BIOST 518
Homework 3

**Total 125 out of 160 credits**

Credits for question1: 8 points

-2 subtracted as it would be more useful to present both stratified descriptive statistics by smoking behavior and SGA

1. **Methods**: Descriptive statistics are presented within groups as defined whether the baby is small for gestational age or not (SGA status). Descriptive statistics are also presented for the total population. Within each group defined by SGA status, values for continuous variables (which include maternal age and height, baby birthweight, parity, and gestational age) are presented as the mean. The standard deviation, minimum value and maximum value are also included. For binary variables (which include maternal smoking status and the child’s sex), the proportions are presented as percentages. In running the analysis, several patients out of the 755 study subjects were missing data for one or more variables. The number of patients missing data for a given variable are also provided in the table below.

**Results**: Descriptive statistics have been tabulated and presented in table 1 below. Several factors do not appear to be associated with SGA status including maternal age, maternal height and parity. The mean gestational ages among the two groups also do not appear to be different between the two groups. From the preliminary analysis, two factors appear associated to SGA status. The observed data indicate that babies that are small for their gestational age tend to be from mothers who smoked during pregnancy. Among the 104 babies who were small for their gestational age and had birth weight data, the mean birth weight was 2231.11 g while for the 650 babies who were not small for their gestational age and had birth weight data, the mean birth weight was 3246.21 g. Also, among babies who were diagnosed as small for their gestational age, 43.27% of the mothers smoked during pregnancy, while only 28.75% of mothers gave births to babies who were small for their age reported being smokers. Gestational age in this population appears uninformative as the mean gestational ages do not appear to be too different between babies born small for their gestational age and babies who are normal for their gestational age. Interestingly, a greater proportion of babies born small for their gestational age are females (57.69%) when compared to babies who were considered normal for their gestational age (47.60%).

**Table 1**

|  |  |
| --- | --- |
|  | Small for gestational age (SGA status) |
|  | **Yes****(N=105)** | **No****(N=650)** | **Total Population****(N=755)** |
| **Maternal Age (years)** | 23.85 (4.90; 16-35) | 24.94 (5.45; 14-43) | 24.79 (5.39; 14-43) |
| **Smokers**1 missing with SGA=Y3 missing with SGA=N | 43.27% | 28.75% | 30.76% |
| **Maternal Height (cm)**6 missing with SGA=Y | 154.56 (5.87; 142-172) | 157.00 (6.54; 106-176) | 156.68 (6.50; 106-176) |
| **Parity** | 0.89 (1.11; 0-6) | 1.13 (1.23; 0-6) | 1.10 (1.21; 0-6) |
| **Female babies**1 missing with SGA=Y3 missing with SGA=N | 57.69% | 47.60% | 49.00% |
| **Birthweight (g)**1 missing with SGA=Y3 missing with SGA=N | 2231.11 (411.60; 1035-3780) | 3246.21 (402.13; 2510-4730) | 3105.63 (534.46; 1035-4730) |
| **Gestational Age (wks)**1 missing with SGA=Y3 missing with SGA=N | 37.92 (2.20; 30-42) | 39.38 (1.24; 38-44) | 39.18 (1.50; 30-44) |

Credits for question2: 9+1+7=17 points

For 2a: -1 subtracted for no mentioning of using the maximum likelihood estimate of the slope from the regression model and we use asymptotically normal distribution for Wald statistics derived from the model

For 2b: -4 subtracted as you didn’t get the right answer and didn’t realize this is a saturated model. So the corresponding fitted values from the regression would agree exactly with the numbers in the descriptive statistics.

