Economics 6400: Econometrics

Lecture 3: Simple Regression Model

CSU, East Bay

October 10, 2017

Last week...

- We derived least squares estimates for β_0 and β_1:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i(y_i - \overline{y})}{\sum_{i=1}^n x_i(x_i - \overline{x})} = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2}$$
$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

- These are derived using the following assumptions:
  1. $E(u) = 0 \Rightarrow E(y - \beta_0 - \beta_1 x) = 0$
  2. $Cov(x, u) = E(xu) = 0 \Rightarrow E[x(y - \beta_0 - \beta_1 x)] = 0$
- Equivalent estimates are obtained if we minimize the sum of squared residuals:

$$\sum_{i=1}^n \hat{u}_i^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

How well does the independent variable explain the dependent variable?

- To motivate a goodness of fit measure, we need a few definitions:
  - Total sum of squares (SST) is:

$$SST \equiv \sum_{i=1}^{n} (y_i - \overline{y})^2$$

  - The explained sum of squares (SSE) is:

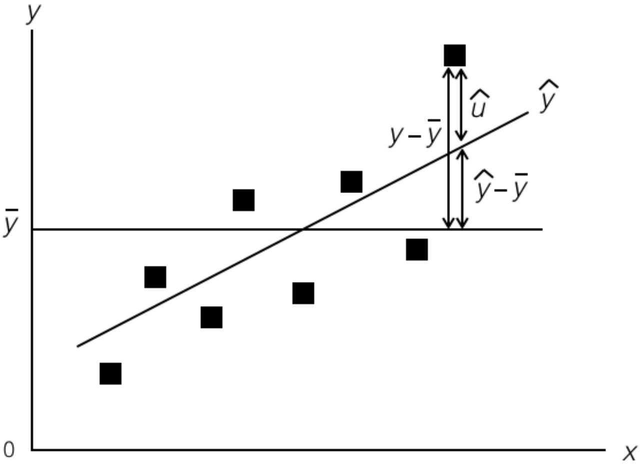
$$SSE \equiv \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$$

  - The residual sum of squares (SSR) is:

$$SSR \equiv \sum_{i=1}^{n} \hat{u}_i^2$$

- SST = SSE + SSR

SST vs. SSE vs. SSR



Last week...

- Total sum of squares (SST) is:

$$SST = SST_y \equiv \sum_{i=1}^{n}(y_i - \overline{y})^2$$

- The residual sum of squares (SSR) is:

$$SSR \equiv \sum_{i=1}^{n} \hat{u}_i^2$$

- The **R-squared** or coefficient of determination is

$$R^2 \equiv \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

- R^2 is the ratio of the explained variation compared to the total variation in y

Statistical properties of the OLS estimators

- An important property of an estimator, W of θ, is that its expected value equals the population value:

$$E(W) = \theta,$$

  which is called unbiasedness.
- Does not mean that the estimate equals the population parameter!
    - Only means that if you were to draw infinite samples from a population and compute an estimate, then the average of these estimates would be equal to the population parameter

Establishing unbiasedness

- To establish unbiasedness, i.e. $E(\hat{\beta}_1) = \beta_1$ and $E(\hat{\beta}_0) = \beta_0$, we require four assumptions:
    1. Linearity: $y = \beta_0 + \beta_1 x + u$
    2. We have a **random** sample of size n, $\{(x_i, y_i) : i = 1, 2, \ldots, n\}$
    3. Sample variation of explanatory variable: The sample outcomes on x are not all the same value
    4. Zero conditional mean: $E(u|x) = 0$

Establishing unbiasedness of β_1

$$
\begin{aligned}
\hat{\beta}_1 =& \frac{\sum_{i=1}^{n}(x_i - \overline{x})y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \\
=& \frac{\sum_{i=1}^{n}(x_i - \overline{x})(\beta_0 + \beta_1 x_i + u_i)}{SST_x} \quad \text{if \# 3 fails then this} = \infty \\
=& \frac{1}{SST_x}\Big(\sum_{i=1}^{n}\beta_0(x_i - \overline{x}) + \sum_{i=1}^{n}\beta_1(x_i - \overline{x})x_i + \sum_{i=1}^{n}(x_i - \overline{x})u_i\Big) \\
=& \frac{1}{SST_x}\Big(\beta_0\sum_{i=1}^{n}(x_i - \overline{x}) + \beta_1\sum_{i=1}^{n}(x_i - \overline{x})x_i + \sum_{i=1}^{n}(x_i - \overline{x})u_i\Big) \\
=& \frac{1}{SST_x}\Big(\beta_0 \cdot 0 + \beta_1 SST_x + \sum_{i=1}^{n}(x_i - \overline{x})u_i\Big) \\
=& \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \overline{x})u_i}{SST_x}
\end{aligned}
$$

