STA 441 S2024 Make-up Test

Please write your answers in the examination booklet. Make sure to write your name and student number on the front cover.

1. (4 points) A market research firm is interested in testing two versions of a television commercial. A large sample of consumers is randomly divided into two groups, by tossing a fair coin. If the coin comes up Heads, the person sees commercial version One. If the coin comes up Tails, the person sees commercial version Two. Each consumer views the commercial alone, in a separate room, and then is given an opportunity to purchase the product. The response variable is binary – purchase versus non-purchase.

   Is there a problem here with potential confounding variables? Briefly explain.

2. (4 points) A market research firm is interested in testing the effect of an advertising campaign (consisting of radio and TV ads, billboards, coupons, etc.) for Jolt Cola. A large random sample of consumers is interviewed before the campaign begins, and asked how much Jolt Cola they have purchased during the past seven days. Then the campaign runs for a month, and the same consumers are interviewed again. Once again, they are asked how much Jolt Cola they have purchased during the past seven days.

   Is there a problem here with potential confounding variables? Briefly explain.

3. (3 Points) A medical study finds that the more organic food a person eats (as a percent of total calories), the better the person's overall health on average (on a 100 point scale as rated by a physician). Assuming they used an *elementary* statistical method, what method do you think they used?

4. (13 Points) In a study of babies' health, the variables included mother's age, mother's race, mother's weight and baby's weight at birth. Because this study was carried out in the United States, the variable `race` had only three values: Black, White and Other.

   (a) Let y denote baby's weight at birth, x_1 denote mother's age, and x_2 denote mother's weight. Representing mother's race by one or more dummy variables, write a regression equation for $E(y|\mathbf{x})$. There are no interactions. The dummy variable coding scheme is up to you, and there is more than one right answer. You do not need to say how your dummy variable(s) are defined. You will do that in the next part.

   (b) Make a table with three rows, one for each race. Make columns showing how your dummy variable or variables are defined. Add another, wider column showing baby's expected weight. The *symbols* for the dummy variable(s) will not appear in this last column.

   (c) Controlling for mother's age and weight, is baby's weight related to mother's race? Give the null hypothesis, using symbols from your regression equation.

   (d) Correcting for mother's age and weight, is baby's weight different for Black and White mothers? Give the null hypothesis, using symbols from your regression equation.

   (e) Allowing for mother's age and race, is mother's weight related to baby's weight? Give the null hypothesis, using symbols from your regression equation.

5. (22 points) Data are collected on university students who were looking for employment immediately following graduation. Three pieces of information are available for each student:

   - Academic Division (Humanities, Sciences or Social Science)
   - Final cumulative Grade Point Average
   - Whether they were employed full time 6 months after graduation: Yes or No.

   Consider a logistic regression model *with* an intercept, indicator dummy variable coding, and GPA centered by subtracting off the mean for the entire sample.

   (a) Denoting centered GPA by x, write a logistic regression equation for $\ln \frac{\pi}{1-\pi}$. Include an intercept. There are no interactions.

   (b) Make a table with three rows, showing how you would set up indicator dummy variables for Academic Division. Make Humanities the reference category. Add a column showing the odds of being employed.

   (c) What is the ratio of the odds of being employed for a Sciences graduate to the odds of being employed for a Social Sciences graduate with the same GPA? Answer in terms of the β symbols of your model.

   (d) One grade point on a four point scale is pretty large. When GPA increases by one point, the odds of being employed are multiplied by _____. Answer in terms of the β symbols of your model.

   (e) Controlling for GPA, you want to test whether students from the different academic divisions have different chances of finding a job. State the null hypothesis in terms of one or more β values.

   (f) You want to know whether, controlling for Academic Division, the chances of finding a job depend on your marks. State the null hypothesis in terms of one or more β values.

   (g) What is the probability of employment for a Sciences graduate with average GPA? Answer in terms of the β symbols of your model.

