**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?
		1. **Among those who survived at least 5 years, there were 8 missing LDL values, yielding a sample size of 606 subjects with a mean LDL of 127.20 (SD=32.92). For those that died within 5 years, after accounting for the 2 missing LDL values giving us a sample size of 119 subjects, the sample mean becomes 118.69 (S.D.=36.15). The sample means are a bit different in magnitude, however the standard deviations seem quite similar.**
	2. 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.
		1. **The point estimate for the true mean LDL in a population of survivors past 5 years is 127.20 (S.E.=1.337). Based on a 95% confidence interval, this observed estimate of 127.20 would not be judged unusual if the true mean LDL among those having survived at least 5 years were anywhere between 124.58 and 129.82 mg/dL. In correspondence, the point estimate of the mean is 118.69 for those who would die within 5 years. The standard error is 3.018. Among those dead within 5 years and based on a 95% confidence interval, it would not be unusual if the true mean LDL were found between 113.60 and 123.80 mg/dL.**
		2. **The estimated means and sample means are equivalent, and this equality is not surprising. The sample mean is our point estimate. However, there does seem to exist a difference in the standard errors, and I would surmise the main source to be the sample size since standard error is simply the standard deviation weighted by sample size. The number of subjects that survived more than 5 years is about 5 times greater than the number that died within 5 years, so it is not surprising that the standard error for those dead within five years is more than 2 times bigger than the standard error for those dead after at least five years.**
	3. 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?
		1. **The point estimate for mean LDL within both populations of 5-year survivors and dead within 5 years will be obtained using the sample mean. Standard errors will be calculated and used for creating confidence intervals for both groups at the 5% significance level.**
		2. **For the population surviving 5 years, an estimate of 127.20 is observed, which would not be judged unusual if the true mean LDL among those having survived at least 5 years were anywhere between 124.58 and 129.82 mg/dL. While for the population dying within 5 years, the point estimate of the mean is 118.69. Among those dead within 5 years and based on a 95% confidence interval, it would not be unusual if the true mean LDL were found between 113.60 and 123.80 mg/dL. These confidence intervals do not overlap, which lead to the conclusion that an estimated difference of the estimated means will be statistically significant at the 5% level. In particular, we can reject the null hypothesis that the mean serum LDL levels are not different by vital status at 5 years in favor of a hypothesis that death within 5 years is associated with lower mean serum LDL.**
	4. 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?)
		1. **A pooled measurement of variance would be the most appropriate combination. In particular, use this pooled variance:**

$$s\_{p}^{2}=\frac{\left(n-1\right)s\_{x}^{2}+(m-1)s\_{y}^{2}}{n+m-2}$$

**, where sample standard deviations for two samples X and Y are weighted. This estimate combines both standard deviations, but it will not test difference between the groups. Furthermore, it is not a standard error estimate.**

* 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?
		1. **A 2-sample t-test assuming equal variances will test the hypothesis that the two populations have the same mean LDL. The standard error formula for unequal variances will be used to estimate standard error.**
		2. **The point estimate for the difference is estimated to be 8.5 mg/dL (95% C.I.: (1.91, 15.09)), with a standard error of 3.36. This difference in means is significantly different at the 5% level (P-value = 0.0115). The data supports an association between higher LDL and greater time until death values.**
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.
		1. **A linear regression model will be fit on LDL as a response. In Model A, an indicator variable for death at 5 years will be used as a predictor (“Dead at 5 Years”). Similarly, another Model B, will consider a predictor which indicates a subject is alive at 5 years (“Alive at 5 Years”).**
		2. **Model B:** $Serum LDL \left(\frac{mg}{dL}\right)=118.70+8.50\*Alive at 5 Years$
			1. **The coefficient for “Alive at 5 Years” presents with a standard error of 3.36, a t-value of 2.532, and p-value of 0.0115.**
		3. **Model A:** $Serum LDL \left(\frac{mg}{dL}\right)=127.20-8.50\*Dead at 5 Years$
			1. **The coefficient for “Dead at 5 Years” presents with a standard error of 3.36, a t-value of 2.532, and p-value of 0.0115.**
		4. **Each model is saturated since the number of predictors in each is 2, while the number of groups within each model is also 2, since our predictors are indicators.**
	2. 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?
