Total score= 14+18.5+10+10+38=90.5

Question 1. (score: 14)

(a)

|  |  |  |  |
| --- | --- | --- | --- |
|   | Survived at least 5 years | Died within 5 years | All subjects |
| Variable | N (%) | Mean (SD) | Missing data | N (%) | Mean (SD) | Missing data | N | Mean (SD) | Missing data |
| LDL (mg/dL) | 606 (83.6%) | 127.2 (32.93) | 8 | 119 (16.4%) | 118.7 (36.16) | 2 | 725 | 125.8 (33.60) | 10 |

N, number

SD, standard deviation

LDL, low-density lipoprotein cholesterol

Score：3

We reverse code the two different outcomes (survived at least 5 years or not; died within 5 years or not). As a result, we get essentially the same answers. They have been condensed to reflect the differences between those that survived at least 5 year and died within 5 years.

There was a total of 735 subject, however, there was missing LDL values for 10 of those subjects. These subjected were excluded from further analysis. Of the 725 evaluable, there were 606 (83.6%) who survived at least 5 years and 119 (16.4%) who died within 5 years.

The mean serum LDLs for those who survived at least 5 years and did not survive within 5 years were 127.20 mg/dL and 118.70 mg/dL, respectively. The standard deviations for those who survived at least 5 years and did not survive within 5 years were 32.9 mg/dL and 36.16 mg/L, respectively.

The mean serum LDLs for those who died within 5 years and did not die within 5 years were 118.70 mg/dL and 127.20 mg/dL, respectively. The standard deviations for those who died within 5 years and did not die within 5 years were 36.16 mg/L and 32.9 mg/dL, respectively.

The values were exactly the same for those who survived at least 5 year and died within 5 years for sample size, sample mean, and same standard deviations.

The average LDL is larger for those subjects who survived at least 5 years compared to those who died within 5 years. The sample means appears to be different in magnitude; the difference in average LDL between those who survived at least 5 years and died within 5 years was 8.50 mg/dL (lower for those who died within 5 years). This appears to be a meaningful difference between both groups.

The standard deviation is smaller for the group that survived at least 5 years compared to those that died within 5 years. The smaller standard deviation is likely due to the larger sample size of the group that survived at least 5 years. The larger standard deviation for the group that died within 5 years is likely due to the smaller sample size. Despite this, the standard deviation comparison is not too different.

(b)

|  |  |  |  |
| --- | --- | --- | --- |
|   | Survived at least 5 yearsN=606 (83.6%) | Died within 5 yearsN=119 (16.4%) | All subjectsN=725 |
| Variable | Point estimate | SE | 95% CI | Point estimate | SE | 95% CI | Point estimate | SE | 95% CI |
| LDL (mg/dL) | 127.2 | 1.34 | 124.57, 129.83 | 118.7 | 3.31 | 112.13, 125.26 | 125.8 | 1.25 | 123.35, 128.25 |

SE, standard error

CI, confidence interval

LDL, low-density lipoprotein

We being by establishing that the point estimate denotes the mean.

There was a total of 735 subject, however, there were missing LDL values for 10 of those subjects. These subjected were excluded from further analysis. Of the 725 evaluable, there were 606 (83.6%) who survived at least 5 years and 119 (16.4%) who died within 5 years.

The point estimate (mean serum LDL level) for those who survived at least 5 years and did not survive at least 5 years were 127.20 and 118.70 mg/dL, respectively. Based on a 95% confidence interval computed with an allowance for unequal variances, the observed point estimate for those subjects who survived at least 5 years would not be judged unusual if the true point estimate of those subjected who died within 5 year was anywhere between 124.57 to 129.83 mg/dL.

The point estimate LDL for those who died within 5 years and did not die within 5 years were 118.70 mg/dL and 127.20, respectively. Based on a 95% confidence interval computed with an allowance for unequal variances, the observed point estimate for those subjects who died within 5 years would not be judged unusual if the true point estimate of those subjected who died within 5 year was anywhere between 112.13 mg/dL to 125.26 mg/dL.

The standard error for those who survived at least 5 years and did not survive at least 5 years were 1.34 and 3.31 mg/L, respectively.

