

Biost 536 / Epi 536
Categorical Data Analysis in Epidemiology

Second Midterm Examination Key
November 13, 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 analyses of the *prevalence of infarct-like lesions on MRI as a function of smoking behaviors*. These analyses use the dataset assigned for the project, and thus these analyses are similar to some of those you might be considering for the project.

Appendix A : Description of the variables and descriptive statistics (**problems 1 - 4**)

Appendix B : Unadjusted analyses of infarct-like lesions by age (**problem 1**)

Appendix C : Unadjusted analyses of infarct-like lesions by smoking history (**problem 2**)

Appendix C : Analyses of infarct-like lesions by smoking history adjusted for age (**problem 3**)

Appendix D : Analyses of infarct-like lesions by smoking history, age, and stroke (**problem 4**)

Problem 5 asks you to consider the impact of multiple comparisons on the generalizability of statistical inference.

1. **Appendix B** provides results from three unadjusted analyses of the prevalence of infarct-like lesions by age.

- a. (9 points) For each of the three models, can the model be used to **test for an association between prevalence of infarct-like lesions and age**? If so, can you tell what it is from the output? If so, provide the P value. If not, explain what further information you would need from the model. (Circle “yes” or “no” and either fill in the blank or supply just enough of an answer to let me know that you know what you would do.)

Model B1: Can be used to test for an association?: Yes No

If so, is the relevant P value is available on output?: Yes (P: <.0005) No

If it can be used, but additional output is needed, what analysis do you need?

(P value from slope for age or the overall chi square test)

Model B2: Can be used to test for an association?: Yes No

If so, is the relevant P value is available on output?: Yes (P: <.0005) No

If it can be used, but additional output is needed, what analysis do you need?

(P value from the overall chi square test)

Model B3: Can be used to test for an association?: Yes No

If so, is the relevant P value is available on output?: Yes (P: <.0005) No

If it can be used, but additional output is needed, what analysis do you need?

(P value from the overall chi square test)

- b. (9 points) For each of the three models, can the model be used to **test whether any association between prevalence of infarct-like lesions and age is nonlinear**? If so, can you tell what it is from the output? If so, provide the P value. If not, explain what further information you would need from the model. (Circle “yes” or “no” and either fill in the blank or supply just enough of an answer to let me know that you know what you would do.)

Model B1: Can be used to test for a nonlinear association?: Yes No

If so, is the relevant P value is available on output?: Yes (P: _____) No

If it can be used, but additional output is needed, what analysis do you need?

(This model estimates the best fitting linear term providing no info about nonlinearity)

Model B2: Can be used to test for a nonlinear association?: Yes No

If so, is the relevant P value is available on output?: Yes (P: 0.555) No

If it can be used, but additional output is needed, what analysis do you need?

(P value from the slope for the squared age term)

Model B3: Can be used to test for a nonlinear association?: **Yes** No

If so, is the relevant P value is available on output?: Yes (P:_____) **No**

If it can be used, but additional output is needed, what analysis do you need?

Ans: Would need to perform a post estimation test that all slopes were equal.

(In Stata: test sage65 = sage72 = sage80)

- c. (9 points) For each of the three models, can the model be used to **test whether any association between prevalence of infarct-like lesions and age is U-shaped?** If so, can you tell what it is from the output? If so, provide the P value. If not, explain what further information you would need from the model. (Circle “yes” or “no” and either fill in the blank or supply just enough of an answer to let me know that you know what you would do.)

Model B1: Can be used to test for a U-shaped association?: Yes **No**

If so, is the relevant P value is available on output?: Yes (P:_____) No

If it can be used, but additional output is needed, what analysis do you need?

(This model estimates the best fitting linear term providing no info about U-shape)

Model B2: Can be used to test for a U-shaped association?: Yes **No**

If so, is the relevant P value is available on output?: Yes (P:_____) No

If it can be used, but additional output is needed, what analysis do you need?

(This model estimates the best fitting U-shape providing no info about whether it had to be)

Model B3: Can be used to test for a U-shaped association?: Yes No

If so, is the relevant P value is available on output?: Yes (P: > 0.5) No

If it can be used, but additional output is needed, what analysis do you need?

