Economics 6400: Econometrics

Lecture 5: More on inference, the F Test, and other topics

CSU, East Bay

October 20, 2016

Last week...

- We derived the **estimated** variation of $\hat{\beta}_j$:

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

where

- $\hat{\sigma}^2 = \frac{SSR}{n-k-1} = \frac{\sum_{i=1}^n \hat{u}_i^2}{n-k-1}$ (from main regression)
- $SST_j = \sum_{i=1}^n (x_{ij} - \overline{x}_j)^2$ is the sample variation in right-hand side variable x_j
- R_j^2 is the R^2 from a regression of x_j on the <u>other</u> right-hand side variables (including a constant)
- The **standard error** is $se(\hat{\beta}_j) = \sqrt{\widehat{Var}(\hat{\beta}_j)} = \sqrt{\frac{\hat{\sigma}^2}{SST_j(1-R_j^2)}}$

Last week...

- We then used the standard error for hypothesis testing
- Suppose we wish to test whether β_j is equal to some value a_j (usually zero):

$$\mathrm{H}_0 : \beta_j = a_j$$
$$\mathrm{H}_1 : \beta_j \neq a_j$$

- Test statistic is:

$$t_{\hat{\beta}_j} \equiv \frac{\hat{\beta}_j - a_j}{se(\hat{\beta}_j)} \sim t_{n-k-1} = t_{df}$$

- The smallest significance level at which the null hypothesis is still rejected is called the *p*-**value** of the hypothesis test

Testing hypotheses about a linear combination of parameters

- Example: Returns to education at 2-year versus a 4-year college:

$$\log(wage) = \beta_0 + \beta_1\, jc + \beta_2 univ + \beta_3 exper + u$$

  where *jc* is years at 2-year college and *univ* is years at 4-year college
- We want to test if $\beta_1 = \beta_2$ (i.e. returns are equal)
- Test $\mathrm{H}_0 : \beta_1 - \beta_2 = 0$ against $\mathrm{H}_1 : \beta_1 - \beta_2 < 0$
- Possible test statistic: $t = \frac{\hat{\beta}_1 - \hat{\beta}_2}{se(\hat{\beta}_1 - \hat{\beta}_2)}$
- Difference in the coefficient is normalized by the variance of the difference
- If the difference is too negative (i.e. return to junior college lower) then the null hypothesis would be rejected

Testing hypotheses about a linear combination of parameters (twoyear.dta)

```
. reg lwage jc univ exper

      Source |       SS       df       MS              Number of obs =    6763
-------------+------------------------------           F(  3,  6759) =  644.53
       Model |  357.752575     3  119.250858           Prob > F      =  0.0000
    Residual |  1250.54352  6759  .185019014           R-squared     =  0.2224
-------------+------------------------------           Adj R-squared =  0.2221
       Total |  1608.29609  6762  .237843255           Root MSE      =  .43014

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          jc |   .0666967   .0068288     9.77   0.000     .0533101    .0800833
        univ |   .0768762   .0023087    33.30   0.000     .0723504    .0814021
       exper |   .0049442   .0001575    31.40   0.000     .0046355    .0052529
       _cons |   1.472326   .0210602    69.91   0.000     1.431041     1.51361
------------------------------------------------------------------------------
```

- Note that $\beta_{jc} - \beta_{univ} = 0.067 - 0.077 = -0.010$

Testing hypotheses about a linear combination of parameters

- Standard regression output will not compute the covariance term

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

- Possible to derive the variance-covariance matrix for $\hat{\beta}$ that contains the covariance of two slope coefficients (see Equation E.14 in Appendix E) but we will not do so in this class

Alternative method: estimate a different model!

