Estimating the Distribution of Usual Daily Energy Expenditure

Dave Osthus

Center for Survey Statistics and Methodology, Iowa State University

February 21, 2014
• The Physical Activity Measurement Survey (PAMS) was designed to obtain information on physical activity patterns of
  • Eligible adult males and females
  • Hispanic and black populations (limited sample size)

• Eligible adults were
  • Between 21-70
  • Capable of physical activity
  • Competent to be interviewed
PAMS cont.

- Fall 2009 - Summer 2011

- Individuals were sampled from four counties in Iowa: Marshall, Black Hawk, Dallas and Polk.
PAMS cont.

- Multi-stage stratified probability sample design
  - 2 strata per county (high/low minority)
  - Systematic sampling of households
  - Simple random sampling eligible adults within households

Table: Selected demographic characteristics of the PAMS sample

<table>
<thead>
<tr>
<th></th>
<th>Males</th>
<th>Females</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>561</td>
<td>786</td>
</tr>
<tr>
<td>Median Age</td>
<td>49</td>
<td>53</td>
</tr>
<tr>
<td>IQR Age</td>
<td>21</td>
<td>18</td>
</tr>
<tr>
<td>Range Age</td>
<td>21-70</td>
<td>20-71</td>
</tr>
<tr>
<td>Count (%) Hispanic</td>
<td>25 (4.5)</td>
<td>28 (3.6)</td>
</tr>
<tr>
<td>Count (%) black</td>
<td>41 (7.3)</td>
<td>73 (9.3)</td>
</tr>
</tbody>
</table>

- Education, townsize, household size, occupation, etc.
• Survey weights \( (w_i) \) adjust for unequal selection probabilities and nonresponse

• 12 post-strata
  • Age group (younger than 50, 50 or older)
  • Race (black, Hispanic, other)
  • Sampling strata ("high" or "low" minority groups)
What was asked of sampled individuals?

- Individuals wore SenseWear Monitor (monitor) for 24 hours.
  - Measures heart rate, heat flux, accelerometer (motion, steps), etc.
  - Proprietary algorithm outputs energy expenditure (EE), MET-minutes, time spent at different intensity levels, etc.

- 24-hour physical activity recall (PAR) administered via phone the following day.
  - Compendium of Physical Activity
  - Calculate energy expenditure, MET-minutes, time spent at different intensity levels, etc.

- Non-consecutive days
PAMS cont.

![Diagram showing EE (kcal/d) for Females with PAR and Monitor categories.](image-url)
PAMS cont.
Goal and Methodology

• Goal: Estimate the distribution of usual, daily EE

• Methodology

1. Notation and assumptions
2. Transformation
3. Estimation of mean effects
4. Group-level measurement error model
5. Population-level measurement error model
6. Estimation of usual daily EE distribution
Notation and Assumptions

- For \( i = 1, 2, \ldots, n \) and \( j = 1, 2 \), let
  - \( X_{ij} \) represents the measurement of EE for individual \( i \) on day \( j \) from the reference instrument (monitor)
  - \( Y_{ij} \) represents the measurement of EE for individual \( i \) on day \( j \) from the self-report instrument (PAR)

- \((X_{i1}, X_{i2}, Y_{i1}, Y_{i2})\) is the complete set of measurements for individual \( i \)

- Assume \( i' \neq i \) independent

- Assume reference instrument unbiased for true, daily EE
Transformation

- Transform measurements to approximate normality
  - PAMS: log transformation
  - \( x_{ij} = \log(X_{ij}) \), \( y_{ij} = \log(Y_{ij}) \)
Mean Effects

- Daily EE measurements reflect nuisance factors such as
  - Day-of-week
  - Time-in-sample
  - Seasonality
- Daily EE measurements reflect demographic factors such as
  - Age
  - Race/Ethnicity
  - Gender
Mean Effects cont.

