THE STATISTICAL PROBLEM OF RELATING DIET AND DISEASE

IMS Special Invited Paper: Indianapolis

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

- I have been fortunate to have many collaborators on this project, but especially I want to mention
  - L. Freedman (Bar–Ilan University)
  - V. Kipnis and D. Midthune (National Cancer Institute)
- See my web pages for relevant papers and detailed references to the work of others
SOME BACKGROUND

- My invitation for the SIP asked for a talk on "Applied Statistics or some interdisciplinary area"

- Be careful what you ask for!

- My first published paper was "Sequential density estimation at an unknown point”, in Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete

- My most recent tentatively accepted one is "The relationship of DNA adduct levels in proximal and distal regions of the colon"
SOME BACKGROUND

- Nutritional epidemiology has attracted interest from many skilled statisticians
  - The problem is interesting statistically, as I hope to convince you
  - The public health and policy implications are enormous
  - and you can talk about your research at parties
- The problem has been a, perhaps the major factor in the development of what is now a large literature on measurement error in nonlinear models.
- Instead of reviewing this progress, or describing recent technical advances, I will talk about the problem motivating all this work.
SOME BACKGROUND

- Here is a partial list of recent statistical researchers in nutritional epidemiology (with apologies in advance!)

  - **Northeast:** B. Rosner, D. Spiegelman, L. Ryan, T. Tosteson, E. Demidenko, K–Y Liang

  - **Northwest:** R. Prentice, C. Y. Wang, L. Shepard, M. Pepe

  - **California:** D. Stram, D. Thomas

  - **Fly–by states:** RJC, N. Wang, W. Fuller, S. Nusser, A. Carriquiry, W. Jiang

  - **Europe:** N. Day, D. Clayton

  - **NIH:** M. Gail, S. Wacholder
SOME BACKGROUND

- Many of the statistical methodology papers are not published in statistical journals.

- The reasons are many

- The outrageous time from submission to publication
  - 2+ years in statistics journals
  - About 1/2 the time in AJE

- The field really is interdisciplinary, and we want to reach the people who will use the methods.

- So, why have so many people worked on the problem?
OUTLINE

- Major issue: contradictory results for nutrient fat intake and breast cancer
- Methods for measuring nutrient intake
- Basic and simple statistical ideas
- Data available
- Old models (Ford Pintos) and new models (Ford Explorers)
- Results, methods, etc.
For some time, nutrient fat intake has been suspected of being a promotor 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 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 CONTROVERSY

- So, where is the controversy?

- **No study that follows women prospectively has ever found a statistically significant effect** of nutrient fat intake on breast cancer.

- One version of NHANES actually found the opposite!

- **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


  • 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)
THE CONTROVERSY

- A great deal of the recent research in the area has focused on trying to reconcile the contradictory results.

- Specifically, there is great interest in trying to understand whether the prospective studies measure diet in a way that diet–disease relationships have reasonable statistical power.

- To understand the issue, you have to understand
  - How is dietary intake measured?
  - What statistical methods/models are used?
THE MEASUREMENT OF 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.
MEASURING DIET

- It is *generally* thought that other instruments are less biased (but more variable) (call them $F'$)

- Multiple **Food diaries/ weighted food records**

- Multiple **24-hour food recalls**

- Multiple **Biomarkers**
  - The list is limited!
  - **Protein**: urinary nitrogen
  - **Caloric intake (energy)**: doubly-labelled water
  - Very, very expensive
MEASURING DIET

- Using many recalls, records or biomarkers is generally not feasible

- Thus the FFQ is used in essentially every large epidemiological study.

- We 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.

• In nutrition, measurement errors are large and consequential

• Usual intake cannot be measured exactly.

• 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.
Effects of Measurement Error

Figure 1: **Observed (solid) and True (open) Data and lines**
BASIC STATISTICAL RESULTS

- $Y$ = disease outcome indicator

- Logistic model for usual intake:

$$\text{pr}(Y = 1|T) = H(\beta_0 + \beta_1 T)$$

- This leads to a logistic model for the FFQ:

$$\text{pr}(Y = 1|Q) = H(\beta_0 + \lambda \beta_1 Q)$$

$\lambda = \text{reliability}$

$= \text{slope in regression of T on Q}$

- Often called regression calibration

- $\lambda < 1$ is called attenuation

- Reduces the problem to estimating the reliability
BASIC STATISTICAL RESULTS

\[
\begin{align*}
\text{pr}(Y = 1|T) &= H(\beta_0 + \beta_1 T) \\
\text{pr}(Y = 1|Q) &= H(\beta_0 + \lambda \beta_1 T)
\end{align*}
\]

\[\lambda = \text{slope in regression of T on Q}\]

- Relative risk attenuated
  \[
  \begin{align*}
  \text{Observed RR} &= (\text{True RR})^\lambda \\
  \text{True RR} &= (\text{Observed RR})^{1/\lambda}
  \end{align*}
  \]

- Sample sizes for fixed power generally go up like
  \[
  n \propto \frac{1}{\lambda^2}
  \]
DESIGNING STUDIES

- The typical way to design large prospective studies is through a **calibration study**

- A small number of people fill out the FFQ and a reference instrument (diary/record/recall)

- From this \( \lambda \) is estimated and the sample size is calculated.

