Perils of Peer Effects Write Up and Simulation

Jacob McGill

2024-05-23

Angrist's "The Perils of Peer Effects" (2014) discusses potential pitfalls when identifying the causal effects of peer characteristics on outcomes. This write up briefly summarizes the potential issues discussed, Angrist's proposed solutions, and provides simulations to highlight both.

Linear In-Means

Angrist first discusses a linear in means model of peer effects. The most basic example of this equation takes the form $y_{ij} = \beta_0 + \beta_1 \bar{y}_j + \epsilon_{ij}$, for individual i in group j, with $\bar{y}_j$ being the mean value of y for group j. However, this equation always produces a β_1 with a value of 1. This can be seen with the following simulation. In this simulation, 3 groups of equal size were assigned a value of "1" for the variable "Yes" at different probabilities (so members of Group A had a probability of 0.5 of being assigned 1, group B had a proability of 0.2, etc.). The mean value of "Yes" for each each group was then calculated and assigned to members of the groups, recorded as "rate". "Yes" was then regressed against "rate" to produce the results below:

```
##
## Call:
## lm(formula = yes ~ rate, data = combined_groups)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8001 -0.1997 -0.1997  0.1999  0.8003
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.226e-13  1.807e-03     0.0        1
## rate        1.000e+00  3.246e-03   308.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4358 on 299998 degrees of freedom
## Multiple R-squared:  0.2403, Adjusted R-squared:  0.2403
## F-statistic: 9.492e+04 on 1 and 299998 DF,  p-value: < 2.2e-16
```

This regression produces a coefficient of 1, as predicted by Angrist.

A more nuanced model controls for both individual and peer characteristics, such as the model $y = \beta\mu_{(y|z)} + \gamma x$, where y is the outcome of interest, x is an individual covariate, and z is group membership. This model attempts to attempt identify the peer effect of the mean of y in group z on an individual's y, holding z constant. This equation can be rewritten to the form $E[y|z] = \frac{\gamma}{(1-\beta)} E[x|z]$, meaning β in this equation acts

as a "social multiplier" that magnifies the effect of individual covariate changes. Angrist shows that multiplier $\frac{1}{1-\beta}$ approximately equals the ratio of the 2SLS to OLS estimand of the effect of individual covariate x on y (with being instrumented by group membership). As a result, β just captures the divergence between the OLS and 2SLS estimates of the effect of x on y, which may exist for reasons other than peer effects, such as weak instruments in 2SLS that do not strongly affect the first stage. To demonstrate how this can occur, I will simulate the Dartmouth high school drinking paper Angrist discusses. In this example, I randomly generated a dummy variable "high_school", indicating whether that individual drank in high school. I used "high_school" to determine the probability of whether someone joined Greek life to capture the influence of high school drinking on joining Greek life, marked by the indicator variable "greek". In this simulation, there are no peer effects on high school drinking. Individuals were then assigned to groups of "dorm", "floor", and "room", with each grouping getting more and more coarse. I then ran an OLS regression of "high_school" against "Greek". I then ran 3 more 2SLS regressions of "high_school" against ""Greek", instrumenting "high_school" with with group membership. The results of these regressions are in the table below, with the coefficients of OLS, 2SLS, their ratio, and their :

```
##          OLS Reg      OLS SE  2SLS Reg   2SLS SE     Ratio
## Dorms  0.2027697 0.009792518 0.2190160 0.8024209 1.0801221
## Floors 0.2027697 0.009792518 0.1609068 0.1798960 0.7935449
## Rooms  0.2027697 0.009792518 0.2085952 0.0376633 1.0287300
```

Here, we see results consistent with what Angrist predicted. As groupings becoming more coarse (moving from rooms to floors to dorms), the ratio of the 2SLS to OLS coefficients increases while the standard error of the 2SLS regression increases. The regression model captures peer effects even though they do not exist in the data generating process.

