STA101 Problem Set 9 - KEY

Summer I, 2021, Duke University

Exercises from the OpenIntro book - 100pts total

Problems include: 9.3, 9.6, 9.8, 9.10, 9.14, 9.19

9.3 (35pts total)

(a) **(5pts)**

$$\widehat{\text{baby weight}} = -80.41 + 0.44\text{gestation} - 3.33\text{parity} - 0.01\text{age} + 1.15\text{height} + 0.05\text{weight} - 8.40\text{smoke}$$

(b) **(5pts for gestation interpretation, 5pts for age interpretation)** $\beta_{\text{gestation}}$: The model predicts a 0.44 ounce increase in the birth weight of the baby for each additional day in length 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) **(5pts for reasonable explanation)** Parity might be correlated with one of the other variables in the model, which introduces collinearity and complicates model estimation.

(d) **(5pts)**

$$\widehat{babyweight} = -80.41 + 0.44(284) - 3.33(0) - 0.01(27) + 1.15(62) + 0.05(100) - 8.40(0) = 120.58$$

$$e = \text{baby weight} - \widehat{babyweight} = 120 - 120.58 = -0.58$$

(e) **(5pts for R^2, 5pts for adjusted R^2)** The R^2 and the adjusted R^2 can be calculated as follows:

$$R^2 = 1 - \frac{\text{Var}(e_i)}{\text{Var}(y_i)} = 1 - \frac{249.28}{332.57} = 0.2504$$

$$R^2_{adj} = 1 - \frac{\text{Var}(e_i)/(n-p-1)}{\text{Var}(y_i)/(n-1)} = 1 - \frac{249.28/(1236-6-1)}{332.57/(1236-1)} = 0.2468$$

9.6 (15pts total)

(a) **(5pts for CI, 5pts for interpretation)** A 95% confidence interval for the slope of height can be calculated as follows:

$$\begin{aligned} df &= n - p - 1 = 31 - 2 - 1 = 28 \\ t^*_{28} &= 2.05 \\ b_{height} \pm t^*_{df} SE_{height} * &= 0.34 \pm 2.05 \times 0.13 \\ &= 0.34 \pm 0.27 \\ &= (0.07, 0.61) \end{aligned}$$

We are 95% confident that for each foot increase in the height of a tree the volume is expected to increase on average by 0.083 to 0.617 cubic feet when controlling for the other variables in the model.

(b) **(3pts for "underestimates", 2pts for the amount)** $\widehat{y} = -57.99 + 0.34 \times 79 + 4.71 \cdot 11.3 = 22.093 < 24.2$. The model underestimates the volume of the tree by $24.2 - 22.093 = 2.107$ cubic feet.

9.8 (10pts total)

Remove learner status.

9.10 (10pts total)

Based on both the p-value and R^2_{adj} ethnicity should be added to the model first.

9.14 (10pts total)

(5pts for "not appropriate", 5pts for justification)

The residuals are right skewed (skewed to the high end). Horror movies seem to show a much different pattern than the other genres. While the residuals plots show a random scatter over years and in order of data collection, there is a clear pattern in residuals for various genres, which signals that this regression model is not appropriate for these data.

9.19 (20 pts total)

(a) **(3pts for "False", 2pts for explanation)** 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) **(5pts for "True")** True.

(c) **(3pts for "False", 2pts for explanation)** 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) **(3pts for "False", 2pts for explanation)** 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 $n < 30$.