Assumptions in OLS Regression

1. $\epsilon$ is a random variable that does not depend on $x$ (i.e., the model is perfect, it properly accounts for the role of $x$ in predicting $y$).
2. $\mathbb{E}[\epsilon] = 0$ (the population mean of the true residual is zero); this will always be true for a model with an intercept.
3. All $\epsilon_i$ are independent of each other (uncorrelated for the population, but not for a sample).
4. All $\epsilon_i$ have the same probability density function (pdf), and thus the same variance (called homoscedasticity).
5. $\epsilon \sim N(0, \sigma_{\epsilon})$ (the residuals, and thus the $y_i$, are normally distributed).
6. The values of each $x_i$ are known exactly.

Checking the Assumptions

- Do the assumptions in OLS regression hold?
  - Which assumptions can you validate?
  - If an assumption is violated, how far off is it?
- If one or more assumptions do not hold, does the observed violation invalidate the statistical procedure used?
  - If so, what next?

Failed Assumptions – the Anscombe Problems


Each of these four data sets produces exactly the same statistical fit (same standard deviation of residuals, same standard errors of model coefficients).

What Happens When OLS Assumptions are Violated?

- At best, the regression becomes inefficient
  - The uncertainty around the estimates is larger than you think: $\text{var}[^\theta] $ for some parameter $\theta$
- At worst, the regression becomes biased
  - The results may be misleading: $\text{bias}[^\theta]$
- We want small mean square error (MSE)
  \[
  \text{MSE}(^\theta) = \text{var}[^\theta] + (\text{bias}[^\theta])^2
  \]

Checking Our Assumptions

- Regression Diagnostics: checking for violations in any of the OLS assumptions
- Topics we’ll address:
  - Normality of residuals
  - Outliers (identically distributed)
  - Leverage and influence
  - Heteroscedasticity (variation in variance)
  - Error in predictor variables
  - The wrong model
  - Correlated residuals
Fixing Problems

- **Regression Remediation**: changing our regression to address diagnostic problems
- Topics we’ll address:
  - Outlier removal or adjustment
  - Data transformation
  - Weighted regression
  - Total regression
  - Model building
  - Autocorrelation analysis

Lecture 8: What have we learned?

- Name the six assumptions in OLS regression
- Define the mean square error (MSE) of a parameter estimate
- Describe the Anscombe graphs and what they teach us about regression