**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 regression model is saturated, because the number of model parameters (2) is matched by the number of possible values of the predictor variable (also 2).

* 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 estimated odds of dying within five years given a subject has low LDL (serum LDL <=160 mg/dl) can be calculated from the following logistic regression equation:

$$Y\_{i}=e^{β\_{0}}\*e^{β\_{1}\*X\_{i}}$$

in which $β\_{0}$ is -1.59 and $β\_{1}$ is -0.307, $Y\_{i}$ represents the odds of death in under 5 years, and $X\_{i}$ represents the status high serum LDL (defined as serum LDL >=160 mg/dl). $X\_{i}$ takes on a value of 1 if the subject has high LDL, and a value of 0 for low LDL, so the value of $Y\_{i}$ for low LDL is exp(-1.59)\*exp(-0.307\*0) which equals 0.204. Thus, according to this model, the odds of dying in 5 years given low serum LDL is 0.204.

The estimated probability of dying within 5 years can be calculated from the estimated odds of dying within 5 years by the following relationship:

$$P=\frac{odds}{1+odds}$$

This yields a probability of death within 5 years given low serum LDL equal to (0.204/(1+0.204)), or 0.169.

The observed proportion of subjects with low LDL that died within 5 years is 0.170, which is equivalent to the probability of dying within 5 years given low LDL. This value is 0.001 (0.69%) higher than the probability estimated by the logistic regression.

* 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 estimated odds of dying within 5 years given high LDL can be calculated from the logistic regression formula by plugging in a value of 1 for $X\_{i}$, which yields the following:

$$odds=e^{-1.59}\*e^{-0.307\*1}$$

Thus, an estimation of the odds of dying within 5 years given high LDL is 0.150.

The estimated probability of dying within 5 years given high serum LDL can be calculated as (0.150/(1+0.150)), or 0.130.

The observed proportion of subjects with high LDL that died within 5 years is 0.126, which is 0.004 (3.1%) lower than the probability estimated by the logistic regression.

* 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 odds of subjects dying within 5 years of study enrollment were compared between subjects who had serum LDL greater than or equal to 160 mg/dL and subjects whose serum LDL was measured to be 159 mg/dL or less. The odds of dying within 5 years of study enrollment was estimated using a simple logistic regression with an indicator of high serum LDL as the predictor variable. An estimate of the odds of death given low LDL levels as well as an estimate of the odds ratio comparing the high LDL group with the low LDL group were produced with 95% confidence intervals and two-sided p-values as computed based on the Wald statistic.

Results: Of the 618 subjects whose serum LDL was less than or equal to 159 mg/dL, the odds of dying within 5 years from study enrollment was 0.204, while for the subjects with serum LDL greater than or equal to 160 mg/dL the odds of 5 year mortality was 0.150. Based on a 95% confidence interval, this observed odds ratio of 0.736 for the comparison of the high LDL group to the low LDL group would not be considered unusual if the true odds ratio were anywhere between 0.403 and 1.34. A two-sided p value of 0.316 suggests that we cannot with high confidence reject the null hypothesis that the odds of 5-year mortality are not associated with serum LDL levels.

This inference is very similar to that of homework #1 problem #6, which yielded an odds ratio of 0.735 (0.13% different) with a 95% confidence interval of [0.373, 1.36] and a p-value of 0.396. The overall conclusion, that the null hypothesis of no association between LDL and 5-year mortality cannot be rejected, is the same between this problem and that of homework #1.

* 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?*

A logistic regression model using an indicator of low LDL as the predictor variable would also be saturated, since the number of parameters and indicator values are not changed. The model would yield an estimated odds ratio of 1.42 comparing odds of low LDL to high LDL, with a 95% confidence interval of [-0.268, 0.972] as opposed to the original model’s estimated odds ratio of 0.736 comparing odds of high LDL to low LDL. This represents a reparamaterization of the original model.

