Biost 536 / Epi 536
Categorical Data Analysis in Epidemiology

Midterm Examination Key
October 16, 2014

Name: _____

Instructions: This exam is closed book, closed notes. You have 110 minutes. You may not use any device that is capable of accessing the internet.

Please provide concise answers to all questions. Rambling answers touching on topics not directly relevant to the question will tend to count against you. Nearly telegraphic writing style is permissible.

NOTE: When you need to make calculations, always use at least four significant digits in your intermediate calculations, and report at least three significant digits. (Example: 1.045 and 0.0001234 and 1234000 each have four significant digits.) Give some indication of how you were calculating your answer. (If you give the wrong answer, but I can determine where you went wrong, you may get partial credit.)

If you come to a problem that you believe cannot be answered without making additional assumptions, clearly state the reasonable assumptions that you make, and proceed.

Please adhere to and sign the following pledge. Should you be unable to truthfully sign the pledge for any reason, turn in your paper unsigned and discuss the circumstances with the instructor.

PLEDGE:
On my honor, I have neither given nor received unauthorized aid on this examination:

Signed:_____

Problems 1-3 consider three different types of study design that might be used to investigate associations between a particular childhood vaccination and autism. In each question you are asked to identify the type of study design and to identify the valid statistical inference that can be made in several alternative analyses.

Problems 4-10 deal with a subset of data from an observational study of pregnancy outcomes in South Africa. The appendices contain results from selected analyses:

Appendix A : Description of the variables and descriptive statistics **(problems 4 through 10))**
Appendix B : Analyses of "Small for Gestational Age (SGA)" by smoking, nulliparity **(problems 1 through 9)**
Appendix C : Stratified and linear regression analysis of SGA by smoking and parity **(problem 10)**

Problems 1-3 are focused on the material presented in Lecture 2, Slides 44-55. As discussed in class, there was apparently some confusion about the wording of this problem relating to their being multiple correct answers for each questions (the intent was that students were to mark every correct answer), the wording prevalence, risk, rates, and "due to", and whether the questions should explicitly mention the use of the slope and its standard error. I present here reworded questions that I will contend should be unambiguous in those regards. Revised wording is displayed in blue font.

Some general comments:

