**Take-Home Final Exam- EPI 740 – Epi Modeling, Fall 2015**

**NAMES (All Group Members)**:

XINYI ZHAO, XI LIU, YONGJIA SONG

**Student ID(s):** 2168379, 2170248, 2168363

**DUE: Thursday December 10, 2015, at 12:00pm**

Contents

* **take home final 2015.docx** This file with instructions, exam scenario, and questions
* **final2015.sas7bdat** SAS dataset for the exam

Grading

There are 16 questions on this exam. You have **one of two choices.** You need to:

**Circle the Choice** you are taking and **write your Student ID** **next to your choice**:

**I. If you turn in answers to all 16 questions, you can earn up to 100 points on this exam and**

**you are eligible for an A or A - grade in the course. Your answer to question 16 (worth 15**

**points) should be limited to no more than 6 typewritten pages; also no computer output is**

**allowed for this question. Student ID #s: 2168379, 2170248, 2168363**

**II. If you turn in answers to questions 1-15 (but not 16), you can earn up to 100 points on this**

**exam, but you are not eligible for an A or A - grade in the course, i.e., you may earn up to a**

**B + in the course. Student ID #s: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

* **The grades for questions 1-15 will be based on a subset of questions that total either 85 points (**choice **I) or 100 points (**choice **II) out of the points specified over all questions. Question 16 counts 15 points if you choose** choice **I.** Since the selection of the questions to be graded will not be made until after you turn in your exam, and because most of the responses build upon one another, you should answer all of questions 1-15 regardless of whether you have chosen choice **I** or **II**.

Groups

* If you are not a PhD student in Epidemiology, you can work in a group of up to three persons.
* The members of your group should be “pre-registered” by sending your names by email to Dr. R by Sunday, November 22nd. You cannot change the members of your group after this date without speaking with us.
* Your group should turn in a single report regardless of whether or not all group members have selected different choices (I or II).
* If you are a PhD student in Epidemiology, you must work on your own.

Completing the exam

* Make sure to turn in this page as the cover page for your answers.
* Put only student ID #s, but no names, on all other pages of your answers.
* Your answers should be **typed**, not handwritten, although formulae and model statements may be hand-written for convenience.
* Make sure to **cut and paste relevant edited computer output and program code where appropriate** in your answers. **Do not** simply attach computer output and expect us to find the relevant portions (i.e., like finding a needle in a haystack). Please provide all other computer code in an appendix.
* When producing such relevant output, particularly graphs, it is important to **label** your output so that we know what figure/curves/table represents what phenomenon. A correct figure with inadequate labeling might lose points!
* Either put your completed exam in the box for final exams outside Dr. K’s office or give your completed exam to Mazie Tinsley (Admin Assistant for Dr K, whose office space is located at CNR 3rd floor Epi Dept. entrance).

Help!

* If you have any questions, you may email either any TA or the instructors (Dr. K: [dkleinb@emory.edu](mailto:dkleinb@emory.edu) and Dr. R: [esrose2@emory.edu](mailto:esrose2@emory.edu))
* **You cannot consult with anyone else outside your group about the exam.**

When turning in your answers, please make sure to sign the following pledge:

**I have neither given nor received aid on this exam. Moreover, I have followed the above instructions and have not communicated about the test or any examples therein with any student(s) who is not my group member(s).**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Signatures of all group members

Scenario

In the United States, men who have sex with men (MSM) represent the majority of people living with HIV and those acquiring new infections. There are a number of behavioral, biological, and historical reasons for this health inequity. A major contributing factor to the HIV risks that MSM face are the high biological transmission risk of anal sex without a condom. MSM who have more male partners for anal sex are at increased HIV risk, particularly when condoms are not used. MSM who live in, and draw one’s sex partners from, a high-prevalence neighborhood (represented as ZIP code in this exam) are also placed at higher risk for acquiring HIV. It has also been generally historically the case that neighborhoods with high levels of MSM HIV prevalence tend to also be areas where more numbers of MSM live. This could potentially lead to an increase in the number of partners that a man might have. This same risk pattern is generally true for bacterial sexual transmitted infections (STI), mostly importantly gonorrhea and chlamydia. Indeed, MSM are also overrepresented in the United States in the incidence of these STI.

There is basic scientific evidence to suggest that existing infection with a bacterial STI also *causes* increased susceptibility to HIV infection among MSM. This is in part because STI may cause inflammation in the anogenital region, which recruits to the region the types of blood cells (T-cells) that HIV infects. Quantifying this causal effect in population-based studies is extremely challenging because the same risk behaviors that result in STI also lead to HIV infection. Unadjusted associations between STI and HIV tend to be very large, and it’s unclear how much of the association is due to causality vs. confounding bias. A second challenge is that both STI and HIV are relatively rare outcomes, even in the most at-risk populations. Thus studying this association requires well-designed and large observational studies.

You work in the Fulton County Health Department, which operates an STI clinic and tests for both STI and HIV, and decided to establish a cohort study of MSM using individuals who come in for STI testing. The purpose of this study is to follow HIV-negative men for HIV incidence, following an initially positive or negative STI test result at the enrollment visit. Over the span of 1 year, the study enrolled 896 MSM who tested positive for an STI (gonorrhea and/or chlamydia) and 938 MSM who tested negative.

The cohort study has a two-year follow-up period, with 5 visits at t = 0 (enrollment visit), 6, 12, 18, and 24 months after enrollment. At each visit, men underwent HIV testing and completed questionnaires. For men who acquired HIV (“seroconverted”), time of acquisition before the study visit was estimated based on a standardized algorithm. At the baseline visit, you collected data on participants’ *age, income*, and *sexual behaviors.*

Relevant variables for this analysis in **final2015.sas7bdat** include:

* **STI** 1 = the participant has an STI at enrollment; 0 = no STI at enrollment
* **HIV\_inc** 1 = the participant acquired HIV (“seroconverted”) during follow-up; 0 = censored.
* **HIV\_PT\_mo** Person-time in months.
* **male\_partners** Number of male anal sex partners in the 6 months before enrollment (range 1-20)
* **condom\_use** 1 = consistent user of condoms, based on a standardized scale of condom use; 0 = not
* **ZIP\_risk** 1 = resident in a ZIP code that has a high prevalence of both HIV and STI; 0 = not

If you’d like extra background information on these topics and this type of study, see:

* Pathela P, Braunstein SL, Blank S, Schillinger JA. HIV incidence among men with and those without sexually transmitted rectal infections: estimates from matching against an HIV case registry. Clin Infect Dis. 2013 Oct;57(8):1203-9.
* Kelley CF, Vaughan AS, Luisi N, Sanchez TH, Salazar LF, Frew PM, Cooper HL, Diclemente R, del Rio C, Sullivan PS, Rosenberg ES. The Effect of High Rates of Bacterial Sexually Transmitted Infections on HIV Incidence in a Cohort of Black and White Men Who Have Sex with Men in Atlanta, Georgia. AIDS Res Hum Retroviruses. 2015 Jun;31(6):587-92.

1. ***(5 points)*  
   a.** Based only on the information provided above, complete the arrows for a directed acyclic graph (DAG) of the hypothesized relationships between the variables **STI, hiv\_inc, condom\_use, ZIP\_risk, and male\_partners.**

*Note:* Presume that the individuals in the study contribute a negligible amount to the prevalence in their ZIP codes:

STI

HIV

male\_partners

condom\_use

ZIP\_risk

**b.** Based on this DAG, which factors would you recommend including for control of confounding when considering the possible causal effect of STI on HIV?

Resident in a ZIP code that has a high prevalence of both HIV and STI; number of male anal sex partners in the 6 months before enrollment; consistent user of condoms.

1. ***(10 points)***  
   Before getting into multivariable analysis of the data, it is important to first understand the distribution of each variable and its bivariate association with incident HIV infection.

Complete the table below using SAS PROC MEANS and/or other tools you’ve learned in this course and earlier ones.

*Note:*

* + For the 2 types of requested confidence intervals, in the space provided after the table, name the methods you used for their calculation.
  + For the continuous variable **male\_partners**, choose a reasonable number of categories and referent group to use for your analysis. In the space provided after the table, briefly describe how you determined cutpoints for the variable and defined the categories. Six lines are provided, but this does not necessarily imply that all should be used! For this question, you only need to report one categorization method.
  + Person-time should be reported in years, not months.
  + As an example, the rows for **condom\_use** have already been completed.

Table 1: Characteristics of and HIV incidence among a cohort of 1,629 MSM in Washington, DC

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Characteristic** | *Participants* | | *HIV Infections* | *Person-years* | *Incidence rate* | | *Incidence rate ratio* | | |
|  | *n* | *(%)* | *#* | *(PY)* | *rate per 100 PY* | *(95% CI)* | *RR* | | *(95% CI)* |
| **Overall** | 1,834 | (100.0) | 91 | 2,687 | 3.39 | (2.74, 4.14) | -- | | -- |
|  |  |  |  |  |  |  |  | |  |
| **STI at enrollment** |  |  |  |  |  |  |  | |  |
| Yes | 896 | (48.9) | 67 | 1,251 | 5.36 | (4.22, 6.80) | 3.20 | | (2.01, 5.11) |
| No | 938 | (51.1) | 24 | 1,436 | 1.67 | (1.12, 2.49) | *ref.* | | |
|  |  |  |  |  |  |  |  | |  |
| **Condom use** |  |  |  |  |  |  |  | |  |
| Yes | 1,041 | (56.8) | 8 | 1,571 | 0.51 | (0.22, 1.00) | *ref.* | | |
| No | 793 | (43.2) | 83 | 1,116 | 7.44 | (5.92, 9.22) | 14.59 | (7.39, 32.38) | |
|  |  |  |  |  |  |  |  | |  |
| **ZIP risk** |  |  |  |  |  |  |  | |  |
| High prevalence | 848 | (46.2) | 62 | 1,267 | 4.89 | (3.82, 6.28) | 2.40 | | (1.54, 3.72) |
| Low prevalence | 986 | (53.8) | 29 | 1,420 | 2.04 | (1.42, 2.94) | *ref.* | | |
|  |  |  |  |  |  |  |  | |  |
| **Male partners** |  |  |  |  |  |  |  | |  |
| 1-3 | 510 | (27.8) | 18 | 754 | 2.39 | (1.50, 3.79) | *ref.* | | |
| 4-6 | 460 | (25.1) | 27 | 644 | 4.19 | (2.88, 6.11) | 1.76 | | (0.97, 3.19) |
| 7-9 | 411 | (22.4) | 20 | 629 | 3.18 | (2.05, 4.93) | 1.33 | | (0.70, 2.52) |
| 10-20 | 453 | (24.7) | 26 | 660 | 3.94 | (2.68, 5.79) | 1.65 | | (0.90, 3.01) |
|  |  |  |  |  |  |  |  | |  |
|  |  |  |  |  |  |  |  | |  |

**2. continued…**

**Method for confidence intervals:**

Poisson regression.

**Male partners categorization:**

The cutpoints are determined as the 1st quartile, medium and the 3rd quartile of number of male partners to enable the sample size of each level relatively similar, so as to avoid problems of sparse sample size when doing analysis.

1. ***(5 points)***  
   In no more than 2 sentences of plain language (i.e. so that your grandparent might understand, presuming he/she has no epidemiological training), interpret the meaning of the two values of the “incidence rate per 100 PY” for those with and without an STI at enrollment. There is no need to explain the 95% CI for these values.

In each year, among 100 persons with sexual transmitted infections, 5.36 persons are at risk to be infected by HIV, while among 100 persons without sexual transmitted infections, 1.67 persons are at risk to be infected by HIV.

