ECO375 Tutorial 5
More Inference

Matt Tudball

University of Toronto Mississauga

October 26, 2017

Including A Regressor: Trade-Offs

- It is not always clear when or when not to include an additional explanatory variable in a regression model.
- There is often a trade-off between bias and variance.
- Recall that the formula for the variance of $\hat{\beta}_j$ from the partialling out interpretation is

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

where

- σ^2 is the variance of the error term u.
- SST_j is the total sample variance in x_j

$$SST_j = \sum_{i=1}^{2}(x_j - \bar{x})^2$$

- R_j^2 is the R^2 from the regression of x_j on all the other independent variables.

Including A Regressor: Trade-Offs

- Also recall that the bias in $\tilde{\beta}_1$ when the explanatory variable x_k has been omitted can be written as

$$\tilde{\beta}_1 - \hat{\beta}_1 = \hat{\beta}_k \tilde{\delta}_1$$

  and the expected bias can be written as $\text{Bias}(\tilde{\beta}_1) = \beta_k \tilde{\delta}_1$.
- The variance of $\tilde{\beta}_1$ can be written as

$$\text{Var}\left(\tilde{\beta}_1\right) = \frac{\tilde{\sigma}^2}{SST_1(1-\tilde{R}_1^2)}$$

  where $\tilde{R}_1^2$ is the R^2 from the regression of x_1 on all the independent variables except for x_k and $\tilde{\sigma}^2$ is the variance of $\tilde{u} = u + \beta_k x_k$
- So what happens when we include x_k in the regression?

Including A Regressor: Trade-Offs

- Let's break it down into several cases.

1. $\hat{\beta}_k = 0$ and $\tilde{\delta}_1 = 0$
   In this case the bias from omitting x_k is zero and x_k contributes nothing to the variance of $\tilde{\beta}_1$ so there no benefit or drawback from including x_k.
2. $\hat{\beta}_k \neq 0$ and $\tilde{\delta}_1 = 0$
   In this case the bias from omitting x_k is zero. If we still include x_k in the regression, however, $\tilde{R}_1^2$ will remain unchanged since $\tilde{\delta}_1 = 0$ while $\tilde{\sigma}^2$ will decrease since we are pulling the variance related to x_k out of the error term. This will reduce the variance of our estimate of β_1.

Including A Regressor: Trade-Offs

1. $\hat{\beta}_k = 0$ and $\tilde{\delta}_1 \neq 0$ In this case the bias from omitting x_k is zero. Since $\tilde{\delta}_1 \neq 0$, if we include x_k in the regression we still increase $\tilde{R}_1^2$ and therefore increase the variance of $\tilde{\beta}_1$, leading to less precise estimates.
2. $\hat{\beta}_k \neq 0$ and $\tilde{\delta}_1 \neq 0$
   In this case there is non-zero bias from omitting x_k. While including x_k in the regression will reduce $\tilde{\sigma}^2$ it will also increase $\tilde{R}_1^2$. Here we must make a judgement call based on the relative loss of precision compared to a reduction in bias.

- What is the main takeaway here? When deciding whether or not to including an additional explanatory variance, frame your decision in terms of the **bias-variance trade-off**.