- Define $\theta_1 = \beta_1 - \beta_2 \Rightarrow \beta_1 = \theta_1 + \beta_2$
- Test $\mathrm{H}_0 : \theta_1 = 0$ against $\mathrm{H}_1 : \theta_1 < 0$
- Substitute the expression for β_1 (that contains θ_1) into the regression equation and group coefficients:

$$\begin{aligned}\log(wage) =& \beta_0 + \beta_1 jc + \beta_2 univ + \beta_3 exper + u \\ =& \beta_0 + (\theta_1 + \beta_2) jc + \beta_2 univ + \beta_3 exper + u \\ =& \beta_0 + \theta_1 jc + \beta_2 (jc + univ) + \beta_3 exper + u\end{aligned}$$

- Create the new variable $jc + univ$ and include in a regression with *jc* and *exper*
- The coefficient on *jc* will be our estimate of θ_1 and Stata will compute its standard error! (Minor) magic!

Testing hypotheses about a linear combination of parameters (twoyear.dta)

```
. gen totcoll = jc + univ

. reg lwage jc totcoll exper

      Source |       SS       df       MS              Number of obs =    6763
-------------+------------------------------           F(  3,  6759) =  644.53
       Model |  357.752575     3  119.250858           Prob > F      =  0.0000
    Residual |  1250.54352  6759  .185019014           R-squared     =  0.2224
-------------+------------------------------           Adj R-squared =  0.2221
       Total |  1608.29609  6762  .237843255           Root MSE      =  .43014

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          jc |  -.0101795   .0069359    -1.47   0.142    -.0237761     .003417
     totcoll |   .0768762   .0023087    33.30   0.000     .0723504    .0814021
       exper |   .0049442   .0001575    31.40   0.000     .0046355    .0052529
       _cons |   1.472326   .0210602    69.91   0.000     1.431041     1.51361
------------------------------------------------------------------------------
```

- $t = \frac{-0.0102}{0.0069} = -1.47$, p-value $= P(T < -1.47) = 0.07$

Testing hypotheses about a linear combination of parameters

- The coefficients on the other variables are unchanged (e.g. *exper*)
  - Provides a way to check if the new model has been properly estimated
- Strategy of rewriting the model so that it contains the parameter of interest works in all cases

Testing multiple linear restrictions

- If we wish to test multiple restrictions, we can no longer use a t test but must use an F test
- Often we wish to test whether a group of variables has no effect on the left-hand side variable
- Consider the following model of baseball player salaries:

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

where

- *gamesyr* is average games played per year
- *bavg* is career batting average
- *hrunsyr* is home runs per year
- *rbisyr* is runs batted in (rbi) per year

Model of baseball player salaries

```
. reg lsalary years gamesyr bavg hrunsyr rbisyr

      Source |       SS       df       MS              Number of obs =     353
-------------+------------------------------           F(  5,   347) =  117.06
       Model |  308.989208     5  61.7978416           Prob > F      =  0.0000
    Residual |  183.186327   347  .527914487           R-squared     =  0.6278
-------------+------------------------------           Adj R-squared =  0.6224
       Total |  492.175535   352  1.39822595           Root MSE      =  .72658

------------------------------------------------------------------------------
     lsalary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       years |   .0688626   .0121145     5.68   0.000     .0450355    .0926898
     gamesyr |   .0125521   .0026468     4.74   0.000     .0073464    .0177578
        bavg |   .0009786   .0011035     0.89   0.376    -.0011918     .003149
     hrunsyr |   .0144295    .016057     0.90   0.369    -.0171518    .0460107
      rbisyr |   .0107657    .007175     1.50   0.134    -.0033462    .0248776
       _cons |   11.19242   .2888229    38.75   0.000     10.62435    11.76048
------------------------------------------------------------------------------
```

- $SSR = 183.186,\ df = n - k - 1 = 347,\ R^2 = 0.628$

Exclusion restrictions

- Suppose we want to test the null hypothesis that *bavg*, *hrunsyr*, and *rbisyr* have no effect on salary, once *years* and *gamesyr* have been controlled for
  - None of these variables is statistically significant on their own due to multicollinearity
  - They are all highly correlated, e.g. $Corr(rbisyr, hrunsyr) = 0.89$
- The null and alternative hypotheses are:

$$\mathrm{H}_0 : \beta_3 = 0, \beta_4 = 0, \beta_5 = 0$$
$$\mathrm{H}_1 : \mathrm{H}_0 \text{ not true}$$