Establishing unbiasedness of β_1

- "Naive proof" where we assume the values of x_i are determined (known) in advance, e.g. in an experiment:

$$
\begin{aligned}
E(\hat{\beta}_1) =& E(\beta_1) + E\Big(\frac{\sum_{i=1}^{n}(x_i - \overline{x})u_i}{SST_x}\Big) \\
=& \beta_1 + \frac{1}{SST_x}\sum_{i=1}^{n} E[(x_i - \overline{x})u_i] \\
=& \beta_1 + \frac{1}{SST_x}\sum_{i=1}^{n}(x_i - \overline{x})E(u_i) \quad \text{since we know } (x_i - \overline{x}) \\
=& \beta_1 + \frac{1}{SST_x}\sum_{i=1}^{n}(x_i - \overline{x}) \cdot 0 \quad \text{since } E(u_i|x) = 0 \\
=& \beta_1
\end{aligned}
$$

Establishing unbiasedness of β_0

- Should be straight-forward:

$$\begin{aligned} E(\hat{\beta}_0) =& \overline{y} - \hat{\beta}_1 \overline{x} \\ =& \beta_0 + \beta_1 \overline{x} + \overline{u} - \hat{\beta}_1 \overline{x} \\ =& \beta_0 + (\beta_1 - \hat{\beta}_1)\overline{x} + \overline{u} \end{aligned}$$

- Once again, conditioning on the values of our sample, x_i:

$$\begin{aligned} E(\hat{\beta}_0) =& E(\beta_0) + E[(\beta_1 - \hat{\beta}_1)\overline{x}] + E(\overline{u}) \\ =& \beta_0 + E[(\beta_1 - \hat{\beta}_1)]\overline{x} \quad \text{since } E(u_i|x) = 0 \\ =& \beta_0 \quad \text{since } E(\hat{\beta}_1) = \beta_1 \end{aligned}$$

Unbiasedness

- Unbiasedness will fail if any of the four assumptions fail
  1. Linearity can fail easily but we can still include non-linear relationships
     - Can include non-linear relationships using logarithms
  2. Random sampling can fail due to selection issues
     - Oversampling wealthy people, men etc.
  3. Variation in explanatory variables unlikely to fail
  4. Zero conditional mean assumption **always a concern**

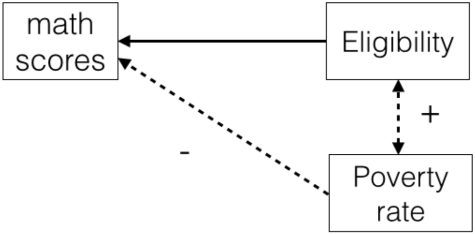
Example where zero conditional mean assumption fails

- Effect of a federally funded school lunch program on student performance
  - *math10* denotes the math performance at a high school receiving a passing score in a standardized math exam
  - *lnchprg* is the percentage of students eligible for a lunch program
- Stata estimates of the model $math10 = \beta_0 + \beta_1 lnchprgr + u$:

math10	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnchprg	-.3188643	.0348393	-9.15	0.000	-.3873523	-.2503763
_cons	32.14271	.9975824	32.22	0.000	30.18164	34.10378

Example where zero conditional mean assumption fails

- Coefficient on *lnchprg* implies a 10 percent rise in the number of students eligible for the program is related to a 3.2 percent decline in the percentage of students that pass the exam
  - This cannot be causal! Known as a "spurious" relationship
  - Eligibility is correlated with other factors such as the poverty rate of children at the score, which is negatively correlated with test performance
  - For school i, $E(u_i | eligibility_i) \neq 0$



