Statistical Rethinking Chapther 6

Shinsuke Nakagami

Research School of Economics, ANU

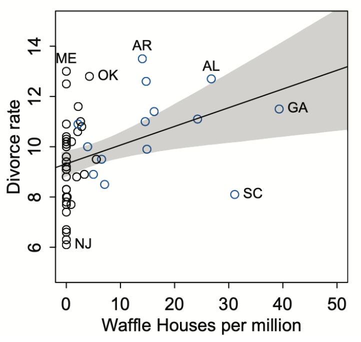
Cont

1. Introduction
2. Spurious association
3. Masked relationship
4. Categorical variables

Intro

Introduction

- Waffle House locations are associated with some of the nation's highest divorce rates.
- States with many Waffle Houses per person have some of the highest divorce rates, but the lowest divorce rates are found where there are zero Waffle Houses.
- This is an example of the spurious correlation, and the common cause is the South.
- Multiple regression is a common tool used to address such cases.
- Key reasons for using multiple regression models include:
    - Statistical "control" for confounds,
    - Multiple and complex causation, and
    - Interactions.



5.1 Spurious association

5.1 Spurious association

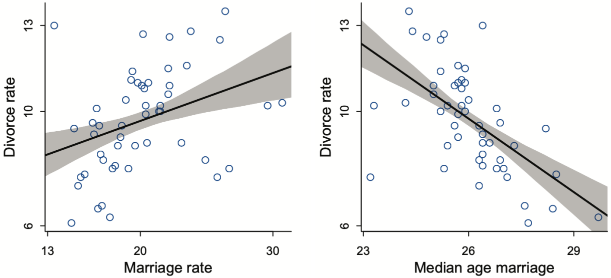
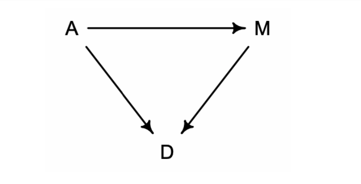


Figure 5.2. Divorce rate is associated with both marriage rate (left) and median age at marriage (right). Both predictor variables are standardized in this example. The average marriage rate across States is 20 per 1000 adults, and the average median age at marriage is 26 years.

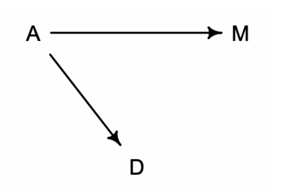
- Is there an association between Divorce rate and Marriage rate?
- Is there an association between Divorce rate and Age?

5.1.1 Think before you regress

- Causal relationships between variables are represented using DAG (Directed Acyclic Graph).
    - Directed: the connections have arrows indicating directions of causal influence.
    - Acyclic: causes do not eventually flow back on themselves.
    - Graph: it consists of nodes and arrows.

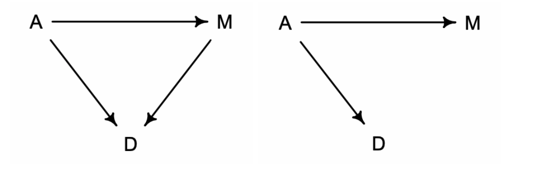


- The DAG represents the relationships among divorce rate (D), marriage rate (M), and the median age at marriage (A).
  - ► A directly influences D
  - ► M directly influences D
  - ► A directly influences M
- Age of marriage influences divorce in two ways: $A \to D$ and $A \to M \to D$.
- In this case, the regression of D on A tells us the total influence of age at marriage to divorce rate, including the indirect effect $A \to M \to D$.
- Even if A has no direct effect on D, it could still be associate with D, and this type of relationship is known as meditation.



- It is possible to consider the case where there is no association between marriage rate (M) and the divorce rate (D).
- The DAG no longer has an association between M and D, but A is a common cause for both M and D, making spurious correlation.
- The posterior distribution from the regression of D on M can be still captured by the DAG.

