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.

Method: Descriptive statistics for the distribution of LDL are presented within groups defined by death within 5 years, survival for 5 years post study entry, and for the entire sample. Mean LDL values are compared between subjects who died within 5 years of study enrollment and those who survived at least 5 years. Mean of LDL in each group defined by vital status were tested using a t-test under the assumption of equal variances.95% confidence intervals for the difference in population means were similarly based on that same handling of variances.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Time to death < 5 yrs.  | Time to death > 5 yrs.  | All patients.  |
| LDL (mg/dL) | 118.70 (119; 36.16) | 127.20 (606; 32.93) | 125.80 (725; 33.60) |

**Descriptive statistics presented are the mean (sample size; standard deviation)**

* 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?

For the group with patients survived at least 5 years, sample size is 606 (8 missing values). Sample mean and sample standard deviation are 127.2 mg/dL and 32.93 respectively. For the group with subjects died within 5 years, the sample size is 116 (2 missing values). And the sample mean and sample standard deviation for LDL value is 118.7 mg/dL and 36.16. Sample mean of LDL from survivor group is 7% higher than the other group so they are not similar. Sample standard deviation from survivor group is 10% lower than the other group so they are not closed to each other as well. ( I think ± 3% would be considered not similar)

* 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?

According to the result of the t-test, point estimate, the estimated standard error of that point estimate, and the 95% confidence interval for mean LDL for subjects who would survive at least 5 years are 127.2 mg/dL, 1.34 and [124.6 mg/dL, 129.8 mg/dL] respectively. For subjects who would die with 5 years, these values are 118.7 mg/dL, 3.31 and [112.1mg/dL, 125.3mg/dL] respectively. Point estimates are not similar as explained from part a. and standard errors are quite distinguished as well. Standard error of LDL for survivor group is about 60% lower than the other group. Point estimates are exactly equal to sample mean whereas estimates of standard errors are equivalent to $\frac{standard deviation}{sample size of group}$.

* 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. CI for the mean LDL in a population surviving 5 years overlaps with the CI for mean LDL in a population dying with 5 years. The C.I for survivor group is [124.6 mg/dL, 129.8 mg/dL] while the C.I for non-survivor group is [112.1mg/dL, 125.3mg/dL]. Since the point estimator for non-survivor group (118.7 mg/dL) is not contained in the CI for survivor group and the point estimator for survivor group (127.2 mg/dL) is not contained in the CI for non-survivor group, no conclusion about significance can be made. It would be better to conduct a 2 sample t-test and find out the p-value to check for the statistical significance. (lecture note 1)

* 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?)

Since we presume the variances are equal for both groups, we would want to standardize the standard deviation (or variance) of both groups. Therefore, choose pooled standard deviation is a good way to combine the two estimated sample standard deviation. The formula is: $pooled SD= \sqrt{\frac{(n\_{1}-1)s\_{1}^{2}+(n\_{2}-2)s\_{2}^{2}}{n\_{1}+n\_{2}-2}}$ and the pooled SD of LDL is 33.602 (calculated by hand).

The combined SD reported by t-test is not what we want to use.

* 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 is 8.5 mg/dL. Thus, survived group has a mean ldl level that is 8.5 mg/dL higher than the ldl level in death group. The estimated standard error of that is 3.36. The 95% confidence interval shows that our estimate of the difference would be reasonable if the true difference is between 1.91 mg/dL and 15.09 mg/dL. P-value is less than 0.05 at 95% significance level (p-value = 0.0115 for 2-tail test; 0.0058 for 1-tail test). Therefore, we can reject the null hypothesis of no difference of mean LDL between patients who died within 5 years and those who survived after 5 years of study in favor of alternative hypothesis that is patients who survived after 5 years tend to have a higher LDL level.

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.

Model A: Let explanatory variable = 1 if the subject died within 5 years. Let explanatory variable = 0 if the subject died after 5 years.

Model B: Let explanatory variable = 0 if the subject died within 5 years. Let explanatory variable = 1 if the subject died after 5 years.

Both models are saturated. In both case, we have two groups (binary, weather or not survived 5 years). And there are two parameters to be estimated in both models as well (intercept and slope). Since the number of parameters and the number of groups are equal, we say the models are saturated.

* 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?

I choose model A. The estimate of the true mean LDL among a population of subjects who survive at least 5 years is 127.2 mg/dL. It’s exactly the same as the point 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 survive at least 5 years? How does this compare to the corresponding estimate from problem 1? Explain the source of any differences.

I choose model A. The confidence interval for the true mean LDL among a population of subjects who survive at least 5 years is [124.5282,129.8679]. This is a little bit difference comparing to the C.I from problem 1. The reason is that regression analysis uses the root MSE (pooled SD) to calculate standard error for CI whereas t-test uses standard deviation within group to get the standard error for 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?

I choose model B. The estimate of the true mean LDL among a population of subjects who die within 5 years is 118.7 mg/dL. It’s exactly the same as the point 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.

I choose model B. The confidence interval for the true mean LDL among a population of subjects who survive at least 5 years is [112.6726, 124.7224]. This is a little bit difference comparing to the C.I from problem 1. The reason is that regression analysis uses the root MSE (pooled SD as assuming equal variance) to calculate standard error for CI whereas t-test uses standard deviation within group to get the standard error for CI.

* 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?

The regression based estimate of the standard deviation within each group for each model is the root MSE or pooled standard deviation. They are different because pooled SD uses a function to combine the SD from each group whereas combined SD is just the SD of combined groups.

