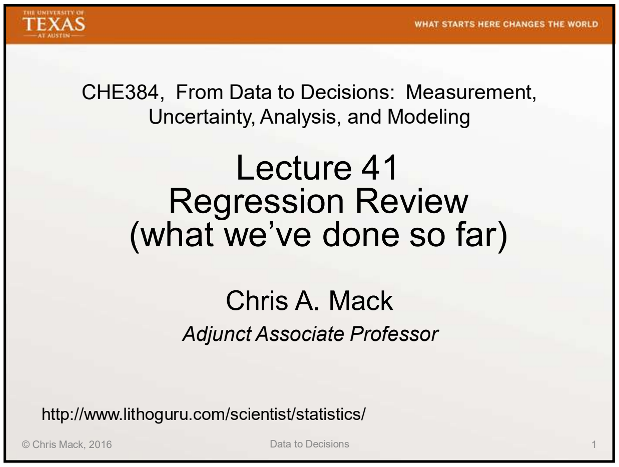


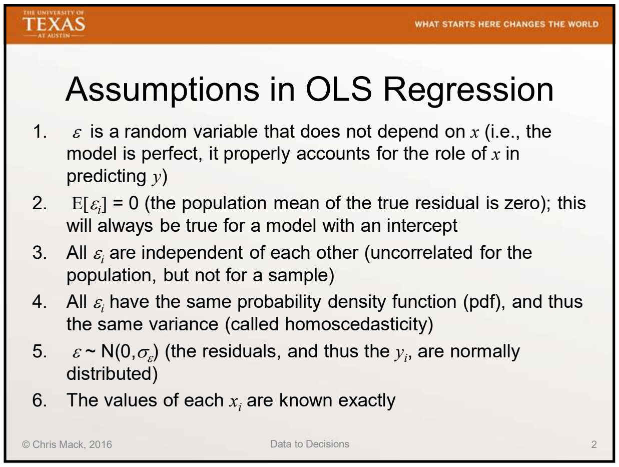
CHE384, From Data to Decisions: Measurement, Uncertainty, Analysis, and Modeling

Lecture 41 Regression Review (what we've done so far)

Chris A. Mack

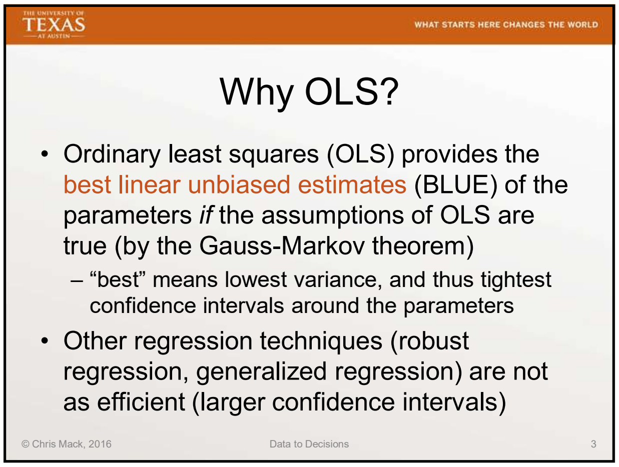
Adjunct Associate Professor

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



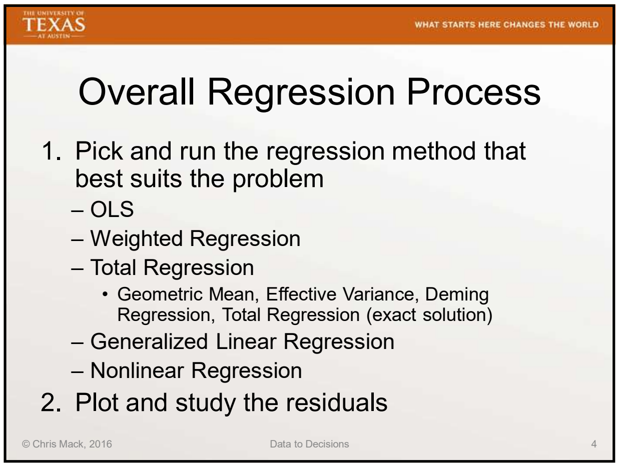
Assumptions in OLS Regression

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. $\mathrm{E}[\varepsilon_i] = 0$ (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



