Monte Carlo Simulation of OLS and Linear Mixed Model Inference of the Effects of Eudaimonic and Hedonic Well-Being on CTRA Gene Expression

Jeffrey A Walker Corresp. 1

1 Department of Biological Sciences, University of Southern Maine, Portland, ME, United States

Corresponding Author: Jeffrey A Walker
Email address: walker@maine.edu

Background This paper evaluates the performance of gene set association methods in a re-analysis of a recent, high-profile dataset which was used to show opposing effects of hedonic and eudaimonic well-being on the expression levels of a set of genes that have been correlated with social adversity (the CTRA gene set). These effects were inferred using a linear model (GLS) of fixed effects with correlated error to estimate partial regression coefficients.

Methods The standardized effects of hedonic and eudaimonic well-being on CTRA gene set expression estimated by GLS was compared to estimates using multivariate (OLS) linear models and generalized estimating equation (GEE) models. The OLS estimates were tested using O'Brien's OLS test, Anderson's permutation r^2_F test, two permutation *F*-tests (including GlobalAncova), and a rotation *z* test (Roast). The GEE estimates were tested using a Wald test with robust standard errors. The performance (type I, II S, and M errors) of all tests was investigated using a Monte Carlo simulation of data modeled on the re-analyzed dataset.

Results Standardized OLS effects (mean partial regression coefficients) of *Hedonia* and *Eudaimonia* on gene expression levels are very small in both the 2013 and 2015 data, as well as the combined data. The GEE estimates and tests are nearly identical to the OLS estimates and tests. By contrast, the GLS estimates are inconsistent between data sets, but in each dataset, at least one coefficient is large and highly statistically significant. Bootstrap and permutation GLS distributions suggest that the GLS model not only results in downward biased standard errors but also inflated coefficients. Both distributions also show the expected, strong, negative correlation between the coefficients for *Hedonia* and *Eudaimonia*. The Monte Carlo simulation of error rates shows highly inflated type I error from the GLS test and slightly inflated type I error from the GEE test. By contrast, type I error for all OLS tests are at the nominal level. Of the OLS tests, the permutation *F*-tests have ~ 1.8X the power of the O'Brien's, Anderson's, and Roast tests. This increased power comes at a cost of high sign error (~ 10\%) if tested on small effects.

Discussion The apparently replicated pattern of hedonic and eudaimonic effects on gene expression is most parsimoniously explained as "correlated noise" due to the geometry of multiple regression. A linear mixed model for estimating fixed effects in designs with many repeated measures or outcomes should be used cautiously because of the potentially inflated type I and type M error. By contrast, permutation *F* tests of OLS estimates have good performance, including moderate power (0.42 --.47) for very small effects.

Monte Carlo Simulation of OLS and Linear Mixed Model Inference of the Effects of Eudaimonic and Hedonic Well-Being on CTRA Gene Expression

Jeffrey A. Walker[1]

[1]**Biological Sciences, University of Southern Maine, 70 Falmouth St., Portland, ME, USA 04103**

ABSTRACT

Background This paper evaluates the performance of gene set association methods in a re-analysis of a recent, high-profile dataset which was used to show opposing effects of hedonic and eudaimonic well-being on the expression levels of a set of genes that have been correlated with social adversity (the CTRA gene set). These effects were inferred using a linear model (GLS) of fixed effects with correlated error to estimate partial regression coefficients.

Methods The standardized effects of hedonic and eudaimonic well-being on CTRA gene set expression estimated by GLS was compared to estimates using multivariate (OLS) linear models and generalized estimating equation (GEE) models. The OLS estimates were tested using O'Brien's OLS test, Anderson's permutation r^2_F test, two permutation F-tests (including GlobalAncova), and a rotation z test (Roast). The GEE estimates were tested using a Wald test with robust standard errors. The performance (type I, II S, and M errors) of all tests was investigated using a Monte Carlo simulation of data modeled on the re-analyzed dataset.

Results Standardized OLS effects (mean partial regression coefficients) of $Hedonia$ and $Eudaimonia$ on gene expression levels are very small in both the 2013 and 2015 data, as well as the combined data. The GEE estimates and tests are nearly identical to the OLS estimates and tests. By contrast, the GLS estimates are inconsistent between data sets, but in each dataset, at least one coefficient is large and highly statistically significant. Bootstrap and permutation GLS distributions suggest that the GLS model not only results in downward biased standard errors but also inflated coefficients. Both distributions also show the expected, strong, negative correlation between the coefficients for $Hedonia$ and $Eudaimonia$. The Monte Carlo simulation of error rates shows highly inflated type I error from the GLS test and slightly inflated type I error from the GEE test. By contrast, type I error for all OLS tests are at the nominal level. Of the OLS tests, the permutation F-tests have $\sim 1.8\times$ the power of the O'Brien's, Anderson's, and Roast tests. This increased power comes at a cost of high sign error ($\sim 10\%$) if tested on small effects.

Discussion The apparently replicated pattern of hedonic and eudaimonic effects on gene expression is most parsimoniously explained as "correlated noise" due to the geometry of multiple regression. A linear mixed model for estimating fixed effects in designs with many repeated measures or outcomes should be used cautiously because of the potentially inflated type I and type M error. By contrast, permutation F tests of OLS estimates have good performance, including moderate power (0.42 –.47) for very small effects.

Keywords: gene set association, self-contained test, type I error, power, type M error, type S error, permutation, multiple outcomes

INTRODUCTION

The motivation for this work is the recent implementation of a linear model (GLS) with correlated error to estimate the mean effect of a phenotype on a set of expression levels for a gene set identified *a priori* (Fredrickson et al., 2015). Specifically, the authors reported a large, negative, highly statistically significant, standardized effect (partial regression coefficient) of eudaimonic well-being on the "conserved transcriptional response to adversity" (CTRA) gene set (Fredrickson et al., 2015). Additionally, using

OLS estimates, the authors reported opposing effects of eudaimonic and hedonic well being on CTRA expression (Fredrickson et al., 2015), a pattern that they argue replicates the results of an earlier 2013 study (Fredrickson et al., 2013). The development of such *gene set association* methods is an active area of research in genomics (Tian et al., 2005; Goeman and Bühlmann, 2007; Hummel et al., 2008; Wu et al., 2010; Tripathi and Emmert-Streib, 2012; Zhou et al., 2013). The effect of a phenotype on the mean response of multiple outcomes, has a long and rich history in applied statistics, especially in the context of clinical outcomes in medicine (O'Brien, 1984; Pocock et al., 1987; Lauter, 1996; Bull, 1998). The development of test methods within the field of genomics has advanced largely without reference to this earlier literature (but see Chen et al., 2007; Tsai and Chen, 2009). The authors of the CTRA gene set association paper (Fredrickson et al., 2013, 2015) failed to draw on either this earlier clinical literature or the more recent gene set association literature.

Here, I re-analyze the dataset of Fredrickson et al. (2015), and the earlier "discovery" dataset (Fredrickson et al., 2013), with two goals in mind. First, and more specifically, I re-analyze the datasets using alternatives to GLS for testing for fixed effects on multiple responses. The alternative methods include O'Brien's OLS test for multiple outcomes (O'Brien, 1984), a permutation r_F^2-test (Anderson and Robinson, 2001), two different permutation partial F-tests, including GlobalAncova (Hummel et al., 2008), a rotation z-test (Roast) (Wu et al., 2010), and a Wald test using robust standard errors (Zeger and Liang, 1986). Second, and more generally, I use Monte Carlo simulation experiments to evaluate the GLS with correlated error and the alternative gene set association methods when used on simulated data explicitly modeled on the dataset from Fredrickson et al. (2015). By simulating a specific dataset, I can compare test methods without extrapolating from more general, theoretical studies. Type I, Type II, Type M and Type S errors errors are investigated, where type S and M errors are errors in the sign and magnitude of the effect when $p \leq 0.05$ (Gelman and Carlin, 2014).

