Biostats 518 Homework 2

TOTAL SCORE: 43/93

1. Question 1SCORE: 8/15

|  |  |  |
| --- | --- | --- |
| **Type** | **Sample Size** | **Mean LDL (sd) mg/dl** |
| Died within 5 years | 119 | 118.7 (36.2) mg/dl |
| Survived at least 5 years | 606 | 127.2 (32.9) mg/dl |

* 1. The samples seem to be similar in magnitude. They are both in the hundreds number set I’m not sure what you mean here. Perhaps you mean to say sample mean?(-0.5). The standard deviations are also similar with an absolute difference of 3.3 mg/dl.
	Final score: 2.5/3

|  |  |  |  |
| --- | --- | --- | --- |
| **Type** | **Mean LDL** | **Standard Error** | **95% CI [low, high]** |
| Died within 5 years | 118.7 mg/dl | 3.3 | [112.1, 125.3] |
| Survived at least 5 years | 127.2 mg/dl | 1.3 | [124.5, 129.8] |

* 1. The point estimates seem to be similar in magnitude Quantify any differences(-0.5). The standard errors are also similar in magnitude, however there is an absolute difference of 2 mg/dl between them. The standard error for the group that died within 5 years is larger because the sample size is smaller. The standard error for the group that survived at least 5 years is smaller because the sample size is larger. Sample size alone does not explain the difference in standard error(-0.5)
	Final Score: 2/3
	2. The overlap between the high end of the 95% CI for the folks who died within 5 years and the low end of the 95% CI for the folks who survived at least 5 years seems to be minimal (0.8 mg/dl). Based on this, we can conclude that there is a significant difference between the groups.
	If they overlap then we cannot conclude a significant different. The overlap is minimal as you mentioned but we do not know how to quantify this or test if this minimal overlap is significant and so we cannot conclude anything from this. (-1)
	Final Score: 2/3
	3. Assuming the means of the two groups are different but the variances are equal, the best estimate for the standard deviation would be the standard deviation from a geometric mean of the LDL.
	This question was looking for the estimated of the **pooled standard deviation** which uses information from both groups. (-3)
	Final Score: 0/3

|  |  |  |  |
| --- | --- | --- | --- |
| **Type** | **Mean LDL mg/dl** | **Standard Error** | **95% CI [low, high]** |
| Difference | 8.5 mg/dl | 3.5 | [1.4, 15.5] |

