MSCR 534: Problem Set 1

Anish Shah

February 28, 2020

Question 1 - Frequency Distribution Data

Use the frequency distribution data in Table 1 to answer questions 1A through 1D.

Table 1:

Outcome	Exposure=1	Exposure=0
Level=0	69	75
Level=1	72	66
Level=2	77	52

1A. State the expressions for a nominal logistic model to estimate the crude association between the exposure and the outcome variable. Choose level 0 as the reference group for the outcome.

$$logit(Outcome = 2) = \beta_0 + \beta_1 Exposure$$
$$logit(Outcome = 1) = \beta_0 + \beta_1 Exposure$$

- Reference for dependent variable is *Outcome=0*
- Nominal levels of *Outcome* include: 0, 1, 2

1B. State the expression for an ordinal logistic model to estimate the crude association between the exposure and outcome variable.

$$ln\left(\frac{P(O \geq g|E)}{P(O < g|E)}\right) = \beta_g + \beta_1 Exposure$$

. . . where g = (1, 2)

1C. Calculate the crude odds ratios (for the effect of the exposure on the outcome) that would result from part 1a.

$$OR_{O=2|E} = 1.39493$$
$$OR_{O=1|E} = 1.02767$$

1D. Calculate three crude odds ratios for an ordinal logistic model: two odds ratios to evaluate the proportional odds assumptions and a third odds ratio that would result from an ordinal logistic model that assumed the odds were proportional.

$$(1) : OR_{\frac{O \geq 2|E}{O<2|E}} = 1.37575$$
$$(2) : OR_{\frac{O \geq 1|E}{O<1|E}} = 1.18951$$
$$(3) : OR_{proportional} = 1.28263$$

Question 2 - Interpreting Interaction

Consider a binary logistic model with 4 treatment levels and 3 levels of ethnicity, with indicator dummy variables coded as follows:

Treatment	T1	T2	T3
1	1	0	0
2	0	1	0
3	0	0	1
4	0	0	0

Ethnicity	E1	E2
1	1	0
2	0	1
3	0	0

$$
\begin{aligned}
(1) : logit(D=1) &= \beta_0 + \beta_1 T1 + \beta_2 T2 + \beta_3 T3 + \beta_4 E1 + \beta_5 E2 \\
(2) : logit(D=1) &= \beta_0 + \beta_1 T1 + \beta_2 T2 + \beta_3 T3 + \beta_4 E1 + \beta_5 E2 \\
&\quad + \beta_6 T1 \times E1 + \beta_7 T1 \times E2 + \beta_8 T2 \times E1 + \beta_9 T2 \times E2 \\
&\quad + \beta_{10} T3 \times E1 + \beta_{11} T3 \times E2
\end{aligned}
$$

2A. What is the interpretation of exp(β_1) using Model 1?

The $e^{\beta_1} = OR$ for the effect of Treatment 1 (reference group: Treatment 4) on the Outcome with adjustment for ethnicity.

2B. What is the interpretation of exp(β_1) using Model 2?

The $e^{\beta_1} = OR$ for the effect of Treatment 1 (reference group: Treatment 4) on the Outcome with adjustment for not only ethnicity, but the interaction between the treatment levels and ethnicitiy levels. The value of β_6 and β_7 should also be considered for their interaction with Treatment 1 with both Ethnicity 1 and Ethnicity 2.

2C. What is the odds ratio comparing Treatment=3 vs Treatment=4 among those with Ethnicity=1 using Model 1?

- T1 = 0
- T2 = 0
- T3 = 1
- E1 = 1
- E2 = 0

$$
\begin{aligned}
OR &= \frac{e^{\beta_0 + \beta_1 T1 + \beta_2 T2 + \beta_3 T3 + \beta_4 E1 + \beta_5 E2}}{e^{\beta_0 + \beta_1 T1 + \beta_2 T2 + \beta_3 T3 + \beta_4 E1 + \beta_5 E2}} \\
&= \frac{e^{\beta_3 (T3=1) + \beta_4 (E1=1)}}{e^{\beta_4 (E1=1)}} \\
&= e^{\beta_3 (T3=1)} \\
&= e^{\beta_3}
\end{aligned}
$$

2D. What is the odds ratio comparing Treatment=3 vs Treatment=4 among those with Ethnicity=1 using Model 2?