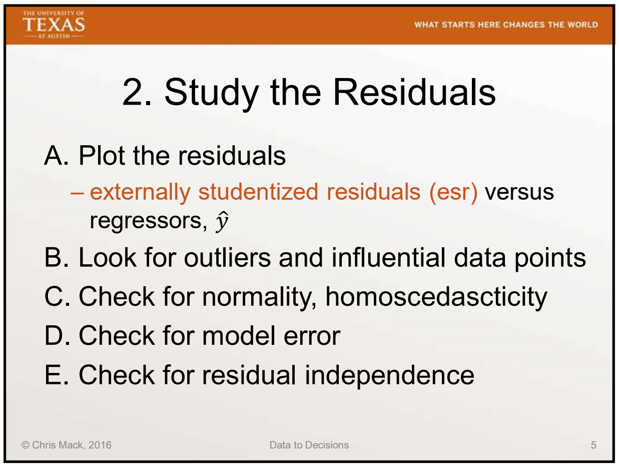
Why OLS?

- Ordinary least squares (OLS) provides the best linear unbiased estimates (BLUE) of the parameters *if* the assumptions of OLS are true (by the Gauss-Markov theorem)
  - "best" means lowest variance, and thus tightest confidence intervals around the parameters
- Other regression techniques (robust regression, generalized regression) are not as efficient (larger confidence intervals)



Overall Regression Process

1. Pick and run the regression method that best suits the problem
   - OLS
   - Weighted Regression
   - Total Regression
     - Geometric Mean, Effective Variance, Deming Regression, Total Regression (exact solution)
   - Generalized Linear Regression
   - Nonlinear Regression
2. Plot and study the residuals



2. Study the Residuals

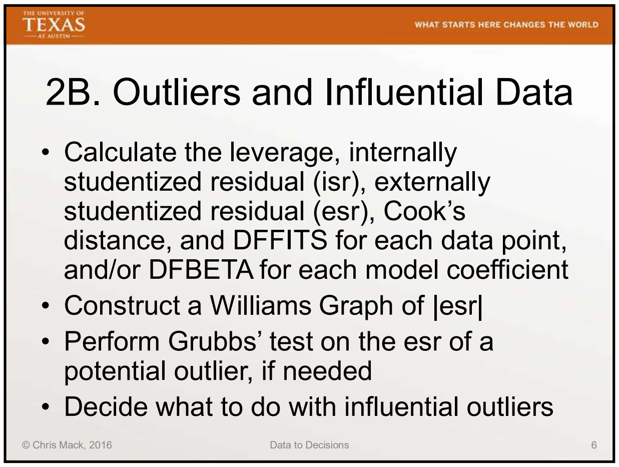
A. Plot the residuals
   - externally studentized residuals (esr) versus regressors, $\hat{y}$

B. Look for outliers and influential data points

C. Check for normality, homoscedascticity

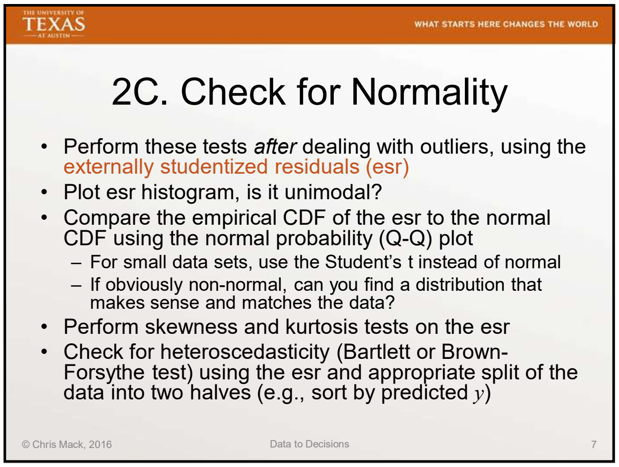
D. Check for model error

E. Check for residual independence



2B. Outliers and Influential Data

- Calculate the leverage, internally studentized residual (isr), externally studentized residual (esr), Cook's distance, and DFFITS for each data point, and/or DFBETA for each model coefficient
- Construct a Williams Graph of |esr|
- Perform Grubbs' test on the esr of a potential outlier, if needed
- Decide what to do with influential outliers

