**Exam 2**

**Stat 512 SP 2020**

**Honor Pledge:** I have not given, received, or used any unauthorized assistance on this exam. By submitting this exam to be graded you agree to this statement.

Kathleen E. Wendt (answers printed in purple Arial font size 12)

**Instructions:**

* **Open book, open notes, calculator required.**
* If an answer is in the computer output, use it; don’t calculate it by hand.
* Show your work where appropriate.
* Make explanations brief and legible.
* 24 questions (4 pts each) + score for legibility (0 – 4pts). Maximum score is 100.
* Computer input/output is provided in the posted handout.
* The exam contains a total of 7 pages.
* There is an additional **7 pages of R output**.

1. How clearly are your answers indicated? Graded on a scale of 0 – 4 pts. Consider color coded answers. Handwritten can receive full credit, but must be legible. This is not a timed exam, so please make my life easier. If handwritten, please see links posted for guidance on how to submit a scan or photos as 1 document.

**Questions 1 through 9:** In 1846 a party of settlers led by the Donner and Reed families left Illinois in covered wagons headed for California, traveling via a new and untested route. After many delays they were stranded in the Sierra Nevada Mountains near Lake Tahoe by an unusually early and intense snowstorm. A few members escaped or were rescued; however, by the spring many people had died. The episode is famous (or infamous) because it was discovered that some of the party members had survived by cannibalism. Grayson (1990) compiled **Age** (in years), **Sex** (male or female) and **Survival** (died or survived) records on 45 adult Donner party members and used that data to address the question of whether women are biologically better able to withstand harsh conditions than men. Two logistic regression models are considered. The R analysis is included at the end of the exam as “**Donner**”. Use a=0.05.

1. A chi-square test is provided in the output. Provide the null hypothesis being tested. Give the test statistic and p-value.

* **H0**: Sex and Survival are independent; no association between categorical variables.
* **Conclusion**: Based on the Chi-squared test with Yates’ continuity correction, we fail to reject the null hypothesis, *chi ts* = 3.2512, *p* = 0.07137 > *alpha* = 0.05. There is not sufficient evidence to suggest an association between sex and survival.

2. Considering Models 1 and 2, what is considered an “event” for the purposes of these logistic regression analyses? In other words, are we modeling the probability of survival or probability of death? Just circle one answer; no need to justify.

DIED SURVIVED



* Because “death” is assigned as “1” in the binary variable, “survival01”.

3. **Consider Model 1,**  What do we learn from the hoslem.test() result? Be specific about your conclusions here.

* H0: The logistic regression model fits.
* Based on the Hosmer-Lemeshow lack of fit test for logistic regression, the model appears to fit, *chi ts* = 5.4048, *p* = 0.7355 > *alpha* = 0.05. This logistic regression model is an appropriate approach for modeling these data.

4. **Using Model 2**, provide an interpretation for the effect associated with Age. Be specific.

* For each additional year in age (i.e., one-unit increase in age), the odds of death are multiplied by *e*0.07820 = 1.0813 within each sex group.

5. Using Model 2, the 95% CI for exp(confint(Model2)) shows the interval associated with age to be (1.014, 1.176). Based on this information, can we conclude that there is a statistically significant effect of Age? Justify your response.

* Yes, the confidence interval does not include 1, thus we can conclude there is a statistically significant effect of age on death.
* Also, the associated p-value has an asterik (\*) that means *p* < 0.05.

6. The data were compiled to study the research hypothesis that women are better biologically adapted to harsh conditions than men. **Using Model2**, discuss whether these results support that hypothesis. Note: To get full credit you should reference (and interpret) a particular estimate(s) and test (or confidence interval).

* Yes, these results support the hypothesis that women are more likely to survive.
  + The odds of men dying are multiplied by *e*1.59729 = 4.9396 compared to women, with age held constant.
    - The associated *p*-value is 0.0345 < *alpha* = 0.05.
  + The corresponding, exponeniated confidence interval for the male group odds of death is (1.215, 25.25).
    - This confidence interval does not contain 1, indicating statistical signficance.
  + The probability of death for women is approximately 32%, compared to approximately 70% for men.

7. Using the prop.table() output, the probability of death for Males is estimated to be 66.6%. Using **Model 2**, the probability of death for Males is estimated to be 70.1% (see the emmeans() output). Explain the difference between these two estimates. Note: I am looking for more than just that the fact that one estimate is from logistic regression.

