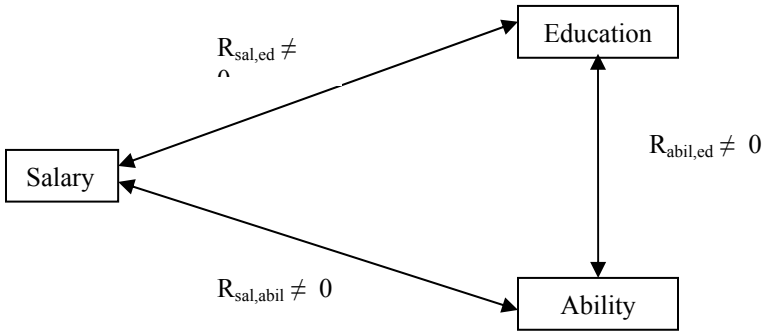
Rockefeller College
University at Albany

PAD 705 Handout: Omitted Variable Bias

Omitted variable bias (OVB) is one of the most common and vexing problems in ordinary least squares regression. OVB occurs when a variable that is correlated with both the dependent *and* one or more *included* independent variables is omitted from a regression equation. Let's think about salary and education; our regression equation is:

$$Salary_i = \beta_0 + \beta_1 education_i + \varepsilon_i$$

In this case, our included independent variable is "education." However, salary is also likely to be related to innate "ability", which has been excluded (possibly because there is no good way to measure it). Ability, we would expect, is also related to the amount of education a person chooses to get – those with greater ability seek more education. This creates what I term the "OVB Triangle":



For OVB to exist with respect to β_1 (the coefficient on education) *each* of the correlations (i.e., Rs) on the Triangle's legs must be non-zero. The amount of OVB depends on how large those correlations are – if any one of them is relatively small, then very little bias will be generated.

To understand why such regressions are biased, we must back-track to the Gauss-Markov Conditions. One of the least-realistic GM Conditions is the requirement that independent variables be non-stochastic – that they not be derived from a sample. In almost all social research, both the dependent and independent variables are measured through samples. If the independent variables are non-stochastic, there is no correlation between the error term and the independent variables, which greatly simplifies the math that proves that our estimate of β is unbiased. When the independents are stochastic, we can no longer be sure the estimate of β is unbiased. The reason this matters is that our derivation of expected value of β depended on the independents being non-stochastic (see the handout "Properties of OLS Estimators Under Hetereoskedascity"). When this assumption fails, the expected value contains a second term:

$$E[\hat{\beta}_1] = E[\beta_1 + \frac{\sum \varepsilon_i (X_i - \overline{X})}{\sum (X_i - \overline{X})^2}]$$

$$E[\hat{\beta}_1] = \beta_1 + E[\frac{Cov(X, \varepsilon)}{\sum (X_i - \overline{X})^2}]$$

The only way for our estimate to be unbiased is for cov(X,ε) to be equal to zero.

Why would leaving out a variable correlated to both the dependent and an included independent create covariance between the independent and error term? So far, we have thought of the error term as a random "add-on" to our regression equation – God dips into a basket and adds or subtracts something from our salary, level of education, etc. The basket contains chits; there are more chits with zero or small numbers on them than large numbers, though every number from positive to negative infinity is possible. The distribution of possible additions and subtractions follows the normal distribution. The error term captures life's inherently random changes to outcomes

However, the error term is also defined as the difference between the actual observations (Y_i) – what we actually earn – and the predicted line found via regression ($\hat{Y}_i$) – what our regression says we should earn. If we omit a variable that is correlated with both an included independent and the dependent variable, we will increase the size of the distance between Y_i and $\hat{Y}_i$ – our regressions will "fit" less well. The "amount" that the error term gets larger depends on the relationship between the omitted variable and both the included independent variable and the dependent variable. Now, recall that this omitted variable and included independent variable are correlated with one another – as the omitted variable gets bigger, the included independent gets bigger, or if the omitted variable gets smaller then the included independent variable gets smaller (assuming positive correlation). If that is the case, then the amount of "extra" distance between Y_i and $\hat{Y}_i$ created by leaving out the omitted variable will also be correlated with the included variable. Hence, there will be correlation between the included independent variable and the error term, creating bias.

