Business Analytics
Key Concepts and Learning Points by Case Session

Session 1: How to measure and describe markets, processes, opinions? – Statistics

Case: Firestone - snowfall promotion

- Types of Data - time series, cross-sectional, qualitative, quantitative
- Data Analysis Toolpak - Excel AddIn
- Descriptive Statistics
    - mean, median, mode (average, median) – central tendency
    - variance, standard deviation (stdev) - spread
    - coefficient of variation (stdev / mean)
    - symmetry, skew
    - percentiles
    - covariance and correlation (how they move together)

Descriptive Statistics: Continued

- Histograms and Pareto Charts (using Toolpak)
- Graphs and Charts (Chart Wizard in Exel)
- probability as relative frequency
- Expected value of a random variable
    - weighted average with probabilities as the weights

Session 2: How to quantify and work with risks and uncertainties? – Probability

Case: OLGC - baseball betting

- what is probability (likelihood, chance, relative frequency, etc)
- sample spaces, trials, outcomes, events
- Venn Diagrams -> probability rules
    - complement, unions, intersections
    - P(A and B) = P(A | B) * P(B)
        - if indep: P(A | B) = P(A) so P(A and B) = P(A) * P(B)
    - P(A or B) = P(A) + P(B) - P(A and B)
- mutually exclusive
- independence

Probability: Continued

- random variables
  - probability distributions:
    - discrete
      - table of X and P(X)
    - continuous
      - probability density function f(x)
      - area under curve
- Discrete Probability Distributions
  - Binomial(n,p), Poisson(lambda)

Probability: Continued

- Continuous Probability Distributions
  - Normal(mu, sigma) - Bell Shaped curve
  - Others: Exponential(mu), t, F, chi-square
- Expected value of a random variable: E[X]
- Variance of a random variable: Var[X] = E[X*X] - (E[X]*E[X])
- Excel Functions: t.dist, norm.dist, binom.dist
- Excel Functions: t.inv, norm.inv, binom.inv, etc.

Session 3. How to make better estimates incorporating multiple factors? – Regression

Case Alfonsos - advertising versus sales

- What is Regression?
  - Least Squares Error Estimation Procedure
  - $Y = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + \ldots + e$
  - Y is a linear function of X's
  - X's need not be linear functions
- How do you run a Regression?
  - Place data in contiguous columns to represent the model
  - Use Data Analysis Toolpack – Regression

Regression Continued:

- Regression Output
  - Overall F test (is model valid - can you reject Ho: all B's are 0)
  - R^2 - measure of percent of variance explained by model
  - Standard Error - estimate of e's standard deviation (sigma)
  - Beta Estimates - Is Y and that X linearly associated?
  - p-values for t-test: Ho: $B_i = 0$ (student t-tests)
- Assumptions Needed for Testing via Regression
  - Y must be actually linearly related to X's
  - e is Normally distributed with mean 0 (normal plots)
  - e has constant variance sigma2 (heteroskedasticity)

Regression Continued:

- e's are independent of each other (autocorrelation)
- X's are not a linear combination of each other(multicollinearity)
- all required X's are in the model (omitted variables)
- outliers are not present (driving results?) tests of high influence
    - remove point and see how much changes

- VIFs - Variance Inflation Factor (should be less than 5)
    - Simply regress each X on all of the other X's and
      calculate VIF for each one
    - VIF = $1 / (1 - R^2)$
    - If R^2 is over 80% then VIF > 5 - an issue exists

Session 4. How to draw valid conclusions from data: Regression Testing

Case: Visteon: viscosity vs tester vs cure time

- What is Sampling?
    - random sampling
    - sampling distribution of an average of n from normal pop
        - variance of average of n is variance of population / n
    - Central Limit Theorem -> bigger n -> normal
- What is testing?
    - Using sample stats to make an inference about a pop. parameter
    - Set up as Ho: the Null - assumed true vs Ha: the alternative
    - Can you reject the Null and therefore accept the alternative

Regression Testing: Continued

- Type I - reject Ho when true (p-value) alpha (1%, 5%, 10%)
- Type II do not reject Ho when false (related to the Power – n)

- Typically use student t-tests (since pop. Sigma is never known)
- t = (Estimate - Ho: Value) / (standard error)
- What is prob of seeing that t if Ho is true (alpha - p-value)

- Statistical Significance versus Economic Significance
- Under real world multivariate conditions - use regression for testing
- Causality vs Association

Regression Testing: Continued

- Using Regressions with Categorical Variables
    - Dummy Variables
        - if categorical variable has k levels create k-1 dummies
        - remaining level captured by "base case"
    - Interactions
- Regression is useful, and widely used
- Need to understand and be able to work with the computer output
- Limitations to regression; be aware of assumptions behind the model
    - (cautions, means of checking)
- Importance of model building (Alfonso's is a great illustration);
    - have a justifiable model in mind

