# Prediction and model performance

Susan Thomas
IGIDR, Bombay

28 November, 2008

# Measures of model performance

- Test statistics with logL or SSE of restricted vs. unrestricted models: *LR* test, $R^2$.
- Forecast performance: in–sample vs. out-of-sample

# Test statistics using logL / SSE

# How to measure the performance of a model?

- Upto now, we have measured “model performance” using the objective function of the optimisation.
  For *MLE:* the likelihood function evaluated at $\hat{\beta}$.
  For *OLS:* the value of the Sum of Squared Errors evaluated at $\hat{\beta}$.
- Typically, these measures are used to compare the performance of alternative models.

# Tests and their critical values

- The standard test in MLE for model comparison is: LR test.
  - Test statistic: LR = $-2log(L_R/L)$, $L_R$ is the likelihood of the "restricted" model.
  - This has a $\chi^2(m)$ distribution where $m$ is the number of restrictions.
- The standard measure in OLS for model comparison is: $R^2$
  - Test statistic: LR = $(RSS_R - RSS_U/m)(RSS_U/N - K)$. $RSS_R$ is the sum of squared residual errors of the "restricted" model.
  - This has a $F(m, N - K)$ distribution where $m$ is the number of restrictions and $K$ is the total number of parameters estimated in the unrestricted model.

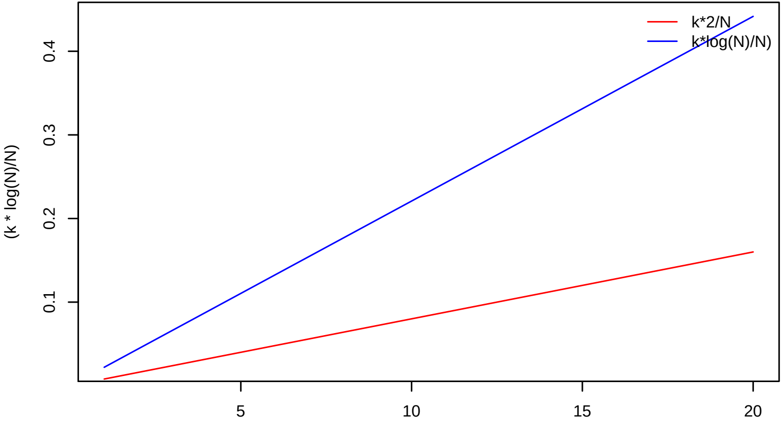
# Adjusting logL measures to accomodate *parsimony*

- MLE: Akaike Information Criteria (AIC), Schwartz-Bayes Information Criteria (SBC)

$$\begin{aligned} \mathrm{AIC(k)} &= logL + \frac{2K}{N} \\ \mathrm{SBC(k)} &= logL + \frac{K \log N}{N} \end{aligned}$$

Accept a model whose AIC(k)/SBC(k) is larger.
Note: sometimes AIC/SBC can give contradictory results.
Choose the more conservative one. Typically SBC.

# Behaviour of $2 * k/N$ vs $k \log N/N$



# Adjusting SSE measures to accomodate *parsimony*

- OLS: Adjusted $R^2$.

$$\bar{R}^2 = 1 - \frac{(N-1)}{(N-K)}(1 - R^2)$$

  Accept a model whose $\bar{R}^2$ is larger.
- The AIC equivalent in OLS is:

$$\mathrm{AIC}(k) = s_y^2(1 - R^2)e^{2k/N}$$

- The SBC equivalent in OLS is:

$$\mathrm{SBC}(k) = s_y^2(1 - R^2)n^{K/N}$$

# Nested vs. non-nested models

- In all the measures above, one constraint is that the "restricted" model is an explicit subset of the unrestricted model.
- However, theory can favour a choice of two linear models, $M_1, M_2$, such that

$$H_0 : M_1, y = X\beta + \epsilon; H_A : M_2, y = Z\gamma + \eta$$

- One approach to compare two non-nested model performance: make a supermodel, which is the sum of both.

$$M_s, y = X\beta + Z\gamma + u$$

  This is called the encompassing approach.

# Nested vs. non-nested models

- If $H_0 : M_1, y = X\beta + \epsilon; H_A : M_2, y = Z\gamma + \eta$
- An encompassing model:

$$Y = X'\beta + Z'\delta + u$$

  where $Z'$ are the variables in $M_2$ which are not in $M_1$.
- Test of $H_0$ is to estimate the model and test if $\gamma = 0$. The critical value is set the F-distribution.
- Two popularly used tests: *J*-test (Davidson-Mackinnon) and the *Cox* test.

# Model selection based on prediction/Forecasting

# Model: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

- What is $E(Y|X_i)$ when $X_i = X_0$?
- We saw earlier that

$$\begin{aligned} E(Y_0|X_0) &= \beta_0 + \beta_1 X_0 \text{ when we know the true parameters} \\ E(\hat{Y}_0|X_0) &= \hat{\beta}_0 + \hat{\beta}_1 X_0 \text{ when we don't} \end{aligned}$$

- We report $E(\hat{Y}_0|X_0)$ with a 95% CI. The CI is determined from the variance of $E(\hat{Y}_0)$.

$$\begin{aligned} \text{var}(Y_0|X_0) &= \text{E}(Y_0|X_0 - E(Y_0|X_0))^2 \\ &= \text{E}(\epsilon_0) = \sigma^2_\epsilon \text{ when we know } (\beta_0, \beta_1) \text{ with certainty} \\ \text{var}(\hat{Y}_0|X_0) &= \hat{\sigma}^2_\epsilon \left[ \frac{1}{N} + \frac{(X_0 - \bar{X})^2}{S_{XX}} \right] \text{ when we do not} \end{aligned}$$

