**2788**

**Biost 515 (Winter 2014)**

**Instructor: Scott Emerson**

**Homework 2**

*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?

**Answer:** The following tablecompares the sample size, sample mean, and sample standard deviation (SD) of serum low density lipoprotein (LDL) values between subjects who survived at least five years and those who died within five years. As shown, the sample means and the sample standard deviations between groups are similar in magnitude. Specifically, the mean LDL among subjects who survived at least 5 years is 1.07 times greater than those who died within 5 years, while the standard deviation of LDL among subjects who died within five years is 1.1 times greater than those who survived at least five years.

|  |
| --- |
| Table 1: Serum Low Density Lipoprotein (LDL) |
|  | n | Mean (mg/dL) | SD (mg/dL) |
| Alive at 5 Years | 606 | 127.2 | 32.9 |
| Death w/in 5 Years | 119 | 118.7 | 36.2 |

* 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.

**Answer:** Table 2compares the point estimate, the estimated standard error of that point estimate (SE), and the 95% confidence interval (CI) for the true mean serum low density lipoprotein (LDL) between those who would survive at least five years and those who would die within five years in a population of subjects similar to the sample. Since the sample mean is the point estimate for a population of similar subjects, we can once again state that we expect a population of similar subjects who would survive at least five years to have a mean LDL approximately 1.07 times greater than the mean LDL in a population whose similar subjects would die within five years. However, the standard errors (unlike the standard deviations) are not as similar in magnitude. That is, the estimated standard error of the point estimates is approximately 2.5 times greater for subjects who would die within five years than subjects who would survive at least five years. This is to be expected, since standard error is positively associated with sample standard deviation and negatively correlated with sample size. As shown in Table 1, the sample of subjects who survived at least five years (n=606) is approximately 5 times larger than those who died within five years (n=119).

|  |
| --- |
| Table 2: Serum Low Density Lipoprotein (LDL) |
|  | Estimate (mg/dL) | SE (mg/dL) | 95% CI (mg/dL) |
| Alive at 5 Years | 127.2 | 1.3 | 124.6 – 129.8 |
| Death w/in 5 Years | 118.7 | 3.3 | 112.1 – 125.3 |

* 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?

**Answer:** As shown in Table 2, the 95% confidence interval (CI) for the mean LDL in a population surviving at least five years (124.6 – 129.8 mg/dL) overlaps by 0.7 mg/dL with the 95% CI for mean LDL in a population dying within five years (112.2 – 125.3 mg/dL). Given this information alone, it would be erroneous to draw any conclusions on the statistical significance of the estimated difference in means between populations.

In general, if we are only given two estimates and their corresponding CIs for analysis, statistical significance can only be determined if the CI do not overlap (i.e. we could conclude the difference in means is statistically significant) or if the CI for one stratum contains the point estimate of the other stratum (i.e. we could conclude the difference in means is not statistically significant). However, if neither of those two situations occur (as is the case for this exercise), we cannot reach any conclusions about the statistical significance of our estimated difference in 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?)

**Answer:** Given that the variances in two populations are equal, but that the means might be different, the best estimate for the standard deviation of LDL measurements in each group is the square root of their pooled variance (otherwise known as the root mean squared error). Observe: if given two populations with variances σ1 = σ2, then the estimated variances, $s\_{1}^{2}$and$ s\_{2}^{2}$, for each population are both estimating the same variable. Therefore, if the data is pooled together, we can create a more precise estimate. The formula for pooled variance is as follows:

$s\_{pooled}^{2}= \frac{\left(n\_{1}-1\right)\left(s\_{1}^{2}\right)+ \left(n\_{2}-1\right)\left(s\_{2}^{2}\right) }{n\_{1}+ n\_{2}-2}$ ≈ $ σ\_{1}= σ\_{2}$

Hence, in for this specific dataset, the best estimate for the standard deviation of LDL measurements in each group is $\sqrt{s\_{pooled}^{2}}$ = 33.4 mg/dL:

$\sqrt{s\_{pooled}^{2}} = \sqrt{\frac{\left(606-1\right)\left(32.9\right)+ \left(119-1\right)\left(36.2\right) }{606 + 119 - 2}}= $ 33.4

* 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?

