Linear Regression Model (1)

Statistical Research Methods I

Seongsoo Choi (최성수)

This week: Linear Regression Model (1)

- Bivariate linear Regression model
- What is the linear regression model?
- How does it look?
- What is it for? (parameters)
- How could it possible? (assumptions)
- How can we estimate the parameters of the linear regression model?
- Ordinary Least Squares (OLS) estimation
- Statistical inference

What is the linear regression model?

What is the linear regression model?

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

- y and x are two variables, representing some population
- We are interested in "explaining y in terms of x or in "studying how y varies with changes in x"
    - How *hourly wage* is related to *years of education*
    - How *political orientation* is explained by *gender*
    - How *crime rate* varies by *the number of CCTV* across neighborhoods
- y: dependent variable, outcome variable, response variable, etc.
- x: indepedent variable, predictor, explanatory variable, regressor, etc.

What is the linear regression model?

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

- The linear regression model is a model in which we describe how y is related to x
- The key assumption of the model is: y and x are linearly related
  - Their relationship is summarized by one parameter, β_1, which describes their linear relationship (slope in other words)
  - We can easily extend this simple regression model to a multivariate one simply by adding other independent variables (x_2, x_3,…) *additively*
  - Note that the slope, β_1 (or β_k), is a (partial) derivative of y with regard to x (or x_k)

Modeling is like map-making

Two parameters characterizing a linear regression model

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

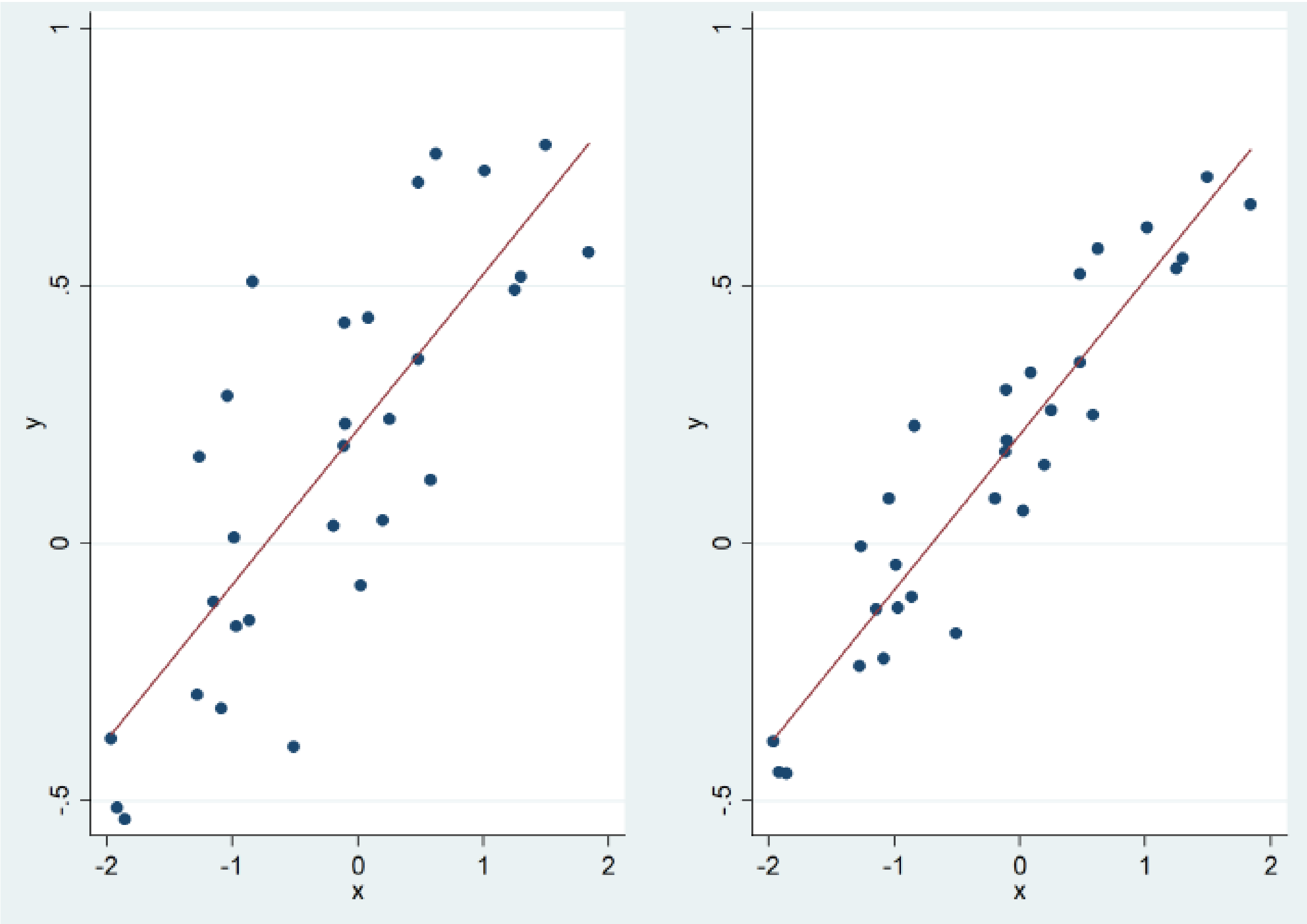
- β_1 (a linear slope) and β_0 (an intercept): how much units of y increases when one unit of x increases
- ε_i: Part of y_i that is not explained by x_i (called *error* or *error term*)
    - The relationship between x and y is not deterministic but probabilistic
- The linear regression model breaks y_i into two parts:
    - explained by x: $\hat{y}_i = \beta_o + \beta_1 x_i$
    - unexplained by x: $\varepsilon_i = y_i - \beta_0 - \beta_1 x_i$

