THE STATISTICAL PROBLEM OF RELATING DIET AND DISEASE

Raymond J. Carroll (http://stat.tamu.edu/~carroll)

Collaboration with

L. Freedman (Bar–Ilan University)

V. Kipnis (National Cancer Institute)
OUTLINE

- My motivation: fat intake and breast cancer
- The controversy
- Statistical Models
- Data
- Potential implications
- Proposed study
THE CONTROVERSY

- For some time, nutrient fat intake has been suspected of being a promoter of breast cancer.
  - *Animal studies* show this clearly.

- *Ecological studies* show it clearly as well.
  - Countries with lower aggregate nutrient fat intake have lower rates of breast cancer.
  - Japanese women who move to the U.S. have higher rates of breast cancer, and higher nutrient fat intakes.

- *Case–control studies*, when pooled in a meta–analysis, also show a nutrient fat intake effect on breast cancer.
THE CONTROVERSY

- These arguments, plus other considerations, have led the National Institutes of Health to undertake the Women’s Health Initiative (WHI).

- The WHI is a huge, expensive clinical trial, one arm of which is to compare two groups prospectively over 10 years for incidence of breast cancer:
  - “Treated” Women who are counseled to undertake a “healthy diet”, one component of which is a massive reduction in nutrient fat intake (from 35% calories coming from fat for a typical American to 20% or less).
The nutrition epidemiology group at Harvard has bitterly opposed the nutrient fat intake hypothesis. This means that they also see as largely a waste of funds the nutrition experiment part of the WHI.

One piece of evidence they cite is that no study that follows women prospectively has ever found a statistically significant effect of nutrient fat intake on breast cancer.

Prospective, nonrandomized studies have their own problems, but they are not subject to the potential confounding of ecological studies. They are also not subject to the clear biases inherent in case–control studies.
THE CONTROVERSY

• Hunter, et al. recently published (1995) a meta-analysis of large prospective cohort studies of fat intake and breast cancer, and found no statistically significant relationship.

  • 7 Studies, many thousands of women combined
  • No individual study or the combined data found a statistically significant relationship.

• More discouraging, ignoring statistical significance in a study of hundreds of thousands of women, the estimated effects of nutrient fat intake were tiny.

  • Estimated relative risks were very small, e.g., 1.10 for % Calories from Fat.

• In contrast, international comparisons show a fat effect (for % Calories from Fat, a relative risk of at least 1.86)
WHAT WE ARE ABOUT

- What I have not yet mentioned is how “nutrient fat intake” is measured.

- Our group has set about the task of trying to understand the statistical properties of the usual instruments for measuring nutrient fat intake.
  - It is this research I want to talk about.

- Studies at the NCI are now underway which will help resolve the controversy.
SOME BACKGROUND IN MEASURING DIET

- The goal is to measure “usual nutrient intake”, which I call $T$ (“truth”).
  - This can only be defined operationally, as the average daily intake of a nutrient over a fixed period of time, e.g., one year.
  - $T$ is unobservable. It is impossible to monitor non–confined populations for their diet.
- The simplest method to ”measure” usual intake is the food frequency questionnaire, called $Q$.
  - It is cheap, fast, and simple.
  - Important: the typical prospective nonrandomized study will have many thousands of participants.
  - The Nurses Health Study has 100,000 participants.
SOME BACKGROUND IN MEASURING DIET

- It is generally thought that other instruments are less biased (but more variable) (call them $F$)
  
  - Food diaries/ weighted food records
  - 24-hour food recalls
  - Biomarkers

- Using many recalls/records or biomarkers in not feasible ($$$).
  
  - Thus the FFQ is used in essentially every large epidemiological study.

- We thus need to answer the question: how well can the FFQ measure usual nutrient intake $T$, and how much does it matter that the FFQ does not measure $T$?
EFFECTS OF MEASUREMENT ERROR

- Since the FFQ does not measure usual nutrient intake exactly, we say that the FFQ is an error–prone instrument, subject to measurement error.

- Usual intake cannot be measured exactly.

- What is the effect of this measurement error?

