Biostatistics 140.623
Third Term, 2014-2015
Final Examination- Answer Key
March 12, 2015

Instructions: You will have two hours for this examination. There are 20 problems. The formula page and Stata output are at the **back** of the exam for your use. Please note that statistical significance is defined by $p < 0.05$.

Questions 1-4 test general knowledge:

1. What is the main purpose of the Cox regression model? (*Circle only one response*).
   a) To estimate the survival function for a time-to-event outcome using binned data.
   b) To estimate the baseline hazard function for a time-to-event outcome under the assumption that the relationship is linear on the log scale.
   c) To test the assumption of proportional event hazards between risk factor groups.
   d) To estimate and make inferences about relative event hazards between risk factor groups.

      The Cox regression model is used when exact time-to-event is known (ungrouped survival data) and it models the log hazard of the event as a function of a log baseline hazard plus covariates. The model assumes nothing about the shape of the relationship between the log(hazard) and time. The goal is to estimate and interpret the regression coefficients in order to make inferences about associations between risk factors and the hazard of the event. An assumption of proportional hazards is made and must be checked, but is not the goal.

   e) To determine whether the number at risk relates to covariates.

2. Suppose that you were interested in assessing differences in time to death by treatment group (drug versus placebo) and that the calculated log-rank test statistic for treatment equals 0.10, which is approximately a chi-squared statistic with one degree of freedom. The null hypothesis that corresponds to this test statistic is: *(Circle only one response.).*
   a) There are more deaths in the drug group.
   b) There are more deaths in the placebo group.
   c) There are equal numbers of deaths in the drug and placebo groups.
   d) There is a difference in median survival between the drug and placebo groups.
   e) There is no difference in the overall hazard of death between the drug and placebo groups.

      The log-rank statistic is testing the null hypothesis of no difference in overall survival between the two groups; this is the same as testing the null hypothesis of no difference in the overall hazard of death between the two groups. It is based on a summed and weighted comparison of observed versus expected numbers of deaths (under the assumption of no difference by group) at each time an event occurs.

3. The following is a Poisson regression model with 4 (follow-up) time bins (1-4) and treatment covariate defined as trt=1 for Treatment A; 0 for Treatment B; and indicator variables for time bins 2, 3, and 4.

$$\log(\text{expected events in bin j}) = \log(\text{person-weeks in binj}) + \beta_0 + \beta_1(\text{time bin 2}) + \beta_2(\text{time bin 3}) + \beta_3(\text{time bin 4}) + \beta_4 \text{trt}$$

This model assumes that: (*Circle only one response*)

a) The hazard of an event is constant within a time bin but varies across time bins.

By adding indicator variables for the 3 time bins into the model, we allow the hazard of the event to vary by time bin. However, within each time bin, the hazard is assumed to be constant.

b) The hazard of an event may vary across time bins but increases within a time bin.
c) The hazard of an event is constant across time bins.
d) The relative hazard of an event for Treatment A versus Treatment B changes across time bin.
e) The relative hazard of an event for time bin j+1 versus time bin j varies by treatment.

4. The AIC (Akaike Information Criterion) is a measure that can be used for: *Circle only one response*)

a) Assessing model goodness of fit.
b) Comparing observed versus expected outcomes.
c) Aiding in model selection based on the model log-likelihood and number of parameters.

For any generalized linear model, the AIC is calculated as –LL + 2(model df) where LL = log- likelihood of the model and model df is the number of parameters.

d) Identifying statistically significant covariates in a model.
e) Checking the underlying model assumption of independence of observations.

Questions 5 through 8 concern data from a study investigating the association between **sleep latency** (the amount of time needed for an individual to fall asleep at night) and **demographic characteristics**. **Models A-D** on pages 11-12 show logistic regression results.

The outcome Y = Slp15 =1 if sleep latency > 15 minutes; = 0 if ≤ 15 minutes

Demographic characteristics are:
age in years

female =1 if female; 0 if male
smk = 0 if never; 1 if current; 2 if former smoker
BMI in kg/m2
bmicat - BMI category
1 if < 18.5 kg/m^2
2 if 18.5 – 24.9 kg/m^2
3 if 25-29.9 kg/m^2
4 if ≥ 30 kg/m^2

5. If age had been centered at the median age of 61 years in **Model A, what would be the values of the estimated regression coefficients**? (*Circle only one response*).

a) b_0= -1.27 and b_1=0.019
b) b_0= -0.13 and b_1=0.019

From Model A, we can write log(odds) = $\beta_0+\beta_1$age where age is not centered and b_0=-1.272739 and b_1=0.018698.

If we center age at 61, then the new b_0 will equal the log(odds|age =61) = $b_0+b_1(61)$= -1.272739 + 0.018698(61) = -0.13.

The slope for age, b_1, will remain the same.

c) b_0= -1.27 and b_1=1.16
d) b_0= 1.16 and b_1=0.019
e) b_0= 0.019 and b_1=1.16

6. From **Model B**, we would conclude that the odds ratio (please note: while it does not change the correct answer for the question, this should have read "we would conclude that the odds") for sleep latency > 15 minutes to fall asleep at night: (*Circle only one response*).

a) Statistically significantly increases with each year of age for all individuals.
b) Statistically significantly increases with each year of age for individuals aged 55-65 years but not in younger nor in older individuals.

