Biostatistics Homework 3

1. Descriptive Statistics

Methods: Descriptive statistics are generated for this cohort study of 755 women with singleton pregnancies in a peri-urban setting in the Western Cape, of South Africa. We describe maternal and fetal characteristics associated with smoking status during pregnancy. Maternal parity was divided according to 1st versus 2nd + pregnancy. Continuous variables (age, height, birth weight) are presented by means, minimum/ maximum and standard deviation. Variables with binary outcomes (infant gender, SGA, first pregnancy) are presented according to percentages. Statistics are also presented according to the outcome of small for gestational age (SGA) babies, defined as weight below the 10% percentile for gestational age in exploratory descriptive statistics.

Descriptive statistics: In the population of 755 women who were enrolled in the study, there were a fairly high proportion of smokers (30.8%). 4 participants did not have smoking status recorded and will be omitted from all further analysis. Overall, smokers tended to be slightly older, more likely to have had more than 1 pregnancy and more likely to have a male infant. Additionally, smokers were more likely to deliver earlier, smaller babies and had a higher proportion of SGA infants. Exploratory descriptive statistics according to outcome of SGA, revealed that women who delivered an infant who was small for gestational age were younger and more likely to be delivering their first child. Infants who were SGA, were more likely to be male with a mean gestational age of 37.9 weeks and birth weight of 2231.1 weeks. As a result, both maternal age, parity and infant gender should be considered as confounders in the analyses below. Because both gestational age and birth weight are included in the definition of SGA, they should not be considered as confounders or effect modifiers in the analyses below.

Of note, many important variables were not considered in this analysis, including maternal chronic medical conditions, recent illness, nutritional status, etc.

Table 1:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Nonsmoker (n = 520) | Smoker (n=231) | All (n=751) |
| Maternal characteristics |  |  |  |
| Age (yrs)1 | 24.6 (5.37,14- 43) | 25.1 (5.35, 15- 42) | 24.8 (5.36, 14-43) |
| Height (cm)1 | 156.6 (6.16, 127-175) | 156.8 (7.12, 106- 176) | 156.7 (6.5, 106-176) |
| First pregnancy (%) | 71.9 | 67.5 | 70.6 |
| Infant characteristics |  |  |  |
| Male (%) | 47.7 | 51.9 | 49 |
| Sga (%) | 11.3 | 19.5 | 13.8 |
| Gestational age (weeks)1 | 39.3 (1.5, 30-44) | 38.9 (1.36, 33- 43) | 39.2 (1.5, 30-44) |
| Birth weight (grams)1 | 3164.9 (533.8, 1035-4730) | 2972.2 (512.38, 1410-4550) | 3105.6 (534.46, 1035-4730) |

1Statistics represent mean (standard deviation, minimum – maximum)

|  |  |  |  |
| --- | --- | --- | --- |
|  | SGA4 (n = 105) | Non-SGA (n = 650)  | All (n= 755) |
| Maternal Characteristics |  |  |  |
| Age (yrs) 1 | 23.8 (4.90, 16-35) | 24.9 (5.45, 14-43) | 24.8 (5.39, 14-43) |
| Height (cm) 1, 2 | 154.6 (5.87, 142-172) | 157.0 (6.54, 106-176) | 156.7 (6.50, 106-176) |
| Tobacco use (%)3 | 43.2 | 28.7 | 30.8 |
| Initial pregnancy (%) | 77.1 | 69.5 | 70.6 |
| Infant characteristics |  |  |  |
| Male (%) | 57.7  | 47.6 | 49.0 |
| Gestational age (weeks)1 | 37.9 (2.2, 30- 42) | 39.4 (1.2, 38-44) | 39.2 (1.5, 30-44) |
| Birth weight (grams)1, 4 | 2231.1 (411.6, 1035-3780) | 3246.2 (402.1, 2510- 4730) | 3105.6 (534.5, 1035-4730) |

1Statistics represent mean (standard deviation, minimum – maximum)

26 individuals are missing data on heigh

34 participants did not have tobacco status recorded

4Infants with SGA are defined as weight < 10% normal for gestational age

2. SGA and maternal smoking status

a. Methods: The relationship between smoking status and the odds of delivering an infant who is SGA was analyzed using simple logistic regression. The Wald estimate was used to construct 95% CI and generate a p-value. The participants without smoking status recorded were excluded from the analysis.

