mpopov1

Test 3

STAT 021

Swarthmore College

Do not flip this page until instructed to do so.

Test organization: There are 12 questions in total on this test and they are organized into three subsections: the first 4 questions are matching or True/False with explanation questions, the next 5 questions are free response short answer and should not require more than a sentence or two to answer. The last section contains 3 long answer free response questions that require more than a couple of sentences to answer fully. There are a total of 60 points possible on this test. The last section explains an extra credit opportunity. If you need additional scratch paper you may come to the front of the class and pick some up.

Instructions: Answer each question to the best of your ability and raise your hand if you are confused by any of the wording in the questions or suspect a typo. For the short and long answer questions show all your work and provide enough justification and/or explanation in order to get full credit or to 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: Maxim Popovlov

Swarthmore Username: mpopov1

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. The purpose of this test is to measure your understanding of the material we have covered this semester. 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.

mpogorel

effect constant + additive, variance of groups the same, independence + randomization, normality

independence, randomization, zero mean, linearity, constant variance, normality

(b) Suppo
0.5, a
a sim
term

Section 1: Matching and True/False problems

1. (5 points)

Suppose we are modeling the weight of birds (in kg) as a linear function of a categorical predictor variable for bird type (with levels pigeon, sparrow, and finch) and a numeric predictor for bird age. Given a "full" model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_3 + \beta_5 x_2 x_3 + \epsilon,$$

where $x_1 = \begin{cases} 1, \text{ if sparrow} \\ 0, \text{ otherwise} \end{cases}$, $x_2 = \begin{cases} 1, \text{ if finch} \\ 0, \text{ otherwise} \end{cases}$ and x_3 is the age of the bird (in months), match the questions below to their corresponding null hypotheses.

a) For newly hatched birds (of age zero months), is there a statistically discernible difference in the weights of these three different bird types?

b) Does the effect of age on a bird's weight depend on what type of bird it is?

c) Given we are only comparing birds of the same age, is there a statistically significant difference in the mean weight of sparrows and pigeons?

d) Given we are only comparing pigeons, is the effect of age on a bird's weight statistically significant?

e) Is there statistically discernible evidence of a linear relationship between bird age and type and bird weight?

1. c $\quad H_0 : \beta_1 = 0$
2. d $\quad H_0 : \beta_3 = 0$
3. e $\quad H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$
4. b $\quad H_0 : \beta_4 = \beta_5 = 0$
5. a $\quad H_0 : \beta_1 = \beta_2 = 0$

2. (5 points)

Determine which of the following statements are true and false. For each statement that is false, provide a brief explanation as to why it is false.

(a) If predictors are collinear, then removing one variable will have no influence on the point estimate of another variable's coefficient.

True

(b) Suppose one numeric predictor variable x_1 has an estimated MLR coefficient of $\hat{\beta}_1 = 0.5$, another numeric predictor x_2 has an estimated coefficient $\hat{\beta}_2 = 5$. If we consider a simultaneous one-unit change in both x_1 and x_2 (while holding any other predictor terms constant), this corresponds to an average change in $\hat{y}$ of at least 5 units.

$\hat{y} = .5x_1 + 5x_2$ 3->4 $\hat{y} = .5(3) + 5(3) = 16.5$ $22 - 16.5 = 5.5$ 1->2 $.5(1) + 5(1) = 5.5$

2->3 $.5(2) + 5(3) = 11$, $.5(3) + 5(3) = 16.5$, $16.5 - 11 = 5.5$ $\hat{y} = .5(4) + 5(4) = 22$ $.5(2) + 5(2) = 11$, $11 - 5.5 = 5.5$

True

(c) As the total sample size increases, the degrees of freedom for the residuals increases as well.

True

3. (5 points)

Determine which of the following continuations of the statement below are true and which are false. For each statement that is false, provide a brief explanation as to why it is false.

If the null hypothesis that the means of four groups are all the same is rejected from an ANOVA model and overall F-test at a 5% significance level, then...

(a) We can then conclude that all the means are different from one another.

False

We can't conclude all the means are different, we can only conclude that at least one of the means is different.

(b) The standardized variability among the group averages is higher than the estimate of the variability of the data within each group.

True

(c) A post-hoc pairwise analysis will identify that there is at least one pair of means that are significantly different.

True

4. (5 points)

Determine if the following statements are true or false. For each statement that is false, provide a brief explanation as to why it is false.

Mayavel

(a) If a given value (for example, the null hypothesized value of a parameter) is within a 95% confidence interval, it will also be within a 99% confidence interval.

