ECN 102: Analysis of Economics Data

Final Review Session Worksheet

Remy Beauregard

Question 1: Choosing Our Standard Errors

Suppose we have estimated the multivariate model:

$$y_i = \beta_1 + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + u_i$$

a) What is the standard error of b_j estimating, for $j = 1, \ldots, k$?

b) Which of our quantities are changed by our choice of standard error? Which are not?

c) Why is our choice of standard error important?

d) Which of our four population assumptions do our default standard errors rely on? Which of these fails most often in practice? What should we do in response?

e) Suppose we plot our residuals against x_2 and find that they get more dispersed as x_2 grows. Which standard errors should we use here? Why?

f) Now suppose we are estimating the effect of a school-level tutoring program on individual student test scores, with students observed within one of many schools. Why might our default errors at the student level fail to be independent? Which standard errors should we use instead and what should we select for G?

g) How does this process change our third and fourth population assumptions? Does this fix any other potential problems for us?

h) Suppose instead we are worried that our outcome is not truly linear in our regressors. Which standard errors should we use and how (generally) does this procedure actually produce a standard error?

i) **[Extension]** Can the process in part (h) accommodate error dependence within clusters? What would we do if we suspected both non-linearity *and* within-cluster dependence, and what would that require of our data?

Question 2: Multivariate Inference with an Interacted Dummy

For this question we use the 1988 National Longitudinal Survey extract of working women (`sysuse nlsw88`). The variables we need are described here.

```
Variable      Storage   Display    Value
    name         type    format    label      Variable label
-------------------------------------------------------------------------------
wage            float   %9.0g                 Hourly wage
ttl_exp         float   %9.0g                 Total work experience (years)
grade           byte    %8.0g                 Current grade completed
union           byte    %8.0g      unionlbl   Union worker
```

Suppose we generate an interaction term with `gen unionXexp = union * ttl_exp` and then run `reg wage ttl_exp grade union unionXexp`, obtaining:

```
      Source |       SS           df       MS      Number of obs   =     1,876
-------------+----------------------------------   F(4, 1871)      =    185.77
       Model |  9267.34596         4  2316.83649   Prob > F        =    0.0000
    Residual |  23334.1286     1,871  12.4714744   R-squared       =    0.2843
-------------+----------------------------------   Adj R-squared   =    0.2827
       Total |  32601.4745     1,875  17.3874531   Root MSE        =    3.5315

------------------------------------------------------------------------------
        wage | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
     ttl_exp |     0.2715     0.0206    13.18   0.000       0.2311      0.3119
       grade |     0.6084     0.0326    18.65   0.000       0.5444      0.6724
       union |     1.4250     0.5766     2.47   0.014       0.2942      2.5558
   unionXexp |    -0.0333     0.0415    -0.80   0.423      -0.1147      0.0481
       _cons |    -4.1632     0.4709    -8.84   0.000      -5.0868     -3.2395
------------------------------------------------------------------------------
```

a) Write down the population model we are estimating.

b) Which regressors are significant at 5%? Which are not?

c) Do our regressors **jointly** explain variation in wages at 5% significance?

d) What type of regressor is *union*? Why? What type is *unionXexp*?

e) How do we interpret the coefficients on the regressors above?

f) How does our regression **avoid** falling into the dummy variable trap? What else could we have done to avoid this?

g) Find the marginal effect of *union* on wages in our regression.

h) Give the MEM and MER for *union* with $\overline{ttl\_exp} = 12.82$ and $ttl\_exp^* = 25$.

i) What is the RMSE of our regression? What is our R-squared? What is our adjusted R-squared? What do these each mean in words?

j) Suppose we want to test if our **variable** union contributes anything to our regression. What kind of test would we run? How many linear restrictions would we be testing? What kind of statistic would we expect Stata to compute?

k) Write down the null and alternative hypotheses of this test. Write down the *restricted* model and the *unrestricted* model. If we ran this test in Stata and found `Prob > F = 0.000`, what could we conclude?

l) How does this test above differ from the default test of overall significance?

m) Could we have used a t-test to answer this question? Why or why not?

n) We estimated the above model with multivariate OLS. What does this mean about our estimators/estimation? (several possible answers)

Question 3: Log Transformations and State Dummies

Suppose we are investigating the dangers of extreme heat on human health. We have monthly data on:

1) The number of days above a given threshold for extreme heat
2) Number of hospitalizations for heat-related illness
3) The state of observation

We want to fit a regression to predict heat-related illness with number of hot days.

a) Write down the linear (population) model we would estimate.

b) What type of data is each of our series?

c) We now think that average heat-related hospitalization may differ by state, due to different preparedness levels, cooling facilities, and other infrastructure. How would we account for the fact that states may differ from each other in means? How would we avoid a particular *trap* while doing this?

d) What is the *total effect* of the number of extreme heat days in our regression if heat days are independent of our state dummies? If they are not?

e) Suppose instead of a level change in the raw number of hospitalizations, we care about the *proportional* (percent) change in hospitalizations associated with an additional extreme heat day. How could we change our regression to accommodate this?

f) If we implemented the transformation above, how would this change our interpretation of the coefficient on the number of extreme heat days? What type of regression is this now?

g) What would be the *marginal effect* of the number of extreme heat days if we implemented the transformation above in part (e)?

h) Suppose we wanted to decide if we should add regressors for the average monthly humidity level and average monthly UV index into our regression. How could we decide if we should add these regressors *individually*? *Jointly*?

i) Suppose we did add the two regressors above and found our new coefficient on extreme heat days. If we find that the partial effect from this regression is different from the b_2 coefficient from a (bivariate) regression of just log hospitalizations on extreme heat days, what does this suggest about extreme heat days and our other regressors?