**Biost 518: Applied Biostatistics II**

**Biost 515: Biostatistics II**

Emerson, Winter 2014

**Homework #4**

January 27, 2014

**Written problems:** To be submitted as a MS-Word compatible file to the class Catalyst dropbox by 9:30 am on Monday, February 3, 2014. See the instructions for peer grading of the homework that are posted on the web pages.

*On this (as all homeworks) Stata / R code and unedited Stata / R output is* ***TOTALLY*** *unacceptable. Instead, prepare a table of statistics gleaned from the Stata output. The table should be appropriate for inclusion in a scientific report, with all statistics rounded to a reasonable number of significant digits. (I am interested in how statistics are used to answer the scientific question.)*

***Unless explicitly told otherwise in the statement of the problem, in all problems requesting “statistical analyses” (either descriptive or inferential), you should present both***

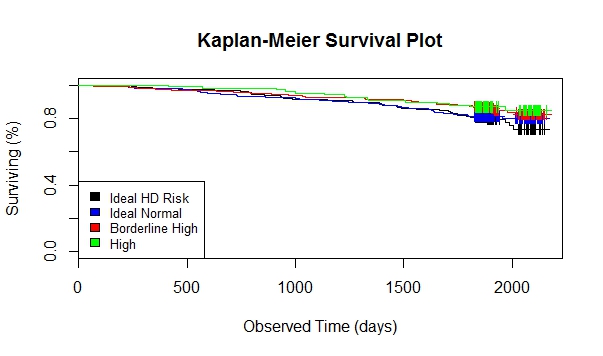
* ***Methods: A brief sentence or paragraph describing the statistical methods you used. This should be using wording suitable for a scientific journal, though it might be a little more detailed. A reader should be able to reproduce your analysis. DO NOT PROVIDE Stata OR R CODE.***
* ***Inference: A paragraph providing full statistical inference in answer to the question. Please see the supplementary document relating to “Reporting Associations” for details.***

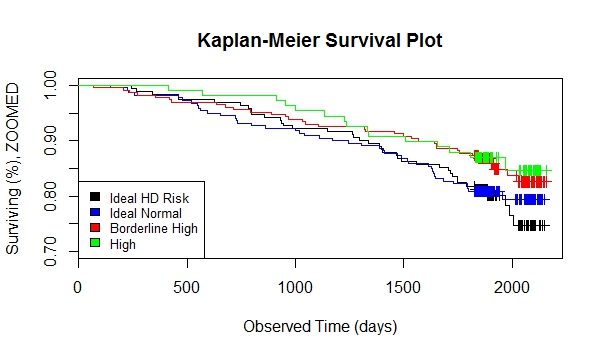
This homework builds on the analyses performed in homeworks #1, #2, and #3. As such, all questions relate to associations among death from any cause, serum low density lipoprotein (LDL) levels, age, and sex in a population of generally healthy elderly subjects in four U.S. communities. This homework uses the subset of information that was collected to examine MRI changes in the brain. The data can be found on the class web page (follow the link to Datasets) in the file labeled mri.txt. Documentation is in the file mri.pdf. See homework #1 for additional information.

