Economics 6400: Econometrics

Lecture 2: Simple Regression Model

CSU, East Bay

October 3, 2017

Simple regression model

- Explains variable y in terms of variable x

$$y = \beta_0 + \beta_1 x + u$$

  where

  - y is the **dependent variable**, explained variable, response variable, predicted variable, regressand
  - x is the **independent variable**, explanatory variable, control variable, predictor variable, regressor
  - u is the **error** or disturbance term, and represents factors other than x that affect y

- $\beta_0 + \beta_1 x$ is the systematic part of y, u is the unsystematic part

Interpreting the simple linear regression model

- Explains variable y in terms of variable x

$$\frac{dy}{dx} = \beta_1$$

as long as

$$\frac{du}{dx} = 0$$

  - $\frac{dy}{dx} = \beta_1$ represents how much the dependent variable changes if the independent variable increases by one unit
  - Only correct interpretation if $\frac{du}{dx} = 0$ such that all other things remain equal when x increases by one unit

Example regression models

1. Soybean yield and fertilizer

$$yield = \beta_0 + \beta_1 fertilizer + u$$

   where

   - β_1 measures the effect of fertilizer on yield, ceteris paribus
   - u might include rainfall, land quality, presence of parasites

2. A simple wage equation

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

   where

   - β_1 measures the change in hourly wage from another year of education, ceteris paribus
   - u might include experience, tenure with current employer, IQ

When can we make a causal interpretation?

- Conditional mean independence assumption

$$E(u|x) = E(u)$$

- If knowledge of x gives us *any* information about u then the condition is violated

- Example: wage equation

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

  - Conditional mean independence unlikely to hold because individuals with more education will be more intelligent on average, e.g. if you knew someone had 16 years of education then you would guess their IQ (contained in u) is probably higher than the IQ of someone with only 8 years of schooling $\Rightarrow E(u|educ = 16) \neq E(u|educ = 8)$

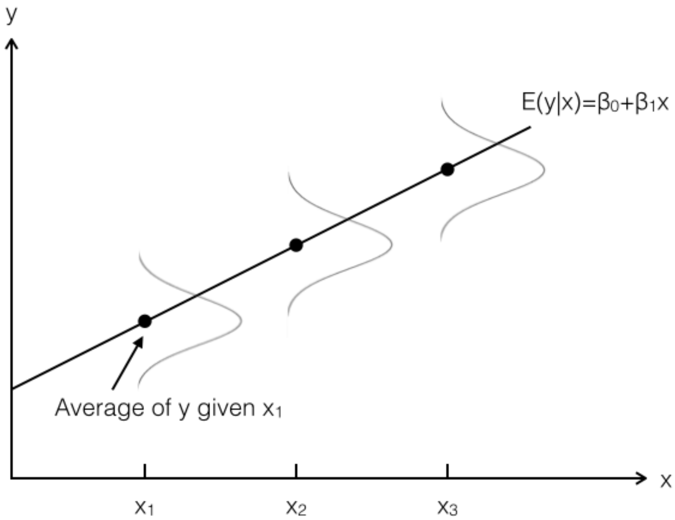
Population regression function

- If $E(u) = 0$ then conditional mean independence assumption implies that

$$\begin{aligned} E(y|x) =& E(\beta_0 + \beta_1 x + u|x) \\ =& \beta_0 + \beta_1 x + E(u|x) \\ =& \beta_0 + \beta_1 x + E(u) \\ =& \beta_0 + \beta_1 x \end{aligned}$$

- **Average** value of the dependent variable can be expressed as a linear function of the explanatory variable
  - Does **not** mean that y equals $\beta_0 + \beta_1 x$ for all units in the population

Population regression function



Estimating β_0 and β_1

- To estimate the parameters β_0 and β_1 we need a random sample of x and y from the population of size n $\{(x_i, y_i) : i = 1, \dots, n\}$
- Econometric model can be written as

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

  where i denotes one of the n observations.
- Example: y_i is housing expenditure of individual i and x_i is individual i's income

Quick review of covariance

- For two random variables X and Y:
  $Cov(X, Y) \equiv E[(X - \mu_X)(Y - \mu_Y)] = E(XY) - \mu_X \mu_Y$
  - The sample analog is: $S_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})$

- Therefore, if either $E(X) = 0$ or $E(Y) = 0$, then $Cov(X, Y) = E(XY)$.
  - Implication: $Cov(x, u) = E(xu)$ since $E(u) = 0$.

