**1.**

**(a)**

Of the 735 observations in the dataset, 10 had missing values of LDL and were excluded from the table below, showing sample size, sample mean LDL, and sample standard deviation of LDL.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Sample Size** | **Sample Mean (mg/dL)** | **Sample SD****(mg/dL)** |
| **Surviving at least 5 years** | 606 | 127.198 | 32.929 |
| **Died within 5 years** | 119 | 118.698 | 36.157 |

The sample mean of LDL is 127.198 mg/dL among 5-year survivors and 118.698 mg/dL among those who died within 5 years. This difference in sample means, 8.500 mg/dL, represents a 7.2% increase in the sample mean LDL among those who died within 5 years. As a result, I believe the sample means are similar in magnitude.

The sample SD of LDL is 32.929 mg/dL among survivors and 36.157 mg/dL among those who died within 5 years. This difference in sample SDs, 3.228 mg/dL, represents a 9.8% increase in sample SD of LDL among those who survived at least 5 years. As a result, I think the sample SDs are similar in magnitude.

**(b)**

The table below reports the point estimate, corresponding estimated standard error, and the 95% confidence interval (based on the normal approximation) for the population mean of LDL in a population of similar elderly subjects. The units of all numbers in the table are mg/dL.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Point****Estimate** | **Estimated****Standard Error** | **95%****Confidence Interval** |
| **Surviving at least 5 years** | 127.198 | 1.338 | [124.571, 129.825] |
| **Died within 5 years** | 118.698 | 3.315 | [112.134, 125.261] |

As in part (a), the sample mean of LDL is 127.198 mg/dL among 5-year survivors and 118.698 mg/dL among those who died within 5 years. This difference in sample means, 8.500 mg/dL, represents a 7.2% increase in the sample mean LDL among those who died within 5 years. I think the sample means are similar in magnitude.

However, the standard error of LDL is 1.338 mg/dL among 5-year survivors and 3.315 mg/dL among those who died within 5 years. This difference in standard errors, represents an increase of 147.8% from the standard error of LDL among 5-year survivors. As a result, I believe the standard errors *are* different in magnitude. This can be explained by the relative sample sizes of the two groups. There are 606 observations who survived at least 5 years compared to only 119 who died within 5 years. Since SE here is defined as sample standard deviation divided by sample size, it is calculated for those who died within 5 years by dividing the sample standard deviation by $\sqrt{119} ≈10.909$, while SE is calculated for 5-year survivors by dividing that group’s sample standard deviation by $\sqrt{606} ≈24.617$, a significantly larger number.

**(c)**

As shown in the table for part (b), the 95% CI for the mean LDL in an elderly population of 5-year survivors is [124.571 mg/dL, 129.825 mg/dL] and the 95% CI for the mean LDL in an elderly population dying within 5 years is [112.134 mg/dL, 125.261 mg/dL]. The 95% CIs between the two groups overlap. However, we *cannot* conclude that, assuming the two populations have equal variances, the estimated difference in mean LDLs would give us sufficient evidence to reject the null hypothesis that the two populations have the same mean LDL at the 0.05 level. We cannot safely reach this conclusion because the standard error used to produce a 95% CI for the difference in means is calculated as $\sqrt{SE\_{Group 1}^{2}+SE\_{Group 2}^{2}}$, while the SE used to compute the confidence interval for each vital group is either $SE\_{Group 1}$ or $SE\_{Group 2}$ and

$\sqrt{SE\_{Group 1}^{2}+SE\_{Group 2}^{2}}< SE\_{Group 1}+ SE\_{Group 2}$.

That is, it’s possible for the confidence intervals computed for individual group means of LDL to overlap while the confidence interval for the difference in mean LDL does not contain 0 mg/dL.

**(d)**

If we assume equal variances in both populations but want to allow for their means to be different, we would combine the two estimated sample standard deviations with the following pooled SD:

$$SD\_{pooled}=\sqrt{\frac{(N\_{S}-1)S^{2}\_{s}+(N\_{D}-1)S^{2}\_{D} }{N\_{S}+N\_{D}-2}}$$

where $N\_{S}$ and $S^{2}\_{s} $are the observation count and sample variance, respectively, of 5-year survivors, while$ N\_{D}$ and $S^{2}\_{D} $are the observation count and sample variance, respectively, of those who died within 5 years.

