Randomization Inference for Outcomes with Clumping at Zero

Randomization Inference for Outcomes with Clumping at Zero*

Luke Keele† Luke Miratrix ‡

August 4, 2015

Abstract

In randomized experiments, randomization forms the "reasoned basis for inference." While randomization inference is well developed for continuous and binary outcomes, there has been comparatively little work for outcomes with nonnegative support and clumping at zero. Typically outcomes of this type have been modeled using parametric models that impose strong distributional assumptions. This article proposes new randomization inference procedures for nonnegative outcomes with clumping at zero. Instead of making distributional assumptions, we propose various assumptions about the nature of response to treatment. Our methods form a set of nonparametric methods for outcomes that are often described as zero-inflated. These methods are illustrated using two randomized trials where job training interventions were designed to increase earnings of participants.

KEYWORDS: Randomization Inference, Zero-Inflated, Tobit

*For comments and suggestions, we thank...

†Associate Professor, Department of Political Science, Penn State University, University Park, PA 16802, Email: ljk20@psu.edu.

‡Assistant Professor, Department of Statistics, Harvard University, Cambridge, MA, 02138 , Email: lmiratrix@stat.harvard.edu

1 Introduction

1.1 Censored Outcomes in Randomized Trials

In many applications, outcome data are nonnegative with a large proportion of recorded zeros. These are often modeled as being a censored version of an underlying continuous outcome, or a mixture of two distributions, one of which is a point mass. An early example of a statistical model for censoring is the "Tobit Model" **?**, first developed to a study expenditures on capital goods such as automobiles and major household appliances. In this study, when households were asked if they have purchases such a capital good in the last 12 months, they often reported zero expenditures. Here, under the assumption that the observed data arise from a censored normal distribution, maximum likelihood-based estimation leads to an appropriate regression model for the censored outcomes.

When modeled as a mixture, outcomes with a clump at zero are often described as zero-inflated. The canonical example of zero-inflation occurs with count data that display a higher proportion of zeros than expected under the Poisson distribution. For a zero-inflated count outcome, analysts might assume that are at least two processes that might produce a nonzero outcome. For example, consider an anti-smoking intervention. We might observe that a person smoked no cigarettes either due to the intervention or because they were already a non-smoker. Zero-inflated models for such processes are based on likelihoods that are typically a mixture of binomial and poisson distributions (**?**). See **?** for a detailed overview of parametric modeling approaches to outcomes with point masses at zero.

Both Tobit models and zero-inflated models rely on strong parametric assumptions. If the underlying errors of the Tobit model are either heteroskedastic or non-normal, the maximum likelihood estimates for the Tobit model are inconsistent (**?**). Zero-inflated models require strong parametric assumptions or the presence of an exclusion restriction for identification (**?**). In a randomized trial, where the goal is to make robust causal claims, this is particularly problematic as the para-

metric assumptions necessary for both models are not implied by the randomization. This could undermine the integrity of any subsequent analysis.

One alternative to parametric based inference is randomization inference which only assumes random assignment of the treatment. Here, randomization serves as the "reasoned basis for inference," and does not require assumptions beyond the randomization itself (**?**). Here under some alternative assumptions arguably weaker than full parametric ones, randomization inference tests of no effect can also be inverted to provide distribution-free confidence intervals, and analysts can use the method of Hodges-Lehmann to obtain point estimate for treatment effects. See **?**, ch. 2 for a review of randomization inference methods.

While randomization inference methods are well developed for continuous outcomes, there has been comparatively little work for outcomes with point masses at zero. Existing work tends to focus on testing without consideration of interval or point estimation (**??**). In this paper, we develop procedures for handling outcomes with clumping at zero within the randomization inference framework with a particular focus on point and interval estimates. While we avoid parametric assumptions, we must make assumptions about the nature of individual level response to treatment. To that end, we also develop procedures for assessing the assumptions we invoke.

After giving an overview of the two empirical examples we use to illustrate this approach, we then review randomization inference and the Tobit model of effects developed by **?**. We then outline the more general testing framework and develop targeted test statistics that focus on different consequences of Tobit-style assumptions. We then extend the Tobit model to allow for scaling in the positive outcomes and discuss how to estimate parameters for this more general approach. We finally illustrate using two empirical jobs training examples, which we describe next.

1.2 Empirical Applications

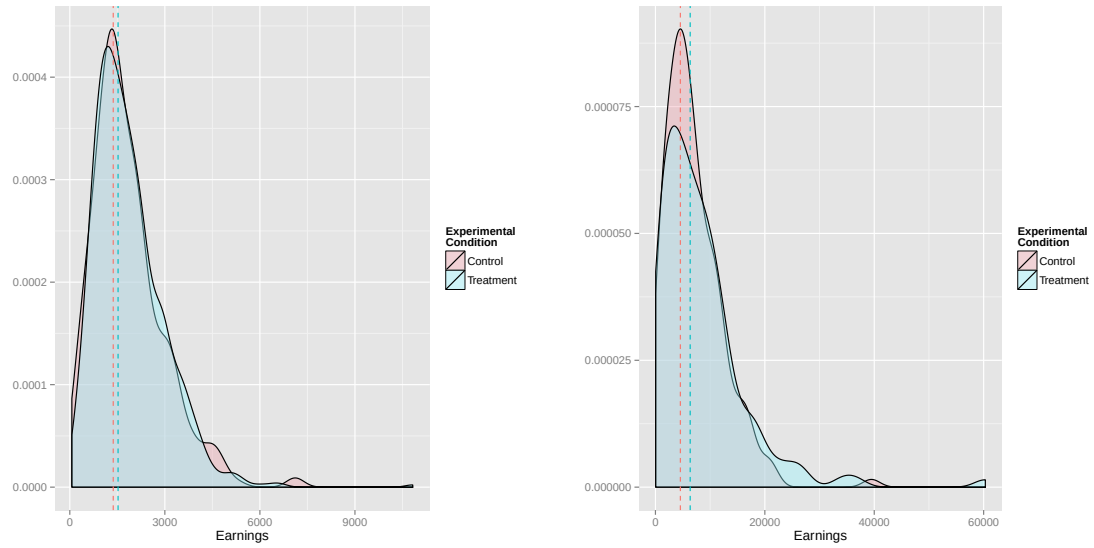
Outcomes with point masses at zero are particularly common in job training RCTs when the outcome is measured as earnings. One example is the JOBS II Intervention Project developed at

the University of Michigan and designed to enhance the reemployment prospects of unemployed workers (**?**). The intervention in the JOBS II RCT was to teach unemployed workers skills related to searching for employment, such as how to prepare job applications and resumes, successfully interview, contact potential employers, and use social networks to obtain job leads. An additional focus of the intervention included activities to enhance self-esteem, increase a sense of self-control, and cope with set-backs.

In the RCT, a large sample of unemployed workers first underwent a screening process. After screening, the researchers randomly assigned 1249 to the job search seminar (treatment) and 552 to a short pamphlet on job search strategies (control). Workers assigned to treatment attended a 20-hour job search seminar over one week. Follow-up interviews on all workers were conducted 6 weeks, 6 months, and 2 years after the intervention. Researchers focused on a number of employment related outcomes including employment status and monthly earnings. After the trial, a number of subjects in both the treated and control remained unemployed and thus had no earnings, producing a point mass at zero in the outcome measure.

Another well known job training intervention is the National Supported Work Demonstration (NSW). In the mid-1970's, the Manpower Demonstration Research Corperation (MDRC) operated the NSW program in ten sites across the United States (**?**). The NSW was designed to move unemployed workers into the job market by giving them work experience and training. Those assigned to the treatment in the NSW were guaranteed employment for 9 to 18 months. However, wages were at below market rates. Once the program expired, trainees were expected to find work in the labor market. Participants assigned to the control condition were not given any aid in finding a job. Participation into treatment and control groups was randomized. One primary endpoint was earnings 27 months after participants left the NSW program.

Figure 1 contains density plots for the subjects with nonzero earnings from both RCTs stratified by treatment status. In JOBS II, 21% of subjects reported zero earnings, and in the NSW RCT, 31% of subjects reported zero earnings. Even after we remove point masses at zero, the earnings



(a) JOBS II (b) NSW

Figure 1: Nonzero earnings for participants in the two job training interventions. The dotted lines represent average earnings by treatment condition in each experiment. In JOBS II, 21% of subjects reported zero earnings. In NSW, 31% of subjects reported zero earnings.

distribution for both treated and control groups clearly depart from Normality, and we may wish to use statistical methods tailored to these distributions.

2 A Review of Randomization Inference and the Tobit Model of Effects

First, we outline randomization inference method with a particular emphasis on the Wilcoxon rank sum test, since existing methods for outcomes with clumping at zero use this test statistic.

2.1 Notation

In our setup, there are n subjects, $i = 1, \ldots, n$. Of the n subjects, m are selected at random to receive treatment, and the remaining $n - m$ are assigned to control. For each subject i observed in the experiment, the indicator $Z_i = z \in \{0, 1\}$, records then randomly assigned treatment status.

