THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

CHE384, From Data to Decisions: Measurement, Uncertainty, Analysis, and Modeling

Lecture 8
Regression Review, part 3

Chris A. Mack
Adjunct Associate Professor

http://www.lithoguru.com/scientist/statistics/

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Assumptions in OLS Regression

1. ε 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. $E[\varepsilon_i] = 0$ (the population mean of the true residual is zero); this will always be true for a model with an intercept
3. All ε_i are independent of each other (uncorrelated for the population, but not for a sample)
4. All ε_i have the same probability density function (pdf), and thus the same variance (called homoscedasticity)
5. $\varepsilon \sim N(0,\sigma_\varepsilon)$ (the residuals, and thus the y_i, are normally distributed)
6. The values of each x_i are known exactly

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Checking the Assumptions

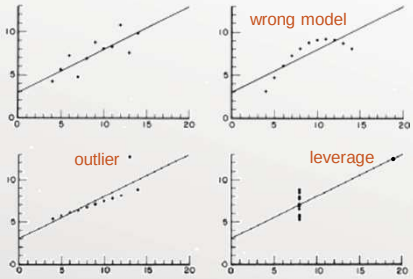
- Do the assumptions in OLS regression hold?
  - Which assumptions can you validate?
  - If an assumption is violated, how far off is it?
- If one or more assumptions do not hold, does the observed violation invalidate the statistical procedure used?
  - If so, what next?

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Failed Assumptions – the Anscombe Problems

F. J. Anscombe, "Graphs in Statistical Analysis", *The American Statistician*, Vol. 27, No. 1 (Feb., 1973) pp. 17 – 21.

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



THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

What Happens When OLS Assumptions are Violated?

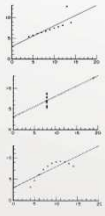
- At best, the regression becomes inefficient
  - The uncertainty around the estimates is larger than you think: $var[\hat{\theta}]$ for some parameter θ
- At worst, the regression becomes biased
  - The results may be misleading: $bias[\hat{\theta}]$
- We want small mean square error (MSE)

$$MSE(\hat{\theta}) = var[\hat{\theta}] + (bias[\hat{\theta}])^2$$

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Checking Our Assumptions

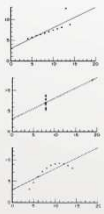
- Regression Diagnostics: checking for violations in any of the OLS assumptions
- Topics we'll address:
  - Normality of residuals
  - Outliers (identically distributed)
  - Leverage and influence
  - Heteroscedasticity (variation in variance)
  - Error in predictor variables
  - The wrong model
  -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



THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

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



THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

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