- The null constitutes three exclusion restrictions

Restricted and unrestricted models

- The **restricted model (r)** without the three variables is:

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

- The **unrestricted model (ur)** with the three variables is:

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

- Basic idea: How much does the SSR increase when we impose the $q = 3$ exclusion restrictions?
    - If SSR_r is sufficiently larger than SSR_{ur} then reject H_0!

Restricted model results

```
. reg lsalary years gamesyr

      Source |       SS       df       MS              Number of obs =     353
-------------+------------------------------           F(  2,   350) =  259.32
       Model |  293.864058     2  146.932029           Prob > F      =  0.0000
    Residual |  198.311477   350  .566604221           R-squared     =  0.5971
-------------+------------------------------           Adj R-squared =  0.5948
       Total |  492.175535   352  1.39822595           Root MSE      =  .75273

------------------------------------------------------------------------------
     lsalary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       years |    .071318    .012505     5.70   0.000     .0467236    .0959124
     gamesyr |   .0201745   .0013429    15.02   0.000     .0175334    .0228156
       _cons |    11.2238    .108312   103.62   0.000     11.01078    11.43683
------------------------------------------------------------------------------
```

- $SSR = 198.311$, $R^2 = 0.597$

F statistic

- The test statistic is

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

- If c is the cutoff given a chosen significance level then reject H_0 if $F > c$
- For the baseball example:

$$F = \frac{198.311 - 183.186}{183.186} \cdot \frac{347}{3} \approx 9.55$$

- With $q = 3$, and $n - k - 1 = 347$, $c_{0.05} = 2.60$ and $c_{0.01} = 3.78$ so we can easily reject H_0 at both of these significance levels

The F distribution

F - Distribution ($\alpha = 0.01$ in the Right Tail)

Denominator Degrees of Freedom df_2 \ df_1	Numerator Degrees of Freedom 1	2	3	4	5	6	7	8	9
1	4052.2	4999.5	5403.4	5624.6	5763.6	5859.0	5928.4	5981.1	6022.5
2	98.503	99.000	99.166	99.249	99.299	99.333	99.356	99.374	99.388
3	34.116	30.817	29.457	28.710	28.237	27.911	27.672	27.489	27.345
4	21.198	18.000	16.694	15.977	15.522	15.207	14.976	14.799	14.659
5	16.258	13.274	12.060	11.392	10.967	10.672	10.456	10.289	10.158
6	13.745	10.925	9.7795	9.1483	8.7459	8.4661	8.2600	8.1017	7.9761
7	12.246	9.5466	8.4513	7.8466	7.4604	7.1914	6.9928	6.8400	6.7188
8	11.259	8.6491	7.5910	7.0061	6.6318	6.3707	6.1776	6.0289	5.9106
9	10.561	8.0215	6.9919	6.4221	6.0569	5.8018	5.6129	5.4671	5.3511
10	10.044	7.5594	6.5523	5.9943	5.6363	5.3858	5.2001	5.0567	4.9424
11	9.6460	7.2057	6.2167	5.6683	5.3160	5.0692	4.8861	4.7445	4.6315
12	9.3302	6.9266	5.9525	5.4120	5.0643	4.8206	4.6395	4.4994	4.3875
13	9.0738	6.7010	5.7394	5.2053	4.8616	4.6204	4.4410	4.3021	4.1911
14	8.8616	6.5149	5.5639	5.0354	4.6950	4.4558	4.2779	4.1399	4.0297
15	8.6831	6.3589	5.4170	4.8932	4.5556	4.3183	4.1415	4.0045	3.8948
16	8.5310	6.2262	5.2922	4.7726	4.4374	4.2016	4.0259	3.8896	3.7804
17	8.3997	6.1121	5.1850	4.6690	4.3359	4.1015	3.9267	3.7910	3.6822
18	8.2854	6.0129	5.0919	4.5790	4.2479	4.0146	3.8406	3.7054	3.5971
19	8.1849	5.9259	5.0103	4.5003	4.1708	3.9386	3.7653	3.6305	3.5225
20	8.0960	5.8489	4.9382	4.4307	4.1027	3.8714	3.6987	3.5644	3.4567
21	8.0166	5.7804	4.8740	4.3688	4.0421	3.8117	3.6396	3.5056	3.3981
22	7.9454	5.7190	4.8166	4.3134	3.9880	3.7583	3.5867	3.4530	3.3458
23	7.8811	5.6637	4.7649	4.2636	3.9392	3.7102	3.5390	3.4057	3.2986
24	7.8229	5.6136	4.7181	4.2184	3.8951	3.6667	3.4959	3.3629	3.2560
25	7.7698	5.5680	4.6755	4.1774	3.8550	3.6272	3.4568	3.3239	3.2172
26	7.7213	5.5263	4.6366	4.1400	3.8183	3.5911	3.4210	3.2884	3.1818
27	7.6767	5.4881	4.6009	4.1056	3.7848	3.5580	3.3882	3.2558	3.1494
28	7.6356	5.4529	4.5681	4.0740	3.7539	3.5276	3.3581	3.2259	3.1195
29	7.5977	5.4204	4.5378	4.0449	3.7254	3.4995	3.3303	3.1982	3.0920

