Simple Linear Regression-II

CIVL 7012/8012

Recap(1)

- Relationship between two variables – regression analysis

- Certain names are used for cause and effect

y	x
Dependent variable	Independent variable
Explained variable	Explanatory variable
Response variable	Control variable
Predicted variable	Predictor variable
Regressand	Regressor

Recap (2)

- Using method of moments we obtained
- Coefficient $$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

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

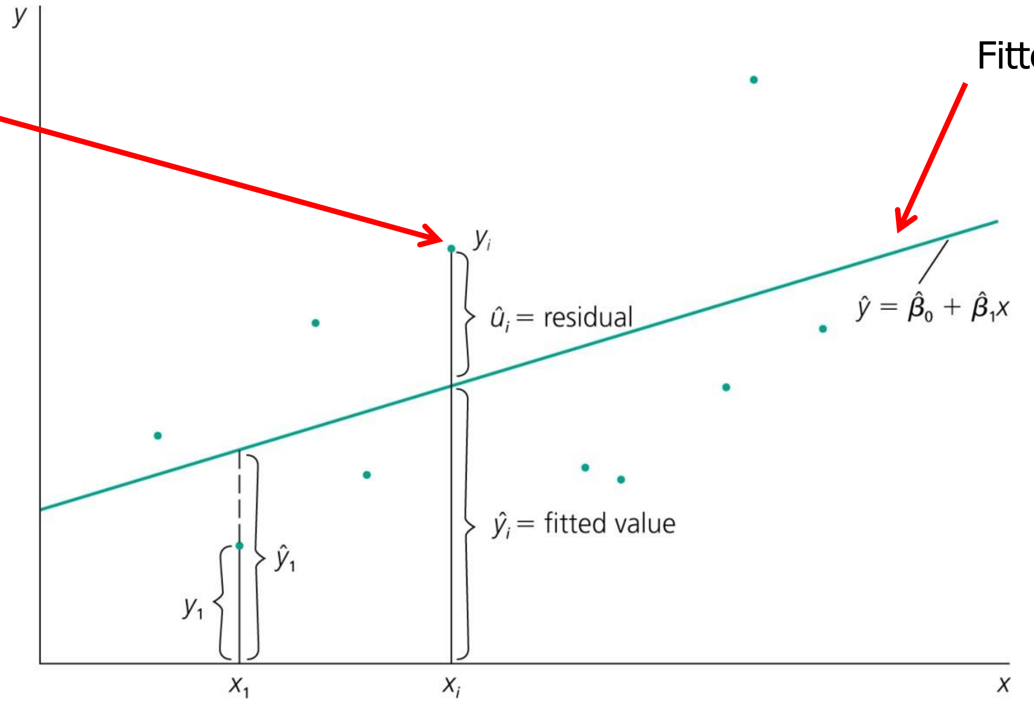
- Intercept $$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Ordinary Least Squares-overview

- Regression-find as good fit as possible
- Resulting in residuals

For example, the i-th data point (x_i, y_i)

Fitted regression line



OLS - estimation

- Regression residuals

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

- Minimize the square of residuals, (named OLS)

$$\min \sum_{i=1}^{n} \widehat{u}_i^2 \quad \rightarrow \quad \widehat{\beta}_0, \widehat{\beta}_1$$

- Parameter estimates (take first order derivative, and equate to zero)

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad \widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$$

Maximum Likelihood -overview

- Conditional p.d.f. of Y for each x

$$p(y|X = x; \beta_0, \beta_1, \sigma^2)$$

- Given any data set

$$\prod_{i=1}^{n} p(y_i|x_i; \beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-(\beta_0+\beta_1 x_i))^2}{2\sigma^2}}$$

- We don't know true parameters, but using estimated parameters, the p.d.f will be

$$\prod_{i=1}^{n} p(y_i|x_i; b_0, b_1, s^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi s^2}} e^{-\frac{(y_i-(b_0+b_1 x_i))^2}{2s^2}}$$

Maximum Likelihood - estimation

- This is called likelihood function

$$\prod_{i=1}^{n} p(y_i|x_i; b_0, b_1, s^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi s^2}} e^{-\frac{(y_i-(b_0+b_1x_i))^2}{2s^2}}$$

