STAT 021 Midterm 2 Correction

- Sumin Byun

#1 (a), (d)

#2 (b), (c)

#3 (b), (c), (d)

#5-c Given that alcohol consumption in drinks is held constant, on average, life span of smokers is 23.44 years shorter than life span of non-smokers.

#5-d Since the alcohol predictor term is fixed, it is reasonable to assume there is a statistically significant relationship between life expectancy and smoking status because p-value is smaller than 0.05. (p-value is <2e-16. in the t-test.)

#6-a I would choose Model 3 because R^2 value is the highest (0.8853) compared to Model 1 (0.8668) and Model 2 (0.8655). Also, residual plot seems better than Mode 1 and 2 because residual plots of Model 1 and 2 seem to have funneling pattern with slightly non-linearity (slightly curved). Model 3 would be better if we want to have further inference such as tests, because the constant variance and linearity assumption are met. Also, p-value is less than 0.05 (1.669e-12) which means it is reasonable to say this model is effective. Therefore, I would go with Model 3.

#7-a null hypothesis (H_0): $\beta_2 = \beta_3 = 0$

alternative hypothesis (H_A): at least one of β_2 or β_3 is not zero.

#7-b In 7-a, we are doing nested F-test, comparing the full model and the reduced model (Model 1). So we should look at both Model 1 and Model 3 residual plots and normal quantile plots to check the assumptions: normal quantile plot for normality assumption and residual plot for constant variance and linearity assumption. On the other hand, in 6.b, we are doing t-test for coefficient a for Model 3. So we should look at the residual plot and normal quantile plot of Model 3 to check the assumptions mentioned above. For both tests, the other assumptions are met, since independance assumption, zero mean, and randomness assumptions are satisfied with the given information. However, when we look at the residual plot of Model 1, the funneling pattern is detected (it becomes more obvious when we exclude upper left one data point) and we can see the data points show slightly curved pattern. With the one data point on the upper left, constant variance assumption seems fine but linearity assumption seems to be not met for Model 1. For Model 3, the residual plot looks a little bit better since it doesn't have a curve pattern and it indicates that constant variance assumption is also met.

Therefore, for 7.a which is a nested F-test, it is less reliable because the Model 1 shows a violation of assumption while Model 3 is fine. (Model 1 and 3 normal quantile plots seem pretty fine, indicating that normality assumption is met for the both model). 6.b test is more reliable because we are only considering Model 3, not Model 1, and Model 3 residual plot seems fine with the assumptions.

In conclusion, test in problem 6.b is more reliable.

8 We can look at the coefficient of determination (how much the variability of the vehicle fuel consumption is explained by the model) to support the decision. We can compare the value of coefficient of determination of each model with and without the new predictor which is the displacement of the vehicle's engine. If the R^2 is higher in the model with the new predictor, it is statistically supported that it is better to add the new predictor.

Also, we can check the residual plot and normal quantile plot and see if the assumptions are met. We can assess whether the new model meets the assumptions as well as the old model (without the new predictor).

Third, we can use standardized / studentized residuals leverage, and cook's distance to see if this new model results in fewer unusual data points compared to the old model. If so, we can consider adding the new predictor to the model.