**Biost 518: Applied Biostatistics II**

**Biost 515: Biostatistics II**

Emerson, Winter 2015

**Homework #3**

January 23, 2015

**Written problems:** To be submitted as a MS-Word compatible file to the class Catalyst dropbox by 9:30 am on Monday, February 2, 2014. See the instructions for peer grading of the homework that are posted on the web pages.

*On this (as all homeworks) Stata / R code and unedited Stata / R output is* ***TOTALLY*** *unacceptable. Instead, prepare a table of statistics gleaned from the Stata output. The table should be appropriate for inclusion in a scientific report, with all statistics rounded to a reasonable number of significant digits. (I am interested in how statistics are used to answer the scientific question.)*

***Unless explicitly told otherwise in the statement of the problem, in all problems requesting “statistical analyses” (either descriptive or inferential), you should present both***

* ***Methods: A brief sentence or paragraph describing the statistical methods you used. This should be using wording suitable for a scientific journal, though it might be a little more detailed. A reader should be able to reproduce your analysis. DO NOT PROVIDE Stata OR R CODE.***
* ***Inference: A paragraph providing full statistical inference in answer to the question. Please see the supplementary document relating to “Reporting Associations” for details.***

This homework considers pregnancy outcomes in an observational study of women attending a prenatal clinic in South Africa. Questions in this homework focus most closely on association with delivery of babies that are small for gestational age (SGA). The data can be found on the class web page (follow the link to Datasets) in the file labeled pregout.txt (you will not need any of the longitudinal measurements in the file preglong.txt). Documentation is in the file pregnancy.pdf.

1. Provide suitable descriptive statistics relevant to this analysis.

**Answer:**

***Methods***: Descriptive statistics were presented based on small for gestational age (SGA) groups. Summary statistics (mean, standard deviation, minimum and maximum) were used for quantitative variables like ht( mother’s height in cm), age(mother’s age at enrollment in years), bweight(birthweight of infant in grams), gesage(gestational age at delivery in weeks). For binary variables like smoker (mothers’ smoking status) and sex (infant’s sex), percentages were presented. Parity(number of prior pregnancies) was also treated like a categorical variable here and percentages were shown. Participants missing data for any variables were excluded only from the analyses involving those variables.

***Results:*** Descriptive statistics of mothers and infants characteristics were shown in Table 1. Data is available on 755 participants with SGA information. 13.91% of their babies were considered small for corresponding gestational age category.

For mothers’ characteristics, participants with available height or age measurements, mothers who had a baby small for gestational age tended to be shorter and younger (mean height 154.6 cm, mean age at enrollment 23.85 yrs old) than those mothers who didn’t had babies small for gestational age(mean height 157.0 cm,mean age at enrollment 24.94 yrs old) . We also found a higher proportion of smokers among mothers who had SAG babies (43.27%) than those who had non-SAG babies(28.75%). Regarding to parity(number of prior deliveries), no obvious trend were detected between mothers who had SAG babies and those who didn’t.

For infants characteristics: infants who were considered as small for corresponding gestational age had a lower mean birthweight(2231 g) and fewer mean gestational age at delivery (37.92 weeks) than non-SAG babies( mean birthweight=3246 g , mean gestational age at delivery=39.38 weeks). There was also a smaller proportion of males in babies who were small for gestational age than those who were not (42.31% vs 52.40%)

