**Biost 518: Applied Biostatistics II**

**Biost 515: Biostatistics II**

Emerson, Winter 2014

**Homework #2**

January 21, 2014

**Question 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 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?

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Sample size** | **Sample mean** | **Sample standard deviation** |
| **Survived at least 5 years** | 606 | 127.2 | 32.9 |
| **Died within 5 years** | 119 | 118.7 | 36.2 |

The sample means not similar in magnitude and sample standard deviations are 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.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Point estimate** | **Estimated standard error** | **95% confidence interval** |
| **Survived at least 5 years** | 127.2 | 1.34 | (124.6, 129.9) |
| **Died within 5 years** | 118.7 | 3.31 | (112.1, 125.3) |

The point estimates are not similar in magnitude and the estimated standard errors are not similar in magnitude. Estimated standard error is calculated by estimated standard deviation divided by square root of sample size. Although two groups have similar sample standard deviations, the sample size for the group with subjects who survive at least 5 years is much higher than the group with subjects who died in 5 years. Thus, the estimated standard error for the group with subjects who survive at least 5 years is much lower than the group with subjects who died in 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?

Yes, the 95%CI for mean LDL in a population surviving 5 years overlap with the 95% CI for mean in a population dying within 5 years. This may be tempting to conclude that there is no significant difference in LDL between a population surviving 5 years and a population dying within 5 years at a 0.05 level of significance. But this conclusion might not be correct.

* 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?)

The pooled sample standard deviation $s\_{p}=\sqrt{\frac{\left(n\_{1}-1\right)s\_{1}^{2}+\left(n\_{2}-1\right)s\_{2}^{2}}{n\_{1}+n\_{2}-2}}$

The best estimate for the standard deviation of LDL measurements in each group is 33.60197.

* 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?

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Point estimate** | **Estimated standard error** | **95% confidence interval** |
| **Difference in means**  | 8.5 | 3.36 | (1.9, 15.1) |

The estimated difference in mean LDL for a population that survives at least 5 years is 8.5 mg/dl statistically higher than a population that dies within 5 years (two sided P = 0.0115). 95% confidence interval shows that the estimated difference is not unusual if the true population mean difference is between 1.9mg/dl and 15.1 mg/dl. Therefore, we reject the null hypothesis of no association between serum LDL and 5 year all cause mortality in favor of a trend that the group who survived at least 5 years has higher mean LDL than the group who died within 5 years.

**Question 2**

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.

Saturated model: number of groups equal to number of parameters

Model A is a saturated model, because the predictor variable used in the analysis only had two values (0, 1) and the regression model has two parameters ($β\_{0}$, $β\_{1}$).

Model B is a saturated model, because the predictor variable used in the analysis only had two values (0, 1) and the regression model has two parameters ($β\_{0}$, $β\_{1}$).

* 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 survive at least 5 years? How does this compare to the corresponding estimate from problem 1?

Model A

The estimated mean LDL for the subjects survived at least 5 years was 127.2 mg/dl. This estimated mean and the sample mean in problem 1 are the same.

* 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.

Model A

The estimated 95% confidence interval for the true mean LDL among a population of subjects who survive at least 5 years was between 124.5 mg/dl and 129.9 mg/dl. This estimation is slightly different from the corresponding estimate in problem 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?

Model B

The estimated mean LDL for the subjects who die within 5 years was 118.7 mg/dl. This estimated mean and the sample mean in problem 1 are the same.

* 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.

Model B

The estimated 95% confidence interval for the true mean LDL among a population of subjects who die within 5 years was between 112.7 mg/dl and 124.7 mg/dl. This estimation is slightly different from the corresponding estimate in problem1.

* 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?

|  |  |
| --- | --- |
| **Model A** | **Estimated standard deviation** |
| Survived at least 5 years | 33.477 |
| Died within 5 years | 33.477 |
| **Model B** | **Estimated standard deviation** |
| Survived at least 5 years | 33.477 |
| Died within 5 years | 33.477 |

The estimate based on regression of the standard deviation within each group is similar to standard deviation of LDL measurements in each group (33.60197) in problem 1.

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

Model B is the model we reparameterize Model A. The inference from Model A and Model B are the same.

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

$β\_{0}$: 127.2

The estimated mean LDL for the group who subjects survives at least 5 years.

