**Homework #2**

January 13, 2014

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?

Table 1. Mean LDL and standard deviation for groups who did or did not survive 5 years.

|  |  |  |
| --- | --- | --- |
|  | Survived 5 years | Died within 5 years |
| Sample size | 593 | 131 |
| Mean LDL (mg/dL) | 127.3 | 118.6 |
| Standard Deviation of LDL (mg/dL) | 32.9 | 36.0 |

Among 724 patients with data available, 593 (81.9%) were still alive at five years, while 131 (18.1%) had died within five years. The general magnitude of both the sample means and sample standard deviations are similar between groups, but a two-sample t-test of the means, assuming equal variances, suggests the mean LDL is significantly higher in patients who survived five years compared to those who died within five years (p=0.0113).

* 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.

Table 2. Estimated mean LDL for a population of patients who do or do not survive 5 years.

|  |  |  |
| --- | --- | --- |
| **Mean LDL (mg/dL)** | Survived 5 years(n=593) | Died within 5 years(n=131) |
| Point estimate | 127.4 | 118.6 |
| Standard error | 1.35 | 3.14 |
| 95% confidence interval | 124.7, 130.0 | 112.4, 124.8 |

The point estimate of the population mean LDL is the sample mean LDL. The magnitudes of the point estimates are similar between groups. The standard error is more than 2-fold larger in the group that died within five years than in those alive. The sample standard deviation is approximately 9.4% larger in the group who died within five years. The standard error is dependent upon the sample standard deviation and the sample size. Since the sample size is much smaller in the group that died within five years, the standard error is larger compared to the standard error for the group that survived.

* 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 five years (124.7, 130.0) overlaps slightly with the 95% CI for the population that died within five years (112.4, 124.8). Even though the confidence intervals overlap, there can still be a statistical difference if the null hypothesis is not contained within the 95% CI, so this observation is not useful in evaluating the 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?)

If the variances are equal, then the sample standard deviations must also be equal. To combine the two estimated sample standard deviations it is most appropriate to take a weighted average that accounts for the difference is sample size between groups. For example, the following calculations, or a similar approach, could be used:

$$\frac{s\_{alive}×n\_{alive}}{N\_{total}}+\frac{s\_{dead}×n\_{dead}}{N\_{total}}$$

where *s* is the sample standard deviation, *n* is the group sample size, and *N* is the total sample size.

* 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 scientific question is whether there is a difference in means between the two groups. The point estimate for the difference in mean LDL between populations surviving five years or dying within five years is 8.76 mg/dL, with standard error 3.42, and the 95% CI of the difference is 2.00 to 15.5 mg/dL. Since the CI of the difference does not contain zero, the difference is statistically significant (p=0.0113) with higher mean LDL among individuals who survive five years compared to those who do not. Therefore, there sufficient evidence to reject the null hypothesis that the two populations have the same mean LDL, suggesting an association between serum LDL and 5-year all-cause mortality.

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 the predictor variable is binary, with values of “dead” or “alive” which suggests the models are saturated. By definition, a saturated model has equal numbers of groups and parameters, and these models have two parameters (LDL and vital status), information for determining the mean is not borrowed across groups, and each group mean can be fit exactly.

* 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, the estimate of the true mean LDL among a population who survive five years determined by the intercept of the linear regression analysis is 127.2 mg/dL, which is nearly the same as the mean value determined in problem 1 (127.4 mg/dL).

* 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, the 95% confidence interval for the true mean LDL among a population who survive five years determined by the confidence interval around the intercept of the linear regression analysis is 124.5 to 129.8 mg/dL, which is nearly the same as the mean value determined in problem 1 (95% CI 124.7, 130.0 mg/dL). One source of the very slight difference may be that the t-test in problem 1 was performed by not assuming equal variances, while the regression model in this problem presumes homoscedasticity. Also, in regression there are n-2 degrees of freedom, while in the t-test there are n-1 degrees of freedom.

* 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, the estimate of the true mean LDL among a population who died within five years determined by the intercept of the linear regression analysis is 118.7 mg/dL, which is nearly the same as the mean value determined in problem 1 (118.6 mg/dL).

* 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, the 95% confidence interval for the true mean LDL among a population who died within five years determined by the confidence interval around the intercept of the linear regression analysis is 112.7 to 124.7 mg/dL, which is nearly the same as the mean value determined in problem 1 (95% CI 112.4, 124.8 mg/dL). One source of the very slight difference may be that the t-test in problem 1 was performed by not assuming equal variances, while the regression model in this problem presumes homoscedasticity. Also, in regression there are n-2 degrees of freedom, while in the t-test there are n-1 degrees of freedom.