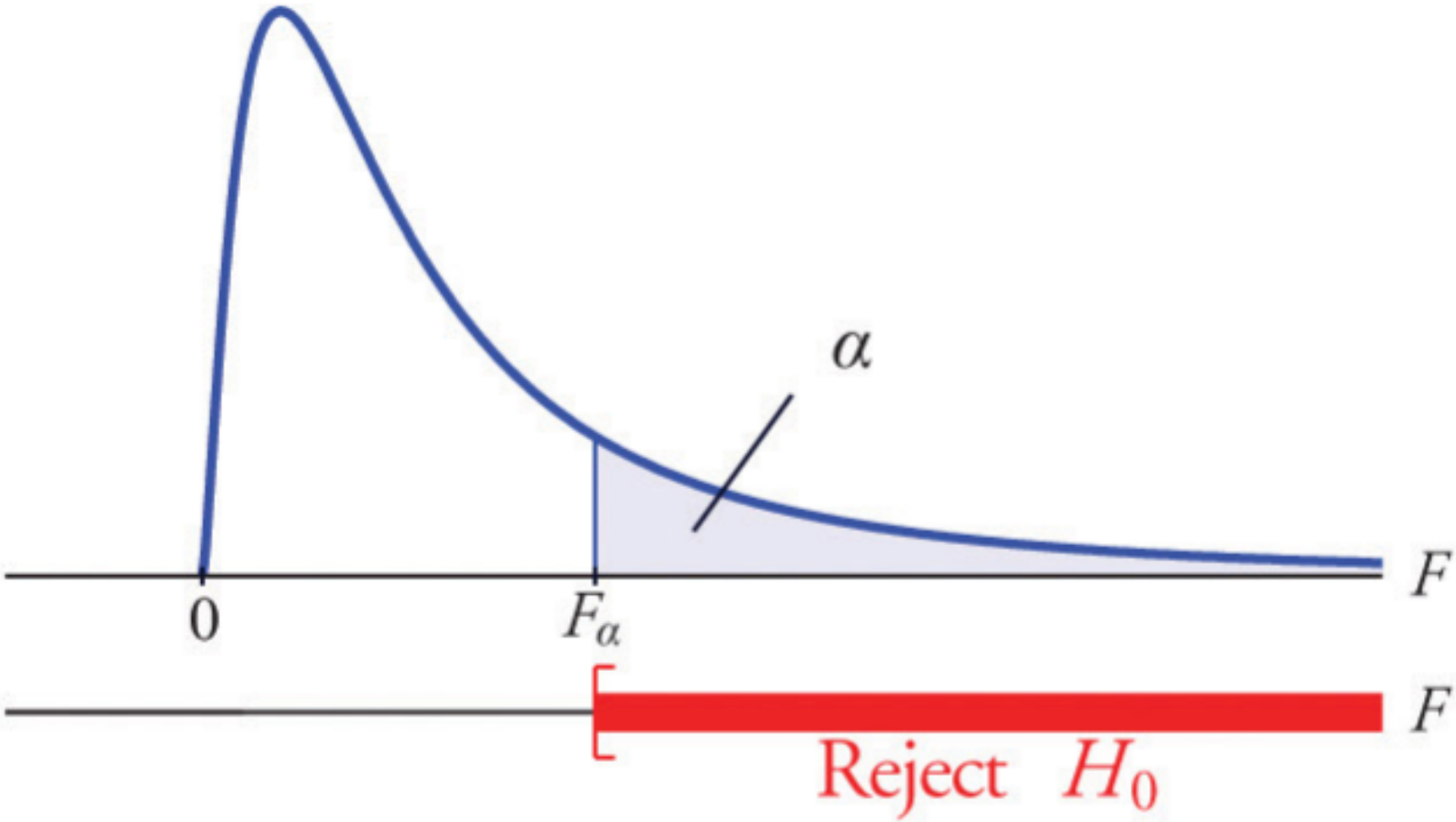
Hypothesis Testing: F-test

- Suppose we want to test hypotheses of the form $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ or even $H_0 : \beta_1 = \beta_2, \beta_3 = 2$.
- For hypotheses of this form we will use an **F-test** (as we discussed in last week's tutorial). The **F-statistic** takes the form,

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

  - SSR_r denotes the sum of squared residuals of the **restricted model** in which we assume the null hypothesis is true. R^2_r is the R^2 from this restricted model.
  - SSR_{ur} denotes the sum of squared residuals of the **unrestricted model** in which we make no assumptions over the coefficients in the null hypothesis. R^2_{ur} is the R^2 from this unrestricted model.
  - q is the number of restrictions being tested. In the example above, it is 3.
  - $n - k - 1$ is the degrees of freedom of the unrestricted model.

Hypothesis Testing: F-test



Hypothesis Testing: F-table

F - Distribution ($\alpha = 0.05$ in the Right Tail)

