ECN 102: Analysis of Economics Data

Final Review Session Worksheet – Answer Key

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 $j = 1, \ldots, k$?

A standard error is our best estimate of the population standard deviation of a statistic. Here s_{b_j} estimates σ_{b_j}, the spread of our distribution of b_j if we drew many samples of size n from the population and estimated β_j from each.

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

The standard errors s_{b_j} themselves, and every quantity computed from them, will change with our choice of standard error:

$$t = \frac{b_j - \beta_{j,0}}{s_{b_j}}, \qquad CI = b_j \pm t^*_{\alpha/2, n-k} \cdot s_{b_j}$$

So our t-statistics, our p-values (`P>|t|`), our 95% confidence intervals, and our F-statistics for joint tests all move.

Our coefficient estimates $b_1, b_2, \ldots, b_k$, our fitted values $\hat{y}_i$, our residuals, and therefore R^2, $\bar{R}^2$, and RMSE will all be unchanged.

c) Why is our choice of standard error important?

Because the wrong standard error gives us improper inference even for an unbiased coefficient. If s_{b_j} is too small, our t-statistics are too large and our confidence intervals too narrow: we reject far more often than α promises and make extra Type I errors. If s_{b_j} is too large, we lose power and commit Type II errors instead, failing to reject the null hypothesis when we should.

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?

The default standard error formula relies on assumptions 3 and 4:

$$3.\ V[u_i|x_{2i},\dots,x_{ki}] = \sigma_u^2 \text{ for all } i \quad \text{(homoskedasticity)}$$

$$4.\ Corr[u_i,u_j|x_{2i},\dots,x_{ki},x_{2j},\dots,x_{kj}] = 0 \text{ for all } i \neq j \quad \text{(independence)}$$

Homoskedasticity is the one that fails most often. In response, it is common practice in economics to simply *always use* robust standard errors. The choice is asymmetric: robust standard errors remain valid *even if* homoskedasticity does hold, so there is no real cost to over-using them, while conventional standard errors are improper when error variance is not constant.

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?

We should use robust standard errors. The spread of our errors depends on the value of x_2, which is exactly a violation of homoskedasticity. Robust standard errors relax assumption 3 to allow each observation its own error variance:

$$V[u_i|x_{2i},\dots,x_{ki}] = \sigma_{ui}^2, \text{ for all } i$$

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?

Students in the same school share teachers, facilities, a principal, a neighborhood, and any school-wide shock that year. Whatever our regressors do not capture about a school hits all of its students at once, so the errors of two students in the same school are very likely correlated and assumption 4 is violated. Errors *between* schools are still plausibly independent, allowing us to use cluster (robust) standard errors. We therefore want clustered standard errors with $G =$ school.

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

3_{clu}: **Cluster heteroskedasticity (allowed):**

$$V[u_i|x_{2i},\dots,x_{ki}] = \sigma_{ui}^2 \text{ for all } i$$

4_{clu}: **Cluster independence:**

$$Corr[u_i, u_j | x_{2i}, \ldots, x_{ki}, x_{2j}, \ldots, x_{kj}] \begin{cases} \neq 0 & i \text{ and } j \text{ in the same cluster} \\ = 0 & i \text{ and } j \text{ in different clusters} \end{cases}$$

Yes, clustering is the stronger correction of the two: it relaxes assumption 4 *and* assumption 3.

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?

A failure of assumption 1 (linearity) means our analytic standard error formula for $b_2, \ldots, b_k$ is no longer trustworthy, since that formula is derived from the linear model. The fix is to generate the sampling variation empirically with bootstrapping:

```
bootstrap, reps(S): reg y x2 x3
```

The procedure treats our sample as if it were the population:

1. **Draw S new samples of size n from our data, resampling rows *with replacement*.**
2. **Estimate b_j^* on each of the S bootstrap samples, and compute the mean $\bar{b_j^*}$ across them.**
3. **The standard deviation of the b_j^* across bootstrap samples is our bootstrapped standard error for b_j.**

The variation across resamples stands in for the sampling variation we cannot observe. This simple version is valid under heteroskedasticity.

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?

No. Resampling *rows* breaks clusters apart, so the simple bootstrap cannot handle clustered errors. If we suspect both, we resample whole *clusters* of observations rather than individual rows:

```
bootstrap, reps(S) cluster(G): reg y x2 x3
```

Since we are now resampling at the *cluster* level rather than the *individual* level, we need enough clusters to generate variation across resamples, generally at least 30 to 50, to produce a reliable standard error.