Two parameters characterizing a linear regression model

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

- In the linear regression model, there are two parameters of our interest (what we want to know)
  1. the relationship between y and x: β_1 and β_0 (or more simply, β as a vector)
  2. how much of y remains unexplained by x: $\sigma_\varepsilon^2 = Var(\varepsilon)$
- Once we know these two parameters, we can reproduce the linear regression model

$$y_i = 0.2 + 0.3x_i + \varepsilon_i$$



Key assumptions in the linear regression model

- Linearity: essential model assumption about x and y
- Zero conditional mean: $E(\varepsilon) = E(\varepsilon|x) = 0$
  - We need to make assumptions about the unknown part ε
  - Its mean is zero (so we can get rid of it easily by taking expectation)
  - Its mean conditional on x is zero
  - This is also called *conditional independence*
    - x and ε are independent of each other ($x \perp \varepsilon$)
    - Since $\varepsilon_i = y_i - E(y_i|x)$ by design, this assumption means that x covers all the systematic part of y and the remaining part, ε, is just pure random

Estimation of the parameters

Ordinary Least Squares (OLS) Estimation

- There can be innumerable slopes (that is, β) we can draw, but what is the best one?
- OLS estimator of β (or $\hat{\beta}_{ols}$ or more simply $\hat{\beta}$) is the slope that minimizes the total sum of error terms (in other words, the difference between $\hat{y}_i$ and actual y_i)

$$\min \hat{\beta}_0, \hat{\beta}_1 \sum_{i=1}^{N} \varepsilon_i^2 = \min \hat{\beta}_0, \hat{\beta}_1 \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i)^2$$

Ordinary Least Squares (OLS) Estimation

- The OLS estimators of β_1 and β_0 are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} = \frac{Cov(x,y)}{Var(x)}$$

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

- Note that:

$$\hat{\beta}_1 = \frac{Cov(x,y)}{Var(x)} = Corr(x,y)\frac{SD(y)}{SD(x)}$$

Ordinary Least Squares (OLS) Estimation

- The OLS estimator of σ_ε^2: the variance of *residual*
  - Residual: $e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i = \hat{\varepsilon}_i$
  - Residual sum of squares (RSS) is the unbiased estimator of σ_ε^2

$$RSS = \hat{\sigma}_\varepsilon^2 = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

- Mean square of Error (MSE)

$$M\hat{S}E = \frac{RSS}{N-k-1} = \frac{RSS}{df}$$

- Stata reports the estimated RMSE ($\sqrt{M\hat{S}E}$)

