**ID: 3197
HW03
BIOST 518 – Winter 2015**

Question 1

**Methods:** In order to assess the relationship between maternal smoking and child small for gestational age (SGA) status, first descriptive statistics were generated for the entire study population as well as by maternal smoking status. SGA was defined as babies whose size or weight is below the tenth percentile for their gestational age. Within the below table the mean, standard deviation (SD), and range are shown for continuous variables (Mother’s age and height, parity and child’s birthweight and gestational age) while percentages are shown for binary variables (Mother’s smoking status, and child’s sex). All computing was done with Stata version 13.1. Participants missing data on smoking status or SGA were not included.

**Results:** Descriptive statistics are shown in the table below. Data were available for 751 mother-child dyads, of which 231 (30.8%) were categorized non-smokers and 520 (69.2%) were non-smokers. Among smokers, mothers on average had a slightly higher parity (previous pregnancies) and their most recent child was slightly less likely to be male. Approximately 13.8% of the study population was characterized as having a SGA child; however, SGA status was increased within the maternal smoking group (19.5%) compared with the non-smoking group (11.3%).

|  |  |  |
| --- | --- | --- |
|  |   | **Smoking Status** |
| **Baseline Characteristics\*** | **All Subjects** | **Smoker** | **Non-smoker** |
| Number | 751 | 231 | 520 |
| Mother’s Age | 24.77 (5.36, 14-43) | 25.13 (5.35, 15-42) | 24.61 (5.36, 14-43) |
| Mother’s Height | 156.69 (6.49, 106-176) | 156.80 (7.19, 106-176) | 156.64 (127-175) |
| Parity  | 1.10 (1.21, 0-6) | 1.19 (1.27, 0-6) | 1.06 (1.19, 0-6) |
| SGA (%) | 13.8% | 19.5% | 11.3% |
| Birthweight  | 3105.63 (534.46, 1035-4730) | 2972.16 (512.38, 1410-4550) | 3164.93 (533.85, 1035-4730) |
| Male (%) | 50.1% | 48.1% | 52.3% |
| Gestational Age | 38.18 (1.50, 30-44) | 38.96 (1.36, 33-43) | 39.28 (1.55, 30-44) |
| \*All baseline characteristics are mean (SD, range) unless otherwise indicated.Missing data within non-smokers: n=5 for heightMissing data within smokers: n=1 for height and gestational age  |

Question 2

**Methods:** A logistic regression analysis was undertaken, where SGA was the binary response outcome of interest and maternal smoking was a binary predictor of interest. 95% Cis were generated from the Wald-statistic. Participants with missing data on maternal smoking status and/or SGA status were not included in the analysis.

**Part A:** From logistic regression analysis of 751 participants with both SGA and maternal smoking status data, it is estimated that among smokers, the odds of having an SGA child is increased by a relative 89%. The estimate is statistically significant at p=0.0032. Based on the 95% confidence interval, the finding is not usual if the odds of SGA were increased by between 24% and 188%.

**Part B**

Logodds SGA = -2.06 + 0.637 x Maternal smoking [0,1]

*Non-smokers*

* Log-odds SGA= -2.06 + 0
* Odds SGA = e^-2.06=0.127
* Probability SGA = odds/1+odds = 0.127/(1.127) = 0.1127

*Smokers*

* Log-odds SGA = -2.06+0.637(1)=-1.423
* Odds SGA = e^-1.423 = 0.241
* Probability SGA = 0.241/1.241= 0.1942

The probabilities are very similar to those given for SGA status within the descriptive statistics table because the descriptives are based on all the data; the model is saturated.

**Part C**

1. Logistic regression with non-smoking [0,1] as the predictor of interest is a reparameterization of the same model. It gives the same estimate of the log-odds and 95%CI (but negative), and the same estimate of the SE. The intercept term, that is the value of the log-odds when non-smoking=0, would not be the same, but the P-values and Wald-chi statistic would be the same. In terms of the OR and 95%CI, this regression would give roughly the inverse of the other regression on smoking [0,1] (OR=0.529, 95%CI: 0.346, 0.808). We could tweak the inference to say: it is estimated that among non-smokers, the odds of having an SGA child is reduced by a relative 47.1% compared with smokers; however, the basic scientific relationship is the same.
2. Logistic regression with not-SGA [0,1] as the response of interest is another reparameterization. It would give the same estimate of the log-odds and 95%CI (but negative), and the same estimate of the SE. The intercept would also be the same (but negative), as it reflects the value of the log-odds when (smoking= 0). The P-values and Wald-chi statistic would be the same. In terms of the OR and 95%CI, this regression would give the roughly the inverse of the other regression on SGA [0,1] (OR=0.529, 95%CI: 0.346, 0.808). We could tweak the inference to say: it is estimated that among smokers, the odds of having a normal weight (not-SGA) child is reduced by a relative 47.1% compared with non-smokers; however, the basic scientific relationship is the same.
3. Logistic regression with not-SGA [0,1] as the response of interest and not-smoker [0,1] as the predictor of interest is yet another reparameterization. It would give the same estimate of the log-odds and 95%CI. The intercept term, that is the value of the log-odds when SGA status=0, would not be the same, but the P-values and Wald-chi statistic would be the same. In terms of the OR and 95%CI, this regression would give the same values as the other regression on SGA and smoking (OR=1.89, 95CI: 1.24-2.89). We could tweak the inference to say: it is estimated that among non-smokers, the odds of having a normal weight (not-SGA) child is increased by 89% compared with smokers; however, the underlying scientific relationship is the same.

