Exam 1

Stat 512 SP 2020

Honor Pledge: I have not given, received, or used any unauthorized assistance on this exam.

Signature:

Printed Name: Kathleen E Wendt

Instructions:

- **Open book, open notes, calculator required.**
- **Time limit is 1 hour, 50 minutes - strictly enforced!**
- If an answer is in the computer output, use it; don't calculate it by hand.
- Show your work where appropriate. Put your final answer in the box (if provided).
- Make explanations brief and legible.
- All questions are worth 4 points except where noted. Maximum score is 100.
- Computer input/output is provided at the end of the exam.
- The exam contains a total of 7 pages (including blank page 7).
- There is an additional **9 pages of R output**.

Questions 1 through 5: 2 pts per problem. $Y \sim X_1 + . . + X_{10}$

For this group of questions, suppose that we have a response variable Y and ten predictor variables (X1 through X10). The investigator is interested in model selection with main effects only (no interaction or polynomial terms). Circle one answer; no need to justify your response.

1. Variables X1 and X3 are highly correlated. This indicates there may be a high value of what? (circle all that apply) collinearity

Cook's Distance R^2 [✓ VIF]

2. The pairwise correlation matrix (from cor()) can be used to determine which variable would be added first using forward selection.

[✓ TRUE] FALSE

3. For this multiple regression, which diagnostic plot is **most useful** for assessing the assumption of equal variance?

[✓ Residuals vs Fitted] QQplot of Residuals Histogram of Residuals Std Residuals vs Leverage

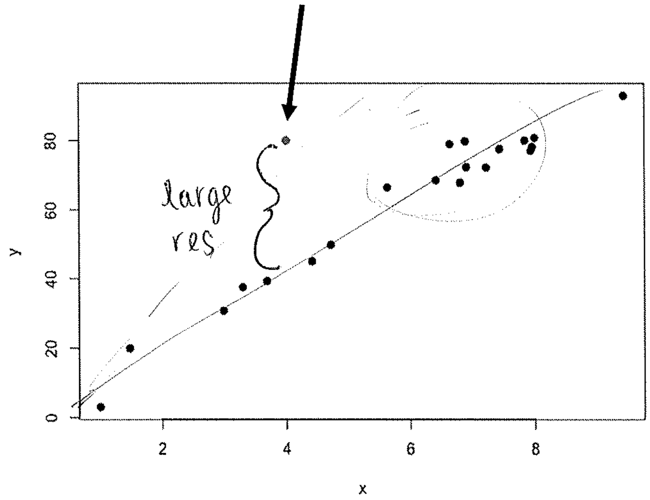
4. The presence of correlation among the predictor variables indicates that an interaction should be considered.

TRUE [✓ FALSE]

5. The point indicated on the below plot has high influence.

TRUE [✓ FALSE]

res + lev



Questions 6 through 16: Fitness

Researchers were interested in developing an equation to predict fitness based on the exercise tests rather than on expensive and cumbersome oxygen consumption measurements. ~~The response~~ variable is Oxygen. A total of 6 potential predictor variables are described below. A total of n=31 subjects participated in the study. The analysis is included at the end of the exam as "**Fitness**". Use α=0.05.

Oxygen = Oxygen intake rate (ml per kg body weight per minute)
Age = Age (years)
Weight = Weight (kg)
RestPulse = Heart rate while resting
RunTime = Time to run 1.5 miles (min)
RunPulse = Heart rate while running (same time oxygen rate was measured)
~~MaxPulse = Maximum heart rate while running~~

6. Prior to starting model selection, the investigators decided to drop MaxPulse from consideration. Looking at the variable descriptions and the output from cor(), discuss why this was a reasonable choice.

Yes, this is reasonable because RunPulse and MaxPulse are highly correlated (r=0.93), indicating possible collinearity.

7. Briefly explain how Model 2 was chosen. Hint: Consider the output for both Models 1 and 2.

Model 1 includes 5 predictors; using all/best subset selection AIC with AIC, the "best" model (for Model 2) is identified as R: Oxygen ~ Age + RunPulse + RunTime. ~~[illegible]~~

8. Using Model 2, interpret the partial regression coefficient for Age. Be specific!

Holding all other variables (RunPulse + RunTime) constant, ~~there~~ for every one-unit increase in Age, there is an estimated 0.25-unit decrease in oxygen intake rate.

9. Consider Model2. What command would you use to get R to provide the 95% confidence interval for the partial regression coefficient for Age.

R: `confint()`

10. For Model 2, interpret the R^2 value.

~~[illegible]~~

81.11% of the variability in oxygen intake rate is explained by the multiple regression (on age, RunTime + RunPulse).

11. In the summary() output for Model2, an F-statistic = 38.64 and p-value < 0.001 are shown. What is being tested here? State the null hypothesis (H_0).

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

12. Using Model 2, predict the oxygen for a subject with Age = 45, RunTime = 12 and RunPulse = 160. Give your answer to 1 decimal place.

$$\underset{\text{Oxygen}}{\hat{Y}} = \overset{\beta_0}{111.71806} + \overset{\beta_1}{-0.2564}(\overset{x_1}{45}) + \overset{\beta_2}{-2.82538}(\overset{x_2}{12}) + \overset{\beta_3}{-0.13091}(\overset{x_3}{160})$$

$\boxed{45.3}$

13. Using Model 2, do the regression assumptions appear to be satisfied? Briefly discuss the information in each of these plots. Your discussion should be specific to this analysis!

$\boxed{\text{Yes}}$ -ish

A. Residuals vs predicted values:

Linearity – There is some curvature, indicating the possibility of violation ~~[illegible]~~ linearity.

Equal variance – Reasonable equal scatter of residuals.