Subject i has two potential outcomes: the outcome r_{Ti} that would be observed if i were assigned to treatment and the outcome r_{Ci} that would be observed if i were assigned to control (**??**). Combining potential responses with the treatment indicator, we define the observed outcome for each unit i: $R_i = Z_i r_{Ti} + (1 - Z_i) r_{Ci}$. Finally, we define $\mathbf{Z} = (Z_1 \dots Z_N)^T$, $\mathbf{Y} = (Y_1 \dots Y_N)^T$, $\mathbf{Y_C} = (Y_{01} \dots Y_{0n})^T$ The causal effect is defined as the counterfactual contrast of $r_{Ti} - r_{Ci}$. Unfortunately, this contrast cannot be observed in the data, since for each subject we only observe one potential outcome as revealed by treatment assignment. This is the fundamental problem of causal inference (**?**).

Under randomization inference, the only stochastic quantity is the assignment of treatment (**?**). The potential outcomes r_{Ti}, r_{Ci} are fixed features of the finite population of n subjects. Randomization creates the distribution used for inference and forms what Fisher termed "the reasoned basis for inference." The observed response, R_i varies with treatment assignment and thus is not fixed. Under this view of inference, parametric distributions such as the t-distribution or Normal distribution are approximations to the randomization distribution and are not models for data.

2.2 Randomization Test of No Effect

With randomization inference, we test the sharp null hypothesis of no individual treatment effect:

$$H_0 : r_{Ti} = r_{Ci} \text{ for } i = 1, \dots n.$$

To test the sharp null, we generate the randomization distribution of a test statistic given the null, and compare the observed value of our test statistic to this distribution. We write the test statistic as $t(\mathbf{Z}, \mathbf{Y})$, a function of the observed outcomes and the treatment assignment. One common such statistic is Wilcoxon's rank sum statistic in which the responses $\mathbf{Y}$ are ranked from 1 to n with average ranks for ties, and the ranks of the treated $(Z_i = 1)$ units are then summed (**?**). Large values of this statistic suggest that r_{Ti} is generally larger than r_{Ci}. The null hypothesis

is then tested by computing $t(\mathbf{Z}, \mathbf{Y})$ and asking whether it falls in the tail of the randomization distribution.

Under H_0, the only randomness is from $\mathbf{Z}$, and we know its distribution. Therefore, we only need to calculate the probability of observing a value of $t(\mathbf{Z}, \mathbf{Y})$ equal to or greater than T, some critical value. The p-value for the test of the sharp null is therefore calculated as follows:

$$\Pr(t(\mathbf{Z}, \mathbf{Y}) \geq T) = \sum_{\mathbf{z} \in \Omega} \mathbf{1}\left\{t(\mathbf{z}, \mathbf{Y}) \geq T\right\} \cdot \Pr(\mathbf{Z} = \mathbf{z}) \tag{1}$$

with Ω being the set of allowed random assignment vectors. For the rank sum statistic without ties, the randomization distribution is the random sum of m numbers randomly selected from $\{1, \ldots, n\}$.

2.3 Confidence Intervals and Point Estimate for Constant Treatment Effects

Both interval and point estimation require two assumptions that are unnecessary for the test of no effect. First, we assume that the stable unit value treatment assumption (SUTVA) holds (**?**). Second, we must make an assumption about the nature of response to the treatment. **?** refers to this second assumption as a "model of effects." A model of effects ties the potential outcomes of a unit together. Under this model, if we observe r_{Ti} we can impute r_{Ci}. First, we use the more common model of a constant response to treatment to review basic principles of interval and point estimation in the randomization inference framework. Later, we use alternative models of effects to account for clumping at zero.

If the treatment effect is constant then $Y_{i1} = Y_{i0} + \tau$ for every $i = 1, 2, \cdots, N$ for some τ. This model of effects implies that $Y_{i0} = R_i - Z_i\tau$. The quantity $R_i - Z_i\tau$ is sometimes referred to as the "adjusted response." and can be calculated for all units with observed data, given tau. Testing the null hypothesis $H_\tau : \tau = \tau_0$ is the same as testing that $\mathbf{Y} - \mathbf{Z}\tau_0$ satisfies the original

sharp null hypothesis of no treatment effect.

Once we have an exact test for any specific τ, the set of τ not rejected by a hypothesis test of level α form a valid $1-\alpha$ confidence interval. Generating confidence intervals in this way is called inverting a test. For example, in the case of an additive constant treatment effect, we construct a 95% confidence interval by testing $H_0 : \tau = \tau_0$ for all possible values of τ_0, and keeping the values of τ_0 that are not rejected at the 5% level.

We can generate point estimates via the method of Hodges-Lehmann. The Hodges-Lehmann estimate, in general, asks the following question: what is the value of the treatment effect such that the test statistic equals its expectation under the null hypothesis of no treatment effect? Under the model of constant-additive effects, the Hodges-Lehmann point estimator for τ is the value of $\hat{\tau}$ such that the adjusted responses, $R_i - \hat{\tau} Z_i$, are exactly without treatment effect. The test-statistic under the null is $t(\mathbf{Z}, \mathbf{Y} - \mathbf{Z}\tau_0)$. We define t_0 as the expectation of the test-statistic under $H_0 : \tau = \tau_0$. The Hodges-Lehmann estimator is the solution to the equation:

$$t_0 = t(\mathbf{Z}, \mathbf{Y} - \mathbf{Z}\widehat{\tau})$$

In practice, $t(\mathbf{Z}, \mathbf{Y} - \mathbf{Z}\widehat{\tau})$ may vary in discrete jumps, so that either no solution or infinitely many solutions may exist. When this occurs, the estimator is computed as:

$$\widehat{\tau} = \frac{\inf\{\tau : t_0 > t(\mathbf{Z}, \mathbf{Y} - \mathbf{Z}\tau)\} + \sup\{\tau : t_0 < t(\mathbf{Z}, \mathbf{Y} - \mathbf{Z}\tau)\}}{2}.$$

As a variant of Hodges-Lehmann, we, instead of solving for the expectation, simply take the value of τ that maximizes the p-value as our point estimate.

In the randomization inference framework, interval and point estimates depend on the model of effects. For outcome measures such as earnings, a constant additive model of effects is unrealistic since under this model negative earnings are possible for those in the treatment group. Next, we demonstrate how we might account for clumping at zero via a ore appropriate model of

effects.

2.4 The Tobit Model of Effects

?, pg. 48 extends the classic Tobit model of effects to a randomization framework with no normality assumption. The Tobit model assumes a latent pre-treatment worth w_i. Under Rosenbaum's Tobit model of effects, τ, the treatment effect, is a constant shift on this latent worth. We do not directly observe this w_i, however, but instead observe 0 if the worth is negative. Under this model we would then have potential outcomes for unit i of $r_{Ci} = \max(0, w_i)$ and $r_{Ti} = \max(0, w_i + \tau)$. Under this model a constant treatment effect on the latent worth will lead to a change in the point mass at 0 as well as a shift in the non-zero distribution.

Given a specific model of effects we can, for any observed r_{Ti} impute the corresponding r_{Ci} (assuming $\tau \geq 0$). We cannot, however, impute r_{Ti} from r_{Ci} in all cases: if $r_{Ci} > 0$ we know r_{Ti} but for $r_{Ci} = 0$ we only know $r_{Ti} \leq \tau$. Nevertheless, for any valid model of effects, under the null, we can impute all control outcomes, and we can conduct valid hypotheses tests for the adjusted outcomes of $\max(R_i - \tau Z_i, 0)$. If tau is correct, and the model of effects is correct, these quantities will be the same regardless of treatment assignment. Given this fact, estimation of a point estimate and confidence interval for τ then proceeds under the methods outlined above.

? uses the Tobit model of effects in conjunction with a rank based test statistic. Specifically, one applies the Tobit model of effects to the observed data to compute adjusted responses, and then applies a rank based test statistic to these adjusted responses. If the adjustment is correct then the treatment and control units should have the same distribution of ranks.

3 Extending the Tobit Model of Effects

3.1 Test Statistics for Clumping at Zero

Here, we propose an alternative set of test statistics that may have greater power than the Wilcoxon rank sum test for specific alternative hypotheses. If the Tobit model of effects holds, two features should be apparent in the distribution of outcomes in the treatment and control arms. First, if the shift of τ is correct, proportion of zeros in the adjusted responses of the treatment group should match the proportion of zeros in the control group. Second, under the correct value for τ, the distribution of adjusted non-zero outcomes in the treated group should have the same shape as the non-zero outcomes in the control group. These two observations motivate different types of test statistics. We then combine these two types of test statistics to form an omnibus test to produce a more powerful test.

The first test statistic we propose is the difference in the proportion of zeros across treated and control arms:

$$\hat{p} = \widehat{p}_C - \widehat{p}_T$$

This test statistic will tend to reject if the adjusted treatment group and the control group have different proportions of zeros, but will be insensitive to any differences in the non-zero outcomes.

The next test statistic can be any measure of difference across the treated and control units that have non-zero outcomes. For example, one test we apply in the application section is the Kolmogorov–Smirnov (KS) test statistic (**?**). Formally, the KS test statistic is defined as follows: let $\widehat{F}_C$ be the empirical CDF of the positive outcomes under the control condition and let $\widehat{F}_T$ be the same for the treatment. Using these quantities, the KS test statistic on the positive outcomes is

$$KS^+ = \max_{y>0} \left| \widehat{F}_T(y) - \widehat{F}_C(y) \right|.$$

We can also look at specific differences in the distributions rather than this overall approach. For example we could focus on the difference in variances, possibly measured as

$$\hat{\sigma} = \hat{\sigma}_C / \hat{\sigma}_T,$$

the ratio of the variance of the control group outcomes and the treatment group outcomes. Alternatively, we might take as our test statistic the difference in medians.

