BIOS 665: Problem Set 1
Assigned: Tuesday, August 29, 2017
Due: Tuesday, September 12, 2017

Note: For all hypothesis tests, please state the method, the null hypothesis, the test statistic, the p-value, and the interpretation of the test using a significance level of 0.05, unless otherwise stated.

Helpful hints: For estimates and tests, simply copying and pasting SAS output without any commentary will not earn full credit, especially on exams. Highlighting is not considered as commentary. However, commentary can be as simple as: 'The 95% CI for the odds ratio is (___, ____).'

1. A randomized study of 160 subjects was conducted to evaluate whether a test treatment had a better outcome than placebo, in which the outcome was either a favorable or unfavorable response. Of the 80 subjects in the test treatment group, 45 had a favorable response, while out of the 80 subjects in the placebo group, 29 subjects had a favorable response.

    a. Construct a contingency table summarizing the association between treatment assignment and response.

    b. Calculate an estimate of the proportion of subjects with a favorable response in each of the treatment groups along with corresponding two-sided 95% confidence intervals.

    c. Calculate an estimate for the difference in proportions and a corresponding two-sided 95% confidence interval comparing test treatment versus placebo.

    d. Conduct a statistical test for whether an association exists between treatment group and response. Please also list two conditions that the data must meet for this test to be valid.

    e. Briefly discuss your results from parts (b.), (c.), and (d.).

2. A multi-center randomized trial was conducted for comparing two doses of a test drug for management of a chronic medical disorder. A contingency table of all subjects classified by the dose (high treatment dose, low treatment dose) and outcome (favorable response, unfavorable response) is given below.

Treatment	Favorable Response	Unfavorable Response
High Dose	45	14
Low Dose	22	33

a. Estimate the odds ratio for favorable response (vs. unfavorable) comparing high dose to low dose.

b. Provide a two-sided 95% confidence interval for this odds ratio.

c. Interpret (b.) in light of a statistical test for the association between dose group and response.

d. The contingency table from one of the centers is given below.

Treatment	Favorable Response	Unfavorable Response
High Dose	7	2
Low Dose	3	4

i. Calculate a corresponding two-sided 95% confidence interval for the odds ratio for favorable response (vs. unfavorable) comparing high dose to low dose for subjects in this center. Interpret the odds ratio estimate.

ii. Apply a two-sided statistical test. Justify your choice for the test.

The data shown below are from a study to evaluate the performance of a screening test for 130 patients for whom a particular disorder was known to be present and for 84 patients for whom it was known to be absent.

Disease	**Test+**	**Test-**	**Total**
Present	106	24	130
Absent	26	58	84

e. Estimate the sensitivity and the specificity for the screening test.

f. Provide a two-sided 99% confidence interval for sensitivity.

g. Provide a two-sided 95% confidence interval for specificity.

h. If 70% of the population has the disease and 30% does not, estimate the proportion of non-diseased patients among those who have a negative test.

3. A randomized study was undertaken to evaluate whether a new eye drop was effective at clearing proteins from the eye lens (*i.e.,* clearing proteins is the desired outcome). Each patient had the placebo drop placed in one eye that was chosen at random and the new treatment drop in the other eye. There were 260 patients enrolled with the results as follows:

	New Treatment	
Placebo	Clear	Not Clear
Clear	136	19
Not Clear	53	52

a. Use a statistical test to assess whether the probability of clearing proteins with the new drop is the same as the probability of clearing proteins with the placebo drop. Justify why you used this test.

b. Is this result good news or bad news for the treatment drop?

c. [For BIOS students and those trying for an H grade]: Create a two-sided 90% confidence interval for the difference in proportions of clearance between the new treatment and placebo.

4. Suppose a new clinical trial is being planned to confirm the findings of the study from Problem #2 (all patients). Suppose the expected probabilities of favorable response are 0.75 for high dose and 0.40 for low dose.

   a. With balanced allocation (i.e., equal sample size for each treatment group), determine the sample sizes needed to provide about 0.90 power at the two-sided 0.05 significance level for this study.

   b. With twice as many patients for high dose as for low dose, determine the sample sizes needed to provide about 0.80 power at the two-sided 0.01 significance level for this study with the same proportions as above.

   c. If at the end of the study there were 75 patients enrolled in each group, what is the power of the study if the proportions are the same as above, under equal allocation and a significance level of 0.05?

Name: ____________________________________ Department: ______________________

BIOS 665: Problem Set 2
Assigned: September 12, 2017
Due: September 28, 2017

Reminder: For all hypothesis tests, please state the method, the null hypothesis, the test statistic, the p-value, and the interpretation of the test using a two-sided significance level of 0.05, unless otherwise stated.

Helpful hints: For estimates and tests, simply copying and pasting SAS output without any commentary will not earn full credit, especially on exams. Highlighting is not considered commentary. However, commentary can be as simple as: The 95% CI for the odds ratio is (___, ____).

I have followed the Honor Code. Signed: ____

1. The following data arise from a health policy study that included interviews with subjects from both rural and urban geographic regions. Use logistic regression to describe the relationship of favorable opinion response (vs. unfavorable response) to stress and residence. Use reference groups of "Urban" for residence and "Low" for stress.

Residence	Stress	Favorable	Unfavorable	Total
Urban	Low	64	20	84
Urban	Medium	76	50	126
Urban	High	122	100	222
Total		262	170	432
Rural	Low	55	30	85
Rural	Medium	68	60	128
Rural	High	115	90	205
Total		238	180	418

a) State the assumptions for using logistic regression to model these data. Mathematically specify the variables used in the model, and interpret all parameters.

b) Provide a quantity that expresses the effect of high stress (as compared to low stress) on favorable response (vs. unfavorable response), and provide a 99% two-sided confidence interval for this quantity. How would you estimate this quantity and its 99% confidence interval by hand, given the computer output? Show your calculations.

c) Formally test the hypothesis that residence has no effect on opinion response.

d) Provide predicted probabilities for favorable response for each of the following:
   a. An individual from an urban area with low stress.
   b. An individual from a rural area with medium stress.

e) Test the hypothesis that the model fit is adequate. Briefly justify your choice for the test.

Name: ______________________________ Department: ____________________

2. The following results for two logistic regression models which were fit to describe the relationship between the probability of no headache pain (versus otherwise) at one hour post-treatment and explanatory variables for treatment (placebo, high dose, low dose), baseline severity (moderate, severe), and (for Model 2) their interaction are presented below. The results for the two models include estimated parameters, their standard errors, and the maximized log-likelihood for the fit of the model.

	Model 1		Model 2	
Parameter	Estimate	Standard Error	Estimate	Standard Error
Intercept	-1.642	0.395	-1.524	0.489
Severe (Baseline Severity)	-1.498	0.364	-1.913	1.168
High dose	2.236	0.512	1.900	0.567
Low dose	1.113	0.386	1.131	0.567
Severe*High	NA	NA	0.860	1.189
Severe*Low	NA	NA	0.061	1.218
Log-likelihood	-109.674		-107.897	

* Explanatory variables equal 1 if the category applies, and 0 if otherwise.
** NA = Not Applicable

a) For Model 1, state the relevant assumptions for the application of logistic regression to these data. Specify the mathematical structure of the model, including mathematical definitions for all explanatory variables.

b) Calculate the odds for each of the following conditions. Please provide numeric solutions.