Leave Out-Mean and Social Returns

Angrist then discusses the leave-out mean model of peer effects, which take the form $y_{ij} = \beta_0 + \beta_1 \bar{y}_{(i)j} + \epsilon_{ij}$ for individual i in group j. Although this regression does not automatically produce $\beta_1 = 1$ (as the $y_{ij} = \beta_0 + \beta_1 \bar{y}_j + \epsilon_{ij}$ did), Angrist argues it does not provide a causal interpretation of peer characteristics, as it just captures intraclass correlation, such as shocks common to groups.We can see this in the example below. Data was generated for a population normally (referred to as "characteristic"), with a mean of 50 and a standard deviation of 1. The observations were grouped into 150 groups of 4 categories and each group received a random "shock". The "shock" and the observation's "characteristic" value were then summed together to create the value "characteristic_mod". The leave out mean for "characteristic_mod" was then calculated for each individual and regressed against "characteristic_mod".

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 32.974675   3.212739 10.2637 < 2.2e-16 ***
## leave_out    0.400804   0.058375  6.8661 1.657e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As can be seen, the regression captures a statistically significant effect of leave out mean. However, this is not capturing any causal effect of leave out mean (since it does not exist in the data generating process). Instead it is capturing intraclass correlation caused by group specific shocks.

Angrist also discusses flaws with the social returns model, $y = \beta_1 \mu_{(x|z)} + \beta_0 x$, which is intended to capture the causal effect of the average value of x for group z on outcome y, controlling for an individual's value of

x. The social returns coefficient, β_1, is proportional to the difference between the 2SLS estimated coefficient of x from of regressing x on y (instrumenting for x with z) and the OLS estimate of the coefficient of x. β_1 then does not necessarily capture the causal effect of $\mu_{(x|z)}$ on y, as the 2SLS estimate can be greater than the OLS estimate for reasons such as measurement error or omitted variable bias. How these can effect 2SLS can be seen in the simulation below (I based it off the Acemoglu and Angrist paper given as an example in the Angrist paper). Data was generated for the "educ" variable and observations were grouped into categories of 10. A "state_effect" was also calculated. The variable "wages_state" was determined by the equation "wage_state = 200 + 25*(educ_state) + 20*state_effect", with "educ_state" being the sum of educ and state_effect. The influence of "state_effect" on wages acts as an omitted variable bias by violating the exclusion of principle of using states as an instrument.

```
##
## Call:
## lm(formula = wage_state ~ educ_state, data = return_groups)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.2847  -5.5934  -0.5476   5.2710  19.2529
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 178.1276     2.2148   80.43   <2e-16 ***
## educ_state   27.7384     0.1359  204.06   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.263 on 998 degrees of freedom
## Multiple R-squared:  0.9766, Adjusted R-squared:  0.9766
## F-statistic: 4.164e+04 on 1 and 998 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## ivreg(formula = wage_state ~ educ_state | group_number, data = return_groups)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -58.7519  -9.9072   0.3483  10.1919  49.7655
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   46.744     40.411   1.157    0.248
## educ_state    35.846      2.493  14.376   <2e-16 ***
##
## Diagnostic tests:
##                  df1 df2 statistic  p-value
## Weak instruments   1 998     13.72 0.000223 ***
## Wu-Hausman         1 997     51.39 1.47e-12 ***
## Sargan             0  NA        NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.52 on 998 degrees of freedom
## Multiple R-Squared: 0.8932,  Adjusted R-squared: 0.8931
## Wald test: 206.7 on 1 and 998 DF,  p-value: < 2.2e-16
```

```
## [1] 8.107327
```

As can be seen, the failure of the state instrument to satisfy the exclusion restriction results in the 2SLS estimate of the effect of education on wages to be bigger than the OLS estimate, creating the appearance of peer effects. To demonstrate the effect of measurement error, I also generated the variable "wages" from the equation "wage = 200 + 25*(educ)" and the variable "educ_noise", which adds some randomly generated noise of mean 0 to the "educ" variable. Regressing wage on "educ_noise" instead of "educ" will attenuate the OLS and 2SLS measures of the effect of education on wage, but the attenuation will be greater for OLS compared to 2SLS, creating a gap between the 2 coefficients and the appearance of peer effects. This effect can be seen below.

```
##
## Call:
## lm(formula = wage ~ educ_noise, data = return_groups)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -75.141 -18.022  -0.247  17.403  95.184
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 384.0671     5.6525   67.95   <2e-16 ***
## educ_noise   11.7210     0.3977   29.47   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.08 on 998 degrees of freedom
## Multiple R-squared:  0.4653, Adjusted R-squared:  0.4648
## F-statistic: 868.6 on 1 and 998 DF,  p-value: < 2.2e-16

##
## Call:
## ivreg(formula = wage ~ educ_noise | group_number, data = return_groups)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -145.8617  -31.6472   -0.7858   31.3338  180.8535
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  648.957    357.574   1.815   0.0698 .
## educ_noise    -7.118     25.430  -0.280   0.7796
##
## Diagnostic tests:
##                  df1 df2 statistic p-value
## Weak instruments   1 998     0.794   0.373
## Wu-Hausman         1 997     1.786   0.182
## Sargan             0  NA        NA      NA
##
## Residual standard error: 47.01 on 998 degrees of freedom
## Multiple R-Squared: -0.7368, Adjusted R-squared: -0.7385
## Wald test: 0.07834 on 1 and 998 DF,  p-value: 0.7796

## [1] -18.83866
```