For any test statistic, we can generate a confidence interval by inverting a sequence of tests. Denote our test statistic of interest t_k. The α-level confidence interval based on t_k is then constructed by testing $H_0 : \tau = \tau_0$ for all possible values of τ_0, and keeping the values of τ_0 that are not rejected by an α-level test using our test statistic t_k. We then have, for t_k, the corresponding confidence interval

$$CI_k \equiv \{\tau_0 : p_k(\tau_0) > \alpha\}$$

where $p_k(\tau_0)$ is the p-value obtained by testing $H_0 : \tau = \tau_0$ with the test statistic t_k. One should think of these intervals as the set of all τ where the predicted control outcomes in the treatment group and the actual control outcomes in the control group are similar as measured by the given test statistic. So, for example, if we used the proportion test, our confidence interval would consist of all possible shifts such that the proportion of zeros is similar. Different test statistics can give different confidence intervals. In particular, depending on the character of the data, we might imagine different test statistics giving shorter intervals, in general, than other test statistics.

Given the Tobit model of effects, all confidence intervals are valid. This is easy to show: for a given interval CI_k, take τ_0 as the truth, i.e., that the null is true for this value of τ. Then the probability that τ_0 is not included in CI_k is the probability that $p_k(\tau_0) \leq \alpha$ which, as these tests are all of correct size, is no more than α.

For the analysis to be fully principled, the test statistic or set of test statistics should be selected as part of the analysis plan before any outcome data are collected. However, if the RCT is to be repeated, an exploratory analysis could be conducted using outcomes from the first trial. This exploratory analysis could then inform the analysis plan for a later trial.

3.2 Goodness of Fit Tests

Some test statistics allow us to test whether the Tobit model of effects is a reasonable fit to the data. Specifically, we can test:

$$H_0^m : \exists\ \tau \text{ s.t. } r_{Ci} = \max(0, r_{Ti} - \tau) \text{ for } i = 1, \ldots, n$$

To do this we generate a confidence interval for τ using a given test statistic. If this test rejects for all τ, i.e., the confidence is the empty set, then we conclude that no shift is appropriate and reject H_0^m. Some test statistics have no power to detect violations of the model, while others do. In particular, there is always a τ such that the corresponding Wilcoxon rank statistic on the adjusted responses equals its expected value. Similarly, the test of an equal proportion of zeros will also have no power to detect violations of the Tobit model of effects, as we can always find a shift to make the proportions of zeros the same. The KS statistic, however, can detect such violations due to its overall focus on shape.

To prove that this goodness of fit test is valid, consider the null. If it were true, then there exists some τ, say τ_0 such that $r_{Ci} = \max(0, r_{Ti} - \tau_0)$ for $i = 1, \ldots, n$. When testing τ_0, we will reject (erroneously) with probability no greater than α, regardless of choice of test statistic. Therefore, our confidence interval will be non-empty with probability greater than or equal to $1 - \alpha$. Since we only reject H_0^m if the confidence interval is empty, we reject with probability no greater than α.

When using these goodness of fit tests, we advocate being conservative. In these circumstances, this means rejecting the goodness of fit test at a higher α, e.g., $\alpha = 0.10$, or calculating confidence

intervals at 90% confidence to assess if they are empty. Goodness of fit tests are known to have low power, in general, but valid confidence intervals under permutation style approaches require correctly specified models. It is therefore important to assess model fit in the process of generating confidence intervals, and one should be wary of even weak evidence against a model of effects specification.

3.3 An Omnibus Test

Above, we suggested several different test statistics when there is clumping at zero. It is natural to jointly use multiple test statistics to take these different consequences into account (**?**). One straightforward method for doing this is to create a test based on a combination of the different test statistics. We outline this omnibus approach next.

To form an omnibus statistic, first take a vector of test statistics, $t_1, \ldots, t_K$, that correspond to K different consequences of our modeling assumptions, with an observed vector of $t^{obs} = (t_1^{obs}, \ldots, t_K^{obs})$. We will combine these statistics to form an omnibus statistic to give an omnibus test.

Consider the case where τ is known and the null is true. We then know all the control outcomes for all our units. In this circumstance, we could obtain the joint null distribution of $t_1, \ldots, t_K$ by, for each possible randomization, calculating $(t_1, \ldots, t_K)$. Call these permutation tuples $t^{(j)*} = (t_1^{(j)*}, \ldots, t_K^{(j)*})$ with $j = 1, \ldots, R$ where R is the number of possible permutations under randomization. In practice, we likely examine only a subset of all possible permutations. For each $t_k^{(j)*}$, we calculate its marginal p-value by calculating

$$p_k^{(j)*} = \mathbf{P}\Big\{t_k^{(j)*} \geq T_k^*\Big\} = \frac{1}{R}\sum_{r=1}^{R} \mathbf{1}\{t_k^{(j)*} \geq t_k^{(r)*}\}.$$

This gives $p^{(j)*} = (p_1^{(j)*}, \ldots, p_K^{(j)*})$. Similarly, we can calculate marginal p-values for t^{obs} by, for each k, calculating $p_k^{obs} = \mathbf{P}\Big\{t_k^{obs} \geq t_k^*\Big\}$. This gives $p^{obs} = (p_1^{obs}, \ldots, p_K^{obs})$.

We now have the joint null distribution of K p-values corresponding to K tests. We also have the observed marginal p-values for our observed data and random assignment. We treat these p-values as test statistics in their own right and calculate an omnibus test statistic such as

$$t_p = \min\{p_1, \ldots, p_K\}$$

or

$$t_p = \sum_{k=1}^{K} \log p_k.$$

The latter is equivalent to the product of the p-values. If we calculate t_p for all the vectors $p^{(j)*}$ we have the null distribution of t_p. We then calculate t_p^{obs} and compare it to this null distribution. Our final p-value is then

$$p = \mathbf{P}\Big\{t_p^{obs} \leq T_p\Big\} = \frac{1}{R}\sum_{r=1}^{R} \mathbf{1}\{t_p^{obs} \leq t_p^{*(r)}\}.$$

By construction, lower values are more extreme since low p-values are indicate greater evidence against the null.

As shown above, the formation of omnibus test statistics, we can combine p-values by either taking the minimum or taking the product. The choice of minimization versus products will have different power against different alternatives. The minimum will have power against alternatives where only one of the primary tests has power. The product will have power against alternatives where all the primary tests have moderate power.

One can also form omnibus test statistics directly, such as taking the sum of the initial t_k; this can be difficult, however, if the t_k are all on different scales. Converting them to p-values first makes them directly comparable. Finally, one might weight one test more heavily than another, such as by taking a weighted sum of the log p-values. Any of these choices provide different tests which will have different power for various alternatives. They will all, however, be valid tests.

For the Tobit model of effects, we might set $K = 2$, $t_1 = \hat{p}$, and $t_2 = KS$. This omnibus test allows us to test whether the proportions of zeroes in the shifted treatment arm is the same as compared to control, and also for whether the shifted treatment arm has the same empirical CDF in the nonzero mass of the distribution as the control arm. Thus we only consider possible τ that align both these quantities, which increases the power to detect departures from the Tobit model such as when we are testing the wrong shift τ.

We note that one could also use the more general KS statistic (denoted KS in our paper) which compares the CDFs of all the outcomes including the point mass at 0. In principle this statistic should be sensitive to the difference in 0s as well as the shape of the positive outcomes. However, this statistic tends to have poor behavior when applied to distributions with point masses and ties.

3.4 Two Simulations

We next illustrate the concepts from above by conducting two simulations: one where the Tobit model is correctly specified, and one where it is not. Before each simulation we illustrate how to generate confidence intervals and point estimates using a single simulated dataset. We then replicate the simulation to illustrate how different testing procedures behave over repeated trials.

3.4.1 Simulation 1: The Tobit Model is correct

In our first simulation, the Tobit model is correctly specified. For each simulation trial, we take 1000 units, each with a latent w drawn from a $N(1, 5^2)$ distribution and a true Tobit shift of 2. We randomize the units to treatment and control and observe their outcomes.

We next generate a confidence interval for τ using four different test statistics. The first, W, is based on on the Wilcoxon rank sum test for the Tobit model as devised by Rosenbaum. The second, $\hat{p}$, is the difference in the proportion of zeros in the treated and control groups, while

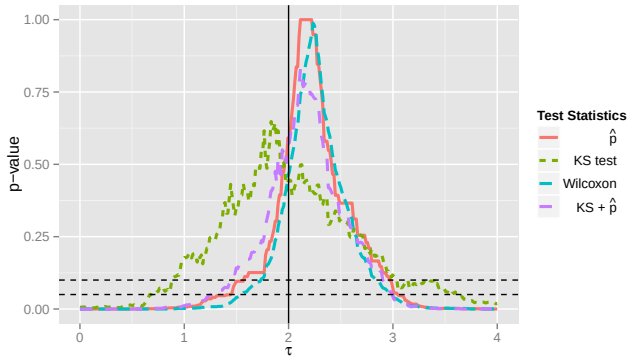


Figure 2: p-values for different possible values of τ under a Tobit model for a single illustrative, synthetic data set. Each line corresponds to the different test statistics of $\hat{p}$, KS, W, and the omnibus test formed from $\hat{p}$ and KS. The true $\tau = 2$ is marked with the solid vertical line. The corresponding 95% confidence intervals for specific test statistics would be the range of τ where the p-value is above the 0.05 line.

the third, KS, is the KS test applied to the nonzero outcomes. Finally, we include an omnibus statistic by adding the log p-values for $\hat{p}$ and KS.

