ECON2271

Business Econometrics

Or: Practical Econometrics for Beginners

Week 7: Multivariate Regression (ii)

Topic 3(ii): Multivariate Regression

Agenda and learning outcomes

Topic 3: Multivariate Regression: (continuous Y, different types of X)

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_nX_n + e$$

Part (i): Essentials: (recap)

a) Model specification:
   - Understand how model specification affects the interpretation of model coefficients.
   - Be aware of key model specification criteria
   - Specification and robustness

Part (ii): Extensions:

b) Allowing for heterogeneity in associations (interaction terms):
   - Understand when, why and how allow for differences in associations between *X* and *Y*;
   - being able to correctly interpret relevant model estimates.

c) Using alternative estimation techniques to cope with endogenous regressors
   - Students able to explain how 2SLS works
   - Students able to perform 2SLS estimation and interpret results correctly

Topic 3(ii): Multivariate Regression
a) Multivariate regression essentials

A recap of Multivariate Regression Essentials:

- **How does model specification affect estimated coefficients and how we interpret them?**
  - E.g. Consider this general model: LifeSatisfaction = f(Education, Income, ***X***)
  - My key interest here is in understanding the relationship between education and life satisfaction.
  - In theory, education is expected to be positively associated with life satisfaction. However, much of this association is expected to occur via income: people with higher education earn higher incomes, on average, and therefore we expect them to have higher life satisfaction. But are there additional benefits of education, beside the benefits which occur via higher income?
  - If I estimate the model without income, the education variable absorbs variation in income, so the education coefficient will reflect the association between education and life satisfaction, including any effects that occur via higher incomes. Hence, I cannot be sure whether this parameter really reflects the effect of education or the effect of income.
  - By including income as well as education I can evaluate the pure association between education and life satisfaction,
  - Here, ***X*** represents a vector of variables which I include in order to account for other things which could cause bias in key model parameters if I omit them. This might include various demographic variables, labour market characteristics, and anything else I think I should account for. That is: these are my

Topic 3(ii): Multivariate Regression
a) Multivariate regression essentials

A recap of Multivariate Regression Essentials:

- **How do we identify the correct model specification?**
  - There is no black&white rulebook to follow here. The model specification deemed to be most appropriate will depend on a number of factors:
    - What is the objective of the research? What are you REALLY trying to find out?
    - What data do you have access to? (obviously a constraint…)
    - What does theory propose?
    - What does the existing literature suggest?
    - What are you prepared to assume about key relationships?
    - What do the data tell you?

Topic 3(ii): Multivariate Regression
a) Multivariate regression essentials

THE UNIVERSITY OF WESTERN AUSTRALIA
Achieve International Excellence

A recap of Multivariate Regression Essentials:

- **How do we identify the correct model specification?**
  - Studenmund distils four key specification criteria:
    1. *Theory*: Is the variable's place in the equation unambiguous and theoretically sound?
    2. *t-Test*: Is the variable's estimated coefficient significant in the expected direction (i.e. correct sign)?
    3. $\bar{R}^2$: Does the overall fit of the equation (adjusted for degrees of freedom) improve when the variable is added to the equation?
    4. *Bias*: Do other variables' coefficients change significantly when the variable is added to the equation?

Topic 3(ii): Multivariate Regression
a) Multivariate regression essentials

THE UNIVERSITY OF WESTERN AUSTRALIA
Achieve International Excellence

A recap of Multivariate Regression Essentials:

- **Specification and robustness**
  - Often, we observe that a key model parameter is quite sensitive to model specification. This raises a number of questions:
    1. What is the source of this sensitivity? Which variables are causing the model parameter to change?
    2. What is the intuitive explanation behind this sensitivity?
    3. With that in mind, what is the most appropriate model specification?