For 2c: -3 subtracted as (didn’t mention the intercept)

**Methods**: Logistic regression analysis was carried out in order to measure the odds of delivery of an infant who is small for their gestational age using maternal smoking behavior as the predictive variable. Since four of the mothers in the study did not have data for smoking status, those patients were excluded from the analysis. Two-sided P-values and 95% confidence intervals calculated using Wald-estimates are reported. It would be better to mention: like an estimated odds ratio was computed using the maximum likelihood estimate of the slope from the regression model and we use asymptotically normal distribution for Wald statistics derived from the model

* 1. **Inference**: From the logistic regression analysis performed, we estimate that the odds of delivering an infant who is small for their gestational age is 89.04 % higher for mothers who smoked during pregnancy compared to mothers who did not smoke during pregnancy. This estimate is statistically significant at a significance level of 0.05 (P=0.003). Based on 95% confidence intervals calculated, this observation is not unusual if the true increase in the odds of delivering an infant who is small for their gestational age is between 23.76% and 188.75% .
	2. **Methods**: Using the logistic regression analysis performed as described above, the intercept obtained is 0.1279826 and the slope is estimated to be 1.890378. This means that the log odds SGA = 0.1279826 + 1.890378\*(Maternal smoker status). The odds were then calculated by exponentiating the log odds when maternal smoking is equal to 0 (non-smoker) or 1 (smoker). The probability was calculated using the formula prob = odds/(1+odds).

	**Results**: Using the equation above, we estimate the odds of a non-smoking mother to have a baby born small for their gestational age to be 1.1365, with the probability being ~~53.19%.~~ In contrast, the odds of mothers having a baby small for their gestational age is 7.5260 and the probability is ~~88.27%.~~ From our descriptive statistics, we see that mothers who smoke during pregnancy do tend to have a higher probability of giving birth to a child who is SGA compared to their non-smoking counterparts. This is supported by the logistic regression analysis performed above.
		1. The slope of the analysis obtained would be the inverse of the slope obtained above. This would mean that the change in odds between groups would decrease, rather than increase since the analysis is now comparing SGA using smokers as the “control” and seeing how non-smokers compared to this group. The significance (P value) does not appear to change for the analysis.(didn’t mention the intercept)
		2. Similarly, to i. the slope of the analysis obtained using logistic regression would be the inverse of the slope obtained in the original analysis. This would indicate a decrease in the odds ratio since the response variable is not conditioned using babies who are SGA as the “control”, meaning that these babies are the point of comparison.(didn’t mention the intercept)
		3. Lastly, conditioning on both non-SGA and non-smokers yields the same change in the odds ratio as the original analysis.(didn’t mention the intercept)

Credits for question3: 9+4+10=23 points

For 3a: -1 subtracted (see comments)

For 3b: -1 subtracted

1. **Methods**: Linear regression analysis was carried out in order to measure the difference in probabilities of delivering a baby who is SGA using maternal smoking behavior as the predictive variable. Since four of the mothers in the study did not have data for smoking status, those patients were excluded from the analysis. Two-sided P-values and 95% confidence intervals calculated using Huber-White estimates are presented.(would be nice to mention as estimated difference in probabilities was computed using the least squared estimate of the slope from the regression model)

	1. **Inference**: From the linear regression analysis performed, we estimate that the difference in probabilities of giving birth to a child who is small for their gestational age between non-smoking and smoking mothers is 0.08134. This means that mothers who smoke have an 8.134% higher probability of having a child who is SGA. This is significant at a statistical significance of 0.05 (two-tailed P-value =0.006) and this observation would not be judged to be unusual if the true risk difference were between 0.02328 and 0.13941.
	2. **Methods**: Using the linear regression analysis performed as described above, the intercept is 0.1134615 and the slope is estimated to be 0.0813437. This means that the probability of SGA = 0.1134615 + 0.0813437\*(Maternal smoker status). The odds were then calculated using odds = probability/(1-probability).

	**Results**: Using the equations above, in non-smoking mothers, we estimate the probability of having an SGA baby to be 0.11346 (11.346%) and the odds to be 0.127982. In mothers who smoked during pregnancy, the probability of having an SGA baby is estimated to be 0.19480 (19.480%), with the odds being 0.241935. As with the logistic regression analysis performed above, the results agree with our findings from our descriptive statistics that more mothers who smoked during pregnancy tended to have babies who were SGA than mothers who did not smoke during pregnancy. (better mention this is a saturated regression model)

		1. The coefficient will be different because it now reflects the probability of having an SGA baby when the mother is a smoker. Additionally, the slope of the regression line will be the negative of the slope of the regression line in the initial analysis. The P-values obtained are still the same.
		2. The coefficient is again different as it now reflects the probability of NOT having an SGA baby when the mother is a non-smoker. The slope of the regression line is also the negative of the slope from the initial analysis as the probability of NOT having an SGA baby decreases with smoking mothers. The P-value is still the same.
		3. The coefficient is different because it reflects the probability of NOT having an SGA baby when the mother is a smoker. The slope, however, is the same as what was obtained from the initial analysis since the probability of NOT having an SGA baby does increase with non-smoking mothers. The P-value is the same.