1. Perform a statistical regression analysis evaluating an association between serum LDL and all-cause mortality by comparing the instantaneous risk (hazard) of death over the entire period of observation across groups defined by serum LDL modeled as a continuous variable.
   1. Include full description of your methods, appropriate descriptive statistics, and full report of your inferential statistics.
      1. **Methods**: Descriptive statistics are presented within groups defined by serum LDL measurements (less than or equal to 99 mg/dL, between 100 and 129 mg/dL inclusive, between 130 and 159 mg/dL inclusive, and greater than or equal to 160 mg/dL), as well as in the entire sample. The cutoffs are scientifically interpreted as ideal for those at risk of heart disease (HD Risk), ideal for normal population, borderline high LDL, and high LDL, respectively. The categorization of LDL will help if features such as effect modification and linearity wish to be diagnosed or determined. Within each group defined by serum LDL level, for continuous variables (age, weight, pack years of smoking) we include the mean and standard deviation. For binary variables (sex and death) we present percentages. A Kaplan-Meier plot is presented to assess the survival curves of the subjects per category of LDL defined above. Furthermore, the same plot is used to consider the linearity of the LDL variable. Note that two Kaplan-Meier plots are given: the first uses a proper survival scale ranging from 0 to 100%. The second Kaplan-Meier plot is the same plot, but zoomed in so that we can see more details of the survival distributions. The instantaneous hazard of death over the entire period of observation across groups defined by continuous serum LDL will be modeled using a Cox proportional hazards regression model. The hazard ratio and 95% CI was computed using Cox proportional hazards regression with the Huber-White sandwich estimator of the (robust) standard errors. Two-sided p-values based on Wald statistics were interpreted at the 5% significance level.
      2. **Results**: Data is available on 735 subjects, however 10 of those subjects (including 2 who died within 5 years) are missing data. Those subjects are omitted from all analyses. None of the 725 subjects were missing data on any other variables of interest for this analysis. Of the 725 subjects with available measurements, 165 had serum LDL measurements less than or equal to 99 mg/dL, 228 had serum LDL measurements between 100 and 129 mg/dL inclusive, 225 had measurements between 130 mg/dL and 159 mg/dL inclusive, and 107 had measurements greater than or equal to 160 mg/dL. The table below presents descriptive statistics within these groups. Subjects having serum LDL in the two lowest levels of serum LDL were more likely to be male than in other intervals. No consistent trend was seen across groups in age, weight, or smoking history. Subjects with the lowest levels of serum LDL appeared to have a higher mortality rate: about 20% of subjects with LDL less than or equal to 99 mg/dL died within 5 years compared to about 13% in subjects with higher serum LDL at study entry. From the first Kaplan-Meier plot, we observe high levels of survival throughout the studies observation period regardless of LDL level with no censoring before 5 years, followed by high concentrations of censoring. In the zoomed in plot of the same survival curves, we observe a high number of crossovers between LDL subgroups before 1500 days. However, after 1500 days, the survival curves split off into distinct higher LDL and lower LDL levels.
      3. **More Results**: From a Cox proportional hazards regression analysis, we estimate that for each 10 mg/dL increase in serum LDL, the instantaneous risk of death for the higher LDL group is 92.8% the risk of the lower LDL group. Based on a 95% confidence interval, it would not be unusual if the true risk of death for a higher (by 10 mg/dL) LDL group were anywhere between 87.8% and 98.2% of the risk for the lower LDL group. At the 5% significance level, we reject the null hypothesis of no association between survival and LDL (P-value=0.009).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Serum Low Density Lipoprotein (LDL) | | | | | |
|  | <=99 mg/dL (n=165) | 100-129 md/dL (n=228 | 130-159mg/dL (n=225) | >=160 mg/dL (n=107) | Any level (n=725) |
| Male (%) | 57.6 | 53.9 | 43.1 | 42.1 | 49.7 |
| Age (yrs) | 74.8 (5.50) | 74.6 (5.08) | 74.2 (5.62) | 74.9 (5.77) | 74.6 (5.45) |
| Weight (lbs) | 160 (31.5) | 160 (28.8) | 158 (32.3) | 163 (30.7) | 160 (30.8) |
| Smoking (pack-years) | 17.5 (24.04) | 21.5 (28.80) | 20.0 (28.83) | 18.1 (24.41) | 19.3 (27.16) |
| Death within 5 years (%) | 20 | 18.9 | 12.9 | 13.1 | 16.4 |

\*Standard Deviations in ( )





* 1. For each population defined by serum LDL value, compute the hazard ratio relative to a group having serum LDL of 160 mg/dL. (This will be used in problem 4). If *HR* is the hazard ratio (use the actual hazard ratio estimate) obtained from your regression model, this can be effected by the Stata code

gen fithrA = *HR ^ (ldl* – 160)