For a sequence of possible τ, we calculate p-values against the null of that τ for each of the four statistics. For each considered τ, the permutation p-values are calculated using 1000 permutations of the treatment vector. We plot the p-values against the values of τ for each test in Figure 2. Plots of p-values like these serve as a useful visual diagnostic for whether a test statistic is behaving as we expect. Also, as we demonstrate later, such plots can usefully reveal when the model of effects is misspecified. We see that, for example, all four tests have very low p-values for the null of $\tau = 4$, so all four tests reject that τ could be 4. Corresponding confidence intervals are shown in Table 1. These are formed by taking all τ that are not rejected. That is, the support of the portion of the curves lying above the horizontal 0.05 line on Figure 2 form our intervals.

Table 1: Confidence intervals and Hodges-Lehmann estimates for a simulated dataset from different test statistics ordered by interval length.

Test Statistic	Lower CI	Upper CI	Hodges-Lehmman Point Est	Length
KS	0.67	3.60	1.83	2.93
$KS + \hat{p}$	1.27	2.98	2.11	1.71
$\hat{p}$	1.46	3.04	2.11	1.58
W	1.55	2.98	2.23	1.43

In Table 1 we observe that, for this particular randomization on this particular dataset, the omnibus test returns a wider interval than the Wilcoxon or difference in the proportion of zeros. This does not necessarily have to occur; since the p-value depends on the joint distribution, its behavior is complex and difficult to characterize. That being said, in general if one or the other of its components has a low p-value, the omnibus will be low. It therefore tends to be low overall when the component curves are not aligned, unlike in this figure, which allows for goodness of fit tests as discussed below.

We can also obtain point estimates by taking the τ with the highest p-value (if multiple τ have the same p-value we take the median of these τ). This is related to the Hodges-Lehmann point estimation discussed above. These are the peaks of the curves on Figure 2 and are also listed on Table 1. In our single example, the point estimates are close to the true value of 2.

We next simulated 100 datasets, and generated the four confidence intervals for each set. The data generation is as described above. Figure 3 shows the distribution of confidence interval lengths for the four different approaches and one additional test statistic: the KS test applied to the entire outcome distribution instead of just the nonzero portion of the distribution. We see that the Wilcoxon test statistic tends to result in the shortest intervals, with an average length of 1.32. The omnibus test has a mean length of 1.50, 14% longer. KS^+, which conditions on the positive units only, has no ability to detect differential rates of 0 outcomes and thus has little power; it therefore does poorly. The test statistic based on the shift in zeros also has poor performance y because it does not exploit anything about the shape of the positive outcomes. The KS test applied to all of the outcome data performs almost as well as the omnibus test, but apparently a sharper focus on the proportion of zeros, which the omnibus test has, is important here. Unfortunately, as the power of each test depends on the underlying distribution, we cannot extend these relationships to other circumstances.

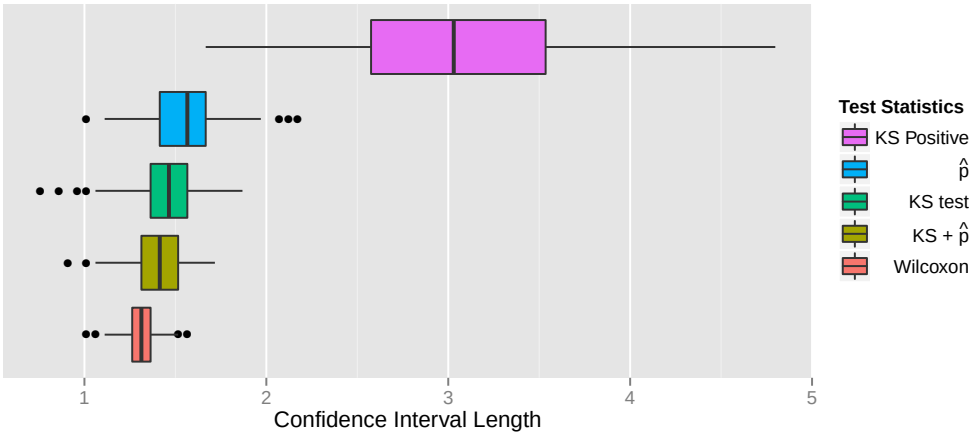


Figure 3: Length of confidence intervals for 100 simulated data sets (with a true Tobit model of effects) for five different test statistics.

3.4.2 Simulation 2: The Tobit Model is misspecified

In our second simulation, we investigate the performance of different test statistics when the Tobit model of effects does not hold. We again simulate the latent w, but now treatment scales w as well as shifts it by 2. In particular, we generate the data under the following misspecified Tobit model of effects:

$$r_T = \max(1.3w + 2, 0).$$

We generated data according to this model and then applied the same set of test statistics to the observed outcomes to generate confidence intervals for all of our statistics. Here, if a confidence interval is the empty set, we reject the overall null that the Tobit model of effects fits the data.

Figure 4, analogous to Figure 2, contains the p-values for a range of τ for each test. First, we note how each of the individual test statistics produce non-empty confidence intervals, erroneously suggesting the Tobit model of effect correctly characterizes response to treatment. However, the omnibus test based on the KS test and the test of the proportion of zeros does not achieve high p-values for any τ, giving an empty set for the confidence interval. The omnibus test detects our misspecification of the Tobit model of effects. Thus while the omnibus test may have lower power

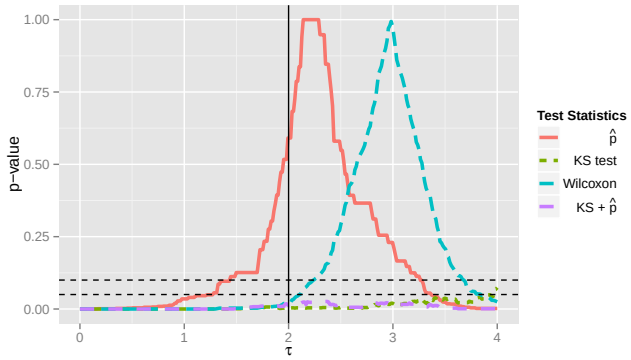


Figure 4: p-values for different possible values of τ under a misspecified model. See Figure 2 for annotation details. Only the omnibus test statistic produces an empty confidence interval.

relative to the more conventional Wilcoxon test statistic, it can clearly provide useful information about the overall suitability of the Tobit model of effects.

We replicated this simulation 200 times to estimate the power of the test using single test statistics and three different omnibus test statistics. Specifically, we combined the $\hat{p}$ test statistic with the KS test, the KS test applied to only the positive part of the response distribution, and the Wilcoxon rank sum statistic. Table 2 contains the results of the simulation.

All the non-omnibus test statistics have very low power to detect misspecification. In general, if there always is some value of the shift that can make the test statistic small, the power to detect misspecification will be reduced to 0. There is always a value of τ to make the Wilcoxon statistic equal to its expected value, for example. The KS test statistic has some power because it only has a low value if the adjusted treatment group has the same CDF as the control group. It is attending to the entire shapes of the two distributions.

The simulation exercise also reveals that the use of an omnibus test statistic is no panacea. The omnibus test based on a combination of $\hat{p}$ and W still have no power to detect poor fit of the Tobit model of effects. A combination of $\hat{p}$ and the KS test is an improvement, although the results are less than acceptable. For this data generation process, the KS statistic on the non-negative outcomes combined with the proportions has the highest power of about 96%. However, this omnibus test statistic did not give the smallest confidence intervals when the model is correctly

specified, as demonstrated above. Therefore, we face a a trade-off between using a test that has no power to detect departures from the modeling assumptions and the expected length of the final confidence interval.

Table 2: Estimates of power, along with uncertainty intervals in brackets, for different test statistics for Tobit model fit. Uncertainty intervals are 95% intervals based on the 200 results of Monte Carlo simulation.

Test Statistic	Est.	Power (%)
$\hat{p}$	0	[0-2]
KS	36	[29-43]
W	0	[0-2]
KS Positive Support	38	[31-45]
Omnibus Tests		
$\hat{p} + KS$	50	[43-57]
$\hat{p} + KS$ Positive Support	96	[92-98]
$\hat{p} + W$	0	[0-2]

4 The Multi-Tobit Model

While the Tobit model of effects outlined above may be an adequate model for many outcomes with zero clumping, we may wish avoid the assumption of a constant shift due to treatment and fit something that is more flexible. In the randomization inference framework, we can introduce a more complex model of effects parameterized by more than a single parameter. Specifically, we conceive of the response to treatment occurring in two stages: a shifting up of one's latent worth by a constant effect, and then a scaling of one's latent worth if it is positive. Specifically, for an individual with latent worth w_i:

$$r_{Ci} = \max(0, w_i) \text{ as before, and}$$

$$r_{Ti} = \max(0, \beta\left(w_i + \tau\right))$$

This additional flexibility allows for the treatment group to have a longer tail than the control group. The order of operations is unimportant here: this model of effects is equivalent to a scale

followed by a shift, with $\beta' = \beta$ and $\tau' = \tau\beta$. We call this model of effects the Multi-Tobit model.

