9. Model Selection

- statistical models
- overview of model selection
- information criteria
- goodness-of-fit measures

Statistical models

- probability distribution models: normal, binomial, poisson, etc.
- conditional distribution models: an RV depends on other variables
- time series model: autoregressive moving average models

Probability distribution model examples

observed data, $y_1, y_2, \ldots, y_N$ are samples from RV Y

problem: construct a probability model based on observed data

- mixture of normal distribution models

$$f(y|m,\theta) = \sum_{j=1}^{m} a_j \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\ -\frac{(y-\mu_j)^2}{2\sigma_j^2}, \quad -\infty < y < \infty$$

- Poisson distribution model

$$f(y|\lambda) = \frac{\lambda^y}{y!} e^{-\lambda}, \quad y = 0, 1, 2, \ldots (\lambda > 0)$$

estimate the **parameters** of the distribution ($e.g.$, μ_j, σ_j, λ) and the resulting density $f(y|m,\hat{\theta})$ is a **statistical model**

Conditional distribution model examples

the distribution of the random variable Y depends on *other* variables, X

we express $f(y|x)$ as the **conditional distribution model**

- linear regression model: y_i is a linear function of x_i (explanatory variables)

$$y = X\beta + u$$

- nonlinear regression model: y is a nonlinear function of x and β

$$y = m(x, \beta) + u$$

$(e.g.,\ y = e^{x^T\beta} + u)$

we estimate the parameter β and the conditional density function $f(y|x, \hat{\beta})$ is a **statistical model**

Time series models

observed data $y_1, y_2, \ldots, y_N$ for events that vary with time is called **time series**

to analyze such time series we consider the conditional distribution

$$f(y_n|y_{n-1}, y_{n-2}, \ldots)$$

given observations up to the time $n-1$

- ARMA model: u is noise

$$y(t) = a_1 y(t-1) + \cdots + a_p y(t-p) + u(t) + b_1 u(t-1) + \cdots + b_q u(t-q)$$

- AR model: y does not depend on past noise

$$y(t) = a_1 y(t-1) + \cdots + a_p y(t-p) + u(t)$$

- MA model: y does not depend on its own past

$$y(t) = u(t) + b_1 u(t-1) + \cdots + b_q u(t-q)$$

when time series data are given and a model order (p or q) is choosen

- model parameters: a_i and b_i and
- noise variance σ^2

can be estimated by LS or ML methods (depending on the model)

Overview

objective of model selection: obtain a good model at a low cost

1. **quality of the model:** defined by a measure of the goodness, e.g., the mean-squared error
   - MSE consists of a *bias* and a *variance* contribution
   - to reduce the bias, one has to use more flexible model structures (requiring more parameters)
   - the variance typically increases with the number of estimated parameters
   - the best model structure is therefore a trade-off between *flexibility* and *parsimony*

2. **price of the model:** an estimation method (which typically results in an optimization problem) highly depends on the model structures, which influences:
   - algorithm complexity
   - properties of the loss function
3. intended use of the model, $e.g.$,
   - summarize the main features of a complex reality
   - predict some outcome
   - test some important hypothesis

Bias-Variance decomposition

assume that the observation Y obeys

$$Y = f(X) + \nu, \quad \mathbf{E}\nu = 0, \quad \mathbf{cov}(\nu) = \sigma^2$$

the mean-squared error of a regression fit $\hat{f}(X)$ at $X = x$ is

$$\begin{aligned}
\text{MSE} &= \mathbf{E}[(Y - \hat{f}(X))^2 | X = x] \\
&= \sigma^2 + [\mathbf{E}\hat{f}(X) - f(X) | X = x]^2 + \mathbf{E}[\hat{f}(X) - \mathbf{E}\hat{f}(X) | X = x]^2 \\
&= \sigma^2 + \text{Bias}^2(\hat{f}(x)) + \text{Var}(\hat{f}(x))
\end{aligned}$$

