Problem 1.

Correct answers: a. d

per unit change in pH, holding other predictors constant, on average increases biomass by 414.9 gm^{-2}. However, per unit change in K level, holding other predictors constant, decreases biomass by 1.01 gm^{-2}. So a is correct.

Problem 7. b.

Both tests are reliable.

The ANOVA test in problem 6. b includes Model 3.

The reduced model in problem 7. a is Model 1, and the full model is model 3.

For both model 1 & 3, since the points in the scatterplots do not form obvious curvature, the linearity condition is met. The points also distribute randomly above and below the zero-residual line, so no issue with constant variance.

In the normal quantile plots, the points align well with the predicted line, so there's no issue with normality.

For both models we assume zero mean, randomness and independence conditions are met.

Since all conditions necessary for estimation and inference are met for the models included in the 2 tests, both tests are reliable.

Problem 8.

① Added variable plots

1. Get residuals for model 3. Record as Res1.
2. Build a model for displacement:

   displacement = $\beta_0' + \beta_1'$ weight + β_2' transmission_typeM +

   β_3' weight × transmission_typeM.

   then get the residuals for this model, record as Res2.
3. Plot added variable plot for Res1 against Res2. Also construct a model for Res1 against Res2.
4. Check the R^2 for the new model in step 3. This R^2 is the additional variability that can be explained by displacement that is not already explained by other 3 predictors in the model. If the R^2 is large (eg. > 0.5), then we might need to include displacement in the model.

② Mallow's Cp

1. Build a model for mpg using all predictors in model 3 and displacement. This is the full model.
2. Then let model 3 be the reduced model.
3. Calculate Mallow's Cp using these 2 models.

Cp gives the amount of information might be available in the predictor "displacement" that are not in model 3. So if Cp is large, then we might need to include displacement as a predictor.

③. Check R^2_{adj} for best subset

1. Get R^2_{adj} for Model 3.
2. Get R^2_{adj} for the model with all predictors in model 3 and displacement.
3. Check the difference between the R^2_{adj} in step 1 and 2. If R^2_{adj} in step 1 is significantly smaller than that for the R^2_{adj} in step 2, then displacement, with other predictors in model 3 present, can explain a significant amount of variability of mpg, so we might need to include "displacement" into the model.

Test 2 - part 1

STAT 021

Swarthmore College

Do not flip this page until instructed to do so.

Test organization: There are 8 questions in total on this test and they are organized into two subsections: the first 4 questions are select all that apply multiple choice questions and the last 4 questions are free response. If you need additional paper you may come to the front of the class and pick some up. There are a total of 30 points possible on this test.

Instructions: The first part of this test are multiple choice questions that do not require any additional explanation or work. No extra work will be considered in the grading of these questions but *you can get partial credit* for many of these questions. The last part of this test involves short answer questions. For these questions, you must show all your work and/or provide enough justification and explain your reasoning in order to get full credit or be considered for partial credit. You do not need a calculator to evaluate any expressions. For any calculation problems, simply writing out the formula to find the answer will suffice.

First and Last Name: Xinxin Li

Swarthmore Username: xli5

Take a deep breath.

You have prepared for this test and with a clear and well-rested mind, you are ready to show me what you have learned this semester. As with the other tests, the purpose of this test is to measure your understanding of the material we have covered. This is nothing more than a metric for me to evaluate your preparedness to think statistically at this particular moment in time and in this particular setting. This is not a perfect measure of your knowledge and does not predict your future statistical skills.

x/75

Multiple choice problems (2 points each)

Use the following information to answer questions 1-4.

Suppose we are trying to understand how the aerial biomass (response variable) production of a certain type of marsh grass is related to the three predictor variables

- pH, which measures the pH level of the soil (from 0-14),
- K, which measures the potassium level of the soil (in ppm), and
- the categorical variable *location* which can be one of three different spots ("OI" is short for Oak Island, "SI" is short for Smith Island, and "SM" is short for Snows Marsh).