The standard error for those who died within 5 years and did not die within 5 years were 3.31 mg/L and 1.34, respectively.

The point estimates, standard errors, and 95% confidence intervals for the two different outcomes (survive at least 5 years and died within 5 years) are exactly the same because they are the reverse code of each other.

The point estimate is larger for those subjects who survived at least 5 years compared to those who died within 5 years. The difference in point estimate between those who survived at least 5 years and died within 5 years was 8.50 mg/dL (lower for those who survived at least 5 years). Based on a 95% confidence interval computed with an allowance for unequal variances, this observed tendency of 8.50 mg/dL lower point estimate among subjects dying within 5 years would not be judged unusual if the true difference population point estimates were anywhere between a 1.44 mg/dL to 15.56 mg/dL lower point estimate among subjected who died within 5 years. This appears to be a meaningful difference between both groups.

The standard error is smaller for the group that survived at least 5 years compared to those that died within 5 years (1.34 versus 3.31 mg/dL, respectively). The smaller standard error is likely due to the larger sample size of the group that survived at least 5 years. The larger standard error for the group that died within 5 years is likely due to the smaller sample size.

The point estimates are essentially the same since they are the average LDLs for the two groups. The SEs are smaller than the SDs for both groups which is reasonable since the SE is calculated as the SD/$\sqrt{n}$. Since the sample size is smaller for the subjects who died within 5 years, the SE is larger compared to the subjects who survived at least 5 years.

For both the standard deviation and standard error, the subjects who died within 5 years have higher values compared to subjects who survived at least 5 years. This is likely due to the smaller sample size of the subjects who died within 5 years.

 Score：3

(c)

The 95% confidence interval for average LDL in subjects who died within 5 years was between 112.13 and 125.26 mg/dL. This overlaps with the 95% CI of the average LDL in subjected who survive at least 5 years (95% CI: 124.57, 129.83 mg/dL).

The overlap in confidence intervals may lead one to believe that there is no statistically significant difference in mean LDL between the two groups.

However, at the 5% significance level, there is a statistically significant difference in mean LDL between those who survived at least 5 years and those who died within 5 years (P-value = 0.0186). In other words, there is a 1.86% probability of obtaining the result equal to or more extreme than what was actually observed given that the null hypothesis is true.

Score：3

(d)

In order to combine the two SDs of LDL measurements from each group we will need to use the following formula:

$$SD\_{combined}^{2}=\frac{SD\_{1}^{2}(n\_{1}-1)+SD\_{2}^{2}(n\_{1}-1)}{\left(n\_{1}-1\right)+(n\_{2}-1)}$$

Therefore, the combined standard deviation is:

$$SD\_{combined}^{2}=\frac{32.93^{2}(606-1)+36.16^{2}(119-1)}{\left(606-1\right)+(119-1)}$$

SD = 44.73 mg/dL （33.48mg/dL）

Score：2

(e)

Method:

Mean serum LDL levels (which was defined as the point estimate) were compared between subjects who died within 5 years and survived at least 5 years. Differences in mean were tested using a t-test without assuming equal variances (Satterwaithe’s approximation). The 95% confidence intervals for the differences in population means were based on the same handling of variances.

Result:

Based on a 95% confidence interval computed with an allowance for unequal variances, this observed tendency of 8.50 mg/dL lower mean serum LDL among subjects dying within 5 years would not be judged unusual if the true difference population means were anywhere between a 1.44 to 15.56 mg/dL lower mean LDL among subjects who died within 5 years. Performing a t-test that does not assume equal variances, the two-tailed, P-value is 0.0186 (t-statistic = 2.3783, df = 158.746) using Satterwaithe’s degrees of freedom. At the 5% significance level, we can with high confidence reject the null hypothesis that the mean serum LDL levels are not different among subjects who died within 5 years and survive at least 5 years in favor of an alternative hypothesis that death within 5 years is associated with a lower mean serum LDL.