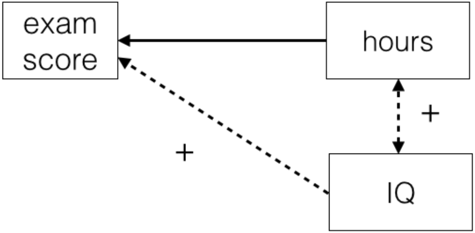
Example from Homework 1 Q1

- Does studying for longer hours improve exam performance?
- Coefficient on *hours*, β_1, in the model:

$$exam\ score = \beta_0 + \beta_1 hours + u$$

is likely to be positive

- More hours means better understanding of the material
- But smarter students may spend more time studying cause they like studying (and they do well on exams regardless)!
- For student i, $E(u_i|hours) \neq 0$ if assignment not random



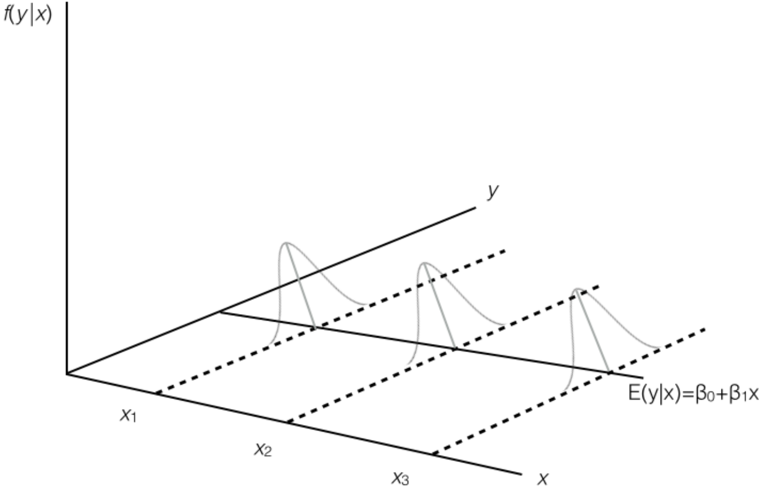
Homoskedasticity

- The variance of estimator ($\hat{\beta}_1$) is also important since it tells us how far from the population parameter (β_1) is likely to be
- An additional assumption will simplify the calculation of $\hat{\beta}_1$'s variance is:
  5. The error u has the same variance given any value of the explanatory variable:

$$\mathrm{Var}(u|x) = \sigma^2$$

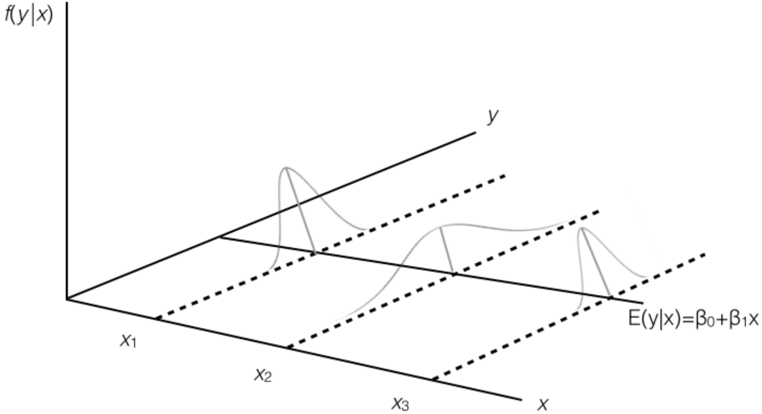
- Since $\text{Var}(u|x) = E(u^2|x) - [E(u|x)]^2$ and $E(u|x) = 0$, $\sigma^2 = E(u^2|x) = E(u^2) = \text{Var}(u)$
  - σ^2 is the unconditional variance of u. A larger σ means that the distribution of the unobservables affecting y is more spread out

Simple regression model under homoskedasticity



Simple regressions model under heteroskedasticity

- Under heteroskedasticity, Var($u|x$)=Var($y|x$) depends on x.