**(e)**

For the true difference in mean LDL between a population of elderly individuals that survives at least 5 years and an elderly population that dies within 5 years, we present the point estimate, standard error estimate, 95% CI, and two-sided p-value from a two-sample t-test that assumes equal variances between groups in the table below. The sample size for this two-sample t-test was 725 (606 5-year survivors and 119 individuals who died within 5 years).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Point Estimate****(mg/dL)** | **Standard Error****(mg/dL)** | **95% CI****(mg/dL)** | **Two-sided** **p-value** |
| **True Difference in mean LDL\*** | 8.501 | 3.357 | [1.911, 15.090] | 0.012 |

\* Expressed as the mean LDL of an elderly population of 5-year survivors minus the mean LDL of an elderly population that dies within 5 years.

The 95% CI from our two-sample t-test that assumes equal variances between groups tells us that our sample is consistent with a true difference in mean LDL of a population of 5-year survivors relative to a population that dies within 5 years of 1.911 mg/dL and 15.090 mg/dL. Our two-sided p-value (which comes from a test of the null hypothesis that the two populations have equal mean LDLs) is 0.012 < 0.05, so our sample contains sufficient evidence to reject the null hypothesis at the 0.05 level of significance. Using the two-sample t-test for a difference in means that presumes equal variance between the two groups, we conclude that there is a statistically significant association between 5-year all-cause mortality and the mean serum LDL.

**2.**

**(a)**

We say that a model is *saturated* if its number of groups is equal to its number of parameters. Both model A and model B are saturated, since in each case the regression model has two parameters ($\hat{β\_{0}}, \hat{β\_{1}}$) and the predictor variable has only two possible levels (dead within 5 years or alive at 5 years). So the number of groups and the number of parameters in the model are both equal to 2.

**(b)**

I use the regression parameter estimates of model A, where the predictor is an indicator that an observation died within 5 years. Here, the estimate of the true mean LDL in a population of individuals who survive at least 5 years is (where the units are mg/dL)

$$\hat{β\_{0}}+ \hat{β\_{1}}×1=127.198-8.501 ×0=127.198 mg/dL$$

This is the same as the corresponding estimate in problem 1.

**(c)**

I use the regression parameter estimates of model A, where the predictor is an indicator that an individual died within 5 years. Here, the 95% CI for the true mean LDL in a population of individuals who survives at least 5 years is the 95% CI ($\hat{β\_{0})}$ = [124.528 mg/dL, 129.868 mg/dL]. This is slightly different than the corresponding estimate from problem 1, [124.571 mg/dL, 129.825 mg/dL].

This can be explained a difference in the underlying estimates of standard errors and critical t-values used by the 95% CIs for the group of 5-year survivors in the t-test in problem 1 (SE=1.338 mg/dL) and the regression estimate in model A (SE=1.360 mg/dL). The regression estimate uses *pooled* SE according to the formula

$${\sqrt{\frac{(N\_{S}-1)S^{2}\_{s}+(N\_{D}-1)S^{2}\_{D} }{N\_{S}+N\_{D}-2}}}/{\sqrt{N\_{S}}}$$

where $N\_{S}$ and $S^{2}\_{s} $are the observation count and sample variance, respectively, of 5-year survivors, while$ N\_{D}$ and $S^{2}\_{D} $are the observation count and sample variance, respectively, of those who died within 5 years.

On the other hand, the SE from problem 1 is given by

$${S\_{S}}/{\sqrt{N\_{S}}}$$

where $N\_{S}$ and $S\_{S}$ are the sample size and sample standard deviation, respectively, of 5-year survivors.

In addition, the degrees of freedom used to obtain the critical t- value for the 95% CI differ between the regression estimate and the estimate from problem 1. For the regression, the critical value is equal to $t\_{.025, N\_{s}+N\_{D}-1}$. For problem 1, the critical value is $t\_{.025, N\_{S}-1}$, a value that is necessarily larger than the critical value used for the CI by the regression estimate since $N\_{D}>0.$

**(d)**

I use the regression parameter estimates of model A, where the predictor is an indicator that an observation died within 5 years. Here, the estimate of the true mean LDL in a population of individuals who die within 5 years is (where the units are mg/dL)

$$\hat{β\_{0}}+ \hat{β\_{1}}×1=127.198-8.501 ×1=118.698 mg/dL$$

This is the same as the corresponding estimate in problem 1.

