The University of Western Australia
Achieve International Excellence

EFMD EQUIS ACCREDITED

ECON2271

Business Econometrics

Or: Practical Econometrics for Beginners

Week 5: Multivariate Regression (i)

Topic 3(i): Multivariate Regression
Agenda and learning outcomes

The University of Western Australia
Achieve International Excellence

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

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

a) Multivariate regression model essentials:
   - Understand how to interpret the parameters of a multivariate function.
   - Understand how to specify and interpret models which include nominal and ordinal *X*-variables.
   - Understand the difference between R-squared and adjusted R-squared in multivariate models.
   - Students to know how to select variables to include, both intuitively and using statistical tools.
   - Be able to compare estimates across alternative model specifications, and use this information to make reasonable inferences about relationships between variables.
   - Understand the key causes and consequences of endogenous regressors.

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(i): Multivariate Regression
a) Multivariate regression essentials

The University of Western Australia
Achieve International Excellence

Multivariate Regression Essentials:

- Most of the time, when we are interested in understanding variation in Y, we have to acknowledge that Y depends not only on one, but on several different variables:

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

- Key consequences of moving to multivariate regression:
  - For a model with *j* independent (*X*) variables, you have to make sure that $n > j$.
  - The *b*-parameters now specifically reflect ***partial*** derivatives: b_j measures the change in *Y* as X_j changes by 1 unit, if all other *X*-variables remain constant.
  - The estimation of OLS is now a little more complex (but unless you're estimating manually you won't need to worry about that).
  - There is one additional problem we have to deal with: multicollinearity, which occurs when two or more X-variables are highly correlated and the model cannot distinguish between which X is causing what variation in Y.
  - We need to introduce the *adjusted* R-squared as a measure of goodness-of-fit, particularly if comparing competing model specifications.
  - If you don't specify your model correctly, you won't be able to know what's really going on.
  - If you can't interpret your model correctly (which can be tricky!) you will draw the wrong conclusions.

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

The University of Western Australia
Achieve International Excellence

Nominal and ordinal regressors: What to do if X a set of mutually exclusive and complementary dummy variables?

- **Example: The relationship between education and health**
  - We want to find out whether people with higher levels of education have better health.
  - Health is measured by the 0-100 physical health index (PH)
  - Education is measured as the highest level qualification obtained:
    - Year 11 of high-school or less (NHS)
    - Year 12 of high-school (HS)
    - Vocational (trade) qualification (VT)
    - Undergraduate tertiary degree (UG)
    - Postgraduate tertiary degree (PG)
  - Could we just assign ascending numbers to these education levels (1-5) and regress PH on education level?
  - We could, but this forces the incremental differences in PH across education levels to be equal… Is this reasonable? Or useful?
  - If not, we could instead create a set of dummy identifier variables for each education level. These would be complementary and mutually exclusive.

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

- **Nominal and ordinal regressors**
  - Example 1: The relationship between education and health

    $PH_i = b_0 + b_1HS_i + b_2VT_i + b_3UG_i + b_4PG_4 + u_i$

  - ➢ How many variables in this model? Why?
  - ➢ What do the b-parameters capture?

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

- **Nominal and ordinal regressors**
  - ➢ What am I testing here?
    - I could perform pairwise hypothesis tests between b-parameters to see if any two groups are different:
      - ➢ $b_1 = 0$; $b_2 = 0$; $b_3 = 0$; $b_4 = 0$ (these test whether any of the groups are different from the NHS group)
      - ➢ $b_1 = b_2$; $b_1 = b_3$; $b_1 = b_4$; $b_2 = b_3$; $b_2 = b_3$; $b_3 = b_4$ (these test whether any of the other groups are different from one another).
    - I could test whether people with different levels of educational report different PH:
      - ➢ H_0: $b_1 = b_2 = b_3 = b_4 = 0$.
      - ➢ This is a joint hypothesis test.

