# Simple Regression Model

# Terminology

Regression model: $y = \beta_0 + \beta_1 x + u$

y is dependent variable, *x* is independent variable (one independent variable for a simple regression), u is error, $\beta_0$ and $\beta_1$ are parameters.

Estimated equation: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

$\hat{y}$ is predicted value, $\hat{\beta}_0$ and $\hat{\beta}_1$ are coefficients

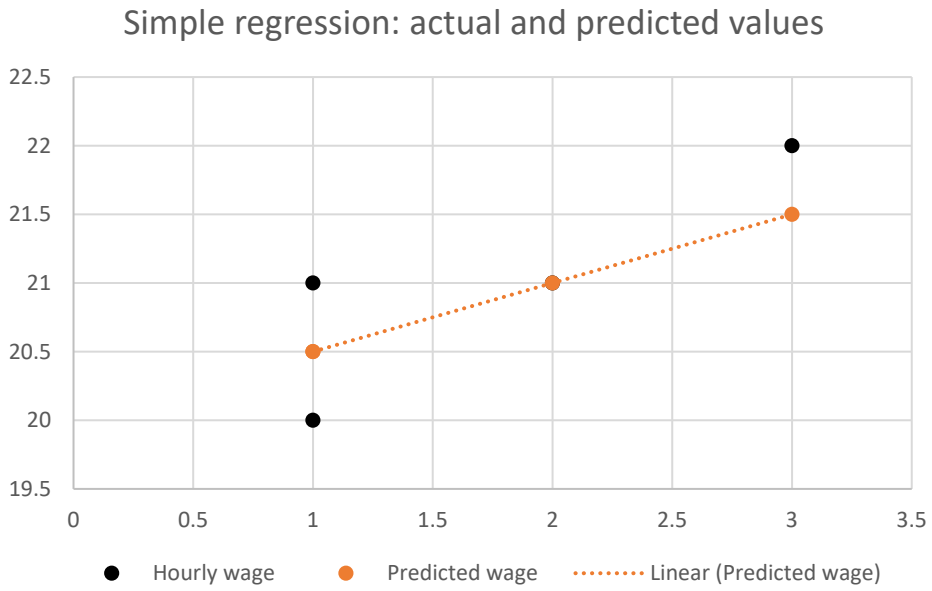
| Population | Sample |
|---|---|
| Parameter $\beta$ | Coefficient $\hat{\beta}$ |
| Error $u$ | Residual $\hat{u}$ |

Residual: $\hat{u} = y - \hat{y}$

$\hat{u}$ = actual value minus predicted value for dependent variable

# Simple regression model example

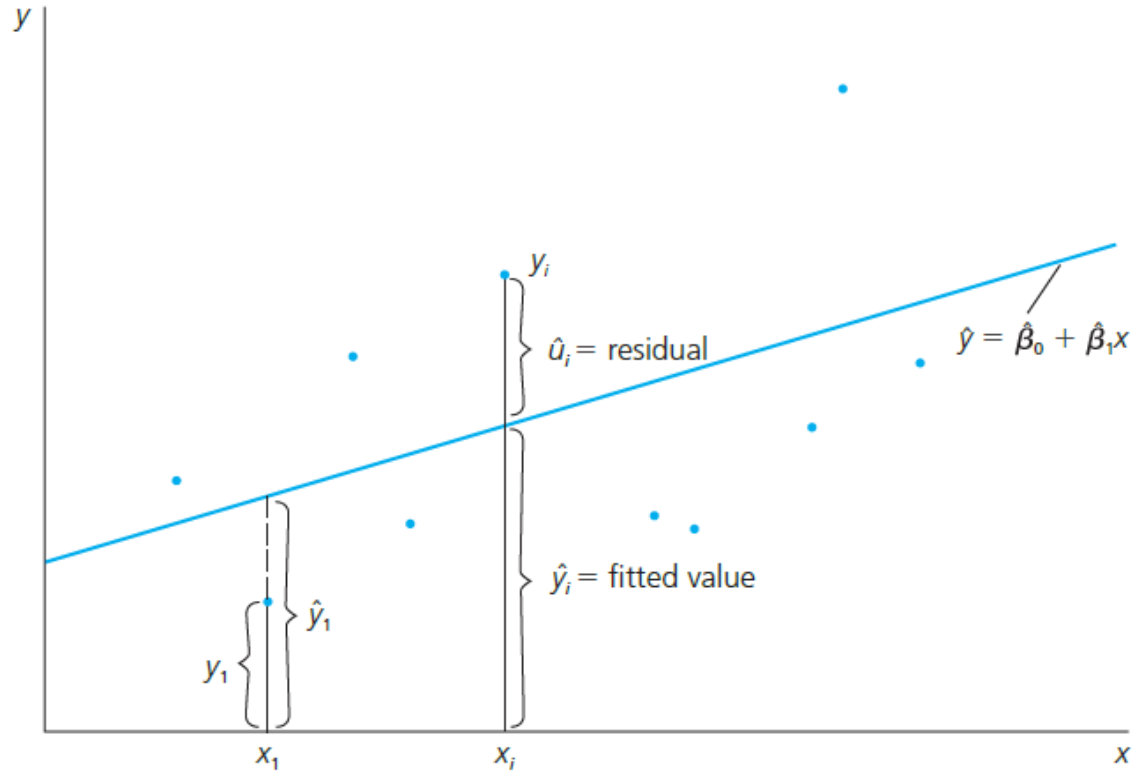
| Dependent variable y | Indep. variable x | Predicted value $\hat{y} = 20 + 0.5x$ | Residual $\hat{u} = y - \hat{y}$ |
|---|---|---|---|
| Hourly wage $ | Years of experience | | |
| 20 | 1 | =20+0.5*1=20.5 | =20-20.5=-0.5 |
| 21 | 2 | =20+0.5*2=21 | =21-21=0 |
| 21 | 1 | =20+0.5*1=20.5 | =21-20.5=0.5 |
| 22 | 3 | =20+0.5*3=21.5 | =22-21.5=0.5 |




Simple regression: hourly wage depends on years of experience.
Figure shows regression line, slope, predicted values, actual values, and residuals.

# Simple regression:
# actual values, predicted values, and residuals

Regression line fits as good as possible through the data points



# Interpretation of coefficients

$$\hat{\beta}_1 = \frac{\Delta y}{\Delta x} = \frac{change\ in\ y}{change\ in\ x}$$

- The coefficient $\hat{\beta}_1$ measures by how much the dependent variable changes when the independent variable changes by one unit.
- $\hat{\beta}_1$ is also called slope in the simple linear regression.
- A derivative of a function is another function showing the slope.
- The formula above is correct if $\frac{\Delta u}{\Delta x}$=0, which means all other factors are fixed.

# Population regression function

Population regression function:

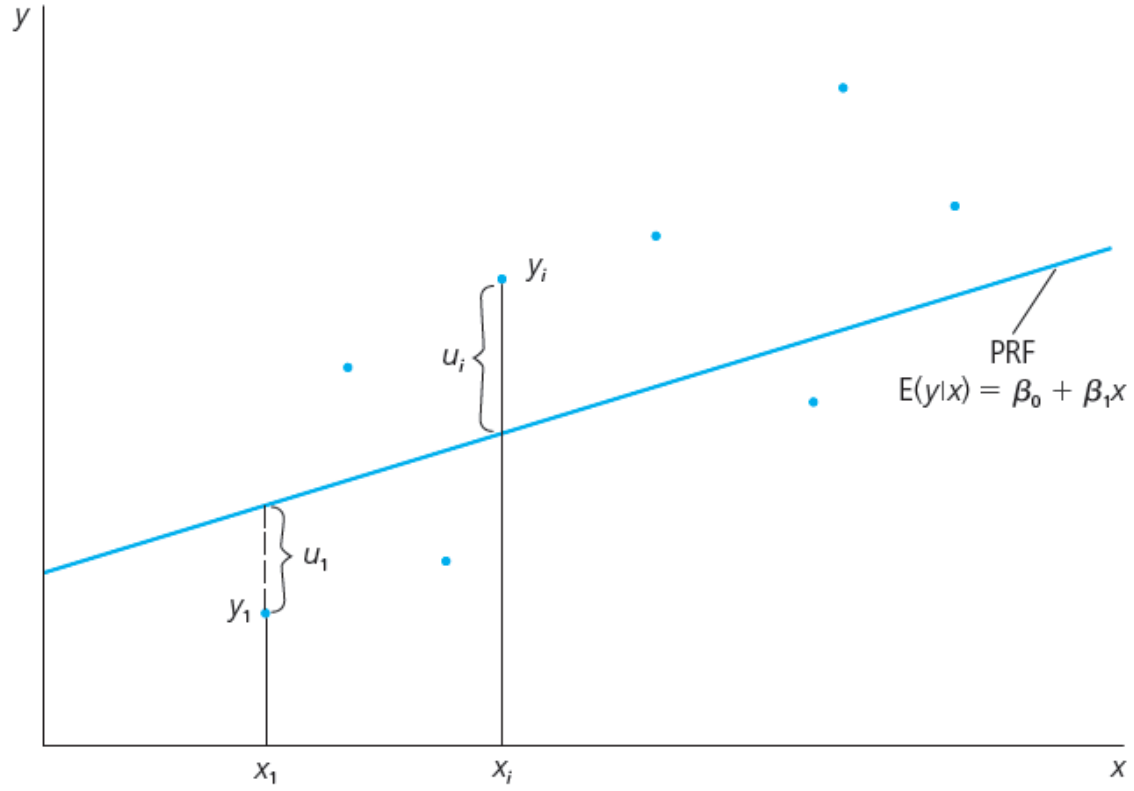
$E(y|x) = E(\beta_0 + \beta_1 x + u|x) = \beta_0 + \beta_1 x + E(u|x) = \beta_0 + \beta_1 x$

if $E(u|x)$=0 (this assumption is called zero conditional mean)

