Name_____
I will adhere to the Hopkins code of academic ethics.
Signature_____
Section (please check one):
( ) Diener-West ( ) McGready

Biostatistics 140.623
Second Term, 2014-2015
Quiz 2
March 5, 2015

Below find a set of years until death or censoring for a group of 5 patients who had a surgical intervention and a second group of 5 who received a medical intervention. The plus notation (+) indicates a censored observation.

Surgical group: 1, 1, 3, 7, 10+
Medical group: 1+, 2, 3+, 5, 8

Now **group** the data into 2 time intervals (bins). To aid you in answering the questions below, complete the following table using the survival data above:

Group	**Interval (years)**	**Deaths**	**Person-years**	**Incidence (Death) rate**
Surgical	0 - 2	2	1+1+3(2)=8	
	> 2 -10	2	1+5+8 =14	2/14 = 0.14
Medical	0 - 2			
	> 2 - 10	2	1+3+6 = 10	2/10 = 0.20

1. The **overall incidence (death) rate** in the **surgical group** is: *(Circle only one response)*

   a) 0.045 deaths per person-year
   b) 0.18 deaths per person-year
   c) 0.20 deaths per person-year
   d) 0.33 deaths per person-year
   e) 0.80 deaths per person-year

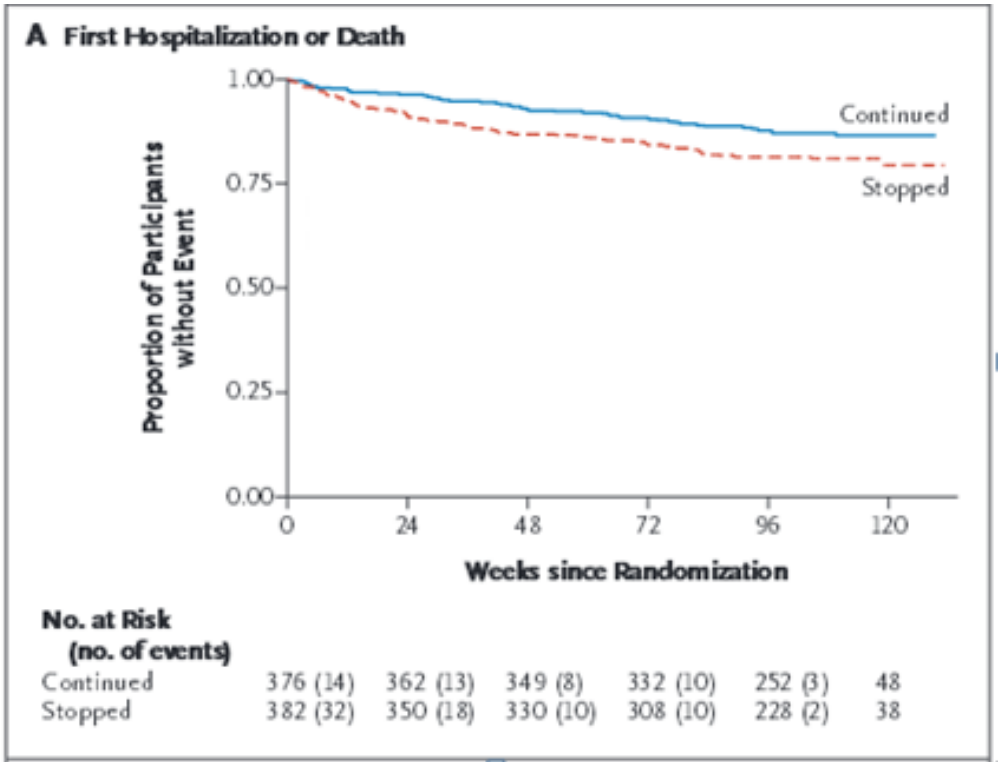
2. The **total person-years** in the **time bin "0- 2 years"** in the **medical group** is: *(Circle only one response)*

   a) 3 person-years
   b) 9 person-years
   c) 0.20 deaths per year
   d) 15 person-years
   e) 19 person-years

Name_____
I will adhere to the Hopkins code of academic ethics.
Signature_____
Lecture Section (please check one):
( ) Diener-West ( ) McGready

Biostatistics 140.623
Third Term, 2013-2014
Quiz 2
March 4, 2014

A recent study[1] compared outcomes between HIV-infected children *randomly assigned* to either **continue** or **stop** prophylaxis therapy after receiving at least 96 weeks of anti-retroviral therapy. Below are the **Kaplan-Meier survival curve estimates** ($\hat{S}(t)$) of time after randomization to *either* first hospitalization or death for the two groups.



1. Approximately what is the estimated proportion of children in the "**Stopped**" group who had been either hospitalized or had died ***within*** 24 weeks after randomization into the study? (*Circle only one response*).

   a) 10%
   b) 90%
   c) 2%
   d) 98%
   e) 25%

[1] Bwakura-Dangarembiz M, et al. A Randomized Trial of Prolonged Co-trimoxazole in HIV-Infected Children in Africa (2014). *New England Journal of Medicine* 370 (1):41-53.

2. The p-value given by the **log rank test** comparing the two survival curves is 0.007. Assuming a significance level (alpha) of 0.05, one can conclude that the observed differences in the Kaplan-Meier curves between the two groups are: (*Circle only one response*).

   a) *Relatively likely* if there is **no difference** in survival over the study follow-up period (at the population level) for the two randomization groups.
   b) *Relatively likely* if there is a scientifically important difference in survival over the study follow-up period (at the population level) for the two randomization groups.
   c) *Relatively unlikely* if there is **no difference** in survival over the study follow-up period (at the population level) for the two randomization groups.
   d) *Relatively unlikely* if there is a scientifically important difference in survival over the study follow-up period (at the population level) for the two randomization groups.
   e) Reflecting a 0.7% chance that the null hypothesis of no survival difference is true.

3. What can be inferred about the **estimated hazard ratio (HR)** of first hospitalization or death for the children randomized to the "**Continued**" group compared to children randomized to the "**Stopped**" group? (*Circle only one response*).
   a) HR =0
   b) HR = 1
   c) HR > 1
   d) HR <1
   e) It is not possible to estimate this from the information given.