The estimated 2SLS estimate of education is larger than the OLS estimate, creating the appearance of peer effects

Solutions

Angrist discusses 2 potential methods to properly identify peer effects. The first requires distinguishing between subjects of peer effects and the peers themselves. Although this approach only captures the effect of peer group manipulation, it eliminates the link between individual characteristics and group characteristics, negating the need to control for individual characteristics. Angrist cites a study examining the effectiveness of housing vouchers, which looked at peer effects on individual outcomes when moving to low poverty neighborhoods. Since the vouchers were randomly assigned, there was no need to control for individual characteristics in the regression. I simulated this in the below regression. In this model "peer_char" is the average of the peer characteristics that a group is assigned too. There is also an individual characteristic that drives wages. I then generated 2 wages, peer_wage and no_peer_wage, to capture a situation where peer effects do not exist (the first regression) and one where they do (the second regression).

```
##
## t test of coefficients:
##
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 2.4999e+02 4.5825e-01 545.5187   <2e-16 ***
## peer_char   5.2656e-04 4.2731e-02   0.0123   0.9902
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 249.986491   0.458255 545.519 < 2.2e-16 ***
## peer_char     0.500527   0.042731  11.713 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As can be seen, this approach captures peer effects when they exist (as the results are statistically significant and close to the actual effect in the model) and does not when they do not exist (as the coefficient on peer effects in the first regression are not statistically significant) and when they do exist (as the results are statistically significant and close to the true effect of 0.5, as seen in the second regression).

The other proposed solution is the "no peer effects" null hypothesis, where OLS and 2SLS parameters are expected to produce the same results if peer effects do not exist. Angrist suggests accomplishing this by using random assignment to create a strong first stage for peer characteristics but ensuring OLS and IV estimates of own effects are the same under the no peer null hypothesis. The example he gives is a job training study that randomly assigned treatment proportions for job search assistance to different labor markets in France. The social returns model for this equation took the form $y_{ic} = \mu + \pi_1 p_c + \pi_0 t_{ic} + v_{ic}$, with t_{ic} being treatment status for individual i in labor market c and p_c the proportion of job hunters receiving aid in job market c. As this experiment does not have measurement error (assuming wages are not self reported) or omitted variable bias (as the instrument, proportion of job aid, is randomly assigned and not correlated with having a job), OLS and 2SLS are not expected to diverge unless peer effects exist. I will demonstrate the effectiveness of this approach with a simulation. 235 "labor markets" were randomly assigned the following proportions of "job aid": (0, 0.25, 0.5, 0.75, 1). Individuals in each labor market had that probability of receving job aid, which increased their probability of being hired from 0.25 to 0.35. In this process, peer effects do no exist.

1000 individuals were generated for each labor market. Receiving job assistance (indicated by a dummy variable) was regressed against job status for both the OLS and 2SLS regressions, with the latter using city to instrument for job assistance start simulating this approach by replicating the data generating process, with a a strong first and second stage, but with no peer effects.

