**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 model is saturated because two distinct groups (those with high (> 160 mg/dL) serum LDL and those with low (<160 mg/dL) serum LDL) are modeled with two regression parameters (the intercept and the slope). They are both saturated models.**

* 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 performed a logistic regression model in order to evaluate the odds of death within 5 years across groups defined by high/low serum LDL levels. The odds of dying within 5 years for subjects with low LDL is obtained from our model by exponentiating the slope coefficient, since this estimated value represents the log-transformed odds of death within 5 years for the low LDL group. Hence, the estimated odds of dying within 5 years among subjects with low LDL = exp(β0) = exp(-1.5863) ≈ 0.205. For subjects with low LDL, the estimated probability of dying within 5 years is the oddslow/(1+oddslow) = exp(β0)/(1+ exp(β0)) ≈ 0.170. The observed proportion of subjects with low LDL dying within 5 years is:** $\frac{N\_{low LDL,dead}}{N\_{low LDL}}$ **=** $\frac{105}{618}$ **≈ 0.170. This observed proportion is exactly the estimated probability of dying within 5 years because of the relationship that exists between odds and probability and the fact that values (number of low LDL subjects who died and number that survived within 5 years) used to compute the observed probability and the estimated odds were equivalent. The estimated odds are about 3.5% higher than the observed proportion.**

* 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 estimated odds of dying within 5 years among subjects with high LDL is obtained from our model by exponentiating the product of the slope and intercept coefficients, since the slope estimate is the log-transformed odds ratio of dying within 5 years for high compared to low LDL subjects. Hence, the estimated odds of dying within 5 years among subjects with low LDL = odds*low*\*(odds*high*/odds*low*) = exp(β0)\*exp(β1) = exp(β0+ β1) = exp(-1.5863 -0.3072) ≈ 0.151. For subjects with high LDL, the estimated probability of dying within 5 years is the odds*high*/(1+odds*high*) = exp(β0+ β1)/(1+ exp(β0+ β1)) ≈ 0.131. This probability is identical to the observed proportion of subjects with low LDL dying within 5 years, which is about 2% lower than the estimated odds of dying within 5 years for subjects with high 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 performed a logistic regression analysis in order to evaluate the odds of death for groups defined by vital status within 5 years (1=dead, 0=alive) with our predictor serum LDL, dichotomized by high (> 160 mg/dL) and low (<160 mg/dL) levels (1=high serum LDL, 0=low serum LDL).**

**Results: Comparing the two groups, the odds of dying within 5 years is estimated to be 26.4% lower (odds ratio 0.7355) for subjects having high LDL compared to those subjects with low LDL. Based on the two-sided Z-test, this observed difference is not statistically different from an odds ratio of 1 (P = 0.315), with a 95% confidence interval suggesting that the observed odds ratio is what might be typically observed if the true odds of dying within 5 years was anywhere between 59.6% lower and 34.0% higher for subjects having high LDL compared to low LDL. We thus fail to reject the hypothesis of no association between survival time and serum LDL; although we do observe a trend toward higher odds of survival among subjects with higher LDL levels. This inference is identical to that of problem 6, and gives the same estimated probabilities and decision rule as problem 5 (both fail to reject the null hypothesis that LDL is associated with risk of death within 5 years).**

* 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?

**Had we used a logistic regression model using indicator of low LDL as our predictor this would remain a saturated model as concluded in part a, since there are still two distinct groups and 2 distinct outcomes. We would have also reached the exact same conclusions and estimates for survival odds among groups defined by high and low serum LDL, the only difference being that our intercept parameter (β0) would be the odds of death within 5 years among subjects having *high* LDL (rather than low LDL), while the slope parameter (β1) would be the odds ratio of death within 5 years for *low LDL to high LDL*, or equivalently the reciprocal of our slope parameter for the model in parts a-c.**

**Had we used an indicator of survival within 5 years instead of death, than our estimated intercept (β0) would be the reciprocal of odds of death within 5 years among subjects having low LDL, or equivalently, it would be the odds of *survival* within 5 years among subjects having low LDL (since odds*survival* = 1/odds*death*). Similarly, our estimate of the slope parameter (β1) would be the reciprocal of our slope parameter for the model in parts a-c, or equivalently, the odds ratio of *survival* within 5 years for high LDL to low LDL.**

* 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?

