Testing Hypotheses about Multiple Population Parameters

Tyler Ransom

Univ of Oklahoma

Feb 12, 2019

Today's plan

1. Review reading topics

   1.1 Hypothesis testing of multiple population parameters

      - Testing equivalence of two β_j's
      - Testing multiple exclusion restrictions
      - The F-test

2. In-class activity: Practice conducting hypothesis tests of multiple population parameters

Test equivalence of two β_j's

Testing single linear restrictions

- So far, we have tested hypotheses that involve only one parameter, β_j
- But some hypotheses involve many parameters
- Example: Are returns to junior college same as university?

$$\log(wage) = \beta_0 + \beta_1 jc + \beta_2 univ + \beta_3 exper + u$$
$$H_0 : \beta_1 = \beta_2$$
$$H_a : \beta_1 \neq \beta_2$$

Testing single linear restrictions

- Recall the familiar t-statistic formula:

$$t = \frac{\text{estimate} - \text{null}}{\text{std. err.}}$$

- Can re-write the null hypothesis to be $H_0 : \beta_1 - \beta_2 = 0$

- Plugging in:

$$t = \frac{\left(\hat{\beta}_1 - \hat{\beta}_2\right) - 0}{se\left(\hat{\beta}_1 - \hat{\beta}_2\right)}$$

Standard Error of $\hat{\beta}_1 - \hat{\beta}_2$

- OLS output gives us $se\left(\hat{\beta}_1\right)$ and $se\left(\hat{\beta}_2\right)$
- but that's not enough to get $se(\hat{\beta}_1 - \hat{\beta}_2)$
- Why? Properties of variances

$$
\begin{aligned}
Var(\hat{\beta}_1 - \hat{\beta}_2) =& Var(\hat{\beta}_1) + Var(\hat{\beta}_2) - 2Cov(\hat{\beta}_1, \hat{\beta}_2), \\
se(\hat{\beta}_1 - \hat{\beta}_2) =& \sqrt{Var(\hat{\beta}_1) + Var(\hat{\beta}_2) - 2Cov(\hat{\beta}_1, \hat{\beta}_2)}
\end{aligned}
$$

- Need to know $Cov(\hat{\beta}_1, \hat{\beta}_2)$ to complete the formula
- This number isn't readily reported by most regression packages

An easy way and a hard way

- **Easy way:** use the `linearHypothesis()` function in the `car` package
  - Syntax: `linearHypothesis(est, ”jc = univ”)`
  - `est` is the output of ‘`lm()`‘
- **Hard way:** re-run slightly different regression

$$\log(wage) = \beta_0 + \theta_1 jc + \beta_2 (jc + univ) + \beta_3 exper + u$$

  - $se\left(\hat{\theta}_1\right) = se\left(\hat{\beta}_1 - \hat{\beta}_2\right)$

Testing multiple exclusion restrictions

Testing multiple exclusion restrictions

- t test is for a **single hypothesis**, whether it involves 1 or 2+ parameters
- But often want to test more than one hypothesis
- We need a statistic that will allow us to test **joint hypotheses**
- Enter: the F-statistic

Example: MLB salaries

- Suppose we use the following model of salaries:

$$\begin{aligned}\log(salary) =& \beta_0 + \beta_1 years + \beta_2 gamesyr + \beta_3 bavg + \\ & \beta_4 hrunsyr + \beta_5 rbisyr + u\end{aligned}$$

- H_0 : Once we control for *years* and *gamesyr*, *rbisyr*, etc. have no effect

$$\begin{aligned}H_0 &: \beta_3 = 0, \beta_4 = 0, \beta_5 = 0 \\ H_a &: \beta_3 \neq 0 \text{ OR } \beta_4 \neq 0 \text{ OR } \beta_5 \neq 0\end{aligned}$$

- **exclusion restrictions:** *bavg*, *hrunsyr*, and *rbisyr* can be excluded

- To test H_0, we need a **joint (multiple) hypotheses test**

Regression output (using `mlb1` data set)

```
tidy(est)
# A tibble: 6 x 5
  term        estimate std.error statistic  p.value
1 (Intercept)  11.2        0.289      38.8   4.19e-128
2 years         0.0689     0.0121      5.68  2.79e-  8
3 gamesyr       0.0126     0.00265     4.74  3.09e-  6
4 bavg          0.000979   0.00110     0.887 3.76e-  1
5 rbisyr        0.0108     0.00717     1.50  1.34e-  1
6 hrunsyr       0.0144     0.0161      0.899 3.69e-  1
```

- Each t-stat for *bavg*, *rbisyr*, *hrunsyr* is insignificant ($|t| < 2$)
- Should we conclude that none of *bavg*, *hrunsyr*, and *rbisyr* affects salaries?