- *With regression analyses, we are describing the conditional distribution of response within groups defined by the predictors. Hence, with binary variables*
  - *in linear regressions we model the probability of Y=1 within groups defined by X*
    - *the intercept is the probability that Y=1 within a group having X=0*
    - *the slope is the difference in probabilities that Y=1, comparing a group with X=1 to a group with X=0*
  - *in Poisson regressions we model the log probability of Y=1 within groups defined by X*
    - *the intercept is the log probability that Y=1 within a group having X=0. The exponentiated intercept thus estimates the probability that Y=1 within a group having X=0.*
    - *the slope is the difference in log probabilities that Y=1, comparing a group with X=1 to a group with X=0. The exponentiated slope estimates the ratio of probabilities that Y=1, comparing a group with X=1 to a group with X=0.*
  - *in logistic regressions we model the log odds that Y=1 within groups defined by X*
    - *the intercept is the log odds that Y=1 within a group having X=0. The exponentiated intercept thus estimates the odds that Y=1 within a group having X=0. (If we so desire, we can estimate the probability that Y=1 by probability = odds / (1 + odds) .)*
    - *the slope is the difference in log odds that Y=1, comparing a group with X=1 to a group with X=0. The exponentiated slope estimates the ratio of odds (the odds ratio) that Y=1, comparing a group with X=1 to a group with X=0. (It is not in general possible to transform an odds ratio to a probability ratio or probability difference, without knowing the "baseline" probability or odds in the reference group. If that "baseline" probability is sufficiently small, however, we can know that the probability ratio is approximately the odds ratio.)*
    - ***VERY SPECIAL CASE**: Owing to an invariance property of the odds ratio, we know that*
      $$Odds(Y=1|X=1) / Odd(Y=1|X=0) = Odds(X=1|Y=1) / Odds(X=1|Y=0)$$
      so the slope from a logistic regression with binary random variables can be interpreted as either odds ratio.
- *In order to have interpretable parameter estimates from a particular regression model, it is important that the sampling scheme support the estimation of the relevant quantities. Basically, in order to estimate the probability distribution conditional on the value of X=x, it is important that the sample sizes not be constrained within subgroups of that group having X=x. Hence*
  - *We can only estimate the overall probability that Y=y if either*
    - *no sample sizes are constrained ("Poisson sampling"), or*
    - *only the overall sample size is constrained ("multinomial sampling").*
  - *We can estimate the conditional probability that Y=y|X=x if either*
    - *the sample size within the group having X=x is unconstrained (such as is true with either Poisson or multinomial sampling), or*
    - *within the group having X=x, only the total sample size within that group is constrained ("binomial sampling")*
- *When considering two binary variables measuring a putative risk factor (vaccination in this problem) and a putative outcome (autism in this problem), we characterize study designs as*
  - *"cross-sectional sampling": we used either Poisson or multinomial sampling*
    - *In this problem, we can always interpret regression parameter estimates when regressing autism (response) on vaccination (predictor), or when regressing vaccination (response) on autism (predictor)*
  - *"cohort design": we used binomial sampling, constraining the sample sizes only within groups defined by one or more putative risk factors or "causes"*
    - *In this problem, we can always interpret regression parameter estimates when regressing autism (response) on vaccination (predictor), but we can also interpret the slope (odds ratio) but not the intercept from a logistic regression of vaccination (response) on autism (predictor)*
  - *"case-control design": we used binomial sampling, constraining the sample sizes only within groups defined by a putative "outcome" or "effect"*

  - *In this problem, we can always interpret regression parameter estimates when regressing vaccination (response) on autism (predictor), but we can also interpret the slope (odds ratio) but not the intercept from a logistic regression of autism (response) on vaccination (predictor)*
- ***SPECIAL NOTE:*** *When only considering hypothesis tests of association between two variables, we can interpret the p values from all regressions irrespective of the sampling scheme.*
  - *In linear regression, this fact is easily seen by considering the equivalences between tests for slopes in linear regression and tests for non-zero correlation, along with the symmetry of the definition of correlation.*
  - *In logistic regression, this fact is easily seen by considering the invariance property of the odds ratios*

1. (15 points) We sample 100,000 kindergarten students and characterize each student with respect to whether they are diagnosed as autistic and whether they were received a particular vaccination.
   a. What term would you use to describe this study?
   Ans: Cross-sectional study *(In this study design, we only constrained the total sample size.)*

   b. Suppose we use study results to perform a linear regression of autism (response) on vaccine use (predictor). Mark the inference that would be **valid**:
      __X__ Estimate a prevalence of autism from the intercept
      __X__ Estimate a difference in autism prevalence by vaccination from the slope
      __X__ Test for an association between autism and vaccination using the slope and its SE

   c. Suppose we use study results to perform a linear regression of vaccine use (response) on autism (predictor). Mark (place an "X" by) the inference that would be **valid**:
      __X__ Estimate a prevalence of vaccination from the intercept
      __X__ Estimate a difference in vaccination prevalence by autism diagnosis from the slope
      __X__ Test for an association between autism and vaccination using the slope and its SE

   d. Suppose we use study results to perform a Poisson regression of autism (response) on vaccine use (predictor). Mark the inference that would be **valid**:
      __X__ Estimate a prevalence of autism from the intercept
      __X__ Estimate a ratio of autism risk due to vaccination from the slope
      __X__ Test for an association between autism and vaccination using the slope and its SE

   e. Suppose we use study results to perform a Poisson regression of vaccine use (response) on autism (predictor). Mark the inference that would be **valid**:
      __X__ Estimate a prevalence of vaccination from the intercept
      __X__ Estimate a ratio of vaccination rates by autism diagnosis from the slope
      __X__ Test for an association between autism and vaccination using the slope and its SE

