Reevaluating Composite Scores with Flexible Regression and Variable Selection

Eli Kravitz

Texas A&M

February 13, 2022

Background

- Follow up to Ma, Ma, Wang, Kravitz & Carroll (2016).
- 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

Background

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)

Background

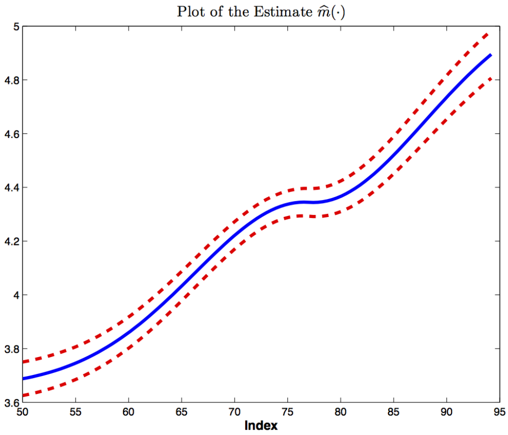
- Improvement: Use **many populations and diseases** to build a more accurate score.
  - single score but more predictive
- To relate 2005-HEI and cancer, Ma et. al. (2016) developed the single index model

$$\Pr(Y_{k\ell} = 1|X) = H\{\beta_{k\ell} m(\sum_j X_{jk}\alpha_j)\}(1)$$

where $H(\cdot)$ is the logistic distribution function, and $m(\cdot)$ is an unknown function.

Background

- On the data of interest (NIH-AARP Study of Diet and Health), $m(\cdot)$ is **nearly linear**.



Setup

- We remove $m(\cdot)$ and work with a flexible GLM.
- We're still able to calibrate HEI, but with **lower variability**, more **numerical stability**, and it is easier to perform **variable selection**.
- Variable selection allows us to see what **HEI components** have **negligable effect** on health status

Setup

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. There are $i = 1, ...n_{k\ell}$ individuals with disease ℓ in population k. 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 $(X_{i1l}, ..., X_{iJ})$ 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.

Setup

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

$$\Pr(Y_{k\ell} = 1 | X_{ijl}, Z_{ik\ell}) = H(\beta_{k\ell} \textstyle\sum_{j=1}^{J} X_{ijk} \alpha_j + Z_{ik\ell} \theta_{k\ell}), \qquad (2)$$

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

- This model needs a constraint for identifiability. Initially set $\beta_{11} = -1$

Important Features

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

- 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.

Model Fitting

- This model falls outside of standard GLM software because of the dependence between α and β
- Parameters are estimated using 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!**)
- 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.
  - This constraint on α forces the **new score** assigned to someone to be **between 0 and 100**.
  - α^* is constrained, so β_{11} is identifiable. Refit to get a value for β_{11}

Variable Selection

- We want to establish if any HEI component has **no effect** on health status
- We add an **adaptive lasso** (Zou, 2006) penalty to the α parameters in our negative log likelihood,

$$n^{-1} L_n(\beta, \alpha, \theta) + \lambda \textstyle\sum_{j=1}^{J} |\widehat{\alpha}_{full,j}|^{-\gamma} |\alpha_j|, \qquad (3)$$

where λ is the tuning parameter, γ is a prespecified positive number, and $\widehat{\alpha}_{full,j}$ is an estimate of α_j which has not been subject to any constraint

Variable Selection

- There are several issues with implementing (3), or any regularization for that matter.
- Typical tools for fitting Lasso problems (*glmnet* or Least Angle Regression Efron (2004)) are not equipt to handle the term $\sum_j X_{ij\ell}\alpha_j$. They cannot penalize the α coefficients without also penalizing the β coefficient.
- Minimizing $n^{-1}L_n$ is **very time consuming**.
  - Solving (3) directly and performing a grid-search for the optimal λ is not realistic

Variable Selection

- For a conceptually simple and computation fast solution, we use Wang and Leng's (2007) Least Squares Approximation (LSA).
- The authors show that under very mild regularity conditions, a loss function, $n^{-1}L_n(\beta) + \sum_p \lambda_p|\beta|$, can be expressed as an asymptotically equivalent least squares problem:

$$Q(\beta) = (\widetilde{\beta} - \beta)^T \hat{\Sigma}^{-1}(\widetilde{\beta} - \beta) + \sum_{j=1}^{d} \lambda_j |\beta_j|,$$

where $\widetilde{\beta}$ is the parameter than minimizes $L_n(\cdot)$ and $\widehat{\Sigma}$ is an asymptotically consistent estimate of the covariance matrix of $\widetilde{\beta}$.

Variable Selection

- We approximate our log likelihood as

$$L_n(\Theta) + \lambda \sum_{j=1}^{J} |\widehat{\alpha}_{full,j}|^{-\gamma}|\alpha_j| \approx$$

$$(\widetilde{\Theta} - \Theta)^T \widehat{\Sigma}^{-1} (\widetilde{\Theta} - \Theta) + \lambda \sum_{j=1}^{J} |\widehat{\alpha}_{full,j}|^{-\gamma}|\alpha_j|, \quad (4)$$

  where $\Theta = (\beta, \alpha, \theta)$.

- (4) can be fit quickly for any value of λ with *glmnet* as a Gaussian family problem.
  - Denote $\widehat{\Theta}_{LSA}(\lambda)$ as the value which minimizes the right hand side of (4) as a function of λ.

