ECO220Y

Regression Analysis:

Introduction and Assumptions

Readings: Chapters 7-8 (Review) and Chapter 18,18.1-18.2

Winter 2012

Lecture 18

Economic Questions

- How much does beauty pay? (see D.Hamermesh "Beauty Pays")
- What is the perfect salary for happiness? (see A.Deaton and D.Kahneman 2010)
- Are there peer effects in binge drinking?
- Who benefits from Universal Child Care in Canada?

Regression Analysis: Preview

Regression analysis is used:

1. To quantify the linear relationship among variables
2. To build a model of economic behavior so that we can conduct what-if analysis
3. For forecasting

Least Squared method is used:

1. To describe data: summarize the linear relationship among variables (Lecture 4)
   - OLS can always be used as a descriptive statistic
   - But cannot infer causality based on observational data
2. To estimate parameters of a model used for what-if analysis and predictions

Review of Least Squares Method (OLS)

- OLS fits the line through the scatter plot minimizing sum of squared errors, ε_i

$$\hat{y} = b_0 + b_1 x$$

$$b_1 = ?$$

$$b_0 = ?$$

- The slope of the line tells us about the direction of the linear association between two variables
- In standardized OLS line, the slope also tells us about the strength of the linear association

$$\hat{z}_y = r z_x$$

Analysis of Variance, or ANOVA

$$SST = SSE + SSR$$

$$SST = \sum(y_i - \bar{y})^2$$

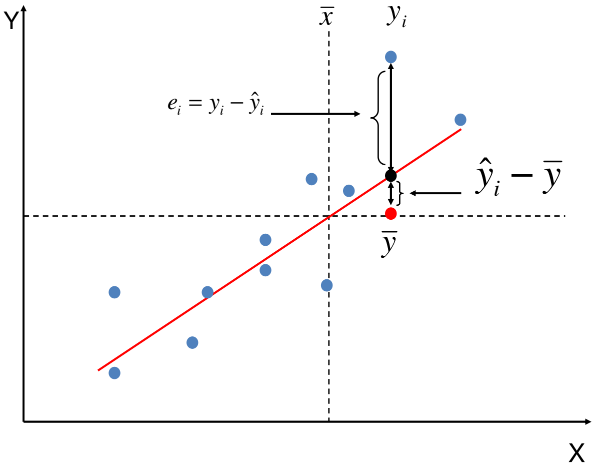
$$SSE = \sum(y_i - \hat{y}_i)^2$$

$$SSR = \sum(\hat{y}_i - \bar{y})^2$$

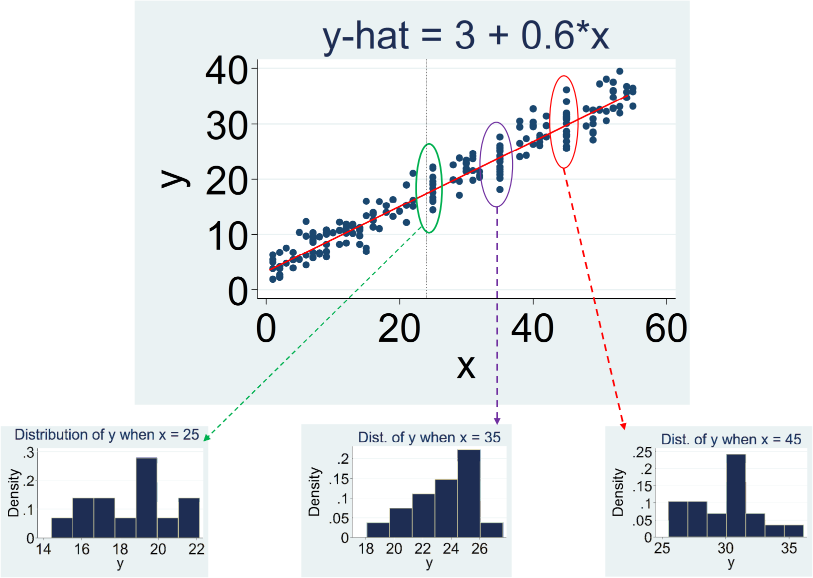
$$\frac{SST}{SST} = \frac{SSE}{SST} + \frac{SSR}{SST}$$

