BIOST 515/518 Winter 2014

Homework 3

ID 3256

1.(a) There are two distinct groups (low and high LDL) and two regression parameters. Since number of groups = number of parameters, this model is saturated.

(b) For the 618 subjects with low LDL (LDL < 160mg/dL), the estimated odds of dying within 5 years is 0.205 (β0 of model). The estimated probability of dying within 5 years is 0.170 (calculated using prob = odds/(1+odds)). 105 out of the 618 subjects with low LDL died within 5 years, so the observed proportion of subjects with low LDL dying within 5 years is 0.170. The estimated odds of dying within 5 years is about 20% higher than the observed proportion, estimated probability of dying within 5 years is exactly the same as the observed proportion.

(c) For the 107 subjects with high LDL (LDL ≥ 160mg/dL), the estimated odds of dying within 5 years is 0.151 (β0 x β1 of model). The estimated probability of dying within 5 years is 0.131 (calculated using prob = odds/(1+odds)). 14 out of the 107 subjects with low LDL died within 5 years, so the observed proportion of subjects with high LDL dying within 5 years is 0.131. The estimated odds of dying within 5 years is about 15.3% higher than the observed proportion, estimated probability of dying within 5 years is exactly the same as the observed proportion.

(d) Of the 618 subjects with low serum LDL (<160mg/dL), the odds of 5-year mortality is 0.205, while the odds of mortality of the 107 subjects with high serum LDL (≥ 160mg/dL) is 0.151 within 5 years of study enrollment. From logistic regression analysis with robust standard errors, we estimate that between subjects with low and high serum LDL levels, the odds of dying within 5 years is 26.5% lower among subjects with high serum LDL levels, although this estimate is not statistically significant (p = 0.316). A 95% confidence interval suggests that this observation is not unusual if a group with high serum LDL levels might have odds of dying within 5 years that was anywhere from 59.6% lower to 34.0% higher than a group with low serum LDL levels. The odds ratios, probabilities of dying, difference in odds are the same as in problem 6 of homework #1 using Wald (Woolf) calculations. The confidence intervals and p-values here are different than the ones in Homework 1 because our standard errors here are computed using robust methods. Nonetheless they both give us statistically non-significant results.

(e) Short answer: **nothing should change** since the two suggested fits are just reparametrizations of the original model. The model will still be saturated. If we fit a logistic regression model using indicator of death within 5 years as the response variable and indicator of low LDL as the predictor (logistic in STATA), the intercept is the odds of dying within 5 years for high LDL instead of low LDL, so to get the odds of dying within 5 years for low LDL we would multiply the intercept and slope. If we used an indicator of survival for at least 5 years as the response variable, the intercept of the model will be the odds of survival within 5 years for low LDL, so to get the odds of dying within 5 years we would just take the reciprocal of it.

(f) If we instead define our predictor variable to be an indicator of dying within 5 years and our response variable to be an indicator of high LDL levels, we should still get the **same results in parts a-c** since the there are still 2 groups and 2 predictors, hence model still saturated, and 2x2 table for odds ratio is symmetric, hence odds and probabilities of dying within 5 years of the two groups are still the same. However it will be difficult to extract information from this new model to answer the questions.

2. (a) There are two distinct groups (low and high LDL) and two regression parameters. Since number of groups = number of parameters, this model is saturated.

(b) For the 618 subjects with low LDL (LDL < 160mg/dL), the estimated probability of dying within 5 years is 0.170 (β0 of model). The estimated probability of dying within 5 years is 0.205 (calculated using odds = prob/(1-prob)). The observed proportion of subjects with low LDL dying within 5 years is 0.170. The estimated odds of dying within 5 years is 20.4% higher than the observed proportion, estimated probability of dying within 5 years is exactly the same as the observed proportion.

(c) For the 107 subjects with high LDL (LDL ≥ 160mg/dL), the estimated probability of dying within 5 years is 0.131 (β0 + β1 of model). The estimated probability of dying within 5 years is 0.151 (calculated using odds = prob/(1-prob)). The observed proportion of subjects with low LDL dying within 5 years is 0.131. The estimated odds of dying within 5 years is 20.6% higher than the observed proportion, estimated probability of dying within 5 years is exactly the same as the observed proportion.

