Advanced hypothesis testing

Gray Calhoun

November 13th, 2014, version 0.6.2

Brief list of topics

- Sequential hypothesis testing testing
- Hypothesis testing with nuisance parameters
- Combined use of these approaches

How to test a boatload of hypotheses

- We should know that testing a bunch of hypotheses at once is not reliable

How to test a boatload of hypotheses

Problems with naive seminar-style testing

- Other settings where this shows up:
  - Forecast evaluation
  - Portfolio selection
  - Every single empirical paper written in economics
- Obvious problem (suppose we have k different tests)

$$\begin{aligned}
&\Pr[\text{at least one test rejects a true null hypothesis}] \\
&\quad = 1 - \Pr[\text{no tests reject a true null}] \\
&\quad = 1 - \prod_{i=1}^{k} \Pr[\text{test } i \text{ does not reject; null } i \text{ is true}] \\
&\quad = 1 - \prod_{i=1}^{k} (1 - \Pr[\text{test } i \text{ rejects; null } i \text{ is true}]) \\
&\quad = 1 - (1-\alpha)^k
\end{aligned}$$

Bonferroni correction

- Obvious solution is the Bonferroni correction: test at α/k

$$
\begin{aligned}
&\Pr[\text{at least one test rejects a true hypothesis}] \\
&\quad = \Pr\left[\bigcup_{i=1}^{k} \{\omega : \text{test } i \text{ rejects a true null}\right] \\
&\quad \leq \sum_{i=1}^{k} \Pr[\{\omega : \text{test } i \text{ rejects}\};\ \text{null } i \text{ is true}] \\
&\quad \leq \sum_{i=1}^{k} \alpha/k \\
&\quad = \alpha
\end{aligned}
$$

- This may be conservative, since it assumes a worst-case dependence structure

Other corrections for multiple testing

- We can estimate the dependence structure between the tests (White, 2000)
- Suppose we have k asymptotically normal test statistics, $S_1, \ldots, S_k$, with

$$(S_1, \ldots, S_k) \to^d Z \sim N(\mu, \Sigma)$$

- Then $\max(|S_1|, \ldots, |S_k|) \to^d \max(|Z_1|, \ldots, |Z_k|)$

$$\begin{aligned}
&\Pr[\text{at least one test rejects a true null hypothesis}] \\
&\qquad = \Pr[\text{at least one } |Z_i| > c \text{ when } \mu_i = 0] \\
&\qquad = \Pr[\max_{i:\mu_i=0} |Z_i| > c] \\
&\qquad \leq \Pr[\max_i |Z_i| > c]
\end{aligned}$$

- choose c so that this last quantity is α
- Other statistics exist too

Stepdown methods for multiple testing

- Here's an interesting algorithm.
- Suppose we specify an order before testing:
  1. Test $\mu_1 = 0$ against $\mu_1 \neq 0$ at size α
  2. If we fail to reject, stop. Otherwise, test $\mu_2 = 0$ against $\mu_2 \neq 0$ at size α.
  3. If we fail to reject, stop. Otherwise test $\mu_3 = 0$ (and so on...)
- Now suppose that j denotes the first true null hypothesis, so $\mu_i \neq 0$ for $i < j$ but $\mu_j = 0$

$$\begin{aligned}
&\Pr[\text{at least one test rejects a true null hypothesis}] \\
&\qquad = \Pr[\text{at least one } |Z_i| > c \text{ when } \mu_i = 0] \\
&\qquad \leq \Pr[|Z_j| > c] \\
&\qquad \leq \alpha
\end{aligned}$$

- If we order the tests <u>in advance</u> and stop when we fail to reject, we control size at α

Holm's variation of the Bonferroni correction

1. Test all k hypotheses at α/k and let R_1 be the number of hypotheses rejected.
2. Test the remaining (nonrejected) hypotheses at $\alpha/(k-R_1)$ and let R_2 be the number of hypotheses rejected.
3. Test again at $\alpha/(k-R_1-R_2)$ (and so on).

Romano and Wolf's (2005) *StepM* procedure does the same thing, but with White's bootstrap procedure

- Either way, this approach lets you find more than one significant result in your paper

More details about the stepdown procedure

- Suppose we specify an order before testing (again)
- In test j, assume that the null hypothesis for **all of the previous tests** is false.
- Again, suppose that j denotes the first true null hypothesis, so $\mu_i \neq 0$ for $i < j$ but $\mu_j = 0$

$$\begin{aligned}
&\Pr[\text{at least one test rejects a true null hypothesis}] \\
&\qquad = \Pr[\text{at least one } |Z_i| > c \text{ when } \mu_i = 0] \\
&\qquad \leq \Pr[|Z_j| > c; \mu_{j-1} \neq 0, \dots, \mu_1 \neq 0] \\
&\qquad \leq \alpha
\end{aligned}$$

- So we can assume all of the previous steps were correct in deriving a statistic for each step.

Next topic: what is a nuisance parameter?