- It is much easier to work with log-likelihood
- Max. the log-likelihood to get parameter est.

$$\begin{aligned} L(b_0, b_1, s^2) &= \log \prod_{i=1}^{n} p(y_i|x_i; b_0, b_1, s^2) \\ &= \sum_{i=1}^{n} \log p(y_i|x_i; b_0, b_1, s^2) \\ &= -\frac{n}{2} \log 2\pi - n \log s - \frac{1}{2s^2} \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2 \end{aligned}$$

In-class example

- Let us use an example to solve for parameter estimates
- All three methods should provide us similar estimates

OLS Assumptions

- Let us stick to OLS for the time being
- Fitted values and residuals

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

Fitted or predicted values

Deviations from regression line (= residuals)

- Algebraic properties of OLS (from last class)

$$\sum_{i=1}^{n} \hat{u}_i = 0 \qquad \sum_{i=1}^{n} x_i \hat{u}_i = 0 \qquad \bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

Deviations from regression line sum up to zero

Correlation between deviations and regressors is zero

Sample averages of y and x lie on regression line

Gauss-Markov Assumptions (*1*)

- <u>Standard assumptions for the linear regression model</u>
- **Assumption SLR.1 (Linear in parameters)**

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

In the population, the relationship between y and x is linear

- **Assumption SLR.2 (Random sampling)**

$$\{(x_i, y_i) : \; i = 1, \ldots n\}$$

The data is a random sample drawn from the population

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

Each data point therefore follows the population equation

Gauss-Markov Assumptions (2)

- <u>Assumptions for the linear regression model (cont.)</u>
- **Assumption SLR.3 (Sample variation in explanatory variable)**

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

The values of the explanatory variables are not all the same (otherwise it would be impossible to study how different values of the explanatory variable lead to different values of the dependent variable)

- **Assumption SLR.4 (Zero conditional mean)**

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

The value of the explanatory variable must contain no information about the mean of the unobserved factors

- **Assumption SLR.5 (Homoskedasticity)**

$$Var(u_i|x_i) = \sigma^2$$

The value of the explanatory variable must contain no information about the <u>variability</u> of the unobserved factors

