Section 5.2 Observational Studies with Multiple Quantitative Variables

Kevin Cummiskey

November 4, 2019

Review

Let's say you are interested in the effects of sleep and coffee consumption on academic performance. You randomly assign cadets in a class to one of three sleep groups (3,5,7 hours) and coffee groups (0, 200, 400 mg) in a 3 x 3 factorial design and record their performance on a WPR. Let y_i be the WPR grade, $x_{1,i}$ be sleep in hours (standardized), and $x_{2,i}$ be the coffee consumption (standardized) of cadet i . Consider the following model:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \epsilon_i \quad \epsilon_i \sim N(0, \sigma^2)$$

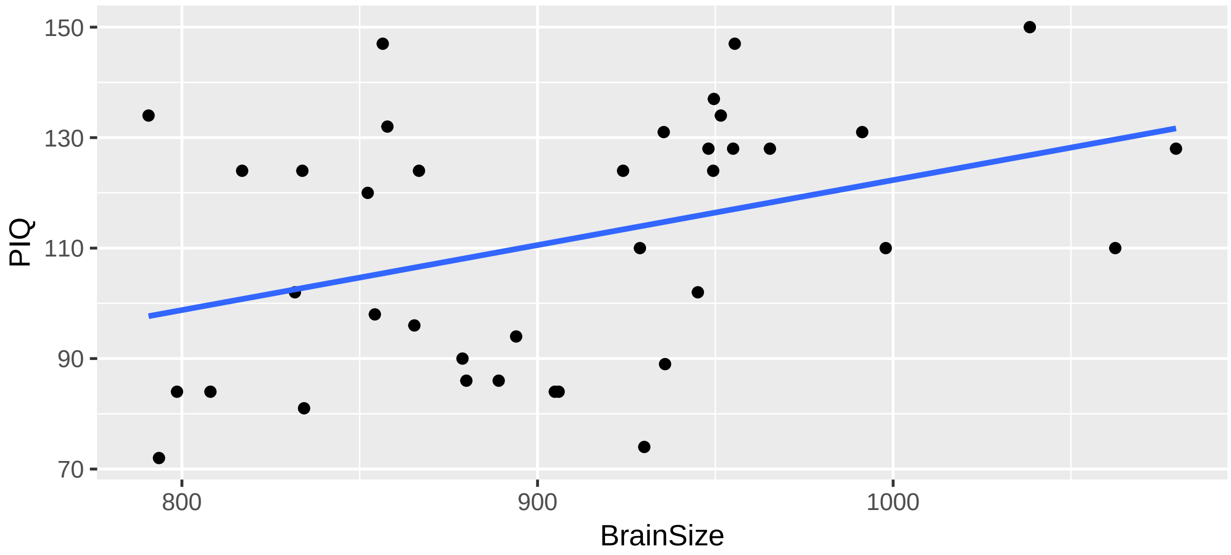
1. Why do we standardize explanatory variables?
2. Interpret β_0.
3. Will the adjusted and unadjusted effect of sleep on WPR performance be the same? Explain.
4. Can we use this model to test whether the effect of sleep on WPR performance depends upon coffee consumption? Explain.

Example 5.2

```
library(tidyverse)
brains = read.table(file = "http://www.isi-stats.com/isi2/data/BrainSize.txt",
                    header = T)
```

