**Biost 518: Applied Biostatistics II**

**Biost 515: Biostatistics II**

Emerson, Winter 2014

**Homework #2**

January 13, 2014

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?

**606 patients survived at least 5 years, with a sample mean LDL of 127 mg/dL and sample standard deviation of LDL values of 32.9 mg/dL. 119 patients survived at least 5 years, with a sample mean LDL of 119 mg/dL and sample standard deviation of LDL values of 36.2 mg/dL. The sample standard deviation for the mean LDL among subjects who survived at least 5 years is more than 3 mg/dL greater than the standard deviation of the mean LDL among nonsurvivors, which represents a large discrepancy. The sample mean LDL of the 5 year survivors is greater than 8 mg/dL than the mean LDL of the 5 year nonsurvivors, which is also a large discrepancy between mean LDL values**.

* 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 point estimate for the true mean LDL for patients who survived at least 5 years post-enrollment is 127.2 mg/dL, with a standard error of 1.134 mg/dL and a 95% confidence interval of (124.6, 129.8) mg/dL. The point estimate for the true mean LDL for patients who survived at least 5 years post-enrollment is 118.7 mg/dL, with a standard error of 3.31 mg/dL and a 95% confidence interval of (112.1, 125.2) mg/dL.**

**The standard error of the point estimate of the true mean LDL in the patients who died within 5 years is over twice as large as the standard error of the point estimate of the true mean LDL 5 year survival group, which is due in part to the sample size in the nonsurvivors being much smaller than the 5-year survival group (119 nonsurvivors and 606 survivors at 5 years post-enrollment).**

**The best estimate of the true mean LDL for a population of similar subjects who would survive at least 5 years is the sample mean, so the answers for the sample mean and point estimates are the same. The standard deviation is greater than the standard error by over an order of magnitude because the standard deviation refers to the variation of LDL levels for an individual in a population of similar individuals, where the standard error is a measure of the error in estimating the true mean given the sample data – the sample standard error is equal to the 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% CI for the mean LDL in a population surviving 5 years (95% CI: [124.6, 129.8]) overlaps with the CI for mean LDL in a population dying within 5 years (95% CI: [112.1, 125.3]) However, the point estimate for the mean LDL of the 5 year nonsurvivors is not contained within the 95% CI for the mean LDL in a population surviving 5 years; similarly, the point estimate for the mean LDL of the 5 year survivors is not contained within the 95% CI for the mean LDL in a population that died within 5 years. Thus, the estimated difference in the estimated mean LDL across 5 year survivorship status at a 0.05 level of significance is significantly different from zero.**

* 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?)

**Our best estimate for the standard deviation of LDL measurements in each group would be to pool the variances together weighting by the degrees of freedom.** = 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?

**The point estimate for the true difference in mean LDL levels between the populations divided by 5-year survivorship is 8.5 mg/dL greater in the 5 year survivor group, with an associated standard error of 3.36 mg/dL. The 95% confidence interval for the true difference in mean LDL levels is (1.91, 15.09) mg/dL. The two-sided p-value for testing the null hypothesis that the two populations have the same mean LDL is 0.012, therefore we conclude that the observed difference is not unlike what would be typically observed if the true mean difference in LDL levels was anywhere between 1.9 mg/dL and 15.09 mg/dL greater in the group of patients who survived at least 5 years relative to the nonsurvivor group.**

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.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ldl | **Coef.** | **Std. Err.** | **T statistic** | **p-value** | **[95% Conf.** | **Interval]** |
| **deadin5** | -8.5 | 3.4 | -2.5 | 0.0120 | -15.1 | -1.9 |
| **intercept** | 127.2 | 1.4 | 93.5 | 0.0000 | 124.5 | 129.9 |

*Table 1: Model A – linear regression of LDL (in mg/dL) on the indicator of 5 year death (deadin5)*

**The regression model is saturated because the number of groups (5 year nonsurvivors and survivors) is equal to the number of parameters (coefficient for indicator of vital status at 5 years and the intercept term).**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ldl | Coef. | Std. Err. | T statistic | P-value | [95% Conf. | Interval] |
| alivein5 | 8.5 | 3.4 | 2.5 | 0.0120 | 1.9 | 15.1 |
| intercept | 118.7 | 3.1 | 38.7 | 0.0000 | 112.7 | 124.7 |

Table 2: Model B – linear regression of LDL (in mg/dL) on the indicator of 5 year survivorship (alivein5)

**The regression model is saturated because the number of groups (5 year survivors and nonsurvivors) is equal to the number of parameters (coefficient for indicator of vital status at 5 years and the intercept term).**