For the population, the average value of the dependent variable can be expressed as a linear function of the independent variable.

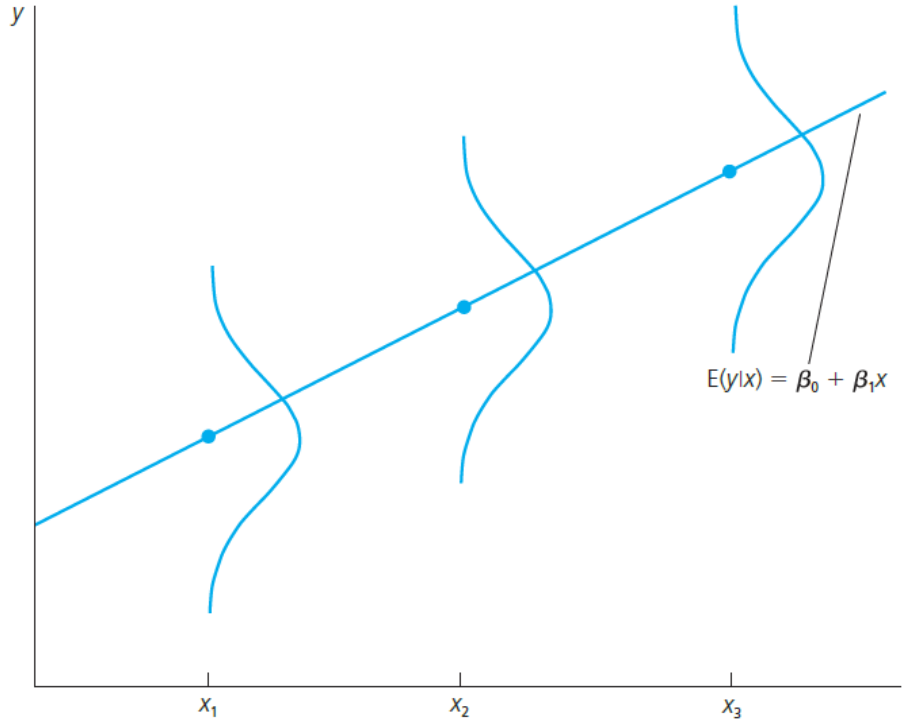
# Population regression function

- Population regression function shows the relationship between y and x for the population



# Population regression function

- For individuals with a particular x, the average value of y is $E(y|x) = \beta_0 + \beta_1 x$




- Note that $x_1$, $x_2$, $x_3$ here refers to $x_i$ and not different variables

# Derivation of the OLS estimates

- For a regression model: $y = \beta_0 + \beta_1 x + u$
- We need to estimate the regression equation: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ and find the coefficients $\hat{\beta}_0$and $\hat{\beta}_1$ by looking at the residuals
- $\hat{u} = y - \hat{y} = y - \hat{\beta}_0 - \hat{\beta}_1 x$
- Obtain a random sample of data with n observations

  $(x_i, y_i)$, where $i = 1 \ldots n$ is the observation
- The goal is to obtain as good fit as possible of the estimated regression equation

# Derivation of the OLS estimates

- Minimize the sum of squared residuals

$$\min \sum_{i=1}^{n} \hat{u}^2 = \sum_{i=1}^{n} (y - \hat{\beta}_0 - \hat{\beta}_1 x)^2$$

We obtain OLS coefficients:

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{cov(x, y)}{var(x)}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

OLS is Ordinary Least Squares, based on minimizing the squared residuals.

# OLS properties

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

The sample average of the dependent and independent variable are on the regression line

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

The residuals sum up to zero (note that we minimized the sum of squared residuals)

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

The covariance between the independent variable and residual is zero.

# Simple regression example: CEO's salary

Simple regression model explaining how return on equity (roe) affects CEO's salary.

Regression model

$$salary = \beta_0 + \beta_1 roe + u$$

Estimated equation for predicted value of wage

$$\widehat{salary} = \hat{\beta}_0 + \hat{\beta}_1 roe$$

Residuals

$$\hat{u} = salary - \widehat{salary}$$

We estimate the regression model to find the coefficients.

$\hat{\beta}_1$ measures the change in the CEO's salary associated with one unit increase in roe, holding other factors fixed.

# Estimated equation and interpretation

- Estimated equation

$$\widehat{salary} = \hat{\beta}_0 + \hat{\beta}_1 roe = 963.191 + 18.501\, roe$$

- Salary is measured in thousand dollars, ROE (return on equity) is measured in %.
- $\hat{\beta}_1$ measures the change in the CEO's salary associated with one unit increase in roe, holding other factors fixed.
- Interpretation of $\hat{\beta}_1$: the CEO's salary increases by \$18,501 for each 1% increase in ROE.
- Interpretation of $\hat{\beta}_0$: if the ROE is zero, the CEO's salary is \$963,191.

# Stata output for simple regression

```
. regress salary roe
```

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

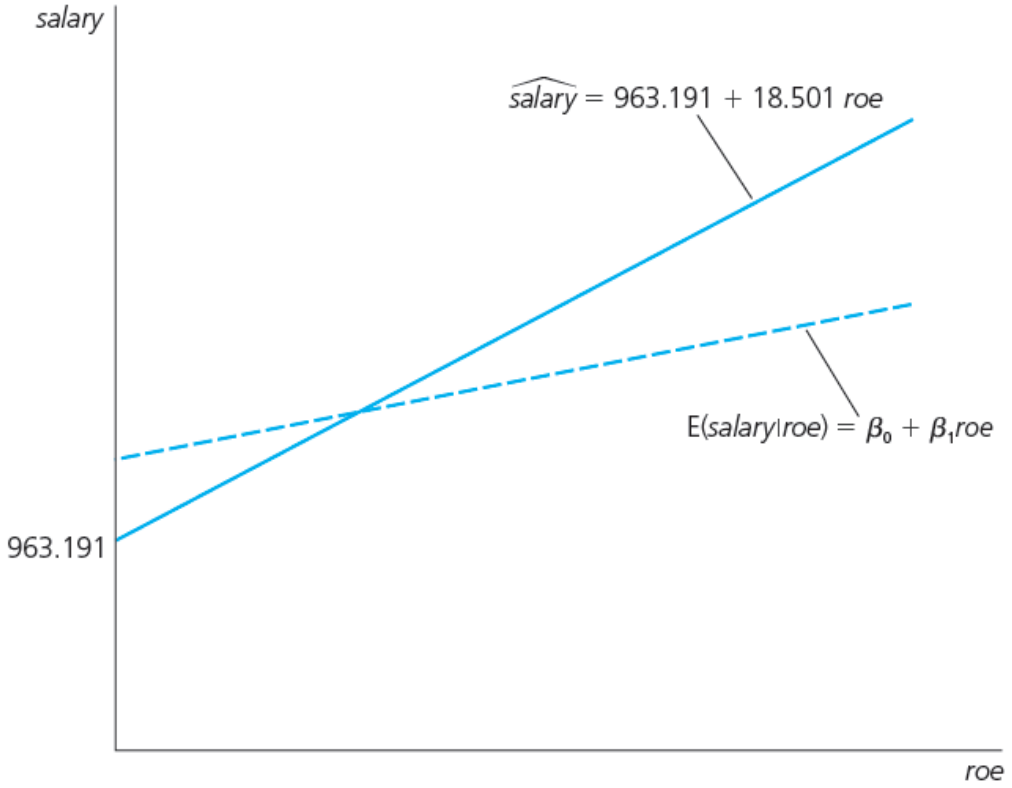
$$\widehat{salary} = \hat{\beta}_0 + \hat{\beta}_1 roe = 963.191 + 18.501\, roe$$

