Homework #2

Anonymizing Code: 2151

1) Methods: Mean serum LDL levels were compared between subjects who died within 5 years of study enrollment and subjects who survived at least 5 years. A t test assuming equality of variances was performed to analyze differences in means between the two groups. 95% population means were based on the same handling of variances. The null hypothesis established was that there is no association between whether a subject survives at least 5 years and mean LDL.

1a)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Sample Size | Sample Mean | Sample Standard Deviation | 95% Confidence Interval |
| Subjects Surviving at least 5 Years | 606 | 127.198 | 32.9289 | 124.571-129.825 |
| Subjects Who Died within 5 Years | 119 | 118.6975 | 36.157 | 112.1338-125.2611 |

The sample means and standard deviations differ somewhat in magnitude. From a t test presuming equal variance, I estimate that the mean serum LDL level in patients who survived for at least 5 years is 127.198 (95% CI: 124.57-129.83) with a standard deviation of 32.93. I estimate that the mean serum LDL level in patients who died within 5 years is 118.698 (95% CI: 112.13-125.26) with a standard deviation of 36.16.

1b)

|  |  |  |  |
| --- | --- | --- | --- |
|  | Point estimate | Estimated Standard Error | 95% Confidence Interval |
| Subjects Surviving at least 5 Years | 127.198 | 1.3376 | 124.571-129.825 |
| Subjects Who Died within 5 Years | 118.6975 | 3.3145 | 112.1338-125.2611 |

The point estimates differ somewhat in magnitude. From a t test presuming equal variance, I estimate that the mean serum LDL level in patients who survived at least 5 years to be 127.198 (95% CI: 124.57-129.83) with an estimated standard error of 1.34. I estimate that the mean serum LDL level in patients who died within 5 years to be 118.698 (95% CI: 112.13-125.26) with an estimated standard error of 3.31. The standard errors for each estimate differ considerably given the large difference in sample size between the two groups. The point estimates provided equal the sample means provided above because the sample mean is the best point estimate available.

1c) Yes, the confidence interval for the mean LDL of those individuals dying within 5 years overlaps with that of those individuals who survived at least 5 years. This information does not suggest any conclusion as to the statistical significance of the estimated difference in means. For that, a p-value is required.

1d)

1e)

|  |  |  |  |
| --- | --- | --- | --- |
|  | Point estimate | Estimated Standard Error | 95% Confidence Interval |
| Difference in mean LDL between a population surviving at least 5 years and a population that dies within 5 years | 8.5005 | 3.3567 | 1.9106-15.0905 |

From a t test presuming equal variance, I estimate that the mean serum LDL level in patients who died within 5 years is 8.5 mg/dL lower than then mean serum LDL level in patients who survived at least 5 years. A 95% confidence interval suggests that this observation is not unusual if the true difference in mean LDL level were between 1.91 mg/dL and 15.09 mg/dL. Because the P value is P=0.0115, I reject the null hypothesis that there is no association between whether a patient survived at least 5 years and serum LDL level.

2) Methods: Mean serum LDL levels were compared between subjects who died within 5 years of study enrollment and subjects who survived at least 5 years. An ordinary least squares regression presuming homoscedasticity was performed to analyze differences in means between the two groups. 95% population means were based on the same handling of variances. Model A is based on an OLS regression using subjects that died within 5 years as the predictor. Model B is based on an OLS regression using subjects that survived at least 5 years as the predictor. The null hypothesis established was that there is no association between whether a subject survives at least 5 years and mean LDL.

2a) A saturated model is a model in which the number of groups equals the number of parameters. Whenever we use a binary predictor, we have a saturated model because B0 and B1 are fit to the two possible values of the predictor. When we have a saturated model, the parameters will estimate the sample descriptive statistics perfectly. In model A we can see this because the intercept is 127.198 (the sample mean for subjects surviving at least 5 years) and the slope is -8.5005 (a negation of the difference in means between those subjects who survived at least 5 years and those who died within 5 years). Similarly in model B, we can see that the intercept is 118.6975 (the sample mean for subjects who died within 5 years) and the slope is again 8.5005 (the difference in means between those subjects who survived at least 5 years and those who died within 5 years). Given that both the slope and the intercept directly predict the sample descriptive statistics and that our predictor of interest is binary and we have two parameters, both models are saturated.

