

Choosing and Defending Specifications: Comparative Model Testing

Kevin A. Clarke
University of Rochester

`kevin.clarke@rochester.edu`

`http://www.rochester.edu/College/PSC/clarke/`

Overview

- Why is comparative model testing necessary?
 - The role of hypothetico-deductivism in political research.
- Traditional approaches: When they do and do not work.
 - Nested v. nonnested models.
- Model selection criteria.
 - AIC and BIC.
- Model selection tests.
 - The Vuong and distribution-free tests.

Based on the following....

- “The Necessity of Being Comparative”
- “Testing Nonnested Models of International Relations”
- “Nonparametric Model Discrimination in International Relations”
- “A Simple Distribution-Free Test for Nonnested Hypotheses”

<http://www.rochester.edu/College/PSC/clarke/>

Why comparative model testing?

Comparative model testing is necessary because political scientists (and other social scientists) have combined probabilistic falsificationism with a hypothetico-deductive approach to research.

Let's spell out each in turn....

First some notation

Symbol	Meaning	Usage
T	theory	
\wedge	and (conjunction)	
K	background conditions	$T \wedge K$
H_0	null hypothesis	
H_1	alternative hypothesis	
\neg	not	$\neg H_0$
\rightarrow	if...then	$A \rightarrow B$
\leftrightarrow	if and only if	$A \leftrightarrow B$

Falsificationism

Classical or frequentist hypothesis testing is based on a probabilistic version of Popperian falsificationism.

$$H_0 \rightarrow \beta = 0$$

$$\beta \neq 0$$

— — — — —

$$\neg H_0$$

where $\beta \neq 0$ means a p-value less than the significance level.

Hypothetico-deductivism

H-D simply means deriving a prediction from a theory and background conditions and then testing the prediction.

A qualitative example from Huth, Gelpi, and Bennett (1993):

Structural realism

Risk-acceptant leaders

Multipolarity

In a multipolar system, risk-acceptant leaders will be more likely to escalate

Now let's combine statistical tests with H-D

Step 1: Theory to hypothesis

$$T \wedge K \rightarrow \neg H_0$$

Step 2: Data to hypothesis

$$H_0 \rightarrow \beta = 0$$

$$\beta \neq 0$$

$$\neg H_0$$

Step 3: Hypothesis to theory

$$T \wedge K \rightarrow \neg H_0$$

$$\neg H_0$$

$$T \wedge K$$

A problem...

The last step is a logical fallacy called “affirming the consequent.”

Step 3: Hypothesis to theory

$$\begin{array}{r} T \wedge K \rightarrow \neg H_0 \\ \neg H_0 \\ \hline T \wedge K \end{array}$$

Another example: father \rightarrow male

Objection 1: We do something else in practice....

Step 1: Theory to hypothesis

$$T \wedge K \rightarrow H_1$$

Step 2: Data to hypothesis

$$H_1 \rightarrow \text{Coefficient is correct}$$

Coefficient is correct

$$H_1$$

Step 3: Hypothesis to theory

$$T \wedge K \rightarrow H_1$$

$$H_1$$

$$T \wedge K$$

Objection 2: Being Bayesian solves the problem

Step 1: Theory to hypothesis

$$T \wedge K \rightarrow H_1$$

Step 2: Data to hypothesis

$$H_1 \rightarrow P(H_1|y) \text{ is high}$$

$$P(H_1|y) \text{ is high}$$

$$H_1$$

Step 3: Hypothesis to theory

$$T \wedge K \rightarrow H_1$$

$$H_1$$

$$T \wedge K$$

The biconditional to save the day!

If the alternative hypothesis is true if and only if the theory is true, and the hypothesis is true if and only if the data comes out right, then then the theory is true if and only if the data come out right.

$$T \wedge K \leftrightarrow H_1$$

$$H_1 \leftrightarrow D$$

— — — — —

$$T \wedge K \leftrightarrow D$$

How does this help us?

The biconditional in action

Step 1: Theory to hypothesis

$$(T_1 \wedge K) \equiv (T_2 \wedge K) \leftrightarrow H_0$$

Step 2: Data to hypothesis

$$H_0 \leftrightarrow \tau \approx 0$$

$$\tau > 0$$

$$\neg H_0$$

Step 3: Hypothesis to theory

$$(T_1 \wedge K) \equiv (T_2 \wedge K) \leftrightarrow H_0$$

$$\neg H_0$$

$$(T_1 \wedge K) > (T_2 \wedge K).$$

Why comparative model testing?