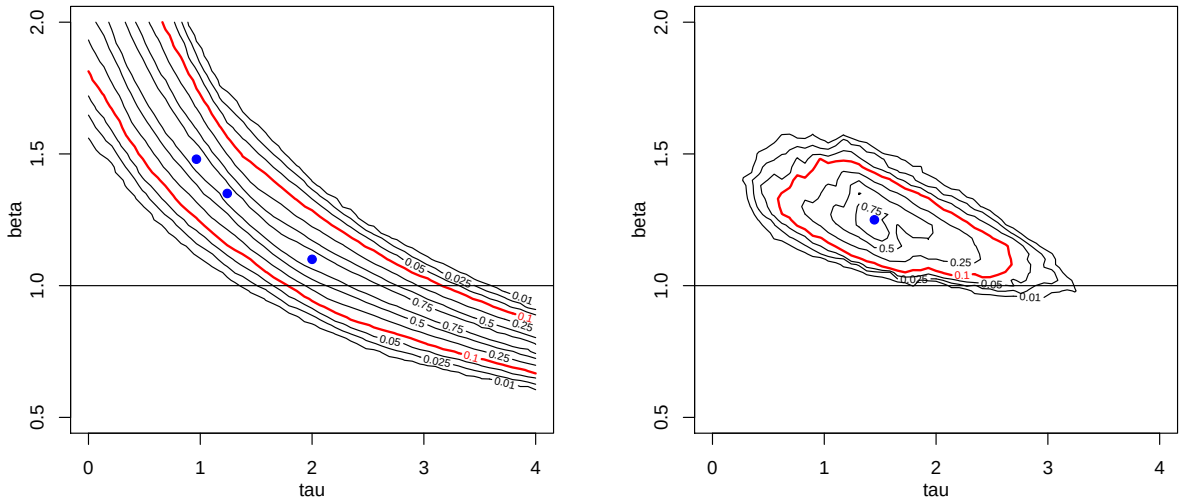
4.1 Testing and estimation

Under this model we can still test the sharp null of no effect under the usual randomization inference framework using permutations of the data and repeatedly applying the test statistic. For this model of effects, $(\tau, \beta) = (0, 1)$ corresponds to no treatment effect. Moreover, $\tau = 0$ and $\beta > 1$ corresponds to a treatment effect that only impacts those who would otherwise have positive outcomes. The test of the sharp null under this model of effects has identical properties and behavior to the Tobit model of effects from above because randomization tests do not depend on the alternative hypothesis. The additional flexibility of this model is apparent in the estimation of treatment effects.

Since this model of effects has two parameters, we produce a confidence region defined by all pairs of values that do not reject the null. Unfortunately, classic tests such as the Wilcoxon rank sum can have perverse behavior for interval estimation since selecting extreme values of τ and β can align a treatment and control distribution's median rank, and thus all pairs of such values will appear in an associated confidence region. To rectify this problem, we need statistics that measure differences in at least two ways, such as both the center and spread of the distribution. The KS test does this by measuring differences of the entire cumulative distribution between the adjusted treatment and control units. Another alternative would be to measure the alignment of the two distributions at two pre-specified quantiles, or by combining multiple tests with an omnibus test as described above.

4.2 Illustration

We illustrate the generation of confidence intervals for the Multi-Tobit model of effects using the same simulated data from the misspecified Tobit model simulation above. In this simulation,



(a) Wilcoxon test statistic

(b) Omnibus ($\hat{p}$ and KS) test statistic

Figure 5: Confidence intervals created by different statistics under the misspecified Tobit model of effects.

we set $\beta = 1.3$ and $\tau = 1.54$. Results from this simulation are shown in Figure 5. Panel 5a demonstrates how using the Wilcoxon rank sum test produces perverse confidence regions. The very small τ coupled with very large β, for example, allow for the overall mean ranks to be aligned, causing the Wilcoxon rank sum test to not reject. On the other hand, if we use statistics that attend to the relative shapes of the two distributions, we can get reasonable regions such as illustrated in Panel 5b. For the results in Panel 5b, we applied the omnibus statistic that tests both the proportion of zeros with $\hat{p}$ and the overall shape of the nonzero part of the response distribution with KS. For this test statistics, the confidence region is relatively compact. Since the confidence region does not include $\beta = 1$ (marked with the horizontal line), we can reject the Tobit model as a reasonable fit.

One drawback of the Multi-Tobit model is that confidence regions of this nature are difficult to interpret. To improve interpretability, it is often worth projecting them into a more interpretable space. Each point in the confidence region, coupled with the observed data, gives full knowledge of the control potential outcomes of all treated units. This means that, for each such point, we could calculate quantities such as average treatment on the treated, the proportion of units

who would be zero under control, or the average shift for treated units with a counterfactual positive outcome. We then take the set of the projected points as a confidence set on the chosen quantity.

5 Applications

We now analyze data from the JOBS II intervention and the NSW study. As we noted earlier, the earnings distributions from both interventions have non-trivial amounts of clumping at zero earnings. We treat this analysis as exploratory in that we fit a number of different tests, primarily to illustrate how these approaches work. Our approach might be viewed as "fishing" in an actual trial. In an actual trial, the analysts would specify which tests would be conducted through the use of an analysis plan before any data are collected.

5.1 The JOBS II Intervention

Let's change the entire analysis in these sections to 90% confidence instead of swapping back and forth?—Luke M

As an initial step, we test for presence of any treatment effect by using the classic Wilcoxon rank test. We are able to reject the sharp null hypothesis ($p = 0.016$). The treatment had some impact. We turn next to assessing the degree of that impact.

By inverting this same test under the assumption of a constant effect assumption, we find that subjects in the treated condition had earnings that were \$88 higher with a 95% confidence interval of $[\$0.01, \$260]$. However, this model of effects is implausible, since it implies that some treated units would have received negative earnings under the control condition. We therefore apply the Tobit model of effects with a Wilcoxon test statistic as recommended in **?**, pg. 48. Recall that the Tobit model of effects implies that the JOBS II intervention raises the marginal value of every worker's labor by the same constant, but that workers are employed only if the marginal value of their labor is positive. Under the assumption of this Tobit model of effects the point

estimate is $\hat{\tau} = \$216$ and the 95% confidence interval for τ is $[\$42, \$394]$. The Wilcoxon statistic gives a fairly wide confidence interval since a wide range of shifts still provides a relatively equal distribution of ranks between the treated and control groups in large part due to the low density of observations in the low hundreds. That is, shifting the treated distribution by a small τ does not change many ranks.

We now use the methods proposed above to see if we can glean any deeper insights into the effectiveness of the JOBS II intervention. First, we apply the test statistic $\hat{p}$ to examine whether there is a difference in the proportion of zeros across the treated and control groups. The point estimate is 6% with a 95% confidence interval of 1% to 11% percentage points, with a p-value against no difference of 0.01, using a simple binomial test of proportions. This test indicates that the JOBS II intervention moved a significant number of workers into employment as they moved from zero to positive earnings. Next, we explore whether whether the intervention increased earnings as well.

As shown in Figure 1, there appears to be a thicker tail of higher incomes in the treatment group. This could be due to a constant shift. To assess this, we applied the KS test to the nonzero values of the earnings distribution, testing a sequence of nulls $\tau_0 \in [0, 400]$. Under this test statistic, $\hat{\tau} = \$41$, and the 95% confidence interval for τ is $[\$0, \$151]$. Under a constant shift model, we have no evidence of an effect when focused on positive earnings with this statistic.

This might signal model misspecficiation, so we turn to a set of omnibus test statistics. While the omnibus test statistics may have wider confidence intervals than the Wilcoxon test statistic, they also have power to detect violations of the Tobit model of effects. We applied three different omnibus tests to the data by combining the $\hat{p}$ test statistic with the KS test statistic, the Wilcoxon test statistic, and a difference in medians statistic. In each case, we calculate the KS, Wilcoxon, and difference in medians test statistics on the non-zero portion of the earnings distribution. We now test at the 90% level to remain sensitive to misspecification. The corresponding p-value curves are on Figure 6.

The tests signal that the Tobit model of effects is a poor fit. F

For the median and KS omnibus tests, the 90% confidence intervals give the empty set, indicating model misfit. The plots of the p-values in Figure 6 are instructive. First, the curves lie below the 0.10 line which is why the corresponding confidence intervals are empty. The range of p-values is generally low and flat along the range of possible value of τ_0. No value of τ_0 makes the adjusted treatment response look like the control group. The omnibus test built from $\hat{p}$ and KS gives a maximum p-value of ≈ 0.05 for $\tau_0 = 33$. The reason for these rejections is the amount of shift necessary to align the proportion of zeros is much larger than what is needed to align the positive distributions. There are no values of τ that satisfy both of these constraints.