It could also be computed by creating a centered LDL variable, and then using the Stata predict command

gen cldl = ldl – 160

stcox cldl

predict fithrA

1. Perform a statistical regression analysis evaluating an association between serum LDL and all-cause mortality by comparing the instantaneous risk (hazard) of death over the entire period of observation across groups defined by serum LDL modeled as a continuous logarithmically transformed variable.
   1. Include full description of your methods, appropriate descriptive statistics (you may refer to problem 1, if the descriptive statistics presented there are adequate for this question), and full report of your inferential statistics.
      1. **Methods**: Due to a lack of scientific interpretation, the table described here will contain non-transformed descriptives for LDL. Descriptive statistics are presented within groups defined by serum LDL measurements (less than or equal to 99 mg/dL, between 100 and 129 mg/dL inclusive, between 130 and 159 mg/dL inclusive, and greater than or equal to 160 mg/dL), as well as in the entire sample. The cutoffs are scientifically interpreted as ideal for those at risk of heart disease (HD Risk), ideal for normal population, borderline high LDL, and high LDL, respectively. The categorization of LDL will help if features such as effect modification and linearity wish to be determined. Within each group defined by serum LDL level, for continuous variables (age, weight, pack years of smoking) we include the mean and standard deviation. For binary variables (sex and death) we present percentages. A Kaplan-Meier plot is presented to assess the survival curves of the subjects per category of LDL. Note that a log transform on LDL has no effect on survival curves, since they only depend on time and death. Furthermore, the same plot is used to consider the linearity of the LDL variable. Note that two Kaplan-Meier plots are given: the first uses a proper survival scale ranging from 0 to 100%. The second Kaplan-Meier plot is the same plot, but zoomed in so that we can see more details of the survival distributions. The instantaneous hazard of death over the entire period of observation across groups defined by continuous serum LDL under a log transformation will be modeled using a Cox proportional hazards regression model. The hazard ratio and 95% CI was computed using Cox proportional hazards regression with the Huber-White sandwich estimator of the (robust) standard errors. Two-sided p-values based on Wald statistics were interpreted at the 5% significance level.
      2. **Results**: Data is available on 735 subjects, however 10 of those subjects (including 2 who died within 5 years) are missing data. Those subjects are omitted from all analyses. None of the 725 subjects were missing data on any other variables of interest for this analysis. Of the 725 subjects with available measurements, 165 had serum LDL measurements less than or equal to 99 mg/dL, 228 had serum LDL measurements between 100 and 129 mg/dL inclusive, 225 had measurements between 130 mg/dL and 159 mg/dL inclusive, and 107 had measurements greater than or equal to 160 mg/dL. The table below presents descriptive statistics within these groups. Subjects having serum LDL in the two lowest levels of serum LDL were more likely to be male than in other intervals. No consistent trend was seen across groups in age, weight, or smoking history. Subjects with the lowest levels of serum LDL appeared to have a higher mortality rate: about 20% of subjects with LDL less than or equal to 99 mg/dL died within 5 years compared to about 13% in subjects with higher serum LDL at study entry. From the first Kaplan-Meier plot, we observe high levels of survival throughout the studies observation period regardless of LDL level with no censoring before 5 years, followed by high concentrations of censoring. In the zoomed in plot of the same survival curves, we observe a high number of crossovers between LDL subgroups before 1500 days. However, after 1500 days, the survival curves split off into distinct higher LDL and lower LDL levels.

**More Results**: From a Cox proportional hazards regression analysis, we estimate that for each 50% increase of serum LDL, the instantaneous risk of death for the higher LDL group is 1.77 times higher than the risk of the lower LDL group. Based on a 95% confidence interval, it would not be unusual if the true risk of death for a higher (by 50%) LDL group were anywhere between 1.36 and 2.32 times higher than the risk for the lower LDL group. At the 5% significance level, we reject the null hypothesis of no association between survival and LDL (P-value<0.0001).   
  
