Forecasting
Lecture 1: Foundations

Bruce E. Hansen

THE UNIVERSITY of WISCONSIN MADISON

Central Bank of Chile
October 29-31, 2013

3-Day Course

- Tuesday
  - Point Forecasting
  - Forecast Selection
  - Leading Indicators
  - Variance Forecasting
  - Interval Forecasting
- Wednesday
  - Combination Forecasts
  - Multi-Step Forecasting
  - Fan Charts
- Thursday: Structural Breaks

Course Website

- `www.ssc.wisc.edu/~bhansen/cbc`
- Slides for all lectures
- Data for the lectures
- R code for empirical analysis for lectures 1 & 2
- Related: `www.ssc.wisc.edu/~bhansen/crete`
    - 5-day forecasting course
    - All the empirical analysis reported here was done in spring 2012, so the "forecasts" appear to be about the past
    - We can compare the forecasts with realizations

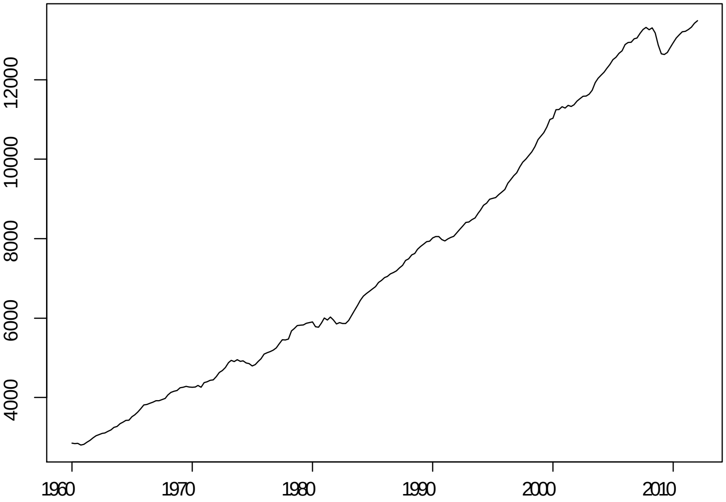
Today's Schedule

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

Example 1

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

Figure: U.S. Real Quarterly GDP



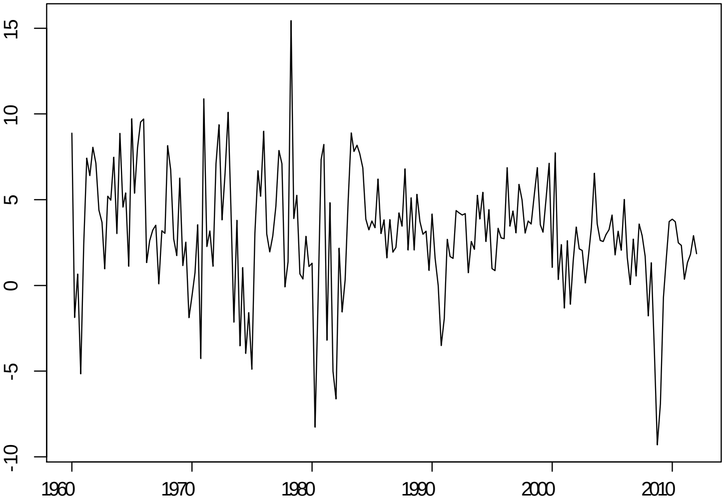
Transformations

- 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

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

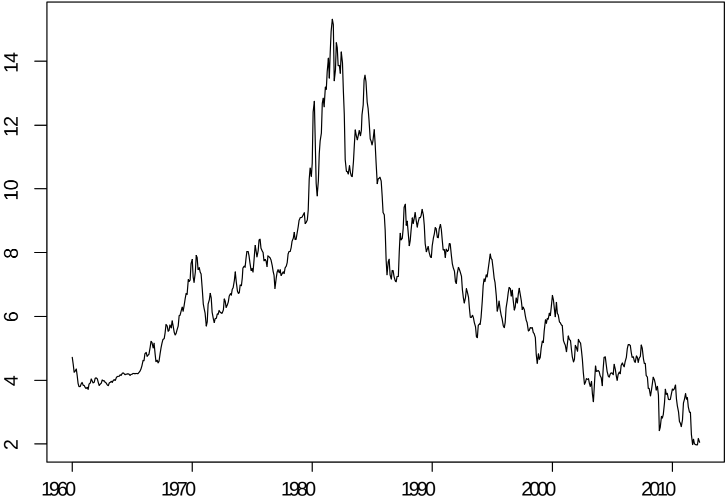
Figure: U.S. Real GDP Quarterly Growth



Example 2

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

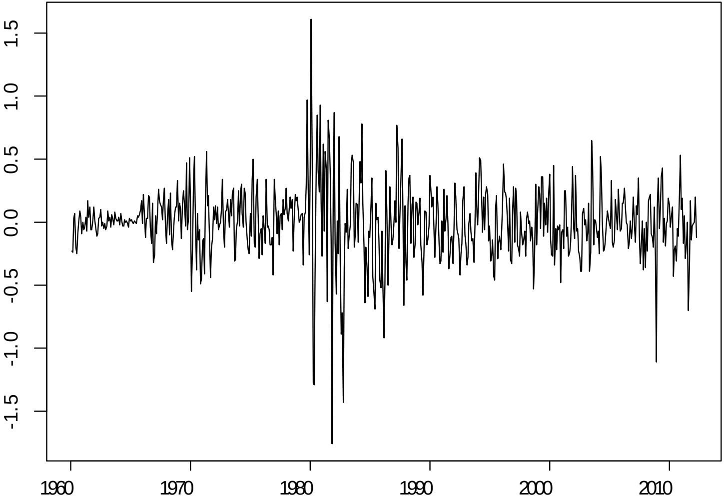
Figure: U.S. 10-Year Treasury Rate



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

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

- 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

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

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

- 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

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

- The conditional variance $\sigma_t^2 = E\left(e_{t+1}^2 | I_t\right)$ may be time-varying

Univariate (Autoregressive) Model

- $x_t = (y_t, y_{t-1}, ..., y_{t-k+1})$
- 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

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

$$\widehat{y}_{n+1|n} = \widehat{f}_{n+1|n} = \widehat{\beta}' \mathbf{x}_n$$

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)

	Data	Forecast	Actual
2011:1	0.36		
2011:2	1.33		
2011:3	1.80		
2011:4	2.91		
2012:1	1.84		
2012:2		2.59	1.20

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)

	Data		Forecast		Actual	
	Level	Change	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	1.82	-0.23

Forecast Selection

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

Point Forecast and MSFE

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

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

