

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

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

## Lecture 34

### The Wrong Model

Chris A. Mack  
Adjunct Associate Professor

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

© Chris Mack, 2016 Data to Decisions 1

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

## Assumptions in OLS Regression

1.  $\varepsilon$  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  $\varepsilon_i$  are independent of each other (uncorrelated for the population, but not for a sample)
4. All  $\varepsilon_i$  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

© Chris Mack, 2016 Data to Decisions 2

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

## Process Modeling

$$y = f(x, \beta) + \varepsilon$$

- Our three tasks in modeling:
  - Find the **equation**  $f(x, \beta)$  that meets our goals
    - Picking the right regressors
    - Picking the right model form
  - Find the **values of the coefficients**  $\beta$  that are “best” in some sense
    - Characterize the nature of  $\varepsilon$  (**distribution of errors**)
- These three tasks are interdependent

© Chris Mack, 2016 Data to Decisions 3

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

## The Wrong Equation

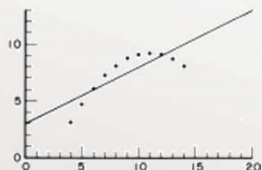
- Also called “model misspecification” or “equation error”
- Picking the wrong equation means that some of the variance in  $y$  isn’t properly explained
  - **Underfitting**: systematic variation in  $y$  is left unexplained
  - **Overfitting**: random variation in  $y$  is fit with a model giving reduced predictive power
  - **Goodness of fit**: various measures of fitting

© Chris Mack, 2016 Data to Decisions 4

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

## First Defense: Graphing

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

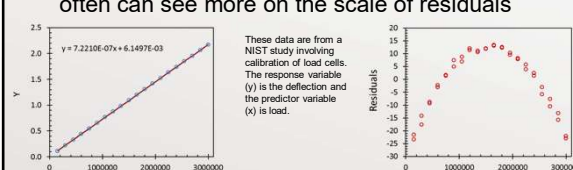


© Chris Mack, 2016 Data to Decisions 5

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

## First Defense: Graphing

- First, graph the data. Does the proposed model make sense given what you see?
- Second, graph the residuals after your fit - you often can see more on the scale of residuals



These data are from a NIST study involving calibration of load cells. The response variable (y) is the deflection and the predictor variable (x) is load.

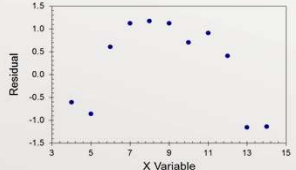
<http://www.itl.nist.gov/div898/strd/ls/data/?ontius.shtml>

© Chris Mack, 2016 Data to Decisions 6

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

## First Defense: Graphing

- The human brain is ideally suited to detecting patterns, **even when there is none**
- Is this an indication of systematic curvature, heteroscedasticity, or just random variation?



© Chris Mack, 2016 Data to Decisions 7

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

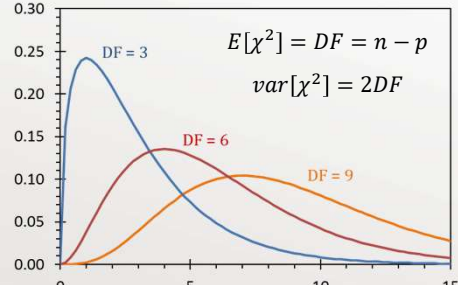
## $\chi^2$ Testing for Equation Error

- In least-squares regression we seek to minimize
 
$$\chi^2 = \sum_{i=1}^n \frac{\varepsilon_i^2}{\sigma_i^2}$$
- If the OLS assumptions are true, then this will be **chi square distributed** with  $n - p$  degrees of freedom
  - We can test for this