**Answer:** The true difference in mean LDL between a population that survives at least five years and a population that dies within five years is estimated to be 8.5 mg/dL (95% CI: 1.9 – 15.1 mg/dL) with a standard error of 3.4 mg/dL. This estimated difference is statistically significant at a 0.05 level of significance (two-sided P = 0.0115), and we can say with high confidence that the distribution of serum LDL differs between those who do or do not have a higher risk of death over a five year period.

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.

**Answer:** In the following two models, $Ε\left(X\right)$ represents mean serum LDL and X = 0, 1 is an indicator of the subject’s vitality status at five years. Since the models each have two groups (i.e. those who were dead within five years and those who survived at least five years) and two parameter estimates (i.e. $β\_{0}$ and $β\_{1}$), both A and B are considered saturated models.

$Model A: Ε\left(X\right)= β\_{0}$ + $β\_{1} ∙ X=127.2 -8.5∙X$

$Model B: Ε\left(X\right)= β\_{0}$ + $β\_{1} ∙ X=118.7+8.5∙X$

* 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?

**Answer:** Using model A, for which the predictor variable (X) is an indicator that the subject died within five years and the response variable ($ Ε\left(X\right) $) is mean serum LDL, the estimate of the true mean LDL among a population of subjects who survive at least five years is 127.2 mg/dL. Therefore, the estimate derived using ordinary least squares regression is identical to the estimate derived in problem 1. Of course, this makes sense since the t test that presumes equal variances across groups and classical linear regression are mathematically equivalent.

* 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.

**Answer:** Again using model A, for which the predictor variable (X) is an indicator that the subject died within five years and the response variable ($ Ε\left(X\right) $) is mean serum LDL, a 95% confidence interval (CI) for the true mean LDL among a population of subjects who survive at least five years is (124.5 – 129.9 mg/dL). This is similar to, but not exactly the same as the result obtained in problem 1. When comparing CIs derived from classical regression verses those derived from a t test which assumes equal variance, the intervals will differ slightly because regression uses the pooled SD and the t test uses the sample SD to calculate CI.

* 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?

**Answer:** Again using model A, for which the predictor variable (X) is an indicator that the subject died within five years and the response variable ($ Ε\left(X\right) $) is mean serum LDL, the estimate of the true mean LDL among a population of subjects who die within five years is 118.7 mg/dL. Therefore, the estimate derived using classical linear regression is identical to the estimate derived in problem 1. (To understand the validity of this result, see question 2b.)

* 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.

**Answer:** This time using model B, for which the predictor variable (X) is an indicator that the subject survived at least five years and the response variable ($ Ε\left(X\right) $) is mean serum LDL, a 95% confidence interval for the true mean LDL among a population of subjects who die within five years is (112.7 – 124.7 mg/dL). This is a similar, but not identical result as was obtained in problem 1. As was stated in part (c), the differences in the CI can be attributed to regression utilizing the pooled SD and the t test utilizing the sample SD in CI calculations.

* 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?

**Answer:** Presuming the variances are equal in the two populations, we can utilize the root mean squared error (MSE) to estimate the within group standard deviation (SD) for each model. Hence, the SD for model A is 33.5 mg/dL and the SD for model B is also 33.5. In problem 1(a), the SD among subjects who survived at least five years was 32.9 mg/dL and the SD among subjects who died within five years was 36.2 mg/dL. In problem 1(f), the root MSE calculated from these SDs was 33.4 mg/dL. Therefore, the regression based SD lies between the two SD estimates and is approximately equal to the calculated root MSE (any difference is due to rounding).

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

**Answer:** Model A is a reparameterized version of model B. If the indicator, X, equals zero in model A, then equivalent results will be achieved with X equal to 1 in model B. In other words, the intercept $(β\_{0})$ in model A is equal to the intercept plus slope $(β\_{0}+β\_{1}) $in model B.

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

**Answer:** The intercept ($β\_{0}$) from regression model A represents the mean serum LDL (mg/dL) for a population of subj ects who would survive at least five years.

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

**Answer:** The slope ($β\_{1}$) from regression model A represents the difference in mean serum LDL (mg/dL) between populations of subjects who would and would not survive at least five 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?

**Answer:** Using the regression parameter estimates, the true difference in mean LDL between a population that survives at least five years and a population that dies within five years is estimated to be 8.5 mg/dL (95% CI: 1.9 – 15.1 mg/dL) with a standard error of 3.4 mg/dL. This estimated difference is statistically significant at a 0.05 level of significance (two-sided P = 0.0120), and we can say with high confidence that the distribution of serum LDL differs between those who do or do not have higher risk of death over a five year period. (Note: I was able to draw the exact same inferences using linear regression as I was using the t test.)