4. Three-hundred seventy-six (376) children were randomized to the "**Continued**" group, and 48 (13%) children were still at risk of hospitalization or death at 120 weeks. However, the corresponding Kaplan-Meier curve estimate at 120 weeks for this group is approximately 90%. How can this have happened? (*Circle only one response*).

   a) Some of the observations in the "Continued" group were censored prior to 120 weeks.
   b) The researchers estimated the Kaplan-Meier curve using only the data on patients who were hospitalized or died during the 120 weeks after randomization.
   c) The researchers estimated the Kaplan-Meier curve using only the data on patients who were not censored during the 120 weeks after randomization.
   d) The Kaplan-Meier estimate at 120 weeks is the risk of surviving beyond 120 weeks among only those who were at risk of having the event at 120 weeks.
   e) The hazard of hospitalization or death for the children in the "Continued" group was assumed to be constant over time.

5. How does the Kaplan-Meier approach to estimating the survival function utilize information from censored observations? (*Circle only one response*).
   a) It drops all censored observations from the sample before the curve is estimated.
   b) It uses the censored observations when considering who is "at risk" of an event at each given time in the follow-up period.
   c) It treats the censoring times as event times.
   d) It treats the event times as censoring times.
   e) It assumes that all censored observations have the event by the end of follow-up.

Name_____
I will adhere to the Hopkins code of academic ethics.
Signature_____
Lecture Section (please check one):
( ) Diener-West ( ) McGready

Biostatistics 140.623
Third Term, 2014-2015
Final Examination
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.
    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.

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
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 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.
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.

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.
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 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 compositve of death from cardiac causes, nonfatal myocardial infarction, or refractory angina) and treatment (PCI versus control). The model used is:

$$\textbf{ln(hazard of primary outcome at time } t\textbf{)} = \ln(\lambda_0[t]) + \beta_1 x_1$$

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.
    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.

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).

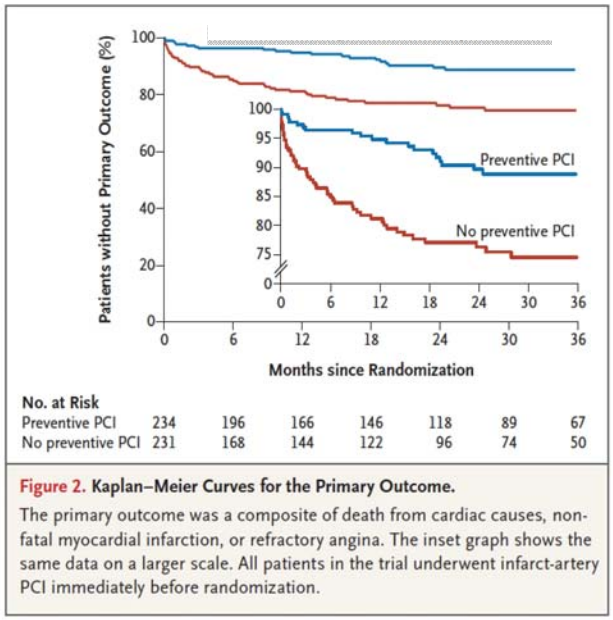


Figure 2. Kaplan–Meier Curves for the Primary Outcome.
The primary outcome was a composite of death from cardiac causes, nonfatal myocardial infarction, or refractory angina. The inset graph shows the same data on a larger scale. All patients in the trial underwent infarct-artery PCI immediately before 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) $b_1 < 0$
    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.
    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)
    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.
    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.).
    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).
    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

