HW4

Matt Kaye

10/3/2018

1

weight	expected.height	5%	94%
46.95	156.338	148.226	164.577
43.72	153.415	145.432	161.558
64.78	172.475	164.157	180.668
32.59	143.386	135.284	151.504
54.63	163.389	155.214	171.547

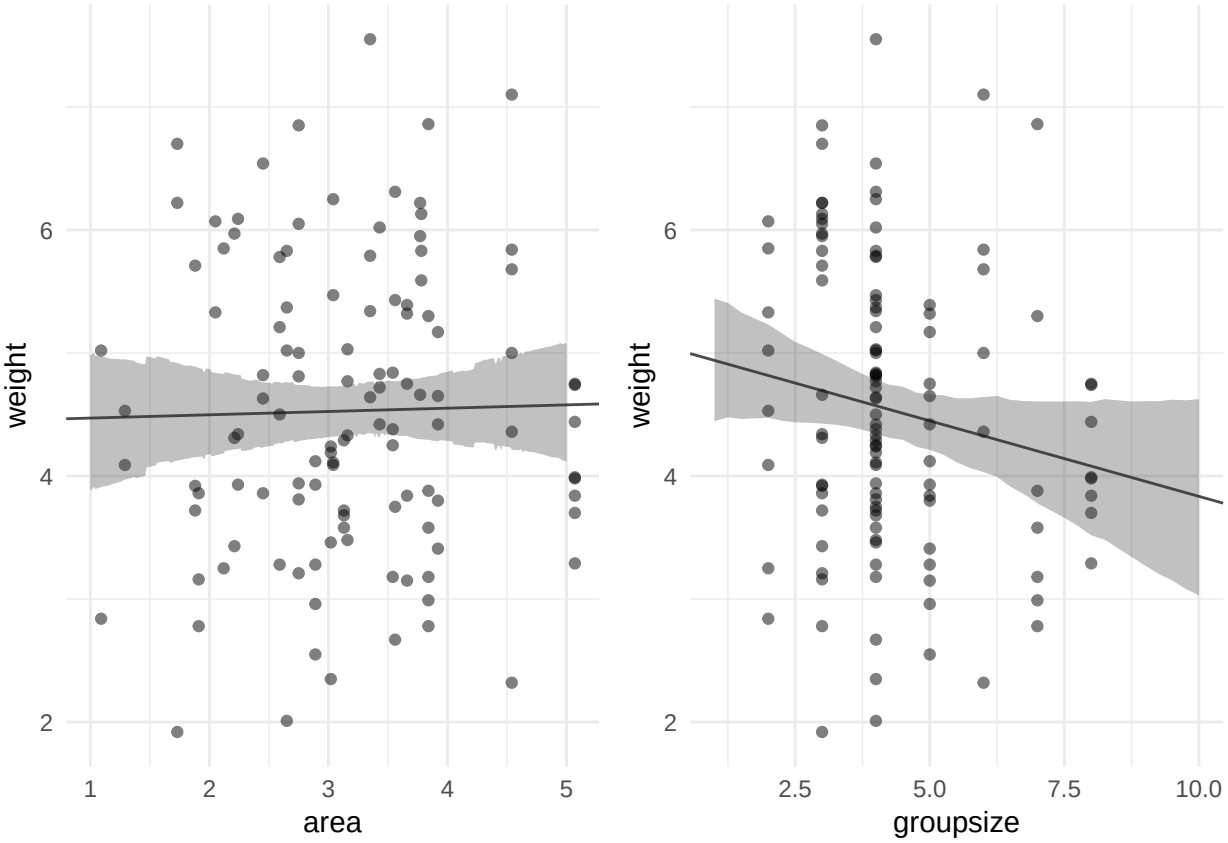
2a

```
##       Mean StdDev  5.5% 94.5%
## a     4.44   0.39  3.82  5.07
## b     0.03   0.12 -0.16  0.22
## sigma 1.18   0.08  1.06  1.30
```

b

```
##        Mean StdDev  5.5% 94.5%
## a      5.06   0.32  4.54  5.58
## b     -0.12   0.07 -0.24 -0.01
## sigma  1.16   0.08  1.04  1.29
```