Score:3

Question 2. Score: 18.5

The simple linear regression models (presuming homoscedasticity) are expressed as:

Model A:

$$E\left(Died within 5 years\right)=β\_{0}+β\_{1}(Death at 5 years)$$

Model B:

$$E\left(Survived at least 5 years\right)=β\_{0}+β\_{1}(Survived at least 5 years)$$

The responder variable is vital status at 5 years and the outcome variable is LDL.

Our scientific question is whether or not there is an association between LDL level and death with 5 years or survived at least 5 years. However, we are testing the statistical question with Models A and B; does the distribution of LDL differ across vital status groups?

In Model A, there are two parameters $(β\_{0} and β\_{1})$ and one predictor variable (dead within 5 years) with two values (No=0, Yes=1), which indicates that the model is saturated by definition. Therefore the intercept $(β\_{0})$ is the mean serum LDL for patients who did not survived at least 5 years (118.70 mg/dL); and the intercept plus the slope $(β\_{0} and β\_{1})$ is the mean serum LDL for patients who survived at least 5 years (127.20 mg/dL).

In Model B, there are two parameters $(β\_{0} and β\_{1})$ and one predictor variable (survived at least 5 years) with two values (No=0, Yes=1), which indicates that the model is saturated by definition. Therefore the intercept is the mean serum LDL for patients who did not die within 5 years (127.20 mg/dL); and the intercept plus the slope $(β\_{0} and β\_{1})$ is the mean serum LDL for patients who died within 5 years (118.70 mg/dL).

Score: 2

(b)

In Model B, there are two parameters $(β\_{0} and β\_{1})$ and one predictor variable (survived at least 5 years) with two values (No=0, Yes=1), which indicates that the model is saturated by definition. The intercept plus the slope $(β\_{0} and β\_{1})$ from Model B is the mean serum LDL for patients who survived at least 5 years (127.20 mg/dL).

$$E\left(Survived at least 5 years\right)=118.70+8.50\left(Survived at least 5 years\right)$$

(units for LDL are mg/dL)

These values are exactly the same as those estimates calculated from Problem 1. Because the model is saturated, we are getting the exact values in descriptive statistics.

Score:2

(c)

Realizing that Model B (reverse coded from Model A in the predictor of interest) answers the question with just the intercept $(β\_{0})$, the confidence is calculated using ordinary least squares regression method presuming homoscedasticity. Based on a 95% confidence interval using classical linear regression method, the mean serum LDL for those who survived at least 5 years would not be judged unusual if the true population mean were anywhere between 124.53 to 129.86 mg/dL.

This is different from the confidence interval calculated in Problem 1, which was 124.57 and 129.83 mg/dL. Although close, these are slightly different. One possible explanation is the use of least square methods to calculate the regression confidence limits. Least squares method is efficient when the data is normally distributed and homoscedastic. There may be some non-linearity, which is giving us a different answer from Problem 1.

Score: 1.5

Note that the preponderance of the data contributing to the pooled SD comes from the survivors, so the pooled SD is closer to the sample SD for the survivors than it is to the sample SD for the subjects who died

(d)

We can use Model B (which is the reverse coded predictor of interest of Model A) and the parameter intercept $(β\_{0})$ to answer this question. Therefore the intercept $(β\_{0})$ is the mean serum LDL for patients who did not survived at least 5 years or who died within 5 years (118.70 mg/dL). This is exactly the same mean serum LDL for patients who died within 5 years after enrollment from Problem 1. Because the model is saturated, we get the exact fit of the group mean.

Score: 2

(e)

We can use Model B (which is the reverse coded predictor of interest of Model A) and the parameter intercept $(β\_{0})$ to answer this question. Based on a 95% confidence interval using ordinary least squares regression method presuming homoscedasticity, the mean serum LDL for those who died within 5 years would not be judged unusual if the true population mean were anywhere between 112.67 to 124.72 mg/dL.

This is different from the confidence interval calculated in Problem 1, which was 112.13 and 125.26 mg/dL. Although close, these are slightly different. One possible explanation is the use of least square methods to calculate the regression confidence limits. Least squares method is efficient when the data is normally distributed and homoscedastic. There may be some non-linearity, which is giving us a different answer from Problem 1.

