Title: Power analysis of common selection bias tests

1. *background*

   (a) *background on meta-analysis, basic model* Meta-analysis is a ...[[copy from biometrics paper]]

   Pairs usually assumed independent, though need not be iid. key assumption is a common mean, the "grand mean" θ.

   $$(y_1, \sigma_1), \ldots, (y_n, \sigma_n)$$
   $$E(y_j) = \theta \in \mathbb{R}, V(y_j) = \sigma_j^2, \quad j = 1, \ldots, n$$

   [maybe mention fixed vs random effects, we are assuming adjustment by between-study variance has been made and σ_j^2 represents the unconditional variance of study y_j.]

   (b) *issue of selection, publication bias, funnel plots, formal tests: egger and begg* [maybe clarification of small sample bias defn should be here]

   [introduce two tests here]

   mention selection mechanisms here, selection on standardized statistic, on raw statistic, randomness added in.

   natural way to look for an association is a correlation between the test statsitics and σ_j, and begg's test is of this type. [define]. Egger's test a little different...

   selection on the statistic of interest can induce selection on the sigmas also.

   for selection on y such as $zs > c$ one expects both z and s to be stochastically larger postselection and negatively dependent.

   definitions/selection on p-value seems not to be a small study effect. 1. term "small study bias" may be ambiguous. No study size bias in selecting according to the unstandardized effect, $\{y > c\}$, is a small study bias when viewed as selection wrt p-value, $\{z > c\sigma\}$. No bias in selecting according to p-value, $\{z > c\}$, is a bias against larger studies when viewed as selection on the raw value, $\{y > c/\sigma\}$, the opposite of a small study effect.

   (c) *justification for power analysis* If selection is present and $\mu = 0$ (true null) then a type 2 error on the selection test may lead to a type 1 error on the meta-analysis. If selection is present and $\mu \neq 0$ (true non-null), then a type 2 error on the selection test may lead to exaggeration of the true non-null meta-analysis effect. Either way, for practical purposes roles of type 1 and type 2 error of the selection test are reversed. The conservative analyst would want to bound the probability of a type 2 error of the selection test, though there is little hope of doing so with such a complex alternative as "selection bias". [sense to use the roles borrowed from usual t-tests.] [similar

issue encountered when using other tests for screening, e.g., meeting a normality assumption using a K-S test with null of normality.]

2. If $0 < x_1 < \ldots < x_n$ then there is no monotonic $v \in C(X)^\perp$. $\sum v_j = 0$, let $v_1 < \ldots < v_k < 0 < v_{k+1} < \ldots < v_n$, divide through by $\sum_{k+1}^n v_j$ so $\sum_1^k -v_j = sum_{k+1}^n v_j = 1$. then $(v, x) = 0$ implies $-\sum_1^k v_j x_j = sum_{k+1}^n v_j x_j$ but the lhs is $\in (x_1, x_k)$ whereas rhs is $in(x_{k+1}, x_n)$. Same proof works if there is some k such that $v_j < 0$ iff $j \leq k$.

   But what vectors v are in fact orthogonal to the columns space? not monontonic vectors, but maybe other permutations. ortho complement is dimension $n-2$ so any of the components other than 2 can be ordered arbitrarily. (11/6) Tried to show that given any $v_1, \ldots, v_n$ with $\sum v_j = 0$ and there is no k with $v_1, \ldots, v_k$ all < 0, $v_{k+1}, \ldots, v_n$ all > 0 its orthogonal complement contains some $0 < x_1 < \ldots < x_n$. But counterexample: $v = (-1/2, 1/2, -1/2, 1/2)$, let $x = (a, a+b, a+b+c, a+b+c+d)$, then $(v, x) = -(2a+b+c)/2 + (2a+2b+c+d)/2 = (b+d)/2 > 0$.

