International Journal of Advanced Statistics and Probability, 2 (2) (2014) 124-134
©Science Publishing Corporation
www.sciencepubco.com/index.php/IJASP
doi: 10.14419/ijasp.v2i2.3698
Research Paper

Certain effects of uncertain models

Brian Knaeble

University of Wisconsin–Stout, USA
Email: knaebleb@uwstout.edu

Abstract

Statistical summaries of multiple regression analyses often state conclusions as if model uncertainty is of little concern. The error due to a mis-specified model, however, can be more significant in practice than the sampling error associated with commonly reported statistics. The true effect of an explanatory variable may be opposite that indicated by a fitted coefficient of a linear model, even if the model is well fit and the coefficient is deemed statistically significant. Here we study the sensitivity of the sign of a fitted coefficient to changes in the model structure. As a consequence of the principle of least squares, we show that a set of covariates with a relatively weak coefficient of determination can not reverse the sign of a relatively strong fitted coefficient of a linear model, given some orthogonality conditions. A consequence of the theory is a necessary condition for Simpson's paradox.

Keywords: *confounding, least squares, model uncertainty, regression, sensitivity analysis.*

1. Introduction

We start with a simple example that is meant to demonstrate the central problem of this paper. The particular example has been chosen for illustrative purposes, and it has been built from readily accessible data. The example highlights a danger of exploratory analysis specifically, and strengthens existing awareness of the difficulties associated with interpretation of observational data generally. It is a striking example of how different models, each well fit to the same data, can lead to statistically significant yet opposite conclusions. The example clearly demonstrates the need for more general theory as requested by Chatfield [1].

The "Swiss" data set within R contains 47 observations on 6 variables: county-level measurements of fertility, agriculture, education, Catholicism, infant mortality, and examination scores. A model of fertility in terms of agriculture alone indicates a positive, statistically significant effect of agriculture on fertility. A model of fertility in terms of agriculture, education, and Catholicism, indicates a negative, statistically significant effect of agriculture on fertility. The details are provided in Table 1.1 and Table 1.2.

The problem is that the two models are incongruous with regards to the effect of agriculture on fertility. In this instance, any conclusion drawn from the data depends strongly on the choice of model. It would thus be misleading to summarize just one of the models, as this could foster a false sense of certainty. Yet scientific articles often

Table 1.1: Fertility as a linear function of agriculture.

variable	$\hat{\beta}_i$	$\mathrm{SE}(\hat{\beta}_i)$	$t = \hat{\beta}_i/\mathrm{SE}(\hat{\beta}_i)$	two-sided p value
agriculture	0.194	0.076	2.532	.015

Table 1.2: Fertility as a linear function of agriculture and covariates.

variable	$\hat{\beta}_i$	$\mathrm{SE}(\hat{\beta}_i)$	$t = \hat{\beta}_i/\mathrm{SE}(\hat{\beta}_i)$	two-sided p value
agriculture	-0.203	0.071	-2.854	0.007
education	-1.072	0.156	-6.881	1.91×10^{-8}
Catholicism	0.145	0.030	4.817	1.84×10^{-5}

summarize results obtained through a single model, without any sensitivity analysis. For further context see some recent examples [2, 3, 4, 5, 6].

By paying attention to issues of model selection, professional communities can guard against publication bias (see [7]) and reduce the prevalence of contradictory claims within scientific literature. For example, Tarino et al have written a paper summarizing the results of a meta study analyzing the effect of saturated fat consumption on cardiovascular health [8], and Scarborough et al have responded critically arguing that some of the studies under consideration in the meta analysis adjusted for covariates inappropriately [9]. At the heart of this controversy is a disagreement regarding model structure. There are also more general concerns about proper analysis of observational data.

Observational studies do continue to play a significant role in health care [10], and observational approaches are re-emerging within ecology [11]. Meanwhile, economists continue to make use of observational data [12], as do social scientists [13]. Interestingly, even the ATLAS particle detector observed thirteen petabytes of data in 2010 [14]. Regarding observational study in general, Rosenbaum states that it is the unmeasured covariates that present the largest difficulties [15]. This article aims to introduce mathematical theory that is meant to address some of these difficulties.

2. Background

Suppose that a sufficient number of high-dimensional observations have been made, to fit a linear model, and that the set of explanatory variables to be used in the model has yet to be determined. An estimate for the qualitative nature of the unique effect of X_i on Y is desired, but the dimension is large enough so that it is not computationally feasible to fit every possible model. Thus, any conclusion reached regarding the effect of X_i on Y must be regarded with some degree of suspicion.

Specifically, suppose that subject matter knowledge has been used to select a linear model, with explanatory variables indexed by I. Denote with ${}_I\hat{\beta}_i$ the ith fitted coefficient within this model, obtained using the principle of least squares. As long as the vector of residuals is nonzero, then it remains possible, through consideration of data associated with additional explanatory variables, indexed by J, with left subscripts indicating model structure, that $\mathrm{sign}({}_{J,I}\hat{\beta}_i) \neq \mathrm{sign}({}_I\hat{\beta}_i)$. We call this a reversal.

Relevant to our study of reversals is general theory regarding the least-squares fitting of linear models. It is possible to mathematically derive a formula that expresses how a single covariate influences an existing model. When an additional column of data is appended to the regression matrix the vector of fitted coefficients changes by

$$(X^tX)^{-1}X^t\mathbf{x}_j\frac{\mathbf{x}_j^t(I - X(X^tX)^{-1}X^t)\mathbf{y}}{\mathbf{x}_j^t(I - X(X^tX)^{-1}X^t)\mathbf{x}_j}, \tag{2.1}$$

where $\mathbf{x}$ is the additional column of data, $\mathbf{y}$ is the vector of response data, and X is the original regression matrix [16].

Hosman et al have shown, for a single coefficient of interest, how the expression in (2.1) decomposes into a ratio of standard errors, a fitted coefficient when X_j is regressed onto the original explanatory variables, and the partial correlation between X_j and Y given the original explanatory variables [17]. It remains unclear, however, how to use such theory pragmatically as part of a model selection procedure.