Score:2

(f)

Assuming equal variances are equal in the two populations for Model A, the regression-based estimate of the standard deviation is the root mean squares of the error (RMSE). Therefore, the SD for Model A is 33.48 mg/dL.

Assuming equal variances are equal in the two populations for Model B, the regression-based estimate of the standard deviation is the root mean squares of the error (RMSE). Therefore, the SD for Model A is 33.48 mg/dL.

Both regression-based estimate of the standard deviations are exactly the same.

Compared to Problem 1, there are differences in the standard deviations. From part Problem 1(d), we calculated the SD to be 44.73 mg/dL. However, this is higher compared to the regression-based estimate from the simple linear regression model which was 33.48 mg/dL. The regression-based estimate is much smaller compared to estimates from pooled standard deviation.

Score: 1

The regression based estimate of the presumed common variance in the two groups is exactly the SD derived from the pooled variance estimate in problem1.

(g)

Model A and B are exactly the same with one exception; the predictor of interest is a reverse coded version of other. If the predictor of interest is death with 5 years, then the coding is defined as 0=No and 1=Yes. If the predictor of interest is survived at least 5 years, then the coding is defined as 0=No and 1=Yes. When these variables are examined, they are essentially the reverse coding of the other. One may also state that Model A is the inverse of Model B.

Score:2

(h)

In Model A, the predictor of interest is death within 5 years and the outcome is serum LDL. The intercept is 127.20 mg/dL which is the mean serum LDL for patients who did not die within 5 years. Based on a 95% confidence interval, it would not be unusual for the true population mean for patients who did not die within 5 years to be between 124.52 and 129.87 mg/dL. The P-value is <0.001 which means that we do not have enough evidence that the association between LDL and death within 5 years are the same; therefore, we reject the null hypothesis in favor for the alternative that there is an association between LDL and death within 5 years.

Score:2

(i)

In Model A, the predictor of interest is death within 5 years and the outcome is serum LDL. The slope is -8.50 mg/dL, which is the change in outcome for a patient who died within 5 years relative to a patient who did not die within 5 years. For a patient who died within 5 years, their mean serum LDL level is decreased on average by 8.50 mg/dL relative to a person that did not die within 5 years.

Based on a 95% confidence interval, it is not unusual for the true population LDL response for a person who dies within 5 years versus not was between -15.09 and -1.91 from a person who did not die within 5 years. The P-value was 0.012 which means that we do not have enough evidence that there is no association between serum LDL and being dead within 5 years; therefore, we reject the null hypothesis in favor of one where there is an association between being dead within 5 years and serum LDL level.

Score: 2

(j)

For Model A where the outcome variable is death within 5 years, we estimated that the average mean serum LDL was lower by 8.50 mg/dL for those who died within 5 years compared to those who did not die within 5 years. The 95% confidence interval suggests that this observation is not unusual if the true difference in mean LDL lowering for a patient who died within 5 years were between a 1.91 and 15.09 mg/dL decrease. Because the P-value is P=0.012, we reject the null hypothesis that there is no linear trend in the average LDL across those who died within 5 years and those who didn’t die within 5 years.

For Model B where the outcome variable is survive at least 5 years, we estimated that the average mean serum LDL was higher by 8.50 mg/dL for those who survived at least 5 years compared to those who did not survive at least 5 years. The 95% confidence interval suggests that this observation is not unusual if the true difference in mean LDL lowering for a patient who died within 5 years were between a 1.91 and 15.09 mg/dL increase. Because the P-value is P=0.012, we reject the null hypothesis that there is no linear trend in the average LDL across those who survived at least 5 years and those who didn’t survive at least 5 years.

In Problem 1, based on a 95% confidence interval computed with an allowance for equal variances, this observed tendency of 8.50 mg/dL lower mean serum LDL among subjects dying within 5 years would not be judged unusual if the true difference population means were anywhere between a 1.91 to 15.09 mg/dL lower mean LDL among subjects who died within 5 years. Performing a t-test that does assume equal variances, the two-tailed, P-value is 0.0115 (t-statistic = 2.5324, df = 723). At the 5% significance level, we can with high confidence reject the null hypothesis that the mean serum LDL levels are not different among subjects who died within 5 years and survive at least 5 years in favor of an alternative hypothesis that death within 5 years is associated with a lower mean serum LDL.

