**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 model because the regression model has two parameters which we need to estimate and we have only two groups in our predictor of interest, the first group which has high serum LDL and the second group which does not have high serum LDL i.e they have levels < 160 mg/dL
	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?

	The estimated odds of dying within five years is the exponentiated intercept of the logistic regression model because we are using an indicator variable for “high” LDL as the response so the intercept of the logistic regression is the odds of dying within 5 years for those with low serum LDL levels. The estimated odds in this case is 0.205 or 20.5%. The second part of the question was not clear if it was asking for the estimated propbability of dying within 5 years for all patients or the probability of dying within 5 years given the patients had low LDL. Based on the regression model the proportion can be evaluated by $prop= \frac{odds}{1+odds}$ and so for patients with low LDL the estimated probability of dying within 5 years is 0.170 or 17.0% and if the question was asking for the unconditional probability of dying within 5 years then that can be estimated as the sample propotion of all patients who died within 5 years which in this case is 0.164 or 16.4%.
	The estimated proportion of patients dying within 5 years with low LDL is exactly the same as the observed proportion of patients with low LDL who died within 5 years.
	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 high LDL dying within 5 years?

	Based on the regression model the estimated odds of dying within 5 years for subjects with high LDL is the sum of the exponentiation of the sum of the intercept and the slope of the log odds logistic regression model. In this case the estimate is 0.151 or 15.1%. Once again, we can estimate the proportion based on the odds and in this case the estimated proportion of subjects with high LDL who died within 5 years is 0.131 or 13.1%. Once again, this estimate is exactly equal to the observed proportion of subjects with high LDL who died within 5 years.
	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?

	**Inference:** From logistic regression analysis, we estimate that for the group with high serum LDL level the odds of dying within 5 years is 15.1% and the odds of dying within 5 years for patients with low LDL is 20.5%. The odds of dying within 5 years is 35.8% higher for patients with low serum LDL than the odds of patients with high serum LDL where we are considering the difference of odds or the odds are 36.0% higher for those with low serum LDL based on the odds ratio, though this estimate is not statistically significant (P = .316). A 95% CI suggests that this observation is not unusual if a group that has high serum LDL might have odds of dying within 5 years that was anywhere from 59.7% lower or 34.2% higher (or if the odds ratio were anywhere between 0.597 to 1.342) than the group with low serum LDL. So we fail to reject the null hypothesis as we don’t have enough statistical evidence to conclude that there is an association between the two variables.

Compared to inference made in assignment 1 we notice that the point estimates for the odds and odds ratio are the same but we get slightly different P-values and confidence intervals. Atleast for the assignment key, we used the Fisher’s exact test and the confidence intervals were generated by an exact method too which explains the differences in the confidence intervals and point estimates. The Wald confidence intervals are much more similar to the estimated confidence intervals above.

* 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?

	The answers would not change for the point estimates, the only change would be how be obtain the point estimates. For both the models described in the question we still only have 2 groups since the predictor is still a binary variable and so we still have a saturated model. When we use the indicator of low LDL as the predictor to get the odds of dying with 5 years for those with low LDL the estimate is not the exponentiated intercept but the exponentiated sum of the intercept and slope and it is the exponentiated intercept of the model to estimate the odds of dying within 5 years for subjects with high LDL. Similarly if we change the response to be an indicator of surviving 5 years the intercept gives us the odds of surviving atleast 5 years and so if we need the odds of dying with 5 years we’d have to consider the reciprocal of the intercept and so on and so forth.
	2. 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?

	If we, effectively switch the predictor and response then we are fitting a different model. In this case if we use indicator of death within 5 years as the predictor then we still have a binary variable and so we still have a saturated model. With this model however, we cannot estimate the odds of dying within 5 years or the odds ratio.
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 model because the regression model has two parameters which we need to estimate and we have only two groups in our predictor of interest, the first group which has high serum LDL and the second group which does not have high serum LDL i.e they have levels < 160 mg/dL
	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?