Ans: Need to see statistically significant slopes in opposite direction. All slopes were in same direction, so we know $P > 0.5$ at least. *(I gave credit if you noted direction or nonsignificance.)*

2. **Appendix C** provides results from three unadjusted analyses of the prevalence of infarct-like lesions by history of smoking behavior.

- a. (9 points) For each of the three models, provide a scientifically useful interpretation of the estimated *intercept*.

Model C1: Prevalence of infarcts among never smokers estimated to be 0.312.

Model C2: Prevalence of infarcts among never smokers estimated to be $e^{-1.1575} = 0.314$.

Model C3: Odds of infarcts among never smokers estimated to be $e^{-0.7883} = 0.4546$.

- b. (9 points) For each of the three models, provide a scientifically useful interpretation of the estimated *slope for pack years of smoking*.

Model C1: Among subjects with similar years since quitting, absolute difference in probability of infarcts per 1 pack year difference in smoking history between groups is estimated to be 0.0019, with higher prevalence with higher pack years.

Model C2: Among subjects with similar years since quitting, relative difference in probability of infarcts per 1 pack year difference in smoking history between groups is estimated to be $e^{0.00489} = 1.0049$ fold higher prevalence, with higher prevalence with higher pack years.

Model C3: Among subjects with similar years since quitting, relative difference in odds of infarcts per 1 pack year difference in smoking history between groups is estimated to be $e^{0.00822} = 1.00825$ fold higher odds, with higher odds with higher pack years.

- c. (9 points) For each of the three models, provide a scientifically useful interpretation of the estimated *slope for years since quitting*.

Model C1: Among subjects with similar pack year history of smoking, absolute difference in probability of infarcts per 1 year difference in time since quitting between groups is estimated to be 0.00168, with higher prevalence with fewer years since quitting.

Model C2: Among subjects with similar pack year history of smoking, relative difference in probability of infarcts per 1 year difference in time since quitting between groups is estimated to be 0.518% lower (RR = $e^{-0.005194} = 0.99482$), with higher prevalence with fewer years since quitting.

Model C3: Among subjects with similar pack year history of smoking, relative difference in odds of infarcts per 1 year difference in time since quitting between groups is estimated to be 0.781% lower (RR = $e^{-0.007836} = 0.99219$), with higher odds with fewer years since quitting.

- d. What was the importance of using the “robust” standard errors in models C1 and C2? Should they also have been used in model C3?

Ans: Linear regression presumes homoscedasticity and Poisson regression presumes the mean-variance relationship of a Poisson distribution. Use of the robust SE will allow for the mean-variance relationship that exists with binary data. Logistic regression presumes the correct mean-variance for binary data when the linear model is correct, and behaves well when it is not.

- e. Briefly discuss the relative advantages and disadvantages of the three regression models for exploring associations (both unadjusted and adjusted for confounding) in this setting. How similar are the results presented in Appendix B, and what issues might you worry about as you explore additional models?

Ans: Points that can be considered:

- Because the prevalence of infarcts is around 0.3, the mean-variance relationship of the binary data does not present much of a problem with linear regression (but we did use robust SE). Furthermore, the constraints on $0 < p < 1$ is unlikely to be violated unless the effects attributable to other variables are especially strong. Hence, the ease of interpreting risk difference provides some advantage.
 - Looking at probability ratios is especially attractive with lower probabilities, because the ratios will accentuate associations compared to RD. However, the prevalence is moderate here. Just the same, there is unlikely to be substantial difference in the qualitative conclusions, and probability ratios are generally well understood.
 - Logistic regression is valid here, though the OR are not approximating RR. Odds and odds ratios are less well-understood.
 - The models in both Appendix B (which I asked about) and Appendix C (which I truly meant to ask about) gave quite similar results.
 - If more complicated models are fit with strongly predictive models, linear regression and Poisson regression may estimate fitted values outside the range of possibility. This need not preclude valid inference about associations, but it does lessen the generalizability of any estimates of the linear trends over the range of data. Logistic regression has no such problem.
3. Is the association between prevalence of infarct-like lesions and smoking history confounded by age? Briefly explain your reasoning. (Appendix D provides results of logistic regression analyses adjusted for age.)

