**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, this is a saturated regression model. The predictor variable of high LDL is binary and thus has two groups (0 and 1). To perform regression with odds, we used a logistic regression model which has two parameters. Since the number of groups equals the number of parameters, this 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 subjects with low LDL, the estimated odds of dying within 5 years is 20.5% and the estimated probability of dying within 5 years is 17.0%. The observed proportion of subjects in this sample with low LDL who died within 5 years is 17.0% (105 out of 618 subjects), which is the same as the estimated probability.*

* 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 subjects with high LDL, the estimated odds of dying within 5 years is 15.9% and the estimated probability of dying within 5 years is 13.7%. The observed proportion of subjects in this sample with high LDL who died within 5 years is 13.7% (16 out of 117 subjects), which is the same as the estimated probability of dying within 5 years amongst high LDL subjects.*

* 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?

*We performed logistic regression analysis amongst 735 subjects using high LDL (LDL > 160 mg/dL) as a response variable and mortality within 5 years as a predictor variable. From the analysis, we estimate that for subjects with high LDL, the odds of death within 5 years is 22.6% lower than the odds of death within 5 years for subjects with low LDL. However, at a level of 0.05 significance, this estimate is not statistically significant (P=0.376). A 95% confidence interval suggests that this observation is not unusual if a group of subjects with high LDL had a true odds of death within 5 years that was 56.1% lower or 36.5% higher than a group with low LDL.*

*The point estimates for the logistic regression are the same as those observed in problems 6. Compared to problem 6, the p-value is slightly larger with logistic regression than with the two-sided chi-square test, though both are still not significant. The p-value is smaller for the chi-square test because there are fewer degrees of freedom when performing that test than when performing the logistic regression. Similarly, the confidence interval from problem 6 is narrower than in this problem, due to the assumption of heteroscedasticity necessary to handle the mean-variance relationships in regression.*

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?

*For death within 5 years as response variable and low LDL as predictor variable:*

* + 1. *It is still a saturated model because of two outcome categories and two parameters*
		2. *For subjects with low LDL, the odds of death within 5 years would be the same as in part b (20.5%) as would the probability of death within 5 years (17.0%).*
		3. *For subjects with high LDL, the odds of death within 5 years would be the same as in part c (15.9%), as would the probability of death within 5 years (13.7%).*

*For survival past 5 years as the response variable and high LDL as predictor variable:*

1. *It is still a saturated model because of two outcome categories and two parameters*
2. *For subjects with low LDL, the odds of survival past 5 years would be 4.88 and the probability of survival past 5 years would be 83.0%, which is the appropriately the difference of the probability of death within 5 years found in part b (17.0%).*
3. *For subjects with high LDL, the odds of survival past 5 years would be 6.31 and the probability of survival past 5 years would be 86.3%, which is the appropriately the difference of the probability of death within 5 years found in part b (13.7%).*
	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?

*For death within 5 years as a predictor variable and high LDL as response variable:*

* + 1. *It is still a saturated model because of two outcome categories and two parameters*
		2. *For subjects with low LDL, the odds of death within 5 years would be the same as in part b (20.5%) as would the probability of death within 5 years (17.0%).*
		3. *For subjects with high LDL, the odds of death within 5 years would be the same as in part c (15.9%), as would the probability of death within 5 years (13.7%).*
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, this is a saturated regression model. The predictor variable of high LDL is binary and thus has two groups (0 and 1). To perform regression of differences in probabilities, we used a simple linear regression model which has two parameters. Since the number of groups equals the number of parameters, this 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 subjects with low LDL, the estimated probability of dying within 5 years is 17.0% and the estimated odds of dying within 5 years is 20.5%. The observed proportion of subjects in this sample with low LDL who died within 5 years is 17.0% (105 out of 618 subjects), which is the same as the estimated probability.*

* 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?

*For subjects with high LDL, the estimated probability of dying within 5 years is 13.7% and the estimated odds of dying within 5 years is 15.9%. The observed proportion of subjects in this sample with high LDL who died within 5 years is 13.7% (16 out of 117 subjects), which is the same as the estimated probability of dying within 5 years amongst high LDL subjects.*

* 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?