Let

\[ x_i = Q_i \gamma_x + \epsilon_{x,i}, \quad (1) \]

where

- \( x_i = (x_{i1}, x_{i2})' \)
- \( Q_i \) is a 2 by \( k \) matrix
- \( \epsilon_{x,i} = (\epsilon_{x,i1}, \epsilon_{x,i2})' \sim (0, \Sigma_{\epsilon\epsilon}) \) is the error vector for \( x_i \)
A generalized least squares (GLS) estimator of the parameter vector $\gamma_x$ is

$$\hat{\gamma}_x = \left( \sum_{i=1}^{n} Q_i' W_i Q_i \right)^{-1} \sum_{i=1}^{n} Q_i' W_i x_i,$$

(2)

where

- $W_i$ is a $2 \times 2$ weight matrix
PAMS: Mean Effects

- $x_i = (0.5(x_{i1} + x_{i2}), 0.5(x_{i1} - x_{i2}))'$

- Covariates of $Q_i$
  - Time-in-sample
  - Weekend
  - Age
  - AQ
  - Black

- AQ is a quadratic function of age with “linear ends”

- Additional covariates considered include season, Hispanic, education, town size, and household size
PAMS: Mean Effects

- Choices of $W_i$
  - $w_i I$
  - $\hat{\Sigma}_{\epsilon \epsilon}^{-1}$

- DuMochel and Duncan (1983)
  - $H_0$: $E\{(Q^*_i'Q^*_i)^{-1}Q^*_i'x^*_i\} = E\{(Q^*_i'w_iIQ^*_i)^{-1}Q^*_i'w_ix^*_i\}$
  - $x^*_i = \hat{\Sigma}_{\epsilon \epsilon}^{-1/2}x_i$
  - $Q^*_i = \hat{\Sigma}_{\epsilon \epsilon}^{-1/2}Q_i$

- Conclusion
  - Mean: weighted estimate
  - Non-intercept: GLS estimate
PAMS: Mean Effects

<table>
<thead>
<tr>
<th></th>
<th>Monitor</th>
<th>PAR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>SE</td>
</tr>
<tr>
<td>Mean* (10^1)</td>
<td>77.99</td>
<td>0.07</td>
</tr>
<tr>
<td>Time-in-Sample** (10^3)</td>
<td>-7.96</td>
<td>3.12</td>
</tr>
<tr>
<td>Weekend** (10^3)</td>
<td>-9.65</td>
<td>9.03</td>
</tr>
<tr>
<td>Age** (10^3)</td>
<td>-3.69</td>
<td>0.62</td>
</tr>
<tr>
<td>AQ** (10^4)</td>
<td>-2.27</td>
<td>0.59</td>
</tr>
<tr>
<td>Black** (10^2)</td>
<td>9.53</td>
<td>2.17</td>
</tr>
</tbody>
</table>

*Survey weighted mean and survey standard error. **Coefficients and standard errors from estimated GLS.
Group-Level Measurement Error Model

The model for the deviations from the regression for the reference instrument \((x_{g,ij,d})\) for group \(g = 1, 2, \ldots, G\) is

\[
x_{g,ij,d} = t_{g,i} + d_{g,ij} + u_{g,ij},
\]

where

- \(t_{g,i}\) is deviation of usual daily EE from pop. mean for individual \(i\)
- \(d_{g,ij}\) is deviation from individual \(i\)'s usual daily EE on day \(j\)
- \(u_{g,ij}\) is measurement error for reference instrument
- \(t_{g,i} \sim (0, \sigma^2_{g,t}), d_{g,ij} \sim (0, \sigma^2_{g,d}), u_{g,ij} \sim (0, \sigma^2_{g,u})\) mutually uncorrelated
- \(\text{Var}(x_{g,ij,d}) = \sigma^2_{g,t} + \sigma^2_{g,d} + \sigma^2_{g,u}\)
The model for the deviations from the regression for the self-report instrument \((y_{g,ij,d})\) for group \(g\) is

\[
y_{g,ij,d} = \beta_g(t_{g,i} + d_{g,ij}) + r_{g,i} + e_{g,ij}
\]  

(4)

where

- \(\beta_g\) is the slope of the relationship between self-report EE deviation and monitor EE deviation
- \(r_{g,i}\) is individual \(i\)'s deviation from the group-level mean
- \(e_{g,ij}\) is the remaining measurement error in the self-report for individual \(i\) on day \(j\)
- \(r_{g,i} \sim (0, \sigma_{g,r}^2), e_{g,ij} \sim (0, \sigma_{g,e}^2)\), mutually uncorrelated with each other and with \(t_{g,i}, d_{g,ij},\) and \(u_{g,ij}\)
- \(\text{Var}(y_{g,ij,d}) = \beta_g^2(\sigma_{g,t}^2 + \sigma_{g,d}^2) + \sigma_{g,r}^2 + \sigma_{g,e}^2\)
Group-Level Measurement Error Model cont.

