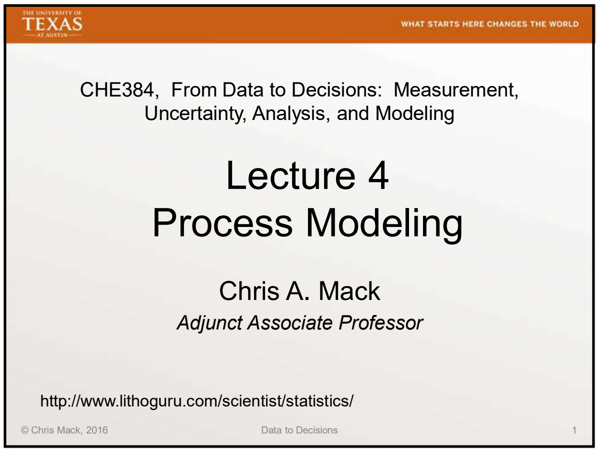


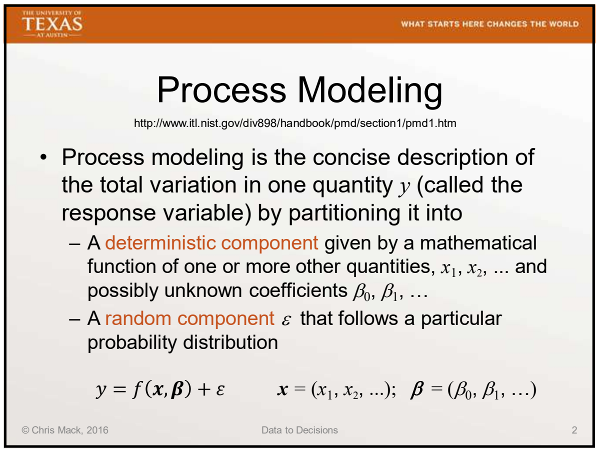
CHE384, From Data to Decisions: Measurement, Uncertainty, Analysis, and Modeling

Lecture 4 Process Modeling

Chris A. Mack

Adjunct Associate Professor

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

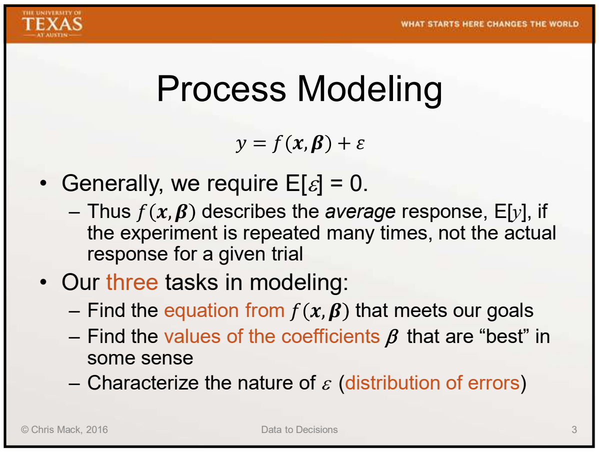


Process Modeling

http://www.itl.nist.gov/div898/handbook/pmd/section1/pmd1.htm

- Process modeling is the concise description of the total variation in one quantity y (called the response variable) by partitioning it into
  - A deterministic component given by a mathematical function of one or more other quantities, x_1, x_2, ... and possibly unknown coefficients β_0, β_1, …
  - A random component ε that follows a particular probability distribution

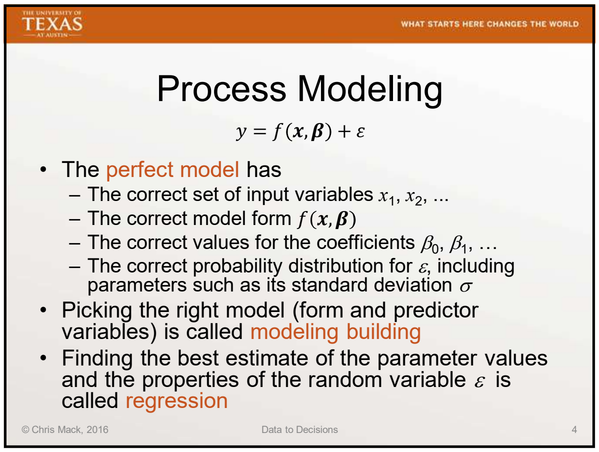
$$y = f(\boldsymbol{x}, \boldsymbol{\beta}) + \varepsilon \qquad \boldsymbol{x} = (x_1, x_2, ...); \;\; \boldsymbol{\beta} = (\beta_0, \beta_1, \ldots)$$