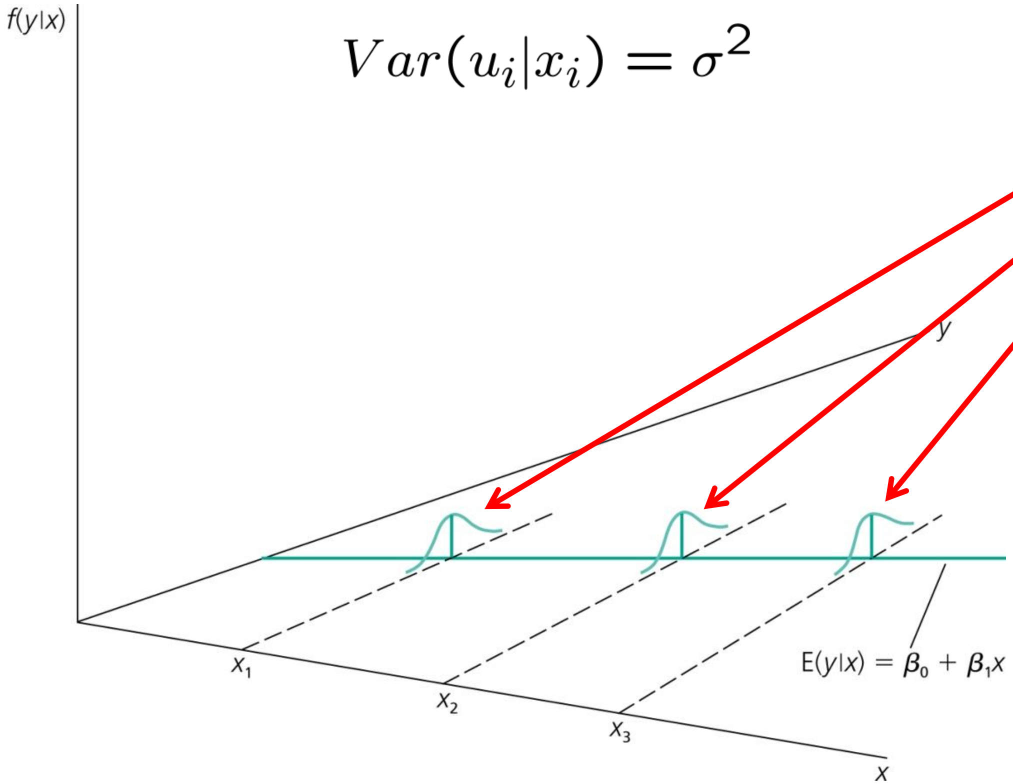
The Gauss-Markov Theorem

- Given our 5 Gauss-Markov Assumptions it can be shown that OLS is "BLUE"
- Best
- Linear
- Unbiased
- Estimator
- Thus, if the assumptions hold, use OLS

More on homoskedasticity

- Graphical illustration of equal variance (homoskedasticity)

$$Var(u_i|x_i) = \sigma^2$$



The variability of the unobserved influences does not dependent on the value of the explanatory variable

Goodness- of- fit: overview

- How well does the explanatory variable explain the dependent variable?

- Measures of variation

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

<u>Total sum of squares</u>, represents total variation in dependent variable

<u>Explained sum of squares</u>, represents variation explained by regression

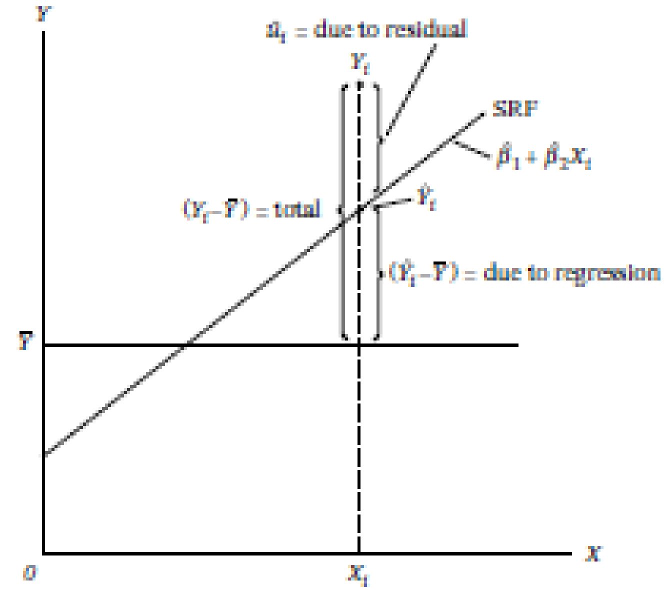
<u>Residual sum of squares</u>, represents variation <u>not</u> explained by regression

Goodness-of-fit : estimation

- Decomposition of total variation

$$SST = SSE + SSR$$

Total variation — Explained part — Unexplained part



- Coefficient of determination

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

R-squared measures the fraction of the total variation that is explained by the regression

Variance of OLS

- Homoskedasticity suggests

$$Var(u_i|x_i) = \sigma^2 = Var(u_i)$$

The variance of u does not depend on x, i.e. is equal to the unconditional variance

- Error variance

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

One could estimate the variance of the errors by calculating the variance of the residuals in the sample; unfortunately this estimate would be biased

- Unbiased estimate of the error variance

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

An unbiased estimate of the error variance can be obtained by substracting the number of estimated regression coefficients
from the number of observations

Variance of OLS

- Homoskedasticity suggests

$$Var(u_i|x_i) = \sigma^2$$

The value of the explanatory variable must contain no information about the variability of the unobserved factors

- Variance of parameter estimates (we omit derivation here)

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

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

Standard errors of the regression coefficients

- Error variance

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

- Unbiased estimate of the error variance

$$se(\hat{\beta}_1) = \sqrt{\widehat{Var}(\hat{\beta}_1)} = \sqrt{\hat{\sigma}^2 / SST_x}$$