Baseline Severity	**Dose**	**Model 1**	**Model 2**
Severe	Placebo		
Severe	Low		
Severe	High		
Moderate	High		

c) Numerically calculate the odds ratios of no headache (vs. otherwise) for each of the following conditions:

	Model 1	**Model 2**
Odds ratio for severe to moderate, for placebo		
Odds ratio for severe to moderate, for low dose		
Odds ratio for high dose to placebo, for moderate baseline		
Odds ratio for high dose to placebo, for severe baseline		

d) Comment on the calculations in c), particularly in regard to the comparisons between rows 1 and 2, and between rows 3 and 4.

e) How would you interpret the parameter corresponding to low dose in Model 1?

f) For Model 1, provide a 90% confidence interval and estimate for the odds ratio of no headache pain (vs. otherwise) for high dose versus placebo, controlling for baseline severity.

g) For Model 2, provide a predicted probability of no headache pain for an individual on low dose who had severe pain at baseline. Please provide a numeric solution.

h) Through the results for Models 1 and 2, apply a statistical test at the $\alpha = 0.05$ significance level to assess the hypothesis that Model 1 has satisfactory goodness of fit in the sense that any association between baseline severity and probability of no headache pain at one hour post-treatment is homogeneous across the placebo, low dose, and high dose groups.

Name: ____ Department: _________________________

3. Consider the following data from a study that evaluated the relationship between dose and dichotomous response concerning pain relief.

Dose	Favorable	Unfavorable	Sample size
1 mg	21	39	60
10 mg	24	36	60
100 mg	42	18	60

a) Use logistic regression to describe the relationship between favorable response (vs. unfavorable response) and dose.

i. Mathematically specify the model, treating dose as categorical, and interpret the parameters. Use dose=1 mg as the reference group.

ii. Now, re-specify the model mathematically, treating dose as continuous and using a log10 transformation. Interpret these model parameters.

~~iii. Provide estimates and 95% Fiducial Limits for the dose values corresponding to ED25, ED50, and ED75. In other words, provide estimates and 95% confidence limits for the dose values which produce a response with 0.25, 0.50, and 0.75 probabilities, respectively.~~

~~b) Use probit analysis to describe the relationship between favorable response (vs. unfavorable response) and dose.~~

~~i. Mathematically specify the model, treating dose as continuous using a log10 transformation. What types of interpretable quantities correspond to either the parameters or functions of the parameters?~~

~~ii. Provide estimates and 95% Fiducial Limits for the dose values corresponding to ED25, ED50, and ED75. In other words, provide estimates and 95% confidence limits for the dose values which produce a response with 0.25, 0.50, and 0.75 probabilities, respectively.~~

~~c) Briefly compare and contrast your results from part (a)(iii) and part (b)(ii).~~

BIOS 665: Problem Set 3
Assigned: September 26, 2017
Due: October 10, 2017

Reminder: For all hypothesis tests, please state the method, the null hypothesis, the test statistic, the degrees of freedom, the p-value, and the interpretation of the test using a two-sided significance level of 5%, unless otherwise stated.

Helpful hints: For estimates and tests, simply copying and pasting SAS output without any commentary will not earn full credit, especially on exams. Highlighting is not considered commentary. However, commentary can be as simple as: The 95% CI for the odds ratio is (_, _).

I have followed the Honor Code. Signed: ____

1. (25 points) Consider the following data from a randomized clinical trial to assess whether an experimental treatment has a safety concern. Investigators fear that the treatment might be associated with the adverse event of a severe headache after 24 hours of treatment (compared to Placebo). Subjects were randomized to either active treatment or placebo. The following table contains data on treatment (Active/Placebo), sex (Male/Female), and occurrence of a severe headache 24 hours.

Treatment	Sex	Occurrence of Severe Headache after 24 Hours		Total
		Yes	No	
Active	Male	22	34	56
	Female	19	37	56
Placebo	Male	37	18	55
	Female	37	20	57

a) (6 points) Under minimal assumptions, conduct a statistical test to assess the association of active treatment vs. placebo with occurrence of severe headache after 24 hours of treatment, controlling for sex. Hint: it may be helpful to write out the 2x2 table for males and the 2x2 table for females.

The null hypothesis is that pooled treatment is not associated with occurrence of headache after 24 hours, controlling for sex. The value of the Mantel-Haenszel test statistic is 19.3553. The test statistic is approximately chi-square with 1 degree of freedom. Since p<0.0001, which is less than 0.05, (or since 19.3553>3.84), reject the null hypothesis. Hence, there is evidence of an association between treatment and occurrence of headache after 24 hours, controlling for sex, with active treatment having significantly less occurrence of headache.

b) (6 points) Considering only females, provide an odds ratio for the effect of active treatment vs. placebo on occurrence of severe headache after 24 hours. Under minimal assumptions, conduct a statistical test to assess the association.

The odds ratio for the females is 0.2776. Females on active treatment had 0.2776 times the odds of having a headache than those on placebo. Alternatively, females on placebo had 1/0.2776 = 3.602 times the odds of having a headache than those on active treatment.

The null hypothesis is that active treatment is not associated with occurrence of headache after 24 hours for females. The Mantel-Haenszel chi-square statistic is 10.7519. The test statistic is approximately chi-square with 1 degree of freedom. Since p = 0.0010, which is less than 0.05, (or since 10.7519 > 3.84), we reject the null hypothesis. Therefore, there is evidence of an association between treatment and occurrence of headache after 24 hours for females, with active treatment having significantly less occurrence of headache.

NOTE: could also refer to CI for OR.

c) (6 points) Under minimal assumptions, provide an odds ratio and a 95% confidence interval for the effect of active treatment vs. placebo on the occurrence of severe headache after 24 hours, controlling for sex. You should assume that the effect of active treatment on presence of rash after 24 hours is the same in both males and females. Do parts a) and c) agree?

The common odds ratio is OR = 0.2956. On average, those with treatment had 0.2956 times the odds of having a severe headache than those with placebo (or those on placebo had 1/0.2956 = 3.383 times the odds of having a severe headache than those on active treatment). The 95% CI is (0.1707, 0.5118). This interval does not include 1, indicating a statistically significant association. This agrees with the results of part a.

d) (7 points) Provide statistical evidence for (or against) the hypothesis that the effect of active treatment on the occurrence of a severe headache after 24 hours is the same for each sex. Provide a sentence explaining your results.

The null hypothesis for this test is that the odds ratios are homogeneous across strata. The Breslow-Day test gives a chi-square value of 0.0504 and a p-value of 0.8224. We fail to reject the null hypothesis and conclude that the odds ratios for males and females can reasonably be considered to be homogeneous.

2. A company is conducting market research on a newly formulated sports drink. The table shown below summarizes the findings from a study comparing male and female athletes with respect to the degree of favorable opinion about this new sports drink.