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

$β\_{1}$: -8.5

The estimated difference in mean LDL for the group who subjects died within 5 years is 8.5 mg/dl lower than for the group who subjects who 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?

|  |  |  |  |
| --- | --- | --- | --- |
| **Model A** | **Point estimate** | **Estimated standard error** | **95% confidence interval** |
| $$β\_{0}$$ | 127.2 | 1.36 | (124.5, 129.9) |
| $$β\_{1}$$ | -8.5 | 3.36 | (-15.1, -1.9) |
| **Model B** | **Point estimate** | **Estimated standard error** | **95% confidence interval** |
| $$β\_{0}$$ | 118.7 | 3.07 | (112.7, 124.7) |
| $$β\_{1}$$ | 8.5 | 3.36 | (1.9, 15.1) |

The mean LDL for the group who survives at least 5 years is 8.5 mg/dl higher than the group who died within 5 years (two sided P = 0.0115). The 95% confidence interval suggests the estimated difference is not unusual if the true population mean difference is between 1.9 mg/dl and 15.1 mg/dl. Therefore, we reject the null hypothesis of no association between serum LDL and 5 year all cause mortality in favor of a trend that the group who survived at least 5 years has higher mean LDL than the group who died in 5 years.

The inference for the linear regression model presumes homoscedasticity and the t test presumes equal variance are nearly the same. The estimated mean difference is the same (8.500541). The standard errors (3.356652) and 95% confidence interval (1.910591, 15.09049) are the same.

**Question 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 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.)

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Sample size** | **Sample mean** | **Sample standard deviation** |
| **Survived at least 5 years** | 606 | 127.2 | 32.9 |
| **Died within 5 years** | 119 | 118.7 | 36.2 |

Same as problem 1

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Point estimate** | **Estimated standard error** | **95% confidence interval** |
| **Difference in means**  | 8.5 | 3.57 | (1.4, 15.6) |

P=0.0186

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Point estimate** | **Estimated standard error** | **95% confidence interval** |
| **Survived at least 5 years** | 127.2 | 1.34 | (124.6, 129.9) |
| **Died within 5 years** | 118.7 | 3.31 | (112.1, 125.3) |

Same as problem 1

The estimated difference in mean LDL for a population that survives at least 5 years is 8.5 mg/dl higher than a population that dies within 5 years (two sided P = 0.0186). The 95% confidence interval suggests the estimated difference is not unusual if the true population mean difference is between 1.4mg/dl and 15.6 mg/dl. Therefore, we reject the null hypothesis of no association between serum LDL and 5 year all cause mortality in favor of a trend that the group who survived at least 5 years has higher mean LDL than the group who died within 5 years.

The sample means, sample standard deviations are exactly the same. The estimated standard errors of the difference in means are slightly different and 95% confidence interval (1.4, 15.6) for t test with unequal variance is wider than (1.9, 15.1) for t test with equal variance.

**Question 4**

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.)

|  |  |  |  |
| --- | --- | --- | --- |
| **Model A** | **Point estimate** | **Estimated standard error** | **95% confidence interval** |
| $$β\_{0}$$ | 127.2 | 1.34 | (124.6, 129.8) |
| $$β\_{1}$$ | -8.5 | 3.57 | (-15.5, -1.5) |
| **Model B** | **Point estimate** | **Estimated standard error** | **95% confidence interval** |
| $$β\_{0}$$ | 118.7 | 3.31 | (112.2, 125.2) |
| $$β\_{1}$$ | 8.5 | 3.57 | (1.5, 15.5) |

The estimated difference in mean LDL for a population that survives at least 5 years is 8.5 mg/dl higher than a population that dies within 5 years (two sided P = 0.0174). The 95% confidence interval suggests the estimated difference is not unusual if the true population mean difference is between 1.5 mg/dl and 15.5 mg/dl. Therefore, we reject the null hypothesis of no association between serum LDL and 5 year all cause mortality in favor of a trend that the group who survived at least 5 years has higher mean LDL than the group who died within 5 years.

The inference for the linear regression model presumes unequal variances (heteroscedasticity) and the t test presumes unequal variance are nearly the same. The estimated mean difference is the same as 8.5 mg/dl (two sided P = 0.0174) and the standard errors (3.574252 vs. 3.565821) and 95% confidence interval ((1.44132, 15.55976) vs. (1.499941, 15.50114)) are slightly different.

