BIS 505b – Problem Set 4 **Due 11:59 PM Eastern, 4/11/2021**

Instructions: Follow the homework instructions outlined in the syllabus. Round your answers to 2 decimal places. Perform all tests at the α = 0.05-level and follow the steps of hypothesis testing.

Assignment

Question 1: This question analyzes data from a study conducted to look at the association between high-risk occupation (Yes/No) and presence of bladder cancer (Yes/No). The study reported the following estimates:

- The estimated risk of bladder cancer (Disease) in those with a high-risk occupation (Exposed) = $\hat{p}$ = 0.5983
- The estimated risk of bladder cancer (Disease) in those without a high-risk occupation (Unexposed) = $\hat{p}$ = 0.4329

a. [12] Report the odds of disease (presence of bladder cancer) in the exposed (those with a high-risk occupation). Report the odds of disease (presence of bladder cancer) in the unexposed (those without a high-risk occupation). Report the odds ratio of bladder cancer in those with a high-risk occupation vs. those without a high-risk occupation.

$$\widehat{odds}_e = \frac{\hat{p}_e}{1-\hat{p}_e} = \frac{0.5983}{1-0.5983} = 1.49$$

$$\widehat{odds}_u = \frac{\hat{p}_u}{1-\hat{p}_u} = \frac{0.4329}{1-0.4329} = 0.76$$

$$\widehat{OR} = \frac{\widehat{odds}_e}{\widehat{odds}_u} = \frac{1.49}{0.76} = 1.96$$

The odds of disease (presence of bladder cancer) in the exposed (those with a high-risk occupation) is 1.49.
The odds of disease (presence of bladder cancer) in the unexposed (those without a high-risk occupation) is 0.76.
The odds ratio of bladder cancer in those with a high-risk occupation vs. those without a high-risk occupation is 1.96.

b. [10] Suppose we were to fit a simple logistic regression model to these data, modeling bladder cancer status where bladder cancer=Yes is the event of interest and high-risk occupation is the only independent variable:

x = High risk occupation (0 = Non-high-risk occupation (reference), 1 = High-risk occupation)

This model does not control for any potential confounders. Report the fitted simple logistic regression model. Use your knowledge of what the intercept and slope represent in a logistic regression model along with your unadjusted results from part **(a)** to estimate the intercept and slope parameters of your fitted logistic regression model.

$b = log(\widehat{OR}) = log(1.96) = 0.67$

$a = log(\widehat{odds_u}) \ = \ log(0.76) = -0.27$

The fitted simple logistic regression model: $log(\frac{\hat{p}}{1-\hat{p}}) = -0.27 + 0.67x$

Question 2: In a study investigating maternal risk factors for congenital syphilis, syphilis is treated as a dichotomous response variable, where 1 represents the presence of disease in a newborn and 0 its absence. The estimated coefficients from a logistic regression model containing the explanatory variables (4 categorical, 1 continuous): cocaine or crack use during pregnancy (1=yes, 0=no (ref)), marital status (1=unmarried, 0=married (ref)), number of prenatal visits to a doctor, alcohol use during pregnancy (1=yes, 0=no (ref)), and level of education (1=less than high school, 0=high school or more (ref)) are listed in the table below.

	Estimate
Intercept	0.080
Cocaine/Crack Use x_1	1.354
Marital Status x_2	0.779
Number of Prenatal Visits x_3	-0.098
Alcohol Use x_4	0.723
Level of Education x_5	0.298

a. [5] As an expectant mother's number of prenatal visits to the doctor increases, does the **probability** that her child will be born with congenital syphilis increase or decrease? Explain.

The probability that her child will be born with congenital syphilis will decrease. Because the slope of Number of Prenatal Visits x_3 is -0.098 < 0, indicates that larger values of x_3 are related to a smaller log-odds, when log-odds decrease, the probability also decreases.

b. [7] Marital status is a dichotomous variable, where the value 1 indicates that a woman is unmarried and 0 indicates that she is married. What are the adjusted relative odds that a newborn will suffer from syphilis for unmarried versus married mothers? Interpret this odds ratio.

$\widehat{OR} = e^{b_2} = e^{0.779} = 2.18$

The adjusted relative odds that a newborn will suffer from syphilis for unmarried versus married mothers is 2.18, which means that the odds of a newborn will suffer from syphilis for unmarried mother is 118% higher compared to married mother when controlling for cocaine/crack use, number of prenatal visits, alcohol use, and level of education.

c. [7] Cocaine or crack use is also a dichotomous variable; the value 1 indicates that a woman used these drugs during pregnancy and 0 indicates that she did not. What is the estimated adjusted odds ratio that a child will be born with congenital syphilis for women who used cocaine or crack versus those who did not? Interpret this odds ratio.

$\widehat{OR} = e^{b_1} = e^{1.354} = 3.87$

The adjusted relative odds that a newborn will suffer from syphilis for women who used cocaine or crack versus those who did not is 3.87, which means that the odds of a newborn will suffer from syphilis for women who used cocaine or crack is 287% higher compared to those who did not when controlling for marital status, number of prenatal visits, alcohol use, and level of education.

d. [10] The estimated coefficient of cocaine or crack use has standard error = 0.162. Construct a 95% confidence interval for the population adjusted odds ratio of cocaine/crack use.