According to Myers there are some situations where certain covariates should not be adjusted for [18], although Rubin tends to disagree [19]. In a medical context, Kurth states that it is often insufficient to adjust only for a few demographic variables [20]. On the other hand, Robins et al point out that adjusting for too many covariates can be problematic [21]. Pearl suggests that practicioners should use graphs to determine an admissible set of covariates for adjustment [22]. It is apparent that there is a lack of consensus on how best to proceed.

In the presence of many unobserved and potentially confounding variables, it is inherently difficult to interpret results. Chatfield has urged statisticians to "stop pretending that model uncertainty does not exist and begin to find ways of coping with it [1]." Our strategy here is based on the intuitive nature of correlation. The theoretical approach

is applicable whenever a coefficient of determination can be estimated, even for unmeasured sets of covariates, where estimates are based on subject matter knowledge. The mathematical theory then leads to strengthened defense of conclusions drawn from a specific model, even in the presence of substantial model uncertainty.

3. Model-independent estimation

We prepare for a theorem that facilitates model-independent estimation for the direction of an effect. Let r denote Pearson's correlation coefficient, and let R denote the positive square root of the coefficient of determination. Let i be an indice within an index I, where I refers to centered, orthogonal columns of data for a subset of explanatory variables. Let I/i representing the index I without i. Let J be a disjoint index that refers to additional columns of data, not necessarily centered nor orthogonal (to themselves). Suppose that centered versions of columns referenced by J are each orthogonal to each column referenced by I/i. Then the following implication holds true.

Theorem 3.1.

$${}_JR < |r(\mathbf{x}_i, \mathbf{y})| \implies \text{sign}({}_{J,I}\hat{\beta}_i) = \text{sign}({}_I\hat{\beta}_i).$$

Theorem 3.1 is reminiscent of a line of reasoning (see [23]) that was used to implicate smoking as a cause of lung cancer in American men [24]. Fisher had earlier argued essentially that 'correlation is not causation', and he maintained that the observed association between smoking and lung cancer could be due to a third factor [25]. Cornfield et al then responded with "the magnitude of the excess lung-cancer risk among cigarette smokers is so great that the results can not be interpreted as arising from an indirect association of cigarette smoking with some other agent or characteristic, since this hypothetical agent would have to be at least as strongly associated with lung cancer as cigarette use; no such agent has been found or suggested [26]."

Theorem 3.1 is formulated to be applicable in much the same way that the argument of Cornfield et al. has been used. Theorem 3.1 regards the sensitivity of a fitted coefficient to expansion of a linear model, assuming the principle of least-squares. The theorem can provide researchers with an argument to use against any claims that their model failed to account for a set of covariates—the covariates can not reverse the observed direction of a unique effect unless they as a whole possess a relatively large coefficient of determination for the response variable. Thus, in conjunction with subject matter knowledge, Theorem 3.1 can apply even to sets of unmeasured covariates. This and the theorem's general formulation within the context of linear modeling distinguish it from other similar results (see [24], [27], [28] or [29]).

A few precautionary remarks are needed. First, note that for Theorem 3.1 to apply, the regression matrix must have orthogonal columns. The counter example in Table 3.1 shows that any weakening of this assumption leaves open the possibility for a covariate associated with neither the response variable nor the explanatory variable of interest, nonetheless, to induce a reversal. In this sense, models fit over principal components, in addition to producing estimators with less variance (see [30]), can lead to more robust interpretations and conclusions. Second, it is not enough to consider potentially confounding covariates individually. As shown in Table 3.2, a set of covariates each correlating arbitrarily weakly with the response data, can together as a whole have an arbitrarily large coefficient of determination, and together they are thus capable of inducing reversals.

With d covariates under consideration, a linear model can be linearly expanded in 2^d possible ways: one expansion for each subset of covariates. If we allow for non linear expansions, say by using higher order combinations of covariates, then the number of possible expansions is greater yet. It may not be computationally feasible to fit all of these models. However, we can compute R^2 for the largest conceivable set of covariates, and if this value is small enough, then we can conclude that this largest extension can not induce a reversal *and* none of the many, smaller, sub extensions can produce a reversal either. These conclusions follow from Theorem 3.1 and the observation that deleting explanatory columns of data from the analysis can not increase R^2. This latter claim can be rigorously justified using Definition 4.1, Lemma 4.2, Proposition 4.3 and Proposition 4.5 of this paper.

The theory of the preceding paragraph is illustrated in Table 3.3. The statistics on display were computed from data associated with an ecological study of mortality, biochemistry, diet and lifestyle that was carried out in rural China in the 1980s and early 1990s (see [31]). For each of sixty four counties, heart disease rates were obtained along with county-level consumption values for each of ten dietary variables. These particular variables were selected for their familiarity and (among the dietary variables) their disparity. The resulting data made for an interesting exploratory analysis.

Wheat is the dietary variable most strongly correlated with heart disease, and the set of remaining dietary variables as a whole possesses a relatively weak coefficient of determination. With an awareness of Theorem 3.1 and a familiarity with R^2 we can thus quickly conclude that none of the $2^9 = 512$ possible linear regression models that

utilize wheat as an explanatory variable can contain an estimate for the unique effect of wheat on heart disease that is negative. We quickly summarize this theoretical conclusion by stating that the data most likely do not indicate a protective effect of wheat consumption on county-level heart disease rates.

Table 3.1: A contrived dataset where $\mathbf{x}_3$ is uncorrelated with both $\mathbf{x}_1$ and $\mathbf{y}$, yet $\text{sign}({}_{1,2,3}\hat{\beta}_1) \neq \text{sign}({}_{1,2}\hat{\beta}_1)$.

$\mathbf{y}$	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
$\sqrt{2}$	$\sqrt{2}$	1	$5+\sqrt{2}$
$-\sqrt{2}$	$-\sqrt{2}$	1	$5-\sqrt{2}$
0	$2\sqrt{2}$	$-5\sqrt{2}-1$	$\sqrt{2}-5$
0	$-2\sqrt{2}$	$5\sqrt{2}-1$	$-\sqrt{2}-5$

Table 3.2: A contrived data set illustrating how the reversal potential of $\mathbf{x}_2$ and $\mathbf{x}_3$ combined can be greater than expected: ${}_1\hat{\beta}_1 = 0.5 = {}_1R$, as $\epsilon \downarrow 0$ both ${}_2R \downarrow 0$ and ${}_3R \downarrow 0$, while ${}_{2,3}R \equiv 0.75 > 0.5$, and ${}_{1,2,3}\hat{\beta}_1 = -1.0$. Incidentally, ${}_1\hat{\beta}_1 = {}_{1,2,3}\hat{\beta}_1 = 0.5$ when $\epsilon = 0$.

