**Biost 518: Applied Biostatistics II**

**Biost 515: Biostatistics II**

Emerson, Winter 2015

**Homework #3**

January 23, 2015

This homework considers pregnancy outcomes in an observational study of women attending a prenatal clinic in South Africa. Questions in this homework focus most closely on association with delivery of babies that are small for gestational age (SGA). The data can be found on the class web page (follow the link to Datasets) in the file labeled pregout.txt (you will not need any of the longitudinal measurements in the file preglong.txt). Documentation is in the file pregnancy.pdf.

1. Provide suitable descriptive statistics relevant to this analysis.

Methods: Descriptive statistics are presented for study participants by whether or not the baby was born small for gestational age in Table 1 below. For the continuous variables of mother’s age, number of prior deliveries, birth weight and gestational age, the mean, standard deviation and minimum and maximum values are presented. For the binary descriptive variables smoking and baby boy percentages are presented. Participants with missing data for certain variables are not included.

Results: All participants (n=755) were characterized as being small for gestational age or not. Several participants were missing values for certain variables: 6 for height, 5 for gestational age, 4 for smoking and baby boy status, those values are missing in the table 1 below. Of 755 participants 105 (13.91%) had a baby small for gestational age. Table 1 below shows a slightly lower mean age (23.85 versus 24.94) and number of prior deliveries (.90 versus 1.13) for mothers of babies small for gestational age. In addition, there appears to be a trend of more mothers with small babies to be smokers (43.3% versus 28.7%), and have a smaller percentage of boy babies (43.3% versus 52.4%). The babies in the small group also weighed less and were younger that the larger babies--the mean birth weight and age for the small babies was 2231.11 grams and 37.92 weeks respectively, compared to 3246.21 grams and 39.38 weeks for the babies not considered small for gestational age.

|  |  |  |
| --- | --- | --- |
|   | **Baby Small for Gestational Age** |  |
|   | **Yes**  | **No** | **Total** |
|   | N=105 | N=650 | N=755 |
| **\*Mother Height (cm)** | 154.56 (5.87, 142-172) | 157.01 (6.53, 106-176) | 156.68 (6.50, 106-176) |
| **Mothers Age (years)** | 23.85 (4.90, 16-35) | 24.94 (5.45, 14-43) | 24.79 (5.39, 14-43) |
| **\*Number prior deliveries** | .90 (1.11, 0-6) | 1.13 (1.23, 0-6) | 26.2 (4.7, 14-58 |
| **Smoker (%)** | 43.3% | 28.7% | 30.8% |
| **\*Birthweight (gms)** | 2231.11(411.60, 1035-3780) | 3246.21 (402.13, 2510-4730) | 3105.63 (534.46, 1035-4730) |
| **Boy baby (%)** | 42.3% | 52.4% | 51.0% |
| **\*Gestational age (weeks)** | 37.92(2.20, 30-42) | 39.38 (1.24, 38-44) | 39.18 (1.50, 30-44) |
| *\*Descriptive statistics include mean (SD, range)* |

1. Perform a statistical regression analysis evaluating an association between the odds of delivery of infants who were small for gestational age (SGA) and maternal smoking behavior. (Only give a formal report of the inference where asked to.)
	1. Give full inference regarding the association between SGA and maternal smoking.

Methods: To evaluate the association between the odds of delivery of infants who were small for gestational age (SGA) and maternal smoking behavior, we performed a logistic regression analysis, with SGA as the outcome and maternal smoking behavior as the predictor. Participants with missing values (n=4) for smoking status were not included in this analysis.

Inference: From logistic regression analysis, we estimate that the odds of having a baby small for gestational age was 1.890 times higher for smokers than nonsmokers and this estimate is statistically significant (p= 0.003). A 95% confidence interval suggests that this observation is not unusual if the odds of having a baby small for gestational age is anywhere from 1.237 to 2.888 times higher for smokers.

* 1. Use the regression model parameter estimates to provide estimates of both the odds and the probability of delivering a SGA infant separately for smokers and nonsmokers. How do these estimates compare with simple descriptive statistics as you might have reported in problem1. Explain any differences or similarities.

Because this is a saturated logistic model, the odds of delivering a SGA infant for smokers must be equal to the odds ratio calculated in the logistic regression analysis **1.890.**

= oddssmokers = psmokers/1-psmokers

oddsnonsmoker pnonsmokers/1-pnonsmokers

And the odds for non-smokers is the reciprocal of the odds for smokers = 1/1.89= **0.529**.