# Simple regression results in a table

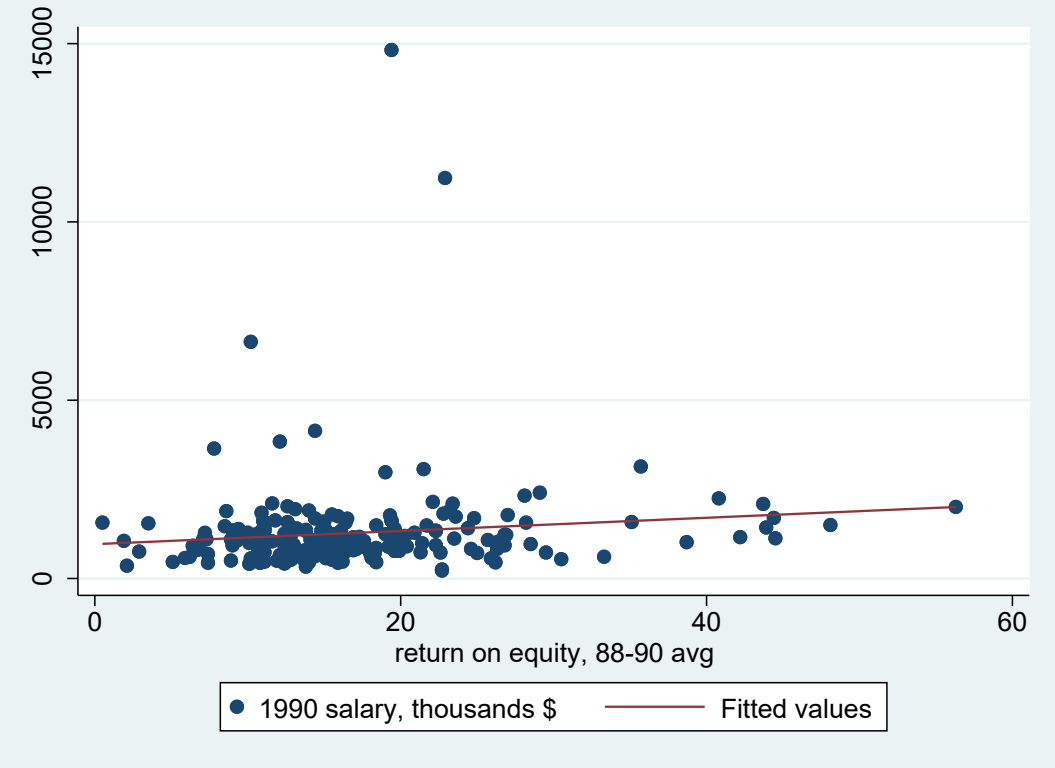
| | (1) |
|---|---|
| VARIABLES | salary |
| | |
| roe | 18.50* |
| | (11.12) |
| Constant | 963.2*** |
| | (213.2) |
| | |
| Observations | 209 |
| R-squared | 0.013 |

$$\widehat{salary} = \hat{\beta}_0 + \hat{\beta}_1 roe = 963.191 + 18.501\, roe$$

# Regression line for sample vs population regression function for population

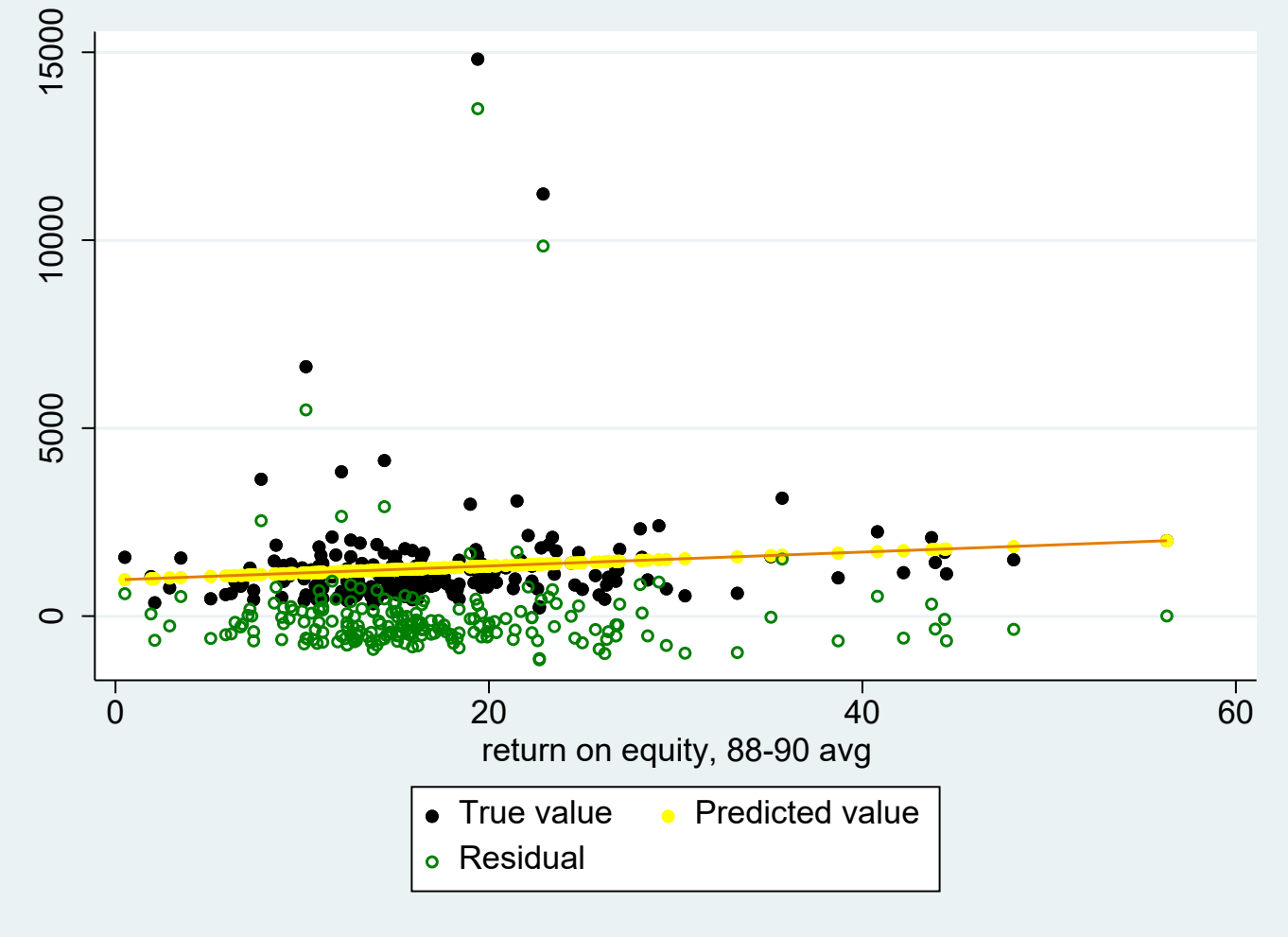


# Estimated regression




Actual and predicted values

# Actual values, predicted values, and residuals



# Actual, predicted values, and residuals

| roe | salary | $\widehat{salary}$ | $\hat{u}$ |
|---|---|---|---|
| | | predicted value $963.191 + 18.501\, roe$ | Residual $salary - \widehat{salary}$ |
| 14.1 | 1095 | 1224 | -129 |
| 10.9 | 1001 | 1165 | -164 |
| 23.5 | 1122 | 1398 | -276 |
| 5.9 | 578 | 1072 | -494 |
| 13.8 | 1368 | 1219 | 149 |
| 20 | 1145 | 1333 | -188 |
| 16.4 | 1078 | 1267 | -189 |
| 16.3 | 1094 | 1265 | -171 |
| 10.5 | 1237 | 1157 | 80 |
| 26.3 | 833 | 1450 | -617 |

The mean salary is 1,281 ($1,281,000). The mean predicted salary is also 1,281.
The mean for the residuals is zero.

# Simple regression example: wage

Simple regression model explaining how education affects wages for workers.

Regression model

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

Estimated equation for predicted value of wage

$$\widehat{wage} = \hat{\beta}_0 + \hat{\beta}_1 educ$$

Residuals

$$\hat{u} = wage - \widehat{wage}$$

We estimate the regression model to find the coefficients.

$\hat{\beta}_1$ measures the change in wage associated with one more year of education, holding other factors fixed.

