3.4 Sensitivity Analysis: People Who Look Comparable May Differ

What is sensitivity analysis?

If the naïve model (3.5)–(3.8) were true, the distribution of treatment assignments $\mathbf{Z}$ in a randomized paired experiment could be reconstructed by matching for the observed covariate, $\mathbf{x}$. It is common for a critic to argue that, in a particular study, the naïve model may be false. Indeed, it may be false. Typically, the critic accepts that the investigators matched for the observed covariates, $\mathbf{x}$, so treated and control subjects are seen to be comparable in terms of $\mathbf{x}$, but the critic points out that the investigators did not measure a specific covariate u, did not match for u, and so are in no position to assert that treated and control groups are comparable in terms of u. This criticism could be dismissed in a randomized experiment — randomization does tend to balance unobserved covariates — but the criticism cannot be dismissed in an observational study. This difference in the unobserved covariate u, the critic continues, is the real reason outcomes differ in the treated and control groups: it is not an effect caused by the treatment, but rather a failure on the part of the investigators to measure and control imbalances in u. Although not strictly necessary, the critic is usually aided by an air of superiority: "This would never happen in my laboratory."

It is important to recognize at the outset that our critic may be, but need not be, on the side of the angels. The tobacco industry and its (sometimes distinguished) consultants criticized, in precisely this way, observational studies linking smoking with lung cancer [103]. In this instance, the criticism was wrong. Investigators and their critics stand on level ground [8].

It is difficult if not impossible to give form to arguments of this sort until one has a way of speaking about the degree to which the naïve model is false. In an observational study, one could never assert with warranted conviction that the naïve model is precisely true. Trivially small deviations from the naïve model will have a trivially small impact on the study's conclusions. Sufficiently large deviations from the naïve model will overturn the results of any study. Because these two facts are always true, they quickly exhaust their usefulness. Therefore, the magnitude of the deviation is all-important. The sensitivity of an observational study to bias from an unmeasured covariate u is the magnitude of the departure from the naïve model that would need to be present to materially alter the study's conclusions.[11]

The first sensitivity analysis in an observational study concerned smoking and lung cancer. In 1959, Jerry Cornfield and his colleagues [15] asked about the magnitude of the bias from an unobserved covariate u needed to alter the conclusion

[11] In general, a sensitivity analysis asks how the conclusion of an argument dependent upon assumptions would change if the assumptions were relaxed. The term is sometimes misused to refer to performing several parallel statistical analyses without regard to the assumptions upon which they depend. If several statistical analyses all depend upon the same assumption — for instance, the naïve model (3.5) — then performing several such analyses provides no insight into consequences of the failure of that assumption.

from observational studies that heavy smoking causes lung cancer. They concluded that the magnitude of the bias would need to be enormous.

The sensitivity analysis model: Quantitative deviation from random assignment

The naïve model (3.5)–(3.8) said that two people, k and ℓ, with the same observed covariates, $\mathbf{x}_k = \mathbf{x}_\ell$, have the same probability of treatment given $(r_T, r_C, \mathbf{x}, u)$, i.e., $\pi_k = \pi_\ell$, where $\pi_k = \Pr(Z_k = 1 \mid r_{Tk}, r_{Ck}, \mathbf{x}_k, u_k)$ and $\pi_\ell = \Pr(Z_\ell = 1 \mid r_{T\ell}, r_{C\ell}, \mathbf{x}_\ell, u_\ell)$. The sensitivity analysis model speaks about the same probabilities in (3.1), saying that the naïve model (3.5)–(3.8) may be false, but to an extent controlled by a parameter, $\Gamma \geq 1$. Specifically, it says that two people, k and ℓ, with the same observed covariates, $\mathbf{x}_k = \mathbf{x}_\ell$, have odds[12] of treatment, $\pi_k/(1-\pi_k)$ and $\pi_\ell/(1-\pi_\ell)$, that differ by at most a multiplier of Γ; that is, in (3.1),

$$\frac{1}{\Gamma} \leq \frac{\pi_k/(1-\pi_k)}{\pi_\ell/(1-\pi_\ell)} \leq \Gamma \quad \text{whenever} \quad \mathbf{x}_k = \mathbf{x}_\ell\,. \tag{3.13}$$

If $\Gamma = 1$ in (3.13), then $\pi_k = \pi_\ell$, so (3.5)–(3.8) is true; that is, $\Gamma = 1$ corresponds with the naïve model. In §3.1, expression (3.1) was seen to be a representation and not a model — something that is always true for suitably defined u_ℓ — but that representation took $\pi_\ell = 0$ or $\pi_\ell = 1$, which implies $\Gamma = \infty$ in (3.13). In other words, numeric values of Γ between $\Gamma = 1$ and $\Gamma = \infty$ define a spectrum that begins with the naïve model (3.5)–(3.8) and ends with something that is hollow in the sense that it is always true, namely (3.1). The hollow statement that is always true, namely (3.1), is the statement that 'association does not imply causation,' that is, a sufficiently large departure from the naïve model can explain away as noncausal any observed association.

If $\Gamma = 2$, and if you, k, and I, ℓ, look the same, in the sense that we have the same observed covariates, $\mathbf{x}_k = \mathbf{x}_\ell$, then you might be twice as likely as I to receive the treatment because we differ in ways that have not been measured. For instance, if your $\pi_k = 2/3$ and my $\pi_\ell = 1/2$, then your odds of treatment rather than control are $\pi_k/(1-\pi_k) = 2$ or 2-to-1, whereas my odds of treatment rather than control are $\pi_\ell/(1-\pi_\ell) = 1$ or 1-to-1, and you are twice as likely as I to receive treatment, $\{\pi_k/(1-\pi_k)\}/\{\pi_\ell/(1-\pi_\ell)\} = 2$ in (3.13).[13]

---

[12] Odds are an alternative way of expressing probabilities. Probabilities and odds carry the same information in different forms. A probability of $\pi_k = 2/3$ is an odds of $\pi_k/(1-\pi_k) = 2$ or 2-to-1. Gamblers prefer odds to probabilities because odds express the chance of an event in terms of fair betting odds, the price of a fair bet. It is easy to move from probability π_k to odds $\omega_k = \pi_k/(1-\pi_k)$ and back again from odds ω_k to probability $\pi_k = \omega_k/(1+\omega_k)$.

[13] Implicitly, the critic is saying that the failure to measure u is the source of the problem, or that (3.5) would be true with $(\mathbf{x}, u)$ in place of $\mathbf{x}$, but is untrue with $\mathbf{x}$ alone. That is, the critic is saying $\pi_\ell = \Pr(Z_\ell = 1 \mid r_{T\ell}, r_{C\ell}, \mathbf{x}_\ell, u_\ell) = \Pr(Z_\ell = 1 \mid \mathbf{x}_\ell, u_\ell)$. As in §3.1, because of the delicate nature of unobserved variables, this is a manner of speaking rather than a tangible distinction. If the formalities are understood to refer to $\pi_\ell = \Pr(Z_\ell = 1 \mid r_{T\ell}, r_{C\ell}, \mathbf{x}_\ell, u_\ell)$, then it is not necessary to

Table 3.3 Sensitivity analysis for the one-sided 95% confidence interval for a constant, additive treatment effect τ on DNA elution rates. As usual, the hypothesis of a constant effect $H_0 : \tau = \tau_0$ is tested by testing no effect on $Y_i - \tau_0$ for the given value of Γ. The one-sided 95% confidence interval is the set of values of τ_0 not rejected in the one-sided, 0.05 level test. As Γ increases, there is greater potential deviation from random treatment assignment in (3.13), and the confidence interval grows longer. For instance, a treatment effect of $\tau_0 = 0.30$ would be implausible in a randomized experiment, $\Gamma = 1$, but not in an observational study with $\Gamma = 2$.

Γ	1	2	3
95% Interval	$[0.37, \infty)$	$[0.21, \infty)$	$[0.094, \infty)$

$$E\left(\overline{T}\middle|\mathscr{F},\mathscr{Z}\right) = \frac{1}{1+\Gamma}\sum_{i=1}^{I} s_i q_i, \tag{3.26}$$

while the variance becomes

$$\text{var}\left(\overline{T}\middle|\mathscr{F},\mathscr{Z}\right) = \text{var}\left(\overline{\overline{T}}\middle|\mathscr{F},\mathscr{Z}\right) = \frac{\Gamma}{(1+\Gamma)^2}\sum_{i=1}^{I}(s_i q_i)^2. \tag{3.27}$$

The remaining calculations are unchanged.

Sensitivity analysis for a confidence interval

Table 3.3 is the sensitivity analysis for the one-sided 95% confidence interval for an additive, constant treatment effect discussed in §2.4.2. As in a randomized experiment, the hypothesis that $H_0 : r_{Tij} = r_{Cij} + \tau_0$ is tested by testing the null hypothesis of no treatment effect on the adjusted responses, $R_{ij} - \tau_0 Z_{ij}$, or equivalently on the adjusted, treated-minus-control pair differences, $Y_i - \tau_0$. The one-sided 95% confidence interval is the set of values of τ_0 not rejected by a one-sided, 0.05 level test.

