Does Life Span Run in Families, and If So, Why?

Grace Wahba

Based on "Using Distance Correlation and SS-ANOVA to Assess Associations of Lifestyle Factors, Diseases, and Mortality" by Jing Kong, Barbara Klein, Ronald Klein, Kristine Lee and Grace Wahba, PNAS, December 11, 2012

Prediction of Complex Traits in Animals, Human and Plants

Department of Animal Sciences, Madison

May 30, 2013

Daniel Gianola, Organizer

Links to these slides (TALKS link), and the above paper (TRLIST link), in my website

`http://www.stat.wisc.edu/~wahba/`

Abstract

The Beaver Dam Eye study began in 1988 and enrolled a group of 4926 people in Beaver Dam, Wisconsin, then aged 43 to 86 years , and followed them for over 20 years. A large fraction of the study population have family members in the study. Many covariates for the study population have been measured at the start and later, and mortality information on this population has been updated to March 2011. We compared the pairwise death ages of relatives who had died by March 2011 and pairwise death ages of unrelated members of this population, and it is evident that there is a familial effect on mortality, agreeing with the common perception that mortality/longevity tends to run in families. In this study, we have several lifestyle variables that are risk factors for mortality. We use the tools of Distance Correlation and Smoothing Spline ANOVA to show that these risky lifestyle factors also tend to run in families, contributing to the Nature-Nurture debate.

Outline

1. Introduction. The Beaver Dam Eye Study (BDES)
2. The Data and the Questions
3. Methodology Outline
4. What is Distance Correlation (DCOR)?
5. Pedigrees and Pedigree Dissimilarity
6. Death Age Scoring
7. The Smoothing Spline ANOVA (SSANOVA) scoring model.
8. Determining Distance Correlation, Results
9. Conclusions

Introduction, The Beaver Dam Eye Study

Beaver Dam is a small, homogenous community about an hour by car from Madison (2011 population about 16,000). The Beaver Dam Eye Study, which began in 1988 enrolled 4926 people aged 43-86 years of age. 2356 people had relatives in the study. By March 2011, 1004 subjects had died, aged from 46 to 101 years, and pedigrees were constructed for them. A subgroup of 843 people came from pedigrees with 2 or more members, resulting in 222 pedigrees with sizes from 2 to 23 people- the population in this study. We compaired pairwise differences in death ages between siblings, and between unrelated pairs in the study population and estimated the average pairwise difference in death age of full sibling pairs was 8.09 while the average pairwise distances in death age of unrelated pairs was 9.67, suggesting agreement with a general perception that mortality/longevity tends to run in families.

The Data and the Questions

The Nature-Nurture debate continues, and BDES provides a unique opportunity to examine some aspects of this debate, since extensive pedigrees are available, as well as a wealth of covariates, including modifiable lifestyle factors related to mortality. Starting with the suggestion that death ages run in families, either younger or older, we decided to develop and apply an innovative combination of two methodological tools, Distance Correlation and Smoothing Spline ANOVA, to see to what extent lifestyle factors and diseases which are risk factors for mortality also run in families.

Table 1. Variable Descriptions: Fixed:Lifestyle:Diseases

variable	units	description
deathage	years	death age
baseage	years	age at baseline
gender	F/M	gender
.....		
edu	years	highest year school/college completed
bmi	kg/m^2	body mass index
smoke	yes/no	history of smoking
inc	yes/no	household personal income $>$ 20T
.....		
diabetes	yes/no	history of diabetes
cancer	yes/no	history of cancer
heart	yes/no	history of cardiovascular disease
kidney	yes/no	history of chronic kidney disease

Variable Descriptions: Fixed:Lifestyle:Diseases, Continued

- Fixed Variables: baseage, deathage and gender: there is a strong cohort effect, that is, for the older cohort at baseage, the data is censored - for a given baseage we do not know anything about those who died before baseline. For the younger cohorts, the data is right censored since we do not know who died after March 2011. The left censoring will be partially accounted for, but the right censoring is a source of bias.
- Lifestyle Variables: edu, bmi, smoke, inc: Of all of the available lifestyle variables observed in BDES baseline, these were most associated with mortality, examining one at a time.
- Diseases: diabetes, cancer, heart, kidney - these are leading causes of death in the US for which we have observations at baseline. We do not have Alzheimer's disease or accident data for this study.

Fixed:Lifestyle:Diseases, Continued

The Questions

- How can you account for the cohort effect?
- How can you compare lifestyle variables to familial relationships?
- To what extent do diseases run in families, and when they do, what are the relative influence of familial and lifestyle factors?