Being able to make “if and only if” statements is a necessary condition for learning. As our theories provide little in the way of such statements, we need to be comparative.

Now let's look at how traditional approaches handle comparative model testing. There are two cases: nested and nonnested.

Definition: Nested models

Two models are *nested* if one model can be reduced to the other model by imposing a set of linear restrictions on the parameter vector.

Consider, for example, these two models,

$$\text{Model 1: } Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon_1$$

$$\text{Model 2: } Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon_2$$

By setting $\beta_3 = \beta_4 = 0$ in model 2, we get model 1.

Nested models and functional form

Models may also be nested in terms of their functional forms. Consider two common duration models:

Distribution	Hazard Function, $\lambda(t)$	Survival Function, $S(t)$
Exponential	λ	$S(t) = e^{-\lambda t}$
Weibull	$\lambda p(\lambda t)^{p-1}$	$S(t) = e^{-(\lambda t)^p}$

The Weibull is the Exponential when $p = 1$.

Tests for discriminating between nested models

Discriminating between nested models is easily accomplished using a variety of standard statistical techniques, such as,

- Z -tests (for a single restriction in OLS or GLM),
- Likelihood ratio tests (for a single or multiple restrictions in GLM),
- F -tests (for a single or multiple restrictions in OLS).

Example: Do nukes make a difference?

Dependent variable: escalation or not (Huth, Gelpi, and Bennett (1993))

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.3739	0.5334	-2.58	0.0100
Balance of forces	1.4628	0.7691	1.90	0.0572
Def. vital interests	-0.3653	0.2995	-1.22	0.2225
Chall. vital interests	0.6136	0.3125	1.96	0.0496
Def. backed down	0.8359	0.3571	2.34	0.0192
Chall. backed down	-0.9565	0.4504	-2.12	0.0337
Def. other dispute	0.7511	0.3063	2.45	0.0142
Chall. other dispute	-0.1457	0.3029	-0.48	0.6304

The log-likelihood is -56.99703 (df=8).

Example: Do nukes make a difference?

Dependent variable: escalation or not (Huth, Gelpi, and Bennett (1993))

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.8857	0.5913	-1.50	0.1341
Balance of forces	1.4650	0.8158	1.80	0.0725
Secure 2nd strike	-1.7609	0.4132	-4.26	0.0000
Def. vital interests	-0.8871	0.3702	-2.40	0.0166
Chall. vital interests	0.8293	0.3651	2.27	0.0231
Def. backed down	1.0274	0.4153	2.47	0.0134
Chall. backed down	-0.7090	0.4911	-1.44	0.1488
Def. other dispute	0.7257	0.3496	2.08	0.0379
Chall. other dispute	-0.0173	0.3430	-0.05	0.9598

The log-likelihood is -45.55808 (df=9).

Example: R commands

```
huth1 <- glm(outcome dispbof+defint+chint+riwhimp+chwhimp
             +riothdis+chothdis,family=binomial(link=probit))
huth2 <- glm(outcome dispbof+rinukes+defint+chint+riwhimp
             +chwhimp+riothdis+chothdis,family=binomial(link=probit))
addterm(huth1,huth2,test="Chisq")
```

	Df	LRT	Pr(Chi)
rinukes	1	22.88	0.00

$$\{\text{LRT} = -2 * [-56.99703 - (-45.55808)] = 22.8779\}$$

Definition: Nonnested models

Two models are *nonnested* if one model cannot be reduced to the other model by imposing a set of linear restrictions on the parameter vector.

$$\text{Model 1: } Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon_1$$

$$\text{Model 2: } Y = \beta_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon_2$$

There is no set of linear restrictions that we can impose on either model 1 or model 2 that will give us the other model.

Nonnested models and functional form

Two models may also be nonnested in terms of their functional forms.

$$\Pr(Y = 1) = \Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2) = \int_{-\infty}^{\beta_0 + \beta_1 X_1 + \beta_2 X_2} \phi(t) dt$$

$$\Pr(Y = 1) = \Lambda(\beta_0 + \beta_1 X_1 + \beta_2 X_2) = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2}}$$

Even when they have the same covariates, probits and logits are nonnested.

