Biost 518: Applied Biostatistics II
Biost 515: Biostatistics II
Emerson, Winter 2014

Homework #4 Key
February 3, 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.
    a. Include full description of your methods, appropriate descriptive statistics, and full report of your inferential statistics.

Answer: *(In providing descriptive and inferential statistics, it is always best to lead off with some overall description of the statistical information that is available. When analyzing differences in mean, ratios of geometric means, differences in proportions, and odds ratios, that is usually summarized by the sample size. When using regression analyses with proportions and odds, we sometimes characterize the total number of events, though this is not as necessary in two group analyses. When analyzing censored time to event, we want to know the overall sample sizes, summary measures of the censoring distribution, and the number of observed events. Note that we estimate the censoring distribution using Kaplan-Meier estimates where censored observations are "events" and deaths (or whatever) represent censored observations of the time that we would have stopped follow-up of the individuals.*

Statistical Methods for descriptive statistics: Descriptive statistics for the censoring distribution included the minimum and maximum observed censoring times and the Kaplan-Meier estimates of the 10^{th}, 50^{th} (median), and 90^{th} percentiles, as well as the mean time of follow-up calculated as the area under the Kaplan-Meier estimate of the censoring distribution's survivor curve.

Descriptive statistics for serum LDL levels included the number of cases with missing data, as well as the minimum, maximum, mean, standard deviation, and the 25th, 50th (median), and 75th percentiles for the cases with available data. For the purposes of descriptive statistics of the survival probabilities by serum LDL level, serum LDL was categorized according to the Mayo Clinic guidelines: less than 70 mg/dL, 70-99 mg/dL, 100-129 mg/dL, 130-159 mg/dL, 160-189 mg/dL, and greater than or equal to 190 mg/dL. Within these categories, Kaplan-Meier estimates of survival were calculated and graphed, and estimates of the 2 and 5 year survival probabilities, as well as the 10th and 20th percentiles of the survival distribution and the restricted mean survival during a period of observation that all LDL strata still had some subjects at risk (5.75 years). *(These descriptive statistics will suffice for all problems. The key issue is that I would want to have multiple strata in order to be able to glean some information about nonlinearity in the association between all cause mortality and LDL. While it is true that the way I might ideally categorize LDL to investigate linearity of log LDL versus the categories I would investigate linearity of untransformed LDL or quadratic relationships, the range of LDL measurements did not really allow substantially different categorizations. Hence, I used the scientifically determined categories based on the Mayo Clinic recommendations.)*

<u>Descriptive statistics</u>: The study consisted of 735 subjects who were followed for death from any cause for a Kaplan-Meier estimated average of 5.33 years (median 5.66 years, range 5.00 to 5.91 years), during which time 133 deaths were observed. Serum LDL measurements at the time of study enrollment were not available on 10 subjects, two of whom were observed to die after 0.189 and 0.657 years of observation, with the remaining subjects still alive after 5.05 to 5.91 years of observation. In the 725 subjects with available serum LDL measurements at enrollment, the mean LDL was 126 mg/dL (SD 33.6 mg/dL, range 11 to 247 mg/dL).

Table 1 presents estimates of the survival distribution within strata defined by serum LDL and in the combined sample from the 725 subjects with available LDL measurements. The greatest difference in survival distributions is apparent when comparing those individuals having the lowest serum LDL levels (less than 70 mg/dL) at times after 2 years of follow-up. The 5 year survival probability is lowest in that group (59.1%) and is observed highest in the subjects having serum LDL between 160 and 189 mg/dL inclusive (88.0%). On average, the subjects in the lowest LDL stratum were estimated to average 4.91 years of life during the first 5.75 years following study enrollment, while the other strata averaged from 5.23 to 5.45 years. Figure 1 presents the Kaplan-Meier survival probability estimates graphically, where it is again the lowest LDL group that shows the most markedly different survival distribution.

Table 1: Kaplan-Meier based estimates of distribution of time from study enrollment to death from any cause for subjects having serum LDL measurements at baseline.

	Serum LDL at Study Enrollment						All Subjects (with LDL Available[3])
	11 – 69 mg/dL	70 – 99 mg/dL	100 – 129 mg/dL	130 – 159 mg/dL	160 – 189 mg/dL	190 – 247 mg/dL	
N Subjects	22	143	228	225	83	24	725
N Deaths	10	28	44	34	11	4	131
2 year Survival Probability[1]	100%	95.8%	93.9%	95.6%	98.8%	95.8%	95.6%
5 Year Survival Probability[1]	59.1%	83.2%	81.1%	87.1%	88.0%	83.3%	83.6%
10th Pctile of Survival[1]	3.46 y	3.80 y	3.41 y	4.30 y	4.53 y	4.13 y	3.66 y
20th Pctile of Survival[1]	3.55 y	5.44 y	5.36 y	NA[1]	NA[1]	NA[1]	5.54 y
5.75 Year Restricted Mean of Survival[2]	4.91 y	5.24 y	5.23 y	5.35 y	5.45 y	5.32 y	5.29 y

[1] Based on Kaplan-Meier estimates computed within strata defined by LDL and overall. NA indicates that the corresponding percentile is not estimable with the available data.