5.1.2. Testable implications.



- The testable implications are used for checking the conditional independencies.
- They are the statements of
  - which variables should be associated with one another (or not) in the data, and
  - which variables become dis-associated when we condition on some other set of variables.
- Informally, conditioning on Z means to fix (learn) the value Z.
- After fixing Z, if X does not give any further information on Y, we say X and Y are independent conditional on Z, denoted by $Y \perp X \mid Z$.
- For the left DAG, $D \not\perp A \quad D \not\perp M \quad A \not\perp M$
  (all pairs of variables should be associated, whatever we condition on.).
- For the right DAG, as A influences both D and M, once we condition on A, M does not give further information on D, which is written as $D \perp M \mid A$.

5.1.3. Multiple regression notation.

An example of the multiple regression is given as follows

$$\begin{aligned} D_i &\sim \text{Normal}\,(\mu_i, \sigma) && \text{[probability of data]} \\ \mu_i &= \alpha + \beta_M M_i + \beta_A A_i && \text{[linear model]} \\ \alpha &\sim \text{Normal}(0, 0.2) && \text{[prior for } \alpha \text{]} \\ \beta_M &\sim \text{Normal}(0, 0.5) && \text{[prior for } \beta_M \text{]} \\ \beta_A &\sim \text{Normal}(0, 0.5) && \text{[prior for } \beta_A \text{]} \\ \sigma &\sim \text{Exponential}(1) && \text{[prior for } \sigma \text{].} \end{aligned}$$

- Once we run the regression, we find that β_M is close to 0, implying the fact that A is a common cause for M and D.

5.1.5. Plotting multivariate posteriors.

Three examples of interpretive plots:

1. Predictor residual plots.
   These plots show the outcome against residual predictor values.
2. Posterior prediction plots.
   These show model-based predictions against raw data, or otherwise display the error in prediction. They are tools for checking fit and assessing predictions. They are not causal tools.
3. Counterfactual plots.
   These show the implied predictions for imaginary experiments. These plots allow you to explore the causal implications of manipulating one or more variables.

5.1.5.1. Predictor residual plots.

- Firstly, we run the regression of Marriage rate (M) on Age at marriage (A) to compute predictor residuals.
- This manipulation corresponds to control/remove the influence from A to M.
- Secondly, we plot the relation between the marriage residuals controlled by A and Divorce rate (D).
- The plot visually shows how marriage rate after controlling A will influence D.

- The regression is given as follows

$$
\begin{aligned}
M_i &\sim \text{Normal}\left(\mu_i, \sigma\right)\\
\mu_i &= \alpha + \beta A_i\\
\alpha &\sim \text{Normal}(0, 0.2)\\
\beta &\sim \text{Normal}(0, 0.5)\\
\sigma &\sim \text{Exponential}(1).
\end{aligned}
$$

  M is marriage rate and A is median age at marriage.
  The predictor residuals can be computed by subtracting the observed marriage rate in each State from the predicted rate.

- The upper left shows the regression with line segments for each residual.
- The residuals are variation that A cannot explain.
- The lower left shows the plot of relationship between the the residuals and D.
- The vertical dashed line represents 0 residuals.
- This figure shows that average divorce rate on both sides of the line is about the same, and little relationship between M and D.
- The intuition behind the right figures is the same but, we observe a negative (causal?) relationship between A and D.

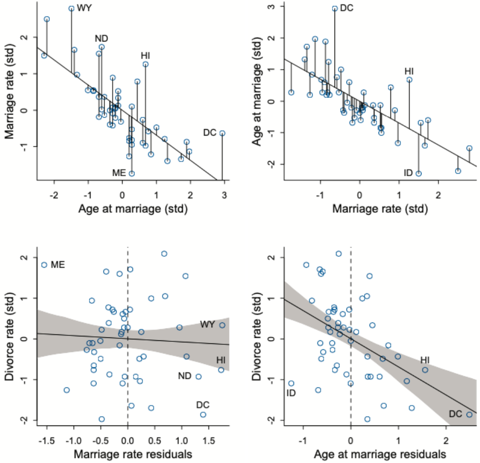


FIGURE 5.4. Understanding multiple regression through residuals. The top row shows each predictor regressed on the other predictor. The lengths of the line segments connecting the model's expected value of the outcome, the regression line, and the actual value are the *residuals*. In the bottom row, divorce rate is regressed on the residuals from the top row. Bottom left: Residual variation in marriage rate shows little association with divorce rate. Bottom right: Divorce rate on age at marriage residuals, showing remaining variation, and this variation is associated with divorce rate.

