MAT 3375
Regression Analysis

Chapter 5
Model Selection

P. Boily (uOttawa)

Summer – 2023

Outline

5 – Model Selection

With reasonable real-world datasets and situations, we can often build tens (if not hundreds) of models related to a specific scenario.

When most of these models are "aligned" with one another (i.e., they give similar results), picking the simplest model is usually the best approach.

In practice, we can also reach a point of **diminishing returns** – including more variables in the model might not yield better predictive power (thanks to the curse of dimensionality).

How do we pick "the" model to work with?

5.1 – Preliminaries

A linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_n)$ should be seen as an attempt to approximate the (**not necessarily linear**) regression function

$$y = f(\mathbf{x}) = \mathrm{E}\left\{Y \mid (X_1, \ldots, X_p) = \mathbf{x}\right\}.$$

In this framework, we assume a **linear relationship** between the response Y and the predictors $X_1, \ldots, X_p$, which we fit using the **OLS** framework:

$$\mathbf{b} = \arg\min_{\boldsymbol{\beta}} \{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2\}.$$

But prediction accuracy suffers when $p > n$; model interpretability can be improved by removing **irrelevant features** (i.e., by **reducing** p).

There are 3 classes of methods to do so:

- **shrinkage/regularization methods** (out-of-scope for the course, see section 4.6);
- **dimension reduction**, in which we project the p predictors onto an $M-$dimensional manifold $\mathcal{H}$, with $M \ll p$, and
- **subset selection**, we identify a subset of the p predictors for which there is evidence of a (strong) association with the response, and we fit a model to this reduced set using the OLS framework – given p predictors (some of which may be interaction terms, binary variables, polynomial powers, etc.), there are 2^p OLS models that can be fit on the data.

  But which of those models should be selected as the **best model**?

5.2 – Best Subset Selection

In the **best subset selection** (BSS) approach, the search for the best model is usually broken down into 3 stages:

1. let $\mathcal{M}_0$ denote the **null model** (no predictor) which simply predicts the sample mean for all observations;

2. for $k = 1, \ldots, p$ (and as long as the model can be fit):

   (a) fit **every** model that contains k predictors (there are $\binom{p}{k}$ of them);
   (b) pick the model with **smallest** SSE (**largest** R^2) and denote it by $\mathcal{M}_k$;

3. select a **unique** model from $\{\mathcal{M}_0, \ldots, \mathcal{M}_p\}$ using C_p (AIC), BIC, R_a^2, or any other appropriate metric.

We cannot use SSE or R^2 as metrics in the last step, as we would always select $\mathcal{M}_p$ since SSE **decreases monotonically** with k and R^2 **increases monotonically** with k.

BSS is conceptually simple, but with 2^p models to try out, it quickly becomes **computationally infeasible** for large p ($p > 40$, say).

When p is large, the chances of finding a model that performs **well** according to step 3 but **poorly** for new data **increase**, which can lead to **overfitting** and **high-variance** estimates.

We are assuming that all models are **OLS** models, but subset selection algorithms can be used for other families of methods; all that is required are appropriate **training** error estimates for step 2b and **test** error estimates for step 3.

5.3 – Stepwise Selection

Stepwise selection (SS) methods attempt to overcome this challenge by only looking at a **restricted** set of models. **Forward stepwise selection** (FSS) starts with the **null model** $\mathcal{M}_0$ and adding predictors one-by-one until it reaches the **full model** $\mathcal{M}_p$:

1. Let $\mathcal{M}_0$ denote the **null model**;

2. For $k = 0, \ldots, p-1$ (and as long as the model can be fit):
   (a) consider the $p-k$ models that add a **single predictor** to $\mathcal{M}_k$;
   (b) pick the model with **smallest** SSE (**largest** R^2), denote it by $\mathcal{M}_{k+1}$;

3. select a **unique** model from $\{\mathcal{M}_0, \ldots, \mathcal{M}_p\}$ using C_p (AIC), BIC, R_a^2, or any other appropriate metric.

Backward stepwise selection (also BSS, unfortunately) works the other way, starting with the **full model** $\mathcal{M}_p$ and removing predictors one-by-one until it reaches the **null model** $\mathcal{M}_0$:

1. Let $\mathcal{M}_p$ denote the **full model**;

2. For $k = p, \ldots, 1$ (and as long as the model can be fit):

   (a) consider the k models that remove a **single predictor** from $\mathcal{M}_k$;
   (b) pick the model with **smallest** SSE (**largest** R^2), denote it by $\mathcal{M}_{k-1}$;

3. select a **unique** model from $\{\mathcal{M}_0, \ldots, \mathcal{M}_p\}$ using C_p (AIC), BIC, R_a^2, or any other appropriate metric.

The computational advantage of SS over B(est)SS is evident: instead of having to fit 2^p models, SS only requires to fit

$$1 + p + (p-1) + \cdots + 2 + 1 = \frac{p^2 + p + 2}{2}.$$

While there is no guarantee that the "**best**" model (among the 2^p B(est)SS models) is found in the SS models, SS can be used in settings where p is **too large** for BSS to be computationally feasible.

For OLS models, **backward SS** only works if $p \leq n$ (otherwise OLS might not have a unique parameter solution); if $p > n$, only **FSS** is viable.

Hybrid selection (HS) methods attempt to mimic BSS while keeping model computation in a manageable range, not unlike in SS.

5.4 – Adjustment Statistics

Commonly, we use one of the following **adjustment statistics**:

- **Mallow's C_p**
- the **Akaike information criterion** (AIC)
- the **Bayesian information criteria** (BIC), or
- the **adjusted coefficience of determination R_a^2**.

