Homework 4

Aji John

Winter Quarter 2018

Question 1

Perform a regression analysis for evaluating an association between serum creatinine level and 5 year all-cause mortality by comparing the odds of death within 5 years across groups defined by whether the subjects have "high" or "low" creatinine levels, where serum creatinine levels greater than 1.2 are considered to be "high", (i.e., "high" corresponds to creatinine $>$ 1.2 mg/dl and "low" corresponds to creatinine less than or equal 1.2 mg/dl). In your logistic regression model, use an indicator of death within 5 years as the response, and an indicator of high serum creatinine level as the predictor. (Only provide a formal report of inference when asked to.)

(a)

Is this a saturated regression model? Explain your answer.

Yes, it is a saturated model, as number of parameters equal to number of coefficients.

(b)

Provide an interpretation of the slope and the intercept in your regression model in terms of the response variable (indicator of death within 5 years) and the predictor variable (high creatinine).

Odds when predictor is 0 – Found by exponentiation of the intercept from the logistic regression: exp(b0) - 0.1561866

Odds ratio between groups differing in the value of the predictor by 1 unit (i.e groups defined by CRT levels) – Found by exponentiation of the slope from the logistic regression: exp(b1) - 2.3673469

(c)

From the logistic regression model, what is the estimated odds of dying within 5 years for subjects with low creatinine levels. What is the estimated probability of dying within 5 years from the logistic regression model for subjects with low creatinine levels ?

Odds of dying within 5 years for subjects with low creatinine levels is e-1.8567=0.1562

Probability of dying within 5 years for subjects with low creatinine levels is 0.1350977.

- Using prob= odds / (1+odds): 0.1562 / (1+0.1562)= 0.1350977.

(d)

For the subset of subjects in the sample with low serum creatinine, calculate the proportion who died within 5 years. Also calculate the sample odds of dying within 5 years for subjects with low creatinine levels. Compare the sample proportion and sample odds to the corresponding estimates in part 1c from the logistic regression model? Briefly explain any similarities or differences.

Proportion who died in 5 years - 0.13 Odds - .15

Compared to 1c , it is the same because it is a saturated model.

(e)

From the logistic regression model, what is the estimated odds of dying within 5 years for subjects with high creatinine levels. What is the estimated probability of dying within 5 years from the logistic regression model for subjects with high creatinine levels ?

Odds = 0.1562 * 2.367 = 0.3697254 - Using prob= odds / (1+odds): 0.3697254/(1+0.3697254) = 0.2699267

(f)

For the subset of subjects in the sample with high creatinine levels, calculate the proportion who died within 5 years. Also calculate the sample odds of dying within 5 years for subjects with high creatinine levels. Compare the sample proportions (or probabilities) and sample odds to the corresponding estimates in part 1e from the logistic regression model? Briefly explain any similarities or differences.

Proportion who died in 5 years - 0.26 Odds of dying in 5 years - .36

It is similar to part 1e because we have a saturated regression model.

(g)

Give full inference regarding an association between 5 year all-cause mortality and serum creatinine levels from the logistic regression model with indicator of death within 5 years as the response and indicator of high creatinine level as the predictor.

From logistic regression analysis, we estimate that for two groups that differed by their CRT level, the odds of dying in 5 years is 136% higher than in the group that had low CRT level, also, this estimate is found to be statistically significant. A 95% CI suggests that this observation is not unusual if a group that differs by CRT level might have odds of dying that was anywhere from 55% lower or 261% higher than the group that had low CRT level. A two-sided p value of 0.00001 suggests that we can with high confidence reject the null hypothesis that the odds of dying in 5-years is not associated with high CRT level.

(h)

How would your answers to part b change if you were instead asked to fit a logistic regression model with indicator of death within 5 years as the response variable, but with indicator of low serum creatinine level as the predictor? Would the statistical evidence for an association between 5 year mortality and serum creatinine levels change? Briefly explain.

Odds when predictor is 0 – Found by exponentiation of the intercept from the logistic regression: exp(b0) i.e .36. It means the odds of a person dying in 5 years is 36% for a person with High CRT.

Odds ratio between groups differing in the value of the predictor by 1 unit (Low CRT/ high CRT) – Found by exponentiation of the slope from the logistic regression: exp(b1) - 0.4224

Here, it is in comparision to the group with high CRT , so odds (Low CRT/ high CRT) is .42 i.e. .42 times less likely than a person with low CRT.

