Biostat 518

1/27/15

HW #3

1. **Methods:** Descriptive statistics (mean, standard deviation, minimum, maximum and percentages for non-continuous variables) were calculated for groups based on smoking status. Of 755 subjects followed during pregnancy and delivery in Western Cape, South Africa, four were missing data on smoking status, birthweight, and gestational age at time of delivery. One additional woman was missing gestational age, and six subjects appeared to have corrupt data (example: smoking status coded as “3” rather than the specified 1 or 2). These observations were excluded from all analyses, resulting in a total of 744 subjects.

**Inference:** In our sample of 744 mothers, 229 were smokers and 515 were classified as non-smokers. Average age, height, parity, and gestational age were similar across groups, but there are larger disparities for birthweight of child and proportion of infants termed small for gestational age (SGA) between smokers and non-smokers. Explicitly, average birthweight of children born to smoking mothers was 2976.38gm compared to 3171.14gm for non-smoking mothers leading to a diagnosis of SGA for 18.78% of smoking mothers and 10.49% of non-smoking mothers. There was also a slight difference in the proportions of the sex of the infants born to smoking and non-smoking mothers, that could be a possible confounder as we explore the effect of smoking on birthweight as male infants tend to be a little heavier than females, and more males were born to the non-smoking group (52.23% male vs. 48.47%).

**Table 1:** Descriptive statistics of sample.

|  |  |  |  |
| --- | --- | --- | --- |
| **Smoking status (n)** | **Smokers (229)** | **Non-smokers (515)** | **Total (744)** |
| Age (yr)\* | 25.06(5.31), 15-42 | 24.58(5.38), 14-43 | 24.73(5.36), 14-43 |
| Height of mother (cm)\* | 156.82(7.20), 106-176 | 156.64(6.16), 127-175 | 156.69(6.49), 106-176 |
| Parity (n)\* | 1.18(1.26), 0-6 | 1.05(1.19), 0-6 | 1.09(1.21), 0-6 |
| Birthweight (gm)\* | 2976.38(511.67), 1410-4550 | 3171.14(532.59), 1035-4730 | 3111.19(533.53), 1035-4730 |
| Sex of infant (% male) | 48.47 | 52.23 | 51.08 |
| Gestational age (weeks)\* | 38.97(1.36), 33-43 | 39.29(1.55), 30-44 | 39.19(1.50), 30-44 |
| SGA (% yes) | 18.78 | 10.49 | 13.04 |

\*Statistics presented are mean(standard deviation), min-max.

1. **Methods:** Logistic regression was used to analyze any differences in the odds of giving birth to an SGA infant for smoking and non-smoking mothers. There were 229 smoking subjects and 515 non-smokers included in this analysis: 11 subjects were excluded for missing or corrupt data as described in problem 1.
2. **Inference:** We estimate with logistic regression that the odds of giving birth to an infant of SGA are 97% higher for smokers than for non-smokers (p<.002). A 95% CI indicates this finding would not be unusual if the true value were between 28% and 204%.
3. **Methods:** The odds of giving birth to an SGA infant were calculated for smoking and non-smoking mothers using the logistic regression model from part a. From those estimates, probabilities were than calculated based on the formula odds = p/(1-p) where p = proportion of SGA infants born. This proportion approximates the probability of having an SGA infant.

**Inference:** The estimated odds of giving birth to an SGA baby for non-smoking mothers is .1171, which is approximately 3 to 25 against. This results in an estimated probability of 10.49%, which is exactly equal to the proportion we found in our descriptive analyses of problem 1. The estimated odds of having an SGA baby for smoking mothers is .2312, which is nearly 1 to 4 against. This translates to a probability of 18.78%, again equal to the proportion found in problem 1. This similarity is due to the fact that our model was derived from these proportions, and thus reflects them accurately.

1. **i. Methods:** Smoking status was recoded such that smokers became group two and non-smokers became group one. Then a logistic regression was performed as before examining probability of SGA infants based on smoking status.

**Inference:** Reversing the order of categories for the predictor variable results in a fitted model where the slope of the line is the negative of that in the original model. The y-intercept also changes from the lowest to the highest predicted value. The odds ratio is then reversed to .5067, so the odds of a non-smoking mother giving birth to an SGA infant are only 50.67% that of smoking mothers.

ii. **Methods:** Data was recoded such that the order of groups based on SGA of infants was reversed during logistic regression.

**Inference:** This reversal led to the same results as in part i.: that is, the resulting logistic model has the negative slope of our original model, and the y-intercept also changes from the lowest to the highest predicted value. This estimates the odds having an SGA infant among non-smoking mothers as 50.67% of that among smoking mothers.

iii. **Methods:** Data was recoded as described in parts i. and ii. such that categories were reversed for both the response and the predictor variables (smoking status and SGA diagnosis) before performing logistic regression.

**Inference:**  Switching both response and predictor categories results in a logistic model identical to the one in our initial analyses with the odds of giving birth to an SGA infant 97% higher for smoking than for non-smoking mothers.

1. **Methods:** A linear regression was performed using robust standard error from the Huber-White sandwich estimator on SGA as predicted by smoking status to estimate the difference in probabilities across groups. Wald-based 95% confidence intervals and a two-sided p value were also computed. The formula p1(1-p2)/p2(1-p1) where p1 represents the proportion of SGA births for smoking mothers and p2 represents the proportion of SGA births for non-smoking mothers was used to calculate the odds ratio for the two groups.

