Econometrics Lecture Notes II

Multiple Regression (close to chapters 3-7, Wooldridge)

M Loecher

Introduction

Ceteris Paribus

Residualizing

Omitted Variable Bias

"Controlling For"

Precision

Type I/II Errors

Testing multiple linear restrictions: The F-test

Introduction

Student Performance and School Size

There is much interest in the effect of school size on student performance. One claim is that, everything else being equal, students at smaller schools fare better than those at larger schools. This hypothesis is assumed to be true even after accounting for differences in class sizes across schools.

$$math10_i = \beta_0 + \beta_1 \, totcomp_i + \beta_2 \, staff_i + \beta_3 \, enroll_i + u_i \qquad (1)$$

Performance is measured by the percentage of students receiving a passing score on the Michigan Educational Assessment Program (MEAP) standardized tenth-grade math test (math10). School size is measured by student enrollment (*enroll*). The null hypothesis is $H_0 : \beta_{enroll} = 0$, and the alternative is $H_1 : \beta_{enroll} < 0$. For now, we will control for two other factors, average annual teacher compensation (*totcomp*) and the number of staff per one thousand students (*staff*). Teacher compensation is a measure of teacher quality, and staff size is a rough measure of how much attention students receive.

OLS Estimate

The estimated equation, with standard errors in parentheses, is

$$\begin{aligned}
\widehat{math10}_i = & \underset{(6.11)}{2.27} + \underset{(1e-04)}{0.000459}\, totcomp_i + \underset{(0.0398)}{0.0479}\, staff_i - \underset{(0.000215)}{0.000198}\, enroll_i \\
& n = 408, R^2 = 0.05
\end{aligned} \tag{2}$$

Ceteris Paribus

Beer & Chips

$$\text{Headache} = 0.5 \cdot \text{Beers} + 0.1 \cdot \text{Chips} + \text{otherFactors}$$

$$y_i = \beta_0 + \beta_1 \cdot x_{i,1} + \beta_2 \cdot x_{i,2} + u_i$$

model I

$$u_i \sim N(0,0), \text{Beers}_i \sim Pois(1.5), \text{Chips}_i = 15 \cdot \text{Beers}_i$$

model Ia

$$u_i \sim N(0,0.5), \text{Beers}_i \sim Pois(1.5), \text{Chips}_i = 15 \cdot \text{Beers}_i$$

model II

$$u_i \sim N(0,0.5), \text{Beers}_i \sim Pois(1.5), \text{Chips}_i = 15 \cdot \text{Beers}_i + v_i$$

$$v_i \sim N(0,2.5)$$

model I

OLS, using observations 1–100
Dependent variable: Headache1

	Coefficient	Std. Error	t-ratio	p-value
const	0.000000	0.000000	undefined	.
NumBeers	2.00000	0.000000	undefined	.

Mean dependent var	2.940000	S.D. dependent var	2.386071
Sum squared resid	0.000000	S.E. of regression	0.000000
R^2	1.000000	Adjusted R^2	1.000000

model Ia, Simple

OLS, using observations 1–100
Dependent variable: Headache2

	Coefficient	Std. Error	t-ratio	p-value
const	−0.0209368	0.0771647	−0.2713	0.7867
NumBeers	1.99596	0.0408399	48.8728	0.0000

Mean dependent var	2.913127	S.D. dependent var	2.429612
Sum squared resid	23.03230	S.E. of regression	0.484792
R^2	0.960588	Adjusted R^2	0.960186
$F(1,98)$	2388.554	P-value(F)	1.26e–70
Log-likelihood	−68.48021	Akaike criterion	140.9604
Schwarz criterion	146.1708	Hannan–Quinn	143.0691

model Ia, Multiple

OLS, using observations 1–100
Dependent variable: Headache2
Omitted due to exact collinearity: NumChips