*We performed simple linear regression analysis assuming heteroscedasticity amongst 735 subjects using high LDL (LDL > 160 mg/dL) as a response variable and mortality within 5 years as a predictor variable. From the analysis, we estimate that for subjects with high LDL, the proportion of death within 5 years (13.7%) is 3.3% lower than the proportion of death within 5 years for subjects with low LDL (17%). However, at a level of 0.05 significance, this estimate is not statistically significant (P=0.347). A 95% confidence interval suggests that this observation is not unusual if a group of subjects with high LDL had a true proportion of death within 5 years that was 10.2% lower or 3.6% higher than a group of subjects with low LDL.*

*The point estimates for the logistic regression are the same as those observed in problems 5. Compared to problem 5, the p-value is larger with linear regression than with the two-sided chi-square test, though both are still not significant. The p-value is smaller for the chi-square test because there are fewer degrees of freedom when performing that test than when performing the logistic regression. Similarly, the confidence interval from problem 6 is narrower than in this problem, due to the assumption of heteroscedasticity necessary to handle the mean-variance relationships in regression.*

* 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?

*For death within 5 years as response variable and low LDL as predictor variable:*

* + 1. *It is still a saturated model because of two outcome categories and two parameters*
		2. *For subjects with low LDL, the odds of death within 5 years would be the same as in part b (20.5%) as would the probability of death within 5 years (17.0%).*
		3. *For subjects with high LDL, the odds of death within 5 years would be the same as in part c (15.9%), as would the probability of death within 5 years (13.7%).*

*For survival past 5 years as the response variable and high LDL as predictor variable:*

1. *It is still a saturated model because of two outcome categories and two parameters*
2. *For subjects with low LDL, the odds of survival past 5 years would be 4.88 and the probability of survival past 5 years would be 83.0%, which is appropriately the difference of the probability of death within 5 years found in part b (17.0%).*
3. *For subjects with high LDL, the odds of survival past 5 years would be 6.31 and the probability of survival past 5 years would be 86.3%, which is appropriately the difference of the probability of death within 5 years found in part b (13.7%).*
	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?

*For death within 5 years as a predictor variable and high LDL as response variable:*

* + 1. *It is still a saturated model because of two outcome categories and two parameters*
		2. *For subjects with low LDL, the odds of death within 5 years would be the same as in part b (20.5%) as would the probability of death within 5 years (17.0%).*
		3. *For subjects with high LDL, the odds of death within 5 years would be the same as in part c (15.9%), as would the probability of death within 5 years (13.7%).*
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, this is a saturated regression model. The predictor variable of high LDL is binary and thus has two groups (0 and 1). To perform regression of ratios of probabilities, we used a poisson regression model which has two parameters. Since the number of groups equals the number of parameters, this 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 subjects with low LDL, the estimated probability of dying within 5 years is 17.0% and the estimated odds of dying within 5 years is 20.5%. The observed proportion of subjects in this sample with low LDL who died within 5 years is 17.0% (105 out of 618 subjects), which is the same as the estimated probability.*

* 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?

*For subjects with high LDL, the estimated probability of dying within 5 years is 13.7% and the estimated odds of dying within 5 years is 15.9%. The observed proportion of subjects in this sample with high LDL who died within 5 years is 13.7% (16 out of 117 subjects), which is the same as the estimated probability of dying within 5 years amongst high LDL subjects.*

* 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?

*Poisson regression allowing for lack of independence over separate intervals was performed on 725 subjects with death within 5 years as the dichotomous response variable and high serum LDL (LDL > 160 mg/dL) as a dichotomous predictor variable. From the analysis, we estimate that for subjects with high LDL, the probability of death within 5 years decreases by 39.3% in comparison to the probability of death in subjects with low LDL. Based on the p-value, this observation is not statistically significant (P=0.383) and we cannot reject the null hypothesis that there is no change in the proportion of subjects that die within 5 years based on high or low LDL level. The 95% CI suggests that this observation is not unusual if a group of subjects with high LDL might have a risk of death within 5 years that was 80.3% lower to 13.5% higher than a group without the increase in LDL.*

*The point estimate provided here makes sense when compared with the point estimate of problem 5 (3.9%) which was a risk difference for a 1 mg/dL change in LDL. In this question, the probability of death is based on a log-scale and thus provides a point estimate that reflects a change in probability over 10 mg/dL changes in LDL. The P-value here is slightly larger than that in problem 5 (P=0.314), thought both suggest the test is non-significant. The confidence interval here is much different than the CI of problem 5 because the numbers are 1) based on a different scale (log-scale) and 2) reflective of a change in ratios of probabilities, rather than differences in probabilities.*

* 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?

