Joint Tests

\[ y_t = \alpha + \beta_1 y_{t-1} + \cdots + \beta_p y_{t-p} + e_t \]

- How do we assess if a subset of coefficients are jointly zero? Example: 3\textsuperscript{rd}+4\textsuperscript{th} lags

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

|      | Coef.   | Robust Std. Err. | t    | P>|t|  | [95% Conf. Interval] |
|------|---------|------------------|------|------|---------------------|
| gdp  |         |                  |      |      |                     |
| L1.  | .327656 | .076895          | 4.26 | 0.000| .1761871            | .479125          |
| L2.  | .1466135| .0858808         | 1.71 | 0.089| -.0225558           | .3157828         |
| L3.  | -.0980287| .0728951       | -1.34| 0.180| -.2416186            | .0455611         |
| L4.  | -.0889209| .0790354       | -1.13| 0.262| -.244606            | .0667641         |
| _cons| 2.378427| .4731312        | 5.03 | 0.000| 1.446447            | 3.310408         |
Joint Hypothesis

• This is a joint test of
  \[ \beta_3 = 0 \]
  \[ \beta_4 = 0 \]

• This can be done with an “F test”

• In STATA, after `regress (reg)` or `newey`
  `.test L3.gdp L4.gdp`

• List variables whose coefficients are tested for zero.
Joint Tests

• “F test” named after R.A. Fisher
  – (1890-1992)
  – A founder of modern statistical theory

• Modern form known as a “Wald test”, named after Abraham Wald (1902-1950)
  – Early contributor to econometrics
F test computation

```
. test L3.gdp L4.gdp

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

          F(  2,  242) =  1.76
Prob > F =  0.1747
```

• You need to list each variable separately
• STATA describes the hypothesis
• The value of “F” is the F-statistic
• “Prob>F” is the p-value
  – Small p-values cause rejection of hypothesis of zero coefficients
  – Conventionally, reject hypothesis if p-value < 0.05
Example: 2-step-ahead GDP AR(4)

```
. newey gdp L(2/5).gdp, lag(2)
```

Regression with Newey-West standard errors
maximum lag: 2

|   | Coef.   | Std. Err. | t    | P>|t|   | [95% Conf. Interval] |
|---|---------|-----------|------|-------|----------------------|
| gdp | .2410617 | .0768239  | 3.14 | 0.002 | .0897296 - .3923938  |
| L2. | -.0368004 | .0703583  | -0.52 | 0.601 | -.1753962 - .1017954 |
| L3. | -.0910108 | .0791053  | -1.15 | 0.251 | -.2468369 - .0648152 |
| L4. | -.1128763 | .0687243  | -1.64 | 0.102 | -.2482533 - .0225006 |
| L5. | .329426   | .5460059  | 6.10 | 0.000 | 2.253873 - 4.404979  |

```
. test L3.gdp L4.gdp L5.gdp
```

( 1)  L3.gdp = 0
( 2)  L4.gdp = 0
( 3)  L5.gdp = 0

```
F(  3,   241) =  1.65
Prob > F =   0.1793
```
Testing after Estimation

• The commands **predict** and **test** are applied to the most recently estimated model

• The command test uses the standard error method specified by the estimation command
  - `reg y x`: classical F test
  - `reg r x, r`: heteroskedasticity-robust F test
  - `newey y x, lag(m)`: correlation-robust F test
    • (The robust tests are actually Wald statistics)
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: \( R^2 = \frac{\sum_{t=1}^{T} \hat{e}_t^2}{\sum_{t=1}^{T} (y_t - \bar{y})^2} \)
- R-bar-squared: \( \overline{R}^2 = \frac{1}{T - p - 1} \frac{\sum_{t=1}^{T} \hat{e}_t^2}{\frac{1}{T - 1} \sum_{t=1}^{T} (y_t - \bar{y})^2} \)
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

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>662.232234</td>
<td>4</td>
<td>165.558059</td>
</tr>
<tr>
<td>Residual</td>
<td>3518.78213</td>
<td>242</td>
<td>14.540422</td>
</tr>
<tr>
<td>Total</td>
<td>4181.01437</td>
<td>246</td>
<td>16.9959934</td>
</tr>
</tbody>
</table>

