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: $I_n = (x_n, x_{n-1}, ...) \text{ where } x_t \text{ includes } y_t, \text{ “leading indicators”, covariates, dummy indicators}$
• 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
  ▶ $y_{n+h} \mid I_n \sim F(y_{n+h} \mid I_n)$

• A complete forecast of $y_{n+h}$ is the conditional distribution $F(y_{n+h} \mid I_n)$
or density $f(y_{n+h} \mid I_n)$

• $F(y_{n+h} \mid I_n)$ contains all information about the unknown $y_{n+h}$
• Since $F(y_{n+h} \mid I_n)$ 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 $F(y_{n+h}|l_n)$
- We can measure its accuracy by a loss function, typically squared error

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

- The risk is the expected loss

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

- The “best” point forecast is the one with the smallest risk

$$f = \arg\min_f E((y_{n+h} - f)^2 | l_n)$$

$$= E(y_{n+h}|l_n)$$

- Thus the optimal point forecast is the true conditional expectation
- Point forecasts are estimates of the conditional expectation
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
  - $I_n = (x_n, x_{n-1}, \ldots)$
- 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 $E (y_{n+1} | x_n, \ldots, x_{n-p})$ is probably a nasty non-linear function, we typically approximate it by a linear function

  $$
  E (y_{n+1} | x_n, \ldots, x_{n-p}) \approx \beta_0 + \beta_1 x_n + \cdots + \beta_p x_{n-p} \\
  = \beta' x_n
  $$

- 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

  $$
  e_{t+1} = y_{t+1} - \beta' x_t
  $$
Linear Forecasting Model

- We now have the linear point forecasting model

\[ y_{t+1} = \beta' x_t + e_{t+h} \]

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

\[
\begin{align*}
\beta &= \left( E \left( x_t x'_t \right) \right)^{-1} \left( E \left( x_t y_{t+1} \right) \right) \\
e_{t+1} &= y_{t+1} - \beta' x_t \\
E \left( x_t e_{t+1} \right) &= 0 \\
\sigma^2 &= E \left( e_{t+1}^2 \right)
\end{align*}
\]
Properties of the Error

- \( E(\mathbf{x}_t e_{t+1}) = 0 \)
  - Projection

- \( E(e_{t+1}) = 0 \)
  - Inclusion of an intercept

- If \( x_t = (y_t, y_{t-1}, ..., y_{t-k+1}) \)
  - \( E(y_{t-j} e_{t+1}) = 0 \), for \( j = 0, ..., k - 1 \)
  - \( E(y_{t-j} e_{t+1}) \neq 0 \) possible for \( j \geq k \)

- \( \sigma^2 = E(e_{t+1}^2) \)
  - This is the unconditional variance
  - The conditional variance \( \sigma^2_t = E(e_{t+1}^2 | l_t) \) 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 unpredictable/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

\[ \hat{\beta} = \left( \sum_{t=0}^{n-1} x_t x_t' \right)^{-1} \left( \sum_{t=0}^{n-1} x_t y_{t+1} \right) \]

\[ \hat{y}_{n+1|n} = \hat{f}_{n+1|n} = \hat{\beta}' x_n \]
Distribution Theory - Consistent Estimation

- If \((y_t, x_t)\) are weakly dependent (stationary and mixing, not trended nor unit roots) then
  - Sample means satisfy the WLLN
    \[
    \frac{1}{n} \sum_{t=0}^{n-1} x_t x'_t \xrightarrow{p} Q = E(x_t x'_t)
    \]
    \[
    \frac{1}{n} \sum_{t=0}^{n-1} x_t y_{t+1} \xrightarrow{p} E(x_t y_{t+1})
    \]
  - Thus by the continuous mapping theory
    \[
    \hat{\beta} = \left( \sum_{t=0}^{n-1} x_t x'_t \right)^{-1} \left( \sum_{t=0}^{n-1} x_t y_{t+1} \right)
    \xrightarrow{p} (E x_t x'_t)^{-1} (E x_t y_{t+1})
    = \beta
    \]
Distribution Theory - Asymptotic Normality

- If \((y_t, x_t)\) are weakly dependent (stationary and mixing) then:
  - Mean-zero random variables satisfy the CLT.
  - If \(u_t = g(y_{t+1}, x_t)\) and \(E(u_t) = 0\), then
    \[
    \frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} u_t \xrightarrow{d} \mathcal{N}(0, \Omega)
    \]
    where
    \[
    \Omega = E(u_t u'_t) + \sum_{j=1}^{\infty} \left( u_t u'_{t+j} + u_{t+j} u'_t \right)
    \]
is the long-run (HAC) covariance matrix
  - If \(u_t\) is serially uncorrelated, then \(\Omega = E(u_t u'_t)\)
  - This occurs when \(u_t\) is a martingale difference sequence \(E(u_t | I_{t-1}) = 0\)
Set \( u_t = x_t e_{t+1} \), which satisfies \( E(x_t e_{t+1}) = 0 \). Thus

\[
\frac{1}{n} \sum_{t=0}^{n-1} x_t e_{t+1} \xrightarrow{d} N(0, \Omega)
\]

\[
\Omega = E(x_t x'_t e_{t+1}^2) + \sum_{j=1}^{\infty} (x_t x'_{t+j} e_{t+1} e_{t+1+j} + x_{t+j} x'_t e_{t+1} e_{t+1+j})
\]