Deriving the variance of $\hat{\beta}_1$

- Given homoskedasticity, we can derive the variance of $\hat{\beta}_1$:

$$
\begin{aligned}
\mathrm{Var}(\hat{\beta}_1) =& \mathrm{Var}\Big(\beta_1 + \frac{\sum_{i=1}^n (x_i - \overline{x}) u_i}{SST_x}\Big) \\
=& \mathrm{Var}\Big(\frac{\sum_{i=1}^n (x_i - \overline{x}) u_i}{SST_x}\Big) \quad \text{since } \mathrm{Var}(\beta_1) = 0 \\
=& \frac{1}{(SST_x)^2} \sum_{i=1}^n (x_i - \overline{x})^2 \mathrm{Var}(u_i) \\
=& \frac{SST_x \sigma^2}{(SST_x)^2} \\
=& \frac{\sigma^2}{SST_x} = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2}
\end{aligned}
$$

Analyzing the formula for $\hat{\beta}_1$'s variance

- $\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{SST_x} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$
- The larger the error variance (σ^2), the *larger* is Var($\hat{\beta}_1$).
  - The larger the variation in unobservables affecting y, the harder it is to isolate the impact of x and precisely estimate $\hat{\beta}_1$.
- The larger the variance in x (SST_x), the *smaller* is Var($\hat{\beta}_1$).
  - The more "spread out" x is, the easier it is to decipher its relationship with E($y|x$), e.g. if we wanted to know the impact of class size on performance, we would ideally randomly assign students to a variety of class sizes

If σ^2 is unknown ...

- $\sigma^2 = \mathrm{E}(u^2)$, so an unbiased estimator of σ^2 is $\frac{1}{n}\sum_{i=1}^{n} u_i^2$.
  - But we do not observe u_i!
- If we replace u_i with $\hat{u}_i$ then we obtain the **biased** estimator $\frac{1}{n}\sum_{i=1}^{n} \hat{u}_i^2$
- **Unbiased** estimator is $\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} \hat{u}_i^2$
  - $\frac{1}{n-2}$ accounts for two restrictions that we imposed to obtain the residuals:

$$\sum_{i=1}^{n} \hat{u}_i = 0, \quad \sum_{i=1}^{n} x_i \hat{u}_i = 0$$

  - Example: suppose we had 3 observations then we only need to know the residual of one observation, $\hat{u}_1 = 3$, to recover the remaining two residuals ($\Rightarrow$ we have one observation only)
  - Can easily solve these two equations for the two unknowns: $3 + \hat{u}_2 + \hat{u}_3 = 0$, and $3x_1 + \hat{u}_2 x_2 + \hat{u}_3 x_3 = 0$

Standard errors

- Standard error of the regression (SER) is $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$
  - Estimate of the standard deviation in y after the effect of x has been removed
- Standard error of $\hat{\beta}_1$ is given by:

$$\mathrm{se}(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{SST_x}} = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

  - Just like standard errors were crucial for conducting hypothesis tests regarding sample and population means, they will be crucial for conducting tests about least squares estimates
  - <u>Quick preview</u>: t statistic for testing whether coefficient is zero under the null hypothesis will be: $t = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$.

Regression through the origin

- Suppose you want to impose the restriction that $\beta_0 = 0$
  - For example, tax revenues (y) will be zero if income (x) is zero
- In such cases, we will be estimating $\tilde{\beta}_1 x$ in the equation:

$$\tilde{y} = \tilde{\beta}_1 x$$

- Least squares will minimize the sum of squared residuals:

$$\sum_{i=1}^{n} (y - \tilde{\beta}_1 x_i)^2$$

which yields solution $\tilde{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$ c.f. $\frac{\sum_{i=1}^{n} y_i (x_i - \overline{x})}{\sum_{i=1}^{n} x_i (x_i - \overline{x})} = \hat{\beta}_1$

  - Two estimates are only the same if $\overline{x} = 0$

Multiple Regression Analysis

- Linear regression model explains y in terms of variables $x_1, x_2, \ldots, x_k$:

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

- By incorporating additional right-hand side variables $x_2 \ldots, x_k$, we can control for factors that:
  1. were previously part of u, and
  2. were likely to be correlated with x_1
- As before, key assumption is that

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

which implies that the independent variables are not correlated with the error term

Multiple regression analysis examples

1. Do smaller class sizes improve student performance?

$$math10 = \beta_0 + \beta_1 class\ size + \beta_2 family\ \ income + u$$

- Effect of class size is measured explicitly holding family wealth fixed

2. Impact of education on wage?

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + u$$

- Effect of education is measured explicitly holding experience fixed

Obtaining the OLS estimates using method of least squares

- Estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ minimize sum of squared residuals:

$$\sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik})^2$$

which leads to $k+1$ first-order conditions/equations with $k+1$ unknowns:

$$\begin{aligned}
\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}) &= 0 \\
\sum_{i=1}^{n} x_{i1}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}) &= 0 \\
&\ldots \\
\sum_{i=1}^{n} x_{ik}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}) &= 0
\end{aligned}$$