	Based on a simple linear regression on the mean (which in this case is a proportion since the response is binary) we have the estimated probability of dying within 5 years for subjects with low LDL as 0.170 or 17.0%. From the estimated probability we estimate the odds of dying within 5 years for subject with low serum LDL as 0.205 or 20.5%. Which is exactly the same as the observed proportion of subjects with low serum LDL who died within 5 years.
	3. 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 high LDL dying within 5 years?

	Based on a simple linear regression on the mean (which in this case is a proportion since the response is binary) we have the estimated probability of dying within 5 years for subjects with high LDL as 0.131 or 13.1%. From the estimated probability we estimate the odds of dying within 5 years for subject with high serum LDL as 0.151 or 15.1%. Which is exactly the same as the observed proportion of subjects with high serum LDL who died within 5 years.
	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?

	**Inference:** From the linear regression analysis, we estimate that for the group with high serum LDL level the probability of dying within 5 years is 13.1% and the estimated probability of dying within 5 years for patients with low LDL is 17.0%. The probability of dying within 5 years is 29.8% higher for patients with low serum LDL than the probability of patients with high serum LDL where we are considering the difference of probability or alternatively we can say the exact difference in probability is 0.039 or 3.9% where in this case we are considering the absolute difference as opposed to the percentage difference though this difference is not significant (P Value= 0.278). A 95% CI suggests that this observation is not unusual if a group that has high serum LDL might have probability of dying within 5 years that was anywhere from 0.110 lower or 0.032 higher than the group with low serum LDL.

Compared to inference made in assignment 1 we notice that the point estimates for the odds and odds ratio are the same but we get slightly different P-values and confidence intervals. At-least for the assignment key, we used the Fisher’s exact test and the confidence intervals were generated by an exact method too which explains the differences in the confidence intervals and point estimates. The Wald confidence intervals are much more similar to the estimated confidence intervals above.

* 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?

	Once again, like question 1e we will not notice any change in the estimates but only in how we evaluate these estimates. We still have a binary predictor as we are looking at low versus high serum LDL so we have a saturated model in both cases. For the first case where we use an indicator of low LDL as our predictor the intercept is the probability of dying within 5 years for high LDL and similarly for low LDL now we need to take the sum of the intercept and slope for the estimated probability of dying within 5 years. And similarly for the second model where we re-parameterize the response we get the probability of surviving at-least 5 years and so we’d evaluate the probability of dying within 5 years by prob = 1 – probility of surviving at-least 5 years
	2. 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?

	If we, effectively switch the predictor and response then we are fitting a different model. In this case if we use indicator of death within 5 years as the predictor then we still have a binary variable and so we still have a saturated model. With this model however, we cannot estimate the risk of death within 5 years or the risk ratio.
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 model because the regression model has two parameters which we need to estimate and we have only two groups in our predictor of interest, the first group which has high serum LDL and the second group which does not have high serum LDL i.e they have levels < 160 mg/dL
	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?

	Based on a Poisson regression model on the mean (which in this case is a proportion since the response is binary) we have the estimated probability of dying within 5 years for subjects with low LDL as 0.170 or 17.0% which is obtained by the exponent of the intercept. From the estimated probability we estimate the odds of dying within 5 years for subject with low serum LDL as 0.205 or 20.5%. Which is exactly the same as the observed proportion of subjects with low serum LDL who died within 5 years.
	3. 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 high LDL dying within 5 years?

	Based on a Poisson regression model on the mean (which in this case is a proportion since the response is binary) we have the estimated probability of dying within 5 years for subjects with high LDL as 0.131 or 13.1% where the estimate is obtained by the exponentiation of the sum of intercept and slope. From the estimated probability we estimate the odds of dying within 5 years for subject with high serum LDL as 0.151 or 15.1%. Which is exactly the same as the observed proportion of subjects with high serum LDL who died within 5 years.
	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?