[2] Average number of years alive during the first 5.75 years following study enrollment, as computed by the area under Kaplam-Meier survival curves computed within strata defined by LDL and overall

[3] Ten of the 735 subjects in the study population were missing baseline serum LDL measurements. Two of those subjects were observed to die after 0.189 y and 0.657 years of observation. The remaining 8 subjects with missing LDL data were still alive at the end of their observation period 5.03 to 5.91 years after study enrollment

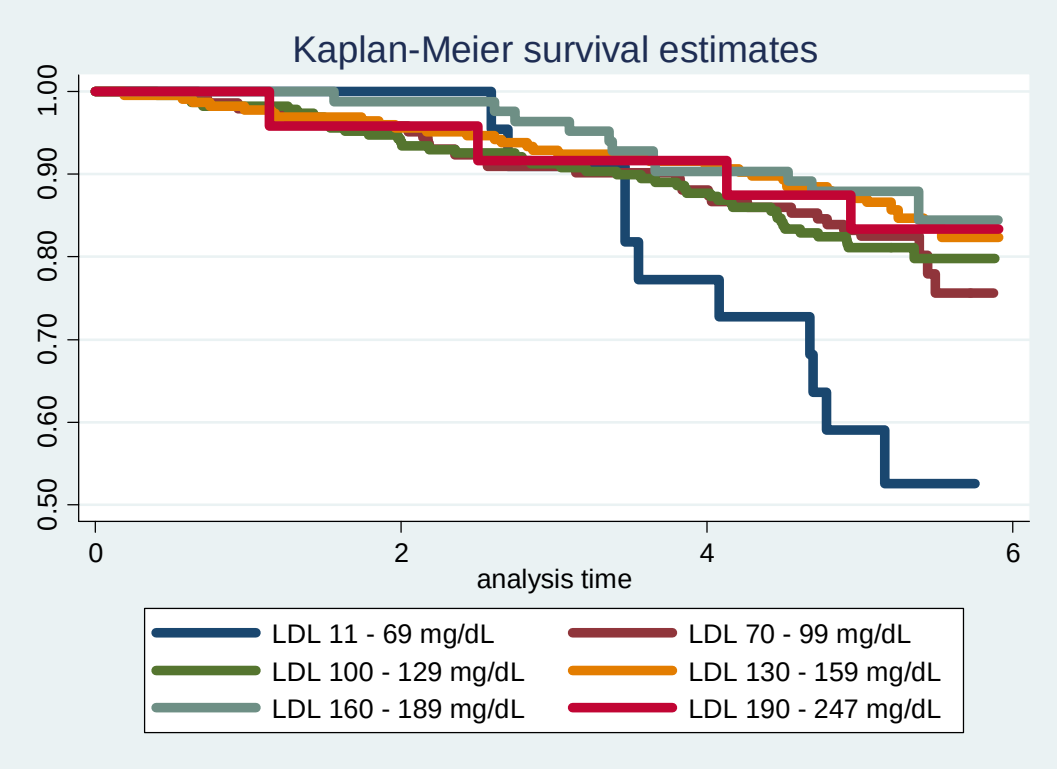


Figure 1: Kaplan-Meier based estimates of distribution of time from study enrollment to death from any cause for 725 subjects having serum LDL measurements at baseline.

<u>Statistical Methods for inferential statistics</u>: Distributions of time to death from any cause was compared across groups defined by serum LDL at baseline using proportional hazards regression modeling serum LDL as a continuous untransformed random variable. Quantification of association between all cause mortality was summarized by the hazards ratio computed from the regression model, with confidence intervals and two-sided p values computed using Wald statistics based on the Huber-White sandwich estimator. Subjects missing data for serum LDL at the time of study accrual were omitted from the analysis. *(Note that I would use this wording whether or not I had centered the log transformed LDL relative to an LDL of 160 mg/dL and whether I had divided LDL values by 10 in order to have Stata report HR on a more convenient scale for me..These are just reparameterizations of the same model, and the results quoted below would obtain from any of these alternative parameterizations. In the Stata output, I did run analyses with both the centered and uncentered LDL values, so you could see the similarity of results.)*

<u>Inferential results</u>: Data was available on 725 subjects having mean serum LDL of 126 mg/dL (SD 33.6 mg/dL; range 11 – 247 mg/dL). During an average of 5.33 years of observation, 131 of those subjects were observed to die. From a proportional hazards regression analysis, we estimate that the instantaneous risk of death is a relative 7.14% lower (hazard ratio 0.929) for each 10 mg/dL higher serum LDL level at baseline. Based on a 95% confidence interval, this observed hazard ratio suggesting lower death rates for groups of patients with higher LDL levels would not be judged unusual if the true instantaneous risk of death were anywhere from 1.80% to 12.2% lower in a group having baseline serum LDL 10 mg/dL higher than that in another group (95% CI for hazard ratio 0.878 to 0.982). A two-sided p value of 0.009 suggests that we can with high confidence reject the null hypothesis that the risk of death from any cause is not associated with serum LDL levels in favor of a tendency for lower mortality with higher serum LDL levels.