Properties of any sample of data

- Fitted values and residuals:

$$\begin{aligned}\hat{y}_i =& \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_k x_{ik} \\ \hat{u}_i =& y_i - \hat{y}_i\end{aligned}$$

- Algebraic properties:
  1. $\sum_{i=1}^{n} \hat{u}_i = 0 \Rightarrow \overline{y} = \overline{\hat{y}}$
  2. $\sum_{i=1}^{n} x_{ij} \hat{u}_i = 0$
  3. $\overline{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \overline{x}_1 + \ldots + \hat{\beta}_k \overline{x}_k$
- The second property means that <u>each</u> independent variable has zero covariance with $\hat{u}_i$

Goodness-of-Fit

- Measures of variation are equivalent to the definitions under the simple regression model:

$$SST \equiv \sum_{i=1}^{n}(y_i - \overline{y})^2,\ SSE \equiv \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2,\ SSR \equiv \sum_{i=1}^{n} \hat{u}_i^2$$

- The coefficient of determination is also equivalent:

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

- R^2 never decreases as you add more right-hand variables
  - If they have no explanatory power, their coefficients can/will be zero

Interpreting the OLS regression equation

- Consider the case where $k = 2$:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

- The intercept $\hat{\beta}_0$ is the predicted value of y when $x_1 = 0$ and $x_2 = 0$.
- Estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ have a **partial effect** interpretation:

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x_1 + \hat{\beta}_2 \Delta x_2$$

so if x_2 is held fixed ($\Delta x_2 = 0$) then

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x_1$$

Meaning of “holding other factors fixed”

- Variation in x_1 and x_2 allows us to estimate the ceteris paribus effect of each even if the data we collect does not explicitly include observations in which x_2 is the same but x_1 varies
    - Non-experimental data rarely have this feature
- If we could collect a sample of individuals with the same values of $x_2, \ldots, x_k$ then we could perform a simple regression of y on x_1 to obtain the relationship between y and x_1
    - Multiple regression analysis allows us to “mimic” this ideal world
    - This power explains why it is the most widely used approach for empirical analysis in economics and other social sciences

“Partialling out” interpretation of multiple regression

- The estimated coefficient of a right-hand side variable in multiple regression can be obtained in two steps:
  1. Regress the right-hand side variable on all the *other* right-hand side variables
  2. Regress y on the residuals from this regression
- Why does this procedure work?
  - Residuals from the first regression are the part of the explanatory variable that is uncorrelated with the other explanatory variables
  - The slope coefficient of the second regression therefore represents the isolated effect of the explanatory variable on the left-hand side variable

Effects of smoking during pregnancy on infant health

- Consider the model to be estimated

$$bwght = \beta_0 + \beta_1 cigs + \beta_2 faminc + u$$

  where

  - *bwght* is birth weight in ounces
  - *cigs* is cigarettes smoked per day while pregnant
  - *faminc* is 1988 family income in \$1000s

- β_2 is likely > 0 since wealthier families have better access to prenatal care

Effects of smoking during pregnancy on infant health (BWGHT.dta)

```
. reg bwght cigs faminc

      Source |       SS       df       MS              Number of obs =    1388
-------------+------------------------------           F(  2,  1385) =   21.27
       Model |  17126.2088     2  8563.10442           Prob > F      =  0.0000
    Residual |  557485.511  1385  402.516614           R-squared     =  0.0298
-------------+------------------------------           Adj R-squared =  0.0284
       Total |   574611.72  1387  414.283864           Root MSE      =  20.063

------------------------------------------------------------------------------
       bwght |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cigs |  -.4634075   .0915768    -5.06   0.000    -.6430518   -.2837633
      faminc |   .0927647   .0291879     3.18   0.002     .0355075    .1500219
       _cons |   116.9741   1.048984   111.51   0.000     114.9164    119.0319
------------------------------------------------------------------------------
```

- $\widehat{bwght} = 116.974 - 0.463 cigs + 0.093 faminc$
- Every cigarette lowers birthweight by 0.463 ounces