Compared to Problem 1, the results for the point estimate (mean difference in LDL across vital status at 5 years) and 95% confidence intervals are very close. Because we performed a t-test without the assumption of equal variances, the 95% confidence intervals do not match exactly. If the t-test was performed assuming equal variances, the 95% confidence intervals were precisely the same as the results from the ordinary least squares regression presuming homoscedasticity. In other words, the conclusions are the same with the t-test of equal variances and the simple linear regression with a saturated model.

Score:2

Question 3.

We are performing a t-test without assuming equal variances to compare the mean LDL values between patients who died within 5 years and did not die within 5 years. Differences in mean were tested using a t-test that did not presume equality of variances. The 95% confidence interval for the difference in population means was based on the same methods for handling variances.

The t-test without presumption of equal variances yielded a mean difference of 8.50 mg/dL between patients who did not die within 5 years and died within 5 years (lower in patients who died within 5 years). Based on the 95% confidence interval the mean difference would not be unusual if the true population mean difference was between 1.44 and 15.56 mg/dL lower for those who died within 5 years relative to those who did not die within 5 years. The t-statistic was 2.3783 and the degrees of freedom (Satterwaithe) was 158.746. The two-side P-value was 0.0186.

Compared to the results of Problem 1 there were differences between the t-test with equal variances presumed and not presumed. For t-test with equal variances presumed, the t-statistic was slightly higher (t-statistic=2.5324) and the degrees of freedom was much higher (df=723). The two-side P-value was 0.0155, which was slightly lower than the results from not presuming unequal variances. The P-value from the t-test that does not presume equal variances was more conservative compared to that of equal variances.

However, the conclusions did not differ. Based on the p-value calculated without presumption of equal variances, the P-value was 0.0186. Based on the P-value, we reject the null in favor of a hypothesis that there is on average a 8.50 mg/dL lower mean serum LDL level in patients who died within 5 years relative to those who did not die within 5 years.

 Score: 10

Question 4.score:10

We performed a simple linear regression that does not presume equality of variances using the Huber-White sandwich estimator. We investigated the difference in mean LDL between patients who died and did not die within 5 years using a simple linear regression method that does not presume equal variances. The 95% confidence interval The 95% confidence interval for the difference in population means were based on the same methods for handling variances.

The regression results not presuming equal variance yielded similar intercept and slope $(β\_{0} and β\_{1})$. However, the standard errors are slightly different. For the intercept, the standard error was 3.305116, which is higher compared to the result of the model assuming equal variances, which was 3.068836. Similarly, the standard error for the slope for the model that does not presume equal variance was higher compared to the model that presumed equal variance (3.565821 versus 3.356652, respectively).

The t-statistic for the intercept is also different with the model that presumed equal variances higher than the one that does not presume equal variances (38.68 versus 35.91, respectively). Similarly, the t-statistic for the slope is also different with the model that presumed equal variances and was higher than the one that does not presume equal variance (2.53 versus 2.38, respectively).

The confidence interval for the intercept was wider for the model that did not presume equal variances (95% CI: 112.21, 125.19) compared to the model that presumed equal variances (95% CI: 112.67, 124.72). Similarly, the confidence interval for the slope was wider for the model that did not presume equal variances (95% CI: 1.50, 15.50) compared to the model that presumed equal variances (95% CI: 1.91, 15.09).

Question 5.

(a) score:5

Methods:

The outcome is LDL in mg/dL and the predictor of interest is Age in years. Both variables will be evaluated as continuous variables for the scatterplot. We will evaluate average LDL across Age groups, which will be categorized into 5-year increments (65 to less than 70 years, 70 to less than 75 years, 75 to less than 80 years, and 80-years +).

We started by visually inspecting the correlation between LDL and Age using a scatterplot with LDL on the y-axis and Age on the x-axis. Differences in mean age was compared between Age groups using simple linear regression without presumption of equal variances. The parameter estimate and 95% CI was computed using simple linear regression with the Huber-White sandwich estimator of the standard errors.