df_2 \ df_1	Numerator Degrees of Freedom								
Denominator Degrees of Freedom	1	2	3	4	5	6	7	8	9
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.385
3	10.128	9.5521	9.2766	9.1172	9.0135	8.9406	8.8867	8.8452	8.8123
4	7.7086	9.9443	6.5914	6.3882	6.2561	6.1631	6.0942	6.0410	6.9988
5	6.6079	5.7861	5.4095	5.1922	5.0503	4.9503	4.8759	4.8183	4.7725
6	5.9874	5.1433	4.7571	4.5337	4.3874	4.2839	4.2067	4.1468	4.0990
7	5.5914	4.7374	4.3468	4.1203	3.9715	3.8660	3.7870	3.7257	3.6767
8	5.3177	4.4590	4.0662	3.8379	3.6875	3.5806	3.5005	3.4381	3.3881
9	5.1174	4.2565	3.8625	3.6331	3.4817	3.3738	3.2927	3.2296	3.1789
10	4.9646	4.1028	3.7083	3.4780	3.3258	3.2172	3.1355	3.0717	3.0204
11	4.8443	3.9823	3.5874	3.3567	3.2039	3.0946	3.0123	2.9480	2.8962
12	4.7472	3.8853	3.4903	3.2592	3.1059	2.9961	2.9134	2.8486	2.7964
13	4.6672	3.8056	3.4105	3.1791	3.0254	2.9153	2.8321	2.7669	2.7144
14	4.6001	3.7389	3.3439	3.1122	2.9582	2.8477	2.7642	2.6987	2.6458
15	4.5431	3.6823	3.2874	3.0556	2.9013	2.7905	2.7066	2.6408	2.5876
16	4.4940	3.6337	3.2389	3.0069	2.8524	2.7413	2.6572	2.5911	2.5377
17	4.4513	3.5915	3.1968	2.9647	2.8100	2.6987	2.6143	2.5480	2.4943
18	4.4139	3.5546	3.1599	2.9277	2.7729	2.6613	2.5767	2.5102	2.4563
19	4.3807	3.5219	3.1274	2.8951	2.7401	2.6283	2.5435	2.4768	2.4227
20	4.3512	3.4928	3.0984	2.8661	2.7109	2.5990	2.5140	2.4471	2.3928
21	4.3248	3.4668	3.0725	2.8401	2.6848	2.5727	2.4876	2.4205	2.3660
22	4.3009	3.4434	3.0491	2.8167	2.6613	2.5491	2.4638	2.3965	2.3419
23	4.2793	3.4221	3.0280	2.7955	2.6400	2.5277	2.4422	2.3748	2.3201
24	4.2597	3.4028	3.0088	2.7763	2.6207	2.5082	2.4226	2.3551	2.3002
25	4.2417	3.3852	2.9912	2.7587	2.6030	2.4904	2.4047	2.3371	2.2821
26	4.2252	3.3690	2.9752	2.7426	2.5868	2.4741	2.3883	2.3205	2.2655
27	4.2100	3.3541	2.9604	2.7278	2.5719	2.4591	2.3732	2.3053	2.2501
28	4.1960	3.3404	2.9467	2.7141	2.5581	2.4453	2.3593	2.2913	2.2360
29	4.1830	3.3277	2.9340	2.7014	2.5454	2.4324	2.3463	2.2783	2.2229
30	4.1709	3.3158	2.9223	2.6896	2.5336	2.4205	2.3343	2.2662	2.2107
40	4.0847	3.2317	2.8387	2.6060	2.4495	2.3359	2.2490	2.1802	2.1240
60	4.0012	3.1504	2.7581	2.5252	2.3683	2.2541	2.1665	2.0970	2.0401
120	3.9201	3.0718	2.6802	2.4472	2.2899	2.1750	2.0868	2.0164	1.9588
∞	3.8415	2.9957	2.6049	2.3719	2.2141	2.0986	2.0096	1.9384	1.8799

In-Class Exercise 1

- Download the dataset LAWSCH85.dta from my website (matthewtudball.com). This is a dataset of law schools. Suppose we want to test the following model to explain the salaries of law school graduates. See Chapter 4, C2.

$$\begin{aligned}\log(salary) =& \beta_0 + \beta_1 LSAT + \beta_2 GPA + \beta_3 \log(libvol) + \beta_4 \log(cost) \\ &+ \beta_5 rank + u\end{aligned}$$

  Suppose you want to test the null hypothesis $H_0 : \beta_1 = 0, \beta_2 = 0$ (i.e. that test scores of incoming students have no effect on salaries).

1. First drop missing observations of *GPA* and *LSAT* by the command `drop if missing(GPA) | missing(LSAT).`
2. Begin by running this regression in Stata and then type the command `test LSAT GPA.`
   This will test the null hypothesis given above. Record the F-statistic. Also record the SSR and call it SSR_{ur}.