From Table 3.2, the hypothesis $H_0 : \tau = \tau_0$ for $\tau_0 = 0$ is barely rejected for $\Gamma = 4$ because the maximum possible one-sided P-value is 0.047. For $\Gamma = 3$, the maximum possible one-sided P-value is 0.04859 for $\tau_0 = .0935$ and is 0.05055 for $\tau_0 = .0936$, so after rounding to two significant digits, the one-sided 95% confidence interval is $[0.094, \infty)$.

Sensitivity analysis for point estimates

For each value of $\Gamma \geq 1$, a sensitivity analysis replaces a single point estimate, say $\widehat{\tau}$, by an interval of point estimates, say $[\widehat{\tau}_{\text{min}}, \widehat{\tau}_{\text{max}}]$ that are the minimum and maximum point estimates for all distributions of treatment assignments satisfying (3.16)–(3.18). Unlike a test or a confidence interval, and like a point estimate, this interval $[\widehat{\tau}_{\text{min}}, \widehat{\tau}_{\text{max}}]$ does not reflect sampling uncertainty; however, it does reflect uncertainty introduced by departures from random treatment assignment in (3.13) or (3.16)–(3.18).

Rosenbaum "sensitivity" site:r-project.org

About 79 results

[PDF] Package 'rbounds' - CRAN
https://cran.r-project.org/web/packages/rbounds/rbounds.pdf - Cached - Similar
Dec 8, 2014 **...** Title Perform **Rosenbaum** bounds **sensitivity** tests for matched and ... Function to calculate **Rosenbaum** bounds for binary data. Usage.

[PDF] Package 'sensitivitymv' - CRAN
https://cran.r-project.org/web/.../**sensitivity**mv/**sensitivity**mv.pdf - Cached - Similar
Aug 19, 2015 **...** Title **Sensitivity** Analysis in Observational Studies. Version 1.3. Date 2015-08-19. Author Paul R. **Rosenbaum**. Maintainer Paul R. **Rosenbaum** ...

[PDF] Package 'SensitivityCaseControl' - CRAN
https://cran.r-project.org/.../**Sensitivity**CaseControl/**Sensitivity**CaseControl.pdf - Cached - Similar
Oct 19, 2012 **...** Title **Sensitivity** Analysis for Case-Control Studies ... scribed in **Rosenbaum** (1991 , ``**Sensitivity** Analysis for Matched Case Control Studies'', ...

[PDF] Package 'sensitivitymw' - CRAN
https://cran.r-project.org/web/.../**sensitivity**mw/**sensitivity**mw.pdf - Cached - Similar
Jul 24, 2014 **...** Maintainer Paul R. **Rosenbaum** <**rosenbaum**@wharton.upenn.edu>. Description **Sensitivity** analysis analysis in matched observational studies ...

R: **Sensitivity** Analysis for Case-Control Studies
search.r-project.org/.../**Sensitivity**CaseControl/.../**Sensitivity**CaseControl-package.html - Cached
This package performs **sensitivity** analysis for matched case-control studies in ... Halloran and **Rosenbaum** (2012, "Case Definition and **Sensitivity** Analysis", ...

Introduction to Propensity Score Methods with R - useR! 2013 Tutorial
https://www.r-project.org/nosvn/conferences/useR.../Bryer.html - Cached - Similar
The use of propensity score methods (**Rosenbaum** & Rubin, 1983) for estimating ... An R Package for **Rosenbaum** bounds **sensitivity** analysis with matched data.

CRAN - Package SensitivityCaseControl
https://cran.r-project.org/package=**Sensitivity**CaseControl - Cached
May 30, 2014 **...** SensitivityCaseControl: **Sensitivity** Analysis for Case-Control Studies ... Halloran and **Rosenbaum** (2013, "Case Definition and **Sensitivity** ...

R: Function to calculate **Rosenbaum** bounds for IV Estimator based...
search.r-project.org/library/rbounds/html/iv_sens.html
Function to calculate **Rosenbaum** bounds for IV Estimator based on Wilcoxon ... iv_sens performs a non-parametric, instrumental variable **sensitivity** analysis on ...

[PDF] Package 'sensitivity2x2xk' - CRAN
https://cran.r-project.org/web/.../**sensitivity**2x2xk/**sensitivity**2x2xk.pdf - Cached - Similar
Dec 9, 2015 **...** Maintainer Paul R. **Rosenbaum** <**rosenbaum**@wharton.upenn.edu> ... table, adaptmh performs an exact, adaptive test and **sensitivity** analysis.

[PDF] ObsSensitivity: An R package for power analysis for **sensitivity** ...
https://www.r-project.org/conferences/.../010411-diamondneil.pdf - Cached - Similar
ObsSensitivity: An R package for power analysis for **sensitivity** ... Keywords: Observational Studies, Power, **Sensitivity** Analysis ... (**Rosenbaum**, 2010, p.4).

Package 'rbounds'

February 20, 2015

Version 2.1

Title Perform Rosenbaum bounds sensitivity tests for matched and unmatched data.

Date 2014-12-7

Author Luke J. Keele

Maintainer Luke J. Keele <ljk20@psu.edu>

Depends R (>= 2.8.1), Matching

Description Takes matched and unmatched data and calculates Rosenbaum bounds for the treatment effect. Calculates bounds for binary outcome data, Hodges-Lehmann point estimates, Wilcoxon signed-rank test for matched data and matched IV estimators, Wilcoxon sum rank test, and for data with multiple matched controls. Package is also designed to work with the Matching package and operate on Match() objects.

License GPL (>= 2)

NeedsCompilation no

Repository CRAN

Date/Publication 2014-12-08 07:23:24

R topics documented:

Observational Studies 1 (2015) 1-17 Submitted 7/14; Published 11/14

Two R Packages for Sensitivity Analysis in Observational Studies

Paul R. Rosenbaum **rosenbaum@wharton.upenn.edu**
Department of Statistics
Wharton School
University of Pennsylvania
Philadelphia, PA 19104-6340 US

Abstract

Two `R` packages for sensitivity analysis in observational studies are described. Package `sensitivitymw` is for matched pairs with one treated subject and one control, or matched sets with one treated subject and a fixed number, $K \geq 2$, of controls. Package `sensitivitymv` is for matched sets with variable numbers of controls. The packages offer conventional statistics, such as the permutational t-test and M-statistics using Huber's weights, but they also offer less familiar test statistics that have higher power in sensitivity analyses. The packages provide several tools useful in sensitivity analyses, such as an aid, `amplify`, to the interpretation of the value of the sensitivity parameter, and a device for combining evidence from several independent sensitivity analyses, `truncatedP`, for instance, several evidence factors or several subgroups.

Keywords: M-test; observational study; permutational t-test; randomization inference; sensitivity analysis.

1. Introduction

1.1 `R` Packages `sensivitymv` and `sensitivitymw`

The two `R` packages `sensivitymv` and `sensitivitymw` perform sensitivity analyses for observational studies with matched pairs or matched sets containing multiple controls. Package `sensitivitymw` is for matched pairs or matching with a fixed number of controls, for instance matching each treated subject to two controls. In contrast, package `sensivitymv` is for matched sets with variable numbers of controls, perhaps some treatment-control pairs together with some triples containing a treated subject and two controls. Also, the packages contain several data sets and several additional functions useful in sensitivity analysis. The packages overlap considerably, but package `sensitivitymw` is faster with additional features for matched pairs and for matching with a fixed number of controls. Both packages are available at CRAN and contain documentation.

My purpose here is to present a gentle introduction to these `R` packages, with pointers to articles for technical detail and pointers to the software documentation for additional options.

1.2 Scope of the current discussion

In an observational study, a sensitivity analysis replaces qualitative claims about whether unmeasured biases are present with an objective quantitative statement about the magnitude of bias that would need to be present to change the conclusions. In this sense, a sensitivity analysis speaks to the assertion "it might be bias" in much the same way that a P-value speaks to the assertion "it might be bad luck". If someone asserted that the higher responses in the treated group in a randomized experiment "might be bad luck," an unlucky randomization with no treatment effect, then a P-value does not deny the logical possibility of bad luck, but objectively measures the quantity of bad luck that would need to be present to alter the impression that the treatment did have an effect. In parallel, a sensitivity analysis measures the magnitude of bias from nonrandom treatment assignment that would need to be present to alter the conclusions of an observational study.

A sensitivity analysis is one tool useful in the large task of designing and interpreting an observational study. The discussion here is rather narrowly focused on carrying out such a sensitivity analysis in `R`.

1.3 What do the packages do?

In an observational study, treated and control subjects may be matched to be similar in terms of observed or measured covariates, but people who look similar in terms of measured covariates may still differ in terms of unmeasured covariates. The packages perform a sensitivity analysis asking about the magnitude of bias from nonrandom treatment assignment that would need to be present to alter the qualitative conclusions of a naive analysis that presumes matching for observed covariates removes all bias.

