

Name: Cheynna Crowley

BIOS 665
Final Exam 2017

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 Cheynna Crowley
PID 730112944
DEPARTMENT BIOS
Campus Box (CB) # BIOS

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	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2
0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3
X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

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.

Tumor Size	Treatment	PN Symptom		Total
		Not at all or Somewhat	Quite a bit or very much	
Small or Med	A	70	65	135
Small or Med	B	58	73	131
Small or Med	C	67	58	125
Large	A	14	42	56
Large	B	10	45	55
Large	C	16	41	57
Total		235	324	559

H0: There is no association between tumor size and PN symptom when controlling for treatment

H1: There is an association between tumor size and PN symptom when controlling for treatment

We use a Mantel-Hanzel test because there is large overall row sample sizes and we can control for treatment.

$$Q_{MH} \sim \chi^2_1$$

$$df=1$$

$$Q_{MH} = 33.0180$$

$$p\text{-value} < 0.0001$$

Since the p-value is less than alpha, we reject H0 and conclude there is an association between tumor size and PN symptoms when controlling for treatment.

By looking at the table of PN symptom and tumor size, we can calculate the marginal probabilities. From this we can see that probability of “Quite a bit” or “very much” (regardless of treatment) for large tumor is $128/168=0.76$ and that the probability of “Not at all” or “Somewhat” (regardless of treatment) is $40/168= 0.24$. This tells us that large tumor size is associated with the more severe PN symptom (“Quite a bit” or “very much”).

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.

OR=3.2288 95% CI=[2.1450,4.8600]

Patients with large tumor size have 3.2228 times the odds of having a PN symptom severity of ‘quite a bit’ or ‘very much’ compared to the odds of a patient with smaller size tumor.

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.

Let PN symptom severity and tumor sizes order be defined by the table given.

To test if there is a trend with the degree of PN symptom severity across tumor sizes use CSMH based on the table scores (in SAS ‘Nonzero Correlation’).

H0: There is no trend in the degree of PN symptom severity across tumor sizes when controlling for treatment.

H1: O.T.W

$$Q_{CSMH} \sim X_1^2$$

$$df=1$$

$$Q_{CSMH} = 35.0877$$

$$p\text{-value} < 0.0001$$

Since the p-value is greater than an alpha of 0.05 we conclude there is a significant trend between PN symptom severity and tumor sizes when controlling for treatment..

Name: Cheynna Crowley

BIOS 665
Final Exam 2017

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).

Treatment	Spearman Rank Correlation	95% CI
A	0.2552	(0.1180,0.3925)
B	0.2649	(0.1269,0.4029)
C	0.2516	(0.1080,0.3952)

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.

		PN Symptom		
Tumor Size	Treatment	Not at all or Somewhat	Quite a bit or very much	Total
Small	A	40	30	70
Small	B	35	33	68
Small	C	39	26	65
Med	A	30	35	65
Med	B	23	40	63
Med	C	28	32	60
Large	A	14	42	56
Large	B	10	45	55
Large	C	16	41	57
Total		235	324	559

Let treatment groups be nominal and PN symptom for $\rightarrow Q_{SMH}$ (row mean score)

H₀: There are no 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.

H₁: OTW

$$Q_{SMH} \sim X_2^2$$

$$Q_{SMH} = 4.0305$$

$$p\text{-value} = 0.1330$$

Since the p-value is greater than an alpha of 0.05, we fail to reject H₀ and conclude that there is not a significant difference among treatment groups controlling for tumor size..

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.

Since we are looking for the progressive location shift with row and column variable ordinal it implies we want to use the Q_{CSMH} test (Non-Zero Correlation Test).

H_0 : There is no progressive shift in the distribution of the PN symptom severity across treatments controlling for tumor size.

H_1 : There is a progressive shift in distribution.

$$Q_{CSMH} \sim X_1^2$$

$$df=1$$

$$Q_{CSMH} = 0.0846$$

$$p\text{-value} = 0.7712$$

Since the p-value is greater than an alpha of 0.05, we fail to reject H_0 and conclude there is not a progressive shift in the distribution.

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.

Assumptions:

- Data is from a stratified random sample
- Observation are independent
- Sufficient sample size
- Independent error terms

$$\text{Model: } \text{logit}(\theta_{ij}) = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + \beta_3 X_{i3} + \beta_4 X_{i4}$$

$$i=1,2 \quad j=1,2$$

θ_{ij} is the probability of more severe PN symptom

$$X_{1j} = I(\text{treatment}=A)$$

$$X_{2j} = I(\text{treatment}=C)$$

$$X_{i1} = I(\text{tumor size}=large)$$

$$X_{i2} = I(\text{tumor size}=medium)$$

Parameter	Estimate (sd error)	Interpretation
α	0.000032 (0.1897)	Log odds of more severe PN symptom, being in treatment B and small tumor size (intercept)
β_1	-0.3297(0.2186)	Incremental in log odds for treatment A
β_2	-0.4237(0.2213)	Incremental in log odds for treatment C
β_3	1.4254(0.2309)	Incremental in log odds for large tumor
β_4	0.5301(0.2051)	Incremental in log odds for medium tumor

$$\text{logit}(\theta_{ij}) = 0.000032 - 0.3297X_{1j} - 0.4237X_{2j} + 1.4254X_3 + 0.5301X_4$$

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”).