In-Class Exercise 1

3. Now run the restricted regression which takes the form.

$$\log(salary) = \beta_0 + \beta_3 \log(libvol) + \beta_4 \log(cost) + \beta_5 rank + u$$

   Record the SSR and call it SSR_r.

4. Look at the formula for the F-statistic on Slide 6. What is q in this exercise? What are n and k? Record them.

5. Now plug SSR_{ur}, SSR_r, q, n and k into the formula and calculate the F-statistic. Is it consistent with the one that Stata calculated in 2)?

In-Class Exercise 2

- Using the same data as before, consider a more complicated null hypothesis of the form $H_0 : \beta_1 = \beta_2$. We can interpret this as saying the effect of *GPA* on salary is the same as the effect of *LSAT*.

1. Note that under the null hypothesis the regression takes the form

$$\begin{aligned}\log(salary) =& \beta_0 + \beta_1(LSAT + GPA) + \beta_3 \log(libvol) + \beta_4 \log(cost) \\ &+ \beta_5 rank + u\end{aligned}$$

   This is the restricted regression. Generate a new variable by running the command `gen GPALSAT = GPA + LSAT`
2. Run this regression and record the SSR and call it SSR_r.

In-Class Exercise 2

3. Note that SSR_{ur} is the same as in the last exercise, as are n and k. Is q different? If so, what is q?
4. Calculate the F-statistic by plugging in SSR_{ur}, SSR_r, q, n and k into the formula.
5. Looking at the F-table in Slide 8, do you reject the null hypothesis at 5% significance?