Simplifies to \( \Omega = E(x_t x'_t e_{t+1}^2) \) when \( x_t e_{t+1} \) 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
  - \( x_t e_{t+1} \) will be close to uncorrelated
  - \( \Omega \approx E(x_t x'_t e_{t+1}^2) \)

- However, this is best thought of as an approximation, not the truth.
Homoskedasticity

- $\sigma_t^2 = E(e_{t+1}^2|I_t) = \sigma^2$ is a constant
- $\Omega = E(x_t x_t' e_{t+1}^2)$ simplifies to $\Omega = E(x_t x_t') E(e_{t+1}^2)$
- Common assumption in introductory textbooks
- Empirically unsound
- Unnecessary for empirical practice
- Avoid!
\[
\sqrt{n} \left( \hat{\beta} - \beta \right) \xrightarrow{d} N(0, V)
\]
\[
V = Q^{-1} \Omega Q^{-1}
\]
\[
\Omega \approx E \left( x_t x_t' e_{t+1}^2 \right)
\]
“Sandwich” variance matrix
Least-Squares Residuals

- $\hat{e}_{t+1} = y_{t+1} - \hat{\beta}' x_t$
- Easy to compute
- Overfit (tend to be too small) when model dimension is large relative to sample size
Leave One-Out Residuals

\[ \tilde{e}_{t+1} = y_{t+1} - \hat{\beta}'_{-t} x_t \]

\[ \hat{\beta}_{-t} = \left( \sum_{j \neq t} x_j x'_j \right)^{-1} \left( \sum_{j \neq t} x_j y_{j+1} \right) \]

- No tendency to overfit
- Easy to compute:
  - \[ \tilde{e}_{t+1} = \frac{\tilde{e}_{t+1}}{1 - h_{tt}} \text{ where } h_{tt} = x'_t (X'X)^{-1} x_t \]
  - Not necessary to actually compute \( n \) regressions!
Computation in R

Regressor Matrix: $x$

- $xxi = \text{solve}(t(x) \times x)$
- $h = \text{rowSums}((x \times xxi) \times x)$

Commands

- $t(x)$ = trace of $x$
- $\times \times$ = matrix multiplication
- $\text{solve}(a)$ = inverse of matrix $a$
- $\text{rowSums} = \text{sum across column by row}$
Sequential Prediction Residuals

- $\bar{e}_{t+1} = y_{t+1} - \hat{\beta}_t' x_t$
- $\hat{\beta}_t = \left( \sum_{j=0}^{t-1} x_j x'_j \right)^{-1} \left( \sum_{j=0}^{t-1} x_j y_{j+1} \right)$

Commonly used for pseudo out-of-sample forecast evaluation

- However, $\hat{\beta}_t$ 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

\[ \hat{V} = \hat{Q}^{-1} \hat{\Omega} \hat{Q}^{-1} \]

\[ \hat{Q} = \frac{1}{n} \sum_{t=0}^{n-1} x_t x_t' \]

\[ \hat{\Omega} = \frac{1}{n} \sum_{t=0}^{n-1} x_t x_t' \hat{e}_{t+1}^2 \]

- Can use least-squares residuals \( \hat{e}_{t+1} \) instead of leave-one-out residuals, but then multiply \( \hat{V} \) by \( n / (n - \text{dim}(x_t)) \).

- Standard errors for \( \hat{\beta} \) are the square roots of the diagonal elements of \( n^{-1} \hat{V} \)

- 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)

\[
y_{t+1} = \beta_0 + \beta_1 y_t + \beta_2 y_{t-1} + \beta_3 y_{t-2} + \beta_4 y_{t-3} + \epsilon_{t+1}
\]

<table>
<thead>
<tr>
<th>Intercept ( \hat{\beta} )</th>
<th>( s(\hat{\beta}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.54 (0.45)</td>
</tr>
<tr>
<td>( \triangle \log(GDP_t) )</td>
<td>0.29 (0.09)</td>
</tr>
<tr>
<td>( \Delta \log(GDP_{t-1}) )</td>
<td>0.18 (0.10)</td>
</tr>
<tr>
<td>( \Delta \log(GDP_{t-2}) )</td>
<td>-0.05 (0.08)</td>
</tr>
<tr>
<td>( \Delta \log(GDP_{t-3}) )</td>
<td>0.06 (0.10)</td>
</tr>
</tbody>
</table>
### Point Forecast - GDP Growth

- **AR(4)**

<table>
<thead>
<tr>
<th>Year</th>
<th>Actual</th>
<th>Forecast</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011:1</td>
<td>0.36</td>
<td></td>
</tr>
<tr>
<td>2011:2</td>
<td>1.33</td>
<td></td>
</tr>
<tr>
<td>2011:3</td>
<td>1.80</td>
<td></td>
</tr>
<tr>
<td>2011:4</td>
<td>2.91</td>
<td></td>
</tr>
<tr>
<td>2012:1</td>
<td>1.84</td>
<td></td>
</tr>
<tr>
<td>2012:2</td>
<td></td>
<td>2.59</td>
</tr>
</tbody>
</table>
Interest Rate Example

- \( y_t = \Delta \text{Rate}_t \)
- AR(12) (reasonable benchmark for monthly data)

