

AE 8 - Simpsons Paradox

Brandon Leslie

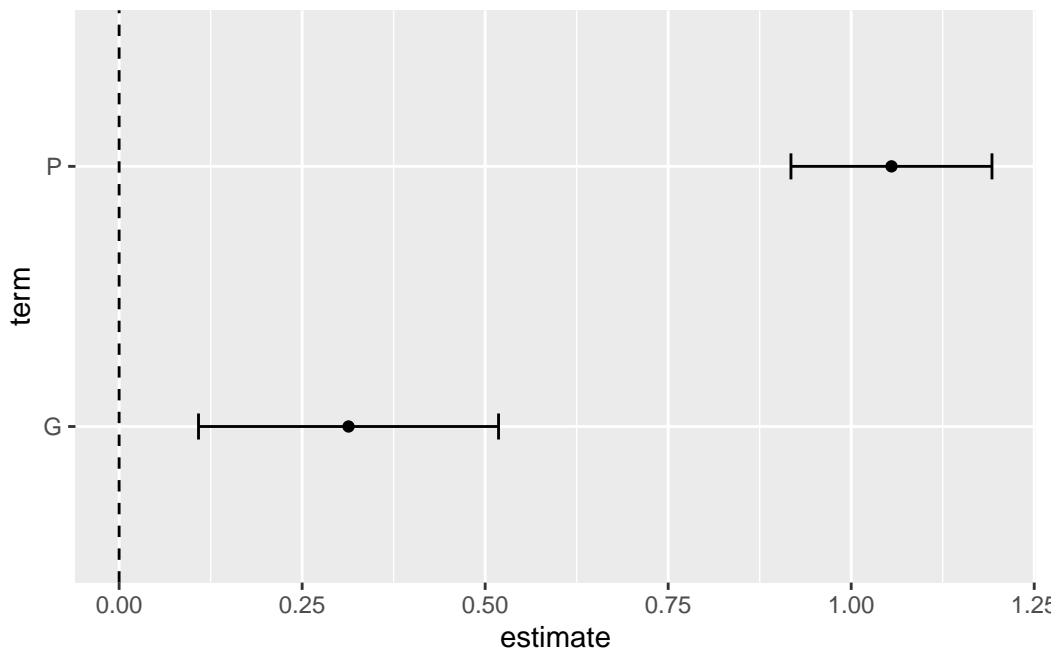
```
#Grandparents, simulated data
```

```
N <- 200
b_GP <- 1
b_GC <- 0.3
b_PC <- 1
b_U <- 2

### Now with positive effect Grandparents, no U
U <- 2*rbinom(N, 1, prob = 0.5) - 1
G <- rnorm(N)
P <- rnorm(N, b_GP*G)
C <- rnorm(N, b_PC*P + b_GC*G)

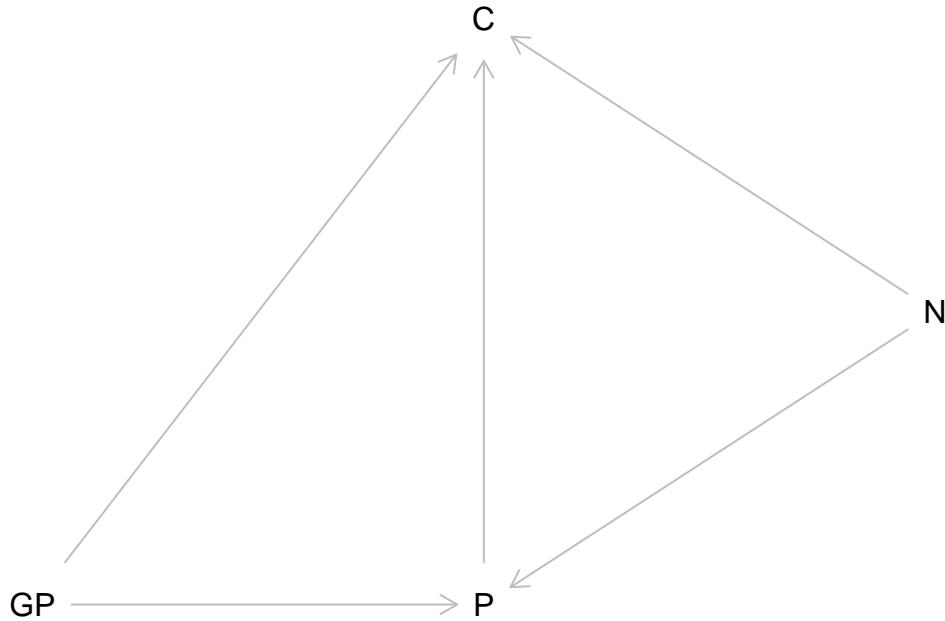
df <- data.frame(C = C, P = P, G = G, U = U)

lm(C ~ P + G, df ) %>%
  tidy() %>%
  filter(term != "(Intercept)") %>%
  ggplot(aes(x = estimate, y = term)) +
  geom_point() +
  geom_errorbar(aes(xmin = estimate - 2*std.error,
                     xmax = estimate + 2*std.error),
                width = 0.1) +
  geom_vline(xintercept = 0, linetype = "dashed")
```