	Coefficient	Std. Error	t-ratio	p-value
const	−0.0209368	0.0771647	−0.2713	0.7867
NumBeers	1.99596	0.0408399	48.8728	0.0000

Mean dependent var	2.913127	S.D. dependent var	2.429612
Sum squared resid	23.03230	S.E. of regression	0.484792
R^2	0.960588	Adjusted R^2	0.960186
$F(1, 98)$	2388.554	P-value(F)	1.26e–70
Log-likelihood	−68.48021	Akaike criterion	140.9604
Schwarz criterion	146.1708	Hannan–Quinn	143.0691

model II, Simple 1

OLS, using observations 1–100
Dependent variable: Headache

	Coefficient	Std. Error	t-ratio	p-value
const	0.113321	0.0778424	1.4558	0.1487
NumBeers	1.94467	0.0411986	47.2022	0.0000

Mean dependent var	2.971979	S.D. dependent var	2.370529
Sum squared resid	23.43863	S.E. of regression	0.489050
R^2	0.957869	Adjusted R^2	0.957439
$F(1, 98)$	2228.052	P-value(F)	3.31e–69
Log-likelihood	−69.35463	Akaike criterion	142.7093
Schwarz criterion	147.9196	Hannan–Quinn	144.8180

model II, Simple 2

OLS, using observations 1–100
Dependent variable: Headache

	Coefficient	Std. Error	t-ratio	p-value
const	0.0638358	0.0758838	0.8412	0.4023
NumChips	0.130339	0.00266228	48.9577	0.0000

Mean dependent var	2.971979	S.D. dependent var	2.370529
Sum squared resid	21.85272	S.E. of regression	0.472215
R^2	0.960719	Adjusted R^2	0.960318
$F(1, 98)$	2396.861	P-value(F)	1.07e–70
Log-likelihood	−65.85161	Akaike criterion	135.7032
Schwarz criterion	140.9136	Hannan–Quinn	137.8119

model II, Multiple

OLS, using observations 1–100
Dependent variable: Headache

	Coefficient	Std. Error	t-ratio	p-value
const	0.0731100	0.0739903	0.9881	0.3256
NumBeers	0.788484	0.313234	2.5172	0.0135
NumChips	0.0779754	0.0209630	3.7197	0.0003

Mean dependent var	2.971979	S.D. dependent var	2.370529
Sum squared resid	20.51274	S.E. of regression	0.459860
R^2	0.963128	Adjusted R^2	0.962368
$F(2, 97)$	1266.857	P-value(F)	3.05e–70
Log-likelihood	−62.68765	Akaike criterion	131.3753
Schwarz criterion	139.1908	Hannan–Quinn	134.5384

Warmup Tasks

1. Write down the expression for the slope in OLS with one variable.
2. Try to simplify this term

$$\sum_{i=1}^{n} (x_i - \bar{x})\bar{y}$$

3. Rewrite 1. by leaving out $\bar{y}$.

Residualizing

Correlations among exogeneous variables

Example with k = 2

$$x_{i,1} = \hat{\delta}_{01} + \hat{\delta}_{11} \cdot x_{i,2} + \hat{r}_{i,1}$$
$$x_{i,2} = \hat{\delta}_{02} + \hat{\delta}_{12} \cdot x_{i,1} + \hat{r}_{i,2}$$

where each “helper” regression yields its own “R-squared”: $R_1^2, R_2^2, \ldots$.

Recall the reason, why the coefficients in the **multiple** regression

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_{i,1} + \hat{\beta}_2 \cdot x_{i,2} + u_i$$

are typically not equal to the individual regressions

$$y_i = \hat{\hat{\beta}}_{01} + \hat{\hat{\beta}}_1 \cdot x_{i,1} + v_{i,1}, y_i = \hat{\hat{\beta}}_{02} + \hat{\hat{\beta}}_2 \cdot x_{i,1} x_{i,2} + v_{i,2}$$

A “Partialling Out” Interpretation of Multiple Regression

Consider again the case with $k = 2$ independent variables. We can write

$$[3.22] \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} \hat{r}_{1,i} y_i}{\sum_{i=1}^{n} \hat{r}_{1,i}^2} \tag{3}$$

where $\hat{r}_{1,i}$ are the OLS residuals from a simple regression of x_1 on x_2.

