Problem 2

Which of the following represents a valid reduced model? (Circle all that apply.)

(a) $\hat{biomass} = \hat{\beta}_0 + \hat{\beta}_1 pH + \hat{\beta}_2 K + \hat{\beta}_4 LocationSM$

n - 1 predictor variable

(b) $\hat{biomass} = \hat{\beta}_0$

(c) $\hat{biomass} = \hat{\beta}_0 + \hat{\beta}_1 pH + \hat{\beta}_2 K$

(d) $\hat{biomass} = \hat{\beta}_0 + \hat{\beta}_1 pH + \hat{\beta}_2 K + \hat{\beta}_3 LocationSI$

Answer : B and C

We cannot simply remove 1 level but instead have to remove the whole categorical predictors

Problem 3

Which of the following data points are likely unusual with respect to soil pH and potassium level (K)? (Circle all that apply.)

(a) Observation 66

(b) Observation 65

(c) Observation 9

(d) Observation 53

Here we are referring to the predictors which the leverage depend on. The values with a high leverage are 65, 9, 53

Answer : B, C, and D

Problem 4

Which of the following data points are likely unusual with respect to their observed biomass? (Circle all that apply.)

(a) Observation 66
(b) Observation 65
(c) Observation 9
(d) Observation 53

Here we are referring to the Response which the Residuals depend on. The values with a high residuals are 53, 65, 66

Answer: A, B, D

5c) What is the average difference in lifespan between smokers and non-smokers? (1 point)

→ $(93.68 - 23.4) - 3.265(\text{Alcohol})$ ← smokers

→ $93.68 - 3.265(\text{Alcohol})$ ← nonsmoker

~~Essentially a difference of 23.4 holding Alcohol constant~~

Average lifespan (smoker) = $(93.68 - 23.4) - 3.265(\text{Alcohol})$

Average lifespan (nonsmoker) = $93.68 - 3.265(\text{Alcohol})$

The avg difference in life span is equation 1 - 2 thus a difference of 23.4. Smokers have on average 23.4 less years.

- 23.4 years

Problem 7 (5 points)

a) State the null and alternative hypotheses for a test of the significance of the categorical predictor variable when using Model 3 as the full model. (2 points)

$H_0: \beta_2 = 0$ categorical predictor $= 0$

$H_a: \beta_2 \neq 0$ categorical " $\neq 0$ (It is significant)

Full model : $MPG = \beta_0 + \beta_1 \text{weight} + \beta_2 \text{transmission\_type M} + \beta_3 \text{weight} \cdot \text{Trans\_Type m} + \varepsilon$

Reduced : $MPG = \beta_0 + \beta_1 \text{weight} + \varepsilon$

$H_0: \beta_2 = \beta_3 = 0$ (categorical predictor $= 0$)

$H_a: \beta_{i=2 \text{ or } 3} \neq 0$ (At least 1 predictor $\neq 0$)

7b) Which of the tests in Problem 6.b or Problem 7.a is most reliable? Briefly explain. (3 points)

In 6b we are testing for the overall model so even if we reject the null it doesn't mean that our categorical predictor is significa

For 7a we are doing an individual t test for the categorical variable and would know right away if the predictor is significant or not

For 6b Full Model → $MPg = \beta_0 + \beta_1 \text{weight} + \beta_2 \text{transmission\_Type M} + \beta_3 \text{weight} \cdot \text{Transmission\_Type M} + \varepsilon$

Reduced Model → $MPg = \beta_0 + \varepsilon$

For 7a Full model → $MPg = \beta_0 + \beta_1 \text{weight} + \beta_2 \text{transmission\_Type M} + \beta_3 \text{weight} \cdot \text{Transmission\_Type M} + \varepsilon$

Reduced model → $MPg = \beta_0 + \beta_1 \text{weight} + \varepsilon$

To answer these we need to examine the conditions. For the reduced model in 7a we observe in the residual vs fitt graph a dip in the residuals. For the normal quantile plot we see a larger deviation from the line around the theoretical quantile **-1.** Although these observation may not be enough to say the assumptions weren't met they allow us to examine the reliability of our ~~test~~. In 6b reduced model $B_0 = \bar{y}$. For this we are using the actual data which in turn reduces variability

As such 6b is more reliable

Problem 8 (6 points)

how exactly do you consider these things? what steps do you need to take and what could be the result?

Suppose someone suggests that we should add another predictor variabl
measures the displacement of the vehicle's engine (in inches). Describe
take to statistically support (or not) this decision without conducting any tests or calculating any confidence intervals. (More points will be awarded for more valid methods with accurate justifications.)

Maybe we could consider the possibility of multicollinearity or an interaction. Generally the more displacement of an engine the more mpg is used. It could, be possible that one of the type of car has a large engine displacement but lower weight and thus we have the effect of displacement on mpg changing over the 3rd variable of weight

— We need to consider the possibility of multicollinearity or an interaction
For Multicollinearity we could use a scatterplot from a matrix to see if there is a linear relationship between engine displacement and MPG
It Seems likely that as engine displacement increases (↑size) the MPG required to run the engine would also increase

We could also compute anova table for the model containing the new predictor. we could examine the MSE for both model or just examine overall variance. Although the MSE wouldn't increase we can judge whether the added variant was significant in explaining any of the model variability