Advanced Time Series and Forecasting Lecture 1 Forecasting

Bruce E. Hansen

THE UNIVERSITY of WISCONSIN MADISON

Summer School in Economics and Econometrics
University of Crete
July 23-27, 2012

5-Day Course

- Monday: Univariate 1-step Point Forecasting, Forecast Selection
- Tuesday: Nowcasting, Combination Forecasts, Variance Forecasts
- Wednesday: Interval Forecasting, Multi-Step Forecasting, Fan Charts
- Thursday: Density Forecasts, Threshold Models, Nonparametric Forecasting
- Friday: Structural Breaks

Each Day

- Lectures: Methods with Illustrations
- Practical Sessions:
  - An empirical assignment
  - You will be given a standard dataset
  - Asked to estimate models, select and combine estimates
  - Make forecasts, forecast intervals, fan charts
  - Write your own programs

Course Website

- `www.ssc.wisc.edu/~bhansen/crete`
- Slides for all lectures
- Data for the lectures and practical sessions
- Assignments
- R code for the many of the lectures

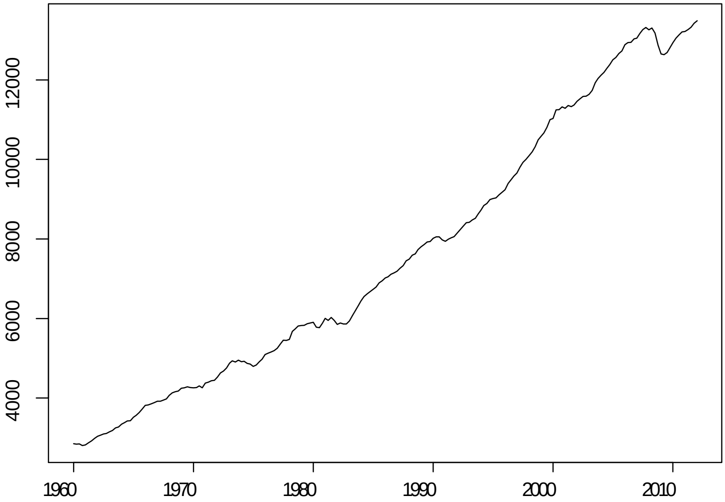
Today's Schedule

- What is Forecasting?
- Point Forecasting
- Linear Forecasting Models
- Estimation and Distribution Theory
- Forecast Selection: BIC, AIC, AIC^c, Mallows, Robust Mallows, FPE, Cross-Validation, PLS, LASSO
- Leading Indicators

Example 1

- U.S. Quarterly Real GDP
  - 1960:1-2012:1

Figure: U.S. Real Quarterly GDP



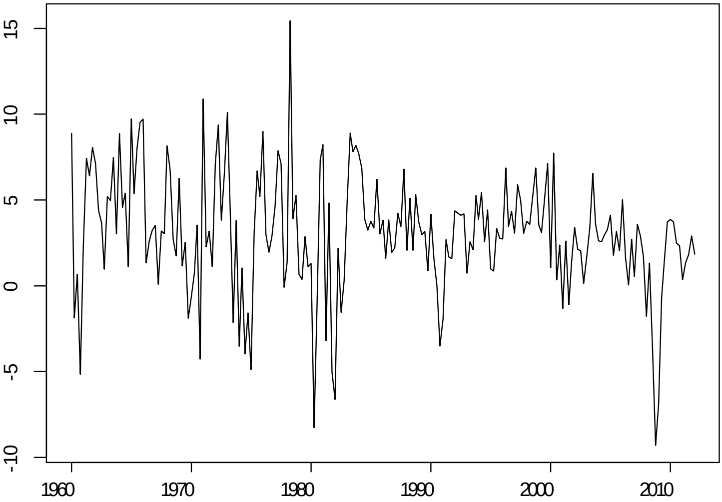
Transformations

- It is mathematically equivalent to forecast y_{n+h} or any monotonic transformation of y_{n+h} and lagged values.
  - It is equivalent to forecast the level of GDP, its logarithm, or percentage growth rate
  - Given a forecast of one, we can construct the forecast of the other.
- Statistically, it is best to forecast a transformation which is close to iid
  - For output and prices, this typically means forecasting growth rates
  - For rates, typically means forecasting changes

Annualized Growth Rate

$$y_t = 400(\log(Y_t) - \log(Y_{t-1}))$$

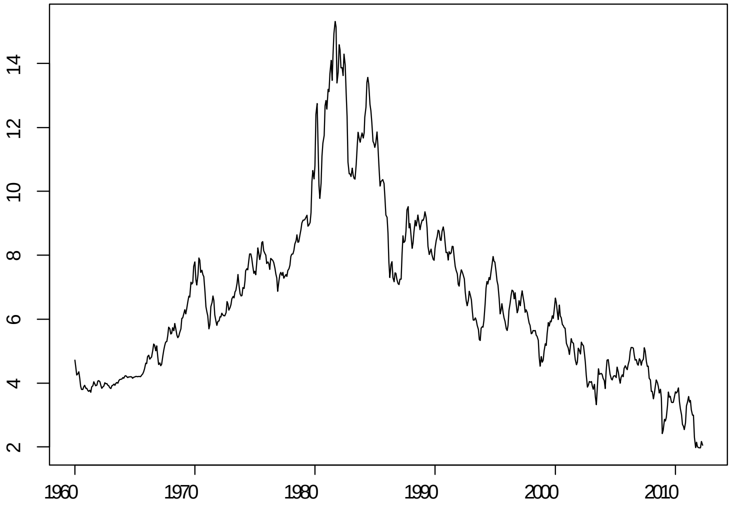
Figure: U.S. Real GDP Quarterly Growth



Example 2

- U.S. Monthly 10-Year Treasury Bill Rate
  - 1960:1-2012:4

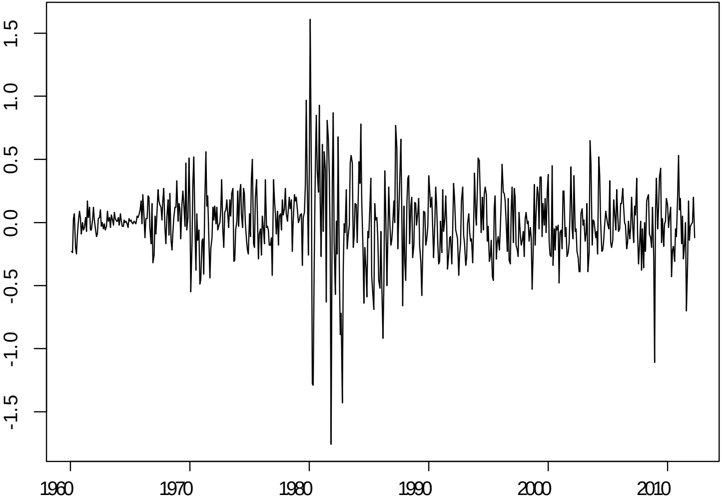
Figure: U.S. 10-Year Treasury Rate



Monthly Change

$$y_t = Y_t - Y_{t-1}$$

Figure: U.S. 10-Year Treasury Rate Change



Notation

- y_t : time series to forecast
- n : last observation
- $n + h$: time period to forecast
- h : forecast horizon
  - We often want to forecast at long, and multiple, horizons
  - For the first days we focus on one-step $(h = 1)$ forecasts, as they are the simplest
- I_n : Information available at time n to forecast y_{n+h}
  - Univariate: $I_n = (y_n, y_{n-1}, ...)$
  - Multivariate: $I_n = (x_n, x_{n-1}, ...)$ where x_t includes y_t, "leading indicators", covariates, dummy indicators

Forecast Distribution

- When we say we want to forecast y_{n+h} given I_n,
  - We mean that y_{n+h} is uncertain.
  - y_{n+h} has a (conditional) distribution
  - $y_{n+h} \mid I_n \sim F(y_{n+h}|I_n)$
- A complete forecast of y_{n+h} is the conditional distribution $F(y_{n+h}|I_n)$ or density $f(y_{n+h}|I_n)$
- $F(y_{n+h}|I_n)$ contains all information about the unknown y_{n+h}
- Since $F(y_{n+h}|I_n)$ is complicated (a distribution) we typically report low dimensional summaries, and these are typically called forecasts

Standard Forecast Objects

- Point Forecast
- Variance Forecast
- Interval Forecast
- Density forecast
- Fan Chart
- All of these forecast objects are features of the conditional distribution
- Today, we focus on point forecasts