The nature of the bias on the included independent variables depends on the nature of the correlation between (a) the dependent and the excluded variable and (b) the included and excluded independent variables. If Y is the dependent variable, X_1 is the included independent and X_2 is the excluded independent, then the coefficient on X_1, β_1, will be biased in the following manner if X_2 is excluded from the regression:

	Negative correlation, X_1 and X_2	Positive correlation, X_1 and X_2
Negative correlation, Y and X_2	β_1 is *overestimated*	β_1 is *underestimated*
Positive correlation, Y and X_2	β_1 is *underestimated*	β_1 is *overestimated*

In practical terms, the requirement that we include all variables that are correlated to both our independent variables and our dependent variable places a heavy burden on our data collection methods. If we wish to know about the relationship between salary and education, for instance, we must be sure to include all variables that could be correlated with both education and salary. In some cases, it may not be possible to

measure a variable that theory or intuition tells you should be included – like ability. To some extent, the requirement is impossible for humans that are boundedly rational: we simply cannot foresee all the possible variables that are correlated with both age and salary. Luckily, the degree of bias decreases as the degree of covariance between the omitted variable and the included independent variable decreases. We can often live with a little bias.

When we have data on multiple variables, we can see the manifestation of OVB by introducing variables one at a time and seeing how much the value of the coefficient changes. We can also use graphical methods to look for evidence of OVB. Below, we return to the salary example from `gender.dta`. Here, we begin with a regression of salary and age.

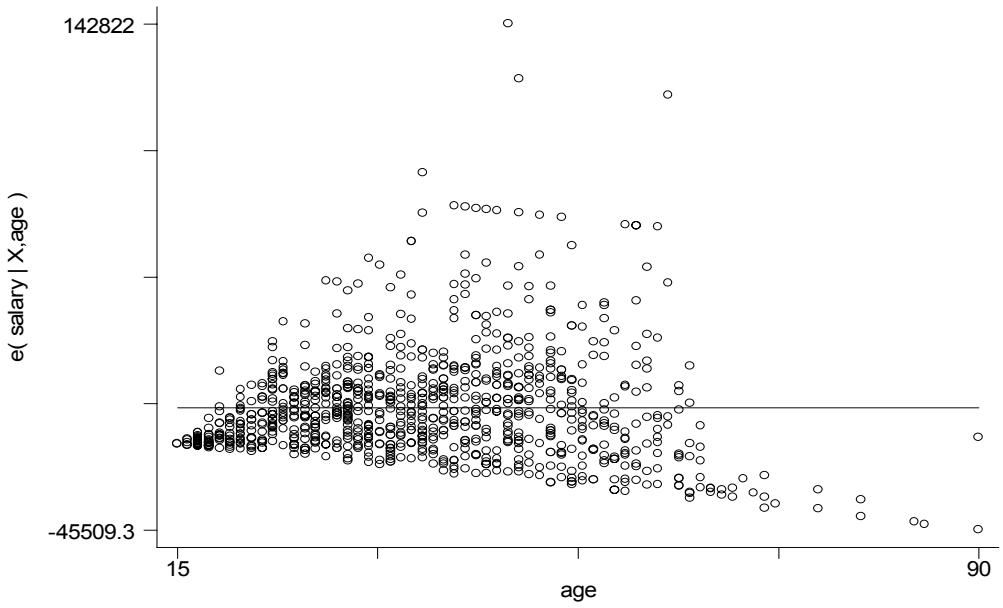
```
. reg salary age

      Source |       SS       df       MS              Number of obs =     950
-------------+------------------------------           F(  1,   948) =   73.52
       Model |  3.0697e+10     1  3.0697e+10           Prob > F      =  0.0000
    Residual |  3.9584e+11   948   417553901           R-squared     =  0.0720
-------------+------------------------------           Adj R-squared =  0.0710
       Total |  4.2654e+11   949   449460305           Root MSE      =   20434

------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    431.424   50.31698     8.57   0.000     332.6785    530.1695
       _cons |   7331.165   2035.786     3.60   0.000     3335.998    11326.33
------------------------------------------------------------------------------
```

