**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.

**The regression model is saturated as there are two groups defined by high serum LDL status (“high” = LDL > 160 mg/dL) and the regression model has two parameters (a slope and intercept term).**

* 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 serum LDL (“low” = LDL < 160 mg/dL) the estimated odds of dying within 5 years is 0.205 and the estimated probability of dying within 5 years is 0.170.**

**Since the regression model is saturated, the estimated probability of dying within 5 years is equal to the observed proportion of subjects with low serum LDL dying within 5 years.**

* 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 serum LDL (“high” = LDL > 160 mg/dL) the estimated odds of dying within 5 years is 0.151 and the estimated probability of dying within 5 years is 0.131.**

**Since the regression model is saturated, the estimated probability of dying within 5 years is equal to the observed proportion of subjects with high serum LDL dying within 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?

**Methods: We fit a logistic regression model to evaluate the association between high serum LDL status (“high” = LDL > 160 mg/dL) and probability of dying within 5 years. We used robust standard error estimates were used since there is not a compelling reason to assume equal variances across groups.**

**Inference: From the logistic regression model, the odds of dying within 5 years in a population of subjects with high serum LDL is 27% lower than a population of subjects with low serum LDL. The odds ratio is not significantly different from 1 at a significance threshold of 0.05 (P=0.316). The 95% confidence interval suggests that the observed odds ratio is not unusual if the true odds of dying within 5 years for a population of subjects with high serum LDL were anywhere between 1-0.4035525% lower and 34% higher compared to a similar population of subjects with low serum LDL.**

* 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?

**The answers to parts a-c would stay the same if the logistic regression model were fit using an indicator of low LDL as the predictor, as it is a reparametrization of the previous logistic regression model. If we had used an indicator of survival for at least 5 years as the response variable instead of death within 5 years, full inferences on the probability of death or survival for both groups can be computed since the model is 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 switched the predictor with the response variable, we would still derive the same values in parts a-c if we used Bayes’ rule. The reparametrized model would generate estimates of the probability of high LDL given vital status; the probability of vital status at 5 years given high LDL could be computed using Bayes’ rule.**

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.

**The regression model is saturated as there are two groups defined by high serum LDL status (“high” = LDL > 160 mg/dL) and the regression model has two parameters (a slope and intercept term).**

* 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 serum LDL (“low” = LDL < 160 mg/dL) the estimated probability of dying within 5 years is 0.170 and the estimated odds of dying within 5 years is 0.205.**

**Since the regression model is saturated, the estimated probability of dying within 5 years is equal to the observed proportion of subjects with low serum 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 high LDL dying within 5 years?

**For subjects with high serum LDL (“high” = LDL > 160 mg/dL) the estimated probability of dying within 5 years is 0.131 and the estimated odds of dying within 5 years is 0.151.**

**Since the regression model is saturated, the estimated probability of dying within 5 years is equal to the observed proportion of subjects with high serum LDL dying within 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?

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **deadin5** | **Coef.** | **Std. Err.** | **t** | **P>t** | **[95% Conf.** | **Interval]** |
| **highldl** | -0.039 | 0.036 | -1.090 | 0.278 | -0.110 | 0.032 |
| **intercept** | 0.170 | 0.015 | 11.230 | 0.000 | 0.140 | 0.200 |

**Methods: We fit a simple linear regression model to evaluate the association between the probability of dying within 5 years and the indicator of high serum LDL (“high” = LDL > 160 mg/dL). Since there was no compelling reason to assume equal variances across groups defined by serum LDL status, confidence intervals were computed using robust standard error estimates.**

**Inference: From the linear regression output, the population of subjects with high serum LDL have a 3.9% lower probability of death within 5 years relative to a population of subjects with low serum LDL. The observed difference in 5 year mortality rates is not significant (P=0.278) at the 0.05 level. The 95% confidence interval asserts the observed probability is not unusual if true difference in probability of dying within 5 years in the high LDL group were anywhere between 11% lower and 3.2% higher relative to the low serum LDL group.**

**The point estimates from the model are the same as in questions 5 and 6 of Homework 1, and the confidence intervals are different. The confidence limits in the linear regression model above were computed using robust standard error estimates, whereas the confidence interval for the risk difference in questions 5 and 6 were calculated using exact confidence interval estimates.**

* 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?

**The answers to parts a-c would stay the same if the linear regression model were fit using an indicator of low LDL as the predictor, as it is a reparametrization of the previous linear regression model. If we had used an indicator of survival for at least 5 years as the response variable instead of death within 5 years, full inferences on the probability of death or survival for both groups can be computed since the model is 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?

**Response A would still be the same as we would still have a saturated model. For parts B and C, we would still be able to infer the same probabilities as we are still fitting a binary response variable with a binary predictor.**

**Since we changed the response variable and the predictor of interest, we can use Bayes’ rule to go back and forth to estimate the probability of dying within 5 years conditional high LDL status from the probability of LDL status given vital status at 5 years.**

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.)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **deadin5** | **Coefficient** | **Robust Std. Err.** | **[95% Conf.** | **Interval]** |
| **LDL > 160mg/dL** | -0.261 | 0.265 | -0.780 | 0.258 |
| **constant** | -1.773 | 0.089 | -1.947 | -1.598 |

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

**The regression model is saturated because the number of groups (divided by high LDL status, defined as serum LDL > 160 mg/dL) is equal to the number of coefficients (the slope parameter and the intercept parameter) in the 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 17%, with the estimated odds of dying within 5 years of 0.2047. Since the regression model is saturated, the estimated probability of dying within 5 years for subjects with low LDL is equal to the observed proportion of subjects with low serum 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 high LDL dying within 5 years?

