Economics 6400: Econometrics

Lecture 6: Interaction terms and revision

CSU, East Bay

October 31, 2017

In the past few weeks...

- We introduced Multiple Regression Analysis:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

- Key assumption that is required to obtain unbiased estimates of $\beta_1, \beta_2 \ldots ...., \beta_k$ is:

$$\mathrm{E}(u|x_1, x_2, \ldots, x_k) = 0$$

  which implies that *all* of the independent variables are not correlated with the error term
- Estimate of β_j is $\hat{\beta}_j$
  - $\Delta\hat{y} = \hat{\beta}_j \Delta x$

In the past few weeks...

- We also derived the **estimated** variation of $\hat{\beta}_j$:

$$\widehat{Var}(\hat{\beta}_j) = \frac{\hat{\sigma}^2}{SST_j(1 - R_j^2)}$$

where

- $\hat{\sigma}^2 = \frac{SSR}{n-k-1} = \frac{\sum_{i=1}^n \hat{u}_i^2}{n-k-1}$
- $SST_j = \sum_{i=1}^n (x_{ij} - \overline{x}_j)^2$ is the sample variation in right-hand side variable x_j
- R_j^2 is the R^2 from a regression of x_j on the other right-hand side variables (including a constant)
- The **standard error** is $se(\hat{\beta}_j) = \sqrt{\widehat{Var}(\hat{\beta}_j)} = \sqrt{\frac{\hat{\sigma}^2}{SST_j(1-R_j^2)}}$

In the past few weeks...

- We then used the standard error for hypothesis testing
- Suppose we wish to test whether β_j is equal to some value (usually zero):

$$\mathrm{H}_0 : \beta_j = a_j$$
$$\mathrm{H}_1 : \beta_j \neq a_j$$

- Test statistic is:

$$t_{\hat{\beta}_j} \equiv \frac{\hat{\beta}_j - a_j}{se(\hat{\beta}_j)} \sim t_{n-k-1} = t_{df}$$

- The smallest significance level at which the null hypothesis is still rejected is called the *p*-**value** of the hypothesis test

Interaction terms

- Sometimes it makes sense for the effect of a right-hand side variable to depend on *another* right-hand side variable
- For example, consider the house pricing equation:

  $$price = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + \beta_3 sqrft \cdot bdrms + \beta_4 bthrms + u$$

  such that the partial effect of another bedroom on price is:

  $$\frac{\Delta price}{\Delta bdrms} = \beta_2 + \beta_3 sqrft$$

  which implies that if $\beta_3 > 0$ then an additional bedroom yields a higher increase in house price in larger houses
- There is an **interaction** effect between square footage and the number of bedrooms

Interaction terms

$$\frac{\Delta price}{\Delta bdrms} = \beta_2 + \beta_3 sqrft$$

- β_2 is equal to the effect of an additional bedroom for a house with zero square footage (!?)
- Need to use sensible values of *sqrft*, e.g. mean (μ)
- Consider the following model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + u$$

which can be re-parameterized as

$$y = \alpha_0 + \delta_1 x_1 + \delta_2 x_2 + \beta_3 (x_1 - \mu_1)(x_2 - \mu_2) + u$$

- This implies that $\delta_2 = \beta_2 + \beta_3 \mu_1$

Interaction terms

- If we expand the last equation:

$$
\begin{aligned}
y =& \alpha_0 + \delta_1 x_1 + \delta_2 x_2 + \beta_3 (x_1 - \mu_1)(x_2 - \mu_2) + u \\
=& \alpha_0 + \delta_1 x_1 + \delta_2 x_2 + \beta_3 x_1 x_2 - \beta_3 \mu_1 x_2 - \beta_3 \mu_2 x_1 + \beta_3 \mu_1 \mu_2 + u \\
=& (\alpha_0 + \beta_3 \mu_1 \mu_2) + (\delta_1 - \beta_3 \mu_2) x_1 + (\delta_2 - \beta_3 \mu_1) x_2 + \beta_3 x_1 x_2 + u
\end{aligned}
$$

  and compare it to the original equation (before adjustment):

$$
y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + u
$$

  you can see that the coefficient on x_2 implies:

$$
\delta_2 - \beta_3 \mu_1 = \beta_2 \Rightarrow \delta_2 = \beta_2 + \beta_3 \mu_1
$$

Attendance and final exam performance example

- Model of standardized final exam performance:

$$
\begin{aligned}
stndfnl = & \beta_0 + \beta_1 atndrte + \beta_2 priGPA + \beta_3 ACT + \beta_4 priGPA^2 \\
& + \beta_5 ACT^2 + \beta_6 priGPA \cdot atndrte + u
\end{aligned}
$$

  where *atndrte* is attendance rate, and *priGPA* is prior GPA
- What is $\frac{\Delta stndfnl}{\Delta atndrte} = \beta_1 + \beta_6 priGPA$?