Compare this to (12) from the Econometrics-I slides.

Thus, $\hat{\beta}_1$ measures the sample relationship between y and x_1 after x_2 has been partialled out.

More on “Partialling Out”

One can show that the estimated coefficient of an explanatory variable in a multiple regression can be obtained in two steps:

1. Regress the explanatory variable x_j on all other explanatory variables $x_{i \neq j}$
2. Regress y on the residuals from this regression

Why does this procedure work?

- The residuals from the first regression is the part of the explanatory variable that is uncorrelated with the other explanatory variables
- The slope coefficient of the second regression therefore represents the **isolated** effect of the explanatory variable on the dep. variable

Wage Data

We confirm the partialling out interpretation of the OLS estimates as follows.

1. Regressing educ on exper and tenure and saving the residuals $\hat{r}_1$
2. Regress log(wage) on the residuals from this regression
3. Compare the coefficient on $\hat{r}_1$ with the coefficient on educ in the regression of log(wage) on educ, exper, and tenure.

Regressing educ on exper and tenure

Model 1: OLS, using observations 1–526
Dependent variable: educ

	Coefficient	Std. Error	t-ratio	p-value
const	13.575	0.184	73.647	0.000
exper	−0.074	0.010	−7.559	0.000
tenure	0.048	0.018	2.600	0.010

Mean dep var	12.56	S.D. dep var	2.77
Sum sq resid	3617.48	S.E. of regr	2.63
R^2	0.1	Adjusted R^2	0.1
$F(2, 523)$	29.49	P-value(F)	0
Log-likelihood	−1253.487	Akaike criterion	2512.974
Schwarz criterion	2525.77	Hannan–Quinn	2517.98

Regress log(wage) on educ-residuals

Model 2: OLS, using observations 1–526
Dependent variable: lwage

	Coefficient	Std. Error	t-ratio	p-value
const	1.623	0.021	78.555	0.000
rhat1	0.092	0.008	11.679	0.000

Mean dep var	1.62	S.D. dep var	0.53
Sum sq resid	117.69	S.E. of regr	0.47
R^2	0.21	Adjusted R^2	0.21
$F(1, 524)$	136.41	P-value(F)	0
Log-likelihood	-352.5903	Akaike criterion	709.1806
Schwarz criterion	717.71	Hannan–Quinn	712.52

Compare the coefficients

Model 1: OLS, using observations 1–526
Dependent variable: lwage

	Coefficient	Std. Error	t-ratio	p-value
const	0.284	0.104	2.729	0.007
educ	0.092	0.007	12.555	0.000
exper	0.004	0.002	2.391	0.017
tenure	0.022	0.003	7.133	0.000

Mean dep var	1.62	S.D. dep var	0.53
Sum sq resid	101.46	S.E. of regr	0.44
R^2	0.32	Adjusted R^2	0.31
$F(3, 522)$	80.39	P-value(F)	0
Log-likelihood	−313.5478	Akaike criterion	635.0956
Schwarz criterion	652.16	Hannan–Quinn	641.78

Regress log(wage) on exp-residuals

Regressing experience on educ and tenure:

Model 2: OLS, using observations 1–526
Dependent variable: lwage

	Coefficient	Std. Error	t-ratio	p-value
const	1.623	0.023	70.237	0.000
rhat2	0.004	0.002	1.989	0.047