**Question 5**

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.)



|  |
| --- |
| Male: LDL ~ age n = 360 P value = 0.3208 Root MSE = 32.144 |
|  | **Point estimate** | **Estimated SE** | **95% CI** |
| $$β\_{0}$$ | 144.1 | 23.6 | (97.7, 190.6) |
| $$β\_{1}$$ | -0.31 | 0.32 | (-0.94, 0.31) |
| Female: LDL ~ age n = 365 P value = 0.5097 Root MSE = 34.279 |
|  | **Point estimate** | **Estimated SE** | **95% CI** |
| $$β\_{0}$$ | 114.2 | 25.4 | (64.3, 164.1) |
| $$β\_{1}$$ | 0.22 | 0.34 | (-.44, 0.89) |

The slopes for male (-0.31) and for female (0.22) are different. It indicates the association between LDL and age might be modified by sex.

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

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Age** | **n** | **Mean** | **SD** | **Min** | **25%** | **Median** | **75%** | **Max** |
| **65-69** | 114 | 127.7 | 32.40 | 51 | 104 | 130.5 | 150 | 217 |
| **70-74** | 303 | 125.3 | 32.50 | 37 | 102 | 126 | 146 | 247 |
| **75-79** | 184 | 126.9 | 35.46 | 11 | 102 | 125 | 150.5 | 225 |
| **80-84** | 80 | 122.8 | 33.49 | 52 | 99 | 119.5 | 145 | 227 |
| **85-89** | 34 | 125.0 | 39.14 | 68 | 97 | 123.5 | 142 | 216 |
| **90-94** | 8 | 124.8 | 35.77 | 57 | 105 | 136.5 | 141.5 | 175 |
| **95-99** | 2 | 132.0 | 1.41 | 131 | 131 | 132 | 133 | 133 |
| **Total** | 725 | 125.8 | 33.60 | 11 | 102 | 125 | 147 | 247 |

725 subjects were grouped into 5 years age interval. The standard deviations among groups seem to be similar. Therefore, I implemented the ordinary least squares regression that presumes homoscedasticity to analyze the association between serum LDL and age.

|  |
| --- |
| LDL ~ age n = 725 P value = 0.6944 Root MSE = 33.622 |
|  | **Point estimate** | **Estimated SE** | **95% CI** |
| $$β\_{0}$$ | 132.5 | 17.34 | (98.9, 166.2) |
| $$β\_{1}$$ | -0.09 | 0.23 | (-0.54, 0.36) |

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

This is not a saturated model. There are 35 age groups but only two parameters (slope and intercept).

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

The estimated mean LDL level among a population of 70 year old subjects is $132.5-0.09×70=126.2$.

* 1. Based on your regression model, what is the estimated mean LDL level among a population of 71 years 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 years old subjects is $132.5-0.09×71=126.11$.

The difference between your answer to this problem and your answer to part c is the same as the slope in magnitude.

* 1. Based on your regression model, what is the estimated mean LDL level among a population of 75 years 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 years old subjects is $132.5-0.09×75=125.75$.

The difference between your answer to this problem and your answer to part c is five times of the slope in magnitude.

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

The root mean squared error is what we use to estimate the standard deviation within group. Root MSE = 33.6 mg/dl.

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

The mean LDL in 0 year old population is estimated to be 132.5 mg/dl. No, it doesn’t have any relevant scientific interpretation.

* 1. What is the interpretation of the slope?

For each year increase in age between two groups, the older group tends to have lower average LDL by 0.09 mg/dl.

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



The ordinary least squares regression that presumes homoscedasticity was implemented to analyze the association between serum LDL and age.

From the linear regression analysis, the estimated mean LDL is -0.09 mg/dl (on average) which is not statistically significant different from 0 (two sided P = 0.6944). 95% confidence interval contains 0 and suggests the observed result would not be unusual if the true difference in mean LDL is between -0.54mg/dl and 0.36 mg/dl. Therefore, we can’t reject the null hypothesis of no association between serum LDL and age based on the regression model.

* 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?

$$-0.09×5=-0.45$$

For each 5 years increase in age between two groups, the older group tends to have lower average LDL by 0.45 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?

$r=-0.0146$ $P=0.6944$

$$R^{2}=0.0002$$

The correlation coefficient between LDL and age is -0.0146, which is square root of R squared reported in the simple regression of age on LDL. The correlation of -0.0146 is not statistical significantly different from 0 (P=0.6944). From the regression analysis, the estimated mean LDL is 0.09 mg/dl and is not statistically significantly different among groups (two sided P=0.6944).

The regression-based conclusion about an association between LDL and age and the inference about correlation are the same. We do not have evidence to conclude that there is an association between LDL and age.