Attendance and final exam performance example

```
. gen priGPA_sq=priGPA*priGPA

. gen ACT_sq=ACT*ACT

. gen priGPA_atndrte=priGPA*atndrte

. reg stndfnl atndrte priGPA ACT priGPA_sq ACT_sq priGPA_atndrte
```

Source	SS	df	MS
Model	152.001032	6	25.3335053
Residual	512.762536	673	.761905701
Total	664.763568	679	.979033237

Number of obs =	680
F(6, 673) =	33.25
Prob > F =	0.0000
R-squared =	0.2287
Adj R-squared =	0.2218
Root MSE =	.87287

stndfnl	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
atndrte	-.0067129	.0102321	-0.66	0.512	-.0268035	.0133777
priGPA	-1.62854	.4810025	-3.39	0.001	-2.572986	-.6840939
ACT	-.1280394	.098492	-1.30	0.194	-.321428	.0653492
priGPA_sq	.2959046	.1010495	2.93	0.004	.0974945	.4943148
ACT_sq	.0045334	.0021764	2.08	0.038	.00026	.0088068
priGPA_atndrte	.0055859	.0043174	1.29	0.196	-.0028913	.0140631
_cons	2.050293	1.360319	1.51	0.132	-.6206863	4.721272

Attendance and final exam performance example

- Coefficient on *atndrte* is negative (?!) but recall this is only relevant when *priGPA* is zero and the lowest GPA in the sample is 0.86
- Though t statistics for the coefficients on the *atndrte* variables are both small, an F test that both are zero is easily rejected at 5% level
- The mean value of *priGPA* is 2.59 so the effect of *atndrte* on *stndfnl* is $-0.0067 + 0.0056 \cdot 2.59 \approx 0.0078$
  - So a 10 percentage point increase in *atndrte* increases $\widehat{stndfnl}$ by 0.078 standard deviations from the mean final exam score
- Is this estimate statistically different from zero?
  - Use technique from Lecture 5 to test linear restrictions!

Attendance and final exam performance example

- We want to test if $\theta = \beta_1 + \beta_6 2.59 = 0$
- Rearranging: $\beta_1 = \theta - \beta_6 2.59$
- Substituting this into the original equation gives:

$$\begin{aligned} stndfnl =& \beta_0 + \beta_1 atndrte + \beta_2 priGPA + \beta_3 ACT + \beta_4 priGPA^2 \\ & + \beta_5 ACT^2 + \beta_6 priGPA \cdot atndrte + u \\ =& \beta_0 + (\theta - \beta_6 2.59) atndrte + \beta_2 priGPA + \dots \\ & + \dots + \beta_6 priGPA \cdot atndrte + u \\ =& \beta_0 + \theta atndrte + \dots + \beta_6 (priGPA - 2.59) \cdot atndrte + u \end{aligned}$$

- Coefficient on *atndrte* is the estimated effect of attendance at $priGPA = 2.59$
  - Standard error of $\hat{\theta} = \hat{\beta}_1 + \hat{\beta}_6 2.59 = 0.0078$ is 0.0026 and t statistic is $\frac{0.0078}{0.0026} \approx 3$

Attendance and final exam performance example

```
. replace priGPA_atndrte=(priGPA-2.59)*atndrte
(680 real changes made)

. reg stndfnl atndrte priGPA ACT priGPA_sq ACT_sq priGPA_atndrte

      Source |       SS       df       MS              Number of obs =     680
-------------+------------------------------           F(  6,   673) =   33.25
       Model |  152.001032     6  25.3335053           Prob > F      =  0.0000
    Residual |  512.762536   673   .7619057           R-squared     =  0.2287
-------------+------------------------------           Adj R-squared =  0.2218
       Total |  664.763568   679  .979033237           Root MSE      =  .87287

------------------------------------------------------------------------------
       stndfnl |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------------+----------------------------------------------------------------
       atndrte |   .0077546   .0026393     2.94   0.003     .0025723    .0129368
        priGPA |   -1.62854   .4810025    -3.39   0.001    -2.572986   -.6840939
           ACT |  -.1280394    .098492    -1.30   0.194     -.321428    .0653492
     priGPA_sq |   .2959046   .1010495     2.93   0.004     .0974945    .4943148
        ACT_sq |   .0045334   .0021764     2.08   0.038       .00026    .0088068
priGPA_atndrte |   .0055859   .0043174     1.29   0.196    -.0028913    .0140631
         _cons |   2.050293   1.360319     1.51   0.132    -.6206861    4.721273
------------------------------------------------------------------------------
```

Revision questions

1. The following table contains the box office revenue (*rev*) and production budgets (*bdgt*) for three movies.

Movie	*rev* (y_i)	*bdgt* (x_i)
1	10	8
2	11	9
3	12	12

- Estimate the intercept and slope coefficients in the equation:

$$\widehat{rev}_i = \hat{\beta}_0 + \hat{\beta}_1 bdgt_i.$$

- Compute the fitted values and residuals for each observation, and verify that the residuals (approximately) sum to zero.
- How much of the variation in *rev* for these three movies is explained by *bdgt*?
- Verify that $\sum_{i=1}^{3} bdgt_i \hat{u}_i = 0$.

Revision questions

2. The model to be estimated is:

$$\log(rev) = \beta_0 + \beta_1 \log(bdgt) + u.$$

   Stata output is provided in the following slides!

