Biost 515 Homework 3

**Methods for every problem:** It is assumed I used these methods unless otherwise specified in the problem answer. I used a .05 level of significance for all inference. Data was provided on 755 subjects with 4 subjects missing data for smoking status. I have eliminated these 4 subjects from ALL analyses. So all of my analyses are based on the 751 subjects for which I have data, except where specifically specified in problem 1, methods section. It should also be noted that I coded in R for all of my answers, so they may not match exactly the same as the STATA output.

**1.** **Methods:** Data was provided for 755 subjects. Of those 755, 4 subjects were missing data on smoking status and were removed from all further analyses. Subjects missing data for other variables are excluded only from analysis involving that variable. We present descriptive statistics stratified on small for gestational age (SGA) status as well as the total sample where smoking status was available. For continuous variables (mother’s height, mother’s age at enrollment, number of prior deliveries, birthweight, and gestational age at delivery) we present the mean (standard deviation; minimum – maximum). For binary variables (mother’s smoking status, male) we present the percentages.

**Results:** Of the 751 subjects with data on smoking status, 6 were missing data for the mother’s height and 1 was missing data on the gestational age at delivery. The subject missing data on gestational age at delivery was also missing data on mother’s height. It should be noted that we cannot assess the impact these omissions might have on the generalizability of our results.

Of the 751 subjects with data on smoking status, 104 fetuses were classified as small for gestational age (SGA Yes) and 647 fetuses were classified as within normal boundaries for gestational age (SGA No). The following table provides descriptive statistics for these groups as well as for the sample as a whole. As you can see there are a couple of trends with regard to SGA status. For the group classified as being small for gestational age (SGA Yes), the mother was more likely to be a smoker, the baby was more likely to be female, and the average birthweight of the baby was lower than in the entire sample.

|  |
| --- |
| Small for Gestational Age Status |
|   | SGA Yes | SGA No | All Subjects |
| Sample Size | 104 | 647 | 751 |
| Mother's Height (cm.) | 154.6 (5.88; 142 - 172) | 157.0 (6.52; 106 - 176) | 156.7 (6.49; 106 - 176) |
| Mother's Age at Enrollment (yrs) | 23.9 (4.91; 16 - 35) | 24.9 (5.42; 14 - 43) | 24.8 (5.36; 14 - 43) |
| Number of prior Deliveries | 0.9 (1.11; 0 - 6) | 1.1 (1.22; 0 - 6) | 1.1 (1.21; 0 - 6) |
| Smoker (%) | 43% | 29% | 31% |
| Birthweight of baby (gm.) | 2231 (411.6; 1035 - 3780) | 3246 (402.1; 2510 - 4730) | 3106 (534.5; 1035 - 4730) |
| Male baby (%) | 42% | 52% | 51% |
| Gestational Age at Delivery (wks) | 37.9 (2.20; 30 - 42) | 39.4 (1.24; 38 - 44) | 39.2 (1.5; 30 - 44) |

\*\*Data for continuous variables presented as: mean (sd; min – max)

**2.** **a) Methods:** We compared the odds of delivery of infants who were small for gestational age (SGA Yes) between the 2 groups: mothers who smoke and mothers who do not smoke using logistic regression. Statistical inference was based on analysis of the odds ratio using the Wald statistic and the regression output. We provide a 95% confidence interval for the odds ratio and two sided p-value using the approximate normal distribution. We did use robust standard error estimates although since we are using logistic regression, it does not make much difference.

**Results:** The odds of SGA Yes among the 231 mothers who smoked was .242 while the odds of SGA Yes among the 520 mothers who did not smoke was .128. The observed odds ratio of 1.89 agrees exactly with the slope of our logistic regression model since we had a saturated model. Based on a 95% confidence interval, this observed odds ratio of 1.89 would not be surprising if the true population odds ratio was between 1.236 and 2.891. With a two-sided p-value = .003, we reject the null hypothesis that the odds of SGA Yes are equal between the mothers who smoke and the mothers who do not smoke. This data provides evidence that mothers who smoke have a higher odds of delivering an infant who was small for gestational age.