Effect	Point Estimate (OR)	95% Wald Confidence Limit
Treatment A vs B	0.719	(0.469,1.104)
Tumor: Large vs Small	4.160	(2.645,6.541)

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.

Assumptions:

- Data is from a stratified random sample
- Agreement is nominal
- Cell counts >5

Model: $\text{logit}(\theta_{hik}) = \alpha_k + x_{1i1}\beta_1 + x_{1i2}\beta_2 + x_{1i3}\beta_3 + x_{2i1}\beta_4 + x_{2i2}\beta_5 + x_{2i3}\beta_6 + x_{h11}\beta_7 + x_{h12}\beta_8 + x_{h13}\beta_9 + x_{h21}\beta_{10} + x_{h22}\beta_{11} + x_{h23}\beta_{12}$

Where:

θ_{hik} is odds of PN symptom severity with h treatment level and i tumor size

h=1 for treatment A, h=0 for treatment B(ref), and h=2 for treatment C

i=0 for small tumor (ref), i=1 for medium tumor, i=2 for large tumor

k=1,2,3 where k=1 is the odds of ‘somewhat’ vs. ‘not at all’, k=2 is the odds of ‘quite a bit’ vs. ‘not at all’, k=3 is the odds of ‘very much’ vs. ‘not at all’

$$\begin{array}{lll} x_{1i1} = I(\text{treatment}=A, k=1) & x_{1i2} = I(\text{treatment}=A, k=2) & x_{1i3} = I(\text{treatment}=A, k=3) \\ x_{2i1} = I(\text{treatment}=C, k=1) & x_{2i2} = I(\text{treatment}=C, k=2) & x_{2i3} = I(\text{treatment}=C, k=3) \\ x_{h11} = I(\text{tumor size}=medium, k=1) & x_{h12} = I(\text{tumor size}=medium, k=2) & \\ x_{h13} = I(\text{tumor size}=medium, k=3) & x_{h21} = I(\text{tumor size}=large, k=1) & \\ x_{h22} = I(\text{tumor size}=large, k=2) & x_{h23} = I(\text{tumor size}=large, k=3) & \end{array}$$

Parameter	Estimate (sd error)	Interpretation
α_1	0.7835(0.7835)	Intercept for the 1 st cumulative logit (reference treatment B, small tumor)
α_2	0.3633(0.3022)	Intercept for the 2 nd cumulative logit (reference treatment B, small tumor)
α_3	0.5476(0.3004)	Intercept for the 3 rd cumulative logit (reference treatment B, small tumor)
β_1	-0.1742(0.3415)	Incremental effect for treatment A 1 st cumulative logit
β_2	-0.1472(0.3353)	Incremental effect for treatment A 2 nd cumulative logit
β_3	-0.7580(0.3395)	Incremental effect for treatment A 3 rd cumulative logit
β_4	-0.0795(0.3443)	Incremental effect for treatment C 1 st cumulative logit
β_5	-0.1642(0.3405)	Incremental effect for treatment C 2 nd cumulative logit
β_6	-0.8186(0.3463)	Incremental effect for treatment C 3 rd cumulative logit
β_7	-0.0545(0.3068)	Incremental effect for medium tumor 1 st

Name: Cheyonna Crowley

BIOS 665
Final Exam 2017

		cumulative logit
β_8	0.5888(0.3129)	Incremental effect for medium tumor 2 nd cumulative logit
β_9	0.3588(0.3353)	Incremental effect for medium tumor 3 rd cumulative logit
β_{10}	-0.5899(0.3742)	Incremental effect for large tumor 1 st cumulative logit,
β_{11}	0.7974(0.3438)	Incremental effect for large tumor 2 nd cumulative logit
β_{12}	1.3473(0.3459)	Incremental effect for large tumor 3 rd cumulative logit

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”.

Effect	PN symptom	OR Point Estimate	95% CI
Treat A vs B	Some What	0.840	(0.430,1.641)
Treat A vs B	Quite a Bit	0.863	(0.447,1.665)
Treat A vs B	Very Much	0.469	(0.241,0.912)
Tumor Large vs Small	Some What	0.554	(0.266,1.154)
Tumor Large vs Small	Quite a Bit	2.220	(1.132,4.354)
Tumor Large vs Small	Very Much	3.847	(1.953,7.577)

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, separately for each follow-up visit, 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.

