Class 17 – Multiple Regression Model Inference (Part III)

Pedram Jahangiry

JON M. HUNTSMAN SCHOOL OF BUSINESS
UtahStateUniversity

Multiple Hypothesis Testing or Joint Hypothesis Testing: Testing multiple linear restrictions (The F-test)

Salary of major league baseball player

Years in the league

Average number of games per year

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

$$+\beta_3 bavg + \beta_4 hrunsyr + \beta_5 rbisyr + u$$

Batting average

Home runs per year

Runs batted in per year

Test whether performance measures have no effect/can be excluded from regression.

$H_0 : \beta_3 = 0, \beta_4 = 0, \beta_5 = 0$ against H_1 : H_0 is not true

Use Venn Diagram to show the exclusion restrictions!

At least one of them is different from zero

Estimation of the unrestricted model (UR)

```r
reg_UR <- lm(log(salary)~years+gamesyr+bavg+hrunsyr+rbisyr, mlb1)
stargazer(reg_UR, type = "text")
```

$$\widehat{\log(salary)} = \underset{(0.29)}{11.19} + \underset{(.0121)}{.0689}\ years + \underset{(.0026)}{.0126}\ gamesyr + \underset{(.00110)}{.00098}\ bavg + \underset{(.0161)}{.0144}\ hrunsyr + \underset{(.0072)}{.0108}\ rbisyr$$

None of these variabels is statistically significant when tested individually

$$n = 353, \qquad SSR_{UR} = 183, \qquad R^2_{UR} = 0.62$$

Idea: How would the model fit be if these variables were dropped from the regression?

Bigger SSR or smaller SSR?

```
===============================================
                        Dependent variable:
                    ---------------------------
                            log(salary)
-----------------------------------------------
years                        0.069***
                             (0.012)

gamesyr                      0.013***
                             (0.003)

bavg                          0.001
                             (0.001)

hrunsyr                       0.014
                             (0.016)

rbisyr                        0.011
                             (0.007)

Constant                     11.192***
                             (0.289)

-----------------------------------------------
Observations                   353
R2                            0.628
Adjusted R2                   0.622
Residual Std. Error      0.727 (df = 347)
F Statistic          117.060*** (df = 5; 347)
===============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

Estimation of the restricted model (R) $\beta_3 = 0, \beta_4 = 0, \beta_5 = 0$

```r
reg_R <- lm(log(salary)~years+gamesyr, mlb1)
stargazer(reg_R, type = "text")
```

```
================================================
                        Dependent variable:
                    ----------------------------
                            log(salary)
------------------------------------------------
years                        0.071***
                             (0.013)

gamesyr                      0.020***
                             (0.001)

Constant                    11.224***
                             (0.108)

------------------------------------------------
Observations                   353
R2                            0.597
Adjusted R2                   0.595
Residual Std. Error     0.753 (df = 350)
F Statistic         259.320*** (df = 2; 350)
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```

$$\widehat{\log(salary)} = \underset{(.11)}{11.22} + \underset{(.0125)}{.0713}\, years + \underset{(.0013)}{.0202}\; gamesyr$$

$n = 353,\qquad SSR_R = 198,\qquad R_R^2 = 0.59$

The sum of squared residuals necessarily increases from 183 to 198,
but is the increase statistically significant?