- Typically, the effect of measurement error is to cause two things:
  - A bias in estimating the nutrient–disease relationship
  - Loss of power, sometimes profound loss of power.

- I’ll illustrate graphically with some simulated data in which I will relate what happens when I regress a quantitative outcome (body mass, blood cholesterol) on an error–prone nutrient measure (FFQ).
Figure 1: **Observed (solid) and True (open) Data and lines**
EFFECTS OF MEASUREMENT ERROR

- As the graph shows, the effect of measurement error is to make the observed relationship between two variables weaker than it really is.

- Thus, in a nutrition context, the effect of errors of ascertainment is to weaken the observed relationship between disease outcome and measured nutrient intake.

- There is a large statistics literature dealing with how to recapture the true relationship.
THE CONTROVERSY REVISITED

- I will propose models for measurement error, and apply them to what are called calibration studies.
  - These have 2 or more FFQ and 2 or more recalls, diaries or records.
- I have access to calibration data from a number of sources.
  - No other group has access to such a broad set of studies, other than Harvard.
- All have at least two FFQ’s given at widely separated times.
- They differ in whether they have recalls or records.
- They differ in whether the recalls/records are or are not given in conjunction with the FFQ.
THE CONTROVERSY REVISITED

- The data sets:
  - Women’s Health Trial Vanguard Study (Records) 
    \[ n = 86 \]
  - American Cancer Society (recalls) \( n = 186 \)
  - AARP (NCI) (recalls) \( n \approx 1,800 \)
  - Nurses Health Study (records) \( n = 168 \)
  - Finnish Smokers Study (recalls)
  - Polyp Prevention Trial (recalls)

- I’ll only use data from the first 4.

- I’ll consider % Calories from Fat.
  - Similar results for \( \log(\text{Total Fat}) \), \( \log(\text{Energy}) \), \( \log(\text{Protein}) \).
Generally, people agree on certain properties of the FFQ (an old model of ours is used).

- **flattened slope**: People who eat large amounts of fat will under-report fat intake, and vice-versa.

- **measurement error**: If you give a person an FFQ multiple times, you will not get the same answers.

- **Equation error or Person–specific bias**: Two people with the same nutrient fat intake will not fill out the FFQ the same way: even if you administer the FFQ multiple times and average out the measurement errors, they won’t agree exactly.

The model then is

\[
\text{FFQ} = \underbrace{\text{flattened slope}} + \underbrace{\text{Person–specific bias}} + \underbrace{\text{measurement error}}
\]
BASIC MODEL

\[ \text{FFQ} = \text{flattened slope} + \text{Person–specific bias} + \text{measurement error} \]

- In symbols, this becomes that the \( j \)th measurement on the \( i \)th person is

\[ Q_{ij} = \beta_0 + \beta_1 T_i + r_i + \epsilon_{ij} \]

- Paradoxically, Food Recalls/Records (F) are called *reference instruments* and reference instruments are thought to be unbiased measures of *usual intake*.

\[ \text{Records/Recalls} = \text{Usual Intake} + \text{measurement error} \]

\[ F_{ij} = T_i + U_{ij} \]
The relative risk is a common epidemiological term. For a relatively rare disease like breast cancer, it signifies the increase in probability of disease when going from one level of usual intake to a higher level of usual intake. In Hunter, et al., the estimated relative risk is 1.10. This means that there is a (non–statistically significant) 10% increase in risk of the disease. Thus, if the risk of disease in the lower group is 5%, and the relative risk is 1.10, then the risk of the disease in the higher group is $5\% \times 1.10 = 5.50\%$. 
SOME TERMINOLOGY

• If we could observe usual nutrient intake, then the large prospective studies could have been used to estimate the true relative risk.

• We can only get an observed relative risk from the FFQ.

• We want to understand how the observed relative risk using the FFQ is related to the true relative risk for usual nutrient intake.
SOME TERMINOLOGY

• Here is a remarkable statistical fact. It was discovered simultaneously and independently in 1987 by Gleser, Whittemore & Keller, Stefanski and myself and Rosner, Willett & Spiegelman.