Assumptions:

Data that arise from a stratified simple random sample, at least 5 observations at each outcome at each level of each main effect, $\beta_k = \beta$ for all k

Month=1:

Model $\text{logit}(\theta_k) = \alpha_1 + \alpha_2 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3$

Where

θ_1 is the P(Status= Good)

θ_2 is the P(Status=Good or Status=fair)

α_1 = Intercept for the 1st cumulative logit (log odds of good status vs fair or poor for males on placebo)

α_2 = Intercept for the 2nd cumulative logit (log odds of good status or fair vs poor for males on placebo)

x_1 =I (sex=Female)

x_2 =I (treatment=Active)

x_3 = age

β_1 = Incremental effect for both types of log odds due to female sex

β_2 = Incremental effect for both types of log odds due to active treatment

β_3 = Incremental effect for both types of log odds per year of age

H0: For Month 1 follow up, $\beta_k = \beta$ for all k

Score Test for Proportional Odds Assumption

Requirement: need 5 observations as each outcome at each level of each main effect

$X^2 = 17.7363$

df=3

p-value=0.0005

Conclusion: Since the p-value is less than alpha, we reject H0 and conclude that for follow up month =1, the overall proportional odds assumption does not hold.

Month=3:

Model $\text{logit}(\theta_k) = \alpha_1 + \alpha_2 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3$

Where

θ_1 is the P(Status= Good)

θ_2 is the P(Status=Good or Status=fair)

α_1 = Intercept for the 1st cumulative logit (log odds of good status vs fair or poor for males on placebo)

α_2 = Intercept for the 2nd cumulative logit (log odds of good status or fair vs poor for males on placebo)

$x_1=I(\text{sex}=\text{Female})$

$x_2=I(\text{treatment}=\text{Active})$

$x_3=\text{age}$

β_1 = Incremental effect for both types of log odds due to female sex

β_2 = Incremental effect for both types of log odds due to active treatment

β_3 = Incremental effect for both types of log odds per year of age

For Month 3, hypothesis test for each to assess whether the overall proportional odds assumption holds for these data at the corresponding follow-up visit.

H0: For Month 3 follow up, $\beta_k = \beta$ for all k

Score Test for Proportional Odds Assumption

Requirement: need 5 observations as each outcome at each level of each main effect

$X^2 = 1.0830$

df=3

p-value=0.7812

Conclusion: Since the p-value is greater than alpha, we fail to reject H0 and conclude that for follow up visit month=3, the overall proportional odds assumption does hold.

Month=5:

Model $\text{logit}(\theta_k) = \alpha_1 + \alpha_2 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3$

Where

θ_1 is the P(Status= Good)

θ_2 is the P(Status=Good or Status=fair)

α_1 = Intercept for the 1st cumulative logit (log odds of good status vs fair or poor for males on placebo)

α_2 = Intercept for the 2nd cumulative logit (log odds of good status or fair vs poor for males on placebo)

$x_1=I(\text{sex}=\text{Female})$

$x_2=I(\text{treatment}=\text{Active})$

$x_3=\text{age}$

β_1 = Incremental effect for both types of log odds due to female sex

β_2 = Incremental effect for both types of log odds due to active treatment

β_3 = Incremental effect for both types of log odds per year of age

H0: For Month 5 follow up, $\beta_k = \beta$ for all k

Name: Cheynna Crowley

BIOS 665
Final Exam 2017

Score Test for Proportional Odds Assumption

Requirement: need 5 observations as each outcome at each level of each main effect

$\chi^2 = 1.4434$

df=3

p-value=0.6954

Conclusion: Since the p-value is greater than alpha, we fail to reject H0 and conclude that for follow up visit month=5, the overall proportional odds assumption does hold.

Summary: The overall proportion odds assumption holds for month 3 and month 5. It does not hold for month 1.

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.

From question 11, we found that month 3 and month 5 are visits where the proportional odds assumptions hold.

Month 3:

Parameter	Estimate (sd error)	Wald Chi-Square	P-value
α_1	-0.1410(0.5336)	0.0698	0.7916
α_2	1.5809(0.5423)	8.4984	0.0036
β_1	0.1531(0.2464)	0.3859	0.5345
β_2	0.5604(0.2208)	6.4423	0.0111
β_3	-0.0137(0.0095)	1.8959	0.1685

Active Vs. Placebo OR=1.751

95% CI=(1.136, 2.7)

Interpretation:

At Month 3 of Follow-up, the odds of good response versus a poor or fair response for patients on active treatment are 1.751 times the odds of a good response versus a poor or fair response for patients on placebo treatment. The 95% Confidence Interval does not contain the null (1) so we can conclude this result is statistically significant.

