**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.

Methods: Variables included in the data are height, age, sga, parity, smoker, birth weight, sex and gestational age at delivery. Subjects with missing values in either of these variables are excluded from analysis when that variable is involved in the analysis. The first table below provides descriptive statistics for non-binary variables including height, age, parity, birth weight and gestational age, both overall, and separately for groups of SGA and nonSGA. Descriptive statistics provided include sample size, mean, standard deviation, minimum and maximum of corresponding variables. The second and third tables provide two-way tables of smoking status and SGA status, and sex and SGA status, respectively.

Results: For the analysis in Table 1, 11 subjects out of 755 that have missing values in at least one of the variables are excluded from analysis. Data are available on 744 subjects out of 755. Among the 744 subjects, 97 (13.04%) have delivery of babies that are small for gestational age (SGA); 647 (86.96%) have delivery of babies that are non-SGA. From Table 1, we note that mothers with delivery of babies that are SGA tend to shorter in height, younger in age, less parity; babies with SGA tend to have smaller birth weight, and gestational age. For the analysis in Table 2, 4 subjects out of 755 that have missing values in either indicator of SGA or smoking status were excluded from analysis. From Table 2, the proportion of delivery of babies with SGA among smokers (0.195) is higher than that among nonsmokers (0.113), which an odds ratio of 1.89. For the analysis in Table 3, 4 subjects out of 755 that have missing values in either indicator of SGA or sex were excluded from analysis. From Table 3, the proportion of male babies with SGA (0.115) is lower than that of female babies (0.163), with an odds ratio of 0.666.

**Table 1.**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **overall (n = 744)** | **SGA (n = 97)** | **non-SGA (n = 647)** |
| **Height (cm)** | 156.69 (6.49; 106-176) | 154.63 (5.90; 142-172) | 157.00 (6.52, 106-176) |
| **Age (yrs)** | 24.73 (5.36; 14-43) | 23.48 (4.74; 16-35) | 24.91 (5.42, 14-43) |
| **Parity** | 1.09 (1.21; 0-6) | 0.825 (1.08; 0-6) | 1.13 (1.22, 0-6) |
| **Birth weight (gms)** | 3111.19 (533.53; 1035-4730) | 2210.63 (415.53; 1035-3780) | 3246.21 (402.13; 2510-4730) |
| **Gestational age (wks)** | 39.19 (1.51; 30-44) | 37.87 (2.23; 30-42) | 39.38 (1.24; 38-44) |

\*Descriptive statistics provided are the mean (standard deviation; minimum – maximum).

**Table 2.**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **SGA status** | P(SGA=1| smoker) | Odds(SGA=1| smoker) |
| SGA (1) | non SGA (0) |
| **Smoking status** | Smoker (1) | 45 | 186 | 0.195 | 0.242 |
| Nonsmoker (0) | 59 | 461 | 0.113 | 0.128 |
| P(smoker=1|SGA) | 0.433 | 0.287 | Odds ratio: 1.89 |
| Odds(smoker=1|SGA) | 0.763 | 0.403 |

**Table 3.**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **SGA status** | P(SGA=1| sex) | Odds(SGA=1| sex) |
| SGA (1) | non SGA (0) |
| **Sex** | Boy (1) | 44 | 339 | 0.115 | 0.130 |
| Girl (2) | 60 | 308 | 0.163 | 0.195 |
| P(Sex=1|SGA) | 0.423 | 0.524 | Odds ratio: 0.666 |
| Odds(Sex=1|SGA) | 0.733 | 1.101 |

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.

Methods: Subjects with missing values in either SGA or smoking behavior were removed from analysis. The odds of delivery of infants who were small for gestational age (SGA) were compared across groups with and without maternal smoking behavior. We used a logistic regression model to test the odds of SGA as a function of binary indicator variable of maternal smoking status. The p-value and 95% confidence interval were calculated for the regression slope based on Wald statistic using maximum likelihood estimation.

Results: We have data on 751 subjects, among whom 231 (30.76%) subjects were smokers and 520 (69.24%) were nonsmokers. 104 (13.85%) infants out of 751 were observed to be SGA, and 647 (86.15%) infants out of 751 were observed to be SGA. We estimate from a logistic regression that the odds of SGA is 89.03% higher in the smokers than in the nonsmokers. A 95% confidence interval suggests that our data would not be unusual if the true odds of SGA in the smokers is anywhere from 23.60% to 189.12% higher than that of the nonsmokers. The two-sided p-value is 0.00336, so this observation is statistically significant at a 0.05 level of significance. Therefore we reject the null hypothesis that the odds of SGA is not associated with 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.