**Table 1.** Descriptive statistics of mothers and infants by SAG categories

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  | **Small for gestational age categories** | | | | | | | | |  |  |  |  |  |
| **Characteristics** | |  | Yes | | | |  | No | | | |  | Total | | | |
| Total (N) | |  | 105(13.91%) | | | |  | 650(86.09%) | | | |  | 755(100%) | | | |
| **Mother** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Height(cm)＊ | |  | 99(154.6;5.873;  142-172) | | | |  | 650(157.0;6.538;  106-176) | | | |  | 749(156.7;6.504;  106-176) | | | |
| Age(yrs)＊ | |  | 105(23.85;4.899;  16-35) | | | |  | 650(24.94;5.449;  14-43) | | | |  | 755(24.79;5.386;  14-43) | | | |
| Smoker(%) | |  | 43.27%(45 of 104) | | | |  | 28.75%(186 of 647) | | | |  | 30.76%(231 of 751) | | | |
| Parity(%) | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | 0 |  | 46.67%(49 of 105) | | | |  | 37.54%(244 of 650) | | | |  | 38.81%(293 of 755) | | | |
|  | 1 |  | 30.48%(32 of 105) | | | |  | 32.00%(208 of 650) | | | |  | 31.79%(240 of 755) | | | |
|  | 2 |  | 13.33%(14 of 105) | | | |  | 18.31%(119 of 650) | | | |  | 17.62%(133 of 755) | | | |
|  | 3 |  | 7.62%(8 of 105) | | | |  | 6.77%(44 of 650) | | | |  | 6.89%(52 of 755) | | | |
|  | 4 |  | 0.95%(1 of 105) | | | |  | 3.38%(22 of 650) | | | |  | 3.05%(23 of 755) | | | |
|  | 5 |  | 0%(0 of 105) | | | |  | 1.23%(8 of 650) | | | |  | 1.06%(8 of 755) | | | |
|  | 6 |  | 0.95%(1 of 105) | | | |  | 0.77%(5 of 650) | | | |  | 0.79%(6 of 755) | | | |
| **Infant** | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Birthweight (g)＊ | |  | 104(2231;411.6;  1035-3780) | | | |  | 647(3246;402.1;  2510-4730) | | | |  | 751(3106;543.5;  1035-4730) | | | |
| Gestational age at delivery (weeks)＊ | |  | 103(37.92;2.204;  30-42) | | | |  | 647(39.38;1.245;  38-44) | | | |  | 750(39.18;1.500;  30-44) | | | |
| Male(%) | |  | 42.31%(44 of 104) | | | |  | 52.40%(339 of 647) | | | |  | 51%(383 of 751) | | | |

**＊Descriptive statistics: number of participants(mean;standard deviation;minimum-maximum)**

1. Perform a statistical regression analysis evaluating an association between the odds of delivery of infants who were small for gestational age (SGA) and maternal smoking behavior. (Only give a formal report of the inference where asked to.)
   1. Give full inference regarding the association between SGA and maternal smoking.

**Answer:**

***Methods:*** Participants with missing values in either SGA or smoking behavior were excluded from analysis. The odds of delivery of SGA infants were compared between mothers with and without maternal smoking. A logistic regression model of binary variable (SGA) as the response and indicator of maternal smoking status(Smoker) as predictor was fitted.Point estimates of the association were based on the slope parameter from the logistic regression analyses, a 95% confidence interval and a two sides P value were computed using Wald statistics from the regression estimates and corresponding standard error.

***Results:***From a logistic regression analysis of 751 participants, We estimate that the odds of having SGA baby is 89.04% higher in the smokers than in the nonsmokers. A 95% confidence interval suggests that this observation is not be unusual if mothers who smoke have odds of getting SAG babies that was anywhere from 23.76% to 188.75% higher than the nonsmokers. The two-sided p-value is 0.003, so this observation is statistically significant at a 0.05 level of significance. We reject the null hypothesis that the odds of getting SGA babies is not associated with mothers’ maternal smoking status.

* 1. Use the regression model parameter estimates to provide estimates of both the odds and the probability of delivering a SGA infant separately for smokers and nonsmokers. How do these estimates compare with simple descriptive statistics as you might have reported in problem 1. Explain any differences or similarities.