Month 5:

Parameter	Estimate (sd error)	Wald Chi-Square	P-Value
α_1	0.3839(0.5424)	0.5011	0.4790
α_2	2.0113(0.5557)	13.099	0.0003
β_1	-0.1448(0.2477)	0.3419	0.5588
β_2	0.5156(0.2228)	5.3545	0.0207
β_3	-0.0158(0.0101)	2.4397	0.1183

Active Vs. Placebo OR=1.675

95% CI=(1.802,2.592)

Interpretation:

At Month 3 of Follow-up, the odds of good response versus a poor or fair response for patients on active treatment are 1.675 times the odds of a good response versus a poor or fair response for patients on placebo treatment. The 95% Confidence Interval does not contain the null (1) so we can conclude this result is statistically significant.

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.

From question 11, we found that month 1 was a follow up visit where the proportional odds assumptions did not hold. Since the proportional of odds assumption does not hold, we can create a contrast test to assess the equality of the three parameters considered separately.

For Month 1:

$$H_0: \text{age_3} = \text{age_2}$$

$$H_1: \text{age_3} \neq \text{age_2}$$

$$X^2=4.0865$$

$$\text{p-value}=0.0432$$

The p-value is less than alpha=0.05 so we reject H0 and conclude that the age effect indicates that the equal slopes assumption is not viable.

$$H_0: \text{SEXF_3} = \text{SEXF_2};$$

$$H_1: \text{SEXF_3} \neq \text{SEXF_2}$$

$$X^2=0.2137$$

$$\text{p-value}=0.6439$$

The p-value is greater than alpha=0.05 so we fail to reject H0 and conclude that the gender effect indicates that the equal slopes assumption is viable.

$$H_0: \text{TRTA_3} = \text{TRTA_2};$$

$$H_1: \text{TRTA_3} \neq \text{TRTA_2};$$

$$X^2=12.9152$$

$$\text{p-value}=0.0003$$

The p-value is less than alpha=0.05 so we reject H0 and conclude that the treatment effect indicates that the equal slopes assumption is not viable.

The new proc logistic statement request a partial proportional odds model, where gender is handled with a single slope for both cumulative logits modeled and treatment and age are handled with a different parameter for each cumulative logit.

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.

Month 1:

Model: $\text{logit}(\theta_k) = \alpha_1 + \alpha_2 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$

α_1 = Intercept for the 1st cumulative logit (log odds of good status vs fair or poor for males on placebo)

α_2 = Intercept for the 2nd cumulative logit (log odds of good status or fair vs poor for males on placebo)

$x_1 = I(\text{sex}=\text{Female})$ $x_2 = I(\text{treatment}=\text{Active})$ $x_3 = \text{age}$

β_1 = Incremental effect for both types of log odds due to female sex

β_2 = Incremental effect for log odds due to active treatment for logit model 1

β_3 = Incremental effect for all types of log odds per year of age

β_4 = Incremental effect for log odds due to active treatment for logit model 2

Assumptions: overall proportional odds assumption does not hold and treatment has unequal slopes. In part 13, age was borderline significant, and since it is a continuous variable, it was decided that age would use the assumption of equal slopes.

Parameter	Estimate (sd error)	Wald Chi-Square	P-value
α_1	-0.1460(0.5522)	0.0699	0.7914
α_2	1.2458(0.5592)	4.9636	0.0259
β_1	-0.2835(0.2471)	1.3162	0.2513
β_2	0.00420	0.2439	0.9863
β_3	-0.00583(0.0101)	0.3308	0.56521
β_4	1.1453(0.3164)	13.1062	0.0003

OR Active vs. Placebo for logit Model 1: 1.004

OR 95% CI for logit model 1: (0.623, 1.620)

OR Active vs. Placebo for logit Model 2: 3.143

OR 95% CI for logit model 1: (1.691, 5.844)

Interpretation:

Since in the previous part, we found that model 1 did not fulfill the equal slopes criteria (specifically treatment and age) we updated our model to have unequal slopes for treatment. It was decided that since age is continuous and was only borderline significant (p-value of 0.0432) that we did not model for different slopes. This creates two Odds Ratios when comparing active

to placebo treatment effect as there are different treatment effects based on the cumulative logit you use in the model. For the first cumulative logit, which compares status good to fair or none, a patient on active has 1.004 times the odds of having a good status outcome compare to a patient on placebo. This result is not significant. For the second cumulative logit, which compares status good or fair to none, a patient on active has 3.143 times the odds of having a good status outcome compare to a patient on placebo. This result is significant.

