Model selection, fit, validation

Emilija Perković

Dept. of Statistics
University of Washington

MSE

So far we have discussed

- how to estimate the OLS model,
- how to do inference on the parameters of our model,
- how to compare two nested models,
- and how to check our model assumptions.

How do we choose a model that will predict well, and not just a model that seems to satisfy our assumptions?

We might consider selecting among models by minimizing the estimated (within-sample) mean squared error (MSE):

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

In-sampe MSE

One issue with this system is that the in-sample MSE is that is **decreases as you add more predictors to the model.**

Remember that we estimate $\underline{\beta}$ by minimizing the residual sum of squares (RSS_p):

$$min_{\beta_0,\beta_1,\ldots,\beta_p} \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 \tag{1}$$

And $MSE_p = RSS_p/n$.

Alternatively, a model that only contains q out of p of the above predictors, $q < p$, minimizes the residual sum of squares RSS_q:

$$\begin{aligned} &min_{\beta_0,\beta_1,\ldots,\beta_q} \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{q}\beta_j x_{ij})^2 \\ &= min_{\beta_0,\beta_1,\ldots,\beta_q,\beta_{q+1}=0,\ldots\beta_p=0} \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 \end{aligned} \tag{2}$$

Notice, that (2) is a special case of (1), when $\beta_{q+1} = 0, \ldots \beta_p = 0$.

Hence, $RSS_p \leq RSS_q$. Since $MSE_q = RSS_q/n$, $MSE_p \leq MSE_q$.

Generalization Error

Ideally, we would want to select a model that has the best predictive properties.

Suppose that our model:

$$\underline{y} = X\underline{\beta} + \underline{\epsilon},$$

was fit using OLS, so that we obtain $\underline{\hat{\beta}} = (X'X)^{-1}X'\underline{y}$.

Now, suppose we have a new *random* point $(x_1^*, \ldots, x_p^*, y^*)$ and we want to evaluate how well our model performs on this new data point.

The **generalization error** (prediction error, $MSE(\hat{y})$) is:

$$\mathrm{E}[(y^* - \hat{y}^*)^2 | X = x] = \mathrm{E}[(y^* - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j x_j^*)^2].$$

The in-sample MSE is an overoptimistic measure for the generalization error.

Generalization Error

The in-sample MSE is an overoptimistic measure for the generalization error. To see this, consider that we observe a n new samples with the same predictor values, that is the same design matrix X, but a different vector of responses $\underline{y}'$.
So now you have observed both

$$\underline{y} = X\underline{\beta} + \underline{\epsilon},$$

and

$$\underline{y}' = X\underline{\beta} + \underline{\epsilon}'.$$

Can the OLS model that we estimated using the first n samples, predict $\underline{y}'$? What is the generalization error of our model? And how does it compare to the in-sample MSE?

Generalization error

We define the estimated out-of-sample prediction error as (out-of-sample MSE):

$$\frac{1}{n}\sum_{i=1}^{n}(y_i' - \hat{y}_i)^2.$$

And we will show that:

$$\mathrm{E}[\frac{1}{n}(\sum_{i=1}^{n} y_i - \hat{y}_i)^2] < \mathrm{E}[\frac{1}{n}(\sum_{i=1}^{n} y_i' - \hat{y}_i)^2]$$

Proof.
For the i^{th} term:

$$\begin{aligned} \mathrm{E}[(y_i - \hat{y}_i)^2] &= \mathrm{Var}[y_i - \hat{y}_i] + (\mathrm{E}[y_i - \hat{y}_i])^2 \\ &= \mathrm{Var}[y_i] + \mathrm{Var}[\hat{y}_i] - 2\mathrm{Cov}[y_i, \hat{y}_i] + (\mathrm{E}[y_i] - \mathrm{E}[\hat{y}_i])^2 \end{aligned} \quad (3)$$

The covariance term is not (usually) zero, because, $\hat{y}_i$ is a function of y_i.
On the other hand, the i^{th} squared error on new data:

$$\begin{aligned} \mathrm{E}[(y_i' - \hat{y}_i)^2] &= \mathrm{Var}[y_i' - \hat{y}_i] + \left(\mathrm{E}[y_i' - \hat{y}_i]\right)^2 \\ &= \mathrm{Var}[y_i'] + \mathrm{Var}[\hat{y}_i] - 2\mathrm{Cov}[y_i', \hat{y}_i] + \left(\mathrm{E}[y_i'] - \mathrm{E}[\hat{y}_i]\right)^2 \end{aligned}$$