- Interpret the coefficient on $\log(bdgt)$. Be specific.
- List three factors (or omitted variables) that are likey to be contained in u? Explain if they are negatively or positively correlated with *bdgt*.
- Based on your answer above, is the coefficient on $\log(bdgt)$ reported in the Stata output likely to be an unbiased estimate of the ceteris paribus effect of $\log(bdgt)$ on (the logarithm of) *rev*? Explain.

Revision questions

- Suppose the "true" model includes one omitted variable, violence (*viol*, score between 0 and 10). The model is:

  $$\log(rev) = \beta_0 + \beta_1 log(bdgt) + \beta_2 viol + u.$$

  Derive an expression that relates the coefficient on log($bdgt$) from the first model, $\tilde{\beta}_1$, to the coefficient on log($bdgt$) from the true model, $\hat{\beta}_1$. Show your workings.
- Is $\tilde{\beta}_1$ likely to be biased upward or downward?
- Perform an F test to determine if $\beta_2 = 0$ against the alternative hypothesis.

```
. reg lrev lbdgt viol

      Source |       SS       df       MS              Number of obs =    1501
-------------+------------------------------           F(  2,  1498) =  367.48
       Model |  799.573466     2  399.786733           Prob > F      =  0.0000
    Residual |  1629.69742  1498   1.0879155           R-squared     =  0.3291
-------------+------------------------------           Adj R-squared =  0.3282
       Total |  2429.27088  1500  1.61951392           Root MSE      =   1.043

------------------------------------------------------------------------------
        lrev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lbdgt |   .7033798   .0261957    26.85   0.000     .6519956     .754764
        viol |   .0034669   .0121618     0.29   0.776     -.020389    .0273228
       _cons |   14.87872   .1072317   138.75   0.000     14.66838    15.08906
------------------------------------------------------------------------------
```

```
. reg lrev lbdgt

      Source |       SS           df       MS      Number of obs   =      1501
-------------+----------------------------------   F(1, 1499)      =    735.33
       Model |  799.485061         1  799.485061   Prob > F        =    0.0000
    Residual |  1629.78582     1,499  1.08724871   R-squared       =    0.3291
-------------+----------------------------------   Adj R-squared   =    0.3287
       Total |  2429.27088     1,500  1.61951392   Root MSE        =    1.0427

------------------------------------------------------------------------------
        lrev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lbdgt |   .7043321   .0259739    27.12   0.000     .6533831    .7552811
       _cons |   14.89336   .0941133   158.25   0.000     14.70875    15.07797
------------------------------------------------------------------------------
```

Revision questions

3. The model to be estimated is:

$$\log(rev) = \beta_0 + \beta_1 bdgt + \beta_2 viol + \beta_3 sex + u.$$

   Stats output is provided in the following slides!

- Interpret the coefficient on *viol*. Be specific.
- Which variables are statistically significant at the 5% level?
- Formally test the null hypothesis that *viol* has no effect on *lrev* against the alternative that *viol* has a positive effect. Carry out the test at the 5% significance level.
- What is the standard error of $\hat{\beta}_3$? What is its estimated variance, $\widehat{Var}(\hat{\beta}_3)$?
- Confirm that $\widehat{Var}(\hat{\beta}_3) = \frac{\hat{\sigma}^2}{SST_3(1-R_3^2)}$ using the attached Stata output.

Revision questions

```
. reg lrev bdgt viol sex

      Source |       SS           df       MS      Number of obs   =     1501
-------------+----------------------------------   F(3, 1497)      =   247.87
       Model |   806.23221         3  268.74407   Prob > F        =   0.0000
    Residual |  1623.03867     1,497  1.08419417   R-squared       =   0.3319
-------------+----------------------------------   Adj R-squared   =   0.3305
       Total |  2429.27088     1,500  1.61951392   Root MSE        =   1.0412

------------------------------------------------------------------------------
        lrev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        bdgt |   .0155791   .0006446    24.17   0.000     .0143147    .0168436
        viol |   .0208567   .0123909     1.68   0.093    -.0034487    .0451621
         sex |  -.0549841   .0126499    -4.35   0.000    -.0797975   -.0301706
       _cons |     16.707   .0878678   190.14   0.000     16.53464    16.87936
------------------------------------------------------------------------------
```

Revision questions

```
. reg sex bdgt viol

      Source |       SS       df       MS              Number of obs =    1501
-------------+------------------------------           F(  2,  1498) =   96.96
       Model |  877.087353     2  438.543676           Prob > F      =  0.0000
    Residual |  6775.35235  1498  4.52293215           R-squared     =  0.1146
-------------+------------------------------           Adj R-squared =  0.1134
       Total |  7652.43971  1500  5.10162647           Root MSE      =  2.1267

------------------------------------------------------------------------------
         sex |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        bdgt |  -.0151214   .0012573   -12.03   0.000    -.0175876   -.0126551
        viol |   .2054019   .0247454     8.30   0.000     .1568625    .2539413
       _cons |   4.011978    .146505    27.38   0.000     3.724601    4.299354
------------------------------------------------------------------------------
```

Next week

- Midterm exam