STAT 3400 - Homework #X

Alex Ojemann

Due February 15, 2023

Problem 9.6.1

a. False; we fit a logistic curve.

b. False; the residuals aren't evenly scattered because the outcome variable for all data points is either 0 or 1.

c. True

Problem 9.6.2

a. False; logistic curves aren't linear so there will be different increases in y for x changing from 4 to 5 and x changing from 6 to 7.

b. True

c. False; observations that aren't independent of each other can lead to a biased model.

d. False; we complete the model using AIC.

Problem 9.6.4

a. The relationship between temperatures and damaged O-rings appears to be moderate, nonlinear and decreasing.

b. The log odds of the probability of a damaged O-ring can be modeled as 11.66-0.22*temperature

c. p_ring = e^(11.66-0.22t) / (1+e^(11.66-0.22t))

d. Yes because temperatures below 53 degrees are common and are associated with a high probability of damaged O-rings

Problem 9.6.6

a. 51: 0.608 53: 0.5 55: 0.392

b.

c. A concern in regards to applying logistic regression here is that the outcome variable isn't binary. There are data values that aren't 0 or 1.

Problem 9.6.7

a. We should drop explain_subj because dropping it results in the lowest AIC of all the possible variables dropped and it's lower than it would be without dropping any variables at all.

b. We should drop cc because dropping it results in the lowest AIC of all the possible variables dropped and it's lower than it would be without dropping any variables at all.

c. We shouldn't drop any of these variables because the AIC without dropping any is lower than any of the potential AICs if one of the variables is dropped.

Problem 9.6.8

a. p_spam = eˆ(-0.81-2.64tm1+1.63w-1.63f1-3.05rs1) / (1+eˆ(-0.81-2.64tm1+1.63w-1.63f1-3.05rs1))

b. 0.694

c. There is a tradeoff between the user missing an important email that's mistakenly classified as spam and the user having to click through spam emails because they were mistakenly not classified as spam. Given that the former is more costly, we may want to set the threshold of probability from the model required to put the email in the spam folder relatively high.