<table>
<thead>
<tr>
<th></th>
<th>( \hat{\beta} )</th>
<th>( s(\hat{\beta}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-0.002</td>
<td>(0.01)</td>
</tr>
<tr>
<td>( \Delta \text{Rate}_t )</td>
<td>0.40</td>
<td>(0.06)</td>
</tr>
<tr>
<td>( \Delta \text{Rate}_{t-1} )</td>
<td>-0.26</td>
<td>(0.07)</td>
</tr>
<tr>
<td>( \Delta \text{Rate}_{t-2} )</td>
<td>0.11</td>
<td>(0.06)</td>
</tr>
<tr>
<td>( \Delta \text{Rate}_{t-3} )</td>
<td>-0.07</td>
<td>(0.07)</td>
</tr>
<tr>
<td>( \Delta \text{Rate}_{t-4} )</td>
<td>0.10</td>
<td>(0.07)</td>
</tr>
<tr>
<td>( \Delta \text{Rate}_{t-5} )</td>
<td>-0.08</td>
<td>(0.07)</td>
</tr>
<tr>
<td>( \Delta \text{Rate}_{t-6} )</td>
<td>-0.05</td>
<td>(0.06)</td>
</tr>
<tr>
<td>( \Delta \text{Rate}_{t-7} )</td>
<td>-0.09</td>
<td>(0.06)</td>
</tr>
<tr>
<td>( \Delta \text{Rate}_{t-8} )</td>
<td>-0.01</td>
<td>(0.07)</td>
</tr>
<tr>
<td>( \Delta \text{Rate}_{t-9} )</td>
<td>0.03</td>
<td>(0.07)</td>
</tr>
<tr>
<td>( \Delta \text{Rate}_{t-10} )</td>
<td>0.09</td>
<td>(0.07)</td>
</tr>
<tr>
<td>( \Delta \text{Rate}_{t-11} )</td>
<td>-0.08</td>
<td>(0.06)</td>
</tr>
</tbody>
</table>
## Point Forecast - 10-year Treasury Rate

- **AR(12)**

<table>
<thead>
<tr>
<th></th>
<th>Actual</th>
<th>Forecast</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Level</td>
<td>Change</td>
</tr>
<tr>
<td>2012:1</td>
<td>1.97</td>
<td>-0.01</td>
</tr>
<tr>
<td>2012:2</td>
<td>1.97</td>
<td>0.00</td>
</tr>
<tr>
<td>2012:3</td>
<td>2.17</td>
<td>0.20</td>
</tr>
<tr>
<td>2012:4</td>
<td>2.05</td>
<td>-0.12</td>
</tr>
<tr>
<td>2012:5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Model</th>
<th>Forecast</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR(0)</td>
<td>2.99</td>
</tr>
<tr>
<td>AR(1)</td>
<td>2.59</td>
</tr>
<tr>
<td>AR(2)</td>
<td>2.65</td>
</tr>
<tr>
<td>AR(3)</td>
<td>2.68</td>
</tr>
<tr>
<td>AR(4)</td>
<td>2.59</td>
</tr>
<tr>
<td>AR(5)</td>
<td>2.83</td>
</tr>
<tr>
<td>AR(6)</td>
<td>2.83</td>
</tr>
<tr>
<td>AR(7)</td>
<td>2.83</td>
</tr>
<tr>
<td>AR(8)</td>
<td>2.78</td>
</tr>
<tr>
<td>AR(9)</td>
<td>2.87</td>
</tr>
<tr>
<td>AR(10)</td>
<td>2.87</td>
</tr>
<tr>
<td>AR(11)</td>
<td>2.91</td>
</tr>
<tr>
<td>AR(12)</td>
<td>3.45</td>
</tr>
</tbody>
</table>
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)
  - $\hat{\beta}(m)$, $m = 1, \ldots, M$
- Corresponding forecasts $\hat{f}_{n+1|n}(m)$
- There is some population criterion $C(m)$ which evaluates the accuracy of $\hat{f}_{n+1|n}(m)$
  - $m_0 = \arg\min_m C(m)$ is infeasible best estimator
- There is a sample estimate $\hat{C}(m)$ of $C(m)$
  - $\hat{m} = \arg\min_m \hat{C}(m)$ is empirical analog of $m_0$
  - $\hat{\beta}(\hat{m})$ is selected estimator
  - $\hat{f}_{n+1|n}(\hat{m})$ 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 $y_t$, $t = 1, \ldots, n$
- Estimation of an AR(k) requires $k$ initial conditions, so the effective sample is for observations $t = 1 + k, \ldots, n$
- The sample varies with $k$, sample size is $n - k$
- For valid comparison of AR(k) models for $k = 1, \ldots, K$
  - Fix sample with observations $t = 1 + K, \ldots, n$
  - $n - K$ 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

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

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

where

- $k(m) = \#$ of parameters
- $\hat{\sigma}^2(m) = n^{-1} \sum_{t=0}^{n-1} \hat{e}_{t+1}^2(m) =$ MLE estimate of $\sigma^2$ 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 \( y = y_1, \ldots, y_n \) is \( f(y) = \prod f(y_i) \)
- A model density \( g(y, \theta) = \prod g(y_i, \theta) \)
- The Kullback-Leibler information criterion (KLIC) is

\[
KLIC(f, g) = \int f(y) \log \left( \frac{f(y)}{g(y, \theta)} \right) dy
\]

\[
= \int f(y) \log f(y) dy - \int f(y) \log g(y, \theta) dy
\]

\[
= C_f - E \log g(y, \theta)
\]

where the constant \( C_f = \int f(y) \log f(y) dy \) 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 $\theta_0$ is the maximizer of $E \log g(y, \theta)$
- Equivalently, $\theta_0$ minimizes $KLIC(f, g(\theta))$. 
The negative log-likelihood function is

\[ L(\theta) = -\log g(y, \theta) \]

The (quasi) MLE is \( \hat{\theta} = \arg\min_\theta L(\theta) \).