In a matched randomized experiment, each subject in a matched set has the same chance of being assigned to treatment or control because randomization has ensured that this is so. Without randomization, two people who look similar may differ in their chances of receiving treatment because they differ in terms of an unmeasured covariate not controlled by matching for measured covariates. The sensitivity analysis assumes that one subject in a matched set may be $\Gamma \geq 1$ times more likely than another to receive treatment because they differ in terms of unobserved covariates. If $\Gamma = 1$, then subjects who look the same are the same: matched subjects have equal chances of treatment, as in a randomized experiment. For $\Gamma = 1$, the sensitivity analysis reports a single answer, for instance a single P-value testing the null hypothesis of no treatment effect, and that single answer is the P-value that would be appropriate in a matched randomized experiment. For $\Gamma > 1$, there is no longer a single P-value, but rather an interval of possible P-values. The sensitivity analysis asks: How large must Γ be before the interval is so long that it is inconclusive, perhaps both accepting and rejecting the null hypothesis of no effect at the 0.05 level? The interval of possible P-values would be inconclusive in this sense if it extended from below 0.05 to above 0.05. The `senmw` and `senmv` functions compute sensitivity bounds for P-values. Specifically, they compute the upper bound on the P-value, for a specific Γ, so if that upper bound is at most 0.05, then a bias of magnitude Γ is too small to lead to acceptance of the null hypothesis. The `senmwCI` function inverts bounds on P-values to obtain sensitivity bounds for confidence intervals and point estimates. For detailed discussion of this model, see Rosenbaum (2002, §4; 2007).

Package 'sensitivitymw'

February 20, 2015

Type Package

Title Sensitivity analysis using weighted M-statistics

Version 1.1

Date 2014-04-29

Author Paul R. Rosenbaum

Maintainer Paul R. Rosenbaum <rosenbaum@wharton.upenn.edu>

Description Sensitivity analysis analysis in matched observational studies with multiple controls using weighted M-statistics to increase design sensitivity.

License GPL-2

LazyLoad yes

NeedsCompilation no

Repository CRAN

Date/Publication 2014-07-24 08:22:45

R topics documented:

105. Trochim, W.M.K.: Pattern matching, validity and conceptualization in program evaluation. Eval Rev **9**, 575–604 (1985)
106. Vandenbroucke, J.P.: When are observational studies as credible as randomized trials? Lancet **363**, 1728–1731 (2004)
107. VanderWeele, T.: The use of propensity score methods in psychiatric research. Int J Methods Psychol Res **15**, 95–103 (2006)
108. Volpp, K.G., Rosen, A.K., Rosenbaum, P.R., Romano, P.S., Even-Shoshan, O., Wang, Y., Bellini, L., Behringer, T., Silber, J.H.: Mortality among hospitalized Medicare beneficiaries in the first 2 years following ACGME resident duty hour reform. J Am Med Assoc **298**, 975–983 (2007)
109. Wang, L.S., Krieger, A.M.: Causal conclusions are most sensitive to unobserved binary covariates. Statist Med **25**, 2257–2271 (2006)
110. Weiss, N.S.: Can the "specificity" of an association be rehabilitated as a basis for supporting a causal hypothesis? Epidemiology **13**, 6–8 (2002)
111. Werfel, U., Langen, V., Eickhoff, I., Schoonbrood, J., Vahrenholz, C., Brauksiepe, A., Popp, W., Norpoth, K.: Elevated DNA single-strand breakage frequencies in lymphocytes of welders. Carcinogenesis **19**, 413–418 (1998)
112. Wintemute, G.J., Wright, M.A., Drake, C.M., Beaumont, J.J.: Subsequent criminal activity among violent misdemeanants who seek to purchase handguns: risk factors and effectiveness of denying handgun purchase. J Am Med Assoc **285**, 1019–1026 (2001)
113. Yu, B.B., Gastwirth, J.L.: Sensitivity analysis for trend tests: Application to the risk of radiation exposure. Biostatistics **6**, 201–209 (2005)
114. Zanutto, E., Lu, B., Hornik, R.: Using propensity score subclassification for multiple treatment doses to evaluate a national antidrug media campaign. J Educ Behav Statist **30**, 59–73 (2005)

for $i = 1, \ldots, I$, whereas in (3.16)–(3.18) the treatment assignment probabilities may vary from pair to pair, are unknown, but are bounded by $1/(1+\Gamma)$ and $\Gamma/(1+\Gamma)$. If $\Gamma = 1.0001$, then (3.14) would differ trivially from a randomized paired experiment, but as $\Gamma \to \infty$ the difference can become arbitrarily large.

Suppose that we had calculated a P-value or a point estimate or confidence interval from a paired observational study matched for observed covariates $\mathbf{x}$, by simply applying conventional statistical methods, that is, the methods in Chapter 2 for a randomized paired experiment. Those inferences would have their usual properties if the naïve model (3.5)–(3.8) were true, that is, if $\Gamma = 1$. How might those inferences change if Γ were some specific number larger than 1, indicating some bias due to failure to control for u? Using (3.16)–(3.18) and a few calculations, we can often deduce the range of possible P-values or point estimates or confidence intervals for a specified Γ. Consider, for instance, the P-value for testing the null hypothesis of no treatment effect. If the naïve model, $\Gamma = 1$, led to a P-value of, say, 0.001, and if $\Gamma = 2$ yields a range of possible P-values from 0.0001 to 0.02, then a bias of magnitude $\Gamma = 2$ creates greater uncertainty but does not alter the qualitative conclusion that the null hypothesis of no effect is not plausible. If the critic is thinking in terms of a moderately large deviation from a randomized trial, in which similar looking people may differ by a factor of $\Gamma = 2$ in their odds of treatment, then the critic is simply mistaken: the bias would have to be considerably larger than $\Gamma = 2$ to make no treatment effect plausible.

Every study is sensitive to sufficiently large biases. There is always a value of Γ such that, for that value and larger values of Γ, the interval of possible P-values includes small values, perhaps 0.0001, and large values, perhaps 0.1. A sensitivity analysis simply displays how the inference changes with Γ. For smoking and lung cancer, the bias would have to be enormous, $\Gamma = 6$; see [85, Chapter 4]. The question answered by a sensitivity analysis is: how large does Γ have to be before one must concede that the critic's criticism might be correct?

It is time to consider an example.

3.5 Welding Fumes and DNA Damage

Sensitivity analysis when testing the hypothesis of no treatment effect

The fumes produced by electric welding contain chromium and nickel and have been judged genotoxic in laboratory tests [39]. Werfel and colleagues [111] looked for evidence of DNA damage in humans by comparing 39 male welders to 39 male controls matched for age and smoking habits. Table 3.1 displays the comparability of the two groups with respect to the three covariates used in matching. Clearly, Table 3.1 is a rather limited demonstration of comparability.

Werfel et al. [111] presented several measures of genetic damage, including the measurement of DNA single strand breakage and DNA-protein cross-links using elution rates through polycarbonate filters with proteinase K. Broken strands are

Table 3.1 Covariate balance in 39 matched welder-control pairs. Covariates are gender, smoking and age.

		Welders	Controls
Male		100%	100%
Smokers		69%	69%
Age	Mean	39	39
	Minimum	23	23
	Lower Quartile	34	32
	Median	38	36
	Upper Quartile	46	46
	Maximum	56	59

expected to pass through filters more quickly, at a higher rate. Figure 3.1 depicts the elution rates and their matched pair differences. The differences are mostly positive, with higher elution rates for welders, and the differences are fairly symmetric about their median, with longer tails than the Normal distribution.

Table 3.2 is the sensitivity analysis for the one-sided P-value using Wilcoxon's signed rank statistic to test the null hypothesis of no treatment effect against the alternative that exposure to welding fumes caused an increase in DNA damage. The first row, $\Gamma = 1$, is the usual randomization inference, which would be appropriate if the 78 men had been paired for age and smoking and randomly assigned to their careers as a welder or a nonwelder. In the first row, the range of possible P-values is a single number, 3.1×10^{-7}, because there would be no uncertainty about the distribution of treatment assignments, $\mathbf{Z}$, in a randomized experiment. The naïve model (3.5)–(3.8) would also lead to $\Gamma = 1$ and the single P-value in the first row of Table 3.2. If this had been a randomized experiment, there would have been strong evidence against the null hypothesis of no effect. However, it was not a randomized experiment. The P-value in the first row of Table 3.2 says that it is implausible that the difference seen in Figure 3.1 is due to chance, the flip of a coin that assigned one man to treatment, another to control. The P-value in the first row of Table 3.2 does not speak to the critic's concern that the difference seen in Figure 3.1 is neither due to chance nor due to an effect caused by welding, but reflects instead some way that the matched welders and controls are not comparable. A small P-value, here 3.1×10^{-7}, computed assuming either randomization or equivalently the naïve model (3.5)–(3.8) does nothing to address the critic's concern. It is, however, possible to speak to that concern.