- The forecast error is

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

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

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

where $Q(m) = E\left(\mathbf{x}_n\mathbf{x}_n'\right)$.

- A good forecast has low MSFE

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 (BIC)

- 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 (AIC)

- Estimates Kullback-Leibler information criterion (KLIC) distance between true and estimated density

AIC

- In the linear regression model

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

- 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.

Corrected AIC

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

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

- 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
- Critical specification assumption: homoskedasticity
  - AIC is a biased estimate of KLIC under heteroskedasticity
- Criterion: KLIC
  - Not a natural measure of forecast accuracy.

Mallows Criterion

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

- Uses a preliminary estimate $\widetilde{\sigma}^2$ of the variance
- $C_n(m)$ is an (approximately) unbiased estimate of the MSFE under homoskedasticity and one-step forecasting
- Model m which minimizes Mallows criterion is an estimate of the lowest MSFE model

Final Prediction Error (FPE) Criterion

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

Relation between 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 are quite similar
- Mallows, FPE, and $\exp\left(AIC(m)/n \right)$ are estimates of MSFE under homoskedasticity

Robust Mallows

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

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

where $\widetilde{e}_{t+1}$ is residual from a preliminary estimate

- An estimate of MSFE, robust to heteroskedasticity and serial correlation (similar to HAC standard errors)

Cross-Validation

- Leave-one-out estimator and prediction residual

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

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

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

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
- $\bar{e}_{t+1}(m)$ 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

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
- Valid comparisons require holding sample constant across models
- All methods except CV and PLS require conditional homoskedasticity
- PLS sensitive to choice of P
- BIC 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)

	Data	Forecast	Actual
2011:1	0.36		
2011:2	1.33		
2011:3	1.80		
2011:4	2.91		
2012:1	1.84		
2012:2		2.65	1.20

