THE UNIVERSITY OF WESTERN AUSTRALIA
Achieve International Excellence

EFMD EQUIS ACCREDITED

ECON2271

Business Econometrics

(Introductory Econometrics)

Week 2:Topic 2 (i)

Topic 2:Univariate Regression and OLS

Agenda and learning outcomes

THE UNIVERSITY OF WESTERN AUSTRALIA
Achieve International Excellence

- **Topic 2(i): Univariate Regression:** $Y = b_0 + b_1X + u$

  a) Introduce the classic univariate linear regression model
     - ➢ Able to distinguish between model variables and parameters;
     - ➢ Interpret meaning of model parameters and error term

  b) Using univariate simple regression with single dummy to test for differences in group means [Continuous Y, single binary X]
     - ➢ Students able to test simple hypotheses comparing means using dummy regression

  c) Using OLS to estimating intercept and slope for a linear function [Continuous Y, single continuous X]
     - ➢ Understand the basic premise of OLS.
     - ➢ Understand how we evaluate goodness of fit (R-squared and F-stat)
     - ➢ Students able to estimate simple univariate regression model and interpret key output correctly.

  d) The classical assumptions behind OLS

  e) Working with nonlinear relationships

Topic 2:Univariate Regression and OLS
a) The Univariate Model

- **The general form linear univariate function:** $Y = b_0 + b_1X$
  - Y: Dependent variable, the value of which depends on X
  - X: Independent or explanatory variable = the value of which determines the value of Y
  - b_0: The constant or intercept term = the value which Y will take when X = 0
  - b_1: Captures the change* in Y when X changes by one unit (= *slope* if X is a continuous variable; *discrete change* if X is a discrete variable)

    * Note on terminology: the term *change* here refers to the *difference* we observe in Y when we change the value of X by 1. We must be careful not to wrongfully imply causation where we really mean association. More on this later.
  - We can use this as a theoretical model for an imagined closed system, where all relevant information is known, there is no room for error: Y only depends on X and nothing else can be going on.
  - This works well in controlled environments ("in the lab") but many environments are messy and complex (e.g. weather, economics). Then, we must try to generalize from a messy and complicated world, about which we know only some information. We dig around to look for patterns in this mess. In economics, we use econometric models for this, which means we need to account for error (denoted *e* or *u*)

Topic 2:Univariate Regression and OLS
a) The Univariate Model

- **The econometric (statistical) version:**

$$Y_i = b_0 + b_1X_i + u_i$$

  - Y_i: The observed value of Y for unit (e.g. individual) *i*
  - X_i: The value of X for unit (e.g. individual) *I*
  - b_0 and b_1: Parameters to be estimated; their true values are not known!
  - u_i: The part of Y_i that is not explained by ($b_0 + b_1X_i$), i.e. the error, disturbance, noise…

➢ We obtain some data for Y and X, and estimate the model parameters**:**

$$\widehat{Y_i} = \widehat{b_0} + \widehat{b_1}X_i \;\;\Rightarrow\;\; Y_i = \widehat{b_0} + \widehat{b_1}X_i + \widehat{u_i}$$

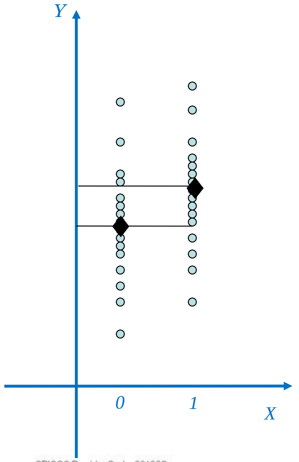
Topic 2:Univariate Regression and OLS
a) The Univariate Model

- We can evaluate the "goodness of fit" of our model by the size of the residuals ($\hat{u}_i$'s): How much of the Y_i's are explained by $\hat{Y}_i = \widehat{b_0} + \widehat{b_1}X_i$?
  - 100%: our model is perfect – we can explain all the variation in *Y*!!
  - 0%: our model can't explain any of the variation in *Y*…
  - BUT: we're not always concerned about explaining as much of *Y* as possible; sometimes we're more interested in finding out what we can and cannot assume about the true value of if b_0 and b_1, given a desired level of confidence (i.e. probability).