Stata outcome table

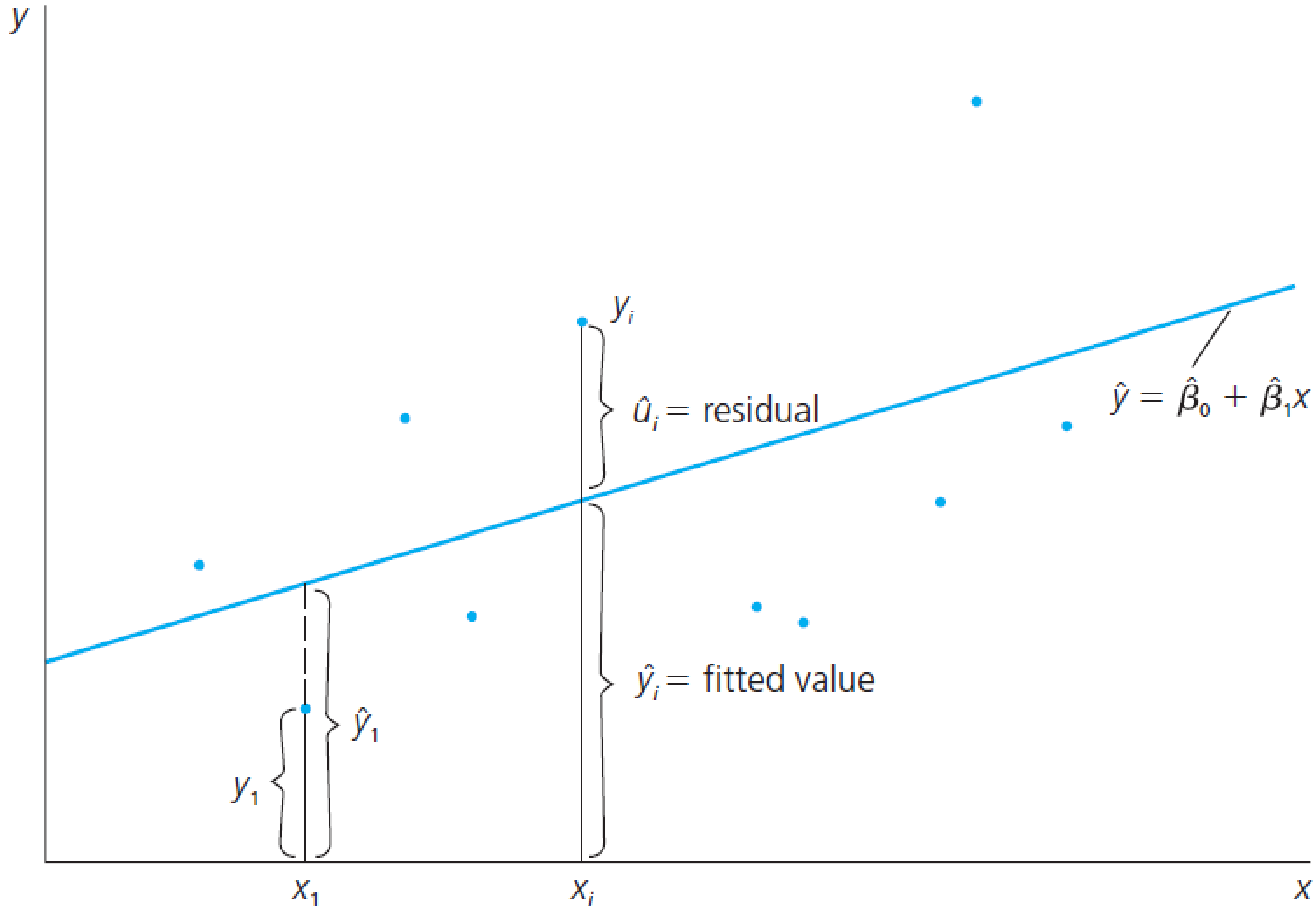
```
. reg y x

      Source |       SS           df       MS      Number of obs   =        30
-------------+----------------------------------   F(1, 28)        =     49.83
       Model |  2.84623978         1  2.84623978   Prob > F        =    0.0000
    Residual |  1.59938664        28  .057120951   R-squared       =    0.6402
-------------+----------------------------------   Adj R-squared   =    0.6274
       Total |  4.44562642        29  .153297463   Root MSE        =      .239

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .3017595   .0427487     7.06   0.000     .2141928    .3893263
       _cons |   .2218985   .0446451     4.97   0.000     .1304472    .3133499
------------------------------------------------------------------------------
```

