**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.

**Answers: Methods:** Subjects are divided into two groups based on whether they are smoker or non-smoker. Descriptive statistics are presented in the table below of each group as well as the entire sample population. For height, age, number of prior deliveries, infant birthweight and gestational age at delivery the mean, standard deviation and the data range are included. For binary variables (small for gestational and sex of the infants), the data is presented in percentages.

**Results:** This set of data includes a total of 755 subjects, but there are 4 subjects are missing data on their smoking status. These subjects are omitted in the data analysis. There are 6 subjects who are missing data on their height; 4 subjects who are missing data on their infant birthweight; 5 subjects who are missing data on gestational age at delivery. But they are not of the main interest of this study, so these subjects are still included.

The rest of the subjects (n=751) are divided into two groups based on their smoking status. The descriptive statistics within two groups are demonstrated in the table below. There are no obvious differences across groups in height, sex, age, number of prior deliveries and gestational age at delivery. However, there is higher percentage of subjects with SGA infants (19.5%) in the smoker group compared to the non-smoker group (11.3%). Also the mean infant birthweight is slightly lower in the smoker group (2972g) compared to the non-smoker group (3164g).

Table1. Descriptive statistics for the pregnancy dataset

|  |  |  |  |
| --- | --- | --- | --- |
|  | Smoker (n= 231) | Non-smoker (n=520) | Total (n=751/755) |
| Height (cm) | 157 (7.19; 106-176; n=230/231) | 157 (6.16;127-175; n=515/520) | 157 (6.49; 106-176; 745/751) |
| Age (years) | 25(5;15-42; n=231/231) | 25 (5; 21-43; n=520/520) | 25 (5 ; 14-43; n=751/751) |
| Small for gestational | 19.5% (n=231/231) | 11.3% (n=520/520) | 13.8%(n=751/751) |
| Number of prior deliveries | 1.19(1.27; 0-6; n=231/231) | 1.05(1.19; 0-6; n=520/520) | 1.10(1.21; 0-6; n=751/751) |
| Infant birthweight (g) | 2972(12.4;1410-4550; n=231/231) | 3164 (534;1035-4730; n=520/520) | 3106(534; 1035-4730; n=751/751) |
| Male | 48.1% (n=231/231) | 52.3% (n=520/520) | 51.0% (n=751/751) |
| Gestational age at delivery (weeks) | 39.0 (1.36; 33-43; n=230/231) | 39.3(1.55; 30-44; n=520/520) | 39.2(1.50; 30-44; n=750/751) |

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.

**Answers: Methods:** Subjects missing data for their smoking status were removed from the dataset. A logistic regression model was fit to describe the odds of delivery of infants who were small for gestational age (SGA) as the response of the maternal smoking behavior. Point estimates of the association were based on the slope parameter from the logistic regression analysis, and 95% confidence intervals and a two sided p value were computed using the Wald statistics (the regression slope estimate and its standard error).

**Results:** 4 of the 755 subjects had missing data on their smoking status, and those 4 subjects were omitted from this analysis. From logistic regression analysis, we estimate that the odds of delivery of infants who were small for gestational age is 89.1 % higher in the smoking group compared to the non-smoking group. With 95% confidence, it would be not unusual if the true odds of delivery of infants who were small for gestational age of the smoking group was anywhere from 23.8% to 188.8% higher than that of the non-smoking group. This observation is statistically significant at a 0.05 level of significance (P= 0.004). Therefore, we can reject the null hypothesis that the maternal smoking is not related to the delivery of infants who were small for gestational age. We conclude that the delivery of infants who were small for gestational age is associated with maternal smoking behavior.

* 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 problem 1. Explain any differences or similarities.

**Answers:** Log odds sga = -2.056+0.637 × smokerst

Prob= odds/(1+odds)

|  |  |  |
| --- | --- | --- |
|  | Smoker | Non-smoker |
| Odds | 0.242 | 0.128 |
| Probability | 0.195 | 0.113 |

The observed probabilities that were reported in Table 1 are 19.5% and 11.3% for smoker and non-smoker group, which agrees to our estimated values. That’s because this logistic regression model is a saturated model, the fitted odds with the regression model would equal to the sample odds.