Choice of Tuning Parameter

- Like all Lasso methods, LSA provides a solution for any λ, however the optimal value of λ must be selected.
- Wang and Leng propose a BIC style criterion, namely

$$BIC(\lambda) = (\widehat{\Theta}_{LSA}(\lambda) - \widetilde{\Theta}_{full})\widehat{\Sigma}^{-1}(\widehat{\Theta}_{LSA}(\lambda) - \widetilde{\Theta}_{full}) + g_n/n \log(n), \quad (5)$$

  where g_n is the number of nonzero coefficients in $\widehat{\Theta}_{LSA}(\lambda)$.
- Can be shown that any λ that **does not select the true subset** of predictors **will not be chosen** by the BIC criterion.

Oracle Properties

- Variable selection procedures should have the *oracle* property.
- Fan and Li (2001): A selection procedure δ has the oracle property if
  - **Selection Consistency**: $\Pr\{\widehat{A}(\lambda) = A\} \to 1$
  - **Optimal Estimation Rate**: $\sqrt{n}(\widehat{\Theta}_{\delta,\widehat{A}_\delta} - \Theta_A) \to N(0, \Sigma_A)$ in distribution, where Θ_A are the nonzero components of Θ and Σ_A is the covariance matrix of the limiting distribution of true subset of predictors
- where $A = \{j : \Theta_j \neq 0\}$ and $\widehat{A}(\lambda) = \{j : \widehat{\Theta}(\lambda)_{LSA,j} \neq 0\}$ The procedure δ should have:

Oracle Properties

- We had unexpected difficulties right before this presentation.
- **Selection consistency** is provided in Wang and Leng (2007) with minor assumptions on λ
- Conditions from Wang and Leng for **optimal estimation rate are not satisfied**.
  - However we can still show optimal estimation rate (unexpected difficulties)

Results

- We apply our methods to the NIH-AARP Study of Diet and Health.
- This study tracks **lung, colorectal, prostate, breast, and ovarian cancer** in adults between the ages of 51-75. As well as **cause of death** for anyone who died during study.
- Mortality is analyzed in two ways: as a mutually exclusive outcome of one of several causes or as the aggregation of *any* type of mortality.

Results

Description	Men # Cases	Men Percentages	Women # Cases	Women Percentages
Sample size	294,673		199,285	
Breast cancer			7,736	3.88%
Ovarian cancer			759	0.38%
Prostate cancer	23,477	7.97%		
Colorectal cancer	4,693	1.59%	2,291	1.15%
Lung cancer	6,135	2.08%	3,630	1.82%

Table: Summary of the NIH-AARP data for cancer occurance.

Results

Description	Men # Cases	Women # Cases
Sample size	219,612	169,480
CVD mortality	8,112	4,028
Cancer mortality	12,247	7,344
Diabetes mortality	269	138
Other cause mortality	10,552	6,349

Table: Summary of the NIH-AARP data for mortality. Cardiovascular disease has been abbreviated as CVD.

Results: Cancer

	se	Unpenalized	Penalized
Whole Grain	0.23	0.85	0.91
Total Fruit	0.26	1.76	2.20
Whole Fruit	0.25	1.01	1.07
Total Grain	0.29	3.74	4.19
Total Veg.	0.28	1.76	1.91
DOL Veg.	0.21	1.00	1.14
Dairy	0.08	0.78	0.82
Meat and Beans	0.13	0.69	0.46
Oils	0.09	0.46	0.48
Sodium	0.11	1.90	1.71
Saturated Fat	0.09	0.77	0.82
Empty Calories	0.06	0.17	**0**

- Ex: A perfect score of 5 for total grains would now received a score of $5 \times 3.74 = 18.7$

Results: Cancer

	Men		Women	
	Estimate	se	Estimate	se
Lung	-0.35	0.02	-0.33	0.024
Colorectal	-0.14	0.019	-0.10	0.029
Prostate	0.05	0.009	·	·
Breast	·	·	-0.013	0.017
Ovarian	·	·	0.016	0.053

Table: Results for $\widehat{\beta}$ when cancer is the outcome of interest.

Results: Mortality

	se	Unpenalized	Penalized
Whole Grain	0.22	0.88	0.84
Total Grain	0.27	5.55	5.90
Whole Fruit	0.23	0.36	**0**
Total Fruit	0.25	-0.13	**0**
Total Veg.	0.27	2.17	2.37
DOL Veg.	0.21	0.80	0.76
Dairy	0.076	0.57	0.50
Meat and Beans	0.12	1.10	0.99
Oils	0.09	0.60	0.52
Sodium	0.10	1.95	1.98
Saturated Fat	0.087	1.05	1.07
Empty Calories	0.056	-0.046	**0**

Table: Mortality Analysis

Results: Mortality

	se	Unpenalized	Penalized
Whole Grain	0.23	0.92	0.88
Total Fruit	0.26	-0.09	**0.00**
Whole Fruit	0.25	0.48	0.36
Total Grain	0.29	5.59	5.58
Total Veg.	0.28	2.11	2.11
DOL Veg.	0.21	0.70	0.67
Dairy	0.08	0.62	0.59
Meat and Beans	0.13	1.04	1.03
Oils	0.09	0.64	0.62
Sodium	0.11	1.85	1.89
Saturated Fat	0.09	1.10	1.07
Empty Calories	0.06	-0.06	**0.00**

Table: All Cause Mortality Analysis

Results: Mortality

	Men		Women	
	Estimate	se	Estimate	se
Cancer	-0.16	0.017	-0.12	0.022
CVD	-0.22	0.021	-0.28	0.028
Other	-0.29	0.019	-0.29	0.023
All- Cause Mortality	-0.25	0.133	-0.23	0.021

Table: Results for $\widehat{\beta}$ when mortality is the outcome of interest. Cardiovascular disease is abbreviated as CVD.

Remarks

- The original HEI is more distorted for **mortality** than for **cancer**
  - The α coefficients are further from 1, and more coefficients are set to 0.
- Nutritionists have focused on cancer over mortality traditionally. This may be a side effect of this.

Thank You