**3080**

**Biost 515 (Winter 2014)**

**Instructor: Scott Emerson**

**Homework 3**

*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.

**Answer:** Yes. The model has two parameters (the slope and the intercept) and two distinct predictor groups (subjects with high serum LDL and subjects with low serum LDL).

* 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?

**Answer:** The logistic regression model estimates that for subjects with low serum LDL, the odds of death within five years is 0.205, leading to an estimated probability of death within five years of 17.0%. The proportion of subjects with low serum LDL for whom we observed death within five years is also 17.0%, so the estimated five year mortality rate equals the observed five year mortality rate.

* 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?

**Answer:** The logistic regression model estimates that for subjects with high serum LDL, the odds of death within five years is 0.151, leading to an estimated probability of death within five years of 13.1%. The proportion of subjects with high serum LDL for whom we observed death within five years is also 13.1%, so the estimated five year mortality rate equals the observed five year mortality rate.

* 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 five 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. An odds ratio different than 1 was tested using logistic regression analysis and allowing for heteroscedasticity. That is, robust standard errors were used to calculate the 95% confidence intervals for the odds ratio.

**Results:** Of the 618 subjects whose serum LDL was less than or equal to 159 mg/dL, the odds of dying within five years of study enrollment was 0.205, while for the 107 subjects with serum LDL greater than or equal to 160 mg/dL, the odds of five year mortality was 0.151. Based on a 95% confidence interval, this observed odds ratio of 0.735 for the comparison of the high LDL group to the low LDL group would not be unusual if the true odds ratio were anywhere between 0.404 and 1.340. Logistic regression suggests that this observed odds ratio is not statistically different from 1 (p-value = 0.316), therefore, we fail to reject the null hypothesis that the survival probabilities are not associated with serum LDL levels.

These results lend to approximately the same inference that was given on problems 5 and 6 of homework #1. There is, however, a slight difference in the confidence intervals and p-values which were obtained (the point estimates are equivalent). These difference derive from the use of Fisher’s exact test on homework #1 and robust logistic regression on this homework when testing the null hypotheses.

* 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?

**Answer:** The models would all be reparameterized versions of each other. So the answers to parts (a) through (c) would all remain the same.

* 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?

**Answer:** If we were to reverse our analysis to describe the distribution of LDL across groups defined by vital status, we would still have a saturated model, but the model’s two distinct predictor groups would now be subjects who were alive at five years versus subjects who were dead. Moreover, one distinct advantage to working with logistic regression is that for a given sample, the odds ratio based on conditioning for the exposure is mathematically equivalent to the odds ratio based on conditioning for the disease. In other words, if we switch the predictor and response variables, our slope estimate remains the same. So although our odds and probabilities will change (the odds of high LDL given observed death within five years = 0.133 with a corresponding 11.8% probability of high LDL, and the odds of high LDL given survival of at least five years = 0.181 with a corresponding 15.3% probability of high LDL.), our odds ratio will remain the same (0.735).

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.

**Answer:** Yes. The model has two parameters (the slope and the intercept) and two distinct predictor groups (subjects with high serum LDL and subjects with low serum LDL).

* 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?

**Answer:** The linear regression model estimates that for subjects with low serum LDL, the probability of death within five years is 17.0%, leading to an odds of death within five years of 0.205. The proportion of subjects with low serum LDL for whom we observed death within five years is also 17.0%, so the estimated five year mortality rate equals the observed five year mortality rate.

* 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?

**Answer:** The linear regression model estimates that for subjects with high serum LDL, the probability of death within five years is 13.1%, leading to an odds of death within five years of 0.151. The proportion of subjects with high serum LDL for whom we observed death within five years is also 13.1%, so the estimated five year mortality rate equals the observed five year mortality rate.

* 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 number of observed deaths within five 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 difference in risk of death between the two groups was tested using linear regression analysis allowing for the possibility of unequal variances. Robust standard errors were used to calculate 95% confidence intervals for the risk difference.

