**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.

This model is saturated, because the number of parameters (diabetes and race) is matched by the number of possible values of the predictor variable (white or non-white).

Methods:

The odds of subjects having been diagnosed with diabetes prior to the MRI at the inception of the study were compared between subjects who were white and subjects who were another race. The odds of having been diagnosed with diabetes prior to the MRI was estimated using a simple logistic regression with an indicator of the subject’s race being white as the predictor variable, using the Huber-White Sandwich estimator to compute robust standard errors. An estimate of the odds of prior diabetes diagnosis given a non-white race as well as an estimate of the odds ratio comparing the white group with the non-white group were produced with 95% confidence intervals and two-sided p-values as computed based on the Wald statistic.

Results:

Of the 162 subjects whose race was non-white, the odds of having been diagnosed with diabetes prior to the MRI was 0.164, while for the subjects whose race was white the odds of prior diabetes diagnosis was 0.109. Based on a 95% confidence interval, this observed odds ratio of 0.661 for the comparison of the white group to the non-white group would not be considered unusual if the true odds ratio were anywhere between 0.0.392 and 1.113. A two-sided p value of 0.119 suggests that we cannot with high confidence reject the null hypothesis that the odds of prior diabetes diagnosis are not associated with being white.

* 1. Using the regression model fit in part (a), provide an interpretation for each of the regression parameters (including the intercept).

A simple logistic regression using an indicator of participant race as white or non-white as a predictor variable and an indicator of a participant’s diagnosis of diabetes before the MRI as a response variable produces a βo value estimate of -1.806 and a β1 value estimate of -0.415, for use in the following equation:

$$Yi= e^{βo}\*e^{β1\*Xi}$$

The term Yi represents the odds of a proir diagnosis of diabetes, and the term Xi represents the indicator variable that takes on the value 1 when the participant’s race is white, and 0 when the participant’s race is not white.The term $e^{βo}$ , estimated in this case to be 0.164, is the intercept and represents the odds of a prior diagnosis of diabetes given an Xi value of 0, which equals the odds of prior diagnosis given a non-white race. The term $e^{β1}$ represents the slope, estimated in this case to be 0.661, and is the odds ratio between white and non-white races and between prior diabetes diagnosis and no prior diabetes diagnosis.

* 1. 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.

The two-sided p value of 0.119 (>0.05) suggests that we cannot with high confidence reject the null hypothesis that the odds of prior diabetes diagnosis are not associated with being white.

* 1. 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)?

The formal inference would differ slightly from that provided in part (a). The odds of prior diabetes diagnosis for those whose race is black is 0.209, and for those whose race is not black the odds of prior diagnosis is 0.107. The odds ratio between blacks and non-blacks and prior diagnosis and non-prior-diagnosis is 1.956, as opposed to the value of 0.661 for the odds ratio comparing whites and non-whites. The p-value is 0.0221, which suggests stastistical significance, that the regression in part (a) does not yield. Since the non-whites category used in part (a) includes the blacks, and the non-blacks category used here includes the whites, these regressions are related due to reparamaterizatin and essentially convey similar information despite apparent differences in statistical significance.

* 1. Using the regression model fit in part (d), provide an interpretation for each of the regression parameters (including the intercept.)

A simple logistic regression using an indicator of participant race as black or non-black as a predictor variable and an indicator of a participant’s diagnosis of diabetes before the MRI as a response variable produces a βo value estimate of -2.235 and a β1 value estimate of 0.671, for use in the following equation:

$$Yi= e^{βo}\*e^{β1\*Xi}$$

The term Yi represents the odds of a proir diagnosis of diabetes, and the term Xi represents the indicator variable that takes on the value 1 when the participant’s race is black, and 0 when the participant’s race is not black.The term $e^{βo}$ , estimated in this case to be 0.107, is the intercept and represents the odds of a prior diagnosis of diabetes given an Xi value of 0, which equals the odds of prior diagnosis given a non-black race. The term $e^{β1}$ represents the slope, estimated in this case to be 1.956, and is the odds ratio between black and non-black races and between prior diabetes diagnosis and no prior diabetes diagnosis.

* 1. 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.

The two-sided p value of 0.0221 (<0.05) suggests that we can with confidence reject the null hypothesis that the odds of prior diabetes diagnosis are not associated with being black.

* 1. 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)?

By using different dummy variables, the difference in model output parameters demonstrates that each variable is important to the model. However, the differences in statistical significance between the two dummy variable regressions may mislead someone into discounting an association that exists with one of the variables (a type II or beta error).

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 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:

*descriptive statistics methods:*

Descriptive statistics for the censoring distribution included the minimum and maximum observed censoring times. Descriptive statistics for serum LDL levels included the number of cases with missing data, as well as the minimum, maximum, mean, standard deviation, and the 25th, 50th (median), and 75th percentiles for the cases with available data.