**b)** Provided below is a table which gives the logistic regression output along with the estimated odds and probability of delivering an infant who was small for gestational age between mothers who smoke and mothers who do not smoke based on the logistic regression output. As you can see, these figures do agree with our descriptive statistics in the sense that both the descriptive statistics and the estimated odds and probability of SGA Yes provide evidence that smoking is positively associated with SGA Yes. In the descriptive statistics, the probability of smoking among SGA Yes was .43 while the probability of smoking among SGA No was only .29. In the table below the odds and probabilities of SGA Yes were both higher for mothers who smoked. However, the actual probabilities in the table below and the descriptive statistics table above do not match. This is because the probabilities were computed by conditioning on different things. In the descriptive statistics table, the probabilities were computed by conditioning on SGA status, while the probabilities in the table below were computed by conditioning on smoking status. This will likely give up different probabilities, however the odds ratio will be equal.

|  |
| --- |
| Logistic Regression Parameters |
| Slope | 1.890 |
| Intercept | 0.128 |
|   |   |
| Odds of SGA Yes |
| Smokers | 0.242 |
| Non-Smokers | 0.128 |
|   |   |
| Probability of SGA Yes |
| Smokers | 0.195 |
| Non-Smokers | 0.113 |

**c) i)** By creating an indicator for non-smoker we essentially reverse the direction of our inference. Our new intercept is the odds of SGA Yes for the smokers while in our previous regression the intercept was the odds of SGA Yes for the non-smokers. Also, our new slope is the reciprocal of the slope from our earlier regression. Our confidence interval are also related. The lower bound for this new model is the reciprocal of the upper bound from the original model while the upper bound is the reciprocal of the original lower bound. The p-values and all inference remains the same.

**ii)** This model is exactly the same as the model in **i)** except that the new intercept is the reciprocal of the intercept of the original model. The p-values and all inference remains the same.

**iii)** This model is the same as the original model except the new intercept is the reciprocal of the intercept in model **i).** The p-values and all inference remains the same.

**3. a) Methods:** We compared the difference in means (probabilities) of delivery of infants who were small for gestational age (SGA Yes) between the 2 groups: mothers who smoke and mothers who do not smoke using linear regression. Statistical inference was based on analysis of the difference in means using the Wald statistic and the regression output. We provide a 95% confidence interval for the difference in means (probabilities) and two sided p-value using the approximate normal distribution. We did use robust standard error estimates. We then use our conclusions about the difference in means (probabilities) to draw conclusions on the association of odds of delivering an infant who was small for gestational age and smoking status.

**Results:** The mean, or probability, of SGA Yes among the 231 mothers who smoked was .195 while the mean, or probability, of SGA Yes among the 520 mothers who did not smoke was .113. The observed difference in probabilities (smoker – non-smoker) of .081 agrees exactly with the slope of our linear regression model since we had a saturated model. Based on a 95% confidence interval, this observed difference in probabilities (smoker – non-smoker) would not be surprising if the true population difference in probabilities (smoker – non-smoker) was between .023 and .139. With a two-sided p-value = .006, we reject the null hypothesis that the probabilities of SGA Yes are equal between the mothers who smoke and the mothers who do not smoke. This data provides evidence that mothers who smoke have a higher probability of delivering an infant who was small for gestational age. Since the odds are a monotone-increasing, 1 to 1 function of probability, our conclusion based on the difference in probabilities can be directly applied to the odds. Therefore, this evidence provides evidence that mothers who smoke have a higher odds of delivering an infant who was small for gestational age than mothers who do not smoke.

