Chapter 3

Multiple Regression Analysis – Gauss-Markov Theorem

Jim Bang

Omitted Variable Bias

Problem: if we misspecify our model, all bets are off!

1. Estimated model: $y = x_1\beta_1 + e$
2. Correct model: $y = x_1\beta_1 + x_2\beta_2 + u$
3. When we omit *relevant* variables $E(Xu) = 0$ but $E(Xe)$ isn't!

The presence of omitted variable bias requires:

1. Omitted variables, x_2 correlate with y *and*
2. Omitted variables, x_2 correlate with x_1

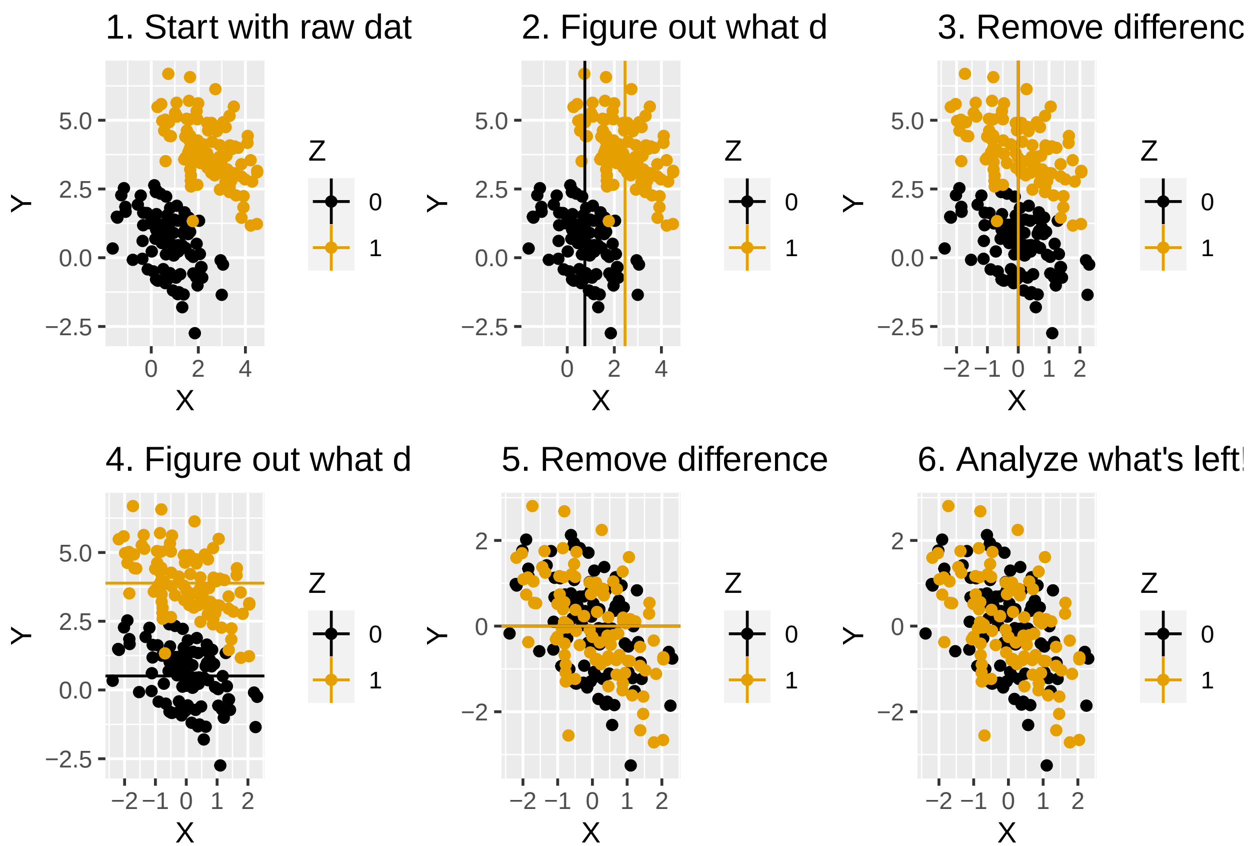
Graphical Demonstration (with Animation!)