Using linear regression analysis we could find the probabilities of delivering an SGA infant for each group. The intercept of this model corresponds to the proportion of SGA infants in the nonsmoker group: **0.113** and the intercept plus the slope corresponds to the added proportion of SGA infants for smokers: 0.1134615 + 1\*(.0813437) = **0.195.**

In terms of similarities to the descriptive analysis in problem 1, if I had dichotomized the group into smokers versus nonsmokers and looked at the proportion of small babies in each group, these values should be equal to the probabilities of SGA from the linear regression parameter estimates.

* 1. There were actually four regression analyses that could have been used to answer this question. I am betting that all students would have fit a regression model with SGA as response and the indicator of maternal smoking as the predictor. Presuming that you did indeed fit that model, explain the similarities and differences between the estimates and inference you would have obtained for the following three additional models (You do not need to run these analyses, if you can tell me how they differ without doing so. It is of course okay to run the analyses if it will help you recognize the more general principles.):
		1. You create an indicator NONSMOKER that the mother was a nonsmoker, and you fit a logistic regression model of response SGA on predictor NONSMOKER.

The parameters for this model are the reciprocals of the regression model we performed in 2a and the p-value is the same.

From logistic regression analysis, we estimate that the odds of having a baby small for gestational age was 0.529 times for nonsmokers than smokers and this estimate is statistically significant (p= 0.003). A 95% confidence interval suggests that this observation is not unusual if the odds of having a baby small for gestational age is anywhere from 0.346 to 0.808 for nonsmokers than smokers.

* + 1. You create an indicator NOTSGA that the infant was not small for gestational age, and you fit a logistic regression model of response NOTSGA on predictor SMOKER.

The parameters for this model are also the reciprocals of the regression model we performed in 2a and the p-value is the same.

From logistic regression analysis, we estimate that the odds of not having a baby small for gestational age was 0.529 times for smokers versus nonsmokers and this estimate is statistically significant (p= 0.003). A 95% confidence interval suggests that this observation is not unusual if the odds of not having a baby small for gestational age is anywhere from 0.346 to 0.808 for smokers than nonsmokers.

* + 1. You fit a regression model of response NOTSGA on predictor NONSMOKER.

The parameters for this model are the same as those for the regression model we performed in 2a and the p-value is the same.

From logistic regression analysis, we estimate that the odds of not having a baby small for gestational age was 1.890 times higher for nonsmokers versus smokers and this estimate is statistically significant (p= 0.003). A 95% confidence interval suggests that this observation is not unusual if the odds of having a baby small for gestational age is anywhere from 1.237 to 2.888 times higher for smokers.

1. Repeat problem 2, except consider a statistical regression analysis evaluating an association between the odds of delivery of infants who were small for gestational age (SGA) and maternal smoking behavior by evaluating the difference in probabilities for SGA across smoking groups.

Methods: To evaluate the association between delivery of infants who were small for gestational age (SGA) and maternal smoking behavior, we evaluated the difference in probabilities for SGA for smoking participants (n=520) and nonsmoking participants (N=231). Using a linear regression model assuming unequal variance between groups we calculate the difference in probabilities of SGA by smoking group and a 95% confidence interval for this risk difference. Participants with missing values (n=4) for smoking status were not included in this analysis.

Inference: From linear regression analysis using the Huber-White sandwich estimator, we estimate that the probability of SGA infants for non-smoking mothers is 0.113 and 0.195 for smokers. The risk difference is 0.081 and the 95% confidence interval is consistent with a true difference in probabilities between 0.028 and 0.135. This difference was statistically significant, p=0.003.

1. Repeat problem 2, except consider a statistical regression analysis evaluating an association between the odds of delivery of infants who were small for gestational age (SGA) and maternal smoking behavior by evaluating the ratio of probabilities for SGA across smoking groups.

Methods: To evaluate the association between delivery of infants who were small for gestational age (SGA) and maternal smoking behavior, we evaluated the ratio of probabilities for SGA across smoking groups using a Poisson regression model assuming unequal variance between groups. Statistical inference on the ratio of the probabilities of SGA was based on the Wald statistic with two-sided p-value and 95% confidence interval. Participants with missing values (n=4) for smoking status were not included in this analysis.