The fitted log-likelihood is \( L(\hat{\theta}) = -\log g(y, \hat{\theta}(y)) \).

Under general conditions, \( \hat{\theta} \to_p \theta_0 \)

The QMLE estimates the best-fitting density, where best is measured in terms of the KLIC.
Asymptotic Theory

\[ \sqrt{n} \left( \hat{\theta}_{QMLE} - \theta_0 \right) \to_d N(0, V) \]

\[ V = Q^{-1} \Omega Q^{-1} \]

\[ Q = -E \frac{\partial^2}{\partial \theta \partial \theta'} \log g(y, \theta) \]

\[ \Omega = E \left( \frac{\partial}{\partial \theta} \log g(y, \theta) \frac{\partial}{\partial \theta} \log g(y, \theta)' \right) \]

If the model is correctly specified \((g(y, \theta_0) = f(y))\), then \(Q = \Omega\) (the information matrix equality). Otherwise \(Q \neq \Omega\).
The MLE $\hat{\theta} = \hat{\theta}(y)$ is a function of the data vector $y$. The fitted model at any $\tilde{y}$ is $\hat{g}(\tilde{y}) = g(\tilde{y}, \hat{\theta}(y))$. The fitted likelihood is $L(\hat{\theta}) = -\log g(y, \hat{\theta}(y))$ (the model evaluated at the observed data). The KLIC of the fitted model is

$$KLIC(f, g) = C_f - \int f(\tilde{y}) \log g(\tilde{y}, \hat{\theta}(y)) d\tilde{y} = C_f - E_{\tilde{y}} \log g(\tilde{y}, \hat{\theta}(y))$$

where $\tilde{y}$ has density $f$, independent of $y$. 
The expected KLIC is the expectation over the observed values $y$

$$E (KLIC(f, g)) = C_f - E_y E_{\tilde{y}} \log g(\tilde{y}, \hat{\theta}(y))$$

$$= C_f - E_{\tilde{y}} E_y \log g(y, \hat{\theta}(\tilde{y}))$$

$$= C_f + T$$

where

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

the second equality by symmetry, and the third setting $\tilde{\theta} = \hat{\theta}(\tilde{y})$, and $y$ and $\tilde{\theta}$ are independent.
Estimating KLIC

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

$$- \log g(y, \tilde{\theta}) \simeq \mathcal{L}(\hat{\theta}) + \frac{n}{2} (\tilde{\theta} - \hat{\theta})' Q (\tilde{\theta} - \hat{\theta})$$

- Asymptotically,

$$\sqrt{n} (\tilde{\theta} - \hat{\theta}) \to_d Z \sim N(0, 2Q^{-1}\Omega Q^{-1})$$

- Take expectations

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

$$\simeq E\mathcal{L}(\hat{\theta}) + \frac{1}{2} E (Z'QZ)$$

$$= E\mathcal{L}(\hat{\theta}) + \text{tr} (Q^{-1}\Omega)$$

- An (asymptotically) unbiased estimate of $T$ is then

$$\hat{T} = \mathcal{L}(\hat{\theta}) + \text{tr} (Q^{-1}\Omega)$$
When $g(x, \theta_0) = f(x)$ (the model is correctly specified) then $Q = \Omega$

- $\text{tr}(Q^{-1}\Omega) = k = \text{dim}(\theta)$
- $\hat{T} = \mathcal{L}(\hat{\theta}) + k$

Akaike Information Criterion (AIC). It is typically written as $2\hat{T}$, e.g.

$$AIC = 2\mathcal{L}(\hat{\theta}) + 2k$$

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

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

\[
TIC = 2 \mathcal{L}(\hat{\theta}) + 2 \text{tr} (\hat{Q}^{-1}\hat{\Omega})
\]

where

\[
\hat{Q} = -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta \partial \theta'} \log g(y_i, \hat{\theta})
\]

\[
\hat{\Omega} = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{\partial}{\partial \theta} \log g(y_i, \hat{\theta}) \frac{\partial}{\partial \theta} \log g(y_i, \hat{\theta})' \right)
\]
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 $\hat{\beta}(m)$ of $\beta$, the point forecast for $y_{n+1}$ is
  \[ f_{n+1|n} = \hat{\beta}(m)'x_n \]

- The forecast error is
  \[ y_{n+1} - f_{n+1|n} = x_n'\beta + e_{t+1} - x_n'\hat{\beta}(m) \]
  \[ = e_{n+1} - x_n'(\hat{\beta}(m) - \beta) \]

- The mean-squared-forecast-error (MSFE) is
  \[ MSFE(m) = E \left( e_{n+1} - x_n'(\hat{\beta}(m) - \beta) \right)^2 \]
  \[ \approx \sigma^2 + E \left( (\hat{\beta}(m) - \beta)'Q(m)(\hat{\beta}(m) - \beta) \right) \]
  where $Q(m) = E(x_nx_n')$.

