Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v8n3p346

Regression and Random Confounding

By Knaeble

Published: 20 December 2015

Electronic Journal of Applied Statistical Analysis
Vol. 08, Issue 03, December 2015, 346-359
DOI: 10.1285/i20705948v8n3p346

Regression and Random Confounding

Brian Knaeble*

Westminster College, Department of Mathematics
1840 South 1300 East, Salt Lake City UT 84105, USA

Published: 20 December 2015

An ordinary least squares regression estimate for the slope, regardless of its strength, can have its sign reversed through adjustment for a random confounding vector of data. The assumption of a rotationally invariant distribution, on the space of centered, random, confounding vectors of data, makes calculation of probabilities for these reversals possible. Here, as the sample size increases, these probabilities are shown to decrease exponentially. This analytic result leads to some asymptotic comparison between regular sampling error and the error due to a mis-specified model.

keywords: least-squares, high-dimensional geometry, gamma function, complementary error function, model uncertainty, omitted-variable bias.

1. Introduction

Chatfield (1995, p. 419) has broadly proclaimed that "model uncertainty is a fact of life and likely to be more serious than other sources of uncertainty which have received far more attention from statisticians." More specifically, Hosman, Hansen, and Holland (2010, p. 849) argue that "when regression results are questioned, it is often the nonconfounding assumption that is the focus of doubt." Despite these concerns, and even in the presence of clear potential for omitted-variable bias, summaries of statistical analyses are often presented as if models were certain. For examples of this practice see Jungert et al. (2012), Nelson et al. (2013), Lignell et al. (2013) and Cervellati et al. (2012). The extent to which the conclusions of such studies are reliable remains indeterminate. Here we clarify the matter, slightly, through study of the confounding of a relationship between two variables due to the presence of a third variable.

*Corresponding author: bknaeble@westminstercollege.edu

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

Assume that n, 3-dimensional observations have been made resulting in three vectors of data, $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{w}$, which we assume to be centered. The observed correlation $r(\mathbf{x}, \mathbf{y})$ provides a crude estimate for an association between X and Y, but it may be confounded by the lurking variable W. Utilizing the information contained within $\mathbf{w}$, multiple regression and the principle of least squares leads to an adjusted estimate, $\hat{\beta}_{x|w}$, for the unique effect of X on Y, while controlling for W. Both $\hat{\beta}_x = r(\mathbf{x}, \mathbf{y})s_y/s_x$ and $\hat{\beta}_{x|w}$ estimate the true, unique effect, β_x. Note that the standard deviations s_y and s_x are positive.

While $r(\mathbf{x}, \mathbf{y})$ is scale invariant, $\hat{\beta}_x$ and $\hat{\beta}_{x|w}$ are not. However, it is possible to concern ourselves simply with the direction of the estimates: specifically $\text{sign}(\hat{\beta}_{x|w})$ and $\text{sign}(\hat{\beta}_x)$ $(= \text{sign}(r(\mathbf{x}, \mathbf{y})))$. For $r(\mathbf{x}, \mathbf{y}) \neq 0$ we assume without loss of generality that $r(\mathbf{x}, \mathbf{y}) > 0$, and then ask for the probability of $\hat{\beta}_{x|w} < 0$. In other words, we ask this question: what is the probability that controlling for a random confounding vector leads to an adjusted estimate opposite the direction of the original estimate?

To answer the question we first consider the set of $\mathbf{w}$ that make $\hat{\beta}_{x|w} = 0$. We shall see how this set is the boundary of an ellipsoidal cone of two nappes (Equation (A.4)). This boundary region can be identified with the space of factor vectors that make reproduced correlation between $\mathbf{x}$ and $\mathbf{y}$ equal to the original correlation between $\mathbf{x}$ and $\mathbf{y}$. The identification reveals a connection between least-squares estimates that adjust to zero (Equation (A.2)) and vanishing residual correlations of factor analysis (Equation (2.28) of Harman (1976)). Here $\mathbf{w}$ is not to be thought of as a factor vector but rather as a generic vector associated with an unobserved variable that may or may not be relevant.

2. Results

Working with centered $\mathbf{y} \neq \mathbf{0}$ and standardized $\mathbf{x}$ and $\mathbf{w}$, meaning $\bar{\mathbf{y}} = \bar{\mathbf{x}} = \bar{\mathbf{w}} = 0$ and $|\mathbf{x}| = |\mathbf{w}| = \sqrt{n-1}$, we fix $\mathbf{y}$ and $\mathbf{x}$ so that $\langle \mathbf{y}, \mathbf{x} \rangle > 0$, and we consider a random $\mathbf{w}$ on the sphere $S^{n-2}_{\sqrt{n-1}}$, distributed uniformly. By this we mean that it is distributed according to the unique, rotationally invariant, probability measure on that sphere. For a proof of existence and uniqueness of such a measure see Khoshnevisan (2010, Theorem 7.23). Note that such a measure would result from spherical projection of vectors, when each vector has entries that are independent observations on a centered, normally distributed W. Note also that the sphere's dimension, $n-2$, is appropriate, since centered vectors lie orthogonal to a vector of ones and thus within an $n-1$ dimensional space. This is the space where $S^{n-2}_{\sqrt{n-1}}$ is situated. Mathematical details are found in the appendix. The probability, P, of $\mathbf{w}$ being positioned so that $\hat{\beta}_{x|w} < 0$, given $0 < r(\mathbf{y}, \mathbf{x}) < 1$, is a function of the sample size n. With $r = r(\mathbf{x}, \mathbf{y})$ and $n \geq 4$ the following theorem can be stated.