* 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.

**Answers:**

= β0 + β1 \*smoker

= β0 + β1 \*(1 – nonsmoker)

= β0 + β1 – β1 \* nonsmoker

= 

The estimated slope from the logistic regression will be exactly equal in absolute value to the one obtained in problem 2a.

The estimated intercept from the logistic regression will be different from the one obtained in problem 2a.

The standard error for the logistic regression will be exactly equal to the one obtained in problem 2a.

The standard error for the intercept from the logistic regression will be different from the one obtained in problem 2a.

The two-sided p value from the logistic regression will be exactly equal to the one obtained in problem 2a.

The 95% confidence intervals for the slope from the logistic regression will be exactly equal in absolute value to the one obtained in problem 2a.

* + 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.

**Answers:**

If = β0 + β1 \*smoker





= –1 \*( β0 + β1 \*smoker)

= –β0 – β1 \*smoker

The estimated slope from the logistic regression will be exactly equal in absolute value to the one obtained in problem 2a.

The estimated intercept from the logistic regression will be exactly equal in absolute value to the one obtained in problem 2a.

The standard error for the logistic regression will be exactly equal to the one obtained in problem 2a.

The standard error for the intercept from the logistic regression will be exactly equal to the one obtained in problem 2a.

The two-sided p value from the logistic regression will be exactly equal to the one obtained in problem 2a.

The 95% confidence intervals for the slope from the logistic regression will be exactly equal in absolute value to the one obtained in problem 2a.

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

**Answers:**

If = β0 + β1 \*smoker





= –1 \* ( β0 + β1 \*smoker)

= –β0 – β1 \*smoker

= –β0 – β1 \* (1– nonsmoker)

= –β0 – β1 + β1 \* nonsmoker

= 

The estimated slope from the logistic regression will be exactly equal to the one obtained in problem 2a.

The estimated intercept from the logistic regression will be different from the one obtained in problem 2a.

The standard error for the logistic regression will be exactly equal to the one obtained in problem 2a.

The standard error for the intercept from the logistic regression will be different from the one obtained in problem 2a.

The two-sided p value from the logistic regression will be exactly equal to the one obtained in problem 2a.

The 95% confidence intervals for the slope from the logistic regression will be exactly equal to the one obtained in problem 2a.

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.
   1. **Answers: Methods:** Subjects missing data for their smoking status were removed from the dataset. A linear regression model using Huber-White estimates of the standard error was fit to describe the difference in probabilities for the delivery of infants who were small for gestational age (SGA) as the response of the maternal smoking behavior. Point estimates of the association were based on the slope parameter from the linear regression analysis, and 95% confidence intervals and a two sided p value were computed using the Wald statistics (the regression slope estimate and its standard error).

**Results:** 4 of the 755 subjects had missing data on their smoking status, and those 4 subjects were omitted from this analysis. From linear regression analysis using Huber-White estimates of the standard error, we estimate that the mean probability of delivery of infants who were small for gestational age is 8.13% higher in the smoking group compared to the non-smoking group. With 95% confidence, it would be not unusual if the probability of delivery of infants who were small for gestational age from the smoking group was anywhere from 2.32% to 13.9% higher than that of the non-smoking group. This observation is statistically significant at a 0.05 level of significance (P= 0.006). Therefore, we can reject the null hypothesis that the maternal smoking is not related to the delivery of infants who were small for gestational age. We conclude that the delivery of infants who were small for gestational age is associated with maternal smoking behavior.

* 1. **Answers:** Prob(sga) = 0.113+0.081 × smokerst

Odds = Prob/(1-Prob)

|  |  |  |
| --- | --- | --- |
|  | Smoker | Non-smoker |
| Probability | 0.194 | 0.113 |
| Odds | 0.241 | 0.127 |

The observed probabilities that were reported in Table 1 are 19.5% and 11.3% for smoker and non-smoker group, which agrees to our estimated values. That’s because this linear regression model is a saturated model, the fitted probabilities with the regression model would equal to the sample probabilities.

* 1. Three other analysis
     1. You create an indicator NONSMOKER that the mother was a nonsmoker, and you fit a linear regression model of response SGA on predictor NONSMOKER.

