Biost 536: Categorical Data Analysis in Epidemiology
Emerson, Fall 2014

Homework #3 Key
October 21, 2014

Written problems: To be submitted as a MS-Word compatible file to the class Catalyst dropbox by 5:30 pm on Tuesday, October 28, 2014. See the instructions for peer grading of the homework that are posted on the web pages.

> *On this (as all homeworks) Stata / R code and unedited Stata / R output is **TOTALLY** unacceptable. Instead, prepare a table of statistics gleaned from the Stata output. The table should be appropriate for inclusion in a scientific report, with all statistics rounded to a reasonable number of significant digits. (I am interested in how statistics are used to answer the scientific question.)*
>
> ***In all problems requesting "statistical analyses" (either descriptive or inferential), you should present both***
>
> - ***Methods:** A brief sentence or paragraph describing the statistical methods you used. This should be using wording suitable for a scientific journal, though it might be a little more detailed. A reader should be able to reproduce your analysis. DO NOT PROVIDE Stata OR R CODE.*
> - ***Inference:** A paragraph providing full statistical inference in answer to the question. Please see the supplementary document relating to "Reporting Associations" for details.*

Questions 1-5 refer to analyses of the data in the file `inflamm.txt` that is located on the class webpages. In those questions we are interested in associations between 4 year mortality and the ankle-arm index (AAI), a marker of peripheral vascular disease. We are interested in exploring different ways of modeling the predictor of interest with respect to ease of statistical inference and ability to fit trends in risk. For this homework, we will presume that ankle-arm index is missing completely at random (MCAR) in this dataset and hence ignorable.

***Instructions for grading:** On this key I place the total points to be awarded for a particular problem just prior to the answer for the problem. I also sometimes provide very specific criteria for awarding points. Remember the purpose of peer grading is to have the grader focus more closely on answers to the problems that were potentially different from those that he/she provided, and to identify areas where the answers on the paper being graded might not be correct. It is not of value to anyone to be overly "lenient" in the grading. So while it is not appropriate to capriciously deduct points, it is equally not appropriate to give points only because "Well, they tried" (unless, of course, such a criterion is specified in the grading instructions). We welcome students' questions regarding the appropriateness of answers during the grading process. And grades assigned by the peer grader can always be appealed.*

In providing answers below, the required answers are in* bold face type*. Further verbage that is in* regular italics type ***is additional information that you are in fact responsible for knowing on exams, but was not really required for the homework assignment. That additional information may help you decide whether the answers you are grading are valid, however.

In each problem requiring both description of methods and results, full credit should only be given when both aspects are appropriately addressed. The grader should be able to determine the exact analysis method from their description.

I note that a major point that I try to make is that there is great similarity among all types of regression. I used "cut-and-paste" extensively. In problem 2, I wrote out the answers completely. In problems 3 and 4, I cut and past the answers and highlight what is different from problem 2 in* blue font*. By doing this I am hoping that you will be able to abstract that commonality of approach across the methods.

At the end of this document, I provide Stata code that I used to produce the answers. This is for your later reference when trying to understand Stata. I did not expect (and in fact do not tolerate) the Stata output as solutions to the homework assignment. It should not have been included in a homework set turned in by a student.

1. We are interested in analyzing associations between 4 year mortality and ankle-arm index at study enrollment using statistical methods appropriate for binary response variables. The observation time for death among these subjects is potentially subject to censoring. Provide a statistical analysis demonstrating that such methods as logistic regression can be used to answer this question.

Instructions for grading: This problem is worth 5 points. Note that you need to consider the minimum observation time among patients whose time to death was censored. It is not sufficient for this problem to consider the patients who did or not die within 4 years, because until you assess the censoring you do not know whether it is valid to dichotomize the sample in that manner.

Methods: **The minimum of the observation time among patients still alive was compared to 4 years.**

Results: **Of patients still alive at the time of data analysis, the shortest observation time was 1480 days, which is 4 years and 19 days. Hence, we have at least 4 years follow-up on all patients, and we will have no incomplete data on a binary indicator of death within 4 years.**

2. Using the risk difference (RD) as a measure of association, provide statistical inference regarding an association between 4 year mortality and baseline measures of ankle-arm index (AAI).

   a. Provide suitable descriptive statistics in support of the analyses performed investigating an association between 4 year mortality and ankle-arm index. (The goal is to have these descriptive statistics support any of the analyses you perform below.)

Instructions for grading: This part of the problem is worth 10 points. Key points to consider in your grading:

- ***The major focus of inference is an association between four year mortality (a binary response variable) and the AAI (a continuous POI). Ultimately, we will have to choose some appropriate method of measuring that association. Because it was not explicitly stated to be otherwise, we should presume that the "burden of proof" is to provide evidence of some sort of association. (If that had already been established, there would almost always be something in the scientific question noting that the primary question related to further detail about the "dose-response".)***
- ***The purpose of descriptive statistics thus falls primarily into the category of "straightforward estimates in support of the primary question of interest". However, depending on the primary method of analysis, we will undoubtedly have secondary purposes falling into the categories of***

  - ***"Examining the validity of inference": This is most pertinent to analyses based on regression models that would only include a single predictor (e.g., untransformed linear continuous or log transformed continuous), because in those cases we might want to see if the data departs so strikingly from the modeled relationship that we would lack precision to detect an association.***
  - ***"Exploratory analyses": After establishing the general trend in any association, it is natural to next consider whether the relationship is linear throughout the range of the POI. Hence, we would like to have descriptive statistics that allow us to examine the pattern of "dose-response".***

  Both of those purposes will lead to assessing the pattern of association.

- ***With a continuous response variable and a continuous POI, we could use a scatterplot and a smooth of some sort to address our primary and secondary questions. However, with a binary response variable, such descriptive statistics are usually relatively uninformative. Hence, we can consider tabular descriptive statistics. (It is okay with me if a student provided an informative plot of the bivariate distribution. It just needs to be a useful depiction of the data that addresses both existence of an association and "linearity" of effect.)***
- ***With a binary response and a continuous POI, we can descriptively address existence of associations by providing descriptive statistics within categories of the response variable (i.e., descriptive statistics of AAI according to four year mortality). While this does provide some information, it suffers from two drawbacks:***
  - ***It is the "backward question" in that it conditions on our putative "effect" and examines the distribution of the putatitve "cause". Esthetically it is far more pleasing to condition on our assumed "cause" and try to estimate the distribution of the "effect".***
  - ***It does not provide information about the "dose-response".***
- ***In my answer, I provided a descriptive analysis that I would typically include no matter which of the analysis models I chose for my primary analysis of an association. This analysis is based on a stratified (dummy variable) analysis on the same scale (RR, OR, or RD) that I was using for my primary analysis. I note that the CI and p values presented in the table are not truly inferential in this case: There is a huge multiple comparison issue in the comparison of the 7 different groups in the categorization. However, those p values and CI can help choose which models might be used in future analyses, in the sense that they identify which groups seemed to have the biggest differences.***
  - ***I note that I use scientifically based intervals for this description. If you used quantiles for categorization you would find it very difficult to judge the "linearity" of the association. Using heptiles the cutpoints were at 0.91, 1.02, 1.06, 1.11, 1.15, and 1.21. Hence, the lowest category spanned 0.28 – 0.91, and the highest category spanned 1.21 – 2.38. The 3^{rd} -6^{th} categories covered 1.02 – 1.21 (so a width of about 0.05 in each of those intervals). We would not likely expect large differences between those groups.***
- ***In my answer, I used the interval containing an AAI of 1.0 as the reference. This was completely arbitrary, but it seemed like a reasonable way to reflect a population that we might have thought as being a little more "normal".***
- ***Because I used a regression model to provide my descriptive statistics, it is important that my description of my methods include that fact. If a student did not provide any similar descriptive statistics, they obviously did not have to include that in their "Methods" section. If the student***

did present CI and p values, the methods by which those were computed should be included. (That is, grade the Methods section according to how well they described what they did.)

- ***The "Results" for the descriptive statistics should provide some information that allows you to assess "linearity" of the association. Note that this need not include the full analysis of the association (RD, OR, or RR along with their CI and p values). Presentation of the estimated four year mortality in each of several strata (at least four such categories) is sufficient. If, however, the student used heptiles for their categorization, they should provide information about the range of AAI in each heptile or their effort was wasted. (That is, grade the "Results" section according to whether they provided adequate description of the data to both assess the existence of an association and its general trend.)***

Ans: ***Methods:*** *I provide description of methods based on my personal preferences. See above for possible interpretations based on other interpretations.* **Subjects were classified with respect to their mortality within four years of study enrollment versus their continued survival at four years. No subjects were censored within four years. The ankle:arm index (AAI) was computed as the ratio of brachial to tibial systolic blood pressure. AAI was missing for 121 subjects. That data was presumed to be missing completely at random (MCAR), and hence ignorable for the purposes of these analyses: all subjects with missing AAI were excluded from analyses of the association between four year mortality and AAI. Descriptive statistics of the mean, standard deviation, minimum and maximum for AAI were computed in groups defined by vital status at four years, as well as in the combined sample. Probabilities of four year survival or mortality were computed using sample proportions within categories of AAI when divided into seven intervals of approximate width of 0.2, with the lowest interval ranging from the observed minimum of 0.28 to 0.55 and the highest interval ranging from 1.55 to the observed maximum of 2.38. For descriptive purposes, estimated mortality risk differences for each such category were compared to individuals with AAI between 0.95 and 1.15 using a linear regression model with dummy variables for each category. Two-sided p values and 95% confidence intervals (CI) were computed using Wald type statistics computed with Huber-White sandwich estimates of the standard errors. However, because such analyses were not the primary statistical analysis and were not adjusted for the multiple comparisons inherent in such an analysis, the CI and p values should be regarded as descriptive and exploratory.**

Results: *NOTE: The student's answers do NOT need to be as detailed as I provide. They do need to provide sufficient description to be able to judge the existence of an association and the "linearity" of any trends..* **Mortality within 4 years of study entry was recorded for 5,000 generally healthy subjects recruited for the Cardiovascular Health Study from Medicare rolls, of whom 495 (9.9%) were observed to die within four years of study enrolment. Ankle-arm index (AAI) measurements were missing for 121 subjects, of whom 13 (10.7%) were observed to die within 4 years. Table 1 contains descriptive statistics for the 4,879 subjects with available ankle-arm index (AAI) measurements. Subjects observed to die within 4 years of study entry tended toward lower average AAI than those still alive at 4 years (0.98 vsersus 1.07). Also included in Table 1 is the observed probability of dying within 4 years within categories of AAI. Descriptively, there appeared to be a marked decreasing probability of death for higher values of AAI up until values of approximately 1.2, though it should be noted that the p values and CI for risk differences shown in Table 1 are not strictly valid, as they do not account for multiple comparisons. Among the categories in Table 1, the highest probability of mortality (36.5%) was observed among the 74 subjects having AAI < 0.55, while the lowest observed probability of 4 year mortality (6.1%) was observed among 1,216 subjects with AAI between 1.15 and 1.35. Subjects with AAI above 1.35 were observed to have approximately 12% probability of mortality within 4 years.**

Table 1: Descriptive statistics for ankle-arm index and four year mortality. Statistics presented for the risk difference are not adjusted for multiple comparisons, and hence they should be regarded as merely descriptive and exploratory.

Ankle:Arm Index (AAI)	Alive at 4 Years (n=4505)	Death in 4 Years (n=495)	Total (n=5000)	Risk Difference; (95% CI), Two-sided P
Mean (SD) **Min – Max**	1.07 (0.165) 0.28 - 2.38 (n= 4397)	0.98 (0.227) 0.30 - 1.89 (n= 482)	1.06 (0.175) 0.28 - 2.38 (n= 4879)	
AAI < 0.55	47 (63.5%)	27 (36.5%)	74 (100.0%)	0.287; (0.177, 0.397); P<0.001
0.55 ≤ AAI < 0.75	174 (76.0%)	55 (24.0%)	229 (100.0%)	0.162; (0.106, 0.218); P<0.001
0.75 ≤ AAI < 0.95	473 (81.8%)	105 (18.2%)	578 (100.0%)	0.104; (0.071, 0.137); P<0.001
0.95 ≤ AAI < 1.15	2387 (92.2%)	202 (7.8%)	2589 (100.0%)	0.000; (reference)
1.15 ≤ AAI < 1.35	1216 (93.9%)	79 (6.1%)	1295 (100.0%)	-0.017; (-0.034, 0.000); P=0.045
1.35 ≤ AAI < 1.55	85 (87.6%)	12 (12.4%)	97 (100.0%)	0.046; (-0.021, 0.112); P=0.177
1.55 ≤ AAI	15 (88.2%)	2 (11.8%)	17 (100.0%)	0.040; (-0.114, 0.193); P=0.613
Missing AAI	108 (89.3%)	13 (10.7%)	121 (100.0%)	NA

b. Answer the question using a continuously modeled term using **untransformed** AAI. Provide a one sentence description of the inference you would make from this analysis (you do not need to fully interpret CI). You will want to save fitted values of the estimated probability of mortality for use in part h.

***Instructions for grading*: This part of the problem is worth 10 points. Key points to consider in your grading:**

- ***Description of the method of analysis should make clear the method of regression analysis, the variables included in the regression model, the form of those variables, the use of asymptotic normal theory as the basis for CI and p values, whether classical SE or "robust" SE were used, whether the Wald (with either classical regression or robust SE) or likelihood ratio (only valid with classical regression) statistics were used, whether one-sided or two-sided p values were used, and the level of CI.***
- ***Comments made in the Key to problem 2 of Homework #2 alI pertain to this problem as well, with the additional notes:***
    - ***Because the only predictors included in the model pertain to the POI, classical linear regression with its assumptions of homoscedasticity is valid for tests of the null hypothesis (in the absence of an association, homoscedasticity must hold). However, under alternative hypotheses (such as are considered when computing a CI), there would be heteroscedasticity. Because the Huber-White sandwich estimator is just as valid in the presence of homoscedasticity, it might be easiest to just use that.***

Ans: ***Methods:*** *(Because part a dealt with descriptive statistics, the answer to this part does not need to address points made in that answer.)* **The binary indicator of death within 4 years was analyzed using linear regression on the untransformed measurement of AAI in order to assess differences in the probability of mortality across groups defined by AAI. The linear regression slope was used to estimate the average linear trend in risk difference associated with every 0.1 difference in AAI. The Huber-White sandwich estimate of the standard error of the regression parameters was used with asymptotic normal theory to compute a two-sided p value from Wald test of association and to compute a 95% confidence interval. A 0.05 threshold was used for statistical significance.**.

