PUBH 501 Biostatistics

STATA: MULTIPLE LINEAR REGRESSION AND CONFOUNDING

Overview

- Multiple linear regression
  - Building a model
  - Assessing model fit
- Confounding

Data

- Using Stata dataset: load data using the following code

```
sysuse auto
```

- Data on automobiles from 1978. Mixture of variable types
- Outcome of interest is gas mileage (mpg)
- Is car head room associated with gas mileage (mpg)?
- …or other features of the car?

Multiple linear regression

Multiple linear regression

- Extension of linear regression
- Looking for a linear relationship between an outcome and multiple independent variables
- Control for multiple independent variables
- $Y = mx + b$ (univariate linear regression)
- $Y = m_1x_1 + m_2x_2 \ldots\ldots + b$ (multivariate linear regression)

-regress- command

- Linear regression is run using the –regress- command
- `regress mpg headroom foreign`
- Where the first variable is the outcome, or the dependent variable
- The next variables are the independent variables in any order

Factor notation

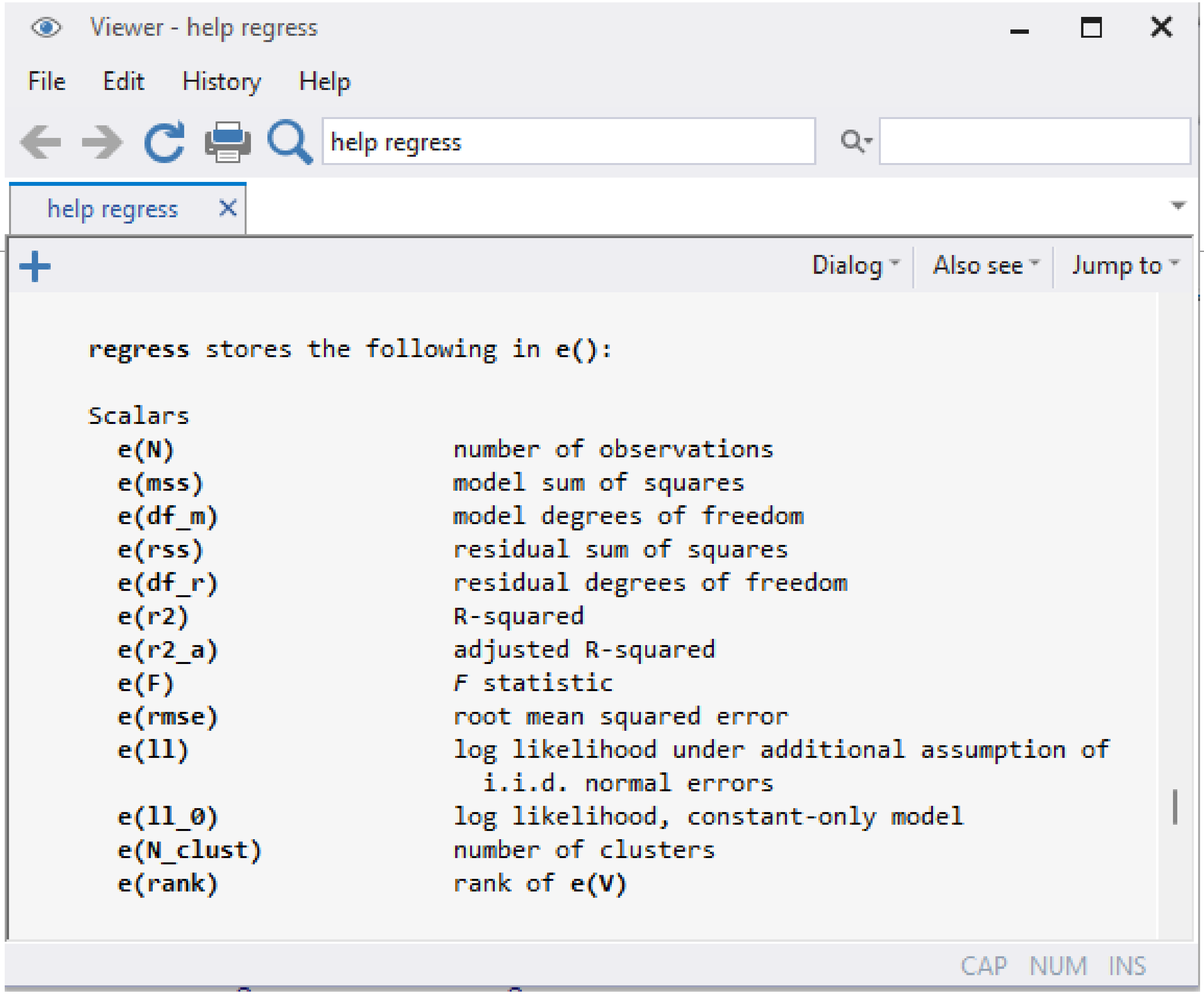
- Can use continuous, categorical, and binary independent variables in regression
- With categorical variables, you can tell Stata they are categorical, otherwise it treats them as continuous
- Do this with factor notation
  - Add “i.” to the beginning of the variable name in the regression command