```
##
## Call:
## lm(formula = job_no_peer ~ job_assist, data = city_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3505 -0.3505 -0.2489  0.6495  0.7511
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.248885   0.001320  188.52   <2e-16 ***
## job_assist  0.101607   0.001878   54.12   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.455 on 234998 degrees of freedom
## Multiple R-squared:  0.01231,    Adjusted R-squared:  0.01231
## F-statistic:  2929 on 1 and 234998 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = job_no_peer ~ assist_rate, data = city_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3515 -0.3256 -0.2737  0.6485  0.7522
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.247825   0.001585  156.31   <2e-16 ***
## assist_rate 0.103700   0.002579   40.21   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4563 on 234998 degrees of freedom
## Multiple R-squared:  0.006833,   Adjusted R-squared:  0.006829
## F-statistic:  1617 on 1 and 234998 DF,  p-value: < 2.2e-16
```

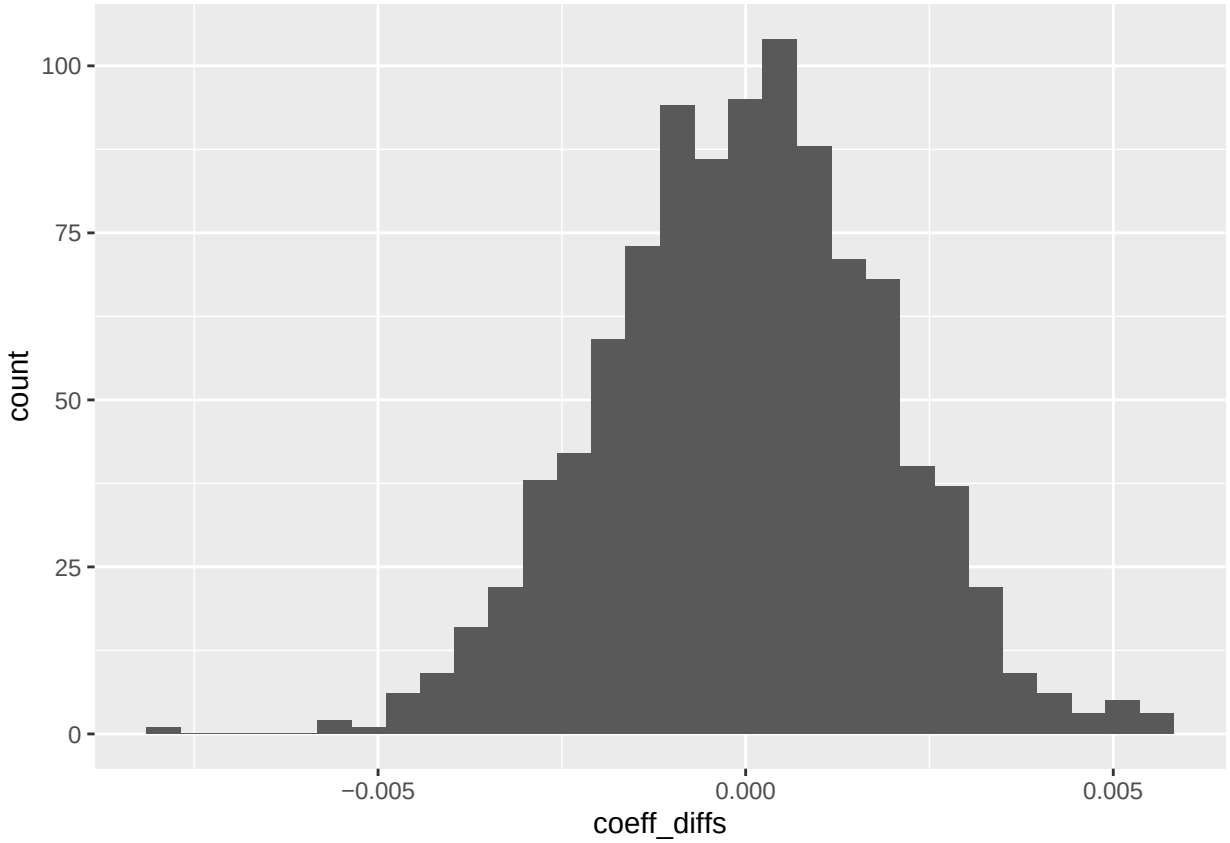
As can be the seen, the coefficients for the OLS regression on job assistance (the first regression) and the 2SLS regression (the second regression) appear to be pretty similar. Testing against the null hypothesis the two coefficients are equal (or alternatively their difference is 0) produces the following z-score.