Results: *(Because part a dealt with descriptive statistics, the answer to this part does not need to address points made in that answer.)* **Linear regression analysis estimates that when comparing two groups that differ in their AAI measurements, the probability of 4 year mortality is a statistically significant absolute 2.85% lower for every 0.1 higher AAI (95% CI 2.22% to 3.47% lower, two-sided p < 0.0005).**

c. Answer the question using a continuously modeled term using **log transformed** AAI. Provide a one sentence description of the inference you would make from this analysis (you do not need to fully interpret CI). You will want to save fitted values of the estimated probability of mortality for use in part h.

Instructions for grading: This part of the problem is worth 10 points. Key points to consider in your grading are the same as in part b, though it may be the case that the student used a different "unit" for comparisons of the log transformation: I used a doubling. You may need to reproduce their analysis to be sure.

Ans: ***Methods:*** *(Because part a dealt with descriptive statistics, the answer to this part does not need to address points made in that answer.)* **The binary indicator of death within 4 years was analyzed using linear regression on the logarithmically transformed measurement of AAI in order to assess differences in the probability of mortality across groups defined by AAI. The linear regression slope was used to estimate the average linear trend in risk difference associated with every two-fold difference in AAI. The Huber-White sandwich estimate of the standard error of the regression parameters was used with asymptotic normal theory to compute a two-sided p value from Wald test of association and to compute a 95% confidence interval. A 0.05 threshold was used for statistical significance.**.

Results: *(Because part a dealt with descriptive statistics, the answer to this part does not need to address points made in that answer.)* **Linear regression analysis estimates that when comparing two groups that differ in their AAI measurements, the probability of 4 year mortality is a statistically significant absolute 19.3% lower for every two-fold higher AAI (95% CI 15.1% to 23.4% lower, two-sided p < 0.0005).**

d. Answer the question using a continuously modeled term using a **quadratic model including both untransformed and squared** AAI. Provide a one sentence description of the inference you would make from this analysis (you do not need to fully interpret CI). You will want to save fitted values of the estimated probability of mortality for use in part h.

Instructions for grading: This part of the problem is worth 10 points. Key points to consider in your grading are the same as in part b.

Ans: ***Methods:*** *(Because part a dealt with descriptive statistics, the answer to this part does not need to address points made in that answer.)* **The binary indicator of death within 4 years was analyzed using linear regression on a quadratic model of AAI that included both a linear continuous term and a term equal to the square of the AAI measurement. Differences in the probability of mortality across groups defined by AAI were then evaluated by simultaneously testing that both regression coefficients were equal to 0. The Huber-White sandwich estimate of the standard error of the regression parameters was used with asymptotic normal theory to compute a two-sided p value from a Wald test of association.** *(This next part was not required for the homework, but I include it in order to show how we might test for linearity.)* ***A hierarchical testing scheme was predefined such that in the presence of a statistically significant primary test for association, a secondary test for linearity of association would be performed using the coefficient for the squared term: if that coefficient for the squared term was significantly different from zero, that would be interpreted as evidence that the association***

between 4 year mortality and AAI was not linear in AAI. Because the overall test of association is used as a "gate-keeper" in this testing strategy, the experiment-wise type 1 error of the test for nonlinearity is preserved.

Results: *(Because part a dealt with descriptive statistics, the answer to this part does not need to address points made in that answer.)* **Linear regression analysis of mortality risk difference across AAI groups using a quadratic model estimates a statistically significant association between 4 year mortality and AAI (two-sided p < 0.0005).** *(Again, the following was not required for this homework.)* ***Because we found a statistically significant association between 4 year mortality and AAI, we further considered whether the regression model presented evidence of a nonlinear association. In that analysis, the regression coefficient for the squared term was found to be highly statistically significant (two-sided p < 0.0005), thus suggesting that the association between 4 year mortality and AAI is not well-described by a purely linear relationship in the risk difference.***

e. Answer the question using a continuously modeled term using **dummy variables with scientifically relevant cutpoints** for AAI. Use 7 intervals with cutpoints at 0.25, 0.55, 0.75, 0.95, 1.15, 1.35, 1.55, 2.4. Provide a one sentence description of the inference you would make from this analysis (you do not need to fully interpret CI). You will want to save fitted values of the estimated probability of mortality for use in part h.

Instructions for grading: This part of the problem is worth 10 points. Key points to consider in your grading are the same as in part b.

Ans: ***Methods:*** *(Because part a dealt with descriptive statistics, the answer to this part does not need to address points made in that answer.)* **The binary indicator of death within 4 years was analyzed using linear regression on dummy variables modeling a categorization of AAI into 7 intervals: AAI less than 0.55, 0.55 ≤ AAI < 0.75, 0.75 ≤ AAI < 0.95, 0.95 ≤ AAI < 1.15, 1.15 ≤ AAI < 1.35, 1.35 ≤ AAI < 1.55, and 1.55 ≤ AAI. Differences in the probability of mortality across groups defined by AAI were then evaluated by simultaneously testing that all regression coefficients for the dummy variables were equal to 0. The Huber-White sandwich estimate of the standard error of the regression parameters was used with asymptotic normal theory to compute a two-sided p value from a Wald test of association.** *(This next part was not required for the homework, but I include it in order to show how we might test for linearity. Notice that this test for linearity requires us to fit a different model than was fit for the test of association.)* ***A hierarchical testing scheme was predefined such that in the presence of a statistically significant primary test for association, a secondary test for linearity of association would be performed in a regression model that included all dummy variables plus a linear continuous term. If the coefficients for one or more of the dummy variables in that augmented model were significantly different from zero in a multiple partial Wald test, that would be interpreted as evidence that the association between 4 year mortality and AAI was not linear in AAI. Because the overall test of association is used as a "gate-keeper" in this testing strategy, the experiment-wise type 1 error of the test for nonlinearity is preserved.***

Results: *(Because part a dealt with descriptive statistics, the answer to this part does not need to address points made in that answer.)* **Linear regression analysis of mortality risk difference across AAI groups using a dummy variable model fit to the categorization of AAI finds a statistically significant association between 4 year mortality and AAI (two-sided p < 0.0005).** *(Again, the following was not required for this homework.)* ***Because we found a statistically significant association between 4 year mortality and AAI, we further considered whether the regression model presented evidence of a nonlinear association by fitting an augmented model including the dummy variables and a linear continuous AAI term.. In that analysis, the regression coefficients for the dummy variables were found to be jointly highly statistically***

significant (two-sided p < 0.0005), thus suggesting that the association between 4 year mortality and AAI is not well-described by a purely linear relationship in the risk difference.

f. Answer the question using a continuously modeled term using **dummy variables with cutpoints derived from quantiles** for AAI. Use 7 intervals. Provide a one sentence description of the inference you would make from this analysis (you do not need to fully interpret CI). You will want to save fitted values of the estimated probability of mortality for use in part h.

Instructions for grading: This part of the problem is worth 10 points. Key points to consider in your grading are the same as in part b.

Ans: ***Methods:*** *(Because part a dealt with descriptive statistics, the answer to this part does not need to address points made in that answer.)* **The binary indicator of death within 4 years was analyzed using linear regression on dummy variables modeling a categorization of AAI into 7 intervals such that equal sample sizes were in each group: cutpoints corresponded to AAI of 0.91, 1.02, 1.06, 1.11, 1.15, and 1.21. Differences in the probability of mortality across groups defined by AAI were then evaluated by simultaneously testing that all regression coefficients for the dummy variables were equal to 0. The Huber-White sandwich estimate of the standard error of the regression parameters was used with asymptotic normal theory to compute a two-sided p value from a Wald test of association.** *(This next part was not required for the homework, but I include it in order to show how we might test for linearity. Notice that this test for linearity requires us to fit a different model than was fit for the test of association.)* ***A hierarchical testing scheme was predefined such that in the presence of a statistically significant primary test for association, a secondary test for linearity of association would be performed in a regression model that included all dummy variables plus a linear continuous term. If the coefficients for one or more of the dummy variables in that augmented model were significantly different from zero in a multiple partial Wald test, that would be interpreted as evidence that the association between 4 year mortality and AAI was not linear in AAI. Because the overall test of association is used as a "gate-keeper" in this testing strategy, the experiment-wise type 1 error of the test for nonlinearity is preserved.***

Results: *(Because part a dealt with descriptive statistics, the answer to this part does not need to address points made in that answer.)* **Linear regression analysis of mortality risk difference across AAI groups using a dummy variable model fit to the categorization of AAI into 7 intervals of equal sample sizes finds a statistically significant association between 4 year mortality and AAI (two-sided p < 0.0005).** *(Again, the following was not required for this homework.)* ***Because we found a statistically significant association between 4 year mortality and AAI, we further considered whether the regression model presented evidence of a nonlinear association by fitting an augmented model including the dummy variables and a linear continuous AAI term.. In that analysis, the regression coefficients for the dummy variables were found to be jointly highly statistically significant (two-sided p < 0.0005), thus suggesting that the association between 4 year mortality and AAI is not well-described by a purely linear relationship in the risk difference.***

g. Answer the question using a continuously modeled term using **linear splines with scientifically relevant cutpoints** for AAI. Use 7 intervals with cutpoints at 0.25, 0.55, 0.75, 0.95, 1.15, 1.35, 1.55, 2.4. Provide a one sentence description of the inference you would make from this analysis (you do not need to fully interpret CI). You will want to save fitted values of the estimated probability of mortality for use in part h.

Instructions for grading: This part of the problem is worth 10 points. Key points to consider in your grading are the same as in part b.

Ans: ***Methods:*** *(Because part a dealt with descriptive statistics, the answer to this part does not need to address points made in that answer.)* **The binary indicator of death within 4 years was analyzed using linear regression on linear splines modeling a categorization of AAI into 7 intervals: AAI less than 0.55, 0.55 ≤ AAI < 0.75, 0.75 ≤ AAI < 0.95, 0.95 ≤ AAI < 1.15, 1.15 ≤ AAI < 1.35, 1.35 ≤ AAI < 1.55, and 1.55 ≤ AAI. Differences in the probability of mortality across groups defined by AAI were then evaluated by simultaneously testing that all regression coefficients for the linear spline variables were equal to 0. The Huber-White sandwich estimate of the standard error of the regression parameters was used with asymptotic normal theory to compute a two-sided p value from a Wald test of association.** *(This next part was not required for the homework, but I include it in order to show how we might test for linearity. Notice that this test for linearity requires us to fit a different model than was fit for the test of association.)* ***A hierarchical testing scheme was predefined such that in the presence of a statistically significant primary test for association, a secondary test for linearity of association would be performed in that regression model by testing for equality among all the linear spline regression coefficients. If statistically significant inequality of two or more of the linear spline coefficients was indicated in a multiple partial Wald test, that would be interpreted as evidence that the association between 4 year mortality and AAI was not linear in AAI. Because the overall test of association is used as a "gate-keeper" in this testing strategy, the experiment-wise type 1 error of the test for nonlinearity is preserved.***

Results: *(Because part a dealt with descriptive statistics, the answer to this part does not need to address points made in that answer.)* **Linear regression analysis of mortality risk difference across AAI groups using a linear spline model fit to the categorization of AAI finds a statistically significant association between 4 year mortality and AAI (two-sided p < 0.0005).** *(Again, the following was not required for this homework.)* ***Because we found a statistically significant association between 4 year mortality and AAI, we further considered whether the regression model presented evidence of a nonlinear association by testing for equality of all the linear spline regression coefficients. In that analysis, a highly statistically significant inequality among at least two of the regression coefficients for the linear spline variables was found (two-sided p < 0.0005), thus suggesting that the association between 4 year mortality and AAI is not well-described by a purely linear relationship in the risk difference.***

h. Graph the fitted values you generated in each of the above analyses and comment on any similarities and differences. Briefly comment on what might be the relative advantages of each of the models for modeling a predictor of interest versus confounders versus precision variables.

Instructions for grading: This part of the problem is worth 10 points.

Ans: Figure 1 plots fitted values from each of parts b-g on the mortality probability scale. Comparisons are probably best made to the linear spline fit, as that is the most flexible model displayed:

- **The linear spline fit is suggestive of a U-shaped function, though it does not appear particularly symmetric about a nadir: The negative slope at lower AAI is of greater magnitude (in absolute value) than is the positive slope at higher AAI.**
- **The quadratic fit appears to present much more of a positive slope at higher AAI than is suggested by the linear splines.**
- **The two "step function" models (dummy variables) give an overall impression similar to that of the linear splines, but neither would tend to be as precise when describing a truly continuous relationship. The fitted values based on the scientific cutpoints tend to agree**

more closely with the linear splines, owing to the fitting of narrower intervals in the extremes of the AAI distribution.

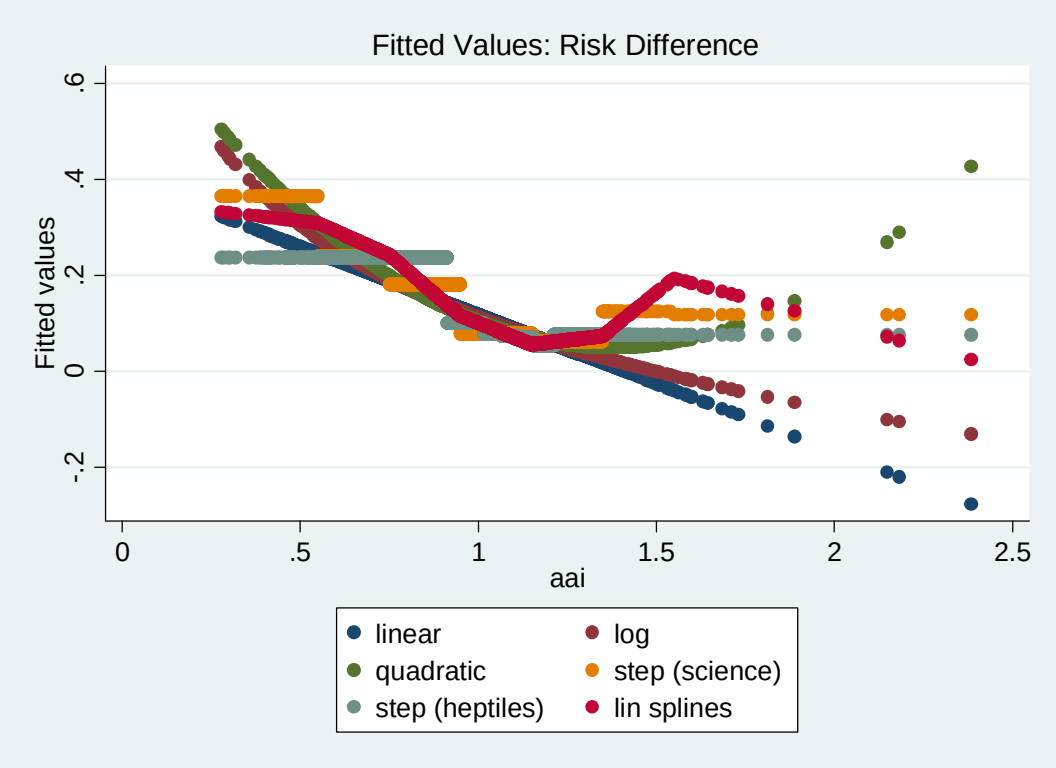
- **Both the linear and logarithmic functions show fairly similar trends, with marked departures from the linear spline models at the higher AAI values where negative probabilities are estimated. Such is "forced" by the monotonic nature of those two "transformations" (i.e., the identity and logarithmic transformations). The higher mortality risk at the lower AAI is apparently influential in predicting a continued higher mortality risk at higher AAI.**

Choices of models for the POI will counterbalance the ability to detect an association (a trend that is not flat) against the parsimony of using fewer terms in the regression model. Certainly in this data, a straight line seems to be able to capture the general trends seen in the more flexible linear splines, though the linear model cannot provide any description of the nonlinear trends. My usual approach would be to test for associations using the linear term, but provide descriptive statistics as I did in the table in part a.