Point Forecast - 10-year Treasury Rate

- AR(2)

	Data		Forecast		Actual	
	Level	Change	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	1.82	-0.23

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

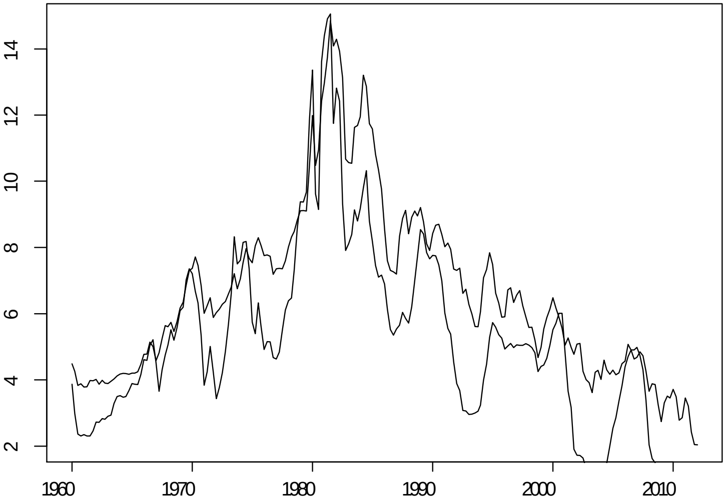
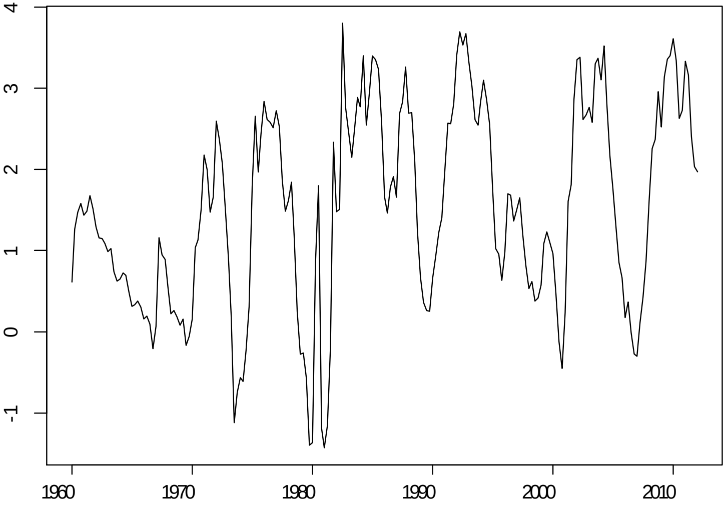


Figure: Term Spread



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

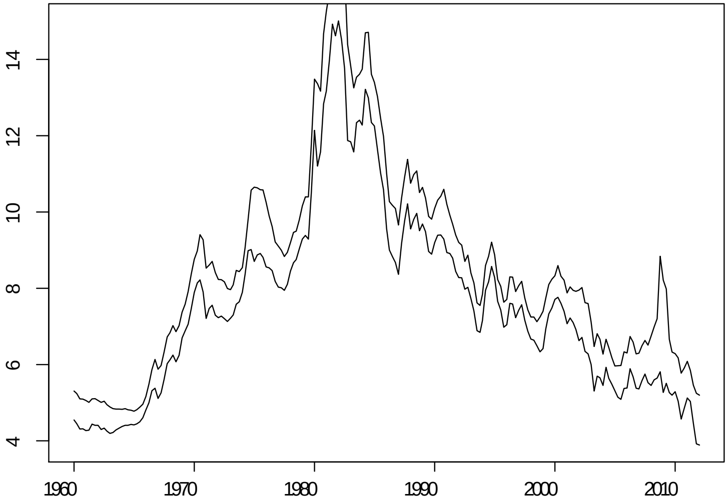
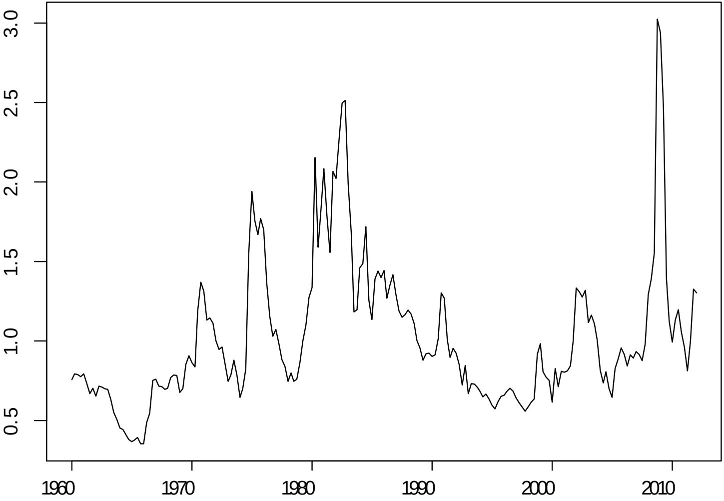


