**Biost 518: Applied Biostatistics II**

**Biost 515: Biostatistics II**

Emerson, Winter 2014

**Homework #3**

January 20, 2014

**Written problems:** To be submitted as a MS-Word compatible file to the class Catalyst dropbox by 9:30 am on Monday, January 27, 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 and #2, 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 5 year all-cause mortality by comparing the odds of death within 5 years across groups defined by whether the subjects have high serum LDL (“high” = LDL > 160 mg/dL). In your regression model, use an indicator of death within 5 years as your response variable, and use an indicator of high LDL as your predictor. (Only give a formal report of the inference where asked to.)
   1. Is this a saturated regression model? Explain your answer.

**In this logistic regression models, two distinct groups (those who have high serum LDL and low serum LDL, dichotomized 160 mg/dL) are modeled with two regression parameters (the intercept and the slope). It is a saturated model.**

* 1. For subjects with low LDL, what is the estimated odds of dying within 5 years? What is the estimated probability of dying within 5 years? How do these estimates compare to the observed proportion of subjects with low LDL dying within 5 years?

**For logistic regression, an indicator of death within 5 years is used as the response variable, and an indicator of high LDL as the predictor. Then, the odds of estimated odds of dying within 5 years for the subjected with low LDL is the exponentiation of the intercept (-1.586315) from the logistic regression: 0.205. The estimated probability of dying within 5 years for the subjected with low LDL is calculated by the equation- odds/(1+odds)= 0.205/1.205=0.170.**

**To calculate the sample proportion for the probability and odds, a two by two table is constructed as the following table. For subject with low LDL, the odds of dying within 5 years is calculated as: 105/513=0.205. The probability of dying within 5 years is 105/618=0.170. These two proportions are the same as the estimated proportions done by logistic regression.**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Death within 5 years | Survival after 5 years | Total |
| LDL<160mg/dL | 105 | 513 | 618 |
| LDL>=160mg/dL | 14 | 93 | 107 |
| Total | 119 | 606 | 725 |

* 1. For subjects with high LDL, what is the estimated odds of dying within 5 years? What is the estimated probability of dying within 5 years? How do these estimates compare to the observed proportion of subjects with high LDL dying within 5 years?

**For logistic regression, an indicator of death within 5 years is used as the response variable, and an indicator of high LDL as the predictor. Then, the odds of estimated odds of dying within 5 years for the subjected with high LDL is the exponentiation of the intercept (-1.586315) times the exponentiation of the slope (-0.3072267): 0.151. The estimated probability of dying within 5 years for the subjected with high LDL is calculated by the equation- odds/(1+odds)= 0.151/1.151=0.131.**

**To calculate the sample proportion for the probability and odds, a two by two table is constructed as the following table. For subject with high LDL, the odds of dying within 5 years is calculated as: 14/93=0.151. The probability of dying within 5 years is 14/107=0.131. These two proportions are the same as the estimated proportions done by logistic regression.**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Death within 5 years | Survival after 5 years | Total |
| LDL<160mg/dL | 105 | 513 | 618 |
| LDL>=160mg/dL | 14 | 93 | 107 |
| Total | 119 | 606 | 725 |

* 1. Give full inference regarding the association between 5 year mortality and high LDL levels. How does this differ from the inference that was made on problems 5 and 6 of homework #1? What is the source of any differences?

**From logistic regression analysis, of the 618 subjects whose serum LDL was less than or equal to 159 mg/dL, the odds of dying within 5 years from study enrollment was 0.205 (the probability is 0.170), while for the 107 subjects with serum LDL greater than or equal to 160 mg/dL the odds of 5 year mortality was 0.151 (the probability is 0.131). Based on a 95% confidence interval, this observed odds ratio of 0.735 for the comparison of the high LDL group to the low LDL group would not be judged unusual if the true odds ratio were anywhere between 0.404 to 1.340. The two-sided p value of 0.316 suggests that we cannot with high confidence reject the null hypothesis that the odds of death within 5 years are not associated with serum LDL levels.**

**The point estimates for odds and probability in the regression analysis are the same to those in problem 5 and 6 in homework 1. However, the p value and 95 confidence interval are different from those in problem 6. The difference comes from logistic regression basing on Wald test; and in homework 1 the p-value is 0.396 and 95% CI is 0.373 to 1.36, which is based on Fisher exact test.**

* 1. How would the answers to parts a-c change if I had instead asked you to fit a logistic regression model using the indicator of death within 5 years as your response variable, but using an indicator of low LDL as your predictor? What if we had used an indicator of survival for at least 5 years as the response variable?