b. 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
```

Answer: *No answer necessary. Fitted values are to be displayed in problem 4.*

2. 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.

    a. 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.

Answer: See problem 1 for methods and report of descriptive statistics.

Statistical Methods for inferential statistics: Distributions of time to death from any cause was compared across groups defined by serum LDL at baseline using proportional hazards regression modeling serum LDL as a continuous logarithmically transformed random variable. Quantification of association between all cause mortality was summarized by the hazards ratio computed from the regression model, with confidence intervals and two-sided p values computed using Wald statistics based on the Huber-White sandwich estimator. Subjects missing data for serum LDL at the time of study accrual were omitted from the analysis. *(Note that I would use this wording whether or not I had centered the log transformed LDL relative to an LDL of 160 mg/dL and whether I had used natural log or the log base 1.1. These are just reparameterizations of the same model, and the results quoted below would obtain from any of these alternative parameterizations. In the Stata output, I did run analyses with both the centered and uncentered LDL values, so you could see the similarity of results.)*

Inferential results: Data was available on 725 subjects having mean serum LDL of 126 mg/dL (SD 33.6 mg/dL; range 11 – 247 mg/dL). During an average of 5.33 years of observation, 131 of those subjects were observed to die. From a proportional hazards regression analysis, we estimate that the instantaneous risk of death is a relative 7.58% lower (hazard ratio 0.924) for each 10% higher serum LDL level at baseline. Based on a 95% confidence interval, this observed hazard ratio suggesting lower death rates for groups of patients with higher LDL levels would not be judged unusual if the true instantaneous risk of death were anywhere from 4.09% to 10.9% lower in a group having baseline serum LDL 10 mg/dL higher than that in another group (95% CI for hazard ratio 0.878 to 0.982). A two-sided p value P < 0.0001 suggests that we can with high confidence reject the null hypothesis that the risk of death from any cause is not associated with serum LDL levels in favor of a tendency for lower mortality with higher serum LDL levels. *(Note that the Stata output only included 3 significant digits for the p value for the regression coefficient, but the overall Wald test p value provided 4 significant digits. Thus I was able to easily provide the maximal precision with which I tend to report p values.)*

    b. 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
```

Answer: *No answer necessary. Fitted values are to be displayed in problem 4.*

3. 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).

    a. 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.

Answer: See problem 1 for methods and report of descriptive statistics.

Statistical Methods for inferential statistics: Distributions of time to death from any cause was compared across groups defined by serum LDL at baseline using proportional hazards regression modeling a quadratic function of serum LDL by including both a continuous untransformed LDL random variable and a term that was the square of the LDL measurements. Quantification of association between all cause mortality was tested by simultaneously testing that both the linear term and the quadratic term had coefficients equal to zero. P values for this two degree of freedom test were based on the Wald statistic computed using standard errors based on the Huber-White sandwich estimator. For descriptive purposes, predicted values for the hazard ratios from this model were plotted for each observed level of LDL relative to a population having serum LDL of 160 mg/LDL, and those fitted values were compared to fitted values from proportional hazards regression models that included only a linear untransformed measure of LDL or only a logarithmically transformed LDL measurement. As a secondary test, a test for nonlinearity of the association between instantaneous risk of death from any cause and serum LDL was performed using a Wald test that the regression coefficient for the squared LDL term was zero. Again, standard errors for this test for nonlinearity were based on the Huber-White sandwich estimator. Subjects missing data for serum LDL at the time of study accrual were omitted from the analysis.

Inferential results: Data was available on 725 subjects having mean serum LDL of 126 mg/dL (SD 33.6 mg/dL; range 11 – 247 mg/dL). During an average of 5.33 years of observation, 131 of those subjects were observed to die. From a proportional hazards regression analysis modeling a quadratic function of LDL (linear and squared terms), we find a statistically significant association between instantaneous risk of death and serum LDL at baseline (two-sided P = 0.0005). Figure 2 displays the fitted hazard ratios relative to a population having serum LDL of 160 mg/dL. From this curve, it can be seen that the quadratic fit estimates a U-shaped curve having a nadir at 171 mg/dL. This fitted curve does not differ markedly from proportional hazards regression fits based on untransformed or logarithmically transformed LDL. A test for nonlinearity based on the squared LDL term in the quadratic proportional hazards regression is barely not statistically significant at the 0.05 level (P = .055) suggesting that we do not have strong evidence that the true association between death from any cause and serum LDL is not adequately described by a log hazard function that is linear in LDL.