R-squared as model's goodness of fit statistic

- R-squared (R^2) shows how much our model explains the variance of the dependent variable y_i
  - This is an intuitive indicator of the estimated error variance, $\hat{\sigma}^2_\varepsilon = Var(e)$
  - The proportion of the explained sum of squares (ESS or Model SS) out of the total sum of squares (TSS) of y_i
    - in other words, it is the proportion of $Var(\hat{y}_i)$ out of $Var(y_i)$

$$R^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} = 1 - \frac{RSS}{TSS}$$

- R^2 falls between 0 and 1



y
y_i
$\hat{u}_i$ = residual
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
$\hat{y}_i$ = fitted value
$\hat{y}_1$
y_1
x_1
x_i
x

R-squared

- R-squared: the proportion of variance in the dependent variable, y, that is explained by our model (x)
  - e.g., $R^2 = 0.3$ means that our model (or independent variables) explains 30% of variation in y
- In social sciences, R^2 is often very low (e.g., $R^2 = 0.03$)
- In a bivariate linear regression, $R^2 = (corr(x, y))^2$, but in multivariate regression, they become apart

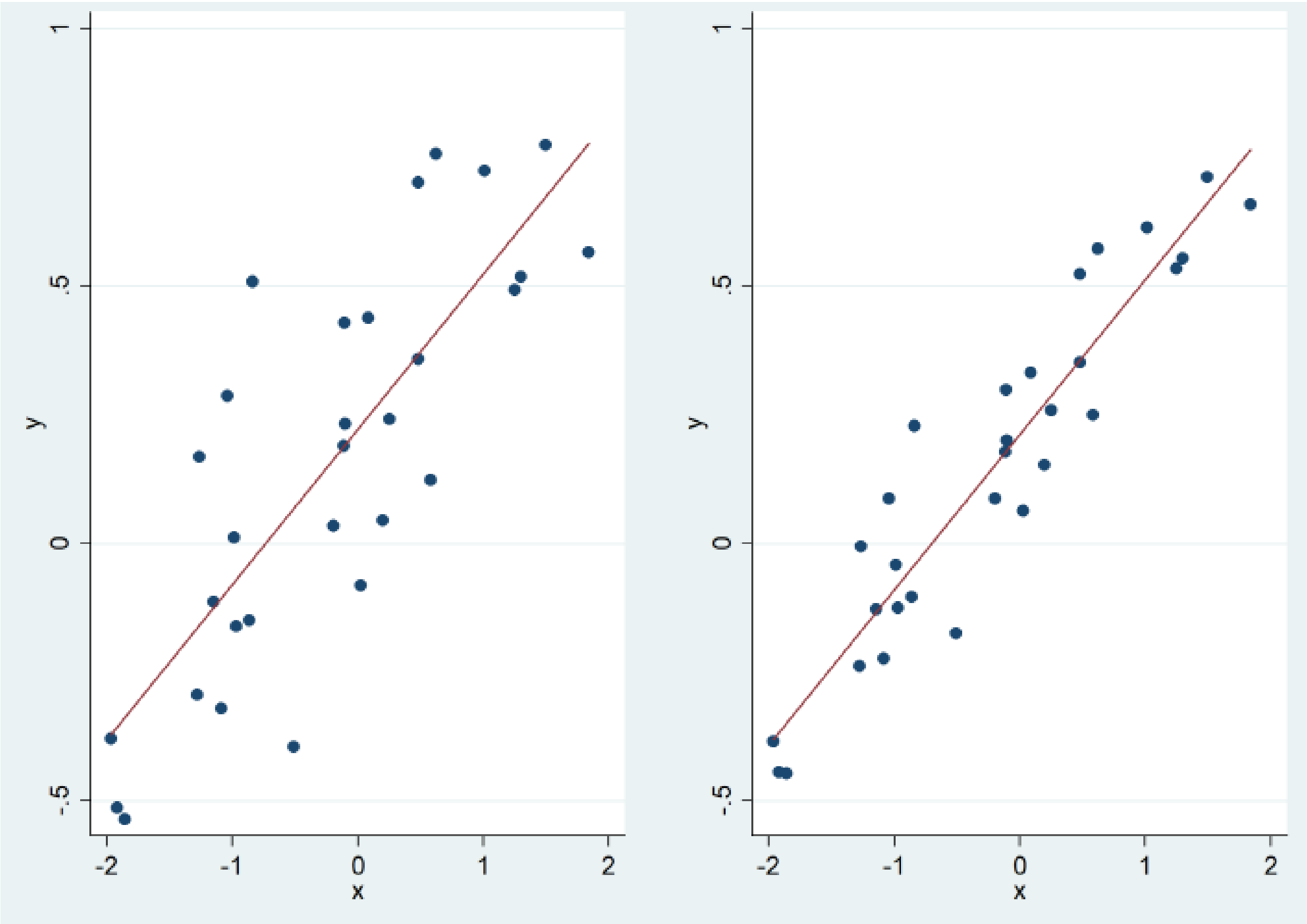
R-squred

- We can also think this way; in $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i$
  - $Var(y_i) = \hat{\beta_1}^2 Var(x_i) + Var(e_i)$
  - where $\hat{\beta_1}^2 Var(x_i)$ captures "between x variation" and $Var(e_i)$ "within x variation"
- This is a basic concept of "Analysis of Variance" (ANOVA) (especially when x is a categorical variable)
  - whether wage inequality lies between education groups or within education groups