Deriving least squares estimates

- We can motivate the derivation of our estimates using two assumptions:
  1. $E(u) = 0 \Rightarrow E(y - \beta_0 - \beta_1 x) = 0$
     - This is not a restrictive assumption since the intercept can always be respecified to make $E(u) = 0$
  2. $Cov(x, u) = E(xu) = 0 \Rightarrow E[x(y - \beta_0 - \beta_1 x)] = 0$
     - Crucially, we assume x and u are not correlated

- Given a sample of data, we choose estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ to solve the sample analogs:
  1. $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$
  2. $\frac{1}{n}\sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$

Deriving least squares estimates: $\hat{\beta}_0$

- Simplifying the first assumption to obtain $\hat{\beta}_0$:

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) =& 0 \\
\Rightarrow \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{1}{n}\sum_{i=1}^{n}\hat{\beta}_0 - \frac{1}{n}\sum_{i=1}^{n}\hat{\beta}_1 x_i =& 0 \\
\Rightarrow \overline{y} - \hat{\beta}_0 - \hat{\beta}_1 \overline{x} =& 0 \\
\Rightarrow \hat{\beta}_0 =& \overline{y} - \hat{\beta}_1 \overline{x}
\end{aligned}
$$

Deriving least squares estimates: $\hat{\beta}_1$

- Simplifying the second assumption to obtain $\hat{\beta}_1$:

$$
\begin{aligned}
\sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) =& 0 \\
\Rightarrow \sum_{i=1}^{n} x_i(y_i - (\overline{y} - \hat{\beta}_1 \overline{x}) - \hat{\beta}_1 x_i) =& 0 \text{ (using solution for } \hat{\beta}_0) \\
\Rightarrow \sum_{i=1}^{n} x_i(y_i - \overline{y}) =& \hat{\beta}_1 \sum_{i=1}^{n} x_i(x_i - \overline{x}) \\
\Rightarrow \hat{\beta}_1 =& \frac{\sum_{i=1}^{n} x_i(y_i - \overline{y})}{\sum_{i=1}^{n} x_i(x_i - \overline{x})}
\end{aligned}
$$

Deriving least squares estimates: $\hat{\beta}_1$

- Solution for $\hat{\beta}_1$ can be rewritten as

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

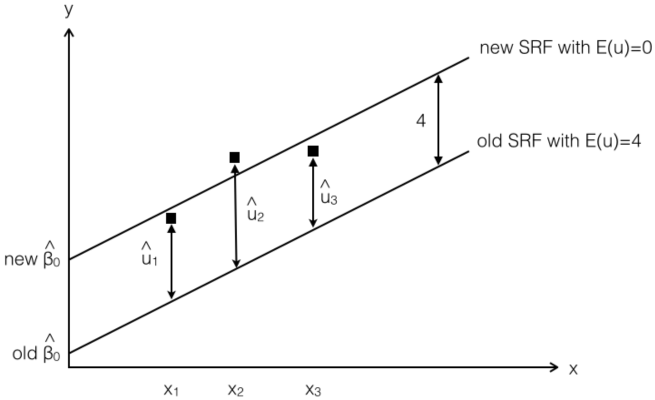
  since

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

- $\hat{\beta}_1$ is the sample covariance between x and y divided by the sample variance of x.

Imposing $E(u) = 0$ is not restrictive

- If the random components had an expected value of 4, i.e. $E(u) = 4$ then we could simply increase the constant by 4 to ensure $E(u) = 4$.