1. ***(10 points)***  
   In studies of disease incidence, such as HIV, it is customary to represent the data as time to disease (“failure”) in a disease-free population (after all, representing the data as “time surviving as HIV-free” is a pessimistic and stigmatizing viewpoint). Thus we will consider HIV infection in terms of failure *F(t)*, rather than survival *S(t)*, which is *1-F(t).* The calculations involved in the survival analysis don’t meaningfully change and hazard ratio estimates are simply inverted. Similarly, survival curves will be inverted to become *failure curves* (aka: incidence curves), and originate at *F(t)=0* and increase to a maximum of 1.

Use the SAS Help page for PROC LIFETEST to find the relevant option that switches the estimates and orientation of the graph to represent time-to-failure, found at (if clicking on the link does not work, then pasting it in your browser should work):

<http://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#statug_lifetest_sect004.htm>

* 1. Provide a plot of the Kaplan-Meier failure curves for both levels of **STI**.

**proc** **lifetest** data=final method=km plots=SURVIVAL(FAILURE);

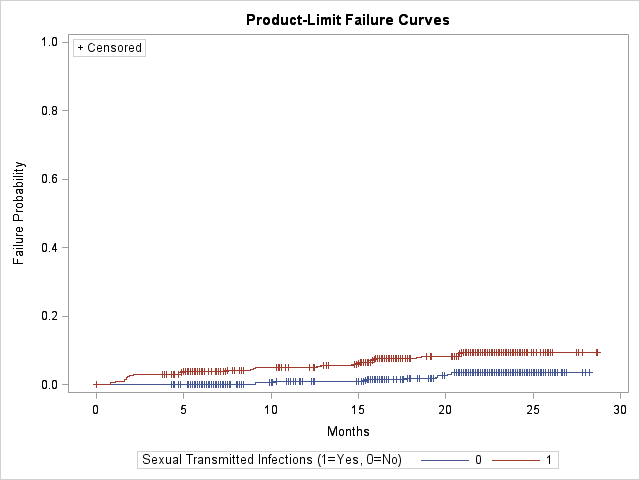
time hiv\_pt\_mo\*hiv\_inc(**0**);

strata sti;

label hiv\_pt\_mo="Months"

sti="Sexual Transmitted Infections (1=Yes, 0=No)";

**run**;



**Figure 1.** Kaplan-Meier failure curves by levels of sexual transmitted infections (STI).

* 1. Carry out a log-rank test to determine if there is a significant difference between the 2 curves.

H0: Two failure curves are the same

H1: Two failure curves are not the same

Log-rank statistics = 25.8506 ~ *Χ* 2df = 1 under H0; p-value < 0.0001.

Under 5% significance level, the data provide sufficient evidence to conclude that there is a significant difference between the two curves.

* 1. Using language suitable for publication, interpret the results of the curves, log-rank test, and the rate ratio’s CI from Question 2.

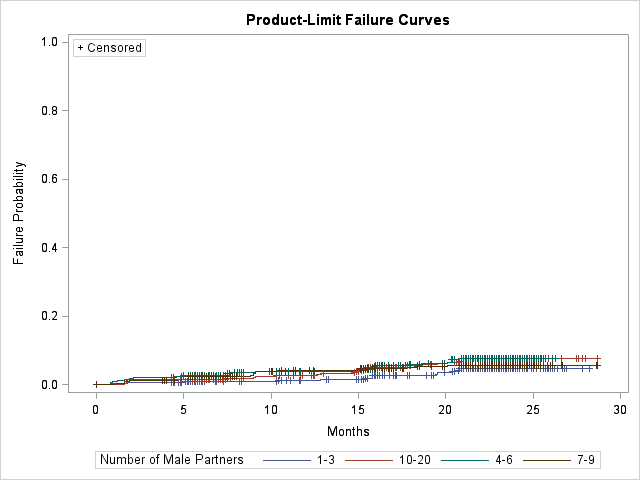
Kaplan-Meier failure curves suggest potential difference in time to HIV infection between two levels of sexual transmitted infections, and log-rank test suggest significant difference in time to HIV infection between two levels of sexual transmitted infections (p < 0.0001). Based on our sample data, we are 95% confident that the true incidence rate ratio for HIV infection comparing those with sexual transmitted infections and those without sexual transmitted infections is between 2.01 and 5.11.

1. ***(10 points)***  
   Similarly, *separately* (i.e. not adjusting for any other covariates) for each of the 3 covariates **male\_partners**, **condom\_use,** and **ZIP\_risk**:

*Note:* Use the same **male\_partners** categories chosen in Question 2

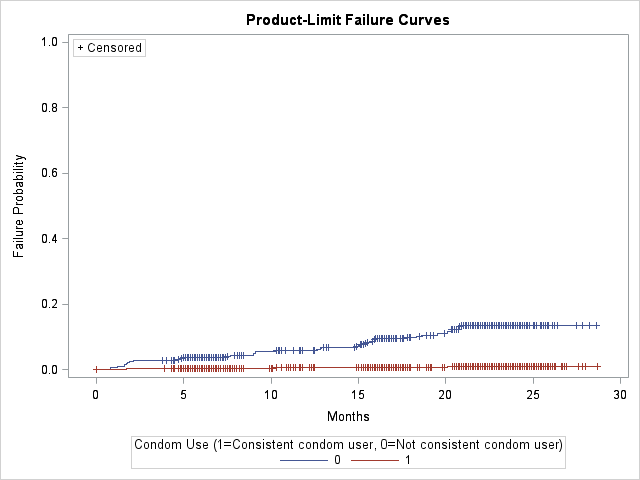
* 1. Provide a plot of the Kaplan-Meier failure curves for the levels of the factor.

(1) Number of Male Partners



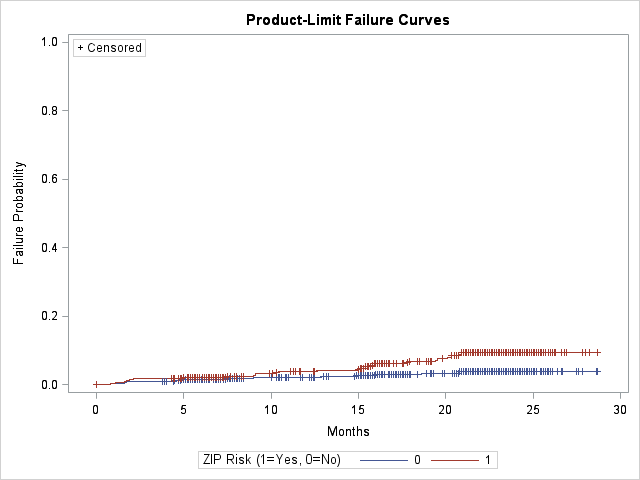
**Figure 2.** Kaplan-Meier failure curves by levels of number of male partners.

(2) Condom Use



**Figure 3.** Kaplan-Meier failure curves by levels of condom use.

(3) ZIP Risk



**Figure 4.** Kaplan-Meier failure curves by levels of ZIP risk.

* 1. Carry out a log-rank test to determine if there is a significant difference between any of the curves.

(1) Number of Male Partners

H0: All failure curves are the same

H1: Not all failure curves are the same

Log-rank statistics = 4.1629 ~ *Χ* 2df = 3 under H0; p-value = 0.2444.

Under 5% significance level, the data do not provide sufficient evidence to conclude that there is a significant difference between any of the curves.

(2) Condom Use

H0: Two failure curves are the same

H1: Two failure curves are not the same

Log-rank statistics = 91.7417 ~ *Χ* 2df = 1 under H0; p-value < 0.0001.

Under 5% significance level, the data provide sufficient evidence to conclude that there is a significant difference between the two curves.

(3) ZIP Risk

H0: Two failure curves are the same

H1: Two failure curves are not the same

Log-rank statistics = 16.2210 ~ *Χ* 2df = 1 under H0; p-value < 0.0001.

Under 5% significance level, the data provide sufficient evidence to conclude that there is a significant difference between the two curves.

* 1. For each covariate, interpret the results of the KM curves, log-rank test, and the rate ratio CI from Question 2

(1) Number of Male Partners

Kaplan-Meier failure curves suggest slight difference in time to HIV infection among four levels of number of male partners, while log-rank test does not suggest significant difference in time to HIV infection among four levels of number of male partners (p = 0.2444). Based on our sample data, we are 95% confident that the true incidence rate ratio for HIV infection comparing those who have 4-6 male partners and those who have 1-3 male is between 0.97 and 3.19, true incidence rate ratio for HIV infection comparing those who have 7-9 male partners and those who have 1-3 male is between 0.70 and 2.52, and true incidence rate ratio for HIV infection comparing those who have more than or equal to 10 male partners and those who have 1-3 male is between 0.96 and 3.01.

(2) Condom Use

Kaplan-Meier failure curves suggest potential difference in time to HIV infection between two levels of condom use, and log-rank test suggest significant difference in time to HIV infection between two levels of condom use (p < 0.0001). Based on our sample data, we are 95% confident that the true incidence rate ratio for HIV infection comparing those who are consistent condom user and those who are not is between 7.39 and 32.38.

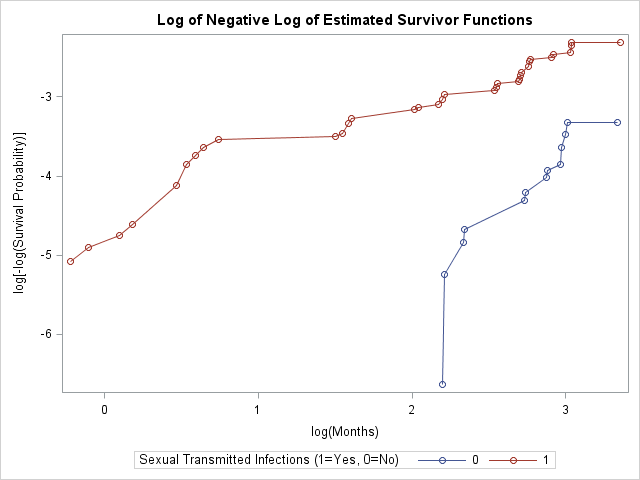
(3) ZIP Risk

Kaplan-Meier failure curves suggest potential difference in time to HIV infection between those who are residents in a ZIP code that has a high prevalence of both HIV and STI and those who are not, and log-rank test suggest significant difference in time to HIV infection between those who are residents in a ZIP code that has a high prevalence of both HIV and STI and those who are not not (p < 0.0001). Based on our sample data, we are 95% confident that the true incidence rate ratio for HIV infection comparing those who are resident in a ZIP code that has a high prevalence of both HIV and STI and those who are not is between 1.54 and 3.72.

1. ***(10 points)***  
   Consider separately all four predictors **sti**, **male\_partners**, **condom\_use,** and **ZIP\_risk**. Using plots of ln-ln Kaplan-Meier curves, evaluate whether each variable satisfies the PH assumption. In answering this question, you need not be restricted to the same categories of **male\_partners** used in Questions 2 and 5.