- The approximation is an equality if $x_n$ is independent of $\hat{\beta}(m)$
  - Ing and Wei (Annals, 2003) show that this holds asymptotically.
Estimation and MSFE

- The MSFE is

\[
MSFE(m) \doteq \sigma^2 + E \left( \left( \hat{\beta}(m) - \beta \right)' Q(m) \left( \hat{\beta}(m) - \beta \right) \right)
\]

\[
= \sigma^2 + MSE(\hat{\beta}(m))
\]

where

\[
MSE(\hat{\beta}(m)) = \text{tr} E \left( Q(m) \left( \hat{\beta}(m) - \beta \right) \left( \hat{\beta}(m) - \beta \right)' \right)
\]

is the weighted mean-squared-error (MSE) of \( \hat{\beta}(m) \) for \( \beta \).

- Given a model \( \beta' x_t \) for the conditional mean, the choice of estimator \( \hat{\beta}(m) \) impacts the MSFE through \( MSE(\hat{\beta}(m)) \).

- The best point forecast (the one with the smallest MSFE) is obtained by using an estimator \( \hat{\beta}(m) \) with the smallest MSE.
Residual Fit

\[ \hat{\sigma}^2 = \frac{1}{n} \sum_{t=0}^{n-1} \hat{e}_{t+1}(m)^2 \]

\[ = \frac{1}{n} \sum_{t=0}^{n-1} e_{t+1}^2 + \frac{1}{n} \sum_{t=0}^{n-1} \left( x_t' \left( \hat{\beta}(m) - \beta \right) \right)^2 \]

\[ - \frac{2}{n} \sum_{t=0}^{n-1} e_{t+1} x_t' \left( \hat{\beta}(m) - \beta \right) \]

- First two terms are estimates of \( MSFE(m) = E \left( e_{n+1} - x_n' \left( \hat{\beta}(m) - \beta \right) \right)^2 \)

- Third term is

\[ \sum_{t=0}^{n-1} e_{t+1} x_t' \left( \hat{\beta}(m) - \beta \right) = e' P(m) e \]

where \( P(m) = X(m) (X(m)'X(m))^{-1} X(m)' \)
Residual Variance as Biased estimate of MSFE

\[ \hat{\sigma}^2 = \frac{1}{n} \sum_{t=0}^{n-1} e_{t+1}^2 + \frac{1}{n} \sum_{t=0}^{n-1} \left( x_t' \left( \hat{\beta}(m) - \beta \right) \right)^2 - \frac{2}{n} e' P(m) e \]

\[ E \left( \hat{\sigma}^2 \right) = \sigma^2 + E \left( x_t' \left( \hat{\beta}(m) - \beta \right) \right)^2 - \frac{2}{n} E \left( e' P(m) e \right) \]

\[ \simeq \text{MSFE}_n(m) - \frac{2}{n} B(m) \]

where

\[ B(m) = E \left( e' P(m) e \right) \]
Relation between Residual variance and MSFE

\[ \hat{\sigma}^2 = \text{MSFE}_n(m) - \frac{2}{n} B(m) \]
\[ B(m) = E(\mathbf{e}' \mathbf{P}(m) \mathbf{e}) \]

- 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

\[ C_n(m) = \hat{\sigma}^2(m) + \frac{2}{n} B(m) \]

- The residual variance plus a optimal penalty \( 2B(m)/n \)
Asymptotic Penalty

From asymptotic theory, for any $m$

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

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

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

Thus

$$\mathbf{e}' \mathbf{P}(m) \mathbf{e} = \left( \frac{1}{\sqrt{n}} \mathbf{e}' \mathbf{X}(m) \right) \left( \frac{1}{n} \mathbf{X}(m)' \mathbf{X}(m) \right)^{-1} \left( \frac{1}{\sqrt{n}} \mathbf{X}(m)' \mathbf{e} \right)$$

$$\rightarrow_d Z(m)' Q(m)^{-1} Z(m)$$

$$= \text{tr} \left( Q(m)^{-1} Z(m) Z(m)' \right)$$
Asymptotic Penalty

\[ \mathbf{e}' \mathbf{P}(m) \mathbf{e} \rightarrow_d \mathbf{Z}(m)' \mathbf{Q}(m)^{-1} \mathbf{Z}(m) \]

\[ = \text{tr} \left( \mathbf{Q}(m)^{-1} \mathbf{Z}(m) \mathbf{Z}(m)' \right) \]

Thus

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

\[ \rightarrow \text{tr} \left( \mathbf{Q}(m)^{-1} E \left( \mathbf{Z}(m) \mathbf{Z}(m)' \right) \right) \]

\[ = \text{tr} \left( \mathbf{Q}(m)^{-1} \mathbf{\Omega}(m) \right) \]
MSFE Criterion for Least-Squares

\[ C_n(m) = \hat{\sigma}^2(m) + \frac{2}{n} \text{tr} \left( Q(m)^{-1} \Omega(m) \right) \]

\[ Q(m) = E \left( x_t(m)x_t(m)' \right) \]
\[ \Omega(m) = E \left( x_t(m)x_t'(m)e_{t+1}^2 \right) \]

