BIS 505b – Problem Set 5 **Due 11:59 PM Eastern, 4/25/2021**

Instructions: Follow the homework instructions outlined in the syllabus. Round your answers to 3 decimal places. Perform all tests at the α = 0.05-level and follow the steps of hypothesis testing. You may use **R** as a calculator to perform intermediate calculations or to compute p-values. However, **R** should **not** be used to run any regression models in this assignment.

Assignment

Question 1: The estimated numbers of new diagnoses of HIV/AIDS for 2018 indicate that young people under the age of 25 account for 21% of new cases and those aged 25-29 account for 20% of new cases (HIV Surveillance Report, Centers for Disease Control and Prevention, 2018)[1]. Given a lag between infection and diagnosis, the estimates suggest that nearly a quarter of those with HIV are infected as adolescents and young adults. Homeless youth have been recognized as one group of young people at particularly high risk for contracting HIV and other sexually transmitted diseases. A study was conducted to identify potential individual and environmental protective factors for sex risk behavior among homeless youth.

a. [5] A Poisson regression model was used to investigate how religious service attendance, expectations for the future, and decision-making skills were related to the number of sexual partners in the past three months, controlling for age, gender and ethnicity. Justify the use of Poisson regression in this setting.
Because the number of sexual partners is a count response, consists of discrete, non-negative integer data, Poisson regression is suitable for modeling the linear relationship between the log of mean count response and regressors.

b. [4] Given the time span over which the number of sexual partners is recorded equals 3 months for all participants, can we directly model the count?

Yes, because the time period is same for all participants, the time at risk is same for all participants.

c. [9] The authors presented the following estimated slopes and 95% confidence intervals for the slopes from a Poisson regression model controlling for age, gender and race/ethnicity. Use the results below to calculate adjusted mean ratios. Interpret each. What can you conclude about each adjusted effect? For each, state your statistical conclusions and your conclusion in the context of the problem.

	Estimate	**95% CI**	**Adjusted mean ratios**	**Statistical conclusions**
Have decision-making skills	-0.44	(-0.86, -0.03)	0.644	Significant
Have expectations for the future	-0.03	(-0.06, 0.01)	0.970	Not significant
Attend religious services	0.27	(0.11, 0.44)	1.310	Significant

[1] https://www.cdc.gov/hiv/library/reports/hiv-surveillance.html

Controlling for age, gender and race/ethnicity:

1) The mean number of sexual partners in the past three months among homeless youth who have decision-making skills is 35.4% lower than those who not (ref), adjusted mean ratio $= e^{b_1} = e^{-0.44} = 0.644$. Because the 95% CI (-0.86, -0.03) does not include null, Wald tests of slope show that there is evidence to fail to reject H_0: $\beta_1 = 0$ and conclude that the mean number of sexual partners in the past three months is significantly different among homeless youth who have decision-making skills vs. those who not.

2) The mean number of sexual partners in the past three months among homeless youth who have expectations for the future is 3.0% lower than those who not (ref), adjusted mean ratio $= e^{b_2} = e^{-0.03} = 0.970$. Because the 95% CI (-0.06, 0.01) includes null, Wald tests of slope show that there is evidence to reject H_0: $\beta_2 = 0$ and conclude that the mean number of sexual partners in the past three months is not significantly different among homeless youth who have expectations for the future vs. those who not.

3) The mean number of sexual partners in the past three months among homeless youth who attend religious services is 31.0% higher than those who not (ref), adjusted mean ratio $= e^{b_3} = e^{0.27} = 1.310$. Because the 95% CI (0.11, 0.44) does not include null, Wald tests of slope show that there is evidence to fail to reject H_0: $\beta_3 = 0$ and conclude that the mean number of sexual partners in the past three months is significantly different among homeless youth who attend religious services vs. those who not.

Question 2: Do No-Smoking-at-Work Policies Keep Smokers at Home?