* 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?

From model A, the regression-based estimate of the standard deviation for the group who died within five years is 35.15, compared to 36.0 in problem 1. From model B, the estimated standard deviation for the group who was alive at five years is 33.12, compared to 32.9 in problem 1. The estimates are slightly different likely due to differences in calculating the standard error. In regression standard error is calculated using n-2 degrees of freedom, and the sum of the squared residual values. For the t-test there are n-1 degrees of freedom. Also, regression uses the pooled standard deviation, which may be different than the standard deviation determined using a t-test.

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

Models A and B are related in that the predictor variables (dead or alive at 5 years) are two groups from the overall study sample, and the response variable is the LDL measured for each individual in the sample.

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

Since this is a model with a binary predictor variable, the intercept for model A is an estimate of the mean LDL value in a population of patients who were alive at five years.

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

The slope for model A is an estimate for the difference in means between the group surviving five years and the group who died within five years. The sum of the intercept and slope in model A is an estimate of the mean LDL value in a population of patients who died within 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?

The point estimate for the true difference in mean LDL between a population that survives five years and a population that dies within five years is 8.45 mg/dL, with a higher mean in the group who survived. The estimated standard error is 3.36, and the 95% confidence interval for the true difference in means is 1.86 to 15.05 mg/dL. The p-value is statistically significant at 0.0120, which suggests the difference in mean LDL is more extreme than would be expected under the null hypothesis that association between serum LDL and 5-year all-cause mortality is the same for both groups. Therefore, there is sufficient evidence to reject the null hypothesis. This conclusion is the same as that drawn in problem 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 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. Comparison of results between analyses assuming or not assuming equal variances.

|  |  |  |
| --- | --- | --- |
|  | **Survived 5 years** | **Died within 5 years** |
|  | Unequal variances | Equal variances | Unequal variances | Equal variances |
| **Sample size** | 593 | 131 |
| **Mean LDL (mg/dL)** | 127.3 | 127.3 | 118.6 | 118.6 |
| **Standard Deviation of LDL (mg/dL)** | 32.9 | 32.9 | 36.0 | 36.0 |
| **Standard error** | 1.35 | 1.35 | 3.14 | 3.14 |
| **95% confidence interval** | 124.7, 130.0 | 124.7, 130.0 | 112.4, 124.8 | 112.4, 124.8 |

The results of the analysis assuming equal variance is the same as the analysis that does not assume equal variances in this case. Since the sample size is large, and the variance is not extremely large, the difference in analyses is not expected to be particularly remarkable. When equal variances are assumed, however, the two-sided p-value for the difference in mean LDL is smaller at 0.0072, compared to p=0.0117 when equal variances are not assumed.

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.)

A regression analysis that allows for unequal variances across groups gives the same values for the variables presented in problem 3. The p-value for the association in this case is p=0.0110, which is slightly different than in problem 3, but still statistically significant.

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.)

Table 4. Descriptive statistics for LDL by age, stratified by sex.

|  |  |  |
| --- | --- | --- |
|  |  | **Serum LDL Measurements (mg/dL)** |
| **Age (years)** | **N** | **Mean**  | **Std. Dev.** | **Median** | **Minimum** | **Maximum** |
| **Female** | 364 | 130.9 | 34.3 | 131 | 11 | 247 |
| **65-69** | 60 | 127.0 | 34.0 | 128.5 | 51 | 217 |
| **70-74** | 152 | 130.7 | 32.8 | 130 | 46 | 247 |
| **75-79** | 92 | 133.5 | 37.5 | 134 | 11 | 225 |
| **80-84** | 42 | 130.2 | 30.4 | 127 | 81 | 201 |
| **85-89** | 16 | 131.9 | 43.2 | 132 | 68 | 216 |
| **90-100** | 2 | 141.2 | 0.71 | 141.5 | 141 | 142 |
| **Male** | 360 | 120.6 | 32.1 | 117 | 37 | 227 |
| **65-69** | 54 | 128.5 | 30.8 | 133.5 | 68 | 206 |
| **70-74** | 150 | 119.6 | 31.3 | 120 | 37 | 188 |
| **75-79** | 92 | 120.2 | 32.1 | 115 | 39 | 218 |
| **80-84** | 38 | 114.6 | 35.3 | 112 | 52 | 227 |
| **85-89** | 18 | 118.8 | 35.2 | 108.5 | 72 | 216 |
| **90-100** | 8 | 122.4 | 34.8 | 132 | 57 | 175 |

Among 364 females and 360 males, the mean LDL values are similar for each group according to sex. Overall, females have higher mean LDL in all age groups compared to males. As shown in Figure 1 below, there appears to be a tendency for LDL values among men to decline slightly with increasing age, while the LDL values remain roughly constant with age among females.



