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.
	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?

|  |  |  |  |
| --- | --- | --- | --- |
| **Vital Status at 5 Years** | **Observations** | **Sample Mean Serum LDL****(mg/dL)** | **Sample SD of Serum LDL****(mg/dL)** |
| **Deceased** | 119 | 118.70 | 36.16 |
| **Alive** | 606 | 127.20 | 32.93 |
| **All Observations** | 725 | 125.80 | 33.60 |

**The sample mean serum LDL values for groups defined by vital status are similar in magnitude. The sample means differ by less than 10 mg/dL, approximately 7%. Additionally, both sample means are included within the 100-129 mg/dL range which is classified as “Near Ideal” by the Mayo Clinic indicating that this difference is not likely to be clinically relevant. The sample standard deviations across groups defined by vital status are also similar in magnitude. However, they show a slightly larger difference of approximately 10%.**

* 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? Explain any differences in your answer about the estimates and estimated SEs compared to your answer about the sample means and sample standard deviations.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Vital Status at 5 Years** | **Observations** | **Estimate of Population Mean Serum LDL****(mg/dL)** | **Standard Error of****Estimate****(mg/dL)** | **95%****Confidence Interval** |
| **Deceased** | 119 | 118.70 | 3.31 | (112.13, 125.26) |
| **Alive** | 606 | 127.20 | 1.34 | (124.57, 129.83) |

**The sample mean serum LDL value is the used as the point estimate for the true mean serum LDL levels found in population groups defined by vital status at 5 years. Therefore, the true mean estimates are still similar in magnitude with a difference of less than 10 mg/dL, approximately 7%. This is also not likely to be clinically relevant as both estimates fall within the same range of serum LDL levels (100-129 mg/dL) as defined by the Mayo Clinic. However, the standard errors for the point estimates are not similar in magnitude. Because the standard error is based on the sample standard deviation and the number of observations the differences in the standard errors reflect the differences in sample size seen between the number of participants who survived five years and those who did not. Therefore, even though the sample standard deviations were similar in magnitude the standard errors of the point estimates for true mean serum LDL are not. There were roughly six fold more participants who were alive at 5 years as participants who died within the 5 year period, because of this large difference in sample size the standard error for the point estimate of mean serum LDL in individuals who survived at least 5 years is significantly smaller.**

* 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?

**The 95% CI for a population surviving 5 years and the 95% CI for a population dying within 5 years overlap by less than 1 mg/dL. Based on this observation the difference in the point estimates for the population mean serum LDL levels for groups defined by vital status at five years are unlikely to be significantly different at a 0.05 level. However, the size of the CI is dependent of the number of observations so the overlap could represent a lack of precision rather than a lack of significance.**

* 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?)

**The best estimate for the standard deviation of LDL measurements in each group, if the variances are presumed to be equal, is calculated using the following equation.**

$$SE\_{equal}= \sqrt{\frac{\left(n\_{1}-1\right)s\_{1}^{2}+(n\_{2}-1)s\_{2}^{2}}{n\_{1}+n\_{2}-2}}$$

**For this population the estimate for the pooled standard deviation is 33.48 mg/dL.**

* 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 serum LDL levels between populations defined by vital status at 5 years is 8.50 mg/dL higher for a population that survives at least 5 years. The standard error of this estimate is given as 3.36 mg/dL. A 95% confidence interval suggests that this observation is not unusual if the true difference in mean serum LDL for populations defined by vital status at 5 years was between 1.91 and 15.09 mg/dL higher for a population surviving at least 5 years. Because the p-value is less than 0.05 (p = 0.0115) we reject the null hypothesis that there is no difference in mean serum LDL levels across populations defined by vital status at 5 years.**

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.

**In both models there are two possible groups defined by vital status at 5 years, it is only the indicator values that differ. A model is considered saturated when the number of groups is equivalent to the number of parameters. For a linear regression there are two parameters that must be fit, β0 and β1, therefore, because both models have two vital status groups and two parameters they are both 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?

**Using Model A the estimate for the true mean LDL among a population of subjects who survive at least 5 years is given by the intercept of the model and is 127.20 mg/dL. This is equivalent to the estimate found using the two-sample t-test assuming equal variances.**

* 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.

