STAT 796: Homework 8

Due Friday, April 5 at 11:59pm on Canvas. Please append your code to your responses.

For this assignment, use data from a study of esophageal cancer among Singapore men. The data come from a case-control study in which individuals with cancer were matched to 4 men without cancer. Matching variables included decade of age. Exposure risk factors, many associated with traditional Chinese culture, were measured to assess their relationship with cancer risk. These data are available on Canvas in the file `signapore.csv` and contain the following variables:

- `set` – Matched set number
- `case` – Binary (0/1) indicator of being a case
- `age` – Age of individual in years
- `dialect` – Binary (0/1) indicator of dialect group. 1 = Hokhien or Teochew, 0 = Cantonese or other
- `sampu` – Binary indicator of sampu wine consumption
- `cigs_day` – Number of cigarettes consumed per day
- `hotbev` – Number of hot beverages (near boiling temperature) consumed per day

1. Summarize the *distribution* of the number of cigarettes smoked per day. (Appropriate summaries could include a table of values, a histogram, or density plot).
2. Using conditional logistic regression, estimate the odds ratio for risk of esophageal cancer comparing those who do drink hot beverages to those who do not. Adjust your estimate for sampu wine consumption and cigarette usage (you may decide the form of the variable for adjustment). Provide a summarizing statement that includes a measure of uncertainty (confidence interval) and a measure of the strength of the statistical evidence (p-value from hypothesis test).
3. Is there evidence of a trend in risk for higher amounts of hot beverage consumption? Evaluate this by fitting a model that uses the number of hot beverages as a continuous exposure. Provide a summarizing statement that includes a measure of uncertainty (confidence interval) and a measure of the strength of the statistical evidence (p-value from hypothesis test).
4. Is there evidence of heterogeneity in the trend in risk for higher amounts of hot beverage consumption? Evaluate this by fitting a model that uses the number of hot beverages as a *categorical* exposure. Provide a summarizing statement that includes a measure of uncertainty (confidence interval) and a measure of the strength of the statistical evidence (p-value from hypothesis test).
5. Using the models from questions 2, 3, and 4, fill in the following table of estimated odds ratios.

Number of Hot Beverages	Estimated OR (Binary Exposure)	Estimated OR (Continuous Exposure)	Estimated OR (Categorical Exposure)
0			
1			
2			
3			