Published in final edited form as:

J R Stat Soc Ser C Appl Stat. 2011 August 1; 60(4): 619–632. doi:10.1111/j.1467-9876.2011.00762.x.

Properties of preliminary test estimators and shrinkage estimators for evaluating multiple exposures – Application to questionnaire data from the SONIC study

Jaya M. Satagopan¹, Qin Zhou¹, Susan A. Oliveria², Stephen W. Dusza², Martin A. Weinstock³, Marianne Berwick⁴, and Allan C. Halpern²

¹ Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center

² Dermatology Service, Department of Medicine, Memorial Sloan-Kettering Cancer Center

³ Department of Dermatology, Brown University

⁴ Division of Epidemiology, Department of Internal Medicine, University of New Mexico

Summary

Epidemiology studies increasingly examine multiple exposures in relation to disease by selecting the exposures of interest in a thematic manner. For example, sun exposure, sunburn, and sun protection behavior could be themes for an investigation of sun-related exposures. Several studies now use pre-defined linear combinations of the exposures pertaining to the themes to estimate the effects of the individual exposures. Such analyses may improve the precision of the exposure effects, but they can lead to inflated bias and type I errors when the linear combinations are inaccurate. We investigate preliminary test estimators and empirical Bayes type shrinkage estimators as alternative approaches when it is desirable to exploit the thematic choice of exposures, but the accuracy of the pre-defined linear combinations is unknown. We show that the two types of estimator are intimately related under certain assumptions. The shrinkage estimator derived under the assumption of an exchangeable prior distribution gives precise estimates and is robust to misspecifications of the user-defined linear combinations. The precision gains and robustness of the shrinkage estimation approach are illustrated using data from the SONIC study, where the exposures are the individual questionnaire items and the outcome is (log) total back nevus count.

Keywords

Empirical Bayes; Minimum risk; Random effects; Exchangeability

1. Introduction

Epidemiology studies increasingly examine multiple exposures in relation to a disease outcome by selecting the exposures in a thematic manner. The ongoing Study of Nevi in Children (SONIC) examines sun related exposure factors contributing to the etiology and evolution of nevus phenotypes. The sun factor data are obtained using a questionnaire that is designed to elicit information about three themes – sunburn, sun exposure, and sun protection behavior – using multiple questions. Oliveria et al (2009) reported the data

Correspondence to: Jaya M. Satagopan, Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, 307, East 63rd Street, 3rd Floor, New York, NY 10065, USA, Phone: (+1) 646-735-8122, FAX: (+1) 646-735-0010, satagopj@mskcc.org.

obtained at the study initiation period (year 2004) when 443 fifth grade students (approximately 10 years old), whose families provided informed consent, were recruited from all the schools in Framingham, Massachusetts, USA. During 2004, each child completed the SONIC questionnaire providing information on sun related factors pertaining to their most recent summer vacation. In this paper, we refer to the responses to the individual questions as “exposures”. The 46 exposures in the SONIC study are given in Table 1. The objective of the analysis reported by Oliveria et al (2009) was to estimate the exposure effects and identify those having significant association with a nevus-related outcome that is described below. One approach to address this objective is to fit a linear or generalized linear regression model for the outcome using all the exposures, ignoring the thematic design of the questionnaire. However, this approach has been criticized for the failure to take account of the themes (Greenland 1993), and on the grounds that incorporating the themes into the analyses may improve the precision of the estimates (i.e., narrower confidence intervals) and perhaps help reduce the false discoveries (Thomas 2005).

One approach for incorporating these themes is to quantify them as linear combinations of the individual exposures, which may be obtained via principal component analysis or factor analysis methods (Bartholomew and Knott 1999). In this paper we focus on pre-defined linear combinations to quantify the themes. Several recent studies of genetic polymorphisms consider the total number of minor alleles in relation to a disease outcome. The increasing availability of gene annotations also continues to accelerate the use of the pre-defined linear combinations in genetic studies. Melanoma studies have also considered cumulative sun exposure by adding the response to multiple questions about time spent in the sun, and similarly for sunburn. Thus, although methods such as factor analysis may be used to elicit the underlying themes using the observed data, pre-defined linear combinations are now seeing a widespread use in modern epidemiology studies. While this approach may be useful for improving the precision, it may lead to excessive false positive findings if the pre-defined linear combinations are inaccurate and the inaccuracies are not accounted for in the analysis.

The contributions of this paper are: (i) the examination of the benefits and limitations of using pre-defined linear combinations to estimate the exposure effects when one does not account for their potential misspecifications; and (ii) the investigation of the properties of a class of empirical Bayes type shrinkage estimators to estimate the exposure effects using pre-defined linear combinations by accounting for their potential misspecifications. We conduct an empirical illustration of these methods by analyzing the data reported by Oliveria et al (2009).

In the SONIC study, the total number of nevi of each child was determined using imaging techniques. The natural logarithm of the total number of nevi was treated as a continuous outcome, and was assumed to follow a normal distribution. The mean and standard deviation of the outcome were 1.81 and 0.94, respectively (on the log scale). The skewness and kurtosis of the sample distribution were -0.03 and 2.61 , respectively. The following host characteristics were recorded for each child: gender, ethnicity/race (from the school records), and a sun sensitivity measure. There were no missing data for the outcome or the host characteristics. Details about the missing exposure data are reported by Oliveria et al (2009), and are briefly summarized here. Both the children and their parent or legal guardian completed the questionnaire. The responses provided by the children were used in all the analyses. The number of missing data was relatively small, ranging from 26 children not responding to the shirt questions to 56 children not responding to the sunscreen questions. These missing data were imputed using the responses provided by the parents that were

available for all the students. Thus, we treated the exposure data as complete in all our analyses.

Oliveria et al (2009) analyzed the data obtained during 2004 using linear regression to identify the exposures that were significantly associated with log total nevus count, after adjusting for the host characteristics. This analysis, which we refer to as the full model analysis and which ignores the thematic design of the questionnaire, is based on the following linear model:

$$\mathbf{Y} = \mathbf{S}\mathbf{a}_{\text{full}} + \mathbf{X}\mathbf{b} + \mathbf{e}_{\text{full}} \quad (\text{Equation 1: Full Model})$$

Here $\mathbf{Y}$ is the vector of outcomes from N independent individuals, $\mathbf{S}$ is the matrix of host characteristics (including a column vector of ones for the intercept) with effects $\mathbf{a}_{\text{full}}$, $\mathbf{X}$ is the $N \times p$ matrix of exposures with effects $\mathbf{b}$, and $\mathbf{e}_{\text{full}}$ is a vector of independent error terms with $N(0, \sigma^2)$ distribution. [The matrix $\mathbf{S}$ may contain other variables besides the host characteristics, as deemed necessary.] In the SONIC study, $N = 443$ and $p = 46$. Our goal is to obtain precise estimates of $\mathbf{b}$. The maximum likelihood estimate (MLE) of $\mathbf{b}$, denoted $\hat{\mathbf{b}}_{\text{full}}$, and its variance can be obtained via least squares regression. Figure 1A shows $\hat{\mathbf{b}}_{\text{full}}$ and their 95% confidence intervals (CIs) for the SONIC exposure. As reported by Oliveria et al (2009), wearing a shirt (rarely, sometimes, often/always) when outside at the beach or pool was significantly associated with increased (log) total nevus count, and similarly for wearing a hat often or always when outdoors, which were contradictory to some other studies. The host characteristics were significantly associated with the outcome ($\mathbf{a}_{\text{full}}$ estimates and CIs not shown).