  - whether math score gap emerges mainly between boys and girls or within each of them

Two parameter estimates of our interest in the OLS model estimation

- $\hat{\beta}$: explained (model) part; the relationship between y and x
- R^2: unexplained part: how much of y remains unexplained (more precisely, 1-R^2); model fit
- $\hat{\beta}$ and R^2 are related to different, separate interests
    - e.g., SES gradient in education or educational reproduction?
        - the relationship between educational achievement and parents' SES measure
        - how much of variation in educational achievement is explained by parents' SES measure

$$y_i = 0.2 + 0.3x_i + \varepsilon_i$$



Stata outcome table

```
. reg y x
```

Source	SS	df	MS
Model	2.84623978	1	2.84623978
Residual	1.59938664	28	.057120951
Total	4.44562642	29	.153297463

Number of obs	=	30
F(1, 28)	=	49.83
Prob > F	=	0.0000
R-squared	=	0.6402
Adj R-squared	=	0.6274
Root MSE	=	.239

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x	.3017595	.0427487	7.06	0.000	.2141928	.3893263
_cons	.2218985	.0446451	4.97	0.000	.1304472	.3133499

Exercise

- Let's think of a linear regression model in which we regress change in math score (y) on change in reading score (x)
- In other words, we predict change in math score using change in reading score

	Change in math score	Change in reading score
Chris	1	3
Jamal	-2	2
Jieun	3	4
Yingyao	0	6
Sho	3	0

Exercise

	Change in math score	Change in reading score
Chris	1	3
Jamal	-2	2
Jieun	3	4
Yingyao	0	6
Sho	3	0

- Build a linear regression model
- Estimate all the parameters of the model
- Draw a scatterplot with the fitted line