c



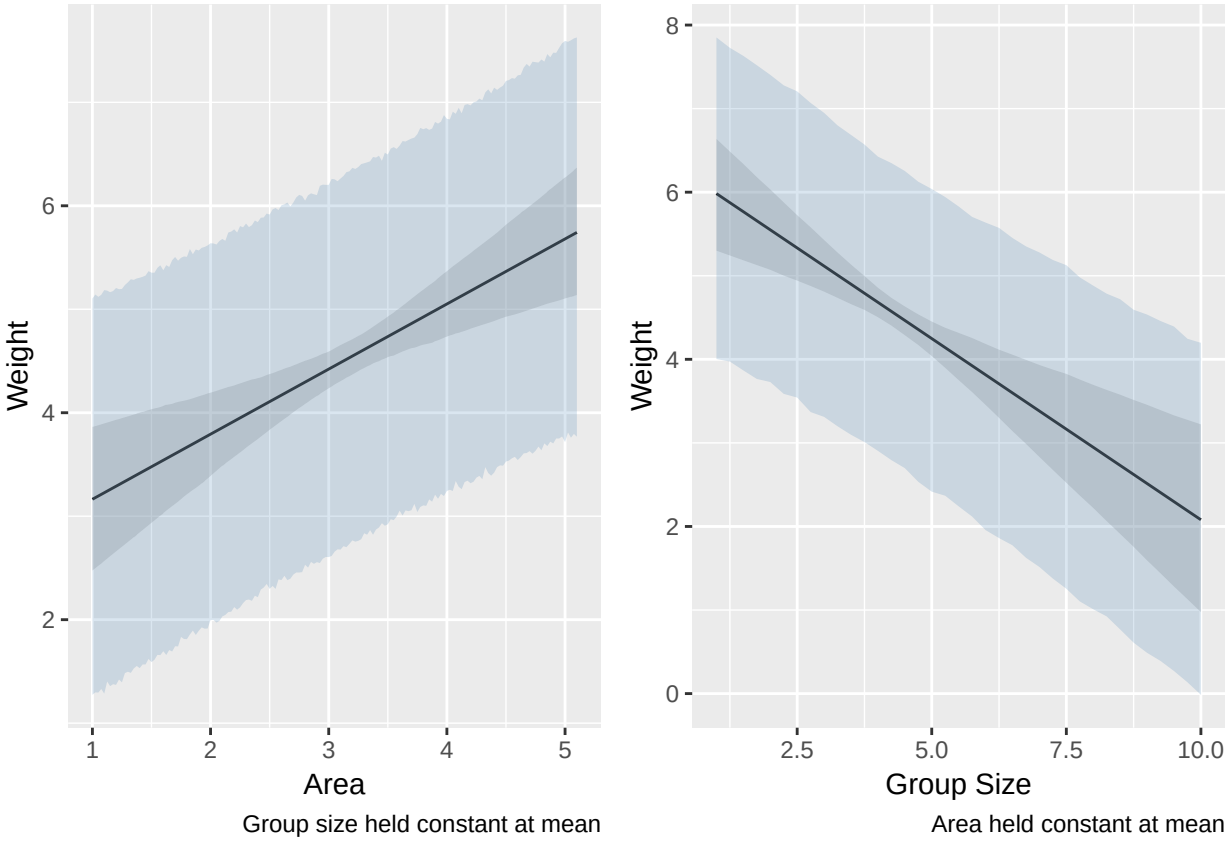
d

No. It does not seem like either variable is statistically significant in predicting weight. This is especially true of area. Zero is virtually in the center of the 89% credible interval for the slope coefficient in the regression of weight on area, and the line plotted above looks about flat. In the case of groupsize, the predictor actually is significant at the 95% level, but, again, it is possible as well that the line is flat. Because this is a fringe case, I would err on the side of caution and say no.

3a

```
##         Mean StdDev  5.5% 94.5%
## a       4.44   0.37  3.85  5.04
## b       0.62   0.20  0.30  0.94
## c      -0.43   0.12 -0.63 -0.24
## sigma   1.12   0.07  1.00  1.24
```

b



c

Both the credible intervals and the counterfactual plots suggest that group size and area are both statistically significant in predicting weight.

d

We get different results in this model than we do in the one from problem two because controlling for other factors through multiple regression allows us to unbias our regression coefficients from the previous model. In other words, these results being different from those of question 2 suggest that there was significant omitted variables bias before because we were using univariate regression, which was throwing off our estimates of the true relationships between weight and age or group size.

4a

```
##
##  Welch Two Sample t-test
##
## data:  calcium$flow by calcium$treatment
## t = 1.8412, df = 47.861, p-value = 0.0718
## alternative hypothesis: true difference in means is not equal to 0
```

```
## 97 percent confidence interval:
##  -0.02613061  0.26939761
## sample estimates:
## mean in group A mean in group B
##        1.144441        1.022807
```

b

Likelihood:

$$flow \mid treatment \sim N(\mu, \sigma)$$

Link (B is a dummy variable for treatment group B):

$$\hat{\mu} = \hat{\beta_0} + \hat{\beta_1} * B$$

c

```
##       Mean StdDev  1.5% 98.5%
## a     1.02   0.05  0.92  1.12
## b     0.12   0.06 -0.02  0.26
## sigma 0.26   0.02  0.21  0.31
```

d

There is a 97% chance that the true difference in the mean of flow between the two groups is between -.02 and .26. This is almost exactly the same as the frequentist confidence interval that we calculated in part a, which is reassuring.

5a

$$\hat{Energy} = \beta_0 + \beta_1 * RedOak + \beta_2 * (.1)$$

To predict the difference in energy between white pine and red oak, we use a dummy variable for red oak (such that 0 is white pine and 1 is red oak) so that we can determine the difference in the intercept coefficients of the two types of wood. We then plug in .1 for the moisture content in our model to determine the difference in energy content of the two types of wood at 10% moisture content. Something to note is that the inclusion of the slope term does not actually have an effect on our estimate of the difference in energy between white pine and red oak, because we assume parallel lines which means that the difference between the two groups is the same at any level of moisture.

b

$$\hat{Energy} = \beta_0 + \beta_1 * RedOak + \beta_2 * (.1) + \beta_3 * RedOak * (.1)$$

Everything in this model is the same as the one above except for the introduction of an interaction term between red oak and moisture content. This allows us to test for a difference in the slope depending on the type of wood.

c

$$Y = \beta_0 + \beta 1 * X * D_{\leq 20} + \beta 1 * (20) * D_{>20}$$

For this model, we want an intercept term and slope term for the regression when X is between 0 and 20. To solve for the slope, we simply run the regression with the dummy for X being greater than 20 as zero. This gives us normal regression estimates for the slope and intercept. Then, to compare the difference in exam scores for people who watch 25 hours of TV and 3 hours of TV, we plug 3 in for X to get an estimate for exam score at 3 hours of TV watching, and we add the first and third terms in the model above to get an estimate of the exam score at 25 hours of TV (because the middle term is 0 because the dummy is 0).