Theorem 2.1. $\left(\dfrac{1-r}{2}\right)^{n-2} < \sqrt{\pi} P(n) < \left(\dfrac{1-r}{1+r}\right)^{(n-2)/2}.$

The probability that a random vector causes an estimate to reverse its direction decreases to zero exponentially as the sample size increases. To suggest a rough comparison, suppose the model $Y = \beta_0 + \beta_x X + \epsilon$ is known with certainty to be true, with $\beta_x > 0$. If we assume that the errors are unbiased, uncorrelated, and have the same variance σ^2, then because we are working with standardized, explanatory vectors of data, the least-squares estimate $\hat{\beta}_x$ has variance $\sigma^2/(n-1)$ (Seber and Lee, 2003, Section 3.2). Assuming normality as well allows us to conclude that $\hat{\beta}_x =^d N(\beta_x, \sigma^2/(n-1))$ (Seber and Lee, 2003, Section 3.4). Since for $x > 0$ the Gaussian, complementary error function satisfies

$$\frac{x}{x^2+1}\frac{1}{\sqrt{2\pi}}e^{-x^2/2} < \int_x^\infty \frac{1}{\sqrt{2\pi}}e^{-t^2/2}dt < \frac{1}{x}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}, \tag{2.1}$$

(Gordan, 1941) we can conclude that $P\left(\hat{\beta}_x < 0 \mid \beta_x > 0\right)$ is decreasing exponentially in the sample size as well.

The upper bound of (2.1) implies

$$P\left(\hat{\beta}_x < 0 \mid \beta_x > 0\right) < \frac{1}{(\beta_x\sqrt{n-1}/\sigma)}\frac{1}{\sqrt{2\pi}}e^{-(\beta_x\sqrt{n-1}/\sigma)^2/2},$$

while the lower bound from Theorem 2.1 ensures

$$\frac{1}{\sqrt{\pi}}\left(\frac{1-r(\mathbf{x},\mathbf{y})}{2}\right)^{n-2} < P\left(\hat{\beta}_{x|w} < 0 \mid \hat{\beta}_x > 0\right),$$

where we have written $P\left(\hat{\beta}_{x|w} < 0 \mid \hat{\beta}_x > 0\right)$ in place of $P(n)$. Thus, asymptotically,

$$P\left(\hat{\beta}_x < 0 \mid \beta_x > 0\right) < P\left(\hat{\beta}_{x|w} < 0 \mid \hat{\beta}_x > 0\right),$$

as long as $e^{-\beta_x^2/(2\sigma^2)} < (1 - r(\mathbf{x},\mathbf{y}))/2$. This observation alludes to the dangers of over-adjustment: adjustment for an independent, normally distributed random variable is not necessarily a benign act; the probability that such adjustment changes the sign of an estimate can be larger than the probability for the sign of the original estimate being correct in the first place.

Conversely, suppose for example, as part of a model selection procedure, that investigators notice $\text{sign}(\hat{\beta}_{x|w}) \neq \text{sign}(\hat{\beta}_x)$, in the presence of a modest $r(\mathbf{x},\mathbf{y})$ value of say $1/3$. According to the theorem, the probability for such a reversal must be less than $1/2^{(n-2)/2}$, where n is the sample size. For large n this probability may be small enough to warrant a rejection of the assumption of a rotationally invariant distribution for $\mathbf{w}$. If $\mathbf{w}$ consists of independent observations on a normally distributed W, then we may reject the assumption that W is independent of X and Y. In this way W can be identified as a possibly confounding variable.

3. Simulations

We have seen, when W is independent of both X and Y, that as the sample size increases then the probability of $\text{sign}(\hat{\beta}_{x|w}) \neq \text{sign}(\hat{\beta}_x)$ decreases exponentially. We have also seen, when assuming $Y = \beta_0 + \beta_x X + \epsilon$ and common assumptions of regression, that as the sample size increases then the probability of $\text{sign}(\hat{\beta}_x) \neq \text{sign}(\beta_x)$ decreases, again, exponentially. Moreover, we have shown asymptotically that the probability of $\text{sign}(\hat{\beta}_{x|w}) \neq \text{sign}(\hat{\beta}_x)$ is greater than the probability of $\text{sign}(\hat{\beta}_x) \neq \text{sign}(\beta_x)$ when $e^{-\beta_x^2/(2\sigma^2)} < (1 - r(\mathbf{x}, \mathbf{y}))/2$.

Here we summarize relevant information obtained from simulations. Table 3.1 gives proportions of simulated trials that satisfy $e^{-\beta_x^2/(2\sigma^2)} < (1 - r(\mathbf{x}, \mathbf{y}))/2$, when X is uniformly distributed from zero to b, and β_x/σ is set at specified values. For simplicity we have fixed $\sigma = 1$ and written just β_x in place of β_x/σ. Our estimate for the proportion is denoted with $\hat{p}$, and the mean correlation across like trials is denoted with $\bar{r}$. Table 3.2 shows when the implication of $e^{-\beta_x^2/(2\sigma^2)} < (1 - r(\mathbf{x}, \mathbf{y}))/2$, namely $P(\text{sign}(\hat{\beta}_{x|w}) \neq \text{sign}(\hat{\beta}_x)) > P(\text{sign}(\hat{\beta}_x) \neq \text{sign}(\beta_x))$, becomes practically significant. The entries of Table 3.2 are simulated estimates for $P(n)$.

In general, $e^{-\beta_x^2/(2\sigma^2)} < (1 - r(\mathbf{x}, \mathbf{y}))/2$ occurs when β_x is relatively large compared to σ. For fixed (β_x, σ), larger b results in larger $\bar{r}$ values and smaller $\hat{p}$ values. Even with $b = 10$ we see that large enough β_x/σ leads to $\hat{p}$ near 1, albeit in the presence of larger r values, and we know via Table 3.2 that reversals are rare when r is large. Nonetheless, large $\hat{p}$ values reveal situations where inaccuracy due to spurious adjustment may be more concerning than inaccuracy due to sampling. For example, with $\beta_x = 3$, $\sigma = 2$, $b = 0.5$, and $n = 10$, we have $\hat{p} = 0.65$ and $\bar{r} = .20$, and with $r = .20$ and $n = 10$ the reversal probability is significant: $P(n) > 0.05$.