PIQ and Brain Size

```
brains %>% ggplot(aes(x = BrainSize, y = PIQ)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```



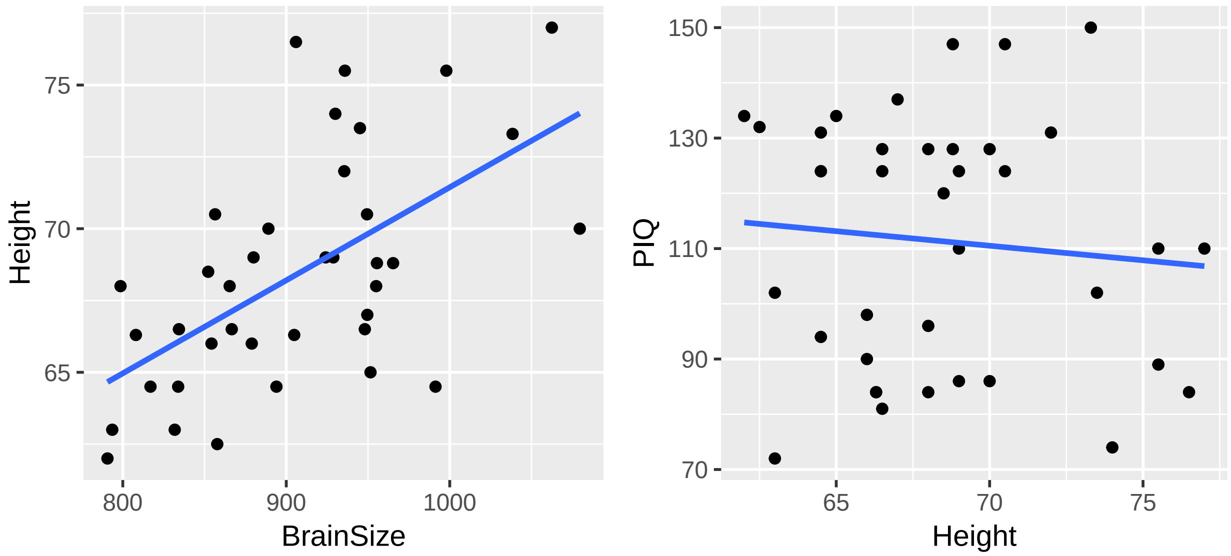
```
model_unadjusted = lm(PIQ ~ BrainSize, data = brains)
summary(model_unadjusted)
```

```
##
## Call:
## lm(formula = PIQ ~ BrainSize, data = brains)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.079 -17.508  -2.096  17.100  41.574
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.66036   43.71288   0.107   0.9157
## BrainSize    0.11765    0.04806   2.448   0.0194 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.21 on 36 degrees of freedom
## Multiple R-squared:  0.1427, Adjusted R-squared:  0.1189
## F-statistic: 5.993 on 1 and 36 DF,  p-value: 0.01937
```