**Answer:**

Log(odds of SGA)= -2.055861+0.636778\*Smoker

Odds of SGA=e(-2.055861+0.636778\*Smoker)

Probability of SGA =[e(-2.055861) \*e(0.636778\*Smoker)]/(1+e(-2.055861) \*e(0.636778\*Smoker))

Odds of SGA for nonsmokers = 0.12798; probability of SGA for nonsmokers = 0.1135.

Odds of SGA for smokers = 0.2419; probability of SGA for smokers = 0.1948.

**In problem 1 we have:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  | **Small for gestational age categories** | | |  |  |  |  |  |
| **Characteristics** | |  | Yes |  | No |  | Total | | | |
| Total (N) | |  | 105(13.91%) |  | 650(86.09%) |  | 755(100%) | | | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Smoker(%) |  | 43.27%(**45** of 104) |  | 28.75%(186 of 647) |  | 30.76%(**231** of 751) |

**Here in problem one, we presented among smokers, probability of SGA=45/231=0.1948 , odds of SGA=0.1948/(1-0.1948)=0.2419**

**Among nonsmokers: probability of SGA=（104-45）/(751-231)=0.1135**

**Odds of SGA=0.1135/(1-0.1135)=0.12799**

**Because this is a saturated model fitted, the estimated odds and probability of event in each group agrees exactly with the sample odds.**

* 1. There were actually four regression analyses that could have been used to answer this question. I am betting that all students would have fit a regression model with SGA as response and the indicator of maternal smoking as the predictor. Presuming that you did indeed fit that model, explain the similarities and differences between the estimates and inference you would have obtained for the following three additional models (You do not need to run these analyses, if you can tell me how they differ without doing so. It is of course okay to run the analyses if it will help you recognize the more general principles.):
     1. You create an indicator NONSMOKER that the mother was a nonsmoker, and you fit a logistic regression model of response SGA on predictor NONSMOKER.

**Answers:** The estimated intercepts in the two regressions are different; the first is the log odds of SGA for nonsmokers, in the new model is the log odds of SGA for smokers. The slopes of two regressions models are the same in absolute values, but their signs are the opposite.

* + 1. You create an indicator NOTSGA that the infant was not small for gestational age, and you fit a logistic regression model of response NOTSGA on predictor SMOKER.

**Answers**: The absolute values of intercepts in the two regressions are same, but their signs are the opposite. the first is the log odds of SGA for nonsmokers, in the new model is the log odds of NOTSGA for nonsmokers, when exponentiated, the latter is the reciprocal of the first one. The absolute values of slopes in the two regressions are same, but their signs are the opposite. .

* + 1. You fit a regression model of response NOTSGA on predictor NONSMOKER.

**Answers**: The estimated intercepts in the two regressions are different; The first one is the log odds of SGA for nonsmokers and the latter one is the log odds of NOTSGA for smokers. The slopes in these two regressions would be the same in absolute value and direction.

1. Repeat problem 2, except consider a statistical regression analysis evaluating an association between the odds of delivery of infants who were small for gestational age (SGA) and maternal smoking behavior by evaluating the difference in probabilities for SGA across smoking groups.

**Answer**

1. ***Methods***: Participants with missing values in either SGA or smoking behavior were excluded from analysis. The probabilities for SGA were compared across groups with and without maternal smoking behavior.A linear regression with robust error that allows for heteroscedasticity was adopted to test difference in the probabilities. Two-sided p-value and a 95% confidence interval were computed for the difference in the population probabilities for SGA based the Wald statistics.

***Results:*** Data were available for 751 participants, among whom 231 (30.76%) were smokers and 520 (69.24%) were nonsmokers. The proportion of SGA among the smokers was 19.48%, and that 11.35% among nonsmokers . A 95% confidence interval suggests that this difference in proportions (point estimate of 8.134% )between the two groups would not be unusual if the true difference in probability of SGA among smokers is anywhere from 2.33% to 13.94% higher than that of nonsmokers. The two-sided p-value is 0.006.Thus this observation is statistically significant at a 0.05 level of significance. We reject the null hypothesis that the probability for SGA is the same for mothers who were smokers and who were nonsmokers.