Background

The 2015 study (Fredrickson et al., 2015) was a replication of an earlier study that reported not only a negative effect of eudaimonic well-being on CTRA expression but also a positive effect of hedonic well-being on CTRA expression (Fredrickson et al., 2013). While the 2015 study did not report a statistically significant effect of hedonic well-being on CTRA expression, it did report plots showing opposing effects of hedonic and eudaimonic scores on mean CTRA gene expression that "replicated" that of Fredrickson et al. (2013). This apparent replication of opposing effects was again emphasized in a published correction (Fredrickson et al., 2016).

The CTRA gene set includes 19 pro-inflammatory, 31 anti-viral, and 3 antibody-stimulating genes. The Fredrickson et al. (2013) data included all 53 genes but the Fredrickson et al. (2015) data is missing IL-6 from the pro-inflammatory subset. Fredrickson et al. (2013) used 53 univariate multiple regressions to estimate the effects (the regression coefficient) of each well-being (hedonic and eudaimonic) score on log2(normalized gene expression) for each gene. The regression model included both well-being scores, seven covariates to adjust for demographic and general health confounding (sex, age, ethnicity, BMI, a measure of alcohol consumption, a measure of smoking, and a measure of recent illness), and eight expression levels of T-lymphocyte markers to adjust for immune status confounding. Hedonic and eudaimonic scores were transformed to z-scores prior to the analysis. The 53 multiple regressions (one for each gene) yielded 53 coefficients for hedonic score and 53 coefficients for eudaimonic score. The coefficients of the 31 anti-viral and 3 antibody genes were multiplied by -1 to make the direction of the effect consistent with the CTRA response. Fredrickson et al. (2013) used a simple one-sample t-test of the 53 coefficients to test for a mean effect of hedonic or eudaimonic score on CTRA expression. A mean coefficient greater than zero reflects a positive CTRA response (increased pro-inflammatory and decreased anti-viral and antibody-stimulating genes).

Fredrickson et al. (2013) used a bootstrap to re-sample the coefficients in order to generate a standard error (the denominator of their t-value and then tested the statistic using $m-1$ degrees of freedom, where m is the number of outcomes (gene expression levels). There are two fundamental problems with this t-test. First, the coefficients are not independent of each other because of the correlated expression levels among genes and as a consequence the standard error in the denominator will be too small, which should result in an inflated Type I error rate. Second, their degrees of freedom does not account for the number of subjects in the study. At the extreme, if only a single gene expression level is measured, Fredrickson's t cannot even be computed. This second error should result in loss of power. The combined effect on Type I and Type II error will depend on the magnitude of the correlations among the expression levels.

Through simulation, however, Brown et al. (2014) discovered an inflated Type-I error in their exploration of the data using the Fredrickson et al. (2013) t-test. O'Brien (1984) developed an appropriate t-test for the effects of an an independent variable on multiple outcomes (see below).

Fredrickson et al. (2015) replicated the 2013 study but treated the 52 gene expression levels as "repeated" measures (or multiple outcomes) of a single expression response and used a linear model with fixed-effects and correlated error to estimate the regression coefficients of expression on hedonic and eudaimonic score. Specifically, Fredrickson et al. (2015) used generalized least squares (GLS) with a heterogenous compound symmetry error matrix to estimate the marginal (population-averaged) fixed effects. Compound symmetry assumes equal correlation (conditional on the set of predictors) among all expression levels. This is not likely to approximate the true error structure for a set of expression levels for different genes, as these expression levels will share different sets of underlying regulatory factors. Fredrickson et al. (2015) re-ran the analysis using an unstructured error matrix, with results contradicting the compound symmetry results, but chose to report this in the supplement and not the main text. Regardless, linear mixed models for repeated measures or multiple outcomes are prone to inflated Type I error due to both upward biased effect estimates and downward biased standard errors (Kackar and Harville, 1984; Kenward and Roger, 1997; Guerin and Stroup, 2000; Littell et al., 2006; Jacqmin-Gadda et al., 2007; Gurka et al., 2011). The amount of bias depends on the true and specified correlation structure, as well as effective sample size (a function of the number of subjects, the number of outcomes, and the correlations among the outcomes), but can be large even with large samples (Gurka et al., 2011). When only the marginal effects are of interest (as here), population-averaged effects are typically estimated using Generalized Estimating Equations (GEE) instead of GLS and standard errors robust to model misspecification are computed using the sandwich estimator (Liang and Zeger, 1986; Zeger and Liang, 1986).

Finally, Cole et al. (2015) replicated the 2015 study, using the same GLS method, but only reported the effect of *Eudaimonia* and not *Hedonia* on CTRA expression. For this study, an unstructured error matrix was specified and the expression levels were not standardized to z-scores. Again, a negative effect of *Eudaimonia* on CTRA expression was reported but the magnitude cannot be easily compared to other values because of the lack of standardization.

METHODS

Data were downloaded as .txt Series Matrix Files from http://www.ncbi.nlm.nih.gov/geo/ using accession numbers GSE45330 (the 2013 dataset, hereafter FRED13), GSE55762 (the focal 2015 dataset, hereafter FRED15) and GSE68526 (the replicate 2015 dataset, hereafter COLE15). The CTRA (response) expression data were log2 transformed. The T-lymphocyte expression data that formed part of the set of covariates were log2 transformed in the downloaded data. The downloaded hedonic and eudaimonic scores in FRED13 had means and variances close but not equal to that expected of z-scores, which suggests that the public data slightly differs from that analyzed by Fredrickson et al. (2013); these were re-standardized to z-scores Three rows of FRED13 had missing covariate data (two rows were completely missing) and were excluded; the number of rows (subjects) in the cleaned matrix was 76. The downloaded hedonic and eudaimonic scores in FRED15 were the raw values and were transformed to z-scores. There was no missing data in FRED15 and the number of subjects was 122. The FRED15 data did not include a measure for *Hedonia*. Excluding rows with missing values left complete data for 108 subjects.

Prior to all analyses, *Hedonia* or *Eudaimonia* scores and the expression levels of all genes were standardized to mean zero and unit variance. Additionally, the 31 anti-viral and 3 antibody genes were multiplied by -1 to make the direction of the effect consistent with the CTRA response (Fredrickson et al., 2013, 2015).

Null hypothesis tests

If β_j is the the effect (partial regression coefficient) of *Hedonia* or *Eudaimonia* on the expression level of the jth gene, the overall effect of *Hedonia* or *Eudaimonia* on expression of the CTRA gene set is simply the averaged coefficient over all genes, $\bar{\beta} = \frac{1}{m}\sum\beta_j$ where m is the number of genes. The three focal null hypotheses that are tested here are $H_0 : \bar{\beta}_{hedonia} = 0$, $H_0 : \bar{\beta}_{eudaimonia} = 0$, and $H_0 : \delta_{hed-eud} = \bar{\beta}_{hedonia} - \bar{\beta}_{eudaimonia} = 0$. All three hypotheses are directional, that is, the mean effect differs from zero. This differs from the general multivariate test that at least one of the coefficients

differs from zero, but the mean response may be zero. While the hypotheses are directional, the tests are two-tailed, that is, the mean response may be up or down regulation of the CTRA gene set.

OLS inferential tests

The effects of *Hedonia* and *Eudaimonia* on the mean of the m gene expression levels are estimated with the multivariate linear model

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \tag{1}$$

where $\mathbf{Y}$ is the $n \times m$ matrix of gene expression levels for the n subjects, $\mathbf{X}$ is the model matrix of dummy variables and covariates, $\mathbf{E}$ is the matrix of residual error, and $\mathbf{B}$ is the $p \times m$ matrix of partial regression coefficients. For the combined data, the model matrix includes a dummy variable indicating dataset (2013 or 2015). The coefficients of the jth column of $\mathbf{B}$ are precisely equal to those from a univariate multiple regression of the jth gene on $\mathbf{X}$ (and why the model is sometimes called a multivariate multiple regression). In R, estimating the m effects of *Hedonia* and *Eudaimonia* is much faster using this multivariate model than looping through m univariate multiple regressions. The mean of the m coefficients is the OLS estimate of the effect of *Hedonia* or *Eudaimonia* on overall CTRA expression level. Because the well-being scores for *Hedonia* and *Eudaimonia* and the m expression levels were mean-centered and variance-standardized, the reported OLS estimates are mean (averaged over the m genes) standard partial regression coefficients.

