Bios 312: Modern Regression Analysis

Midterm Examination

March 1, 2012

Name: __________________________________________

Instructions: Please provide concise answers to all questions. Questions are of varying levels of difficult, so you may find it advantageous to skip questions you find especially difficult, and return to these questions at the end of the exam.

You are allowed one (1) page of your own notes to assist you when taking the exam.

You may use a calculator to assist with arithmetic. When making intermediate calculations, always use at least four significant digits; report at least three significant digits. If you are running short of time, leave answers in unsimplified form to receive the majority of credit.

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

Please adhere to the following pledge. If you are 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

Signature: __________________________________________

This exam consists of 9 pages plus an Appendix of Results. There are 175 total points.

- Question 1: Parts a-m, 75 points
- Question 2: Parts a-f, 35 points
- Question 3: Parts a-k, 65 points
Question 1 (75 points). Use the results in Part 1 of the Appendix to answer the following questions about birth weight. A description of the dataset is provided on page 1 of the Appendix.

1.a) Based on the regression model including black and asian (Model 1.1 or 1.2), what is the estimated mean birth weight for black, white, and asian? (5 points)

White: 1130.8 grams (the constant)

Black: 1130.8 - 93.8 = 1037 grams (constant minus black)

Asian: 1130.8 - 137.1 = 993.7 grams (constant minus asian)

1.b) Using the regression model including black and asian (Model 1.1 or 1.2) and the descriptive statistics provided, give an estimate for the mean birth weight in a combined group consisting of all asian and all white subjects. Show how you obtained this estimate using the regression results; do not just give the final answer. (5 points)

From 1.a, the mean in white is 1130.8 and the mean in asian is 993.7. In the descriptive stats, there n=24 asian and n=215 white. If we combine these groups, the mean birth weight will be a weighted average. (1130.8*215 + 993.7*24)/(215+24) = 1117 grams. Note that this is given in the t-test results.

1.c) Using the regression model including pltct (Model 1.3 or 1.4), what is the estimated mean birth weight for a subject with a platelet count of 100? (5 points)

888.2 + 100*.9483 = 983 grams
1.d) Using the regression model including pltct (Model 1.3 or 1.4), what is the estimated change in mean birth weight for a subject with a platelet count of 200 compared to a subject with a platelet count of 210? (5 points)

This is just 10 times the slope estimate; 9.483

1.e) Using the regression model including pltct (Model 1.3 or 1.4), provide an estimate of the intercept. What scientific use would you make of this quantity? (5 points)

888.3 grams. This is the expected birth weight when platelet count is 0. Since 0 is outside of the observed range of platelet count, it has no scientific utility.

1.f) Using the regression model including pltct stratified by black (Models 1.5 and 1.6), sketch the estimated association between platelet count and birth weight in the black group and in the non-black group. Explain how you could determine if black modifies the association between birth weight and platelet count. Refer to your graph in your explanation. (10 points)

Show a plot with two regression lines, properly labeled so I can determine if it is correct. The lines should cross, the y-intercept or other points could be labeled, etc.

Effect modification will exist if the slopes of the lines are different.

1.g) Suppose that instead of stratifying by black, we fit one regression model including pltct and black. If the correlation between black and pltct is 0 ($r_{xw}=0$), what would be the estimate of the pltct coefficient in such a model? If the correlation between black and pltct is not 0 ($r_{xw} \neq 0$), what would be the estimate of the pltct coefficient in this model? If either or both of these quantities are not able to be calculated, explain the difficulty. (10 points)

If the correlation is 0, then there is no confounding. The estimate of pltct in this case would be a weighted average of the two coefficients. $(0.8131*239 + 1.028*287) / (239+287) = 0.930$
If the correlation is not 0, then black counfounds the relationship between pltct and birth weight. When there is confounding, we cannot predict what the pltct coefficient will be. It could increase, decrease, or stay the same.

1.h) Using the regression model including loge of pltct (Model 1.7), provide an estimate of the expected birth weight if platelet count is 100. (5 points)

\[ 372.2 + 135.7 \times \log(100) = 997.1 \text{ grams; note this is similar to 1.c} \]

1.i) Using the regression model including loge of pltct and black (Model 1.8), provide an interpretation for the intercept. What scientific use would you make of this quantity? (5 points)

The intercept is the expected birth weight when pltct is 1 (log pltct is 0) for white/asian (black is 0). A pltct of 1 is outside the range of our data, so it has no scientific use.

1.j) Using the regression model including loge of pltct and black (Model 1.8), provide an interpretation of the pltct coefficient. Give the 95% confidence interval for the coefficient and interpret this quantity in a scientifically meaningful way. (5 points)

The pltct coefficient is the expected change in birth weight per 1-unit increase in log(pltct) among subjects who have the "same" race (in the model, race is coded as black/non-black, so I use the term "same" loosely). A 95% CI for the log(pltct) coefficient is [70.8, 195.6]. To interpret scientifically, we would want to, at least, recognize that also corresponds to an e-fold (2.72 fold) increase in pltct. Holding race constant, a 2.72 fold increase in pltct is associated with a 133 gram increase in birth weight (95% CI 70.8 grams to 195.6 grams). A change of base could also be used. For example, to talk about a doubling, we would multiply the coefficient and CI by log(2). Holding race constant, a 2 fold increase in pltct is associated with a 92.1 gram increase in birth weight (95% CI 49.0 grams to 135.6 grams).