Several regular smokers were selected from a larger Minnesota survey about smoking habits. All reported smoking about the same number of cigarettes per day in recent months. Each was asked to report on the number of cigarettes smoked the day before (January 26) during the same 2-hour period. They were also asked whether they had been at work or at home or elsewhere during that 2-hour period.

Six of the smokers were either at work or at home during the 2-hour period. Data on the number of cigarettes smoked by these 6 smokers in the 2-hour period and whether at home or at the office (which requires smokers to go outside in Minnesota in January) are shown in the table below.

Table. Data on six smokers for one 2-hour period. The response variable is the number of cigarettes smoked during the period and the explanatory variable [Location] equals 0 if the person is at home or 1 if the person is at work. Note that since exposure is identically equal to one (i.e., one 2-hour period) for everyone, no offset is needed in the Poisson regression model.

Number of Cigarettes	3	0	0	1	2	1
Location	0	1	1	1	0	0

a. [4] Write the equation of the population simple Poisson regression model.

$$log(\mu) = \alpha \ + \beta_1 \ Location$$

b. [3] Hypothesize the direction for the slope of "Location" in this model. That is, if "Location" is coded as 1=work, 0=home, what do you expect the sign (i.e., positive or negative) of the estimated slope (b_1) for the covariate "Location" to be if we were to fit a Poisson regression model to these data? Explain.
The sign of estimated slope (b_1) for the covariate "Location" would be negative because that means the mean ratio < 1, indicating that the mean number of cigarettes smoked in the 2-hour period among smokers at office is smaller than those who at the home because of the No-Smoking-at-Work Policies.

c. [3] Calculate the average number of cigarettes smoked during a 2-hour period at each location. Does the observed difference support your hypothesis in part **(b)**?

Location	Mean Number of Cigarettes/2h ($\hat{\mu}$)
At work 1	$\frac{1}{3} = 0.333$
At home 0	$\frac{6}{3} = 2.000$

Yes, because the mean number of cigarettes smoked during a 2-hour period at work is smaller than that at home ($0.333\ vs.\ 2$)

d. [10] Use the group sample means that you computed in part **(c)** to find the equation of the fitted simple Poisson regression model for these data. Show your work.

$$b_1 = log(\hat{\mu}_1) - log(\hat{\mu}_0) = log(\frac{\hat{\mu}_1}{\hat{\mu}_0}) = log(\frac{0.333}{2.000}) = -1.791$$
$$a = log(\hat{\mu}_0) = log(2) = 0.693$$
$$log(\hat{\mu}) = a + b_1 x = 0.693 - 1.791x$$

e. [4] Interpret the exponentiated slope estimate from your model in the context of these data.
The mean number of cigarettes smoked during a 2-hour period at work is smaller than that at home is 83.3% lower than those who at home, with mean ratio $= e^{b_1} = e^{-1.791} = 0.167$.

Question 3: Consider the small data set consisting of responses $Y = \{2, 3, 8, 9, 1\}$.

a. [5] What is the mean? What is the log of the mean? What is the mean of the logs of each data point?

$$\hat{\mu} = \frac{2+3+8+9+1}{5} = 4.600$$
$$log(\hat{\mu}) = log(4.600) = 1.526$$
$$\text{Mean}[log(Y)] = \frac{log2+log3+log8+log9+log1}{5} = 1.214$$

The mean is 4.600, log of the mean is 1.526, the mean of the logs of each data point is 1.214

b. [5] Suppose we fit a Poisson regression model of the count Y. The model includes only an intercept term and no independent variables. Report the equation of this fitted model.

$$log(\hat{\mu}) = a_1 = 1.526$$

c. [5] Now we fit a simple linear regression model with $\log(Y)$ as the response. This model also includes only an intercept term and no independent variables. Report the equation of this fitted model.

$\log(Y) = a_2 = 1.214$

d. [5] Are the estimated models described in **(b)** and **(c)** the same? Why or why not?