4. Application

At the end of Section 2 it has been described how the upper inequality from Theorem 2.1 can be used to identify W as a possible confounder. The reasoning described there is applicable and readily demonstrated with an example here. Note that when $r < 0$ we can apply the theorem with $|r|$ in place of r.

Plant scientist Paul Frater has studied the roots of little bluestem (*Schizachyrium scoparium*), a North American prairie grass.[1] He has observed 197 plants throughout the Central United States and taken measurements on variables such as plant root length (root length), soil carbon-nitrogen ratio (C:N), and percentage of root colonized by mycorrhizal fungi (colonization).

The bivariate correlation between C:N and root length is $r = -.080$. However, when root length is linearly modeled as a function of C:N and colonization, then the least-squares fitted coefficient for C:N is positive. The histogram for colonization is shown in Figure 1, along with a Q–Q plot.

Suppose that the observations of colonization are independent and normally distributed. If colonization were independent of both C:N and root length then by either Theorem 2.1 (or an extension of Table 3.2) the observed reversal would occur with probability less than 10^{-7}. This value is small enough to warrant a rejection of the

[1] Data available upon request

Table 3.1: For each set of specified values for the parameters β_x, n, and b, a fixed set of $100,000$ synthetic samples was obtained, so as to determine the proportion, $\hat{p}$, of synthetic samples that satisfy $e^{-\beta_x^2/(2\sigma^2)} < (1 - r(\mathbf{x}, \mathbf{y}))/2$, and also to compute the mean correlation, $\bar{r}$.

		β_x						
		.1	.5	1	2	3	5	10
$(n = 6, b = 1)$	$\hat{p}$	0	.021	.154	.717	.991	1	1
	$\bar{r}$	.025	.132	.260	.476	.630	.807	.940
$(n = 30, b = 1)$	$\hat{p}$	0	0	.004	.975	1	1	1
	$\bar{r}$	.030	.141	.275	.505	.652	.821	.945
$(n = 6, b = 10)$	$\hat{p}$	0	0	0	.001	.043	.999	1
	$\bar{r}$	.258	.807	.940	.984	.993	.997	.999
$(n = 30, b = 10)$	$\hat{p}$	0	0	0	0	0	1	1
	$\bar{r}$	.274	.821	.945	.985	.994	.998	.999

Table 3.2: Given two, positively correlated, length-n vectors $\mathbf{x}$ and $\mathbf{y}$, we produced $100,000$ synthetic, length-n, $\mathbf{w}$ vectors, each with entries consisting of independent observations from the standard, normal distribution, and for each such $\mathbf{w}$, we computed $\hat{\beta}_{x|w}$. Each table entry gives the proportion of the $100,000$ trials that satisfied $\hat{\beta}_{x|w} < 0$, and as such the table entries are simulated estimates for $P(n)$.

n	$r = .1$	$r = .25$	$r = .5$	$r = .9$
6	.267	.132	.041	.001
10	.153	.038	.004	$< .001$
20	.044	.002	$< .001$	$< .001$
50	.002	$< .001$	$< .001$	$< .001$
100	.001	$< .001$	$< .001$	$< .001$

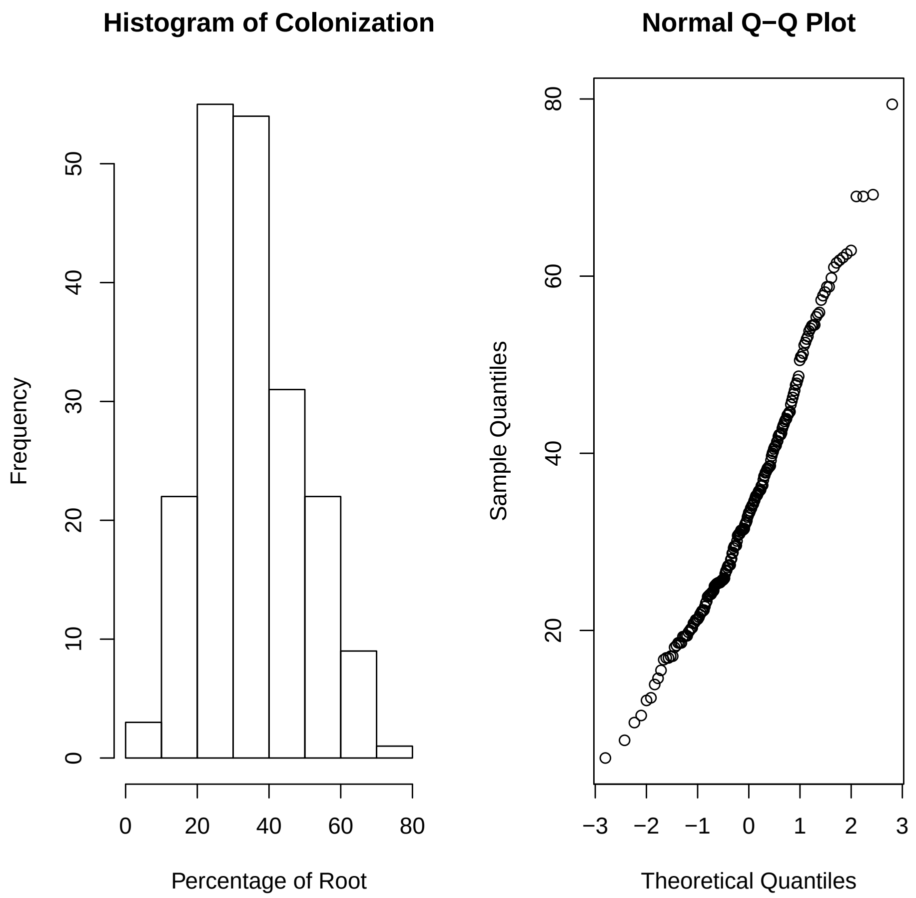

Figure 1: The graphics do not indicate extreme departure from normality

independence assumption. This rejection combined with subject matter knowledge then leads to consideration of colonization as a confounding variable for the effect of C:N on root length.