Number of obs = 247
F( 4, 242) = 11.39
Prob > F = 0.0000
R-squared = 0.1584
Adj R-squared = 0.1445
Root MSE = 3.8132
```

- SSR=3518.78
- $s^2 = 14.54$
- $R^2 = 0.158$
- R-bar-squared=0.144
- SER=3.8132
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
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 increases (and $R^2$ decreases) 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
  – Sequential F-tests
**Example: GDP**

| gdp   | Coef.    | Std. Err. | t       | P>|t|   | [95% Conf. Interval] |
|-------|----------|-----------|---------|-------|---------------------|
| gdp L1. | .327656  | .076895   | 4.26    | 0.000 | .1761871 - .479125  |
| L2.    | .1466135 | .0858808  | 1.71    | 0.089 | -.0225558 - .3157828|
| L3.    | -.0980287| .0728951  | -1.34   | 0.180 | -.2416186 - .0455611|
| L4.    | -.0889209| .0790354  | -1.13   | 0.262 | -.244606 - .0667641 |

- Sequential F tests do not reject 4\textsuperscript{th} lag, 3\textsuperscript{rd}+4\textsuperscript{th}, and 2\textsuperscript{nd}+3\textsuperscript{rd}+4\textsuperscript{th}
- Rejects 1\textsuperscript{st}+2\textsuperscript{nd}+3\textsuperscript{rd}+4\textsuperscript{th}
- Testing method selects AR(1)
Example: GDP

|     | Coef.    | Robust Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|-----|----------|------------------|-------|-----|----------------------|
| gdp |          |                  |       |     |                      |
| L1. | 0.327656 | 0.076895         | 4.26  | 0.000 | 0.1761871 0.479125    |
| L2. | 0.1466135| 0.0858808        | 1.71  | 0.089 | -0.0225558 0.3157828  |
| L3. | -0.0980287| 0.0728951       | -1.34 | 0.180 | -0.2416186 0.0455611  |
| L4. | -0.0889209| 0.0790354       | -1.13 | 0.262 | -0.244606 0.0667641   |

. test L3.gdp L4.gdp

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

F(  2,  242) =  1.76  
Prob > F =  0.1747

. test L2.gdp L3.gdp L4.gdp

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

F(  3,  242) =  1.36  
Prob > F =  0.2552

. test L1.gdp L2.gdp L3.gdp L4.gdp

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

F(  4,  242) =  8.85  
Prob > F =  0.0000
### Sequential t-tests

|       | Coef.  | Robust Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|-------|--------|------------------|-------|-----|---------------------|
| gdp   |        |                  |       |     |                     |
| L1.   | .3412071 | .0764232        | 4.46  | 0.000 | .1906738 - .4917405 |
| L2.   | .1327376 | .0826814        | 1.61  | 0.110 | -.0301228 - .2955981 |
| L3.   | -.1293765 | .0731709       | -1.77 | 0.078 | -.2735037 - .0147508 |

|       | Coef.  | Robust Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|-------|--------|------------------|-------|-----|---------------------|
| gdp   |        |                  |       |     |                     |
| L1.   | .3268403 | .0760611        | 4.30  | 0.000 | .1770265 - .476654  |
| L2.   | .0870349 | .0742668        | 1.17  | 0.242 | -.059245 - .2333148 |

|       | Coef.  | Robust Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|-------|--------|------------------|-------|-----|---------------------|
| gdp   |        |                  |       |     |                     |
| L1.   | .3604753 | .0690582        | 5.22  | 0.000 | .22446 - .4964907   |

- 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

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

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

- AR(p) with normal errors and uniform priors

\[ P(M_1 | D) \propto \exp \left( -\frac{T}{2} \cdot BIC \right) \]

where

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

is known as the Bayes Information Criterion or Schwarz Information Criterion (SIC). The number \( N \) is the total number of observations, while \( T \) is the number used for estimation of the AR(p).
Bayes Selection

• The Bayes method is to select the model with the highest posterior probability
  — the model with the smallest value of BIC
• Sometimes BIC is written a bit differently
• But are all equivalent for model selection

\[
BIC_1 = N \ln \left( \frac{SSR}{T} \right) + (p + 1) \ln(N)
\]

\[
BIC_2 = \ln \left( \frac{SSR}{T} \right) + (p + 1) \frac{\ln(N)}{N}
\]
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 $p$

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

• $N=$total number of observations
• For every AR($p$) model

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

• As you change the AR order, the number of observations used for estimation $T$ changes.
  — Do not change $N$ as you vary AR models
Computation

• For a baseline model, record $N$ (example $N=250$)
• Direct calculation
  \[
  \text{.dis ln(_result(4)/_result(1))*250+(1+_result(3))*ln(250)}
  \]
  or
  \[
  \text{.dis ln(e(rss)/e(N))*250+e(rank)*ln(250)}
  \]
  \[
  \_result(1)=e(N)=T
  \]
  \[
  \_result(3)=p
  \]
  \[
  e(rank)=p+1
  \]
  \[
  \_result(4)=e(rss)=SSR
  \]
• Warning:
  – STATA has \textbf{estimates} and \textbf{estat} commands which report “BIC”, but they assume $N=T$ which is not appropriate for AR comparisons
  – Use the direct calculation
Example: AR for GDP

- There are $N=251$ observations
- An AR(0) uses $T=251$
- An AR(1) uses $T=250$ observations
- An AR($p$) uses $T=251-p$ observations
Example: AR(1) for GDP