- T1 = 0
- T2 = 0
- T3 = 1
- E1 = 1
- E2 = 0

$$
\begin{aligned}
OR &= \frac{e^{\beta_3(T3{=}1)+\beta_4(E1{=}1)+\beta_6 T1\times(E1{=}1)+\beta_8 T2\times(E1{=}1)+\beta_{10}(T3{=}1)\times(E1{=}1)+\beta_{11}(T3{=}1)\times E2}}{e^{\beta_4(E1{=}1)+\beta_6 T1\times(E1{=}1)+\beta_8 T2\times(E1{=}1)}} \\
&= \frac{e^{\beta_3(T3{=}1)+\beta_4(E1{=}1)+\beta_{10}(T3{=}1)\times(E1{=}1)}}{e^{\beta_4(E1{=}1))}} \\
&= e^{\beta_3(T3{=}1)+\beta_{10}(T3{=}1)\times(E1{=}1)} \\
&= e^{\beta_3+\beta_{10}}
\end{aligned}
$$

Question 3 - Nilton Data

Use the permanent SAS dataset named "Nilton" on Canvas (in Problem Set 1 assignment) to answer questions 3A through 3E. The data came from a cross-sectional study of inpatients with Methicillin-resistant staph aureus (MRSA). For simplicity, ignoring missing values in all answers below.

The dataset contains the following variables: METHICSE (dichotomous outcome of interest, coded 1 for MRSA), AGE (continuous), AGECAT (dichotomous coded 1 if age >= 55), PREVHOSP (dichotomous coded 1 if hospitalized in the previous 6 months), SEX (dichotomous coded 1 for male), PREANTBU (dichotomous coded 1 for antibiotic use in the previous 3 months). The dichotomous variables are all coded 1 or 0.

3A. Suppose you model AGE and SEX as a predictor of METHISCE in a logistic regression. State the model in terms of the prevalence of MRSA.

$$ln(P = MRSA) = \beta_0 + \beta_1 AGE + \beta_2 SEX$$

3B. Run the model from 3A to estimate the predicted prevalence of MRSA for a 50 year-old male.

$$P_{MRSA} = 0.72593$$

3C. Run a binary logistic model to estimate the association between previous hospitalization with prevalent MRSA. In an adjusted model, control of AGECAT, SEX, and PREANTBU. What are the crude and adjusted odds ratios? Interpret the adjusted odds ratio in one sentence.

Association between prior hosptialization and MRSA

	Dependent variable:	
	METHICSE	
	Crude	Adjusted
	(1)	(2)
Previous Hospitalization	11.700***	4.850***
	(6.190, 23.800)	(2.250, 11.000)
Age > 55		3.400***
		(1.810, 6.570)
Sex		2.290**
		(1.210, 4.460)
Prior Antibiotics		5.330***
		(2.650, 11.100)
Constant	0.118***	0.028***
	(0.061, 0.205)	(0.010, 0.069)
Observations	292	289
Log Likelihood	−160.000	−140.000
Akaike Inf. Crit.	323.000	290.000
Note:		*p<0.1; **p<0.05; ***p<0.01

The $OR = 4.85228$ for prevalent MRSA in those with prior hospitalization after adjustment for age > 55 years, sex, and prior antibiotic usage.

3D. Run a model with three product terms between the exposure (previous hospitalization) with age category, sex, and antibiotic use. The additional three terms should be the product of PREVHOSP with AGECAT, SEX, and PREANTBU respectively. Include all three product terms in one model. What is the odds ratio for prevalent MRSA in patients with previous hospitalization in the past 6 months compared to those without hospitalization among women less than 55 years old who did not use antibiotics in the past three months?

$$OR_{MRSA} = 2.47278$$

3E. Run a likelihood ratio test for the addition of three product terms for interaction in part 3D (between previous hospitalization with age category, sex, and antibiotic use). Test the null hypothesis that all three beta coefficients for the interaction terms = 0). What is the -2ln likelihood for the full and reduced models? How many degrees of freedom are there for the likelihood ratio test? What is the p-value for the result? What is your interpretation about interaction?