   f. Suppose we use study results to perform a logistic regression of autism (response) on vaccine use (predictor). Mark the inference that would be **valid**:
      __X__ Estimate an odds of prevalent autism from the intercept
      __X__ Estimate an odds ratio for prevalence of autism by vaccination from the slope
      __X__ Test for an association between autism and vaccination using the slope and its SE

   g. Suppose we use study results to perform a logistic regression of vaccine use (response) on autism (predictor). Mark the inference that would be **valid**:
      __X__ Estimate an odds of vaccination from the intercept
      __X__ Estimate an odds ratio for prevalence of vaccination by autism diagnosis from the slope
      __X__ Test for an association between autism and vaccination using the slope and its SE

2. (15 points) We sample 10,000 babies who were vaccinated and 10,000 babies who were not vaccinated and follow them until their 6th birthday to assess whether they are diagnosed as autistic.

   a. What term would you use to describe this study?

   Ans: Cohort study *(In this study design, we constrained the sample size within groups defined by vaccination.)*

   b. Suppose we use study results to perform a linear regression of autism (response) on vaccine use (predictor). Mark the inference that would be **valid**:

      __X__ Estimate a prevalence of autism from the intercept
      __X__ Estimate a difference in autism prevalence by vaccination from the slope
      __X__ Test for an association between autism and vaccination using the slope and its SE

   c. Suppose we use study results to perform a linear regression of vaccine use (response) on autism (predictor). Mark (place an "X" by) the inference that would be **valid**:

      _____ Estimate a prevalence of vaccination from the intercept
      _____ Estimate a difference in vaccination prevalence by autism diagnosis from the slope
      __X__ Test for an association between autism and vaccination using the slope and its SE

   d. Suppose we use study results to perform a Poisson regression of autism (response) on vaccine use (predictor). Mark the inference that would be **valid**:

      __X__ Estimate a prevalence of autism from the intercept
      __X__ Estimate a ratio of autism risk due to vaccination from the slope
      __X__ Test for an association between autism and vaccination using the slope and its SE

   e. Suppose we use study results to perform a Poisson regression of vaccine use (response) on autism (predictor). Mark the inference that would be **valid**:

      _____ Estimate a prevalence of vaccination from the intercept
      _____ Estimate a ratio of vaccination rates by autism diagnosis from the slope
      __X__ Test for an association between autism and vaccination using the slope and its SE

   f. Suppose we use study results to perform a logistic regression of autism (response) on vaccine use (predictor). Mark the inference that would be **valid**:

      __X__ Estimate an odds of prevalent autism from the intercept
      __X__ Estimate an odds ratio for prevalence of autism by vaccination from the slope
      __X__ Test for an association between autism and vaccination using the slope and its SE

   g. Suppose we use study results to perform a logistic regression of vaccine use (response) on autism (predictor). Mark the inference that would be **valid**:

      _____ Estimate an odds of vaccination from the intercept
      __X__ Estimate an odds ratio for prevalence of vaccination by autism diagnosis from the slope
      __X__ Test for an association between autism and vaccination using the slope and its SE

3. (15 points) We sample 1,000 autistic 5 year olds and 1,000 non-autistic 5 year olds and review their medical records to assess whether they had received the particular vaccination.
   a. What term would you use to describe this study?
   Ans: Case-control study *(In this study design, we constrained sample size within groups defined by autism.)*

   b. Suppose we use study results to perform a linear regression of autism (response) on vaccine use (predictor). Mark the inference that would be **valid**:
      ____ Estimate a prevalence of autism from the intercept
      ____ Estimate a difference in autism prevalence by vaccination from the slope
      __X__ Test for an association between autism and vaccination using the slope and its SE

