Biost 536 / Epi 536
Categorical Data Analysis in Epidemiology

Midterm Examination
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)**

1. 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?

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

        _____ Estimate prevalence of autism from the intercept

        _____ Estimate difference in autism risk due to vaccination from the slope

        _____ Test for an association between autism and vaccination using the slope

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

        _____ Estimate prevalence of vaccination from the intercept

        _____ Estimate difference in vaccination rates by autism diagnosis from the slope

        _____ Test for an association between autism and vaccination using the slope

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

        _____ Estimate prevalence of autism from the intercept

        _____ Estimate ratio of autism risk due to vaccination from the slope

        _____ Test for an association between autism and vaccination using the slope

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

        _____ Estimate prevalence of vaccination from the intercept

        _____ Estimate ratio of vaccination rates by autism diagnosis from the slope

        _____ Test for an association between autism and vaccination using the slope

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

        _____ Estimate odds of autism from the intercept

        _____ Estimate odds ratio for autism due to vaccination from the slope

        _____ Test for an association between autism and vaccination using the slope

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

        _____ Estimate odds of vaccination from the intercept

        _____ Estimate odds ratio for vaccination by autism diagnosis from the slope

        _____ Test for an association between autism and vaccination using the slope

2. 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?

    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 prevalence of autism from the intercept

        _____ Estimate difference in autism risk due to vaccination from the slope

        _____ Test for an association between autism and vaccination using the slope

    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 prevalence of vaccination from the intercept

        _____ Estimate difference in vaccination rates by autism diagnosis from the slope

        _____ Test for an association between autism and vaccination using the slope

    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 prevalence of autism from the intercept

        _____ Estimate ratio of autism risk due to vaccination from the slope

        _____ Test for an association between autism and vaccination using the slope

    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 prevalence of vaccination from the intercept

        _____ Estimate ratio of vaccination rates by autism diagnosis from the slope

        _____ Test for an association between autism and vaccination using the slope

    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 odds of autism from the intercept

        _____ Estimate odds ratio for autism due to vaccination from the slope

        _____ Test for an association between autism and vaccination using the slope

    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 odds of vaccination from the intercept

        _____ Estimate odds ratio for vaccination by autism diagnosis from the slope

        _____ Test for an association between autism and vaccination using the slope

3. 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?

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

        _____ Estimate prevalence of autism from the intercept

        _____ Estimate difference in autism risk due to vaccination from the slope

        _____ Test for an association between autism and vaccination using the slope

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

        _____ Estimate prevalence of vaccination from the intercept

        _____ Estimate difference in vaccination rates by autism diagnosis from the slope

        _____ Test for an association between autism and vaccination using the slope

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

        _____ Estimate prevalence of autism from the intercept

        _____ Estimate ratio of autism risk due to vaccination from the slope

        _____ Test for an association between autism and vaccination using the slope

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

        _____ Estimate prevalence of vaccination from the intercept

        _____ Estimate ratio of vaccination rates by autism diagnosis from the slope

        _____ Test for an association between autism and vaccination using the slope

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

        _____ Estimate odds of autism from the intercept

        _____ Estimate odds ratio for autism due to vaccination from the slope

        _____ Test for an association between autism and vaccination using the slope

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

        _____ Estimate odds of vaccination from the intercept

        _____ Estimate odds ratio for vaccination by autism diagnosis from the slope

        _____ Test for an association between autism and vaccination using the slope

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***.

$$p = \beta_0 + \beta_1 \times smo\,\text{ker} + \beta_2 \times nulliparous + \beta_3 \times smok\_nullip$$

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

   b. 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.

   c. 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.

   d. 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.

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***.

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

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

   b. 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.

   c. 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.

   d. 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.

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

   b. 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.

   c. 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.

   d. 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.

7. **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.

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

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

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

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

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

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

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

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. What scientific interpretation can you place on the choice of weights for the stratified analysis?

b. 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.

c. 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.

d. 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.)

e. 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.)

f. 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.)*

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
```