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

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

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 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
```


Name_____
I will adhere to the Hopkins code of academic ethics.
Signature_____

Section (please check one):
( ) Diener-West ( ) McGready

Biostatistics 140.623
Third Term, 2015-2016
Final Examination
March 10, 2016

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. Choose one best response for each question.

Questions 1-4 address general knowledge.

Questions 5 – 9 refer to data from the most recent National Hospital Ambulatory Medical Care Survey regarding factors associated with waiting time (in minutes) for nearly 25,000 emergency department visits at U.S. Hospitals in 2011.

The primary outcome of interest is:
waittime = waiting time in minutes

Predictors of interest include:
age = age in years
agecat = four age categories (age quartiles):
1: 0- 20 years
2: > 20- 40 years
3: > 40 – 60 years
4: >= 60 years
white: 1 if white, 0 if non-white
private: 1 if private insurance, 0 if public

A series of linear regression models, **Models A-E** are given at the back of the exam.

5. Based on a comparison of **Models A and B,** which of the following can be concluded? (*Circle only one response*)
   a) Age categories contribute to the prediction of waiting times beyond that contributed by age in years.
   b) The relationship between average waiting times and age is linear because the p-value for the slope of age in Model A is less than 0.05.
   c) Model A is statistically significantly better than Model B because F(1, 24676) is less than F(3, 24674) .
   d) The relationship between average waiting times and age is not linear because the mean waiting time differences for two consecutive age categories is not consistent across age categories.
   e) Age is not a statistically significant predictor of waiting times.

6. Which is the following conclusion can be made from the results for **Model C**? (*Circle only one response*)

   a) The relationship between waiting times and race (white vs. non-white) is confounded by age differences between the race groups.
   b) The relationship between waiting times and race (white vs. non-white) is not confounded by age differences between the race groups.
   c) The relationship between waiting times and race (white vs. non-white) is modified by age.
   d) The relationship between waiting times and race (white vs. non-white) is not modified by age.
   e) White subjects had wait times of 16.4 minutes less that black subjects of comparable age.

7. After fitting **Model C** in Stata, which of the following commands could be used to get an estimated mean waiting time, and 95% confidence interval, for 65 year old white subjects? (*Circle only one response*)

a) lincom agecat4 + white
b) test agecat4 white
c) lincom _cons+ agecat4 + white
d) test _cons agecat4 white
e) No command is necessary. The estimated mean is 64.9+-16.4+-0.07 with 95% confidence interval (62.4+-18.6+-2.9, 67.3+-14.2+2.8)

8. Based on the results of **Model E**, what it the estimated mean difference in waiting time for **white subjects with private insurance** compared to **white subjects with public insurance** of the same age? (*Circle only one response*)

a) -8.8 (-13.0, -4,5)
b) -16.6 (-19.1, -14.0)
c) -13.0 (-17.3, -8.8)
d) -5.3 (-7.7, -2.8)
e) Mean differences cannot be estimated with linear regression.

9. The R^2 value for **Model B** is 0.002. How can this be interpreted in conjunction with the rest of the information given in the output for Models A-E? (*Circle only one response*)

a) Model B predicts waiting times well for observations not used in fitting Model B.
b) Taken together race, and insurance type (private or public) are statistically significant predictors of waiting time after accounting for age.
c) While the mean waiting times are tatistically different across (at least some of) the age categories, there is still substantial variation in the individual waiting times within each age group.
d) Model B does not fit the observed data well.
e) The association between waiting time and age is not modified by other factors.

Questions 10 through 14 refer to a randomized controlled study[1] that was performed to assess the efficacy of financial incentives for quitting smoking. A total of 878 employees (who smoked) of a multinational company based in the United States were randomized to either a:

control group: received information about smoking-cessation programs (442 employees)
or
intervention group: received information about smoking-cessation programs plus financial incentives (436 employees)

The primary outcome for this study was quitting smoking within 12 months after randomization. (***referred to as "quitting smoking")***

The results from a simple logistic regression are as follows:

$$\ln(\text{odds of quitting smoking}) = -2.95 + 1.20x_1,$$

where $x_1 = 1$ for the intervention group, and 0 for the control group. The standard error of the intercept is 0.20, and the standard error of the slope for x_1 is 0.25

10. What is the odds ratio (and 95% CI) of quitting smoking for the intervention group compared to the control group? (*Circle only one response*)

a) -2.95 (-3.34, 2.55)
b) 0.50 (0.20, 0.80)
c) 1.20 (0.70, 1.70)
d) 3.3 (2.0, 5.5)
e) Odds ratios cannot be estimated from logistic regression.

11. What proportion of the persons randomized to the intervention group quit smoking? (i.e: what is the probability that a person randomized to the intervention group quit smoking?) (*Circle only one response*)

a) 5%
b) 15%
c) 77%
d) 20%
e) Logistic regression can only be used to estimate odds and odds ratios.

[1] Volpp K, et al. A Randomized, Controlled Trial of Financial Incentives for Smoking Cessation (2009) New England Journal of Medicine; Vol 360 (7); pps 699-709.

12. How does randomization minimize the potential of the relationship between quitting smoking and the intervention group versus control group being confounded by person's age? (*Circle only one response*)

a) Randomization minimizes the potential for an association between quitting smoking and age.
b) Randomization minimizes the potential for an association between the intervention (versus control) and age.
c) Randomization minimizes the potential for an association between quitting smoking and the intervention (versus control).
d) Randomization minimizes the potential that the association between quitting smoking and the intervention (versus control) differs by age.
e) Randomization minimizes the potential that the relationship between quitting smoking and person's age is statistically significant.

In order to investigate whether the relationship between quitting smoking and the financial incentives intervention is modified by sex, the researchers examined the results from the following logistic regression model:

$$\ln(\text{odds of quitting smoking}) = -3.1 + 1.40x_1 + 0.15x_2 + -0.30x_3$$

where $x_1 = 1$ for intervention group, and 0 for persons randomized to the control group
$x_2 = 1$ for males and 0 for females
$x_3 = x_1 * x_2$ (the interaction term)