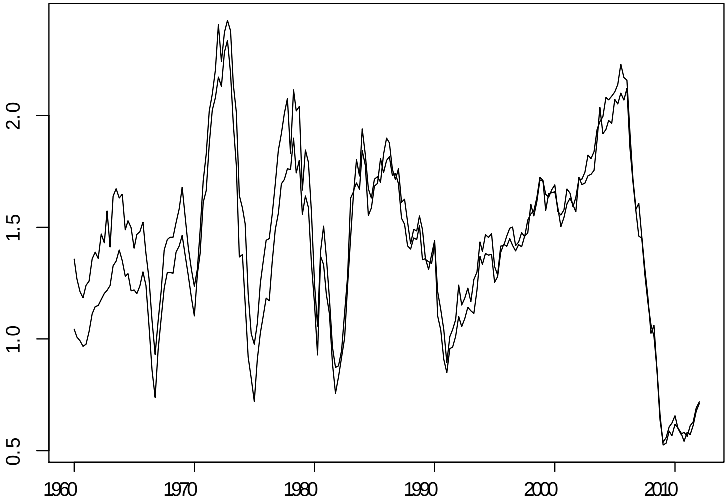
Figure: High Yield Spread



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
Sp+HY+HS+BP	10.0[illegible]	1.1

CV-Selected Forecast: 1.3%
Actual: 1.2%

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)

Variance Forecasting

Variance Forecasts

- Forecast uncertainty
  - Point forecasts insufficient!
- $\sigma^2_{t+1} = \operatorname{var}(y_{t+1} | I_t)$
- In the model $y_{t+1} = \boldsymbol{\beta}' \mathbf{x}_t + e_{t+1}$
  - $\sigma^2_{t+1} = \operatorname{var}(e_{t+1} | I_n) = E\left(e^2_{t+1} | I_t\right)$

10-Year Bond Rate

- Prediction Residuals
- Squares

Figure: Leave-One-Out Prediction Residuals

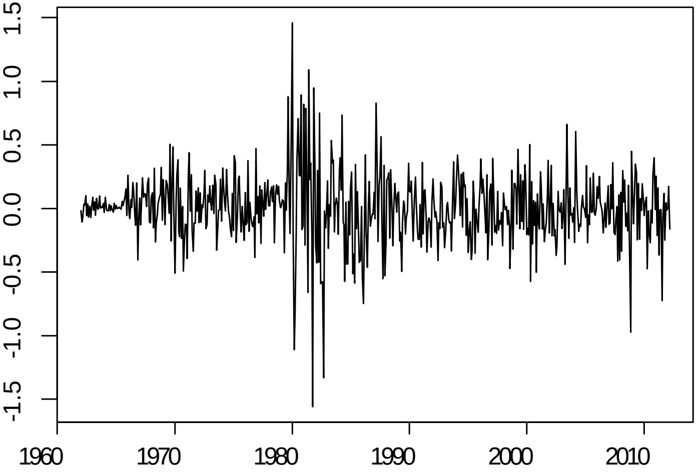
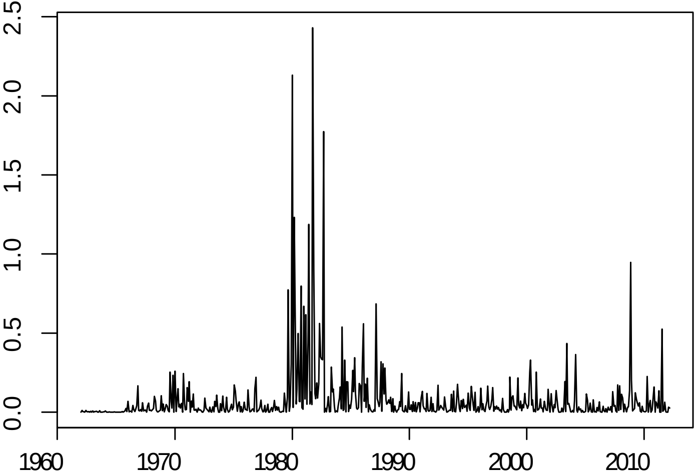