**For subjects with high LDL, the estimated probability of dying within 5 years is 13.1%, with estimated odds of dying within 5 years of 0.1505. Since the regression model is saturated, the estimated probability of dying within 5 years for subjects with high LDL is equal to the observed proportion of subjects with high serum LDL dying within 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?**

**From Poisson regression, we estimate that for a population of subjects dichotomized by high serum LDL status (where high LDL = serum LDL > 160 mg/dL), the probability of death within 5 years is 0.170 for subjects with low serum LDL and 0.131 for subjects with high serum LDL. The result is not significant at a 0.05 level significance threshold (P=0.3237); the 95% confidence interval for the risk ratio of death within 5 years suggests the observed risk ratio of 0.77 is not unusual if the true risk of death within 5 years for high LDL subjects is anywhere between 54% lower and 29% higher than a population with low serum LDL.**

* 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?

**The answers to parts a-c would stay the same; the new proposed model is a reparametrization of the previous model using the indicator of high LDL as the predictor. The new model If we had used an indicator of survival for at least 5 years as the response variable instead of death within 5 years, full inferences on the probability of death or survival for both groups can be computed since the model is 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?

**Response A would still be the same as we would still have a saturated model. For parts B and C, we would still be able to infer the same probabilities as we are still fitting a binary response variable with a binary predictor.**

**Since we changed the response variable and the predictor of interest, we can use Bayes’ rule to go back and forth to estimate the probability of dying within 5 years conditional high LDL status from the probability of LDL status given vital status at 5 years.**

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).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **deadin5** | **Coef.** | **Std. Err.** | **t** | **P>t** | **[95% Conf.** | **Interval]** |
| **LDL (mg/dL)** | -0.001 | 0.000 | -2.390 | 0.017 | -0.002 | 0.000 |
| **intercept** | 0.294 | 0.058 | 5.070 | 0.000 | 0.180 | 0.408 |

**Methods: We fit a simple linear regression model to evaluate the association between the probability of dying within 5 years and serum LDL level. Robust standard error estimates were used since there was no compelling reason to assume equal variances across varying LDL levels.**

**Inference: The linear regression model found a significant association between serum LDL and the probability of dying within 5 years. The regression model estimates that for every 1 mg/dL increase in serum LDL, the estimated probability of dying within 5 years is reduced by 0.1%. The results are significant at the 0.05 level (P=0.017) and the 95% confidence interval asserts that the observed difference in probability of 5 year mortality is not unusual if the true difference in probability is anywhere from 0.2% lower to 0% lower per 1 mg/dL increase in serum LDL.**

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

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **deadin5** | **Coef.** | **Std. Err.** | **z** | **P>z** | **[95% Conf.** | **Interval]** |
| **LDL (mg/dL)** | -0.006 | 0.003 | -2.370 | 0.018 | -0.012 | -0.001 |
| **intercept** | -1.016 | 0.330 | -3.080 | 0.002 | -1.662 | -0.370 |

Methods: We fit a Poisson regression **model to evaluate the association between the probability of dying within 5 years and serum LDL level. The null hypothesis is the risk ratio of the probability of death within 5 years is equal to 1.**

**For each 1 mg/dL increase in serum LDL, the probability of death within 5 years decreases by 0.6%. Based on the 95% confidence interval, the observed difference in probability of death within 5 years is not unusual if the true difference in probability was anywhere between 1.2% lower and 0.1% lower for each 1 mg/dL increase in serum LDL. The observed difference in probability of death within 5 years is significant (P=0.018) at the 0.05 level, therefore we reject the null hypothesis of no association between serum LDL status and the probability of death within 5 years.**

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

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **deadin5** | **Odds Ratio** | **Std. Err.** | **z** | **P>z** | **[95% Conf.** | **Interval]** |
| **LDL (mg/dL)** | 0.992 | 0.003 | -2.340 | 0.019 | 0.986 | 0.999 |
| **intercept** | 0.511 | 0.209 | -1.640 | 0.101 | 0.229 | 1.141 |

**Methods: We fit a logistic regression model to evaluate the association between the probability of dying within 5 years and serum LDL level. We use robust standard errors as there is no compelling reason to assume equal variances across serum LDL concentrations.**

**Inference: When comparing populations varying by serum LDL concentrations of 1 mg/dL, the odds of dying within 5 years is 0.8% lower in the group with higher serum LDL. The difference in odds ratio is significant at the 0.05 level (P=0.019). Based on the 95% confidence interval, the observed difference in odds ratio is not unusual if the true difference in odds ratios is anywhere from 1.4% lower to 0.1% lower in a group with average serum LDL concentration 1 mg/dL higher relative to another 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 conclusions about the association between serum LDL and 5 year mortality rate are similar to the conclusions reached in problems 1-3 of this homework in that they supported a tendency for subjects with higher LDL to have a reduced probability of 5 year mortality, though the models treating LDL as a continuous variable found a significant association between serum LDL level and 5 year all-cause mortality and the models dichotomizing LDL by “high” LDL status did not find a significant association. From this, it appears best to treat LDL as a continuous variable than to dichotomize the data by an arbitrary cutoff, as treating LDL as a continuous variable does not achieve the same loss of information as separating the groups into two categories does.**

**It is more clinically useful to use serum LDL status to predict the probability of dying within 5 years rather than to predict the serum LDL of patients that have survived or died after 5 years.**

**Additionally, I would have preferred to use logistic regression since the outcome of interest is binary (vital status at 5 years) as to avoid significant nonlinearities.**