This is a case of comparing nested models, as the full model is the reduced model with the addition of interaction terms. Testing against the null hypothesis, that the interaction terms all have a beta-coefficient of 0, we can use the likelihood ratio test.

- The -2LogLikelihood for the reduced model is -140.21493
- The -2LogLikelihood for the full model is -139.29045

For the likelihood ratio test, there are 3 degrees of freedome. The $\chi^2 = 1.84896$, with a $p = 0.60434$. This suggests we do not have enough evidence to reject the null hypothesis, and can accept that the interaction terms to have $\beta = 0$.

There are no significant interactions between PREVHOSP and AGECAT/SEX/PREANTBU. The full model including interactions is not more informative or a better fit than the reduced model.

Question 4 - Kaplan Meier Curves with Censoring

A prospective student is considering her course schedule and trying to determine if she should take MSCR534. Use the tables below to determine the two-year survival of two groups, those who recently took MSCR 534 (534=1) and those that did not (534=0).

4A. Fill in the missing cells in Table 2 and Table 3.

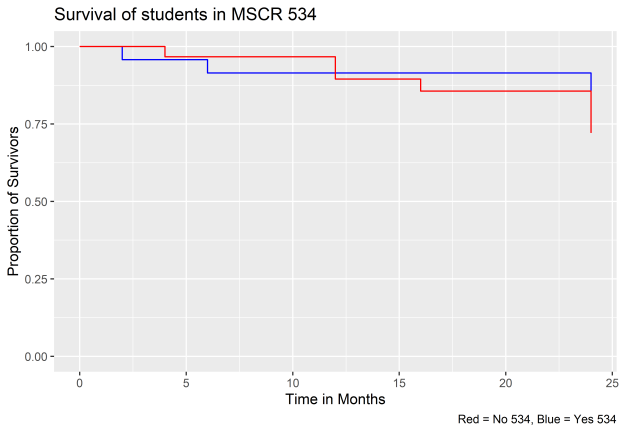
Table 2:

Event time (Months)	Number of s urvivors	Number of deaths	Number censored	Pr oportion s urviving	534 status
0	24	0	0	1	1
2	24	1	1	0.958	1
6	22	1	2	0.915	1
12	19	0	2	0.915	1
24	17	2	3	0.807	1

Table 3:

Event time (Months)	Number of s urvivors	Number of deaths	Number censored	Pr oportion s urviving	534 status
0	30	0	0	1	0
4	30	1	2	0.967	0
12	27	2	2	0.895	0
16	23	1	3	0.856	0
24	19	3	4	0.721	0

4B. Graph the Kaplan Meier curve for each of the tables (by hand is fine). Graph the survival curves on the same graph. Hint: the Y axis should be from 0 to 1.0 and the X axis should be from 0 to 24 months.



4C. If the data above were true, would you take MSCR 534 again?

It appears that the survival trend for those that have not taken MSCR 534 is initially higher, but by two years, they are doing what appears to be significantly worse. I think I would take MSCR 534 again for that ~8% survival benefit it appears to confer over the long-term.

Question 5 - Midterm Question with Solution

Write a midterm exam question focused on nominal/ordinal logistic regression, survival analysis, or Cox proportional hazards models. Also include a solution to the question.

There is a longitudinal study on patients with lung cancer amongst several institutions. The event of interest was dying of lung cancer. Other measures included patient age, sex, weight loss, average meal size, and functional status (both ECOG and Karney scales). The investigation is evaluating the most important predictors of death by lung cancer with age, sex, weight loss, and functional status (**age**, **sex**, **wt.loss**, **ph.ecog**). This dataset can be download as **lung.csv**, (*meta: attached with assignment submission*).

