**Biost 518: Applied Biostatistics II**

**Biost 515: Biostatistics II**

Emerson, Winter 2015

**Homework #3**

January 23, 2015

**Written problems:** To be submitted as a MS-Word compatible file to the class Catalyst dropbox by 9:30 am on Monday, February 2, 2014. See the instructions for peer grading of the homework that are posted on the web pages.

***Unless explicitly told otherwise in the statement of the problem, in all problems requesting “statistical analyses” (either descriptive or inferential), you should present both***

* ***Methods: A brief sentence or paragraph describing the statistical methods you used. This should be using wording suitable for a scientific journal, though it might be a little more detailed. A reader should be able to reproduce your analysis. DO NOT PROVIDE Stata OR R CODE.***
* ***Inference: A paragraph providing full statistical inference in answer to the question. Please see the supplementary document relating to “Reporting Associations” for details.***

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:* As stated in the documentation, the low birth weight, pre-term delivery and SGA (small for gestational age) are the three common complications of pregnancy. The pre-term delivery and SGA are in the pathway of causing low birth weight. In this analysis we mainly focus on SGA. Therefore the outcome of this study is SGA.**

**Variables other than birth weight, pre-term delivery or ID are considered possible predictors of interest.**

 **There are 755 observations of 9 variables in our data set.**

**The variables in the following table are arranged according to their effects on SGA scientifically:**

***Table 1.1*** Overall descriptive statistics

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Name** | **Missing** | **Mean** | **SD** | **Max** | **Median**  | **Min**  |
| **Height**  | **6/755** | **156.68cm** | **6.50** | **176cm** | **156cm** | **106cm** |
| **Age** | **0/755** | **24.79 yrs** | **5.39** | **43yrs** | **24yrs** | **14yrs** |
| **SGA** | **0/755** | **13.9%** | **-** | **-** | **-** | **-** |
| **Parity**  | **0/755** | **1.10** | **1.21** | **6** | **1** | **0** |
| **Non-smoker**  | **4/755** | **69.2%** | **-** | **-** | **-** | **-** |
| **Girl**  | **4/755** | **49.0%** | **-** | **-** | **-** | **-** |

 ***Table 1.12*** Descriptive statistics divided by SGA

|  |  |  |
| --- | --- | --- |
|  | **SGA=0 mean(missing; sd; range) (n=650)** | **SGA=1 mean(missing; sd;range)(n=105)** |
| **Height** | **157cm (0; 6.54; (106,176))** | **154.6cm (6; 5.87; (142,172))** |
| **Age**  | **24.9yrs (0; 5.45; (14,43))** | **23.8yrs (0; 4.90; (16,35))** |
| **Parity**  | **1.13 (0;1.23;(0,6))** | **0.90 (0;1.11;(0,6))** |
| **Non-Smoker** | **71.3% (missing=3)** | **56.7% (missing=1)** |
| **Girl** | **47.6% (missing=3)** | **57.7% (missing=1)** |

***Table 1.13*** Descriptive statistics divided by maternal smoking

|  |  |  |
| --- | --- | --- |
|  | **Smoker** **mean(missing; sd; range) (n=231)** | **Non-smoker mean(missing;sd;range)(n=520)** |
| **Height** | **157cm (1; 7.19; (106,176))** | **157cm (5; 6.16; (127,175))** |
| **Age**  | **25.1yrs (0; 5.35; (15,42))** | **24.6yrs (0; 5.37; (14,43))** |
| **Parity**  | **1.19 (0;1.27;(0,6))** | **1.06(0;1.19;(0,6))** |
| **SGA** | **19.5% (missing=0)** | **11.3% (missing=0)** |
| **Girl** | **51.9% (missing=0)** | **47.7% (missing=0)** |

  **There are 4 missing value in maternal smoking.**

***Inference*:**

**From table 1.1 we can see that there are 6 missing values in height and 4 each in smoker and the sex of infant. Notice there are only 13.9% of the observations in our data set had infants that are small for gestational age, thus maybe risk difference would not be a good choice to compare SGA rates. Also, the ranges of height and age are both very wide.**

**From table 1.2, there are 650 infants that are normal for gestational age, 105 that are small for gestational age. It seems that observations that mothers who are taller, older with prior delivers, non-smokers and deliver boys are less likely to have infants that are small for gestational age. But this is only a very crude deduction.**