Answers:

Odds of SGA for nonsmokers = 0.128; probability of SGA for nonsmokers = 0.113.

Odds of SGA for smokers = 0.242; probability of SGA for smokers = 0.195.

These estimated values agree exactly with the sample proportion and sample odds in problem 1, since this is a *saturated* logistic regression model.

* 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: Originally we fitted the model log odds (SGA) = b0 + b1\*x, where X is an indicator variable of smoker. If we change it into an indicator variable of nonsmoker denoted by Y, then Y = 1 – X. Then log odds (SGA) = b0 + b1\*(1-y) = (b0+b1) + (-b1)\*y. Hence the new intercept is (b0+b1), and the new slope is (-b1). The estimated intercepts in the two regressions are different; however, we note that the exponentiation of estimated intercept in the first model is 0.12798, which is odds of SGA for nonsmokers, whereas that in the new model is 0.2419, which is odds of SGA for smokers. The absolute values of slopes in the two regressions are same (both 0.63678), but their signs differ. We note that the exponentiation of the estimated slope in the first regression is 1.89038, which is the estimated odds ratio, and that in the new regression is 0.52899, which is reciprocal of 1.89038. Therefore though some of the estimates differ, but the inference is essentially the same.

* + 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: Originally we fitted the model log odds (SGA) = b0 + b1\*x, where X is an indicator variable of smoker. If we change the response into notSGA, then we have log odds (1/notSGA)) = b0 + b1\*x = -b0 + (-b1)\*y. Hence the new intercept is –b0, and the new slope is (-b1). The absolute values of intercepts in the two regressions are same (both 2.0559), but their signs differ. We note that the exponentiation of the estimated intercept in the first regression is 0.12798, which is odds of SGA for nonsmokers, and that in the new regression is 7.8136, which is reciprocal of 0.12798. The absolute values of slopes in the two regressions are same (both 0.63678), but their signs differ. We note that the exponentiation of the estimated slope in the first regression is 1.89038, which is the estimated odds ratio, and that in the new regression is 0.52899, which is reciprocal of 1.89038.

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

Answers: If we change the response into notSGA, and predictor into nonsmoker, then we have log odds (1/notSGA)) = b0 + b1\*(1-x) = -(b0+b1) + b1\*y. Hence the new intercept is –(b0+b1), and the new slope is b1. The estimated intercepts in the two regressions are different; however, we note that the exponentiation of estimated intercept in the first model is 0.12798, which is odds of SGA for nonsmokers, where as that in the new model is 0.2419, which is reciprocal of odds of SGA for smokers. The estimated slopes in the two regressions are same (both 0.63678). The exponentiation of the estimated slope is the estimated odds ratio of SGA between smokers and nonsmokers.

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.
2. Methods: Subjects with missing values in either SGA or smoking behavior were removed from analysis. The probabilities for SGA were compared across groups with and without maternal smoking behavior. We used a robust linear regression that allows heteroscedasticity 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 on the Huber-White sandwich estimator using approximate normal distribution.

Results: We have data on 751 subjects, among whom 231 (30.76%) subjects were smokers and 520 (69.24%) were nonsmokers. The proportion of SGA among the 231 subjects was 19.48%, and that among the 510 nonsmokers was 11.35%. A 95% confidence interval suggests that this difference of proportions 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 among nonsmokers. The two-sided p-value is 0.0061, 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 smoking status.

1. Answers:

Odds of SGA for nonsmokers = 0.128; probability of SGA for nonsmokers = 0.113.

Odds of SGA for smokers = 0.242; probability of SGA for smokers = 0.195.

These estimated values agree exactly with the sample proportion and sample odds in problem 1, since this is a *saturated* linear regression model.

1. *Compare the four regressions.*
	* 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: Originally we fitted the model SGA = b0 + b1\*x, where X is an indicator variable of smoker. If we change it into an indicator variable of nonsmoker denoted by Y, then Y = 1 – X. Then SGA = b0 + b1\*(1-y) = (b0+b1) + (-b1)\*y. Hence the new intercept is (b0+b1), and the new slope is (-b1). The estimated intercepts in the two regressions are different; however, we note that the estimated intercept in the first regression is the estimated probability for SGA in nonsmokers, which is 11.35%, whereas that in the new model is the estimated probability for SGA in smokers, which is 19.48%. The absolute values of slopes in the two regressions are same (both 0.08134), but their signs differ. In the first regression, the estimated slope is 0.08134, which suggests that the proportion for SGA among smokers is estimated to be higher than that among nonsmokers by a difference of proportions of 8.134%. In the new regression, the estimated slope is -0.08134, which suggests that the proportion for SGA among nonsmokers is estimated to be lower than that among smokers by a difference of proportions of 8.134%.