Gender	Response			Total
	Unfavorable opinion about new sports drink	Neutral opinion about new sports drink	Favorable opinion about new sports drink	
Female	15	25	60	100
Male	15	30	55	100
Total	30	55	115	200

a) (7 points) Under minimal assumptions, assess the association between gender and (ordinal) degree of favorable opinion with an appropriate statistical test at the two-sided 0.05 level, assigning integer scores to the response categories. Briefly interpret your results in 1-2 sentences; be sure to address whether it would be reasonable for the company to market preferentially to males or to females.

The null hypothesis is that there is no association between gender and degree of favorable opinion. We obtained a test statistic of 0.229, 1 df and a p-value 0.633. We failed to reject the null hypothesis and conclude that there is no association between gender and favorable opinion in terms of a location shift. Therefore, there is no evidence that the company should market preferentially to either males or females.

b) (6 points) Repeat (a), but assign rank scores for the response categories.

The null hypothesis is that there is no association between gender and degree of favorable opinion. We obtained a test statistic of 0.343, 1 df and a p-value 0.558. We failed to reject the null hypothesis and conclude that there is no association between gender and favorable opinion. Therefore, there is no evidence that would lead the company to market preferentially to either males or females.

c) (6 points) Repeat (a), but assign modified ridit scores for the response categories.

The null hypothesis is that there is association between gender and degree of favorable opinion. We obtained a test statistic of 0.343, 1 df and a p-value 0.558. We failed to reject the null hypothesis and conclude that there is no association between gender and favorable opinion. Again, there is no evidence to lead the company to market preferentially to either males or females.

d) (6 points) Briefly compare your results across parts (a), (b), and (c). Comment on any noteworthy differences, as well as any similarities (or equivalencies).

We observed that the estimates are very similar across parts a, b and c. Part c is exactly the same as part b because the modified ridit scores are a scaled factor of the rank scores.

3. Consider the data in table below. These data are from a clinical trial conducted in two centers for the comparison of two treatments for a gastrointestinal disorder with respect to a dichotomous response (Good vs. Poor).

		Response		
		Good	Poor	Total
Center 1	Test Treatment	32	11	43
	Placebo	23	20	43
Center 2	Test Treatment	29	5	34
	Placebo	15	17	32

a) (9 points) For Center 1, provide an estimate for the odds ratio (and its 95% confidence interval) describing the relationship between test treatment and placebo for good versus poor response. Repeat for Center 2. Briefly justify your methods, and interpret the results.

Center 1:

The estimated OR is 2.530 and its 95% CI is (1.018, 6.285). The test treatment is significantly better than the placebo at the 0.05 significance level because the CI does not contain 1. These methods arewell justified as each cell count is at least 10.

Center 2:

The estimated OR is 6.573 and its 95% CI is (2.028, 21.305). The test treatment is significantly better than the placebo at the 0.05 significance level because the CI does not contain 1. These methods are well justified as each cell count is at least 5.

b) Provide a (common) odds ratio and a 95% confidence interval describing the relationship between test treatment and placebo for good versus poor response, controlling for center. You may assume that the effect of treatment on response is homogeneous across centers.

The estimated OR is 3.6565 and its 95% CI is (1.7984, 7.4341). The test treatment is significantly better than the placebo, controlling for center, at the 0.05 significance level because the CI does not contain 1. These methods are well justified as each cell count is at least 5.

c) (9 points) For Center 1, provide and interpret the results of a statistical test for the association between treatment and response using the two-sided 0.05 significance level. Repeat for Center 2.

Center 1:

We hypothesized that there is no association between treatment and response. We obtained a 1 df test with chi square statistic of 4.038 with p-value 0.0445<0.05. We reject the null hypothesis and conclude that there is an association between treatment and response for center 1, with test treatment significantly better than placebo.

Center 2:

We hypothesized that there is no association between treatment and response. We obtained a 1 df test with chi square statistic of 10.784 with p-value 0.001<0.05. We reject the null hypothesis and conclude that there is an association between treatment and response for center 2, with test treatment significantly better than placebo.

Note: Fisher's exact test results are also acceptable for both parts.

d) (7 points) Under minimal assumptions, assess the association between treatment and response, controlling for center, with a statistical test at the two-sided 0.05 level. Briefly justify your methods, and interpret the results.

Controlling for center:

We hypothesized that there is no association between treatment and response. We obtained a 1 df test with chi square statistic of 13.437 with p-value 0.0002<0.05. We reject the null hypothesis and conclude that there is an association between treatment and response controlling for center, with test treatment significantly better than placebo.

This is consistent with parts a and b. Notice that the Mantel-Fleiss criterion is met, and so it is appropriate to use the CMH estimates.

e) Briefly compare and contrast your center-specific results to your overall results (controlling for center), i.e., compare/contrast a) with b) & compare/contrast c) with d).

Since the association between treatment and response are in the same direction and pattern for each center, the statistic results from c and d are consistent;

The overall OR (controlling for center) is a function of weighted average of differences in the mean scores of the two treatments for center strata. The center 1, center 2, and overall ORs were all in the same direction and the CI's didn't contain 1 and the p-values were all <0.05, so all results aligned.

Name: Department:

BIOS 665: Problem Set 4
Assigned: October 26, 2017
Due: November 9, 2017

Reminder: For all hypothesis tests, please state the method, the null hypothesis, the test statistic, the degrees of freedom, the p-value, and the interpretation of the test using a two-sided significance level of 5%, unless otherwise stated.

Helpful hints: For estimates and tests, simply copying and pasting SAS output without any commentary will not earn full credit, especially on exams. Highlighting is not considered commentary. However, commentary can be as simple as: The 95% CI for the odds ratio is (__, __).

I have followed the Honor Code. Signed: ____________________________________

1. A social scientist is interested in how region and education level are associated with people's opinions regarding physical activity. She conducts a survey on randomly selected individuals, asking them their geographic region (West Coast, Midwest, and East Coast) and education level (college graduate, high school graduate, less than high school). She also asks how strongly they agree with the following statement: "U.S. employers should offer greater incentives to have physically active employees." Participants could choose one of these options: "Disagree," "Neutral," or "Agree." Where applicable, let "less than high school" education level and "West Coast" region be the reference groups.

		Level of agreement with statement regarding physical activity		
Education Level	Region	Disagree	Neutral	Agree
College Graduate	West Coast	18	15	48
	Midwest	13	19	21
	East Coast	28	28	52
High School Graduate	West Coast	46	23	24
	Midwest	22	20	21
	East Coast	48	18	23
Less than high school	West Coast	13	15	28
	Midwest	15	16	17
	East Coast	15	17	24

a) Mathematically specify a proportional odds regression model for level of agreement (ordered from more agreement to less agreement) with main effects for region and education level (using all 3 levels). State assumptions and interpret all model parameters. Then, assess goodness of fit of the proportional odds model, being sure to justify these methods.

b) Conduct a test to assess whether proportional odds is a reasonable assumption for these data. State your null hypothesis, the p-value of your test, the criteria for making the decision of your test, and write a sentence explaining the results of your test. Then,

   i) test whether education level has an effect on agreement at the two-sided 0.05 significance level. State your null hypothesis, the p-value of your test, the criterion for making the decision of your test, and write a sentence explaining the results of your test.

   ii) provide an estimate and 95% confidence interval for the odds ratio of agree vs. (neutral or disagree) comparing college graduates with those having less than high school education. What do you conclude about the statistical significance of this effect from the confidence interval? Briefly discuss how this estimate compares to a comparable estimate for (agree or neutral) vs. disagree.