- A nuisance parameter affects the (asymptotic) distribution of the statistic we want to study, but is not of interest on its own
- If we want to test hypotheses about b in

$$y_i = a + bx_i + gz_i + e_i$$

where $e_i \sim (0, \sigma)$, then a, g, and σ are all potentially nuisance parameters

- If we want to estimate IRFs for the VAR

$$y_t = a_0 + \sum_{i=1}^{p} A_i y_{t-i} + e_t$$

then information about order of integration and cointegrating relationships can be thought of as nuisance parameters

Dealing with simple nuisance parameters

- Often we have a consistent estimator that we can plug in ($\hat{\sigma}$ in a t-test)
- If not, we can take the supremum over the possible values of the nuisance parameter
    - i.e. in testing for a break, the date is often a nuisance parameter
    - This could lead to a "test in levels and test in differences" approach to time-series (we'll see that that's too simplistic next time)
- Even if we have a consistent estimator, we may still want to take the second approach
    - The asymptotic distribution may be well behaved, but the finite-sample distribution may be much worse.
- We can use the asymptotic distribution to limit the region that we need to consider for the supremum (McCloskey, 2012)

Basic idea for using asymptotic distributions of nuisance parameters

- Suppose we have a nuisance parameter θ_1 and a parameter of interest μ.
- θ_1 can be vector valued.
- Assume we reject if $\hat{\mu} > c_\alpha$ for some critical value c_α
- The procedure:
  1. Construct a $1-\epsilon$ confidence interval for θ_1 and call it $\hat{\Theta}_1$.
  2. For any value α, let $c(\alpha, \theta_1)$ be the critical value for a hypotheses is test on μ assuming θ_1 is the true value. Now find

     $$c^* = \sup_{\theta_1 \in \hat{\Theta}_1} c(\alpha - \epsilon, \theta_1)$$

  3. Reject if $\hat{\mu} > c^*$
- The probability of rejecting the null under the alternative is less than or equal to α
- Step 2 may be computationally difficult

Proof of basic idea

- Setup is exactly the same as in the sequential testing example
- Assume that the null hypothesis is true
- We have

$$
\begin{aligned}
\Pr[\hat{\mu} > c^*] &= \Pr\left[\hat{\mu} > c^* \cap (\theta_1 \in \hat{\Theta}_1 \cup \theta_1 \notin \hat{\Theta}_1)\right] \\
&\leq \Pr\left[(\hat{\mu} > c^* \cap \theta_1 \in \hat{\Theta}_1) \cup \theta_1 \notin \hat{\Theta}_1\right] \\
&\leq \Pr[\hat{\mu} > c(\alpha - \epsilon, \theta_1)] + \Pr[\theta_1 \notin \Theta_1\big] \\
&\leq \alpha - \epsilon + \epsilon
\end{aligned}
$$

- Note that we can then iterate: bound other nuisance parameters and test other hypotheses

Some more recommended reading

- Leeb and Pötscher, 2005, “Model Selection and Inference: Facts and Fiction”
- Rosenbaum, 2008, “Testing hypotheses in order”
- More details (in a different setting): Paul Rosenbaum’s *Design of observational studies*
- Also look at McCloskey’s paper

How does this translate into a research strategy?

- As a research strategy:
- **Step 1:** decide on a sequence of hypotheses relevant for your paper
  - Order them: the first should be the main question you want to address in your paper
  - The next should be the second most important question.
  - Subsequent hypotheses should be refinements/sensitivity analysis, etc.
- **Step 2:** What are the nuisance parameters needed to get asymptotic distributions for each of those tests?
- **Step 3** Apply the sequential procedure from above:
  1. First level-α step:
     - Construct $1-\epsilon$ CI for the first nuisance parameters
     - Test the first hypothesis at $\alpha-\epsilon$, choosing the worst critical values of the nuisance parameters over the $1-\epsilon$ confidence interval.
  2. Second level-α step:
     - Construct $1-\epsilon$ CI for the second nuisance parameters
     - Test the second hypothesis at $\alpha-\epsilon$, choosing the worst critical values of the nuisance parameters over the $1-\epsilon$ confidence interval.
  3. Continue, and stop when you fail to reject a hypothesis

Estimation

- This approach can work (in theory) if you are interested in testing hypotheses about parameters or constructing confidence intervals
- Kind of doesn't work if you want to do estimation; for estimation in this setting you probably want to do Bayesian inference (which we'll talk about in more detail soon)
- Actually getting confidence intervals for the nuisance parameters can be tricky

Why not just...?

- Why not just pretest?
  - Pretesting affects the asymptotic distribution of potentially all of the coefficient estimators
  - Let's look at example code...
- Why not just do "model selection"?
  - Model selection behaves like a pretest
  - Assume we test $\beta = 0$
  - As $n \to \infty$, power $\to 1$ unless $\beta = b/\sqrt{n}$ for some b

$$t = \frac{\sqrt{n}\hat{\beta}}{\hat{\tau}} = \frac{\sqrt{n}(\hat{\beta} - \beta)}{\hat{\tau}} + \frac{\sqrt{n}\beta}{\hat{\tau}} \Rightarrow N(b/\tau, 1)$$

  - *Consistent* model selection: do t-test but use $c_n = o(\sqrt{n})$ as cutoff (there is more nuance, but this gives the gist)
  - *Conservative* model selection: just use fixed cutoff (sometimes chooses too large of a model)
  - Problem: for any n, there exists a b_n that causes the exact same problems as in the pretest scenario

Grid bootstrap (dealing with potential unit roots)

- Same ideas can apply when test statistics/confidence intervals depend on choice between I(0) and I(1), etc.
- See Grid Bootstrap example (Hansen, 1999, Mikusheva, 2007, 2012)

Grid bootstrap (dealing with potential unit roots)

- Same ideas can apply when test statistics/confidence intervals depend on choice between I(0) and I(1), etc.
- See Grid Bootstrap example (Hansen, 1999, Mikusheva, 2007, 2012)
- Obviously, following through on this advice is really annoying
  - Potentially impossible at this stage, too
  - I don't know of uniformly valid confidence intervals for the error-correction terms, for example
- Model selection in time-series settings is a really hard problem
  - It seems obvious that we should use the data to choose a model
  - You can easily show that this backfires (in simulations and theory)
  - "Shrinkage," etc. has similar properties
  - Inference is at least something we can understand conceptually, but usually controlling size properly destroys power
- Rule of thumb:
  1. If a model selection statistic implies your model is **bad**, you should probably listen
  2. If a statistic implies your model is **good**, proceed very cautiously

License and copying