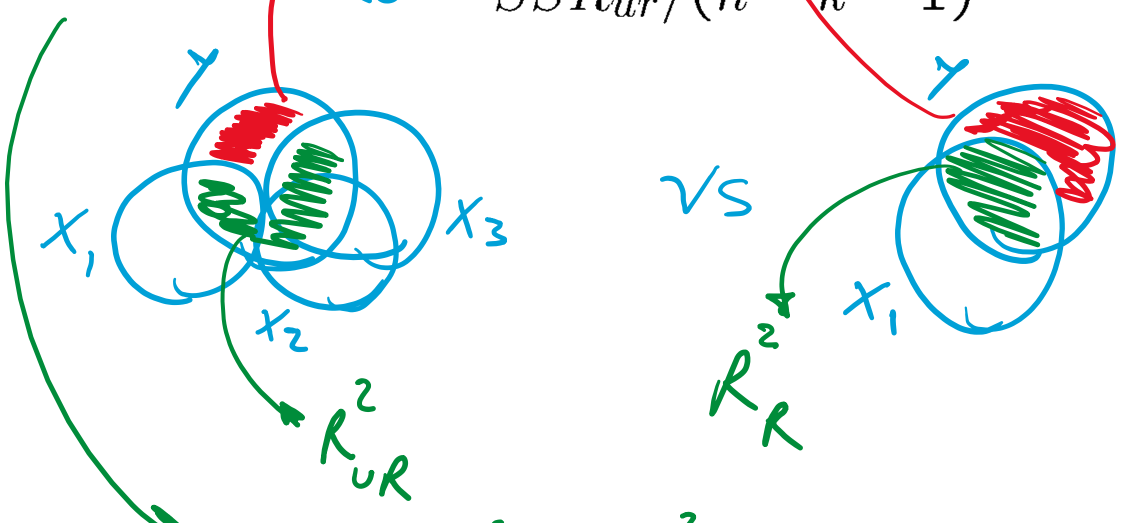
What is a good test statistic for testing $\beta_3 = 0, \beta_4 = 0, \beta_5 = 0$?

Joint significance test statistic:

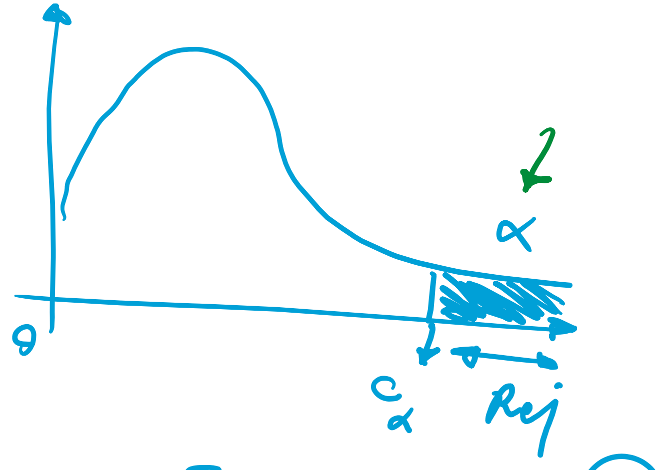
Number of restrictions

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} \sim F_{q,n-k-1}$$

The relative increase of the SSR when dropping variables (going from H_1 to H_0) **follows a F-distribution** (if the null hypothesis H_0 is correct)



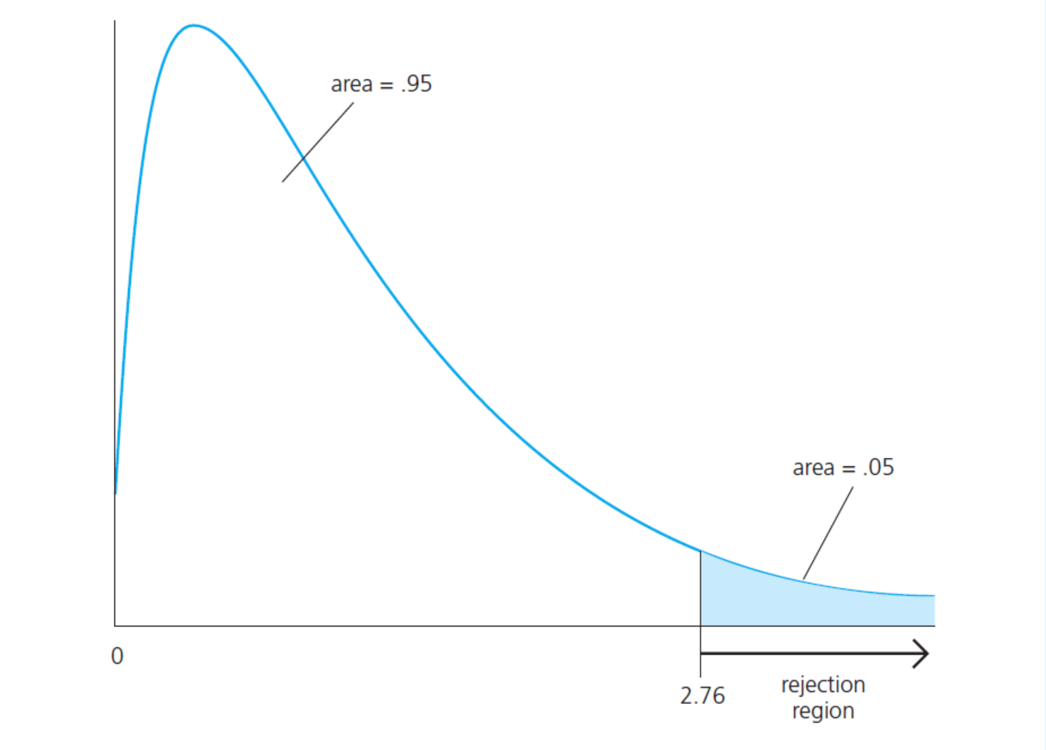
$$F = \frac{(R^2_{UR} - R^2_R)/q}{(1-R^2_{UR})/n-k-1}$$