For the purposes of descriptive statistics of the survival probabilities by serum LDL level, serum LDL was categorized according to the Mayo Clinic guidelines: less than 70 mg/dL, 70-99 mg/dL, 100-129 mg/dL, 130-159 mg/dL, 160-189 mg/dL, and greater than or equal to 190 mg/dL. Within these categories, Kaplan-Meier estimates of survival were calculated and graphed, and estimates of the 2 and 5 year survival probabilities were determined.

*inferential statistics methods:*

Distributions of time to death from any cause was compared across groups defined by serum LDL at baseline using proportional hazards regression modeling serum LDL as a continuous untransformed random variable fit as dummy variables by categorizing according to the Mayo Clinic guidelines (<70, 70-100, 100-130, 130-160, 160-190, and >190 mg/dL) . 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 missing data for serum LDL at the time of study accrual were omitted from the analysis.

Descriptive Statistics:

The minimum censored observation time (the observation time at which the participant left the study without having died) is 1827 days (about 5.0 years), the maximum is 2159 days (about 5.9 years). 10 cases were missing data for serum LDL levels. Of those remaining, the minimum serum LDL level was 11 mg/dL and the maximum was 247 mg/dL, with a mean of 125.8 mg/dL, and a standard deviation of 33.6. The 25th percentile for serum LDL was 102 mg/dL, the median was 125 mg/dL and the 75th percentile was 147 mg/dL.

The Kaplan–Meier estimates of survival for the Mayo Clinic recommended categories of serum LDL levels are plotted below, and the number of subjects total, number of subject deaths, and 2 and 5 year survival probabilities are summarized in the following table.





Inferential Statistics:

An analysis of total all-cause mortality within the study period as predicted by various categories of serum LDL levels finds that the observed difference in instantaneous risk of death is no greater than what might reasonably be expected when serum LDL level had no true effect (P= 0.091; >0.05). The group having serum LDL between 70 and 100 mg/dL is estimated to have a hazard ratio of 0.476 between groups differing in serum LDL level by 1 mg/dL (95% confidence interval unadjusted for multiple comparisons: 0.232 to 0.978). The group having serum LDL between 100 and 130 mg/dL is estimated to have a hazard ratio of 0.458 between groups differing in serum LDL level by 1 mg/dL (95% confidence interval unadjusted for multiple comparisons: 0.0.230 to 0.911). The group having serum LDL between 130 and 160 mg/dL is estimated to have a hazard ratio of 0.377 between groups differing in serum LDL level by 1 mg/dL (95% confidence interval unadjusted for multiple comparisons: 0.186 to 0.761). The group having serum LDL between 160 and 190 mg/dL is estimated to have a hazard ratio of 0.294 between groups differing in serum LDL level by 1 mg/dL (95% confidence interval unadjusted for multiple comparisons: 0.122 to 0.705). The group having serum LDL greater than 190 mg/dL is estimated to have a hazard ratio of 0.381between groups differing in serum LDL level by 1 mg/dL (95% confidence interval unadjusted for multiple comparisons: 0.119 to 1.216).

* 1. Provide an interpretation for each parameter in your regression model, including the intercept.

The exponentiated slope estimated by the regression, e^(β1), represents for each stratum of serum LDL values the hazard ratio between groups differing in value of the predictor by 1 unit (in this case, 1 mg/dL of serum LDL), calculated by which each stratum is treated as a dummy variable with the value of 1 and all other stratums are considered “0”. For example, the value of 0.476 returned by the regression model for the exponentiated coefficient for stratum [100-130 mg/dL] is the hazard ratio for a cox proportional hazards regression using a dummy variable as a predictor with the value of “1” if the participant has serum LDL between 100 and 130 mg/dL, and a value of “0” otherwise.

The intercept is not generated by the model, and is difficult to interpret. It represents the baseline hazard, when for the following equation:

$$λ\left(Xi\right)=λ\_{0}\left(t\right)\*e^{β1\*Xi}$$

Xi is equal to zero, and the instantaneous risk of death is $λ\_{0}$(t).

* 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?

I would perform an analysis comparing the root mean squared error (RMSE) of the fitted model using dummy variables with that of the fit treating LDL as a continuous linear term, or compare the R^2 values between the two fits.

The R^2 value for the linear fit (a cox proportional hazards regression using LDL as a continuous predictor variable) is 0.011, and the R^2 for the fit using dummy variables is 0.011. The two fits appear to have equivalent coefficients of variation.

* 1. 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.

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.

Methods:

*descriptive statistics methods:*

Descriptive statistics for the censoring distribution included the minimum and maximum observed censoring times. Descriptive statistics for serum LDL levels included the number of cases with missing data, as well as the minimum, maximum, mean, standard deviation, and the 25th, 50th (median), and 75th percentiles for the cases with available data.

For the purposes of descriptive statistics of the survival probabilities by serum LDL level, serum LDL was categorized according to the Mayo Clinic guidelines: less than 70 mg/dL, 70-99 mg/dL, 100-129 mg/dL, 130-159 mg/dL, 160-189 mg/dL, and greater than or equal to 190 mg/dL. Within these categories, Kaplan-Meier estimates of survival were calculated and graphed, and estimates of the 2 and 5 year survival probabilities were determined.

