BIOS 665: Final Problem Set (Take Home Exam)
Assigned: November 20, 2017
Due: **December 14, 2017 at 12:00 noon (HC 0001)**

Please note attendance is required at this session to submit your exam in person unless prior permission has been granted. Please note that you may not ask a classmate to submit your exam on your behalf.

If you are **not** a BIOS student and are attempting to earn a **P**, complete any 14 problems.

If you **are** a BIOS student or are attempting to earn an **H**, complete all 23 problems.

****

FOR GRADING PURPOSES,
PLEASE SUBMIT YOUR SOLUTIONS TO EACH PROBLEM *SEPARATELY*,
STAPLING WHERE NECESSARY;
PLEASE MAKE SURE YOUR *NAME APPEARS AT THE TOP OF EACH PROBLEM.*

****

ALL students: Please complete this checklist, and submit this facesheet with your exam.

BIOS 665 Final Exam Checklist
Fall 2017

NAME _____

PID _____

DEPARTMENT _____

Campus Box (CB) # _____

Check One: I am trying for an 'H' grade: [ ]
I am trying for a 'P' grade: [ ]

Please check the problems you are submitting below:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23

Honor Code: Please remember that the Honor Code is in effect, and ***all*** work must be done independently. You may only consult the instructors, and no other individuals.

All submitted work has been completed independently by me, and I am bound by the Honor Code.

Signed _____ Date_____

For all hypothesis tests, please state the method, the null hypothesis, the test statistic, the distribution to which you will compare the test statistic, and the p-value; use a two-sided significance level of 0.05, unless otherwise stated.

For estimates and tests, simply copying and pasting SAS output without any commentary will not earn full credit for this exam. Highlighting is not considered commentary. However, commentary can be as simple as, "The 95% CI for the odds ratio is (___ , ___)."

Part I

For Problems 1-10, consider the following data from a randomized clinical trial comparing treatment regimen A, treatment regimen B, or treatment regimen C in women with node-positive, early-stage breast cancer. Peripheral neuropathy (PN) is one of the most frequent toxicities associated with these regimens for the treatment of early-stage breast cancer. The researchers aimed to investigate the impact of the three different docetaxel-based regimens and tumor sizes on patient-reported outcomes of PN. Table 1 contains data on tumor size, treatment, and patient-reported symptoms of PN.

Table 1

Tumor Size	Treatment Regimen	PN Symptom				Total
		Not at all	Somewhat	Quite a bit	Very much	
Small (0 – 2.0 cm)	A	15	25	18	12	70
	B	10	25	15	18	68
	C	13	26	16	10	65
Medium (2.1 – 4.0 cm)	A	11	19	23	12	65
	B	8	15	20	20	63
	C	9	19	22	10	60
Large (≥ 4.1 cm)	A	6	8	20	22	56
	B	5	5	15	30	55
	C	8	8	19	22	57

1. Under minimal assumptions, conduct a statistical test to assess the association of tumor size (i.e., large size vs. smaller size (pooling small and medium)) with PN symptom (["Quite a bit" or "Very much"] vs. ["Not at all" or "Somewhat"]), controlling for treatment regimen. Justify your method. If statistically significant, discuss whether large tumor size is associated with more severe or less severe PN symptom.

2. Provide an estimate of the common odds ratio and its associated 95% confidence interval for the effect of tumor size (i.e., large size vs. smaller size (pooling small and medium)) on dichotomous severity of PN symptom (["Quite a bit" or "Very much"] vs. ["Not at all" or "Somewhat"]), controlling for treatment regimen.

3. Under minimal assumptions, conduct a test to determine whether there is a trend in the degree of PN symptom severity (as distinct levels) across tumor sizes (as distinct levels), controlling for treatment regimen. In a sentence, describe your results in terms of the problem.

4. Report the Spearman rank correlation coefficients and associated 95% confidence intervals, separately by treatment, as a measure of association between ordered tumor size (as distinct levels) and degree of PN symptom severity (as distinct levels).

5. Under minimal assumptions, conduct a statistical test to determine whether there are any differences in the proportions of PN symptom (["Quite a bit" or "Very much"] vs. ["Not at all" or "Somewhat"]) among the treatment groups (as distinct levels), controlling for tumor size. Consider the treatment groups as nominal. Write a sentence that explains your results in terms of the problem.

6. Under minimal assumptions, conduct a statistical test to determine whether there is a progressive shift of the distribution of the PN symptom severity (as distinct levels) across treatments (as distinct levels and ordered from A to C), controlling for tumor size. In a sentence, describe your results in terms of the problem.

7. Mathematically specify a logistic model for more severe PN symptom (["Quite a bit" or "Very much"] vs. ["Not at all" or "Somewhat"]), with main effects for treatment (considered as nominal; treating "B" as the reference) and tumor size (as distinct levels and treating "small" as the reference). State assumptions, and mathematically define all variables in the model. Interpret all model parameters.

