1. (a)

The sample size, sample mean, and sample standard deviation of LDL values for subjects surviving less than or at least 5 years are presented in the table below

|  |  |  |
| --- | --- | --- |
|  | Less than 5 years | At least 5 years |
| Sample size | 119 | 606 |
| Sample mean | 118.70 mg/dl | 127.20 mg/dl |
| Sample standard deviation | 36.16 mg/dl | 32.93 mg/dl |

We observe that the sample mean of ldl for subjects surviving less than 5 year is less than the sample mean of ldl for subjects surviving at least 5 years. The sample standard deviation for subjects surviving less than 5 years is bigger than the sample standard deviation for subjects surviving at least 5 years.

(b)

The point estimate, the estimated standard error of the point estimate, and the 95% CI for the true mean LDL in a population of similar subjects who would survive less than or at least 5 years are presented in the table below.

|  |  |  |
| --- | --- | --- |
|  | Less than 5 years | At least 5 years |
| Point estimate | 118.70 | 127.20 mg/dl |
| Standard error | 3.315 mg/dl | 1.338 mg/dl |
| 95% CI | (112.13, 125.26) mg/dl | (124.57, 129.83) mg/dl |

For subjects surviving less than 5 years, the point estimate of mean ldl is less than that for subjects surviving at least 5 years while the standard error is larger compared to those subjects surviving at least 5 years.

The point estimate for the population and the sample means are the same. However, the standard error is different from the sample standard deviation. This is because the sample size is taken into account when estimating the standard error. That is, standard error hence it’s smaller than the sample standard deviation.

(c)

The confidence interval for the mean ldl in the population surviving less than 5 years is (112.13, 125.26) while for those surviving at least 5 years is (124.57, 129.83). The two confidence intervals overlap albeit slightly.

I can’t make a conclusion on the statistical significance at an alpha level of 0.05 based on the observed overlap. This is because there is a possibility that the confidence intervals overlap but the mean difference is still statistically significant.

(d)

Assuming equal variances but allowing the possibility of different means, we use pooled standard deviation as the estimate of standard deviation of LDL. This is given by where, and , are sample sizes and sample standard deviation for the two groups.

In this problem, the estimate would be 33.602 mg/dl

(e)

Point estimate=8.501, standard error=3.357, and the 95% CI is (1.911, 15.09).

The two sided p-value is 0.0115 which is statistically significant at 0.05 alpha level. With high confidence, we reject the null hypothesis that the mean serum LDL levels are not different by vital status at 5 years in favor of a hypothesis that death within 5 years is associated with lower mean serum LDL.

1. (a)

Both models are saturated since the predictor of interest is binary and there are two parameter estimates from the regression analysis.

(b)

Considering model B, the estimate of the true mean LDL among a population of subjects who survive at least 5 years is118.7 mg/dl. This is exactly the same as the estimate from problem 1.

(c)

Considering model B, the confidence interval for the true mean LDL among population of subjects who survive at least 5 years is (112.67, 124.72) mg/dl. There is a slight variation from estimate in problem 1. The variation arises from the use of pooled standard deviation in computing the confidence intervals unlike in problem one where the sample standard deviation is used.

(d)

Considering model A, the estimate of the true mean LDL among a population of subjects who survive less than 5 years is 127.2 mg/dl. This is exactly the same as the estimate from problem 1.

(e)

Considering model A, the confidence interval for the true mean LDL among population of subjects who survive less than 5 years is (124.53, 129.87) mg/dl. There is a slight variation from the estimate in problem 1. This difference arises as a result of using pooled standard deviation in computing the confidence intervals. In problem 1, the sample standard deviation is used.

(f)

Assuming equal variances, the regression-based estimate of the standard deviation within each group for each model is 33.48mg/dl. This is the root mean square error (RMSE) while for problem one it’s the pooled standard deviation.

(g)

Models A and B are similar since they both involve re-parameterization of the predictor variable comparing the mean difference in LDL between the two vital survival groups.

(h)

Model A intercept: The intercept represents the mean LDL when the indicator is 0 which corresponds to the subjects who survived at least 5 years. Therefore we can interpret the intercept as average LDL value for the subjects who survived at least 5 years.



Model A slope: The mean LDL is estimated to differ between the two vital groups by 8.501mg/dl (on average), with the group surviving less than 5 years tending to have a lower average LDL.

