**Biost 518: Applied Biostatistics II**

**Biost 515: Biostatistics II**

Emerson, Winter 2014

30/108

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

**Yes, this is a saturated regression model because the number of possible values in the predictor of interest modeled is the same as the number of parameters resulting from the regression analysis. In this case, the predictor variable (High LDL) has two possible values (High LDL and Low LDL) and the regression produces two parameters (slope and intercept).**

Complete and correct.

3/3

* 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 estimated odds of dying within 5 years is .205. The estimated probability of dying within 5 years is .17. The observed proportion of subjects with low LDL dying within 5 years is also .17 and exactly matches the estimated probability of dying within 5 years. This is then equal to the estimated odds/(1+estimated odds).**

The reported estimates are correct, but method of acquiring the odds from a logistic regression was not stated. (-1) Minor note: report 0.170 instead of 0.17.

2/3

* 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 .158. The estimated probability of dying within 5 years is .137. The observed proportion of subjects with high LDL dying within 5 years is also .137 and exactly matches the estimated probability of dying within 5 years. This is then equal to the estimated odds/(1+estimated odds).**

The reported figures are incorrect (-1). It is possible author incorrectly interpreted the slope of the logistic regression (-0.15863) as odds. Sample estimates are also incorrect: the sample proportion of subjects with high LDL dying within 5 years is 0.131, while the sample odds is 0.151. Method for acquiring estimates not stated (-1).

1/3

* 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?

**From logistic regression analysis, I estimate that those individuals with high LDL (defined as LDL greater than or equal to 160 mg/dL), the odds of death are 32.7% lower than they are for those individuals with low LDL, though this is not statistically significant (P=.376). A 95% CI suggests that this observation is not unusual if subjects with high LDL might have odds of death that was anywhere from 56.13% lower to 36.53% higher than the group with low LDL.**

Methods for the inference were not stated (-2). Differences and sources of differences with problem 5 of homework 1 were not reported, nor for problem 6 (-5). Although it is mentioned that a result is not significant, the consequence on the null hypothesis is not stated (-1). Since the logistic regression considers odds ratios, the analysis would have been most appropriately based on the odds ratio in this case. (-1).

1/10

* 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?

**Because this data is dichotomized between high LDL and low LDL, the change in whether high LDL was used as a predictor or low LDL would not have changed the results that we are able to draw from the data. Similarly, given that the 5-year-survival was dichotomous as well, shifting from death within 5 years to survival for at least 5 years would not have changed the results that we were able to draw from the data. If we had used an indicator of survival for at least 5 years, we could have calculated the same results by substituting 1-p for p.**

Author states correctly the equivalences. The complement does not produce similar fittings, but the inverse does. (-1)

2/3

* 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?

**Because this data is based on a cohort study, such an inversion would have produced invalid results. While we would have been able to calculate an odds ratio, we would not have been able to calculate any other calculations.**

Author correctly states that the odds ratio would still be found in the model. However, the connection to Logistic Regression invariance is not stated (-1). Regardless of cohort status, the answer should be specific to the model. The calculation becomes a minor point to the lack of equivalence in interpretation from one model to the other (-1).

1/3

1. Perform a statistical regression analysis evaluating an association between serum LDL and 5 year all-cause mortality by comparing the differences in the probability of death within 5 years across groups defined by whether the subjects have high serum LDL (“high” = LDL > 160 mg/dL). In your regression model, use an indicator of death within 5 years as your response variable, and use an indicator of high LDL as your predictor. (Only give a formal report of the inference where asked to.)
	1. Is this a saturated regression model? Explain your answer.

**Yes, this is a saturated regression model because the number of possible values in the predictor of interest modeled is the same as the number of parameters resulting from the regression analysis. In this case, the predictor variable (High LDL) has two possible values (High LDL and Low LDL) and the regression produces two parameters (slope and intercept).**

3/3

* 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 .17. The estimated odds of dying within 5 years is .205. The observed proportion of subjects with low LDL dying within 5 years is also .17 and exactly matches the estimated probability of dying within 5 years. This is then equal to the estimated odds/(1+estimated odds).**

The method of attaining estimates is not stated (-1). The linear regression needed to be stated in some way. Otherwise, the figures are correct. Again, minor note: 0.170 not 0.17.

