SAT data exercises:

15.10. The least-squares regression line is

$$y = 602.4 - 1.142x,$$

or, in other words,

$$\text{Average math SAT score} = 602.4 - (1.142 \times \% \text{ seniors taking the SAT}).$$

(a) The slope $b = -1.142$ is negative. This means the relationship between average math SAT score (y) and the percentage of seniors taking the SAT (x) is negative. More specifically, $b = -1.142$ is interpreted as follows:

> "For a one-unit increase in the percentage of seniors taking the SAT, we would expect the average math SAT score to **decrease** by 1.142.

(b) To make the prediction, simply plug in "61%" for "x" in the regression equation:

$$\begin{aligned} y &= 602.4 - 1.142x \\ &= 602.4 - (1.142 \times 61) \;=\; 532.7. \end{aligned}$$

The regression equation (model) would predict Georgia to have an average SAT math score of 532.7. This is slightly larger than what Georgia actually had.
(c) This is a question about extrapolation, that is, making predictions using an x value that outside the range of the data. However, the graph shows that basically all percentages are included−ranging from 0 to 100%. There do not appear to be any values of x (% seniors taking the SAT) that would be an extrapolation. That is, using any percentage in the range 0-100% would be justified for prediction based on the observed data.

15.12. (a) The correlation $r = -0.86$ implies a strong negative linear relationship between between average math SAT score (y) and the percentage of seniors taking the SAT (x). This is clear from the figure (Moore and Notz, pp 331).
(b) The square of the correlation is

$$r^2 = (-0.86)^2 \approx 0.74.$$

Interpretation: Approximately 74% of the variability in the average SAT math score state-level data is explained by the straight-line relationship with the percentage of seniors taking the SAT exam.

- This means that approximately 26% of the variability in the average SAT math score state-level data is explained by other sources of variation; e.g., state-level resources for teaching, differences in intelligence/test-taking ability of students in different states, differences in state-level education quality/policies, etc.

Manatee data exercises:

Note: The data set used to make the scatterplot shown in the text (Figure 14.12, Moore and Notz, pp 333) used data only up through 2016. The manatee data we looked at in the notes went up through 2018.

15.11. The authors are asking you to calculate r^2 and interpret it. The square of the correlation is

$$r^2 = (0.94)^2 \approx 0.88.$$

Interpretation: Approximately 88% of the variability in the number of manatee deaths due to boats is explained by the straight-line relationship with the number of boat registrations in Florida.

- This means that approximately 12% of the variability in the number of manatee deaths due to boats is explained by other sources of variation; e.g., weather-related sources, heavy/light traffic days, time of day, manatee mating seasons, etc.

15.13. The least-squares regression line is

$$y = -47.16 + 0.136x,$$

or, in other words,

$$\text{number of manatee deaths} = -47.16 + (0.136 \times \text{number of boat registrations}).$$

The slope $b = 0.136$ is interpreted as follows:

> "For a one-unit increase in the number of boat registrations (in 1000s), we would expect the number of manatee deaths to **increase** by 0.136."

To make the prediction, simply plug in "1000" for "x" in the regression equation:

$$\begin{aligned} y &= -47.16 + 0.136x \\ &= -47.16 + (0.136 \times 1000) \;=\; 88.84. \end{aligned}$$

The regression equation (model) would predict the number of manatee deaths to be about 89.

Wine data exercises:

15.15. (a) The scatterplot is on the next page (top).
(b) There is a moderate negative linear relationship between the liters of alcohol consumed from wine and the death rate (per 100,000 people).
(c) The correlation is $r = -0.645$, which I checked using R:

```
> cor(liters.alcohol,death.rate)
[1] -0.645
```

This value agrees with the description in part (b). The correlation is negative, so the linear relationship between the variables is negative (e.g., as one variable increases, the other tends to decrease). And, there is a linear relationship between the variables, and this relationship is moderately strong.

15.18. I'm just going to use R to superimpose the least-squares regression line (like we did in the notes), and then I will make the two predictions separately. See next page (bottom).