When modeling confounders, the greater flexibility of the linear splines might be of some value to ensure that residual confounding is minimized, though having to use 7 parameters may lead to a decrease in precision without really providing markedly better adjustment.

When modeling "precision variables", there is less need to model the exact relationship, because all POI groups will be treated equally. Thus parsimony may again prove advantageous.

Figure 1: Comparison of fitted values from the 6 different regression models.



3. Using the odds ratio (OR) as a measure of association, answer all parts of question 2.

a. Provide suitable descriptive statistics in support of the analyses performed investigating an association between 4 year mortality and ankle-arm index. (The goal is to have these descriptive statistics support any of the analyses you perform below.)

***Instructions for grading*: This part of the problem is worth 10 points. Key points to consider in your grading are the same as in problem 2.**

Ans: ***Methods:*** *I provide description of methods based on my personal preferences. See above for possible interpretations based on other interpretations.* **Subjects were classified with respect to their mortality within four years of study enrollment versus their continued survival at four years. No subjects were censored within four years. The ankle:arm index (AAI) was computed as the ratio of brachial to tibial systolic blood pressure. AAI was missing for 121 subjects. That data was presumed to be missing completely at random (MCAR), and hence ignorable for the purposes of these analyses: all subjects with missing AAI were excluded from analyses of the association between four year mortality and AAI. Descriptive statistics of the mean, standard deviation, minimum and maximum for AAI were computed in groups defined by vital status at four years, as well as in the combined sample. Probabilities of four year survival or mortality were computed using sample proportions within categories of AAI when divided into seven intervals of approximate width of 0.2, with the lowest interval ranging from the observed minimum of 0.28 to 0.55 and the highest interval ranging from 1.55 to the observed maximum of 2.38. For descriptive purposes, estimated mortality odds ratios for each such category were compared to individuals with AAI between 0.95 and 1.15 using a logistic regression model with dummy variables for each category. Two-sided p values and 95% confidence intervals (CI) were computed using Wald type statistics ~~computed with Huber-White sandwich estimates of the standard errors~~. However, because such analyses were not the primary statistical analysis and were not adjusted for the multiple comparisons inherent in such an analysis, the CI and p values should be regarded as descriptive and exploratory.**

Results: *NOTE: The student's answers do NOT need to be as detailed as I provide. They do need to provide sufficient description to be able to judge the existence of an association and the "linearity" of any trends..* **Mortality within 4 years of study entry was recorded for 5,000 generally healthy subjects recruited for the Cardiovascular Health Study from Medicare rolls, of whom 495 (9.9%) were observed to die within four years of study enrolment. Ankle-arm index (AAI) measurements were missing for 121 subjects, of whom 13 (10.7%) were observed to die within 4 years. Table 2 contains descriptive statistics for the 4,879 subjects with available ankle-arm index (AAI) measurements. Subjects observed to die within 4 years of study entry tended toward lower average AAI than those still alive at 4 years (0.98 vsersus 1.07). Also included in Table 1 is the observed probability of dying within 4 years within categories of AAI. Descriptively, there appeared to be a marked decreasing probability of death for higher values of AAI up until values of approximately 1.2, though it should be noted that the p values and CI for odds ratios shown in Table 2 are not strictly valid, as they do not account for multiple comparisons. Among the categories in Table 2, the highest probability of mortality (36.5%) was observed among the 74 subjects having AAI < 0.55, while the lowest observed probability of 4 year mortality (6.1%) was observed among 1,216 subjects with AAI between 1.15 and 1.35. Subjects with AAI above 1.35 were observed to have approximately 12% probability of mortality within 4 years.**

Table 2: Descriptive statistics for ankle-arm index and four year mortality. Statistics presented for the odds ratios are not adjusted for multiple comparisons, and hence they should be regarded as merely descriptive and exploratory.

Ankle:Arm Index (AAI)	Alive at 4 Years (n=4505)	Death in 4 Years (n=495)	Total (n=5000)	Odds Ratio; (95% CI), Two-sided P
Mean (SD) Min – Max	1.07 (0.165) 0.28 - 2.38 (n= 4397)	0.98 (0.227) 0.30 - 1.89 (n= 482)	1.06 (0.175) 0.28 - 2.38 (n= 4879)	
AAI < 0.55	47 (63.5%)	27 (36.5%)	74 (100.0%)	6.79; (4.14, 11.1); P<0.001
0.55 ≤ AAI < 0.75	174 (76.0%)	55 (24.0%)	229 (100.0%)	3.74; (2.67, 5.22); P<0.001
0.75 ≤ AAI < 0.95	473 (81.8%)	105 (18.2%)	578 (100.0%)	2.62; (2.03, 3.39); P<0.001
0.95 ≤ AAI < 1.15	2387 (92.2%)	202 (7.8%)	2589 (100.0%)	1.000; (reference)
1.15 ≤ AAI < 1.35	1216 (93.9%)	79 (6.1%)	1295 (100.0%)	0.77; (0.587, 1.005); P=0.054
1.35 ≤ AAI < 1.55	85 (87.6%)	12 (12.4%)	97 (100.0%)	1.67; (0.896, 3.105); P=0.106
1.55 ≤ AAI	15 (88.2%)	2 (11.8%)	17 (100.0%)	1.58; (0.358, 6.938); P=0.548
Missing AAI	108 (89.3%)	13 (10.7%)	121 (100.0%)	NA

b. Answer the question using a continuously modeled term using **untransformed** AAI. Provide a one sentence description of the inference you would make from this analysis (you do not need to fully interpret CI). You will want to save fitted values of the estimated probability of mortality for use in part h.

***Instructions for grading*: This part of the problem is worth 10 points. Key points to consider in your grading:**

- ***Description of the method of analysis should make clear the method of regression analysis, the variables included in the regression model, the form of those variables, the use of asymptotic normal theory as the basis for CI and p values, whether classical SE or "robust" SE were used, whether the Wald (with either classical regression or robust SE) or likelihood ratio (only valid with classical regression) statistics were used, whether one-sided or two-sided p values were used, and the level of CI.***
- ***Comments made in the Key to problem 3 of Homework #2 alI pertain to this problem as well, ~~with the additional notes:~~***
  - ***~~Because the only predictors included in the model pertain to the POI, classical linear regression with its assumptions of homoscedasticity is valid for tests of the null hypothesis (in the absence of an association, homoscedasticity must hold). However, under alternative hypotheses (such as are considered when computing a CI), there would be heteroscedasticity. Because the Huber-White sandwich estimator is just as valid in the presence of homoscedasticity, it might be easiest to just use that.~~***

Ans: *Methods:* *(Because part a dealt with descriptive statistics, the answer to this part does not need to address points made in that answer.)* **The binary indicator of death within 4 years was analyzed using logistic regression on the untransformed measurement of AAI in order to assess odds ratios of mortality across groups defined by AAI. The logistic regression slope was used to estimate the average linear trend in the log odds ratios associated with every 0.1 difference in AAI. The ~~Huber-White sandwich~~ estimate of the standard error of the regression parameters was used with asymptotic normal theory to compute a two-sided p value from Wald test of association and to compute a 95% confidence interval. A 0.05 threshold was used for statistical significance.**.

Results: *(Because part a dealt with descriptive statistics, the answer to this part does not need to address points made in that answer.)* **Logistic regression analysis estimates that when comparing two groups that differ in their AAI measurements, the odds of 4 year mortality is**

a statistically significant 24.2% lower for every 0.1 higher AAI (95% CI 20.5% to 27.7% lower, two-sided p < 0.0005).

c. Answer the question using a continuously modeled term using **log transformed** AAI. Provide a one sentence description of the inference you would make from this analysis (you do not need to fully interpret CI). You will want to save fitted values of the estimated probability of mortality for use in part h.

***Instructions for grading*: This part of the problem is worth 10 points. Key points to consider in your grading are the same as in part b, though it may be the case that the student used a different "unit" for comparisons of the log transformation: I used a doubling. You may need to reproduce their analysis to be sure.**

Ans: *Methods:* *(Because part a dealt with descriptive statistics, the answer to this part does not need to address points made in that answer.)* **The binary indicator of death within 4 years was analyzed using logistic regression on the logarithmically transformed measurement of AAI in order to assess the ratio of the odds of mortality across groups defined by AAI. The logistic regression slope was used to estimate the average linear trend in log odds ratio associated with every two-fold difference in AAI. The ~~Huber-White sandwich~~ estimate of the standard error of the regression parameters was used with asymptotic normal theory to compute a two-sided p value from Wald test of association and to compute a 95% confidence interval. A 0.05 threshold was used for statistical significance.**.

Results: *(Because part a dealt with descriptive statistics, the answer to this part does not need to address points made in that answer.)* **Logistic regression analysis estimates that when comparing two groups that differ in their AAI measurements, the odds of 4 year mortality is a statistically significant 79.9% lower (OR=0.201) for every two-fold higher AAI (95% CI 73.6% to 84.6% lower, two-sided p < 0.0005).**

d. Answer the question using a continuously modeled term using a **quadratic model including both untransformed and squared** AAI. Provide a one sentence description of the inference you would make from this analysis (you do not need to fully interpret CI). You will want to save fitted values of the estimated probability of mortality for use in part h.

***Instructions for grading*: This part of the problem is worth 10 points. Key points to consider in your grading are the same as in part b.**

Ans: *Methods:* *(Because part a dealt with descriptive statistics, the answer to this part does not need to address points made in that answer.)* **The binary indicator of death within 4 years was analyzed using logistic regression on a quadratic model of AAI that included both a linear continuous term and a term equal to the square of the AAI measurement. Ratios of the odds of mortality across groups defined by AAI were then evaluated by simultaneously testing that both regression coefficients were equal to 0. The ~~Huber-White sandwich~~ estimate of the standard error of the regression parameters was used with asymptotic normal theory to compute a two-sided p value from a ~~Wald~~ likelihood ratio test of association.** *(This next part was not required for the homework, but I include it in order to show how we might test for linearity.)* ***A hierarchical testing scheme was predefined such that in the presence of a statistically significant primary test for association, a secondary test for linearity of association would be performed using the coefficient for the squared term: if that coefficient for the squared term was significantly different from zero, that would be interpreted as evidence that the association between 4 year mortality and AAI was not linear in AAI. Because the overall test of association is used as a "gate-keeper" in this testing strategy, the experiment-wise type 1 error of the test for nonlinearity is preserved.***

Results: *(Because part a dealt with descriptive statistics, the answer to this part does not need to address points made in that answer.)* **Logistic regression analysis of mortality odds ratio across AAI groups using a quadratic model estimates a statistically significant association between 4 year mortality and AAI (two-sided p < 0.0005).** *(Again, the following was not required for this homework.)* ***Because we found a statistically significant association between 4 year mortality and AAI, we further considered whether the regression model presented evidence of a nonlinear association. In that analysis, the regression coefficient for the squared term was found to be highly statistically significant (two-sided p = 0.005), thus suggesting that the association between 4 year mortality and AAI is not well-described by a purely linear relationship in the log odds ratio.***

e. Answer the question using a continuously modeled term using **dummy variables with scientifically relevant cutpoints** for AAI. Use 7 intervals with cutpoints at 0.25, 0.55, 0.75, 0.95, 1.15, 1.35, 1.55, 2.4. Provide a one sentence description of the inference you would make from this analysis (you do not need to fully interpret CI). You will want to save fitted values of the estimated probability of mortality for use in part h.

Instructions for grading: This part of the problem is worth 10 points. Key points to consider in your grading are the same as in part b.

Ans: ***Methods:*** *(Because part a dealt with descriptive statistics, the answer to this part does not need to address points made in that answer.)* **The binary indicator of death within 4 years was analyzed using logistic regression on dummy variables modeling a categorization of AAI into 7 intervals: AAI less than 0.55, 0.55 ≤ AAI < 0.75, 0.75 ≤ AAI < 0.95, 0.95 ≤ AAI < 1.15, 1.15 ≤ AAI < 1.35, 1.35 ≤ AAI < 1.55, and 1.55 ≤ AAI. Ratios of the odds of mortality across groups defined by AAI were then evaluated by simultaneously testing that all regression coefficients for the dummy variables were equal to 0. The ~~Huber-White sandwich~~ estimate of the standard error of the regression parameters was used with asymptotic normal theory to compute a two-sided p value from a ~~Wald~~ likelihood ratio test of association.** *(This next part was not required for the homework, but I include it in order to show how we might test for linearity. Notice that this test for linearity requires us to fit a different model than was fit for the test of association.)* ***A hierarchical testing scheme was predefined such that in the presence of a statistically significant primary test for association, a secondary test for linearity of association would be performed in a regression model that included all dummy variables plus a linear continuous term. If the coefficients for one or more of the dummy variables in that augmented model were significantly different from zero in a multiple partial Wald test, that would be interpreted as evidence that the association between 4 year mortality and AAI was not linear in AAI. Because the overall test of association is used as a "gate-keeper" in this testing strategy, the experiment-wise type 1 error of the test for nonlinearity is preserved.***

Results: *(Because part a dealt with descriptive statistics, the answer to this part does not need to address points made in that answer.)* **Logistic regression analysis of mortality odds ratios across AAI groups using a dummy variable model fit to the categorization of AAI finds a statistically significant association between 4 year mortality and AAI (two-sided p < 0.0005).** *(Again, the following was not required for this homework.)* ***Because we found a statistically significant association between 4 year mortality and AAI, we further considered whether the regression model presented evidence of a nonlinear association by fitting an augmented model including the dummy variables and a linear continuous AAI term.. In that analysis, the regression coefficients for the dummy variables were found to be jointly highly statistically significant (two-sided p < 0.0005), thus suggesting that the association between 4 year mortality and AAI is not well-described by a purely linear relationship in the log odds ratio.***

f. Answer the question using a continuously modeled term using **dummy variables with cutpoints derived from quantiles** for AAI. Use 7 intervals. Provide a one sentence description of the inference you would make from this analysis (you do not need to fully interpret CI). You will want to save fitted values of the estimated probability of mortality for use in part h.