**Had we fit a logistic regression where LDL (high compared to low groups) is analyzed based on the predictor vital status (dead/alive within 5 years) this would remain a saturated model as concluded in part a, since there are still two distinct groups and 2 distinct outcomes. We would obtain different estimated odds for individual response groups (i.e. we would get the odds of high serum LDL for both groups alive and dead within 5 years, as opposed to the odds of death for both groups high LDL and low LDL). The odds ratio interesting remains the same in this analysis, since the odds ratio of high serum LDL among patients dead compared to alive within 5 years is equivalent to the odds ratio of death within 5 years among patents with high LDL compared to low LDL. This is most easily observed by the following equivalent odds ratio formulae:**

**Odds ratio of death within 5 years among groups defined by high compared to low LDL =**

$ \frac{N\_{high LDL \& dead}}{N\_{high LDL \& alive}}/\frac{N\_{low LDL \& dead}}{N\_{low LDL \& alive}}$ **=** $\frac{N\_{high LDL \& dead}}{N\_{high LDL \& alive}}\*\frac{N\_{low LDL \& alive}}{N\_{low LDL \& dead}}$ **=** $\frac{N\_{high LDL \& dead}}{N\_{low LDL \& dead}}/\frac{N\_{high LDL \& alive}}{N\_{low LDL \& alive}}$ **=**

**Odds ratio of high LDL among groups defined by death compared to survival within 5 years.**

 **The confidence interval remains the same since the degrees of freedom are equivalent and the standard error is computed using exactly the same values, i.e. standard error for log odds ratio =**

$$\sqrt{\frac{1}{N\_{high LDL \& alive}}+\frac{1}{N\_{low LDL \& dead}}+\frac{1}{N\_{high LDL \& alive}}+\frac{1}{N\_{low LDL \& dead}}}$$

**for all parameterizations and forms of logistic regression with a binary outcome and a binary response.**

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 model is saturated because two distinct groups (those with high (> 160 mg/dL) serum LDL and those with low (<160 mg/dL) serum LDL) are modeled with two regression parameters (the intercept and the slope). They are both saturated models.**

* 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 the estimated intercept parameter β0 = 0.16990 ≈ 0.170. The estimated probability of dying within 5 years is plow LDL/(1-plow LDL) = 0.16990/(1-0.16990) = 0.205. The estimated probability of dying within 5 years for low LDL subjects is exactly the observed proportion of subject with low LDL dying within 5 years, which is about 3.5% lower than the estimated odds of dying within 5 years for subjects with low LDL, based on the regression analysis.**

* 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 the estimated intercept parameter plus the estimated slope parameter: β0 + β1 = 0.16990-0.03906 = 0.13084 ≈ 0.131. The estimated odds of dying within 5 years is phigh ­LDL/(1-phigh LDL) = 0.13084/(1-** **0.13084) = 0.1505361 ≈ 0.151 . The estimated probability of dying within 5 years for high LDL subjects is exactly the observed proportion of subject with high LDL dying within 5 years, which is about 2% lower than the estimated odds of dying within 5 years for subjects with low LDL, based on the regression analysis.**

* 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 performed a binomial regression analysis using the identity link in order to evaluate the difference in probability of death for groups defined by vital status within 5 years (1=dead, 0=alive) with our predictor serum LDL, dichotomized by high (> 160 mg/dL) and low (<160 mg/dL) levels (1=high serum LDL, 0=low serum LDL). This utilizes a Wald test statistic based on the normal approximation to the binomial distribution.**

**Results: Comparing the two groups, the probability of dying within 5 years is estimated to be 3.91% lower for subjects having high LDL (probability = 0.16990-0.03906 ≈ 0.131) compared to those subjects with low LDL (0.170). Based on the Z-test, this observed difference is not statistically different from difference in probability of 0 (two-sided P = 0.277), with a 95% confidence interval suggesting that the observed difference is what might be typically observed if the true difference in probability of dying within 5 years was anywhere between 0.109 lower and 0.031 higher for subjects having high LDL compared to low LDL. We thus fail to reject the hypothesis of no association between survival time and serum LDL; although we do observe a trend toward higher probability of survival among subjects with higher LDL levels. This inference is identical to that of problem 5, and gives the same estimated probabilities and decision rule as problem 6 (both fail to reject the null hypothesis that LDL is associated with risk of death within 5 years).**

* 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?