- In this topic we restrict ourselves to situations where *Y* is approximately continuous (and unlimited and cardinal…)
  1. First, we look at a model where *X* is binary,
  2. Second, we look at a model where *X* is (approximately) continuous and unlimited

Topic 2:Univariate Regression and OLS
b) The Single Dummy Variable Model

- **Movements in conditional means:** Continuous Y, binary X



$$Y_i = b_0 + b_1 X_i + u_i$$

Y_i = A continuous variable (e.g. wages)

X_i = Binary variable (e.g. male = 1, female = 0)

b_0 = Conditional mean (expected) Y when X = 0:

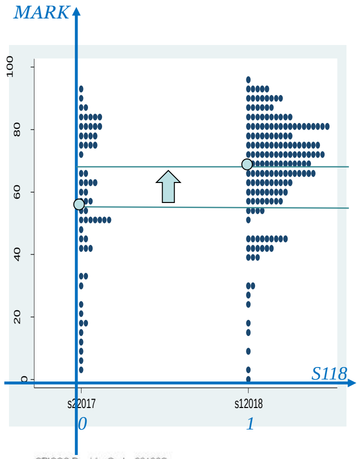
b_1 = Shift in the conditional mean (expected) Y as we move from X = 0 to X = 1 (e.g. from female to male)

u_i = The difference between the expected Y for unit (e.g. individual) *i*, given the value of X, and the actual value of Y for unit *i*

Topic 2:Univariate Regression and OLS

b) The Single Dummy Variable Model

- **Example 1**: Revisit Example 2 from Lecture 1 (ECON1111 marks across two semesters)
  - I combine these sets of marks into one variable (MARK) and create a binary (0/1) dummy variable to indicate whether the mark is from semester 1 2018 (S118).



$$MARK_i = b_0 + b_1(S118_i) + u_i$$

$MARK_i$ = ECON1111 mark for student *i*.

$S118_i$ = Binary variable where 1 indicates the student *i* did ECON1111 in sem 1 2018, and 0 otherwise (i.e. sem 2 1017).

b_0 = Conditional mean (expected) mark when S118 = 0;

b_1 = Shift in the conditional mean (expected) mark as we move from S118 = 0 to S118 = 1

u_i = The difference between the expected mark for student *i*, given the value of S118, and the actual mark for student *i*

Topic 2:Univariate Regression and OLS

b) The Single Dummy Variable Model

THE UNIVERSITY OF WESTERN AUSTRALIA
Achieve International Excellence

$$MARK_i = b_0 + b_1(S118_i) + u_i$$

We can estimate this model and generate: $MARK_i = \hat{b}_0 + \hat{b}_1(S118) + u_i$

$\hat{b}_0$ = Our estimate for b_0 = the conditional mean ($\overline{MARK}$ | S118 = 0)

$\hat{b}_1$ = Our estimate for b_1 = the difference in the conditional means when we move from S118 = 0 to S118 = 1: ($\overline{MARK}$ | S118 = 1) - ($\overline{MARK}$ | S118 = 0)

$\widehat{MARK}_i$ = Our estimate of student *i*'s mark, given the semester s/he studied in and the mean marks for each semester.

$$\widehat{MARK}_i = \hat{b}_0 + \hat{b}_1(S118_i) \;\Rightarrow\; u_i = MARK_i - \widehat{MARK}_i$$

Recall: H_0: these means are not different; b1 = 0
H_1: these means are different; b1 $\neq$ 0