* 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 the regression parameter estimates from model A (which uses the indicator of death at 5 years as the predictor), the estimate of the true mean LDL among a population of subjects who survive at least 5 years is 127.2 mg/dL (corresponding to the intercept coefficient in model A). This is the same estimate for the true mean LDL for the population of subjects who survive at least 5 years as question 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 the intercept parameter from model A, the 95% confidence interval for the true mean LDL among a population of subjects who survive at least 5 years is (124.5, 129.8) mg/dL, which differs slightly from the corresponding confidence interval estimate in question 1 (95% CI: [112.1, 125.3] mg/dL). The difference is due from the t-test using the estimate of the standard deviation of LDL levels only from the population of subjects who survived at least 5 years, where the regression parameter estimate uses the pooled standard deviation from both groups when estimating the bounds of the 95% confidence interval.**

* 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 the intercept parameter from model B (which uses the indicator of survival at 5 years as the predictor), the estimate of the true mean LDL among a population of subjects who die within 5 years is 118.7 mg/dL, which is equivalent to the corresponding estimate from 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 the intercept parameter from model B, the 95% confidence interval for the true mean LDL among a population of subjects who die within 5 years is [112.7, 124.7] mg/dL, which differs slightly from the estimate in question 1 (95% CI: [112.1, 125.3] mg/dL). The difference is due from the t-test using the estimate of the standard deviation of LDL levels only from the population of subjects who survived at least 5 years, where the regression parameter estimate uses the pooled standard deviation from both groups when estimating the bounds of the 95% confidence interval. The standard deviation in LDL among subjects who survive at least 5 years is higher than the non-survivor group; hence the 95% confidence interval for the regression model (which uses the pooled standard deviation estimate) is wider than the corresponding t-test confidence interval.**

* 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?

**In the regression model assuming equal variances, the estimate of the standard deviation within each group is the root mean square error (33.48 mg/dL), which is the same for both models. This estimate is equal to the corresponding estimate from problem 1.**

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

**Models A and B are reparametrizations of the same model – model A uses the indicator of 5 year survival as the predictor and model B uses the indicator of 5 year death as the predictor**

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

**The intercept (127.2 mg/dL) corresponds to the estimated mean LDL among a population of subjects who survived after 5 years.**

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

**The slope parameter (-8.5 mg/dL) corresponds to the average change in mean LDL in a population that dies before 5 years, relative to a population that dies within 5 years. According to the model, the mean LDL in the population that dies within 5 years is 8.5 mg/dL lower than a population of individuals that survives at least 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?

**Using regression model A, the point estimate for the true difference in mean LDL between a population that survives at least 5 years and a population that dies within least 5 years is 8.5 mg/dL lower, with the group dying within 5 years tending towards lower mean LDL. The p-value testing the hypothesis that the two populations have the same mean LDL is 0.012, so at a significance threshold of 0.05 we reject the null hypothesis that the true difference in mean LDL levels is zero, and that the observed difference in mean LDL levels is what might be typically observed if the true difference in mean LDL is anywhere between 15.1 mg/dL lower and 1.9 mg/dL lower in the population that died within 5 years compared to the population that survived at least 5 years.**

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.)

**Table 3 displays the result for the two sample t test comparing mean LDL across populations separated by 5 year vital status assuming equal variances, and table 4 displays the results for the corresponding two-sample t test assuming unequal variances.**

**The two approaches return the same point estimate for the estimated mean difference in LDL across the two groups, but they differ in their estimation of standard error and p-values.**

 **The t-test assuming equal variances uses the pooled standard error (the square root of the weighted sum of the sample variances, weighting by the degrees of freedom), while the t-test assuming unequal variances estimates the standard error as the square root of the sum of the sample variances divided by their respective sample sizes.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **5 year vital status** | **Obs** | **Mean** | **Std. Err.** | **Std. Dev.** | **[95% Conf.** | **Interval]** |
| **Survivor** | 606 | 127.20 | 1.34 | 32.93 | 124.57 | 129.83 |
| **Non-survivor** | 119 | 118.70 | 3.31 | 36.16 | 112.13 | 125.26 |
| **diff** |   | 8.50 | 3.36 |   | 1.91 | 15.09 |
| **two-sided p-value** | 0.0115 |  |  |  |  |  |

*Table 3: Two-sample t test comparing mean LDL (mg/dL) of populations separated by vital status at 5 years post study enrollment, assuming equal variances*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **5 year vital status** | **Obs** | **Mean** | **Std. Err.** | **Std. Dev.** | **[95% Conf.** | **Interval]** |
| **Survivor** | 606 | 127.20 | 1.34 | 32.93 | 124.57 | 129.83 |
| **Non-survivor** | 119 | 118.70 | 3.31 | 36.16 | 112.13 | 125.26 |
| **difference** |   | 8.50 | 3.57 |   | 1.44 | 15.56 |
| **two-sided p-value** | 0.0186 |  |  |  |  |  |

*Table 4: Two-sample t test comparing mean LDL (mg/dL) of populations separated by vital status at 5 years post study enrollment, assuming unequal variances*