In order to incorporate the themes into the analysis, we defined linear combinations using the pre-defined weights given in Table 1, which can be interpreted as follows. Sun exposure is the cumulative number of hours spent in the sun. Weights of 1 and 2 are assigned if a person spends 3 to 4 hours and more than 4 hours, respectively, by the pool/beach during weekdays, and similarly for weekend days and for time spent outdoors. The weight is 0 if a person reports spending less than one hour by the beach/pool. These weights are chosen to reflect a trend type effect, which is used in several epidemiology studies (for example, genotype data are recorded as 0, 1, or 2 to represent the number of copies of a minor allele). Similarly, sunburn is the cumulative number of redness and painful burns. Sun protection behavior is defined in a similar manner. Let $\mathbf{W}$ denote a $p \times K$ matrix of weights, where K is the number of themes ($K = 3$ in the SONIC study). Thus, $\mathbf{XW}$ is an $N \times K$ matrix, where each column quantifies a pre-defined theme. The linear model based on the themes, referred to as the reduced model, is given by:

$$\mathbf{Y} = \mathbf{S}\mathbf{a}_{\text{red}} + \mathbf{XW}\mathbf{d} + \mathbf{e}_{\text{red}} \quad (\text{Equation 2: Reduced Model})$$

Here $\mathbf{a}_{\text{red}}$ are the effects of the variables contained in $\mathbf{S}$, and $\mathbf{d}$ denotes the effects of the themes. Its MLE $\hat{\mathbf{d}}$ and variance can be obtained via least squares regression. Under the reduced model, the exposure effect is $\mathbf{Wd}$. Its MLE is $\hat{\mathbf{b}}_{\text{red}} = \mathbf{W}\hat{\mathbf{d}}$, and the variance is $\mathbf{W}\text{Var}(\hat{\mathbf{d}})\mathbf{W}^T$. Figure 1B shows $\hat{\mathbf{b}}_{\text{red}}$ and their 95% CIs for the SONIC data. The exposures representing sunburn and sun protection behaviors were significantly associated with increased nevus counts, although the magnitudes of the effects were small relative to Figure 1A. Further, lengths of the CIs of $\hat{\mathbf{b}}_{\text{red}}$ were narrower than those of $\hat{\mathbf{b}}_{\text{full}}$, demonstrating the precision the gains of this approach. The host characteristics were significantly associated with the outcome, as before ($\mathbf{a}_{\text{red}}$ estimates and CIs not shown). However, the conclusions regarding the exposures are substantially different from those of the full model analysis.

In the absence of any external information or *a priori* knowledge about the themes, one approach for estimating the exposure effects would be to use the full model. The MLE of the full model provides unbiased estimates of the exposure effects if the model is correctly specified and in the absence of any unmeasured confounders. The reduced model is appealing since it can potentially improve the precision relative to the full model. However, the effects may be biased and the type I error may be inflated if $\mathbf{W}$ is inaccurate, even in the absence of any unmeasured confounders. Therefore, $\hat{\mathbf{b}}_{\text{red}}$ is not robust to misspecifications of the linear combinations. The estimates from the full and reduced model would be the same i.e., $\mathbf{b} = \mathbf{W}\mathbf{d}$ only when $\mathbf{W}$ is correctly specified. In this paper we will examine alternative methods to obtain estimates having better precision than those obtained from the full model and less false positive findings than the reduced model.

The remainder of this paper is organized as follows. In Section 2 we describe a preliminary hypothesis testing procedure that tests the accuracy of the pre-defined linear combinations (Sen 1986), and show that this approach may not be effective for the SONIC data. In Section 3, we propose a class of empirical Bayes type shrinkage estimators that are a weighted average of the estimates obtained from the full and reduced models. We illustrate their properties using simulations in Section 4, analyze the SONIC data in Section 5, and conclude the paper with a discussion in Section 6. Our investigations shed light on several important properties of the shrinkage estimators. The difference between the exposure effect estimates obtained from the full and reduced models can be treated as a nuisance parameter. We show that some members of the proposed class of shrinkage estimators are intimately related to the preliminary hypothesis test, while the other members are related to a random effects specification for the nuisance parameter. We also show that specifying an exchangeable prior distribution for the nuisance parameter leads to a shrinkage estimator that is robust to misspecifications of the predefined linear combinations.

2. Preliminary testing approach

One approach to assess the accuracy of $\mathbf{W}$ is to conduct a preliminary test of the general global null hypothesis $H_0: \mathbf{b} = \mathbf{W}\mathbf{d}$ (Sen 1986). Let $\mathbf{r} = \mathbf{b} - \mathbf{W}\mathbf{d}$. Its MLE is $\hat{\mathbf{r}} = \hat{\mathbf{b}}_{\text{full}} - \hat{\mathbf{b}}_{\text{red}}$. The calculation of the variance $\Sigma = \text{Var}(\hat{\mathbf{r}})$ is described in the Appendix. When the sample size N is large, Σ is $O(1/N)$. Therefore, the test statistic is $\Lambda = \hat{\mathbf{r}}^T \Sigma^{-1} \hat{\mathbf{r}}$, which follows a central chi-squared distribution with p degrees of freedom when H_0 is true. If H_0 is rejected, this would be an indication that $\mathbf{W}$ is incorrect, and we can obtain the preliminary test estimator (PTE) of $\mathbf{b}$ as $\hat{\mathbf{b}}_{\text{PT}} = \hat{\mathbf{b}}_{\text{full}}$. If H_0 is not rejected, $\mathbf{W}$ may be considered accurate and the PTE is $\hat{\mathbf{b}}_{\text{PT}} = \hat{\mathbf{b}}_{\text{red}}$.

