Eco311 Optional Reading: Type I Error (TIE) and Publication Bias

(Jing Li, Miami University)

1. Recall: in class we learn that two variables can be correlated thanks to an omitted variable (lurking variable, confounder). One extreme case is that two variables are independent or unrelated (no causation at all), but because of omitted variable bias, the regression may falsely indicate the two variables are related.

2. One example is that a regression may imply the number of shark attacks is related to sales of ice cream. That regression is spurious because there is no direct link between the two variables. They are shown by the regression to be correlated just because both are related to temperature. The misleading correlation between shark attack and ice cream sale disappears once temperature is included (controlled) in the regression.

3. The goal of this note is to show that we can get spurious result for a different reason—if we run too many regressions, a few of them would indicate two variables are related even though they are actually unrelated. Unfortunately, those few eye-catching results (autism is related to MMR vaccine, eating ear wax helps sleep, Dr. Li attacks a dog, you name it) are more likely to be published or reported by social media than other more common results (autism is unrelated to vaccine, eating ear wax does not help sleep, a dog attacks Dr. Li).

4. In this note, the cause of spurious results is not an omitted variable, but type I error (TIE)

5. Simply put, type I error happens when a correct null hypothesis is rejected (the test statistics can end up in the rejection zone, or in either tail of the bell, even if null hypothesis is true). We choose to use 1.96 as the critical value so that the likelihood of TIE is as small as 0.05. But 0.05 is not zero. That means still there is a small chance we reject a null hypothesis that should not be rejected. This situation is similar to that a judge may mistakenly send an innocent person to jail.

6. For a simple regression model

$$y = \beta_0 + \beta_1 x + e,$$

the default null hypothesis is that x and y are unrelated (x does not matter for y), or

$$H_0 : \beta_1 = 0.$$

Type I error means that when we run too many regressions using unrelated y and x, about 5 percent of those regressions would produce significant t value, which (mistakenly) rejects H_0 and (mistakenly) implies that y and x are related. Unfortunately, because they are often eye-catching, those 5 percent studies are more likely to be published than other 95 percent studies. This undesirable situation is called publication bias. The fact is, most editors prefer publishing results that are statistically significant.

7. We use simulation to illustrate TIE in the regression setting

```
> set.seed(12345)
> n = 1000
> iter = 100
> v.t = rep(0, iter)
> for (i in 1:iter) {
+ y = rnorm(n)
+ x = rnorm(n)
+ co = coef(summary(lm(y~x)))
+ v.t[i] = co[2, "t value"]
+ if (abs(v.t[i])>1.96) {cat("regression id is ", i, ", t value is", v.t[i], "\n")}
+ }
regression id is  8 , t value is 1.989501
regression id is  22 , t value is 2.192319
regression id is  26 , t value is 2.393193
regression id is  40 , t value is -2.350241
regression id is  56 , t value is 2.185919
regression id is  77 , t value is -2.009841
> sum(abs(v.t)>1.96)/iter
[1] 0.06
```

Notice that by construction y and x are unrelated. In the ideal world, all regressions should indicate they are indeed unrelated.

8. However, in reality, when we run 100 regressions using those unrelated y and x, 6 out of the 100 regressions imply y and x are related since their t values exceed 1.96 in absolute value (and therefore reject the correct null hypothesis that they are unrelated).

9. I believe that infamous study relating autism to MMR vaccine belongs to those few spurious regressions. That author finds a statistically significant correlation just by chance. Sadly, that study got published, and received too much unwarranted attention.

10. If too many researchers try to find relationship between autism and vaccine, there must be some of them who can find a correlation and get published. Another way to look at this issue is, if there are many autism patients, there must be some samples and everyone in the sample has received MMR vaccine.

11. To guard against those nonsense results, you have to ask "what is the mechanism through which vaccine can cause autism". Always keep in mind,

    `correlation is not causation`

    `correlation can be found by chance`

12. Next time you read something "revolutionary" on social media, first ask,

    `Is that type I error?`

    `Is that correlation or causation?`

    `Is there any mechanism or story behind the statistics?`