Month 3:

Model: $\text{logit}(\theta_k) = \alpha_1 + \alpha_2 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

Assumptions: equal slopes (overall proportional odds assumption)

Parameter	Estimate (sd error)	Wald Chi-Square	P-value
α_1	-0.1410(0.5336)	0.0698	0.7916
α_2	1.5809(0.5423)	8.4984	0.0036
β_1	0.1531(0.2464)	0.3859	0.5345
β_2	0.5604(0.2208)	6.4423	0.0111
β_3	-0.0137(0.0095)	1.8959	0.1685

Active Vs. Placebo OR=1.751

95% CI=(1.136, 2.7)

Interpretation:

At Month 3 of Follow-up, the odds of good response verses a poor or fair response for patients on active treatment are 1.751 times the odds of a good response verses a poor or fair response for patients on placebo treatment. The 95% Confidence Interval does not contain the null (1) so we can conclude this result is statistically significant.

Month 5:

Model:

Assumptions:

Parameter	Estimate (sd error)	Wald Chi-Square	P-Value
α_1	0.3839(0.5424)	0.5011	0.4790
α_2	2.0113(0.5557)	13.099	0.0003
β_1	-0.1448(0.2477)	0.3419	0.5588
β_2	0.5156(0.2228)	5.3545	0.0207
β_3	-0.0158(0.0101)	2.4397	0.1183

Active Vs. Placebo OR=1.675

95% CI=(1.802,2.592)

Interpretation:

At Month 3 of Follow-up, the odds of good response verses a poor or fair response for patients on active treatment are 1.675 times the odds of a good response verses a poor or fair response for patients on placebo treatment. The 95% Confidence Interval does not contain the null (1) so we can conclude this result is statistically significant.

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 \times 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.

Assumptions:

- The model relates a marginal mean to the linear predictor $x'_{ij}\beta$ through a link function.

Model:

$$\text{logit}(\theta_k) = \alpha_1 + \alpha_2 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + x_4\beta_4 + x_5\beta_5 + x_6\beta_6 + x_7\beta_7 + x_8\beta_8 + x_9\beta_9$$

α_1 = First Cummulative Logit

α_2 = Second Cummulative Logit

x_1 = I (treatment=Active)

x_2 = I (sex=Female)

x_3 = age

x_4 = I (base status=2)

x_5 = I (base status=3)

x_6 = I (month=3)

x_7 = I (month=5)

x_8 = I (treatment=Active, month=3)

x_9 = I (treatment=Active, month=5)

Score Statistics for Type 3 GEE Analysis			
Source	DF	Chi-Square	P-Value
TRT	1	9.18	0.0024
AGE	1	0.82	0.3663
SEX	1	0.26	0.6119
BASE STATUS	2	47.44	<0.0001
MONTH	2	8.29	0.0158
TRT*MONTH	2	0.81	0.666

GEE fit criteria: 1721.1055

Test Model of $\text{logit}(\theta_k) = \alpha_1 + \alpha_2 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + x_4\beta_4 + x_5\beta_5 + x_6\beta_6 + x_7\beta_7$
 *exclusion of interaction effect of treatment and month since p-value is 0.666

Score Statistics for Type 3 GEE Analysis			
Source	DF	Chi-Square	P-Value
TRT	1	9.14	0.0025
AGE	1	0.82	0.3660
SEX	1	0.26	0.6093
BASE STATUS	2	47.41	<0.0001
MONTH	2	8.44	0.0147

GEE fit criteria: 1719.3129

Final Model: $\text{logit}(\theta_k) = \alpha_1 + \alpha_2 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + x_4\beta_4 + x_5\beta_5 + x_6\beta_6 + x_7\beta_7$

α_1 = First Cummulative Logit

α_2 = Second Cummulative Logit

x_1 = I (treatment=Active)

x_2 = I (sex=Female)

x_3 = age

x_4 = I (base status=2)

x_5 = I (base status=3)

x_6 = I (month=3)

x_7 = I (month=5)

Analysis of GEE Parameter Estimates					
Parameter	Estimate	SE	CI	Z	p-value
α_1	-1.2271	0.4873	-2.1822, -0.2720	-2.52	0.0118
α_2	0.7117	0.4844	-0.2377, 1.6610	1.47	0.1418
β_1	0.5447	0.1794	0.1931, 0.8964	3.04	0.0024
β_2	-0.0926	0.1855	-0.4562, 0.2710	-0.50	0.6176
β_3	-0.0072	0.0080	-0.0229, 0.0086	-0.89	0.3721
β_4	0.7327	0.2076	0.3258, 1.1397	3.53	0.0004
β_5	2.1660	0.2888	1.600, 2.7320	7.50	<0.0001
β_6	-0.0376	0.1273	-0.2871, 0.2118	-0.30	0.7675
β_7	0.2703	0.1195	0.0361, 0.5045	2.26	0.0237

The model assumes that there is no significant interaction between month and treatment.

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.