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

Model A and B are exactly the same but reparameterized. Intercept parameter in model A represents the mean LDL value of survivor group whereas it represents the mean LDL of non-survivor group in model B. Slope parameter in model A means how much the true mean LDL among a population of subjects will drop if we switch a subject who survive at least 5 years to a subject who die within 5 years while slope parameter in model B means how much the true mean LDL among a population of subjects will increase if we switch a subject who die within 5 years to a subject who survive at least 5 years.

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

The intercept from model A is the estimate of the true mean LDL among a population of subjects who survive at least 5 years.

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

The slope from model A is the estimate of how much the true mean LDL among a population of subjects will drop if we switch a subject who survive at least 5 years to a subject who die within 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?

I choose to use model B. Point estimate for true difference in means is the slope estimate (or absolute slope estimate), which is 8.5 mg/dL. The estimated standard error of the point estimate is 3.36 and the 95% confidence interval is [1.91 mg/dL, 15.09 mg/dL]. P-value (2-tail) is 0.012. According to the p-value <0.05, we would reject the null hypothesis of two population have the same mean LDL in favor of alternative hypothesis that is there is a significant association between serum LDL and 5 year all cause mortality. The inference is the same as the one from 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.)

Standard error (3.574) and confidence interval ([1.44, 15.56]) for difference of population mean in two groups are different by using the t-test with unequal variances. T statistics, p-value and degree of freedom for testing the difference of mean between two groups are different as well. This is because we difference method to calculate the variance of difference of means. If we assume equal variance, we would use the pool SD whereas if we assume unequal variance, we would use $\sqrt{\frac{sd\_{1}^{2}}{n\_{1}}+\frac{sd\_{2}^{2}}{n\_{2}}}$ to find the standard deviation of difference in means.

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.)

By using the regression with unequal variances, standard error and confidence interval and p-value for difference of population mean in two groups are slightly different to the result of t-test that assumes unequal variance. This is because the regression using robust standard error approximates the t-test that allows unequal variance. That’s why the estimates are very closed but slightly changed from the ones when using t-test.

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.)

One appropriate descriptive statistics for binary continuous variables would be scatterplot.



The scatterplot above shows us the relationship between age and LDL. The explanatory variable, age is on the x-axis and the response variable, LDL (mg/dL) is on the y-axis. There are two fitted lines on the top of the scatterplot. The red one gives us the expected LDL across the age for male only and the green line gives us the expected LDL across the age for female only. Based on the fitted lines, we can see that female has higher LDL level on average than male does. The difference becomes more significant as age goes up.

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

The association between serum LDL and age is evaluated by performing a linear regression model assuming unequal variance across the age (heteroscedasticity). The estimates of slope and intercept parameter and their 95% confidence interval would be calculated. P-value would be obtained to evaluate if age has an impact or effect on LDL value.

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

No, this is not a saturated model because age as our explanatory variable is continuous in this case. This is say we might need to borrow information when using the model for predicting purpose. In short, the model here doesn’t contain complete information about the population.

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

The linear regression model we get is: $E\left(Age\right)=132.53-0.09×Age. $Therefore, if age is 70, mean LDL level predicted would be 126.23 mg/dL. This is saying on average, we expect that the population of 70-year-old subjects would have mean LDL level equal to 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 (d?) relate to the slope?

If we plug in 71 into the regression model, we would get E(LDL|Age=71) = 126.14 mg/dL. The difference of the expected mean LDL level between 70-year-old and 71-year-old is 0.09 which is just the absolute value of the slope. This matches with the interpretation of the slope in the linear regression model.

* 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?

If we plug in 75 into the regression model, we would get E(LDL|Age=75) = 125.78 mg/dL. The difference of the expected mean LDL level between 70-year-old and 75-year-old is 0.45 which is exactly the absolute value of the 5×slope estimate.

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

Root mean squared error is the sample standard deviation of residuals under the assumption of unequal variance.

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

Intercept reflects the estimated mean value of response variable (LDL level in this case) when explanatory variable is 0 (age = 0). Here, it doesn’t have a relevant scientific meaning because our interest is mainly focusing on the LDL of the elder people instead of newborn babies. Meanwhile, it doesn’t make sense that newborn babies would have the highest LDL than any other group of subjects.

* 1. What is the interpretation of the slope?

Slope tells us how much the value of response variable (LDL) will change if we change the explanatory variable (age) by 1 unit.

* 1. Provide full statistical inference about an association between serum LDL and age based on your regression model. 95 C.I slope p-value

Since the intercept parameter estimate is not scientifically meaningful, we wouldn’t want to include it in the inference. The estimate of slope parameter is -0.09. This means we would expect to have 0.09 mg/dL decrease in LDL level when we have the estimate of mean age of population increased by 1. The 95% confidence interval shows that the slope estimate would be reasonable if the true slope is between -0.547 and 0.367. The p-value for the slope is 0.698. This is based on the hypothesis of testing if the true slope value is equal to 0 or not. Since the p-value is greater than 0.05, we can’t reject the null hypothesis that is: there is no association between serum LDL and age.

* 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?

We would need change the slope estimate to 5×(-0.09) = -0.45. This follows directly from part f. Part f told us that mean LDL level would drop 0.45 mg/dL if we increased age by 5. Here we would just treat 5-year increase as one segment (unit). The new 95% CI would be [-2.735, 1.835]. To get it, we just multiply the 95% CI from part j by 5. This is because the formula to calculate CI is: point estimate ± t ×($\frac{SD}{\sqrt{n}}$). Since both the new point estimate and estimate of SD increased by 5 times, we can get the new CI by just multiplying 5.

* 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 correlation between age and LDL level is -0.0146. This is saying the correlation between age and LDL is almost 0 which can be ignored. Therefore, we can conclude that there’s no association between LDL and age. The regression-based result gives us the same conclusion (p-value shows insignificance).