- \((\bar{x}_{g,i,d}, \bar{y}_{g,i,d}, x_{g,i1,d} - x_{g,i2,d}, y_{g,i1,d} - y_{g,i2,d})\)

- The method of moments was used to derive estimators for 
  \(\theta_g = (\beta_g, \sigma_{g,t}^2, \sigma_{g,d}^2, \sigma_{g,u}^2, \sigma_{g,e}^2, \sigma_{g,r}^2)'\)

- The variance of \(\hat{\theta}_g\) was computed via a delete-one sub group jackknife procedure (Kott, 2001)
PAMS: Group-Level Measurement Error Model

- Parameters in deviation may vary with age

- $4 (= G)$ age groups:
  - Group 1: $< 40$
  - Group 2: $40 – 49$
  - Group 3: $50 – 59$
  - Group 4: $\geq 60$
PAMS: Group-Level Measurement Error Model cont.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
<th>Group 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_g$</td>
<td>0.83 (0.11)</td>
<td>0.75 (0.08)</td>
<td>0.74 (0.07)</td>
<td>0.61 (0.06)</td>
</tr>
<tr>
<td>$100\sigma^2_{g,t}$</td>
<td>2.40 (0.52)</td>
<td>2.46 (0.40)</td>
<td>2.22 (0.34)</td>
<td>2.30 (0.30)</td>
</tr>
<tr>
<td>$100\sigma^2_{g,d}$</td>
<td>1.14 (0.30)</td>
<td>0.31 (0.24)</td>
<td>0.85 (0.20)</td>
<td>0.65 (0.18)</td>
</tr>
<tr>
<td>$100\sigma^2_{g,u}$</td>
<td>0.87 (0.40)</td>
<td>1.19 (0.31)</td>
<td>0.89 (0.26)</td>
<td>0.42 (0.23)</td>
</tr>
<tr>
<td>$100\sigma^2_{g,e}$</td>
<td>0.80 (0.21)</td>
<td>0.88 (0.17)</td>
<td>0.78 (0.14)</td>
<td>0.73 (0.12)</td>
</tr>
<tr>
<td>$100\sigma^2_{g,r}$</td>
<td>1.43 (0.32)</td>
<td>0.73 (0.25)</td>
<td>0.86 (0.21)</td>
<td>0.76 (0.19)</td>
</tr>
</tbody>
</table>

- Standard errors smoothed across age groups
- $\sigma_t^2/(\sigma_t^2 + \sigma_d^2 + \sigma_u^2) \approx 0.6$
- $(\sigma_r^2 + \sigma_e^2)/\sigma_u^2 \approx 2$
- Trend with respect to age in all variance parameters and $\beta$
Population-Level Measurement Error Model

- One possible population-level model is

\[ x_{ij} = \mu_x + z_{1,i} \alpha_1 + t_i + d_{ij} + u_{ij}, \]
\[ y_{ij} = \mu_y + z_{2,i} \alpha_2 + \beta(t_i + d_{ij}) + r_i + e_{ij}, \] (5)

- \( \mu_x + z_{1,i} \alpha_1 \) analogous to \( \mathbf{Q}_i \gamma_{x,i} \) in (1)
- \( \mu_y + z_{2,i} \alpha_2 \) analogous to \( \mathbf{Q}_i \gamma_{y,i} \)
- \( (t_i, d_{ij}, u_{ij}, e_{ij}, r_i) \) are assumed to be mutually independent
- \( \beta \) and variances can be functions of explanatory variables
Population-Level Measurement Error Model cont.

- Estimation can be carried out in two steps

1. Perform a regression of the type (1) and compute the second moments of the deviations from the mean regression

2. GLS regression where $i$-th individual's vector of observations is

$$
\begin{bmatrix}
0.5(x_{i1} + x_{i2}) \\
0.5(y_{i1} + y_{i2}) \\
(0.5(x_{i1,d} + x_{i2,d}))^2 \\
(0.5(y_{i1,d} + y_{i2,d}))^2 \\
0.25(x_{i1,d} + x_{i2,d})(y_{i1,d} + y_{i2,d}) \\
(x_{i1,d} - x_{i2,d})^2 \\
(y_{i1,d} - y_{i2,d})^2 \\
(x_{i1,d} - x_{i2,d})(y_{i1,d} - y_{i2,d})
\end{bmatrix}
$$
PAMS: Population-Level Measurement Error Model

\[ x_{ij} = \mu_x + z_{1,i} \alpha_1 + t_i + d_{ij} + u_{ij}, \]
\[ y_{ij} = \mu_y + z_{2,i} \alpha_2 + \beta(t_i + d_{ij}) + r_i + e_{ij}, \]