# Estimated equation and interpretation

- Estimated equation

$$\widehat{wage} = \hat{\beta}_0 + \hat{\beta}_1 educ = -0.90 + 0.54\, educ$$

- Wage is measured in $/hour. Education is measured in years.
- $\hat{\beta}_1$ measures the change in person's wage associated with one additional year increase in education, holding other factors fixed.
- Interpretation of $\hat{\beta}_1$: the hourly wage increases by $0.54 for additional year of education.
- Interpretation of $\hat{\beta}_0$: if education is zero, person's wage is -$0.90 (but no one in the sample has zero education).

# Stata output for simple regression

```
. reg wage educ
```

| Source | SS | df | MS | Number of obs | = | 526 |
|---|---|---|---|---|---|---|
| | | | | F(1, 524) | = | 103.36 |
| Model | 1179.73205 | 1 | 1179.73205 | Prob > F | = | 0.0000 |
| Residual | 5980.68226 | 524 | 11.4135158 | R-squared | = | 0.1648 |
| | | | | Adj R-squared | = | 0.1632 |
| Total | 7160.41431 | 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 | -.9048517 | .6849678 | -1.32 | 0.187 | -2.250472 | .4407687 |

# Variations

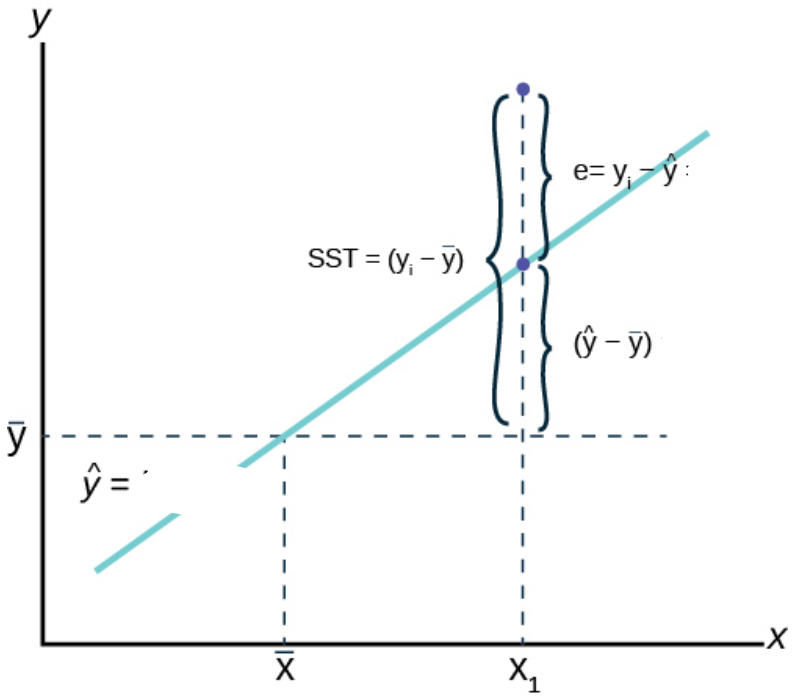
$$SST = \sum_{i=1}^{n}(y - \bar{y})^2 \quad SSE = \sum_{i=1}^{n}(\hat{y} - \bar{y})^2 \quad SSR = \sum_{i=1}^{n}(y - \hat{y})^2 = \sum_{i=1}^{n}\hat{u}^2$$

$$SST = SSE + SSR$$

- SST is total sum of squares and measures the total variation in the dependent variable
- SSE is explained sum of squares and measures the variation explained by the regression
- SSR is residual sum of squares and measures the variation not explained by the regression

Note: some call SSE error sum of squared and SSR regression sum of squares, where R & E are confusingly reversed.

# Variations



# Goodness of fit measure
# R-squared

- $R^2$ = SSE/SST = 1 – SSR/SST
- R-squares is explained sum of squares divided by total sum of squares.
- R-squared is a goodness of fit measure.  It measures the proportion of total variation that is explained by the regression.
- An R-squared of 0.7 is interpreted as 70% of the variation is explained by the regression and the rest is due to error.
- R-squared that is greater than 0.25 is considered good fit.

# R-squared calculated

```
. reg wage educ
```

| Source | SS | df | MS | Number of obs | = | 526 |
|---|---|---|---|---|---|---|
| | | | | F(1, 524) | = | 103.36 |
| Model | 1179.73205 | 1 | 1179.73205 | Prob > F | = | 0.0000 |
| Residual | 5980.68226 | 524 | 11.4135158 | R-squared | = | 0.1648 |
| | | | | Adj R-squared | = | 0.1632 |
| Total | 7160.41431 | 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 | -.9048517 | .6849678 | -1.32 | 0.187 | -2.250472 | .4407687 |

R-squared = SS Model /SS Total = 1179.73 / 7160.41 = 0.1648
16% of the variation in wage is explained by the regression and the rest is due to error.
This is not a very good fit.

# Log transformation (logged variables)

- Sometimes variables (y or x) are expressed as logs, log(y) or log(x)
- With logs, interpretation is in percentage/elasticity
- Variables such as age and education that are measured in units such as years should not be logged
- Variables measured in percentage points (e.g. interest rates) should not be logged
- Logs cannot be used if variables have zero or negative values
- Taking logs often reduces problems with large values or outliers
- Taking logs helps with homoskedasticity and normality

# Log-log form

- Linear regression model: $y = \beta_0 + \beta_1 x + u$
- log-log form: $log(y) = \beta_0 + \beta_1 \log(x) + u$
- Instead of the dependent variable, use log of the dependent variable.
- Instead of the independent variable, use log of the independent variable.

$$\hat{\beta}_1 = \frac{\Delta \log(y)}{\Delta \log(x)} = \frac{\Delta y}{y} \frac{x}{\Delta x} = \frac{percent\ change\ in\ y}{percent\ change\ in\ x}$$

- The dependent variable changes by $\hat{\beta}_1$ percent when the independent variable changes by one percent.

# Log-linear form (also called semi-log)

- Linear regression model: $y = \beta_0 + \beta_1 x + u$
- Log-linear form: $log(y) = \beta_0 + \beta_1 x + u$
- Instead of the dependent variable, use log of the dependent variable.

$$\hat{\beta}_1 = \frac{\Delta \log(y)}{\Delta x} = \frac{\Delta y}{y}\frac{1}{\Delta x} = \frac{percent\ change\ in\ y}{change\ in\ x}$$

- The dependent variable changes by $\hat{\beta}_1$*100 percent when the independent variable changes by one unit.

# Linear-log form

- Linear regression model: $y = \beta_0 + \beta_1 x + u$
- Linear-log form: $y = \beta_0 + \beta_1 \log(x) + u$
- Instead of the independent variable, use log of the independent variable.

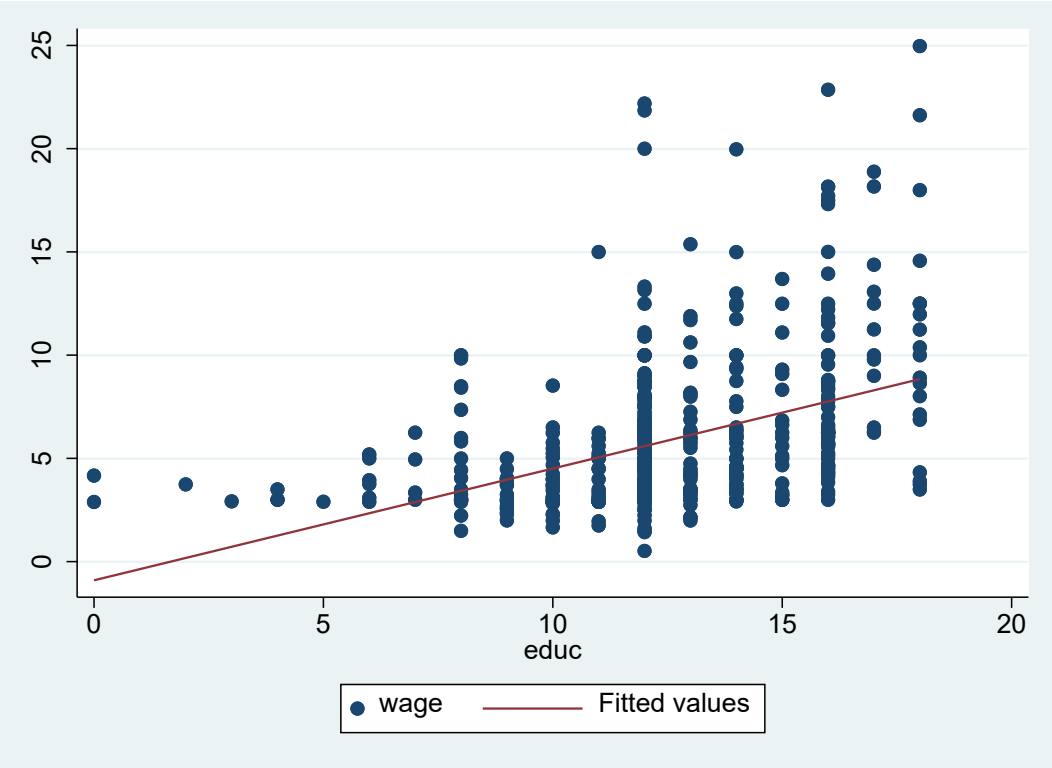
$$\hat{\beta}_1 = \frac{\Delta y}{\Delta \log(x)} = \Delta y \frac{x}{\Delta x} = \frac{change\ in\ y}{percent\ change\ in\ x}$$

- The dependent variable changes by $\hat{\beta}_1/100$ units when the independent variable changes by one percent.