2. Consider the data from Question 1 above.

   a) Mathematically specify and fit a generalized logits regression model for level of agreement, treating the level of agreement as nominal. Include main effects for education level and region. Let "neutral" be your reference level for the outcome variable. State assumptions, and interpret all model parameters.

   b) Use the model from Part 2.a. to test whether education level has an effect on agreement at the two-sided 0.05 significance level. State your null hypothesis, the p-value of your test, the criterion for making the decision of your test, and write a sentence explaining the results of your test.

   c) Using this model, provide an estimate and 95% confidence interval for the odds ratio of agree vs. neutral comparing college graduates with those having less than high school education. Repeat for disagree vs. neutral, as well as for agree vs. disagree. What do you conclude about the statistical significance of each effect from these confidence intervals? *Hint: You might invoke a separate PROC LOGISTIC by changing the reference level to obtain the odds ratio for agree vs disagree.*

3. The following data are from a study to compare two treatments with respect to the relationship between dose and dichotomous response concerning pain relief.

Treatment	Dose	Favorable	Unfavorable	Sample Size
A	1 mg	21	39	60
A	10 mg	24	36	60
A	100 mg	42	18	60
B	2 mg	23	37	60
B	20 mg	31	29	60
B	200 mg	50	10	60

a) For Treatment A only, use logistic regression to describe the relationship between favorable response and log10(dose). Note: log10(.) denotes the log base 10 transformation. *Hint: recall Problem Set #2, Problem 3.*

i) State assumptions, and mathematically specify the model.

ii) Provide estimates and 95% confidence limits (or fiducial limits) for the doses at which each of 25%, 50%, and 75% favorable response are predicted by the model (i.e. ED25, ED50, and ED75, respectively).

iii) Use a probit analysis to calculate all the estimates and confidence intervals requested in Part 3.a.ii. How might your assumptions change when using a probit model vs a logistic model?

iv) Briefly compare and contrast your results from Part 3.a.ii. and Part 3.a.iii.

b) For the data from both Treatment A and Treatment B, use logistic regression to describe the relationship between favorable response and log10(dose) for the data, allowing for separate effects for each treatment group, as illustrated in class.

i) State assumptions, and mathematically specify the model.

ii) Evaluate goodness of fit of the model.

iii) Provide a point estimate and its 95% confidence interval for the relative potency of Treatment B relative to Treatment A.

4. The table shown below displays the cross-classification of maternal age groups (in years) and the number of births with a particular disorder in a specific geographic area during a specific time period, as well as the corresponding numbers of all births. When necessary, use ‘20-24 years’ as the reference group for maternal age group and ‘1’ as the reference group for birth order.

Birth Order	1	2	3	1	2	3
Maternal Age	Number of Births with Disorder			Total Number of Births		
20-24	128	152	71	329,462	326,735	175,682
25-29	54	112	101	114,987	208,692	207,060
30-34	41	79	109	39,473	83,224	117,312
35-39	38	89	99	14,202	28,478	45,015
40+	22	44	83	3,046	5,381	8,654

a) Specify the mathematical structure of a statistical model to describe the variation in the rates of the disorder per 100,000 live births with respect to maternal age group and birth order.

b) Interpret the estimated parameters of this model, and provide appropriate two-sided 95% confidence intervals for those pertaining to birth order.

c) Use the model from Part 4.a. to obtain predicted values for the rates of the birth disorder for the respective birth order subpopulations corresponding to ‘30-34 years’ for maternal age group.

BIOS 665: Problem Set 4; **Solution Sketch for Problem 3**
Assigned: October 24, 2017
Due: November 7, 2017

3. The following data are from a study to compare two treatments with respect to the relationship between dose and dichotomous response concerning pain relief.

Treatment	Dose	Favorable	Unfavorable	Sample Size
A	1 mg	21	39	60
A	10 mg	24	36	60
A	100 mg	42	18	60
B	2 mg	23	37	60
B	20 mg	31	29	60
B	200 mg	50	10	60

a) For Treatment A only, use logistic regression to describe the relationship between favorable response and log10(dose). Note: log10(.) means log transformation with base 10.

i. State assumptions and specify the model.

(1) responses of subjects determined through a tolerance distribution
(2) tolerances follow a logistic distribution
(3) observations in the data set are independent
(4) at least five instances of favorable response and five instances of unfavorable response for each dose (in group A)
(5) sampling compatible with a simple random sample
(6) model fits the data adequately

The model is: logit(p_i) = $\alpha + \beta x_i$ where

p_i= probability of favorable response at level x_i
x_i= log10(dose)

Note that in SAS, *log* is the natural log (base *e*). *log10* is the transformation for the base-10 logarithm.

Also note that the probability of *favorable* response is being modeled here.

ii. Provide 95% Fiducial Limits for the doses at which each of 25%, 50% and 75% favorable response are predicted by the model (i.e. ED25, ED50, and ED75, respectively).

For ED25, let $\mathbf{x_{25}} = \log_{10}(ED25)$ and $\mathbf{p_{25}}$=probability of response at upper quartile of tolerance distribution. Then:

$$\ln\left\{\frac{p_{25}}{1-p_{25}}\right\} = \ln\left\{\frac{0.25}{1-0.25}\right\} = -\ln 3 = \hat{\alpha} + \hat{\beta}x_{25}$$

We can obtain $x_{25} = \frac{-ln3-\hat{\alpha}}{\hat{\beta}}$.

Similarly, for ED50, $x_{50} = \frac{-\hat{\alpha}}{\hat{\beta}}$.

For ED75, $x_{75} = \frac{ln3-\hat{\alpha}}{\hat{\beta}}$.

You can calculate these point estimates from PROC LOGISTIC. Confidence intervals for x_{50} can be calculated with the PROC IML code, or with the Taylor Series linearization formula provided in the notes. However, the variance for x_{25}, x_{75}, or any other percentile, would require adjustment to the given formulas. Please note that the correct Taylor Series approximations for estimating the variance of x_{25} and x_{75} were provided in the supplemental materials for this chapter on Sakai.

Alternatively, the point estimates and confidence intervals can be calculated in SAS directly from PROC PROBIT when specifying the *logistic* distribution. (PROBIT is the name of the PROCEDURE, a program that SAS has written. However, it has the flexibility to fit not only *probit models* but also *logistic models*). This syntax was provided in the notes and is shown below.

