

Tonight you will learn to run a simple R script. R can be run on Linux, Windows, and Mac machines; it can be run as batch files or on a variety of GUIs. In this class we will use the standard Windows GUI, which is adequate for most purposes. For large data sets, I usually run R in batch mode on the economics department's linux workstation (beast). At home, on my Ubuntu machine, I use a GUI called [Rstudio](#).

### Getting help for R

There are two good online search engines for R:

1. <http://finzi.psych.upenn.edu/nmz.html>
2. <http://www.rseek.org/>

Using the Windows GUI, one can access help pages for a function in any of the *loaded* packages by typing a question mark before the name of the function:

```
?lm
```

To search for a keyword in the loaded packages' help pages, use two question marks:

```
??lm
```

### Using R to read or write data

The "load" command allows one to bring in an R format data set (a "workspace"):

```
load(file="exampdat.Rdata", .GlobalEnv)
```

one can write an R data set using the "save" command:

```
save(exampdat, file="exampdat.Rdata")
```

The package "foreign" provides capability to read or write in a number of popular formats. This is the "csv" format, which can be directly opened by spreadsheet software:

```
uu<-read.csv("exampdat.csv")
```

```
write.csv(uu, file="exampdat.csv")
```

The command "read.table" can be modified to read tab-delimited data; "read.lines" is often useful for pulling in large amounts of text for parsing into variables; the package "xml" will scrape html tables from the web.

---

## HYPOTHESIS TESTING

### Steps for conducting a hypothesis test:

- 1) Set up a null hypothesis (i.e., posit that the true value is equal to a specific number).
- 2) Create a test statistic.
- 3) Make a decision rule (i.e., reject the null hypothesis if the test statistic exceeds some cutoff value).

### P-Value, Critical Value, Size of Test

- The *size of test* is the probability that you are rejecting the null hypothesis when in fact it is true.
- The *critical value* of a t-statistic or f-statistic is constructed assuming a certain size of test (usually 0.05).
- The *p-value* gives the size of test at which the estimated t-statistic or f-statistic becomes the critical value.

### t-tests for one parameter

- 1)  $H_0: b = b_{\text{hypothesized}}$
- 2)  $t\text{-stat} = (b_{\text{estimated}} - b_{\text{hypothesized}}) / \text{standard error}(b_{\text{estimated}})$
- 3) Reject  $H_0$  if  $|\text{abs}(t\text{-stat})| > t\text{-critical}$

### t-tests for linear combination of parameters

- 1)  $H_0: b_1 + b_2 = b_{\text{hypothesized}}$
- 2)  $t\text{-stat} = (b_{1,\text{estimated}} + b_{2,\text{estimated}} - b_{\text{hypothesized}}) / \text{standard error}(b_{1,\text{estimated}} + b_{2,\text{estimated}})$
- 3) Reject  $H_0$  if  $|\text{abs}(t\text{-stat})| > t\text{-critical}$

### F-tests for group of parameters

- 1)  $H_0: b_1 = b_2 = 0$
- 2)  $F\text{-stat} = ((\text{error sum of squares in restricted regression} - \text{error sum of squares in unrestricted regression}) / \text{number of restrictions}) / (\text{error sum of squares in unrestricted regression} / \text{degrees of freedom in unrestricted regression})$
- 3) Reject  $H_0$  if  $F\text{-stat} > F\text{-critical}(\text{numerator degrees of freedom} = \text{number of parameters set equal to zero}; \text{denominator degrees of freedom set equal to degrees of freedom in the unrestricted regression})$

### F-test to Drop Irrelevant Independent variables

Create a model to explain variation in home values. You should follow these steps:

- 1) Run a regression in which you include all the independent variables that you think—*a priori*—are relevant. Call this the **unrestricted** regression.
- 2) Store the sum of squared residuals and the degrees of freedom from this regression.
- 3) Make a note of the variables which have a p-value above 0.10.
- 4) Set up a new regression, which omits all those independent variables with a high p-value
- 5) Store the sum of squared residuals and the degrees of freedom from this regression. Call this the **restricted** regression.
- 6) Use your stored values to carry out a hypothesis test
  - a) Null Hypothesis: the *omitted* variables do *not* belong in the model
  - b) Test Statistic: F-test

- c) Decision rule: if the F-statistic is high enough, then *reject* the null hypothesis
  - i) How do we determine if the F-statistic is high enough? If it *exceeds* the critical value.
  - ii) How do we calculate the critical value? Set it equal to an F-statistic with numerator degrees of freedom equal to the number of omitted variables and denominator degrees of freedom equal to the degrees of freedom in the unrestricted regression. Set the size of test equal to .05.

