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

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

Lecture 42
Multiple Regression

Chris A. Mack
Adjunct Associate Professor

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

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

Multiple Regression

- Multiple Regression is regression against more than one predictor variable
- Case 1: The same input variable in different forms (bivariate data)
  - $\hat{y} = \sum_k \beta_k f_k(x)$ for bivariate data
  - Example: $\hat{y} = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 \ln(x)$
  - Polynomial models can be justified as Taylor series expansions of an unknown function

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

Multiple Regression

- Multiple Regression is regression against more than one predictor variable
- Case 2: More than one input variable (multivariate data)
  - $\hat{y} = \sum_k \beta_k f_k(x_1, x_2, \dots, x_m)$ for multivariate data
  - Example: $\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 \ln(x_2)$
  - Holding all other variables constant, β_j is the change in $\hat{y}$ (mean Y) per unit change in x_j

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

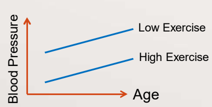
Multiple Regression Math

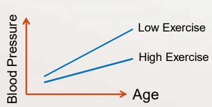
- Multiple Linear Regression uses the same mathematical techniques as standard linear regression
  - Maximum likelihood estimator (MLE): define the likelihood function, take the derivative with respect to each parameter, solve p equations simultaneously to give an exact solution
  - OLS: same solution as the MLE solution when the proper assumptions are met

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

What's New in Multiple Regression

- Interaction: often predictor variables interact in their influence on the response
  - Example: $\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$
  - The influence of x_1 an y depends on the magnitude of x_2 (and vice versa)
    - For fixed x_2, $\hat{y} = (\beta_0 + \beta_2 x_2) + (\beta_1 + \beta_3 x_2) x_1$
    - For fixed x_1, $\hat{y} = (\beta_0 + \beta_1 x_1) + (\beta_2 + \beta_3 x_1) x_2$





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

What's New in Multiple Regression

- Multicollinearity: often predictor variables are correlated with each other – they are not independent
  - Also called confounding
  - More on this in later lectures
- Example: what body measures predict strength in a certain fitness test?
  - Height is correlated with strength
  - Weight is correlated with strength
  - But height and weight are correlated with each other! Are they really two different predictors?

A Quick Example

- A shampoo manufacturer has purchased a data set on population characteristics and wants to build model
  - Length of hair is used as a proxy for how much shampoo will be consumed: the response
  - Surprisingly, hair length is negatively correlated with height. Why?
  - Missing predictor variable: gender

Building a Model

- Which predictor variables should be included in the model (and which should be ignored)?
  - Have we left out an important predictor?
  - Have we included an unnecessary predictor?
- What is the functional form of the model?
- Theory often guides us to answer these questions, but doesn't always leave us without choices
  - The answers usually depend on the purpose of our model

What is the Purpose of our Model?

- Predictive ability: Predicting the model output
  - We want a small SE($\hat{y}$), and we don't care about SE(β_k)
  - We need to avoid overfitting, but the form of the model is not too important
  - We need to think about model scope (the range of predictor variables over which our predictions are valid)
- Control: Given a measured output, how much should we change an input to move the output?
  - We want small SE(β_k)
- Interpretive ability: Testing or validating a theory, providing explanation
  - We want a both small SE($\hat{y}$) and small SE(β_k)

Building a Model

- In general, we strive for parsimony
  - Find the simplest model consistent with the data and our knowledge of the problem
- If a simple model is not good enough, we can
  - Add more predictor variables
  - Add more complex functions of the predictor variables
  - Add interaction terms
- How do we know if the added terms are really helping, or just fitting the noise (overfitting)?
  - R^2 always improves when new model terms are added
  - We need something else to understand overfitting

Lecture 42: What have we learned?

- Define multiple regression, and how it applies to both bivariate and multivariate data
- What mathematical techniques are used for OLS multiple regression?
- Explain *interaction* and what it means for multiple regression
- Explain *multicollinearity* and what it means for multiple regression
- Provide three unique purposes for a model