* ***Methods:*** Descriptive statistics of LDL treated as a continuous variable, which included sample size, sample mean, and sample standard deviation, presented within groups defined by death within 5 years, survival for 5 years post study entry, and for the entire sample.
* ***Results:*** The total sample size is 735 subjects, however, 10 of the subjects have missing data on LDL values. (Two of them are in the group with death within 5 years and 8 in the group with survival after 5 years.) Those subjects are omitted from the analysis. The details of descriptive statistics are provided in the following table. The group with death within 5 years has smaller sample size (119 compared to 606 in the group with survival after 5 years), smaller sample mean (118.7 vs 127.2) and larger standard deviation (36.16 vs 32.93).

|  |  |  |  |
| --- | --- | --- | --- |
|  | Death within 5 years | Survival for 5 years | All subjects |
|  | Number (missing data) | Mean (SD) | Number (missing data) | Mean (SD) | Number (missing data) | Mean (SD) |
| Serum LDL (mg/dl) | 119 (2) | 118.7 (36.16) | 606(8) | 127.2 (32.93) | 725 (10) | 125.8 (33.60) |

* ***Methods:*** We used the sample distribution of LDL values to estimate a population parameter, stratifying by the observation time below and above 5 years. The sample mean is considered as the best point estimate for true value of the population mean. Then point estimate, critical value (α=0.05), and standard error (by using Student’s t-distribution) are used to construct 95% confidence interval, which represents a range of possible values of true values of the population mean.
* ***Results***: The details of point estimate, the estimated standard error of that point estimate, and the 95% confidence interval for two groups stratifying whether the subjects survived for at least 5 years are provided in the following table. The group with death within 5 years has smaller sample mean (118.7 vs 127.2) and larger standard error (2.21 vs 1.34).

The point estimates are the same as the sample mean. However, the standard error depends on both the standard deviation and the sample size. The standard error falls as the sample size increases and the variation of sample distribution of the mean decreases. (SE = SD/√(sample size)). Therefore, it is much smaller compared to standard deviation.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Sample size | Mean (SD) | Standard error | 95% Confidence interval |
| Survival for 5 years | 606 | 127.2 (32.93) | 1.34 | 124.6 | 129.8 |
| Death within 5 years | 119 | 118.7 (36.16) | 2.21 | 112.1 | 125.3 |

