**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.
* Yes. Because no matter in high or low LDL group, there are subjects dying within 5 years, and the estimated odds of dying within 5 years for high or low LDL groups will equal to the sample odds. (see b. & c.)

|  |  |  |  |
| --- | --- | --- | --- |
|  | LDL $\geq $ 160 | LDL < 160 | Total |
| Died within 5 years | 16 | 105 | 121 |
| Survived at least 5 years | 101 | 513 | 614 |
| Total | 117 | 618 | 735 |

* 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?
* We use the logistic regression to get the model: $log odds(dying|LDL)=-1.59-0.26×LDL$ (use STATA command: logit)
* From the model, the odds of dying within 5 years for those who with low LDL is $e^{-1.59}$ = 0.2; the probability of dying within 5 years is 0.2/(1+0.2) = 0.17. With low LDL, the proportion of dying within 5 years is 105/618=0.17. The results are the same.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | $$β$$ | robust SE | z | P- value | 95% CI |
| LDL | -0.26 | 0.29 | -0.88 | 0.377 | (-0.82, 0.31) |
| Intercept | -1.59 | 0.11 | -14.80 | 0 | (-1.80, -1.38) |

* 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 low LDL dying within 5 years?
* The odds of dying within 5 years for those who with high LDL is $e^{-1.59-0.26}$ = 0.16; the probability of dying within 5 years is 0.16 / (1+0.16) = 0.14. With high LDL, the proportion of dying 5 years is 16/117=0.14. The results are the same.
	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?
* The odd of dying 5 year for those who has higher LDL is 23% ($e^{-0.26}=0.77)$ lower than the other group, and this is no statistically significant (P-value = 0.377 > alpha = 0.05). With 95 % CI, we are not surprised if the true odd of dying within 5 year for whom has higher LDL is 56% lower or 36% higher than those subjects with lower LDL.
* Recall from hw#1, problem 5 & 6: the probability is 3.91% lower for subjects who with higher LDL; the odds ratio is 0.735 for comparing higher LDL to lower LDL, 95% CI: 0.373 to 1.36, and p-value is 0.396. The results are not exactly the same because the method is different.
	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?
* Low LDL=1
* The odds ratio became 1.292; that means the odds of dying within 5 year is 1.29 times when compared lower LDL to higher LDL, and it is also the reciprocal of 0.77.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | $$β$$ | robust SE | z | P- value | 95% CI |
| LDL | 1.292 | 0.374 | 0.88 | 0.377 | (0.732, 2.280) |
| Intercept | 0.158 | 0.043 | -6.84 | 0 | (0.093, 0.269) |

* Survived at least 5 year =1
* Compare to the original model, both intercept and slope are changed. The odds ratio is 1.292 that means the odds of survival at least 5 years for those whom with higher LDL is 1.292 times than those subjects with lower LDL.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | $$β$$ | robust SE | z | P- value | 95% CI |
| LDL | 1.292 | 0.374 | 0.88 | 0.377 | (0.732, 2.280) |
| Intercept | 4.886 | 0.524 | 14.8 | 0 | (3.960, 6.028) |

* 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?
* We let LDL as our response variable and survival status as the predictor. We can know the subject who dying within5 year is $e^{-0.256} =0.77$ times more likely to have higher LDL. With 95% confidence, it is not unusual if the true odds ratio is from 0.43 to 1.37.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | $$β$$ | robust SE | z | P- value | 95% CI |
| Survival status | -0.256 | 0.290 | -0.88 | 0.377 | (-0.824, 0.312) |
| Intercept | -1.625 | 0.109 | -14.92 | 0 | (-1.839, -1.412) |

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.
* Yes. Because the number of groups (high or low LDL) is as same as the number of parameters
	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?
* We use linear regression to get model: $E(dying|LDL)=0.17-0.033×LDL$
* The probability of died within 5 years for who have low LDL is 0.17. The odds of that is 0.17/(1-0.17) = 0.20. The results are the same.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | $$β$$ | Robust SE | t | P- value | 95% CI |
| LDL | -0.033 | 0.035 | -0.94 | 0.347 | (-0.102, 0.040) |
| Intercept | 0.170 | 0.015 | 11.23 | 0 | (0.140, 0.200) |