Point Forecasts

- $f_{n+h|h}$, the most common forecast object
- "Best guess" for y_{n+h} given the distribution $F(y_{n+h}|I_n)$
- We can measure its accuracy by a loss function, typically squared error

$$\ell(f, y) = (y - f)^2$$

- The risk is the expected loss

$$E_n \ell(f, y_{n+h}) = E\left((y_{n+h} - f)^2 | I_n\right)$$

- The "best" point forecast is the one with the smallest risk

$$\begin{aligned} f &= \operatorname*{argmin}_f E\left((y_{n+h} - f)^2 | I_n\right) \\ &= E(y_{n+h}|I_n) \end{aligned}$$

- Thus the optimal point forecast is the true conditional expectation
- Point forecasts are estimates of the conditional expectation

Estimation

- The conditional distribution $F(y_{n+h}|I_n)$ and ideal point forecast $E(y_{n+h}|I_n)$ are unknown
- They need to be estimated from data and economic models
- Estimation involves
  - Approximating $E(y_{n+h}|I_n)$ with a parametric family
  - Selecting a model within this parametric family
  - Selecting a sample period (window width)
  - Estimating the parameters
- The goal of the above steps is not to uncover the "true" $E(y_{n+h}|I_n)$, but to construct a good approximation.

Information Set

- What variables are in the information set I_n?
- All past lags
  - $I_n = (x_n, x_{n-1}, ...)$
- What is x_t?
  - Own lags, “leading indicators”, covariates, dummy indicators

Markov Approximation

- $E\left(y_{n+1} | I_n\right) = E\left(y_{n+1} | x_n, x_{n-1}, \ldots\right)$
  - Depends on infinite past
- We typically approximate the dependence on the infinite past with a Markov (finite memory) approximation
- For some p,

$$E\left(y_{n+1} | x_n, x_{n-1}, \ldots\right) \approx E\left(y_{n+1} | x_n, \ldots, x_{n-p}\right)$$

- This should not be interpreted as true, but rather as an approximation.

Linear Approximation

- While the true $E\left(y_{n+1}|x_n, ..., x_{n-p}\right)$ is probably a nasty non-linear function, we typically approximate it by a linear function

$$\begin{aligned} E\left(y_{n+1}|x_n, ..., x_{n-p}\right) &\approx \beta_0 + \beta_1' x_n + \cdots + \beta_p' x_{n-p} \\ &= \boldsymbol{\beta}' \mathbf{x}_n \end{aligned}$$

- Again, this should not be interpreted as true, but rather as an approximation.
- The error is **defined** as the difference between y_{n+h} and the linear function

$$e_{t+1} = y_{t+1} - \boldsymbol{\beta}' \mathbf{x}_t$$

Linear Forecasting Model

- We now have the linear point forecasting model

$$y_{t+1} = \boldsymbol{\beta}'\mathbf{x}_t + e_{t+h}$$

- As this is an approximation, the coefficient and eror are defined by projection

$$\begin{aligned}
\boldsymbol{\beta} &= \left(E\left(\mathbf{x}_t\mathbf{x}_t'\right)\right)^{-1}\left(E\left(\mathbf{x}_t y_{t+1}\right)\right) \\
e_{t+1} &= y_{t+1} - \boldsymbol{\beta}'\mathbf{x}_t \\
E\left(\mathbf{x}_t e_{t+1}\right) &= 0 \\
\sigma^2 &= E\left(e_{t+1}^2\right)
\end{aligned}$$

Properties of the Error

- $E\left(\mathbf{x}_t e_{t+1}\right) = 0$
  - Projection
- $E\left(e_{t+1}\right) = 0$
  - Inclusion of an intercept
- If $x_t = \left(y_t, y_{t-1}, ..., y_{t-k+1}\right)$
  - $E\left(y_{t-j} e_{t+1}\right) = 0$, for $j = 0, ..., k-1$
  - $E\left(y_{t-j} e_{t+1}\right) \neq 0$ possible for $j \geq k$
- $\sigma^2 = E\left(e_{t+1}^2\right)$
  - This is the unconditional variance
  - The conditional variance $\sigma_t^2 = E\left(e_{t+1}^2 | I_t\right)$ may be time-varying

Univariate (Autoregressive) Model

- $x_t = (y_t, y_{t-1}, \ldots, y_{t-k+1})$
- A linear forecasting model is

$$y_{t+1} = \beta_0 + \beta_1 y_t + \beta_2 y_{t-1} + \cdots + \beta_k y_{t-k+1} + e_{t+1}$$

- AR(k) – Autoregression of order k
  - Typical AR(k) **models** add a stronger **assumption** about the error e_{t+1}
    - IID (independent)
    - MDS (unpredictable)
    - white noise (linearly unpredicatable/uncorrelated)
  - These assumptions are convenient for analytic purpose (calculations, simulations)
  - But they are unlikely to be true
    - Making an assumption does not make the assumption **true**
    - Do not confuse assumptions with truth

Least Squares Estimation

$$
\begin{aligned}
\widehat{\beta} &= \left(\sum_{t=0}^{n-1} \mathbf{x}_t \mathbf{x}_t'\right)^{-1} \left(\sum_{t=0}^{n-1} \mathbf{x}_t y_{t+1}\right) \\
\widehat{y}_{n+1|n} &= \widehat{f}_{n+1|n} = \widehat{\beta}' \mathbf{x}_n
\end{aligned}
$$

Distribution Theory - Consistent Estimation

- If $(y_t, \mathbf{x}_t)$ are weakly dependent (stationary and mixing, not trended nor unit roots) then
  - Sample means satisfy the WLLN

$$\frac{1}{n}\sum_{t=0}^{n-1} \mathbf{x}_t \mathbf{x}_t' \xrightarrow{p} Q = E\left(\mathbf{x}_t \mathbf{x}_t'\right)$$

$$\frac{1}{n}\sum_{t=0}^{n-1} \mathbf{x}_t y_{t+1} \xrightarrow{p} E\left(\mathbf{x}_t y_{t+1}\right)$$

  - Thus by the continuous mapping theory

$$\widehat{\beta} = \left(\sum_{t=0}^{n-1} \mathbf{x}_t \mathbf{x}_t'\right)^{-1} \left(\sum_{t=0}^{n-1} \mathbf{x}_t y_{t+1}\right)$$

$$\begin{aligned} &\xrightarrow{p} \left(E\mathbf{x}_t \mathbf{x}_t'\right)^{-1} \left(E\mathbf{x}_t y_{t+1}\right) \\ = \quad & \beta \end{aligned}$$

Distribution Theory - Asymptotic Normality

- If $(y_t, \mathbf{x}_t)$ are weakly dependent (stationary and mixing) then:
  - Mean-zero random variables satisfy the CLT.
    If $\mathbf{u}_t = g(y_{t+1}, \mathbf{x}_t)$ and $E(\mathbf{u}_t) = 0$, then

    $$\frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} \mathbf{u}_t \xrightarrow{d} N(0, \Omega)$$

    where

    $$\Omega = E(\mathbf{u}_t \mathbf{u}_t') + \sum_{j=1}^{\infty} \left(\mathbf{u}_t \mathbf{u}_{t+j}' + \mathbf{u}_{t+j} \mathbf{u}_t' \right)$$

    is the long-run (HAC) covariance matrix
  - If $\mathbf{u}_t$ is serially uncorrelated, then $\Omega = E(\mathbf{u}_t \mathbf{u}_t')$
  - This occurs when $\mathbf{u}_t$ is a martingale difference sequence
    $E(\mathbf{u}_t | I_{t-1}) = 0$

- Set $\mathbf{u}_t = \mathbf{x}_t e_{t+1}$, which satisifes $E\left(\mathbf{x}_t e_{t+1}\right) = 0$. Thus

$$\frac{1}{n}\sum_{t=0}^{n-1} \mathbf{x}_t e_{t+1} \xrightarrow{d} N(0, \Omega)$$

$$\Omega = E\left(\mathbf{x}_t \mathbf{x}_t' e_{t+1}^2\right) + \sum_{j=1}^{\infty} \left(\mathbf{x}_t \mathbf{x}_{t+j}' e_{t+1} e_{t+1+j} + \mathbf{x}_{t+j} \mathbf{x}_t' e_{t+1} e_{t+1+j}\right)$$

- Simplifies to $\Omega = E\left(\mathbf{x}_t \mathbf{x}_t' e_{t+1}^2\right)$ when $\mathbf{x}_t e_{t+1}$ serially uncorrelated
  - A sufficient condition is that e_{t+1} is a MDS
    - When the linear forecasting model is the true conditional expectation
    - Otherwise, e_{t+1} is not a MDS
  - If the forecasting model is a good approximation, then
    - e_{t+1} will be close to a MDS
    - $\mathbf{x}_t e_{t+1}$ will be close to uncorrelated
    - $\Omega \approx E\left(\mathbf{x}_t \mathbf{x}_t' e_{t+1}^2\right)$
  - However, this is best thought of as an approximation, not the truth.

