**Biost 518: Applied Biostatistics II**

**Biost 515: Biostatistics II**

Emerson, Winter 2014

**Homework #5**

February 3, 2014

**Written problems:** To be submitted as a MS-Word compatible file to the class Catalyst dropbox by 9:30 am on Monday, February 10, 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.)*

***Unless explicitly told otherwise in the statement of the problem, 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.***

Problems 2 and 3 of the homework build on the analyses performed in homeworks #1 through #4. As such, all questions relate to associations among death from any cause, serum low density lipoprotein (LDL) levels, age, and sex in a population of generally healthy elderly subjects in four U.S. communities. This homework uses the subset of information that was collected to examine MRI changes in the brain. The data can be found on the class web page (follow the link to Datasets) in the file labeled mri.txt. Documentation is in the file mri.pdf. See homework #1 for additional information. Problem 1 of this homework uses the same dataset to explore associations between prevalence of diabetes and race in the population from which that sample was drawn.

1. Perform a statistical regression analysis evaluating an association between prevalence of diabetes and race by comparing the odds of a diabetes diagnosis across.
	1. Fit a logistic regression model that uses whites as a reference group. Is this a saturated model? Provide a formal report (methods and inference) about the scientific question regarding an association between diabetes and race.
	**METHODS: We fit a logistic regression model on 735 subjects with diabetes as a binary response variable and 3 predictors of interest which correspond to each race: “Black”, “Asian” and “Other”. Each of these predictors were binary variables using the indicator for their respective race. For example for the variable black each test subject was labelled 1 if “Black” and 0 if not. And similarly binary variables were created for “Asian” and “Other”. Since we have 3 predictors and 1 intercept we have 4 parameters to estimate and 4 groups to model, this is a saturated model. Statistical inference on the odds ratio of diabetes was based on the Wald statistic computed from the regression slope parameter and its standard error as estimated using the Huber-White sandwich estimator, with two-sided p value and 95% confidence interval computed using the approximate normal distribution for linear regression parameter estimates.
	INFERENCE: For the 735 subjects we had 79 (10.7%) subjects with diabetes and 656 (89.3%) without diabetes. Among the 572 “whites” the odds of diabetes is 0.1085, among the 104 “Blacks” the odds of diabetes is 0.2093, among the 47 “Asians” the odds of diabetes is 0.06818 and among the 12 subjects of “Other” races the odds of diabetes is 0.2000. Our inference is based on comparing the odds ratio to the base group which in this case is “Whites”. The odds ratio of diabetes when comparing “Blacks” and “Whites” is 1.929 which shows the odds of diabetes is estimated to be 92.9% higher for “Blacks” than for “Whites” which is a statistically significant difference based on a two-sided P-value of 0.0267 < 0.05. A 95% confidence interval shows that this observation would not be unusual if the true odds ratio was between 1.07918 and 3.446, in other words if the odds of diabetes was anywhere between 7.918% higher to 244.6% higher. Similarly we can obtain the estimated odds ratio for comparing “Whites” with “Asians” and “other” races as 0.6283 and 1.843 corresponding to 37.17% lower odds for Asians compared to whites and 84.30% higher odds for “other” races compared to whites. However, these differences were not statistically significant (two-sided P-values 0.4498 and 0.4390). For our final conclusion we would have to look at the two sided chi square p-value for the full model which is 0.1096 which implies that we fail to reject null hypothesis that there is no association between diabetes and race.

