Homework 5 - Biost 536

John Schoof

11/11/2020

1. *Fit a logistic model relating binary cider exposure (> 10 g/day) to the risk of esophageal cancer, using grouped-linear adjustment for age and the six age groups in the variable agegp. Write a sentence appropriate for the results section of a scientific paper reporting the results of this analysis. Note, here and below: your sentence should include both a point estimate and CI for the parameter(s) of interest.*

$$\text{Logit}(Case = 1|cider, agegroup) = \beta_0 + \beta_1 * cider + \beta_2 * agegroup$$

I ran the above logistic regression of the log odds of being diagnosed with esophageal cancer based on a sample of 975 individuals using robust standard errors. I estimate that, on average, the odds of a high consumer of cider being diagnosed are 2.187 (95% CI: 1.553, 3.08) times greater than the odds of a low consumer of cider being diagnosed for individuals of the same age group. This finding is statistically significant at the alpha = 0.05 confidence level.

2. *Fit a logistic model relating binary cider exposure (> 10 g/day) to the risk of esophageal cancer using dummy variables to adjust for age and the six age groups in the variable "agegp". Write a sentence appropriate for the results section of a scientific paper reporting the results of this analysis.*

$$\text{Logit}(Case = 1|cider, age2, age3, age4, age5, age6) \\ = \beta_0 + \beta_1 * cider + \beta_2 * age2 + \beta_3 * age3 + \beta_4 * age4 + \beta_5 * age5 + \beta_6 * age6$$

I ran the above logistic regression of the log odds of being diagnosed with esophageal cancer based on a sample of 975 individuals using robust standard errors. I estimate that, on average, the odds of a high consumer of cider being diagnosed are 2.027 (95% CI: 1.444, 2.847) times greater than the odds of a low consumer of cider being diagnosed for individuals of the same age group. This finding is statistically significant at the alpha = 0.05 confidence level.

3a. Compare the results for the exposure variable in the Q1 and Q2 analyses. Are results similar or very different?

Adjusting for the grouped linear age variable in question 1 provides an estimated OR comparing low and high consumers of cider of 2.19 (95% CI: 1.55, 3.08). Adjusting for age using dummy variables results in an estimated OR of 2.03 (95% CI: 1.44, 2.85). While we wouldn't expect these estimates to be the same because the grouped linear adjust forces more structure on the model, the two estimates are still quite similar.

3b. Which result would you prefer to report in a scientific article, and why?

I prefer the result from the age dummy variable adjustment approach in question 2 because it forces less structure on the model. Both approaches force the effect of age on cider consumption to be the same across all age groups, but in the dummy variable approach the effects across age groups are able to vary.

4. *Fit a logistic model relating binary cider exposure (> 10 g/day) to the risk of esophageal cancer using continuous linear adjustment for age. Write a sentence appropriate for the results section of a scientific paper reporting the results of this analysis.*

$$\text{Logit}(Case = 1|cider, age) = \beta_0 + \beta_1 * cider + \beta_2 * age$$

I ran the above logistic regression of the log odds of being diagnosed with esophageal cancer based on a sample of 975 individuals using robust standard errors. I estimate that, on average, the odds of a high consumer of cider being diagnosed are 2.168 (95% CI: 1.539, 3.055) times greater than the odds of a low consumer of cider being being diagnosed for individuals of the same age. This finding is statistically significant at the alpha = 0.05 confidence level.

5. *Fit a logistic model relating binary cider exposure (> 10 g/day) to the risk of esophageal cancer using continuous quadratic adjustment for age. Write a sentence appropriate for the results section of a scientific paper reporting the results of this analysis.*

$$\text{Logit}(Case = 1|cider, age) = \beta_0 + \beta_1 * cider + \beta_2 * age + \beta_3 * age^2$$

I ran the above logistic regression of the log odds of being diagnosed with esophageal cancer based on a sample of 975 individuals using robust standard errors. I estimate that, on average, the odds of a high consumer of cider being diagnosed are 1.969 (95% CI: 1.4, 2.77) times greater than the odds of a low consumer of cider being being diagnosed for individuals of the same age. This finding is statistically significant at the alpha = 0.05 confidence level.

6. *Fit a logistic model relating binary cider exposure (> 10 g/day) to the risk of esophageal cancer using linear spline adjustment for age. Use the same age groups as for the ageg variable for your splines.*

$$\text{Logit}(Case = 1|cider, age) = \beta_0 + \beta_1 * s_1 + \beta_2 * s_2 + \beta_3 * s_3 + \beta_4 * s_4 + \beta_5 * s_5 + \beta_6 * s_6$$

6a. Write 1-2 sentences describing the analysis (not the results) that would be appropriate for the methods section of a scientific paper.

To adjust for age using linear splines I created six intervals of age using the same age groups defined by the variable “agegp.” The five knots are at ages 35, 45, 55, 65, and 75.

6b. Write a sentence appropriate for the results section of a scientific paper reporting the results of this analysis.

I ran the above logistic regression of the log odds of being diagnosed with esophageal cancer based on a sample of 975 individuals using robust standard errors. I estimate that, on average, the odds of a high consumer of cider being diagnosed are 1.983 (95% CI: 1.41, 2.788) times greater than the odds of a low consumer of cider being diagnosed among those in the lowest age interval. This finding is statistically significant at the alpha = 0.05 confidence level.

7a. Compare the OR estimates from Q4, Q5, and Q6. Are your results similar or very different?
My age-adjusted OR estimates comparing high and low cider consumers from Q4, Q5, and Q6 are 2.17, 1.97, and 1.98, respectively. It makes sense that the model with the quadratic term and the linear spline model are almost the same because they allow for more flexibility. While

OR estimate from the model with continuous linear adjustment is slightly different that the other two, it is still very similar and captures the general trend.

7b. Which approach would you prefer if you were studying this exposure and wanted to adjust for age as a potential confounder, and why?

I prefer the model from Q4 with continuous linear adjustment for the confounding variable, age. I prefer this model because it captures the same trend, is easier to interpret, and there is less risk that it will overfit the data. Therefore, it is likely more generalizable. I do not expect any of the three models to estimate the true OR and therefore I value the overall trend and generalizability over flexibility in this case.

8a. Are the models in Q1 and Q2 nested? No

8b. Are the models in Q1 and Q4 nested? No

8c. Are the models in Q4 and Q5 nested? Yes, 4 is nested in 5. If the quadratic term in the Q5 model was eliminated then we would be left with the model from Q4.

8d. Are the models in Q4 and Q6 nested? No

8e. Are the modest in Q5 and Q6 nested? No