Reevaluating Composite Scores: Fast Quasi-Netwton Optimization and Bootstrapping

Eli Kravitz

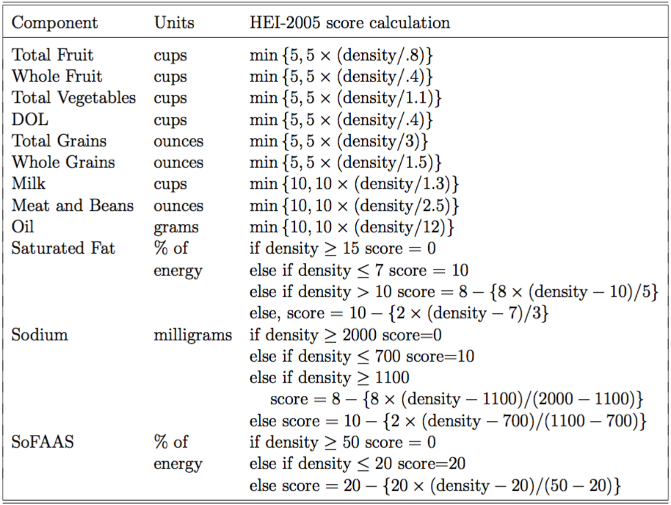
Texas A&M

February 13, 2022

Introduction

- It is common in epidemiology to use composite scores to assess health behavior.
  - Healthy Eating Index (next slide), Mediterranean Diet Score, Physical and Mental Health Composite Scores, etc.
- Assign individuals' health behavior a **single interpretable score** between 0 and 100. Use that score to model disease risk

Introduction



Component	Units	HEI-2005 score calculation
Total Fruit	cups	$\min\{5, 5 \times (\text{density}/.8)\}$
Whole Fruit	cups	$\min\{5, 5 \times (\text{density}/.4)\}$
Total Vegetables	cups	$\min\{5, 5 \times (\text{density}/1.1)\}$
DOL	cups	$\min\{5, 5 \times (\text{density}/.4)\}$
Total Grains	ounces	$\min\{5, 5 \times (\text{density}/3)\}$
Whole Grains	ounces	$\min\{5, 5 \times (\text{density}/1.5)\}$
Milk	cups	$\min\{10, 10 \times (\text{density}/1.3)\}$
Meat and Beans	ounces	$\min\{10, 10 \times (\text{density}/2.5)\}$
Oil	grams	$\min\{10, 10 \times (\text{density}/12)\}$

Saturated Fat	% of energy	if density ≥ 15 score $= 0$
		else if density ≤ 7 score $= 10$
		else if density > 10 score $= 8 - \{8 \times (\text{density} - 10)/5\}$
		else, score $= 10 - \{2 \times (\text{density} - 7)/3\}$

Sodium	milligrams	if density ≥ 2000 score=0
		else if density ≤ 700 score=10
		else if density ≥ 1100
		score $= 8 - \{8 \times (\text{density} - 1100)/(2000 - 1100)\}$
		else score $= 10 - \{2 \times (\text{density} - 700)/(1100 - 700)\}$

SoFAAS	% of energy	if density ≥ 50 score $= 0$
		else if density ≤ 20 score=20
		else score $= 20 - \{20 \times (\text{density} - 20)/(50 - 20)\}$

Figure: 2005 Healthy Index Index (HEI) developed by U.S. Department of Agriculture (USDA)

Introduction

- Improvement: Use **many populations and diseases** to build a more accurate score.
  - single score but more predictive
- The data of interest NIH-AARP Study of Diet and Health
- Relate **cancer** and **mortality** with **quality of diet** as measured by the HEI.

Introduction

Denote $j = 1, ..., J$ as the index of the HEI component. There are $k = 1, ...K$ populations and $\ell = 1, ...L_K$ diseases in each population. The data are observed as follows:

- $Y_{ik\ell} = 1$ is a binary indicator of disease ℓ for the i^{th} person in population k .
- Let $(X_{i1}, ..., X_{iJ})$ be the HEI score for person i with components $j = 1, ..., J$. $J = 12$ in the 2005-HEI
- Covariates are denoted as $Z_{ik\ell}$. This includes age, ethnicity, education, body mass index, smoking status, etc.