This is an (asymptotically) unbiased estimate of the MSFE
Homoskedastic Case

When \( E (e^2_{t+1} \mid l_t) = \sigma^2 \)

then

\[
\Omega(m) = E (x_t(m)x'_t(m)e^2_{t+1}) = Q(m)\sigma^2
\]

\[
\text{tr} \left( Q(m)^{-1}\Omega(m) \right) = \sigma^2 \text{tr} \left( I(m) \right) = \sigma^2 k(m)
\]

\[
C_n(m) = \hat{\sigma}^2(m) + \frac{2}{n} \sigma^2 k(m)
\]

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

\[ C_n(m) = \tilde{\sigma}^2(m) + \frac{2}{n} \sigma^2 k(m) \]

- Replace the unknown \( \sigma^2 \) with a preliminary estimate \( \tilde{\sigma}^2 \)
  - bias-corrected residual variance from a “large” model

\[
\tilde{\sigma}^2 = \frac{1}{n - K} \sum_{t=0}^{n-1} \hat{e}_{t+1}(K)^2
\]

\[ C_n(m) = \tilde{\sigma}^2(m) + \frac{2}{n} \tilde{\sigma}^2 k(m) \]

- Sometimes written as

\[ C_n(m) = \sum_{t=0}^{n-1} \hat{e}_{t+1}(m)^2 + 2\tilde{\sigma}^2 k(m) \]
Final Prediction Error (FPE) Criterion

\[ C_n(m) = \hat{\sigma}^2(m) + \frac{2}{n} \sigma^2 k(m) \]

- Replace the unknown \( \sigma^2 \) with \( \hat{\sigma}^2(m) \)

\[ FPE_n(m) = \hat{\sigma}^2(m) \left( 1 + \frac{2}{n} k(m) \right) \]
Relations between Mallows, FPE, and Akaike

- Take log of FPE and multiply by $n$

\[
n \log (FPE_n(m)) = n \log (\hat{\sigma}^2(m)) + n \log \left(1 + \frac{2}{n} k(m)\right)
\]

\[
\simeq n \log (\hat{\sigma}^2(m)) + 2k(m)
\]

\[
= AIC(m)
\]

- Thus Mallows, FPE and Akaike model selection is quite similar

- Mallows, FPE, and $\exp \left(\frac{AIC(m)}{n}\right)$ are estimates of MSFE under homoskedasticity
Robust Mallows

- Ideal Criterion

\[ C_n(m) = \hat{\sigma}^2(m) + \frac{2}{n} \text{tr} \left( Q(m)^{-1} \Omega(m) \right) \]

\[ Q(m) = E \left( x_t(m) x_t(m)' \right) \]

\[ \Omega(m) = E \left( x_t(m) x_t'(m) e_t^2 \right) \]

- Sample estimate

\[ C_n^*(m) = \hat{\sigma}^2(m) + \frac{2}{n} \text{tr} \left( \hat{Q}(m)^{-1} \hat{\Omega}(m) \right) \]

\[ \hat{Q}(m) = \frac{1}{n} \sum_{t=0}^{n-1} x_t x_t' \]

\[ \hat{\Omega}(m) = \frac{1}{n} \sum_{t=0}^{n-1} x_t x_t' \tilde{e}_{t+1}^2 \]

where \( \tilde{e}_{t+1} \) is residual from a preliminary estimate

- Robust Mallows similar to TIC, not
Cross-Validation

- **Leave-one-out estimator**

\[ \hat{\beta}_{-t}(m) = \left( \sum_{j \neq t} x_j(m)x_j(m)' \right)^{-1} \left( \sum_{j \neq t} x_j(m)y_{j+1} \right) \]

- **Leave-one-out prediction residual**

\[ \tilde{e}_{t+1}(m) = y_{t+1} - \hat{\beta}_{-t}(m)'x_t(m) \]
\[ = \frac{\hat{e}_{t+1}(m)}{1 - h_{tt}(m)} \]

- \( \tilde{e}_{t+1}(m) \) is a forecast error based on estimation without observation \( t \)

- \( E\tilde{e}_{t+1}(m)^2 \approx MSFE_n(m) \)

- \( CV_n(m) = \frac{1}{n} \sum_{t=0}^{n-1} \tilde{e}_{t+1}(m)^2 \) 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, \( \frac{1}{(1-a)^2} \simeq 1 - 2a \)

\[
CV_n(m) = \frac{1}{n} \sum_{t=0}^{n-1} \tilde{e}_{t+1}(m)^2 \\
= \frac{1}{n} \sum_{t=0}^{n-1} \frac{\hat{e}_{t+1}(m)^2}{(1 - h_{tt}(m))^2} \\
\simeq \frac{1}{n} \sum_{t=0}^{n-1} \hat{e}_{t+1}(m)^2 + 2 \frac{1}{n} \sum_{t=0}^{n-1} \hat{e}_{t+1}(m)^2 h_{tt}(m) \\
= \hat{\sigma}^2(m) + \frac{2}{n} \sum_{t=0}^{n-1} \hat{e}_{t+1}(m)^2 x_t' (X'X)^{-1} x_t \\
= \hat{\sigma}^2(m) + \frac{2}{n} \text{tr} \left( (X'X)^{-1} \sum_{t=0}^{n-1} \hat{e}_{t+1}(m)^2 x_t x_t' \right) \\
= C_n^*(m)
\]
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

\[ \hat{\beta}_t(m) = \left( \sum_{j=0}^{t-1} x_j(m)x_j(m)' \right)^{-1} \left( \sum_{j=0}^{t-1} x_j(m)y_{j+1} \right) \]