Statistical Inference of $\hat{\beta}$

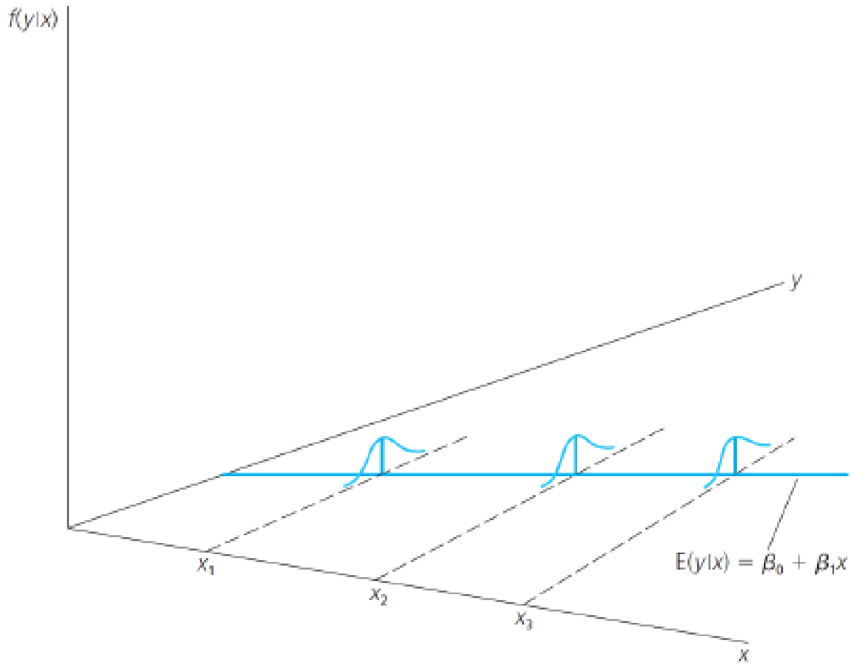
Statistical Inference of $\hat{\beta}$

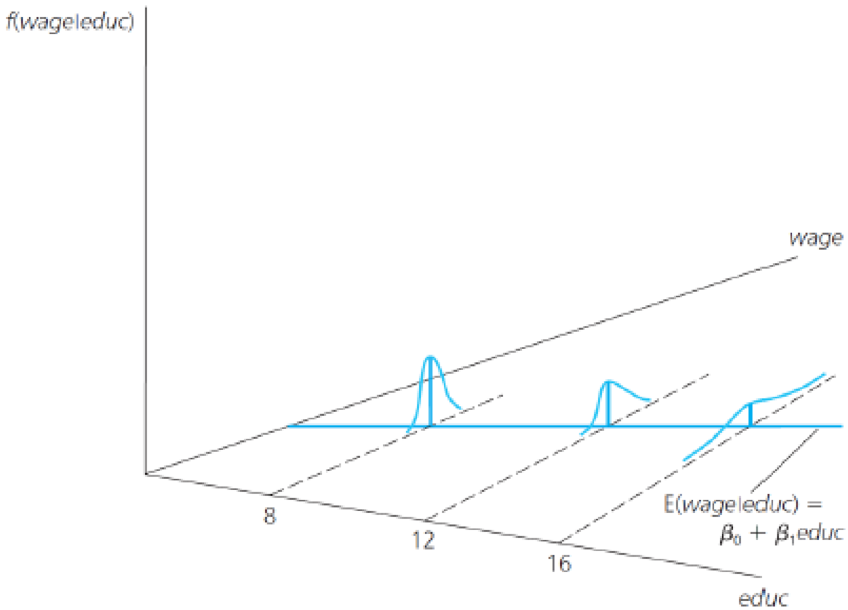
- We want to learn whether our estimate, $\hat{\beta}$ is a chance product or reflects a statistical tendency generalizable to the population
- Two ways (as we learned previously)
  - to construct the CI (e.g., 95%)
  - to use p-value
- For both ways, the key statistic we need is the *standard error (SE)* of $\hat{\beta}$
- Two things we need to be sure about
  - Whether $\hat{\beta}$ has a normal distribution (e.g., Central Limit Theorem)
  - How can we get $SE(\hat{\beta})$

Whether $\hat{\beta}$ has a normal distribution

- Yes, we can apply the CLT to $\hat{\beta}$
  - Using jargon, we can say "$\hat{\beta}$ is asymptotically normal" (normal when N is large enough, e.g., $N > 30$)
  - $\hat{\beta} \sim N(\beta, Var(\hat{\beta}))$ where $Var(\hat{\beta}) = SE(\hat{\beta})^2$
- How can we get $SE(\hat{\beta})$? Under the assumption of *homoskedasticity*,
  - $Var(\hat{\beta}) \equiv \hat{SE}(\hat{\beta})^2 = \frac{Var(e)}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$
  - where $Var(e) = \frac{\sum_{i=1}^{N} e_i^2}{(N-k-1)} = \frac{RSS}{(N-k-1)}$, So $\hat{SE}(\hat{\beta}) = \sqrt{\frac{Var(e)}{\sum_{i=1}^{N}(x_i - \bar{x})^2}}$
- Note that $\hat{SE}(\hat{\beta})$ decreases as N increases

Homoskedasticity and heteroskedasticity





- The assumption of homoskedasticity (등분산성) affects our estimate of $Var(e)$ and $\hat{SE}(\hat{\beta})$ (but not $\hat{\beta}$)

Homoskedasticity and heteroskedasticity

- The default setting for STATA is to assume homoskedasticity but easily can relax it
- The rule of thumb suggested by Angrist and Pischke (2009) is to choose the conservative one:
  - A larger $\widehat{SE}(\hat{\beta})$ between the one under the homoskedasticity assumption and the one without assuming it

