***Biost 518: Applied Biostatistics II***

***Biost 515: Biostatistics II***

*Emerson, Winter 2014*

***Homework #2***

*January 13, 2014*

***Written problems:*** *To be submitted as a MS-Word compatible file to the class Catalyst dropbox by 9:30 am on Tuesday, January 21, 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.***

*This homework builds on the analyses performed in homework #1, 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.*

1. *Perform statistical analyses evaluating an association between serum LDL and 5 year all-cause mortality by comparing mean LDL values across groups defined by vital status at 5 years using a t test that presumes equal variances across groups. Depending upon the software you use, you may also need to generate descriptive statistics for the distribution of LDL within each group defined by 5 year mortality status. As this problem is directed toward illustrating correspondences between the t test and linear regression, you do not need to provide full statistical inference for this problem. Instead, just answer the following questions.*
	1. *What are the sample size, sample mean and sample standard deviation of LDL values among subjects who survived at least 5 years? What are the sample size, sample mean and sample standard deviation of LDL values among subjects who died within 5 years? Are the sample means similar in magnitude? Are the sample standard deviations similar?*

For subjects who survived at least 5 years, the sample size is n=606, with 8 samples missing. The sample mean is 127.198 (mg/dl), and the standard deviation is 32.929. For subjects who died within 5 years, the sample size is n=119, with 2 samples missing. The sample mean is 118.697 (mg/dl), and the standard deviation is 36.157. The sample means have a percent difference of 6.91% and are both within the serum LDL range recommended by the Mayo Clinic as “Near ideal” (100-129 mg/dl), and are thus similar in magnitude. The sample standard deviations have a percent difference of 9.34% and are thus similar in magnitude.

* 1. *What are the point estimate, the estimated standard error of that point estimate, and the 95% confidence interval for the true mean LDL in a population of similar subjects who would survive at least 5 years? What are the corresponding estimates and CI for the true mean LDL in a population of similar subjects who would die within 5 years? Are the point estimates similar in magnitude? Are the standard errors similar in magnitude? Explain any differences in your answer about the estimates and estimated SEs compared to your answer about the sample means and sample standard deviations.*

In a population of subjects similar to those within the study sample, of those who would survive at least 5 years, the point estimate for the true mean LDL is 127.198 mg/dl, with a 95% confidence interval of [124.6, 129.8] and a standard error of 1.338. In a population of similar subjects who would die within 5 years, the point estimate for the true mean LDL is 118.697 mg/dl with a 95% confidence interval of [112.1, 125.3] and a standard error of 3.315. The point estimates are similar in magnitude, with a percent difference of 6.91% and common residence within the “Near ideal” range for serum LDL values recommended by the Mayo Clinic (100-129 mg/dl). The standard errors have a percent difference of 85.0% (a difference of 1.98), though as they relate to the means their magnitudes are similar (values of mean serum LDL within 1 standard error still fall approximately within the “Near ideal” serum LDL range for either 5 year survivors or those that die within 5 years). The point estimates for the mean serum LDL values of entire populations are taken directly from the means of the sample values, so they are exactly equal to the sample means. However, the standard deviation of the sample means are calculated directly from the sample as the square root of the variance of the sample, whereas the standard error is calculated using the sample standard deviation divided by the square root of the sample size.

* 1. *Does the CI for the mean LDL in a population surviving 5 years overlap with the CI for mean LDL in a population dying with 5 years? What conclusions can you reach from this observation about the statistical significance of an estimated difference in the estimated means at a 0.05 level of significance?*

The 95% confidence interval for mean LDL values among those that survive 5 years ([124.6, 129.8] overlaps with the 95% confidence interval for mean LDL values among those that die within 5 years ([112.1, 125.3]). This suggests that even with a p-value less than 0.05, which would indicate statistical significance of the finding to an alpha level of 0.05, there is a possibility that rejecting the null hypothesis that the difference between the two sample means is zero would lead to a type I alpha error in which the null hypothesis reflects reality and is incorrectly rejected.

