Hypothesis testing counterexample

Kellie Ottoboni

Draft May 4, 2016

1 Set-Up and General Case

Suppose we observe N units with outcomes Y. The response schedule is

$$Y_i = aX_i$$

for some $a \in \mathbb{R}$. Instead of X, we observe $\tilde{X} = X - \delta Z$, where $X \perp Z$. We observe Z. We are interested in whether Z has an effect on Y. According to the response schedule, any association is purely through association with X.

We will use the absolute value of correlation between Z and the residuals as our test statistic:

$$\rho(Z, Y - \hat{Y}) = \frac{\langle Z, Y - \hat{Y} \rangle}{\|Z\| \left\| Y - \hat{Y} \right\|}$$

Since the denominator does not depend on the order of elements of Z, we may disregard it, since it will be constant when we permute Z. Thus our test statistic will be $\tau = |\langle Z, Y - \hat{Y} \rangle|$.

1.1 No stratification

Suppose we do not attempt to stratify observations. Assume that we know a, but instead of observing X we observe Then our prediction will be $\hat{Y} = a\tilde{X}$. The observed test statistic will be

$$
\begin{aligned}
\tau^{obs} &= |\langle Z, Y - a\tilde{X}\rangle| \\
&= |\langle Z, Y - a(X - \delta Z)\rangle| \\
&= |\langle Z, a\delta Z\rangle| \\
&= |a\delta|\|Z\|^2
\end{aligned}
$$

Let Π be the set of all permutations of Z. For any permutation $Z^* \in \Pi$,

$$
\begin{aligned}
\tau^* &= |\langle Z^*, Y - a\tilde{X}\rangle| \\
&= |\langle Z*, a\delta Z\rangle| \\
&\leq |a\delta|\|Z\|\|Z^*\| && \text{by Cauchy-Schwarz} \\
&= |a\delta|\|Z\|^2
\end{aligned}
$$

with equality if and only if $Z = Z^*$. Therefore the p-value of the test will be

$$
p = \mathbb{P}(\tau^* \geq \tau^{obs} \mid\mid H_0) = \frac{\#\{Z^* \in \Pi \mid Z^* = Z\}}{N!}
$$

The p-value is anti-conservative if $p \leq \alpha$ more than $\alpha(100)\%$ of the time. If Z has a continuous distribution, all elements of Z will be distinct with probability 1, so $p\langle\alpha$ whenever $N!\rangle\alpha^{-1}$. This is problematic: for example, if $\alpha = 0.05$, the test is anti-conservative when $N \geq 4$.

Suppose instead that Z has a discrete distribution. Let D be the number of distinct elements in Z, and suppose the ith distinct element appears d_i times, so that $d_1 + \cdots + d_D = N$. Then the numerator is $\#\{Z^* \in \Pi \mid Z^* = Z\} = d_1! \cdots d_D!$. Therefore the test is anti conservative whenever $d_1! \cdots d_D! \leq \lfloor \alpha N! \rfloor$. Suppose that $\alpha = 0.05$ and $N = 10$. The lefthand side is $181,440$. The test will only have the correct level if there are only 2 distinct elements, one of which appears 9 times, or all elements are the same (in which case the test is not informative).

1.1.1 Issue

The test goes wrong because the elements of Z are not exchangeable given $\hat{Y}$. In particular, we introduce a bit of Z into $\hat{Y}$ because we model the outcome incorrectly, only observing the noisy measure $\tilde{X}$ instead of the true X.

1.2 Stratified

Suppose we stratify by $\hat{Y}$. Let the subscript s index stratum, for $s = 1, \ldots, S$. Let τ_s be the inner product of Z and $Y - \hat{Y}$ for individuals in stratum s. The test statistic will be $\tau = \sum_{s=1}^{S} \tau_s$. We have the same problem as before, but now $\tau_s^{obs} \geq \tau_s^*$ for all permutations of Z for which $Z_s = Z_s^*$.

$$\begin{aligned}
\tau^{obs} &= \sum_{i=1}^{S} \tau_s^{obs} \\
&= \sum_{i=1}^{S} |a\delta| \|Z_s\|^2 \\
&= |a\delta| \|Z\| \\
\tau^* &= \sum_{i=1}^{S} \tau_s^* \\
&= \sum_{i=1}^{S} |\langle Z_s *, a\delta Z_s \rangle| \\
&\leq \sum_{i=1}^{S} |a\delta| \|Z_s\| \|Z_s^*\| && \text{by Cauchy-Schwarz} \\
&= |a\delta| \|Z\|^2
\end{aligned}$$

We have the same issue: $\tau^{obs} \geq \tau^*$, with equality if and only if $Z_s^* = Z_s$ for each $s = 1, \ldots, S$.