Topic 3(ii): Multivariate Regression

a) Multivariate regression essentials

A recap of Multivariate Regression Essentials:

- **Specification and robustness**
  - We also use alternative model specifications to check our robust our baseline results are.
  - For example: Say you are working in the productivity commission and are asked to estimate the relationship between wealth and health using cross-sectional data. You estimate a significant positive relationship between wealth and health. Your supervisor is sceptical, however, and asks: How robust is this estimate?
  - A robust estimate is one that is not sensitive to alternative model specifications, methods of measurement, and data sources. If there really is a positive relationship between wealth and health, then we should be able to observe this:
    - Regardless of what control variables are included.
    - Regardless of how health and wealth is measured.
    - Regardless of which data set you use.
  - So you need to be sure you're measuring what you think you are measuring, and not something else.

Topic 3(ii): Multivariate Regression

a) Multivariate regression essentials

The University of Western Australia
Achieve International Excellence

```
. reg ph lnwealth hs vt ug pg lnreinc age marrdef ue nil

      Source |       SS           df       MS      Number of obs   =    14,879
-------------+----------------------------------   F(10, 14868)    =    465.01
       Model |  1732222.35        10  173222.235   Prob > F        =    0.0000
    Residual |  5538560.38    14,868  372.515495   R-squared       =    0.2382
-------------+----------------------------------   Adj R-squared   =    0.2377
       Total |  7270782.72    14,878  488.693556   Root MSE        =    19.301

------------------------------------------------------------------------------
          ph |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lnwealth |   1.816976   .0970777    18.72   0.000     1.626692    2.007261
          hs |   2.003178    .526927     3.80   0.000     .9703362     3.03602
          vt |   .9350519   .4258943     2.20   0.028     .1002465    1.769857
          ug |   3.050033   .4936564     6.18   0.000     2.082406    4.017661
          pg |   3.766319   .7822309     4.81   0.000      2.23305    5.299588
     lnreinc |   1.328053   .2264577     5.86   0.000     .8841683    1.771938
         age |  -.3921182   .0101397   -38.67   0.000    -.4119932   -.3722431
     marrdef |   1.489094   .3597226     4.14   0.000     .7839936    2.194195
          ue |  -4.509817   .8520219    -5.29   0.000    -6.179885   -2.839749
         nil |  -9.147464   .4030142   -22.70   0.000    -9.937421   -8.357506
       _cons |   56.27447   2.391308    23.53   0.000     51.58721    60.96173
------------------------------------------------------------------------------
```

Topic 3(ii): Multivariate Regression
a) Multivariate regression essentials

THE UNIVERSITY OF WESTERN AUSTRALIA
Achieve International Excellence

A recap of Multivariate Regression Essentials:

- **Specification and robustness**
  - For example, your supervisor may ask whether there is some omitted variable bias here: people who accumulate wealth faster than others may have some sort of characteristics that also promotes good health behaviours.
    - ➢ If you have information about personality characteristics, you can include these in your model as additional controls and see whether the estimated wealth-coefficient changes.
  - Or, your supervisor may be critical of your health and/or wealth variables: could your estimates be affected by bias from measurement error?
    - ➢ Investigate the basis for these concerns: how exactly are these variables measured?
    - ➢ Find alternative measures of health (maybe something more objective, like BMI?)
    - ➢ Find alternative measures of wealth (maybe home value?). Is the lin-log specification reasonable?
  - Or, your supervisor may be critical of your sample. How representative are these data?
    - ➢ Find alternative samples and estimate your model based on those data.

```
reg ph lnwealth hs vt ug pg lnreinc age marrdef ue nil pnextrv pnagree pnconsc pnemote pnopene

      Source |       SS           df       MS      Number of obs   =    13,090
-------------+----------------------------------   F(15, 13074)    =    328.14
       Model |  1768438.27        15  117895.884   Prob > F        =    0.0000
    Residual |  4697259.41    13,074    359.2825   R-squared       =    0.2735
-------------+----------------------------------   Adj R-squared   =    0.2727
       Total |  6465697.68    13,089    493.9795   Root MSE        =    18.955

------------------------------------------------------------------------------
          ph |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lnwealth |   1.778926   .1061554    16.76   0.000     1.570846    1.987006
          hs |   1.945465   .5651678     3.44   0.001     .8376538    3.053276
          vt |   1.180441   .4514458     2.61   0.009     .2955414     2.06534
          ug |   2.520112   .5291261     4.76   0.000     1.482948    3.557276
          pg |   3.733946   .8239828     4.53   0.000      2.11882    5.349072
     lnreinc |   1.369488   .2370122     5.78   0.000     .9049093    1.834066
         age |  -.4499304    .011433   -39.35   0.000    -.4723407     -.42752
     marrdef |   1.603539   .3816161     4.20   0.000     .8555162    2.351562
          ue |  -3.057277   .9570725    -3.19   0.001    -4.933278   -1.181275
         nil |  -8.794305   .4262293   -20.63   0.000    -9.629777   -7.958834
     pnextrv |   .5476807   .1597488     3.43   0.001     .2345499    .8608115
     pnagree |    .426691   .2021037     2.11   0.035     .0305384    .8228436
     pnconsc |   1.548499   .1816136     8.53   0.000     1.192511    1.904488
     pnemote |   2.866799   .1762728    16.26   0.000     2.521278    3.212319
     pnopene |  -.2429965   .1760103    -1.38   0.167    -.5880021    .1020092
       _cons |    32.3876   2.834419    11.43   0.000     26.83173    37.94348
------------------------------------------------------------------------------
```