Comparing models

- Make a table with R^2, adjusted R^2, F statistic and associated df, and coefficients
- Store model estimates using: `estimates store`
- `regress mpg headroom foreign`
  - `estimates store model1`
- `regress mpg headroom foreign weight`
  - `estimates store model2`
- Create table of estimates using: `estimates table`
- `estimates table model1 model2, b(%4.3f) p(%4.3f) stats(r2 r2_a F df_m df_r)`

b(%4.3f) for beta coeff. Keeps 4 digits to the left and 3 to the right of the decimal place: xxxx.xxx

p(%4.3f) for p-value: xxxx.xxx



```
regress stores the following in e():

Scalars
  e(N)                   number of observations
  e(mss)                 model sum of squares
  e(df_m)                model degrees of freedom
  e(rss)                 residual sum of squares
  e(df_r)                residual degrees of freedom
  e(r2)                  R-squared
  e(r2_a)                adjusted R-squared
  e(F)                   F statistic
  e(rmse)                root mean squared error
  e(ll)                  log likelihood under additional assumption of
                           i.i.d. normal errors
  e(ll_0)                log likelihood, constant-only model
  e(N_clust)             number of clusters
  e(rank)                rank of e(V)
```

Variable	model1	model2	model3	model4	model5
headroom	-2.232	-0.219	-0.192	-0.046	-0.086
	0.003	0.687	0.732	0.940	0.883
foreign	3.740	-1.655	-1.803	-3.129	-2.934
	0.007	0.131	0.172	0.047	0.034
weight		-0.006	-0.007	-0.006	-0.006
		0.000	0.000	0.000	0.000
price			0.000	0.000	
			0.837	0.788	
rep78					
2				-0.240	-0.148
				0.934	0.959
3				-0.008	0.093
				0.998	0.972
4				0.715	0.801
				0.798	0.771
5				3.861	3.973
				0.194	0.174
_cons	26.866	41.993	42.158	40.855	40.533
	0.000	0.000	0.000	0.000	0.000

legend: b/p

estimates table model1 model2 model3 model4 model5, b(%7.3f) p(%4.3f) stats(r2 r2_a F df_m df_r)

```
    Variable | model1    model2    model3    model4    model5
-------------+--------------------------------------------------
          r2 |   0.252     0.663     0.664     0.695     0.695
        r2_a |   0.231     0.649     0.644     0.655     0.660
           F |  11.966    46.005    34.043    17.129    19.867
        df_m |   2.000     3.000     4.000     8.000     7.000
        df_r |  71.000    70.000    69.000    60.000    61.000
----------------------------------------------------------------
                                                  legend: b/p
```

Summary

- Oops, real life data. There are several observations missing for rep78
- Because of this, and the lack of statistical relationship between rep78 and mpg, think I will go with model 2 as the final model
- For each increase in pound of car weight, there is a 0.006 (95% CI -0.008, -0.005) decrease in car mpg, controlling for car headroom and car foreign/domestic status.
- $Mpg = 31.99 + -0.219x_{headroom} + -1.655x_{foreign} + -0.006x_{weight}$

Confounding

Confounding

- Many of these car variables seem related.
- Is the apparent association of some of them with mpg really due to another variable?
- Do they distort a true relationship between mpg and something else?

Suspected confounding

- Suspect there is a true relationship between weight and mpg, based on prior knowledge
- Suspect car length may be a confounder
  - Is it related to weight?
  - Is it related to mpg?
- Car length is probably related to weight, and that is likely where its relationship to mpg comes from

Car length and weight

- Use linear regression and visually a scatter plot to assess relationship between length and weight
- Look at correlation
- Seems very much related

Car length and mpg

- Use linear regression and visually a scatter plot to assess relationship between length and mpg
- Correlation
- Again is related

Does car length alter relationship of weight and mpg?

- Run two linear regression models, one controlling for length and one not
- Does the estimate for weight change by more than 10%?
  - Yes, it changes by

Immediate Commands

- ttesti #obs1 #mean1 #sd1 #obs2 #mean2 #sd2 , options

`ttesti` is like `ttest`, except that we specify summary statistics rather than variables as arguments. For instance, we are reading an article that reports the mean number of sunspots per month as 62.6 with a standard deviation of 15.8. There are 24 months of data. We wish to test whether the mean is 75:

```
. ttesti 24 62.6 15.8 75

One-sample t test
------------------------------------------------------------------------------
         |     Obs        Mean    Std. err.   Std. dev.   [95% conf. interval]
---------+--------------------------------------------------------------------
       x |      24        62.6    3.225161        15.8    55.92825    69.27175
------------------------------------------------------------------------------
    mean = mean(x)                                                t =  -3.8448
H0: mean = 75                                    Degrees of freedom =       23

    Ha: mean < 75               Ha: mean != 75                 Ha: mean > 75
 Pr(T < t) = 0.0004         Pr(|T| > |t|) = 0.0008          Pr(T > t) = 0.9996
```

Immediate Commands

Stata's `tabulate` command makes tables and calculates various measures of association. The immediate form, `tabi`, does the same, but we specify the contents of the table following the command:

```
. tabi 5 10 \ 2 14
           |         col
       row |         1          2 |     Total
-----------+----------------------+----------
         1 |         5         10 |        15
         2 |         2         14 |        16
-----------+----------------------+----------
     Total |         7         24 |        31
           Fisher's exact =                 0.220
   1-sided Fisher's exact =                 0.170
```

The `tabi` command is slightly different from most immediate commands because it uses '\' to indicate where one row ends and another begins.

Tip of the day: storing estimates

- Remember graphing residuals last week from a postestimation command?
- Stata keeps results in short-term memory after running a command
- You can “store” these results for long-term use

```
estimates store NameEstimate
```

- We will do this with our regressions
- Comes in handy for model comparison and for recalling results without re-running analysis
  - Sometimes models can take a long time, so this is helpful