Generalization error

y_i' is independent of y_i, and has the same distribution. Hence, $\mathrm{E}[y_i'] = \mathrm{E}[y_i]$, $\mathrm{Var}[y_i'] = \mathrm{Var}[y_i]$, and $\mathrm{Cov}[y_i', \hat{y}_i] = 0$.

$$\begin{aligned} \mathrm{E}[(y_i' - \hat{y}_i)^2] &= \mathrm{Var}[y_i] + \mathrm{Var}[\hat{y}_i] + (\mathrm{E}[y_i] - \mathrm{E}[\hat{y}_i])^2 \\ &= \mathrm{E}[(y_i - \hat{y}_i)^2] + 2\mathrm{Cov}[y_i, \hat{y}_i], \end{aligned}$$

where the last line follows using (3) from the previous slide.
Averaging over data points,

$$\mathrm{E}[\frac{1}{n}\sum_{i=1}^{n}(y_i' - \widehat{y}_i)^2] = \mathrm{E}[\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2] + \frac{2}{n}\sum_{i=1}^{n}\mathrm{Cov}[y_i, \hat{y}_i]$$

It also holds that: $\mathrm{Cov}[y_i, \hat{y}_i] = \sigma^2 H_{ii}$, so

$$\mathrm{E}[\frac{1}{n}\sum_{i=1}^{n}(y_i' - \widehat{y}_i)^2] = \mathrm{E}[\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2] + \frac{2}{n}\sigma^2\mathrm{tr}[\mathbf{H}]$$

and we know that with p predictors and one intercept, $\mathrm{tr}[\mathbf{H}] = p + 1$. Thus,

$$\mathrm{E}[\frac{1}{n}\sum_{i=1}^{n}(y_i' - \widehat{y}_i)^2] = \mathrm{E}[\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2] + \frac{2}{n}\sigma^2(p+1).$$

Generalization error

Then

$$\mathrm{E}[\frac{1}{n}\sum_{i=1}^{n}(y_i' - \widehat{y}_i)^2] \approx \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 + \frac{2}{n}\sigma^2(p+1)$$

The second term indicates how far away our in-sample MSE estimate is from estimating the **true expected squared error**.

We see that the second term:

- Grows with σ^2.
- Shrinks with n.
- Grows with p.

We usually do not know σ^2, but we can estimate it. This is how we come to our first model selection criteria.

Mallow's C_p statistic, Mallows 1973

For a linear model with $p+1$ coefficients fit by OLS,

$$C_p = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 + \frac{2}{n}\hat{\sigma}^2(p+1)$$

- The Mallows C_p statistic just substitutes in an estimator of σ^2 ($\hat{\sigma}^2$) *always from the largest model we consider*.
- $\hat{\sigma}^2$ is an unbiased estimator of σ^2 if the true model contains a strict subset of the predictor variables.
- The selection rule is to **pick the model which minimizes** C_p.

We can think of C_p as having two parts,

$$C_p = MSE + (\text{penalty})$$

- The penalty can be seen as a cost which we are imposing on models for having extra parameters.
- Every new parameter has to pay that cost by reducing the MSE by a certain amount; and if it doesn't, the extra parameter isn't worth it.

Example: State data

Data frame `state.x77` is a matrix with 50 rows and 8 columns:

- Population - population estimate as of July 1, 1975
- Income - per capita income (1974)
- Illiteracy - illiteracy (1970, percent of population)
- Life Exp - life expectancy in years (1969-71)
- Murder - murder rate per 100,000 population (1976)
- HS Grad - percent high-school graduates (1970)
- Frost - mean days with minimum temperature below freezing (1931-1960)
- Area - land area in square mile

Let's fit a few linear models and compare them using Mallow's C_p.