Results: From simple logistic regression analysis, we estimate that in women who smoke, the odds are 89% higher of having a SGA infant compared to women who do not smoke. The constructed 95% CI indicate that this odds ratio would not be unusual if women who smoke had between a 23.8% and 88.8% odds of having an infant with SGA compared to women who do not smoke. With α level set at 0.05, we reject the null hypothesis that the odds ratio of having a SGA baby between smokers and nonsmokers is 1.

b. The odds of a nonsmoker having an infant with SGA is 12.8%. This was obtained from the intercept of the logistic regression model . The probability that a nonsmoker will have an infant with low SGA is 11.3% (odds/ (1+odds). This exactly matches the odds and probabilities calculated from table 1: probability = 59/520 = 11.3 with odds of 12.8. The odds of a smoker having an infant with SGA is 24.2%. This was obtained by exponentiating the fitted model: -2.056 + 0.637β where Β = 1 (i.e. smoker). The probability that a smoker will have an infant with SGA is 19.5%. Once again, our fitted model gives the same answers seen in the descriptive statistics with probability of a smoker having low SGA of 45/231 = 19.4% and odds of 24.2%. These answers correlate with each other because our logistic regression model is a fitted model.

|  |  |  |
| --- | --- | --- |
|  | Fitted model (%) | Descriptive statistics (%) |
| Nonsmoker: probability SGA infant | 11.3 | 11.3 |
| Nonsmoker: odds SGA infant | 12.8 | 12.8 |
| Smoker: probability SGA infant | 19.5 | 19.4 |
| Smoker: odds SGA infant | 24.2 | 24.2 |

c. Hypothetical Regressions:

i. Here, the alternative model is a re- parameterization of the initial model by reversing the order of your binary predictor variable. The resulting slope and intercept will be the probability of the complementary event. For example, in this new logistic regression, the intercept will represent the odds of a smoker having a SGA infant (i.e. the constant is 24.2). The slope will represent the decreasing odds ratio of having a SGA infant in nonsmokers compared to smokers. The width of the CI and the p-value will remain the same.

ii. This once again is a re-parameterization of the initial model by reversing the order of the binary response variable. The width of the CI and the p-value will remain the same.

iii. This final model switches the role of the response variable and predictor of interest in the logistic regression model. The invariance properties of the odds ratio means that the slope which estimates the odds ratio of having a SGA infant (response) in a smoker (predictor) will be the same as the odds ratio of being a smoker (response) if you have a SGA infant. The width of the CI and p-value will remain the same.

3. Probability of SGA based on maternal smoking status

a. Methods: The relationship between the maternal smoking status and delivering an infant who was small for gestational age (SGA) was explored using classical linear regression. The difference in the probability of delivering a SGA infant between mothers who smoked and mothers who did not smoke, was generated by making inference on the regression model. Robust estimates of SE were used to construct 95% confidence intervals and generate a p-value. The 4 participants without smoking status recorded were excluded from all analysis.

Inference: From classic linear regression, the probability of a nonsmoker delivering an infant with SGA was 11.3% and smokers were 8.1% more likely to likely to deliver a infant with SGA. From a robust calculation of 95% CI, we would not be surprised to get this result if smokers were anywhere from 2.3% to 13.9% more likely to deliver an infant who was SGA. Using an α level of 0.05, we can reject with great confidence the null hypothesis that smokers and nonsmokers have the same probability of delivering an infant with SGA (p =0.006)

b. The probability of a nonsmoker delivering an infant who is SGA is given by the intercept of the regression model as 11.3%. The odds can thus be calculated as (p/ (1-p) or 12.8%. The probability for a smoker to deliver an infant with SGA is found using the fitted model of 0.11346 + 0.0813437(1) or 19.5%. The odds for a smoker to deliver an infant with SGA is found by plugging into equation p/(1-p) or 24.2%. Just as in question 2b, the probabilities and odds that we calculated using the fitted model from our linear regression are the same as the probabilities and odds that we calculated from our descriptive statistics. The linear regression gives the same answers as the probability and odds derived from the descriptive statistics because this is a saturated fitted model.

|  |  |  |
| --- | --- | --- |
|  | Fitted model (%) | Descriptive statistics (%) |
| Nonsmoker: probability SGA infant | 11.3 | 11.3 |
| Nonsmoker: odds SGA infant | 12.8 | 12.8 |
| Smoker: probability SGA infant | 19.5 | 19.4 |
| Smoker: odds SGA infant | 24.2 | 24.2 |

c. Hypothetical regressions

i. The hypothetical regression that uses nonsmoker as the predictor of interest will just be a re-parameterization of our initial model. Because the model is saturated, we should be able to predict the new intercept and slope. The intercept should represent the probability of a smoker having an infant with SGA and the slope will represent the decrease in probability of having an infant with SGA if the mother is a nonsmoker. The slope will be the negative of our initial model slope or -0.0813. The p-value and the width of the CI should remain the same.