*Note: PROC LIFETEST only produces ln-ln survival curves, not ln-ln failure curves. However, either may be equivalently used to assess the PH assumption.*

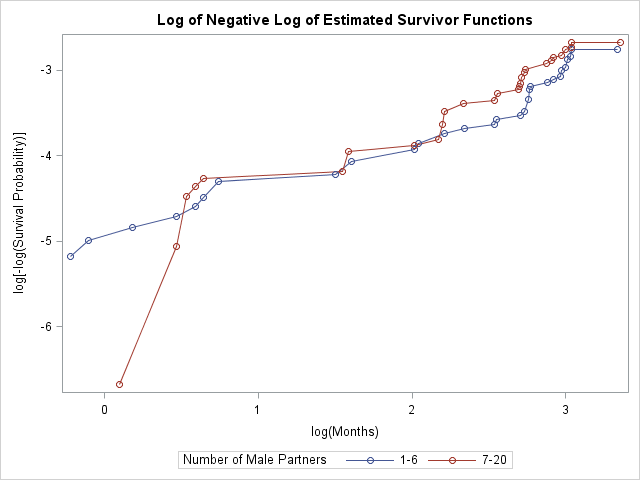
(1) STI



**Figure 5.** Ln-ln Kaplan-Meier survival curves by levels of STI.

The two ln-ln Kaplan-Meier curves do not cross, suggesting that the PH assumption is probably satisfied. (Since graphical method is relatively subjective, we cannot conclude whether it is exactly satisfied before checking the other two methods.)

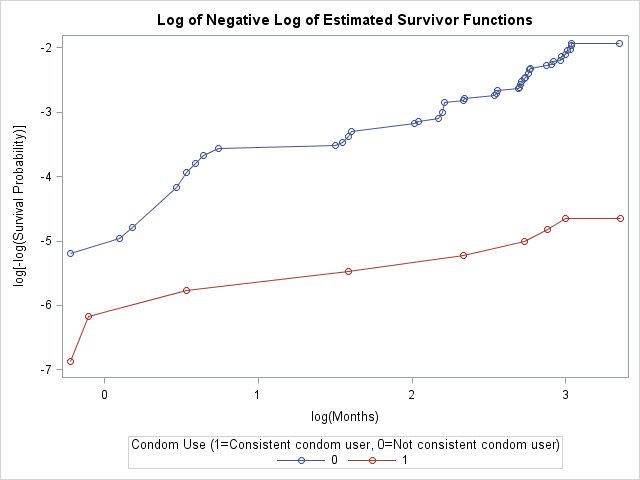
(2) Male Partners (re-categorized as 1-6 male partners and 7-20 male partners; ref = 1-6)



**Figure 6.** Ln-ln Kaplan-Meier survival curves by levels of number of male partners.

The two ln-ln Kaplan-Meier curves cross at several points, suggesting that the PH assumption could be violated. (Since graphical method is relatively subjective, we cannot conclude whether it is exactly violated before checking the other two methods.)

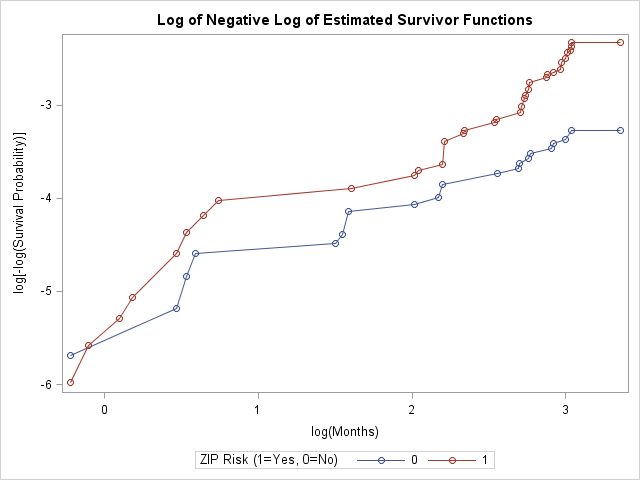
(3) Condom Use



**Figure 7.** Ln-ln Kaplan-Meier survival curves by levels of condom use.

The two ln-ln Kaplan-Meier curves do not cross, suggesting that the PH assumption is probably satisfied. (Since graphical method is relatively subjective, we cannot conclude whether it is exactly satisfied before checking the other two methods.)

(4) ZIP Risk



**Figure 8.** Ln-ln Kaplan-Meier survival curves by levels of ZIP risk.

The two curves cross only at the left extreme of survival time while other parts do not cross. More methods should be checked to conclude if the PH assumption is satisfied or violated.

1. ***(10 points)***  
   Consider separately all 4 predictors again (one-at-a-time). Using Schoenfeld residuals GOF tests, evaluate whether each variable satisfies the PH assumption.

(1) STI

H0: PH assumption satisfied

H1: PH assumption not satisfied

Pearson correlation coefficient = -0.4049; p-value < 0.0001.

Under 5% significance level, PH assumption is violated for STI.

(2) Number of Male Partners (use the same categorization as in Question 6)

H0: PH assumption satisfied

H1: PH assumption not satisfied

Pearson correlation coefficient = -0.0702; p-value = 0.5088.

Under 5% significance level, PH assumption is satisfied for number of male partners.

(3) Condom Use

H0: PH assumption satisfied

H1: PH assumption not satisfied

Pearson correlation coefficient = -0.1007; p-value = 0.3422.

Under 5% significance level, PH assumption is satisfied for condom use.

(4) ZIP Risk

H0: PH assumption satisfied

H1: PH assumption not satisfied

Pearson correlation coefficient = 0.2043; p-value = 0.0521.

Under 5% significance level, PH assumption is satisfied for ZIP risk.

1. ***(10 points)***  
   Consider separately all 4 predictors once more. Using an extended Cox model that contains each predictor and a product term of the form **V** × **t**, where **V** denotes a given predictor and **t** denotes month (a continuous variable), evaluate whether each predictor variable satisfies the PH assumption.

(1) STI

H0: PH assumption satisfied (δ = 0)

H1: PH assumption not satisfied (δ ≠ 0)

Test statistics = 12.7875 ~ *Χ* 2df = 1 under H0; p-value = 0.0003.

Under 5% significance level, PH assumption is violated for STI.

(2) Number of Male Partners (use the same categorization as in Question 6)

H0: PH assumption satisfied (δ = 0)

H1: PH assumption not satisfied (δ ≠ 0)

Test statistics = 0.3302 ~ *Χ* 2df = 1 under H0; p-value = 0.5655.

Under 5% significance level, PH assumption is satisfied for number of male partners.

(3) Condom Use

H0: PH assumption satisfied (δ = 0)

H1: PH assumption not satisfied (δ ≠ 0)

Test statistics = 0.7382 ~ *Χ* 2df = 1 under H0; p-value = 0.3902.

Under 5% significance level, PH assumption is satisfied for condom use.

(4) ZIP Risk

H0: PH assumption satisfied (δ = 0)

H1: PH assumption not satisfied (δ ≠ 0)

Test statistics = 3.4130 ~ *Χ* 2df = 1 under H0; p-value = 0.0647.

Under 5% significance level, PH assumption is satisfied for ZIP risk.

1. ***(10 points)***  
   Based on your results for questions 6, 7, and 8, what can you conclude about which of the four variables satisfy the PH assumption and which do not? (Note: A summary table would help here.) How would you criticize the conclusions based on these methods?

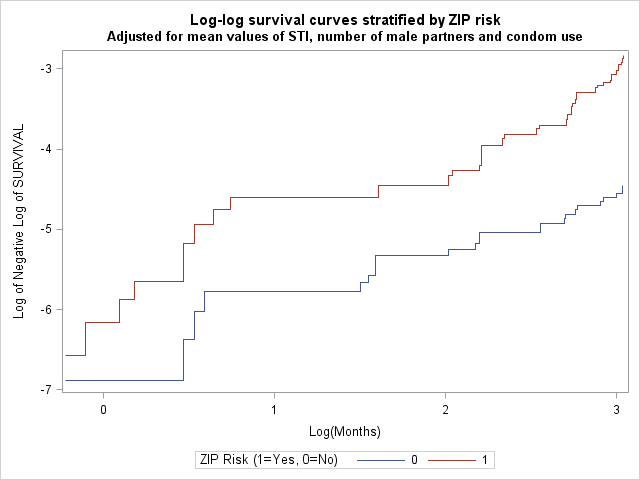
|  |  |  |  |
| --- | --- | --- | --- |
| Predictors | PH assumption | | |
| ln-ln KM Curves | GOF Test | Time-dependent variables |
| STI | satisfied | violated | violated |
| Number of Male Partners | violated | satisfied | satisfied |
| Condom Use | satisfied | satisfied | satisfied |
| ZIP Risk | slightly violated | satisfied | satisfied |

Based on the summary table, PH assumption is satisfied for condom use, since all three methods result in satisfied conclusion; PH assumption is possible to be violated for STI since two out of three methods result in violated conclusion; PH assumptions are possible to be satisfied for number of male partners and ZIP risk since two out of three methods result in satisfied conclusions for them. However, the conclusions based on these methods are questionable since each predictor was evaluated separately but not jointly (i.e. not adjusted for other predictors).

1. ***(10 points)***  
   In a no-interaction Cox model that contains **STI** and controls for **male\_partners**, **condom\_use,** and **ZIP\_risk,** evaluate the PH assumption for **ZIP\_risk** ,using:

*(note: for question 10, dichotomize* ***male\_partners*** *at <7 and ≥7.)*

* 1. Adjusted ln-ln survival curves.   
     *note: be sure to choose appropriate values for the covariates when estimating these!*



**Figure 9.** Adjusted ln-ln survival curves by levels of ZIP risk.

The two ln-ln survival curves do not cross, suggesting that the PH assumption is probably satisfied. (Since graphical method is relatively subjective, we cannot conclude whether it is exactly satisfied before checking the other two methods.)

* 1. A linear interaction term with time. That is, use an extended Cox model that contains all predictors and a product term of the form **V** × **t**, where **V** denotes **ZIP\_risk** and **t** denotes month.

H0: PH assumption satisfied (δ = 0)

H1: PH assumption not satisfied (δ ≠ 0)

Test statistics = 4.1597 ~ *Χ* 2df = 1 under H0; p-value = 0.0414.

Under 5% significance level, PH assumption is violated for ZIP risk.

* 1. Schoenfeld residuals.

H0: PH assumption satisfied

H1: PH assumption not satisfied

Pearson correlation coefficient = 0.2248; p-value = 0.0321.

Under 5% significance level, PH assumption is violated for ZIP risk.

* 1. How do the results of parts a and b compare with their unadjusted counterparts above? What do you conclude about whether **ZIP\_risk** satisfies the proportional hazards assumption?

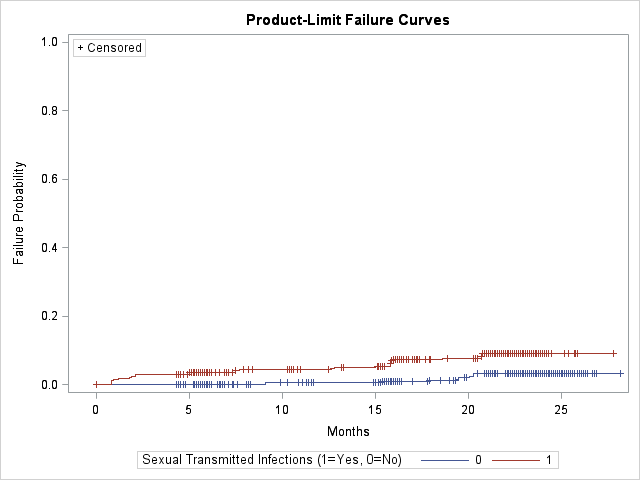
The result of part a (satisfied) is opposite to its unadjusted counterpart (violated), and the result of part b (violated) is opposite to its unadjusted counterpart (satisfied). In the adjusted model, two out of three methods result in violated conclusion, suggesting ZIP risk does not satisfy PH assumption in the adjusted model. And since an adjusted model is more accurate than an unadjusted model, conclusion should be based on the adjusted model. Consequently, ZIP risk does not satisfy the PH assumption.