A logistic regression model using an indicator of survival for at least 5 years as the response variable would also be saturated, since the number of parameters and indicator values are not changed. The model would yield an estimated odds ratio of 1.36 comparing odds of high LDL to low LDL with a 95% confidence interval of [0.745, 2.482] as opposed to the original model’s odds ratio of 0.736. The new model’s odds ratio of 1.36 reflects a ratio of the odds of surviving past 5 years to the odds of dying within 5 years, which is the inverse of the odds ratio for the original model. This represents a reparamaterization of the original 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?*

To describe the distribution of LDL across groups defined by vital status would involve fitting a logistic regression model with the predictor variable an indicator of death within 5 years and the response variable an indicator of high LDL (switching the predictor and response variables from the model used in parts a-c). The model would yield an estimated odds ratio of 0.736 with a 95% confidence interval of [0.4030, 1.3423], which exactly matches the estimated odds ratio and 95% confidence interval of the original model. The new model odds ratio represents the ratio of the odds of having high LDL given a vital status of “dead” at 5 years to the odds of having low LDL given a vital status of “dead” at 5 years, which ends up equivalent to the ratio of the odds of dying within 5 years given high LDL to the odds of surviving 5 years given high LDL.

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 regression model is saturated, because the number of model parameters (2) is matched by the number of possible values of the predictor variable (also 2).

* 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 estimated probability of dying within five years given a subject has low LDL (serum LDL <=160 mg/dl) can be calculated from the following linear regression equation:

$$Y\_{i}=β\_{0}+β\_{1}\*X\_{i}$$

in which $β\_{0}$ is 0.153 and $β\_{1}$ is -0.0358, $Y\_{i}$ represents the probability of death in under 5 years, and $X\_{i}$ represents the status high serum LDL (defined as serum LDL >=160 mg/dl). $X\_{i}$ takes on a value of 1 if the subject has high LDL, and a value of 0 for low LDL, so the value of $Y\_{i}$ for low LDL is (0.153)+(-0.0358\*0) which equals 0.153. Thus, according to this model, the probability of dying in 5 years given low serum LDL is 0.153.

The estimated odds of dying within 5 years can be calculated from the estimated odds of dying within 5 years by the following relationship:

$$odds=\frac{p}{1-p}$$

This yields a probability of death within 5 years given low serum LDL equal to (0.153/(1-0.153)), or 0.181.

The observed proportion of subjects with low LDL that died within 5 years is 0.170, which is equivalent to the probability of dying within 5 years given low LDL. This value is 0.017 (10.5%) higher than the probability estimated by the linear regression.

* 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 high LDL dying within 5 years?*

The estimated probability of dying within 5 years given high LDL can be calculated from the linear regression formula by plugging in a value of 1 for $X\_{i}$, which yields the following:

$$odds=0.153-0.0358\*(1)$$

Thus, an estimation of the odds of dying within 5 years given high LDL is 0.117.

The estimated probability of dying within 5 years given high serum LDL can be calculated as (0.117/(1-0.117)), or 0.193.

The observed proportion of subjects with high LDL that died within 5 years is 0.126, which is 0.067 (42%) lower than the probability estimated by the linear regression.

* 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 probability of subjects dying within 5 years of study enrollment were compared between subjects who had serum LDL greater than or equal to 160 mg/dL and subjects whose serum LDL was measured to be 159 mg/dL or less. The probability of dying within 5 years of study enrollment was estimated using a simple linear regression with an indicator of high serum LDL as the predictor variable and an indicator of death within 5 years as the response variable. An estimate of the probability of death given low LDL levels as well as an estimate of the difference in proportions comparing the high LDL group with the low LDL group were produced with 95% confidence intervals and two-sided p-values as computed based on the Wald statistic.

Results: Of the 618 subjects whose serum LDL was less than or equal to 159 mg/dL, the probability of dying within 5 years from study enrollment was 0.170, while for the subjects with serum LDL greater than or equal to 160 mg/dL the probability of 5 year mortality was 0.131. Based on a 95% confidence interval, this observed difference in proportions of -0.0391 for the comparison of the high LDL group to the low LDL group would not be considered unusual if the true difference in proportions were anywhere between -0.120 and 0.0316. A two-sided p value of 0.0.278 suggests that we cannot with high confidence reject the null hypothesis that the odds of 5-year mortality are not associated with serum LDL levels.

This inference is very similar to that of homework #1 problem #5, which yielded a difference of proportions of -0.0391 with a 95% confidence interval of [-0.109, 0.0314] and a p-value of 0.314. The overall conclusion, that the null hypothesis of no association between LDL and 5-year mortality cannot be rejected, is the same between this problem and that of homework #1.