Multicollinearity and single hypothesis tests

- What's going on? Severe multicollinearity
- $Corr\left(hrunsyr, rbisyr\right) \approx 0.9$
- Theoretically, one can't hit a home run without getting at least +1 RBI
- Coeffs on *hrunsyr* and *rbisyr* are imprecisely estimated, so need joint test

Unrestricted vs. restricted model

- Original model (with all variables) is the **unrestricted model**:

$$\log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr + \beta_3 bavg + \beta_4 hrunsyr + \beta_5 rbisyr + u$$

- Imposing $H_0 : \beta_3 = 0, \beta_4 = 0, \beta_5 = 0$ gives the **restricted model**:

$$\log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr + u$$

- We want to see how the fit deteriorates as we remove the three variables

- Use SSR as the measure of fit (or lack thereof)

Intuition of the hypothesis test

- Let SSR_{ur} denote the SSR from the unrestricted model
- In this example, $df_{ur} = 353 - 6 = 347$
- Let SSR_r be the SSR from the restricted model; $df_{ur} = 353 - 3 = 350$
- Also recall that SSR must (weakly) decrease when more variables are added:

$$SSR_r \geq SSR_{ur}$$

- Want to know: Does SSR increase by enough to conclude that H_0 is false?

The F-statistic

- Let our general model be written as

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$$

- What to test if last q variables can be excluded:

$$H_0 : \beta_{k-q+1} = 0, \ldots, \beta_k = 0$$

- Compute SSR_{ur} and SSR_r from each model

- F statistic formula:

$$F = \frac{(SSR_r - SSR_{ur})/(df_r - df_{ur})}{SSR_{ur}/df_{ur}} = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(N-k-1)}$$

More on the F test

- While the t-stat has one degrees-of-freedom, the F-stat has two:
    - Numerator (q) and Denominator ($N - k - 1$)
- We say "Our test statistic has an F distribution with $(q, N - k - 1)$ d.f."

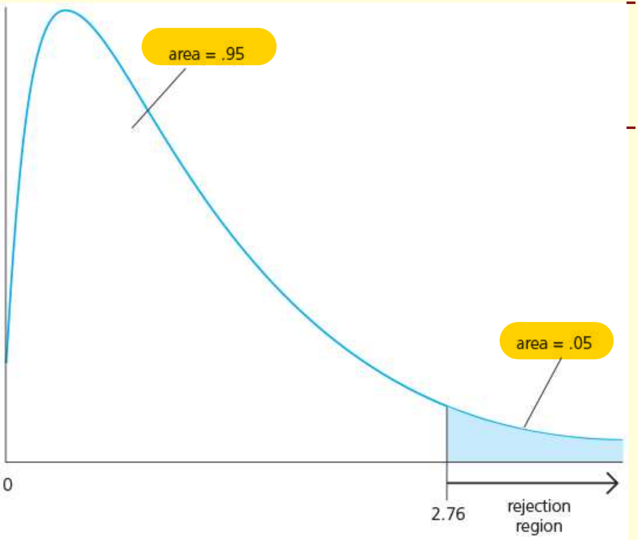
R^2 form of F test

- Can also compute the F-stat from the R^2 of the regression
- Simple algebra shows F can also be written as

$$F = \frac{(R^2_{ur} - R^2_r)/q}{(1 - R^2_{ur})/(n - k - 1)}$$

- Notice: R^2_{ur} comes first in the numerator
- We know $R^2_{ur} \geq R^2_r$ so this ensures $F \geq 0$.

Visualizing the *F* distribution



- F distribution is ratio of χ^2 distributions
- $F > 0$ always

Performing the *F* test in R

The code to do the *F* test in R is below:

```
linearHypothesis(est,c("bavg=0", "rbisyr=0", "hrunsyr=0"))

Linear hypothesis test

Hypothesis:
bavg = 0
rbisyr = 0
hrunsyr = 0
Model 1: restricted model
Model 2: lsalary ~ years + gamesyr + bavg + rbisyr + hrunsyr
  Res.Df    RSS Df Sum of Sq      F    Pr(F)
1    350 198.31
2    347 183.19  3    15.125 9.5503 4.474e-06 ***
```

The overall F test

- By default, R also reports an “overall” F test with every regression

- Conducts the following hypothesis test:

$$
\begin{aligned}
H_0 &: \beta_1 = 0, \ \ldots, \ \beta_k = 0 \\
H_a &: \text{any slope coefficient } \neq 0
\end{aligned}
$$

- From our MLB example:

```
>glance(est)

A tibble: 1 x 11
r.sq adj.r.sq sigma statistic p.value df logLik AIC BIC
0.628 0.622 0.727 117. 2.94e-72 6 -385. 784. 811.
```

Relationship between F and t tests

- In the F test setting, nothing rules out $q = 1$

- So, we can use the F statistic to test, say,

$$H_0 : \beta_j = 0$$
$$H_a : \beta_j \neq 0$$

- Of course, we can use a t statistic, too

- Does that mean there are now two ways to test a single restriction?

- No. It can be shown that, when $q = 1$, then $F = t^2$