**Biost 518: Applied Biostatistics II**

**Biost 515: Biostatistics II**

Emerson, Winter 2015

**Homework #3**

January 23, 2015

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.

**Method:** Descriptive analysis were used to examine study participants’ birth growth measurements and demographic characteristics such as gestational age at birth, birth weight and sex and maternal factors that can be associated with small for gestational age (SGA) such as mother’s age, height, smoker status and parity. The descriptive statistics for growth measurements/demographic characteristics and maternal factors were reported for the total study sample and within two groups defined by babies who are either small for gestational age or not. The descriptive statistics used were number of subjects, sample’s mean, standard deviation, minimum and maximum provided for continuous data such as gestational age at birth (weeks), birth weight (grams), maternal age (years), height (cm) and the frequency mean for categorical data.

**Result:** The study sample is for 755 subjects with four having missing values in smoking status and sex variables, while gestational age at birth and height variables had five and six missing values respectively. All the missing data (n=11) was excluded from this data set for the descriptive analysis and all subsequent analysis (including questions 2 to 8) used this complete data set. Out of the 744 analyzed, 97 (13.0%) were small for gestational age. Table 1 below shows that babies who are small for gestational age are likely to be female, have lower birth weight and are born to mothers with lower height and smoking.

**Table 1:** Study population growth measurements/demographic characteristics and maternal risk factors within total study sample and stratified within two groups defined by being either small for gestational age or not.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Babies who are Small for Gestational Age1[97] | Babies who are not Small for Gestational Age1[647] | Total Study Sample1 [744] |
| Population Growth Measurements/demographic characteristics at birth (units) |
| Gestational Age at delivery (weeks) 2 | 37.9 (2.2; 30-42) | 39.4 (1.2; 38-44) | 39.2 (1.5; 30-44) |
| Birth weight (grams) 2 | 2231.1 (411.6; 1035- 3780)  | 3246.2 (402.1; 2510- 4730)  | 3105.6 (534.5; 1035- 4730)  |
| Males 3 | 42.3% [41] | 52.4% [339] | 51.0%  |
| Maternal Risk Factors (units) |
| Age (years) 2 | 23.8 (4.9: 16-35) | 24.9 (5.4; 14-43) | 24.79 (5.4; 14-43) |
| Height (cm) 2 | 154.56 (5.87; 142- 172) | 157.01 (6.54; 106- 176) | 156.68 (6.50; 106-176) |
| Smoker status (yes) 3 | 44.3% [43] | 28.7% [186] | 30.8% |
| Parity (number of prior deliveries) 2 | 0.9 (1.1; 0-6) | 1.1 (1.2; 0-6) | 1. 1 (1.2; 0-6) |

1 Study sample subgroups and total sample number depicted as [N]

2 Descriptive statistics for each stratum of continuous variables include the Mean (SD; min – max), [N]

3 Descriptive statistics for each stratum of categorical variables include the Percentage %, [N]

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.)

**Method:** The odds of delivering a baby with SGA among mothers who smoke was compared to the odds of delivering a baby with SGA among mothers who do not smoke using a logistic regression model. The small for gestational age (SGA) variable coded as 1=with SGA and 0=with no SGA, while maternal smoking status recoded to 1=Smoker and 0=Nonsmoker in the complete data set that has all missing data excluded as explained in question were used to generate the estimates. The Wald statistic was used to estimate the regression slope and its standard error. The two-sided p value and 95% confidence interval were computed using normal distribution for logistic regression parameter estimates. Because this model is saturated i.e. the number of the regression parameters in the model (intercept and slope) is equal to sample subgroups (smokers and non smokers) I did not use the robust SE.

Model used to generate estimates

Log odds of SGA = intercept/($β\_{0}$) + Slope/($β\_{1})$\*Smokers

* 1. Give full inference regarding the association between SGA and maternal smoking.