* + 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: Originally we fitted the model SGA = b0 + b1\*x, where X is an indicator variable of smoker. If we change the response into notSGA, then we have (1 – notSGA) = b0 + b1\*x, so notSGA = (1-b0) – b1\*x. Hence the new intercept is (1–b0), and the new slope is (-b1). The estimated intercepts in the two regressions differ, but sum up to 1. The estimated intercept in the first regression is the estimated probability for SGA among nonsmokers, which is 11.35%, whereas that in the new model is the estimated probability for *non*-SGA among nonsmokers, which is 88.65%. The absolute values of slopes in the two regressions are same (both 0.08134), but their signs differ. In the first regression, the estimated slope is 0.08134, which suggests that the proportion for SGA among smokers is estimated to be *higher* than that among nonsmokers by a difference of proportions of 8.134%. In the new regression, the estimated slope is -0.08134, which suggests that the proportion *non*-SGA among smokers is estimated to be *lower* than that among nonsmokers by a difference of proportions of 8.134%.

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

Answers: If we change the response into notSGA, and predictor into nonsmoker, then we have (1- notSGA) = b0 + b1\*(1-y) = (1-b0-b1) + b1\*y. Hence the new intercept is (1-b0-b1), and the new slope is b1. The estimated intercepts in the two regressions differ. The estimated intercept in the first regression is the estimated probability for SGA among *nonsmokers*, which is 11.35%, whereas that in the new model is the estimated probability for *non*-SGA among *smokers*, which is 80.52%. The estimated slopes in the two regressions are same, but could be interpreted differently. In the first regression, the estimated slope is 0.08134, which suggests that the proportion for SGA among *smokers* is estimated to be *higher* than that among nonsmokers by a difference of proportions of 8.134%. In the new regression, the estimated slope is 0.08134, which suggests that the proportion *non*-SGA among *nonsmokers* is estimated to be *higher* than that among smokers by a difference of proportions of 8.134%.

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.
2. Methods: Subjects with missing values in either SGA or smoking behavior were removed from analysis. The probabilities for SGA were compared across groups with and without maternal smoking behavior. We used a Poisson regression model to test the ratio of the probabilities. The p-value and 95% confidence interval were calculated for the regression slope based on Wald statistic using maximum likelihood estimation.

Results: We have data on 751 subjects, among whom 231 (30.76%) subjects were smokers and 520 (69.24%) were nonsmokers. We estimate from a Poisson regression that the rate for SGA is 71.69% higher in the smokers than that in the nonsmokers. A 95% confidence interval suggests that our data would not be unusual if the rate for SGA in the smokers is anywhere from 20.19% to 145.27% higher than that in the nonsmokers. The two-sided p-value is 0.00302, so this observation is statistically significant at a 0.05 level of significance. Therefore we reject the null hypothesis that the rate of SGA is not associated with maternal smoking status.

1. Answers:

Odds of SGA for nonsmokers = 0.128; probability of SGA for nonsmokers = 0.113.

Odds of SGA for smokers = 0.242; probability of SGA for smokers = 0.195.

These estimated values agree exactly with the sample proportion and sample odds in problem 1, since this is a *saturated* Poisson regression model.

1. *Compare the four regressions.*
	* 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: Originally we fitted the model log(SGA) = b0 + b1\*x, where X is an indicator variable of smoker. If we change it into an indicator variable of nonsmoker denoted by Y, then Y = 1 – X. Then log(SGA) = b0 + b1\*(1-y) = (b0+b1) + (-b1)\*y. Hence the new intercept is (b0+b1), and the new slope is (-b1). The estimated intercepts in the two regressions are different; however, we note that exponentiation of the estimated intercept in the first regression is the estimated rate for SGA in nonsmokers, which is 11.35%, whereas that in the new model is the estimated rate for SGA in smokers, which is 19.48%. The absolute values of slopes in the two regressions are same (both 0.54054), but their signs differ. In the first regression, the estimated slope is 0.54054, which suggests that the rate for SGA among smokers is estimated to be higher than that among nonsmokers by 71.69% (exp(0.54054)=1.7169). In the new regression, the estimated slope is -0.54054, which suggests that the rate for SGA among *non*smokers is estimated to be only 58.24% that among smokers.

