Biostat 518

HW #3

ID: 3502 Score: 144 out of 160

1. Methods: Per each group defined by newborn size as small for gestational age or not small for gestational age, descriptives continuous variables (mean, standard deviation, minimum and maximum) are computed for mother’s height and age at enrollment, birth weight of infant and gestational age; and frequencies are used to describe mother smoker status, parity and infant’s gender.

Results: The dataset used for the analysis contains 755 women. Overall 11 (1.5%) women hasn’t data at least on one variable: 6 miss mother’s height; 4 miss value on smoking status, baby’s birth weight, gender and gestational age; 1 miss mother miss 1 gestational age. All records have the small for gestational age filled.

As shown on table 1 over the 755 pregnant women 105 (13.9%) delivered a baby classified as small for gestational age (SGA). Regarding mother characteristics mothers, among those with SGA newborn 43.3% smoked compared to 28.8% who smoked among the non SGA newborn mothers. On parity, seems that on the non SGA there is a slight tendency for higher parity compared to the SGA group. Other characteristics are almost similar namely mother’s height, age at enrollment and parity.

Regarding babies’ characteristics, among the SGA group there are 42.3% males compared to 52.4% males among the non-SGA. Babies from SGA group weighted at birth on average 2.2 Kg compared to 3.2 Kg on non SGA; and SGA babies were born on average with 37.9 weeks of gestation whereas the non SGA had 39.4 weeks.

Table 1 – Descriptives of mother and infant characteristics per newborn size



There were 45/231 (19.5%) SGA babies among smoking mothers compared to 59/520 (11.4%) among non smoking mothers. Score: 10. Very nice work.

2.a) Methods: a logistic regression with robust standard errors is used to compute the association between with the odds for small for gestational age (SGA) and maternal smoking status. 95% confidence intervals and p-values from Wald-test are reported. The significance level is set to 5%.

Results: Only 751 records are used (refer to table 2), due to missing values on maternal smoking status. Smoking mothers have 1.89 times higher odds for SGA baby than the non-smoking mothers. This odds-ratio wouldn’t unusual to be found between 1.24 and 2.89 and we reject the null hypothesis of an odds-ratio equal to 1 with p-value of 0.003.

Table 2 – Logistic regression coefficients with outcome SGA baby and predictor as smoker mother

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Odds Ratio** | **95% CI** | **p** |
| Total cases in the model | 751 |  |  |
|  |  |  |  |
| Mother smoke | 1.890 | 1.237 - 2.888 | 0.003 |
| Intercept | 0.128 | 0.098 - 0.168 | < 0.001 |
|  |  |  |  |
| Wald test = 8.67 |  |  | 0.003 |

Score: 10; It would be nice to mention odd of SGA in each group before comparing the two groups.

2.b) Methods: a logistic regression with robust standard errors is used to compute the association between with the odds for SGA and maternal smoking status. The estimated equation is used to compute the odds for SGA according to the maternal smoking status. Then the proportion of SGA in each level o smoking status is computed using the formula p=odds/(1+odds).

Results: The odds for SGA on non smokers mothers was 0.128 compared to the odds of 0.242 on smoking mothers. The proportion of SGA babies on non smoking mothers is 11.4% compared to 19.5% on smoking mothers.

The odds-ratio reported on 2a) could be computed from the descriptives table (0.433/(1-0.433))/(0.288/(1 -0.288)) = 1.89.

On the descriptive table no proportion of SGA is reported for non smoking mothers. But the above proportions computed from the regression are similar to those obtained by counting frequencies. Among the 520 smoking mothers 59 (59/520 = 11.4%) had a SGA baby. And among the 231 non smoking mothers 45 (45/231 = 19.5%) had a SGA baby.

Score: 4; -1 for not mentioning any reason of agreement. The reason of agreement is that the logistics model we fit is saturated.

2.c) i. If instead of smoker we use non smoker dummy variable for the predictor. We can see the relationship by following these equations:

*(SGA and NoSGA represent probabilities for SGA and No SGA respectively).*

From this it is clear that the new intercept in odds scale is the product of the previous coefficients (2a) on odds-scale: 0.128\*1.890 = 0.242.

The new slope is the inverse of the previous slope 1/1.890378 = 0.529.