95% CI for $log(OR), \beta_1$: $b_1 \pm z_{1-\frac{\alpha}{2}} s_{b_1} = 1.354 \pm 1.96 \times 0.162 = (1.04, 1.67)$

95% CI for OR, β_1: $(e^{c_L}, e^{c_U}) = (e^{1.04}, e^{1.67}) = (2.83,\ 5.31)$

The 95% confidence interval for the population adjusted odds ratio of cocaine/crack use is $(2.83,\ 5.31)$

e. [5] Estimate the probability that a newborn will be born with congenital syphilis if the child's mother is unmarried, has less than a high school level of education, uses cocaine and alcohol during pregnancy, and has 2 prenatal visits to the doctor. Assume these values are in the range of the data used to fit the model.

The fitted multiple logistic regression model is:

$$log(\frac{\hat{p}}{1-\hat{p}}) = 0.080 + 1.354x_1 + 0.779x_2 - 0.098x_3 + 0.723x_4 + 0.298x_5$$

Probability that a newborn will be born with congenital syphilis is

$$\hat{p} = \frac{e^{0.080+1.354x_1+0.779x_2-0.098x_3+0.723x_4+0.298x_5}}{1+e^{0.080+1.354x_1+0.779x_2-0.098x_3+0.723x_4+0.298x_5}}$$

Plug in $x_1 = 1, x_2 = 1,\ x_3 = 2,\ x_4 = 1,\ x_5 = 1$

$$e^{0.080+1.354x_1+0.779x_2-0.098x_3+0.723x_4+0.298x_5}$$
$$= e^{0.080+1.354\times1+0.779\times1-0.098\times2+0.723\times1+0.298\times1} = 50.15$$

$$\hat{p} = \frac{50.15}{1+50.15} = 0.98$$

The estimated probability that a newborn will be born with congenital syphilis is 0.98 with the stated conditions.

f. [5] Estimate the probability that a newborn will be born with congenital syphilis if the child's mother is married, has greater than a high school education, does not use crack/cocaine during pregnancy, does not consume alcohol during pregnancy, and has 14 prenatal visits to the doctor. Assume these values are in the range of the data used to fit the model.
Plug in $x_1 = 0, x_2 = 0,\ x_3 = 14,\ x_4 = 0,\ x_5 = 0$ to the same model as question e

$$e^{0.080+1.354x_1+0.779x_2-0.098x_3+0.723x_4+0.298x_5}$$
$$= e^{0.080-0.098\times 14} = 0.27$$

$$\hat{p} = \frac{0.27}{1+0.27} = 0.22$$

The estimated probability that a newborn will be born with congenital syphilis is 0.22 with the stated conditions.

Question 3: Suppose you are interested in studying intravenous (IV) drug use among high school students in the United States. Drug use is characterized as a dichotomous variable, where 1 indicates that an individual has injected drugs within the past year and 0 that he/she has not. (4 binary + 1 continuous) Factors that might be related to drug use are: instruction about the human immunodeficiency virus (HIV) in school (where 1 indicates that instruction was received and 0 indicates that instruction was not received (ref)), age of the student (years), sex (1=male, 0=female (ref)), and general knowledge about HIV, including the various modes of transmission and ways to reduce risk (1=possesses good general knowledge, 0=does not (ref)). The estimated coefficients and standard errors from a logistic regression model containing each of these explanatory variables as well as the interaction between HIV instruction and sex are displayed in the table below.

	Estimate	Standard Error
Intercept	-1.183	0.859
HIV INSTRUCTION x_1	0.039	0.421
AGE x_2	-0.164	0.092
SEX x_3	1.212	0.423
HIV KNOWLEDGE x_4	-0.187	0.048
HIV INSTRUCTION*SEX x_5	-0.663	0.301

a. [10] Determine if the effect of HIV instruction on IV drug use is significantly different in males vs. females.

(1) State the null and alternative hypotheses
$H_0: \beta_5 = 0$ vs. $H_1: \beta_5 \neq 0$

(2) Specify the significance level, α = 0.05

(3) Compute the test statistic

$$z = \frac{b}{s_b} = \frac{-0.663}{0.301} = -2.20 \sim N(0,1)$$

(4) Generate the decision rule
Given α = 0.05,
Reject H_0 if $|z| \geq z_{1-\frac{\alpha}{2}} = z_{0.975} = 1.96$ or if $p \leq 0.05$

(5) Draw a statistical conclusion and state the conclusion in words in the context of the problem.

$|z| = 2.20 > 1.96 \ \rightarrow Reject\ H_0$
$or\ p = P(Z \geq 2.20) = 0.028 \leq 0.05 \ \rightarrow Reject\ H_0$

Conclusion: There is evidence to reject H_0 and conclude that the effect of HIV instruction on IV drug use is significantly different in males vs. females.

b. [8] Write the equation of the fitted logit (log odds) of IV drug use for males. Write the equation of the fitted logit (log odds) of IV drug use for females.