- this relation is known as **bias-variance decomposition**
- no matter how well we estimate $f(x)$, σ^2 represents *irreducible error*
- typically, the more complex we make model $\hat{f}$, the lower the bias, but the higher the variance

proof of bias-variance decomposition: note that

- the true f is deterministic
- $\mathbf{var}(Y|X=x) = \sigma^2$ and $\mathbf{E}[Y|X=x] = f(x)$
- $\hat{f}(x)$ is random

we will omit the notation of conditioning on $X = x$

$$\begin{aligned}
\mathbf{E}[(Y - \hat{f}(X))^2] &= \mathbf{E}[Y^2] + \mathbf{E}[\hat{f}(x)^2] - \mathbf{E}[2Y\hat{f}(x)] \\
&= \mathbf{var}(Y) + \mathbf{E}[Y]^2 + \mathbf{var}\,\hat{f}(x) + \mathbf{E}[\hat{f}(x)]^2 - 2f(x)\mathbf{E}[\hat{f}(x)] \\
&= \mathbf{var}(Y) + f(x)^2 + \mathbf{var}\,\hat{f}(x) + \mathbf{E}[\hat{f}(x)]^2 - 2f(x)\mathbf{E}[\hat{f}(x)] \\
&= \sigma^2 + \mathbf{var}\,\hat{f}(x) + (f(x) - \mathbf{E}[\hat{f}(x)])^2 \\
&= \sigma^2 + \mathbf{var}\,\hat{f}(x) + (\mathbf{E}[f(x) - \hat{f}(x)])^2 \\
&= \sigma^2 + \mathbf{var}\,\hat{f}(x) + [\mathrm{Bias}(\hat{f}(x))]^2
\end{aligned}$$

Example

consider a problem of fitting polynomial of degree 0 and 1 to data:

$$\hat{y}(t) = a_0 \quad \text{VS} \quad \hat{y}(t) = c_0 + c_1 t$$

a linear regression model $y = X\beta$ gives the error covariance

$$\mathbf{cov}\,\hat{\beta} = (X^T X)^{-1} \quad \text{(assume } X \text{ is nonrandom and error is homoskedastic)}$$

therefore, the variances of a_0 and c_0 are given by

$$1/N = \mathbf{var}(\hat{a}_0) \quad < \quad \mathbf{var}(\hat{c}_0) = \frac{\sum_i t_i^2}{N \sum_i t_i^2 - (\sum_i t_i)^2}$$

(more complex model gives higher variance)

Overfitting

we will explain by an example of AR model with white noise ν

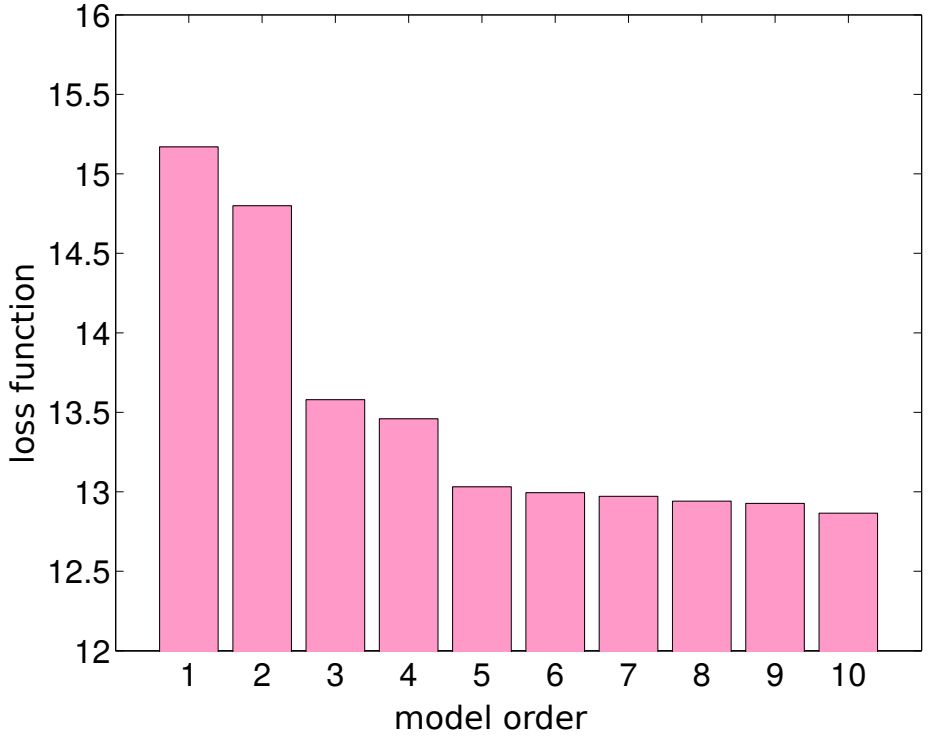
$$y(t) + a_1 y(t-1) + \cdots + a_p y(t-p) = \nu(t)$$