Example: State data

A function for calculating the Mallows C_p of a linear model is provided in your R script.

```
fit1 <- lm(Life.Exp~., data=statedata) ## all predictors
fit2 <- lm(Life.Exp~Population + Income + Murder,data=statedata)
fit3 <- lm(Life.Exp~Population + Income,data=statedata)

## Let us compare the Mallow's C_p of these three models
Cp.lm(list(fit1,fit2,fit3))
[1] 0.6434448 0.6788404 1.5924723
```

- We prefer the model that minimizes Mallow's C_p.
- Hence, we would choose the model with all predictors.
- The model with only Population, Income and Murder as predictors also performs pretty well.

R squared and adjusted R squared

Recall: In the beginning of the course, we mentioned R squared as a measure of model goodness-of-fit.
Note that

$$R^2 = 1 - \frac{RSS}{SYY} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{n \cdot MSE}{SYY}.$$

So that R^2 suffers from a similar problem as the in sample MSE.
It will always grow with the number of predictors.
As such it is also not a very useful tool for model selection.
An alternative called the **adjusted R squared** was suggested:

$$R^2_{adj} = 1 - \frac{MSE \frac{n}{n-p-1}}{SYY}$$

- the adjusted R squared accounts for the number of parameters
- it is definitely a better model selection and goodness-of-fit criteria compared to R squared, but it is still generally not considered a very good criteria.

Example: State data

How do the three models we fit before perform in terms of R^2 and R^2_{adj}?

```
> summary(fit1)$r.sq
[1] 0.7361563
> summary(fit2)$r.sq
[1] 0.6658573
> summary(fit3)$r.sq
[1] 0.1359421

> summary(fit1)$adj.r.sq
[1] 0.6921823
> summary(fit2)$adj.r.sq
[1] 0.6440654
> summary(fit3)$adj.r.sq
[1] 0.09917365
```

As expected, the largest model has the largest R^2.
We would choose the model that maximizes **the adjusted R square**.Based on the adjusted R square we would also choose the largest model, but the model from fit2 (Income, Population, Murder) perfroms similarly well.

Akaike information criterion (AIC), Akaike 1973

The great Japanese statistician Hirotugu Akaike proposed a famous model selection rule which also has the form of "in-sample performance plus penalty".

What has come to be called the **Akaike information criterion** (AIC) is

$$AIC(model) = 2\dim(model) - 2L_{model},$$

where L_{model} is the maximum value of the log-likelihood for the model.

- Akaike's rule is to pick the model which minimized the AIC.
- The AIC is based on information theory. Akaike showed that AIC can be used to asymptotically estimate, how much more (or less) information is lost by using model f1 compared to model f2 to represent the true model f.

AIC

In linear models with Gaussian noise, the maximum value of the log-likelihood of a model is equal to:

$$L = -\frac{n}{2}(1 + \log 2\pi) - \frac{n}{2}\log RSS/n$$

Hence, for OLS the AIC criterion is of the form:

$$AIC = 2(p+1) + n(1 + log(2\pi)) + n\log RSS/n.$$

And we choose the model which minimizes the AIC.

- An attractive quality of AIC is that it can be applied whenever we have a likelihood.
- It is therefore used for tasks like comparing models of probability distributions, or predictive models where the whole distribution is important.
- C_p, by contrast, really only makes sense if we're trying to do regression and want to use squared error.

Example: State data

How do the three models we fit before perform in terms of AIC? Implemented in R function `AIC()`.

```
> AIC(fit1,fit2,fit3)
     df      AIC
fit1  9 121.7092
fit2  5 125.5198
fit3  4 171.0234
```

- We will prefer the model that minimizes the AIC value.
- In this case, it is again the biggest model (containing all predictors) that wins the comparison.

Bayesian information criterion (BIC), Schwarz 1978

The Bayesian information criterion (BIC) is a similar criterion to AIC. It has the following form:

$$BIC(model) = \log n \cdot \dim(model) - 2L_{model},$$

where L_{model} is again the maximum value of the log-likelihood for the model. We will also choose the model that minimizes the BIC criterion.

- The BIC has a stronger penalty than AIC, so it will tend to choose smaller models than AIC
- The stronger penalty means that as $n \to \infty$, if the true model is among those BIC can select among, BIC will tend to pick the true model. (under some assumptions).
- Unfortunately, the model selected by BIC will tend to predict less well than the one selected by AIC or leave-one-out cross-validation (next criterion).