Also, please note that LD50 is not the log of ED50. ED50 is "The dosage that is effective for, or produces a response in, 50% of individuals." In the specific case when the response variable is death, then we call this quantity by the name of LD50, for "the dosage that is lethal for 50% of the individuals".

```
proc probit data=pain2 log10;
   model favor/total = dose / dist=logistic lackfit inversecl(prob=.25 .50 .75);
run;
```

Analysis of Maximum Likelihood Estimates

Parameter	*DF*	*Estimate*	*Standard Error*	*Wald Chi-Square*	*Pr > ChiSq*
Intercept	1	-0.8042	0.2520	10.1834	0.0014
ldose	1	0.7316	0.1950	14.0704	0.0002

~~*Estimated Covariance Matrix*~~

~~*Parameter*~~	~~*Intercept*~~	~~*ldose*~~
~~*Intercept*~~	~~0.063513~~	~~-0.01679~~
~~*ldose*~~	~~-0.01679~~	~~0.007176~~

Commented [STA1]: This is the Cov matrix when using ln(dose).

Estimated Covariance Matrix

Parameter	**Intercept**	**ldose**
Intercept	0.063513	-0.03866
ldose	-0.03866	0.038044

Probability	**Log10(dose)**	**95% Fiducial Limits**	
0.25	-0.4020	-2.0644	0.1959
0.50	1.0993	0.6379	1.6229
0.75	2.6005	1.9563	4.4339

Probability	**dose**	**95% Fiducial Limits**	
0.25	0.39630	0.00862	1.56994
0.50	12.56763	4.34417	41.97031

```
Probability        dose  95% Fiducial Limits

       0.75  398.54498   90.42638       27160
```

iii. Use a probit analysis to calculate all the estimates and confidence intervals as part (a)(ii). Compare and contrast your results from part (a)(iii) and part (a)(ii).

Note: PROC PROBIT will fit a *probit model* by default. This was the desired model for this part of the question. A *probit analysis* is fitting a *probit model*, which is a model that uses a *probit link function*. Fitting a *logistic model* in PROC PROBIT is not considered a *probit analysis*, as it uses the *logit link function*.

```
proc probit data=pain2 log10;
                  model favor/total = dose /    lackfit
                  inversecl(prob=.25 .50.75);
run;
```

```
             Probit Analysis on Log10(dose)

Probability  Log10(dose)  95% Fiducial Limits

       0.25      -0.3929     -1.9746     0.1928

       0.50       1.0941      0.6409     1.6142

       0.75       2.5811      1.9441     4.3479


Probability        dose  95% Fiducial Limits

       0.25     0.40468    0.01060     1.55893

       0.50    12.41894    4.37454    41.13011

       0.75   381.11672   87.91580       22281
```

We can see that although their assumed underlying distributions are different, they provide similar estimates and confidence limits for ED25, ED50, and ED75, respectively.

b) For the data from both Treatment A and Treatment B, use logistic regression to describe the relationship between favorable response and log(dose) for the data, accounting for separate effects for each treatment group, as illustrated in class.

i. State assumptions and specify the model.

(1) responses of subjects are determined through a tolerance distribution
(2) tolerances follow a logistic distribution
(3) observations in the data set are independent
(4) at least five instances of favorable response and five instances of unfavorable response for each dose and in each group
(5) sampling compatible with a simple random sample
(6) model fits the data adequately
(7) dilution assumption: doses of drug A (z_{Ai}) and doses of drug B (z_{Bi}) are related by the relative potency ρ. In other words, $z_{Bi}/z_{Ai}= \rho$.

The model is: logit(p_{Ai}) =$\alpha_A+\beta_A x_{Ai}$ and logit(p_{Bi}) =$\alpha_B+\beta_B x_{Bi}$ where

p_{Ai}= probability of favorable response at level x_{Ai} for drug A
p_{Bi}= probability of favorable response at level x_{Bi} for drug B
x_{Ai}= log10(z_{Ai})
x_{Bi}= log10(z_{Bi})

$\alpha_{Ai}=\alpha_{Bi}+\beta$log10($\rho$)
=log odds of favorable response on drug A when log10(dose)=0 (i.e., dose = 1mg)

α_{Bi}=log odds of favorable response on drug B when log10(dose)=0

β_A=increment in log odds of favorable response (or log odds ratio) for every 1 unit increase in log10(dose) of drug A (i.e., 10 fold increase in dose)

β_B =increment in log odds of favorable response for every 1 unit increase in log10(dose) of drug B

Note again that you should use the *log10* syntax in SAS for a base-10 logarithm.

Also, note that the modeled probability is that of *favorable* response.

ii. Evaluate goodness of fit of the model.

This involves 3 steps:

1. Check necessity of the quadratic terms via model selection

```
proc logistic data=pain descending;
        freq count;
        model response= int_a int_b ldose_a ldose_b sqldose_a sqldose_b /
                noint scale=none aggregate start=4 selection=forward details;
        eq_slope: test ldose_a=ldose_b;
  run;
```

Analysis of Effects Eligible for Entry

Effect	*DF*	*Score Chi-Square*	*Pr > ChiSq*
sq_ldose_A	1	2.5278	**0.1119**
sq_ldose_B	1	2.2055	**0.1375**

2. test equality of the slopes

Linear Hypotheses Testing Results

Label	*Wald Chi-Square*	*DF*	*Pr > ChiSq*
eq_slope	**0.8774**	**1**	**0.3489**

3. fit the final model with separate intercepts but common slope

```
proc logistic data=pain outest=estimate (drop=intercept _link_ _lnlike_)
covout;
        freq count;
        model response= int_a int_b dose /noint scale=none aggregate covb;
run;
```

Analysis of Maximum Likelihood Estimates

Parameter	*DF*	*Estimate*	*Standard Error*	*Wald Chi-Square*	*Pr > Chi Sq*
int_A	1	-0.9351	0.2145	19.0095	<.0001
int_B	1	-0.7672	0.2351	10.6527	0.0011
ldose	1	0.8602	0.1423	36.5571	<.0001

iii. Provide a point estimate and its 95% confidence interval for the relative potency of Treatment B relative to Treatment A.

The log relative potency is estimated as

$$\boldsymbol{log_{10}\hat{p} = \frac{\hat{\alpha}_A - \hat{\alpha}_B}{\hat{\beta}} = \frac{-0.9351 - (-0.7672)}{0.8602} = -0.1951}$$

use the PROC IML code provided on Sakai to implement Fieller's Theorem:

Estimate	Log10(Value)	95% CI for Log10(Value)	Value	95% CI for Value
Potency	-0.1951	(-0.7998, 0.3183)	0.6381	(0.1586, 2.0811)
$LD50_A$	1.0871	(0.7137, 1.4765)	12.2199	(5.1720, 29.9543)
$LD50_B$	0.8919	(0.4526, 1.2560)	7.7974	(2.8356, 18.0293)

The meaning for the relative potency is that a dose of treatment B must be 0.6381 time as a dose of treatment A to have the same effect. A 95% confidence interval of the relative potency is (0.1586, 2.0811), which contains 1.

Name: ____ Dept:____________________________

BIOS 665: Problem Set 5
Assigned: November 9, 2017
Due: November 28, 2017

Reminder: For all hypothesis tests, please state the method, the null hypothesis, the test statistic, the degrees of freedom, the p-value, and the interpretation of the test using a two-sided significance level of 5%, unless otherwise stated.

