ • $Y \sim X3$	• $Y \sim X1 + X2$ • $Y \sim X1 + X3$ • $Y \sim X2 + X3$	• $Y \sim X1 + X2 + X3$

Searching Over All Possible Subsets (1)

The most direct and ideal approach is to examine **all possible subsets** of potential predictors.

What are "all possible subsets"?

- e.g., if there are 3 potential predictors: `X1`, `X2`, and `X3`, the models we could consider include

0 predictor	1 predictor	2 predictors	3 predictors
• $Y \sim 1$	• $Y \sim X1$ • $Y \sim X2$ • $Y \sim X3$	• $Y \sim X1 + X2$ • $Y \sim X1 + X3$ • $Y \sim X2 + X3$	• $Y \sim X1 + X2 + X3$

With q possible predictors $X_1, X_2, \ldots, X_q$, there are

$$\underbrace{2 \times 2 \times \ldots \times 2}_{q \text{ times}} = 2^q$$

subsets of $\{X_1, X_2, \ldots, X_q\}$ as each X_i can be included or not included in the model. There would be 2^q possible models

Searching Over All Possible Subsets (2)

- If $q = 20$, there are $2^{20} > 1$ million candidate models
- If $q = 30$, there are $2^{30} > 1$ billion candidate models
  not practical to examine all of them unless q is quite small.

If it's possible to search over all possible subsets, then a good strategy is

1. Choose the best models in each class of p-term models, for $p = 0, 1, \ldots, q$
2. Analyze these best models more closely, including diagnostic plots, possible transformations, etc.
3. Select the best model(s) (including transformations) by comparing both diagnostics and scores.

Stepwise Precedures

As it's not practical to search all possible subsets when q is large, here are a few commonly used algorithms aimed to find the "best" model without look at all possible subsets

- Forward selection (FS)
- Backward elimination (BE)
- Stepwise selection (SW)

Forward Selection (FS)

The **Forward Selection** algorithm consider all candidate subsets consisting of one additional term beyond the current subset

1. It begins with an intercept-only model.
2. At each iteration, add the predictor with the smallest P-value and then fit the new model
3. Stop if:
    - The new selected variable is not significant, or
    - All variables have been selected (all variables added). 4 Otherwise, fit the new model with the new added variable, and go to Step 2

The FS algorithm considers at most

$$q + (q - 1) + \cdots + 2 + 1 = q(q + 1)/2$$

subsets, not all 2^q possible subsets.

Backward Elimination (BE)

1. The **Backward Elimination (BE)** algorithm begins with the full set of variables.
2. Eliminate the least significant variable (the one with the largest P-value)
3. Stop if:
   - All variables are significant, or
   - All variables have been eliminated (intercept-only model).
4. Otherwise, eliminate the least significant variable and go to Step 2.

The BE algorithm considers at most

$$q + (q - 1) + \cdots + 2 + 1 = q(q + 1)/2$$

subsets, not all 2^q possible subsets.

Stepwise Selection (SW)

1. At each iteration, the **Stepwise Selection (SW)** algorithm consider all models obtained by either adding or deleting one term to or from the current model.
2. At each iteration, choose the model with the lowest AIC or BIC
3. Stop if the model with the lowest AIC/BIC is the current model
4. Otherwise, let the model with the lowest AIC/BIC be the new current model and then go back to Step 1

Using the SW algorithm, a term added to a model might be removed at a later step

FS, BE, SW Algorithms with AIC and BIC

- Instead of using P-values, we can use scoring methods like AIC and BIC.
- At any iteration, compare models based on the chosen scoring method.
- Among the models we are considering, we choose the model with the lowest AIC (or BIC).
- We stop the procedure when no candidates reduce the score.
- The major difference is that we are not judging variables based on significance levels, but only on the basis of how they affect the score.

Cautions About the FS, BE, SW Algorithms

- The indicator variables of a **categorical predictor** should be included or removed altogether
  - better using AIC/BIC rather than P-values since adding/removing a categorical predictor may involve more than 1 parameter but a numerical predictor involves just 1 parameter
  - May simplify a categorical predictor by merging some categories
- If the possible pool of model terms include **interactions**, note
  - an interaction is never added unless all the lower order effects in the interaction are already included.
  - if an interaction is in the current model, none of its component variables or lower order interaction should be removed

Final Remarks about FS, BE, Stepwise Methods

- These methods should **NOT be used mechanically**:
  - Do chosen variables make sense according to domain-specific knowledge?
  - Do the diagnostic plots indicate that model assumptions are valid?
  - Be open to other models that may be approximately as adequate.
- The order in which we add/remove variables do not indicate relative importance.
- These methods may not give the “best” model
- All three methods usually give similar results for non-collinear data.

Problem & Goals

- When we have many predictors (with many possible interactions), it can be difficult to find a good model.
- Which main effects do we include?
- Which interactions do we include?
- Model selection procedures try to simplify / automate this task.
- Election data has $2^6 = 64$ different models with just main effects!

General comments

- This is generally an “unsolved” problem in statistics: there are no magic procedures to get you the “best model.”
- Many machine learning methods look for good “sparse” models: selecting a “sparse” model.
- “Machine learning” often work with very many predictors.
- Our model selection problem is generally at a much smaller scale than “data mining” problems.
- Still, it is a hard problem.