5.1.5.2. Posterior prediction plots.

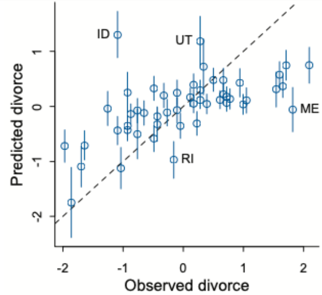


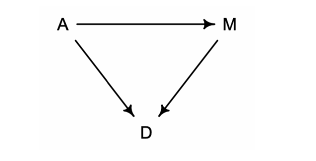
Figure 5.5. Posterior predictive plot for the multivariate divorce model, `m5.3`. The horizontal axis is the observed divorce rate in each State. The vertical axis is the model's posterior predicted divorce rate, given each State's median age at marriage and marriage rate. The blue line segments are 89% compatibility intervals. The diagonal line shows where posterior predictions exactly match the sample.

- We verify how well the statistical model explains the observed data.
- The figure plots the predicted divorce rate against the actual observed divorce rate.
- There is a tendency for states with high divorce rates to be underestimated and states with low divorce rates to be overestimated (regression toward the mean).
- Some states (ID, UT) deviate significantly from the model's predictions.

5.1.5.3. Counterfactual plots.

If the median age of marriage in Utah had been higher, what would the divorce rate have been?

- Consider manipulating A that influences D in two ways: $A \to D$ and $A \to M \to D$.
- Firstly, we consider the regression of A on M, and the regression of A and M on D.
- Next, determine the range of values for A.
- We simulate how M will change if we manipulated variable A.
- Using manipulated A and simulated M, we can simulate how D will change.
- Note that we need to simulate the influence of A on M before we simulate the joint influence of A and M on D.



5.2. Masked relationship

- Consider the problem where there are two predictor variables that are correlated with one another. However, one of these is positively correlated with the outcome and the other is negatively correlated with it.
- Multiple regression can measure the direct effects of multiple factors that are not apparent in simple bivariate relationships.
- The question:
  To what extent is the energy content of milk, measured here by kilocalories, is related to the percent of the brain mass that is neocortex.
- We consider three variables: Kilocalories per gram of milk K, average body weight (mass) of females M, and percentage of neocortex relative to total brain mass N.
- First, we look at the relationships from the regression of K on N and K on M respectively, and then consider the multiple regression of K on both M and N.

- The top two plots show the bivariate regression of K with N and K with M, revealing the weak association with wide compatibility intervals.
- The bottom two plots show the counterfactual results from the multiple regression incorporating both N and M:

$$
\begin{aligned}
K_i &\sim \text{Normal}\left(\mu_i, \sigma\right) \\
\mu_i &= \alpha + \beta_N N_i + \beta_M M_i \\
\alpha &\sim \text{Normal}(0, 0.2) \\
\beta_N &\sim \text{Normal}(0, 0.5) \\
\beta_M &\sim \text{Normal}(0, 0.5) \\
\sigma &\sim \text{Exponential}(1).
\end{aligned}
$$

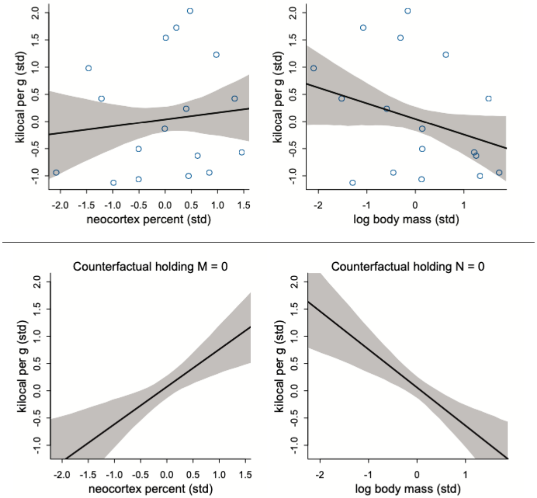
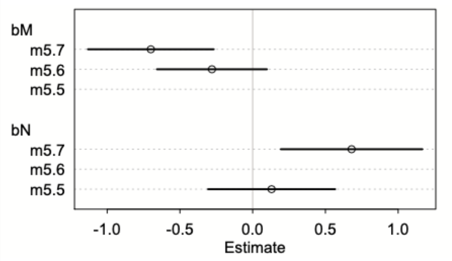
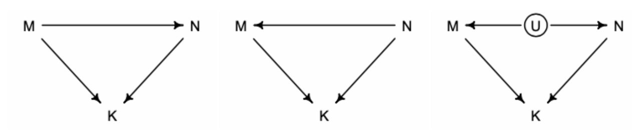


