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

Lecture 25 Design of Experiments and Blocking

Chris A. Mack
Adjunct Associate Professor

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

 1

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
- Replication
  - To compute a model-independent estimate of the process standard deviation
- Randomization and blocking
  - Allows the detection of drift, reduces influence of unimportant variations (called nuisance variables)

 2

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

 3

Sample Where the Variation Is

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

 4

Optimal Design

- For general multiple regression models, there are no simple designs that can be applied
- 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
  - The model must be specified ahead of time, as well as the range for each predictor variable
  - With multiple predictor variables, there can be trade-offs between parameter variances

 5

What to Optimize?

- A-optimality (average): minimize the average variance of the estimates of the regression coefficients
- 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 of the design
- 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

 6

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

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
  - This is called blocking

Blocking or Covariate Analysis

- For uncontrolled, measured inputs, use blocking to remove an effect we are not interested in, reducing known variability
  - Different measurement tools, process batches
  - Spatial or temporal variations
- Blocking: grouping the experimental results that you wish to compare into blocks that are similar to one another
- Covariate Analysis: put the uncontrolled but measured inputs into the model. Then ignore them when the modeling is complete

Classic Blocking Example

- I have a new shoe sole that I claim will last longer, but is otherwise identical to the existing sole
- I've recruited 100 volunteers to test the new versus old soles. What experimental design should I use?
  - Randomization: randomly assign half of the participants to wear the new shoes, and half to wear the old shoes
  - Blocking: every participant is given one shoe with the new sole and one with the old (randomly assigned to left or right foot)

Randomized Complete Block Design

- The complete experiment is performed for each block
  - Within each block, the testing order is randomized
- Examples of blocks
  - Raw material batches
  - People (operators)
  - Process or measurement tools
  - Time
- If it is not possible to run a complete experiment in each block, use a balanced incomplete block design
  - Any two treatments appear together an equal number of times

Latin Square Design

- For the case of one primary variable and two nuisance variables (blocking variables)
  - Graeco-Latin square design: 3 nuisance factors
  - Hyper-Graeco-Latin square design: 4 nuisance factors
- The number of levels of each blocking variable must equal the number of levels of the primary variable
- The Latin square model assumes that there are no interactions between variables

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

Latin Square Design

- Example: We test amount of gasoline additive on emissions using four cars and four drivers

	Car 1	Car 2	Car 3	Car 4
Driver 1	T1	T2	T4	T3
Driver 2	T4	T3	T1	T2
Driver 3	T2	T4	T3	T1
Driver 4	T3	T1	T2	T4

T1 through T4 are the four treatment levels (amount of additive)

4X4 Latin Square Design

