Biostatistics 515/518

Winter 2015

Homework #3

3763

1. Planned analysis: Looking at associations with the outcome of delivery of babies that are small for gestational age. Descriptive statistics.

Table 1 presents the descriptive statistics for mothers who did and did not have babies who were small for gestational age (SGA) and for the cohort as a whole. There were a total of 755 mothers in our cohort. One-hundred five (13.9%) delivered babies that were SGA and the remaining 650 (86.1%) delivered infants who were normal for gestational age (NGA). There were a few (≤ 6) missing values for several variables and the missing values were not included in the analysis. The total number of available values for each variable is included in the table. If not listed, there were no missing values for that variable. The categories for the variable parity were created to divide the population roughly equally into mothers who had no previous deliveries, 1 previous delivery, or 2 or more previous deliveries.

Based on Table 1, factors that appear to be associated with having an SGA infant include lower age at enrollment, no previous deliveries, being a smoker, lower gestational age at delivery, female sex, and lower birthweight. Factors that seem to be associated with having an NGA infant include older age at enrollment, having ≥ 2 previous deliveries, being a non-smoker, male sex, higher gestational age at delivery and higher birthweight.

When looking at infant characteristics, birthweight is part of the definition of SGA and is thus not a confounder because it is in the pathway of interest. Gestational age at delivery is a possible confounder in the association between any maternal factor and SGA because maternal factors that may be associated with pre-term delivery may also be associated with having an SGA infant. If it is more common for SGA infants to be female, the sex of the child is likely a precision variable since it likely has no association with any maternal factors but appears associated with our outcome of interest.

When looking at interactions between maternal variables, age is probably a confounder in the association between parity and SGA. Younger mothers are more likely to be delivering their first child and older mothers more likely to be delivering their 2nd or more child. Other interactions between maternal variables are not as obvious and would require further analysis to determine.

 **Table 1: Maternal and infant characteristics in women with and without SGA infants.**

|  |  |  |  |
| --- | --- | --- | --- |
| ***Maternal characteristics*** | **SGA infants****N=105** | **Non-SGA infants****N=650** | **All infants****N=755** |
|  **Age at enrollment (years)** | Mean (SD) | 23.8 (4.9) | 25 (5.4) | 24.8 (5.4) |
|  **Height (cm) N=749** | Mean (SD) | 154.6 (5.9) | 157 (6.5) | 156.7 (6.5) |
|  **Previous deliveries** **0** **1** **≥ 2** | N (%) | 49 (46.7%)32 (30.5%)24 (22.9%) | 244 (37.5%)208 (32%)198 (30.5%) | 293 (38.8%)240 (31.8%)222 (29.4%) |
|  **Smoker N=751** | N (%) | 45 (42.9%) | 186 (28.6%) | 231 (30.6%) |
| ***Infant characteristics*** |
|  **Male N=751** | N (%) | 44 (41.9%) | 339 (52.2%) | 383 (50.7%) |
|  **Gestational age at delivery (weeks) N=750** | Mean (SD) | 37.9 (2.2) | 39.4 (1.2) | 39.2 (1.5) |
|  **Birthweight (gms) N=751** | Mean (SD) | 2231.1 (411.6) | 3236.2 (402.1) | 3105.6 (534.5) |