***Instructions for grading*: This part of the problem is worth 10 points. Key points to consider in your grading are the same as in part b.**

Ans: *Methods:* *(Because part a dealt with descriptive statistics, the answer to this part does not need to address points made in that answer.)* **The binary indicator of death within 4 years was analyzed using logistic regression on dummy variables modeling a categorization of AAI into 7 intervals such that equal sample sizes were in each group: cutpoints corresponded to AAI of 0.91, 1.02, 1.06, 1.11, 1.15, and 1.21. Ratios of the odds of mortality across groups defined by AAI were then evaluated by simultaneously testing that all regression coefficients for the dummy variables were equal to 0. The ~~Huber–White sandwich~~ estimate of the standard error of the regression parameters was used with asymptotic normal theory to compute a two-sided p value from a ~~Wald~~ likelihood ratio test of association.** *(This next part was not required for the homework, but I include it in order to show how we might test for linearity. Notice that this test for linearity requires us to fit a different model than was fit for the test of association.)* ***A hierarchical testing scheme was predefined such that in the presence of a statistically significant primary test for association, a secondary test for linearity of association would be performed in a regression model that included all dummy variables plus a linear continuous term. If the coefficients for one or more of the dummy variables in that augmented model were significantly different from zero in a multiple partial Wald test, that would be interpreted as evidence that the association between 4 year mortality and AAI was not linear in AAI. Because the overall test of association is used as a "gate-keeper" in this testing strategy, the experiment-wise type 1 error of the test for nonlinearity is preserved.***

Results: *(Because part a dealt with descriptive statistics, the answer to this part does not need to address points made in that answer.)* **Logistic regression analysis of mortality risk difference across AAI groups using a dummy variable model fit to the categorization of AAI into 7 intervals of equal sample sizes finds a statistically significant association between 4 year mortality and AAI (two-sided $p < 0.0005$).** *(Again, the following was not required for this homework.)* ***Because we found a statistically significant association between 4 year mortality and AAI, we further considered whether the regression model presented evidence of a nonlinear association by fitting an augmented model including the dummy variables and a linear continuous AAI term.. In that analysis, the regression coefficients for the dummy variables were found to be jointly highly statistically significant (two-sided $p < 0.0005$), thus suggesting that the association between 4 year mortality and AAI is not well-described by a purely linear relationship in the log odds ratio.***

g. Answer the question using a continuously modeled term using **linear splines with scientifically relevant cutpoints** for AAI. Use 7 intervals with cutpoints at 0.25, 0.55, 0.75, 0.95, 1.15, 1.35, 1.55, 2.4. Provide a one sentence description of the inference you would make from this analysis (you do not need to fully interpret CI). You will want to save fitted values of the estimated probability of mortality for use in part h.

***Instructions for grading*: This part of the problem is worth 10 points. Key points to consider in your grading are the same as in part b.**

Ans: *Methods:* *(Because part a dealt with descriptive statistics, the answer to this part does not need to address points made in that answer.)* **The binary indicator of death within 4 years was**

analyzed using logistic regression on linear splines modeling a categorization of AAI into 7 intervals: AAI less than 0.55, 0.55 ≤ AAI < 0.75, 0.75 ≤ AAI < 0.95, 0.95 ≤ AAI < 1.15, 1.15 ≤ AAI < 1.35, 1.35 ≤ AAI < 1.55, and 1.55 ≤ AAI. Ratios of the odds of mortality across groups defined by AAI were then evaluated by simultaneously testing that all regression coefficients for the linear spline variables were equal to 0. The ~~Huber-White sandwich~~ estimate of the standard error of the regression parameters was used with asymptotic normal theory to compute a two-sided p value from a ~~Wald~~ likelihood ratio test of association. *(This next part was not required for the homework, but I include it in order to show how we might test for linearity. Notice that this test for linearity requires us to fit a different model than was fit for the test of association.)* ***A hierarchical testing scheme was predefined such that in the presence of a statistically significant primary test for association, a secondary test for linearity of association would be performed in that regression model by testing for equality among all the linear spline regression coefficients. If statistically significant inequality of two or more of the linear spline coefficients was indicated in a multiple partial Wald test, that would be interpreted as evidence that the association between 4 year mortality and AAI was not linear in AAI. Because the overall test of association is used as a "gate-keeper" in this testing strategy, the experiment-wise type 1 error of the test for nonlinearity is preserved.***

Results: *(Because part a dealt with descriptive statistics, the answer to this part does not need to address points made in that answer.)* **Logistic regression analysis of mortality risk difference across AAI groups using a linear spline model fit to the categorization of AAI finds a statistically significant association between 4 year mortality and AAI (two-sided p < 0.0005).** *(Again, the following was not required for this homework.)* ***Because we found a statistically significant association between 4 year mortality and AAI, we further considered whether the regression model presented evidence of a nonlinear association by testing for equality of all the linear spline regression coefficients. In that analysis, a highly statistically significant inequality among at least two of the regression coefficients for the linear spline variables was found (two-sided p < 0.0005), thus suggesting that the association between 4 year mortality and AAI is not well-described by a purely linear relationship in the log odds ratio.***

h. Graph the fitted values you generated in each of the above analyses and comment on any similarities and differences. Briefly comment on what might be the relative advantages of each of the models for modeling a predictor of interest versus confounders versus precision variables.

Instructions for grading: This part of the problem is worth 10 points.

Ans: Figure 2 plots fitted values from each of parts b-g on the mortality probability scale. Comparisons are probably best made to the linear spline fit, as that is the most flexible model displayed:

- **The linear spline fit is suggestive of a U-shaped function, though it does not appear particularly symmetric about a nadir: The negative slope at lower AAI is of greater magnitude (in absolute value) than is the positive slope at higher AAI.**
- **The quadratic fit appears very similar to the logarithmic fit, and not too dissimilar from the linear fit ~~to present much more of a positive slope at higher AAI than is suggested by the linear splines~~.**
- **The two "step function" models (dummy variables) give an overall impression similar to that of the linear splines, but neither would tend to be as precise when describing a truly continuous relationship. The fitted values based on the scientific cutpoints tend to agree more closely with the linear splines, owing to the fitting of narrower intervals in the extremes of the AAI distribution.**

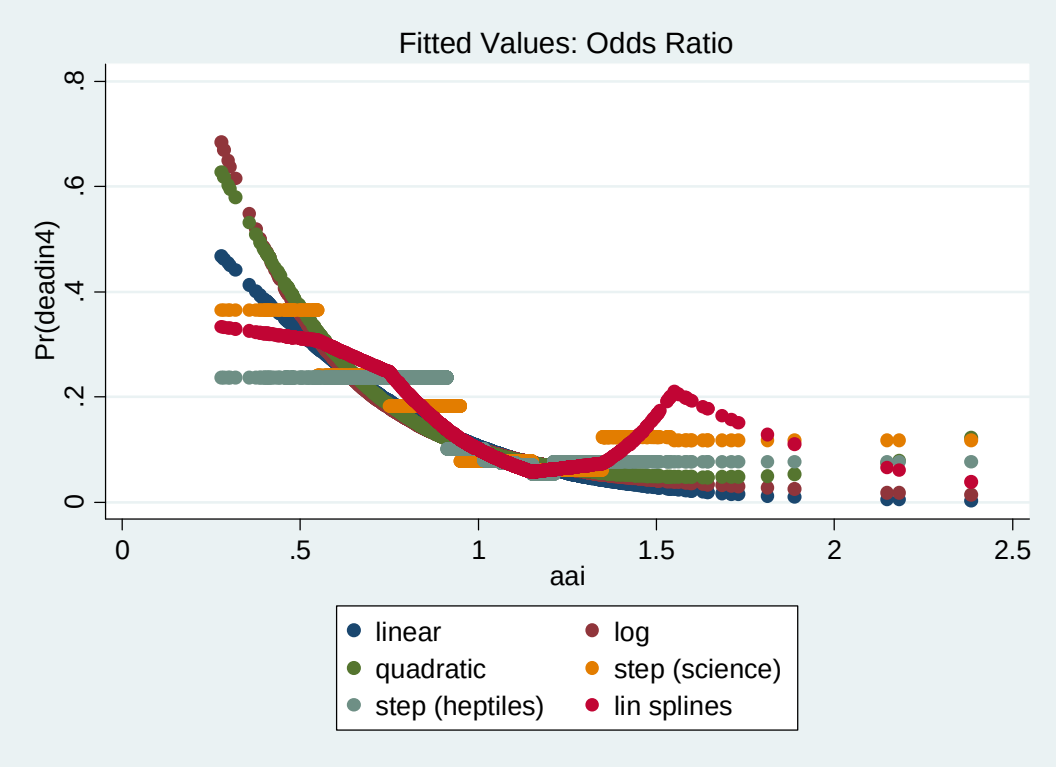
- **Both the linear and logarithmic functions show fairly similar trends, with marked departures from the linear spline models at the higher AAI values, though the estimates are always within the range 0 to 1 ~~where negative probabilities are estimated~~. Such is "forced" by the monotonic nature of those two "transformations" (i.e., the identity and logarithmic transformations). The higher mortality risk at the lower AAI is apparently influential in predicting a continued higher mortality risk at higher AAI. It is notable that owing to the nonlinear logit link, the "linear" fit on the log odds scale does not appear linear on the probability scale.**

Choices of models for the POI will counterbalance the ability to detect an association (a trend that is not flat) against the parsimony of using fewer terms in the regression model. Certainly in this data, a straight line seems to be able to capture the general trends seen in the more flexible linear splines, though the linear model cannot provide any description of the nonlinear trends. My usual approach would be to test for associations using the linear term, but provide descriptive statistics as I did in the table in part a.

When modeling confounders, the greater flexibility of the linear splines might be of some value to ensure that residual confounding is minimized, though having to use 7 parameters may lead to a decrease in precision without really providing markedly better adjustment.

When modeling "precision variables", there is less need to model the exact relationship, because all POI groups will be treated equally. Thus parsimony may again prove advantageous.

Figure 2: Comparison of fitted values from the 6 different regression models.



4. Using the risk ratio (RR) as a measure of association, answer all parts of question 2.

a. Provide suitable descriptive statistics in support of the analyses performed investigating an association between 4 year mortality and ankle-arm index. (The goal is to have these descriptive statistics support any of the analyses you perform below.)

***Instructions for grading*: This part of the problem is worth 10 points. Key points to consider in your grading are the same as in problem 2.**

Ans: ***Methods:*** *I provide description of methods based on my personal preferences. See above for possible interpretations based on other interpretations.* **Subjects were classified with respect to their mortality within four years of study enrollment versus their continued survival at four years. No subjects were censored within four years. The ankle:arm index (AAI) was computed as the ratio of brachial to tibial systolic blood pressure. AAI was missing for 121 subjects. That data was presumed to be missing completely at random (MCAR), and hence ignorable for the purposes of these analyses: all subjects with missing AAI were excluded from analyses of the association between four year mortality and AAI. Descriptive statistics of the mean, standard deviation, minimum and maximum for AAI were computed in groups defined by vital status at four years, as well as in the combined sample. Probabilities of four year survival or mortality were computed using sample proportions within categories of AAI when divided into seven intervals of approximate width of 0.2, with the lowest interval ranging from the observed minimum of 0.28 to 0.55 and the highest interval ranging from 1.55 to the observed maximum of 2.38. For descriptive purposes, estimated mortality risk ratios for each such category were compared to individuals with AAI between 0.95 and 1.15 using a Poisson regression model with dummy variables for each category. Two-sided p values and 95% confidence intervals (CI) were computed using Wald type statistics computed with Huber-White sandwich estimates of the standard errors. However, because such analyses were not the primary statistical analysis and were not adjusted for the multiple comparisons inherent in such an analysis, the CI and p values should be regarded as descriptive and exploratory.**

Results: *NOTE: The student's answers do NOT need to be as detailed as I provide. They do need to provide sufficient description to be able to judge the existence of an association and the "linearity" of any trends..* **Mortality within 4 years of study entry was recorded for 5,000 generally healthy subjects recruited for the Cardiovascular Health Study from Medicare rolls, of whom 495 (9.9%) were observed to die within four years of study enrolment. Ankle-arm index (AAI) measurements were missing for 121 subjects, of whom 13 (10.7%) were observed to die within 4 years. Table 3 contains descriptive statistics for the 4,879 subjects with available ankle-arm index (AAI) measurements. Subjects observed to die within 4 years of study entry tended toward lower average AAI than those still alive at 4 years (0.98 vsersus 1.07). Also included in Table 1 is the observed probability of dying within 4 years within categories of AAI. Descriptively, there appeared to be a marked decreasing probability of death for higher values of AAI up until values of approximately 1.2, though it should be noted that the p values and CI for risk ratios shown in Table 3 are not strictly valid, as they do not account for multiple comparisons. Among the categories in Table 1, the highest probability of mortality (36.5%) was observed among the 74 subjects having AAI < 0.55, while the lowest observed probability of 4 year mortality (6.1%) was observed among 1,216 subjects with AAI between 1.15 and 1.35. Subjects with AAI above 1.35 were observed to have approximately 12% probability of mortality within 4 years.**

Table 3: Descriptive statistics for ankle-arm index and four year mortality. Statistics presented for the risk ratios are not adjusted for multiple comparisons, and hence they should be regarded as merely descriptive and exploratory.

Ankle:Arm Index (AAI)	Alive at 4 Years (n=4505)	Death in 4 Years (n=495)	Total (n=5000)	Risk Ratio; (95% CI), Two-sided P
Mean (SD) Min – Max	1.07 (0.165) 0.28 - 2.38 (n= 4397)	0.98 (0.227) 0.30 - 1.89 (n= 482)	1.06 (0.175) 0.28 - 2.38 (n= 4879)	
AAI < 0.55	47 (63.5%)	27 (36.5%)	74 (100.0%)	4.678; (3.37, 6.50); P<0.001
0.55 ≤ AAI < 0.75	174 (76.0%)	55 (24.0%)	229 (100.0%)	3.08; (2.360, 4.02); P<0.001
0.75 ≤ AAI < 0.95	473 (81.8%)	105 (18.2%)	578 (100.0%)	2.33; (1.872, 2.90); P<0.001
0.95 ≤ AAI < 1.15	2387 (92.2%)	202 (7.8%)	2589 (100.0%)	1.00; (reference)
1.15 ≤ AAI < 1.35	1216 (93.9%)	79 (6.1%)	1295 (100.0%)	0.78; (0.608, 1.005); P=0.055
1.35 ≤ AAI < 1.55	85 (87.6%)	12 (12.4%)	97 (100.0%)	1.59; (0.918, 2.74); P=0.098
1.55 ≤ AAI	15 (88.2%)	2 (11.8%)	17 (100.0%)	1.51; (0.407, 5.58); P=0.539
Missing AAI	108 (89.3%)	13 (10.7%)	121 (100.0%)	NA

b. Answer the question using a continuously modeled term using **untransformed** AAI. Provide a one sentence description of the inference you would make from this analysis (you do not need to fully interpret CI). You will want to save fitted values of the estimated probability of mortality for use in part h.