b. 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)) * HR2^(ldl^2 - 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
```

Answer: *No answer necessary. Fitted values are to be displayed in problem 4.*

4. 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.

Answer: Figure 2 display the fitted values from the three models. In each case, the model predicts a trend that is predominantly downward with higher LDL. There is not tremendous difference between the three models over the mid range of LDL, and the greatest differences between the models occur in the lowest ranges of LDL, where our data is relatively sparse. The fits from the logarithmic transformation and the quadratic function are remarkably similar. To the extent that this is an accurate representation of the true association, the greater interpretability of the logarithmic transformation makes that more attractive to me. It should be noted that the U-shape seen in the quadratic fit cannot be taken as proof that the highest LDL groups actually have increase risk over the groups with moderate levels: A quadratic curve ultimately has to be U-shaped over the whole real line, just as the linear and logarithmic curves must be monotonic (steadily increasing or steadily decreasing). It would require a much larger sample size and a more careful analysis to establish whether there is really a tendency for both lower and higher LDL to be associated with higher risk of death in this population..

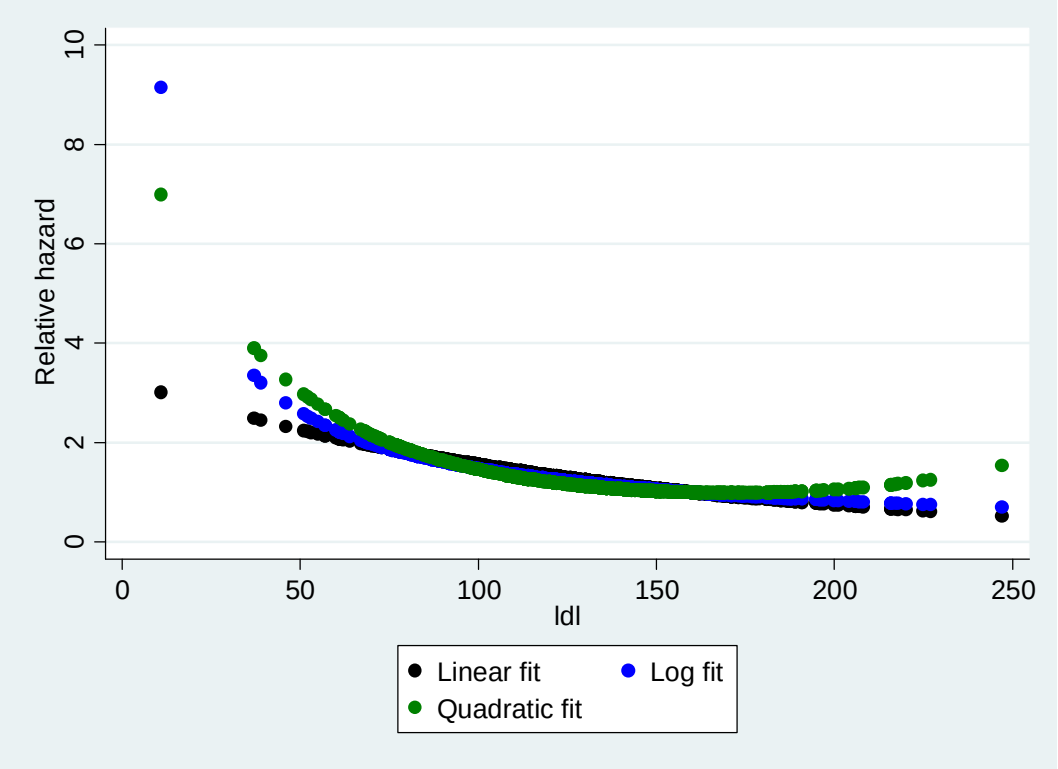


Figure 2: Plots of the fitted hazard ratios from proportional hazards regression models including only a linear continuous term for LDL, only a logarithmically transformed term for LDL, or a quadratic function of LDL that includes both a linear and squared term. All hazard ratio estimates are relative to a group having serum LDL of 160 mg/dL..

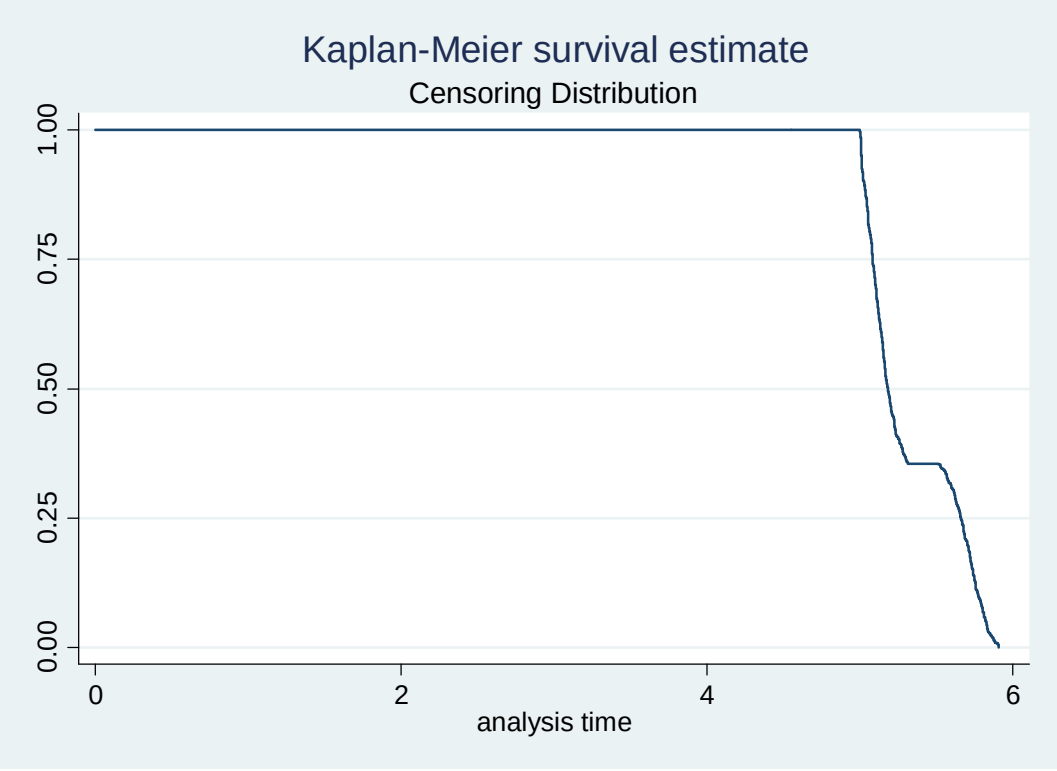
Stata Code and Output