Helpful hints: For estimates and tests, simply copying and pasting SAS output without any commentary will not earn full credit, especially on exams. Highlighting is not considered commentary. However, commentary can be as simple as: The 95% CI for the odds ratio is (_, _).

I have followed the Honor Code. Signed: ____

1. The contingency table shown below displays data from a randomized study to compare an existing dental treatment to an experimental dental treatment for the prevention of dental caries (cavities). An event was defined as the first occurrence of at least one cavity detected at a follow-up visit (such that a participant could only have experienced the event once during the study). Following treatment, there were biannual follow-ups lasting for a period of 2 years.

		Follow-up period for dental caries (months)				Follow-up period for withdrawal (months)				
Treatment	No caries by 2 years	0-6	6-12	12-18	18-24	0-6	6-12	12-18	18-24	Total
Existing	39	2	5	14	35	2	2	7	10	116
Experimental	60	1	2	8	17	2	1	10	15	116

a) Present the data in life table format separately for each treatment group.

	Treatment= Existing			
Interval	No Caries	Dental Caries	Withdrawal	At Risk
0-6	112	2	2	116
6-12	105	5	2	112
12-18	84	14	7	105
18-24	39	35	10	84

Name: ____ Dept:____________________________

	Treatment= Experimental			
Interval	No Caries	Dental Caries	Withdrawal	At Risk
0-6	113	1	2	116
6-12	110	2	1	113
12-18	92	8	10	110
18-24	60	17	15	92

b) Provide life table estimates for the cumulative probabilities (and standard errors) of no occurrence of a cavity by the end of each of the four periods for each treatment group. State the assumptions for these estimates, and assume subjects who withdraw may be treated as not having caries at the time of withdrawal.

Assumptions:

- The first observation for each interval/stratification level includes subjects who withdrew and the second level will include subjects who has the event
- Patients who experienced the event (cavities) are not censored
- Withdrawal is independent of condition being studied
- Multiple withdrawals occur uniformly throughout the interval

Interval	Estimated Survival Rates	Standard Errors
Treatment= Existing		
0-6	0.9826	0.0122
6-12	0.9383	0.0226
12-18	0.8089	0.0376
18-24	0.4505	0.0498
Treatment=Experimental		
0-6	0.9913	0.00866
6-12	0.9737	0.0150
12-18	0.8995	0.0288
18-24	0.7185	0.0455

c) Which treatment has a more favorable outcome? Provide a statistical test at the two-sided 0.05 significance level.

Ho: The distribution of time to cavity is the same for existing and experimental treatment.

H1: Otherwise

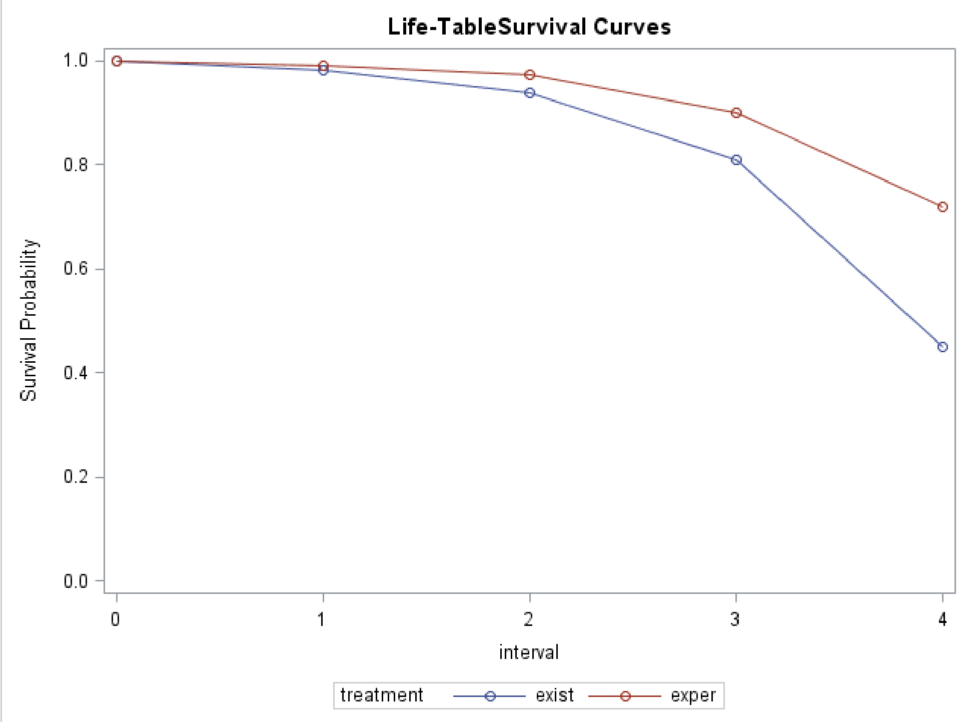
Mantel-Cox Test

DF=1

Q_{MC}= 13.4851

P-Value=0.0002

Name: ____ Dept:________________________



Conclusion:

The p-value is less than alpha of 0.05 so we reject Ho.

There is a difference in survival between existence and experimental treatment.

The experimental treatment has the more favorable outcome.

2. Fit a piecewise exponential model to the data presented in Problem 1 in order to describe the relationship of time to occurrence of dental caries with regard to the main effects for treatment and follow-up period, as well as considering their interaction. When necessary, use 0-6 months as the reference group for follow-up period, and use the Existing treatment as the reference group for treatment.

	Treatment= Existing			
Interval	No Caries	Dental Caries	Withdrawal	Total person month exposure
0-6	112	2	2	6(112+0.5*2+0.5*2)=684
6-12	105	5	2	6(105+5*0.5+0.5*2)=651
12-18	84	14	7	6(84+0.5*14+0.5*7)=567
18-24	39	35	10	6(39+35*0.5+10*0.5)=369

	Treatment= Experimental			
Interval	No Caries	Dental Caries	Withdrawal	Total person month exposure

0-6	113	1	2	6(113+0.5*1+0.5*2)=687
6-12	110	2	1	6(110+0.5*2+0.5*1)=669
12-18	92	8	10	6(92+8*0.5+10*0.5)=606
18-24	60	17	15	6(60+0.5*17+0.5*15)=456

a) Mathematically specify the structure of the model, including the interaction, defining all variables used in the model. You do not need to interpret model parameters.