The following graph shows how omitting a single *relevant* variable can introduce quite a bit of bias.

```
df <- data.frame(Z = as.integer((1:200>100))) %>%
  mutate(X =  .5   + 2*Z + rnorm(200)) %>%
  mutate(Y = -.5*X + 4*Z + 1 + rnorm(200),time="1") %>%
  group_by(Z) %>%
  mutate(mean_X=mean(X),mean_Y=mean(Y)) %>%
  ungroup()
before_cor <- paste("1. Start with raw data. Correlation between X and Y: ",round(cor(df$X,df$Y),3),sep='')
after_cor <- paste("6. Analyze what's left! Correlation between X and Y controlling for Z: ",round(cor(df$X-df$mean_X,df$Y-df$mean_Y),3),s
dffull <- rbind(
  df %>% mutate(mean_X=NA,mean_Y=NA,time=before_cor),
  df %>% mutate(mean_Y=NA,time='2. Figure out what differences in X are explained by Z'),
  df %>% mutate(X = X - mean_X,mean_X=0,mean_Y=NA,time="3. Remove differences in X explained by Z"),
```

```r
  df %>% mutate(X = X - mean_X,mean_X=NA,time="4. Figure out what differences in Y are explained by Z"),
  df %>% mutate(X = X - mean_X, Y = Y - mean_Y,mean_X=NA,mean_Y=0,
    time="5. Remove differences in Y explained by Z"),
  df %>% mutate(X = X - mean_X,Y = Y - mean_Y,mean_X=NA,mean_Y=NA,time=after_cor))
p1 <- ggplot(subset(dffull, time == names(table(dffull$time))[1]),
  aes(y=Y,x=X,color=as.factor(Z)))+geom_point()+
  geom_vline(aes(xintercept=mean_X,color=as.factor(Z)), na.rm = TRUE)+
  geom_hline(aes(yintercept=mean_Y,color=as.factor(Z)), na.rm = TRUE)+
  guides(color=guide_legend(title="Z"))+
  scale_color_colorblind()+
  labs(title = names(table(dffull$time))[1])
p2 <- ggplot(subset(dffull, time == names(table(dffull$time))[2]),
  aes(y=Y,x=X,color=as.factor(Z)))+geom_point()+
  geom_vline(aes(xintercept=mean_X,color=as.factor(Z)), na.rm = TRUE)+
  geom_hline(aes(yintercept=mean_Y,color=as.factor(Z)), na.rm = TRUE)+
  guides(color=guide_legend(title="Z"))+
  scale_color_colorblind()+
  labs(title = names(table(dffull$time))[2])
p3 <- ggplot(subset(dffull, time == names(table(dffull$time))[3]),
  aes(y=Y,x=X,color=as.factor(Z)))+geom_point()+
  geom_vline(aes(xintercept=mean_X,color=as.factor(Z)), na.rm = TRUE)+
  geom_hline(aes(yintercept=mean_Y,color=as.factor(Z)), na.rm = TRUE)+
  guides(color=guide_legend(title="Z"))+
  scale_color_colorblind()+
  labs(title = names(table(dffull$time))[3])
p4 <- ggplot(subset(dffull, time == names(table(dffull$time))[4]),
  aes(y=Y,x=X,color=as.factor(Z)))+geom_point()+
  geom_vline(aes(xintercept=mean_X,color=as.factor(Z)), na.rm = TRUE)+
  geom_hline(aes(yintercept=mean_Y,color=as.factor(Z)), na.rm = TRUE)+
  guides(color=guide_legend(title="Z"))+
  scale_color_colorblind()+
  labs(title = names(table(dffull$time))[4])
p5 <- ggplot(subset(dffull, time == names(table(dffull$time))[5]),
  aes(y=Y,x=X,color=as.factor(Z)))+geom_point()+
  geom_vline(aes(xintercept=mean_X,color=as.factor(Z)), na.rm = TRUE)+
  geom_hline(aes(yintercept=mean_Y,color=as.factor(Z)), na.rm = TRUE)+
  guides(color=guide_legend(title="Z"))+
  scale_color_colorblind()+
```

```
  labs(title = names(table(dffull$time))[5])
p6 <- ggplot(subset(dffull, time == names(table(dffull$time))[6]),
  aes(y=Y,x=X,color=as.factor(Z)))+geom_point()+
  geom_vline(aes(xintercept=mean_X,color=as.factor(Z)), na.rm = TRUE)+
  geom_hline(aes(yintercept=mean_Y,color=as.factor(Z)), na.rm = TRUE)+
  guides(color=guide_legend(title="Z"))+
  scale_color_colorblind()+
  labs(title = names(table(dffull$time))[6])
ggarrange(p1, p2, p3, p4, p5, p6, nrow = 2, ncol = 3)
```



