Controlling Survey Response Bias with Range Regression

John T. Chen*  Yuanting Zhang**
* Bowling Green State University
** U.S. Food and Drug Administration
Outline

• Errors and Biases in HDS (2002, 2008)
• Range regression
• Nutrition usage and BMI (2008), Nutrition usage and age (2002)
• Comparing different methods
• Results and Conclusion
Response bias

• Definition: Response bias is a type of cognitive bias which can affect the results of a statistical survey if respondents answer questions in the way they think the questioner wants them to answer rather than according to their true beliefs.

• Possible Causes:
  – wording of the question (leading questions)
  – Social desirability (responses may be biased toward what respondents believe is socially desirable
  – Convenience
  – Undetectable

• Meaning of response bias for this paper: cognitive bias + data entry error
Data and Research Question

• Data: Collected by the Food and Drug Administration periodically every 4-5 years, the Health and Diet Survey (HDS) was random-digit-dial (RDD) telephone survey that supports national estimates of consumer perceptions, knowledge, and behavior related to diet, nutrition, and health.

• Research question: We want to understand the relationships between food label usage and other factors (BMI for 2008 data, age for 2002 data)

  (BMI= [(body weight in lb)/ ((body height in inches)^2)]x703.07)
HDS food label usage questions
2002   2008

People tell us they use food product labels in many different ways. When you look at food labels, either in the store or at home, how often, if at all, do you use the labels in the following ways? Would you say you often, sometimes, rarely or never use the label?

a. To help you decide which brand of a particular food item to buy
b. To figure out how much of the food product you or your family should eat
c. To compare different food items with each other
d. To see if something said in advertising or on the package is actually true
e. To get a general idea of the nutritional content of the food
f. To see how high or low the food is in things like calories, salt, vitamins, or fat
g. To help you in meal planning
h. To see if there is an ingredient that you or someone in your family should avoid
Survey Response bias and error

• Cognitive bias/data entry error example: 2008 FDA HDS (n=2,584)

<table>
<thead>
<tr>
<th>Height</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>21 inches</td>
<td>162 pounds</td>
</tr>
<tr>
<td>48 inches</td>
<td>170 pounds</td>
</tr>
<tr>
<td>66 inches</td>
<td>615 pounds</td>
</tr>
<tr>
<td>64 inches</td>
<td>500 pounds</td>
</tr>
</tbody>
</table>

• Bias/error due to convenience example—some respondents may use the same value (e.g. 3) throughout the survey questions. In the HDS 2008 data, 297 out of 1584 have response errors (either the respondent reads label components in a consistent pattern or cognitive bias due to convenience)
Possible Methods for Dealing with Response bias

• Delete unreasonable data
• Impute missing/no response
• Use median to avoid extreme values

<table>
<thead>
<tr>
<th>Data</th>
<th>Sample mean</th>
<th>Sample median</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.5, 3, 5, 7, 8.5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>1.5, 3, 5, 7, 33.5</td>
<td>10</td>
<td>5</td>
</tr>
</tbody>
</table>

Outliers do not affect the sample median

• Group data by ranges
Pros and cons of grouping data

• Ranging partly reduces response bias, however, the chunked data may fall in few non-informative ranges.

<table>
<thead>
<tr>
<th>HDS 2008</th>
<th>weight</th>
<th>100 – 149 lbs</th>
<th>26%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>150 --199 lbs</td>
<td>40%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>200 – 249 lbs</td>
<td>21%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>HDS 2008</th>
<th>height</th>
<th>5 to less than 6 ft</th>
<th>75%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>6 to less than 7 ft</td>
<td>23%</td>
<td></td>
</tr>
</tbody>
</table>

• Chunked data is unable to serve the goal:
  – Identifying the pattern connecting food label usage frequency and obesity measuring by BMI
Mean Likert Scale

- Likert scale: ordering but no intrinsic meaning, *often* vs *sometimes* (1 < 2) but the difference between *often* and *sometimes* may not be the same as the difference between *sometimes* versus *rarely* (2 < 3).