1. ***(10 points)***  
   Next, we will examine the possible interaction of each factor (male partner number, condom use, and ZIP code risk) with STI, one-at-a-time in unadjusted analyses. There might be important effect-modification to consider. For example, the HIV risk for men with an STI might be greater among men who do not use condoms (compared to those use condoms).

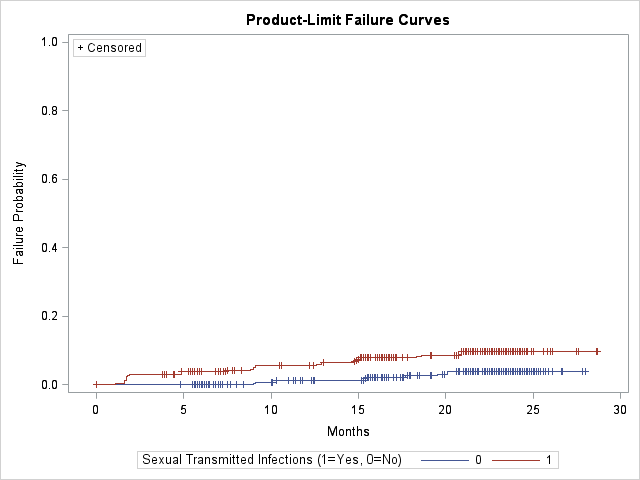
.   
*(note: for this question, again dichotomize* ***male\_partners*** *at <7 and ≥7.)*

* 1. Provide stratified KM failure plots (ie: without adjustment in a regression model), showing for each level of each potential effect modifier (considered separately), the comparison between those with and without an STI.

(1) Number of Male Partners

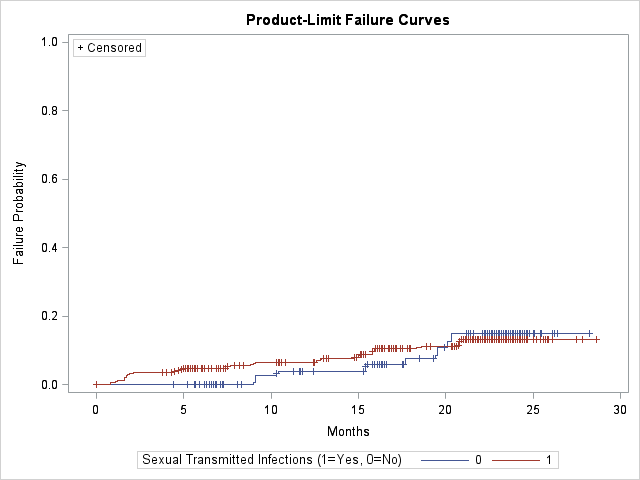


**Figure 10-1.** ln-ln Kaplan-Meier curves by levels of STI, for MSM who have < 7 male partners.

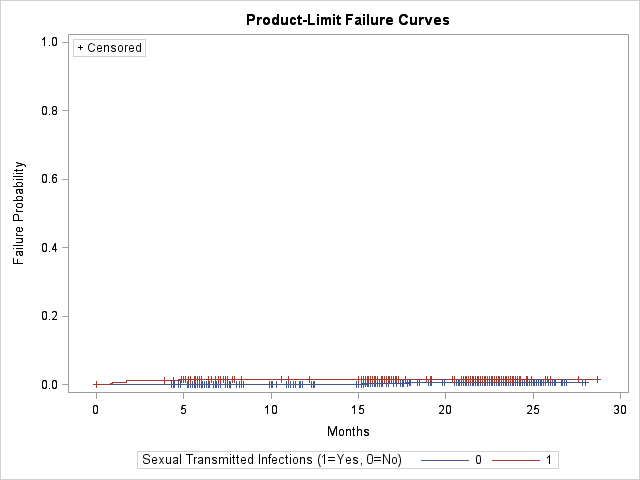


**Figure 10-2.** ln-ln Kaplan-Meier curves by levels of STI, for MSM who have ≥ 7 male partners.

(2) Condom Use

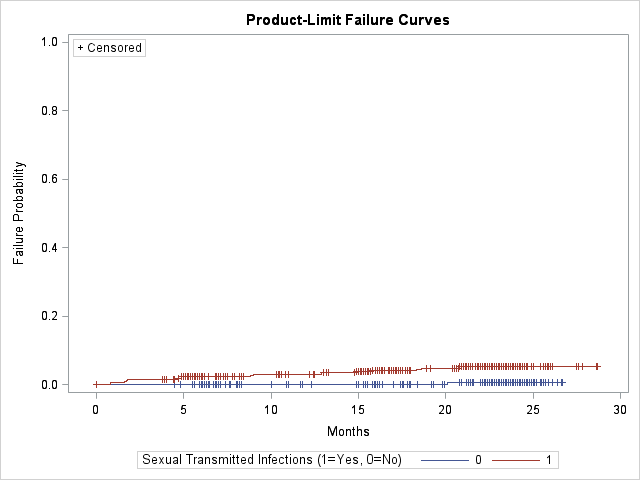


**Figure 11-1.** ln-ln Kaplan-Meier curves by levels of STI, for MSM who are not consistent user of condom.

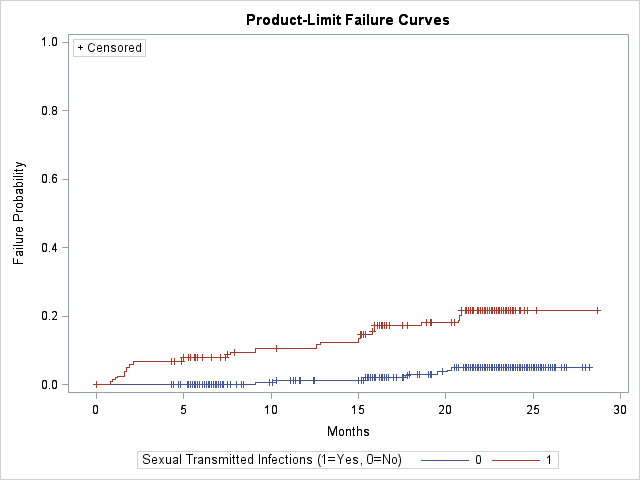


**Figure 11-2.** ln-ln Kaplan-Meier curves by levels of STI, for MSM who are consistent user of condom.

(2) ZIP Risk:



**Figure 12-1.** ln-ln Kaplan-Meier curves by levels of STI, for MSM who do not have ZIP risk.



**Figure 12-2.** ln-ln Kaplan-Meier curves by levels of STI, for MSM who have ZIP risk.

* 1. Conduct appropriate log-rank tests within the strata that compare those with and without STI.

(1) Number of male partners:

*Number of male partners < 7:*

H0: Two failure curves are the same

H1: Two failure curves are not the same

Log-rank statistics = 13.8286 ~ *Χ* 2df = 1 under H0; p-value = 0.0002.

Under 5% significance level, the data provide sufficient evidence to conclude that there is a significant difference between the two curves.

Number of male partners ≥ 7:

H0: Two failure curves are the same

H1: Two failure curves are not the same

Log-rank statistics = 12.1058 ~ *Χ* 2df = 1 under H0; p-value = 0.0005.

Under 5% significance level, the data provide sufficient evidence to conclude that there is a significant difference between the two curves.

(2) Condom use:

*Not consistent condom user:*

H0: Two failure curves are the same

H1: Two failure curves are not the same

Log-rank statistics = 0.0021 ~ *Χ* 2df = 1 under H0; p-value = 0.9630.

Under 5% significance level, the data do not provide sufficient evidence to conclude that there is a significant difference between the two curves.

*Consistent condom user:*

H0: Two failure curves are the same

H1: Two failure curves are not the same

Log-rank statistics = 2.2573 ~ *Χ* 2df = 1 under H0; p-value = 0.1330.

Under 5% significance level, the data do not provide sufficient evidence to conclude that there is a significant difference between the two curves.

(3) ZIP risk:

*Not resident in a ZIP code that has a high prevalence of both HIV and STI:*

H0: Two failure curves are the same

H1: Two failure curves are not the same

Log-rank statistics = 11.1110 ~ *Χ* 2df = 1 under H0; p-value = 0.0009.

Under 5% significance level, the data provide sufficient evidence to conclude that there is a significant difference between the two curves.

*Resident in a ZIP code that has a high prevalence of both HIV and STI:*

H0: Two failure curves are the same

H1: Two failure curves are not the same

Log-rank statistics = 52.0864 ~ *Χ* 2df = 1 under H0; p-value < 0.0001.

Under 5% significance level, the data provide sufficient evidence to conclude that there is a significant difference between the two curves.

* 1. Using the log-rank results and the cumulative “failure probability” in the plots, comment on the possibility of significant effect modification for each of the 3 covariates. How might you critique the results?

Based on the KM failure plot, number of male partners could not be an effect modifier, since the KM plots for its two levels are very similar, and the p-value of log-rank tests for its two levels are close (0.0002 vs. 0.0005); condom use is possible to be an effect modifier since the KM plots for its two levels seem to be very different, and the p-value of log-rank tests for its two levels are very different (0.9630 vs. 0.1330); ZIP code risk is possible to be an effect modifier since the KM plots for its two levels seem to be very different, and the of log-rank statistics for its two levels are very different (11.1110 vs. 52.0864). However, the conclusions based on these methods are questionable since each potential effect modifier was evaluated separately but not jointly (i.e. not adjusted for other predictors), and the evaluations are subjective when judging “similar” and “different”.

1. ***(15 points)***  
   Regardless of your conclusions from earlier questions, fit a stratified Cox PH model with **STI** as the E variable and controlling simultaneously for **male\_partners**, **condom\_use,** and **ZIP\_risk** as both V and W variables, where all two-way (E\*W) effect modification terms are considered. Assume that **condom\_use** does not satisfy the PH assumption and needs to be stratified upon, but that **STI**, **male\_partners** and **ZIP\_risk** each satisfy the PH assumption.

*(note: again, for this question, dichotomize* ***male\_partners*** *at <7 and ≥7.)*

* 1. State the form of the hazard function you have used to fit this model.

hg(t, **X**) = h0g(t)exp[β1STI + β2 male\_partners + β3ZIP\_risk + β4(male\_partners × STI) + β5(ZIP\_risk × STI) + β6(condom\_use × STI)]

where g = 1 (consistent user of condoms), 2 (not consistent user of condoms)

male\_partners = 0 if number of male partners < 7;

= 1 if number of male partners ≥ 7

* 1. Obtain collinearity diagnostics (based on the inverse of the information matrix)for this model. Proceed from this point to examine collinearity for these data, modifying the model and obtaining additional collinearity diagnostics as you consider appropriate.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **VARIABLE** | **VDP1** | **VDP2** | **VDP3** | **VDP4** | **VDP5** | **VDP6** |
| **EIGENVAL** | 0.0162 | 0.11538 | 0.75508 | 1.38578 | 1.56646 | 2.16113 |
| **CONDINDX** | 11.5606 | 4.32781 | 1.69178 | 1.24880 | 1.17458 | 1.00000 |
|  | . | . | . | . | . | . |
| **STI** | 0.9464 | 0.02677 | 0.01220 | 0.00833 | 0.00015 | 0.00616 |
| **Number of Male Partners** | 0.0049 | 0.88389 | 0.01340 | 0.07356 | 0.00459 | 0.01968 |
| **ZIP Risk** | 0.9713 | 0.00651 | 0.00254 | 0.00249 | 0.01711 | 0.00003 |
| **STI × Number of Male Partners** | 0.0060 | 0.94093 | 0.00077 | 0.01450 | 0.00397 | 0.03388 |
| **STI × ZIP Risk** | 0.9777 | 0.00683 | 0.00177 | 0.00068 | 0.01008 | 0.00290 |
| **STI × Condom Use** | 0.0014 | 0.00085 | 0.83302 | 0.13021 | 0.00126 | 0.03325 |

The highest CI is 11.5606 and there are three VDP’s that are > 0.5 for this CI, which corresponds to STI, ZIP risk and the interaction between STI and ZIP risk (VDP = 0.9464, 0.9713 and 0.9777 respectively). This suggests STI, ZIP risk and the interaction between them are involved in the potential collinearity problem.

