Section 6: Cross — Validation

Yotam Shem-Tov

Fall 2014
There are two types of Prediction errors: In sample prediction error and out of sample prediction error.

In sample prediction error: how well does the model explain the data which is used in order to estimate the model.

Consider a sample, \((y, X)\), and fit a model \(f(\cdot)\) (for example a regression model), and denote the fitted values by \(\hat{y}_i\).

In order to determine how well the model fits the data, we need to choose some criterion, which is called the loss function, i.e \(L(y_i, \hat{y}_i)\).

Standard loss functions:

\[
MSE = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2, \quad RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2}
\]
Out of sample prediction error

- How well can the model predict a value of $y_j$ given $x_j$ where observation $j$ is not in the sample. This is referred to as the out of sample prediction error.
- How can we estimate the out of sample prediction error?
- The most commonly used method is Cross-Validation.
Summary of the approach:

1. Split the data into a training set and a test set
2. Build a model on the training data
3. Evaluate on the test set
4. Repeat and average the estimated errors

Cross-Validation is used for:

1. Choosing model parameters
2. Model selection
3. Picking which variables to include in the model
There are 3 common CV methods, in all of them there is a trade-off between the bias and variance of the estimator.

1. Random sub-sampling CV

2. K-fold CV

3. Leave one out CV (LOOCV)

My preferred method is *Random sub-sampling CV*. 
Random sub-sampling CV

1. Randomly split the data into a test set and training set.
2. Fit the model using the training set, *without using the test set at all!*
3. Evaluate the model using the test set.
4. Repeat the procedure multiple times and average the estimated errors (RMSE).

What is the tuning parameter in this procedure? The *fraction* of the data which is used as a test set. There is no common choice of *fraction* to use. My preferred choice is 50%, however this is arbitrary.
Recall the dilemma of choosing a P-score model: with or without interactions.

**No interactions**

**With interactions**
Random sub-sampling CV: Example

We can use CV in order to choose between the two competing models.

L0=100 # number of repetitions
rmse.model.1 <- rmse.model.2 <- rep(NA,L0)
a = data.frame(treat=treat,x)
for (j in c(1:L0)){
  id = sample(c(1:dim(d)[1]),round(dim(d)[1]*0.5))
  ps.model1 <- glm(treat~(.),data=a[id,],family=binomial(link=logit))
  ps.model2 <- glm(treat~(.)^2,data=a[id,],family=binomial(link=logit))
  rmse.model.1[j]=rmse(predict(ps.model1,newdata=a[-id,], type="response"),a$treat[-id])
  rmse.model.2[j]=rmse(predict(ps.model2,newdata=a[-id,], type="response"), a$treat[-id])
}
The results are in the table below:

<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th>Model 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.29</td>
<td>0.33</td>
</tr>
<tr>
<td>Median</td>
<td>0.29</td>
<td>0.32</td>
</tr>
</tbody>
</table>

It is clear that model 1, no interactions, has a lower out of sample prediction error.

Model 2 (with interactions) over fits the data, and generates a model with a wrong P-score. The model includes too many covariates.

Note, it is also possible to examine other models that include some of the interactions, but not all of them.
Randomly split the data into $K$ folds (groups)

Estimate the model using $K - 1$ folds

Evaluate the model using the remaining fold.

Repeat the process by the number of folds, $K$ times

Average the estimated errors across folds

The choice of $K$, is a classic problem of bias-variance trade-off.

What is the tuning parameter in this method? The *number of folds*, $K$. There is no common choice of $K$ to use. Commonly used choices are, $K = 10$, and $K = 20$. The choice of $K$ depends on the size of the sample, $N$. 

Yotam Shem-Tov

STAT 239/ PS 236A
The tuning parameter

- $K$ folds,

**Choosing the number of folds, $K$**

\[ \uparrow K \text{ lower bias, higher variance} \]
\[ \downarrow K \text{ higher bias, lower variance} \]

- Random sub-sampling,

**Choosing the fraction of the data in the test set**

\[ \downarrow \text{fraction lower bias, higher variance} \]
\[ \uparrow \text{fraction higher bias, lower variance} \]
Leave one out CV (LOOCV)

- LOOCV is a specific case of $K$ folds CV, where $K = N$

- Example in which there is an analytical formula for the LOOCV statistic

- The model: $Y = X\beta + \varepsilon$

- The OLS estimator: $\hat{\beta} = (X'X)^{-1}X'y$

- Define the hat matrix as, $H = X(X'X)^{-1}X'$

- Denote the elements on the diagonal of $H$, as $h_i$

- The LOOCV statistic is,

$$CV = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{e_i}{1 - h_i}\right)^2$$

where $e_i = y_i - x_i'\hat{\beta}$, and $\hat{\beta}$ is the OLS estimator over the whole sample
The CV methods discussed so far do not work when dealing with time series data.

The dependence across observations generates a structure in the data, which will be violated by a random split of the data.