Figure: Squared Prediction Residuals



Variance Forecast Methods

- Constant Variance $\sigma_t^2 = \sigma^2$
  - Uncertainty not state-dependent
- GARCH
  - Common in financial data
  - Estimated by MLE
- Regression Approach
  - $\sigma_t^2 = E\left(e_{t+1}^2 | I_n\right) \approx \boldsymbol{\alpha}'\mathbf{x}_t$

2-Step Variance Estimation

- Start with residuals $\widehat{e}_{t+1}$
  - Better choice: leave-one-out residuals $\widetilde{e}_{t+1}$
- Estimate variance model (constant, ARCH, or regression)
- Obtain $\widehat{\sigma}_n^2$ from fitted model

Which Residuals?

- Least-squares residual variance biased toward zero
  - Forecast variance biased towards zero
- Leave-one-out residual variance estimates out-of-sample MSFE
  - Better choice

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

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

Constant Variance Model

- $\sigma_t^2 = \sigma^2$
- $\widehat{\sigma}_n^2 = \widehat{\sigma}^2 = \dfrac{1}{n-1} \sum_{t=1}^{n-1} \widetilde{e}_{t+1}^2$

Regression Variance Model

- $\sigma_t^2 \approx \boldsymbol{\alpha}'\mathbf{x}_t$
- $e_{t+1}^2 = \boldsymbol{\alpha}'\mathbf{x}_t + \eta_t$
- $\widehat{\boldsymbol{\alpha}} = \left(\sum_{t=1}^{n-1} \mathbf{x}_t\mathbf{x}_t'\right)^{-1} \left(\sum_{t=1}^{n-1} \mathbf{x}_t\widetilde{e}_{t+1}^2\right)$
- $\widehat{\sigma}_n^2 = \widehat{\boldsymbol{\alpha}}'\mathbf{x}_n$
    - Easy, but not constrained to $(0, \infty)$

GARCH Models

- $\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \alpha e_t^2$
- Conditional variance of e_{t+1}
- Specifies conditional variance as function of recent squared innovations
- Large innovations (in magnitude) raise conditional variance
- Lagged variance smooths σ_t^2
- Non-negativity constraints: $\omega > 0$, $\beta \geq 0$, $\alpha > 0$

GARCH with Regressors

- $\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \alpha e_t^2 + \gamma x_t$
- $x_t > 0$ useful to constrain regressor to be positive

Estimation by Quasi-LIkelihood

- Numerical optimization

Model Selection

- Model with 2 ARCH lags and 2 regressors

$$\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \alpha_1 e_t^2 + \alpha_2 e_{t-1}^2 + \gamma_1 x_{1t} + \gamma_2 x_{2t}$$

- How many lags? How many regressors?
- Presence of lagged σ_{t-1}^2 complicates issues
  - β not identified when $\alpha_1 = \alpha_2 = \gamma_1 = \gamma_2 = 0$
  - This means conventional tests and information criterion are not correct when the process is close to constant variance
  - We typically ignore this complication
- Since estimation is nonlinear MLE much of model selection & combination literature is not relevant
  - AIC appropriate
  - Unfortunately, not easy to compute with standard packages

AIC for GARCH models

If model m has parameter vector $\theta(m)$ with $k(m)$ elements

- $AIC(m) = 2\mathcal{L}(\widehat{\theta}(m)) + 2k(m)$
- Not standard output

Variance Forecast from GARCH model

- $\sigma_{n+1}^2 = \omega + \beta\sigma_n^2 + \alpha_1 e_n^2$
- $\widehat{\sigma}_{n+1}^2 = \widehat{\omega} + \widehat{\beta}\widehat{\sigma}_n^2 + \widehat{\alpha}_1 \widetilde{e}_n^2$
- $\widehat{\sigma}_{n+1}^2$ is estimated conditional variance of y_{n+1}
- Standard deviation $\sqrt{\widehat{\sigma}_{n+1}^2}$