O'Brien's OLS t-test

O'Brien's OLS test (O'Brien, 1984; Logan and Tamhane, 2004; Dallow et al., 2008) was developed explicitly for testing the directional hypothesis that the mean effect of multiple outcomes differs from zero, which is precisely the question pursued in both Fredrickson papers. Given m standardized regression coefficients and associated t-values, O'Brien's test statistic is

$$t_{Obrien} = \frac{\mathbf{j}^\top \mathbf{t}}{\sqrt{\mathbf{j}^\top \mathbf{R} \mathbf{j}}} \tag{2}$$

where $\mathbf{j}$ is a m vector of 1s, $\mathbf{t}$ contains the t-values associated with each of the m partial regression coefficients, and $\mathbf{R}$ is the conditional correlation matrix of the m expression levels. $\mathbf{R}$ was computed separately for *Hedonia* and *Eudaimonia* from the residuals of the multivariate linear model (equation 1) with all covariates in the model except the focal covariate (the reduced model). A t distribution with $n-m$ degrees of freedom was used to test t_{Obrien} against the null.

Anderson's permutation r^2_F-test

As an alternative to the parametric O'Brien's OLS test, I use four different permutation, or permutation-like tests. The first of these is Anderson's permutation r^2_F-test of the partial correlation between the focal predictors (Z) and the outcomes (Y) conditional on the covariates (X) (any of these can be univariate or multivariate) (Anderson and Robinson, 2001; Anderson, 2001). This test uses a permutation of the residuals of the null model ("permutation under the null") as these residuals, but not the Y, X, or Z, are exchangeable. Freedman and Lane (1983), initially developed the permutation under the null using a t-value of the effect (for a gene set with $m > 1$, this statistic would be the mean or sum of the t-values for each gene). Anderson provided both theoretical and empirical evidence for the superior performance of the permutation under the null, but used the squared partial correlation coefficient $r^2_F = \rho^2_{zy.x}$ in place of t. For a permutation under the null, the predictor variables are divided into main effects $\mathbf{Z}$ (hedonic or eudaimonic scores were tested independently) and nuisance covariates $\mathbf{X}$ (the demographic and immune variables plus the well-being score not being tested) and the model becomes

$$\mathbf{Y} = \mathbf{XA} + \mathbf{ZB} + \mathbf{E} \tag{3}$$

and the two-sided test statistic is

$$r^2_F = \frac{(\sum \mathbf{E}_{\pi(F)} \mathbf{E}_{Z|X})^2}{\sum \mathbf{E}^2_{\pi(F)} \sum \mathbf{E}^2_{Z|X}} \tag{4}$$

Where $\mathbf{E}$ are the residuals from different fit models. The computations for this are

1. Compute $\mathbf{E}_{Z|X}$, which are the residuals of $\mathbf{Z}$ regressed on $\mathbf{X}$
2. Set $\mathbf{B} = 0$ and fit model 3. The residuals from this model are the estimated residuals under the null ($\mathbf{E}_{Y|X}$) and the predicted values are the estimated Y under the null ($\hat{\mathbf{Y}}$).
3. Permute the residuals under the null and add to the predicted values under the null, $\mathbf{Y}_\pi = \hat{\mathbf{Y}} + \mathbf{E}^\pi{}_{Y|X}$, where π indicates permutation
4. Fit model model 3, again with $\mathbf{B} = 0$ but substitute $\mathbf{Y}_\pi$ for $\mathbf{Y}$, and compute the residuals from this fit, which are the $\mathbf{E}_{\pi(F)}$

To generate the null distribution of the test statistic, 10,000 permutations were run, including an iteration of non-permuted data. The two-sided p-value of each hypothesis was computed as the fraction of $r_F^2 \geq$ the observed r_F^2.

Permutation F_{ga}-test (GlobalAncova)

Because it is implemented in the GlobalAncova package (Mansmann et al., 2010), the permutation F-test described in Hummel et al. (2008) is an attractive alternative to the permutation r_F^2-test. The GlobalAncova test statistic, F_{ga} compares the residual sums of squares of the reduced model not including the predictor(s) of interest (Z) to the residual sums of squares of the full model. The full model 3 is fit but substituting the residuals of the reduced model ($\mathbf{E}^\pi{}_{Y|X}$) for $\mathbf{Y}$. GlobalAncova permutes the main effects ($\mathbf{Z}$), and thus does not preserve the covariance relationship between the main effects and the nuisance effects. While the Z are exchangeable under the null in experimental designs when they are assigned randomly, they are not exchangeable under the null in observational designs (Freedman and Lane, 1983; Anderson, 2001). The consequences of this violation of exchangeability is unknown but may be minor for these data given the small covariances between the main and nuisance effects. 10,000 permutations were run.

Permutation F_{pun}-test

The GlobalAncova F_{ga}-test had high power with these data (see results) but a concern was this high power was due to the violation of exchangeability. Consequently, I implemented a modification of the GlobalAncova test by permuting the residuals under the null to generate a permuted response ($\mathbf{Y}_\pi$) (see above description for the permutation r_F^2-test). Each iteration, the full models and reduced models were fit to $\mathbf{Y}_\pi$ and the respective residual sums of squares were computed (note that in GlobalAncova, the reduced-model residual sums of squares are only computed once, for the observed data). I refer to this as the permutation F_{pun}-test (for "permutation under the null"). 10,000 permutations were run, including an iteration of non-permuted data. The two-sided p-value of each hypothesis was computed as the fraction of $F_{pun} \geq$ the observed F_{pun}.

Rotation z-test (ROAST)

The rotation-test described in Wu et al. (2010) is an attractive alternative to the permutation tests as it is small-sample exact and is implemented in the function roast from the limma package (Ritchie et al., 2015) (by contrast, permutation tests are only asymptotically exact). The test statistic, z_{rot}, is a mean z-score computed from the set of m moderated t-statistics computed for each gene. Using a hierarchical model, the moderated t-statistic uses information on the error of all genes in the set to estimate the gene specific standard error. A p-value for the test statistic is evaluated in a very similar manner to the permutation tests described above, but, instead of permutation, the n-vector of reduced-model residuals is rotated by a random vector r, which is constant for all genes within each iteration but variable among iterations. The observed and rotated z-scores from 10,000 rotations were used to generate the null distribution. The p-value for the "UpOrDown" test was used as this is the test of the two-tailed directional hypothesis.

Inference using linear model with fixed effects and correlated error

The model used by Fredrickson et al. (2015) is

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i \tag{5}$$

where $\mathbf{y}_i$ is the vector of m responses for subject i, $\mathbf{X}_i$ is the model matrix for subject i, which includes the main effect *Gene* to identify the jth element of $\mathbf{y}_i$, and $\boldsymbol{\beta}$ is the vector of fixed (or population-averaged) effects. In this model, $\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is the within subject error covariance matrix representing

the correlated errors. The correlated errors result from random effects due to subjects but the model does not explicitly model these. To implement this model, the data matrix with separate columns for each gene is stacked into long format by combining the m expression levels into a single response variable (*Expression*) and the variable *Gene* is created to identify the gene associated with a specific response (expression value). The univariate regression of *Expression* on the set of predictors results in the same OLS estimates as in the multivariate model described above. These estimates are unbiased but the standard errors for the estimates are incorrect because of the correlated errors. As in the multivariate model for the combined data, the model matrix includes a dummy variable indicating dataset (2013 or 2015).

