3293

BIOST 518

HW03

2 February 2015

1. **Methods**: Mean, standard deviation, minimum, and maximum are presented for the continuous variables (mother’s height, age, parity and infant’s birth weight, gestational age) and percentages are presented for the binary variables (sex and mother’s smoking status). The summary statistics are presented for the overall sample and stratified by small for gestational age (SGA) outcome and smoking status, the primary predictor of interest. Individuals who were missing data were only excluded for the analyses that involved the missing data.

**Inference:** There were 755 mother-infant pairs. There was no missing data on SGA status, but there were 4 missing data on smoking status and thus were excluded from all further analysis that involved smoking status. There was a trend towards taller mothers among infants who were not SGA, but mean height was not different between smoking groups and therefore not a potential confounder. Infants who were SGA tended to have mothers who were shorter, younger, smoked, and had fewer prior pregnancies and the infants tended to be female, younger gestational age, and birth weight.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Mean** | **Std. Dev.** | **Min.** | **Max.** |
| Height1 | 156.68 | 6.50 | 106 | 176 |
| Age | 24.79 | 5.38 | 14 | 43 |
| Parity | 1.10 | 1.21 | 0 | 6 |
| Birthweight2 | 3105.63 | 534.46 | 1035 | 4730 |
| Gestational age3 | 39.18 | 1.50 | 30 | 44 |
|  | **%** |  |  |  |
| Smoked2 | 30.76 |  |  |  |
| Male2 | 51.00 |  |  |  |
| SGA | 13.91 |  |  |  |

1. missing 6 observations 2. Missing 4 observations 3. Missing 5 observations

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Small for gestational age** | | | |  | **Not small for gestational age** | | | |
| **Variable** | **Mean** | **Std. Dev.** | **Min.** | **Max.** |  | **Mean** | **Std. Dev.** | **Min.** | **Max.** |
| Height1 | 154.56 | 5.87 | 142 | 172 |  | 157.01 | 6.54 | 106 | 176 |
| Age | 23.85 | 1.90 | 16 | 35 |  | 24.94 | 5.45 | 14 | 43 |
| Parity | 0.90 | 1.11 | 0 | 6 |  | 1.13 | 1.23 | 0 | 6 |
| Birthweight2 | 2231.11 | 411.60 | 1035 | 3780 |  | 3246.21 | 402.13 | 2510 | 4730 |
| Gestational age | 37.92 | 2.20 | 30 | 42 |  | 39.38 | 1.24 | 28 | 41 |
|  | **%** |  |  |  |  | **%** |  |  |  |
| Smoked2 | 43.27 |  |  |  |  | 28.75 |  |  |  |
| Male2 | 42.31 |  |  |  |  | 52.40 |  |  |  |

1. missing 6 observations 2. Missing 4 observations

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Smoked during pregnancy** | | | |  | **Did not smoke during pregnancy** | | | |
| **Variable** | **Mean** | **Std. Dev.** | **Min.** | **Max.** |  | **Mean** | **Std. Dev.** | **Min.** | **Max.** |
| Height1 | 156.80 | 7.19 | 106 | 176 |  | 156.64 | 6.16 | 127 | 175 |
| Age | 25.13 | 5.35 | 15 | 42 |  | 24.61 | 5.37 | 14 | 43 |
| Parity | 1.19 | 1.27 | 0 | 6 |  | 1.06 | 1.19 | 0 | 6 |
| Birthweight2 | 2972.16 | 512.38 | 1410 | 4550 |  | 3164.93 | 533.85 | 1035 | 4730 |
| Gestational age | 38.96 | 1.36 | 33 | 43 |  | 39.28 | 1.55 | 30 | 44 |
|  | **%** |  |  |  |  | **%** |  |  |  |
| SGA | 19.48 |  |  |  |  | 11.35 |  |  |  |
| Male2 | 48.05 |  |  |  |  | 52.031 |  |  |  |

Missing 4 observations that did not report smoking status during pregnancy

2a. **Methods:** Logistic regression was used to compare the ratio of odds of SGA between mothers who smoked during pregnancy and those who did not. A 2-sided p-value and a 95% confidence interval were calculated using Wald-based estimates. Significance was determined using an alpha level of .05. There were 4 individuals missing data on smoking who were excluded.