• Suppose that you run a large prospective study using the FFQ, and you find an observed relative risk of $R_{obs}$.

• Then the true relative risk $R_{true}$ is very nearly

$$R_{true} = R_{obs}^{1/\gamma},$$

where $\gamma = \text{slope of the regression of usual intake } T \text{ on the FFQ.}$

• This reduces the statistical problem to estimating $\gamma$. 
SUMMARY OF BASIC MODEL

\[ FFQ = \text{flattened slope} + \text{Person-specific bias} \]
\[ + \text{measurement error} \]

\[ \text{Records/Recalls} = \text{Usual Intake} + \text{measurement error} \]

\[
Q_{ij} = \beta_0 + \beta_1 T_i + r_i + \epsilon_{ij} \\
F_{ij} = T_i + U_{ij}
\]

- Then the true relative risk \( R_{true} \) is very nearly

\[
R_{true} = R_{obs}^{1/\gamma};
\]

\[
\gamma = \text{slope of the regression of usual intake } T \text{ on the FFQ.}
\]

- \( \gamma \) is the slope in the regression of \( T \) on \( Q \).

- It can be estimated by maximum likelihood (method of moments).
RESULTS OF THE BASIC MODEL

- We fit the basic model to the four main calibration data sets.

- We assumed that the observed relative risk is 1.10, and estimated what the true relative risk would be.

  WHTVS: 1.32
  Nurses Health Study: 1.41
  AARP: 1.21
  ACS: 1.27

- This is much in keeping with the results of Hunter, et al.

  - Even accounting for errors and biases in the FFQ, the true relative risk is not very large.
WHAT’S MISSING FROM THE BASIC MODEL

\[ FFQ = \text{flattened slope} + \text{Person–specific bias} + \text{measurement error} \]

\[ \text{Records/Recalls} = \text{Usual Intake} + \text{measurement error} \]

- You do not have to be in love with formulae to ask the question: *where is the flattened slope and the person–specific bias in records/recalls?*
  - Should the model not be:
    \[ FFQ = \text{flattened slope} + \text{Person–specific bias} + \text{measurement error} \]
    \[ \text{Records/Recalls} = \text{flattened slope} + \text{Person–specific bias} + \text{measurement error} \]
  - We think yes. More on why later, but for now, let’s entertain the new model.
THE NEW MODEL

- The new model has flattened slope, Person–specific bias and measurement error for the FFQ and for records/recalls.

- We hypothesize that the Person–specific biases are probably correlated, if they exist.

- There is a complication:
  - The four calibration data sets do not allow the estimation of all parameters (statisticians call this non–identifiability).

- We are thus reduced to a sensitivity analysis.
  - These are still useful, because if anything interesting is found, then it suggests a need for further studies.
For this sensitivity analysis, today we assume that

- No flattened slope in recalls/records
- Equal variances for Person–specific biases
- We vary the correlation between the Person–specific biases.
- Models fit using maximum likelihood.
Estimated True Relative Risk

Figure 2: True Relative Risks from various data sets when the usual model is applied and a relative risk of 1.10 is observed.
Figure 3: **True Relative Risks from various data sets when the new model is applied and a relative risk of 1.10 is observed.**
EFFECTS ON RELATIVE RISK

- Old model and New model estimates of true relative risk when the correlation in the **Person–specific biases** is 0.70.

<table>
<thead>
<tr>
<th>Study</th>
<th>Old Model</th>
<th>New Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>WHTVS</td>
<td>1.32, ≥ 2.0</td>
<td>2.0</td>
</tr>
<tr>
<td>NHS</td>
<td>1.41, ≥ 2.0</td>
<td>2.0</td>
</tr>
<tr>
<td>AARP</td>
<td>1.21, ≥ 2.0</td>
<td>2.0</td>
</tr>
<tr>
<td>ACS</td>
<td>1.27, 1.40</td>
<td></td>
</tr>
</tbody>
</table>

- What we see is that measurement error may be hiding a real risk effect, if the correlation between the **Person–specific biases** is at all high.