**b)** Provided below is a table which gives the linear regression output along with the estimated odds and probability of delivering an infant who was small for gestational age between mothers who smoke and mothers who do not smoke based on the linear regression output. As you can see, these figures do agree with our descriptive statistics in the sense that both the descriptive statistics and the estimated odds and probability of SGA Yes provide evidence that smoking is positively associated with SGA Yes. In the descriptive statistics, the probability of smoking among SGA Yes was .43 while the probability of smoking among SGA No was only .29. In the table below the odds and probabilities of SGA Yes are both higher for mothers who smoked. However, the actual probabilities in the table below and the descriptive statistics table above do not match. This is because the probabilities were computed by conditioning on different things. In the descriptive statistics table, the probabilities were computed by conditioning on SGA status, while the probabilities in the table below were computed by conditioning on smoking status. This will likely give us different probabilities, however the odds ratio will be equal.

|  |
| --- |
| Linear Regression Parameters |
| Slope | 0.081 |
| Intercept | 0.114 |
|   |   |
| Odds of SGA Yes |
| Smokers | 0.242 |
| Non-Smokers | 0.128 |
|   |   |
| Probability of SGA Yes |
| Smokers | 0.195 |
| Non-Smokers | 0.114 |

**c) i)** By creating an indicator for non-smoker we essentially reverse the direction of our inference. Our new intercept is the original slope plus the original intercept, and our new slope is the negative of our original slope. The confidence intervals are also related. The new lower bound is the negative of the original upper bound, and the new upper bound is the negative of the original lower bound. The p-values and all inference remains the same.

**ii)** This model is exactly the same at the model in **i)** except the new intercept is 1 – intercept from our original model. The p-values and all inference remains the same.

**iii)** This model is exactly the same as the original model except the new intercept is 1 – intercept from the model in **i)**. The p-values and all inference remains the same.

**4. a) Methods:** We compared the ratio of means (probabilities) of delivery of infants who were small for gestational age (SGA Yes) between the 2 groups: mothers who smoke and mothers who do not smoke using Poisson regression. Statistical inference was based on analysis of the ratio of means using the Wald statistic and the regression output. We provide a 95% confidence interval for the ratio of means (probabilities) and two sided p-value using the approximate normal distribution. We did use robust standard error estimates. We then use our conclusions about the ratio of means (probabilities) to draw conclusions on the association of odds of delivering an infant who was small for gestational age and smoking status.

**Results:** The mean, or probability, of SGA Yes among the 231 mothers who smoked was .195 while the mean, or probability, of SGA Yes among the 520 mothers who did not smoke was .113. The observed ratio of probabilities (smoker – non-smoker) of 1.717 agrees with the slope of our Poisson regression model since we had a saturated model (accounting for rounding differences). Based on a 95% confidence interval, this observed ratio of probabilities (smoker – non-smoker) would not be surprising if the true population ratio of probabilities (smoker – non-smoker) was between 1.202 and 2.453. With a two-sided p-value = .003, we reject the null hypothesis that the probabilities of SGA Yes are equal between the mothers who smoke and the mothers who do not smoke. This data provides evidence that mothers who smoke have a higher probability of delivering an infant who was small for gestational age. Since the odds are a monotone-increasing, 1 to 1 function of probability, our conclusion based on the difference in probabilities can be directly applied to the odds. Therefore, this evidence provides evidence that mothers who smoke have a higher odds of delivering an infant who was small for gestational age than mothers who do not smoke.

**b)** Provided below is a table which gives the Poisson regression output along with the estimated odds and probability of delivering an infant who was small for gestational age between mothers who smoke and mothers who do not smoke based on the Poisson regression output. As you can see, these figures do agree with our descriptive statistics in the sense that both the descriptive statistics and the estimated odds and probability of SGA Yes provide evidence that smoking is positively associated with SGA Yes. In the descriptive statistics, the probability of smoking among SGA Yes was .43 while the probability of smoking among SGA No was only .29. In the table below the odds and probabilities of SGA Yes are both higher for mothers who smoked. However, the actual probabilities in the table below and the descriptive statistics table above do not match. This is because the probabilities were computed by conditioning on different things. In the descriptive statistics table, the probabilities were computed by conditioning on SGA status, while the probabilities in the table below were computed by conditioning on smoking status. This will likely give us different probabilities, however the odds ratio will be equal.