*inferential statistics methods:*

Distributions of time to death from any cause was compared across groups defined by serum LDL at baseline using proportional hazards regression modeling serum LDL as a continuous untransformed random variable fit as linear splines by categorizing according to the Mayo Clinic guidelines (<70, 70-100, 100-130, 130-160, 160-190, and >190 mg/dL) . 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 missing data for serum LDL at the time of study accrual were omitted from the analysis.

Descriptive Statistics:

The minimum censored observation time (the observation time at which the participant left the study without having died) is 1827 days (about 5.0 years), the maximum is 2159 days (about 5.9 years). 10 cases were missing data for serum LDL levels. Of those remaining, the minimum serum LDL level was 11 mg/dL and the maximum was 247 mg/dL, with a mean of 125.8 mg/dL, and a standard deviation of 33.6. The 25th percentile for serum LDL was 102 mg/dL, the median was 125 mg/dL and the 75th percentile was 147 mg/dL.

The Kaplan–Meier estimates of survival for the Mayo Clinic recommended categories of serum LDL levels are plotted below, and the number of subjects total, number of subject deaths, and 2 and 5 year survival probabilities are summarized in the following table.





Inferential Statistics:

An analysis of total all-cause mortality within the study period as predicted by various categories of serum LDL levels treated as linear splines finds that the observed difference in instantaneous risk of death is greater than what might reasonably be expected when serum LDL level had no true effect (P<0.05). The group having serum LDL less than 70 mg/dL is estimated to have a hazard ratio of 0.978 between groups differing in serum LDL by 1 mg/dL (95% confidence interval unadjusted for multiple comparisons: 0.960 to 0.996). The group having serum LDL between 70 and 100 mg/dL is estimated to have a hazard ratio of 0.980 between groups differing in serum LDL level by 1 mg/dL (95% confidence interval unadjusted for multiple comparisons: 0.953 to 1.01). The group having serum LDL between 100 and 130 mg/dL is estimated to have a hazard ratio of 0.998 between groups differing in serum LDL level by 1 mg/dL (95% confidence interval unadjusted for multiple comparisons: 0.976 to 1.03). The group having serum LDL between 130 and 160 mg/dL is estimated to have a hazard ratio of 1.00 between groups differing in serum LDL level by 1 mg/dL (95% confidence interval unadjusted for multiple comparisons: 0.979 to 1.03). The group having serum LDL between 160 and 190 mg/dL is estimated to have a hazard ratio of 0.971 between groups differing in serum LDL level by 1 mg/dL (95% confidence interval unadjusted for multiple comparisons: 0.930 to 1.01). The group having serum LDL greater than 190 mg/dL is estimated to have a hazard ratio of 1.03 between groups differing in serum LDL level by 1 mg/dL (95% confidence interval unadjusted for multiple comparisons: 0.979 to 1.08).

* 1. Provide an interpretation for each parameter in your regression model, including the intercept.

The exponentiated slope estimated by the regression, e^(β1), represents for each stratum of serum LDL values the hazard ratio between groups differing in value of the predictor by 1 unit (in this case, 1 mg/dL of serum LDL) and between the same knots, calculated by which between each pair of knots a linear fit is estimated.

The intercept is not generated by the model, and is difficult to interpret. It represents the baseline hazard, when for the following equation:

$$λ\left(Xi\right)=λ\_{0}\left(t\right)\*e^{β1\*Xi}$$

Xi is equal to zero, and the instantaneous risk of death is $λ\_{0}$(t).

* 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?

I would perform an analysis comparing the root mean squared error (RMSE) of the fitted model using dummy variables with that of the fit treating LDL as a continuous linear term, or compare the R^2 values between the two fits.

The R^2 value for the linear fit (a cox proportional hazards regression using LDL as a continuous predictor variable) is 0.011, and the R^2 for the fit using linear splines is 0.018. The linear splines fit has a higher coefficient of variation and thus appears to be a better fit.

* 1. 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.
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?

In Homeworks 1-3, various analyses were performed on the association between serum LDL and mortality by dichotomizing mortality into death in under 5 years and survival past 5 years. This dichotomization results in a loss of information about specific timepoints of mortality throughout the entire study period. The strategies for regression used in HW 4 and 5 treat the output variable of mortality as continuous and thus capture more of the mortality trends across all timepoints during the observation period.

* 1. 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.

This is the part where, if I had had enough time, I would have included a plot of fitted hazard ratios as predicted by each Cox Proportional Hazards Regression for LDL=160 for problems 1-3 of HW#4, and problems 2 and 3 of this HW. I would then compare the apparent trends in the fitted curves.

* 1. *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?

For the sake of scientific relevance, I would prefer an undichotomized (multiplicative) Cox Proportional Hazards regression model, using serum LDL as an untransformed linear predictor term. I have no scientific reason to suspect a quadratic or otherwise nonlinear relationship, and I do not want to risk the potential loss in statistical significance resulting from multiple comparisons of covariates.

**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.