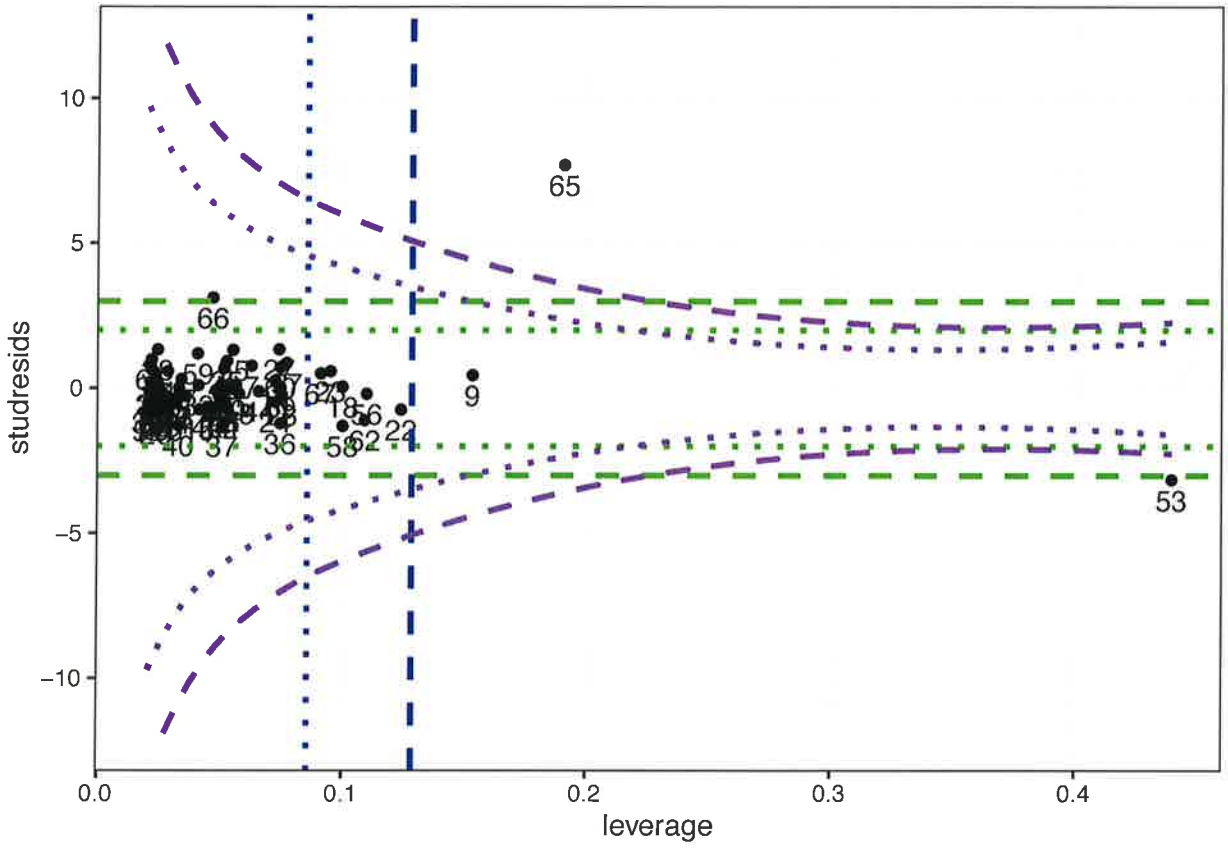
Aerial biomass is called `BIO` and is measured with units gm^{-2}.

Consider the main effects model shown below. The summary output is shown on the next page. Below the model summary is a plot of the studentized residuals and leverage values for each data point based on this model. The heavy dashed lines represent "extreme" cutoffs and the short dashed lines represent "moderate" cutoffs for leverage (blue), studentized residuals (green), and Cook's distance (purple), respectively.

$$\widehat{biomass} = \hat{\beta}_0 + \hat{\beta}_1 pH + \hat{\beta}_2 K + \hat{\beta}_3 LocationSI + \hat{\beta}_4 LocationSM + \epsilon, \text{ where}$$

$$LocationSI = \begin{cases} 1, & \text{if at Smith Island} \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad LocationSM = \begin{cases} 1, & \text{if at Snows Marsh} \\ 0, & \text{otherwise} \end{cases}$$

```
##
## Call:
## lm(formula = BIO ~ pH + K + Location, data = biomass)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -811.15 -190.99  -37.70   96.78 1056.67
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   43.0122   299.1682   0.144   0.8864
## pH           414.9021    43.3381   9.574 6.68e-12 ***
## K             -1.0095     0.2324  -4.344 9.32e-05 ***
## LocationSI  -497.6490   163.4252  -3.045   0.0041 **
## LocationSM    58.1814   131.6870   0.442   0.6610
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 354.2 on 40 degrees of freedom
## Multiple R-squared:  0.7382, Adjusted R-squared:  0.712
## F-statistic: 28.19 on 4 and 40 DF,  p-value: 3.613e-11
```