8. Using your model from Problem 7, provide estimates and corresponding 95% confidence intervals for the odds ratios of regimens A vs. B and of "large" vs. "small" tumor size for more severe PN symptom ("Quite a bit" or "Very much) compared to less severe ("Not at all" or "Somewhat").

9. Mathematically specify a generalized logits model for each level of degree of PN symptom severity (treating "Not at all" as the reference), with main effects for treatment (considered as nominal; treating "B" as the reference) and tumor size (as distinct levels and treating "small" as the reference). State assumptions, and mathematically define all variables in the model. Interpret all model parameters.

10. Using your model from Problem 9, provide estimates and corresponding 95% confidence intervals for the odds ratios of "A" vs. "B" and of "large" vs. "small" tumor size separately for each PN symptom severity level compared to "Not at all".

Part II

For Problems 11 – 20, a randomized clinical trial was conducted to examine the effects of active treatment versus placebo on the self-reported current status of patients' unremitting rheumatoid arthritis. For the purposes of this exam, the self-reported arthritis status is a trichotomous variable consisting of (1) poor, (2) fair, and (3) good ordinal levels. An additional dichotomous variable was calculated to determine whether patients had improved from baseline at each follow-up visit. This subset of the data is presented in RHEUMARTH.SAS7BDAT and is comprised of $n = 289$ patients with complete data for follow-up visits conducted at each of one month, three months, and five months post-randomization. A description of the variables is as follows:

- ID: Unique patient identification code
- TRT: The study arm to which the patient was randomized
  - A = Active treatment
  - P = Placebo
- MONTH: The follow-up visit month (1, 3, 5)
- SEX: The patient's sex
  - F = Female
  - M = Male
- AGE: The patient's age in years
- BASE_STATUS: Self-reported rheumatoid arthritis status at baseline
  - 1 = Poor
  - 2 = Fair
  - 3 = Good
- STATUS: Self-reported rheumatoid arthritis status at the respective follow-up visit
  - 1 = Poor
  - 2 = Fair
  - 3 = Good
- IMPROVE: Whether or not the patient has improved since baseline
  - 1 = Improved
  - 0 = Did not improve

11. Mathematically specify a proportional odds model, <u>separately for each follow-up visit</u>, where the outcome of interest is the patient's self-reported rheumatoid arthritis status (ordered from good to poor). Include main effects for treatment, sex, and age of the patient as explanatory variables, with a male patient randomized to placebo representing the reference categories for the categorical explanatory variables. State your assumptions, and

mathematically define all variables in the model. Fit your specified models and perform a hypothesis test for each to assess whether the overall proportional odds assumption holds for these data at the corresponding follow-up visit. Describe your findings in one to two sentences. Hint: To restrict the outcome of interest, '`WHERE MONTH = X;`' can be included as a statement in a SAS procedure with `X` taking the value 1, 3, or 5.

12. In reference to the models in Problem 11, if there are any follow-up visits where the proportional odds assumption holds, fit your specified model(s) and report parameter estimates (including intercept(s)), their standard errors, test statistics, and p-values. Provide the odds ratio(s) and corresponding 95% confidence interval(s) comparing active treatment to placebo. Briefly interpret the treatment effect(s) on patients' arthritis status at follow-up.

13. In reference to the models in Problem 11, if there are any follow-up visits where the proportional odds assumption does not hold, investigate a more suitable model by performing hypothesis tests to assess the proportional odds assumption separately for each explanatory variable (treatment, sex, and age). Be sure to state your null hypotheses in terms of model parameters. Describe your findings in one to three sentences. Does the proportional odds assumption apply to any of the explanatory variables? Briefly state how this might change the interpretation of model parameters.

14. In reference to applicable models in Problem 13, mathematically specify your final recommended model(s) after having assessed the overall proportional odds assumption(s) in Problem 11 for each visit and the possible separate proportional odds assumptions in Problem 13. Be sure to address the main effects of treatment, sex, and age of the patient, and maintain the same reference levels as Problem 11. State your assumptions, and mathematically define all variables in the model(s). Fit your specified model(s) and report parameter estimates (including intercept(s)), their standard errors, tests statistics, and p-values. Provide the odds ratio(s) and corresponding 95% confidence interval(s) for better response comparing active treatment to placebo. Briefly interpret the treatment effect(s) on patients' arthritis status at follow-up.

15. Regardless of your findings about whether the proportional odds assumption does or does not apply to one or more months of follow-up, proceed to fit a repeated measures proportional odds model for patients' self-reported rheumatoid arthritis status at each month of follow-up (ordered from good to poor) using generalized estimating equation (GEE) methodology. Include treatment group, age, sex, baseline status, and month of follow-up (as a CLASS variable) as main effects in the model, with a male patient randomized to placebo with poor baseline status and assessed at one month follow-up representing the reference levels. Additionally, include an interaction term for treatment

group × month of follow-up and assume an independent working correlation structure. For the purposes of this exam, determine goodness-of-fit for the main effects model by only considering exclusion of the interaction between treatment group and month of follow-up. Present your final model by reporting the GEE parameter estimates (including intercept(s)) and their associated standard errors, test statistics, and p-values. Justify your choice for the final model, incorporating a brief discussion of any assumptions or limitations of the model in one to three sentences.