	Since we have a saturated model the point estimates form our regression model would be the same for a test of comparing the odds. The SE and consequently the confidence intervals would be slightly different even though we use robust standard errors.**
	2. Using the regression model fit in part (a), provide an interpretation for each of the regression parameters (including the intercept).
	**ANSWER: Under the regression model in part A we fit a logistic regression with 3 variables which are binary for Races “Blacks”, “Asians” and “Other races”. The race “Whites” is the reference group and so for our model the intercept corresponds to the odds of developing diabetes for “white” subjects (or the log odds if we have not taken the exponential of the estimate). Every other intercept refers to the ODDS RATIO (or log odds ratio) of developing diabetes between whites and the corresponding race. The slope of “Asians” is the odds ratio (log odds ratio) for Asians and whites, slope for “Blacks” is the odds ratio (log odds ratio) for the Blacks and Whites and finally the slope for “Other Races” is the odds ratio (log odds ratio) between “whites” and other races.**
	3. If we were to ignore issue related to multiple comparisons, what conclusions would you reach based on the p values reported in the regression output from part (a) using a 0.05 level of significance.
	**ANSWER: Based on the regression output we observed the P-value for the intercept as < 2\*10^-16, which is the P-value for testing if the odds of getting diabetes for whites is 0. Based on this at 0.05 level we can conclude that there is strong evidence to reject the null hypothesis in favor of the alternative that the odds of developing diabetes among Whites is not 0.
	For the slope of Blacks a P-value of 0.0267 was reported from which we conclude that there is evidence to reject the null hypothesis that the odds ratio of diabetes between whites and Blacks is 1.
	The estimated slope of Asians is reported as 0.4498 from which we conclude that there is not enough statistical evidence to reject the null hypothesis that the odds ratio between Asians and whites is 1.
	The estimated slope of Other races is reported as 0.4390 from which we conclude that there is not enough statistical evidence to reject the null hypothesis that the odds ratio between “Other” races and whites is 1. Here “Other” races refers to races that are not White, Black or Asian.**
	4. Now fit a logistic regression model that uses blacks as a reference group. How would your report of formal inference differ from that that you provided in part (a)? How does this regression model relate to that in part (a)?
	**ANSWER: In this case the point estimates are all different for the intercept and slopes. Now the intercept is the odds of diabetes for Blacks and similarly each slope is the odds ratio for race: Blacks and thus testing each parameter would be testing for differences in the odds between Blacks and any other race. So we can extract the information of the previous regression model ONLY for comparing whites and Blacks. In this case the estimated slope for Whites is 0.5185 where this is the odds ratio odds-white:odds-black and thus we can obtain the slope of black in the previous model by taking the reciprocal of the slope of whites in this model, 1/0.05185 = 1.929 which is the odds ratio odds-black:odds-white and similarly the rest of the analysis does the not change in-fact we observe the two sided P-value 0.0267 for testing differences in odds between blacks and whites. And similarly the confidence intervals will be the reciprocal of the CI reported in part A for the slope of blacks. There is no straightforward relationship between the other estimated slopes for both models and thus we would report different things there however the final conclusion will still be based on the observed significant difference between blacks and whites and thus our final conclusion will be the same that we have enough statistical evidence to reject null hypothesis that there is no association between race and odds of developing diabetes.**
	5. Using the regression model fit in part (d), provide an interpretation for each of the regression parameters (including the intercept.)
	**ANSWER: For the regression model in part D the slope is the estimated odds (or log odds) of getting diabetes for “Blacks” and similarly the estimates for the slopes can be interpreted as odds ratios of getting diabetes between Blacks and the races Whites, Asians and Other races which are neither Black, White or Asian. Thus the slope of Whites in this model is the estimated odds (log odds) of diabetes between Whites and blacks were the odds ratio is defined by
	And similarly the slope of Asians is the odds ratio of diabetes between Asians and blacks and the slope of “other races” is the odds of diabetes between Blacks and other races excluding Asians and whites.**
	6. If we were to ignore issue related to multiple comparisons, what conclusions would you reach based on the p values reported in the regression output from part (d) using a 0.05 level of significance.
	**ANSWER: Based on the regression output we observed the P-value for the intercept as < 2.8\*10^-8, which is the P-value for testing if the odds of getting diabetes for Blacks is 0. Based on this at 0.05 level we can conclude that there is strong evidence to reject the null hypothesis in favor of the alternative that the odds of developing diabetes among Blacks is not 0.
	For the slope of Whites a P-value of 0.0267 was reported from which we conclude that there is evidence to reject the null hypothesis that the odds ratio of diabetes between Whites and Blacks is 1.
	The estimated slope of Asians is reported as 0.0860 from which we conclude that there is not enough statistical evidence to reject the null hypothesis that the odds ratio between Asians and Blacks is 1.
	The estimated slope of Other races is reported as 0.9558 from which we conclude that there is not enough statistical evidence to reject the null hypothesis that the odds ratio between “Other” races and Blacks is 1. Here “Other” races refers to races that are not White, Black or Asian.**
	7. What do your results from parts (c) and (f) say about the dangers of using the p values for individual regression parameters from a dummy variable regression to decide whether to include or exclude those variables in a regression model (i.e., in a “stepwise model building” procedure)?
	**ANSWER: For parts C and F we notice that the slope of the P-value reported for the slope of the dummy variable Asian changes drastically from 0.4498 to 0.0860 which would be significant if we were testing at a 0.1 level this can be seen from the fact that the odds of diabetes in Whites Blacks and Asians is 0.1085, 0.2093 and 0.06818. This shows that there is much larger observed difference between the blacks and Asians as opposed to difference in odds between the whites and Asians. This shows that using dummy variable regression is NOT always a good procedure to decide whether we want to include or exclude a variable as it may be statistically significant using one reference group but not statistically significant if we change our reference group.