5/5 for performing an appropriate analysis

2.5/5 for reporting the association appropriately

Wrong interpretation of point estimate (-1)

Wrong direction(-0.5)

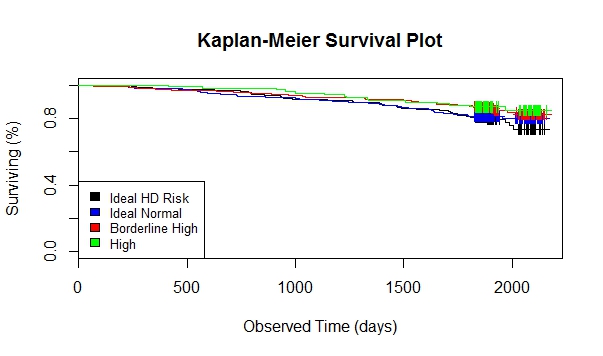
Wrong interpretation of CI (-1)

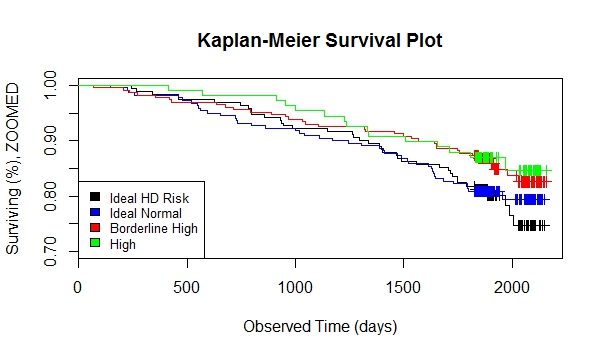
1. Total: 7.5

The HR for 50% increase in LDL is 1.5^(beta1) = 0.7152

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Serum Low Density Lipoprotein (LDL) | | | | | |
|  | <=99 mg/dL (n=165) | 100-129 md/dL (n=228 | 130-159mg/dL (n=225) | >=160 mg/dL (n=107) | Any level (n=725) |
| Male (%) | 57.6 | 53.9 | 43.1 | 42.1 | 49.7 |
| Age (yrs) | 74.8 (5.50) | 74.6 (5.08) | 74.2 (5.62) | 74.9 (5.77) | 74.6 (5.45) |
| Weight (lbs) | 160 (31.5) | 160 (28.8) | 158 (32.3) | 163 (30.7) | 160 (30.8) |
| Smoking (pack-years) | 17.5 (24.04) | 21.5 (28.80) | 20.0 (28.83) | 18.1 (24.41) | 19.3 (27.16) |
| Death within 5 years (%) | 20 | 18.9 | 12.9 | 13.1 | 16.4 |

\*Standard Deviations in ( )





* 1. For each population defined by serum LDL value, compute the hazard ratio relative to a group having serum LDL of 160 mg/dL. (This will be used in problem 4). If *HR* is the hazard ratio (use the actual hazard ratio estimate) obtained from your regression model, this can be effected by the Stata code

gen logldl = log(ldl)

stcox logldl

fithrB = *HR ^ (logldl* – log(160))

It could also be computed by creating a centered logarithmically transformed LDL variable, and then using the Stata predict command

gen clogldl = log(ldl / 160)