- true AR model has order $p = 5$
- the parameters to be estimated are $\theta = (a_1, a_2, \ldots, a_p)$ with p unknown
- question: how to choose a proper value of p ?
- define a quadratic loss function

$$f(\theta) = \sum_{t=p+1}^{N} |y(t) - (a_1 y(t-1) + \ldots + a_p y(t-p))|^2$$

and obtain $\hat{\theta}$ by using the LS method:

$$\hat{\theta} = (\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_p) = \underset{\theta}{\operatorname{argmin}}\, f(\theta)$$



- the minimized loss is a decreasing function of the model structure
- f begins to decreases as the model picks up the relevant features
- as p increases, the model tends to *over fit* the data
- in practice, we look for the "knee" in the curve (around $p = 5$)

Model selection scheme

- simple approach: enumerate a number of different models and to compare the resulting models
- what to compare ? how well the model is capable of reproducing these data
- how to compare ? comparing models on fresh data set: cross-validation
- model selection criterions

Information criteria

criteria for evaluating statistical models

- Kullback-Leibler (KL) information
- Akaike information criteria (AIC)
- Bayesian information criteria (BIC)
- others $e.g.$, GIC, FPE, consistent AIC

two models are **nested** if one is a special case of the other

- AR model of order 2 is nested with AR model of order 3

otherwise, they are called **nonnested**

Kullback-Leibler divergence

assumption:

- let $y_1, \ldots, y_N$ are drawn from a true density $g(x)$
- let $f(x)$ be the density of our specified model

Akaike proposed that a goodness of model should be assesed in terms of

the distance (the closeness) of $f(x)$ to the true density $g(x)$

KL divergence is defined as the expectation of $\log(g/f)$ w.r.t to g

$$I(g; f) = \mathbf{E}\left[\log\left(\frac{g(x)}{f(x)}\right)\right] = \int_{-\infty}^{\infty} \log\left(\frac{g(x)}{f(x)}\right) g(x)dx$$

with properties: (i) $I(g; f) \geq 0$ and (ii) $I(g; f) = 0 \Leftrightarrow g(x) = f(x)$

example: let $g(x)$ and $f(x)$ be n-dimensional Gaussian with parameters

$$(\mu_1, \Sigma_1) \quad \text{and} \quad (\mu_2, \Sigma_2) \quad \text{respectively}$$

it can be shown that KL divergence is given by

$$I(g; f) = (1/2) \left\{ \log \left(\frac{|\Sigma_2|}{|\Sigma_1|} \right) + \mathbf{tr}(\Sigma_2^{-1} \Sigma_1) + (\mu_1 - \mu_2)^T \Sigma_2^{-1} (\mu_1 - \mu_2) - n \right\}$$

Parsimony Principle

idea: among competing models which all explain the data well, the model with the smallest number of parameters should be chosen

In the previous example on page 9-13, how to determine model order p ?