The second row permits a substantial departure from random treatment assignment or (3.5)–(3.8). It says that two men of the same age and smoking status — the same $\mathbf{x}$ — may not have the same chance of a career as a welder: one such man may be twice as likely as another to choose a career as a welder, $\Gamma = 2$, because they differ in terms of a covariate u that was not measured. This introduces a new source of uncertainty beyond chance. Using (3.16)–(3.18), we may determine every possible P-value that could be produced when $\Gamma = 2$, and it turns out that the smallest possible P-value is 3.4×10^{-12} and the largest possible P-value is 0.00064. Although

```
#### Week 5, Sensitivity calculations, DOS Ch 3, Rosenbaum vignette

> install.packages("sensitivitymw")
> data(erpcp) # Welders DNA damage
> dim(erpcp) # # data are outcomes in wide form; each row is a subclass
[1] 39  2
> head(erpcp)
  welder control
1  0.899   0.751
2  1.233   0.875
3  1.733   0.161
4  3.156   0.630
5  1.749   1.462
6  0.431   0.702
> attach(erpcp)
> boxplot(welder,control)# matches DOS Fig 3.1

> t.test(erpcp$welder, erpcp$control)
        Welch Two Sample t-test
data:  erpcp$welder and erpcp$control
t = 5.1442, df = 54.368, p-value = 3.785e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.3502495 0.7974940
sample estimates:
mean of x mean of y
1.3957436 0.8218718

> wilcox.test(erpcp$welder, erpcp$control)
        Wilcoxon rank sum test with continuity correction
data:  erpcp$welder and erpcp$control
W = 1251.5, p-value = 9.497e-07
alternative hypothesis: true location shift is not equal to 0
> wilcox.test(erpcp$welder - erpcp$control)
        Wilcoxon signed rank test
data:  erpcp$welder - erpcp$control
V = 715, p-value = 6.247e-07
alternative hypothesis: true location is not equal to 0

#### p.4 Gamma and p-values
>  senmw(erpcp, gamma = 1, method = "t")$pval
[1] 2.048115e-05
>  senmw(erpcp, gamma = 2, method = "t")$pval
[1] 0.003737467
>  senmw(erpcp, gamma = 3, method = "t")$pval
[1] 0.02275942
>  senmw(erpcp, gamma = 4, method = "t")$pval
[1] 0.0579339
> # I think doubling 'p' is right



> #### now to senmwCI page 5
>  senmwCI(erpcp, gamma = 1, method = "t", one.sided = TRUE)
$PointEstimate
minimum maximum
 0.5739  0.5739
$Confidence.Interval
minimum maximum
  0.394     Inf

>  senmwCI(erpcp, gamma = 1, method = "t", one.sided = FALSE) # get two-sided CI
$PointEstimate
minimum maximum
 0.5739  0.5739
```

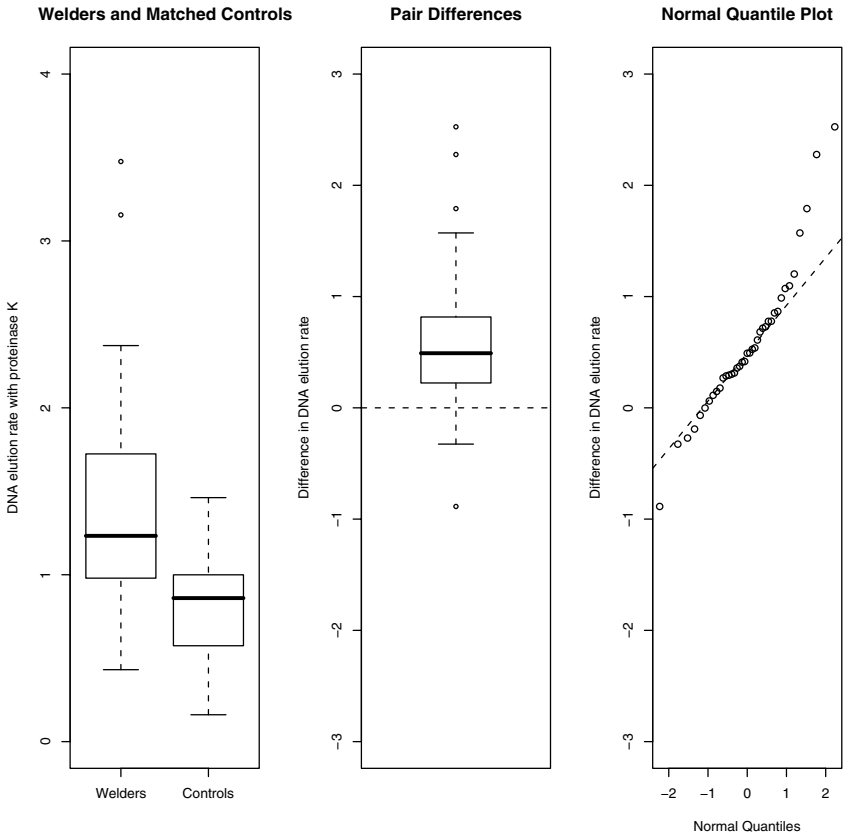

Fig. 3.1 DNA elution rates through polycarbonate filters with proteinase K for 39 male welders and 39 male controls matched for age and smoking. This assay is a measure of DNA single strand breakage and DNA-protein cross-links. In the boxplot of differences, there is a line at zero. In the Normal quantile plot, the line is fitted to the median and quartiles.

Table 3.2 Sensitivity analysis for the one-tailed P-value for testing the null hypothesis of no treatment effect on DNA elution rates with proteinase K in 39 pairs of a male welder and a male control matched for age and smoking. The table gives the lower (min) and upper (max) bounds on the one-sided P-value for departures from random assignment of various magnitudes, Γ. For $\Gamma = 1$, the two P-values are equal to each other and equal to the randomization P-value from Chapter 2. For $\Gamma > 1$, there is a range $[P_{\text{min}}, \quad P_{\text{max}}]$ of possible P-values. This study is sensitive only to very large biases, for instance $\Gamma = 5$, because at this point the range includes both small and large, significant and insignificant, P-values.

Γ	P_{min}	P_{max}
1	3.1×10^{-7}	3.1×10^{-7}
2	3.4×10^{-12}	0.00064
3	$< 10^{-15}$	0.011
4	$< 10^{-15}$	0.047
5	$< 10^{-15}$	0.108

In a randomized experiment, the permutational t-test is the randomization test that uses as its test statistic either the total, $T = \sum_{i=1}^{I} Y_i$, or the mean, $(1/I)\sum_{i=1}^{I} Y_i$, where these two statistics give the same permutational P-value. The permutation distribution of the mean, or the permutational t-test, is of historical and conceptual importance, in part because, in a randomized experiment, the expectation of $(1/I)\sum_{i=1}^{I} Y_i$ is the average treatment effect, $\{1/(2I)\}\sum_{i=1}^{I}\sum_{j=1}^{2}(r_{Tij} - r_{Cij})$.

2.2 Using the permutational t-test in matched pairs

Werfel et al. (1998) matched 39 welders exposed to chromium and nickel to 39 unexposed controls, measuring DNA damage in lymphocytes by DNA elution rates through polycarbonate filters with proteinase K (or ERPC+). Pairs were matched for age and smoking habits. The data frame `erpcp` in both packages has two columns, welder and control, and it contains the ERPC+ values for 39 pairs or rows.

The following calculations obtain the upper bound on the one-sided P-value testing the null hypothesis of no treatment effect using the permutational t-test (method="t"). For $\Gamma = 1$, this is the usual randomization P-value for the mean difference, namely 2.048×10^{-5}. For $\Gamma = 3$, the upper bound is 0.0228. For $\Gamma = 4$, the upper bound is 0.0579, so P-values well below and slightly above the conventional 0.05 level are possible under H_0 if the bias could be as large as $\Gamma = 4$. In other words, rejection of H_0 is sensitive to unmeasured biases of magnitude $\Gamma = 4$.

```
> library(sensitivitymw)
> data(erpcp)
> senmw(erpcp, gamma = 1, method = "t")$pval
[1] 2.048115e-05
> senmw(erpcp, gamma = 3, method = "t")$pval
[1] 0.02275942
> senmw(erpcp, gamma = 4, method = "t")$pval
[1] 0.0579339
```

Association does not imply causation, and that is always true, but logical implication tells us less than sensitivity analysis of the data at hand. The sensitivity analysis says that the observed association between welding and DNA elution rates is too strong to be explained by a bias of $\Gamma = 3$, because the maximum possible P-value from a bias of $\Gamma = 3$ is 0.0228, so a bias of that magnitude would not make the null hypothesis of no effect plausible. However, a bias of $\Gamma = 4$ would make the null hypothesis barely plausible, because with a bias that large, the P-value could be as large as $0.0579 > 0.05$. Saying that association does not imply causation is essentially the same as saying that the upper bound on the P-value tends to 1 as $\Gamma \to \infty$.