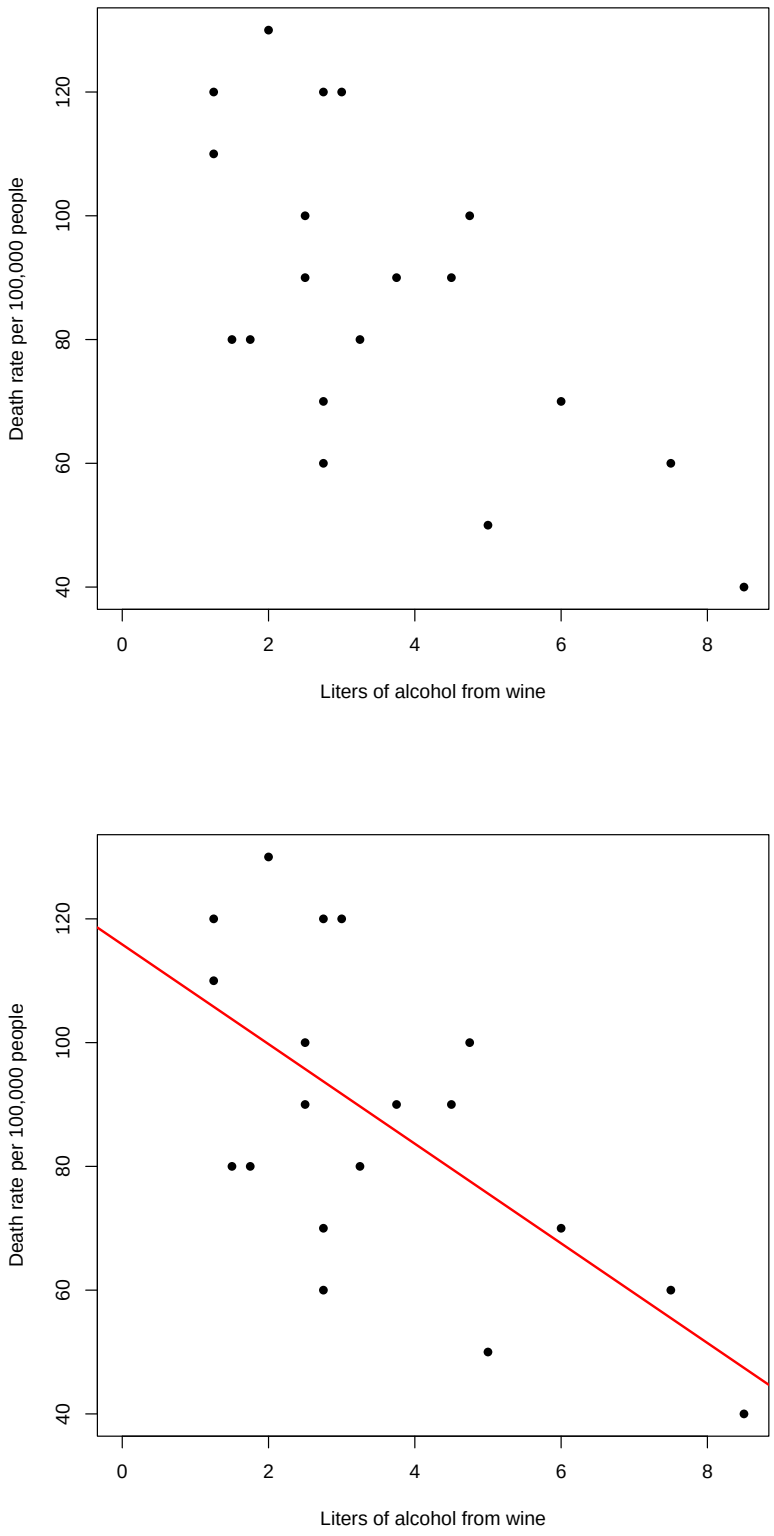

Figure 1: Wine data. Liters of alcohol consumed from wine (x) and death rate per 100,000 people (y) for 19 developed countries.

R code:

```
# Wine data
liters.alcohol = c(3.25,4.75,2.75,1.5,4.5,3,8.5,3.75,1.25,2,7.5,2.75,2.5,1.75,
    5,2.5,6,2.75,1.25)
death.rate = c(80,100,60,80,90,120,40,90,110,130,60,70,100,80,50,90,70,120,120)
plot(liters.alcohol,death.rate,xlab="Liters of alcohol from wine",
    ylab="Death rate per 100,000 people",xlim=c(0,max(liters.alcohol)),pch=16)
cor(liters.alcohol,death.rate) # correlation

fit = lm(death.rate~liters.alcohol) # least squares regression line
fit
plot(liters.alcohol,death rate,xlab="Liters of alcohol from wine",
    ylab="Death rate per 100,000 people",xlim=c(0,max(liters.alcohol)),pch=16)
abline(fit,col="red",lwd=2) # superimpose line
```

Implementation in R:

```
> fit = lm(death.rate~liters.alcohol)
> fit

Coefficients:
   (Intercept)  liters.alcohol
        115.86           -8.05
```

The least-squares regression line is

$$y = 115.86 - 8.05x,$$

or, in other words,

$$\text{Death rate} = 115.86 - (8.05 \times \text{amount of alcohol from wine}).$$

To make the predictions, simply plug in the value of x" in the regression equation. When $x = 1$ liters per year, we would predict the death rate (per 100,000 people) to be

$$\begin{aligned} y &= 115.86 - 8.05x \\ &= 115.86 - (8.05 \times 1) \;=\; 107.81. \end{aligned}$$

When $x = 8$ liters per year, we would predict the death rate (per 100,000 people) to be

$$\begin{aligned} y &= 115.86 - 8.05x \\ &= 115.86 - (8.05 \times 8) \;=\; 51.46. \end{aligned}$$

15.20. The analyses in the previous two exercises might suggest that "the more wine you drink, the better." There are major flaws with this reasoning.