5. Discussion

Subject matter knowledge is needed to definitively identify a covariate as a confounding variable (Pearl, 2009a, p. 100). As just shown, Theorem 2.1 leads to a test for confounding, and the approach is consistent with the practice of characterizing confounding variables as those variables that correlate with both treatment (or an explanatory variable of interest) and outcome (Frank, 2000, p. 150). Pearl (2009b) gives a causal definition for confounding in his book *Causality*. For further reading relating to confounding see Greenland and Morganstern (2001), Rosenbaum and Rubin (1983), McNamee (2003), and Howards et al. (2012).

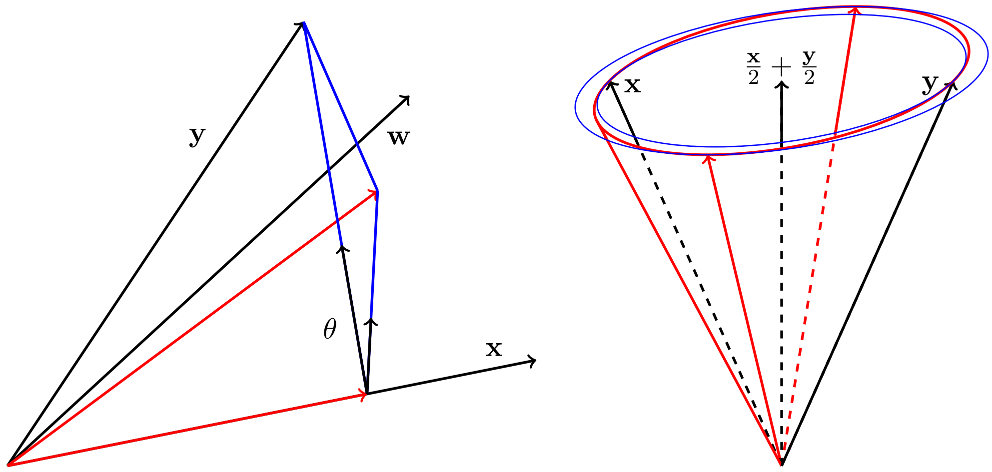

Figure 2: $t \propto \cot\theta, F \propto \cot^2(\theta)$, and $\text{sign}(\hat{\beta}_{x|w}) \neq \text{sign}(\hat{\beta}_x)$ when $\mathbf{w}$ is positioned within the red, elliptical cone.

Here we have taken advantage of symmetry about zero and focused on reversals of directions of effects. There is some related literature on this topic within the field of econometrics. Leamer (1975) has used t and F statistics to show how reversals can occur during model contraction only if the set of variables to be dropped is more significant than the variable of interest. There are similar results (McAleer, Pagan, and Visco, 1986; Oksanen, 1987; Visco, 1988) that are useful to the applied econometrician (Giles, 1989, p. 465). See also Knaeble (2014). All these results, including the results of this paper (which are distinct in their emphasis on probabilities), can be used to improve intuition for the applied practicioner of regression.

Testing for independence of W, as we have done here in Section 4, is not the same as testing for significance of W, as done through a t test within multiple regression analysis. The former test checks the degree to which centered $\mathbf{w}$ points along the sum of standardized $\mathbf{x}$ and standardized $\mathbf{y}$. The latter test, which in our case is equivalent to an F test of significance for a model extension, checks the change in $\hat{\mathbf{y}}$ due to $\mathbf{w}$. See Figure 2 for an illustration. Note how the t or F statistics are determined by the position of $\mathbf{x}$ and the position of the plane spanned by $\{\mathbf{x}, \mathbf{w}\}$, whereas $\hat{\beta}_{x|w}$ depends on the particular position of $\mathbf{w}$ itself.

Hoeffding (1948) has introduced a non-parametric test for the independence of two continuous random variables. More recently, Szekély, Rizzo, and Bakirov (2007, Theorem 7) have introduced a test based on distance covariance. Distance covariance can be used to test the independence of W from (X, Y). Szekély, Rizzo, and Bakirov (2007) have run simulations to compare the power of their test, to the power of the likelihood ratio test that uses Wilks Lambda (Wilks, 1935), and to the power of the related Puri-Sen rank or sign tests (Puri and Sen, 1971, Chapter 8). Margaritis and Thrun (2001) discretize at multiple resolutions and take a Bayesian approach.

Reasoning based on knowledge of the magnitudes for the probabilities from Theorem

2.1 may be used to support informal testing of independence during exploratory analysis, as shown in Section 4. It may be possible to sharpen this reasoning, with tighter inequalities from improved geometric analysis, or through consideration of the exact values of $\hat{\beta}_{x|w}$ and $\hat{\beta}_x$ (not just their directions). Hosman, Hansen, and Holland (2010) and Frank (2000) have formulas that could prove useful in this regard.

W has been assumed to be normally distributed during production of Table 3.2. The variable colonization has been observed to be approximately normally distributed in Section 4. Theorem 2.1 requires $\mathbf{w}$ to have a rotationally invariant distribution. Simulations with binomial ($p = .5$), uniform, and exponential ($\lambda = 1$) distributions for W have produced tables analogous and similar to Table 3.2. These modified tables generally differ only in the thousandths digit. This fact is supportive of the claim that our results are not sensitive to departures from normality. It remains to see precisely how sensitive Theorem 2.1 is to departures from a perfectly rotationally-invariant distribution for $\mathbf{w}$.

