# Lecture 15:
# Linear Model Selection

## Big Data and Machine Learning for Applied Economics
## Econ 4676

Ignacio Sarmiento-Barbieri

Universidad de los Andes

September 29, 2020

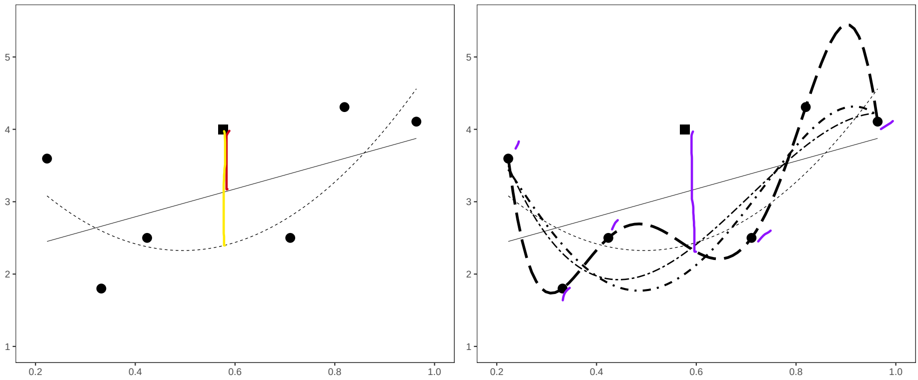
# Agenda

1. Motivation
   - Recap: Overfit
2. Classical Framework for Model Selection
   - AIC: Akaike Information Criterion
   - SIC/BIC: Schwarz/Bayesian Information Criterion
   - Connection to Classical Hypothesis Testing: General
   - AIC/SIC in the linear regression model
3. Model Selection in Practice
   - Best Subset Selection
   - Stepwise Selection
4. Review & Next Steps
5. Further Readings

# Overfit and out of Sample Prediction

- ML we care about prediction out of sample
- Overfit: complex models predict very well inside a sample but ”bad” outside
- Choose the right complexity level
- How do we measure the out of sample error?
- $R^2$ doesn’t work: measures prediction in sample, it’s non decreasing in complexity (PS1)

# Overfit and out of Sample Prediction



# Motivation

- Estimating test error: two approaches
  1. We can directly estimate the test error, using either a validation set approach or a cross-validation approach
  2. We can indirectly estimate test error by making an adjustment to the training error to account for overfitting.
     - SIC Mallows
       AIC, BIC, $C_p$ and Adjusted $R^2$ ad hoc
     - These techniques adjust the training error for the model size, and can be used to select among a set of models with different numbers of variables.
     - I'll focus on AIC and BIC. They are intimately related to more classical notions of hypothesis testing.

# Classical Framework for Model Selection

- The framework for model selection can be described as follows.
- We have a collection of parametric models

$$f_i(x_i,\theta) \tag{1}$$

- where $\theta \in \Theta_j$ for $j = 1, \ldots, J$.
- Some linear structure is usually imposed on the parameter space, so typically $\Theta_j = m_j \cap \theta_J$ , where $m_j$ is a linear subspace of $\mathcal{R}^{p_J}$ of dimension $p_j$ and $p_1 < p_2 < \cdots < p_J$.
- e.g.

$$y = X_{n\times p_j}\beta + u \tag{2}$$

$$y = x_1\beta_1 + x_2\beta_2 + \quad + x_p\beta_p + u$$

# AIC

- Akaike (1969) was the first to offer a unified approach to the problem of model selection.
- His point of view was to choose a model from the set $f_i$ which performed well when evaluated on the basis of forecasting performance.
- His criterion, which has come to be called the Akaike information criterion is

AIC = likelihood(θ) + penalidad

parametros

$$AIC(j) = l_j(\hat{\theta}) - p_j \tag{3}$$

- where $l_j(\theta)$ the log likelihood corresponding to the $j$ model maximized over $\theta \in \Theta_j$.

# AIC

$$AIC(j) = l_j(\hat{\theta}) - p_j \tag{4}$$

- Akaike's model selection rule was simply to maximize AIC over the $j$ models, that is to choose the model $j^*$ which maximizes $AIC(j)$.
- This approach seeks to balance improvement in the fit of the model, as measured by the value of the likelihood, with a penalty term, $p_j$.
- Thus one often sees this and related procedures referred to as penalized likelihood methods.
- The trade-off is simply: does the improvement which comes inevitably from expanding the dimensionality of the model compensate for the increased penalty?