True

(b) Decreasing the significance level (α) will increase the probability of making a Type 1 Error. Ho rejected when Ho true

False

Decreasing α makes it harder to reject the null which makes it harder to make a Type I error.

(c) Suppose the null hypothesis is $\beta_1 = 0$ and we fail to reject H_0. Under this scenario, the true slope of x_1 is zero.

False

We can't say for certain that the true slope is zero. Failing to reject the null just means that there isn't significant evidence that β_1 is not 0. It doesn't say that it IS zero.

Section 2: Short answer questions

5. (4 points)

When computing a confidence interval for the slope, β_j, of a numeric predictor variable, x_j, in a multiple linear regression model, provide a reason why someone might prefer to use a bootstrap procedure rather than rely on the formula $\hat{\beta}_j \pm t^*_{(n-k-1),\alpha/2} \times SE(\hat{\beta}_j)$?

A bootstrap procedure gives you more data. Maybe the original sample has problems and has a lot of data that would be considered outliers in other samples. Bringing in bootstrap samples gives you more data which creates a higher likelihood of creating a more accurate confidence interval. This is better than just relying on one sample's data.

6. (3 points)

If you could only use one measure (among the studentized residuals, leverage values, and Cook's distance values) to identify potentially influential data points, which would you choose and why?

I would choose leverage since that's what leverage is specifically used for. If a point has a very large leverage point ($>3(k+1)/n$), then that means it has large leverage over the data set. That point is very influential towards the data set and the values that come out of it.

is within a

For questions 7-9 consider the following random sample of $n = 250$ Minor League baseball hitters who were drafted for the Major League between the years 1992 and 2006. We are going to consider modeling the average number of times each hitter appeared at the plate per game (`ave_plate_counts`) as the response variable and player position as a categorical predictor (with levels: Catcher, First Base, Outfield, Second Base, Shortstop, Third Base). Below is the R summary output for this one-way ANOVA model.

```
##
## Call:
## lm(formula = ave_plate_counts ~ position, data = baseball_dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.97080 -0.16440  0.06234  0.30217  0.87561
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            3.48288    0.07475  46.592  < 2e-16 ***
## positionFirst Base     0.42739    0.12242   3.491  0.00057 ***
## positionOutfield       0.19625    0.08726   2.249  0.02541 *
## positionSecond Base    0.28036    0.12074   2.322  0.02105 *
## positionShortstop      0.30913    0.10311   2.998  0.00300 **
## positionThird Base     0.38201    0.11772   3.245  0.00134 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4547 on 244 degrees of freedom
## Multiple R-squared:  0.07043,    Adjusted R-squared:  0.05138
## F-statistic: 3.697 on 5 and 244 DF,  p-value: 0.003039
```

mpgover1

Section 3:

10. (9 points) Suppose you

7. (3 points)

(a) What are the error degrees of freedom based on this model?

(b) What is the reference level?

a- The error degrees of freedom is 244 based on this model

b- The reference level is Catchers since they're not included in the model. Their value is the intercept which is 3.48188 plate appearances per game.

8. (6 points)

Suppose the average number of plate appearances per game is 3.72 over all 250 data points. What is the estimated group effect for Shortstop players?

Shortstop: 3.48188
+0.39913
3.79201

Effect: 3.79201 - 3.72 = 0.07201

The estimated group effect for Shortstop players is positive .07201 plate appearances per game towards the model.

9. (4 points)

Consider two additional numeric predictors: the age each player was drafted (`age_drafted`) and their overall pick order when they were drafted (`overall_pick`). If we were to fit a regression model including each of the three predictor variables and an interaction between the two numeric variables, explain the meaning of the coefficient for the interaction term within the context of this data. (You should be able to answer this in no more than two sentences.)

The coefficient of this interaction term shows the estimated amount of plate appearances per game a player with a certain draft pick multiplied by their draft age would have over or under a catcher.

* Sorry for the weird wording

mpgsorel

Section 3: Long answer questions

10. (9 points)

Suppose you have access to a data set on a random sample of Swarthmore students. The variables included in this data set are a numeric variable for the average amount of time (in hr) spent studying each week school is in session, a binary categorical variable distinguishing students who are double majors from those who aren't double majors, a numeric variable for each student's current, cumulative GPA, and a categorical variable indicating if a student has ever worked more than two student jobs in a semester or if the most number of student jobs they've worked in a semester is one or if they have not worked any student jobs for any semesters.

State a research question that can be answered with the overall F-test for each of the following models. Also provide a mathematical representation of the model and state the null hypothesis based on the notation you define for each model.

