Measures of Fit from AR(p)

- Residual Sum of Squared Errors $SSR = \sum_{t=1}^{T} \hat{e}_t^2$
- Residual Mean Squared Error $s^2 = \frac{1}{T-p-1} \sum_{t=1}^{T} \hat{e}_t^2$
- Root MSE (Standard Error of Regression)

$$SER = \sqrt{\frac{1}{T-p-1} \sum_{t=1}^{T} \hat{e}_t^2}$$

- R-squared

$$R^2 = \sum_{t=1}^{T} \hat{e}_t^2 \Big/ \sum_{t=1}^{T} (y_t - \bar{y})^2$$

- R-bar-squared

$$\bar{R}^2 = \frac{\frac{1}{T-p-1} \sum_{t=1}^{T} \hat{e}_t^2}{\frac{1}{T-1} \sum_{t=1}^{T} (y_t - \bar{y})^2}$$

Uses

- SSR is a direct measure of the fit of the regression
  - It decreases as you add regressors
- s^2 is an estimate of the error variance
- SER is an estimate of the error standard deviation
- R^2 and R-bar-squared are measures of in-sample forecast accuracy

Example

```
. reg gdp L(1/4).gdp, r

Linear regression                               Number of obs     =        275
                                                F(4, 270)         =       9.86
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1597
                                                Root MSE          =     3.6147
```

- R^2 =0.16
- SER (Root MSE)=3.61

Access after estimation

- STATA stores many of these numbers in "_result"
- _result(1)=T
- _result(2)=MSS (model sum of squares)
- _result(3)=k (number of regressors)
- _result(4)=SSR
- _result(5)=T-k-1
- _result(6)=F-stat (all coefs=0)
- _result(7)=R^2
- _result(8)=R-bar-squared
- _result(9)=SER

Example

```
. dis _result(1)
275

. dis _result(2)
670.39341

. dis _result(3)
4

. dis _result(4)
3527.8675

. dis _result(5)
270
```

```
. dis _result(6)
9.8565788

. dis _result(7)
.1596836

. dis _result(8)
.14723447

. dis _result(9)
3.6147166
```

- (1) T = 275
- (2) MSS = 670
- (3) k = 4
- (4) SSR = 3528
- (5) T-k-1 = 270
- (6) F-stat = 9.9
- (7) R^2 = 0.16
- (8) R-bar-squared = 0.147
- (9) SER = 3.61

Model Selection

- Take the GDP example. Should we use an AR(1), AR(2), AR(3),…?
- How do we pick a forecasting model from among a set of forecasting models?
- This problem is called *model selection*
- There are sets of tools and methods, but there is no universally agreed methodology.

Selection based on Fit

- You could try and pick the model with the smallest SSR or largest R^2.
- But the SSR decreases (and R^2 increases) as you add regressors.
- So this idea would simply pick the largest model.
- Not a useful method!

Selection Based on Testing

- You could test if some coefficients are zero.
- If the test accepts, then set these to zero.
- If the test rejects, keep these variables.
- This is called “selection based on testing”
- You could either use
  - Sequential t-tests
    - From regression output
  - Sequential F-tests
    - .test L3.gdp L4.gdp
    - .testparm L(3/4).gdp

Example: GDP

gdp	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
gdp						
L1.	.3363865	.0743514	4.52	0.000	.1900044	.4827686
L2.	.1422228	.0834467	1.70	0.089	-.0220661	.3065118
L3.	-.0682526	.0721168	-0.95	0.345	-.2102354	.0737302
L4.	-.0728112	.0741333	-0.98	0.327	-.218764	.0731416

```
. testparm L(3/4).gdp

 ( 1)  L3.gdp = 0
 ( 2)  L4.gdp = 0

       F(  2,   270) =    1.09
            Prob > F =    0.3365
```

- Sequential F tests do not reject 4th lag, 3rd+4th, and 2nd+3rd+4th
- Rejects 1st+ 2nd+3rd+4th
- Testing method selects AR(1)