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

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



R^2 measures how well the line fits the data



y-hat = 3 + 0.6*x
y
40
30
20
10
0
0
20
40
60
x
Distribution of y when x = 25
Density
.3
.2
.1
0
14
16
18
20
22
y
Dist. of y when x = 35
Density
.2
.15
.1
.05
0
18
20
22
24
26
y
Dist. of y when x = 45
Density
.25
.2
.15
.1
.05
0
25
30
35
y

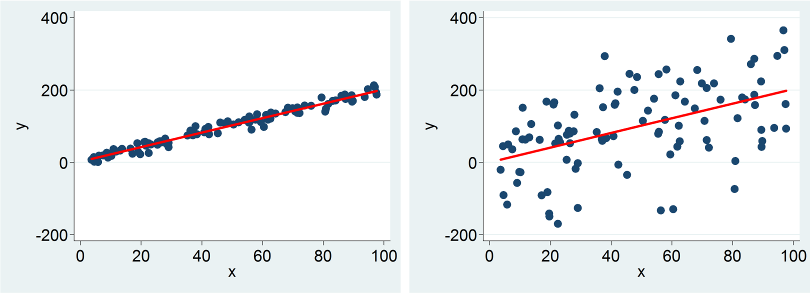
Population Regression Function

Observation Index

$$\underbrace{y_i}_{\text{Dependent Variable}} = \underbrace{\beta_0}_{\text{Constant Term}} + \underbrace{\beta_1}_{\text{Slope Coefficient}} \cdot \underbrace{x_i}_{\text{Independent Variable}} + \underbrace{\varepsilon_i}_{\text{Error Term}}$$

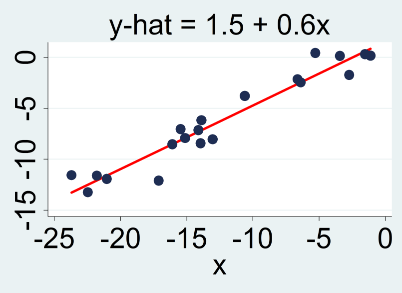
Residual, or error term, ε

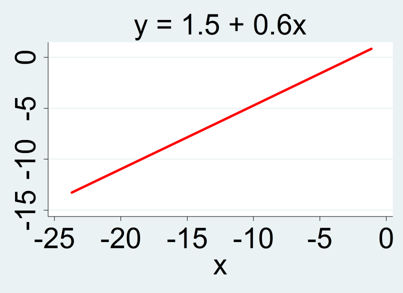
- Residual, or error term: remainder, what is left over
- ε_i picks up all unobserved factors
- Error term (ε) includes all other factors that affect y aside from x_i
  - for practical reasons, it is impossible to collect data on everything
  - reflects reality: model cannot control for everything
  - $\varepsilon_i = y_i - \beta_0 - \beta_1 x_i$, or $\varepsilon_i = y_i - \hat{y}_i$



Error Term is Important

- Probabilistic model: Contains an unobservable term ε, which makes only probabilistic statements possible:
  - $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
  - Our goal is to estimate parameters and determine the precision of estimate
- Deterministic model: All terms are observed, which means no uncertainty





Ordinary Least Squares

- OLS is a method for estimating β_0 and β_1
- OLS returns estimates: b_0 and b_1
- Can be shown that $E[b_0] = \beta_0$ and $E[b_1] = \beta_1$
- OLS produces unbiased, consistent and relatively efficient estimates

Compare relationship between β_1 and b_1 and μ and $\bar{X}$. What is the counterpart of ε in OLS?