**Results:** Among the 618 subjects whose serum LDL was less than or equal to 159 mg/dL, the observed five year mortality rate was 17.0%, while for the 107 subjects with serum LDL greater than or equal to 160 mg/dL, the observed five year mortality rate was 13.1%. The absolute difference in risk of death within five years is estimated to be 3.91%, with high LDL subjects having the lower risk. A 95% confidence interval suggests that these results might be typically observed if the true difference was anywhere between 11.0% lower and 3.16% higher, and robust linear regression suggests that this observed difference is not statistically different from 0 (p-value = 0.278). Therefore, we fail to reject the null hypothesis that the survival probabilities are not associated with serum LDL levels.

These results lend to approximately the same inference that was given on problems 5 and 6 of homework #1. There is, however, a slight difference in the confidence intervals and p-values which were obtained (the point estimates are equivalent). The difference derives from the use of Wald statistics to test the null hypothesis in homework #1 and the use of robust linear regression on this assignment.

* 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?

**Answer:** The models would all be reparameterized versions of each other. So the answers to parts (a) through (c) would all remain the same.

* 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?

**Answer:** If we were to reverse our analysis to describe the distribution of LDL across groups defined by vital status, we would still have a saturated model, but the model’s two distinct predictor groups would now be subjects who were alive at five years versus subjects who were dead. However, in linear regression, the slope of our model (unlike logistic regression) will change and we cannot estimate the odds or probabilities within a group in a valid fashion.

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.

**Answer:** Yes. The model has two parameters (the slope and the intercept) and two distinct predictor groups (subjects with high serum LDL and subjects with low serum LDL).

* 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?

**Answer:** The Poisson regression model estimates that for subjects with low serum LDL, the probability of death within five years is 17.0%, leading to an odds of death within five years of 0.205. The proportion of subjects with low serum LDL for whom we observed death within five years is also 17.0%, so the estimated five year mortality rate equals the observed five year mortality rate.

* 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?

**Answer:** The Poisson regression model estimates that for subjects with high serum LDL, the probability of death within five years is 13.1%, leading to an odds of death within five years of 0.151. The proportion of subjects with high serum LDL for whom we observed death within five years is also 13.1%, so the estimated five year mortality rate equals the observed five year mortality rate.

* 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 number of observed deaths within five 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 ratios of the probability of death within five years between the two groups was tested using Poisson regression analysis allowing for the possibility of unequal variances. Robust standard errors were used to calculate 95% confidence intervals for the risk ratio.

**Results:** Among the 618 subjects whose serum LDL was less than or equal to 159 mg/dL, the observed five year mortality rate was 17.0%, while for the 107 subjects with serum LDL greater than or equal to 160 mg/dL, the observed five year mortality rate was 13.1%. The ratio of these probabilities of death is estimated to be 0.770 for subjects with high LDL compared to those with low LDL and a 95% confidence interval suggests that these results might be typically observed if the true risk ratio was anywhere between 0.458 and 1.294. Poisson regression suggests that this observed ratio is not statistically different from 1 (p-value = 0.324), therefore, we fail to reject the null hypothesis that the survival probabilities are not associated with serum LDL levels.

These results lend to approximately the same inference that was given on problems 5 and 6 of homework #1. There is, however, a slight difference in the confidence intervals and p-values which were obtained (the point estimates are equivalent). These difference derive from the use of Fisher’s exact test on homework #1 and robust logistic regression on this homework to test the null hypotheses.

* 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?

**Answer:** The models would all be reparameterized versions of each other. So the answers to parts (a) through (c) would all remain the same.

* 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?

**Answer:** If we were to reverse our analysis to describe the distribution of LDL across groups defined by vital status, we would still have a saturated model, but the model’s two distinct predictor groups would now be subjects who were alive at five years versus subjects who were dead. However, in Poisson regression, the slope of our model (unlike logistic regression) will change and we cannot estimate the odds within a group in a valid fashion.