3. *model* Data are independent random pairs $(y_j, \sigma_j), \sigma_j > 0, E(y_j \mid \sigma_j) = \mu, var(y_j \mid \sigma_j) = \sigma_j^2, j = 1, \ldots, n$. Independence implies the distribution of $(y_1, \ldots, y_n) \mid (\sigma_1, \ldots, \sigma_n)$ is the same as the distribution of $(y_1 \mid \sigma_1, \ldots, y_n \mid \sigma_n)$ (implied by bayes rule in case densities exist, and if not, how to define conditional distribution anyway?) The actual property we need. allows us to replace the conditional test statistic of a meta-analysis data set say $T((y_1, \sigma_1), \ldots, (y_n, \sigma_n)) \mid \sigma_1, \ldots, \sigma_n$ with $T((y_1 \mid \sigma_1, \sigma_1, \ldots, (y_n \mid \sigma_n, \sigma_n)))$. (this property might be better than independence. e.g. analysists might choose sample sizes hence σ_j based off of what other analysts have chosen.)

   Selection on the raw value of the type $\{Z > cS\}$ affects the distribution of the precisions S as well. The class of postselection distributions S^* is in principle only limited by the support of Z. Given $c \in \mathbb{R}$, mean-zero $Z \sim f_Z$ and a density on f on the nonnegative reals, such that $F_Z(cs) < 1$ for any s in the support of f. Then f is a distribution for S^* given $\{Z > cS\}$ under the scale model [ref], when the unconditional distributional of S is given by $g(s) = f(s)/(1 - F_Z(cs))$.

   Proof. Since in that case,

$$\begin{aligned} f_{Z^*,S^*}(a,b) &\propto f_{Z,S}(a,b)\{a > cb\} \\ &= f_Z(a) f_S(b)\{a > cb\} \\ &= f_Z(a)\{a > cb\}\frac{f(b)}{1 - F_Z(cb)} \end{aligned}$$

   and integrating out a gives $f_{S^*}(b) = f(s)$. □

   [may combine these small facts inone place. most are related to scale family]

The conditional density of Y^* given S^* given a selection event $A \in \sigma(Y,S)$ is related to the conditional density of Y given S as

$$f_{Y^*|S^*=b}(a) \propto f_{Y,S}(a,b)\mathbb{1}_A(a,b)/P(A) \propto f_{Y|S=b}(a)\mathbb{1}_A(a,b)/P(A) \text{ i.e.,}$$
$$Y^* \mid (S^* = s^*) \sim Y \mid (S = s^*)\Big| A$$

Ie, the two types of conditioning commute, as one can take select on A then take the conditional density of Y given S, or one can take the conditional density and then select on A. For example, if the observed data (Z^*, S^*) comes from selection on the raw value, $\{Z > cS\}$, then one can compute the density of Z^* as $Z \mid Z > cS^*$.

4. *test intuition* Four testing scenarios: $\theta = 0, \theta \neq 0$ and null/alt ie selection/no selection. The observations are (y_j, σ_j), not necessarily iid [cite above display].

   (a) *null/no selection* 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. According to usual random design OLS theory, the test is consistent under the null, asymptotic normality valid for inference.[given ols assumptions eg variance must exist][do OLS assumptions hold if (x,y) arent necessarily iid, just meet the moment conditions? check wooldridge. consistency proof prob goes through but maybe not asy normality.]

   Begg's test. Must first establish that Begg's actual standardized residuals, $(y - \hat{\mu})/\hat{\sigma}$, is asymptotically equivalent to $(y - \mu)/\sigma$, ie $\sqrt{n}$ times the difference tends to 0. Tried on reverse of p.13, couldnt finish. Once established, can refer to U-statistics asymptotic normality.

   Next, must likely use scale family assumption. Under this assumption, then under the null $\mu = 0$, so the mean of the kernel of the U-statistic is just $P(\{y_1/\sigma_1 < y_2/\sigma_2\}\{\sigma_1 < \sigma_2\}) = P(\{z_1 < z_2\}\{\sigma_1 < \sigma_2\})$. Under scale family assumption, z_j indepdendent of σ_j so the above is $1/4$, as required for consistency.

   Without the scale family assumption, z_1 and z_2 though both mean zero may have different distributions. Conssitency requires the median of $z_1 - z_2$ be 0, which does not [?] follow from the mean being zero and independence.