- Sequential prediction residuals

\[ \bar{e}_{t+1}(m) = y_{t+1} - \hat{\beta}_t(m)'x_t(m) \]

- Predictive Least Squares. For some \( P \)

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

- 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
- $\bar{e}_{t+1}(m)$ 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
  - $\bar{e}_{t+1}(m)$ 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 $\beta = (\beta_1, \ldots, \beta_P)$
- $\hat{\beta}$ minimizes the penalized least-squares criterion

$$S(\beta) = \sum_{t=0}^{n-1} (y_{t+1} - \beta' x_t)^2 + \lambda \sum_{j=1}^{P} |\beta_j|$$

- 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 $\lambda$ important
Comments on LASSO

- Theory for time-series and forecasting not well developed
- Current theory suggests LASSO appropriate for **sparse** models
  - Most coefficients 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 $\hat{m}$ be the selected model
- Let $MSFE_n(\hat{m})$ denote the MSFE using the selected estimator
- We say that selection is asymptotically optimal if
  \[
  \frac{MSFE_n(\hat{m})}{\inf_m MSFE_n(m)} \xrightarrow{p} 1
  \]
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
- AIC, Mallows, FPE, CV are asymptotically optimal for selection under homoskedasticity
- CV is asymptotically optimal for selection 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, AICₕ, Robust Mallows, CV

<table>
<thead>
<tr>
<th>Model</th>
<th>BIC</th>
<th>AICₕ</th>
<th>Cₙ*</th>
<th>CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR(1)</td>
<td>473</td>
<td>466</td>
<td>10.7</td>
<td>10.7</td>
</tr>
<tr>
<td>AR(2)</td>
<td>472</td>
<td>462</td>
<td>10.6</td>
<td>10.5</td>
</tr>
<tr>
<td>AR(3)</td>
<td>477</td>
<td>464</td>
<td>10.7</td>
<td>10.7</td>
</tr>
<tr>
<td>AR(4)</td>
<td>481</td>
<td>465</td>
<td>10.8</td>
<td>10.8</td>
</tr>
<tr>
<td>AR(5)</td>
<td>483</td>
<td>464</td>
<td>10.8</td>
<td>10.8</td>
</tr>
<tr>
<td>AR(6)</td>
<td>489</td>
<td>466</td>
<td>11.0</td>
<td>10.9</td>
</tr>
<tr>
<td>AR(7)</td>
<td>494</td>
<td>468</td>
<td>11.1</td>
<td>11.1</td>
</tr>
<tr>
<td>AR(8)</td>
<td>498</td>
<td>470</td>
<td>11.3</td>
<td>11.2</td>
</tr>
<tr>
<td>AR(9)</td>
<td>500</td>
<td>469</td>
<td>11.3</td>
<td>11.2</td>
</tr>
<tr>
<td>AR(10)</td>
<td>505</td>
<td>471</td>
<td>11.4</td>
<td>11.4</td>
</tr>
<tr>
<td>AR(11)</td>
<td>511</td>
<td>473</td>
<td>11.5</td>
<td>11.5</td>
</tr>
<tr>
<td>AR(12)</td>
<td>511</td>
<td>471</td>
<td>11.4</td>
<td>11.3</td>
</tr>
</tbody>
</table>

Methods select AR(2)
### 10-Year Treasury Rate

<table>
<thead>
<tr>
<th>Model</th>
<th>BIC</th>
<th>$\text{AIC}_c$</th>
<th>$C_n^*$</th>
<th>CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR(1)</td>
<td>−1518</td>
<td>−1527</td>
<td>0.0798</td>
<td>0.0798</td>
</tr>
<tr>
<td>AR(2)</td>
<td>−1541*</td>
<td>−1554</td>
<td>0.0768*</td>
<td>0.0768*</td>
</tr>
<tr>
<td>AR(3)</td>
<td>−1538</td>
<td>−1555</td>
<td>0.0769</td>
<td>0.0769</td>
</tr>
<tr>
<td>AR(4)</td>
<td>−1532</td>
<td>−1554</td>
<td>0.0773</td>
<td>0.0773</td>
</tr>
<tr>
<td>AR(6)</td>
<td>−1531</td>
<td>−1561</td>
<td>0.0772</td>
<td>0.0770</td>
</tr>
<tr>
<td>AR(8)</td>
<td>−1522</td>
<td>−1562</td>
<td>0.0777</td>
<td>0.0774</td>
</tr>
<tr>
<td>AR(10)</td>
<td>−1513</td>
<td>−1561</td>
<td>0.0784</td>
<td>0.0781</td>
</tr>
<tr>
<td>AR(12)</td>
<td>−1506</td>
<td>−1563</td>
<td>0.079</td>
<td>0.0787</td>
</tr>
<tr>
<td>AR(20)</td>
<td>−1471</td>
<td>−1561</td>
<td>0.081</td>
<td>0.080</td>
</tr>
<tr>
<td>AR(22)</td>
<td>−1470</td>
<td>−1570*</td>
<td>0.081</td>
<td>0.080</td>
</tr>
<tr>
<td>AR(24)</td>
<td>−1458</td>
<td>−1565</td>
<td>0.081</td>
<td>0.081</td>
</tr>
</tbody>
</table>