# BIC

- Schwarz (1978) showed that while the *AIC* approach may be quite satisfactory for selecting a forecasting model
- However had the unfortunate property that it was inconsistent, in particular, as $n \to \infty$, it tended to choose too large a model with positive probability.
- Schwarz (1978) formalized the model selection problem from a Bayesian standpoint:

AIC Pj

$$SIC(j) = l_j(\hat{\theta}) - \overbrace{\frac{1}{2} p_j \underline{log(n)}} \tag{5}$$

- It has the property that as $n \to \infty$, presuming that there was a true model, $j^*$, then $\hat{j} = argmax\ SIC(j)$, satisfied

$$p(\hat{j} = j^*) \to 1 \tag{6}$$

# AIC vs BIC

$$AIC(j) = l_j(\hat{\theta}) - p_j \tag{7}$$

$$SIC(j) = l_j(\hat{\theta}) - p_j \frac{1}{2} log(n) \tag{8}$$

$n > 8$

- Note that

$$\frac{1}{2} log(n) > 1 \;\; for \;\; n > 8 \tag{9}$$

- The SIC penalty is larger than the AIC penalty,
- SIC tends to pick a smaller model.
- In effect, by letting the penalty tend to infinity slowly with n, we eliminate the tendency of AIC to choose too large a model.

# Connection to Classical Hypothesis Testing: General

- Recall the likelihood ratio tests, that we classically use to assess goodness of fit/ compare models.
- Suppose that we are comparing a larger model $j$ to a smaller model $i$

$$\underbrace{T_n = 2(l_j(\hat{\theta}_j) - l_i(\hat{\theta}_i))} \quad 2 \log \left( \frac{\mathcal{L}_i}{\mathcal{L}_j} \right) \tag{10}$$

- It can be shown that $T_n \to \chi^2_{\underbrace{p_j - p_i}}$ for $p_j > p_i = p^*$.
- So classical hypothesis testing would suggest that we should reject an hypothesized smaller model $i$, in favor of a larger model $j$ iff $T_n$ exceeds an appropriately chosen critical value from the $\chi^2_{p_j - p_i}$ table

$H_0$ $i$ true $\qquad$ $H_1$ $j$ true

$T_n \gtrless CV_{\chi^2_{p_j - p_i}}$ } test

# Connection to AIC

AIC chooses $j$ over $i$, iff

model j — model i

$$\overbrace{l_j(\hat{\theta}) - p_j}^{AIC(j)} > \overbrace{l_i(\hat{\theta}) - p_i}^{AIC(i)} \tag{11}$$

some algebra

$$l_j(\hat{\theta}) - l_i(\hat{\theta}) > \underbrace{p_j - p_i}_{+} \tag{12}$$

$P_j > P_i$

$$2\frac{l_j(\hat{\theta}) - l_i(\hat{\theta})}{p_j - p_i} > 2 \tag{13}$$

valores criticos

$\chi^2$

AIC → valor critico es 2

$$\frac{Tn}{(P_j - P_i)} > 2$$

# Connection to SIC

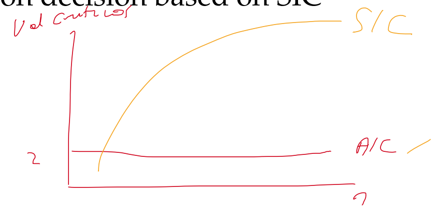
$$l_j - \frac{1}{2} p_j \log(n) > l_i - \frac{1}{2} p_i \log(n)$$

Same algebra (HW)

In contrast Schwarz would choose *j* over *i*, iff

$$\frac{\overbrace{2(l_j(\hat{\theta}_j) - l_i(\hat{\theta}_i))}^{T_n}}{p_j - p_i} > \boxed{\underline{log(n)}} \tag{14}$$

Then $log(n)$ can be interpreted as an implicit critical value for the model selection decision based on SIC



Error tipo I
Reject H0 | H0 true
α / size

# AIC/SIC in the linear regression model

$y = X\beta + u \quad u \sim N(0, \sigma^2 I)$

Recall that for the for the Normal/Gaussian linear regression model the log likelihood function is

$$l(\beta, \sigma^2) = \overbrace{-\frac{n}{2} log(2\pi)}^{u} - \underbrace{\frac{n}{2} log(\sigma^2)}_{} - \frac{1}{2\sigma^2}\underbrace{(y - X\beta)'(y - X\beta)}_{u' u} \tag{15}$$

evaluating at $\hat{\beta}$, and $\hat{\sigma}^2 = (y - X\hat{\beta})'(y - X\hat{\beta})$ we get the concentrated/profile log-likelihood

$$l(\cancel{\hat{\beta}}, \underset{(\hat\beta)}{\hat{\sigma}^2}) = -\frac{n}{2} log(2\pi) - \frac{n}{2} log(\hat{\sigma}^2) - \frac{n}{2} \tag{16}$$