**Answers:**

If = β0 + β1 \*smoker

 = β0 + β1 \*smoker

= β0 + β1 \*(1 – nonsmoker)

= β0 + β1 – β1 \* nonsmoker

= 

The estimated slope from the linear regression will be exactly equal in absolute value to the one obtained in problem 3a.

The estimated intercept from the logistic regression will be different from the one obtained in problem 3a.

The standard error for the linear regression will be exactly equal to the one obtained in problem 3a.

The standard error for the intercept from the linear regression will be different from the one obtained in problem 3a.

The two-sided p value from the linear regression will be exactly equal to the one obtained in problem 3a.

The 95% confidence intervals for the slope from the linear regression will be exactly equal in absolute value to the one obtained in problem 3a.

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

**Answers:**

If = β0 + β1 \*smoker



= 1– ( β0 + β1 \*smoker)

= 1 – β0 – β1 \*smoker

The estimated slope from the linear regression will be exactly equal in absolute value to the one obtained in problem 3a.

The estimated intercept from the linear regression will be different from the one obtained in problem 3a.

The standard error for the linear regression will be exactly equal to the one obtained in problem 3a.

The standard error for the intercept from the linear regression will be exactly equal to the one obtained in problem 3a.

The two-sided p value from the linear regression will be exactly equal to the one obtained in problem 3a.

The 95% confidence intervals for the slope from the logistic regression will be exactly equal in absolute value to the one obtained in problem 3a.

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

**Answers:**

If = β0 + β1 \*smoker



= 1– ( β0 + β1 \*(1– nonsmoker)) = 1 – β0 – β1 + β1 \*nonsmoker

The estimated slope from the logistic regression will be exactly equal to the one obtained in problem 3a.

The estimated intercept from the logistic regression will be different from the one obtained in problem 3a.

The standard error for the logistic regression will be exactly equal to the one obtained in problem 3a.

The standard error for the intercept from the logistic regression will be different from the one obtained in problem 3a.

The two-sided p value from the logistic regression will be exactly equal to the one obtained in problem 3a.

The 95% confidence intervals for the slope from the logistic regression will be exactly equal to the one obtained in problem 3a.

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.
   1. **Answers: Methods:** Subjects missing data for their smoking status were removed from the dataset. A poisson regression model was fit to describe the ratio of delivery of infants who were small for gestational age (SGA) as the response of the maternal smoking behavior. Point estimates of the association were based on the slope parameter from the logistic regression analysis, and 95% confidence intervals and a two sided p value were computed using the Wald statistics (the regression slope estimate and its standard error).

**Results:** 4 of the 755 subjects had missing data on their smoking status, and those 4 subjects were omitted from this analysis. From poisson regression analysis, we estimate that the risk ratio of delivery of infants who were small for gestational age is 71.7% higher in the smoking group compared to the non-smoking group. With 95% confidence, it would be not unusual if the probability of delivery of infants who were small for gestational age from the smoking group was anywhere from 16.5% to 153.0% higher than that of the non-smoking group. This observation is statistically significant at a 0.05 level of significance (P= 0.006). Therefore, we can reject the null hypothesis that the maternal smoking is not related to the delivery of infants who were small for gestational age. We conclude that the delivery of infants who were small for gestational age is associated with maternal smoking behavior.

* 1. **Answers:** Log rate sga = -2.176+0.5405 × smokerst

|  |  |  |
| --- | --- | --- |
|  | Smoker | Non-smoker |
| Probability | 0.195 | 0.113 |
| Odds | 0.242 | 0.127 |

The observed probabilities that were reported in Table 1 are 19.5% and 11.3% for smoker and non-smoker group, which agrees to our estimated values. That’s because this poisson regression model is a saturated model, the fitted proportion with the regression model would equal to the sample proportion.

* 1. Three other analysis
     1. You create an indicator NONSMOKER that the mother was a nonsmoker, and you fit a poisson regression model of response SGA on predictor NONSMOKER.

**Answers:**

= β0 + β1 \*smoker

= β0 + β1 \*(1 – nonsmoker)

= β0 + β1 – β1 \* nonsmoker

= 

The estimated slope from the poisson regression will be exactly equal in absolute value to the one obtained in problem 4a.