For the tabular descriptive analysis: indicator variables were created for Age (1=65 to less than 70 years, 2=70 to less than 75 years, 3=75 to less than 80 years, and 4=80 years or greater) and Sex category (0=female and 1=male).

Tabular descriptive statistics were presented within groups defined by Age categories, and for the entire sample. Within each group defined by Age category, for continuous variable (LDL) we included the mean, standard deviation, minimum and maximum. For binary variable (male), we presented frequency and percentages.

For the stratified groups (by sex), we included mean and standard deviation for continuous variables. For binary variables we presented frequency and percentages.

Results:

Data was available for 735 subjects; however, 10 of those subjects are missing data on serum LDL. Those subjects were omitted from the final analysis. We were unable to predict the impact of these missing data on the study’s eventual conclusions. None of the 735 subjects were missing data on any of the other descriptive variables of interest.

Of the 725 subjects with available measurements, patients in the 80-year old or older age category had the lower LDL relative to the other groups. The proportion of males appear similar across all age categories ranging from 47.37% to 51.61%).

Table 1. Descriptive analysis of LDL and male proportion across all age categories.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Age category | 65 to less than 70 | 70 to less than 75 | 75 to less than 80 | 80 + | Total |
| N, number (%) | 114 (15.72) | 303 (41.79) | 184 (25.38) | 124 (17.10) | 725 (100) |
| LDL (mg/dL)1 | 127.70 (32.40) | 125.32 (32.50; 37-247) | 126.85 (35.46; 11-225) | 123.68 (34.73; 52-227) | 125.80 (33.60; 11-247) |
| Male, number (%) | 54 (47.37) | 150 (49.50) | 92 (50.00) | 64 (51.61) | 360 (49.66) |
| 1 Descriptive statistics presented are the mean (standard deviation; minimum-maximum) |

Among the 10 missing serum LDL data, there were 3 in the 65 to less than 70 year age category, 2 in the 70 to less than 75 year age category, 3 in the 75 to less than 80 year age category, and 2 in the 80 year or greater age category. Six were male and 4 were female. Missing data may bias our results since the proportion is higher among males with missing data.

Table 2. Descriptive analysis of missing data.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Missing data | 65 to less than 70 years | 70 to less than 75 years | 75 to less than 80 years | 80 + years | Total |
| N, number (%) | 3 (30) | 2 (20) | 3 (30) | 2 (20) | 10 |
| Male, number (%) | 2 (33.3) | 1 (16.7) | 2 (33.3) | 1 (16.7) | 6 |
| Female, number (%) | 1 (25) | 1 (25) | 1 (25) | 1 (25) | 4 |

In the stratified analyses, males who were 80 years old and older had a lower average serum LDL level compared to females who were 80 years old and older (116.77 versus 131.05 mg/dL, respectively). Males who were 75 to less than 80 years old had a lower average serum LDL level compared to females who were 75 to less than 80 years old (120.18 versus 133.52 mg/dL, respectively). Males who were 70 to less than 75 years old had a lower average serum LDL level compared to females who were 70 to less than 75 years old (119.63 versus 130.90 mg/dL, respectively). Males who were 65 to less than 70 years old had slightly higher average serum LDL levels compared to females who were 65 to less than 70 years old (128.50 versus 126.98, respectively). In total, males had lower average serum LDL levels compared to females (120.59 versus 130.94 mg/dL, respectively)

Although males have a lower average serum LDL level compared to females; the proportion of males and females were roughly the same for all age categories. The lower average LDL level for males relative to females indicate some potential for confounding. By the definition of confounding, it’s possible that sex may affect LDL level; but it is unclear whether or not sex affects age. Since we are interested in the Age to LDL relationship, sex may not be a confounder by definition because it has no effect on Age.