We may also test the component-wise hypotheses $H_{0j}: b_j = (\mathbf{W}\mathbf{d})_j$ for $j = 1, \dots, p$. The test

statistic is $T_j = \frac{\hat{r}_j^2}{\sum_{jj}}$, which follows a central chi-squared distribution with 1 degree of freedom when H_{0j} is true. The subscripts j and jj denote the j th component of a vector and the j th diagonal entry of a matrix, respectively. The result of this hypothesis test can be used to obtain the corresponding component-wise PTEs, denoted $\hat{\mathbf{b}}_{j,\text{PT}}$ ($j = 1, \dots, p$). Since p hypotheses are tested, a Bonferroni correction or alternative methods such as the false discovery rate may be used.

We used the $\mathbf{W}$ matrix given in Table 1 to analyze the SONIC data. The global test statistic was not significant at the 5% level ($\Lambda = 40.21$ with 46 d.f; p -value = 0.71). The component-wise test statistics were not significant based on a Bonferroni correction for conducting 46 tests (the T_j s, each having 1 d.f, ranged from 0.001 to 5.89; p -values ranged from 0.01 to 0.97, all of which are greater than $0.0011 \approx 0.05/46$). These suggest that $\mathbf{W}$ may be deemed accurate, and we may estimate the exposure effects using $\hat{\mathbf{b}}_{\text{red}}$. In order to examine whether

this conclusion is reliable, we generated a 46×3 matrix $\mathbf{W}$ where each entry was drawn from a $N(0,1)$ distribution, and obtained the corresponding $\hat{\mathbf{b}}_{\text{red}}$. Clearly, this random $\mathbf{W}$ is inaccurate. However, neither the global test nor the component-wise test was rejected. Similar results were obtained when the entries of $\mathbf{W}$ were drawn randomly from $N(0,0.5)$, $N(10,1)$, and $\text{Uniform}(0,1)$ distributions. This suggests that the preliminary test approach may not be effective for the SONIC study, motivating the need to investigate alternative methods for estimating $\mathbf{b}$ by incorporating the themes.

3. Empirical Bayes estimators

We consider empirical Bayes type shrinkage estimators, denoted $\hat{\mathbf{b}}_{\text{eb}}$, that are pooled estimators based on the full and reduced models (Huntsberger 1955), defined as: $\hat{\mathbf{b}}_{\text{eb}} = \Theta \times \hat{\mathbf{b}}_{\text{full}} + (\mathbf{I}_p - \Theta) \times \hat{\mathbf{b}}_{\text{red}}$. Here $\mathbf{I}_p$ is the identity matrix of dimension p , and Θ is an unknown square matrix. We obtain Θ as the matrix that minimizes the mean squared error (MSE) of $\hat{\mathbf{b}}_{\text{eb}}$.

3.1 $\mathbf{r}$ is a fixed effect

Suppose $\mathbf{r} = \mathbf{b} - \mathbf{W}\mathbf{d}$ is an unknown fixed effect. Let Θ be a square matrix of dimension p . In the Appendix we show that the MSE is minimized when $\Theta = \mathbf{r}\mathbf{r}^T \times (\Sigma + \mathbf{r}\mathbf{r}^T)^{-1}$. The

resulting estimator, denoted $\hat{\mathbf{b}}_{\text{eb,gen}}$, is given by $\hat{\mathbf{b}}_{\text{eb,gen}} = \frac{\Lambda}{1+\Lambda} \hat{\mathbf{b}}_{\text{full}} + \frac{1}{1+\Lambda} \hat{\mathbf{b}}_{\text{red}}$.

Suppose we assume that Θ is a diagonal matrix with different diagonal entries. The MSE of $\hat{\mathbf{b}}_{\text{eb}}$ is minimized when the j th diagonal entry of Θ is $\theta_j = T_j/(1 + T_j)$. This provides the component-wise empirical Bayes estimator, denoted $\hat{\mathbf{b}}_{\text{eb,comp}}$, having j th component: $\hat{b}_{j,\text{eb,comp}} = \theta_j \times \hat{b}_{j,\text{full}} + (1 - \theta_j) \times \hat{b}_{j,\text{red}}$.

These estimators also have striking parallels to empirical Bayes estimators considered in recent genetic epidemiology applications (Luo et al 2009; Chen et al 2009). Our investigations establish a connection between the weights of these empirical Bayes estimators and the preliminary hypothesis tests.

3.2 $\mathbf{r}$ is a random effect

The reduced model provides erroneous estimates of $\mathbf{b}$ when $\mathbf{W}$ is inaccurate. In this case, the effects of the exposures are: $\mathbf{b} = \mathbf{W}\mathbf{d} + \mathbf{r}$, where $\mathbf{r}$ is an unknown nuisance parameter, and can be interpreted as the residual effect. When there is uncertainty about the accuracy of $\mathbf{W}$, $\mathbf{r}$ is not equal to $\mathbf{0}$. It is a random quantity that varies around $\mathbf{0}$. We specify an $N(\mathbf{0}, \mathbf{A})$ distribution for this random effect. The marginal distribution of $\hat{\mathbf{r}} = \hat{\mathbf{b}}_{\text{full}} - \hat{\mathbf{b}}_{\text{red}}$ is $N(\mathbf{0}, \Sigma + \mathbf{A})$. The minimum mean squared estimator of $\mathbf{r}$ is the posterior mean $E(\mathbf{r}|\hat{\mathbf{r}}) = \mathbf{A}(\Sigma + \mathbf{A})^{-1} \hat{\mathbf{r}}$. Note that $\mathbf{I}_p - \mathbf{A}(\Sigma + \mathbf{A})^{-1} = \Sigma(\Sigma + \mathbf{A})^{-1}$. Therefore, if $\mathbf{A}$ can be estimated empirically from the observed data, we have:

$$\begin{aligned} \hat{\mathbf{b}}_{\text{eb}} &= \hat{\mathbf{b}}_{\text{red}} + E(\mathbf{r}|\hat{\mathbf{r}}) \\ &= \hat{\mathbf{b}}_{\text{full}} - \hat{\mathbf{r}} + \mathbf{A}(\Sigma + \mathbf{A})^{-1} \hat{\mathbf{r}} \\ &= \hat{\mathbf{b}}_{\text{full}} - \Sigma(\Sigma + \mathbf{A})^{-1} \hat{\mathbf{r}} \end{aligned} \quad (\text{Equation 3})$$

We obtain three estimators based on the following assumptions: (i) $\mathbf{A}$ is a general symmetric matrix. A conservative estimate of $\mathbf{A}$ is $\hat{\mathbf{r}}\hat{\mathbf{r}}^T$. Using this in Equation 3, it can be seen that the resulting estimator is indeed $\hat{\mathbf{b}}_{\text{eb,gen}}$. (ii) $\mathbf{A}$ is a diagonal matrix with unequal diagonal entries. Its conservative estimator is $\text{diag}\{\hat{\mathbf{r}}\hat{\mathbf{r}}^T\}$, and the corresponding estimator is denoted $\hat{\mathbf{b}}_{\text{eb,non-ex}}$. (iii) $\mathbf{A}$ is a diagonal matrix with equal diagonal entries i.e., $\mathbf{A} = t^2\mathbf{I}_p$. A

conservative estimate of the prior variance t^2 is $\frac{\hat{\mathbf{r}}^T \hat{\mathbf{r}}}{p}$. The corresponding estimator is denoted $\hat{\mathbf{b}}_{\text{eb,ex}}$.