**Because the new model is just the reparameterizations of the old one used in parts a-c, the answers will not be changed. (The LDL is dichotomized into low and high group. It is a binary variable.)If the indicator of survival for at least 5 years is used as the response variable, the answers will not be changed too. (For it is a binary variable. The inverse of odds of surviving after 5 years= odds of dying within 5 years)**

* 1. In parts a-d of this problem, we described the distribution of death within 5 years across groups defined by LDL level. What if we fit a logistic regression model mimicking the approach used in problems 1 – 4 of homework #2, where we described the distribution of LDL across groups defined by vital status? How would our answers to parts a-c change?

**The estimated odd ratio, it’s 95% CI, and p-value won’t be changed because it can be calculated in either direction. However, the exponentiation of intercept will be for the odds for having low or high LDL for the subjects defined by vital status. And we cannot calculated the odds and probability of dying within 5 years for high and low LDL groups. Instead, what we can calculate is the odds and probability of having high or low LDL groups given dying within 5 years or surviving after 5 years.**

1. Perform a statistical regression analysis evaluating an association between serum LDL and 5 year all-cause mortality by comparing the differences in the probability of death within 5 years across groups defined by whether the subjects have high serum LDL (“high” = LDL > 160 mg/dL). In your regression model, use an indicator of death within 5 years as your response variable, and use an indicator of high LDL as your predictor. (Only give a formal report of the inference where asked to.)
   1. Is this a saturated regression model? Explain your answer.

**In this linear regression models, two distinct groups (those who have high serum LDL and low serum LDL, dichotomized 160 mg/dL) are modeled with two regression parameters (the intercept and the slope). It is saturated model.**

* 1. For subjects with low LDL, what is the estimated probability of dying within 5 years? What is the estimated odds of dying within 5 years? How do these estimates compare to the observed proportion of subjects with low LDL dying within 5 years?

**For linear regression, an indicator of death within 5 years is used as the response variable, and an indicator of high LDL as the predictor. Then, the estimated probability of dying within 5 years for the subjected with low LDL is the intercept from the linear regression: 0.170. The estimated odds of dying within 5 years for the subjected with low LDL is calculated by the equation- probability/(1-probability)= 0.170/0.830 =0.205.**

**To calculate the sample proportion for the probability and odds, a two by two table is constructed as the following table. For subject with low LDL, the odds of dying within 5 years is calculated as: 105/513=0.205. The probability of dying within 5 years is 105/618=0.170. These two proportions are the same as the estimated proportions done by linear regression.**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Death within 5 years | Survival after 5 years | Total |
| LDL<160mg/dL | 105 | 513 | 618 |
| LDL>=160mg/dL | 14 | 93 | 107 |
| Total | 119 | 606 | 725 |

* 1. For subjects with high LDL, what is the estimated probability of dying within 5 years? What is the estimated odds of dying within 5 years? How do these estimates compare to the observed proportion of subjects with high LDL dying within 5 years?

**For linear regression, an indicator of death within 5 years is used as the response variable, and an indicator of high LDL as the predictor. Then, estimated probability of dying within 5 years for the subjected with high LDL is the intercept plus the slope (0.1699029-0.0390618): 0.131. The estimated odds of dying within 5 years for the subjected with high LDL is calculated by the equation- probability/(1-probability)= 0.131/1.131=0.151.**

**To calculate the sample proportion for the probability and odds, a two by two table is constructed as the following table. For subject with high LDL, the odds of dying within 5 years is calculated as: 14/93=0.151. The probability of dying within 5 years is 14/107=0.131. These two proportions are the same as the estimated proportions done by linear regression.**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Death within 5 years | Survival after 5 years | Total |
| LDL<160mg/dL | 105 | 513 | 618 |
| LDL>=160mg/dL | 14 | 93 | 107 |
| Total | 119 | 606 | 725 |

* 1. Give full inference regarding the association between 5 year mortality and high LDL levels. How does this differ from the inference that was made on problems 5 and 6 of homework #1? What is the source of any differences?