Generalized Estimating Equations

Fredrickson et al. (2015) used maximum likelihood with the GLS model with a heterogenous compound symmetry error matrix to estimate the fixed effects in equation 5. To replicate these results, I estimated the fixed effects and their standard errors and p-values using the gls function from the nmle package (Pinheiro et al., 2015) (with a heterogenous compound symmetry error matrix and using maximum likelihood method). Additionally, because only the fixed effects are of interest, and because of the known bias in the standard errors of the GLS with correlated errors model, I used Generalized Estimating Equations (GEE) with an exchangeable error matrix to estimate the fixed effects using the function geeglm in the geepack package (Yan, 2002). The default sandwich estimator was used to compute the standard errors of the effects, which is robust to error covariance misspecification (Liang and Zeger, 1986). Nevertheless, GEE is less efficient if the error covariance is misspecified (Sammel and Ryan, 2002).

Permutation and bootstrap GLS

Exploration of the behavior of the GLS as implemented by Fredrickson et al. (2015) suggested partial regression coefficients that were more unstable than implied by the standard error. To explore the consequences of this instability on inference, I implemented both a bootstrap procedure to compute approximate standard errors and the Freedman and Lane permutation procedure (Anderson and Robinson, 2001) described above to compute permutation-GLS p-values. Each iteration of either the bootstrap or the permutation, the data were resampled (or the residuals permuted) in wide format, rescaled, and reshaped to long format. Coefficients were estimated using the gls function as described above. The first iteration used the observed (not resampled) data. The standard partial regression coefficients and associated t-values for $Hedonia$ and $Eudaimonia$ were saved each iteration and used to generate standard errors for the bootstrap and a null distribution of t-values for the permutation. Because the time required to fit the GLS, and the exploratory goal of this analysis, I used only 299 iterations, which is sufficient for approximate, exploratory values. The regressor *Smoke* was excluded from the bootstrap analysis because some bootstrap samples had zero cases with level $Smoke = 1$, which leads to an unsolvable model. Because of this slightly different specification, I limited the bootstrap analysis to the FRED15 dataset. The GLS coefficients for FRED15 with and without *Smoke* in the model are 0.086 and 0.032 for $Hedonia$ and -0.511 and -0.568 for $Eudaimonia$.

Monte Carlo simulations of errors

I used Monte Carlo simulation to explore type I, type II, type S, and type M errors with data similar in structure to the focal Fred15 dataset. Type S and type M error are the errors in the sign and magnitude of the estimated coefficient when $p < 0.05$ (Gelman and Carlin, 2014). For type S error, I used the (frequentist) probability that the sign of the coefficient is wrong when $p < 0.05$, which is $\frac{N(p<0.05|\beta_{estimated}<0,\beta_{true}>0)}{N(p<0.05)}$, where $N(x)$ is the number of iterations with condition x (Gelman and Carlin, 2014). For type M error, I used the exaggeration ratio ($\frac{\beta_{estimated}}{\beta_{true}}$) (Gelman and Carlin, 2014). In each run of the simulation, a random $n \times p$ matrix $\mathbf{X}$ of independent variables (n samples of p covariates) and a random $n \times m$ matrix $\mathbf{Y}$ of response variables (n samples of m responses) were generated using the function rmvnorm from the mvtnorm package (Genz et al., 2015). All simulated independent variables were modeled as continuous variables sampled from $\mathcal{N}(\mathbf{0}, \mathbf{S}_X)$, where $\mathbf{S}_X$ is the covariance matrix of the 17 regressor variable from FRED15. The 52 response variables were modeled as continuous variables sampled from $\mathcal{N}(\mathbf{0}, \mathbf{S}_Y)$, where $\mathbf{S}_Y$ is the covariance matrix of the 52 gene expression levels from FRED15. For the power simulations (including type S and M errors), the standardized effect of $Eudaimonia$ on the mean response was set to 0.067, which is the estimated, standardized effect for the FRED15 dataset. The effect of all other covariates, including that of $Hedonia$ was set to zero. For the type I simulations, all effects were set to zero. Sample size (the number of subjects n) was set to that for FRED15 (122) for all runs. To

Table 1. GLS estimates of the variance-standardized coefficients for the 2013, 2015, and combined data. The GLS-permutation p-values are also given. $\delta_{hed-eud}$ is the difference in the estimates: $\beta_{hedonia} - \beta_{eudaimonia}$

Type	Data	Estimate	SE	p	$p_{gls-perm}$
Hedonia	FRED13	0.537	0.172	0.002	0.29
	FRED15	0.086	0.122	0.484	0.74
	FRED13+15	0.073	0.042	0.081	0.36
Eudaimonia	FRED13	0.135	0.177	0.443	0.83
	FRED15	-0.511	0.126	< 0.001	0.16
	FRED13+15	-0.116	0.043	0.007	0.19
$\delta_{hed-eud}$	FRED13	0.401	0.331	0.225	0.73
	FRED15	0.596	0.231	0.01	0.3
	FRED.13+15	0.189	0.079	0.017	0.26

explore the consequences of increasing increasing m on error rates, the simulation was run with three levels of m (10, 30, 52). The $m \times m$ covariance matrix used to generate $\mathbf{Y}$ using the rmvnorm function was a random sample of $\mathbf{S}_Y$ each iteration. In each iteration of the simulation, the permutation and rotation null distributions were generated from 2000 permuted samples.

Correlated estimation error

Regressors with a high positive correlation, as with *Hedonia* and *Eudaimonia*, have negatively correlated partial regression coefficients. I give a brief mathematical explanation of this in the discussion but also show this empirically using Monte Carlo simulation. The simulation was implemented precisely as described for the type I simulation above, except that I only simulated all $m = 52$ gene expression levels in the FRED15 dataset. Each run of the simulation, the coefficients for *Hedonia* and *Eudaimonia* were estimating using GLS, OLS (multivariate), and GEE. 100 iterations were run to generate 100 pairs of points for the correlation.

The COLE15 dataset did not include a measure of *Hedonia* and was not analyzed using the alternative gene set methods but was analyzed using the GLS only to compare three different datasets using this method. All analyses were performed using R (R Core Team, 2015). All data cleaning and analysis scripts are available at the public GitHub repository https://github.com/middleprofessor/happiness.

RESULTS

GLS replication of previous analyses

The variance-standardized effects and p-values for hedonic and eudaimonic scores estimated from the GLS for each dataset are given in Table 1. My estimates for FRED15 and the combined FRED13+15 are within 0.002 standard units of those reported in Fredrickson et al. (2015). Fredrickson et al. (2015) do not report the GLS results for the 2013 data alone and the reported results for COLE15 are from unstandardized expression levels and specifying an unstructured error matrix. My estimates of the coefficients for the FRED13 data show a pattern opposite to that for FRED15. That is, with the 2013 data, the effect of *Hedonia* is large and has a very small p-value (0.002) while the effect for *Eudaimonia* is small and not statistically significant ($p = 0.44$). My FRED13 coefficients are the same as those reported in the exploratory re-analysis of the FRED13 and FRED15 datasets by Brown et al. (2016), who also note the opposite pattern from the 2015 results. Following z-score standardization of the expression levels, the effect of *Eudaimonia* on mean expression level for the COLE15 data is trivially negative and not statistically significant. By contrast the standardized effects of *Eudaimonia* on mean expression level when re-analyzed without *Hedonia* as a covariate (to conform to the COLE15 analysis) are large and positive for FRED13 and large and negative for FRED15 (Table 2). A diagnostic plot of residual versus fitted values from the GLS model for FRED15 suggests strongly biased estimates (Fig. 1A).