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

$$y_i = b_0 + b_1 x_i + e_i$$

$$e_i = ?$$

Assumptions

- Six assumptions underlie the linear regression model:
  - We cannot interpret results of regression analysis without knowing underlying assumptions
- Econometrics addresses violations of the underlying assumptions:
  - ECO374H Applied Econometrics (for Commerce)
  - ECO375H Applied Econometrics I and
  - ECO375H Applied Econometrics II
- ECO220 reviews the assumptions informally, mostly relying on graphical techniques to detect the violations of the assumptions

Assumptions

- Model is linear in parameters and errors
- $E[\varepsilon] = 0$
- $V[\varepsilon_i] = \sigma^2$, or homoscedasticity
- $Cov[\varepsilon_i, \varepsilon_j] = 0$, or no autocorrelation
- $Cov[x_i, \varepsilon_i] = 0$, or exogeneity
- $\varepsilon \sim N(0, \sigma^2)$

Assumption # 1

Functional form of regression equation is linear in the error and parameters:

$$\boxed{} = \beta_0 + \beta_1 \boxed{} + \varepsilon_i$$

Note: what is in red boxes does not have to be linear, i.e. can be any function of y and/or x. For instance, $ln(y)$ or x^2.

$$y_i = \beta_0 + \beta_1 x^2 + \varepsilon_i$$

$$ln(y_i) = \beta_0 + \beta_1 x + \varepsilon_i$$

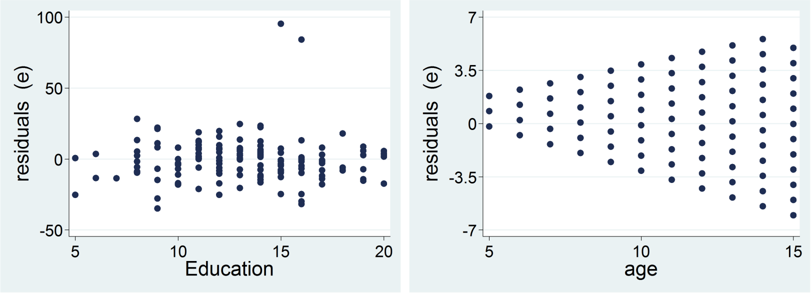
$$ln(y_i) = \beta_0 + \beta_1 x^2 + \varepsilon_i$$

Assumption # 2

- Error has mean zero: $E[\varepsilon_i] = 0$ for $i = 1, 2, ..., n$
- Constant term, β_0 picks up any systematic, constant effects
- Recall that ε_i captures all other factors that we do not observe. We call them random or white noise
  - Some errors are positive, some are negative, but on average these are zero
  - If we include a constant term, it should soak up all the systematic differences between observations and leave the random one
  - Zero mean of a random component also means that $\hat{y} = b_0 + b_1 x$

Assumption # 3

- Homoscedasticity: $V[\varepsilon_i] = \sigma^2$ for all $i = 1, 2, ..., n$
- We expect that the noise is just as "noisy" for all our data
- When heteroscedasticity exists, it affects the standard errors of parameters estimates and, hence, the inference
- Can test for the presence of heteroscedasticity:
  - Formally, using Breusch-Pagan test
  - Informally, check the pattern of the scatter plot of residuals against predicted values or against x



Assumption # 4

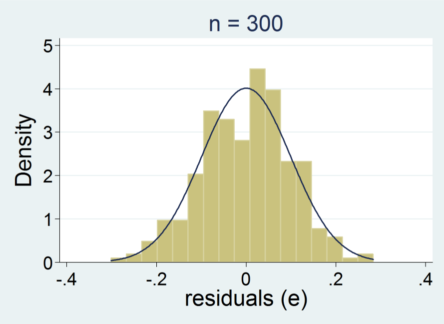
- No autocorrelation/no serial correlation
- $COV[\varepsilon_i, \varepsilon_j] = 0$ if $i \neq j$
- Also can be stated as $E[\varepsilon_i \varepsilon_j] = 0$
- Problem for time-series data
- Plot residuals against time variable to see whether there is a pattern, or trend in the residuals

Assumption # 5

- $COV[X, \varepsilon] = 0$ or $E[X, \varepsilon] = 0$
- This assumption is also referred to as exogeneity assumption
- As a rule, assumption # 5 is violated with observational data
- Violation of Assumption #5 $\Rightarrow E[b_1] \neq \beta_1$
- When assumption #5 is violated, endogeneity exists
- Advantage of experimental data - exogeneity assumption holds

Assumption # 6

- Assumption # 6 is often referred to as Normality assumption
- ε_i is normally distributed



-
- Together, Assumptions #2, #3 and #6 imply that $\varepsilon_i \sim N(0, \sigma^2)$
- Note that Assumptions #2-#6 are all about the unobserved error term