   c. Suppose we use study results to perform a linear regression of vaccine use (response) on autism (predictor). Mark (place an "X" by) the inference that would be **valid**:
      __X__ Estimate a prevalence of vaccination from the intercept
      __X__ Estimate a difference in vaccination prevalence by autism diagnosis from the slope
      __X__ Test for an association between autism and vaccination using the slope and its SE

   d. Suppose we use study results to perform a Poisson regression of autism (response) on vaccine use (predictor). Mark the inference that would be **valid**:
      ____ Estimate a prevalence of autism from the intercept
      ____ Estimate a ratio of autism risk due to vaccination from the slope
      __X__ Test for an association between autism and vaccination using the slope and its SE

   e. Suppose we use study results to perform a Poisson regression of vaccine use (response) on autism (predictor). Mark the inference that would be **valid**:
      __X__ Estimate a prevalence of vaccination from the intercept
      __X__ Estimate a ratio of vaccination rates by autism diagnosis from the slope
      __X__ Test for an association between autism and vaccination using the slope and its SE

   f. Suppose we use study results to perform a logistic regression of autism (response) on vaccine use (predictor). Mark the inference that would be **valid**:
      ____ Estimate an odds of prevalent autism from the intercept
      __X__ Estimate an odds ratio for prevalence of autism by vaccination from the slope
      __X__ Test for an association between autism and vaccination using the slope and its SE

   g. Suppose we use study results to perform a logistic regression of vaccine use (response) on autism (predictor). Mark the inference that would be **valid**:
      __X__ Estimate an odds of vaccination from the intercept
      __X__ Estimate an odds ratio for prevalence of vaccination by autism diagnosis from the slope
      __X__ Test for an association between autism and vaccination using the slope and its SE

4. **Appendix B** provides data on the prevalence of "small for gestational age" (SGA) at birth within groups defined by smoking status and nulliparity. Suppose we let p be the prevalence of SGA at birth, and we are interested in performing **linear regression** involving main effects for ***smoker*** and ***nulliparous*** and multiplicative interaction ***smok_nullip = smoker*** x ***nulliparous***.

$$p = \beta_0 + \beta_1 \times smoker + \beta_2 \times nulliparous + \beta_3 \times smok\_nullip$$

a. (5 points) Can you calculate the estimated ***intercept*** that would be obtained from that regression? If so provide the estimate. If not, explain why not.

Ans: 0.1026 *(a saturated model, so the sample proportion among nonsmoking women with prior live birth)*

b. (5 points) Can you calculate the estimated slope for ***smoker*** that would be obtained from that regression? If so provide the estimate. If not, explain why not.

Ans: .1622 - .1026 = .0596 *(a saturated model, so the difference of sample proportions for smokers minus nonmoskers among the women with prior live birth)*

c. (5 points) Can you calculate the estimated slope for ***nulliparous*** that would be obtained from that regression? If so provide the estimate. If not, explain why not.

Ans: .1298 - .1026 = .0272 *(a saturated model, so the difference of sample proportions for nulliparous women minus that for women with prior live birth among the nonsmokers)*

d. (5 points) Can you calculate the estimated slope for the interaction ***smok_nullip*** that would be obtained from that regression? If so provide the estimate. If not, explain why not.

Ans: (.2530 - .1298) – (.1622 - .1026) = .0636 *(a saturated model, so the "difference of differences" comparing the difference of sample proportions for smokers minus nonsmokers among the women with prior live births to the difference of sample proportions for smokers minus nonsmokers among the nulliparous women. The same answer could be obtained as the "difference of differences" comparing the difference of sample proportions for nulliparous women minus that for women with prior live birth among smokers to the difference of sample proportions for nulliparous women minus that for women with prior live birth among nonsmokers.)*

5. **Appendix B** provides data on the prevalence of "small for gestational age" (SGA) at birth within groups defined by smoking status and nulliparity. Suppose we let p be the prevalence of SGA at birth, and we perform **Poisson regression** involving main effects for ***smoker*** and ***nulliparous*** and multiplicative interaction ***smok_nullip = smoker*** x ***nulliparous***.

$$\log(p) = \beta_0 + \beta_1 \times smo\,\mathrm{ker} + \beta_2 \times nulliparous + \beta_3 \times smok\_nullip$$

a. (5 points) Can you calculate the estimated ***intercept*** that would be obtained from that regression? If so provide the estimate. If not, explain why not.