***Instructions for grading*: This part of the problem is worth 10 points. Key points to consider in your grading:**

- ***Description of the method of analysis should make clear the method of regression analysis, the variables included in the regression model, the form of those variables, the use of asymptotic normal theory as the basis for CI and p values, whether classical SE or "robust" SE were used, whether the Wald (with either classical regression or robust SE) or likelihood ratio (only valid with classical regression) statistics were used, whether one-sided or two-sided p values were used, and the level of CI.***
- ***Comments made in the Key to problem 4b of Homework #2 alI pertain to this problem as well, with the additional notes:***
  - ***Because the death probabilities are not particularly rare in some groups, it will be especially important to allow for departures from the Poisson mean-variance relationship by using the Huber-White sandwich estimator.*** ~~***Because the only predictors included in the model pertain to the POI, classical linear regression with its assumptions of homoscedasticity is valid for tests of the null hypothesis (in the absence of an association, homoscedasticity must hold). However, under alternative hypotheses (such as are considered when computing a CI), there would be heteroscedasticity. Because the Huber-White sandwich estimator is just as valid in the presence of homoscedasticity, it might be easiest to just use that.***~~

Ans: ***Methods:*** *(Because part a dealt with descriptive statistics, the answer to this part does not need to address points made in that answer.)* **The binary indicator of death within 4 years was analyzed using Poisson regression on the untransformed measurement of AAI in order to assess ratios of the probability of mortality across groups defined by AAI. The Poisson regression slope was used to estimate the average linear trend in risk ratio associated with every 0.1 difference in AAI. The Huber-White sandwich estimate of the standard error of the regression parameters was used with asymptotic normal theory to compute a two-sided p value from Wald test of association and to compute a 95% confidence interval. A 0.05 threshold was used for statistical significance.**.

Results: *(Because part a dealt with descriptive statistics, the answer to this part does not need to address points made in that answer.)* **Poisson regression analysis estimates that when comparing two groups that differ in their AAI measurements, the probability of 4 year mortality is a statistically significant relative 20.8% lower (RR = 0.792) for every 0.1 higher AAI (95% CI 17.7% to 23.9% lower, two-sided p < 0.0005).**

c. Answer the question using a continuously modeled term using **log transformed** AAI. Provide a one sentence description of the inference you would make from this analysis (you do not need to fully interpret CI). You will want to save fitted values of the estimated probability of mortality for use in part h.

Instructions for grading: This part of the problem is worth 10 points. Key points to consider in your grading are the same as in part b, though it may be the case that the student used a different "unit" for comparisons of the log transformation: I used a doubling. You may need to reproduce their analysis to be sure.

Ans: ***Methods:*** *(Because part a dealt with descriptive statistics, the answer to this part does not need to address points made in that answer.)* **The binary indicator of death within 4 years was analyzed using Poisson regression on the logarithmically transformed measurement of AAI in order to assess ratios of the probability of mortality across groups defined by AAI. The Poisson regression slope was used to estimate the average linear trend in risk ratio associated with every two-fold difference in AAI. The Huber-White sandwich estimate of the standard error of the regression parameters was used with asymptotic normal theory to compute a two-sided p value from Wald test of association and to compute a 95% confidence interval. A 0.05 threshold was used for statistical significance.**.

Results: *(Because part a dealt with descriptive statistics, the answer to this part does not need to address points made in that answer.)* **Poisson regression analysis estimates that when comparing two groups that differ in their AAI measurements, the probability of 4 year mortality is a statistically significant relative 72.2% lower (RR = 0.278) for every two-fold higher AAI (95% CI 66.3% to 77.0% lower, two-sided p < 0.0005).**

d. Answer the question using a continuously modeled term using a **quadratic model including both untransformed and squared** AAI. Provide a one sentence description of the inference you would make from this analysis (you do not need to fully interpret CI). You will want to save fitted values of the estimated probability of mortality for use in part h.

Instructions for grading: This part of the problem is worth 10 points. Key points to consider in your grading are the same as in part b.

Ans: ***Methods:*** *(Because part a dealt with descriptive statistics, the answer to this part does not need to address points made in that answer.)* **The binary indicator of death within 4 years was analyzed using Poisson regression on a quadratic model of AAI that included both a linear continuous term and a term equal to the square of the AAI measurement. Ratios of the probability of mortality across groups defined by AAI were then evaluated by simultaneously testing that both regression coefficients were equal to 0. The Huber-White sandwich estimate of the standard error of the regression parameters was used with asymptotic normal theory to compute a two-sided p value from a Wald test of association.** *(This next part was not required for the homework, but I include it in order to show how we might test for linearity.)* ***A hierarchical testing scheme was predefined such that in the presence of a statistically significant primary test for association, a secondary test for linearity of association would be performed using the coefficient for the squared term: if that coefficient for the squared term was significantly different from zero, that would be interpreted as evidence that the association***

between 4 year mortality and AAI was not linear in AAI. Because the overall test of association is used as a "gate-keeper" in this testing strategy, the experiment-wise type 1 error of the test for nonlinearity is preserved.

Results: *(Because part a dealt with descriptive statistics, the answer to this part does not need to address points made in that answer.)* **Poisson regression analysis of mortality risk ratio across AAI groups using a quadratic model estimates a statistically significant association between 4 year mortality and AAI (two-sided p < 0.0005).** *(Again, the following was not required for this homework.)* ***Because we found a statistically significant association between 4 year mortality and AAI, we further considered whether the regression model presented evidence of a nonlinear association. In that analysis, the regression coefficient for the squared term was found to be not statistically significant (two-sided p = 0.112), thus there is insufficient evidence to suggest that the association between 4 year mortality and AAI is not well-described by a purely linear relationship in the risk ratio.***

e. Answer the question using a continuously modeled term using **dummy variables with scientifically relevant cutpoints** for AAI. Use 7 intervals with cutpoints at 0.25, 0.55, 0.75, 0.95, 1.15, 1.35, 1.55, 2.4. Provide a one sentence description of the inference you would make from this analysis (you do not need to fully interpret CI). You will want to save fitted values of the estimated probability of mortality for use in part h.

Instructions for grading: This part of the problem is worth 10 points. Key points to consider in your grading are the same as in part b.

Ans: ***Methods:*** *(Because part a dealt with descriptive statistics, the answer to this part does not need to address points made in that answer.)* **The binary indicator of death within 4 years was analyzed using Poisson regression on dummy variables modeling a categorization of AAI into 7 intervals: AAI less than 0.55, 0.55 ≤ AAI < 0.75, 0.75 ≤ AAI < 0.95, 0.95 ≤ AAI < 1.15, 1.15 ≤ AAI < 1.35, 1.35 ≤ AAI < 1.55, and 1.55 ≤ AAI. Ratios of the probability of mortality across groups defined by AAI were then evaluated by simultaneously testing that all regression coefficients for the dummy variables were equal to 0. The Huber-White sandwich estimate of the standard error of the regression parameters was used with asymptotic normal theory to compute a two-sided p value from a Wald test of association.** *(This next part was not required for the homework, but I include it in order to show how we might test for linearity. Notice that this test for linearity requires us to fit a different model than was fit for the test of association.)* ***A hierarchical testing scheme was predefined such that in the presence of a statistically significant primary test for association, a secondary test for linearity of association would be performed in a regression model that included all dummy variables plus a linear continuous term. If the coefficients for one or more of the dummy variables in that augmented model were significantly different from zero in a multiple partial Wald test, that would be interpreted as evidence that the association between 4 year mortality and AAI was not linear in AAI. Because the overall test of association is used as a "gate-keeper" in this testing strategy, the experiment-wise type 1 error of the test for nonlinearity is preserved.***

Results: *(Because part a dealt with descriptive statistics, the answer to this part does not need to address points made in that answer.)* **Poisson regression analysis of mortality risk ratio across AAI groups using a dummy variable model fit to the categorization of AAI finds a statistically significant association between 4 year mortality and AAI (two-sided p < 0.0005).** *(Again, the following was not required for this homework.)* ***Because we found a statistically significant association between 4 year mortality and AAI, we further considered whether the regression model presented evidence of a nonlinear association by fitting an augmented model including the dummy variables and a linear continuous AAI term.. In that analysis, the regression coefficients for the dummy variables were found to be jointly highly statistically***

significant (two-sided p < 0.0005), thus suggesting that the association between 4 year mortality and AAI is not well-described by a purely linear relationship in the risk ratio.

f. Answer the question using a continuously modeled term using **dummy variables with cutpoints derived from quantiles** for AAI. Use 7 intervals. Provide a one sentence description of the inference you would make from this analysis (you do not need to fully interpret CI). You will want to save fitted values of the estimated probability of mortality for use in part h.

Instructions for grading: This part of the problem is worth 10 points. Key points to consider in your grading are the same as in part b.

Ans: ***Methods:*** *(Because part a dealt with descriptive statistics, the answer to this part does not need to address points made in that answer.)* **The binary indicator of death within 4 years was analyzed using Poisson regression on dummy variables modeling a categorization of AAI into 7 intervals such that equal sample sizes were in each group: cutpoints corresponded to AAI of 0.91, 1.02, 1.06, 1.11, 1.15, and 1.21. Ratios of the probability of mortality across groups defined by AAI were then evaluated by simultaneously testing that all regression coefficients for the dummy variables were equal to 0. The Huber-White sandwich estimate of the standard error of the regression parameters was used with asymptotic normal theory to compute a two-sided p value from a Wald test of association.** *(This next part was not required for the homework, but I include it in order to show how we might test for linearity. Notice that this test for linearity requires us to fit a different model than was fit for the test of association.)* ***A hierarchical testing scheme was predefined such that in the presence of a statistically significant primary test for association, a secondary test for linearity of association would be performed in a regression model that included all dummy variables plus a linear continuous term. If the coefficients for one or more of the dummy variables in that augmented model were significantly different from zero in a multiple partial Wald test, that would be interpreted as evidence that the association between 4 year mortality and AAI was not linear in AAI. Because the overall test of association is used as a "gate-keeper" in this testing strategy, the experiment-wise type 1 error of the test for nonlinearity is preserved.***

Results: *(Because part a dealt with descriptive statistics, the answer to this part does not need to address points made in that answer.)* **Poisson regression analysis of mortality risk difference across AAI groups using a dummy variable model fit to the categorization of AAI into 7 intervals of equal sample sizes finds a statistically significant association between 4 year mortality and AAI (two-sided p < 0.0005).** *(Again, the following was not required for this homework.)* ***Because we found a statistically significant association between 4 year mortality and AAI, we further considered whether the regression model presented evidence of a nonlinear association by fitting an augmented model including the dummy variables and a linear continuous AAI term.. In that analysis, the regression coefficients for the dummy variables were found to be jointly highly statistically significant (two-sided p < 0.0005), thus suggesting that the association between 4 year mortality and AAI is not well-described by a purely linear relationship in the risk ratio.***

g. Answer the question using a continuously modeled term using **linear splines with scientifically relevant cutpoints** for AAI. Use 7 intervals with cutpoints at 0.25, 0.55, 0.75, 0.95, 1.15, 1.35, 1.55, 2.4. Provide a one sentence description of the inference you would make from this analysis (you do not need to fully interpret CI). You will want to save fitted values of the estimated probability of mortality for use in part h.

Instructions for grading: This part of the problem is worth 10 points. Key points to consider in your grading are the same as in part b.

Ans: ***Methods:*** *(Because part a dealt with descriptive statistics, the answer to this part does not need to address points made in that answer.)* **The binary indicator of death within 4 years was analyzed using Poisson regression on linear splines modeling a categorization of AAI into 7 intervals: AAI less than 0.55, 0.55 ≤ AAI < 0.75, 0.75 ≤ AAI < 0.95, 0.95 ≤ AAI < 1.15, 1.15 ≤ AAI < 1.35, 1.35 ≤ AAI < 1.55, and 1.55 ≤ AAI. Ratios of the probability of mortality across groups defined by AAI were then evaluated by simultaneously testing that all regression coefficients for the linear spline variables were equal to 0. The Huber-White sandwich estimate of the standard error of the regression parameters was used with asymptotic normal theory to compute a two-sided p value from a Wald test of association.** *(This next part was not required for the homework, but I include it in order to show how we might test for linearity. Notice that this test for linearity requires us to fit a different model than was fit for the test of association.)* ***A hierarchical testing scheme was predefined such that in the presence of a statistically significant primary test for association, a secondary test for linearity of association would be performed in that regression model by testing for equality among all the linear spline regression coefficients. If statistically significant inequality of two or more of the linear spline coefficients was indicated in a multiple partial Wald test, that would be interpreted as evidence that the association between 4 year mortality and AAI was not linear in AAI. Because the overall test of association is used as a "gate-keeper" in this testing strategy, the experiment-wise type 1 error of the test for nonlinearity is preserved.***

Results: *(Because part a dealt with descriptive statistics, the answer to this part does not need to address points made in that answer.)* **Poisson regression analysis of mortality risk difference across AAI groups using a linear spline model fit to the categorization of AAI finds a statistically significant association between 4 year mortality and AAI (two-sided p < 0.0005).** *(Again, the following was not required for this homework.)* ***Because we found a statistically significant association between 4 year mortality and AAI, we further considered whether the regression model presented evidence of a nonlinear association by testing for equality of all the linear spline regression coefficients. In that analysis, a highly statistically significant inequality among at least two of the regression coefficients for the linear spline variables was found (two-sided p < 0.0005), thus suggesting that the association between 4 year mortality and AAI is not well-described by a purely linear relationship in the risk ratio.***

h. Graph the fitted values you generated in each of the above analyses and comment on any similarities and differences. Briefly comment on what might be the relative advantages of each of the models for modeling a predictor of interest versus confounders versus precision variables.

Instructions for grading: This part of the problem is worth 10 points..

Ans: Figure 3 plots fitted values from each of parts b-g on the mortality probability scale. Comparisons are probably best made to the linear spline fit, as that is the most flexible model displayed:

- **The linear spline fit is suggestive of a U-shaped function, though it does not appear particularly symmetric about a nadir: The negative slope at lower AAI is of greater magnitude (in absolute value) than is the positive slope at higher AAI.**
- **The quadratic fit appears very similar to the logarithmic fit, and not too dissimilar from the linear fit ~~to present much more of a positive slope at higher AAI than is suggested by the linear splines~~.**
- **The two "step function" models (dummy variables) give an overall impression similar to that of the linear splines, but neither would tend to be as precise when describing a truly continuous relationship. The fitted values based on the scientific cutpoints tend to agree**

more closely with the linear splines, owing to the fitting of narrower intervals in the extremes of the AAI distribution.