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.)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ldl** | **Coef.** | **Std. Err.** | **t** | **P>t** | **[95% Conf.** | **Interval]** |
| **deadin5** | -8.50 | 3.57 | -2.38 | 0.02 | -15.50 | -1.50 |
| **intercept** | 127.20 | 1.34 | 95.04 | 0.00 | 124.57 | 129.83 |

*Table 5: Linear regression of mean LDL (mg/dL) on the indicator of 5 year all-cause mortality allowing for unequal variances*

**The results from the linear regression of LDL on the indicator of 5 year all-cause mortality allowing for unequal variances across groups returns the same point estimate of the mean difference in LDL across groups, but the 95% confidence interval for the mean difference in LDL is narrower and the p-value is larger using the robust standard error estimates. In the dataset, the smaller group of patients that died within 5 years has a higher variance – the t-test assuming equal variances is anti-conservative; the p-values are too small and the reported CI is too narrow.**

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.)



*Figure 1: Two-way scatterplot of serum LDL by age separated by gender. Least squares best fit lines for both genders are superimposed.*

**The least squares fit lines over the two-way scatterplot of serum LDL across age demonstrate an effect of gender on LDL, with females tending to have higher average serum LDL concentrations compared to males of similar age. The slight unequal slopes is suggestive of potential effect modification of gender on the relationship of serum LDL by age, though the tendency may be weak due to the lack of data points at higher values of age (>85 years). For sex to be a confounder on the association between age and LDL, it must be associated with the predictor of interest in the sample and causally related with the response. The mean ages for the female and male population in the sample are relatively close, with mean ages of 74.4 years and 74.7 years respectively. Since we do not have outside evidence suggesting that sex is causally related to serum LDL and the distribution of ages across sex are approximately equal, it does not suggest that sex is a confounder for the association between age and LDL.**

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

**A linear regression model was fit to assess the linear effect of age on serum LDL concentration allowing for unequal variances of LDL by age.**

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

**The model is not saturated as the number of groups defined by continuous variable age is infinite while the number of model parameters is fixed at a slope and intercept term.**

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

**Based on the regression model, the estimated mean LDL level is equal to 132.5 – 0.09 \* age. For a population of 70 year old subjects the estimated mean LDL level is 126.23 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?

**Based on the regression model, the estimated mean LDL level is equal to 132.5 – 0.09 \* age; for a population of 71 year old subjects, the estimated mean LDL level is 126.14 mg/dL. The estimated mean LDL level for the population of 71 year old subjects differs from the estimated mean LDL level for the population of 70 year old subjects by one slope coefficient – or 0.09 mg/dL lower estimated mean LDL level.**

* 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?

**Based on the regression model, the estimated mean LDL level is equal to 132.5 – 0.09 \* age; for a population of 71 year old subjects, the estimated mean LDL level is 125.78 mg/dL. The estimated mean LDL level for the population of 75 year old subjects differs from the estimated mean LDL level for the population of 70 year old subjects by five times the slope coefficient – or 5\*0.09 = 0.45 mg/dL lower estimated mean LDL level than a population of 70 year old subjects.**

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

**The root mean square error is an estimate of the standard deviation of the residuals. In our regression model, we assumed the groups defined by age have unequal variances, so the root mean square error is based on a weighted average of the square root of the within group variances.**

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

**The interpretation of the intercept is the average mean LDL level for a population of 0 year olds, which does not have a relevant scientific interpretation.**

* 1. What is the interpretation of the slope?

**The slope is an estimate of the average change in mean LDL levels across groups 1 year apart. The regression estimate for the slope is -0.09, so the estimated average mean LDL levels is 0.09 mg/dL lower in a population of individuals for each year of age difference.**

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

**The linear regression of serum LDL level on age yielded an average difference of 0.09 mg/dL reduction in mean serum LDL concentration for every year of age difference. The two-sided p-value evaluating the association between age and LDL is nonsignificant at the 0.05 significance level (P=0.694). The 95% confidence interval asserts the observed difference in mean serum LDL for each year of age is typical for what might be observed if the true difference in mean serum LDL is anywhere between 0.54 mg/dL lower and 0.36 mg/dL higher per year of age difference.**

* 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?

**Given that the estimated difference in mean LDL across groups that differ by one year of age is Normal with mean -0.09 and standard error 0.229, the estimated difference in mean LDL across groups that differ by 5 years of age is Normal(5\*-0.09, 5\*0.229) = Normal(-0.45, 1.15). I would report that the estimated difference in mean LDL across groups that differ by 5 years of age is 0.45 mg/dL lower per 5 year difference, with a 95% confidence interval of [-2.7, 1.8] mg/dL.**

* 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 test for nonzero correlation between LDL and age returns an estimated correlation of -0.0146, which is not significantly different from a correlation of zero. The conclusion of nonzero correlation between LDL and age is similar to the regression-based conclusion about a nonsignificant slight negative correlation between age and LDL.**

**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.