 Figure 1. Scatterplot of LDL values for individuals according to age and stratified 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.

Method: A linear regression model was fit to the data to compare the mean LDL values of subjects according to age. The predictor variable is age, and the response variable is the LDL value, and both are continuous variables.

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

This model is not saturated. Since the predictor variable is continuous, and the response variable is also continuous, the number of groups and parameters are not equal. In this case we are borrowing information across ages to determine the mean, so the mean for a specific age cannot necessarily be fit exactly.

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

By linear regression, the fitted value for estimating the mean LDL for a subject according to age is defined by the following equation:

$$E\left[LDL\_{i}|Age\_{i}\right]=132.5-0.0902×Age\_{i}$$

To estimate the mean LDL for a population of 70-year-old subjects, the age (70 yr) is substituted into the above equation, giving an estimated mean LDL 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?

To estimate the mean LDL for a population of 71-year-old subjects, the age (71 yr) is substituted into the above equation, giving an estimated mean LDL of 126.1 mg/dL. Since the model is not saturated, the mean values for a given age cannot be calculated exactly, so the estimation is based on the linear equation above. The expected difference in LDL for an interval is based on the slope such that for every 1-year interval in age, the LDL declines by 0.0902 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 c relate to the slope?

To estimate the mean LDL for a population of 75-year-old subjects, the age (75 yr) is substituted into the above equation, giving an estimated mean LDL of 125.7 mg/dL. Since the model is not saturated, the mean values for a given age cannot be calculated exactly, so the estimation is based on the linear equation above. The expected difference in LDL for an interval is based on the slope such that for a 5-year increase in age, the LDL declines by the factor (0.0902 mg/dL/year) x (5 years).

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

In this linear regression model, the root mean squared error is the estimated standard deviation in predicted mean LDL at each age. Stated differently, it is the average within-group variance for LDL according to age.

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

The intercept is the predicted mean value of LDL when age = 0, thus for a newborn. This is not scientifically relevant mainly because the youngest age in our sample is 65 years old, so the LDL of a newborn is irrelevant for this scientific question.

* 1. What is the interpretation of the slope?

The slope is the estimated amount that the mean LDL changes with each one-year increase in age.

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

On average, the mean LDL levels are estimated to be 0.0902 mg/dL lower for every year increase in age with a 95% confidence interval (CI) from 0.547 mg/dL lower to 0.367 mg/dL higher per year. The association described by this linear regression analysis suggests LDL levels do not vary with age, and the results are within the range that may be reasonably expected to be true under the null hypothesis that LDL level is independent of age (p = 0.6984).

* 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?

The difference in mean LDL across groups that differ by 5 years of age is described by the slope of the linear regression line defined above as a 0.0902 mg/dL decrease for each 1-year increase in age. Therefore the estimated 5-year difference in mean LDL is determined by the following calculation: (0.0902 mg/dL/year)(5 years) = 0.451 mg/dL. Since the 95% confidence interval for each 1-year age increase is (-0.547, 0.367 mg/dL), the 95% confidence interval will be considerably wider and can be estimated by multiplying each endpoint of the interval by 5 years, giving a 95% confidence interval for the estimated 5-year difference in mean LDL from 2.735 mg/dL lower, to 1.835 mg/dL higher than the mean difference in LDL of 0.451 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?

Method: To test for a nonzero correlation between LDL and age, a simple correlation test was performed to identify the estimated strength and direction of a linear relationship between the two variables.

Inference: The correlation coefficient for the relationship between LDL and age is r = -0.0146, with a p-value of 0.6944. This suggests that while there appears to be a negative relationship between LDL and age, such that LDL decreases as age increases, the correlation is not statistically significant. Therefore, there is insufficient evidence to reject the null hypothesis that correlation is zero. This general result is similar to that identified by linear regression such that there does not appear to be an association between LDL and age.