**(e)**

I use the regression parameter estimates of model B, where the predictor is an indicator that an individual survived at least 5 years. Here, the 95% CI for the true mean LDL in a population of individuals who dies within 5 year is [112.672 mg/dL, 124.722 mg/dL]. This is slightly different than the corresponding estimate from problem 1, [112.134 mg/dL, 125.261 mg/dL].

This can be explained a difference in the underlying estimates of standard errors and critical t-values used by the 95% CIs for the group of individuals who died within five in the two-sample t-test in problem 1 (3.315 mg/dL) and the regression estimate in model A (3.357 mg/dL). The regression estimate uses *pooled* SE according to the formula

$${\sqrt{\frac{(N\_{S}-1)S^{2}\_{s}+(N\_{D}-1)S^{2}\_{D} }{N\_{S}+N\_{D}-2}}}/{\sqrt{N\_{D}}}$$

where $N\_{S}$ and $S^{2}\_{s} $are the observation count and sample variance, respectively, of 5-year survivors, while$ N\_{D}$ and $S^{2}\_{D} $are the observation count and sample variance, respectively, of those who died within 5 years.

On the other hand, the SE from problem 1 is given by

$${S\_{D}}/{\sqrt{N\_{D}}}$$

where $N\_{D}$ and $S\_{D}$ are the sample size and sample standard deviation, respectively, of observations that died within 5 years.

In addition, the degrees of freedom used to obtain the critical t- value for the 95% CI differ between the regression estimate and the estimate from problem 1. For the regression, the critical value is equal to $t\_{.025, N\_{s}+N\_{D}-1}$. For problem 1, the critical value is $t\_{.025, N\_{D}-1}$, a value that is necessarily larger than the critical value used for the CI by the regression estimate since $N\_{S}>0.$

**(f)**

If we assume the variances are equal in the two populations, the regression-based estimated of the standard deviation for each group in each model is

$$S\_{pooled}=\sqrt{\frac{(N\_{S}-1)S^{2}\_{s}+(N\_{D}-1)S^{2}\_{D} }{N\_{S}+N\_{D}-2}}=33.477 mg/dL$$

The corresponding estimate in problem 1 for 5-year survivors would be

$$S\_{s}=32.929 mg/dL$$

and the corresponding estimate in problem 1 for individuals who died within 5-years is

$$S\_{D}=36.157 mg/dL$$

The regression-based estimate of the within-group standard deviation is a weighted average of the sample standard deviations of the two groups.

**(g)**

Models A and B both estimate the association between mean LDL values (measured in mg/dL) and vital status at 5 years by fitting a least squares regression line under the assumption that the variance of LDL is the same in both vital status groups. The models are the same except that the predictor of interest (LDL) has been re-parameterized between the models.

**(h)**

Since model A is saturated, the intercept of 127.198 mg/dL is exactly the sample mean LDL for observations that survived for at least 5 years.

**(i)**

Since model A is saturated, the slope of -8.501 mg/dL is exactly the difference in sample mean LDL for individuals who died within 5 years relative to individuals who survived for at least 5 years (i.e. the sample mean LDL for survivors was 8.501 mg/dL higher).

**(j)**

The table on the following page gives the point estimate, corresponding standard error estimate, and 95% CI for the true difference in mean LDL between a population of elderly individuals that survives at least 5 years relative to a population of elderly individuals that dies within 5 years. The table also provides the two-sided p-value for the null hypothesis that the true mean LDL is the same in both populations. I use the regression parameter estimates from ordinary least squares regression that assumes homoscedasticity (from model B) to fill in this table. The sample size for this regression was 725 (606 5-year survivors and 119 individuals who died within 5 years).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Point Estimate****(mg/dL)** | **Standard Error****(mg/dL)** | **95% CI****(mg/dL)** | **Two-sided** **p-value** |
| **True Difference in mean LDL\*** | 8.501 | 3.357 | [1.911, 15.090] | 0.012 |

\* Expressed as the mean LDL of an elderly population of 5-year survivors minus the mean LDL of an elderly population that dies within 5 years.