The traditional approach: “Super” models

- Model 1 includes X_1 , X_2 , and X_3 .
- Model 2 includes X_3 , X_4 , and X_5 .

The combined “super” model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon$$

Will any of the traditional approaches taken to nested models work here?

The traditional approaches and “super” models

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon$$

- Z -tests?
- Likelihood ratio tests?
- F -tests?

More about the f-test and nonnested models

$$\text{Model 1 : } \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}_1$$

$$\text{Model 2 : } \mathbf{Y} = \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}_1$$

Let $\tilde{\mathbf{X}}$ be the variables in \mathbf{X} , but not in \mathbf{Z} .

Let $\tilde{\mathbf{Z}}$ be the variables in \mathbf{Z} , but not in \mathbf{X} .

Let \mathbf{W} be the variables the two models have in common.

$$\text{Combined : } \mathbf{Y} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \tilde{\mathbf{Z}}\boldsymbol{\gamma} + \mathbf{W}\boldsymbol{\sigma} + \boldsymbol{\epsilon}$$

More about the f-test and nonnested models

$$\text{Combined : } \mathbf{Y} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \tilde{\mathbf{Z}}\boldsymbol{\gamma} + \mathbf{W}\boldsymbol{\sigma} + \boldsymbol{\epsilon}$$

- Testing $\boldsymbol{\beta}$ or $\boldsymbol{\gamma}$ does not test the full models.
- Testing $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}$ or $\boldsymbol{\gamma}$ and $\boldsymbol{\sigma}$ does not test model 1 against model 2.
- The f-test discriminates between either model 1 or model 2 and a hybrid model that is neither model 1 or model 2.
- The model, itself, is atheoretic.

Traditional approaches

So traditional approaches to model discrimination do not work for the nonnested case.

While there are numerous techniques we might use, let's consider model selection criteria and model selection tests.

Model selection criteria

AIC (Akaike's information criteria)

SIC (Schwarz's Bayesian information criteria)

Model selection tests

Vuong test and my distribution-free test

Kullback-Leibler discrepancy

The criteria and tests we will look at are all based upon this quantity, which is similar to a measure of “distance.”

$$I(g : f) = E_Y \ln \left\{ \frac{g(Y)}{f(Y)} \right\} = \int_{-\infty}^{\infty} \ln \left\{ \frac{g(Y)}{f(Y)} \right\} g(Y) dY$$

$$I(g : f) > 0 \quad \text{for } f \neq g$$

$$I(g : f) = 0 \quad \text{for } f = g$$

More on the Kullback-Leibler discrepancy

$$\begin{aligned} I(g : f) &= E_Y \ln \left\{ \frac{g(Y)}{f(Y|\theta)} \right\} \\ &= E_Y \ln g(Y) - E_Y \ln f(Y|\theta) \end{aligned}$$

If $g(Y)$ is the true sampling distribution of Y and $f(Y|\theta)$ is a particular model, we want an $f(Y|\theta)$ that is as large as possible to minimize the KL discrepancy. That is, we want a model that is as close to the true model as possible.

Model selection criteria

$$I(g : f) = E_Y \ln g(Y) - E_Y \ln f(Y)$$

Model selection criteria estimate $-E_Y \ln f(Y)$. The model for which this estimate is minimized is the “best” model.

AIC: $-2 \ln(\text{maximum likelihood}) + 2(\text{number of parameters})$

SIC: $-2 \ln(\text{maximum likelihood}) + \ln T(\text{number of parameters})$

Example: Realism v. rational deterrence

Huth, Gelpi, and Bennett (1993) test the relative explanatory power of structural realism and rational deterrence theory on the escalation of militarized disputes among great powers from 1816 to 1985.

Structural realism

Focuses on the attributes of the international system.

Rational deterrence

Focuses on the resolve and relative military capabilities of adversarial states.