(d) Of the 618 subjects with low serum LDL (<160mg/dL), 17.0% were observed to die within 5 years, while 13.1% of the 107 subjects with high serum LDL (≥ 160mg/dL ) died within 5 years of study enrollment. From linear regression on binary variables analysis with robust standard errors, we estimate that between the subjects with low and high serum LDL levels, the absolute difference in probability of death within 5 years is 3.91% lower among subjects with high serum LDL levels, although this estimate is not statistically significant (p = 0.278). A 95% confidence interval suggests that this observation is not unusual if a group with high serum LDL levels might have probabilities of dying within 5 years that was anywhere from 11.0% lower to 3.16% higher than a group with low serum LDL levels. The odds ratios, probabilities of dying, difference in odds are the same as in problem 5 of homework #1 (p = 0.314 from chi-squared test). The confidence intervals and p-values here are different than the ones in Homework 1 because our standard errors here are computed using robust methods. Also we are not doing a chi-squared test here. Nonetheless they both give us statistically non-significant results.

(e) Short answer: **nothing should change** since the two suggested fits are just reparametrizations of the original model. The model will still be saturated. If we fit a linear regression model using indicator of death within 5 years as the response variable and indicator of low LDL as the predictor, the intercept is the probability of dying within 5 years for high LDL instead of low LDL, so to get the probability of dying within 5 years for low LDL we would add the intercept and slope. If we used an indicator of survival for at least 5 years as the response variable, the intercept of the model will be the probability of survival within 5 years for low LDL, so to get the probability of dying within 5 years we would just take the 1 minus that number.

(f) If we instead define our predictor variable to be an indicator of dying within 5 years and our response variable to be an indicator of high LDL levels, we should still get the **same results in parts a** since the there are still 2 groups and 2 predictors, hence model still saturated, and 2x2 table for probabilities is symmetric, thus odds and probabilities of dying within 5 years of the two groups are still the same. However we cannot answer the questions b-c directly from the model. We will need information such as P(5-year mortality) and P(High LDL) to do a Bayes rule calculation.

3. (a) There are two distinct groups (low and high LDL) and two regression parameters. Since number of groups = number of parameters, this model is saturated.

(b) For the 618 subjects with low LDL (LDL < 160mg/dL), the estimated probability of dying within 5 years is 0.170 (exp(β0) of model). The estimated probability of dying within 5 years is 0.205 (calculated using odds = prob/(1-prob)). The observed proportion of subjects with low LDL dying within 5 years is 0.170. The estimated odds of dying within 5 years is 20.4% higher than the observed proportion, estimated probability of dying within 5 years is exactly the same as the observed proportion.

(c) For the 107 subjects with high LDL (LDL ≥ 160mg/dL), the estimated probability of dying within 5 years is 0.131 (exp(β0) x exp(β1) of model). The estimated probability of dying within 5 years is 0.151 (calculated using odds = prob/(1-prob)). The observed proportion of subjects with low LDL dying within 5 years is 0.131. The estimated odds of dying within 5 years is 20.4% higher than the observed proportion, estimated probability of dying within 5 years is exactly the same as the observed proportion.

(d) Of the 618 subjects with low serum LDL (<160mg/dL), 17.0% were observed to die within 5 years, while 13.1% of the 107 subjects with high serum LDL (≥ 160mg/dL ) died within 5 years of study enrollment. From Poisson regression on analysis, we estimate that subjects with high serum LDL have 33.0% lower 5-year mortality rates compared to subjects with low serum LDL (risk ratio = 0.770), although this estimate is not statistically significant (p = 0.324). A 95% confidence interval suggests that this observation is not unusual if a group with high serum LDL levels might have anywhere between 54.2% lower to 29.4% higher 5-year mortality rates compared to a group with low serum LDL levels (ratio of rates between 0.458 and 1.294). The point estimates match up exactly with the table of homework 1. However note that we are dealing with risk ratios here while problems 5 & 6 of homework 1 dealt with odds ratios and risk differences. The confidence intervals are different and p-value here is lower than in homework 1, since we used robust standard errors here and Wald in homework 1. Nonetheless they both give us statistically non-significant results.

(e) Short answer: nothing should change since the two suggested fits are just reparametrizations of the original model. The model will still be saturated. If we fit a linear regression model using indicator of death within 5 years as the response variable and indicator of low LDL as the predictor, the intercept is the log of probability of dying within 5 years for high LDL instead of low LDL, so to get the probability of dying within 5 years for low LDL we would multiply the exponents of the intercept and slope. If we used an indicator of survival for at least 5 years as the response variable, the intercept of the model will be the log probability of survival within 5 years for low LDL, so to get the probability of dying within 5 years we would just take the 1 minus the exponent of the number.

(f) If we instead define our predictor variable to be an indicator of dying within 5 years and our response variable to be an indicator of high LDL levels, we should still get the **same results in parts a-c** since the there are still 2 groups and 2 predictors, hence model still saturated, and 2x2 table for cases and controls is symmetric, hence odds and probabilities of dying within 5 years of the two groups are still the same.