2b) Using model A, the estimate of true mean LDL among a population of subjects who survive at least 5 years is 127.198. This is exactly the same as the corresponding estimate in problem 1.

2c) Using model A, the 95% confidence interval for the true mean LDL among a population of subjects who survive at least 5 years is 124.528-129.868. This interval differs ever so slightly from the corresponding estimate in problem 1 because it is using the pooled standard deviation rather than the in group estimated standard deviation derived from the sample of subjects who survive at least 5 years. It borrows this information because it assumes equal variance between groups.

2d) Using model B, the estimate of true mean LDL among a population of subjects who die within 5 years is 118.698. This is exactly the same as the corresponding estimate in problem 1.

2e) Using model B, the 95% confidence interval for the true mean LDL among a population of subjects who die within 5 years is 112.673-124.722. This interval differs ever so slightly from the corresponding estimate in problem 1 because it is using the pooled standard deviation rather than the in group estimated standard deviation derived from the sample of subjects who died within 5 years. It borrows this information because it assumes equal variance between groups.

2f) The regression based estimate of standard deviation within each group is 33.477 for each model. This is slightly lower than the corresponding estimate from problem 1.

2g) Models A and B are simply reparameterizations of one another.

2h) From ordinary least squares regression presuming homoscedasticity, I estimate that the intercept is equal to 127.198 (95% CI: 124.528-129.868). In other words, a population in which all modeled covariates are equal to zero would have a mean serum LDL level of 127.198 mg/dL. A 95% confidence interval suggests that this observation is not unusual if the true mean LDL in the population of individuals who survive at least 5 years were between 124.528 and 129.868. Because the P value is P<0.001, I reject the null hypothesis that there is no association between serum LDL level and whether or not an individual dies within 5 years.

2i) From ordinary least squares regression presuming homoscedasticity, I estimate that the slope is equal to -8.5005 (95% CI: -15.09— -1.91). In other words, a population that survives at least 5 years tends to have a serum LDL level 8.5005 mg/dL higher than a population that dies within 5 years. In other words the true difference in mean LDL between the two populations is 8.5005 mg/dL. A 95% confidence interval suggests that this observation is not unusual if the true difference in mean LDL between these populations were between -15.09 and -1.91. Because the P value is P=0.012, I reject the null hypothesis that there is no association between serum LDL level and whether or not an individual dies within 5 years.

2j) From ordinary least squares regression presuming homoscedasticity, I estimate that the true difference in mean serum LDL level is 8.5005 mg/dL between a population that survives at least 5 years and a population that dies within 5 years. A 95% confidence interval suggests that this observation is not unusual if the true difference in mean LDL between these populations were between 1.91 and 15.09. Because the P value is P=0.012, I reject the null hypothesis that there is no association between serum LDL level and whether or not an individual dies within 5 years. This conclusion further upholds the inference drawn in problem 1.

3) Method: Mean serum LDL levels were compared between subjects who died within 5 years of study enrollment and subjects who survived at least 5 years. A t test allowing for unequal variances was performed to analyze differences in means between the two groups. 95% population means were based on the same handling of variances. The null hypothesis established was that there is no association between whether a subject survives at least 5 years and mean LDL.

The means, standard deviations, point estimates, and estimated standard errors for each group do not change. While the mean difference between groups does not change, the standard error and thus the 95% confidence interval around the mean difference does change, becoming wider on account of the larger estimated standard error. As a result, the p value also changes (p=0.0186), however remains statistically significant.