|  |
| --- |
| Poisson Regression Parameters |
| Slope | 1.717 |
| Intercept | 0.114 |
|   |   |
| Odds of SGA Yes |
| Smokers | 0.242 |
| Non-Smokers | 0.128 |
|   |   |
| Probability of SGA Yes |
| Smokers | 0.195 |
| Non-Smokers | 0.114 |

**c) i)** By creating an indicator for non-smoker we essentially reverse the direction of our inference. Our new intercept is the original slope \* original intercept. The new slope is the reciprocal of the original slope. The confidence intervals are also related. The new lower bound is the reciprocal of the original upper bound, and the new upper bound is the reciprocal of the original lower bound. The p-values and all inference remains the same.

**ii)** Due to how the exponential function works, in this model, we get a new slope of .9082. The new intercept is 1 – original intercept. We also get a new confidence interval for our slope of (.8463, .9748). Interestingly, the p-value changes to p-value = .008, but the inference remains the same.

**iii)** In this model, our slope is the reciprocal of the slope in **ii)**, and our intercept is 1 – intercept in **i).** Our confidence interval is also related. The new lower bound is the reciprocal of the upper bound in **ii)**, and the new upper bound is the reciprocal of the lower bound in **ii)**. The p-values and all inference remains the same as in **ii)**.

**5.** In Biost 514 we would stratify our sample into two groups: mothers who smoke and mothers who do not smoke and compare the means (probabilities) of SGA in mothers who smoke to SGA in mothers who do not smoke. We would compare the difference means using a two-sample t-test without assuming equal variances. This analysis most closely resembles the analysis we did in question 3. In both the two sample t-test and question 3 we were able to obtain the exact same means (probabilities) of SGA in both groups: .195 for mothers who smoke and .113 in mothers who do not smoke. We also get similar confidence intervals for the difference in means (smokers – non-smokers) which was (.023, .139). Because all of our estimates were the same, we also get the same two-sided p-value of .006. All methods from 2 – 4 imply the same means (probabilities) of SGA and odds of SGA for each group defined by smoking status as does the two sample t-test. The inference conclusions are also all equivalent among methods 2 – 4 and the t-test. The p-values are slightly lower (.003 difference) in the methods used for questions 2 & 4 than for method 3 and the t-test. Overall, all of these methods provide similar results.

**6. a) Methods:** We compared the difference in means (probabilities), equivalently the risk difference (RD), of delivery of infants who were small for gestational age (SGA Yes) stratified by age of the mother at the time of study enrollment. Statistical inference was based on analysis of the difference in means using the Wald statistic and linear regression output. We provide a 95% confidence interval for the difference in means (probabilities), the risk difference, and two sided p-value using the approximate normal distribution. We did use robust standard error estimates.

**Results:** The slope of our linear regression of SGA Yes stratified on age of the mother at time of study enrollment was -.00427. Based on a 95% confidence interval, this observed risk difference of -.00427 in mothers who are 1 year older than another mother would not be surprising if the true population risk difference in mothers who are 1 year older than another mother was between -.00854 and .000003. Our two-sided p-value was .0502. Our findings are right on the verge of significance, but are not significant at the .05 level of significance. Therefore, we must conclude that age of the mother and probability of SGA Yes are not associated. However, I do think more research is needed to determine this association or lack thereof since our results are right on the border of significance.

**b) Methods:** We analyzed the ratio of means (probabilities), equivalently the risk ratio (RR), of delivery of infants who were small for gestational age (SGA Yes) stratified by age of the mother at the time of study enrollment. Statistical inference was based on analysis of the ratio of means using the Wald statistic and Poisson regression output. We provide a 95% confidence interval for the ratio of means (probabilities), the risk ratio, and two sided p-value using the approximate normal distribution. We did use robust standard error estimates.