Topic 3(ii): Multivariate Regression
a) Multivariate regression essentials

The University of Western Australia
Achieve International Excellence

A recap of Multivariate Regression Essentials:

- **Specification and robustness**
  - Should the model include personality variables? Why/ why not?
  - Is the association between wealth and health robust with respect to whether or not we control for differences in these personality characteristics?

Topic 3(ii): Multivariate Regression
b) Interaction Terms

The University of Western Australia
Achieve International Excellence

Interaction terms: In a standard regression model, Y = b0 + b1X + u, we estimate one relationship between X and Y, which is captured by the X-coefficient (b1).

- Suppose we suspect that the relationship between X and Y varies systematically for different groups.
  - For example, maybe the relationship between income and financial satisfaction differs between men and women? How do we test this?
  - We could estimate the same model twice, once for men then for women, and then compare the income coefficient.
  - However, it would be neat to estimate the difference between the income coefficient for men and women in the same model. For various reasons…
  - So, instead of:

    $FS_i = b_0 + b_1(\ln Y) + \mathbf{b}_j\mathbf{X}_{ji} + u_i$ estimated for men and women separately…
  - We estimate:

    $FS_i = b_0 + b_1(\ln Y) + b_2(\ln Y)(\text{female}) + \mathbf{b}_j\mathbf{X}_{ji} + u_i$ estimated for everyone.

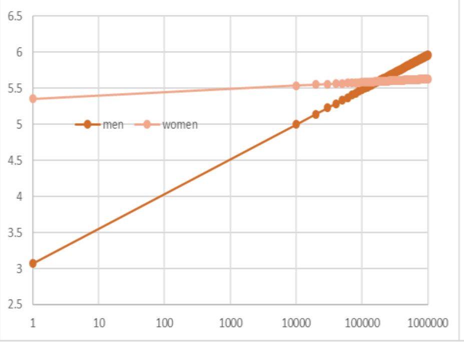
```
. gen pyfemale=female*(lnrealpy)
(2 missing values generated)

. regress fs lnrealpy pyfemale female lnrew age agesq marrdef kids ue nil

      Source |       SS           df       MS      Number of obs   =    10,667
-------------+----------------------------------   F(10, 10656)    =    242.52
       Model |  9137.75362        10  913.775362   Prob > F        =    0.0000
    Residual |   40149.584    10,656  3.76779129   R-squared       =    0.1854
-------------+----------------------------------   Adj R-squared   =    0.1846
       Total |  49287.3376    10,666  4.62097671   Root MSE        =    1.9411

------------------------------------------------------------------------------
          fs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lnrealpy |   .2094968   .0211668     9.90   0.000     .1680059    .2509877
    pyfemale |  -.1886499   .0252828    -7.46   0.000     -.238209   -.1390908
      female |   2.279951   .2945464     7.74   0.000     1.702585    2.857317
       lnrew |   .4018594   .0123933    32.43   0.000     .3775662    .4261525
         age |  -.1829727   .0151252   -12.10   0.000    -.2126209   -.1533245
       agesq |   .0019246   .0001697    11.34   0.000      .001592    .0022572
     marrdef |   .5379621   .0490265    10.97   0.000     .4418611    .6340631
        kids |  -.2305425   .0451601    -5.11   0.000    -.3190646   -.1420204
          ue |   -1.80793    .108816   -16.61   0.000    -2.021229    -1.59463
         nil |  -.4221049     .05445    -7.75   0.000    -.5288371   -.3153726
       _cons |   3.071175   .3968713     7.74   0.000     2.293234    3.849117
------------------------------------------------------------------------------
```