Example: 10-Year Bond Rate

GARCH(1,1)

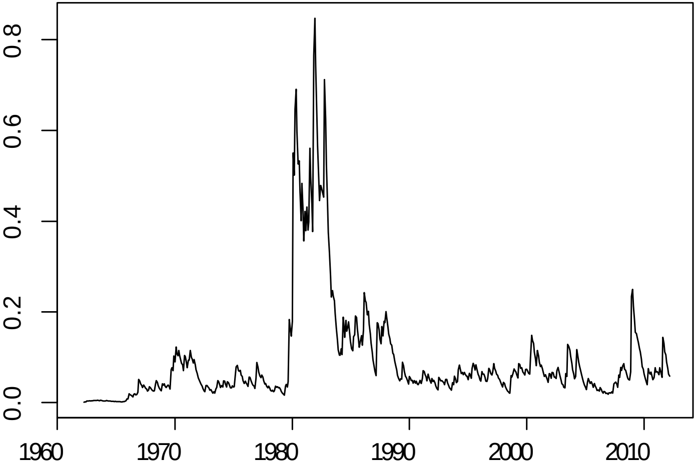
$$\sigma_t^2 = \omega + \alpha e_t^2 + \beta \sigma_{t-1}^2$$

	Estimate	*s.e.*
ω	0.0001	0.0001
α	0.200	0.041
β	0.835	0.025

Variance Forecast

- Conditional variance
  - $\widehat{\sigma}^2_{n+1} = 0.054$
  - $\widehat{\sigma}_{n+1} = 0.23$
- Unconditional
  - $\widehat{\sigma}^2 = 0.076$
  - $\widehat{\sigma} = 0.28$
- The conditional variance at present is similar, but somewhat smaller than the unconditional

Figure: Estimated Variance



Example: GDP Growth

Figure: GDP: Leave-One-Out Prediction Residuals

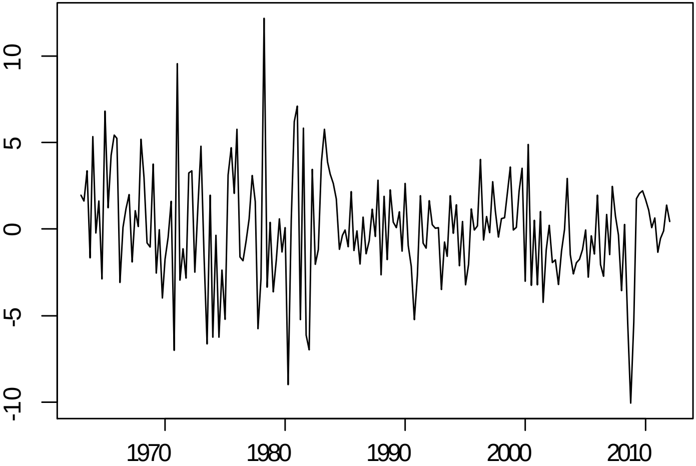
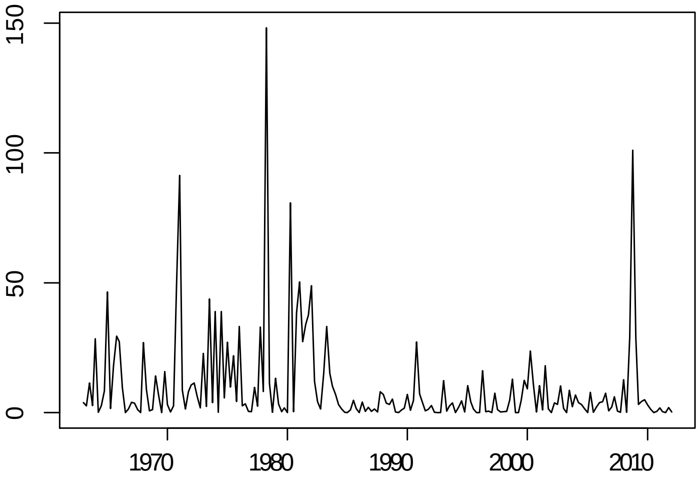


