Predictive Analytics

Week 6: Model Selection and Estimation III

Semester 2, 2018

Discipline of Business Analytics, The University of Sydney Business School

Week 6: Model Selection and Estimation III

1. Maximum likelihood (continued)
2. Inference for the ML estimator (optional)
3. Analytical criteria
4. Comparison of model selection methods
5. Limitations of model selection
6. Optimism (optional)

Reading: Chapter 6.1 of ISL.

Exercise questions: Chapter 6.1 of ISL, Q1. Try to answer this question based on your existing knowledge of regression variable selection, which will help you be prepared for the next lecture.

Assignment Project Exam Help

Maximum likelihood (continued)

https://powcoder.com

Add WeChat powcoder

Maximum likelihood estimation (MLE), which we have discussed in the context of linear regression is one of the most important concepts in statistics. We now present it more generally for inference.

ML for discrete distributions (key concept)

Let $p(y;\theta)$ be a discrete probability distribution. The likelihood function is

$$
\begin{aligned}
\ell(\theta) &= P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_N = y_N) \\
&= P(Y_1 = y_1)\, P(Y_2 = y_2) \ldots P(Y_N = y_N) \\
&= \prod_{i=1}^{N} p(y_i; \boldsymbol{\theta})
\end{aligned}
$$

The maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$ is the value of $\boldsymbol{\theta}$ that maximises $\ell(\boldsymbol{\theta})$.

ML for continuous distributions (key concept)

Let $p(y;\theta)$ be a density function. The likelihood function is

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= p(y_1;\boldsymbol{\theta})\,p(y_2;\boldsymbol{\theta})\,\ldots\,p(y_N;\boldsymbol{\theta})\\ &= \prod_{i=1}^{N} p(y_i;\boldsymbol{\theta})\end{aligned}$$

The maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$ is the value of $\boldsymbol{\theta}$ that maximises $\ell(\boldsymbol{\theta})$.

Maximum likelihood

- Even though $\ell(\boldsymbol{\theta})$ equals an expression that involves $p(y_i; \boldsymbol{\theta})$, we think of these functions in different ways.

- When considering a probability mass function or density $p(y; \boldsymbol{\theta})$, we consider y to be a variable, and $\boldsymbol{\theta}$ to be fixed.

- In the likelihood, $\boldsymbol{\theta}$ is a variable, and y is fixed.

Log-likelihood (key concept)

The log-likelihood is

$$L(\boldsymbol{\theta}) = \log \left(\prod_{i=1}^{N} p(y_i; \boldsymbol{\theta}) \right)$$

$$= \sum_{i=1}^{N} \log p(y_i; \boldsymbol{\theta})$$

Because the log-likelihood is a **monotonic** transformation of the likelihood, maximising it is the same as maximising the likelihood.

Example: Bernoulli distribution

Suppose that $Y_1, \ldots, Y_N$ follow the Bernoulli distribution with parameter θ (the probability of a success).

$$p(y_i;\theta) = \theta^{y_i}(1-\theta)^{(1-y_i)}$$

$$\ell(\theta) = \prod_{i=1}^{N} \theta^{y_i}(1-\theta)^{(1-y_i)}$$

$$\begin{aligned} L(\theta) &= \sum_{i=1}^{N} \left[y_i \log(\theta) + (1-y_i)\log(1-\theta) \right] \\ &= \left(\sum y_i\right)\log(\theta) + (N - \sum y_i)\log(1-\theta) \end{aligned}$$

Example: Bernoulli distribution

Derivative of the log-likelihood with respect to θ:

$$\frac{dL(\theta)}{d\theta} = \frac{\sum y_i}{\theta} - \frac{N - \sum y_i}{1-\theta}$$

The ML estimate therefore satisfies

$$\frac{\sum y_i}{\widehat{\theta}} = \frac{N - \sum y_i}{1-\widehat{\theta}}.$$

The solution is the sample proportion:

$$\widehat{\theta} = \frac{\sum_{i=1}^{n} y_i}{N}.$$

What about the 2nd order derivative?

Inference for the ML estimator (optional)

Inference for the ML estimator

The **score function** is

$$s(\widehat{\boldsymbol{\theta}}) = \nabla L(\boldsymbol{\theta})\big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}$$

For example, when the parameter is a scalar

$$s(\widehat{\theta}) = \sum_{i=1}^{n} \frac{d \log p(y_i;\theta)}{d\theta}\bigg|_{\theta=\widehat{\theta}}$$

The **observed information matrix** is the negative of the second derivative (the Hessian matrix) of the log-likelihood:

$$\boldsymbol{J}(\widehat{\boldsymbol{\theta}}(\mathcal{D})) = -\nabla^2_{\boldsymbol{\theta}} L(\boldsymbol{\theta})\big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}$$

When the parameter is a scalar,

$$J(\widehat{\theta}) = \sum_{i=1}^{N} \frac{d^2 \log p(y_i)}{d\theta^2}\bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}.$$

We define the **Fisher information matrix** as the expected value of the observed information matrix

$$\boldsymbol{I}_n(\widehat{\boldsymbol{\theta}}) = E_{\boldsymbol{\theta}}\left[\boldsymbol{J}(\widehat{\boldsymbol{\theta}}(\mathcal{D}))\right]$$

So the **observed information matrix** is a sample-based version of the Fisher information matrix.