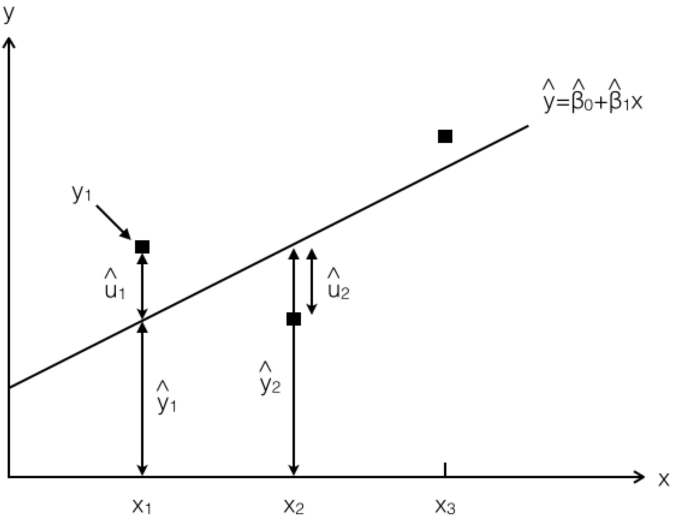


Least squares estimates

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are called the least squares estimates because they **minimize** the **sum of squared residuals**
- The **residual** for observation i is the difference between the actual value and the fitted value:

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

Fitted values and residuals



Least squares estimates

- Least squares procedure chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the following sum:

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

- First order conditions are (differentiating with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$ using the chain rule):

$$\frac{d}{d\hat{\beta}_0} \sum_{i=1}^{n} \hat{u}_i^2 = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\frac{d}{d\hat{\beta}_1} \sum_{i=1}^{n} \hat{u}_i^2 = -2 \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

which are (for practical purposes) identical to the two assumptions above.

Population versus sample regression function

- The regression line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ is the **sample regression function (SRF)** because it is the estimated version of the **population regression function (PRF)** $E(y|x) = \beta_0 + \beta_1 x$
  - Slope estimate implies that $\Delta\hat{y} = \hat{\beta}_1 \Delta x$
- The PRF is something fixed, but unknown, in the population.
- With a different sample of data we would estimate a different (but hopefully very similar) SRF

CEO salary and return on equity (roe)

- Data (CEOSAI1.dta) consists of salaries (in thousands \$) and return equity for 209 CEOs in 1990 from *Business Week*
  - ROE is average for years 1988, 1989, and 1990
- Descriptive statistics from Stata for these two variables:

```
summ salary roe
```

Variable	Obs	Mean	Std. Dev.	Min	Max
salary	209	1281.12	1372.345	223	14822
roe	209	17.18421	8.518509	.5	56.3

CEO salary and return on equity (roe)

- Stata estimates of the model $salary = \beta_0 + \beta_1 roe + u$:

```
. reg salary roe

      Source |       SS           df       MS      Number of obs   =       209
-------------+----------------------------------   F(1, 207)       =      2.77
       Model |  5166419.04         1  5166419.04   Prob > F        =    0.0978
    Residual |   386566563       207  1867471.32   R-squared       =    0.0132
-------------+----------------------------------   Adj R-squared   =    0.0084
       Total |   391732982       208  1883331.64   Root MSE        =    1366.6

------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         roe |   18.50119   11.12325     1.66   0.098    -3.428196    40.43057
       _cons |   963.1913   213.2403     4.52   0.000     542.7902    1383.592
------------------------------------------------------------------------------
```