Example: Results on structural realism

Dependent variable: escalation or not

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.5668	0.7777	-2.01	0.0439
System uncertainty 1	0.7772	0.1866	4.16	0.0000
System size*risk	-0.6900	0.2766	-2.49	0.0126
System uncertainty 2	-0.0244	0.1760	-0.14	0.8898
System diffusion*risk	0.1750	0.2589	0.68	0.4990
Risk-acceptant	0.8184	1.2580	0.65	0.5153

Example: Results on rational deterrence

Dependent variable: escalation or not

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.8857	0.5913	-1.50	0.1341
Balance of forces	1.4650	0.8158	1.80	0.0725
Secure 2nd strike	-1.7609	0.4132	-4.26	0.0000
Def. vital interests	-0.8871	0.3702	-2.40	0.0166
Chall. vital interests	0.8293	0.3651	2.27	0.0231
Def. backed down	1.0274	0.4153	2.47	0.0134
Chall. backed down	-0.7090	0.4911	-1.44	0.1488
Def. other dispute	0.7257	0.3496	2.08	0.0379
Chall. other dispute	-0.0173	0.3430	-0.05	0.9598

Example: Model comparison

```
huth.test <- mod.sel(huth1,huth2)
```

Selection Criteria

	Model One	Model Two
LogLik	-55.933	-45.558
AIC	123.865	109.116
BIC	139.313	132.289

Problems with model selection criteria

- Sampling properties are unknown.
- No statement of uncertainty.
- No ability to choose neither model.

Model selection tests

Model selection tests are also based on the Kullback-Leibler discrepancy. Written a little differently to emphasize the covariates:

$$\text{KLIC} \equiv E_0[\ln h_0(Y_i|X_i)] - E_0[\ln f(Y_i|X_i; \beta_*)]$$

We want to minimize this “distance,” and we do that by choosing the model that maximizes $E_0[\ln f(Y_i|X_i; \beta_*)]$.

To compare two models, then, it is natural to look at the ratio of their likelihoods.

Log-likelihoods

Before going further, let's consider the log-likelihood that is reported every time you run a generalized linear model. That log-likelihood is the sum of the log-likelihoods for each individual observation.

	Log-likelihood
Obs. 1	-0.54498
Obs. 2	-0.55547
Obs. 3	-0.42883
⋮	⋮
Obs. 97	-0.18510
Total	-55.93255

Log-likelihoods

How do we get these individual log-likelihoods?

Probit:

$$\log\text{-Lik}_i = y_i * \log(\hat{p}) + (1 - y_i) * \log(1 - \hat{p})$$

Poisson:

$$\log\text{-Lik}_i = \beta' \mathbf{x}_i * y_i - \exp(\beta' \mathbf{x}_i) - \log(\Gamma(y_i + 1))$$

Log-likelihood ratios

From the individual log-likelihoods, we can get the individual log-likelihood ratios by simply differencing.

	Model 1	Model 2	Difference
Obs. 1	-0.54498	-0.50802	-0.03696
Obs. 2	-0.55547	-0.37072	-0.18476
Obs. 3	-0.42883	-0.38157	-0.04726
⋮	⋮	⋮	⋮
Obs. 97	-0.18510	-0.14135	-0.04375
Total	-55.93255	-45.55808	-10.37447

Log-likelihood ratios

If the two models are equivalent, the difference between their log-likelihoods should be zero. If one model is closer to the true specification, then the difference between their log-likelihoods will be significantly different from zero.

The difference between the two tests we're going to talk about lies in whether we consider the average or the median.

- Vuong: is the average individual log-likelihood ratio different from zero?
- Distribution-free: is the median individual log-likelihood ratio different from zero?

Vuong test

Vuong proves under general conditions that the expected value given in the null hypothesis can be consistently estimated by $(1/n)$ times the likelihood ratio statistic,

$$\frac{1}{n}LR_n(\hat{\beta}_n, \hat{\gamma}_n) \xrightarrow{a.s.} E_0 \left[\ln \frac{f(Y_i|X_i; \beta_*)}{g(Y_i|Z_i; \gamma_*)} \right]$$

After standardization, the likelihood ratio statistic is asymptotically normally distributed. So the actual test is,

$$\text{under } H_0 : \frac{LR_n(\hat{\beta}_n, \hat{\gamma}_n)}{(\sqrt{n})\hat{\omega}_n} \xrightarrow{D} N(0, 1)$$

Vuong test

The variance is calculated in the normal way — the sum of the squares minus the square of the sum.

$$\hat{\omega}_n^2 \equiv \frac{1}{n} \sum_{i=1}^n \left[\ln \frac{f(Y_i|X_i; \hat{\beta}_n)}{g(Y_i|Z_i; \hat{\gamma}_n)} \right]^2 - \left[\frac{1}{n} \sum_{i=1}^n \ln \frac{f(Y_i|X_i; \hat{\beta}_n)}{g(Y_i|Z_i; \hat{\gamma}_n)} \right]^2$$

So what does the Vuong test amount to?

