Name:

BIOS 6612: Midterm Examination Solutions

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**)

This is the empirical proportion of males aged 50–62 in the study developing CHD. We add up the total number of CHD cases in this group (across levels of cholesterol) and divide by the total number of individuals in this group (across levels of cholesterol): $(13 + 33 + 35 + 49)/(123 + 176 + 174 + 183) = 0.1982$.

(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**)

This model only contains categorical covariates, so the intercept may be interpreted as the estimated log odds of response for someone with reference levels of all covariates. In this case, this would be a female, aged 30–49, with cholesterol level < 190.

(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**)

The estimated odds ratio is $\exp(1.100) = 3.004$. This means that the odds of CHD among males are approximately 3 times higher than for females, adjusting for age and cholesterol level. A 95% confidence interval for this odds ratio is $\exp(1.100 \pm 1.9600(0.116)) = (2.392, 3.772)$.

(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**)

The reference group for cholesterol level is < 190, so each estimated odds ratio in the model is comparing risk of CHD in those with the indicated cholesterol level with someone with cholesterol < 190 (with the same age and sex). The estimated odds ratios are 1.279 for cholesterol 190–219, 2.022 for cholesterol 220–249, and 3.194 for cholesterol > 250. The comparison between < 190 and the next-lowest group is not significant ($p = 0.2318$), but the higher levels (220–249 and > 250) are significant ($p < 0.001$). This suggests an increasing level of risk of CHD with increasing cholesterol levels, adjusting for sex and age.

(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**)

30–49 is the reference level for age group, so in this interaction model the main effect coefficient estimate for sex gives the odds ratio comparing risk of CHD between male and female patients in this age group: $\exp(1.384) = 3.992$.

(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**)

Since sex and age group are involved in an interaction term in this model, we need to add the β estimates for sex and the sex-by-age interaction to get the estimated odds ratio for male vs. female risk of CHD in the age group 50–62: $\exp(1.384 - 0.4911) = 2.443$.

(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**)

The presence of this interaction term means that the relationship between sex and risk of CHD depends on age. Specifically, this model estimates a separate odds ratio for this relationship among younger and older people: we found an OR of 3.992 for younger patients and 2.443 for older patients. This means that the odds of CHD is 3.9 times higher in younger men than in younger women, but only 2.4 times higher in older men relative to older women. This is an interaction between two binary covariates, so we just need to look at the Wald p-value for the interaction coefficient to test significance: with $p = 0.0429$ we reject the null hypothesis and conclude that there is a statistically significant interaction between sex and age.

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**)

   From the table, the OR is estimated to be $55 \cdot 42/(59 \cdot 27) = 1.45$. This means that the odds of under-reporting height are increased by 45% in women compared with men. The log odds ratio is $\log(1.45) = 0.3716$; standard error on the log scale is $\sqrt{1/55 + 1/42 + 1/59 + 1/27} = 0.3098$. The estimated 95% CI is therefore $\exp(0.3716 \pm 1.9600 \cdot 0.3098) = (0.7901, 2.6613)$. This interval includes 1, so we would fail to reject the null and conclude that this is not a statistically significant relationship.

(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). (**15 points**)

The estimated odds ratio from the table is $\exp(1.3413) = 3.824$. This means that women's odds of under-reporting height are increased by 3.8 times compared to men, adjusting for measured height. An associated 95% confidence interval is $\exp(1.3413 \pm 1.9600 \cdot 0.4706) = (1.520, 9.618)$. This estimate and inference are different from part (a) because now we are adjusting for measured height: this means that this is the odds ratio comparing men and women *at the same measured height* on odds of under-reporting height.

(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**)

The odds ratio estimate is $\exp(0.0508) = 1.052$. This means that for each 1-kg increase in height, the odds of under-reporting height increase by 5%. Judging by the Wald p-value ($p = 0.0031$), this is a statistically significant effect.

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 $1/p$.
Our function $g(p) = 1/p$, so we find $g'(p) = -1/p^2$. We know that as $n \to \infty$, the distribution of $\hat{p}$ approaches $\mathcal{N}(p, p(1-p)/n)$. The delta method tells us that

$$\sqrt{n}\left(g(\hat{p}) - g(p)\right) \xrightarrow{d} \mathcal{N}\left(0, \left[g'(p)\right]^2 p(1-p)\right).$$

Applying this to our $g(p) = 1/p$, we have the approximate variance of $1/\hat{p}$ as

$$\left(-\frac{1}{p^2}\right)^2 \frac{p(1-p)}{n} = \frac{1-p}{p^3 n}.$$

Therefore, the asymptotic distribution of $1/\hat{p}$ is

$$\sqrt{n}\left(\frac{1}{\hat{p}} - \frac{1}{p}\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{1-p}{p^3}\right).$$

This might not be the best choice for how to construct a confidence interval because the sampling distribution of $1/\hat{p}$ is likely further from normality for a given sample size n than the distribution of $\hat{p}$. In particular, since $\hat{p} \in [0,1]$, $1/\hat{p} \in [1, \infty)$ will be skewed for smaller n. What would make more sense would be to construct a CI the way we do for odds ratios: we have the usual Wald-type CI for p based on $\hat{p}$,

$$\hat{p} \pm 1.96 \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

so we can get a CI for $1/p$ by applying $g(p) = 1/p$ to the endpoints, remembering that we need to reverse the endpoints since g is a decreasing function:

$$\frac{1}{\hat{p} \mp 1.96 \cdot \sqrt{\hat{p}(1-\hat{p})/n}}.$$