The first two assumptions correspond to specifying a non-exchangeable prior distribution for $\mathbf{r}$, while the third assumption is an exchangeable prior. These assumptions are also reflected in the subscripts “non-ex” and “ex” for the estimators. The choice of an exchangeable versus a non-exchangeable prior is not a testable hypothesis but simply a user-defined presumption since the likelihood does not contain information about the nature of the variance $\mathbf{A}$ (Lindley and Smith 1972). In Section 4 we conduct simulations to compare the operating characteristics of the various estimators.

3.3 Properties of the empirical Bayes estimators

The empirical Bayes estimators effectively shrink the difference between the estimates of the full and reduced model towards 0. Clearly, they provide biased estimates of $\mathbf{b}$. When the sample size increases, the variance Σ converges in probability to 0, and the bias converges to 0. The distributions of the empirical Bayes estimators are required to obtain their variances and confidence intervals. When $\mathbf{r} = \mathbf{0}$, the empirical Bayes estimators have an asymptotic mixture distribution (Sen 1986, Beran 1997), and obtaining their variances is fairly complicated. When $\mathbf{r} \neq \mathbf{0}$, they have an asymptotic normal distribution. In principle, a bootstrap procedure may be used to obtain the variances. However, this approach does not consistently estimate the distribution of shrinkage estimators (Beran 1997). Following Chen et al (2009), in Appendix C we derive approximations to the variances of the empirical Bayes estimators when $\mathbf{r} \neq \mathbf{0}$.

4. Simulations

We examined the operating characteristics of the various estimators using simulations as follows. We sampled $N = 400$ individuals with replacement from the SONIC data and retained their host characteristics and exposure data. Exposures 31 to 38 and 39 to 46 in Table 1 were designated as sunburn and sun exposure questions, respectively. We defined two binary latent causal variables L_1 and L_2 as follows. If a person spent 3 or more hours outdoors or by the pool or beach, then $L_1=1$ with probability 0.80. If a person spent 1 or 2 hours by the pool or beach then $L_1=1$ with a small probability such that the marginal probability of L_1 was 0.05 (= frequency of SONIC subjects spending 3 or more hours outdoors/pool/beach). Similarly, $L_2=1$ with probability 0.80 if a person had 2 or more episodes of redness or 2 or more painful sunburns. If a person had only 1 episode of redness or pain, $L_2=1$ with a small probability such that the marginal probability of L_2 was 0.06 (= frequency of SONIC subjects with 2 or more episodes of redness or pain). We generated the outcome as $Y = u_0 + u_1 L_1 + u_2 L_2 + u_{12} L_1 L_2 + N(0,1)$, with $u_0 = 0$, $u_1 = u_2 \in \{1, 1.5\}$, and $u_{12} \in \{-2, -0.25, 0, 2\}$. Thus, neither the full model nor the reduced model was the correct model generating the outcome. When analyzing the data, we assumed that only the questionnaire responses were observed, but not L_1 and L_2 . Questions 31 to 46 were considered to be the true non-null exposures when evaluating the operating characteristics (described below). Questions 1 to 30 were considered null. We analyzed the data using correct and incorrect $\mathbf{W}$. The incorrect $\mathbf{W}$ was a column vector of 1s. The correct $\mathbf{W}$ consisted of 3 columns: a column vector of 1s, and $\mathbf{W}_2$ and $\mathbf{W}_3$ given by Table 1. We refer to this as the correct $\mathbf{W}$ since questions 31 to 38 and 39 to 46 were correctly designated as sunburn and sun exposure questions, respectively, with non-zero weights.

We generated 1000 data sets under each parametric configuration, and estimated the exposure effects and their 95% CIs using various methods. An exposure was declared significant if its 95% CI excluded zero. The following operating characteristics were

examined. Type I error was the fraction of null exposures that were significant, averaged across all simulations. Since the effect of a non-null exposure may impact the significance of a null exposure, we also examined type I error under non-zero values of u_1 , u_2 and u_{12} . Detection probability (DP) was the fraction of simulations where at least one non-null exposure was significant. False discovery rate i.e., FDR (or true positive fraction i.e., TPF) was the fraction of null (or non-null) exposures among those significant, averaged across all the simulations. When none of the exposures were significant in a simulation, we set TPF = 0 for that simulation. Hence, $FDR + TPF \neq 1$ in our investigations.

Figure 2 shows the results for $u_1 = u_2 = 1$. When $\mathbf{W}$ was correctly specified, $\hat{\mathbf{b}}_{\text{red}}$ had excellent overall performance, as expected. Nonetheless, $\hat{\mathbf{b}}_{\text{eb,ex}}$ was a good competitor and had the best detection probability. When $\mathbf{W}$ was incorrect, $\hat{\mathbf{b}}_{\text{red}}$ had poor overall performance, including a substantially inflated type I error. In contrast, $\hat{\mathbf{b}}_{\text{eb,ex}}$ had the best overall performance. Further, the operating characteristics of $\hat{\mathbf{b}}_{\text{eb,ex}}$ were fairly stable regardless of the choice of $\mathbf{W}$, suggesting that $\hat{\mathbf{b}}_{\text{eb,ex}}$ is robust to misspecification of $\mathbf{W}$ and has a remarkable overall performance. The PTEs performed similar to $\hat{\mathbf{b}}_{\text{red}}$.

Similar results were obtained for $u_1 = u_2 = 1.5$. We also examined the distribution of the empirical Bayes estimators and compared their empirical variance to the estimated variance. Our investigations showed that the distributions of $\hat{\mathbf{b}}_{\text{eb,gen}}$ and $\hat{\mathbf{b}}_{\text{eb,ex}}$ did not have severe departures from normality, and that the variance estimates given in the Appendix are reasonable for these shrinkage estimators alone (results not shown, but available upon request from the corresponding author).