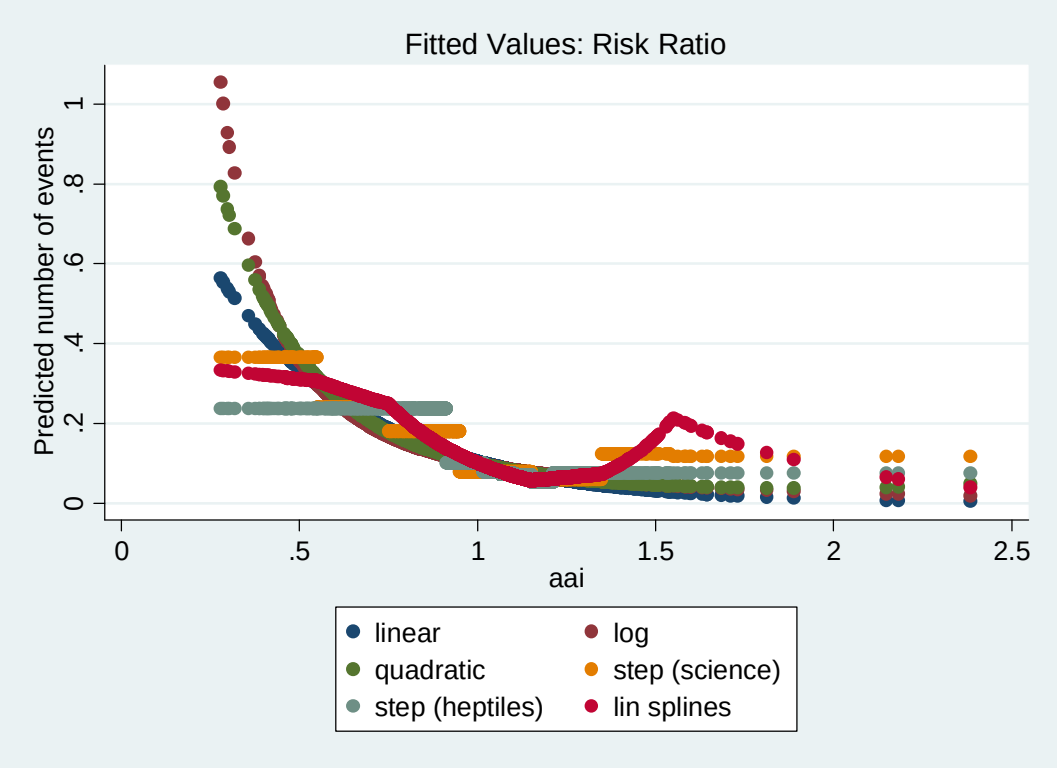
- **Both the linear and logarithmic functions show fairly similar trends, with marked departures from the linear spline models at the lowest and highest AAI values, though it is the estimates at the lower AAI for the logarithmic fit that exceed the maximum possible value of 1 ~~where negative probabilities are estimated~~. Such is "forced" by the monotonic nature of those two "transformations" (i.e., the identity and logarithmic transformations). The higher mortality risk at the lower AAI is apparently influential in predicting a continued higher mortality risk at higher AAI. It is notable that owing to the nonlinear log link, the "linear" fit on the log odds scale does not appear linear on the probability scale.**

Choices of models for the POI will counterbalance the ability to detect an association (a trend that is not flat) against the parsimony of using fewer terms in the regression model. Certainly in this data, a straight line seems to be able to capture the general trends seen in the more flexible linear splines, though the linear model cannot provide any description of the nonlinear trends. My usual approach would be to test for associations using the linear term, but provide descriptive statistics as I did in the table in part a.

When modeling confounders, the greater flexibility of the linear splines might be of some value to ensure that residual confounding is minimized, though having to use 7 parameters may lead to a decrease in precision without really providing markedly better adjustment.

When modeling "precision variables", there is less need to model the exact relationship, because all POI groups will be treated equally. Thus parsimony may again prove advantageous.

Figure 3: Comparison of fitted values from the 6 different regression models.

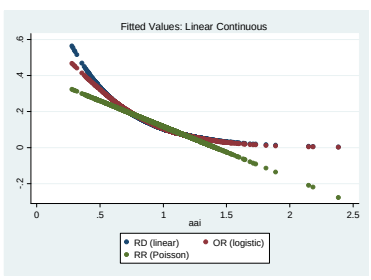


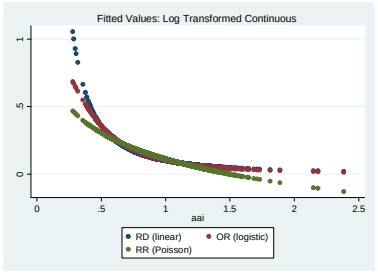
5. Comment on the similarity of fitted values from problems 2, 3, and 4 as a function of the models that were fit (e.g., how do the fitted values from the untransformed AAI models compare, etc.)

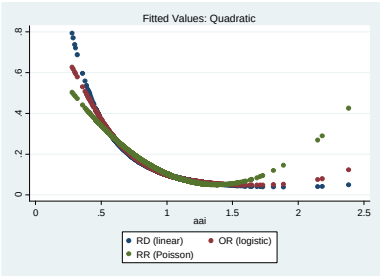
***Instructions for grading*: This problem is worth 10 points.**

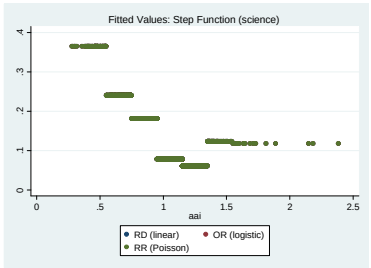
Ans: The following figures overlay the fitted values from the RD, OR, and RR regressions. Key observations:

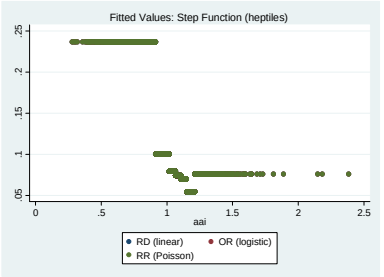
- **The dummy variable models are saturated for their respective categorizations, and thus the fitted values agree between the RD, OR, and RR regressions.**
- **The linear, log transformation, and quadratic fits are less flexible in fitting the markedly nonlinear association between mortality and AAI. They all tend to agree for the RD, OR, and RR regressions where the bulk of the data lies (AAI between 0.75 and 1.35), with the curves beyond that interval largely driven by the shape imposed by the regression model and the link function.**
- **The linear splines are markedly more flexible, and even though they are not saturated, they are all able to fit the nonlinear association fairly well. Hence, the RR, OR, and RR fitted values show marked agreement.**

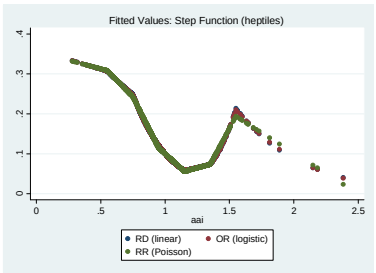












Appendix
Stata Commands and Output

```
. ******************
. * Read in the data
. ******************
. quietly: infile id site age male bkrace smoker estrogen prevdis diab2 bmi ///
>        systBP aai cholest crp fib ttodth death cvddth using ///
>            http://www.emersonstatistics.com/datasets/inflamm.txt

. * The first row was variable names
. drop in 1
(1 observation deleted)

.
. ******************
. * Problem 1: Descriptive statistics of observation time among censored
. ******************
. summ ttodth if death==0

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      ttodth |      3879    2603.711    413.5922       1480       2942

. ** Creating dichotomized time to death
. g deadin4=0
. replace deadin4=1 if ttodth<4*365.25
(495 real changes made)

.
. ******************
. * Creating transformations of AAI
. ******************
.
. ** Log transformation using log base 2 (so 1 unit equals a doubling of AAI)
. g logaai= log(aai) / log(2)
(121 missing values generated)

. ** Squared term for use in quadratic modeling
. g aaisqr= aai^2
(121 missing values generated)
```

```
. ** Categorization based on scientifically determined cutpoints (approx every 0.2)
. ** (I multiply by 100, because Stata requires integers for making dummy variables)
. egen aaictg= cut(aai), at(0.25,0.55,0.75,0.95,1.15,1.35,1.55,2.4)
(121 missing values generated)

. replace aaictg= 100*aaictg
(4879 real changes made)

. ** Categorization based on heptiles
. egen aaiQ= cut(aai), group(7)
(121 missing values generated)

. ** Linear splines
. mkspline saai025 0.55 saai055 0.75 saai075 0.95 saai095 1.15 saai115 ///
>         1.35 saai135 1.55 saai155= aai

.
. ******************
. * Problems 2a, 3a, 4a: Descriptive statistics
. ******************
. ** Finding the descriptive statistics for AAI within the categories
. tabstat aai, stat(n mean sd min q max) by(aaictg)

Summary for variables: aai
     by categories of: aaictg

  aaictg |         N      mean        sd       min       p25       p50       p75       max
---------+--------------------------------------------------------------------------------
      25 |        74  .4645838  .0672174     .2778     .4229    .47555     .5203     .5495
      55 |       229  .6591092  .0564939     .5503     .6099     .6623     .7097     .7485
      75 |       578   .874631  .0561317       .75     .8344    .88675     .9231     .9497
      95 |      2589  1.065618  .0517809     .9503     1.027    1.0694    1.1081      1.15
     115 |      1295  1.215699  .0478892    1.1503    1.1765    1.2055    1.2481    1.3486
     135 |        97  1.411403  .0540712      1.35    1.3652    1.3969      1.44    1.5405
     155 |        17  1.761135   .248509    1.5509    1.5962    1.6443    1.8121    2.3846
---------+--------------------------------------------------------------------------------
   Total |      4879   1.06393  .1745873     .2778         1    1.0847    1.1667    2.3846
--------------------------------------------------------------------------------------------

. tabstat aai, stat(n mean sd min q max) by(aaiQ)
```

```
Summary for variables: aai
     by categories of: aaiQ

    aaiQ |         N      mean        sd       min       p25       p50       p75       max
---------+--------------------------------------------------------------------------------
       0 |       697  .7448215  .1395772     .2778     .6512     .7784     .8623     .9133
       1 |       697  .9735373  .0295345     .9134     .9485     .9767         1    1.0168
       2 |       694   1.04202  .0137405    1.0169    1.0294     1.043    1.0547    1.0643
       3 |       699  1.084815  .0121164    1.0645    1.0738    1.0846    1.0949    1.1061
       4 |       698  1.128731  .0130128    1.1062     1.117    1.1291    1.1397    1.1513
       5 |       697  1.180378   .016889    1.1517    1.1667    1.1795     1.194    1.2124
       6 |       697  1.292957  .1053007    1.2126     1.234    1.2615     1.312    2.3846
---------+--------------------------------------------------------------------------------
   Total |      4879   1.06393  .1745873     .2778         1    1.0847    1.1667    2.3846
--------------------------------------------------------------------------------------------

.
. ** Descriptive statistics for AAI according to 4 year vital status
. tabstat aai, stat(n mean sd min q max) by(deadin4)

Summary for variables: aai
     by categories of: deadin4

 deadin4 |         N      mean        sd       min       p25       p50       p75       max
---------+--------------------------------------------------------------------------------
       0 |      4397  1.073553  .1650962     .2778    1.0116    1.0903    1.1707    2.3846
       1 |       482   .976145  .2268047     .2978     .8473    1.0262    1.1263    1.8881
---------+--------------------------------------------------------------------------------
   Total |      4879   1.06393  .1745873     .2778         1    1.0847    1.1667    2.3846
--------------------------------------------------------------------------------------------

. tabulate aaictg deadin4, row col missing

+--------------------+
| Key                |
|--------------------|
|     frequency      |
|  row percentage    |
| column percentage  |
+--------------------+
```