Unmeasured Neighborhood Affect

```
dag1 <- dagitty("dag{
  GP -> P;
  GP -> C;
  P -> C ;
  N -> P;
  N -> C}"
)
coordinates( dag1 ) <-
  list( x = c(GP = -1, P = 0, C = 0, N = 1),
       y = c(GP = 2, P= 2, C = 0 , N = 1))
plot(dag1)
```

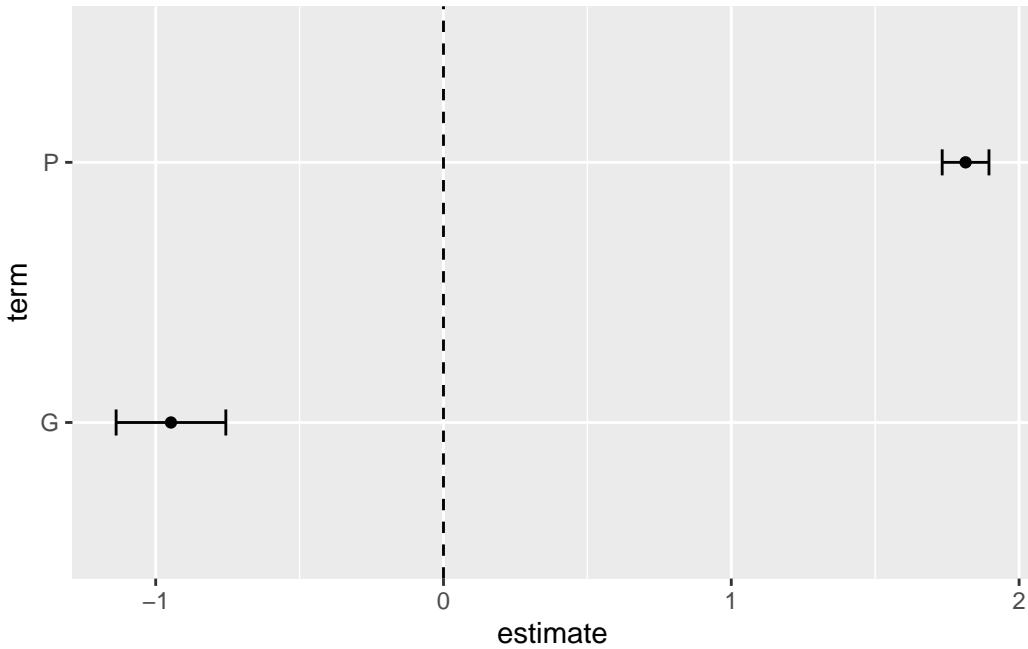


```

U <- 2*rbinom(N, 1, prob = 0.5) - 1
G <- rnorm(N)
P <- rnorm(N, b_GP*G + b_U*U)
C <- rnorm(N, b_PC*P + b_GC + b_U*U )

df <- data.frame(C = C, P = P, G = G, U = U)

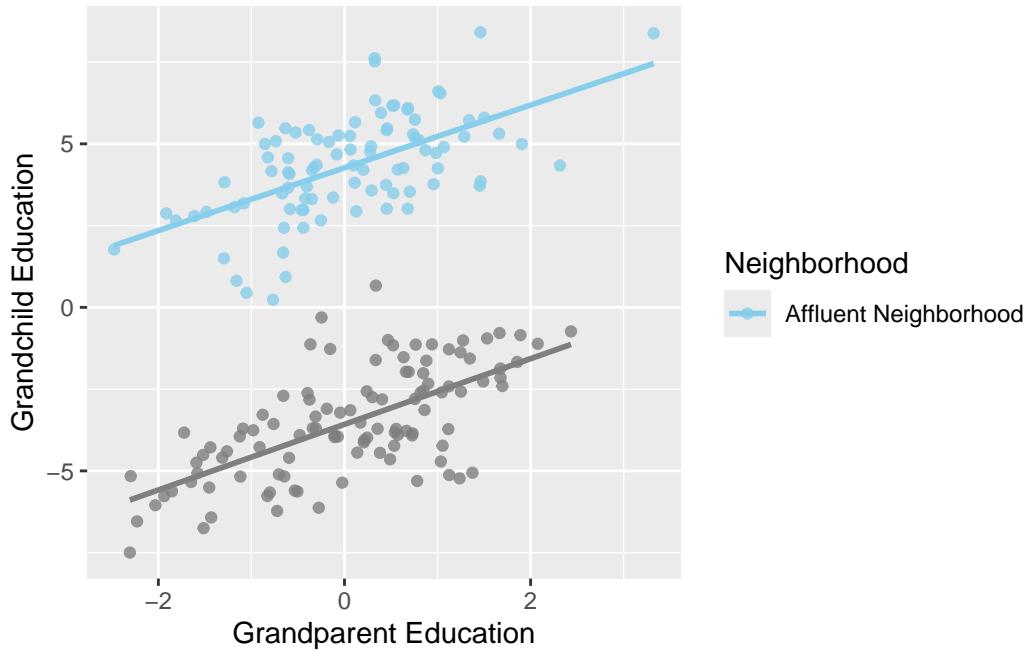
lm(C ~ P + G, df) %>%
  tidy() %>%
  filter(term != "(Intercept)") %>%
  ggplot(aes(x = estimate, y = term)) +
  geom_point() +
  geom_errorbar(aes(xmin = estimate - 2*std.error,
                     xmax = estimate + 2*std.error),
                width = 0.1) +
  geom_vline(xintercept = 0, linetype = "dashed")
    
```



```
df$Neighborhood <- ifelse(df$U > 0, "Affluent Neighborhood", "Poor Neighborhood")
```