* 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 low LDL dying within 5 years?
* With high LDL, the probability of dying within 5 years is 0.17-0.033 = 0.137. The odds of that is 0.158. To round off, the results are the same.
	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?
		+ The risk difference of dying within 5 year between high LDL and low LDL is 0.033. The subjects with higher LDL tend to have lower probability of dying within 5 year by 0.03. There is no statistically significant (P-value = 0.347 > alpha =0.05). With 95 % CI, it is not unusual if the risk difference from 0.102 lower to 0.04 higher in high LDL level than low LDL level.
		+ As mentioned in Q1-d: the probability is 3.91% lower for subjects whom with higher LDL in problem 5 & 6 in HW#1. The result is slightly different between these two questions because of using different method to estimate in the logistic regression.
	2. 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?
		+ Low LDL = 1
		+ The absolute value of slope is the same but different sign (negative to positive) and the intercept is different. When we calculate the probability of dying within 5 year for lower LDL is 0.137+0.033 =0.17, which exactly agree with the answer of Q1-b.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | $$β$$ | Robust SE | t | P- value | 95% CI |
| LDL | 0.033 | 0.035 | 0.94 | 0.347 | (-0.036, 0.102) |
| Intercept | 0.137 | 0.032 | 4.3 | 0 | (0.074, 0.199) |

* Survived at least 5 year =1
* In this model, both intercept and slope is changed from the original model. The probability of survived at least 5 year increase by 0.033 when the LDL increase by 1 unit.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | $$β$$ | Robust SE | t | P- value | 95% CI |
| LDL | 0.033 | 0.035 | 0.94 | 0.347 | (-0.036, 0.102) |
| Intercept | 0.830 | 0.015 | 54.87 | 0 | (0.800, 0.860) |

* 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 result shows that compare to those who survival at least 5 year, the probability for subjects whom died within 5 year to have higher LDL decrease by 0.03.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | $$β$$ | Robust SE | t | P- value | 95% CI |
| Survival status | -0.0323 | 0.0343 | -0.94 | 0.347 | (-0.0996, 0.035) |
| Intercept | 0.1645 | 0.0150 | 10.98 | 0 | (0.1351, 0.1939) |

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.
* Yes. There is no any group without event. From the following results, we can know the estimated odds exactly agree with the observed proportion.
	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?
* The Poisson regression model: $log rate =3.75-0.366×LDL$
* For lower LDL, the probability is $e^{-1.773}$=0.17; the odds is 0.17/0.83=0.2. The result is as same as the observed proportion.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | $$β$$ | robust SE | z | P- value | 95% CI |
| LDL | -0.217 | 0.249 | -0.87 | 0.383 | (-0.705, 0.271) |
| Intercept | -1.773 | 0.089 | -19.92 | 0 | (-1.947, -1.598) |

* 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 low LDL dying within 5 years?
* The probability of dying within 5 year for those with high LDL, $e^{-1.773-0.217}=0.14$; the odds is 0.14/ 0.86 = 0.16. The results are the same.
	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 this model, we can know the risk of dying within 5 year can decrease by 20% ($e^{-0.217}$=0.8) for those who have higher LDL. However, there is no statistically significant (p-value=0.383> alpha=0.05). With 95% CI, we are not surprised if the true risk ratio is from 0.49 to 1.31. That is 51% lower to 31% higher for higher LDL to be dying within 5 year.
* HW#1 is about odds ratio and risk different, not risk ratio.
	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?
* Low LDL= 1
* The result shows that low LDL is $e^{0.217}=1.24 $times more likely to die within 5 year. The reciprocal of 1.24 = 0.8; the interpretation of this result is as same as Q3.d

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | $$β$$ | robust SE | z | P- value | 95% CI |
| LDL | 0.217 | 0.249 | 0.217 | 0.249 | (-0.271, 0.705) |
| Intercept | -1.990 | 0.232 | -1.990 | 0.232 | (-2.445, -1.534) |