The P-value bounds are one-sided. In a sensitivity analysis, it is safe though somewhat conservative to obtain a two-sided P-value by doubling the smaller of two one-sided P-values, reporting a two-sided bound of $0.02275942 \times 2 = 0.04551884$ for $\Gamma = 3$. The reason doubling the one-sided P-value is conservative in a sensitivity analysis is that the bias that pushes the test statistic T into the upper tail is different from the bias that pushes it into

```
#### Week 5, Sensitivity calculations, DOS Ch 3, Rosenbaum vignette


> install.packages("sensitivitymw")
> data(erpcp) # Welders DNA damage
> dim(erpcp) # # data are outcomes in wide form; each row is a subclass
[1] 39  2
> head(erpcp)
  welder control
1  0.899   0.751
2  1.233   0.875
3  1.733   0.161
4  3.156   0.630
5  1.749   1.462
6  0.431   0.702
> attach(erpcp)
> boxplot(welder,control)# matches DOS Fig 3.1

> t.test(erpcp$welder, erpcp$control)
        Welch Two Sample t-test
data:  erpcp$welder and erpcp$control
t = 5.1442, df = 54.368, p-value = 3.785e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.3502495 0.7974940
sample estimates:
mean of x mean of y
1.3957436 0.8218718

> wilcox.test(erpcp$welder, erpcp$control)
        Wilcoxon rank sum test with continuity correction
data:  erpcp$welder and erpcp$control
W = 1251.5, p-value = 9.497e-07
alternative hypothesis: true location shift is not equal to 0
> wilcox.test(erpcp$welder - erpcp$control)
        Wilcoxon signed rank test
data:  erpcp$welder - erpcp$control
V = 715, p-value = 6.247e-07
alternative hypothesis: true location is not equal to 0

#### p.4 Gamma and p-values
>  senmw(erpcp, gamma = 1, method = "t")$pval
[1] 2.048115e-05
>  senmw(erpcp, gamma = 2, method = "t")$pval
[1] 0.003737467
>  senmw(erpcp, gamma = 3, method = "t")$pval
[1] 0.02275942
>  senmw(erpcp, gamma = 4, method = "t")$pval
[1] 0.0579339
> # I think doubling 'p' is right


> #### now to senmwCI page 5
>  senmwCI(erpcp, gamma = 1, method = "t", one.sided = TRUE)
$PointEstimate
minimum maximum
 0.5739  0.5739
$Confidence.Interval
minimum maximum
  0.394     Inf

>  senmwCI(erpcp, gamma = 1, method = "t", one.sided = FALSE) # get two-sided CI
$PointEstimate
minimum maximum
 0.5739  0.5739
```

Arguments

y	If y is an n by J matrix, then: (i) the rows are n matched sets, (ii) the first column is the treated response in a set, columns 2 to J contain the responses of controls in the same matched set. Every set must have J-1 controls, and NAs are not allowed in y. If y is a vector, then y is the vector of treated-minus-control pair differences in outcomes in n=length(y) matched pairs.
gamma	gamma is the sensitivity parameter, gamma=1 for a randomization test, gamma>1 for sensitivity bounds. Use of gamma<1 will generate an error. This parameter gamma is denoted by the upper case Greek letter gamma in the cited literature, for instance Rosenbaum (2007, 2014).
method	If method is NULL, then the method is determined by the parameters, namely inner, trim, lambda, m1, m2, and m. If method is not NULL, then these parameters are set according to the selected method and stated values of the parameters are ignored. The default values of the parameters are equivalent to method="h". (i) method = "h" (Huber, unweighted) is unweighted and sets inner=0, trim=3, lambda = 1/2, m1=m2=m=1. Method "h" is equivalent to the default settings. Its psi function levels off at 3 times the median (lambda = 1/2) of the absolute pair differences. The unweighted method h is often a good choice in small samples with few pairs or sets (say 20 sets). Unweighted method h is often a reasonable choice when the number of controls in each matched set is 6 or more. (Method "h" is almost the same as the default method for the senmv function in the sensitivitymv package, except: (a) senmv permits variable numbers of controls, (b) senmv uses trim = 2.5, not trim = 3.) (ii) method = "w" (weighted). Method "w" sets inner=0, trim=3, lambda=1/2, m1=12, m2=20, m=20. These weights are sturdy, all-purpose weights, often better than method="h" with 2-4 controls per matched set. Method="s" will often perform better for short-tailed Normal errors and method="l" will often perform better for long-tailed errors such as the t with 4 degrees of freedom. (iii) method = "f" (fixed choice weights). Method "f" sets inner=0, trim=3, lambda=1/2, m1=14, m2=20, m=20. Similar to method="w", method="f" uses all-purpose weights that were suggested, based on various calculations, in section 7.2 of Rosenbaum (2014) as the choice of a person who wants a "fixed choice" of weights. (iv) method = "s" (weighted for short tails) has weights appropriate for short tailed distributions, such as the Normal distribution. Method "s" sets inner=0, trim=3, lambda=1/2, m1=16, m2=20, m=20. (v) method = "l" (i.e., lower case letter L, weighted for long tails) has weights appropriate for long tailed distributions, such as the t-distribution with 4 degrees of freedom. It sets inner=0, trim=3, lambda=1/2, m1=12, m2=19, m=20. These weights redescend. The senmwCI function does not permit weights that redescend, and in particular does not permit method = "l". (vi) method = "q" (Quade) ranks sets using ordinary ranks (1, 2, ..., n) applied to ranges of M-scores within sets, in parallel with Quade (1979) and Tardiff (1987). It sets inner=0, trim=3, lambda=1/2, m1=2, m2=2, m=2. (vii) method = "t" (permutational t-test) is unweighted and permutes the observations themselves without ranking or scoring. It sets inner=0, trim=Inf,

the lower tail; see the related discussion of use of the Bonferroni inequality in sensitivity analyses in Rosenbaum and Silber (2009a, §4.5).

The function `senmwCI` computes point estimates and confidence intervals for an additive effect τ. For $\Gamma = 1$, there is a single point estimate, which for `method`= "`t`" is the mean difference, `mean(erpcp$welder-erpcp$control) = 0.5739`. The default is a one-sided 0.05-level confidence interval. (The level is controlled by `alpha` and one-or-two sided is controlled by `one.sided`.)

```
> senmwCI(erpcp, gamma = 1, method = "t", one.sided = TRUE)
$PointEstimate
minimum maximum
 0.5739  0.5739

$Confidence.Interval
minimum maximum
  0.394     Inf
```

For $\Gamma = 2$, there is no longer a single point estimate, 0.5739, but rather an interval of point estimates, [0.4167, 0.7487] and a longer 95% confidence interval, $\tau \geq 0.2081$. Notably, with a bias of at most $\Gamma = 2$, the smallest possible point estimate of τ, namely 0.4167, is still fairly large.

```
> senmwCI(erpcp, gamma = 2, method = "t", one.sided = TRUE)
$PointEstimate
minimum maximum
0.4167  0.7487

$Confidence.Interval
minimum maximum
 0.2081     Inf
```

In a sensitivity analysis, it is safe but somewhat conservative to form a 95% two-sided confidence interval as the intersection of two one-sided 97.5% confidence intervals, for the same reason that two-sided P-values are safe but somewhat conservative; see Rosenbaum (1995, §2.1) for some details.

2.3 M-statistics for matched pairs

An M-statistic gives each Y_i a controlled degree of influence. Let s be the median of the $|Y_i| = |R_{i1} - R_{i2}|$, as in Maritz (1979). For matched pairs, the M-statistic is $T = \sum_{i=1}^{I} \psi(Y_i/s)$ where $\psi(\cdot)$ is a suitable function. Taking $\psi(y) = y$ yields the same P-values as the permutational t-test. Huber (1981) proposed a $\psi(\cdot)$ that tops out at a constant $h > 0$ and bottoms out at $-h$, specifically $\psi(y) = \max\{-h, \min(y, h)\} = \text{sign}(y) \cdot \min(|y|, h)$, thereby limiting to $\pm hs$ the influence one observation Y_i can have on the statistic T.