The estimated intercept from the poisson regression will be different from the one obtained in problem 4a.

The standard error for the poisson regression will be exactly equal to the one obtained in problem 4a.

The standard error for the intercept from the poisson regression will be different from the one obtained in problem 4a.

The two-sided p value from the poisson regression will be exactly equal to the one obtained in problem 4a.

The 95% confidence intervals for the slope from the poisson regression will be exactly equal in absolute value to the one obtained in problem 4a.

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

**Answers:**

If = β0 + β1 \*smoker

= e β0 + β1 \* smoker

= 1-e β0 + β1 \* smoker

= log(1-e β0 + β1 \* smoker)

The estimated slope from the poisson regression will be different from the one obtained in problem 4a.

The estimated intercept from the poisson regression will be different from the one obtained in problem 4a.

The standard error for the poisson regression will be different from the one obtained in problem 4a.

The standard error for the intercept from the poisson regression will different the one obtained in problem 4a.

The two-sided p value from the poisson regression will be different from the one obtained in problem 4a.

The 95% confidence intervals for the slope from the poisson regression will be different from the one obtained in problem 4a.

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

**Answers:**

If = β0 + β1 \*smoker

From above, = log(1-e β0 + β1 \* smoker)

= log(1-e β0 + β1 \* (1-nonsmoker)) = 

The estimated slope from the poisson regression will be different from the one obtained in problem 4a.

The estimated intercept from the poisson regression will be different from the one obtained in problem 4a.

The standard error for the poisson regression will be different from the one obtained in problem 4a.

The standard error for the intercept from the poisson regression will be different from the one obtained in problem 4a.

The two-sided p value from the poisson regression will be exactly equal to the one obtained in problem 4a.

The 95% confidence intervals for the slope from the poisson regression will be different from the one obtained in problem 4a.

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?

**Answers:** The estimated slope from the linear regression will be exactly equal to risk difference obtained in the chi-squared test.

The estimated intercept from the linear regression will be exactly equal to the risk of the unexposed group from the chi square test.

The two-sided p value from the linear regression will be almost equal to the one obtained in the chi-squared test.

The 95% confidence intervals for the slope from the linear regression will be almost equal to the one obtained in the chi-squared test.

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).

**Answers: Methods:** A linear regression model using Huber-White estimates of the standard error was fit to describe the linear trend in in probabilities for the delivery of infants who were small for gestational age (SGA) as a function of age. Point estimates of the association were based on the slope parameter from the linear regression analysis, and 95% confidence intervals and a two sided p value were computed using the Wald statistics (the regression slope estimate and its standard error).

**Results:** From linear regression analysis using Huber-White estimates of the standard error, we estimate that for each year difference in age, the probability of delivery of SGA infants is 0.45% lower in the older group. With 95% confidence, it would be not unusual if a group that is one year older have the probability of delivery of SGA infants that was anywhere from 0.03% to 0.87 % lower than the younger group. This observation is statistically significant at a 0.05 level of significance (P= 0.036).

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

**Answers: Methods:** A poisson regression model was fit to describe the linear trend in in risk ratio for the delivery of infants who were small for gestational age (SGA) as a function of age. Point estimates of the association were based on the slope parameter from the linear regression analysis, and 95% confidence intervals and a two sided p value were computed using the Wald statistics (the regression slope estimate and its standard error).

**Results:** From poisson regression analysis, we estimate that for each year difference in age, the risk ratio of delivery of SGA infants decrease by 3.38% in the older group, though this estimate is not statistically significant (P = 0.07). With 95% confidence, it would be not unusual if a group that is one year older have the risk ratio of delivery of SGA infants that was anywhere from 7.00% lower to 0.33 % higher than the younger group.

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

**Answers: Methods:** A logistic regression model was fit to describe the linear trend in in odds for the delivery of infants who were small for gestational age (SGA) as a function of age. Point estimates of the association were based on the slope parameter from the logistic regression analysis, and 95% confidence intervals and a two sided p value were computed using the Wald statistics (the regression slope estimate and its standard error).

