**Biost 518: Applied Biostatistics II**

**Biost 515: Biostatistics II**

Emerson, Winter 2014

**Homework #3**

January 20, 2014

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 is a saturated regression model. LDL is dichotomized to “high” and “low” values with 160 mg/dL being the cutoff point, and 5-year all-cause mortality are dichotomized to either alive for 5 years, or dead within 5 years. Therefore, there are two parameters and two groups and the model is saturated.

* 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?

All 618 patients with LDL < 160mg/dL had recorded measurements. Of these patients, 105 died within 5 years. The estimated odds of death within 5 years of the initial MRI among patients with low LDL is 0.205, or 20.5%. The probability of death within 5 years in this subset of patients is 0.1699, or approximately 17%. Similarly, the observed proportion of patients with low LDL who died within 5 years was 17%. Since both the response and predictor variables are binary, we expect the probability of death to equal the observed proportion of death among patients with low LDL since there was no missing data. The odds of death is related to the proportion by the equation: odds = p/(1-p) so since the proportion, *p*, is sufficiently large, the odds are greater than the 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?

Among the remaining 117 patients a total of 107 had recorded measurements with LDL ≥ 160mg/dL. Of these patients, 14 died within 5 years. The estimated odds of death within 5 years of the initial MRI among patients with high LDL is 0.151, or 15.1%. The probability of death within 5 years in this subset of patients is 0.131, or 13.1%. The observed proportion of patients with high LDL who died within 5 years was 13.1%. Since both the response and predictor variables are binary, we expect the probability of death to equal the observed proportion of death among patients with high LDL. The odds of death is related to the proportion by the equation: odds = p/(1-p) so since the proportion, *p*, is sufficiently large, the odds are greater than the proportion.

* 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: Odds of death within 5 years were evaluated between groups of patients with high LDL (≥ 160 mg/dL) and those with low LDL (< 160 mg/dL). The analysis was performed using logistic regression, with a binary indicator of death within 5 years as the response variable, and dichotomized high LDL groups as the predictor variable. A robust standard error estimator (Huber-White sandwich estimator) was used for the analysis, and 95% confidence intervals (CI) and two-sided p-values were Wald-based estimates with significance at 0.05. Two models were fit, one using high LDL as the predictor to find estimates for odds of death among patients with low LDL (i.e. the intercept provides the estimate for low LDL since “high LDL” = 0), and a second using low LDL as the predictor to find estimates for odds of death among patients with high LDL. The intercept from the models were used to estimate the odds of death, the 95% CI, and standard error for the appropriate groups.

Inference: A total of 618 patients had recorded low LDL levels, and 107 patients had recorded high LDL levels. Data was missing for 10 patients. Among patients with high LDL (≥ 160 mg/dL), the odds of death within 5 years of their MRI is estimated to be 15.1%, which is consistent with a population odds between 8.5% and 26.4%, with 95% confidence, the probability of death in this group is 13.1%, and the observed proportion of patients who died is 13.1%. Among patients with low LDL (< 160 mg/dL), the odds of death within 5 years of their MRI is estimated to be 20.5%, which is consistent with a population odds between 16.6% and 25.3%, with 95% confidence, and the probability of death and observed proportion who died in this group is 17%. The odds ratio is 0.735, which is consistent with the true population odds ratio between 0.404 and 1.34 between groups with low LDL compared to those with high LDL, though the estimate is not statistically significant with p = 0.3158. Assuming significance at 0.05, there is insufficient evidence of a difference in odds of death based on LDL levels, and therefore we fail to reject the null hypothesis assuming no difference between groups.

In problems 5 and 6 of homework #1, the odds of death within 5 years of their MRI was 15.1% for patients with high LDL (≥ 160 mg/dL) and 20.5% for those with low LDL (< 160 mg/dL). The odds ratio was estimated to be 0.735 (95% CI 0.373, 1.36). The p-value was 0.396 and we failed to reject the null hypothesis that survival is not associated with LDL levels. The conclusions from homework #1 are the same as with this analysis using logistic regression, but some values vary slightly. In homework #1, Fisher’s exact test was used to evaluate the odds ratio. The estimated odds of death among patients with high LDL was the same as homework #1 (15.1%). The 95% CI for the estimated odds ratio was narrower using logistic regression. One source of difference is likely reduced precision achieved by dichotomizing the LDL measurements, which is naturally a continuous variable. The differences are also likely due to differing standard errors and slightly narrower confidence intervals determined through regression, which result in part from n-2 degrees of freedom in regression compared to n-1 degrees of freedom in non-regression methods. With narrower confidence intervals, we expect a smaller p-value as occurs here with the logistic regression (p=0.316) compared to Fisher’s exact test for the odds ratio (p=0.396).