What conclusions should we draw from the unadjusted model?

Confounding by Height?

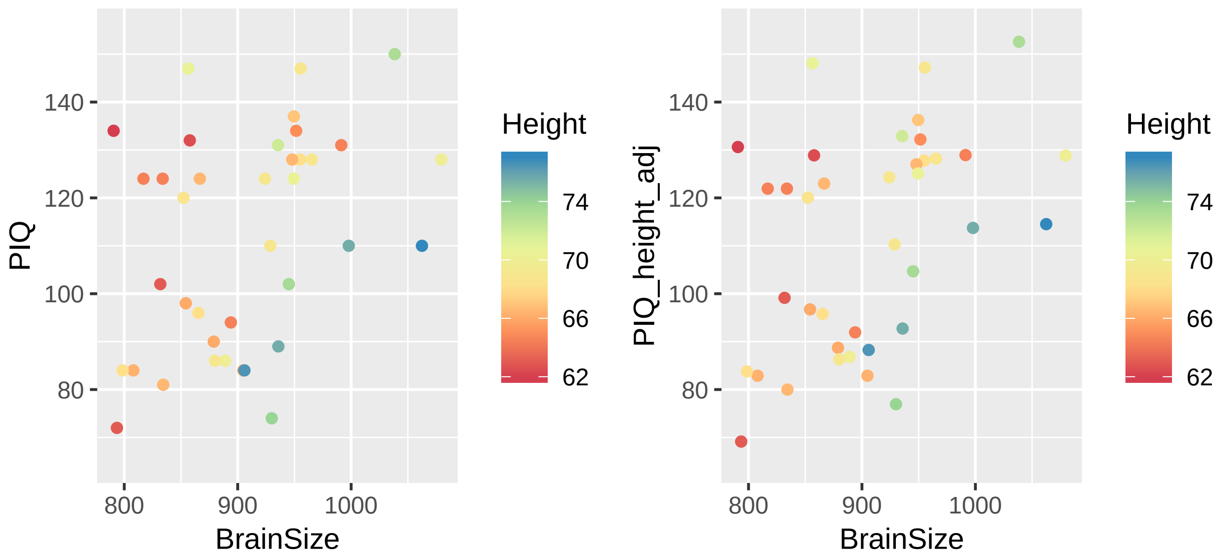
```
library(gridExtra)
p1 = brains %>% ggplot(aes(x = BrainSize, y = Height)) +
  geom_point() + geom_smooth(method = "lm", se = F)
p2 = brains %>% ggplot(aes(x = Height, y = PIQ)) +
  geom_point() + geom_smooth(method = "lm", se = F)
grid.arrange(p1,p2, ncol = 2)
```



Is there evidence of confounding? How will adjusting for height change the effect of BrainSize on PIQ?

Let's look at the height-adjusted PIQ.

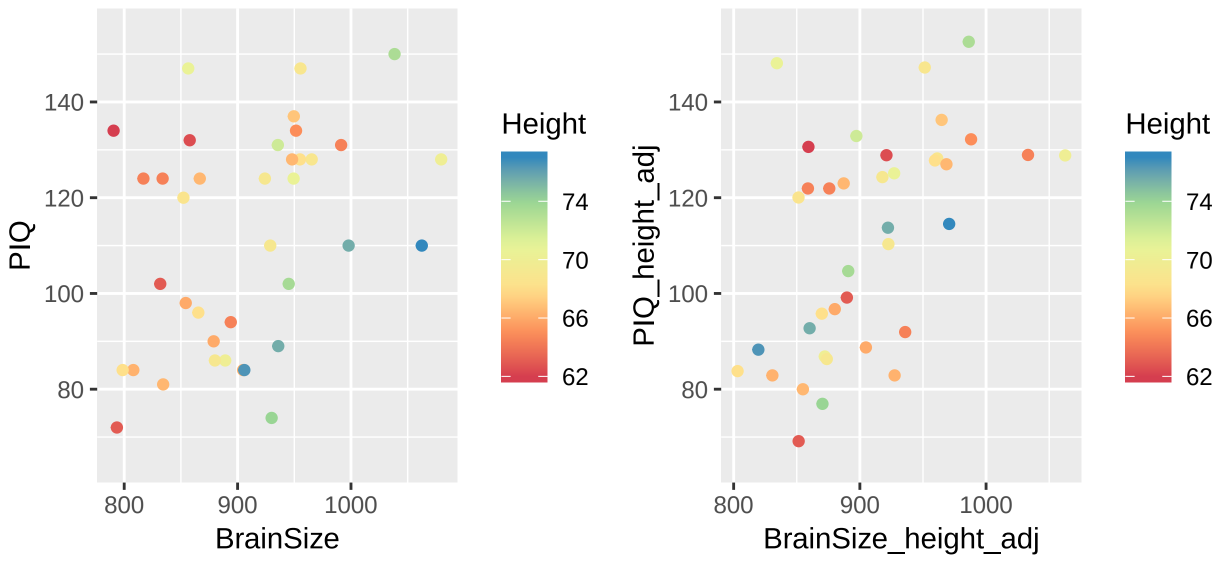
```
model_PIQheight = lm(PIQ ~ Height, data = brains)
summary_PIQheight = model_PIQheight %>% fortify()