5A. Assume that the proportional hazard model is correct. What are the most important predictors of death by lung cancer?

```
# Require libraries
library(tidyverse)
library(survival)
library(survminer)

# Data set
df <- as_tibble(lung)
write_csv(lung, "lung.csv")

# Set up survival object
df$survobj <- with(df, Surv(time, status))

# Model
m <- coxph(survobj ~ age + sex + wt.loss + ph.ecog, data = df)

# Display HR
stargazer(m, type = "latex", header = FALSE,
          title = "Association between Predictors of Lung Cancer Death",
          no.space = FALSE, single.row = FALSE,
          apply.coef = exp,
          ci = TRUE, p.auto = FALSE, report = "vc*s",
          ci.custom = list(exp(confint(m))),
          covariate.labels = c("Age", "Sex", "Weight Loss", "ECOG"),
          column.labels = c("Adjusted Model"),
          table.placement = "H"
)
```

Association between Predictors of Lung Cancer Death

	Dependent variable:
	survobj
	Adjusted Model
Age	1.013
	(0.995, 1.033)
Sex	0.554***
	(0.393, 0.781)
Weight Loss	0.991
	(0.978, 1.004)
ECOG	1.674***
	(1.308, 2.143)
Observations	213
R^2	0.136
Max. Possible R^2	0.998
Log Likelihood	−659.510
Wald Test	29.940*** (df = 4)
LR Test	31.021*** (df = 4)
Score (Logrank) Test	30.646*** (df = 4)
Note:	*p<0.1; **p<0.05; ***p<0.01

5B. Evaluate if the proportional hazard assumption can be made using Schoenfeld residuals. Plot the residuals. Interpret the findings.

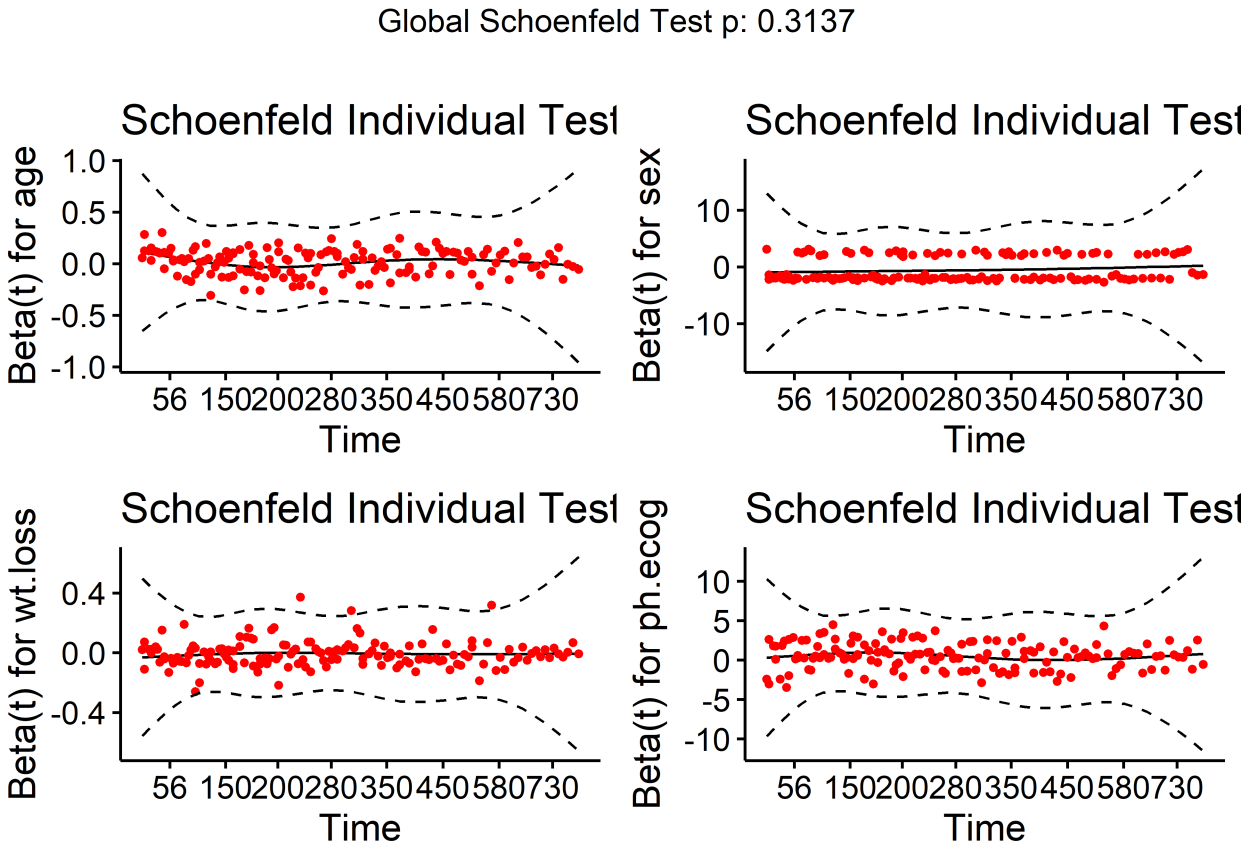
```
# Test PH assumption
t <- cox.zph(m)

# Display findings
t$table %>%
    kable("latex", booktabs = TRUE,
          caption = "Tests of PH Assumption for each Predictor") %>%
    kable_styling(latex_options = "hold_position")
```