* 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?

Since there are two dichotomized groups of LDL, performing the same logistic regression with the response variable of death within 5 years but using low LDL as the predictor variable will not change the results of parts a-c, but will only affect how the computer output is interpreted. The same is true if survival for 5 years us used instead of death within 5 years.

* 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?

Changing LDL to a continuous variable, while keeping the response variable as dichotomized based on 5-year all-cause mortality, results in a model that is no longer saturated. In effect we are changing the sampling scheme from a cohort study where LDL is dichotomized and set, while 5-year mortality is evaluated, to a case control study where the exposure, 5-year mortality, is set and LDL evaluated as a continuous variable. This is a much more awkward analysis for determining the odds of death in groups with high or low LDL. By adjusting the analysis in this way, the scientific question is much more appropriately changed to determining an association with death and mean LDL as in homework #2. That said, the analysis performed in this way give results that for each 1.0 mg/dL increase in ldl the odds of death decreases by 0.77% (95% CI -1.4, -0.12).

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 is a saturated regression model. LDL is dichotomized to “high” and “low” values with 160 mg/dL being the cutoff point, and 5-year all-cause mortality are dichotomized to either alive for 5 years, or dead within 5 years. Therefore, there are two parameters and two groups and the model is saturated.

* 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?

All 618 patients with LDL < 160mg/dL had recorded measurements. Of these patients, 105 died within 5 years. The probability of death within 5 years in this subset of patients is 0.1699, or approximately 17%. The estimated odds of death within 5 years of the initial MRI among patients with low LDL is 0.205, or 20.5%. Similarly, the observed proportion of patients with low LDL who died within 5 years was 17%. Since both the response and predictor variables are binary, we expect the probability of death to equal the observed proportion of death among patients with low LDL since there was no missing data. The odds of death is related to the proportion by the equation: odds = p/(1-p) so since the proportion, *p*, is sufficiently large, the odds are greater than the proportion.

* 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?

Among the remaining 117 patients with LDL ≥ 160mg/dL a total of 107 had recorded measurements. Of these patients, 14 died within 5 years. The probability of death within 5 years in this subset of patients is 0.131, or 13.1%. The estimated odds of death within 5 years of the initial MRI among patients with high LDL is 0.151, or 15.1%. The observed proportion of patients with high LDL who died within 5 years was 13.1%. Since both the response and predictor variables are binary, we expect the probability of death to equal the observed proportion of death among patients with high LDL. The odds of death is related to the proportion by the equation: odds = p/(1-p) so since the proportion, *p*, is sufficiently large, the odds are greater than the proportion.

* 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, or risk, of death within 5 years was evaluated between groups of patients with high LDL (≥ 160 mg/dL) and those with low LDL (< 160 mg/dL). The analysis was performed using linear regression to evaluate the risk difference, with a binary indicator of death within 5 years as the response variable, and dichotomized high LDL groups as the predictor variable. A robust standard error estimator (Huber-White sandwich estimator) was used for the analysis. Two models were fit, one using high LDL as the predictor to find estimates for risk of death among patients with low LDL (i.e. the intercept provides the estimate for low LDL since “high LDL” = 0), and a second using low LDL as the predictor to find estimates for risk of death among patients with high LDL. The intercept from the models were used to estimate the risk of death, the 95% confidence interval (CI), and standard error for the appropriate groups, and the slope was used to predict the difference in probability and the associated 95% confidence interval. Reported p-values are two-sided, with significance at 0.05.

Inference: A total of 618 patients had recorded low LDL levels, and 107 patients had recorded high LDL levels. Data was missing for 10 patients. Among patients with high LDL (≥ 160 mg/dL), the probability of death within 5 years of their MRI is estimated to be 13.1%, which is consistent with a population risk between 6.7% and 19.5%, with 95% confidence. Among patients with low LDL (< 160 mg/dL), the probability of death within 5 years of their MRI is estimated to be 17.0%, which is consistent with a population risk between 14.0% and 20.0%, with 95% confidence. The absolute risk difference is 3.91% higher with low LDL, which is consistent with a true population risk difference from 3.2% lower and 11.0% higher between groups with low LDL compared to those with high LDL, though the estimate is not statistically significant with p = 0.2780. Assuming significance at 0.05, there is insufficient evidence of a difference in odds of death based on LDL levels, and therefore we fail to reject the null hypothesis assuming no difference between groups.