From Model B, we can write:
log(odds) = $\beta_0 + \beta_1$age1 + $\beta_2$age2 +$\beta_3$age3
where the age spline terms are defined by:
log(odds) = $\beta_0 + \beta_1$age + β_2(age-55)$^+$ +β_3(age-65)$^+$
such that (age-55)$^+$= 0 if age ≤ 55; = (age-55) if age > 55
and (age-65)$^+$= 0 if age ≤ 65; = (age-65) if age > 65

Thus the "slopes" or the change in log odds of sleep latency > 15 is:

β_1 for ages ≤ 55	**β_1 is estimated by b_1, the coefficient for age1 (Z=-0.36, p=0.722 ns)**
$\beta_1 + \beta_2$ for ages 55 – 65	**$\beta_1 + \beta_2$ is estimated by $b_1 + b_2$, the sum of the coefficients for age1 and age 2 (Z=2.67, p=0.008 using lincom)**
$\beta_1 + \beta_2 + \beta_3$ for ages > 65	**$\beta_1 + \beta_2 + \beta_3$ is estimated by $b_1 + b_2 + b_3$, the sum of the coefficients for age1 and age 2and age3 (Z= -1.42, p=0.154 using lincom, ns)**

c) Statistically significantly decreases with each year of age for individuals aged > 65 years but not in younger individuals.
d) Statistically significantly decreases with each year of age for individuals < 55 years and > 65 years but not in individuals aged 55-65 years.
e) Is not statistically significantly associated with age in these individuals.

7. The results of the Likelihood Ratio Test of the Extended **Model D** to the Null **Model C** suggest that: (*Circle only one response*).

a) BMI category does not contribute to the model of sleep latency beyond what is predicted by smoking status.
b) Smoking status does not contribute to the model of sleep latency beyond what is predicated by BMI.
c) Neither BMI nor smoking status contributes to the model of sleep latency.
d) Taken together, BMI and smoking statistically significantly contribute to the model of sleep latency.
e) Taken together, BMI and smoking status statistically significantly contribute to the model of sleep latency beyond what is predicted by age and its spline terms, and sex.

Model C is the null model and includes age plus age spline terms and sex.
Model D is the extended model and includes age plus age spline terms, sex, smoking categories and BMI categories.

The LRT tests the Ho: None of the variables in the extended model contribute beyond that contributed by the variables in the null model. (i.e: the coefficients for these additional variables are all equal to 0)

8. Suppose that, instead of handling BMI as a categorical variable, that BMI was used as a continuous variable using spline terms with knots at 18.5, 25, and 30 kg/m^2 using the following Stata command:

```
.mkspline bm1 18.5 bm2 25 bm3 30 bm4= bmi, marginal
```

The interpretation of the coefficient for **bm3** would be: (*Circle only one response*).

a) The adjusted difference, between individuals with BMI 25-29 kg/m^2 and those with BMI 18.5 – 24.9 kg/m^2, in the log odds of sleep latency > 15 minutes.

b) The difference, between individuals with BMI 25-29.9 kg/m^2 and those with BMI 18.5-24.9 kg/m^2, in the adjusted change in the log odds of sleep latency > 15 minutes with each kg/m^2 increase in BMI.

Using the mkspline command above with the marginal option, the model is:
log(odds) = $\beta_0 + \beta_1$bmi1 + β_2 bmi2 +$\beta_3$bmi3+$\beta_4$bmi4+ ….
which is interpreted as
log(odds) = $\beta_0 + \beta_1$bmi1 + β_2(bmi-18.5)$^+$ +β_3(bmi-25)$^+$ +β_4(bmi-30)$^+$
β_1 is the change in the log odds of sleep latency > 15 with each unit increase in BMI, for individuals with BMI < 18.5.
$\beta_1 + \beta_2$ is the change in the log odds of sleep latency > 15 with each unit increase in BMI for individuals with BMI 18.5 -24.9
$\beta_1 + \beta_2 + \beta_3$ is the change in the log odds of sleep latency > 15 with each unit increase in BMI for individuals with BMI 25 -29.9
$\beta_1 + \beta_2 + \beta_3 + \beta_4$ is the change in the log odds of sleep latency > 15 with each unit increase in BMI for individuals with BMI ≥ 30

Thus, β_3 is the difference in "slope" (or change in log odds per unit increase in BMI) between individuals with BMI 25-29.9 and individuals with BMI 18.5-24.9.

c) The adjusted change in log odds of sleep latency > 15 minutes with each kg/m^2 increase in BMI among individuals with BMI 25-29.9 kg/m^2.

d) The adjusted log odds of sleep latency > 15 minutes in individuals with BMI ≥ 30 kg/m^2.

e) The adjusted change in average log odds of sleep latency > 15 minutes with each kg/m2 increase in BMI in individuals with BMI ≥ 30 kg/m^2.

Questions 9 -11 reflect Poisson regression models of lung cancer deaths by age groups and population at risk in each age group. Variables are:

Age in age categories: < 45, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 85-80, > 80

Smoking category (**smoke**):
1-smokes **neither** cigarettes nor cigars/pipes
2- smokes cigars/pipes **only**
3- smokes **both** cigarettes and cigars/pipes
4- smokes cigarettes **only**

Poisson regression **Models 1-3** are found on **pages 13-14.**

9. Comparing **Models 1-3**, it can be concluded that: (*Circle only one response*).

a) Smoking confounds the relationship between age and lung cancer death rate.
b) Age confounds the relationship between smoking and lung cancer death rate.