- a trade-off curve between the loss function and the model order
- model selection criterions

a model selction criterion consists of two parts:

$$\text{Loss function} + \text{Model complexity}$$

- the first term is to assess the quality of the model, e.g., quadratic loss, likelihood function
- the second term is to penalize the model order and grows as the number of parameters increases

Model selection criteria

Akaike Information Criterion (AIC)

$$\text{AIC} = -2\mathcal{L} + 2n$$

Bayesian Information Criterion (BIC)

$$\text{BIC} = -2\mathcal{L} + n \log N$$

- $\mathcal{L}$ is the (maximized) loglikelihood function
- n is the number of effective parameters
- N is the number of sample size

some known properties:

- BIC tends to penalize complex models more heavily (due to the term $\log N$)
- BIC is asymptotically consistent

  (the probability that BIC will select the correct model approaches one as the sample size $N \to \infty$)
- AIC tends to choose models which are too complex as $N \to \infty$

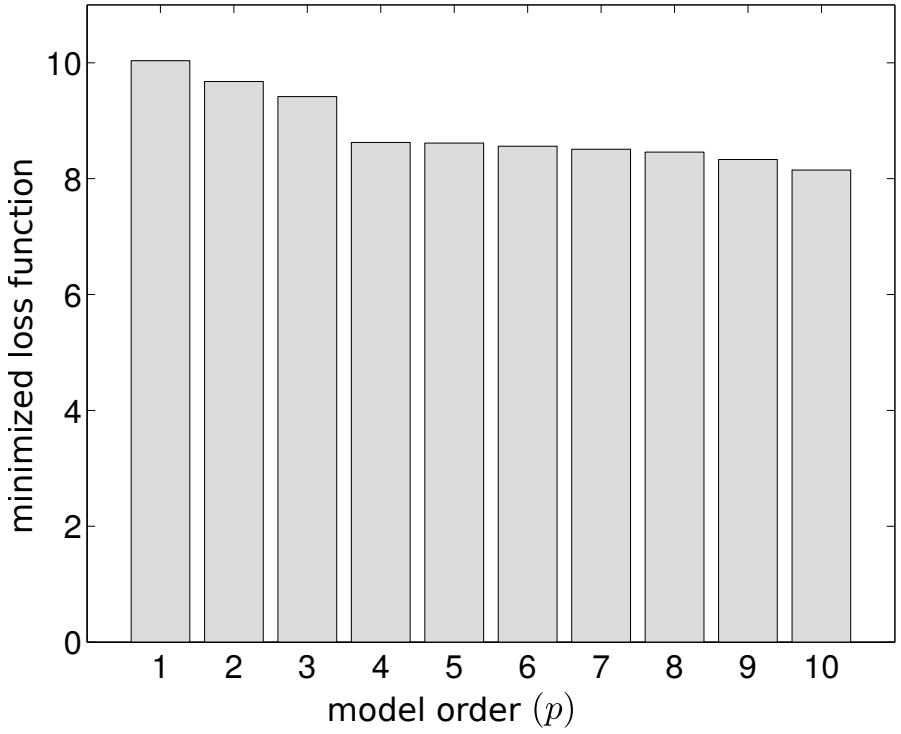
consider two nested models with n_1 and n_2 parameters and $n_2 = n_1 + m$