These concerns may be amplified because of the small sample size.

B. QQplot of residuals (Residuals vs Quantiles): overall – OK but check for problematic observations (maybe 15, 10, 17)

Normality – Evidence of heavy tails + possible issue w/ normality assumption; but, again, ~~[illegible]~~ this issue may be amplified by problematic observations + small N.

14. Using Model 2, based on the Cook's distance criteria are you concerned that any observations have high influence? Discuss. Note: Use the rule of thumb from class.

Any observations have high influence? Yes [✓ No]

Discuss: Obs 10 has highest Cook's Distance (~0.3), but this value does not exceed 1.

15. Considering the results for Model 2, a colleague suggests that since the R^2 value would be higher for the full model (all predictors) that the full model will be better for prediction. Do you agree that the full model will be better (than Model 2) for making predictions for new observations? Discuss.

Do you agree? Yes [✓ No]

Discuss:

R2 increases when new predictors are added to a model, regardless of their predictive value. More is not always better.

16. Suppose the investigators had wanted to include sex (M or F) as a predictor in the model. Explain how the design matrix (or model.matrix) would have been modified if this variable had been included.

The new design matrix would have 31 rows (n) and 5 columns (K+1) – instead of 31 rows and 4 columns.

lsmeans() is the same as emmeans()

4 levels

adg ~ trt + iwt

Questions 17 through 26: Average Daily Gain pg 4 R code

An experiment was conducted over a 160 day period to evaluate the effects of a feed additive (TRT) on the growth of cattle. Thirty-two cows (n = 32) were randomly assigned to one of four feed additive treatment levels (TRT = 0, 10, 20 or 30). **NOTE: TRT is a categorical predictor in all models considered here!** The response variable is average daily gain (ADG) over the treatment period. Initial weight (IWT) of each animal was also used as a covariate in some analyses. The analysis is included at the end of the exam as "**Average Daily Gain**". Use α=0.05.

There are 4 models shown in the output:

Model 1A: ANCOVA WITH Interaction
Model 1B: ANCOVA WITH Interaction (alternate parameterization)
Model 2: ANCOVA NO Interaction
Model 3: One-way ANOVA

17. Briefly describe the difference between the ANCOVA models WITH and WITHOUT interaction. Hint: Think in terms of slopes and intercepts.

ANCOVA with interaction allows for different intercepts + slopes for each group (trt).

without? -2

Questions 18 through 21 refer to the ANCOVA WITH Interaction (Models 1A and 1B).

18. Using Model 1A, in the table "Anova Table (Type III tests)" look at the line labeled "IWT" with F=0.0024 and p-value= 0.9617. What is being tested here? State the null hypothesis (H_0) using words. Hint: Think in terms of slopes and intercepts.

4 AM

H_0: Intercepts of Initial weight do not differ.

* → but #22

19. Test the null hypothesis that the slope for TRT 30 is equal to zero.

-4 R: summary (Model1A)

Test Statistic:	0.557
P-value:	0.5829

20. Test for a difference between the slopes for TRT 10 vs TRT 20. Give the test statistic and p-value. Hint: Notice the lht() statements used with Model 1B

-2

R: car::lht (Model1B, C1, rhs = c(0))

Test Statistic:	1.3075
P-value:	0.2641

-12

21. Calculate the emmean for TRT=30 with IWT=390. In other words, calculate the predicted value. Give your answer to 1 decimal place.

By hand

$$\underset{\beta_0}{} \quad \underset{\beta_1\ IWT}{} \quad \underset{\beta_2\ TRT30}{}$$

$$\hat{Y} = \overset{\beta_0}{1.3818} + \overset{\beta_1\ IWT}{-0.0001(390)} + \overset{\beta_2\ TRT30}{-0.2856} + 0.002$$

$$\boxed{1.85}$$

Questions 22 and 23 refer to the ANCOVA NO Interaction (Model 2).

22. Using Model 2, in the table "Anova Table (Type III tests)" look at the line labeled "TRT" with F= 4.04 and p-value= 0.0170. What is being tested here? State the null hypothesis (H_0) using words.
Hint: Think in terms of slopes and intercepts.

w/o int

H_0: Intercepts of treatment do not differ.

23. Adjusting for IWT, which TRTS have mean ADG values that are significantly different from each other? Make your conclusions based on Tukey adjusted p-values with α=0.05.

trts 0 -vs.- 30 , $p = 0.0229 < \alpha = 0.05$

24. Complete the following table. **(6 pts)** Note that there are n=32 observations from 4 TRTS.

Model	**p**	**SSResid**	**AIC**
1 (ANCOVA w Interaction) ~~Model 1A~~	8	2.95228	-60.26
2 (ANCOVA NO Interaction)	5	3.424569	-61.51
3 (ANOVA)	4	4.405425	−55.456

$$= 32 \cdot \ln\left(\frac{4.405423}{32}\right) + 2 \cdot 4$$

$$32 \cdot -1.983 + 8 = -55.456$$

25. Using a backward elimination approach, which model would be selected? Circle one answer. **(2 pts)**

Model 1 [✓ Model 2] Model 3

ANCOVA w/o interaction ((b/c ~~int not of interest~~))

26. Considering the table from #24, a colleague says that your AIC selection process is flawed because you did not consider the simple linear regression model (including just IWT). Give the research goals stated in the problem description, does the simple linear regression model need to be considered? Justify your response.

Regression needs to be considered? Yes [✓ No]

Discuss: The goal of the study is to test the effects of a feed additive on cattle growth. Initial cattle weight is not of primary interest, but it is an important covariate, to determine the unique contributions the feed additive could make to cattle growth (i.e., not just that big cattle got bigger).

Thank you!

-12 88