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

- **Nominal and ordinal regressors**
  - **Example: The relationship between education and health**

    $PH_i = b_0 + b_1HS_i + b_2VQ_i + b_3UG_i + b_4PG_i + u_i$

```
. regress ph hs vt ug pg

      Source |       SS           df       MS      Number of obs   =    15,577
-------------+----------------------------------   F(4, 15572)     =    163.33
       Model |  307861.213         4  76965.3032   Prob > F        =    0.0000
    Residual |  7338138.73    15,572  471.239322   R-squared       =    0.0403
-------------+----------------------------------   Adj R-squared   =    0.0400
       Total |  7645999.94    15,576  490.883407   Root MSE        =    21.708

------------------------------------------------------------------------------
          ph |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          hs |   10.71983   .5553488    19.30   0.000     9.631283    11.80838
          vt |   5.548865   .4515231    12.29   0.000     4.663827    6.433903
          ug |   10.81406   .5091926    21.24   0.000     9.815988    11.81214
          pg |   11.65467   .8303855    14.04   0.000     10.02701    13.28232
       _cons |    68.6909   .3293638   208.56   0.000      68.0453    69.33649
------------------------------------------------------------------------------
```

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

- **Nominal and ordinal regressors**
  - **Example: The relationship between education and health**
    - All group (HS, VT, UG and PG) are statistically different from the reference group (NHS)
    - Seems we have 3 broad levels: NHS has the lowest PH, VT score about 5 points higher, and HS, UG and PG all score about 5 points higher again.
    - First, perhaps we should test whether we are right to lump HS, UG and PG together:

```
. test (_b[hs] = _b[ug]= _b[pg])

 ( 1)  hs - ug = 0
 ( 2)  hs - pg = 0

       F(  2, 15572) =    0.59
            Prob > F =    0.5518
```

    - We use the F-distribution to test two-sided hypotheses about more than one regression coefficient at a time (i.e. joint hypotheses). The critical value for the F-statistic depends on the number of restrictions implied by the null hypothesis (K, here K=2), and the number of degrees-of-freedom (n – K – 1, here df= 15572). You can refer to a table (e.g. the critical value at the 95% level of confidence is 19.5), or you can use the p-value provided by your statistical package.

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

- **The mixed-variable multivariate regression model**
  - Often, if we want to understand the relationship between two variables, we need to know much more about what drives our dependent variable.
  - E.g. We know that health is determined by a whole range of variables:
    - Income, wealth, age, marital status…
  - So if education groups are different with respect to these characteristics, education may be masking the true characteristics determining differences in health.
  - That is, we are right to conclude that people with different levels of education report different health, but we might be wrong in concluding that this is actually about education. It might be that people in the NHS group are much older than anybody else, for example.
  - In other words, we need to compare the physical health of people with different education *but for whom all other such characteristics are the same*! Ceteris paribus!
  - Therefore, we need to specify and estimate a larger model, where we introduce any variables we know (or suspect) will affect health, and we want to hold constant when observing differences across education levels:

$$PH_i = b_0 + b_1HS_i + b_2VT_i + b_3UG_i + b_4PG_i + b_5Female_i + b_6Age_i + b_7\ln(Inc{+}1)_i + b_7\ln(Wth{+}1)_i + b_8Partn_i + u_i$$

  - Here, we have added gender *(Female)*, age, log of income and wealth, and a dummy indicator for individuals who are married or in a de-facto relationship (*Partn*).

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


```
. regress ph hs vt ug pg female age lninc lnwth partn

      Source |       SS           df       MS      Number of obs   =    14,877
-------------+----------------------------------   F(9, 14867)     =    449.33
       Model |  1554728.08         9  172747.565   Prob > F        =    0.0000
    Residual |  5715696.39    14,867  384.455263   R-squared       =    0.2138
-------------+----------------------------------   Adj R-squared   =    0.2134
       Total |  7270424.48    14,876  488.735176   Root MSE        =    19.608

------------------------------------------------------------------------------
          ph |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          hs |   3.172676   .5325074     5.96   0.000     2.128896    4.216456
          vt |   2.492377   .4269131     5.84   0.000     1.655575     3.32918
          ug |   4.993481     .49378    10.11   0.000     4.025611    5.961351
          pg |   5.960783   .7879749     7.56   0.000     4.416255    7.505311
      female |   -1.62763   .3248575    -5.01   0.000    -2.264391   -.9908696
         age |  -.4898287   .0094001   -52.11   0.000    -.5082541   -.4714034
       lninc |   2.195591   .2261228     9.71   0.000     1.752362    2.638819
       lnwth |   2.034765   .1001552    20.32   0.000     1.838449    2.231082
       partn |    2.97935   .3614856     8.24   0.000     2.270794    3.687907
       _cons |   44.79887   2.363374    18.96   0.000     40.16637    49.43138
------------------------------------------------------------------------------
```

