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

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

Lecture 65 Regression Design

Chris A. Mack

Adjunct Associate Professor

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

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

Six Principles for Regression Design

(NIST/SEMATECH e-Handbook of Statistical Methods, section 4.3.3)

- Capacity for the primary model
- Capacity for the alternate model
- Minimum variance of estimated coefficients or predicted values
  - Except for simple cases, must search for optimal design
- Sample where the variation is
- Repeats and replication
  - To compute a model-independent estimate of the process standard deviation
- Randomization and blocking
  - Allows the detection of drift, reduces influence of known/unknown but unimportant variations

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

Primary vs. Alternate Model

- For exploratory work, we may not have a clear idea of what our model could be
- In some cases, we have a clear primary and alternate model in mind
- Simple case: one predictor variable, linear vs. quadratic models
  - Optimizing the design for linear (dumbbell design) means we are insensitive to quadratic variation
  - Optimizing for quadratic gives us reasonable efficiency for a linear model

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

Sample Where the Variation Is

- For non-constant variance, make number of data/repeats $n_i \propto \sigma_i^2$
- For curves, sample more in the steep regions
  - Think about evenly spaced y-values rather than evenly spaced x-values

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

Optimal Design

- For general multiple regression models, designs can be complicated!
- Optimal Design is an algorithmic approach for searching the design space and optimizing some statistical metric of the model
  - Non-optimal designs require a greater number of data points to estimate parameters with the same precision
  - With multiple predictor variables, there can be trade-offs between parameter variances
  - Limitation: The model must be specified ahead of time, as well as the range for each predictor variable

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

What to Optimize?

- A-optimality (average): minimize the average variance of the estimates of the regression coefficients (trace of covariance matrix)
- C-optimality (combination): minimize the variance of a predetermined linear combination of model parameters
- D-optimality (determinant): maximize the determinant of the information matrix X^TX (minimize determinant of covariance matrix)
- E-optimality (eigenvalue): maximize the minimum eigenvalue of the information matrix
- T-optimality: maximize the trace of the information matrix
- G-optimality: minimize the maximum h_{ii} (hat matrix diagonal), minimizing the maximum variance of the predicted values
- I-optimality (integrated): minimize the average prediction variance over the design space
- V-optimality (variance): minimize the average prediction variance over a set of m specific points

Optimal Design Examples

- Linear and quadratic regression models with uncorrelated observations
  - D-optimal design is dumbbell for linear model and equal thirds for quadratic model
- Linear and quadratic regression models with highly correlated observations (an autoregressive error structure)
  - D-optimal design is close to equally spaced

Replicates versus Repeats

- Repeats are duplications of the experiment on some data within the same experimental run
  - Example: repeat the generation of one data point five times to independently assess variability
  - Repeats often do not include all sources of variation
- Replicates are repeated experimental runs where the entire experimental run is repeated at a separate time
  - Each replicate is subject to the same (full) variability, but independently (i.e., a complete block)

Repeat/Replicate Examples

- We wish to test the effect of a tool setting (3 levels) on the quality of a manufactured product
- Repeats:
  - Set the tool to a randomly selected level, run five products through and measure each
  - Set the tool to the next level, and repeat
- Replicates:
  - Set the tool to a randomly selected level, run one product and measure. Set the tool to the next level, and repeat until all levels have been measured.
  - Replicate the above experiment five times, each with a randomized order of levels

Randomization

- For uncontrolled, unmeasured inputs, use randomization to prevent an unknown effect from biasing our results
  - Turn systematic errors into random errors, which average to zero
  - Allows for time-series analysis to detect drift
- Note: randomization is not as effective as blocking when trying to remove known variation (uncontrolled, measured inputs)

Blocking

- Let X be the results from treatment 1, and Y the results from treatment 2. We wish to measure X – Y (the difference in results)

$$var[X-Y] = var[X] + var[Y] - 2cov[X,Y]$$

- We can reduce the variance of X – Y by increasing the covariance of X and Y
  - An error that is the same for X and Y will cancel
  - We can increase the covariance with blocking

Lecture 65: What have we learned?

- Can you name all six principles of designing for regression?
- What is optimal design?
- What is the difference between repeats and replicates?
- What is randomization used for?