```
## [1] 0.6560868
```

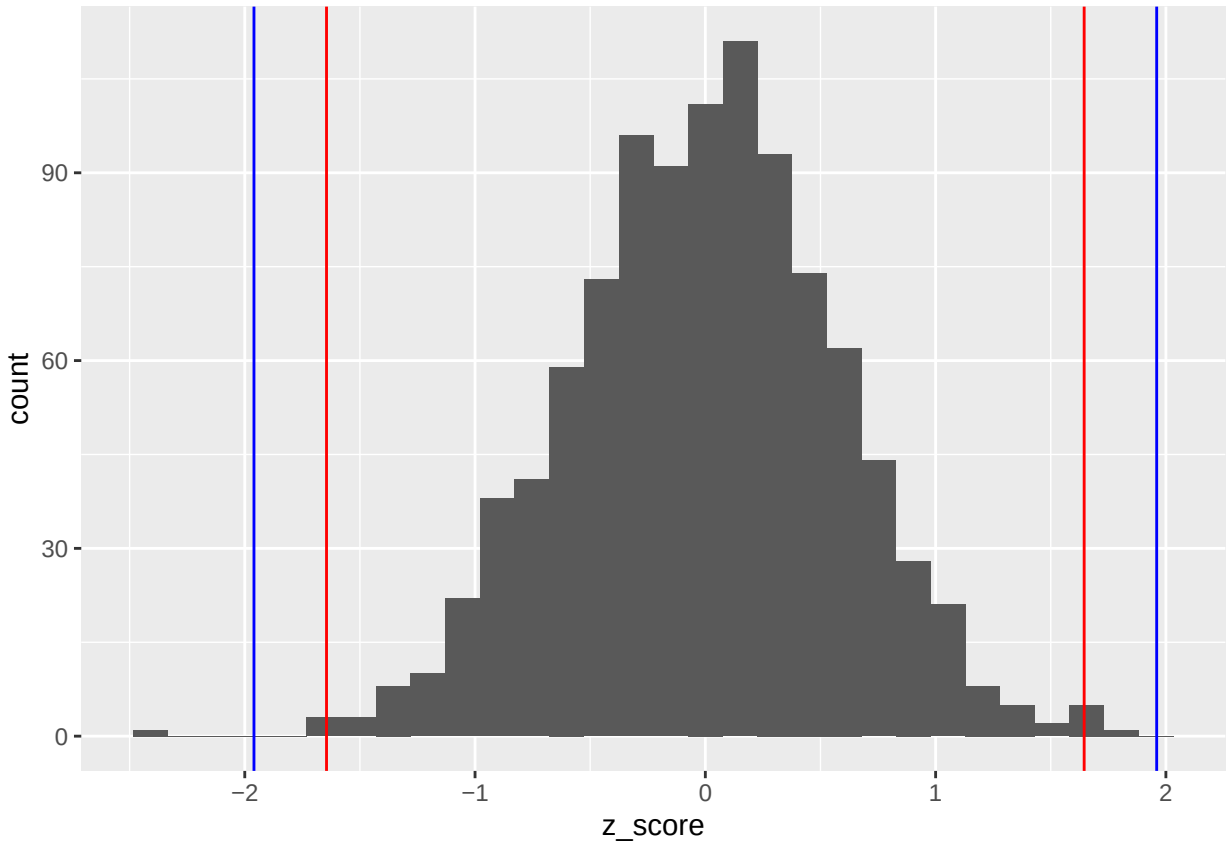
We would then fail to reject a null hypothesis that the OLS and 2SLS parameters have a difference that is not equal to 0 at the 5% level. To test the robustness of this approach, I also conducted a Monte Carlo simulation. The monte carlo simulation repeated the above data generation and regressions 1000 times. The difference in coefficients between the 2SLS and OLS regressions and the z-scores of the difference in

coefficients against a null hypothesis of equaling 0 are also plotted on histograms. The z-score histograms also include a blue vertical line at 1.96 and -1.96 (indicating statistical significance at 5%) and a red vertical line for 1.645 and -1.645 (indicating statistical significance at the 10% level).

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



As can be seen in the histogram of coefficient difference, most of the coefficients in the no peer null have a difference no greater than 0.005. Further, as seen in the histogram of the z-scores of the coefficient differences, the majority of the estimated coefficients do not have a difference that is statistically different from 0, indicating that this model satisfies the no peer null assumption.