5. Application to the SONIC study

We estimated the exposures in the SONIC study using the robust estimator $\hat{\mathbf{b}}_{\text{eb,ex}}$ using the $\mathbf{W}$ given in Table 1. Figure 1C shows the estimates and the 95% CIs. As expected, the values of $\hat{\mathbf{b}}_{\text{eb,ex}}$ were between $\hat{\mathbf{b}}_{\text{full}}$ and $\hat{\mathbf{b}}_{\text{red}}$, illustrating the shrinkage. The lengths of the confidence intervals of $\hat{\mathbf{b}}_{\text{eb,ex}}$ were smaller than those of $\hat{\mathbf{b}}_{\text{full}}$, demonstrating the efficiency gains of the empirical Bayes type shrinkage estimator based on the exchangeable prior relative to the full model estimator. None of the exposures were significant at the 5% level. This suggests that the cross-sectional information on sun-related factors ascertained at the young age of 10 in the SONIC study does not explain the variation in the nevus counts. In both the nevus and melanoma literature, it is anticipated that the intensity and duration of sun-related exposures over time impact a nevus-related outcome rather than exposure based solely on a single time point. The results of our shrinkage estimation approach are consistent with these anticipated characteristics. A longitudinal follow-up of the SONIC cohort is currently underway to understand the effects of the intensity and duration of the exposures on nevus counts.

We also calculated the relative precision of each exposure as the ratio of the standard error of the exposure effect based on the full model estimator to a shrinkage estimator (Figure 1D). When the relative precision is greater than 1, the shrinkage estimator is more precise. The estimator $\hat{\mathbf{b}}_{\text{eb,ex}}$ had the best overall performance since the relative precisions of all the 46 exposures were greater than 1.

We conducted an empirical evaluation of the robustness of $\hat{\mathbf{b}}_{\text{eb,ex}}$ by considering $\mathbf{W}$ matrices with entries randomly generated from a $N(0, 1)$ distribution. Figure 3 plots the estimates and variances of $\hat{\mathbf{b}}_{\text{red}}$, $\hat{\mathbf{b}}_{\text{eb,ex}}$, $\hat{\mathbf{b}}_{\text{eb,non-ex}}$, and $\hat{\mathbf{b}}_{\text{eb,comp}}$ using $\mathbf{W}$ from Table 1 versus the random $\mathbf{W}$. All the empirical Bayes estimators were robust to the misspecification of $\mathbf{W}$, and $\hat{\mathbf{b}}_{\text{eb,ex}}$ had the smallest variance among all the empirical Bayes estimators considered here. In contrast, $\hat{\mathbf{b}}_{\text{red}}$ was sensitive to the choice of $\mathbf{W}$, highlighting the importance of

accounting for the potential misspecifications of pre-specified linear combinations. Similar results were obtained for other random choices of $\mathbf{W}$.

6. Discussion

Although methods such as factor analysis can be used to quantify the exposure themes based on the observed data (for example, Bartholomew and Knott 1999), the use of pre-defined linear combinations is now becoming popular in modern epidemiology studies due to the increasing availability of external data (such as gene annotation databases or other studies). Our investigations highlight the importance of accounting for the potential misspecifications of the pre-defined linear combinations, which has been ignored in some recent studies.

Our results show that the benefits of shrinkage estimation are best realized when $\mathbf{r} = \mathbf{b} - \mathbf{Wd}$ is treated as a random effect with an exchangeable prior. This phenomenon can be explained as follows. There are two sources of variation corresponding to the estimated value of $\mathbf{r}$. The first is explained variation, which arises due to modeling the effects of the questionnaire items based on the user-defined composite summaries. This variation (Σ) can be obtained using the variance of the estimated effects under the full and reduced models. The second source of variation is the unexplained variation, which arises due to potential misspecification of the user-defined composite summaries. Treating $\mathbf{r}$ as a fixed effect essentially ignores the unexplained variation, which leads to an inefficient empirical Bayes estimator. The random effects assumption accounts for the unexplained variation by specifying a prior distribution for $\mathbf{r}$. A non-exchangeable prior involves estimating several components of the unknown prior variance. In contrast, an exchangeable prior requires estimating a single parameter for the prior variance (i.e., t^2). Consequently, $\hat{\mathbf{b}}_{\text{eb,ex}}$ is more efficient.

We also examined the power of the estimators, defined as the fraction of significant non-null exposures averaged across all the simulations. Surprisingly, $\hat{\mathbf{b}}_{\text{red}}$ was the most powerful estimator even when $\mathbf{W}$ was inaccurate. This may be attributed to its inflated type I error i.e., when an exposure corresponding to a theme was identified as being significant under the reduced model analysis, all the other exposures corresponding to the same theme were also identified as being significant. However, $\hat{\mathbf{b}}_{\text{red}}$ had the worst detection probability when $\mathbf{W}$ was inaccurate. Our investigations also show that $\hat{\mathbf{b}}_{\text{eb,gen}}$ is similar to $\hat{\mathbf{b}}_{\text{full}}$, which appears to be counterintuitive since we would anticipate $\hat{\mathbf{b}}_{\text{eb,gen}}$ to behave like $\hat{\mathbf{b}}_{\text{red}}$ when $\mathbf{r} = \mathbf{0}$. However, Λ has a central chi-squared distribution with p degrees of freedom when $\mathbf{r} = \mathbf{0}$. When p is large, this is equivalent to a $N(p, 2p)$ distribution. Hence, on an average, Λ will be such that $\Lambda/(1+\Lambda)$ is near unity when p is large ($p = 46$ in the SONIC study). Consequently, $\hat{\mathbf{b}}_{\text{eb,gen}}$ is similar to $\hat{\mathbf{b}}_{\text{full}}$. Our shrinkage estimators are derived by minimizing the mean squared error. This also has parallels to penalized regression estimators based on a quadratic penalty (Hastie et al 2001). Further work is needed to conduct a comprehensive evaluation of estimators based on alternative penalties and their performance in practical applications.

All the SONIC data analyses reported in this paper are adjusted for the host characteristics (gender, ethnicity/race, and SSI). In our analyses, we have assumed that there are no unmeasured confounders. This assumption is currently being scrutinized in the ongoing SONIC study in several ways, including the planned examination of any unknown underlying population stratification, and the examination of various behavioral factors including, but not limited to, the desire to get a tan, going to get a tan as desired, and indeed getting a tan. The issue of effect modification is outside the scope of the current analyses, and is not considered in this paper. Under the auspices of the ongoing SONIC study, we are currently investigating different methods of collapsing the observed data to address effect

modification issues (such as effect modification by gender), and these results will be reported elsewhere.

Acknowledgments

This work was supported in part by research grants R01CA137420 and R01AR049342 from the National Institutes of Health, USA, and UL1RR024996 of the Clinical and Translational Science Center at Weill Cornell Medical College, New York, USA.