(j)

The point estimate for true difference in means between a population that survives at least 5 years and a population dies within 5 years using regression is 8.501 (s.e=3.3566) mg/dl and the 95% CI is (1.911, 15.09). The p-value is 0.012 which leads us to reject with high confidence the null hypothesis that the mean serum LDL levels are not different by vital status at 5 years in favor of a hypothesis that death within 5 years is associated with lower mean serum LDL. This inference is similar to the one in problem 1.

1. Comparing the results from the t test assuming equal variances and t test assuming unequal variances, I observe that the point estimate for true difference is the same but the standard error, p-value, and the 95% CI are different. In either case, the overall inference decision stays the same.
2. Comparing the results using robust linear regression and t test with unequal variances, I observe differences in the p-value, standard error and the 95% confidence. The point estimate for the true difference in mean between the two vital groups is the same. The differences observed are as a result of the approximation of the t test using Huber-White sandwich estimator.
3. (a)

Below is a table for descriptive statistics of ldl and age obtained by classical linear regression method

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| LDL | Estimate | Standard error | 95% CI | p-value |
| age | -0.0902 | 0.2294 | -0.5406, 0.3603 | 0.694 |
| Intercept | 132.53 | 17.15 | 98.85, 166.21 |  |

With high confidence, we fail to reject the null hypothesis of no association between ldl and age at an alpha level of 0.05.

To evaluate whether this association may be confounded or modified by sex, we add sex to the model and compare the estimates of the new model with the estimates from the model containing age as the only covariate. The descriptive statistics from the new model are presented in the table below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| LDL | Estimate | Standard error | 95% CI | p-value |
| age | -0.0616 | 0.2270 | -0.5072, 0.3840 | 0.786 |
| Intercept | 135.52 | 16.98 | 102.19, 168.86 |  |

The new estimates from this model don’t show any significant difference from the previous estimates in the model consisting of age as the only covariate. We therefore, with high confidence conclude that the association between ldl and age is not confounded by sex.

Below is a scatter plot of ldl and age fitted for each sex.



The two fitted lines representing both males and females don’t show a big variation in the slopes which further indicates that the effect of sex on the association between ldl and age is not significant.

(b)

The model fit allows age to have a distinct average LDL and I use linear regression model with the continuous age variable. I choose to perform classical linear regression instead of robust regression. Both methods yield the same estimates but the standard error, confidence intervals and p-value are different.

(c)

This model is not saturated since it involves a continuous variable as the predictor of interest. Hence there can be infinitely many data points as our predictors of interest.

(d)

The mean LDL level at 70 years is estimated to be 132.53-0.0902\*70=126.22mg/dl

(e)

The mean LDL level at 71 years is estimated to be 132.53-0.0902\*71=126.13mg/dl and the difference with respect to the 70 year old subjects is 126.12-126.22=-0.09mg/dl. This difference is similar to the slope.

(f)

The mean LDL level at 75 years is estimated to be 132.53-0.0902\*75=125.77mg/dl and the difference with respect to the 70 year old subjects is 125.77-126.22=-0.45mg/dl. This difference is 5 times the value of the slope.

(g)

The root mean squared error is the standard deviation of LDL within each age group.

(h)

The intercept can be interpreted as the average LDL among newborns (age 0) and is estimated to be 132.53 mg/dl. An age group of 0 is way outside the range of our data (min age=65) and it would make to no scientific sense using the estimate.

(i)

The mean LDL is estimated to differ between two age groups by 0.0902mg/dl (on average) for each year difference in age, with the younger group tending toward higher LDL. It is a measure of association, interpretable as a first order trend in the means.

(j)

From linear regression analysis, we estimate that mean LDL differs between two age groups by 0.0902mg/dl (on average) for each year difference in age, with the younger group tending towards slightly higher LDL average. This result is not significantly different from 0 (p=0.694), with a 95% CI suggesting that such observed results would not be unusual if the true difference in mean LDL between age groups were anywhere between 0.5406 mg/dl lower and 0.3602 mg/dl higher for each year difference in age, with the younger group tending toward higher LDL. We thus fail to reject the null hypothesis that mean LDL difference doesn’t differ across age groups.

(k)

For groups differing by 5 years of age, we would multiply the point estimate by 5 and the confidence interval as well. Therefore the point estimate for the mean reported in this case would be 0.451mg/dl on average and the CI is (-2.703, 1.801) mg/dl.

(l)

We estimate a correlation of -0.0146 which is statistically not significant from 0 (p=0.6944). This estimate is similar to the signed square root of the R squared reported in a simple linear regression of LDL on age. The p-value agrees exactly with the statistical significance of the LDL parameter in classical simple linear regression.