To conclude, here we summarize what has been done. Geometric analysis on vectors of centered data has led to mathematical results that quantify the relationship between error probabilities associated with statistical adjustment and probabilities associated with sampling error. Simulations have shown that the error probabilities associated with adjustment can be practically significant and greater than the probabilities associated with sampling error. The analysis has shown, that as the sample size increases, both types of probabilities decrease exponentially. The main analytic result can be used to support an argument that a covariate is a confounding variable.

A. Appendix

A.1. Caps on a Sphere

For the purpose of proving Theorem 2.1 we can assume that all three vectors, $\mathbf{y}$, $\mathbf{x}$ and $\mathbf{w}$, have each been *geometrically* standardized, meaning that they are centered and positioned on a *unit* sphere.

Definition A.1. The $(d-1)$-dimensional sphere of radius a is

$$S_a^{d-1} = \left\{ (u_1, u_2, ..., u_d) \in \mathbb{R}^d : u_1^2 + u_2^2 + ... + u_d^2 = a^2 \right\}.$$

Definition A.2. The d-dimensional ball of radius a is

$$B_a^d = \left\{ (u_1, u_2, ..., u_d) \in \mathbb{R}^d : u_1^2 + u_2^2 + ... + u_d^2 < a^2 \right\}.$$

Remark A.1. We use the same notation for embedded such objects.

Remark A.2. When $a = 1$ we drop the subscript.

Definition A.3. For nonzero vectors $\mathbf{u}_1$ and $\mathbf{u}_2$, define the angle between them, θ, to be

$$\theta(\mathbf{u}_1, \mathbf{u}_2) = \cos^{-1}\left(\frac{\langle \mathbf{u}_1, \mathbf{u}_2 \rangle}{|\mathbf{u}_1||\mathbf{u}_2|}\right).$$

Definition A.4. Given a point, $\mathbf{p}$, on the sphere S^{k-1}, and an angle ϕ, where $0 < \phi < \pi/2$, define the *cap*, $C(p, \phi)$, as

$$C(p, \phi) = \left\{ \mathbf{s} \in S^{k-1} : \theta(\mathbf{s}, \mathbf{p}) < \phi \right\}.$$

Definition A.5. Given a cap, $C(p, \phi)$, of the sphere S^{k-1}, define its negative, $-C(p, \phi)$, as

$$-C(p, \phi) = \left\{ \mathbf{s} \in S^{k-1} : \theta(-\mathbf{s}, \mathbf{p}) < \phi \right\}.$$

Definition A.6. Given a cap, $C(p, \phi)$, of the sphere S^{k-1}, and also the cap's negative, $-C(p, \phi)$, define their union, $K(p, \phi)$, as

$$K(p, \phi) = C(p, \phi) \cup -C(p, \phi).$$

Definition A.7. With n denoting the sample size, set $k = n - 1$ and fix $\mathbf{x}, \mathbf{y} \in S^{k-1}$ so that $0 < r(\mathbf{x}, \mathbf{y}) < 1$. Define the subset of centered, unit length, reversal-inducing vectors as

$$R = \{\mathbf{w} \in S^{k-1} : \hat{\beta}_{x|w} < 0\}.$$

Proposition A.1. *Within the context of Definition A.7,*

$$K\left(\frac{\mathbf{x}+\mathbf{y}}{|\mathbf{x}+\mathbf{y}|}, \frac{\theta(\mathbf{x}, \mathbf{y})}{2}\right) \subset R \subset K\left(\frac{\mathbf{x}+\mathbf{y}}{|\mathbf{x}+\mathbf{y}|}, \cos^{-1}\left(\sqrt{\frac{2r(\mathbf{x}, \mathbf{y})}{r(\mathbf{x}, \mathbf{y})+1}}\right)\right). \tag{A.1}$$

Proof. A formula of Cohen et al. (2003, Equation (3.24)) gives

$$\hat{\beta}_{x|w} = \frac{r(\mathbf{x}, \mathbf{y}) - r(\mathbf{x}, \mathbf{w})r(\mathbf{w}, \mathbf{y})}{1 - r(\mathbf{x}, \mathbf{w})^2}. \tag{A.2}$$

Set $r = r(\mathbf{x}, \mathbf{y})$ and select k orthonormal basis vectors for the ambient space of S^{k-1}, so that $\mathbf{x}$ and $\mathbf{y}$ can be expressed with new coordinates as

$$\mathbf{x} = \left(-\sqrt{\frac{1-r}{2}}, \sqrt{\frac{1+r}{2}}, 0, ..., 0\right) \text{ and } \mathbf{y} = \left(\sqrt{\frac{1-r}{2}}, \sqrt{\frac{1+r}{2}}, 0, ..., 0\right).$$

Temporarily let $\mathbf{w} = (w_1, ..., w_k)$ vary throughout this same ambient space. It is thus a centered vector of arbitrary magnitude, while $\mathbf{x}$ and $\mathbf{y}$ are centered and unit length. Because any correlation coefficient is an inner product of centered vectors divided by their magnitudes, and since $(1 - r(\mathbf{x}, \mathbf{w})^2) > 0$, we have via (A.2) that $\hat{\beta}_{x|w} = 0$ if and only if

$$\begin{aligned}
2r &= 2r(\mathbf{x}, \mathbf{w})r(\mathbf{w}, \mathbf{y}) \\
2r &= 2\left(-w_1\sqrt{\frac{1-r}{2}} + w_2\sqrt{\frac{1+r}{2}}\right)\frac{1}{|\mathbf{w}|}\left(w_1\sqrt{\frac{1-r}{2}} + w_2\sqrt{\frac{1+r}{2}}\right)\frac{1}{|\mathbf{w}|}
\end{aligned}$$

$$\begin{aligned} 2r &= 2\left(w_2^2\frac{(1+r)}{2} - w_1^2\frac{(1-r)}{2}\right)\frac{1}{|\mathbf{w}|^2} \\ 2r|\mathbf{w}|^2 &= \left(w_2^2(1+r) - w_1^2(1-r)\right). \end{aligned} \tag{A.3}$$

