Homework #3 **125/160**

02 February 2015

**Question 1 8/10**

**8/10 - discusses trends but not comparison of actual values in results section - it would be helpful to highlight specific values from the table for variables analyzed in the assignment such as range and mean of age, which we will later look at in graph form**

Methods: Summary statistics are shown in the table below. For binary variables (smoking, sex of infant) the count and proportion are shown. For continuous and ordered categorical variables (mother’s height, mother’s age, parity, birth weight, and gestational age), the mean, standard deviation, minimum, and maximum are shown.

Results: Of the 751 subjects with non-missing data on smoking status (the exposure of interest), 231 were smokers and 520 did not smoke. The mean of mother’s height, mother’s age, parity, and gestational age are similar across groups defined by smoking status. Mean birth weight was lower among mothers who smoked and there were a slightly higher proportion of male babies in the non-smoking group. There was a higher proportion of small for gestational age babies in the group of mothers who smoked. Only small numbers of subjects are missing data for any of these categories. Of the 755 subjects with non-missing data for the small gestational age variable (the outcome of interest), 105 were small for gestational age and 650 were not small for gestational age. The mean of mother’s height, mean of mother’s age, and mean parity are similar in both groups. There is a higher proportion of smokers and females in the small for gestational age group. There is also a lower mean birth weight and earlier mean gestational age in the small for gestational age group. Only small numbers of subjects are missing data for any of these variables, with the largest group of missing being the mother’s height variable among those with babies small for gestational age.

There is no strong evidence here to suggest potential confounding of an association between smoking during pregnancy and small for gestational age babies, by mother’s height, mother’s age, parity, gestational age, sex, or birth weight.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Smoked During Pregnancy** | | |
|  | **Yes (n=231)** | **No (n=520)** | **Overall (n=751)** |
| **Mother’s height (cm)** | 157 (7.19; 106-176; n=230) | 157 (6.16; 127-175; n=515) | 157 (6.49; 106-176; n=745) |
| **Mother’s age** | 25.1 (5.35; 15-42; n=231) | 24.6 (5.37; 14-43; n=520) | 24.8 (5.36; 14-43; n=751) |
| **Parity** | 1.19 (1.27; 0-6; n=231) | 1.06 (1.18; 0-6; n=520) | 1.10 (1.21; 0-6; n=751) |
| **Birth weight (g)** | 2972 (512; 1410-4550; n=231) | 3165 (534; 1035-4730; n=520) | 3106 (534; 1035-4730; n=751) |
| **Male (%)** | 48.1% (n=231) | 52.3% (n=520) | 51.0% (n=751) |
| **Gestational age (weeks)** | 39.0 (1.36; 33-43; n=230) | 39.3 (1.55; 30-44; n=520) | 39.2 (1.50; 30-44; n=750) |
| **Small for gestational age** | 19.5% (n=231) | 11.3% (n=520) | 13.8% (n=751) |
|  | **Small for Gestational Age (Below 10th Percentile)** | | |
|  | **Yes (n=105)** | **No (n=650)** | **Overall (n=755)** |
| **Mother’s height (cm)** | 155 (6.87; 142-172; n=99) | 157 (6.54; 106-176; n=650) | 157 (6.50; 106-176; n=749) |
| **Mother’s age** | 23.8 (4.90; 16-35; n=105) | 24.9 (5.45; 14-43; n=650) | 24.8 (5.39; 14-43; n=755) |
| **Parity** | 0.90 (1.11; 0-6; n=105) | 1.13 (1.23; 0-6; n=650) | 1.10 (1.21; 0-6; n=755) |
| **Smoking (%)** | 43.3% (n=104) | 28.7% (n=647) | 30.8% (n=751) |
| **Birth weight (g)** | 2231 (412; 1035-3780; n=104) | 3246 (402; 2510-4730; n=647) | 3106 (534; 1035-4730; n=751) |
| **Male (%)** | 42.3% (n=104) | 52.4% (n=647) | 51.0% (n=751) |
| **Gestational age (weeks)** | 37.9 (2.20; 30-42; n=103) | 39.4 (1.24; 38-44; n=647) | 39.2 (1.50; 30-44; n=750) |

aFor binary variables, % (n) are shown. For continuous variables, mean (SD; min-max; n) are shown.