The first three of these must be **minimized**, while the last must be **maximized**.

The **adjustment statistics** require the following quantities:

- n, p, and $d = p + 2$
- $\hat{\sigma}^2$, the estimate of $\sigma^2\{\varepsilon\}$;
- SSE and SST.

Mallow's C_p statistic is

$$C_p = \frac{1}{n}(\mathrm{SSE} + 2d\hat{\sigma}^2) = \frac{1}{n}\mathrm{SSE} + \underbrace{\frac{2d\hat{\sigma}^2}{n}}_{\text{adjustment}}.$$

As d increases, so does the adjustment term. Note that if $\hat{\sigma}^2$ is an unbiased estimate of $\sigma^2\{\varepsilon\}$, C_p is an unbiased estimate of MSE.

The **Akaike information criterion** (AIC) is

$$\text{AIC} = -2\ln L + \underbrace{2d}_{\text{adjustment}},$$

where L is the maximized value of the likelihood function for the estimated model. If the errors are **normally distributed**, this requires maximizing

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\hat{\sigma}} \exp\left(-\frac{(Y_i - \mathbf{X}_i\boldsymbol{\beta})^2}{2\hat{\sigma}^2}\right) = \frac{1}{(2\pi)^{n/2}\hat{\sigma}^n} \exp\left(-\frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n}(Y_i - \mathbf{X}_i\boldsymbol{\beta})^2\right),$$

or, upon taking the logarithm,

$$\ln L = \text{constant} - \frac{1}{2\hat{\sigma}^2}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2,$$

and so

$$\arg\max_{\boldsymbol{\beta}}\{\ln L(\boldsymbol{\beta})\} = \arg\min_{\boldsymbol{\beta}}\{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2\}.$$

However,

$$\begin{aligned}
\text{AIC} &= -2\ln L + 2d = \text{constant} + \frac{1}{\hat{\sigma}^2}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + 2d \\
&= \text{constant} + \frac{\text{SSE}}{\hat{\sigma}^2} + 2d \\
&= \text{constant} + \frac{n}{\hat{\sigma}^2}\cdot\frac{1}{n}\left(\text{SSE} + 2d\hat{\sigma}^2\right) = \text{constant} + \frac{n}{\hat{\sigma}^2}C_p.
\end{aligned}$$

Evidently, when the error structure is normal, **minimizing** AIC is equivalent to **minimizing** C_p.

The **Bayesian information criterion** uses a different adjustment term:

$$\text{BIC} = \frac{1}{n}(\text{SSE} + d\hat{\sigma}^2 \ln n) = \frac{1}{n}\text{SSE} + \underbrace{d\hat{\sigma}^2 \frac{\ln n}{n}}_{\text{adjustment}}.$$

This adjustment penalizes models with a **large** number of predictors; **minimizing** BIC results in selecting models with fewer variables than those obtained by **minimizing** C_p, in general.

The **adjusted coefficient of determination** R^2_a of a $k-$parameter model is

$$R^2_{a,k} = 1 - \frac{\text{SSE}/(n-k-1)}{\text{SST}/(n-1)} = 1 - (1-R^2)\frac{n-1}{n-k-1}.$$

Maximizing $R^2_{a,k}$ **minimizes** $\frac{\text{SSE}}{n-k-1}$, penalizing **unnecessary** variables.

TL;DR: if p is the number of parameters in the **full model** (F), we want to find a **reduced model** (R) with k parameters that also fits the data well.

1. R_p^2-**criterion:** for each $k-$subset of parameters, compute the coefficient of determination $R_k^2 = 1 - \frac{\text{SSE}_k}{\text{SST}}$; we find a $k-$subset such that if we increase k, the highest R_k^2 does not change significantly (to 2 decimal places, say).

2. $R_{a,p}^2-$**criterion:** for each $k-$subset of parameters, compute the adjusted coefficient of determination $R_{a,k}^2 = 1 - \frac{n-1}{n-k}\frac{\text{SSE}_k}{\text{SST}}$; we find a $k-$subset that maximizes $\{R_{a,k}^2\}$.

3. **Mallow's** C_p-**criterion:** $C_p = \frac{\text{SSE}_k}{\text{MSE}(F)} - (n-2k)$; we find a $k-$subset such that C_p is small and close to k. This criterion might produce numerous appropriate reduced models.

Example: for a certain data set with three predictors, we obtain the corresponding Mallow's C_p and R_p^2 for all subsets of the predictors.

p	C_p	R_p^2	Variables in model
4	4.0000	0.8548	X_1, X_2, X_3
3	22.4041	0.7527	X_1, X_2
3	29.1518	0.7189	X_1, X_3
2	42.3306	0.6429	X_1
3	52.8666	0.6002	X_2, X_3
2	81.6508	0.4461	X_2
2	146.8485	0.1197	X_3

Using either Mallow's C_p criterion or the R_p^2 criterion, can we find any "good" reduced models?

Solution: except for the first selection (which turns out to be the full model, not a reduced one), none of the C_p are really small and near p, so Mallow's C_p-criterion is unlikely to be useful.

As for the other criterion, we have

p	Highest R_p^2
2	0.6429
3	0.7527
4	0.8548

Going from $p = 2$ to $p = 3$, the difference is $0.7527 - 0.6429 = 0.1098$; going from $p = 3$ to $p = 4$, the difference is $0.8548 - 0.7527 = 0.1021$. The second difference is slightly smaller than the first; given this, if we absolutely had to chose a reduced model, we ought to opt for the model for which $R_3^2 = 0.7527$ (the one with variables X_1 and X_2).