Process Modeling

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

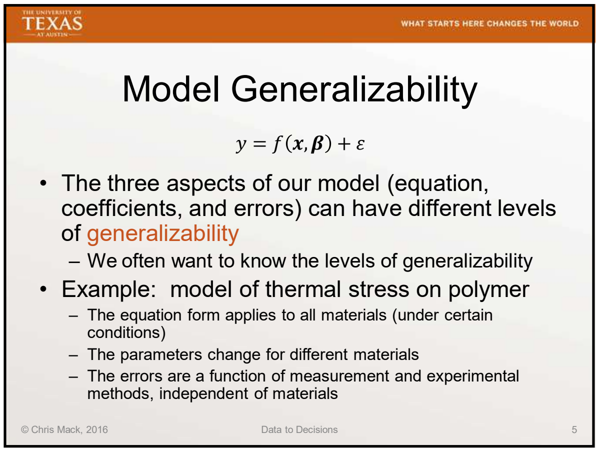
- Generally, we require E[ε] = 0.
  - Thus $f(\boldsymbol{x}, \boldsymbol{\beta})$ describes the *average* response, E[y], if the experiment is repeated many times, not the actual response for a given trial
- Our three tasks in modeling:
  - Find the equation from $f(\boldsymbol{x}, \boldsymbol{\beta})$ that meets our goals
  - Find the values of the coefficients $\boldsymbol{\beta}$ that are “best” in some sense
  - Characterize the nature of ε (distribution of errors)



Process Modeling

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

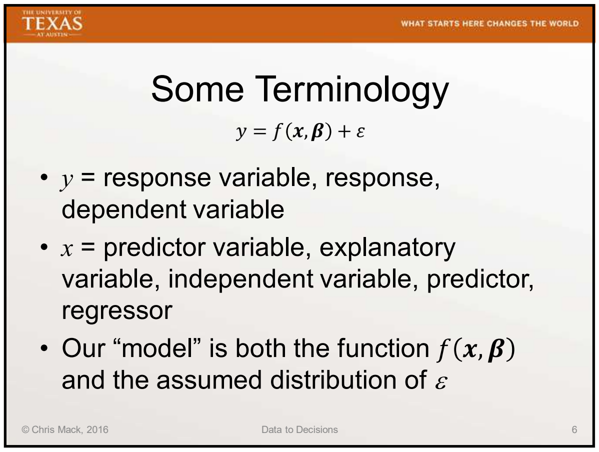
- The perfect model has
  - The correct set of input variables x_1, x_2, ...
  - The correct model form $f(\boldsymbol{x}, \boldsymbol{\beta})$
  - The correct values for the coefficients β_0, β_1, …
  - The correct probability distribution for ε, including parameters such as its standard deviation σ
- Picking the right model (form and predictor variables) is called modeling building
- Finding the best estimate of the parameter values and the properties of the random variable ε is called regression



