**BIOST 518/515: Applied Biostatistics II/Biostatistics II**

Emerson, Winter 2015

**Homework 03**

January 29, 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:**

As the scientific question focuses mostly on the association with delivery of babies that are small for gestational age (SGA), descriptive statistics were provided within groups defined by whether or not the mother delivered a baby that was SGA as well as for the entire sample. Descriptive statistics included for continuous variables (maternal height, maternal age, parity, infant birthweight, and gestational age) were (arithmetic) mean, standard deviation, minimum and maximum. Descriptive statistics included for binary variables (maternal smoking behavior, and sex of infant) were percentages.

Note that in the raw dataset, maternal smoking behavior was coded as 1 for smoker and 2 for non-smoker, as well as the sex of infant was coded as 1 for boy and 2 for girl. New indicators were created for these two variables to bring them back to the standard scales: maternal smoking behavior with 1 as smoker and 0 for non-smoker, and sex of infant with 1 as boy and 0 as girl.

**Results:**

The total number of observations in the dataset is 755. None of the observations was missing data on whether the babies delivered were small for gestational age (SGA) or not. Of the 755 observations, 105 mothers delivered babies that were SGA, and 650 mothers delivered babies that were not SGA. Table 1a shows the descriptive statistics within these two subgroups (referred to as the SGA group and non-SGA group, hereafter) and all observations.

Out of the entire 755 observations, 6 were missing data on maternal height, 4 were missing data on maternal smoking behavior, 4 were missing data on infant birthweight, 4 were missing data on sex of infant, and 5 were missing data on gestational age at delivery. Table 1b summaries the number of observations with missing data by subgroups. The number of missing data was small and there was no additional information to determine the pattern of the missing data.

Mothers in the SGA group tended to be slightly shorter on average compared to their counterparts, however, the mean difference was just roughly 1.5 cm. Mothers in the non-SGA group tended to be slightly older and experienced slightly more from prior deliveries, however, the differences were small. 42.9% of the mothers in the SGA group were smokers, while only 28.6% of their counterparts were smokers. It is obvious that on average infants in the SGA group tended to be much smaller than the infants in the non-SGA group (with more 1,000 grams of difference in mean birthweights) and a slightly smaller gestational age at delivery. 68.6% of SGA babies were baby boys, while this proportion in the non-SGA group was only 42.9%.

**Table 1a. Descriptive Statistics within Groups by Delivery of Babies that are SGA and Sample**

|  |  |
| --- | --- |
| **Variables** | **Delivery of Babies** |
| **SGA (n = 105)** | **Not SGA (n = 650)** | **All Sample (n = 755)** |
| **Mother’s Characteristics:** |  |  |  |
| **Height (cm)1** | 154.6 (5.9; 142.0 – 172.0) | 157.0 (6.5; 106.0 – 176.0) | 156.7 (6.5; 106.0 – 176.0) |
| **Age at enrollment (years)1** | 23.9 (4.9, 16.0 – 35.0) | 24.9 (5.5; 14.0 – 43.0) | 25.0 (5.4; 14.0 – 43.0) |
| **Prior deliveries (counts)1** | 0.9 (1.11; 0.0 – 6.0) | 1.13 (1.23; 0.0 – 6.0) | 1.1 (1.2; 0.0 – 6.0) |
| **Smoker (%)** | 42.9% | 28.6% | 30.6% |
| **Infant’s Characteristics:** |  |  |  |
| **Birthweight (grams)1** | 2,231 (412; 1,035 – 3,780) | 3,246 (402.1; 2,510 – 4,730) | 3,106 (534; 1,035 – 4,730) |
| **Sex – Boy (%)** | 68.6% | 42.9% | 46.5% |
| **Gestational age (weeks)1** | 37.9 (2.2; 30.0 – 42.0) | 39.4 (1.2; 38.0 – 44.0) | 39.2 (1.5; 30.0 – 44.0) |

*1 Descriptive statistics presented are: mean (standard deviation; minimum – maximum)*