**Question 2 19/25 a. 10/10 b. 5/5 c. 4/10 - 4 pts for discussion of reparameterization and the fact that each model is scientifically the same. Points off for: no mention of exactly *how* the odds ratio is different between models, no mention of relations between slopes and intercepts between models**

(a) Methods: Logistic regression was used to compare the odds of having a baby small for gestational age between mothers who smoked during pregnancy and mothers who did not smoke during pregnancy, using the odds ratio. The 95% confidence interval and two-sided p-value are Wald-based estimates. An alpha level of 0.05 was used to determine statistical significance. Subjects missing data for smoking status were excluded from this analysis.

log(odds SGA=1) = β0 + β1 (smoker)

Results: The odds of having an infant that is small for gestational age is 1.89 times higher among mothers who smoked compared to mothers who did not smoke. With a 95% confidence interval, this observed odds ratio would not be considered unusual if the true odds are between 1.24 and 2.89 times higher in mothers who smoked compared to those who did not. Given a two-sided p value of 0.003, we reject the null hypothesis of an odds ratio equal to 1 in favor of the hypothesis that the odds of a baby of small gestational age are higher in mothers who smoke than in mothers who do not smoke.

(b) Probability is calculated as odds/(1+odds). The odds of having a baby that is small for gestational age among mothers who smoked during pregnancy is 0.242 and the probability is 0.195. The odds of small for gestational age in the non-smoking group is 0.128 and the probability is 0.113. These probabilities correspond exactly to the proportion of babies small for gestational age among smokers and non-smokers, respectively.

(c) (i) Fitting the model with the indicator NONSMOKER as the predictor instead, where mother’s who smoked=0 and mothers who did not smoke=1, is a reparameterization of the original analysis, which is a saturated model. While we would get a **different odds ratio** (and a different confidence interval), the underlying relationship in our sample will not have changed and thus the inference about the relationship is the same.

log(odds SGA=1) = β0 + β1 (non-smoker)

(ii) Had we fit a model with the indicator NOTSGA as the response instead of sga where babies small for gestational age=0 and babies not small for gestational age=1, keeping the original smoking predictor variable (smoking=1, non-smoking=0), we again are simply reparameterizing a saturated model and thus the underlying relationship between smoking and sga and the inference about that relationship are the same.

log(odds NOTSGA=1) = β0 + β1 (smoker)

(iii) Had we fit a model using both new indicator variables, NONSMOKE and NOTSGA, we get the **same odds ratio, confidence interval, and p-value** as the original model with SMOKE and SGA when we switched the predictor to nonsmoker. The underlying relationship and inference about smoking and sga are the same.

log(odds NOTSGA=1) = β0 + β1 (non-smoker)

**Question 3 20/25 a. 10/10 b. 5/5 c. 5/10 - i./iii. should explain what intercept represents and how it is different from original model ii. note the typo, but I know what you meant**

(a) Methods: Linear regression with robust standard errors was used to compare the probability of having a baby small for gestational age between mothers who smoked during pregnancy and mothers who did not smoke during pregnancy (because when the outcome is binary for simple linear regression, the expectation of SGA is equal to the probability that SGA=1, so we compare probabilities using the difference in means). The Huber-White sandwich estimator was used to compute standard errors and to estimate 95% confidence intervals and a two-sided p-value. An alpha level of 0.05 was used to determine significance. Subjects missing data for smoking status were excluded from this analysis.

Results: The probability of having a baby that is small for gestational age among mothers who smoked during pregnancy is 0.195. The probability of having a baby that is small for gestational age in the non-smoking group is 0.113. The probability of having an infant that is small for gestational age is an absolute 8.13% higher among mothers who smoked compared to mothers who did not smoke. With a 95% confidence interval, this observed difference in mean (equivalent to difference in probability in this case of a binary outcome) would not be considered unusual if the true difference is between 2.34% and 13.9%. Given a two-sided p value of 0.006, we reject the null hypothesis of a difference in means equal to zero in favor of the hypothesis that the probability of a baby of small gestational age is higher in mothers who smoke than in mothers who do not smoke.