	**Inference:** From the Poisson regression analysis, we estimate that for the group with high serum LDL level the probability of dying within 5 years is 13.1% and the estimated probability of dying within 5 years for patients with low LDL is 17.0%. We estimate that the probability of dying within 5 years decreases by 23.0% for patients with high serum LDL though this difference is not significant (P Value= 0.324). A 95% CI suggests that this observation is not unusual if the risk ratio was between 0.458 and 1.30, i.e. of the probability of dying within 5 years for the group of subjects with high LDL was 54.2% lower and 29.5% higher.

Compared to inference made in assignment 1 we notice that the point estimates for the odds and odds ratio are the same but we get slightly different P-values and confidence intervals. At-least for the assignment key, we used the Fisher’s exact test (or the chi squared test) and the confidence intervals were generated by an exact method too which explains the differences in the confidence intervals and point estimates. The Wald confidence intervals are much more similar to the estimated confidence intervals above

* 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?

	Once again, like question 1e we will not notice any change in the estimates but only in how we evaluate these estimates. We still have a binary predictor as we are looking at low versus high serum LDL so we have a saturated model in both cases. For the first case where we use an indicator of low LDL as our predictor the exponentiation of the intercept is the probability of dying within 5 years for high LDL and similarly for low LDL now we need to take the exponentiation of the sum of the intercept and slope for the estimated probability of dying within 5 years. And similarly for the second model where we re-parameterize the response we get the probability of surviving at-least 5 years and so we’d evaluate the probability of dying within 5 years by prob = 1 – probility of surviving at-least 5 years
	2. 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?

	If we, effectively switch the predictor and response then we are fitting a different model. In this case if we use indicator of death within 5 years as the predictor then we still have a binary variable and so we still have a saturated model. With this model however, we cannot estimate the risk of death within 5 years or the risk ratio.
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).

	**Inference:** From linear regression analysis, we estimate that for each mg/dL difference in serum LDL, the risk difference for dying within 5 years versus surviving atleast 5 years is -0.001 mg/dL. A 95% CI suggests that this observation is not unusual if the true risk difference were between -0.00188 and -0.0001847 mg/dL. Because the P value (0.0171) is P < .05, we reject the null hypothesis at a 95% confidence level that there is no linear trend in the probability of dying within 5 years and serum LDL levels.”
	2. Evaluate associations between 5 year mortality and LDL using risk ratio (RR: ratios of probabilities).

	**Inference:** “From Poisson regression analysis, we estimate that for each 1 fold increase in serum LDL levels, the probability of dying within 5 years decreases by 0.645%, statistically significant observation (P < 0.05). A 95% CI suggests that this observation is not unusual if for a 1 fold increase in serum LDL levels the probability of dying within 5 years decrease by 1.16% to 0.111% decrease. Based on these results we reject the null hypothesis in favor of the alternative that there is an association between serum LDL levels and probability of dying within 5 years.
	3. Evaluate associations between 5 year mortality and LDL using odds ratio (OR: ratios of odds)

**Inference:** “From logistic regression analysis, we estimate that for each mg/dL difference in serum LDL level, the odds of dying within 5 years is 0.777% lower in the group woth higher serum LDL. This estimate is statistically significant (P = .0198 < 0.05). A 95% CI suggests that this observation is not unusual if a group that has one mg/dL serum LDL higher might have odds of dying with 5 years that was anywhere from 1.42% lower or 0.124% lower than the group with lower 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*.?

	These results seems to exhibit a different conclusion compared to the first 3 questions and the previous assignment. The regression analysis performed in this question shows there is statistically significant evidence to suggest an association between serum LDL levels and probability of dying within 5 years of the study whereas before when we dichotomized the predictor of interest there seemed to be no evidence of association. A priori I would say the choice of analysis really depends on the scientific question. If we are primarily concerned with the probability of surviving if patients have high serum LDL levels based on some clinically determined threshold. But if we wish to recognize or test for a general association between the two variables I would perform the analysis used in this question considering the predictor as a continuous variable mostly because of statistical reasons because dichotomizing a continuous variable leads to loss of power for our test. In other words, keeping it as continuous variable preserves all the information in our data.

**Discussion Sections: January 22 – 24, 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.