4. (a) Methods: Distributions of death within 5 years were compared across groups defined by the continuous measure of LDL. Linear regression with robust standard error was performed on an indicator of whether the subject died within 5 years against LDL levels. 95% confidence intervals for the risk difference were constructed based on that same handling of variances.

Inference: From a linear regression analysis of the 618 subjects with low LDL (< 160mg/dL) and 107 subjects with high LDL (≥ 160mg/dL), we estimate a 5-year mortality risk difference of 0.103% lower with each unit increase of LDL levels. Based on a 95% confidence interval computed with robust standard errors, we find that observing such an estimated difference is not unusual if the true difference in 5-year mortality risk differences were anywhere between a tendency towards high serum LDL groups having risk differences 0.188% lower to 0.0185% lower per mg/dL difference in serum LDL.These results are statistically significant at a 0.05 level of significance (two-sided P=0.017), thus we can with high confidence reject the null hypothesis that there is no difference in risk differences between subjects with low and high serum LDL levels in favor of a hypothesis that high serum LDL levels is associated with a lower 5-year mortality rate. However since the difference in risk differences are so small it may not be of scientific interest.

(b) Methods: Distributions of death within 5 years were compared across groups defined by the continuous measure of LDL. Poisson regression with robust standard error was performed on an indicator of whether the subject died within 5 years against LDL levels. 95% confidence intervals for the risk difference were constructed based on that same handling of variances.

Inference: From a Poisson regression analysis of the 618 subjects with low LDL (< 160mg/dL) and 107 subjects with high LDL (≥ 160mg/dL), we estimate a 5-year mortality risk ratio of 0.645% lower with each mg/dL difference of LDL levels with high LDL levels having a lower risk ratio. Based on a 95% confidence interval computed with robust standard errors, we find that observing such an estimated difference is not unusual if the true difference in 5-year mortality rate ratios were anywhere between a tendency towards high serum LDL groups having risk ratios 0.112% lower to 1.17% lower per mg/dL increase in serum LDL. These results are statistically significant at a 0.05 level of significance (two-sided P=0.018), thus we can with high confidence reject the null hypothesis that there is no difference in risk ratios between subjects with low and high serum LDL levels in favor of a hypothesis that high serum LDL levels is associated with a lower 5-year mortality rate. However since the difference in risk ratios are so small it may not be of scientific interest.

(c) Methods: Distributions of death within 5 years were compared across groups defined by the continuous measure of LDL. Logistic regression with robust standard error was performed on an indicator of whether the subject died within 5 years against LDL levels. 95% confidence intervals for the odds ratio were constructed based on that same handling of variances.

Inference: From a logistic regression analysis of the 618 subjects with low LDL (< 160mg/dL) and 107 subjects with high LDL (≥ 160mg/dL), we estimate an odds ratio of dying in 5 years of 0.774% lower with each unit increase of LDL levels. Based on a 95% confidence interval computed with robust standard errors, we find that observing such an estimated difference is not unusual if the true difference in odds ratio were anywhere between a tendency towards high serum LDL groups having odds ratios 0.125% lower to 1.42% lower per mg/dL difference in serum LDL. These results are statistically significant at a 0.05 level of significance (two-sided P=0.019), thus we can with high confidence reject the null hypothesis that there is no difference in odds ratios between subjects with low and high serum LDL levels in favor of a hypothesis that high serum LDL levels is associated with a lower odds of dying within 5 years. However since the difference in odds ratios are so small it may not be of scientific interest.

(d) In all parts of this problem, the p-values are significant and we reject the null hypothesis that there is no association between serum LDL levels and 5-year mortality rates. In problems 1-3 of this homework, none of the p-values were significant and we cannot reject the null hypothesis based on dichotomizing the predictor (serum LDL levels) and outcome (5-year mortality). In problems 2 and 4 of homework 2, we got significant p-values and reject the null hypothesis, and the p-values for the unequal variance test (hw 2 Q4) was the same compared to the ones in this problem. When assuming equal variances (hw 2 Q2), the p-values were lower.

A priori, I would perform a Poisson regression with robust standard errors to characterize the risk ratios of death in 5 years across groups defined by continuous serum LDL measures. It is best to not dichotomize a continuous measure (of serum LDL), as we lose precision, thus the tests in problems 1-3 are not as attractive. It is more of scientific interest to think of serum LDL levels as the “predictor” and 5-year mortality as the “outcome”, thus the tests in homework 2 are not as attractive. Finally, among the tests in problem 4 of this homework, odds ratio has a difficult interpretation thus not as attractive. It may be more intuitive for patients to understand the likelihood (ratio) of mortality per ratio change in serum LDL compared to the risk difference. Furthermore, for biological reasons, a multiplicative model makes more sense than an additive model. However there are some cautions for using Poisson regression: if there is a lack of linearity then the model may not be easily interpretable.