Since STI is the main exposure of our study, it cannot be dropped to address the collinearity problem. We first drop the interaction between ZIP risk and STI to refit the model and assess the collinearity again:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **VARIABLE** | **VDP1** | **VDP2** | **VDP3** | **VDP4** | **VDP5** |
| **EIGENVAL** | 0.11028 | 0.61310 | 0.90173 | 1.37300 | 2.00189 |
| **CONDINDX** | 4.26052 | 1.80699 | 1.48998 | 1.20749 | 1.00000 |
|  | . | . | . | . | . |
| **STI** | 0.56201 | 0.31978 | 0.00035 | 0.07054 | 0.04732 |
| **Number of Male Partners** | 0.87211 | 0.03662 | 0.00050 | 0.05877 | 0.03201 |
| **ZIP Risk** | 0.00353 | 0.28023 | 0.55455 | 0.13948 | 0.02220 |
| **STI × Number of Male Partners** | 0.94086 | 0.00516 | 0.00006 | 0.00978 | 0.04414 |
| **STI × Condom Use** | 0.00228 | 0.39608 | 0.43772 | 0.13110 | 0.03282 |

The highest CI is 4.26052 which is less than 10, suggesting there is no collinearity problem now.

* 1. What do you conclude about collinearity for these data, and what, if anything do you recommend be done to remedy any collinearity problem found?

According to the results of part b, there is one collinearity problem for these data and the interaction between STI and ZIP risk is the main resource for this problem. To remedy this problem we would recommend dropping the interaction term between STI and ZIP risk.

* 1. How would you criticize this approach to assessing collinearity for these data?

Since this is a Cox model, the cut-off of CI > 30 to judge collinearity problem is not appropriate here. Thus we just look at the highest CI and see if it is > 10. However, there is no agreed on cut-point for Cox model, so this alternative to assess collinearity is relatively subjective.

1. ***(10 points)***   
   Assess the presence of significant effect modification in the model from Question 12, showing the appropriate tests and results along the way.

Likelihood Ratio Test:

H0: δ1 = δ2 = δ3 = 0

H1: one of the δ’s is not 0

LR = -2logLR – (-2logLF) = 1118.497 - 1116.126 = 2.371 ~ *Χ* 2df = 3 under H0; p-value = 0.5.

Under 5% significance level, the data do not provide sufficient evidence to conclude that any of the effect modifications is significant.

1. ***(15 points)***  
   Despite what you were told for questions 12 and 13, a divine vision has informed you that **condom\_use** *does* satisfy the proportional hazards assumption and therefore no longer needs to be stratified on in the model you fit for Question 12. Furthermore, regardless of the results of Question 13, the vision tells you to fit this unstratified Cox PH model, assuming significant effect modification between **STI** and **condom\_use**. Using the all possible subsets change in estimate approach, carry out confounding/precision assessment, summarizing the results in a table. Make a conclusion about which model is ‘best’.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **V variables in model** | **Condom use = 0** | | | **Condom use = 1** | | |
| **aHR** | **Change in aHR** | **95% CI Ratio for aHR** | **aHR** | **Change in aHR** | **95% CI Ratio for aHR** |
| Condom use, Number of male partners, ZIP risk (*GS model*) | 1.943 | - | 2.86 | 5.070 | - | 16.22 |
| Condom use, Number of male partners | 1.009 | 48.07% | 2.74 | 2.761 | 45.54% | 15.98 |
| Condom use, ZIP risk | 1.949 | 0.31% | 2.86 | 5.079 | 0.18% | 16.22 |
| Condom use | 1.012 | 47.92% | 2.74 | 2.769 | 45.39% | 15.98 |

There are three possible subset models. Among all the possible subset models, only the model containing STI, condom use, ZIP risk and interaction between condom use and STI has a change in adjusted HR that is less than 10% for each level of condom use (0.31% and 0.18% for condom use = 0 and condom use = 1, respectively). This suggests that the number of male partners is not a confounder and can be dropped. However, this model does not gain more precision than the GS model in each level of condom use (2.86 vs. 2.86 for condom use = 0; 16.22 vs. 16.22 for condom use = 1). For validity, we still select the GS model as the “best” model.

*SAS output (for the “estimate” statement calculating aHR and 95% CI):*

Model 1 (GS model):

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Label** | **Estimate** | **Standard Error** | **z Value** | **Pr > |z|** | **Alpha** | **Lower** | **Upper** | **Exponentiated** | **Exponentiated Lower** | **Exponentiated Upper** |
| **condom use = 0** | 0.6641 | 0.2682 | 2.48 | 0.0133 | 0.05 | 0.1385 | 1.1896 | 1.9427 | 1.1485 | 3.2859 |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Label** | **Estimate** | **Standard Error** | **z Value** | **Pr > |z|** | **Alpha** | **Lower** | **Upper** | **Exponentiated** | **Exponentiated Lower** | **Exponentiated Upper** |
| **condom use = 1** | 1.6233 | 0.7108 | 2.28 | 0.0224 | 0.05 | 0.2302 | 3.0164 | 5.0699 | 1.2589 | 20.4178 |

Model 2 (Condom use, Number of male partners):

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Label** | **Estimate** | **Standard Error** | **z Value** | **Pr > |z|** | **Alpha** | **Lower** | **Upper** | **Exponentiated** | **Exponentiated Lower** | **Exponentiated Upper** |
| **condom use = 0** | **0.008814** | **0.2568** | **0.03** | **0.9726** | **0.05** | **-0.4944** | **0.5121** | **1.0089** | **0.6099** | **1.6687** |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Label** | **Estimate** | **Standard Error** | **z Value** | **Pr > |z|** | **Alpha** | **Lower** | **Upper** | **Exponentiated** | **Exponentiated Lower** | **Exponentiated Upper** |
| **condom use = 1** | **1.0157** | **0.7070** | **1.44** | **0.1508** | **0.05** | **-0.3701** | **2.4014** | **2.7612** | **0.6907** | **11.0391** |

Model 3 (Condom use, ZIP risk):

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Label** | **Estimate** | **Standard Error** | **z Value** | **Pr > |z|** | **Alpha** | **Lower** | **Upper** | **Exponentiated** | **Exponentiated Lower** | **Exponentiated Upper** |
| **condom use = 0** | **0.6675** | **0.2679** | **2.49** | **0.0127** | **0.05** | **0.1424** | **1.1925** | **1.9493** | **1.1531** | **3.2954** |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Label** | **Estimate** | **Standard Error** | **z Value** | **Pr > |z|** | **Alpha** | **Lower** | **Upper** | **Exponentiated** | **Exponentiated Lower** | **Exponentiated Upper** |
| **condom use = 1** | **1.6252** | **0.7107** | **2.29** | **0.0222** | **0.05** | **0.2322** | **3.0182** | **5.0792** | **1.2613** | **20.4535** |

Model 4 (Condom use only):

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Label** | **Estimate** | **Standard Error** | **z Value** | **Pr > |z|** | **Alpha** | **Lower** | **Upper** | **Exponentiated** | **Exponentiated Lower** | **Exponentiated Upper** |
| **condom use = 0** | **0.01185** | **0.2567** | **0.05** | **0.9632** | **0.05** | **-0.4912** | **0.5149** | **1.0119** | **0.6119** | **1.6735** |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Label** | **Estimate** | **Standard Error** | **z Value** | **Pr > |z|** | **Alpha** | **Lower** | **Upper** | **Exponentiated** | **Exponentiated Lower** | **Exponentiated Upper** |
| **condom use = 1** | **1.0184** | **0.7070** | **1.44** | **0.1497** | **0.05** | **-0.3673** | **2.4041** | **2.7689** | **0.6926** | **11.0690** |

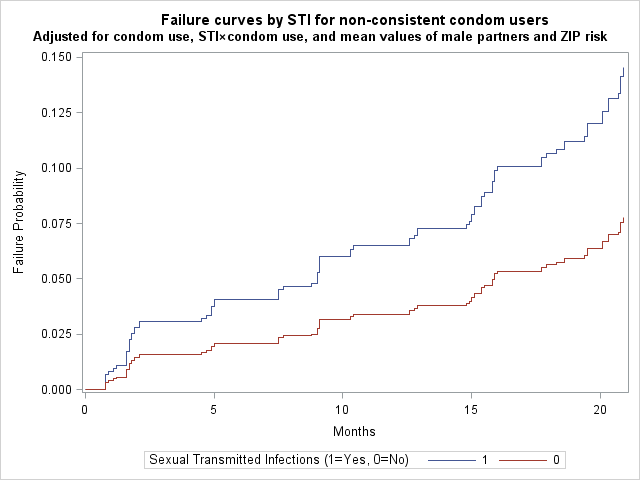
1. ***(10 points)***  
   Interpret the hazard ratio findings from the ‘best’ model identified in question 14, bearing in mind the stated results from clinical trials. Provide plots of adjusted failure curves that demonstrate the effect of having and not having an STI, considering effect modification, at average values of any confounders in your best model. Describe what you see in these plots and put in context with the hazard ratio findings.

*Note: The effect modifier* ***condom\_use*** *is not considered a confounder, even though its main effect is in the model.*

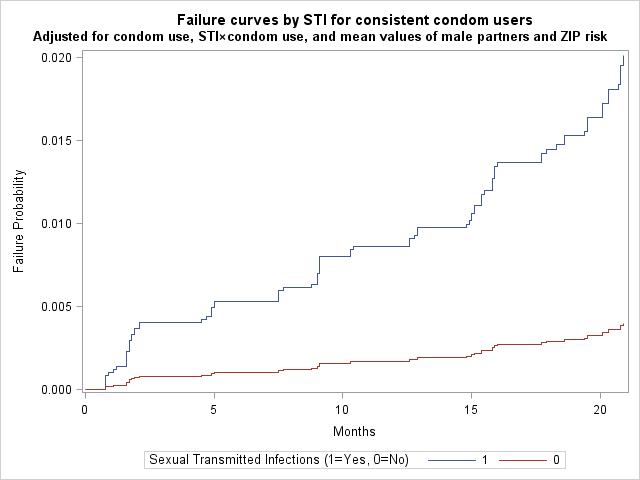
*The built-in plotting options for PHREG can only produce survival and not failure curves. Therefore use the BASELINE statement in conjunction with SGPLOT to produce your curves. This method was demonstrated for adjusted survival curves in lab. Because you are plotting failure, use a data step to compute adjusted failure estimates as 1-(adjusted survival estimates).*

Interpretation: the ratio of the hazard in HIV infection among the MSM who tested positive for an STI at enrollment versus the MSM who tested negative STI at enrollment is estimated to be 5.07 (CI95%:1.26, 20.42) for those who are consistent user of condoms, and 1.94 (CI95%: 1.15, 3.29) for those who are not.