(b) The odds in each smoking group can be calculated using the probabilities calculated from linear regression in the formula: odds = probability/(1-probability).

The probability of having a baby that is small for gestational age among mothers who smoked during pregnancy is 0.195 and the odds are 0.242. The probability of having a baby that is small for gestational age in the non-smoking group is 0.113 and the odds are 0.128. These probabilities correspond exactly to the proportion of babies small for gestational age among smokers and non-smokers, respectively.

(c) (i) The reparameterization of the saturated model by converting the predictor to NONSMOKER does not change the inference about the underlying relationship. In this model the **estimate of the slope and corresponding confidence intervals are of the same magnitude but opposite sign** compared to the original analysis. The **estimate of the intercept is different**, but the p-value is the same as the original analysis.

(ii) When the outcome is converted to NOTSGA and using the original SMOKER predictor, the **slope and confidence intervals are also of the same magnitude** but of **opposite slope** compared to the original analysis. The estimate of the intercept is different, but the p-value is the same as the original analysis. Again, the reparameterization of a saturated model does not change the inference about the relationship.

(iii) When both the exposure and the outcome have been converted to NONSMOKER and NOTSGA, the **slope and confidence intervals are the same (both in sign and magnitude)** as the original analysis. The p-value is also the same, though the **intercept is different**. Again, the reparameterization of the saturated model does not change the inference about the relationship.

**Question 4. 19/25 a. 9/10 - p-val does not agree b. 5/5 c. 5/10 Are robust SE used? Answer makes no reference to differences between poisson regression and other models, nor what the slope and intercepts specifically refer to. Poisson regression estimates depend on how you model the variables even though the interpretation is the same scientifically.**

(a) Methods: Poisson regression was used to model the ratio of probabilities of small for gestational age babies among mothers who smoked and mothers who did not smoke. Wald-based 95% confidence intervals and two-sided p-values are reported. An alpha level of 0.05 was used to determine statistical significance. Subjects missing data on smoking status were excluded from this analysis.

log(Pr(SGA=1)) = β0 + β1 (smoker)

Results: The probability of a baby small for gestational age is **1.72 times higher** among mothers who smoked compared to mothers who did not smoke. **The rate** (equivalent to the probability in this case) of small for gestational age babies among non-smokers the rate is 0.113 and among smokers is 0.195. Based on the 95% confidence interval, this estimated ratio of probabilities would not be considered unusual if the true probability is between 1.20 and 2.45 higher among mothers who smoke. A p-value of **0.006** leads us to reject the null hypothesis that the ratio of probabilities is equal to 1 in favor of the alternative hypothesis that the probability of a small gestational age baby is higher among mothers who smoke compared to those who do not smoke (ratio is greater than 1).

(b) The rate is the same as the probability of small for gestational age babies, which is exactly equal to the probability of small for gestational age babies. The odds can be calculated using: odds=probability/(1-probability). The probability of having a baby that is small for gestational age among mothers who smoked during pregnancy is 0.195 and the odds are 0.242. The probability of having a baby that is small for gestational age in the non-smoking group is 0.113 and the odds are 0.128. These probabilities correspond exactly to the proportion of babies small for gestational age among smokers and non-smokers, respectively.

(c) (i) Using SGA as the response and NONSMOKER as the predictor, the estimate of the ratio of probabilities **(0.582) is the inverse of the ratio calculated in part a**. The reparaterization of this saturated model does not change the inference about the relationship between smoking and sga.

log(Pr(SGA=1)) = β0 + β1 (non-smoker)

(ii) Using NOTSGA on the original predictor SMOKER, **estimates the ratio of probabilities** of a baby not small for gestational age which is different than the previous reparameterization of this model. However, the underlying relationship between smoking and sga does not change and thus the inference remains the same.

log(Pr(NOTSGA=1)) = β0 + β1 (smoker)

(iii) Using NOTSGA with the predictor NONSMOKER is a reparameterization of the saturated model. The **exponentiated intercept** is exactly equal to the proportion of smokers who did not have a small for gestational age baby. While the estimates may change, the underlying relationship does not and thus the inference does not change.

log(Pr(NOTSGA=1)) = β0 + β1 (non-smoker)

**Question 5**. **5/10 - Answer addresses how t test is similar to linear regression and that we are testing probabilities, but no mention of chi squared test.**

A t-test allowing for the possibility of unequal variances could also be used to make a simple two sample comparison of SGA by smoking status. The t-test would estimate the same probabilities because we have a saturated model. The magnitude of the confidence intervals are slightly different due to the standard error assumption (allowing for unequal variances).