**Had we fit a regression model using indicator of low LDL as our predictor this would remain a saturated model as concluded in part a, since there are still two distinct groups and 2 distinct outcomes. We would have reached the exact same conclusions and estimates for survival probability among groups defined by high and low serum LDL, the only difference being that our intercept parameter (β0) would be the probability of death within 5 years among subjects having *high* LDL (instead of low LDL), while the slope parameter (β1) would be the difference in probability of death within 5 years for *low LDL minus high LDL*, or equivalently the negation of our slope parameter for the model in parts a-c.**

**Had we used an indicator of survival within 5 years instead of death, than our estimated intercept (β0) would be one minus the probability of death within 5 years among subjects having low LDL, or equivalently, it would be the probability of *survival* within 5 years among subjects having low LDL (since p*survival* = 1-p*death*). Similarly, our estimate of the slope parameter (β1) would be the negation of the slope parameter estimate from the model in parts a-c, or equivalently, the difference in probability of death within 5 years for low LDL minus high LDL since:**

**Palive|highLDL –Palive|lowLDLa**

**= (1– Pdead|highLDL) – (1 – Pdead|lowLDL) = Pdead|lowLDL -Pdead|highLDLb**

**= - ( Pdead|highLDL – Pdead|lowLDL)c**

**a: our estimate of the slope parameter (β1)**

**b: difference in probability of death within 5 years for low LDL minus high LDL**

**c: negation of the slope parameter estimate from the model in parts a-c**

* 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?

**Had we fit a regression where LDL (high compared to low groups) is analyzed based on the predictor vital status (dead/alive within 5 years) this would remain a saturated model as concluded in part a, since there are still two distinct groups and 2 distinct outcomes. We would obtain different estimated probabilities for individual response groups (i.e. we would get the probability of high serum LDL for both groups alive and dead within 5 years, as opposed to the probability of death for both groups high LDL and low LDL). Our decision rule would not change- we would fail to reject the hypothesis of no association between survival time and serum LDL, based on an identical 2-sided P value of 0.277.**

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 model is saturated because two distinct groups (those with high (> 160 mg/dL) serum LDL and those with low (<160 mg/dL) serum LDL) are modeled with two regression parameters (the intercept and the slope). They are both saturated models.**

* 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 the exponentiated estimated intercept parameter exp(-1.77253) = 0.16990 ≈ 0.170. The estimated probability of dying within 5 years is plow LDL/(1-plow LDL) = 0.16990/(1-0.16990) = 0.205. The estimated probability of dying within 5 years for low LDL subjects is exactly the observed proportion of subject with low LDL dying within 5 years, which is about 3.5% lower than the estimated odds of dying within 5 years for subjects with low LDL, based on the regression analysis.**

* 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 the exponentiated sum of the estimated intercept parameter plus the estimated slope parameter: exp(-1.77253+** **-0.26124)= 0.13084 ≈ 0.131. The estimated odds of dying within 5 years is phigh ­LDL/(1-phigh LDL) = 0.13084/(1-** **0.13084) = 0.1505361 ≈ 0.151 . The estimated probability of dying within 5 years for high LDL subjects is exactly the observed proportion of subject with high LDL dying within 5 years, which is about 2% lower than the estimated odds of dying within 5 years for subjects with low LDL, based on the regression analysis.**

* 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 performed a poisson regression analysis in order to evaluate the ratio of probability of death for groups defined by vital status within 5 years (1=dead, 0=alive) with our predictor serum LDL, dichotomized by high (> 160 mg/dL) and low (<160 mg/dL) levels (1=high serum LDL, 0=low serum LDL).**

**Results: Comparing the two groups, the probability of dying within 5 years is estimated to be 23.0% lower (risk ratio 0.770) for subjects having high LDL compared to those subjects with low LDL. Based on the two-sided Z-test, this observed difference is not statistically different from a risk ratio of 1 (P = 0.359), with a 95% confidence interval suggesting that the observed risk ratio is what might be typically observed if the true risk of dying within 5 years was anywhere between 55.9% lower and 34.5% higher for subjects having high LDL compared to low LDL. We thus fail to reject the hypothesis of no association between survival time and serum LDL; although we do observe a trend toward higher risk of survival among subjects with higher LDL levels. This inference gives the same estimated probabilities and decision rule as in problems 5 and 6 (i.e. equivalent estimates for probabilities and odds and all fail to reject the null hypothesis that LDL is associated with risk of death within 5 years). The two-sided P-value is higher in this case since the test statistic is based on a higher standard error.**

* 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?