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.

$$wage_i = \beta_1 + \beta_2 \,\text{ttl\_exp}_i + \beta_3 \, grade_i + \alpha_1 \, union_i + \alpha_2 \, (union_i \times \text{ttl\_exp}_i) + u_i$$

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

ttl_exp **($t = 13.18$), *grade* ($t = 18.65$), and *union* ($t = 2.47$) are all statistically significantly different from zero at $\alpha = 5\%$; each has `P>|t|` < 0.05. The interaction *unionXexp* is not: `P>|t|` $= 0.423 > 0.05$, so we fail to reject $H_0 : \alpha_2 = 0$.**

Equivalently, we could compare each $|t|$ against the critical value $t^*_{0.025,1871} \approx 1.96$.

`invttail(1871,0.025) = 1.9612`

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

Yes. The default test of overall significance reports $F(4, 1871) = 185.77$ with `Prob > F` $= 0.0000 < 0.05$, so we reject $H_0 : \beta_2 = \beta_3 = \alpha_1 = \alpha_2 = 0$. At least one of our regressors statistically explains variation in *wage*.

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

***union* is a dummy (binary indicator) variable: it takes only the values 0 and 1, encoding the qualitative characteristic "is a union worker". *unionXexp* is an interaction term, the product of our dummy and a continuous regressor, which lets the slope on experience differ between union and non-union workers.**

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

Interpret the interacted model as two lines, one for each value of the dummy:

- $b_1 = -4.1632$**: the predicted hourly wage of a non-union worker with zero years of schooling and zero experience. This is outside the range of our data, so it is not economically meaningful on its own.**
- $b_2 = 0.2715$**: among non-union workers ($union = 0$), one additional year of total work experience is associated with a \$0.27 higher hourly wage, holding schooling constant.**
- $b_3 = 0.6084$**: one additional year of schooling completed is associated with a \$0.61 higher hourly wage, holding experience and union status constant. *grade* is not interacted, so this applies to union and non-union workers alike.**
- $a_1 = 1.4250$**: the intercept shift for union workers. At zero years of experience, union workers are predicted to earn \$1.43 more per hour than non-union workers with the same schooling.**
- $a_2 = -0.0333$**: the slope shift for union workers. Each additional year of experience is associated with \$0.03 *less* of a wage gain for union workers than for non-union workers, so the return to experience is $0.2715 - 0.0333 = 0.2382$ for union workers.**

Writing out the two fitted lines makes this concrete:

$$\text{Non-union } (union = 0): \quad \widehat{wage} = -4.1632 + 0.2715 \cdot \text{ttl\_exp} + 0.6084 \cdot grade$$

$$\text{Union } (union = 1): \quad \widehat{wage} = (-4.1632 + 1.4250) + (0.2715 - 0.0333) \cdot \text{ttl\_exp} + 0.6084 \cdot grade$$

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

Union status is a mutually exclusive and exhaustive two-category variable (union, non-union). We include only *one* of the two possible dummies along with the constant term, so no regressor is a perfect linear combination of the others. Non-union is our leave-out group, absorbed into the constant, and every coefficient on *union* is interpreted relative to that group.

Alternatively, we could include dummies for *both* categories and drop the constant term (`reg wage ttl_exp grade union nonunion, nocons`) or include the dummy for non-union instead of union. What we cannot do is include both dummies *and* the constant, since $nonunion = 1 - union$ would be perfectly collinear with the constant.

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

$$\frac{\partial \widehat{wage}}{\partial\, union} = a_1 + a_2 \cdot \text{ttl\_exp} = 1.4250 - 0.0333 \cdot \text{ttl\_exp}$$

Because *union* appears in two places (on its own and inside the interaction), its marginal effect depends on the value of *ttl_exp* at which we evaluate it.

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

MEM: plug in the sample mean of experience.

$$MEM = 1.4250 - 0.0333 \times 12.82 = 1.4250 - 0.4269 = 0.9981 \approx \$1.00$$

MER: plug in the representative value $ttl\_exp^* = 25$.

$$MER = 1.4250 - 0.0333 \times 25 = 1.4250 - 0.8325 = 0.5925 \approx \$0.59$$

At mean experience, union membership is associated with about $1.00 more per hour; for a worker with 25 years of experience, the union premium shrinks to about $0.59 per hour.

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?

