Regression III

Model Testing and Discrimination

Dave Armstrong

Goals for Today

1. Discuss Information Theory as it relates to model selection and discrimination.
   - Identify the difference between AIC and BIC.
2. Describe Clarke's distribution free test for non-nested models.
3. Consider model selection uncertainty and methods for controlling it - Model averaging.

What do we Mean by 'Model Selection'

- Testing competing models against each other (i.e., relative fit).
  - Nested model tests
  - Non-nested model tests
- Multi-model inference
- How to deal with model selection uncertainty in a principled way.

Notes

Type notes here...

Options for Comparative Model Fit

- Direct tests of nested models - F (ANOVA), χ^2 (Analysis of Deviance, LR-Test)
- Information Criteria measures (e.g., AIC and BIC)
- Tests for Non-nested Models (e.g., Clarke and Vuong)

Notes

Type notes here...

Nested Model Tests

Tests like the LR test and F-test require nested models because,

- They are considering the different between two statistics (RSS or LR)
- This difference follows an F or χ^2 distribution under the null (neither distribution permits negative values).
- So, the model with more parameters *must* provide a fit not worse than the model with fewer parameters.
- The only way to ensure this is the case is to ensure that the models are nested

Notes

Type notes here...

Likelihood Ratio Test

The LR Test uses the statistic defined by the difference in the log-likelihoods of the models.

$$LR = -2\left(ll_{\text{restricted}} - ll_{\text{unrestricted}}\right) \sim \chi^2_{p-q}$$

where there are p parameters in the unrestricted model and q parameters in the restricted model.

- The distribution is asymptotically right, but will not be exactly χ^2 in finite samples.
- Deviance is often taken as $-2ll_{\text{model}}$, though this is not always the case (take, for example, the linear model case).

Notes

Type notes here...

Information Theory

Information theorists believe in reality, but not in the notion of "true" models.

- Models are necessarily simplified constructions that try to approximate reality.
- There is more information in large datasets than small.
- Information amounts to the ability to identify interesting, though substantively small effects

Notes

Type notes here...

Principles for Model-based Inference

Parsimony

- Encapsulates the bias-variance tradeoff.

Multiple Working Hypotheses

- There is no single null hypothesis against which an alternative is to be tested.
- rather, there is a (small-ish) set, well-specified and theoretically derived working hypotheses.

Strength of Evidence

- We must be able to quantify the "strength of evidence" supporting various working hypotheses if science is to progress in the usual way.

Notes

Type notes here...

K-L Information

Kullback and Leibler (1951) quantified the meaning of "information".

$$I(f,g) = \int f(x) log \left(\frac{f(x)}{g(x|\theta)} \right) dx$$

where:

- f denotes a fixed (i.e., constant) reality (reality is non-parametric [i.e., it has no parameters])
- g is a model approximating f with parameters θ.
- $I(f,g)$ is the information lost when using g to approximate f.

Notes

Type notes here...

Expected Information

We cannot use $I(f, g)$ in model selection because it requires knowledge of f and θ (the parameters in g.) Instead, consider the *Expected Infromation*

$$\begin{aligned} E_f\left[I(f,g)\right] &= E_f\left[log(f(x))\right] - E_f\left[log(g(x|\theta))\right] \\ &= C - E_f\left[log(g(x|\theta))\right] \end{aligned}$$

If we wanted to compare model $g(x|\theta)$ with model $m(x|\gamma)$, we could calculate:

$$\begin{aligned} E_f\left[I(f,g)\right] - E_f\left[I(f,m)\right] &= \big(E_f\left[log(f(x))\right] - E_f\left[log(g(x|\theta))\right]\big) \\ &\quad - \big(E_f\left[log(f(x))\right] - E_f\left[log(m(x|\gamma))\right]\big) \\ &= E_f\left[log(g(x|\theta))\right] - E_f\left[log(m(x|\gamma))\right] \\ &\approx E_f\left[log(g(x|\hat{\theta}))\right] - E_f\left[log(m(x|\hat{\gamma}))\right] \end{aligned}$$

Notes

Type notes here...

Expected Information in Model Comparisons

As described above, we could use expected information to compare models. Some things to note:

1. We usually describe a "candidate set" of models - the set of models over which the comparison is to be made.
2. There is no assumption that the "true" model is in the candidate set of models.
3. In fact, there is no assumption that a true model exists at all.
    - The guiding principle that reality is complex and nonparametric eschews the idea of a "true" model - all models are simplifications, abstractions and approximations and those none is in any sense "true".