The third test, the omnibus Wilcoxon test, has a non-empty confidence interval, but the low shape of the curve on Figure 6 suggests the model of effects is a poor fit; no τ makes the adjusted responses look much like the control responses. If we truly believe the model of effects, then our results are roughly similar to the Wilcoxon test alone with a Hodges-Lehmann point estimate of \$228 with a 95% confidence interval of $[\$176, \$355]$. The key difference is the length of the confidence interval for the omnibus test is nearly half the length of the Wilcoxon test alone. Under our modeling assumption, the omnibus gives a narrow interval, but this is likely due to all the tests signaling model misspecification. This underscores the importance of model checking when generating confidence intervals by test inversion.

In general, the empirical analysis suggests that the JOBS II intervention moved workers from being unemployed to being employed. Overall, there is little evidence that those who would have obtained employment in the absence of the program would have had higher earnings. Since the Tobit model of effects assumes that the treatment raises the marginal value of every worker's labor by the same constant, it fits the data from JOBS II poorly. Moreover, given that there is little evidence that wages were increased by the intervention, we see little reason to attempt to fit the Multi-Tobit model to the JOBS II data. However, this evidence is consistent with the goals of the JOBS II intervention which focused on job-search skills. That is, the treatment was

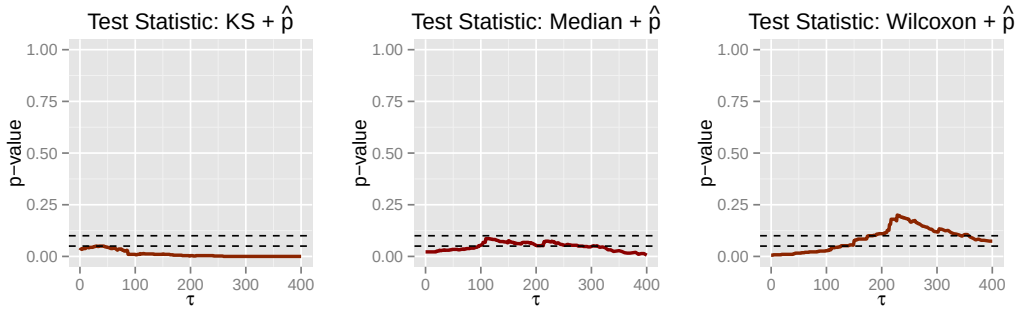


Figure 6: P-value curves for different three omnibus test statistics. Horizontal lines mark the 0.05 and 0.10 α levels. The low curves all suggest that the Tobit model of effects is poor. The two left curves lying below 0.10 indicate actual rejection of the goodness of fit test at the 90% level.

designed to help job-seekers find jobs, but the intervention did not focus on skills that might boost earnings.

5.2 The NSW Intervention

Next, we examine the data from the NSW program. We again begin with a plot of earnings by treatment status for those workers with nonzero earnings. Figure 7 contains density plots for nonozero earnings. In the NSW data, the earnings distribution has a longer and thicker right tail among the treated than the control, even discounting a single treated worker with earnings in excess of $60,000. This suggests that for some treated workers earnings were much higher.

Furthermore, the standard deviation of outcomes among the treated is substantially larger than the in control group ($7867 as compared to $5484). The standard deviation among the workers with positive earnings is also larger ($8042 vs. $5380).[1] This difference in standard deviations might be due to a causal effect of the NSW intervention on those who would have positive earnings regardless of treatment status. Or it could be due to shifting and pulling observations away from 0. We investigate this next.

We first fit a series of test statistics to the data, including a series of omnibus tests, under

[1] Although this is not, strictly speaking, a fair comparison since some workers moved from no earnings are in the treatment group which could inflate the standard deviation.

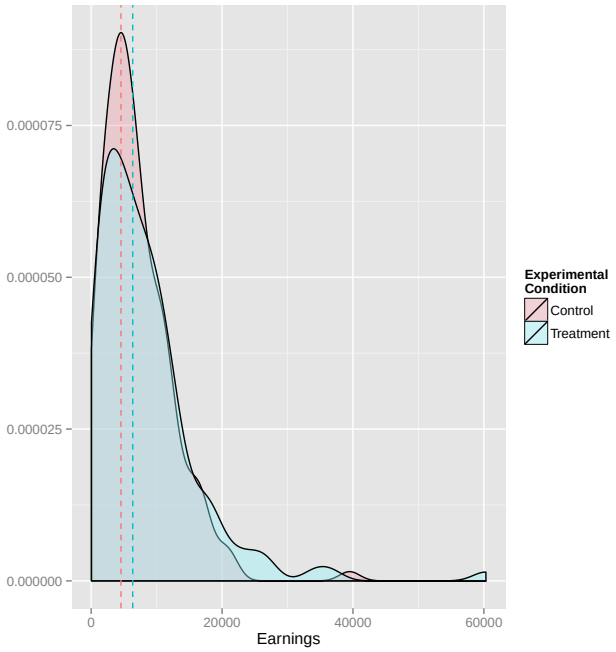


Figure 7: Density of positive earnings for treated and control groups in the NSW program.

the Tobit model. We include a test statistic of the ratio of standard deviations on the positive outcomes to directly assess whether the difference can be solely due to a shift. The corresponding intervals and point estimates are in Table 3. First, using the rank based test statistic, we reject the sharp null ($p = 0.006$). There is an effect. Again, we next address the question of how to characterize the effect.

Assuming the Tobit model of effects is correct we have a series of point estimates and confidence intervals corresponding to the different statistics. The Wilcoxon rank sum test gives a Hodges-Lehmman point estimate of $\hat{\tau} = 1695$ with a 95% confidence interval of $[\$448, \$3326]$. The Wilcoxon statistic gives a fairly wide confidence interval since a wide range of shifts still provides relatively equal distribution of ranks between the groups in large part due to the low density of observations in the low thousands: shifting by a low-valued τ does not change many ranks. Potentially because of this the joint Wilcoxon & proportion test gives a slightly shorter confidence interval than the test based on Wilcoxon rank sum test alone, although it may be the case that the shorter confidence interval signals model misspecification. Using $\hat{p}$ as a test statistic produces

results that are quite similar to those based on the Wilcoxon test statistic. The point estimate is $\hat{\tau} = 1470$ with a 95% confidence interval of $[\$660, \$3180]$. Figure 8 contains a plot of the p-values against values of τ. Both test statistics appear well-behaved, but, as we noted above, neither has any power to detect a failure of the Tobit model of effects.

Table 3: 90% Confidence Intervals from different test statistics ordered by length.

	Point Estimate	95% CI	Length
Single Test Statistics			
$\hat{\sigma}$	2490	[2490, 3585]	1096
W	1695	[600, 3075]	2476
$\hat{p}$	1470	[660, 3180]	2521
KS^+ (Positive Support)	2505	[0, 3990]	3991
Omnibus Test Statistics			
$W + \hat{p}$	1650	[660, 3090]	2431
$KS + \hat{p}$	1650	[495, 3585]	3091

Next, we review results based on the KS statistic applied to the positive support of the earnings distribution. This test statistic has no power to speak of here; it fails to reject for any considered shift. As it never rejects, the confidence interval is actually wider than the reported interval of $[0, 3990]$. The lack of power is so severe that an omnibus statistic using the KS statistic is worse than using $\hat{p}$ alone. Since KS^+ is focused on the positive outcomes only, this hints that the positive outcomes between treatment and control are in fact relatively aligned. An omnibus test statistic composed of $KS + \hat{p}$ produces a point estimate of $\hat{\tau} = 1470$, similar to the omnibus test of the Wilcoxon and $\hat{p}$. This is driven by $\hat{p}$.

We also used the ratio of standard deviations on the positive outcomes as a test statistic, since, as we noted above, the standard deviations were unequal between the groups, potentially more so than what could be accounted for by a Tobit model. This test statistic produces a much larger point estimate of $\hat{\tau} = 2490$ and a 95% confidence interval of $[\$2490, \$3585]$, which is the shortest of any interval. This test statistic, however, suggests model misspecification, as shown on Figure 8: no considered shift gives a high p-value. That is, none of the possible shifts due to

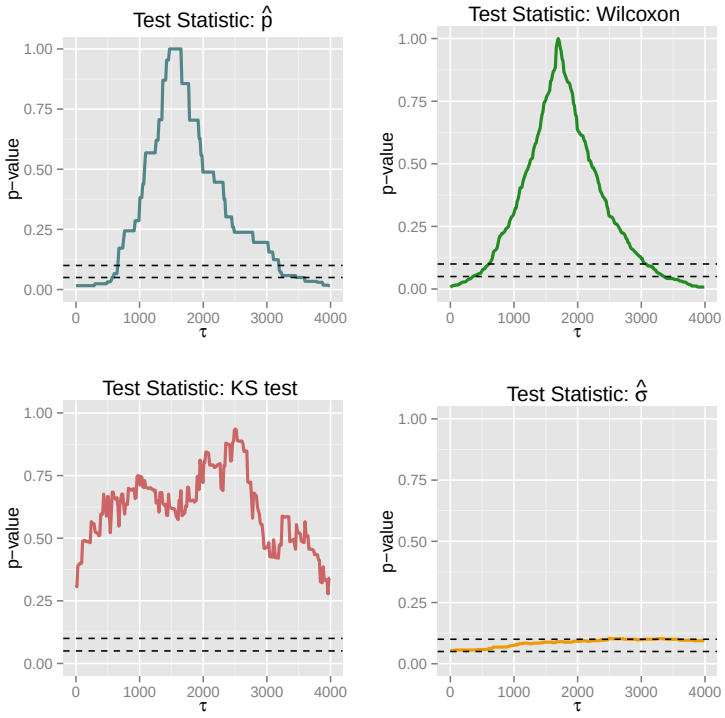


Figure 8: P-value curves for different test statistics. Horizontal lines mark the 0.05 and 0.10 α levels. The $\hat{\sigma}$ statistic suggest model misfit: no constant shift makes the treatment and control groups comparable in terms of spread.

treatment makes the standard deviation of the adjusted treatment group response comparable to the responses of the control group (although implausibly high shifts in the 10000 range would align this statistic). If the Tobit model were true, then some plausible shift should align the standard deviations. However, there appears to be greater variability in the adjusted treatment outcomes than control. We note, however, that it does achieve p-values of $0.12 > 0.10$, which means no official rejection of the goodness-of-fit test. Various omnibus statistics incorporating $\hat{\sigma}$ have maxima of around 0.30. Regardless, we next apply the Multi-Tobit model to see if rescaling the positive outcomes could improve fit.