Log(odds SGA=1)= B0 + B1(smoker)

**Inference:** We estimate that among mothers who smoked, the odds of SGA are a relative 89.04% higher than among mothers who did not smoke during pregnancy and statistically significant, p=0.003. Given the 95% confidence interval, our results are consistent with a true difference in odds between 23.76 and 188.75%.

2b. Odds = probability/(1-probability)

Probability=odds/(1+odds)

**Odds of SGA for non-smokers = 0.12798**

**Probability of SGA for non-smokers = 0.12798/(1+0.12798) = 0.1135**

Odds of SGA for smokers/odds of SGA for non-smokers = 1.89038

Odds of SGA for smokers/0.12798 = 1.89038

**Odds of SGA for smokers = 0.24193**

**Probability of SGA for smokers = 0.24193/(1+0.24193) = 0.1948**

In the description of the sample we calculated that the proportion of SGA among smokers and non-smokers was 19.48% and 11.35% respectively. These are the same probabilities estimated by our linear regression, which is what we would expect since the model is saturated.

2ci. The model has just been reparameterized, so the underlying relationship between SGA and smoking has not changed. Therefore the inference is the same. The output is different because we’re comparing different groups, but the composition of the groups has not changed and so the numbers are related. The y-intercept is now the odds of SGA given the mother is a smoker.

Log(odds SGA=1)= B0 + B1(nonsmoker)

2cii. The model has just been reparameterized, so the underlying relationship between SGA and smoking has not changed. Therefore the inference is the same. The output is different because we’re comparing different groups, but the composition of the groups has not changed and so the numbers are related. The y-intercept is the odds of SGA not occurring given non-smoking mother.

Log(odds NotSGA=1)= B0 + B1(smoker)

2ciii. The model has just been reparameterized, so the underlying relationship between SGA and smoking has not changed. Therefore the inference is the same. The output is the same, except for the y-intercept, because we’re comparing binary outcomes and have inverted both. The y-intercept is the odds of SGA not occurring if the mother smokes.

Log(odds NotSGA=1)= B0 + B1(nonsmoker)

3a. **Methods:** Linear regression was used to compare the difference in proportions of SGA between mothers who smoked during pregnancy and those who did not. A 2-sided p-value and a 95% confidence interval were calculated using the Huber-White sandwich estimator for robust standard error. Significance was determined using an alpha level of .05. 4 individuals were excluded due to missing data on smoking status.

E(SGA|)=B0 + B1(smoker)

**Inference:** We estimate that among mothers who smoked, the probability of SGA is an absolute 8.13% higher than among mothers who did not smoke during pregnancy and the difference in probabilities is statistically significant, p=0.006. The estimate for the probability of SGA among non-smokers is 11.35%, the y-intercept. Given the 95% confidence interval, our results are consistent with a true absolute difference in probabilities between 2.33 and 13.94%.

3b. Odds = probability/(1-probability)

Probability=odds/(1+odds)

**Probability of SGA for non-smokers = 0.1135**

**Odds of SGA for non-smokers = 0.1135/(1-0.1135) = 0.1280**

**Probability of SGA for smokers = 0.1135 + 0.0813(1) = 0.1948**

**Odds of SGA for smokers = 0.1948/(1-0.1948) = 0.2419**

In the description of the sample we calculated that the proportion of SGA among smokers and non-smokers was 19.48% and 11.35% respectively. These are the same probabilities estimated by our linear regression, which is what we would expect since the model is saturated.

3ci. The model has just been reparameterized, so the underlying relationship between SGA and smoking has not changed. Therefore the inference is the same. The output is different because we’re comparing different groups, but the composition of the groups has not changed and so the numbers are related. The y-intercept will be the probability estimated above for SGA among smokers.