aaictg	deadin4 0	1	Total
25	47	27	74
	63.51	36.49	100.00
	1.04	5.45	1.48
55	174	55	229
	75.98	24.02	100.00
	3.86	11.11	4.58
75	473	105	578
	81.83	18.17	100.00
	10.50	21.21	11.56
95	2,387	202	2,589
	92.20	7.80	100.00
	52.99	40.81	51.78
115	1,216	79	1,295
	93.90	6.10	100.00
	26.99	15.96	25.90
135	85	12	97
	87.63	12.37	100.00
	1.89	2.42	1.94
155	15	2	17
	88.24	11.76	100.00
	0.33	0.40	0.34
.	108	13	121
	89.26	10.74	100.00
	2.40	2.63	2.42
Total	4,505	495	5,000
	90.10	9.90	100.00
	100.00	100.00	100.00

```
.
. ** Descriptive statistics for RD, OR, RR across AAI categories
. regress deadin4 ib95.aaictg, robust

Linear regression                               Number of obs =    4879
                                                F(  6,  4872) =   17.93
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0387
                                                Root MSE      =  .29276

------------------------------------------------------------------------------
             |               Robust
    deadin4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     aaictg |
         25  |   .2868425   .0562488     5.10   0.000     .1765695    .3971154
         55  |   .1621523    .028738     5.64   0.000     .1058128    .2184917
         75  |   .1036385   .0168935     6.13   0.000     .0705195    .1367575
        115  |  -.0170185   .0084925    -2.00   0.045    -.0336676   -.0003695
        135  |   .0456889   .0338678     1.35   0.177    -.0207072    .1120851
        155  |   .0396247   .0783763     0.51   0.613    -.1140283    .1932776
             |
      _cons  |   .0780224   .0052749    14.79   0.000     .0676812    .0883636
------------------------------------------------------------------------------

. logistic deadin4 ib95.aaictg

Logistic regression                             Number of obs   =       4879
                                                LR chi2(6)      =     150.50
                                                Prob > chi2     =     0.0000
Log likelihood = -1497.8239                     Pseudo R2       =     0.0478

------------------------------------------------------------------------------
    deadin4 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     aaictg |
         25  |   6.788406   1.713088     7.59   0.000     4.139643    11.13199
         55  |   3.735206   .6393465     7.70   0.000     2.670646    5.224115
         75  |   2.623187   .3420962     7.39   0.000     2.031524    3.387166
        115  |   .7677053   .1054024    -1.93   0.054     .5865817    1.004756
        135  |   1.668259   .5287813     1.61   0.106     .8963165    3.105027
```

```
        155  |   1.575578   1.191658     0.60   0.548     .3578079    6.937925
------------------------------------------------------------------------------

. poisson deadin4 ib95.aaictg, robust irr

Poisson regression                                      Number of obs   =       4879
                                                        Wald chi2(6)    =     188.57
                                                        Prob > chi2     =     0.0000
Log pseudolikelihood = -1532.3228                       Pseudo R2       =     0.0409

------------------------------------------------------------------------------
             |               Robust
     deadin4 |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      aaictg |
         25  |   4.676413   .7838192     9.20   0.000     3.366991    6.495069
         55  |   3.078278   .4173653     8.29   0.000     2.359927    4.015293
         75  |   2.328317    .258857     7.60   0.000     1.872438    2.895188
        115  |   .7818762   .1002926    -1.92   0.055      .608069    1.005364
        135  |   1.585587   .4417055     1.65   0.098     .9184731    2.737247
        155  |   1.507863    1.00681     0.62   0.539     .4073886    5.581034
------------------------------------------------------------------------------

.
. ******************
. * Problem 2: Inference based on risk difference (RD)
. ******************
. ** Part b: Note that it is easy to transform linear regression coefficients
. ** to another scale, so I do not need to fit a scaled aai
. regress deadin4 aai, robust

Linear regression                                       Number of obs =    4879
                                                        F(  1,  4877) =   79.03
                                                        Prob > F      =  0.0000
                                                        R-squared     =  0.0277
                                                        Root MSE      =  .29428

------------------------------------------------------------------------------
             |               Robust
     deadin4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
```

```
         aai |  -.2845762   .0320112    -8.89   0.000    -.3473327   -.2218198
       _cons |   .4015599   .0354806    11.32   0.000      .332002    .4711178
------------------------------------------------------------------------------

. predict linRD
(option xb assumed; fitted values)
(121 missing values generated)

. ** Part c: Note that I defined the log transform on base 2 to improve interpretability
. regress deadin4 logaai, robust

Linear regression                                      Number of obs =    4879
                                                       F(  1,  4877) =   81.73
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0314
                                                       Root MSE      =  .29372

------------------------------------------------------------------------------
             |               Robust
     deadin4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      logaai |  -.1927068   .0213156    -9.04   0.000     -.234495   -.1509186
       _cons |   .1115365   .0048112    23.18   0.000     .1021044    .1209687
------------------------------------------------------------------------------

. predict logRD
(option xb assumed; fitted values)
(121 missing values generated)

. ** Part d: Note that I can read the test for linearity from the coefficient table
. regress deadin4 aai aaisqr, robust

Linear regression                                      Number of obs =    4879
                                                       F(  2,  4876) =   43.26
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0349
                                                       Root MSE      =  .29322

------------------------------------------------------------------------------
             |               Robust
     deadin4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
```

```
-------------+----------------------------------------------------------------
         aai |  -1.030173    .179547    -5.74   0.000    -1.382166   -.6781798
      aaisqr |    .373014   .0850627     4.39   0.000     .2062527    .5397753
       _cons |   .7612228   .0962579     7.91   0.000     .5725139    .9499317
------------------------------------------------------------------------------

. predict quadRD
(option xb assumed; fitted values)
(121 missing values generated)

. ** Part e: I have to fit an extra model (and use testparm) to get the test for linearity
. regress deadin4 i.aaictg, robust

Linear regression                               Number of obs =    4879
                                                F(  6,  4872) =   17.93
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0387
                                                Root MSE      =  .29276

------------------------------------------------------------------------------
             |               Robust
     deadin4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      aaictg |
         55  |  -.1246902   .0627228    -1.99   0.047    -.2476552   -.0017252
         75  |   -.183204   .0582552    -3.14   0.002    -.2974104   -.0689975
         95  |  -.2868425   .0562488    -5.10   0.000    -.3971154   -.1765695
        115  |   -.303861    .056395    -5.39   0.000    -.4144207   -.1933013
        135  |  -.2411535   .0652327    -3.70   0.000     -.369039    -.113268
        155  |  -.2472178   .0961828    -2.57   0.010    -.4357794   -.0586562
             |
       _cons |   .3648649   .0560009     6.52   0.000     .2550778    .4746519
------------------------------------------------------------------------------

. predict ctgRD
(option xb assumed; fitted values)
(121 missing values generated)

. regress deadin4 i.aaictg aai, robust

Linear regression                               Number of obs =    4879
```

```
                                                F(  7,  4871) =   15.61
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0393
                                                Root MSE      =   .2927

------------------------------------------------------------------------------
             |               Robust
     deadin4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      aaictg |
          55 |   -.098617   .0650294    -1.52   0.129    -.2261039    .0288699
          75 |  -.1282433   .0679134    -1.89   0.059    -.2613843    .0048976
          95 |  -.2062829     .07492    -2.75   0.006    -.3531599   -.0594059
         115 |  -.2031854   .0831152    -2.44   0.015    -.3661287   -.0402421
         135 |  -.1142467   .0997378    -1.15   0.252    -.3097777    .0812843
         155 |  -.0734347   .1449838    -0.51   0.613    -.3576683    .2107989
             |
         aai |  -.1340349   .0801993    -1.67   0.095    -.2912616    .0231919
       _cons |   .4271353   .0669375     6.38   0.000     .2959075     .558363
------------------------------------------------------------------------------
```

```
. testparm i.aaictg

 ( 1)  55.aaictg = 0
 ( 2)  75.aaictg = 0
 ( 3)  95.aaictg = 0
 ( 4)  115.aaictg = 0
 ( 5)  135.aaictg = 0
 ( 6)  155.aaictg = 0

       F(  6,  4871) =    7.15
            Prob > F =    0.0000
```

```
. ** Part f: I have to fit an extra model (and use testparm) to get the test for linearity
. regress deadin4 i.aaiQ, robust

Linear regression                               Number of obs =    4879
                                                F(  6,  4872) =   17.91
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0374
                                                Root MSE      =  .29296
```

```
------------------------------------------------------------------------------
             |               Robust
     deadin4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        aaiQ |
          1  |  -.1362984   .0197336    -6.91   0.000    -.1749852   -.0976117
          2  |  -.1574781   .0191025    -8.24   0.000    -.1949276   -.1200286
          3  |  -.1623368   .0189278    -8.58   0.000    -.1994438   -.1252299
          4  |  -.1665283   .0187951    -8.86   0.000    -.2033752   -.1296813
          5  |  -.1822095   .0182667    -9.97   0.000    -.2180204   -.1463985
          6  |  -.1606887   .0189883    -8.46   0.000    -.1979143   -.1234631
             |
       _cons |   .2367288   .0161124    14.69   0.000     .2051413    .2683164
------------------------------------------------------------------------------
```

```
. predict ctgQRD
(option xb assumed; fitted values)
(121 missing values generated)

. regress deadin4 i.aaiQ aai, robust

Linear regression                               Number of obs =    4879
                                                F(  7,  4871) =   15.49
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0383
                                                Root MSE      =  .29285

------------------------------------------------------------------------------
             |               Robust
     deadin4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        aaiQ |
          1  |  -.1062054    .026474    -4.01   0.000    -.1581064   -.0543044
          2  |  -.1183746   .0301827    -3.92   0.000    -.1775463   -.0592028
          3  |  -.1176025   .0329542    -3.57   0.000    -.1822076   -.0529975
          4  |  -.1160158   .0358777    -3.23   0.001    -.1863522   -.0456793
          5  |  -.1249015   .0394716    -3.16   0.002    -.2022836   -.0475195
          6  |  -.0885684   .0474095    -1.87   0.062    -.1815123    .0043756
             |
         aai |  -.1315739   .0836265    -1.57   0.116    -.2955196    .0323719
```

```
       _cons |   .3347279   .0657212     5.09   0.000     .2058847    .4635711
------------------------------------------------------------------------------

. testparm i.aaiQ

 ( 1)  1.aaiQ = 0
 ( 2)  2.aaiQ = 0
 ( 3)  3.aaiQ = 0
 ( 4)  4.aaiQ = 0
 ( 5)  5.aaiQ = 0
 ( 6)  6.aaiQ = 0

       F(  6,  4871) =    7.05
            Prob > F =    0.0000

. ** Part g: I can test linearity from this model using post-estimation test
. regress deadin4 saai*, robust

Linear regression                                      Number of obs =    4879
                                                       F(  7,  4871) =   15.66
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0391
                                                       Root MSE      =  .29272

------------------------------------------------------------------------------
             |               Robust
     deadin4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     saai025 |  -.0849833   .5751133    -0.15   0.883    -1.212465    1.042498
     saai055 |  -.3295612   .3192676    -1.03   0.302    -.9554698    .2963473
     saai075 |  -.6384884   .1711107    -3.73   0.000    -.9739426   -.3030341
     saai095 |  -.2931897   .0798071    -3.67   0.000    -.4496476   -.1367319
     saai115 |   .0860971   .1077805     0.80   0.424    -.1252014    .2973956
     saai135 |   .6027501   .3870952     1.56   0.120    -.1561311    1.361631
     saai155 |  -.2042239    .210094    -0.97   0.331     -.616103    .2076552
       _cons |   .3550068   .2922303     1.21   0.224    -.2178964    .9279101
------------------------------------------------------------------------------

. predict splRD
(option xb assumed; fitted values)
(121 missing values generated)
```

```
. test saai025 =  saai055 = saai075 = saai095 = saai115 = saai135 = saai155

 ( 1)  saai025 - saai055 = 0
 ( 2)  saai025 - saai075 = 0
 ( 3)  saai025 - saai095 = 0
 ( 4)  saai025 - saai115 = 0
 ( 5)  saai025 - saai135 = 0
 ( 6)  saai025 - saai155 = 0

       F(  6,  4871) =    7.23
            Prob > F =    0.0000

. ** Part h: Note that scatter can take multiple Y variables
. twoway (scatter linRD logRD quadRD ctgRD ctgQRD splRD aai), ///
> t1("Fitted Values: Risk Difference") ///
> legend(label(1 "linear") label(2 "log") label(3 "quadratic") ///
> label(4 "step (science)") label(5 "step (heptiles)") label(6 "lin splines"))

.
. ******************
. * Problem 3: Inference based on odds ratio (OR)
. ******************
. ** Part b: Note that it is not easy to transform logistic regression coefficients
. ** to other units, so I do need to raise the estimates to the power 0.1 for better interpretability
. ** (see the code for problem 4b for another approach)
. logistic deadin4 aai

Logistic regression                               Number of obs   =       4879
                                                  LR chi2(1)      =     122.23
                                                  Prob > chi2     =     0.0000
Log likelihood = -1511.9612                       Pseudo R2       =     0.0389

------------------------------------------------------------------------------
     deadin4 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         aai |   .0625627   .0152232   -11.39   0.000     .0388325    .1007943
------------------------------------------------------------------------------

. * To calculate OR for difference of 0.1 in AAI
. di .0625627^0.1, .0388325^0.1,  .1007943^0.1
```

```
.75793428 .7226359 .79495693

. di 1-.0625627^0.1, 1-.0388325^0.1,  1-.1007943^0.1
.24206572 .2773641 .20504307

. predict linOR
(option pr assumed; Pr(deadin4))
(121 missing values generated)

. ** Part c: Note that I defined the log transform on base 2 to improve interpretability
. logistic deadin4 logaai

Logistic regression                               Number of obs   =       4879
                                                  LR chi2(1)      =     122.28
                                                  Prob > chi2     =     0.0000
Log likelihood = -1511.9347                       Pseudo R2       =     0.0389

------------------------------------------------------------------------------
     deadin4 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      logaai |   .2013589   .0278569   -11.58   0.000     .1535364    .2640767
------------------------------------------------------------------------------

. predict logOR
(option pr assumed; Pr(deadin4))
(121 missing values generated)

. ** Part d: Note that I can read the test for linearity from the coefficient table
. logistic deadin4 aai aaisqr

Logistic regression                               Number of obs   =       4879
                                                  LR chi2(2)      =     128.46
                                                  Prob > chi2     =     0.0000
Log likelihood = -1508.8463                       Pseudo R2       =     0.0408

------------------------------------------------------------------------------
     deadin4 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         aai |   .0020905   .0025915    -4.98   0.000     .0001841    .0237395
      aaisqr |   6.516378   4.301692     2.84   0.005      1.78689    23.76373
------------------------------------------------------------------------------
```

```
. predict quadOR
(option pr assumed; Pr(deadin4))
(121 missing values generated)

. ** Part e: I have to fit an extra model (and use testparm) to get the test for linearity
. logistic deadin4 i.aaictg

Logistic regression                                     Number of obs   =       4879
                                                        LR chi2(6)      =     150.50
                                                        Prob > chi2     =     0.0000
Log likelihood = -1497.8239                             Pseudo R2       =     0.0478

------------------------------------------------------------------------------
     deadin4 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      aaictg |
          55 |   .5502331   .1577956    -2.08   0.037     .3136458    .9652815
          75 |   .3864216   .1022022    -3.60   0.000     .2301085    .6489187
          95 |     .14731   .0371744    -7.59   0.000     .0898312    .2415668
         115 |   .1130907    .030302    -8.13   0.000     .0668886     .191206