**Had we fit a regression model using indicator of low LDL as our predictor this would remain a saturated model as concluded in part a, since there are still two distinct groups and 2 distinct outcomes. We would have reached the exact same conclusions and estimates for survival probability among groups defined by high and low serum LDL, the only difference being that our intercept parameter (β0) would be the probability of death within 5 years among subjects having *high* LDL (instead of low LDL), while the slope parameter (β1) would be the ratio of probability of death within 5 years of *low LDL to high LDL*, or equivalently the reciprocol of our slope parameter for the model in parts a-c.**

**Had we used an indicator of survival within 5 years instead of death, than our estimated intercept (β0) would be one minus the probability of death within 5 years among subjects having low LDL, or equivalently, it would be the probability of *survival* within 5 years among subjects having low LDL (since p*survival* = 1-p*death*). Also, our estimate of the slope parameter (β1) would be related to the slope and intercept parameters estimate from the model in parts a-c, by the following formula:**

**Palive|highLDL/Palive|lowLDLa**

**= (1– Pdead|highLDL) /(1 – Pdead|lowLDL)**

**= (1 – exp(β0(a-c) + β1(a-c)) )/(1-exp(β0(a-c)) )b**

**a: our estimate of the slope parameter (β1) using indicator of survival**

**b: β0(a-c) & β1(a-c) represent the intercept and slope estimates from model using indicator of death (parts a-c)**

* 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?

**Had we fit a regression where LDL (high compared to low groups) is analyzed based on the predictor vital status (dead/alive within 5 years) this would remain a saturated model as concluded in part a, since there are still two distinct groups and 2 distinct outcomes. We would obtain different estimated probabilities for individual response groups (i.e. we would get the probability of *high serum LDL* for both groups alive and dead within 5 years, as opposed to the probability of *death* for both groups high LDL and low LDL). Our decision rule would not change- we would fail to reject the hypothesis of no association between survival time and serum LDL, but the test statistic and P-value change in the direction of a less conservative test (higher absolute value of test statistic and lower P-value of 0.354 compared to 0.359).**

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 performed a linear regression analysis in order to evaluate the difference in probability of death for groups defined by vital status within 5 years (1=dead, 0=alive) with the continuous predictor serum LDL (mg/dL).**

**Results: Comparing the two groups, the estimated difference in probability of dying within 5 years per 1.0 mg/dL increase in serum LDL is -0.001, which is estimated by the slope parameter obtained from our model. Based on the t-test, this observed difference in probability is statistically different from 0 (two-sided P = 0.0115), with a 95% confidence interval suggesting that the observed difference is what might be typically observed if, for each 1.0 mg/dL increase in serum LDL, the true difference in probability of dying within 5 years was anywhere between 0.0018 to 0.0002 lower. We thus reject the hypothesis of no association between survival time and serum LDL.**

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

**Methods: We performed a poisson regression analysis in order to evaluate the ratio of risk of death for groups defined by vital status within 5 years (1=dead, 0=alive) with the continuous predictor serum LDL (mg/dL).**

**Results: Comparing the two groups, the probability of dying within 5 years is estimated to be 0.64% lower (risk ratio 0.9936) for subjects for each 1.0 mg/dL increase in serum LDL. Based on the Z-test, this observed difference is statistically different from a risk ratio of 1 (two-sided P = 0.02091), with a 95% confidence interval suggesting that the observed risk ratio is what might be typically observed if, for each 1.0 mg/dL increase in serum LDL, the true risk of dying within 5 years was anywhere between was anywhere between 1.19% and 0.10% lower for subjects having high LDL compared to low LDL. We thus reject the hypothesis of no association between survival time and serum LDL.**

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

**Methods: We performed a logistic regression model in order to evaluate the odds of death for groups defined by vital status within 5 years (1=dead, 0=alive) with the continuous predictor serum LDL (mg/dL).**

**Results: Comparing the two groups, the odds of dying within 5 years is estimated to be 0.10% lower (odds ratio 0.9990) for subjects having high LDL compared to those subjects with low LDL. Based on the two-sided Z-test, this observed difference is statistically different from an odds ratio of 1 (P = 0.315), with a 95% confidence interval suggesting that the observed odds ratio is what might be typically observed if the true odds of dying within 5 years was anywhere between 1.18% and 0.02% lower for subjects having high LDL compared to low LDL. We thus reject the hypothesis of no association between survival time and 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*.?

**The conclusions of association from these models (linear, poisson, and logistic regression of death status within 5 years on continuous serum LDL) all indicate a significant association between serum LDL and risk of death within 5 years. This is contrary to the conclusions obtained in problems 1-3, where we do not detect such an association. As in problems 2 and 4 of hw #2, we conclude that there is a significant association between serum LDL and 5 year all-cause mortality, however in this case we do not use LDL as the response variable but the predictor of interest. Were I conducting this analysis, I would choose to use serum LDL as the predictor variable since this is a cross sectional cohort study where the measured exposure (serum LDL) is known, and we wish to make inference about its effect on the outcome mortality. Since this study was designed by randomly sampling older adults on medicaid, the risk ratio would be generalizable to this population of older adults, and I would prefer to use poisson regression to estimate the risk ratio of 5 year death status. I would use the continuous variable serum LDL as the predictor of interest in order to increase the precision of the association.**

**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.