Statistical Rethinking: Chapter 5 - Multivariate Linear Models

Chris Hayduk

January 13, 2019

1 Easy

Problem 5E1.
Which of the linear models below are multiple linear regressions?

1. $\mu_i = \alpha + \beta x_i$
2. $\mu_i = \beta_x x_i + \beta_z z_i$
3. $\mu_i = \alpha + \beta(x_i - z_i)$
4. $\mu_i = \alpha + \beta_x x_i + \beta_z z_i$

Linear models 2 and 4 are multiple linear regressions.

Problem 5E2.
Write down a multiple regression to evaluate the claim: *Animal diversity is linearly related to latitude, but only after controlling for plant diversity.* You just need to write down the model definition.

$$\begin{gathered}
\text{animal diversity}_i \sim \text{Normal}(\mu_i, \sigma) \\
\mu_i = \beta_{\text{latitude}}\text{latitude}_i + \beta_{\text{diversity}}\text{diversity}_i \\
\beta_{\text{latitude}} \sim \text{Normal}(0, 10) \\
\beta_{\text{diversity}} \sim \text{Normal}(0, 10) \\
\sigma \sim \text{Uniform}(0, 10)
\end{gathered}$$

Problem 5E3.
Write down a multiple regression to evaluate the claim: *Neither amount of funding nor size of laboratory is by itself a good predictor of time to PhD degree; but together these variables are both positively associated with time to degree.* Write down the model definition and indicate which side of zero each slope parameter should be on.

$$
\begin{aligned}
\text{time}_i &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_i &= \beta_{\text{lab size}}\text{lab size}_i + \beta_{\text{funding}}\text{funding}_i \\
\beta_{\text{lab size}} &\sim \text{Normal}(0, 10) \\
\beta_{\text{funding}} &\sim \text{Normal}(0, 10) \\
\sigma &\sim \text{Uniform}(0, 10)
\end{aligned}
$$

Both parameters should have slopes greater than zero since the problem specifies that "together the variables are both positively associated with time to degree".

Problem 5E4.
Suppose you have a single categorical predictor with 4 levels (unique values), labeled A, B, C, and D. Let A_i be an indicator variable that is 1 where case i is in category A. Also suppose B_i, C_i, and D_i for the other categories. Now which of the following linear models are inferentially equivalent ways to include the categorical variable in a regression? Models are inferentially equivalent when it's possible to compute one posterior distribution from the posterior distribution of another model.

1. $\mu_i = \alpha + \beta_A A_i + \beta_B B_i + \beta_D D_i$
2. $\mu_i = \alpha + \beta_A A_i + \beta_B B_i + \beta_C C_i + \beta_D D_i$
3. $\mu_i = \alpha + \beta_B B_i + \beta_C C_i + \beta_D D_i$
4. $\mu_i = \alpha_A A_i + \alpha_B B_i + \alpha_C C_i + \alpha_D D_i$
5. $\mu_i = \alpha_i(1 - B_i - C_i - D_i) + \alpha_B B_i + \alpha_C C_i + \alpha_D D_i$