Problem 1

Which of the following statements are supported by the R output for this model? (Circle all that apply)

(a) The effect of changing pH level of the soil has a greater impact on the biomass than changing the potassium level, given the location of the grass is the same.

(b) The effect of changing the potassium level of the soil has a greater impact on the biomass than changing the pH level, given the location of the grass is the same.

(c) The effect of changing the pH level of the soil on the biomass depends on the potassium level of the soil.

(d) Comparing grass from soil with matching pH levels and matching potassium levels, grass from the Snows Marsh tends to have higher biomass than grass from Oak Island.

Problem 2

Which of the following represents a valid reduced model if we want to determine whether or not to include the categorical predictor for location? (Circle all that apply)

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

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

x1/5

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

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

Problem 3

Which of the following data points are likely most 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

Problem 4

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

(a) Observation 66

(b) Observation 65

(c) Observation 9

(d) Observation 53

Short answer questions

Problem 5 (6 points)

The R-output below contains the results from a regression model fit to a data set concerning life expectancy in years, alcohol consumption in drinks per day, and smoking status (smoker - Yes vs. non-smoker - No). Answer the following questions pertaining to the R output below.

```
## Call:
## lm(formula = Lifespan ~ Alcohol + Smoker)

## Residuals:
##     Min      1Q  Median      3Q     Max
## -30.796  -7.139   0.125   6.949  19.578

## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 93.6818   2.1024     44.56   <2e-16 ***
## Alcohol     -3.2656   0.3147    -10.38   <2e-16 ***
```

```
## SmokerYes  -23.4392  1.9922    -11.77   <2e-16 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Residual standard error: 9.959 on 97 degrees of freedom
## Multiple R-squared: 0.7129, Adjusted R-squared: 0.7069
## F-statistic: 120.4 on 2 and 97 DF, p-value: < 2.2e-16
```

a) How does drinking alcohol affect average life expectancy? (Explain in a full sentence.) (2 points)

On average, 1 drink of increase in alcohol consumption per day would reduce the life expectancy by 3.2656 years given that the smoker status is held constant.

b) Suppose someone consumes about 2.5 alcoholic drinks per day and smokes regularly. What is the average life expectancy for this individual? (1 point)

Lifespan = 93.6818 − 3.2656×2.5 − 23.4392

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

On average, smokers' lifespan is shorter than non-smokers for 23.4392 years. Lifespan for smokers minus Life span for non-smoker is −23.4392 on average.

d) Suppose we consider only individuals who consume the same amount of alcohol; is there a statistically significant relationship between life expectancy and smoking status? Justify your answer. (2 points)

t-test for slopes:

H_0: $\beta_2 = 0$

H_a: $\beta_2 \neq 0$

β_2 is the slope for predictor "SmokerYes".

t-value = −11.77

p-value = < 2e-16 ⇒ this is really small

So p-value < alpha = 0.05. We reject H_0 that $\beta_2 = 0$, and there is statistically significant evidence that the slope for "SmokerYes" is not zero, and there is a statistically significant relationship between life span and smoking status given alcohol is in the MLR.

6.b. (continued) predicting vehicle fuel consumption (in mile per gallon) from weight, transmission type, and their interaction is somehow effective since at least one of weight, transmission type, and their interaction term is statistically significant for predicting mpg.

x1.5

For Problem 6-8 we are going to consider three different MLR models for vehicle fuel consumption (in miles per gallon) as predicted by the vehicles weight (in lbs) and possibly also by the transmission type where $transmission\_typeM = \begin{cases} 1, & \text{if manual vehicle} \\ 0, & \text{otherwise} \end{cases}$.

You can assume the vehicles were selected as a simple random sample.