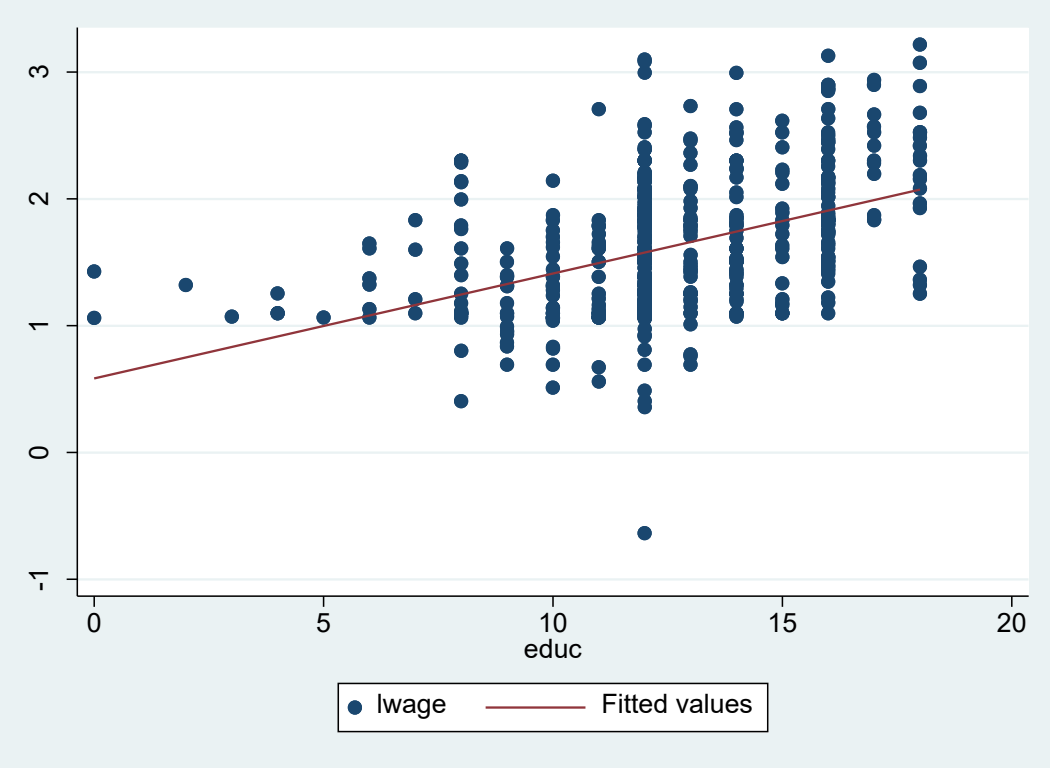
# Example of data with logs

| wage | lwage | educ |
|---|---|---|
| 3.10 | 1.13 | 11 |
| 3.24 | 1.18 | 12 |
| 3.00 | 1.10 | 11 |
| 6.00 | 1.79 | 8 |
| 5.30 | 1.67 | 12 |
| 8.75 | 2.17 | 16 |
| 11.25 | 2.42 | 18 |
| 5.00 | 1.61 | 12 |
| 3.60 | 1.28 | 12 |
| 18.18 | 2.90 | 17 |

# Linear vs log-linear form




Linear form: wage on education




Log-linear form: log wage on education

# Linear vs log-linear form

| | (1) | (2) |
|---|---|---|
| VARIABLES | wage | lwage |
| educ | 0.541*** | 0.0827*** |
| | (0.0532) | (0.00757) |
| Constant | -0.905 | 0.584*** |
| | (0.685) | (0.0973) |
| Observations | 526 | 526 |
| R-squared | 0.165 | 0.186 |

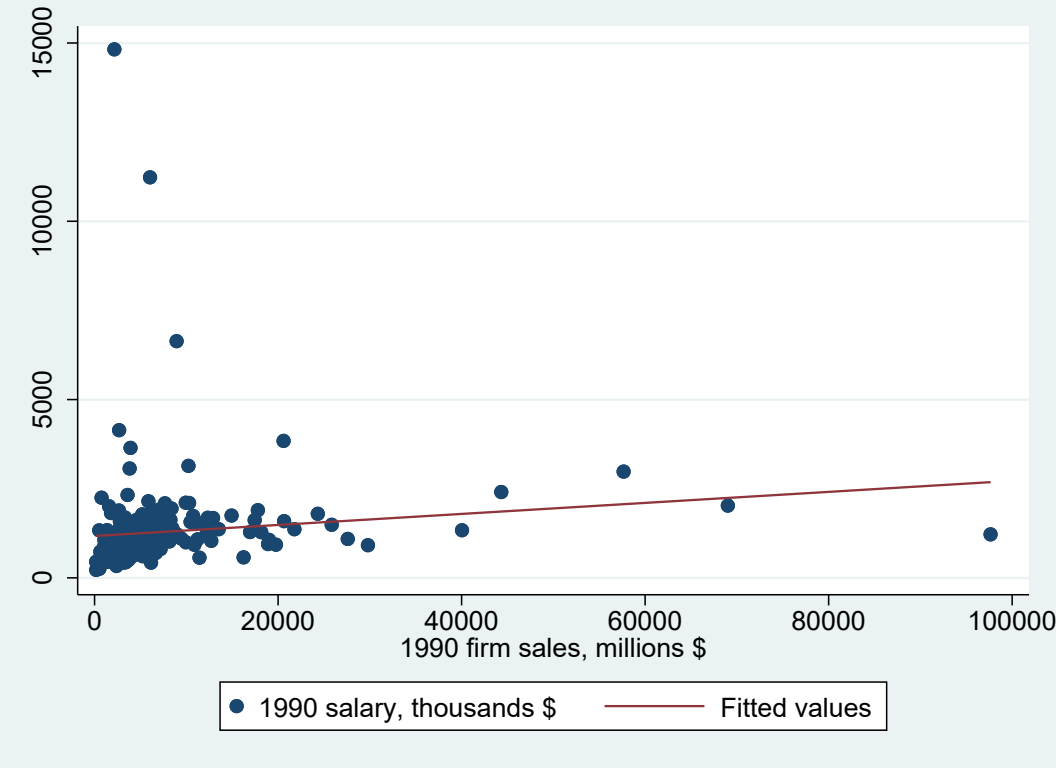
Linear form: wage increases by $0.54 for each additional year of education.
Log-linear form: wage increases by 8.2% for each additional year of education.