   (b) *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, $E(u(g_j(y/\sigma_j))v(1/\sigma_j)) = E(E(\dots \mid \sigma_j)) = E(E(u(g_j(y/\sigma_j)) \mid \sigma_j)v(1/\sigma_j)) = E(u(g_j(z)))E(v(1/\sigma_j))$ 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 $E(g(z)) \neq 0$].

If distributions are different, possibly non-iid postselection distribution. Can lead to inconsistent test. #23 in egger.R. would be nice to establish analytically, need estimate of t tails.

If selection mechanisms can vary ...

If $g(z) = 0$ ...

analogous analysis for begg?

Selection present, $\mu \neq 0$. egger regression response is then $\sim (\mu_j/\sigma_j, 1)$. selection will induce larger μ_j and smaller σ_j. what can be said about postselection response? at least in normal case, and simple thresholding as selection mechanism? will set this case aside.

5. *criterion for consistency of egger's test* [this is the scale family situation, assumption needs to be introduced] Focused on the situation where there is selection that can depend on both vector components, eg, $g_j(y_j, \sigma_j)$, but the postselection distribution is the iid. e.g., the preselection (y_j, σ_j) arent just independent but iid, plus the selection stragy is the same $g_j = g$, all j. In this case (other case, not iid, requires entire vector be considered), Egger's test is inconsistent when the plim of $\hat{\beta}_0 = \overline{ys} - \bar{s}\hat{Cov}(ys, s)/\hat{V}(s)$ is 0:

$$E^*(ys)/E^*(s) = Cov(ys, s)/V(s)$$

Letting $s' = s/\sqrt{V^*(s)}$ so that $V^*(s') = V^*(ys) = 1$, rewrite as $Corr^*(ys, s') = E^*(ys)/E(s')$. When s' and ys are independent, get 0 on both sides, this is just the null case. At other extreme, when the conditional outcome mean is the identity, $E^*(ys \mid s') = s'$, get 1 on both sides. [simulation at 24a.]

Suppose the responses are normal [ref] and selection is hard thresholding on the standardized statistic, where the threshold may depend on s. The conditional outcome mean $E(ys \mid s) = \int_{c(s)}^{\infty} z\phi(z)dz/(1 - \Phi(c(s)))$ is the gaussian hazard function $\phi(c(s))/(1 - \Phi(c(s)))$. this being the identity requires the cutoff function $c(s)$ to be the inverse of the gaussian hazard function $\phi/(1 - \Phi)$ [inverse exists, function is convex increasing], a concave increasing selection function. [describe growth rate, should be fast. unlikely this type of selection] The authors of the primary studies would have to be more selective when the studies are larger/have smaller sd.

Not too crazy to have selection increasing in $s = 1/\sigma$, if using selection on raw value rather than p-value. Flat selection on raw value $y_j > c$ ie $z_j\sigma_j > c$ with $z_j = y_j/\sigma_j \sim (0,1)$ is selection increasing in s viewed as selection on the p-vlue, $z_j > c/\sigma_j$.

selection on raw value Setting where given iid (y_j, σ_j) and selecting on $y_j = \sigma_j z_j$. Test is asymptotically null iff $E(sz)/E(z) = E(s^2)/E(s)$ where now s and z are the post-selection distributions. this criterion is obtained by taking plims in $0 = \hat{\beta}_0 = \overline{y} - \overline{x}\hat{\beta}_1$. Provided $E(z) \neq 0$ (ie provided not in the true null case) rewrite as

$$0 = E\left(s(s/E(s) - z/E(z))\right) = E\left(s(s/E(s) - \mu(s)/E(\mu(s)))\right)$$

where $\mu(s) = E(z \mid s)$ is the conditional mean of $z = ys$ given s. The RV in parens, say $u(s)$ is mean zero. If $u(s)$ is monotonic in s, then this is not possible unless $u(s)$ is constant.