**Inference:** Among mothers who do not smoke (n= 515), the odds of SGA is 0.1171, while among mothers who smoke (n= 229), the odds is 0.2312. The odds ratio (OR) of SGA between mothers who smoke and those who do not is 1.97. The 95% CI shows that this estimate (OR) will not be unusual if the true population OR is between 1.28 and 3.05. The p value for the z statistic is significant at 0.05 level (two sided p=0.002) and thus we can confidently reject the null hypothesis that states there is no difference in the odds (i.e. OR=1) of babies with SGA between mothers who smoke compared to mothers who do not smoke and conclude that, in this sample, the odds of SGA among mothers who smoke is 97.36% higher than those who do not.

* 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.

**Method:**

Regression Model Estimates:

The log odds estimates of the model (Log odds of SGA = -2.144414 + 0.6798673 \*Smokers) were used to calculate the odds as follows:

1. The odds of SGA among mothers who do not smoke = e-2.144414 = 0.11714
2. The odds of SGA among mothers who smoke = e -2.144414 \*e 0.6798673 = 0.23118

The proportion of babies with SGA were then calculated using the formula p = (odds/1+odds) as follows:

* + 1. The proportion of SGA among mothers who do not smoke =(0.1171 / 1.1171) = 0.105
		2. The proportion of SGA among mothers who smoke = (0.2312/ 1.2312) = 0.188

Descriptive Statistic:

Table 2: The number of babies with SGA and without SGA stratified by the smoking status of their mothers.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Without SGA | With SGA | Total |
| Non Smoker | 461 | **54\*** | 515 |
| Smoker | 186 | **43\*** | 229 |
| Total | 647 | 97 | 744 |

\* Number of babies with SGA delivered from mothers who smoke and do not smoke. This was displayed in Table 1 as part of descriptive statistics.

Method to calculate the proportion and odds from descriptive statistics table

The proportion was calculated by dividing the number of babies with SGA delivered from mothers who smoke and do not smoke divided the total number of mothers who smoke or do not smoke respectively:

1. Proportion of babies with SGA among mothers who do not smoke = 54/515 = 0.105
2. Proportion of babies with SGA among mothers who smoke = 43/229 = 0.188

 The odds of SGA in the sample was calculated as follows:

1. The odds of SGA among mothers who do not smoke by dividing the number of babies with SGA delivered from mothers who do not smoke divided by the number babies without SGA delivered from women who do not smoke = 54/461 = 0.1171
2. The odds of SGA among mothers who smoke by dividing the number of babies with SGA delivered from mothers who smoke divided by the number babies without SGA delivered from women who smoke = 43/186 = 0.23118

**Results:** The proportion calculated from regression model estimates exactly matches the sample’s proportion of babies with SGA born from mothers who smoke (n=43) and proportion of babies with SGA born from mothers who do not smoke (n=54). Similarly, the fitted odds from the model matched the odds of the sample. This finding is expected because the logistic regression model is saturated i.e. the number of the regression parameters in the model (intercept and slope) is equal to sample subgroups (smokers and non smokers).

* 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.):

My Model (0) is a logistic regression model fitted for response SGA on predictor SMOKER

Log odds SGA/SMOKER = Log (Pr (SGA=1/SMOKER)/ Pr (SGA=0/ SMOKER)) = $β\_{0 }+ β\_{1}\*SMOKER$

$β\_{0 }$: Is the log odds of SGA for mothers who do not smoke

$β\_{1}$: Is the difference in log odds of SGA for mothers who smoke compared to log odds of SGA for mothers who do not smoke

* + 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.

Model (1) has the predictor re-parameterized and the logistic regression model is expressed as:

Log odds SGA/NONSMOKER = Log (Pr (SGA/NONSMOKER)/ Pr (SGA/NONSMOKER)) = $α\_{0}+ α\_{1}\*NONSMOKER$

$α\_{0}$: Is the log odds of SGA for mothers who smoke

$α\_{1}: $ Is the difference in logs odds of SGA of mothers who do not smoke compared to SGA of mothers who smoke

* + 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.

