Estimating the Probability of Compliance with Physical Activity Guidelines from a Bayesian Perspective, Revisited

Dave Osthus and Bryan Stanfill

Center for Survey Statistics and Methodology, Iowa State University

March 9, 2012
Previously on PAMS
Adding Survey Statistics to Analysis
Model 2.0
Results
Model Assessment
PAMS is a two-stage survey designed to obtain information on physical activity patterns of

- Adult women and men (21-70)
- Hispanic and African American populations
Eligible adults from Black Hawk, Dallas, Marshall, and Polk counties.

Eligibility criteria:

- Not pregnant or lactating
- Able to complete an interview in either English or Spanish
- No physical limitation or medical restrictions preventing the adult from participating in physical activity
- Reside in a household with a land line phone
Stratification and Sampling Design

Stratification

- Counties were partitioned into a low minority and high minority tract group
- Done to improve chances of recruiting African Americans and Hispanics
- Total of 8 strata

Two stage sampling within strata

- First stage: Systematic sampling of households (sampling frame: listing of land line phone numbers)
- Second stage: SRS of adults within household sampled
Data Collection Process

- Selected eligible individuals wore SenseWear Monitor for 24 hours on two non-consecutive days.
- Minute by minute energy expenditure information is recorded.
- Data collection was evenly distributed over two years, partitioned into eight quarters.
Here is a break-down of our data by gender for BMI and Age.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Statistic</th>
<th>Females (n=269)</th>
<th>Males (n=201)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Q1</td>
<td>45.00</td>
<td>37.00</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>54.00</td>
<td>48.00</td>
</tr>
<tr>
<td></td>
<td>Q3</td>
<td>62.00</td>
<td>60.00</td>
</tr>
<tr>
<td>BMI</td>
<td>Q1</td>
<td>25.76</td>
<td>25.91</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>29.62</td>
<td>28.89</td>
</tr>
<tr>
<td></td>
<td>Q3</td>
<td>36.02</td>
<td>32.95</td>
</tr>
</tbody>
</table>
And a break-down of the ethnic make-up by strata of our data:

<table>
<thead>
<tr>
<th>Minority Tract</th>
<th>County</th>
<th>Strata</th>
<th>Other</th>
<th>African-American</th>
<th>Hispanic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>Black Hawk</td>
<td>1</td>
<td>84</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Dallas</td>
<td>2</td>
<td>7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Marshall</td>
<td>3</td>
<td>15</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Polk</td>
<td>4</td>
<td>100</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>High</td>
<td>Black Hawk</td>
<td>5</td>
<td>46</td>
<td>19</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Dallas</td>
<td>6</td>
<td>60</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Marshall</td>
<td>7</td>
<td>24</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Polk</td>
<td>8</td>
<td>80</td>
<td>18</td>
<td>9</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>416</td>
<td>40</td>
<td>14</td>
<td></td>
</tr>
</tbody>
</table>
Def: A **metabolic equivalent of a task**, or MET, is a concept expressing energy cost as a multiple of resting metabolic heart rate. 1 MET is considered one’s resting metabolic rate.

- For example, engaging in an activity with a MET value of 3 would require three times the energy that person consumes at rest.

Def: **MET-minutes** are simply the MET levels multiplied by the minutes engaged in that physical activity.

- For example, someone who engaged in a physical activity with a MET value of 3 for 30 minutes engaged in a $3 \times 30 = 90$ MET-minute physical activity.
The Center for Disease Control states on their website that adults need at least:

1. 150 minutes of moderate-intensity aerobic activity (3-6 METs) every week.

2. 75 minutes of vigorous-intensity aerobic activity (over 6 METs) every week.