	If we wish to decide whether to include the entire variable “RACE” in a model or not then again this method can be unreliable. In this case, consider a categorical variable with categories 1,2,3 and assume there is a linear trend. This linear trend could be significant if we use our variable as an ordered predictor of interest but if we use the middle category as a reference variable we may not find any statistically significant slopes.**
2. Perform a statistical regression analysis evaluating an association between all-cause mortality and serum by comparing the instantaneous risk (hazard) of death over the entire period of observation across groups defined by serum LDL when fit as dummy variables using the categories suggested by the Mayo Clinic as reported on Homework #1. The Stata egen command can be used to categorize the LDL levels

egen ldlCTG = cut(ldl), at(0 70 100 130 160 190 250)

* 1. Include full description of your methods, appropriate descriptive statistics, and full report of your inferential statistics.
	**METHODS FOR DESCRIPTIVE STATISTICS: Firstly, we provide Kaplien-Mier curves for each category of serum LDL along with table for showing the survival probabilities for 1,3,5 years. The survival probabilities are evaluated using Kaplien-Mir estimates. We also provide the sample size or number of participants for each category and we present the total number of deaths within each category for the duration of the study.
	For serum LDL as a categorical variable we provide descriptive statistics for the mean, minimum, maximum and standard deviation of each serum LDL in each category in mg/dL. There were 10 missing values for serum LDL and so we excluded those 10 missing observations from our analysis.

	DESCRIPTIVE STATISTICS: We have the following table of descriptive statistics:**

|  |  |  |
| --- | --- | --- |
|  | Serum LDL at study enrollment | All Subjects |
|  | 11-69mg/dL | 70-99mg/dL | 100-129mg/dL | 130-159mg/dL | 160-189mg/dL | 190-250mg/dL |
| N subjects | 22 | 143 | 228 | 225 | 83 | 24 | 725 |
| N Deaths | 10 | 28 | 44 | 34 | 11 | 4 | 131 |
| 1 year Survival probability | 1.00 | 0.979 | 0.983 | 0.978 | 1.00 | 1.00 | 0.981 |
| 3 years survival probability | 0.909 | 0.909 | 0.912 | 0.929 | 0.964 | 0.917 | 0.921 |
| 5 years Survival probability | 0.591 | 0.832 | 0.811 | 0.871 | 0.80 | 0.833 | 0.835 |