References

- Bartholomew, DJ.; Knott, M. Latent variable models and factor analysis. New York: Wiley; 1999.
- Beran R. Diagnosing bootstrap success. *Annals of the Institute of Statistical Mathematics*. 1997; 49:1–24.
- Chen Y-H, Chatterjee N, Carroll RJ. Shrinkage estimators for robust and efficient inference in haplotype-based case-control studies. *Journal of the American Statistical Association*. 2009; 104:220–233. [PubMed: 19430598]
- Greenland S. Methods for epidemiologic analyses of multiple exposures: A review and comparative study of maximum-likelihood, preliminary testing, and empirical-Bayes regression. *Statistics in Medicine*. 1993; 12:717–736. [PubMed: 8516590]
- Hastie, TJ.; Tibshirani, RJ.; Friedman, J. *The Elements of Statistical Learning*. New York: Springer; 2001.
- Huntsberger DV. A generalization of preliminary testing procedure for pooling data. *Annals of Mathematical Statistics*. 1955; 26:734–743.
- Lindley DV, Smith AFM. Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B*. 1972; 34:1–41.
- Luo S, Mukherjee B, Chen J, Chatterjee N. Shrinkage estimation for robust and efficient screening of single-SNP association from case-control genome-wide association studies. *Genetic Epidemiology*. 2009; 33:740–750. [PubMed: 19434716]
- Oliveria AA, Satagopan JM, Geller AC, Dusz SW, Weinstock MA, Berwick M, Bishop M, Heneghan MK, Halpern AC. Study of nevi in children (SONIC): Baseline findings and predictors of nevus count. *American Journal of Epidemiology*. 2009; 169:41–53. [PubMed: 19001133]
- Sen PK. On the asymptotic distributional risks of shrinkage and preliminary test versions of maximum likelihood estimators. *Sankhya*. 1986; 48:354–371.
- Thomas DC. The need for a systematic approach to complex pathways in molecular epidemiology. *Cancer Epidemiology, Biomarkers and Prevention*. 2005; 14:557–559.

APPENDIX

Appendix A: Calculating $\Sigma = \text{Var}(\hat{\mathbf{b}}_{\text{full}} - \hat{\mathbf{b}}_{\text{red}})$

$\text{Var}(\hat{\mathbf{b}}_{\text{full}} - \hat{\mathbf{b}}_{\text{red}}) = \text{Var}(\hat{\mathbf{b}}_{\text{full}}) + \text{Var}(\hat{\mathbf{b}}_{\text{red}}) - 2\text{Cov}(\hat{\mathbf{b}}_{\text{full}}, \hat{\mathbf{b}}_{\text{red}})$. Since $\hat{\mathbf{b}}_{\text{full}}$ and $\hat{\mathbf{b}}_{\text{red}}$ are maximum likelihood estimates, we can write:

$$\text{Cov}(\hat{\mathbf{b}}_{\text{full}}, \hat{\mathbf{b}}_{\text{red}}) \approx \frac{1}{N^2} \text{Var}(\hat{\mathbf{b}}_{\text{full}}) \left(\sum_{i=1}^N \psi_{i,\text{full}}(\mathbf{b}) \psi_{i,\text{red}}^T(\mathbf{d}) \right) \mathbf{W} \text{Var}(\hat{\mathbf{b}}_{\text{red}})$$

Here $\psi_{i,\text{full}}(\mathbf{b})$ and $\psi_{i,\text{red}}(\mathbf{d})$ are the contributions of person i ($i = 1 \dots N$) to the score functions of the full and reduced models, respectively. Conditional upon $\mathbf{S}$, $\mathbf{X}$, and $\mathbf{W}$, we have $\sigma_{\text{full}}^2 = \text{Var}(\mathbf{Y}|\mathbf{S}, \mathbf{X}, \mathbf{W}) = \sigma_{\text{red}}^2$. Therefore, it follows that the expected $\text{Cov}(\hat{\mathbf{b}}_{\text{full}}, \hat{\mathbf{b}}_{\text{red}})$ is equal to $\text{Var}(\hat{\mathbf{b}}_{\text{red}})$. Hence, the expected value of Σ is $\text{Var}(\hat{\mathbf{b}}_{\text{full}}) - \text{Var}(\hat{\mathbf{b}}_{\text{red}})$.

Appendix B: Derivation of the empirical Bayes estimators of Section 3.1

We shall obtain Θ that minimizes the risk function $\text{Trace}\{E(\hat{\mathbf{b}}_{\text{eb}} - \mathbf{b})(\hat{\mathbf{b}}_{\text{eb}} - \mathbf{b})^T\}$ as follows.

$$\begin{aligned} & E\{(\hat{\mathbf{b}}_{\text{eb}} - \mathbf{b})(\hat{\mathbf{b}}_{\text{eb}} - \mathbf{b})^T\} \\ &= E\left\{\left[\Theta(\hat{\mathbf{b}}_{\text{full}} - \mathbf{b}) + (\mathbf{I} - \Theta)(\hat{\mathbf{b}}_{\text{full}} - \mathbf{W}\mathbf{d} - \mathbf{r})\right]\left[\Theta(\hat{\mathbf{b}}_{\text{full}} - \mathbf{b}) + (\mathbf{I} - \Theta)(\hat{\mathbf{b}}_{\text{full}} - \mathbf{W}\mathbf{d} - \mathbf{r})\right]^T\right\} \\ &= \Theta\Sigma\Theta^T + \text{Var}(\hat{\mathbf{b}}_{\text{red}}) + (\mathbf{I} - \Theta)\mathbf{r}\mathbf{r}^T(\mathbf{I} - \Theta)^T \end{aligned}$$

This risk function is minimized when $\Theta = \mathbf{r}\mathbf{r}^T(\Sigma + \mathbf{r}\mathbf{r}^T)^{-1}$. We can estimate Θ by plugging in the estimated values of $\mathbf{r}$ and Σ . Since $\mathbf{r} = \mathbf{b} - \mathbf{W}\mathbf{d}$, we can write $\hat{\mathbf{b}}_{\text{eb}} = \hat{\mathbf{b}}_{\text{full}} - (\mathbf{I} - \Theta)\hat{\mathbf{r}}$.

$$\begin{aligned} \text{Further, } (\Sigma + \mathbf{r}\mathbf{r}^T)^{-1} &= \Sigma^{-1} - \frac{\Sigma^{-1}\mathbf{r}\mathbf{r}^T\Sigma^{-1}}{1 + \mathbf{r}^T\Sigma^{-1}\mathbf{r}}. \text{ Therefore, } \hat{\mathbf{b}}_{\text{eb}} = \hat{\mathbf{b}}_{\text{full}} - \frac{1}{1 + \Lambda}\hat{\mathbf{r}}. \text{ Equivalently,} \\ \hat{\mathbf{b}}_{\text{eb}} &= \frac{\Lambda}{1 + \Lambda}\hat{\mathbf{b}}_{\text{full}} - \frac{1}{1 + \Lambda}\hat{\mathbf{b}}_{\text{red}}. \end{aligned}$$

Suppose Θ is a diagonal matrix with the j th diagonal entry denoted θ_j . The j th component of the empirical Bayes estimator of $\mathbf{b}$ is: $\hat{b}_{j,\text{eb}} = \theta_j \hat{b}_{j,\text{full}} + (1 - \theta_j) \hat{b}_{j,\text{red}}$. Let $\Sigma_{jj} = \text{Var}(\hat{b}_{j,\text{full}} - \hat{b}_{j,\text{red}})$ denote the j th diagonal entry of Σ . The contribution of $\hat{b}_{j,\text{eb}}$ to the risk function is

given by: $E(\hat{b}_{j,\text{eb}} - b_j)^2 = \theta_j^2 \Sigma_{jj} + \text{Var}(\hat{b}_{j,\text{red}}) + (1 - \theta_j)^2 r_j^2$. This risk function is minimized when $\theta_j = (r_j)^2 / \{\Sigma_{jj} + (r_j)^2\} = T_j / (1 + T_j)$ where $T_j = (r_j)^2 / \Sigma_{jj}$.