**Using Model A the 95% confidence interval for the true mean LDL level among a population of subjects who survive at least 5 years is given as 124.53 mg/dL to 129.87 mg/dL. This is equivalent up to the second decimal place to the 95% confidence interval obtained from the two-sample t-test assuming equal variances. This slight difference is a result of how the confidence intervals are calculated. Unlike the t-test, linear regression models use the pooled standard deviation for determination of the 95% 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?

**Using Model B the estimate for the true mean LDL among a population of subjects who die within 5 years is given by the intercept of the model and is 118.70 mg/dL. This is equivalent to the estimate found using the two-sample t-test assuming equal variances.**

* 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.

**Using Model B the 95% confidence interval for the true mean LDL level among a population of subjects who survive at least 5 years is given as 112.67 mg/dL to 124.72 mg/dL. This differs from the 95% confidence interval obtained from the two-sample t-test assuming equal variances. This difference is a result of how the confidence intervals are calculated. Unlike the t-test, linear regression models use the pooled standard deviation for determination of the 95% CI. Because there were fewer observations for subjects who died within five years the pooled standard deviation is less than the sample standard deviation resulting in a narrower confidence interval.**

* 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?

**By assuming that the variances are equal in the two populations the regression based estimate of the standard deviation borrows information across groups and uses the pooled standard deviation as the estimate for both populations. For both models the estimate of the standard deviation is given as 33.48 mg/dL. This is not equal to either standard deviation estimate found using the t-test. It is close to but slightly larger than the standard deviation estimate for a population of subjects surviving at least 5 years and less than the standard deviation estimate for a population of subjects who die within 5 years. This is because there are more subjects observed to survive for at least 5 years so the pooled standard deviation estimate is weighted towards this value. However, because both standard deviation estimates are of a similar magnitude the pooled standard deviation is also of a similar magnitude.**

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

**Models A and B are related by their intercepts and slope. The slope of Model A is the negative of the slope for Model B and vice versa. Additionally the intercept of Model A is equal to the intercept of Model B plus the slope of Model B. This relationship is also seen for Model B, where the slope of Model A plus the intercept of Model A is equal to the intercept of Model B.**

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

**The estimated difference in mean serum LDL for two groups differing in vital status at 5 years is 8.50 mg/dL lower for a population of subjects who die within 5 years.**

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

**The estimated difference in mean serum LDL for two groups differing in vital status at 5 years is 8.50 mg/dL higher for a population of subjects surviving at least 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?

**The point estimate for the true difference in mean serum LDL levels between populations defined by vital status at 5 years is 8.50 mg/dL higher for a population that survives at least 5 years. The standard error for this estimate is given as 3.36 mg/dL. A 95% confidence interval suggests that this observation is not unusual if the true difference in mean serum LDL for populations defined by vital status at 5 years was between 1.91 and 15.09 mg/dL higher for a population surviving at least 5 years. Because the p-value is less than 0.05 (p = 0.012) we reject the null hypothesis that there is no difference in mean serum LDL levels across populations defined by vital status at 5 years. This is equivalent to the inference made using the two-sample t-test assuming equal variances.**

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.)

|  |  |  |
| --- | --- | --- |
|  | **Two-Sample t-test Assuming Equal Variance** | **Two-Sample t-test Allowing for Unequal Variance** |
| **Vital Status at 5 years** | **Deceased** | **Alive** | **Deceased** | **Alive** |
| **Observations** | 119 | 606 | 119 | 606 |
| **Estimate of Population Mean Serum LDL (mg/dL)** | 118.70 | 127.20 | 118.70 | 127.20 |
| **Standard Error of Estimate (mg/dL)** | 3.31 | 1.34 | 3.31 | 1.34 |
| **95% Confidence Interval** | (112.13, 125.26) | (124.57, 129.83) | (112.13, 125.26) | (124.57, 129.83) |
| **Estimate of True Difference in Population Means\*****(mg/dL)** | 8.50 | 8.50 |
| **Standard Error of Estimate****(mg/dL)** | 3.36 | 3.57 |
| **95% Confidence Interval** | (1.91, 15.09) | (1.44, 15.56) |
| **p-value** | 0.0115 | 0.0186 |