The 2 models are different, because the intercept term of the Poisson regression model is the log of the mean, while the intercept term of the simple linear regression model is the mean of the logs of each data point ($1.526\ vs.\ 1.214$).

Question 4: A Poisson regression analysis of the rate of episodes of severe hypoglycemia 低血糖 was performed using data from the long-term follow-up of patients enrolled in the Diabetes Control and Complications Trial. The number of severe hypoglycemia events [episodes] was modeled and the years of study follow-up [followup] was accounted for through an offset term. The following variables were included as predictor variables in the model:

- intensive*:* Treatment group assignment (1 = intensive blood glucose management, 0 = conventional blood glucose management)
- insulin: Baseline daily insulin dose
- duration: Number of months duration of diabetes upon entry into the study (range: 12 to 180)
- female: An indicator for female (1) vs. male (0)
- adult: An indicator variable for adult (1, ≥ 18 years) vs. adolescent (0) on entry
- hba: HbA1c level, which is a measure of the overall level of blood glucose control. The lower the level of HbA1c, the greater the risk of hypoglycemia when the blood glucose falls too low.
- hxcoma: An indicator variable for history of coma and/or seizure 昏迷、发作 prior to entry (1 = yes, 0 = no)

The fitted Poisson regression model is reported below:

	Estimate	Standard Error	*z*-statistic	P-value
Intercept	-0.957	0.217	-4.410	<0.0001
Intensive	-1.074	0.049	-21.918	<0.0001
Insulin	0.005	0.099	0.052	0.959
Duration	0.002	0.0006	2.500	0.012
Female	0.178	0.042	4.238	<0.0001
Adult	-0.598	0.066	-9.061	<0.0001
Hba	-0.034	0.015	-2.267	0.023
Hxcoma	0.601	0.069	8.710	<0.0001

a. [5] Complete the table of results above by filling in the z-statistic and the p-value columns.

b. [5] Write the equation of the fitted Poisson regression model.

$$log(\hat{\lambda}) = a + \beta_1 Intensive + \beta_2 Insulin + \beta_3 Duration + \beta_4 Female + \beta_5 Adult + \beta_6 Hba + \beta_7 Hxcoma$$

$$= -0.957 - 1.074 Intensive + 0.0051 Insulin + 0.0015 Duration + 0.178 Female - 0.598 Adult - 0.034 Hba + 0.601 Hxcoma$$

c. [5] Interpret the effect of "Intensive" in the model above. Perform a hypothesis test to determine if this is an important predictor in the model. Perform the steps of hypothesis testing. Controlling for all the other variables in the model, the rate of episodes of severe hypoglycemia in the intensive blood glucose management group is 65.8% lower than that in the conventional blood glucose management group (ref), adjusted rate ratio = $e^{b_1} = e^{-1.074} = 0.342.$

(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

$$z = \frac{b}{s_b} = \frac{-1.074}{0.049} = -21.918 \sim N(0, 1)$$

(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.
$|z| = 21.918 > 1.96 \rightarrow Reject\ H_0$
$or\ p = P(Z \geq 21.918) =< 0.0001 \rightarrow Reject\ H_0$

Conclusion: There is evidence to reject H_0 and conclude that rate of episodes of severe hypoglycemia is significantly different in the intensive blood glucose management group vs. conventional blood glucose management group.

d. [10] Interpret the remaining exponentiated slope estimates from the fitted model in the context of these data and interpret the result of the hypothesis test (i.e., the p-value you computed above) in the context of this problem. The full steps of hypothesis testing are not required for this question.

Controlling for all the other variables in the model:
1) As baseline daily insulin dose increases, the rate of episodes of severe hypoglycemia increases. A 1-dose increase in baseline daily insulin dose increases the rate of episodes of severe hypoglycemia by 0.5%, adjusted rate ratio = $e^{b_2} = e^{0.005} = 1.005$. The Wald test of the slope show that there is evidence to **fail to reject** $H_0: \beta_2 = 0$ and conclude that there is **not** a significant linear relationship between the rate of episodes of severe hypoglycemia and baseline daily insulin dose (p-value 0.959).