1. The data we have are averages for countries−they are not for individual people. When averages are used, this necessarily reduces variation and so the relationship will be stronger on the country level than it would for individuals. This is a phenomenon known as "ecological association."

2. Correlation does not imply causation! We cannot conclude that drinking more wine *causes* a decrease in the death rate−even at the country level. There are many lurking variables in the way. Wine is more commonly associated with "the affluent," and these populations may have better health practices overall, better health care systems overall, etc. These variables are probably strongly associated with the death rate.

15.29. If we were blindly make this prediction, we would obtain

$$\begin{aligned} y &= 115.86 - 8.05x \\ &= 115.86 - (8.05 \times 150) \;=\; -1091.64. \end{aligned}$$

First of all, death rates cannot be negative, so this doesn't even make sense. Second, the value $x = 150$ is well outside the range of the data for countries in the study. Look at the scatterplot. Countries with alcohol consumption ranging from 1-8.5 liters per person were used. Making a prediction for a country with 150 liters per person is a severe **extrapolation**.

Additional exercises:

15.16. The exercise uses Figure 15.6 (Moore and Notz, pp 363).

(a) There is a strong positive linear relationship between the state percentages voting for President Obama in 2008 and the state percentages voting for President Obama in 2012. There is an outlier in the upper right corner, but this is most likely Washington DC which votes overwhelmingly for Democrat candidates.
(b) I graphed the least-squares regression line using R; see next page.
(c) The authors are asking you to calculate r^2 and interpret it. The square of the correlation is

$$r^2 = (0.983)^2 \approx 0.966.$$

Interpretation: Approximately 96.6% of the variability in the state percentages voting for President Obama in 2012 is explained by the straight-line relationship with the state percentages voting for President Obama in 2008.

- This means that approximately 3.4% of the variability in the state percentages voting for President Obama in 2012 is explained by other sources of variation, for example, differences in state-level opinions about President Obama between 2008-2012, differences in the voting electorate, etc.

15.25. No, it is not possible. The correlation r and the slope of the least-squares regression line b will always have the same sign. Remember, the correlation measures the strength and the direction of the linear relationship between two quantitative variables.

- If the correlation r is positive, then the two variables have a positive linear relationship. This will mean the slope b is also positive.

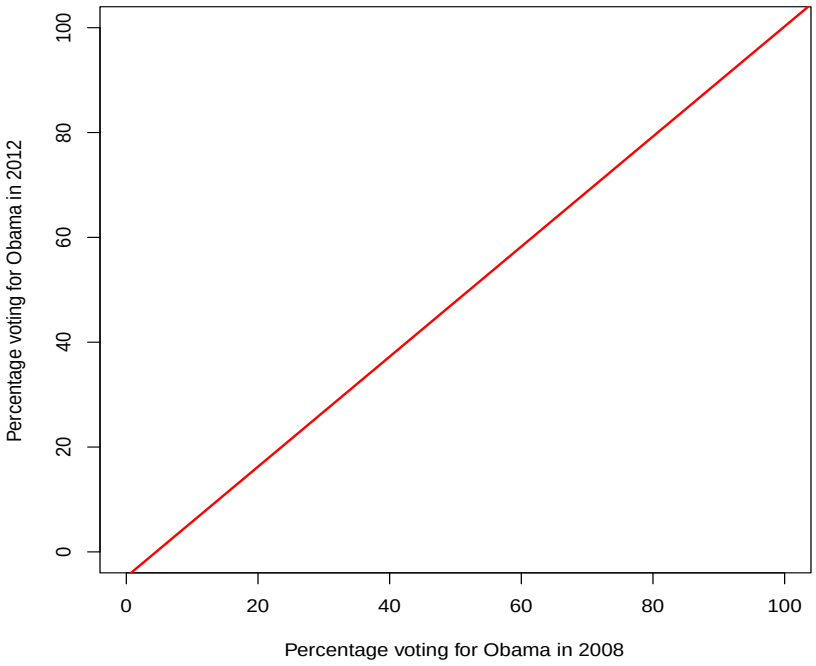

Figure 2: Least squares regression line: $y = -4.75 + 1.05x$.

- If the correlation r is negative, then the two variables have a negative linear relationship. This will mean the slope b is also negative.