#height-adjusted PIQ
brains = brains %>%
  mutate(PIQ_height_adj = mean(PIQ) + summary_PIQheight$.resid)

#plots on page 382
# color scale
sc = scale_color_distiller(palette = "Spectral", direction = 1)

p3 = brains %>% ggplot(aes(x = BrainSize, y = PIQ, color = Height)) +
  geom_point() + ylim(65,155) + sc
p4 = brains %>% ggplot(aes(x = BrainSize, y = PIQ_height_adj, color = Height)) +
  geom_point() + ylim(65,155) + sc
grid.arrange(p3,p4,ncol = 2)
```



How did we obtain height-adjusted PIQ? What happened to PIQ after adjusting for height?

Let's look at height-adjusted brainsize.

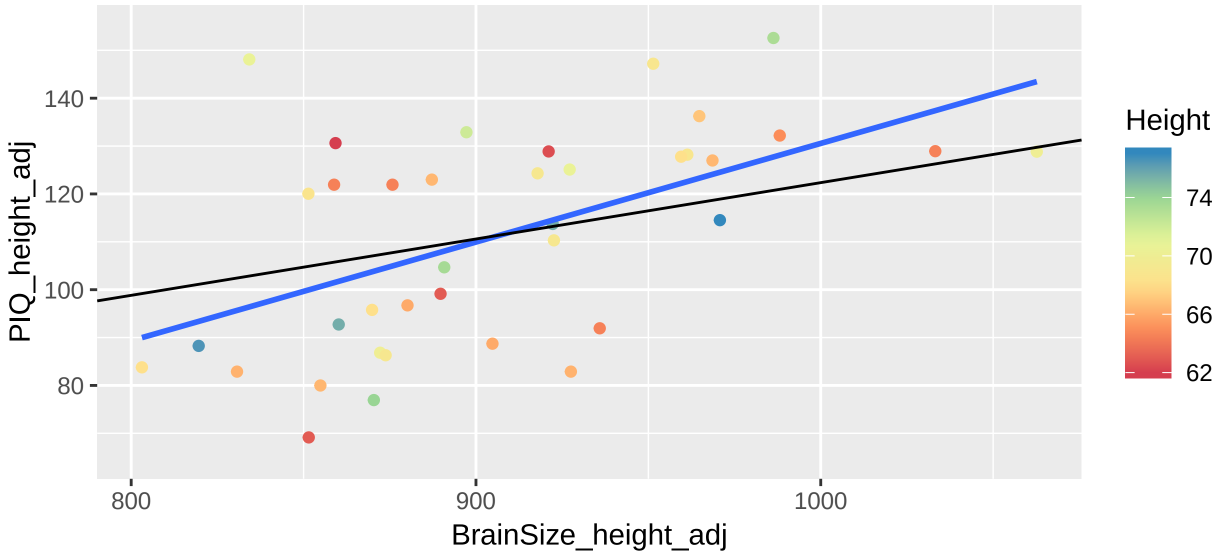
```
model_Brainheight = lm(BrainSize ~ Height, data = brains)
summary_Brainheight = model_Brainheight %>% fortify()

#height-adjusted PIQ
brains = brains %>%
  mutate(BrainSize_height_adj = mean(BrainSize) + summary_Brainheight$.resid)

#plots on page 382
p5 = brains %>% ggplot(aes(x = BrainSize, y = PIQ, color = Height)) +
  geom_point() + ylim(65,155) + sc
p6 = brains %>%
  ggplot(aes(x = BrainSize_height_adj, y = PIQ_height_adj, color = Height)) +
  geom_point() +
  ylim(65,155) +
  sc
grid.arrange(p5,p6,ncol = 2)
```



How do we obtain height-adjusted BrainSize?

Let's look at an added variable plot predicting PIQ from BrainSize before (black line) and after (blue line) adjusting for height.

```
p6   + geom_smooth(method = "lm", se = FALSE) +
  geom_abline(aes(slope = 0.1177, intercept = 4.66))
```



Let's fit the main effects model.

```
model_height_brainsize = lm(PIQ ~ BrainSize + Height, data = brains)
summary(model_height_brainsize)
```

```
## 
## Call:
## lm(formula = PIQ ~ BrainSize + Height, data = brains)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -32.74 -12.09  -3.84  14.18  51.69 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 111.27847   55.86881   1.992 0.054243 .  
## BrainSize     0.20606    0.05467   3.769 0.000605 ***
## Height       -2.72984    0.99322  -2.748 0.009403 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 19.51 on 35 degrees of freedom
## Multiple R-squared:  0.2949, Adjusted R-squared:  0.2546 
## F-statistic: 7.319 on 2 and 35 DF,  p-value: 0.00221
```

What would we conclude from these results?

Interaction

If we test for an interaction, what research question are we answering?

Why should we standardize BrainSize and Height?

```r
brains = brains %>% mutate(std.BrainSize = scale(BrainSize),
                           std.Height = scale(Height))
model_interaction = lm(PIQ ~ std.BrainSize * std.Height,
                       data = brains)
summary(model_interaction)
```

```
## 
## Call:
## lm(formula = PIQ ~ std.BrainSize * std.Height, data = brains)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -32.779 -12.001  -3.871  14.209  51.604 
## 
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              111.39048    3.69138  30.176  < 2e-16 ***
## std.BrainSize             14.95030    4.02498   3.714 0.000728 ***
## std.Height               -10.88380    4.08511  -2.664 0.011712 *
## std.BrainSize:std.Height  -0.08444    3.17809  -0.027 0.978958
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.79 on 34 degrees of freedom
## Multiple R-squared:  0.2949, Adjusted R-squared:  0.2327
## F-statistic:  4.74 on 3 and 34 DF,  p-value: 0.007219
```

What do we conclude from these results?