Homework 1

Aji John

Q1. For the scientific question of interest, what are the two primary variables in this study? What is the response variable and what is the predictor variable? What types of variables are they: quantitative or qualitative? Discrete or continuous? Nominal, ordinal, interval, or ratios? Censored (right, left, or both) or uncensored? Explain briefly.

Answer. The scientific question explores magnetic resonance imaging (MRI) changes and its relationship with aging, cardiovascular disease, cerebrovascular disease, and mortality. One of forays is to look at the behavior or lifesyle choice of smoking behavior, and asking whether it explains death by any cause. The two primary variables are 'packyrs'(smoking history), and 'death'. Response variable is 'death', and predictor variable is 'packyrs'. Predictor variable is quantitative, and response variable is discrete categorical. The study is right censored.

Q2. What is the population of interest for this study? What is the sample? What is the size of the sample? Are there any individuals in the sample who have missing data on smoking history? If so, provide the participant identification numbers for any study individuals who have missing data on smoking history.

Answer. Adults aged 65 years and older is the population of interest. Sample is 735 individuals randomly selected from medicare rolls.Yes, one individual has missing smoking history, and the id is 545.

Q3. Problems 4 – 7 ask you to dichotomize the time to death according to death within 5 years of study enrollment or death after 5 years. Why is this valid? Provide descriptive statistics that support your answer.

Answer. Approx. 18% of the individuals who were part of the study were classified as observed to die while on study. Analysis has to be partitioned as 18% of the samples were not present till the end of obserbation period.

Table 1: Breakdown with respect to participants vital status

Status	Number of individuals
Not observed to die	602
Observed to die	133

Q4. Provide a suitable descriptive statistical analysis for selected variables in this dataset as might be presented in a table to be included in a manuscript exploring the association between smoking history in pack years and 5 year all-cause mortality in the medical literature. In addition to the two variables of primary interest, you may restrict attention to age, sex, weight, and prior history of cardiovascular disease, e.g., coronary heart disease (CHD), congestive heart failure (CHF), and stroke.

Answer. We are studying the association between overall mortality and smoking history in pack years.

Table 2: Descriptive statistics for selected variables with respect to participants smoking history.

	0-(n=321)	<=10-(n=73)	11-25-(n=99)	> 25-(n=241)
Age(yrs)	75.3(5.9, 66-99)	74.8(5.8, 68-95)	74.1(4.5, 65-89)	73.6(4.7,65-91)
Male(%)	38%	63%	47%	62%
Weight(lbs)	155.6(30.2, 74-264)	162.2(27.3, 107-242)	157.6(26.1, 107-243)	166(33.1,86-258)
Prior Stroke Diag(%)	3%	2%	9%	4%

	0-(n=321)	<=10-(n=73)	11-25-(n=99)	> 25-(n=241)
Prior CHF Diag(%)	5%	4%	6%	5%
Prior CHD Diag(%)	9%	6%	5%	9%
Death(%)	15%	12%	14%	24%

```
* one row excluded for missing smoking record
* descriptive statistics are presented in mean(std dev, min-max)
```

Q5. Perform a statistical analysis evaluating an association between smoking history in pack years and 5-year all-cause mortality by comparing mean pack years across groups defined by vital status at 5 years.

Answer. We do a 2 sample regular t-test as we have two groups, and we have a continous variable as response(dependent), and categorical variable as predictor.

As p-value is 0.004 (using critical value of 0.05), we reject the null hypothesis that mean difference across groups defined by vital status at 5 years would be the same. With 95% Confidence interval, the difference in mean pack years between the survivability groups is 2.89 to 15.666566. The best estimate of the difference, i.e. the point estimate is 9.28245 units, not that unsual to see.

Q6. Perform a statistical analysis evaluating an association between smoking history and 5-year all-cause mortality by comparing the probability of death within 5 years across groups defined by whether or not study participants have ever smoked. Note that a participant who never smoked has 0 pack years, as is indicated in the documentation file.

Answer. First, the observations follow random sampling, and sample size(735) is satifactory. Given the nature of analysis i.e. two categorical variables with expected counts as equal, we employ chi-square test of homogeneity.

	Obs to die	Otherwise	
Smoker	83	330	413
Non-Smoker	49	272	321
	132	602	734

```
* one row excluded for missing smoking record
```

As p-value is 0.11(much higher than our critical value of 0.05), we fail to reject our null hypothesis that the distribution(or proportion) would be same in 5-year all-cause mortality given the subject had smoked in the past or not.

Q7. Perform a statistical analysis evaluating an association between smoking history and 5-year all-cause mortality by comparing the odds of death within 5 years across groups defined by whether or not the study participants ever smoked.

Answer. As we are testing the independence between two nominal variables viz. being a smoker or not and 5-year all-cause mortality, we use the preferred Fisher's exact test, and other reason is that we started with fixed number of subjects.

Subsequently, we get a p-value of 0.09(higher than our critical value of 0.05) by using the same proportions as previously, hence, we fail to reject our null hypothesis that being a smoker/not and mortality is independent of each other.

Q8. Compare the association results of questions 5, 6 and 7, and briefly discuss any similarities and/or differences in the statistical inference.

Q9. Briefly discuss limitations of the above association analyses that you conducted in questions 5 – 7 for detecting associations between smoking history and death within 5 years. What analysis would you have

preferred a priori in order to answer the question about an association between smoking history and time to death? Why?

Answer.

Limitation of above associated analysis

- The analysis does not look into the variance inside the pack years
- One of the assumptions we made in t-test in Q5 was that variances are same across two samples; infact, on subsequent analysis using variance test, it happens to be not the same (p-value = 1.499e-08) for two groups.
- We have more than two groups as we identified, and t-test is not ideal in such scenarios, higher chances of Type I error

What analysis is preferred

- Linear regression or ANOVA is preferred. We would have preferred ANOVA as we have finer groupings of smoking behaviors which would help us to identify the sub-group which could be contributing to the rejection. Furthermore, we could also explore the interaction effects of other variables like age/sex/weight. Linear regression would accomplish similar results but could give us narrow analysis results whien using pack years variable, and additionally, having multiple predictor variables is a good motivation to use linear regression model.