**Biost 515/518 HW#3**

January 20, 2014

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.)

***Parts (a) – (d) of this question use a logistic regression model with an indicator of high LDL as the predictor variable and an indicator of death within 5 years as the response variable, with a robust estimation of the standard error.***

* 1. Is this a saturated regression model? Explain your answer.

**In this model, there are two groups (serum LDL < 160 mg/dL and serum LDL ≥ 160 mg/dL) being modeled with two regression parameters (β0 and β1, or the equivalent of the slope and intercept). Therefore, this is a saturated model.**

* 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?

**For subjects with low LDL, the exponentiated intercept estimates that the odds of dying within 5 years is 0.205. The estimated probability of dying within 5 years is 0.2047/1.2047 = 0.170. Of the 618 subjects with low LDL, 105 died within five years, for an observed proportion of 105/618 = 0.170. Therefore, the estimated odds is about 20% higher (relative difference) than the observed proportion dying within 5 years, but the estimated proportion matches the observed proportion exactly.**

* 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?

**For subjects with high LDL, the estimated odds of dying within 5 years is 0.151. The estimated probability of dying within 5 years is 0.131. In our data, of the 107 subjects with high LDL, 14 died, for an observed proportion of 0.131. Therefore, the estimated probability of death within 5 years is exactly the same as the observed proportion among subjects with high LDL, and the odds are about 15% higher (relative difference) than the observed value.**

* 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?

**A logistic regression analysis of 725 available observations from a sample of 735 subjects was used to compare the odds of subjects dying within 5 years of study enrollment between subjects who had serum LDL greater than or equal to 160 mg/dL (“high”) and subjects whose serum LDL was 159 mg/dL or less (“low”). A robust estimator of the standard error was used, and odds were compared using the odds ratio of the high LDL group to the low LDL group.**

**Among the 618 subjects with low LDL, the odds of dying within 5 years from study enrollment was 0.205, while among the 107 subjects with high LDL, the odds of dying within 5 years was 0.151. From logistic regression analysis, we estimate that the odds ratio is 0.735 for the comparison of the high LDL group to the low LDL group; that is, the odds of dying within 5 years are 26.5% lower among subjects with high LDL. A 95% confidence interval suggests that this would not be unusual if the group with high LDL has odds of death within 5 years between 59.7% lower and 34.0% higher than the odds of death within 5 years among subjects with low LDL. These differences are not statistically significant (two-sided p=0.316), so we cannot reject our null hyopthesis that the odds of death within 5 years of study enrollment are equal among subjects with low versus high LDL. The point estimates and the odds ratio from the logistic regression model exactly match the estimates calculated in problem 6 from homework #1; the confidence intervals are slightly narrower (0.403 to 1.340 compared to 0.373 to 1.36), and our p-value of 0.316 is lower than the previously calculated p-value of 0.396. This difference is likely due to differences in calculation of the standard error: here, we use the robust standard error, and in problem 6 we used Woolf’s method. However, both methods lead to the same conclusion.**

* 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?

**All of the estimates would have been the same regardless of the model parameterization, though they would have been located differently in the model. These models are all saturated.**

* 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?

**If we had fit a logistic regression model using an indicator of death within 5 years as the predictor variable and an indicator of high LDL as the response variable, again using a robust estimation of the standard error, our odds ratios would have been the same. The model would still have been saturated.**

**Our estimates would have been the same as (b) and (c) above for the probabilities requested. However, the natural estimates based on this model would have been, for example, the odds of having high LDL among a population of subjects who survive at least 5 years past study enrollment (0.181) and among a population of subjects who die within 5 years (0.133).**

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.)

***Parts (a) – (d) of this question use a linear regression model with an indicator of high serum LDL as the predictor variable and an indicator of death within 5 years as the response variable, using the robust standard error estimate.***

* 1. Is this a saturated regression model? Explain your answer.

**In this model, there are two groups (serum LDL < 160 mg/dL and serum LDL ≥ 160 mg/dL) being modeled with two regression parameters (β0 and β1, or the slope and intercept). Therefore, this is a saturated model.**

* 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 0.170, and the estimated odds of dying within 5 years is 0.1699/0.8301 = 0.205. The estimated probability of death within 5 years among subjects with low LDL exactly matches the observed proportion of subjects with low LDL that died within 5 years. The estimated odds of death within 5 years among subjects with low LDL is about 20% higher (relative difference) than the observed 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 high LDL dying within 5 years?

**For subjects with high LDL, the estimated probability of dying within 5 years is 0.1699 – 0.0391 = 0.131. The estimated odds of dying within 5 years is 0.151. Therefore, the estimated probability of death within 5 years is exactly the same as the observed proportion among subjects with high LDL, and the odds are about 15% higher (relative difference) than the observed value.**

* 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?