$\mathbf{y}$	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
$\sqrt{2}+\sqrt{3}$	$2\sqrt{2}$	$\sqrt{2}\sqrt{3}\epsilon+\epsilon$	$\sqrt{2}\sqrt{3}\epsilon+\epsilon$
$-\sqrt{2}+\sqrt{3}$	$-2\sqrt{2}$	$-\sqrt{2}\sqrt{3}\epsilon+\epsilon$	$-\sqrt{2}\sqrt{3}\epsilon+\epsilon$
$-\sqrt{3}$	$2\sqrt{2}$	$2\sqrt{2}-\epsilon$	$-2\sqrt{2}-\epsilon$
$-\sqrt{3}$	$-2\sqrt{2}$	$-2\sqrt{2}-\epsilon$	$2\sqrt{2}-\epsilon$

Table 3.3: Correlations between dietary variables and county level heart disease rates; with wheat excluded $R^2 = 0.30$.

dietary variable	observed correlation with *Heart Disease*
Cholesterol	-.15
Saturated Fat	-.18
Fish	-.21
Nuts	.01
Salt	.00
Spices	.33
Wheat	.64
Beans	-.33
Fruits	-.03
Vegetables	-.13

4. Simpson's paradox

The reversal of a fitted coefficient's sign brings to mind Simpson's paradox. Wagner has described Simpson's paradox as "the designation for a surprising situation that may occur when two populations are compared with respect to the incidence of some attribute: if the populations are separated in parallel into a set of descriptive categories, the population with higher overall incidence may yet exhibit a lower incidence within each such category [32]." A mathematical definition is provided in Table 4.1. See Good and Mittal's article [33] for an overview of related concepts and terminology in the literature.

Julious has described Simpson's paradox in a medical setting [34], and Bickel et al have spoken of a related phenomenon when analyzing admissions data from the University of California at Berkeley [35]. They have written the following: "Examination of aggregate data on graduate admissions to the University of California, Berkeley, for fall 1973 shows a clear but misleading pattern of bias against female applicants. . . . If the data is properly pooled, taking into account the autonomy of departmental decision making, thus correcting for the tendency of women to

Table 4.1: Simpson's paradox occurs when $\frac{\sum a_j}{\sum b_j} > \frac{\sum c_j}{\sum d_j}$, yet $\forall j \ \frac{a_j}{b_j} < \frac{c_j}{d_j}$.

	category 1	category 2	$\cdots$	category s
population 1	a_1/b_1	a_2/b_2	$\cdots$	a_s/b_s
population 2	c_1/d_1	c_2/d_2	$\cdots$	c_s/d_s

apply to graduate departments that are more difficult for applicants of either sex to enter, there is a small but statistically significant bias in favor of women."

The details show that not *every* department had a higher acceptance rate for females. Nonetheless, the authors have chosen to describe the reversal as "a paradox, sometimes referred to as Simpson's", likely because after adjusting for a confounding variable, namely 'department', an opposite interpretation of the data becomes possible. To recognize their use of the term, and other similar usage (see [36]), an alternative, weaker definition of Simpson's paradox should be considered. The terminology of linear modeling can apply.

Let Y indicate the presence or absence of an attribute, taking the values one or zero. Let X_i indicate membership within one population or another, taking the values zero or one. Let the s indicator variables $X_{j_1}, X_{j_2}, ..., X_{j_s}$ together indicate category. With $I = \{i\}$ and $J = \{j_1, j_2, ..., j_s\}$, Simpson's paradox, in its weaker sense, can be said to occur when $\text{sign}({}_{J,I}\hat{\beta}_i) \neq \text{sign}({}_{I}\hat{\beta}_i)$. We say that Simpson's paradox, in its stronger sense, occurs when Wagner's previously stated definition is satisfied.

Lemma 4.1. *Occurrence of the strong Simpson's paradox implies occurrence of the weak Simpson's paradox.*

Proof. Let $\mathbf{y}$, $\mathbf{x}_1$, and $\{\mathbf{x}_j\}_{j=2,3,...k}$ be vectors each taking only the values zero and one, with the latter set associated with a single categorical variable. For each j in $\{2, 3, ..., k\}$, let $\hat{\beta}_1(j)$ represent the first, least-squares fitted coefficient when the model is fit over only those observations with $x_j = 1$. Let $\hat{\beta}_1(1)$ represent the first, least-squares fitted coefficient when the model is fit over only those observations where for every j, $x_j = 0$. It suffices to show that if for all $j = 1, 2, 3, ..., k$, $\hat{\beta}_1(j) > 0$, then with $J = \{2, 3, ..., k\}$ we have ${}_{J,1}\hat{\beta}_1 > 0$.

Start by considering a length-n vector of real-valued observations, namely $\mathbf{x}$. Consider the quantity $\sum_{i=1}^{n}(x_i - z)^2$, as a function of z, and note that it is concave up. Its derivative with respect to z is $-2\sum_{i=1}^{n}(x_i - z)$, which is equal to zero precisely when $z = \sum_{i=1}^{n} x_i/n = \bar{x}$. We state these observations for future reference. We also purposefully shift the entries of $\mathbf{x}_1$ so that instead of being 0 or 1 they are $-.5$ or $.5$. This does not effect ${}_{J,1}\hat{\beta}_1$.