Model 1contains only age; age is associated with lung cancer death rate.
Model 2 contains only smoking; smoking is associated with lung cancer death rate.
Model 3 contains both age and smoking. We observe that the estimate smoking coefficients substantially change in Model 3 as compared to Model 2, suggesting that age appears to confound the relationship between smoking and lung cancer death rate since age is associated with lung cancer death rate (and we assume that it is also associated with smoking).

c) Smoking modifies the relationship between age and lung cancer death rate.
d) Age modifies the relationship between smoking and lung cancer death rate.
e) Both smoking and age are mediators of the relationship between lung cancer deaths and the population at risk.

10. In **Model 2** which contains only smoking categories, it is assumed that: (*Circle only one response*).
a) The incidence rate ratio of lung cancer death by smoking status is the same across age categories.
By not including age categories, the Poisson model assumes a constant incident rate across ages and the incidence rate ratio would be proportional and independent of age.

b) The incidence rate ratio of lung cancer death by smoking status varies by age category.
c) The incidence (hazard) of lung cancer death is not constant across age categories.
d) The incidence of lung cancer death changes linearly with age.
e) The incidence of lung cancer death is proportional to age.

11. From the **Model 3 output** we can see that, after controlling for age, the incidence rate of lung cancer death is significantly greater in individuals smoking cigarettes only as compared to both cigarettes and cigars/pipe. This is supported by: (*Circle only one response*).

a) log(IRR) =0.218, Z=5.63, p=0.0
b) log(IRR)=0.417, Z=10.45, p=0.0
c) IRR=1.22, Z=8.39, p=0.0
This is given by exponentiating $b_{11} - b_{10}$ in Model 3. Using the lincom command (with the irr option), we see that the IRR is estimated by exp(0.4169596 – 0.2179552) = 1.22, Z=8.39, p=0.000.
d) IRR=1.44, Z=9.74, p=0.0
e) LR χ_1^2= 4034, p=0.0

Questions 12 through 15 concern the results from a randomized clinical trial of percutaneous coronary intervention (PCI) in patients with STEMI (acute ST-segment elevation myocardial infarction).

The researchers used simple **Cox regression** to measure the association between the primary outcome (a composite of death from cardiac causes, nonfatal myocardial infarction, or refractory angina) and treatment (PCI versus control). The model used is:

ln(hazard of primary outcome at time t) = ln(λ_0[t]) + $\beta_1 x_1$
where x_1= 1 for PCI intervention group and 0 for control group,
and t represents time in the follow-up period (0 – 36 months).

12. What does the function $\lambda_0(t)$ represent in the Cox regression equation? (*Circle only one response*).

a) The hazard of the primary outcome in the PCI group at time = 0.
b) The hazard of the primary outcome in the control group at time=0.
c) The hazard ratio of the primary outcome for the PCI group compared to the control group at time=0.
d) The hazard of the primary outcome in the control group as a function of time across the follow-up period.

Recall, the "intercept" for a Cox proportional hazards model is a function of time: it tracks the ln(hazard) of the outcome over time, for the group whose predictor values are all 0. In this case, there is only one dichotomous predictor, x_1: $x_1 = 0$ for those randomized to the control group.

e) The difference in the ln(hazard) of primary outcome between the PCI and control groups at an time in the follow-up period.

13. What assumption did the researchers have to make in order to use Cox regression to quantify the relationship between the primary outcome and PCI (versus control)? (*Circle only one response*).
    a) The relationship between the primary outcome and PCI is statistically significant.
    b) The hazard of the primary outcome is constant over time in both the PCI and control groups.
    c) PCI will reduce the hazard of the primary outcome by at least 20%.
    d) The relationship between the ln(hazard) of the primary outcome and time is linear.
    e) The ratio of the hazard of the primary outcome for the PCI group compared to the control group is constant over the 36 month follow-up period.

    This is the assumption of proportional hazards.

The following shows the Kaplan-Meier curve estimates of the cumulative probability of being without the primary outcome in the **PCI** (Preventive PCI) and **control** groups (No Preventive PCI).

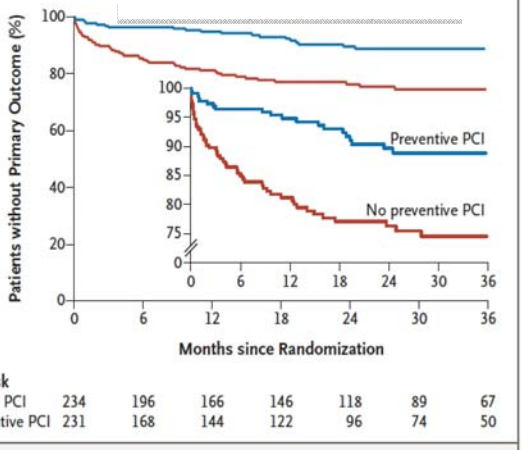

Figure 2. Kaplan–Meier Curves for the Primary Outcome.
The primary outcome was a composite of death from cardiac causes, nonfatal myocardial infarction, or refractory angina. The inset graph shows the same data on a larger scale. All patients in the trial underwent infarct-artery PCI immediately before randomization.