```
. /// Reading in and preparing data

. quietly:                                                                    ///
> infile ptid mridate age male race weight height packyrs yrsquit alcoh physact   ///
>        chf chd stroke diabetes genhlth ldl alb crt plt sbp aai fev dsst atrophy ///
>            whgrd numinf volinf obstime death                                ///
>            using http://www.emersonstatistics.com/datasets/mri.txt

. list in 1
     +-------------------------------------------------------------------------------------------+
  1. | ptid | mridate | age | male | race | weight | height | packyrs | yrsquit | alcoh |
     |    . |       . |   . |    . |    . |      . |      . |       . |       . |     . |
     |-------------------------------------------------------------------------------------------|
     | physact | chf | chd | stroke | diabetes | genhlth | ldl | alb | crt | plt | sbp  |
     |       . |   . |   . |      . |        . |       . |   . |   . |   . |   . |   . |
     |-------------------------------------------------------------------------------------------|
     | aai  | fev  | dsst  | atrophy  | whgrd  | numinf  | volinf  | obstime  |  death  |
     |   .  |   .  |    .  |        .  |    .  |      .  |       .  |        .  |       .  |
     +-------------------------------------------------------------------------------------------+

. drop in 1
(1 observation deleted)

. replace obstime= obstime / 365.25
(735 real changes made)

. egen ldlCTG = cut(ldl), at(0 70 100 130 160 190 250)
(10 missing values generated)
```

```
. /// Descriptive statistics for censoring distribution

. bysort death: tabstat obstime, stat(n max) col(stat) by(ldlCTG)

-> death = 0
Summary for variables: obstime
     by categories of: ldlCTG

  ldlCTG |         N       max
---------+--------------------
       0 |        12  5.754962
      70 |       115  5.878166
     100 |       184  5.883641
     130 |       191  5.908282
     160 |        72  5.905544
     190 |        20  5.908282
---------+--------------------
   Total |       594  5.908282

-> death = 1
Summary for variables: obstime
     by categories of: ldlCTG

  ldlCTG |         N       max
---------+--------------------
       0 |        10  5.166325
      70 |        28  5.494866
     100 |        44  5.357974
     130 |        34  5.535934
     160 |        11   5.38809
     190 |         4  4.944559
---------+--------------------
   Total |       131  5.535934

. g censor = 1 - death

. stset obstime censor

     failure event:  censor != 0 & censor < .
obs. time interval:  (0, obstime]
 exit on or before:  failure
```

```
      735  total obs.
        0  exclusions
      735  obs. remaining, representing
      602  failures in single record/single failure data
 3630.376  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =   5.91102
```

```
. sts graph, t1("Censoring Distribution")

        failure _d:  censor
  analysis time _t:  obstime
```



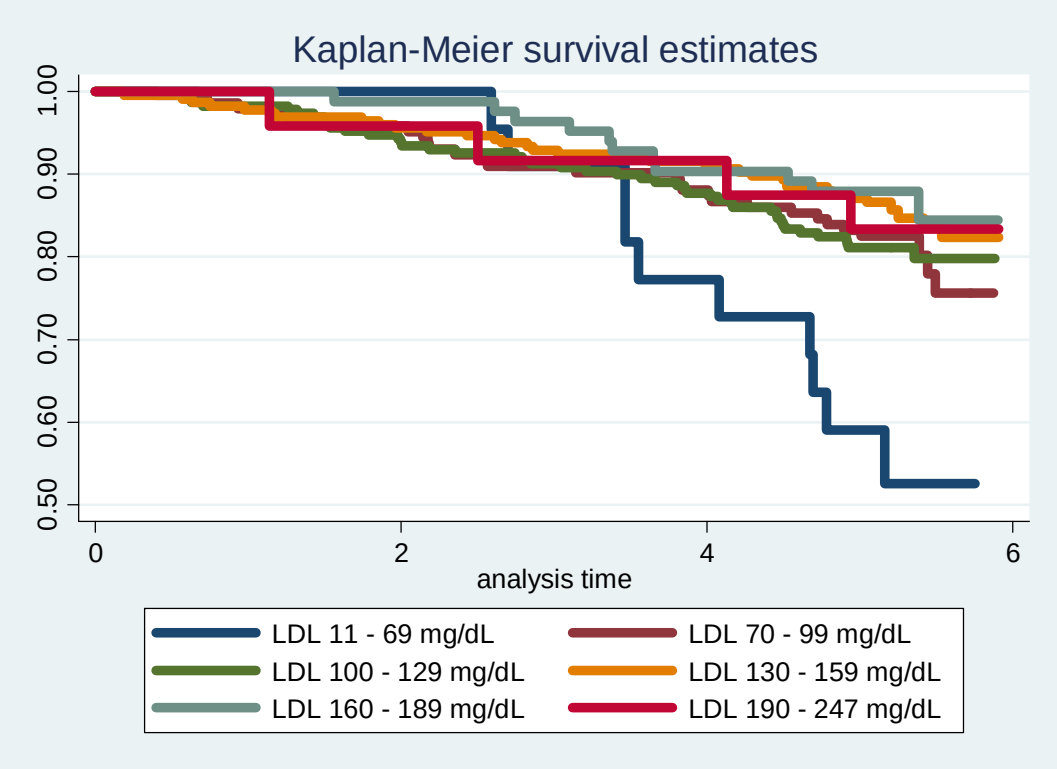
```
. stci, p(10)

        failure _d:  censor
  analysis time _t:  obstime

             |    no. of
             |  subjects         10%     Std. Err.     [95% Conf. Interval]
-------------+-------------------------------------------------------------
       total |       735    5.029432      .0047079      5.01848    5.04038