Inference for the ML estimator

A standard result shows that the sampling distribution of the ML converges to the normal distribution

$$\widehat{\boldsymbol{\theta}} \to N(\theta, \boldsymbol{I}_n^{-1}(\boldsymbol{\theta}))$$

as $n \to \infty$.

That suggests the large sample approximations

$$N(\boldsymbol{\theta}, \boldsymbol{I}_N(\widehat{\boldsymbol{\theta}})^{-1}) \text{ or } N(\boldsymbol{\theta}, \boldsymbol{J}(\widehat{\boldsymbol{\theta}})^{-1})$$

Example: Bernoulli distribution

Continuing the example, the observed information matrix is

$$-\frac{d^2L(\theta)}{d\theta^2} = \frac{\sum y_i}{\theta^2} + \frac{n - \sum y_i}{(1-\theta)^2}$$

Since $E(Y) = \theta$,

$$E(J(\theta)) = \frac{N}{\theta(1-\theta)},$$

so that

$$I_N^{-1} = \frac{\theta(1-\theta)}{N},$$

which is familiar as the variance of a sample proportion from basic statistics.

Inference for the ML estimator

The corresponding estimates for the standard errors of individual parameters are

$$\mathsf{SE}(\widehat{\theta}_j) = \sqrt{\boldsymbol{I}_N(\widehat{\boldsymbol{\theta}})^{-1}_{jj}} \text{ or } \mathsf{SE}(\widehat{\theta}_j) = \sqrt{\boldsymbol{J}(\widehat{\boldsymbol{\theta}})^{-1}_{jj}}.$$

A large sample $100 \times (1-\alpha)\%$ confidence interval is

$$\widehat{\theta}_j \pm z_{\alpha/2} \times \mathsf{SE}(\widehat{\theta}_j)$$

The following large sample approximation leads to accurate confidence intervals and hypothesis tests

$$2\left(L(\widehat{\boldsymbol{\theta}}) - L(\boldsymbol{\theta})\right) \sim \chi^2_d,$$

where d is the number of parameters in $\boldsymbol{\theta}$.

Analytical criteria

Analytical criteria

Analytic criteria provide estimate the generalisation error based on theoretical arguments. They have the form:

criterion = training loss + penalty for number of parameters

Mallow's C_p statistic

The **Mallow's** C_p statistic applies to linear regression. It directly implements the recipe suggested by our calculation of the optimism:

$$C_p = \frac{\text{RSS}}{N} + \frac{2}{N}\widehat{\sigma}^2(p+1),$$

In the formula, $\widehat{\sigma}^2$ is an estimate of variance of the errors based on the largest model under consideration.

Mallow's C_p statistic

We select the model with the lowest C_p. To compare two specifications,

$$\Delta C_p = \text{MSE}_1 - \text{MSE}_2 + \frac{2}{N}\widehat{\sigma}^2(p_1 - p_2).$$

The **Akaike information criterion** (**AIC**) applies to models estimated by maximum likelihood.

$$\text{AIC} = -2L(\widehat{\theta}) + 2p,$$

where $L(\widehat{\theta})$ is the maximised log-likelihood and p is the number of estimated parameters. We select the model with the lowest AIC.

Akaike Information Criterion

- The AIC is one of the most popular and versatile strategies for model selection.

- The formula follows the in-sample performance plus penalty for complexity structure.

- The AIC has a rigorous theoretical justification which we not address here. However, keep in mind that it is an asymptotic approximation ($N \to \infty$).

AIC for linear regression

In the special case of comparing linear regression specifications under Gaussian errors, the AIC simplifies to (up to proportionality and ignoring constant terms in the log-likelihood):

$$\text{AIC} = \log\left(\frac{\mathsf{RSS}}{N}\right) + \frac{2}{N}(p+2),$$

The number of parameters is $d = p + 2$ because the parameter vector includes the constant and the variance of errors. Note that this is different from the formula in the book, which is a simplification with unknown error.