Mallows, $\text{AIC}_c$, 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)**

<table>
<thead>
<tr>
<th>Year</th>
<th>Actual</th>
<th>Forecast</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011:1</td>
<td>0.36</td>
<td></td>
</tr>
<tr>
<td>2011:2</td>
<td>1.33</td>
<td></td>
</tr>
<tr>
<td>2011:3</td>
<td>1.80</td>
<td></td>
</tr>
<tr>
<td>2011:4</td>
<td>2.91</td>
<td></td>
</tr>
<tr>
<td>2012:1</td>
<td>1.84</td>
<td></td>
</tr>
<tr>
<td>2012:2</td>
<td></td>
<td>2.65</td>
</tr>
</tbody>
</table>
## Point Forecast - 10-year Treasury Rate

- **AR(2)**

<table>
<thead>
<tr>
<th>Year</th>
<th>Actual Level</th>
<th>Change</th>
<th>Forecast Level</th>
<th>Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012:1</td>
<td>1.97</td>
<td>-0.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2012:2</td>
<td>1.97</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2012:3</td>
<td>2.17</td>
<td>0.20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2012:4</td>
<td>2.05</td>
<td>-0.12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2012:5</td>
<td></td>
<td></td>
<td>1.96</td>
<td>-0.09</td>
</tr>
</tbody>
</table>
Recall, the ideal forecast is

\[ E (y_{n+1} | I_n) = E (y_{n+1} | x_n, x_{n-1}, \ldots) \]

where \( I_n \) contains all information

\( x_n = \text{lags + leading indicators} \)

- 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

<table>
<thead>
<tr>
<th>Model</th>
<th>CV</th>
<th>Forecast</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spread</td>
<td>10.4</td>
<td>2.8</td>
</tr>
<tr>
<td>HY Spread</td>
<td>10.6</td>
<td>2.5</td>
</tr>
<tr>
<td>Housing Starts</td>
<td>10.3</td>
<td>1.4</td>
</tr>
<tr>
<td>Building Permits</td>
<td>10.3</td>
<td>1.7</td>
</tr>
<tr>
<td>Sp+HY</td>
<td>10.3</td>
<td>2.7</td>
</tr>
<tr>
<td>Sp+HS</td>
<td>10.02</td>
<td>1.5</td>
</tr>
<tr>
<td>Sp+BP</td>
<td>10.1</td>
<td>1.9</td>
</tr>
<tr>
<td>HY+HS</td>
<td>10.4</td>
<td>1.4</td>
</tr>
<tr>
<td>HY+BP</td>
<td>10.4</td>
<td>1.6</td>
</tr>
<tr>
<td>HS+BP</td>
<td>10.4</td>
<td>1.4</td>
</tr>
<tr>
<td><strong>Sp+HY+HS</strong></td>
<td><strong>10.00</strong></td>
<td><strong>1.3</strong></td>
</tr>
<tr>
<td>Sp+HY+BP</td>
<td>10.1</td>
<td>1.7</td>
</tr>
<tr>
<td>Sp+HS+BP</td>
<td>10.05</td>
<td>1.3</td>
</tr>
<tr>
<td>HY+HS+BP</td>
<td>10.5</td>
<td>1.3</td>
</tr>
<tr>
<td><strong>Sp+HY+HS+BP</strong></td>
<td><strong>10.02</strong></td>
<td><strong>1.1</strong></td>
</tr>
</tbody>
</table>
## Coefficient Estimates

<table>
<thead>
<tr>
<th></th>
<th>$\hat{\beta}$</th>
<th>$s(\hat{\beta})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta \log(GDP_{t+1})$</td>
<td>$-0.33$</td>
<td>$(1.03)$</td>
</tr>
<tr>
<td>Intercept</td>
<td>$-0.33$</td>
<td>$(1.03)$</td>
</tr>
<tr>
<td>$\Delta \log(GDP_t)$</td>
<td>$0.16$</td>
<td>$(0.10)$</td>
</tr>
<tr>
<td>$\Delta \log(GDP_{t-1})$</td>
<td>$0.09$</td>
<td>$(0.10)$</td>
</tr>
<tr>
<td>Bond Spread$_t$</td>
<td>$0.61$</td>
<td>$(0.23)$</td>
</tr>
<tr>
<td>High Yield Spread</td>
<td>$-1.10$</td>
<td>$(0.75)$</td>
</tr>
<tr>
<td>Housing Starts$_t$</td>
<td>$1.86$</td>
<td>$(0.65)$</td>
</tr>
</tbody>
</table>
Alternative Specifications

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

- **Data Set: U.S. macro data**
  - 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](http://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 variables
  - Estimate a set of models, select via information criteria
  - Compute point forecast for next period
Figure: U.S. Unemployment Rate