1. **Answers**:

Odds of SGA for nonsmokers = 0.12799; probability of SGA for nonsmokers = 0.1135.

Odds of SGA for smokers = 0.2419; probability of SGA for smokers = 0.1948.

Because this is a saturated model fitted, the estimated odds and probability of event in each group agrees exactly with the sample odds and probabilities in problem 1. And the rate difference in problem 1: 0.1948-0.1135=0.0813, agreeing with our rate difference point estimate from linear regression exactly.

1. **Answers**:
   * 1. You create an indicator NONSMOKER that the mother was a nonsmoker, and you fit a logistic regression model of response SGA on predictor NONSMOKER.

The estimated intercepts in the two regressions are different; the first is the probability of SGA for nonsmokers, in the new model is the probability of SGA for smokers. The slopes of two regressions models are the same in absolute values, but their signs are the opposite.

* + 1. You create an indicator NOTSGA that the infant was not small for gestational age, and you fit a logistic regression model of response NOTSGA on predictor SMOKER.

The estimated intercepts in the two regressions are different; the first is the probability of SGA for nonsmokers, in the new model is the probability of NOTSGA for nonsmokers, but they add up to 1. The slopes of two regressions models are the same in absolute values, but their signs are the opposite.

* + 1. You fit a regression model of response NOTSGA on predictor NONSMOKER.

The estimated intercepts in the two regressions are different; the first is the probability of SGA for nonsmokers, in the new model is the probability of NOTSGA for smokers. The slopes of two regressions models are the exactly same.

1. Repeat problem 2, except consider a statistical regression analysis evaluating an association between the odds of delivery of infants who were small for gestational age (SGA) and maternal smoking behavior by evaluating the ratio of probabilities for SGA across smoking groups.

**Answer:**

1. **Methods**: .Participants with missing values in either SGA or smoking behavior were excluded from analysis. The rates ratios for SGA were compared across groups with and without maternal smoking behavior.A poisson regression with robust error that allows for heteroscedasticity was adopted to test difference in the rate ratios. 95% confidence intervals and a two sides P value were computed using Wald statistics from the regression estimates and corresponding standard errors

**Results**: Data were available for 751 participants, among whom 231 (30.76%) were smokers and 520 (69.24%) were nonsmokers. We estimate from the Poisson regression analysis that the rate for SGA is 71.69% higher in smokers than in nonsmokers. A 95% confidence interval suggests that our data would not be unusual if the rate for SGA among smokers is anywhere from 20.19% to 145.07% higher than that in nonsmokers. The two-sided p-value is 0.003, the result is statistically significant at a 0.05 level of significance. So we could reject the null hypothesis that the rate of SGA is not associated with maternal smoking status.

b. Odds of SGA for nonsmokers = 0.12798; probability of SGA for nonsmokers = 0.1135.

Odds of SGA for smokers = 0.2419; probability of SGA for smokers = 0.1948.

Because this is a saturated model fitted, the estimated odds and probability of event in each group agrees exactly with the sample odds and probabilities in problem 1. And the rate ratio in problem 1: 0.1948/0.1135=1.716, agreeing with our rate ratio point estimate from poisson regression exactly.

c . **Answers**:

* + 1. You create an indicator NONSMOKER that the mother was a nonsmoker, and you fit a logistic regression model of response SGA on predictor NONSMOKER.

The estimated intercepts in the two regressions are different; the first is the rate of SGA for nonsmokers, in the new model is the rate of SGA for smokers. The slopes of two regressions models are the same in absolute values, but their signs are the opposite.

* + 1. You create an indicator NOTSGA that the infant was not small for gestational age, and you fit a logistic regression model of response NOTSGA on predictor SMOKER.

The estimated intercepts in the two regressions are different; the first is the rate of SGA for nonsmokers, in the new model is the rate of NOTSGA for nonsmokers. The slopes of two regressions models are different in values.