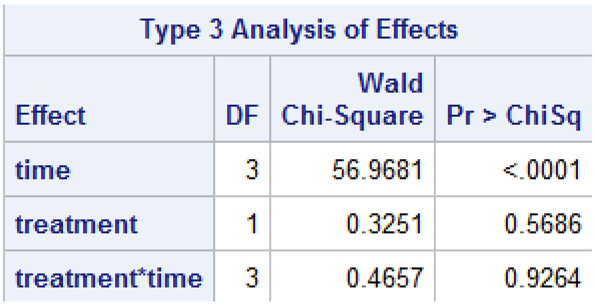
$\text{Logit}(\frac{\lambda}{n})= \alpha+X_1\beta_1+X_2\beta_2+X_3\beta_3+X_4\beta_4+X_5\beta_5+X_6\beta_6+X_7\beta_7$

Parameter	Value	Definition
α	-5.8319	Intercept: existing treatment and 0-6 month follow up
β_1	0.9705	6-12 month follow up
β_2	2.1556	12-18 month follow up
β_3	3.5761	18-24 month follow up
β_4	-0.6990	Experimental treatment
β_5	-0.2493	Interaction for time 6-12 and experimental treatment
β_6	0.0611	Interaction for time 12-18 and experimental treatment
β_7	-0.2965	Interaction for time 18-24 and experimental treatment

*with λ is the number of caries and n is the person month exposure

b) Fit the model you specified in (a). Evaluate goodness of fit of the main effects model by assessing whether there is evidence of interaction between treatment and follow-up period. State your null hypothesis in terms of parameters from the model in (a), the value of your test statistic, the approximate distribution of your test statistic, your criteria for making your decision, and the conclusion from your test.

H0: $\beta_5 = \beta_6 = \beta_7 = 0$ H1=otherwise



Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
time	3	56.9681	<.0001
treatment	1	0.3251	0.5686
treatment*time	3	0.4657	0.9264

Wald Chi-Square ~ χ_3^2

Wald Chi-Square=0.4657

p-value=0.9264

Since the p-value is greater than alpha=0.05, we fail to reject H0 and conclude that the time*treatment interaction is not significant.

c) Regardless of your conclusion from (b), fit the model with main effects for treatment and follow-up period, but not including their interaction. Interpret the estimated model parameters. Provide a 95% confidence interval for the model parameter (or an appropriate transformation of this parameter) corresponding to the treatment variable.

Logit($\frac{\lambda}{n}$)= $\alpha+X_1\beta_1+X_2\beta_2+X_3\beta_3+X_4\beta_4$

Parameter	Definition	Value	Interpretation
α	Intercept: existing treatment and 0-6 month follow up	-5.7734	Log incidence density for existent treatment and 0-6 month follow up
β_1	6-12 month follow up	0.8933	Increment for 6-12 month follow up
β_2	12-18 month follow up	2.1808	Increment for 12-18 month follow up
β_3	18-24 month follow up	3.4785	Increment for 18-24 month follow up
β_4	Experimental treatment	-0.8845	Increment for treatment experimental

Odds Ratio	Estimate	95% CI
Treatment(experimental vs existing)	0.413	(0.260,0.657)

d) Provide model-predicted values using the model in (c) for cumulative probabilities of no occurrence of a cavity by 6 months, 12 months, 18 months, and 24 months, respectively, for each treatment.

Cumulative probabilities of no occurrence of a cavity		
Time	Estimated within failure	Estimated Survival
6 months, Experimental treatment	$e^{(-5.7734-0.8845)}$=0.0013	1$e^{(-0.0013*6)}$=0.992
12 months, Experimental treatment	$e^{(-5.7734-0.8845+0.8933)}$=0.0031	0.992*$e^{(-0.0031*6)}$=0.9737
18 months, Experimental treatment	$e^{(-5.7734-0.8845+2.1808)}$=0.01136	0.9737*$e^{(-0.01136*6)}$=0.9095
24 months, Experimental treatment	$e^{(-5.7734-0.8845+3.4785)}$=0.0416	0.90995*$e^{(-0.0416*6)}$=0.7089
6 months, Existence treatment	$e^{(-5.7734)}$=0.00311	1$e^{(-0.00311*6)}$=0.9815
12 months, Existence treatment	$e^{(-5.7734+0.8933)}$=0.0076	0.9815$e^{(-0.0076*6)}$=0.9377
18 months, Existence treatment	$e^{(-5.7734+2.1808)}$=0.0275	0.9377$e^{(-0.0275*6)}$=0.7950
24 months ,Existence treatment	$e^{(-5.7734+3.4785)}$=0.101	0.7950$e^{(-0.101*6)}$=0.434

Name: ____ Dept:___________________________

3. For this problem, you will be analyzing a dataset for a randomized, controlled trial among women of childbearing age to evaluate the longitudinal effects of an educational intervention. The primary response variable is the participants' self-rating of health as either "good" or "poor". The researchers would like to assess the effect of the intervention on self-rated health across the follow-up period.

REPEATED.sas7bdat contains data on $n = 80$ women enrolled in this trial. These data were measured at 4 points in time: at the time of randomization, then 3, 6, and 12 months post-randomization.

Each observation in the dataset contains values for the following variables:

- ID: unique participant identification code

- TIME: the visit number for this observation of this participant
  - 2 corresponds to the 3 month post-randomization visit
  - 3 corresponds to the 6 month post-randomization visit
  - 4 corresponds to the 12 month post-randomization visit

- RX: the group to which the participant has been randomized
  - control
  - intervention

- HEALTH: participant's self-rated level of health for this visit
  - Good
  - Poor

- AGE_GROUP: participant's age group at time of randomization
  - 15-24 (years old)
  - 25 to 34 (years old)
  - 35+ (years old)

- BASE: participant's self-rated level of health at randomization
  - Good
  - Poor

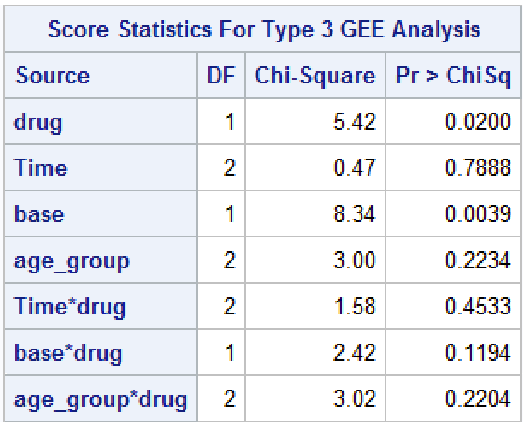
a) Fit a GEE repeated measures logistic regression model across all study follow-up visits (but not including time of randomization) to describe the marginal relationship of the log odds of participants' self-rating of good health to the main effects of randomized group, visit (as a class variable), health self-rating at time of randomization, and age group as explanatory variables. Use 15-24 year old women in the control group with a good health assessment at randomization as your reference group, and use 3 months post-randomization as your reference visit. Assume an exchangeable working correlation structure. Present a table of all parameter estimates, their standard errors, the score statistics, and the corresponding p-values.

Parameter	Estimate	Standard Error	ZScore Statistics	p-value
Intercept*	0.3809	0.4101	0.93	0.3530
Drug 0	1.9493	0.4662	4.18	<0.0001
Time 3	-0.0756	0.3090	0.24	0.8066
Time 4	-0.2234	0.3199	-0.70	0.4849
Base Poor	-1.6875	0.4475	-3.77	0.0002
Age 25-34	0.9629	0.4400	2.19	0.0286
Age 35+	1.0800	0.6632	1.63	0.1034

*reference group drug=1,time=2,base=good, age_group=15-24

b) Assess goodness of fit through consideration of all pairwise interactions with intervention (only), using manual backward selection to eliminate interactions (one-at-a-time) that are not significant at the 0.05 level. If any interactions are significant at the 0.05 level, include them and re-fit the model. Present and justify your choice for the final model.