Model Generalizability

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

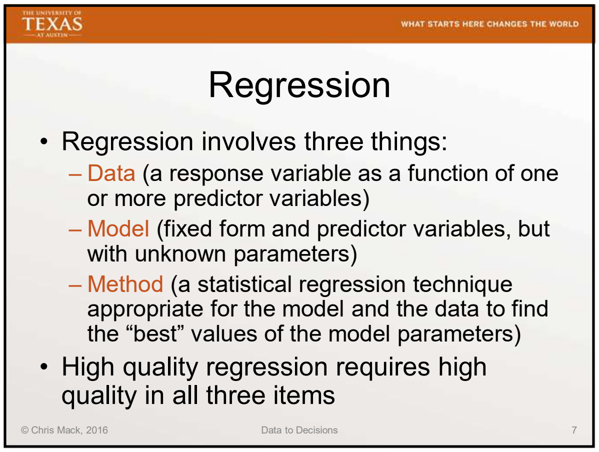
- The three aspects of our model (equation, coefficients, and errors) can have different levels of generalizability
  - We often want to know the levels of generalizability
- Example: model of thermal stress on polymer
  - The equation form applies to all materials (under certain conditions)
  - The parameters change for different materials
  - The errors are a function of measurement and experimental methods, independent of materials



Some Terminology

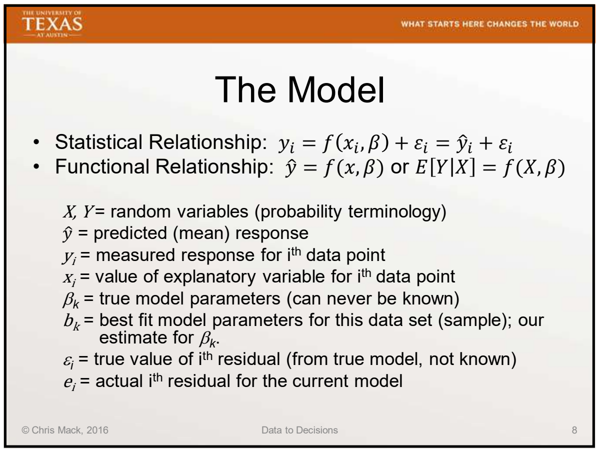
$$y = f(\boldsymbol{x}, \boldsymbol{\beta}) + \varepsilon$$

- y = response variable, response, dependent variable
- x = predictor variable, explanatory variable, independent variable, predictor, regressor
- Our “model” is both the function $f(\boldsymbol{x}, \boldsymbol{\beta})$ and the assumed distribution of ε



Regression

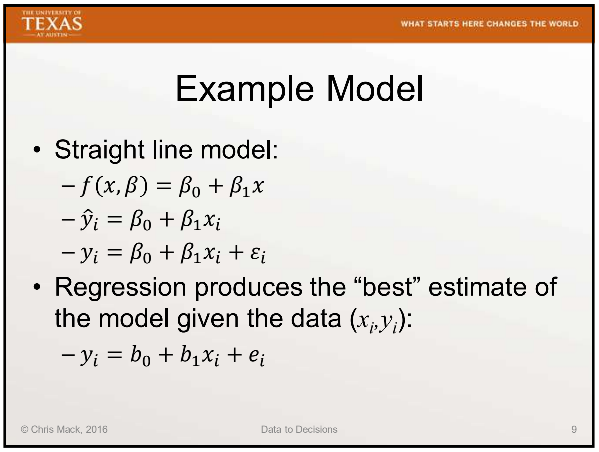
- Regression involves three things:
  - Data (a response variable as a function of one or more predictor variables)
  - Model (fixed form and predictor variables, but with unknown parameters)
  - Method (a statistical regression technique appropriate for the model and the data to find the "best" values of the model parameters)
- High quality regression requires high quality in all three items



The Model

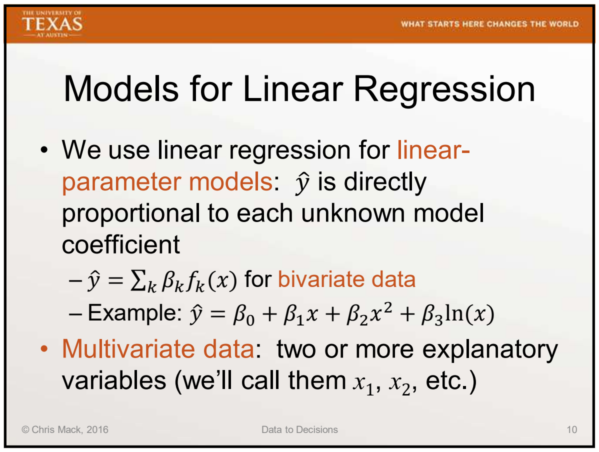
- Statistical Relationship: $y_i = f(x_i, \beta) + \varepsilon_i = \hat{y}_i + \varepsilon_i$
- Functional Relationship: $\hat{y} = f(x, \beta)$ or $E[Y|X] = f(X, \beta)$

