**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?

|  |  |
| --- | --- |
|  | **Vital Status** |
|  | **Alive within 5 years** | **Dead within 5 years** |
| **Number Total** | 614 | 121 |
| **Number Missing**  | 8 | 2 |
| **Mean(Standard Deviation) (mg/dL)** | 127.2(32.9) | 118.7(36.2) |

**Among the 614 subjects who were alive within 5 years (1826.25 days) of the study, 8 had missing serum LDL values, and so were removed from the analysis. Among the 121 subjects who were dead within 5 years of the study, 2 had missing values and were also removed from the analysis. Mean serum LDL values (mg/dL) were higher among those subjects who were alive within 5 years (mean of 127.2 mg/dL with standard deviation 32.9 mg/dL; compared to subjects dead within 5 years: mean of 118.7 mg/dL with standard deviation 36.2 mg/dL), with slightly less variation, as indicated by the lower standard deviation. Despite this difference, the magnitudes of the means and standard deviations are rather similar.**

* 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.

**The estimated mean serum LDL among subjects surviving 5 years was 127 mg/dL, with a standard error of 1.34 mg/dL. Based on a 95% confidence interval using the student’s t-test, mean serum LDL among subjects surviving within 5 years would not be judged unusual if the true population mean was anywhere between 124.5 and 129.8 mg/dL.**

**The estimated mean serum LDL among subjects dead within 5 years was 119 mg/dL, with a standard error of 3.31mg/dL. Based on a 95% confidence interval using the student’s t-test, mean serum LDL among subjects surviving within 5 years would not be judged unusual if the true population mean was anywhere between 112.1 and 125.3 mg/dL.**

**These estimates are not so different in magnitude. The standard errors for subjects dead within 5 years was about 2.5 times larger than that of subjects who survived 5 years into the study, this difference is likely due to the larger sample size of survivors compared to those who do did not survive within 5 years, as standard error is inversely proportional to sample size, and there were about 5 times more subjects who were alive, compared to dead, within 5 years.**

* 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 CI’s for both populations overlap, though not by very much. Since the differences do not overlap, we cannot confidently state that mean LDL levels differ between populations. However this result does not yield conclusive evidence that mean serum LDL level is *the same* in either group, since a two sided t-test may still result in a difference in mean serum LDL that is significantly greater than 0. Hence, we are unable to make any definitive conclusions about the estimated means.**

* 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?)

**Under the assumption of equal variance, we would use pooled standard error, i.e.:** $\hat{SD}= \sqrt{\frac{(N\_{A}-1)S\_{A}^{2}+ (N\_{D}-1)S\_{D}^{2}}{N\_{A}+N\_{D}-2}}$ **Where SA and SD are the standard deviations of both populations: alive within 5 years, and dead within 5 years; NA and ND are the number of subjects alive and dead within 5 years, respectively. Here the estimated standard deviation in each group is 33.48 mg/dL.**

* 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?

**Using a student’s t-test assuming equal variance, the estimated difference in mean serum LDL was 8.50 mg/dL less in the group that died within 5 years, with a standard error of 3.36 mg/dL. Based on a 95% confidence interval, 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 in population means were anywhere between a 1.91 mg/dL to 15.1 mg/dL lower mean LDL among subjects who die within 5 years and the variances were identical in the two groups. Hence at the 0.05 significance level (two-sided P = 0.01154) we can reject the hypothesis that mean serum LDL is equivalent between the two populations.**

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.

**For both models, we have two groups (alive in 5 years, and dead in 5 years) and two parameters (slope and intercept measurements), hence both models are saturated.**

* 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?

**Using Model A (using as our predictor an indicator that the subject died within 5 years) the estimate of true mean LDL among a population of subjects who survive at least 5 years is 127.2 mg/dL (that is, it is the estimated intercept for model A). This estimate is identical to that of 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 survive at least 5 years? How does this compare to the corresponding estimate from problem 1? Explain the source of any differences.

**Using Model A, I obtain a 95% confidence interval of 124.5 to 129.9 mg/dL for true mean LDL among a population of subjects who survive at least 5 years. This confidence interval is nearly identical to the one obtained in problem 1. The reason these confidence intervals are not exactly the same is because the regression-estimated confidence interval takes into account the error from the entire sample population (both subjects alive and dead within 5 years) by using the pooled sample standard deviation and a corresponding t-test statistic reflecting the entire sample size minus 1, while problem 1 uses only the sample standard deviation for those subjects who survive at least 5 years, and a corresponding t-test statistic reflecting only the sample size of subjects surviving at least 5 years minus 1.**

* 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?

**Using Model B (using as our predictor an indicator that the subject died within 5 years) the estimate of true mean LDL among a population of subjects who survive at least 5 years is 127.2 mg/dL (that is, it is the estimated intercept for model B). This estimate is identical to that of 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.