Final Model: $\text{logit}(\theta_k) = \alpha_1 + \alpha_2 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + x_4\beta_4 + x_5\beta_5 + x_6\beta_6 + x_7\beta_7$

α_1 = First Cummulative Logit

α_2 = Second Cummulative Logit

$x_1=I(\text{treatment}=\text{Active})$

$x_2=I(\text{sex}=\text{Female})$

$x_3=\text{age}$

$x_4=I(\text{base status}=2)$

$x_5=I(\text{base status}=3)$

$x_6=I(\text{month}=3)$

$x_7=I(\text{month}=5)$

$$\text{Month 1, Status Good vs fair or poor trt vs placebo} = \frac{e^{\alpha_1 + \beta_1}}{e^{\alpha_1}} = e^{\beta_1} = 1.72$$

$$\text{Month 3, status good vs fair or poor trt vs placebo} = \frac{e^{\alpha_1 + \beta_1 + \beta_6}}{e^{\alpha_1 + \beta_6}} = e^{\beta_1} = 1.72$$

$$\text{Month 5, status good vs fair or poor, trt vs placebo} = \frac{e^{\alpha_1 + \beta_1 + \beta_7}}{e^{\alpha_1 + \beta_7}} = e^{\beta_1} = 1.72$$

$$\text{Month 1, Status Good or fair vs poor trt vs placebo} = \frac{e^{\alpha_2 + \beta_1}}{e^{\alpha_2}} = e^{\beta_1} = 1.72$$

$$\text{Month 3, status good or fair vs poor trt vs placebo} = \frac{e^{\alpha_2 + \beta_1 + \beta_6}}{e^{\alpha_2 + \beta_6}} = e^{\beta_1} = 1.72$$

$$\text{Month 5, status good or fair vs poor, trt vs placebo} = \frac{e^{\alpha_2 + \beta_1 + \beta_7}}{e^{\alpha_2 + \beta_7}} = e^{\beta_1} = 1.72$$

Follow Up Month	Status	OR	95% CI
1	Good vs fair or poor	1.72	1.21,2.45
3	Good vs fair or poor	1.72	1.21,2.45
5	Good vs fair or poor	1.72	1.21,2.45
1	Good or fair vs poor	1.72	1.21,2.45
3	Good or fair vs poor	1.72	1.21,2.45
5	Good or fair vs poor	1.72	1.21,2.45

Since we are only comparing the treatment effect for each specific circumstance, you can see that the month of follow up does not impact the treatment effect. If we were comparing treatment effect month 1 verse month 3 we would see a difference.

Patients on active treatment have 1.72 times the odds of good status verses fair or poor at follow up compared to those on placebo. Similarly, Patients on active treatment have 1.72 times the odds of good or fair status verses poor at follow up compared to those on placebo

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:

Final Model: $\text{logit}(\theta_k) = \alpha_1 + \alpha_2 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + x_4\beta_4 + x_5\beta_5 + x_6\beta_6 + x_7\beta_7$

α_1 = First Cummulative Logit

α_2 = Second Cummulative Logit

$x_1=I(\text{treatment}=\text{Active})$

$x_2=I(\text{sex}=\text{Female})$

$x_3=\text{age}$

$x_4=I(\text{base status}=2)$

$x_5=I(\text{base status}=3)$

$x_6=I(\text{month}=3)$

$x_7=I(\text{month}=5)$

a. randomized to the active treatment arm.

Arthritis Status	Model	Predicted Probability
Good	$\alpha_1 + \beta_1 + 50\beta_3 + \beta_7$	0.316
Good or Fair	$\alpha_2 + \beta_1 + 50\beta_3 + \beta_7$	0.763
Fair	$0.763 - 0.316 = 0.447$	0.447
Poor	$1 - (0.316 + 0.447) = 0.237$	0.237

*by bayes theory and independence

b. randomized to the placebo arm.

Arthritis Status	Model	Predicted Probability
Good	$\alpha_1 + 50\beta_3 + \beta_7$	0.211
Good or Fair	$\alpha_2 + 50\beta_3 + \beta_7$	0.651
Fair	$0.651 - 0.211 = 0.44$	0.44
Poor	$1 - 0.651 = 0.349$	0.349

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.

First considering all main effects and pairwise interaction between treatment and follow-up visit

$$\text{logit}(\theta) = \alpha_1 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + x_4\beta_4 + x_5\beta_5 + x_6\beta_6 + x_7\beta_7 + x_8\beta_8 + x_9\beta_9$$

α_1 =intercept (treatment=placebo, sex=male,month=1,baseline=1)

x_1 =I (treatment=Active)

x_2 = I (month=3)

x_3 = I(month=5)

x_4 =I (base status=2)

x_5 =I (base status=3)

x_6 =I (sex=female)

x_7 =age

x_8 =I(treatment A, month=3)

x_9 =I(treatment A, month=5)

Score Statistics For Type 3 GEE Analysis			
	DF	Chi Square	PValue
TRT	1	0.80	0.3697
MONTH	2	8.53	0.0140
BASE_STATUS	2	69.00	<0.0001
SEX	1	0.73	0.3926
AGE	1	1.85	0.1743
TRT*MONTH	2	1.25	0.5340

GEE FIT CRITERIA : 1033.4505

→ By looking at the Score Statistics /Type 3 information with the full model, we see that the interaction variable TRT*MONTH is not significant. We can proceed to drop it from the model and see the new parameter estimates.