* + 1. You fit a regression model of response NOTSGA on predictor NONSMOKER.

The estimated intercepts in the two regressions are different; the first is the rate of SGA for nonsmokers, in the new model is the rate of NOTSGA for smokers. The slopes of two regressions models are also different in values.

1. How do the analyses performed in problems 2-4 compare to that that would be obtained in a simple two sample comparison of SGA by smoking status (i.e., using methods covered in Biost 517/514.) Explicitly mention where they would be similar or different?

**Answer**: We could perform a two-sample test of proportions of SAG among mothers who were smokers and mothers who were not.

***Methods***: Participants with missing values in either SGA or smoking behavior were excluded from analysis. The proportions of SGA were compared across groups with and without maternal smoking behavior. We used t test that allows for the possibility of unequal variances to test difference in the means(the proportion of SGA). A two-sided p-value and 95% confidence interval were computed using a sample variance estimates from each group.

***Results***: Data were available for 751 participants, among whom 231 (30.76%) were smokers and 520 (69.24%) were nonsmokers. The proportion of SGA among the 231 smokers was 0.1948, and that among the nonsmokers was 0.1135. A 95% confidence interval that allows for unequal variances suggests that this difference of 0.0813 in proportions(point estimate) would not be unusual if the true proportion of SGA among smokers is anywhere between 0.0231 and 0.1395 higher than that among nonsmokers. Based on a t test allowing for unequal variances, this observation is statistically significant at a 0.05 level of significance (two-sided p-value = 0.003). So we can reject the null hypothesis that the proportion of SGA is not associated with the maternal smoking status.

***Comparisons***: The two sample test for proportions using t test is exactly same as those tests performed in problems 2-4 for estimated parameters(Wald based). We can get the odds ratio , rate difference and rate ratio,of SGA between smokers and nonsmokers from this simple two sample test for proportions, which will agree with the exponentiation of estimated parameters(intercepts and slopes) from logistic regression in problem 2 and from poisson regression in problem 4, and the estimated parameters of linear regression in problem 3(not exponentiated).

1. Perform a regression analysis of the distribution of the prevalence of SGA infants across groups defined by the continuous measure of maternal age. In all cases we want formal inference. (Note: In problem 7, I am asking you to plot the estimated probabilities of SGA infants from each of these regression models. Hence, you will want to make sure you estimate those fitted values following each regression.)
   1. Evaluate associations using risk difference (RD: difference in probabilities).

**Methods**: Participants with missing values in either SGA or age were excluded from analysis. The probabilities for SGA were compared across groups defined by the continuous measure of maternal age.A linear regression with robust error that allowing for heteroscedasticity was adopted to test difference in the probabilities. Two-sided p-value and a 95% confidence interval were computed for the difference in the population probabilities for SGA based the Wald statistics.

**Results**: Data were available on 755 participants, with a mean age at 24.79 years old. The probability for SGA is estimated to be 0.4515% lower for each year increase in age between two groups. A 95% confidence interval suggests that this difference in probability for SAG (point estimate -0.4515%) between two groups differing in 1 year age would not be unusual if the true difference in probability for SGA in the older group is anywhere from 0.8745% to 0.02856% lower than that in the younger group . The two-sided p-value is 0.036, so this observation is statistically significant at a 0.05 level of significance. We can reject the null hypothesis that the probability for SGA is not associated with the maternal age.

* 1. Evaluate associations between risk ratio (RR: ratios of probabilities).