CEO salary and return on equity (roe): Stata output

- Stata estimates of the model $salary = \beta_0 + \beta_1 roe + u$:

y x

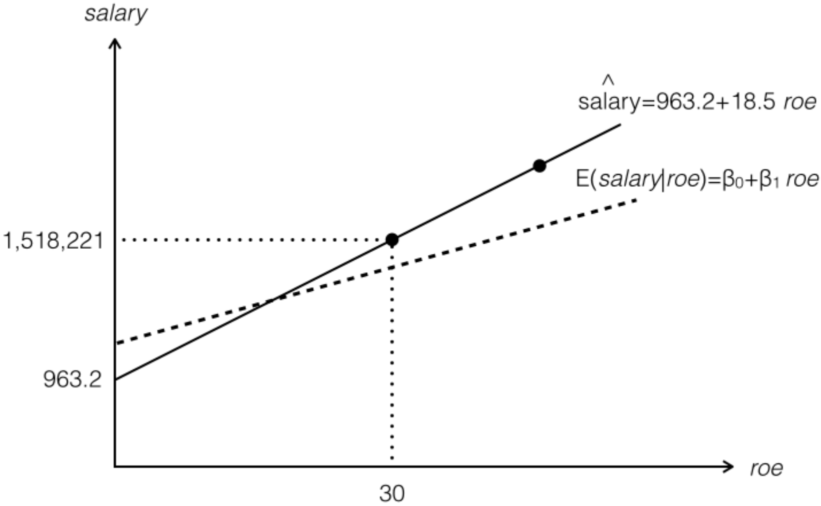
```
. reg salary roe
```

Source	SS	df	MS
Model	5166419.04	1	5166419.04
Residual	386566563	207	1867471.32
Total	391732982	208	1883331.64

Number of obs =	209
F(1, 207) =	2.77
Prob > F =	0.0978
R-squared =	0.0132
Adj R-squared =	0.0084
Root MSE =	1366.6

	salary	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
$\widehat{\beta}_1$	roe	18.50119	11.12325	1.66	0.098	-3.428196	40.43057
$\widehat{\beta}_0$	_cons	963.1913	213.2403	4.52	0.000	542.7902	1383.592

Sample regression function (SRF, thick line) and (unknown) population regression function (PRF, dotted)



Wage and education (using WAGE1.dta)

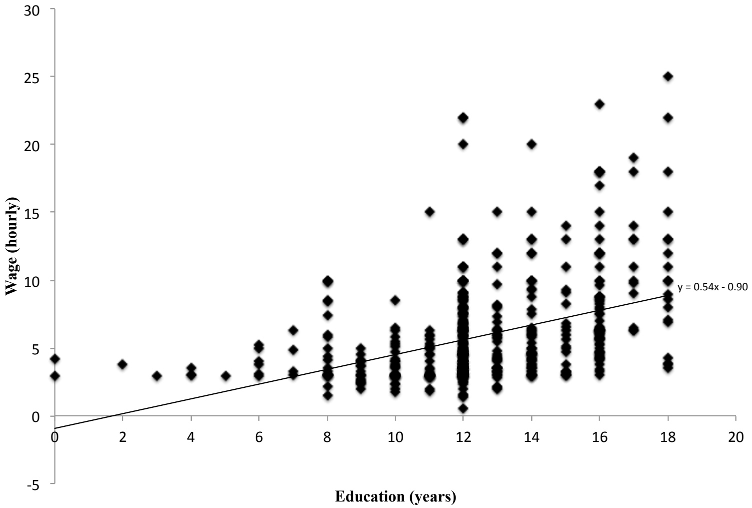
- Stata estimates of the model $wage = \beta_0 + \beta_1 education + u$ using data on 526 individuals:

```
. reg wage educ

      Source |       SS           df       MS      Number of obs   =       526
-------------+----------------------------------   F(1, 524)       =    103.36
       Model |  1179.73204         1  1179.73204   Prob > F        =    0.0000
    Residual |  5980.68225       524  11.4135158   R-squared       =    0.1648
-------------+----------------------------------   Adj R-squared   =    0.1632
       Total |  7160.41429       525  13.6388844   Root MSE        =    3.3784

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .5413593    .053248    10.17   0.000     .4367534    .6459651
       _cons |  -.9048516   .6849678    -1.32   0.187    -2.250472    .4407687
------------------------------------------------------------------------------
```