- $RMSE = 3.5315$. **This is $s_e = \sqrt{ResSS/(n-k)}$, the estimated standard deviation of our residuals, measured in the units of the outcome (dollars per hour). It tells us the typical size of a prediction error but is not standardized, so it can only be compared across models fit to the same outcome.**

- $R^2 = 0.2843$. **Our regressors explain 28.43% of the sample variation in *wage*:** $R^2 = 1 - ResSS/TSS$. **It never falls when we add a regressor, so it cannot tell us whether a new regressor is worth including.**
- $\bar{R}^2 = 0.2827$. **The same measure of fit, penalized for the number of parameters k we estimate:** $\bar{R}^2 = 1 - \frac{ResSS/(n-k)}{TSS/(n-1)}$. **Because the penalty grows with k, adjusted R-squared can fall when we add a regressor that contributes little explanatory power, which makes it the appropriate measure for comparing models with different numbers of regressors (but the same outcome variable).**

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?

This is a joint F-test of the *variable* union status. Union enters our model through two *regressors*, *union* and *unionXexp*, so the variable contributes nothing only if both of their coefficients are zero at once. Setting each of those two coefficients to zero is $q = 2$ linear restrictions.

With more than one restriction we cannot use a t-test, so Stata computes an F-statistic with $q = 2$ numerator and $n - k = 1876 - 5 = 1871$ denominator degrees of freedom: $F_{2,1871}$.

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?

$$H_0 : \alpha_1 = 0 \ \cap \ \alpha_2 = 0$$

$$H_A : \alpha_1 \neq 0 \ \cup \ \alpha_2 \neq 0$$

Note the $\cap$ and the $\cup$: rejecting tells us that *at least one* of the two coefficients is non-zero, not that both are, and does not tell us which one.

Restricted vs. unrestricted. The unrestricted model is the one we actually estimated; the restricted model is what is left after imposing H_0, the same model with both union terms deleted:

Unrestricted: $wage_i = \beta_1 + \beta_2 \, \text{ttl\_exp}_i + \beta_3 \, grade_i + \alpha_1 \, union_i + \alpha_2 \, (union_i \times \text{ttl\_exp}_i) + u_i$

Restricted: $wage_i = \beta_1 + \beta_2 \, \text{ttl\_exp}_i + \beta_3 \, grade_i + u_i$

The F-statistic compares the two by asking how much worse the restricted model explains variation in wages than the unrestricted model. A large F means that dropping the union terms cost us a lot of explanatory power.

Conclusion. `Prob > F` **is the p-value of our F-test: the probability of seeing an F-statistic this large if H_0 were true. Since $0.000 < \alpha = 0.05$, we reject H_0 at 5% significance. The unrestricted model explains statistically significantly more variation in wages than the restricted model, so the *variable* union does contribute explanatory power to our regression.**

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

The default test of overall significance imposes restrictions on *all* $k-1=4$ slope coefficients ($H_0 : \beta_2 = \beta_3 = \alpha_1 = \alpha_2 = 0$, so $q = 4$), so our restricted model is only a constant:

$$\text{Restricted (overall significance): } wage_i = \beta_1 + u_i$$

That is the $F(4, 1871) = 185.77$ Stata reports in the header of our regression output. Our test in part (k) imposes only the $q = 2$ restrictions involving union status, so its restricted model still contains *ttl_exp* and *grade*, and it is an $F_{2,1871}$. Both are F-tests of the same form (the unrestricted model is identical) but they answer different questions: "does anything in our model explain wages?" versus "does union status specifically explain wages?"

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

No. A t-test can impose only a single linear restriction ($q = 1$), and our null involves two coefficients simultaneously ($q = 2$). A t-test on *union* alone would test only the partial effect of the standalone *union* term, the union premium evaluated at $ttl\_exp = 0$, and would ignore the channel running through the interaction entirely.

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

Our four population assumptions are (1) linearity, (2) unbiasedness $E[u_i|x_{2i}, \dots, x_{ki}] = 0$, (3) homoskedasticity $V[u_i|x_{2i}, \dots, x_{ki}] = \sigma_u^2$, and (4) independence of errors. Under these assumptions, we can say any of the following:

- **Our estimator is unbiased: $E[b_j] = \beta_j$ (from assumptions 1 and 2).**
- **Our estimator is consistent: $V[b_j] \to 0$ as $n \to \infty$, so b_j converges to β_j.**
- **Our estimator is BLUE: of all linear unbiased estimators, OLS has the minimum variance (Gauss-Markov), which is what keeps our Type II error probability as low as possible.**
- **Our estimator is asymptotically normal: $b_j \sim N(\beta_j, \sigma^2_{b_j})$ by the CLT.**
- **OLS minimizes the sum of squared residuals, and our reported standard errors $s_{b_j} = s_e / \sqrt{\sum_i \tilde{x}^2_{ji}}$ are valid estimates of the true sampling variation.**

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.

$$hosp_{it} = \beta_1 + \beta_2\, hotdays_{it} + u_{it}$$

where i indexes states and t indexes months.