Relation between Mallow's C_p and the AIC

For a linear regression with Gaussian errors and known variance:

$$\text{AIC} = \frac{1}{\widehat{\sigma}^2}\left(\frac{\text{RSS}}{N} + \frac{2}{N}\widehat{\sigma}^2(p+1)\right),$$

which compares to

$$C_p = \frac{\text{RSS}}{N} + \frac{2}{N}\widehat{\sigma}^2(p+1).$$

Hence, the AIC and C_p lead to the same decision in this case. For practical purposes, the AIC and C_p are regarded as the same for linear regression.

Bayesian information criterion

The **Bayesian information criterion** (**BIC**) also applies to models estimated by maximum likelihood.

$$\text{BIC} = -2L(\widehat{\theta}) + \log(N)p$$

where $L(\widehat{\theta})$ is the maximised log-likelihood, p is the number of estimated parameters and N is the sample size. We select the model with the lowest BIC.

Bayesian information criterion

- The BIC formula is comparable to the AIC case, but with 2 penalty factor replaced by $\log(N)$. Hence, the BIC penalises complexity more heavily when $N \geq 8$. The BIC has a very different theoretical justification to the AIC.

- The BIC is an asymptotic approximation to a Bayesian approach to model selection.

BIC: Gaussian linear regression case

In the special case of a linear regression with Gaussian errors, the BIC simplifies to (ignoring constant terms)

$$\text{BIC} = \log\left(\frac{\text{RSS}}{N}\right) + \frac{\log(N)}{N}(p+2)$$

If we assume that the variance of the errors is known, we have instead

$$\text{BIC} = \frac{1}{\hat{\sigma}^2}\left(\frac{\text{RSS}}{N} + \frac{\log(N)}{N}\hat{\sigma}^2(p+1)\right),$$

In this case the BIC is proportional to AIC and C_p, but with a $\log(N)$ penalty factor instead of 2.

Comparison of model selection methods

Model selection properties

Consistency. In a collection of models that includes the correct model, the probability that the model selection criterion chooses the correct one approaches one when $N \to \infty$.

Efficiency. The selected model predicts as well as the theoretically best model under consideration in terms of expected loss when $N \to \infty$.

It is not possible to combine these properties (Claeskens and Hjort, 2008, Section 4.9).

Properties of model selection methods

LOOCV, AIC, and Mallow's C_p. Efficient but not consistent. The efficiency follows because they construct unbiased estimators of the test error. However, they select models that are strictly more complex than the true model when $N \to \infty$.

BIC. Consistent under some conditions but not efficient. It often chooses models that are too simple because of its heavier penalty on complexity.

LOOCV, AIC and C_p

- LOOCV, AIC, and C_p are equivalent when $N \to \infty$. They will pick the same model in practice when N is large.

- In finite samples, we can view AIC and C_p as theoretical approximations to LOOCV.

- The advantage of AIC and C_p over LOOCV is mainly computational. CV should be preferred to AIC when the assumptions of the model (e.g., constant error variance) are likely to be wrong.

- LOOCV is universally applicable, while this is not the case for AIC and C_p.

Limitations of model selection

Limitations of model selection

- Standard statistical inference is no longer valid after model selection.

- The reason is that standard inference assumes a fixed model, whereas model selection will by definition pick the specific model that best fits the sample. This will lead to optimistic estimates of sample variation based on the chosen model.

- In our context, the only way around this difficulty would be data splitting: using one part of the sample for model selection, and another for inference.

Limitations of model selection

Model selection is an important tool in your data analysis process, but should not be a replacement to model building through EDA, diagnostics, and domain knowledge.

Review questions

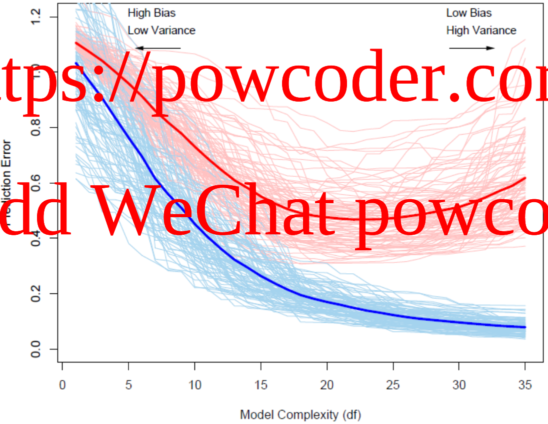
- What are the Akaike Information Criterion, Bayesian Information Criterion and Mallow's C_p?

- What are the relationships between the above 3 metrics?
- Why is it incorrect to conduct statistical inference after model selection (using the same data)?

Optimism (optional)

Optimism

Our objective in this section is to develop a better understanding of overfitting. This discussion will inform our understanding of analytical criteria in the next section.