         135 |   .2457512   .0962553    -3.58   0.000     .1140509    .5295323
         155 |   .2320983   .1834869    -1.85   0.065     .0492889    1.092937
------------------------------------------------------------------------------

. predict ctgOR
(option pr assumed; Pr(deadin4))
(121 missing values generated)

. logistic deadin4 i.aaictg aai

Logistic regression                                     Number of obs   =       4879
                                                        LR chi2(7)      =     153.29
                                                        Prob > chi2     =     0.0000
Log likelihood =  -1496.429                             Pseudo R2       =     0.0487

------------------------------------------------------------------------------
     deadin4 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      aaictg |
          55 |    .734467    .246271    -0.92   0.357     .3806795     1.41705
```

```
         75  |   .7104699   .3206675    -0.76   0.449     .2933321    1.720806
         95  |   .3597241   .2131478    -1.73   0.084     .1126175    1.149035
        115  |   .3453478   .2494273    -1.47   0.141     .0838449    1.422449
        135  |   1.003858   .9354283     0.00   0.997     .1616197    6.235198
        155  |   1.529213   2.087409     0.31   0.756     .1053318    22.20121
             |
        aai  |   .2259356   .2020593    -1.66   0.096     .0391504    1.303867
------------------------------------------------------------------------------
```

```
. testparm i.aaictg

 ( 1)  [deadin4]55.aaictg = 0
 ( 2)  [deadin4]75.aaictg = 0
 ( 3)  [deadin4]95.aaictg = 0
 ( 4)  [deadin4]115.aaictg = 0
 ( 5)  [deadin4]135.aaictg = 0
 ( 6)  [deadin4]155.aaictg = 0

           chi2(  6) =   38.30
         Prob > chi2 =    0.0000
```

```
. ** Part f: I have to fit an extra model (and use testparm) to get the test for linearity
. logistic deadin4 i.aaiQ

Logistic regression                               Number of obs   =       4879
                                                  LR chi2(6)      =     149.34
                                                  Prob > chi2     =     0.0000
Log likelihood =  -1498.404                       Pseudo R2       =     0.0475

------------------------------------------------------------------------------
     deadin4 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        aaiQ |
          1  |   .3599633   .0555568    -6.62   0.000     .2660012    .4871165
          2  |    .277517   .0461772    -7.70   0.000     .2002878     .384525
          3  |   .2591354   .0439131    -7.97   0.000     .1859012    .3612196
          4  |   .2434328   .0420859    -8.17   0.000     .1734678    .3416168
          5  |   .1859199   .0351647    -8.90   0.000     .1283307    .2693526
          6  |   .2653491   .0446867    -7.88   0.000      .190752    .3691189
------------------------------------------------------------------------------
```

```
. predict ctgQOR
(option pr assumed; Pr(deadin4))
(121 missing values generated)

. logistic deadin4 i.aaiQ aai

Logistic regression                                     Number of obs   =       4879
                                                        LR chi2(7)      =     152.19
                                                        Prob > chi2     =     0.0000
Log likelihood = -1496.9822                             Pseudo R2       =     0.0484

------------------------------------------------------------------------------
     deadin4 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        aaiQ |
          1  |   .4509134   .0929235    -3.86   0.000     .3010779    .6753164
          2  |   .3714963   .0900123    -4.09   0.000     .2310528    .5973073
          3  |   .3615319    .094963    -3.87   0.000     .2160542    .6049652
          4  |   .3543337   .1007886    -3.65   0.000     .2029047    .6187752
          5  |   .2844446   .0904099    -3.96   0.000     .1525618    .5303342
          6  |   .4508593   .1613383    -2.23   0.026     .2235845    .9091601
             |
         aai |   .3807237   .2184786    -1.68   0.092     .1236366    1.172392
------------------------------------------------------------------------------

. testparm i.aaiQ

 ( 1)  [deadin4]1.aaiQ = 0
 ( 2)  [deadin4]2.aaiQ = 0
 ( 3)  [deadin4]3.aaiQ = 0
 ( 4)  [deadin4]4.aaiQ = 0
 ( 5)  [deadin4]5.aaiQ = 0
 ( 6)  [deadin4]6.aaiQ = 0

           chi2(  6) =   31.86
         Prob > chi2 =    0.0000

. ** Part g: I can test linearity from this model using post-estimation test
. logistic deadin4 saai*

Logistic regression                                     Number of obs   =       4879
```

```
                                                LR chi2(7)      =     152.99
                                                Prob > chi2     =     0.0000
Log likelihood = -1496.5831                     Pseudo R2       =     0.0486

------------------------------------------------------------------------------
     deadin4 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     saai025 |   .6503547   1.861117    -0.15   0.880     .0023838    177.4308
     saai055 |   .2213336   .3702856    -0.90   0.367      .008337     5.87606
     saai075 |   .0125442   .0148146    -3.71   0.000     .0012393    .1269703
     saai095 |   .0179553   .0172003    -4.20   0.000     .0027465    .1173826
     saai115 |   3.673454   5.670126     0.84   0.399     .1783241    75.67267
     saai135 |   426.1017   1381.808     1.87   0.062     .7398561    245402.7
     saai155 |   .1063655   .3683617    -0.65   0.518     .0001199    94.33133
------------------------------------------------------------------------------

. predict splOR
(option pr assumed; Pr(deadin4))
(121 missing values generated)

. test saai025 =  saai055 = saai075 = saai095 = saai115 = saai135 = saai155

 ( 1)  [deadin4]saai025 - [deadin4]saai055 = 0
 ( 2)  [deadin4]saai025 - [deadin4]saai075 = 0
 ( 3)  [deadin4]saai025 - [deadin4]saai095 = 0
 ( 4)  [deadin4]saai025 - [deadin4]saai115 = 0
 ( 5)  [deadin4]saai025 - [deadin4]saai135 = 0
 ( 6)  [deadin4]saai025 - [deadin4]saai155 = 0

           chi2(  6) =   39.82
         Prob > chi2 =    0.0000

. ** Part h: Note that scatter can take multiple Y variables
. twoway (scatter linOR logOR quadOR ctgOR ctgQOR splOR aai), ///
> t1("Fitted Values: Odds Ratio") ///
> legend(label(1 "linear") label(2 "log") label(3 "quadratic") ///
> label(4 "step (science)") label(5 "step (heptiles)") label(6 "lin splines"))

.
. ******************
. * Problem 4: Inference based on risk ratio (RR)
```

```
. ******************
. ** Part b: Note that it is not easy to transform Poisson regression coefficients
. ** to other units, so I do need to raise the estimates to the power 0.1 for better interpretability
. ** To have Stata report RR for difference of 0.1 in AAI, I create a scaled version of AAI
. ** (see the code for problem 3b for another approach)
. g aai1= aai/0.1
(121 missing values generated)

. poisson deadin4 aai1, robust irr

Poisson regression                                      Number of obs   =       4879
                                                        Wald chi2(1)    =     135.90
                                                        Prob > chi2     =     0.0000
Log pseudolikelihood = -1544.6506                       Pseudo R2       =     0.0332

------------------------------------------------------------------------------
             |               Robust
     deadin4 |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        aai1 |   .7915076    .015875   -11.66   0.000     .7609968    .8232417
------------------------------------------------------------------------------

. predict linRR
(option n assumed; predicted number of events)
(121 missing values generated)

. ** Part c: Note that I defined the log transform on base 2 to improve interpretability
. poisson deadin4 logaai, robust irr

Poisson regression                                      Number of obs   =       4879
                                                        Wald chi2(1)    =     171.91
                                                        Prob > chi2     =     0.0000
Log pseudolikelihood = -1546.2746                       Pseudo R2       =     0.0322

------------------------------------------------------------------------------
             |               Robust
     deadin4 |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      logaai |   .2781437   .0271456   -13.11   0.000     .2297184    .3367772
------------------------------------------------------------------------------
```

```
. predict logRR
(option n assumed; predicted number of events)
(121 missing values generated)

. ** Part d: Note that I can read the test for linearity from the coefficient table
. poisson deadin4 aai aaisqr, robust irr

Poisson regression                                      Number of obs   =       4879
                                                        Wald chi2(2)    =     191.11
                                                        Prob > chi2     =     0.0000
Log pseudolikelihood = -1543.3149                       Pseudo R2       =     0.0340

------------------------------------------------------------------------------
             |               Robust
     deadin4 |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         aai |   .0134064   .0161159    -3.59   0.000     .0012708    .1414298
      aaisqr |   3.073506   2.169097     1.59   0.112     .7707383    12.25635
------------------------------------------------------------------------------

. predict quadRR
(option n assumed; predicted number of events)
(121 missing values generated)

. ** Part e: I have to fit an extra model (and use testparm) to get the test for linearity
. poisson deadin4 i.aaictg, robust irr

Poisson regression                                      Number of obs   =       4879
                                                        Wald chi2(6)    =     188.57
                                                        Prob > chi2     =     0.0000
Log pseudolikelihood = -1532.3228                       Pseudo R2       =     0.0409

------------------------------------------------------------------------------
             |               Robust
     deadin4 |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      aaictg |
          55 |   .6582564   .1272091    -2.16   0.030     .4507119    .9613712
          75 |   .4978853   .0881182    -3.94   0.000     .3519488    .7043347
          95 |   .2138391   .0358418    -9.20   0.000      .153963    .2970011
         115 |   .1671957   .0314652    -9.50   0.000     .1156204    .2417774
```

```
        135  |   .3390606   .1053641    -3.48   0.001     .1844013    .6234343
        155  |     .32244   .2198264    -1.66   0.097      .084749     1.22677
------------------------------------------------------------------------------
```

```
. predict ctgRR
(option n assumed; predicted number of events)
(121 missing values generated)
```

```
. poisson deadin4 i.aaictg aai, robust irr

Poisson regression                                Number of obs   =       4879
                                                  Wald chi2(7)    =     190.84
                                                  Prob > chi2     =     0.0000
Log pseudolikelihood = -1531.1331                 Pseudo R2       =     0.0417

------------------------------------------------------------------------------
             |               Robust
     deadin4 |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      aaictg |
          55 |   .8430776   .2085415    -0.69   0.490     .5191784    1.369048
          75 |      .8378   .3076733    -0.48   0.630     .4078922    1.720819
          95 |   .4584711   .2271653    -1.57   0.116     .1736021    1.210791
         115 |   .4336584   .2625082    -1.38   0.168     .1323995    1.420395
         135 |   1.126252   .8766904     0.15   0.879     .2449334    5.178725
         155 |   1.605626   1.859394     0.41   0.683     .1659248    15.53737
             |
         aai |   .2818366   .2146734    -1.66   0.096     .0633347     1.25416
------------------------------------------------------------------------------
```

```
. testparm i.aaictg

 ( 1)  [deadin4]55.aaictg = 0
 ( 2)  [deadin4]75.aaictg = 0
 ( 3)  [deadin4]95.aaictg = 0
 ( 4)  [deadin4]115.aaictg = 0
 ( 5)  [deadin4]135.aaictg = 0
 ( 6)  [deadin4]155.aaictg = 0

           chi2(  6) =   37.80
         Prob > chi2 =    0.0000
```

```
. ** Part f: I have to fit an extra model (and use testparm) to get the test for linearity
. poisson deadin4 i.aaiQ, robust irr

Poisson regression                                      Number of obs   =       4879
                                                        Wald chi2(6)    =     177.41
                                                        Prob > chi2     =     0.0000
Log pseudolikelihood =  -1532.436                       Pseudo R2       =     0.0409

------------------------------------------------------------------------------
             |               Robust
     deadin4 |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        aaiQ |
          1  |   .4242424   .0560906    -6.49   0.000     .3273961    .5497364
          2  |   .3347742   .0489402    -7.49   0.000     .2513714    .4458493
          3  |   .3142497   .0470647    -7.73   0.000     .2343104    .4214618
          4  |   .2965442    .045562    -7.91   0.000     .2194361    .4007474
          5  |    .230303   .0395646    -8.55   0.000     .1644632    .3225005
          6  |    .321212   .0477123    -7.65   0.000     .2400799    .4297618
------------------------------------------------------------------------------

. predict ctgQRR
(option n assumed; predicted number of events)
(121 missing values generated)

. poisson deadin4 i.aaiQ aai, robust irr

Poisson regression                                      Number of obs   =       4879
                                                        Wald chi2(7)    =     181.50
                                                        Prob > chi2     =     0.0000
Log pseudolikelihood = -1531.3269                       Pseudo R2       =     0.0415

------------------------------------------------------------------------------
             |               Robust
     deadin4 |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        aaiQ |
          1  |   .5060666   .0880197    -3.92   0.000      .359881    .7116335
          2  |   .4204038   .0865891    -4.21   0.000      .280768    .6294854
          3  |   .4074659   .0908705    -4.03   0.000     .2631852    .6308426
```

```
          4  |    .397341   .0953459    -3.85   0.000     .2482614    .6359423
          5  |   .3207234   .0869907    -4.19   0.000     .1884764    .5457633
          6  |   .4852989   .1382871    -2.54   0.011     .2776242    .8483232
             |
         aai |   .4734539   .2193229    -1.61   0.107     .1909708    1.173785
------------------------------------------------------------------------------
```

```
. testparm i.aaiQ

 ( 1)  [deadin4]1.aaiQ = 0
 ( 2)  [deadin4]2.aaiQ = 0
 ( 3)  [deadin4]3.aaiQ = 0
 ( 4)  [deadin4]4.aaiQ = 0
 ( 5)  [deadin4]5.aaiQ = 0
 ( 6)  [deadin4]6.aaiQ = 0

           chi2(  6) =   30.67
         Prob > chi2 =    0.0000
```

```
. ** Part g: I can test linearity from this model using post-estimation test
. poisson deadin4 saai*, robust irr

Poisson regression                                Number of obs   =       4879
                                                  Wald chi2(7)    =     189.72
                                                  Prob > chi2     =     0.0000
Log pseudolikelihood = -1531.1744                 Pseudo R2       =     0.0416

------------------------------------------------------------------------------
             |               Robust
     deadin4 |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     saai025 |    .740026    1.35254    -0.16   0.869     .0205832    26.60611
     saai055 |   .3492157    .425252    -0.86   0.388     .0321038    3.798668
     saai075 |   .0273697    .025369    -3.88   0.000     .0044493    .1683634
     saai095 |   .0246066   .0214134    -4.26   0.000       .00447    .1354541
     saai115 |   3.369746   4.977595     0.82   0.411     .1863165    60.94571
     saai135 |   208.6296   565.5872     1.97   0.049     1.027542    42359.65
     saai155 |   .1365988   .4107253    -0.66   0.508     .0003768    49.52681
------------------------------------------------------------------------------
```

```
. predict splRR
```

```
(option n assumed; predicted number of events)
(121 missing values generated)

. test saai025 =  saai055 = saai075 = saai095 = saai115 = saai135 = saai155

 ( 1)  [deadin4]saai025 - [deadin4]saai055 = 0
 ( 2)  [deadin4]saai025 - [deadin4]saai075 = 0
 ( 3)  [deadin4]saai025 - [deadin4]saai095 = 0
 ( 4)  [deadin4]saai025 - [deadin4]saai115 = 0
 ( 5)  [deadin4]saai025 - [deadin4]saai135 = 0
 ( 6)  [deadin4]saai025 - [deadin4]saai155 = 0

           chi2(  6) =   41.93
         Prob > chi2 =    0.0000

. ** Part h: Note that scatter can take multiple Y variables
. twoway (scatter linRR logRR quadRR ctgRR ctgQRR splRR aai), ///
> t1("Fitted Values: Risk Ratio") ///
> legend(label(1 "linear") label(2 "log") label(3 "quadratic") ///
> label(4 "step (science)") label(5 "step (heptiles)") label(6 "lin splines"))

.
. ******************
. * Problem 5: Comparison of fitted values on the various scales
. ******************
. twoway (scatter linRR linOR linRD aai), ///
> t1("Fitted Values: Linear Continuous") ///
> legend(label(1 "RD (linear)") label(2 "OR (logistic)") label(3 "RR (Poisson)"))

.
. twoway (scatter logRR logOR logRD aai), ///
> t1("Fitted Values: Log Transformed Continuous") ///
> legend(label(1 "RD (linear)") label(2 "OR (logistic)") label(3 "RR (Poisson)"))

.
. twoway (scatter quadRR quadOR quadRD aai), ///
> t1("Fitted Values: Quadratic") ///
> legend(label(1 "RD (linear)") label(2 "OR (logistic)") label(3 "RR (Poisson)"))

.
. twoway (scatter ctgRR ctgOR ctgRD aai), ///
```

```
> t1("Fitted Values: Step Function (science)") ///
> legend(label(1 "RD (linear)") label(2 "OR (logistic)") label(3 "RR (Poisson)"))

.
. twoway (scatter ctgQRR ctgQOR ctgQRD aai), ///
> t1("Fitted Values: Step Function (heptiles)") ///
> legend(label(1 "RD (linear)") label(2 "OR (logistic)") label(3 "RR (Poisson)"))

.
. twoway (scatter splRR splOR splRD aai), ///
> t1("Fitted Values: Step Function (heptiles)") ///
> legend(label(1 "RD (linear)") label(2 "OR (logistic)") label(3 "RR (Poisson)"))
```