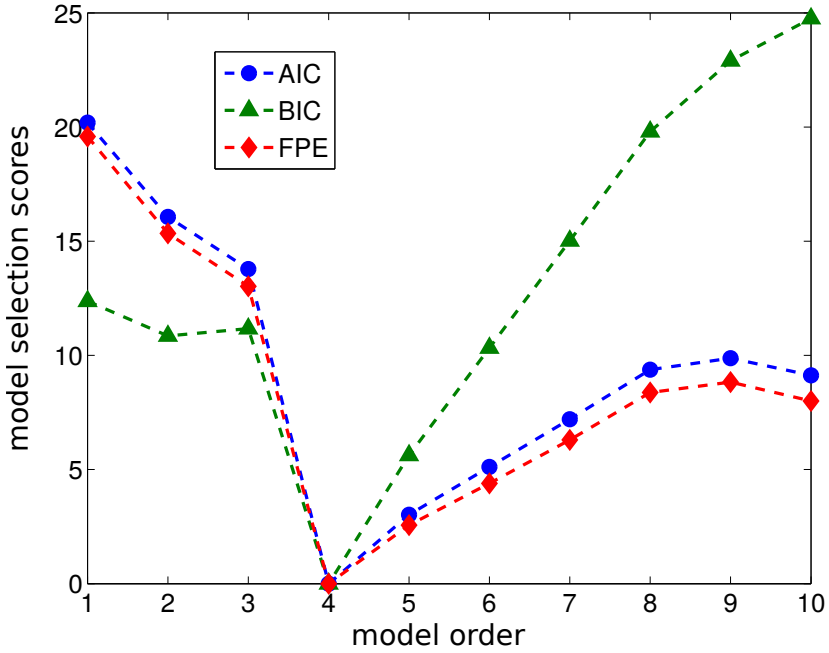
- an LR test favors the larger model if $2\mathcal{L}$ increases by $\chi^2_{0.05}(m)$ (at 5% signifiance)
- AIC favors the larger model if $2\mathcal{L}$ increases by more than $2m$, which is a lesser penalty for model size than the LR test if $m < 7$

- for example, for one restriction $(m = 1)$, the LR test uses the critical value of 3.84 whereas AIC uses a lower value of 2

- BIC favors the larger model if $2\mathcal{L}$ increases by $m \log N$, a much larger penalty than either AIC or LR test of size 0.05

Example





- the true system is AR model of order 4 with white noise of variance 1
- generate data of 100 points and estimate θ using LS

Goodness-of-fit measures

for linear models with n regressors

- **standard error of the regression**

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

  which is the estimated standard deviation of the error term

- **mean absolute error**

$$\frac{1}{N-n}\sum_{i}|y_i - \hat{y}_i|$$

- R**-squared** denoted by R^2

R^2 measure

R^2 is based on the decomposition of the total sum of squares (TSS)

$$\underbrace{\sum_i (y_i - \bar{y})^2}_{\text{TSS}} = \underbrace{\sum_i (y_i - \hat{y}_i)^2}_{\text{RSS}} + \underbrace{\sum_i (\hat{y}_i - \bar{y})^2}_{\text{ESS}} + 2\sum_i (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$$

- TSS: total sum of squares
- RSS: residual sum of squares
- ESS: explained sum of squares

for OLS, the last term on RHS is zero if the model has a constant term, so

$$\text{TSS} = \text{RSS} + \text{ESS}$$

R^2 is defined as

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

- the proportion of the total variation in Y that can be linearly predicted by X
- R^2 is termed as **coefficient of determination**
- it is not adjusted for the degree of freedom
- for linear model, $0 \leq R^2 \leq 1$; if $R^2 \approx 1$, the model fits the data well

for nonlinear model, $\sum_i (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) \neq 0$, two possible measures are

$$R^2_{\text{RES}} = 1 - \frac{\text{RSS}}{\text{TSS}}, \quad R^2_{\text{EXP}} = \frac{\text{ESS}}{\text{TSS}} \qquad \text{and} \quad R^2_{\text{RES}} \neq R^2_{\text{EXP}}$$

it is possible that $R^2_{\text{RES}} < 0$ and $R^2_{\text{EXP}} > 1$

this extension is called **Pseudo-R^2**

Adjusted R^2

R^2 is biased because it is always increased by using larger models

the bias in R^2 can be reduced by using the following adjustment

$$\text{adjusted } R^2 = 1 - \frac{\text{RSS}/(N-n-1)}{\text{TSS}/(N-1)}$$

- the presence of n penalizes the criterion for the number of predictor variables
- so adjusted R^2 can either increase or decrease when using larger models
- adjusted R^2 increases if the added predictor variables decrease RSS enough to compensate for the increase in n