**The table of descriptive statistics for serum LDL is below:**

|  |  |  |
| --- | --- | --- |
|  | Serum LDL at study enrollment | All Subjects |
|  | 11-69mg/dL | 70-99mg/dL | 100-129mg/dL | 130-159mg/dL | 160-189mg/dL | 190-250mg/dL |
| N subjects | 22 | 143 | 228 | 225 | 83 | 24 | 725 |
| Mean mg/dL | 56.18 | 86.69 | 114.7 | 142.7 | 172.3 | 208.3 | 125.8 |
| Min mg/dL | 11 | 70 | 100 | 130 | 160 | 191 | 11 |
| Maxmg/dL | 69 | 99 | 129 | 159 | 189 | 247 | 147 |
| SDmg/dL | 13.64 | 8.246 | 8.365 | 8.528 | 9.215 | 13.48 | 33.60 |

**Based on the KM curves and the table of survival probabilities we notice that for the category with the lowest serum LDL we observe the worst survival experience, and the survival curve is significantly different from the survival curves of other categories. The intersection of many survival curves suggests that we should explore the possibility of non-linear trends and simply looking at the KM curves and table of survival probabilities we observe a different survival experience for different levels of serum LDL.
The table of descriptive statistics for the LDL variable is not very informative as we observe mostly what we would expect that the mean and standard deviation changes for serum LDL. We notice from the table that for the extreme LDL levels i.e. for very low serum LDL at 11-69 mg/dL and very high 190-250 mg/dL there are two main differences compared to the other categories. Firstly, those groups have a much smaller sample size which is reasonable because we would expect to see most of the subjects in ‘average’ range for serum LDL and secondly, we observe that the variance of serum LDL in these two categories is higher than the other categories in the ‘medium’ range.

METHODS: Distributions of time to death from any cause was compared across groups defined by serum LDL using proportional hazards regression modeling serum LDL as a categorical dummy variable where the categories are created based on the quantification by Mayo clinic. We use the category with the lowest serum LDL as our reference group. Quantification of association between all cause mortality was summarized by the hazards ratio computed from the regression model, with confidence intervals and two-sided p values computed using Wald statistics based on the Huber-White sandwich estimator. Subjects with missing data for serum LDL at the time of study accrual were omitted from the analysis.
INFERENCE: Based on the cox regression using the lowest LDL category as the reference group we observe that the hazards are 60.2%, 60.7%, 70.6%, 74.3% 68.3% lower than the category of serum LDL 11-69 mg/dL. The percentages correspond to the groups of 70-99 mg/dL, 100-129 mg/dL, 130-159 mg/dL and 190-250 mg/dL serum LDL. For each of them we observe 95% CI which cover the region <1 and the highest upper limit of a confidence interval is 0.9892 <1. The two sided P-value using a chi-squared distribution with 5 df for testing the significance of this variable is 0.0087 < .05 and so we reject the null hypothesis in favor of the alternative that there is some association between serum LDL and all-cause mortality**

**Even if we look at individual parameters in this case we observe that all of them show a low p-value.**

* 1. Provide an interpretation for each parameter in your regression model, including the intercept.
	**ANSWER: For the regression model above the intercept is when we set all the other variables =0 which in this case corresponds to being in the group with serum LDL 11-69 mg/dL and in this case the intercept represents the hazard for the subjects with serum LDL levels between 11 mg/dL to 69 mg/dL. The estimates of the slopes can be interpreted as the hazard ratio between the two groups. So for a given category, say category i, the slope estimates**