\[ (\mu_x, \alpha'_1) = \gamma_x \]
\[ (\mu_y, \alpha'_2) = \gamma_y \]
\[ \text{Var}(t_i) = \sigma_t^2 \]
\[ \text{Var}(d_{ij}) = \sigma_d^2(1 + \lambda_1 \text{Age}_i) \]
\[ \text{Var}(u_{ij}) = \sigma_u^2(1 + \lambda_1 \text{Age}_i) \]
\[ \text{Var}(e_{ij}) = \sigma_e^2(1 + \lambda_1 \text{Age}_i) \]
\[ \text{Var}(r_i) = \sigma_r^2(1 + \lambda_1 \text{Age}_i) \]
**Table:** Variance components, $\beta$ and $\lambda_1$ are estimated based on deviations from the mean effects regression via a generalization of the EGLS procedure described in the group-level section.

<table>
<thead>
<tr>
<th></th>
<th>Est</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta$</td>
<td>0.71</td>
<td>0.04</td>
</tr>
<tr>
<td>$100\sigma^2_t$</td>
<td>2.32</td>
<td>0.16</td>
</tr>
<tr>
<td>$100\sigma^2_d$</td>
<td>0.74</td>
<td>0.10</td>
</tr>
<tr>
<td>$100\sigma^2_u$</td>
<td>0.78</td>
<td>0.12</td>
</tr>
<tr>
<td>$100\sigma^2_e$</td>
<td>0.82</td>
<td>0.06</td>
</tr>
<tr>
<td>$100\sigma^2_r$</td>
<td>0.92</td>
<td>0.09</td>
</tr>
<tr>
<td>$100\lambda_1$</td>
<td>-1.75</td>
<td>0.32</td>
</tr>
</tbody>
</table>

- $\beta$ is less than 1
- Negative $\lambda_1 \rightarrow$ larger variance components for younger individuals
• Usual daily EE for individual $i$ is

$$E(x_{ij}|t_i) = \mu_x + z_{1,2,i} \alpha_{1,2} + t_i$$

• $z_{1,2,i} \alpha_{1,2}$ is the non-nuisance effects

• Given a one-day PAR observation, $y_{i,obs}$, an estimator of usual daily EE for individual $i$ is

$$\hat{EE}_i = \mu_x + z_{1,2,i} \alpha_{1,2} + \beta^{-1}(y_{i,obs} - \mu_y - z_{2,i} \alpha_{2})$$

$$E(\hat{EE}_i|t_i) = \mu_x + z_{1,2,i} \alpha_{1,2} + t_i$$

$$V_{PAR}(\hat{EE}_i - t_i) = \sigma^2_{d,i} + \beta^{-2}(\sigma^2_{r,i} + \sigma^2_{e,i})$$
• For a 55 year old individual
  • \( \hat{V}_{\text{PAR}}\{\hat{EE}_i - t_i\} = 0.0397 \)
  • \( \hat{V}_{\text{monitor}}\{\hat{EE}_i - t_i\} = 0.0143 \)

• PAR variance is almost 3 times larger than monitor
Estimated Usual Daily EE Distribution

- Goal: Estimate the usual daily EE distribution
  - $\mu_x + z_{1,2,i} \alpha_{1,2} + t_i$
- Not assuming $t_i$ normally distributed
- Estimate distribution $t_i$ via an extrapolation technique
Construct three quantities with different levels of error variance

\[
\epsilon_{3,i,j} = [\hat{\sigma}_t^2 + \hat{\sigma}_{d,i}^2 + \hat{\sigma}_{u,i}^2]^{-0.5} x_{i,j,d}
\]

\[
\epsilon_{2,i} = [\hat{\sigma}_t^2 + 0.5(\hat{\sigma}_{d,i}^2 + \hat{\sigma}_{u,i}^2)]^{-0.5} [0.5(x_{i,1,d} + x_{i,2,d})]
\]

\[
\epsilon_{1,i} = [\hat{\sigma}_t^2 + \hat{V}(\hat{t}_i)]^{-0.5} \hat{t}_i
\]
Estimated Usual Daily EE Distribution cont.