*For death within 5 years as response variable and low LDL as predictor variable:*

* + 1. *It is still a saturated model because of two outcome categories and two parameters*
		2. *For subjects with low LDL, the odds of death within 5 years would be the same as in part b (20.5%) as would the probability of death within 5 years (17.0%).*
		3. *For subjects with high LDL, the odds of death within 5 years would be the same as in part c (15.9%), as would the probability of death within 5 years (13.7%).*

*For survival past 5 years as the response variable and high LDL as predictor variable:*

1. *It is still a saturated model because of two outcome categories and two parameters*
2. *For subjects with low LDL, the odds of survival past 5 years would be 4.88 and the probability of survival past 5 years would be 83.0%, which is appropriately the difference of the probability of death within 5 years found in part b (17.0%).*

*For subjects with high LDL, the odds of survival past 5 years would be 6.31 and the probability of survival past 5 years would be 86.3%, which is appropriately the difference of the probability of death within 5 years found in part b (13.7%).*

* 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?

*For death within 5 years as a predictor variable and high LDL as response variable:*

* + 1. *It is still a saturated model because of two outcome categories and two parameters*
		2. *For subjects with low LDL, the odds of death within 5 years would be the same as in part b (20.5%) as would the probability of death within 5 years (17.0%).*
		3. *For subjects with high LDL, the odds of death within 5 years would be the same as in part c (15.9%), as would the probability of death within 5 years (13.7%).*
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).

*Simple linear regression allowing for heteroscedasticity was performed on 725 subjects with death within 5 years as the dichotomous response variable and serum LDL (mg/dL) as a continuous predictor variable. From linear regression analysis, we estimate that for each 1 mg/dL difference in LDL between groups, the difference in risk of death within 5 years is 0.103% with subjects with higher LDL having a lower risk of death within 5 years. A 95% CI suggests that this observation is not unusual if the true difference in risk per 1 mg/dL difference in LDL was between 0.019%, 0.188%. With a p-value of 0.017, we reject the null hypothesis that there is no linear trend in the risk of death within 5 years across levels of LDL.*

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

*Poisson regression allowing for lack of independence over separate intervals was performed on 725 subjects with death within 5 years as the dichotomous response variable and serum LDL (mg/dL) as a continuous predictor variable. From the analysis, we estimate that for each 10-fold increase in serum LDL, the probability of death within 5 years decreases by 1.49%. Based on the p-value, this observation is statistically significant (P=0.018). The 95% CI suggests that this observation is not unusual if a group of subjects with a 10-fold increase in serum LDL might have a risk of death within 5 years that was 2.68% to 0.260% lower than a group without the increase in LDL.*

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

*Simple logistic regression allowing for heteroscedasticity was performed on 725 subjects with death within 5 years as the dichotomous response variable and serum LDL (mg/dL) as a continuous predictor variable. From the logistic regression analysis, we estimate that for each 1 mg/dL difference in LDL between groups, the odds of death within 5 years is 0.8% lower in subjects with higher LDL. A 95% CI suggests that this observation is not unusual if a group of subjects with 1 mg/dL increase in LDL had an odds of death within 5 years that was between 0.130% and 1.42% lower. With a p-value of 0.019, we reject the null hypothesis that there is no linear trend in the odds of death within 5 years across levels of 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*?

*By using a continuous predictor variable, rather than a dichotomous one (as in problems 1-3), we get much better precision with all three regression models. Though the point estimates were all accurate with LDL as a dichotomous variable (insofar as they matched one another and the corresponding t-tests), the p-values were too large to deem the tests significant. The confidence intervals were also very wide for each regression model. By keeping LDL as a continuous variable, we were able obtain statistically significant regression tests with much narrower confidence intervals.*

*In problems 2 and 4 of Homework 2, we used linear regression with LDL as the response variable rather than the predictor variable. The p-value for HW2 is slightly smaller (0.0115) than for this HW (0.017), but both suggest a significant test. However, because we are scientifically interested in how LDL impacts mortality in a causal sense, it is more appropriate for LDL to be treated as the predictor variable. While LDL can be adjusted throughout someone’s life, mortality is set. Thus, to set mortality as the predictor variable would require “looking back in time” to get the appropriate LDL.*

*Out of the three regression models used in this homework, linear regression is a good choice because the output (risk difference) is easier to communicate than the odds ratio. LDL is something that people can alter with diet and other potential interventions, so communicating its impact on mortality may have important public health implications. Also, the outcome isn’t rare, so relative risk isn’t necessary to emphasize the association. We’re also dealing with a cohort study, so logistic regression and odds ratio are less appropriate than with a case control study. The linear regression model in this homework is also preferable to the model used in HW02, because the scientific question is better answered by using mean LDL as the predictor variable as LDL can be adjusted by an individual, unlike mortality.*

**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.