Biostat 518

3/5/2015

HW 06

1. A. **Methods:** Ten patients were missing data on LDL and were excluded from analysis, leaving us with 725 patients. We used proportional hazards regression to examine the association between time to death and level of LDL within an elderly population, modeling LDL as both a continuous and as a dummy variable. LDL levels were centered for the continuous variable, by subtracting 1mg/dl from every measurement. The categories for the step-wise function were: 11-70, 70-100, 100-130, 130-160, and 160-247mg/dl. We used a post-hoc Wald test to check if the dummy variables improved the fit of the model.

**Inference:** 131 of 725 patients were observed to die during the study period. Overall, we found a significant association between LDL levels and mortality (two-sided p<.01), but no evidence that this association is non-linear in a way that a step-wise function would improve the model (two-sided p=.3609).

B. **Methods:** Ten patients were missing data on LDL and were excluded from analysis, leaving us with 725 patients. We used proportional hazards regression to examine the association between time to death and level of LDL within an elderly population, modeling LDL as a quadratic function. LDL levels were centered before analysis, by subtracting 1mg/dl from every measurement. We used a post-hoc Wald test to see if adding the quadratic term improved our model.

**Inference:** 131 of 725 patients were observed to die during the study period. Overall, we found a significant association between LDL levels and mortality (two-sided p<.001), but just narrowly missed establishing a statistically significant impact from the quadratic term. As it was, the effect of the quadratic term was not statistically distinguishable from zero and we fail to reject the null hypothesis that the association between mortality and LDL is non-linear (p=.055).

C. **Methods:** Ten patients were missing data on LDL and were excluded from analysis, leaving us with 725 patients. We used proportional hazards regression to examine the association between time to death and level of LDL within an elderly population, modeling LDL as a cubic function. LDL levels were centered before analysis, by subtracting 1mg/dl from every measurement. We used a post-hoc Wald test to see if modeling LDL as a cubic was an improvement over a linear model.

**Inference:** 131 of 725 patients were observed to die during the study period. Overall, we found a significant association between LDL levels and mortality (two-sided p<.0001), and evidence that the association is non-linear and that modeling LDL as a cubic function improves the model (two-sided p<.01).

d. **Methods:** Ten patients were missing data on LDL and were excluded from analysis, leaving us with 725 patients. We used proportional hazards regression to examine the association between time to death and level of LDL within an elderly population, modeling LDL as both a continuous variable and with linear splines. LDL levels were centered for the continuous variable, by subtracting 1mg/dl from every measurement. The nodes for the spline were at: 11, 70, 100, 130, 160, and 247mg/dl. We used a post-hoc Wald test to see if modeling LDL with linear splines was an improvement over a linear model.

**Inference:** 131 of 725 patients were observed to die during the study period. Overall, we found a significant association between LDL levels and mortality (two-sided p<.0001), but found no evidence that this association was non-linear (two-sided p=.119), though we still cannot rule out the possibility of non-linearity.

e. **Methods:** Ten patients were missing data on LDL and were excluded from analysis, leaving us with 725 patients. We used proportional hazards regression to examine the association between time to death and level of LDL within an elderly population, modeling LDL as both a continuous and continuous log-transformed variable. LDL levels were centered for the continuous variable, by subtracting 1mg/dl from every measurement. We used a post-hoc Wald test to see if log-transforming LDL improved the linear fit of our regression.

**Inference:** 131 of 725 patients were observed to die during the study period. Overall, we found a significant association between LDL levels and mortality (two-sided p<.0001), and evidence that log-transforming LDL improved the fit (two-sided p<.005). We reject the null hypothesis that the association between LDL and mortality is purely linear.

f. The spline, quadratic, and cubic models are all roughly similar, and as the spline allows for the most flexibility, it is likely to be the one that most closely fits the actual data. The log-transformed model is unable to bend with the apparent curve in the data, and the step-wise model shows a far more drastic jump in the hazard ratio for the lower end of the ldl spectrum than the other models, leading me to think that perhaps the spline, cubic, and quadratic functions are not quite flexible enough to show the true hazrd of low LDL. What this graph clearly demonstrates, is that the choice you make for modeling your predictor can lead to some big differences in your outcome. The predicted difference for an LDL level of 1 is around .1 in the log-transformed model, and nearly 1 in the model with dummy variables!