\( \epsilon_{1,i} \) is based on the model

\[
\begin{pmatrix}
\bar{x}_{i,d} \\
\bar{y}_{i,d}
\end{pmatrix} = \begin{pmatrix} 1 \\
\beta \end{pmatrix} t_i + \begin{pmatrix} \psi_{x,i} \\
\psi_{y,i} \end{pmatrix},
\]

where

\[
(\psi_{x,i}, \psi_{y,i})' = [\bar{d}_i + \bar{u}_i, \beta \bar{d}_i + r_i + \bar{e}_i]'
\]

\[
(\psi_{x,i}, \psi_{y,i})' \sim (0, V_{\psi\psi})
\]

\[
V_{\psi\psi} = \begin{pmatrix}
0.5(\sigma^2_{d,i} + \sigma^2_{u,i}) & 0.5\beta \sigma^2_{d,i} \\
0.5\beta \sigma^2_{d,i} & 0.5\beta^2 \sigma^2_{d,i} + \sigma^2_{r,i} + 0.5\sigma^2_{e,i}
\end{pmatrix}
\]

\[
\hat{t}_i = [(1, \beta)V_{\psi\psi}^{-1}(1, \beta)']^{-1}(1, \beta)V_{\psi\psi}^{-1}(\bar{x}_{i,d}, \bar{y}_{i,d})'
\]

\[
V(\hat{t}_i) = [(1, \beta)V_{\psi\psi}^{-1}(1, \beta)']^{-1}
\]
Estimated Usual Daily EE Distribution cont.

\[ \varepsilon_1: Q-Q Plot \]

\[ \varepsilon_2: Q-Q Plot \]

\[ \varepsilon_3: Q-Q Plot \]
Estimated Usual Daily EE Distribution cont.

$$\epsilon_{k,i} = b_{0,k} + b_{1,k}q + b_{2,k}z_q$$

- $k = 1, 2, 3$
- $q$ is the standard normal quantile

$$z_q = 0 \quad \text{if } q < 1$$
$$= (q - 1)^2 \quad \text{if } 1 \leq q \leq 2$$
$$= 1 + 2(q - 2) \quad \text{if } 2 \leq q$$
Estimated Usual Daily EE Distribution cont.

Epsilons vs. Normal Quantiles
Red Lines are Fitted Quantile Functions

Epsilon_1

Epsilon_2

Epsilon_3

Epsilon Quantiles

Standard Normal Quantiles

-2  0  2

-2  0  2
• What would $b_0$, $b_1$, and $b_2$ be were the error variance 0?
Estimated Usual Daily EE Distribution cont.

\[ Q(\epsilon_0, i) = Q\left(\frac{t_i}{\sigma_t^2}\right) = b_{0,0} + b_{1,0}q + b_{2,0}z_q \equiv q_{\epsilon_0} \]
If $\tilde{X} \sim N(0,1)$,

$$P(\tilde{X} \leq q) = P\left(\frac{t_i}{\sigma_t^2} \leq q_{\epsilon_0}\right)$$

$$= P\left(\mu_x + z_{1,2,i} \alpha_{1,2} + t_i \leq \mu_x + z_{1,2,i} \alpha_{1,2} + \sigma_t^2 q_{\epsilon_0}\right)$$

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>0.05</th>
<th>0.25</th>
<th>0.5</th>
<th>0.75</th>
<th>0.95</th>
<th>.99</th>
</tr>
</thead>
<tbody>
<tr>
<td>35</td>
<td>No</td>
<td>7.61</td>
<td>7.75</td>
<td>7.85</td>
<td>7.95</td>
<td>8.11</td>
<td>8.26</td>
</tr>
<tr>
<td>35</td>
<td>Yes</td>
<td>7.71</td>
<td>7.85</td>
<td>7.95</td>
<td>8.05</td>
<td>8.21</td>
<td>8.36</td>
</tr>
<tr>
<td>65</td>
<td>No</td>
<td>7.46</td>
<td>7.60</td>
<td>7.70</td>
<td>7.80</td>
<td>7.96</td>
<td>8.11</td>
</tr>
<tr>
<td>65</td>
<td>Yes</td>
<td>7.55</td>
<td>7.69</td>
<td>7.79</td>
<td>7.89</td>
<td>8.05</td>
<td>8.20</td>
</tr>
</tbody>
</table>
Estimated Usual Daily EE Distribution cont.

Estimated PDF of Usual Energy Expenditure

- NonBlack35
- Black35
- NonBlack65
- Black65

Energy Expenditure (kcal)
Density
Variable

Energy Expenditure (kcal)
Future Work

• Jackknife variance for distribution of usual, daily EE

• Estimate distribution of usual, daily EE in the original scale (?)

  • How do we define this?
Questions?