4) Method: Mean serum LDL levels were compared between subjects who died within 5 years of study enrollment and subjects who survived at least 5 years. A linear regression allowing for unequal variances was performed to analyze differences in means between the two groups. 95% population means were based on the same handling of variances. The null hypothesis established was that there is no association between whether a subject survives at least 5 years and mean LDL.

The only differences between this analysis and that in problem 3 is that there is a slight difference in standard errors and thus confidence intervals and p values differ slightly as well.

5) Method: Mean serum LDL levels were compared with the age of the subjects. A linear regression allowing for unequal variances was performed to analyze the association between the two variables. 95% population means were based on the same handling of variances. The null hypothesis established was that there is no association between age and serum LDL level.

5a)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | Serum Low Density Lipoprotein (LDL) | | | |
|  |  | **≤129 mg/dL**  **(n=393)** | **130-159 mg/dL**  **(n=225)** | **≥ 160 mg/dL**  **(n=107)** | **Any Level**  **(n=725)** |
| Age (years) | Mean | 74.70 | 74.20 | 74.88 | 74.57 |
| Standard Deviation | 5.25 | 5.62 | 5.77 | 5.45 |
| Minimum | 65 | 67 | 65 | 65 |
| Maximum | 92 | 99 | 94 | 99 |
| Male (%) | | 55.5% | 43.1% | 42.1% | 49.7% |

Data is available for 735 subjects, though due to missing data, ten subjects have been removed from the study. Note that we cannot assess the impact that the removal of those ten subjects might have on the generalizability of our results. Above is a table depicting the relevant variables stratified into three LDL levels.

5b) Mean serum LDL levels were compared with the age of the subjects. A linear regression allowing for unequal variances was performed to analyze the association between the two variables. 95% population means were based on the same handling of variances. The null hypothesis established was that there is no association between age and serum LDL level.

5c) Given that both the predictor and the outcome are continuous variables this model is not saturated. Given that ages are calculated in years and range from 65 to 99 we would need 35 different parameters, something that we do not have, nor want.

5d) Based on the regression model, estimated mean LDL level among a population of 70 year old subjects is 126.21 mg/dL.

5e) Based on the regression model, estimated mean LDL level among a population of 71 year old subjects is 126.12 mg/dL. The difference between this and the estimated mean LDL in a population of 70 year olds is .09, or the slope of the linear regression. The slope of the linear regression is the difference between mean LDL levels for two populations separated by 1 year of age.

5f) Based on the regression model, estimated mean LDL level among a population of 75 year old subjects is 125.76 mg/dL. The difference between this and the estimated mean LDL in a population of 70 year olds is .45 or the slope of the line multiplied by the 5 year age difference between the two populations.

5g) The root mean squared error is the sample standard deviation of the residuals. Based on the regression model, the variance of the residuals, measured as the sample standard deviation of the residuals is 33.622.

5h) Based on the regression model the intercept is 132.5281 (95% CI: 98.48-166.58). Given that the youngest person in the study is 65, this intercept is scientifically meaningless. If however we reparamiterize our age variable to age above 65, our intercept shifts to 126.6657 (95% CI: 121.74-131.59) which holds meaning as the estimated mean LDL for a population of 65 year olds.

5i) Based on the regression model the slope is -.0901904 (95% CI: -.55-.37). In other words for each additional average year of age of a population, the mean LDL of the population decreases by .0901904.

5j) From linear regression analysis, I estimate that for each year difference in age, the differene in mean LDL is .09 mg/dL. A 95% confidence interval suggests that this observation is not unusual if the true difference in mean LDL per year difference in age were between -.55 and .37 mg/dL. Given a p value of p=0.698, I have failed to reject the null hypothesis that there is no linear trend between age and mean LDL.

5k) The point estimate and CI for the difference in mean LDL across groups that differ for 5 years in age would simply be 5 times greater than the above listed point estimates and confidence intervals. The point estimate would be -.450952 with a 95% confidence interval of -2.735-1.833.

5l)