There are k categories and within each there are two values for X_1. We thus divide the sample of observations of Y into $2k$ sub samples, and compute each mean. Our assumption that for $j = 1, 2, 3, ..., k$, $\hat{\beta}_1(j) > 0$ ensures that paired means within a given category are different. We can thus set each of $\{{}_{J,1}\beta_0, {}_{J,1}\beta_2, {}_{J,1}\beta_3, ..., {}_{J,1}\beta_k\}$ within the closed interval bounded by the differing means of its associated category. Note that the least-squares estimates must come from such a subset of the parameter space. Observe, given our setup, that for any $\alpha > 0$, due to the observations of the opening paragraph, that ${}_{J,1}\beta_1 = \alpha$ results in a lower sum of the squares of the residuals than ${}_{J,1}\beta_1 = -\alpha$. We thus rule out the possibility of a negative value for ${}_{J,1}\hat{\beta}_1$. □

Lemma 4.1 combined with the contrapositive statement of Theorem 3.1 (after squaring the inequality) thus results in the following necessary condition for (the strong or weak) Simpson's paradox. The coefficients of determination refer to the coefficients of determination for the incidence of the attribute of interest.

Corollary 4.1. *For Simpson's paradox to occur it is necessary for the set of indicator variables associated with categorization to possess a coefficient of determination that is larger than the coefficient of determination possessed by the variable indicating population.*

5. Mathematical theory

This section develops some mathematics that can be used to prove Theorem 3.1. There is a geometric flavor to the definitions that is best embraced before moving on to the lemmas and propositions. Attention should be drawn to Proposition 5.1 in particular as it may prove useful during future in depth study of the least-squares fitting procedure. A solid understanding of this proposition leads to a thorough understanding of the proof of the theorem.

5.1. Notation

The existence of a general data set as depicted in Table 5.1 is assumed. There are n, m-dimensional observations. Let I index a subset of $\{1, 2, ..., m\}$, J index a disjoint subset, and K index a generic subset. Let i stand for a generic element of I, j stand for a generic element of J, and k stand for a generic element of K.

Bold symbols indicate observed vectors of data within $\mathbb{R}^n$. Also, $\langle \cdot, \cdot \rangle$ is used for the standard inner product, $|\cdot|$ for the associated, Euclidean norm, and $\perp$ to indicate orthogonality.

With $\mathbf{e}$ denoting a vector of n ones, the vectors $\{\mathbf{e}, \mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_m}\}$ are assumed to be a linearly independent set. The span of $\mathbf{e}$, and a subset of vectors indexed by K, is a vector subspace denoted with ${}_KV$. For every K, both $\mathbf{y} \notin {}_KV$ and $\mathbf{y} \not\perp {}_KV$ are assumed.

In general, V stands for a vector subspace. Also, left subscripts indicate a subset of explanatory variables, and a post subscript typically indicates a variable of interest.

Table 5.1: A sufficiently general data set that illustrates the notation.

$\mathbf{y}$	$\mathbf{x}_1$	$\mathbf{x}_2$	...	$\mathbf{x}_m$
y_1	$x_{1,1}$	$x_{2,1}$	$\cdots$	$x_{m,1}$
y_2	$x_{1,2}$	$x_{2,2}$	...	$x_{m,2}$
y_3	$x_{1,3}$	$x_{2,3}$	...	$x_{m,3}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
y_n	$x_{1,n}$	$x_{2,n}$	...	$x_{m,n}$

5.2. Definitions

In this subsection $K = \{k_1, k_2, ..., k_p\}$.

Definition 5.1. Denote the projection of $\mathbf{y}$ onto V with

$$p_V(\mathbf{y}) = \underset{\mathbf{v} \in V}{\operatorname{argmin}}(|\mathbf{y} - \mathbf{v}|).$$

Definition 5.2. The vector of fitted coefficients, $({}_K\hat{\beta}_0,\ {}_K\hat{\beta}_{k_1},\ {}_K\hat{\beta}_{k_2}, ...,\ {}_K\hat{\beta}_{k_p})$, is the unique solution of

$$p_{\,{}_KV}(\mathbf{y}) = {}_K\hat{\beta}_0\mathbf{e} + {}_K\hat{\beta}_{k_1}\mathbf{x}_{k_1} + {}_K\hat{\beta}_{k_2}\mathbf{x}_{k_2} + ... + {}_K\hat{\beta}_{k_p}\mathbf{x}_{k_p}.$$

Definition 5.3. ${}_Ky$ is the function

$${}_Ky : \mathbb{R}^p \to \mathbb{R}$$

$${}_Ky : (\alpha_{k_1}, \alpha_{k_2}, ..., \alpha_{k_p}) \mapsto {}_K\hat{\beta}_0 + {}_K\hat{\beta}_{k_1}\alpha_{k_1} + {}_K\hat{\beta}_{k_2}\alpha_{k_2} + ... + {}_K\hat{\beta}_{k_p}\alpha_{k_p}.$$

Definition 5.4. The qth fitted value is

$${}_K\hat{y}_q = {}_Ky(x_{k_1,q}, x_{k_2,q}, ..., x_{k_p,q}).$$

Definition 5.5. The vector of fitted values is

$$\mathbf{{}_K\hat{y}} = ({}_K\hat{y}_1,\ {}_K\hat{y}_2, ...,\ {}_K\hat{y}_n).$$

Remark 5.1. Within $\mathbb{R}^n$, $\mathbf{{}_K\hat{y}} = p_{\,{}_KV}(\mathbf{y})$.

Definition 5.6. Define ${}_KR$ as the positive square root of the coefficient of determination:

$${}_KR = +\sqrt{{}_KR^2} = +\sqrt{\frac{\sum_{q=1}^{n}({}_K\hat{y}_q - \bar{\mathbf{y}})^2}{\sum_{q=1}^{n}(y_q - \bar{\mathbf{y}})^2}}.$$

Definition 5.7. For generic vectors $\mathbf{x} = (x_1, x_2, ..., x_n)$ and $\mathbf{y} = (y_1, y_2, ..., y_n)$, and with s denoting the sample standard deviation, define the Pearson correlation coefficient r as

$$r(\mathbf{x}, \mathbf{y}) = \frac{1}{n-1}\sum_{q=1}^{n}\left(\frac{x_q - \bar{\mathbf{x}}}{s_{\mathbf{x}}}\right)\left(\frac{y_q - \bar{\mathbf{y}}}{s_{\mathbf{y}}}\right).$$

5.3. Geometry

The following lemmas are stated without proof, as they can be surmised to be true or derived from the material in books on mathematical analysis (e.g. Cheney's text, [37]). See the appendix of this article for a proof of Proposition 5.1.