Example

fitting a polynomial of order $n = 0 : 7$ to the data (true order is 2)

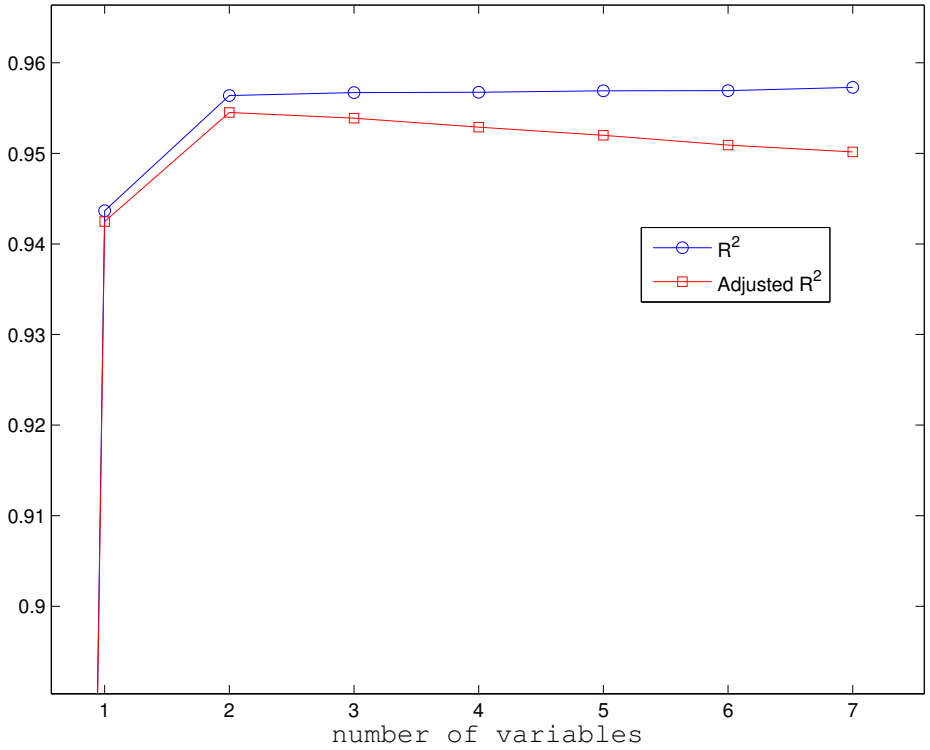
the true value is $\beta = (-2, 2, 1)$

```
beta_hat =

  3.99  -3.47  -2.40  -2.59  -2.54  -2.68  -2.62  -2.43
     0   4.98   2.78   3.60   3.20   4.85   3.77  -0.96
     0      0   0.73   0.04   0.65  -3.34   0.46  23.56
     0      0      0   0.15  -0.16   3.43  -1.76 -45.99
     0      0      0      0   0.05  -1.30   1.97  43.18
     0      0      0      0      0   0.18  -0.78 -20.71
     0      0      0      0      0      0   0.10   4.92
     0      0      0      0      0      0      0  -0.45
```

$\hat{\beta}$ explained in each column corresponds to each order

R^2 and adjusted R^2 plot VS polynomial order



- R^2 is always increasing and goes to 1 for larger models
- adjusted R^2 is slightly decreasing when more variables is added to the model

References

Chapter 8 in

A.C. Cameron and P.K. Trivedi, *Microeconometircs: Methods and Applications*, Cambridge, 2005

S. Konishi and G. Kitagawa, *Information Criteria and Statistical Modeling*, Springer, 2008

Chapter 9 in

D. Ruppert and D.S. Matteson, *Statistical and Data Analysis for Financial Engineering with R examles*, Springer, 2015