**Table 1b. Number of Observations with Missing Data by Variables and Subgroups**

|  |  |
| --- | --- |
| **Variables** | **Delivery of Babies** |
| **SGA (n = 105)** | **Not SGA (n = 650)** | **All Sample (n = 755)** |
| **Maternal height**  | 6 | 0 | 6 |
| **Maternal age at enrollment**  | 0 | 0 | 0 |
| **Prior deliveries**  | 0 | 0 | 0 |
| **Maternal smoking behavior (smoker)** | 1 | 3 | 4 |
| **Infant birthweight**  | 1 | 3 | 4 |
| **Sex of infant (boy)** | 1 | 3 | 4 |
| **Gestational age**  | 2 | 3 | 5 |

2. 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.)

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

**Method:**

The odds of delivery of infants who were SGA were compared between two groups defined by maternal smoking behavior, i.e. smokers vs. non-smokers, using a classical logistic regression model, where the outcome of interest was the odds of delivery SGA infants and the predictor of interest was whether the mother was a smoker. Due to the mean-variance relationship in logistic model, it is not necessary to use robust standard errors. Statistical inference was then based on the Wald statistic computed from the regression results. The two-sided p-value and the 95% confidence interval were constructed using the asymptotic normal distribution for logistic regression.

4 observations with missing data on maternal smoking behavior were omitted from this analysis. As a result, there was a total of 751 observations, with 231 smokers and 520 non-smokers.

**Results:**

There were 751 observations with complete data on maternal smoking behavior and whether the babies were SGA, among which 231 were smokers and 520 were non-smokers. Among non-smokers, the odds of delivery of infants who were SGA was 0.128. For smokers, the odds of delivering SGA infants was 0.242. The observed odds ratio between the smokers and non-smokers was 1.890. With 95% confidence, this is consistent with such true odds ratio being anywhere between 1.233 and 2.883. This result is statistically significant at the 5% critical threshold with a two-sided p-value of 0.003. As a result, we reject the null hypothesis, in favor of an alternative hypothesis that the odds of delivering SGA babies were associated with maternal smoking behavior.

b. 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:**

Using the regression model parameter estimates, the estimated odds of delivering a SGA infant for non-smokers is e-2.0559 = 0.128

The estimated probability of delivering a SGA infant for non-smokers is e-2.0559/(1+ e-2.0559) = 0.113

The estimated odds of delivering a SGA infant for smokers is e-2.0559+0.6368 = 0.242

The estimated probability of delivering a SGA infant for smokers is e-2.0559+0.6368/(1+ e-2.0559+0.6368) = 0.195

In Problem 1, I reported the probability of being smokers in each subgroup defined by whether or not the mother delivered an SGA infant. The proportion of smokers in the SGA group is 42.9%, while it is only 28.6% in the non-SGA group. These probabilities are inherently different from what we found in this question – the probability of delivering a SGA infant among non-smokers and smokers, because we stratified on different variables. However, these results somewhat agree with each other in the sense that smokers tend to have higher risk of delivering SGA, or mothers delivering SGA babies were more likely to be smokers. If I were to report the odds ratio in question 1, it should be exactly the same as the observed odds ratio found in the logistic regression.

c. 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.):

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

ii. 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.

iii. You fit a regression model of response NOTSGA on predictor NONSMOKER.

**Answer:**

The logistic regression I performed in part a used an indicator for SGA as the outcome of interest and an indicator for SMOKER as the predictor of interest. We essentially fit the following model:

log[odds(SGA|SMOKER)] = b0+ b1SMOKER [1]

The estimated intercept of b0= -2.0559 implies that the log-odds of delivering a SGA infant for non-smokers is -2.0599, while the estimated slope of b1=0.6368 implies that the log-odds of delivering a SGA infant for smokers is b0+ b1= (-2.0559) + 0.6368 = -1.4191

*Part i:*

If instead, an indicator NONSMOKER was created, and a logistic regression model of response SGA on predictor NONSMOKER was performed, we then have the following model:

log[odds(SGA|NONSMOKER)] = a0+ a1NONSMOKER [2]

Theoretically, because of linear transformation, we should get similar statistical inference from this model, i.e. statistical significance, although z-statistics, confidence intervals and p-values may vary very slightly, due to the fact that the logistic regression uses numerical iterations to find the maximum likelihood estimators.

