Title: Power analysis of common selection bias tests

selection models tests dont just test for constancy of mean, they are oriented toward a trend, presumably for increased power, leveraging assumptions about researchers' practices.

testing for publication bias before conductingn meta analysis is itself a selection event.

mu=0, thresholding on p-value. easy to analyze, as z and s are independent.

$\theta \neq 0$, thresholding on raw value, study heterogeneity: induce selection on s as well as z, and dependence between z and s.

test intuition If no selection, i.e., null, and $y \sim (\theta, \sigma_j)$. Egger's test is the regression $y_j/\sigma_j \sim (\theta/\sigma_j, 1)$ on $1/\sigma$. The linear model $y_j/\sigma = (1, 1/\sigma)^T(\beta_0, \beta_1)+\epsilon$ is satisfied with $\beta = (0, \theta)$ and $\epsilon_j = y_j/\sigma_j - \theta/\sigma_j$ independent with equal variance 1, conditionally on S's. According to usual [[random design]] OLS theory, the test is consistent under the null, asymptotic normality valid for inference. Under the gaussian model [ref] valid finite-sample inference conditional on §'s. Common to use a t-test but under the null the error variances are known to be 1 and could just use a normal reference distribution. [maybe an acknowledgment of possible violations of the model [ref basic meta analysismodel], e.g., not accounting for sampling variability in the sigmas].

alternative/Selection present, $\theta = 0$. preselection, the observations are $y_j \sim (0, \sigma_j)$. If selection is on the p-value/z-stat $y/\sigma \sim (0, 1)$. If $y_j/\sigma_j \mid \sigma_j \sim f$ ie in addition to $E(y_j/\sigma_j \mid \sigma_j) = \theta/\sigma_j, V(y/\sigma_j \mid \sigma_j) = 1$, the entire conditional distribution is specified, ie the conditional distributions $y_j \mid \sigma_j$ are a scale family $f(y/\sigma)/\sigma$. (show can assume any reasonable selection mechanism (ie symmetric in arguments) $g(y_1/\sigma_1, \ldots, y_n/\sigma_n, U)$ can be reduced to a function $g(y_j/\sigma_j, U)$ in this case since the y_j/σ_j are iid. then eg hard thresholding is given by..., probabilistic threshold is given by...) Then postselection response is independent of postselection regressor: given u, v and selection mechanism g_j, $E(u(g_j(y/\sigma_j))v(1/\sigma_j)) = E(E(\ldots \mid \sigma_j)) = E(E(u(g_j(y/\sigma_j)) \mid \sigma_j)v(1/\sigma_j)) = E(u(g_j(z)))E(v(1/\sigma_j))$ with $z \sim f$. If all the selection mechanisms are the same say g, then $E(g(y_j/\sigma_j) \mid 1/\sigma_j) = E(g(z))$ is constant, and as before conditional variance is constant. Again a wellspecified homoskedastic linear model $y_j/\sigma_j \sim (1, 1/\sigma_j)^T\beta + \epsilon_j$, now with $\beta = (E(g(z)), 0)$. So test is consistent. [provided $E(g(z)) = E(z^*) \neq 0$]. [[$\theta \neq 0$, raw value thresholding]] [[analogous test intuition for begg test. use idealized test statistic.]]

test slopes under pval selection, $\theta = 0$ slope of egger test, p-val thresholding

Parameterize the p-value selection models $\{(Z, S) \mid Z > c : c \in \operatorname{supp} Z\}$ by the mean of Z^*. This is possible since the mean

$$\mu(c) = \int_c^\infty z f_Z(z) dz/(1 - F_Z(c))$$

is a strictly monotonic function of the cutoff c [need to assume density $f_Z(z) \neq 0$ for all $z \in \operatorname{supp} Z$ ie support is convex. throughout, perhaps clarify focus is on