- Obvious question: **does it matter what reference instrument you use in setting sample size?**
ESTIMATING RELIABILITY

\[ \lambda = \text{reliability} \]

\[ = \text{slope in regression of } T \text{ on } Q \]

- We need to estimate the reliability

- Assume we have a reference instrument unbiased for usual intake

  \[ F = T + U \]

  - Error U is uncorrelated with the FFQ

  \[ \lambda = \text{slope in regression of } T \text{ on } Q \]

  \[ \lambda = \text{slope in regression of } F \text{ on } Q \]

- The literature typically assumes that records, diaries, recalls, biomarkers are reference instruments

- **Does it matter which you choose?**
Figure 2: **Reliability estimates when the Reference Instrument is a protein biomarker or a diary/record/recall. Note how the use of diary/record/recall underestimates the effect of measurement error.**
Figure 3: Sample size factors when the Reference Instrument is a protein biomarker or a diary/record/recall. If the factor = 3 for example, a study designed for 80% power using a diary/record/recall should really by increased 3 times in sample
WHAT IS GOING ON?

- The basic upshot of the previous slides is that if you use a diary/record/recall as a reference instrument, then you
  - Underestimate the effects of measurement error
  - Underpower your study
- It is of immense interest to understand why this is happening.
  - What makes diaries/records/recalls underperform as reference instruments for protein intake?
BASIC MODEL

- General agreement on certain properties of the FFQ.
  - 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

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

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

- Since they are also based on self–report, the same model can be entertained for diaries/records/recalls

  - Flattened slope
  - Person–specific bias, and this correlated with that of the FFQ
  - Measurement error
We add to the previous discussion that the biomarker is a true reference instrument.

- No **Flattened slope** = unbiasedness
- No **Person–specific bias**
- Classical **Measurement error**
BASIC MODEL

- Here is the model, in symbols

\[
Q_{ij} = \beta_0 + \beta_1 T_i + r_i + \epsilon_{ij}
\]
\[
F_{ij} = \alpha_0 + \alpha_1 T_i + s_i + \eta_{ij}
\]
\[
B_{ij} = T_i + U_{ij}
\]
\[
\text{corr}(r_i, s_i) \neq 0
\]

- The nutrition literature assumes \( \alpha_0 = 0, \alpha_1 = 1, s_i = 0, \text{corr}(r_i, s_i) = 0 \)

- These assumptions are testable!
SOME DATA

- We have access to 2 data sets with multiple biomarkers
  - One has a standard 7–day diary
  - One has a weighed food record
- I will display the results for the diary
  - The results for the weighed food record are similar
- For both data sets, our model
  - Not stat. significantly different from any model in which it is nested
  - Stat. significantly different from any of our sub-models (including standard model)
  - Has highest AIC and BIC
Figure 4: AIC for various models in the literature. Our model has highest value, also of BIC.
Figure 5: Likelihood ratio significance levels for model fit against nested hypotheses.
IS THE BIOMARKER UNBIASED?

- We have assumed that the biomarker urinary nitrogen has
  - No Flattened slope = unbiasedness
  - No Person–specific bias
  - Classical Measurement error

- This is not a testable hypothesis from the data we have.

- Person–specific bias seems not a serious issue

- Unbiasedness is considered ”common” knowledge among nutritionists
  - Standard textbooks
  - (very) Small feeding studies
WHAT IS LEFT?

- Nutritional epidemiology regresses disease outcomes on a nutrient and \textit{energy = caloric intake simultaneously}
  
  - Effects of measurement error complex
  - Correlated variables, correlated errors, can cause almost anything to happen

- Our studies do not have the energy biomarker (doubly labelled water) measured. If they did:
  
  - there is still a ”reliability” for the nutrient
  - our models can be generalized

- An NCI study will be completed in about a year, and a European study in about 2 years
  
  - Then we will ”know” whether our univariate results generalize to the bivariate regression
WHAT ABOUT FAT INTAKE?

- There is no biomarker for fat

- Energy = Fat + Protein + Carbs + Alcohol

- Not too unreasonable to suppose that if FFQ’s are much less precise than previously suspected for protein, then they are much less precise for fat
WHAT ABOUT BODY MASS?

• A body mass index (BMI) effect on attenuation has been hypothesized
  • The idea is that lean people are less biased and variable, and their reliability should therefore be higher.

• We use a Prentice–idea and split data into 3 BMI ter-tiles and compute the reliability

• We have also developed a **lovely nonparametric regression local estimating equation** estimation and testing methodology and theory
  • Under review since February
  • You may see it in 2 years

• The results are simple, see next graph
Figure 6: *Average reliability of 11 cohorts of men and women, for body mass index groups.*
CONCLUSIONS

- Most nutritional epidemiology studies are designed on the basis of and correct for measurement error using diaries/records/recalls as reference instruments

- These methods, like the FFQ, and based on self–report

- Our results for protein confirm that these have two major undesirable properties

  - **Flattened slope**
  
  - **Person–specific bias**, and this correlated with that of the FFQ ($\rho(r, s) > 0.6$)

- Univariate results suggest that diaries/records/recalls

  - Overestimate the reliability of the FFQ
  
  - Understate the sample size needed to get 80% power by factors of 2–9
CONCLUSIONS

- The univariate results then suggest that the FFQ is a particularly insensitive instrument

- **Worst Case:** In a major UK study, with 100 women, we estimate the reliability to be less than 10%
  - This means that an observed relative risk of 1.10 is consistent with a true relative risk of 3.10

- **Worst Case:** In the same UK study, we estimate that the sample size required for fixed power is 9 times larger than what would be designed if the diary were the reference instrument

- If these results hold up in the bivariate (protein + energy) analysis, profound implications for nutritional epidemiology
  - Can only disease–nutrient relationships with huge relative risks be detected?