The Wald-test wouldn’t change because we are using the same variables, model and records. We just relabeled one of the variables.

ii. If instead of the SGA we used its complementary for the dependent variable. We can see the relationship by following these equations:

*(SGA and NoSGA represent probabilities for SGA and No SGA respectively).*

This means the new intercept on odds scale is just the inverse of that on 2a: 1/0.128 = 7.81.

And the new slope on odds scale is just the inverse of that obtained on 2a: 1/1.89 = 0.53.

The Wald-test wouldn’t change because we are using the same variables, model and records. We just relabeled one of the variables.

iii. If we modeled with the complementary of SGA and smoking we would proceed successfully from these equations.

*(SGA and NoSGA represent probabilities for SGA and No SGA respectively).*

This means the new intercept, on odds scale, is the inverse of multiplication the coefficients on 2a (they are already in odds scale): 1/(0.128\*1.890)= 4.13.

The slope in odds scale is the same as on 2a.

The Wald-test wouldn’t change because we are using the same variables, model and records. We just relabeled one of the variables.

Score: 10;

3.a) Methods: a linear regression with robust standard errors is used to compute the association between with the probabilities for small for gestational age (SGA) and maternal smoking status. 95% confidence intervals and p-values from Fisher-test are reported. The significance level is set to 5%.

Results: Only 751 records are used (refer to table 3), due to missing values on maternal smoking status. Smoking mothers have 8.13% higher probabilities for SGA baby than the non-smoking mothers. This probability-difference wouldn’t unusual to be found between 2.3 and 13.9% and we reject the null hypothesis of a probability difference of zero p-value of 0.006.

Table 3 – Linear regression coefficients with outcome SGA baby and predictor as smoker mother

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Probability diference** | **95% CI** | **p** |
| Total cases in the model | 751 |  |  |
|  |  |  |  |
| Mother smoke | 0.081 | 0.023 - 0.139 | 0.006 |
| Intercept | 0.113 | 0.086 - 0.141 | < 0.001 |
|  |  |  |  |
| Fisher test = 7.56 |  |  | 0.006 |

Score: 9; -1. The p value and CI is computed using Wald statistics.

3.b) Methods: a linear regression with robust standard errors is used to compute the association between with the probabilities for small for gestational age (SGA) and maternal smoking status. The estimated equation is used to compute the probabilities of a SGA outcome per maternal smoking status.

Results: According to the regression non smokers mothers had a probability of 11.3% of delivering a SGA baby whereas smokers mothers had 11.3% + 8.1% = 19.4% probability of having a SGA baby. These are the same probabilities computed under descriptive (question 1). So the odds will be the same as computed on 2b.

Score: 5; Note: This is because that the model is saturated.

3.c)

i. If instead of smoker we use non smoker dummy variable for the predictor. We can see the relationship by following these equations:

*(SGA and NoSGA represent probabilities for SGA and No SGA respectively).*

The new intercept is the sum of the intercept and slope on table 3: 0.081 + 0.113 = 0.194.

The slope is the same as on table 3 but with negative sign: -0.081.

ii. If instead of the SGA we used its complementary for the dependent variable. We can see the relationship by following these equations:

*(SGA and NoSGA represent probabilities for SGA and No SGA respectively).*

The new intercept will be complementary of the intercept on table 3: 1 – 0.113 = 0.887.

The slope is the same as on table 3 but with negative sign: -0.081.

iii. If we modeled with the complementary of SGA and smoking we would proceed successfully from these equations:

*(SGA and NoSGA represent probabilities for SGA and No SGA respectively).*

The new intercept will be complementary of the sum of the intercept and slop on table 3: 1 – 0.113 – 0.081 = 0.806.

The slope is the same as on table 3: 0.081.

Score: 9; -1 for not mentioning p value.

4.a) Methods: a Poisson regression with robust standard errors is used to compute the association between with the probabilities for small for gestational age (SGA) and maternal smoking status. 95% confidence intervals and p-values from Wald-test are reported. The significance level is set to 5%.

Results: Only 751 records are used (refer to table 4), due to missing values on maternal smoking status. Smoking mothers have 1.72 times higher probability for SGA baby than the non-smoking mothers. This ratio of probabilities wouldn’t unusual to be found between 1.20 and 2.45 and we reject the null hypothesis of a probability difference of zero p-value of 0.003.