Homoskedasticity

```
. reg y x

      Source |       SS           df       MS      Number of obs   =        30
-------------+----------------------------------   F(1, 28)        =     49.83
       Model |  2.84623978         1  2.84623978   Prob > F        =    0.0000
    Residual |  1.59938664        28  .057120951   R-squared       =    0.6402
-------------+----------------------------------   Adj R-squared   =    0.6274
       Total |  4.44562642        29  .153297463   Root MSE        =      .239

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .3017595   .0427487     7.06   0.000     .2141928    .3893263
       _cons |   .2218985   .0446451     4.97   0.000     .1304472    .3133499
------------------------------------------------------------------------------
```

Heteroskedasticity

```
. reg y x, vce(robust)

Linear regression                               Number of obs =         30
                                                F(  1,     28) =     74.20
                                                Prob > F      =    0.0000
                                                R-squared     =    0.6402
                                                Root MSE      =      .239

------------------------------------------------------------------------------
             |               Robust
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .3017595   .0350314     8.61   0.000     .2300011     .373518
       _cons |   .2218985   .0434006     5.11   0.000     .1329963    .3108007
------------------------------------------------------------------------------
```

Statistical inference: using p-value

- Once we get $\hat{SE}(\hat{\beta})$, next steps are quite straightforward:
  - Compute the test statistic (t or z-statistic):

$$z = \frac{\hat{\beta} - \beta}{\hat{SE}(\hat{\beta})}$$

- The true β is set based on our hypothesis:
  - Usually, $H_0 : \beta = 0$ (No relationship between x and y; when x increases by one unit, we find no statistical evidence that y also changes systematically)
  - In that case, $z = \frac{\hat{\beta}}{\hat{SE}(\hat{\beta})}$, and go to the z(or t) table to get the p-value that correponds to the z value

Confidence intervals

- $\hat{\beta} \pm \mathbf{1.96} \times \hat{SE}(\hat{\beta})$ for 95% CI
  - If we conduct the same analysis repeatedly with different samples that are randomly sampled in the exactly same way from our population many times,
  - the true β will fall into this CI with a 95% of probability
  - or our CI will miss the true β with a 5 in 100 chance
- $\hat{\beta} \pm \mathbf{2.58} \times \hat{SE}(\hat{\beta})$ for 99% CI

Regression result table

```
. reg y x
```

Source	SS	df	MS
Model	2.84623978	1	2.84623978
Residual	1.59938664	28	.057120951
Total	4.44562642	29	.153297463

Number of obs	=	30
F(1, 28)	=	49.83
Prob > F	=	0.0000
R-squared	=	0.6402
Adj R-squared	=	0.6274
Root MSE	=	.239

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x	.3017595	.0427487	7.06	0.000	.2141928	.3893263
_cons	.2218985	.0446451	4.97	0.000	.1304472	.3133499

Exercise

	Change in math score	Change in reading score
Chris	1	3
Jamal	-2	2
Jieun	3	4
Yingyao	0	6
Sho	3	0

- Under homoskedasticity, get the standard error of $\hat{\beta}$, t, and p-value
- Report 95% CI
- What can we say about the statistical inference of our estimate $\hat{\beta}$?

Several important cautions about statistical significance

- Statistical significance is not practical (or substantive) significance
  - Any small differences can be statistically significant if N is very large
  - Statistical significance doesn't guarantee practical significance of a finding
- Statistically insignificant $\hat{\beta}$ doesn't mean no relationship between x and y
  - It just tells us that we can't be certain enough about the presence of a systematic relationship between x and y
  - Statistical significance (p-value) is a continuous measure, not a all-or-nothing measure ("The difference between statistical significance and insignificance is statistically insignificant")

Useful Stata commands

- regress y x
- predict
  - without an option: it generates a new variable for the predicted value
    - $\hat{y}_i = \hat{\beta} x_i$
  - with the option (, residual): it generate a new variable for the residual
    - $y_i - \hat{y}_i = y_i - \hat{\beta} x_i = e_i$

Next

- Let's extend the bivariate linear regression model to the multivariate one
- How can we link our theory to a linear regression model?
  - Modeling strategy: building nesting and nested models
  - Confounders and mediators
- How can we incorporate categorical independent variables? (answer: dummy variables)
- How can we incorporate nonlinear relationships between x and y?
- How can we incorporate interactions between two independent variables?