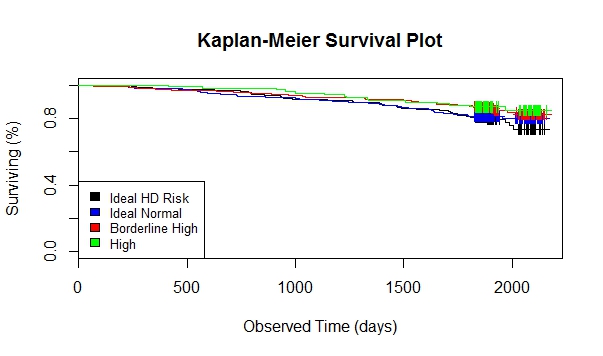
stcox clogldl

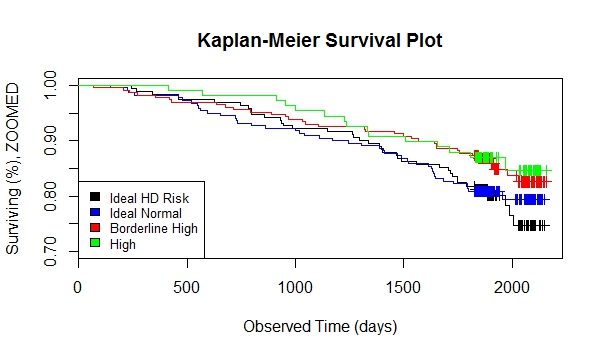
predict fithrB

1. Perform a statistical regression analysis evaluating an association between serum LDL and all-cause mortality by comparing the instantaneous risk (hazard) of death over the entire period of observation across groups defined by serum LDL modeled quadratically (so include both a term for serum LDL modeled continuously and a term for the square of LDL).
   1. Include full description of your methods, appropriate descriptive statistics (you may refer to problem 1, if the descriptive statistics presented there are adequate for this question), and full report of your inferential statistics. In the inferential statistics, include your conclusion regarding the linearity of the association of serum LDL and the log hazard.
      1. **Methods**: Due to a lack of scientific interpretation, the table described here will contain non-transformed descriptives for LDL. Descriptive statistics are presented within groups defined by serum LDL measurements (less than or equal to 99 mg/dL, between 100 and 129 mg/dL inclusive, between 130 and 159 mg/dL inclusive, and greater than or equal to 160 mg/dL), as well as in the entire sample. The cutoffs are scientifically interpreted as ideal for those at risk of heart disease (HD Risk), ideal for normal population, borderline high LDL, and high LDL, respectively. The categorization of LDL will help if features such as effect modification and linearity wish to be determined. Within each group defined by serum LDL level, for continuous variables (age, weight, pack years of smoking) we include the mean and standard deviation. For binary variables (sex and death) we present percentages. A Kaplan-Meier plot is presented to assess the survival curves of the subjects per category of LDL. Note that a log transform on LDL has no effect on survival curves, since they only depend on time and death. Furthermore, the same plot is used to consider the linearity of the LDL variable. Note that two Kaplan-Meier plots are given: the first uses a proper survival scale ranging from 0 to 100%. The second Kaplan-Meier plot is the same plot, but zoomed in so that we can see more details of the survival distributions. The instantaneous hazard of death over the entire period of observation across groups defined by continuous serum LDL under a quadratic transformation will be modeled using a Cox proportional hazards regression model. Note that an un-transformed LDL variable will also be placed in the regression. The hazard ratio and 95% CI was computed using Cox proportional hazards regression with the Huber-White sandwich estimator of the (robust) standard errors. Two-sided p-values based on Wald statistics were interpreted at the 5% significance level.
      2. **Results**: Data is available on 735 subjects, however 10 of those subjects (including 2 who died within 5 years) are missing data. Those subjects are omitted from all analyses. None of the 725 subjects were missing data on any other variables of interest for this analysis. Of the 725 subjects with available measurements, 165 had serum LDL measurements less than or equal to 99 mg/dL, 228 had serum LDL measurements between 100 and 129 mg/dL inclusive, 225 had measurements between 130 mg/dL and 159 mg/dL inclusive, and 107 had measurements greater than or equal to 160 mg/dL. The table below presents descriptive statistics within these groups. Subjects having serum LDL in the two lowest levels of serum LDL were more likely to be male than in other intervals. No consistent trend was seen across groups in age, weight, or smoking history. Subjects with the lowest levels of serum LDL appeared to have a higher mortality rate: about 20% of subjects with LDL less than or equal to 99 mg/dL died within 5 years compared to about 13% in subjects with higher serum LDL at study entry. From the first Kaplan-Meier plot, we observe high levels of survival throughout the studies observation period regardless of LDL level with no censoring before 5 years, followed by high concentrations of censoring. In the zoomed in plot of the same survival curves, we observe a high number of crossovers between LDL subgroups before 1500 days. However, after 1500 days, the survival curves split off into distinct higher LDL and lower LDL levels.
      3. **More Results**: From a Cox proportional hazards regression analysis, we estimate that for each 10 mg/dL increase in serum LDL, the instantaneous risk of death for the higher LDL group is 77.8% the risk of the lower LDL group. Based on a 95% confidence interval, it would not be unusual if the true risk of death for a higher (by 10 mg/dL) LDL group were anywhere between 63.6% and 93.5% of the risk for the lower LDL group. At the 5% significance level, we reject the null hypothesis of no association between survival and LDL (P-value<0.001). After adjusting for the quadratic term, the linear term remains statistically significant (P=0.007). However, at the 5% significance level, there is not sufficient evidence to support the importance of a quadratic LDL term after adjusting for a linear LDL term (P=0.054). This lack of evidence for nonlinearity between LDL and death at 5 years does not prove a linear relationship, since the relationship could be nonlinear in a way that a quadratic polynomial cannot detect.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Serum Low Density Lipoprotein (LDL) | | | | | |
|  | <=99 mg/dL (n=165) | 100-129 md/dL (n=228 | 130-159mg/dL (n=225) | >=160 mg/dL (n=107) | Any level (n=725) |
| Male (%) | 57.6 | 53.9 | 43.1 | 42.1 | 49.7 |
| Age (yrs) | 74.8 (5.50) | 74.6 (5.08) | 74.2 (5.62) | 74.9 (5.77) | 74.6 (5.45) |
| Weight (lbs) | 160 (31.5) | 160 (28.8) | 158 (32.3) | 163 (30.7) | 160 (30.8) |
| Smoking (pack-years) | 17.5 (24.04) | 21.5 (28.80) | 20.0 (28.83) | 18.1 (24.41) | 19.3 (27.16) |
| Death within 5 years (%) | 20 | 18.9 | 12.9 | 13.1 | 16.4 |