Table 4 – Poisson regression coefficients with outcome SGA baby and predictor as smoker mother

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Rate Ratio** | **95% CI** | **p** |
| Total cases in the model | 751 |  |  |
|  |  |  |  |
| Mother smoke | 1.717 | 1.203 - 2.451 | 0.003 |
| Intercept | 0.113 | 0.089 - 0.144 | < 0.001 |
|  |  |  |  |
| Wald test = 8.86 |  |  | 0.003 |

Score: 10;

4.b) Methods: a Poisson regression with robust standard errors is used to compute the association between with the probabilities for small for gestational age (SGA) and maternal smoking status. The estimated equation is used to compute the probabilities of a SGA outcome per maternal smoking status.

Results: According to the regression non smokers mothers had a probability of 11.3% of delivering a SGA baby whereas smokers mothers had 0.113 \* 1.717 = 19.4% probability of having a SGA baby. These are the same probabilities computed under descriptives (question 1). So the odds will be the same as computed on 2b.

Score: 5; Note: This is because that the model is saturated.

4.c)

i. If instead of smoker we use non smoker dummy variable for the predictor. We can see the relationship by following these equations:

*(SGA and NoSGA represent probabilities for SGA and No SGA respectively).*

The new intercept on probability scale is the product of the intercept and the rate-ratio on table 4: 0.113 \* 1.717 = 0.194.

The slope on probability scale is the inverse of that reported on table 4: 1/1.717 = 0.582.

ii. If instead of the SGA we used its complementary for the dependent variable. We can see the relationship by following these equations:

*(SGA and NoSGA represent probabilities for SGA and No SGA respectively).*

This equation can not be transformed into factors that guarantee β0 separation from β1. But we can see that if smoke dummy variable is 0 the new intercept on ratio scale is : 1 - 0.1134615 = 0.8865.

And if smoke is 1 we compute the slope on ratio scale; : (1 - 0.1134615\*1.716927)/(1 - 0.1134615) = 0.9082

We can compare to the software output:

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Rate Ratio** | **95% CI** | **p** |
| Total cases in the model | 751 |  |  |
|  |  |  |  |
| Mother smoke | 0.908 | 0.846 - 0.975 | 0.007 |
| Intercept | 0.887 | 0.860 - 0.914 | < 0.001 |
|  |  |  |  |
| Wald test = 7.15 |  |  | 0.008 |

Score: 9; -1 for not comparing p values across models.

iii. If we modeled with the complementary of SGA and smoking we would proceed successfully from these equations:

*(SGA and NoSGA represent probabilities for SGA and No SGA respectively).*

This equation can not be transformed into factors that guarantee β0 separation from β1. But if dummy nosmoke is zero that implies the new intercept on ratio scale is = 1 – (.1134615\*1.716927) = 0.8052.

And for the slope on ration scale is = [1 – (.1134615)]:0.8052 = 1.1012.

Which are comparable to the coefficients computed from the software:

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Odds Ratio** | **95% CI** | **p** |
| Total cases in the model | 751 |  |  |
|  |  |  |  |
| Mother smoke | 1.102 | 1.028 - 1.182 | 0.007 |
| Intercept | 0.804 | 0.755 - 0.857 | < 0.001 |
|  |  |  |  |
| Wald test = 7.39 |  |  | 0.007 |

5. Under Biost 517 we covered tests for independence distribution of one binary variable over the other binary. We could have used:

- Chi-squared test – the result would be similar to the Wald test we computed on logistic and Poisson regression.

- z-test for proportions – the result from this would be the square-root of the Wald test

Also we learned to use t-test. If we use t-test with unequal variances assumption to compare the probabilities of SGA on each group defined by smoking status we obtain a mean difference of probabilities that corresponds to slope of linear regression; standard error of that difference corresponds standard error with robust estimation of slope; also the t-test value is the same as the t-test for the slope.

Score: 8. -2 for no explicit comparison for the first two methods;

6a). Methods: a linear regression with robust standard errors is used to compute the association between with the probabilities for small for gestational age (SGA) and maternal age at enrollment. 95% confidence intervals and p-values from Fisher-test are reported. The significance level is set to 5%.