In problems 5 and 6 of homework #1, the probability of death within 5 years of study enrollment was 13.1% for patients with high LDL, and 17.0% for patients with low LDL. The absolute difference in survival probability was 3.91% higher with low LDL (95% CI -3.14%, 10.9%). Using a chi squared test, this difference in probability was not statistically significant (p=0.314) and we failed to reject the null hypothesis that survival is not associated with LDL levels. The conclusions from homework #1 are the same as with this analysis using linear regression, but some values vary slightly. The estimated absolute risk difference with linear regression was the same (3.91%), but the 95% CI for the difference was slightly different narrower and the p-value slightly lower (p=0.278) than in homework #1. One source of difference is likely reduced precision achieved by dichotomizing the LDL measurements, which is naturally a continuous variable. The differences are also likely due to differing standard errors, which result in part from n-2 degrees of freedom in regression compared to (# of cells-1) degrees of freedom in in chi-squared methods. With narrower confidence intervals, we expect a smaller p-value as occurs here with linear regression (p=0.278) compared to the chi squared test for the difference in probabilities (p=0.314).

* 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?

Since there are two dichotomized groups of LDL, performing the same logistic regression with the response variable of death within 5 years but using low LDL as the predictor variable will not change the results of parts a-c, but will only affect how the computer output is interpreted. The same is true if survival for 5 years us used instead of death within 5 years.

* 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?

Changing LDL to a continuous variable, while keeping the response variable as dichotomized based on 5-year all-cause mortality, results in a model that is no longer saturated. In effect we are changing the sampling scheme from a cohort study where LDL is dichotomized and set, while 5-year mortality is evaluated, to a case control study where the exposure, 5-year mortality, is set and LDL evaluated as a continuous variable. This is a much more awkward analysis for determining the probability of death in groups with high or low LDL. By adjusting the analysis in this way, the scientific question is much more appropriately changed to determining an association with death and mean LDL as in homework #2. That said, the analysis performed in this way give results that for each 1.0 mg/dL increase in ldl the probability of death decreases by 0.10% (95% CI -0.18, -0.02).

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 is a saturated regression model. LDL is dichotomized to “high” and “low” values with 160 mg/dL being the cutoff point, and 5-year all-cause mortality are dichotomized to either alive for 5 years, or dead within 5 years. Therefore, there are two parameters and two groups and the model is saturated.

* 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?

All 618 patients with LDL < 160mg/dL had recorded measurements. Of these patients, 105 died within 5 years. The probability of death within 5 years in this subset of patients is 0.1699, or approximately 17%. The estimated odds of death within 5 years of the initial MRI among patients with low LDL is 0.205, or 20.5%. Similarly, the observed proportion of patients with low LDL who died within 5 years was 17%. Since both the response and predictor variables are binary, we expect the probability of death to equal the observed proportion of death among patients with low LDL since there was no missing data. The odds of death is related to the proportion by the equation: odds = p/(1-p) so since the proportion, *p*, is sufficiently large, the odds are greater than the proportion.

* 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?

Among the remaining 117 patients with LDL ≥ 160mg/dL a total of 107 had recorded measurements. Of these patients, 14 died within 5 years. The probability of death within 5 years in this subset of patients is 0.131, or 13.1%. The estimated odds of death within 5 years of the initial MRI among patients with high LDL is 0.151, or 15.1%. The observed proportion of patients with high LDL who died within 5 years was 13.1%. Since both the response and predictor variables are binary, we expect the probability of death to equal the observed proportion of death among patients with high LDL. The odds of death is related to the proportion by the equation: odds = p/(1-p) so since the proportion, *p*, is sufficiently large, the odds are greater than the proportion.

