Final Exam Sample
KEY

STA209-04: Applied Statistics

May 4, 2019

1) [40 pts] The National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) aims to conduct and support research on several common chronic conditions with hopes of improving the health and quality of life of those afflicted. One primary condition of interest to the NIDDK is diabetes.

One NIDDK study was interested in determining factors associated with diabetes among females at least 21 years old of Pima Indian heritage.

i) One factor commonly found to be associated with diabetes is BMI. The CDC defines an individual as obese if their BMI is greater than or equal to thirty. The following table cross-tabulates obese and non-obese patients by their diabetes diagnosis.

	Diabetic	**Non-Diabetic**
Obese	219	253
Non-Obese	49	247

Using the provided table, compute and interpret both the odds ratio and relative risk for having diabetes given a BMI ≥ 30.

$\hat{OR} = \frac{219*247}{49*253} = 4.36$, the odds of being diabetic are 4.36 times as high among obese individuals relative to non-obese individuals; $\hat{RR} = \frac{219/(219+253)}{49/(49+247)} = 2.80$, obese individuals are 2.8 times as likely to be diabetic as non-obese individuals.

ii) Suppose that these data were collected by recruiting individuals based on their diabetes diagnosis, as opposed to through a random sampling from the population of 21+ year old females of Pima Indian hertiage. Are both of the quantities computed in the previous question appropriate to use to quantify the strength of the association between obesity and diabetes? If not, explain why not and state which should be used.

Only the odds ratio should be used. The design described is a case-control study, for which relative risks are inappropriate to compute.

iii) Suppose that these data were collected by recruiting from among individuals without diabetes that were then followed forward in time for a number of years. Which of the quantities computed in (i) - odds ratio, relative risk, or both - could be used to quantify the strength of the association between obesity and diabetes? How would you characterize this study design?

Both quantities may be used. This describes a prospective study.

iv) Suppose that investigators were interested in determining whether being obese caused diabetes. Which of the previously described study designs would you recommend the investigators use? Why?

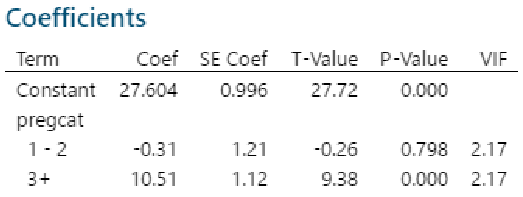
A prospective design should be recommended. In contrast to the case-control (i.e. restrospective) design, a prospective study tracks the development of diabetes among an initially non-diabetic population and classifies recruits according to some exposure, or cause. By following the recruits forward in time and recording how many develop the condition, we can better infer the causal relationship between the exposure and condition than if we were to retroactively assess exposure among a sample of diabetic and non-diabetic subjects.

v) In addition to BMI, investigators found that both the number of pregnancies (”None”, ”1-2”, ”3+”) and age of the subject were associated with a diabetes diagnosis. Given these two associations, researchers were interested in assessing whether age confounded the relationship between the number of pregnancies and a diabetes diagnosis. An ANOVA was then performed to determine whether number of pregnancies and age were associated. Complete the ANOVA table provided below to determine whether age is a confounding variable. Be sure to explain how the ANOVA results support or refute the idea that age is a confounder. (Note that the 95% critical value for the appropriate F-distribution is 3).

Source	DF	SS	MS	F-Value	P-Value
Group	2	106078-84192 = 21886	$\frac{21886}{2} = 10943$	$\frac{10943}{110.05} = 99.44$	Way smaller than 0.05
Error	765	84192	$\frac{84192}{765} = 110.05$		
Total	767	106078			

The F-test performed using the ANOVA table above indicates that there is an association between number of pregnancies and age. Since age is associated with both the number of pregnancies and diabetes diagnosis, it is a confounder for the association between number of pregnancies and diabetes diagnosis.

vi) An alternate approach to determining whether there is an association between age and number of pregnancies would be to fit a regression model with age as the outcome and number of pregnancies as a covariate. Doing so yields the following output:



Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	27.604	0.996	27.72	0.000	
pregcat					
1 - 2	-0.31	1.21	-0.26	0.798	2.17
3+	10.51	1.12	9.38	0.000	2.17

Using this output, determine the equation for the fitted regression line.

Age = 27.604 - 0.31∗(pregcat = ”1-2”) + 10.51∗(pregcat = ”3+”)

vii) Interpret the coefficient corresponding to ”1-2”. Your answer should directly include or reference the idea of a ”reference category”.

The age of an individual whose had 1-2 pregnancies is 0.31 years less than the reference, which is an individual with no pregnancies.

viii) Based on this model, what is the predicted age for an individual who has never been pregnant?

$27.604 - 0.31 * 0 + 10.51 * 0 = 27.604$