# Example of data with logs

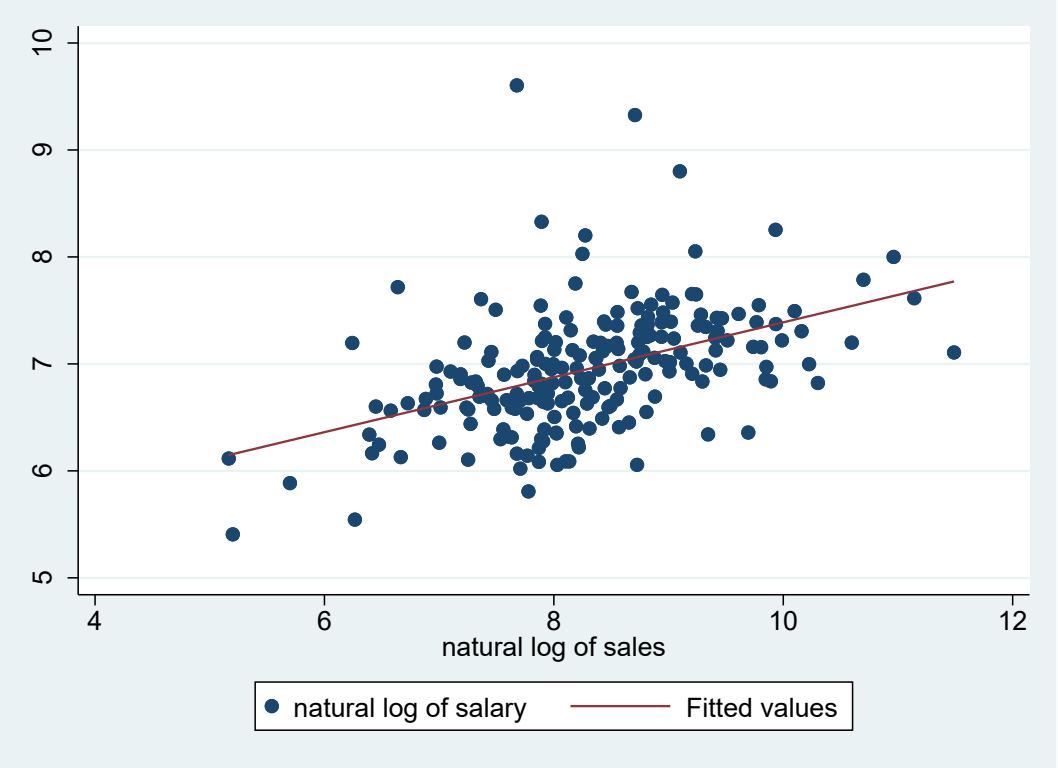
| Salary (thousand dollars) | lsalary | Sales (Million dollars) | lsales |
|---|---|---|---|
| 1095 | 7.0 | 27595 | 10.2 |
| 1001 | 6.9 | 9958 | 9.2 |
| 1122 | 7.0 | 6126 | 8.7 |
| 578 | 6.4 | 16246 | 9.7 |
| 1368 | 7.2 | 21783 | 10.0 |
| 1145 | 7.0 | 6021 | 8.7 |
| 1078 | 7.0 | 2267 | 7.7 |
| 1094 | 7.0 | 2967 | 8.0 |
| 1237 | 7.1 | 4570 | 8.4 |
| 833 | 6.7 | 2830 | 7.9 |

Note that one unit is thousand dollars for salary and million dollars for sales.

# Linear vs log-log form

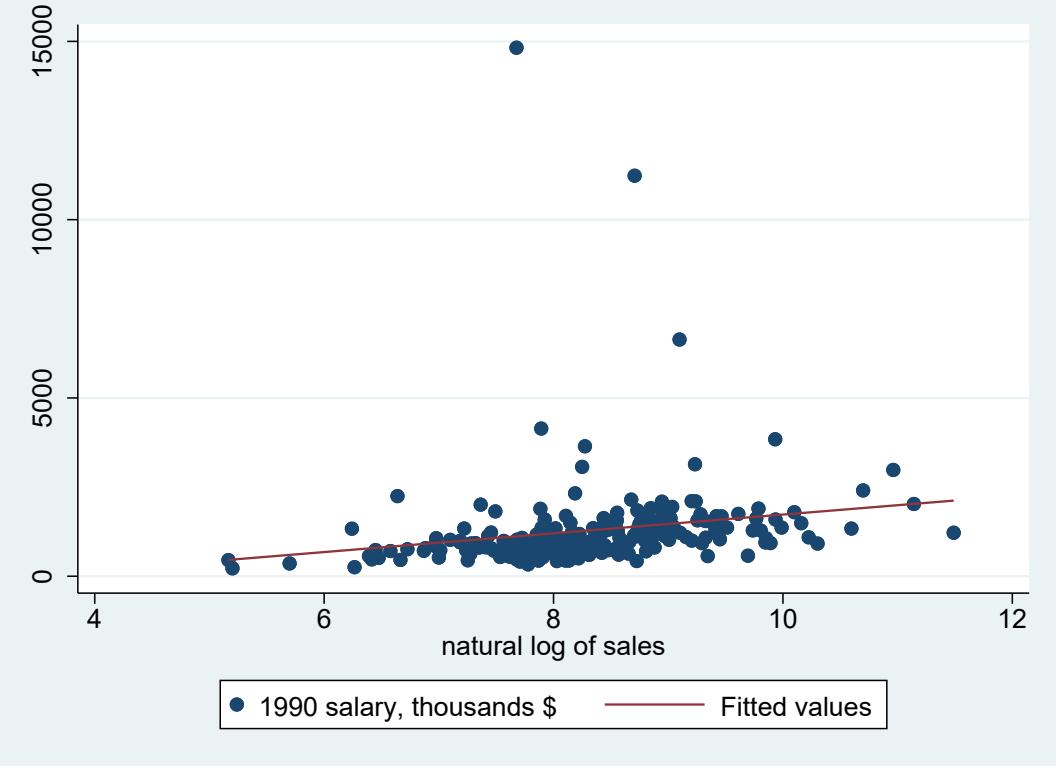



Linear form: salary on sales




Log-log form: log salary on log sales

# Log-linear vs linear-log form


Log-linear form: log salary on sales

Linear-log form: salary on log sales

# Interpretation of coefficients

| VARIABLES | Linear<br>salary | Log-log<br>lsalary | Log-linear<br>lsalary | Linear-log<br>salary |
|---|---|---|---|---|
| sales | 0.0155* | | 1.50e-05*** | |
| | (0.00891) | | (3.55e-06) | |
| lsales | | 0.257*** | | 262.9*** |
| | | (0.0345) | | (92.36) |
| Constant | 1,174*** | 4.822*** | 6.847*** | -898.9 |
| | (112.8) | (0.288) | (0.0450) | (771.5) |

Linear form: salary increases by 0.155 thousand dollars ($155 dollars) for each additional one million dollars in sales.
Log-log form: salary increases by 0.25% for every 1% increase in sales.
Log-linear form: salary increases by 0.0015% (=0.000015*100) for each additional one million dollar increase in sales.
Linear-log form: salary increases by 2.629 (=262.9/100) thousand dollars for each additional 1% increase in sales.

# Gauss Markov assumptions

- Gauss Markov assumptions are standard assumptions for the linear regression model

1. Linearity in parameters

2. Random sampling

3. No perfect collinearity (or sample variance in the independent variable)

4. Exogeneity or zero conditional mean – regressors are not correlated with the error term

5. Homoscedasticity – variance of error term is constant

# Assumption 1: linearity in parameters

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

- The relationship between *y* and *x* is linear in the population.
- Note that the regression model can have logged variables (e.g. log sales), squared variables (e.g. education$^2$) or interactions of variables (e.g. education*experience) but the $\beta$ parameters are linear.

# Assumption 2: random sampling

$(x_i, y_i)$, where $i$= 1….n

- The data are a random sample drawn from the population.
- Each observation follows the population equation $y = \beta_0 + \beta_1 x + u$
- Data on workers (y=wage, x=education).
- Population is all workers in the U.S. (150 million)
- Sample is workers selected for the study (1,000)
- Drawing randomly from the population – each worker has equal probability of being selected
- For example, if young workers are oversampled, this will not be a random/representative sample.

# Assumption 3: no perfect collinearity

$$SST_x = \sum_{i=1}^{n} (x - \bar{x})^2 > 0$$

- In the simple regression model with one independent variable, there needs to be sample variation in the independent variable (variance of *x* must be positive).
- If there is no variation, the independent variable will be a constant and a separate coefficient cannot be estimated because there is perfect collinearity with the constant in the model.
- Note that $SST_x$ is in the denominator of $\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$

# Assumption 4: zero conditional mean (exogeneity)

$$E(u_i|x_i) = 0$$

- Expected value of error term *u* given independent variable *x* is zero.
- The expected value of the error must not differ based on the values of the independent variable.
- The errors must sum up to zero for each *x*.