The two-sided p-value (which comes from a test of the null hypothesis that the two populations have equal means) is 0.012 < 0.05, so our sample contains sufficient evidence to reject the null hypothesis at the 0.05 level of significance. Using ordinary least squares regression that assumes homoscedasticity, we conclude that there is a statistically significant association between 5-year all-cause mortality and the mean serum LDL. This is the same inference we made in problem 1.

**3.**

For the true difference in mean LDL between a population of elderly individuals that survives at least 5 years and an elderly population that dies within 5 years, we present the point estimate, standard error estimate and two-sided p-value from a two-sample t-test that does *not* assume equal variances between groups in the third row of the table below. The sample size for this two-sample t-test was 725 (606 5-year survivors and 119 individuals who died within 5 years). The first two rows provide the descriptive statistics given in problem 1 (b).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Sample Size** | **Point Estimate****(mg/dL)** | **Standard Error****(mg/dL)** | **95% CI****(mg/dL)** | **Two-Sided** **p-value** |
| **Surviving at least 5 years** | 606 | 127.198 | 1.338 | [124.571, 129.825] | -- |
| **Died within 5 years** | 119 | 118.698 | 3.315 | [112.134, 125.261] | -- |
| **True Difference in mean LDL\*** | -- | 8.501 | 3.574 | [1.441, 15.560] | 0.019 |

\* Expressed as the mean LDL of an elderly population of 5-year survivors minus the mean LDL of an elderly population that dies within 5 years. Does *not* assume equal variance between groups

Our two-sided p-value (which comes from a test of the null hypothesis that the two populations have equal mean LDLs) is 0.019 < 0.05, so our sample contains sufficient evidence to reject the null hypothesis at the 0.05 level of significance. Using the two-sample t-test for a difference in means that does *not* assume equal variance between the two groups, we conclude that there is a statistically significant association between 5-year all-cause mortality and the mean serum LDL.

Analysis using the two-sample t-test for a difference in means that does *not* assume equal variance yields a different estimated standard error, 95% CI and two-sided p-value than in problem 1. The point estimate for the true difference in mean LDL between the two population groups is the same. It turns out that our estimated 95% CI in problem 1 was anti-conservative; the observed sample variance in the larger group (5-year survivors) was smaller than that of individuals who died within 5 years. As a result, the larger sample variance of those who died within 5 years was dominated by the smaller estimated variance for 5-year survivors in our estimated SE in problem 1 (e) that assumed equal variances between groups. Our two-sided p-value of 0.019 here is larger than the two-sided p-value of 0.012 in problem 1 (e), also due to a larger SE estimate for the true difference in mean LDL when we do not assume equal variances between groups. As a result, we still have evidence to reject the null hypothesis (at the .05 level of significance) that the population mean LDLs for the two groups are equal.

**4.**

I conduct ordinary least squares regression that does not assume homoscedasticity, using the model A specification (where my predictor indicates that an observation died within 5 years). The sample size for this regression was 725 (606 5-year survivors and 119 individuals who died within 5 years). I present results in the table below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Point Estimate****(mg/dL)** | **Robust Standard Error\*\*****(mg/dL)** | **95% CI****(mg/dL)** | **Two-sided** **p-value** |
| **Intercept** | 127.198 | 1.338 | [124.570, 129.826] | $$<10^{-5}$$ |
| **True Difference in mean LDL\*** | -8.501 | 3.566 | [1.500, 15.051] | 0.017 |

\* Expressed as the mean LDL of an elderly population of 5-year survivors minus the mean LDL of an elderly population that dies within 5 years. Does *not* assume equal variance between groups

Using ordinary least squares regression that does *not* assume homoscedasticity, we find a two-sided p-value (which comes from a test of the null hypothesis that the two populations have equal means) of 0.017 < 0.05, so our sample contains sufficient evidence to reject the null hypothesis at the 0.05 level of significance.

