PhD Qualifying Examination–Part II

Department of Statistics
University of South Carolina
August 11, 2020 - 9:00AM–11:00AM

READ FIRST THESE INSTRUCTIONS

1. DO NOT write your name on any of your answer sheets. Instead, write your pre-assigned codename.
2. There are two (2) problems on this examination.
3. You are **not allowed** to use search engines during the examination. But you may use the HELP manuals of the statistical packages that you use. Please adhere to the HONOR CODE in this instance. Any violation of the HONOR CODE (such as using search engines) will lead to a zero for the exam.
4. SAS OnDemand can be accessed through the following link https://odamid.oda.sas.com.
5. You have two hours for this examination. Both problems will be graded and are of equal weight.

The Problems

1. A local health clinic sent fliers to its clients to encourage everyone, but especially older persons at high risk of complications, to get a flu shot in time for protection against an expected flu epidemic. In a pilot follow-up study, 159 clients were randomly selected and asked whether they actually received a flu shot. A client who received a flu shot was coded $Y = 1$, and a client who did not receive a flu shot was coded $Y = 0$. In addition, data were collected on their age ($\mathbf{X}_1$) and their health awareness. The latter data were combined into a health awareness index ($\mathbf{X}_2$), for which higher values indicate greater awareness. Also included in the data was client gender, where males were coded $\mathbf{X}_3 = 1$ and females were coded $X_3 = 0$. Let $\mathbf{X} = [X_1, X_2, X_3]$.

i:	**1**	**2**	**3**	...	**157**	**158**	**159**
Y_i:	0	0	1	...	1	1	1
X_{i1}:	59	61	82	...	76	68	73
X_{i2}:	52	55	51	...	22	32	56
X_{i3}:	0	1	0	...	1	0	1

   (a) (i) Consider a simple model with only one predictor, X_3. Use the Delta method to derive the formula for the standard error of the estimate for e^{β_1}.

   (ii) Next, multiple logistic regression model with three predictor variables in first-order terms is assumed to be appropriate. Write down the regression model and the likelihood function.

   (iii) Find the maximum likelihood estimated value of β_0, β_1, β_2, and β_3 using the data provided.

   (iv) Obtain estimates for e^{β_1}, e^{β_2}, e^{β_3} and interpret these numbers.

   (b) (i) Assess the adequacy of the fit of the logistic regression model in (a)(ii). Propose a remedy if you identify problems in your diagnosis.

   (ii) Derive the hat matrix using the logistic model described in (a)(ii), reduced as much as possible. Demonstrate and explain the importance of the hat matrix in model diagnosis. Use the values in the hat matrix to produce a regression diagnostic plot and explain what you see.

   (c) For a logistic regression model, let $\hat{\boldsymbol{\mu}}^{(-)} = (\hat{\boldsymbol{\mu}}^{(-1)}, \hat{\boldsymbol{\mu}}^{(-2)}, ..., \hat{\boldsymbol{\mu}}^{(-n)})$, where $\hat{\boldsymbol{\mu}}^{(-i)}$ denotes the estimate of $E(Y_i)$ for observation i after fitting the model without that observation (leave-one-out cross-validation). Let $\hat{\boldsymbol{\mu}}$ denote the estimate of $E(Y_i)$ using the full data.

   Consider an intercept only logistic regression model for all i, derive the formulas for calculating $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\mu}}^{(-i)}$. Derive the formula for the correlation between $\mathbf{Y}$ and $\hat{\boldsymbol{\mu}}^{(-)}$ (simplify as much as possible).

2. The table below shows the sodium content (mg) for 5 different brands of beers in half-gallon jars sold in restaurants.

A	B	C	D	E
75	57	58	58	62
67	58	61	59	66
70	60	56	58	65
75	59	58	61	63
65	62	57	57	64
71	60	56	56	62
67	60	61	58	65
67	57	60	57	65
76	59	57	57	62
68	61	58	59	67

(a) Perform a hypothesis test to determine whether or not the mean sodium content is the same in all brands at $\alpha = 0.05$ using ANOVA. Explain your notation and be specific about the hypotheses and the model.

(b) Express the total sum of squares, the error sum of squares, and the regression sum of squares as three quadratic forms in $\mathbf{Y}$ ($\mathbf{Y}$ is the vector of all sodium content measurements). According to the model described in (a), what distribution does each sum of squares have?

(c) Several beer drinkers mentioned that brand A beer tasted saltier than the other four beer brands (B, C, D, E). Let μ_i be the mean of the ith brand and let $\psi_c = \mu_A - \frac{1}{4}(\mu_B + \mu_C + \mu_D + \mu_E)$. Perform a hypothesis test to determine whether $\psi_c = 0$ at $\alpha = 0.05$. Also, derive the formula for calculating the 95% confidence interval for ψ_c.

(d) (i) If we obtain several 95% confidence intervals for the following linear combination $\psi_1, \psi_2, ..., \psi_g$ separately, argue why the probability that each ψ_i will be in its interval **simultaneously** is less than 0.95? That is,

$$P(\psi_1 \in I_1, \psi_2 \in I_2, \psi_3 \in I_3, ...\psi_g \in I_g) \leq 0.95.$$

(ii) Consider the following four linear combinations $c_1 = (1, -\frac{1}{4}, -\frac{1}{4}, -\frac{1}{4}), c_2 = (1, -1, 0, 0), c_3 = (1, 0, -1, 0), c_4 = (1, 0, 0, -1)$, derive the **simultaneous** $(1-\alpha)\%$ confidence intervals $(I^*_{c_i})$ using the Bonferroni procedure. Show that

$$P(\psi_{c_1} \in I^*_{c_1}, \psi_{c_2} \in I^*_{c_2}, \psi_{c_3} \in I^*_{c_3}, \psi_{c_4} \in I^*_{c_4}) \geq 0.95.$$