E(SGA|X)=B0 + B1(nonsmoker)

3cii. The model has just been reparameterized, so the underlying relationship between SGA and smoking has not changed. Therefore the inference is the same. The output is different because we’re comparing different groups, but the composition of the groups has not changed and so the numbers are related. The y-intercept will be the probability of not SGA among non-smoker.

E(NotSGA|X)=B0 + B1(smoker)

3ciii. The model has just been reparameterized, so the underlying relationship between SGA and smoking has not changed. Therefore the inference is the same. The output is the same, except for the y-intercept, because we’re comparing binary outcomes and have inverted both. The y-intercept will be the probability estimated above for not SGA among smokers.

E(NotSGA|X)=B0 + B1(nonsmoker)

4a. **Methods:** Poisson regression was used to compare the difference in proportions of SGA between mothers who smoked during pregnancy and those who did not. A 2-sided p-value and a 95% confidence interval were calculated using the Huber-White sandwich estimator for robust standard error. Significance was determined using an alpha level of .05. 4 individuals were excluded due to missing data on smoking status

**Inference:** We estimate that among mothers who smoked, the rate of SGA is 1.71 times higher than among mothers who did not smoke during pregnancy and the difference in rate is statistically significant, p=0.006. We estimate the rate of SGA among non-smokers is 11.35%, the y-intercept exponentiated. Given the 95% confidence interval, our results are consistent with a true difference in rates between 1.16 and 2.53.

Log[E(SGA|X)= B0 + B1(smoker)

4b. Odds = probability/(1-probability)

Probability=odds/(1+odds)

**Probability of SGA for non-smokers = exp(-2.176) = 0.1135**

**Odds of SGA for non-smokers = 0.1135/(1-0.1135) = 0.1280**

**Probability of SGA for smokers = exp(-2.176 + 0.5405(1)) = 0.1948**

**Odds of SGA for smokers = 0.1948/(1-0.1948) = 0.2419**

In the description of the sample we calculated that the proportion of SGA among smokers and non-smokers was 19.48% and 11.35% respectively. These are the same probabilities estimated by our Poisson regression, which is what we would expect since the model is saturated.

4ci. The model has just been reparameterized, so the underlying relationship between SGA and smoking has not changed. Therefore the inference is the same. The output is different because we’re comparing different groups, but the composition of the groups has not changed and so the numbers are related. The y-intercept will be the rate estimated above for SGA among smokers.

Log[E(SGA|X)= B0 + B1(nonsmoker)

4cii. The model has just been reparameterized, so the underlying relationship between SGA and smoking has not changed. Therefore the inference is the same. The output is different because we’re comparing different groups, but the composition of the groups has not changed and so the numbers are related. The y-intercept will be the rate of not SGA among non-smoker.

Log[E(NotSGA|X)= B0 + B1(smoker)

4ciii. The model has just been reparameterized, so the underlying relationship between SGA and smoking has not changed. Therefore the inference is the same. The output is the same, except for the y-intercept, because we’re comparing binary outcomes and have inverted both. The y-intercept is the rate of SGA not occurring given a mother who smokes.

Log[E(NotSGA|X)= B0 + B1(nonsmoker)

5. Methods: Individuals were stratified based on smoking status and the difference in the means of SGA were compared using a two-sample t-test that did not assume equal variance. 95% confidence intervals and two-sided p-value were calculated using Wald-based estimates.

Inference: Because our regression models were all saturated, we obtain the same probabilities of SGA given a particular smoking status. The standard errors are calculated differently, so we do not expect them or the 95% confidence intervals to be identical. If we had used the t-test that assumes equal variance and simple linear regression (no robust standard error) the results would be identical.

6a. **Methods:** Linear regression was used to compare the difference in proportions of SGA across maternal ages. A 2-sided p-value and a 95% confidence interval were calculated using the Huber-White sandwich estimator for robust standard error. Significance was determined using an alpha level of .05.