The Results

- A partial accounting for the cohort effect is proposed.
- A novel way to compare lifestyle variables and disease variables familial distances is proposed.
- Only a rough answer here. Can't rule out genetic effects that influence lifestyle factors that are disease risk factors.

Methodology

Our approach is to combine two methological tools, one fairly recent, and one older.

Two methodological tools:

- Distance Correlation: G. Szekely and M. Rizzo (2009) Brownian Distance covariance, *Ann. Appl. Stat* **3**(4)1236-1265. aka "DCOR".
- Smoothing Spline ANOVA, G. Wahba, Y. Wang, C. Gu, R. Klein, B. Klein (1995), Smoothing Spline ANOVA for exponential families, with application to the Wisconsin Epidemiological Study of Diabetic Retinopathy. *Ann. Stat.* **23**(6)1865-1895, books by Chong Gu and Yuedong Wang. aka "SSANOVA".

Methodology Continued

What DCOR and SSANOVA do:

- DCOR looks at the joint distribution of two random Euclidean vectors X and Y and tests, in a completely nonparametric way, whether they are independent or not. It depends only on pairwise distances among the samples from X and among the samples from Y, via a correlation-like statistic. It is ideal for familial effects where only pairwise familial distances are observed. (Need to project familial effects into a Euclidean space)
- SSANOVA is a general method for predicting an outcome nonparametrically as a function of several possibly interacting variables. SSANOVA is here used to determine Lifestyle and Disease scores relating to deathage and use their pairwise distances in DCOR to obtain correlation-like estimates between mortality, familial effects, lifestyle factors and diseases.

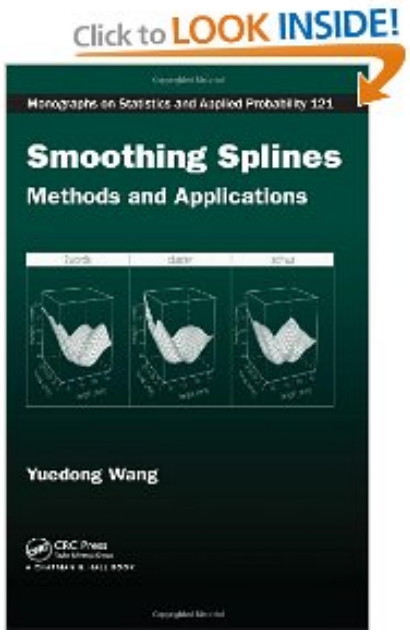


Figure 1: Yuedong Wang, Smoothing Splines Methods and Applications (2011) Rcode: assist

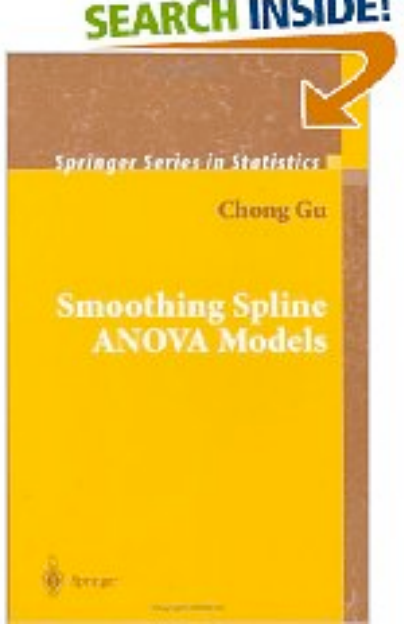


Figure 2: Chong Gu, Smoothing Spline ANOVA Models (2002) Rcode: gss

X. Lin et. al. Smoothing spline ANOVA models for large data sets with Bernoulli observations and the randomized GACV. *Ann. Statist*, 28:1570–1600, 2000.