Notes

Type notes here...

Akaike's Information Criterion (AIC)

The goal was to estimate: $E_y E_x \left[log(g(x|\hat{\theta}(y))) \right]$, essentially the relative information with θ replaced with the MLE estimates $\hat{\theta}$.

- Akaike found that $log(\mathcal{L}(\hat{\theta}|\text{data}))$ was a biased estimator of $E_y E_x \left[log(g(x|\hat{\theta}(y))) \right]$, but that asymptotically the bias is approximately equal to K, the number of parameters in $\hat{\theta}$. Thus,

$$log(\mathcal{L}(\hat{\theta}|\text{data})) - K \approx C - \hat{E}_{\hat{g}} \left[I(f, \hat{g}) \right]$$

K is not arbitrary, but chosen to minimize bias in the estimated expected information.

$$\begin{aligned} AIC &= -2(log(\mathcal{L}(\hat{\theta}|\text{data})) - K) \\ &= -2log(\mathcal{L}(\hat{\theta}|\text{data})) + 2K \end{aligned}$$

Notes

Type notes here...

Small-sample Correction

When K is large relative to n or for any value of K for small n, there is a correction to AIC.

$$AIC_c = -2log(\mathcal{L}(\hat{\theta}|\text{data})) + 2K + \frac{2K(K+1)}{n-K-1}$$

- This should be used probably always, but especially if $n/K \leq 40$ for the largest K in the model set.
- AIC_c converges to AIC as $n \to \infty$.

Notes

Type notes here...

Delta values

Often, for AIC_c or AIC to be interpretable, Δ_i should be calculated such that for each model i in the model set,

$$\Delta_i = AIC_i - AIC_{\min}$$

This gives the "best" model $\Delta_i = 0$

- This captures the information loss due to using model g_i rather than the best model, g_{min}.
- The large Δ_i, the less likely model i is the best approximation of reality f.

Conventional cut-off values for Δ_i are:

- $\Delta_i \leq 2$ indicates substantial support,
- $4 \leq \Delta_i \leq 7$ indicates less support,
- $\Delta_i \geq 10$ indicates essentially no support.

Notes

Type notes here...

BIC

The BIC is defined as:

$$BIC = -2\log(\mathcal{L}) + K\log(n)$$

- BIC is not technically based in "information theory" and as such is not an information criterion measure.
- The BIC is meant to approximate the Bayes Factor (or rather its log):

$$\frac{\Pr(D|M_1)}{\Pr(D|M_2)} = \frac{\int \Pr(\theta_1|M_1)\Pr(D|\theta_1, M_1)d\theta_1}{\int \Pr(\theta_2|M_2)\Pr(D|\theta_2, M_2)d\theta_2}$$

Models need not be nested and we need not appeal to the idea that there exists a "true" model, much less that the true model is in our set of candidate models.

Notes

Type notes here...

AIC or BIC

The question of whether to use AIC or BIC is often left to how much you want to penalize additional model parameters. In actuality, the question is one of performance in picking the K-L best model.

- When there are "tapering effects", AIC is better
- When reality is simple with a few big effects captured by the highest posterior probability models, then BIC is often better.

Notes

Type notes here...

Non-nested Model Tests

Both AIC and BIC work for non-nested models, but neither is a *test* per se (i.e., they don't have sampling distributions which can be evaluated to produce p-values). There are a set of tests for non-nested models that do have known sampling distributions. Consider the following set of models:

$$\begin{aligned} H_1 &: \mathbf{y} = \mathbf{X}\beta + \mathbf{u}_1, \quad E(\mathbf{u}_1'\mathbf{u}_1) = \sigma_1^2\mathbf{I} \\ H_2 &: \mathbf{y} = \mathbf{Z}\gamma + \mathbf{u}_2, \quad E(\mathbf{u}_2'\mathbf{u}_2) = \sigma_2^2\mathbf{I} \end{aligned}$$

where β and γ are vectors of length k_1 and k_2, respectively.

Models are nested if there exists a vector of values a such that $a\beta = \gamma$ or $a\gamma = \beta$ depending on which model is bigger. Otherwise, the models are non-nested.

- Usually we talk about models being nested when one contains a subset of variables in the other, but that is not the only way to define nesting.