Results: According to the regression, for each 5 years old higher on maternal age at enrollment there is a tendency of 2.3% lower probability of SGA and it wouldn’t be unusual if this probability difference were found between 0.1 and 4.4%. The null hypothesis of no association is rejected with p-value 0.036.

Score: 9; -1 for the implicit interpretation of CI. When talking the true difference, the direction of the difference should be stated clearly.

6b). Methods: a Poisson regression with robust standard errors is used to compute the association between with the probabilities for small for gestational age (SGA) and maternal age at enrollment. 95% confidence intervals and p-values from Wald-test are reported. The significance level is set to 5%.

Results: According to the regression, for each 5 years old higher on maternal age at enrollment there is a tendency of 15.8% relative lower probability of delivery a SGA baby and it wouldn’t be unusual if this probability ratio were found between 0.5 and 28.9%. The null hypothesis of no association is rejected with p-value 0.046.

Score: 9; -1 for wrong numbers.

6c). Methods: a logistic regression with robust standard errors is used to compute the association between with the odds for small for gestational age (SGA) and maternal age at enrollment. 95% confidence intervals and p-values from Wald-test are reported. The significance level is set to 5%.

Results: According to the regression, for each 5 years old higher on maternal age at enrollment there is a tendency of 18.0% relative lower odds of delivery a SGA baby and it wouldn’t be unusual if this odds-ratio were found between 0.4 and 32.6%. The null hypothesis of no association is rejected with p-value 0.046.

Score: 9; -1 for wrong numbers and the wrong rejection conclusion.

6.d) Methods: equations estimated on 6a-6c are used to compute the probability of having a baby SGA given the mother has 20 years old at enrollment. Results are summarized in a table.

Results: Table 5 shows probabilities for SGA given the mother had 20 years old at enrollment. It is for note the high similarities of the 3 probabilities.

Table 5 – Probabilities computed through regression of SGA on mother age at enrollment

|  |  |  |
| --- | --- | --- |
|  | **Probability (%)** | **95% CI** |
| Linear regression | 16.07 | 12.62 - 19.51 |
| Poisson regression | 16.13 | 12.98 - 20.00 |
| Logistic regression | 16.13 | 12.92 - 19.94 |
|  |  |  |

Score: 6; -2 for not providing sample proportion of SGA.;-2 for no comment about the similarity and difference between these three estimated and sample proportion;

7.a) Methods: Probability fitted values from linear, poisson and logistic regression with robust standard errors are computed for each mother discrete age in years at enrollment. Then a scatter plot is produced is used to display the 3 fits.

Results: Refer to figure 1. The linear regression gives a line whereas the logistic and poisson regressions produces each a curve with upper concavity. The 3 estimates are quite similar especially at the center of the age (around 25 years) and then at extremes linear regression departs from both poisson a logistic regression. Both poisson and logistic regression give much closer results between them, than to linear regression.

q7a.tif

Figure 1 – Fitted values from linear, poisson and logistic regression for SGA from mother age at enrollment.

7b) Methods: Probability fitted values from linear, poisson and logistic regression with robust standard errors are computed for each mother continuous age in years at enrollment. Then a line plot is produced is used to display the 3 fits.

Results: Refer to figure 2. The 3 estimates are quite similar especially at the center of the age (around 25 years) and then at extremes linear regression departs from both poisson a logistic regression. Both poisson and logistic regression give much closer results between them, than to linear regression.

q7b.tif

Figure 2 – Fitted values from linear, poisson and logistic regression for SGA from mother age at enrollment.

Score: 8; -2 for not correctly showing sample proportions of SGA within each unique age.

8.a) Methods: a logistic regression with robust standard errors is used to compute the association between with the odds for small for gestational age (SGA) and logarithmically transformed maternal age at enrollment. 95% confidence intervals and p-values from Wald-test are reported. The significance level is set to 5%.

Results: According to the regression, for each 10% higher maternal age there is a tendency of 8.7% relative lower odds. It wouldn’t unusual to find this relative lower odd between 0 and 16.7%. We don’t reject null hypothesis of no association, p-value 0.052.

Score: 9; -1 for the wrong p value; Even though the key used different log base, the p value should be the same.