Inference: From Poisson regression analysis, we estimate that the proportion of SGA for smoking mothers was 0.195 and 0.113 for non-smoking mothers. The observed ratio of SGA was 1.717 times higher for smokers was statistically significant p= 0.003. The 95% confidence interval is consistent with a true difference in ratios between 1.203 and 2.451 times higher for smoking versus non-smoking mothers.

1. How do the analyses performed in problems 2-4 compare to that that would be obtained in a simple two sample comparison of SGA by smoking status (i.e., using methods covered in Biost 517/514.) Explicitly mention where they would be similar or different?

Comparing the analyses performed above in 2-4 to that obtained using simple two sample comparisons of SGA by smoking status, using a two-sample t-test assuming unequal variance we could calculate the mean number of SGA across smoking groups, which is equal to the proportions of SGA calculated by linear and Poisson regression analyses above. The t-test statistic and z-test statistic in the linear regression analyses evaluating the association in each of the regression analyses is significant and the p-value is 0.003 for both the one-sided t-test analysis and z-test (though the test statistics are different).

Using a chi-squared test we could calculate the risk difference calculated in the linear regression analyses. The point estimates for the risk difference is the same (.081), as are the confidence intervals for the chi-squared test and linear regression using robust standard error estimates: 0.023 to 0.139.

The chi-squared analysis also provides a point estimate the same as logistic regression analyses 1.890, though the 95% confidence interval is slightly wider for the logistic regression analysis using robust standard errors than the chi-squared test 1.240 to 2.882 and 1.237 to 2.888 respectively.

The chi-squared analysis provides a point estimate for the risk ratio 1.717 equal to that obtained with Poisson regression and the 95% confidence intervals are equal: 1.203 to 2.450.

1. Perform a regression analysis of the distribution of the prevalence of SGA infants across groups defined by the continuous measure of maternal age. In all cases we want formal inference. (Note: In problem 7, I am asking you to plot the estimated probabilities of SGA infants from each of these regression models. Hence, you will want to make sure you estimate those fitted values following each regression.)
	1. Evaluate associations using risk difference (RD: difference in probabilities).

Methods: The association between SGA and a continuous measure of maternal age was analyzed using a linear regression analysis with robust standard error estimates. The two- sided p-value and 95% confidence interval were computed using the approximate normal distribution for the linear regression model. There were no participants excluded from this analysis, since all participants (n=755) had age and SGA reported.

 Inference: From linear regression analysis using the Huber-White sandwich estimator, we estimate that the probability of having an SGA infants is 0.4% lower for each additional year of age. The 95% confidence interval is consistent with a true probability difference of 0.9% lower to 0.01% lower for each additional year of age. The p-value for this analysis is 0.036 allows us to reject the null hypothesis that there is no association between SGA and maternal age at alpha 0.05.

* 1. Evaluate associations between risk ratio (RR: ratios of probabilities).

Methods: The association between SGA and a continuous measure of maternal age was analyzed using a Poisson regression model with robust standard error estimates. The two-sided p-value and 95% confidence interval were computed using the approximate Poisson distribution. There were no participants excluded from this analysis, since all participants (n=755) had age and SGA reported.

Inference: From Poisson regression analysis using the Huber-White sandwich estimator, we estimate that the risk ratio is 0.966 for mothers increasing in age. The 95% confidence interval is consistent with a true ratio between 0.934 and .999 for increasing age on a continuous 1 year scale. The p-value for this analysis is 0.046 which allows us to reject the null hypothesis that there is no association between SGA and maternal age at alpha 0.05.

* 1. Evaluate associations using odds ratio (OR: ratios of odds)

Methods: The association between SGA and a continuous measure of maternal age was analyzed using a logistic regression model with robust standard error estimates. The two-sided p-value and 95% confidence interval were computed using the approximate normal distribution. There were no participants excluded from this analysis, since all participants (n=755) had age and SGA reported.

Inference: From logistic regression analysis using the Huber-White sandwich estimator, we estimate that the odds ratio of SGA for is .961 for mothers increasing in age. The 95% confidence interval is consistent with a true ratio between 0.923 and 1.001 for increasing age on a continuous 1 year scale. The p-value for this analysis is 0.054, which does not allow us to reject the null hypothesis that there is no association between SGA and maternal age at alpha 0.05.