Mean dep var	1.62	S.D. dep var	0.53
Sum sq resid	147.22	S.E. of regr	0.53
R^2	0.01	Adjusted R^2	0.01
$F(1, 524)$	3.96	P-value(F)	0.05
Log-likelihood	-411.4614	Akaike criterion	826.9227
Schwarz criterion	835.45	Hannan–Quinn	830.26

Omitted Variable Bias

Omitted Variable Bias

Population version

$$
\begin{aligned}
y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \\
x_2 &= \delta_0 + \delta_1 x_1 + v \\
\Rightarrow y &= \beta_0 + \beta_1 x_1 + \beta_2(\delta_0 + \delta_1 x_1 + v) + u \\
y &= (\beta_0 + \beta_2 \delta_0) + (\beta_1 + \beta_2 \delta_1) x_1 + (u + \beta_2 v)
\end{aligned}
$$

$$\Rightarrow \beta_2 \delta_1 = OVB$$

OVB, II

We can also write the bias without the need for additional Greek symbols:

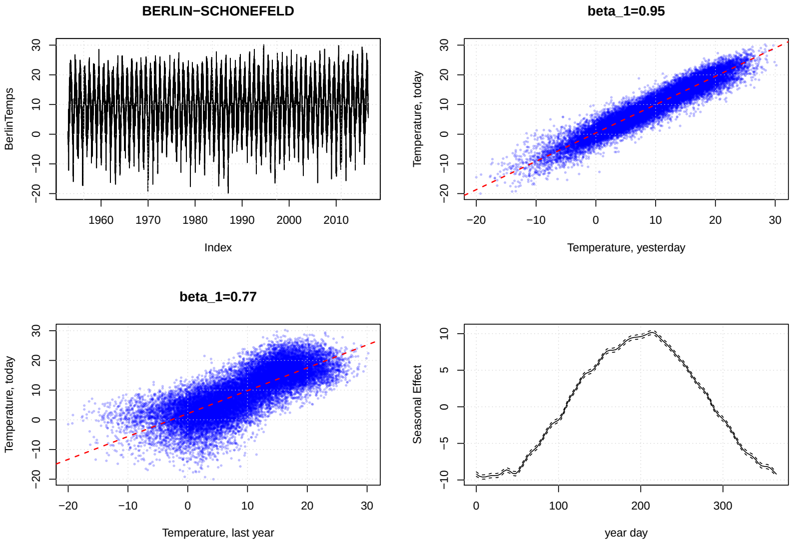
$$OVB = \beta_2 \cdot \frac{cov(x_1, x_2)}{var(x_1)} = \beta_2 \cdot \frac{s_{x_1,x_2}}{s^2_{x_1}}$$

TABLE 3.2 Summary of Bias in $\widetilde{\beta}_1$ When x_2 Is Omitted in Estimating Equation (3.40)

	$\text{Corr}(x_1, x_2) > 0$	$\text{Corr}(x_1, x_2) < 0$
$\beta_2 > 0$	Positive bias	Negative bias
$\beta_2 < 0$	Negative bias	Positive bias

“Controlling For”

Control Variables I



Control Variables II

	Model 1	Model 2	Model 3
(Intercept)	2.10^{***}	-0.00	-0.00
	(0.05)	(0.04)	(0.04)
lastYear	0.77^{***}	0.03^{***}	
	(0.00)	(0.01)	
Season		0.97^{***}	1.00^{***}
		(0.01)	(0.00)
R^2	0.59	0.76	0.76
Adj. R^2	0.59	0.76	0.76
Num. obs.	22950	22950	22950
RMSE	5.00	3.81	3.81

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 1: Statistical models

Financial Markets

What exactly do “they” mean by β?