# Example of zero conditional mean

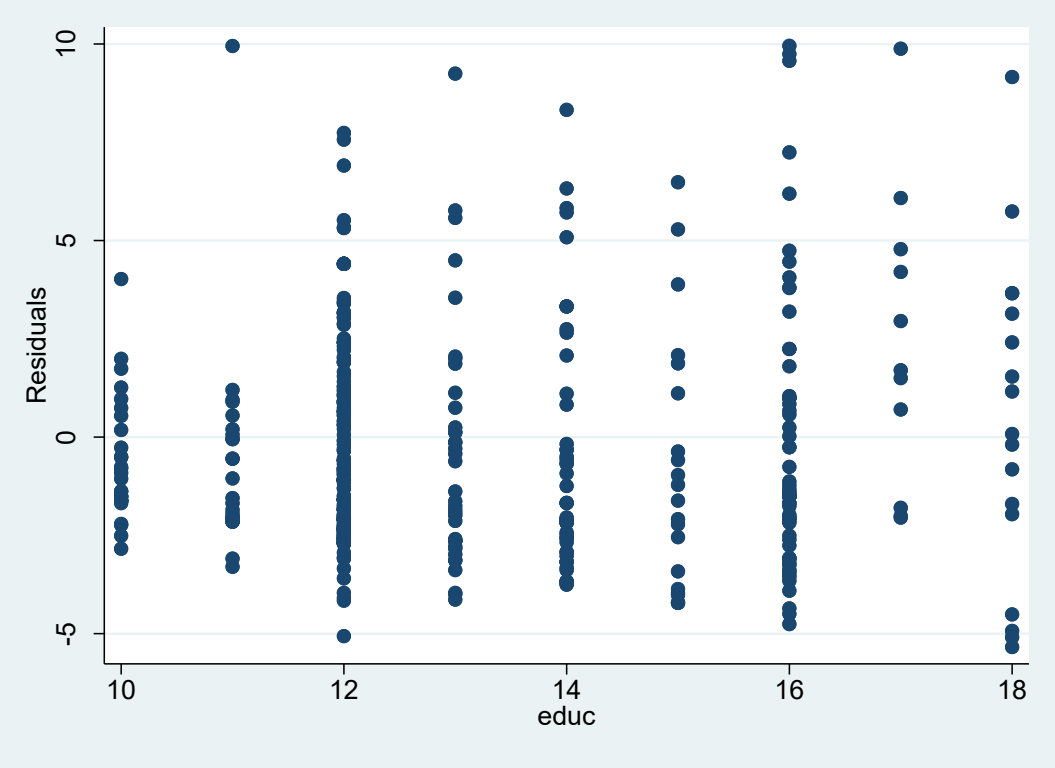
Regression model

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

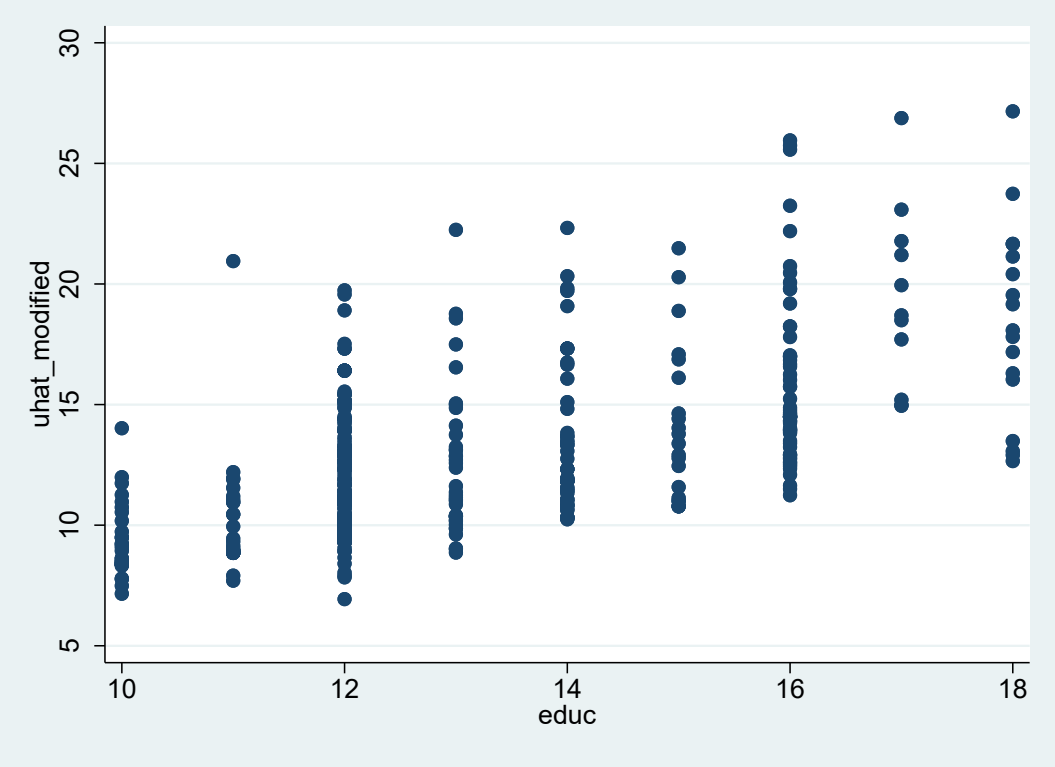
- In the example of wage and education, when ability (which is unobserved and part of the error) is higher, education would also be higher.
- This is a violation of the zero conditional mean assumption.

# Example of exogeneity vs endogeneity

Exogeneity - zero conditional mean




Endogeneity - conditional mean is not zero




E(u|x)=0 error term is the same given education

E(u|x)>0 ability/error is higher when education is higher

# Unbiasedness of the OLS estimators

- Gauss Markov Assumptions 1-4 (linearity, random sampling, no perfect collinearity, and zero conditional mean) lead to the unbiasedness of the OLS estimators.

  $E(\hat{\beta}_0) = \beta_0$ and $E(\hat{\beta}_1) = \beta_1$

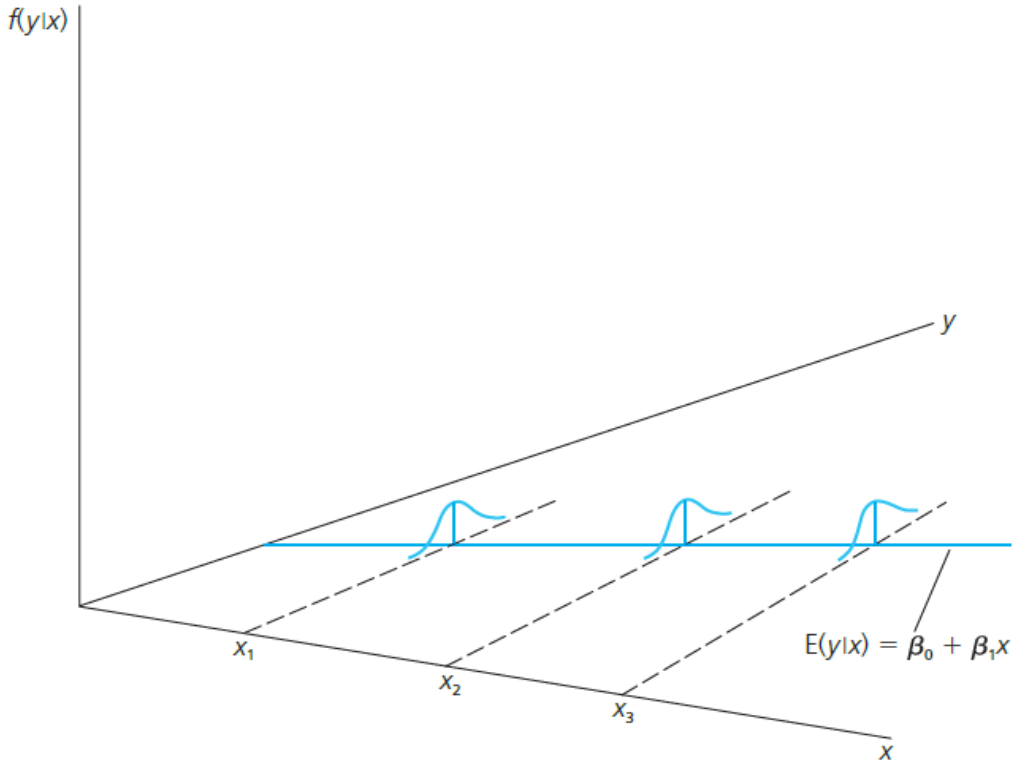
- Expected values of the sample coefficients $\hat{\beta}$ are the population parameters $\beta$.
- If we estimate the regression model with many random samples, the average of these coefficients will be the population parameter.
- For a given sample, the coefficients may be very different from the population parameters.