Reasoning behind the F statistic

- Let $Z_i, i = 1, 2, \ldots, n$, be independent random variables, each distributed as standard normal
- Define new random variable as the sum of the squares of the Z_i:

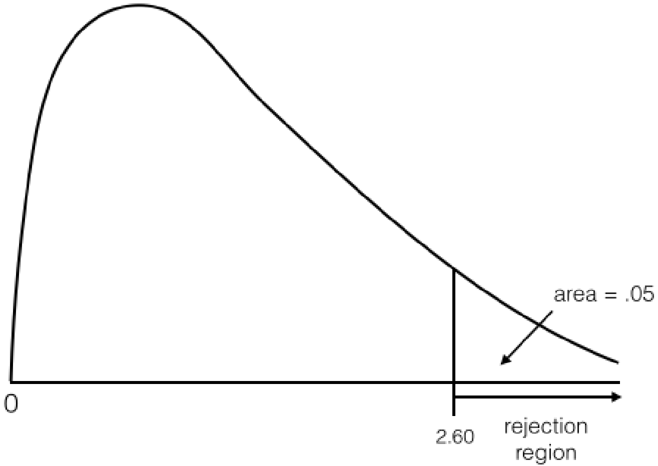
$$X = \sum_{i=1}^{n} Z_i^2$$

- X has a **chi-square distribution** with n **degrees of freedom**
- If X_1 and X_2 are independent then the random variable

$$F = \frac{X_1/k_1}{X_2/k_2}$$

  has an **F distribution** with (k_1, k_2) degrees of freedom
- Can be shown that $\frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)}$ is the ratio of two independent chi-square random variables divided by their respective degrees of freedom

5% critical value in an $F_{3,347}$ distribution



R-squared version of the F statistic

- Since $SSR_r = SST(1 - R_r^2)$ and $SSR_{ur} = SST(1 - R_{ur}^2)$:

$$F = \frac{(R_{ur}^2 - R_r^2)/q}{(1 - R_{ur}^2)/(n - k - 1)}$$

- For the baseball example:

$$F = \frac{0.6278 - 0.5971}{1 - 0.6278} \cdot \frac{347}{3} \approx 9.54$$

- Cannot be used for testing *all* linear restrictions

F statistic for overall significance of a regression

- Common set of exclusion restrictions involves testing the **overall significance of the regression**

$$\mathrm{H}_0 : \ \beta_1 = \beta_2 = \ldots = \beta_k = 0$$

- $R_r^2 = 0$ since none of the variation in y is being explained because there are no explanatory variables so F statistic becomes:

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

  where R^2 is the usual R-squared from the unrestricted model (i.e. R_{ur})
- If we fail to reject H_0 then there is no evidence that any of the independent variables help to explain y

Overall significance of baseball regression

```
. reg lsalary years gamesyr bavg hrunsyr rbisyr

      Source |       SS       df       MS              Number of obs =     353
-------------+------------------------------           F(  5,   347) =  117.06
       Model |  308.989208     5  61.7978416           Prob > F      =  0.0000
    Residual |  183.186327   347  .527914487           R-squared     =  0.6278
-------------+------------------------------           Adj R-squared =  0.6224
       Total |  492.175535   352  1.39822595           Root MSE      =  .72658

------------------------------------------------------------------------------
     lsalary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       years |   .0688626   .0121145     5.68   0.000     .0450355    .0926898
     gamesyr |   .0125521   .0026468     4.74   0.000     .0073464    .0177578
        bavg |   .0009786   .0011035     0.89   0.376    -.0011918     .003149
     hrunsyr |   .0144295    .016057     0.90   0.369    -.0171518    .0460107
      rbisyr |   .0107657    .007175     1.50   0.134    -.0033462    .0248776
       _cons |   11.19242   .2888229    38.75   0.000     10.62435    11.76048
------------------------------------------------------------------------------
```

- $F = \frac{R^2/k}{(1-R^2)/(n-k-1)} = \frac{0.6278/5}{(1-0.6278)/(347)} = 117.06$

Testing general linear restrictions with the F test

- Example: Test whether house price assessments are rational

$$\begin{aligned}\log(price) = &\beta_0 + \beta_1 \log(assess) + \beta_2 \log(lotsize) \\ &\beta_3 \log(sqrft) + \beta_4 bdrms + u\end{aligned}$$

where *assess* is the assessed value of the house

- Seek to test whether a 1% change in assessment is associated with a 1% change in price:

$$\mathrm{H}_0: \ \beta_1 = 1, \beta_2 = 0, \beta_3 = 0, \beta_4 = 0$$

Testing general linear restrictions with the F test

- Unrestricted regression:

$$\begin{aligned}\log(price) =& \beta_0 + \beta_1 \log(assess) + \beta_2 \log(lotsize) \\ & \beta_3 \log(sqrft) + \beta_4 bdrms + u\end{aligned}$$

- Restricted regression (tricky because we need to impose a non-zero restriction on β_1):

$$\begin{aligned}\log(price) =& \beta_0 + \log(assess) + u \\ \Rightarrow \log(price) - \log(assess) =& \beta_0 + u\end{aligned}$$

- Test statistic

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

- **Cannot** use the R^2 form of the F test since the dependent variable is different so SST will be different!

Testing general linear restrictions with the F test

```
. reg lprice lassess llotsize lsqrft bdrms

      Source |       SS       df       MS              Number of obs =      88
-------------+------------------------------           F(  4,    83) =   70.58
       Model |  6.19607473     4  1.54901868           Prob > F      =  0.0000
    Residual |  1.82152879    83  .02194613           R-squared     =  0.7728
-------------+------------------------------           Adj R-squared =  0.7619
       Total |  8.01760352    87  .092156362           Root MSE      =  .14814

------------------------------------------------------------------------------
      lprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lassess |   1.043065    .151446     6.89   0.000     .7418453    1.344285
    llotsize |   .0074379   .0385615     0.19   0.848    -.0692593    .0841352
      lsqrft |  -.1032384   .1384305    -0.75   0.458     -.378571    .1720942
       bdrms |   .0338392   .0220983     1.53   0.129    -.0101135    .0777918
       _cons |    .263743   .5696647     0.46   0.645    -.8692972    1.396783
------------------------------------------------------------------------------
```