We define the **training error** as the empirical loss for the training data,

$$\overline{\text{err}}_{\mathcal{D}} = \frac{1}{N}\sum_{i=1}^{N} L(Y_i, \widehat{f}(\boldsymbol{x}_i)).$$

We focus on our standard regression setting,

$$\overline{\text{err}}_{\mathcal{D}} = \frac{1}{N}\sum_{i=1}^{N} \left(Y_i - \widehat{f}(\boldsymbol{x}_i)\right)^2,$$

which is RSS$/N$ for linear regression estimated by least squares.

Optimism

The **expected prediction error** (EPE) is

$$\begin{aligned}\text{EPE}(\boldsymbol{x}_0) &= E_{\mathcal{D}}\left[\left(Y_0 - \widehat{f}(\boldsymbol{x}_0)\right)^2\right] \\ &= E_{\mathcal{D}}\left[\left(f(\boldsymbol{x}_0) + \varepsilon_0 - \widehat{f}(\boldsymbol{x}_0)\right)^2\right] \\ &= \sigma^2 + E_{\mathcal{D}}\left[\left(f(\boldsymbol{x}_0) - \widehat{f}(\boldsymbol{x}_0)\right)^2\right]\end{aligned}$$

where $\boldsymbol{x}_0$ is fixed.

Optimism

Now, consider the estimation error for a training case i,

$$E_{\mathcal{D}}\left(Y_i - \widehat{f}(\boldsymbol{x}_i)\right)^2 = E_{\mathcal{D}}\left[\left(f(\boldsymbol{x}_i) + \varepsilon_i - \widehat{f}(\boldsymbol{x}_i)\right)^2\right]$$

$$= \sigma^2 + E_{\mathcal{D}}\left[\left(f(\boldsymbol{x}_i) - \widehat{f}(\boldsymbol{x}_i)\right)^2\right] - 2E_{\mathcal{D}}\left[\widehat{f}(\boldsymbol{x}_i)\varepsilon_i\right]$$

$$= \sigma^2 + E_{\mathcal{D}}\left[\left(f(\boldsymbol{x}_i) - \widehat{f}(\boldsymbol{x}_i)\right)^2\right] - 2\text{Cov}(\widehat{f}(\boldsymbol{x}_i), \varepsilon_i)$$

Unlike in the EPE, the last term appears because the estimator $\widehat{f}(\boldsymbol{x}_i)$ is a function of $\mathcal{D}$, which includes training case i itself.

Optimism

Averaging over the data, the expected value of the training error is

$$
\begin{aligned}
E\left[\mathsf{err}_{\mathcal{D}}\right] &= \frac{1}{N}\sum_{i=1}^{N} E_{\mathcal{D}}\left[\left(Y_i - \widehat{f}(\boldsymbol{x}_i)\right)^2\right] \\
&= \sigma^2 + \frac{1}{N}\sum_{i=1}^{N} E_{\mathcal{D}}\left(f(\boldsymbol{x}_i) - \widehat{f}(\boldsymbol{x}_i)\right)^2 \\
&\quad - \frac{2}{N}\sum_{i=1}^{N} \mathsf{Cov}(\widehat{f}(\boldsymbol{x}_i), \varepsilon_i).
\end{aligned}
$$

Because of the last term, the training error is not a good estimate of the expected prediction error.

Optimism

We define the out of sample error as

$$\overline{\mathsf{Err}}_{\mathsf{out}} = \frac{1}{N}\sum_{i=1}^{N} E\left[\left(Y_i^0 - \widehat{f}(x_i)\right)^2\right],$$

$$= \sigma^2 + \frac{1}{N}\sum_{i=1}^{N} E_{\mathcal{D}}\left[\left(f(x_i) - \widehat{f}(x_i)\right)^2\right],$$

where $Y_i^0 = f(\boldsymbol{x}_i) + \varepsilon_i^0$ indicates an independent case for a given $\boldsymbol{x}_i$.

Optimism

The **optimism** of the training error is

$$\overline{\text{Err}}_{\text{out}} - E(\overline{\text{Err}}_{\mathcal{D}}) = \frac{2}{N}\sum_{i=1}^{N} \text{Cov}(\widehat{f}(x_i), \varepsilon_i)$$

The more we overfit, the higher Cov$(\widehat{f}(x_i), \varepsilon_i$ will be, increasing the optimism.

Example: linear regression

For the linear regression model, we can show that

$$\text{Optimism} = \frac{2}{N}\sum_{i=1}^{N}\text{Cov}\left(\widehat{f}(x_i), \varepsilon_i\right) = \frac{2}{N}\sigma^2(p+1)$$

Interpretation:

- The larger the sample size (N), the harder it is to overfit.
- The larger the variance of the errors (σ^2), the larger the overfitting.
- The optimism is proportional to the number of predictors.