SSANOVA Continued - for Penalized Least Squares Fits

$$
\begin{aligned}
y_i &= f(x(i)) + \epsilon_i, \quad i = 1, 2, ..., n \\
x(i) &= (x_1(i), x_2(i)), \ldots, x_p(i)). \\
f(x) &= \mu + \sum_{\alpha} f_\alpha(x_\alpha) + \sum_{\alpha<\beta} f_{\alpha\beta}(x_\alpha, x_\beta) + \cdots . \\
min_f I_\lambda(f, y) &= \sum_{i=1}^{n} (y_i - f(x(i))^2 + \lambda J_\theta(f), \\
J_\theta(f) &= \sum_{\alpha} \theta_\alpha J_\alpha(f_\alpha) + \sum_{\alpha<\beta} \theta_{\alpha\beta} J_{\alpha\beta}(f_{\alpha\beta}) + \cdots .
\end{aligned}
$$

The $J_\alpha, J_{\alpha\beta}, \cdots$ are quadratic penalty functionals on the smooth terms, the series is truncated somewhere, and λ and $\theta_\alpha, \theta_{\alpha\beta}, \cdots$, $(\sum \theta \log \theta = 0)$ are smoothing parameters to be chosen, usually by GCV, (Generalized Cross Validation) or other related methods.

Distance Correlation (DCOR)

For a random sample $(X, Y) = \{(X_k, Y_k) : k = 1, ..., n\}$ of n i.i.d random vectors (X, Y) from the joint distribution of random vectors X in R^p and Y in R^q, the Euclidean distance matrices $(a_{ij}) = (|X_i - X_j|_p)$ and $(b_{ij}) = (|Y_i - Y_j|_q)$ are computed. Define the double centering distance matrices

$$A_{ij} = a_{ij} - \overline{a}_{i\cdot} - \overline{a}_{\cdot j} + \overline{a}_{\cdot\cdot}, \quad i, j = 1, ..., n,$$

where

$$\overline{a}_{i\cdot} = \frac{1}{n}\sum_{j=1}^{n} a_{ij}, \quad \overline{a}_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n} a_{ij}, \quad \overline{a}_{\cdot\cdot} = \frac{1}{n^2}\sum_{i,j=1}^{n} a_{ij},$$

similarly for $B_{ij} = b_{ij} - \overline{b}_{i\cdot} - \overline{b}_{\cdot j} + \overline{b}_{\cdot\cdot}, \quad i, j = 1, ..., n.$

The sample distance covariance $\mathcal{V}_n(X, Y)$ is defined by

$$\mathcal{V}_n^2(X, Y) = \frac{1}{n^2} \sum_{i,j=1}^{n} A_{ij} B_{ij}.$$

The sample distance correlation $\mathcal{R}_n(X, Y)$ (DCOR) is defined by

$$\mathcal{R}_n^2(X, Y) = \begin{cases} \dfrac{\mathcal{V}_n^2(X, Y)}{\sqrt{\mathcal{V}_n^2(X)\mathcal{V}_n^2(Y)}}, & \mathcal{V}_n^2(X)\mathcal{V}_n^2(Y) > 0; \\ 0, & \mathcal{V}_n^2(X)\mathcal{V}_n^2(Y) = 0, \end{cases}$$

where the sample distance variance is defined by

$$\mathcal{V}_n^2(X) = \mathcal{V}_n^2(X, X) = \frac{1}{n^2} \sum_{i,j=1}^{n} A_{ij}^2.$$

Pedigree Distance and the Kinship Coefficient 2ϕ

Degree	Genetic relation		Relation
	%	No Genes	
-	100%	30,000	Identical twins
1st	50%	15,000	Parents, siblings, children, fraternal twins
2nd	25%	7,500	Grandparents, grandchildren, aunts, uncles, nieces, nephews, half-siblings, double-cousins (children of 2 siblings x 2 other siblings), identical twin cousins (children of identical twins)
3rd	12.5%	3,750	First cousins, great grandparents, great grandchildren, great aunts, great uncles, grandnieces, grandnephews, half-aunts, half-uncles, half-nephews, half-neices

The Table contains 2ϕ, the kinship coefficient (%). The pedigree distance a_{ij} between person i and person j is defined here as $1 - 2\phi$

Death Age Scoring

Death age as a function of fixed, lifestyle and disease variables will be modeled as

$$\begin{aligned} death\ age_i =& g_0(baseline\ age_i, gender_i) + \\ & g_1(lifestyle\ factors_i) + g_2(diseases_i), \end{aligned}$$

where g_0 is a term involves fixed characteristics, baseline age and gender for individual i, g_1 is a term that includes only lifestyle factors, namely edu, bmi, smoke, inc, and g_2 is a term containing only disease variables, namely diabetes, cancer, cardiovascular disease and chronic kidney disease. In the paper, the fitted values of g_1 and g_2 are treated as scores for the individuals and to be used to assess the association with familial relationships. Do g_1 and g_2 scores, both high and low, run in families, thus partially explaining why mortality runs in families?