. stci, p(25)

        failure _d:  censor
  analysis time _t:  obstime

             |    no. of
             |  subjects         25%     Std. Err.     [95% Conf. Interval]
-------------+-------------------------------------------------------------
       total |       735    5.086927      .0069813      5.07598    5.09788

. stci, p(50)

        failure _d:  censor
  analysis time _t:  obstime

             |    no. of
             |  subjects         50%     Std. Err.     [95% Conf. Interval]
-------------+-------------------------------------------------------------
       total |       735    5.185489      .0100226      5.16632    5.20465

. stci, p(75)

        failure _d:  censor
  analysis time _t:  obstime

             |    no. of
             |  subjects         75%     Std. Err.     [95% Conf. Interval]
-------------+-------------------------------------------------------------
       total |       735    5.664613      .0152841      5.62902    5.68652
```

```
. stci, p(90)

        failure _d:  censor
  analysis time _t:  obstime

             |    no. of
             |  subjects         90%     Std. Err.     [95% Conf. Interval]
-------------+-------------------------------------------------------------
       total |       735    5.776865      .0106913       5.7577    5.80424

. stci, rmean

        failure _d:  censor
  analysis time _t:  obstime

             |    no. of  restricted
             |  subjects        mean     Std. Err.     [95% Conf. Interval]
-------------+-------------------------------------------------------------
       total |       735    5.332998      .0120774      5.30933    5.35667
```

. /// Descriptive statistics for LDL

```
. tabstat ldl, stat(n mean sd min q max) col(stat)

    variable |         N      mean        sd       min       p25       p50       p75       max
-------------+--------------------------------------------------------------------------------
         ldl |       725  125.8028  33.60197        11       102       125       147       247

. list obstime death if ldl==.

     |  obstime   death |
-----+------------------+
 17. |  5.91102       0 |
 34. | 5.054072       0 |
 87. | 5.670089       0 |
236. | 5.546885       0 |
511. |   5.7577       0 |
529. | 5.790555       0 |
555. | 5.262149       0 |
589. |  5.14716       0 |
608. | .6570842       1 |
700. | .1889117       1 |
```

```
. stset obstime death

     failure event:  death != 0 & death < .
obs. time interval:  (0, obstime]
 exit on or before:  failure

      735  total obs.
        0  exclusions
      735  obs. remaining, representing
      133  failures in single record/single failure data
 3630.376  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =   5.91102

. sts graph, by(ldlCTG) plotopts(lwid(1.25)) ylabel(0.5(0.1)1.0) ///
>         legend(label(1 "LDL 11 - 69 mg/dL") label(2 "LDL 70 - 99 mg/dL") ///
>                label(3 "LDL 100 - 129 mg/dL") label(4 "LDL 130 - 159 mg/dL") ///
>                label(5 "LDL 160 - 189 mg/dL") label(6 "LDL 190 - 247 mg/dL"))

         failure _d:  death
   analysis time _t:  obstime
```


Kaplan-Meier survival estimates
1.00
0.90
0.80
0.70
0.60
0.50
0
2
4
6
analysis time
LDL 11 - 69 mg/dL
LDL 70 - 99 mg/dL
LDL 100 - 129 mg/dL
LDL 130 - 159 mg/dL
LDL 160 - 189 mg/dL
LDL 190 - 247 mg/dL

```
. sts list, by(ldlCTG) at(2 5)
         failure _d:  death
   analysis time _t:  obstime
```

Time	Beg. Total	Fail	Survivor Function	Std. Error	[95% Conf.	Int.]
ldlCTG=0						
2	0	0	1.0000	.	.	.
5	14	9	0.5909	0.1048	0.3610	0.7621
ldlCTG=70						
2	138	6	0.9580	0.0168	0.9090	0.9809
5	120	18	0.8322	0.0313	0.7601	0.8842
ldlCTG=100						
2	215	14	0.9386	0.0159	0.8985	0.9632
5	186	29	0.8114	0.0259	0.7543	0.8565
ldlCTG=130						
2	216	10	0.9556	0.0137	0.9190	0.9758
5	197	19	0.8711	0.0223	0.8199	0.9086
ldlCTG=160						
2	83	1	0.9880	0.0120	0.9175	0.9983
5	74	9	0.8795	0.0357	0.7876	0.9333
ldlCTG=190						
2	24	1	0.9583	0.0408	0.7392	0.9940
5	21	3	0.8333	0.0761	0.6148	0.9339

```
Note:  survivor function is calculated over full data and evaluated at
       indicated times; it is not calculated from aggregates shown at left.
```

```
. sts list, at(2 5)
         failure _d:  death
   analysis time _t:  obstime
```

Time	Beg. Total	Fail	Survivor Function	Std. Error	[95% Conf.	Int.]
2	702	34	0.9537	0.0077	0.9359	0.9667
5	615	87	0.8354	0.0137	0.8065	0.8603

```
Note:  survivor function is calculated over full data and evaluated at
       indicated times; it is not calculated from aggregates shown at left.
```

```
. stci, p(10) by(ldlCTG)
        failure _d:  death
  analysis time _t:  obstime
```

ldlCTG	no. of subjects	10%	Std. Err.	[95% Conf.	Interval]
0	22	3.463381	.4700722	2.59001	3.55099
70	143	3.800137	.6540937	2.14374	4.72279
100	228	3.411362	.5369233	1.99316	4.16975
130	225	4.295688	.5982186	2.65572	5.20739
160	83	4.533881	.7268575	2.74333	.
190	24	4.128679	2.056744	1.13621	.
total	725	3.663244	.2847362	3.02533	4.16975

```
. stci, p(20) by(ldlCTG)
        failure _d:  death
  analysis time _t:  obstime
```

ldlCTG	no. of subjects	20%	Std. Err.	[95% Conf.	Interval]
0	22	3.550992	.4054107	2.59001	4.69541
70	143	5.442847	.	4.03012	.
100	228	5.357974	.	4.45722	.
130	225	.	.	5.2512	.
160	83	.	.	4.6872	.
190	24	.	.	1.13621	.
total	725	5.535934	.	5.20739	.

```
. /// To compute restricted mean I restrict observations to 5.75 years based on censoring
> /// distribution within the LDL strata

. g obs575= obstime

. replace obs575= 5.75 if obstime > 5.75
(80 real changes made)

. g death575= death
```

```
. replace death575= 0 if obstime > 5.75
(0 real changes made)

. stset obs575 death575

     failure event:  death575 != 0 & death575 < .
obs. time interval:  (0, obs575]
 exit on or before:  failure

      735  total obs.
        0  exclusions
      735  obs. remaining, representing
      133  failures in single record/single failure data
 3625.311  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =      5.75

. stci, rmean by(ldlCTG)

         failure _d:  death575
   analysis time _t:  obs575

             |    no. of  restricted
ldlCTG       |  subjects        mean      Std. Err.    [95% Conf. Interval]
-------------+-------------------------------------------------------------
           0 |        22    4.905046(*)   .2297088      4.45483    5.35527
          70 |       143    5.237235(*)   .1016258      5.03805    5.43642
         100 |       228    5.226007(*)    .081784      5.06571     5.3863
         130 |       225    5.351399(*)   .0746197      5.20515    5.49765
         160 |        83    5.445542(*)   .0929914      5.26328     5.6278
         190 |        24    5.321327(*)   .2308693      4.86883    5.77382
-------------+-------------------------------------------------------------
       total |       725    5.285568(*)   .0428303      5.20162    5.36951

(*) largest observed analysis time is censored, mean is underestimated
```

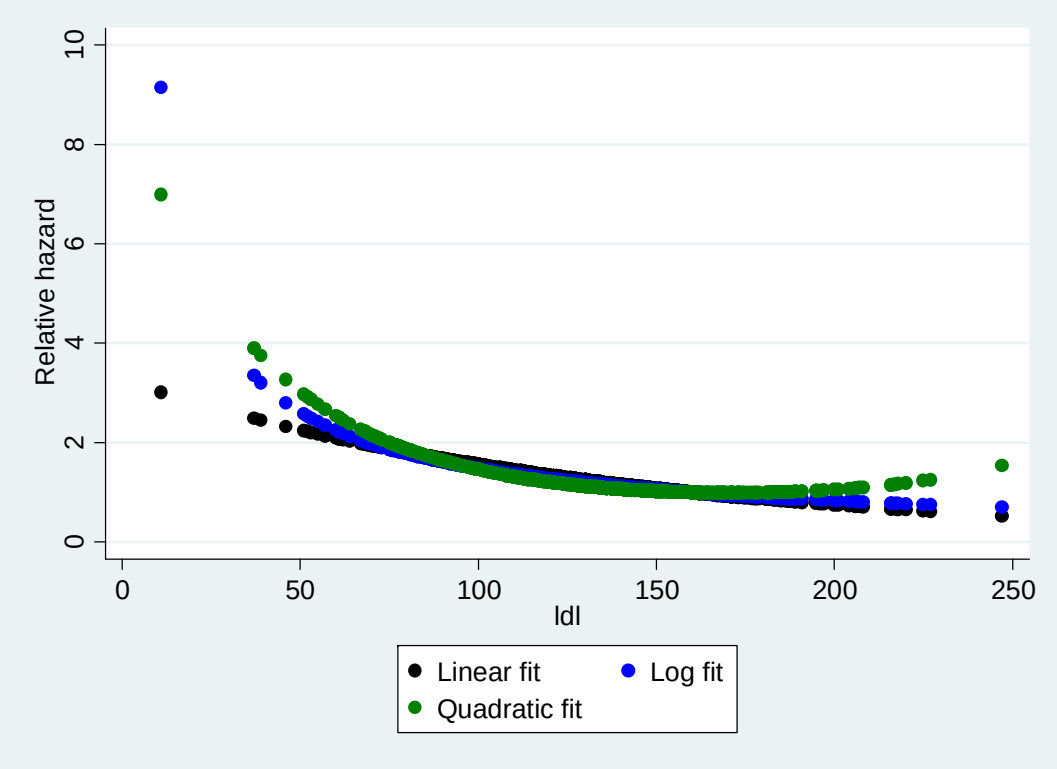
```
. /// Problem 1