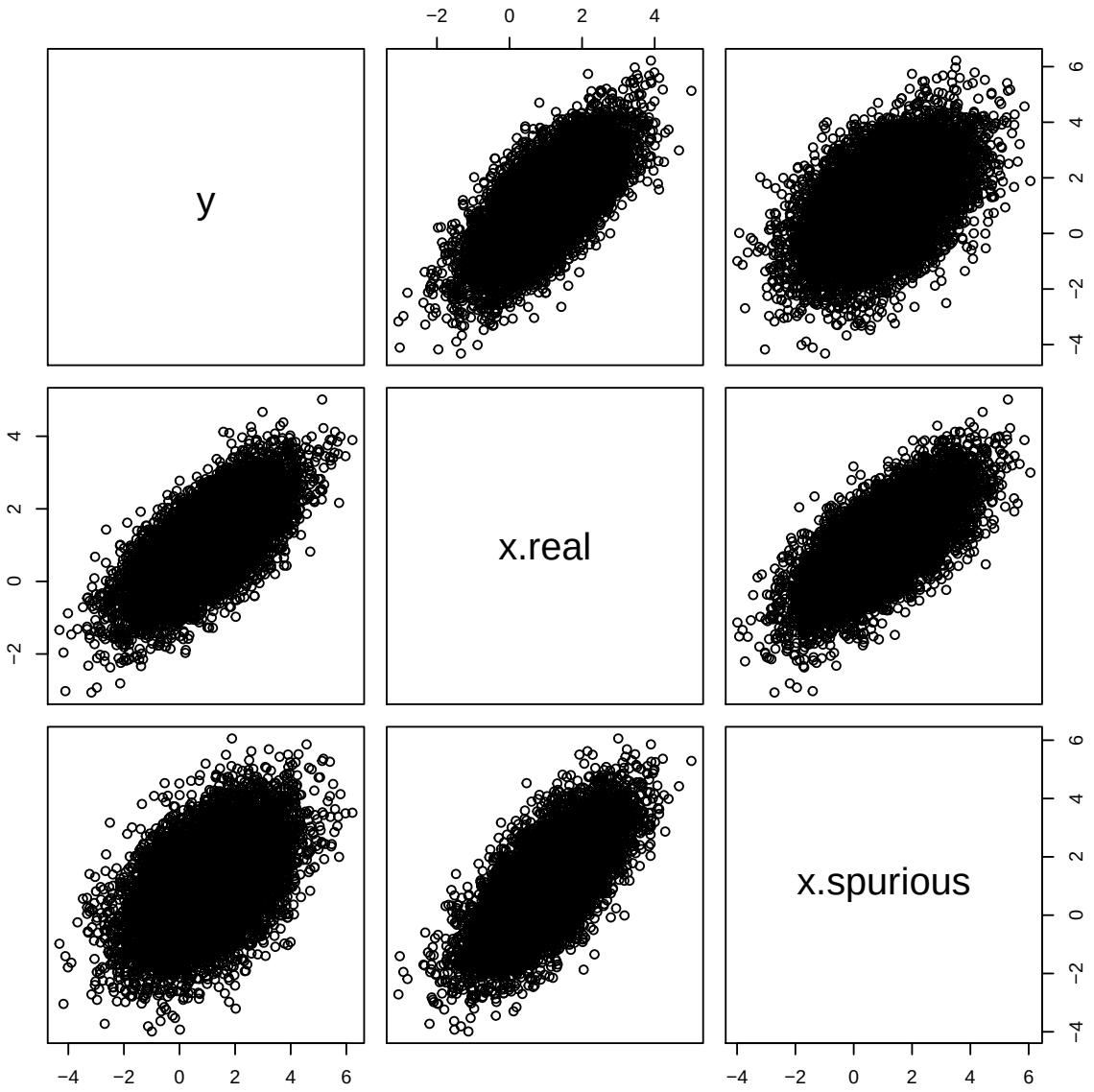
Models 1, 3, 4, and 5 are all inferentially equivalent.

2 Medium

Problem 5M1.
Invent your own example of a spurious correlation. An outcome variable should be correlated with both predictor variables. But when both predictors are entered in the same model, the correlation between the outcome and one of the predictors should mostly vanish (or at least be greatly reduced).

```
N <- 1e4
x.real <- rnorm(n = N, mean = 1, sd = 1)
x.spurious <- rnorm(n = N, mean = x.real)
y <- rnorm(n = N, mean = x.real)

df <- data.frame(y = y, x.real = x.real, x.spurious = x.spurious)
pairs(df)
```




```
model <- lm(y ~ x.real + x.spurious)
precis(model)

##               Mean StdDev  5.5% 94.5%
## (Intercept)  0.01   0.01 -0.01  0.04
## x.real       0.99   0.01  0.97  1.01
## x.spurious  -0.01   0.01 -0.02  0.01
```