**There are some points worth mentioning:**

* **The smoking status differs between two groups by 14.6% in difference and 25.7% in ratio, which seem to be significant and worth looking into.**
* **The range difference in height is due to the small sample size of mothers with height less than 142cm, and the larger group is more likely to contain these extreme values.**
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:***

**These data come from a cohort study without any sample size fixed by design. We are interested in evaluation association between the odds of delivery of infants who were SGA and maternal smoking behavior.**

**We can compare the odds of infants who were SGA given the maternal smoking status, or we can compare the odds of maternal smoking status given the infants SGA condition. Because the odds ratios are mathematically the same, the slope of these two logistic regressions should be the same.**

 **In this analysis, the odds of the infants being small for gestational age are compared between mothers with smoking habit and mothers without smoking habit by a logistic regression. The point estimator, standard error, two-sided p-value and a 95% confidence interval were given by Wald based estimates.**

***Inference:***

**There were 231 smokers and 520 non-smokers in our study with 4 subjects missing in smoking status. The observations with missing value in smoking were excluded in the analysis. There were no missing values in SGA in our data set.**

**The odds of infants that are SGA with a smoking mother was 0.242, and the odds of infants that are SGA with a non-smoking mother was 0.128. Based on a 95% confidence interval, the observed odds ratio 0.529 comparing between the non-smoker group to the smoker group would not be unusual if the true odds ratio fell between 0.347 and 0.811. The two-sided p-value is 0.00322 (< 0.05), thus we reject with high confidence the null hypothesis that the odds of getting infants that are SGA is not associated with maternal smoking, in favor of the alternative hypothesis that the odds of getting infants that are SGA is associated with maternal smoking.**

* 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.

Table 2.b

|  |  |  |
| --- | --- | --- |
|  | **Smokers**  | **Non-smokers** |
| **Odds** | 0.242 | 0.128 |
| **Probability**  | 0.195 | 0.113 |

**The probabilities of the delivering a SGA infant for smokers and nonsmokers in this table agree with** **the** **probabilities of delivering a SGA infant for smokers and nonsmokers given in table 1.13.**

**In this regression model, we have two parameters and two groups of predictor of interest. Therefore this is a saturated model, so the estimated odds of event in each group agree with the sample odds. The relationship between odds and probability is a 1-1 mapping, so the probabilities agree.**

* 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.



* + 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.



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



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.

**a.**

***Methods:***

**These data come from a cohort study without any sample size fixed by design. We are interested in evaluation association between the probability of delivery of infants who were SGA and maternal smoking behavior.**

**In this analysis, the probability of the infants being small for gestational age are compared between mothers with smoking habit and mothers without smoking habit by a linear regression. The point estimator, standard error, two-sided p-value and a 95% confidence interval were given. I did use the robust SE because there was a relationship between mean and variance.**

***Inference:***

**There were 231 smokers and 520 non-smokers in our study with 4 subjects missing in smoking status. The observations with missing value in smoking were excluded in the analysis. There were no missing values in SGA in our data set.**

**The proportion of infants that are SGA with a smoking mother was 19.5%, and the proportion of infants that are SGA with a non-smoking mother was 11.3%. Based on a 95% confidence interval, the observed proportion difference 8.13% lower in the non-smoker group to the smoker group would not be unusual if the true difference fell between 13.92% lower in the non-smoker group and 2.34% lower in the non-smoker group. The two-sided p-value is 0.00286 (< 0.05), thus we reject with high confidence the null hypothesis that the proportion of getting infants that are SGA is not associated with maternal smoking, in favor of the alternative hypothesis that the proportion of getting infants that are SGA is associated with maternal smoking.**

**b.**

Table 3.b

|  |  |  |
| --- | --- | --- |
|  | **Smokers**  | **Non-smokers** |
| **Odds** | 0.242 | 0.128 |
| **Probability**  | 0.195 | 0.113 |

**The probabilities of the delivering a SGA infant for smokers and nonsmokers in this table agree with** **the** **probabilities of delivering a SGA infant for smokers and nonsmokers given in table 1.13.**

**In this regression model, we have two parameters and two groups of predictor of interest. Therefore this is a saturated model, so the estimated proportion of event in each group agrees with the sample proportion. The relationship between odds and probability is a 1-1 mapping, so the odds agree.**

**c.**

* + 1. 
		2. 
		3. 
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. ***Methods:***