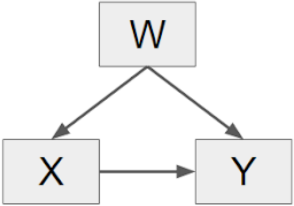
Bad Controls

To illustrate, suppose we are interested in the effects of a college degree on earnings and that people can work in one of two occupations, white collar and blue collar. A college degree clearly opens the door to higher-paying white collar jobs. Should occupation therefore be seen as an omitted variable in a regression of wages on schooling? After all, occupation is highly correlated with both education and pay. Perhaps it's best to look at the effect of college on wages for those within an occupation, say white collar only. The problem with this argument is that once we acknowledge the fact that college affects occupation, comparisons of wages by college degree status within an occupation are no longer apples-to-apples, *even if college degree completion is randomly assigned*.

Bad Controls II

Confounders vs. Colliders

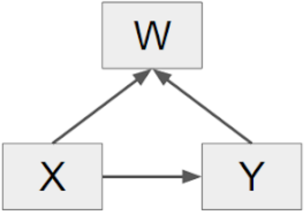
Sentences that begin with *"Controlling for [factors X, Y, and Z], ..."* are reassuring amidst controversial subject matter. But less reassuring is the implicit assumption that X, Y, and Z are indeed the things we call "confounders." We review the definition of a confounder via the following causal graph:



An example would be the effect of educational attainment (x) on earnings (y) where mental ability (w) is a confounder.

Confounders vs. Colliders

Suppose that x is still the amount of education and y is still earned income, but instead of being mental ability, w is now annual dollars spent on decorative artwork. In this scenario, education and income probably cause art purchases rather than vice versa. Supposing this is truth, below is what the new causal graph looks like.



Unlike in previous section, the simpler regression without w recovers the true coefficient of x, while the regression with w has a horribly biased estimate.

Precision

Sampling Variance

Under assumptions MLR.1 - MLR.5, the (square of the) standard error of the slope estimate is:

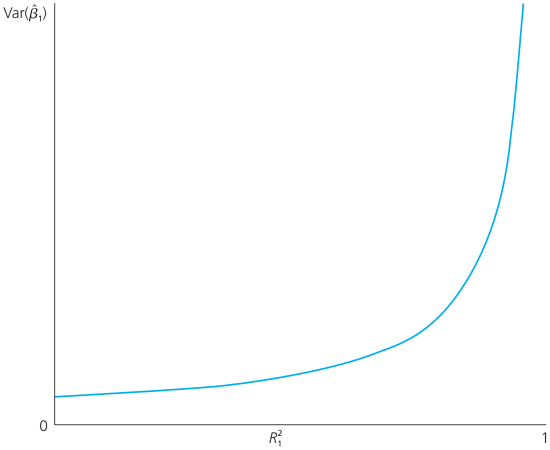
$$(SE)^2 = Var(\hat{\beta}_j) = \frac{\sigma_u^2}{SST_j(1 - R_j^2)}$$

with

- σ_u^2: variance of error term
- SST_j:Total sample variation in explanatory variable x_j: $\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$
- R_j^2:R-squared from a regression of explanatory variable x_j on all other independent variables (including a constant)
- $(1 - R_j^2)$ is often referred to as **Variance Inflation Factor**.

$Var(\hat{\beta}_j)$ as a function of R_j^2

FIGURE 3.1 $\text{Var}(\hat{\beta}_1)$ as a function of R_1^2.



Exercises, OVB, VIF

Type I/II Errors

False Negatives

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.02453    0.14973  -0.164    0.870
## NumBeers    1258.54262 1469.53106   0.856    0.394
## NumChips     -83.76791   97.96879  -0.855    0.395
##
## Residual standard error: 0.9403 on 97 degrees of freedom
## Multiple R-squared:  0.8708, Adjusted R-squared:  0.8682
## F-statistic:   327 on 2 and 97 DF,  p-value: < 2.2e-16
```

R_1^2

How correlated are the two x variables?