 **where the denominator is the hazard of the reference group, i.e. 11-69 mg/dL serum LDL.**

* 1. What analysis would you perform to assess whether the regression model used in this problem provides a “better fit” than does a model that uses only a continuous linear term for LDL? What is the result of such an analysis?
	**ANSWER: To test linearity we create a nested model scenario by considering the null hypothesis as the linear model with modelling serum LDL as a linear continuous predictor and the alternative as the full model with dummy variables for serum LDL and a continuous linear predictor for the serum LDL. We then compare the two models using a Wald type-test. For this test we observed a two sided p-value of 0.3988 which means that we do not enough statistical evidence to reject the null hypothesis that the model is linear and so we do not have enough evidence to suggest the model is not non-linear.**
	2. For each population defined by serum LDL value, compute the hazard ratio relative to a group having serum LDL of 160 mg/dL. (This will be used in problem 4). This can be effected by generating fitted hazard ratio estimates for each individual in the sample, and then dividing that fitted value by the fitted value for a subject having a LDL of 160 mg/dL.
	**ANSWER:** **The relative hazard was evaluated in STATA**
1. Perform a statistical regression analysis evaluating an association between all-cause mortality and serum by comparing the instantaneous risk (hazard) of death over the entire period of observation across groups defined by serum LDL when fit as linear splines using the categories suggested by the Mayo Clinic as reported on Homework #1. The Stata mkspline command can be used to create the predictors that can be used in a regression

mkspline ldl0 70 ldl70 100 ldl100 130 ldl130 160 ldl160 190 ldl190 = ldl

* 1. Include full description of your methods, appropriate descriptive statistics, and full report of your inferential statistics.
	**ANSWER: For descriptive statistics please refer to question 2 which shows KM curves for each category of the serum LDL variable.
	METHODS: Distributions of time to death from any cause was compared across groups defined by serum LDL using proportional hazards regression with linear splines modeling serum LDL as a continuous variable where the knots for the splines were created based on the quantification by Mayo clinic. Quantification of association between all cause mortality was summarized by the hazards ratio computed from the regression model, with confidence intervals and two-sided p values computed using Wald statistics based on the Huber-White sandwich estimator. Subjects with missing data for serum LDL at the time of study accrual were omitted from the analysis.
	INFERENCE: Based on the cox regression we obtained the estimates for the changes in hazard for two groups of observation which differ by one and are within the same two knot points. Based on this for patients with serum LDL between 11-69 mg/dL we observed that for every 1 mg/dL the hazard was 2.20% lower for the group with higher LDL (95% CI 3.98 to 0.367 lower for the higher group). Similarly, for patients with serum LDL between 70-99 mg/dL we observed that for every 1 mg/dL the hazard was 2. 03% lower for the group with higher LDL (95% CI 4.65 lower to 0. 670 higher for the higher group). For patients with serum LDL between 100-129 mg/dL we observed that for every 1 mg/dL the hazard was 0.229% lower for the group with higher LDL (95% CI 2.36 lower to 1.95 higher for the higher LDL group). For patients with serum LDL between 130-159 mg/dL we observed that for every 1 mg/dL the hazard was 0.361% higher for the group with higher LDL (95% CI 2.06 lower to 2.84 higher for the higher LDL group). For patients with serum LDL between 160-189 mg/dL we observed that for every 1 mg/dL the hazard was 2.91% lower for the group with higher LDL (95% CI 7.02 lower to 1.38 higher for the higher LDL group). For patients with serum LDL between 190-250 mg/dL we observed that for every 1 mg/dL the hazard was 2.88% higher for the group with higher LDL (95% CI 2.09 lower to 8.10 higher for the higher LDL group). Since serum LDL is a single predictor we are interested in we look at the two-sided P-value for the entire group of variables where the grouping refers to the different knots/splines. The highly significant two-sided P-value < 0.00005 and thus we reject the null hypothesis in favor of the alternative that serum LDL is associated with the hazard.**
	2. Provide an interpretation for each parameter in your regression model, including the intercept.
	**ANSWER: The intercept in this model is simply the hazard for the group of subjects with serum LDL 0 mg/dL. In this case the intercept does not have a scientifically relevant interpretation as we did not sample anyone with less than 11 mg/dL serum LDL level and it is not possible to imagine anyone with 0 mg/dL serum LDL esp. for the subjects aged 65-99 years which is the range of our sample for age.
	The first parameter estimate is the hazard ratio which comparing two subjects who differ by 1 mg/dL and but have total serum LDL within the range 11-69 mg/dL. Similarly, the other parameters are the hazard ratio comparing two subjects differing by 1 mg/dL but have total serum LDL in the ranges 70-99 mg/dL, 100-129 mg/dL, 130-159 mg/dL, 160- 189 mg/dL and 190-247 mg/dL respectively.**
	3. What analysis would you perform to assess whether the regression model used in this problem provides a “better fit” than does a model that uses only a continuous linear term for LDL? What is the result of such an analysis?
	**ANSWER: To test linearity we notice that a straight line is a special case of a linear spline where we only have 2 knots which are the range of the data for the predictor of interest in which case the estimated coefficients for each spline will equal to each other and thus we can test linearity by testing the null hypothesis that all the coefficients are zero versus the alternative that they are not equal. Based on a Chi square test statistic with 5 degrees of freedom (Wald type test) we get a 2 sided P-value 0.0788 >0.05 form which we conclude that we don’t have enough statistical evidence to reject the null hypothesis that the relationship can be explained by a linear model, i.e. we don’t have enough evidence to conclude the relationship is non-linear.**
	4. For each population defined by serum LDL value, compute the hazard ratio relative to a group having serum LDL of 160 mg/dL. (This will be used in problem 4). This can be effected by generating fitted hazard ratio estimates for each individual in the sample, and then dividing that fitted value by the fitted value for a subject having a LDL of 160 mg/dL.
	**ANSWER:** **The relative hazard was evaluated in STATA**
1. By answering the following questions, compare the relative advantages and disadvantages of the various statistical analysis strategies we have considered in Homeworks 1-4 and problems 2 and 3 in this homework.
	1. What advantages do the regression strategies used in Homeworks 4 and 5 provide over the approaches used in Homeworks 1-3?
	**ANSWER: In Home works 1-3 we dichotomized death into a binary variable for death within 5 years or after 5 years. Dichotomizing the variable leads to loss of information and is usually undesirable. Whereas in 4 and 5 we treat death as a continuous censored variable which would lead to an analysis with the highest statistical power and would allow us to best address the scientific question of detecting association between serum LDL and all cause mortality. Also using the dichotomization we divided the variable death at 5 years which was an arbitrary distinction.
	In homework 4-5 we not only had the added advantage of treating the response as a continuous censored variable but also flexible methods which allowed us to capture non-linear trends.**
	2. Comment on any similarities or differences of the fitted values from the three models fit in Homework 4 and the two models fit in problems 2 and 3 of this homework.
	**ANSWER: Below are the plots from HW4 AND the plots for this HW for the plots for the two different fits in this model.**
	