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

- **What can we learn from this?**

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

The University of Western Australia
Achieve International Excellence

- **What can we learn from this?**
  - More specifically:
    - If we thought that people with higher education are healthier because they have higher incomes and wealth levels, we are wrong… When we hold these and other variables constant, the association between education and health weakens, so it is possible that these factors account for some of this effect, but some differences persist across education levels.
    - If we thought that people with higher education are healthier because they tend to be younger, we are wrong… When we hold age and other variables constant, the association between education and health weakens, so it is possible that that age accounts for some of this effect, but some differences across education levels persist.

Topic 3(i): Multivariate Regression

a) Multivariate regression essentials

THE UNIVERSITY OF WESTERN AUSTRALIA
Achieve International Excellence

- **What can we learn from this?**
  - If we want to find out whether it is age or economic circumstances which are the key drivers of the observed difference in health across education levels, we need to introduce these variables separately into the model, and determine when and how the education coefficients change.

```
. regress ph hs vt ug pg age

      Source |       SS           df       MS      Number of obs   =    15,577
-------------+----------------------------------   F(5, 15571)     =    630.37
       Model |  1287138.69         5  257427.739   Prob > F        =    0.0000
    Residual |  6358861.25    15,571  408.378476   R-squared       =    0.1683
-------------+----------------------------------   Adj R-squared   =    0.1681
       Total |  7645999.94    15,576  490.883407   Root MSE        =    20.208

------------------------------------------------------------------------------
          ph |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          hs |    5.17271   .5292482     9.77   0.000     4.135322    6.210098
          vt |   4.795491   .4206117    11.40   0.000     3.971043    5.619939
          ug |   9.104614   .4752994    19.16   0.000     8.172972    10.03626
          pg |   10.88378   .7731797    14.08   0.000     9.368258     12.3993
         age |  -.4296662   .0087742   -48.97   0.000    -.4468647   -.4124677
       _cons |   89.56126   .5250258   170.58   0.000     88.53215    90.59037
------------------------------------------------------------------------------
```

Topic 3(i): Multivariate Regression

a) Multivariate regression essentials

THE UNIVERSITY OF WESTERN AUSTRALIA
Achieve International Excellence

- **What can we learn from this?**

```
. regress ph hs vt ug pg age lninc

      Source |       SS           df       MS      Number of obs   =    15,577
-------------+----------------------------------   F(6, 15570)     =    597.17
       Model |  1430370.13         6  238395.021   Prob > F        =    0.0000
    Residual |  6215629.81    15,570  399.205511   R-squared       =    0.1871
-------------+----------------------------------   Adj R-squared   =    0.1868
       Total |  7645999.94    15,576  490.883407   Root MSE        =     19.98

------------------------------------------------------------------------------
          ph |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          hs |   4.558207   .5242752     8.69   0.000     3.530567    5.585847
          vt |    3.83879   .4189169     9.16   0.000     3.017664    4.659916
          ug |   7.108442   .4816026    14.76   0.000     6.164445    8.052439
          pg |   8.419096   .7754418    10.86   0.000      6.89914    9.939052
         age |  -.4171162   .0087004   -47.94   0.000      -.43417   -.4000623
       lninc |   3.909969   .2064203    18.94   0.000     3.505361    4.314577
       _cons |   48.06245   2.251517    21.35   0.000     43.64921    52.47569
------------------------------------------------------------------------------
```