13. Based on the estimated regression coefficients given above, what is the odds ratio of quitting smoking for males randomized to the intervention group compared to males randomized to the control group? (*Circle only one response*)

a) 3.0
b) 4.1
c) 0.86
d) 1.10
e) -0.15

14. The **standard error** for the estimated coefficient of the interaction term (x_3) is 0.55. Given this information, what conclusion can be made? (*Circle only one response*)

a) Because the 95% confidence interval for the coefficient (β_3) of the interaction term **does not include 1,** the relationship between quitting smoking and the intervention is (statistically significantly) confounded by sex.
b) Because the 95% confidence interval for the coefficient (β_3) of the interaction term **does not include 1**, the relationship between quitting smoking and the intervention is (statistically significantly) modified by sex.
c) Because the 95% confidence interval for the coefficient (β_3) of the interaction term **includes 0**, the relationship between quitting smoking and the intervention is not (statistically significantly) confounded by sex.
d) Because the 95% confidence interval for the coefficient (β_3) of the interaction term **includes 0**, the relationship between quitting smoking and the intervention is not (statistically significantly) modified by sex.
e) This cannot be answered without the results from a likelihood ratio test.

Questions 15 - 20 refer to an investigation of all-cause mortality (time to death) for a subset of participants enrolled in the longitudinal Framingham Heart Study.

15. **Figure 1** displays the estimated **cumulative survival function (Kaplan-Meier curve)** by baseline diabetes status (diabetes versus no diabetes) by time **in years.** The difference in the estimated survivor function at **10 years** between participants without diabetes versus with diabetes, S(10|without diabetes) – S(10|with diabetes) , is approximately: *(Circle only one response)*

    a) .90 – 0.70
    b) .10 – 0.30
    c) 0.90/0.70
    d) 0.10/0.30
    e) 0.70 – 0.90

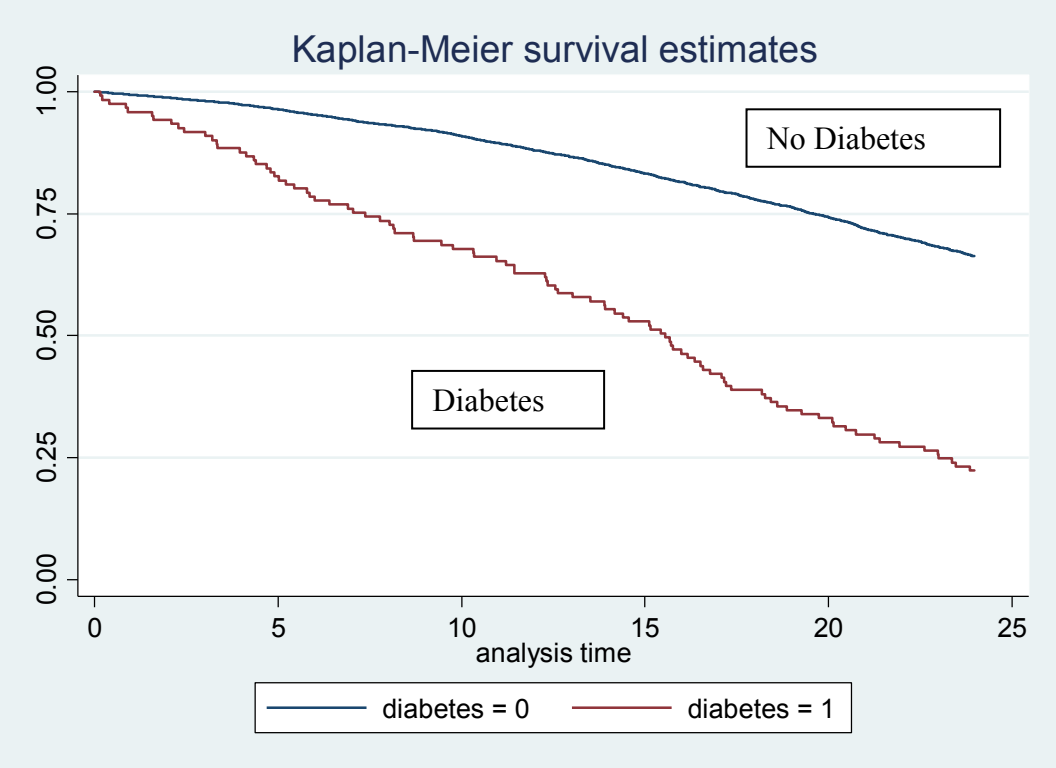


Figure 1

16. The **logrank test statistic** to accompany Figure 1 is calculated as 176.32 with an associated p-value of approximately zero. One could conclude that: *(Circle only one response)*

    a) There is a statistically significant difference in overall survival by diabetes status.
    b) There is no significant difference in overall survival by diabetes status.
    c) There is a statistically significant difference in survival at 10 years by diabetes status.
    d) There is no difference in survival at time zero by diabetes status.
    e) The hazard of death is constant in the two groups.

Table 1 shows the unadjusted and adjusted hazard ratios of all cause-mortality from Cox regression models.

	Unadjusted HR	**95% CI**	**Adjusted* HR**	**95% CI**
Covariate				
Diabetes	3.74	3.03, 4.61	2.50	2.01, 3.09
Sex	0.59	0.54, 0.65	-	-
BMI category				
< 25 kg/m^2	1.00		1.00	-
25- 29	1.32	1.18, 1.47	1.05	0.93, 1.17
> 29	1.73	1.50, 1.99	1.44	1.24, 1.67
Age (years)				
≤ 60	1.08	1.08, 1.09	1.09	1.08,1.10
> 60	1.14	1.11, 1.17	1.14	1.11, 1.17
Current Smoker	1.09	0.99, 1.20	1.50	1.34, 1.67
BP Medication	2.69	2.18, 3.31	-	-
Females				
No BP Meds	-	-	1.0	-
BP Meds	-	-	1.55	1.18, 2.02
Males				
No BP Meds	-	-	1.73	1.55, 1.92
BP Meds	-	-	?	

*Adjusted for all covariates listed plus the **interaction** of sex and BP medication.

17. Using the output from **Model Z** at the back of the exam, what is the **Adjusted HR** for the Males using BP medications compared to Females not using BP medications (fill in the cell with the ? in the last line of Table 1): (*Circle only one response*)

a) 1.64
b) 1.73
c) 3.37
d) 4.38
e) 6.02

18. An assessment of the multivariable-adjusted hazard ratios versus the unadjusted hazard ratios for all-cause mortality provided in **Table 1** and **Model Z** suggests: (*Circle only one response*)

    a) No evidence of confounding of the relationship between age and the log hazard of all-cause-mortality by any of the other multiple adjustment variables.
    b) No evidence of confounding of the relationship between diabetes and the log hazard of all-cause mortality by any of the other multiple adjustment variables.
    c) Possible interaction between diabetes and sex on the log hazard of all-cause mortality, controlling for the other adjustment variables.
    d) No evidence of interaction between sex and the use of blood pressure medications on the log hazard of all-cause mortality.
    e) Evidence of a possible effect modification of the relationship between diabetes and the log hazard of all-cause mortality by time.

19. A test of the null hypothesis of no interaction between sex and use of blood pressure medications in **Model Z** can be performed using: (*Circle only one response*)

    a) An F-test of the null hypothesis that the regression coefficient for the interaction term equals zero.
    b) A Likelihood- Ratio Test of the null hypothesis that the regression coefficient for the interaction term equals zero.
    c) A log-rank test of the null hypothesis that the survival distribution is the same by sex.
    d) A comparison of the AIC values for the multivariable-adjusted versus unadjusted models.
    e) A Z-test resulting from the sum (linear combination) of the regression coefficients for sex, use of blood pressure medication, and the interaction term.

20. A correct interpretation of the **exponentiated interaction coefficient**, $\exp(\beta_9)$ in **Model Z** is: (*Circle only one response*)

    a) The difference, between males and females, in the log hazard ratio for death for those taking BP medication versus not taking BP medication.
    b) The difference, between those taking BP medication and those not taking BP medication, in the log hazard ratio for death for males versus females.
    c) The factor by which the hazard ratio for death for females versus males differs between those taking BP medication and those not taking BP medication.
    d) The factor by which the hazard ratio for death for those taking BP medication versus not taking BP medication differs for males versus females.
    e) The hazard ratio of death for those taking BP medication versus those not taking BP medication, adjusting for sex.

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

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

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

$$AIC = RSS + 2(\text{model df})$$

$$\ln = \log_e$$

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

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

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

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

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:

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

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

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

Questions 5- 9 pertain to **Models A – E** where:

The primary outcome of interest is: **waittime** = waiting time in minutes

Predictors of interest include:
age = age in years
agecat = four age categories (age quartiles):
1: 0- 20 years
2: > 20- 40 years
3: > 40 – 60 years
4: >= 60 years
white: 1 if white, 0 if non-white
private: 1 if private insurance, 0 if public

Model A