. stset obstime death

     failure event:  death != 0 & death < .
obs. time interval:  (0, obstime]
 exit on or before:  failure

      735  total obs.
        0  exclusions
      735  obs. remaining, representing
      133  failures in single record/single failure data
 3630.376  total analysis time at risk, at risk from t =         0
                           earliest observed entry t =         0
                                last observed exit t =   5.91102

. stcox ldl, robust
        failure _d:  death
  analysis time _t:  obstime

Cox regression -- Breslow method for ties

No. of subjects      =          725                 Number of obs   =       725
No. of failures      =          131
Time at risk         =  3585.390827
                                                    Wald chi2(1)    =      6.76
Log pseudolikelihood =   -836.48275                 Prob > chi2     =    0.0093
------------------------------------------------------------------------------
             |               Robust
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ldl |   .9926246   .0028267    -2.60   0.009     .9870997    .9981804
------------------------------------------------------------------------------

. g cldl= ldl - 160
(10 missing values generated)
```

```
. stcox cldl, robust
        failure _d:  death
  analysis time _t:  obstime

Cox regression -- Breslow method for ties

No. of subjects      =          725                 Number of obs    =       725
No. of failures      =          131
Time at risk         =  3585.390827
                                                    Wald chi2(1)     =      6.76
Log pseudolikelihood =   -836.48275                 Prob > chi2      =    0.0093
------------------------------------------------------------------------------
             |               Robust
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cldl |   .9926246   .0028267    -2.60   0.009     .9870997    .9981804
------------------------------------------------------------------------------

. predict fithrA
(option hr assumed; relative hazard)
(10 missing values generated)
```

```
. /// Problem 2 (I use base 1.1 logarithm to be able to talk about 10% difference)

. g logldl= log(ldl) / log(1.1)
(10 missing values generated)

. stcox logldl, robust
         failure _d:  death
   analysis time _t:  obstime

Cox regression -- Breslow method for ties

No. of subjects      =          725                 Number of obs   =       725
No. of failures      =          131
Time at risk         =  3585.390827
                                                    Wald chi2(1)    =     17.39
Log pseudolikelihood =   -835.39585                 Prob > chi2     =    0.0000
------------------------------------------------------------------------------
             |               Robust
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      logldl |   .9242374   .0174635    -4.17   0.000     .8906355     .959107
------------------------------------------------------------------------------

. g clogldl= log(ldl / 160) / log(1.1)
(10 missing values generated)

. stcox clogldl, robust
         failure _d:  death
   analysis time _t:  obstime

Cox regression -- Breslow method for ties

No. of subjects      =          725                 Number of obs   =       725
No. of failures      =          131
Time at risk         =  3585.390827
                                                    Wald chi2(1)    =     17.39
Log pseudolikelihood =   -835.39585                 Prob > chi2     =    0.0000
------------------------------------------------------------------------------
             |               Robust
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     clogldl |   .9242374   .0174635    -4.17   0.000     .8906355     .959107
------------------------------------------------------------------------------
```

```
. predict fithrB
(option hr assumed; relative hazard)
(10 missing values generated)


. /// Problem 3

. g ldlsqr= ldl^2
(10 missing values generated)

. stcox ldl ldlsqr, robust
        failure _d:  death
  analysis time _t:  obstime

Cox regression -- Breslow method for ties

No. of subjects      =          725                 Number of obs   =       725
No. of failures      =          131
Time at risk         =  3585.390827
                                                    Wald chi2(2)    =     15.28
Log pseudolikelihood =   -835.23455                 Prob > chi2     =    0.0005
------------------------------------------------------------------------------
             |               Robust
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ldl |   .9742307   .0095294    -2.67   0.008     .9557314     .993088
      ldlsqr |   1.000076   .0000398     1.92   0.055     .9999984    1.000154
------------------------------------------------------------------------------

. g cldlsqr= cldl^2
(10 missing values generated)
```

```
. stcox cldl cldlsqr, robust

         failure _d:  death
   analysis time _t:  obstime

Cox regression -- Breslow method for ties

No. of subjects      =          725                 Number of obs   =       725
No. of failures      =          131
Time at risk         =  3585.390827
                                                    Wald chi2(2)    =     15.28
Log pseudolikelihood =   -835.23455                 Prob > chi2     =    0.0005
------------------------------------------------------------------------------
             |               Robust
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cldl |   .9983156   .0039568    -0.43   0.671     .9905903    1.006101
     cldlsqr |   1.000076   .0000398     1.92   0.055     .9999984    1.000154
------------------------------------------------------------------------------

. predict fithrC
(option hr assumed; relative hazard)
(10 missing values generated)

. twoway (scatter fithrA fithrB fithrC ldl, color(black blue green)), ///
>         legend(label(1 "Linear fit") label(2 "Log fit") label(3 "Quadratic fit"))
```


Relative hazard
0
2
4
6
8
10
0
50
100
150
200
250
ldl
Linear fit
Log fit
Quadratic fit