Lemma 5.1. *For any* $\mathbf{y}$ *and for any* V

$(\mathbf{y} - p_V(\mathbf{y})) \perp V$.

Lemma 5.2. *For any* $\mathbf{y}$ *and for any* V

$|p_V(\mathbf{y})|^2 + |\mathbf{y} - p_V(\mathbf{y})|^2 = |\mathbf{y}|^2$.

Lemma 5.3. *For any vectors* $\mathbf{x}, \mathbf{y}$

$\mathbf{x} \perp \mathbf{y} \implies |\mathbf{x}|^2 + |\mathbf{y}|^2 = |\mathbf{x} + \mathbf{y}|^2$.

Lemma 5.4. *For* $V_1 \perp V_2$ *and* $V = \text{span}\{V_1, V_2\}$

$p_V(\mathbf{y}) = p_{V_1}(\mathbf{y}) + p_{V_2}(\mathbf{y})$.

Definition 5.8. For nonzero vectors $\mathbf{y} \in \mathbb{R}^n$ and $\mathbf{v} \in V$, define $\theta(\mathbf{y}, \mathbf{v})$, with $0 \leq \theta \leq \pi$, via

$$\cos(\theta) = \frac{\langle \mathbf{y}, \mathbf{v} \rangle}{|\mathbf{y}||\mathbf{v}|}.$$

Proposition 5.1. *Let* V *be a vector subspace of* $\mathbb{R}^n$. *For a fixed vector* $\mathbf{y} \notin V$, *with* $\mathbf{y} \not\perp V$, *and for a fixed, nonzero vector* $\mathbf{w} \in V$:

(i) If $\mathbf{w}$ *is a scalar multiple of* $p_V(\mathbf{y})$, *then* $\theta(\mathbf{y}, p_V(\mathbf{y}) + t\mathbf{w})$ *is non decreasing on* $\{t : t > 0, p_V(\mathbf{y}) + t\mathbf{w} \neq 0\}$.

(ii) If $\mathbf{w}$ *is not a scalar multple of* $p_V(\mathbf{y})$, *then* $\theta(\mathbf{y}, p_V(\mathbf{y}) + t\mathbf{w})$ *is a strictly increasing function of* $t > 0$.

5.4. Simplifications

Proofs of the propositions in this section are left to the reader.

Definition 5.9. A vector of data $\mathbf{x}$ is *centered* if $\bar{\mathbf{x}} = 0$.

Definition 5.10. A vector of data $\mathbf{x}$ is *geometrically standardized* if $\bar{\mathbf{x}} = 0$ and $|\mathbf{x}| = 1$.

Definition 5.11. Given a vector of data $\mathbf{x}$ we use the term *standardization* to describe the process

$$\mathbf{x} \mapsto \frac{\mathbf{x} - \bar{\mathbf{x}}\mathbf{e}}{|\mathbf{x} - \bar{\mathbf{x}}\mathbf{e}|}.$$

Remark 5.2. Standardization results in geometrically standardized data.

Proposition 5.2. *Standardization preserves the orthogonality of a set of centered vectors.*

Proposition 5.3. *For any* K, *standardization preserves the signs of* $\{ {}_K\hat{\beta}_k\}_{k \in K}$ *and the value of* ${}_KR$.

Proposition 5.4. *For any* K, *if the data is geometrically standardized, then* ${}_K\hat{\beta}_0 = 0$.

Proposition 5.5. *For any* K, *if the data is geometrically standardized, then* ${}_KR = \cos(\theta(\mathbf{y}, p_{{}_KV}(\mathbf{y})) = |p_{{}_KV}(\mathbf{y})|$.

Proposition 5.6. *For* $k = 1, 2, ..., m$, ${}_kR = |r(\mathbf{x}_k, \mathbf{y})|$.

5.5. Proof of Theorem 3.1

By Proposition 5.3, geometrically standardized data can be assumed, and by Proposition 5.2, orthogonality of the vectors indexed by I is retained.

Proposition 5.6 allows us to state the contrapositive of the implication from Theorem 3.1 as

$$\text{sign}({}_{J,I}\hat{\beta}_i) \neq \text{sign}({}_{I}\hat{\beta}_i) \implies {}_{J}R > {}_{i}R.$$

By assumption each vector indexed by I/i is orthogonal to both the vector indexed by i and every (now centered) vector indexed by J. Therefore it suffices to demonstrate

$$\text{sign}({}_{J,i}\hat{\beta}_i) \neq \text{sign}({}_{i}\hat{\beta}_i) \implies {}_{J}R > {}_{i}R.$$

The hypothesis, $\text{sign}({}_{J,i}\hat{\beta}_i) \neq \text{sign}({}_{i}\hat{\beta}_i)$, implies that within ${}_{J,i}V$

${}_{J}V$ separates $p_{{}_{J,i}V}(\mathbf{y})$ from $p_{{}_{i}V}(\mathbf{y})$.

Thus the straight line from $p_{{}_{J,i}V}(\mathbf{y})$ to $p_{{}_{i}V}(\mathbf{y})$ intersects ${}_{J}V$ at a point $\mathbf{q}$.

Consider the two-stage path: from $p_{{}_{J}V}(\mathbf{y})$ to $\mathbf{q}$ within ${}_{J}V$, and then from $\mathbf{q}$ to $p_{{}_{i}V}(\mathbf{y})$ within ${}_{J,i}V$, along two straight line segments. Using Proposition 5.1 we can conclude that

$$\theta(\mathbf{y}, p_{{}_{J}V}(\mathbf{y})) \leq \theta(\mathbf{y}, \mathbf{q}) < \theta(\mathbf{y}, p_{{}_{i}V}(\mathbf{y})). \tag{5.1}$$

This conclusion is valid for the following reasons. We have assumed in Section 5.1 that for any K, $\mathbf{y} \not\perp {}_{K}V$, which implies, even for geometrically standardized data, and again for any K, that ${}_{K}\hat{\beta}_i \neq 0$. Also, if $p_{{}_{i}V}(\mathbf{y}) - p_{{}_{J,i}V}(\mathbf{y})$ is a scalar multiple of $p_{{}_{J,i}}(\mathbf{y})$, then $\mathbf{q} = \mathbf{0}$ and Proposition 5.4 ensures that $p_{{}_{J,i}}(\mathbf{y})$ is a scalar multiple of $\mathbf{x_i}$. This contradicts either ${}_{J,i}\hat{\beta}_i \neq 0$ or ${}_{J,i}\hat{\beta}_j \neq 0$ for $j \in J$. Thus we conclude that $p_{{}_{i}V}(\mathbf{y}) - p_{{}_{J,i}V}(\mathbf{y})$ is not a scalar multiple of $p_{{}_{J,i}}(\mathbf{y})$, and we are justified in using part (ii) of Proposition 5.1 along the first segment. Finally, note that Proposition 5.1 applies along the second segment because the segment lies along a ray emanating from $p_{{}_{J,i}V}(\mathbf{y})$.

