**Biost 518: Applied Biostatistics II**

**Biost 515: Biostatistics II**

Emerson, Winter 2014

**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.

This is a saturated regression model. For the 2 grouping variable (high LDL and Not high LDL), there are 2 parameters to be fit in the regression model. So it’s 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?

By constructing the logistic regression model, the estimated odds of dying within 5 years for subject with low ldl is e^(-1.5863)=0.2047.

So the estimated probability is odds/(1+odds)=0.2047/(1+0.2047)=0.1699.

The we can get the observed proportion of subjects dying within 5 years with low LDL, they are exactly the same(0.1699).

* 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?

By constructing the logistic regression model, the estimated odds of dying within 5 years for subjects with high LDL is e^(-1.5863-0.2562)=0.1584.

The estimated probability is odds/(1+odds)=0.1584/(1+0.1584)=0.1368

This value is exactly the same with the observed proportion of subjects with high LDL dying in 5 years.

* 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?

The association between 5 year mortality and high LDL levels are assessed by using the logistic regression model. The odds ratio of dying in 5 years with high LDL level and low LDL level is tested.

From the logistic regression model, the odds ratio of subjects dying in 5 years between high LDL level and low LDL level is ~~e^(-0.2562)=~~0.7740. with a 95% confidence interval, it will be not unusual if the odds ratio of subjects dying in 5 years with high LDL level to the group with low LDL level is anywhere between 0.4288 to 1.3653 . But this is not a statistically significant result (P=0.3658), we fail to reject the null hypothesis that there is no association between LDL level and 5 year mortality.

 By comparing the result of HW#1, the point estimate of odds ratio and the 95% confidence interval is different. This is because for the logistic regression model, it use the maximum likelihood estimation for the point estimate and use wald based method to estimate the P value and CI.

* 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?

They are reparameterizations. For the first case, the slope of the model will not change because it indicates the odds ratio, but the intercept value will change, it indicates the (log) odds of different group.For the second case, if we use the indicator of survival for at least 5 years, the slope and the intercept both stay 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?

a won’t change, it’s still 2 2 parameter fit 2 groups saturated model. When we answer the question from b to c, we probably want to change the question to: the odds of low/high ldl level in the group who survive/die in 5 years. They follow the same schedule to find the odds from the intercept.

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. In this regression model, there are only 2 parameters, and the grouping variable(high LDL and not high LDL level.) has 2 values. So it’s 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?

Probability(low ldl)=0.1699.

This value is the same with the observed proportion of subjects with low LDL dying within 5 years.

* 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?

Probability(high LDL)=0.1699-0.0332\*1=0.1367

They are the same.

* 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?

The association of LDL level and 5 year mortality is tested by performing the robust linear regression model. The proportion of dying in 5 years of group with low LDL level is 0.1699. And the proportion of dying in 5 years of group with high LDL level is 0.1368. The proportion difference between group who have high LDL level and group who have not high LDL level is- 0.1331. With a 95% confidence interval, it will not be unusual if the proportion difference is anywhere between -0.1022983 and 0.0359968. We cannot reject the null hypothesis that there is no association between LDL level and 5 year mortality(P=0.3469.)

* 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?

They are reparameterizations. For the first case, the slope of the regression model have the same absolute value(opposite sign). And the intercept is different, because it shows the proportion of dying in different LEL group. For the second case, the slope also has same absolute value(opposite sign). And the intercept is different.

* 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?

a won’t change, it’s still 2 2 parameter fit 2 groups saturated model.

For the following questions trying to find the proportion of high/low LDL level in dead/survive group, the proportion will differ from the d~c questions. But we can get the value from the intercept. From example, we get the proportion of high LDL level in survival group is 0.1645. we can also use the model to calculate proportion in other group. But they differ from the d~c anwers.

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.

It is a saturated regression model. Because there are 2 parameters in this model to fit 2 groups.

* 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?

P(low LDL)=0.1699

Odds=0.2047.

They are the same with the observed value.

* 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?

P(high LDL)=0.1368.

Odds=0.1584.

 They are the same.

* 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 passion regression analysis, we estimate that the proportion ration of different group is 0.8049. A 95% CI indicates that the observation will not be unusual if the risk ratio is anywhere from 0.4941779 to 1.310941. This observation is not a statistically significant result(P=0.3831)

* 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?

For the first case, after using different predictor, the slope\*previous slope =1. The intercept stands for the proportion of dying in 5 years among different groups. For the second case, both the slope and the intercept differ from the original one.

* 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 change the predictor of interest and response variable, then the intercept stands for the proportion of high/low LDL level in people who die/survive in 5 years. The the value will not be the same.

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 simple linear regression analysis, we estimate that for each unit difference in ldl level, the difference in proportion of dying in 5 years is -8.500541. A 95% CI suggest that this observation is not unusal if the true difference in proportion is anywhere between -15.50114 and 1.499941. We have the confidence to reject the null hypothesis that the proportion of dying in 5 years is the same among all LDL levels.(P= 0.0174).

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

From poission regression analysis, we estimate that for each 10 fold increase in ldl level, he probability of dying decrease by 85.29%. A 95% CI suggest that the observation will not be unusual if the dying probability in 5 years decrease from 81.83% to 88.88%. This is highly statistically significant observation. (P<0.0001)

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

From logistic regression analysis, we estimate that for each unit difference in ldl, the odds of death in 5 years is 0.7744% lower in the higher ldl level group, and this estimate is statistically significant(P=0.0108). A 95% CI suggests that this observation is not unusual if a group that is one unit ldl lever higher might have odds of dead in 5 years that was anywhere from 1.374% to 0.1712% lower than the younger group

* 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 P value change, then the result change from “not statistically significant” to statistically significant. I would prefer treating the ldl as continuous variable, because if we dichotomize the data, we lose some information then we cannot make the correct judgment.