- Put another way: **if true**, of course the large prospective studies cannot find a statistically significant relationship between nutrient fat intake and breast cancer.

  - The instrument, the FFQ, is too imprecise to be of much use.
BIOMARKERS

- There are two intakes which can, in principle, be measured without a flattened slope and without Person-specific biases.
  - Energy (caloric intake), via double-labelled water (DLW)
  - Protein, via urinary nitrogen.
- There are a number of small studies \((n \approx 10)\) on DLW which suggest the flattened slope, but because our model is new the suggestion is not definitive.
  - No one will pass on their data for DLW to us to check.
- There is no biomarker for fat. We can only infer from energy and protein.
EVIDENCE FOR THE NEW MODEL

- Recently, we received data from the MRC on 160 women: four weighed food records, four urinary nitrogen, one Willett–style FFQ.

- We did model checking, and compared our model to all the other models that have been proposed in the literature.

  - **Our model is not statistically significantly different from any more complex published model.**

  - **All simpler models included in ours provide a statistically significantly worse fit to the data.**

  - **Our model had the highest AIC and BIC among all models.**
### Table 1: Nested models

<table>
<thead>
<tr>
<th>Model Name</th>
<th>$-2 \times \text{LL}$</th>
<th>df</th>
<th>p</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unstructured</td>
<td>-1223.3</td>
<td>104</td>
<td>-</td>
<td>507.2</td>
<td>224.7</td>
</tr>
<tr>
<td>Plummer &amp; Clayton</td>
<td>-1173.2</td>
<td>56</td>
<td>0.43</td>
<td>530.6</td>
<td>378.5</td>
</tr>
<tr>
<td>Our model</td>
<td><strong>-1134.4</strong></td>
<td>19</td>
<td>0.39</td>
<td><strong>548.2</strong></td>
<td><strong>496.6</strong></td>
</tr>
<tr>
<td>Our model, uncorrelated person–specific biases</td>
<td><strong>-1123.7</strong></td>
<td>18</td>
<td>0.17</td>
<td>543.9</td>
<td>495.0</td>
</tr>
<tr>
<td>Standard Model</td>
<td>-1050.1</td>
<td>15</td>
<td>&lt; 0</td>
<td>510.1</td>
<td>469.3</td>
</tr>
</tbody>
</table>

**EVIDENCE FOR THE NEW MODEL**
EVIDENCE FOR THE NEW MODEL

- There is a flattened slope in weighed food records.
- There is Person–specific bias.
- There is a correlation between the two Person–specific biases, ≥ 0.50.
- We used the MRC data, starting from an observed relative risk of 1.10.
  - Old model: true relative risk estimate of 1.35
  - New model: true relative risk estimate of 1.80.
  - The new model suggests that studies that were designed to have 80% power for detecting an effect should be 2.5 times larger.
- These results thus suggest that the FFQ is a far less precise instrument for measuring protein–disease relationships than had been thought previously.
EVIDENCE FOR THE NEW MODEL

- We have recently obtained protein biomarker data from a series of European centers.
- These results suggest that the FFQ is even worse than we had seen in the MRC data.
FUTURE EVIDENCE FOR THE NEW MODEL

- The NCI is undertaking a large \(n = 400\) study of DLW and urinary nitrogen.

- We will thus be able to understand the ability of the FFQ to measure energy–disease, protein–disease and protein density (% Calories from Protein) – disease relationships.

- The AARP study may have such a component.
  - So too may the EPIC study in Europe.
CONCLUSIONS

\[ \text{FFQ} = \text{flattened slope} + \text{Person–specific bias} + \text{measurement error} \]

\[ \text{Records/Recalls} = \text{flattened slope} + \text{Person–specific bias} + \text{measurement error} \]

- It is possible that the lack of an observed dietary fat–breast cancer relationship is due to the poor qualities of the FFQ.

- Future studies (including WHI) will help us understand this, although there is no biomarker for fat.

- If the FFQ is as poor an instrument as suggested by the MRC data, and we hope it is not, then there are major implications for nutritional epidemiology, where the FFQ is the instrument of choice.