		1. **To estimate the true mean LDL among subjects having survived at least 5 years I will use Model B by plugging in the appropriate value into the predictor (“Alive at 5 Years”=1).**
		2. **Model B estimates the true mean LDL among those having survived at least 5 years to be about 118.70+8.50\*1 = 127.20. This estimate is precisely the same as the one found in problem 1.**
	3. 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.
		1. **First, an estimate for the true mean LDL among subjects having survived at least 5 years will be obtained using a regression model (Model B) where serum LDL is a response of being alive at 5 years. The standard error for the variable “Alive at 5 Year’s” coefficient will be used to build a 95% confidence interval around the estimate of the true mean stated previously.**
		2. **The true mean LDL among those having survived at least 5 years is estimated as 127.20. Based on a 95% confidence interval, this observed estimate of 127.20 would not be judged unusual if the true mean LDL among those having survived at least 5 years were anywhere between 120.61 and 133.70 mg/dL. This confidence interval is different from the one found in problem 1. This difference is a result of the standard errors being calculated in different manners. In particular, it should be noted that the standard error in problem 1 considers only the sample relevant to itself (i.e. those alive at 5 years). However, the standard error for the coefficient is an estimate of the standard deviation across all cases.**
	4. 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?
		1. **To estimate the true mean LDL among subjects having died within 5 years I will use for consistency’s sake Model B by plugging in the appropriate value into the predictor (“Alive at 5 Years”=0).**
		2. **Model B estimates the true mean LDL among those having survived at least 5 years to be about 118.70. This estimate is precisely the same as the one found in problem 1.**
	5. 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.
		1. **First, an estimate for the true mean LDL among subjects having died within 5 years will be obtained using a regression model (Model B) where serum LDL is a response of being alive at 5 years. The standard error for the variable “Alive at 5 Year’s” coefficient will be used to build a 95% confidence interval around the estimate of the true mean stated previously.**
		2. **The true mean LDL among those having survived at least 5 years is estimated as 118.70. Based on a 95% confidence interval, this observed estimate of 118.70 would not be judged unusual if the true mean LDL among those having died within 5 years were anywhere between 112.11 and 125.29 mg/dL. This confidence interval is different from the one found in problem 1. This difference is a result of the standard errors being calculated in different manners. In particular, it should be noted that the standard error in problem 1 considers only the sample relevant to itself (i.e. those alive at 5 years). However, the standard error for the coefficient is an estimate of the standard deviation across all cases.**
	6. 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?
		1. **Among those who survived at least 5 years, standard deviation 32.92. For those that died within 5 years, the standard deviation is 36.15. The regression does not change the estimated standard deviation.**
	7. How do models A and B relate to each other?
		1. **The two models are essentially the same. Each can calculate the other precisely given the corresponding value for the predictor, or indicator. For example, the two models output the same value when “Alive at 5 Years”=1 and “Dead at 5 Years”=0, and the same will hold when switching these two values.**
	8. Provide an interpretation of the intercept from the regression model A.
		1. **Model A will output only the intercept when “Dead at 5 Years” = 0. Hence, the intercept is the estimated mean for those having survived at least 5 years.**
	9. Provide an interpretation of the slope from the regression model A.
		1. **The slope parameter is the estimated difference in mean serum LDL between two populations differing in their survival status at 5 years. This estimates the association between serum LDL and survival at 5 years.**
	10. 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?