Plug in $\hat{\sigma}^2$ for the unknown σ^2

$$se(\hat{\beta}_0) = \sqrt{\widehat{Var}(\hat{\beta}_0)} = \sqrt{\hat{\sigma}^2 n^{-1} \sum_{i=1}^{n} x_i^2 / SST_x}$$

Simple Regression Estimation

Term	Formulae
Coefficient of x	$\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$
Intercept coefficient	$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$
Coefficient of Determination (r^2)	$r^2 = \dfrac{\sum_i(\hat{y}_i - \bar{y})^2}{\sum_i(y_i - \bar{y})^2}$
Standard error (β_1)	$\mathrm{se}(\hat{\beta}_1) = \hat{\sigma} / \left(\sum(x_i - \bar{x})^2\right)^{1/2}$ where, $\hat{\sigma}^2 = \dfrac{\sum_i \widehat{u_i}^2}{n-2}$
Standard error (β_0)	se(β_0)=sqrt ($\dfrac{\hat{\sigma}^2 * \sum_i x_i^2}{n\sum_i(x_i - \bar{x})^2}$)
t-stat	Coefficient/standard error

Example: In class

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.980847369
R Square	0.96206156
Adjusted R Square	0.957319256
Standard Error	6.493003227
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	8552.727273	8552.727273	202.8679245	5.75275E-07
Residual	8	337.2727273	42.15909091		
Total	9	8890			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	24.45454545	6.413817299	3.812791091	0.005142172	9.664256241	39.24483467	9.664256241	39.24483467
x	0.509090909	0.035742806	14.24317115	5.75275E-07	0.42666785	0.591513968	0.42666785	0.591513968

Interpretation of Regression Results

- Y = 24.54 + 0.50 X
- Weekly Expenditure = 24.54 + 0.50 Weekly Income
- Within sample averages of expenditure of 60 families when weekly income (X) increases by $1, the weekly expenditure is expected to increase by $0.50.
- 0.50 is slope of X
- The intercept represents when weekly income is zero, the weekly expenditure is $24.54.

Interpretation of Regression Results (2)

- Often the interpretation of the intercept may not have any viable interpretation.
- Sometimes the intercept is interpreted even when there is no such input data
  - For example in the data we do not have any family weekly income of $0
  - So the interpretation is somewhat limiting.
- Interpretation of coefficient of variable has strong implications.

Example-1

- CEO salary as a function of return on equity
- $\widehat{salary} = 963.191 + 18.501\, roe$
  - Where salary is measured in \$1,000
  - roe is in %
- If the return on equity increases by 1% then the salary is predicted to increase by 18.5 or \$18,500
- If the roe is 0, then the predicted salary is 963.191 or \$963,191

Example-2

- Wage as a function of education
- $\widehat{wage} = -0.90 + 0.54educ$
  - Wage is measured in $/hr
  - Education denote years of schooling
- 1 more year of education increase hourly wage by $0.54/hr (this is not the final model though!)
- Person with no education earns $0.90/hr (not so accurate because the 526 sample has only 8 individuals with no education, so the regression estimate is poor)

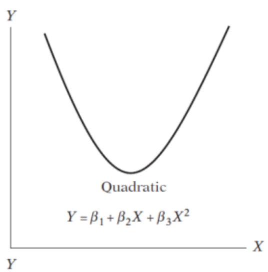
Linearity in parameters

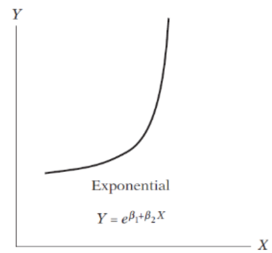
- Relationship between y and x is linear

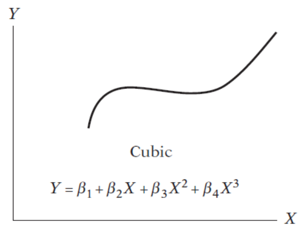
$$y = \beta_0 + \beta_1 x + u$$

In the population, the relationship between y and x is linear

- Other examples of linear in parameters







- Examples of non-linearity in parameters
  - y= 1/ $\beta_0 + \beta_1 x$
  - y = $\beta_0 + \beta_1^2 x$