However, our robust standard error estimate for the true difference in mean LDL (3.566 mg/dL) is slightly smaller than our estimated standard error for this difference in problem 3 (3.574 mg/dL). As a result, our 95% for the difference in mean LDLs from these regression results, [1.500, 15.051], is narrower than the 95% CI from problem 3, [1.441, 15.560]. In addition, our two-sided p-value for problem 4 (p=0.017) is slightly smaller than the two-sided p-value from problem 3 (p=0.019). As a result, we still have evidence to reject the null hypothesis (at the .05 level of significance) that the population mean LDLs for the two groups are equal.

**5.**

**(a)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Arithmetic Mean** | **SD** | **Min** | **Max** |
|  | ***Overall (N=725)*** |
| Serum LDL level (mg/dL) | 125.80 | 33.60 | 11 | 247 |
| Male | 49.7% | -- | -- | -- |
|  | ***Individuals 65-74 Years Old (N = 417)*** |
| Serum LDL level (mg/dL) | 125.97 | 32.45 | 37 | 247 |
| Male | 48.9% | -- | -- | -- |
|  | ***Individuals 75-84 Years Old (N=264)*** |
| Serum LDL level (mg/dL) | 125.63 | 34.87 | 11 | 227 |
| Male | 49.2% | -- | -- | -- |
|  | ***Individuals 85 Years and Older (N=44)*** |
| Serum LDL level (mg/dL) | 125.25 | 37.23 | 57 | 216 |
| Male | 59.1% | -- | -- | -- |

The descriptive statistics presented in the table presented above indicate that the mean LDL level does not vary considerably between the three age groups defined as 65-74 years old, 75-84 years old, and above 84 years old. In each age group, the mean LDL is between 125 mg/dL and 126 mg/dL, while its value in the full sample is 125.8 mg/dL.

The descriptive statistics suggest that sex may be associated with age and, thus, sex may potentially confound the association between LDL and age. Among the 44 observations aged 85 or older, 59% (26 observations) are male, while in the full sample only 49.7% of observations are male. It is worth noting that 85+ age group has a small sample size compared to the other two groups (44 individuals compared to 417 individuals aged 65-74 and 264 individuals aged 75-84). In addition, the full sample share of males, 49.7%, in this age group would be roughly 22 individuals, and the difference between 22 males and 26 males in this particular age group may not be scientifically meaningful.

**(b)**

To evaluate a potential association between serum LDL and age, I conduct ordinary least squares regression that does *not* assume homoscedasticity. This model is shown below:

$E\left[X\_{i}\right]= β\_{0}+ β\_{1}X\_{1, i}$,

where $Y\_{i}$ is the LDL level of the *i*th observation and $X\_{1, i}$ is age in the *i*th observation. Since my predictor of interest is age, my coefficient of interest is $β\_{1}$ and I will make inference using my estimate, $\hat{β\_{1}}$, which is the estimated difference in mean LDL between individuals whose ages differ by 1 year.

**(c)**

We say that a model is *saturated* if its number of groups is equal to its number of parameters. This model is *not* saturated because it contains 2 parameters ($β\_{0}, β\_{1}$), but there is many more than 1 level for “continuous” age variable. As a result, we cannot fit the sample mean LDL value for each age group exactly.

**(d)**

Estimating the model specified in part (b) yields the following

$$E\left[LDL\_{i}\right| Age\_{i}]=132.528-0.090 ×Age\_{i}$$

As a result, we can write

$$E\left[LDL\_{i}\right| Age\_{i}=70]=132.528-0.090 ×70=126.215 mg/dL$$

So our estimated mean LDL in a population of 70-year-old subjects is $126.215 mg/dL.$

**(e)**

Estimating the model specified in part (b) yields the following

$$E\left[LDL\_{i}\right| Age\_{i}]=132.528-0.090 ×Age\_{i}$$

As a result, we can write

$$E\left[LDL\_{i}\right| Age\_{i}=71]=132.528-0.090 ×71=126.125 mg/dL$$

So our estimated mean LDL in a population of 70-year-old subjects is $126.125 mg/dL. $The difference between this and my answer to part (d) is $126.215\frac{mg}{dL}- 126.125\frac{mg}{dL} =0.090\frac{mg}{dL}$, which is equal to the estimated slope, $β\_{1}$.