**Methods**: Participants with missing values in either SGA or age were excluded from analysis. The rates ratios for SGA were compared across groups defined by the continuous measure of maternal age. A poisson regression with robust error that allows for heteroscedasticity was adopted to test difference in the rate ratios. 95% confidence intervals and a two sides P value were computed using Wald statistics from the regression estimates and corresponding standard errors

**Results**: Data were available on 755 participants, with a mean age at 24.79 years old. We estimated from a Poisson regression that the probability for SGA is a relative 3.384% lower for each year increase in age between two groups. A 95% confidence interval suggests that our data would not be unusual if the probability for SGA in the older group is anywhere from 0.06074% to 6.596% lower than that in the younger group with one year difference in age (95% confidence interval for risk ratio of SGA is 0.93403 to 0.99939). The two-sided p-value is 0.046, so this observation is statistically significant at a 0.05 level of significance. Therefore we reject the null hypothesis that the probability for SGA is not associated with the maternal age.

* 1. Evaluate associations using odds ratio (OR: ratios of odds)

**Methods**: Participants with missing values in either SGA or age were excluded from analysis. The odds of delivery of SGA infants were compared across groups defined by the continuous measure of maternal age. A logistic regression model of binary variable (SGA) as the response and continuous variable (age) as predictor was fitted. Point estimates of the association were based on the slope parameter from the logistic regression analyses, a 95% confidence interval and a two sides P value were computed using Wald statistics from the regression estimates and corresponding standard error.

**Results:** Data were available on 755 participants, with a mean age at 24.79 years old. We estimated from a logistic regression that the odds for SGA is a relative 3.900% lower for each year increase in age between two groups (odds ratio = 0.9610). A 95% confidence interval suggests that our data would not be unusual if the odds for SGA in the older group is anywhere from 0.07617% higher to 7.718% lower than that in the younger group with one year difference in age (95% confidence interval for odds ratio of SGA is 0.922820 to 1.00076169). The two-sided p-value is 0.054, so this observation is not statistically significant at a 0.05 level of significance. Therefore we cannot reject the null hypothesis that the odds of SGA is not associated with the maternal age.

* 1. Using the regression parameter estimates from each of these regressions, provide an estimate of the probability that a 20 year old mother would have a SGA infant. Explain any similarities or differences these estimates might have when compared to the sample proportion of SGA infants among 20 year olds.

**Answers**:

1. Using risk difference (RD, linear regression), the estimated probability that a 20 year olds mother would have a SGA infant is 0.1607.
2. Using risk ratio (RR, linear regression), the estimated probability that a 20 year olds mother would have a SGA infant is 0.1613.
3. Using risk difference (RD, linear regression), the estimated probability that a 20 year olds mother would have a SGA infant is 0.1613.
4. Comparison: The sample proportion of SGA infants among 20 year olds was 0.075(3/40). The estimates from the three regressions above are similar to each other, but quite different from the sample proportion. This is so because we used a continuous measurement for age, so our regression models were not saturated , they need to borrow information between two discrete age groups with 1 year interval, that’s how we get the regression parameters (the intercept and the slope). And we would expect that the fitted probability of SGA would not agree exactly with the sample proportion.
5. Produce a plot of the estimated probability of an SGA infant by age as derived by each of the following methods. Comment on the similarity and difference among the various fitted values form the various analyses performed in problem 6. (Note that Stata allows you to specify multiple Y variables for a single X variable: scatter y1 y2 y3 y4 age)
   1. Sample proportions within each unique age: This can be obtained in Stata using the command egen *varname*= mean(sga), by(age).

**Answer:**

**Methods**: The plot of sample proportions within each unique age for SGA is presented as below.



* 1. Estimated probabilities for each age in the data as derived from each of the regression analyses. In Stata, this can be obtained using the simple “post-estimation” command: predict *varname.* (But use a different variable name for each fitted value.)
     1. After performing a linear regression, the default action of the “predict” function is to create a variable that contains the estimated “linear predictor”, which corresponds to the regression based estimate of the mean. With a binary response variable, the mean response is the proportion.

**Methods:** The plot of estimated probabilities of SGA for each age derived from the linear regression is shown as below. Sample proportions within each unique age for SGA is as presented in the same figure.



