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

Lecture 59
Other Regression Topics

Chris A. Mack
Adjunct Associate Professor

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

Indicator Variables

- Categorical (qualitative) variables can be included in a regression using indicator variables
  - Did the sample receive a certain treatment?
  - Which tool was used to perform the process?
  - What material was the part made from?
- Indicator variables are binary, taking values of 0 or 1 only

$$x_2 = \begin{cases} 0 & \text{If the observation is from tool A} \\ 1 & \text{If the observation is from tool B} \end{cases}$$

Indicator Variable Example

- Consider an etch process with material removed (y) versus time (x_1) for different etch tools (indicator variable x_2)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

- For tool A (x_2 = 0), $\hat{y} = \beta_0 + \beta_1 x_1$
- For tool B (x_2 = 1), $\hat{y} = (\beta_0 + \beta_2) + \beta_1 x_1$
  - Result: same slope (etch rate), but different intercepts (i.e., parallel lines)

With Interactions

- Now add an interaction term:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

- For tool A (x_2 = 0), $\hat{y} = \beta_0 + \beta_1 x_1$
- For tool B (x_2 = 1), $\hat{y} = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_1$
  - Result: β_2 is the change in intercept and β_3 is the change in slope when the tool changes from A to B
  - The fit is the same as if two separate regressions are performed (one for tool A and one for tool B)

Indicators for Multiple Levels

- Suppose there are three different tools

x_2	x_3	
0	0	If the observation is from tool A
1	0	If the observation is from tool B
0	1	If the observation is from tool C

- For a levels, we need a – 1 binary indicator variables
  - Making the mistake of using a indicator variables for a levels produces perfect multicollinearity!

Nonlinear Regression

- We call our regression nonlinear if it is nonlinear in the coefficients
  - Linear regression: $\hat{y} = \beta_0 + \beta_1 \ln(x)$
  - Nonlinear regression: $\hat{y} = \beta_0 e^{\beta_1 x}$
- To find the best fit coefficients, we must use an iterative technique
  - There are problems with local minima vs. global minimum
  - The solution may depend on our initial guess
  - Convergence can be slow (or may diverge)

Nonlinear Model Examples

- Michaelis-Menten model for enzyme kinetics

$$\hat{y} = \frac{\beta_0 x}{\beta_1 + x}$$

$\hat{y}$ = reaction rate
x = enzyme substrate concentration

- Logistic Growth model

$$P(t) = \frac{K}{1 + \left(\frac{K}{P_0} - 1\right)e^{-r}}$$

P = population at time t
P_0 = initial population
K = carrying capacity (max population)
r = growth rate

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

Transforming from Nonlinear to Linear

- Sometimes, a nonlinear model can be transformed to a linear model

$$\hat{y} = \beta_0 e^{\beta_1 x} \rightarrow ln(\hat{y}) = ln(\beta_0) + \beta_1 x$$

- The big question: What will be the distribution of residuals in the transformed model?

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

Nonlinear Regression Techniques

- Steepest Descent: take the gradient, find the direction where χ^2 is declining fastest, go in that direction
  - Problem: we don't know how far to go, so we take small steps, resulting in slow convergence
- Taylor Series: use a second order fit to the local region and jump to the minimum of that parabola
  - Only works if we are near the minimum
- Levenberg-Marquadt: start with steepest descent, then switch to Taylor series
  - Problem: we may get caught in a local minimum, depending on the starting point

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

Lecture 59: What have we learned?

- What is an indicator variable and how is it used in regression?
- If I have a categorical variable with four levels, how many indicator variables are needed to represent it?
- What are the problems with non-linear regression?