Final Model:

$$\text{logit}(\theta) = \alpha_1 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + x_4\beta_4 + x_5\beta_5 + x_6\beta_6 + x_7\beta_7$$

α_1 =intercept (treatment=placebo, sex=male, month=1, baseline=1)

x_1 =I (treatment=Active)

x_2 = I (month=3)

x_3 = I(month=5)

x_4 =I (base status=2)

x_5 =I (base status=3)

x_6 =I (sex=female)

x_7 =age

Parameter	Estimate(SE)	95%CI	Z	p-value
α_1	1.0883(0.5259)	(0.0576,2.1190)	2.07	0.0385
β_1	0.4090(0.1978)	(0.0213,0.7967)	2.07	0.0387
β_2	0.2133(0.1588)	(-0.0979,0.5245)	1.34	0.1792
β_3	0.6064(0.1515)	(0.3095,0.9033)	4.00	<0.0001
β_4	-1.5463(0.2264)	(-1.9899,-1.1026)	-6.83	<0.0001
β_5	-2.4908(0.2943)	(-3.0675,-1.9140)	-8.46	<0.0001
β_6	-0.1937(0.2272)	(-0.6390,0.2517)	-0.85	0.3941
β_7	-0.0119(0.0089)	(-0.0294,0.0056)	-1.33	0.1830

Score Statistics for Type 3 GEE			
Source	DF	Chi-Square	P-value
TRT	1	4.31	0.0378
MONTH	2	16.05	0.0003
BASE_STATUS	2	68.99	<0.0001
SEX	1	0.73	0.3929
AGE	1	1.79	0.1806

GEE FIT CRITERIA: 1031.5143

Justification for the model:

By looking at the Score Statistics /Type 3 information, we see that the interaction variable TRT*MONTH is not significant. By dropping it from the model, we can see trt became more significant in the model, as well as base status. A limitation of this model, is that if interaction is actually significant it obviously doesn't account for it.

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.

Model: $\text{logit}(\theta) = \alpha_1 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + x_4\beta_4 + x_5\beta_5 + x_6\beta_6 + x_7\beta_7$

α_1 =intercept (treatment=placebo, sex=male,month=1,baseline=1)

x_1 =I (treatment=Active)

x_2 = I (month=3)

x_3 = I(month=5)

x_4 =I (base status=2)

x_5 =I (base status=3)

x_6 =I (sex=female)

x_7 =age

OR for the effect active trt has on improvement from baseline (relative to placebo)= 1.5053

95% CI: (1.0216,2.2181)

At each follow up visit, a patient on active treatment has 1.5053 times the odds of having an improved status verses a patient on placebo treatment.

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.

Statistic	Value	95% CI
Simple Kappa	0.2735	0.1530,0.3940
Weighted Kappa	0.2734	0.1242,0.4226

Based on the contingency table above the rows are the ratings of the patients rheumatoid arthritis determined by the physician, the columns are the ratings of the patients rheumatoid arthritis determined by the patient, and the diagonal is where the patient and the physician agree. If the kappa statistic=1 it implies the physician and patient are in agreement. The weighted kappa statistic takes into account the order of the severity classification. The simple kappa is equal to 0.2735 which according to our text book is indicative of a slight agreement. Since the confidence bounds do not contain the value 0, you can reject the hypothesis that the kappa is 0 for this data (no agreement) at the alpha 0.05 level of significance. The weighted and simple kappa's are basically the same so it implies that if you consider disagreement close to the diagonals less heavily than disagreement further away from the diagonals, you get the same agreement.

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	404 4057	354 4359
	Female	362 3331	369 6481
	Male	718 4412	631 4438
	Female	602 3538	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.

$$\log \left\{ \frac{\mu(x)}{N(x)} \right\} = x' \beta$$

$$\log\{\lambda_i\} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

Where:

$\mu(x)$ is the number of deaths

$N(x)$ is the total number of admissions

λ is the rate of incidence (death rate)

α = log death rate for males during the time period 1848-1850 that are >15 years old

$x_1 = I(\text{time period is 1846-1848})$

$x_2 = I(\text{sex is female})$

Name: Cheynna Crowley

BIOS 665
Final Exam 2017

$x_3 = I(\text{age is 2-15})$

Parameter	Estimate	Category
α	-2.0819	Males, >15 , 1848-1850
β_1	-0.5127	time period 1846-1848
β_2	-0.1464	Female
β_3	0.3644	Age 2-15

$$\log\{\lambda_i\} = 2.0819 - 0.5127x_1 - 0.1464x_2 + 0.3644x_3$$

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.

