Name:

BIOS 6612: Practice Midterm Examination
Solutions

March 10, 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

- You may use a computers, **but no statistical model fitting procedures or internet access is permitted**.
- The exam is open-book and open-notes.
- 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.

1. (**15 points**) Answer the following questions.

   (a) Circle true or false: (**5 points**)

   (i) TRUE FALSE The Wald test is based on the distance between estimate and true parameter, measured in units of standard errors.

   (ii) TRUE FALSE The Wald, score, and likelihood ratio tests are equivalent in small samples.

   (iii) TRUE FALSE The likelihood ratio test is generally more powerful than the Wald test.

   (iv) TRUE FALSE The likelihood ratio test may be used to compare non-nested models.

   (v) TRUE FALSE The score test is based on the derivative of the log-likelihood at the value of the parameter under the alternative hypothesis.

   Wald test is based on distance between estimate and true parameter, and relies on the log-likelihood surface being roughly quadratic in a neighborhood of the MLE, so you should be more confident applying this in large samples (where the CLT operates reliably) than in small samples. Score test is based on the slope of the log-likelihood at the null parameter value and likelihood ratio test directly compares the likelihood of two models (which model is more likely given the observed data?); both of these work better in small samples. All three are equivalent in large samples under the null hypothesis, and have a limiting chi-square distribution. All of them can only be used to compare nested models.

2. (**40 points**) An analysis of historical data on 1309 passengers in the *Titanic* disaster of 1912 was conducted to determine the effects of several demographic variables on probability of passengers' survival. The data set consists of the following variables:

   - `sex`, factor with two levels, `female` and `male`.
   - `age`, in years; missing for 263 of the passengers.
   - `passengerClass`, factor with three levels `1st`, `2nd`, or `3rd` class.
   - `survived` (**outcome**), factor with two levels, `yes` if the passenger survived the sinking and `no` if not.

   Below are some summary statistics for this data set.

```
      age              sex      passengerClass survived
Min.   : 0.1667   female:466   1st:323        no :809
1st Qu.:21.0000   male  :843   2nd:277        yes:500
Median :28.0000                3rd:709
Mean   :29.8811
3rd Qu.:39.0000
Max.   :80.0000
NA's   :263
```

   We are interested in modeling the probability that `survived==yes`. Some critical values that may be useful as you answer the following questions are $\chi^2_{0.95,1} = 3.8415$, $\chi^2_{0.95,2} = 5.9915$, $\chi^2_{0.95,3} = 7.8147$.

   (a) The table below gives the cross-tabulation for the outcome and passenger class.

	survived	
passengerClass	no	yes
1st	123	200
2nd	158	119
3rd	528	181

   (i) What is the probability of survival for all passengers? (**2 points**)
   We have to sum up the columns to get the probability of survival overall: $(200 + 119 + 181)/1309 = 0.381971$.

   (ii) Compute the log-likelihood for the intercept-only logistic regression model. (**4 points**)
   Then the log-likelihood is $\log L = (123+158+528)\log(1-0.381971)+(200+119+181)\log(0.381971) = -870.5122$.

(iii) Compute the log-likelihood for the logistic regression model treating `passengerClass` as a categorical covariate with three levels. (**6 points**)

First we need to calculate the probability of survival for each level of the covariate: we have $200/(123+200) = 0.6192$ for 1st class, $119/(119+158) = 0.4296$ for 2nd class, and $181/(181+528) = 0.2553$ for 3rd class. Then the overall log-likelihood is $\log L = 200\log(0.6192) + 123\log(1-0.6192) + 119\log(0.4296) + 158\log(1-0.4296) + 181\log(0.2553) + 528\log(1-0.2553) = -806.6295$

(iv) Conduct a likelihood ratio test at the 5% level of significance of the null hypothesis that passenger class is not associated with odds of survival. Be sure to state the reference distribution under the null. (**6 points**)
The difference in number of parameters between the intercept-only and passenger class models is 2, so the reference distribution is chi-square with 2 degrees of freedom. The test statistic is $-2(-870.5122+806.6295) = 127.7655 > \chi^2_{0.95,2} = 5.9915$, so we reject the null hypothesis and conclude that passenger class is a significant predictor of survival.