Tests of PH Assumption for each Predictor

	chisq	df	p
age	0.43529	1	0.50941
sex	2.67313	1	0.10206
wt.loss	0.04571	1	0.83071
ph.ecog	1.63551	1	0.20094
GLOBAL	4.75158	4	0.31375

The p-values reported for the individual predictors suggest there is not enough evidence to reject the null hypothesis, as well as the global test, suggesting that the proportional hazards assumption can be accepted. This can be visualized with the graphical interpretation of the Schoenfeld residuals.

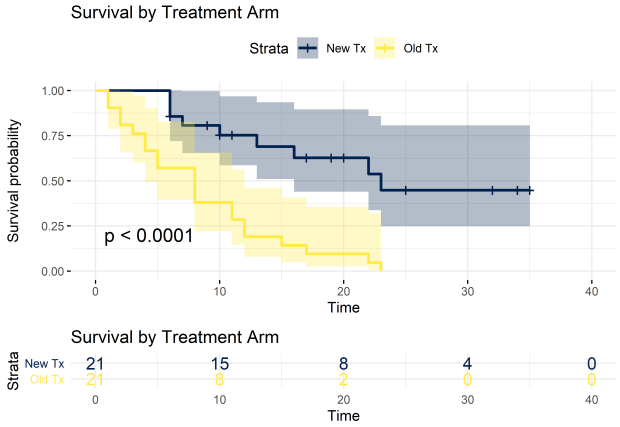
```
# Plot residuals
ggcoxzph(t)
```



Question 6 - Anderson Data

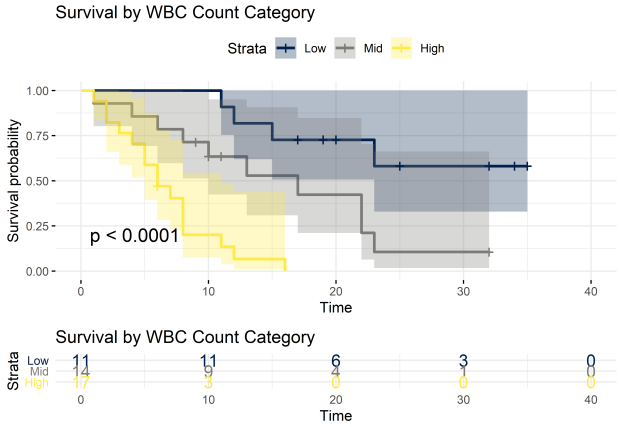
Use the permanent dataset on Canvas named "Anderson" (in Problem Set 1 assignment) to answer questions 6A through 6D. The dataset consists of remission survival times (SURVT), measured in weeks, on 42 leukemia patients, half of whom get a new therapy (RX=0) and half of whom get a standard therapy (RX=1). Control variables are SEX (1=male, 0=female) and log white blood cell count. The log white blood cell count is in two forms, (LOGWBC - continuous) and a three-level categorical variable (LWBC3). The variable STATUS indicates event (out of remission - coded 1) or censorship (coded 0).