The SSANOVA Death Age Scoring Model

The SSANOVA death age scoring model is:

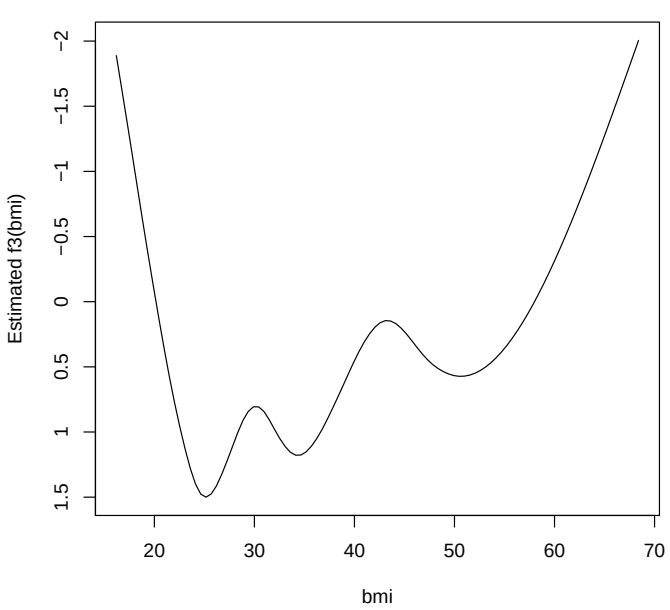
$$
\begin{aligned}
deathage =& \mu + f_1(baseage) + \beta_{gender} I_{\{gender=F\}} && \Big\} fixed \\
&+ f_2(edu) + f_{12}(baseage : edu) + f_3(bmi) && \Big\} lifestyle \\
&+ \beta_{smoke} I_{\{smoke=no\}} + \beta_{inc} I_{\{inc>20T\}} && \\
&+ \beta_{diabetes} I_{\{diabetes=no\}} + \beta_{cancer} I_{\{cancer=no\}} && \Big\} disease \\
&+ +\beta_{heart} I_{\{heart=no\}} + \beta_{kidney} I_{\{kidney=no\}} &&
\end{aligned}
$$

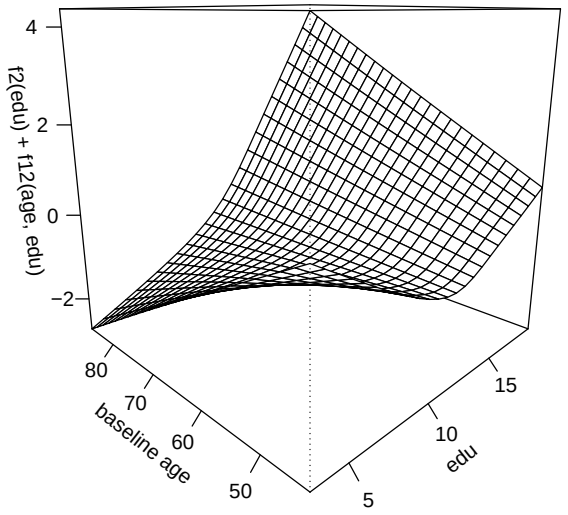
Fitted effects of linear terms in the SS-ANOVA model

gender = F	smoke = no	inc > 20T	
1.141	1.349	0.546	
diabetes = no	cancer = no	heart = no	kidney = no
2.000	0.888	1.131	1.303

Thus, considering only yes-no variables, in this population you are expected to live longer if you are female, don't smoke, aren't poor, don't have diabetes, cancer, heart or kidney problems at baseline.

Fitted effects of continuous terms in the SS-ANOVA model





Left $f_3(\text{bmi})$ vs bmi.

Right: $f_2(\text{edu}) + f_{12}(\text{baseline age, edu})$ vs baseline age and edu.