In this model, intercept a0 now represents the log-odds for smokers, while (a0 + a1) represents the log-odds for non-smokers. Because the estimated log-odds in model [2] should agree with those in model [1], if we were to fit this model, we should observe the following:

Log-odds for smokers: a0 = b0+ b1

Log-odds for non-smokers: a0 + a1 = b0 🡪 a1= b0 - a0 = - b1

Conclusion: **a0 = b0+ b1** and **a1= - b1** where b0= -2.0559, b1=0.6368

*Part ii*

Instead, an indicator NOTSGA was created, and a logistic regression model of response NOTSGA on predictor SMOKER was performed, we then have the following model:

log[odds(NOTSGA|SMOKER)] = c0+ c1SMOKER

Note that, from probability, we know odds(NOTSGA|SMOKER) = 1/odds(SGA|SMOKER), and thus, by properties of logarithm, we have:

log[1/odds(SGA|SMOKER)] = c0+ c1SMOKER

equivalent to: log[1] – log[odds(SGA|SMOKER)] = c0+ c1SMOKER

equivalent to: log[odds(SGA|SMOKER)] = - c0- c1SMOKER [3]

In this model, intercept (-c0) now represents the log-odds (SGA) for non-smokers, while (-c0- c1) represents the log-odds for smokers. Because the estimated log-odds in model [3] should agree with those in model [1], if we were to fit this model, we should observe the following:

Log-odds for non-smokers: - c0 = b0 🡪 c0 = -b0

Log-odds for smokers: - c0- c1 = b0+ b1 🡪 c1 = -b1

Conclusion: **c0 = - b0**and **c1 = - b1**where b0= -2.0559, b1=0.6368

*Part iii*

Instead, a logistic regression model of response NOTSGA on predictor NOTSMOKER was performed, we then have the following model:

log[odds(NOTSGA|NONSMOKER)] = d0+ d1NONSMOKER

Using similar proof as part (i), this model is equivalent to:

log[odds(SGA|NONSMOKER)] = -d0- d1NONSMOKER [4]

In this model, intercept (-d0) now represents the log-odds (SGA) for smokers, while (-d0 – d1) represents the log-odds for non-smokers. Because the estimated log-odds in model [4] should agree with those in model [1], if we were to fit this model, we should observe the following:

Log-odds for smokers: - d0 = b0+ b1 🡪 d0 = - (b0+ b1)

Log-odds for non-smokers: - d0 – d1 = b0 🡪 d1= - d0 – b0 = b1

Conclusion: **d0 = - (b0+ b1)** and **d1= b1** where b0= -2.0559, b1=0.6368

3. 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.

Part a:

**Method:**

The probabilities for SGA across two groups defined by maternal smoking behavior, i.e. smokers vs. non-smokers were compared using a linear regression model with robust standard error estimates, where the outcome of interest was the probabilities of SGA and the predictor of interest was whether the mother was a smoker. Statistical inference was then based on the Wald statistic computed from the regression result and the Huber-White sandwich estimated standard errors. The two-sided p-value and the 95% confidence interval were constructed using the asymptotic normal distribution for linear regression.

The fitted model was performed using basic “lm” function in R and the robust standard errors covariance matrix, confidence intervals, and test statistics were constructed based on the “sandwich” and “lmtest” packages in R, instead of the “regress” command in the “uwIntroStats” package in R or Stata. Due to the difference in mechanism behind the screen, these results may not match such outputs from Stata completely.

4 observations with missing data on maternal smoking behavior were omitted from this analysis. As a result, there was a total of 751 observations, with 231 smokers and 520 non-smokers.

