Rank Tests for Heterogeneous Treatment Effects with Covariates

Roger Koenker[*]

Department of Economics
410 David Kinley Hall
1407 W. Gregory, MC-707
Urbana, IL 61801, USA *e-mail:* rkoenker@uiuc.edu

Abstract: Employing the regression rankscore approach of Gutenbrunner and Jurečková (1992) we consider rank tests designed to detect heterogeneous treatment effects concentrated in the upper tail of the conditional response distribution given other covariates.

AMS 2000 subject classifications: Primary 62G10; secondary 62J05.

1. Introduction

Heterogeneous treatment response has long been recognized as an essential feature of randomized controlled experiments. The Neyman (1923) framework of "potential outcomes" foreshadows modern developments by Rubin (1978) and others acknowledging the right of each experimental subject to have a distinct response to treatment. Statistical inference based on ranks has played an important role in these developments. Lehmann (1953) describes several heterogeneous treatment effect models and derives locally optimal rank tests for them. Rosenbaum (2007) has reemphasized the relevance of heterogeneity of treatment effects in biomedical applications and stressed the rank based approach to inference. He, Hsu and Hu (2009) have recently proposed tests based on "expected shortfall" designed to detect response in the upper or lower tail of the response distribution after adjusting for covariate effects.

Rank tests for the treatment-control model have focused almost exclusively on the two sample problem without considering possibly confounding covariate effects. In this paper we will describe some new rank tests designed for several heterogeneous treatment effect models. The tests employ the regression rankscores introduced by Gutenbrunner and Jurečková (1992) and therefore are able to cope with additional covariate effects.

[*] Partially supported by NSF Grant SES 08-50060. The author would like to thank Xuming He and Ya-Hui Hsu for valuable conversations on the subject of this paper.

2. Quantile Treatment Effects

For the two sample setting Lehmann (1974) introduced a general model of treatment response in the following way:

> Suppose the treatment adds the amount $\Delta(x)$ when the response of the untreated subject would be x. Then the distribution G of the treatment responses is that of the random variable $X + \Delta(X)$ where X is distributed according to F.

Thus, $F(x) = G(x + \Delta(x))$ so $\Delta(x)$ is the horizontal distance between the control distribution, F, and the treatment distribution, G,

$$\Delta(x) = G^{-1}(F(x)) - x.$$

Plotting $\Delta(x)$ versus x yields what is sometimes called the "shift plot." For present purposes we find it more convenient to evaluate $\Delta(x)$ at $x = F^{-1}(\tau)$ and define the quantile treatment effect as

$$\delta(\tau) = G^{-1}(\tau) - F^{-1}(\tau)$$

The *average* treatment effect can be obtained by simply integrating:

$$\bar{\delta} = \int_0^1 \delta(\tau) d\tau = \int (G^{-1}(\tau) - F^{-1}(\tau)) d\tau \equiv \mu(G) - \mu(F).$$

But mean treatment may obscure many important features of $\delta(\tau)$. Only in the pure location shift case do we not lose something by the aggregation. We now consider three simple models of the quantile treatment effect.

Partial Location Shift: Rather than assuming that the treatment induces a constant effect $\delta(\tau) = \delta_0$ over the entire distribution we may instead consider a partial form of the location shift restricted to an interval

$$\delta(\tau) = \delta_0 I(\tau_0 < \tau < \tau_1).$$

Thus, the shift may occur only in the upper tail, or near the median, or of course, over all of $(0, 1)$. □

Partial Scale Shift: Similarly, we may consider treatment effects that correspond to scale shifts of the control distribution over a restricted range,

$$\delta(\tau) = \delta_0 I(\tau_0 < \tau < \tau_1) F^{-1}(\tau).$$

Imagine stretching the right tail of the control distribution beyond some specified τ_0 quantile, while leaving the distribution below $F^{-1}(\tau_0)$ unperturbed. □

Lehmann Alternatives: The family of Lehmann (1953) alternatives may be expressed as

$$G(x) = F(x)^{\gamma} \qquad \text{or} \qquad 1 - G(x) = (1 - F(x))^{1/\gamma},$$

and has been widely considered in the literature in part perhaps because it is closely associated with the Cox proportional hazard model. In the two sample version of the Cox model, when $1/\gamma = k$, an integer, the treatment distribution is that of a random variable taking the minimum of k trials from the control distribution. The quantile treatment effect for the Cox form of the Lehmann alternative is easily seen to be,

$$\delta(\tau) = F^{-1}(1 - (1 - \tau)^{\gamma}) - F^{-1}(\tau). \tag{1}$$

Rosenbaum (2007) and Conover and Salsburg (1988) argue that the Lehmann family offers an attractive model for two sample treatment-control experiments in which a substantial fraction of subjects fail to respond to treatment, but the remainder exhibit a significant response. □