14. Based on the Kaplan-Meier curves above, what can be stated about the estimated value of β_1 from the Cox regression model given on page 5? (*Circle only one response*).
    a) $b_1 > 0$
    b) $b_1 < 0$

    Based on the KM curves, the Preventive PCI group has better survival (a larger percentage of persons in this group do not have the primary outcome of death over the follow-up period, as compared to the control group). As such, the Preventive PCI group has a lower hazard than the control group. The estimated hazard ratio comparing the hazard of the primary outcome for the Preventive PCI group to the control group over the follow-up period is less than 1. Since e^{b1} is this estimated hazard ratio, then b_1 , the ln(hazard ratio), will be less than 0.

    c) $b_1 = 0$
    d) This cannot be answered without being given a specific time, and value of $\hat{\lambda}_o[t]$ at this specified time.
    e) This cannot be answered because there is no relationship between Kaplan-Meier curve estimates and the hazards of the primary outcome.

15. There were 234 patients randomized to the treatment group, and 67 still at risk of mortality at 36 months. In other words, 29% of the treatment group was still at risk of death at 36 months. However, the corresponding Kaplan-Meier curve estimate at 36 months for the treatment group is nearly 90%. How can this have happened? (*Circle only one response*).

    a) Some of the observations in the treatment group were censored (lost to follow-up or completed the study alive) in the Kaplan-Meier estimates.

    If there were no censoring in these data, then the proportion of the treatment group surviving beyond 36 month, $\hat{S}(36)$, would be 67/234 = 29%. However, here the estimate of $\hat{S}(36)$ is nearly 90%. Recall that

    $$\hat{S}(36) = \Pr(\text{Surviving beyond 36 months} \mid \text{still around right before 36 months}) \times \hat{S}(35+).$$

    b) The researchers estimated the Kaplan-Meier curve using only the data on patients who died in the follow-up period.
    c) The researchers do not know how to properly estimate Kaplan-Meier curves.
    d) The Kaplan-Meier curve estimate at 36 months ($\hat{S}(36)$) is the risk of surviving among only those who were still alive and enrolled in the study at 36 months.
    e) The researchers grouped the data into one-week time bins prior to plotting the survival curves.

Questions 16-20 involve data from the UMARU impact study, a randomized trial of 595 subjects between 20 and 50 years old, with a substance abuse issue to assess the relative efficacy of long term versus short term residential drug treatment programs. Subjects were followed for up to 39 months after the start of treatment.
The following are baseline covariates in Cox regression **Models W- Z** which are found on **pages 15-17.**

treat: 1 for long-term, 0 for short-term treatment
age_cat: takes on values 1-6 for 5-year age intervals; the age range in each of the intervals are [20, 25), [25, 30), [30, 35), [35,40), [40, 45) and [45, 50].
white: 1 if subject identifies as white, 0 if non-white.
iv_druguse: 1 if subject was using intravenous (IV) drug at time of enrollment, 0 if not

16. Based on the result from **Model W**, what is the unadjusted hazard ratio (and 95% CI) of relapse for the long-term treatment group compared to the short term treatment group at 24 months after randomization? (*Circle only one response*).

a) -0.24 (-0.42, -0.06)
b) 0.24 (0.06, 0.42)
c) 0.79 (0.66, 0.94)

In model W, b_1 = ln(hazard ratio) = -.23 with 95%CI (-.42, -.06). Exponentiating these results gives the estimated hazard ratio of 0.79, and corresponding 95% CI of 0.66 to 0.94.

d) 1.27 (1.06, 1.52)
e) This cannot be answered without being given the value of $\hat{\lambda}_o$[t=24 months].

17. Based on the results for **Models W- Y**, which of the following statements is true? (*Circle only one response*).
a) The proportional hazards assumption with regard to the treatment groups is violated.
b) The relationship between time to relapse and treatment group is modified by age at enrollment.
c) The relationship between time to relapse and treatment group is substantially confounded by at least one of the following: IV drug use, age, and race.
d) The relationship between time to relapse and treatment group is not confounded by IV drug use, age, and race.

The unadjusted (Model W) and adjusted (Models X and Y) slopes for treatment, i.e. ln(hazard ratios) for treatment (versus control), and hence the ln hazard ratios are similar (less than 15% difference) across the models indicating that the relationship between (time to) relapse and treatment is not confounded by other subject characteristics. Given that subjects were randomized to the treatment and control groups, this is not surprising.

e) IV drug use is not a statistically significant predictor of time to relapse after accounting for treatment group.

18. Based on the result from **Model Y**, does the relationship between the hazard of relapse and age at enrollment appear to be linear on the log scale (after adjusting for treatment group, IV drug use and race)? (*Circle only one response*).

a) No, because the AIC value for Model Y is smaller than the AIC values for Models W and X.
b) This cannot be answered without seeing the results of a Cox regression that includes age as a continuous predictor (as well as treatment group, IV drug use, and race as predictors)
c) This cannot be answered without having the p-value from a Likelihood ratio test comparing model Y to model X.
d) No, because the differences in the adjusted ln(hazard) are not similar in value for each consecutive pair of age categories (2 vs 1, 3 vs 2, etc.).

As the age categories are equal in width (5 years), if the relationship between the ln(hazard) and age were linear, then the difference in the ln(hazard) between any 2 consecutive categories would be similar in value across the 6 age categories.

e) Yes, because some of the age category coefficients are statistically significant.

19. Which of the following is true based on the results from **Model Z**? (*Circle only one response*).

a) Long-term treatment is more effective that short-term treatment in reducing the hazard of relapse, but only for white subjects (after adjusting for IV drug use and age at enrollment).