* Model 2 considers age as a covariate, but prop.table() does not.
* prop.table() provides raw proportions of death and survival for each sex group.
* Model 2 provides a model-based (logistic regression model) estimate of survival accounting for sex AND age.
  + Age is an important covariate because men had greater variability in age and were, on average, older than women.

**Questions 8 through 14 (Wood Trts):** A study was done to compare wood hardness after 5 different beetle- repellant treatments (**Trt** = A, B, C, D or E). The research goal is to look for differences between treatments. A total of **8 logs** are used in the study. Each log is cut into five pieces and the treatments are randomly assigned to pieces such that each Trt is represented exactly once with each Log. The hardness of the wood (Y) is measured two-weeks after trt application. The R analysis is included at the end of the exam as “**Wood Trts**”. Use α=0.05.

8. Write a statistical model corresponding to Model1. You can use greek letters (or not), but indicate which factor corresponds to each letter.

* Balanced Randomized Complete Block design
* 5 treatments randomly assigned in 8 blocks
  + Yij = mu + ti + bj + errorij
    - t = 5 wood treatments; A – E
    - b = 8 logs; 1 – 8
    - *i =* 1,…*t*
    - *j* = 1,…*b*
    - Yij = response (mean wood hardness) for *i*th treatment in *j*th block
    - mu = overall population mean of wood hardness (intercept)
    - ti = the *i*th treatment effect (5 treatments; A – E)
    - bj = the *j*th block effect (8 logs; 1 – 8)
    - errorij = random error; independent, normal

9. Considering Model1, a colleague considers your analysis and says “Since differences between logs are not of research interest, you should drop log from the model.” Justify why log should be kept in the model. I am looking for more than just a statement about significance!

* Because the treatments were randomized within blocks (logs 1-8), they should be analyzed in these blocks. Even if block was a nonsignificant term, it is important to retain to account for the experimental design.

10. Briefly explain the difference between Models 1 and 2. Be specific!

* Model 1 treats block as fixed effect; uses lm()
* Model 2 treats block as random effect; uses lmer()

11. Comparing Models 1 and 2, the test of Log is different. Specifically, for Model1, F= 12.569, p-value < 0.001. For Model2, Chi.sq = 24.2, p-value < 0.001. State the null hypothesis corresponding to the test of Log for each model.

H0 model 1: beta1 = beta2 = … = beta8; no difference in coefficient estimates.

H0 model 2: variance of block (log) = 0.

12. If the goal of the analysis is the compare treatments, are the results affected by whether Model 1 or 2 is used? Justify your response.

* No, there is no difference across treatment results in either model. The resulting *F* statistics (1.9838) and *p* values (0.1244) for Model 1 and Model 2 are the same for treatment. Treating block (logs) as fixed (Model1) or random (Model2) does not change the presence of a difference in treatments. In both models, no differences across treatments are detected.
* When comparing means, models will give the same results because block effects “cancel out” as long as there is no missing data.

13. If the goal of the analysis is to construct confidence intervals for the individual treatment means, are the results affected by whether Model 1 or 2 is used? Justify your response.

* Yes. These models produce different confidence intervals for individual treatment means. When making inference about individual means, treating blocks as random will give wider confidence intervals (and larger *p*-values) because we are accounting for block-to-block variability. There’s more uncertainty because we are trying to generalize a larger population; we sacrifice power and precision when using a random effects model.

14. Suppose the investigators are planning another experiment, and would like to compare 6 treatments with 12 logs (but still 5 pieces per log). Is a Balanced Incomplete Block design possible for this scenario? Justify your response.

BIBD possible: Yes No



* Both conditions are satisfied.

Justification:

1. **N = tr = bk**

t = 6

r = r

b = 12

k = 5

6r = 60

r = 10

tr = bk

6\*10 = 12\*5

**60 = 60**

1. **lambda = r(k-1)/(t-1)** = 10(5-1)/(6-1) = 40/5 = 8 = **integer**

**Questions 15 through 25 :** An investigator is interested in comparing the **Yield** for 5 wheat **Varieties** (V1, V2, V3, V4, V5) and 3 **Fertilizers** (F1, F2, F3). A single field was divided into 45 uniform plots and the 15 treatment combinations (5 Varieties x 3 Fertilizers) were randomly assigned to plots with the only restriction that each treatment be assigned to three plots (n=3 plots per treatment combination). It is known that there are differences between the Varieties, so the primary interest is in comparing Fertilizer treatments.

