Lecture 12: More complicated tests

Pierre Biscaye

Fall 2022

Testing other hypotheses

- Tests of statistical significance so far mostly for $H_0 : \beta_j = 0$
- Economic significance suggests we might want to test some other hypotheses too
  - Is this effect large enough that we should care about it?
- Consider testing $H_0 : \beta_j = b$ against $H_1 : \beta_j \neq b$
- Can test this easily with our test statistic

$$t = \frac{\hat{\beta}_j - \beta_j}{s.e.(\hat{\beta}_j)} \tag{1}$$

- Can conduct hypothesis tests for *any* b.
- Not necessarily very useful; would like a range of probable values for true β.

Confidence Intervals

$$t = \frac{\hat{\beta}_j - \beta_j}{s.e.(\hat{\beta}_j)} \sim t_{n-k-1} \tag{2}$$

$$Pr(\hat{\beta}_j - c_{\frac{\alpha}{2}} * s.e.(\hat{\beta}_j) < \beta_j < \hat{\beta}_j + c_{\frac{\alpha}{2}} * s.e.(\hat{\beta}_j)) = 1 - \alpha \tag{3}$$

- 95% confidence interval shows values where for 95% of samples, the interval would contain the true parameter.

$$\left[\hat{\beta}_j - c_{\frac{\alpha}{2}} * s.e.(\hat{\beta}_j), \hat{\beta}_j + c_{\frac{\alpha}{2}} * se(\hat{\beta}_j)\right] \tag{4}$$

- This gives us an estimated likely range of the magnitude of β_j, and therefore some indication of economic significance.

Confidence Intervals

- A confidence interval for β_j contains all values of β_j where the probability of observing an estimate as large as $\hat{\beta}_j$ is at least α.
  - In other words, the CI contains all b such that we would fail to reject $H_0 : \beta_j = b$ at the α significance level.
- Any H_0 outside of the confidence interval would be rejected with $1 - \alpha$ confidence.
- So if $H_0 : \beta_j = 0$, and 0 is outside of the CI, we reject H_0 at significance level α.
- But if the CI contains 0, we fail to reject H_0.

To Jupyter!

Testing hypotheses on linear combinations of parameters

- Suppose we know that education is related to earnings
- But we want to test something more complicated, e.g., does the *type* of education influence earnings?
- Consider $log(wage_i) = \beta_0 + \beta_1 jc_i + \beta_2 univ_i + \beta_3 exper_i + u_i$
- What are H_0 and H_1 if we want to test whether a year spent in a two-year/junior college (*jc*) or a 4- year college/university (*univ*) has the same relationship with earnings?

Testing hypotheses on linear combinations of parameters

- Suppose we know that education is related to earnings
- But we want to test something more complicated, e.g., does the *type* of education influence earnings?
- Consider $log(wage_i) = \beta_0 + \beta_1 jc_i + \beta_2 univ_i + \beta_3 exper_i + u_i$
- What are H_0 and H_1 if we want to test whether a year spent in a two-year/junior college (jc) or a 4-year college/university ($univ$) has the same relationship with earnings?

$$H_0 : \beta_1 = \beta_2 \tag{5}$$

$$H_1 : \beta_1 \neq \beta_2 \tag{6}$$

Reformulate this is a linear combination of parameters

Similar process as for testing for a difference in means: we have two random variables, so want to define a new random variable as the difference between them and test whether it equals 0.

$$H_0 : \beta_1 = \beta_2 \tag{7}$$

$$H_0 : \beta_1 - \beta_2 = 0 \tag{8}$$

$$H_1 : \beta_1 - \beta_2 \neq 0 \tag{9}$$

Construct a t-statistic

$$t = \frac{(\hat{\beta_1} - \hat{\beta_2} - (\beta_1 - \beta_2))}{s.e.(\hat{\beta_1} - \hat{\beta_2})} \tag{10}$$

$$t = \frac{\hat{\beta_1} - \hat{\beta_2}}{s.e.(\hat{\beta_1} - \hat{\beta_2})} \tag{11}$$

Can we do this with the regression output? To Jupyter!