* 1. *If we presume that the variances are equal in the two populations, but we want to allow for the possibility that the means might be different, what is the best estimate for the standard deviation of LDL measurements in each group? (That is, how should we combine the two estimated sample standard deviations?)*

When it can be assumed that the two populations have the same variance but unique means, the standard deviations can be obtained by calculating a pooled standard deviation by the following formula:

$$\sqrt{s\_{x1x2}^{2}}=\sqrt{\frac{\left(N1-1\right)s\_{x1}^{2}+(N2-1)s\_{x2}^{2}}{N1+N2-2}}$$

The value for the pooled standard deviation of the LDL measurements in each group is 33.47872.

* 1. *What are the point estimate, the estimated standard error of the point estimate, the 95% confidence interval for the true difference in means between a population that survives at least 5 years and a population that dies with 5 years? What is the P value testing the hypothesis that the two populations have the same mean LDL? What conclusions do you reach about a statistically significant association between serum LDL and 5 year all cause mortality?*

For a true difference in mean serum LDL values between a population that survives at least 5 years and a population that dies within 5 years, the point estimate is -8.5005 with a 95% confidence interval of [-15.090491, -1.910591] and a standard error of 8.311797. A Welch 2-sample t-test was performed assuming equal variances to test the 2-sided null hypothesis the difference in means is equal to zero, and yielded a p-value of 0.01154(<0.05) which indicates statistical significance at a 0.05 alpha level. The null hypothesis that the difference in means is equal to zero can be rejected, and the statistical significance suggests there is not a lack of association between all-cause 5-year mortality and serum LDL level.

1. *Perform statistical analyses evaluating an association between serum LDL and 5 year all-cause mortality by comparing mean LDL values across groups defined by vital status at 5 years using ordinary least squares regression that presumes homoscedasticity. As this problem is directed toward illustrating correspondences between the t test and linear regression, you do not need to provide full statistical inference for this problem. Instead, just answer the following questions.*
	1. *Fit two separate regression analyses. In both cases, use serum LDL as the response variable. Then, in model A, use as your predictor an indicator that the subject died within 5 years. In model B, use as your predictor an indicator that the subject survived at least 5 years. For each of these models, tell whether the model you fit is saturated? Explain your answer.*

Both Model A and Model B are saturated. Saturation of a linear regression model requires that the number of groups is equal to the number of parameters. The predictor variable for Model A, an indicator of death within 5 years (a dummy variable with the value of 0 if the subject is alive at 5 years, and 1 if dead at 5 years) has two possible values, or groups. The same is true with Model B, for which the predictor of interest is a dummy variable that is 1 if the subject is a live at 5 years, and 0 if the subject is dead by 5 years. Each linear regression has two parameters, the slope and the intercept. Thus, both models are saturated with two groups and two model parameters.

* 1. *Using the regression parameter estimates from one of your models (tell which one you use), what is the estimate of the true mean LDL among a population of subjects who survive at least 5 years? How does this compare to the corresponding estimate from problem 1?*

According to Model A, when the predictor variable is equal to 0 the resulting LDL values calculated using the model correspond to subjects who have survived at least 5 years. An ordinary least squares regression that assumes homoscedasticity yields the following equation, where Yi is the serum LDL level (in mg/dl), and Xi is the variable indicating whether the subject survived 5 years (0) or died within 5 years (1):

$$Y\_{i}=127.198-8.501X\_{i}$$

By plugging in a value of “0” to this equation in place of Xi, Yi is found equal to 127.2 mg/dl, which is both the intercept of the linear regression equation and the estimate of the true mean LDL among a population of subjects who survive at least 5 years. This is exactly equal to the value obtained from problem 1’s point estimate of the mean serum LDL values in a similar population of people who survive at least 5 years.

* 1. *Using the regression parameter estimates from one of your models (tell which one you use), what is a confidence interval for the true mean LDL among a population of subjects who survive at least 5 years? How does this compare to the corresponding estimate from problem 1? Explain the source of any differences.*

According to Model A, a 95% confidence interval for the true mean LDL among a population of subjects who survive at least 5 years is [124.528, 129.868]. Compared to that calculated in problem 1, [124.6, 129.8], the values are similar, differing by only 0.1 or 0.2 mg/dl. This difference is due to the use of pooled standard deviations for each sample group used in the regression compared with individual standard deviations used for a t-test.