Adjusted curves:



**Figure 13-1.** Adjusted failure curves by levels of STI, for MSM who are not consistent users of condom.



**Figure 13-2.** Adjusted failure curves by levels of STI, for MSM who are consistent users of condom.

Based on the two plots, it is clear that the overall failure probabilities for MSM who are consistent users of condom are less than the overall failure probabilities for MSM who are not consistent users of condom –among MSM who are consistent users of condom, maximum failure probabilities are around 0.020 and 0.0025 for STI = 1 and STI = 0, respectively, while among MSM who are not consistent users of condom, maximum failure probabilities are around 0.150 and 0.075 for STI = 1 and STI = 0, respectively. In addition, the failure probability increases much faster for STI = 1 than that for STI = 0 among MSM who are consistent users of condom, while the difference in changes of failure probabilities between two STI levels among MSM who are not consistent users of condom are not as noticeable as that among MSM who are. This finding is consistent with the HR findings where HR for those who are consistent user of condoms is much higher than HR for those who are not (5.07 vs. 1.94).

**Note for Question 16: In writing up you answer, you are limited to no more than 6 type-written, single-spaced pages (minimum font size 11). Note that many parts of the question can be answered with brief responses!**

***(15 points)***

**16.** This question deals with several alternative modeling approaches to the STI and HIV study.

Additional Survival Analyses

Regardless of your conclusions from earlier questions, assume that the variables **STI** and **ZIP\_risk** do not satisfy the PH assumption, but that **male\_partners (***dichotomized at <7 and ≥7)* and **condom\_use** each satisfy the PH assumption.

1. State the hazard function form of an extended Cox model that would contain **STI, male\_partners**, and **condom\_use** as time-independent predictors, **STI\*g(t)** as a time-dependent predictor where **g(t)** is some function of time, and stratified on **ZIP\_risk** (assuming no interaction of **ZIP\_risk** with the predictors in the model).

Hg(t, **X**) = h0g(t) \* e [β1\*STI + β2\*male\_partners + β3\*condom\_use + δ\*STI\*g(t)]

g=0, 1, strata defined from zip\_risk.

1. For the hazard model in part a, what is the formula for the hazard ratio for the effect of **STI**, that controls for **male\_partners, condom\_use** and **ZIP\_risk**?

HR(STI, controlling for male\_partners, condom\_use, zip\_risk) = e [β1 + δ\*g(t)]

1. Describe how you would carry out a likelihood ratio test to determine if the time-dependent variable **STI\*g(t)** is statistically significant in the hazard model of part a. (In your answer, make sure to describe the null hypothesis, the form of the test statistic, and its distribution under the null hypothesis.

The null hypothesis is whether the participants has an STI at enrollment or not is not dependent on time (δ=0). In order to perform the likelihood ratio test, we need to run both the full model and the reduced model. The full model is the hazard function including all the time-independent predictors and the STI\*g(t) interaction term, while the reduced model is the hazard function that only consists of all the time-independent variables, not STI\*g(t) interaction term.

Full model: Hg(t, **X**) = h0g(t) \* e [β1\*STI + β2\*male\_partners + β3\*condom\_use + δ\*STI\*g(t)]

Reduced model: Hg(t, **X**) = h0g(t) \* e [β1\*STI + β2\*male\_partners + β3\*condom\_use]

Two log final likelihood statistics were used to calculate the likelihood ratio based on the following formula: LR = -2 ln(Lreduced)-(-2ln(Lfull)). This statistic h6as a chi-square distribution with degrees of freedom equal to 1.

1. Based on the data, including previous analyses/survival curves, how would you define the Heaviside function g(t)? Briefly justify your choice.

g(t)=0, t ≤ 3 month

g(t)=1, t > 3 month

Based on previous survival curves, the survival curves among people who has STI at enrollment and those who hasn’t went apart after 3 months.

1. How would you modify the model you stated in part a to allow for possible interaction of **ZIP\_risk** with thevariables **condom\_use** and **male\_partners?**

Hg(t, **X**) = h0g(t) \* e [β1\*STI + β2g\*male\_partners + β3g\*condom\_use + δ\*STI\*g(t)]

g = 0, 1

Logistic regression

Consider the model at the beginning of Question 14 (before confounding assessment) that included **STI** and all 3 control variables (**condom\_use, ZIP\_risk,** *dichotomized* **male\_partners** at 7***)***, and the interaction between **STI** and **condom\_use**. Suppose it is fit in SAS as both a survival analysis (**HIV\_INC** and **HIV\_PT\_MO** as outcome) and as a logistic regression model (**HIV\_INC**) as outcome.

1. Under what conditions for the conduct of the cohort study, might it be appropriate to fit a logistic regression model to the resulting data?   
   (That is, assuming that a cohort study is still done, but perhaps the follow-up might have occurred differently)

When outcome is a categorical variable, it is appropriate to fit a logistic regression model to the resulting data. When fitting a logistic regression model, we do not need Person time in months. That is, we do not need to have as much as 5 visits in the follow-up period to get the time of HIV infection, we just need test for HIV infection at the the end of the study to determine if the patient acquired HIV during follow up. Therefore, if there is no loss of follow up in cohort study, it is appropriate to fit a logistic regression model to the resulting data.

1. Given the cohort study that, was conducted, would you expect the estimates of the 2-year risk of HIV from the sample of 1,834 MSM to be larger, smaller, or equivalent to the “true” 2-year risk in the underlying population? If ‘larger’ or ‘smaller’, name the biases involved.

We can not decide whether the estimates of the 2-year risk of HIV from from the sample of 1,834 MSM to be larger, smaller, or equivalent to the “true” 2-year risk in the underlying population. There are potential selection bias and information of bias exist in this study. Because this study enrolled individuals who come in for STI testing, and they may be not representative of the source population. They may have different characteristics like more high-risk activities than the source population have, which will result in a selection bias. Besides, subjects who got HIV infection may get STI during the follow-up period rather than at the enrollment of this study. In this situation, these subjects would introduce a bias of misclassification of exposure.

1. Fit this interaction model as a Cox PH survival analysis and as a logistic regression analysis. Fill in the following table with the HR, OR, and conditional-margins RR (just point estimates are ok) and comment on how these estimates compare to one another.

|  |  |  |
| --- | --- | --- |
|  | **Condom\_use = Yes** | **Condom\_use = No** |
| **Model, measure** |  |  |
| Survival, HR | 5.0699 | 1.9427 |
| Logistic, OR | 5.2539 | 1.9804 |
| Logistic, RR | 5.1907 | 1.8671 |

Comments: HR, OR, and conditional-margins RR are approximately the same and they are all > 1, indicating STI is associated with increasing risk of HIV infection. Besides, this association is stronger among people who use condom compared to those who don’t. However, we can see the ORs inflate the RR in this cohort study.

1. Explain the reason for how the estimated OR compares to the estimated RR, in question 16h, given the observed outcomes in this study?

The estimated ORs inflate the estimated RRs among condom use group and non-condom use group. Since HIV infection is not a rare disease in both condom use group and non-condom use group, ORs will overestimate the RRs in cohort study.

1. Given your “theoretical” responses to questions 16f-g, the estimates in question 16h, and knowing that this was a study with variable follow-up times on participants, what do you conclude about the potential bias of using a logistic regression model on these data?

When using a logistic regression model on these data, we do not use the variable follow-up times on participants and thus lead to the loss of information. For example, those who left the study are simply regarded as missing in outcome HIV infection while there is a period of time he contributed to the time of non-infection. So the estimate risk may be larger in logistic model. There might be information bias of using a logistic regression model on these data.

Poisson Regression

Suppose, you now decided to fit a Poisson regression survival model, using the same interaction model considered at the start of Question 14 and Question 16f that included **STI** and all 3 control variables (**condom\_use, ZIP\_risk,** *dichotomized* **male\_partners** at 7***)***, and the interaction between **STI** and **condom\_use**.

This analysis would ideally use a summary dataset of all participants and their HIV infection “counts”, rather than an individual line-listing (one line per participant) style of data.

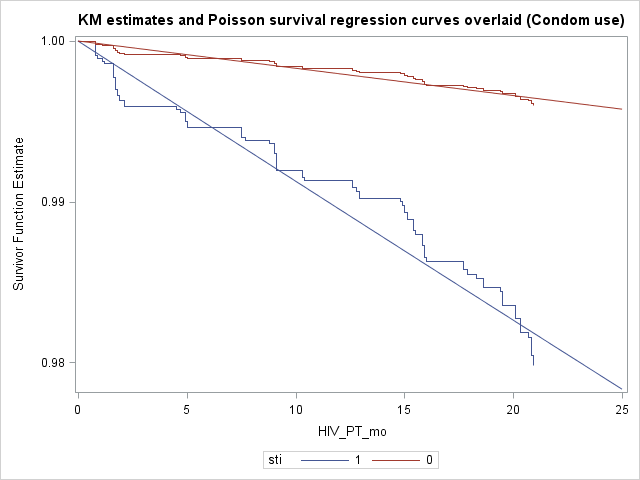
1. Complete the 16-row grid below for the summary dataset you would construct to conduct this analysis. The two columns **HIV\_inc\_total** and **HIV\_PT\_mo\_total** represent the sum of **HIV\_INC** and **HIV\_PT\_mo**

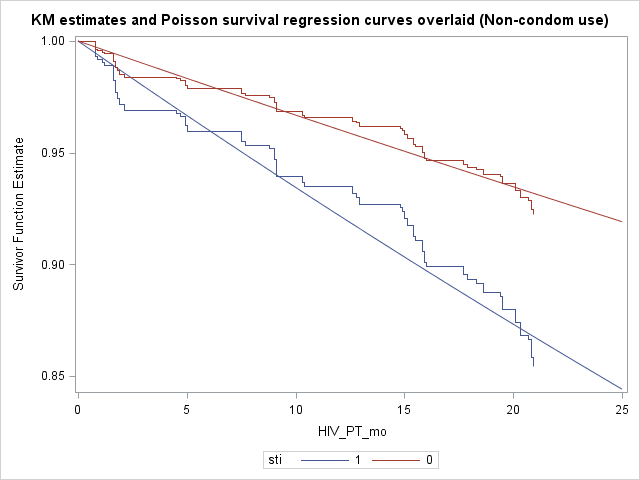
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **STI** | **condom\_use** | **male\_partner\_bin** | **ZIP\_risk** | **HIV\_inc\_total** | **HIV\_PT\_mo\_total** |
| 0 | 0 | 0 | 0 | 0 | 511.2 |
| 0 | 0 | 0 | 1 | 11 | 1292.5 |
| 0 | 0 | 1 | 0 | 1 | 443.4 |
| 0 | 0 | 1 | 1 | 8 | 1084.7 |
| 0 | 1 | 0 | 0 | 0 | 2451.6 |
| 0 | 1 | 0 | 1 | 1 | 4940.2 |
| 0 | 1 | 1 | 0 | 0 | 2208.7 |
| 0 | 1 | 1 | 1 | 3 | 4300.7 |
| 1 | 0 | 0 | 0 | 7 | 3954.2 |
| 1 | 0 | 0 | 1 | 24 | 1152.4 |
| 1 | 0 | 1 | 0 | 19 | 3695.1 |
| 1 | 0 | 1 | 1 | 13 | 1254.0 |
| 1 | 1 | 0 | 0 | 1 | 1863.4 |
| 1 | 1 | 0 | 1 | 1 | 611.1 |
| 1 | 1 | 1 | 0 | 1 | 1908.8 |
| 1 | 1 | 1 | 1 | 1 | 569.2 |

1. Fit the Poisson survival model and compute the appropriate estimated rate-ratio for the effect of STI on HIV incidence. Comment on how these results compare to those obtained in the Cox proportional hazards model from Question 16h. What does this suggest about each respective hazard assumption for the Cox and Poisson models?

According to the above output, the estimated rate ratio of acquiring HIV among people who have STI at enrollment compared to those who don’t is 5.1510 among condom use group and 2.0088 among non-condom use group. According to the result of 16h, the hazard ratios of acquiring HIV compared people who have STI at enrollment with those who don’t are 5.0699 and 1.9427 respectively among condom use group and non- condom use group. The estimated rate ratios based on poisson survival model are close to the estimated hazard ratios using Cox proportional hazard model.