Model 1: $mpg = \beta_0 + \beta_1 weight + \epsilon$

Model 2: $mpg = \beta_0 + \beta_1 weight + \beta_2 transmission\_typeM + \epsilon$

Model 3: $mpg = \beta_0 + \beta_1 weight + \beta_2 transmission\_typeM + \beta_3 weight \cdot transmission\_typeM + \epsilon$

The summary for each model and the corresponding residual plots and Normal quantile plots are shown on last three pages after the statement of each problem.

Problem 6 (5 points)

a) Which of the three models would you choose to predict vehicle mileage? Justify your answer.

I would choose model 3.

For Model 2: p-value for t-test for "transmission_type M" = 0.398 > 0.05 (alpha), so it has a non-significant predictor.

For Model 3 and Model 1: p-values for t-tests for all predictors are less than alpha (0.05), So all predictors are significant based on t-test. No issue with Residual plots from linearity, constant variance, and normality. However, R^2_{adj} for model 3 is 0.8853, and R^2 for model 1 is 0.8716. So Model 3 does best in explaining the variability of fuel consumption (mpg).

b) State the null and alternative hypotheses for a test of a linear association among the predictors and response based on your answer to part (a). For an $\alpha = 0.05$ significance level, report the p-value and interpret the conclusion of this test in the context of the problem. (3 points)

ANOVA F-test.

$H_0: \beta_1 = \beta_2 = \beta_3 = 0$

H_A: at least one of the β_i is not zero.

Where β_1, β_2, β_3 are the slopes for weight, transmission_type_M, and the interaction term between them respectively.

F-statistic = 73.02, and p-value = 1.669e-12 → really small.

So p-value < alpha = 0.05, and we reject H_0 that all predictors have slope = 0. There is at least 1 predictor that has slope ≠ 0 and has correlation with mpg. This model (see top)

8. (continued) there are a large amount of variability for displacement that can be explained by displacement while other predictor terms are present, and we may need to consider displacement as a predictor variable. x/15

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)

Nested F-test.

$H_0: \beta_2 = \beta_3 = 0$

H_a: at least one of β_i above $\neq 0$.

Where β_2 is the slope for the predictor "transmission_typeM".

β_3 is the slope for the interaction term transmission_typeM x weight.

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

7.a is more reliable.

ANOVA F-test from 6.b can only say that at least one predictor terms is statistically significant for this model of predicting mpg, but we don't know which specific predictor is significant.

Nested F-test, however, clearly states the predictor terms that we would like to test on for significance in the model. So this is more reliable.

Problem 8 (6 points)

Suppose someone suggests that we should add another predictor variable displacement, which measures the displacement of the vehicle's engine (in inches). Describe what steps you could take to statistically support (or not) this decision without conducting any tests or calculating any confidence intervals.

① Get the residuals (res1) for Model 3:

$$\hat{mpg} = \hat{\beta}_0 + \hat{\beta}_1 \text{weight} + \hat{\beta}_2 \text{transmission\_typeM} + \hat{\beta}_3 \text{weight} \times \text{transmission\_typeM}$$

② Build a model for

$$\text{displacement} = \beta_0' + \beta_1' \text{weight} + \beta_2' \text{transmission\_typeM} + \beta_3' \text{weight} \times \text{transmission\_typeM}.$$

(this is the displacement in inches)

and get the residuals. (res2)

③ Plot added variable plot for the residuals in step ① against the residuals in step ②. Also construct a model for res1 based on res2.

④ Check the R^2 for the new model in ③. This R^2 is the additional variability that can be explained by displacement that is not already explained by other 3 predictor terms in the model. If this R^2 is relatively large (eg. > 0.5) then (see top)

Model 1: $mpg = \beta_0 + \beta_1 weight + \epsilon$

```
##
## Call:
## lm(formula = mpg ~ weight, data = car_dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1418 -1.1597  0.4131  1.1759  4.1569
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 39.9659944  1.5594404   25.63  < 2e-16 ***
## weight      -0.0067480  0.0004985  -13.54  1.5e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.189 on 27 degrees of freedom
## Multiple R-squared:  0.8716, Adjusted R-squared:  0.8668
## F-statistic: 183.2 on 1 and 27 DF,  p-value: 1.502e-13
```