New results

Standardized mean effects ($\bar{\beta}$) estimated by multivariate regression (OLS) are very small and positive for *Hedonia* and very small and negative for *Eudaimonia* for both 2013 and 2015 datasets and the

Table 2. GLS estimates of the variance-standardized coefficients for the FRED13, FRED15, and COLE15 data. *Hedonia* was excluded from the FRED13 and FRED15 analyses and *hispanic* and $ln(hh.income)$ were excluded from the COLE15 analysis so that all analyses had the same set of covariates.

Data	Estimate	SE	p
FRED13	0.558	0.107	< 0.001
FRED15	-0.519	0.086	< 0.001
COLE15	-0.007	0.076	0.931

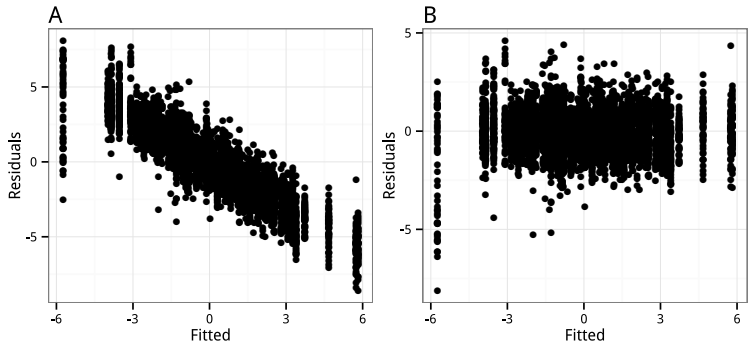

Figure 1. Residual versus fitted values from the A. fixed effects linear model with correlated error (GLS), and B. Generalized Estimating Equations (GEE) model. The pattern in *A* indicates strongly biased estimates.

combined dataset (Table 3). The bootstrap SE for each effect is too large, relative to the signal, to have any confidence in the direction of either of the effects for any dataset. p-values from all OLS tests (with one exception) for all three hypotheses, for all three datasets fail to reject the null. The exception is the p-value from the GlobalAncova test for *Eudaimonia* for the combined (FRED13+15) dataset. The p-values from O'Brien's OLS t-test, Anderson's r^2_F-test, and the Roast test are very similar to each other across all tests. Similarly, the p-values from the two permutation F-tests (GlobalAncova and F_{pun}) are very similar to each other across all tests. The GEE estimates are the same as the OLS estimates to the 2nd decimal place for all three datasets and the robust standard errors are large relative to the coefficients (Table 4). The GEE p-values are very similar to those from the OLS tests (especially the grouping of O'Brien's OLS t-test, Anderson's r^2_F-test, and the Roast test) for all three datasets and fail to reject any of the nulls. A diagnostic plot of residual versus fitted values from the GEE model for FRED15 does not suggest biased estimates (Fig. 1B).

The GLS permutation p-values fail to reject the nulls for any of the tests (Table 1). Compared to the GEE p-values, the GLS permutation p-values are much less similar to those from the OLS tests. The GLS bootstrap distributions of standardized effects for *Hedonia* and *Eudaimonia* for FRED15 are shown in Fig. 2. The standard errors of the effects computed from these distributions are 0.27 for *Hedonia* and 0.36 for *Eudaimonia*, which are 2–3 times the standard errors computed by the GLS model.

Test size, power, sign error, and magnitude error

The GLS test has inflated type I error rates that increases with the number of outcomes (m). At the size of the FRED15 data ($m = 52$), type I error is above 25% for the GLS test (Fig. 3A). By contrast, Type I error for all alternative methods are relatively stable as m increases. Notably, Type I error for all OLS tests are near the nominal level (0.05) while that for GEE is slightly elevated (0.08). Power of all tests increases from $m = 10$ to $m = 30$ but has variable behavior between $m = 30$ and $m = 52$ (Fig. 3B). GlobalAncova and F_{pun}-tests have over $\sim 1.8\times$ the power of O'Brien's, Anderson's R^2_F and Roast tests when $m = 52$, without inflation of type I error. GLS also has relatively high power when $m = 52$ but this comes at a large cost of type I error. GEE has about $1.2\times$ the power of O'Brien's, Anderson's R^2_F, and Roast tests when $m = 52$, but at a small cost of type I error. Type S errors are trivially low (0.001 – 0.002) for GEE,

Table 3. OLS estimates of mean effects (Estimate) on CTRA gene expression. The estimates are the mean variance-standardized partial regression coefficients from the multivariate regression over the m responses (genes). $\delta_{hed-eud}$ is the difference in mean effect. The SEs were estimated using a bootstrap in which entire rows of the dataset were re-sampled to preserve all covariances (2000 bootstrap samples, including the observed sample, were used for the standard error). The p-values are from O'Brien's OLS t-test, Anderson's r^2_F-test, GlobalAncova test, F_{pun}-test, and Roast.

Type	Data	Estimate	SE	$p_{O'Brien}$	p_{r2}	p_{GA}	p_F	p_{Roast}
Hedonia	FRED13	0.026	0.121	0.80	0.77	0.86	0.84	0.78
	FRED15	0.062	0.044	0.24	0.21	0.28	0.28	0.23
	FRED13+15	0.052	0.041	0.22	0.24	0.38	0.41	0.22
Eudaimonia	FRED13	-0.063	0.128	0.50	0.48	0.78	0.78	0.48
	FRED15	-0.067	0.048	0.20	0.19	0.12	0.12	0.20
	FRED13+15	-0.064	0.038	0.14	0.23	0.04	0.06	0.13
$\delta_{hed-eud}$	FRED13	0.089	0.242	0.64				0.60
	FRED15	0.129	0.085	0.22				0.19
	FRED13+15	0.116	0.074	0.17				0.14

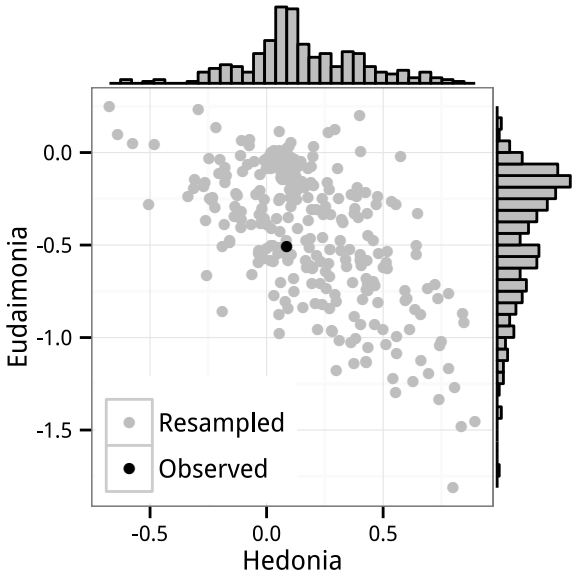

Figure 2. Distribution of GLS bootstrap resampled standard partial regression coefficients for *Hedonia* and *Eudaimonia*. The data are the FRED15 dataset and the coefficients were estimated by the linear model with correlated error (GLS). Also shown is the observed value for FRED15(black).

Table 4. Generalized Estimating Equations estimates of the effects and difference in effects ($\delta_{hed-eud}$). SE is a robust standard error.

Type	Data	Estimate	SE	p
Hedonia	FRED13	0.026	0.098	0.79
	FRED15	0.062	0.04	0.12
	FRED13+15	0.052	0.039	0.19
Eudaimonia	FRED13	-0.063	0.098	0.52
	FRED15	-0.067	0.046	0.14
	FRED13+15	-0.064	0.039	0.1
$\delta_{hed-eud}$	FRED13	0.089	0.189	0.64
	FRED15	0.129	0.079	0.1
	FRED13+15	0.116	0.073	0.11

O'Brien's, Anderson's R^2_F and Roast tests (Fig. 3C). Type S errors are relatively high for GLS above $m > 10$ (0.06 when $m = 52$) and high (~ 0.1) for the permutation F tests across all m. The Exaggeration Ratio (ER) generally decreased with m in all methods except GLS (Fig. 3D). As a consequence, at $m = 52$, statistically significant effect sizes estimated by GLS were close to $3\times$ the true size. By contrast, when $m = 52$, statistically significant effect sizes estimated by GEE, O'Brien's, Anderson's R^2_F, and Roast were only to $2\times$ the true size, while the ERs for GlobalAncova and F_{pun}-tests are ~ 1.3.