**Results:** The slope of our Poisson regression of SGA Yes stratified on age of the mother at time of study enrollment was .968. Based on a 95% confidence interval, this observed risk ratio of .968 in mothers who are 1 year older than another mother would not be surprising if the true population risk ratio in mothers who are 1 year older than another mother was between .936 and 1.002. Our two-sided p-value was .0616. Our findings are not significant at the .05 level of significance. Therefore, we must conclude that age of the mother and probability of SGA Yes are not associated. However, I do think more research is needed to determine this association or lack thereof since our results are close to significant.

**c) Methods:** We analyzed the odds ratio of delivery of infants who were small for gestational age (SGA Yes) stratified by age of the mother at the time of study enrollment. Statistical inference was based on analysis of the odds ratio using the Wald statistic and Logistic regression output. We provide a 95% confidence interval for the odds ratio, and two sided p-value using the approximate normal distribution. We did use robust standard error estimates.

**Results:** The slope of our Logistic regression of SGA Yes stratified on age of the mother at time of study enrollment was .963. Based on a 95% confidence interval, this observed odds ratio of .963 in mothers who are 1 year older than another mother would not be surprising if the true population odds ratio in mothers who are 1 year older than another mother was between .926 and 1.002. Our two-sided p-value was .0611. Our findings are not significant at the .05 level of significance. Therefore, we must conclude that age of the mother and probability of SGA Yes are not associated. However, I do think more research is needed to determine this association or lack thereof since our results are close to significant.

**d)** Provided below is a table which gives the estimated probabilities of a 20 year old mother delivering an infant who was small for gestational age based on 3 different regression methods and the actual sample proportion of 20 year old mothers who did deliver an infant who was small for gestational age. As you can see, the estimates are all relatively close together, but far different from the actual sample proportion. This may be an indication that our linear model is mis-specified, i.e., we do not actually have a linear relationship between age of the mother and probability of delivering an SGA Yes infant. This difference between the actual sample proportion and the estimates can occur, because the linear regression model borrows information about SGA Yes from all ages and all age comparisons. Therefore, for any given age, the estimates can differ drastically from the sample, especially when there is no linear relationship.

|  |
| --- |
| Regression Estimates for the Probability of delivering an SGA Yes infant for a 20 Year Old Mother |
| Linear (RD) | Poisson (RR) | Logistic (OR) | Sample Proportion |
| 0.1589 | 0.1594 | 0.1895 | 0.0513 |

**7. a) Discussion of similarities/differences for parts a and b:** As you can see from the scatterplot of sample proportions of SGA Yes by age, it does not appear that SGA and age are associated. There may be a slight downward linear trend, but it is not clear. The linear regression line seems to split all of the points down the middle. The Poisson and Logistic regression lines look like they overestimate the sample proportions at the boundaries and underestimate the sample proportions in the middle.



**b)**







**8. a) Methods:** We analyzed the odds ratio of delivery of infants who were small for gestational age (SGA Yes) stratified by log(age) (base e) of the mother at the time of study enrollment. Statistical inference was based on analysis of the odds ratio using the Wald statistic and Logistic regression output. We provide a 95% confidence interval for the odds ratio, and two sided p-value using the approximate normal distribution. We did use robust standard error estimates.

**Results:** The slope of our Logistic regression of SGA Yes stratified on age of the mother at time of study enrollment was .406. Based on a 95% confidence interval, this observed odds ratio of .406 in mothers who are e (2.718) years older than another mother would not be surprising if the true population odds ratio in mothers who are e (2.718) years older than another mother was between .153 and 1.074. Our two-sided p-value was .0693. Our findings are not significant at the .05 level of significance. Therefore, we must conclude that age of the mother and probability of SGA Yes are not associated. However, I do think more research is needed to determine this association or lack thereof since our results are close to significant.

**b)** This is silly to have performed such an analysis rather than in problem 6c. In both cases we are measuring the same association. We are estimating how SGA Yes probability changes as we change the age. By log transforming age, we are only changing the scale by which age increases. This alters our slope, but does not alter our inference about the association between SGA Yes and age. Also, it is confusing to think about how the probability of SGA Yes changes with respect to an additional e (2.718) years. It is much easier to interpret the slope of the regression in the original scale of 1 additional year.