**From linear regression analysis, of the 618 subjects whose serum LDL was less than or equal to 159 mg/dL, the probability of dying within 5 years from study enrollment was 0. 170 (the odds is 0. 205), while for the 107 subjects with serum LDL greater than or equal to 160 mg/dL the probability of 5 year mortality was 0. 131 (the odds is 0.151).**

**The probability of dying within 5 years from study enrollment among subjects with high LDL is estimated to be 3.9% lower than in subjects with low LDL levels. The estimated robust SE is 0.0360, which leads to a 95% CI for the true difference in probability of death within 5 years from lower 10.97 to higher 3.16% in subjects with high LDL compared to those with low LDL. Based on a two-sided p value of 0.278, we cannot with high confidence reject the null hypothesis that the probability of death within 5 year are not associated with serum LDL levels.**

**The point estimates for odds and probability and the difference of probability in the linear analysis are the same to those in problem 5 and 6 in homework 1 (probability is the same concept to proportion). However, the p value and 95 confidence interval are different from those in problem 5. The difference comes from linear regression basing on Wald test; and in homework 1 the p-value is 0.314 and 95% CI is -10.9% to 3.14%, which is based on chi squared test.**

* 1. How would the answers to parts a-c change if I had instead asked you to fit a regression model using the indicator of death within 5 years as your response variable, but using an indicator of low LDL as your predictor? What if we had used an indicator of survival for at least 5 years as the response variable?

**Because the new model is just the reparameterizations of the old one used in parts a-c, the answers will not be changed. (The LDL is dichotomized into low and high group. It is a binary variable.) If the indicator of survival for at least 5 years is used as the response variable, the answers will not be changed too. (For it is also a binary variable. 1- probability of surviving after 5 years= probability of dying within 5 years)**

* 1. In parts a-d of this problem, we described the distribution of death within 5 years across groups defined by LDL level. What if we fit a regression model mimicking the approach used in problems 1 – 4 of homework #2, where we described the distribution of LDL across groups defined by vital status? How would our answers to parts a-c change?

**The p-value is only the one that won’t be changed. If we take on the new model, we cannot answer the previous questions. The new intercept and slope will be the probability and the difference of probability (and its 95% CI) between having high or low LDL groups given dying within 5 years or surviving after 5 years.**

1. Perform a statistical regression analysis evaluating an association between serum LDL and 5 year all-cause mortality by comparing the ratios of the probability of death within 5 years across groups defined by whether the subjects have high serum LDL (“high” = LDL > 160 mg/dL). In your regression model, use an indicator of death within 5 years as your response variable, and use an indicator of high LDL as your predictor. (Only give a formal report of the inference where asked to.)
   1. Is this a saturated regression model? Explain your answer.

**In this Poisson regression models, two distinct groups (those who have high serum LDL and low serum LDL, dichotomized 160 mg/dL) are modeled with two regression parameters (the intercept and the slope). It is a saturated model.**

* 1. For subjects with low LDL, what is the estimated probability of dying within 5 years? What is the estimated odds of dying within 5 years? How do these estimates compare to the observed proportion of subjects with low LDL dying within 5 years?