$F > c_\alpha \Rightarrow$ Rej H_0

Rejection rule

Figure 4.7 The 5% critical value and rejection region in an $F_{3,60}$ distribution.



- ❑ we reject H_0 in favor of H_1 at the chosen significance level if

$$F > c$$

- ❑ How do you find the critical value c?

✓ Use table G.3

✓ Use R: $qf(1-\alpha, q, n-k-1)$

$$qf(0.95, 3{,}60) = 2.76$$

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

Calculating the F-statistic

Number of restrictions to be tested

$$F = \frac{(198.311 - 183.186)/3}{183.186/\boxed{(353-5-1)}} \approx 9.55$$

Degrees of freedom in the unrestricted model

$$F \sim F_{3,347} \;\Rightarrow\; c_{0.01} = 3.83$$

$$p_{value} = P(F_{statistic} > 9.55) = 0.000$$

The null hypothesis is **overwhelmingly** rejected (even at very small significance levels).

R code: $1 - pf(9.55, 3{,}347) =$ 4.475229e−06

- ✓ The three variables are "jointly significant"
- ✓ They were not significant when tested individually
- ✓ The likely reason is **multicollinearity** between them (use Venn diagram to show it!)

Testing multiple linear restrictions Using R

$$H_0 : \beta_3 = 0, \beta_4 = 0, \beta_5 = 0 \qquad \text{against} \qquad H_1 : H_0 \text{ is not true}$$

At least one of them is different from zero

```
  H0 <- c("bavg=0","hrunsyr=0","rbisyr=0")
  linearHypothesis(reg_UR, H0)
Linear hypothesis test

Hypothesis:
bavg = 0
hrunsyr = 0
rbisyr = 0

Model 1: restricted model
Model 2: log(salary) ~ years + gamesyr + bavg + hrunsyr + rbisyr

  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 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Test of overall significance of a regression

Unrestricted Model: $$y = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + u$$

Restricted Model: $$y = \beta_0 + u$$
(regression on constant)

The null hypothesis states that the explanatory variables are not useful at all in explaining the dependent variable

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$$

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} = \frac{R^2/k}{(1-R^2)/(n-k-1)} \sim F_{k,n-k-1}$$

- ✓ The test of overall significance is reported in most regression packages
- ✓ the null hypothesis is usually overwhelmingly rejected

Test of overall significance of a regression Using R

```
reg_UR <- lm(log(salary)~years+gamesyr+bavg+hrunsyr+rbisyr, mlb1)
stargazer(reg_UR, type = "text")

===============================================
                        Dependent variable:
                    ---------------------------
                            log(salary)
-----------------------------------------------
years                        0.069***
                             (0.012)

gamesyr                      0.013***
                             (0.003)

bavg                          0.001
                             (0.001)

hrunsyr                       0.014
                             (0.016)

rbisyr                        0.011
                             (0.007)

Constant                     11.192***
                             (0.289)

-----------------------------------------------
Observations                    353
R2                             0.628
Adjusted R2                    0.622
Residual Std. Error      0.727 (df = 347)
F Statistic          117.060*** (df = 5; 347)
===============================================
Note:              *p<0.1; **p<0.05; ***p<0.01
```

```
H0 <- c("years","gamesyr","bavg=0","hrunsyr=0","rbisyr=0")
linearHypothesis(reg_UR, H0)
Linear hypothesis test