With $w_2 = 1$, line (A.3) can be rewritten as

$$\begin{aligned} 2r(w_1^2 + 1 + w_3^2 + ... + w_k^2) &= (1+r) + (r-1)w_1^2 \\ (r+1)w_1^2 + 2r(w_3^2 + ... + w_k^2) &= 1 - r \\ \frac{1+r}{1-r}w_1^2 + \frac{2r}{1-r}(w_3^2 + ... + w_k^2) &= 1. \end{aligned} \tag{A.4}$$

Since scaling of $\mathbf{w}$ does not affect $\hat{\beta}_{x|w}$, the zero set $\{\mathbf{w} : \hat{\beta}_{x|w} = 0\}$ is conical, of two nappes, with ellipsoidal cross-sections. The reversal region $R = \{\mathbf{w} \in S^{k-1} : \hat{\beta}_{x|w} < 0\}$ is thus the intersection of the interior of this ellipsoidal cone with the standard sphere. Elementary trigonometry ensures that the minimum opening angle for the cone is $\theta(\mathbf{x}, \mathbf{y})$, and half the maximum opening angle for the cone is $\cos^{-1}\left((2r(\mathbf{x}, \mathbf{y})/(r(\mathbf{x}, \mathbf{y}) + 1))^{1/2}\right)$. It remains only to observe that any spherical cap is defined with an angle ϕ that is half the opening angle for an associated spherical cone. □

A.2. Exponentially Decreasing Probabilities

In order to estimate the proportion of the sphere taken up by R, we seek first a lower estimate for the volume of the smallest set in (A.1) and then an upper estimate for the volume of the largest set in (A.1). We proceed using techniques known to convex geometers (Ball, 1997). We let "∂" denote the boundary of a set, and we let μ denote the Hausdorff measure of a set (where normalizing constants have been chosen for consistency with Lebesgue measure). See Federer (1969) for measure theoretic details. Note that $\partial B_a^d = S_a^{d-1}$. We write Γ for the gamma function.

Lemma A.1. *Let* $\mathbf{x}, \mathbf{y} \in S^{k-1}$*:* $0 < r(\mathbf{x}, \mathbf{y}) < 1$*. For all* $k \geq 3$*,*

$$\frac{1}{\sqrt{\pi}}\left(\frac{(1 - r(\mathbf{x}, \mathbf{y}))}{2}\right)^{k-1} < \frac{\mu\left(K\left(\frac{\mathbf{x}+\mathbf{y}}{|\mathbf{x}+\mathbf{y}|}, \frac{\theta(\mathbf{x},\mathbf{y})}{2}\right)\right)}{\mu(S^{k-1})}. \tag{A.5}$$

Proof. Without loss of generality assume that $(\mathbf{x} + \mathbf{y})/2 = (t, 0, 0, ..., 0)$ for some $t > 0$. Elementary trigonometry ensures that $t = \sqrt{(r(\mathbf{x}, \mathbf{y}) + 1)/2}$ and $|(\mathbf{x} - \mathbf{y})/2| = \sqrt{\frac{1-r(\mathbf{x},\mathbf{y})}{2}}$. Let s stand for $\sqrt{\frac{1-r(\mathbf{x},\mathbf{y})}{2}}$.

Due to the curvature of S^{k-1},

$$\mu\left(K\left(\frac{\mathbf{x}+\mathbf{y}}{|\mathbf{x}+\mathbf{y}|}, \frac{\theta(\mathbf{x},\mathbf{y})}{2}\right)\right) > 2\mu\left(B_s^{k-1}\right).$$

Also, $\mu(B^d) = \frac{\pi^{d/2}}{\Gamma(1+d/2)}$, $\mu(\partial B^d) = \frac{d\pi^{d/2}}{\Gamma(1+d/2)}$, and $\mu(B_a^d) = a^d \frac{\pi^{d/2}}{\Gamma(1+d/2)}$ (Ball, 1997, Lecture 1). Therefore,

$$\begin{aligned}
\frac{\mu\left(K\left(\frac{\mathbf{x}+\mathbf{y}}{|\mathbf{x}+\mathbf{y}|}, \frac{\theta(\mathbf{x},\mathbf{y})}{2}\right)\right)}{\mu(S^{k-1})} > \frac{2\mu\left(B_s^{k-1}\right)}{\mu(\partial B^k)} &= \frac{\frac{2s^{k-1}\pi^{(k-1)/2}}{\Gamma(1/2+k/2)}}{\frac{k\pi^{k/2}}{\Gamma(1+k/2)}} \\
&= \frac{2s^{k-1}}{\sqrt{\pi}k} \frac{\Gamma(1+k/2)}{\Gamma(1/2+k/2)} \\
&> \frac{2}{\sqrt{\pi}} \frac{s^{k-1}}{k} \frac{3k-5}{2k-2} \\
&= \frac{1}{\sqrt{\pi}} \frac{3k-5}{k(k-1)} \left(\frac{1-r(\mathbf{x},\mathbf{y})}{2}\right)^{(k-1)/2} \\
&\geq \frac{1}{\sqrt{\pi}} \left(\frac{1-r(\mathbf{x},\mathbf{y})}{2}\right)^{(k-1)/2} \left(\frac{1-r(\mathbf{x},\mathbf{y})}{2}\right)^{(k-1)/2} \\
&= \frac{1}{\sqrt{\pi}} \left(\frac{1-r(\mathbf{x},\mathbf{y})}{2}\right)^{k-1}
\end{aligned}$$