Homoskedasticity

- $\sigma_t^2 = E\left(e_{t+1}^2 | I_t\right) = \sigma^2$ is a constant
- $\Omega = E\left(\mathbf{x}_t \mathbf{x}_t' e_{t+1}^2\right)$ simplifies to $\Omega = E\left(\mathbf{x}_t \mathbf{x}_t'\right) E\left(e_{t+1}^2\right)$
- Common assumption in introductory textbooks
- Empirically unsound
- Unnecessary for empirical practice
- Avoid!

Distribution Theory

- $\sqrt{n}\left(\widehat{\beta}-\beta\right) \xrightarrow{d} N(0, V)$
- $V = Q^{-1}\Omega Q^{-1}$
- $\Omega \approx E\left(\mathbf{x}_t \mathbf{x}_t' e_{t+1}^2\right)$
- "Sandwich" variance matrix

Least-Squares Residuals

- $\widehat{e}_{t+1} = y_{t+1} - \widehat{\boldsymbol{\beta}}' \mathbf{x}_t$
- Easy to compute
- Overfit (tend to be too small) when model dimension is large relative to sample size

Leave One-Out Residuals

- $\widetilde{e}_{t+1} = y_{t+1} - \widehat{\boldsymbol{\beta}}'_{-t}\mathbf{x}_t$
- $\widehat{\boldsymbol{\beta}}_{-t} = \left(\sum_{j \neq t} \mathbf{x}_j \mathbf{x}_j'\right)^{-1} \left(\sum_{j \neq t} \mathbf{x}_j y_{j+1}\right)$
- No tendency to overfit
- Easy to compute:
  - $\widetilde{e}_{t+1} = \dfrac{\widehat{e}_{t+1}}{1 - h_{tt}}$ where $h_{tt} = \mathbf{x}_t' \left(X'X\right)^{-1} \mathbf{x}_t$
  - Not necessary to actually compute n regressions!

Computation n R

Regressor Matrix: x

- `xxi=solve(t(x)%*%x)`
- `h=rowSums((x%*%xxi)*x)`

Commands

- `t(x)`= trace of x
- `%*%` = matrix multiplication
- `solve(a)`= inverse of matrix a
- `rowSums` = sum across column by row

Sequential Prediction Residuals

- $\overline{e}_{t+1} = y_{t+1} - \widehat{\boldsymbol{\beta}}_t' \mathbf{x}_t$
- $\widehat{\boldsymbol{\beta}}_t = \left(\sum_{j=0}^{t-1} \mathbf{x}_j \mathbf{x}_j' \right)^{-1} \left(\sum_{j=0}^{t-1} \mathbf{x}_j y_{j+1} \right)$
- Commonly used for pseudo out-of-sample forecast evaluation
- However, $\widehat{\boldsymbol{\beta}}_t$ is highly variable for small t (small initial sample sizes)
- Can be noisy

Variance Estimator/Standard Errors

- Asymptotic variance (White) estimator with leave-one-out residuals

$$\begin{aligned}
\widehat{V} &= \widehat{Q}^{-1}\widehat{\Omega}\widehat{Q}^{-1} \\
\widehat{Q} &= \frac{1}{n}\sum_{t=0}^{n-1} \mathbf{x}_t \mathbf{x}_t' \\
\widehat{\Omega} &= \frac{1}{n}\sum_{t=0}^{n-1} \mathbf{x}_t \mathbf{x}_t' \widetilde{e}_{t+1}^2
\end{aligned}$$

- Can use least-squares resiudals $\widehat{e}_{t+1}$ instead of leave-one-out residuals, but then multiply $\widehat{V}$ by $n/(n-\dim(\mathbf{x}_t))$.
- Standard errors for $\widehat{\boldsymbol{\beta}}$ are the square roots of the diagonal elements of $n^{-1}\widehat{V}$
- Report standard errors to interpret precision of coefficient estimates.

GDP Example

- $y_t = \Delta \log(GDP_t)$, quarterly
- AR(4) (reasonable benchmark for quarterly data)

$$y_{t+1} = \beta_0 + \beta_1 y_t + \beta_2 y_{t-1} + \beta_3 y_{t-2} + \beta_4 y_{t-3} + e_{t+1}$$

	$\widehat{\beta}$	$s(\widehat{\beta})$
Intercept	1.54	(0.45)
$\Delta \log(GDP_t)$	0.29	(0.09)
$\Delta \log(GDP_{t-1})$	0.18	(0.10)
$\Delta \log(GDP_{t-2})$	-0.05	(0.08)
$\Delta \log(GDP_{t-3})$	0.06	(0.10)

Point Forecast - GDP Growth

- AR(4)

	Actual	Forecast
2011:1	0.36	
2011:2	1.33	
2011:3	1.80	
2011:4	2.91	
2012:1	1.84	
2012:2		2.59

Interest Rate Example

- $y_t = \Delta Rate_t$
- AR(12) (reasonable benchmark for monthly data)

	$\widehat{\beta}$	$s(\widehat{\beta})$
Intercept	-0.002	(0.01)
$\Delta Rate_t$	0.40	(0.06)
$\Delta Rate_{t-1}$	-0.26	(0.07)
$\Delta Rate_{t-2}$	0.11	(0.06)
$\Delta Rate_{t-3}$	-0.07	(0.07)
$\Delta Rate_{t-4}$	0.10	(0.07)
$\Delta Rate_{t-5}$	-0.08	(0.07)
$\Delta Rate_{t-6}$	-0.05	(0.06)
$\Delta Rate_{t-7}$	-0.09	(0.06)
$\Delta Rate_{t-8}$	-0.01	(0.07)
$\Delta Rate_{t-9}$	0.03	(0.07)
$\Delta Rate_{t-10}$	0.09	(0.07)
$\Delta Rate_{t-11}$	-0.08	(0.06)

Point Forecast - 10-year Treasury Rate

- AR(12)

	Actual		Forecast	
	Level	Change	Level	Change
2012:1	1.97	-0.01		
2012:2	1.97	0.00		
2012:3	2.17	0.20		
2012:4	2.05	-0.12		
2012:5			1.93	-0.12

Forecast Selection

- We used (arbitrarily) an AR(4) for GDP, and an AR(12) for the 10-year rate
- The forecasts will be sensitive to this choice
- GDP Example

Model	Forecast
AR(0)	2.99
AR(1)	2.59
AR(2)	2.65
AR(3)	2.68
AR(4)	2.59
AR(5)	2.83
AR(6)	2.83
AR(7)	2.83
AR(8)	2.78
AR(9)	2.87
AR(10)	2.87
AR(11)	2.91
AR(12)	2.45

Forecast Selection - Big Picture

- What is the goal?
  - Accurate Forecasts
    - Low Risk (low MSFE)
- Finding the "true" model is irrelevant
  - The true model may be an AR(∞) or have a very large number of non-zero coefficients

Testing

- It is common to use statistical tests to select empirical models
- This is inappropriate
  - Tests answer the scientific question: Is there sufficient evidence to reject the hypothesis that this coefficient is zero?
  - Tests are not designed to answer the question: Which estimate yields the better forecast?
- This is not a minor issue
  - Lengthy statistics literature documenting the poor properties of "post selection" estimators.
  - Estimators based on testing have particularly bad properties
- Tests are appropriate for answering scientific questions about parameters
- Standard errors are appropriate for measuring estimation precision
- For model selection, we want something different

Model Selection: Framework

- Set of estimates (models)
  - $\widehat{\beta}(m)$, $m = 1, ..., M$
- Corresponding forecasts $\widehat{f}_{n+1|n}(m)$
- There is some population criterion $C(m)$ which evaluates the accuracy of $\widehat{f}_{n+1|n}(m)$
  - $m_0 = \operatorname{argmin}_m C(m)$ is infeasible best estimator
- There is a sample estimate $\widehat{C}(m)$ of $C(m)$
- $\widehat{m} = \operatorname{argmin}_m \widehat{C}(m)$ is empirical analog of m_0
- $\widehat{\beta}(\widehat{m})$ is selected estimator
- $\widehat{f}_{n+1|n}(\widehat{m})$ selected forecast

