101/108

1.

a. This is a saturated model because there are two response variable groups (those dead within 5 years and those alive at 5 years), which are modeled with two regression parameters (the intercept and the slope).

b.

For subjects with low LDL, the estimated odds of dying within 5 years is .205, and the estimated probability of dying within 5 years is .170. The observed proportion of subjects with low LDL dying within 5 years is also .170, the same as the estimated probability (as is the case for saturated models).

c. For subjects with high LDL, the estimated odds of dying within 5 years is .151, and the estimated probability of dying within 5 years is .131. The observed proportion of subjects with high LDL dying within 5 years is also .131, the same as the estimated probability (as is the case for saturated models).

d.

Logistic regression analysis allowing for heteroscedasticity was performed by comparing the odds of death within 5 years across groups defined by whether the subjects have high serum LDL.

Groups that have low LDL (<160 mg/dL) have an estimated odds of being dead in 5 years that is 73.5% of a group that has high LDL (estimated odds 26.5% lower in the group that has low LDL), though this estimate is not statistically significant (P = 0.316). A 95% confidence interval suggests this observation is not unusual if a group that has low LDL has an odds of death within 5 years which is between values that are 59.5% lower and 34.0% higher that that of a group which has high LDL.

This estimated odds ratio and proportions are the same as that computed in HW1, #5 and #6, though the CI here is slightly smaller (and the SE is smaller here). The p value is also slightly smaller. This has to do with the fact that regression does not quantify an error distribution and uses maximum likelihood estimation.

e.

Since the predictor is already dichotomized, changing it from low LDL to high LDL does not change the scientific inference at all (though the regression parameters are changed). Similarly, since the response variable is already dichotomized, it does not matter if we change it from death within 5 years to alive at 5 years; you still get the same scientific inference (though regression parameters are changes). For each, the y intercept will change and the OR will be inversed.

f.

If we switch the predictor of interest and response variable, the p value and conclusions about scientific inference will stay the same since both are dichotomized. However, the slope and intercept (and their SEs and CIs) will obviously change, as they would then give information the change in LDL between the groups rather than the change is death odds.

2.

a.

 This is a saturated model because there are two response variable groups (those dead within 5 years and those alive at 5 years), which are modeled with two regression parameters (the intercept and the slope).

b. For subjects with low LDL, the estimated probability of dying within 5 years is .170, and the estimated odds is .205. The observed proportion of subjects with low LDL dying within 5 years is also .170, the same as the estimated probability (as is the case for saturated models).

c.

For subjects with high LDL, the estimated probability of dying within 5 years is .131, and the estimated odds is .151. The observed proportion of subjects with high LDL dying within 5 years is also .131, the same as the estimated probability (as is the case for saturated models).

d.

Linear regression analysis allowing for heteroscedasticity was performed by comparing the risk of death within 5 years across groups defined by whether the subjects have high serum LDL.

Groups that have low LDL have an estimated risk of dying within 5 years that is 3.9% lower than a group who has high LDL (estimated risk of death 17.0% vs. 13.1%), however, this is not statistically significant (P value = .278). A 95% confidence interval suggests that this observation would be unsurprising if the true risk difference was between 10.9% lower and 03.15% higher in the low LDL group than the high LDL group. This is exactly the estimated difference in risks and confidence interval from HW1, #5, but we have a smaller P value here.

e.

Since the predictor is already dichotomized, changing it from low LDL to high LDL does not change the scientific inference at all (though the regression parameters are changed). Similarly, since the response variable is already dichotomized, it does not matter if we change it from death within 5 years to alive at 5 years; you still get the same scientific inference (though regression parameters are changes). For each, the y intercept will change and the RD will be inversed.

f.

If we switch the predictor of interest and response variable, the p value and conclusions about scientific inference will stay the same since both are dichotomized. However, the slope and intercept (and their SEs and CIs) will obviously change, as they would then give information the change in LDL between the groups rather than the change is death odds.

3.

a.