Notes

Type notes here...

Distribution Free Test

Clarke (2003) puts forth a distribution-free test that is really a "paired sign test". The statistic is calculated as:

$$d_i = \log(\mathcal{L}_{\beta,x_i}) - \log(\mathcal{L}_{\gamma,z_i}) + (p-q)\left(\frac{log(n)}{2n}\right)$$

$$B = \sum_{i=1}^{n} I_{0,+\infty}(d_i)$$

- The d_i are the difference in individual log-likelihoods for the two models
- The second equation above counts up the number of positive d_i values.
- We are testing to see whether B is significantly bigger than a random binomial variable that has a $p = .5$ and n the same as the number of rows in $\mathbf{X}$ and $\mathbf{Z}$.

Notes

Type notes here...

Examples in R

You can produce AIC, AICc and BIC in the following ways:

```r
library(car)
library(AICcmodavg)
data(Prestige)
mod1 <- lm(prestige ~ income + women,
  data=na.omit(Prestige), y=T)
mod2 <- lm(prestige ~ education + type + women,
  data=na.omit(Prestige), y=T)
AIC(mod1)
```

```
## [1] 763.8879
```

```r
c(AICc(mod1), AICc(mod2))
```

```
## [1] 764.3180 685.4286
```

```r
c(AIC(mod1), AIC(mod2))
```

```
## [1] 763.8879 684.5055
```

```r
c(BIC(mod1), BIC(mod2))
```

Notes

Type notes here...

Clarke Tests in R

```r
library(clarkeTest)
clarke_test(mod1, mod2)
```

```
##
## Clarke test for non-nested models
##
## Model 1 log-likelihood: -378
## Model 2 log-likelihood: -336
## Observations: 98
## Test statistic: 24 (24%)
##
## Model 2 is preferred (p = 4.2e-07)
```

Notes

Type notes here...

Model Selection Uncertainty

Generally we present what we think to be the single best model after a more or less extensive model search.

- Our estimates of sampling variability of parameters is often too small because we "forget" to include model selection uncertainty (the fact that we didn't know initially exactly the right model).
- This uncertainty captures the extent to which we are unsure about this model and have considered other alternatives.
- Others have proposed solutions to the problem (e.g., Leamer, with the idea of "Leamer Bound"), but we will consider an alternative - AIC/BIC weights and Model Averaging.

Notes

Type notes here...

Akaike Weights

We can construct Akaike weights in the following way:

$$w_i = \frac{exp\left(\frac{-\Delta_i}{2}\right)}{\sum_i exp\left(\frac{-\Delta_i}{2}\right)}$$

- $exp\left(\frac{-\Delta_i}{2}\right)$ is the likelihood of the model given the data.
- w_i gives (essentially) the probability that model i is the K-L best model approximation of f.

Notes

Type notes here...

AIC Weights (2)

These estimates can be used to give us measures of effects and sampling variance that are *unconditional* on the model selected.

Average Effect

$$\hat{\bar{\theta}} = \sum_i w_i \hat{\theta}_i$$

Sampling Variability

$$\widehat{var}\left(\hat{\bar{\theta}}\right) = \left[\sum_i w_i \left[\widehat{var}(\hat{\theta}_i | g_i) + \left(\hat{\theta}_i - \hat{\bar{\theta}}\right)^2\right]^{\frac{1}{2}}\right]^2$$

Notes

Type notes here...

Example

Let's think about the Ericksen Dataset and estimating different models. One method would be to specify the models we wanted directly.

```r
library(MuMIn)
mods <- list()
mods[[1]] <- lm(undercount ~ crime + highschool, data=Ericksen)
mods[[2]] <- lm(undercount ~ poverty + language, data=Ericksen)
mods[[3]] <- lm(undercount ~ housing + crime, data=Ericksen)
modavg <- model.avg(mods)
```

Notes

Type notes here...

Multi-model Sampling Variance in R

```r
sma <- summary(modavg)
printCoefmat(sma$coefmat.full, digits=3)
```

```
##             Estimate Std. Error Adjusted SE z value Pr(>|z|)
## (Intercept) -2.78927    1.05254     1.06876    2.61   0.0091 **
## crime        0.06561    0.01264     0.01279    5.13    3e-07 ***
## highschool   0.01767    0.02551     0.02585    0.68   0.4942
## housing     -0.00434    0.01795     0.01827    0.24   0.8124
## poverty      0.00247    0.02195     0.02199    0.11   0.9106
## language     0.00658    0.05719     0.05724    0.11   0.9084
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The `coefmat.full` element averages over all models including zeros for the coefficient when the variable was excluded.