Mathematical Proof

$$E(\tilde{\beta}_1) = E[(X_1'X_1)^{-1}X_1'(X_1\beta_1 + X_2\beta_2 + u)]$$

$$E(\tilde{\beta}_1) = \beta_1 + \beta_2 E[(X_1'X_1)^{-1}X_1'X_2]$$

We can even rewrite this as:

$$E(\hat{\beta}_1) = \beta_1 + \beta_2\delta$$

where δ is the coefficient from regressing X_2 on X_1, $X_2 = X_1\delta + e$.

Irrelevant Variables

1. Generate a completely random variable, sunspots, which takes values from a Poisson distribution with parameter $\lambda = 10$, as a new column in *wage1*.
2. Regress *wage* on *educ*, *exper*, and *sunspots* (with a constant) and call this regression *wage.lm3*.
3. Compare this to our previous regression, *wage.lm2* (which is identical but excludes *sunspots*) using stargazer(…, type = 'text').

```r
set.seed(8675309)
wage1$sunspots <- rpois(length(wage1$wage), lambda = 10)
wage.lm3 <- lm(wage ~ educ + exper + sunspots, data = wage1)
stargazer(wage.lm2, wage.lm3, type = 'text')
```

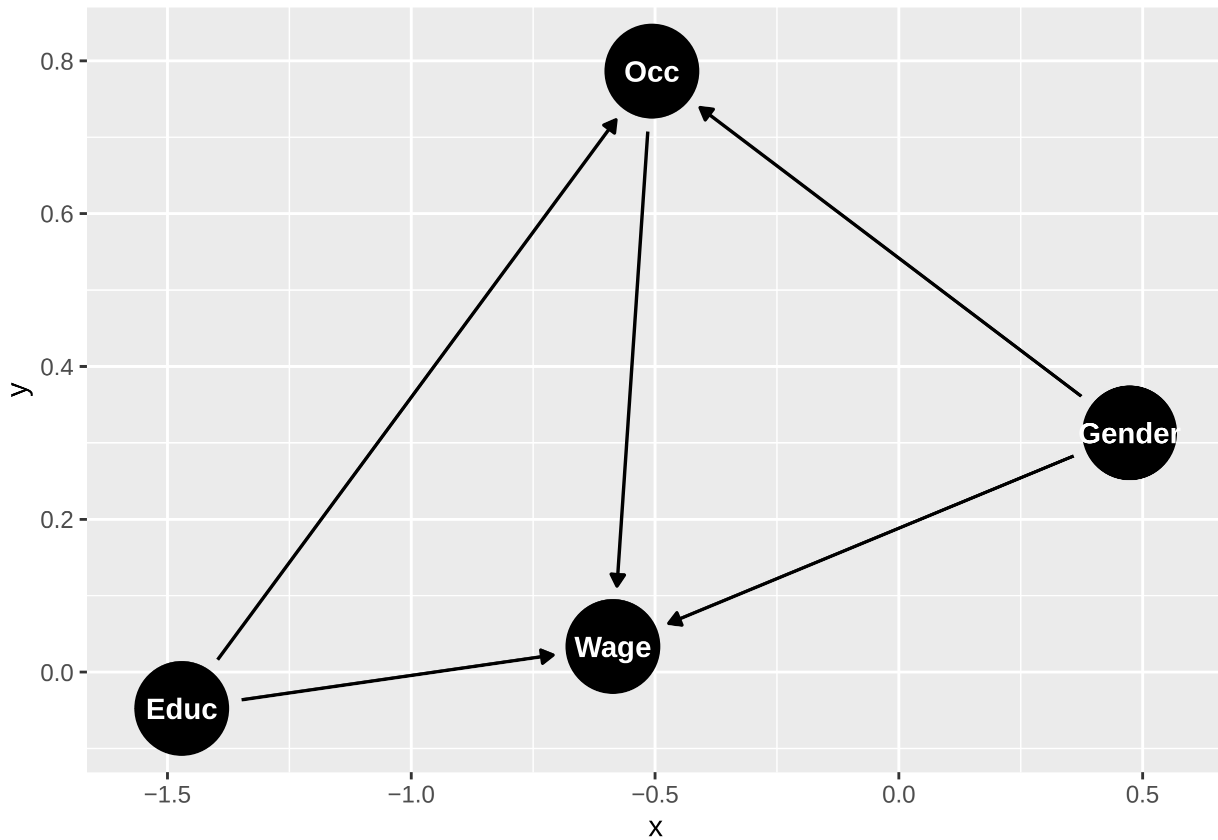
```
##
## ===================================================================
##                                   Dependent variable:
##                       ---------------------------------------------
##                                          wage
##                                (1)                    (2)
## -------------------------------------------------------------------
## educ                        0.644***               0.642***
##                              (0.054)                (0.054)
##
## exper                       0.070***               0.070***
##                              (0.011)                (0.011)
##
## sunspots                                            -0.029
##                                                     (0.044)
##
## Constant                    -3.391***              -3.075***
##                              (0.767)                (0.906)
##
## -------------------------------------------------------------------
## Observations                   526                    526
## R2                            0.225                  0.226
## Adjusted R2                   0.222                  0.221
## Residual Std. Error     3.257 (df = 523)       3.259 (df = 522)
## F Statistic         75.990*** (df = 2; 523) 50.748*** (df = 3; 522)
## ===================================================================
## Note:                                   *p<0.1; **p<0.05; ***p<0.01
```