A paired z -test on the individual log-likelihoods!

Vuong test

	Model 1	Model 2	Difference
Obs. 1	-0.54498	-0.50802	-0.03696
⋮	⋮	⋮	⋮
Obs. 97	-0.18510	-0.14135	-0.04375
Total	-55.93255	-45.55808	-10.37447

$$Z = \frac{\text{Observed difference} - 0}{\text{S.D. diff} / \sqrt{n}} = \frac{LR_n(\hat{\beta}_n, \hat{\gamma}_n) - 0}{(\sqrt{n})\hat{\omega}_n} \underset{D}{\approx} N(0, 1)$$

Vuong test: Adjustment

We can always make the log-likelihood larger by adding additional variables to our model. Just like with AIC and BIC, we need a penalty adjustment for a model with too many variables. For the Vuong test, this adjustment is:

$$L\tilde{R}_n(\hat{\beta}_n, \hat{\gamma}_n) \equiv LR_n(\hat{\beta}_n, \hat{\gamma}_n) - \left[\binom{p}{2} \ln n - \binom{q}{2} \ln n \right],$$

where p and q are the number of parameters estimated in the two models. For the Huth et al. case, $p = 6$ and $q = 9$.

Vuong test: Problem

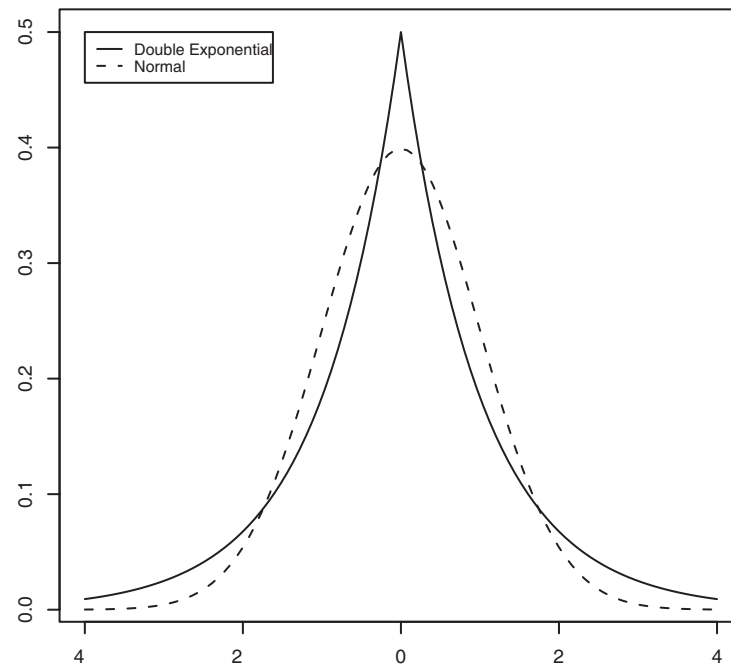
It can be shown that the Vuong test is very powerful if the underlying distribution of the individual log-likelihood ratios is normally distributed.

If the underlying distribution is not normal, then we can do better. If we measure the kurtosis of the distribution...

Distance	Sample Size				
	50	100	200	500	1000
0.3	5.19	5.44	5.49	5.50	5.49
0.4	5.26	5.52	5.58	5.57	5.59
0.5	5.24	5.49	5.62	5.68	5.73
0.6	5.32	5.62	5.81	5.88	5.90
0.7	5.40	5.70	5.89	6.09	6.14
0.8	5.47	5.87	6.10	6.34	6.38
0.9	5.58	5.98	6.35	6.58	6.70

Vuong test: Problem

...it looks much more like a double exponential or Laplace distribution.