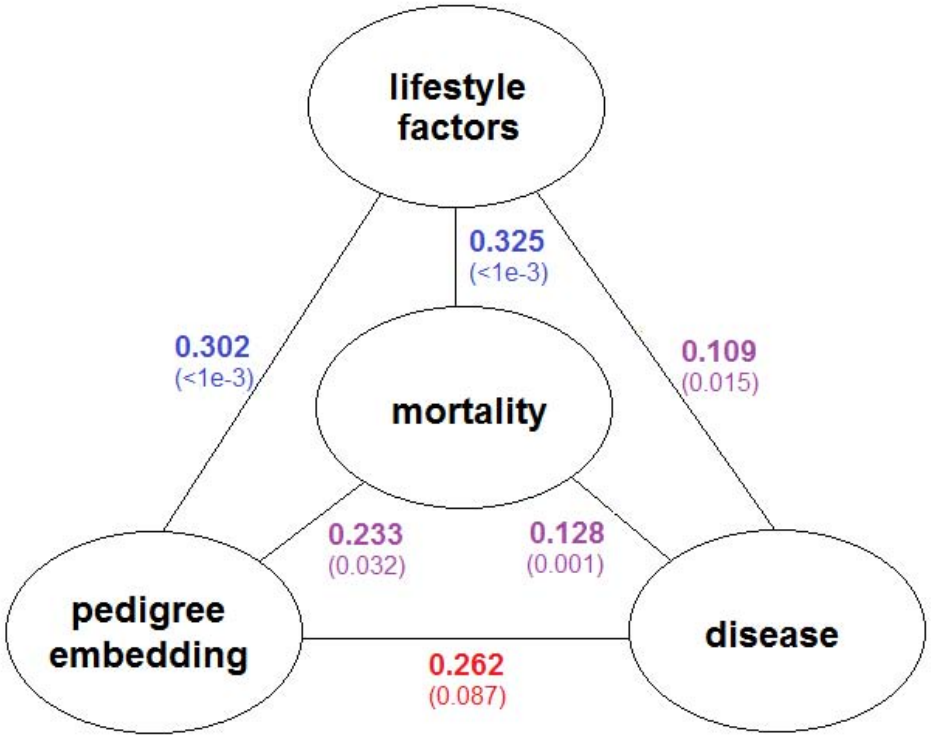
Determining Distance Correlation (DCOR)

All six DCOR values between mortality, pedigree, lifestyle factors and diseases will be computed.

The lifestyle factor score g_1 for an individual is based on the four-vector of the fitted effects for smoke, bmi, edu and inc. Similarly the disease score g_2 is based on the four-vector of fitted effects for the four disease variables.

Pedigree distances are not real distances, they do not in general satisfy the triangle inequality, so they are not even metric. The elegant theory in DCOR assumes that distances are Euclidean or some very specific metric distances. We embedded the pairwise distances into a Euclidean space, and found the results similar. We present the results based on the embedded pedigree distances.

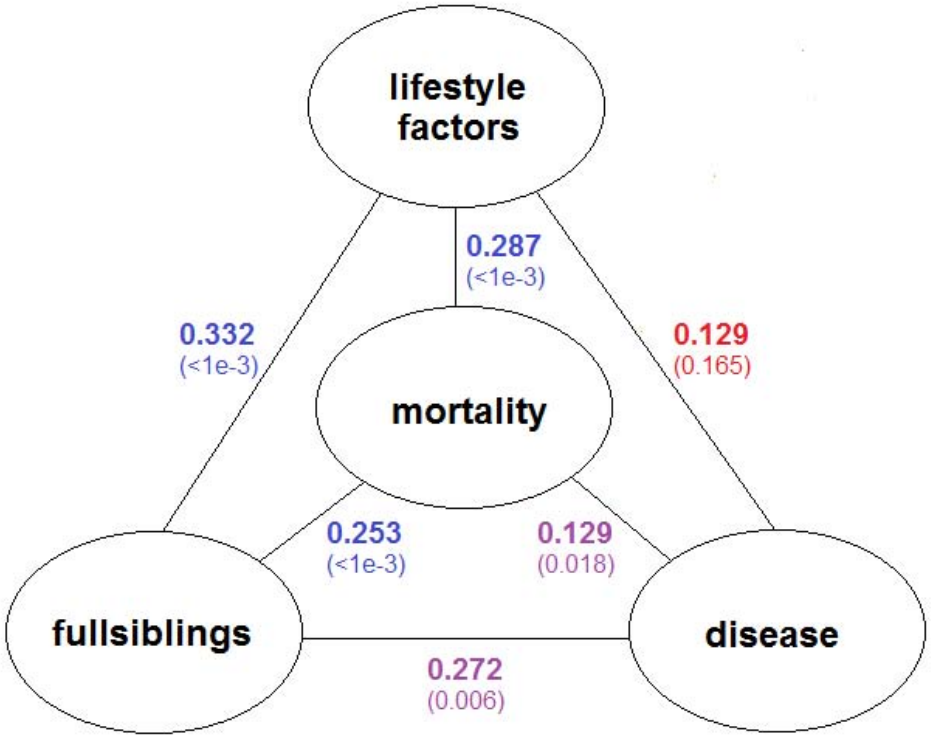
DCOR Results, Entire Pedigrees



very signif-signif
lifestyle:pedigree
lifestyle:mortality
disease:mortality
mortality:pedigree
disease:lifestyle
disease:pedigree

DCOR results using embedded pedigree distance. Estimates agree to two figures with the original pedigree distances. Numbers in parens are significance levels to test independence, based on a permutation test with 1000 replicates.