Mode Z includes an interaction term between treatment and race. In order to assess the differential in the relapse/treatment association for black and white subjects, write out the model results for each race, adjusted for age:

Black (white = 0) :

$\ln(\hat{\lambda}[t]) = \ln(\hat{\lambda}_0[t]) + b_1(treatment) + (\text{adjustment for age category})$

So the slope for treatment, i.e: the ln(hazard ratio) comparing black subjects in long-term to black subject in short-term is b_1.
b_1 = -0.03, and is not statistically significant.

White (white =1):

$\ln(\hat{\lambda}[t]) = \ln(\hat{\lambda}_0[t]) + b_1(treatment) + b_2 + b_3(1 * treatment) + (\text{adjustment for age category})$

So the slope for treatment, ie: the ln(hazard ratio) comparing white subjects in long-term to white subjects in short-term is b_1+b_3. This sum is statistically significant, and negative: when exponentiated the estimated hazard ratio and the confidence interval endpoints are all less than 1.

```
. lincom treat+ white_treat

 ( 1)  treat + white_treat = 0

------------------------------------------------------------------------------
   _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------+-----------------------------------------------------------------
  (1) |  -.2800709   .1058858    -2.65   0.008    -.4876031   -.0725386
------------------------------------------------------------------------------
```

Notice that the results with the inclusion of this interaction term are still interpretable in terms of proportional hazards. The hazard ratio of relapse for blacks in long-term treatment compared to blacks in short-term treatment is $e^{-0.03}$ = 0.97 at any time in the following period. The same hazard ratio for whites is $e^{-0.28}$ = 0.76, at any time in the following period. The hazard ratios for both blacks anad whites are constant across the entire follow-up period.

b) Long-term treatment is more effective that short-term treatment in reducing the hazard of relapse, but only for non-white subjects (after adjusting for IV drug use and age at enrollment).
c) The assumption of proportional hazards is violated because the interaction term (white_treat) is statistically significant.
d) There is no difference in the hazards of relapse between the long-term and short term treatment programs after adjusting for race, IV drug use and age.
e) The relationship between time-to-relapse and race is modified by IV drug use.

20. Based on the results from **Model Z**, which of the following is the log hazard ratio of relapse at 24 months after randomization for 23- year old white subjects in long term treatment who used IV drugs versus (minus) 42- year old non-white subjects in short-term treatment who used IV drugs? (*Circle only one response*).

a) $\ln(\hat{\lambda}_o[24]) - 0.03 + 0.35 - 0.25 + 0.47$

b) $\ln(\hat{\lambda}_o[24]) - 0.03 + 0.35 - 0.25 + 0.47 + 0.38$

c) -0.03 + 0.35- 0.25 + 0.47

For a 23- year old white subject in long term treatment who used IV drugs, we can write:

$\ln(\hat{\lambda}[t]) = \mathrm{l}(\hat{\lambda}_0[t]) + b_1 + b_2 + b_3 + b_4 + 0$

For a 42- year old non-white subject in short term treatment who used IV drugs, we can write:

$\ln(\hat{\lambda}[t]) = \mathrm{l}(\hat{\lambda}_0[t]) + b_4 + b_8$

By subtraction, the difference =

$b_1 + b_2 + b_3 - b_8 = -0.03 + 0.35 - 0.25 + 0.47$

d) -0.03 + 0.35- 0.25 + 0.47 +0.38

e) -0.03 + 0.35- 0.25 + 23 - 42

Biostatistics 140.623
Final Exam Formula Sheet

Tabled chi-squared values: (α=0.05)

df=1, χ2= 3.84
df=2, χ2= 5.99
df=3, χ2= 7.81
df=200, χ2= 233.99

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ...\varepsilon$$

$$F_{s,\ n-p-s-1} = \frac{(RSS_{Null} - RSS_{Extended})/s}{RSS_{Extended}/(n-p-s-1)}$$

AIC = RSS + 2(model df)

$$\ln = \log_e$$

$$\ln\left(\frac{a}{b}\right) = \ln(a) - \ln(b)$$

$$\frac{e^{a+b}}{e^{a}} = e^{b}$$

$$\text{log odds} = \text{logit}[\Pr(Y=1)] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ...\beta_s X_s$$

$$\Pr(Y=1) = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ...\beta_s X_s}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ...\beta_s X_s}} = \frac{\text{odds}}{1 + \text{odds}}$$

LRT (Likelihood Ratio Test) = -2 (LL_{Null} - $LL_{Extended}$)
where LL = log likelihood
AIC = - 2 LL + 2(model df)

Poisson Regression (LLR) Model:

$$\log(\mu_i) = \log N_i + \beta_1 X_1 + ... + \beta_p X_p$$

$$\log(\lambda_i) = \beta_1 X_1 + ... + \beta_p X_p$$

Proportional Hazards Model:

$$\log \lambda(t;X) = \log \lambda_0(t:X) + \beta_1 X_1 + ... + \beta_p X_p$$

$$\lambda(t;X) = \lambda_0(t;X) e^{\beta_1 X_1 + ... + \beta_p Xp}$$

$$S(t;X) = [S_0(t)]^{e^{X\beta}}$$

Models A-D concern questions 5-8:

The outcome Y = Slp15 =1 if sleep latency > 15 minutes; = 0 if ≤ 15 minutes

Demographic characteristics are: age in years
female =1 if female; 0 if male **smk** = 0 if never; 1 if current; 2 if former smoker
BMI in kg/m^2
bmicat - BMI category: 1 if < 18.5 kg/m^2; 2 if 18.5 – 24.9 kg/m^2; 3 if 25-29.9 kg/m^2

