Dulce Ventura TEST CORRECTIONS

1) a) & d

5.3)
(5.C) $\hat{y} = 93.68 - 3.2(\text{alcohol}) - 23.44(1)$ (smoker)

$- \hat{y} = 93.68 - 3.2\text{alcohol} - 23.\cancel{44}(0)$ (nonsmoker)

$\hat{y} = \cancel{93.68} - \cancel{3.2}(\text{alcohol}) - 23.44 - \cancel{93.68} + \cancel{3.2}\text{alcohol}$

when controlling for alcohol consumption (keeping alcohol the same), the average difference in lifespan is -23.44 between smokers and non-smokers. In other words, when controlling for alcohol consumption, non-smokers have an average lifespan that is on average 23.44 years longer than smokers.

6.2)
$H_0: \beta_0 = \beta_1 = \beta_2 = \beta_3 = 0$
$H_a: \beta_1 \neq 0$

The p.value is $1.669e^{-12}$ for an $\alpha = 0.05$. The p.value is less than 0.05, so we can reject the null hypothesis. since the p.value represents the probability of $\beta_i = 0$, the chance of that being true is very small (i.e our p.value). So we can reject the null & conclude that at least one of the β_i do not equal 0. In other words, there is a linear relationship between y & at least one of the predictors.

7.1) $H_0: \beta_2 = \beta_3 = 0$ → The reduced model is better.

H_a: At least one $\beta_j \neq 0$ (for $j = 2, 3$) → The full model is better.

7.2) The ~~more reliable tests is the~~
Both an f. test & a t. test depend on the assumptions that the residuals are independent & identically distributed & come from a ~~rand~~ normally distributed sample that has equal variances. Based on the residual plot in model 3, there does appear to be linearity & constant variance. The error also appears to normally distributed for the most part, although the points in the center of the center of the quantiles plot is lifting. This might indicate that the normality assumption may be questioned. The t. test assumes that the data is normally distributed, so the questionable quantile plot might indicate that the f. test is slightly more reliable (the f. test is question 7a). Although, technically if the normality plot is not met, statistical inferences cannot be made including f. tests.

8) In order to decide whether or not to include an additional variable, I would first look at a general scatterplot & look at the individual correlation between variable displacement & mpg. If there seems to be correlation then I will create an addition variable plot. I will use the residual plot of a model without the displacement model (model 3) & then calculate the residual of a model 1 predict displacement using all other predictors. With this plot I will be able to see if this individaal predictor says something more about the model & if it messes up any of the conditions of linearity & constant variance. In addition, you can take a look at the adjusted R squared value of the models with and without the additional variable. If the adj r-squared model improves with the additional variable, then this additional variable may help the model explain more of the variance. Another useful analysis can be to look at a VIF. This will analyze multicollinearity & you can compare the VIF values of a model with & without the additional predictor. If the VIF is greater than a threshold of 5 with the addition of the variable, I would reconsider adding it to the model. Another method would be to calculate Mallows's Cp. If the additional variable increases Mallows's Cp, than we might reconsider including the additional variable as the new predictor may not be explaining a lot of variability.