# An alternative for var($\hat{Y}_0|X_0$)

- If $\hat{Y}_0|X_0 = \hat{\beta}_0 + \hat{\beta}_1 X_0 + \hat{\epsilon}_0$, then
- var($\hat{Y}_0|X_0$) can also be written as:

$$\mathrm{var}(\hat{Y}_0|X_0) = \mathrm{E}(\hat{Y}_0|X_0 - \bar{Y}_0|X_0)^2 = \mathrm{E}(\hat{\epsilon}_0)^2 = \mathrm{var}(\hat{\epsilon}_0)$$

- var($\hat{\epsilon}_0$) is:

$$\begin{aligned}
\mathrm{var}(\hat{\epsilon}_0) &= \mathrm{var}(\hat{Y}_0 - Y_0) \\
&= \mathrm{var}(\hat{Y}_0) + \mathrm{var}(Y_0) - 2cov(\hat{Y}_0, Y_0) \\
\mathrm{var}(\hat{\epsilon}_0) &= \hat{\sigma}^2 \left[1 + \frac{1}{N} + \frac{(X_0 - \bar{X})^2}{S_{XX}}\right] > \mathrm{var}(\hat{Y}_0)
\end{aligned}$$

- In forecasting ($\hat{Y}_0|X_0$), use $\sigma(\hat{\epsilon}_0)$ which results in a fatter CI, than $\sigma(\hat{Y}_0)$.

# Predicting $Y_0|X_0$ for a multiple regression model

- Model for investment (1967-1982):
  $\text{real}\,Inv_t = \beta_0 + \beta_1 t + \beta_2 \text{real}\,Y_t + \beta_3 i_t + \beta_4 inf_t + u_t$

| Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|
| (Intercept) | -0.5091 | 0.0551 | -9.234 |
| $t$ (t=1 in 1967) | -0.0165 | 0.0019 | -8.409 |
| real $Y_t$ (in trillions) | 0.6704 | 0.0549 | 12.189 |
| $i_t$ (in %) | -0.0023 | 0.0012 | -1.908 |
| $inf_t$ (in %) | -0.00009 | 0.0013 | -0.070 |
| Sum of squared residuals = 0.0004507 | | Number of obs. = 15 | |
| Std. Err of residuals = 0.006703 | | t(10), 5% = 2.228 | |

Estimated var-cov of estimates

| | Intercept | $t$ | $Y_t$ | $i_t$ | $inf_t$ |
|---|---|---|---|---|---|
| Intercept | 0.00303 | | | | |
| $t$ | 0.0001 | 3.88e-6 | | | |
| $Y_t$ | -0.0030 | -0.0010 | 0.0030 | | |
| $i_t$ | 5.59e-6 | 2.29e-7 | 7.78e-6 | 1.49e-6 | |
| $inf_t$ | 3.21e-6 | 4.27e-8 | 2.28e-6 | 7.51e-7 | 1.82e-6 |

- What is
  $E(Inv_{1983}|t = 16, Y_{16} = 1.5 \text{ trillion}, i_{16} = 10\%, inf_{16} = 4\%?$

# Predicting $Y_0|X_0$ for a multiple regression model

- $(\hat{\beta})'$ = (-0.509, -0.017, 0.670, -0.002, -0.0001)
- $(X_0)$ = (1, 16, 1.5, 10, 4)
- $\hat{Inf_t} = \hat{Y}_0|X_0 = X_0\hat{\beta} = 0.2036$
- $\hat{v}ar(\hat{Inf_t}) = \hat{v}ar(\hat{\epsilon}_0)^2 = \sigma^2 + X_0' \left[(\sigma^2(X'X)^{-1}\right] X_0$
- $\hat{v}ar(\hat{\epsilon}_0)^2 = 0.00009772$
- 95% CI for $\hat{Inf_t}$ =

$$0.2036 \pm 2.228(0.009885) = (0.1811, 0.2262)$$

- If this comes out as comparable to the actual data in 1983, then the model works – forecasts as a tool for *model performance*.

# Using prediction in model selection

- Typically, the estimation is applied on the whole set of data. Under this situation, the measure of model performance becomes the SSE of the whole data set.
- "In sample prediction" is how well the estimated model fits the data used in the estimation itself.
  Typically, printed as the root-mean-squared-error (RMSE) of the model: $\sqrt{\sum_i \hat{\epsilon}_i^2 / N}$
- "Out of sample prediction" is how well the estimated model fits the data that has *not* been included in the estimation. Here the measure is the same; however, the data is not.

# Calculating "out-of-sample" RMSE

- The procedure to calculate the "out of sample prediction".
  1. Partition the dataset (size $N$) into two: $N_1$, $N_2$. Typically, $N_1 >> N_2$.
  2. Estimate the model using $N_1$ "in-sample" observations.
  3. For model $M_1$, calculate $RMSE_{M_1}$, using $N_2$ "out-of-sample" observations.
- This can be replicated for all competing models. The same partitioned data should be used:
  1. $N_1$ to estimate alternative models, $M_2$, $M_3$
  2. Using the estimated coefficients to make predictions for $N_2$ oservations to calculate "out-of-sample" RMSE for each model: $RMSE_{M_2}, RMSE_{M_3}, \ldots.$
- The model that has the "smallest" out-of-sample RMSE is considered the best.

# Model selection based on “out-of-sample” RMSE

- This approach can be used to select across all manner of different models.
- Care has to be taken on the partitioning of data:
    1. For cross-sectional data: the partitioned datasets have to be random.
    2. This is not a choice for time-series data.
       Time series data depends upon simulation methods when using prediction as an alternative to traditional methods. “MonteCarlo”, “bootstrap”, “block-bootstrap”.