STAT 21 Midterm II Corrections

Stat 21 Directory

Problem 1

Which of the following statements are supported by the R output for this model?

I selected just "(a) Comparing grass from soil..." I continue to think this is correct, since:

- "(b) The effect of changing the pH level of the soil on the biomass depends on the potassium level of the soil" implies that there is an interaction term, but there isn't.
- For both (c) and (d), "The effect of changing the potassium level of the soil has a greater impact on the biomass than changing pH level..." or vice versa, I believe the claims are ill-formed. It's true that the coefficient for a one unit increase in pH is much, much larger than for K. But a one unit increase on a scale from 0 to 14 (pH) is just not comparable with a one unit increase in K parts per million, which is a scale from 0 to 1,000,000. Of course going from some pH x to $x+1$ has a much bigger impact than going from $5,000,000$ to $5,000,001$. But that is because the unit sizes are vastly different.
- Another interpretation could be that a "greater impact" means a greater *positive* impact, in which case we'd say the effect of pH > than K just because the coefficient is positive vs negative. But this interpretation seems wrong; I think the question as written is about magnitude, not sign.
- So I think both statements are false.

Problem 2

... which of the following are valid reduced models?

I selected all options. In fact, only (a) and (b) are valid. The other two contain one indicator variable each. But the categorical variable in question has three levels, so it is not legitimate to only have one indicator variable.

Problem 5

What is the average difference in lifespan between smokers and non-smokers?

I wrote: $(93.6818) - (93.6818 - 23.4392) = 23.4392$ years difference in average lifespan. I'm told I didn't have the correct interpretation.

The interpretation: On average, our sample shows that there is a 23.4392 year difference in lifespan between smokers and non-smokers. Non-smokers lived longer in this dataset. We should note that there is sampling variability due to random chance, so we cannot say for certain that, in the population at large, the average difference is in fact 23.4392 years.

Problem 7

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.

I wrote that our alternative hypothesis is: $\mathrm{H_A}: \beta_2 < 0$. In fact, the significance codes returned by R in the `lm()` function are for two-sided *t*-tests by default. So the correct alternative should be: $\mathrm{H_A}: \beta_2 \neq 0$.

b) Which of the tests in Problem 6.b or Problem 7.a is most reliable?

I used a definition of "reliable" that was more akin to "useful."

Knowing now that "reliable" means "appropriate to use given the degree to which assumptions are met," I would select the *t*-test for slope as "more reliable." That's because it is trying to do a narrower thing, as opposed to the ANOVA overall *F*-test test. The *F*-test assumes that the population (here, of cars) is normally distributed. On the other hand, the *t*-test for slope is about an estimator. Consequently, we're able to make fewer assumptions about the data and just see if the specific slope we've found is statistically significant, given the standard error of the estimator.