Ans: log(.1026) = -2.277 *(a saturated model, so the log sample proportion…)*

b. (5 points) Can you calculate the estimated slope for ***smoker*** that would be obtained from that regression? If so provide the estimate. If not, explain why not.

Ans: log(.1622 / .1026) = .4580

c. (5 points) Can you calculate the estimated slope for ***nulliparous*** that would be obtained from that regression? If so provide the estimate. If not, explain why not.

Ans: log(.1298 / .1026) = .2352

d. (5 points) Can you calculate the estimated slope for the interaction ***smok_nullip*** that would be obtained from that regression? If so provide the estimate. If not, explain why not.

Ans: log(.2530 / .1298) – log (.1622 / .1026) = .2094

6. **Appendix B** provides data on the prevalence of "small for gestational age" (SGA) at birth within groups defined by smoking status and nulliparity. Suppose we let p be the prevalence of SGA at birth, and we perform **logistic regression** involving main effects for ***smoker*** and ***nulliparous*** and multiplicative interaction ***smok_nullip = smoker*** x ***nulliparous***.

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \times smoker + \beta_2 \times nulliparous + \beta_3 \times smok\_nullip$$

a. (5 points) Can you calculate the estimated ***intercept*** that would be obtained from that regression? If so provide the estimate. If not, explain why not.

Ans: log (.1026 / .8924) = -2.169 *(a saturated model, so the log odds among nonsmoking women with prior live birth)*

b. (5 points) Can you calculate the estimated slope for ***smoker*** that would be obtained from that regression? If so provide the estimate. If not, explain why not.

Ans: log (.1622 / .8378) – log (.1026 / .8974) = .5267 *(a saturated model, so the log odds ratio…)*

c. (5 points) Can you calculate the estimated slope for ***nulliparous*** that would be obtained from that regression? If so provide the estimate. If not, explain why not.

Ans: log (.1298 / .8702) – log (.1026 / .8974) = .2664

d. (5 points) Can you calculate the estimated slope for the interaction ***smok_nullip*** that would be obtained from that regression? If so provide the estimate. If not, explain why not.

Ans: log (.2530 / .7470) – log (.1298 / .8702) – log (.1622 / .8378) + log (.1026 / .8974) = .2933

7. (5 points) **Appendix B** provides data on the prevalence of "small for gestational age" (SGA) at birth within groups defined by smoking status and nulliparity. Suppose we let p be the prevalence of SGA at birth, and we perform **logistic regression** involving main effects for ***smoker*** and ***nulliparous*** (without the interaction). Can you similarly calculate what the regression coefficients would be for this model? If so provide the estimates. If not, explain why not.

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \times smoker + \beta_2 \times nulliparous$$

Ans: No, because this is not a saturated model. *(There are four distinct groups, but only 3 parameters)*

8. Again using **Appendix B**, does nulliparity confound the estimation of an association between maternal smoking and prevalence of "small for gestational age"?
    a. (5 points) Answer the question assuming you are using risk difference as a measure of association.

    Ans: It is reasonable that nulliparous women might be at risk for higher SGA (perhaps in part due to age and body size). Supporting evidence for this is that among nonsmokers 32 / 280 = .1026 of women with prior live birth have SGA, while 27 / 181 = .1298 of nulliparous have SGA.

    Smoking does appear to be associated with parity in the sample: 148 / 460 = .322 of women with prior live birth smoke, while 83 / 291 = .285 of nulliparous women smoke.

    Thus there does seem to be evidence of confounding.

    b. (5 points) Answer the question assuming you are using risk ratio as a measure of association.

