Regression #8: Loose Ends

Econ 671

Purdue University

- In this lecture we investigate a variety of topics that you are probably familiar with, but need to touch on nonetheless. These include:

  1. Multicollinearity
  2. Coefficient interpretation with log transformations.
  3. Dummy / Indicator Variables
  4. Nonlinearities.

Multicollinearity

So, what is *multicollinearity*?

-

So, why is this important?

1.
2.
3.

Multicollinearity

In a very real sense, the importance of multicollinearity, and its perception as a "problem" in econometrics, is overblown.

For example, OLS estimators are still unbiased, consistent and efficient in the presence of high (but not perfect) collinearity.

Goldberger (1991) likens the problem of multicollinearity to *micronumerosity* - the "problem" of having a small sample size.

A series of interesting (and entertaining) quotes on this issue are taken from his book:

Multicollinearity

> *"The extreme case, 'exact micronumerosity' arises when $n = 0$, in which case the sample estimate of μ is not unique ... The extreme case is easy enough to recognize. "Near micronumerosity" is more subtle, and yet very serious. It arises when the rank condition $n > 0$ is barely satisfied ... "*

He continues by noting the similarity of consequences with multicollinearity ...

> *"The consequences of micronumerosity are serious. Precision of estimation is reduced ... Investigators will sometimes be led to acccept the hypothesis $\mu = 0$ ... even though the true situation may be not that $\mu = 0$ but that the sample data have not enablesd us to pick μ up. The estimate of μ will [also] be very sensitive to the sample data ...*

Multicollinearity

Finally, he suggests some tests for micronumerosity ...

> *Tests for the presence of micronumerosity require the judicious use of various fingers. Some researchers prefer a single finger, others use their toes, still others let their thumbs rule. A generally reliable guide may be obtained by counting the number of observations. Most of the time in econometric analysis, when n is close to zero, it is also far from infinity.*

While these are entertaining, they illustrate that:

1. The problems associated with small sample sizes are like those associated with multicollinearity.
2. Multicollinearity does not violate any of our fundamental assumptions; it is simply a feature of the regression model itself. Large standard errors are not "wrong" or "misleading" as the coefficient estimates *should* vary a lot from sample to sample.

While multicollinearity may be bad, in the sense that individual t-statistics are small, leading the applied researcher to update his/her beliefs about the publishability of the work and subsequently wanting to throw himself/herself out the window, it can also aid in inference, (e.g., prediction), as the following example suggests:

Suppose:

-

and, to fix ideas, set $\sigma^2 = 1$. In addition, suppose that the explanatory variables have been scaled so that:

-

In this case,

-

since the off-diagonal, $\sum_i x_{1i} x_{2i}$ is the sample correlation, denoted as ρ.

Multicollinearity: Example

It follows that

-

Thus, for the purposes of getting a "small" variance for $\hat{\beta}_1$ and $\hat{\beta}_2$, we would want to set $\rho = 0$.
However, consider the parameter

-

With very little work, we can show:

-

For this parameter, it is clear that $\rho > 0$ leads to increased precision!

Multicollinearity: Example

Note that this logic translates to the exercise of prediction. To this end, suppose we wish to predict y when x_1 and x_2 equal the same value, say c. (Note this is not completely unreasonable given our initial scaling of the data). Then,

-

at which point the preference for a small variance associated with

-

becomes clear. This argument also extends to general problems of prediction.

Consider two different regressions:

$$
\begin{aligned}
y_i &= \beta_1 + \beta_2 Educ_i + u_i \\
y_i &= \theta_1 + \theta_2 Educ_i + \theta_3 Educ_i^2 + \theta_4 Educ_i^3 + \theta_5 Educ_i^4 + \epsilon_i
\end{aligned}
$$

```
. regress lwage educat

      Source |       SS       df       MS              Number of obs =    1260
-------------+------------------------------           F(  1,  1258) =   95.89
       Model |  31.5149966     1  31.5149966           Prob > F      =  0.0000
    Residual |  413.464976  1258  .328668502           R-squared     =  0.0708
-------------+------------------------------           Adj R-squared =  0.0701
       Total |  444.979972  1259  .353439215           Root MSE      =   .5733

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      educat |   .0602839   .0061563     9.79   0.000     .0482061    .0723616
       _cons |   .9014239   .0790132    11.41   0.000     .7464117    1.056436
------------------------------------------------------------------------------

. regress lwage educat educat2 educat3 educat4

      Source |       SS       df       MS              Number of obs =    1260
-------------+------------------------------           F(  4,  1255) =   26.63
       Model |  34.8081566     4  8.70203915           Prob > F      =  0.0000
    Residual |  410.171816  1255  .326830132           R-squared     =  0.0782
-------------+------------------------------           Adj R-squared =  0.0753
       Total |  444.979972  1259  .353439215           Root MSE      =  .57169

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      educat |  -.1293852   .9497547    -0.14   0.892    -1.992667    1.733897
     educat2 |   .0233402   .1400233     0.17   0.868    -.2513654    .2980458
     educat3 |  -.0015552   .0087018    -0.18   0.858    -.0186269    .0155166
     educat4 |    .000042   .0001941     0.22   0.829    -.0003388    .0004229
       _cons |   1.602221   2.250264     0.71   0.477    -2.812474    6.016916
------------------------------------------------------------------------------
```

- In the simple regression model, education looks “clearly significant.”