Topic 2:Univariate Regression and OLS

b) The Single Dummy Variable Model

THE UNIVERSITY OF WESTERN AUSTRALIA
Achieve International Excellence

- **Model estimate from Stata:**

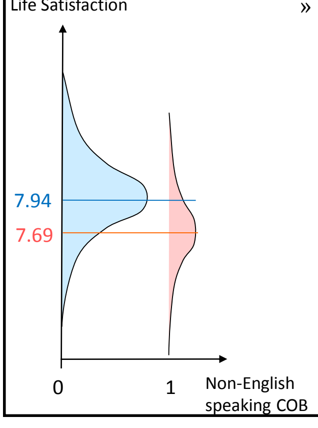
```
. regress MARK S118

      Source |       SS           df       MS      Number of obs   =       240
-------------+----------------------------------   F(1, 238)       =     12.57
       Model |  4781.46497         1  4781.46497   Prob > F        =    0.0005
    Residual |  90534.5184       238  380.397136   R-squared       =    0.0502
-------------+----------------------------------   Adj R-squared   =    0.0462
       Total |  95315.9833       239  398.811646   Root MSE        =    19.504

------------------------------------------------------------------------------
        MARK |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        S118 |   10.25167   2.891564     3.55   0.000     4.555344      15.948
       _cons |    57.2623   2.497202    22.93   0.000     52.34285    62.18174
------------------------------------------------------------------------------
```


Topic 2:Univariate Regression and OLS

b) The Single Dummy Variable Model

THE UNIVERSITY OF WESTERN AUSTRALIA
Achieve International Excellence

- **Example 2:** Revisit the last exercise from Lecture 1
  - A sample of individuals for whom we have data on life satisfaction and information about whether they are born in an English-speaking country or not.
  - We found that these two groups are different, and that we can be 95% confident that this difference is not explained by randomness.



» $LS_i = b_0 + b_1(NES_i) + u_i$

LS_i = Life satisfaction of individual *i*.

NES_i = Binary variable where 1 indicates Non-English-Speaking COB, and 0 otherwise (i.e. English-speaking COB), for individual *i*.

b_0 = $\overline{LS}$ | NES = 0 (= mean LS when NES = 0)

b_1 = ($\overline{LS}$ | NES = 1) - ($\overline{LS}$ | NES = 0) (= shift in conditional mean when moving from NES=0 to NES=1)

u_i = The difference between the expected LS for individual *i*, given the value of NES, and the actual LS of individual *i*

Topic 2:Univariate Regression and OLS

b) The Single Dummy Variable Model


```
. reg LS NES

      Source |       SS           df       MS      Number of obs   =    17,498
-------------+----------------------------------   F(1, 17496)     =     52.93
       Model |  112.002229         1  112.002229   Prob > F        =    0.0000
    Residual |  37024.5106    17,496  2.11617001   R-squared       =    0.0030
-------------+----------------------------------   Adj R-squared   =    0.0030
       Total |  37136.5128    17,497  2.12245029   Root MSE        =    1.4547

------------------------------------------------------------------------------
          LS |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         NES |  -.2460357   .0338189    -7.28   0.000    -.3123242   -.1797472
       _cons |   7.939331   .0117243   677.17   0.000      7.91635    7.962312
------------------------------------------------------------------------------
```


Topic 2:Univariate Regression and OLS

b) The Single Dummy Variable Model

The University of Western Australia
Achieve International Excellence

- **A note on model diagnostics**: Marks example

```
. regress MARK S118

      Source |       SS           df       MS      Number of obs   =       240
-------------+----------------------------------   F(1, 238)       =     12.57
       Model |  4781.46497         1  4781.46497   Prob > F        =    0.0005
    Residual |  90534.5184       238  380.397136   R-squared       =    0.0502
-------------+----------------------------------   Adj R-squared   =    0.0462
       Total |  95315.9833       239  398.811646   Root MSE        =    19.504

------------------------------------------------------------------------------
        MARK |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        S118 |   10.25167   2.891564     3.55   0.000     4.555344      15.948
       _cons |    57.2623   2.497202    22.93   0.000     52.34285    62.18174
------------------------------------------------------------------------------
```

