Practice Exercises: Lesson 6.3 Solutions

Diez, D. M., Çetinkaya-Rundel, M., Barr, C. D. (2019). OpenIntro Statistics (4th ed.). OpenIntro. https://www.openintro.org/book/os/ CC BY-SA 3.0

STAT 1201
Introduction to Probability and Statistics

8.39 (a) The point estimate and standard error are $b_1 = 0.9112$ and $SE = 0.0259$. We can compute a T-score: $T = (0.9112 - 1)/0.0259 = -3.43$. Using $df = 168$, the p-value is about 0.001, which is less than $\alpha = 0.05$. That is, the data provide strong evidence that the average difference between husbands' and wives' ages has actually changed over time. (b) $\widehat{age}_W = 1.5740 + 0.9112 \times age_H$. (c) Slope: For each additional year in husband's age, the model predicts an additional 0.9112 years in wife's age. This means that wives' ages tend to be lower for later ages, suggesting the average gap of husband and wife age is larger for older people. Intercept: Men who are 0 years old are expected to have wives who are on average 1.5740 years old. The intercept here is meaningless and serves only to adjust the height of the line. (d) $R = \sqrt{0.88} = 0.94$. The regression of wives' ages on husbands' ages has a positive slope, so the correlation coefficient will be positive. (e) $\widehat{age}_W = 1.5740 + 0.9112 \times 55 = 51.69$. Since R^2 is pretty high, the prediction based on this regression model is reliable. (f) No, we shouldn't use the same model to predict an 85 year old man's wife's age. This would require extrapolation. The scatterplot from an earlier exercise shows that husbands in this data set are approximately 20 to 65 years old. The regression model may not be reasonable outside of this range.

8.41 There is an upwards trend. However, the variability is higher for higher calorie counts, and it looks like there might be two clusters of observations above and below the line on the right, so we should be cautious about fitting a linear model to these data.

8.43 (a) $r = -0.72 \rightarrow$ (2) (b) $r = 0.07 \rightarrow$ (4) (c) $r = 0.86 \rightarrow$ (1) (d) $r = 0.99 \rightarrow$ (3)

9 Multiple and logistic regression

9.1 (a) $\widehat{baby\_weight} = 123.05 - 8.94 \times smoke$ (b) The estimated body weight of babies born to smoking mothers is 8.94 ounces lower than babies born to non-smoking mothers. Smoker: $123.05 - 8.94 \times 1 = 114.11$ ounces. Non-smoker: $123.05 - 8.94 \times 0 = 123.05$ ounces. (c) H_0: $\beta_1 = 0$. H_A: $\beta_1 \neq 0$. $T = -8.65$, and the p-value is approximately 0. Since the p-value is very small, we reject H_0. The data provide strong evidence that the true slope parameter is different than 0 and that there is an association between birth weight and smoking. Furthermore, having rejected H_0, we can conclude that smoking is associated with lower birth weights.

9.3 (a) $\widehat{baby\_weight} = -80.41 + 0.44 \times gestation - 3.33 \times parity - 0.01 \times age + 1.15 \times height + 0.05 \times weight - 8.40 \times smoke$. (b) $\beta_{gestation}$: The model predicts a 0.44 ounce increase in the birth weight of the baby for each additional day of pregnancy, all else held constant. β_{age}: The model predicts a 0.01 ounce decrease in the birth weight of the baby for each additional year in mother's age, all else held constant. (c) Parity might be correlated with one of the other variables in the model, which complicates model estimation. (d) $\widehat{baby\_weight} = 120.58$. $e = 120 - 120.58 = -0.58$. The model over-predicts this baby's birth weight. (e) $R^2 = 0.2504$. $R^2_{adj} = 0.2468$.

9.5 (a) (-0.32, 0.16). We are 95% confident that male students on average have GPAs 0.32 points lower to 0.16 points higher than females when controlling for the other variables in the model. (b) Yes, since the p-value is larger than 0.05 in all cases (not including the intercept).

9.7 Remove age.

9.9 Based on the p-value alone, either gestation or smoke should be added to the model first. However, since the adjusted R^2 for the model with gestation is higher, it would be preferable to add gestation in the first step of the forward- selection algorithm. (Other explanations are possible. For instance, it would be reasonable to only use the adjusted R^2.)

9.11 She should use p-value selection since she is interested in finding out about significant predictors, not just optimizing predictions.