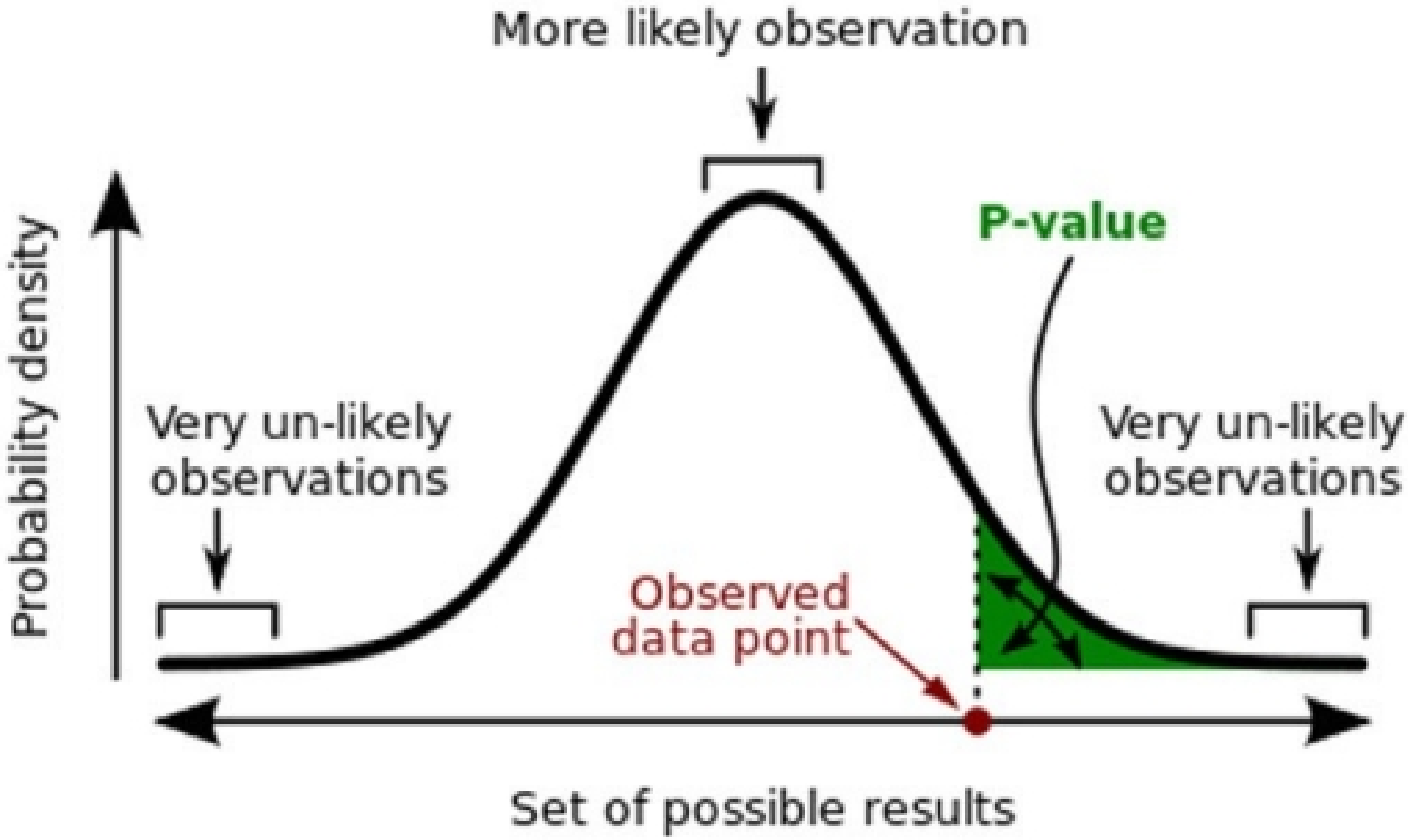
p-values

- Let's return to the simple t-tests we talked about last week.
- Let's consider the two-sided hypothesis test $H_0 : \beta_1 = \gamma$ against the alternative hypothesis $H_0 : \beta_1 \neq \gamma$. We will test this using a t-test and the t-statistic takes the form

$$t = (\hat{\beta}_1 - \gamma)/\widehat{se}(\hat{\beta}_1)$$

- At $\alpha\%$ significance we would reject the null hypothesis if $|t| > t_{n-k-1,\alpha/2}$ where $t_{n-k-1,\alpha/2}$ is the critical value from the t-distribution with $n - k - 1$ degrees of freedom.

p-values: Visualisation



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

p-values

- Let's suppose we have our t-statistic $t = \hat{\beta}_1 / \widehat{se}(\hat{\beta}_1)$ for a null hypothesis $H_0 : \beta_1 = 0$ against a two-sided alternative $H_0 : \beta_1 \neq 0$.
- Then the **p-value** is the probability, under the null hypothesis, of observing a t-statistic more extreme than the one we actually observed. To put this concretely,

$$p\text{-value} = \mathbb{P}(|T_{n-k-1}| > |t|)$$

where T_{n-k-1} is a random variable following the t-distribution on $n - k - 1$ degrees of freedom.

p-values

- We can use p-values to do hypothesis testing. If p-value $< \alpha$ then this says that the probability of observing a test statistic more extreme than ours is smaller than our significance level.
- Recall that our critical value $t_{n-k-1,\alpha/2}$ is defined as the value on the t-distribution such that $\mathbb{P}(|T_{n-k-1}| > |t_{n-k-1,\alpha/2}|) = \alpha$.
- Therefore, if p-value $< \alpha$ it must mean that

$$\alpha = \mathbb{P}(|T_{n-k-1}| > |t_{n-k-1,\alpha/2}|) < p\text{-value} = \mathbb{P}(|T_{n-k-1}| > |t|)$$

- This will only be true if $|t| > |t_{n-k-1,\alpha/2}|$, meaning that we will reject the null hypothesis.
- What is the takeaway here? Checking whether p-value $< \alpha$ is equivalent to checking whether $|t| > |t_{n-k-1,\alpha/2}|$.

p-values

- If the two are equivalent, why use p-values at all?
- p-values are useful because they give us a sense of the confidence with which we are rejecting the null hypothesis.
- If the p-value is something very small like 0.001, then we know that, under the null hypothesis, there is a 0.1% chance of observing a result more extreme than the one we actually observed.
- This provides us with more compelling evidence against the null than if the p-value were, for example, 0.04, indicating that there is a 4% chance of observing a result more extreme than the one we actually observed.