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:** Five year mortality was estimated using linear regression and treating serum LDL as a continuous predictor of interest. Differences in survival distributions were tested using a model which allowed for unequal variance through use of the Huber-White sandwich estimator of the standard errors. Confidence intervals were assessed at an alpha=0.05 level of significance.

**Results:** From a linear regression analysis of 725 available observations from a sample of 735 elderly subjects between ages 65 and 99, we estimate an absolute difference in five year mortality of 0.103% for each 1 mg/dL increase in serum LDL, with the higher LDL tending to have longer survival. Based on a 95% confidence interval, we find that observing such an estimated difference is not unusual if the true difference in survival were anywhere between
0.188% and 0.018% for each 1 mg/dL increase (again, with tendency of longer survival for higher LDL) . These results are statistically significant evidence of an association between five year mortality and serum LDL (p-value = 0.017) and we can confidently reject the null hypothesis of no association between serum LDL and survival in favor of a hypothesis that death within 5 years is associated with lower mean serum LDL.

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

**Methods:** Five year mortality was estimated using Poisson regression and treating serum LDL as a continuous predictor of interest. Ratios in survival distributions were tested using a model which allowed for unequal variance through use of the Huber-White sandwich estimator of the standard errors. Confidence intervals were assessed at an alpha=0.05 level of significance.

**Results:** From a Poisson regression analysis of 725 available observations from a sample of 735 elderly subjects between ages 65 and 99, we estimate that for each 1 mg/dL increase in serum LDL, the probability of death within five years decreases by 0.645%. Based on a 95% confidence interval, this observation is not unusual if a 1 mg/dL increase in serum LDL truly decreases the probability of five year morbidity anywhere from 0.112% to 1.175%. These results are statistically significant evidence of an association between five year mortality and serum LDL (p-value = 0.018) and we can confidently reject the null hypothesis of no association between serum LDL and survival in favor of a hypothesis that five year morbidity is associated with lower mean serum LDL.

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

**Methods:** Five year mortality was estimated using logistic regression and treating serum LDL as a continuous predictor of interest. Odds ratios in survival distributions were tested using a model which allowed for unequal variance through use of the Huber-White sandwich estimator of the standard errors. A confidence interval for the odds ratio was assessed at an alpha=0.05 level of significance.

**Results:** From a logistic regression analysis of 725 available observations from a sample of 735 elderly subjects between ages 65 and 99, we estimate that for each 1 mg/dL increase in serum LDL, the odds of five year mortality decreases by 0.774% (OR = 0.992). Based on a 95% confidence interval, we find that this observation is not unusual if the true odds of death within five years were decrease anywhere between 0.125% and 1.42% for each 1 mg/dL increase. These results are statistically significant evidence of an association between five year mortality and serum LDL (p-value = 0.019) and we can confidently reject the null hypothesis of no association between serum LDL and survival in favor of a hypothesis that death within 5 years is associated with lower mean 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*.?

**Answer:** When comparing a model which dichotomizes serum LDL into high and low levels (as we did in problems 1-3) versus one which treats serum LDL continuously, we are able to draw more precise conclusions if we avoid dichotomization. That is, we could reject the null hypothesis of no association between survival probability and serum LDL in favor of a hypothesis that death within 5 years is associated with lower mean serum LDL. When comparing to the models developed in problems 2 and 4 from homework #2, we are able to draw similar conclusions based on reparameterized models and different assumptions on variance between groups.

A priori, I would have preferred the analyses done in the last question of this homework assignment for the following reasons: it is statistically much more precise to not dichotomize a continuous measure; it does not intuitively make as much sense to treat a subject with serum LDL of 161 mg/dL the same as a subject with 190 mg/dL (and not the same as a subject with 159 mg/dL); it is more scientifically relevant to compare the risk of an outcome given a predictor than to compare the risk of predictor given the outcome; and finally, we can avoid making false assumptions on the sample variance if we use a model which allows for heteroscedasticity.