Credits for question4: 9+4+7=20 points

For 4a: -1 subtracted (see comments)

For 4b: -1 subtracted

For 4c: -3 subtracted

1. **Methods**: Poisson regression analysis was performed in order to measure the ratio of the probabilities of delivering a baby that is small for their gestational age using maternal smoking behavior as the predictive variable. Since the interest is in the risk ratios, we report the incidence risk ratios between smoking and non-smoking mothers. Since four of the mothers in the study did not have data for smoking status, those patients were excluded from the analysis. Two-sided P-values and 95% confidence intervals calculated using Wald estimates are presented.(two sided 95% CI for true risk ratio were computed using the Huber-White sandwich estimate of the SE to account for the mean-variance relationship in these binary data)

	1. **Inference**: From the Poisson regression analysis, we estimate that the probability of having a baby born small for their gestational age is 71.69% higher in mothers who smoked during pregnancy compared to mothers who did not smoke during pregnancy. This observation is significant at a statistical level of 0.05 (P=0.003) and according to Wald estimates, would not be judged unusual if the true increase in rate ratios were between 20.28% and 145.07%.
	2. **Results**: Among nonsmokers, 59/520 babies were born SGA. This gives a probability of 0.1135 of having an SGA baby among nonsmoking mothers, with an odds of 0.1280. Among smoking mothers, however, there were 45/231 babies born SGA. This yields a probability of 0.1948 of having an SGA baby among smoking mothers. The odds are 0.2419 among this population. This is in agreement with the trend seen that babies who were born small for their gestational age had a higher proportion that were born to smoking mothers than babies who were not small for their gestational age.( would be nicer realizing this is a saturated poisson regression model)
	3. When analyzing by conditioning on NONSMOKING (i), the slope observed is the negative of the(log probability) slope from the initial analysis since the observation for the association will switch from being positive (increase in risk ratios) to negative (decrease in risk ratio)(no mentioning of the log probability intercept). Similarly, when conditioning for NOTSGA (ii), the (log probability)slopes will be different because the comparison will be different and conditioned using babies with SGA (NOTSGA==0). Finally, the analysis when conditioned to both NOTSGA and NONSMOKER(no mentioning of the log probability intercept) (iii) will also have different (log probability)slopes, though it will be similar to the analysis done in part ii having the same value but the positive rather than the negative(no mentioning of the log probability intercept).

Credits for question5: 5 points

-5 points subtracted(see comments)

1. For a test of means, such as the t-test(allowing for unequal variances?), the “means” reported by the software are the same as the probabilities expressed by linear regression and Poisson regression analyses. When looking at the chi squared tests, the risk difference, risk ratio and the odds ratio reported are the same values obtained as the slope of linear regression, incidence risk ratio of the Poisson regression and the slope of the logistic regression, respectively. This is likely because these analyses were carried out using a binary predictor. This may not hold true for predictors with continuous variables.Should presented in a more clear format and the chi square test for association(the score test from the regression) correspond roughly to the analysis in problem 2(logistic regression).

 Problem 3(linear regression)~t test allowing for unequal variances

 Problem 4(poisson regression)~2 sample test of probability ratios from likelihood theory

Credits for question6: 33 points

For 6a: -2

For 6b: -2

For 6c: -1

For 6d:-2

* 1. **Methods**: Linear regression analysis was performed to evaluate the association between the prevalence of SGA infants across groups defined by maternal age, using the risk difference. Two-sided P-values and 95% confidence intervals calculated using Huber-White estimates are presented..(would be nice to mention as estimated difference in probabilities was computed using the least squared estimate of the slope from the regression model)