* 1. The p-value that the groups don’t have equal mean LDLs is 0.0186, so the p-value that the groups do have equal mean LDLs is 0.9814. Based on all of the data and analysis we’ve conducted thus far, I can conclude that there is a significant association between serum LDL and 5-year all cause mortality.
	You need to specify if the difference is lower or higher for one group versus the other(-0.5) . Secondly, stating that the p –value is 0.9814 for groups do have equal mean is not valid. I’m not sure what you were trying to show with this(-1).
	Final Score: 1.5/3
1. Question 2 SCORE: 11.5/20
	1. Both of the models are not saturated because we have more data points (119 for died within 5 years and 606 for survived at least 5 years) This is not the definition of a saturated model. Both models are saturated because they are binary so we have 2 groups and only 2 parameters. (-1) than estimated parameters in our polynomial equation 🡪 E(ldl|deadin5) = 127.2-8.5\*deadin5
	Providing an explicit model is not part of the answer key but otherwise I would suggest presenting the second model as well. None the less you need to specify what model A is and what model B is(-0.5).
	Final Score: 0.5/2
	2. Using the model based off of those who died in 5 years (deadin5), the group who survived at least 5 years have an LDL value of 127.2 mg/dl. This is the exact same answer we got when we ran a t-test in question 1.
	Final Score: 2/2
	3. Using the model based off of those who died in 5 years (deadin5), the group who survived at least 5 years have a 95% CI of 124.5 mg/dl on the low end and 129.8 mg/dl on the high end. These are the exact estimates derived from the t-test in question 1.
	Not sure how you got the same answer but there should be a slight difference which is explained by the fact that one uses the pooled standard error and the other does not.(-1)
	Final Score: 1/2
	4. Using the model based off of those who died in 5 years (deadin5), the group that died within 5 years have an LDL value of 118.7 mg/dl. This is the exact same answer we got when we ran a t-test in question 1.
	How did you get this answer from the model. Need to explain it is obtained by setting deadin5 ==1(-0.5)
	Final Score: 1.5/2
	5. Using the model based off of those who died in 5 years (deadin5), the group that died within 5 years have a 95% CI of 109.4 mg/dl on the low end and 124.5 mg/dl The point estimates are incorrect. Please explain how you obtained these estimates from Model A.(-0.5) on the high end. These estimates, however, seem to be a little different from the values I derived in question 1. There seems to be an absolute difference of 2.7 mg/dl on the low end and 0.8 mg/dl on the high end.Please explain the source of this difference (-0.5)
	Final Score: 1/2
	6. For the first model, the deadwithin5 group seems to have a standard deviation of 4.65 and the ones who survived at least 5 years seem to have a standard deviation of 1.3. The standard error we obtained in question 1 is the same for the group that survived at least 5 years, but greater than the value for the deadwithin5 group (3.3). For the second model, the deadwithin5 group seems to have a standard deviation of 6.36 and the ones who survived at least 5 years seem to have a standard deviation of 9.66. These results are different from what we obtained in question 1. The estimates from this model are much higher, probably because there are fewer data points so the standard error is higher.
	Assuming equal variance the estimated standard deviation should be the same for each group and this is the RMSE.
	Final Score: 0.5/2
	7. Both of the models seem to be the identical on the surface, but in reality they give us different results from each other. It seems like a lot hinges I belive this is a typo but because of this unfortunately I can’t understand what you mean by this sentence on the data the model was generated from.
	The question asked about how they are related which I felt, was not adequately answered.
	Final Score: 0.5/2
	8. E(LDL|deadin5) = 127.2 – 8.5(X)

When X (aka deadin5) is equal to 0 (the person survived at least 5 years), the mean LDL is 127.2 mg/dl.
Final Score: 2/2

* 1. When X (aka deadin5) is equal to 1 (the person died within 5 years of the first MRI), the mean LDL is 8.5 mg/ml less.
	Final Score: 2/2

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Mean LDL mg/dl | Standard Error | 95% CI [low, high] |
| Model1 | 127.2 mg/dl | .3 | [-15,.1 -1.9] |

* 1. The p-value testing whether the populations are the same is equal to the square of the r-squared value (0.0088^2) = 7.7e-5. We can again conclude that there is a significant association between LDL and 5-year all cause mortality. This matches the conclusion that was drawn in question 1.
	The estimated standard error is incorrect(-0.5) The P-value is incorrect as it is not the square of the R-square value. I believe Stata reports the P-value separately. (-0.5). You also need to provide a point estimate of the difference or state if one group has a higher mean than the other and if so then by how much. (-0.5)
	Final Score: 0.5/2
1. For the deadin5 group, the p-value seems to be smaller and the confidence intervals seem to be more narrow (compared to the results from question 1). For the group that survived at least 5 years, the p-value seems to be larger and the confidence interval seems to be wider (compared to the results from question 1).
This question asked about t-test on the difference of means and so this was a two sample t-test. The comparison should be between the one value obtained for SE, CI and P-value etc versus the SE,CI, P-value obtained in question 1. (-5) Furthermore, even if you decide to state the results for difference groups you did not provide the point estimates and confidence intervals or state how much the difference was (-2)
Final Score: 3/10
2. For the deadin5 group, the 95% CI seems to be more narrow than the results we saw in question 3. For the group that survived at least 5 years, the 95% CI seems to be wider than compare to question 3. These trends seem to match what we expect based on the analyses we’ve done so far and the observation that t-test and regression can be interchangeable for this data set.
I’m not sure I understand your conclusion. First you stated the differences and then conclude that the two test are interchangeable. In terms of our final decision they can be interchanged but they will still provide different confidence intervals and hence different results(-1). Again, you need to specify either the point estimates or quantify the differences. (-2) The question asked for the inference for the difference in means not means of each group.(-5)
Final Score: 2/10
3. Question 5 SCORE: 18.5/38