Example: GDP

gdp	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
gdp						
L1.	.3363865	.0743514	4.52	0.000	.1900044	.4827686
L2.	.1422228	.0834467	1.70	0.089	-.0220661	.3065118
L3.	-.0682526	.0721168	-0.95	0.345	-.2102354	.0737302
L4.	-.0728112	.0741333	-0.98	0.327	-.218764	.0731416

```
. testparm L(2/4).gdp

 ( 1)  L2.gdp = 0
 ( 2)  L3.gdp = 0
 ( 3)  L4.gdp = 0

       F(  3,   270) =    1.09
            Prob > F =    0.3518

. testparm L(1/4).gdp

 ( 1)  L.gdp = 0
 ( 2)  L2.gdp = 0
 ( 3)  L3.gdp = 0
 ( 4)  L4.gdp = 0

       F(  4,   270) =    9.86
            Prob > F =    0.0000
```

- F test does not reject that 2nd, 3rd, and 4th lags are jointly zero
- F test rejects that all lags are jointly zero
- Sequential tests select AR(1) model

Sequential t-tests

gdp	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
gdp						
L1.	.3457996	.0739665	4.68	0.000	.2001801	.4914192
L2.	.1309085	.0805715	1.62	0.105	-.0277145	.2895314
L3.	-.094816	.071968	-1.32	0.189	-.2365011	.0468692
L1.	.3337403	.0738275	4.52	0.000	.188399	.4790815
L2.	.0959558	.0713902	1.34	0.180	-.0445872	.2364988
L1.	.3714752	.0661944	5.61	0.000	.2411651	.5017853

- Sequential t-tests also select AR(1)

Select based on Tests?

- Somewhat popular, but *testing* does not lead to good *forecasting* models
- Testing asks if there is strong statistical evidence against a restricted model
- If the evidence is not strong, testing selects the restricted model
- Testing does not attempt to evaluate which model will lead to a better forecast.

Bayes Criterion

- Thomas Bayes (1702-1761) is credited with inventing *Bayes Theorem*
  - M_1 =model 1
  - M_2 =model 2
  - D=Data

$$P(M_1 \mid D) = \frac{P(D \mid M_1)}{P(D \mid M_1)P(M_1) + P(D \mid M_2)P(M_2)}$$

Bayes Selection

- The probabilities $P(M_1)$ and $P(M_2)$ are "priors" believed by the user
- The probabilities $P(D|M_1)$ and $P(D|M_2)$ come from probability models.
- We can then compute the posterior probability of model 1

$$P(M_1 \mid D) = \frac{P(D \mid M_1)}{P(D \mid M_1)P(M_1) + P(D \mid M_2)P(M_2)}$$

Simplification

- AR(p) with normal errors and uniform priors

$$P(M_1 \mid D) \propto \exp\left(-\frac{BIC}{2}\right)$$

where

$$BIC = T\ln\left(\frac{SSR}{T}\right) + (p+1)\ln(T)$$

is known as the *Bayes Information Criterion* or *Schwarz Information Criterion* (SIC). The number *p+1* is the number of estimated coefficients, while *T* is the sample size. SSR/T is the same as the residual variance

Alternative Versions

- BIC is also called SIC (Schwarz Information Criterion) by many authors.
- Our formula is a simplification. The correct formula is

$$BIC = -2L + (p+1)\ln(T)$$

  where *L* is the log-likelihood. This is the formula reported in Stata
- *L* is the joint density of the data. With normal errors, it equals

$$2L = -T(\ln(2\pi)+1) + T\ln\left(\frac{SSR}{T}\right)$$

- The difference in definitions is just the first constant, which does not vary across models and is thus unimportant.
- Also, sometimes BIC is written as

$$BIC = \ln\left(\frac{SSR}{T}\right) + (p+1)\frac{\ln(T)}{T}$$

Bayes Selection

- The Bayes method is to select the model with the highest posterior probability
  - the model with the smallest value of BIC
- Adding/subtracting constants, or multiplying/dividing by constants does not alter the selected model
- The different definitions select the same model

Trade-off

- When we compare models, the larger model (the AR with more lags) will have
  - Smaller SSR
  - Larger p
- The BIC trades these off.
  - The first term is decreasing in p
  - The second term is increasing in p

$$BIC = T\ln\left(\frac{SSR}{T}\right) + (p+1)\ln(T)$$

Computation

- *T*=total number of observations
- For every AR(p) model

$$BIC(p) = T \ln\left(\frac{SSR}{T}\right) + (p+1)\ln(T)$$

- You want to compute BIC(p) for each model.
- Compare the values and select the model with smallest value of BIC

PROBLEM!