A Caveat about Omitted Variables

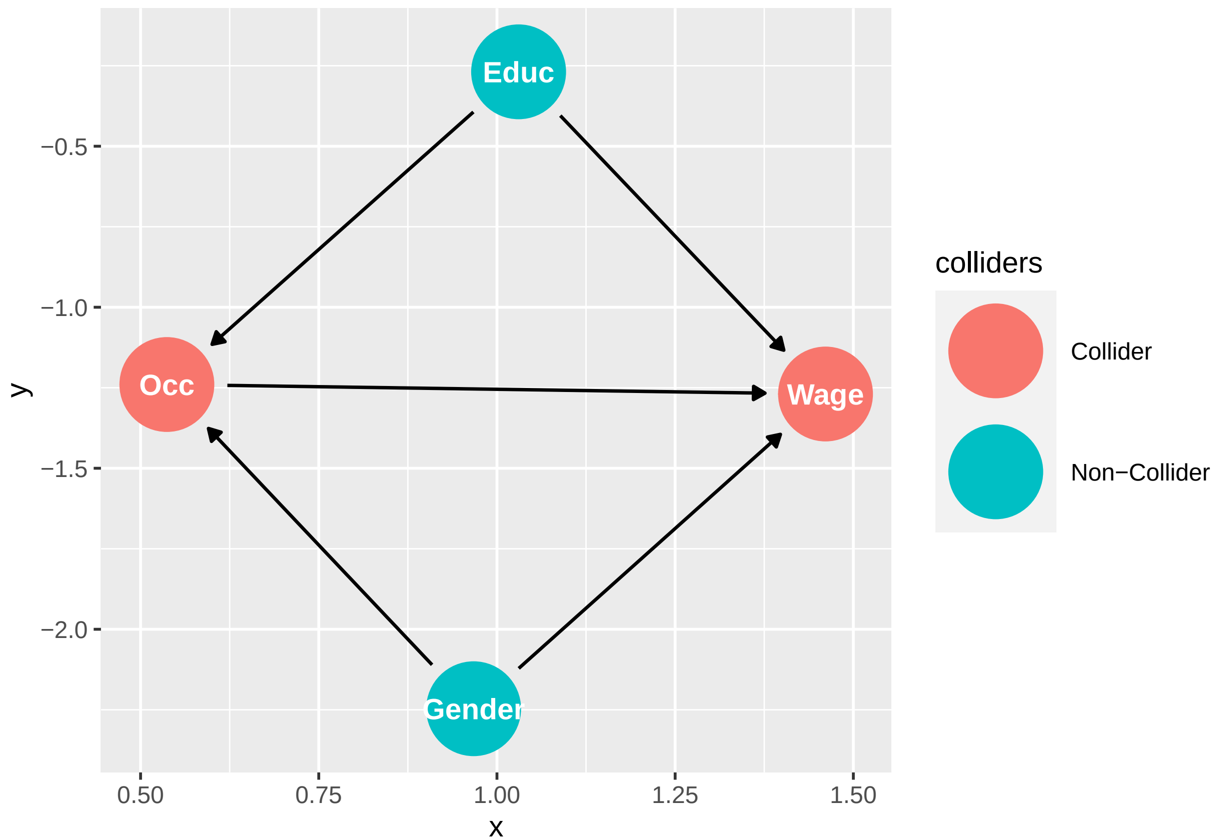
- Should tests for gender wage discrimination control for occupational choice?
    a. Yes
    b. No