* 1. *Using the regression parameter estimates from one of your models (tell which one you use), what is the estimate of the true mean LDL among a population of subjects who die within 5 years? How does this compare to the corresponding estimate from problem 1?*

According to Model B, an estimate of the true mean LDL among a population of subjects who die within 5 years is obtained by plugging in a value of 0 for Xi in the following equation, obtained via ordinary least squares linear regression assuming homoscedasticity:

$$Y\_{i}=118.697+8.501X\_{i}$$

This yields the intercept for Model B, 118.697 mg/dL, which is equal to an estimate of the mean LDL among a population of subjects who die within 5 years. This occurs because the indicator variable for Model B has a value of 0 if the sample subject died within 5 years, and a value of 1 if the sample subject survived past 5 years. This is exactly equal to the point estimate for the mean LDL value in a population of people who die within 5 years found in problem 1.

* 1. *Using the regression parameter estimates from one of your models (tell which one you use), what is a confidence interval for the true mean LDL among a population of subjects who die within 5 years? How does this compare to the corresponding estimate from problem 1? Explain the source of any differences.*

According to Model B, a 95% confidence interval for the true mean LDL value among a population of subjects who die within 5 years is [112.673, 124.722]. Compared to that calculated in problem 1, [112.1, 125.3], the values are similar, differing by 0.4 or 0.5 mg/dL. This difference is due to the use of pooled standard deviations for each sample group used in the regression compared with individual standard deviations used for a t-test.

* 1. *If we presume the variances are equal in the two populations, what is the regression based estimate of the standard deviation within each group for each model? How does this compare to the corresponding estimate from problem 1?*

For Model A, the regression based estimate of the standard deviation within each group is 33.45392. For Model B, the value is also 33.45392. For problem 1, the estimated value for the standard deviation was found to be 33.47872. This value is close but slightly different likely due to rounding discrepancies in the two methods.

* 1. *How do models A and B relate to each other?*

Models A and B both utilize predictor variables that are indicators of 5-year vital status. However, Model A has a predictor variable that has a value of 0 for a subject who survives past 5 years, and a value of 1 if the subject dies within 5 years; Model B has a predictor variable that has a value of 0 for a subject who dies within 5 years, and a value of 1 if the subject survives past 5 years. Thus, when the predictor variable for Model A is equal to 0, yielding an outcome variable equal to the intercept. that intercept represents the mean LDL value for subjects who survive past 5 years. For Model B, the intercept represents the mean LDL value for subjects who die within 5 years. However, the mean LDL value for the other group can be calculated from either model by plugging in a value of “1” in the place of the predictor variable in the linear regression model equation. The mean LDL values for each group are the same for both models.

* 1. *Provide an interpretation of the intercept from the regression model A.*

The intercept from the regression Model A is equal to the point estimate of the mean serum LDL value in a population of subjects that survive past 5 years.

* 1. *Provide an interpretation of the slope from the regression model A.*

The slope from regression Model A is equal to the difference in the point estimates of the means between the group of subjects who die within 5 years and those who survive past 5 years.

* 1. *Using the regression parameter estimates, what are the point estimate, the estimated standard error of the point estimate, the 95% confidence interval for the true difference in means between a population that survives at least 5 years and a population that dies within 5 years? What is the P value testing the hypothesis that the two populations have the same mean LDL? What conclusions do you reach about a statistically significant association between serum LDL and 5 year all cause mortality? How does this compare to the corresponding inference from problem 1?*

The point estimate for the true difference in means between a population that survives at least 5 years and a population that dies within 5 years is equal to the slope provided by regression Model B, which is 8.501 (mg/dl), with a 95% confidence interval of [1.911, 15.090] and a standard error of 3.357. A P-value testing the null hypothesis that the two populations have a difference in mean LDL equal to zero is 0.01154 (<0.05) which indicates statistical significance at a 0.05 alpha level. The null hypothesis that the difference in means is equal to zero can be rejected, and the statistical significance suggests there is not a lack of association between all-cause 5-year mortality and serum LDL level. This inference is the same as in problem 1.