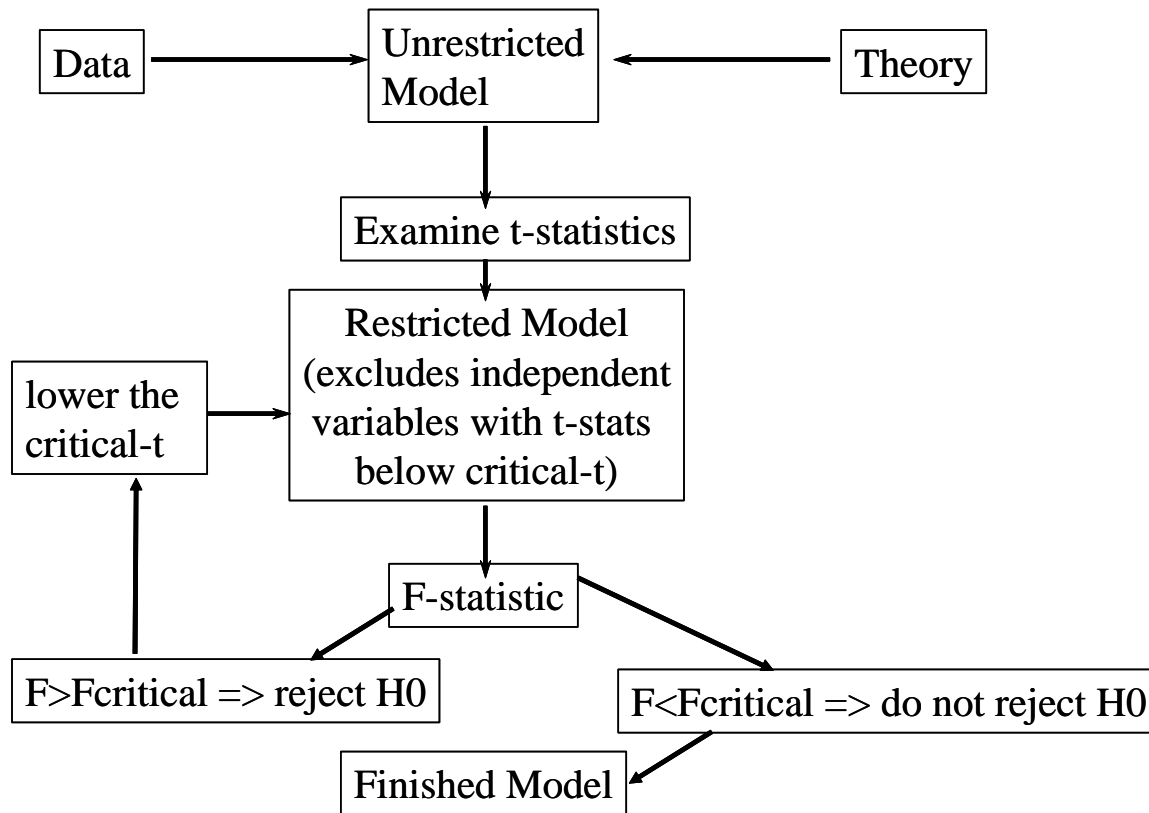
identifies these insignificant coefficients by examining the t-test, given in the `summary(lm(...))` output; the null hypothesis of the t-test is that the coefficient equals zero. If the p-value is less than 0.1 then you can reject the null hypothesis.

- Remove the insignificant variables from the model and perform an F-test to see if these independent variables jointly fail to explain the variation in the dependent variable. The null hypothesis is that these variables do *not* belong in the model. If the p-value is less than 0.1 then you can reject the null hypothesis.

**General modeling procedure:**

- In building a model, one begins with theory, selecting independent variables that theory suggests explain the variation in the dependent variable.
- The first (unrestricted) regression will usually show that some coefficients are not different from zero. One

**Schema of General Model-Building Strategy**



**Homework assignment:** The homework is due by noon next Thursday (this gives me time to look at it before class).

- 1) Review any material you might have from previous courses on hypothesis testing and regression analysis. (nothing to be turned in)
- 2) Build a model explaining home prices in Williamson County. Your group may wish to analyze a submarket (e.g., all homes costing between \$140,000 and \$190,000)—just be sure to make the submarket broad enough to include at least 1,000 observations. Be creative and use as many independent variables as you think reasonable. Remove the insignificant variables following the procedure outlined above. Use the help files to find some new feature of R and include this in your program. Write a half page essay describing what your results mean. Write comments that explain what each step in your R program does. Turn in to me your edited output, your commented program, and your half page essay.
- 3) Install R on your personal computer. Go to <http://cran.r-project.org/> and find the “binaries” for the operating system on your computer. Make a directory on your personal computer for your R work related to this class. Copy to this directory the data and R program used in this first class. Change the `setwd()` command in the R program to your directory name. Make sure that you can run the program on your own machine. (Nothing to turn in).