

Mathematics 231

Lecture 8
Liam O'Brien

1

Announcements

- Reading
 - Today M&M 2.3 117-119
 - M&M 2.4 125-132
 - Next class M&M 2.5 142-151

2

Evaluating the Regression Model

- Checking Assumptions
- Residual Plots
- R^2 and Correlation

3

Assumptions

- The regression line estimates the conditional mean of Y given $X=x$ for any point x if the following assumptions are met.
 1. Conditional mean of Y is a linear function of X .
 2. Conditional SD of Y is constant for all X .
- We often make an additional assumption:
 3. The conditional distribution of Y is a normal distribution for any value of x .

4

Checking Assumptions

■ Model and Residuals

$$\text{Data} = \text{Predicted Values} + \text{Residuals}$$

(Pattern) (Deviation)

- **Predicted values:** the part of the data that is explained by the regression model.
- **Residuals:** the part of the data that is not explained by the regression model.

5

Checking Assumptions

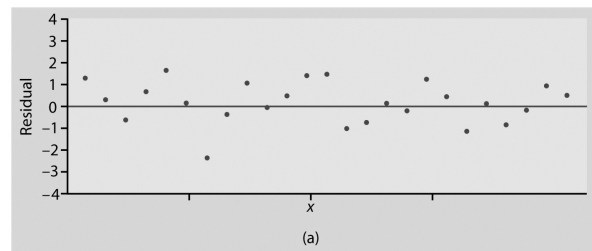
- In linear regression,
 - Fitted line represents the pattern.
 - Residuals represent deviations from the pattern.
- Examine the scatter plot of the original data with the regression line superimposed on it.
- Do you see any marked deviations from the line?

6

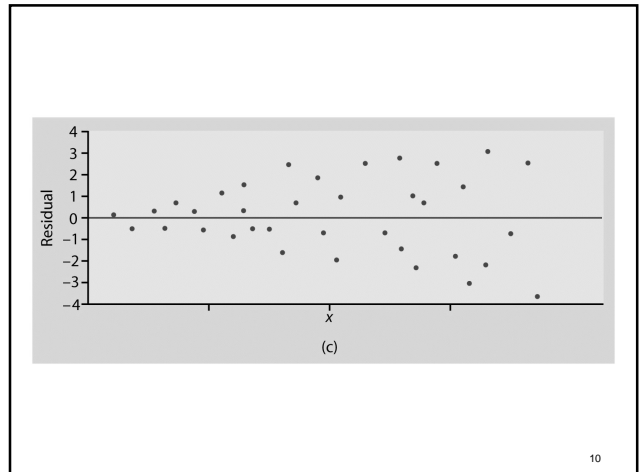
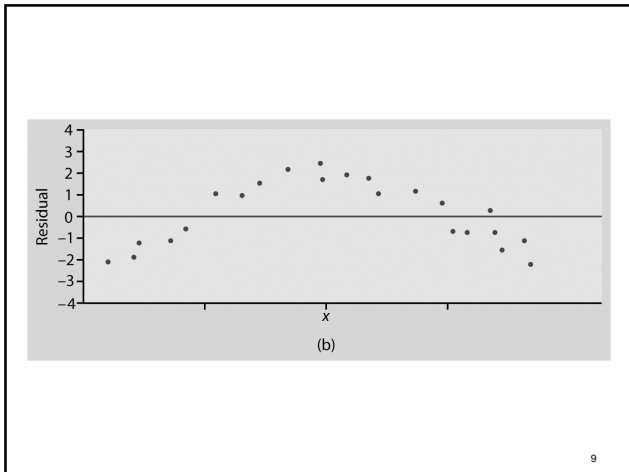
Residual Plots

- You always need to use a **residual plot** to check to see if model assumptions hold. Plot the residuals against the x variable (or predicted values).
- What to look for in a residual plot:
 - There should be no obvious patterns (random scatter about 0).
 - Vertical spread of the points should be approximately the same over the entire range of x -values.

7



8



Residual Plots in Stata

- After you run the regression, click on **Statistics > Linear regression and related > Regression diagnostics > Residual versus predictor plot**
- Enter the name of the explanatory variable in the box.

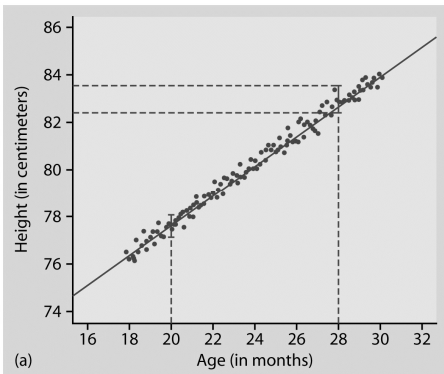
11

R²: Measure of Fit

- R² = squared value of the correlation coefficient
- Since the correlation must be between -1 and 1, R² must be between 0 and 1.
- R² has the interpretation of being the proportion of variation in Y that is explained by the variation in X.

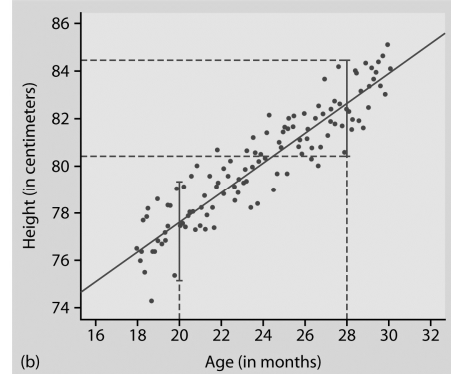
12

Example: $R^2 = 0.989$



13

Example: $R^2 = 0.849$



14

R^2 : Measure of Fit

- If $R^2 = 0.80$, that does not mean that 80% of Y is explained by X .
- This is a common mistake that is made.

15

Pizza Example



- Small pizza costs \$5.90 plus \$1 per topping.
- 100% of the variation in small pizza prices is due to differences (variation) in the number of toppings.
- Not the same as saying that 100% of the price of small pizzas is due to toppings.

16