"continuous scale families" not just "scale families"] ,

$$\begin{aligned}\mu'(c) &= \frac{-cf_Z(c)}{1-F_Z(c)} + \frac{\left(\int_c^\infty z f_Z(z)dz\right) f_Z(c)}{(1-F_Z(c))^2} \\ &= \frac{f_Z(c)}{(1-F_Z(c))^2}\left(\int_c^\infty (z-c) f_Z(z)dz\right) > 0.\end{aligned}$$

Let $h > 0$, let $\theta_n = h/\sqrt{n}$, let P_n denote the law of (S^*, Z^*) conditional on $\{Z > c(\theta_n)\}$. Assume $V(S) < \infty$, $f_Z(z) \neq 0$ for $z \in \operatorname{supp} Z$. Then

$$\lim_n P_n\left(\frac{\hat\beta_0}{\sqrt{V(\hat\beta_0)}} > t_{n-1,1-\alpha}\right) = 1 - \Phi\left(z_{1-\alpha} - \frac{h}{\sqrt{V(Z)E(S^2)/V(S)}}\right).$$

So the test slope is $\frac{V(S)}{V(Z)E(S^2)}$. [proof just uses lindeberg clt, can probably drop identically distributed assumption. but would complicate the parameterization of the alternatives. might be able to state parameterization in terms of cutoff for the mean of the z.]

Proof. [0. formula for betahat. this will probably be used elsewhere, maybe set off.] Egger's test is to reject when $\hat\beta_0/\sqrt{\hat V(\hat\beta_0)} > t_{n-1,1-\alpha}$.

Let X_n be the $n \times 2$ design matrix, ie a column of $1's$ and a column of the regressors s_j. Let ζ_n by the column vector of measurements $z_j^* = y_j^*/\sigma_j^*$. Then $\hat\beta_0$ is the first component of $(X_n^t X_n)^{-1} X_n^t \zeta_n$, which computes to $\hat\beta_0 = (\hat V(s))^{-1} n^{-1} \sum_{j=1}^n (\overline{s^2} - \overline{s} s_j) z_j$, where $\hat V(s) = n^{-1}\sum_j (s_j - \overline{s})^2$. The variance estimate is $\hat V(\hat\beta_0) = \frac{SSE \cdot \overline{s^2}}{n(n-2)\hat V(S)}$ $[[\overline{s^2}\hat V(s) SSE/(n-2)]]$, where $SSE = ||(I - X(X^tX)^{-1}X^t)\zeta_n||^2$. So the test statistic is

$$\frac{n^{-1/2}\sum_{j=1}^n (\overline{s^2} - \overline{s} s_j) z_j}{\sqrt{\overline{s^2}\hat V(s) SSE/(n-2)}}.$$

[1. First step: make iid: 1–variance term, 2–means] To show: Asymptotic equivalence of

$$n^{-1/2}\frac{\sum_{j=1}^n \left(\hat V(S)^{-1}(\overline{S^2} - \overline{S} S_j) Z_j - \mu_n\right)}{\sqrt{\overline{S^2}/\hat V(S) SSE/(n-2)}}.$$

and

$$n^{-1/2}\frac{\sum_{j=1}^n \left(V(S)^{-1}(E(S^2) - E(S) S_j) Z_j - \mu_n\right)}{\sqrt{V(Z)E(S^2)/V(S)}}.$$

To show: [need to assume $V(S) < \infty$]

$$\sqrt{n}\left(\left(E(S^2) - \overline{S^2}\right)\overline{Z^*} - \overline{SZ^*}\left(E(S) - \overline{S}\right)\right) \underset{P_n}{\to} 0$$

where convergence is in probability under the sequence of laws P_n of $(Z^*, S^*) = (Z^*, S)$ along θ_n. Distribution of S does not change with P_n.

First term: $\sqrt{n}\left(E(S^2) - \overline{S^2}\right)$ is $O_{P_n}(1)$ by the CLT. and $\overline{Z^*} \to 0$ using a weak LLN for triangular arrays and $Z^* \sim_{\theta_n} Z\{Z > c_n\}/(1 - F_Z(c_n))$. So $\sqrt{n}\left(E(S^2) - \overline{S^2}\right)\overline{Z^*} \to 0$ in probability along P_n. Second term: $\sqrt{n}\left(E(S) - \overline{S}\right)$ is $O_{P_n}(1)$ by CLT. Orthogonality of S and Z^* and domination of Z^* implies $\overline{SZ^*} \to_{P_n} E_0(SZ) = 0$.