**Using Model B, I obtain a 95% confidence interval of 112.7 to 124.7 mg/dL for true mean LDL among a population of subjects who died within 5 years. This confidence interval is close to the one obtained in problem 1. The reason these confidence intervals are not exactly the same is because the regression-estimated confidence interval takes into account the error from the entire sample population (both subjects alive and dead within 5 years) by using the pooled sample standard deviation and a corresponding t-test statistic reflecting the entire sample size minus 1, while problem 1 uses only the sample standard deviation for those subjects who were dead within 5 years, and a corresponding t-test statistic reflecting only the sample size of subjects dead within 5 years minus 1. This difference in confidence intervals was more pronounced than that which was compared in parts 2c and 1b, because the sample sizes used to estimate these two confidence intervals were much more different.**

* 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?

**Assuming equal variance, the regression based estimate of standard deviation for both models (A and B) is the root mean squared error, or 33.48. This is identical to the estimated standard deviation from problem 1.**

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

**Models A and B yield equivalent conclusions about the relationship between serum LDL and vital status, in that from either model we can derive equivalent estimates of mean serum LDL for either group (alive or dead 5 years post enrollment), as well as the estimated difference in mean serum LDL between these two groups. Both models yield equivalent p-values for the test of a slope estimate that is significantly different from 0 (indicative of a difference in mean serum LDL between groups defined by vital status). They are different in that the intercept of Model A gives an estimation of mean serum LDL for subjects surviving after 5 years, and the slope plus the intercept gives an estimation of mean serum LDL for subjects who had died within 5 years; while the intercept of Model B gives an estimation of mean serum LDL for subjects who had died within 5 years, and the slope plus the intercept gives an estimation of mean serum LDL for subjects who had survived 5 years.**

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

**The intercept from regression model A is the estimated value of serum LDL for those subjects who were alive after 5 years of study enrollment (predictor of interest = 0).**

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

**The slope from regression model A is the estimated value of difference in mean serum LDL when comparing subjects who were dead as opposed to alive after 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?

**We conducted a linear regression model assuming equal variance of serum LDL regressed on the predictor, vital status within 5 years, using an indicator variable of whether or not the subject died within 5 years. We obtain an estimated difference in mean serum LDL of 8.50 mg/dL lower in subjects dead within 5 years, with standard error 3.36 mg/dL, based on the Huber-White sandwich estimator. Based on a 95% confidence interval, 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 1.9 mg/dL to 15.1 mg/dL lower mean LDL among subjects who die within 5 years and the variances were identical in the two groups. Using a t test that presumes equal variances, this observation is statistically significant at a 0.05 level of significance (two-sided P= 0.0115), and we can thus conclude with high confidence that the distribution of serum LDL differs between those who do or do not have higher risk of death over a 5 year period. This estimate, confidence interval, and P value are identical to that obtained in the analysis of problem 1, which uses a t-test assuming equal variance.**

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.)

**Using a student’s t-test assuming unequal variance, the estimated difference in mean serum LDL was 8.50 mg/dL less in the group that died within 5 years, with a standard error of about 3.57 mg/dL. Based on a 95% confidence interval, 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 in population means were anywhere between a 1.4 mg/dL to 15.6 mg/dL lower mean LDL among subjects who die within 5 years and the variances were identical in the two groups. Hence at the 0.05 significance level (two-sided P = 0.01858) we can reject the hypothesis that mean serum LDL is equivalent between the two populations. These results are a bit more conservative than those in problem 1 (indicated by the larger confidence interval and greater P value).**

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.)

**We conducted a linear regression model assuming unequal variance of serum LDL regressed on the predictor, vital status within 5 years, using an indicator variable of whether or not the subject died within 5 years. We obtain an estimated difference in mean serum LDL of 8.50 mg/dL lower in subjects dead within 5 years, with standard error 3.56 mg/dL, based on the Huber-White sandwich estimator. Based on a 95% confidence interval, 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 1.5 mg/dL to 15.5 mg/dL lower mean LDL among subjects who die within 5 years and the variances were unequal in the two groups. Using a t test that presumes unequal variances, this observation is statistically significant at a 0.05 level of significance (two-sided P= 0.01723), and we can thus conclude with high confidence that the distribution of serum LDL differs between those who do or do not have higher risk of death over a 5 year period. The results of this analysis are close to that obtained in problem 3, however the standard error computed using the Huber-White sandwich estimator in this case, is not as conservative as the t-test under the assumption of unequal variance (indicated by the smaller confidence interval and the lesser P value).**

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.)