Question 3

**Methods:** A linear regression analysis was undertaken, where SGA was the binary response outcome of interest and maternal smoking was a binary predictor of interest. The analysis was conducted using Huber-White sandwich estimator for the standard error. Participants with missing data on maternal smoking status and/or SGA status were not included in the analysis.

**Part A:** From linear regression analysis of 751 participants with both SGA and maternal smoking status data, among smoking mothers, the mean probability of SGA is increased by an absolute 8.13% relative to non-smoking mothers. The estimate is statistically significant at p=0.0061. Based on the 95% confidence interval, the finding is not usual if the difference in probability of SGA was between 2.3% and 13.9%.

**Part B**

Probability of SGA = 0.1135 + 0.08134 x Maternal smoking [0,1]

*Non-smokers*

* Probability of SGA= 0.1135
* Odds of SGA = Prob SGA/ (1- Prob SGA) = 0.1135 / (1-0.1135) = 0.128

*Smokers*

* Probability of SGA= 0.1135 + 0.08134 (1) = 0.195
* Odds of SGA = Prob SGA/ (1- Prob SGA) = 0.242

The probabilities are very similar to those given for SGA status within the descriptive statistics table because the descriptives are based on all the data; the model is saturated.

**Part C**

1. Linear regression with non-smoking [0,1] as the predictor of interest is a reparameterization of the same model. It would give the same estimate of the difference in probabilities and 95%CI (but negative), and the same estimate of the SE. The intercept term, that is the value of the difference when non-smoking=0, would not be the same, but the P-values and t-statistic would be the same (t-statistic will be negative). We could weak the inference to say: it is estimated that among non-smokers, the probability of SGA is reduced by an absolute 8.13% relative to smoking mothers; however, the basic scientific relationship is the same.
2. Linear regression with not-SGA [0,1] as the response of interest is another reparameterization of the same model. It would give the same estimate of the mean difference in probabilities and 95%CI (but negative), and the same estimate of the SE. The intercept would not be the same as it reflects the value of the probability of not-SGA when (smoking= 0). It would be equal to the 1 minus the probability in the main regression. P-values and t-statistic would be the same (t-statistic will be negative). We could tweak the inference to say: it is estimated that among smokers, the probability of having a normal weight (not-SGA) child is reduced by an absolute 8.13% relative to non-smokers; however, the basic scientific relationship is the same..
3. Linear regression with not-SGA [0,1] as the response of interest and not-smoker [0,1] as the predictor of interest is yet another parameterization. It would give the same estimate of the difference in probabilities and 95%CI. The intercept term would not be the same as it reflects the value of the probability of not-SGA when (not-smoking= 0). It would be equal to the 1 minus the probability in the second regression (part i). We could tweak the inference to say: it is estimated that among non-smokers, the probability of having a normal weight (not-SGA) child is increased by an absolute 8.13% relative to smokers; however, the basic scientific relationship is the same.

Question 4

**Methods:** A Poisson regression analysis was undertaken, where SGA was the binary response outcome of interest and maternal smoking was a binary predictor of interest. Participants with missing data on maternal smoking status and/or SGA status were not included in the analysis.

**Part A:** From Poisson regression analysis of 751 participants with both SGA and maternal smoking status data, among smoking mothers, the probability of SGA increases by 71.7% relative to non-smoking mothers. The estimate is statistically significant at p=0.0029. Based on the 95% confidence interval, the finding is not usual if the probability was increased by between 20.2% and 145%.

**Part B**

Log (E(SGA|Smoker) = Log (Probability of SGA) = -2.176 + 0.541 x Maternal smoking [0,1]

*Non-smokers*

* Log (Probability of SGA) = -2.176 + 0
* Probability of SGA = e^-2.176 = 0.1135
* Odds of SGA = Prob SGA/ (1- Prob SGA) = 0.1135 / (1-0.1135) = 0.128

*Smokers*

* Log (Probability of SGA)= -2.176 + 0.541 (1) = -1.635
* Probability of SGA = e^-1.635 = 0.195
* Odds of SGA = Prob SGA/ (1- Prob SGA) = 0.242

The probabilities are very similar to those given for SGA status within the descriptive statistics table because the descriptives are based on all the data; the model is saturated.

