5. Chapter 5 Solutions

5E1. Only (2) and (4) are multiple linear regressions. Both have more than one predictor variable and corresponding coefficients in the linear model. The model (1) has only a single predictor variable, x. The model (3) has two predictor variables, but only their difference for each case enters the model, so effectively this is a uni-variate regression, with a single slope parameter.

5E2. A verbal model statement like this will always be somewhat ambiguous. That is why mathematical notation is needed in scientific communication. However, the conventional interpretation of the statement would be:

$$A_i \sim \text{Normal}(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta_L L_i + \beta_P P_i$$

where A is animal diversity, L is latitude, and P is plant diversity. This linear model "controls" for plant diversity, while estimating a linear relationship between latitude and animal diversity.

5E3. Define T as time to PhD degree, the outcome variable implied by the problem. Define F as amount of funding and S as size of laboratory, the implied predictor variables. Then the model (ignoring priors) might be:

$$T_i \sim \text{Normal}(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta_F F_i + \beta_S S_i$$

The slopes β_F and β_S should both be positive.

How can both be positively associated with the outcome in a multiple regression, but neither by itself? If they are negatively correlated with one another, then considering each alone may miss the positive relationships with the outcome. For example, large labs have less funding per student. Small labs have more funding per student, but poorer intellectual environments. So both could be positive influences on time to degree, but be negatively associated in nature.

5E4. This question is tricky. First, the answer will actually depend upon the priors, which aren't mentioned in the problem. But assuming weakly informative or flat priors, the answer is that (1), (3), (4), and (5) are inferentially equivalent. They'll make the same predictions, and you can convert among them after model fitting. (2) stands out because it has a redundant parameter, the intercept α.

5M1. The are many good answers to this question. The easiest approach is to think of some context that follows the divorce rate pattern in the chapter: one predictor influences both the outcome and the other predictor. For example, we might consider predicting whether or not a scientific study replicates. Two predictor variables are available: (1) sample size and (2) statistical significance. Sample size influences both statistical significance and reliability of a finding. This induces a correlation between significance and successful replication, even though significance is not associated with replication, once sample size is taken into account.

5M2. Again, many good answers are possible. The pattern from the milk energy example in the chapter is the simplest. Consider for example the influences of income and drug use on health. Income is positively associated, in reality, with health. Drug use is, for the sake of the example, negatively associated with health. But wealthy people consume more drugs than the poor, simply because the wealthy can afford them. So income and drug use are positively associated in the population. If this positive association is strong enough, examining either income or drug use alone will show only a weak relationship with health, because each works on health in opposite directions.

5M3. Divorce might lead to, or be in expectation of, remarriage. Thus divorce could cause marriage rate to rise. In order to examine this idea, or another like it, the data would need to be structured into more categories, such as remarriage rate versus first marriage rate. Better yet would be longitudinal data. In many real empirical contexts, causation involves feedback loops that can render regression fairly useless, unless some kind of time series framework is used.

5M4. It is worth finding and entering the values yourself, for the practice at data management. But here are the values I found, scraped from Wikipedia and merged into the original data:

R code 5.1

```
library(rethinking)
data(WaffleDivorce)
d <- WaffleDivorce
d$pct_LDS <- c(0.75, 4.53, 6.18, 1, 2.01, 2.82, 0.43, 0.55, 0.38,
  0.75, 0.82, 5.18, 26.35, 0.44, 0.66, 0.87, 1.25, 0.77, 0.64, 0.81,
  0.72, 0.39, 0.44, 0.58, 0.72, 1.14, 4.78, 1.29, 0.61, 0.37, 3.34,
  0.41, 0.82, 1.48, 0.52, 1.2, 3.85, 0.4, 0.37, 0.83, 1.27, 0.75,
  1.21, 67.97, 0.74, 1.13, 3.99, 0.92, 0.44, 11.5 )
d$L <- standardize( d$pct_LDS )
d$A <- standardize( d$MedianAgeMarriage )
d$M <- standardize( d$Marriage )
d$D <- standardize( d$Divorce )
```

A first regression model including this variable might be:

R code 5.2

```
m_5M4 <- quap(
    alist(
        D ~ dnorm(mu,sigma),
        mu <- a + bM*M + bA*A + bL*L,
        a ~ dnorm(0,0.2),
        c(bA,bM,bL) ~ dnorm(0,0.5),
        sigma ~ dexp(1)
    ), data=d )
precis( m_5M4 )
```

```
       mean   sd  5.5% 94.5%
a      0.00 0.09 -0.15  0.15
bA    -0.69 0.14 -0.92 -0.46
bM     0.04 0.15 -0.20  0.27
bL    -0.31 0.12 -0.50 -0.12
sigma  0.73 0.07  0.62  0.85
```

As expected, there is a negative association between percent LDS and divorce rate. This model assumes the relationship between divorce rate and percent LDS is linear. This makes sense if the LDS

community has a lower divorce rate within itself only, and so as it makes up more of a State's population, that State's divorce rate declines. This is to say that the expected divorce rate of State i is a "convex" mix of two average divorce rates:

$$D_i = (1 - P)D_{\text{G}} + PD_{\text{LDS}}$$

where D_i is the divorce rate for State i, P is the proportion of the State's population that is LDS, and the two divorce rates D_{G} and D_{LDS} are the divorce rates for gentiles (non-LDS) and LDS, respectively. If $D_{\text{G}} > D_{\text{LDS}}$, then as P increases, the value of D_i increases linearly as well.

But maybe the percent LDS in the population has a secondary impact as a marker of a State-level cultural environment that has lower divorce in more demographic groups than just LDS. In that case, this model will miss that impact. Can you think of a way to address this?

5M5. This is an open-ended question with many good, expanding answers. Here's the basic outline of an approach. The first two implied variables are the rate of obesity O and the price of gasoline P. The first proposed mechanism suggests that higher price P reduces driving D, which in turn increases exercise X, which then reduces obesity O. As a set of regressions, this mechanism implies:

(1) D as a declining function of P
(2) X as a declining function of D
(3) O as a declining function of X

In other words, for the mechanism to work, each predictor above needs be negatively associated with each outcome. Note that each outcome becomes a predictor. That's just how these causal chains look. A bunch of reasonable control variables could be added to each of the regressions above. Consider for example that a very wealthy person will be more insensitive to changes in price, so we might think to interact P with income. The second proposed mechanism suggests that price P reduces driving D which reduces eating out E which reduces obesity O. A similar chain of regressions is implied:

(1) D as a declining function of P
(2) E as an increasing function of D
(3) O as an increasing function of E

5H1. For the graph $M \rightarrow A \rightarrow D$, the implications are that M is independent of D when conditioning on A. You can check this with `dagitty` if you aren't sure:

R code 5.3

```
library(dagitty)
dag_5H1 <- dagitty("dag{M->A->D}")
impliedConditionalIndependencies(dag_5H1)
```

```
D _||_ M | A
```

We can check these with the data, provided we are willing to make some additional statistical assumptions about the functions that relate each variable to the others. The only functions we've used so far in the book are linear (additive) functions. The implication above suggests that a regression of D on both M and A should show little association between D and M. You know from the chapter that this is true in the divorce data sample.

So the data are consistent with this graph. But can you think of a way that marriage rate M would causally influence median age of marriage A? If you cannot, then maybe this graph fails on basic scientific grounds. No data are required.