**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.
		1. Two distinct groups (those who died within 5 years and those who survived at least 5 years) are modeled with two regression parameters (the intercept and the slope). Hence, the model is saturated.
	2. 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?
		1. Consider the following logistic regression model which will provide estimates: . For subjects with low LDL, the estimated odds of dying within 5 years is . Again, for subjects with low LDL, the estimated probability of dying within 5 years is . The observed proportion of subjects with low LDL dying within 5 years is 17.0%, which is equal to the aforementioned estimate of probability, but completely dissimilar to the estimated odds.
	3. 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?
		1. For subjects with high LDL, the estimated odds of dying within 5 years is . Again, for subjects with high LDL, the estimated probability of dying within 5 years is calculated as . The observed proportion of subjects with high LDL dying within 5 years is 13.1%, which is exactly equal to the aforementioned estimate of probability, but completely dissimilar to the estimated odds.
	4. 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?
		1. From a logistic regression analysis of 725 available observations from a sample of 735 subjects, we estimate the odds ratio comparing the odds of dying within 5 years between high LDL and low LDL subjects 0.735, though this estimate is not statistically significant (P=0.315). That high p-value suggests that we cannot with high confidence reject the null hypothesis that 5 year mortality is not associated with high serum LDL levels. Based on a 95% confidence interval, this observed odds ratio is not unusual if the true odds ratio of death within 5 years between high LDL and low LDL were anywhere between 0.404 and 1.340.
		2. Problem 5 of Homework #1 was concerned with an inference about the proportion surviving for 5 years. In problem 5’s results, estimated proportions were reported. These same proportions can be used to estimate each of the odds for dying within 5 years for high LDL and low LDL. Hence, the odds ratio reported in problem 1d of this homework could be estimated. Next, the inference just reported and the inference for problem 5 have similar p-values, 0.315 and 0.314, respectively. Problem 5 used the chi-squared test for independence for its p-value, while the inference in the logistic regression used a Wald test. The conclusion is the same for both problem 5 and our current problem: we cannot with high confidence reject the null hypothesis that the survival probabilities are not associated with serum LDL levels.
		3. Problem 6 of Homework #1 is quite similar to the inference reported in Problem 1di of this homework. The point estimates of interest in problem 6 are the odds of dying within 5 years, in both cases of high LDL and low LDL, and also the odds ratio. The main difference is that while the logistic regression model we used in problem 1 of this homework used Wald statistics, problem 6 used exact methods (Fisher’s exact). This change to exact methods altered the confidence interval as well as the p-value. Using the Fisher’s exact, the confidence interval for the odds ratio ranged from 0.373 to 1.36 (compare to our problem 1 CI: [0.404,1.340]). Furthermore, the p-value using the exact method is 0.396, while our problem 1 Wald based p-value is 0.315. Problem 6 could have used Wald statistics, and thus would have been the same inference as reported in 1di. The conclusion is the same for both problem 6 and our current problem: we cannot with high confidence reject the null hypothesis that the survival probabilities are not associated with serum LDL levels.
	5. 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?
		1. If a logistic regression model using the indicator of death within 5 years as a response variable and indicator of low LDL level as a predictor was fit, then the answers for parts a-c would not change at all. The model would still be saturated. The values of the model coefficients would change, but would result in the same estimates of odds and probabilities. As for changing the indicator for the response, again, the same answers attained in parts a-c could be replicated, but the calculations gain a step. Our models have as coefficients log-odds values, so to create odds we exponentiate. When switching the indicator in the response, the same first step applies, but then the inverse operation is required.
	6. 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?
		1. The regression suggested would be switching the response and predictor as they were placed to begin with in parts a-e. Hence, the model is different. Parts a-c would be unanswerable with this new model. Instead of asking what the odds of dying are for someone with low LDL, the question would have to be posed as: what is the odds of having low LDL, given that they died within five years of the study?
2. 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.
		1. Two distinct groups (those who died within 5 years and those who survived at least 5 years) are modeled with two regression parameters (the intercept and the slope). Hence, the model is saturated.
	2. 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?
		1. Consider the regression model:

E(Dying within 5 Years | High LDL)=0.17-0.039\*High LDL.

Based on a linear regression using an indicator of death within 5 years as a response variable and an indicator of high LDL as a predictor, subjects with low LDL (“High LDL” = 0) have an estimate probability of dying within 5 years of about 17%, as estimated from the regression coefficients. Hence, the odds of dying within 5 years for a subject with low LDL is estimated as 0.205. The probability estimate is equivalent to the proportion of subjects with low LDL dying within 5 years, i.e. approximately 17%.