DCOR Results, Full Sibling Pairs



very signif-signif
lifestyle:fullsibs
lifestyle:mortality
disease:mortality
mortality:fullsibs
disease:lifestyle
disease:fullsibs

DCOR results, full siblings and unrelated pairs. If sib pairs are coded 0 and unrelated pairs coded 1, pedigree data is Euclidean. Mortality more strongly associated with full siblings compared with entire pedigrees. Disease now significantly associated with sibs.

Quantifying Full Sibling Effects

For the full siblings study, we quantify the effects of the relationships between siblings and mortality, lifestyle and disease. Let group 0 be the collection of all pairs of full siblings, and let group 1 be the collection of all unrelated pairs - we get the following table for the means. 95% Bootstrap percentile confidence intervals for the mean differences are based on 10,000 replications.

Quantifying Full Sibling Effects, Continued

Mean differences Bootstrap percentile confidence intervals for the mean differences in the full siblings study, in years

variable	mortality	lifestyle	disease
group 0 mean	8.091	1.405	1.119
group 1 mean	9.662	1.654	1.229
difference	1.571	0.249	0.110
95% CI	(0.919, 2.211)	(0.167, 0.331)	(0.020, 0.202)

We can see that the mortality:fullsibs effect estimate is 1.571 years, the lifestyle:fullsibs effect estimatee is 0.249 years and the disease:fullsibs effect is 0.110 years. An effect purely due to inherited genes is not exactly the same as familial relationships. It is certainly an open question what, beyond lifestyle and (specific) diseases contribute to the total mortality:fullsibs effect.

Summary and Conclusions

Does Life Span Run in Families, and If So, Why? Its nothing new that some families tend to have longer life spans than others. This study suggests that four of the most significant lifespan correlates for mortality, namely edu, bmi, smoke and inc, do as a group run in families, and contribute to the pairwise differences observed in lifespan that are shorter in relatives than in the study population as a whole. Similarly four diseases that we have that are the most frequent cause of death, considered as a group, do run in families, but the combined disease score explains somewhat less than lifestyle variables in this study.

Summary and Conclusions, Continued

Considering diabetes, the strongest cause of death in this study, things become complicated. We do not, for example know whether bmi and smoke, which are known to be major risk factors for diabetes, have genetic polymorphism risk factors in addition to a behavioral component. The American Diabetes Association website says "it may be difficult to figure out whether your diabetes is due to lifestyle factors or genetic suseptibility".. or both. Besides bmi and smoking, diet, which likely runs in families and is likely a risk factor, is remarkably hard to obtain accurately, however. WHI

Final Remarks

- The Beaver Dam Eye Study provided an ideal opportunity to apply some emerging statistical tools to examine questions regarding relationships between various kinds of information collected at the start of the study and mortality, because it has both extensive pedigree information as well as a wealth of covariates of interest. David DeMets was in on the design early, and the BDES study shows the benefits of having a highly qualified statistician in on the study early on.
- The methodological approach we have proposed is easily adaptable to other studies for exploring relations between attributes of subjects with multiple clusters of observable attributes, simultaneously with other factors for which only pairwise relationships are observed.

Caveats

- Some caveats with respect to the mortality data here are: The mortality data is censored at both ends, that is, we do not see cohorts of the oldest subjects who have died before the study began, and, at the other end, we have access to death ages only to those in the study who have died by March 2011. The left censoring is, to some extent accounted for in the presence of baseage in the SS-ANOVA model for deathage–note that there is an interaction term for baseage and education, since it was observed that the oldest cohort in the study clearly had fewer years of formal education than younger members. This study does not use the subjects who would otherwise be included who do not have a recorded death age prior to March 2011. This is a possible source of bias in the conclusions.

More Questions Than Answers

- We have shown that pairwise differences in lifestyle factors that run in families correlate well with pairwise differences in death age that also run in families, partially accounting for the familial death age effect. This leads to new questions to be asked about the complex relations between genetics, family structure, lifestyle factors, and other variables. We provide here an overall methodological approach which shows promise to help in answering these questions in future studies.