Introduction

- We model the probability of someone of population ℓ having disease k as

$$\Pr(Y_{k\ell} = 1 | X_{ijl}, Z_{ik\ell}) = H(\beta_{k\ell} \textstyle\sum_{j=1}^{J} X_{ijk} \alpha_j + Z_{ik\ell} \theta_{k\ell}), \qquad (1)$$

  where $H(\cdot)$ is the logistic function.

- This model needs a constraint for identifiability. Initially set $\beta_{11} = -1$. After model converges, set $\alpha_j^* = \alpha_j / \alpha^T c_{max}$ where c_{max} is the highest value assigned to a component in the HEI.

Introduction

- Written less compactly, we fit a system of nonlinear equations:

$$H(\beta_{11}\sum_{j=1}^{J} X_{i11}\alpha_j + Z_{i11}\theta_{11})$$

$$\mathrm{H}(\beta_{12}\sum_{j=1}^{J} X_{i12}\alpha_j + Z_{i12}\theta_{i12})$$

$$\vdots$$

$$\vdots$$

$$\mathrm{H}(\beta_{k\ell}\sum_{j=1}^{J} X_{ijk}\alpha_j + Z_{ik\ell}\theta_{k\ell}),$$

where each population, disease combination is modeled with **separate** β and θ but a **single** α

Introduction

$$\Pr(Y_{ik\ell} = 1 | X_{ij}, Z_{ik\ell}) = H(\beta_{k\ell} \sum_{j=1}^{J} X_{ij}\alpha_j + Z_{ik\ell}\theta_{k\ell}),$$

- Three unknown vectors
  - α: The new weights assigned to the 12 HEI components. When $\alpha \equiv 1$, the HEI is unchanged.
  - β: The effect of diet on disease ℓ in population k
  - θ: Covariate effect
- Single dietary score, $\sum_{j=1}^{J} X_i\alpha_j$, that **does not depend on population or disease**
- Similarly, β and θ have no dependence on diet.

Introduction

- There are two important areas that the methods from STAT605 can be used.
- Model Fitting
  - Reduce the **computation time** required to fit the model.
- Assessing relative risk (RR) of disease
  - RR has a complication asymptotic distribution that may not be valid.
  - **Bootstrap** can give SE and CI's.

Model Fitting

- This model falls outside of standard GLM software because of the dependence between α and β
- Parameters are estimated using iterative profile likelihood procedure.
  - Fix α and estimate β, θ with standard GLM methods
  - Fix β and θ and maximize likelihood with respect to α. (**very slow!**)

Model Fitting

- Maximize likelihood with the Broyden–Fletcher–Goldfarb-Shanno (BFGS) algorithm
  - BFGS is a quasi-Newton algorithm.
- Recall that a quasi-Newton algorithm alters the traditional Newton method,

$$x_{t+1} = x_t - \alpha_t H(x_t)^{-1} \nabla f(x_t) \tag{2}$$

  by replacing the inverse Hessian matrix, $H(\cdot)^{-1}$, by an approximation.

Model Fitting

- In BFGS, H_{t+1}^{-1} is given by

$$(I - \frac{\Delta x_t y_t^T}{y_t^T \Delta x_t}) H_t^{-1} (I - \frac{y_t \Delta x_t}{y_t^T \Delta x_t}) + \frac{\Delta x_t \Delta x_t}{y_t^T \Delta x_t}. \tag{3}$$

  where

  - y_t is the difference in
  - Δx_t is the step size
  - H_t^{-1} is the previous approximation of the Hessian.

Model Fitting

- BFGS is implemented in C++ using the *Rcpp* package along with the matrix algebra library *Armadillo*.

Model Fitting

	HEI Score	Cancer Weight	Mortality Weight
Total Grain	5	0.85	0.88
Total Fruit	5	1.76	-0.13
Whole Fruit	5	1.00	0.36
Whole Grain	5	3.74	5.55
Total Vegetables	5	1.77	2.16
DOL Vegetables	5	0.99	0.80
Dairy	10	0.78	0.57
Meat and Beans	10	0.69	1.09
Oils	10	0.46	0.6
Sodium	10	1.903	1.95
Saturated Fats	10	0.77	1.06
Empty Calories	20	0.17	-0.045

Table: Results from fitting our model Interpretation: A perfect score of 5 for total grains in the HEI would now received a score of $5 \times 3.74 = 18.7$