16. Restate your final model from Problem 15 at the top of the page for the grader's reference. Using this final model, provide the odds ratio and corresponding 95% confidence interval comparing active treatment to placebo at each month of follow-up for (good status) versus (fair or poor status) and for (good or fair status) versus (poor status). Briefly interpret the treatment effect(s) on arthritis status over the follow-up period.

17. Restate your final model from Problem 15 at the top of the page for the grader's reference. Provide model-predicted probabilities of good arthritis status, fair arthritis status, and poor arthritis status for a 50 year old male patient with poor baseline arthritis status at the five month follow-up visit and who is:

    a. randomized to the active treatment arm.

    b. randomized to the placebo arm.

18. As a follow-up analysis and using generalized estimating equation (GEE) methodology, fit a repeated measures model which describes the response variable of the proportion of patients who experience improvement from baseline across all study follow-up visits. In addition to the main effects for treatment and follow-up visit (as a CLASS variable), the model should also be adjusted for baseline status, sex, and age. A male patient randomized to placebo with poor baseline status and assessed at one month follow-up would comprise the reference levels for the categorical explanatory variables. Assuming an exchangeable correlation structure, assess goodness-of-fit for the main effects model by considering the pairwise interaction of treatment with follow-up visit. Present your final model by listing a table of parameter estimates, their standard errors, test statistics, and p-values. Justify your choice for the final model, incorporating a brief discussion of any assumption or limitation of the model in one to three sentences.

19. Restate your final model from Problem 18 at the top of the page for the grader's reference. Using this model, provide and interpret the estimated odds ratio(s), along with the corresponding 95% confidence interval(s), for the effect active treatment has on improvement from baseline (relative to placebo) at each follow-up visit.

20. In order to further examine how well physicians' evaluations of patients' rheumatoid arthritis agreed with the responses of patients, a subset of patients were asked to classify the severity of their rheumatoid arthritis as very poor, poor, fair, good, or very good at a particular clinic visit. The physicians were also asked to evaluate the patients' rheumatoid arthritis independently based on a physical exam. The resulting data are presented below in Table 2.

Table 2

		Patient Rating				
		Very Poor	**Poor**	**Fair**	**Good**	**Very Good**
Physician Rating	**Very Poor**	12	2	4	2	3
	Poor	3	7	3	4	2
	Fair	2	3	9	2	4
	Good	2	5	1	6	4
	Very Good	4	3	3	2	8

Compute and report the unweighted and weighted kappa statistics with their 95% confidence intervals. In two to three sentences, summarize your conclusions about the degree of agreement for the physicians' and patients' assessments of rheumatoid arthritis.

Part III

From 1845 to 1850, over a million people living in Ireland starved to death when potato blight decimated the staple crop upon which they subsisted. While about a million and a half Irish people had the means to emigrate to America, Britain, or Australia, the poorest among them often faced the prospect of unemployment and eviction from their homes when unable to pay rent. Once faced with such calamity, many of the starving poor entered "workhouses" established by the government to house, feed, and clothe the most destitute. Living conditions in these public houses were frequently so bad they often lacked basic sanitation, and disease was widespread. The data in Table 3 show mortality rates in workhouses for one region of Ireland. They are taken from "The Great Famine in Nenagh Poor Law Union, Co. Tipperary" by D.Grace (2000).

Table 3. Distribution of deaths and admissions by time period, sex, and age group for the Nenagh workhouses during the Irish Famine, 1846-1850.
Values in cells are number of deaths over number of admissions.

Time period	Sex	Age group (years)	
		2-15	**>15**
1846-1848	**Male**	$\frac{404}{4057}$	$\frac{354}{4359}$
	Female	$\frac{362}{3331}$	$\frac{369}{6481}$
1849-1850	**Male**	$\frac{718}{4412}$	$\frac{631}{4438}$
	Female	$\frac{602}{3538}$	$\frac{634}{6416}$

21. Fit a main effects Poisson regression model for death rate within the workhouses, treating time period, sex, and age group as explanatory variables (all treated as nominal variables). Mathematically specify the fitted Poisson regression model; define all variables in the model.

22. Determine goodness-of-fit for the main effects model relative to the saturated model with all 2- and 3-way interactions among the explanatory variables. For all parameters in the model you determine to be most appropriate (i.e., the 'final' model), report the estimates, standard errors, test statistics, and p-values. Briefly defend your choice of final model.

23. Regardless of your final model in Problem 22, use the main effects Poisson regression model and provide a table of expected death rates (per 100 admissions) for all combinations of time period, sex, and age group. Did the death rate change over time? Did the death rate vary according to sex or by age group? Briefly summarize your findings in language a non-statistician can readily understand.