The R analysis is included at the end of the exam as “**Wheat**”. Use a=0.05.

15. Do the model assumptions (of equal variance and normally distributed residuals) appear to be satisfied? Discuss specific evidence from the output.

Equal Variance: Based on the Residuals vs. Fitted values plot, the assumption of equal variance appears to be satisfied; the residuals seem randomly scattered across levels of fitted values.

Normally Distributed Residuals: Based on the Normal Q-Q plot, the residuals seem normally distributed; they are tightly fitted to the line and do not show heavy tails (i.e., diverging ends).

16. In the Type 3 ANOVA table, look at the line labeled “Fertilizer” with F= 17.38 and p-value <0.0001. Explain in words what is being tested here. Note: I am looking for more than “main effect” of Fertilizer.

* **H0**: No difference in Yield by Fertilizer, averaging over Variety.
* **Conclusion**: We reject the null hypothesis. There is at least one significant difference in Yield by Fertilizer, averaging over Variety, *F* = 17.38, *p* < 0.0001.

17. In the Type 3 ANOVA table, look at the line labeled “Variety\*Fertilizer” with F= 1.50 and p-value = 0.1992. Explain in words what is being tested here. Note: I am looking for more than “interaction between Fertilizer and Variety”.

* Does the effect of Fertilizer depend on Variety?
* **H0**: No evidence of an interaction between Fertilizer and Variety.
* **Conclusion**: We fail to reject the null hypothesis. There is not sufficient evidence to suggest that the effect of Variety depends on Fertilizer, *F* = 1.50, *p* = 0.1992.

18. Looking at the interaction plot and the Type 3 ANOVA table, a colleague expresses surprise that the Variety\*Fertilizer interaction is not significant (p-value = 0.1992) given that the lines in the interaction plot are not parallel. Explain how this is possible.

* It is possible to have intersecting lines in a plot without a significant interaction result. It depends on the level of uncertainty in the estimates. A plot that does not illustrate estimates of uncertainty (e..g, standard errors) can be misleading.

19. Considering the Type 3 ANOVA table, what follow-up pairwise comparisons would be of interest to compare the Fertilizer treatments? Give an appropriate emmeans statement that could be used. Justify briefly.

* Because the interaction (Variety\*Fertilizer) is not significant and the research question is focused on Fertilizers, it is appropriate to examine the pairwise comparisons for Fertilizer type, averaging over Variety.
* R: emmeans::emmeans(Model, pairwise ~ Fertilizer)

20. Suppose (just for this question) that there had been a significant Variety\*Fertilizer interaction. What follow-up pairwise comparisons would be of interest to compare the Fertilizer treatments. Give an appropriate emmeans statement that could be used. Justify briefly why this is either the same or different from your answer in 19.

* This is a different emmeans analysis because of the hypothetically significant interaction. In this case, it is appropriate to examine the pairwise comparisons for Fertilizer type at each level of variety.
* R: emmeans::emmeans(Model, pairwise ~ Fertilizer|Variety)

21. Suppose (just for this question) that there had been some missing data. For example, 3 (out of 15) treatment combinations had 2 observations (instead of 3) for a total of 3 missing observations for the experiment. What modifications to the R code would be required?

* None.
  + The contrasts should still be changed.
  + R: car::Anova( , type = 3) and emmeans::emmeans() should still be used.
    - These can handle the unbalanced design.
  + The dfResid calculation should be: dfResid = Total # obs – 1 – dfA – dfB – dfAB.
    - This is automatic in R and robust to the unbalanced design.

22. Would it have been acceptable to run the analysis as a one-way ANOVA with t=15 treatments? What would be a disadvantage of this approach?

* Yes, but this approach ignores the factorial structure (experimental design), which reduces power and precision.

23. Suppose (just for this question), that the experiment had been run in an RCB design. Specifically, 3 Fields (blocks) were used. Within each field, each of the 15 treatment combinations were randomly assigned to a single plot. Provide the one line of R code for fitting an appropriate model.

* Two-way mixed model with crossed effects (randomized complete block design with replicates)
* R: lmer(Yield ~ Fertilizer\*Variety + (1|Field))

24. Since this study has 2 factors, a split-plot design could have been used. Discuss one reason why the investigator might have chosen a split-plot design.

* To gain accuracy on sub-plot factor B (Fertilizer), at the expense of less accuracy on whole-plot factor A (Variety).
* Convenience in agricultural experimentation processes.