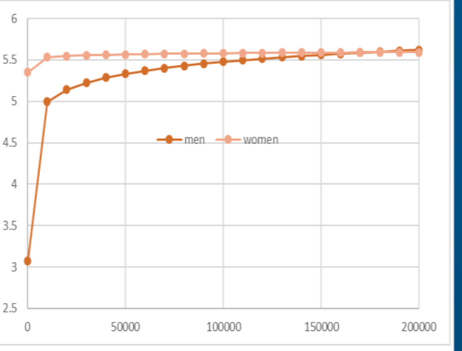
Topic 3(ii): Multivariate Regression
b) Interaction Terms

THE UNIVERSITY OF WESTERN AUSTRALIA
Achieve International Excellence

- Is the relationship between income and financial satisfaction different for women, compared to for men?
  - Oh yes!
    - The coefficient for *female* is positive and statistically significant, estimated at 2.28.
      - Hence, the intercept term is 2.28 higher for women than for men
      - The intercept term for men is the _*cons* coefficient = 3.07
      - So the intercept term for women =
    - The coefficient for *pyfemale* is negative and statistically significant, estimated at -0.19
      - Hence, the slope parameter is lower for women than for men
      - The slope parameter for men is the *lnrealpy* coefficient = 0.21
      - So the slope parameter for women =

Topic 3(ii): Multivariate Regression
b) Interaction Terms





- Do women care about their personal income at all??
  - We can perform a t-test for the total (net) slope for women; or
  - We can estimate the model just for women and evaluate the income slope

The University of Western Australia
Achieve International Excellence

```
. test (lnrealpy+ pyfemale=0)

 ( 1)  lnrealpy + pyfemale = 0

       F(  1, 10656) =    1.89
            Prob > F =    0.1687

. regress fs lnrealpy lnrew age agesq marrdef kids ue nil if female==1
```

Source	SS	df	MS		
Model	4459.69713	8	557.462141	Number of obs =	5,600
Residual	22216.1384	5,591	3.97355364	F(8, 5591) =	140.29
				Prob > F =	0.0000
				R-squared =	0.1672
				Adj R-squared =	0.1660
Total	26675.8355	5,599	4.76439284	Root MSE =	1.9934

fs	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnrealpy	.0311958	.0160354	1.95	0.052	-.0002397	.0626314
lnrew	.3956604	.0169073	23.40	0.000	.3625155	.4288053
age	-.1497585	.0213776	-7.01	0.000	-.1916668	-.1078502
agesq	.0015686	.0002403	6.53	0.000	.0010975	.0020397
marrdef	.6100033	.0675275	9.03	0.000	.4776232	.7423833
kids	-.2465008	.0647594	-3.81	0.000	-.3734545	-.1195471
ue	-1.726368	.1640104	-10.53	0.000	-2.047892	-1.404844
nil	-.3287921	.0692704	-4.75	0.000	-.464589	-.1929952
_cons	4.330118	.5080301	8.52	0.000	3.334181	5.326054

Topic 3(ii): Multivariate Regression
b) Interaction Terms

- Do women care about their personal income at all??
  - The F-test is a joint hypothesis test, testing the null hypothesis that the sum of the two slope coefficients equals zero, i.e. that the total (net) association between income and financial satisfaction = 0 for females.
    - p-value > 0.05
    - Apparently, women's financial satisfaction is not significantly related to their own personal income…
  - When we estimate the model for women only, the income coefficient is very small (0.03) and not statistically significant at the 95% level of confidence.
  - What is going on??