$l(\hat\sigma^2) = cons - \frac{n}{2} \log(\hat\sigma^2)$

# AIC/SIC in the linear regression model

Thus maximizing SIC / BIC

$$SIC = l_i - \frac{1}{2}p_i log(n) \tag{17}$$

is equivalent to minimize

$$SIC_i = \frac{n}{2}log(\hat{\sigma}_i^2) + \frac{1}{2}p_i log(n) \tag{18}$$

or minimizing

$$\boxed{log(\hat{\sigma}_i^2) + \frac{p_i}{n}log(n)} \tag{19}$$

$y = x_1\beta_1 + \beta_2 x_2 + u$

$p = 2$

$X(1)$ → cont de ceros

→ ISLR

otra forma q' es equiv

HW $(d = p)$

- Similarity for AIC
- When using software is important to check what is being computed. In R, the function `AIC` minimizes and not maximizes, and defines AIC as $-2l_i + kp_i$ with $k = 2$ as default that can be changed,e.g. $k = log(n)$ gives SIC

# Comparison LR, t, AIC, BIC in the linear regression model

Example of adding one more covariate $p_j - p_i = 1$

$$T_n = \underbrace{2(l_j(\hat{\theta}_j) - l_i(\hat{\theta}_i))} \to \chi^2_{\underbrace{p_j-p_i}_{=1}} \tag{20}$$

$$\frac{T_n}{p_j - p_i} \to \boxed{\frac{\chi^2_{p_j-p_i}}{p_j - p_i}} \approx F_{p_j-p_i,\infty} \tag{21}$$

$Q = \sum^{k} z_i^2$

$z_i \sim N(0,1)$

$Q \sim \chi^2_{(k)}$

$F_{d_1,d_2} = \frac{\chi^2_{d_1}/d_1}{\chi^2_{d_2}/d_2}$

$\frac{\chi^2_\infty}{\infty} \to 1$

$\to E(z^2) = 1$

$\frac{1}{n}\chi^2_\infty = \lim_{n\to\infty} \frac{1}{n}\sum^{n} z_i^2$ LLN

# Comparison LR, t, AIC, BIC in the linear regression model

$p_j - p_i = 1$

$$\sqrt{2(l_j(\hat{\theta}_j) - l_i(\hat{\theta}_i))} \to \sqrt{F_{\boxed{1},\infty}} = t_\infty$$

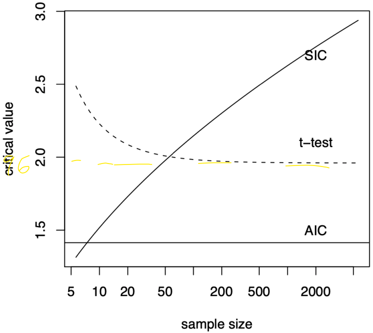
AIC

$$\sqrt{\boxed{2l_j(\hat{\theta}) - l_i(\hat{\theta})}} > \sqrt{2}$$

SIC

BIC

$$\sqrt{\boxed{2(l_j(\hat{\theta}_j) - l_i(\hat{\theta}_i))}} > \sqrt{log(n)}$$



1,96

# Model Selection in Practice

- We have $M_k$ models
- We want to find the model that best predicts out of sample
- We have a number of ways to go about it
    - Best Subset Selection
    - Stepwise Selection
        - Forward selection
        - Backward selection

# Best Subset Selection

1. Let $M_0$ denote the null model, which contains no predictors. This model simply predicts the sample mean for each observation.

2. For $k = 1, 2, \ldots, p$:

   1. Fit all $\binom{p}{k}$ models that contain exactly k predictors

   2. Pick the best among these $\binom{p}{k}$ models, and call it $M_k$. Where *best* is the one with the smallest $SSR$ → $\underbrace{R^2 \text{ max}}$ Training data

3. Select a single best model from among $M_0, \ldots, M_p$ using cross-validated prediction error, AIC ($C_p$), BIC, or adjusted $R^2$.

$$\underset{1+}{M_0}, \underbrace{\hat{M}_1, \hat{M}_2, \quad \hat{M}_p}_{p} \Big| \; ?$$

# Stepwise Selection

$p \leq 20$ $2^p - 1,048,576$

- For computational reasons, best subset selection cannot be applied with very large p.
- Best subset selection may also suffer from statistical problems when p is large: larger the search space, the higher the chance of finding models that look good on the training data, even though they might not have any predictive power on future data.
- Thus an enormous search space can lead to overfitting and high variance of the coefficient estimates.
- For both of these reasons, stepwise methods, which explore a far more restricted set of models, are attractive alternatives to best subset selection.

# Forward Stepwise Selection

- Forward stepwise selection begins with a model containing no predictors, and then adds predictors to the model, one-at-a-time, until all of the predictors are in the model.