- In the second model, however, education does not appear related to wages, as none of the coefficients are statistically significant.

- This, however, is an artifact of *multicollinearity*. The relevant question to ask is if all of the education variables are *jointly* equal to zero.

In fact, we calculate a χ_4^2 statistic equal to 106.5 for the joint null hypothesis that $\theta_2 = \theta_3 = \theta_4 = \theta_5 = 0$.

Multicollinearity

We close this discussion with a general derivation that cleanly reveals the "problem" of multicollinearity. Consider a regression equation that has been transformed into deviation from means:

-

where x_1 is a scalar and z_i a vector. We then have:

-

Using the *partitioned inverse theorem* to select off the (1,1) element of this matrix, we obtain:

-

Multicollinearity

Continuing,

-

where R_1^2 is the R-squared value from a regression of x_1 on all the Z's. (To see this last point, recall our earlier derivation of R^2 in the lecture notes).

Thus, high values of R_1^2 lead to high variances (the multicolinearity problem). Conversely, lots of variation in x_1 mitigates the variance.

Interpretations with Common Transformations

Models with logarithmic transformations on the dependent and independent variables are ubiquitous in applied work and thus it is useful to pause and explain how to interpret coefficients in such cases.

As a benchmark, consider a model without any transformations:

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

It is clear in such a situation that β_j represents a *marginal effect* - the expected change in y corresponding to a unit change in x_j (holding all else constant):

-

thus producing its interpretation.

Interpretations with Common Transformations

Now, consider a model in which both the dependent and all the explanatory variables have logarithmic transformations:

-

To see the interpretation here, it is useful to take the *differential* of both sides of this equation, noting that $d[f(x)] = f'(x)dx$. Thus,

-

or

-

yielding

-

Interpretations with Common Transformations

The left hand side of the last equation is a partial *elasticity* - the percentage change in y corresponding to a percentage change in x_j (holding all else constant). Thus, in log-log models, the coefficients represent (partial) elasticities.

$$\text{Example}: \ \widehat{logCoffeeDemanded} = .77 - .253 logPrice.$$

This would indicate that the demand for coffee is *price-inelastic*.

Perhaps equally common is the case where the dependent variable has a log transformation, but the independent variables do not:

-

Performing a similar operation, we obtain:

-

which rearranges to:

-

or $100\beta_j$ represents the *percentage change in y corresponding to a unit change in x*.

$$\text{Example}: \widehat{LogWage} = 2.2 + .12 Education.$$

That is, an added year of schooling increases your wages by 12 percent, on average.

Interpretations with Common Transformations

Also note that other parameters of interest can be obtained via simple manipulations of this formula.

For example, in the log-levels model, we can re-arrange things to obtain:

-

Evaluated at $y = 10$, for example, this would imply that an added year of schooling increases your (hourly) wage by about \$1.20 on average.

Finally, similar (and obvious) manipulations can be performed to provide the correct interpretation when the explanatory variables are measured in logs while the dependent variable is measured in levels.

Dummy Variables and Interactions

- Dummy variables (or indicator variables) represent a useful way to represent qualitative information in a *quantitive* way.
- For example, one can control for variation across race, gender or region of residence through the creation of dummy variables.
- Often data sets will code such information in a way that is not directly useful to the econometrician. For example, gender may be listed as "F" or "M" while region of residence may be coded as, say, 1-4, denoting the East, West, North and South, respectively.

Dummy Variables and Interactions

When dummy variables are used, one must take care to interpret the parameters correctly. To see this, consider two models:

$$\begin{aligned} y_i &= \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 D_{3i} + u_i \\ y_i &= \alpha_1 + \alpha_2 \tilde{D}_{2i} + \alpha_3 \tilde{D}_{3i} + v_i \end{aligned}$$

where

Model 1	**Model 2**
$D_1 = I(Ed < 12)$	
$D_2 = I(Ed = 12)$	$\tilde{D}_2 = I(Ed \geq 12)$
$D_3 = I(Ed > 12)$	$\tilde{D}_3 = I(Ed > 12)$

Dummy Variables and Interactions

The conditional expectations reveal the interpretation of the coefficients in each model:

$$
\begin{aligned}
E(y|Dropout, Model1) &= \beta_1 \\
E(y|Dropout, Model2) &= \alpha_1
\end{aligned}
$$

$$
\begin{aligned}
E(y|HSGrad, Model1) &= \beta_2 \\
E(y|HSGrad, Model2) &= \alpha_1 + \alpha_2
\end{aligned}
$$

Dummy Variables and Interactions

$$\begin{aligned} E(y|MorethanHS, Model1) &= \beta_3 \\ E(y|MorethanHS, Model2) &= \alpha_1 + \alpha_2 + \alpha_3 \end{aligned}$$

Thus, in Model 2, the α's are interpreted as the gains (or losses) from moving to the higher education group while the β's are the average wages for the given group.