2/3

* 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 .137. The estimated odds of dying within 5 years is .158. The observed proportion of subjects with high LDL dying within 5 years is also .137 and exactly matches the estimated probability of dying within 5 years. This is then equal to the estimated odds/(1+estimated odds).**

The reported figures are incorrect (-1). Sample estimates are also incorrect: the sample proportion of subjects with high LDL dying within 5 years is 0.131, while the sample odds is 0.151. Method for acquiring estimates not stated (-1). Since the same mistake as before was done, it is possible that author’s subsets of data were incorrectly created.

1/3

* 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?

**From Linear regression analysis, I estimate that those individuals with high LDL (defined as LDL greater than or equal to 160 mg/dL), have a the risk of death that is .033 lower than it is for those individuals with low LDL, though this is not statistically significant (P=.376). A 95% CI suggests that this observation is not unusual if subjects with high LDL might have a difference in risk of death that was anywhere from .107 lower to .04 higher than the group with low LDL.**

Methods for the inference were not stated (-2). Differences and sources of differences with problem 5 of homework 1 were not reported, nor for problem 6 (-5). Although it is mentioned that a result is not significant, the consequence on the null hypothesis is not stated (-1). The analysis completed has incorrect point estimates, p-value, and confidence interval. (-2)

0/10

* 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?

**Because this data is dichotomized between high LDL and low LDL, the change in whether high LDL was used as a predictor or low LDL would not have changed the results that we are able to draw from the data. Similarly, given that the 5-year-survival was dichotomous as well, shifting from death within 5 years to survival for at least 5 years would not have changed the results that we were able to draw from the data. If we had used an indicator of survival for at least 5 years, we could have calculated the same results by substituting 1-p for p.**

Author states correctly the equivalences and adjustments.

3/3

* 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?

**Because this data is based on a cohort study, such an inversion would have produced invalid results. While we would have been able to calculate an odds ratio, we would not have been able to calculate any other calculations.**

This answer was copy pasted without adjustment for difference in the model considered nor the point estimates of interest. (-3)

0/3

1. Perform a statistical regression analysis evaluating an association between serum LDL and 5 year all-cause mortality by comparing the ratios of the probability of death within 5 years across groups defined by whether the subjects have high serum LDL (“high” = LDL > 160 mg/dL). In your regression model, use an indicator of death within 5 years as your response variable, and use an indicator of high LDL as your predictor. (Only give a formal report of the inference where asked to.)
	1. Is this a saturated regression model? Explain your answer.

**Yes, this is a saturated regression model because the number of possible values in the predictor of interest modeled is the same as the number of parameters resulting from the regression analysis. In this case, the predictor variable (High LDL) has two possible values (High LDL and Low LDL) and the regression produces two parameters (slope and intercept).**

Correct.

3/3

* 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 .17. The estimated odds of dying within 5 years is .205. The observed proportion of subjects with low LDL dying within 5 years is also .17 and exactly matches the estimated probability of dying within 5 years. This is then equal to the estimated odds/(1+estimated odds).**

The method of attaining estimates is not stated (-1). The poisson regression needed to be stated in some way.

2/3

* 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 .137. The estimated odds of dying within 5 years is .158. The observed proportion of subjects with high LDL dying within 5 years is also .137 and exactly matches the estimated probability of dying within 5 years. This is then equal to the estimated odds/(1+estimated odds).**

The method of attaining estimates is not stated (-1). Reported values are not correct (-1).

1/3

* 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?

**From Poisson regression analysis, I estimate that for those individuals with high LDL (defined as LDL greater than or equal to 160 mg/dL), the probability of death is 19.51% lower than it is for those individuals with low LDL, though this is not statistically significant (P=0.419). A 95% CI suggests that this observation is not unusual if subjects with high LDL might have a risk of death that was anywhere from 52.43% lower to 36.20% higher than the group with low LDL.**

Methods for the inference were not stated (-2). Differences and sources of differences with problem 5 of homework 1 were not reported, nor for problem 6 (-5). Although it is mentioned that a result is not significant, the consequence on the null hypothesis is not stated (-1). The Poisson regression results are incorrectly interpreted. Without knowing the method, it is not clear what analysis was done (-2).