1. *Perform statistical analyses evaluating an association between serum LDL and 5 year all-cause mortality by comparing mean LDL values across groups defined by vital status at 5 years using a t test that allows for the possibility of unequal variances across groups. How do the results of this analysis differ from those in problem 1? (Again, we do not need a formal report of the inference.)*

A Welch two-sample t-test that allows for the possibility of unequal variances testing the two-sided null hypothesis that the true difference in mean serum LDL values is equal to zero was performed across groups defined by vital status at 5 years. The point estimate for the difference in mean LDL values is -8.5005 mg/dl with a 95% confidence interval of [-15.55976, -1.44132] and a p-value of 0.01858 (<0.05). This suggests statistical significance (at the 0.05 alpha level) sufficient enough to allow rejection of the null hypothesis that the difference in means is equal to zero. The analysis of problem 1 yielded a slightly different 95% confidence interval and p-value ([-15.090491, -1.910591] and 0.01154, respectively). The difference here is due to the assumption of homoscedasticity made for problem 1, and the relaxing of that assumption for the current problem (3).

1. *Perform statistical analyses evaluating an association between serum LDL and 5 year all-cause mortality by comparing mean LDL values across groups defined by vital status at 5 years using a linear regression model that allows for the possibility of unequal variances across groups. How do the results of this analysis differ from those in problem 3? (Again, we do not need a formal report of the inference.)*

A robust least-squares linear regression using the Huber-White sandwich estimator that allows for heteroscedasticity was performed, using as a predictor variable an indication that a subject died within 5 years. The estimate for the intercept (and the estimate for the mean LDL value in a population of subjects who survive past 5 years) is 127.198 mg/dl with a 95% confidence interval of [124.570, 129.826] and a standard error of 1.338. The estimate for the slope (and the estimate for the difference in the mean LDL values between a population of subjects who die within 5 years and those that survive past 5 years) is -8.501 mg/dl with a 95% confidence interval of [-15.501, -1.500] and a standard error of 3.566. The point estimate for the difference in means is the same value as calculated for problem 3, but the 95% confidence interval is slightly different due to the use of a pooled standard deviation for the regression that is not done as part of the t-test in problem 3.

1. *Perform a regression analysis evaluating an association between serum LDL and age by comparing the distribution of LDL across groups defined by age as a continuous variable. (Provide formal inference where asked to.)*
	1. *Provide descriptive statistics appropriate to the question of an association between LDL and age. Include descriptive statistics that would help evaluate whether any such association might be confounded or modified by sex. (But we do not consider sex in the later parts of this problem.)*

|  |  |
| --- | --- |
|  | Serum LDL (mg/dl) |
|  | <70 | 70-100 | 100-129 | 130-159 | 160-189 | >190 |
| Age (years) | 75 | 74.59 | 74.64 | 74.2 | 74.57 | 75.96 |
| (mean, (sd, min-max)) | (5.37, 69.0-92.0) | (5.41, 67.0-90.0) | (5.08, 65.0-90.0) | (5.62, 67.0-99.0) | (5.67, 65.0-94.0) | (6.11, 67.0-87.0) |
| Sex (% male) | 65.38 | 57.24 | 53.95 | 43.11 | 48.19 | 20.83 |

The mean ages for each group of serum LDL (thresholds defined by the Mayo Clinic as pertaining to risk levels of coronary heart disease) are similar, with similar min-max ranges. However, the percent of male subjects within each serum LDL stratification decreases as serum LDL increases. Therefore, sex appears to be independently associated with the outcome, but not with the predictor of interest, and acting outside the pathway of interest, so it may introduce confounding to the association between serum LDL and age.

* 1. *Provide a description of the statistical methods for the model you fit to address the question of an association between LDL and age.*

In order to evaluate an association between serum LDL and age, a robust least-squares linear regression was conducted using the Huber-White sandwich estimator using the continuous variable of age as a predictor variable and the continuous variable of ldl as the output; heteroscedasticity was allowed for.

* 1. *Is this a saturated model? Explain your answer.*

This is not a saturated model. There are two parameters to the model, the intercept and the slope, but there are more than two possible values for the predictor variable since age is treated continuously by the model.

* 1. *Based on your regression model, what is the estimated mean LDL level among a population of 70 year old subjects?*

The estimated mean LDL level among a population of 70 year old subjects can be found by plugging in “70” in place of Xi in the following equation:

$$Y\_{i}=132.5281-0.09019X\_{i}$$

This yields an estimated mean of 126.2148 mg/dl for a population of 70 year old subjects.