9.13 Nearly normal residuals: With so many observations in the data set, we look for particularly extreme outliers in the histogram and do not see any.
variability of residuals: The scatterplot of the residuals versus the fitted values does not show any overall structure. However, values that have very low or very high fitted values appear to also have somewhat larger outliers. In addition, the residuals do appear to have constant variability between the two parity and smoking status groups, though these items are relatively minor.
Independent residuals: The scatterplot of residuals versus the order of data collection shows a random scatter, suggesting that there is no apparent structures related to the order the data were collected.
Linear relationships between the response variable and numerical explanatory variables: The residuals vs. height and weight of mother are randomly distributed around 0. The residuals vs. length of gestation plot also does not show any clear or strong remaining structures, with the possible exception of very short or long gestations. The rest of the residuals do appear to be randomly distributed around 0.
All concerns raised here are relatively mild. There are some outliers, but there is so much data that the influence of such observations will be minor.

9.15 (a) There are a few potential outliers, e.g. on the left in the `total_length` variable, but nothing that will be of serious concern in a data set this large. (b) When coefficient estimates are sensitive to which variables are included in the model, this typically indicates that some variables are collinear. For example, a possum's gender may be related to its head length, which would explain why the coefficient (and p-value) for `sex_male` changed when we removed the `head_length` variable. Likewise, a possum's skull width is likely to be related to its head length, probably even much more closely related than the head length was to gender.

9.17 (a) The logistic model relating $\hat{p}_i$ to the predictors may be written as $\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = 33.5095 - 1.4207 \times sex\_male_i - 0.2787 \times skull\_width_i + 0.5687 \times total\_length_i - 1.8057 \times tail\_length_i$. Only `total_length` has a positive association with a possum being from Victoria. (b) $\hat{p} = 0.0062$. While the probability is very near zero, we have not run diagnostics on the model. We might also be a little skeptical that the model will remain accurate for a possum found in a US zoo. For example, perhaps the zoo selected a possum with specific characteristics but only looked in one region. On the other hand, it is encouraging that the possum was caught in the wild. (Answers regarding the reliability of the model probability will vary.)

9.19 (a) False. When predictors are collinear, it means they are correlated, and the inclusion of one variable can have a substantial influence on the point estimate (and standard error) of another. (b) True. (c) False. This would only be the case if the data was from an experiment and x_1 was one of the variables set by the researchers. (Multiple regression can be useful for forming hypotheses about causal relationships, but it offers zero guarantees.) (d) False. We should check normality like we would for inference for a single mean: we look for particularly extreme outliers if $n \geq 30$ or for clear outliers if $n < 30$.

9.21 (a) `exclaim_subj` should be removed, since it's removal reduces AIC the most (and the resulting model has lower AIC than the None Dropped model). (b) Removing any variable will increase AIC, so we should not remove any variables from this set.

9.23 (a) The equation is:

$$\begin{aligned}\log\left(\frac{p_i}{1-p_i}\right) = -0.8124 \\ &- 2.6351 \times \texttt{to\_multiple} \\ &+ 1.6272 \times \texttt{winner} \\ &- 1.5881 \times \texttt{format} \\ &- 3.0467 \times \texttt{re\_subj}\end{aligned}$$

(b) First find $\log\left(\frac{p}{1-p}\right)$, then solve for p:

$$\begin{aligned}&\log\left(\frac{p}{1-p}\right) \\ &\quad = -0.8124 - 2.6351 \times 0 + 1.6272 \times 1 \\ &\qquad - 1.5881 \times 0 - 3.0467 \times 0 \\ &\quad = 0.8148 \\ &\frac{p}{1-p} = e^{0.8148} \quad \rightarrow \quad p = 0.693\end{aligned}$$

(c) It should probably be pretty high, since it could be very disruptive to the person using the email service if they are missing emails that aren't spam. Even only a 90% chance that a message is spam is probably enough to warrant keeping it in the inbox. Maybe a probability of 99% would be a reasonable cutoff. As for other ideas to make it even better, it may be worth building a second model that tries to classify the importance of an email message. If we have both the spam model and the importance model, we now have a better way to think about cost-benefit tradeoffs. For instance, perhaps we would be willing to have a lower probability-of-spam threshold for messages we were confident were not important, and perhaps we want an even higher probability threshold (e.g. 99.99%) for emails we are pretty sure are important.