This is a saturated model because there are two response variable groups (those dead within 5 years and those alive at 5 years), which are modeled with two regression parameters (the intercept and the slope).

b. For subjects with low LDL, the estimated probability of dying within 5 years is .170, and the estimated odds is .205. The observed proportion of subjects with low LDL dying within 5 years is also .170, the same as the estimated probability (as is the case for saturated models).

c.

For subjects with high LDL, the estimated probability of dying within 5 years is .131, and the estimated odds is .151. The observed proportion of subjects with high LDL dying within 5 years is also .131, the same as the estimated probability (as is the case for saturated models).

d.

Poisson regression analysis allowing for heteroscedasticity was performed by comparing the risk ratio of death within 5 years across groups defined by whether the subjects have high serum LDL.

The estimated risk ratio of death within 5 years between a group who has a mean serum LDL > 160 mg/dL and a group that has a mean serum LDL <160 mg/dL is .770, with the group that has a mean serum LDL <160 mg/dL having the higher risk (estimated risk of death 17.0% vs. 13.1%). However, this is not statistically significant (P value = .324). A 95% confidence interval suggests that this observation would be unsurprising if the true risk ratio was between .458 and 1.294. This is exactly the estimated difference in risks and confidence interval from HW1, #5, but we have a smaller P value here.

e.

Since the predictor is already dichotomized, changing it from low LDL to high LDL does not change the scientific inference at all (though the regression parameters are changed). Similarly, since the response variable is already dichotomized, it does not matter if we change it from death within 5 years to alive at 5 years; you still get the same scientific inference (though regression parameters are changes). For each, the y intercept and RR will change accordingly.

f.

If we switch the predictor of interest and response variable, the p value and conclusions about scientific inference will stay the same since both are dichotomized. However, the slope and intercept (and their SEs and CIs) will obviously change, as they would then give information the change in LDL between the groups rather than the change is risk of death.

4.

a

Linear regression analysis allowing for heteroscedasticity was performed by comparing the risk of death within 5 years across groups defined by the serum LDL level.

From a linear regression analysis of 725 available observations from a sample of 735 elderly subjects between ages 65 and 99, we estimate a difference in mean risk of death within 5 years of -.00103 between groups with a mean serum LDL level 1 mg/dL different, with the group with the higher mean serum LDL having a lower risk of death within 5 years. Based on a 95% confidence interval, the observed results would not be surprising if the true risk difference between these groups is between -.00188 and -.000185. These results are statistically significant with a p value of 0.017.

b.

Poisson regression analysis allowing for heteroscedasticity was performed by comparing the risk of death within 5 years across groups defined by the serum LDL level.

From a linear regression analysis of 725 available observations from a sample of 735 elderly subjects between ages 65 and 99, we estimate a risk ratio of death within 5 years of .994 between groups with a mean serum LDL level 1 mg/dL different, with the group with the higher mean serum LDL having a lower risk of death within 5 years. Based on a 95% confidence interval, the observed results would not be surprising if the true risk ratio between these groups is between .988 and .999. These results are statistically significant with a p value of 0.018.

c.

Logistic regression analysis allowing for heteroscedasticity was performed by comparing the risk of death within 5 years across groups defined by the serum LDL level.

From a linear regression analysis of 725 available observations from a sample of 735 elderly subjects between ages 65 and 99, we estimate a odds ratio of death within 5 years of .992 between groups with a mean serum LDL level 1 mg/dL different, with the group with the higher mean serum LDL having a lower risk of death within 5 years. Based on a 95% confidence interval, the observed results would not be surprising if the true risk ratio between these groups is between .986 and .999. These results are statistically significant with a p value of 0.019.

d.

The overall conclusion of these three analyses is that there is an association between risk of death within 5 years and LDL that is statistically significant at the p=.05 level, which is similar to the conclusions drawn by problems 2 and 4 of HW2. However, this is different than the conclusion from problems 1-3, which did not find a statistically significant association between risk of death within 5 years and LDL when LDL was dichotomized based on having high LDL vs low LDL. *A priori*, I would have selected to examine the risk difference using robust linear regression of the distribution of death within 5 years and groups defined by the continuous measure of their LDL. This ensures I do not lose any information from dichotomizing, and examining the risk difference accentuates the public health impact in question – the risk of death within 5 year with increased LDL.