Model (2) has the outcome re-parameterized and the logistic regression model is expressed as:

Log odds NOTSGA/SMOKER = Log (Pr (NOTSGA/SMOKER)/ Pr (NOTSGA/SMOKER)) = $γ\_{0}+ γ\_{1}\*SMOKER$

$γ\_{0}$: Is the log odds of No SGA of mothers who do not smoke

$γ\_{1}$: Is the difference in log odds of No SGA of mothers who smoke compared to the log odds of No SGA to mothers who do not smoke

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

Model (3) has both the outcome and predictor re-parameterized and the logistic regression model is expressed as:

Log odds NOTSGA/NONSMOKER = Log (Pr (NOTSGA/NONSMOKER)/ Pr (NOTSGA/NONSMOKER)) = $δ\_{0}+ δ\_{1}\*NONSMOKER$

$δ\_{0}$ : Is the log odds of NOTSGA among mothers who smoke

$δ\_{1}$: Is the difference in log odds of No SGA among mothers who do not smoke compared to the log odds of No SGA among mothers who do not smoke

Comparisons:

Model 0 = Log odds SGA/SMOKER = Log (Pr (SGA/SMOKER)/ Pr (SGA/ SMOKER)) = $β\_{0 }+ β\_{1}\*SMOKER$

Model 2 = Log odds NOTSGA/SMOKER = Log (Pr (NOTSGA/SMOKER)/ Pr (NOTSGA/SMOKER)) = $γ\_{0}+ γ\_{1}\*SMOKER$

Both these models will have exact coefficients but will have different signs because the reference group is reversed. I would expect based on the estimates of model 0 which were Log odds of SGA = -2.144414 + 0.6798673 \*Smokers to be for model 2 as Log odds of NOTSGA = 2.144414 - 0.6798673 \*Smokers

Similarly, models 1 and 3 will have a similar relationship because the reference group comparing the odds of as the reference group for comparison of odds is reversed by parameterization.

Model 1 = Log odds SGA/NONSMOKER = Log (Pr (SGA/NONSMOKER)/ Pr (SGA/NONSMOKER)) = $α\_{0}+ α\_{1}\*NONSMOKER$

Model 3= Log odds NOTSGA/NONSMOKER = Log (Pr (NOTSGA/NONSMOKER)/ Pr (NOTSGA/NONSMOKER)) = $δ\_{0}+ δ\_{1}\*NONSMOKER$

I ran models 1 and 3 and the estimates were as follows:

Model 1 = Log odds of SGA/NONSMOKER = - 0.6798673 - 0.784793 \*NONSMOKERS

Model 3 = Log odds of NON SGA/NONSMOKER = 0.6798673 + 0.784793 \*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.

**Method:** To evaluate differences in probabilities for SGA across smoking groups, a linear regression model that assumes unequal variance was used to compare the probability of having a baby with SGA among mothers who smoke to the probability of having a baby with SGA among mothers who do not. The Wald statistic was used to estimate the difference in probabilities (slope/$β\_{1}$) and the standard error using Huber-White sandwich estimator. The approximate normal distribution for linear regression estimates was used to compute the two-sided p value and 95% CI.

**Inference:** The proportion of babies with SGA is 0.105 among mothers who do not smoke (n= 515), in comparison the proportion is 0.188 among mothers who smoke (n= 229). The observed higher difference of 8.29% suggests poorer pregnancy outcome i.e. higher SGA probability among mothers who smoke. The 95% CI shows that this finding will not be unusual if the true probability difference is anywhere between 3.07% and 13.51%. We can with high confidence, based on the two-sided p value (p=0.002) of the t statistic reject the null hypothesis that states that there is no association between the probability of having a baby with SGA and mothers who smoke.

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.

a.