* Survived at least 5 year =1
* With high LDL level, the rate is $e^{0.039}=1.04$ to survival at least 5 year.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | $$β$$ | robust SE | z | P- value | 95% CI |
| LDL | 0.039 | 0.041 | 0.95 | 0.34 | (-0.041, 0.120) |
| Intercept | -0.186 | 0.018 | -10.23 | 0 | (-0.222, -0.151) |

* 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?
* We know the risk for subjects who died within 5 years is $e^{-0.218}=0.80$ times to have higher LDL than those survived at least 5 year. The result is as same as the previous analysis.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | $$β$$ | Robust SE | z | P- value | 95% CI |
| Survival status | -0.218 | 0.250 | -0.87 | 0.383 | (-0.709, 0.272) |
| Intercept | -1.805 | 0.091 | -19.83 | 0 | (-1.983, -1.626) |

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).
* To find out the association between 5 year mortality and the continuous LDL, we use linear regression to see the risk different.

Model: $E(dying|LDL) = 0.29426-0.001034×LDL$

* The risk different per unit of LDL is -0.001034. That is, when LDL increases by 1 unit, the risk probability decreases by 0.001034. The p-value =0.017 < alpha =0.05, the beta of LDL has statistics significant. With 95% confidence, it is not unusual if the true probability is lower between 0.000185 and 0.001884 when LDL increases by 1 unit.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | $$β$$ | robust SE | t | P- value | 95% CI |
| LDL | -0.001034 | 0.000433 | -2.39 | 0.017 | (-0.001884, -0.000185)  |
| Intercept | 0.294260 | 0.058031 | 5.07 | 0 | (0.180330, 0.408190) |

* 1. Evaluate associations between 5 year mortality and LDL using risk ratio (RR: ratios of probabilities).
* By using Poisson regression with robust SE, we get the model:$ log rate =-1.01637-0.00647×LDL\_{i}$.
* The risk of mortality will decrease by 0.66% ( $1-e^{-0.00647}=1-0.9934=0.0066)$, which is statistics significant (p-value =0.012 < alpha =0.05). With 95% confidence, it is not unusual if the true risk of mortality decreases by 0.11% to 1.18%.

$( 1-e^{-0.00112}=1-0.9989=0.0011 and 1-e^{-0.01182}=1-0.9882=0.0118$ )

* If used 10 base, the risk of mortality will decrease by 1.5% ($1-10^{-0.0067}=1-0.985=0.015)$, and 95 % CI suggested that we are not surprised if the true risk ratio is between 0.26% and 2.7%.

 $( 1-10^{-0.00112}=1-0.9974=0.0026 and 1-10^{-0.01182}=1-0.973=0.027$ )

* logit

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | $$β$$ | robust SE | z | P- value | 95% CI |
| LDL | -0.00647 | 0.00273 | -2.37 | 0.018 | (-0.01182, -0.00112)  |
| Intercept | -1.01637 | 0.32954 | -3.08 | 0.002 | (-1.66226, -0.37049) |

* Logistic

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | $$β$$ | robust SE | z | P- value | 95% CI |
| LDL | 0.9936 | 0.0027 | -2.37 | 0.018 | (0.988254, 0.998878 ) |
| Intercept | 0.3619 | 0.1193 | -3.08 | 0.002 | (0.189710, 0.690398) |

* 1. Evaluate associations between 5 year mortality and LDL using odds ratio (OR: ratios of odds)
* We calculate odds ratio by using logistic regression. From the result, we know the odds of dying within 5 years is 1% lower when LDL increases per unit. The p-value is 0.019 < alpha = 0.05, has statistically significant. With 95% confidence, we are not surprised if the true probability is between1% and 1.5% lower when LDL increases 1 unit.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | $$β$$ | robust SE | z | P- value | 95% CI |
| LDL | 0.9923 | 0.0033 | -2.34 | 0.019 | (0.9858, 0.999) |
| Intercept | 0.5106 | 0.2094 | -1.64 | 0.101 | (0.2286, 1.1405) |

* 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*.?
* From the previous analysis, we know the subjects with higher LDL level are less likely to die within 5 year; Compared to low LDL level, the odds ratio is 0.77, the risk different is -0.033 and the risk ratio is 0.8; however, there is no statistical significant. I prefer to use Poisson regression because the question is interested in mortality rate.

**Discussion Sections: January 22 – 14, 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.