With the default settings (or `method`= "`h`") in the `erpcp` data, the upper bounds on P-values using Huber's weights are similar to those from the permutational t-test in §2.2,

```
  #### Week 6, Sensitivity calculations, DOS Ch 3, Rosenbaum vignette


> install.packages("sensitivitymw")
> data(erpcp) # Welders DNA damage
> dim(erpcp) # # data are outcomes in wide form; each row is a subclass
[1] 39  2
> head(erpcp)
  welder control
1  0.899   0.751
2  1.233   0.875
3  1.733   0.161
4  3.156   0.630
5  1.749   1.462
6  0.431   0.702
> attach(erpcp)
> boxplot(welder,control)# matches DOS Fig 3.1

> t.test(erpcp$welder, erpcp$control)
        Welch Two Sample t-test
data:  erpcp$welder and erpcp$control
t = 5.1442, df = 54.368, p-value = 3.785e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.3502495 0.7974940
sample estimates:
mean of x mean of y
1.3957436 0.8218718

> wilcox.test(erpcp$welder, erpcp$control)
        Wilcoxon rank sum test with continuity correction
data:  erpcp$welder and erpcp$control
W = 1251.5, p-value = 9.497e-07
alternative hypothesis: true location shift is not equal to 0
> wilcox.test(erpcp$welder - erpcp$control)
        Wilcoxon signed rank test
data:  erpcp$welder - erpcp$control
V = 715, p-value = 6.247e-07
alternative hypothesis: true location is not equal to 0

#### p.4 Gamma and p-values
>  senmw(erpcp, gamma = 1, method = "t")$pval
[1] 2.048115e-05
>  senmw(erpcp, gamma = 2, method = "t")$pval
[1] 0.003737467
>  senmw(erpcp, gamma = 3, method = "t")$pval
[1] 0.02275942
>  senmw(erpcp, gamma = 4, method = "t")$pval
[1] 0.0579339
> # I think doubling 'p' is right


> #### now to senmwCI page 5
>  senmwCI(erpcp, gamma = 1, method = "t", one.sided = TRUE)
$PointEstimate
minimum maximum
 0.5739  0.5739
$Confidence.Interval
minimum maximum
  0.394     Inf

>  senmwCI(erpcp, gamma = 1, method = "t", one.sided = FALSE) # get two-sided CI
$PointEstimate
minimum maximum
 0.5739  0.5739
```

```
$Confidence.Interval
minimum maximum
 0.3561  0.7916

>  senmwCI(erpcp, gamma = 2, method = "t", one.sided = TRUE)
$PointEstimate
minimum maximum
 0.4167  0.7487
$Confidence.Interval
minimum maximum
 0.2081     Inf

>  senmwCI(erpcp, gamma = 2, method = "t", one.sided = FALSE)
$PointEstimate
minimum maximum
 0.4167  0.7487
$Confidence.Interval
minimum maximum
 0.1552  1.0414
```

matches two-sided t.test interval

```
### now try the "amplify stuff" (2009 JASA paper)
> install.packages("sensitivitymv")
##### page 6  sec 2.4
### Gamma = f(Lamda [bias in assignment], Delta [bias in outcome])
### amplify takes arguments Gamma, Lamda(values >Gamma)  produces Delta
> amplify(3,c(4:7))
        4         5         6         7
11.000000  7.000000  5.666667  5.000000
## discussion p.7



> ## mercury in RQ
```

lambda	Observations are scaled by the lambda quantile of the absolute pair differences. See the help file for senmw for more information.
m1	One of three parameters that determine the weights. See the discussion of m below.
m2	One of three parameters that determine the weights. See the discussion of m below.
m	One of three parameters that determine the weights. See the help file for senmw for more information. m2<m is not available for confidence intervals.
alpha	1-alpha is the coverage of the confidence interval.
one.sided	If TRUE, the confidence interval is one sided. If FALSE, the confidence interval is two-sided. The default is one-sided.
tol	The senmwCI function calls the R function uniroot, and tol is the tol (or tolerance) parameter in that call. If tol=NULL, senmwCI picks a reasonable tolerance.
interval	The senmwCI function calls the R function uniroot, and interval is the interval parameter in that call. If interval=NULL, senmwCI picks a reasonable interval.
detail	If detail=FALSE, the interval of point estimates and the confidence interval are reported after rounding based on tol. If detail=TRUE, then the results are not rounded, the tol and interval are reported.

Value

PointEstimate	An interval of point estimates allowing for a bias of gamma in treatment assignment. Rounded if detail=FALSE.
CI	An confidence interval allowing for a bias of gamma in treatment assignment. Rounded if detail=FALSE.
search.interval	If detail=TRUE, the interval of parameter values searched to find the estimates and confidence intervals.
tolerance	If detail=TRUE, the tolerance used in solving for estimates and confidence intervals.

Note

senmwCI inverts a test to obtain confidence intervals and point estimats; so, it calls senmw many times, solving several equations, and senmwCI is much slower than a single call to senmw. senmwCI finds point estimates and confidence intervals by searching for a value of the parameter tau in "interval" determining the solution tau.hat to an estimating equation with an error of "tol" in solving the equation. If interval=NULL and tol=NULL, senmwCI tries to pick a reasonable finite interval and tol>0. If concerned about these "reasonable values", set detail=TRUE, make the interval longer, the tol smaller, and wait longer for program to run. As illustrated in the examples, if there is reason for concern, the solutions produced by senmwCI can be checked by running senmw with tau set to the endpoints of the various intervals.

Unlike senmw, senmwCI does not permit redescending rank scores, m2<m or method="l".

but this will vary from one data set to another. In parallel, `senmwCI` may be used to obtain a sensitivity analysis for point estimates and confidence intervals.

```
> senmw(erpcp, gamma = 1, method = "h")$pval
[1] 6.402131e - 06
> senmw(erpcp, gamma = 2, method = "h")$pval
[1] 0.002410713
> senmw(erpcp, gamma = 3, method = "h")$pval
[1] 0.01859188
> senmw(erpcp, gamma = 4, method = "h")$pval
[1] 0.05304687
```

(Some comments about default settings follow. By default, `senmv`, `senmw` and `senmwCI` use the median of $|Y_i|$ to define s, but the user can select a different quantile by changing the value of `lambda`, the default being `lambda` $= 1/2$ for the median. By default, $h = 2.5$ in `senmv` and $h = 3$ in `senmw` and `senmwCI`, but the user can select different values by changing the value of `trim`. If the Y_i are discrete and most Y_i equal zero, the median $|Y_i|$ is not useful for scaling, and it may be reasonable to take `lambda` $= .90$ and $h =$ `trim` $= 1$, which resembles a trimmed mean.)

2.4 Amplification: an aid to interpreting Γ

When computing or reporting a sensitivity analysis, it is often convenient to have an analysis indexed by a single parameter, Γ. As discussed in §1.3, the sensitivity analysis reports the range of possible inferences when an unobserved bias alters the odds of treatment by a factor of at most Γ. The extremes of that range are produced by a bias strongly related to the outcome. An amplification interprets the single parameter Γ in terms of two parameters, one Λ controlling the relationship between the unobserved bias and treatment assignment Z_{ij}, the other Δ controlling the relationship between the unobserved bias and the outcome Y_i. Here, Λ is the maximum impact of the bias on the odds of treatment, $Z_{i1} - Z_{i2} = 1$, and Δ is the maximum impact of the unobserved bias on the odds of a positive response difference, $Y_i > 0$. A bias of Γ is equivalent to the curve defined by $\Gamma = (\Lambda\Delta + 1) / (\Lambda + \Delta)$. More precisely, under a certain semiparametric model for Y_i and $Z_{i1} - Z_{i2}$, a sensitivity analysis at Γ gives exactly the same P-value bounds as all sensitivity analyses at (Λ, Δ) such that $\Gamma = (\Lambda\Delta + 1) / (\Lambda + \Delta)$. In other words, one can calculate and report using one parameter Γ but have available the equivalent interpretations involving two parameters (Λ, Δ). See Rosenbaum and Silber (2009b) for a precise discussion.

The function `amplify` in the `sensitivitymv` package performs the required elementary calculations. Specifically, the call `amplify(gamma, lambda)` takes a scalar $\Gamma > 1$ and a vector of Λ's and computes the corresponding vector of Δ's. The analyses in §2.2 and §2.3 were insensitive to $\Gamma = 3$. The following call considers $\Lambda = (4, 5, 6, 7)$.

```
> library(sensitivitymv)
> amplify(3, c(4 : 7))
```

The result is:

4	5	6	7
11.00	7.00	5.67	5.00

For example, an unobserved covariate that increases the odds of treatment, $Z_{i1} - Z_{i2} = 1$, by at most $\Lambda = 5$ and the odds of a positive response difference, $Y_i > 0$, by at most $\Delta = 7$ is equivalent to $\Gamma = 3$. However, $\Gamma = 3$ is also equivalent to $(\Lambda, \Delta) = (7, 5)$, to $(\Lambda, \Delta) = (4, 11)$, to $(\Lambda, \Delta) = (11, 4)$, and to $(\Lambda, \Delta) = (6, 5.67)$. That is, a bias of $\Gamma = 3$ is quite a large bias, the omission of a covariate strongly related to both treatment assignment and response.

Similarly, `amplify(1.5,2)` yields 4, so $\Gamma = 1.5$ corresponds with both $(\Lambda, \Delta) = (2, 4)$ and $(\Lambda, \Delta) = (4, 2)$, while `amplify(1.25,2)` yields 2, so $\Gamma = 1.25$ corresponds $(\Lambda, \Delta) = (2, 2)$. In words, $\Gamma = 1.25$ corresponds with a doubling of the odds of treatment and a doubling of the odds of a positive response difference, not a trivially small bias. In $\Gamma = (\Lambda\Delta + 1) / (\Lambda + \Delta)$, as $\Delta \to \infty$, the corresponding Λ approaches Γ.