Figure: GDP: Squared Prediction Residuals



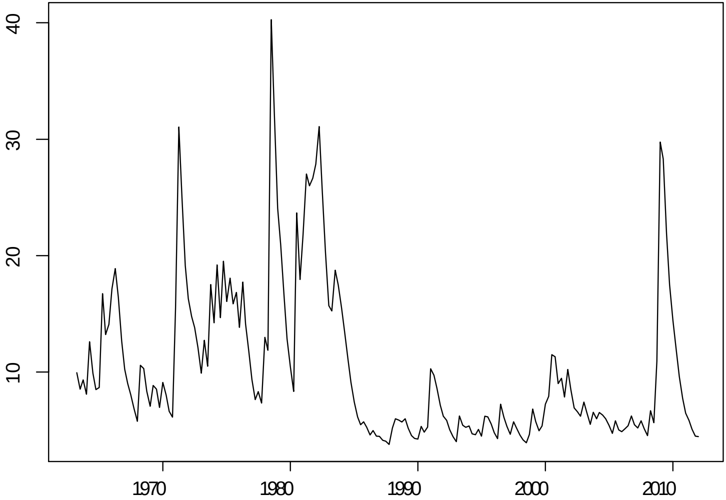
GARCH(1)

$$\sigma_t^2 = \omega + \alpha e_t^2 + \beta \sigma_{t-1}^2$$

	Estimate	*s.e.*
ω	0.81	0.46
α	0.21	0.06
β	0.72	0.06

- Conditional variance
  - $\widehat{\sigma}_{n+1}^2 = 4.1$
  - $\widehat{\sigma}_{n+1} = 2.0$
- Unconditional
  - $\widehat{\sigma}^2 = 9.8$
  - $\widehat{\sigma} = 3.1$

Figure: GDP: Estimated Variance



Interval Forecasting

Interval Forecasts

- Take the form $[a, b]$
- Should contain y_{n+1} with probability $1 - 2\alpha$

$$
\begin{aligned}
1 - 2\alpha &= P_n\left(y_{n+1} \in [a, b]\right) \\
&= P_n\left(y_{n+1} \leq b\right) - P_n\left(y_{n+1} \leq a\right) \\
&= F_n(b) - F_n(a)
\end{aligned}
$$

where $F_n(y)$ is the forecast distribution

- It follows that

$$
\begin{aligned}
a &= q_n(\alpha) \\
b &= q_n(1 - \alpha)
\end{aligned}
$$

- $a = \alpha$'th and $b = (1 - \alpha)$'th quantile of conditional distribution

Interval Forecasts are Conditional Quantiles

- The ideal 80% forecast interval, is the 10% and 90% quantile of the conditional distribution of y_{n+1} given I_n
- Our feasible forecast intervals are estimates of the 10% and 90% quantile of the conditional distribution of y_{n+1} given I_n
- The goal is to estimate conditional quantiles.

Mean-Variance Model

- Write

$$
\begin{aligned}
y_{t+1} &= \mu_t + \sigma_t \varepsilon_{t+1} \\
\mu_t &= E\left(y_{t+1} | I_t\right) \\
\sigma_t^2 &= \operatorname{var}\left(y_{t+1} | I_t\right)
\end{aligned}
$$

- Assume that ε_{t+1} is independent of I_t.
- Let $q_t(\alpha)$ and $q^{\varepsilon}(\alpha)$ be the α'th quantiles of y_{t+1} and ε_{t+1}. Then

$$
q_t(\alpha) = \mu_t + \sigma_t q^{\varepsilon}(\alpha)
$$

- Thus a $(1-2\alpha)$ forecast interval for y_{n+1} is

$$
\left[\mu_n + \sigma_n q^{\varepsilon}(\alpha), \quad \mu_n + \sigma_n q^{\varepsilon}(1-\alpha)\right]
$$

Mean-Variance Model

- Given the conditional mean μ_n and variance σ_n^2, the conditional quantile of y_{n+1} is a linear function $\mu_n + \sigma_n q^{\varepsilon}(\alpha)$ of the conditional quantile $q^{\varepsilon}(\alpha)$ of the normalized error

$$\varepsilon_{n+1} = \frac{e_{n+1}}{\sigma_n}$$

- Interval forecasts thus can be summarized by μ_n, σ_n^2, and $q^{\varepsilon}(\alpha)$