    Ans: Same as in part a.

    c. (5 points) Answer the question assuming you are using odds ratio as a measure of association.

    Ans: Same as in part a.

9. Again using **Appendix B**, does nulliparity modify any association between maternal smoking and "small for gestational age"?
    a. (5 points) Answer the question assuming you are using risk difference as a measure of association.

    Ans: Yes, based on the estimated interaction term of 0.0626 in problem 4d.

    b. (5 points) Answer the question assuming you are using risk ratio as a measure of association.

    Ans: Yes, based on the estimated interaction term of 02093 in problem 5d.

    c. (5 points) Answer the question assuming you are using odds ratio as a measure of association.

    Ans: Yes, based on the estimated interaction term of 0.2933 in problem 6d.

10. **Appendix C** contains the results of stratified and regression analyses investigating the parity adjusted association between maternal smoking and prevalence of "small for gestational age" using risk difference (RD) as a measure of association.
    a. (5 points) What scientific interpretation can you place on the choice of weights for the stratified analysis?

    Ans: Impact if smokers became nonsmokers.

    b. (5 points) In the **stratified analysis**, can you provide a p value for an association between maternal smoking and "small for gestational age" after adjustment for parity? If so, do so. If not, explain why not.

    Ans: Using the CI: The 99.52% CI for the standardized risk difference just barely excludes 0, so two-sided P value is approximately 0.0048.

c. (5 points) Using the **linear regression analysis**, provide full statistical inference for any parity adjusted association between maternal smoking and "small for gestational age". Be sure to provide an interpretation of the confidence interval.

Ans: The risk of SGA in smokers is estimated an absolute 8.4% higher than in nonsmokers of similar parity. Based on 95% CI, such a statistically significant observation (two-sided P = 0.005) is not unusual if the true RD were between 2.6% and 14.2% absolute higher.

d. (5 points) How might you explain any potential differences between the inference you might obtain from the stratified and regression analyses when using RD as the measure of association? Include issues that might not have arisen in this particular analysis, but could potentially arise. (But please be brief. Just write enough so that I know that you know the issues.)

Ans: Differences between the analysis models include

- **parity is modeled continuously in the regression analysis, but as "dummy variables" in the stratified analysis;**
- **no interaction is modeled in the regression analysis, but a stratified analysis necessarily models the interaction;**
- **the regression analysis would weight the observations according to "efficiency weights" assuming a constant effect across strata, while the stratified analysis used "importance weights" based on the distribution of parity across the smoking population.**

e. (5 points) What additional issues might you need to consider when assessing differences between stratified analyses and regression analyses when using risk ratio (RR) or odds ratios (OR) as the measure of association? (Again, brevity is highly desired.)

Ans: The role of effect modification can be different across the models, as the definition of effect modification must consider the measure of association.

For Poisson regression, we are basically considering the weighted geometric means of log risk ratios, while the stratified risk ratio will first take weighted arithmetic means of the risk across parity groups separately for smokers and nonsmokers, and then take the ratio of those arithmetic means.

f. (5 points) How different would you expect to be the quantification of parity adjusted association between smoking and SGA based on logistic regression from that based on Poisson regression in this data? Why? *(Note: I do not provide these analyses in the Appendices.)*

Ans: In the nonsmoking women with prior live births, the probability of SGA is estimated to be .1026, corresponding to an odds of SGA of .1143. Hence, this is not sufficiently rare event probability to make the OR approximate the RR to a high level of accuracy. (It may not be too bad, however.)

Grade distribution:

Highest achieved	**:**	**155**	**(of 160 possible)**
Mean (SD)	**:**	**130 (21.2)**	
80th percentile	**:**	**145.2**	
50th percentile	**:**	**135.5**	
80th percentile	**:**	**116.6**	

APPENDIX A: **Description of variables and descriptive statistics**

These data come from a South African observational study of maternal risk factors and pregnancy outcomes. We are particularly interested in babies that are "Small for Gestational Age" as a sign of intrauterine growth restriction. Risk factors that we consider are maternal smoking during pregnancy and maternal "parity" (number of prior live births).