* 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?
		1. Based on a linear regression using an indicator of death within 5 years as a response variable and an indicator of high LDL as a predictor, subjects with high LDL (“High LDL” = 1) have an estimate probability of dying within 5 years of about 13.1%. Hence, the odds of dying within 5 years for a subject with high LDL is estimated as 0.151. The probability estimate is equivalent to the proportion of subjects with low LDL dying within 5 years, i.e. approximately 13.1%.
	2. 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?
		1. From a linear regression analysis of 725 available observations from a sample of 735 subjects, we estimate the probability of dying within 5 years is 3.9% lower in the high LDL group, though this estimate is not statistically significant (P=0.315). That low p-value suggests that we cannot with high confidence reject the null hypothesis that 5 year mortality is not associated with high serum LDL levels. Based on a 95% confidence interval, this observation is not unusual if a group that has high LDL has probability of dying within 5 years that is anywhere from 11.5% lower or 3.7% higher than the low LDL group.
		2. Problem 5 of Homework #1 was concerned with an inference about the proportion surviving for 5 years. In problem 5’s results, estimated proportions were reported and absolute difference in survival probability between high LDL and low LDL subjects was estimated as 3.91%, which is the same estimate reported in the recent full inference. Next, the inference just reported and the inference for problem 5 have similar p-values, 0.315 and 0.314, respectively. Problem 5 used the chi-squared test for independence for its p-value, while the inference in the linear regression used a t-test. The conclusion is the same for both problem 5 and our current problem: we can not with high confidence reject the null hypothesis that the survival probabilities are not associated with serum LDL levels.
		3. Problem 6 of Homework #1 is different to the inference reported in Problem 2di of this homework. The point estimates of interest in problem 6 are the odds of dying within 5 years, in both cases of high LDL and low LDL, and also the odds ratio. Thus, the methods employed are going to be different. The confidence interval and p-value are completely different. The conclusion is the same for both problem 6 and our current problem: we can not with high confidence reject the null hypothesis that the survival probabilities are not associated with serum LDL levels.
	3. 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?
		1. If a linear regression model using the indicator of death within 5 years as a response variable and indicator of low LDL level as a predictor was fit, then the answers for parts a-c would not change at all. The model would still be saturated. The values of the model coefficients would change, but would result in the same estimates of odds and probabilities. In particular, where before the intercept contained the probability of dying within 5 years for low LDL, now with the new parameterization the intercept will contain the probability of dying within 5 years for subjects with high LDL. As for changing the indicator for the response, again, the same answers attained in parts a-c could be replicated, but the calculations gain a step. Our models have as coefficients probability/proportion values. When switching the indicator in the response, we continue to use the probabilities stated as coefficients, but the complement has to be taken. For example, if a linear regression model using the indicator of survival for at least 5 years as the response variable with an indicator of high LDL as the predictor, the intercept is estimated as 83%, which is the probability of surviving for at least 5 years given a subject has low LDL. For the probability of dying within 5 years for a subject with low LDL, the complement is taken: 1-0.83=17%, exactly as given above. All values in parts a-c can be calculated as such.
	4. 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?
		1. The answers for the questions asked would not change, but obtaining the same estimates with the regression model mimicking the approach used in problems 1-4 of homework #2 becomes more involved. In particular, the Bayes rule would have to be enacted. The model with switched predictor and response would give us P( Low LDL | Dying within 5 Years). However, to answer, for instance part a, we require P( Dying with 5 Years | Low LDL), which can be estimated using the following formula:

 .

 All of these values can be estimated.

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.
		1. Two distinct groups (those who died within 5 years and those who survived at least 5 years) are modeled with two regression parameters (the intercept and the slope). Hence, the model is saturated.
	2. 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?
		1. Consider the regression model:

Log(λ) = -1.77-0.26\*High LDL, where λ is the risk ratio.

Based on a Poisson regression using an indicator of death within 5 years as a response variable and an indicator of high LDL as a predictor, subjects with low LDL have an estimate probability of dying within 5 years of about . Hence, the odds of dying within 5 years for a subject with low LDL is estimated as . The probability estimate is equivalent to the proportion of subjects with low LDL dying within 5 years, i.e. approximately 17%.