(a) a simple linear regression model;

(b) an ANOVA model;

(c) a multiple linear regression model (not SLR or ANOVA).

a- Is there a statistically significant effect of hours spent studying each week school is in on a student's cumulative GPA?

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \hat{x}$ $\hat{y}$ = cumulative GPA $\hat{\beta}_1$ = average amount of hours spent studying $\hat{\beta}_0$ = intercept

$H_0: \beta_1 = 0$

b- Is there a statistically significant effect of number of jobs worked in a semester on a student's cumulative GPA?

$\hat{y} = \hat{\mu}_0 + \hat{\mu}_1 X_1 + \hat{\mu}_2 X_2$ $\hat{y}$ = cumulative GPA $\hat{\mu}_0$ = reference level: >=2 jobs worked $\hat{\mu}_1$ = 1 job worked $\hat{\mu}_2$ = no jobs worked

$H_0: \mu_1, \mu_2 = 0$

c- Is there a statistically significant effect of at least one of hours spent studying each week school is in and whether or not a student is a double major on a student's cumulative GPA?

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$ $\hat{y}$ = cumulative GPA $\hat{\beta}_1$ = average amount of hours spent studying $\hat{\beta}_2$ = whether or not student is double major β_0 = intercept

$H_0: \beta_1, \beta_2 = 0$

mpgcrel

11. (8 points)

Consider the ANOVA model for the baseball data you used in questions 7-8. Reference the R output on pg 5 and the plots on pg 10 to answer the following questions about this model.

(a) Check the conditions necessary for conducting a test to determine if the average plate appearance (per game) is significantly different for players of these six different positions.

(b) Write out in words and in symbols the hypotheses that would be tested in part (a). (Clearly define your notation.)

(c) What can you conclude about the test in part (b)? Write a paragraph discussing your conclusions and reference any relevant statistics and/or plots as part of your discussion.

a- Effect is constant and additive: The effects of each level seem to be constant and they're all positive so they're additive.

Variance of groups are the same: There's a little bit of fluctuation with the variances but the variances seem fairly similar.

Independence: There's no real way to determine if the residuals are independent with the information given.

Randomization: Stated that the sample of 250 was randomly selected.

Normality: This condition is not met as the normal plot clearly has a curved relationship with the data points

b- H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = 0$
H_A: at least one $\mu_i \neq 0$

H_0: The mean effect of first base, outfield, second base, short stop, and third base on average plate appearances per game is 0 for all of them.
H_A: At least one of the positions mentioned has an effect on average plate appearances per game greater than 0.

c- You can't really conclude much due to the normality condition not being upheld. The normal plot has a clear curve, which means that the current model isn't very viable. If the normality condition was okay though, then I would conclude that we should reject H_0, which means that there is statistically significant evidence that at least one of the levels of the predictor variable has an effect on average plate appearances per game greater than 0. This is because the p-value is .003039 which is less than $\alpha = .05$.

12. (8 points) Suppose two people in local wells. This samples from across

eference the
this model.

mayorel

12. (8 points)

Suppose two people are studying the historic data set about the amount of arsenic (Arsenic) in local wells. This data contains $n = 70$ observations from a random selection of well water samples from across the state. In addition to the levels of arsenic, the data also records the year the data was collected (Year) and the distance from the well to the nearest mining site (Miles).

Person A fits the following MLR model to the data:

$$Arsenic = \beta_0 + \beta_1 Year + \beta_2 Miles + \epsilon$$

and computes an adjusted R^2 value of 0.26. ← low

Person B considers the following correlations:

$$Cor(Arsenic, Year) = \rho_1; \quad Cor(Arsenic, Miles) = \rho_2$$

and estimates each with their sample correlations $r_1 = 0.77$ and $r_2 = -0.34$. Are the two people's conclusions contradictory? Explain your answer.

Not necessarily. Person A finds that their model has a lot of variation that isn't explained by the model. Person B finds that there's some variation not explained by the first predictor (Year) and there's a good amount of variation not explained by the second predictor (Miles). So, the fact that there isn't a lot of variation explained in Person A's model is supported by Person B's correlations. Also, the R^2 value of person A's model is low and positive. The R^2 values of Person B's correlations are .77 for Year and -.34 for Miles. Considering that the positive correlation is stronger than the negative correlation, it makes sense that the full model has a low but positive correlation.

Section 4: Extra credit opportunity

If the response rate to my end of the semester evaluation form (on Moodle under Week 13 and 14) is at least 85% of our class size (over both sections), two percentage points will be added to everyone's Test 3 grade (up to 100 total possible points). **Hint:** You may not know how to or want to contact everyone in my class but you do know your group mates pretty well.