1. a. When I tested the association between mortality and LDL using linear splines with a proportional hazards regression in STATA 13.1, I did not get an intercept in my output but if I had it would have represented mortality associated with having LDL of 0mg/dl (a physically improbable if not impossible value). The slopes that I did get in my output represent the slope of the line fit between the nodes of the spline. The nodes for the spline were at: 70, 100, 130, 160, and 400mg/dl, thus the estimated hazard ratio for two groups differing by 1 mg/dl in LDL with an LDL between 0 and 70 is .978, two groups differing by 1mg/dl in LDL with an LDL between 70 and 100 is .979, the slope for people with LDL 100-130mg/dl is .999, an LDL of 130-160mg/dL is .998, and 160-400 is .994.

b. We found no evidence of a u-shaped curve with this sample. The slope of the line for all levels of LDL remained positive, and only increased slightly with increasing LDL. It could be that this is one half of a U-shaped curve, and that the other part of the curve lies in a range beyond the scope of our data or physical possibility.

3. a. The estimated hazard ratio for whites differing in 1 log mg/dl of LDL is .461. The hazard ratio for a black person with an LDL level of 0 compared to a white person with LDL 0 is .154. For an Asian person compared to a white person where both have LDL = 0, their HR is 304.98, and for someone in an other racial category compared to a white where both have LDL = 0, their HR is 3.33\*10^8. 1.55 is the hr for comparing a black and a white person with LDL levels of 0 to a black and white person of LDL = 1 (a difference of differences). .31 is the similar difference of differences for an Asian and a white person, and .0179 is the difference of differences for someone in an “other” racial category and a white person.

b. I tested the interaction term post-hoc to see if race modified the association between LDL and time to death. The resulting p-value was right on the edge of statistical significance at two-sided p=.045. We fail to reject the null hypothesis that there race does not modify the association between LDL and time to death.

c. I performed a post-hoc Wald test examining log LDL as it relates to time to death with no other factors, and found a significant association (two-sided p<.001). We reject the null hypothesis of no association between LDL and mortality.

d. Testing the association between race and time to death without any other interactions revealed a statistically significant effect of race on mortality, two-sided p<.023.

e. A post-hoc Wald test was used to test if the coefficient in our model for blacks was equivalent to the coefficient for whites. We did not find a statistically significant difference, two-sided p=.644.

4. a. No, from a linear regression examining differences in mean monthly salary for 485 faculty hired between 1990-1995 who were within 1 year of finishing their studies there was not a statistically significant difference in salary between males and females, taking into account a possible interaction with year (p=.895). We also used a clustered analysis, as we were looking at the salary for many faculty members over time and an individual’s salary at one point in time should be highly correlated with their salary a month or even many months later. Year had a statistically significant effect on monthly salary, but there was no interaction between year and sex (p=.987).

b. Using methods similar to those above, but substituting log-transformed monthly salary for the response still did not find a significant effect of sex on monthly salary when controlling for year, repeated measures, and an effect modification of sex on year (p=.602). There was also no interaction between sex and year (p=.684).

c. With part a, I’ve calculated the average dollar amount difference between men and women’s salaries, which would make it easy to adjust pay scales if that were my goal. I know how much more I should be paying female faculty members. With part b, knowing the percent difference gives better perspective on how large the difference is relative to average salaries, and is what I would use if reporting the results as just general evidence of sex bias.

d. My point estimates would not change, but failing to adjust for correlated measurements does affect the standard error, which in turn affects the confidence interval and p values. Failing to adjust for correlated measurements can lead to conservative or anti-conservative estimates; in this case it would have led to conservative estimates, and the CIs broadened and p values got bigger when analyses were run without clustering.