Chapter 9 Textbook exercises

Solutions to even-numbered questions
Statistics and statistical programming
Northwestern University
MTS 525

Aaron Shaw

November 11, 2020

All exercises taken from the *OpenIntro Statistics* textbook, 4^{th} edition, Chapter 9.

9.4 Absenteeism in Australian schools

(a) Here's a way to write the equation:

$$\widehat{days} = 18.93 - (9.11 \times ethnicity) + (3.10 \times sex) + (2.15 \times learner\ status)$$

(b) Let's go through these one-by-one. I should note that I think the labeling of these variables encode racial and gender biases in some quiet ways that are irresponsible:

- b_{eth}: On average, not-aboriginal students are estimated to be absent 9.11 days less than aboriginal students.
- b_{sex}: On average, male students are estimated to be absent 3.10 days more than female students.
- b_{lrn}: On average, students classified as "slow learners" are estimated to be absent 2.15 days more than those classified as "average learners."

(c) A residual for some observation i is the observed outcome minus the fitted value ($y_i - \hat{y}_i$). I can calculate $\hat{y}_i$ by plugging the observed predictor values into the regression equation above in part (a):

```
18.93 - (9.11*0) + (3.10 * 1) + (2.15 * 1)
```

```
## [1] 24.18
```

The observed outcome for this student (y_i) was 2 days absent. So the residual is $2 - 24.18 = -22.18$.

(d) Formulas for this appear below. Note that I denote the variance as σ^2, the residuals as e, and the outcome y.

$$R^2 = 1 - \frac{\sigma_e^2}{\sigma_y^2} = 1 - \frac{240.57}{264.17} = 0.0893$$

$$R^2_{adj} = 1 - \frac{\frac{\sigma_e^2}{(n-p-1)}}{\frac{\sigma_y^2}{n-1}} = 1 - \frac{\frac{240.57}{146-3-1}}{\frac{264.17}{146-1}} = 0.0701$$

9.16 O-rings

(a) The damaged O-rings almost all occurred at the lower launch-time temperatures, with the lowest launch temperature accounting for nearly half of the total number of damaged O-rings.

(b) The model suggests that lower launch-time temperatures result in a higher probability of O-ring damage. The coefficient of the "Temperature" term is negative with a very small (proportionally speaking) standard error. It is statistically significant with a p-value near 0 ($H_0: \beta_{temp} = 0$, $H_A: \beta_{temp} \neq 0$), indicating that the data provide evidence that the coefficient is likely different from 0. By exponentiating the coefficient (see the R-code below), we see that a one degree farenheit increase in launch-time temperature is associated with 81% as large odds of O-ring damage occurring. In other words, the model indicates that higher launch temperatures associate with reduced odds of O-ring damage.

```
exp(-.2162)
```

```
## [1] 0.8055742
```

(c) The corresponding logistic model where $\hat{p}_{damage}$ represents the probability of a damaged O-ring:

$$log(\frac{\hat{p}_{damage}}{1-\hat{p}_{damage}}) = 11.663 - 0.2162 \times \ Temperature$$

(d) Given the high stakes in terms of human lives and vast costs involved, concerns about the relationship between O-rings and launch-time temperature seem more than justified from this data. The significant negative association between temperature and O-ring damage suggest increased potential for failures at low launch temperatures. That said, several limitations of the data, modeling strategy, and estimates should be kept in mind. See my answer to part (c) of 9.18 below for more on this.

9.18 More O-rings

(a) Let's do this in R. Note that we'll need to plug in the values for temperature *and* do some algebra with the natural logarithm parts of the formula (on the left hand side above) to find the predicted probability of O-ring damage. I'll solve it by writing a little function `probs` that takes fitted values from the model and runs them through the inverse logistic function to return probabilities (see the textbook for some of the algebraic details here). I can test my function on some of the example model-estimated probabilities provided in the textbook:

```
probs <- function(x){
  p.hat <- exp(11.663-(0.2162*x) )
  pred <- p.hat / (1 + p.hat) # inverse logit
  return(round(pred, 3))
}

## examples
probs(57)
```

```
## [1] 0.341
```

```
probs(65)
```

```
## [1] 0.084
```

Both of those look good, so now I can plug in the values the problem asks me to solve for:

```
vals <- c(51, 53, 55)

probs(vals)
```

```
## [1] 0.654 0.551 0.443
```

(b) I'll use my little function above to build a data frame with the predicted values and plot everything in ggplot.