3. Matched sets with multiple controls

3.1 M-statistics with multiple controls

With $n_i \geq 2$ subjects in set i, there are $n_i - 1$ treated-minus-control pair differences, Y_{ik}, $k = 1, \ldots, n_i - 1$, all with the same treated subject, $Z_{ij} = 1$, but each with a different control, $Z_{i\ell} = 0$. The scale factor, s, is now defined to be the median of the $\sum_{i=1}^{I} \binom{n_i}{2}$ absolute differences, $\left|R_{ij} - R_{ij'}\right|$ with $j < j'$. The M-statistic is then $T = \sum_{i=1}^{I} w_i \sum_{k=1}^{n_i - 1} \psi\left(Y_{ik}/s\right)$, summing over all $\sum_{i=1}^{I} (n_i - 1)$ pair differences Y_{ik}, where set i is given weight w_i. See Rosenbaum (2007, 2014) for technical discussion of sensitivity analyses using these statistics.

There are various ways to attach weights w_i to matched sets, and `senmv` and `senmw` provide several options. Before discussing weights, consider an example with constant weights, essentially an unweighted example, in which every treated subject is matched to $n_i - 1 = 2$ controls.

3.2 Example with two controls

Fish often contains mercury. Does eating large quantities of fish increase levels of mercury in the blood? Data set `mercury` in the `sensitivitymw` package is from the 2009-2010 National Health and Nutrition Examination Survey (NHANES) and is the example in Rosenbaum (2014). There are 397 rows or matched triples and three columns, one treated with two controls. The values are methylmercury levels in blood in μg/dL. Column 1, "Treated", describes an individual who had at least 15 servings of fish or shellfish in the previous month. Column 2, "Zero", describes an individual who had 0 servings of fish or shellfish in the previous month. Column 2, "One", describes an individual who had 1 serving of fish or shellfish in the previous month. In the comparison here, Zero and One are not distinguished; both are controls. Sets were matched for gender, age, education, household income, black race, Hispanic, and cigarette consumption; see Table 1 in Rosenbaum (2014). A description of the data follows.

In a straightforward way, the senmv package may be used in calculations for approximate evidence factors in the sense of Rosenbaum (2011); see documentation for the truncatedP or truncatedPbv functions.

There are six data sets, erpcp, tbmetaphase, mercury, mtm, lead150 and lead250 . As noted in the documentation for senmv and truncatedP, these three data sets may be used to reproduce analyses from the cited literature, as illustrated in the examples for senmv and truncatedP. The documentation for mscorev shows how to reproduce an intermediate result, specifically Table 3 in Rosenbaum (2007).

Author(s)

Paul Rosenbaum

Maintainer: Paul R. Rosenbaum <rosenbaum@wharton.upenn.edu>

References

Huber, P. (1981) Robust Statistics. New York: Wiley, 1981.

Maritz, J. S. (1979) Exact robust confidence intervals for location. Biometrika 1979, 66, 163-166.

Rosenbaum, P. R. (2002) Observational Studies (2nd edition). New York: Springer.

Rosenbaum, P. R. (2007) Sensitivity analysis for m-estimates, tests and confidence intervals in matched observational studies. Biometrics, 2007, 63, 456-464.

Rosenbaum, P. R. and Silber, J. H. (2009) Amplification of sensitivity analysis in observational studies. Journal of the American Statistical Association, 104, 1398-1405.

Rosenbaum, P. R. (2011) Some approximate evidence factors in observational studies. Journal of the American Statistical Association, 106, 285-295.

Rosenbaum, P. R. (2013) Impact of multiple matched controls on design sensitivity in observational studies. Biometrics, 2013, 69, 118-127.

Zaykin, D. V., Zhivotovsky, L. A., Westfall, P. H. and Weir, B. S. (2002) Truncated product method of combining P-values. Genetic Epidemiology, 22, 170-185.

`amplify` *Amplification of sensitivity analysis in observational studies.*

Description

Uses the method in Rosenbaum and Silber (2009) to interpret a value of the sensitivity parameter gamma, for instance the parameter in the senmv function. Each value of gamma amplifies to a curve (lambda,delta) in a two-dimensional sensitivity analysis, the inference being the same for all points on the curve. That is, a one-dimensional sensitivity analysis in terms of gamma has a two-dimensional interpretation.

Usage

```
amplify(gamma, lambda)
```

Arguments

`gamma`	gamma > 1 is the value of the sensitivity parameter, for instance the parameter in senmv. length(gamma)>1 will generate an error.
`lambda`	lambda is a vector of values > gamma. An error will result unless lambda[i] > gamma > 1 for every i.

Details

A single value of gamma, say gamma = 3.5 in the example, corresponds to a curve of values of (lambda, delta), including (4, 26), (6,8), (8,6), and (11,5) in the example. An unobserved covariate that is associated with a lambda = 6 fold increase in the odds of treatment and a delta = 8 fold increase in the odds of a positive pair difference is equivalent to gamma = 3.5.

The curve is gamma = (lambda*delta + 1)/(lambda+delta). Amplify is given one gamma and a vector of lambdas and solves for the vector of deltas. The calculation is elementary.

This interpretation of gamma is developed in detail in Rosenbaum and Silber (2009), and it makes use of Wolfe's (1974) family of semiparametric deformations of an arbitrary symmetric distribuiton.

Strictly speaking, the amplification describes matched pairs, not matched sets. The senmv function views a k-to-1 matched set with k controls matched to one treated individual as a collection of k correlated treated-minus-control matched pair differences; see Rosenbaum (2007). For matched sets, it is natural to think of the amplification as describing any one of the k matched pair differences in a k-to-1 matched set.

The curve has asymptotes that the function amplify does not compute: gamma corresponds with (lambda,delta) = (gamma, Inf) and (Inf, gamma).

A related though distict idea is developed in Gastwirth et al (1998). The two approaches agree when the outcome is binary, that is, for McNemar's test.

Value

Returns a vector of values of delta of length(lambda) with names lambda.

Note

The example expands the discussion of Table 1 in Rosenbaum (2007). The study is insensitive to a bias of gamma = 3.5. An unobserved covariate associated with a lambda = 6 fold increase in the odds of treatment and a delta= 8 fold increase in the odds of positive pair difference is equivalent to gamma = 3.5. Also, gamma = 3.5 is equivalent to (lambda,delta) = (4,26), (8,6) and (11,5).

Author(s)

Paul R. Rosenbaum

References

Gastwirth, J. L., Krieger, A. M., Rosenbaum, P. R. (1998) Dual and simultaneous sensitivity analysis for matched pairs. Biometrika, 85, 907-920.

Rosenbaum, P. R. (2007) Sensitivity analysis for m-estimates, tests and confidence intervals in matched observational studies. Biometrics, 2007, 63, 456-464.

Rosenbaum, P. R. and Silber, J. H. (2009) Amplification of sensitivity analysis in observational studies. Journal of the American Statistical Association, 104, 1398-1405.

Wolfe, D. A. (1974) A charaterization of population weighted symmetry and related results. Journal of the American Statistical Association, 69, 819-822.

Examples

```
data(erpcp)
senmv(erpcp,gamma=3.5,trim=1)
amplify(3.5,6)
amplify(3.5,c(4,6,8,11))
```

`erpcp` *DNA Damage Among Welders*

Description

Matched pairs of a welder and a control, matching for age and smoking. The values are DNA elution rates through polycarbonate filters with proteinase K (or erpcp). Data are originally from Werfel et al. (1998) and were used as an example in Rosenbaum (2007). Data are used to illustrate the senmv function in the sensitivitymv package.

Usage

```
data(erpcp)
```

Format

A data frame with 39 observations on the following 2 variables.

`welder` erpcp value for the welder

`control` erpcp value for the matched control

Source

Werfel et al. (1998).

References

Rosenbaum, P. R. Sensitivity analysis for m-estimates, tests and confidence intervals in matched observational studies. Biometrics, 2007, 63, 456-464.

Werful, U., Langen, V., Eickhoff, I. et al. Elevated DNA strand breakage frequencies in lymphocytes of welders exposed to chromium and nickel. Carcinogenesis, 1998, 19, 413-418.

Examples

```
data(erpcp)
```

```
$Confidence.Interval
minimum maximum
 0.3561  0.7916

>  senmwCI(erpcp, gamma = 2, method = "t", one.sided = TRUE)
$PointEstimate
minimum maximum
 0.4167  0.7487
$Confidence.Interval
minimum maximum
 0.2081     Inf

>  senmwCI(erpcp, gamma = 2, method = "t", one.sided = FALSE)
$PointEstimate
minimum maximum
 0.4167  0.7487
$Confidence.Interval
minimum maximum
 0.1552  1.0414
```

```
### now try the "amplify stuff" (2009 JASA paper)
> install.packages("sensitivitymv")
##### page 6  sec 2.4
### Gamma = f(Lamda [bias in assignment], Delta [bias in outcome])
### amplify takes arguments Gamma, Lamda(values >Gamma)  produces Delta
> amplify(3,c(4:7))
        4         5         6         7
11.000000  7.000000  5.666667  5.000000
## discussion p.7
```

```
> ## mercury in RQ
```

```
# Save data objects
Y  <- lalonde$re78   # the outcome of interest
Tr <- lalonde$treat  # the treatment of interest

# Match - without replacement
mDW  <- Match(Y=Y, Tr=Tr, X=DWglm$fitted, replace=FALSE)

# One should check balance, but let's skip that step for now.

# Sensitivity Test:
hlsens(mDW, pr=.1, Gamma=1.5, GammaInc=.25)
```

`iv_sens` *Function to calculate Rosenbaum bounds for IV Estimator based on Wilcoxon sign rank test.*

Description

`iv_sens` performs a non-parametric, instrumental variable sensitivity analysis on matched pairs following the logic of the Neyman-Rubin framework for causal inference. The function supports a variable-valued instrument.