Topic 3(i): Multivariate Regression

a) Multivariate regression essentials

- **What can we learn from this?**

```
. regress ph hs vt ug pg age lninc lnwth

      Source |       SS           df       MS      Number of obs   =    14,885
-------------+----------------------------------   F(7, 14877)     =    559.00
       Model |  1515566.38         7  216509.483   Prob > F        =    0.0000
    Residual |  5762118.63    14,877  387.317243   R-squared       =    0.2082
-------------+----------------------------------   Adj R-squared   =    0.2079
       Total |  7277685.01    14,884  488.960293   Root MSE        =     19.68

------------------------------------------------------------------------------
          ph |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          hs |   3.279219   .5340505     6.14   0.000     2.232414    4.326024
          vt |   3.255181   .4212224     7.73   0.000     2.429534    4.080829
          ug |   5.640601   .4891421    11.53   0.000     4.681822     6.59938
          pg |   6.813405   .7854116     8.67   0.000     5.273901    8.352908
         age |  -.4746127   .0092186   -51.48   0.000    -.4926823   -.4565431
       lninc |   2.335761   .2240628    10.42   0.000      1.89657    2.774952
       lnwth |   2.058722   .1004704    20.49   0.000     1.861787    2.255656
       _cons |    42.9883   2.339191    18.38   0.000      38.4032     47.5734
------------------------------------------------------------------------------
```

Topic 3(i): Multivariate Regression

a) Multivariate regression essentials

- **What can we learn from this?**
  - Seems that we can't blame health differences across education levels on age.. However, some of it can be explained by income and wealth.
  - I think that what is interesting about this is that there appears to be a direct positive effect of education on health, even after we control for all these factors.
  - If we compare two individuals of same age, gender, marital status, income and wealth – but where one has a higher level of education than the other, the person with the higher education will report better health.

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

THE UNIVERSITY OF WESTERN AUSTRALIA
Achieve International Excellence

- **But there are some very important caveats:**
  - We can only compare ACROSS individuals here.
  - We can only say that people who have different education have different health, after controlling for age, gender, income, wealth, and marital status.
  - Suggesting that this implies that more education actually leads to better health outcomes is a long shot – at least in this context (cross-sectional analysis in well-developed economy).
  - We need to ask: are there other things that could explain an association between education and health?
    - Education is not a random variable: The amount of education a person pursues is determined by:
      - access - to what extent is education a feasible option?; and
      - choice - if you have the option to invest in more education, what determines your choice of whether or not to do so?
    - It is possible that people who take particular education paths have particular characteristics which are associated also with health.
    - If so, it would be wrong to conclude that more education produces better health outcomes.
    - This is called **omitted variable bias**, and it is an important potential problem!

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

THE UNIVERSITY OF WESTERN AUSTRALIA
Achieve International Excellence

- **Omitted variable bias:**
  - Say we estimate a model Y = b0 + b1X + u.
  - You can imagine that Y is health and X is education.
  - Let's say there is another key variable determining health, which is not observed: Z.
    - You can imagine that Z is some personal characteristic, such as conscientiousness: people who are characterise as conscientious are well-organised, dedicated to their work, take their responsibilities seriously, and are able to commit to things. They have an internal (rather than external) locus of control.
    - People with higher education are more likely to score higher on "conscientiousness" than others. This makes sense: conscientious people would seem better suited to pursuing education.
    - People who score higher on "conscientiousness" tend also to be healthier. This also makes sense: conscientious people are probably more likely to take their health seriously, commit to healthy behaviours and develop healthy habits.
  - This means that, in the simple model above, this important information (Z) is omitted. Hence, the error term will contain this key information. However, some of this information is correlated with X (education). Consequently, X is correlated with u, which means it is endogenous. This yields a biased estimate of b1: the model will attribute variation in health to variation in education, when some of this variation is in fact attributable to Z – the omitted variable.

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

THE UNIVERSITY OF WESTERN AUSTRALIA
Achieve International Excellence

- **Other possible problems: reverse causality or simultaneity**
  - Are there *yet* other things that could explain an association between education and health?
    - Is it possible that health can determine education?
      - What if people with health issues fare worse at school, or find their health problems to be a significant barrier to pursuing further training or education?
    - If so, we have a problem of reverse causality or simultaneity: this happens when X causes Y, but Y also causes X, or when X and Y move simultaneously.
      - This problem will also make it difficult to determine exactly how much of the association between X and Y is attributed to the causal effect of X on Y.
      - This problem also causes X to be correlated with u, and the coefficient for X to be biased.
  - So, omitted variables and reverse causality (and simultaneity) can cause some serious problems. What can be done?
    - If the problem is omitted information, the most obvious solutions is to try to get hold of this information…
    - If this is not possible, or the problem is reverse causality (or simultaneity), we need to look for other statistical techniques…