Demonstrating the "partialling out" effect of *cigs*

1. Regress *cigs* on *faminc*:

```
. reg cigs faminc

      Source |       SS       df       MS              Number of obs =    1388
-------------+------------------------------           F(  1,  1386) =   42.78
       Model |  1481.60979     1  1481.60979           Prob > F      =  0.0000
    Residual |  47996.8419  1386  34.6297561           R-squared     =  0.0299
-------------+------------------------------           Adj R-squared =  0.0292
       Total |  49478.4517  1387  35.6730005           Root MSE      =  5.8847

------------------------------------------------------------------------------
        cigs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      faminc |  -.0551538   .0084321    -6.54   0.000    -.0716948   -.0386129
       _cons |   3.688107   .2912973    12.66   0.000     3.116676    4.259538
------------------------------------------------------------------------------
```

2. Obtain residuals $\hat{u}$ using command "predict residuals, r"

Demonstrating the "partialling out" effect of *cigs*

3. Regress *bwght* on *residuals*:

```
. reg bwght residuals

      Source |       SS       df       MS              Number of obs =    1388
-------------+------------------------------           F(  1,  1386) =   25.32
       Model |  10307.1563     1  10307.1563           Prob > F      =  0.0000
    Residual |  564304.563  1386   407.14615           R-squared     =  0.0179
-------------+------------------------------           Adj R-squared =  0.0172
       Total |   574611.72  1387  414.283864           Root MSE      =  20.178

------------------------------------------------------------------------------
       bwght |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   residuals |  -.4634075    .092102    -5.03   0.000    -.6440818   -.2827333
       _cons |   118.6996   .5416022   219.16   0.000     117.6371     119.762
------------------------------------------------------------------------------
```

- Same coefficient on *cigs* from full regression and *residuals* in this regression (MAGIC!!!!!)

Expected value of the OLS estimators

- Standard assumptions for the multiple regression model
  1. Linear in parameters: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$
  2. Random sampling
  3. No perfect collinearity: none of the independent variables is constant and there are no *perfect* linear relationships among the independent variables, e.g. if one variable is a constant multiple of another
  4. Zero conditional mean: $E(u|x_1, x_2, \ldots, x_k) = 0$
- The above assumptions $\Rightarrow E(\hat{\beta}_j) = \beta_j,\ j = 0, 1, \ldots, k$

Violation of zero conditional mean assumption

- When might Assumption 4 fail:
  - Mis-specifying functional form, e.g. including *inc* but omitting inc^2
  - Omitting an important variable that is correlated with the included variables $x_1, x_2, ..., x_k$ violates this assumption
  - Failing to account for joint determination of y and an explanatory variable, e.g. p and q are jointly determined by the intersection of demand and supply so a regression of p on q will lead to biased estimates
- If Assumption 4 holds then we say the right-hand side variables are **exogenous**
- If x_j is correlated with u then we say x_j is **endogenous**

Omitted variable bias

- Consider a population model with two right-hand side variables x_1 and x_2 that are correlated with linear relationship:

$$x_2 = \delta_0 + \delta_1 x_1 + v$$

- Inserting the equation for x_2 into the "true model" yields:

$$\begin{aligned} y =& \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \\ =& \beta_0 + \beta_1 x_1 + \beta_2(\delta_0 + \delta_1 x_1 + v) + u \\ =& (\beta_0 + \beta_2 \delta_0) + (\beta_1 + \beta_2 \delta_1) x_1 + (\beta_2 v + u) \end{aligned}$$

- If y is only regressed on x_1 (i.e. x_2 omitted) then:
  - First term is the estimated intercept: $\tilde{\beta}_0 = \hat{\beta}_0 + \hat{\beta}_2 \delta_0$
  - Second term is the estimated slope coefficient on x_1, and $\hat{\beta}_2 \delta_1$ is the **omitted variable bias**: $\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \delta_1$
  - Third term is the error term
- All estimated coefficients will be biased!

When are you not in trouble?