**Method:** To evaluate the ratio in probabilities for SGA across smoking groups, a Poisson regression model that assumes unequal variance to compare the probability of having a baby with SGA among mothers who smoke relative to the probability of having a baby with SGA among mothers who do not. The Wald statistic was used to estimate the difference in probabilities (slope/$β\_{1}$) and the standard error using Huber-White sandwich estimator. The approximate normal distribution for linear regression estimates was used to compute the two-sided p value and 95% CI.

**Inference:** The proportion of babies with SGA is 0.105 among mothers who do not smoke (n= 515); in comparison to the proportion is 0.188 among mothers who smoke (n= 229). This observed ratio of 79.08% (rate ratio = 1.79) relative increase in proportion of babies with SGA among mothers who smoke compared to mothers who do smoke suggests poorer pregnancy outcome i.e. higher SGA probability among mothers who smoke. This finding will not be unusual if the true probability rate ratio is anywhere between 23.78% and 259.09% relative increase in SGA probability among mothers who smoke compared to mothers who do not. We can with high confidence, based on the two-sided p value (p=0.002) of the z statistic reject the null hypothesis that states that there is no association between the probability of having a baby with SGA and mother who smokes.

4b. **Method:**

Regression Model Estimates:

The proportion estimates of the model = 0.1048544+ 1.790797\*Smokers were used to calculate the odds as follows:

1. The proportion of SGA among mothers who do not smoke = 0.1048544
2. The proportion of SGA among mothers who smoke = 0.1048544 \*1.790797= 0.188

To calculate the odds, we use the formula = odds = p/1-p. Therefore the odds for the above proportions is:

1. The odds of SGA among mothers who do not smoke = 0.1049/ 1 - 0.1049 = 0.117
2. The odds of SGA among mothers who smoke = 0.1877/ 1- 0.1877 = 0.231

Descriptive Statistic:

The proportion that was calculated as explained in question 2 (b) generated the following:

1. Proportion of babies with SGA among mothers who do not smoke = 54/515 = 0.105
2. Proportion of babies with SGA among mothers who smoke = 43/229 = 0.188

Similarly the odds of SGA that was calculated in question 2 (b) generated the following:

1. The odds of SGA among mothers who do not smoke = 54/461 = 0.117
2. The odds of SGA among mothers who smoke = 0.231

**Results:** The proportion calculated from regression model estimates exactly matches the sample’s proportion of babies with SGA born from mothers who smoke (n=43) and proportion of babies with SGA born from mothers who do not smoke (n=54). Similarly, the fitted odds from the model matched the odds of the sample. This finding is expected because the Poisson regression model is saturated i.e. the number of the regression parameters in the model (intercept and slope) is equal to sample subgroups (smokers and non smokers).

**4c.**

My Model (0) in a Poisson regression model fitted for response SGA on predictor SMOKER

E(SGA\T, SMOKER) = log ($λ$T) = log(T)+ intercept/($β\_{0}$)/nonsmokers + Slope/($β\_{1})$\*SMOKER

$β\_{0 }$: Is the proportion of SGA for mothers who do not smoke

$β\_{1}$: Is the ratio of SGA proportion for mothers who smoke compared to mothers who do not smoke

* + 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.

Model (1) has the predictor re-parameterized and the Poisson regression model is expressed as:

E (SGA\T, NONSMOKER) = log ($λ$T) = log (T)+ intercept/($α\_{0}$)/smokers + Slope/($α\_{1})$\*NONSMOKER

$α\_{0}$: Is the proportion of SGA for mothers who smoke

$α\_{1}: $ Is the ratio of SGA proportion for mothers who do not smoke compared to SGA of mothers who smoke

* + 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.