Note that the interpretation of the coefficients changes across models, even though $\tilde{D}_3$ and D_3, for example are the *same variable*.

Dummies and Interactions: Example 1

The STATA output clearly shows that the α's are estimated as differences between the β's. Wages are significantly larger when moving to the higher education group.

```
. regress wage D1 D2 D3, noconst
------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          D1 |     774.25   41.29344    18.75   0.000     693.2111    855.2889
          D2 |   862.6718   19.54007    44.15   0.000     824.3241    901.0194
          D3 |   1076.024   18.18003    59.19   0.000     1040.346    1111.703
------------------------------------------------------------------------------

. regress wage tildeD2 tildeD3
------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     tildeD2 |   88.42176   45.68329     1.94   0.053    -1.232276    178.0758
     tildeD3 |   213.3525   26.68947     7.99   0.000     160.9741    265.7309
       _cons |     774.25   41.29344    18.75   0.000     693.2111    855.2889
------------------------------------------------------------------------------
```

Interaction

What is an *interaction*?

-

This is sometimes done to add flexibility to a regression model. However, most of the time, the interaction is added to enable the researcher to test some hypothesis of interest. Consider, for example the regression model:

$$LogWage_i = \beta_1 + \beta_2 Fem_i + \beta_3 Educ_i + \beta_4 Fem_i * Educ_i + \beta_5 Exper_i + u_i.$$

What potentially interesting hypotheses would this enable us to test?

What would you conclude based on these results?

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
female	-.2958839	.1787929	-1.65	0.099	-.6471275	.0553597
educ	.0927793	.0089777	10.33	0.000	.0751423	.1104163
femaleEd	-.0038152	.013975	-0.27	0.785	-.0312694	.0236391
exper	.0094302	.0014518	6.50	0.000	.0065781	.0122823
_cons	.4614994	.1267468	3.64	0.000	.2125017	.7104971

Nonlinearities

Nonlinearities are typically handled in a regression framework by including powers of the explanatory variables and including these as separate regressors. In a sense, this might be thought of as a special interaction.

Sometimes these are included to make the regression model more flexible, but if this is the case, researchers often tend to prefer *nonparametric* methods.

However, economic theory (and common sense) often suggests the inclusion of such variables to allow, for example, quadratic profiles of certain covariates. Consider, for example, the model below:

$$LogWage_i = \beta_1 + \beta_2 Fem_i + \beta_3 Educ_i + \beta_4 Exper_i + \beta_5 Exper_i^2 + u_i,$$

with output presented on the following page:

When are the returns to experience at a maximum?

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
female	-.3371868	.0363214	-9.28	0.000	-.4085411	-.2658324
educ	.0841361	.0069568	12.09	0.000	.0704692	.0978029
exper	.03891	.0048235	8.07	0.000	.029434	.0483859
exper2	-.000686	.0001074	-6.39	0.000	-.000897	-.0004751
_cons	.390483	.1022096	3.82	0.000	.1896894	.5912767

What if you wanted to test if return to experience profiles are different for men and women? To this end, you might want to estimate a model like:

$$\begin{aligned} LogWage_i &= \beta_1 + \beta_2 Educ_i + \beta_3 Exper_i + \beta_4 Exper_i^2 \\ &+ \beta_5 F_i + \beta_6 F_i * Exper_i + \beta_7 F_i * Exper_i^2 + u_i, \end{aligned}$$

where F represents the Female Dummy.

What would be some hypothesis tests of interest?

When are wages highest for men? For Women? Do we see the same experience profiles for both genders?

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.0857402	.0068921	12.44	0.000	.0722004	.0992801
exper	.0543082	.0065378	8.31	0.000	.0414643	.067152
exper2	-.000929	.0001447	-6.42	0.000	-.0012134	-.0006447
female	-.035051	.0800504	-0.44	0.662	-.1923137	.1222117
fexper	-.0320967	.0094877	-3.38	0.001	-.0507357	-.0134578
fexper2	.0005158	.0002093	2.46	0.014	.0001046	.000927
_cons	.2186471	.1078223	2.03	0.043	.0068253	.4304688

Maximum Experience, Males: 29.2, Maximum Experience, Females: 26.9