* 1. The 95% confidence intervals for two groups overlap. If confidence intervals on individual parameters do not overlap, we know a statistically significant difference exists. However, if the confidence intervals do overlap, then the conclusions are unclear, especially the 95% confidence interval for one group doesn’t contain the point estimate of the other stratum.
	2. The pooled standard deviation, SP, is calculated from the above equation. Therefore, the best estimate is 33.48.
* ***Methods:*** The mean of LDL values were compared across two groups who died within 5 years of study enrollment and those who survived at least 5 years. Differences in the mean were tested using a t test with equal variances assumption. 95% confidence intervals for the difference in population means were based on that same treatment of variances.
* ***Results:*** The point estimate of difference in means of LDL values between a population that survives at least 5 years and a population that dies with 5 years is 8.50 mg/dL with a higher mean LDL levels among the group with at least 5-year-survial. The estimated standard error is 3.36. Based on the 95% confidence interval constructed from the equal variance assumption, this observed 8.50 mg/dL higher mean LDL among the group with least 5-year-survial would not be judged unusual if the true difference in the population means were anywhere between a 1.91 to 15.09 mg/dL. Using a t test that presumes equal variances, this observation is statistically significant at a 0.05 level of significance (two-sided P= 0.0115). We can conclude that the distribution of LDL values differs between those two groups.
	1. Both models are saturated because the predictor variable used in both the analyses only has two values and the regression model has two parameters. That means no extra information is borrowed across the groups for calculating the mean of LDL.
	2. Using model A in which the subject died within 5 years is used as predictor, the estimate of the true mean LDL among a population of subjects who survive at least 5 years would be the intercept, of which the value is 127.20. The estimate is the same to the corresponding point estimate from problem 1.
	3. Using model A, 95% confidence interval for the true mean LDL among a population of subjects who survive at least 5 years is the 95% confidence interval for the intercept, which is between 124.5 and 129.9. It is slightly different from the previous estimate (124,6, 129.8) because the CI for the intercept is calculated from the pooled standard deviation instead of individual standard deviation of each group. The latter one is used in t test.
	4. Using model A in which the subject died within 5 years is used as predictor, the estimate of the true mean LDL among a population of subjects who die within 5 years would be the intercept plus the slop, of which the value is 118.7. The estimate is the same to the corresponding point estimate from problem 1.
	5. Using model B in which the subject survived at least 5 years is used as predictor, 95% confidence interval for the true mean LDL among a population of subjects who die within 5 years is the 95% confidence interval for the intercept, which is between 112.7 and 124.7. It is slightly different from the previous estimate (112.1, 125.3) because the CI for the intercept is calculated from the pooled standard deviation instead of individual standard deviation of each group. The latter one is used in t test.
	6. The regression based estimate of the standard deviation within each group for each model is the pooled standard deviation (Root MSE). The value is 33.48, which is the same in two models and the same to the pooled standard deviation calculated in problem 1.
	7. Each model can be reparameterized to another model by creating the new predictor as 1- original predictor. And value of the new intercept plus the slop would be the value of old intercept.
	8. The intercept is the estimate of the true mean LDL among a population of subjects who survive at least 5 years; and its 95% confidence interval is the 95% confidence interval for the true mean LDL among a population of subjects who survive at least 5 years.
	9. The slop is the estimate of difference in means of LDL values between a population that survives at least 5 years and a population that dies with 5 years. Its 95% confidence interval is the 95% confidence interval for the true difference in the means between a population that survives at least 5 years and a population that dies with 5 years.
	10. The point estimate of difference in means of LDL values between a population that survives at least 5 years and a population that dies with 5 years is 8.50 mg/dL with a higher mean LDL levels among the group with at least 5-year-survial. The estimated standard error is 3.36. Based on the 95% confidence interval, this observed 8.50 mg/dL higher mean LDL among the group with least 5-year-survial would not be judged unusual if the true difference in the population means were anywhere between a 1.91 to 15.09 mg/dL. Using classical linear regression, this observation is statistically significant at a 0.05 level of significance (two-sided P= 0.012). We can conclude that the distribution of LDL values differs between those two groups. The point estimate, the estimated standard error of the point estimate, the 95% confidence interval are the same to the corresponding inference from problem 1.
1. The point estimate of the means of each group and the difference in the mean between two groups are the same in these two analyses. The 95% confidence intervals of the mean for each group and their standard deviations are the same in these two analyses. However, the standard errors and the 95% confidence intervals of the difference in means of LDL values between two groups are different between these two analyses. In this new analysis, the standard error is 3.57 (in problem 1: 3.36) and the 95% confidence 1.44-15.56 (in problem 1: 1.91-15.09). These differences come from the different assumption of standard error. In this analysis, the equation of standard error is



1. The answer is similar to that in question 3. The point estimate of the means of each group and the difference in the mean between two groups are the same in these two analyses. However, the 95% confidence intervals of the mean for each group and their standard deviations are different between these two analyses. Besides, the standard errors and the 95% confidence intervals of the difference in means of LDL values between two groups are different between these two analyses. In this new analysis, the standard error is 3.57 (in problem 3: 3.36) and the 95% confidence 1.50-15.50 (in problem 3: 1.91-15.09). These differences come from the different assumption of standard error. In this analysis, the equation of standard error is
* ***Methods:*** Descriptive statistics of age and LDL treated as a continuous variable, which included sample size, mean, standard deviation, minimum and maximum, presented within groups defined by male and female, as well as in the entire sample. Beside, p-values are also included to evaluate the association between age and sex, as well as LDL and sex.
* ***Results:*** The total sample size is 735 subjects; however, 10 of the subjects have missing data on LDL values. (Six of them are in the male group and 4 in the female group.) Those subjects are omitted from the analysis. The details of descriptive statistics are provided in the following table. The means and standard deviations of age between males and females are similar. In addition, the p-value of the association between age and sex is 0.418. We could conclude that there is no significant evidence to prove there is association between age and sex. The means of LDL among the male group is lower than the female group (120.6 mg/dl vs 130.9 mg/dl), though the standard deviations are similar. However, the p-value of the association between LDL and sex is <0.0001, which is significant. We could conclude that there is a significant association between LDL and sex. In sum, from the above data, we cannot determine whether sex is a confounder in the association between LDL and age.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Males (n=360) | Females (n=365) |  | All subjects (n=725) |
|  | Mean (SD) | Min, Max | Mean (SD) | Min, Max | P-value | Mean (SD) | Min, Max |
| Age (year) | 74.7 (5.63) | 66, 99 | 74.4 (5.26) | 65, 91 | 0.418 | 74.6 (5.45) | 65, 99 |
| Serum LDL (mg/dl) | 120.6 (32.15) | 37, 227 | 130.9 (34.25) | 11, 247 | 0.0001 | 125.8 (33.60) | 11, 247 |