Selection Criterion

- Bayesian Information Criterion (BIC)
  - $C(m) = P\,(m \text{ is true})$
- Akaike Information Criterion (AIC), Corrected AIC (AIC$_c$)
  - $C(m) = KLIC$
- Mallows, Predictive Least Squares, Final Prediction Error, Leave-one-out Cross Validation:
  - $C(m) = MSFE$
- LASSO
  - Penalized LS

Important: Sample must be constant when comparing models

- This requires careful treatment of samples
- Suppose you observe y_t, $t = 1, ..., n$
- Estimation of an AR(k) requires k initial conditions, so the effective sample is for obserations $t = 1 + k, ..., n$
- The sample varies with k, sample size is $n - k$
- For valid comparison of AR(k) models for $k = 1, ..., K$
  - Fix sample with observations $t = 1 + K, ..., n$
  - $n - K$ observations
  - Estimate all AR(k) models using this same $n - K$ observations

Bayesian Information Criterion

- M models, equal prior probability that each is the "true" model
- Compute posterior probability that model m is true, given data
- Schwarz showed that in the normal linear regression model the posterior probability is proportional to

$$p(m) \propto \exp\left(-\frac{BIC(m)}{2}\right)$$

$$BIC(m) = n \log \widehat{\sigma}^2(m) + \log(n) k(m)$$

where

  - $k(m) = \#$ of parameters
  - $\widehat{\sigma}^2(m) = n^{-1} \sum_{t=0}^{n-1} \widehat{e}_{t+1}^2(m) =$ MLE estimate of σ^2 in model m

- The model with highest probability maximizes $p(m)$, or equivalently minimizes $BIC(m)$

Bayesian Information Criterion - Properties

- Consistent
  - If true model is finite dimensional, BIC will identify it asymptotically
- Conservative
  - Tends to pick small models
- Inefficient in nonparametric settings
  - If there is no true finite-dimensional model, BIC is sub-optimal
  - It does not select a finite-sample optimal model
- We are not interested in "truth", rather we want good performance

Akaike Information Criterion

- Motivated to minimize KLIC distance
- The true density of $\mathbf{y} = y_1, , ..., y_n$ is $\mathbf{f}(\mathbf{y}) = \prod f(y_i)$
- A model density $\mathbf{g}(\mathbf{y}, \theta) = \prod g(y_i, \theta)$.
- The Kullback-Leibler information criterion (KLIC) is

$$KLIC(\mathbf{f}, \mathbf{g}) = \int \mathbf{f}(\mathbf{y}) \log \left(\frac{\mathbf{f}(\mathbf{y})}{\mathbf{g}(\mathbf{y}, \theta)} \right) d\mathbf{y}$$

$$\begin{aligned} &= \int \mathbf{f}(\mathbf{y}) \log \mathbf{f}(\mathbf{y}) d\mathbf{y} - \int \mathbf{f}(\mathbf{y}) \log \mathbf{g}(\mathbf{y}, \theta) d\mathbf{y} \\ &= C_f - E \log \mathbf{g}(\mathbf{y}, \theta) \end{aligned}$$

  where the constant $C_f = \int \mathbf{f}(\mathbf{y}) \log \mathbf{f}(\mathbf{y}) d\mathbf{y}$ is independent of the model g.
- $KLIC(f, g) \geq 0$, and $KLIC(f, g) = 0$ iff $g = f$. Thus a "good" approximating model g is one with a low KLIC.

Pseudo-True

- The pseudo-true value θ_0 is the maximizer of $E \log g(y, \theta)$
- Equivalently, θ_0 minimizes $KLIC(f, g(\theta))$.

Estimation

- The negative log-likelihood function is

$$\mathcal{L}(\theta) = -\log \mathbf{g}(\mathbf{y}, \theta)$$

- The (quasi) MLE is $\widehat{\theta} = \operatorname{argmin}_{\theta} \mathcal{L}(\theta)$.
- The fitted log-likelihood is $\mathcal{L}(\widehat{\theta}) = -\log \mathbf{g}(\mathbf{y}, \widehat{\theta}(\mathbf{y}))$
- Under general conditions, $\widehat{\theta} \longrightarrow_p \theta_0$
- The QMLE estimates the best-fitting density, where best is measured in terms of the KLIC.

Asymptotic Theory

$$\sqrt{n}\left(\widehat{\theta}_{QMLE}-\theta_0\right)\rightarrow_d N(0,V)$$

$$\begin{aligned} V &= Q^{-1}\Omega Q^{-1} \\ Q &= -E\frac{\partial^2}{\partial\theta\partial\theta'}\log g(y,\theta) \\ \Omega &= E\left(\frac{\partial}{\partial\theta}\log g(y,\theta)\frac{\partial}{\partial\theta}\log g(y,\theta)'\right) \end{aligned}$$

If the model is correctly specified $(g(y,\theta_0)=f(y))$, then $Q=\Omega$ (the information matrix equality).
Otherwise $Q\neq\Omega$.

KLIC of Fitted Model

The MLE $\widehat{\theta} = \widehat{\theta}(\mathbf{y})$ is a function of the data vector $\mathbf{y}$.
The fitted model at any $\tilde{\mathbf{y}}$ is $\hat{\mathbf{g}}(\tilde{\mathbf{y}}) = \mathbf{g}(\tilde{\mathbf{y}}, \widehat{\theta}(\mathbf{y}))$.
The fitted likelihood is $\mathcal{L}(\widehat{\theta}) = -\log \mathbf{g}(\mathbf{y}, \widehat{\theta}(\mathbf{y}))$ (the model evaluated at the observed data).
The KLIC of the fitted model is is

$$\begin{aligned} KLIC(\mathbf{f}, \hat{\mathbf{g}}) &= C_f - \int \mathbf{f}(\tilde{\mathbf{y}}) \log \mathbf{g}(\tilde{\mathbf{y}}, \widehat{\theta}(\mathbf{y})) d\tilde{\mathbf{y}} \\ &= C_f - E_{\tilde{\mathbf{y}}} \log \mathbf{g}(\tilde{\mathbf{y}}, \widehat{\theta}(\mathbf{y})) \end{aligned}$$

where $\tilde{\mathbf{y}}$ has density $\mathbf{f}$, independent of $\mathbf{y}$.

Expected KLIC

The expected KLIC is the expectation over the observed values $\mathbf{y}$

$$\begin{aligned} E\left(KLIC(\mathbf{f}, \hat{\mathbf{g}})\right) &= C_f - E_{\mathbf{y}} E_{\tilde{\mathbf{y}}} \log \mathbf{g}(\tilde{\mathbf{y}}, \widehat{\theta}(\mathbf{y})) \\ &= C_f - E_{\tilde{\mathbf{y}}} E_{\mathbf{y}} \log \mathbf{g}(\mathbf{y}, \widehat{\theta}(\tilde{\mathbf{y}})) \\ &= C_f + T \end{aligned}$$

where

$$T = -E \log \mathbf{g}(\mathbf{y}, \widetilde{\theta})$$

the second equality by symmetry, and the third setting $\widetilde{\theta} = \widehat{\theta}(\tilde{\mathbf{y}})$, and $\mathbf{y}$ and $\widetilde{\theta}$ are independent.