* + 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: Originally we fitted the model log(SGA) = b0 + b1\*x, where X is an indicator variable of smoker. If we change the response into notSGA, then we have log(1 – notSGA) = b0 + b1\*x. The estimated intercepts in the two regressions differ, but sum of exponentiation of two is 1. The exponentiation of estimated intercept in the first regression is the estimated rate for SGA among nonsmokers, which is 11.35%, whereas that in the new model is the estimated rate for *non*-SGA among nonsmokers, which is 88.65%. The estimated slopes in the two regressions also differ. In the first regression, the estimated slope is 0.54054, which suggests that the rate for SGA among smokers is estimated to be higher than that among nonsmokers by 71.69% (exp(0.54054)=1.7169). In the new regression, the estimated slope is -0.09624, which suggests that the rate for *non-*SGA among smokers is estimated to be only 90.83% (exp(-0.09624)=0.9083) that among *non*smokers.

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

Answers: If we change the response into notSGA, and predictor into nonsmoker, then we have log(1- notSGA) = b0 + b1\*(1-y). The estimated intercepts in the two regressions differ. The exponentiation of estimated intercept in the first regression is the estimated rate for SGA among *nonsmokers*, which is 11.35%, whereas that in the new model is the estimated rate for *non*-SGA among *smokers*, which is 80.52%. The estimated slopes in the two regressions also differ. In the first regression, the estimated slope is 0.54054, which suggests that the rate for SGA among smokers is estimated to be higher than that among nonsmokers by 71.69% (exp(0.54054)=1.7169). In the new regression, the estimated slope is 0.09624, which suggests that the rate for *non-*SGA among *non*smokers is estimated to be 10.1% (exp(0.09624)=1.101) that among smokers.

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?

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

Results: We have data on 751 subjects, among whom 231 (30.76%) subjects were smokers and 520 (69.24%) were nonsmokers. The mean of indicator variable of SGA among the 231 subjects was 0.1948, and that among the 510 nonsmokers was 0.1135. A 95% confidence interval that allows for unequal variances suggests that this difference of 0.0813 in means would not be unusual if the true mean of indicator variable of SGA among smokers is anywhere between 0.0231 and 0.1395 higher than that among nonsmokers. Based on a t test that allows for unequal variances, this observation is statistically significant at a 0.05 level of significance (two-sided p-value = 0.00628). Therefore we reject the null hypothesis that the mean of indicator variable of SGA is not associated with the maternal smoking status.

Comparisons: The analysis using t test is essentially same as those performed in problems 2-4, especially problem 3. From the estimated samples means of two groups in the t test, we can calculate the odds ratio of SGA between smokers and nonsmokers, which agrees with the exponentiation of estimated slope from logistic regression in problem 2. The two sample t test that allows for unequal variances is a special case of robust linear regression, so the estimates agree: the estimated intercept in the linear regression is the sample mean of SGA for nonsmokers in the t test, and the estimated slope in the linear regression is the estimated difference in means in the t test. (Note that the standard errors in the robust linear regression and t test differ slightly.) From the t test, we can also get estimated rate ratio of SGA between smokers and nonsmokers, which agrees with the exponentiation of the estimated slope from the Poisson regression.

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: The probabilities for SGA were compared across groups defined by the continuous measure of maternal age. We used a robust linear regression that allows heteroscedasticity to test difference in the probabilities. Standard error was estimated using Huber-White sandwich estimator. Two-sided p-value and a 95% confidence interval were computed for the difference in the population probabilities for SGA based the Wald statistic using approximate normal distribution.

Results: We have data on 755 subjects, whose mean age is 24.79.105 (13.91%) infants out of 755 were observed to be SGA, and 650 (86.09%) infants out of 755 were observed to be SGA. The difference in the mean probability for SGA is estimated to be 0.00452 lower for each year higher in age between two groups. A 95% confidence interval suggests that this difference of proportions of 0.00452 between the two groups would not be unusual if the true difference in probability for SGA in the older group is anywhere from 0.0286% to 0.874% (absolute difference) lower than that in the younger group per year difference in age. The two-sided p-value is 0.0364, 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 between risk ratio (RR: ratios of probabilities).