Each of the foregoing semi-parametric alternatives are intended to capture to some degree the idea that the treatment strongly influences the response, but in some restrictive way that makes conventional tests for a full location shift unsatisfactory. As in the motivating example of He, Hsu and Hu (2009) involving treatments for rheumatoid arthritis there is a need for a more targeted approach capable of detecting a more localized effect.

3. Rank Tests for QTEs

We very briefly review some general theory of rank tests in the regression setting based on the regression rankscores introduced by Gutenbrunner and Jurečková (1992). For further details see, Gutenbrunner *et al.* (1993) or Koenker (2005). Consider the linear quantile regression model

$$Q_{Y|X,Z}(\tau|x, z) = x^{\top}\beta(\tau) + z\delta(\tau). \tag{2}$$

We have a binary treatment variable, z, and p other covariates, denoted by the vector x. We would like to test the hypothesis $H_0 : \delta(\tau) \equiv 0$ versus local alternatives $H_n : \delta_n(\tau) = \delta_0(\tau)/\sqrt{n}$ in the presence of other covariate effects represented by the linear predictor $x^{\top}\beta(\tau)$ terms. Of course, in the two sample setting the latter term is simply an intercept. We will write X to denote the matrix with typical row x_i of the observed covariates.

Under the null hypothesis the regression rankscores are defined as,

$$\hat{a}(\tau) = \text{ argmax } \{a^\top y | X^\top a = (1-\tau)X^\top 1, \quad a \in [0,1]^n\}$$

This n-vector constitutes the dual solution to the quantile regression problem

$$\hat{\beta}(\tau) = \text{ argmin } \sum \rho_\tau(y_i - x_i^\top \beta).$$

The function $\hat{a}_i(\tau) = 1$ when $y_i > x_i^\top \hat{\beta}(\tau)$ and $\hat{a}_i(\tau) = 0$ when $y_i < x_i^\top \hat{\beta}(\tau)$ and integrating,

$$\hat{b}_i = \int_0^1 \hat{a}_i(\tau) d\tau \qquad i = 1, \ldots, n,$$

yields "ranks" of the observations. In the two sample setting these $\hat{a}_i(\tau)$'s are exactly the rankscores of Hájek (1965). Generalizing, we may consider integrating with another score function to obtain,

$$\hat{b}_i^\varphi = \int_0^1 \hat{a}_i(\tau) d\varphi(\tau).$$

As described in Hájek and Šidák (1967) the choice of φ is dictated by the form of the alternative H_n. When $\delta_0(\tau)$ is of the pure location shift form $\delta_0(\tau) = \delta_0$, there are three classical options for φ: normal (van der Waerden) scores $\varphi(\tau) = \Phi^{-1}(\tau)$, Wilcoxon scores $\varphi(\tau) = \tau$, and sign scores $\varphi(\tau) = |\tau - \frac{1}{2}|$. These choices are optimal iid error models

$$y_i = x_i^\top \beta + u_i \tag{3}$$

when the u_i's are Gaussian, logistic and double exponential, respectively. In this form the model is a special case of (2) in which the coordinates of $\beta(\tau)$ are all independent of τ except for the "intercept" component that takes the form $\beta_0(\tau) = F_u^{-1}(\tau)$, the quantile function of the iid errors. For simplicity of exposition, we will maintain this iid error model in the next subsection, with the understanding that eventually it may be relaxed.

4. Noncentralities and Scores

Choice of the score function, φ can be motivated by examining the noncentrality parameter of the corresponding rank tests under local alternatives. Our test statistic is

$$T_n = s_n^\top Q_n^{-1} s_n / A^2(\varphi)$$

where $s_n = (z - \hat{z})^\top \hat{b}_n, Q_n = (z - \hat{z})^\top (z - \hat{z}), \hat{z} = P_X z$, the projection of z onto the space spanned by the x covariates, and $A^2(\varphi) = \int (\varphi(t) - \bar{\varphi})^2 dt$, with $\bar{\varphi} = \int \varphi(t) dt$.