ii. Once again, this is just a re- parameterization of our initial model by reversing our response variable. As the model continues to be fitted, we once again will be able to predict the intercept and slope. Here the intercept represents the probability of a nonsmoker having an infant who not SGA, or 1 – intercept, in our initial model. The slope will be the same as the slope in question c, part i – i.e. the probability of having a non- SGA baby will decrease if the mother is a smoker. The p-value and width of the confidence intervals stay the same.

iii. Here we have done a third re- parameterization by reversing the response variable and the predictor of interest. Once again, the model remains saturated, so we should be able to predict our intercept, slope, width of CI and p-value. The intercept here is the probability of a smoker having a non-SGA baby (1- 0.195) and the slope will be the increase in probability for a non-smoker to have a nonSGA baby (same slope as our initial model).

4. Ratio of Probability of delivering a SGA infant

a. Methods: A Poisson regression model was used to compare the probability of delivering an infant who was SGA between mothers who smoked and mother who did not smoke. Statistical inference on the ratio of the probability of having an infant who was SGA was computed from the regression slope parameter using the Wald statistic. Robust estimates of the SE were generated with the Huber-White sandwich estimator to construct 95% CI and a p-value. The 4 participants without smoking status were excluded from analysis.

Results: The proportion of nonsmokers delivering a SGA baby was 0.113, while the proportion of smokers delivering a SGA baby was 0.195. Based on 95% CI, the observed ratio of probability of delivering a SGA infant of 0.579, which suggests a 71.7% increase in the probability of delivering a SGA infant in mothers who smoke compared to mothers who don’t smoke, would not be judged unusual if the true increase in probability was between 20.3% and 145%. The p- value of 0.003 means that with high confidence we can reject the null hypothesis that probability of delivering a SGA infant is the same regardless of maternal smoking status.

b. The probability of a nonsmoker delivering an SGA baby is estimated by exponentiating the intercept of our fitted model at 0.113. The resultant calculated odds p/(1-p) is estimated at 12.8%. The probability of a smoker delivering a SGA infant is determined by exponentiating the fitted model 0.5405(1) – 2.176, as 19.5%. The odds of a smoker delivering a SGA infant are found by p/(1-p) at 24.2. Because this is a saturated model, the probability and odds are exactly the same as those estimated by the descriptive statistics.

|  |  |  |
| --- | --- | --- |
|  | Fitted model (%) | Descriptive statistics (%) |
| Nonsmoker: probability SGA infant | 11.3 | 11.3 |
| Nonsmoker: odds SGA infant | 12.8 | 12.8 |
| Smoker: probability SGA infant | 19.5 | 19.4 |
| Smoker: odds SGA infant | 24.2 | 24.2 |

c. Hypothetical regressions:

i. The hypothetical regression that uses nonsmoker as our predictor of interest will not result in the similar measures of association found in re-parameterizations problems 2 and 3. Because it is a saturated model, the intercept will be the proportion of smoker’s with low SGA babies (19.5) and fitted values will be the same. However, the measure of association (i.e. ratio of SGA infants in nonsmokers/ smokers) will be different than found in part a. The p-value should be quite similar.

ii. By switching the response variable to be not- SGA instead of SGA, you will still have a saturated model, and get the expected intercept - i.e. log (1- 0.113), the proportion of nonsmokers who have non-SGA babies. However, the ratio/ measure of association is a non-linear transformation and will not be related to the slope in part a. The p-value should be quite similar.

iii. By reversing both the predictor of interest and the response variable, you still have a fitted model. The intercept will be the proportion of smokers with non-SGA infants log (1-0.195). Once again, the ratio of non-SGA in nonsmokers/ non-SGA infants in smokers will not be similar to part a.

5. Methods: The proportion of infants born with SGA was compared between maternal nonsmokers and maternal smokers. Differences in the probability of delivering a low SGA infant between the two groups was tested using Pearson’s chi squared test for independence. 95% CI were constructed using the Wald test. The 4 patients who did not have smoking status recorded were removed from the analysis.

Results: Of the 231 smokers, 45 had infants with SGA with a risk of 19.5% and of the 520 nonsmokers, 59 had SGA infants with a risk of 11.3%. Based on 95% confidence intervals, the observed difference in the risk of having an SGA infant between smokers and nonsmokers of 8.1% would not be unusual if the true difference in risk were between 2.3% and 13.9%. With a p-value of 0.003, we can reject the null hypothesis that SGA infants is not associated with maternal smoking history.