The fitted logistic model is:

$$log(\frac{\hat{p}}{1-\hat{p}}) = -1.183 + 0.039\ HIV\ Instruction\ - 0.164\ Age\ + 1.212 \times Sex - 0.187\ HIV\ Knowledge - 0.663\ HIV\ Instruction \times Sex$$

The fitted logit (log odds) of IV drug use for males ($x_3 = 1$)

$$log(\frac{\hat{p}}{1-\hat{p}}) = -1.183 + 0.039\ HIV\ Instruction\ - 0.164\ Age\ + 1.212 \times (1) - 0.187\ HIV\ Knowledge - 0.663\ HIV\ Instruction \times (1)$$
$$= 0.029 - 0.624\ HIV\ Instruction\ - 0.164\ Age\ - 0.187\ HIV\ Knowledge$$

The fitted logit (log odds) of IV drug use for females ($x_3 = 0$)

$$log(\frac{\hat{p}}{1-\hat{p}}) = -1.183 + 0.039\ HIV\ Instruction\ - 0.164\ Age\ + 1.212 \times (0) - 0.187\ HIV\ Knowledge - 0.663\ HIV\ Instruction \times (0)$$
$$= -1.183 + 0.039\ HIV\ Instruction\ - 0.164\ Age - 0.187\ HIV\ Knowledge$$

c. [10] Report the adjusted OR for the effect of HIV instruction on IV drug use in males. Report the adjusted OR for the effect of HIV instruction on IV drug use in females. In the model above, is HIV instruction significantly associated with IV drug use in females?

The adjusted OR for the effect of HIV instruction on IV drug use in males.
$\widehat{OR} = e^{-0.624} = 0.54$

The adjusted OR for the effect of HIV instruction on IV drug use in females.
$\widehat{OR} = e^{0.039} = 1.04$

Hypothesis test of if HIV instruction significantly associated with IV drug use in females:

(1) State the null and alternative hypotheses
$H_0: \beta_1 = 0$ vs. $H_1: \beta_1 \neq 0$

(2) Specify the significance level, α = 0.05

(3) Compute the test statistic

$$z = \frac{b}{s_b} = \frac{0.039}{0.421} = 0.09\ \sim N(0,1)$$

(4) Generate the decision rule
Given α = 0.05,
Reject H_0 if $|z|\ \geq z_{1-\frac{\alpha}{2}} = z_{0.975} = 1.96$ or if $p \leq 0.05$

(5) Draw a statistical conclusion and state the conclusion in words in the context of the problem.

$|z| = 0.09 < 1.96 \ \rightarrow Fail\ to\ reject\ H_0$
$or\ p = P(Z \geq 0.09) = 0.93 > 0.05 \ \rightarrow Fail\ to\ reject\ H_0$
Conclusion: There is evidence to fail to reject H_0 and conclude that HIV instruction is not significantly associated with IV drug use in females

d. [5] Controlling for all the other variables included in the model, as a student becomes older, does the probability that he or she has used intravenous drugs in the past year increase or decrease? Explain.

The probability that he or she has used intravenous drugs in the past year will decrease because the estimated slope of age x_2 is -0.164 <0, indicates that larger values of x_2 are related to a smaller log-odds, when log-odds decrease, the probability also decreases.

e. [6] Using the model reported above, what is the estimated probability of IV drug use for a 13-year old male who did not receive HIV instruction and who does not possess good general knowledge about HIV? Assume these values are in the range of the data used to fit the model.

The fitted logistic model is:

$$log(\frac{\hat{p}}{1-\hat{p}}) = -1.183 + 0.039\ HIV\ Instruction\ - 0.164\ Age\ + 1.212 \times Sex - 0.187\ HIV\ Knowledge - 0.663\ HIV\ Instruction \times Sex$$

$$= -1.183 + 0.039\ x_1\ - 0.164\ x_2\ + 1.212 \times x_3 - 0.187\ x_4 - 0.663 x_1 x_3$$

Probability of IV drug use is

$$\hat{p} = \frac{e^{-1.183+0.039\ x_1\ -0.164\ x_2\ +1.212\times x_3-0.187\ x_4-0.663x_1x_3}}{1 + e^{-1.183+0.039\ x_1\ -0.164\ x_2\ +1.212\times x_3-0.187\ x_4-0.663x_1x_3}}$$

Plug in $x_1 = 0, x_2 = 13,\ x_3 = 1,\ x_4 = 0$

$$e^{-1.183+0.039\ x_1\ -0.164\ x_2\ +1.212\times x_3-0.187\ x_4-0.663x_1x_3}$$
$$= e^{-1.183\ -0.164\ \times 13\ +1.212\times 1} = 0.12$$

$$\hat{p} = \frac{0.12}{1 + 0.12} = 0.11$$

The estimated probability of IV drug use for a 13-year old male who did not receive HIV instruction and who does not possess good general knowledge about HIV is 0.11.