- Mean likert scale: index of the purchasing behavior (frequency) of a group.

<table>
<thead>
<tr>
<th># of customers</th>
<th>Likert scores</th>
<th>Mean likert score</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>6 -often, 4-sometimes</td>
<td>(6+2x4)/10=1.4</td>
</tr>
<tr>
<td>10</td>
<td>5 -often, 5-sometimes</td>
<td>(5+2x5)/10=1.5</td>
</tr>
</tbody>
</table>
Use of Nutrition label and BMI
Regression Methods

• Normal linear regression, $y=ax+b+\varepsilon$
  implausible model assumption for categorical data

• Logistic regression, $\text{logit}(p)=ax+b$
  modeling likelihood for association, BMI error,
  exact likert scale may also contain survey errors

• Range regression: group mean scale=$ay+b+\varepsilon$
  group mean scale: average out error or bias;
  range $y$: group out error/bias;
  $\varepsilon$: normal model
  modeling two primary endpoints directly.
Range regression

- **Definition:** Range regression models the mean responses in terms of the range of the explanatory variable. It reduces the survey response error/bias by ranging the values of the explanatory variable and taking the average of the dependent variable within each range. By the law of large number and the central limit theorem, the normality assumption is valid.

- **References:**
Reducing Response Bias by Averaging

• Taking the mean of nutrition scores to average out response biases of a few individuals. For example, rolling a die 1000 times, 20 bias/error readings will not affect the overall average.

\[ \lim_{n \to \infty} \left( \frac{n-20}{n} \times \text{true mean} + \frac{20}{n} \times \text{bias} \right) = \text{true mean} \]

• Using ranges for extreme errors; using average to reduce the impact of response errors and errors of putting a response into a wrong range.
Range regression---step 1

- Rescale the variable using ranges to reduce the impact of response biases. For example,

```
if BMI<=18.5    then BMIclass=17.25;
if 18.5<BMI<=25 then BMIclass=21.75;
if 25<BMI<=30   then BMIclass=27.5;
if 30<BMI<=35   then BMIclass=32.5;
if 35<BMI<=40   then BMIclass=37.5;
if 40<BMI       then BMIclass=42.5;
```
Range regression--- step 2

• Relationship: computing the mean response within each subgroup

• For each value of \textit{BMI class}, find the mean of getting general idea of the nutritional content of the food

<table>
<thead>
<tr>
<th>BMI class</th>
<th>17.25</th>
<th>21.75</th>
<th>27.5</th>
<th>32.5</th>
<th>37.5</th>
<th>42.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean score of getting idea of the nutritional content of the food</td>
<td>1.519</td>
<td>1.544</td>
<td>1.589</td>
<td>1.619</td>
<td>1.627</td>
<td>1.754</td>
</tr>
</tbody>
</table>
Range regression--- step 3

• Modeling the mean-responses versus the median of each range.

• Linear regression -> *BMIclass* as the explanatory variable and *mean nutrition score* as the dependent variable

| variable   | Parameter estimate | Standard error | t-value | Pr > |t| |
|------------|--------------------|----------------|---------|------|---|
| intercept  | 1.36478            | 0.04621        | 29.54   | < 0.0001 |
| BMIclass   | 0.00817            | 0.00149        | 5.50    | 0.0053 |
BMI range and mean likert score

R-square=0.8830 (BMI range--proxy of various related factors attributable to nutrition label usage behavior;
Baseline mean likert score=1.36478+0.00817x17.25
Range regression Result
Interpretation

DV: use food label to get a general idea of the nutritional content of the food, Mean likert value (1-often 2-sometimes 3-rarely 4-never);
IV: BMI ranges

As shown in the diagram, the observed and the predicted values agree with each other. On average, customer groups who are in the lower BMI ranges shopping with more frequent attention on nutrition content of the food. The linear pattern is significant.
The outcome of theqqplot strongly suggests the plausibility of the normal model. This is due to the central limit theorem and the law of large number in each range where the mean likert score is computed.
Correcting Convenience-type of Errors

- Assume that 100 respondents in the HDS 2008 data (2584 respondents), incorrectly reported their purchasing frequency as rarely (response bias of respondents who simply put the value 3 for every survey question).