Theorem 1. *(Gutenbrunner, Jurečková, Koenker and Portnoy) Under the local alternative, $H_n : \delta_n(u) = \delta_0(u)/\sqrt{n}$ to the null model (3), T_n is asymptotically $\chi_1^2(\eta)$ with noncentrality parameter*

$$\eta = [Q_n A^2(\varphi)]^{-\frac{1}{2}} \int_0^1 f(F^{-1}(u))\delta_0(u)d\varphi(u).$$

A general strategy for selecting score functions, φ, is to optimize this noncentrality parameter given choices of $\delta_0(u)$ and f. In the case of location shift, $\delta_0(u) = \delta_0$,

$$\begin{aligned} \eta &= \delta_0 \int_0^1 f(F^{-1}(u))d\varphi(u) \\ &= -\delta_0 \int_0^1 \frac{f'}{f}(F^{-1}(u))\varphi(u)du, \end{aligned}$$

and optimal performance of the test is achieved by choosing $\varphi(u) = f'/f(F^{-1}(u))$, thereby achieving the same asymptotic efficiency as the likelihood ratio test. In the case of partial location shifts we may consider trimmed score functions of the form,

$$\varphi(u) = \frac{f'}{f}(F^{-1}(u))I(\tau_0 < u < \tau_1).$$

In particular we will consider the trimmed Wilcoxon scores $\varphi(u) = uI(\tau_0 < u < \tau_1)$ in the next section. Hettmansperger (1968) has previously considered symmetrically trimmed Wilcoxon tests motivated by robustness considerations.

For scale shift alternatives we have local alternatives of the form

$$\delta_n(u) = \delta_0 F^{-1}(u)/\sqrt{n}$$

and noncentrality parameter

$$\eta = [Q_n A^2(\varphi)]^{-\frac{1}{2}}\delta_0 \int f(F^{-1}(u))F^{-1}(u)d\varphi(u)$$

and again integrating by parts we have optimal score functions of the form,

$$\varphi(u) = -(1 + F^{-1}(u) \cdot \frac{f'}{f}(F^{-1}(u)))$$

which for the Gaussian distribution yields $\varphi(u) = (\Phi^{-1}(u))^2 - 1$. Again, we may consider partial scale shifts and obtain restricted forms.

Finally, for alternatives of the Lehmann type (1) we will consider localized versions with $\gamma_n = 1 + \gamma_0/\sqrt{n}$, so expanding,

$$\delta_n(u) = \gamma_0(f(F^{-1}(u)))^{-1}[-(1-u)\log(1-u)]/\sqrt{n} + o(1/\sqrt{n}),$$

and again integrating by parts in the noncentrality expression we have,

$$\begin{aligned}\eta &= -[Q_n A^2(\varphi)]^{-\frac{1}{2}}\gamma_0 \int [(1-u)\log(1-u)]d\varphi(u) \\ &= -[Q_n A^2(\varphi)]^{-\frac{1}{2}}\gamma_0 \int [\log(1-u)+1]\varphi(u)du,\end{aligned}$$

so the optimal score function is $\varphi(u) = \log(1-u)+1$. (An alternative derivation of this result can be found in Conover and Salsburg (1988)). An apparent advantage of this class of alternatives is that the score function is independent of the error distribution F.

5. Simulation Evidence

Throughout this section we will consider models that under the null hypothesis take the form,

$$y_i = \beta_0 + x_i\beta_1 + v_i$$

with v_i iid from some distribution, F, with Lebesgue density, f. The covariate, x will be standard normal. Three families of alternatives will be considered, one from each of the three general classes already discussed:

Location Shift	$\delta_n(u) = \gamma_n I(\tau_0 < u < \tau_1)$
Scale Shift	$\delta_n(u) = \gamma_n F^{-1}(u) I(\tau_0 < u < \tau_1)$
Lehmann Shift	$\delta_n(u) = F^{-1}(1-(1-u)^{\gamma_n}) - F^{-1}(u)$

where in the location and scale shift cases, $\gamma_n = \gamma_0/\sqrt{n}$ while in the Lehmann case $\gamma_n = 1+\gamma_0/\sqrt{n}$. Having specified quantile functions for the alternatives, it is straightforward to generate data according to these specifications. Under the alternatives we have,

$$y_i = \beta_0 + x_i\beta_1 + z_i\delta_n(U_i) + F^{-1}(U_i),$$

where the U_i are iid $U[0,1]$ random variables. The treatment indicator, z_i is generated as Bernoulli with probability $1/2$ throughout the simulations.