* 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 risk of death within 5 years was evaluated between groups of patients with high LDL (≥ 160 mg/dL) and those with low LDL (< 160 mg/dL). The analysis was performed using poisson regression (with maximum likelihood estimation) to evaluate the risk ratio, with a binary indicator of death within 5 years as the response variable, and dichotomized high LDL groups as the predictor variable. A robust standard error estimator (Huber-White sandwich estimator) was used for the analysis. P-values and 95% confidence intervals (CI) are Wald-based estimates. Two models were fit, one using high LDL as the predictor to find estimates for risk of death among patients with low LDL (i.e. the intercept provides the estimate for low LDL since “high LDL” = 0), and a second using low LDL as the predictor to find estimates for risk of death among patients with high LDL. The intercept from the models were used to estimate the risk of death, the 95% CI, and standard error for the appropriate groups, and the slope was used to predict the risk ratio and the associated 95% confidence interval. Reported p-values are two-sided, with significance at 0.05.

Inference: A total of 618 patients had recorded low LDL levels, and 107 patients had recorded high LDL levels. Data was missing for 10 patients. Among patients with high LDL (≥ 160 mg/dL), the probability of death within 5 years of their MRI is estimated to be 13.1%, which is consistent with a population risk between 6.7% and 19.5%, with 95% confidence. Among patients with low LDL (< 160 mg/dL), the probability of death within 5 years of their MRI is estimated to be 17.0%, which is consistent with a population risk between 14.3% and 20.2%, with 95% confidence. The risk ratio is 0.7701 between groups, with higher risk with low LDL, which is consistent with a true population risk difference from 0.458 to 1.294 higher between groups. This estimate is not statistically significant with p = 0.3237. Assuming significance at 0.05, there is insufficient evidence of a difference in risk ratio based on LDL levels, and therefore we fail to reject the null hypothesis assuming no difference between groups.

In problems 5 and 6 of homework #1, the probability of death within 5 years of study enrollment was 13.1% for patients with high LDL, and 17.0% for patients with low LDL. The odds of death within 5 years of their MRI was 15.1% for patients with high LDL (≥ 160 mg/dL) and 20.5% for those with low LDL (< 160 mg/dL). The odds ratio was estimated to be 0.735 (95% CI 0.373, 1.36). The p-value was 0.396 and we failed to reject the null hypothesis that survival is not associated with LDL levels. The conclusions from homework #1 are the same as with this analysis using poisson regression, but some values vary slightly. In homework #1, Fisher’s exact test was used to evaluate the odds ratio. The 95% CI for the estimated odds ratio was narrower using poisson regression, and the p-value is correspondingly smaller. One source of difference is likely reduced precision achieved by dichotomizing the LDL measurements, which is naturally a continuous variable. The differences are also likely due to differing standard errors and slightly narrower confidence intervals determined through regression, which result in part from n-2 degrees of freedom in regression compared to n-1 degrees of freedom in non-regression methods. With narrower confidence intervals, we expect a smaller p-value as occurs here with the poisson regression (p=0.3237) compared to Fisher’s exact test for the odds ratio (p=0.396).

* 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?

Since there are two dichotomized groups of LDL, performing the same poisson regression with the response variable of death within 5 years but using low LDL as the predictor variable will not change the results of parts a-c, but will only affect how the computer output is interpreted. The same is true if survival for 5 years us used instead of death within 5 years.

* 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?

Changing LDL to a continuous variable, while keeping the response variable as dichotomized based on 5-year all-cause mortality, results in a model that is no longer saturated. In effect we are changing the sampling scheme from a cohort study where LDL is dichotomized and set, while 5-year mortality is evaluated, to a case control study where the exposure, 5-year mortality, is set and LDL evaluated as a continuous variable. This is a much more awkward analysis for determining the probability of death in groups with high or low LDL. By adjusting the analysis in this way, the scientific question is much more appropriately changed to determining an association with death and mean LDL as in homework #2. That said, the analysis performed in this way give results that for each 1.0 mg/dL increase in ldl the probability of death decreases by 0.00645 (95% CI -0.00112, -0.0117).

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: The risk of death within 5 years was evaluated between groups of patients with high LDL (≥ 160 mg/dL) and those with low LDL (< 160 mg/dL). The analysis was performed using linear regression to evaluate the risk difference, with a binary indicator of death within 5 years as the response variable, and continuous serum LDL measurements as the predictor variable. A robust standard error estimator (Huber-White sandwich estimator) was used for the analysis. The slope was used to predict the risk difference and the associated 95% confidence interval (CI). Reported p-values are two-sided, with significance at 0.05.