**Question 6. 36/40 a. 10/10 b. 6/10 - methods ok, but numbers in results do not agree with key - 0.0338\*5 = 0.169 which does agree with the key. However, your p value indicates that we should not reject the null, but the key has a p value of 0.046. The p value should be the same for a difference of 1 yr or 5 yrs of age. c. 10/10 numbers different based on 1 yr difference, but trend appears similar d. 10/10 - nice table - easy to read/understand answer**

(a) Methods: Linear regression with robust standard errors (computed using the Huber-White sandwich estimator) was used to compare the probability of SGA infants across groups defined by maternal age. The robust standard errors were calculated using the Huber-White sandwich estimator. An alpha level of 0.05 was used to determine statistical significance. All subjects had non-missing data for sga and maternal age and were included in analysis.

E(SGA|Age) = β0 + β1 (age)

Results: For the 755 subjects in this dataset, the estimated decrease in probability of SGA with each **10 year increase** in maternal age is **4.52%**. Based on the 95% confidence interval, this decrease would not be considered unusual if the true decrease is between **0.286% and 8.74%**. The intercept does not have a scientifically meaningful interpretation in this case. A p-value of **0.036** suggests that we reject the null hypothesis of no difference in probability of sga with increasing maternal age in favor of the alternative hypothesis.

(b) Methods: Poisson regression was used to estimate the ratio of probabilities of sga across groups defined by maternal age. 95% confidence intervals and two-sided p-values were based on Wald statistics. An alpha level of 0.05 was used to determine statistical significance. All subjects had non-missing data for sga and maternal age and were included in analysis.

log[E(SGA|Age)] = β0 + β1 (age)

Results: For each 1 year increase in age, the probability of sga decreases by **3.38%**. Based on the 95% confidence interval, this estimate would not be judged unusual if the true value were between a decrease of 6.96% and an increase in the probability of sga of 3.32%. A p-value of **0.074** suggests that the null hypothesis cannot be rejected.

(c) Methods: Logistic regression was used to compare the ratio of odds of sga across groups defined by maternal age. 95% confidence intervals and two-sided p-values are based on Wald statistics. An alpha level of 0.05 was used to determine statistical significance. All subjects had non-missing data for sga and maternal age and were included in analysis.

log(odds[SGA=1]) = β0 + β1 (age)