1.k) Using the regression model including loge of pltct and black (Model 1.8), provide an estimate of the expected birth weight if platelet count is 100 and race is asian. (5 points)

\[ 426.1 + 133.2 \times \log(100) = 1039 \text{ grams (similar to 1.c and 1.h)} \]
1.l) Is there any evidence that black confounds the relationship between log<sub>e</sub> of pltct and birth weight? What would you have to consider for black to be a confounder? (5 points)

The main symptom of confounding is the unadjusted and adjusted association differ by a scientifically relevant amount. In this case, the estimates for pltct in the two models are very similar. There is not strong evidence for confounding. To formally evaluate confounding, we would need to consider the science and evaluate causal relationships. Black will be a confounder if it is related to the birth weight (causal) and is associated with pltct (in the sample).

1.m) The standard error estimate for the log<sub>e</sub> pltct coefficient is similar in models 1.7 and 1.8 (31.07 versus 31.76). Explain this finding. Would you consider black to be a precision variable? Why or why not? (5 points)

Strictly speaking, a precision variable is only related to the outcome. Black does not fit this definition. The standard error for log(pltct) is being impacted by two factors: the correlation of black with pltct (which will inflate the std error) and the association of black with birth weight (which will decrease the RMSE, and thus decrease the std error). It appears as if these two factors are, approximately, balancing each other out resulting in a similar std error.
**Question 2 (35 points).** Use the results in Part 1 and Figures 1 and 2 of the Appendix to answer the following questions about birth weight.

2.a) Based on the results of Model 1.3, is there evidence that the slope for pltct is significantly different from 0? (5 points)

Yes; $p < 0.001$ and the CI does not contain 0

2.b) Comment on the reliability of your answer to 2.a. Use Figure 1 and the relevant summary statistics to help your evaluation. (5 points)

Our answer will be reliable if the constant variance assumption holds. From the summary statistics, there is some evidence that the standard deviation of birth weight decreases as the platelet count increases. I would choose a model with robust standard errors. [Note: I mainly was interested if you discussed homoskedasticity or constant variance. If you discussed it well, I gave credit regardless of your conclusion about whether constant variance held or not.]

2.c) Based on the results of Models 1.5 and 1.6, what is the estimated mean birth weight in black with a platelet count of 200? What is the estimated mean birth weight in white/asian with a platelet count of 200? (5 points)

Black: $837.9 + 1.028 \times 200 = 1043.5$ grams

White/Asian: $956.8 + 0.8131 \times 200 = 1119.4$ grams

2.d) Comment on the reliability of your answer to 2.c. Use Figure 1 and relevant summary statistics to help your evaluation. (5 points)

Estimates about means require an additional assumption about linearity. The linear relationship holds better for the black group than the other group. For the other group, there is a particularly noticeable flattening of the curve at counts around 200. The linear relationship for black does not hold as well at low platelet counts. [Note: The discussion of linearity was important for this question.]
2.e) Suppose that in Models 1.5 and 1.6 I had used $\log_e$ of platelet count instead of untransformed platelet count. Would your estimates of the mean birth weight be more or less accurate compared to untransformed platelet count? Explain your reasoning. (5 points)

The log transformation seems to have opposite effects in black and non-black. In black, the relationship is more U-shape; however this could be due to one subject with a very low platelet count and high birth weight. In the other group, the relationship between log platelet count and birth weight is more close to being linear than when using (untransformed) platelet count.

2.f) Suppose that in Model 1.4 we failed to show a significant association between platelet count and mean birth weight. Give 4 distinct reasons why we could have obtained such a result. State your reasons specific to the model being considered here. (10 points)

There may have been an association with platelet count, but we lacked the power to detect it (a type II error)

There may have been an association with platelet count, but not in the parameter considered-- the difference in mean birth weight across groups (e.g. it may have been manifested in ratios of geometric means across groups)

There may have been an association with platelet count, but the best fitting line had 0 slope (e.g. a U-shaped relationship).

There was no association with platelet count (the null hypothesis was correct)

[Many of you said that confounding was a possible reason, but this is inherently a different scientific question. When we adjust for other covariates, we are no longer measuring the association between platelet count and birth weight-- we are estimating the association while holding all other adjustment variables constant].
Question 3 (65 points). Use the results in Part 2 of the Appendix to answer the following questions about IVH (intraventricular hemorrhage).