		1. **First, a 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 will be obtained using a linear regression of serum LDL (mg/dL) as indicated by survival at5 years. The standard error for the variable “Dead at 5 Year’s” coefficient will be used to build a 95% confidence interval around the estimate of the true mean difference stated previously. The regression uses a 2-sample t-test assuming equal variances to test the hypothesis that the two populations have the same mean LDL.**
		2. **The true difference in means between a population that survives at least 5 years and a population that dies within 5 years is estimated as 8.5. Based on a 95% confidence interval computed with an allowance for equal variances, this observed tendency of 8.50 mg/dL lower mean serum LDL among subjects dying earlier would not be judged unusual if the true difference population means were anywhere between a 1.91 mg/dL to 15.09 mg/dL lower mean LDL among subjects who die within 5 years. Using a t test that similarly allows for the possibility of equal variances, this observation is statistically significant at a 0.05 level of significance (two-sided P= 0.0115), and we can with high confidence reject the null hypothesis that the mean serum LDL levels are not different by vital status at 5 years in favor of a hypothesis that death within 5 years is associated with lower mean serum LDL. These p-value and confidence interval estimates are different from our problem 1 correspondents since where before we assumed unequal variance, we assume equal variance in the analysis reported here.**
2. 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.)
	1. **Using a t test that allows for the possibility of unequal variances across groups, the possible association between serum LDL and 5 year all-cause mortality will be evaluated. Mean LDL values across groups defined by vital status at 5 years will be compared.**
	2. **The point estimate for mean LDL value of subjects dying within 5 years is 118.70, while mean LDL for those surviving more than 5 years is estimated as 127.20. Hence, the difference between group means is estimated at 8.5. Based on a 95% confidence interval, it would not be unusual if the true difference in mean LDL values defined by vital status was found anywhere between 1.44 and 15.56 mg/dL. Furthermore, at the 5% significance level, with a p-value = 0.0185, we reject the null hypothesis for the lack of an association and instead find evidence for an association between serum LDL and 5 year all-cause mortality. These results are the same as the ones found in problem 1. The same assumptions were made, so this similarity is not surprising.**
3. 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.)
	1. **Using a linear regression model with unequal variances in this setting of a binary predictor resulting in a saturated model is equivalent to a 2 sample t test assuming unequal variances. Hence, this problem is a repetition of the previous problem. I will reiterate to be thorough. A t-test that allows for the possibility of unequal variances across groups will be used to test the possible association between serum LDL and 5 year all-cause mortality.**
	2. **The point estimate for mean LDL value of subjects dying within 5 years is 118.70, while mean LDL for those surviving more than 5 years is estimated as 127.20. Hence, the difference between group means is estimated at 8.5. If we would have done the regression, then this estimate of the mean difference would be the slope of the model. Based on a 95% confidence interval, it would not be unusual if the true difference in mean LDL values defined by vital status was found anywhere between 1.44 and 15.56 mg/dL. Furthermore, at the 5% significance level, with a p-value = 0.0185, we reject the null hypothesis for the lack of an association and instead find evidence for an association between serum LDL and 5 year all-cause mortality.**
4. 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.)
		1. **Male and female subjects are separated for their mean LDL and mean Age to be accounted separately. Furthermore, a combined group is created so that both male and female subjects can be pooled.**
		2. **Data is available on 735 subjects, although 10 subjects LDL data are missing (4 female, 6 male). These subjects are omitted from all analyses regarding mean LDL. All other variables were complete. The presents 369 females and 366 males. Females tend to have greater mean LDL than males, while age is similar between the two groups.**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Female** | **Male** | **Combined** |
| **Sample Size** | **369** | **366** | **735** |
| **LDL (mg/dL)** | **130.9 (34.3)** | **120.6 (32.2)** | **125.8 (33.6)** |
| **Age (years)** | **74.4 (5.3)** | **74.7 (5.6)** | **74.6 (5.5)** |
| **Table for Descriptive Statistics. Counts are given for Sample Size. Mean and standard deviation are given for LDL and Age. There are 10 LDL values missing: 4 for females and 6 for males.**  |

* 1. Provide a description of the statistical methods for the model you fit to address the question of an association between LDL and age.