3. An equivalent mix of moderate- and vigorous-intensity aerobic activity every week.

Key: These minutes of physical activity of specified intensity levels must come in at least 10 continuous minute intervals, known as bouts.
Operational Definition of CDC guidelines:

- On 5 or more days a week, individuals should engage in 90 MET-minutes of physical activity at an intensity of at least 3 METs observed during bouts of at least 10 minutes.
Goal:

Our goal is to develop a statistical model that can plausibly “produce” the observed data while at the same time, aides us in answer two primary questions. They are:

1. What proportion of Iowans comply with the CDC PA guidelines?

2. What covariates are associated with compliance?
Future Work from the Past

There are several things we planned to do from here:

- Use survey statistics
- Add random variables for multiple observations
- Add more covariates to our data set
How do we deal with the two stage sampling, i.e., the systematic sampling of households by phone number in stage one and the simple random sampling of an eligible individual within that house in stage two? Here are a few approaches.
**Source:** Ch. 14 of Gelman and Hill’s *Data Analysis Using Regression and Multilevel/Hierarchical Models* (2007)

**Objective:** Estimate $P($Voting for Bush$)$ for each state.

**Sample:** National polling data is collected and then corrections for nonresponse based on sex, ethnicity, age and education are made.

**Model:** Bayesian logistic regression model, with sex, ethnicity, age and education as covariates.

**Accounting for Sample Design:** Post-stratify based on variables included in the model.
Source: *Small Area Inference for Binary Variables in the National Health Interview Survey* by Malec et. al. (1997)

Objective: Estimate $P($Visited the doctor in the last 12 months$)$ for subpopulations within each state.

Sample: Two stage design. Households selected in first stage. Individuals within households selected in second stage.

Model: Hierarchical Bayesian logistic model.

Accounting for Sample Design: Post-stratify data based on variables included in the model.
• **Source:** *Physical Activity in the United States Measured by Accelerometer* by Troiano, Dodd, et. al. (2007)

• **Objective:** Estimate prevalence of adherence to physical activity recommendations.

• **Sample:** Complex, multistage probability design.

• **Model:** Beta-binomial model.

• **Accounting for Sample Design:** Use survey weights and post-stratification after model was fit.
Our Course of Action

1. Fit model.

2. Apply the model to estimate proportion of compliance to CDC guidelines.

3. Account for complex sample design through incorporation of survey weights followed by post-stratification.
Bayesian Mixture Model

Let $Y_{kij(i)} \equiv Y_{kij}$, $k = 1, 2, \ldots, 8$, $i = 1, 2, \ldots, n_k$, where $n_k$ is the number of sampled individuals from stratum $k$, $j(i) \in \{1, J(i)\}$, where $J(i) = 1$ if person $i$ provided 1 day of data and $J(i) = 2$ if person $i$ provided 2 days of data. Then $Y_{kij}$ is a random variable associated with the MET-minutes of individual $i$ on day $j$ in stratum $k$. We will define $Y_{kij}$ as follows:

$$Y_{kij} = (1 - V_{kij}) + V_{kij}W_{kij}$$

where,

$$V_{kij} \sim \text{Bernoulli}(p_{ki})$$

and,

$$W_{kij} \sim \text{Shifted Gamma}(\mu_{ki}, \phi_k)$$
More specifically, if $V_{kij} \sim \text{Bernoulli}(p_{ki})$, then

$$
\begin{align*}
    f_1(v_{kij} | p_{ki}) &= p_{ki}^{v_{kij}}(1 - p_{ki})^{1 - v_{kij}} \\
    E(V_{kij}) &= p_{ki} \\
    \text{Var}(V_{kij}) &= p_{ki}(1 - p_{ki}) \\
    \text{logit}(p_{ki}) &\equiv x_{ki}^T \beta_{1k} + \gamma_{1ki} \\
    \Rightarrow f_1(v_{kij} | \beta_{1k}, \gamma_{1ki}) &= \left( \frac{\exp(x_{ki}^T \beta_{1k} + \gamma_{1ki})}{1 + \exp(x_{ki}^T \beta_{1k} + \gamma_{1ki})} \right)^{v_{kij}} \left( 1 - \frac{\exp(x_{ki}^T \beta_{1k} + \gamma_{1ki})}{1 + \exp(x_{ki}^T \beta_{1k} + \gamma_{1ki})} \right)^{1 - v_{kij}} \\
    \pi(\beta_{1k}) &\propto 1 \\
    \gamma_{1ki} &\sim \text{iid N}(0, \sigma_{1p}^2) \\
    \sigma_{1p}^2 &\sim \text{Unif}(0, 10)
\end{align*}
$$
For the shifted gamma, if $W_{kij}^* \sim \text{Gamma}(\mu_{ki}, \phi_k)$ and