```
. regress waittime age

      Source |       SS           df       MS      Number of obs   =    24,678
-------------+----------------------------------   F(1, 24676)     =      9.93
       Model |  63197.0214         1  63197.0214   Prob > F        =    0.0016
    Residual |   157026495    24,676  6363.53118   R-squared       =    0.0004
-------------+----------------------------------   Adj R-squared   =    0.0004
       Total |   157089693    24,677  6365.83428   Root MSE        =    79.772

------------------------------------------------------------------------------
    waittime |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0669103   .0212321    -3.15   0.002    -.1085265    -.025294
       _cons |   59.00096   .9478167    62.25   0.000     57.14318    60.85873
------------------------------------------------------------------------------
```

Model B

```
. regress waittime i.agecat

      Source |       SS           df       MS      Number of obs   =    24,678
-------------+----------------------------------   F(3, 24674)     =     18.10
       Model |  344932.847         3  114977.616   Prob > F        =    0.0000
    Residual |   156744760    24,674  6352.62866   R-squared       =    0.0022
-------------+----------------------------------   Adj R-squared   =    0.0021
       Total |   157089693    24,677  6365.83428   Root MSE        =    79.703

------------------------------------------------------------------------------
    waittime |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      agecat |
          2  |   6.846341   1.450474     4.72   0.000     4.003324    9.689358
          3  |   5.766656   1.432259     4.03   0.000     2.959342    8.573969
          4  |  -1.989739   1.437056    -1.38   0.166    -4.806456    .8269778
             |
       _cons |   53.82912   1.024621    52.54   0.000      51.8208    55.83744
------------------------------------------------------------------------------
```



```
------------------------------------------------------------------------------
```

Model C

```
. regress waittime i.agecat white

      Source |       SS           df       MS      Number of obs   =    24,678
-------------+----------------------------------   F(4, 24673)     =     68.01
       Model |  1713048.79         4  428262.198   Prob > F        =    0.0000
    Residual |   155376644    24,673  6297.43621   R-squared       =    0.0109
-------------+----------------------------------   Adj R-squared   =    0.0107
       Total |   157089693    24,677  6365.83428   Root MSE        =    79.356

------------------------------------------------------------------------------
    waittime |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      agecat |
           2 |   6.876619   1.444161     4.76   0.000     4.045976    9.707261
           3 |   5.965783   1.426088     4.18   0.000     3.170566    8.761001
           4 |  -.0672839   1.436733    -0.05   0.963    -2.883367    2.748799
             |
       white |  -16.41755   1.113856   -14.74   0.000    -18.60078   -14.23433
       _cons |   64.85284   1.264948    51.27   0.000     62.37346    67.33221
------------------------------------------------------------------------------
```

Model D

```
. regress waittime i.agecat white private

      Source |       SS           df       MS      Number of obs   =    24,678
-------------+----------------------------------   F(5, 24672)     =     60.72
       Model |   1909591.9         5   381918.38   Prob > F        =    0.0000
    Residual |   155180101    24,672  6289.72522   R-squared       =    0.0122
-------------+----------------------------------   Adj R-squared   =    0.0120
       Total |   157089693    24,677  6365.83428   Root MSE        =    79.308

------------------------------------------------------------------------------
    waittime |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      agecat |
           2 |   6.658733   1.443803     4.61   0.000     3.828792    9.488674
           3 |   6.125212     1.4255     4.30   0.000     3.331147    8.919277
           4 |  -.9506666   1.444523    -0.66   0.510    -3.782018    1.880685
             |
       white |  -15.60523   1.122618   -13.90   0.000    -17.80563   -13.40483
     private |  -6.138852   1.098181    -5.59   0.000    -8.291353    -3.98635
       _cons |   66.49571   1.297886    51.23   0.000     63.95178    69.03965
------------------------------------------------------------------------------
```


Model E

```
. gen white_private = white*private

. regress waittime i.agecat white private white_private

      Source |       SS           df       MS      Number of obs   =    24,678
-------------+----------------------------------   F(6, 24671)     =     50.93
       Model |  1921835.77         6  320305.962   Prob > F        =    0.0000
    Residual |   155167857    24,671  6289.48388   R-squared       =    0.0122
-------------+----------------------------------   Adj R-squared   =    0.0120
       Total |   157089693    24,677  6365.83428   Root MSE        =    79.306

------------------------------------------------------------------------------
     waittime |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
       agecat |
           2  |   6.682855   1.443879     4.63   0.000     3.852766    9.512945
           3  |   6.148449   1.425569     4.31   0.000     3.354247    8.942651
           4  |  -.8586839   1.445999    -0.59   0.553    -3.692929    1.975561
              |
        white |   -16.5572   1.313675   -12.60   0.000    -19.13208   -13.98231
      private |  -8.765908   2.179704    -4.02   0.000    -13.03826   -4.493558
white_private |   3.511582   2.516813     1.40   0.163    -1.421522    8.444686
        _cons |   67.09411   1.366888    49.09   0.000     64.41493    69.77329
------------------------------------------------------------------------------


. lincom white + white_private
 ( 1)  white + white_private = 0

------------------------------------------------------------------------------
     waittime |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
          (1) |  -13.04562   2.150743    -6.07   0.000     -17.2612   -8.830029
------------------------------------------------------------------------------

. test white white_private
 ( 1)  white = 0
 ( 2)  white_private = 0

       F(  2, 24671) =   97.59
            Prob > F =    0.0000

. lincom private+ white_private
 ( 1)  private + white_private = 0

------------------------------------------------------------------------------
     waittime |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
          (1) |  -5.254326   1.268012    -4.14   0.000    -7.739706   -2.768945
------------------------------------------------------------------------------

. test private white_private
 ( 1)  private = 0
 ( 2)  white_private = 0

       F(  2, 24671) =   16.60
            Prob > F =    0.0000
```


Questions 17- 20 pertain to investigating the log hazard of death and the following covariates :

Model Z $\log[\lambda(t;X)] = \log[\lambda_0(t;X)] + \beta_1 sex + \beta_2 bmicat2 + \beta_3 bmicat3 + \beta_4 age1 + \beta_5 age2 + \beta_6 diabetes + \beta_7 cursmoke + \beta_8 bpmeds + \beta_9 bpmeds*sex$

sex= 0 if female ; 1 if male
bmicat= 1if < 25 ; 2 if 25-29 ; 3 if > 29 kg/m2
age1 = age-60**,**
and **age2**=0 if age ≤ 60 years or **age2** = (age-60) if age > 60 years.
diabetes=0 if no ; 1 if yes
cursmoke=0 if no ; 1 if yes
bpmeds= 0 if no ; 1 if yes