	
	 **The main difference we note in the above plots is that the one from HW4 fitted more smooth functions for the relative hazard as opposed to the fits above. Using dummy variables we clearly see a step function and we observe that the step function captures a very slight u-shaped trend like the quadratic fit from HW4 however this is likely not significant. The linear spline captures the rise in hazard at extreme values of serum LDL like the quadratic fit however it recognizes an increasing gradient from serum LDL 100-160 mg/dL which was not observed in any of the other plots.**
	3. *A priori*, of all the analyses we have considered for exploring an (unadjusted) association between all cause mortality and serum LDL in an elderly population, which one would you prefer and why?
	**ANSWER: A priori I would have used the linear splines to model the hazard because it has the advantage of not losing information by categorizing the predictor or the response as opposed to q2 where we categorized the predictor of interest hence losing power or HW1-3 where we dichotomized the response. And secondly, this method is the most flexible as it allows us to detect non-linear trends which a linear model does not and it does not impose a specified shape to the relationship i.e. quadratic or log. We noticed in the quadratic model an increase in hazard for very high serum LDL levels and this was enforced by the fact that we were using a quadratic model. In conclusion, I would use the linear splines method because it uses the most information in the data while making the least amount of assumptions.**

**Discussion Sections: February 3 - 7, 2014**

We continue to discuss the dataset regarding FEV and smoking in children. Come do discussion section prepared to describe descriptive statistics, especially as they relate to confounding, precision, effect modification, and the impact of heteroscedasticity.