The Pearson’s Chi squared test for independence gives us the exact same answers we got in the saturated regression models performed above in questions 2-4. The risk difference of 8.1% corresponds to the slope in the linear regression model. The risk ratio of 71.7% corresponds to inference on the slope of the poisson regression and the odds ratio corresponds to inference on the slop of the logistic regression. The estimated measure of association with a p-value of 0.003 is also similar to the regression analyses.

6. a. Risk Difference:

Methods: The relationship between maternal age and the risk of having a SGA infant were compared using linear regression. To make the intercept more interpretable, maternal age was linear transformed in to age + 14. Point estimates of the linear association of having a SGA infant and maternal age were based on the slope parameter. The Huber-White estimator was used to compute SE which did not assume equal variance across groups. 95% CI and p- value were constructed using the Wald statistics. Participants missing smoking status or age were excluded from analysis.

Results: From linear regression analysis the mean probability of having a SGA infant at the age of 14 (the youngest participant in the study) is 25.1%. We estimate that the mean probability of having a SGA infant decreases by 0.45% for every 1 year that maternal age increases. By 95% CI, this decrease in probability of having a SGA infant would not be surprising if the true decrease were anywhere between 0.87% and 0.03%. Using an α level of 0.05, we reject the null hypothesis that maternal age and low SGA are not associated (p= 0.04).

b. Risk Ratio

Methods: The relationship between maternal age and having a SGA infant was explored using a Poisson regression model. Statistical inference on the ratio of the probability of having a SGA infant with increasing maternal age as a continuous variable was based on the Wald statistic computed from the regression slope parameter. 95% confidence intervals and 2 –sided p-value were constructed based on the robust estimate of standard error from the Huber- White sandwich estimator. Participants without maternal smoking status age recorded were excluded from the analysis.

Results: From Poisson regression, we estimate that for each 1 year increase in maternal age, the risk ratio of delivering an SGA infant decreases by 3.4%. By 95% CI, this decrease in risk ratio would not be surprising if the real decrease were between 6.6% and 0.03%. With a p-value of 0.04, we reject the null hypothesis that the risk ratio of having a SGA infant with advancing maternal age is 1.

c. Odds Ratio

Methods: The odds of having a SGA infant were compared between mothers with different ages using logistic regression. Statistic inference on the ratio of odds of having a SGA infant with advancing maternal age modeled as a continuous variable was based on the Wald statistic computed from the regression parameter. 95% CI and the p-value were computed using the logistic regression parameter estimate. To make results more interpretable, estimates are rescaled according to a 10 year increase in age. Participants with smoking status or age recorded were excluded from analysis.

Results: From logistic regression, the odds of having a SGA infant decrease by 6.4% by every 10 year increase in maternal age. By 95% CI, this decrease in odds would not be unusual if the true odds of having a SGA infant were anywhere between 77% lower and 0.7% higher with every 10 year increase in maternal age. Based on p value of 0.05, we reject the null hypothesis that the odds ratio of delivering a SGA infant with advancing maternal age is 1.

d.

|  |  |
| --- | --- |
|  | Probability for 20 year old to have SGA-infant |
| Linear Regression a) | 16.1% |
| Poisson Regression b) | 16.1% |
| Logistic Regression c) | 16.1% (odds was 19.3%) |

Each of the fitted models give approximately equal estimates for the probability of having an SGA infant at the age of 20. Of note, the fitted model for the logistic regression in part c gave the odds of having a SGA infant at age 20, so that probability needed to be calculated from the formula odds/ (1+ odds). They are all accurate estimates of the model parameter because they are saturated models.

7. a.



b. i. Linear Regression

ii. Poisson Regression Estimates



iii. Logistic Regression Estimates



8.a. Methods: The odds of delivering a SGA infant were compared across log –transformed (base 10) maternal age using logistic regression. Statistical inference was based on the Wald statistic computed from the slope regression parameter and its standard error was used to generate 95% confidence intervals and a p-value. The 1 individual without maternal age recorded was excluded from analysis.

Results: There were 754 individuals available for analysis, 105 had SGA infants and 650 had non-SGA infants. From logistic regression analysis, we estimate that the odds of having a SGA infant are 88.9% lower for each 10 fold increase in maternal age. Based on 95% confidence intervals, this estimate would not be unusual if the odds of having a SGA infant were between 98.9% and lower and 8% higher for each 10 fold increase in age. A p-value of 0.06 means that we can not reject the null hypothesis that the odds ratio of having a SGA infant for each 10 fold increase in maternal age is 1.

b. It is slightly ridiculous to log transform age because biologically age is not multiplicative. Furthermore, it is easier for the reader to think of age increasing by 1 year rather than a 2 fold or 10 fold or exponential increase in age. However, transforming on a logarithmic scale does reduce the influence of outliers and may fit our non-linear data slightly better.