The important feature in Cox PH model is that the baseline function is a hazard of t , but not involve any time-independent predictors, whereas the exponential expression involves the predictors, not t. Therefore, the Cox PH model assumes that the hazard ratio comparing two specifications of any predictors is constant over time. Besides, the Cox PH model is robust since it is a semi-parametric model of unspecified baseline hazard, which indicates Cox PH model usually gives good results regardless of the correct parametric model. For the Poisson survival model, it is parametric and it assumes that the hazard is constant over time. It will give equivalent answers to Cox PH model when all the predictors are categorical and the assumption is met. Therefore, since the estimates using two models are close, it suggests that the hazard rate of acquiring HIV is constant over time.





By checking this assumption, overlaid plots of KM adjusted survival curves and Poisson survival curves were generated by condom use. Based on the graphs above, it shows that the assumption of Poisson survival regression is met among both condom use and non-condom use. (SAS code showed in Appendix)

**Appendix: SAS Code**

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\* EPI 740 Take-home Final Exam

\* Programmer: Xinyi Zhao, Xi Liu, Yongjia Song

\* Date: 12/09/2015

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

libname hitomi "H:\Fall 2015\EPI 740";

\*-----------------------------------------;

\* Question 2 --Descriptive statistics ;

\*-----------------------------------------;

\*\*\* Data cleaning;

**proc** **univariate** data=hitomi.final2015;

var male\_partners;

**run**;

\* categorize male partners by quartiles;

**data** final;

set hitomi.final2015;

if male\_partners<=**3** then npartners="1-3 ";

if male\_partners>**3** and male\_partners<=**6** then npartners="4-6";

if male\_partners>**6** and male\_partners<=**9** then npartners="7-9";

if male\_partners>**9** then npartners="10-20";

py=hiv\_pt\_mo/**12**;

**run**;

**proc** **means** data=final max min;

var male\_partners;

class npartners;

**run**;

\*\*\* Descriptive statistics;

\*\* Frequencies;

**proc** **means** data=final;

var sti zip\_risk;

**run**;

**proc** **freq** data=final;

tables sti zip\_risk npartners;

**run**;

\*\* # HIV infections and rates;

\* overall;

**proc** **means** data=final sum;

var hiv\_inc py;

**run**;

\* STI;

**proc** **means** data=final sum;

var hiv\_inc py;

class sti;

**run**;

\* ZIP risk;

**proc** **means** data=final sum;

var hiv\_inc py;

class zip\_risk;

**run**;

\* number of male partners;

**proc** **means** data=final sum;

var hiv\_inc py;

class npartners;

**run**;

\*\* Rate ratios;

**%macro** idr(group, ref);

proc sql;

create table hiv\_infections as

select &group,

count(hiv\_inc) as hiv\_infections

from final

where hiv\_inc=**1**

group by &group;

quit;

proc print data=hiv\_infections;

run;

proc sql;

create table person\_years as

select &group,

round(sum(py),**1**) as person\_years

from final

group by &group;

quit;

proc print data=person\_years;

run;

proc sql;

create table twobytwo as

select \*, log(person\_years/**100**) as ln\_n

from hiv\_infections,person\_years

where hiv\_infections.&group=person\_years.&group;

quit;

proc print data=twobytwo;

run;

proc genmod data=twobytwo;

class &group (ref="&ref");

model hiv\_infections=&group/link=log dist=poisson offset=ln\_n;

lsmeans &group/ ilink cl diff exp;

run;

**%mend**;

options mprint symbolgen;

\* STI;

%***idr***(sti, **0**);

\* ZIP risk;

%***idr***(zip\_risk, **0**);

\* number of male partners;

%***idr***(npartners, **1**-**3**);

\*------------------------------------------------------;

\* Question 4 --KM failure curves and log-rank test ;

\*------------------------------------------------------;

**proc** **lifetest** data=final method=km plots=SURVIVAL(FAILURE);

time hiv\_pt\_mo\*hiv\_inc(**0**);

strata sti;

label hiv\_pt\_mo="Months"

sti="Sexual Transmitted Infections (1=Yes, 0=No)";

**run**;

\*------------------------------------------------------;

\* Question 5 --KM failure curves and log-rank test ;

\*------------------------------------------------------;

\* number of male partners;

**proc** **lifetest** data=final method=km plots=SURVIVAL(FAILURE);

time hiv\_pt\_mo\*hiv\_inc(**0**);

strata npartners;

label hiv\_pt\_mo="Months"

npartners="Number of Male Partners";

**run**;

\* condom use;

**proc** **lifetest** data=final method=km plots=SURVIVAL(FAILURE);

time hiv\_pt\_mo\*hiv\_inc(**0**);

strata condom\_use;

label hiv\_pt\_mo="Months"

condom\_use="Condom Use (1=Consistent condom user, 0=Not consistent condom user)";

**run**;

\* ZIP risk;

**proc** **lifetest** data=final method=km plots=SURVIVAL(FAILURE);

time hiv\_pt\_mo\*hiv\_inc(**0**);

strata zip\_risk;

label hiv\_pt\_mo="Months"

zip\_risk="ZIP Risk (1=Yes, 0=No)";

**run**;

\*-------------------------------------------------------------------------;

\* Question 6 --checking PH assumption using log-log survival curves ;

\*-------------------------------------------------------------------------;

\* STI;

**proc** **lifetest** data=final method=km plots=(lls);

time hiv\_pt\_mo\*hiv\_inc(**0**);

strata sti;

label hiv\_pt\_mo="Months"

sti="Sexual Transmitted Infections (1=Yes, 0=No)";

**run**;

\* number of male partners;

**data** temp;

set final;

if male\_partners<=**6** then npart="1-6 ";

if male\_partners>**6** then npart="7-20";

if male\_partners<=**6** then np=**0**;

if male\_partners>**6** then np=**1**;

/\* re-categorized as 1-6 and 7-20\*/

**run**;

**proc** **means** data=temp max min;

var male\_partners;

class npart;

**run**;

**proc** **freq** data=temp;

tables npart\*np/list;

**run**;

**proc** **lifetest** data=temp method=km plots=(lls);

time hiv\_pt\_mo\*hiv\_inc(**0**);

strata npart;

label hiv\_pt\_mo="Months"

npart="Number of Male Partners";

**run**;

\* condom use;

**proc** **lifetest** data=final method=km plots=(lls);

time hiv\_pt\_mo\*hiv\_inc(**0**);

strata condom\_use;

label hiv\_pt\_mo="Months"

condom\_use="Condom Use (1=Consistent condom user, 0=Not consistent condom user)";

**run**;

\* ZIP risk;

**proc** **lifetest** data=final method=km plots=(lls);

time hiv\_pt\_mo\*hiv\_inc(**0**);

strata zip\_risk;

label hiv\_pt\_mo="Months"

zip\_risk="ZIP Risk (1=Yes, 0=No)";

**run**;

\*---------------------------------------------------------;

\* Question 7 -- checking PH assumption using GOF test ;

\*---------------------------------------------------------;

**%macro** gof(dsn, group);

proc phreg data=&dsn;

model hiv\_pt\_mo\*hiv\_inc(**0**)=&group;

output out=SR\_DATA ressch=SR\_E;

run;

data fonly;

set SR\_DATA;

if hiv\_inc=**1**;

run;

proc rank data=fonly out=ranked\_f ties=mean;

var hiv\_pt\_mo;

ranks timerank;

run;

proc corr data=ranked\_f nosimple;

var SR\_E;

with timerank;

run;

**%mend**;

options mprint symbolgen;

\* STI;

%***gof***(final, sti);

\* number of male partners;

%***gof***(temp, np);

\* condom use;

%***gof***(final, condom\_use);

\* ZIP risk;

%***gof***(final, zip\_risk);

\*-------------------------------------------------------------------------;

\* Question 8 -- checking PH assumption using time-dependent variables ;

\*-------------------------------------------------------------------------;

\* STI;

**proc** **phreg** data=temp;

model hiv\_pt\_mo\*hiv\_inc(**0**)=sti stit;

stit=sti\*hiv\_pt\_mo;

**run**;

\* number of male partners;

**proc** **phreg** data=temp;

model hiv\_pt\_mo\*hiv\_inc(**0**)=np npt;

npt=np\*hiv\_pt\_mo;

**run**;

\* condom use;

**proc** **phreg** data=temp;

model hiv\_pt\_mo\*hiv\_inc(**0**)=condom\_use condomt;

condomt=condom\_use\*hiv\_pt\_mo;

**run**;

\* ZIP risk;

**proc** **phreg** data=temp;

model hiv\_pt\_mo\*hiv\_inc(**0**)=zip\_risk zipt;

zipt=zip\_risk\*hiv\_pt\_mo;

**run**;

\*----------------------------------------;

\* Question 10 --adjusted Cox model ;

\*----------------------------------------;

\* re-categorize number of male partners;

**data** temp3;

set final;

if male\_partners<**7** then nps=**0**;

if male\_partners>=**7** then nps=**1**;

**run**;

**proc** **means** data=temp3 max min;

var male\_partners;

class nparts;

**run**;

**proc** **freq** data=temp3;

tables nparts\*nps/list;

**run**;

\*\*\* ln-ln KM curves;

\* obtain means for each predictor;

**proc** **univariate** data=temp3;

var sti nps condom\_use zip\_risk;

histogram sti nps condom\_use zip\_risk/normal;

**run**;

**proc** **means** data=temp3;

var sti nps condom\_use;

**run**;

**data** one\_level;

input sti nps condom\_use;

datalines;

0.4885496 0.4711014 0.5676118

;

**run**;

**proc** **print** data=one\_level;

**run**;

\* Create adjusted ln-ln curves;

**proc** **phreg** data=temp3 plots(overlay=stratum)=survival;

model hiv\_pt\_mo\*hiv\_inc(**0**)=sti nps condom\_use;

strata zip\_risk;

baseline covariates=one\_level out=out1 survival=survgraph loglogs=llsurvgraph;

**run**;

**data** log;

set out1;

ln\_mo=log(hiv\_pt\_mo);

**run**;

**proc** **sgplot** data=log;

Title1 "Log-log survival curves stratified by ZIP risk";

Title2 "Adjusted for mean values of STI, number of male partners and condom use";

step x=ln\_mo y=llsurvgraph/group=zip\_risk;

label ln\_mo="Log(Months)"

zip\_risk="ZIP Risk (1=Yes, 0=No)";

**run**;

\*\*\* Time-dependent variables;

**proc** **phreg** data=temp3;

model hiv\_pt\_mo\*hiv\_inc(**0**)=sti nps condom\_use zip\_risk zipt;

zipt=zip\_risk\*hiv\_pt\_mo;

**run**;

\*\*\* GOF test;

**proc** **phreg** data=temp3;

model hiv\_pt\_mo\*hiv\_inc(**0**)=sti nps condom\_use zip\_risk;

output out=SR\_DATA ressch=SR\_sti SR\_nps SR\_condom SR\_zip;

**run**;

**data** fonly;

set SR\_DATA;

if hiv\_inc=**1**;

**run**;

**proc** **rank** data=fonly out=ranked\_f ties=mean;

var hiv\_pt\_mo;

ranks timerank;

**run**;

**proc** **corr** data=ranked\_f nosimple;

var SR\_sti SR\_nps SR\_condom SR\_zip;

with timerank;

**run**;

\*-------------------------------------------;

\* Question 11 -- examine interactions ;

\*-------------------------------------------;

\*\*\* KM failure plot;

