**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 $\chi^2$ 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**)

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 $1/p$.