Quantitative Research Methods

Jee Hwang, PhD Candidate in Economics
December 1, 2012

What is Quantitative Research?
The use of methods or techniques that employ numerical data for the purpose of investigating and understanding patterns of behavior or natural phenomena.

Characteristics of Quant. Research
- Objective, i.e. results tend to be independent from researcher bias.
- Provides numerical representations of behavior or phenomena being studied
- Emphasis on consistency (and efficiency) of results and not its validity
- Statistical probabilities are at the core of most (if not all) formal analyses

Quant. Research Methods & Uses
- Examining central tendency and general patterns in data
  - Moments (e.g. mean, variance)
  - Quantiles (e.g. 50 percentile or median)
  - Correlations (degree of association)
  - Histograms (distribution of variable)
  - Plots (e.g. scatter, connected lines, time plots)
Quant. Research Methods & Uses

- Examining differences between 2 or more groups
  - Student t test (difference in mean values)
  - Mann-Whitney U test (difference in median values)
  - Chi-Squared test (difference in proportions)
  - ANOVA (>2 groups, 1 variable)
  - ANCOVA (ANOVA with covariates)
  - MANOVA (>2 groups, 2 or more variables)
  - MANCOVA (MANOVA with covariates)
  - Bayesian approaches (the first three above are classical or “frequentist” approaches. The remaining four can be used in both Bayesian and classical contexts)

- Exploring relationships
  - Cluster Analysis (forms groups using a set of attributes)
  - Principal Component Analysis (identify patterns in data with many dimensions, e.g. facial recognition)
  - Factor Analysis (models the observed variables using its relation to unobserved/latent factors, e.g. intelligence, health)
  - Canonical Correlation (uses correlations between variables to derive linear combos, e.g. responses from two different personality tests)
  - Finite Mixture Models (regression based cluster analysis, e.g. identify and explain the behavior of the “sickly” vs. “healthy” amongst health care users)

- Testing a priori relationships and making predictions
  - Method of Least Squares (minimize the sum of squared residuals)
  - Maximum Likelihood Estimation (maximize the likelihood function)
  - The above are common techniques for conducting regression analyses

Types of data

- Nominal (e.g. gender, occupation, ethnicity/race)
- Dichotomous (e.g. Yes/No responses, gender)
- Ordinal (e.g. income range, education, level of satisfaction)
- Integer/count (e.g. visitation to park, class absences, number of natural disasters, accidents, and other events/outcome data)
- Continuous (e.g. age, height, weight)
Data for Quant. Research

- Types of data sets
  - Cross-sectional (many individuals or events in a single time period)
  - Time series (one individual or event across many time periods)
  - Panel (many individuals or events across many time periods)

Process of Quant. Research

- Formulate a specific question that you want answered (e.g. “What are the monetary benefits from advanced degrees?”)
- Identify the variables you need to conduct an investigation (wage, education, experience, other factors)
- Formulate a model with the variables that can be used to help answer your question, e.g. wage = f(education, experience, other factors)
- Hypothesis: wage and education share a positive relationship, i.e. benefits are positive.
- Use QA methods to test hypothesis
  - In classical inference, the statistical or “null” hypothesis being tested is: ‘no relationship’, against the alternative hypothesis: ‘positive relationship’ using p-values as a reference
  - In Bayesian inference, the ‘positive relationship’ hypothesis is tested directly against other hypotheses by evaluating its relative likelihood.