**These data come from a cohort study without any sample size fixed by design. We are interested in evaluation association between the proportion of delivery of infants who were SGA and maternal smoking behavior.**

**In this analysis, the probability of the infants being small for gestational age are compared between mothers with smoking habit and mothers without smoking habit by a poisson regression. The point estimator, standard error, two-sided p-value and a 95% confidence interval were given. I did use the robust SE because there was a relationship between mean and variance.**

***Inference:***

**There were 231 smokers and 520 non-smokers in our study with 4 subjects missing in smoking status. The observations with missing value in smoking were excluded in the analysis. There were no missing values in SGA in our data set.**

**The probability of infants that are SGA with a smoking mother was 19.5%, and the probability of infants that are SGA with a non-smoking mother was 11.3%. Based on a 95% confidence interval, the observed probability ratio 0.582, indicating 41.8% relative decrease in the non-smoker group to the smoker group, would not be unusual if the true probability ratio fell between 0.408 (59.2% relative decrease in the non-smoking group) and 0.831 (16.9% relative decrease in the non-smoking group). The two-sided p-value is 0.00631 (< 0.05), thus we reject with high confidence the null hypothesis that the probability of getting infants that are SGA is not associated with maternal smoking, in favor of the alternative hypothesis that the probability of getting infants that are SGA is associated with maternal smoking.**

Table 3.b

|  |  |  |
| --- | --- | --- |
|  | **Smokers**  | **Non-smokers** |
| **Odds** | 0.242 | 0.128 |
| **Probability**  | 0.195 | 0.113 |

**The probabilities of the delivering a SGA infant for smokers and nonsmokers in this table agree with** **the** **probabilities of delivering a SGA infant for smokers and nonsmokers given in table 1.13.**

**In this regression model, we have two parameters and two groups of predictor of interest. Therefore this is a saturated model, so the estimated proportion of event in each group agrees with the sample proportion. The relationship between odds and probability is a 1-1 mapping, so the odds agree.**

* 1. 





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?

**Ｉwould use a Chi-square test to do the comparison. And I would use inference for proportions to get the point estimates and CIs.**

* **The 2\*2 contingency table is the same:**

|  |  |  |
| --- | --- | --- |
|  | **Smoker** | **Non-smoker** |
| **SGA=0** | **186** | **461** |
| **SGA-1** | **45** | **59** |

* **The Chi-square test is the square of the Z-test, which corresponds to the linear regression in problem 3. The two-sided p-value should be exactly the same in these two analyses. But due to different way of handling standard error, the confidence intervals may vary a little.**
* **The point estimates in problem 2-4 should be exactly the same with the estimates I would get from the inference for proportions from BIOST514. In problem 2-4, the estimates are based on normal approximation, thus if we use CIs that computed under normal approximation, we would get the same CIs. If we use standard error computed under the mean-variance relationship. The CIs would be different. Due to the same reasons, under large sample, the two-sided p-value should be very similar (even when we use robust SE).**
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:***

**These data come from a cohort study without any sample size fixed by design. We are interested in evaluation association between the proportion of delivery of infants who were SGA and maternal age.**

**In this analysis, the probabilities of the infants being small for gestational age are compared across different maternal age groups by a poisson regression. The estimators of the parameters, two-sided p-value and a 95% confidence interval were given. I did use the robust SE because there was a relationship between mean and variance.**

***Inference:***

**There were no missing values in age or SGA.**

**From the poisson regression analysis on SGA using robust standard error, we estimate that with every 1 year increase in mother’s age, the probability of infants being small for gestational age decrease by 0.45%. According to the 95% confidence interval, this observation would not be unusual if the true relationship were between 0.030% decrease and 0.87% increase in the probability of infants being small for gestational age with every 1 year increase in mother’s age. The two-sided p-value is 0.054 (>0.05), thus we failed to reject the null hypothesis that SGA infants is not associated with mother’s age in favor of a tendency for lower SGA rate with higher maternal age.**

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

***Methods:***

**These data come from a cohort study without any sample size fixed by design. We are interested in evaluation association between the proportion of delivery of infants who were SGA and maternal age.**

**In this analysis, the probabilities of the infants being small for gestational age are compared across different maternal age groups by a linear regression. The estimators of the parameters, two-sided p-value and a 95% confidence interval were given. I did use the robust SE because there was a relationship between mean and variance.**