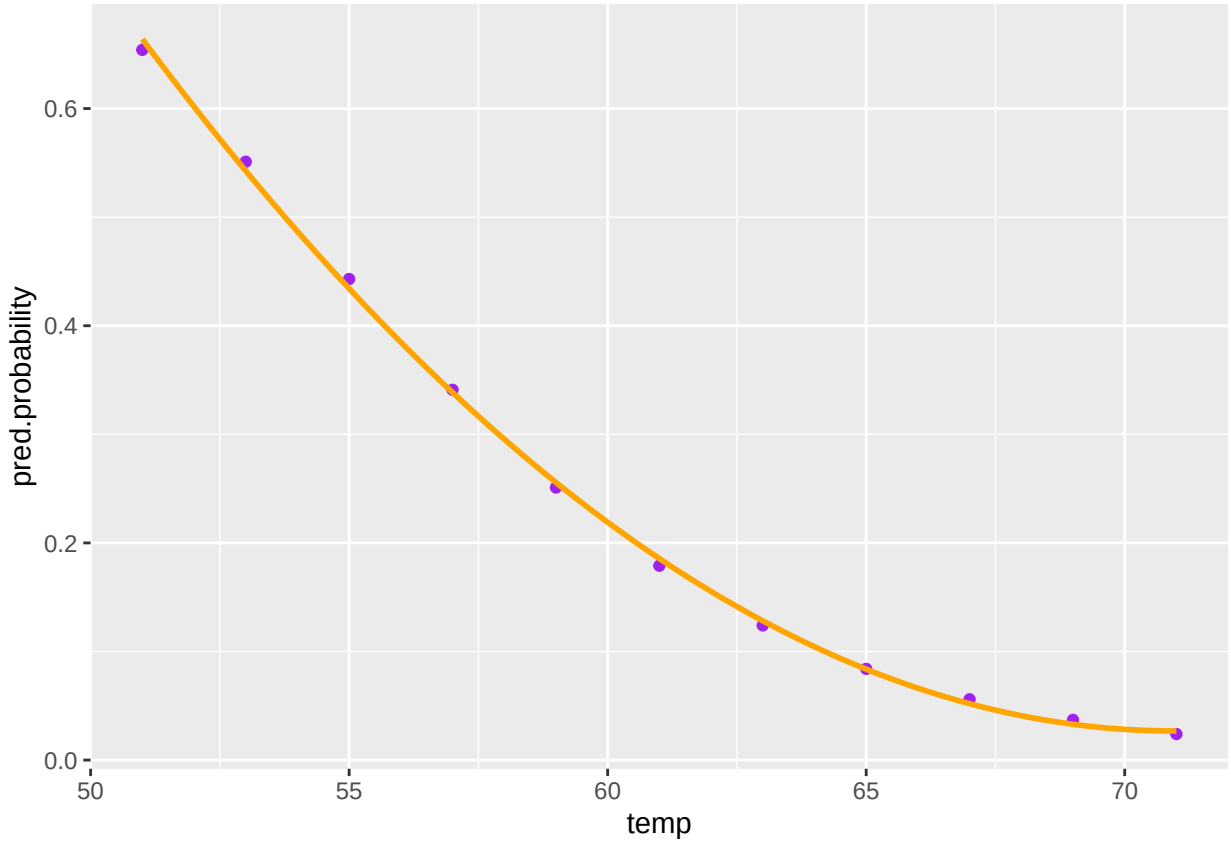
Note that the question asks for a "smooth curve" fit to the dots. There are many ways to do this in ggplot. I demonstrate one here using `geom_smooth()` that fits a quadratic function ($y = x^2$) to the points. You might experiment with the different options for `geom_smooth()` or, for a simpler solution, just try `geom_line()` (with no arguments) instead.

```
temp =  seq(51, 71, by=2) # This creates a vector from 51 to 71, counting by twos

preds <- data.frame( # I'll want the data frame for ggplot
  temp,
  pred.probability = probs(temp) # store the probabilities as another vector
)

library(ggplot2)

ggplot(data=preds, aes(x=temp, y=pred.probability)) +
  geom_point(color="purple") + # Plot the points
  geom_smooth(color="orange", # Add a smooth line
              method="glm", # Create a line fit to the data
              formula = y ~ poly(x, 2), # Using this formula
              se=FALSE) # Don't show standard errors
```



(c) I have several concerns about this application of logistic regression to the problem and data at hand. First, this is a fairly small observational dataset with a lot of potential confounders and threats to the assumptions necessary to identify the model. For instance, the textbook is unclear whether each

mission was treated as an independent trial or each O-ring was treated as an independent trial. Either assumption is problematic. The O-rings within any given mission are probably more similar to each other than to the O-rings on other missions. In addition, it is possible that the O-ring production or installation procedures may have changed across the missions over time. Likewise any of the flight and/or launch procedures may have varied in subtle ways correlated (or not) with the temperature and/or the O-ring outcomes. Any such clustered or time-dependent structures lurking in the data could lead to unobserved bias in the estimates when we model each set of mission-specific or O-ring-specific outcomes as independent events without accounting for these added sources of covariance/clustering.

Furthermore, if the model treats each O-ring as an indpendent trial, about 50% of the observed failures occurred in a single mission—the mission with the lowest observed launch-time temperature. The result is that this one mission with its one launch temperature could drive the model results disproportionately (it generates observations that exert "high leverage" on the model fit to the data). Without knowing ahead of time that temperature was a likely explanation (as compared against any of the other infinite details of that one mission), it's hard to see how NASA analysts necessarily should have drawn this conclusion on the basis of evidence like this.