**For Poisson regression, an indicator of death within 5 years is used as the response variable, and an indicator of high LDL as the predictor. Then, the estimated probability of dying within 5 years for the subjected with low LDL is the exponentiation of intercept (-1.772528) from the Poisson regression: 0.170. The estimated odds of dying within 5 years for the subjected with low LDL is calculated by the equation- probability/(1-probability)= 0.170/0.830 =0.205.**

**To calculate the sample proportion for the probability and odds, a two by two table is constructed as the following table. For subject with low LDL, the odds of dying within 5 years is calculated as: 105/513=0.205. The probability of dying within 5 years is 105/618=0.170. These two proportions are the same as the estimated proportions done by Poisson regression.**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Death within 5 years | Survival after 5 years | Total |
| LDL<160mg/dL | 105 | 513 | 618 |
| LDL>=160mg/dL | 14 | 93 | 107 |
| Total | 119 | 606 | 725 |

* 1. For subjects with high LDL, what is the estimated probability of dying within 5 years? What is the estimated odds of dying within 5 years? How do these estimates compare to the observed proportion of subjects with high LDL dying within 5 years?

**For Poisson regression, an indicator of death within 5 years is used as the response variable, and an indicator of high LDL as the predictor. Then, the estimated probability of dying within 5 years for the subjected with high LDL is the exponentiation of the intercept (-1.772528) times the exponentiation of the slope (-0.2612434): 0.131. The estimated odds of dying within 5 years for the subjected with high LDL is calculated by the equation- probability/(1-probability)= 0.131/1.131=0.151.**

**To calculate the sample proportion for the probability and odds, a two by two table is constructed as the following table. For subject with high LDL, the odds of dying within 5 years is calculated as: 14/93=0.151. The probability of dying within 5 years is 14/107=0.131. These two proportions are the same as the estimated proportions done by Poisson regression.**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Death within 5 years | Survival after 5 years | Total |
| LDL<160mg/dL | 105 | 513 | 618 |
| LDL>=160mg/dL | 14 | 93 | 107 |
| Total | 119 | 606 | 725 |

* 1. Give full inference regarding the association between 5 year mortality and high LDL levels. How does this differ from the inference that was made on problems 5 and 6 of homework #1? What is the source of any differences?

**From Poisson regression analysis, of the 618 subjects whose serum LDL was less than or equal to 159 mg/dL, the probability of dying within 5 years from study enrollment was 0. 170 (the odds is 0. 205), while for the 107 subjects with serum LDL greater than or equal to 160 mg/dL the probability of 5 year mortality was 0. 131 (the odds is 0.151).**

**Based on a 95% confidence interval, this observed ratio of probability of 0.770 for the comparison of the high LDL group to the low LDL group would not be judged unusual if the true ratio of probability were anywhere between 0.458 to 1.294. The two-sided p value of 0.324 suggests that we cannot with high confidence reject the null hypothesis that the probability of death within 5 years are not associated with serum LDL levels.**

**The point estimates for odds and probability in the regression analysis are the same to those in problem 5 and 6 in homework 1. However, the figure of ratio of probability and its 95% CI is different from the figure of odds ratio and the associated 95% CI calculated in homework problem 6. Besides, the p value is different from those in problem 5 and 6. The difference comes from Poisson regression basing on Wald test; and in homework 1 the p-value is is based on Fisher exact test and chi squared test.**

* 1. How would the answers to parts a-c change if I had instead asked you to fit a regression model using the indicator of death within 5 years as your response variable, but using an indicator of low LDL as your predictor? What if we had used an indicator of survival for at least 5 years as the response variable?

**Because the new model is just the reparameterizations of the old one used in parts a-c, the answers will not be changed. (The LDL is dichotomized into low and high group. It is a binary variable.) If the indicator of survival for at least 5 years is used as the response variable, the answers will not be changed too. (For it is a binary variable. 1- probability of surviving after 5 years= probability of dying within 5 years)**

* 1. In parts a-d of this problem, we described the distribution of death within 5 years across groups defined by LDL level. What if we fit a regression model mimicking the approach used in problems 1 – 4 of homework #2, where we described the distribution of LDL across groups defined by vital status? How would our answers to parts a-c change?

**If we take on the new model, we cannot answer the question a-c. The new intercept and slope will be the probability and the ratio of probability (and its 95% CI) between having high or low LDL groups given dying within 5 years or surviving after 5 years.**

1. Perform a regression analysis of the distribution of death within 5 years across groups defined by the continuous measure of LDL. (In all cases we want formal inference.)
   1. Evaluate associations between 5 year mortality and LDL using risk difference (RD: difference in probabilities).