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

THE UNIVERSITY OF WESTERN AUSTRALIA
Achieve International Excellence

- **Can we improve our health model?**
  - It just so happens that wave 13 of the HILDA data contains information about the Big5 personality traits, which includes:
    - Extraversion
    - Agreeableness
    - Conscientiousness
    - Emotional stability
    - Openness to experience
  - So, for all the individuals in wave 14 who also were captured in wave 13, I can "borrow" this information from the previous wave. This information is then not current, but it is only one year old, and there is lots of research demonstrating that personality is fairy stable in the short-to-medium run.
  - Will the positive association between education and health disappear if I control for personality?

Topic 3(i): Multivariate Regression

a) Multivariate regression essentials

THE UNIVERSITY OF WESTERN AUSTRALIA
Achieve International Excellence

```
. regress ph hs vt ug pg female age lninc lnwth partn pnextrv pnagree pnconsc pnemote pnopene

      Source |       SS           df       MS      Number of obs   =    13,089
-------------+----------------------------------   F(14, 13074)    =    318.06
       Model |  1642678.82        14  117334.202   Prob > F        =    0.0000
    Residual |  4823017.32    13,074  368.901431   R-squared       =    0.2541
-------------+----------------------------------   Adj R-squared   =    0.2533
       Total |  6465696.14    13,088  494.017126   Root MSE        =    19.207

------------------------------------------------------------------------------
          ph |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          hs |   2.607395   .5714861     4.56   0.000     1.487199    3.727591
          vt |   2.281519   .4536271     5.03   0.000     1.392344    3.170694
          ug |   4.141362   .5304874     7.81   0.000      3.10153    5.181195
          pg |   5.591682    .829528     6.74   0.000     3.965687    7.217677
      female |  -2.733549   .3566742    -7.66   0.000    -3.432682   -2.034416
         age |  -.5567844   .0105021   -53.02   0.000      -.57737   -.5361987
       lninc |   2.146555   .2361381     9.09   0.000      1.68369     2.60942
       lnwth |   2.006655   .1092507    18.37   0.000     1.792508    2.220803
       partn |   2.812935   .3829812     7.34   0.000     2.062237    3.563634
     pnextrv |   .7194922   .1623273     4.43   0.000      .401307    1.037677
     pnagree |   .8866264   .2118651     4.18   0.000     .4713399    1.301913
     pnconsc |   1.732453   .1841087     9.41   0.000     1.371574    2.093333
     pnemote |   2.713672   .1791543    15.15   0.000     2.362503     3.06484
     pnopene |  -.5038454    .180377    -2.79   0.005    -.8574106   -.1502801
       _cons |   21.18897   2.792746     7.59   0.000     15.71478    26.66316
```

Topic 3(i): Multivariate Regression

a) Multivariate regression essentials

THE UNIVERSITY OF WESTERN AUSTRALIA
Achieve International Excellence

- **What can we learn from this?**
  - The model R-squared…
  - But the education coefficients..
  - However, it is clear that personality…
  - But it is not conscientiousness which dominates: it is…
  - And openness to experience has a negative association with health… Why?

Topic 3(i): Multivariate Regression

a) Multivariate regression essentials

THE UNIVERSITY OF WESTERN AUSTRALIA
Achieve International Excellence

- **Some important notes about model specification:**
  - If the model isn't specified correctly, we may draw the wrong conclusions.
  - A well specified model will not be *underspecified*: it will include all relevant variables, and not omit any important information (especially if this information could be correlated with the model regressors)
  - But a model that includes variables which aren't key to explaining variation in the regressand (Y) is *overspecified*: This will most commonly cause problems with variance (i.e. loss of efficiency/estimation precision).
  - How do we choose the "prefect" model specification?
    - Use theory, prior research and intuition: include what SHOULD matter.
    - Use statistics: include variables which appear to be important, and discard those that appear irrelevant (unless there is a good reason for including them anyway).
    - If you have competing models, you can compare the adjusted R-squared directly: the model with the highest adjusted R-squared is best at explaining Y.
      - You can also use a special F-test to compare the fit of two alternative models, but if the adjusted R-squared is noticeably different the F-stat will almost always give you the same conclusion.