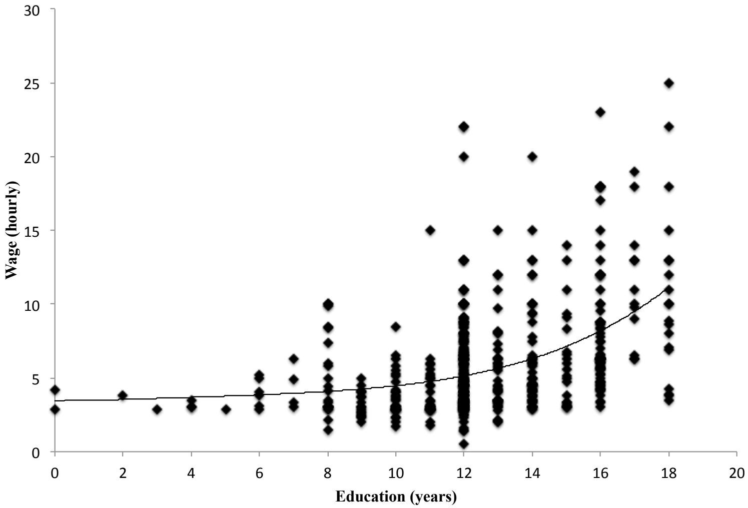
Wage and education

- OLS regression line: $wage = -0.90 + 0.54\ education$
- Predicted hourly wage for a person with 0 years of education is -$0.90 an hour (?!?)
  - Regression line tries to minimize the sum of (squared) residuals
  - Most of the people in the data have education levels of 8 years or above (only 18 of 526 people have 7 or fewer years of education), so the regression line concentrates on these people
- Linear nature implies that every four years of education increases salary by 4*0.54=$2.16, regardless of initial level of education
  - We likely need a specification with non constant effects of education

Wage and education with linear regression



Wage and education with more flexible regression line



Algebraic properties of OLS statistics

- The sum and the sample average of the OLS residuals is zero:

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

- The sample covariance between the regressors and OLS residuals is zero

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

- The point $(\overline{x}, \overline{y})$ is on the OLS regression line
  - For *salary* and *roe* regression: $\overline{y} = 963.1913 + 18.50119\overline{x} = 963.1913 + 18.50119 * 17.18421 = \$1281.12$

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:

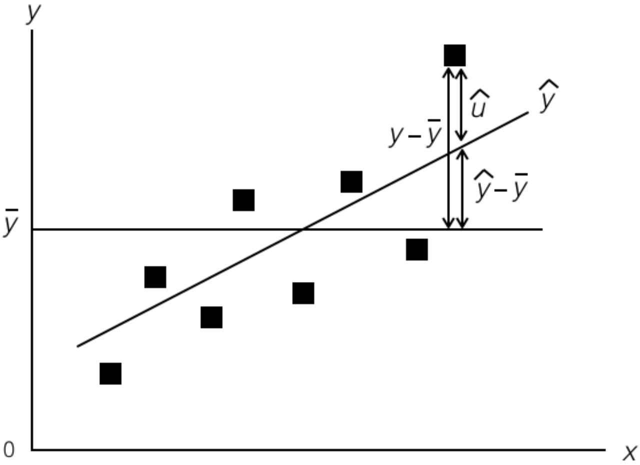
$$SSE \equiv \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$$

  - 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



Goodness of fit

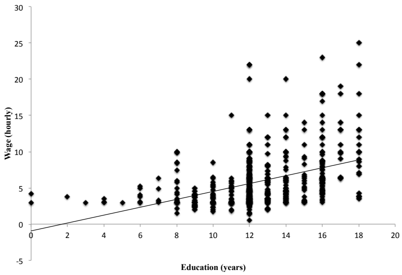
- The **R-squared** or coefficient of determination is