Estimating KLIC

- Ignore C_f, goal is to estimate $T = -E \log \mathbf{g}(\mathbf{y}, \tilde{\theta})$
- Second-order Taylor expansion about $\hat{\theta}$,

$$-\log \mathbf{g}(\mathbf{y}, \widetilde{\theta}) \simeq \mathcal{L}(\widehat{\theta}) + \frac{n}{2}\left(\widetilde{\theta} - \widehat{\theta}\right)' Q \left(\widetilde{\theta} - \widehat{\theta}\right)$$

- Asymptotically,

$$\sqrt{n}\left(\widetilde{\theta} - \widehat{\theta}\right) \to_d Z \sim N\left(0, 2Q^{-1}\Omega Q^{-1}\right)$$

- Take expectations

$$\begin{aligned} T &= -E \log \mathbf{g}(\mathbf{y}, \widetilde{\theta}) \\ &\simeq E\mathcal{L}(\widehat{\theta}) + \frac{1}{2}E\left(Z'QZ\right) \\ &= E\mathcal{L}(\widehat{\theta}) + \operatorname{tr}\left(Q^{-1}\Omega\right) \end{aligned}$$

- An (asymptotically) unbiased estimate of T is then

$$\widehat{T} = \mathcal{L}(\widehat{\theta}) + \operatorname{tr}\left(Q^{-1}\Omega\right)$$

AIC

- When $g(x, \theta_0) = f(x)$ (the model is correctly specified) then $Q = \Omega$
  - $\text{tr}\left(Q^{-1}\Omega\right) = k = \dim(\theta)$
  - $\widehat{T} = \mathcal{L}(\widehat{\theta}) + k$
- Akaike Information Criterion (AIC). It is typically written as $2\widehat{T}$, e.g.

$$\begin{aligned} AIC &= 2\mathcal{L}(\widehat{\theta}) + 2k \\ &= n \log \widehat{\sigma}^2(m) + 2k(m) \end{aligned}$$

  in the linear regression model
- Similar in form to BIC, but "2" replaces $\log(n)$
- Picking a model with the smallest AIC is picking the model with the smallest estimated KLIC.

TIC

Takeuchi (1976) proposed a robust AIC, and is known as the Takeuchi Information Criterion (TIC)

$$TIC = 2\mathcal{L}(\widehat{\theta}) + 2\,\mathrm{tr}\left(\hat{Q}^{-1}\hat{\Omega}\right)$$

where

$$\begin{aligned}
\hat{Q} &= -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta\partial\theta'}\log g(y_i,\widehat{\theta}) \\
\hat{\Omega} &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial}{\partial\theta}\log g(y_i,\widehat{\theta})\frac{\partial}{\partial\theta}\log g(y_i,\widehat{\theta})'\right)
\end{aligned}$$

Corrected AIC

- In the normal linear regression model, Hurvich-Tsai (1989) calculated the exact AIC

$$AIC_c(m) = AIC(m) + \frac{2k(m)\,(k(m)+1)}{n-k(m)-1}$$

- Works better in finite samples than uncorrected AIC
- It is an exact correction when the true model is a linear regression, not time series, with iid normal errors.
- In time-series or non-normal errors, it is not an exact correction.

Comments on AIC Selection

- Widely used, partially because of its simplicity
- Full justification requires correct specification
  - normal linear regression
- TIC allows misspecification, but not widely known
- Critical specification assumption: homoskedasticity
  - AIC is a biased estimate of KLIC under heteroskedasticity
- Criterion: KLIC
  - Not a natural measure of forecast accuracy.

Point Forecast and MSFE

- Given an estimate $\widehat{\boldsymbol{\beta}}(m)$ of $\boldsymbol{\beta}$, the point forecast for y_{n+1} is

$$f_{n+1|n} = \widehat{\boldsymbol{\beta}}(m)'\mathbf{x}_n$$

- The forecast error is

$$\begin{aligned} y_{n+1} - f_{n+1|n} &= \mathbf{x}_n'\boldsymbol{\beta} + e_{t+1} - \mathbf{x}_n'\widehat{\boldsymbol{\beta}}(m) \\ &= e_{n+1} - \mathbf{x}_n'\left(\widehat{\boldsymbol{\beta}}(m) - \boldsymbol{\beta}\right) \end{aligned}$$

- The mean-squared-forecast-error (MSFE) is

$$\begin{aligned} MSFE(m) &= E\left(e_{n+1} - \mathbf{x}_n'\left(\widehat{\boldsymbol{\beta}}(m) - \boldsymbol{\beta}\right)\right)^2 \\ &\simeq \sigma^2 + E\left(\left(\widehat{\boldsymbol{\beta}}(m) - \boldsymbol{\beta}\right)' Q(m)\left(\widehat{\boldsymbol{\beta}}(m) - \boldsymbol{\beta}\right)\right) \end{aligned}$$

  where $Q(m) = E\left(\mathbf{x}_n\mathbf{x}_n'\right)$.
- The approximation is an equality if $\mathbf{x}_n$ is independent of $\widehat{\boldsymbol{\beta}}(m)$
  - Ing and Wei (Annals, 2003) show that this holds asymptotically

Estimation and MSFE

- The MSFE is

$$\begin{aligned} MSFE(m) &\simeq \sigma^2 + E\left(\left(\widehat{\boldsymbol{\beta}}(m) - \boldsymbol{\beta}\right)' Q(m)\left(\widehat{\boldsymbol{\beta}}(m) - \boldsymbol{\beta}\right)\right) \\ &= \sigma^2 + MSE(\widehat{\boldsymbol{\beta}}(m)) \end{aligned}$$

  where

$$MSE(\widehat{\boldsymbol{\beta}}(m)) = \operatorname{tr} E\left(Q(m)\left(\widehat{\boldsymbol{\beta}}(m) - \boldsymbol{\beta}\right)\left(\widehat{\boldsymbol{\beta}}(m) - \boldsymbol{\beta}\right)'\right)$$

  is the weighted mean-squared-error (MSE) of $\widehat{\boldsymbol{\beta}}(m)$ for $\boldsymbol{\beta}$
- Given a model $\boldsymbol{\beta}'\mathbf{x}_t$ for the conditional mean, the choice of estimator $\widehat{\boldsymbol{\beta}}(m)$ impacts the MSFE through $MSE(\widehat{\boldsymbol{\beta}}(m))$
- The best point forecast (the one with the smallest MSFE) is obtained by using an estimator $\widehat{\boldsymbol{\beta}}(m)$ with the smallest MSE

Residual Fit

$$
\begin{aligned}
\widehat{\sigma}^2 &= \frac{1}{n}\sum_{t=0}^{n-1}\widehat{e}_{t+1}(m)^2 \\
&= \frac{1}{n}\sum_{t=0}^{n-1}e_{t+1}^2 + \frac{1}{n}\sum_{t=0}^{n-1}\left(\mathbf{x}_t'\left(\widehat{\beta}(m)-\beta\right)\right)^2 \\
&\quad -\frac{2}{n}\sum_{t=0}^{n-1}e_{t+1}\mathbf{x}_t'\left(\widehat{\beta}(m)-\beta\right)
\end{aligned}
$$

- First two terms are estimates of

$$MSFE(m) = E\left(e_{n+1} - \mathbf{x}_n'\left(\widehat{\beta}(m)-\beta\right)\right)^2$$

- Third term is

$$\sum_{t=0}^{n-1}e_{t+1}\mathbf{x}_t'\left(\widehat{\beta}(m)-\beta\right) = \mathbf{e}'\mathbf{P}(m)\mathbf{e}$$

where $\mathbf{P}(m) = \mathbf{X}(m)\left(\mathbf{X}(m)'\mathbf{X}(m)\right)^{-1}\mathbf{X}(m)'$

Residual Variance as Biased estimate of MSFE

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{t=0}^{n-1} e_{t+1}^2 + \frac{1}{n}\sum_{t=0}^{n-1}\left(\mathbf{x}_t'\left(\widehat{\boldsymbol{\beta}}(m) - \boldsymbol{\beta}\right)\right)^2 - \frac{2}{n}\mathbf{e}'\mathbf{P}(m)\mathbf{e}$$

$$\begin{aligned} E\left(\widehat{\sigma}^2\right) &= \sigma^2 + E\left(\mathbf{x}_t'\left(\widehat{\boldsymbol{\beta}}(m) - \boldsymbol{\beta}\right)\right)^2 - \frac{2}{n}E\left(\mathbf{e}'\mathbf{P}(m)\mathbf{e}\right) \\ &\simeq MSFE_n(m) - \frac{2}{n}B(m) \end{aligned}$$

where

$$B(m) = E\left(\mathbf{e}'\mathbf{P}(m)\mathbf{e}\right)$$

Relation between Residual variance and MSFE

$$\begin{aligned} \widehat{\sigma}^2 &= MSFE_n(m) - \frac{2}{n}B(m) \\ B(m) &= E\left(\mathbf{e}'\mathbf{P}(m)\mathbf{e}\right) \end{aligned}$$