**%macro** itr(itr);

proc lifetest data=temp3 method=km plots=SURVIVAL(FAILURE);

time hiv\_pt\_mo\*hiv\_inc(**0**);

where &itr=**0**;

strata sti;

label hiv\_pt\_mo="Months"

sti="Sexual Transmitted Infections (1=Yes, 0=No)";

run;

proc lifetest data=temp3 method=km plots=SURVIVAL(FAILURE);

time hiv\_pt\_mo\*hiv\_inc(**0**);

where &itr=**1**;

strata sti;

label hiv\_pt\_mo="Months"

sti="Sexual Transmitted Infections (1=Yes, 0=No)";

run;

**%mend**;

options mprint symbolgen;

\* Male partners;

%***itr***(nps);

\* Condom use;

%***itr***(condom\_use);

\* ZIP risk;

%***itr***(zip\_risk);

\*-------------------------------------------;

\* Question 12 -- examine collinearity ;

\*-------------------------------------------;

filename collin "S:\course\epi750\SAS macros\collin\_2011.sas";

%include collin;

**data** temp4;

set temp3;

stinps=sti\*nps;

stizip=sti\*zip\_risk;

sticon=sti\*condom\_use;

**run**;

\* original model;

**proc** **phreg** data=temp4 covout outest=file;

model hiv\_pt\_mo\*hiv\_inc(**0**)=sti nps zip\_risk stinps stizip sticon/covb;

strata condom\_use;

**run**;

%***collin***(covdsn=file, output=col)

\* drop stizip;

**proc** **phreg** data=temp4 covout outest=file1;

model hiv\_pt\_mo\*hiv\_inc(**0**)=sti nps zip\_risk stinps sticon/covb;

strata condom\_use;

**run**;

%***collin***(covdsn=file1, output=col)

\*-------------------------------------------;

\* Question 13 -- examine interaction ;

\*-------------------------------------------;

\* interaction model;

**proc** **phreg** data=temp4;

model hiv\_pt\_mo\*hiv\_inc(**0**)=sti nps zip\_risk stinps stizip sticon;

strata condom\_use;

**run**;

\* non-interaction model;

**proc** **phreg** data=temp4;

model hiv\_pt\_mo\*hiv\_inc(**0**)=sti nps zip\_risk;

strata condom\_use;

**run**;

\*-------------------------------------------;

\* Question 14 -- examine confounding ;

\*-------------------------------------------;

**proc** **phreg** data=temp4;

model hiv\_pt\_mo\*hiv\_inc(**0**)=sti nps zip\_risk condom\_use sticon/rl covb;

estimate "condom use = 0" sti **1**/exp cl;

estimate "condom use = 1" sti **1** sticon **1**/exp cl;

**run**;

\* if drop zip risk;

**proc** **phreg** data=temp4;

model hiv\_pt\_mo\*hiv\_inc(**0**)=sti nps condom\_use sticon/rl covb;

estimate "condom use = 0" sti **1**/exp cl;

estimate "condom use = 1" sti **1** sticon **1**/exp cl;

**run**;

\* if drop male partners;

**proc** **phreg** data=temp4;

model hiv\_pt\_mo\*hiv\_inc(**0**)=sti condom\_use zip\_risk sticon/rl covb;

estimate "condom use = 0" sti **1**/exp cl;

estimate "condom use = 1" sti **1** sticon **1**/exp cl;

**run**;

\* if drop both zip risk and male partners;

**proc** **phreg** data=temp4;

model hiv\_pt\_mo\*hiv\_inc(**0**)=sti condom\_use sticon/rl covb;

estimate "condom use = 0" sti **1**/exp cl;

estimate "condom use = 1" sti **1** sticon **1**/exp cl;

**run**;

\*-----------------------------------------------;

\* Question 15 -- adjusted failure curves ;

\*-----------------------------------------------;

**proc** **means** data=temp4;

var nps zip\_risk;

**run**;

\* covariate sets for condom = 1;

**data** one\_level\_1;

input sti nps zip\_risk condom\_use sticon;

datalines;

1 0.4711014 0.4623773 1 1

0 0.4711014 0.4623773 1 0

;

**run**;

**proc** **print** data=one\_level\_1;

**run**;

\* covariate sets for condom = 0;

**data** one\_level\_2;

input sti nps zip\_risk condom\_use sticon;

datalines;

1 0.4711014 0.4623773 0 0

0 0.4711014 0.4623773 0 0

;

**run**;

**proc** **print** data=one\_level\_2;

**run**;

**%macro** adj(group, dsn, status);

proc phreg data=temp4 plots(overlay=row)=(survival);

model hiv\_pt\_mo\*hiv\_inc(**0**)=sti nps zip\_risk condom\_use sticon;

baseline covariates=&dsn out=out1 survival=survgraph loglogs=llsurvgraph/id=sti;

run;

data failure1;

set out1;

f=**1**-survgraph;

run;

proc sgplot data=failure1;

where condom\_use=&status;

Title1 "Failure curves by STI for &group";

Title2 "Adjusted for condom use, STI×condom use, and mean values of male partners and ZIP risk";

step x=hiv\_pt\_mo y=f/group=sti;

label hiv\_pt\_mo="Months"

sti="Sexual Transmitted Infections (1=Yes, 0=No)"

f="Failure Probability";

run;

**%mend**;

options mprint symbolgen;

\* condom use = 1;

%***adj***(consistent condom users, one\_level\_1, **1**);

\* condom use = 0;

%***adj***(non-consistent condom users, one\_level\_2, **0**);

\*-------------------;

\* Question 16 ;

\*-------------------;

**data** final;

set final;

sti\_condom=sti\*condom\_use;

**run**;

\*\*\*\* Survival analysis;

**proc** **phreg** data=interaction;

model hiv\_pt\_mo\*hiv\_inc(**0**)=sti condom\_use zip\_risk np sti\_condom/rl;

contrast "condom" sti **1** sti\_condom **1**/estimate=exp;

contrast "no condom use" sti **1** sti\_condom **0**/estimate=exp;

**run**;

\*\*\*\* Logistic regression;

\* OR;

**proc** **logistic** data=interaction descending;

model hiv\_inc=sti condom\_use zip\_risk np sti\_condom;

contrast "condom" sti **1** sti\_condom **1**/estimate=exp;

contrast "no condom" sti **1** sti\_condom **0**/estimate=exp;

**run**;

\* RR;

**proc** **means** data=final;

var sti condom\_use zip\_risk np sti\_condom;

**run**;

**proc** **logistic** data=final descending;

model hiv\_inc=sti condom\_use zip\_risk np sti\_condom;

estimate "P(D|E=1) condom use" intercept **1** sti **1** condom\_use **1** zip\_risk **0.4623773** np **0.4711014** sti\_condom **1**/ilink;

estimate "P(D|E=0) condom use" intercept **1** sti **0** condom\_use **1** zip\_risk **0.4623773** np **0.4711014** sti\_condom **0**/ilink;

estimate "P(D|E=1) non condom use" intercept **1** sti **1** condom\_use **0** zip\_risk **0.4623773** np **0.4711014** sti\_condom **0**/ilink;

estimate "P(D|E=0) non condom use" intercept **1** sti **0** condom\_use **0** zip\_risk **0.4623773** np **0.4711014** sti\_condom **0**/ilink;

**run**;

\* 16i;

\* Look at the distribution of hiv\_inc among condom\_use and non-condom\_use;

**proc** **freq** data=final;

tables hiv\_inc\*sti;

where condom\_use=**1**;

**run**;

**proc** **freq** data=final;

tables hiv\_inc\*sti;

where condom\_use=**0**;

**run**;

\* 16k;

**proc** **means** data=final n sum maxdec=**2**;

class sti condom\_use np zip\_risk;

var hiv\_inc hiv\_pt\_mo;

output out=final\_sum sum(hiv\_inc hiv\_pt\_mo)=hiv\_inc\_sum hiv\_pt\_mo\_sum;

**run**;

**data** final\_sum;

set final\_sum;

where \_TYPE\_ =**15**;

ln\_survt\_sum = log(hiv\_pt\_mo\_sum);

**run**;

**proc** **print** data=final\_sum;

**run**;

**proc** **genmod** data=final\_sum;

class sti(ref = "0") / param=ref;

model hiv\_inc\_sum = sti condom\_use np zip\_risk sti\*condom\_use/ offset = ln\_survt\_sum link=log dist=poisson;

estimate "RR" sti **1** sti\*condom\_use **1** /exp;

estimate "RR" sti **1** sti\*condom\_use **0** /exp;

**run**;

\* Comparing the KM curves vs Poisson regression model;

\*\*\* Condom use group;

\* Plot the adjusted KM plot;

**proc** **means** data=final;

var condom\_use np zip\_risk condom\_use sti\_condom;

**run**;

**data** km\_plot;

input sti condom\_use np zip\_risk sti\_condom;

cards;

1 1 0.4711014 0.4623773 1

0 1 0.4711014 0.4623773 0

;

**run**;

**proc** **phreg** data=final plots(overlay)=(survival);

model hiv\_pt\_mo\*hiv\_inc(**0**)=sti condom\_use np zip\_risk sti\_condom;

baseline covariates=km\_plot out=out1 survival=s1;

**run**;

**proc** **sgplot** data=out1;

step y=s1 x=hiv\_pt\_mo;

**run**;

\* Plot the Poisson regression model;

**data** exp\_plot;

do t=**0** to **25** by **0.1**;

sti=**1**;

s=exp(-t\*exp(-**6.4298**+**0.6975**-**2.9876**\***1**+**0.0514**\***0.4711014**+**1.5459**\***0.4623773**+**0.9416**\***1**));

output;

sti=**0**;

s=exp(-t\*exp(-**6.4298**-**2.9876**\***1**+**0.0514**\***0.4711014**+**1.5459**\***0.4623773**));

output;

end;

**run**;

**data** exp\_plot;

set exp\_plot out1;

**run**;

title "KM estimates and Poisson survival regression curves overlaid (Condom use)";

**proc** **sgplot** data=exp\_plot;

series y=s x=t/group=sti;

step y=s1 x=hiv\_pt\_mo/group=sti;

**run**;

title;

\*\*\* Non-Condom use group;

\* Plot the adjusted KM plot;

**proc** **means** data=final;

var condom\_use np zip\_risk condom\_use sti\_condom;

**run**;

**data** km\_plot;

input sti condom\_use np zip\_risk sti\_condom;

cards;

1 0 0.4711014 0.4623773 0

0 0 0.4711014 0.4623773 0

;

**run**;

**proc** **phreg** data=final plots(overlay)=(survival);

model hiv\_pt\_mo\*hiv\_inc(**0**)=sti condom\_use np zip\_risk sti\_condom;

baseline covariates=km\_plot out=out1 survival=s1;

**run**;

**proc** **sgplot** data=out1;

step y=s1 x=hiv\_pt\_mo;

**run**;

\* Plot the Poisson regression model;

**data** exp\_plot;

do t=**0** to **25** by **0.1**;

sti=**1**;

s=exp(-t\*exp(-**6.4298**+**0.6975**+**0.0514**\***0.4711014**+**1.5459**\***0.46237730**));

output;

sti=**0**;

s=exp(-t\*exp(-**6.4298**+**0.0514**\***0.4711014**+**1.5459**\***0.4623773**));

output;

end;

**run**;

**data** exp\_plot;

set exp\_plot out1;

**run**;

title "KM estimates and Poisson survival regression curves overlaid (Non-condom use)";

**proc** **sgplot** data=exp\_plot;

series y=s x=t/group=sti;

step y=s1 x=hiv\_pt\_mo/group=sti;

**run**;

title;