Model (2) has the outcome re-parameterized and the logistic regression model is expressed as:

E (NOTSGA\T, SMOKER) = log ($λ$T) = log (T)+ intercept/($γ\_{0}$)/nonsmokers + Slope/($γ\_{0})$\*SMOKER

$γ\_{0}$: Is the proportion of No SGA of mothers who do not smoke

$γ\_{1}$: Is the ratio in proportion of No SGA of mothers who smoke compared to mothers who do not smoke

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

Model (3) has both the outcome and predictor re-parameterized and the logistic regression model is expressed as:

E (NOTSGA\T, NONSMOKER) = log ($λ$T) = log (T)+ intercept/($δ\_{0}$)/smokers + Slope/($δ\_{0})$\*NONSMOKER

$δ\_{0}$ : Is the proportion of NOTSGA among mothers who smoke

$δ\_{0}$: Is the ratio in proportion of No SGA among mothers who do not smoke compared to the rates No SGA among mothers who do not smoke

Comparisons:

Model 0 = E(SGA\T, SMOKER) = log ($λ$T) = log(T)+ intercept/($β\_{0}$)/nonsmokers + Slope/($β\_{1})$\*SMOKER

Model 2 = E (NOTSGA\T, SMOKER) = log ($λ$T) = log (T)+ intercept/($γ\_{0}$)/nonsmokers + Slope/($γ\_{0})$\*SMOKER

Both these models will have exact intercept coefficients but the slope will be inversely related to each other because the reference group is reversed.

Similarly, models 1 and 3 will have a similar relationship, the intercept will be the same but the slope will be inversely related because the reference group comparing the rates is reversed by parameterization.

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?

**Method:** The unconditional exact test based on chi square test was used to compare two SGA samples (with and without SGA) by smoking status. Risk differences and risk ratios were generated, with 95% CI and two-sided p values.

**Inference:** The risk /probability of baby with SGA is 0.188 among mothers who smoke. In comparison, the risk of baby with SGA among mothers who do not smoke is 0.105. Both the values for proportion match results on proportions for questions 2 (b) obtained through logistic regression and manually and using a linear regression in question 3.

The difference in risk of having a baby with SGA among women who smoke is 8.29% higher compared to mothers who do not smoke. Based on the 95% CI, this difference will not be unusual if the true risk difference is anywhere between 2.58% and 14.00%. The estimate here matches the slope of linear regression (question 3) but the computed 95% CI are slightly different from question 3 which was between 3.07% and 13.51%.

The risk of SGA among mothers who smoke is 79.08% (this estimate matches exactly the estimate in question 4) higher relative to mothers who do not smoke and will not be unusual if risk ratio is between 23.81% and 259.03% (which is slightly different from question 4 estimate: 23.78% and 259.09%) higher for mothers who smoke compared to mothers who do not. This finding is significant (two sided p = 0.0019) which is the same to the z statistic and p value in question 4, therefore rejecting the null hypothesis (RD=0 and RR=1) that states that there is no association between baby with SGA and mother’s smoking status.

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).

**Method:** To evaluate differences in probabilities for SGA across groups defined by maternal age as a continuous variable, a linear regression model that assumes unequal variance was used, with SGA as an outcome variable (binary) and Age as the predictor. To have a more meaningful intercept, age was parameterized by subtracting 14 years (minimum age) from all ages thus reference group will be mothers at age 14 years. The Wald statistic was used to estimate the difference in probabilities (slope/$β\_{1}$) and the standard error using Huber-White sandwich estimator. The approximate normal distribution for linear regression estimates was used to compute the two-sided p value and 95% CI.

Model used to generate estimates

SGA = intercept/($β\_{0}$)/mothers at age 14 years + Slope/($β\_{1})$\*Age

Fitted Estimates: Linear model SGA = 0.1909324 - 0.0056458 \*Age

**Inference:** The predicted proportion of babies with SGA delivered to mothers’ age 14 years is 19.09% and this finding ~~will be~~ would not be unusual if the true proportion is between 13.62% and 24.56%. The difference in proportions of babies with SGA between two groups of mothers above 14 years who differ by 1 year of age is 5.65% less for the older age group compared to they younger group. The 95% CI shows that this finding will not be unusual if the true probability difference is anywhere between 9.74% and 1.55% less for the older age. We can with high confidence, based on the two-sided p value (p=0.007) of the t statistic reject the null hypothesis that states that there is no association between the probability of having a baby with SGA and mothers age 14 years and above.