The third inequality is due to the hypothesis $k \geq 3$. The middle inequality is due to the following argument. In general, $\Gamma(x) = (x-1)\Gamma(x-1)$, resulting in $\Gamma(x) - \Gamma(x-1) = (x-1)\Gamma(x-1) - \Gamma(x-1) = \Gamma(x-1)((x-1)-1) = \Gamma(x-1)(x-2)$. By convexity, $\Gamma(x+1/2)$ is thus greater than $\Gamma(x) + \Gamma(x-1)(x-2)/2$, and with the substitution of $\Gamma(x)/(x-1)$ for $\Gamma(x-1)$ in the latter expression, the inequality simplifies to $\Gamma(x+1/2) > \Gamma(x) + \Gamma(x)(x-2)/(2(x-1))$. Dividing by $\Gamma(x)$ then results in $\Gamma(x+1/2)/\Gamma(x) > 1 + (x-2)/(2(x-1)) = (3x-4)/(2x-2)$. With $x = 1/2 + k/2$, the result is $\Gamma(k/2)/\Gamma((k-1)/2) > (3k-5)/(2k-2)$. □

Lemma A.2. *Let $\mathbf{x}, \mathbf{y} \in S^{k-1}$. Let $0 < r(\mathbf{x},\mathbf{y}) < 1$. For all $k \geq 3$,*

$$\frac{\mu\left(K\left(\frac{\mathbf{x}+\mathbf{y}}{|\mathbf{x}+\mathbf{y}|}, \cos^{-1}\left(\sqrt{\frac{2r(\mathbf{x},\mathbf{y})}{r(\mathbf{x},\mathbf{y})+1}}\right)\right)\right)}{\mu(S^{k-1})} < \frac{1}{\sqrt{\pi}} \left(\frac{1-r(\mathbf{x},\mathbf{y})}{1+r(\mathbf{x},\mathbf{y})}\right)^{(k-1)/2}. \tag{A.6}$$

Proof. Without loss of generality assume that $\frac{\mathbf{x}+\mathbf{y}}{|\mathbf{x}+\mathbf{y}|}\sqrt{\frac{2r(\mathbf{x},\mathbf{y})}{r(\mathbf{x},\mathbf{y})+1}} = (t, 0, 0, ..., 0)$ where $t = \sqrt{\frac{2r(\mathbf{x},\mathbf{y})}{r(\mathbf{x},\mathbf{y})+1}}$. Set $s = \sqrt{\frac{1-r(\mathbf{x},\mathbf{y})}{1+r(\mathbf{x},\mathbf{y})}}$. Radial projection of $C\left(\frac{\mathbf{x}+\mathbf{y}}{|\mathbf{x}+\mathbf{y}|}, \cos^{-1}\left(\sqrt{\frac{2r(\mathbf{x},\mathbf{y})}{r(\mathbf{x},\mathbf{y})+1}}\right)\right)$ (from the origin) onto $\{(u_1, u_2, ..., u_k) \in \mathbb{R}^k : u_1 = 1\} \cong \mathbb{R}^{k-1}$ results in the ball B_s^{k-1}. Thus,

$$\mu\left(K\left(\frac{\mathbf{x}+\mathbf{y}}{|\mathbf{x}+\mathbf{y}|}, \cos^{-1}\left(\sqrt{\frac{2r(\mathbf{x},\mathbf{y})}{r(\mathbf{x},\mathbf{y})+1}}\right)\right)\right) < 2\mu\left(B_s^{k-1}\right).$$

Also, $\mu(B^d) = \frac{\pi^{d/2}}{\Gamma(1+d/2)}$, $\mu(\partial B^d) = \frac{d\pi^{d/2}}{\Gamma(1+d/2)}$, and $\mu(B_a^d) = a^d \frac{\pi^{d/2}}{\Gamma(1+d/2)}$ (Ball, 1997, Lecture 1). Therefore,

$$\frac{\mu\left(K\left(\frac{\mathbf{x}+\mathbf{y}}{|\mathbf{x}+\mathbf{y}|}, \cos^{-1}\left(\sqrt{\frac{2r(\mathbf{x},\mathbf{y})}{r(\mathbf{x},\mathbf{y})+1}}\right)\right)\right)}{\mu(S^{k-1})} < \frac{2\mu\left(B_s^{k-1}\right)}{\mu(\partial B^k)} = \frac{\frac{2s^{k-1}\pi^{(k-1)/2}}{\Gamma(1/2+k/2)}}{\frac{k\pi^{k/2}}{\Gamma(1+k/2)}}$$

$$
\begin{aligned}
&= \frac{2s^{k-1}}{\sqrt{\pi}k}\frac{\Gamma(1+k/2)}{\Gamma(1/2+k/2)} \\
&< \frac{2}{\sqrt{\pi}}\frac{s^{k-1}}{k}\frac{k+3}{4} \\
&= \frac{2}{\sqrt{\pi}}\frac{s^{k-1}}{4}\frac{k+3}{k} \\
&\leq \frac{2}{\sqrt{\pi}}\frac{s^{k-1}}{4}\frac{2}{1} \\
&= \frac{s^{k-1}}{\sqrt{\pi}}.
\end{aligned}
$$

The third inequality is due to the hypothesis $k \geq 3$. The middle inequality is due to the following argument. In general, $\Gamma(x+1) = x\Gamma(x)$. By convexity, $\Gamma(x+1/2) < (\Gamma(x)+\Gamma(x+1))/2$. Thus, $\Gamma(x+1/2) < (\Gamma(x)+x\Gamma(x))/2$. Dividing by $\Gamma(x)$ results in $\Gamma(x+1/2)/\Gamma(x) < (1+x)/2$. With $x = 1/2+k/2$, $\Gamma(k/2)/\Gamma((k-1)/2) < (k+3)/4$. □

To complete the proof of Theorem 2.1, divide (A.1) by $\mu(S^{k-1})$, appeal to the inequalities in $(A.5)$ and $(A.6)$, and write the result in terms of n, before multiplying by $\sqrt{\pi}$.