How to test this? Rewrite the model

$$t = \frac{\hat{\beta_1} - \hat{\beta_2}}{s.e.(\hat{\beta_1} - \hat{\beta_2})} \tag{12}$$

- Calculating $s.e.(\hat{\beta_1} - \hat{\beta_2})$ is tricky; what to do?
- Define $\theta = \beta_1 - \beta_2$
- Create a new variable $allcoll_i = jc_i + univ_i$

$$\begin{aligned}
log(wage_i) &= \beta_0 + \beta_1 jc_i + \beta_2 univ_i + \beta_3 exper_i + u_i \\
log(wage_i) &= \beta_0 + (\theta + \beta_2) jc_i + \beta_2 univ_i + \beta_3 exper_i + u_i \\
log(wage_i) &= \beta_0 + \theta jc_i + \beta_2 (jc_i + univ_i) + \beta_3 exper_i + u_i \\
log(wage_i) &= \beta_0 + \theta jc_i + \beta_2 allcoll_i + \beta_3 exper_i + u_i
\end{aligned}$$

$$t = \frac{\hat{\theta}}{s.e.(\hat{\theta})}$$

- $s.e.(\hat{\theta}) = s.e.(\hat{\beta_1} - \hat{\beta_2})$ given directly by regression output

Describe hypothesis and test

- $log(wage_i) = \beta_0 + \theta jc_i + \beta_2 allcoll_i + \beta_3 exper_i + u_i$

$$H_0 : \theta = 0 \tag{13}$$
$$H_1 : \theta \neq 0 \tag{14}$$
$$Reject\ H_0\ if\ |t| > 1.96 \tag{15}$$

To Jupyter!

Describe hypothesis and test

- $log(wage_i) = \beta_0 + \theta jc_i + \beta_2 allcoll_i + \beta_3 exper_i + u_i$

$$H_0 : \theta = 0 \quad (16)$$
$$H_1 : \theta \neq 0 \quad (17)$$
$$Reject\ H_0\ if\ |t| > 1.96 \quad (18)$$

- $t = -1.468$: fail to reject.
- Can also do this test in R with the 'car' package (though output is not as intuitive).
- What if we wanted to test that $\beta_2 > \beta_1$? Can't do this with 'car'.

One-sided test with linear combination of parameters: $\beta_2 > \beta_1$

- $\theta = \beta_1 - \beta_2$
- $log(wage_i) = \beta_0 + \theta jc_i + \beta_2 allcoll_i + \beta_3 exper_i + u_i$
- Set ourselves up to learn by rejecting the null:

$$H_0 : \theta \geq 0 \tag{19}$$
$$H_1 : \theta < 0 \tag{20}$$
$$Reject\ H_0\ if\ t < -1.645 \tag{21}$$

- We have $t = -1.468$.
- We fail to reject at the 5% significance level that the effect of one year of junior college on wages is greater than the effect of one year of university, holding total years of college and years of experience constant.

Multiple linear restrictions

- Suppose we wanted to test whether a set of variables is *jointly* significant
- For example

$$log(wage_i) = \beta_0 + \beta_1 jc_i + \beta_2 univ_i + \beta_3 exper_i + \beta_4 tenure_i + u_i \quad (22)$$

$$H_0: \beta_3 = 0 \; and \; \beta_4 = 0 \quad (23)$$

- Why do this?
  1. Might be interested in whether years of work variables together are associated with wages
  2. Individual coeffs might not be statistically significant or variables might be highly correlated, so might test whether they jointly explain some of the variation in wages
  3. There are tradeoffs to adding variables to your model: is it worth adding these variables, in the sense that they explain significant variation in wages?

Example: cigarette consumption and infant birth weight

- Suppose we want to estimate the effects of maternal cigarette consumption on birthweight.
- We have data on daily maternal cigarette consumption, birthweight in ounces, birth order (parity), and family income (100s of $) from the 1988 National Health Interview Survey (along with some other variables).
- We could regress

$$bwght_i = \beta_0 + \beta_1 cigs_i + \beta_2 parity_i + \beta_3 faminc_i + u_i \tag{24}$$

- We may have a null $H_0 : \beta_1 = 0$. To Jupyter!
- Any potential concerns with interpreting this result causally?

Concern: SES as omitted variable

$$bwght_i = \beta_0 + \beta_1 cigs_i + \beta_2 parity_i + \beta_3 faminc_i + \beta_4 motheduc_i + \beta_5 fatheduc_i + u_i \tag{25}$$

- Do these family background variables (mother and father years of education) explain meaningful differences in birthweight? To Jupyter!
- We have missing values for *motheduc* and espec. for *fatheduc* - what about MLR2?
- Want to see statistical evidence that it is important to include these variables in the model.
- We have two p-values - can we use these to test whether *either* education variable matters?