Proof. Given: $S > 0$ a.s., $s_2 \geq s_1$ implies $u(s_2) \geq u(s_1)$, and $E(u(S)) = 0$. Then $E(u(S); u(S) < 0) = -E(u(X); u(S) > 0)$, and

$$\begin{aligned} E(Su(S)) &= E(Su(S); u(S) < 0) + E(Su(S); u(S) \geq 0) \\ &\geq E(Su(S); u(S) < 0) + \sup\{s : u(s) < 0\} E(u(S); u(S) \geq 0) \\ &= E\left((S - \sup\{s : u(s) < 0\})u(S); u(S) < 0\}\right) \geq 0, \end{aligned}$$

and $= 0$ iff $S\{u(S) > 0\} = \sup\{s : u(s) < 0\} = S - \sup\{s : u(s) < 0\}$ iff S is consant or $u(S) = 0$ a.s. □

The RV $u(s)$ is degenerate iff the conditional mean $\mu(s)$ is a linear function of s. Very aggressive thresholding, like $c(s)$ described above.

Turn therefore to when $u(s)$ is monotonic. Sufficient condition is that the integral of the surival function $\int_x^\infty (1 - F(y))dy$ of the studies be log concave (egger p15).

Proof. Given $\mu(s) = \int_{cs}^\infty z f_Z(z)dz/(1 - F_Z(cs))$. Then $\mu(s) > cs$ [why strict], $E\mu(s) > cEs$. Then

$$\begin{aligned} \frac{\partial}{\partial s}\left(\frac{s}{Es} - \frac{\mu(s)}{E\mu(s)}\right) &= \frac{1}{Es} - \frac{\mu'(s)}{E\mu(s)} \\ &> (Es)^{-1}\left(1 - c^{-1}\mu'(s)\right). \end{aligned}$$

Substitute

$$\begin{aligned} \mu'(s) &= c\frac{-csf_Z(cs)}{1 - F_Z(cs)} + \frac{\left(\int_{cs} z f_Z(z)dz\right)(cf_Z(cs))}{(1 - F_Z(cs))^2} \\ &= \frac{cf_Z(cs)}{1 - F_Z(cs)}(\mu(s) - cs) \end{aligned}$$

to get

$$\frac{1}{Es}\left(1-\frac{f_Z(cs)}{1-F_Z(cs)}(\mu(s)-cs)\right).$$

Let $x=cs$, then enough to show

$$f_Z(x)\left(\frac{\int_x^\infty z f_Z(z)dz}{1-F_Z(x)}-x\right)<1-F(x).$$

Integrating by parts, $\int_x^\infty z f_Z(z)dz - x + xF_Z(x) = \int_x^\infty (1-F_Z(z))dz$, provided that $x(1-F_Z(x))\to 0$ as $x\to\infty$. The latter is implied by $V(X) = cV(1/\sigma) < \infty$. Substituting, the condition becomes $\frac{\int_x (1-F_Z(z))dz}{1-F_Z(x)} < \frac{1-F_Z(x)}{f_Z(x)}$, $\frac{\partial}{\partial x}\log\int_x(1-F_Z(z))dz < \frac{\partial}{\partial x}\log(1-F_Z(x))$, $\frac{\partial}{\partial x}\log\frac{1-F_Z(x)}{\int_x(1-F_Z(z))dz} > 0$, implying since the argument to log is nonnegative

$$\begin{aligned}0 &< \frac{\partial}{\partial x}\frac{1-F_Z(x)}{\int_x^\infty(1-F_Z(z))dz}\\ &= -\frac{\partial^2}{\partial x^2}\log\int_x^\infty(1-F_Z(z))dz.\end{aligned}$$

□

Log-concavity of the tail integral of $1-F(x)$ is implied by log concavity of $1-F(x)$, in turn implied by log concavity of the density f [Bagnoli Thrm 2. try to find reference in marshall olkin or elsewhere]. This includes many commonly used distributions. A commonly used distribution it does not include is the pareto [egger.R #26; bagnoli notes]. As for converse, proof shows $\mu' < 1$ iff log concave right integral. If $\mu' < 1$ condition doesnt hold, monotonicity of u will depend on joint distribution, give power law example.

(12/5) 1. The sufficienct for inconsistency of egger's test with thresholding on the raw value is, for some s, $0 = u'(s) = \frac{d}{ds}\frac{(1-F_Z(cs))^\alpha}{\int_{cs}^\infty(1-F_Z(z))dz}$ where $\alpha = cE(S)/E(\mu_c(s))$ is ≤ 1. The monotonicity condition is probably sufficient as well, in the sense that there is a nondegenerate distribution of $s=1/\sigma$, such that $E(Su(S))=0$ if $u'(s)=0$ for some s.