2) As number of months duration of diabetes upon entry into the study increases, the rate of episodes of severe hypoglycemia increases. A 1-month increase in duration of diabetes upon entry into the study increases the rate of episodes of severe hypoglycemia by 0.2%, adjusted rate ratio = $e^{b_3} = e^{0.002} = 1.002$. The Wald test of the slope show that there is evidence to **reject** $H_0: \beta_3 = 0$ and conclude that there is a significant linear relationship between the rate of

episodes of severe hypoglycemia and number of months duration of diabetes upon entry into the study (p-value 0.012).

3) The rate of episodes of severe hypoglycemia in females is 19.5% higher than that in males (ref), adjusted rate ratio = $e^{b_4} = e^{0.178} = 1.195$. The Wald test of the slope show that there is evidence to **reject** $H_0: \beta_4 = 0$ and conclude that the rate of episodes of severe hypoglycemia is significantly different in females vs. in males (p-value <0.0001).

4) The rate of episodes of severe hypoglycemia in adult is 45.0% lower than that in adolescent (ref), adjusted rate ratio = $e^{b_5} = e^{-0.598} = 0.550$. The Wald test of the slope show that there is evidence to **reject** $H_0: \beta_5 = 0$ and conclude that the rate of episodes of severe hypoglycemia is significantly different in adult vs. in adolescent (p-value <0.0001).

5) As HbA1c level increases, the rate of episodes of severe hypoglycemia decreases. A 1-unit increase in HbA1c level decreases the rate of episodes of severe hypoglycemia by 3.3%, adjusted rate ratio = $e^{b_6} = e^{-0.034} = 0.967$. The Wald test of the slope show that there is evidence to **reject** $H_0: \beta_6 = 0$ and conclude that there is a significant linear relationship between the rate of episodes of severe hypoglycemia and HbA1c level (p-value 0.023).

6) The rate of episodes of severe hypoglycemia in those have history of coma and/or seizure prior to entry is 82.4% higher than that in those without (ref),
adjusted rate ratio = $e^{b_7} = e^{0.601} = 1.824$. The Wald test of the slope show that there is evidence to **reject** $H_0: \beta_7 = 0$ and conclude that the rate of episodes of severe hypoglycemia is significantly different in those have history of coma and/or seizure prior to entry vs. those without (p-value <0.0001).

e. [5] Estimate the fitted response for the following patient below:

Subject ID	Intensive	Insulin	Duration	Female	Adult	Hba	Hxcoma
001	1	0.336	110	0	1	9.74	1

$log(\hat{\lambda}) = -0.957 - 1.074 \times 1 + 0.0051 \times 0.336 + 0.0015 \times 110 + 0.178 \times 0 - 0.598 \times 1 - 0.034 \times 9.74 + 0.601 \times 1 = -2.15$

$\hat{\lambda} = e^{-2.15} = 0.116$

According to our model, this patient is expected to have episodes of severe hypoglycemia 0.116 times/year.

f. [8] Suppose the follow-up time variable was instead recorded in months instead of years. Would this affect the effect of "Intensive" in the model that you described in part **(c)**? If so, how? If not, why not? Would this affect the fitted response for the individual considered in part **(e)**? If so, how? If not, why not?

1) Different unit of follow-up time variable will not affect the effect of “Intensive” in the model. Because the effect is measured by slope $\beta_1 \; = \; log(\frac{\lambda_1}{\lambda_0})$ which is the log rate ratio that offsets the effect of unit of follow-up time.

2) Different unit of follow-up time variable will affect the fitted response for the individual considered in part **(e).** Because the fitted response is the $\hat{\lambda} \; = \frac{y_i}{t_i}$, which is the mean ratio that depends on the unit of follow-up time.