```
.stcox sex i.bmicat age1 age2 diabetes cursmoke bpmeds bpmeds_sex

     Cox regression -- Breslow method for ties

No. of subjects =        4,373                 Number of obs    =       4,373
No. of failures =        1,518
Time at risk    =     32858668
                                               LR chi2(9)       =     1156.18
Log likelihood  =   -11847.173                 Prob > chi2      =      0.0000
------------------------------------------------------------------------------
         _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------+-----------------------------------------------------------------
         sex |   1.730151    .096061     9.87   0.000     1.551758    1.929053
             |
      bmicat |
          2  |   1.045331   .0607683     0.76   0.446     .9327622    1.171485
          3  |    1.43955   .1091144     4.81   0.000     1.240817    1.670113
             |
        age1 |    1.09016   .0047743    19.71   0.000     1.080842    1.099558
        age2 |   1.042639   .0173954     2.50   0.012     1.009096    1.077297
    diabetes |   2.495091   .2732532     8.35   0.000     2.013103     3.09248
    cursmoke |   1.496908   .0823015     7.34   0.000     1.343987    1.667228
      bpmeds |   1.546226   .2130261     3.16   0.002     1.180325    2.025557
  bpmeds_sex |   1.636712   .3607016     2.24   0.025     1.062632    2.520935
------------------------------------------------------------------------------

. lincom age1+age2, hr
 ( 1)  age1 + age2 = 0

------------------------------------------------------------------------------
         _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------+-----------------------------------------------------------------
        (1) |   1.136643   .0160758     9.06   0.000     1.105568    1.168592
------------------------------------------------------------------------------