Topic 3(ii): Multivariate Regression
c) 2-Stage Least Squares Estimation

The problem of endogenous regressors (recap):

- Let's say we want to estimate how income (Y) affects health (H). That is, we want to estimate how a change in income will change health, all other things held constant:
  - $H_i = b_0 + b_1(\ln Y_i) + b_j\boldsymbol{X}_j + u_i$
  - Recall: The income variable (lnY) will be endogenous if:
    - The model omits an important variable which is correlated both with income and health;
    - There causality runs in both directions (reverse causality); or
    - The data suffers from non-random measurement error (e.g. people with high income systematically under-report their income).
  - Any of these things will cause income to be correlated with the error term.
  - This will again cause the income coefficient (b_1) to be biased.
  - This happens because the income coefficient will capture the total association between income and health.
    - If health has a positive effect on income, and income has a positive effect on health, b_1 will capture both of these effects and be biased upward.
    - If, for argument's sake, health has a negative effect on income, and income has a positive effect on health, b_1 will capture the net effect and be biased downward

Topic 3(ii): Multivariate Regression
c) 2-Stage Least Squares Estimation

2-Stage Least Squares Estimation: (text reference: Gujarati Ch 20, Section 20.4)

- Estimating how income explains health (b_1):
  - $H_i = b_0 + b_1 \ln Y_i + b_2 \boldsymbol{X}_i + u_i$; but $\ln Y_i = b_3 + b_4 H_i + b_5 \boldsymbol{Z}_i + e_i$
  - We need to be able to separate out the effect of *H* on *Y* to correctly estimate the effect of *Y* on *H*.
  - We may be able to use 2-Stage Least Squares (2SLS), which involves constructive an instrumental variable (IV) within the model:
    1. Stage 1 – construct an IV for Y by regressing Y on all exogenous variables in the model:

       $Y_i = \hat{b}_6 + \hat{b}_7 \boldsymbol{X}_i + \hat{b}_8 \boldsymbol{Z}_i + v_i$

       $\Rightarrow$ $\hat{Y}_i$ = the IV for Y = the part of Y explained by *X*'s and *Z*'s (but not by H; hopefully, that part of the variation, which we don't want to include, is contained in the error term v_i)
    2. Stage 2 – estimate $H_i = b_9 + b_{10}\hat{Y}_i + b_{11}\boldsymbol{X}_i + e_i$
  - If this all goes to plan, the b_{10} estimate is now an unbiased ("pure") estimate of the effect of *Y* on *H*.
  - BUT: this depends on how well the first stage (reduced form) regression is estimated.

Topic 3(ii): Multivariate Regression
c) 2-Stage Least Squares Estimation

2-Stage Least Squares Estimation:

- Let's try…:
  - $H_i = b_0 + b_1 \ln Y_i + b_2 \boldsymbol{X}_i + u_i$; but $\ln Y_i = b_3 + b_4 H_i + b_5 \boldsymbol{Z}_i + e_i$
  - We need to identify the set of exogenous explanatory variables for health (***X***) and income (***Z***). These can overlap, but they can't be equivalent: ***Z*** must include at least one variable which is not included in ***X***.
  - Hard to find a variable that explains income but should not be included in the health model…
  - Here is one possibility: personal income is quite different for married women with children than others, all other things held constant. However, if we compare the health of two women who are otherwise comparable but where one is married with kids and the other is not, one would not expect their health to be all that different. Hence, we can try to generate dummies to identify gender/marital status/kids combinations, and include these in Z but not in X.