- $SSR_{ur} = 1.82$

Testing general linear restrictions with the F test

```
. gen newvar = lprice - lassess

. reg newvar

      Source |       SS       df       MS              Number of obs =      88
-------------+------------------------------           F(  0,    87) =    0.00
       Model |           0     0           .           Prob > F      =       .
    Residual |  1.88014885    87  .021610906           R-squared     =  0.0000
-------------+------------------------------           Adj R-squared =  0.0000
       Total |  1.88014885    87  .021610906           Root MSE      =  .14701

------------------------------------------------------------------------------
      newvar |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |  -.0848135   .0156709    -5.41   0.000    -.1159612   -.0536658
------------------------------------------------------------------------------
```

- $SSR_r = 1.88$
- $F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} = \frac{(1.880-1.822)/4}{1.822/83} = 0.661$
- $c_{0.05} = 2.50$, which indicates that there is no evidence against the null hypothesis that assessed values are rational!

Reporting regression results

- Things that must be reported...
  1. Estimated OLS coefficients
     - The main coefficients should be interpreted somewhere in the paper
  2. Standard errors (usually beside or below the coefficient)
     - Usually preferable to t statistics since you can construct confidence intervals and test hypotheses other than H_0: $\beta_j = 0$
  3. R-squared
     - Provides a goodness-of-fit measure and makes calculation of F statistics for exclusion restrictions simple
  4. Number of observations

Testing the salary-benefits tradeoff

- If only a couple of models are being estimated then the results can be reported in equation form:
- Consider a model of teacher salaries and benefits

$$\log(\textit{salary}) = \beta_0 + \beta_1(b/s) + \textit{other factors}$$

- Testing the salary-benefits tradeoff is the same as the test of H_0: $\beta_1 = -1$ against H_1: $\beta_1 \neq -1$
  - A 1% or 0.01 rise in b/s should lead to a $-1 \times 0.01 = 0.01\%$ drop in *salary* if benefits and salary are equivalent
- Other factors: size of the school (*enroll*), staff per thousand students (*staff*), and measures of school dropout and graduation rates (*droprate* and *gradrate*)

Testing the salary-benefits tradeoff

- If only two variants of this model are estimated then two sample regression functions can be reported in equation form:

$$\widehat{\log(salary)} = 10.523 - 0.825(b/s) \tag{1}$$

$$\widehat{\log(salary)} = 10.884 - 0.605(b/s) + 0.0874\log(enroll) - 0.222\log(staff) \tag{2}$$

- When several equations are estimated, the coefficient estimates are reported in different columns

Testing the salary-benefits tradeoff

	(1) log(salary)	(2) log(salary)	(3) log(salary)
benefits/salary	-0.825**	-0.605**	-0.589**
	(0.200)	(0.165)	(0.165)
log(enroll)		0.087**	0.088**
		(0.007)	(0.007)
log(staff)		-0.222**	-0.218**
		(0.050)	(0.050)
school dropout rate, perc			-0.000
			(0.002)
school graduation rate, perc			0.001
			(0.001)
Constant	10.523**	10.844**	10.738**
	(0.042)	(0.252)	(0.258)
Observations	408	408	408
R^2	0.040	0.353	0.361

Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$

Identifying necessary components of a regression table

	(1)	(2)	(3)
	log(salary)	log(salary)	log(salary)
benefits/salary	$\boxed{-0.825^{**}}$	-0.605**	-0.589**
	$\boxed{(0.200)}$	(0.165)	(0.165)
log(enroll)		$0.087\boxed{^{**}}$	0.088**
		(0.007)	(0.007)
log(staff)		-0.222**	-0.218**
		(0.050)	(0.050)
school dropout rate, perc			-0.000
			(0.002)
school graduation rate, perc			0.001
			(0.001)
Constant	10.523**	10.844**	10.738**
	(0.042)	(0.252)	(0.258)
Observations	408	408	408
R^2	0.040	0.353	0.361

Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$

Testing the salary-benefits tradeoff

- Without controlling for any other factors, $\hat{\beta}_1$ is -0.825
  - t statistic is $\frac{-0.825+1}{0.200} = 0.875$, which is below below 1.96 so we cannot reject H_0
- Adding controls for school size and staff reduces the magnitude of the coefficient
  - t statistic is now $\frac{-0.605+1}{0.165} = 2.39$, which is above 1.96 so we can reject H_0

How to use columns

- Use columns to include control variables to demonstrate the robustness of your main coefficient (e.g. the coefficient on (b/s))
  - Add more important controls earlier to create a "triangle" shape in the bottom left-hand corner of the table
  - Inclusion of controls will indicate any omitted variables bias from excluding these controls
- Estimate the equation with different subsamples:
  - Different cohorts, genders, countries
  - Different time periods

Different columns for gender: Dependent variable is log(*wage*)

	(1) All	(2) Men	(3) Women
years of education	0.092**	0.096**	0.080**
	(0.007)	(0.009)	(0.010)
years potential experience	0.004*	0.008**	0.002
	(0.002)	(0.002)	(0.002)
years with current employer	0.022**	0.018**	0.010
	(0.003)	(0.004)	(0.005)
Constant	0.284**	0.322*	0.356*
	(0.104)	(0.139)	(0.141)
Observations	526	274	252
R^2	0.316	0.365	0.212

Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$

Different columns for different years: Dependent variable is log(*price*)

	(1) All	(2) 1978 only	(3) 1981 only
square footage of house	0.000**	0.000**	0.000**
	(0.000)	(0.000)	(0.000)
# rooms in house	0.031	0.051*	0.061*
	(0.024)	(0.024)	(0.029)
# bathrooms	0.235**	0.220**	0.293**
	(0.032)	(0.036)	(0.033)
Constant	10.176**	10.015**	10.304**
	(0.128)	(0.130)	(0.153)
Observations	321	179	142
R^2	0.527	0.561	0.704

Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$

Example from Levitt and Syverson (2008)

TABLE 2.—THE IMPACT OF AGENT-OWNERSHIP STATUS ON SALE PRICE AND TIME-TO-SALE

	(1)	(2)	(3)	(4)
	Dependent Variable: ln(Sale Price of Home)			
Coefficient on agent-owned home	0.048	0.042	0.038	0.037
(Standard error)	(0.004)	(0.004)	(0.003)	
R^2	0.856	0.886	0.896	0.958
	Variable: Days to Sale			
Coefficient on agent-owned home	16.89	11.03	10.25	9.47
(Standard error)	(2.42)	(2.40)	(2.39)	(2.25)
R^2	0.123	0.130	0.139	0.384
Controls included:				
City × year interactions	Yes	Yes	Yes	Yes
Basic house characteristics	Yes	Yes	Yes	Yes
Indicators of house quality	No	Yes	Yes	Yes
Keywords in description	No	No	Yes	Yes
Block fixed effects	No	No	No	Yes
"Excess return" of agent assuming a 20% annual discount rate	0.039	0.036	0.032	0.032

Notes: Regression coefficients are reported in the table, along with standard errors in parentheses. Results are based on a sample of 98,038 single-family home sales in 34 Cook County, Illinois, suburbs over the period 1992–2002. The dependent variable in the top panel of the table is the natural log of the sale price; the dependent variable in the bottom panel is the number of days on the market. Each coefficient reported in the table is from a separate regression. The other variables included in each specification are noted in the table, but the coefficients on these other variables are not reported here (table 3 presents a subset of coefficient estimates for these controls). See the appendix for a complete list. The table's bottom row reports the implied "excess return" accruing to agents selling their own homes, computed as the additional price received for a home adjusted for the extra time on the market, under the assumption of a 20% annual discount rate.