For linear Gaussian models:

$$BIC = \log n(p+1) + n(1 + log(2\pi)) + n \log RSS/n.$$

Example: State data

How do the three models we fit before perform in terms of BIC? Implemented in R function `BIC()`.

```
> BIC(fit1,fit2,fit3)
     df      BIC
fit1  9 138.9174
fit2  5 135.0799
fit3  4 178.6715
```

- We prefer the model that minimizes the BIC.
- In this case, we would choose the model in fit2 as the best model according to BIC. This is the model containing predictors: Income, Population and Murder.
- The stronger penalty of the BIC means that the BIC does not consider adding the other predictors “worth it” for the small decrease in RSS.

Leave-one-out cross-validation (LOOCV)

- Recall: Studentized residuals, Cook's distance.
  When looking at influential points and outliers, we considered omitting one point from the data set, estimating the model, and then trying to predict that one data point.
- The **leave-one-out** fitted value for data point i is $\hat{y}_{i(i)}$, where the subscript (i) indicates that point i was left out in calculating this fit.

The **leave-one-out cross-validation score** of the model is

$$LOOCV = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_{i(i)})^2$$

We choose the model that **minimizes the LOOCV score.**

- Many older regression textbooks look at $n \cdot LOOCV$, and call it PRESS, "predictive residual sum of squares".
- The story for cross-validation is pretty compelling: we artificially create "new" data to see how well our model can generalize to it.
- In fact, LOOCV is an unbiased estimate of the generalization error.

Leave-one-out cross-validation (LOOCV)

Re-estimating the model n times would be seriously time-consuming, but there is fortunately a short-cut for OLS:

$$LOOCV = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - H_{ii}}\right)^2$$

- The numerator inside the square is just the residual of the model fit to the full data.
- This gets divided by $1 - H_{ii}$, which is also something we can calculate with just one fit to the model. (The denominator says that the residuals for high-leverage points count more. Similar to Cook's distance.)
- The gap between LOOCV and the MSE can be thought of as a penalty, just like with C_p, AIC or BIC.
- $LOOCV \approx \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2(1 - 2H_{ii}) \approx MSE + 2\hat{\sigma}^2 \mathrm{tr}[\mathbf{H}]$
- LOOCV can also be performed with other types of loss functions, for example L1-norm, or zero-one loss function.

Example: State data

How do the three models we fit before perform in terms of *LOOCV*?

```
> loocv.lm(fit1)
[1] 0.8264418
> loocv.lm(fit2)
[1] 0.7030285
> loocv.lm(fit3)
[1] 1.862332
```

- According to the LOOCV score, we should opt for the second model.
- That is, the model containing Income, Population and Murder as predictors of Life Expectancy.

K-fold cross-validation

In leave-one-out cross-validation, we omitted each data point in turn, and tried to predict it. K-fold cross-validation is somewhat different:

- Randomly divide the data into k equally-sized parts, or "folds".
- For each fold
  - Temporarily hold back that fold, calling it the "testing set".
  - Call the other $k-1$ folds, taken together, the "training set".
  - Estimate each model on the training set.
  - Calculate the MSE of each model on the testing set.
- Average MSEs over folds.

We then pick the model with the lowest MSE, averaged across testing sets.
The point of this is just like the point of leave-one-out: the models are compared only on data which they didn't get to see during estimation.
In fact, leave-one-out cross-validation can be seen as as special case of K-fold cross-validation for $K = n$.
Like leave-one-out CV, k-fold cross-validation can be applied to any loss function, such as the proportion of cases miss-classified, or negative log-likelihood.

Example: State data

How do the three models we fit before perform in terms of 5-fold CV?

```
> fit1.cv5
Mean squared error         :  1.003933

> fit2.cv5
Mean squared error         :  0.779691

> fit3.cv5
Mean squared error         :  1.903795
```

- According to the 5-fold CV score, we should opt for the second model again.

Model selection criteria: summary

Shao (1997) classified the model selection criteria as follows:

- Class 1: AIC, C_p, LOOCV;
- Class 2: BIC, K-fold CV

Shao (1997) showed:

- criteria in class 1 are asymptotically valid if there is no fixed-dimensional correct model,
- criteria in class 2 are asymptotically valid if there exists a fixed-dimensional correct model.