© Chris Mack, 2016 Data to Decisions 8

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

## Chi Square Distribution



$E[\chi^2] = DF = n - p$   
 $var[\chi^2] = 2DF$

© Chris Mack, 2016 Data to Decisions 9

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

## $\chi^2$ Testing for Equation Error

- In order to calculate  $\chi^2$  for our fit we must know  $\sigma_i$  ahead of time (**independent of our regression**) through independent information
  - What is the measurement uncertainty in  $y$ ?
- Calculate the probability of getting this value of  $\chi^2$  assuming no equation error
  - This statistic is very sensitive to the assumption of normal, homoscedastic residuals
- Compare the resulting p-value to our designated significance level ( $\alpha$ )
  - If we fail the test due to large  $\chi^2$ , then **we reject the hypothesis that the model accounts for all variation in  $y$  except measurement uncertainty as described by the  $\sigma_i$**

© Chris Mack, 2016 Data to Decisions 10

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

## $\chi^2$ Goodness of Fit Test

- Reduced chi-square:  $\chi_{red}^2 = \chi^2 / (n - p)$
- For  $\chi_{red}^2 \gg 1$ : poor model
  - There is variance in  $y$  that is not explained by the model or by measurement error, or our error variance estimate is too low
- For  $\chi_{red}^2 \ll 1$ : fitting the noise
  - The model is fitting the noise, or our error variance estimate is too high
- For  $\chi_{red}^2 \approx 1$ : good model fit
  - The variation in  $y$  is well explained by the model + measurement error

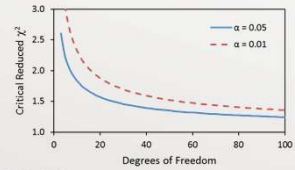
© Chris Mack, 2016 Data to Decisions 11

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

## Simple Case: $\sigma_i = \text{constant}$

- Compare residual variance to error estimate
 
$$\chi_{red}^2 = \frac{1}{n - p} \sum_{i=1}^n \frac{\varepsilon_i^2}{\sigma_i^2} = \frac{1}{\sigma^2} \left( \frac{\sum_{i=1}^n \varepsilon_i^2}{n - p} \right) = \frac{s_\varepsilon^2}{\sigma^2}$$

Alternate view:  
What amount of measurement error would allow us to pass this test?  
Is that amount reasonable?



© Chris Mack, 2016 Data to Decisions 12

THE UNIVERSITY OF TEXAS  
AT AUSTIN

WHAT STARTS HERE CHANGES THE WORLD

## $\chi^2$ for Weighted Regression

- Recall that for a weighted regression
 
$$[SE(\sqrt{w}\varepsilon)]^2 = \frac{\sum_{i=1}^n w_i (y_i - \hat{y}_i)^2}{(n-p)}$$
- Using the usual weights ( $w_i = 1/\sigma_i^2$ ),
 
$$[SE(\sqrt{w}\varepsilon)]^2 = \frac{\chi^2}{(n-p)} = \chi_{red}^2$$
- Thus the expected value of  $SE(\sqrt{w}\varepsilon) = 1$

© Chris Mack, 2016 Data to Decisions 13

THE UNIVERSITY OF TEXAS  
AT AUSTIN

WHAT STARTS HERE CHANGES THE WORLD

## $\chi^2$ for Total Regression

- For a total regression, the sum of square errors is
 
$$S = \sum_{i=1}^n \left[ \left( \frac{\hat{y}_i - y_i}{\sigma_{yi}} \right)^2 + \left( \frac{\hat{x}_i - x_i}{\sigma_{xi}} \right)^2 \right] \quad \begin{array}{l} \hat{y}_i = \text{predicted } y \text{ value} \\ \hat{x}_i = \text{predicted } x \text{ value} \end{array}$$
- If  $x$  and  $y$  are normally distributed, then  $S$  will be  $\chi^2$  distributed with  $n - p$  degrees of freedom
  - $-p$  = number of model parameters excluding the  $n$  predicted  $x$  values

© Chris Mack, 2016 Data to Decisions 14

THE UNIVERSITY OF TEXAS  
AT AUSTIN

WHAT STARTS HERE CHANGES THE WORLD

## Lecture 34: What have we learned?

- How can we use graphing to detect equation error?
- What is required to perform a chi-square test of the residuals to find equation error?
- What is the reduced chi-square and what is its expected value?
- What are the assumptions inherent in the chi-square test?

© Chris Mack, 2016 Data to Decisions 15