* + 1. After performing a Poisson regression, the default action of the “predict” function is to create a variable that contains the exponentiated estimated “linear predictor”, which corresponds to the regression based estimate of the mean. With a binary response variable, the mean response is the proportion. (The linear predictor in Poisson regression corresponds to the log “rate”, because Poisson regression uses a log link function.

**Methods**: The plot of estimated “rate” of SGA for each age derived from Poisson regression is is shown as below. Sample proportions within each unique age for SGA is as presented in the same figure.



* + 1. In logistic regression, the estimated “linear predictor” corresponds to the log odds. Exponentiating that would correspond to the odds. By default, Stata figures that you would really rather have the estimated probability, which is computed as prob = odds / (1 + odds). So, after performing a logistic regression, the default action of the “predict” function is to create a variable that contains the the regression based estimate of the mean.

**Methods**:The plot of estimated probabilities of SGAfor each age derived from logistic regression is presented as below. Sample proportions within each unique age for SGA and as presented in the same figure. Sample proportions within each unique age for SGA and the other two fitted lines( from linear regression and poisson regression) are also presented in the same figure

.



**Comparisons:** All of the linear regression, Poisson regression and logistic regression provide similar fitted values across age groups. But because these three regression models are all unsaturated models, they differ with the sample proportions at corresponding ages greatly. This is so because we used a continuous measurement for age, so our regression models need to borrow information between two discrete age groups with 1 year interval, that’s how we get the regression parameters (the intercept and the slope). And we would expect that the fitted probability of SGA would not agree exactly with the sample proportion. This difference agrees with the results we get in problem 6.

1. Perform a logistic regression analyses of the distribution of the prevalence of SGA infants across groups defined by the logarithmically transformed maternal age.
   1. Provide formal inference for associations using odds ratio (OR: ratios of odds) and log transformed age.

**Methods**: Participants with missing values in either SGA or age were excluded from analysis. The odds of delivery of SGA infants were compared across groups defined by the logarithmically transformed maternal age. A logistic regression model of binary variable (SGA) as the response and logarithmically transformed variable (logage) as predictor was fitted. Point estimates of the association were based on the slope parameter from the logistic regression analyses, a 95% confidence interval and a two sides P value were computed using Wald statistics from the regression estimates and corresponding standard error.

**Results**: Data were available on 755 participants, with a mean age at 24.79 years old. We estimated from the logistic regression that the odds ratio for SGA between the older group and younger group is 0.3853 (odds ratio = -0.9536) for two groups differing by 1 unit in log-transformed age. A 5-fold increase in age is estimated to get an odds ratio of SGA only 2-0.9536 = 0.2155, the odds of SGA decease by 78.45%. A 95% confidence interval suggests that our data would not be unusual if the odds for SGA in the older group is anywhere from 85. 646% lower to 3.444% higher than that in the younger group with one unit difference in log-tranformed age(0r 2.71828 years difference in age). (95% confidence interval for odds ratio of SGA is 0.1435 to 1.0344). The two-sided p-value is 0.058, so this observation is not statistically significant at a 0.05 level of significance., we can not reject the null hypothesis that the odds of SGA is not associated with the logarithmically transformed maternal age.

* 1. Why might it be reasonable or silly to have performed such an analysis rather than the analysis in problem 6c?

**Answers:** I don’t think it is reasonable to log transform age. Firstly, it is very hard to understand and interpret the log transformed data, 1 unit difference in log transformed age is not commonly used in scientific communication, not to speak in daily life. And there is no scientific ground that log transformed age will provide a better fit.

Besides, the rang of age in this sample was 14-43 years old, a histogram will show that it is not highly skewed and there is no obvious outliers. Plus, there is also no scientific ground for the assumption that age behaves multiplicatively on the risk of SGA.

Furthermore, a scatter plot will show the logrithm of age is approximatedly linear over the range of age sample. So it does not matter that much which we use. I just prefer age than logage, as it makes more sense to me.