Model A

```
. logit Slp15 age

Logistic regression                                     Number of obs   =        821
                                                        LR chi2(1)      =       4.27
                                                        Prob > chi2     =     0.0387
Log likelihood = -565.09366                             Pseudo R2       =     0.0038
------------------------------------------------------------------------------
       Slp15 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    .018698   .0090718     2.06   0.039     .0009176    .0364784
       _cons |  -1.272739   .5571782    -2.28   0.022    -2.364788    -.18069
------------------------------------------------------------------------------
. est store A
```

Model B

```
. mkspline age1 55 age2 65 age3 = age, marginal

. logit Slp15 age1 age2 age3

Logistic regression                                     Number of obs   =        821
                                                        LR chi2(3)      =       9.25
                                                        Prob > chi2     =     0.0261
Log likelihood = -562.60413                             Pseudo R2       =     0.0082
------------------------------------------------------------------------------
       Slp15 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        age1 |  -.0128556   .0361538    -0.36   0.722    -.0837158    .0580046
        age2 |   .0819046   .0547896     1.49   0.135    -.0254811    .1892903
        age3 |  -.1132878   .0509072    -2.23   0.026     -.213064   -.0135115
       _cons |   .2584022   1.886719     0.14   0.891    -3.439499    3.956303
------------------------------------------------------------------------------
. est store B

. lincom age1 +age2
 ( 1)  [Slp15]age1 + [Slp15]age2 = 0
------------------------------------------------------------------------------
       Slp15 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |    .069049   .0258665     2.67   0.008     .0183515    .1197465
------------------------------------------------------------------------------

. lincom age1 +age2+ age3
 ( 1)  [Slp15]age1 + [Slp15]age2 + [Slp15]age3 = 0
------------------------------------------------------------------------------
       Slp15 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -.0442388   .0310547    -1.42   0.154    -.1051049    .0166273
------------------------------------------------------------------------------
```


Model C

```
. logit Slp15 age1 age2 age3 female
Logistic regression                                     Number of obs   =        821
                                                        LR chi2(4)      =       9.59
                                                        Prob > chi2     =     0.0479
Log likelihood = -562.43374                             Pseudo R2       =     0.0085
------------------------------------------------------------------------------
       Slp15 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        age1 |  -.0129809   .0361621    -0.36   0.720    -.0838574    .0578956
        age2 |   .0828356   .0548255     1.51   0.131    -.0246204    .1902917
        age3 |  -.1144016    .050957    -2.25   0.025    -.2142754   -.0145278
      female |   .0823713   .1411278     0.58   0.559    -.1942341    .3589768
       _cons |   .2214594   1.888178     0.12   0.907    -3.479301     3.92222
------------------------------------------------------------------------------
. est store C
```

Model D

```
. logit Slp15 age1 age2 age3 female i.smk i.bmicat

Logistic regression                                     Number of obs   =        821
                                                        LR chi2(9)      =      22.84
                                                        Prob > chi2     =     0.0066
Log likelihood =  -555.8081                             Pseudo R2       =     0.0201
------------------------------------------------------------------------------
       Slp15 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        age1 |  -.0289882   .0371392    -0.78   0.435    -.1017796    .0438032
        age2 |   .0963857   .0557898     1.73   0.084    -.0129603    .2057317
        age3 |  -.0963259   .0516255    -1.87   0.062    -.1975101    .0048582
      female |   .0921986   .1426108     0.65   0.518    -.1873135    .3717106
             |
         smk |
     Current |   .2092384   .2036174     1.03   0.304    -.1898445    .6083212
      Former |  -.1891028   .1599506    -1.18   0.237    -.5026002    .1243945
             |
      bmicat |
           2 |   .1040141   .9355023     0.11   0.911    -1.729537    1.937565
           3 |  -.0053195   .9291496    -0.01   0.995    -1.826419     1.81578
           4 |   .4902065   .9266137     0.53   0.597    -1.325923    2.306336
             |
       _cons |   .8475727   2.131638     0.40   0.691    -3.330362    5.025507
------------------------------------------------------------------------------
.est store D

. lrtest D C

Likelihood-ratio test                                 LR chi2(5)  =     13.25
(Assumption: C nested in D)                           Prob > chi2 =    0.0211

. est stats *
Akaike's information criterion and Bayesian information criterion
-----------------------------------------------------------------------------
       Model |    Obs    ll(null)   ll(model)     df          AIC         BIC
-------------+---------------------------------------------------------------
           A |    821   -567.2302   -565.0937      2     1134.187    1143.608
           B |    821   -567.2302   -562.6041      4     1133.208     1152.05
           C |    821   -567.2302   -562.4337      5     1134.867     1158.42
           D |    821   -567.2302   -555.8081     10     1131.616    1178.721
-----------------------------------------------------------------------------
```


Models 1-3 concerns questions 9 -11. **Variables** are:
dead (number of deaths) in each age group;
population (number at risk) in each age group.
Age in age categories: < 45, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 85-80, > 80 years
Smoking category (**smoke**):

1-smokes **neither** cigarettes nor cigars/pipes
2- smokes cigars/pipes **only**
3- smokes **both** cigarettes and cigars/pipes
4- smokes cigarettes **only**