Correlated coefficients

The expected, large negative correlation between the partial regression coefficients for *Hedonia* and *Eudaimonia* are shown using the GLS bootstrap distribution (Fig. 2) and using the GLS Monte Carlo simulation results (Fig. 4). Despite modeling the empirical correlations among the regressors and among the response variables, the distribution of standardized coefficients from the GLS Monte Carlo simulation have a much smaller range than that from the GLS permutation (e.g. 95% confidence interval for $\beta_{eudaimonia}$ from the Monte Carlo simulation is -0.20 to 0.23 while that from the GLS permutation is -0.62 to 0.57), which suggests there is something about the structure of the actual data that is inflating the coefficient estimates (Littell et al., 2006).

DISCUSSION

A causal association between well-being components and CTRA expression levels would be an important discovery. Certainly, some association between well-being scores and CTRA expression levels must exist because of common shared paths within the complex network of causal paths of the underlying physiology. Nevertheless, observational studies like that of Fredrickson et al. (2013, 2015) are poor designs for discovering knowledge (Walker, 2014). The re-analysis of the CTRA gene expression data in subjects scored for hedonic and eudaimonic well-being highlights several important results: 1) any effect of hedonic and eudaimonic well-being on CTRA expression is very small and the noise is too large relative to the signal to reliably estimate the sign and magnitude of these mean effects, 2) the apparent replication of opposing effects is most parsimoniously explained by correlated noise due to the high correlation between *Hedonia* and *Eudaimonia*, 3) the GLS with correlated error test has high error rates and inflated effect estimates for simulated data modeled on the focal dataset, and 4) all of the OLS tests have appropriate error rates and the permutation F-tests have high power.

The association between well-being and CTRA expression

Standardized mean effects of *Hedonia* and *Eudaimonia* are very small (Table 3) but the standardization effectively precludes easy comparison to published effect sizes on expression levels. The multivariate (OLS) regression was re-run on the unstandardized expression levels of FRED13 and FRED15 and the mean coefficients were back-transformed to a fold change per four standard deviation change in the predictor (effectively comparing someone at the high and low ends of the well-being axis), using $FC = 2^{4*\bar{\beta}}$. For *Eudaimonia*, I used the reciprocal of this fold change to make the value greater than one and multiplied it by -1 to indicate a decreasing effect. The FC values were 1.036 and 1.072 for *Hedonia*

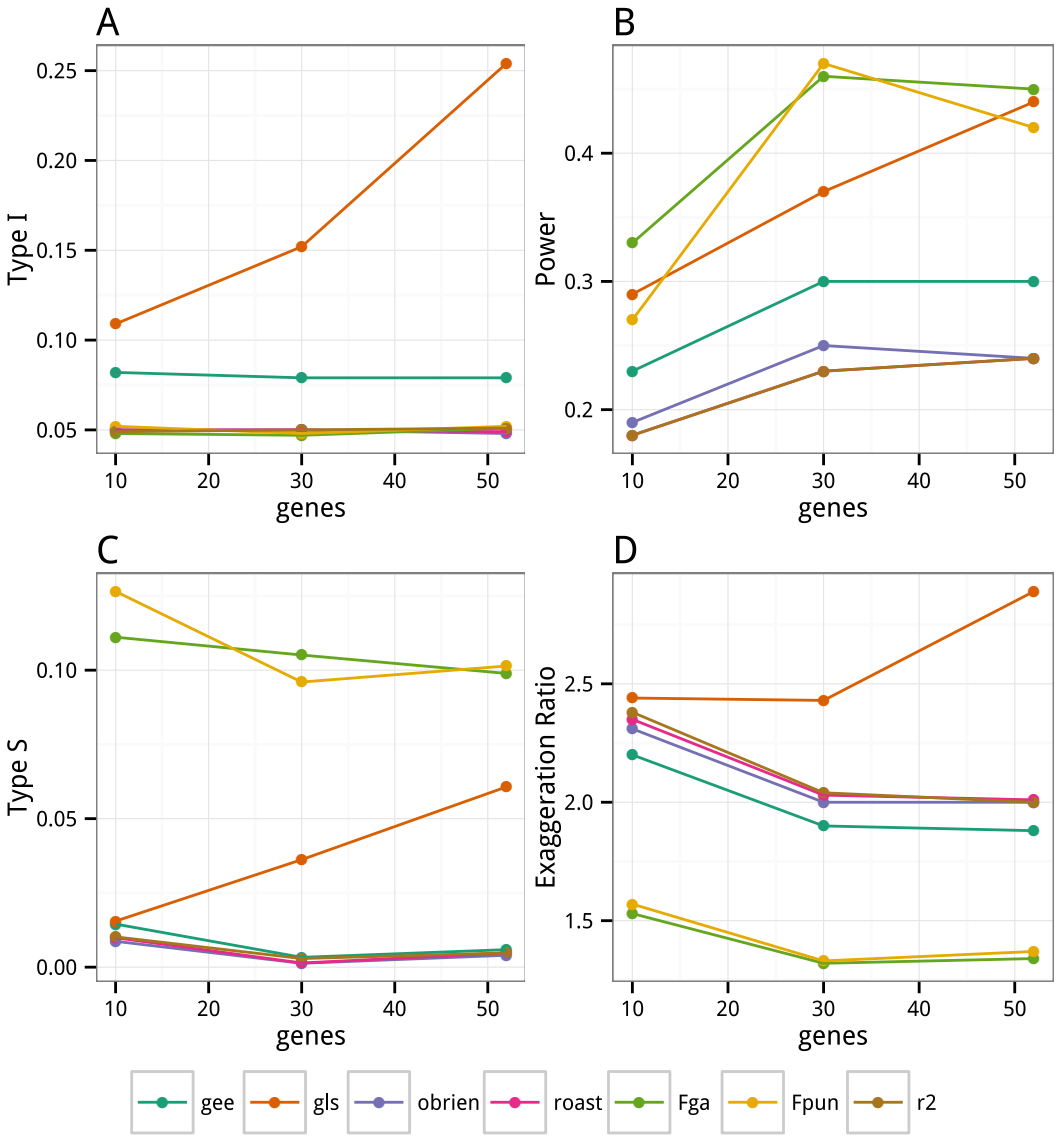

Figure 3. Errors for the different test methods based on Monte Carlo simulation of FRED15 dataset. A) Type I error when the true effect is zero. B) Power when the true effect is 0.067 (the absolute value of the OLS estimated effect of *Eudaimonia* on mean expression for the FRED15 dataset). C) Type S (”sign”) error when the true effect is 0.067. Type S error is the fraction of statistically significant effects in which the estimate has the opposite sign of the true effect. D) Type M (”magnification”) error when the true effect is 0.067, illustrated by the Exaggeration Ratio (ER). ER is the ratio of the estimated to true effect when $p \leq 0.05$. The rates are based on ≥ 6000 values for all methods except GLS, which, because of the time necessary to compute the statistics, are based on 2000 values for 10 and 30 genes and 1000 values for the full 52 genes.