- The residual variance is smaller than the MSFE by $\frac{2}{n}B(m)$
- This is a classic relationship
- It suggests that "estimates" of the MSFE need to be equivalent to

$$C_n(m) = \widehat{\sigma}^2(m) + \frac{2}{n}B(m)$$

- The residual variance plus a optimal penalty $2B(m)/n$

Asymptotic Penalty

From asymptotic theory, for any m

$$\frac{1}{n}\mathbf{X}(m)'\mathbf{X}(m) \to_p Q(m) = E\left(\mathbf{x}_t(m)\mathbf{x}_t(m)'\right)$$

$$\frac{1}{\sqrt{n}}\mathbf{X}(m)'\mathbf{e} \to_d Z(m) \sim N(0, \Omega(m))$$

$$\Omega(m) = E\left(\mathbf{x}_t(m)\mathbf{x}_t'(m)e_{t+1}^2\right)$$

Thus

$$\begin{aligned}\mathbf{e}'\mathbf{P}(m)\mathbf{e} &= \left(\frac{1}{\sqrt{n}}\mathbf{e}'\mathbf{X}(m)\right)\left(\frac{1}{n}\mathbf{X}(m)'\mathbf{X}(m)\right)^{-1}\left(\frac{1}{\sqrt{n}}\mathbf{X}(m)'\mathbf{e}\right) \\ &\to_d Z(m)'Q(m)^{-1}Z(m) \\ &= \operatorname{tr}\left(Q(m)^{-1}Z(m)Z(m)'\right)\end{aligned}$$

Asymptotic Penalty

$$
\begin{aligned}
\mathbf{e}'\mathbf{P}(m)\mathbf{e} &\rightarrow_d Z(m)'Q(m)^{-1}Z(m) \\
&= \operatorname{tr}\left(Q(m)^{-1}Z(m)Z(m)'\right)
\end{aligned}
$$

Thus

$$
\begin{aligned}
B(m) &= E\left(\mathbf{e}'\mathbf{P}\mathbf{e}\right) \\
&\longrightarrow \operatorname{tr}\left(Q(m)^{-1}E\left(Z(m)Z(m)'\right)\right) \\
&= \operatorname{tr}\left(Q(m)^{-1}\Omega(m)\right)
\end{aligned}
$$

MSFE Criterion for Least-Squares

$$C_n(m) = \widehat{\sigma}^2(m) + \frac{2}{n}\operatorname{tr}\left(Q(m)^{-1}\Omega(m)\right)$$

$$\begin{aligned} Q(m) &= E\left(\mathbf{x}_t(m)\mathbf{x}_t(m)'\right) \\ \Omega(m) &= E\left(\mathbf{x}_t(m)\mathbf{x}_t'(m)e_{t+1}^2\right) \end{aligned}$$

This is an (asymptotically) unbiased estimate of the MSFE

Homoskedastic Case

When

$$E\left(e_{t+1}^2 \mid I_t\right) = \sigma^2$$

then

$$\Omega(m) = E\left(\mathbf{x}_t(m)\mathbf{x}_t'(m)e_{t+1}^2\right) = Q(m)\sigma^2$$

$$\operatorname{tr}\left(Q(m)^{-1}\Omega(m)\right) = \sigma^2 \operatorname{tr}\left(\mathbf{I}(m)\right) = \sigma^2 k(m)$$

$$C_n(m) = \widehat{\sigma}^2(m) + \frac{2}{n}\sigma^2 k(m)$$

Under homoskedasticity, the MSFE can be estimated by the residual variance, plus a penalty which is proportional to the number of estimated parameters

Mallows Criterion

$$C_n(m) = \widehat{\sigma}^2(m) + \frac{2}{n}\sigma^2 k(m)$$

- Replace the unknown σ^2 with a preliminary estimate $\widetilde{\sigma}^2$
  - bias-corrected residual variance from a "large" model

$$\widetilde{\sigma}^2 = \frac{1}{n-K}\sum_{t=0}^{n-1}\widehat{e}_{t+1}(K)^2$$

$$C_n(m) = \widehat{\sigma}^2(m) + \frac{2}{n}\widetilde{\sigma}^2 k(m)$$

- Sometimes written as

$$C_n(m) = \sum_{t=0}^{n-1}\widehat{e}_{t+1}(m)^2 + 2\widetilde{\sigma}^2 k(m)$$

Final Prediction Error (FPE) Criterion

$$C_n(m) = \widehat{\sigma}^2(m) + \frac{2}{n}\sigma^2 k(m)$$

- Replace the unknown σ^2 with $\widehat{\sigma}^2(m)$

$$FPE_n(m) = \widehat{\sigma}^2(m)\left(1 + \frac{2}{n}k(m)\right)$$

Relations betwees Mallows, FPE, and Akaike

- Take log of FPE and multiply by n

$$\begin{aligned} n\log\left(FPE_n(m)\right) &= n\log\left(\widehat{\sigma}^2(m)\right) + n\log\left(1+\frac{2}{n}k(m)\right) \\ &\simeq n\log\left(\widehat{\sigma}^2(m)\right) + 2k(m) \\ &= AIC(m) \end{aligned}$$

- Thus Mallows, FPE and Akaike model selection is quite similar
- Mallows, FPE, and $\exp\left(AIC(m)/n\right)$ are estimates of MSFE under homoskedasticity

Robust Mallows

- Ideal Criterion

$$C_n(m) = \widehat{\sigma}^2(m) + \frac{2}{n}\operatorname{tr}\left(Q(m)^{-1}\Omega(m)\right)$$

$$\begin{aligned} Q(m) &= E\left(\mathbf{x}_t(m)\mathbf{x}_t(m)'\right) \\ \Omega(m) &= E\left(\mathbf{x}_t(m)\mathbf{x}_t'(m)e_{t+1}^2\right) \end{aligned}$$

- Sample estimate

$$C_n^*(m) = \widehat{\sigma}^2(m) + \frac{2}{n}\operatorname{tr}\left(\widehat{Q}(m)^{-1}\widehat{\Omega}(m)\right)$$

$$\begin{aligned} \widehat{Q}(m) &= \frac{1}{n}\sum_{t=0}^{n-1}\mathbf{x}_t\mathbf{x}_t' \\ \widehat{\Omega}(m) &= \frac{1}{n}\sum_{t=0}^{n-1}\mathbf{x}_t\mathbf{x}_t'\widetilde{e}_{t+1}^2 \end{aligned}$$

  where $\widetilde{e}_{t+1}$ is residual from a preliminary estimate
- Robust Mallows similar to TIC, not

Cross-Validation

- Leave-one-out estimator

$$\widehat{\beta}_{-t}(m) = \left(\sum_{j \neq t} \mathbf{x}_j(m)\mathbf{x}_j(m)' \right)^{-1} \left(\sum_{j \neq t} \mathbf{x}_j(m) y_{j+1} \right)$$

- Leave-one-out prediction residual

$$\begin{aligned} \widetilde{e}_{t+1}(m) &= y_{t+1} - \widehat{\beta}_{-t}(m)'\mathbf{x}_t(m) \\ &= \frac{\widehat{e}_{t+1}(m)}{1 - h_{tt}(m)} \end{aligned}$$

- $\widetilde{e}_{t+1}(m)$ is a forecast error based on estimation without observation t
- $E\widetilde{e}_{t+1}(m)^2 \simeq MSFE_n(m)$
- $CV_n(m) = \frac{1}{n} \sum_{t=0}^{n-1} \widetilde{e}_{t+1}(m)^2$ is an estimate of $MSFE_n(m)$
- Called the leave-one-out cross-validation (CV) criterion

CV is Similar to Robust Mallows

By a Taylor expansion, $\dfrac{1}{(1-a)^2} \simeq 1-2a$

$$
\begin{aligned}
CV_n(m) &= \frac{1}{n}\sum_{t=0}^{n-1} \widetilde{e}_{t+1}(m)^2 \\
&= \frac{1}{n}\sum_{t=0}^{n-1} \frac{\widehat{e}_{t+1}(m)^2}{(1-h_{tt}(m))^2} \\
&\simeq \frac{1}{n}\sum_{t=0}^{n-1} \widehat{e}_{t+1}(m)^2 + 2\frac{1}{n}\sum_{t=0}^{n-1} \widehat{e}_{t+1}(m)^2 h_{tt}(m) \\
&= \widehat{\sigma}^2(m) + \frac{2}{n}\sum_{t=0}^{n-1} \widehat{e}_{t+1}(m)^2 \mathbf{x}_t' \left(X'X\right)^{-1} \mathbf{x}_t \\
&= \widehat{\sigma}^2(m) + \frac{2}{n}\operatorname{tr}\left(\left(X'X\right)^{-1} \sum_{t=0}^{n-1} \widehat{e}_{t+1}(m)^2 \mathbf{x}_t \mathbf{x}_t' \right) \\
&= C_n^*(m)
\end{aligned}
$$