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

Model used to generate estimates

E(SGA\T, Age) = log ($λ$T) = log(T)+ intercept/($β\_{0}$)/mothers at age 14 years + Slope/($β\_{1})$\*Age

Fitted Estimates: SGA Rates = 0.1909324 - 0.0056458 \*Age (Rate Ratio)

**Method:** To evaluate the ratio in probabilities for SGA across age groups, a Poisson regression model that assumes unequal variance was used. SGA was defined as a binary event and age as predictor (continuous variable). To have a more meaningful intercept, age was parameterized by subtracting 14 years (minimum age) from all ages thus reference group will be mothers at age 14 years. The Wald statistic was used to estimate the difference in probabilities (slope/$β\_{1}$) and the standard error using Huber-White sandwich estimator. The approximate normal distribution for linear regression estimates was used to compute the two-sided p value and 95% CI.

**Inference:** The proportion of babies with SGA is 20.94% among mothers age 14 years. The observed ratio of 4.59% (rate ratio = 0.954) relative decrease in proportion of babies with SGA among older mothers suggests that poorer pregnancy outcome is associated with lower maternal age ~~of mother~~. This finding will not be unusual if the true probability rate ratio is anywhere between 1.07% and 7.98% relative decrease in SGA probability among one year older mothers group. We can with high confidence, based on the two-sided p value (p=0.011) of the z statistic reject the null hypothesis that states that there is no association between the probability of having a baby with SGA and mothers’ age.

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

 **Method:** The odds of delivering a baby with SGA was compared between mothers of different age groups using a logistic regression model. The small for gestational age (SGA) variable coded as 1=with SGA and 0=with no SGA was outcome variable, while the predictor was age as a continuous variable was parameterized by subtracting 14 years (minimum age) from all ages to have a meaningful intercept/reference group. The Wald statistic was used to estimate the regression slope and its standard error. The two-sided p value and 95% confidence interval were computed using normal distribution for logistic regression parameter estimates. Because this model is saturated i.e. the number of the regression parameters in the model (intercept and slope) is equal to sample subgroups (smokers and non smokers) I did not use the robust SE.

Model used to generate estimates

Log odds of SGA = intercept/($β\_{0}$)/mother at age 14 years + Slope/($β\_{1})$\*Age

Fitted Estimates: SGA odds = 0.2590214 \* 0.9476836\*Age (Odds Ratio)

**Inference:** Among mothers of age 14 years, the odds of SGA is 0.259, for age groups differing by one year there is a 5.23% lower odds for older age group. The 95% CI shows that this estimate (OR) will not be unusual if the true population OR is between 0.908 and 0.990. The p value for the z statistic is significant at 0.05 level (two sided p=0.015) and thus we can confidently reject the null hypothesis that states there is no difference in the odds (i.e. OR=1) of babies with SGA among mothers who differ in age groups.

* 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.

\*\*Please note that my models had age parameterized (age – 14 years), therefore the constant/intercept coefficients will be different from estimates generated from a model with non-parameterized age.

Linear model SGA = 0.1909324 - 0.0056458 \*6 = 0.157

To convert to proportion = odds/1+odds = 0.157/1.157 = 13.57%

Poisson model SGA = 0.2093789 \* 0.95412596 = 0.157

To convert to proportion = odds/1+odds = 0.157/1.157 = 13.57%

Logistic Model SGA = 0.2590214 \* 0.94768366= 0.188

To convert to proportion = odds/1+odds = 0.188/1.188 = 15.82%

The sample proportions for linear and Poisson models are the same and the sample proportion for logistic regression is slightly different.

1. 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).