Usage

```
iv_sens(Rt, Rc, Dt, Dc, Gamma = 6, GammaInc = 1)
```

Arguments

`Rt,Rc`	Vectors of observed response outcomes for matched treatment and control observations, respectively.
`Dt,Dc`	Vectors of observed doses for matched observations, respectively. This is level of dose encouraged by the instrument.
`Gamma`	Upper-bound on gamma parameter.
`GammaInc`	To set user specified increments for gamma parameter.

Details

Given matched pairs of observations on an instrument `Z`, which encourages dose `D`, this function performs a Rosenbaum's bounds sensitivity analysis. Note that matching is done on levels of the instrument. See example below.

Value

Returns an object of class `rbounds`.

In a randomized experiment, the permutational t-test is the randomization test that uses as its test statistic either the total, $T = \sum_{i=1}^{I} Y_i$, or the mean, $(1/I)\sum_{i=1}^{I} Y_i$, where these two statistics give the same permutational P-value. The permutation distribution of the mean, or the permutational t-test, is of historical and conceptual importance, in part because, in a randomized experiment, the expectation of $(1/I)\sum_{i=1}^{I} Y_i$ is the average treatment effect, $\{1/(2I)\}\sum_{i=1}^{I}\sum_{j=1}^{2}(r_{Tij} - r_{Cij})$.

2.2 Using the permutational t-test in matched pairs

Werfel et al. (1998) matched 39 welders exposed to chromium and nickel to 39 unexposed controls, measuring DNA damage in lymphocytes by DNA elution rates through polycarbonate filters with proteinase K (or ERPC+). Pairs were matched for age and smoking habits. The data frame `erpcp` in both packages has two columns, welder and control, and it contains the ERPC+ values for 39 pairs or rows.

The following calculations obtain the upper bound on the one-sided P-value testing the null hypothesis of no treatment effect using the permutational t-test (method="t"). For $\Gamma = 1$, this is the usual randomization P-value for the mean difference, namely 2.048×10^{-5}. For $\Gamma = 3$, the upper bound is 0.0228. For $\Gamma = 4$, the upper bound is 0.0579, so P-values well below and slightly above the conventional 0.05 level are possible under H_0 if the bias could be as large as $\Gamma = 4$. In other words, rejection of H_0 is sensitive to unmeasured biases of magnitude $\Gamma = 4$.

```
> library(sensitivitymw)
> data(erpcp)
> senmw(erpcp, gamma = 1, method = "t")$pval
[1] 2.048115e - 05
> senmw(erpcp, gamma = 3, method = "t")$pval
[1] 0.02275942
> senmw(erpcp, gamma = 4, method = "t")$pval
[1] 0.0579339
```

Association does not imply causation, and that is always true, but logical implication tells us less than sensitivity analysis of the data at hand. The sensitivity analysis says that the observed association between welding and DNA elution rates is too strong to be explained by a bias of $\Gamma = 3$, because the maximum possible P-value from a bias of $\Gamma = 3$ is 0.0228, so a bias of that magnitude would not make the null hypothesis of no effect plausible. However, a bias of $\Gamma = 4$ would make the null hypothesis barely plausible, because with a bias that large, the P-value could be as large as $0.0579 > 0.05$. Saying that association does not imply causation is essentially the same as saying that the upper bound on the P-value tends to 1 as $\Gamma \to \infty$.

The P-value bounds are one-sided. In a sensitivity analysis, it is safe though somewhat conservative to obtain a two-sided P-value by doubling the smaller of two one-sided P-values, reporting a two-sided bound of $0.02275942 \times 2 = 0.04551884$ for $\Gamma = 3$. The reason doubling the one-sided P-value is conservative in a sensitivity analysis is that the bias that pushes the test statistic T into the upper tail is different from the bias that pushes it into

Author(s)

Luke Keele, Penn State University, <ljk20@psu.edu>

Jason W. Morgan, Ohio State University, <morgan.746@osu.edu>

References

Angrist, Joshua D., Imbens, Guido W., and Rubin, Donald B. (1996). "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91/434, pp. 444–455.

Rosenbaum, Paul R. (1996). "Comment." *Journal of the American Statistical Association* 91/434, pp. 465–468.

Rosenbaum, Paul R. (2002). *Observational Studies*. Springer-Verlag.

Rosenbaum, Paul R. (2010). *Design of Observational Studies*. Springer-Verlag.

See Also

See also data.prep, binarysens, hlsens, Match, mcontrol

Examples

```
## Example from Rosenbaum (2010, ch. 5).

data(AngristLavy)

#Match on Economic Status Across Levels of the Instrument
rr <- Match(Y=AngristLavy$avgmath, Tr=AngristLavy$z, X=AngristLavy$pct_disadv,
estimand ="ATC", M=2,  replace=FALSE)

#Extract Matched Outome Data
ctrl <- AngristLavy$avgmath[rr$index.control]
trt <- AngristLavy$avgmath[rr$index.treated]

#Extract Matched Doses
#Doses Encouraged By Instrument - Here Class Size
csize.trt <- AngristLavy$classize[rr$index.treated]
csize.ctrl <- AngristLavy$classize[rr$index.control]

#Run Sensitivity Analsyis
iv_sens(trt, ctrl, csize.trt, csize.ctrl, Gamma=2, GammaInc=.1)
```

---

mcontrol *Sensitivity Analysis For Multiple Matched Controls*

Description

Function to calculate Rosenbaum bounds for continuous or ordinal outcomes based on Wilcoxon sign rank test *p*-value when there are multiple matched control units.

Package 'ivpack'

February 20, 2015

Type Package

Title Instrumental Variable Estimation.

Version 1.2

Date 2014-10-24

Author Yang Jiang and Dylan Small

Maintainer Dylan Small <dsmall@wharton.upenn.edu>

Description This package contains functions for carrying out instrumental variable estimation of causal effects and power analyses for instrumental variable studies.

Depends AER, sandwich, lmtest

License GPL-2

NeedsCompilation no

Repository CRAN

Date/Publication 2014-10-25 08:07:54

R topics documented:

See Also

ivreg

Examples

```
### This is the IV model in panel A, column (5) of Table 3 from Card, 1995, "Using
### Geographic Variation in College Proximity to Esimate the Return from Schooling"
data(card.data)
ivmodel=ivreg(lwage ~ educ + exper + expersq + black + south + smsa + reg661 + reg662 +
reg663 + reg664 + reg665+ reg666 + reg667 + reg668 + smsa66, ~ nearc4 + exper +
expersq + black + south + smsa + reg661+ reg662 + reg663 + reg664 + reg665 + reg666 +
reg667 + reg668 + smsa66, x=TRUE, data=card.data)
anderson.rubin.ci(ivmodel)
```

ARsensitivity.ci *ARsensitivity.ci*

Description

Calculates the confidence interval for the effect of a treatment (endogenous) variable using an instrumental variable, which is based on an extension of Anderson-Rubin test and allows IV be possibly invalid within a certain range.

Usage

```
ARsensitivity.ci(ivmodel, Delta=NULL, conflevel=.95)
```

Arguments

ivmodel	Instrumental variable (IV) model fit using ivreg. Make sure to use the option x=TRUE when fitting the ivreg model.
Delta	The allowance of sensitivity parameter for possibly invalid IV. If Delta=NULL, the ARsensitivity.ci function will calculate the confidence interval for a standard Anderson-Rubin test with valid IV.
conflevel	Confidence level for confidence interval.

Value

confidence.interval	Confidence interval for effect of treatment. If it's a 2*2 matrix, the confidence interval is consisted of two disjoint intervals, each row of the matrix is one interval.
printinfo	Report the confidence interval in one printing sentence.
ci.type	If ci.type=1, the confidence interval is finite. If ci.type=2, the confidence interval is infinite. If ci.type=3, the confidence interval is an empty set.

Package 'ivmodel'

July 30, 2015

Type Package

Title Statistical Inference and Sensitivity Analysis for Instrumental Variables Model

Version 1.1

Date 2015-07-29

Author Yang Jiang, Hyunseung Kang, and Dylan Small

Maintainer Hyunseung Kang <hskang@stanford.edu>

Description Contains functions for carrying out instrumental variable estimation of causal effects, including power analysis, sensitivity analysis, and diagnostics.

Depends Matrix

License GPL-2

LazyData true

NeedsCompilation no

Repository CRAN

Date/Publication 2015-07-30 07:07:08

R topics documented: