BIOST 513 Homework #4

Due date: February 3, 2014

Total score: 7+8+7+7=29

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.

Score: 7

1. Include full description of your methods, appropriate descriptive statistics, and full report of your inferential statistics.

Comments: 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 should be listed

**ANSWER:**

**Methods:**

**Proportional hazard regression was performed using death as the binary outcome and time in years as the time element. The predictor of interest is LDL level. The summary measure is the instantaneous risk (hazard) of death at each time among those subjects who have not died across groups defined by serum LDL as a continuous variable. We assume that the missing data is right-censored and non-informative. Descriptive analysis was performed using Kaplan-Meier survival curves among different LDL categories (LDL < 130, LDL 130 to < 160, LDL 160 to < 190, and LDL >/= 190 mg/dL).**

**Proportional hazards assumes that the ratio of these instantaneous rates is constant in time between two groups. Parameter estimates were determined using maximum partial likelihood estimation. P-values (two-tailed) and confidence intervals were derived using Wald-based estimates. Huber-White sandwich estimator was used to estimate the standard error.**

**Results:**

**Of the 735 patients in the study, 725 had values for serum LDL. A total of 133 (18.10%) patients died in the study. Patients with an LDL >/= 190 mg/dL was older compared to the other groups by approximately 1 year. Patients in the LDL 160 to 190 mg/dL group was slightly heavier compared to the other LDL groups. Patients in the LDL group >/= 190 mg/dL was shorter compared to the other groups. Patients in the LDL 130 to 160 mg/dL group had lower physical activity compared to the other groups. The proportion of males in the LDL >/= 190 mg/dL group was lower compared to the other groups. A larger proportion of Asians were in the LDL groups that had 130-160 mg/dL and >/= 190 mg/dL categories. Patients in the LDL >/= 190 mg/dL group had a lower proportion of history of smoking; but they had higher alcohol use. General health was different between the groups with the LDL 130 to 160mg/dL group having the highest score (Table 1).**

**Comparison between patients across different LDL categories (LDL < 130, LDL 130 to < 160, LDL 160 to < 190, and LDL >/= 190 mg/dL) was illustrated using Kaplan-Meier survival curves (Figure 1). From the KM curve, the LDL <130 group had a lower probability of survival at 5 years followed by the LDL 130 to < 160 and LDL 160 to < 190 mg/dL groups. The LDL >/= 190 group had the highest survival probability at 5 years.**

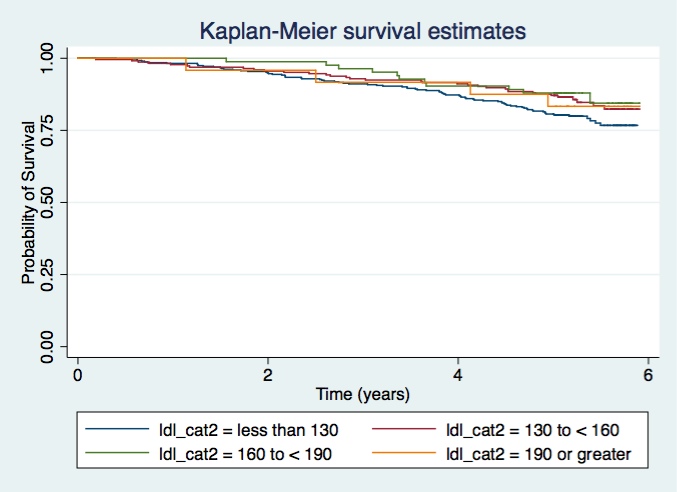
**The result of the proportional hazard regression model for the non-centered analysis yielded a slope coefficient of 0.9926 (95% CI: 0.9871, 0.9982). From the proportional hazards regression analysis we estimate that for each 1 mg/dL unit difference in serum LDL, the risk of death is 0.74% lower in the group with the higher serum LDL. This estimate is statistically significant (two-tailed, P=0.006); hence we reject the null hypothesis of no association. A 95% CI suggests that this observation is not unusual if a group that has 1 mg/dL higher serum LDL might have risk of death that was anywhere from 0.18% lower to 1.29% lower than the group with the lower serum LDL.**

**In the re-centered analysis, the results of the proportional hazard model yielded an HR of 0.9926; 95% CI: 0.9871, 0.9982. This is the exact same result as the non-centered proportional model. The only difference is in the interpretation of the intercept. We same the same similarities between models when the non-robust method was used where the HR=0.9926; 95% CI: 0.9874, 0.9979.**

**Table 1. Descriptive analysis among LDL groups.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | LDL<130 | 130 </= LDL <160 | 160 </= LDL < 190 | LDL >/=190 | Missing\* |
| N | 393 | 225 | 83 | 24 | 10 |
| Variables | Mean (SD), range | Mean (SD), range | Mean (SD), range | Mean (SD), range | Mean (SD), range |
| Age, year | 74.70 (5.35), 65-92 | 74.20 (5.62), 67-99 | 74.57 (5.67), 65-94 | 75.96 (6.11), 67-87 | 74.40 (6.13), 68-86 |
| Weight, lbs | 159.91 (29.93), 86-264 | 158.38 (32.27), 96-245 | 165.13 (32.94), 74-257 | 154.46 (19.42), 126-200 | 166.80 (29.07), 130-213 |
| Height, inches | 166.98 (9.68), 114-189.5 | 164.57 (10.01), 140.5-190.5 | 164.67 (8.96), 139-183 | 160.43 (7.22), 150.9-171.5 | 166.84 (7.82), 154.6-177.8 |
| LDL | 101.25 (19.30), 11-129 | 142.73 (8.53), 130-159 | 172.28 (9.21), 160-189 | 208.33 (13.48), 191-247 | --- |
| Physical activity (1000 kcal) | 2.04 (2.17), 0-13.81 | 1.67 (1.61), 0-9.84 | 2.29 (2.52), 0-13.04 | 1.59 (2.00), 0-7.43 | 0.77 (0.65), 0-1.82 |
|  | N (%) | N (%) | N (%) | N (%) | N (%) |
| Male | 218 (55.47) | 97 (43.11) | 40 (48.19) | 5 (20.83) | 6 (60.00) |
| Race |  |  |  |  |  |
| White | 309 (78.63) | 173 (76.89) | 64 (77.11) | 18 (75.00) | 8 (80.00) |
| Black | 56 (14.25) | 31 (13.78) | 12 (14.46) | 4 (16.67) | 1 (10.00) |
| Asian | 22 (5.60) | 18 (8.00) | 4 (4.82) | 2 (8.33) | 1 (10.00) |
| Other | 6 (1.53) | 3 (1.33) | 3 (3.61) | 0 (0.00) | 0 (0.00) |
| History of smoking | 217 (55.22) | 125 (55.56) | 47 (56.63) | 10 (41.67) | 6 (60.00) |
| History of Alcohol use | 195 (49.62) | 113 (50.22) | 42 (50.60) | 15 (62.50) | 4 (40.00) |
| General Health |  |  |  |  |  |
| Excellent | 43 (10.94) | 39 (17.33) | 13 (15.66) | 2 (8.33) | 0 (0.00) |
| Very Good | 129 (32.82) | 67 (29.78) | 21 (25.30) | 11 (45.83) | 4 (40.00) |
| Good | 155 (39.44) | 89 (39.56) | 41 (49.40) | 8 (33.33) | 4 (40.00) |
| Fair | 56 (14.25) | 27 (12.00) | 8 (9.64) | 2 (8.33) | 2 (20.00) |
| Poor | 10 (2.54) | 3 (1.33) | 0 (0.00) | 1 (4.17) | 0 (0.00) |

**Figure 1. Kaplan-Meier survival curve for patients with LDL<160 and LD>/= 160.**



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:

**Method:**

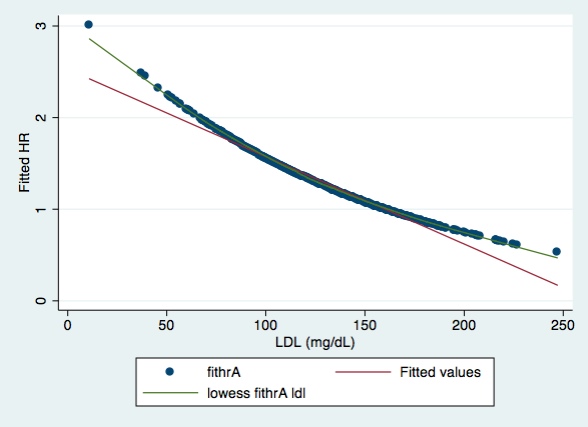
**A fitted hazard ratio was performed by re-centering LDL at 160 mg/dL. Fitted HR was plotted on the Y-axis with LDL on the X-axis.**

**Results:**

**The fitted HR was calculated as 1.3277 with a 95% CI (1.3040, 1.3514). From the proportional hazards regression analysis we estimate that for each 1 mg/dL unit difference in serum LDL, the risk of death is 32.77% higher in the group with the higher serum LDL. A 95% CI suggests that was not unusual if a group that had 1 mg/dL higher LDL had hazard anywhere from 30.4% to 35.1% higher compared to a lower LDL group.**

**We notice that the Fitted HR has a slight concave curvature that is almost U-shape at the tails.**

**Figure 2. Fitted HR for LDL re-centered at 160 mg/dL using robust SE estimates.**

****

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.

Score: 8

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:**

**Methods:**

**Proportional hazard regression was performed using death as the binary outcome and time in years as the time element. The predictor of interest is LDL level that is log transformed. The summary measure is the instantaneous risk (hazard) of death at each time among those subjects who have not died across groups defined by serum log LDL as a continuous variable. We assume that the missing data is right-censored and non-informative. Descriptive analysis was performed using Kaplan-Meier survival curves across log LDL categories (log LDL < 130, log LDL 130 to < 160, log LDL 160 to < 190, and log LDL >/= 190 mg/dL).**

**Proportional hazards assumes that the ratio of these instantaneous rates is constant in time between two groups. Parameter estimates were determined using maximum partial likelihood estimation. P-values (two-tailed) and confidence intervals were derived using Wald-based estimates. Huber-White sandwich estimator was used to estimate the standard error.**

**We also re-centered log LDL to log LDL160 mg/dL. Comparison in the proportional hazard model was done between data that was centered and not centered.**

**Results:**

**Of the 735 patients in the study, 725 had values for serum LDL. See the response from Question 1 Part a for detailed descriptive analysis.**

**The result of the proportional hazard model was a HR=0.4375; 95% CI: 0.2967, 0.6453. From the proportional hazards regression analysis that was not centered, we estimate that for each 1-fold log LDL mg/dL difference, the risk of death is 56.3% lower in the group with the higher serum log LDL. This estimate is highly statistically significant (two-tailed, P<0.0001); hence, we reject the null hypothesis of no association. A 95% CI suggests that this observation is not unusual if a group that has 1 unit higher of serum log LDL might have risk of death that was anywhere from 30.08% lower to 72.62% lower than the group with the lower serum log LDL. (instead of 1-fold log LDL difference, it better to interpret 10% increase)**

**The results of the proportion hazards model that was re-centered on 160 was: HR=0.4375; 95% CI: 0.2767, 0.6453. This was exactly the same as the above model. We got same the same results when the non-robust method was used where the HR=0.4375; 95% CI: 0.2738, 0.6992.**

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 affected 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:**

**Method:**

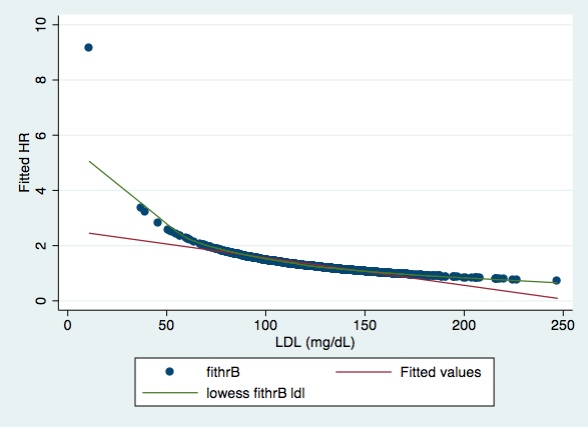
**A fitted hazard ratio was performed by re-centering log LDL at 160 mg/dL. Fitted HR was plotted on the Y-axis against the LDL on the X-axis.**

**Results:**

**The fitted HR was calculated as 1.3027 with a 95% CI (1.2709, 1.3345). From the proportional hazards regression analysis we estimate that for each 1-fold increase in log LDL, the risk of death is 30.27% higher in the group with the higher serum LDL. A 95% CI suggests that was not unusual if a group that had a 1 fold log increase in LDL was anywhere between 27.09% and 33.45% higher compared to a lower LDL group.**

**We see a slight concave curvature with the fitted HR against LDL at the extremities. More so, at the lower LDL range (Figure 3).**

**Figure 3. Fitted HR for log LDL re-centered at 160 mg/dL using robust SE estimates.**

****

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).

Score:7

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:**

**Method:**

**Proportional hazard regression was performed using death as the binary outcome and time in years as the time element. The predictor of interest is LDL level. The summary measure is the instantaneous risk (hazard) of death at each time among those subjects who have not died across groups defined by serum LDL as a continuous variable. We assume that the missing data is right-censored and non-informative. Descriptive analysis was performed using Kaplan-Meier survival curve for LDL groups were squared for the following categories: LDL < 130, LDL 130 to < 160, LDL 160 to < 190, and LDL >/= 190 mg/dL.**

**Proportional hazards assumes that the ratio of these instantaneous rates is constant in time between two groups. Parameter estimates were determined using maximum partial likelihood estimation. P-values (two-tailed) and confidence intervals were derived using Wald-based estimates. Huber-White sandwich estimator was used to estimate the standard error.**

**In the proportional hazards model, we re-centered LDL to (LDL-160 mg/dL)^2. We included the re-centered LDL and its squared version in to the model in order to create a quadratic expression. We also compared this to not centering.**

**Results:**

**Of the 735 patients in the study, 725 had values for serum LDL. Please see Question 1 Part a for the descriptive analysis.**

**The results from the proportional hazard model that was not centered was HR=0.9742; 95% CI: 0.9557, 0.9931. From the proportional hazards regression analysis estimate that for each 1 mg/dL unit difference in serum LDL, the risk of death is 2.58% lower in the group with the higher serum LDL. This estimate was statistically significant (two-tailed, P=0.008); hence we reject the null hypothesis of no association. A 95% CI suggests that this observation is not unusual if a group that has 1 mg/dL higher serum LDL might have risk of death that was anywhere from 0.069% and 4.43% lower.**

**In the centered model, the HR was 0.9983; 95% CI: 0.9906, 1.0061. From the proportional hazards model, the risk of death estimate for each 1 mg/dL unit difference in serum LDL was 0.17% lower in the group with the higher serum LDL. This estimate was not statistically significant (P=0.671, two-tailed); hence we do not have enough evidence to reject the null hypothesis of no association. A 95% CI suggests that this observation is not unusual if a group that has 1 mg/dL higher serum LDL might have a risk of death that was anywhere from 0.94% lower and 0.61% higher.**

**The slope parameter for the LDL-squared term is not statistically different from 0 (two-tailed, P=0.055); hence, we have no strong evidence for a departure from a straight-line model that can be detected by a quadratic model.**

**In the non-robust model, the HR for the non-centered LDL was HR=0.9742; 95% CI: 0.9529, 0.9960 (P=0.021). For the centered LDL model, the HR was 0.9983; 95% CI: 0.9903, 1.0064 (P=0.683, two-tailed). Similar to the robust model, the slope estimate for the LDL variable in the non-centered model was significant; however, the centered model was not statistically significant.**

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)) \* *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

**Method:**

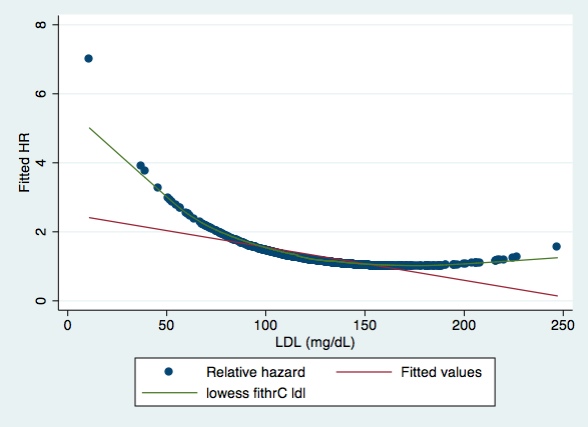
**A fitted hazard ratio was performed by re-centering log LDL at 160 mg/dL and then squaring it. Fitted HR was plotted on the Y-axis against the LDL on the X-axis.**

**Results:**

**The fitted HR was estimated as 1.3072 (95% CI: 1.2761, 1.3382). From the proportional hazards regression analysis we estimate that for each 1-fold increase in log LDL, the risk of death is 30.27% higher in the group with the higher serum LDL. A 95% CI suggests that was not unusual if a group that had a 1-fold log increase in LDL was anywhere between 27.61% to 33.82% higher compared to a lower LDL group.**

**In the fitted log HR, there is a concave curvature that almost appears U-shape at the tail ends (Figure 4).**

**Figure 4. Fitted HR for LDL + LDL^2 re-centered at 160 mg/dL using robust SE estimates.**



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.score：7

please turn to the key for more details

**ANSWER:**

**With each transformation, we begin to notice a change in the fitted HR at low LDL levels. In Figure 2, there is curve from the fitted straight-line that is noticeable. We see the same thing happening in figure 3; although this is a little more pronounced. Again, if Figure 4, we see the same concavity at the ends, but more pronounced at the lower LDL range. Figure 5 combines all three figures with the fitted HRs. The fithrC has the least linearity of the three fitted HR curves. However, most of the points could be fitted linearly. Strangely, when we run the robust model and then the non-robust model, the P-value changes from 0.089 to 0.055, which is close to significance.**

**Overall, it appears that the points fall mostly on a straight light, suggesting a linear relationship at most LDL points; with the exception of the lower LDL range.**

**Figure 5. Fitted HR for LDL, log LDL and quadratic formula for LDL re-centered at 160 using robust SE estimates.**