E(SGA|Age)= B0+B1Age

**Inference:** We estimate that the probability of SGA declines an absolute 4.51% per 10 years older the mother is, though it is not statistically significant, p=0.054. Given the 95% confidence interval, our results are consistent with a true difference in probabilities between an absolute increase of 0.07% and an absolute decrease of 9.10%. The y-intercept does not have a useful interpretation as it corresponds to a newborn mother. We conclude that there is not a statistically significant linear relationship between maternal age and probability of SGA.

6b. **Methods:** Poisson regression was used to compare the multiplicative differences in rates of SGA across maternal ages. A 2-sided p-value and a 95% confidence interval were calculated using Wald-based statistics. Significance was determined using an alpha level of .05.

log[E(SGA|Age)]= B0+B1Age

**Inference:** We estimate that the rate of SGA declines a relative 3.38% per year increase in maternal age, though it is not statistically significant, p=0.074. Given the 95% confidence interval, our results are consistent with a true difference in probabilities between a decline of 6.96% and an increase of 0.33%. The y-intercept does not have a useful interpretation as it corresponds to a newborn mother. We conclude that there is not a statistically significant relationship between maternal age and probability of SGA.

6c. **Methods:** Logistic regression was used to compare the ratio of odds of SGA across maternal ages. A 2-sided p-value and a 95% confidence interval were calculated using Wald-based statistics. Significance was determined using an alpha level of .05.

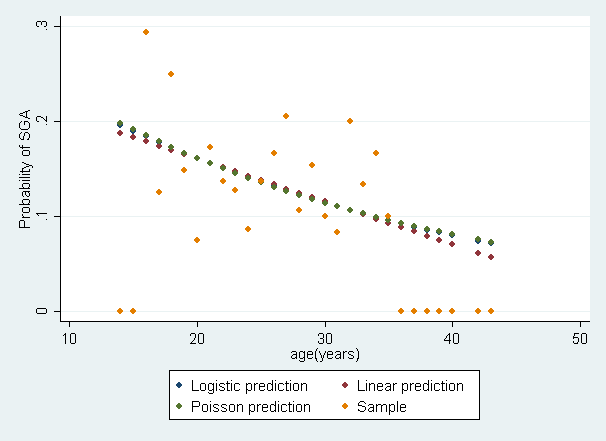
log(odds SGA=1) = B0+B1Age

**Inference:** We estimate that the odds of SGA declines a relative 3.90% per one year increase in maternal age, which is not statistically significant, p=0.054. Given the 95% confidence interval, our results are consistent with a true relative decline in the odds of SGA between -0.08 and 7.72%. The y-intercept does not have a useful interpretation as it corresponds to a newborn mother.

6d. The model is not saturated because there are only 2 parameters and more than two ages present in the sample. The regression borrow from the rest of the data and that is why the estimates are different from what is seen in the sample.

|  |  |
| --- | --- |
|  | Proportion of SGA among 20 year olds |
| Sample | 7.50% |
| Linear regression | 16.07% |
| Poisson regression | 16.13% |
| Logistic regression | 16.13% |

7. None of the models exactly predict what is seen in the sample because the models are not saturated. The logistic and Poisson regression lines a slightly curved du to the loglink.



8a. **Methods:** Logistic regression was used to compare the ratio of odds of SGA across log-base 2 transformed maternal ages. A 2-sided p-value and a 95% confidence interval were calculated using Wald-based statistics. Estimates were exponentiated for reporting.

Log(odds(SGA=1)= B0+B1log(Age)

**Inference:** We estimate that the odds of SGA are a relative 48.37% lower per 2-fold increase in maternal age, which is not statistically significant, p=0.058. Given the 95% confidence interval, our results are consistent with a true relationship in odds between a relative 73.96% lower and 2.38% higher per 2-fold increase in maternal age. The y-intercept does not have a useful interpretation as it corresponds to a newborn mother. We conclude that there is not statistically significant relationship between maternal age and odds of SGA.

8b. Fold increases don’t make sense for reporting changes in odds of SGA across ages. We are not interested in the ratio of odds of SGA between 20 and 40 years olds.