A convenient property of the regression rankscores is that they are invariant to the parameter, β, so we can take $\beta = 0$ for purposes of generating the data for the simulations. Of course, test statistics are based on inclusion of the covariate, x_i in estimation of the rankscores under the null model. Dependence between x_i and the treatment indicator is potentially a serious problem. Asymptotically, this is seen in the appearance of Q_n in the noncentrality parameter. But to keep things simple, we will maintain independence of x and z mimicking full randomization of treatment.

We consider the following collection of tests for “treatment effect:”

T	Student t-test
N	Normal (van der Waerden) rank test
S	Sign (median) rank test
W$[\tau_0, \tau_1]$	Trimmed Wilcoxon rank test
H$[\tau_0, \tau_1]$	Trimmed normal scale rank test
L	Lehmann Alternative rank test

All the rank tests are computed as described in Section 3, following Gutenbrunner *et al.* (1993). The piecewise linearity of the $\hat{a}_i(u)$ functions can be exploited, so

$$\hat{b}_i^{\varphi} = \int_0^1 \hat{a}_i(u)d\varphi(u) = \sum_{j=1}^{J} \frac{\hat{a}_i(\tau_j) - \hat{a}_i(\tau_{j-1})}{\tau_j - \tau_{j-1}} \int_{\tau_{j-1}}^{\tau_j} \varphi(u)du.$$

The last integral can be computed in closed form for all of our examples. See the function `ranks` in Koenker (2009) for further details.

In Table 1 we report results of a simulation with Gaussian F. Entries in the table represent empirical rejection frequencies for 10,000 replications. There are three sample sizes, three settings of the local alternative parameter, γ_0, and three distinct forms for the alternative hypothesis. Eight tests are evaluated: two versions of the Wilcoxon test one trimmed, one untrimmed; and two of the normal scale test one trimmed, one untrimmed. The first three columns of the table evaluate size of the test. These entries generally lie with experimental sampling accuracy for the nominal 0.05 level of the tests. Power of the tests for $\gamma_0 = 0.5$ and $\gamma_0 = 1$ are reported in the next six columns.

The restricted location shift alternative is specified as $\delta_n(u) = \gamma_n I(0.6 < u < 1)$ so there is no signal at the median and the poor performance of the sign test reflects this handicap. The other classical tests of global location shift also perform rather badly, even worse than the global normal scale test. The best performance is achieved by the trimmed Wilcoxon test, but the trimmed normal scale tests is also quite a strong contender.

The restricted scale shift alternative is specified as $\delta_n(u) = \gamma_n \Phi^{-1}(u) I(0.5 < u < 1)$ so again there is no signal at the median and the sign test is a disaster. The Student t test, the Wilcoxon, and the normal scores tests perform even worse than their lackluster showing for the location shift alternative. Here, not surprisingly given that it was designed for this situation, the trimmed normal scale test is the clear winner.

The Lehmann alternative affords an opportunity for all the tests to demonstrate some strength; these alternatives combine features of global location and scale shift with a more pronounced effect in the right tail so all the tests have something to offer. Again, not surprisingly, the Lehmann test designed for this situation is the clear winner, but the classical location shift tests are not far behind. Only the global normal scale test is poor in this case.

The banal conclusion that may be drawn from Table 1 seems to be that it pays to know what the alternative is before choosing a test. But if we delve slightly deeper we may be led to the conclusion that the Lehmann alternatives are quite adequately countered by traditional rank tests, while the asymmetric forms of the Wilcoxon and normal scale tests are better for stronger forms of asymmetric response captured in the partial location and scale shift alternatives. Before jumping to such conclusions, however, it would be prudent to consider whether the normality assumption that underlies all of the simulation results of Table 1 is critical.