Figure 1: Estimated probability of having a Small for Gestational Age babies for Maternal Ages (14- 43 years)



Figure 1 shows that overall the SGA probabilities estimated by Poisson and Logistic regression models for all maternal age groups are more similar compared to the estimates of the linear regression model. Of note, the probability estimates for SGA are very similar for all three models for maternal ages between 16 to 35 years.

The fitted values that were generated in question 6 (with age parameterized Age-14 years) were as follows:

Linear model SGA = 0.1909324 - 0.0056458\*Age14

Poisson model SGA = 0.2093789 \* 0.9541259\*Age14

Logistic Model SGA = 0.2590214 \* 0.9476836\*Age14

The fitted values between the Poisson and Logistic model are closer compared to linear model as shown in the plot graphs. However, the estimate for female age 20, the estimated proportion was exactly the same between linear and Poisson model (13.57%) compared to the logistic model (15.82%). In the plot the estimated probabilities at this point are all similar.

* 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.
		2. 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.
		3. 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.

The estimated probabilities using this method also produced same results to the previous graph (see Figure 2). The probability estimates from Poisson and Logistic model are generally similar for all maternal age groups. However, estimated probabilities for SGA are similar for all models between maternal age ranges of 16 to 35 years.

The fitted values that were generated in question 6 (with age parameterized Age-14 years) were as follows:

Linear model SGA = 0.1909324 - 0.0056458\*Age14

Poisson model SGA = 0.2093789 \* 0.9541259\*Age14

Logistic Model SGA = 0.2590214 \* 0.9476836\*Age14

The fitted values between the Poisson and Logistic model are closer compared to linear model as shown in the plot graphs. However, the estimate for female age 20, the estimated proportion was exactly the same between linear and Poisson model (13.57%) compared to the logistic model (15.82%). In the plot the estimated probabilities at this point are all similar.

Figure 2: Estimated probability of having a Small for Gestational Age babies for Maternal Ages (14- 43 years)



1. Perform a logistic regression analyses of the distribution of the prevalence of SGA infants across groups defined by the logarithmically transformed maternal age.

Method:

**Method:** The odds of delivering a baby with SGA was compared between mothers of different age groups using a logistic regression model. The small for gestational age (SGA) variable coded as 1=with SGA and 0=with no SGA was outcome variable, while the predictor (age) was log transformed to the base of 10. The $β\_{1} $estimate was multiplied by log base 10 (1.1) and expnentiated to generate meaningful estimates (i.e. 10% increase in age). The Wald statistic was used to estimate the regression slope and its standard error. The two-sided p value and 95% confidence interval were computed using normal distribution for logistic regression parameter estimates. Because this model is saturated i.e. the number of the regression parameters in the model (intercept and slope) is equal to sample subgroups (smokers and non smokers) I did not use the robust SE.

Model used to generate estimates

Log odds of SGA = intercept/($β\_{0}$)/mother + Slope/($β\_{1})$\*logAge

Fitted Estimates: SGA odds = 2.143883 \* -1.277279 \*logAge (Odds Ratio)

* 1. Provide formal inference for associations using odds ratio (OR: ratios of odds) and log transformed age.

**Inference:** The intercept is not interpreted here because it is nonsensical and our inference will be based on the slope. Mothers who differ by 10% age difference have a lower 11.46% relative difference in odds of SGA, with older age groups having lower odds compared to the younger age group. The 95% CI shows that this difference in odds will not be unusual if the true relative difference is between 2.26% and 19.79% lower for the older age group compared to the younger age group. The p value for the z statistic is significant at 0.05 level (two sided p=0.016) and thus we can confidently reject the null hypothesis that states there is no relative difference in the odds of babies with SGA among mothers who differ in age groups.

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

This analysis will not make sense for two reasons:

* + 1. Biological/scientific relationship between outcome and predictor is not on a multiplicative scale
		2. Interpretation of model is hard, using age changes in a multiplicative rather than linear scale does is difficult to comprehend. Interpretation of age using a % increase is hard to understand and a simpler linear scale makes more sense.