To look for evidence of OVB, we will plot the residuals (i.e., $Y_i - \hat{Y}_i$) against age. If there seems to be a pattern of residuals becoming either smaller or larger as age increases or decreases, this would be *prima facie* evidence of OVB. Below, there seems to be evidence of just that.

```
. rvpplot age, yline(0)
```



Now, let's see if there are variables that are correlated to both salary and age:

```
. corr
(obs=950)

             |      sex      age   salary    hours    weeks     educ
-------------+------------------------------------------------------
         sex |   1.0000
         age |   0.0088   1.0000
      salary |  -0.2529   0.2683   1.0000
       hours |  -0.2259   0.0861   0.4234   1.0000
       weeks |  -0.0155   0.2301   0.4148   0.3661   1.0000
        educ |   0.0861   0.0995   0.4143   0.2090   0.1931   1.0000
```

At least one candidate here for OVB would be leaving out education – it has a small amount of correlation with age and a greater correlation to salary.

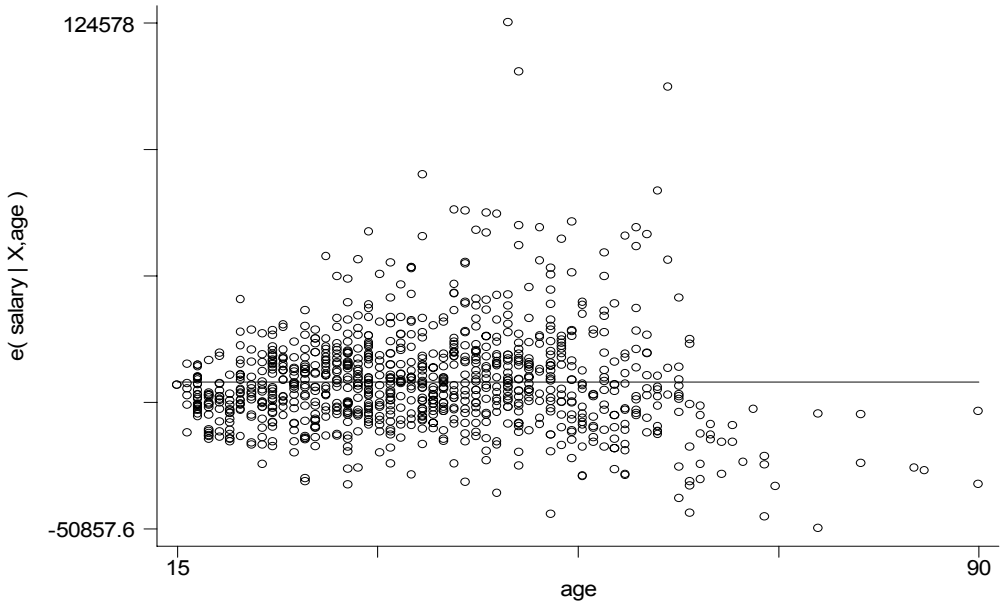
We will now add in several variables that could be correlated with age. Note how much age changes as a result – by about $63 per additional year of age. A change of more than 3 to 5% is, informally speaking, indicative of OVB. However, our residuals-to-age plot still seems to have a pattern.

```
. reg salary age sex educ

      Source |       SS       df       MS              Number of obs =     950
-------------+------------------------------           F(  3,   946) =  140.12
       Model |  1.3122e+11     3  4.3741e+10           Prob > F      =  0.0000
    Residual |  2.9531e+11   946   312171930           R-squared     =  0.3076
-------------+------------------------------           Adj R-squared =  0.3055
       Total |  4.2654e+11   949   449460305           Root MSE      =   17668

------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   368.8599   43.72361     8.44   0.000     283.0534    454.6663
         sex |  -12326.02   1151.183   -10.71   0.000    -14585.18   -10066.85
        educ |   3259.901   213.5539    15.27   0.000     2840.807    3678.995
       _cons |  -114982.6   8549.233   -13.45   0.000    -131760.3   -98204.97
------------------------------------------------------------------------------
```

```
. rvpplot age, yline(0)
```



A reasonable supposition would be that we are still missing a key variable. Mis-specification can be a form of OVB – given the "up and down" pattern of residuals, it may be that we need to have a quadratic term to account for the decelerating increase in salary with age. Below is the regression and residual-to-age plot – it again seems to have some pattern.