	$\gamma_0 = \mathbf{0}$			$\gamma_0 = \mathbf{0.5}$			$\gamma_0 = \mathbf{1}$		
	n=50	n=100	n=500	n=50	n=100	n=500	n=50	n=100	n=500
Location									
T	0.0518	0.0560	0.0523	0.1234	0.1448	0.1566	0.3212	0.3402	0.4468
N	0.0540	0.0561	0.0516	0.1133	0.1359	0.1531	0.2030	0.2577	0.4090
W[0,1]	0.0559	0.0576	0.0524	0.1045	0.1188	0.1262	0.1693	0.1982	0.3070
S	0.0678	0.0649	0.0542	0.0752	0.0510	0.0536	0.0519	0.0460	0.0534
W[.6,.95]	0.0547	0.0514	0.0527	0.2906	0.3667	0.4504	0.5341	0.7156	0.9175
H[0,1]	0.0363	0.0432	0.0467	0.1473	0.2179	0.2538	0.3882	0.4926	0.7166
H[.5,1]	0.0300	0.0434	0.0514	0.2211	0.3376	0.3844	0.6654	0.7827	0.9055
L	0.0460	0.0529	0.0531	0.1846	0.2612	0.2970	0.4481	0.5744	0.7831
Scale									
T	0.0496	0.0569	0.0514	0.1033	0.1382	0.1593	0.2671	0.2984	0.4277
N	0.0506	0.0573	0.0507	0.0903	0.1123	0.1451	0.1557	0.1867	0.3368
W[0,1]	0.0531	0.0565	0.0500	0.0798	0.0894	0.0974	0.1290	0.1395	0.2066
S	0.0698	0.0580	0.0520	0.0709	0.0473	0.0553	0.0569	0.0507	0.0562
W[.6,.95]	0.0536	0.0554	0.0491	0.1665	0.2077	0.2635	0.3610	0.4593	0.6506
H[0,1]	0.0346	0.0412	0.0453	0.1118	0.2026	0.3093	0.2561	0.3817	0.7460
H[.5,1]	0.0318	0.0440	0.0475	0.1385	0.3205	0.4873	0.4336	0.6418	0.9326
L	0.0460	0.0539	0.0493	0.1307	0.2175	0.3282	0.2999	0.4208	0.7556
Lehmann									
T	0.0545	0.0534	0.0488	0.3866	0.4347	0.5420	0.7795	0.8618	0.9612
N	0.0559	0.0547	0.0493	0.3719	0.4215	0.5379	0.7388	0.8403	0.9585
W[0,1]	0.0568	0.0544	0.0507	0.3700	0.4093	0.5093	0.7273	0.8291	0.9457
S	0.0717	0.0594	0.0570	0.3145	0.2830	0.3698	0.5395	0.6520	0.8246
W[.6,.95]	0.0555	0.0514	0.0508	0.3802	0.4402	0.5512	0.7885	0.8601	0.9662
H[0,1]	0.0366	0.0364	0.0483	0.0459	0.0841	0.1397	0.0812	0.1022	0.3149
H[.5,1]	0.0336	0.0433	0.0468	0.2709	0.4081	0.5494	0.7111	0.8240	0.9616
L	0.0500	0.0529	0.0474	0.3892	0.4808	0.6111	0.8034	0.8920	0.9823

TABLE 1

Rejection Frequencies for Several Rank Tests: Nominal level of significance for all tests is 0.05, table entries are each based on 10,000 replications, all models have standard normal iid errors under the null and local alternatives with the indicated γ_0 parameters.

Table 2 reports simulation results for an almost identical experimental setup except that Gaussian error is replaced everywhere by Student t_3 error. Most of the features of the two tables are very similar. Especially in the partial location shift setting one sees even worse performance of the classical global rank tests and the t test. Performance of the Lehmann test deteriorates somewhat for both the location and scale alternatives under Student errors, but remains strong for the Lehmann alternative.