**\*meansurvived – meandeceased**

**Values in red are not equivalent between the two tests.**

**The descriptive statistics for each group (point estimate of mean, standard error, and 95% CI) are equivalent to those found using the two-sample t-test assuming equal variance. The point estimate for the true difference in population mean serum LDL levels for groups defined by vital status at 5 years is also equivalent to what was determined using the two-sample t-test that assumes equal variance. However, the standard error of this point estimate, 95% CI, and p-value were different from those obtained for the t-test that assumes equal variances. This is due to the assumptions about variance and how this impacts the calculation of the pooled standard error. Because the two-sample t-test that allows for the possibility that the variances could be unequal a different method is used to determine the pooled standard error resulting in a different 95% CI and p-value. In this case the assumption that the variances were equal resulted in anti-conservative estimates of the true population parameters.**

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.)

|  |  |  |
| --- | --- | --- |
|  | **Linear Regression Assuming Homoscedasticity** | **Linear Regression Allowing for Heteroscedasticity** |
| **Vital Status at 5 years** | **Deceased** | **Alive** | **Deceased** | **Alive** |
| **Estimate of Population Mean Serum LDL (mg/dL)** | 118.70 | 127.20 | 118.70 | 127.20 |
| **Standard Error of Estimate (mg/dL)** | 3.07 | 1.36 | 3.31 | 1.34 |
| **95% Confidence Interval** | (112.67, 124.72) | (124.53, 129.87) | (112.21, 125.19) | (124.57, 129.83) |
| **Estimate of True Difference in Population Means****(mg/dL)** | 8.50 | 8.50 |
| **Standard Error of Estimate****(mg/dL)** | 3.36 | 3.57 |
| **95% Confidence Interval** | (1.91, 15.09) | (1.50, 15.50) |
| **p-value** | 0.012 | 0.017 |

**Values in red are not equivalent between the two models.**

**The only equivalent values between the two regression models are the point estimates for the mean serum LDL levels in populations defined by vital status at 5 years and the point estimate for the true difference in the mean serum LDL levels across these populations. However, the standard errors, 95% CIs, and p-value obtained for the linear regression using robust standard errors were different from those obtained for the linear regression that assumes homoscedasticity. This is due to the assumptions about variance and how this impacts the calculation of the pooled standard error. Because the regression model using robust standard errors allows for the possibility that the variances could be unequal a different method is used to determine the pooled standard error resulting in different 95% CIs and p-value. In this case the assumption that the variances were equal resulted in anti-conservative estimates of the true population parameters.**

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.)

|  |  |  |
| --- | --- | --- |
|  |  | **Age (years)** |
|  |  | **65-69****(n = 117)** | **70-74****(n = 305)** | **75-79****(n = 187)** | **80-84****(n = 81)** | **85-89****(n = 35)** | **90-94****(n = 8)** | **≥95****(n = 2)** | **All Ages****(n = 735)** |
| **Male** | missing | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| (%) | 47.86% | 49.51% | 50.27% | 46.91% | 54.29% | 75.00% | 100.00% | 49.80% |
| **Serum LDL1****(mg/dL)** | missing | 3 | 2 | 3 | 1 | 1 | 0 | 0 | 10 |
| mean | 127.7 | 125.3 | 126.9 | 122.8 | 125.0  | 124.8  | 132.0 | 125.8 |
| SD | 32.4 | 32.5 | 35.5 | 33.5 | 39.1 | 35.8 | 1.4 | 33.6 |
| min-max | 51 – 217 | 37 – 247 | 11 – 225 | 52 – 227 | 68 – 216 | 57 – 175 | 131 – 133 | 11 – 247 |

**The mean serum LDL levels are similar in magnitude across all age groups. However, there are significant differences in the percentage of each age group that is male. Therefore, if there is an association between gender and serum LDL this could cause effect modification or confound the analysis.**