. lincom sex +bpmeds +bpmeds_sex, hr
 ( 1)  sex + bpmeds + bpmeds_sex = 0

------------------------------------------------------------------------------
         _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------+-----------------------------------------------------------------
        (1) |   4.378541     .76663     8.43   0.000     3.106663    6.171128
------------------------------------------------------------------------------
```


Name_____
I will adhere to the Hopkins code of academic ethics.
Signature_____
Lecturer (please check one):
( ) Diener-West ( ) McGready

Biostatistics 140.623
Third Term, 2016-2017
Final Examination
March 16, 2017

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.

Questions 1 through 5 concern general knowledge.

1. Suppose that a Likelihood Ratio Test (LRT) was performed for a comparison on an extended model with p+s covariates versus a null model with p covariates. If the observed p-value for the LRT is p=0.22, one would conclude that: (*Circle only one response*)

   a) Taken together, the p covariates X_1, ….,X_p do not contribute to the model.
   b) Taken together, the s covariates X_{p+1}, ….,X_{p+s} do not contribute to the model.
   c) Taken together, the p+s covariates X_1, ….,X_{p+s} do not contribute to the model.
   d) None of the individual s covariates X_{p+1}, ….,X_{p+s} are statistically significantly associated with the outcome.
   e) None of the p+s covariates X_1, ….,X_{p+s} are statistically significantly associated with the outcome.

2. Consider the log odds [obesity] = $\beta_0+\beta_1$age + $\beta_2(\text{age-40})^+ + \beta_3(\text{age-65})^+ + \beta_4$ exercise
   + $\beta_5$age*exercise +$\beta_6(\text{age-40})^+$*exercise+ β_7 $(\text{age-65})^+$*exercise

   where $(\text{age-40})^+ = 0$ if age ≤ 40; or = (age-40) if age > 40
   where $(\text{age-65})^+ = 0$ if age ≤ 65; or = (age-65) if age > 65
   and exercise=1 for daily; 0 for not daily

   In this model, the **log odds ratio** for obesity per unit increase in age among those who are over 65 years of age and exercise daily is: (*Circle only one response*)

   a) $\beta_0+\beta_1+\beta_2+\beta_3+\beta_4+\beta_5+\beta_6+\beta_7$
   b) $\beta_3+\beta_5+\beta_6+\beta_7$
   c) $\beta_1+\beta_2+\beta_3+\beta_5+\beta_6+\beta_7$
   d) $\beta_3+\beta_4+\beta_7$
   e) $\beta_0+\beta_3+\beta_4+\beta_7$

4. n/a

5. The generic formulation of a multiple regression model including age (x_1=age in years), smoking status (x_2 = 1 if smoker, and 0 if non-smoker) and sex (x_3= 1 if female, 0 if male) is as follows:

$$LHS = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_2 * X_3$$

What comparison is being made by the coefficient, β_2? (*Circle only one response*)

a) Smokers to non-smokers.
b) Smokers to non-smokers of the same sex and age.
c) Smokers to non-smokers of the same age, but only among females.
d) Smokers to non-smokers of the same age, but only among males.
e) The difference, between females and males, between smokers to non-smokers.

Questions 6 through 9 refer to the results of a survey of married women aged 18 to 49 years who were interviewed at an outpatient service in a hospital in Uganda. Of interest was the association between HIV status and desire for future children. Women also reported their partner's desire for future children. (*Am J Public Health 2013; 103: 278-285*).

Table 1 shows the tabulation of desire for future children in 784 married partners by woman's HIV status:

Future Children	HIV status of Woman HIV+	HIV-	Total
Both want	120	275	395
	34.29	63.36	50.38
Both do not want	162	111	273
	46.29	25.58	34.82
Only man wants	49	25	74
	14.00	5.76	9.44
Only woman wants	19	23	42
	5.43	5.30	5.36
Total	350	434	784
	100.00	100.00	100.00

Pearson chi2(3) = 70.3223 Pr = 0.000

6. Suppose that a model was specified in the following way to investigate this association:

$$\textbf{logit}(P(Y=1)) = \beta_0 + \beta_1 X$$

where Y= 1 if **both** partners want future children; 0 otherwise
and X = 1 if HIV+; 0 if HIV-

From the data in **Table 1,** what is the estimate of the regression coefficient, β_1 ? (*Circle only one response*)

a) [(120/230)/(275/159)]
b) $\log_e$[(120/230)/(275/159)]
c) [(120/350)/(275/434)]
d) $\log_e$(275/159)
e) [(275/159)/(120/230)]
f) $\log_e$[(275/159)/(120/230)]

The authors used **logistic regression models** to investigate the association between a woman's desire for future children and her HIV status, along with other characteristics of the individual.

The outcome was defined as: **desire for future children**; 1 for yes, 0 for no.
The covariates are defined in Table 2. In addition, the unadjusted and adjusted results are presented in **Table 2**: (95% CIs are given in parentheses next to each estimated odds ratio).

Table 2

Characteristics	Unadjusted OR (95% CI)	Adjusted* OR (95% CI)
HIV status		
HIV -	1.000	1.000
HIV +	0.295 (0.228, 0.382)	0.461 (0.326, 0.653)
Age, years		
≤ 24	1.000	1.000
25-29	0.559 (0.394, 0.793)	1.108 (0.772, 1.700)
30-34	0.287 (0.199, 0.414)	0.979 (0.609, 1.573)
35-39	0.214 (0.046, 0.204)	1.193 (0.622, 2.289)
≥ 40	0.071 (0.040, 0.125)	0.346 (0.160, 0.755)
Educational attainment		
Any primary	1.000	1.000
≥ Secondary	1.892 (1.453, 2.465)	1.004 (0.709, 1.420)
Parity	0.473 (0.425, 0.526)	0.505 (0.439, 0.581)
Foster child < 18 years		
0	1.000	1.000
≥1	0.525 (0.399, 0.691)	0.638 (0.450, 0.904)
Household income, UGX		
0 – 50,000	1.000	1.000
50,001-150,000	1.430 (1.039, 1.969)	1.306 (0.872, 1.958)
≥ 150,001	1.990 (1.461, 2.711)	2.006 (1.325, 3.036)
HIV + child in household		
No	1.000	1.000
Yes	0.320 (0.190, 0.539)	0.740 (0.387, 1.415)
Current marriage		
First marriage	1.000	1.000
Second marriage	0.602 (0.444,0.818)	0.805 (0.664,1.467)

***Adjusted for all variables in Table 2**

7. Based on the results in **Table 2**, what is the estimated **adjusted odds ratio** of desiring future children between HIV+ women aged 30-34 and HIV- women aged 25-29 (i.e. HIV+ women aged 30-34 as compared to HIV- women aged 25-29) who are otherwise similar with respect to the other characteristics? (*Circle only one response*)

a) 0.15
b) 0.41
c) 0.88
d) 1.51
e) 2.46

8. What **assumption** is made about the relationship between parity (number of times the woman previously has given birth) and the desire for future children in the logistic regression models in **Table 2**? (*Circle only one response*)

a) The odds ratio of the desire for future children is multiplied by the number of previous births.
b) The change in the log odds of desire for future children for each additional previous birth is constant.
c) The odds of desire for future children is not associated with number of previous births.
d) The variability in the odds of desire for future children is constant.
e) Parity is independent of the desire for future children.

9. What **conclusion** can be made about the adjusted relationships between each covariate and the desire for future children in the logistic regression models in **Table 2**? (*Circle only one response*)

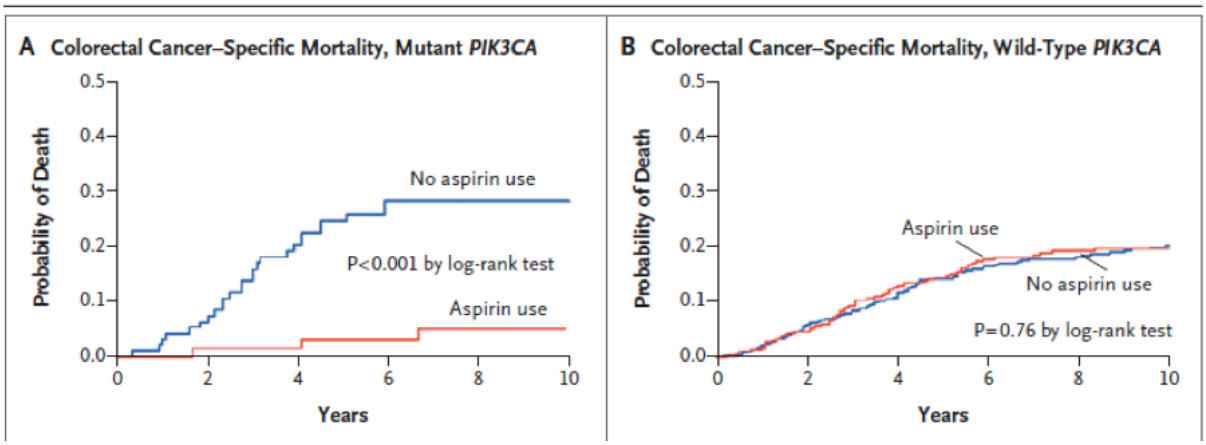
a) The adjusted odds of the desire for future children significantly **decreased** in women with positive HIV status, older age, higher parity, one or more foster child, HIV+ child in the household, and/or second marriage and significantly **increased** in women with secondary education and/or from households with > 50,000 UGX income.
b) The adjusted odds of the desire for future children significantly **decreased** in women with positive HIV status, decreased in women having higher parity, and/or more than one foster child and significantly **increased** in women from households with > 150,000 UGX income.