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

	Estimate	Std. Error	z value
`(Intercept)`	3.5221	0.3267	10.7807
`sex male` (ref. `female`)	-2.4978	0.1660	-15.0439
`age`	-0.0344	0.0063	-5.4325
`passengerClass 2nd` (ref. `1st`)	-1.2806	0.2255	-5.6778
`passengerClass 3rd` (ref. `1st`)	-2.2897	0.2258	-10.1401

(i) Provide an interpretation for the intercept in this model, or explain why you do not think the intercept is interpretable. (**4 points**)
Although age is not centered in this model, there is at least one passenger with age close to 0. Thus, the intercept can be interpreted as the log odds of survival for a newborn female in first class.

(ii) Calculate the estimated odds ratio for the association between survival and passenger sex; provide an interpretation for the estimate. Construct a 95% confidence interval for this odds ratio. (**6 points**)
The estimated odds ratio is $\exp(-2.4978) = 0.0823$. This means that the odds of survival for male passengers were 0.082 times the odds of survival for female passengers, adjusting for age and passenger class. Equivalently, we could say that the odds of survival for male passengers were approximately 91% lower than for female passengers, adjusting for age and passenger class. A 95% confidence interval for this odds ratio is $\exp(-2.4978 \pm 1.9600(0.1660)) = (0.0594, 0.1139)$.

(iii) Calculate the estimated odds ratio for the association between survival and passenger age; provide an interpretation for the estimate. Construct a 95%

confidence interval for this odds ratio. (**6 points**)
The estimated odds ratio is $\exp(-0.0344) = 0.9662$. This means that the odds of survival decreases by approximately 3.3% for each additional year of age, adjusting for sex and passenger class. Equivalently, we could say that the odds of survival decrease by a factor of 0.966 for each 1 year increase in age, adjusting for sex and passenger class. A 95% confidence interval for this odds ratio is $\exp(-0.0344 \pm 1.9600(0.0063)) = (0.9543, 0.9783)$.

(c) A second model is fitted to the data, adding an interaction term between `age` and `sex`; these two main effects remain in the model as does `passengerClass`. The following estimates and Wald p-values are obtained:

	Estimate	Pr(>|z|)
`sex male` (ref. `female`)	-1.0298	0.0041
`age`	-0.0041	0.6660
`sex*age`	-0.0529	0.0000

(i) Is there a significant interaction between age and sex with respect to odds of survival? Provide a p-value to support your conclusion. (**2 points**)
There is a significant interaction between age and sex, $p < 0.0001$ from the last line of the table.

(ii) Interpret the effect of sex on odds of survival, given that age and the sex $\times$ age interaction are included in the model. (**4 points**)
The coefficient estimate for sex is -1.0298, so the odds ratio for survival of a male passenger at age 0 compared with a female passenger at age 0 is $\exp(-1.0298) = 0.3571$. If we want to compare the survival odds between a male and female passenger at specific ages, then we need to look at the odds ratio

$$\begin{aligned} OR &= \frac{\exp(\cdots + \texttt{male}\beta_1 + \texttt{age}\beta_2 + \texttt{male}\cdot\texttt{age}\beta_3)}{\exp(\cdots + \texttt{female}\beta_1 + \texttt{age}\beta_2 + \texttt{female}\cdot\texttt{age}\beta_3)} \\ &= \frac{\exp(\cdots + \beta_1 + \texttt{age}\beta_2 + \cdot\texttt{age}\beta_3)}{\exp(\cdots + 0\cdot\beta_1 + \texttt{age}\beta_2 + 0\cdot\texttt{age}\beta_3)} \\ &= \exp(\beta_1 + \texttt{age}\beta_3) \end{aligned}$$

This quantity is estimated from our data as $\exp(-1.0298 - \texttt{age}0.0529)$, meaning that, for example, the odds ratio comparing survival between males and females at age 10 is $\exp(-1.0298 - 10 \cdot 0.0529) = 0.2104$ and at age 30 it is $\exp(-1.0298 - 30 \cdot 0.0529) = 0.0730$. This means that there is a larger difference in odds of survival between older males and females than between younger males and females.

3. (**5 points**) Suppose you have used maximum likelihood estimation to estimate a parameter $\hat{\theta}$ that you know to be asymptotically normally distributed with mean θ and variance σ_θ^2. Derive the asymptotic distribution of $\log \hat{\theta}$.
   This is a transformation of the MLE, so by the invariance property of MLEs $\log \hat{\theta}$ is the MLE of $\log \theta$, and therefore asymptotically normal with mean $\log \theta$. The variance (from the delta method) is $\left(\frac{d}{d\theta} \log \theta\right)^2 \sigma_\theta^2 = \sigma_\theta^2/\theta^2$.