- In particular, at each step the variable that gives the greatest additional improvement to the fit is added to the model.

# Forward Stepwise Selection

1. Let $M_0$ denote the null model, which contains no predictors. This model simply predicts the sample mean for each observation.

2. For $k = 0, 1, \ldots, p-1$:

   1. Consider all $p-k$ models that augment the predictors in $M_k$ with one additional predictor.

   2. Choose the best among these $p-k$ models, and call it $M_{k+1}$. Where *best* is the one with the smallest *SSR* → $R^2$

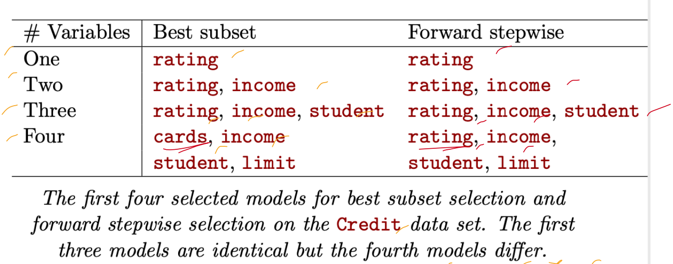
3. Select a single best model from among $M_0, \ldots, M_p$ using cross-validated prediction error, AIC ($C_p$), BIC, or adjusted $R^2$.

Best  $p=20$  $2^p = 1,048,576$

Forw & Back  $p = 20$  $1 + \frac{(p+1)p}{2} = 211$

# Forward Stepwise Selection

- Computational advantage over best subset selection is clear.
- It is not guaranteed to find the best possible model out of all $2^p$ models containing subsets of the p predictors.
- ISLR Example



| # Variables | Best subset | Forward stepwise |
|---|---|---|
| One | rating | rating |
| Two | rating, income | rating, income |
| Three | rating, income, student | rating, income, student |
| Four | cards, income<br>student, limit | rating, income,<br>student, limit |

*The first four selected models for best subset selection and forward stepwise selection on the* Credit *data set. The first three models are identical but the fourth models differ.*

ISLR → R tiene todos los datos.

# Backward Stepwise Selection

- Like forward stepwise selection, backward stepwise selection provides an efficient alternative to best subset selection.

- However, unlike forward stepwise selection, it begins with the full least squares model containing all p predictors, and then iteratively removes the least useful predictor, one-at-a-time.

# Backward Stepwise Selection

1. Let $M_0$ denote the null model, which contains no predictors. This model simply predicts the sample mean for each observation.

2. For $k = p, p-1, \ldots, p$:

   1. Consider all $k$ models that contains all but one of the predictors in $M_k$, for a total of $k-1$ predictors
   2. Choose the best among these $k$ models, and call it $M_{k-1}$. Where *best* is the one with the smallest $\underline{SSR}$ → $R^2$ not

- Fernon

3. Select a single best model from among $M_0, \ldots, M_p$ using cross-validated prediction error, AIC ($C_p$), BIC, or adjusted $R^2$.

# Backward Stepwise Selection

- Like forward stepwise selection, the backward selection approach searches through only $1 + p(p+1)/2$ models, and so can be applied in settings where p is too large to apply best subset selection
- Like forward stepwise selection, backward stepwise selection is not guaranteed to yield the best model containing a subset of the p predictors.
- Backward selection requires that the number of samples n is larger than the number of variables p (so that the full model can be fit). In contrast, forward stepwise can be used even when $n < p$, and so is the only viable subset method when p is very large.

# Validation and Cross-Validation

- Each of the procedures returns a sequence of models $M_k$ indexed by model size $k = 0, 1, 2, \ldots$ .
- Our job here is to select $\hat{k}$. Once selected, we will return model $M_{\hat{k}}$
- We compute the validation set error or the cross-validation error for each model $M_k$ under consideration, and then select the $k$ for which the resulting estimated test error is smallest.
- This procedure has an advantage relative to AIC ($C_p$), BIC, and adjusted $R^2$, in that it provides a direct estimate of the test error, and doesn't require an estimate of the error variance $\sigma^2$
- It can also be used in a wider range of model selection tasks, even in cases where it is hard to pinpoint the model degrees of freedom (e.g. the number of predictors in the model) or hard to estimate the error variance $\sigma^2$

# Review & Next Steps

- Today:
  - Basic Classical Framework for Model Selection AIC, SIC/BIC
  - Model Selection in Practice
    - Best Subset Selection regsubset package Leaps
    - Stepwise Selection
- Next class: Regularization/Shrinkage Methods
- Questions? Questions about software?

# Further Readings

- Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning (Vol. 1, No. 10). New York: Springer series in statistics.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.
- Koenker, R. (2013) Economics 508: Lecture 4. Model Selection and Fishing for Significance. Mimeo