Normal Error Quantile Forecasts

- Make the approximation $\varepsilon_{t+1} \sim N(0,1)$
  - Then $q^{\varepsilon}(\alpha) = Z(a)$ are normal quantiles
  - Useful simplification, especially in small samples
- 0.10, 0.25, 0.75, 0.90 quantiles are
  - -1.285, -0.675, 0.675, 1.285
- Forecast intervals

$$\left[\widehat{\mu}_n + \widehat{\sigma}_n Z(\alpha), \quad \widehat{\mu}_n + \widehat{\sigma}_n Z(1-\alpha)\right]$$

Nonparametric Error Quantile Forecasts

- Let $\varepsilon_{t+1} \sim F$ be unknown
  - We can estimate $q^{\varepsilon}(\alpha)$ as the empirical quantiles of the residuals
  - Set
$$\widehat{\varepsilon}_{t+1} = \frac{\widetilde{e}_{t+1}}{\widehat{\sigma}_t}$$
  - Sort $\widehat{\varepsilon}_1, ..., \widehat{\varepsilon}_n$.
  - $\widehat{q}^{\varepsilon}(\alpha)$ and $\widehat{q}^{\varepsilon}(1-\alpha)$ are the α'th and $(1-\alpha)$'th percentiles
$$\left[\widehat{\mu}_n + \widehat{\sigma}_n \widehat{q}^{\varepsilon}(\alpha), \quad \widehat{\mu}_n + \widehat{\sigma}_n \widehat{q}^{\varepsilon}(1-\alpha)\right]$$
- Computationally simple
- Reasonably accurate when $n \geq 100$
- Allows asymmetric and fat-tailed error distributions

Constant Variance Case

- If $\widehat{\sigma}_t = \widehat{\sigma}$ is a constant, there is no advantage for estimation of $\widehat{\sigma}$ for forecast interval
- Let $\widehat{q}^e(\alpha)$ and $\widehat{q}^e(1-\alpha)$ be the α'th and $(1-\alpha)$'th percentiles of original residuals $\widetilde{e}_{t+1}$
- Forecast Interval:

$$[\widehat{\mu}_n + \widehat{q}^{\varepsilon}(\alpha), \quad \widehat{\mu}_n + \widehat{q}^e(1-\alpha)]$$

- When the estimated variance is a constant, this is numerically identical to the definition with rescaled errors $\widehat{\varepsilon}_{t+1}$

Example: Interest Rate Forecast

- $n = 603$ observations
- $\widehat{\varepsilon}_{t+1} = \dfrac{\widetilde{e}_{t+1}}{\widehat{\sigma}_t}$ from GARCH(1,1) model
- 0.10, 0.25, 0.75, 0.90 quantiles
- -1.16, -0.59, 0.62, 1.26
- Point Forecast = 1.96
- 50% Forecast interval = $[1.82, \quad 2.10]$
- 80% Forecast interval = $[1.69, \quad 2.25]$
- Actual: 1.82

Example: GDP

- $n = 207$ observations
- $\widehat{\varepsilon}_{t+1} = \dfrac{\widetilde{e}_{t+1}}{\widehat{\sigma}_t}$ from GARCH(1,1) model
- 0.10, 0.25, 0.75, 0.90 quantiles
- -1.18, -0.63, 0.57, 1.26
- Point Forecast $= 1.31$
- 50% Forecast interval $= [0.04, \quad 2.4]$
- 80% Forecast interval $= [-1.07, \quad 3.8]$
- Actual: 1.20

Mean-Variance Model Interval Forecasts - Summary

- The key is to break the distribution into the mean μ_t, variance σ_t^2 and the normalized error ε_{t+1}

$$y_{t+1} = \mu_t + \sigma_t \varepsilon_{t+1}$$

- Then the distribution of y_{n+1} is determined by μ_n, σ_n^2 and the distribution of ε_{n+1}
- Each of these three components can be separately approximated and estimated
- Typically, we put the most work into modeling (estimating) the mean μ_t
  - The remainder is modeled more simply
  - For macro forecasts, this reflects a belief (assumption?) that most of the predictability is in the mean, not the higher features.