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

**Serum LDL levels were selected as the response variable and age as the predictor of interest. A linear regression model allowing for heteroscedasticity was used to determine if there was a linear trend in mean serum LDL across groups defined by age. Both variables were treated as continuous. An alpha level of 0.05 was used and the 95% confidence interval was calculated. The null hypothesis was that there is no linear trend in the mean serum LDL across age groups, β1= 0. The alternative hypothesis is that there is a linear trend in the mean serum LDL across age groups, β1≠ 0.**

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

**This model is not saturated. In order for a model to be saturated the number of groups must be equivalent to the number of parameters. For a linear regression there are two parameters that must be fit, β0 and β1, therefore, because age is a continuous variable there are an infinite number of possible groups and the model is unsaturated.**

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

**The estimated mean serum LDL level among a population of 70 year olds is given as 126.22 mg/dL by this model. This was calculated using the estimated linear relationship:**

$$E\left[Age\_{i}\right]=132.53-0.09×Age\_{i}$$

* 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 relate to the slope?

**The estimated mean serum LDL level among a population of 71 year olds is given as 126.13 mg/dL by this model. This was also calculated using the estimated linear relationship given above. The difference in estimated mean serum LDL levels for populations of 70 and 71 year olds is 0.09 mg/dL lower for a population of 71 year old subjects. This is exactly the slope given by the linear regression model. The slope tells the estimated difference in mean serum LDL levels per year in age. This relates to saturation, because if a model is unsaturated the parameters given by the model are not an exact fit for the sample because information must be borrowed across groups for determining the means and this value will likely differ from the true sample mean serum LDL level among 71 year old study subjects.**

* 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?

**The estimated mean serum LDL level among a population of 75 year olds is given as 125.76 mg/dL by this model. This was also calculated using the estimated linear relationship given above. The difference in estimated mean serum LDL levels for populations of 70 and 75 year olds is 0.45 mg/dL lower for a population of 75 year old subjects. This is exactly 5 times the slope given by the linear regression model. The slope tells the estimated difference in mean serum LDL levels per year in age, so because 70 to 75 represents a 5 year increase in age the difference is 5 times the slope. This relates to saturation, because if a model is unsaturated the parameters given by the model are not an exact fit for the sample because information must be borrowed across groups for determining the means. Therefore this value will likely differ from the true sample mean serum LDL level among 75 year old study subjects.**

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

**The root mean squared error in the regression model is the pooled standard deviation for mean serum LDL levels across all age groups. This is used as an estimate of the within group standard deviation. Because this model did not assume homoscedasticity this value is based on the average within group variance.**

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

**The intercept gives the estimated mean serum LDL level for newborns (age = 0) as 132.53 mg/dL. This estimate is well outside the range of the sample, the youngest subject(s) sampled were 65 years of age. Therefore, this estimate has no predictive value and its interpretation is not scientifically relevant, especially as LDL is expected to be quite low in infants.**

* 1. What is the interpretation of the slope?

**The slope tells the estimated difference in mean serum LDL levels per year in age. Using this model, there is an estimated decrease of 0.09 mg/dL in average serum LDL between subjects of populations separated by one year of age.**

* 1. Provide full statistical inference about an association between serum LDL and age based on your regression model.

**From linear regression analysis, we estimate that for each year difference in age, the difference in mean serum LDL is -0.09 mg/dL, with older groups averaging a lower serum LDL level. A 95% confidence interval suggests that this observation is not unusual if the true difference in mean serum LDL per year is between -0.55 and 0.37 mg/dL. Because the p-value is greater than 0.05 (p = 0.698) we fail to reject the null hypothesis that there is no linear trend in the mean serum LDL across age groups.**

* 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?

**The estimate for the difference in mean serum LDL across groups differing by 5 years in age is given by -0.45 mg/dL, with older groups averaging a lower serum level. This is 5 times the estimate for the difference in mean serum LDL across groups differing by 1 year in age. The 95% confidence interval for this estimate is given by -2.73 to 1.83 mg/dL.**

* 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 coefficient for serum LDL and age is r = -0.01, this is equivalent to the square root of the R2 value given with the regression model. This value is also consistent with the inference made based on the regression model, that we cannot with confidence reject the null hypothesis of no linear trend between serum LDL levels and age. An r value of 0 indicates no linear correlation, so a value of -0.01 is consistent with no or very weak linear correlation which indicates that we cannot reject the null.**