Hypothesis:
years = 0
gamesyr = 0
bavg = 0
hrunsyr = 0
rbisyr = 0

Model 1: restricted model
Model 2: log(salary) ~ years + gamesyr + bavg + hrunsyr + rbisyr

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    352 492.18
2    347 183.19  5    308.99 117.06 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Confirm this:

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} = \frac{R^2/k}{(1-R^2)/(n-k-1)} \sim F_{k,n-k-1}$$

Example: Test whether house price assessments are rational

Actual house price

The assessed housing value (before the house was sold)

Size of lot (in square feet)

$$\log(price) = \beta_0 + \beta_1 \log(assess) + \beta_2 \log(lotsize) + \beta_3 \log(sqrft) + \beta_4 bdrms + u$$

Square footage

Number of bedrooms

- If house price assessments are rational, a 1% change in the assessment should be associated with a 1% change in price.
- In addition, other known factors should not influence the price once the assessed value has been controlled for.

$$H_0 : \beta_1 = 1, \beta_2 = 0, \beta_3 = 0, \beta_4 = 0$$

Example: Test whether house price assessments are rational (cont'd)

Unrestricted regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + u$$

$$\log(price) = \beta_0 + \beta_1 \log(assess) + \beta_2 \log(lotsize) + \beta_3 \log(sqrft) + \beta_4 bdrms + u$$

Restricted regression

$$H_0 : \beta_1 = 1, \beta_2 = 0, \beta_3 = 0, \beta_4 = 0$$

$$y = \beta_0 + x_1 + u$$

$$\Rightarrow \quad [y - x_1] = \beta_0 + u$$

The restricted model is actually a regression of (y-x_1) on a constant

$$\boldsymbol{log(price) - log(assess) = \beta_0 + u}$$

Example: Test whether house price assessments are rational (cont'd)
Regression output for the UnRestricted regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + u$$

$$\log(price) = \beta_0 + \beta_1 \log(assess) + \beta_2 \log(lotsize) + \beta_3 \log(sqrft) + \beta_4 bdrms + u$$

```
reg_UR <- lm(log(price)~log(assess)+log(lotsize)+log(sqrft)+bdrms, hprice1)
stargazer(reg_UR, type = "text")
```

```
=================================================
                        Dependent variable:
                    -----------------------------
                             log(price)
-------------------------------------------------
log(assess)                   1.043***
                              (0.151)

log(lotsize)                   0.007
                              (0.039)

log(sqrft)                    -0.103
                              (0.138)

bdrms                          0.034
                              (0.022)

Constant                       0.264
                              (0.570)

-------------------------------------------------
Observations                    88
R2                             0.773
Adjusted R2                    0.762
Residual Std. Error       0.148 (df = 83)
F Statistic           70.583*** (df = 4; 83)
=================================================
Note:                 *p<0.1; **p<0.05; ***p<0.01
```

$$\begin{aligned}\widehat{\log(price)} &= \underset{(.570)}{.264} + \underset{(.151)}{1.043} \log(assess) + \underset{(.0386)}{.0074} \log(lotsize) \\ &\quad - \underset{(.1384)}{.1032} \log(sqrft) + \underset{(.0221)}{.0338}\, bdrms \\ n &= 88,\ \mathrm{SSR} = 1.822,\ R^2 = .773.\end{aligned}$$

Example: Test whether house price assessments are rational (cont'd)

```
reg_UR <- lm(log(price)~log(assess)+log(lotsize)+log(sqrft)+bdrms, hprice1)

H0 <- c("log(assess)=1","log(lotsize)=0","log(sqrft)=0","bdrms=0")
linearHypothesis(reg_UR, H0)
Linear hypothesis test

Hypothesis:
log(assess) = 1
log(lotsize) = 0
log(sqrft) = 0
bdrms = 0

Model 1: restricted model
Model 2: log(price) ~ log(assess) + log(lotsize) + log(sqrft) + bdrms

  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     87 1.8801
2     83 1.8215  4   0.05862 0.6678 0.6162
```

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} = \frac{(1.880 - 1.822)/4}{1.822/(88-4-1)} \approx .661$$

Conclusion: We fail to reject the null i.e. everything is reflected in variable "assess" already!