Model selection criteria: summary

Summary:

- AIC, C_p and LOOCV will tend to prefer models which are bigger than the true model, even when the true model is available to them. They are "not consistent for model selection".
- Conversely, BIC or K-fold CV (where K depends on the sample size), will tend to choose the correct model if it is available to them. However, their model choice may have somewhat worse predictive performance.
- Since C_p and AIC involve less calculation than leave-one-out, they have advantages when n is large.
- The same can be said for BIC as compared to K-fold LOOCV.
- There is generally no consensus on what criteria is best to use: the cross-validation methods are the most flexible, but they can take a very long time to apply to large data sets.

Model selection procedures

One way to automatically select a model is to begin with the largest model and then prune it:

- Pick your favorite model selection criterion, consider deleting each coefficient in turn, and pick the sub-model with the best value of the criterion.
- Having eliminated a variable, one then re-estimates the model, and repeats the procedure.

Stop when nothing can be eliminated without worsening the criterion.

- This is an example of **backwards** stepwise model selection.
- **Forward** stepwise model selection starts with the intercept-only model and adds variables in the same fashion. That is, at each time adding the variable that most improves the criterion.

Model selection procedures

- There are, naturally, forward-backward hybrids. For these you can start with any model and at each step either choose to **delete** a variable from the model, or **add** a variable to the model based on which action would most improve your criterion.
- Implemented with AIC and BIC in R function `step()`.
- All of these procedures are “greedy”. That is, they will not necessarily select the best possible model.

Another option is **all subsets regression**.

- For p predictors there are 2^p possible subsets of predictors.
- So this method is not feasible for a large p.
- Function `regsubsets` in R package `leaps`.

Inference after model selection

All of the inferential statistics in earlier lectures presumed that our choice of model was fixed, and not dependent on the data.
Because we now pick the model to predict well on our data, if we then run hypothesis tests on that same data, they’ll be too likely to tell us everything is significant, and our confidence intervals will be too narrow.

- Post selection inference - Currently an active area of research on how we should “modify” p-values and hypothesis test for after model selection.

For us, the suggestion to deal with this issue: **Data splitting**.

- Randomly divide your data set into two parts.
- Calculate your favorite model selection criterion for all your candidate models using only the first part of the data. Pick one model as the winner.
- Re-estimate the winner, and calculate all your inferential statistics, using only the other half of the data.

(Division into two equal halves is optional, but usual.)

Why not model selection based on p-values?

For a single coefficient β_i, testing whether $\beta_i = 0$ involves the t-statistic:

$$\frac{\hat{\beta}_i}{SE(\hat{\beta}_i)} = \frac{\hat{\beta}_i}{\frac{\hat{\sigma}}{\sqrt{n\widehat{\text{Var}}[x_i]}}VIF_i} = \frac{\hat{\beta}_i}{\frac{\hat{\sigma}}{\sqrt{\sum_{j=1}^{n}(x_{ji}-\bar{x}_i)^2}}VIF_i}$$

- Larger coefficients will have larger test statistics and be more significant ($\hat{\beta}_i$ in the numerator).
- Reducing the noise around the regression line will increase all the test statistics, and make every variable more significant ($\hat{\sigma}$ in the denominator).
- Increasing the sample size will increase all the test statistics, and make every variable more significant ($\sqrt{n}$ in the denominator).
- More variance in a predictor variable will increase the test statistic and make the variable more significant.
- More correlation between x_i and the other predictors will decrease the test statistic and make the variable less significant (VIF_i in the denominator).

Why not model selection based on p-values?

- The test statistic, and thus the p-value, runs together an estimate of the actual size of the coefficient with how well we can measure that particular coefficient.
- That is a very, very different question from “Is this variable truly relevant to the response?”
- Utterly trivial variables can show up as having highly significant coefficients, if the predictor has a lot of variance and isn’t very correlated with the other predictors.
- Every variable whose coefficient isn’t exactly zero will eventually (as $n \to \infty$) have an arbitrarily large test statistic, and an arbitrarily small p-value.
- This is why we prefer to use the model selection procedures described in this lecture to select predictors which will yield a model that has a good predictive performance.