2. From the monotonicity condition $u'(s)>0$ or $u'(s)<0$, gronwall's condition gives necessary conditions that must be satisfied in order that egger's test be consistent for raw thresholding.

4. For distributions of S with a mean given by c, m, $u'(s)=0$ on an interval when the outcome distribution follows a power law with exponent m and thresholding on the raw value at c. This seems to be the only type of distribution where $u'(s)$ vanishes on the interval, by solving the diff eq. (egger 16). But the criterion for inconsistency can be met without u vanishing on an interval.

6. *consistency of begg test*

3. Begg test well motivated by power. Whether selecting on raw value or p-value, clear trend of y against σ.

(12/2) a. Begg test is inconsistent for iid (y, σ) iff $E(\tilde{s}(\tilde{y} - E(\tilde{y}))) = 0$. The analogous criterion for the egger test is $E(\tilde{s}(\tilde{y} - E(\tilde{y})/E(\tilde{s})\tilde{s}))$ [but before this was z not s–check]. b. The density of $\tilde{y}$ given $\tilde{\sigma} = s$ is the same as the density of y given $\sigma = s$ given the selection event (egger 16). c. The begg test is consistent when thresholding on the raw value ie $\{y > c\}$ or p-value $\{y/\sigma > c\}$ (egger 16), ie the criterion in a) is never met.

Given a slope family F_Z for the outcome distribution,

$$y \mid \sigma \sim F_Z(y/\sigma)$$

Then $y/\sigma = ys$ is independent of s, as $y/\sigma \mid \sigma \sim f_Z$ whatever σ. Egger and begg tests sufficient statistics are similar, $(y^*/\sigma^*, \sigma^*) = (z^*, \sigma^*)$.

Consistency under p-value selection. Given iid observations from a mean zero scale family $y \mid \sigma \sim F_Z(y/\sigma)$, suppose selection event A is on the the standardized measurement $z = y/\sigma$ alone. Then the joint post selection density factors,

$$f_{z^*,\sigma^*}(a, b) = \frac{f_{z,\sigma}(a, b)\mathbb{1}\{a \in A\}}{P(A)} = f_Z(a)\frac{\{a \in A\}}{P(A)}f_\sigma(b)$$

and so z^* and s^* are independent. Moreover, integrating out a in [above display] shows that $s^* \sim s$, the distribution of σ is unchanged by the selection event.

Begg's test $\sqrt{9n/4}|\tau| > z_{1-\alpha}$ is consistent under the alternative, ie rejects with probability 1, if the mean of the kernel of Kendall's τ, viewed as a U-statistic, ie the concordance probability [relate below to beggs actual statistic]

$$P\left(\left\{\frac{y_1^* - \mu}{\sigma_1^*} < \frac{y_1^* - \mu}{\sigma_1^*}\right\} \cap \{\sigma_1 \ < \sigma_2\}\right)$$

tends in probability to a value other than $1/4$ as $n \to \infty$. This holds in light of the preceding paragraph since

$$P\left(\{z_1^* - \mu/\sigma_1^* < z_2^* - \mu/\sigma_2^*\}\{\sigma_1^* < \sigma_2^*\}\right) = P(\sigma_1^* < \sigma_2^*)E(F_{z_1^* - z_2^*}(\mu(1/\sigma_2^* - 1/\sigma_1^*))\{\sigma_1^* < \sigma_2^*\})$$

$$= P(\sigma_1 < \sigma_2)E(F_{z_1^* - z_2^*}(\mu(1/\sigma_2 - 1/\sigma_1))\{\sigma_1 < \sigma_2\}).$$

Since σ_1, σ_2 are iid, $P(\sigma_1 < \sigma_2) = 1/2$, since z_1^*, z_2^* are iid, $F_{z_1^* - z_2^*}$ is symmetric about 0, and on $\{\sigma_1 < \sigma_2\}$, $1/\sigma_2 - 1/\sigma_1 < 0$. Then above

is $< 1/4$ so long as $\mu = E(z^*) > 0$. In particular, for p-val thresholding, where $(1 - F_Z(c))E(z^*) = \int_c^\infty z f_Z(z)dz = -\left(\int_{-\infty}^c z f_Z(z)dz\right) > \int_{-\infty}^\infty z f_Z(z)dz = 0$, assuming c is in the interior of the support of Z.