X, Y= random variables (probability terminology)
$\hat{y}$ = predicted (mean) response
y_i = measured response for ith data point
x_i = value of explanatory variable for ith data point
β_k = true model parameters (can never be known)
b_k = best fit model parameters for this data set (sample); our estimate for β_k.
ε_i = true value of ith residual (from true model, not known)
e_i = actual ith residual for the current model



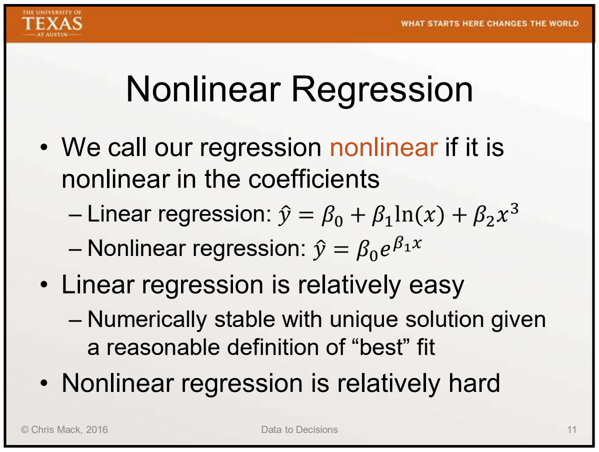
Example Model

- Straight line model:
  - $f(x, \beta) = \beta_0 + \beta_1 x$
  - $\hat{y}_i = \beta_0 + \beta_1 x_i$
  - $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
- Regression produces the "best" estimate of the model given the data (x_i,y_i):
  - $y_i = b_0 + b_1 x_i + e_i$



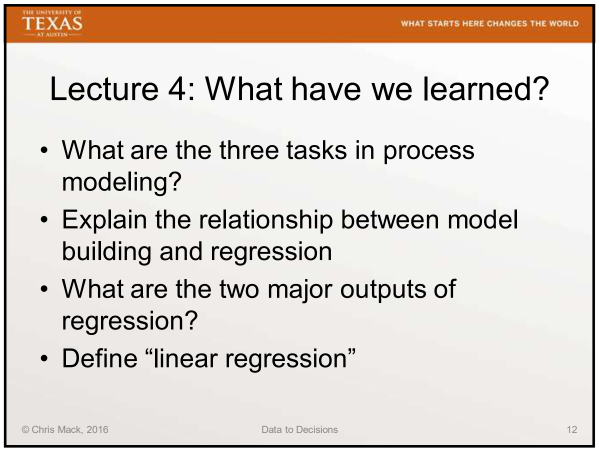
Models for Linear Regression

- We use linear regression for linear-parameter models: $\hat{y}$ is directly proportional to each unknown model coefficient
  - $\hat{y} = \sum_k \beta_k f_k(x)$ for bivariate data
  - Example: $\hat{y} = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 \ln(x)$
- Multivariate data: two or more explanatory variables (we'll call them x_1, x_2, etc.)



Nonlinear Regression

- We call our regression nonlinear if it is nonlinear in the coefficients
  - Linear regression: $\hat{y} = \beta_0 + \beta_1 \ln(x) + \beta_2 x^3$
  - Nonlinear regression: $\hat{y} = \beta_0 e^{\beta_1 x}$
- Linear regression is relatively easy
  - Numerically stable with unique solution given a reasonable definition of "best" fit
- Nonlinear regression is relatively hard



Lecture 4: What have we learned?

- What are the three tasks in process modeling?
- Explain the relationship between model building and regression
- What are the two major outputs of regression?
- Define "linear regression"