Ans: Age is positively correlated with pack years and negatively correlated with years since quitting. Age is also associated with prevalence of infarcts. Hence, age likely confounds the association between smoking and prevalence of infarcts. Comparing the adjusted and unadjusted logistic regressions reveals estimates of smoking effects that are further from the null in the adjusted models. Such behavior can be observed when adjusting for a nonconfounding “precision variable” or when adjusting for a confounder, so comparison of these two models does not help us in this problem.

4. Appendix E contains results of analyses of prevalence of infarct-like lesions by prior smoking history, age, and stroke.
- a. (20 points) Provide scientifically useful interpretations for each of the regression parameters in the regression model. (Extreme brevity is OK)

Ans: (This problem was graded by looking for evidence that you could interpret the intercept, main effects that were not involved in interactions, dummy variables, and dummy variables involved in interactions. I did not require the correct interpretation of the stroke * yrsquit term to get full credit.)

- Intercept: $e^{-4.806} = .00818$ is estimated odds of infarcts among newborn never smokers without history of TIA or stroke.

- Slope for age: 1.053 is odds ratio for infarcts per 1 year difference in age between groups having similar smoking and TIA/stroke history.
- Slope for stroke.1: 2.292 is odds ratio for infarcts between TIA only and no TIA/stroke groups who are nonsmokers of similar age.
- Slope for stroke.2: 6.282 is odds ratio for infarcts between stroke and no TIA/stroke groups who are nonsmokers of similar age.
- Slope for packyrs: 1.00859 is odds ratio for infarcts per 1 pack year difference in smoking history between groups that are currently smoking, are of similar age, and who have no prior history of TIA/stroke.
- Slope for yrsquit: 0.9903 is odds ratio for infarcts per 1 year difference in time since quitting between groups that are smokers with a similar pack year history, are of similar age, and who have no prior history of TIA/stroke.
- Slope for packyrs * stroke.1: 1.0066 is the “ratio of OR” comparing the OR for infarcts per 1 pack year difference in smoking history between groups that are currently smoking, are of similar age, and who have prior history of TIA only to the OR for infarcts per 1 pack year difference in smoking history between groups that are currently smoking, are of similar age, and who have no prior history of TIA/stroke
- Slope for packyrs * stroke.2: 1.0054 is the “ratio of OR” comparing the OR for infarcts per 1 pack year difference in smoking history between groups that are currently smoking, are of similar age, and who have prior history of stroke to the OR for infarcts per 1 pack year difference in smoking history between groups that are currently smoking, are of similar age, and who have no prior history of TIA/stroke
- Slope for yrsquit * stroke.1: 1.0101 is a mathematical construct tending to describe the “ratio of OR” comparing the OR for infarcts per 1 year difference in time since quitting between groups of smokers that have a negligible pack year history, are of similar age, and who have prior history of TIA only to the OR for infarcts per 1 pack year difference in smoking history between groups of smokers that have a negligible pack year history, are of similar age, and who have no prior history of TIA/stroke
- Slope for yrsquit * stroke.2: 0.9918 is a mathematical construct tending to describe the “ratio of OR” comparing the OR for infarcts per 1 year difference in time since quitting between groups of smokers that have a negligible pack year history, are of similar age, and who have prior history of stroke to the OR for infarcts per 1 pack year difference in smoking history between groups of smokers that have a negligible pack year history, are of similar age, and who have no prior history of TIA/stroke
 - b. (10 points) What parameters would you test in order to answer the question of whether prior history of stroke modifies the association between infarct-like lesions and smoking?

Ans: Slopes for stroke.1#packyrs, stroke.2#packyrs, stroke.1#yrsquit, stroke.2#yrsquit

- c. (10 points) What parameters would you test in order to answer the question of an association between infarct-like lesions and smoking?

Ans: All of those in part b plus the packyrs and yrsquit main effects

- d. (5 points) How might the analysis results from the model in Appendix E that pertain to the group of patients without prior history of stroke compare to an analysis that restricted only to that group?

Ans: Because no interaction is fit for stroke by age, we cannot be sure exactly. However, if there is no substantial effect modification between stroke and age, the regression parameters might be similar to the main effects in the models from Appendix E (i.e. just omitting any terms that include the stroke variable).