***Inference:***

**There were no missing values in age or SGA.**

**From the linear regression analysis on SGA using robust standard error, we estimate that with every 1 year increase in mother’s age, the probability of infants being small for gestational age has a relative decrease of 3.4% (risk ratio=0.966). According to the 95% confidence interval, this observation would not be unusual if the true relationship were between 0.062% (risk ratio=0.99937) relative decrease and 6.6% (risk ratio=0.934) relative decrease in the probability of infants being small for gestational age with every 1 year increase in mother’s age. The two-sided p-value is 0.074 (>0.05), thus we failed to reject the null hypothesis that there is no linear trend in the probability of SGA infants across mother’s age groups in favor of a tendency for lower SGA rate with higher maternal age.**

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

***Methods:***

**These data come from a cohort study without any sample size fixed by design. We are interested in evaluation association between the proportion of delivery of infants who were SGA and maternal age.**

**In this analysis, the odds of the infants being small for gestational age are compared across different maternal age groups by a logistic regression. The estimators of the parameters, two-sided p-value and a 95% confidence interval were given. I did not use the robust SE although there was a relationship between mean and variance, for the robust SE would produce a SE very similar to classic SE.**

***Inference:***

**There were no missing values in age or SGA.**

**From the logistic regression analysis on SGA, we estimate that with every 1 year increase in mother’s age, the odds of infants being small for gestational age is 3.9% (odds ratio=0.961) lower. According to the 95% confidence interval, this observation would not be unusual if the true odds were between 0.00062% (odds ratio: 0.9999376) lower and 7.8% (odds ratio: 0.92) lower of infants being small for gestational age with every 1 year increase in mother’s age. The two-sided p-value is 0.055 (>0.05), thus we failed to reject the null hypothesis that SGA infants is not associated with mother’s age in favor of a tendency for lower SGA rate with higher maternal age.**

* 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.

|  |  |
| --- | --- |
|  | **Probability that a 20 yrs mother have SGA infant** |
| **Linear regression on RD** | **0.16069** |
| **Poisson regression on RR** | **0.16131** |
| **Logistic regression on OR** | **0.16128** |
| **Sample proportion** | **0.075** |

 **The three estimates are the same, and are relatively larger than sample proportion. The regression models are all fitting the expected value of Y (the proportion) with normal approximation; hence they have the similar predicted value (mean value) at a given age. The reason that the probability is greater than the sample proportion is that in these models, the number of grouping variables is greater than the number of parameters, thus they are not saturated models, which means they borrow information from other age groups resulting in a different probability with the sample.**

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).



* 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.



* + 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.



* + 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.



1. Perform a logistic regression analyses of the distribution of the prevalence of SGA infants across groups defined by the logarithmically transformed maternal age.

Provide formal inference for associations using odds ratio (OR: ratios of odds) and log transformed age.

a.

 ***Methods:***

**These data come from a cohort study without any sample size fixed by design. We are interested in evaluation association between the proportion of delivery of infants who were SGA and maternal age.**

**In this analysis, the odds of the infants being small for gestational age are compared across different maternal 1.1-fold log-transformed age groups by a logistic regression. The estimators of the parameters, two-sided p-value and a 95% confidence interval were given. I did not use the robust SE although there was a relationship between mean and variance, for the robust SE would produce a SE very similar to classic SE.**

***Inference:***

**There were no missing values in age or SGA.**

**From the logistic regression analysis on SGA, we estimate that with every 10% increase in mother’s age, the odds of infants being small for gestational age is 8.7% (odds ratio=0.913) lower. According to the 95% confidence interval, this observation would not be unusual if the true odds were between 0.17% (odds ratio: 0.83) lower and 0.25% (odds ratio: 1.0025) higher of infants being small for gestational age with every 1 year increase in mother’s age. The two-sided p-value is 0.058 (>0.05), thus we failed to reject the null hypothesis that SGA infants is not associated with mother’s age in favor of a tendency for lower SGA rate with higher maternal age.**

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

**Ｉthink it is not necessary to perform such an analysis.**

 **In problem 6c, we have a constant odds ratio between groups differing in 1 years of age.**

**But if we do log-transformation:**

****

**The odds ratio will be depending on x, which is not ideal for analysis. Also if we use such analysis, we will have to use a k-fold comparison between age groups, which is hard to understand and complicated.**