6A. Run PROC LIFETEST three times to perform log rank tests for the effects of 1) treatment, 2) log white blood cell count, and 3) for the effect of gender. State the null hypotheses (there are three) and report the p-values and decision. Examine the survival plots. Why do you think the log rank test for gender was not significant even though the estimated survival curves (for gender) look different? Examine the log (-log) survival plots; does the PH assumption seem violated for any of the predictors?



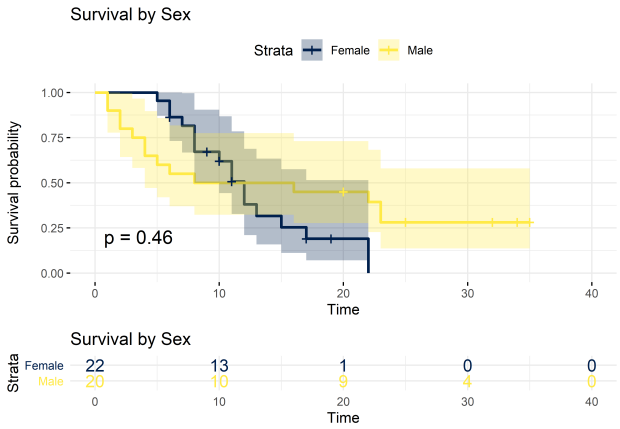
- H_0: The survival curves are the same between treatment arms.
- $\chi^2 = 16.79294$
- p = 0.00004

We can see that treatment has a visually different survival curve (the new treatment arm has much better survival). The test statistics suggests that we have enough evidence to reject the null and conclude that the two curves are different.



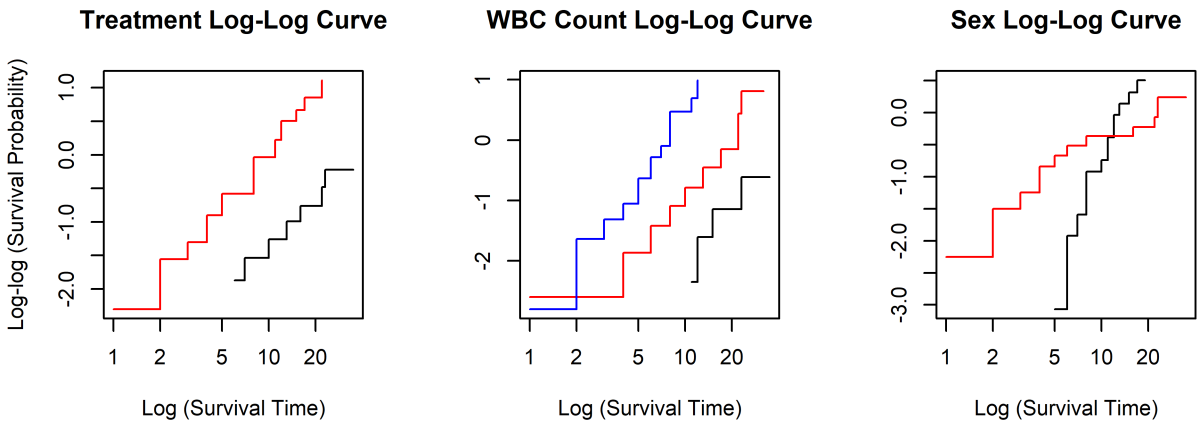
- H_0: The survival curves are the same between WBC categories.
- $\chi^2 = 26.39063$
- p = 0

We can see that WBC category each visually have different curves. The test statistics suggests that we have enough evidence to reject the null and conclude that the three curves are different.