Solutions:

1. An iterated approach of CV
2. Bootstrap 0.632 (?)
Summary of the iterated approach:

1. Build a model using the first $M$ periods
2. Evaluate the model on period $t = (M + 1) : T$
3. Build a model using the first $M + 1$ periods
4. Evaluate the model on period $t = (M + 2) : T$
5. Continue iterating forward until, $M + 1 = T$
6. Average over the estimated errors
We want to predict the GDP growth rate in California in 2014.

The available data is *only* the growth rates in the years 1964 – 2013.

Consider the following three possible Auto-regression models:

1. \( y_t = \alpha + \beta_1 y_{t-1} \)
2. \( y_t = \alpha + \beta_1 y_{t-1} + \beta_2 y_{t-2} \)
3. \( y_t = \alpha + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \beta_3 y_{t-3} \)
Example: The data

![Graph showing the relationship between growth and year from 1970 to 2010. The data points are scattered, and a trend line is drawn through them. There is a shaded area indicating confidence intervals.]
Example: estimation of the three models

<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.954*</td>
<td>1.935*</td>
<td>1.411</td>
</tr>
<tr>
<td></td>
<td>(0.841)</td>
<td>(0.919)</td>
<td>(0.977)</td>
</tr>
<tr>
<td>Lag 1</td>
<td>0.717***</td>
<td>0.710***</td>
<td>0.716***</td>
</tr>
<tr>
<td></td>
<td>(0.103)</td>
<td>(0.149)</td>
<td>(0.149)</td>
</tr>
<tr>
<td>Lag 2</td>
<td>0.014</td>
<td>−0.145</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.150)</td>
<td>(0.182)</td>
<td></td>
</tr>
<tr>
<td>Lag 3</td>
<td></td>
<td></td>
<td>0.217</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.150)</td>
</tr>
<tr>
<td>R²</td>
<td>0.505</td>
<td>0.509</td>
<td>0.534</td>
</tr>
<tr>
<td>Adj. R²</td>
<td>0.495</td>
<td>0.487</td>
<td>0.502</td>
</tr>
<tr>
<td>Num. obs.</td>
<td>49</td>
<td>48</td>
<td>47</td>
</tr>
</tbody>
</table>

***p < 0.001, **p < 0.01, *p < 0.05
Example: choice of model

- Which of the models will you choose?
- Will you use an F-test?
- What is your guess: which of the models will have a lower out of sample error, using CV?
Example: F-test I

- Note, in order to conduct an F-test, we need to drop the first 3 observations. This is in order to have the same data used in the estimation of all three models.
- Dropping the first 3 observations, might biased our results in favour of models 2 and 3, relative to model 1.

Analysis of Variance Table

Model 1: \( y \sim \text{lag1} + \text{lag2} \)
Model 2: \( y \sim \text{lag1} + \text{lag2} + \text{lag3} \)

<table>
<thead>
<tr>
<th>Res.Df</th>
<th>RSS</th>
<th>Df</th>
<th>Sum of Sq</th>
<th>F</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>330.02</td>
<td>44</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>314.58</td>
<td>43</td>
<td>15.438</td>
<td>2.1102</td>
<td>0.1536</td>
</tr>
</tbody>
</table>
Example: F-test II

Analysis of Variance Table

Model 1: y ~ lag1  
Model 2: y ~ lag1 + lag2  

<table>
<thead>
<tr>
<th>Res.Df</th>
<th>RSS</th>
<th>Df</th>
<th>Sum of Sq</th>
<th>F</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>330.03</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>330.02</td>
<td>1</td>
<td>0.012439</td>
<td>0.0017</td>
<td>0.9677</td>
</tr>
</tbody>
</table>

Yotam Shem-Tov  
STAT 239/ PS 236A
### Analysis of Variance Table

**Model 1:** $y \sim \text{lag1}$  
**Model 2:** $y \sim \text{lag1} + \text{lag2} + \text{lag3}$

<table>
<thead>
<tr>
<th>Res.Df</th>
<th>RSS</th>
<th>Df</th>
<th>Sum of Sq</th>
<th>F</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>330.03</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>314.58</td>
<td>2</td>
<td>15.45</td>
<td>1.0559</td>
<td>0.3567</td>
</tr>
</tbody>
</table>

Yotam Shem-Tov  
STAT 239/ PS 236A
We used the iterative approach, as this is time series data. $M$ is the number of periods used for fitting the model before starting the CV procedure. The average RMSE are,

<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M = 5$</td>
<td>27.266</td>
<td>27.078</td>
<td>26.994</td>
</tr>
<tr>
<td>$M = 10$</td>
<td>29.770</td>
<td>29.586</td>
<td>29.474</td>
</tr>
<tr>
<td>$M = 15$</td>
<td>33.106</td>
<td>32.924</td>
<td>32.797</td>
</tr>
</tbody>
</table>

Among Model 1 and Model 2 only, which is preferable?
The tuning parameter in time series CV

What is the bias-variance trade-off in the choice of $M$?

<table>
<thead>
<tr>
<th>Choice of $M$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\uparrow M$</td>
</tr>
<tr>
<td>$\downarrow M$</td>
</tr>
</tbody>
</table>
For a survey of cross-validation results, see Arlot and Celisse (2010),
http://projecteuclid.org/euclid.ssu/1268143839