**Results:** From logistic regression analysis, we estimate that for each year difference in age, the odds of delivery of SGA infants is 3.90% lower in the older group, though this estimate is not statistically significant (P = 0.05). With 95% confidence, it would be not unusual if a group that is one year older have the odds of delivery of SGA infants that was anywhere from 7.72% lower to 0.08 % higher than the younger group.

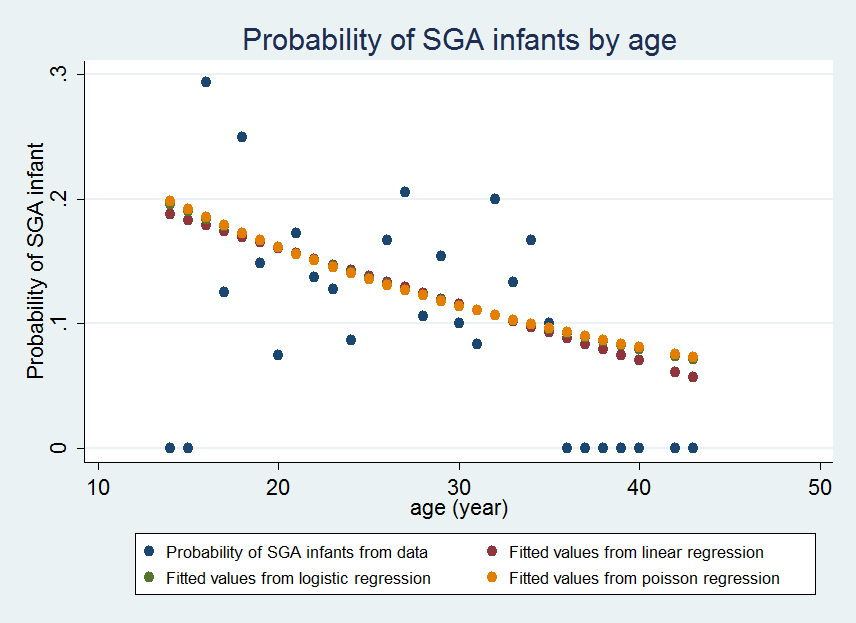
* 1. Using the regression parameter estimates from each of these regressions, provide an estimate of the probability that a 20 year old mother would have a SGA infant. Explain any similarities or differences these estimates might have when compared to the sample proportion of SGA infants among 20 year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

**Answers:**

|  |  |
| --- | --- |
|  | Probability of SGA infant at 20 years old |
| Linear regression  Prob(sga) = 0.2510-0.0045× age | 16.1% |
| Logistic regression  Log odds sga = -0.8532-0.0400 × age | 16.1% |
| Poisson regression  Log rate sga = -1.136-0.0344 × age | 16.1% |
| Dataset | 7.50% |

With all three regression models, the estimate of the probability of SGA infants is 16.1% for a 20 year old mother. However, the sample proportion of SGA infants among 20-year old (sample size = 40) is 7.50%. The estimated value is different from the sample proportion, since the regression models we used here are not saturated model.

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)

**Answers:**

The linear regression model estimates the average difference in probabilities per year. Its estimates are slightly different from the other two regression models. The estimates from the logistic regression and the poisson regression model are identical. The logistic regression model estimates the average odds ratio between two groups differing one year and the poission regression estimates the average probabilities ratio.

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.

**Answers: Methods:** A logistic regression model was fit to describe the linear trend in in odds for the delivery of infants who were small for gestational age (SGA) as a function of the logarithmically transformed age. Point estimates of the association were based on the slope parameter from the logistic regression analysis, and 95% confidence intervals and a two sided p value were computed using the Wald statistics (the regression slope estimate and its standard error).

**Results:** From logistic regression analysis, we estimate that for each 10% difference in age, the odds of delivery of SGA infants is 8.69% lower in the older group, though this estimate is not statistically significant (P = 0.058). With 95% confidence, it would be not unusual if a group that is one year older have the odds of delivery of SGA infants that was anywhere from 16.9% lower to 0.323 % higher than the younger group.

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

**Answers:** Age is usually count by years. Ten percent difference in age is difficult to understand and for the some parameter that people are interested, for example, the odds ratio at each age is hard to estimate in logarithmically transformed age.