Relative Risk and Bootstrap

- Replaced old dietary score, $\sum_J X_{ij}$, with new dietary score $T = \sum_J X_{ij}\alpha_j$
- We want to assess how effective new dietary suggestions are.
  - We want to see if the disease risk of someone with a poor diet differs from the disease risk of someone with a good diet.
- Formally, we want to estimate the relative risk of moving from the 90^{th} quantile of dietary scores to the 10^{th} quantile:

$$exp\{\beta_{k\ell}(T_{k\ell,90^{th}} - T_{k\ell,10^{th}})\}, \tag{4}$$

where $T_{k\ell,q^{th}}$ is the q^{th} quantile of $\sum_J X_{ij}\alpha_j$

Relative Risk and Bootstrap

- I asses the variability in (4) using the non parametric bootstrap.
- In practice, models have a huge number of nusiance parameters, θ.
  - Reedy et. all (2008) suggest **25 covariates** per population-disease combination
  - **100 nuisance parameters** in our analysis.
- These can expand asymptotic confidence intervals well past nominal coverage.
- Diciccio and Efron (1996) suggest that bootstrap CI's are **"more dependable"** than delta-method CI's

Relative Risk and Bootstrap

- Getting an adequate number of samples is challenging. Fitting even a single model can be very time consuming.
- This limits the number of bootstraps samples that can be taken in a reasonable amount of time.
- To **decrease variability** in sampling, I implement the **during-sampling balanced** bootstrap suggested in Efron and Tibshirani (1993).
  - This allows the relatively small number of bootstrap sample to more accurately estimate the variability of the relative risk

Relative Risk and Bootstrap

- **Balanced Bootstrap:** Generate $b = 1, \ldots, B$ copies each population. Permute each of the subgroups separately to assure that samples from each population appear the same number of times.

  1. **Model Fitting:** For the b generated data set fit the regression $(Y_{ik\ell,b} = 1|X_b, Z_b) = H(\beta_{k\ell} \sum_j X_{ij,b}\alpha_j + Z_{ikl,b}\theta_{kl})$ with $\beta_{11} = -1$
  2. **Impose Constraint:** $\widehat{\alpha}_b$ is set to $\widehat{\alpha}^*_{j,b} = \alpha_{j,b}/\alpha_b c_{max}$ and calculate a value for β_{11}.
  3. **Recalculate Dietary Score:** Using the original dataset, X define $T_{i,b} = \sum_j X_{ij}\widehat{\alpha}^*_{j,b}$
  4. **Compute log relative risk** Take the 10th and 90th percentile of $T_{i\ell,b}$ and calculate log relative risk as $\mathcal{L}_{kl,b} = \beta^*_{kl,b}(T_{i,b_{90th}} - T_{il,b_{10th}})$.

- Repeat steps $(1 \sim 4)$ for all B data sets.

Relative Risk and Bootstrap

- Steps ($1 \sim 4$) are completely parallelized.
- The bootstrap samples are asymptotically distributed as a Gaussian centered at the true log RR. (Not shown here)

Relative Risk and Bootstrap

	Men		Women	
	Relative Risk	CI	Relative Risk	CI
Lung	0.46	(0.42 ,0.48)	0.52	(0.47, 0.55)
Colorectal	0.74	(0.67, 0.79)	0.82	(0.74, 0.92)
Prostate	1.12	(1.06, 1.16)	•	•
Breast	•	•	0.97	(0.91, 1.03)
Ovarian	•	•	1.03	(0.87, 1.26)

Table: Relative risk for men and women when cancer risk of interest.

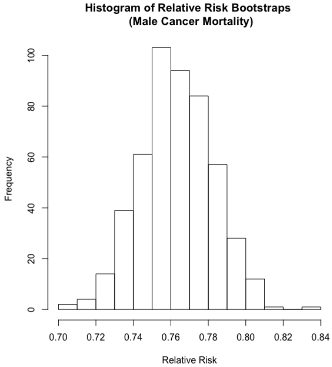
Relative Risk and Bootstrap

	Men		Women	
	Relative Risk	CI	Relative Risk	CI
Cancer	0.77	(0.73 ,0.80)	0.82	(0.77, 0.86)
CVD	0.69	(0.65, 0.73)	0.64	(0.59, 0.70)
Other	0.62	(0.59, 0.65)	0.62	(0.58, 0.67)

Table: Relative risk for men and women when mortality risk of interest.

Thank You

Relative Risk and Bootstrap



- Histograms of bootstraps are fairly symmetric.
- BC_a may have been helpful, but the distribution isn't extremely skewed.