- H_0: The survival curves are the same sex curves.
- $\chi^2 = 0.55686$
- p = 0.45553

We can see that the sex curves seem to cross, but that males appear to have a better survival. Although they appear to be visually different, the test statistics suggest that there is insufficient evidence to reject the null hypothesis, and thus we can conclude that the curves are not different (at least in this model). The issue at hand may be the fact hte curves cross in the middle, suggesting differences in survival at different time periods, which may be confounding the relationship.



When examining the log-log plots, we can visually assess the proportional hazards assumption. The plots appear proportional for both Treatment and WBC, but not for Sex. There appears to be a cross over at ~ 10 months, and thus we should consider additional analysis after separting by time.

6B. For part 6B assume the PH assumption is not violated for any of the variables. Run a Cox model with RX, LOGWBC, and SEX in the model. State the model in terms of the hazard

function. What are the estimated crude hazard ratio and adjusted hazard ratio for treatment? Interpret the adjusted hazard ratio in one sentence.

$$Crude: h(t) = h_0 t \times e^{\beta_1 RX}$$
$$Adjusted: h(t) = h_0 t \times e^{\beta_1 RX + \beta_2 LOGWBC + \beta_3 SEX}$$

Association between Treatment and Survival

	Dependent variable:	
	survobject	
	Crude	Adjusted
	(1)	(2)
Treatment	4.817***	4.498***
	(2.147, 10.809)	(1.820, 11.113)
WBC Count		5.376***
		(2.779, 10.398)
Sex		1.370
		(0.562, 3.338)
Observations	42	42
R^2	0.322	0.675
Max. Possible R^2	0.988	0.988
Log Likelihood	−85.008	−69.590
Wald Test	14.530*** (df = 1)	33.540*** (df = 3)
LR Test	16.352*** (df = 1)	47.188*** (df = 3)
Score (Logrank) Test	17.247*** (df = 1)	48.015*** (df = 3)
Note:		*p<0.1; **p<0.05; ***p<0.01

The adjusted hazard ratio suggests that those that received standard treatment (RX=1) have a hazard rate 4.5 times that of the new treatment (RX=0), after adjustment for WBC count and sex.

6C. Suppose it is decided that the PH assumption is violated just for the gender variable. Run a stratified Cox model for the effect of treatment, adjusted for logwbc, with gender as the stratified variable (assuming no interaction with treatment). What is the estimated hazard ratio for the effect of treatment?

Association between Treatment and Survival with Gender Stratification

	Dependent variable:
	survobject Stratified Model
Treatment	2.713**
	(1.072, 6.864)
WBC Count	4.279***
	(2.180, 8.398)
Observations	42
R^2	0.534
Max. Possible R^2	0.967
Log Likelihood	−55.735
Wald Test	22.750*** (df = 2)
LR Test	32.060*** (df = 2)
Score (Logrank) Test	30.798*** (df = 2)
Note:	*p<0.1; **p<0.05; ***p<0.01

6D. Now include a treatment-gender interaction term in the model. Create this interaction term in a SAS data step. Note that SEX should not be in the model statement, but in the strata statement. What is the Wald test p-value for the product term coefficient? What are the estimated hazard ratios for treatment using the model with the product term?

Association between Tx and Survival with Sex Interaction

	Dependent variable:
	survobject
RX	1.332
	(0.437, 4.059)
LOGWBC	4.361***
	(2.188, 8.689)
RX:SEX	5.166*
	(0.861, 30.991)
Observations	42
R^2	0.568
Max. Possible R^2	0.967
Log Likelihood	−54.127
Wald Test	23.880*** (df = 3)
LR Test	35.276*** (df = 3)
Score (Logrank) Test	33.149*** (df = 3)
Note:	*p<0.1; **p<0.05; ***p<0.01

The adjusted hazard ratio is seen in the above table. The p-value for the product term coefficient is 0.07243.