**Result:**

There were 751 observations with complete data on maternal smoking behavior and whether the babies were SGA, among which 231 were smokers and 520 were non-smokers. Among non-smokers, the proportion of SGA babies delivered was 11.3%. Meanwhile, among smokers, the proportion of SGA babies delivered was 19.5%. With 95% confidence, this observed absolute difference of 8.13% is consistent with a true difference of the probability of SGA babies delivered between smokers and non-smokers being anywhere between 2.34% and 13.93%. This result is statistically significant at the 5% critical threshold with a two-sided p-value of 0.006. As a result, we reject the null hypothesis, in favor of an alternative hypothesis that the probability of delivering SGA babies were associated with maternal smoking behavior.

Part b:

**Answer:**

Using the regression model parameter estimates, the estimated probability of delivering a SGA infant for non-smokers is 11.3%

Thus estimated odds of delivering a SGA infant for non-smokers is 0.113462/(1-0.113462) = 0.128

The estimated probability of delivering a SGA infant for smokers is 19.5%

Thus estimated odds of delivering a SGA infant for smokers is (0.113462+0.081344)/(1-0.113462-0.081344) = 0.242

In Problem 1, I reported the probability of being smokers in each subgroup defined by whether or not the mother delivered an SGA infant. The proportion of smokers in the SGA group is 42.9%, while it is only 28.6% in the non-SGA group. These probabilities are inherently different from what we found in this question – the probability of delivering a SGA infant among non-smokers and smokers, because we stratified on different variables. However, these results somewhat agree with each other in the sense that smokers tend to have higher risk of delivering SGA, or mothers delivering SGA babies were more likely to be smokers. If I were to report the odds ratio in question 1, it should be exactly the same as the observed odds ratio constructed from the estimated odds above.

Part c:

**Answer:**

The linear regression I performed in part a used an indicator for SGA as the outcome of interest and an indicator for SMOKER as the predictor of interest. We essentially fit the following model:

Pr[SGA=1|SMOKER] = b0+ b1SMOKER [1]

Note that because of the binary outcome, the conditional expected value is equivalent to the probability. The estimated intercept of b0= 0.113462 implies that the probability of delivering a SGA infant for non-smokers is 0.113462, while the estimated slope of b1=0.081344 implies that the probability of delivering a SGA infant for smokers is b0+ b1= 0.113462 + 0.081344 = 0.194806

*Part i:*

If instead, an indicator NONSMOKER was created, and a linear regression model of response SGA on predictor NONSMOKER was performed, we then have the following model:

Pr[SGA=1|NONSMOKER] = a0+ a1NONSMOKER [2]

Theoretically, because of linear transformation, we should get similar statistical inference from this model, i.e. statistical significance. The t-statistics and confidence intervals will of course be different if the point estimates are different. However, the underlying robust standard errors should be similar, as this is only linear transformation of variables.

In this model, intercept a0 now represents the probability of SGA for smokers, while (a0 + a1) represents the probability for non-smokers. Because the estimated probabilities in model [2] should agree with those in model [1], if we were to fit this model, we should observe the following:

Probability for smokers: a0 = b0+ b1

Probability for non-smokers: a0 + a1 = b0 🡪 a1= b0 - a0 = -b1

Conclusion: **a0 = b0+ b1** and **a1= - b1** where b0= 0.113462, b1=0.081344

*Part ii:*

Instead, an indicator NOTSGA was created, and a linear regression model of response NOTSGA on predictor SMOKER was performed, we then have the following model:

Pr[NOTSGA=1|SMOKER] = c0+ c1SMOKER

which is equivalent to:

1 - Pr[NOTSGA=0|SMOKER] = c0+ c1SMOKER

Or equivalent to:

Pr[NOTSGA=0|SMOKER] = 1 - c0- c1SMOKER [3]

In this model, (1 - c0) now represents the probability of SGA (NOTSGA=0) for non-smokers, while (1 - c0 - c1) represents the probability for smokers. Because the estimated probabilities in model [3] should agree with those in model [1], if we were to fit this model, we should observe the following:

Probability for non-smokers: 1 - c0= b0🡪 c0 = 1 - b0

Probability for smokers: 1 - c0 - c1 = b0+ b1 🡪 c1 = - b1

Conclusion: **c0 = 1 - b0**and **c1 = - b1**where b0= 0.113462, b1=0.081344

*Part iii:*

Instead, a linear regression model of response NOTSGA on predictor NOTSMOKER was performed, we then have the following model:

Pr[NOTSGA=1|NONSMOKER] = d0+ d1NONSMOKER

which is equivalent to: 1 - Pr[NOTSGA=0|NONSMOKER] = d0+ d1NONSMOKER

equivalent to: Pr[NOTSGA=0|NONSMOKER] = 1 - d0- d1NONSMOKER [4]

In this model, (1 - d0) now represents the probability of SGA (hence NOTSGA=0) for smokers, while (1 - d0 - d1) represents the probability for non-smokers. Because the estimated probabilities in model [4] should agree with those in model [1], if we were to fit this model, we should observe the following:

Probability for smokers: 1 - d0 = b0+ b1 = 0.194806 🡪 d0 = 1 - (b0+ b1)

Probability for non-smokers: 1 - d0 – d1 = b0 🡪 d1= (1 - d0) – b0 = b1

Conclusion: **d0 = 1 - (b0+ b1)** and **d1= b1** where b0= 0.113462, b1=0.081344

4. 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.

Part a:

**Method:**

The probabilities for SGA across two groups defined by maternal smoking behavior, i.e. smokers vs. non-smokers were compared using a Poisson regression model with robust standard error estimates, where the outcome of interest was the probabilities of SGA and the predictor of interest whether the mother was a smoker. Statistical inference was then based on the Wald statistic computed from the regression result and the Huber-White sandwich estimated standard errors. The two-sided p-value and the 95% confidence interval were constructed using the asymptotic normal distribution for Poisson regression.

The fitted model was performed using basic “glm” function in R and the robust standard errors covariance matrix, confidence intervals, and test statistics were constructed based on the “sandwich” and “lmtest” packages in R, instead of the “regress” command in the “uwIntroStats” package in R or Stata. Due to the difference in mechanism behind the screen, these results may not match such outputs from Stata completely.

4 observations with missing data on maternal smoking behavior were omitted from this analysis. As a result, there was a total of 751 observations, with 231 smokers and 520 non-smokers.

**Result:**

There were 751 observations with complete data on maternal smoking behavior and whether the babies were SGA, among which 231 were smokers and 520 were non-smokers. Among non-smokers, the proportion of SGA babies delivered was 11.3%. Meanwhile, among smokers, the proportion of SGA babies delivered was 19.5%. With 95% confidence, this observed rate ratio of 1.717 is consistent with a true probability of SGA babies delivered among smokers being anywhere between 1.203 and 2.451 times higher the probabilities of SGA among non-smokes. This result is statistically significant at the 5% critical threshold with a two-sided p-value of 0.003. As a result, we reject the null hypothesis, in favor of an alternative hypothesis that the probability of delivering SGA babies were associated with maternal smoking behavior.

Part b:

**Answer:**

Using the regression model parameter estimates, the estimated probability of delivering a SGA infant for non-smokers is e-2.17629 = 0.113462 = 11.3%

Thus estimated odds of delivering a SGA infant for non-smokers is 0.113462/(1-0.113462) = 0.128

The estimated probability of delivering a SGA infant for smokers is e-2.17629+0.54054 = 0.1948062 = 19.5%

Thus estimated odds of delivering a SGA infant for smokers is 0.1948062 /(1-0.1948062) = 0.242

In Problem 1, I reported the probability of being smokers in each subgroup defined by whether or not the mother delivered an SGA infant. The proportion of smokers in the SGA group is 42.9%, while it is only 28.6% in the non-SGA group. These probabilities are inherently different from what we found in this question – the probability of delivering a SGA infant among non-smokers and smokers, because we stratified on different variables. However, these results somewhat agree with each other in the sense that smokers tend to have higher risk of delivering SGA, or mothers delivering SGA babies were more likely to be smokers. If I were to report the odds ratio in question 1, it should be exactly the same as the observed odds ratio constructed from the estimated odds above.