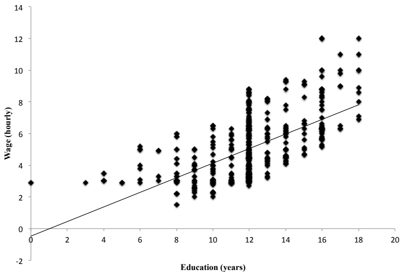
$$R^2 \equiv \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

- R^2 is the ratio of the explained variation compared to the total variation in y
    - Ranges from 0 to 1
    - If all data points lie on the regression line (so that $\hat{u}_i = 0$ for all i and $SSR = 0$) then $R^2 = 1$.
- How much should we care about R-squared values?

Different R^2 values

- R^2 on left is 16.5; R^2 on the right is 33.2
  - To construct the chart on the right, I removed some of the largest $\hat{u}$ values from the chart on the left





Example: Computing OLS estimates by hand

- The following table contains the budgets and box office revenue for six movies:

Movie	Revenue	Budget
Chicken Run	107	45
Fantastic Four	155	100
Frida	26	12
From Hell	32	35
Lord of the Rings	313	93
Mission: Impossible II	215	125

Example: Computing OLS estimates by hand

movie	*rev* (y)	*bdgt* (x)	$y-\overline{y}$	$x-\overline{x}$	$(y-\overline{y})\cdot(x-\overline{x})$	$(x-\overline{x})^2$
Chicken	107	45	-34.33	-23.33	801.11	544.44
Fant. 4	155	100	13.67	31.67	432.78	1002.78
Frida	26	12	-115.33	-56.33	6497.11	3173.44
F. H.	32	35	-109.33	-33.33	3644.44	1111.11
LOTR	313	93	171.67	24.67	4234.44	608.44
M:i-2	215	125	73.67	56.67	4174.44	3211.11
$\sum_{i=1}^{6}$	848	410	0.00	0.00	19784.33	9651.33
$\frac{1}{6}\sum_{i=1}^{6}$	141.33	68.33				

- $\hat{\beta}_1 = \frac{19{,}784.33}{9{,}651.33} = 2.05, \quad \hat{\beta}_0 = 141.33 - 2.05 \cdot 68.33 = 1.26$

Example: Computing OLS estimates by hand

- Estimated equation is:

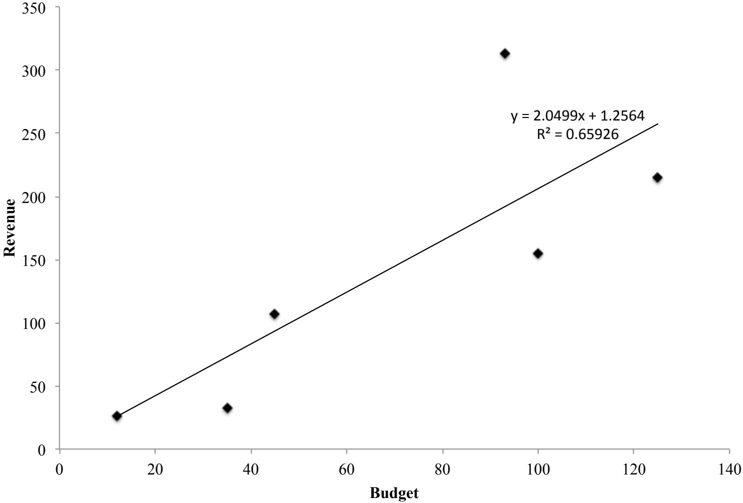
$$\widehat{rev} = 1.26 + 2.05 bdgt$$

Movie	y	$\hat{y}$	$\hat{u}$	$\hat{u}^2$	$(y - \overline{y})^2$
Chicken...	107	93.50	13.50	182.19	1178.78
Fantastic 4	155	206.25	-51.25	2626.26	186.78
Frida	26	25.86	0.14	0.02	133031.78
From Hell	32	73.00	-41.00	1681.25	11953.78
LOTR	313	191.90	121.10	14665.77	29469.44
M:i-2	215	257.49	-42.49	1805.80	5426.78
$\sum_{i=1}^{6}$			0.00	20961.29	61517.33