5.2.1 A Multi-Tobit Model

Recall that the multi-tobit model implies that response to treatment occurs in two stages. First, it assumes that the marginal value of a worker's labor is shifted up by a constant amount τ.

Second, if the marginal value of an employed worker is positive that constant effect is scaled by a second amount β. This allows for greater variability among the earnings in the treated condition. We begin by fitting a multi-tobit model using five individual test statistics: $\hat{p}$, KS, KS^+, the Wilcoxon rank sum test statistic, and $\hat{\sigma}$. Figure 9 shows the fit for these five test statistics using contour plot to show the possible joint values of the two parameters. In all cases we observe large, diffuse regions for the parameters, which suggests low power and motivates omnibus statistics.

For example, the Wilcoxon rank sum test statistic is not sensitive to differences in distributional shape. Due to this insensitivity, there is a relatively large region that allows large values of β to offset low values of τ. This is equivalent to letting large ranks in the treatment group compensate for the low ranks from those with zero earnings. This pattern is even more pronounced when $\hat{p}$ is the test statistic. The KS statistic, with a focus on overall shape of the earnings distribution, produces a tighter confidence region than the Wilcoxon or $\hat{p}$ test statistics.

Omnibus test statistics, unlike single test statistics, may be sensitive to multiple parts of the distribution. Figure 10 shows the confidence regions for four different omnibus test statistics. The confidence regions based on omnibus statistics inherit their general shape from the simpler statistics. For example, the test statistic based on a combination of $\hat{p}$ and the Wilcoxon rank sum test still has a relatively large region that does not allow us to reject the possibility of large values of β with low values of τ. For the omnibus test statistic which is a combination of $\hat{\sigma}$ and the Wilcoxon rank sum test, the confidence region clearly includes $\tau < 0$, suggesting that large scalings with potentially negative shifts are possible. In general, it appears that we are not constraining enough aspects of the two distributions and are thus getting nonsensical fits. Even an omnibus statistic that combines all four individual test statistics, while providing evidence that β is unlikely to be larger than 1.5, does not well determine τ, the shift in earnings.

We fit the Multi-Tobit model because the Tobit model was clearly misspecified. We next examine whether the Multi-Tobit model itself has good fit. To assess model fit, we calculated the Hodges-

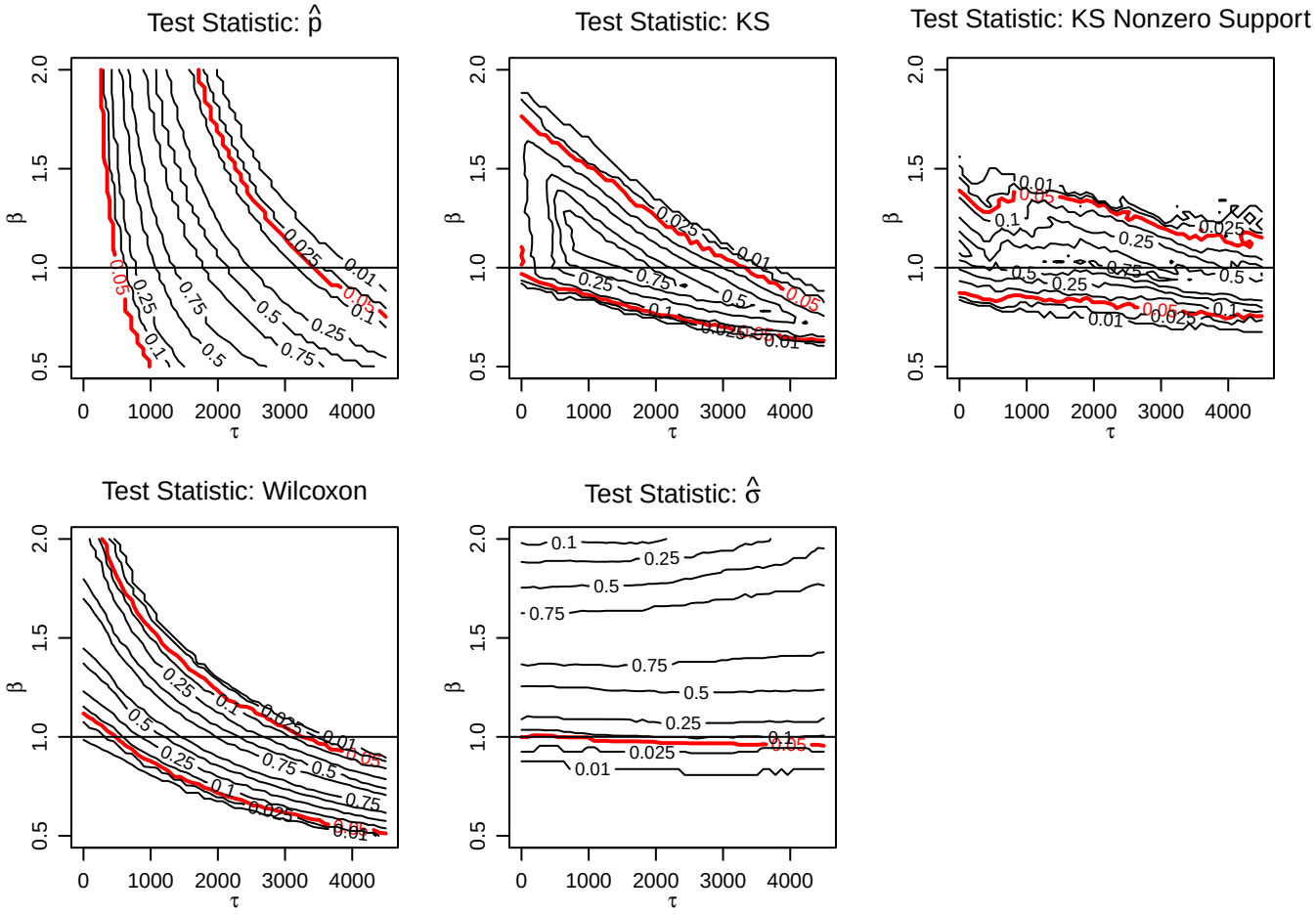
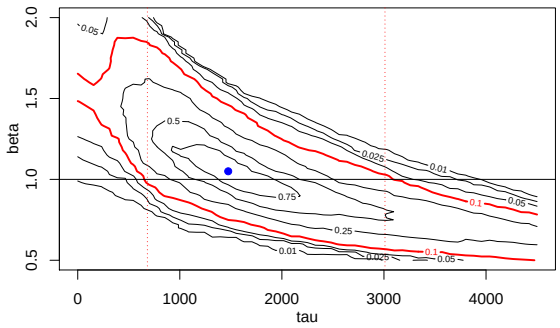
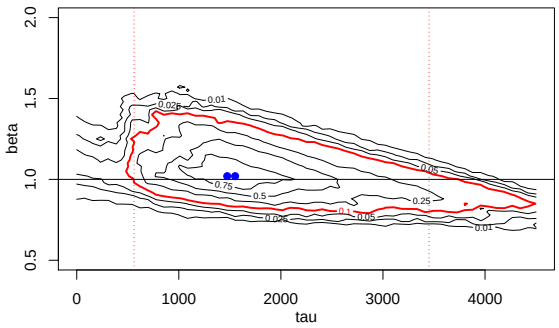


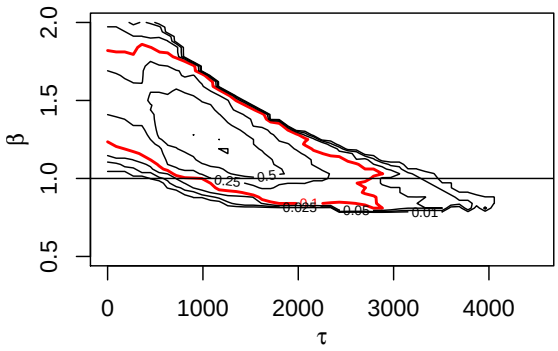
Figure 9: Confidence regions for five test statistics with the Multi-Tobit model of effects. In all cases, confidence regions are wide and diffuse.