Tables allow a quick overview of data

<table>
<thead>
<tr>
<th>Highest Level of Education</th>
<th>OBS</th>
<th>Mean</th>
<th>Std. Dev</th>
<th>Median</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO HIGH SCHOOL DEGREE</td>
<td>3752</td>
<td>11.18</td>
<td>6.05</td>
<td>9.75</td>
<td>0.13</td>
<td>76.04</td>
</tr>
<tr>
<td>HIGH SCHOOL DIPLOMA</td>
<td>12846</td>
<td>15.23</td>
<td>8.79</td>
<td>13.02</td>
<td>0.00</td>
<td>90.00</td>
</tr>
<tr>
<td>SOME COLLEGE BUT NO DEGREE</td>
<td>8518</td>
<td>16.20</td>
<td>10.14</td>
<td>13.50</td>
<td>0.00</td>
<td>89.00</td>
</tr>
<tr>
<td>ASSOCIATES DEGREE(OCCUP/VOC)</td>
<td>2305</td>
<td>18.43</td>
<td>9.61</td>
<td>16.34</td>
<td>0.59</td>
<td>76.04</td>
</tr>
<tr>
<td>ASSOCIATES DEGREE(ACADEMIC)</td>
<td>2224</td>
<td>18.92</td>
<td>10.49</td>
<td>16.42</td>
<td>0.01</td>
<td>95.00</td>
</tr>
<tr>
<td>BACHELOR'S DEGREE</td>
<td>8876</td>
<td>26.28</td>
<td>15.71</td>
<td>22.20</td>
<td>0.00</td>
<td>99.00</td>
</tr>
<tr>
<td>MASTER'S DEGREE</td>
<td>3343</td>
<td>31.60</td>
<td>16.94</td>
<td>27.87</td>
<td>0.00</td>
<td>104.39</td>
</tr>
<tr>
<td>PROFESSIONAL DEGREE</td>
<td>657</td>
<td>41.43</td>
<td>22.17</td>
<td>36.99</td>
<td>2.20</td>
<td>95.96</td>
</tr>
<tr>
<td>DOCTORATES DEGREE</td>
<td>615</td>
<td>39.79</td>
<td>19.11</td>
<td>36.73</td>
<td>0.06</td>
<td>89.00</td>
</tr>
<tr>
<td>TOTAL</td>
<td>43136</td>
<td>19.72</td>
<td>13.80</td>
<td>15.46</td>
<td>0.00</td>
<td>104.39</td>
</tr>
</tbody>
</table>

Compiled using data from the October Current Population Survey (2005 to 2007)

Not everyone in this sample were employed
Here are the findings

- Based on the results, an individual would receive (on average) an additional:
  - $19.91/hr for a master’s degree
  - $28.25/hr for a doctoral degree
  - $30.48/hr for a professional degree

- Compared to someone who does not have a high school degree, i.e. benefits are positive.

- We should be cautious in taking the results too seriously since we ignored “other factors” such as occupation and gender, which have been shown in previous studies to also affect wages, i.e. results may be sensitive to model specification.

Some tips for Quant. Research

- When transforming variables in your data, be aware of the mathematical consequences of:
  - Taking logarithm of zero or negative values
  - Dividing by zero values
  - Taking the square root of negative values

- The above appear as missing values when attempted with most statistical packages. Try rescaling variables when possible.
Some tips for Quant. Research

- Check the validity of your results using common sense or findings from similar studies. This may be harder to do when conducting exploratory analysis.
- Using Qualitative Analysis in tandem can help in choosing the best approach and translating the results into information that can be applied in the real world, e.g. public policy.

Unfortunately, formal mixing of the two is not yet common at advanced levels perhaps because both methods require a certain degree of training.

Playing the Lottery

- One approach to help us decide is to evaluate what kind of return we could expect from the $2 investment given the overall odds of winning one of the nine possible prizes (1 in 31.8 or about 3%). For simplicity, let's assume that there are no other participants which would effect our expected return. In general the expected return from participating can be expressed as:

\[
\text{Expected Return} = \Pr(\text{win}) \times (\text{prize minus cost}) + (1 - \Pr(\text{win})) \times (\text{zero minus cost})
\]

\[
= \text{Gain(win)} + \text{Loss(not win)}
\]

If Expected Return greater than zero (>0) then play; if less than or equal to zero (≤0) then don't play.

Should You Play?

- Gain(win) can be found by taking the weighted sum of all prizes less the $2 cost, i.e. each prize minus $2 times the probability of winning the particular prize all added together.

\[
\text{Gain(win)} = \frac{1}{1,752,235,100} \times ($40 mil - $2) + \frac{1}{5,153,633} \times ($1 mil - $2) + \frac{1}{648,976} \times ($10K - $2) + \frac{1}{19,088} \times ($100 - $2) + \frac{1}{12,245} \times ($100 - $2) + \frac{1}{360} \times ($7 - $2) + \frac{1}{706} \times ($7 - $2) + \frac{1}{111} \times ($4 - $2) + \frac{1}{55} \times ($4 - $2)
\]

\[
= 0.53
\]
Should You Play?

- Loss (not win) is simply \(0.97 \times (\$0 - \$2) = -\$1.04\), i.e. the 97% chance you will not win times the loss of $2.

Your expected return is therefore:

\[
\text{Expected Return} = 0.53 - 1.94 = -1.41 \text{ ($< 0$)}
\]

So in the long run, you can expect to lose $1.41 from playing this lottery and therefore, you should not play.

If the secondary prizes are fixed, how large does the jackpot need to be in order for this lottery to be worth playing?
Answer: more than $287,728,678
Questions or Comments?