Methods: The probabilities for SGA were compared across groups defined by the continuous measure of maternal age. We used a Poisson regression model to test the ratio of the probabilities. The p-value and 95% confidence interval were calculated for the regression slope based on Wald statistic using maximum likelihood estimation.

Results: We have data on 755 subjects, whose mean age is 24.79.105 (13.91%) infants out of 755 were observed to be SGA, and 650 (86.09%) infants out of 755 were observed to be SGA. We estimated from a Poisson regression that the probability for SGA is a relative 3.38% 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 0.053% to 6.604% lower than that in the younger group with one year difference in age (95% confidence interval for risk ratio of SGA is 0.93396 to 0.99947). The two-sided p-value is 0.0465, 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: The odds of SGA were compared across groups defined by the continuous measure of maternal age. We used a logistic regression model to test the odds of SGA as a function of continuous measure of maternal age. The p-value and 95% confidence interval were calculated for the regression slope based on Wald statistic using maximum likelihood estimation.

Results: We have data on 755 subjects, whose mean age is 24.79.105 (13.91%) infants out of 755 were observed to be SGA, and 650 (86.09%) infants out of 755 were observed to be SGA. We estimated from a logistic regression that the odds for SGA is a relative 3.90% lower for each year increase in age between two groups (odds ratio = 0.961). A 95% confidence interval suggests that our data would not be unusual if the odds for SGA in the older group is anywhere from 0.069% to 7.584% lower than that in the younger group with one year difference in age (95% confidence interval for odds ratio of SGA is 0.92416 to 0.99931). The two-sided p-value is 0.0461, so this observation is statistically significant at a 0.05 level of significance. Therefore we 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 is 0.075. We note that the estimates from the three regressions are pretty similar, but different from the sample proportion. This is so because we used a continuous measurement for age, so we modeled the two regression parameters (the intercept and the slope) with many groups (number of distinct ages). 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).

Methods: A plot of sample proportions within each unique age by age is presented.



* 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: A plot of estimated probabilities of SGA by age derived from linear regression is presented. Sample proportions within each unique age are also superimposed.



* + 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: A plot of estimated probabilities of SGA by age derived from Poisson regression is presented. Sample proportions within each unique age are also superimposed.



* + 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: A plot of estimated probabilities of SGA by age derived from logistic regression is presented. Sample proportions within each unique age are also superimposed.



*(A plot combing the four plots is presented below.)*

**

*Comparisons:* We notice that all of linear regression, Poisson regression and logistic regression provide similar fitted values across age groups. They differ with the sample proportions at some ages, since they model a general trend of probability of SGA across age groups. The plot agrees with our analysis in problem 6: the fitted values from the three regressions are very similar (around 0.161), but the sample proportion at the age of 20 is around 0.72 as shown in the plot.

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: The odds of SGA were compared across groups defined by the continuous measure of logarithmically transformed maternal age. We used a logistic regression model to test the odds of SGA as a function of continuous measure of log-transformed maternal age. The p-value and 95% confidence interval were calculated for the regression slope based on Wald statistic using maximum likelihood estimation.

Results: We have data on 755 subjects, whose mean age is 24.79.105 (13.91%) infants out of 755 were observed to be SGA, and 650 (86.09%) infants out of 755 were observed to be SGA. We estimated from a 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 in log-transformed age. A 2-fold increase in age is estimated to cause an *odds* of SGA only 2-0.9536 = 0.516 as large. A 95% confidence interval suggests that our data would not be unusual if the odds for SGA in the older group is anywhere from 85.33% lower to 1.23% higher than that in the younger group with one year difference in log-tranformed age (95% confidence interval for odds ratio of SGA is 0.1467 to 1.0123). The two-sided p-value is 0.053, so this observation is not statistically significant at a 0.05 level of significance. Therefore we do 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?

Scatterplots of log odds of SGA from samples against age and log-transformed age are plotted separately below.





Answers: I think it is *not* reasonable to have performed such analysis using log-transformed age. Although the log odds of SGA show a general linear trend across groups defined by log-transformed age, it is unnatural to understand age on a log-transformed/multiplicative scale. It would be easier to interpret age with certain years of difference, rather than a difference by 1 in log-transformed age. In addition, the range of log-transformed age is from about 2.6 to 3.8, which is too small. So I think we should use untransformed age as in problem 6c.