3.a ) Based on the results from Model 2.1, calculate the probability of IVH in asian and the probability of IVH in black. (5 points)

Asian: log odds = -1.4394 - .5065 = -1.9459; odds = .1429; prob=0.125 (which equals 3/24)

Black: log odds = -1.4394; odds = .2371; prob=.1916 (55/287)

3.b) Based on the results from Model 2.1, calculate the odds of IVH in asian divided by the odds of IVH in black. Also provide a 95% confidence interval for your estimate. (5 points)

OR = exp(-.5065) = 0.60 (.1429/.2371)

95% CI = [exp(-1.751), exp(.7384)] = [0.17, 2.1]

3.c) Model 2.2 uses the same predictors (asian and white) as Model 2.2, but Model 2.1 uses robust standard errors. Explain why the robust standard error for white, asian, and the intercept are all slightly larger than the standard error estimates obtained using classical logistic regression. (5 points)

In logistic regression there is a mean-variance relationship. If we model the mean well, then we will also estimate the variance well. In this model we have perfectly modeled the probability/odds/log odds of IVH in our three groups. Thus, we have perfectly modeled the variance. The robust standard errors will be larger than the classical errors in this case because the classical approach must be correct.

3.d) Based on the results from Model 2.3, calculate the probability of IVH in asian and the probability of IVH in black. (5 points)

Asian: log odds = -1.941; odds = .1436; prob=.1256

Black: log odds = -1.941 + .5017; odds = 0.2371; prob=.1916
3.e) Based on the results from Model 2.4, calculate the probability of IVH in asian and the probability of IVH in black. (5 points)

Asian: log odds = -1.447 + 1*-0.2541 = -1.7011; odds = .1825; prob=0.154

Black: log odds = -1.447; odds = .2352; prob=0.19

3.f) Compare the predicted probabilities of IVH in asian and black from models 3.a, 3.d, and 3.e. Explain any similarities or differences. (5 points)

Compare A to D: The estimates for black are the same, the estimates for asian are different.

Compare A to E, D to E: The estimates for both black and asian are different

Models A, D, and E differ in their assumptions. Model A assumes that the probability of IVH is different in each of the three race groups, and estimates the probability of IVH separately by race. Model D estimates the probability of IVH in black and estimates the probability of IVH in the asian/white group; in other words, it assumes that asian and white have the same probability of IVH. Therefore, in model D, the estimate for black is the same as model A (because both models include a parameter to explicitly estimate the probability) while the estimate for asian is actually a weighted average of the probability in white and the probability in asian. Model E assumes that race is an ordered categorical variable so that the probability of IVH is a linear function of race. It thus borrows information (albeit poorly) across race to estimate the probabilities.

3.g) Why is Model 2.4 inappropriate for modeling race? (5 points)

Model 2.4 assumes that race is an ordered categorical variable; it is not.

3.h) Using the results from Model 2.5, provide full statistical inference and interpretation for the black coefficient. Your answer should be on a scientifically-meaningful scale and include confidence intervals, p-values, and a (brief) written summary of the results. (10 points)

Comparing two subjects who have the same birth weight, but differ in race, the black subject will have a 1.4 fold increased odds of IVH compared to the non-black (white or asian) subject. We are 95% confident that the true odds ratio is between 0.86 and 2.31. Because the
confidence interval for the odds ratio contains 1 (and p=0.176), we have no evidence to conclude that the odds ratio differs by black.

3.i) Using the results from Model 2.5, provide full statistical inference and interpretation for the \textit{bwt} coefficient. Your answer should be on a scientifically-meaningful scale and include confidence intervals, p-values, and a (brief) written summary of the results. Furthermore, report your results per \textbf{100 g increase in birth weight} (rather than a 1 g increase). (10 points)

Comparing two subjects who have the same value of black, a subject who is 100 grams heavier at birth has a 0.82 fold decreased odds \((e^{100\times-.0019983})\) of IVH compared to a subject who is lighter at birth. We are 95% confidence that the odds ratio for a 100 gram increase in birth weight is between 0.75 and 0.90. Because this confidence interval does not contain zero (and \(p < 0.001\)), there is an association between birth weight and IVH among subjects of the "same" race (I am using the term "same race" loosely here).

3.j) Using the results of Model 2.5, provide an interpretation for the intercept. What scientific use would you make of this quantity? (5 points)

This is the predicted probability of IVH if birth weight is zero and race is asian or white (\textit{black}==0). Since birth weight cannot be zero, it has no scientific use.

3.k) Are there any symptoms that birth weight confounds the association between black and IVH? Explain your reasoning. (5 points)

To answer this question, I compared model 2.3 to model 2.5. Adjusting for \textit{bwt} attenuates the black coefficient from 0.50 (unadjusted) to 0.34 (adjusted). This is a fairly large change. Furthermore, the coefficient is getting closer to the null value of 0 rather than becoming more extreme. If \textit{bwt} were a precision variable, it would cause the black coefficient to become more extreme as precision variables in logistic regression drive probabilities towards zero or one. Thus, \textit{bwt} is not a precision variable. Since the black coefficient changed by a large amount, and we can rule out that this change was due to improved precision, there are symptoms that \textit{bwt} confounds the association between black and IVH.