1. Regression analysis looking at association between maternal smoking and SGA using odds/ORs.
	1. Full inference regarding the association between SGA and maternal smoking.
		1. Methods: Logistic regression with robust standard deviations was used to calculate the odds ratio with a 95% confidence interval and 2-sided p-value for the odds ratio for SGA between smoking mothers and non-smoking mothers.
		2. Results: The odds of a smoking mother having an SGA infant was 1.89 times the odds for a non-smoking mother. With a p-value of 0.003, we can reject the hypothesis that there is no difference in odds of having an SGA infant in smoking mothers when compared to non-smoking mothers. These results would not be unusual if the true population odds were between 1.24 and 2.89 times higher in smoking mothers.
	2. Estimates of the odds and probability of delivering an SGA infant in smokers and non-smokers.
		1. In non-smokers:
			1. Odds of SGA infant= e-2.0559 = 0.128
			2. Probability = [0.128/(1+0.128)] = 0.113 = 11.3%
		2. In smokers
			1. Odds = e(-2.0559 + 0.6368) = 0.242
			2. Probability = [0.242/(1+0.242)] = 0.195 = 19.5%
		3. When comparing these results to the numbers from table 1, we can see that in 231 smokers had 45 SGA infants for a proportion of 45/231 = 19.5% and odds of 45/(231-45) = 0.242(same as probability and odds above) and 520 non-smokers had 60 SGA infants for a proportion of 11.3% and odds of 60/(520-60) = 13% (again, same as probability and odds above).
	3. Explain the similarities and differences between the estimates and inference for the following additional models: These are used to trade the places of the predictor and the outcome, and illustrate the invariance properties of the odds ratio.
		1. NONSMOKER (predictor), SGA (response)
			1. The p-value will be the same and the confidence interval will still not cross 1.
			2. The OR will be < 1 (likely almost half the risk since the previous analysis gave us almost twice the risk)
		2. SMOKER (predictor), NOTSGA (response)
			1. The p-value will not change and the confidence interval will still not cross 1.
			2. The OR will be < 1 (likely almost half the risk since the previous analysis gave us almost twice the risk)
		3. NONSMOKER (predictor), NOTSGA (response)
			1. This will give us the exact same results as looking at smoker(predictor) and sga(response).
			2. The odds of not having an sga infant if you are a nonsmoker are the same as having an sga infant if you are a smoker.
2. Regression analysis evaluating association between smoking and sga by using probabilities of SGA across smoking groups.
	1. Full inference regarding SGA and Maternal Smoking.
		1. Methods: A linear regression with robust standard errors was performed to calculate the probability having an SGA infant in smoking mothers when compared with the non-smoking mothers. Inference on the difference in probabilities was determined from the standard error for the slope, which was estimated using the Huber-White sandwich estimator with 95% confidence interval and two-sided p-value for the difference in probabilities between the groups. The odds within each group was then calculated from the probability using (odds=p/(1-p)).
		2. Results: Because this is a saturated model, the probability of SGA in the smoking group will be equal to the proportion of SGA in the smoking group in the sample, and the same for the non-smoking group. The intercept from the linear regression is 0.113, which corresponds to an 11.3% probability of SGA in non-smokers. The slope is 0.0813, which corresponds to an 8.13% higher probability of SGA in the smoking versus non-smoking group. When converted to odds, the odds of SGA in the smoking group are 24.1% and in the non-smoking group 12.7%, for a difference of 11.4% higher in the smoking mothers. These results would not be unusual if the true population difference in probabilities were between 2.3% and 13.9% higher in the smoking mothers. With a p-value of 0.006, we can reject the null hypothesis that there is no difference in the probability or odds of SGA in smoking mothers when compared to non-smoking mothers.
	2. Estimates of the odds and probability of delivering an SGA infant in smokers and non-smokers.
		1. In non-smokers:
			1. Odds of SGA infant= 0.113/(1-0.113) = 12.7%
			2. Probability = 11.3%
		2. In smokers
			1. Odds = 0.194/(1-0.194)= 24.1%
			2. Probability = 19.4%
		3. When comparing these results to the numbers from table 1, the results are essentially identical.
	3. Explain the similarities and differences between the estimates and inference for the following additional models:
		1. NONSMOKER (predictor), SGA (response)
			1. The intercept will give the probability of SGA in smokers (19.4%) and the slope (negative in this case) will give the difference between the probability in smokers and non-smokers.
			2. The p-value is identical.
		2. SMOKER (predictor), NOTSGA (response)
			1. The intercept gives the probability of having a non-SGA infant if you are a non-smoker (1-0.113) and the slope gives the difference in probability of having a non-SGA infant if you are a smoker (which is the same as the difference in the probability of having an SGA infant if you are a smoker).
			2. The P-value is identical.
		3. NONSMOKER (predictor), NOTSGA (response)
			1. The intercept gives the probability of having a non-SGA infant if you are a smoker and the slope (identical but in the opposite direction as the previous model) the difference between the probabilities in smokers and non-smokers.
			2. The p-value and confidence interval are identical.
3. Association between the odds of delivery of SGA infants and maternal smoking by looking at the ratio of probabilities for SGA across smoking groups.
	1. Full inference regarding SGA and Maternal Smoking.
		1. Methods: A poisson regression with robust standard errors was performed to calculate the ratio of the probability of having an SGA infant in smoking mothers when compared with the non-smoking mothers. Inference on the ratio was determined from the standard error for the slope, which was estimated using the Huber-White sandwich estimator with 95% confidence interval and two-sided p-value for the ratio of probabilities.
		2. Results: The intercept from the poisson regression is -2.176 which, when exponentiated, gives the probability of an SGA infant in non-smokers (11.3%). The slope of 1.717 tells us that the probability of having an SGA infant in a smoker is 1.7 times higher than in a non-smoker (0.113 X 1.717), or 19.5%. These results would not be unusual if the true population ratio of probabilities were between 1.2 and 2.5 times higher in the smoking mothers. With a p-value of 0.003, we can reject the null hypothesis that there is no difference in the probability or odds of SGA in smoking mothers when compared to non-smoking mothers.