\*Standard Deviations in ( )





* 1. For each population defined by serum LDL value, compute the hazard ratio relative to a group having serum LDL of 160 mg/dL. (This will be used in problem 4). If *HR* is the hazard ratio (use the actual hazard ratio estimate) obtained from your regression model for the LDL term and *HR2* is the hazard ratio (use the actual hazard ratio estimate) obtained from your regression model for the squared LDL term, this can be effected by the Stata code

gen fithrC = *HR^((ldl* - 160)) \* *HR^((ldl* - 160)^2)

It could also be computed by creating a centered LDL variable, and then using the Stata predict command

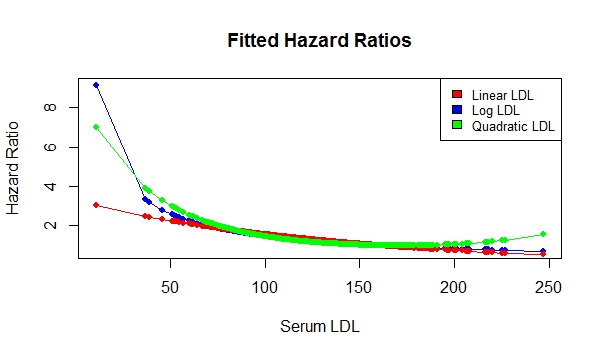
gen cldl = ldl – 160

gen cldlsqr= cldl ^ 2

stcox cldl cldlsqr

predict fithrC

1. Display a graph with the fitted hazard ratios from problems 1 – 3. Comment on any similarities or differences of the fitted values from the three models.
   1. Using the generated fitted hazard ratios per proportional hazards regression model each with its uniform transform of LDL from part b of the previous problems, the fitted hazard ratios in each model are plotted against LDL. Three LDL variable models are represented: LDL as linear, log LDL, and quadratic LDL.
   2. It is expected that the quadratic LDL model produces the most curved fitted hazard ratios, although the log LDL model produces the most skewed distribution of fitted hazard ratios. The three models behave similarly between 100 and 160 mg/dL serum LDL. For the highest LDL category (>=160 mg/dL), linear and log LDL models behave the most similarly. However, in the lowest LDL category (<100 mg/dL), the three models behave dissimilarly, which is likely due to lack of sample size within that low stratum.



**Discussion Sections: January 27 – 31, 2014**

We continue to discuss the dataset regarding FEV and smoking in children. Come do discussion section prepared to describe the approach to the scientific question posed in the documentation file fev.doc.