Code and first-stage regression…

```
. ivregress 2sls ph age agesq marrdef hs vt ug pg (lnrealpy = marrfemkids marrmalekids), first

First-stage regressions
-----------------------

                                                Number of obs     =      9,919
                                                F(   9,    9909)  =      76.70
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0651
                                                Adj R-squared     =     0.0643
                                                Root MSE          =     1.5632

------------------------------------------------------------------------------
    lnrealpy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0635932   .0123313     5.16   0.000     .0394212    .0877652
       agesq |  -.0007622   .0001378    -5.53   0.000    -.0010323   -.0004922
     marrdef |  -.1514424   .0421208    -3.60   0.000    -.2340076   -.0688771
          hs |   .3246656   .0585747     5.54   0.000     .2098472     .439484
          vt |   .3925045    .044327     8.85   0.000     .3056144    .4793945
          ug |   .7480557   .0481133    15.55   0.000     .6537438    .8423677
          pg |    .826885   .0697852    11.85   0.000     .6900918    .9636782
 marrfemkids |  -.4482989   .0479037    -9.36   0.000       -.5422   -.3543979
marrmalekids |   .5107121   .0482655    10.58   0.000     .4161018    .6053223
       _cons |   8.884731   .2645168    33.59   0.000     8.366224    9.403237
------------------------------------------------------------------------------
```


Second-stage: the 2SLS regression

```
Instrumental variables (2SLS) regression          Number of obs   =      9,919
                                                  Wald chi2(8)    =     932.21
                                                  Prob > chi2     =     0.0000
                                                  R-squared       =     0.0948
                                                  Root MSE        =     19.896

------------------------------------------------------------------------------
          ph |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lnrealpy |   1.565239   .7128784     2.20   0.028     .1680228    2.962455
         age |  -.1745518   .1609002    -1.08   0.278    -.4899104    .1408069
       agesq |  -.0020708    .001788    -1.16   0.247    -.0055752    .0014337
     marrdef |   5.437294   .5016244    10.84   0.000     4.454129     6.42046
          hs |    5.22169   .7783838     6.71   0.000     3.696086    6.747294
          vt |   3.821847   .6421972     5.95   0.000     2.563164     5.08053
          ug |    7.04732   .8047281     8.76   0.000     5.470082    8.624558
          pg |   8.938279   1.070978     8.35   0.000       6.8392    11.03736
       _cons |   63.70818   7.053942     9.03   0.000     49.88271    77.53365
------------------------------------------------------------------------------
Instrumented:  lnrealpy
Instruments:   age agesq marrdef hs vt ug pg marrfemkids marrmalekids
```

Compare this to the standard OLS regression:

The University of Western Australia
Achieve International Excellence

```
. regress ph lnrealpy age agesq marrdef hs vt ug pg

      Source |       SS           df       MS      Number of obs   =     9,921
-------------+----------------------------------   F(8, 9912)      =    130.09
       Model |  412185.903         8  51523.2379   Prob > F        =    0.0000
    Residual |   3925645.9     9,912  396.049828   R-squared       =    0.0950
-------------+----------------------------------   Adj R-squared   =    0.0943
       Total |   4337831.8     9,920  437.281432   Root MSE        =    19.901

------------------------------------------------------------------------------
          ph |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lnrealpy |   1.343885   .1258804    10.68   0.000     1.097133    1.590636
         age |  -.1559312   .1533726    -1.02   0.309    -.4565726    .1447102
       agesq |  -.0022838   .0016943    -1.35   0.178     -.005605    .0010374
     marrdef |   5.412142    .490818    11.03   0.000     4.450039    6.374245
          hs |   5.293894   .7466961     7.09   0.000     3.830218     6.75757
          vt |   3.913244    .566136     6.91   0.000     2.803502    5.022985
          ug |   7.210399    .619329    11.64   0.000     5.996388     8.42441
          pg |   9.124533   .8946471    10.20   0.000     7.370843    10.87822
       _cons |    65.5571   3.498783    18.74   0.000     58.69877    72.41543
------------------------------------------------------------------------------
```


Topic 3(ii): Multivariate Regression

c) 2-Stage Least Squares Estimation

The University of Western Australia
Achieve International Excellence

What did we find?

- If we have correctly identified what variables are exogenous and endogenous, and not broken any other rules, then this experiment appears to show that there really is a positive effect of income on health, even when we try to instrument income by using what we think is a set of exogenous variables.
- However, the first-stage regression is not great… so our instrument (the estimate of PH from the first-stage regression) is pretty weak. This is not a good thing.
- This was just an illustration of how it is supposed to work. It's hard to come up with a really good example that works really well, but textbooks do provide some.