Example: Computing OLS estimates by hand

- So $R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{20961.29}{61517.33} = 0.6593$
- Therefore, about 65.9% of the variation in revenues is explained by budgets in this very small sample of movies.
- Note that:
  - $\sum_{i=1}^{6} \hat{u}_i = 0$
  - It can be shown that $\sum_{i=1}^{6} x_i \hat{u}_i = 0$

Example: Computing OLS estimates by hand



Changing units of measurement

- If the dependent (LHS) variable is multiplied by a constant c then the intercept and slope term will also be multiplied by that same constant
  - Example: if we specified *salary* in dollars, rather than thousands of dollars, the estimated equation becomes:

$$salary\_in\_dollars = 963,191 + 18,501\ roe + u$$

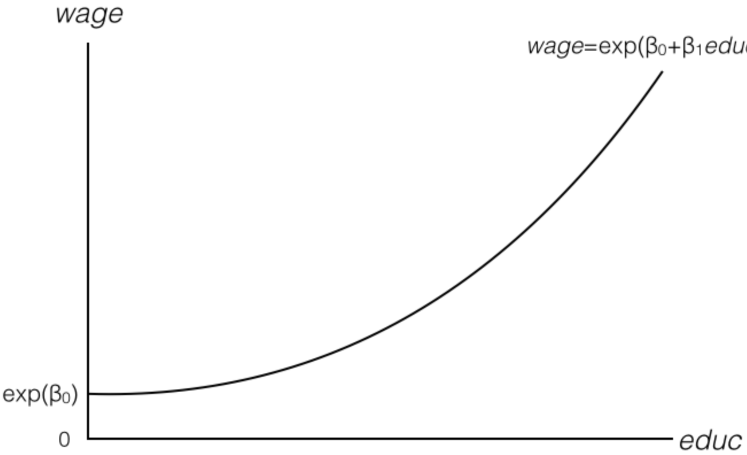
- If the independent (RHS) variable is multiplied or divided by a constant c then the slope term is also divided or multiplied by that same constant
  - Example: if we divide *roe* by 100, the slope termed is multiplied by 100:

$$salary = 963.191 + 1,850.1\ roe\_in\_decimal + u$$

Incorporating nonlinearities

- It is easy to incorporate *some* types of nonlinear relationships between left- and right-hand side variables
  - Note that *linear* regression means that the regression is linear in the parameters β_0 and β_1. An example of a *nonlinear* regression model is $y = \frac{\beta_1 x}{\beta_2 + x}$
- If we wanted a right-hand side variable to increase (or decrease) the left-hand side by a constant **percentage** then we could replace y with the natural logarithm of y:
  - Example: $\log(wage) = \beta_0 + \beta_1 educ + u$
  - A one year increase in education increases wages by $100 \cdot \beta_1$ percent

Wage and education with a more flexible specification



Wage and education with a more flexible specification

- If we wanted a right-hand side variable to have a constant proportional increase (or decrease) on the left-hand side then we could replace both y and x with their natural logarithms:
  - Example: $\log(salary) = \beta_0 + \beta_1 \log(sales) + u$
  - A one percent increase in firm sales increases CEO salary by β_1 percent

Functional forms involving logarithms

Model	Dependent variable	Independent variable	Interpretation of β_1
Level-level	y	x	$\Delta y = \beta_1 \Delta x$
Level-log	y	$\log(x)$	$\Delta y = (\beta_1/100)\%\Delta x$
Log-level	$\log(y)$	x	$\%\Delta y = (100 \cdot \beta_1)\Delta x$
Log-log	$\log(y)$	$\log(x)$	$\%\Delta y = \beta_1 \%\Delta x$

Next week

- Simple regression model continued (Chapter 2)
- Multiple regression analysis (Chapter 3)