```
##               Estimate Std. Error   t value Pr(>|t|)
## (Intercept) 2.796e-06  1.029e-05 2.720e-01    0.786
## NumChips    6.667e-02  3.630e-07 1.836e+05   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
## Residual standard error: 6.464e-05 on 98 degrees of free
## Multiple R-squared:  0.999999997,    Adjusted R-squared:
## F-statistic: 3.373e+10 on 1 and 98 DF,  p-value: < 2.2e-
```

False Positives

Create some VERY random data:

FalsePositives-Ftest.pdf

Testing multiple linear restrictions: The F-test

Intro

We already know how to test whether a particular variable has no partial effect on the dependent variable: use the t statistic. Now, we want to test whether a group of variables has no effect on the dependent variable. More precisely, the null hypothesis is that a set of variables has no effect on y, once another set of variables has been controlled.

Example

As an illustration of why testing significance of a group of variables is useful, we consider the following model that explains major league baseball players' salaries:

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

where *salary* is the 1993 total salary, *years* is years in the league, gamesyr is average games played per year, *bavg* is career batting average (for example, bavg 5 250), *hrunsyr* is home runs per year, and *rbisyr* is runs batted in per year. Suppose we want to test the null hypothesis that, once years in the league and games per year have been controlled for, the statistics measuring performance-*bavg*, *hrunsyr*, and *rbisyr*-have no effect on salary. Essentially, the null hypothesis states that productivity as measured by baseball statistics has no effect on salary.

The F statistic

In terms of the parameters of the model, the null hypothesis is stated as

$$[4.29] \quad H_0 : \beta_3 = \beta_4 = \beta_5 = 0 \tag{4}$$

A test of multiple restrictions is called a **multiple hypotheses test** or a **joint hypotheses test**.

$$[4.30] \quad H_A :?? \tag{5}$$

$$[4.37] \quad F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} \tag{6}$$

Example

Unrestricted ("big") model:

$$\widehat{\mathrm{lsalary}} = \underset{(0.289)}{11.192} + \underset{(0.012)}{0.06886}\,\mathsf{years} + \underset{(0.00265)}{0.01255}\,\mathsf{gamesyr} + \underset{(0.0011)}{0.0009786}\,\mathsf{bavg}$$
$$+ \underset{(0.0161)}{0.01443}\,\mathsf{hrunsyr} + \underset{(0.00717)}{0.01077}\,\mathsf{rbisyr}$$
$$T = 353 \quad \bar{R}^2 = 0.6224 \quad F(5, 347) = 117.06 \quad \hat{\sigma} = 0.72658$$

(standard errors in parentheses)

Restricted ("small") model:

$$\widehat{\mathrm{lsalary}} = \underset{(0.108)}{11.22} + \underset{(0.0125)}{0.0713}\,\mathsf{years} + \underset{(0.00134)}{0.0202}\,\mathsf{gamesyr}$$
$$T = 353 \quad \bar{R}^2 = 0.5948 \quad F(2, 350) = 259.32 \quad \hat{\sigma} = 0.75273$$

(standard errors in parentheses)

Example, cont

$SSR_{ur} =$

$SSR_r =$

$n, k, q =$