- R-squared: The proportion of variation in *Y* which is explained by *X*
- Model F-statistic: tests whether the model explains a significant proportion of the variation in Y (so it's a test statistic for the R-squared, essentially)

  - Here: R-squared is quite low – only 5% of the variation in marks is explained by which semester you're studying in.
  - But is the objective here to explain as much variation in students marks as possible? No!
  - The objective here is only to see if the conditional means are statistically different (i.e. to test if $b_1 = 0$)

Topic 2:Univariate Regression and OLS

b) The Single Dummy Variable Model

➢ **A note on model diagnostics**: Life Satisfaction example:

```
. reg LS NES

      Source |       SS           df       MS      Number of obs   =    17,498
-------------+----------------------------------   F(1, 17496)     =     52.93
       Model |  112.002229         1  112.002229   Prob > F        =    0.0000
    Residual |  37024.5106    17,496  2.11617001   R-squared       =    0.0030
-------------+----------------------------------   Adj R-squared   =    0.0030
       Total |  37136.5128    17,497  2.12245029   Root MSE        =    1.4547

------------------------------------------------------------------------------
          LS |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         NES |  -.2460357   .0338189    -7.28   0.000    -.3123242   -.1797472
       _cons |   7.939331   .0117243   677.17   0.000      7.91635    7.962312
------------------------------------------------------------------------------
```

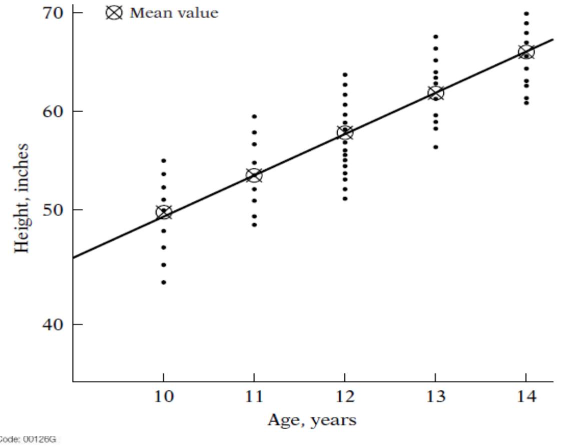
– R-squared is pretty low… what does this signify?
  - Whether or not you are born in an English-speaking country matters, statistically, in terms of your life satisfaction, but it doesn't explain much of the variation in life satisfaction across individuals on its own (<1%).
– Does it matter?
  - Not unless the point of the exercise is to try to explain as much variation in LS as possible.
  - Here, the point is rather to see whether there is a significant difference between two groups.

