**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.

This is a saturated model since the number of groups (high serum LDL group and low serum LDL group) equals the number of parameters (the intercept and the slope).

* 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?

The logistic regression model of the indicator of death within 5 years on serum LDL estimates that subjects having serum LDL lower than 160mh/dL have an odds of death within 5 years of 0.205, which leads to an estimated probability of death within 5 years of 17.0%. The observed proportion of subjects with low LDL dying within 5 years is the same as an estimated probability of death within 5 years given a subject has low serum LDL.

* 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?

The logistic regression model of the indicator of death within 5 years on serum LDL estimates that subjects having serum LDL greater than 160mh/dL have an odds of death within 5 years of 0.151, which leads to an estimated probability of death within 5 years of 13.1%. The observed proportion of subjects with high LDL dying within 5 years is the same as an estimated probability of death within 5 years given a subject has high serum LDL.

* 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?

**Methods**. We dichotomize our data according to serum LDL with high serum LDL being greater or equal to 160mg/dL and low serum LDL being less than 160mg/dL. We also split the data according to subject’s vital status at 5 years into two groups – the group of subjects that died within 5 years and the group of subjects that survived at least 5 years. We perform robust logistic regression to evaluate an association between 5 year mortality and high LDL levels. We use an indicator of death within 5 years as our response variable, and an indicator of high LDL as our predictor.

**Results.** From the estimates of logistic regression, we conclude that the subjects with high LDL have 26.5 % lower mortality than the subjects with low LDL level

(95% CI 34.1% lower to 59.6% higher). The performed logistic regression suggests that the odds ratio 95% CI includes 1 (P=0.316), therefore we fail to reject the null hypothesis and we say that there is not enough evidence to conclude anything about the association between 5 year mortality and high LDL levels.

The point estimate is the same as in homework #1 but the 95% CIs are different due to the methods being used (Fischer’s exact test in homework #1 vs. logistic regression (Wald test) in homework #3).

* 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?

In both cases the answers would be the same since we would just be reparameterizing the same model.

* 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 model would still be saturated. The odds ratio will still be the same. However, the odds would be different. The odds of having high LDL given that the subject died within 5 years would be 0.133 with probability of having high LDL given that they died within 5 years being 0.118. The odds of having high LDL given that the subject survived at least 5 years would be 0.181 with probability of having high LDL given that they survived at least 5 years being 0.153.

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.

This is a saturated model since the number of groups (high serum LDL group and low serum LDL group) equals the number of parameters (the intercept and the slope).

* 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 linear regression model (allowing for unequal variances) with the indicator of death within 5 years on serum LDL estimates that subjects having serum LDL lower than 160mg/dL have an odds of death within 5 years of 0.205, which leads to an estimated probability of death within 5 years of 17.0%. The observed proportion of subjects with low LDL dying within 5 years is the same as an estimated probability of death within 5 years given a subject has low serum LDL.

* 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 linear regression model (allowing for unequal variances) with the indicator of death within 5 years on serum LDL estimates that subjects having serum LDL greater than 160mh/dL have an odds of death within 5 years of 0.151, which leads to an estimated probability of death within 5 years of 13.1%. The observed proportion of subjects with high LDL dying within 5 years is the same as an estimated probability of death within 5 years given a subject has high serum LDL.

* 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?

**Methods.** The proportion of subjects dying within 5 years of study enrollment were compared between subjects who had high serum LDL (greater than or equal to 160mg/dL) and subjects whose serum LDL was measured to be 159mg/dL or less. We evaluate the 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. We calculate the differences using robust linear regression using an indicator of high LDL as our predictor.

**Results.** We estimate the difference in probability of death to be 3.9% with the group with higher LDL having lower mortality. Based on a 95% CI, this difference would not be unusual if the true difference in probability of dying within 5 years were anywhere between 11.0% lower to 3.15% higher in the subjects with lower LDL vs subjects with higher LDL. This observation is not statistically significant at a 0.05 level of significance (P=0.278) and we cannot with high confidence reject the null hypothesis that the survival probabilities are not associated with serum LDL levels.

The point estimate is the same as in homework #1 but the 95% CIs are different due to the methods being used (Fisher’s exact test vs Wald).

* 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?

In both cases the answers would be the same since we would just be reparameterizing the same model.

* 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 model is still saturated. However, we cannot compute the probabilities b-c from this model since it provides the information about risk difference.

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.

This is a saturated model since the number of groups (high serum LDL group and low serum LDL group) equals the number of parameters (the intercept and the slope).

* 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 presuming unequal variances with the indicator of death within 5 years on serum LDL estimates that subjects having serum LDL lower than 160mh/dL have an odds of death within 5 years of 0.205, which leads to an estimated probability of death within 5 years of 17.0%. The observed proportion of subjects with low LDL dying within 5 years is the same as an estimated probability of death within 5 years given a subject has low serum LDL.