# Assumption 5: homoscedasticity

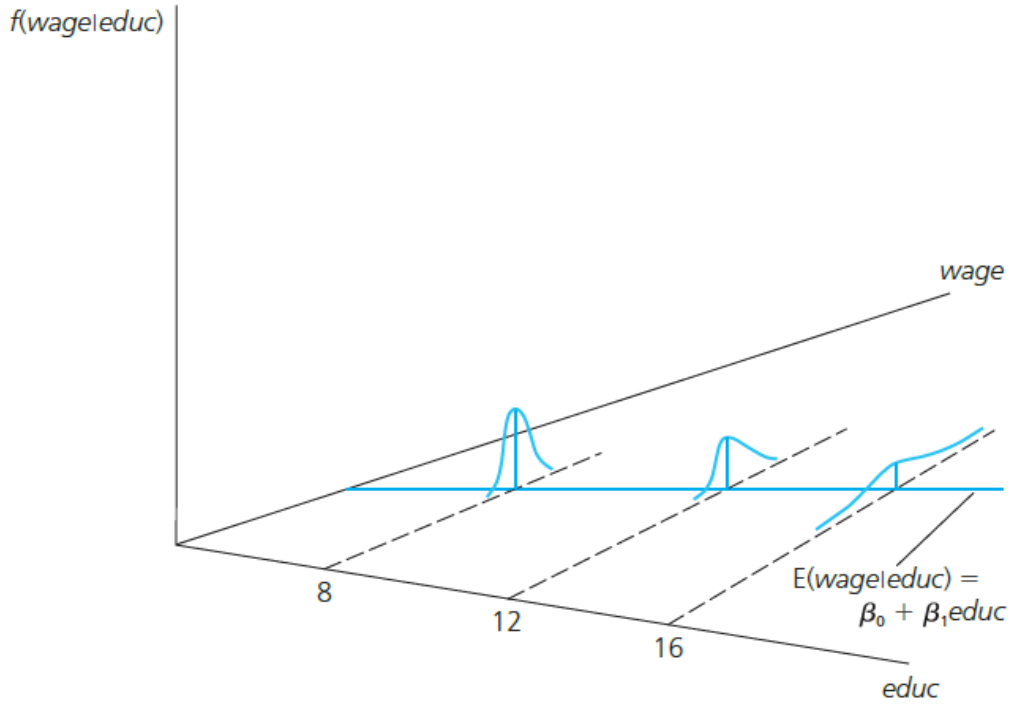
- Homoscedasticity $var(u_i|x_i) = \sigma^2$
- Variance of the error term $u$ must not differ with the independent variable $x$.
- Heteroscedasticity $var(u_i|x_i) \neq \sigma^2$ is when the variance of the error term $u$ is not constant for each $x$.

# Homoscedasticity vs heteroscedasticity

Homoscedasticity
$var(u|x) = \sigma^2$

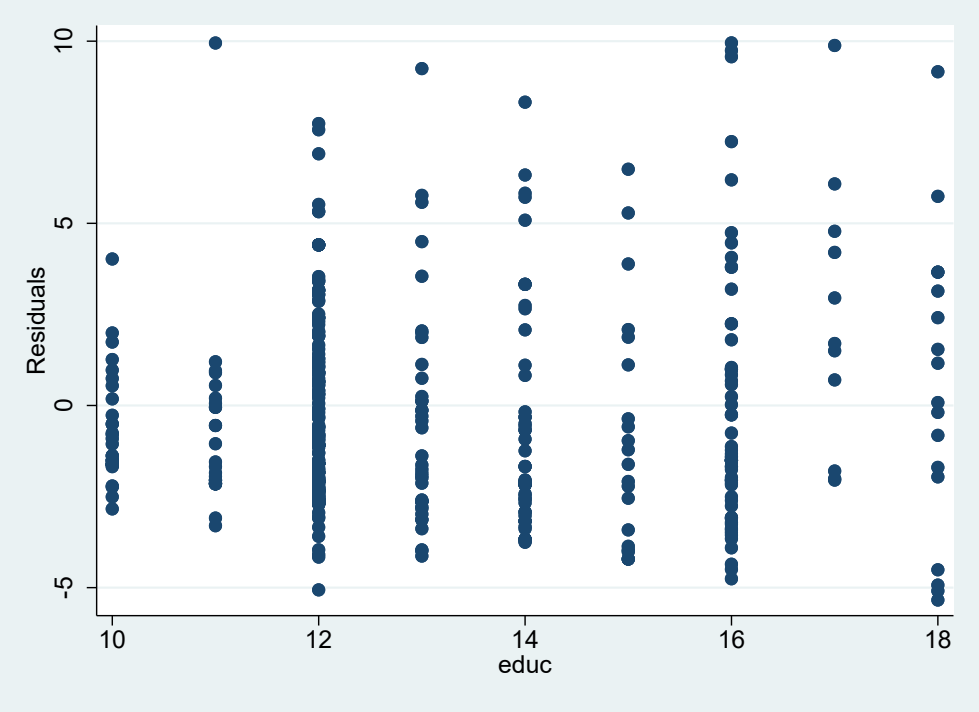



Heteroscedasticity
$var(u|x) \neq \sigma^2$

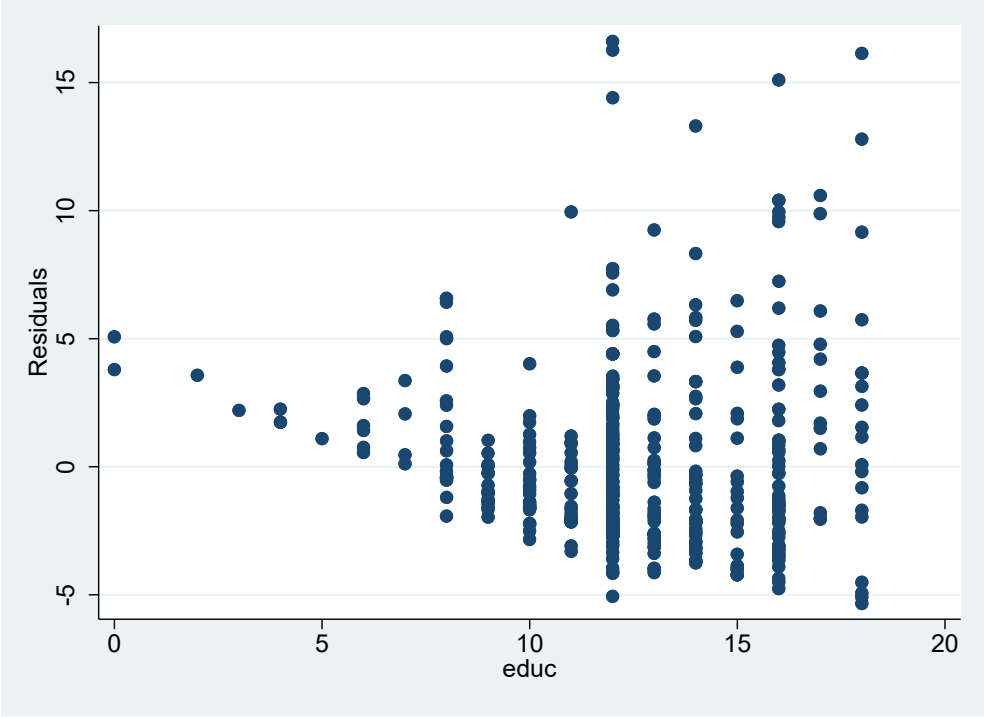


# Homoscedasticity vs heteroscedasticity

Homoscedasticity
$var(u|x) = \sigma^2$




Heteroscedasticity
$var(u|x) \neq \sigma^2$



# Unbiasedness of the error variance

We can estimate the variance of the error term as:

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2$$

- The degrees of freedom (*n*-*k*-1) are corrected for the number of independent variables *k*=1.

- Gauss Markov Assumptions 1-5 (linearity, random sampling, no perfect collinearity, zero conditional mean, and homoscedasticity) lead to the unbiasedness of the error variance.

$$E(\hat{\sigma}^2) = \sigma^2$$

# Variances of the OLS estimators

- The estimated regression coefficients are random, because the sample is random. The coefficients will vary if a different sample is chosen.
- What is the sample variability in these OLS coefficients? How far are the coefficients from the population parameters?

# Variances of the OLS estimators

$$var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma^2}{SST_x}$$

$$var(\hat{\beta}_0) = \frac{\sigma^2 n^{-1} \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n}(x - \bar{x})^2} = \frac{\sigma^2 n^{-1} \sum_{i=1}^{n} x_i^2}{SST_x}$$

The variances are higher if the variance of the error term is higher and if the variance in the independent variable is lower.

Estimators with lower variance are desirable. This means low variance in error term but high variance in the independent variable is desirable.

# Standard errors of the regression coefficients

$$se\left(\hat{\beta}_1\right) = \sqrt{var\left(\hat{\beta}_1\right)} = \sqrt{\frac{\hat{\sigma}^2}{SST_x}}$$

$$se\left(\hat{\beta}_0\right) = \sqrt{var\left(\hat{\beta}_0\right)} = \sqrt{\frac{\hat{\sigma}^2 n^{-1} \sum_{i=1}^{n} x_i^2}{SST_x}}$$

- The standard errors are square root of the variances.
- The unknown population variance of error term $\sigma^2$ is replaced with the sample variance of the residuals $\hat{\sigma}^2$.
- The standard errors measure how precisely the regression coefficients are calculated.