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Type** | **Sample Size** | **Mean** | **Standard Deviation** | **Min** | **Max** |
| LDL (mg/dl) | 393 | 101.25 | 19.29 | 11 | 129 |
| Age (years) | 225 | 74.69 | 5.25 | 65 | 92 |
| Male  | 107 | 0.55 | 0.49 | 0 | 1 |

* 1. Descriptive stats above communicate the very basic knowledge required to assess whether or not there is a relationship between LDL and age (and whether sex is a confounder)
	The table above does not allow us to easily identify any possible associations in the sample. Secondly, it is not clear how you picked your sample size and how they differ for LDL and Age. Reporting the SD, min and Max was redundant for Sex.
	Final Score: 1/5
	2. I used simple regression to create a model that relates LDL and age. The process includes plotting age on the X-axis and LDL (mg/dl) level on the Y-axis, and then creating a best-fit line to summarize the data points. Best being defined by having the lowest sum of square residuals value as possible (because we are using Ordinary Least Squares). Sometimes RMSE or root mean squared error is also used to communicate how “well” a model is fit to its data.
	You need to state if you are assuming homoscedasticity or not. (-0.5). It is not clear if the variables are treated as continuous or categorical(-0.5)
	Final Score: 2/3
	3. A saturated model is one where the number of parameters is equal to the number of data points. The answer is no, this model is not saturated. There are 725 data points in this model and only one parameter.
	The definition of a saturated model is incorrect(-0.5) but the conclusion is true.
	Final Score: 2.5/3
	4. E(LDL|age) = 132.5 – 0.09(AGE)
		1. 132.5 – 0.09\*70 = 126.2 mg/dl

Final Score :3/3

* 1. 132.5 – 0.09\*71 = 126.11 mg/dl. The answer to this question is a little lower (by 0.09 mg/dl, or 1 unit) than the previous answer. The slope is an averaged estimation of all of the data points used to fit the model from part c.
	The point estimate is correct and difference is correct but interpretation is incorrect of regarding what the difference is compared to the slope.
	Final Score: 2/3
	2. 132.5 – 0.09\*75 = 125.75 mg/dl. As age increases, the mean LDL value seems to decrease at a rate of 0.09 mg/dl. Again, the slope is an average of the general trend of the data. You need to state that the difference is 5 times the slope (-1)
	Final Score: 2/3
	3. The root mean squared error (RMSE) in my model is 33.62 (mg/dl). In general, RMSE refers to how close the model fit line is to the actual data points. What this means in particular to my model is that the trend line has a leeway of 33.62 mg/dl.
	Need to state that RMSE is standard deviation within each group.
	Final Score: 2/3
	4. There is meaning in the intercept. It means every person in our model will have a mean LDL value of at least 132.5 mg/dl.
	That is not correct interpretation of the intercept.
	Final Score: 0/3
	5. There is also meaning in the slope. It means a person in the model will have 0.09 mg/dl lower mean LDL serum values for each year they age.
	The slope is the estimated difference in mean LDL levels between two groups that differ by 1 year. It is not a deterministic model which states the any person will have a lowered LDL level as they grow older.
	Final Score: 0/3
		1. Method: I used simple linear regression to fit the model and used ordinary least squares to understand how well the model was fit to the data.
		2. Inference: Based on the model, there seems to be the trend that the older a person is, the lower their mean LDL serum value is. The p-value is about 4e-8 (r-squared\*r-squared), which is definitely denotes a significant association.
		The method has already been stated in the question. Again, please note that R^4 is NOT the P-value. You need to state the point estimates and confidence intervals. The conclusion is incorrect.
		Final Score: 0/3
	6. When calculating the point estimate and 95% CI for the mean difference across groups that differ by 5 years, I would report
		1. Point estimate difference = 0.45 mg/dl
		2. 95% CI = [-2.7, 1.8]

Final Score: 3/3

* 1. The correlation between LDL serum value and age seems to be -0.0146. This seems to indicate that there is a very weak relationship between those two variables. This is contrary to the conclusion I previously came to when viewing the data solely through the lens of regressions and descriptive statistics.
	Correct correlation however there was no reported P-value and a conclusion could not have been drawn from the point estimate alone.
	Final Score: 1/3