Part c:

**Answer:**

The Poisson regression I performed in part a used an indicator for SGA as the outcome of interest and an indicator for SMOKER as the predictor of interest. We essentially fit the following model:

log{Pr[SGA=1|SMOKER]} = b0+ b1SMOKER [1]

The estimated intercept of b0= -2.17629 implies that the log-probability of delivering a SGA infant for non-smokers is -2.17629, while the estimated slope of b1= 0.54054 implies that the log-probability of delivering a SGA infant for smokers is b0+ b1= -2.17629 + 0.54054 = -1.63575

*Part i:*

If instead, an indicator NONSMOKER was created, and a linear regression model of response SGA on predictor NONSMOKER was performed, we then have the following model:

log{Pr[SGA=1|NONSMOKER]} = a0+ a1NONSMOKER [2]

In this model, intercept a0 now represents the log-probability of SGA for smokers, while (a0 + a1) represents the log-probability for non-smokers. Because the estimated log-probabilities in model [2] should agree with those in model [1], if we were to fit this model, we should observe the following:

Log-probability for smokers: a0 = b0+ b1

Log-probability for non-smokers: a0 + a1 = b0 🡪 a1= b0 - a0 = -b1

Conclusion: **a0 = b0+ b1** and **a1= - b1** where b0= -2.17629, b1= 0.54054

*Part ii:*

Instead, an indicator NOTSGA was created, and a linear regression model of response NOTSGA on predictor SMOKER was performed, we then have the following model:

log{Pr[NOTSGA=1|SMOKER]} = c0+ c1SMOKER

🡪 Pr[NOTSGA=1|SMOKER] = exp(c0+ c1SMOKER)

🡪 1 - Pr[SGA=1|SMOKER] = exp(c0+ c1SMOKER)

🡪 Pr[SGA=1|SMOKER] = 1 - exp(c0+ c1SMOKER)

🡪 log{Pr[SGA=1|SMOKER]} = log{1 - exp(c0+ c1SMOKER)} [3]

In this model, log{1 – exp(c0)} now represents the log-probability of SGA for non-smokers, while log{1 – exp(c0 + c1)} represents the log-probability for smokers. Because the estimated log-probabilities in model [3] should agree with those in model [1], if we were to fit this model, we should observe the following:

Log-probability for non-smokers: log{1 – exp(c0)}= b0🡪 c0 = log{1 – exp(b0)}

Log-probability for smokers: log{1 – exp(c0 + c1)} = b0+ b1 🡪 1 – exp(c0 + c1) = exp(b0+ b1) 🡪 c1 = log{1 – exp(b0 + b1)} - c0 = log{1 – exp(b0 + b1)} - log{1 – exp(b0)} = log{[1 – exp(b0 + b1)]/[ 1 – exp(b0)]}

Conclusion: **c0 = log{1 – exp(b0)}**and **c1 = log{[1 – exp(b0 + b1)]/[ 1 – exp(b0)]}** where b0= -2.17629, b1= 0.54054

A nicer representation can be: **exp(c0) = 1 – exp(b0)** and **exp(c0 + c1) = 1 - exp(b0+ b1)**

*Part iii:*

Instead, a linear regression model of response NOTSGA on predictor NOTSMOKER was performed, we then have the following model:

log{Pr[NOTSGA=1|NONSMOKER]} = d0+ d1NONSMOKER

Using similar algebra as above, we have:

log{Pr[SGA=1|NONSMOKER]} = log{1 - exp(d0+ d1NONSMOKER)} [4]

In this model, log{1 – exp(d0)} now represents the log-probability of SGA for smokers, while log{1 – exp(d0 + d1)} represents the log-probability for non-smokers. Because the estimated log-probabilities in model [4] should agree with those in model [1], if we were to fit this model, we should observe the following:

Log-probability for smokers: log{1 – exp(d0)}= b0+ b1🡪 d0 = log{1 – exp(b0+ b1)} or exp(d0) = 1 – exp(b0+ b1)

Log-probability for non-smokers: log{1 – exp(d0 + d1)} = b0 🡪 exp(d0 + d1) = 1 – exp(b0)

Conclusion in nice form: **exp(d0) = 1 – exp(b0+ b1)** and **exp(d0 + d1) = 1 – exp(b0)**, where b0= -2.17629, b1= 0.54054

5. 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**:

If we were to use a simple two-sample comparison of SGA by smoking status (methods covered in BIOST 517/514), we would use a t-test for the difference in the mean SGA, which is equivalent to the proportion of SGA babies, among groups defined by smoking status, smokers and non-smokers.