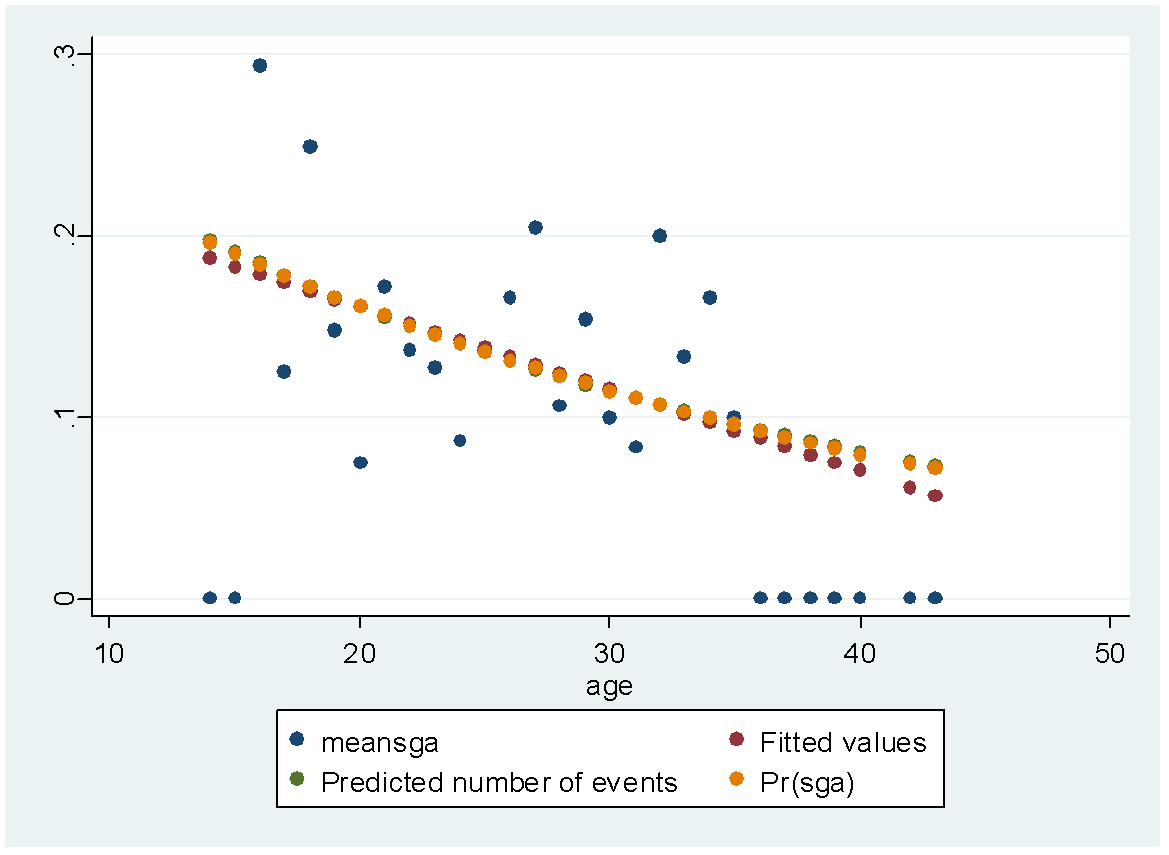
Results: For each 1 year increase in age, the odds of sga decrease by 3.90%. Based on the 95% confidence interval, this estimate would not be judged unusual if the odds in a group one year older is between 7.72% lower and 0.0764% higher than the younger group. A p-value of **0.054** suggests we cannot reject the null hypothesis in favor of the alternative.

(d) Fitted values from each regression were all similar to each other, though they were all higher than the sample probability. The sample mean of 20-year olds is much lower than the sample means in age groups surrounding 20. Because age is a continuous variable with a large number of values, the regression models are not saturated and therefore do not exactly predict the sample mean.

|  |  |
| --- | --- |
| **Model** | **Fitted value at age=20** |
| Sample probability | 7.50% |
| Linear | 16.07% |
| Poisson | 16.13% |
| Logistic | 16.13% |

**Question 7. 8/10 - would be best to label the fitted values according to either regression type, or interpretation of regression output (Risk Ratio, Odds Ratio, etc). I can’t see a y axis label**

The plot below shows both the sample probabilities and the fitted values from each regression model; the logistic and poisson regression models are slightly curvilinear. Because the regression models are not saturated they do not predict the sample probabilities exactly.



**Question 8**. **15/15** **a. 10/10 b. 5/5 - A 10-fold increase is a bit extreme - what about a 2-fold increase, or a 50% increase (as in the key)?**

(a) Methods: Logistic regression was used to evaluate the odds ratio of a baby small for gestational age across groups defined by a 10-fold difference maternal age. Maternal age was log-transformed using base 10. All subjects had non-missing data for both sga and maternal age and thus no subjects were excluded from analysis. An alpha of 0.05 was used to determine statistical significance with a two-sided p-value.

log[odds(SGA=1)] = β0 + β1 log10(age)

Results: The odds of sga is a relative 88.9% decrease for a 10-fold increase in maternal age. This odds ratio would not be judged unusual if the true value for a 10-fold increase in maternal age is between a 98.9% decrease and an 8.11% increase. A p-value of P=0.058 suggests we cannot reject the null hypothesis that the odds ratio is equal to 1.

(b) If we thought that a k-fold increase in age (rather than an absolute increase in age by k) was the relevant comparison scientifically, then we might have chosen to use the log of maternal age. However, in this case reporting a change for an absolute difference in age is much more reasonable. Given the range of ages over which women bear children, discussion of a 10-fold increase in age is not scientifically meaningful.