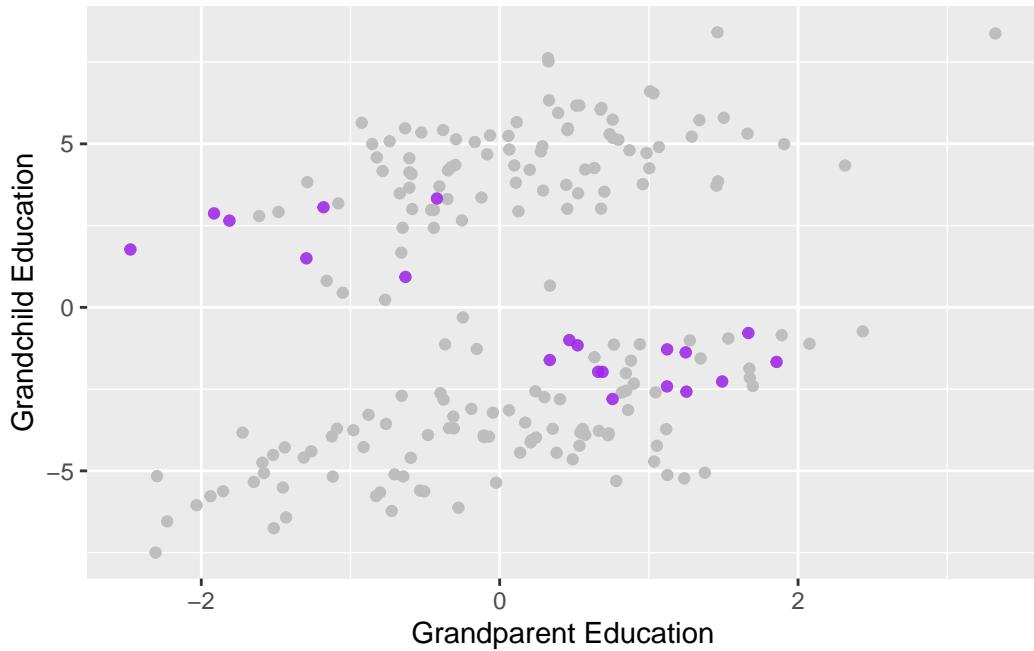
```
ggplot(df, aes(x = G, y = C, color = Neighborhood)) +
  geom_point(alpha = 0.8) +
  geom_smooth(method = "lm", se = FALSE, aes(group = Neighborhood)) +
  scale_color_manual(values = c(
    "Affluent Neighborhood" = "skyblue",
    "Poor Neighborhood" = "red")) +
  labs(
    x = "Grandparent Education",
    y = "Grandchild Education"
  )
```

```
`geom_smooth()` using formula = 'y ~ x'
```



```
P_percentiles <- quantile(df$P, probs = c(0.45, 0.55))
df$Selected <- ifelse(df$P >= P_percentiles[1] & df$P <= P_percentiles[2], "Selected", "not")

ggplot(df, aes(x = G, y = C)) +
  geom_point(color = "grey") +
  geom_point(data = df %>% filter(Selected == "Selected"), color = "purple", alpha = 0.8) +
  labs(
    x = "Grandparent Education",
    y = "Grandchild Education")
```



```

ggplot(df, aes( x = G, y = C)) +
  geom_point(color = "grey", alpha = 0.6) +
  geom_point(data = df %>% filter(Selected == "Selected"), color = "purple", alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE, aes(group = Neighborhood), color = "grey") +
  geom_smooth(data = df %>% filter(Selected == "Selected"), aes(x = G, y = C), method = "lm")
  labs(
    x = "Grandparent Education",
    y = "Grandchild Education")

`geom_smooth()` using formula = 'y ~ x'
`geom_smooth()` using formula = 'y ~ x'

```