Topic 2:Univariate Regression and OLS

c) The Classic Linear Regression Model

- **Movements in conditional means:** Continuous Y, discrete X

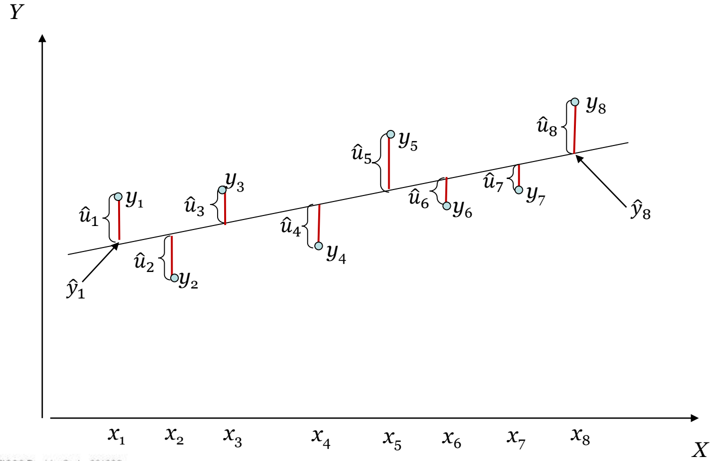
$$Y_i = b_0 + b_1 X_i + u_i$$



Topic 2:Univariate Regression and OLS
c) The Classic Linear Regression Model

- **The classical linear regression model:** Continuous Y, continuous X

$$Y_i = b_0 + b_1 X_i + u_i$$



Topic 2:Univariate Regression and OLS
c) The Classic Linear Regression Model

- **The classical linear regression by Ordinary Least Squares:**

$$Y_i = b_0 + b_1 X_i + u_i$$

- How it works:
  - We have some data for X and Y, which we assume are linearly related.
  - The aim of OLS is to estimate the value of the model parameters (the b's) by identifying the regression line that fits the data the best
  - Any estimate of the b-parameters will yield a predicted value of Y, given the corresponding value of X (or "conditional X"):

Topic 2:Univariate Regression and OLS

c) The Classic Linear Regression Model

The University of Western Australia
Achieve International Excellence

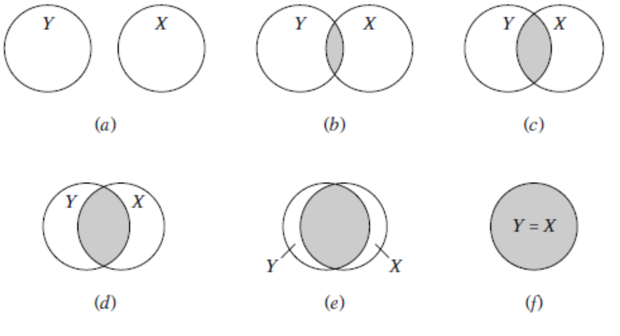
- **The Error (disturbance or residual) Term:**
  - The error term of a regression model reflects, or captures, the unexplained variation in the dependent variable: the portion of variation in Y (the dependent variable) that is not explained by X (the explanatory variable/s).
  - Therefore, the observed (total) variation in the dependent variable can be split into the:
    - Explained component, $\hat{Y}$,
    - Unexplained component, $\hat{u}$.
  - Hence, we can obtained explained, unexplained and total values for each pair of X and Y, square these, and thus obtain total sum of squares (SST), explained sum of squares (SSE), and residual sum of squares (SSR).

Topic 2:Univariate Regression and OLS

c) The Classic Linear Regression Model

The University of Western Australia
Achieve International Excellence

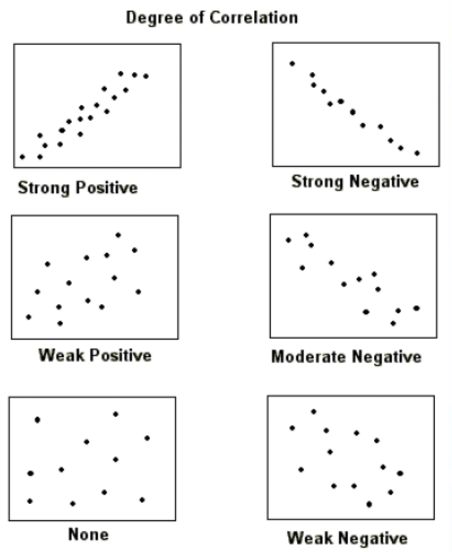
- **The Coefficient of Determination: R-squared**
  - The R-squared (R^2 *or* r^2) statistic of the univariate regression estimate tells you the proportion of variation in Y which is explained by variation in X:
    - R-squared = SSR/SST.



Topic 2:Univariate Regression and OLS

c) The Classic Linear Regression Model

- **Slope and correlation:**
  - The slope of the regression line = $\frac{\Delta Y}{\Delta X}$
  - The coefficient of correlation (*r*) between Y and X gives you a measure of how closely they are aligned.
  - *r* is bounded by by -1 (=perfectly negatively correlated) and +1 (=perfectly positively correlated).
  - $r = \pm\sqrt{r^2}$
  - So *r* and r^2 are closely connected but not the same!