To finish this proof we apply the cosine function to (5.1), reversing the ordering, resulting in

$$\cos(\theta(\mathbf{y}, p_{{}_{J}V}(\mathbf{y}))) \geq \cos(\theta(\mathbf{y}, \mathbf{q})) > \cos(\theta(\mathbf{y}, p_{{}_{i}V}(\mathbf{y}))).$$

Proposition 5.5 then allows us to substitute ${}_{J}R$ for $\cos(\theta(\mathbf{y}, p_{{}_{J}V}(\mathbf{y})))$ and ${}_{i}R$ for $\cos(\theta(\mathbf{y}, p_{{}_{i}V}(\mathbf{y})))$, resulting in

${}_{J}R > {}_{i}R,$

which is the desired conclusion from line (**??**). □

A. Appendix: Proof of proposition 5.1

For part (i), with $\alpha \neq 0$, it suffices to show that

$$\cos(\theta) = \frac{\langle \mathbf{y}, p_V(\mathbf{y}) + t(\alpha p_V(\mathbf{y}))\rangle}{|\mathbf{y}||p_V(\mathbf{y}) + t(\alpha p_V(\mathbf{y}))|}$$

is non increasing on $\{t : t > 0, t \neq -1/\alpha\}$. For $\alpha > 0$, or for $\alpha < 0$ and $t < -1/\alpha$,

$$\frac{\langle \mathbf{y}, p_V(\mathbf{y}) + t(\alpha p_V(\mathbf{y}))\rangle}{|\mathbf{y}||p_V(\mathbf{y}) + t(\alpha p_V(\mathbf{y}))|} = \frac{\langle \mathbf{y}, (1+t\alpha)p_V(\mathbf{y})\rangle}{|\mathbf{y}||(1+t\alpha)p_V(\mathbf{y})|} = \frac{(1+t\alpha)}{(1+t\alpha)}\frac{\langle \mathbf{y}, p_V(\mathbf{y})\rangle}{|\mathbf{y}||p_V(\mathbf{y})|} = \frac{\langle \mathbf{y}, p_V(\mathbf{y})\rangle}{|\mathbf{y}||p_V(\mathbf{y})|},$$

which is constant. For $\alpha < 0$ and $t > -1/\alpha$ then

$$\frac{\langle \mathbf{y}, p_V(\mathbf{y}) + t(\alpha p_V(\mathbf{y}))\rangle}{|\mathbf{y}||p_V(\mathbf{y}) + t(\alpha p_V(\mathbf{y}))|} = \frac{\langle \mathbf{y}, (1+t\alpha)p_V(\mathbf{y})\rangle}{|\mathbf{y}||(1+t\alpha)p_V(\mathbf{y})|} = \frac{(1+t\alpha)}{-(1+t\alpha)}\frac{\langle \mathbf{y}, p_V(\mathbf{y})\rangle}{|\mathbf{y}||p_V(\mathbf{y})|} = -\frac{\langle \mathbf{y}, p_V(\mathbf{y})\rangle}{|\mathbf{y}||p_V(\mathbf{y})|},$$

which is also constant. Furthermore,

$$-\frac{\langle \mathbf{y}, p_V(\mathbf{y})\rangle}{|\mathbf{y}||p_V(\mathbf{y})|} \leq \frac{\langle \mathbf{y}, p_V(\mathbf{y})\rangle}{|\mathbf{y}||p_V(\mathbf{y})|}$$

because Lemma 5.2 states

$$|p_V(\mathbf{y})|^2 + |\mathbf{y} - p_V(\mathbf{y})|^2 = |\mathbf{y}|^2,$$

which expands to give

$$\langle p_V(\mathbf{y}), p_V(\mathbf{y})\rangle + \langle \mathbf{y}, \mathbf{y}\rangle - 2\langle \mathbf{y}, p_V(\mathbf{y})\rangle + \langle p_V(\mathbf{y}), p_V(\mathbf{y})\rangle = \langle \mathbf{y}, \mathbf{y}\rangle,$$

which implies

$$\langle \mathbf{y}, p_V(\mathbf{y})\rangle \geq 0.$$

For part (ii), with $\alpha \in \mathbb{R}$, write $\mathbf{w} = \alpha p_V(\mathbf{y}) + \mathbf{u}$, where $\mathbf{u} \perp p_V(\mathbf{y})$. $\cos(\theta)$ thus becomes

$$\frac{\langle \mathbf{y}, p_V(\mathbf{y}) + t(\alpha p_V(\mathbf{y}) + \mathbf{u})\rangle}{|\mathbf{y}||p_V(\mathbf{y}) + t(\alpha p_V(\mathbf{y}) + \mathbf{u})|} = \frac{\langle \mathbf{y}, (1+t\alpha)p_V(\mathbf{y}) + t\mathbf{u}\rangle}{|\mathbf{y}||(1+t\alpha)p_V(\mathbf{y}) + t\mathbf{u}|} = \frac{\langle \mathbf{y}, (1+t\alpha)p_V(\mathbf{y})\rangle + \langle \mathbf{y}, t\mathbf{u}\rangle}{|\mathbf{y}||(1+t\alpha)p_V(\mathbf{y}) + t\mathbf{u}|}.$$

The $\langle \mathbf{y}, t\mathbf{u}\rangle$ term can be dropped since

$$\langle \mathbf{y}, t\mathbf{u}\rangle = t\langle \mathbf{y}, \mathbf{u}\rangle = t\langle p_V(\mathbf{y}) + (\mathbf{y} - p_V(\mathbf{y})), \mathbf{u}\rangle = t\langle p_V(\mathbf{y}), \mathbf{u}\rangle + \langle (\mathbf{y} - p_V(\mathbf{y})), \mathbf{u}\rangle = 0 + 0,$$

where the final zero is due to Lemma 5.1. Thus, it suffices to show that

$$L = \frac{\langle \mathbf{y}, (1+t\alpha)p_V(\mathbf{y})\rangle}{|\mathbf{y}||(1+t\alpha)p_V(\mathbf{y}) + t\mathbf{u}|}$$

is decreasing for $t > 0$.