Appendix C: Asymptotic distribution and approximate variances of the empirical Bayes estimators when $\mathbf{r} \neq \mathbf{0}$

Define $\mathbf{b}_{\text{eb}} = \mathbf{b} - \Sigma(\Sigma + \mathbf{A})^{-1}\mathbf{W}\mathbf{d}$. The true value of the empirical Bayes estimator is given by $\mathbf{b}_{\text{eb,true}} = \mathbf{b} - \Sigma_{\text{true}}(\Sigma_{\text{true}} + \mathbf{A})^{-1}\mathbf{W}\mathbf{d}$, where Σ_{true} is the true value of $\Sigma = \text{Var}(\hat{\mathbf{b}}_{\text{full}} - \hat{\mathbf{b}}_{\text{red}})$. Note that Σ is $O_p(1/N)$. Therefore, $N^{1/2}(\mathbf{b}_{\text{eb}} - \mathbf{b}_{\text{eb,true}}) = o_p(1)$. Suppose $\mathbf{r} \neq \mathbf{0}$. A one-step Taylor's series expansion of $\mathbf{b}_{\text{eb}} = f(\mathbf{b}, \mathbf{W}\mathbf{d})$ in the neighbourhood of $(\mathbf{b}, \mathbf{W}\mathbf{d}) = (\hat{\mathbf{b}}_{\text{full}}, \hat{\mathbf{b}}_{\text{red}})$ is:

$$f(\mathbf{b}, \mathbf{W}\mathbf{d}) = f(\hat{\mathbf{b}}_{\text{full}}, \hat{\mathbf{b}}_{\text{red}}) + \begin{pmatrix} \frac{\partial f}{\partial \mathbf{b}} & \frac{\partial f}{\partial \mathbf{W}\mathbf{d}} \end{pmatrix} \begin{pmatrix} \mathbf{b} - \hat{\mathbf{b}}_{\text{full}} \\ \mathbf{W}\mathbf{d} - \hat{\mathbf{b}}_{\text{red}} \end{pmatrix}.$$

The matrix of derivatives on the right hand side, which we shall denote as $\mathbf{M}$, is evaluated at $(\mathbf{b}, \mathbf{W}\mathbf{d}) = (\hat{\mathbf{b}}_{\text{full}}, \hat{\mathbf{b}}_{\text{red}})$. Therefore, we have:

$$\sqrt{N}(\hat{\mathbf{b}}_{\text{eb}} - \mathbf{b}_{\text{eb}}) = \sqrt{N}\mathbf{M} \begin{pmatrix} \hat{\mathbf{b}}_{\text{full}} - \mathbf{b} \\ \hat{\mathbf{b}}_{\text{red}} - \mathbf{W}\mathbf{d} \end{pmatrix} + o_p(1).$$

Since the maximum likelihood estimators are consistent and have an asymptotic normal distribution, $\hat{\mathbf{b}}_{\text{eb}}$ is an approximately consistent estimator of $\mathbf{b}_{\text{eb,true}}$ and has an asymptotic normal distribution. Therefore, the covariance matrix of $\hat{\mathbf{b}}_{\text{eb}}$ is given by $\mathbf{M}\mathbf{R}\mathbf{M}^T$, where $\mathbf{R}$ is the covariance matrix of $(\hat{\mathbf{b}}_{\text{full}}, \hat{\mathbf{b}}_{\text{red}})$. The block diagonal element of $\mathbf{R}$ are the variances of $\hat{\mathbf{b}}_{\text{full}}$ and $\hat{\mathbf{b}}_{\text{red}}$. The off-block-diagonal matrices are $\text{Cov}(\hat{\mathbf{b}}_{\text{full}}, \hat{\mathbf{b}}_{\text{red}})$. Further, $\mathbf{M} = (\mathbf{I} - \mathbf{M}_2 \quad \mathbf{M}_2)$, where $\mathbf{M}_2$ is obtained by taking the derivative of $\hat{\mathbf{b}}_{\text{eb}}$ with respect to $\hat{\mathbf{b}}_{\text{full}}$ and $\hat{\mathbf{b}}_{\text{red}}$.

For $\hat{\mathbf{b}}_{\text{eb,gen}}$, $\mathbf{M}_2 = \frac{\Lambda}{1+\Lambda} \mathbf{I} + \frac{2\mathbf{r}\mathbf{r}^T \Sigma^{-1}}{(1+\Lambda)^2}$. For $\hat{\mathbf{b}}_{\text{eb, non-ex}}$, $\mathbf{M}_2 = (\mathbf{I} - 2\mathbf{C}) \times (\Sigma + \mathbf{A})^{-1} \Sigma$. Let V_j denote the j th element of the vector $\mathbf{V} = (\Sigma + \mathbf{A})^{-1} \mathbf{r}$. $\mathbf{C}$ is a diagonal matrix with diagonal entry given by $C_{jj} = V_j \times r_j$.

For $\hat{\mathbf{b}}_{\text{eb,ex}}$, $\mathbf{M}_2 = (\mathbf{I} - 2/p \times \mathbf{C}^T) \times (\Sigma + \mathbf{A})^{-1} \Sigma$. Here $\mathbf{C}$ is a square matrix of dimension p with

the (i,j) th entry given by $C_{ij} = V_i \times r_j$. Finally, for $\hat{\mathbf{b}}_{\text{eb,comp}}$, $\mathbf{M}_2 = \text{diag} \left\{ \frac{\Sigma_{jj}(\Sigma_{jj} - r_j^2)}{(\Sigma_{jj} + r_j^2)^2} \right\}$.

Figure 1A: Full Model Estimates

Figure 1B: Reduced Model Estimates

Figure 1C: EB.EX Estimates

Figure 1D: Relative Precision

Figure 1.