If a two-sample t-test that does not assume equal variances were to be used in this case, we would get estimated SGA proportions of 0.195 and 0.113 for smokers and non-smokers, respectively. The 95% confidence interval for the difference in mean, or proportions, is from 0.023 to 0.140, with a two-sided p-value of 0.006.

These results match exactly with the linear regression model using robust standard errors in problem 3. The linear regression model essentially estimate exactly the same thing: difference in mean SGA, i.e. proportion of SGA because SGA is a binary variable, across smokers and non-smokers group. The confidence intervals could be slightly different due to the actual method of handling sample size and degrees of freedom. The inference result should be the same. Confidence intervals, p-values, etc would be exactly the same if we were to fit a classical linear regression in problem 3, and then compare it with a t-test that assumes equal variances.

The point estimates from such two-sample t-test match the back-transformed results from the logistic regression in problem 3 and Poisson regression in problem 4. However, untransformed coefficients, confidence intervals, p-values would not be the same because these two models estimate different things. These two models also use z-statistics instead of t-statistics as in the two-sample t-test or the linear model regression.

If we were to use a test for equal probability (prop.test), because of the fact that both variables are binary, we will get the exact same results as the t-test (except for potential slight difference in test statistics and p-values). Therefore, the comparisons between such model and those in 2-4 will be the same as written above.

6. 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.)

a. Evaluate associations using risk difference (RD: difference in probabilities).

**Method**:

The probabilities of delivering SGA babies across groups defined by maternal age were compared using a linear regression model with robust standard error estimates, where the outcome of interest was the probabilities of SGA and the predictor of interest was maternal age. Statistical inference was then based on the Wald statistic computed from the regression result and the Huber-White sandwich estimated standard errors. The two-sided p-value and the 95% confidence interval were constructed using the asymptotic normal distribution for linear regression.

None of the 755 observations had missing data on either SGA or maternal age and thus they were all included in the analysis.

**Results:**

755 observations had complete data on maternal age and whether the babies were SGA. The estimated probability of SGA for a mother of 14 years old (the youngest age in our dataset) was 18.88%. From the linear regression model with robust standard error, 1 year increase in age was corresponding to an absolute decrease of 0.45% in the probability of SGA. With 95% confidence interval, this observed risk difference is consistent with a true absolute decrease in probability anywhere from 0.03% to 0.87% per 1 year increase in age. This result is statistically significant at the 5% critical threshold with a two-sided p-value of 0.036. As a result, we reject the null hypothesis, in favor of an alternative hypothesis that the probability of delivering SGA babies were associated with maternal age.

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

**Method**:

The probabilities of delivering SGA babies across groups defined by maternal age were compared using a Poisson regression model with robust standard error estimates, where the outcome of interest was the probabilities of SGA and the predictor of interest was maternal age. Statistical inference was then based on the Wald statistic computed from the regression result and the Huber-White sandwich estimated standard errors. The two-sided p-value and the 95% confidence interval were constructed using the asymptotic normal distribution for Poisson regression.

None of the 755 observations had missing data on either SGA or maternal age and thus they were all included in the analysis.

**Results:**

755 observations had complete data on maternal age and whether the babies were SGA. The estimated probability of SGA for a mother of 14 years old (the youngest age in our dataset) was 19.83%. From the Poisson regression model with robust standard error, 1 year increase in age was corresponding to a ratio of 0.9662, which is equivalent to a relative decrease of 3.38% in the probability of SGA. With 95% confidence interval, this observed risk ratio is consistent with a true risk ratio anywhere from 0.934 to 0.999 per 1 year increase in age, or a relative decrease of anywhere from 0.06% to 6.60% per 1 year increase in age. This result is statistically significant at the 5% critical threshold with a two-sided p-value of 0.046. As a result, we reject the null hypothesis, in favor of an alternative hypothesis that the probability of delivering SGA babies were associated with maternal age.