- Range regression output hardly changed. (consistent)

<table>
<thead>
<tr>
<th>Parameter Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Variable</strong></td>
</tr>
<tr>
<td>Intercept</td>
</tr>
<tr>
<td>BMIclass (100 errors)</td>
</tr>
<tr>
<td>BMIclass (original)</td>
</tr>
</tbody>
</table>
Example of Convenience error among 100 respondents

Convenience error: assume 100 respondents answered “1” throughout regardless of true value.

Convenience error: assume 100 respondents answered “2” throughout.

Convenience error: assume 100 respondents answered “3” throughout.

Convenience error: assume 100 respondents answered “4” throughout.
Central Limit Theorem

Let $X_1 \ldots X_n \ldots$ be a sequence of i.i.d. random variables with $E(X_i) = \mu$ and $0 < \text{Var}(X_i) = \sigma^2 < \infty$. Define $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$. Let $G_n(x)$ denote the cumulative distribution function of $\sqrt{n}(\bar{X}_n - \mu)/\sigma$. Then for any $x$, $-\infty < x < \infty$,

$$\lim_{n \to \infty} G_n(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \, dy$$

- Counting processes for the occurrence of response errors, martingale convergence theorem
## Comparisons

<table>
<thead>
<tr>
<th>Linear regression</th>
<th>Logistic regression</th>
<th>Range regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likert scores vs BMI score</td>
<td>Logit of risk vs BMI score</td>
<td>Mean likert score vs BMI score</td>
</tr>
<tr>
<td>Normality invalid for discrete data</td>
<td>Binomial model</td>
<td>Valid normality noises</td>
</tr>
<tr>
<td>Including response error in BMI</td>
<td>Including erroneous BMI</td>
<td>Average out data errors; range out response errors</td>
</tr>
<tr>
<td>Correlation coefficient not interpretable</td>
<td>Odds ratio for association</td>
<td>Correlation coefficient for linear relationship</td>
</tr>
</tbody>
</table>

Selected Findings

• HDS 2008, the relation between food label usage and BMI ranges (individual effect)

<table>
<thead>
<tr>
<th>stat\mean</th>
<th>nutrition</th>
<th>brand</th>
<th>calories</th>
<th>accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-square</td>
<td>0.8830</td>
<td>0.8545</td>
<td>0.5174</td>
<td>0.7096</td>
</tr>
<tr>
<td>p-value</td>
<td>0.0053</td>
<td>0.0084</td>
<td>0.1071</td>
<td>0.0353</td>
</tr>
</tbody>
</table>

• HDS 2002, the relation between food label usage and age ranges (individual effect)

<table>
<thead>
<tr>
<th>Stat\mean</th>
<th>brand</th>
<th>avoidance</th>
<th>nutrition</th>
<th>plan</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-square</td>
<td>0.6982</td>
<td>0.7421</td>
<td>0.2578</td>
<td>0.7536</td>
</tr>
<tr>
<td>P-value</td>
<td>0.0098</td>
<td>0.0060</td>
<td>0.1990</td>
<td>0.0052</td>
</tr>
</tbody>
</table>
More than two variables

• For independent variables, taking ranges for each variable, then combining all the ranges into strata to define a set of new “ranges”;  
• For dependent variable, taking average within each stratum to formulate the range regression.
Conclusion

Range regression averages out the response errors of the dependent variable and ranges out response errors of the explanatory variable. It is an effective and robust method to identify the linear pattern governing two variables in survey data.