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.

Methods: The instantaneous risk of subjects dying were compared over the entire period of study enrollment across groups defined by serum LDL as a continuous variable using a simple proportional hazards regression model. Descriptive statistics were generated and displayed as Kaplan Meier table and plot within groups defined by serum LDL measurements (less than or equal to 129 mg/dL, between 130 and 159 mg/dL inclusive, and greater than or equal to 160 mg/dL). Statistical inference on the hazards ratio was computed from the regression slope parameter, with two-sided p value and 95% confidence interval.

Descriptive statistics:



Summary table

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| LDL level (mg/dL) | N | mean | SD | min | max |
| < 130 | 393 | 101.255 | 19.297 | 11 | 129 |
| 130-160 | 225 | 142.733 | 8.528 | 130 | 159 |
| > 160 | 107 | 180.365 | 18.260 | 160 | 247 |
| Total | 725 | 125.803 | 33.602 | 11 | 247 |

Kaplan-Meier table

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Time (month) | Number of subjects at risk of death | Death | Probability to survive | Std. Error | [95% C.I.] |  |
| ldl < 130(mg/dL) |  |  |  |  |  |  |
| 12 | 387 | 7 | 0.982 | 0.007 | [0.963, | 0.992] |
| 24 | 374 | 13 | 0.949 | 0.011 | [0.922, | 0.967] |
| 36 | 359 | 15 | 0.911 | 0.014 | [0.878, | 0.935] |
| 48 | 344 | 15 | 0.873 | 0.017 | [0.836, | 0.902] |
| 60 | 318 | 26 | 0.807 | 0.020 | [0.764, | 0.842] |
| 130<ldl<160(mg/dL) |  |  |  |  |  |  |
| 12 | 221 | 5 | 0.978 | 0.010 | [0.947, | 0.991] |
| 24 | 216 | 5 | 0.956 | 0.014 | [0.919, | 0.976] |
| 36 | 210 | 6 | 0.929 | 0.017 | [0.887, | 0.956] |
| 48 | 206 | 4 | 0.911 | 0.019 | [0.866, | 0.942] |
| 60 | 197 | 9 | 0.871 | 0.022 | [0.820, | 0.909] |
| ldlCTG>160 (mg/dL) |  |  |  |  |  |  |
| 12 | 0 | 0 | 1.000 | . | . | . |
| 24 | 106 | 2 | 0.981 | 0.013 | [0.927, | 0.995] |
| 36 | 103 | 3 | 0.953 | 0.020 | [0.891, | 0.980] |
| 48 | 98 | 5 | 0.907 | 0.028 | [0.833, | 0.949] |
| 60 | 94 | 4 | 0.869 | 0.033 | [0.789, | 0.920] |

The graph and table above depicts Kaplan-Meier estimates of survival probability for the 393 subjects whose serum LDL was less than 130 mg/dL, 225 subjects whose serum LDL was between 130-160 mg/dL and the 107 subjects with serum LDL greater than or equal to 160 mg/dL. Kaplan-Meier curves shows that it’s hard to distinguish the survival probability between 130-160 LDL group and >160 LDL group but <130 LDL group likely to have a lower survival probability towards the end of the study. Group with LDL less than 130 mg/dL has the biggest sample size (393), which is almost 4 times as the sample size for LDL > 160 group. Also, most of the death in the study occurred in the last two years.

Results:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Hazard Ratio | SE | Z | p-value | 95% C.I. |  |
| LDL level | 0.9926246 | 0.00282 | -2.6 | 0.009 | [0.987, | 0.998] |

The instantaneous risk of death is estimated to be 7.14% lower (hazard ratio 0.9286) for each 10 mg/dl difference in LDL level, with the group having the higher level of LDL tending toward a lower instantaneous risk of death. This observed difference is statistically different from an hazard ratio of 1 (P = 0.009), with a 95% confidence interval suggesting that the observed hazard ratio is what might be typically observed if the true instantaneous risk of dying was anywhere between 1.8% and 12.18% lower for each 10/mg/dl higher LDL level. We thus reject the null hypothesis of no association between survival time and LDL level at study entry in favor of a trend toward lower risk of death among subjects with higher LDL levels.

\*(A 1mg/dl difference in LDL is not clinically important. I think it’d be useful to use a 10 mg/dl difference, and I used the estimates by exponentiating the HR based on a 1 mg/dl by 10.)

* 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.

Methods: The instantaneous risk of subjects dying were compared over the entire period of study enrollment across groups defined by serum LDL as a continuous logarithmically transformed variable using a simple proportional hazards regression model. Descriptive statistics were generated and displayed as Kaplan Meier table and plot within groups defined by serum LDL measurements (less than or equal to 129 mg/dL, between 130 and 159 mg/dL inclusive, and greater than or equal to 160 mg/dL). Statistical inference on the hazards ratio was computed from the regression slope parameter, with two-sided p value and 95% confidence interval.

Descriptive statistics: See question 1 part a.

Results:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Hazard Ratio | SE | Z | p-value | 95% C.I. |  |
| Log LDL level | 0.438 | 0.087 | -4.17 | 0.000 | [0.297 | 0.645] |

When comparing two groups with different LDL levels, the instantaneous risk of dying is estimated to be 7.57% lower (hazard ratio 0.924) for each 10% difference in LDL level, with the group having the higher level of cholesterol tending toward a lower instantaneous risk of death. This observed difference is statistically different from an odds ratio of 1 (P < 0.0005), with a 95% confidence interval suggesting that the observed hazard ratio is what might be typically observed if the true instantaneous risk of death was anywhere between 4.09% and 10.93% lower for each 10% higher LDL level. We thus reject the null hypothesis of no association between survival time and LDL level at study entry in favor of a trend toward risk of death among subjects with higher LDL levels is lower.

\*(The HR estimated per 1 unit of log LDL is not of much clinical interest. I used 10% difference for the purpose of greater interest, and I used the estimates by exponentiating the HR based on a 1% by natural logarithm of 1.1)

3/5 for performing an appropriate analysis

Did not report whether using Huber-White sandwich estimator or not (-1)

Did not report which statistic the statistical inference is based on (-1)

4/5 for reporting the association appropriately

Did not report the study population (-1)

Total: 7

* 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.

Methods: The instantaneous risk of subjects dying were compared over the entire period of study enrollment across groups by fitting a proportional hazards model with both LDL and LDL squared variable. Descriptive statistics were generated and displayed as Kaplan Meier table and plot within groups defined by serum LDL measurements (less than or equal to 129 mg/dL, between 130 and 159 mg/dL inclusive, and greater than or equal to 160 mg/dL). Statistical inference on the hazards ratio was computed from the regression slope parameter, with two-sided p value and 95% confidence interval.

Descriptive statistics: See question 1 part a.

Results:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Hazard Ratio | Robust SE | z | P-value | 95% C.I. |  |
| LDL | 0.9742 | 0.0095294 | -2.67 | 0.008 | 0.9557314 | 0.993088 |
| LDL^2 | 1.000076 | 0.0000398 | 1.92 | 0.055 | 0.9999984 | 1.000154 |
| chi-square statistic = 15.28 | p-value (prob > chi-sq) = 0.0005 |

According to the output, p-value of chi-square is 0.0005 < 0.05. Thus, we conclude there’s significant association between LDL level and log hazard function. And the p-value for squared LDL term is 0.055 > 0.05. Therefore, we cannot say there exists a significant non-linear relationship between LDL and mortality.

(The output above is from the model of LDL and hazard instead of log hazard. But since they are 1-to-1 transformed, the model for log hazard gives us the same p-values).

* 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)) \* *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

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.



First of all, since the all the hazard ratio is centered at LDL = 160, we see that fitted hazard ratio is 1 when LDL = 160 mg/dl. And there also exists a clear trend that as LDL level goes down, hazard ratio would increase. At the lower bound of LDL, fitted hazard ratio under quadratic model > log model > original model. As LDL level goes up to the extreme, fitted hazard ratio under quadratic model would rise as well whereas fitted values of hazard ratio under the other two models tend to keep going downward.