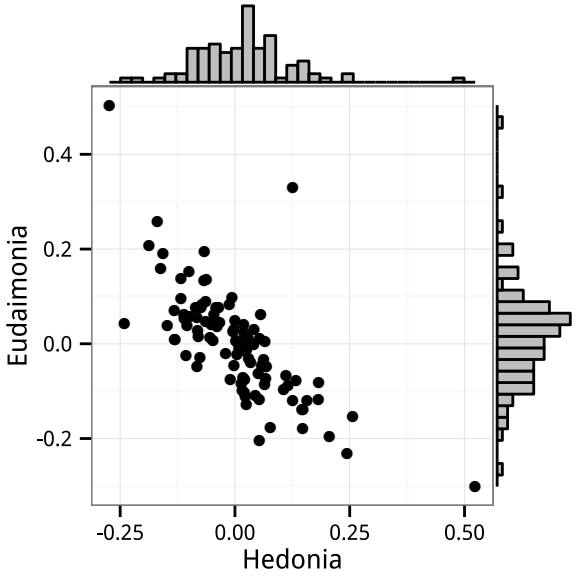

Figure 4. Bivariate distribution of standard partial regression coefficients for *Hedonia* and *Eudaimonia* from the Monte Carlo experiments. The Monte Carlo simulated the FRED15 data but with zero expected effect of any of the regressors on the gene expression levels. The coefficients were estimated by the linear model with correlated error (GLS).

and -1.078 and -1.06 for *Eudaimonia* (note that Fredrickson et al. (2015) reported this fold change as a percent). The biological significance of such a small mean effect awaits experimental evidence.

Several features of the GLS results suggest unstable and inflated coefficient estimates resulting from the GLS model. First, the highly variable pattern of effects among the three datasets (FRED13, FRED15, COLE15) when estimated using the same error structure suggests that either a large lack of generalizability among samples or the coefficients are more unstable than suggested by their (non-robust) standard error. Second, the GLS coefficients are very different from and generally much larger than the OLS coefficients (Tables 1 and 3). Third, at least one of the GLS coefficients in each of the datasets is very large relative to what we'd expect from a gene set association given observational data and the stated hypotheses. Fourth, in a supplementary table, Fredrickson et al. (2015) report strikingly different results (small, negative coefficients for both *Hedonia* and *Eudaimonia*) for the FRED15 dataset using an unstructured error matrix for the GLS computation ($Hedonia: \beta = -0.014, p = 0.17; Eudaimonia: -0.0026, p = 0.81$. Compare these to Table 1). Fredrickson et al. (2015) failed to identify or address any of these concerns, including why the 2013 dataset was not analyzed using the updated (GLS) analysis, or how to interpret the differing results using a compound symmetry error matrix, which was the focus of Fredrickson et al. (2015), or an unstructured error matrix, which was the focus of Cole et al. (2015). The results reported here support the conclusion of inflated coefficient estimates from the GLS. These results include the large coefficients that commonly occurred in the GLS with permuted data despite the expected effects of zero and the diagnostic plot of the residual vs. fitted values that indicates biased estimates (Fig. 1).

The replication in the pattern of effects between datasets

The apparent replication of opposing signs for hedonic and eudaimonic effects on CTRA expression (Fredrickson et al., 2015, 2016) can be inferred only from the OLS estimates; the GLS estimates are strikingly inconsistent with a replicated pattern of expression. This failure of the GLS estimates to replicate was not noted by Fredrickson et al. (2015) because they used the OLS estimates to illustrate the replication but GLS to infer effects. Regardless, any replication in the sign of the mean effect should not be surprising given only two replicates of two coefficients.

The pattern of opposing signs for hedonic and eudaimonic effects on CTRA expression is consistent with very small effects in combination with the high empirical correlation between hedonic and eudaimonic scores (0.80 in FRED13 and 0.74 in FRED15). Partial regression coefficients of regressors that are positively correlated are themselves negatively correlated because their estimation shares common com-

ponents that are of opposite sign. This is easily shown using the data from FRED15 where, disregarding all predictors but hedonic and eudaimonic scores, the partial regression coefficient of any gene expression level on $Hedonia$ (X_1) and $Eudaimonoia$ (X_2) are

$$\begin{aligned}\beta_1 &= 0.018\mathbf{x}_1^\top\mathbf{y} - 0.013\mathbf{x}_2^\top\mathbf{y}\\ \beta_2 &= -0.013\mathbf{x}_1^\top\mathbf{y} + 0.018\mathbf{x}_2^\top\mathbf{y}\end{aligned} \tag{6}$$

where the 0.018 and -0.013 are the diagonal and off-diagonal elements of the inverse of the $\mathbf{X}^\top\mathbf{X}$ matrix of FRED15 (again disregarding all other predictors to simplify the explanation). Because of the high correlation between hedonic and eudaimonic scores, both β_1 and β_2 include a large contribution from the covariance of the other X with Y but the sign of this contribution is negative. Consequently, if the true effects are trivially small, then the pair of β coefficients will tend to have opposite signs because of the negative correlation of estimates centered near zero. Random noise creates negatively correlated coefficients that tend to be opposite in sign. Linear mixed models do not adjust for this correlation. The negative correlation between coefficients is easily seen in the distribution of bootstrap GLS estimates of $\beta_{hedonia}$ and $\beta_{eudaimonia}$ (Fig. 2). The tendency for the coefficients to have opposite signs if the expected effects are zero is seen in the Monte Carlo simulation of the FRED15 data (Fig. 4). While I've shown the negative correlation using GLS estimates, this correlation would also appear in OLS estimates. The most parsimonious explanation of the apparent replication of opposing effects of hedonic and eudaimonic scores on CTRA gene expression is correlated noise arising from the geometry of multiple regression.

Comparison of method performance

The Monte-Carlo simulations of the GLS with correlated error for repeated measures or multiple outcome data are consistent with other studies demonstrating inflated Type I error due to downward biased standard errors (Guerin and Stroup, 2000; Littell et al., 2006; Jacqmin-Gadda et al., 2007; Gurka et al., 2011). By contrast, all of the OLS methods maintain error rates close to the expected value (0.05). The permutation F-tests (F_{pun} and GlobalAncova) have much higher power than the O'Brien's OLS, Anderson's r_F^2, and Roast tests and, unlike the moderately high power for GLS, this power does not trade-off with type I error.

In designs with low power because of small effect sizes, type S and M errors are more likely to emerge (Gelman and Carlin, 2014). That is, with low power, only unusually large estimates are large enough to reach statistical significance. And with a true effect size near zero, an estimate with unusually large error from the true value has a high probability of being the wrong sign. In the simulation here, the true effect is small but the tests with the highest power are associated with the highest rate of type S error. Sign error is a cost of a higher powered test. This type S error affects the permutation F-tests, which have $10\times$ the type S error as the other OLS tests. Indeed, type S error, even with a very small effect, is trivial in the O'Brien's, Anderson's, and Roast tests. The exaggeration ratio (ER), a measure of type M error (Gelman and Carlin, 2014), is a good indicator of the expected inflation of an estimate when the design or test has low power. The expected inflation is nearly $3\times$ the true value for the GLS estimate under the conditions of the FRED15 dataset. By contrast, the expected inflation is less than $1.4\times$ for the F_{pun} and GlobalAncova tests. The high powered tests result in the (perhaps paradoxical) negative relationship between type M and type S error among the tests.

Conclusions

The OLS estimates combined with the permutation F-tests provide some evidence of a very small negative association between $Eudaimonia$ and mean CTRA expression, although the Monte Carlo results of these F tests raise some concern about the sign of this effect. As I've stated above, however, there must be some association between well-being components and CTRA expression, so an observational design with a statistically significant p-value should not cause much excitement. What we want to know are the causal pathways that explain this association — does decreased CTRA cause eudaimonic well-being, or does eudaimonic well-being cause decreased CTRA, or is the correlation jointly determined by an unknown causal pathway? And we want to know if the effect magnitude has meaningful physiological consequences.

The linear model with correlated errors (GLS) has few merits for the estimation of mean fixed effects across multiple responses. The estimation is time consuming and estimation with an unstructured error matrix is plagued with difficulties in convergence. Simulations of the model (here and elsewhere)

with repeated measures or multiple outcomes show a high frequency of inflated coefficient estimates and downward biased standard errors. As expected, the Generalized Estimating Equations with robust standard errors perform much better than the GLS, but even this estimator has inflated type I error. While all the OLS methods maintain type I error at the nominal rate, the tests using the F-ratio (F_{pun} and GlobalAncova) have a relatively high power and small exaggeration ratio. A concern of the GlobalAncova test for observational data is the violoation of the exchangeability assumption. How the F_{pun}-test and GlobalAncova perform with simulated data with moderate to large correlations between predictors and nuisance covariates remains to be investigated.