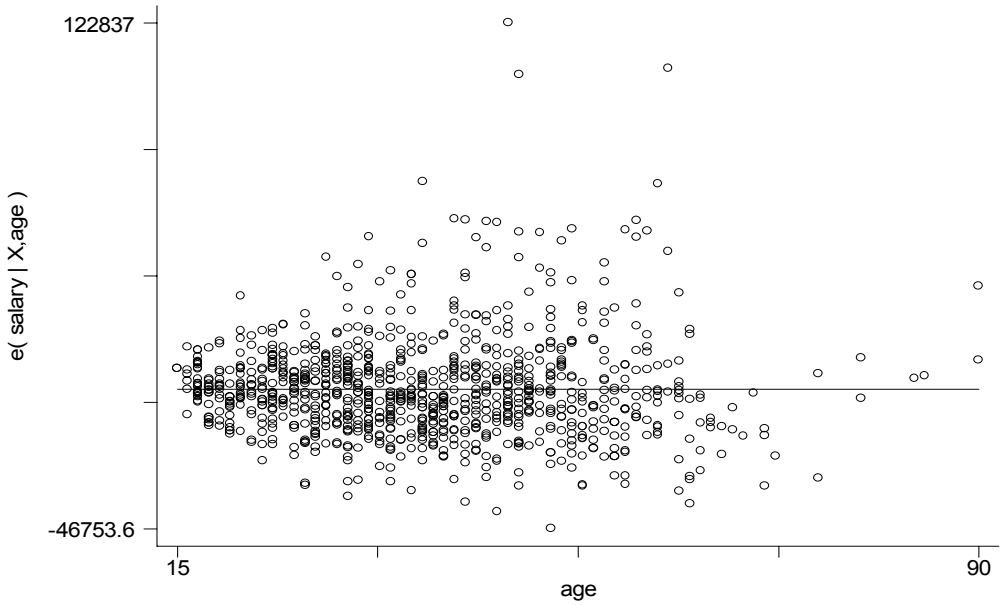
```
. gen age2=age^2

. reg salary age age2 sex educ

      Source |       SS       df       MS              Number of obs =     950
-------------+------------------------------           F(  4,   945) =  130.91
       Model |  1.5208e+11     4  3.8020e+10           Prob > F      =  0.0000
    Residual |  2.7446e+11   945   290430140           R-squared     =  0.3565
-------------+------------------------------           Adj R-squared =  0.3538
       Total |  4.2654e+11   949   449460305           Root MSE      =   17042

------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   2109.888   209.7258    10.06   0.000     1698.306     2521.47
        age2 |  -20.95795   2.473042    -8.47   0.000    -25.81123   -16.10466
         sex |  -12424.52   1110.432   -11.19   0.000    -14603.72   -10245.32
        educ |   2811.098    212.682    13.22   0.000     2393.714    3228.482
       _cons |  -129234.2   8415.879   -15.36   0.000    -145750.2   -112718.2
------------------------------------------------------------------------------
```

```
. rvpplot age, yline(0)
```



Stata also includes a command that tests for omitted variables – `ovtest`. This command is run post-regression and tests the hypothesis that the model has no omitted variables.

```
. ovtest

Ramsey RESET test using powers of the fitted values of salary
       Ho:  model has no omitted variables
                 F(3, 942) =     28.18
                  Prob > F =      0.0000
```

Despite our earlier inclusion of variables, `ovtest` still found OVB. We no longer have any additional variables to leverage in this model. We might improve the model by changing the specification, but we may have run into the limits of our data. There may be some other variable that should be included in the regression that is lacking in our dataset (for instance, experience). The only solution is to seek a richer dataset.