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

**Method**:

The odds of delivering SGA babies across groups defined by maternal age were compared using a logistic regression model, where the outcome of interest was the odds of delivering SGA and the predictor of interest was maternal age. Statistical inference was then based on the Wald statistic computed from the regression result and the Huber-White sandwich estimated standard errors. The two-sided p-value and the 95% confidence interval were constructed using the asymptotic normal distribution for logistic regression.

None of the 755 observations had missing data on either SGA or maternal age and thus they were all included in the analysis.

**Results:**

755 observations had complete data on maternal age and whether the babies were SGA. The estimated odds of delivering SGA babies for a mother of 14 years old (the youngest age in our dataset) was 0.244. From the logistic regression model, 1 year difference in age was corresponding to a ratio of 0.961 of the odds of delivering SGA babies between the older-age group and the younger-age group. With 95% confidence, this observed odds ratio is consistent with a true odds ratio anywhere between 0.922 and 1.000. This result is not statistically significant at the 5% critical threshold with a two-sided p-value of 0.055. As a result, we fail to reject the null hypothesis that the probability of delivering SGA babies were not associated with maternal age.

d. 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.

**Answer:**

Using the linear regression model, the estimated probability that a 20-year-old mother would have a SGA infant is 16.069%.

Using the Poisson regression model, the estimated probability that a 20-year-old mother would have a SGA infant is 16.131%. From the estimated coefficients, one needs to exponentiate the linear combination to get to the correct probability, since the Poisson regression estimates the expected log-probability.

Using the logistic regression model, the estimated probability that a 20-year-old mother would have a SGA infant is 16.128%. From the estimated coefficients, one needs to exponentiate the linear combination to get the correct odds of SGA for the 20-year-old mother, and then use the odds-probability relationship to find the correct probability.

7. 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)

a. Sample proportions within each unique age: This can be obtained in Stata using the command egen *varname*= mean(sga), by(age).

**Answer:**



b. 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.)

i. 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.

ii. 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.

iii. 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.

**Answer:**

The estimated probabilities for each age in the data derived from the linear regression model followed a straight line, as we can see from the figure below.

The estimated probabilities for each age in data derived from the Poisson regression model did not follow a straight line. If we were to take the natural log of the fitted values, they would follow a straight line.

The estimated probabilities for each age in data derived from the logistic regression model also did not follow a straight line. If we were to take the natural log of the odds of SGA, they would follow a straight line.



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

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

**Method:**

The odds of delivering SGA babies across groups defined by maternal age were compared using a logistic regression model, where the outcome of interest was the odds of delivering SGA and the predictor of interest was log-transformed maternal age. Statistical inference was then based on the Wald statistic computed from the regression result and the Huber-White sandwich estimated standard errors. The two-sided p-value and the 95% confidence interval were constructed using the asymptotic normal distribution for logistic regression.

None of the 755 observations had missing data on either SGA or maternal age and thus they were all included in the analysis.

**Results:**

755 observations had complete data on maternal age and whether the babies were SGA. The estimated odds of delivering SGA babies for a mother of 14 years old (the youngest age in our dataset) was 0.269. From the logistic regression model, 1 unit difference in log-transformed age (2.72-fold increase in age) was corresponding to a ratio of 0.385 of the odds of delivering SGA babies between the older-age group and the younger-age group. With 95% confidence, this observed odds ratio is consistent with a true odds ratio anywhere between 0.142 and 1.026 per 1 unit difference in log-transformed age. This result is not statistically significant at the 5% critical threshold with a two-sided p-value of 0.058. As a result, we fail to reject the null hypothesis that the probability of delivering SGA babies were not associated with maternal age.

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

**Answer:**

If we had a scientific reason or belief that age would have a multiplicative effect on the chance of delivering an SGA babies, then it would be reasonable to perform such analysis rather than the analysis in problem 6c. In reality, one could assume so, since aging potentially has multiplicative effect on biological processes in the human body. It might also be of interest to reduce the heteroscedasticity level by log-transforming age.