Sometimes it's helpful to draw the pattern of causality. There are some neat tools for this in the *ggdag* package.

```
wageGapDag  <- dagify(Wage ~ Educ + Gender + Occ,
  Occ ~ Educ + Gender)
ggdag(wageGapDag)
```



```
ggdag_collider(wageGapDag)
```



Multicollinearity

Perfect Multicollinearity

When a group of variables are perfectly collinear (correlated), we cannot invert $X'X$. It becomes akin to dividing by zero. Example: White/Nonwhite.

1. Using the wage1 data create the variable *white* equal to one minus *nonwithe*.
2. Regress *wage* on *educ*, *exper*, *white* and *nonwhite*. Call this regression *wage.lm4* and summarize the results.

```
wage1$white <- 1 - wage1$nonwhite
wage.lm4 <- lm(wage ~ educ + exper + white + nonwhite, data = wage1)
summary(wage.lm4)
```

```
##
## Call:
## lm(formula = wage ~ educ + exper + white + nonwhite, data = wage1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.538 -1.982 -0.709  1.205 15.835
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.40304    0.84860  -4.010 6.95e-05 ***
## educ         0.64412    0.05405  11.917  < 2e-16 ***
## exper        0.07009    0.01099   6.378 3.95e-10 ***
## white        0.01621    0.47006   0.034    0.972
## nonwhite          NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.26 on 522 degrees of freedom
## Multiple R-squared:  0.2252, Adjusted R-squared:  0.2207
## F-statistic: 50.56 on 3 and 522 DF,  p-value: < 2.2e-16
```

Notice how the lm() function drops *nonwhite* since it is a perfect linear function of *white* and the constant (reverse the order they appear and it will drop whichever is last).

Perfect Multicollinearity without a Constant

Suppose I want to know the absolute magnitude of the intercept for each group (white/nonwhite).

1. One way I could do this is by adding. The baseline for the omitted group (here, *nonwhite*) is the regular intercept.
2. Another way I could do this is to regress the model with both variables, excluding the intercept.

Do #2, name it *wage.lm5*, and summarize the results.

```
wage.lm5 <- lm(wage ~ educ + exper + white + nonwhite - 1, data = wage1)
summary(wage.lm5)
```

```
##
## Call:
## lm(formula = wage ~ educ + exper + white + nonwhite - 1, data = wage1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.538 -1.982 -0.709  1.205 15.835
##
## Coefficients:
##          Estimate Std. Error t value Pr(>|t|)
## educ      0.64412    0.05405  11.917  < 2e-16 ***
## exper     0.07009    0.01099   6.378 3.95e-10 ***
## white    -3.38683    0.77481  -4.371 1.49e-05 ***
## nonwhite -3.40304    0.84860  -4.010 6.95e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.26 on 522 degrees of freedom
## Multiple R-squared:  0.782,  Adjusted R-squared:  0.7803
## F-statistic:   468 on 4 and 522 DF,  p-value: < 2.2e-16
```

Notice that the coefficient for *nonwhite* is the same as the previous example's intercept; the value for *white* is the intercept of the previous regression plus its marginal effect for the *white* group. ### Imperfect Multicollinearity

Variance Formula - Scalar Form

Simple Regression Model

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_x}$$

Multiple Regression Model

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)} = \frac{\sigma^2}{SST_j} \cdot V.I.F$$

Effect of Multicollinearity 1. $SST_j = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$ 2. R_j^2 is the R^2 obtained from regressing x_j on all other x's and a constant. 3. The more correlated x_j is with the other x's, the more inflated the variance becomes.

Bias-variance trade-off.