Distribution-free test

If faced with data from a double-exponential or Laplace distribution, you would replace the normal z or t test with a distribution-free test such as the sign test, the paired sign test, or a signed-rank test.

The paired sign test is used to test the null hypothesis that the probability of a random variable from the population of paired differences being greater than zero is equal to the probability of the random variable being less than zero.

In our case,

$$H_0 : \Pr_0 \left[\ln \frac{f(Y_i|X_i; \beta_*)}{g(Y_i|Z_i; \gamma_*)} > 0 \right] = 0.5.$$

Distribution-free test

Letting $d_i = \ln f(Y_i|X_i; \hat{\beta}_n) - \ln g(Y_i|Z_i; \hat{\gamma}_n)$, the test statistic is

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

which is simply the number of positive differences, and it is distributed Binomial with parameters n and $\theta = 0.5$.

If model f is “better” than model g , B will be significantly larger than its expected value under the null hypothesis ($n/2$).

Distribution-free test: How to....

1. Run model f , saving the individual log-likelihoods.
2. Run model g , saving the individual log-likelihoods.
3. Compute the differences and count the number of positive and negative values.
4. The number of positive differences is distributed binomial($n, p = .5$).

Distribution-free test: Adjustment

As we are working with the individual log-likelihood ratios, we cannot apply the same correction to the “summed” log-likelihood ratio as Vuong did for his test.

We can, however, apply the *average* correction to the individual log-likelihood ratios. So we subtract the following factors from each individual log-likelihood.

$$\text{Model } f: [(p/2n) \ln n]$$

$$\text{Model } g: [(q/2n) \ln n]$$

Huth et al. (1993) data revisited

Selection criteria

	Model One	Model Two
LogLik	-55.933	-45.558
AIC	123.865	109.116
BIC	139.313	132.289

Model Section Tests*

	Statistic	P-value
Vuong	-0.781	0.435
Clarke	45.000	0.543

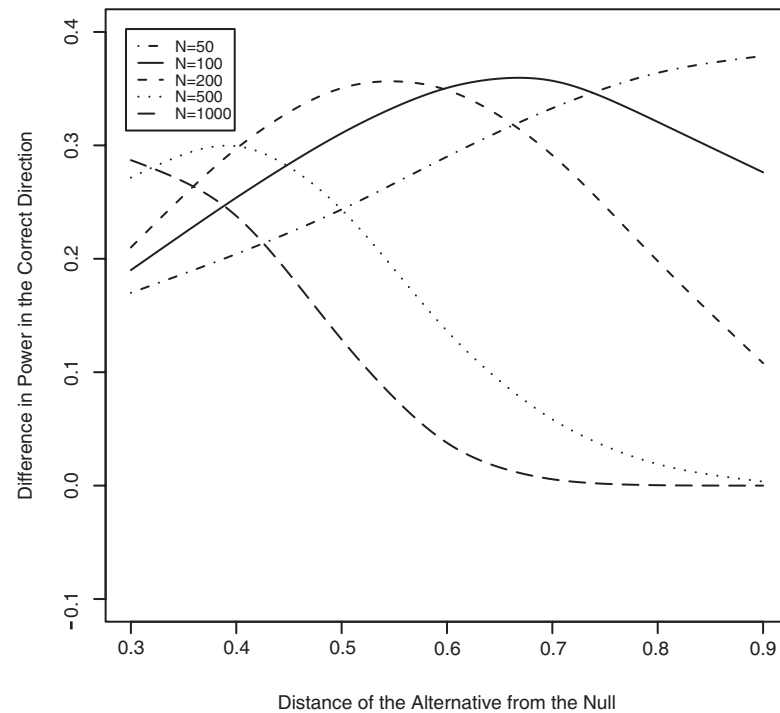
*The log-likelihoods for model two are subtracted from the log-likelihoods for model one.