```
. reg gdp L.gdp

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>548.5238</td>
<td>1</td>
<td>548.5238</td>
</tr>
<tr>
<td>Residual</td>
<td>3663.91099</td>
<td>248</td>
<td>14.7738347</td>
</tr>
<tr>
<td>Total</td>
<td>4212.43479</td>
<td>249</td>
<td>16.9174088</td>
</tr>
</tbody>
</table>

Number of obs = 250
F( 1, 248) = 37.13
Prob > F = 0.0000
R-squared = 0.1302
Adj R-squared = 0.1267
Root MSE = 3.8437

| gdp  | Coef.  | Std. Err. | t     | P>|t|   | [95% Conf. Interval] |
|------|--------|-----------|-------|-------|----------------------|
| gdp  | 0.3604753 | 0.0591595 | 6.09  | 0.000 | 0.2439562 to 0.4769944 |
| L1.  |        |           |       |       |                      |
| _cons | 2.147687 | 0.312436  | 6.87  | 0.000 | 1.532321 to 2.763054  |
```

`. dis ln(_result(4)/_result(1))*251+(1+_result(3))*ln(251)
684.94211

\[
BIC = N \ln \left( \frac{SSR}{T} \right) + (1 + p) \ln(N) = 251 \times \ln \left( \frac{3664}{250} \right) + 4 \ln(251) = 684.9
\]
BIC picks AR(1) for GDP Growth

<table>
<thead>
<tr>
<th>AR order</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>P=0 (no lag)</td>
<td>714.4</td>
</tr>
<tr>
<td>P=1</td>
<td>684.9*</td>
</tr>
<tr>
<td>P=2</td>
<td>689.2</td>
</tr>
<tr>
<td>P=3</td>
<td>690.2</td>
</tr>
<tr>
<td>P=4</td>
<td>694.4</td>
</tr>
<tr>
<td>P=5</td>
<td>698.8</td>
</tr>
</tbody>
</table>
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 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

\[
R(\hat{y}) = E(y - \hat{y})^2
\]

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

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

- 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)

\[ E(\text{SSR}) = E(\text{MSFE}) - 2\sigma^2(p+1) \]

and

\[ E(\text{MSFE}) = T\sigma^2 \]

- Shibata (1980) suggested the bias adjustment

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

- Known as the Shibata criteria.
Akaike

• If you take Shibata’s criterion, divide by $T$, take the log, and multiply by $N$, then

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

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

$$= AIC$$

• This looks somewhat like BIC, but “2” has replaced “$\ln(N)$”.

• Called the “Akaike Information criterion” (AIC)
Formulas and Comparison

\[
\begin{align*}
AIC &= N \ln\left( \frac{SSR}{T} \right) + 2(p + 1) \\
BIC &= N \ln\left( \frac{SSR}{T} \right) + \ln(N)(p + 1)
\end{align*}
\]

- Intuitively, both trade-off 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 \( \ln(N) > 2 \) (if \( N > 7 \))
    - For example, if \( N = 240 \), \( \ln(N) = 5.5 \) 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.
**Computation**

- For given N (e.g. N=251)
- Direct calculation
  
  \[
  \text{.dis } \ln(_\text{result}(4)/_\text{result}(1))\times251+(1+_\text{result}(3))\times2
  \]
  
  Or
  
  \[
  \text{.dis } \ln(\text{e(rss)/e(N)})\times251+\text{e(rank)}\times2
  \]

  \_result(1)=e(N)=T

  \_result(3)=p

  e(rank)=p+1

  \_result(4)=e(rss)=SSR