* 1. *Based on your regression model, what is the estimated mean LDL level among a population of 71 year old subjects?* *How does the difference between your answer to this problem and your answer to part c relate to the slope?*

The estimated mean LDL level among a population of 71 year old subjects can be found by plugging in “71” in place of Xi in the following equation:

$$Y\_{i}=132.5281-0.09019X\_{i}$$

This yields an estimated mean of 126.1246 mg/dl for a population of 71 year old subjects. The slope in this equation represents the estimated difference in mean LDL values over a unit difference in age. Thus, adding 0.09019 to the mean LDL calculated for age=71 yields the mean LDL calculated for age=70.

* 1. *Based on your regression model, what is the estimated mean LDL level among a population of 75 year old subjects? How does the difference between your answer to this problem and your answer to part c relate to the slope?*

The estimated mean LDL level among a population of 75 year old subjects can be found by plugging in “75” in place of Xi in the following equation:

$$Y\_{i}=132.5281-0.09019X\_{i}$$

This yields an estimated mean of 125.7639 mg/dl for a population of 75 year old subjects. The slope in this equation represents the estimated difference in mean LDL values over a unit difference in age. For a difference of 5 units in age, adding [0.09019]\*5 to the mean LDL calculated for age=75 yields the mean LDL calculated for age=70.

* 1. *What is the interpretation of the “root mean squared error” in your regression model?*

The root mean squared error of the regression model is the standard deviation of the residual , which is the error distribution. For the linear equation in the form y=mx+b, an error term $e\_{i}$ is added to account for the effect of noise, yielding the following form for the linear regression equation:

$$Y\_{i}=132.5281-0.09019X\_{i}+ e\_{i}$$

The root mean squared error for this regression is 33.59838.

* 1. *What is the interpretation of the intercept? Does it have a relevant scientific interpretation?*

The intercept is the estimated mean value for the output variable, serum LDL level, that corresponds with the minimum age included in the regression’s calculation, which in this case is an age of zero years. Theoretically, infants under a year old may have LDL content in their serum, so the intercept is possibly relevant scientifically. However, this regression was formed using data from a sample of subjects age 65 years and older so it is possible that the linear relationship cannot be reasonably extrapolated to subjects in early development, from a scientific standpoint.

* 1. *What is the interpretation of the slope?*

The slope represents the estimated difference in mean LDL values over a unit difference in age. In this case, the slope of -0.09019 mg/dl/year means that a subject one year older than a second subject can be expected to have a serum LDL level 0.9019 mg/dl lower than that of the second subject.

* 1. *Provide full statistical inference about an association between serum LDL and age based on your regression model.*

According to the robust least-squares linear regression allowing for heteroscedasticity conducted using age as a predictor variable for serum LDL levels, the estimated mean difference over a unit (one year) increase in age of serum LDL level is -0.09019 mg/dl with a 95% confidence interval of [-0.54697, 0.36659] and a p-value testing the two-sided null hypothesis that the difference in means is equal to zero of p=0.6944 (>0.05). These results are not statistically significant at the 0.05 alpha level; additionally, the 95% confidence interval does not exclude the null value of zero. There is not sufficient evidence to reject the null hypothesis that the difference in means is equal to zero.

* 1. *Suppose we wanted an estimate and CI for the difference in mean LDL across groups that differ by 5 years in age. What would you report?*

The slope produced by the least-squares linear regression represents the difference in mean LDL across groups that differ by 1 year in age, and if this value is multiplied by 5 it would represent the difference in mean LDL across groups that differ by 5 years in age. The robust standard error produced by the regression for the slope is 0.23266, and this can be used to compute 95% confidence intervals for both the difference due to 1 year and due to 5 years.

* 1. *Perform a test for a nonzero correlation between LDL and age. How does your regression-based conclusion about an association between LDL and age compare to inference about correlation?*

The Pearson correlation coefficient was calculated between LDL and age, and the multiple R squared is 0.0002137. This indicates a very weak positive correlation, if any. This conclusion agrees well with the inference gleaned from the linear regression analysis that there is not sufficient evidence to assert that there is no lack of association.

***Discussion Sections: January 13 – 17, 2014***

*We will discuss the dataset regarding FEV and smoking in children. Come do discussion section prepared to describe the approach to the scientific question posed in the documentation file fev.doc.*