Comments on CV Selection

- Selecting one-step forecast models by cross-validation is computationally simple, generally valid, and robust to heteroskedasticity
- Does not require correct specification
- Similar to robust Mallows
- Similar to Mallows, AIC and FPE under homoskedasticity
- Conceptually easy to generalize beyond least-squares estimation

Predictive Least Squares (Out-of-Sample MSFE)

- Sequential estimates

$$\widehat{\beta}_t(m) = \left(\sum_{j=0}^{t-1} \mathbf{x}_j(m)\mathbf{x}_j(m)'\right)^{-1} \left(\sum_{j=0}^{t-1} \mathbf{x}_j(m) y_{j+1}\right)$$

- Sequential prediction residuals

$$\overline{e}_{t+1}(m) = y_{t+1} - \widehat{\beta}_t(m)'\mathbf{x}_t(m)$$

- Predictive Least Squares. For some P

$$PLS_n(m) = \frac{1}{P} \sum_{t=n-P}^{n-1} \overline{e}_{t+1}(m)^2$$

- Major Difficulty: PLS very sensitive to P

Comments on Predictive Least Squares

- Conceptually simple, easy to generalize beyond least-squares
  - Can be applied to actual forecasts, without need to know forecast method
- $\overline{e}_{t+1}(m)$ are fully valid prediction errors
- Possibly more robust to structural change than CV
  - Intuitive, but this claim has not been formally justified
- Very common in applied forecasting
  - Frequently asserted as "empirical performance"
- On the negative side, PLS over-estimates MSFE
  - $\overline{e}_{t+1}(m)$ is a prediction error from a sample of length $t < n$
  - PLS will tend to be overly-parsimonious
  - Very sensitive to number of pseudo out-of-sample observations P

LASSO

- L1 constrained optimization
- Least-Angle regression
- Let $\boldsymbol{\beta} = (\beta_1, ..., \beta_P)$
- $\widehat{\boldsymbol{\beta}}$ minimizes the penalized least-squares criterion

$$S(\boldsymbol{\beta}) = \sum_{t=0}^{n-1} \left(y_{t+1} - \boldsymbol{\beta}'\mathbf{x}_t\right)^2 + \lambda \sum_{j=1}^{P} \left|\beta_j\right|$$

- Many coefficient estimates $\hat{\beta}_j$ will be zero
  - LASSO is effectively a variable selection method
- Even if $P > n$, LASSO is still feasible!
- Choice of λ important

Comments on LASSO

- Theory for time-series and forecasting not well developed
- Current theory suggests LASSO appropriate for **sparse** models
  - Most coeffients are zero
  - A few, fixed, coefficients are non-zero
  - (Adaptive) LASSO can consistently select the non-zero coefficients
  - LASSO has similarities with BIC selection, but better
- A huge advantage is that LASSO allows for extremely large P, without need for ordering.

Theory of Optimal Selection

- $MSFE_n(m)$ is the MSFE from model m
- $\inf_m MSFE_n(m)$ is the (infeasible) best MSFE
- Let $\widehat{m}$ be the selected model
- Let $MSFE_n(\widehat{m})$ denote the MSFE using the selected estimator
- We say that selection is asymptotically optimal if

$$\frac{MSFE_n(\widehat{m})}{\inf_m MSFE_n(m)} \xrightarrow{p} 1$$

Theory of Optimal Selection

- A series of papers have shown that AIC, Mallows, FPE are asymptotically optimal for selection
- Assumptions
  - Autoregressions
  - Errors are iid, homoskedastic
  - True model is AR(∞)
- Shibata (Annals, 1980), Ching-Kang Ing with co-authors (2003, 2005, etc)
- Proof Method: Show that the selection criterion is uniformly close to MSFE

Theory of Optimal Selection - Regression Case

- In regression (iid date) case
- Li (1987), Andrews (1991), Hansen (2007), Hansen and Racine (2012)
- AIC, Mallows, FPE, CV are asymptotically optimal for seletion under homoskedasticity
- CV is asymptotically optimal for seletion under heteroskedasticity

Forecast Selection - Summary

- Testing inappropriate for forecast selection
- Feasible selection criteria: BIC, AIC, AIC_c, Mallows, Robust Mallows, FPE, PLS, CV, LASSO
- Valid comparisons require holding sample constant across models
- All methods except CV and PLS require conditional homoskedasticity
- PLS sensitive to choice of P
- BIC and LASSO appropriate when true structure is sparse
- CV quite general and flexible
  - Recommended method

GDP Example

Methods: BIC, AIC$_c$, Robust Mallows, CV

Model	BIC	AIC$_c$	C_n^*	CV
AR(1)	473	466	10.7	10.7
AR(2)	**472**	**462**	**10.6**	**10.5**
AR(3)	477	464	10.7	10.7
AR(4)	481	465	10.8	10.8
AR(5)	483	464	10.8	10.8
AR(6)	489	466	11.0	10.9
AR(7)	494	468	11.1	11.1
AR(8)	498	470	11.3	11.2
AR(9)	500	469	11.3	11.2
AR(10)	505	471	11.4	11.4
AR(11)	511	473	11.5	11.5
AR(12)	511	471	11.4	11.3

Methods select AR(2)

10-Year Treasury Rate

Model	BIC	AIC_c	C_n^*	CV
AR(1)	-1518	-1527	0.0798	0.0798
AR(2)	$-\mathbf{1541}^*$	-1554	$\mathbf{0.0768}^*$	$\mathbf{0.0768}^*$
AR(3)	-1538	-1555	0.0769	0.0769
AR(4)	-1532	-1554	0.0773	0.0773
AR(6)	-1531	-1561	0.0772	0.0770
AR(8)	-1522	-1562	0.0777	0.0774
AR(10)	-1513	-1561	0.0784	0.0781
AR(12)	-1506	-1563	0.079	0.0787
AR(20)	-1471	-1561	0.081	0.080
AR(22)	-1470	$-\mathbf{1570}^*$	0.081	0.080
AR(24)	-1458	-1565	0.081	0.081

Mallows, AIC_c, FPE select AR(22)

Robust Mallows, CV select AR(2)

Difference due to conditional heteroskedasticity

AR(2) through AR(6) near equivalent with respect to C_n^* and CV

Point Forecast - GDP Growth

- AR(2)

	Actual	Forecast
2011:1	0.36	
2011:2	1.33	
2011:3	1.80	
2011:4	2.91	
2012:1	1.84	
2012:2		2.65

Point Forecast - 10-year Treasury Rate

- AR(2)

	Actual		Forecast	
	Level	Change	Level	Change
2012:1	1.97	-0.01		
2012:2	1.97	0.00		
2012:3	2.17	0.20		
2012:4	2.05	-0.12		
2012:5			1.96	-0.09

Forecasting with Leading Indicators

- Recall, the ideal forecast is

$$E\left(y_{n+1} | I_n\right) = E\left(y_{n+1} | x_n, x_{n-1}, \ldots\right)$$

  where I_n contains all information
- $x_n =$ lags + leading indicators
  - Variables which help predict y_{t+1}
  - We have focused on univariate lags
  - Typically more information in related series
  - Which?

Good Leading Indicators

- Measured quickly
- Anticipatory
- Varies by forecast variable

Interest Rate Spreads

- Difference between Long and Short Rate
- Measured immediately
- Indicate monetary policy, aggregate demand
- Term Structure of Interest Rates:
- Long Rate is the market expectation of the average future short rates
- Spread is the market expectation of future short rates
- I use U.S. Treasury rates, difference between 10-year and 3-month