$$F = \frac{(\quad - \quad)/ \quad}{\quad /(\quad \quad)}$$

Critical F value from table

$$F_{\quad , \quad , \quad} =$$

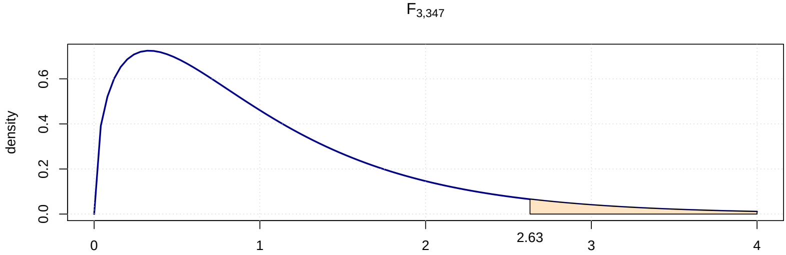
Example, solution

$$SSR_{ur} = 0.72658^2 \cdot (353 - 5 - 1) = 183.19$$

$$SSR_r = 0.75273^2 \cdot (353 - 2 - 1) = 198.31$$

$$n = 353, k = 5, q = 3$$

$$F = \frac{(198.31 - 183.19)/3}{183.19/(353 - 5 - 1)} = 9.55 > F_{0.05,3,347} = 2.63$$



The 5% critical value and rejection region in an F(3,347) distribution

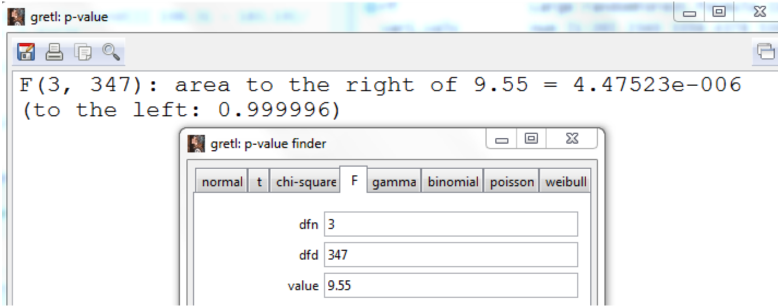
Example, solution, cont.

The F statistic is well above the 1% critical value in the F distribution with 3 and 347 degrees of freedom, and so we soundly reject the hypothesis that *bavg*, *hrunsyr*, and *rbisyr* have no effect on salary.

The outcome of the joint test may seem surprising in light of the insignificant t statistics for the three variables. What is happening is that the two variables *hrunsyr* and *rbisyr* are highly correlated, and this multicollinearity makes it difficult to uncover the partial effect of each variable; this is reflected in the individual t statistics. The F statistic tests whether these variables (including *bavg*) are jointly significant, and multicollinearity between *hrunsyr* and *rbisyr* is much less relevant for testing this

p-values for F statistic

1. Using gretl



2. Tables:

TABLE G.3c 1% 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	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63
90	6.93	4.85	4.01	3.54	3.23	3.01	2.84	2.72	2.61	2.52
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32

Adjusted R-Squared

$$R^2 = 1 - \frac{RSS}{TSS} \stackrel{?}{=} 1 - \frac{\sigma_u^2}{\sigma_y^2}$$

$$R_{adj}^2 = 1 - \frac{RSS/(n-k-1)}{TSS/(n-1)}$$

The R-Squared Form of the F Statistic

For testing exclusion restrictions, it is often more convenient to have a form of the F statistic that can be computed using the R-squareds from the restricted and unrestricted models. Using the fact that $SSR_r = SST \cdot (1 - R_r^2)$ and $SSR_{ur} = SST \cdot (1 - R_{ur}^2)$ we can substitute into (4.37) to obtain

$$[4.41] \quad F = \frac{(R_{ur}^2 - R_r^2)/q}{(1 - R_{ur}^2)/(n - k - 1)} \tag{7}$$

Example, F test

Consider the following model to explain child birth weight in terms of various factors:

$$[4.42] \quad \begin{aligned} bwght &= \beta_0 + \beta_1\, cigs + \beta_2\, parity + \beta_3\, faminc \\ &+ \beta_4\, motheduc + \beta_5\, fatheduc + u \end{aligned}$$

bwght: birth weight, in pounds. *cigs*: average number of cigarettes the mother smoked per day during pregnancy. *parity*: the birth order of this child. *faminc*: annual family income. *motheduc*: years of schooling for the mother. *fatheduc*: years of schooling for the father.

Let us test the null hypothesis that,

1. after controlling for *cigs* and *parity*, parents feature have no effect on bwght.
2. after controlling for *cigs*, *parity*, and *faminc*, parents education has no effect on bwght.