		1. **A linear regression model is fit with serum LDL as a response to Age. Each variable is in its continuous form. The resulting regression will allow us to model the estimated effect an increase or decrease in age per year has on each unit (mg/dL) of LDL.**
		2. **Model:** $Serum LDL \left(\frac{mg}{dL}\right)=132.53-0.09\*Age$
			1. **The coefficient for Age presents with a standard error of 0.23 and p-value=0.69.**
	2. Is this a saturated model? Explain your answer.
		1. **This model is not saturated. For it to be saturated, there would not to be an equal number of parameters to groups of interest. This model has two parameters with only one group of interest, namely subjects with ages ranging from 65 to 99.**
	3. Based on your regression model, what is the estimated mean LDL level among a population of 70 year old subjects?
		1. **The estimated mean LDL level among a population of 70 year old subjects is given by: 132.52-0.09\*70 = 126.22 mg/dL.**
	4. 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?
		1. **The estimated mean LDL level among a population of 71 year old subjects is given by: 126.22-0.09 = 126.13 mg/dL. This estimate is easily calculated by using the answer from part (d) and subtracting to slope one more time, since 71 and 70 differ only by one.**
	5. 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?
		1. **The estimated mean LDL level among a population of 75 year old subjects is given by: 126.13-0.09\*4 = 125.77. This estimate is easily calculated by using the answer from part € and subtracting the slope four more times, since 75 and 71 differ by 4.**
	6. What is the interpretation of the “root mean squared error” in your regression model?
		1. **Our model has an RMSE of 33.62. The RMSE gives an impression of the magnitude of the model’s inability to fit the observations exactly. A model with lower RMSE is preferred. This value can be used for building confidence intervals as well.**
	7. What is the interpretation of the intercept? Does it have a relevant scientific interpretation?
		1. **The intercept is estimated mean LDL for newborns. However, it is scientifically irrelevant since our data did not fit any subjects with age less than 65 years. To estimate newborns would be severe extrapolation.**
	8. What is the interpretation of the slope?
		1. **The slope can only reasonably apply to subjects with age in our observed range in the predictor. Once that requirement is met, the slope can be interpreted as the decrease in mg/dL LDL per year increase of age. The model suggests that older subjects will tend towards lower mean LDL.**
	9. Provide full statistical inference about an association between serum LDL and age based on your regression model.
		1. **A linear regression model is fit with serum LDL as a response to Age. Each variable is in its continuous form. The resulting regression will allow us to model the estimated effect an increase or decrease in age per year has on each unit (mg/dL) of LDL.**
		2. **From a linear regression we estimate that when comparing two populations differing in age per years there is a 0.09 decrease of mg/dl in serum cholesterol per 1 year difference in age, with older subjects having lower cholesterol (95% CI 0.54 lower to 0.36 higher). These results are well within what might be attributed to a chance observation in the absence of a true association (P = 0.69).**
	10. 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?
		1. **Based on the above regression model, the best estimate for the difference in mean cholesterol between two groups of subjects who differ in age by 5 years is -0.45 mg/dl. Based on a 95% confidence interval, it would not be unusual if the true mean difference in mean cholesterol between two groups differing in age by 5 was found anywhere between -0.90 and 0.0008 mg/dL.**
	11. 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?
		1. **Pearson’s correlation coefficient will be used for a test of association with the paired samples of age and LDL. The test will be two-sided and evaluated at the 5% significance level.**
		2. **The sample estimate for the Pearson correlation coefficient is -0.0146. Based on a 95% confidence interval, it would not be unusual if the true correlation coefficient for age and LDL were found anywhere between -0.087 and 0.058. At the 5% significance level, the sample given does not support a non-zero correlation of age and LDL (p-value = 0.69).**
		3. **The regression-based conclusion about an association between LDL and age agrees with the inference about correlation. Both inferences have the same p-value, because they are essentially the same test.**

**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.