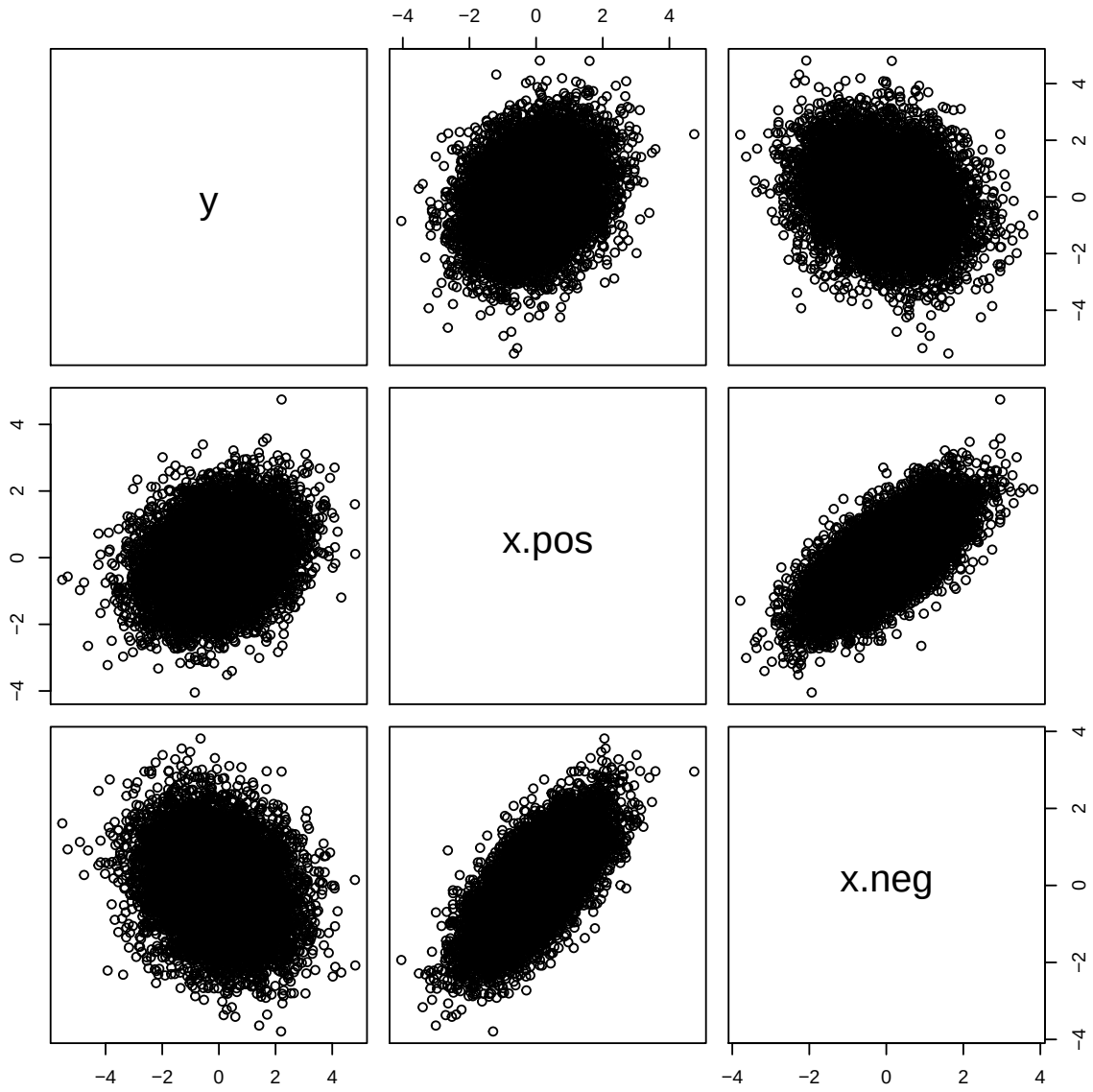
Problem 5M2.
Invent your own example of a masked relationship. An outcome variable should be correlated

with both predictor variables, but in opposite directions. And the two predictor variables should be correlated with one another.

```
N <- 1e4
rho <- 0.7

x.pos <- rnorm(N)
x.neg <- rnorm(N, rho*x.pos, sqrt(1-rho^2))
y <- rnorm(N, x.pos - x.neg)

df <- data.frame(y = y, x.pos = x.pos, x.neg = x.neg)
pairs(df)
```




```
model <- lm(y ~ x.pos + x.neg)
precis(model)

##              Mean StdDev  5.5% 94.5%
## (Intercept)  0.00   0.01 -0.02  0.02
## x.pos        1.04   0.01  1.02  1.06
## x.neg       -1.01   0.01 -1.03 -0.99
```

Problem 5M3.
It is sometimes observed that the best predictor of fire risk is the presence of firefighters - States and localities with many firefighters also have more fires. Presumably firefightes do not

cause fires. Nevertheless, this is not a spurious correlation. Instead fires cause firefighters. Consider the same reversal of causal inference in the context of the divorice and marriage data. How might a high divorce rate cause a higher marriage rate? Can you think of a way to evaluate this relationship, using multiple regression?

A high divorce rate may cause a higher marriage rate because people who are divorced may get married multiple times throughout their lives.

Problem 5M4.
In the divorce data, States with high numbers of Mormons have much lower divorce rates than the regression models expected. Find a list of LDS population by State and use those numbers as a predictor variable, predicting divorce rate using marriage rate, median age at marriage, and percent LDS population (possibly standardized). You may want to consider transformations of the raw percent LDS variable.