**A linear regression analysis of 725 available observations from a sample of 735 subjects was used to compare the probability of death between subjects with high serum LDL (≥160 mg/dL) and subjects with low serum LDL, using the robust estimate of the standard error. The risk difference was used to compare the probability of death and was calculated as (risk among subjects with high LDL) – (risk among subjects with low LDL).**

**Among the 618 subjects with low LDL, the probability of dying within 5 years of study enrollment was 0.170, while among the 119 subjects with high LDL, the probability of dying within 5 years was 0.131. From linear regression analysis, we estimate that the probability of death within 5 years is 3.91% lower (absolute difference) among the high LDL group than among the low LDL group. A 95% confidence interval suggests that this observed difference would not be unusual if the true difference is between 11.0% lower and 3.16% higher in the subjects with high LDL relative to the subjects with low LDL. This is not a statistically significant result (two-sided p=0.278), so we cannot reject our null hyopthesis that the probability of death within 5 years of study enrollment is equal among subjects with low versus high serum LDL. The point estimates and risk difference from the linear regression model exactly match the estimates calculated in problem 5 from homework #1; the confidence interval is slightly wider (-0.110 to 0.0316 compared to -0.109 to 0.0314). This difference is likely due to differences in calculation of the standard error: here we again use the robust standard error, while before we used the t-test allowing for unequal variances. We note that the difference is smaller than in question 1 above, since both tests adjust for the possibility of unequal variances. Our p-value of 0.278 is lower than the p-value of 0.314 calculated using a chi squared test, again because of different standard error and test statistic calculations.**

* 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?

**All of the estimates would have been the same regardless of the model parameterization, though they would have been located differently in the model. These models are all saturated.**

* 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?

**If we had fit a linear regression model using an indicator of death within 5 years as the predictor variable and an indicator of high LDL as the response variable, again using a robust estimation of the standard error, the model would still have been saturated.**

**Our estimates would have been the same as (b) and (c) above for the probabilities requested. However, the natural estimates based on this model would have been, for example, the probability of having high LDL among a population of subjects who survive at least 5 years past study enrollment (0.153) or among a population of subjects who die within 5 years (0.118).**

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.)

***Parts (a) – (d) of this question use a Poisson regression model with an indicator of high serum LDL as the predictor variable and an indicator of death within 5 years as the response variable, using the robust standard error estimate.***

* 1. Is this a saturated regression model? Explain your answer.

**In this model, there are two groups (serum LDL < 160 mg/dL and serum LDL ≥ 160 mg/dL) being modeled with two regression parameters (β0 and β1 in the Poisson regression model). Therefore, this is a saturated model.**

* 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 exp(-1.7725) = 0.170, and the estimated odds of dying within 5 years is 0.1699/0.8301 = 0.205. The estimated probability of death within 5 years among subjects with low LDL exactly matches the observed proportion of subjects with low LDL that died within 5 years (105/618 = 0.170). The estimated odds of death within 5 years among subjects with low LDL is about 20.4% higher (relative difference) than the observed 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 high LDL dying within 5 years?

**For subjects with high LDL, the estimated probability of dying within 5 years is exp(-0.2612) \* exp(-1.7725) = 13.1%. The estimated odds of dying within 5 years is 0.151. Again, the estimated probability of death within 5 years is exactly the same as the observed proportion among subjects with high LDL, and the odds are about 15% higher (relative difference) than the observed value.**

* 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?

**A Poisson regression analysis of 725 available observations from a sample of 735 subjects was used to compare the probability of death between subjects with high serum LDL (≥160 mg/dL) and subjects with low serum LDL, using the robust estimate of the standard error. The risk ratio was used to compare the probability of death and was calculated as (risk among subjects with high LDL) / (risk among subjects with low LDL). A risk ratio different from 1 was tested using Poisson regression, and p-values and 95% confidence intervals were computed for the slope in the regression model.**

**Among the 618 subjects with low LDL, the probability of dying within 5 years of study enrollment was 0.170, while among the 119 subjects with high LDL, the probability of dying within 5 years was 0.131. Based on the 95% confidence interval, the estimated risk ratio of 0.770 for the comparison of the high LDL group to the low LDL group would not be considered unusual if the true risk ratio were between 0.458 and 1.294. A two-sided p-value of 0.324 suggests that we cannot with high confidence reject our null hypothesis that the risk of 5 year mortality is not associated with LDL levels. The within-group estimates from the Poisson regression model exactly match the estimates calculated in problem 5 from homework #1. However, the risk ratio was calculated here rather than the risk difference, unlike in homework 1 problem 5. The two-sided p-value from this test is slightly higher than the p-value obtained using the chi-squared test for problems 5 and 6 (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?

**All of the estimates would have been the same regardless of the model parameterization, though they would have been located differently in the model. These models are all saturated.**

* 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?