**Methods: A variable was created for age category, grouping individuals into ten year age categories (60-69, 70-79, 80-89, and 90-99 years of age). Descriptive statistics are presented within groups defined by sex, and for the entire sample. Within each group defined by sex we present age percentages.**

**Results: Data is available on 735 subjects, however 10 of those subjects (including 2 who died within 5 years) are removed from the descriptive statistics as they are missing data on serum low density lipoprotein (LDL). It should be remembered that we can not assess the impact that such omissions might have on the generalizability of our results. None of the 725 subjects were missing data on any other variables of interest for this analysis. Of the 725 subjects with available measurements, 117 were between the ages of 61-70, 492 were between the ages of 71-80, 116 were between the ages of 81-90, and 10 were between the ages 91-99. The following table presents descriptive statistics regarding sex within these groups. Subjects between the ages 61-70 were slightly less likely to be male than in any other age category, yet this tendency was very small. All distributions of sex within age categories are similar and close to 50%, with the exception of the single age category of 91-99 years, where we observe a much higher probability of being male (80% of subjects are male). This striking difference is most likely due to the small number of subjects within this age category, so we conclude that there is no overall confounding occurring due to sex.**

|  |  |
| --- | --- |
|  | **Age Category** |
|  | **61-70 years****(n=114)** | **71-80 years****(n=487)** | **81-90 years****(n=114)** | **91-99 years****(n=10)** | **All Ages****(n=10)** |
| **Male (%)** | 47.4% | 49.7% | 49.1% | 80.0% | 49.7% |

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

**Methods: In order to address the association between serum LDL and age, we use a linear regression model with a continuous age variable, in order to allow each age to have distinct average serum LDL estimation. We use robust Huber-White estimation in order to account for the potential unequal variances of LDL across the range of age values.**

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

**This model is not a saturated model, since we are estimating 2 parameters across all possible values of age. Hence, since there are greater than 2 age values, we do not have a saturated model.**

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

**Using the estimated linear regression intercept and slope parameters of 132.52810 mg/dL and -0.09019 mg/dL, I obtain an estimated mean LDL among a population of 70 year old subjects of 126.2 mg/dL.**

* 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?

**Using the estimated linear regression intercept and slope parameters of 132.52810 mg/dL and -0.09019 mg/dL, I obtain an estimated mean LDL among a population of 71 year old subjects of 126.1 mg/dL. The difference between this estimate and the estimate of mean LDL among a population of 70 year old subjects is exactly the slope estimated in this regression model; this makes sense because the linear regression slope represents the change in mean serum LDL per unit increase in the linear predictor (in this case age).**

* 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?

**Using the estimated linear regression intercept and slope parameters of 132.52810 mg/dL and -0.09019mg/dL, I obtain an estimated mean LDL among a population of 71 year old subjects of 125.8 mg/dL. The difference between this estimate and the estimate of mean LDL among a population of 70 year old subjects is exactly 5 times the slope estimated in this regression model; this again makes sense because the difference between 70 and 75 is a five unit increase, so with a linear predictor we expect the difference to be 5 times the slope, given the increment in predictor value of 5 units.**

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

**In our regression model, we allow the possibility of heteroscadicity, and so the root mean squared error is the square root of average variance across all age values. This provides a very rough idea of the magnitude of overall variance, though not with respect to any particular age range.**

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

**The intercept provides an estimation of mean serum LDL among newborns, age 0, and is far outside of the range of our data. For this reason, the intercept should not be used to extrapolate an estimate of mean serum LDL for this age. We conclude that it has no scientific interpretation.**

* 1. What is the interpretation of the slope?

**Because our model was created using age as a linear predictor, with no higher order terms, the slope is an estimate of difference in mean serum LDL between groups differing by 1 year in age.**

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

**Results: We conducted a linear regression model assuming unequal variance of serum LDL regressed on the continuous predictor, age. We estimate that for each year increase in age, the in mean serum LDL decreases by 0.09 mg/dL, with standard error 0.23 mg/dL, based on the Huber-White sandwich estimator. A 95% confidence interval suggests that this observation is not unusual if the true difference in mean serum LDL per year difference in age were between -0.55 and 0.37 mg/dL. Using a t test that presumes unequal variances, this observation is not statistically significant at a 0.05 level of significance (two-sided P= 0.698), and we thus fail to reject the hypothesis that serum LDL changes with age.**

* 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?

**Methods: In order to analyze groups that differ by 5 years in age, we create age categories of years (65-69, 70-74, 75-79,… ,95-99 years). We then conducted a linear regression model assuming unequal variance of serum LDL regressed on this 5 year age response, which we treat as a continuous variable. By doing this, the slope of this model is an estimate of difference in mean serum LDL between groups differing by 5 years in age.**

**Results: We estimate that for each 5 year increase in age, the in mean serum LDL decreases by 0.61 mg/dL, with standard error 1.12 mg/dL, based on the Huber-White sandwich estimator. A 95% confidence interval suggests that this observation is not unusual if the true difference in mean serum LDL per 5 year difference in age were between -** **2.81 and 1.58 mg/dL. Using a t test that presumes unequal variances, this observation is not statistically significant at a 0.05 level of significance (two-sided P= 0.5828), and we thus fail to reject the hypothesis that serum LDL changes with age.**

* 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?

**Methods: We conducted a test of correlation between serum LDL (mg/dL) and age in years, using Pearson’s product moment correlation test.**

**Results: We observe an estimated Pearson’s r= -0.01, indicating a very slight negative correlation between age and serum LDL. A 95% confidence interval suggests that this observation is not unusual if the true correlation were between -0.09 and 0.06. This observation is not statistically significant at a 0.05 level of significance (two-sided P= 0.6944), and we thus fail to reject the hypothesis that serum LDL is correlated with age.**

**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.