First we state and prove a Lemma.

Lemma A1. *For $(1 + t\alpha) \neq 0$, $t/(1 + t\alpha)$ is a strictly increasing function of t.*

Proof. $\frac{d}{dt}\frac{t}{1+t\alpha} = \frac{1(1+t\alpha)-\alpha t}{(1+t\alpha)^2} = \frac{1}{(1+t\alpha)^2} > 0.$

Now for t such that $(1 + t\alpha) > 0$,

$$L = \frac{\langle \mathbf{y}, (1+t\alpha)p_V(\mathbf{y})\rangle}{|\mathbf{y}||(1+t\alpha)p_V(\mathbf{y}) + t\mathbf{u}|} = \frac{1/(1+t\alpha)}{1/(1+t\alpha)}\frac{\langle \mathbf{y}, (1+t\alpha)p_V(\mathbf{y})\rangle}{|\mathbf{y}||(1+t\alpha)p_V(\mathbf{y}) + t\mathbf{u}|} = \frac{\langle \mathbf{y}, p_V(\mathbf{y})\rangle}{|\mathbf{y}||p_V(\mathbf{y}) + t\mathbf{u}/(1+t\alpha)|}.$$

Note that $t/(1 + t\alpha)$ is positive because $t > 0$ and $(1 + t\alpha) > 0$, and note also that $t/(1 + t\alpha)$ is increasing by Lemma A1. Thus, as a consequence of Lemma 5.3, $|p_V(\mathbf{y}) + t\mathbf{u}/(1 + t\alpha)|$ is increasing in t, which implies that L is decreasing in t as desired.

For t such that $(1 + t\alpha) < 0$,

$$L = \frac{\langle \mathbf{y}, (1+t\alpha)p_V(\mathbf{y})\rangle}{|\mathbf{y}||(1+t\alpha)p_V(\mathbf{y}) + t\mathbf{u}|} = \frac{1/(1+t\alpha)}{1/(1+t\alpha)}\frac{\langle \mathbf{y}, (1+t\alpha)p_V(\mathbf{y})\rangle}{|\mathbf{y}||(1+t\alpha)p_V(\mathbf{y}) + t\mathbf{u}|} = \frac{\langle \mathbf{y}, p_V(\mathbf{y})\rangle}{-|\mathbf{y}||p_V(\mathbf{y}) + t\mathbf{u}/(1+t\alpha)|}.$$

Note that $t/(1+t\alpha)$ is negative because $t > 0$ and $(1+t\alpha) < 0$, and note also that $t/(1+t\alpha)$ is increasing by Lemma A1. Thus, as a consequence of Lemma 5.3, $|p_V(\mathbf{y})+t\mathbf{u}/(1+t\alpha)|$ is decreasing in t, so that $-|\mathbf{y}||p_V(\mathbf{y})+t\mathbf{u}/(1+t\alpha)|$ is increasing in t, which implies that L is decreasing in t, again as desired.

For t such that $(1+t\alpha) = 0$, note that $\alpha < 0$ so that $0 < t < -1/\alpha \iff (1+t\alpha) > 0$, $t = -1/\alpha \iff (1+t\alpha) = 0$, and $t > -1/\alpha \iff (1+t\alpha) < 0$. Note also that since $\mathbf{y} \notin V$ and $\mathbf{y} \not\perp V$, Lemma 5.2 implies not only $\langle \mathbf{y}, p_V(\mathbf{y})\rangle \geq 0$ as derived previously, but also the strict inequality $\langle \mathbf{y}, p_V(\mathbf{y})\rangle > 0$. Thus for $\{(t_1, t_2, t_3) : 0 < t_1 < t_2 = -1/\alpha < t_3 < \infty\}$,

$$\frac{\langle \mathbf{y}, (1+t_1\alpha)p_V(\mathbf{y})\rangle}{|\mathbf{y}||(1+t_1\alpha)p_V(\mathbf{y}) + t_1\mathbf{u}|} = \frac{\langle \mathbf{y}, p_V(\mathbf{y})\rangle}{|\mathbf{y}||p_V(\mathbf{y}) + t_1\mathbf{u}/(1+t_1\alpha)|} > 0, \quad \frac{\langle \mathbf{y}, (1+t_2\alpha)p_V(\mathbf{y})\rangle}{|\mathbf{y}||(1+t_2\alpha)p_V(\mathbf{y}) + t_2\mathbf{u}|} = 0,$$

and

$$\frac{\langle \mathbf{y}, (1+t_3\alpha)p_V(\mathbf{y})\rangle}{|\mathbf{y}||(1+t_3\alpha)p_V(\mathbf{y}) + t_3\mathbf{u}|} = \frac{\langle \mathbf{y}, p_V(\mathbf{y})\rangle}{-|\mathbf{y}||p_V(\mathbf{y}) + t_3\mathbf{u}/(1+t_3\alpha)|} < 0.$$

This shows that

$$L = \frac{\langle \mathbf{y}, (1+t\alpha)p_V(\mathbf{y})\rangle}{|\mathbf{y}||(1+t\alpha)p_V(\mathbf{y}) + t\mathbf{u}|}$$

must be decreasing at any positive t satisfying $(1 + t\alpha) = 0$. □

References

[1] C. Chatfield, Model uncertainty, data mining and statistical inference, Journal of the Royal Statistical Society: Series A, 158, part 3, (1995), pp. 419466.

[2] Davis et al, Rice consumption and urinary arsenic concentrations in U.S. children, Environmental Health Perspectives, vol.120, issue 10, (2012), p1418-1424.