Inference: A total of 618 patients had recorded low LDL levels, and 107 patients had recorded high LDL levels. Data was missing for 10 patients. From linear regression, it is estimated that for each 1 mg/dL increase in serum LDL, there is a 0.103% decrease in the risk of death. This is consistent, with 95% confidence, with a true population risk decrease between 0.0185% and 0.188%. This result is statistically significant with p=0.0171, giving sufficient evidence to reject the null hypothesis that 5-year mortality is not associated with serum LDL levels.

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

Methods: The risk of death within 5 years was evaluated between groups of patients with high LDL (≥ 160 mg/dL) and those with low LDL (< 160 mg/dL). The analysis was performed using poisson regression (with maximum likelihood estimation) to evaluate the risk ratio, with a binary indicator of death within 5 years as the response variable, and continuous serum LDL measurements as the predictor variable. A robust standard error estimator (Huber-White sandwich estimator) was used for the analysis. P-values and 95% confidence intervals (CI) are Wald-based estimates. The slope was used to predict the risk ratio and the associated 95% confidence interval. Reported p-values are two-sided, with significance at 0.05.

Inference: A total of 618 patients had recorded low LDL levels, and 107 patients had recorded high LDL levels. Data was missing for 10 patients. From poisson regression, it is estimated that for each 1 mg/dL increase in serum LDL, there is a 0.00645 (0.65%) decrease in the probability of death. This estimate is consistent with a true population risk difference between 0.00112 (0.11%) and 0.0117 (1.2%) lower for each LDL unit increase. This suggests the relative risk of death decreases with increasing serum LDL. This estimate is statistically significant with p = 0.0177. Assuming significance at 0.05, there is sufficient evidence of a difference in risk ratio based on LDL levels, and therefore we reject the null hypothesis assuming no difference between groups.

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

Methods: Odds of death within 5 years were evaluated between groups of patients with high LDL (≥ 160 mg/dL) and those with low LDL (< 160 mg/dL). The analysis was performed using logistic regression, with a binary indicator of death within 5 years as the response variable, continuous serum LDL as the predictor variable. A robust standard error estimator (Huber-White sandwich estimator) was used for the analysis, and 95% confidence intervals (CI) and two-sided p-values were Wald-based estimates with significance at 0.05. The slope was used to estimate the odds ratio, the 95% CI, and standard error. Reported p-values are two-sided, with significance at 0.05.

Inference: A total of 618 patients had recorded low LDL levels, and 107 patients had recorded high LDL levels. Data was missing for 10 patients. From logistic regression, it is estimated that for each 1 mg/dL increase in serum LDL, the odds ratio decreases by 0.00774. This estimate is consistent with a true population decrease in the odds ratio between 0.00126 (0.13%) and 0.0143 (1.4%) lower for each LDL unit increase. This suggests the relative odds of death decreases with increasing serum LDL. This estimate is statistically significant with p = 0.0194. Assuming significance at 0.05, there is sufficient evidence of a difference in risk ratio based on LDL levels, and therefore we reject the null hypothesis assuming no difference between groups.

* 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*?

In prior problems, the results were not statistically significant (p>0.05) so there was insufficient evidence to reject the null hypothesis that 5-year all-cause mortality is not associated with serum LDL levels. In those problems LDL was dichotomized into clinically relevant groups with a cutoff point of 160 mg/dL serum LDL. By dichotomizing, some precision is lost, and the trend between small intervals (i.e. 1 mg/dL LDL). When the data are analyzed by keeping LDL as a continuous variable the models predict incremental changes in the probability of death with increasing LDL. From a clinical standpoint, it is not informative to know changes in the risk of death per 1 mg/dL increase in LDL since we have broader epidemiologic cutoff points of LDL categories that are used in practice. To answer the question of whether or not there is an association between 5-year all-cause mortality and serum LDL, I would chose to determine the relative risk among patients within groups based upon clinically relevant LDL measurements corresponding to “ideal” (LDL < 100 mg/dL), “low” (LDL 100-129 mg/dL), “moderate” (LDL 130-159 mg/dL), “high” (LDL 160-189 mg/dL) and “very high” (LDL ≥ 190 mg/dL). People often prefer predicting death in terms of risk or probability of death, so, *a priori*, I would fit a linear regression model to the data according to the categories defined here. Alternatively, I would use logistic regression to determine the odds of death in each group.