**Inference:** The estimated difference in probability of giving birth to an SGA infant for smoking and non-smoking mothers was .0823, p<.005. A 95% CI suggests this difference would not be unusual if the true difference lies between .026 and .140. This estimated difference is exactly equal to the difference between observed proportions in our descriptive analyses, as again, our linear model was built off of these values. The estimated odds ratio predicts that the odds for smoking mothers of giving birth to a SGA infant are 97% higher than for non-smoking mothers.

c. Much like with logistic regression, the slope of our model would change signs if we reversed the order of the groups for either the predictor or response variables, and as the slope represents our estimated difference, that would also change signs. Similarly, the odds ratio would be reversed to indicate the odds of a non-smoking mother giving birth to an SGA infant to be 50.67% that of the odds of a smoking mother. Again, if we reversed the order for both groups this would get the same results as in our original analysis.

1. **Methods:** A Poisson regression was performed on SGA as predicted by smoking status in order to evaluate the ratio of probabilities between groups. Missing values were discarded as described in problem 1.

**Inference:** The estimated ratio of probabilities of giving birth to an SGA infant between smoking and non-smoking mothers was 1.79, so the risk of having an SGA infant is 79% higher for smoking mothers with p<.004 which corresponds exactly to the proportions observed in our descriptive analyses as the Poisson model was based on these observations. A 95% Wald-based confidence interval indicates this ratio would not be unusual if the true risk ratio were between .18 and .98.

c. As in the cases listed above, reversing the order of the groups for either the response or predictor variables would result in a negative slope in our fitted model. This would reverse our estimated risk ratio, so non-smoking mothers would be estimated to be at only 56% the risk of smoking mothers of giving birth to an SGA infant. Again, reversing the order of groups in both the response and the predictor variables would obtain the same results as our original model.

1. A simple two-sample comparison of SGA based on smoking status could be done with a Chi-squared test with one degree of freedom. This test is similar in that it examines observed differences between the two groups in order to determine whether these differences are statistically significant, but unlike regression analyses, does not create a model for estimating the response variable based on the predictor variable. Using the proportions of observed values from each group, one could calculate difference in proportions, risk ratio, and odds ratio and obtain point estimates identical to those reported above, but without a confidence interval.
2. A. **Methods:** We used linear regression with standard error based on the Huber-White sandwich estimator to model the probability of SGA births based on the mother’s age. Wald-based confidence intervals and a 2-sided p value were also calculated.

**Inference:** Our estimated difference in probability of SGA for groups differing by one year of age is -.0056 (p<.007), such that for a group their probability of giving birth to an SGA infant is .0056 less than for a group of women one year younger. We estimate that this difference would not be unusual if the true difference were between -.0097 and -.0016.

**b.** **Methods:** A Poisson regression was performed on SGA as predicted by age in order to evaluate the ratio of probabilities between groups.

**Inference:** For a group of women the same age, their estimated risk of having an SGA infant is 4.6% lower than a group of women one year younger (p<.023). A 95% CI suggests this data would not be unusual if the true risk decreased from between 91.63% to 99.35% for groups differing by 1 year of age.

**c.** **Methods:** We used Logistic regression to model SGA as predicted by age in order to evaluate the odds ratio between groups differing by age.

**Inference:** The estimated odds ratio for groups differing by 1 year of age is 2.58, indicating that the odds of giving birth to an SGA infant are 158% higher for women in the younger group. (p<.015). A 95% CI indicates this finding would not be unusual if the true ratio were between 2.48 and 2.69.

1. **Table 2:** Modeled and observed probability of giving birth to an SGA infant in 20-year-old mothers using different methods of regression.

|  |  |
| --- | --- |
| **Method** | **Proportion** |
| Linear Regression | .158 |
| Poisson Regression | .1578 |
| Logistic Regression | .18775 |
| Observed value | .05128 |

**Methods:** Models were fitted as described in problem 6, and then probability estimates were made using the age of 20 as the predictor.

**Inference:** There was high agreement between the linear and Poisson regressions, and the estimate obtained by logistic regression differed by .03. The actual observed proportion of SGA births in this sample was 2 out of 39, resulting in a proportion of .051, a full .1 difference from our estimates.

1. **A. Methods:** The proportion of SGA infants born stratified by mother’s age was calculated and plotted using Stata software ver. 13.1.

**Inference:** The highest proportions were found at the lower end of the age spectrum, at ages 16 and 18, then ranged between .05 and .2 for mothers aged 19-35. No SGA infants were born to mothers older than 35 or under 15, but this may be due more to smaller sample size at these ages than a true difference in likelihood.

**Plot 1:** Proportion of SGA infant born stratified by age of mother.



**b. Methods:** Methods for statistical analysis were as described in problem 6. The graphic was produced with Stata version 13.1.

**Inference:** You can see from Plot 2 that the predicted probabilities of an SGA infant based on mother’s age are roughly in agreement, with a more marked departure at the extremes of the predictor values. All three models reflect what appears to be a genuine downward trend, though the data itself does not appear to fit any of the models particularly well.

**Plot 2:** Comparison of predicted probabilities of giving birth to an SGA infant based on age using different statistical models with observed proportions.



1. A. **Methods:** Age was log-transformed with the natural log, and then a logistic regression was performed using log age as the predictor and SGA as the response variable.

**Inference:** Our model estimates an odds ratio of 1.32 for groups differing by 1 year of age (p<.016). A 95% CI indicates this would not be unusual if the true odds ratio lay between 1.10 and 2.19.

b. There is no justification for examining age on a non-linear scale: the range of ages for fertile women should never exceed 50 years, and the significance of a difference in ages holds similar significance across age groups (that is, the difference between a 14 and a 15 year old is similar to the difference between a 30 and 31 year old).