	Model 1	Model 2	Model 3
(Intercept)	117.8035^{***}	115.4699^{***}	114.5243^{***}
	(1.2147)	(1.6559)	(3.7285)
cigs	-0.6324^{***}	-0.5979^{***}	-0.5959^{***}
	(0.1076)	(0.1088)	(0.1103)
parity	1.7630^{**}	1.8323^{**}	1.7876^{**}
	(0.6576)	(0.6575)	(0.6594)
faminc		0.0671^{*}	0.0560
		(0.0324)	(0.0366)
motheduc			-0.3705
			(0.3199)
fatheduc			0.4724
			(0.2826)
R^2	0.0329	0.0364	0.0387
Adj. R^2	0.0313	0.0340	0.0347
Num. obs.	1191	1191	1191
RMSE	19.8234	19.7961	19.7888

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 2:

Example, F test solution

$$F_{3,1} = \frac{(0.0387 - 0.0329)/3}{(1 - 0.0387)/1185} = 2.38 < F_{0.05,3,1185} = 2.61$$

$$F_{3,2} = \frac{(0.0387 - 0.0364)/2}{(1 - 0.0387)/1185} = 1.42 < F_{0.05,2,1185} = 3$$

Is the decision w.r.t faminc surprising ? Look at the following:

$$F_{2,1} = \frac{(0.0364 - 0.0329)/1}{(1 - 0.0364)/(1191 - 3 - 1)} = 4.31 > F_{0.05,1,1187} = 3.85$$

Logarithmic Transformations

Summary, Logarithms (Again!)

TABLE 2.3 Summary of Functional Forms Involving Logarithms

Model	Dependent Variable	Independent Variable	Interpretation of β_1
Level-level	y	x	$\Delta y = \beta_1 \Delta x$
Level-log	y	$\log(x)$	$\Delta y = (\beta_1/100)\%\Delta x$
Log-level	$\log(y)$	x	$\%\Delta y = (100\beta_1)\Delta x$
Log-log	$\log(y)$	$\log(x)$	$\%\Delta y = \beta_1 \%\Delta x$

Example, Elasticity

For a sample of 506 communities in the Boston area, we estimate a model relating median housing price (*price*) in the community to various community characteristics: *nox* is the amount of nitrogen oxide in the air, in parts per million; dist is a weighted distance of the community from five employment centers, in miles; *rooms* is the average number of rooms in houses in the community; and *stratio* is the average student-teacher ratio of schools in the community. The population model is

$$log(price) = \beta_0 + \beta_1 log(nox) + \beta_2 log(dist) + \beta_3 rooms + \beta_4 stratio + u \tag{8}$$

Thus, β_1 is the elasticity of *price* with respect to *nox*. We wish to test $H_0 : \beta_1 = -1$ against the alternative $H_1 : \beta_1 \neq -1$

	Model 1	Model 2	Model 3
(Intercept)	11.084^{***}	11.005^{***}	10.873^{***}
	(0.318)	(0.290)	(0.299)
lnox	-0.954^{***}	-0.901^{***}	-0.825^{***}
	(0.117)	(0.106)	(0.115)
log(dist)	-0.134^{**}	-0.215^{***}	-0.214^{***}
	(0.043)	(0.040)	(0.040)
rooms	0.255^{***}	0.246^{***}	0.247^{***}
	(0.019)	(0.017)	(0.017)
stratio	-0.052^{***}	-0.042^{***}	-0.039^{***}
	(0.006)	(0.005)	(0.006)
crime		-0.015^{***}	-0.014^{***}
		(0.001)	(0.002)
proptax			-0.002
			(0.001)
R^2	0.584	0.656	0.659
Adj. R^2	0.581	0.653	0.654
Num. obs.	506	506	506
RMSE	0.265	0.241	0.241

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Example, house prices, cont.

The t statistic is $(-0.954 + 1)/.117 = .393$. Controlling for the factors we have included, there is little evidence that the elasticity is different from -1.