Acknowledgements

Thanks to Seth Dutter for his assistance with the proof of Proposition A.1. Thanks to Paul Frater for providing data on *Schizachyrium scoparium.*

References

Ball, K. (1997). An elementary introduction to modern convex geometry. *Flavors of geometry*, MSRI Publications, Volume 31.

Cervellati, C., Bonaccorsi, G., Cremonini, E., Bergamini, C.M., Patella, A., Castaldini, C., Ferrazzini, S., Capatti, A., Picarelli, V., Pansini, F.S., and Massari, L. (2010). Bone mass density selectively correlates with serum markers of oxidative damage in post-menopausal women. *Clinical Chemistry and Laboratory Medicine,* Volume 51, Issue 2, Pages 333-338.

Chatfield, C. (1995). Model Uncertainty, Data Mining and Statistical Inference. *Journal of the Royal Statistical Society: A*, 158, Part 3, pp. 419466.

Cohen, J., Cohen, P., West, S.G., and Aiken, L.S. (2003). *Applied Multiple Regression / Correlation Analysis for the Behavioral Sciences, Third Edition.* Lawrence Erlbaum Associates, Publishers, Mahwah, New Jersey, London.

Federer, H. (1969). *Geometric measure theory.* Springer, New York.

Frank, K.A. (2000). Impact of a Confounding Variable on a Regression Coefficient. *Sociological Methods & Research*, Vol. 29, No. 2, 147-194.

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

Gordon, R.D. (1941). Values of Mills' ratio of area to bounding ordinate and of the normal probability integral for large values of the argument. *Annals of Mathematical Statistics*, Vol. 12, No. 3, Sept.

Greenland, S., and Morgenstern, H. (2001). Confounding in Health Research. *Annual Review of Public Health*, 22:189-212.

Harman, H.H. (1976) *Modern Factor Analysis, Third Edition.* University of Chicago Press.

Hawkins, D. (2004). The Problem of Overfitting. *Journal of Chemical Information and Modeling,* 44,1-12.

Hoeffding, W. (1948). A Non-Parametric Test of Independence. *The Annals of Mathematical Statistics,* Volume 19, Number 4, 546-557.

Hosman, C.A., Hansen, B.B., and Holland, P.W. (2010). The sensitivity of linear regression coefficients' confidence limits to the omission of a confounder. *Annals of Applied Statistics*, Vol. 4, No. 2, 849-870.

Howards, P.P., Schisterman, E.F., Poole, C., Kaufman, J.S., Weinberg, C.R. (2012). "Toward a clearer definition of confounding" revisited with directed acyclic graphs. *American Journal of Epidemiology*, 176(6):506-11.

Jungert, A., Roth, H.J., and Neuhäuser-Berthold, M. (2012). Serum 25-hydroxyvitamin D_3 and body composition in an elderly cohort from Germany: a cross-sectional study. *Nutrition & Metabolism*, 9:42.

Khoshnevisan, D. (2010). *Probability.* American Mathematical Society.

Knaeble, B. (2014). Certain effects of uncertain models. *International Journal of Advanced Statistics and Probability*, 2,2,124-134.

Leamer, E.E. (1975). A result on the sign of restricted least-squares estimates. *Journal of Econometrics*, 3, 387-390.

Lignell, S., Aune, M., Darnerud, P.O., Hanberg, A., Larsson, S.C., Glynn, A. (2013). 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.

Margaritis, D. and Thrun, S. (2001). A Bayesian Multiresolution Independence Test for Continuous Variables. *Proceedings of the 17th Annual Conference on Uncertainty in AI (UAI).*

McAleer, M., Pagan, A., and Visco, I. (1986). A further result on the sign of restricted least-squares estimates. *Journal of Econometrics,* 32, 287-290.

McNamee, R. (2003). Confounding and confounders. *Occupational & Environmental Medicine,* 60:227-234.

Nelson, R.K., Horowitz, J.F., Holleman, R.G., Swartz, A.M., Strath, J.S., Kriska, A.M., and Richardson, C.R. (2013). 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.

Oksanen, E.H. (1987). On sign changes upon deletion of a variable in linear regression analysis. *Oxford Bulletin of Economics and Statistics,* 49, 227-229.

Pearl, J (2009a). Causal inference in statistics: An overview. *Statistical Surveys*, Vol. 3, p 96-146.

Pearl, J (2009b). *Causality: models, reasoning and inference.* Cambridge University Press.

Puri, M.L. and Sen, P.K. (1971). *Nonparametric Methods in Multivariate Analysis.* Wiley, New York.

Rosenbaum, P.R., and Rubin, D.B. (1983b). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society, Series B*, Vol. 45, No. 2, p 212-218.

Seber, G. and Lee, A. (2003). *Linear Regression Analysis.* John Wiley & Sons, Inc., New Jersey.

Szekély, G.J., Rizzo, M.L., and Bakirov, N.K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics.*

Visco, I. (1988). Again on sign changes upon deletion of a variable from a linear regression. *Oxford Bulletin of Economics and Statistics,* 50, 225-227.

Wilks, S.S. (1935). On the independence of k sets of normally distributed statistical variables. *Econometrica,* 3, 309-326.