   (h) Some people say that a Humanities graduate with average marks has no better than a 50% chance of finding work within 6 months of graduation. What null hypothesis would you use to test this claim? State the null hypothesis in terms of one or more β values.

6. (14 Points) In a study of math education in elementary school, equal numbers of boys and girls were randomly assigned to one of three training programs designed to improve spatial reasoning. After five school days of training, the students were given a standardized test of spatial reasoning. Score on this test is the response variable. The table below shows population treatment means.

	Training Program		
	One	Two	Three
Girls	μ_{11}	μ_{12}	μ_{13}
Boys	μ_{21}	μ_{22}	μ_{23}

In terms of the μ_{ij} values in the table, give the null hypothesis you would test to answer each question.

(a) Is there a main effect of Training Program?

(b) Averaging across Training program, is there difference between the average scores of girls and boys?

(c) Is there a Training Program by Gender interaction?

(d) Averaging across Gender, does Training Program affect the mean spatial reasoning score?

(e) Is the effect of Training Program the same for boys and girls?

(f) Does Training Program have any effect for girls?

(g) Is there a difference beween the marginal means for girls and boys?

7. (23 Points) This question uses the Diet data from Assignment 4. See the handout with formula sheet and printouts. Recall that the variables were

   - `Person`: Identification code
   - `gender`: F or M
   - `Age`: In years
   - `Height`: In cm.
   - `pre_weight`: Weight in kg. before starting the diet
   - `Diet`: 1, 2, or 3, randomly assigned
   - `weight6weeks`: Weight in kg. after 6 weeks on the diet

   (a) Denoting `weight6weeks` by y and `pre_weight` by x, write $E(y|\mathbf{x})$ for the model used in `proc reg`. Of course the vector $\mathbf{x}$ includes other variables in addition to $x =$ `pre_weight`.

   (b) Here is the main question. Controlling for gender and weight before starting, did diet affect weight after six weeks?

      i. Give the null hypothesis in symbols from your regression model.

      ii. For the test of this null hypothesis, copy the table below into your answer book, and fill it in.

Test Statistic (a number)	p-value (a number)	Reject Null Hypothesis? (Yes or No)	Statistically Significant? (Yes or No)

      iii. What proportion of the remaining variation in weight after 6 weeks is explained by diet, after taking into account gender and weight before the study? The answer is a number. Show a little work. **Circle your answer.**

      iv. Now consider Bonferroni-corrected pairwise comparisons of the diets, controlling for gender and weight before starting. Copy the table below into your answer book and fill it in.

Comparison	Test statistic value	Uncorrected p-value	Bonferroni corrected p-value
Diet 1 vs. 2			
Diet 1 vs. 3			
Diet 2 vs. 3			

      v. In plain, non-statistical language, what do you conclude?

8. (17 Points) This question uses the Birdkeeping data from Assignment 9. See the handout with formula sheet and printouts. Here, the focus is on smoking and cancer rather than on bird keeping and cancer. The variables we will use are pretty obvious from the SAS code, except possibly for `ses`. This stands for "SocioEconomic Status," a combination of how much money you make, how educated you are and how respectable your job is. See the `proc format` for what the values 0 and 1 mean.

   (a) You wish to test whether, controlling for the other variables, number of cigarettes per day is related to getting lung cancer.

      i. Copy the table below into your answer book, and fill it out.

Test Statistic (a number)	p-value (a number)	Reject Null Hypothesis? (Yes or No)	Statistically Significant? (Yes or No)

      ii. In plain, non-statistical language, what do you conclude? No marks for this without the first part right. Start your answer with "Allowing for other possible risk factors . . . "

   (b) For fixed sex, socioeconomic status and age, smoking 20 more cigarettes per day multiplies the estimated odds of cancer by — what? The answer is a number. Show a little work. **Circle your answer.**

   (c) Estimate the probability of cancer for a 50 year old non-smoking male of low socioeconomic status. The answer is a number. Show a little work. **Circle your answer.**