0/10

* 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?

**Because this data is dichotomized between high LDL and low LDL, the change in whether high LDL was used as a predictor or low LDL would not have changed the results that we are able to draw from the data. Similarly, given that the 5-year-survival was dichotomous as well, shifting from death within 5 years to survival for at least 5 years would not have changed the results that we were able to draw from the data. If we had used an indicator of survival for at least 5 years, we could have calculated the same results by substituting 1-p for p.**

The aspect of reparameterization is correctly acknowledged. However, the reasoning for the complement is not given in the answer (-1). In the linear regression it is obvious enough, but in the Poisson regression it is not, just as in the logistic regression.

2/3

* 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?

**Because this data is based on a cohort study, such an inversion would have produced invalid results. While we would have been able to calculate an odds ratio, we would not have been able to calculate any other calculations.**

This answer was copy pasted without adjustment for difference in the model considered nor the point estimates of interest. (-3)

0/3

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).

**From Linear regression analysis, I estimate that the mean LDL among subjects dying within 5 years is .001mg/dL lower than in subjects who survived at least 5 years. The estimated SE is .0004 mg/dL which leads to a 95% CI for the true difference in mean LDL between .0019 and .0018 mg/dL, with lower mean LDL in subjects who die within 5 years. Based on a two-sided p value of 0.017, I reject the null hypothesis that surviving subjects would tend toward a higher mean serum LDL.**

Methods were not stated (-2). Need more digits in all estimates except the p-value (-2). The confidence is incorrect regardless of number of significant digits (-1) and it does not even make sense (-1). Author’s estimate 0.001 is not found between 0.0019 and 0.0018. Some information is needed on the data itself, such as number of subjects, etc. (-1). We are interested in estimating the probability of death, not mean LDL (-3).

0/10

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

**From Poisson regression analysis, I estimate that the mean LDL among subjects dying within 5 years is 0.65% lower than in subjects who survived at least 5 years. The estimated SE is .28% which leads to a 95% confidence interval that the true mean LDLs of those subjects who died withn 5 years is between 1.20% lower and 0.10% lower than those who survived at least 5 years. Based on a two-sided p value of 0.021, I reject the null hypothesis that surviving subjects would tend toward a higher mean serum LDL.**

Methods were not stated (-2). Need more digits in all estimates except the p-value (-2). The confidence is incorrect regardless of number of significant digits (-1). Some information is needed on the data itself, such as number of subjects, etc. (-1). The analysis is incorrectly stated: we are not interested in estimating mean LDL, instead we desire an estimate for the probability of death, or some rate (Poisson regression), for inference. (-4)

0/10

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

**From logistic regression analysis, I estimate that for each mg/dL difference in LDL, the odds of death are 0.77% lower in the higher LDL group. The estimated standard error is 0.33% which leads to a confidence interval that the true difference for 1mg/dL difference in LDL is between 1.42% and 0.13% lower in the higher LDL group. Based on a two-sided p value of 0.019, I reject the null hypothesis that surviving subjects would tend toward a higher mean serum LDL.**

Methods were not stated (-2). A relative standard error can be a percentage, otherwise it really should not be (-1). The estimates provided are not correct. It is likely the incorrect analysis was performed. (-6).

1/10

* 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*.?

**The conclusions reached in 4a-c of this homework, and problems 2 and 4 of homework #2, all produce statistically significant, though relatively small associations between LDL and 5-year-survival. By contrast, in problems 1-3 all analyses failed to reject the null hypothesis that there was an association between LDL and 5-year-survival. Given that dichotomizing the LDL variable looses significant data, I would have preferred an analysis of an association relying on continuous data for LDL. Further, given that this data is the result of a cohort study, I would have preferred a logistic regression analysis producing an odds ratio that can then be used to show an association in either direction as whether high LDL leads to longer life, or whether longer life leads to higher LDL cannot be established from the data.**

Need to be more specific about the “significant data lost.” (-1).

2/3