	2. Estimates of the odds and probability of delivering an SGA infant in smokers and non-smokers.
		1. In non-smokers:
			1. Odds of SGA infant= 0.113/(1-0.113) = 12.7%
			2. Probability = e-2.176291 = 11.3%
		2. In smokers
			1. Odds = 0.194/(1-0.194)= 24.1%
			2. Probability = e(-2.176291+0.5405362) = 19.4%
		3. When comparing these results to the numbers from table 1, the results are essentially identical.
	3. Explain the similarities and differences between the estimates and inference for the following additional models:
		1. NONSMOKER (predictor), SGA (response)
			1. The intercept will give the probability of SGA in smokers (e-1.635755 = 19.4%) and the slope (negative in this case) will give the probability ratio in smokers and non-smokers. The slope is identical to the initial model but in the opposite direction.
			2. The p-value is identical.
		2. SMOKER (predictor), NOTSGA (response)
			1. The intercept gives the probability of having a non-SGA infant if you are a non-smoker e-.0.1204308 = 88.7% and the slope gives the probability ratio of having a non-SGA infant if you are a smoker (which is the same as the probability ratio of having an SGA infant if you are a smoker).
			2. The Z-statistic is less negative, P-value is slightly larger, at 0.007 instead of 0.003, and confidence interval slightly wider than in the previous test.
		3. NONSMOKER (predictor), NOTSGA (response)
			1. The intercept gives the probability of having a non-SGA infant if you are a smoker and the slope (identical but in the opposite direction as the model in ii) gives the ratio of probabilities between smokers and non-smokers.
			2. p-value and confidence interval are identical to the model in ii but different than the p-value in confidence intervals in i and the original model.
4. When comparing the tests above to the t-test, Chi-square test, risk difference, risk ratio, and odds ratio that we have learned previously, the following similarities and differences are apparent
	1. Chi-squared test:
		1. All of the tests give the probabilities of SGA obtained via a Chi-squared test (11.4% in non-smokers and 19.5 in smokers).
		2. When comparing the linear regression to the chi-squared test, the slope is equal to the risk difference and the confidence interval for the slope identical to the confidence interval for the risk difference. The p-value for the linear regression is 0.006, compared to the p-value of 0.003 for the chi-squared test.
		3. The logistic regression gives the same odds ratio as the odds ratio from the two-sample test, the same confidence interval around the odds ratio, and the p-value is also the same (0.003). It is identical to the chi-squared test.
		4. The slope in the poisson regression is identical to the risk ratio from the two-sample test. The 95% confidence interval around the slope is also identical to the 95% confidence interval around the risk ratio. The p-value in poisson is 0.007, compared to 0.003 for the chi-squared test.
	2. T-test
		1. The probabilities are identical in all tests above (linear, logistic, poisson regression).
		2. The difference in means in the t-test (a difference in proportions given this is a binary variable) is the same as the slope in the linear regression and the 95% CI and p-value are identical to the linear regression. These tests are identical as both use the t-distribution as their reference distribution.
5. Regression analyses of distribution of prevalence of SGA in infants across groups defined by continuous measure of mother’s age.
	1. Using risk difference (difference in probabilities).
		1. Methods: Simple linear regression was performed for the probability of SGA infant across groups defined by maternal age. A 95% confidence interval and twp-sided p-value were calculated for the difference.
		2. Results: For every 1 year increase in maternal age, the risk of having an SGA infant decreased by 0.43%. These results would not be unusual if the true risk difference were between 0.0003% and 0.85% lower for every 1-year increase in age. With a p-value of 0.05, we can reject the null hypothesis that there is no difference in the risk of having an SGA infant among women of different ages.
	2. Using risk ratio (ratio of probabilities).
		1. Methods: Poisson regression with robust standard deviation was used to calculate the proportional risk (risk ratio) of having an SGA infant in groups defined by maternal age. A 95% confidence interval and two-sided p-value were calculated for the ratio.
		2. Results: For every year increase in age, there is proportional risk reduction of having an SGA infant of 3.2%. This result would not be unusual if the true proportional risk were between 6.4% lower and 0.2% higher for every year increase in age. With a p-value of 0.061, we cannot reject the null hypothesis that there is no difference in risk of SGA infant in groups of mothers that differ by 1 year in maternal age.
	3. Using odds ratio.
		1. Methods: Logistic regression was performed to estimate the odds of having an SGA infant in groups defined by differences in mother’s age. A 95% CI and two-sided p-value were calculated for the odds.
		2. Results: For every 1 year increase in age, there was a 3.7% decrease in the odds of having an SGA infant. This difference would not be unusual if the difference in odds were between 7.5% lower and 0.3% higher in women who are 1 year older. With a p-value of 0.07, we cannot reject the null hypothesis that there is no difference in the odds of having an SGA infant in groups of mothers that differ by 1 year in maternal age.
	4. Estimate of risk/odds for a 20-year old mother given by each of the models above.
		1. Model 1: Linear regression: Risk = 0.2441626 + (20 (0.-0042665)) = 15.9%
		2. Model 2: Poisson regression: Risk = e(-1.185715 + (20X(-0.0325385)) = 15.9%
		3. Model 3: Logistic regression: Odds = 0.4019534 X (0.9631149)20 = 0.19, probability = 0.19(1+0.19) = 15.9%.
		4. The estimates are identical for all of the groups. I believe that this is because this is a saturated model and therefore the risk and odds in any regression model should match the risk and odds in the sample exactly.
6. Estimated probability of SGA infants by maternal age.
	1. Proportion of SGA infants by age