Table 3. Descriptive analysis of mean LDL across age groups stratified by sex.

|  |  |
| --- | --- |
|   | Male |
|  Age category | 65 to less than 70 | 70 to less than 75 | 75 to less than 80 | 80 + | Total |
| N | 54 (15.00) | 150 (41.67) | 92 (25.56) | 64 (17.78) | 360 (100) |
| LDL (mg/dL)1 | 128.50 (30.84) | 119.63 (31.34) | 120.18 (32.11) | 116.77 (34.73) | 120.59 (32.15) |
|   | Female |
|  Age category | 65 to less than 70 | 70 to less than 75 | 75 to less than 80 | 80 + | Total |
| N | 60 (16.44) | 153 (41.92) | 92 (25.21) | 60 (16.44) | 365 (100) |
| LDL (mg/dL)1 | 126.98 (33.98) | 130.90 (32.76) | 133.52 (37.53) | 131.05 (34.25) | 130.94 (34.25) |

1 Descriptive statistics presented as mean and standard deviation.

In Figure 1, we plot the LDL on the y-axis and Age on the x-axis with a best-fit straight line and a loess line along with some jitter. The relationship appears elastic across the Age, but the LDL levels remains constant. Although there were several outliers in the age range of 90 to 100, this had the effect of stabilizing the best-fit line since equal number of outliers were above the line.

Figure 1. Scatter plot of the age and serum LDL level relationship.



Figure 2 illustrates the same relationship between Age and LDL stratified by sex. There does not appear to be too much difference between the sexes; however, males tended to show some movement on the loess line to go upward. This is on the extreme range and my not be very helpful. The best-fit line appears to slope downward for males compared to females.

Figure 2. Scatter plot of the age and serum LDL level relationship stratified by sex.



Based on the stratified tabulated and graphical relationship between Age and serum LDL, there does not appear to be any confounding by sex on the Age to LDL relationship.

(b) score:3

Methods:

We compared the linear association between Age as a predictor variable on a continuous scale with serum LDL level as the outcome variable on a continuous scale. Least squares estimation was used to estimate the parameter coefficients and intercept. Difference in means (slope) LDL between patients differing by 1 year in age was estimated using least square estimation without presuming homoscedasticity. We did not presume equal variances when fitting the regression model. The 95% confidence interval for the parameter coefficients $(β\_{0} and β\_{1})$ were computed using least square estimation with the Huber-White sandwich estimator of the standard errors.

(c) score 3

In simple linear regression model, there are two parameters $(β\_{0} and β\_{1})$ and one predictor variable (Age) as a continuous scale, which indicates that the model is not saturated by definition.

$$E\left(Age\right)=132.53-0.09\left(Age\right)$$

 (units for Age are year and LDL are mg/dL)

(d) score:3

The estimated mean LDL for 70-year old subjects is:

$$E\left(Age\right)=132.53-0.09\left(70\right)=126.23 mg/dL$$

(e) score 3

The estimated mean LDL for 71-year old subjects is:

$$E\left(Age\right)=132.53-0.09\left(71\right)=126.14 mg/dL$$

As the patient population gets older by 1 year, there is an incremental decrease in the mean LDL for the population by a rate of 0.09 mg/dL per year. Therefore, when the patient population is older by 1 year (say, 70 to 71 years), we observe a drop in average serum LDL level by approximately 0.09 mg/dL. The slope from part (c) is the rate at which this incremental decrease in LDL level occurs with every increase in age by 1 unit or year.

If the model were saturated, the estimates would estimate the sample descriptive statistics perfectly. In the case of sex, a binary variable, this would mean that all women would have the same average LDL and all men would have the same average LDL. Because our model is not saturated, there are variations across different Ages. The slope, however, remains constant, and we can simply multiple the Age with the slope to see how the one unit change in Age would affect the LDL level.

(f) score 3

The estimated mean LDL for 75-year old subjects is:

$$E\left(Age\right)=132.53-0.09\left(75\right)=125.78 mg/dL$$

As the patient population gets older by 1 year, there is an incremental decrease in the mean LDL for the population by a rate of 0.09 mg/dL per year. Therefore, when the patient population is older by 1 year (say, 70 to 71 years), we observe a drop in average serum LDL level by approximately 0.09 mg/dL. The slope is the rate at which this incremental decrease in LDL level occurs with every increase in age by 1 unit or year. If the patient were 75 years old, the LDL would be 125.78 mg/dL. The difference in average LDL between a 75 year old and a 70 year old is the product of the slope and the difference in years, 5 years (0.09 X 5 = 0.45 mg/dL). Therefore, a 75-year old person will have, on average, a reduction in serum LDL level of 0.45 mg/dL relative to a 70-year old person.