**Part C**

1. Poisson regression with non-smoking [0,1] as the predictor of interest is a reparameterization of the same model. It would give the same estimate of log-probabilities and 95%CI (but negative), and the same estimate of the SE. The intercept term, that is the value of the difference when non-smoking=0, would not be the same, but the P-values and Wald-Chi statistic will be. We could tweak the inference to say: it is estimated that among non-smokers, the probability of SGA is reduced by 41.8% relative to smoking mothers; however, the basic scientific relationship is the same.
2. Poisson regression with not-SGA [0,1] as the response of interest is another reparameterization of the same model. None of the estimates would be the same. We could tweak the inference to say: it is estimated that among smokers, the probability of having a normal weight (not-SGA) child is reduced by 9.2% relative to non-smokers; however, the basic scientific relationship is the same.
3. Poisson regression with not-SGA [0,1] as the response of interest and not-smoker [0,1] as the predictor of interest is yet another parameterization. None of the estimates would be the same. We could tweak the inference to say: it is estimated that among non-smokers, the probability of having a normal weight (not-SGA) child is increased by 10.1% relative to smokers; however, the basic scientific relationship is the same.

Question 5

Values obtained from linear regression would be the same as those obtained from the t-test because this model is saturated. In particular, the difference of means would be equivalent to the slope in the regression function. Values obtained from logistic regression would be the same as the Chi-squared test for the same reason.

Question 6

**Part A**

**Methods:** A linear regression analysis was undertaken, where SGA was the binary response outcome of interest and maternal age was a continuous predictor of interest. The analysis was conducted using Huber-White sandwich estimator for the standard error. Participants with missing data on maternal age and/or SGA status were not included in the analysis.

**Results:** From linear regression analysis of 755 participants with both SGA and maternal age data, for two groups of mothers differing by one year of age, the mean probability of SGA is decreased by an absolute 0.45%, with the older mothers having a slightly reduced probability of having an SGA child. The estimate is statistically significant at p=0.04. Based on the 95% confidence interval, the finding is not usual if the difference in probability of SGA was between 0.029% and 0.87%.

**Part B**

**Methods:** A Poisson regression analysis was undertaken, where SGA was the binary response outcome of interest and maternal age was a continuous predictor of interest. Participants with missing data on maternal age and/or SGA status were not included in the analysis.

**Results:** From Poisson regression analysis of 755 participants with both SGA and maternal age data, for each increase in maternal age by one year, the probability of SGA decreases by a relative 3.4%. The estimate does not achieve statistical significance at p=0.05. Based on the 95% confidence interval, the finding is not usual if the true reduction in probability of SGA is between 0% and 6.6%.

**Part C**

**Methods:** A logistic regression analysis was undertaken, where SGA was the binary response outcome of interest and maternal age was a continuous predictor of interest. 95% Cis were generated from the Wald-statistic. Participants with missing data on maternal age and/or SGA status were not included in the analysis.

**Results:** From logistic regression analysis of 755 participants with both SGA and maternal age data, for each increase in maternal age by one year, the odds of SGA decreases by a relative 3.9%. The estimate does not achieve statistical significance at p=0.05. Based on the 95% confidence interval, the finding is not usual if the true reduction in odds of SGA is between 0% and 7.6%.

**Part D**

Linear regression: E(SGA | Age) = Probability of SGA = 0.251 - 0.00452 x Maternal Age

* Probability of SGA = 0.161

Poisson regression: Log (E(SGA | Age) = Log (Probability of SGA) = -1.136 – 0.0344 x Maternal Age

* Probability of SGA = 0.161

Logistic regression: Log (Odds SGA | Age) = Log (Odds of SGA) = -0.853 – 0.0398 x Maternal Age

* Probability of SGA = 0.161

The previous estimates are based fitted, saturated regressions across all ages in the sample; thus, they are different from the estimate of the proportion of SGA among 20 year old mothers in the study sample. These estimates are based on information “borrowed” from the entire sample distribution. This estimate of the probability of SGA is 0.075 (SD= 0.267) based on 20 year old mothers in the sample (N=40) only.

Question 7



Question 8

**Part A**

**Methods:** A logistic regression analysis was undertaken, where SGA was the binary response outcome of interest and maternal age was a log-basee-transformed continuous predictor of interest. 95% Cis were generated from the Wald-statistic. Participants with missing data on maternal age and/or SGA status were not included in the analysis.

**Results:** From logistic regression analysis of 755 participants with both SGA and maternal age data, for each increase in 1 unit increase in log-maternal age (~2.72 years), the odds of SGA decreases by a relative 61.5%. The estimate does not achieve statistical significance at p=0.05. Based on the 95% confidence interval, the finding is not usual if the true reduction in odds of SGA is between 0% and 85.3%.

**Part B**

It does not make sense to talk multiplicative changes on the predictor maternal ages. For example, if we used a log-base 2, our inference would be on a 2-fold increase in age, which does not make sense in this population of mothers with an age range of 14-43. Thus, the analysis in part A is silly.