1.2.1 Issue

We have the same problem as in the last section. Now, the issue would be solved with a weaker condition: exchangeability of the residuals *within strata*. Intuitively, stratum assignment may depend on Z in some way, but within strata, there should be no dependence on Z. A stronger condition is the following.

Assumption 1. *Conditional independence between potential outcomes and treatment given stratum assignment:* $Y(Z) \perp\!\!\!\perp Z \mid S$.

For Assumption 1 to hold, the stratification must be so fine that it variations in $\hat{Y}$ within strata are unrelated to treatment assignment. Intuitively, the results within each strata must act as though they had come from a randomized experiment. This condition implies that $\mathbb{P}(Z_i = z \mid S_i = s, Y_i(Z)) = \mathbb{P}(Z_i = z \mid S_i = s)$, so the probability of any particular unit receiving level z of treatment is constant within strata.[1] In an actual randomized experiment, the analogous condition is that the treatment is assigned at random within strata.

2 With Orthogonalization

Suppose now that $Y_i = aX_i$, and we observe X and Z. However, X and Z are correlated. We'll begin by orthogonalizing the covariate. Define $\tilde{X}$ to be X minus the projection of X onto Z:

$$\tilde{X} = X - \frac{\langle X, Z\rangle}{\|Z\|^2} Z$$

This is a special case of the previous section, setting $\delta = \frac{\langle X,Z\rangle}{\|Z\|^2}$. Now, we predict Y by $\hat{Y} = \hat{a}\tilde{X}$ for an appropriately chosen $\hat{a}$.

3 Noise In Prediction

Return to the general case where $\tilde{X} = X - \delta Z$. Suppose that we don't know a and we estimate it by $\hat{a}$. It is unclear whether the test has the right level: there doesn't seem to be any general rule about it.

$$\begin{aligned}
\tau^{obs} &= |\langle Z, Y - \hat{Y}\rangle| \\
&= |\langle Z, aX - \hat{a}\tilde{X}\rangle| \\
&= \left| a\langle Z, \tilde{X} + \delta Z\rangle - \hat{a}\langle Z, \tilde{X}\rangle \right| \\
&= |a - \hat{a}||\langle Z, \tilde{X}\rangle| + |a\delta| \, \|Z\|^2
\end{aligned}$$

In the orthogonal case, the first term is 0. In terms of X instead of $\tilde{X}$,

[1] Wager and Athey [2015] make this assumption about the leaves of their causal trees: the leaves are so fine that they approximate a randomized experiment within them.

$$
\begin{aligned}
\tau^{obs} &= |\langle Z, Y - \hat{Y}\rangle| \\
&= |\langle Z, aX - \hat{a}\tilde{X}\rangle| \\
&= |a\langle Z, X\rangle - \hat{a}\langle Z, X - \delta Z\rangle| \\
&= |a - \hat{a}||\langle Z, X\rangle| + |\hat{a}\delta| \, \|Z\|^2
\end{aligned}
$$

Now, the permuted test statistic will be

$$
\begin{aligned}
\tau^* &= |\langle Z^*, Y - \hat{Y}\rangle| \\
&= \left| a\langle Z^*, \tilde{X} + \delta Z\rangle - \hat{a}\langle Z^*, \tilde{X}\rangle \right| \\
&= |a - \hat{a}||\langle Z^*, \tilde{X}\rangle| + |a\delta| \, |\langle Z^*, Z\rangle|
\end{aligned}
$$

Equivalently,

$$
\begin{aligned}
\tau^* &= |\langle Z^*, Y - \hat{Y}\rangle| \\
&= |a\langle Z^*, X\rangle - \hat{a}\langle Z^*, X - \delta Z\rangle| \\
&= |a - \hat{a}||\langle Z^*, X\rangle| + |\hat{a}\delta| \, |\langle Z^*, Z\rangle|
\end{aligned}
$$

We can use Cauchy-Schwarz to bound $\langle Z^*, Z\rangle$, but we don't know anything about how $\langle Z^*, X\rangle$ and $\langle Z^*, \tilde{X}\rangle$ vary across $Z^* \in \Pi$.

