Stat 21 Test 2 Corrections

Elaine Kim

2. When solving for a nested F-test, we compare a full model to a reduced (nested) model. The reduced model eliminates the group of predictors that we are testing. With categorical predictors, a reduced model should either include or exclude all related categorical predictors. Options (c) and (d) only have one of the predictor variables for "Location"; therefore, these two are not valid.

   Answer) a, b

4. Soil pH and potassium levels are predictors. We can use leverage to assess how much of an influence data points have on the slope of a regression line based on how far they are from $\bar{x}$. The graph shows very unusual points on the right side of the three dashed lines. So, observation points 65, 9, and 53 are likely very unusual w/ respect to soil pH and potassium level.

   Ans) b, c, d

5. a) Drinking alcohol has a negative impact on life expectancy. When smoker status is controlled for, every additional alcoholic drink per day causes a 3.2656 decrease in life expectancy in years.

   c) On average, non-smokers have a lifespan that is 23.2392 years longer than the lifespan of smokes.

6. a) I would choose Model 3. First, in the data for Model 2, we can see that transmission-type M has a non-significant p-value, indicating that it may not have a significant relationship with mpg. Model 1 and Model 3 have relatively similar (and high!) R^2 values. In the studentized residual plot of Model 1, we can see a slight "dip" around 18 ~ 20 (fitted value). This dip is even more noticeable if we remove the point at 10 (fitted value) and 2 (studentized residual). This dip may indicate that the constant variance condition may not be met (= may not have even distribution above & below). There is also slight trailing off at the ends of the residual quantile plot. We do not see these problems (or, they are not as obvious) in Model 3. So, I would use Model 3.

7. (b) Problem 6b indicates an overall ANOVA test of model 3 ($H_0: \beta_1 = \beta_2 = \beta_3 = 0$, H_a: At least one of $\beta_1, \beta_2, \beta_3$ is not 0). Problem 7a indicates a Nested F-test for the categorical predictors in Model 3 ($H_0: \beta_2 = \beta_3 = 0$, H_a: At least one of $\beta_2, \beta_3 = 0$). For the nested F-test, the reduced model is actually Model 1. Because we identified potential errors in the residual plot for Model 1, the nested F-test in problem 7a may be slightly less reliable. ↳ in question 6a

8. Some steps we could take:

- Added variable plot: By removing the predictor of interest, we can see how well the remaining predictors work to model the response. This tells us what unique information may be in the new variable (so, in this case, the displacement of the vehicle's engine).

- Cook's Distance: This takes into account both the standardised / studentized residual and the leverage. With this, we can quantify the amount of influence that a particular data case has on the estimated regression. So, we could perhaps identify the Cook's D_i values for data points in the model with the new predictor and see if any of the values are moderately / very unusual.

- Scatterplot matrix: A scatterplot matrix would allow you to see if there was any multicollinearity between your new predictor and the existing predictors. It would also give you information about the relationship between your new predictor and the response variable

- Mallow's C_p: Takes into account what info might be available in other potential predictors. So, if Mallow's C_p is lowered by adding the predictor, we may want to consider using that predictor in our model.