Saturated Model: $\log\{\lambda_i\} = \alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 + \beta_5x_5 + \beta_6x_6 + \beta_7x_7$

Log Likelihood: 21475.871

Main Effects Model: $\log\{\lambda_i\} = \alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_3$

Log Likelihood: 21452.4799

H0: Main effects model is of adequate fit

H1: o.t.w

→ Likelihood Ratio Test= $-2(\text{LogLikelihood(model1)} - \text{LogLikelihood(model2)}) \sim \chi^2_{df}$

df=difference in the number of parameters=4

$\chi^2_{4,0.05} = 9.488$

LRT= $-2(21452.4799 - 21475.871) = 46.7862 \sim \chi^2_4$

Since 46.7862 is greater than 9.488 we reject H0 and conclude that the main effects model is not an adequate fit.

Now to determine all best model, first start with the saturated model and then individually remove insignificant parameters then re-run.

Model: $\log\{\lambda_i\} = \alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 + \beta_5x_5 + \beta_6x_6 + \beta_7x_7$

$x_1 = I(\text{time period is 1846-1848})$

$x_2 = I(\text{sex is female})$

$x_3 = I(\text{age is 2-15})$

$x_4 = I(\text{time period is 1846-1848, sex is female})$

$x_5 = I(\text{time period is 1846-1848, age is 2-15})$

$x_6 = I(\text{sex is female, age is 2-15})$

$x_7 = I(\text{sex is female, age is 2-15, time period is 1846-1848})$

Parameter	Estimate	p-value
α	-1.9507	<0.0001
β_1	-0.5600	<0.0001
β_2	-0.3638	<0.0001
β_3	0.1350	0.0133
β_4	0.0087	0.9255
β_5	0.0689	0.4490
β_6	0.4084	<0.0001
β_7	0.0341	0.7934

→ Remove sex*time*age interaction and re-run

$$\text{Model: } \log\{\lambda_i\} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6$$

Parameter	Estimate	p-value
α	-1.9475	<0.0001
β_1	-0.5689	<0.0001
β_2	-0.3702	<0.0001
β_3	0.1291	0.0092
β_4	0.0262	0.6877
β_5	0.0855	0.1894
β_6	0.4209	<0.0001

→ Remove time*sex interaction and re-run

$$\text{Model: } \log\{\lambda_i\} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_6 x_6$$

Parameter	Estimate	p-value
α	-1.9639	<0.0001
β_1	-0.5235	<0.0001
β_2	-0.3694	<0.0001
β_3	0.1598	0.0003
β_6	0.4219	<0.0001

→ Remove time*age interaction and re-run

$$\text{Model: } \log\{\lambda_i\} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_6 x_6$$

Parameter	Estimate	SE	Wald Chi Sq	p-value
α	-1.9680	0.0341	3337.34	<0.0001
β_1	-0.5126	0.0325	248.15	<0.0001
β_2	-0.3610	0.0449	64.15	<0.0001
β_3	0.1600	0.0437	13.43	0.0002
β_6	0.4215	0.0628	45.08	<0.0001

→ Final Model: $\log\{\lambda_i\} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$

$$x_1 = I(\text{time period is 1846-1848})$$

$$x_2 = I(\text{sex is female})$$

$$x_3 = I(\text{age is 2-15})$$

$$x_4 = I(\text{sex is female, age is 2-15})$$

This is a good choice of model because all terms are significant.

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.

Time Period	Sex	Age		Death Rate per 100 admissions
1848-1850	Male	2-15	$e^{\alpha+\beta_3}$	17.95
1848-1850	Male	>15	e^{α}	12.47
1848-1850	Female	2-15	$e^{\alpha+\beta_3+\beta_2}$	15.51
1848-1850	Female	>15	$e^{\alpha+\beta_2}$	10.77
1846-1848	Male	2-15	$e^{\alpha+\beta_3+\beta_1}$	10.75
1846-1848	Male	>15	$e^{\alpha+\beta_1}$	7.47
1846-1848	Female	2-15	$e^{\alpha+\beta_3+\beta_2+\beta_1}$	9.29
1846-1848	Female	>15	$e^{\alpha+\beta_2+\beta_1}$	6.45

Regardless of time period and sex, the death rates are consistently lower for the higher age group (>15) compared to the lower age group (2-15).

Regardless of time period and age group, the death rates are consistently lower for the female group compared to the male group.

Regardless of sex and age group, the death rates are consistently lower for the earlier time period (1846-1848) compared to the later time period (1848-1850).

The group with the lowest death rate is Female, Age group >15, and time period 1846-1848. The group with the highest death rate is Male, Age group 2-15, and time period 1848-1850.