Consistency under raw selection. Suppose selection event A is a function of both z and σ. Then no longer the case that z^* and σ^* are independent or that s^*igma is distributed as σ. Instead we only have

$$f_{z^*,s^*}(a,b) = \frac{f_{z,s}(a,b)\{(a,b) \in A\}}{P(A)} = \frac{f_z(a)f_s(b)\{(a,b) \in A\}}{P(A)}$$

Suppose the postselection measurement z^* is monotonic in σ^*...[got stuck, sim 36 suggest might not always be consitent...]

7. *slope of egger and begg test, p-val thresholding* (12/30) Verified by simulation (egger #34) the slope of begg test is $\mu'(0)/\sigma(0) = \int_\infty^\infty f_Z^2(z)dz * E(s_1 - s2; s1 < s2)/(\sqrt{4/9}$. Local power function approximation is not too bad for uniform and normal f_Z. (1/1) can rewrite $E(s_1 - s2; s1 < s2)$ as $(1/2) * E(|s_1 - s_2|)$; maybe relate to mean absolute deviation?

   (1/1) It is perhaps to be expected that the power to detect a trend depends on the dispersion of σ relative to the dispersion of y. Oddly egger test/p-value thresholding power depends on location of σ distribution. The power curve will be better or worse than begg's depending on this location. Is it also odd that the power of begg's test depends on the dispersion of σ (not relative to that of z

   (1/7) test slope for egger test under raw thresholding see egger p 20

8. *slope, raw thresholding* (1/7) test slope for egger test under raw thresholding see egger p 20

9. *discussion* random truncation model though. must find a parametrization though. exponential tilting.

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"] ,

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

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 $h/\sqrt{V(Z)E(S^2)/V(S)}$.

Proof. [0. formula for betahat] 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 $\hat{\beta}_0 = (\hat{V}(s))^{-1} \sum_{j=1}^n (\overline{s^2} - \overline{s} s_j) z_j$, where $\hat{V}(s) = n^{-1} \sum_j (s_j - \overline{s})^2$. The variance estimate is $\hat{V}(\hat{\beta}_0) = \overline{s^2} \hat{V}(s) RSS/(n-2)$, where $RSS = ||(I - X(X^tX)^{-1}X^t)\zeta_n||^2$. So the test statistic is

$$\frac{\hat{\beta}_0 = (\hat{V}(s))^{-1} \sum_{j=1}^n (\overline{s^2} - \overline{s} s_j) z_j}{\overline{s^2} \hat{V}(s) RSS/(n-2)}.$$

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

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

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:

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

where convergence is in probability under the sequence of laws P_n of (Z^*, S^*) at θ_n.

First term: $\sqrt{n} \left(E(S^2) - \overline{S^2} \right)$ is $O_{P_n}(1)$ by the CLT, also distribution of S does not change with P_n. [need $V(S) < \infty$] 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: $RSS/n = \zeta_n^t (I - X(X^tX)^{-1}X^t)\zeta_n/n \to_{P_n} V(Z)$, 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 $RSS/n = o_{P_n}(1) + n^{-1}\zeta_n^t\zeta_n$, 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:

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

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

$$E_n\left(\left(\frac{(E(S^2) - E(S)S_1)Z_1^*}{V(S)} - E_n(Z^*)\right)^2\right) = \frac{E_n((Z^*)^2)((E(S^2))^2 - E(S^2)(E(S))^2)}{V(S)^2} - (E_n(Z^*))^2$$

$$= E_n((Z^*)^2)E(S^2)/V(S) - (E_n(Z^*))^2$$

$$\to \frac{V(Z)}{V(S)}E(S^2)$$

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

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

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

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

□