- Key relationships:
  - True model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$
  - Relationship between x_1 and x_2: $x_2 = \delta_0 + \delta_1 x_1 + v$
  - Incorrect and correct estimates of β_1: $\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \delta_1$

1. If $\beta_2 = 0$ then x_2 is not part of u so $\tilde{\beta}_1 = \hat{\beta}_1$
2. If $\delta_1 = 0$ then x_1 & x_2 are uncorrelated so $\tilde{\beta}_1 = \hat{\beta}_1$

Examples where omitting x_2 is not a problem

1. x_2 is not part of u (i.e. $\beta_2 = 0$):
   - Probability of lung cancer (y) is related to cigarette consumption (x_1) and cigarette consumption is correlated with alcohol consumption (x_2) but lung cancer and alcoholic consumption are probably not causally related ($\beta_2 = 0$)
   - Home price (y) is related to the time it takes to commute to the city (x_1) and commuting time is (strongly) correlated with distance to the city (x_2) but $\beta_2 = 0$ if x_1 is already included in the regression (i.e. home owners care about commuting time, not distance)
2. x_1 & x_2 are uncorrelated (i.e. $\delta_1 = 0$):
   - Airfares (y) are related to concentration (competition) on a route (x_1) and whether the ticket is a first or economy class ticket (x_2) but x_1 and x_2 are not obviously correlated

Omitted variable bias example

- Consider the wage equation

$$wage = \beta_0 + \beta_1 educ + \beta_2 abil + u$$

$$abil = \delta_0 + \delta_1 educ + v$$

- If *abil* omitted from regression then:

$$wage = (\beta_0 + \beta_2\delta_0) + (\beta_1 + \beta_2\delta_1)educ + (\beta_2 v + u)$$

- The return to education β_1 will **upward biased** ($\tilde{\beta}_1 > \hat{\beta}_1$) because $\beta_2\delta_1 > 0$. It will appear that people with many years of education earn very high wages, but this is partly due to the fact that people with more education tend to have greater ability on average

Omitted variable bias example

- For simple case above, bias is $\beta_2 \delta_1$
- Tracking the direction of bias:

	$\delta_1 > 0$ Corr$(x_1, x_2) > 0$	$\delta_1 < 0$ Corr$(x_1, x_2) < 0$
$\beta_2 > 0$	Positive bias	Negative bias
$\beta_2 < 0$	Negative bias	Positive bias

Omitted variable bias and birth weight example

- Consider the birth weight equation from above:

$$bwght = \beta_0 + \beta_1 cigs + \beta_2 faminc + u$$

$$cigs = \delta_0 + \delta_1 faminc + v$$

- If *faminc* is omitted from regression then:

$$bwght = (\beta_0 + \beta_2\delta_0) + (\beta_1 + \beta_2\delta_1)cigs + (\beta_2 v + u)$$

- β_2 likely > 0 and $\text{Corr}(cigs, faminc) < 0$ so we would expect $\tilde{\beta}_1$ to be **downward biased** (i.e. $\tilde{\beta}_1 < \hat{\beta}_1$)
- People who consume a lot of cigarettes tend to be less wealthy and have less access to prenatal care so birth weights are more negatively correlated with cigarette consumption

Omitted variable bias and birth weight example

```
. reg bwght cigs

      Source |       SS       df       MS              Number of obs =    1388
-------------+------------------------------           F(  1,  1386) =   32.24
       Model |  13060.4194     1  13060.4194           Prob > F      =  0.0000
    Residual |    561551.3  1386  405.159668           R-squared     =  0.0227
-------------+------------------------------           Adj R-squared =  0.0220
       Total |   574611.72  1387  414.283864           Root MSE      =  20.129

------------------------------------------------------------------------------
       bwght |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cigs |  -.5137721   .0904909    -5.68   0.000    -.6912861   -.3362581
       _cons |   119.7719   .5723407   209.27   0.000     118.6492    120.8946
------------------------------------------------------------------------------
```

- $\tilde{\beta}_1 = -0.514 < \hat{\beta}_1 = -0.463$

Omitted variable bias and birth weight example

- If *faminc* omitted from regression then:

$$bwght = (\beta_0 + \beta_2\delta_0) + (\beta_1 + \beta_2\delta_1)cigs + (\beta_2 v + u)$$

- The coefficient on *cigs*, $\tilde{\beta}_1$, will be equal to:

$$\begin{aligned}\hat{\tilde{\beta}}_1 =& \hat{\beta}_1 + \hat{\beta}_2\delta_1 \\ =& -0.463 + 0.093 \times -0.543 \\ =& -0.514\end{aligned}$$

where δ_1 is obtained from a regression of *faminc* on *cigs*

Next week

- Multiple regression analysis continued (Chapter 3)