c) There is no confounding by any of the other covariates of the relationship between a woman's HIV status and her desire for future children
d) There is no interaction between HIV status and any of the other covariates on a woman's desire for future children.
e) There is no statistically significant association between HIV status and desire for future children, after adjusting for all other covariates.

Questions 15 through 20 focus on the results of a study examining aspirin and survival among patients with colorectal cancer. (*NEJM* 367 (17): 1596-1606, 2012. As per the authors:

"We obtained data on 964 patients with rectal or colon cancer from the Nurses' Health Study and the Health Professionals Follow-up Study, including data on aspirin use after diagnosis and the presence or absence of PIK3CA mutation."

"We used data from two prospective cohort studies, the Nurses' Health Study (NHS, involving 121,700 women who were enrolled in 1976) and the Health Professionals Follow-up Study (HPFS, involving 51,500 men who were enrolled in 1986)."

The authors present the following **Figure A** and **Figure B** as part of the article. The starting point (time 0) for each patient was the year of colorectal cancer diagnosis. Each is a Kaplan-Meier estimate show the **proportion who had died** by the given follow-up time (i**.e. 1- S(t))**



15. What is the null hypothesis for the **log-rank test** with p-value <0.001 presented in Figure A? (*Circle only one response*)

    a) The population level Survival Curves (and, hence, hazards of death over-time) are different for the aspirin and no-aspirin groups.
    b) The population level Survival Curves (and, hence, hazards of death over-time) are the same for the aspirin and no-aspirin groups.
    c) The hazard of death is constant over time in both the aspirin and non-aspirin groups.
    d) The hazard of death is not constant over time in both the aspirin and non-aspirin groups.
    e) The ratio of the hazard of death in the aspirin group versus the non-aspirin group changes over time.

16. Suppose the researchers had fit the following Cox regression model, using only the data on subjects with the PIK3CA mutation (the data used to create the curves in **Figure A**):

    $\log(\lambda(t,x_1)) = \log(\lambda_o(t)) + \beta_1 x_1$,
    where $x_1 = 1$ for the aspirin group, 0 for the non-aspirin group

    What can be said about b_1, the estimate of β_1 for this model? (*Circle only one response*)

    a) $b_1=0$
    b) $b_1>0$
    c) $b_1<0$
    d) $b_1>1$
    e) $b_1<1$

17. The validity of the **proportional hazards assumption** for the model depicted in **question 16** can be confirmed when: (*Circle only one response*)

    a) Observing that the plot of the log(-log S(t)) versus log t results in approximately parallel straight lines for the aspirin and non-aspirin groups.
    b) Observing that the plot of the log(-log S(t)) versus log t results in diverging straight lines for the aspirin and non-aspirin groups.
    c) The p-value for the log-rank statistic is less than 0.05.
    d) The AIC achieves the minimum value.
    e) There is a statistically significant interaction between aspirin status (aspirin versus non-aspirin) and time.

18. Suppose the researchers had fit the following Cox regression model, using only the data on subjects with the PIK3CA mutation (the data used to create the curves in **Figure A**):

    $\log(\lambda(t,x_1)) = \log(\lambda_o(t)) + \beta_1 x_1$,
    where $x_1 = 1$ for the aspirin group, 0 for the non-aspirin group

    What does the function $\lambda_o(t)$ quantify? (*Circle only one response*)

    a) The hazard of death for the aspirin group at t=0.
    b) The hazard of death for the aspirin group as a function of time over the follow-up period.
    c) The hazard of death for the non-aspirin group at t=0.
    d) The hazard of death for the non-aspirin group as a function of time over the follow-up period.
    e) The hazard ratio of death the aspirin group compared to the non-aspirin group.

19. What assumption did the researchers have to make when fitting the model from **question 18**? (*Circle only one response*)

    a) The hazard of death is constant over time in both the aspirin and non-aspirin groups.
    b) The hazard of death is constant over time in only the non-aspirin group.
    c) The difference in log(hazard of death) between the aspirin and non-aspirin groups is constant over time.
    d) The difference in hazard of death between the aspirin and non-aspirin groups is constant over time.
    e) The log(hazard) of death is a linear function of time in both the aspirin and non-aspirin groups.

20. Which of the following Cox regression models, *based on data from all subjects* (with or without the PIK3CA mutation) would correspond to the results presented in **Figure A and Figure B**? (*Circle only one response*)

    For each of the following models, the variables are defined as:
    $x_1 = 1$ for the aspirin group, 0 for the non-aspirin group
    $x_2 = 1$ for those with the PIK3CA mutation , 0 for those with the Wild-Type PIK3CA (no mutation)
    t= follow-up time

    a) $\log(\lambda(t,x_1)) = \log(\lambda_o(t)) + \beta_1 x_1$
    b) $\log(\lambda(t,x_1, x_2)) = \log(\lambda_o(t)) + \beta_1 x_2$
    c) $\log(\lambda(t,x_1\ x_2,)) = \log(\lambda_o(t)) + \beta_1 x_1 + \beta_2 x_2$
    d) $\log(\lambda(t,x_1\ x_2)) = \log(\lambda_o(t)) + \beta_1 x_1 + \beta_2 x_2 + \beta_3 t + \beta_4 (x_1 * t)$
    e) $\log(\lambda(t,x_1,x_2,x_3)) = \log(\lambda_o(t)) + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 * x_2)$

Biostat 623 Midterm Exam Formula Sheet

One Sample

$$H_0 : \mu = \mu_0 \quad z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

$$n = \frac{(z_{\alpha/2} + z_\beta)^2 \sigma^2}{\Delta^2}$$

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

$$H_0 : p = p_0 \quad z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$$

$$\boldsymbol{n = \left[\frac{z_{\alpha/2}\sqrt{p_0 q_0} + z_\beta \sqrt{p_a q_a}}{\Delta} \right]^2}$$

Two Samples

$$\mathbf{H_0 : \mu_1 - \mu_2 = \mu_0} \quad \boldsymbol{z = \frac{(\bar{x}_1 - \bar{x}_2) - \mu_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}}$$

$$n = \frac{\left(z_{\alpha/2} + z_\beta \right)^2 \left(\sigma_1^2 + \sigma_2^2 \right)}{\Delta^2}$$

$$\boldsymbol{t = \frac{(\bar{x}_1 - \bar{x}_2) - \mu_0}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}}$$

$$\mathbf{where}\ s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - \mu_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$H_0 : \mu_d = \mu_{d_0} \qquad t = \frac{\bar{d} - \mu_{d_0}}{s_d/\sqrt{n}}$$

$$H_0 : p_1 - p_2 = 0 \qquad z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\frac{\bar{p}\bar{q}}{n_1} + \frac{\bar{p}\bar{q}}{n_2}}}$$

$$n = \frac{\left[z_{\alpha/2}\sqrt{2\bar{p}\bar{q}} + z_\beta \sqrt{p_1 q_1 + p_2 q_2} \right]^2}{\Delta^2}$$

	Tail Probability	
Z	1-sided	2-sided
0.65	0.26	0.52
0.75	0.23	0.46
0.84	0.20	0.40
1.28	0.10	0.20
1.645	0.05	0.10
1.96	0.025	0.05 (same as χ^2=3.84)
2.58	0.005	0.010

Linear Regression Model: $E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ...\beta_s X_s$

$\ln = \log = \log_e$

$\ln\left(\frac{a}{b}\right) = \ln(a) - \ln(b)$ $\qquad \frac{e^{a+b}}{e^a} = e^b$

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

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

Proportional Hazards Model:

$\log h(t;X) = \log h_0(t:X) + \beta_1 X_1 + ... + \beta_p X_p$

$h(t;X) = h_0(t;X) e^{\beta_1 X_1 + ... + \beta_p Xp}$

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

LRT (Likelihood Ratio Test) $= -2\,(LL_{Null} - LL_{Extended})$

where LL = log likelihood

AIC = - 2 LL + 2(model df)