Topic 2:Univariate Regression and OLS

c) The Classic Linear Regression Model

- **Example 3:** Estimating the health-income gradient
  - The health-income gradient refers to the association between income and health
  - I obtain some income and health data from a large sample of Australian households (2014):
    - PH_i = The physical health index for individual *i*; PH is a 0-100 index aggregated from responses to a set of survey questions;
    - INC_i = Equivalised annual disposable (after-tax) household income of individual *i*;
  - I estimate the health-income gradient by estimating the simple linear regression model:

$$PH_i = b_0 + b_1(INC_i) + u_i$$

Topic 2:Univariate Regression and OLS

c) The Classic Linear Regression Model

- **Example 3:** Estimating the health-income gradient

```
. regress ph realeqinc

      Source |       SS           df       MS      Number of obs   =    15,586
-------------+----------------------------------   F(1, 15584)     =    388.09
       Model |  185890.094         1  185890.094   Prob > F        =    0.0000
    Residual |  7464497.05    15,584  478.984667   R-squared       =    0.0243
-------------+----------------------------------   Adj R-squared   =    0.0242
       Total |  7650387.15    15,585  490.881434   Root MSE        =    21.886

------------------------------------------------------------------------------
          ph |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   realeqinc |   .0000751   3.81e-06    19.70   0.000     .0000676    .0000826
       _cons |   70.67294   .2749388   257.05   0.000     70.13402    71.21185
------------------------------------------------------------------------------
```


Topic 2:Univariate Regression and OLS

c) The Classic Linear Regression Model

- **Example 3:** What happens if we scale the variables?
  - Sometimes we might want to scale a variable to reflect very large or very small values. Here, PH is measured on a 0-100 scale, while INC has a huge range:

```
. summarize realeqinc, detail

                          realeqinc
-------------------------------------------------------------
      Percentiles      Smallest
 1%       7953.2              0
 5%     17694.17              0
10%     21796.15              0       Obs              23,107
25%     30518.57              0       Sum of Wgt.      23,107

50%     44927.33                      Mean           53547.47
                        Largest       Std. Dev.      42748.76
75%      64288.8         834071
90%        91050         834071       Variance       1.83e+09
95%       112484         834071       Skewness       5.994644
99%     197658.3         834071       Kurtosis       71.47739
```

  - See what happens if I generate a new variable I call INC = realequinc/10,000, and use this instead:

Topic 2:Univariate Regression and OLS

c) The Classic Linear Regression Model

- **Example 3:** What happens if we scale the variables?

```
. generate inc = realeqinc/10000

. regress ph inc

      Source |       SS           df       MS      Number of obs   =    15,586
-------------+----------------------------------   F(1, 15584)     =    388.09
       Model |  185890.095         1  185890.095   Prob > F        =    0.0000
    Residual |  7464497.05    15,584  478.984667   R-squared       =    0.0243
-------------+----------------------------------   Adj R-squared   =    0.0242
       Total |  7650387.15    15,585  490.881434   Root MSE        =    21.886

------------------------------------------------------------------------------
          ph |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         inc |   .7510782   .0381257    19.70   0.000     .6763474     .825809
       _cons |   70.67294   .2749388   257.05   0.000     70.13402    71.21185
------------------------------------------------------------------------------
```

  – Note: Now the "inc" coefficient reflects the change in mean PH as we compare people whose incomes are $10,000 higher. (SE and CI also change to reflect the new scale)
  – Nothing else changes.