[3] Jungert et al, Serum 25-hydroxyvitamin D_3 and body composition in an elderly cohort from Germany: a cross-sectional study, Nutrition & Metabolism, 9,42, (2012), Accessed in 2013 from http://www.nutritionandmetabolism.com/content/9/1/42.

[4] Nelson et al, Daily physical activity predicts degree of insulin resistance: a cross-sectional observational study using the 2003–2004 National Health and Nutrition Examination Survey, International Journal of Behavioral Nutrition and Physical Activity, 10, 10, (2013), Accessed in 2013 from http://www.ijbnpa.org/content/10/1/10.

[5] Lignell et al, Prenatal exposure to polychlorinated biphenyls and polybriminated diphenyl ethers may influence birth weight among infants in a Swedish cohort with background exposure: a cross-sectional study, Environmental Health, 12, 44, (2013), Accessed in 2013 from http://www.ehjournal.net/content/12/1/44.

[6] Cervellati et al, Bone mass density selectively correlates with serum markers of oxidative damage in post-menopausal women, Clinical Chemistry and Laboratory Medicine, volume 51, issue 2, (2012), pages 333-338.

[7] K. Dickersin, The existence of publication bias and risk factors for its occurence, The Journal of the American Medical Association, (1990), 1385-1389.

[8] Tarino et al, Meta-analysis of prospective cohort studies evaluating the association of saturated fat with cardiovascular disease, The American Journal of Clinical Nutrition, 91, 3, (2010), 535-546.

[9] Scarborough et al, Meta-analysis of effect of saturated fat intake on cardiovascular disease: overadjustment obscures true associations, The American Journal of Clinical Nutrition, vol. 92, no. 2, (2010), 458-459.

[10] C.Y. Lu, Observational studies: a review of study designs, challenges and strategies to reduce confounding, The International Journal of Clinical Practice, Blackwell Publishing Ltd., 63, 5, (2009), 691-697.

[11] R. Sagarin, A. Pauchard, Observational approaches in ecology open new ground in a changing world, Frontiers in Ecology and the Environment, 8, (2010), 379-386.

[12] J. Wooldridge, Introductory Econometrics, A Modern Approach, South-Western Cengage Learning, USA, (2013).

[13] S.L. Morgan, C Winship, Counterfacutals and Causal Inference: Mathods and Principles for Social Research, Cambridge University Press, New York USA, (2007).

[14] G. Brumfiel, High-energy physics: down the petabyte highway, Nature, 469, (2011), 282-283.

[15] P.R. Rosenbaum, Observational study, Encyclopedia of Statistics in Behavioral Science, volume 3, (2005), pp. 1451-1462.

[16] G. Seber, A. Lee, Linear Regression Analysis, John Wiley & Sons, Hoboken USA, (2003), Equation (3.32).

[17] C.A. Hosman, B.B. Hansen, P.W. Holland, The sensitivity of linear regression coefficients' confidence limits to the omission of a confounder, The Annals of Applied Statistics, vol. 4, no. 2, (2010), 849-870, Proposition 2.1.

[18] Myers et al, Effects of adjusting for instrumental variables on bias and precision of effect estimates, American Journal of Epidemiology, 174, 11, (2011), 1213-1222.

[19] D. Rubin, Author's reply: Should observational studies be designed to allow lack of balance in covariate distributions across treatment groups?, Statistics in Medicine, 28, 9, (2009), 1420-123.

[20] D. Kurth, J. Sonis, Assessment and control of confounding in trauma research, Journal of Traumatic Stress, vol. 20, no. 5, (2007), pp. 807820.

[21] J.M. Robins, S. Greenland, The role of model selection in causal inference from nonexperimental data, American Journal of Epidemiology, vol. 123, no. 3, (1986).

[22] J. Pearl, Causal inference in statistics: an overview, Statistical Surveys, (2009), 96-146.

[23] Cornfield et al, Smoking and lung cancer: recent evidence and a discussion of some questions, Journal of the National Cancer Institute, 22, (1959), 173-203, Appendix A.

[24] D.Y Lin, B.M. Psaty, R.A. Kronmal, Assessing the sensitivity of regression results to unmeasured confounders in observational studies, Biometrics, 54, (1998), 948-963.

[25] R.A. Fisher, Cigarettes, cancer and statistics, Centennial Rev Arts and Sciences, Michigan State University, 2, 151, (1958).

[26] Cornfield et al, Smoking and lung cancer: recent evidence and a discussion of some questions, Journal of the National Cancer Institute, 22, (1959), 173-203.

[27] D. Giles, Coefficient sign changes when restricting regression models under instrumental variables estimation, Oxford Bulletin of Economics and Statistics, 51, (1989), 465-467.

[28] McAleer et al, A further result on the sign of restricted least-squares estimates, Journal of Econometrics, 32, (1986), 287-290.

[29] P.R. Rosenbaum, D.B. Rubin, Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome, Journal of the Royal Statistical Society, Series B, 11, (1983), 212-218.

[30] G. Seber, A. Lee, Linear Regression Analysis, John Wiley & Sons, Hoboken USA, (2003), Section 3.6.

[31] Chen et al; Geographic study of mortality, biochemistry, diet and lifestyle in rural China; Epidemiological Studies Unit, Oxford; http://www.ctsu.ox.ac.uk/ china/monograph/; Revised (1990); Accessed 2009.

[32] C.H. Wagner, Simpson's paradox in real life, The American Statistician, 36, 1, (1982), 4648.

[33] I.J. Good, Y. Mittal, The amalgamation and geometry of two-by-two contingency tables, The Annals of Statistics, vol. 15, no. 2, (1987), pp. 694-711.

[34] S.A. Julious, M.A. Mullee, Confounding and Simpson's paradox. British Medical Journal, 309, 6967, (1994), 14801481.

[35] P.J. Bickel, E.A. Hammel, J.W. O'Connell, Sex bias in graduate admissions: data from Berkeley, Science, 187, 4175, (1975), 398404.

[36] D.R. Appleton, J.M. French, M. Vanderpump, Ignoring a covariate: an example of Simpson's paradox, The American Statistician, volume 50, issue 4, (1996), 340-341.

[37] W. Cheney, Analysis for Applied Mathematics, Springer, New York USA, (2001).