Model 1

```
. poisson dead i.age, exposure(pop)
```

$$\log(\lambda_j) = \beta_0 + \beta_1 age_2 + \beta_2 age_3 + \beta_3 age_4 + \beta_4 age_5 + \beta_5 age_6 + \beta_6 age_7 + \beta_7 age_8 + \beta_8 age_9$$

```
Poisson regression                                Number of obs   =         36
                                                  LR chi2(8)      =    3864.26
                                                  Prob > chi2     =     0.0000
Log likelihood =  -215.8728                       Pseudo R2       =     0.8995
------------------------------------------------------------------------------
        dead |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |
       45-49 |   .5560324   .0799878     6.95   0.000     .3992593    .7128056
       50-54 |   .9881493   .0768149    12.86   0.000     .8375948    1.138704
       55-59 |   1.371452   .0652555    21.02   0.000     1.243553     1.49935
       60-64 |   1.628995   .0625358    26.05   0.000     1.506427    1.751563
       65-69 |   1.957145   .0626921    31.22   0.000     1.834271    2.080019
       70-74 |   2.205774   .0641042    34.41   0.000     2.080132    2.331416
       75-79 |   2.457785   .0671346    36.61   0.000     2.326204    2.589367
         80+ |   2.687489   .0708023    37.96   0.000     2.548719    2.826259
             |
       _cons |  -3.395722   .0584206   -58.13   0.000    -3.510224   -3.281219
      ln(pop) |          1  (exposure)
------------------------------------------------------------------------------
```

Model 2

```
. poisson dead  i.smoke, exposure(pop)
```

$$\log(\lambda_j) = \beta_0 + \beta_1 smoke_2 + \beta_2 smoke_3 + \beta_3 smoke_4$$

```
Poisson regression                                Number of obs   =         36
                                                  LR chi2(3)      =     145.28
                                                  Prob > chi2     =     0.0000
Log likelihood = -2075.3636                       Pseudo R2       =     0.0338
------------------------------------------------------------------------------
             dead |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------------+----------------------------------------------------------------
            smoke |
          2.smoke |   .3667831   .0466918     7.86   0.000     .2752688    .4582974
          3.smoke |  -.0633457    .038233    -1.66   0.098    -.1382809    .0115896
          4.smoke |    .054597   .0392158     1.39   0.164    -.0222646    .1314587
                  |
            _cons |  -1.839969   .0349215   -52.69   0.000    -1.908414   -1.771524
           ln(pop) |          1  (exposure)
------------------------------------------------------------------------------
```

Model 3

```
. poisson dead i.age i.smoke, exposure(pop)
```

$$\log(\lambda_j) = \beta_0 + \beta_1 age_2 + \beta_2 age_3 + \beta_3 age_4 + \beta_4 age_5 + \beta_5 age_6 + \beta_6 age_7 + \beta_7 age_8 + \beta_8 age_9 +$$

$$+ \beta_9 smoke_2 + \beta_{10} smoke_3 + \beta_{11} smoke_4$$

```
Poisson regression                              Number of obs   =         36
                                                LR chi2(11)     =    4034.50
                                                Prob > chi2     =     0.0000
Log likelihood = -130.75483                     Pseudo R2       =     0.9391

------------------------------------------------------------------------------
        dead |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |
       45-49 |   .5538766   .0799886     6.92   0.000     .3971019    .7106514
       50-54 |   .9803869   .0768183    12.76   0.000     .8298257    1.130948
       55-59 |   1.379458   .0652606    21.14   0.000      1.25155    1.507367
       60-64 |   1.654229   .0625688    26.44   0.000     1.531596    1.776861
       65-69 |   1.998171   .0627875    31.82   0.000      1.87511    2.121232
       70-74 |   2.271406   .0643537    35.30   0.000     2.145275    2.397537
       75-79 |   2.558575   .0677844    37.75   0.000      2.42572     2.69143
         80+ |   2.846925   .0724225    39.31   0.000     2.704979     2.98887
             |
       smoke |
     2.smoke |   .0478065   .0469926     1.02   0.309    -.0442972    .1399103
     3.smoke |   .2179552   .0386942     5.63   0.000      .142116    .2937945
     4.smoke |   .4169596   .0399121    10.45   0.000     .3387333    .4951859
             |
       _cons |  -3.680024   .0682382   -53.93   0.000    -3.813769    -3.54628
    ln(pop) |          1  (exposure)
------------------------------------------------------------------------------

. lincom 4.smoke - 3.smoke, irr
 ( 1)  - [dead]3.smoke + [dead]4.smoke = 0

------------------------------------------------------------------------------
        dead |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   1.220187    .028954     8.39   0.000     1.164738    1.278276
------------------------------------------------------------------------------


. lincom 4.smoke - 2.smoke, irr
 ( 1)  - [dead]2.smoke + [dead]4.smoke = 0

------------------------------------------------------------------------------
        dead |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   1.446509   .0548414     9.74   0.000     1.342918    1.558091
------------------------------------------------------------------------------


. lincom 3.smoke - 2.smoke, irr
 ( 1)  - [dead]2.smoke + [dead]3.smoke = 0

------------------------------------------------------------------------------
        dead |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   1.185481   .0431857     4.67   0.000      1.10379    1.273218
------------------------------------------------------------------------------
```


Models W-Z concern questions 16-20**:**

treat: 1 for long-term, 0 for short-term treatment
age_cat: takes on values 1-6 for 5-year age intervals; the age range in each of the intervals are [20, 25), [25, 30), [30, 35), [35,40), [40, 45) and [45, 50].
white: 1 if subject identifies as white, 0 if non-white.
iv_druguse: 1 if subject was using intravenous (IV) drug at time of enrollment, 0 if not