$$f_2^*(w_{kij}^* | \mu_{ki}, \phi_k) = \exp \left\{ \phi_k \left[ - \frac{1}{\mu_{ki}} w_{kij}^* - \log(\mu_{ki}) \right] + c(w_{kij}^*, \phi_k) \right\}$$

$$c(w_{kij}^*, \phi_k) = \phi_k \log(\phi_k) - \log(\Gamma(\phi_k)) + (\phi_k - 1) \log(w_{kij}^*)$$

$$E(W_{kij}^*) = \mu_{ki}$$

$$\text{Var}(W_{kij}^*) = \frac{\mu_{ki}^2}{\phi_k}$$

$$\log(\mu_{ki}) \equiv x_{ki}^T \beta_{2k} + \gamma_{2ki}$$

$$\Rightarrow f_2^*(w_{kij}^* | \beta_{2k}, \phi_k, \gamma_{2ki}) = \exp \left\{ \phi_k \left[ - \frac{1}{\exp(x_{ki}^T \beta_{2k} - \gamma_{2ki})} w_{kij}^* - x_{ki}^T \beta_{2k} - \gamma_{2ki} \right] \right\}$$

$$+ \ c(w_{kij}^*, \phi_k)$$
Then define $W_{kij} = W_{kij}^* + \delta$, ($\delta = 30$ in our case), where $\delta$ is a fixed, known constant.

$$f_2(w_{kij} | \beta_{2k}, \phi_k, \gamma_{2ki}) = \exp \left\{ \phi_k \left[ - \frac{(w_{kij} - \delta)}{\exp(x_{ki}^T \beta_{2k} - \gamma_{2ki})} - x_{ki}^T \beta_{2k} - \gamma_{2ki} \right] + c(w_{kij}, \phi_k) \right\}$$

$$c(w_{kij}, \phi_k) = \phi_k \log(\phi_k) - \log(\Gamma(\phi_k)) + (\phi_k - 1)\log(w_{kij} - \delta)$$

$$E(W_{kij}) = \mu_{ki} + \delta$$

$$Var(W_{kij}) = \frac{\mu_{ki}^2}{\phi_k}$$

$$\pi(\beta_{2k}) \propto 1$$

$$\phi_k \overset{iid}{\sim} \text{Unif}(0, 100)$$

$$\gamma_{2ki} \overset{iid}{\sim} \text{N}(0, \sigma_{2p}^2)$$

$$\sigma_{2p}^2 \sim \text{Unif}(0, 10)$$
Posterior is proportional to prior times likelihood:

\[ \pi(\beta_1, \beta_2, \gamma_1, \gamma_2, \phi, | y) \propto \pi(\beta_1, \beta_2, \gamma_1, \gamma_2, \phi) L(y | \beta_1, \beta_2, \gamma_1, \gamma_2, \phi) \]

where we assume,

\[ \pi(\beta_1, \beta_2, \gamma_1, \gamma_2, \phi) = \pi(\beta_1) \pi(\beta_2) \pi(\gamma_1) \pi(\gamma_2) \pi(\phi) \]

and,

\[ L(y | \beta_1, \beta_2, \gamma_1, \gamma_2, \phi) = \prod_{k=1}^{8} \prod_{i=1}^{n_k} \prod_{j=1}^{J(i)} L_{kij}(\beta_{1k}, \beta_{2k}, \gamma_{1ki}, \gamma_{2ki}, \phi_k | y_{kij}) \]

\[ L_{kij}(\beta_{1k}, \beta_{2k}, \gamma_{1ki}, \gamma_{2ki}, \phi_k | y_{kij}) = f_1(v_{kij} | \beta_{1k}, \gamma_{1ki}) I(y_{kij} = 0) + f_2(w_{kij} | \beta_{2k}, \phi_k, \gamma_{2ki}) I(y_{kij} \neq 0) \]
We consider these covariates: Age, Gender, BMI, Race, Education, House Hold Size (Trost et. al. (2002))

WinBugs was used to sample from each joint posterior using a standard Metropolis algorithm.