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

- ***hotdays* is numerical and discrete (a count of days in the month, bounded between 0 and 31).**
- ***hosp* is numerical and discrete (a count of hospitalizations).**
- ***state* is categorical and unordered (qualitative), so it cannot enter a regression as a single numeric variable but must be converted into a set of dummy variables.**

In terms of data structure, this is panel (longitudinal) data: repeated observations on the same states over multiple time periods, denoted x_{it}.

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?

Include a set of state dummy variables, which shift the intercept for each state and thereby allow states to differ in mean hospitalizations:

$$hosp_{it} = \beta_1 + \beta_2\, hotdays_{it} + \gamma_2 d_{2i} + \gamma_3 d_{3i} + ... + \gamma_S d_{Si} + u_{it}$$

With S states we include at most $S-1$ dummies along with the constant. Including all S dummies plus a constant would be the dummy variable trap: the dummies sum to exactly 1 for every observation, making one of them a perfect linear combination

of the others and the constant, and our coefficients could not be estimated. The omitted state is the leave-out group, and each γ_s is interpreted as the difference in mean hospitalizations between state s and that leave-out state.

The alternative fix is to include all S dummies and drop the constant term.

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?

If *hotdays* is independent of the state dummies, there are no indirect channels, so the total effect equals the partial effect:

$$\frac{d\,\widehat{hosp}}{d\,hotdays} = \frac{\partial\,\widehat{hosp}}{\partial\,hotdays} = b_2$$

If they are not independent, the total effect adds the indirect effects running through the correlated state dummies:

$$\frac{d\,\widehat{hosp}}{d\,hotdays} = \underbrace{b_2}_{\text{direct}} + \underbrace{\sum_s g_s \frac{\partial d_s}{\partial\,hotdays}}_{\text{indirect}}$$

Equivalently, the total effect is the b_2 we would get from a *bivariate* regression of hospitalizations on heat days alone, while the multivariate b_2 is the partial effect holding state fixed. Here we would expect them to differ: hot states are hot every year, so *hotdays* and the state dummies are strongly correlated.

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?

We use a log transformation when we care about a proportional change rather than a level change. Take the natural log of the outcome variable and estimate a log-linear model:

$$\ln(hosp_{it}) = \beta_1 + \beta_2\, hotdays_{it} + \gamma_2 d_{2i} + ... + \gamma_S d_{Si} + u_{it}$$

In Stata, `gen ln_hosp = ln(hosp)` and then regress *ln_hosp* on our regressors.

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?

This is a log-linear regression; we interpret b_2 as a semi-elasticity: one additional extreme heat day is associated with approximately a $100 \times b_2$ percent change in hospitalizations, holding state fixed. So $b_2 = 0.04$ would mean one more hot day is associated with roughly a 4% increase in heat-related hospitalizations.

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

For a log-linear model, the coefficient gives us $\Delta \ln \hat{y}/\Delta x$; to recover the marginal effect in levels we multiply by the fitted value:

$$\frac{\Delta \widehat{hosp}}{\Delta\, hotdays} = b_2 \cdot \widehat{hosp}$$

Since this depends on $\widehat{hosp}$, we would report it at a specific point, the MEM (evaluated at $\overline{hosp}$) or an MER (evaluated at some representative $hosp^*$).

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*?

Individually: add each regressor and run a t-test of association on its coefficient ($H_0 : \beta_j = 0$, $q = 1$). We could also check whether adjusted R-squared rises or RMSE falls, though these are not hypothesis tests and neither is sufficient evidence of statistical significance on its own.

Jointly: run an F-test of the two coefficients together, $H_0 : \beta_{humidity} = 0 \cap \beta_{UV} = 0$ with $q = 2$ restrictions, comparing the unrestricted model (with both) against the restricted model (with neither). This is the proper test when the two regressors are correlated with each other (as UV and humidity might be), since each individual t-test can fail to reject while the pair jointly explains a significant amount of variation.

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?

It tells us that extreme heat days are correlated with humidity and/or the UV index. The bivariate coefficient measures the *total effect*: the direct effect of heat days plus every indirect effect running through omitted regressors that covary with heat days. The multivariate coefficient measures the *partial effect*: the effect of heat days holding humidity and UV fixed. These two are equal only when heat days are independent of the other regressors, so a difference between them is exactly the evidence of correlation.