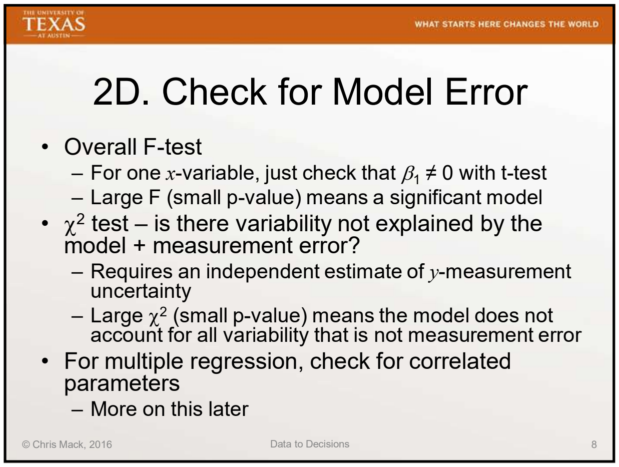


WHAT STARTS HERE CHANGES THE WORLD

2C. Check for Normality

- Perform these tests *after* dealing with outliers, using the externally studentized residuals (esr)
- Plot esr histogram, is it unimodal?
- Compare the empirical CDF of the esr to the normal CDF using the normal probability (Q-Q) plot
  - For small data sets, use the Student’s t instead of normal
  - If obviously non-normal, can you find a distribution that makes sense and matches the data?
- Perform skewness and kurtosis tests on the esr
- Check for heteroscedasticity (Bartlett or Brown-Forsythe test) using the esr and appropriate split of the data into two halves (e.g., sort by predicted y)

© Chris Mack, 2016 Data to Decisions 7

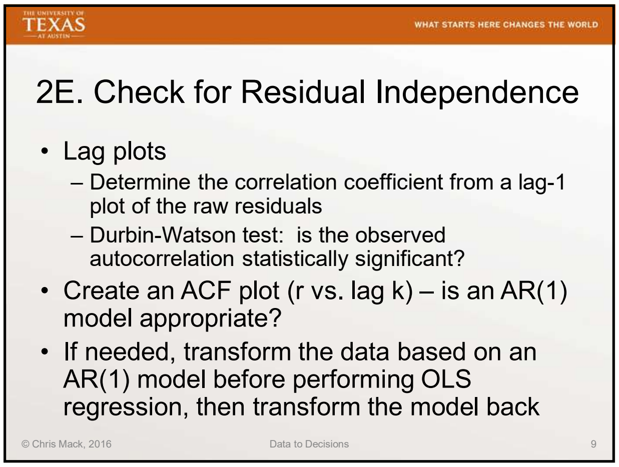


WHAT STARTS HERE CHANGES THE WORLD

2D. Check for Model Error

- Overall F-test
  - For one x-variable, just check that $\beta_1 \neq 0$ with t-test
  - Large F (small p-value) means a significant model
- χ^2 test – is there variability not explained by the model + measurement error?
  - Requires an independent estimate of y-measurement uncertainty
  - Large χ^2 (small p-value) means the model does not account for all variability that is not measurement error
- For multiple regression, check for correlated parameters
  - More on this later

© Chris Mack, 2016 Data to Decisions 8

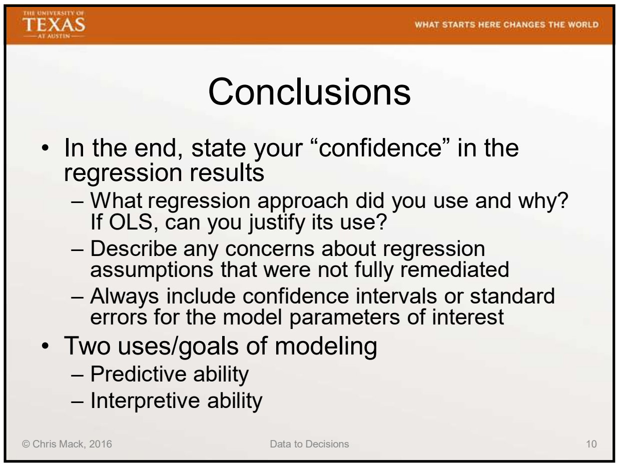


WHAT STARTS HERE CHANGES THE WORLD

2E. Check for Residual Independence

- Lag plots
  - Determine the correlation coefficient from a lag-1 plot of the raw residuals
  - Durbin-Watson test: is the observed autocorrelation statistically significant?
- Create an ACF plot (r vs. lag k) – is an AR(1) model appropriate?
- If needed, transform the data based on an AR(1) model before performing OLS regression, then transform the model back

© Chris Mack, 2016 Data to Decisions 9



WHAT STARTS HERE CHANGES THE WORLD

Conclusions

- In the end, state your “confidence” in the regression results
  - What regression approach did you use and why? If OLS, can you justify its use?
  - Describe any concerns about regression assumptions that were not fully remediated
  - Always include confidence intervals or standard errors for the model parameters of interest
- Two uses/goals of modeling
  - Predictive ability
  - Interpretive ability

© Chris Mack, 2016 Data to Decisions 10