Name:

BIOS 6612: Midterm Examination

March 18, 2021

Academic integrity: ***All graduate educational programs and courses taught at the CSPH are conducted under the honor system.***
I understand that my participation in this examination and in all academic and professional activities as a UC Anschutz Medical Campus student is bound by the provisions of the UC AMC Honor Code. I understand that work on this exam and other assignments are to be done independently unless specific instruction to the contrary is provided.

Signature: _____

Instructions

- This exam is worth a total of **100 points**.
- There is one extra credit question worth **5 points**.
- Check to make sure your exam has 8 pages (not including this cover sheet).
- The exam is open-book and open-notes. You may use a computer, **but no internet access is permitted**.
- Write your name at the top of this page and write your initials at the top of each subsequent page in the spaces indicated.
- Attempt all questions and show your work for partial credit.
- Write answers in the space provided below each question; if you need more space, use the back of the page, clearly indicating which question the continuing answer corresponds to.
- Unless otherwise indicated, hypothesis testing should be conducted at the 5% level of significance.

Table 1: Critical values for the χ^2 distribution (5% level of significance)

DF	Critical value
1	3.84
2	5.99
3	7.81
4	9.49
5	11.07
6	12.59
7	14.07
8	15.51
9	16.92
10	18.31
11	19.68
12	21.03
13	22.36
14	23.68
15	25.00

1. (**15 points**) Answer the following questions. Circle true or false: (**5 points**)

   (a) TRUE FALSE The three components of GLMs are the link function, the data distribution, and the linear predictor.

   (b) TRUE FALSE Wald test statistics are calculated using exact variance.

   (c) TRUE FALSE The inverse of any cumulative distribution function can be used as a link function to model binary data.

   (d) TRUE FALSE Grouped data can be modeled using the binomial distribution.

   (e) TRUE FALSE The Bayesian Information Criterion cannot be used to compare nested models.

2. (**45 points**) The Framingham Heart Study, designed to examine the effect of various factors on risk of coronary heart disease (CHD), includes data on 4856 individuals aged 30–62 years at baseline. Individuals were then followed for up to 12 years; at the end of follow-up, each participant was assessed to determine whether he or she had developed CHD. The full data set appears below: for each covariate pattern, the column `chd` gives the number of individuals determined to have developed CHD, while the column `total` gives the total individuals.

sex	age.group	cholesterol	chd	total
Male	30-49	<190	13	340
Male	30-49	190-219	18	408
Male	30-49	220-249	40	421
Male	30-49	>=250	57	362
Male	50-62	<190	13	123
Male	50-62	190-219	33	176
Male	50-62	220-249	35	174
Male	50-62	>=250	49	183
Female	30-49	<190	6	542
Female	30-49	190-219	5	552
Female	30-49	220-249	10	412
Female	30-49	>=250	18	357
Female	50-62	<190	9	58
Female	50-62	190-219	12	135
Female	50-62	220-249	21	218
Female	50-62	>=250	48	395

We are interested in modeling the probability of developing CHD. Assume throughout this question that reference levels for the covariates are as follows:

- `cholesterol`: levels < 190
- `sex`: female
- `age.group`: 30–49

(a) Estimate the probability of CHD in a male, aged 50–62, based on the results of this study. (**5 points**)

(b) A logistic regression model including `sex`, `age.group`, and `cholesterol` is fitted to the data, resulting in the following maximum likelihood coefficient estimates:

	Estimate	Std. Error	z value	Pr(>|z|)
`(Intercept)`	-4.1831	0.1902	-21.9934	0.0000
`cholesterol>=250`	1.1614	0.1843	6.3008	0.0000
`cholesterol190-219`	0.2462	0.2059	1.1958	0.2318
`cholesterol220-249`	0.7040	0.1928	3.6522	0.0003
`sexMale`	1.1000	0.1162	9.4674	0.0000
`age.group50-62`	1.1345	0.1113	10.1947	0.0000

(i) Provide an interpretation for the intercept in this model, or explain why you do not think the intercept is interpretable. (**5 points**)

(ii) Calculate the estimated odds ratio for the association between CHD and sex based on this model; provide an interpretation for the estimate. Construct a 95% confidence interval for this odds ratio. (**10 points**)

(iii) Describe the relationship between risk of CHD and cholesterol level in the context of this model; be sure to include odds ratio estimates and appropriate statements about statistical significance in your answer. (**10 points**)

(c) A second model is fitted to the data, adding the interaction between age group and sex to the model containing only the main effects. The following table of coefficient estimates is obtained:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.3608	0.2162	-20.1718	0.0000
cholesterol>=250	1.1117	0.1860	5.9776	0.0000
cholesterol190-219	0.2351	0.2059	1.1419	0.2535
cholesterol220-249	0.6725	0.1933	3.4794	0.0005
sexMale	1.3844	0.1871	7.3977	0.0000
age.group50-62	1.4656	0.2009	7.2967	0.0000
sexMale:age.group50-62	-0.4911	0.2425	-2.0250	0.0429

(i) Based on this model, what is the estimated odds ratio for CHD comparing male to female patients in the age group 30–49, adjusting for cholesterol? (**5 points**)

(ii) Based on this model, what is the estimated odds ratio for CHD comparing male to female patients in the age group 50–62, adjusting for cholesterol? (**5 points**)

(iii) Interpret the interaction between age group and sex. (*Hint: You may want to make reference to your answers to parts (i) and (ii) in your response to this question.*) Is this effect statistically significant? Give a p-value to support your conclusion. (**5 points**)

3. (**40 points**) A study is conducted to determine the association between sex and whether or not someone under-reports their height. The 200 participants were asked to self-report their height in inches (recorded as `repht`); then their height was measured by study personnel (recorded as `height`).

   (a) The 2×2 table below shows the number of participants by sex and whether their reported height was less than their measured height:

	`repht<height`	
Female?	`FALSE`	`TRUE`
`FALSE`	55	27
`TRUE`	59	42

   Estimate and interpret the odds ratio comparing risk of under-reporting height between men and women. Calculate a 95% confidence interval for this odds ratio. Is this a statistically significant association? (**15 points**)

(b) A logistic regression model is fitted to the data, including sex and measured height as covariates. Estimated coefficients are reported in the table below. *Note: the reference level for sex in this model is male.*

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.6239	1.3576	-3.4058	0.0007
I(sex == "F")TRUE	1.3413	0.4706	2.8502	0.0044
height	0.0508	0.0172	2.9557	0.0031

(i) Estimate and interpret the odds ratio comparing risk of under-reporting height between men and women based on this model. Calculate a 95% confidence interval for this odds ratio. Explain why this estimate differs from the estimate found in part (a). *Hint: the variance of the log odds ratio can be calculated by* $Var(\log(OR)) = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$*, where a, b, c, and d are cell values in the 2x2 table.* (**15 points**)

(ii) Estimate and interpret the odds ratio for the effect of measured height on under-reporting of height; include in your answer a test of the significance of this result. (**10 points**)

4. (**+5 points extra credit**) Suppose you have a sample of n iid Bernoulli random variables, each with success probability p. Let $\hat{p}$ be the MLE of p based on this sample. Give the asymptotic distribution of $1/\hat{p}$ based on the delta method. Explain why this might *not* be the best way to construct a confidence interval for $1/p$.