p-values: Calculation

- How do we go about calculating p-values?
- Recall that for a two-sided test, p-value $= \mathbb{P}(|T_{n-k-1}| > |t|)$.
- Example: Suppose we know that our t-statistic $t = 1.74$, $n = 36$, $k = 5$ and $\alpha = 0.05$. Then we need to calculate

$$\begin{aligned}\mathbb{P}(|T_{30}| > |1.74|) &= \mathbb{P}(T_{30} > |1.74|) + \mathbb{P}(T_{30} < -|1.74|) \\ &= 2\,\mathbb{P}(T_{30} > |1.74|)\end{aligned}$$

  with the last equality coming from the symmetry of the t-distribution.
- We cannot calculate this probability from the t-table. We will need to use a computational package like Stata. For this example, we can use the command `display 2*ttail(30,1.74)`.
- This returns a p-value of 0.092. Since $0.092 > 0.05$ we do not reject the null at 5% significance.
- Note that in general, for degrees of freedom $n - k - 1$ and t-statistic t you can calculate the probability of observing T_{n-k-1} greater than t by `display ttail(n-k-1,t)`.

p-values: F-tests

- Calculating p-values for F-statistics is not too dissimilar to calculating them for t-statistics. The general form of the p-value is

$$p\text{-value} = \mathbb{P}(F_{q,n-k-1} > F)$$

  where $F_{q,n-k-1}$ is a random variable from the F-distribution with degrees of freedom $(q,\ n-k-1)$ and F is the F-statistic from some hypothesis test.
- Not that we do not need an absolute value here since an F-test is always one-sided (see Slide 7).
- Example: Suppose we have an F-statistic of 4, $q = 3$ and $n-k-1 = 30$.
  Then p-value $= \mathbb{P}(F_{3,30} > 4)$. We can calculate this using the Stata command `display Ftail(3,30,4).`
- This returns a p-value of 0.017 and so we would reject the null hypothesis at 5% significance but not at 1% significance.

In-Class Exercise 3

- Now load in the dataset 401KSUBS.dta from my website (matthewtudball.com). This dataset contains information on net financial wealth (*nettfa*), age of the survey respondent (*age*), annual family income (*inc*), family size (*fsize*) and participation in certain pension plans for people in the United States. The wealth and income variables are both recorded in thousands of dollars. See Chapter 4, C8.

1. Suppose we are only interested in single person households. Drop all households except those for which $fsize = 1$.
2. Use OLS to estimate the model:

$$nettfa = \beta_0 + \beta_1 inc + \beta_2 age + u$$

What is the coefficient on *inc*? What is the *p*-value on *inc*? Using only the *p*-value in the Stata output, would you reject the null hypothesis that $\beta_1 = 0$?

In-Class Exercise 3

3. Does the intercept from the regression in the previous regression have an interesting meaning? Recall that this is the predicted value of *nettfa* when $inc = 0$ and $age = 0$.
4. **Tough Question:** Find the p-value for the test $H_0 : \beta_2 = 1$ against $H_1 : \beta_2 < 1$. Do you reject H_0 at the 1% significance level?
   Hint 1: Remember the command in Slide 18 `ttail`.
   Hint 2: Keep in mind that this is a one-sided test. What do you need to calculate for a one-sided test? Look at the derivation on Slide 18 for inspiration here.

In-Class Exercise 4

- Now load in the dataset DISCRIM.dta from my website (matthewtudball.com). This is a dataset on fast food restaurants where *psoda* is the price of a medium soda; *prpblck* is the proportion of people in the surrounding zip code who are black; *income* is median income in the surrounding zip code; and *prppov* is the proportion of people in the surrounding zip code who are in poverty. Use OLS to estimate the model:

$$\log(psoda) = \beta_0 + \beta_1 prpblck + \beta_2 \log(income) + \beta_3 prppov + u$$

Use only Stata output to answer these questions (i.e. no manual calculations). See Chapter 4, C9.

1. Is $\hat{\beta}_1$ statistically different from 0 at the 5% level against a two-sided alternative? What about at the 1% level?

In-Class Exercise 4

2. What is the correlation between log(*income*) and *prppov*? Is each variable statistically significant in any case? Report the two-sided *p*-values.
3. To the original regression in Slide 21, add the variable log(*hseval*). Interpret its coefficient and report the two-sided *p*-value for $H_0 : \beta_{\log(hseval)} = 0$. *hseval* is median housing value in the surrounding zipcode.
4. In the regression in 3), what happens to the individual statistical significance of log(*income*) and *prppov*? Are these variables jointly significant? (Compute a *p*-value). How do you interpret your answers?
5. Given the results of the previous two regressions, which one would you report as most reliable in determining whether racial composition of a zip code influences local fast food prices?