(i)

How would your answers to part b change if you were instead asked fit a logistic regression model with indicator of surviving at least 5 years as the response variable and indicator of high creatinine level as the predictor? Would the statistical evidence for an association between 5 year mortality and serum creatinine levels change? Briefly explain.

Odds are 15% of a person dying in 5 years having a low CRT level. And, the odds ratio is 2.3 i.e. a person with low CRT is 2.3 times more likely to die than the person with high CRT.

Our statiscal inference does not change.

Question 2

In question 1, a prospective association analysis was conducted where we investigated differences in the distribution of death within 5 years across groups defined by serum creatinine level. In this question, you will now conduct a retrospective analysis and fit a logistic regression model for the distribution of serum creatinine across groups defined by vital status at 5 years. In your retrospective logistic regression model, use an indicator for high serum creatinine level as the response, and indicator of death within 5 years as the predictor. (Only provide a formal report of inference when asked to.)

(a)

Provide an interpretation of the slope and the intercept in your regression model in terms of the response variable (indicator of high creatinine level) and the predictor variable (indicator of death within 5 years).

Our model here is log odds = -1.4214 + 0.8618 * Deathin5i

when predictor is 0(Does not die in 5 years) –Found by exponentiation of the intercept from the logistic regression: exp(b0) = 0.2413759

Odds ratio between groups differing in the value of the predictor by 1 unit–Found by exponentiation of the slope from the logistic regression: exp(b1) = 2.367418

(b)

From the logistic regression model, what is the estimated odds of high creatinine level for subjects who die within 5 years? What is the estimated probability of having high serum creatinine for subjects who die within 5 years.

Odds is 0.57 Using prob = odds / (1+odds): 0.36364

(c)

From the logistic regression model, what is the estimated odds of having a high creatinine level for subjects who survive at least 5 years? What is the estimated probability of having a high serum creatinine for subjects who survive at least 5 years.

Odds is 0.2413759 Using prob= odds / (1+odds): 0.19

(d)

Give full inference regarding an association between 5 year all-cause mortality and serum creatinine levels from the logistic regression model with indicator of high serum creatinine as the response and an indicator of death within 5 years as the predictor.

From logistic regression analysis, we estimate that for two groups that differed by their 5 year all-cause mortality, the odds of having high CRT is 136% higher than in the group that survived at the end of years, also, this estimate is found to be statistically significant. A 95% CI suggests that this observation is not unusual if a group that differed by 5-year all-cause mortality might have odds of high CRT that was anywhere from 55% lower or 261% higher than the group that survived at the end of 5 years. A two-sided p value of 0.00001 suggests that we can with high confidence reject the null hypothesis that the odds of having a hight CRT is not associated with 5 year all-cause mortality.

(e)

Compare the association results in part 2d from the retrospective logistic model to the association results in part 1g from the prospective logistic regression model. Briefly describe any similarities or differences.

The results are similar in a sense that the group which has higher CRT is found to be most likely not to survive at the end of 5 years.

Question 3

Perform a regression analysis to evaluate an association between odds of death within 5 years and the continuous measure of serum creatinine levels (i.e., do not use a dichotomized variable for serum creatinine levels in this analysis).

(a)

Provide an interpretation of the slope and the intercept in your logistic regression model.

The model we get is log odds of DeathIn5 = -3.605 + 1.789 * CRT

Intercept, when predictor is 0 (Creatinine level) – Found by exponentiation of the intercept from the logistic regression: exp(b0) = 0.02718744

Odds ratio between groups differing in the value of the predictor by 1 unit (mg/dl) –Found by exponentiation of the slope from the logistic regression: exp(b1) = 5.98

- Two rows deleted because of NA

(b)

Give full inference for an association between 5 year all-cause mortality and serum creatinine levels from your logistic regression model.

From logistic regression analysis, we estimate that for two groups that differ by one unit(mg/dl) in CRT, the odds of death in 5 years is 498% higher in the group that differs by 1 unit(mg/dl), also, this estimate is found to be statistically significant. A 95% CI suggests that this observation is not unusual if a group that is one unit(mg/dl) higher might have odds of death in 5 years that was anywhere from 210% lower or 1049% higher than the group that has a higher CRT level. A two-sided p value of 0.00005 suggests that we can with high confidence reject the null hypothesis that the odds of death in 5 years is not associated with CRT.