Chains were checked to confirm the limiting distributions were reached and desirable mixing was achieved.
What covariates are significant?

- No covariates are significant in all strata
- Stratum 2 has 12 observations, hence the large uncertainty
- Gender and BMI prove to be the most informative covariates
Estimating the Proportion of Compliance, $p$

For $l$ from $l = 1, 2, \ldots, L$, do the following:

1. For individual $i$ in stratum $k$, do the following:
2. Sample $\beta_{1ki}(m), \beta_{2ki}(m), \gamma_{1ki}(m), \gamma_{2ki}(m)$, and $\phi_{ki(m)}^{(l)}$ from the posterior distribution for $m = 1, 2, \ldots, 7$.
3. Simulate $y_{ki(m)}^{(l)}$, based on the parameter draws from the previous step.
4. Calculate $p_{ki(m)}^{(l)}$, where $p_{ki(m)}^{(l)} = 1$ if $y_{ki(m)}^{(l)} \geq 90$, and 0 otherwise.
5. Calculate $p_{ki}^{(l)}$, where $p_{ki}^{(l)} = 1$ if $\sum_{m=1}^{7} p_{ki(m)}^{(l)} \geq 5$, and 0 otherwise.
6. Calculate $p_{k}^{(l)} = \frac{1}{N_{k}} \sum_{i=1}^{n_{k}} w_{ki} p_{ki}^{(l)}$, where $N_{k}$ is the population size of stratum $k$, $n_{k}$ is the sample size in stratum $k$, and $w_{ki}$ is the post-stratified survey weight for individual $i$ in stratum $k$.
7. Calculate $p^{(l)} = \frac{1}{N} \sum_{k=1}^{8} N_{k} p_{k}^{(l)}$, where $N$ is the population size. $p^{(l)}$ is the estimated proportion of compliance for the $l^{th}$ simulated draw.
Estimating \( p \) continued.

Then \( p \), the proportion of weekly compliance with CDC recommendations, is estimated as \( \frac{1}{L} \sum_{l=1}^{L} p^{(l)} \).

- **Simple Random Sample (red):** \( \hat{p} = 0.389 \) with 95% CI (0.362, 0.419)
- **Survey Weighted (blue):** \( \hat{p} = 0.402 \) with 95% CI (0.358, 0.446)
A quantity we’re interested in is the proportion of zeros

We compare an estimated distribution for that quantity against the observed quantity to get a Bayes p-value of 0.25

We expected this to not be near 0 or 1 because this quantity is a feature of our model
We compare the estimated distribution of % people between 30 and 90 MET-Mins against the empirical estimate and calculate a Bayes p-value of 0.10.

Similarly, we compare % people with more than 90 MET-Mins against our data and calculate a Bayes p-value of 0.39.
We estimated the distribution of the maximum MET-Mins generated by our model and compared it against the observed maximum.

The smallest simulated maximum is larger than the observed maximum giving a Bayes p-value of 0.00.

This feature of the model needs reevaluation and could potentially be solved by removing insignificant covariates with large uncertainty.
Though a greater breadth of covariates were included in our model, it appears only Gender, BMI, and possibly Age are significantly related to MET-Mins.

The estimated proportion of compliance with CDC guidelines for eligible adults in these four counties, \( p \), was estimated as 0.402 (0.358, 0.446).

Had we not taken into account the sampling design, we would have underestimated the uncertainty in our estimate of \( p \).
Thanks you for listening.