Model W: ln(hazard of relapse at time t) = $\ln(\lambda_0[t]) + \beta_1 x_1$

```
. stcox treat, nohr
        failure _d:  censor == 1
  analysis time _t:  time
Cox regression -- Breslow method for ties
No. of subjects =          585                     Number of obs   =       585
No. of failures =          471
Time at risk    =       141923
                                                   LR chi2(1)      =      6.86
Log likelihood  =   -2710.1336                     Prob > chi2     =    0.0088
------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |  -.2419827   .0923941    -2.62   0.009    -.4230717   -.0608936
------------------------------------------------------------------------------
. est store W
```

Model X: ln(hazard of relapse at time t) = $\ln(\lambda_0[t]) + \beta_1 x_1 + \beta_2 x_2$

```
. stcox treat iv_druguse, nohr
        failure _d:  censor == 1
  analysis time _t:  time
Cox regression -- Breslow method for ties
No. of subjects =          585                     Number of obs   =       585
No. of failures =          471
Time at risk    =       141923
                                                   LR chi2(2)      =     18.49
Log likelihood  =   -2704.3199                     Prob > chi2     =    0.0001
------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |  -.2302159   .0924643    -2.49   0.013    -.4114426   -.0489893
  iv_druguse |   .3255089   .0967982     3.36   0.001      .135788    .5152299
------------------------------------------------------------------------------
. est store X
```


Model Y:

ln(hazard of relapse at time t) =

$$\ln(\lambda_0[t]) + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8$$

```
. stcox treat white  iv_druguse i.age_cat, nohr

        failure _d:  censor == 1
  analysis time _t:  time

Iteration 0:   log likelihood = -2713.5637
Iteration 1:   log likelihood = -2696.0106
Iteration 2:   log likelihood = -2695.9678
Iteration 3:   log likelihood = -2695.9678
Refining estimates:
Iteration 0:   log likelihood = -2695.9678

Cox regression -- Breslow method for ties

No. of subjects =          585                     Number of obs   =       585
No. of failures =          471
Time at risk    =       141923
                                                   LR chi2(8)      =     35.19
Log likelihood  =   -2695.9678                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |  -.2227576   .0926809    -2.40   0.016    -.4044088   -.0411064
       white |   .2096038    .112931     1.86   0.063    -.0117368    .4309444
  iv_druguse |    .386416   .1053305     3.67   0.000      .179972    .5928599
             |
     age_cat |
      25-29  |  -.0605183   .1692869    -0.36   0.721    -.3923145     .271278
      30-34  |  -.2179493   .1678659    -1.30   0.194    -.5469603    .1110618
      35-39  |  -.1843114   .1770157    -1.04   0.298    -.5312558    .1626329
      40-44  |  -.4944738   .2132287    -2.32   0.020    -.9123943   -.0765533
      45-50  |  -.7889832   .3270934    -2.41   0.016    -1.430074   -.1478919
------------------------------------------------------------------------------

.est store Y
```

Model Z:

ln(hazard of relapse at time t) =

$$\ln(\lambda_0[t]) + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9$$

```
.gen white_treat = white*treat

. stcox treat white  white_treat iv_druguse i.age_cat, nohr

        failure _d:  censor == 1
  analysis time _t:  time

Iteration 0:   log likelihood = -2713.5637
Iteration 1:   log likelihood = -2695.3932
Iteration 2:   log likelihood = -2695.3293
```



```
Iteration 3:   log likelihood = -2695.3293
Refining estimates:
Iteration 0:   log likelihood = -2695.3293

Cox regression -- Breslow method for ties

No. of subjects =          585                     Number of obs   =       585
No. of failures =          471
Time at risk    =       141923
                                                   LR chi2(9)      =     36.47
Log likelihood  =   -2695.3293                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |  -.0282404   .1965929    -0.14   0.886    -.4135554    .3570745
       white |    .347035   .1691883     2.05   0.040      .015432    .6786379
 white_treat |  -.2518304   .2236506    -1.13   0.260    -.6901775    .1865166
  iv_druguse |   .3809726   .1053208     3.62   0.000     .1745477    .5873975
             |
     age_cat |
       25-29 |  -.0466745   .1695572    -0.28   0.783    -.3790005    .2856515
       30-34 |  -.2019194   .1682786    -1.20   0.230    -.5317394    .1279006
       35-39 |  -.1649912   .1775157    -0.93   0.353    -.5129156    .1829332
       40-44 |  -.4664019   .2142489    -2.18   0.029     -.886322   -.0464817
       45-50 |   -.782055   .3270817    -2.39   0.017    -1.423123   -.1409866
------------------------------------------------------------------------------

. lincom treat+ white_treat

 ( 1)  treat + white_treat = 0

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -.2800709   .1058858    -2.65   0.008    -.4876031   -.0725386
------------------------------------------------------------------------------

. est stats *
Akaike's information criterion and Bayesian information criterion

-----------------------------------------------------------------------------
       Model |    Obs    ll(null)   ll(model)     df          AIC         BIC
-------------+---------------------------------------------------------------
           W |    585   -2713.564   -2710.134      1     5422.267    5426.639
           X |    585   -2713.564    -2704.32      2      5412.64    5421.383
           Y |    585   -2713.564   -2695.968      8     5407.936    5442.908
           Z |    585   -2713.564   -2695.329      9     5408.659    5448.003---------
           ---------------------------------------------------------------
               Note:  N=Obs used in calculating BIC; see [R] BIC note
```