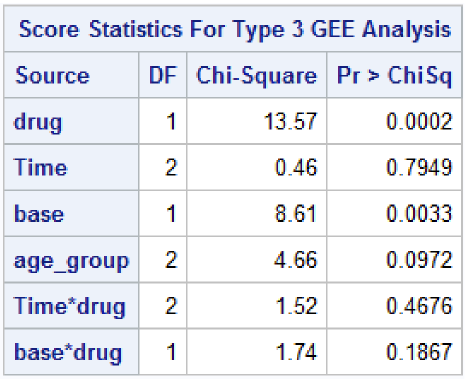
Trial 1 Model: health =drug time base age_group drug*time drug*base drug*age_group



Score Statistics For Type 3 GEE Analysis			
Source	DF	Chi-Square	Pr > ChiSq
drug	1	5.42	0.0200
Time	2	0.47	0.7888
base	1	8.34	0.0039
age_group	2	3.00	0.2234
Time*drug	2	1.58	0.4533
base*drug	1	2.42	0.1194
age_group*drug	2	3.02	0.2204

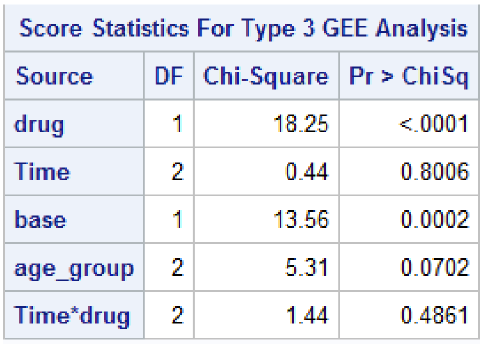
Trial 1 Conclusion: No significant interaction effects

Trial 2 Model: health =drug time base age_group drug*time drug*base



Score Statistics For Type 3 GEE Analysis			
Source	DF	Chi-Square	Pr > ChiSq
drug	1	13.57	0.0002
Time	2	0.46	0.7949
base	1	8.61	0.0033
age_group	2	4.66	0.0972
Time*drug	2	1.52	0.4676
base*drug	1	1.74	0.1867

Trial 2 Conclusion: No significant interaction effects
Trial 3 Model: health =drug time base age_group drug*time



Score Statistics For Type 3 GEE Analysis			
Source	DF	Chi-Square	Pr > ChiSq
drug	1	18.25	<.0001
Time	2	0.44	0.8006
base	1	13.56	0.0002
age_group	2	5.31	0.0702
Time*drug	2	1.44	0.4861

Trial 3 Conclusion: No significant interaction effects
Final Model: health =drug time base age_group

c) Regardless of your final model in (b), use the main effects model from part (a), and:

i) provide the odds ratio and corresponding 95% confidence interval that pertain to the overall intervention effect. Interpret this result in one sentence.

Model: health =drug time base age_group

Log OR: 1.9493 log OR 95% CI ={1.0356,2.8629}

OR=exp(log OR)=7.0235 95% CI OR={2.8167,17.5129}

The subjects on intervention had odds of good health that were 7.0235 times the odds of good health for those on control treatment. The 95% confidence interval does not include the null value so we can conclude that patients on intervention do significantly better than patients on control.

ii) provide model-predicted probabilities of self-rated good health -- one for each follow-up visit -- for an individual who is 25-34 years of age, has poor health at randomization, and is randomized to the intervention arm.

Model	Equation	Value
Follow Up 2, Intervention, Age 25-34, Poor health baseline	$\frac{e^{0.3809+0.9629+1.9493-1.6875}}{1+e^{0.3809+0.9629+1.9493-1.6875}}$	$=\frac{e^{1.6056}}{1+e^{1.6056}}=0.833$
Follow Up 3, Intervention, Age 25-34, Poor health baseline	$\frac{e^{0.3809+0.9629+1.9493-1.6875-0.0756}}{1+e^{0.3809+0.9629+1.9493-1.6875-0.0756}}$	$=\frac{e^{1.53}}{1+e^{1.53}}=0.822$
Follow Up 4, Intervention, Age 25-34, Poor health baseline	$\frac{e^{0.3809+0.9629+1.9493-1.6875-0.2234}}{1+e^{0.3809+0.9629+1.9493-1.6875-0.2234}}$	$=\frac{e^{1.3822}}{1+e^{1.3822}}=0.799$

4. For this question, consider the data from Question 1 of Problem Set 4 (restated below).

   A social scientist is interested in how region and education level are associated with people's feelings towards physical activity. She conducts a survey on randomly selected individuals, asking them their geographic region (West Coast, Midwest, and East Coast) and education level (college graduate, high school graduate, less than high school). Then she asked how strongly they agree with the following statement: "U.S. employers should offer greater incentives to have physically active employees." Participants could choose one of these options: "Disagree", "Neutral", or "Agree".

		Level of agreement with statement regarding physical activity		
Education Level	Region	Disagree	Neutral	Agree
College Graduate	West Coast	18	15	48
	Midwest	13	19	21
	East Coast	28	28	52
High School Graduate	West Coast	46	23	24
	Midwest	22	20	21
	East Coast	48	18	23
Less than high school	West Coast	13	15	28
	Midwest	15	16	17
	East Coast	15	17	24

	Level of agreement with statement regarding physical activity (pooled across regions)		
Education Level	Disagree (ordinal 1)	Neutral (ordinal 2)	Agree (ordinal 3)
Less than high school (ordinal 1)	43	48	69
High School (ordinal 2)	116	61	68
College (ordinal 3)	59	62	121

a) Under minimal assumptions, and pooling across regions, conduct a statistical test to assess the association of education level (treated as a nominal variable) with the level of agreement (also treated as a nominal variable). Justify your method.

Since we are looking for general association and with no scale (nominal column and row variables) it implies we want to use the Q_{GMH} test.

Q_{GMH}=38.6521
df=4
p-value=<0.0001

b) Under minimal assumptions, and pooling across regions, conduct a statistical test to assess the association of education level (treated as a nominal variable) with the level of agreement (treated as an ordinal variable), in terms of a location shift. Justify your method.
Since we are looking for the location shift with ordinal column variables it implies we want to use the Q_{SMH} test.
Q_{SMH}=36.8894
df=2
p-value=<0.0001

c) Under minimal assumptions, and pooling across regions, conduct a statistical test to assess whether increasing level of education (treated as an ordinal variable) provides a progressive location shift in the level of agreement (also treated as an ordinal variable). Justify your method.
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.
Q_{CSMH}=3.3328
df=1
p-value=0.0679

d) For each of the tests conducted in (a), (b), and (c), compare and contrast your results in terms of the nature of the association between education level and the level of agreement. Please provide this summary in a short paragraph of 3-5 sentences.

For pooling across regions, we can conclude that when education level and level of agreement are nominal variables they are statistically associated. When education is treated as a nominal variable and level of agreement is treated as a ordinal variable, we can conclude there is a statistical location shift. When level of education and level of agreement are both ordinal, we can conclude that there is not a progressive location shift.