* 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 poisson regression model presuming unequal variances with the indicator of death within 5 years on serum LDL estimates that subjects having serum LDL greater than 160mh/dL have an odds of death within 5 years of 0.151, which leads to an estimated probability of death within 5 years of 13.1%. The observed proportion of subjects with high LDL dying within 5 years is the same as an estimated probability of death within 5 years given a subject has high serum LDL.

* 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?

**Methods.** The proportion of subjects dying within 5 years of study enrollment were compared between subjects who had high serum LDL (greater than or equal to 160mg/dL) and subjects whose serum LDL was measured to be 159mg/dL or less. The ratios of the probability of death within 5 years across groups were compared using poisson regression that presumes unequal variances. 95% confidence intervals for the ratio were computed using Wald statistics.

**Results.** Of the subjects with low serum LDL, 17% were observed to die within 5 years, while 13.1% of the subjects with high serum LDL died within 5 years of enrollment. The ratio of the probability of death within 5 years is 0.770. Based on a 95% CI, this ratio would not be unusual if the true ratio of the probability of death within 5 years were anywhere between 0.458 and 1.29. This observation is not statistically significant at a 0.05 level of significance (P=0.324) and we cannot with high confidence reject the null hypothesis that the survival probabilities are not associated with serum LDL levels.

The point estimate is the same as in homework #1 but the 95% CIs are different due to the methods being used.

* 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?

The model would still be saturated. However, we cannot compute the probabilities b-c from this model since it provides the information about risk ratios.

* 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 probabilities would be different. The probability of having high LDL given that the subject died within 5 years would be 0.118 and the probability of having high LDL given that the subject survived at least 5 years would be 0.153.

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.** We dichotomize the data according to subjects’ vital status at 5 years into the groups of the ones who died within 5 years and the ones that survived at least 5 years. We evaluate the association between 5 year mortality and LDL by calculating risk difference using linear regression that presumes unequal variances. Standard errors were computed using the Huber-White sandwich estimator.

**Results.** From a linear regression analysis of 725 available observations from a sample of 735 subjects, we estimate the probability of dying to decrease by 0.103% for each increase in LDL by 1mg/dL, i.e. subjects with higher LDL having lower mortality. Based on a 95% confidence interval, we find that observing an estimated difference is not unusual if the true association between 5 year mortality and LDL were such that the difference in probabilities of dying within 5 years were anywhere between 0.018% and 0.188% lower for every 1mg/dL of LDL (with subjects with higher LDL having lower mortality). These results are statistically significant evidence of an association between 5 year mortality and LDL (P=0.017) .

* 1. Evaluate associations between 5 year mortality and LDL using risk ratio (RR: ratios of probabilities).

**Methods.** We dichotomize the data according to subjects’ vital status at 5 years into the groups of the ones who died within 5 years and the ones that survived at least 5 years. We evaluate the association between 5 year mortality and LDL by calculating risk ratio using poisson regression that presumes unequal variances.

**Results.** From a poisson regression analysis of 725 available observations from a sample of 735 subjects, we estimate that for every 1mg/dL increase in LDL the probability of dying decreases by 0.994 times (the risk ratio is 0.994), i.e. subjects with higher LDL having lower mortality. Based on a 95% confidence interval, we find that observing an estimated risk ratio is not unusual if the true association between 5 year mortality and LDL were such that the risk ratio were anywhere between 0.988and 0.999 (subjects with higher LDL having lower mortality). These results are statistically significant evidence of an association between 5 year mortality and LDL (P=0.018) .

* 1. Evaluate associations between 5 year mortality and LDL using odds ratio (OR: ratios of odds)

**Methods.** We dichotomize the data according to subjects’ vital status at 5 years into the groups of the ones who died within 5 years and the ones that survived at least 5 years. We evaluate the association between 5 year mortality and LDL by comparing the odds between the dying within 5 years across the groups with different LDL levels. We use logistic regression that allows for unequal variances to calculate the odds ratio and 95% confidence for it.

**Results.** When comparing two groups with different LDL levels, the odds of dying within 5 years is estimated to be 7.48% lower (odds ratio 0.9252) for each 10mg/dL difference in LDL, with the group having higher serum LDL tending toward a lower odds of death within 5 years. This observed difference is statistically different from an odds ration of 1 (P=0.019), with a 95% confidence interval suggestion that the observed odds ratio is what might be typically observed if the true odds ratio of dying within 5 years were anywhere between 1.25% and 13.3% lower for each 10mg/dL higher serum LDL. We reject the null hypothesis of no association between mortality and serum LDL and conclude that there is higher odds of survival among subjects with higher serum LDL.

* 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*.?

When we lose dichotomize a continuous variable, we lose information. Therefore, it is advantageous to perform regression analysis of the distribution of death within 5 years across groups defined by the continuous measure of LDL. It also makes sense to use logistic regression since odds ratio is easier to interpret.

**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.