5. Suppose that historically approximately 10% of the hypotheses generated from exploratory analyses eventually pan out to be true. Suppose that we design confirmatory studies with a type 1 error of 0.05 and 90% statistical power. Explain the impact that performing multiple comparisons might have on the generalizability of the eventual analysis results. Consider both the cases of failure to adjust for inflation of the type 1 error the setting in which appropriate adjustment is made. (It is perfectly acceptable to illustrate the issues with hypothetical results. Just make clear those hypothetical assumptions.)

Ans: Of major concern is the Bayes factor, which is the ratio of power to type 1 error.

If we do multiple comparisons, we might increase the type 1 error from 0.05 to, say, 0.075—a relative 50% increase.

To have the same positive predictive value of a “significant result”, we would then need to have increased power to $1.5 * 90\% = 135\%$, which is impossible.

If we control our type 1 error while performing the multiple comparisons, we just need to ensure that we also control our power. As a general rule, controlling for type 1 error in such a way will decrease our power, unless we also increase our sample size.

Bottom line: We should plan for the impact that multiple comparisons will have on both the type 1 error and the power, and try to maintain the Bayes factor.

Grade distributions are:

Total possible : 139
Max achieved : 133
Mean (SD) : 107 (16.8)
80th %ile : 125
50th %ile : 109
20th %ile : 92

APPENDIX A: Description of variables and descriptive statistics

These data come from a cross-sectional study of cerebral MRI findings in a population based sample of 3,775 generally healthy older adults in four regions of the U.S. We are particularly interested in associations between smoking and the prevalence of infarct-like lesions on MRI. This exam considers the following variables (all measured at time of study enrolment).

<i>age</i>	age of the subject in years
<i>packyrs</i>	pack year history (packs per day times years smoking) for the subject (never smokers = 0)
<i>yrsquit</i>	years since quitting smoking for the subject (never smokers = 0 and current smokers = 0)
<i>stroke</i>	categorical variable indicating prior history of cerebrovascular disease (0= none, 1= TIAs only, 2= clinical stroke)
<i>infarcts</i>	indicator of infarct-like lesions found on MRI (0= no, 1= yes)

Univariate statistics both overall and within groups defined by presence of infarct-like lesions:

```
. tabstat age packyrs yrsquit stroke, by(infarcts) stat(n mean sd min q max)
```

infarcts	vrbl	N	mean	sd	min	p25	p50	p75	max
0	age	2529	74.67	4.96	65	71	74	78	94
	packyrs	2439	15.69	23.99	0	0	0.50	26.0	204
	yrsquit	2529	8.67	13.83	0	0	0	15.0	83
	stroke	2529	0.058	0.313	0	0	0	0	2
1	age	1246	75.93	5.38	65	72	75	80	97
	packyrs	1209	21.19	29.21	0	0	6.75	36.0	240
	yrsquit	1246	7.67	13.18	0	0	0	12.0	74
	stroke	1246	0.295	0.684	0	0	0	0	2
Total	age	3775	75.09	5.14	65	71	74	78	97
	packyrs	3648	17.51	25.96	0	0	1.87	29.2	240
	yrsquit	3775	8.34	13.63	0	0	0	15.0	83
	stroke	3775	0.136	0.482	0	0	0	0	2

Correlations among selected variables:

```
. corr packyrs yrsquit age
```

```
(obs=3648)
```

	packyrs	yrsquit	age
packyrs	1.0000		
yrsquit	0.0990	1.0000	
age	-0.0975	0.0634	1.0000

APPENDIX B: Selected analyses of prevalence of infarct-like lesions by age.**Creation of variables used in some analyses: squared age and linear splines with two knots**

```
. g agesqr= age^2
. mkspline sage65 72 sage72 80 sage80 = age
```

Model B1: Logistic regression of infarcts on age

```
. logistic infarcts age
```

```
Logistic regression                Number of obs   =       3775
                                   LR chi2(1)        =       49.26
                                   Prob > chi2         =       0.0000
Log likelihood = -2369.5705        Pseudo R2      =       0.0103
```

infarcts	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
age	1.047906	.0069919	7.01	0.000	1.034291 1.061699