Figure: 10-Year and 3-Month T-Bill Rates

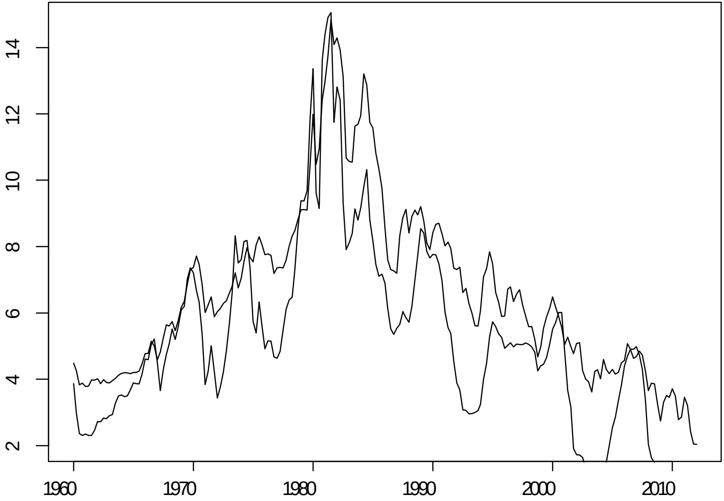
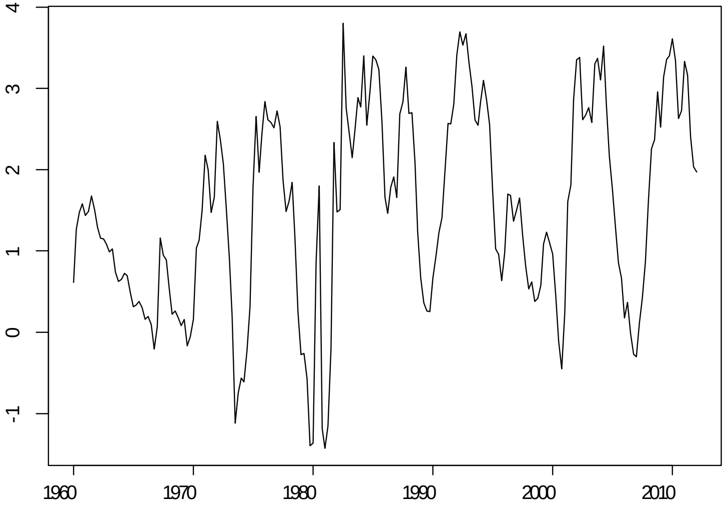


Figure: Term Spread



High Yield Spread

- "Riskless" rate: U.S. Treasury
- Low-risk rate: AAA grade corporate bond
- High Yield rate: Low grade corporate bond
- Theory: high-yield rate includes premium for probability of default
- Low grade bond rates increase with probability of default – when real activity is expected to fall
- Spread: Difference between corporate bond rates
- I use difference between AAA and BAA bond rates

Figure: AAA and BAA Corporate Bond Rates

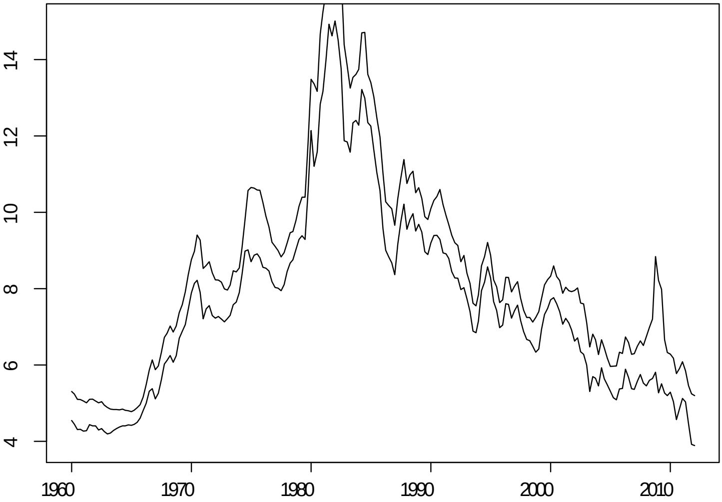
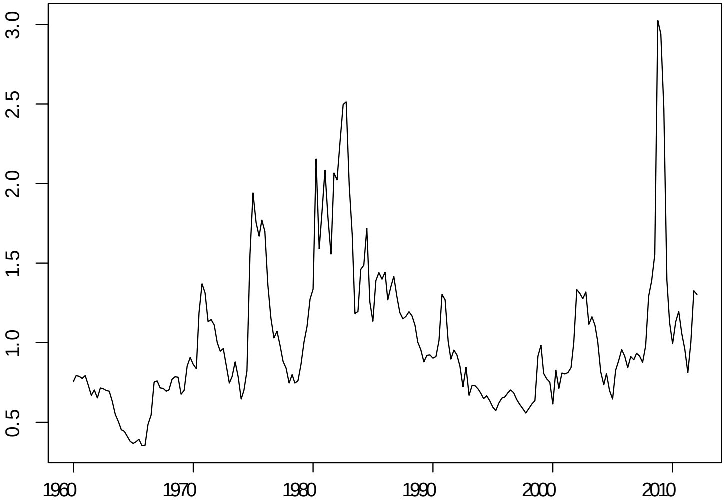


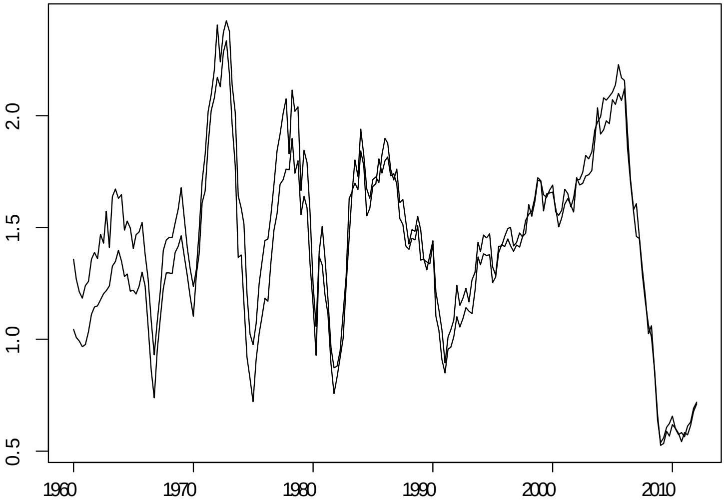
Figure: High Yield Spread



Construction Indicators

- Building Permits
- Housing Starts
- Anticipate construction spending

Figure: Housing Starts, Building Permits



Mixed Frequency Data

- U.S. GDP is measured quarterly
- Interest rates: Daily
- Permits: Monthly
- Simplest approach: Quarterly aggregation
  - Aggregate (average) daily and monthly variables to quarterly level
- Mixed Frequency approach
  - Use lower frequency data as predictors
- For now, we use aggregate (quarterly) data

Timing

- Variables reported in separate sequences
- Should we use only "quarter 1" variables to forecast "quarter 2"?
- Or should we use whatever is available?
  - E.g., use quarter 2 interest rates to forecast quarter 1 GDP?
- Let's use quarter 1 data to forecast quarter 2

Models Selection by CV

- All estimates include intercept plus two lags of GDP growth

Model	CV	Forecast
Spread	10.4	2.8
HY Spread	10.6	2.5
Housing Starts	10.3	1.4
Bulding Permits	10.3	1.7
Sp+HY	10.3	2.7
Sp+HS	10.02	1.5
Sp+BP	10.1	1.9
HY+HS	10.4	1.4
HY+BP	10.4	1.6
HS+BP	10.4	1.4
Sp+HY+HS	**10.00**	**1.3**
Sp+HY+BP	10.1	1.7
Sp+HS+BP	10.05	1.3
HY+HS+BP	10.5	1.3

Coefficient Estimates

$\Delta\log(GDP_{t+1})$	$\widehat{\beta}$	$s(\widehat{\beta})$
Intercept	-0.33	(1.03)
$\Delta\log(GDP_t)$	0.16	(0.10)
$\Delta\log(GDP_{t-1})$	0.09	(0.10)
Bond Spread$_t$	0.61	(0.23)
High Yield Spread	-1.10	(0.75)
Housing Starts$_t$	1.86	(0.65)

Alternative Specifications

- Lags of Leading Indicators
- Transformations (Changes, Growth Rates, Logs, Differences)

Practical Session

- Data Set: U.S. macro data
  - Monthly 1960:1 - 2012:4
  - Unemployment Rates
  - 10-year Treasury Rate
  - 3-month Treasury Rate
  - AAA bond rate
  - BAA bond rate
  - Housing Starts
  - Building Permits
  - Industrial Production Index
  - CPI Index (less food and energy)
- `www.ssc.wisc.edu/~bhansen/crete`

Assignment 1

- Estimate model for Unemployment Rate
  - Write your own programs!
- First model: Autoregression
  - Estimate a set of autoregressions
  - Compute model selection criteria:
    - CV
    - Optional: BIC, AIC, AIC_c, Mallows, Robust Mallows, FPE
  - Select model
  - Compute point forecast for next period
- Second model add leading indicators
  - Select and transform relevant varibles
  - Estimate a set of models, select via information criteria
  - Compute point forecast for next period

Figure: U.S. Unemployment Rate