* 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?
		1. Based on a Poisson regression using an indicator of death within 5 years as a response variable and an indicator of high LDL as a predictor, subjects with high LDL have an estimate probability of dying within 5 years of about . Hence, the odds of dying within 5 years for a subject with high LDL is estimated as . The probability estimate is equivalent to the proportion of subjects with low LDL dying within 5 years, i.e. approximately 13.1%.
	2. 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?
		1. From a Poisson regression analysis of 725 available observations from a sample of 735 subjects, we estimate the risk ratio of dying within 5 years comparing high LDL to low LDL is 0.77, though this estimate is not statistically significant (P=0.359). That low p-value suggests that we cannot with high confidence reject the null hypothesis that 5 year mortality is not associated with high serum LDL levels. Based on a 95% confidence interval, this observation is not unusual if the true risk ratio of dying within 5 years from high LDL to low LDL were found anywhere between 0.44 and 1.34.
		2. Problem 5 of Homework #1 was concerned with an inference about the proportion surviving for 5 years as defined by high or low LDL. In problem 5’s results, estimated proportions were reported as 13.1% for high LDL and 17% for low LDL. The ratio of these proportions is precisely the point estimate of interest in this current problem, namely 0.77. The confidence interval and p-value are quite different between the two, mainly because problem 5 is concerned with difference in probabilities, while our Poisson regression is concerned with the ratio of these same probabilities (interpreted as risks). The conclusion is the same for both problem 5 and our current problem: we can not with high confidence reject the null hypothesis that the survival probabilities are not associated with serum LDL levels.
		3. Problem 6 of Homework #1 is different to the inference reported in Problem 3di of this homework. The point estimates of interest in problem 6 are the odds of dying within 5 years, in both cases of high LDL and low LDL, and also the odds ratio (compare to our interest in the risk ratio).Thus, the methods employed are going to be different. The confidence interval and p-value are completely different. The conclusion is the same for both problem 6 and our current problem: we can not with high confidence reject the null hypothesis that the survival probabilities are not associated with serum LDL levels.
	3. 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?
		1. If a Poisson regression model using the indicator of death within 5 years as a response variable and indicator of low LDL level as a predictor was fit, then the answers for parts a-c would not change at all. The model would still be saturated. The values of the model coefficients would change, but would result in the same estimates of odds and probabilities. As for changing the indicator for the response, again, the same answers attained in parts a-c could be replicated, but the calculations become more involved.
	4. 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?
		1. The regression suggested would be switching the response and predictor as they were placed to begin with in parts a-e. Hence, the model is different. Parts a-c would be answerable with this new model, but only after a few transformations. However, a similar set of questions could be answered. Instead of asking what the odds of dying are for someone with low LDL, the question could be posed as: what is the risk ratio of having low LDL, given that they died within five years of the study?
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).
		1. Using a linear regression, the association between 5 year mortality across groups defined by the continuous measure of LDL will be evaluated. The response of the regression is designated as an indicator of subjects dying within 5 years, while the predictor is simply serum LDL (mg/dL). A 95% confidence interval around the risk difference per 1 mg/dL increase in LDL will be estimated using the standard error.
		2. From a linear regression analysis of 725 available observations from a sample of 735 subjects, we estimate the probability of dying within 5 years decreases by 0.103% per 1 mg/dL increase in LDL level. Based on a 95% confidence interval, it would not be unusual if the true difference in probability of dying within 5 years per 1 mg/dL increase of LDL were decreased anywhere between 0.0954% and 0.111%. At the 5% significance level, there is evidence to support an association between mortality at 5 years and serum LDL level (P=0.0115). However, it should be noted that when dealing with small numbers, it is quite easy to get a significant value. Hence, this association should be investigated much further.
	2. Evaluate associations between 5 year mortality and LDL using risk ratio (RR: ratios of probabilities).
		1. Using a Poisson regression, the association between 5 year mortality across groups defined by the continuous measure of LDL will be evaluated. The response of the regression is designated as an indicator of subjects dying within 5 years, while the predictor is simply serum LDL (mg/dL). A 95% confidence interval around the risk ratio per 1 mg/dL increase in LDL will be estimated using the standard error.
		2. From a Poisson regression analysis of 725 available observations from a sample of 735 subjects, we estimate the risk ratio of dying within 5 years comparing high LDL to low LDL as designated per 1 mg/dL increase in LDL is 0.993. Based on a 95% confidence interval, this observed estimate is not unusual if the true risk ratio of dying within 5 years from high LDL to low LDL as designated by a 1 mg/dL increase in LDL were found anywhere between 0.988 and 0.999. At the 5% significance level, there is evidence to support an association between mortality at 5 years and serum LDL level (P=0.02091).
	3. Evaluate associations between 5 year mortality and LDL using odds ratio (OR: ratios of odds)
		1. Using a logistic regression, the association between 5 year mortality across groups defined by the continuous measure of LDL will be evaluated. The response of the regression is designated as an indicator of subjects dying within 5 years, while the predictor is simply serum LDL (mg/dL). A 95% confidence interval around the odds ratio per 1 mg/dL increase in LDL will be estimated using the standard error.
		2. From a logistic regression analysis of 725 available observations from a sample of 735 subjects, we estimate the odds ratio of dying within 5 years comparing high LDL to low LDL as designated per 1 mg/dL increase in LDL is 0.992. Based on a 95% confidence interval, this observed estimate is not unusual if the true odds ratio comparing the odds of dying within 5 years from high LDL to low LDL as designated by a 1 mg/dL increase in LDL were found anywhere between 0.986 and 0.998. At the 5% significance level, there is evidence to support an association between mortality at 5 years and serum LDL level (P=0.0119).
	4. 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*.?
		1. Problems 1-3 each concluded that we could not with high confidence reject the null hypothesis about no association between serum LDL and mortality at 5 years. However, problems 2 and 4 of homework #2 each rejected the null hypothesis of no association between serum LDL and risk of death in favor of a hypothesis that surviving subjects would tend toward higher mean serum LDL. The main difference between the two sets of problem are the response and predictor considered. Problems 1-3 considered mortality at 5 years a response to serum LDL level (as an indicator, high or low), while problems 2 and 4 of homework #2 considered serum LDL level a response to mortality at 5 years. A priori I would have preferred to test mortality at 5 years as a response to serum LDL on the grounds of scientific relevance. Perhaps there is a useful interpretation of a cadavers LDL level, but I certainly would not weigh it more useful than the applications possible once we understand the response a living human’s mortality will have with respect to their LDL level.

**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.