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.)

**Answer:** Using a t test that allows for the possibility of unequal variances across groups, the true difference in mean LDL between a population that survives at least five years and a population that dies within five years is estimated to be 8.5 mg/dL (95% CI: 1.4 – 15.6 mg/dL) with a standard error of 3.6 mg/dL. This estimated difference is statistically significant at a 0.05 level of significance (two-sided P = 0.0186), and we can say with high confidence that the distribution of serum LDL differs between those who do or do not have higher risk of death over a five year period.

When compared to problem 1 for which we assumed equal variances across groups, our inferences from both results remains equivalent. However, the confidence intervals lost some precision and the standard error and p value (while still remaining significant) also slightly increased when the test allowed for heteroscedasticity. This observation aligns with the rule that the greatest precision will be obtained if using the t test which justifiably assumes homoscedasticity.

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.)

**Answer:** Using a linear regression model that allows for the possibility of unequal variances across groups, the true difference in mean LDL between a population that survives at least five years and a population that dies within five years is estimated to be 8.5 mg/dL (95% CI: 1.5 – 15.5 mg/dL) with a standard error of 3.6 mg/dL. This estimated difference is statistically significant at a 0.05 level of significance (two-sided P = 0.0170), and we can say with high confidence that the distribution of serum LDL differs between those who do or do not have higher risk of death over a five year period.

When we compare this analysis to problem 3 for which we used a t test which allowed for unequal variances across groups, our inferences from both results remains equivalent. However, the confidence intervals in this problem have higher precision and the standard error and p value also slightly decreased.

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.)

**Answer:** Table 3 (next page) presents descriptive statistics for serum low density lipoprotein (LDL) (mg/dL) within groups defined by sex and age (years) of subjects and for the entire sample. Age was dichotomized into three intervals: 65-75 years, 76-85 years, and 86-100 years. Each group includes the number of subjects, mean, standard deviation (SD), median, minimum and maximum.

Data is available on 735 subjects, however 10 of those subjects are missing data on serum LDL and have therefore been omitted from all analyses. Females tended to have higher serum LDL levels than males within each age interval. Overall, the mean serum LDL was 130.9 mg/dL in females and 120.6 mg/dL in males. In female subjects, mean serum LDL was highest for those between the ages of 76 and 86 (132.7 mg/dL) and lowest for those between the ages of 65 and 75 (130.1 mg/dL). Among male subjects, mean serum LDL was negatively associated with increasing age intervals: mean serum LDL was highest in 65 to 75 year old males (121.4 mg/dL) and lowest in 86 to 100 year old males (117.2 mg/dL). If further analyses which take sex into consideration were to be carried out, this data suggests that sex may be an effect modifier or confounder.

|  |
| --- |
| Table 3: Serum Low Density Lipoprotein (LDL) (mg/dL) dichotomized by Sex and Age (yrs) |
|  | n | Mean (SD) | Median | Min - Max |
| Female |  |  |  |  |
|  65 - 75 | 239 | 130.1 (34.1) | 130 | 46 – 247 |
|  76 - 85 | 111 | 132.7 (34.0) | 133 | 11 – 225 |
|  86 - 100 | 15 | 131.7 (41.1) | 141 | 68 – 216 |
|  All Ages | 365 | 130.9 (34.3) | 131 | 11 – 247 |
| Male |  |  |  |  |
|  65 - 75 | 230 | 121.4 (32.3) | 121.5 | 37 – 206 |
|  76 - 85 | 110 | 119.5 (31.2) | 113.5 | 52 – 227 |
|  86 - 100 | 20 | 117.2 (36.4) | 111.5 | 57 – 216 |
|  All Ages | 360 | 120.6 (32.1) | 117 | 37 – 227 |
| All Subjects |  |  |  |  |
|  65 - 75 | 469 | 125.8 (33.5) | 126 | 37 – 247 |
|  76 - 85 | 221 | 126.1 (33.2) | 124 | 11 – 227 |
|  86 - 100 | 35 | 123.4 (38.6) | 128 | 57 – 216 |
|  All Ages | 725 | 125.8 (33.6) | 125 | 11 - 247 |