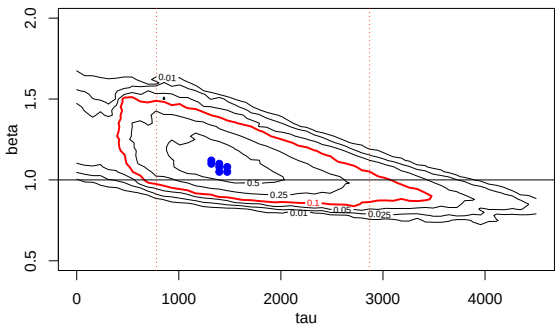
(a) Test Statistic: Wilcoxon+$\hat{p}$



(b) Test Statistic: KS+$\hat{p}$



(c) Test Statistic: Wilcoxon+$\hat{\sigma}$



(d) Test Statistic: Wilcoxon+KS+$\hat{p}+\hat{\sigma}$

Figure 10: Multi-tobit confidence region using the proportion and Wilcoxon omnibus and the proportion and KS omnibus. Vertical lines mark the Tobit confidence interval corresponding to fixing β at 1 (up to granularity of the grid and simulation error).

Lehman point estimates by maximizing the p-value over the confidence region for the various statistics and omnibus statistics examined. Because all these p-values are generated with a monte-carlo simulation we, to reduce Monte-Carlo simulation error, first calculated an upper 95% confidence interval for the p-value at the maximum found point, and then took all points that had p-value estimates within this interval. Finally, to get a single point estimate, we reported the median of each cloud, reasonable when the points are tightly clustered as they usually are. Table 4 contains the results for a variety of test statistics. The clouds are depicted on the figures. We generally see estimates for τ in the \$1000-\$1500 range and scalings of 1.02 to 1.18.

Table 4: Hodges-Lehman style point estimates for the Multi-Tobit model under different test statistics. Above line are single statistics. Below line are omnibus.

Test Statistic	$\hat{\tau}$	$\hat{\beta}$	p-value
$\hat{p}$	1552	1.00	1.00
KS	1086	1.10	0.98
KS Positive Support	0	1.10	0.96
W	1280	1.10	1.00
$\hat{\sigma}$	3414	1.53	1.00
$W + \hat{p}$	1474	1.05	1.00
$W + \hat{p} + \hat{\sigma}$	1241	1.18	0.74
$KS + \hat{p}$	1513	1.02	0.96
$\hat{p} + \hat{\sigma}$	1047	1.49	1.00
All	1397	1.08	0.76

We then adjusted the outcomes for treated workers using the point estimate corresponding to the omnibus of all four statistics and compared these adjusted treated outcomes to actual control outcomes using boxplots. If the multi-tobit model of effects is correctly specified, the boxplot of the adjusted treatment units should look more similar to the control units than the unadjusted treatment units do. The results are in Figure 11.They indeed look more similar, although we are possibly over-shrinking with a somewhat too large value of β.

Overall, there is considerable uncertainty due to the sample size and the difficulty of assessing goodness of fit in general. The tails, in particular, are hard to pin down due to having so few points. Even under the very structured Tobit and Multi-Tobit models, we cannot truly differentiate

between amplification of positive earnings and simply shifting workings into employment.

An alternate approach would be to investigate the scaling and the zero inflation separately. The zero-inflation is relatively straight forward: the difference in proportions in those with positive earnings is 11%, with a 95% confidence interval of 2%-20%. We can reject the null for a test of equal proportions ($p = 0.02$), indicating that the NSW program did move people into positive earnings. To estimate impacts on the positive earnings, however, we need a model of effects to describe them. We cannot simply compare positive earnings in each group, since the positive earnings in the treatment group is a mixture of those who would otherwise have obtained positive earnings, and those who would not. The Tobit and Multi-Tobit model allow for this separation. Other models are possible, such as principle stratification (**?**). All of these approaches usually rely on strong distributional assumptions. We recommend a sensitivity approach of trying many, and attempting to assess model fit.

As a final note, we acknowledge that model checking and fitting is more an art than a science. We show many different tests and combinations to illustrate that different test statistics focus on different aspects of model-checking. Taken as a whole, they can be considered a sensitivity analysis in the style of ensuring overall impacts are consistent across a range of different estimation procedures. Here, we do seem to see a suggestion of impact on both positive earnings and moving people into positive earnings, but the former claim is not statistically significant, likely a function of it not being directly testable without strong assumptions.

6 Summary

Outcome data with clumping at zero invariably require strong assumptions for analysis. Parametric approaches to these data rely on distributional assumptions about the data. The motivation for these assumptions is the likelihood that distinct behaviors produce values that can only be observed in a single distribution. In job training applications where the outcome is wages, we can easily imagine a process whereby the intervention moves some subjects into positive earnings but

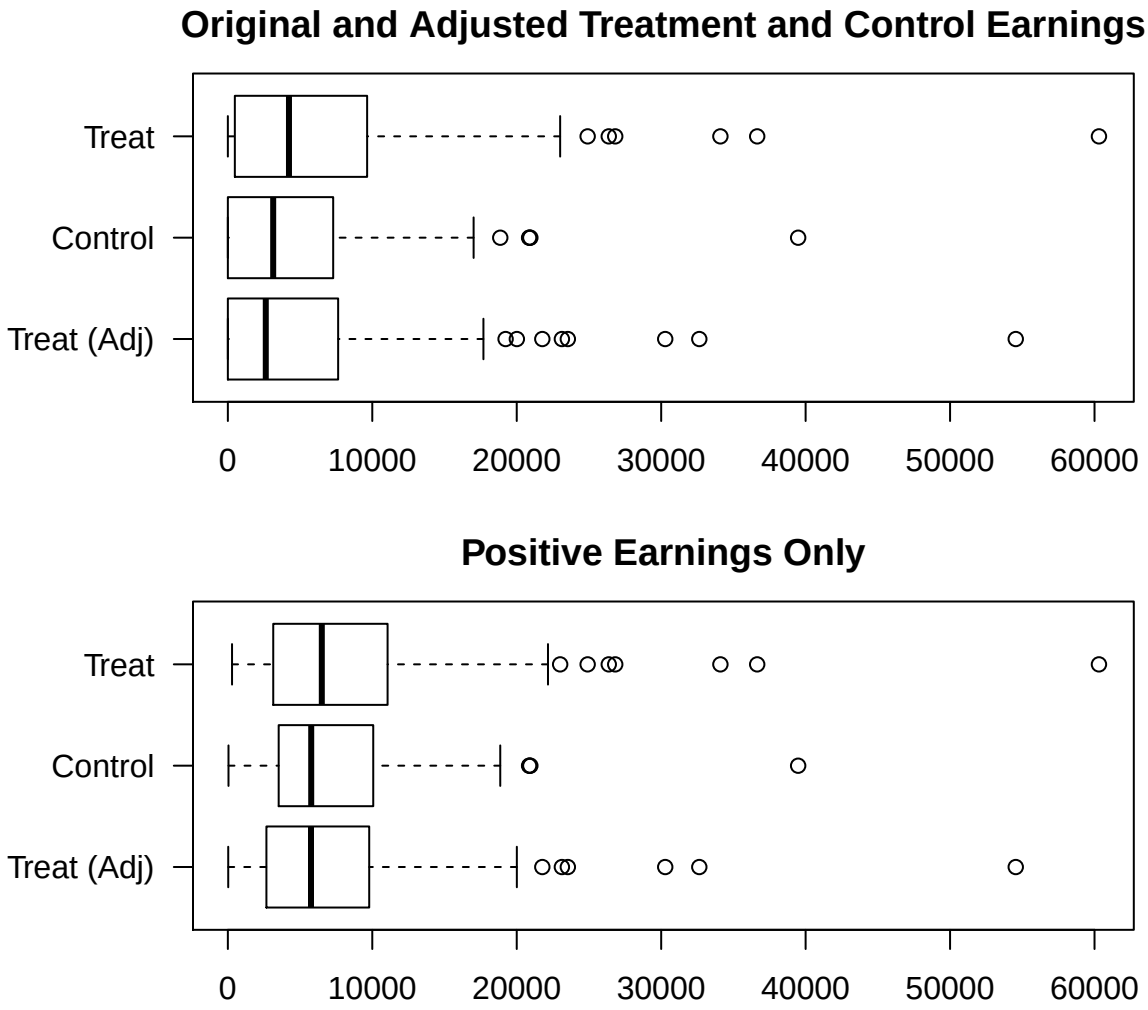


Figure 11: Boxplots comparing Multi-Tobit adjusted treatment outcomes to control outcomes with $\hat{\beta} = 1.08$ and $\hat{\tau} = \$1397$. Adjusted treated outcomes should be similar to control outcomes (so in the set of three, the lower boxplot should match the middle more than the upper). Bottom set are positive outcomes only.

for other subjects increases earnings.

Our approach still requires modeling assumptions but these assumptions are different arguably far weaker than those need for parametric modeling. Moreover, we develop and emphasize the need for goodness of fit tests, to allow for model checking. In both applications, we found evidence that the Tobit model of effects was a poor fit. Moreover, the large number of tests that are possible suggest careful pre-specification in an analysis plan to avoid fishing and over-fitting the data. Because the assumptions are different, conducting analysis such as these as well as the full parametric approach could be worthwhile as a final sensitivity check.