- As you vary the AR order *p*, the number of observations changes
  - The sample changes
- This invalidates the BIC comparison. You want only to compare BIC, or any other criterion, on exactly the same sample

Example: AR for GDP

- Available observations: 1947q2 : 2016q4
  - 271 time periods
- An AR(0) uses all T=275 observations
  - 1947q2 : 2016q4
- An AR(1) uses T=274 observations (omits first)
  - 1947q3 : 2016q4
- An AR(2) uses T=273 observations
  - 1947q4 : 2016q4
- An AR(p) uses T=275-p observations
- These are not directly comparable

Solution: Restrict Sample Periods

- First, determine the set of models to compare
  - E.g., AR(0), AR(1), AR(2), AR(3), AR(4)
- Determine a unified sample period where all can be estimated with same number of observations
- Available observations: 1947q2 : 2016q4
- Restricted sample period: 1948q2 : 2016q4
- Omits first 4 observations
  - Uses them only as initial conditions for the AR models
- Use the option “if time>=tq(1948q2)”
  - Where “time” is the time index

GDP Example: AR(1)

- Estimate AR(1) omitting first 4 observations, and save estimates as “ar1”

```
. reg gdp L(1/1).gdp if time>=tq(1948q2), r

Linear regression                               Number of obs     =        275
                                                F(1, 273)         =      31.18
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1387
                                                Root MSE          =     3.6393

------------------------------------------------------------------------------
             |               Robust
         gdp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gdp |
         L1. |   .3722104   .0666621     5.58   0.000     .2409733    .5034474
             |
       _cons |   2.020271   .3307502     6.11   0.000     1.369125    2.671416
------------------------------------------------------------------------------

. estimates store ar1
```

GDP Example: AR(2)

```
. reg gdp L(1/2).gdp if time>=tq(1948q2), r

Linear regression                               Number of obs     =        275
                                                F(2, 272)         =      17.48
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1477
                                                Root MSE          =     3.6271

------------------------------------------------------------------------------
             |               Robust
         gdp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gdp |
         L1. |   .3341523   .0743135     4.50   0.000     .1878495    .4804552
         L2. |   .1016604   .0720406     1.41   0.159    -.0401676    .2434884
             |
       _cons |   1.813011   .3626172     5.00   0.000     1.099118    2.526905
------------------------------------------------------------------------------

. estimates store ar2
```

GDP Example: AR(3)

```
. reg gdp L(1/3).gdp if time>=tq(1948q2), r

Linear regression                               Number of obs     =        275
                                                F(3, 271)         =      12.08
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1552
                                                Root MSE          =     3.6177

------------------------------------------------------------------------------
             |               Robust
         gdp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gdp |
         L1. |   .3434043   .0743793     4.62   0.000     .1969695    .4898391
         L2. |   .1328605   .0809129     1.64   0.102    -.0264373    .2921584
         L3. |  -.0934446   .0720728    -1.30   0.196    -.2353383    .0484491
             |
       _cons |   1.984848   .3752152     5.29   0.000     1.246141    2.723556
------------------------------------------------------------------------------

. estimates store ar3
```

GDP Example: AR(4)

```
. reg gdp L(1/4).gdp if time>=tq(1948q2), r

Linear regression                               Number of obs     =        275
                                                F(4, 270)         =       9.86
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1597
                                                Root MSE          =     3.6147
```

gdp	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
gdp						
L1.	.3363865	.0743514	4.52	0.000	.1900044	.4827686
L2.	.1422228	.0834467	1.70	0.089	-.0220661	.3065118
L3.	-.0682526	.0721168	-0.95	0.345	-.2102354	.0737302
L4.	-.0728112	.0741333	-0.98	0.327	-.218764	.0731416
_cons	2.131395	.4227888	5.04	0.000	1.299012	2.963777

```
. estimates store ar4
```

Display BIC (and AIC)

- For latest estimated model
  - .estimate stats
- For a set of estimated & stored models
  - .estimate stats ar1 ar2 ar3 ar4

GDP Example

- It is convenient to display all models together.
- BIC is displayed in the right column. The AR(1) has the smallest value

```
. estimates stats ar1 ar2 ar3 ar4

Akaike's information criterion and Bayesian information criterion
```

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
ar1	275	-764.9856	-744.4497	2	1492.899	1500.133
ar2	275	-764.9856	-743.0177	3	1492.035	1502.886
ar3	275	-764.9856	-741.8014	4	1491.603	1506.07
ar4	275	-764.9856	-741.0638	5	1492.128	1510.211

Note: N=Obs used in calculating BIC; see **[R] BIC note**.