```r
printCoefmat(sma$coefmat.subset, digits=3)
```

```
##             Estimate Std. Error Adjusted SE z value Pr(>|z|)
## (Intercept)  -2.7893     1.0525      1.0688    2.61   0.0091 **
## crime         0.0665     0.0101      0.0103    6.47  < 2e-16 ***
## highschool    0.0289     0.0272      0.0277    1.04   0.2981
## housing      -0.0116     0.0279      0.0284    0.41   0.6834
## poverty       0.1809     0.0551      0.0562    3.22   0.0013 **
## language      0.4825     0.1006      0.1026    4.70  2.6e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The `coefmat.subset` element averages across only those models that included the variable of interest, disregarding those where the variable was not included.

Notes

Type notes here...

Option 2

Another option would be to let the function sort out which variables ought to be in there:

```r
library(leaps)
library(MuMIn)
E <- na.omit(Ericksen)
fm1 <- lm(undercount ~ ., data=E,
    na.action="na.fail")
out <- dredge(fm1)
ma <- model.avg(out)
```

The `dredge()` function estimates all 2^k models and `ma()` averages over them based on their AIC weights.

Notes

Type notes here...

Model Selection Summary

```r
sma <- summary(ma, subset=delta < 4)
```

```r
printCoefmat(sma$coefmat.full, digits=3)
```

```
##              Estimate Std. Error Adjusted SE z value Pr(>|z|)
## (Intercept)  -0.30929    1.78497     1.80078    0.17  0.86363
## conventional  0.02940    0.00904     0.00919    3.20  0.00138 **
## crime         0.02111    0.01527     0.01541    1.37  0.17081
## language      0.16050    0.11549     0.11662    1.38  0.16876
## minority      0.09130    0.02470     0.02503    3.65  0.00026 ***
## poverty      -0.08804    0.08853     0.08930    0.99  0.32422
## citystate    -0.52924    0.78402     0.79087    0.67  0.50337
## highschool    0.00461    0.03027     0.03059    0.15  0.88034
## housing      -0.00621    0.01696     0.01717    0.36  0.71747
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
printCoefmat(sma$coefmat.subset, digits=3)
```

```
##              Estimate Std. Error Adjusted SE z value Pr(>|z|)
## (Intercept)  -0.30929    1.78497     1.80078    0.17  0.86363
## conventional  0.02965    0.00867     0.00883    3.36  0.00078 ***
## crime         0.02662    0.01214     0.01236    2.15  0.03123 *
## language      0.20199    0.09168     0.09347    2.16  0.03069 *
## minority      0.09139    0.02454     0.02487    3.68  0.00024 ***
## poverty      -0.13473    0.07552     0.07691    1.75  0.07980 .
## citystate    -1.13177    0.79535     0.80971    1.40  0.16219
## highschool    0.01537    0.05377     0.05439    0.28  0.77748
## housing      -0.02045    0.02561     0.02606    0.78  0.43260
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Notes

Type notes here...

More Multi-model Inference

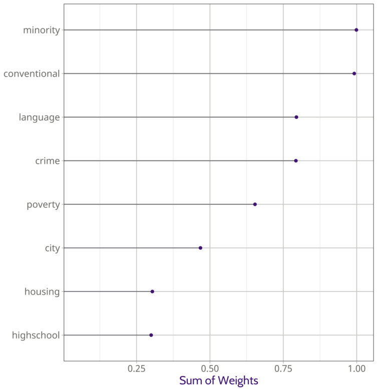
We could also think about two other uses of Akaike weights:

- We could actually use $\hat{\hat{\theta}}$ as our point estimate that will be less biased due to model selection.
- Summing w_i across all of the models including variable j can give a sense of how "important" variables are.

Notes

Type notes here...

Important Variables in R



Notes

Type notes here...

Multi-model Averaging Cautions

- Difficult if candidate models have different functional forms for the same variable (e.g., additive in one model and conditional in another.)
- Only really takes care of the model if all of the models you ever want to estimate are in the candidate set.
- Won't probably do what you want if you've got different ways to operationalize a single concept.

Notes

Type notes here...