This exam considers the following variables (all measured at time of study enrolment) on a subset of 751 subjects from that study.

smoker: indicator that the mother smoked during pregnancy (**0**= no, **1**= yes)
parity: number of prior live births (**0**= none, **1**= one, **2**= two or more)
nulliparous: indicator that this was the mother's first pregnancy (so *parity* = 0*)* (**0**= no, **1**= yes)
sga*:* indicator that the baby was small for gestational age at birth (**0**= no, **1**= yes)

The following tables present crosstabulation of prevalence of "small for gestational age" by maternal smoking and parity.

```
. table sga parity smoker, col row
```

	smoker and parity							
	0				1			
sga	0	1	2	Total	0	1	2	Total
0	181	148	132	461	62	59	65	186
1	27	18	14	59	21	14	10	45
Total	208	166	146	520	83	73	75	231

APPENDIX B: Prevalence of "small for gestational age" (SGA) by maternal smoking overall and within groups defined by *nulliparous*

```
. tabulate smoker sga, row chi
```

Key
frequency
row percentage

smoker	sga 0	sga 1	Total
0	461	59	520
	88.65	11.35	100.00
1	186	45	231
	80.52	19.48	100.00
Total	647	104	751
	86.15	13.85	100.00

Pearson chi2(1) = 8.8708 Pr = 0.003

```
. bysort nulliparous: tabulate smoker sga, row chi
```

-> nulliparous = 0

Key
frequency
row percentage

smoker	sga 0	sga 1	Total
0	280	32	312
	89.74	10.26	100.00
1	124	24	148
	83.78	16.22	100.00
Total	404	56	460
	87.83	12.17	100.00

Pearson chi2(1) = 3.3348 Pr = 0.068

-> nulliparous = 1

Key
frequency
row percentage

smoker	sga 0	sga 1	Total
0	181	27	208
	87.02	12.98	100.00
1	62	21	83
	74.70	25.30	100.00
Total	243	48	291
	83.51	16.49	100.00

Pearson chi2(1) = 6.5379 Pr = 0.011

<u>APPENDIX C</u>: Parity adjusted analyses of the association between maternal smoking and prevalence of "small for gestational age" using risk difference (RD) as the measure of association.

```
. cs sga smoker, by(parity) rd istandard

          parity |       RD       [95% Conf. Interval]       Weight
-----------------+------------------------------------------------
               0 |    .1232044     .0191205   .2272882           83
               1 |    .0833471    -.0186028    .185297           73
               2 |    .0374429    -.0531096   .1279955           75
-----------------+------------------------------------------------
           Crude |    .0813437      .023451   .1392363
 I. Standardized |    .0827641       .02531   .1402183

. cs sga smoker, by(parity) rd istandard level(99.52)

          parity |       RD    [99.52% Conf. Interval]       Weight
-----------------+------------------------------------------------
               0 |    .1232044      -.02656   .2729688           83
               1 |    .0833471    -.0633468    .230041           73
               2 |    .0374429    -.0928516   .1677374           75
-----------------+------------------------------------------------
           Crude |    .0813437     -.001957   .1646443
 I. Standardized |    .0827641     .0000943    .165434

. regress sga smoker parity, robust

Linear regression                               Number of obs =     751
                                                F(  2,   748) =    5.46
                                                Prob > F      =  0.0044
                                                R-squared     =  0.0171
                                                Root MSE      =  .34313

------------------------------------------------------------------------------
             |               Robust
         sga |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      smoker |   .0839281   .0295777     2.84   0.005     .0258629    .1419933
      parity |  -.0305493   .0151863    -2.01   0.045     -.060362   -.0007365
       _cons |   .1403684   .0200977     6.98   0.000     .1009139    .1798229
```