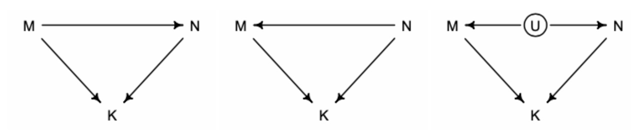
FIGURE 5.9. Milk energy and neocortex among primates. In the top two plots, simple bivariate regressions of kilocalories per gram of milk (K) on (left) neocortex percent (N) and (right) log female body mass (M) show weak associations. In the bottom row, a model with both neocortex percent (N) and log body mass (M) shows stronger associations.



- If the regression incorporates both N and M, the means of the posterior distributions of β_N and β_M clearly move away from the 0.0.

- In this context, N is positively correlated with K, and M is negatively correlated with K.
- Also, both explanatory variables are positively correlated with one another.
- In such cases, the regression on only one predictor canceled out the two effect.
- For example, if M increases, K will decrease.
  At the same time, N will increase, which lead to an increase in K.
- By incorporating both predictors, we can evaluate how one predictor influences the outcome while accounting for the influence of the other predictor.
- These graphs are consistent with data.





- These DAGs have the same conditional independencies, which is known as a Markov equivalence.
- This implies that all pairs of variables are associated, regardless of what we condition on. On the other words, any pair of variables is not independent (connected by some path), regardless of what we condition on.

5.3. Categorical variables

- A common question for statistical methods is to what extent an outcome changes as a result of presence or absence of a category.
- A category means discrete and unordered, e.g., sex, developmental status, and geographical region.

5.3.1. Binary categories

- The first approach uses a dummy variable/indicator variable, for example, one that takes the value 1 for male and 0 for female.

$$\begin{aligned} h_i &\sim \text{Normal}\left(\mu_i, \sigma\right) \\ \mu_i &= \alpha + \beta_m m_i \\ \alpha &\sim \text{Normal}(178, 20) \\ \beta_m &\sim \text{Normal}(0, 10) \\ \sigma &\sim \text{Uniform}(0, 50) \end{aligned}$$

  where h is the height and m is the dummy variable.

- α represents the average height for female.
  β_m represents the expected difference between males and females in height.
- In this approach, we need to set two priors for α and β_m.
  The prior prediction for male height will have more uncertainty because it depends on two priors, α and β_m.

- Another approach is to use an index variable.
- This approach uses an index variable to assign a different integer to each category. The model then uses parameters indexed by these integers, such as α_1 for the avr height for females and α_2 for the avr height for males.
- The model becomes:

$$\begin{aligned} h_i &\sim \text{Normal}\left(\mu_i, \sigma\right) \\ \mu_i &= \alpha_{\text{sex }[i]} \\ \alpha_j &\sim \text{Normal}(178, 20) \quad \text{ for } j = 1..2 \\ \sigma &\sim \text{Uniform}(0, 50). \end{aligned}$$

- This approach allows us to reflect the idea that the prior uncertainty about the mean height for each category is the same.
- The difference in average height between females and males can be obtained by sampling from the posterior distribution and calculating the difference. This calculated difference is called a contrast.
- This approach can easily extended when there are more than two categories.

5.5. Practice

5H1. In the divorce example, suppose the DAG is: $M \to A \to D$. What are the implied conditional independencies of the graph? Are the data consistent with it?

- $M \not\perp A$, $A \not\perp D$, and $M \not\perp D$.
- $M \perp D \mid A$, which is consistent with the lower left plot of the Figure 5.9.