If the model were saturated, the estimates would estimate the sample descriptive statistics perfectly. In the case of sex, a binary variable, this would mean that all women would have the same average LDL and all men would have the same average LDL. Because our model is not saturated, there are variations across different Ages. The slope, however, remains constant, and we can simply multiple the Age with the slope to see how the one unit change in Age would affect the LDL level.

(g) score 3

RMSE is the sample standard deviation of the residuals. We use the RMSE instead of the SD from the sample to calculate our confidence intervals for the parameter estimates. RMSE measures the differences between values predicted by the model and values observed from the model. It borrows information to get within-group differences.

(h) score: 3

The intercept is the Y-intercept (or estimated mean of the population) when all the other variables in the model have no influence on the parameter estimates. In other words, when all the variables are 0 or if the slopes are 0. In our model, the intercept is the estimated mean when Age = 0 or if the slope is 0.

The intercept is commonly not scientifically interesting because it may be outside the range of the data or what is considered plausible. In our simple linear regression model, the intercept is 132.53 mg/dL, which is not implausible. But given that the average LDL of the total is 125.80 mg/dL, the intercept is on the high end and possibly outside the reasonable range.

(i) score 2

The slope is estimate difference in mean serum LDL level between patients differing by 1 year in age. For instance, and increase in 1 year is associated with an average decrease in mean serum LDL level by 0.09 mg/dL.

(j) score 3

Method:

We tested the association between the predictor of interest (Age as a continuous variable) and the outcome of interest (serum LDL as a continuous variable) using simple linear regression model. We tested whether or not there is a linear trend in the average LDL across age groups using ordinary least squares regression. An association will exist if the slope $(β\_{1})$ is nonzero; in other words, the average LDL would be different across age groups. Parameter estimates will be estimated using least squares method allowing for heteroscedasticidity. The 95% confidence intervals was computed using least square estimation with the Huber-White sandwich estimator of standard errors.

Results:

From linear regression analysis, we estimate that for each year increase in age, the difference in mean LDL is lower by 0.09 mg/dL. A 95 % confidence interval suggests that this observation is not unusual if the true lowering in mean LDL per year difference in age were between -0.37 and 0.55 mg/dL. Because the P=0.698, we do not have enough evidence to reject that null hypothesis that there is no linear trend in the average LDL across age groups.

(k) score 3

We calculate the confidence interval for the difference in mean LDL across groups that differ by 5 years of age by using the following expression:

$$\left(Age\right)=132.53-0.09\left(70\right)=126.23 mg/dL$$

$$ \left(Age\right)=132.53-0.09\left(75\right)=125.78 mg/dL$$

Subtracting the difference will yield a mean difference of 0.45 mg/dL differing by 5 years in age. The 95% confidence interval for the slope is exactly the confidence interval for the difference in means for classical linear regression. Therefore, using the classical linear regression presuming equal variances, we can estimate the 95% confidence interval.

The 95% confidence interval for the slope in the simple linear regression presuming equal variances is -0.54, 0.36; and it is almost similar to the one yielded by the model not presuming equal variances.

Since the estimate difference is 5 years, the 95% confidence interval is -0.54\*5 for the lower limit and 0.36\*5 for the upper limit: -2.7, 1.8.

(l) score 3

Correlation between LDL (outcome variable, continuous) and Age (predictor of interest, continuous) was performed using Pearson correlation. We evaluated there was a significant correlation between LDL and Age using a two-sided significance level of 5%. LDL level is weakly negatively correlated with age. The Pearson correlation coefficient was -0.0146 indicating that as a person increased in age, the LDL drops. The P-value was 0.6944.

The P-value for the nonzero correlation is very close to the test for slope generated by the simple regression model, which was P=0.6984.

In both cases, the conclusions were similar. There was no statistically significant correlation between LDL and Age at the 5% significance level.