	$\gamma_0 = \mathbf{0}$			$\gamma_0 = \mathbf{0.5}$			$\gamma_0 = \mathbf{1}$		
	n=50	n=100	n=500	n=50	n=100	n=500	n=50	n=100	n=500
Location									
T	0.0447	0.0491	0.0455	0.0718	0.0883	0.0962	0.1521	0.1880	0.2047
N	0.0498	0.0532	0.0477	0.0756	0.0889	0.1054	0.1160	0.1627	0.2187
W[0,1]	0.0534	0.0538	0.0493	0.0791	0.0860	0.1013	0.1132	0.1438	0.1991
S	0.0645	0.0448	0.0537	0.0647	0.0586	0.0493	0.0544	0.0507	0.0520
W[.6,.95]	0.0502	0.0523	0.0507	0.1746	0.2314	0.3133	0.3774	0.5304	0.7467
H[0,1]	0.0354	0.0421	0.0496	0.0757	0.1038	0.1246	0.1693	0.2411	0.3226
H[.5,1]	0.0304	0.0419	0.0519	0.0930	0.1520	0.1720	0.2677	0.3839	0.4714
L	0.0445	0.0486	0.0488	0.0964	0.1314	0.1575	0.2055	0.3043	0.3987
Scale									
T	0.0494	0.0475	0.0498	0.0753	0.1075	0.1163	0.1430	0.2264	0.2937
N	0.0566	0.0518	0.0505	0.0750	0.0902	0.0976	0.1068	0.1529	0.2157
W[0,1]	0.0598	0.0538	0.0521	0.0707	0.0805	0.0802	0.0975	0.1237	0.1527
S	0.0713	0.0491	0.0531	0.0656	0.0642	0.0496	0.0582	0.0528	0.0545
W[.6,.95]	0.0542	0.0532	0.0507	0.1289	0.1688	0.1952	0.2568	0.3697	0.5176
H[0,1]	0.0339	0.0426	0.0497	0.0741	0.1246	0.1709	0.1315	0.2521	0.4378
H[.5,1]	0.0293	0.0415	0.0508	0.0776	0.1902	0.2493	0.1831	0.4136	0.6404
L	0.0469	0.0510	0.0512	0.0919	0.1448	0.1815	0.1647	0.3004	0.4724
Lehmann									
T	0.0436	0.0459	0.0465	0.2645	0.4320	0.5146	0.4851	0.7550	0.9345
N	0.0497	0.0488	0.0474	0.3319	0.4286	0.5270	0.6928	0.8440	0.9551
W[0,1]	0.0536	0.0495	0.0490	0.3361	0.4129	0.4979	0.6994	0.8360	0.9426
S	0.0703	0.0468	0.0560	0.2698	0.3174	0.3462	0.5261	0.6585	0.8242
W[.6,.95]	0.0525	0.0513	0.0504	0.3447	0.4540	0.5401	0.7158	0.8699	0.9578
H[0,1]	0.0366	0.0435	0.0490	0.0393	0.0894	0.1377	0.0320	0.0927	0.2932
H[.5,1]	0.0336	0.0412	0.0473	0.2144	0.4222	0.5439	0.4772	0.8293	0.9614
L	0.0468	0.0488	0.0509	0.3417	0.4884	0.6057	0.7082	0.8982	0.9799

TABLE 2

Rejection Frequencies for Several Rank Tests: Nominal level of significance for all tests is 0.05, table entries are each based on 10,000 replications, all models have iid Student t_3 errors under the null and local alternatives with the indicated γ_0 parameters.

6. Conclusions

Rank tests continue to play an important role in many domains of statistical application like survival analysis, but their potential value in the context of linear models remains under-appreciated. The regression rankscore methods of Gutenbrunner and

Jurečková (1992) have opened a wide vista of new opportunities for rank based inference in the regression setting. More targeted inference is particularly important in the context of heterogeneous treatment models. We have taken a few steps in this direction, but there are interesting new paths ahead.

References

Conover, W. and Salsburg, D. (1988). Locally Most Powerful Tests for Detecting Treatment Effects When only a Subset of Patients Can Be Expected to 'Respond' to Treatment. *Biometrics* **44** 189–196.

Gutenbrunner, C. and Jurečková, J. (1992). Regression Quantile and Regression Rank Score Process in the Linear Model and Derived Statistics. *Ann. Statist.* **20** 305-330.

Gutenbrunner, C., Jurečková, J., Koenker, R. and Portnoy, S. (1993). Tests of Linear Hypotheses Based on Regression Rank Scores. *J. Nonparametric Statistics* **2** 307–331.

Hájek, J. and Šidák, Z. (1967). *Theory of Rank Tests*. Academia, Prague.

He, X., Hsu, Y.-H. and Hu, M. (2009). Detection of Treatment Effects by Covariate Adjusted Expected Shortfall. preprint.

Hettmansperger, T. (1968). On the Trimmed Mann-Whitney Statistic. *Annals of Math. Stat.* **39** 1610–1614.

Koenker, R. (2005). *Quantile Regression*. Cambridge U. Press: Cambridge.

Koenker, R. (2009). quantreg. R package version 4.45, available from = http://CRAN.R-project.org/package=quantreg.

Lehmann, E. (1953). The Power of Rank Tests. *Ann. Math. Stat.* **24** 23–43.

Lehmann, E. (1974). *Nonparametrics: Statistical Methods based on Ranks*. Holden-Day.

Neyman, J. (1923). On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. *Statistical Science* **5** 465–472. (In translation from the original Polish).

Rosenbaum, P. R. (2007). Confidence Intervals for Uncommon but Dramatic Responses to Treatment. *Biometrics* **63** 1164–1171.

Rubin, D. B. (1978). Bayesian Inference for Causal Effects: The Role of Randomization. *The Annals of Statistics* **6** 34–58.