To show: $(n-2)/nRSS = \zeta_n^t(I - X(X^tX)^{-1}X^t)\zeta_n/n \to_{P_n} V(Z)$, convergence is in probability along P_n.

$$\begin{aligned}\zeta_n^t X(X^tX)^{-1}X^t\zeta_n &= (\hat{V}(s))^{-1}\left(\overline{z^*}(\overline{s^2}\,\overline{z^*} - \overline{s}\,\overline{sz^*}) + \overline{z^*s}(\overline{z^*s} - \overline{z^*}\,\overline{s})\right)\\ &= (\hat{V}(s))^{-1}\left(\overline{s^2}(\overline{z^*})^2 + (\overline{z^*s})^2 - 2\overline{s}\,\overline{z^*sz^*}\right).\end{aligned}$$

[fix overbar leaking over like unibrow] Converges in probability to 0 as above [verify], with each monomial convering to $E(S^2)E(Z^2)$. So $RSS = o_{P_n}(1) + n^{-1}\zeta_n^t\zeta_n$, which tends along P_n to $E(Z^2) = V(Z)$, as above.

[2. Second step: CLT application]

Let Z_n^* denote the postselection distribution of Z under $\theta_n = h/\sqrt{n}$, i.e., conditional on $\{Z > Sc(\theta_n)\}$ [causes confusion with indexing subscript]. Let E_n denote expectation under $\theta_n = h/\sqrt{n}$, i.e., conditional on $\{Z > Sc(\theta_n)\}$. Only relevant for Z^* since the distribution of $S^* \sim S$ does not change with n. Let $\mu_n = E(Z_n^*) = h/\sqrt{n}$.

Apply Lindeberg-Feller CLT to conclude asy normality:

$$n^{-1/2}\frac{\sum_{j=1}^n\left(V(S)^{-1}(E(S^2) - E(S)S_j)Z_j - \mu_n\right)}{\sqrt{V(Z)E(S^2)/V(S)}} \underset{\Rightarrow}{\theta_n} N(0,1)$$

Since the (Z_j^*, S_j) are iid, the Lindeberg condition is

$$E_n\left(\left(\frac{V(S)^{-1}(E(S^2) - E(S)S_1)Z_1^* - E_n(Z^*)}{\sqrt{V(Z)E(S^2)/V(S)}}\right)^2 ; n^{-1/2}|\ldots| > \epsilon\right) \to 0$$

for all $\epsilon > 0$. Ellipses represent the term in parenthesis. The family $\{V(S)^{-1}(E(S^2) - E(S)S_1)Z_1^* - E_n(Z^*)\}$ over the probabilities P_n is in fact uniformly integrable. Since the distribution of S_1 does not depend on P_n, S_1 and Z_1^* are independent, and $Z_1^* \sim_{\theta_n} Z\{Z > c_n\}/(1 - F_Z(c_n)) \to_{a.s.} Z$.

Also

$$\begin{aligned}E_n\left(\left(\frac{(E(S^2) - E(S)S_1)Z_1^*}{V(S)} - E_n(Z^*)\right)^2\right) &= \frac{E_n((Z^*)^2)((E(S^2))^2 - E(S^2)(E(S))^2)}{V(S)^2} - (E_n(Z^*))^2\\ &= E_n((Z^*)^2)E(S^2)/V(S) - (E_n(Z^*))^2\\ &\to \frac{V(Z)}{V(S)}E(S^2)\end{aligned}$$

as $n \to \infty$. By definition of E_n, $E_n(Z^*) \to 0$ and as above domination of Z^* by bounded multiples of Z implies $E_n(Z^*) \to E(Z^2)$. So the summands are standardized.

[3. Third step: obtain slope]

The local limiting power at the null $\theta = 0$ is then

$$\lim_n P_n \left(\frac{\hat{\beta}_0}{\sqrt{V(\hat{\beta}_0)}} > t_{n-1,1-\alpha} \right) = \lim_n P_n \left(n^{-1/2} \frac{\sum_{j=1}^n V(S)^{-1}(E(S^2) - E(S)S_j)Z_j}{\sqrt{V(Z)E(S^2)/V(S)}} > t_{n-1,1-\alpha} \right)$$

$$= \lim_n P_n \left(n^{-1/2} \frac{\sum_{j=1}^n \left(V(S)^{-1}(E(S^2) - E(S)S_j)Z_j - \mu_n \right)}{\sqrt{V(Z)E(S^2)/V(S)}} > t_{n-1,1-\alpha} - \frac{n^{-1/2}\mu_n}{\sqrt{V(Z)E(S^2)/V(S)}} \right)$$

$$= 1 - \Phi \left(z_{1-\alpha} - \frac{h}{\sqrt{V(Z)E(S^2)/V(S)}} \right).$$

□

describe pearson begg statistic, the begg statistic [[ref a definition above but make sure it is the equivalent version using S not σ in the pairs]] except instead of using Kendall's τ, using the Pearson correlation coefficient:

...

similar variations of begg's statistic have been considered previously, e.g., [[bmc paper]] suggesting spearman's rho. the definition [[ref pearson begg above]] also omits the $O(1/n)$ variance adjustments .... A test of publication bias rejects if [[ref pearson correlation]] is large. Given IID normal data [[which the above is not due to thetahat]] one would usually refer to the t_{n-2} distribution. Doing so is equivalent to testing for a slope of 0 in the linear model

$$Z - \theta S \sim 1 + S$$

relate this test to egger's test. θ is the projection of $Y/\sigma = YS = Z$ onto the S's, and the regressand in [[ref above regression]], $(Y - \theta)S = Z - \theta S$, is the residual. ... Therefore the intercept in [[above regression]] is the same as the intercept in the Egger regression [[ref egger regression defn]]. The RSS is also the same.So testing the intercept being 0 is the same as carrying out Egger's regression, and testing the slope being 0 carries out the pearson version of Begg's test. The joint distribution of the slope and intercept is readily available under the gaussian model, or asymptotically more generally. Therefore it would be reasonable to test for both being equal to 0 to improve the FPR or for either being $\neq 0$ to improve power. The regression residuals are orthogonal and following lin their skewness under typical selection models gives additional information that may be used to boost power.

Two related problem with this program:

first, the correlation test [[ref]] is not a consistent test, as second, the test statistic is redundant to the egger test.

1. linear relationship between betahats
2. the difference in the t-statistics is:
3. correct distribution of f-stat (which is t-stat)

coefficients are related by $\hat{\beta}_0 m_1 = -\hat{\beta}_1 m_2$. some linear relationship would have to hold between the coefficients since the projection of the regressand onto the design column space $M(1, s)$ lies in the one dimensional subspace $M(1, s) \cap s^{\perp}$.

item 1 shows that the power function of the pearson begg test cannot exceed that of egger's test. Since the egger statistic [[ref inuition section]] is consistent [[under what model]], the pearson begg statistic cannot be, if

$$diff \not\to 0$$

. the observations are not IID due to the common term θ, so that the usual consistency assumptions for OLS are unmet. This problem is not specific to the pearson begg test. The original begg test also suffers from bias in the same direction due to θ [[cite other manuscript]]. The pearson begg test is like the original begg test in both good and bad ways. linear relationship between the two stats.

by item 2 the correct distribution is, which gives a test identical to the egger test. Therefore the difference between the pearson begg test and egger's test is due entirely to the uncorrected bias. As the pearson begg test and original begg test differ only in the choice of correlation measure and the $O(1/n)$ variance corrections, conclude that the difference between Egger's test and Begg's test arise due to 1) the robustness/efficiency issues of using a rank correlation measure rather than the Pearson correlation, 2) an unaccounted-for bias in Begg's test, and 3) $O(1/n)$ variance corrections.