Quadratic functional forms

- Quadratic functional forms allow you to incorporate non-linear effects when you have zero or negative values ($\log(x)$ doesn't work if $x \leq 0$)
- Consider the following wage function:

$$\widehat{wage} = 3.73 + 0.298 exper - 0.0061 exper^2$$

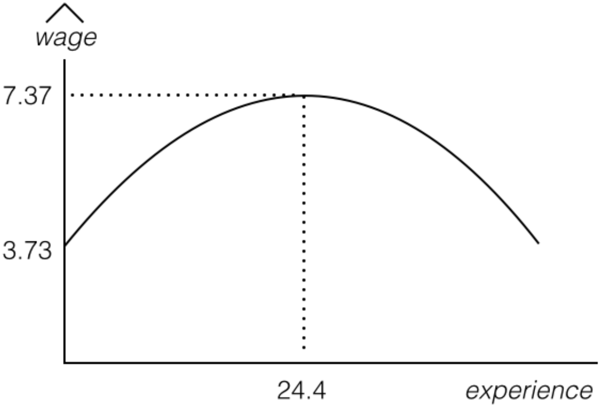
- Marginal effect of experience:

$$\frac{d\widehat{wage}}{dexper} = 0.298 - 2 \cdot 0.0061 exper$$

- First year of experience increases the wage by $0.298 - 2 \times 0.0061 \times 0 \approx 30$ cents.
- The second year by $0.298 - 2 \times 0.0061 \times 1 \approx 29$ cents.

Wage maximum with respect to work experience

$$\frac{d\widehat{wage}}{dexper} = 0 \Rightarrow 0.298 - 2 \cdot 0.0061 exper = 0 \Rightarrow exper^* = \frac{0.298}{0.0122} \approx 24.4$$



Wage maximum with respect to work experience

- Does this mean the return to experience becomes negative after 24.4 years?
- Not necessarily. It depends on how many observations in the sample lie to the right of the turning point (there may be none)
  - In fact, in this dataset, 28% of the observations lie to the right suggesting there may be a specification problem (e.g. omitted variables)

Effects of pollution on housing prices

- Suppose (the log of) house price is estimated to be:

$$\widehat{\log(price)} = 13.39 - 0.902 \log(nox) - 0.087 \log(dist) \\ - 0.545 rooms + 0.062 rooms^2 - 0.048 stratio$$

  where *nox* is nitrogen oxide in air, *dist* is distance from employment centers, and *stratio* is student/teacher ratio

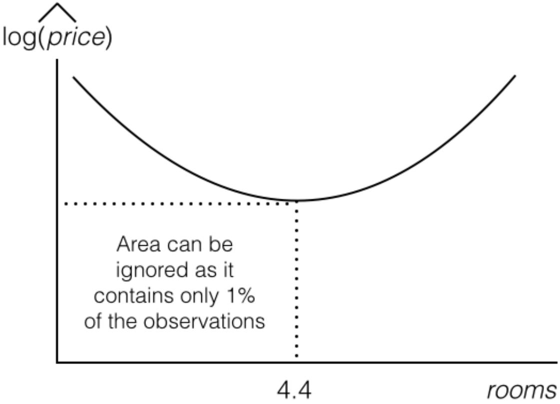
- Marginal effect of rooms:

$$\frac{d\widehat{\log(price)}}{drooms} = \frac{\% dprice}{drooms} = -0.545 + 0.124 rooms$$

  - Increase rooms from 5 to 6: $-0.545 + 0.124 \times 5 = 7.5\%$
  - Increase rooms from 6 to 7: $-0.545 + 0.124 \times 6 = 19.9\%$

Effects of pollution on housing prices

$$\frac{d\widehat{\log(price)}}{drooms} = 0 \Rightarrow rooms* = \frac{0.545}{0.124} \approx 4.4$$



Next week

- Further Issues (Chapter 6) and Midterm Revision