2023 August Qualifying Exam

Day 2

1. Anorexia is an eating disorder characterized by abnormally low body weight. A study on anorexia consists of weight data for 72 young women who were divided into three groups: *control group with standard treatment* (Cont), *cognitive behavior treatment group* (CBT), in which the participants met with a therapist, and *family therapy group* (FT), in which the parents intervened when they observed anorexia. You can access this dataset in R using the command:

```
load(url("https://people.stat.sc.edu/gregorkb/data/anorexia.Rdata"))
```

The data set looks like this:

```
head(anorexia)

##   Treat Prewt Postwt
## 1  Cont  80.7   80.2
## 2  Cont  89.4   80.1
## 3  Cont  91.8   86.4
## 4  Cont  74.0   86.3
## 5  Cont  78.1   76.1
## 6  Cont  88.3   78.1
```

The dataset contains the following three variables:

- Treat: Factor of three levels: "Cont" (control), "CBT" (Cognitive behavioural treatment) and "FT" (family treatment).
- Prewt: Weight of patient before study period, in lbs.
- Postwt: Weight of patient after study period, in lbs.

The outcome variable of interest is weight gain (Postwt-Prewt).

(a) Make plots to examine the relationship between weight gain ((Postwt-Prewt) and before weight (Prewt) by treatment (Treat) groups. Describe what you observe in these plots. Are there differences in weight gain between the three treatment groups? Which treatment option seems to be the most/the least effective? Do you observe any individual variation in response to the treatment?

(b) Let y_{ij} denote the weight gain for the j-th woman in the i-th treatment group. Let n_i denote the number of observations in the i-th treatment group. Consider the model

$$y_{ij} = \mu_i + \epsilon_{ij}, \qquad i \in \{\text{Cont, CBT, FT}\}, \quad j = 1, 2, ..., n_i,$$

where ϵ_{ij} are independent Normal$(0, \sigma^2)$ random variables.

i. Write down a test statistic and the corresponding decision rule for testing H_0: $\mu_i = \mu_{i'}$ versus H_1: $\mu_i \neq \mu_{i'}$, for a given pair of treatments $i \neq i'$. Be sure to define any notation you introduce.

ii. Write down a test statistic and the corresponding decision rule for testing

$$H_0\text{: } \mu_{\text{Cont}} = \mu_{\text{CBT}} = \mu_{\text{FT}} \quad \text{versus} \quad H_1 : \mu_i \neq \mu_{i'} \text{ for some } i \neq i'.$$

Be sure to define any notation you introduce.

(c) Is there a significant difference in weight gain between the three treatments based on the model in (b)? If so, identify the pairs in which the treatments are significantly different at level 0.05.

(d) (i) Consider modeling after weight (Postwt) as a function of Treatment (Treat) and before weight (Prewt). The table below lists potential models for the data. Use appropriate tests to determine which of these models is most appropriate for the data.

	Covariates
1	
2	Prewt
3	Treat
4	Treat, Prewt
5	Treat×Prewt
6	Treat, Prewt, Treat×Prewt

(ii) Is it possible to compare all pairs of models in the above table using F-tests? If not, list the pairs which cannot be compared with an F-test.

(e) Create a new variable indicating whether the weight increased or not. Write down a logistic regression model using this new variable to examine the effect of the three treatments. Write a short conclusion regarding the effectiveness of the three treatments based on the results from the logistic regression.

2. In a proteomics experiment, counts of the protein TGF-β in tumor samples taken from a breast cancer patient before and after treatment were recorded according to this table:

	Before treatment	After treatment
TGF-β Protein	a	b
All other protein	c	d
Total protein count	$t_a = a + c$	$t_b = b + d$

Table 1: Proteomics data from one patient

Let π_a and π_b be the true fractions of TGF-β protein before and after treatment, respectively. In this paired sample testing, the parameter of interest is the treatment effect θ defined as

$$\theta = \frac{\pi_a}{\pi_b}.$$

The parameter θ is the fold change in the TGF-β protein abundance after normalization for total protein count. Notice that the calculation of θ is identical to relative risk for a 2×2 contingency table.

(a) Use the appropriate test to determine the significance of the treatment effect using the observed data presented in Table 2. Be sure to state your hypotheses, testing procedure, decision rule, and conclusion.

	Before treatment	After treatment
Protein of interest	13	10
All other protein	176	94
Total protein count	189	104

Table 2: Observed data from one patient

(b) Let X_1 and X_2 be independent random variables representing the counts of the TGF-β protein in the before- and after-treatment samples, respectively, and assume

$$X_1 \sim \text{Poisson}(\pi_a t_a) \quad \text{and} \quad X_2 \sim \text{Poisson}(\pi_b t_b).$$

Write down the joint probability mass function of X_1 and X_2.