**(f)**

Estimating the model specified in part (b) yields the following

$$E\left[LDL\_{i}\right| Age\_{i}]=132.528-0.090 ×Age\_{i}$$

And we can write

$$E\left[LDL\_{i}\right| Age\_{i}=75]=132.528-0.090 ×75=125.764 mg/dL$$

The difference between this and my answer to part (d) is $126.215\frac{mg}{dL}- 125.764\frac{mg}{dL} =0.450\frac{mg}{dL}$, which is equal to 5 times estimated slope, or $5× \hat{β}\_{1}$.

**(g)**

The root mean squared error in my regression model is an estimate of the standard deviation within age “groups” of the continuous age variable. If we assume equal variance between groups, RMSE can be expressed as

$$root MSE= \sqrt{\frac{RSS}{df\_{R}}}$$

where RSS is the residual sum of squares, and $df\_{R} $is the degrees of freedom of the residual. RMSE can also be interpreted as the weighted average of the sample standard deviations within the age groups.

**(h)**

The value of the intercept (132.528 mg/dL) estimates the mean of LDL for observations with an age of 0 (newborns). This does not have a relevant scientific interpretation. Note also that the sample which gives rise to this interval is only composed of elderly individuals (aged 65 or above) and we are interested in making inference on this elderly population, not younger individuals or newborns.

**(i)**

The slope from this model (-0.090 mg/dL) estimates the true difference in mean LDL among a population of elderly individuals with a common, fixed age relative to a population of elderly individuals whose age is 1 year lower. The magnitude of the slope estimates the true difference in mean LDL between populations of elderly individuals whose ages differ by 1 year.

**(j)**

The regression model estimates a true difference in population mean LDL between elderly individuals whose ages differ by 1 year of -0.090 mg/dL. That is, it estimates that a 1-year increase in the age of an elderly population is associated with a decrease in the population mean LDL of 0.090 mg/dL. However, the standard error of this estimate is 0.229 mg /dL and the corresponding 95% confidence interval is [-0.541 mg/dL, 0.360 mg/dL]. That is, our observed sample would not be surprising if the true difference in population mean LDL between elderly individuals whose ages differ by 1 year were between -0.541 mg/dL and 0.360 mg/dL. We note that 0 mg/dL is contained in this interval. The model also reports a two-sided p-value of 0.694; that is, given that there was no true difference in population mean LDL for elderly individuals of different ages, there is roughly a 69.4% chance of observing a sample as or more extreme than what we actually observe. As a result, we do not have evidence of a statistically significant association between age and serum LDL at the .05 level of significance (two-sided p-value = 0.694 > .05).

**(k)**

If we wanted an estimate and CI for the true difference in mean LDL across observations whose ages differ by 5 years, I would report the regression model’s estimate of $β\_{1}$ multiplied by 5.This is because we would model this case as

$$E\left(Y\_{i} | X\_{i}=x+5\right)= β\_{0}+β\_{1}\left(x+5\right)= β\_{0}+ β\_{1}×x+5×β\_{1}$$

and our coefficient of interest becomes $5×β\_{1}.$

For our data, I would report an estimate and CI for the true difference in mean LDL (both measured in mg/dL) across observations whose ages differ by 5 years as

Estimate:$ 5×\hat{β\_{1}}=-0.451$

$$95\% CI= 5×\hat{β\_{1}} \pm t\_{.025,723}×SE(5×\hat{β\_{1})}$$

$$= [-2.702{mg}/{dL} , 1.801 mg/dl]$$

**(l)**

Note that in our regression model the hypothesis test of

$$H\_{0}:β\_{1}=0 vs. H\_{A}:β\_{1}\ne 0$$

is the same test we conduct for nonzero correlation between LDL and age. We find an estimated correlation between LDL and age of -0.146. However, the two-sided p-value associated with this estimate is 0.694. That is, given that there was no correlation in population LDL and age, there is roughly a 69.4% chance of observing a sample as or more extreme than what we actually observe. As a result, we do not have evidence of a nonzero correlation between serum LDL and age at the .05 level of significance (two-sided p-value = 0.694 > .05). We reach the same conclusion about an association between LDL and age with this inference about correlation as we did with our regression-based inference in part (j).