4/19/22 Stats Test Corrections

Problem # 3

b) Observation 65 and d) Observation 53
and c) Observation 9

Problem # 4

a) Observation 66
b) Observation 65
d) Observation 53

Problem # 5

a) Holding other predictors constant the model estimates that each drink per day decreases your life span by approximately 3.2656 years.

b) The average life expectancy for this individual will be roughly 93.68 − 3.27(2.5) − 23.44 years.

☆ c) The average difference in lifespan between smokers and non-smokers is that smokers experience a decrease by approximately 23.4392 years in lifespan when you keep other predictors constant in comparison to non-smokers

d) There is a statistically significant relationship between life expectancy and smoking status when you control for the amount alcohol consumed. The Pr(>|t|) column displays that the p-value for SmokerYes is <2e-16, which is much smaller than 0.05 and thus shows that smoking status has a statistically significant relationship with life expectancy.

Problem # 6 - a) curved pattern - linear assump val
resid plot more not in mod. 3
in l'il

a) I would pick Model 3 because a) domain-wise, a model with an interraction term makes sense because cars with certain weights predominantly use one transmission type over another, and b) model 3, unlike model 1 and 2, meets the linearity assumption, as you can see on the stud. resids vs. fit. values graph. The same graph for model 2 and 1 shows curvature, which violates the mlr linearity assumption. No curvature for model 3

b) $H_0: \beta_1 = \beta_2 = \beta_3 = 0$

$H_a: \beta_1 \neq \beta_2 \neq \beta_3 \neq 0$

Null hypothesis: coefficients for weight, transmission type, and the interaction term are all 0.
Alt hypothesis: The predictor terms aren't equal to 0.

The p-value that the summary for model 3 reports is 1.669×10^{-12}, which means that weight, transmission type, and the interaction between the two have a statistically significant relationship with mpg when combined into one model.

Problem # 7

a) ~~[illegible]~~

$H_0: \beta_2 = \beta_3 = 0$

H_a: at least β_2 or β_3 or both not equal to 0.

b) The test that is more reliable is the one in 6B. ~~[illegible]~~ 6b ~~[illegible]~~ is the test of model 3, while 7a is a nested F test of model 1 and model 3. The residuals vs fits plot of model 1, however, is not favorable, as it gives us a funneling pattern, which violates the linearity assumption of MLR. Therefore it would be best if we just performed the test in 6B. ~~[illegible]~~

(mod 1 - reduced
mod 3 - full)

Problem # 8

We could probably add this predictor and perform backward elimination by identifying the term that has the largest P-value and if the P-value is large (e.g. more than 5%), eliminate it and re-fit the new model until all predictors are "significant." Or, we could eliminate predictors that give largest drop in Mallow's C_p until C_p doesn't get smaller after predictor removal. On the other hand, we could try forward selection, by starting with a model with no predictors and finding the best predictor (biggest initial R^2). Add the predictor to the model, if p-value small enough (below %5, e.g.), keep the predictor. If not, drop it. Or, stepwise regression, which starts with forward selection, but uses backward elimination to get rid of predictors that have become redundant.