Check the Samples!

- All models should have the same number of observations (in this case 275)
- If not, you need to alter the samples

```
. estimates stats ar1 ar2 ar3 ar4

Akaike's information criterion and Bayesian information criterion
```

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
ar1	275	-764.9856	-744.4497	2	1492.899	1500.133
ar2	275	-764.9856	-743.0177	3	1492.035	1502.886
ar3	275	-764.9856	-741.8014	4	1491.603	1506.07
ar4	275	-764.9856	-741.0638	5	1492.128	1510.211

Note: N=Obs used in calculating BIC; see **[R] BIC note**.

Problem with BIC

- This is the theory behind the BIC
- If one of the models is true, and the others false,
  - Then BIC selects the model most likely to be true
- If none of the models are true, all are approximations
  - BIC does not necessarily pick a good *forecasting* model
- **BIC selection is not designed to produce a good forecast**

Selection to Minimize MSFE

- Our goal is to produce forecasts with low MSFE (mean-square forecast error).
- If $\hat{y}$ is a forecast for y, the MSFE is

$$R(\hat{y}) = \mathrm{E}\left(y - \hat{y}\right)^2$$

- If we had a good estimate of the MSFE, we could pick the model (forecast) with the smallest MSFE.
- Consider the estimate: The in-sample sum of square residuals, SSR

SSR

- In-sample MSFE

$$SSR = \sum_{T=1}^{T} \left(y_t - \hat{y}_t \right)^2$$

$$= \sum_{T=1}^{T} \hat{e}_t^2$$

- Two troubles
  - It is a biased estimate (overfitting in-sample)
  - It decreases as you add regressors, it cannot be used for selection

Bias

- It can be shown that (approximately)

$$E(SSR) = E(MSFE) - 2\sigma^2(p+1)$$

and

$$E(MSFE) = T\sigma^2$$

- Shibata (1980) suggested the bias adjustment

$$S_p = SSR \cdot \left(1 + \frac{2(p+1)}{N}\right)$$

- Known as the Shibata criteria.

Akaike

- If you take Shibata's criterion, divide by *T*, take the log, and multiply by *T*, then

$$T\ln\left(\frac{S_p}{T}\right) = T\ln\left(\frac{SSR}{T}\right) + T\ln\left(1+\frac{2(p+1)}{T}\right)$$

$$\cong T\ln\left(\frac{SSR}{T}\right) + 2(p+1)$$

$$= AIC$$

- This looks somewhat like BIC, but "2" has replaced "ln(T)".
- Called the "Akaike Information criterion" (AIC)

Formulas and Comparison

$$AIC = T\ln\left(\frac{SSR}{T}\right) + 2(p+1)$$

$$BIC = T\ln\left(\frac{SSR}{T}\right) + \ln(T)(p+1)$$

- Intuitively, both make similar trade-offs
  - Larger models have smaller SSR, but larger p
  - The difference is that BIC puts a higher penalty on the number of parameters
    - The AIC penalty is 2
    - The BIC penalty is ln(T)>2 (if T>7)
    - For example, if T=240, ln(T)=5.5 is much larger than 2

Hirotugu Akaike

- 1927-2009
- Japanese statistician
- Famous for inventing the AIC

Motivation for AIC

- Motivation 1: The AIC is an approximately unbiased estimate of the MSFE
- Motivation 2 (Akaike's): The AIC is an approximately unbiased estimate of the Kullback-Liebler Information Criterion (KLIC)
  - A loss function on the density forecast
  - Suppose f(y) is a density forecast for y, and g(y) is the true density. The KLIC risk is

$$KLIC(f,g) = E\ln\left(\frac{f(y)}{g(y)}\right)$$

Akaike's Result

- Akaike showed that in a normal autoregression the AIC is an approximately unbiased estimator of the KLIC
- So Akaike recommended selecting forecasting models by finding the one model with the smallest AIC
- Unlike testing or BIC, the AIC is designed to find models with low forecast risk.