* ***Methods:* The probabilities of subjects dying within 5 years of study enrollment were compared across groups defined by the continuous LDL values. Differences in the probabilities were tested using a linear regression with robust method. 95% confidence intervals for the difference in population probabilities were based on Wald statistics computed using that same treatment of variances.**
* ***Results:* From linear regression method, 725 available observations from a sample of 735 elderly subjects between ages 65 and 99, we estimate that for each mg/dL difference in LDL value, the difference of probability dying within 5 years from study enrollment is 0.0010, with higher LDL group tending to have lower probability of dying within 5 years. The estimated robust SE is 0.00043. A 95% CI suggests that this observation is not unusual if the true difference in probability of death within 5 years is between 0.0019 lower per mg/dL difference in LDL level to 0.0002 lower per mg/dL difference in LDL level. This observation is statistically significant at a 0.05 level of significance (two-sided p value of 0.017), and we can with high confidence reject the null hypothesis that the probability of death within 5 year is not associated with serum LDL levels.**
  1. Evaluate associations between 5 year mortality and LDL using risk ratio (RR: ratios of probabilities).
* ***Methods:* The probabilities of subjects dying within 5 years of study enrollment were compared across groups defined by the continuous LDL values. Ratio of the probabilities were tested using a Poisson regression with robust method. 95% confidence intervals for the ratio of probabilities in population were based on Wald statistic computed using that same treatment of variances.**
* ***Results:* From Poisson regression method, 725 available observations from a sample of 735 elderly subjects between ages 65 and 99, we estimate that for each mg/dL difference in LDL value, the ratio of probability dying within 5 years from study enrollment of a higher LDL group compared to a lower LDL group is 0.9936. A 95% CI suggests that this observation is not unusual if the true ratio of probability of death within 5 years per each mg/dL difference in LDL value when a higher LDL group is compared to a lower LDL group is between 0.9883 and 0.9989. This observation is statistically significant at a 0.05 level of significance (two-sided p value of 0.018), and we can with high confidence reject the null hypothesis that the probability of death within 5 year is not associated with serum LDL levels.**
  1. Evaluate associations between 5 year mortality and LDL using odds ratio (OR: ratios of odds)
* ***Methods:* The odds of subjects dying within 5 years of study enrollment were compared across groups defined by the continuous LDL values. Odds ratio was tested using a logistic regression with robust method. 95% confidence intervals for the odds ratio in population were based the Wald statistic calculated using that same treatment of variances.**
* ***Results:* From logistic regression method, 725 available observations from a sample of 735 elderly subjects between ages 65 and 99, we estimate that for each mg/dL difference in LDL value, the odds ratio of dying within 5 years from study enrollment of a higher LDL group compared to a lower LDL group is 0.9923. A 95% CI suggests that this observation is not unusual if the true odds ratio of death within 5 years per each mg/dL difference in LDL value when a higher LDL group is compared to a lower LDL group is between 0.9858 and 0.9987. This observation is statistically significant at a 0.05 level of significance (two-sided p value of 0.019), and we can with high confidence reject the null hypothesis that the odds of death within 5 year is not associated with serum LDL levels.**
  1. How do your conclusions about such an association from this model compare to your conclusions reached in problems 1-3 of this homework and problems 2 and 4 of homework #2? Which analyses would you prefer *a priori*.?

**I would choose the analysis that was used in problems 4 of homework #2, which is performing statistical analyses evaluating an association between serum LDL and 5 year all-cause mortality by comparing mean LDL values across groups defined by vital status at 5 years using a linear regression model that allows for the possibility of unequal variances across group. Because of the following reasons:**

**1) the analyses for problem 1-3 of this homework dichotomize a continuous measurement of LDL. This will lead to loss of information and be less precise.**

**2) the concepts of comparisons of risk difference, risk ratio and odds ratio are much more difficult to understand compared to the simpler comparison of means.**

**Discussion Sections: January 22 – 24, 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.