* 1. Using the regression parameter estimates from each of these regressions, provide an estimate of the probability that a 20 year old mother would have a SGA infant. Explain any similarities or differences these estimates might have when compared to the sample proportion of SGA infants among 20 year olds.

Using regression parameter estimates from each of the analyses above we can estimate the probabilities that a 20 year old mother would have a SGA infant. The fitted values for each of these regressions are included below:

Linear regression: .2509966 + (20\*-.0045152) = **.1606926**

Poisson regression: exp (-1.135976+(20\*-.0344235))= **.16130698**

Logistic regression: exp(-.8531571+(20\*-.0397786))= .19229414, .19229414/(1+.19229414)= **.16128079**

When comparing these estimates to the sample proportion of SGA infants among 20 year olds, we should expect to get a similar proportion in the sample as compared to the point estimates.

1. Produce a plot of the estimated probability of an SGA infant by age as derived by each of the following methods. Comment on the similarity and difference among the various fitted values form the various analyses performed in problem 6. (Note that Stata allows you to specify multiple Y variables for a single X variable: scatter y1 y2 y3 y4 age)
	1. Sample proportions within each unique age: This can be obtained in Stata using the command egen *varname*= mean(sga), by(age).

**Figure 1: Proportion of SGA Infants by Maternal Age**



* 1. Estimated probabilities for each age in the data as derived from each of the regression analyses. In Stata, this can be obtained using the simple “post-estimation” command: predict *varname.* (But use a different variable name for each fitted value.)
		1. After performing a linear regression, the default action of the “predict” function is to create a variable that contains the estimated “linear predictor”, which corresponds to the regression based estimate of the mean. With a binary response variable, the mean response is the proportion.

**Figure 2: Proportion of SGA Infants Predicted by Maternal Age by Linear Regression Analysis**



* + 1. After performing a Poisson regression, the default action of the “predict” function is to create a variable that contains the exponentiated estimated “linear predictor”, which corresponds to the regression based estimate of the mean. With a binary response variable, the mean response is the proportion. (The linear predictor in Poisson regression corresponds to the log “rate”, because Poisson regression uses a log link function.

**Figure 3: Proportion of SGA Infants Predicted by Maternal Age by Poisson Regression Analysis**



* + 1. In logistic regression, the estimated “linear predictor” corresponds to the log odds. Exponentiating that would correspond to the odds. By default, Stata figures that you would really rather have the estimated probability, which is computed as prob = odds / (1 + odds). So, after performing a logistic regression, the default action of the “predict” function is to create a variable that contains the the regression based estimate of the mean.

**Figure 4: Proportion of SGA Infants Predicted by Maternal Age by Logistic Regression Analysis**



1. Perform a logistic regression analyses of the distribution of the prevalence of SGA infants across groups defined by the logarithmically transformed maternal age.
	1. Provide formal inference for associations using odds ratio (OR: ratios of odds) and log transformed age.

Methods: The association between SGA and a continuous measure of log transformed maternal age was analyzed using a logistic regression model with robust standard error estimates. The two-sided p-value and 95% confidence interval were computed using the approximate normal distribution. There were no participants excluded from this analysis, since all participants (n=755) had age and SGA reported.

 Inference: From logistic regression analysis using the Huber-White sandwich estimator, we estimate that the odds ratio of SGA for is 0.385 for mothers increasing in log age. The 95% confidence interval is consistent with a true ratio between 0.147 and 1.010 for increasing log transformed age on a continuous 1 year scale. The p-value for this analysis is 0.052 which does not allow us to reject the null hypothesis that there is no association between SGA and maternal age at alpha 0.05.

* 1. Why might it be reasonable or silly to have performed such an analysis rather than the analysis in problem 6c?

It would be reasonable to perform this analysis rather than the analysis in 6c if we expected large outliers in age extremely young or extremely old mothers. By log transforming this predictor we could potentially gain precision in our predicted model.

However, we know from our descriptive analysis in #1 that age does not appear to be strongly skewed, so log transforming age may not be helpful. It also makes the model more difficult to interpret since age is easier to understand “age” than “log age.” Finally, the model presented in 6c was not statistically significant at alpha 0.05, which indicates we cannot reject the null hypothesis that there is no relationship between SGA and age, so it is difficult to understand why we would transform our predictor to add to a non-significant model (we would not want to perform this analysis after performing the analysis in 6c, we should decide a priori which to perform).