Tests for *joint* significance

- Need a new null hypothesis: $H_0 : \beta_4 = 0$ and $\beta_5 = 0$
  - Not the same as testing separate hypotheses for each variable
  - For the joint hypothesis, $H_1 : \beta_4 \neq 0$ *or* $\beta_5 \neq 0$
- Consider 2 statistical models:

$$bwght_i = \beta_0 + \beta_1 cigs_i + \beta_2 parity_i + \beta_3 faminc_i + u_i \tag{26}$$

$$bwght_i = \beta_0 + \beta_1 cigs_i + \beta_2 parity_i + \beta_3 faminc_i + \beta_4 motheduc_i + \beta_5 fatheduc_i + u_i \tag{27}$$

- The first is the *restricted* model, and the second is the *unrestricted* model.
- If H_0 is true, both models should fit the data similarly: use this to design a test statistic around model fit.

Unrestricted and restricted models

- Consider an *unrestricted* model
- We allow β_j to take any value for all j, but want to test that some of them are jointly 0

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + u_i \tag{28}$$

$$H_0 : \beta_{k-q+1} = \beta_{k-q+2} = ... = \beta_k = 0 \tag{29}$$

- We then consider the *restricted* model where we restrict the effects of $\beta_k - q + 1$ to $\beta_k = 0$

$$y_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_{k-q} x_{k-q,i} + u_i \tag{30}$$

- If H_0 is true, then the restricted model is the correct model.

Construct a test statistic based on model fit

- Model fit is reflected by the SSR.
- Consider the sum of squared residuals in the restricted model (SSR_r)
- and the sum of squared residuals in the unrestricted model (SSR_u).
- We define our test statistic, F:

$$F = \frac{(SSR_r - SSR_u)/q}{SSR_u/(n-k-1)} \tag{31}$$

- F is the difference in sum of squares divided by the number of restrictions (in the numerator) as a share of the unrestricted sum of squares normalized by d.f.
- q is the number of restrictions, and also the d.f. in the difference between the two sums of squares
- If H_0 is true, will have SSR_r close to SSR_u and therefore a small F.

How is F distributed?

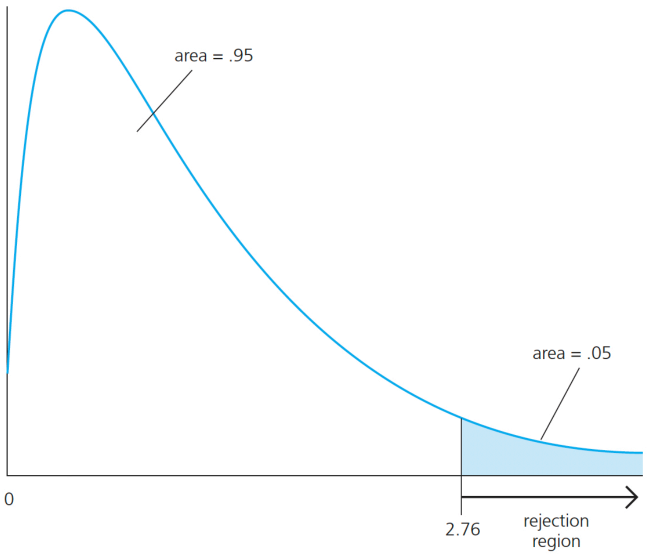
$$F = \frac{(SSR_r - SSR_u)/q}{SSR_u/(n-k-1)} \tag{32}$$

- By definition, $SSR_r \geq SSR_u \Rightarrow F \geq 0$
- Under MLR6, $\hat{u} \sim N(0, \sigma^2)$ so SSRs are sums of squared normal variables, and so is $SSR_r - SSR_u$.
- Squared normal variables are distributed χ^2, and so are sums of χ^2 variables.
- Ratio of two χ^2 variables, each divided by their d.f., is distributed $F(q, n-k-1)$.
- Since F>0, we reject H_0 if $F > c_\alpha$, taken from the $F(q, n-k-1)$ distribution (all tests are one-sided).

F distribution

FIGURE 4.7

The 5% critical value and rejection region in an $F_{3,60}$ distribution.



F critical values for 5% significance level

TABLE G.3b

5% Critical Values of the F Distribution

Denominator Degrees of Freedom	Numerator Degrees of Freedom									
	1	2	3	4	5	6	7	8	9	10
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83

Example: The 5% critical value for numerator $df = 4$ and large denominator df (∞) is 2.37.