* 1. Estimated probabilities for each age in the data as derived from each of the three regression analyses.
		1. Linear regression: Estimated risk of SGA by maternal age



* + 1. Poisson regression: Estimated proportion of SGA by maternal age



* + 1. Logistic regression: Estimated proportion of SGA by maternal age



iv: Multiple models on the same graph of proportion or estimated proportion of SGA infants by maternal age. When comparing these results to the results from problem 6, we can see that by choosing an age in the middle of the age distribution, rather than the edge of the distribution, we chose a point at which all 3 regression models give essentially the same result. If we had chosen an age at one of the tails of the distribution, the logistic or poisson models (which give a curvilinear association) would have been different (higher) than the linear estimate.



* + 1. Methods: Logistic regression was performed to estimate the odds of having an SGA infants on groups defined by log-transformed age. A 95% confidence interval and two-sided p-value were calculated for the difference in odds.
		2. Results: For every 1 unit difference in log-transformed age, there was a 59.4% relative decrease in the odds of having an SGA infant. This finding would not be unusual of the true difference in odds was between 85% lower and 9.5% higher per unit of log-transformed age. With a p-value of 0.075, we cannot reject the null hypothesis that there is no difference in the odds of having an SGA infant across groups defined by log-transformed age.
	1. I would not have chosen to log-transform age because I do not believe that age acts multiplicatively on the outcome of having an SGA infant. It seems to act more additively.