Results of the SONIC questionnaire analysis. Figures 1A, 1B, and 1C correspond to $\hat{\delta}_{\text{full}}$, $\hat{\delta}_{\text{red}}$ and $\hat{\delta}_{\text{eb}\cdot\text{ex}}$, respectively. Each black horizontal line denotes the 95% confidence interval of an exposure (organized from top to bottom as given in Table 1), and shows the parameter estimate as a diamond. For ease of visualization, the questions are separated by red horizontal lines. Figure 1D is the precision of the shrinkage estimators relative to the full model estimator.

Figure 2. Type I error, detection probability (DP), false discovery rate (FDR) and true positive fraction (TPF) of the various estimators based on the simulation study. The preliminary test estimators performed similarly to the reduced model estimator, and are not shown.

Figure 3.

An empirical comparison of the estimated effects (left column) and the variances (right column) of the 46 questionnaire items in the SONIC data when the estimates are obtained using $\mathbf{W}$ given in Table 1 (horizontal axis) versus when the components of $\mathbf{W}$ are chosen according to $N(0, 1)$ (vertical axis). The rows correspond to $\hat{\mathbf{b}}_{\text{red}}$ (row 1), $\hat{\mathbf{b}}_{\text{eb,ex}}$ (row 2), $\hat{\mathbf{b}}_{\text{eb,nonex}}$ (row 3), and $\hat{\mathbf{b}}_{\text{eb,comp}}$ (row 4), respectively. Each figure also shows the 45 degree line.

Table 1

Exposures in the SONIC study. Column 1 is a generic question number. Column 2 shows the categorized questions. The response to each question is binary. Column 3 is the question type, which used to label the exposures in Figure 1. Column 4 is the theme— sun protection behavior, sunburn, and sun exposure. Columns 5, 6, and 7 show the three columns of the **W** matrix

Number	Question	Type	Theme	W		
				W ₁	W ₂	W ₃
1	Often or always limit time in the sun when not at beach/pool	LIMIT	SUN PROT.	1	0	0
2	Sometimes limit time in sun when not at beach/pool	LIMIT	SUN PROT.	2	0	0
3	Rarely limit time in sun when not at beach/pool	LIMIT	SUN PROT.	3	0	0
4	Often or always limit time in the sun while at beach/pool	LIMIT	SUN PROT.	1	0	0
5	Sometimes limit time in the sun while at beach/pool	LIMIT	SUN PROT.	2	0	0
6	Rarely limit time in the sun while at beach/pool	LIMIT	SUN PROT.	3	0	0
7	Often or always use sunscreen on the back	SUN SCR	SUN PROT.	1	0	0
8	Sometimes or rarely use sunscreen on the back	SUN SCR	SUN PROT.	2	0	0
9	Often or always use sunscreen (other anatomic sites)	SUN SCR	SUN PROT.	1	0	0
10	Sometimes or rarely use sunscreen (other anatomic sites)	SUN SCR	SUN PROT.	2	0	0
11	Often or always use sunscreen on face while at not at beach/pool	SUN SCR	SUN PROT.	1	0	0
12	Sometimes or rarely use sunscreen on face while not at beach/pool	SUN SCR	SUN PROT.	2	0	0
13	Often or always use sunscreen on back while not at beach/pool	SUN SCR	SUN PROT.	1	0	0
14	Sometimes or rarely use sunscreen on back while not at beach/pool	SUN SCR	SUN PROT.	2	0	0
15	Often or always use sunscreen on face while at beach/pool	SUN SCR	SUN PROT.	1	0	0
16	Sometimes or rarely use sunscreen on face while at beach/pool	SUN SCR	SUN PROT.	2	0	0
17	Often or always use sunscreen on back while at beach/pool	SUN SCR	SUN PROT.	1	0	0
18	Sometimes or rarely use sunscreen on back while at beach/pool	SUN SCR	SUN PROT.	2	0	0
19	Often or always wear hat when not at beach/pool	HAT	SUN PROT.	1	0	0
20	Sometimes wear hat when not at beach/pool	HAT	SUN PROT.	2	0	0
21	Rarely wear hat when not at beach/pool	HAT	SUN PROT.	3	0	0
22	Often or always wear hat while at beach/pool	HAT	SUN PROT.	1	0	0
23	Sometimes wear hat while at beach/pool	HAT	SUN PROT.	2	0	0
24	Rarely wear hat while at beach/pool	HAT	SUN PROT.	3	0	0

Number	Question	Type	Theme	W		
				W ₁	W ₂	W ₃
25	Often or always wear shirt to cover the back while not at beach/pool	SHIRT	SUN PROT.	1	0	0
26	Sometimes wear shirt to cover the back while not at beach/pool	SHIRT	SUN PROT.	2	0	0
27	Rarely wear shirt to cover the back while not at beach/pool	SHIRT	SUN PROT.	3	0	0
28	Often or always wear shirt to cover the back while at beach/pool	SHIRT	SUN PROT.	1	0	0
29	Sometimes wear shirt to cover the back while at beach/pool	SHIRT	SUN PROT.	2	0	0
30	Rarely wear shirt to cover the back while at beach/pool	SHIRT	SUN PROT.	3	0	0
31	Have painful sunburn on the back 2 or more times	PAIN	SUNBURN	0	0	1
32	Have painful sunburn on the back once	PAIN	SUNBURN	0	0	2
33	Have painful sunburn (other anatomic sites) 2 or more times	PAIN	SUNBURN	0	0	1
34	Have painful sunburn (other anatomic sites) once	PAIN	SUNBURN	0	0	2
35	Back stayed red or pink for more than 12 hours after being in the sun: more than once	RED	SUNBURN	0	0	1
36	Back stayed red or pink for more than 12 hours after being in the sun: only once	RED	SUNBURN	0	0	2
37	Skin (other anatomic sites) stayed red or pink for more than 12 hours after being in the sun: > once	RED	SUNBURN	0	0	1
38	Skin stayed red or pink for more than 12 hours after being in the sun: only once	RED	SUNBURN	0	0	2
39	Spend more than 2 hours outdoors on a typical weekend day	OUT	SUN EXP	0	1	0
40	Spend between 1 to 2 hours outdoors on a typical weekend day	OUT	SUN EXP	0	2	0
41	Spend more than 2 hours outdoors on a typical week day	OUT	SUN EXP	0	1	0
42	Spend between 1 to 2 hours outdoors on a typical week day	OUT	SUN EXP	0	2	0
43	Spend more than 2 hours by the beach/pool on a typical weekend day	POOL	SUN EXP	0	1	0
44	Spend 1 to 2 hours by the beach/pool on a typical weekend day	POOL	SUN EXP	0	2	0
45	Spend more than 2 hours by the beach/pool on a typical week day	POOL	SUN EXP	0	1	0
46	Spend 1 to 2 hours by the beach/pool on a typical week day	POOL	SUN EXP	0	2	0