Example: AR(3) for GDP

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 248</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>639.828998</td>
<td>3</td>
<td>213.276333</td>
<td>F(3, 244) = 14.65</td>
</tr>
<tr>
<td>Residual</td>
<td>3551.16846</td>
<td>244</td>
<td>14.5539691</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>4190.99745</td>
<td>247</td>
<td>16.967601</td>
<td>R-squared = 0.1527</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Adj R-squared = 0.1422</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Root MSE = 3.815</td>
</tr>
</tbody>
</table>

| gdp      | Coef.     | Std. Err. | t    | P>|t|  | [95% Conf. Interval] |
|----------|-----------|-----------|------|------|----------------------|
| gdp      |           |           |      |      |                      |
| L1.      | .3412071  | .0634035  | 5.38 | 0.000| .2163191 .4660952    |
| L2.      | .1327376  | .0664123  | 2.00 | 0.047| .001923 .2635523     |
| L3.      | -.1293765 | .0633675  | -2.04| 0.042| -.2541935 -.0045595  |
| _cons    | 2.193251  | .361578   | 6.07 | 0.000| 1.481039 2.905464   |

\[
AIC = N \ln \left( \frac{SSR}{T} \right) + 2(1 + p) = 251 \times \ln \left( \frac{3551}{248} \right) + 2 \times 4 = 676.1
\]

\[
.3 \ln(_\text{result}(4)/_\text{result}(1)) + 251 \times (1 + _\text{result}(3)) \times 2 = 676.06241
\]
AIC picks AR(3) for GDP Growth

<table>
<thead>
<tr>
<th>AR order</th>
<th>BIC</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>P=0 (no lag)</td>
<td>714.4</td>
<td>710.8</td>
</tr>
<tr>
<td>P=1</td>
<td>684.9*</td>
<td>677.9</td>
</tr>
<tr>
<td>P=2</td>
<td>689.2</td>
<td>678.6</td>
</tr>
<tr>
<td>P=3</td>
<td>690.2</td>
<td>676.1*</td>
</tr>
<tr>
<td>P=4</td>
<td>694.4</td>
<td>676.8</td>
</tr>
<tr>
<td>P=5</td>
<td>698.8</td>
<td>677.7</td>
</tr>
</tbody>
</table>
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(N) 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 \([1, \ldots, T]\)

• For each \(y_t\), you construct an out-of-sample forecast \(\hat{y}_t\).
  – This is typically done on the observations \([R+1, \ldots, T]\)
  – \(R\) 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 \([1,\ldots,t-1]\)
• 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 \([1,\ldots,t-1]\)
• Also called “Pseudo Out-of-Sample” forecasting
  – Diebold, Section 10.3
  – Stock-Watson, Key Concept 14.10
• 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**

<table>
<thead>
<tr>
<th>AR order</th>
<th>BIC</th>
<th>AIC</th>
<th>PLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>P=0 (no lag)</td>
<td>714.4</td>
<td>710.8</td>
<td>3.58</td>
</tr>
<tr>
<td>P=1</td>
<td>684.9*</td>
<td>677.9</td>
<td>3.435</td>
</tr>
<tr>
<td>P=2</td>
<td>689.2</td>
<td>678.6</td>
<td>3.432*</td>
</tr>
<tr>
<td>P=3</td>
<td>690.2</td>
<td>676.1*</td>
<td>3.47</td>
</tr>
<tr>
<td>P=4</td>
<td>694.4</td>
<td>676.8</td>
<td>3.53</td>
</tr>
<tr>
<td>P=5</td>
<td>698.8</td>
<td>677.7</td>
<td>3.52</td>
</tr>
</tbody>
</table>