8b) The model on 6c just looks for a first order relationship between odds for SGA and maternal age at enrollment whereas the model on 8a seeks for a non linear relationship. Both models have a quite the same Wald test, but 8a model is includes the null value in it (so no non linear relationship, well at least with age logarithmilly transformed). Also the fitted values from both models as shown on the figure below are quite similar. Moreover to my knowledge there is no scientific reason for SGA to be related to log of mother age. So doing 8a we add difficulties without benefit for interpretation whereas 6c is simpler and easier to interpret.

Score: 5;

q8b.tif

**APPENDIX**

**Stata do file**

/\* read the dataset

set more off

clear

infile mcode ht age sga parity smoker bweight sex gesage ///

using pregout.txt

drop in 1

\*/

// Question 1

recode smoker (2 = 0)

recode parity (4 5 6 = 4), gen(parity\_cat)

recode sex (1 = 1) (2 = 0), gen(male)

sum ht age sga parity smoker bweight male gesage

ctabstat ht age sga parity smoker bweight male gesage, stat(mean sd min max n) by(sga) ///

col(stat) long format(%5.2f)

mvpatterns ht age sga parity smoker bweight sex gesage // There are missings

tab parity\_cat sga, miss col nofreq row

tab smoker sga , nofreq col row

tab male sga , col row

g nonsmoker = ! smoker

g nonsga = ! sga

// Question 2

// a

logit sga smoker, or robust

test smoker

// b on non smokers

lincom \_cons, eform

di %6.4f r(estimate)/(1 + r(estimate))

// on smokers

lincom \_cons + smoker, eform

di %6.4f r(estimate)/(1 + r(estimate))

//

// Question 3

// a

regress sga smoker, robust

test smoker

// Question 4

poisson sga smoker, robust irr

test smoker

// 4cii

poisson nonsga smoker, robust irr

test smoker

// 4ciii

poisson nonsga nonsmoker, robust irr

test nonsmoker

// 5

prtest sga , by(smoker)

// 6

// 6a

regress sga age, robust

predict psga\_rd

lincom 5\*age

// 6b

poisson sga age, robust irr

predict psga\_rr

lincom 5\*age, eform

test age

// 6c

logit sga age, robust or

predict psga\_or

lincom 5\*age, eform

test age

// 6d

regress sga age, robust

lincom \_cons + 20\*age

poisson sga age, robust irr

lincom \_cons + 20\*age, eform

logit sga age, robust or

lincom \_cons + 20\*age, or

di r(estimate)/(r(estimate) + 1)

di .1484096/(1 + .1484096)

di .2491551/(1 + .2491551)

// 7a

egen psga\_rdmean = mean(psga\_rd), by(age)

egen psga\_rrmean = mean(psga\_rr), by(age)

egen psga\_ormean = mean(psga\_or), by(age)

scatter psga\_rdmean psga\_rrmean psga\_ormean age, ///

legend(order(1 "Linear regression" 2 "Poisson regression" 3 "Logistic regression") ///

cols(1)) ///

ytitle(Probability for SGA delivery) ///

xtitle(Mother age at enrollment (years)) ///

ylabel(0 (0.05) 0.2, angle(horizontal) format(%3.2f)) ///

xlabel(10(5)45)

graph export q7a.tif, replace

// 7b

twoway (line psga\_rd age, sort lwidth(thick)) (line psga\_rr age, sort lwidth(thick)) ///

(line psga\_or age, sort lwidth(thick)), ///

legend(order(1 "Linear regression" 2 "Poisson regression" 3 "Logistic regression") ///

cols(1)) ///

ytitle(Probability for SGA delivery) ///

xtitle(Mother age at enrollment (years)) ///

ylabel(0 (0.05) 0.2, angle(horizontal) format(%3.2f)) ///

xlabel(10(5)45)

graph export q7b.tif, replace

// 8a

gen logage = log(age)

logit sga logage, robust or

predict psga\_orlogage

lincom .09531018\*logage, eform

// 8b

scatter (psga\_ormean psga\_orlogage age) || (lowess sga age, bwidth(0.95)), ///

legend(order(1 "Logistic regression untransformed age (6c)" 2 "Logistic regression transformed age (8a)" 3 "Lowess based") ///

cols(1)) ///

ytitle(Probability for SGA delivery) ///

xtitle(Mother age at enrollment (years)) ///

ylabel(0 (0.05) 0.25, angle(horizontal) format(%3.2f)) ///

xlabel(10(5)45)

graph export q8b.tif, replace