4 Instrumental Variables Set-up

This is the typical two-stage structural model for instrumental variables.

$$Y = aX + \varepsilon$$
$$X = bZ + \delta$$

where X, Z, and (δ, ε) are all mutually independent, but $\mathrm{Cov}(\delta, \varepsilon)\rangle 0$. Predict Y by $\hat{Y} = \hat{a}X$ for some appropriate $\hat{a}$. Then,

$$\begin{aligned}
\tau^{obs} &= |\langle Z, Y - \hat{Y}\rangle| \\
&= |\langle Z, aX + \varepsilon - \hat{a}X\rangle| \\
&= |\langle Z, (a - \hat{a})(\gamma Z + \delta) + \varepsilon\rangle| \\
&= |\langle Z, (a - \hat{a})\delta + \varepsilon\rangle| + |(a - \hat{a})\gamma| \|Z\|^2
\end{aligned}$$

And similarly,

$$\begin{aligned}
\tau^{*} &= |\langle Z^{*}, (a - \hat{a})\delta + \varepsilon\rangle| + |(a - \hat{a})\gamma| |\langle Z^{*}, Z\rangle| \\
&\leq |\langle Z^{*}, (a - \hat{a})\delta + \varepsilon\rangle| + |(a - \hat{a})\gamma| \, \|Z\|^2
\end{aligned}$$

While we don't know exactly the first term in finite samples, we know it ought to be small because of the independence between Z and (δ, ε). However, it seems that the observed test statistic may be systematically larger than the permuted statistic because of the second term involving Z.

5 Philip's email

I just tried to do the math, and I'm not getting what I expected...

To simplify the notation let $T = \tilde{X} = X - \delta Z$. We have $Y = aX$ and $\langle X, Z\rangle = 0$, which implies that $\langle Y, Z\rangle = a \times 0 = 0$. First, $\langle Y, T\rangle = \langle Y, X - \delta Z\rangle = \langle Y, X\rangle - \delta\langle Y, Z\rangle = \langle Y, X\rangle = a\|X\|^2$. So,

$\hat{a} = \langle Y, T\rangle / \|X\|^2 = a$.

But if we don't know X (or at least $\|X\|^2$), we can't compute that.

If we defined $U = T - Z\langle T, Z\rangle/\|Z\|^2$, then $U = X - \delta Z - Z\langle X - \delta Z, Z\rangle/\|Z\|^2 = X - \delta Z + \delta Z\|Z\|^2/\|Z\|^2 = X$, so we can find X.

It remains to show that the OLS estimate of a is $\hat{a}$.

The proof strategy is to suppose that the OLS estimate is $a + \gamma$, then show that $\gamma = 0$.

Now, $\|Y - (a+\gamma)T\|^2 = \|Y\|^2 - 2(a+\gamma)\langle Y, T\rangle + (a+\gamma)^2\|T\|^2$. Since $\|Y\|^2$ is fixed, we minimize that by minimizing the last two terms.

Substitute $\langle Y, T\rangle = a\|X\|^2$. Then the sum of the last two terms is

$-2(a+\gamma)a\|X\|^2 + (a+\gamma)^2\|T\|^2$.

Differentiating wrt γ to find a stationary point gives (if I got the algebra right)

$\gamma = a\left(\frac{\|X\|^2}{\|T\|^2} - 1\right)$.

Since $\|X\|^2 < \|X\|^2 + \delta^2\|Z\|^2 = \|T\|^2$, that seems to show that $\gamma \neq 0$, and hence that $\hat{a}$ isn't the OLS estimator...which doesn't make sense to me.

References

Stefan Wager and Susan Athey. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *arXiv:1510.04342 [math, stat]*, October 2015. URL `http://arxiv.org/abs/1510.04342`. arXiv: 1510.04342.