GDP Example

- Use the estimates stats command
- AIC is displayed in the next-to-last right column. The AR(3) has the smallest value

```
. estimates stats ar1 ar2 ar3 ar4

Akaike's information criterion and Bayesian information criterion
```

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
ar1	275	-764.9856	-744.4497	2	1492.899	1500.133
ar2	275	-764.9856	-743.0177	3	1492.035	1502.886
ar3	275	-764.9856	-741.8014	4	1491.603	1506.07
ar4	275	-764.9856	-741.0638	5	1492.128	1510.211

Note: N=Obs used in calculating BIC; see **[R] BIC note**.

Comments

- BIC picks AR(1), AIC picks AR(3)
- This is common
  - AIC typically selects a larger model than BIC
  - Mechanically, it is because BIC puts a larger penalty on the dimension of the model
    - ln(T) versus 2
  - Conceptually, it is because
    - BIC assumes that there is a true finite model, and is trying to find the true model
    - AIC assumes all models are approximations, and is trying to find the model which makes the best forecast.
      - Extra lags are included if (on balance) they help to forecast

Selection based on Prediction Errors

- A sophisticated selection method is to compute true out-of-sample forecasts and forecast errors, and pick the model with the smallest out-of-sample forecast variance
  - Instead of forecast variance, you can apply any loss function to the forecast errors

Forecasts

- Your sample is $[y_1, y_T]$ for observations [1,…,T]
- For each y_t, you construct an out-of-sample forecast $\hat{y}_t$.
  - This is typically done on a the observations [R+1,…,T]
  - R is a start-up number
  - P=T-R is the number of out-of-sample forecasts

Out-of-Sample Forecasts

- By out-of sample, $\hat{y}_t$ must be computed using only the observations [1,…,t-1]
- In an AR(1)

$$\hat{y}_t = \hat{\alpha}_{t-1} + \hat{\beta}_{t-1} y_{t-1}$$

- Where the coefficients are estimated using only the observations [1,…,t-1]
- Also called “Pseudo Out-of-Sample” forecasting
- The out-of-sample forecast error is

$$\tilde{e}_t = y_t - \hat{y}_t$$

Forecast error

- The out-of-sample (OOS) forecast error is different than the full-sample least-squares residual
- It is a true forecast error
- An estimate of the mean-square forecast error is the sample variance of the OOS errors

$$\tilde{\sigma}^2 = \frac{1}{P}\sum_{t=R+1}^{T}\tilde{e}_t^2$$

Selection based on pseudo OOS MSE

- The predictive least-squares (PLS) criterion is the estimated MSFE using the OOS forecast errors

$$PLS = \sqrt{\frac{1}{P}\sum_{t=R+1}^{T}\tilde{e}_t^2}$$

- PLS selection picks the model with the smallest PLS criterion
- This is very popular in applied forecasting

Comments on PLS

- PLS has the advantage that it does not depend on approximations or distribution theory
- It can be computed for **any** forecast method
  - You just need a time-series of actual forecasts
  - You can use it to compare published forecasts
- Disadvantages
  - It requires the start-up number of observations R
  - The forecasts in the early part of the sample will be less precise than in the later part
    - Averaging over these errors can be misleading
    - Will therefore tend to select smaller models than AIC
  - Less strong theoretical foundation for PLS than for AIC

Jorma Rissanen

- The idea of PLS is due to Jorma Rissanen, a Finnish information theorist

Computation

- Numerical computation of PLS in STATA is unfortunately tricky
- We will discuss it later when we discuss recursive estimation

PLS picks AR(2) for GDP Growth

AR order	PLS
P=0 (no lag)	3.58
P=1	3.435
P=2	3.432*
P=3	3.47
P=4	3.53
P=5	3.52

Assignments

- Diebold, Chapter 15
- Problem Set #8
  - Due Tuesday (4/4)
- Read Chapter 8 from *The Signal and the Noise*
  - Reading Reflection
  - Thursday (3/30)