Model B2: Logistic regression of infarcts on age and agesqr

```
. logistic infarcts age agesqr
```

```
Logistic regression                Number of obs   =       3775
                                   LR chi2(2)        =       49.61
                                   Prob > chi2         =       0.0000
Log likelihood = -2369.3959        Pseudo R2      =       0.0104
```

infarcts	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
age	1.147666	.1769108	0.89	0.372	.8484068 1.552483
agesqr	.9994127	.0009943	-0.59	0.555	.9974659 1.001363

Model B3: Logistic regression of infarcts on linear splines

```
. logistic infarcts sage*
```

```
Logistic regression                Number of obs   =       3775
                                   LR chi2(3)        =       49.91
                                   Prob > chi2         =       0.0000
Log likelihood = -2369.2477        Pseudo R2      =       0.0104
```

infarcts	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
sage65	1.064267	.0365936	1.81	0.070	.9949078 1.138461
sage72	1.050369	.0152387	3.39	0.001	1.020922 1.080665
sage80	1.035128	.0213013	1.68	0.093	.9942086 1.077731

APPENDIX C: Unadjusted analyses of prevalence of infarct-like lesions by pack year history and years since quitting smoking.

Model C1: Linear regression of infarcts on packyrs, yrsquit

```
. regress infarcts packyrs yrsquit, robust
```

```
Linear regression                                Number of obs =      3648
                                                F( 2, 3645) =      21.46
                                                Prob > F       =      0.0000
                                                R-squared     =      0.0122
                                                Root MSE     =      .46803
```

infarcts	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
packyrs	.0018941	.0003086	6.14	0.000	.001289	.0024991
yrsquit	-.001678	.000576	-2.91	0.004	-.0028073	-.0005487
_cons	.312111	.0102084	30.57	0.000	.2920963	.3321256

Model C2: Poisson regression of infarcts on packyrs, yrsquit

```
. poisson infarcts packyrs yrsquit, robust
```

```
Poisson regression                                Number of obs =      3648
                                                Wald chi2(2) =      55.65
                                                Prob > chi2   =      0.0000
Log pseudolikelihood = -2530.5488                Pseudo R2      =      0.0054
```

infarcts	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
packyrs	.0048917	.0006839	7.15	0.000	.0035513	.0062321
yrsquit	-.0051946	.0019799	-2.62	0.009	-.0090751	-.0013141
_cons	-1.157517	.0313646	-36.91	0.000	-1.218991	-1.096044

Model C3: Logistic regression of infarcts on packyrs, yrsquit

```
. logit infarcts packyrs yrsquit
```

```
Logistic regression                                Number of obs =      3648
                                                LR chi2(2)    =      43.39
                                                Prob > chi2   =      0.0000
Log likelihood = -2295.4261                       Pseudo R2    =      0.0094
```

infarcts	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
packyrs	.0082182	.001333	6.17	0.000	.0056055	.0108309
yrsquit	-.007836	.0027611	-2.84	0.005	-.0132477	-.0024243
_cons	-.7882922	.047226	-16.69	0.000	-.8808536	-.6957309

APPENDIX D: Age adjusted analyses of the association between prevalence of infarct-like lesions and smoking history.

`. logit infarcts packyrs yrsquit age`

```

Logistic regression                                Number of obs   =       3648
                                                    LR chi2(3)      =       105.68
                                                    Prob > chi2     =       0.0000
Log likelihood = -2264.2824                        Pseudo R2      =       0.0228
    
```

infarcts	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
packyrs	.0095282	.0013604	7.00	0.000	.0068618	.0121946
yrsquit	-.0095575	.0027866	-3.43	0.001	-.0150192	-.0040957
age	.0547071	.0069528	7.87	0.000	.0410799	.0683343
_cons	-4.916088	.5284127	-9.30	0.000	-5.951758	-3.880419

`. logistic infarcts packyrs yrsquit age`

```

Logistic regression                                Number of obs   =       3648
                                                    LR chi2(3)      =       105.68
                                                    Prob > chi2     =       0.0000
Log likelihood = -2264.2824                        Pseudo R2      =       0.0228
    
```

infarcts	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
packyrs	1.009574	.0013735	7.00	0.000	1.006885	1.012269
yrsquit	.9904881	.0027601	-3.43	0.001	.985093	.9959126
age	1.056231	.0073437	7.87	0.000	1.041935	1.070723