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

**Answer:** To address the question of an association between serum LDL and age, a model was fit to the dataset with linear regression allowing for heteroscedasticity (that is, the model utilized robust standard error estimates). With age (years) as a continuous variable and the predictor of interest, and serum LDL (mg/dL) as the outcome of interest, the model determined statistical significance at the alpha = 0.05 level. The model generated in this analysis is as follows:

$Model: Ε\left(X\right)= β\_{0}$ + $β\_{1} ∙ X=132.5 -0.9∙X, $65 ≤ X ≤ 100

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

**Answer:** The model described in part (b) is not a saturated model. There are two parameters in the model (i.e. the slope and intercept), but since we treat age continuously, the predictor variable has an infinite number of values.

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

**Answer:** Based on my regression model, the estimated mean LDL level among a population of 70 year old subjects is 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 d relate to the slope?

**Answer:** Based on my regression model, the estimated mean LDL level among a population of 71 year old subjects is 126.1 mg/dL. This implies the mean difference in serum LDL between a 70 and 71 year old is 0.1 mg/dL. Then, the slope in our model ($β\_{1})$ represents the difference in mean serum LDL across populations differing in age by 1 year. In other words, for every 1 year increase in age, mean serum LDL will decrease by 0.1 mg/dL.

* 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 d relate to the slope?

**Answer:** Based on my regression model, the estimated mean LDL level among a population of 75 year old subjects is 125.8 mg/dL. The difference in mean serum LDL between a 70 and 75 year old is 0.5 mg/dL. This difference relates to the slope
($β\_{1}= -0.1)$ in that a five year increase in age decreases mean serum LDL by (0.1)\*(5) = 0.5 mg/dL.

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

**Answer:** In my regression model, the “root mean squared error” can be interpreted as the average within group standard deviation.

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

**Answer:** In my regression model, the intercept ($β\_{0})$ represents the estimated serum LDL for a population of newborn babies (subjects who are 0 years old). Although this is not scientifically relevant since our dataset included subjects between 65 and 100 years old, it is also not an entirely senseless estimation of mean serum LDL.

* 1. What is the interpretation of the slope?

**Answer:** In my regression model, the intercept ($β\_{1})$ represents the difference in estimated serum LDL (mg/dL) across groups differing in age by 1 year. In other words, since
$β\_{1}$= -0.1, for every 1 year increase in age, mean serum LDL will decrease by 0.1 mg/dL.

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

**Answer:** From a linear regression analysis with robust standard errors, we estimate that mean serum low density lipoprotein (LDL) differs by 0.1 mg/dL (on average) among adults 65 years of age and older and whose age differs by 1 year, with the older populations tending toward lower mean serum LDL. Based on a 95% confidence interval, this 0.1 mg/dL difference in mean serum LDL between subjects with one year age gap would not be judged unusual if the true difference in subjects one year apart were anywhere between -0.5 mg/dL to 0.4 mg/dL Unfortunately, this observation is not statistically significant at a 0.05 level of significance (p = 0.698), therefore, we fail to reject the null hypothesis that a 1 year difference in age is not associated with serum LDL levels.

* 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?

**Answer:** From a linear regression analysis with robust standard errors, we estimate that for every 5 year difference in age, mean serum low density lipoprotein (LDL) differs by 0.5 mg/dL (on average) among adults 65 years of age and older, with the older populations tending toward lower mean serum LDL. Based on a 95% confidence interval, this 0.5 mg/dL difference in mean serum LDL between subjects with a five year age gap would not be judged unusual if the true difference in subjects five years apart were anywhere between -2.7 mg/dL to 1.8 mg/dL Unfortunately, this observation is not statistically significant at a 0.05 level of significance (p = 0.698), therefore, we fail to reject the null hypothesis that a 5 year difference in age is not associated with serum LDL levels.

* 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?

**Answer:** Since the test for significant correlation is exactly the test for slope in classical simple linear regression, I performed simple linear regression (i.e. assumed homoscedasticity) and calculated R2 to equal 0.0002 with a corresponding p value of 0.6944. This implies the correlation, R, between age and LDL is $\sqrt{0.0002}$ = 0.0141, and there is little to no correlation between age and serum LDL. These results agree with the regression-based conclusions about an association between LDL and age, which were given in part (j) and failed to reject the hypothesis that no association exists.