* 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?*

If the linear regression model was fitted using an indicator of low LDL as the predictor (as opposed to an indicator of high LDL), it would remain saturated (since the numbers of parameters and predictor values would not change) and yield a difference in proportions of 0.0442 with a 95% confidence interval of [-0.0266, 0.115], which represents the proportion of subjects who died within 5 years of enrollment with low LDL minus the proportion of subjects who survived past 5 years of enrollment with low LDL, as opposed to the original model’s result that represents the difference in proportions between death and survival given high LDL. This is a reparameterization of the original model.

If the response variable used was an indicator of survival past 5 years (as opposed to an indicator of death under 5 years), it would remain saturated (since the numbers of parameters and predictor values would not change) and yield a difference in proportions of 0.0391 with a 95% confidence interval of [-.0316, 0.110], which represents the proportion of subjects who survived within 5 years of enrollment with high LDL minus the proportion of subjects who died withhin 5 years of enrollment with high LDL, as opposed to the original model’s result that represents the proportion who died minus the proportion who survived (so the absolute value of the result is the same). This is a reparameterization of the original 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?*

To describe the distribution of LDL across groups defined by vital status would involve fitting a logistic regression model with the predictor variable an indicator of death within 5 years and the response variable an indicator of high LDL (switching the predictor and response variables from the model used in parts a-c). The model would yield an estimated difference in proportions of -0.0358 with a 95% confidence interval of [-0.101, 0.0290], which represents the proportion of subjects with high LDL who died within 5 years of study enrollment minus the proportion of subjects with low LDL who died within 5 years of study enrollment, as opposed to the original model’s result meaning the proportion of subjects who died within 5 years given high LDL minus the proportion of subjects who survived past 5 years given high LDL.

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 regression model is saturated, because the number of model parameters (2) is matched by the number of possible values of the predictor variable (also 2).

* 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 calculated by raising *e* to the power of the intercept of a Poisson regression model that uses as a predictor variable an indicator that a subject has high LDL and uses as a response variable an indicator that a subject has died withinh 5 years of study enrollment. This result becomes *e^(-1.773)* or 0.170.

The estimated odds of dying within 5 years given low LDL is equal to p/(1-p), and becomes (0.170/(1-0.170)) or 0.205.

The observed proportion of subjects with low LDL that die within 5 years is 0.170, which is exactly matched by the value estimated by the Poisson regression model.

* 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 high LDL dying within 5 years?*

For subjects with high LDL, the estimated probability of dying within 5 years is calculated by multiplying *e* raised to the power of the slope of the Poisson regression model (which represents the ratio of the proportion who died given high LDL to the proportion who died given low LDL) by the estimated proportion that died within 5 years given low LDL, which becomes the following calculation: *e^(-0.261)\*0.170* which equals 0.131.

The estimated odds of dying within 5 years given high LDL is equal to p/(1-p), and becomes 0.151.

The observed proportion of subjects with high LDL that die within 5 years is 0.126, which is 0.005 (3.9%) different from that estimated by the Poisson regression.

* 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 probability of subjects dying within 5 years of study enrollment were compared between subjects who had serum LDL greater than or equal to 160 mg/dL and subjects whose serum LDL was measured to be 159 mg/dL or less. The probability of dying within 5 years of study enrollment was estimated using a Poisson regression with an indicator of high serum LDL as the predictor variable and an indicator of death within 5 years as the response variable. An estimate of the probability of death given low LDL levels as well as an estimate of the ratio of proportions comparing the high LDL group with the low LDL group were produced with 95% confidence intervals and two-sided p-values as computed based on the Wald statistic.

* 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?*

Both of these changes would be a reparamaterization of the original model, and the differences between these results and the original results would be on a log scale.

* 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?*

This change would entail switching the predictor and response variables, representing a reparamaterization of the original model.

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).*
	2. *Evaluate associations between 5 year mortality and LDL using risk ratio (RR: ratios of probabilities).*
	3. *Evaluate associations between 5 year mortality and LDL using odds ratio (OR: ratios of odds)*
	4. *How do your conclusions about such an association from this model compare to your conclusions reached in problems 1-3 of this homework and problems 2 and 4 of homework #2? Which analyses would you prefer a priori.?*

***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.*