* 1. In the data, LDL and age are both continuous variable. Because the means and the standard deviations are different, we will run the robust regression analysis to evaluate the association between LDL and age. The LDL are used as the response variable and age as predictor variable. In this analysis, mean LDL values will be compared across different age groups.
	2. No, this is not a saturated model. Because the predictor variable used in this analysis is a continuous variable. We also have to borrow information across the groups for the mean.
	3. In the regression analysis, we estimated there is a linear relationship between LDL and age. E(LDL|age)= 132.5-0.090\*age

From the above model, the estimated mean LDL among a population of 70-year-old subjects is 132.5-0.09\*70=126.2 (mg/dl)

* 1. In the regression analysis, we estimated there is a linear relationship between LDL and age. E(LDL|age)= 132.5-0.090\*age

From the above model, the estimated mean LDL among a population of 71-year-old subjects is 132.5-0.09\*71=126.11 (mg/dl). The difference between the answer to part e and part d is the estimated slope.

* 1. In the regression analysis, we estimated there is a linear relationship between LDL and age. E(LDL|age)= 132.5-0.090\*age

From the above model, the estimated mean LDL among a population of 71-year-old subjects is 132.5-0.09\*75=125.75 (mg/dl). The difference between the answer to part f and part d is the estimated slope (-0.09) times 5 (the difference between these two age groups).

* 1. It is the squared root of the mean of variance of LDL within the each age groups. It represents the average within group standard deviation of LDL levels.
	2. If we extrapolated the estimated LDL value to a newborn, the value of newborn would be the intercept (132.5). However, in the current data, we only had the age range from 65 to 99. Therefore, the intercept doesn’t have a relevant scientific interpretation here; it is just a mathematical construct to build up the model.
	3. In the scatterplot, we can find that a straight line relationship doesn’t exist. Thus, we interpret the slope as an average difference in mean LDL per one year difference in age.



* 1. From linear regression analysis, we estimate that for each year difference in age for the 725 subjects, the difference in mean LDL values is -0.09 mg/dl. A 95% CI suggests that this observation is not unusual if the true difference in mean LDL per year difference in age were between -0.55 and 0.37 mg/dl. Because the two-side P value is 0.698, we cannot have enough evidence to reject the null hypothesis that there is no linear trend in the average LDL across age groups.
	2. From linear regression analysis, we estimate that for each 5 years difference in age for the 725 subjects, the difference in mean LDL values is -0.45 mg/dl. A 95% CI suggests that this observation is not unusual if the true difference in mean LDL per 5 years difference in age were between -2.70 and 1.80 mg/dl. Because the two-side P value is 0.694, we cannot have enough evidence to reject the null hypothesis that there is no linear trend in the average LDL across groups that differ by 5 years in age.
	3. From correlations analysis for the 725 subjects, the two continuous variables, LDL and age, are not strongly correlated.( R=-0.0146 ) The two-sided p value of 0.6944 suggests that we can not with high confidence reject the null hypothesis that there is correlation in the LDL values and age.

The squared correlation is the value of R-squared in the regression analysis. And the p-value of correlation is the same to the p-value of classical linear regression.