ACKNOWLEDGMENTS

I am grateful to three reviews and editor for greatly improving this manuscript.

REFERENCES

Anderson, M. J. (2001). Permutation tests for univariate or multivariate analysis of variance and regression. *Canadian journal of fisheries and aquatic sciences*, 58(3):626–639.

Anderson, M. J. and Robinson, J. (2001). Permutation tests for linear models. *Australian & New Zealand Journal of Statistics*, 43(1):75–88.

Brown, N. J. L., MacDonald, D. A., Samanta, M. P., Friedman, H. L., and Coyne, J. C. (2014). A critical reanalysis of the relationship between genomics and well-being. *Proceedings of the National Academy of Sciences of the United States of America*, 111(35):12705–12709.

Brown, N. J. L., MacDonald, D. A., Samanta, M. P., Friedman, H. L., and Coyne, J. C. (2016). More questions than answers: continued critical reanalysis of fredrickson et al.'s studies of genomics and well-being. *PLOS ONE*, 11(6):e0156415.

Bull, S. B. (1998). Regression models for multiple outcomes in large epidemiologic studies. *Statistics in Medicine*, 17(19):2179–2197.

Chen, J. J., Lee, T., Delongchamp, R. R., Chen, T., and Tsai, C.-A. (2007). Significance analysis of groups of genes in expression profiling studies. *Bioinformatics*, 23(16):2104–2112.

Cole, S. W., Levine, M. E., Arevalo, J. M. G., Ma, J., Weir, D. R., and Crimmins, E. M. (2015). Loneliness, eudaimonia, and the human conserved transcriptional response to adversity. *Psychoneuroendocrinology*, 62:11–17.

Dallow, N. S., Leonov, S. L., and Roger, J. H. (2008). Practical usage of O'Brien's OLS and GLS statistics in clinical trials. *Pharmaceutical statistics*, 7(1):53–68.

Fredrickson, B. L., Grewen, K. M., Algoe, S. B., Firestine, A. M., Arevalo, J. M. G., Ma, J., and Cole, S. W. (2015). Psychological well-being and the human conserved transcriptional response to adversity. *PLoS ONE*, 10(3):e0121839.

Fredrickson, B. L., Grewen, K. M., Algoe, S. B., Firestine, A. M., Arevalo, J. M. G., Ma, J., and Cole, S. W. (2016). Correction: psychological well-being and the human conserved transcriptional response to adversity. *PLOS ONE*, 11(6):e0157116.

Fredrickson, B. L., Grewen, K. M., Coffey, K. A., Algoe, S. B., Firestine, A. M., Arevalo, J. M. G., Ma, J., and Cole, S. W. (2013). A functional genomic perspective on human well-being. *Proceedings of the National Academy of Sciences of the United States of America*, 110(33):13684–13689.

Freedman, D. and Lane, D. (1983). A nonstochastic interpretation of reported significance levels. *Journal of Business & Economic Statistics*, 1(4):292–298.

Gelman, A. and Carlin, J. (2014). Beyond power calculations cssessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, 9(6):641–651.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., and Hothorn, T. (2015). *mvtnorm: Multivariate Normal and t Distributions*. R package version 1.0-3.

Goeman, J. J. and Bühlmann, P. (2007). Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980–987.

Guerin, L. and Stroup, W. (2000). A simulation study to evaluate PROC MIXED analysis of repeated measures data. *Annual Conference on Applied Statistics in Agriculture*.

Gurka, M. J., Edwards, L. J., and Muller, K. E. (2011). Avoiding bias in mixed model inference for fixed effects. *Statistics in Medicine*, 30(22):2696–2707.

503 Hummel, M., Meister, R., and Mansmann, U. (2008). GlobalANCOVA: exploration and assessment of
504 gene group effects. *Bioinformatics*, 24(1):78–85.
505 Jacqmin-Gadda, H., Sibillot, S., Proust, C., Molina, J.-M., and Thiébaut, R. (2007). Robustness of
506 the linear mixed model to misspecified error distribution. *Computational Statistics & Data Analysis*,
507 51(10):5142–5154.
508 Kackar, R. N. and Harville, D. A. (1984). Approximations for standard errors of estimators of fixed and
509 random effect in mixed linear models. *Journal of the American Statistical Association*, 79(388):853–
510 862.
511 Kenward, M. G. and Roger, J. H. (1997). Small sample inference for fixed effects from restricted
512 maximum likelihood. *Biometrics*, 53(3):983.
513 Lauter, J. (1996). Exact t and F tests for analyzing studies with multiple endpoints. *Biometrics*, 52(3):964–
514 970.
515 Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models.
516 *Biometrika*, 73(1):13–22.
517 Littell, R. C., Stroup, W. W., Milliken, G. A., Wolfinger, R. D., and Schabenberger, O. (2006). *SAS for*
518 *mixed models*. SAS institute.
519 Logan, B. R. and Tamhane, A. C. (2004). On O'Brien's OLS and GLS tests for multiple endpoints.
520 *Lecture Notes-Monograph Series*, 47:76–88.
521 Mansmann, U., Meister, R., Hummel, M., Scheufele, R., and with contributions from S. Knueppel (2010).
522 *GlobalAncova: Calculates a global test for differential gene expression between groups*. R package
523 version 3.36.0.
524 O'Brien, P. C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics*, 40(4):1079–
525 1087.
526 Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and R Core Team (2015). *nlme: Linear and Nonlinear*
527 *Mixed Effects Models*. R package version 3.1-122.
528 Pocock, S. J., Geller, N. L., and Tsiatis, A. A. (1987). The analysis of multiple endpoints in clinical trials.
529 *Biometrics*, 43(3):487–498.
530 R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for
531 Statistical Computing, Vienna, Austria.
532 Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015). limma powers
533 differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*,
534 43:doi: 10.1093/nar/gkv007.
535 Sammel, M. D. and Ryan, L. M. (2002). Effects of covariance misspecification in a latent variable model
536 for multiple outcomes. *Statistica Sinica*, pages 1207–1222.
537 Tian, L., Greenberg, S. A., Kong, S. W., Altschuler, J., Kohane, I. S., and Park, P. J. (2005). Discovering
538 statistically significant pathways in expression profiling studies. *Proceedings of the National Academy*
539 *of Sciences of the United States of America*, 102(38):13544–13549.
540 Tripathi, S. and Emmert-Streib, F. (2012). Assessment method for a power analysis to identify differen-
541 tially expressed pathways. *PLOS ONE*, 7(5):e37510.
542 Tsai, C.-A. and Chen, J. J. (2009). Multivariate analysis of variance test for gene set analysis. *Bioinfor-*
543 *matics*, 25(7):897–903.
544 Walker, J. A. (2014). The effect of unmeasured confounders on the ability to estimate a true performance
545 or selection gradient (and other partial regression coefficients). *Evolution*, 68(7):2128–2136.
546 Wu, D., Lim, E., Vaillant, F., Asselin-Labat, M.-L., Visvader, J. E., and Smyth, G. K. (2010). ROAST:
547 rotation gene set tests for complex microarray experiments. *Bioinformatics*, 26(17):2176–2182.
548 Yan, J. (2002). geepack: Yet another package for generalized estimating equations. *R-News*, 2/3:12–14.
549 Zeger, S. L. and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes.
550 *Biometrics*, 42(1):121–130.
551 Zhou, Y.-H., Barry, W. T., and Wright, F. A. (2013). Empirical pathway analysis, without permutation.
552 *Biostatistics*, 14(3):573–585.