Topic 2:Univariate Regression and OLS

c) The Classic Linear Regression Model

THE UNIVERSITY OF WESTERN AUSTRALIA
Achieve International Excellence

- **Example 3:** What happens if we restrict the sample?
  – Suppose I want to restrict the sample to only include individuals with an income less than $100,000. Why might I want to do this?

```
. regress ph inc if inc<10

      Source |       SS           df       MS      Number of obs   =    14,278
-------------+----------------------------------   F(1, 14276)     =    919.39
       Model |  436002.383         1  436002.383   Prob > F        =    0.0000
    Residual |  6770086.28    14,276  474.228515   R-squared       =    0.0605
-------------+----------------------------------   Adj R-squared   =    0.0604
       Total |  7206088.67    14,277  504.734095   Root MSE        =    21.777

------------------------------------------------------------------------------
          ph |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         inc |    2.61757   .0863272    30.32   0.000     2.448357    2.786782
       _cons |   62.17131   .4391525   141.57   0.000     61.31051    63.03211
------------------------------------------------------------------------------
```

  – Income coefficient is now 3.5 times larger!!!! Why??
  – t-stat for income coefficient = 30.32 $\Rightarrow$ even higher than before
  – The R-squared = 0.06 $\Rightarrow$ income explains about 6% of the variation in PH observed across individuals in this sample, which 2.5 times more than with the full sample. Why?

Topic 2:Univariate Regression and OLS

c) The Classic Linear Regression Model

THE UNIVERSITY OF WESTERN AUSTRALIA
Achieve International Excellence

- **Example 3:** What happens if I restrict the sample further?

```
. regress ph inc if inc<5

      Source |       SS           df       MS      Number of obs   =     8,751
-------------+----------------------------------   F(1, 8749)      =    506.14
       Model |  277249.223         1  277249.223   Prob > F        =    0.0000
    Residual |  4792436.25     8,749  547.769602   R-squared       =    0.0547
-------------+----------------------------------   Adj R-squared   =    0.0546
       Total |  5069685.47     8,750  579.392625   Root MSE        =    23.404

------------------------------------------------------------------------------
          ph |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         inc |    5.18068    .230277    22.50   0.000     4.729283    5.632077
       _cons |   54.09793   .7860433    68.82   0.000      52.5571    55.63876
------------------------------------------------------------------------------
```

  - The income coefficient almost doubles in size!!
  - What does this mean?
  - What can I conclude?

Topic 2:Univariate Regression and OLS

c) The Classic Linear Regression Model

THE UNIVERSITY OF WESTERN AUSTRALIA
Achieve International Excellence

- **Some really, really important things to bear in mind:**

CAUTION

Beware of the OLS small print!

- The reliability of OLS estimates rests on a set of assumptions – many of them concerning the error terms. If these are violated, then problems ensue. Different contexts are associated with different common problems. We'll talk more about this next week.

CAUTION

The slope parameter does not measure correlation!

- Correlation is a term which refers to the strength or degree of linear association between two variables. If all points lie ON the linear regression line, we have zero error and perfect (100%) correlation. The correlation coefficient is then 1. If X and Y are all over the place, with no systematic linear pattern, then we have 0% correlation (correlation coefficient = 0).
- The slope parameter or coefficient measures $\frac{\Delta Y}{\Delta X}$, which can take any value at all; its interpretation depends on the scales on which Y and X are measured.

CAUTION

Association is NOT the same as causation!

- The fact that Y is observed to increase along with X does not mean that X causes Y!

Topic 2:Univariate Regression and OLS

Guide to further reading

The University of Western Australia
Achieve International Excellence

- **Topic 2(i): Univariate Regression:** $Y = b_0 + b_1X + e$
  a) Introduce the classic univariate linear regression model
     - Woodridge Ch 4.1; Gujarati Ch 1
  b) Using univariate simple regression with single dummy to test for differences in group means [Continuous Y, single binary X]
     - Gujarati Ch 2.2 (at a pinch)
  c) Using OLS to estimating intercept and slope for a linear function [Continuous Y, single continuous X]
     - Gujarati Ch 3.1. Woodridge Ch 4.2 – 4.4