	**Inference**: Using linear regression, we observe that the difference in the risk of having a baby that is SGA decreases by 0.4515 %for every incremental increase in maternal age. This observation is significant at a level of 0.05 (P=0.036) and based on the 95% confidence intervals calculated, would not be judged as an unusual observation if the true difference in risks of having a baby that is SGA decreased from 0.0286% to 0.8745%.
	2. **Methods**: Poisson regression analysis was performed to evaluate the association between the prevalence of SGA infants across groups defined by maternal age, using the risk ratios. Two-sided P-values and 95% confidence intervals calculated using Wald estimates are presented.(two sided 95% CI for true risk ratio were computed using the Huber-White sandwich estimate of the SE to account for the mean-variance relationship in these binary data)

	**Inference**: Using Poisson regression, we observe a decrease of ~~3.4423%~~ in risk ratios of having a baby born with SGA status for each incremental increase in maternal age. This observation is significant at a level of 0.05 (P=0.046). Based on the calculated 95% confidence intervals, this would not be judged unusual if the true decrease in risk ratios between non-smoking and smoking mothers were between ~~0.0608% and 6.824%~~.(numbers don’t agree with the keys , even the keys are reporting 5 year difference)
	3. **Methods**: Logistic regression analysis was performed to evaluate the association between the prevalence of SGA infants across groups defined by maternal age, using the odds ratios. Two-sided P-values and 95% confidence intervals calculated using Wald estimates are presented.It would be better to mention: like an estimated odds ratio was computed using the maximum likelihood estimate of the slope from the regression model and we use asymptotically normal distribution for Wald statistics derived from the model

	**Inference**: Using logistic regression, we observe that the odds of having a baby with SGA status decreases by 3.9% for each incremental increase in maternal age. This observation is not significant at a level of 0.05 (P=0.054) and is not judged to be unusual if the true decrease in the odds of having a baby that small for gestational age was between -.0764 % (really an increase in odds by 0.0764%) and 7.718%.
	4. **Methods**: After performing linear, Poisson, and logistic regression as described above, the software was used to predict probabilities of having a baby with SGA status when the maternal age is 20.

	**Results**: Using linear regression, Poisson regression, and logarithmic regression, the probabilities of having a baby with SGA status when the maternal age is 20 was found to be 0.160692, 0.161307, and 0.161281, respectively. The corresponding odds as obtained from each analysis are found to be 0.191458, 0.191598, and 0.191592, respectively. These values are very similar to each other, only varying slightly. For example, the probabilities calculated are all around 16% and the variation is likely to be within the standard errors of the estimates. Similarly, the odds calculated are around 0.191, and the slight variations are likely to be within the standard errors of these estimates.(didn’t address the question about why these estimates are different from the sample proportion )

Credits for question 7: 8 points

-2 points

1. Would be better included the sample proportions in the graph

The fitted values are similar in some ranges, from about 18 years of age to about 30 years of age. Outside of this range, the linear prediction varies noticeably from the other two predictions while the Poisson and logistic predictions retain similar predictions on the probability of having a baby with SGA status. This is likely because these two predictions rely on a log transform and are then exponentiated in order to obtain the predictive probabilities.

Credits for question 8: 8 +3=11points

For 8a: -2 points

For 8b: -2

* 1. **Methods**: Maternal age was first log transformed (base 2) for the purposes of this problem. Logistic regression analysis was then performed to evaluate the association between the prevalence of SGA infants across groups defined by the log transformed maternal age, using the odds ratios. Two-sided P-values and 95% confidence intervals calculated using Wald estimates are presented. It would be better to mention: like an estimated odds ratio was computed using the maximum likelihood estimate of the slope from the regression model and we use asymptotically normal distribution for Wald statistics derived from the model

	**Inference**: Based on the logistic regression carried out, we observe that the odds of having a baby with SGA status decreases by 48.3674% for each two-fold change in maternal age. This observation is not significant at a level of 0.05 (P=0.058) and is not judged to be unusual if the true decrease in the odds of having a baby that small for gestational age was between -2.3751 % (an increase in odds by 2.3751%) and 73.9592%.
	2. This analysis would seem silly because it is not likely that we look for the changes in the probability of having a child with SGA status as age doubles (or whatever base is used; eg, every tenfold change in age, etc). Only addressed the this would not be ideal in terms of clarity of communication. No mentioning whether it would make a difference .