Log level interpretation

- Interpretation of coefficient for different logarithm forms

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	$\%\Delta y = (100\beta_1)\Delta x$
Log-log	$\log(y)$	$\log(x)$	$\%\Delta y = \beta_1 \%\Delta x$

Semi-log form example (1)

- Regression of log wages on years of eduction

$$\log(wage) = \beta_0 + \beta_1 educ + u$$

Natural logarithm of wage

- This changes the interpretation of the regression coefficient:

$$\beta_1 = \frac{\partial \log(wage)}{\partial educ} = \frac{1}{wage} \cdot \frac{\partial wage}{\partial educ} = \frac{\frac{\partial wage}{wage}}{\partial educ}$$

Percentage change of wage

... if years of education are increased by one year

Semi-log form example (2)

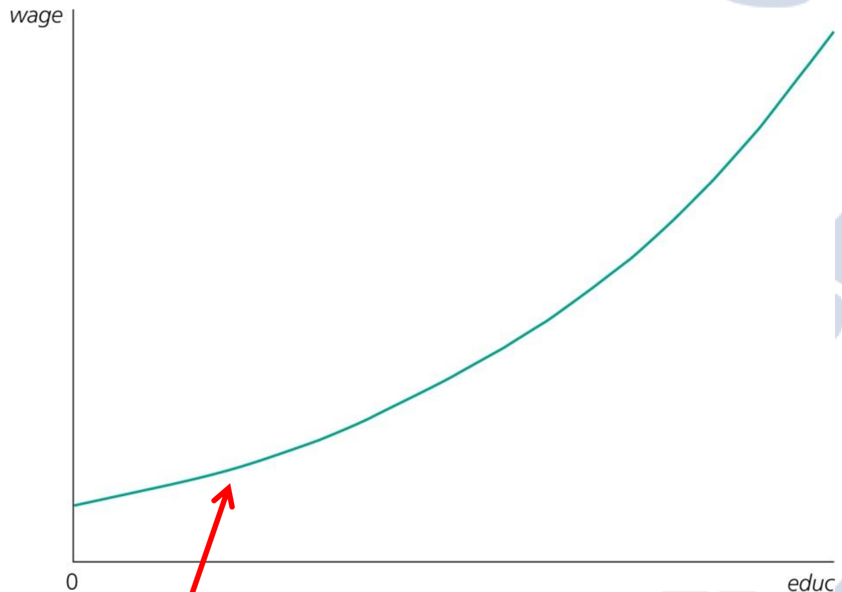
- Regression estimate

$$\widehat{\log}(wage) = 0.584 + 0.083 \ educ$$

The wage increases by 8.3 % for every additional year of education (= return to education)

For example:

$$\frac{\frac{\partial wage}{wage}}{\partial educ} = \frac{\frac{+0.83\$}{10\$}}{+1 \text{ year}} = 0.083 = +8.3\%$$



Growth rate of wage is 8.3 % per year of education

Log-log form example (1)

- CEO salary and firm sales

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + u$$

Natural logarithm of CEO salary

Natural logarithm of his/her firm's sales

- Interpretation of regression coefficient

$$\beta_1 = \frac{\partial \log(salary)}{\partial \log(sales)} = \frac{\frac{\partial salary}{salary}}{\frac{\partial sales}{sales}}$$

Percentage change of salary

... if sales increase by 1 %

Logarithmic changes are always percentage changes

Log-log form example (2)

- Fitted regression equation

$$\widehat{\log}(salary) = 4.822 + 0.257 \log(sales)$$

+ 1 % sales ! + 0.257 % salary

- Coefficient interpretation

$$\frac{\frac{\partial salary}{salary}}{\frac{\partial sales}{sales}} = \frac{\frac{+2{,}570\$}{1{,}000{,}000\$}}{\frac{+10{,}000{,}000\$}{1{,}000{,}000{,}000\$}} = \frac{+0.257\% \text{ salary}}{+1\% \text{ sales}} = 0.257$$