Session 5. How to incorporate uncertainty and risk to make better decisions? - Decision Trees

Case: Research at ICI

- What is a Decision Tree
  - decisions, decision nodes, states of nature, event nodes, and payoffs
  - strategic tool, what-if analysis,
  - competitive response, decision analysis
- Treeplan Excel Add-in

Decision Trees. Continued

- Measures:
  - risk neutral - expected value (repetition of similar projects)
  - non-risk neutral (one-time, bet-the farm)
    - maximax, maximin, minimax, minimin, utility functions
- certainty, uncertainty and risk
- dealing with uncertainty in your estimates

Session 6. How to better plan shipments, choose paths, schedule, etc? – Optimization

Case: Northwest Newsprint - shipping from mills to markets

- Optimization / Math Programming
  - use math to describe the problem to be solved
  - Linear Program (Simplex) with Constrained Resources
    - decision variables
    - objective function: min vs max
    - constraints <. <=, >=, >, =
    - Lego Example
    - feasible region

Optimization: Continued

    - integer vs continuous vs mixed
    - non-negativity constraints
  - Non-Linear Optimization
    - search technique for best answer
- Solver Excel Add-In
  - how to use it
  - how to interpret its output: Answer Report
  - rerun for sensitivity analyses

Session 7. How to estimate impact of proposed changes on processes that exhibit uncertainty? – Simulation

Case: Blood Mobile - time-based (dynamic) Monte Carlo Simulation

Case: Revisit Research at ICI (static) Monte Carlo Simulation

- descriptive and NOT prescriptive
- output of a simulation is a random variable
    - need multiple runs / repetitions
    - statistically test/compare results (t-tests)
- may need run-in period for time-based / dynamic case
- how to set number of replication or run-length

Simulation: Continued

- Excel MCSim Add-In
    - replications
    - define output cell to record
- Excel:
    - rand() generates 0 <= x < 1 uniform random variable
    - generating uniform(a,b): a + (rand() * (b-a))
    - generating normal(mu, sigma)
        - norm.inv(rand(),mu,sigma)
    - using vlookup() and rand() to generate discrete rvs
    - turning on and off "automatic recalc"
        - copy and paste special as values

Session 8. How to make better forecasts and therefore better plans? – Forecasting

Case: Northern Napa Valley Winery

- Forecasting Methods - Qualitative vs Quantitative
- Decomposition: Base (B), Trend (T), Seasonality (S), Random (R)
- Time Series Regression
- Autoregressive Models
    - regression using lagged Y-values as X vars
- Moving Averages - issues with trend and seasonality, oversmooth?
- Centred Moving Averages - compensates for trend
    - over a seasonal period can remove seasonality

Forecasting: Continued

- Exponential Smoothing
    - no T model: $S_t = S_{t-1} + \text{alpha} * (\text{Actual}_t - S_{t-1})$
        - choice of alpha (0.01 to 0.5, 0.20 typical)
    - Holts exponential smoothing for B and T
    - Winters exponential smoothing for B, T, and S*
- Forecasting when you have historic data.
    - Plot the data - examine the plot.
        - T present? S present?
        - range of seasonality swings constant S+ (additive)

Forecasting: Continued

- range of seasonal swings changing with level of data

S* (multiplicative)

- Choose the simplest model that does the job
  - B, no T, no S: exponential smoothing, simple moving averages
  - B, no T, S+: regression with dummy variables for seasons
  - B, no T, S*: mov avg over seas. period to remove seasonality and estimate the seasonal indexes (factors)
  - B, T, no S:
    - time-series reg. or Holt's 2 parameter smoothing approach
  - B, T, S+:
    - time-series regression with seasonal dummy variables

Forecasting: Continued

- B, T, S*:
  - centred mov. avg over seas. period to remove seasonality
  - calculate seasonal indexes (factors).
  - time-series regressions on deseasonalized data
  - forecast by re-seasonalizing your fitted data.
  - or Winter's 3 parameter smoothing approach

- Measures of Goodness
  - use holdout sample if possible
  - Mean Squared Error (MSE),
  - Mean Absolute Deviation (MAD)

Forecasting: Continued

- Mean Absolute Percent Error (MAPE)
- Max Absolute Error
- Max Absolute Percent Error

Session 9. How to Put it All Together?

Case: Park River – SO2 Reduction, the Scrubber Technology

- how to combine multiple methods and tools to solve a real issue
- how to analyze problems
- making good recommendations

Session 10. How to manage revenue and yield to improve profitability? - Revenue Management

Case: Yield Management at American Airlines Game

- Revenue Management and Optimum Dynamic Pricing
  - estimating the demand vs price curve
  - high data requirements
  - overbooking, yield management