(c) If $X_1 + X_2$ is regarded as fixed, show that the distribution of X_1 is given by

$$P(X_1 = a | X_1 + X_2 = a + b) = \frac{(a+b)!}{a!b!} P^a Q^b$$

and give P and Q.

(d) Derive the maximum likelihood estimates $\widehat{P}$ and $\widehat{\theta}$ for P and θ, respectively. Report the values of $\widehat{P}$ and $\widehat{\theta}$ based on the data presented in Table 2.

(e) Now consider the data presented in Table 3 and Table 4 from two individuals and assume that the treatment effect θ is constant across individuals. Propose (just describe, no need to implement) an approach that could account for between-subject heterogeneity while estimating θ.

	Before treatment	After treatment
Protein of interest	33	51
All other protein	176	94
Total protein count	500	500

Table 3: Observed data from patient #1

	Before treatment	After treatment
Protein of interest	86	10
All other protein	149	94
Total protein count	1000	1000

Table 4: Observed data from patient #2

3. Uterine leiomyomata (also called fibroids) are the leading cause of hysterectomy for women approaching the age of menopause and are also found to be associated with adverse pregnancy outcomes, such as difficulty conceiving, preterm birth, and cesarean delivery. A prospective cohort study of early pregnancy screened pregnant women within the first 13 weeks of gestation with endovaginal ultrasound. All participants are independent in the study. The goal of this study is to investigate significant risk factors for fibroids and estimate their effects. You can access the dataset with this R command:

```
load(url("https://people.stat.sc.edu/gregorkb/data/FibroidData.Rdata"))
```

The dataset contains the following variables from the study:

- Column 1: the study number
- Column 2: age at study (ultrasound)
- Column 3: race (1 for black and 0 for white)
- Column 4: parity status (whether a participant has given birth before; 1 for yes)
- Column 5: age of menarche (when a participant had her first period)
- Column 6: obese status (body mass index greater than 30; 1 for yes)
- Column 7: Fibroid (the presence of fibroid at ultrasound; 1 for yes)

(a) There are two sub-studies in the data, as seen in Column 1. Focus on the data in sub-study 2 for all of part (a). Conduct a regression analysis with "Fibroid" as the binary response and the age at study, race, parity, age of menarche, and obese as covariates. Use a probit model, which specifies the relationship between the binary response Y_i and the vector of covariates X_i for subject i as

$$P(Y_i = 1|X_i) = \Phi(X_i^T\beta) = \Phi(\beta_0 + \sum_{j=1}^{5} x_{ij}\beta_j), \tag{1}$$

where $X_i = (1, x_{i1}, ..., x_{i5})^T$, $\beta = (\beta_0, \beta_1, \cdots, \beta_5)^T$, and $\Phi(\cdot)$ is the cumulative distribution function of a standard normal random variable. Label this model M1. You can either use an existing package

or write your own code for this analysis. Both frequentist and Bayesian approaches are acceptable. Answer the following questions.

i. Compute estimates of the parameters in the model M1 expressing the effects of race and parity and give careful interpretations of the estimated values.

ii. List the significant covariates at significance level 0.05. Describe the characteristics of a subgroup of women who have a high risk of fibroids.

iii. Calculate the estimated probabilities of contracting Fibroids for all combinations of subgroups of women with different race, parity status, and obesity status whose age at ultrasound was 25 and age at menarche was 12.

iv. The study investigators would like to consider a more flexible model with a nonlinear effect of the age at study. Construct a new model, M2, by adding a quadratic term of age at study to M1. Fit M2 and comment on whether a quadratic term is needed. Be sure to provide evidence to support your conclusion.

(b) It was reported that the ultrasonographers in sub-study 1 lacked intensive systematic training and likely missed some fibroids. To address this under-reporting problem, consider introducing a latent true fibroid status variable R_i for participant i, and let α be the probability of missing a fibroid in sub-study 1. That is, let

$$P(Y_i = 0|R_i = 1, G_i = 1) = \alpha,$$

where G_i is the number of the sub-study in which subject i participated. Assume that there is no over-reporting, so that $P(Y_i = 1|R_i = 0, G_i = 1) = 0$. Assume no under- or over-reporting exists in sub-study 2, so that $Y_i = R_i$ if $G_i = 2$. We would like to analyze the fibroid data from both sub-studies using the same model (M1). Note that R_i follows a probit model instead of Y_i in equation (1) in this case.

i. Derive $P(Y_i = 1|X_i)$ and $P(Y_i = 0|X_i)$ for subject i in sub-study 1.

ii. Write down the observed likelihood based on all of the observed data $\{(Y_i, X_i, G_i), i = 1, \ldots, n\}$ under the probit model M1 in the case of under-reporting.

iii. Derive the conditional probabilities

$$P(R_i = 1|Y_i = 1, G_i = 1, X_i) \quad \text{and} \quad P(R_i = 1|Y_i = 0, G_i = 1, X_i)$$

for subject i in sub-study 1.