R commands

```
huth1 <- glm(outcome~ nuncp1+runcp13+nuncp2+runcp23+risk23pm,  
             family=binomial(link=probit))  
  
huth2 <- glm(outcome~ dispbof+rinukes+defint+chint+riwhimp  
             +chwhimp+riothdis+chothdis,family=binomial(link=probit))  
  
huth.test <- mod.sel(huth1,huth2)  
  
summary(huth.test)
```

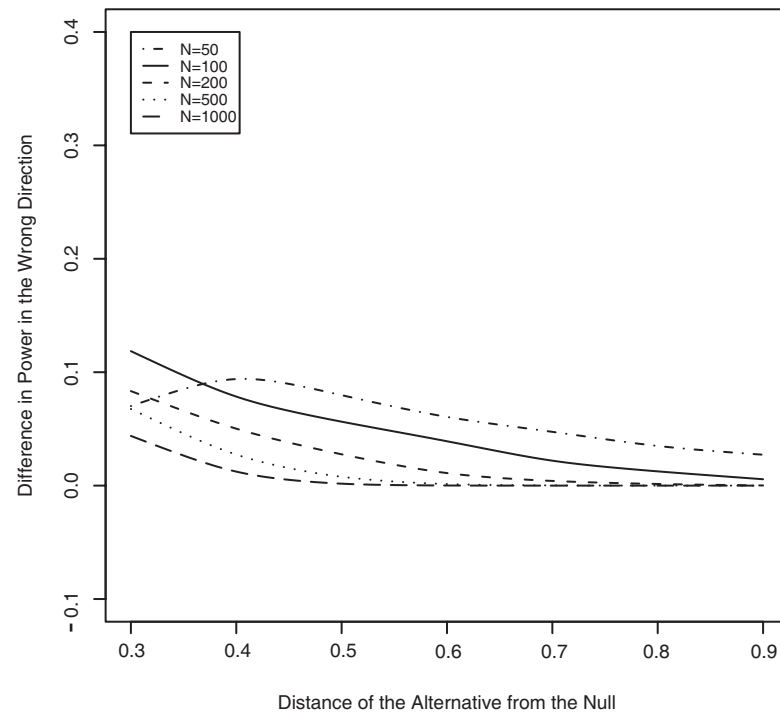
Test comparison: Getting it right

Difference in the probability of choosing the right model.



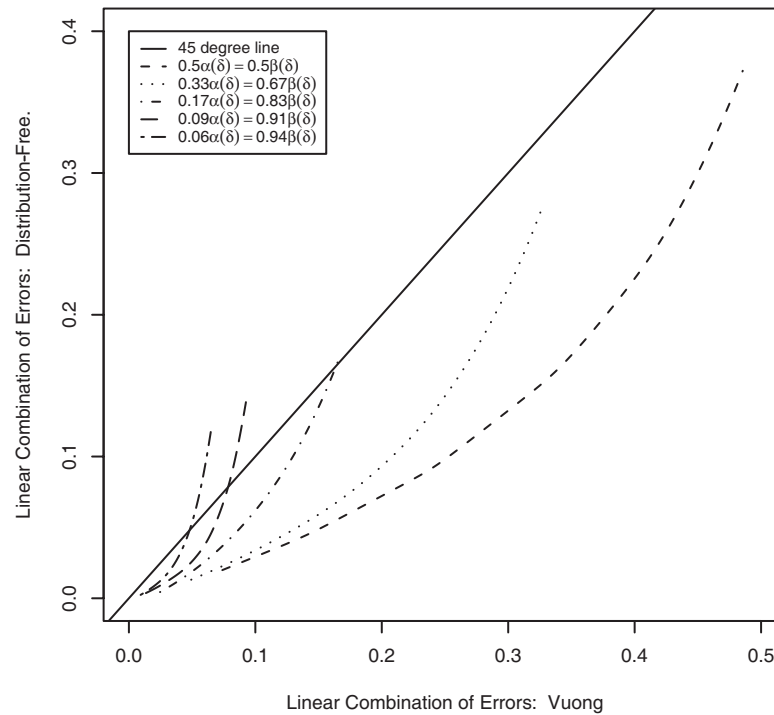
Test comparison: Getting it wrong

Difference in the probability of choosing the wrong model.



Test comparison: Balancing the good and bad

Linear combination of errors.



Take home message

- Being comparative is a necessary condition for making reliable inferences.
- Whether the rival models are nested or nonnested, there are techniques available for discriminating between them.
- The distribution-free test has greater power than the Vuong when the underlying distribution is not normal.