**If we had fit a Poisson regression model using an indicator of death within 5 years as the predictor variable and an indicator of high LDL as the response variable, again using a robust estimation of the standard error, the model would still have been saturated.**

**Our estimates would have been the same as (b) and (c) above for the probabilities requested. However, the natural estimates based on this model would have been, for example, the probability of having high LDL among a population of subjects who survive at least 5 years past study enrollment (0.153) or among a population of subjects who die within 5 years (0.118), and the risk ratio based on these values.**

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 proportion of subjects dying within 5 years of study enrollment were compared across serum LDL measurements using linear regression analysis with a robust estimate of the standard error. The slope represents the difference in the probability of death between groups differing by 1 mg/dL in serum LDL, and inference on the slope (to obtain a two-sided p-value and 95% confidence interval) tests whether there is a linear trend between serum LDL levels and the proportion of subjects dying within 5 years of study enrollment.**

**Results: From a linear regression analysis of 725 available observations from a sample of 735 subjects, we estimate a difference in probability of death within 5 years of -1.03% per 10 mg/dL increase in LDL. Based on the 95% confidence interval, this estimated difference would not be considered unusual if the true difference was between -1.89% and -0.185% per 10 mg/dL increase in serum LDL. A two-sided p-value of 0.0171 suggests that we can with high confidence reject our null hypothesis of no linear trend between probability of death and serum LDL levels and conclude that higher LDL is associated with reduced probability of death within 5 years of study enrollment.**

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

**Methods: The proportion of subjects dying within 5 years of study enrollment were compared across serum LDL measurements using Poisson regression analysis with a robust estimate of the standard error. The exponentiated slope represents the ratio in the probability of death between groups differing by 1 mg/dL in serum LDL, and inference on the slope (to obtain a two-sided p-value and 95% confidence interval) tests whether this ratio is equal to 1, indicating equal risk of death across serum LDL levels.**

**Results: From Poisson regression analysis of 725 available observations from a sample of 735 subjects, we estimate that for each 1 mg/dL increase in serum LDL levels, the probability of death within 5 years is 0.645% lower. Based on a 95% confidence interval, this observation is not unusual if the true probability of death with a 1 mg/dL difference in serum LDL levels is between 1.17% and 0.112% lower in the group with higher LDL. The two-sided p-value of 0.0177 suggests that with high confidence, we can reject the null hypothesis of equal risk of death across serum LDL levels and conclude that higher serum LDL is associated with a reduced risk of death within 5 years of study enrollment. However, based on the confidence interval, this risk reduction is quite small, so only questionably scientifically relevant.**

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

**Methods: The odds of death within 5 years of study enrollment were compared across serum LDL measurements using logistic regression analysis with a robust estimate of the standard error. The exponentiated slope represents the ratio comparing the odds of death between groups differing by 1 in serum LDL, and inference on the slope (to obtain a two-sided p-value and 95% confidence interval) tests whether this ratio is equal to 1, indicating equal odds of death across serum LDL levels.**

**Results: From logistic regression analysis of 725 available observations from a sample of 735 subjects, we estimate that for each 1 mg/dL difference in serum LDL, the odds of death within 5 years of study enrollment is 0.774% lower in the group with 1 mg/dL higher LDL. A 95% confidence interval suggests that this observation is not unusual if the odds of death within 5 years of study enrollment is between 1.42% lower and 0.125% lower in the group with higher LDL. This estimate is statistically significant (two-sided p=0.019), so we can with high confidence reject our null hypothesis of equal odds of death across serum LDL levels and conclude that higher serum LDL is associated with reduced odds of death within 5 years of study enrollment.**

* 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*?

**Based on each of the models that use the continuous measurement of LDL, statistically significant results allow us to reject our null hypothesis and conclude that higher LDL levels are associated with lower 5 year all-cause mortality rates. In contrast, each of the models treating LDL as a dichotomous variable give statistically insignificant results. Problem 4 of homework 2 gives the same (or a very similar) p-value as the model in problem 4a on on this assignment – 0.0171 and 0.017 – since it is a regression model that allows possibility of unequal variances. Problem 2 of homework 2 assumes equal variances but still comes to the same conclusion as the linear regression models here.**

***A priori*, we would need to consider study design, desire for public health impact, frequency of the event, and expected effect modification or nonlinearity. Since our data come from a cross-sectional survey, the odds ratio, the relative risk, and the risk difference are all valid. I would prefer not to dichotomize LDL, since statistical precision is lost when dichotomizing a continuous variable. This suggests that one of the models in question 4 is most appropriate. The odds ratio is not very easily interpretable to the public. Thus, though the odds ratio allows a greater possibility of avoiding major nonlinearities and effect modification because it is not subject to the range constraints of probabilities, I would probably choose to make the additional assumptions in favor of the public health impact and greater interpretability of risk ratios and risk differences. Between these two models (the linear and Poisson regression models from #4 (a) and (b)), I prefer the linear regression model; death is not a rare enough event in this elderly population that comparing ratios is significantly superior, and using linear regression to estimate the risk difference allows the results to be scaled to estimate the people affected in a larger population.**