Calculating F in birth weight example

$$
\begin{aligned}
bwght_i =& \beta_0 + \beta_1 cigs_i + \beta_2 parity_i + \beta_3 faminc_i + u_i \\
bwght_i =& \beta_0 + \beta_1 cigs_i + \beta_2 parity_i + \beta_3 faminc_i + \\
& \beta_4 motheduc_i + \beta_5 fatheduc_i + u_i
\end{aligned}
$$

$$F = \frac{(SSR_r - SSR_u)/q}{SSR_u/(n-k-1)}$$

To Jupyter!

Interpreting our test

- We fail to reject the null that parental education does not influence birth weight.
  - No evidence from this dataset that parental education is important in birthweight, holding other variables constant.
  - Looking at individual *p*-values would have been misleading!
  - Influenced by correlation between mother and father education.
- Note: if we *had* rejected the null, we would not know which of the restrictions mattered statistically, just that at least one of the restrictions failed.
  - Remember that $H_1 : \beta_4 \neq 0 \textit{ or } \beta_5 \neq 0$
- Should we include parental education? Maybe... but, MLR2 concerns might lead to preferring the restricted model

$$y_i = \beta_0 + \beta_1 cigs_i + \beta_2 parity_i + \beta_3 faminc_i + u_i \tag{33}$$

- Decision is more complicated if we had rejected the null.

Alternate approach to F-statistic

$$R^2 = 1 - \frac{SSR}{SST} \tag{34}$$

$$SSR = SST(1 - R^2) \tag{35}$$

- This applies to both SSR_r and SSR_u, and SST is the same across models, so

$$\begin{aligned}
F &= \frac{(SSR_r - SSR_u)/q}{SSR_u/(n-k-1)} \\
&= \frac{(SST(1-R_r^2) - SST(1-R_u^2))/q}{SST(1-R_u^2)/(n-k-1)} \\
&= \frac{(SST(R_u^2 - R_r^2))/q}{SST(1-R_u^2)/(n-k-1)} \\
&= \frac{(R_u^2 - R_r^2)/q}{(1-R_u^2)/(n-k-1)}
\end{aligned}$$

Calculating F in R

Three ways to calculate F in R

1. $F = \frac{(SSR_r - SSR_u)/q}{SSR_u/(n-k-1)}$
2. $F = \frac{(R_u^2 - R_r^2)/q}{(1-R_u^2)/(n-k-1)}$
3. Using the 'car' package and *linearHypothesis* function

To Jupyter!

p-values for F-statistics

- Unlike t statistics, it can be hard to know what to make of the level of an F statistic.
  - This is because F critical values vary with numerator *and* denominator d.f.
- p-values often more informative - generated by *linearHypothesis* function in R.
- p-value: Probability you observe F at least as large as the one we observed if H_0 is true.
- Example with parent education: $F = 1.437, p = Pr(> F) = 0.238$

Overall F-statistic

- Another use of F-statistics is the overall F of the regression
- The overall F-statistic calculates whether *any* variables have explanatory power

$$
\begin{aligned}
y_i &= \beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki} + u_i \\
H_0 &: \beta_1 = \beta_2 = ... = \beta_k = 0 \\
H_1 &: \beta_1 \neq 0 \, or \, \beta_2 \neq 0 \, or \ldots or \, \beta_k \neq 0
\end{aligned}
$$

- Under this null, the restricted model is $y_i = \beta_0^r + u_i^r$.
- $R_r^2 = 0$
- The test is thus whether the model we specified explains significantly more than 0% of the variation in y.

Overall F-statistic

$$F = \frac{(R_u^2 - R_r^2)/q}{(1 - R_u^2)/(n-k-1)} = \frac{R_u^2/q}{(1 - R_u^2)/(n-k-1)} \tag{36}$$

- Ex with birthweight model: $R_u^2 = 0.0388$, $q = 5$, $n - k - 1 = 1185$

$$F = \frac{0.0388/5}{1 - 0.0388/1185} = \frac{0.008}{0.0008} \approx 10 \tag{37}$$

To Jupyter!

Overall F-statistic

- Ex with birthweight model: $R_u^2 = 0.0388$, $q = 5$, $n - k - 1 = 1185$

$$F = \frac{0.0388/5}{1 - 0.0388/1185} = \frac{0.008}{0.0008} \approx 10 \tag{38}$$

- $p < 0.01$: can reject null at high level of confidence
- True even though R^2 *seems* somewhat small.
- Don't need a big R^2 to nevertheless be explaining a statistically significant amount of the variation.

Other hypotheses

- We can also study more complicated hypotheses.
- Covered in section 4.6 of your book.
- Many can be written as a combination of the two methods here:
  - One approach: substitute some parameters for others (re-writing your model with linear combinations)
  - Then take an F-test on the remaining.