HW2

1. Mean serum LDL was compared between those who died within 5 years of study enrollment and those who survived for at least 5 years. A t-test assuming equal variances was used to test differences in the mean. 95% confidence intervals were also based on the same handling of variances as previously mentioned.

1a. There were 606 subjects who survived for at least 5 years, and their sample mean of LDL was 127.2 mg/dL (SD=32.9). There were 119 subjects who died within 5 years, and their sample mean of LDL was 118.7 mg/dL (SD=36.2). The sample means are similar in magnitude, and the sample standard deviations are similar.

1b. The point estimate of the true mean LDL in a population of similar subjects who would survive for at least 5 years is 127.2 mg/dL (SE=1.34, 95% CI: 124.6-129.8). For similar subjects who would die within 5 years, the point estimate of the true mean LDL is 118.7 mg/dL (SE=3.31, 95% CI: 112.1-125.3). The point estimates are similar in magnitude, although the standard error for the death within 5 years subjects is about 2.5 times the standard error for the survival of at least 5 years subjects. Standard error is influenced by the sample size (SD/√n). Although the standard deviations between the two groups are similar, the death within 5 years group is about 1/5th the size of the survival past 5 years group, so its standard error in the death within 5 years group is thus larger.

1c. The confidence intervals of mean LDL for those surviving 5 years (124.6-129.8) and those dying within 5 years (112.1-125.6) do overlap. Although they do overlap, the overlap is relatively small and we cannot be sure about the statistical significance of an estimated difference of means with 0.05 level of significance looking solely at the confidence intervals.

1d. If the variances in the two populations are equal, then we can estimate the standard deviation of LDL in each group using the pooled variance. Thus we can use the following formula using the sample size and SD of each group:



sp=√(118\*36.162+605\*32.932)/(725-2)=33.48.

1e. The true difference in means between a population surviving at least 5 years and one that dies within 5 years has a point estimate of 8.5 mg/dL (95% CI: 1.9, 15.1) and estimated standard error of 3.36. The p-value testing the same mean LDL is 0.0115. Since our p-value is statistically significant at 0.05 level of significance, we conclude with high confidence that that the distribution of serum LDL is different between those who die within 5 years and those who survive for at least 5 years.

2a. Linear regression using ordinary least squares and presuming homoscedasticity was used to compare mean LDL across groups defined by vital status at 5 years. Both of these are saturated models, since the predictor variable in each only has two values (died within 5 years or survived at least 5 years).

2b. Using the intercept from Model A, we get an estimate of true mean LDL for those surviving at least 5 years of 127.2 mg/dL. This is the same as estimate from problem 1.

2c. Using Model A, the 95% confidence interval for the true mean LDL of those surviving at least 5 years is (124.53, 129.87). This confidence interval is a little wider than that in problem 1. It is wider because it uses a sample size of the entire population, instead of one group as in problem 1.

2d. Using the intercept from Model B, we get that the estimate of the true mean LDL among a population of subjects who die within 5 years is 118.7 mg/dL. This is the same as the estimate in problem 1.

2e. The confidence interval for the true mean LDL among a population of subjects who die within 5 years (112.67, 124.72) is slightly wider than interval from problem 1. It is wider because it uses a sample size of the entire population, instead of one group as in problem 1.

2f. The root MSE of the model is 33.477, so it is the regression based estimate of the standard deviation within each group. This is the same as the estimate in problem 1.

2g. The slopes in Models A and B are the additive inverses of each other, and represent the difference in mean LDL between the 5 year mortality groups. The intercepts in Models A and B represent the estimated true mean LDL in those surviving at least 5 years and those dying within 5 years, respectively. The true mean of the other group can be calculated by adding the slope to the intercept.

2h. The intercept in Model A is the estimate of the mean LDL among a population of subjects who survive at least 5 years. Thus, the estimated mean LDL among a population of subjects who survive for at least 5 years is 127.2 mg/dL.

2i. The slope of Model A is the estimate of the difference in mean LDL among a population subjects who survive at least 5 years and a population that dies within 5 years. Thus, estimated mean LDL among a population of subjects who die within 5 years is 8.5 mg/dL lower than the estimated mean LDL among a population of subjects who survive for at least 5 years.

2j. The true difference in mean LDL between a population that survives at least 5 years and one that dies within 5 has a point estimate of 8.5 mg/dL (95% CI: 1.91, 15.09) and a standard error of 3.36. The p-value testing if the two populations have the same mean LDL is 0.012. Since our p-value is statistically significant at 0.05 level of significance, we conclude with high confidence that that the distribution of serum LDL is different between those who die within 5 years and those who survive for at least 5 years. This is the same inference as in problem 1.

3. The 5-year survival group-specific sample size, sample mean, sample standard deviation, sample standard error, 95% confidence interval, and point estimate of the true difference are the same as in problem 1. The confidence intervals overlap with each other as well. However, the standard error of the true difference’s point estimate and its 95% confidence interval are different from problem 1 (specifically larger SE and wider CI). This is because problem 1 uses the pooled variance, whereas this problem uses the sample variance separately, to calculate the standard error and 95% CI of the mean difference. This results in a larger standard error, smaller t-statistic and larger p-value, and wider confidence interval for the unequal variance t-test. The t-test presuming equal variance is anti-conservative, and the variances of the two groups are not the same.

4. The point estimate of the true difference in mean LDL between a population that survives at least 5 years and one that dies within 5 years is the same using unequal variance t-test or robust linear regression. However, the standard error and p-value is smaller using robust regression, and also has narrower 95% confidence interval for the difference in mean LDL. These are not the same because the robust regression approximates the t-test of unequal variances using the Huber-White sandwich estimator.

5a. Below is a jittered scatterplot of LDL level by age with a superimposed smooth. Under that is are jittered scatterplots of LDL level by age for females and males separately.





5b. Simple linear regression was used to look at the association between LDL and age, with LDL being the response variable and age being the predictor variable. LDL and age were both continuous variables. The significance of age was tested using a t-statistic presuming equal variance. 95% confidence intervals were also based on the same handling of variances.

5c. This is not a saturated model since the number of groups is not equal to the number of parameters (we only have one predictor, and it is continuous).

5d. The regression model is E(LDLi | Agei)=132.53-0.09\*Agei

Putting in a value of 70 for age, we get 132.53-0.09\*70=126.23. Thus, the estimated mean LDL level among a population of 70-year-old subjects is 126.23 mg/dL.

5e. Putting 71 for age in our regression model of E(LDLi | Agei)=132.53-0.09\*Agei

132.53-0.09\*71=126.14. The estimated mean LDL level among a population of 71-year-old subjects is 126.14 mg/dL. The difference between the estimated mean LDL among 71 year olds and 70 year olds is 0.09, the slope of the model.

5f. Putting 75 for age in our regression model of E(LDLi | Agei)=132.53-0.09\*Agei

132.53-0.09\*75=125.78. The difference between the estimated mean LDL among 75 year olds and 70 year olds is 0.45, five times the slope of the model.

5g. The root mean squared error is 33.62. The estimated within group standard deviation is 33.62.

5h. The intercept is the estimated mean LDL level among a population of 0 year olds. This does not have a relevant scientific interpretation since a 0 year old would deterministically score 0.

5i. The slope is the difference in mean LDL between two groups differing by one year in age. Thus, the one year difference in mean LDL is 0.09 mg/dL.

5j. **Method**: Simple linear regression was used to look at the association between LDL and age, with LDL being the response variable and age being the predictor variable. LDL and age were both continuous variables. The t-statistic was used to test whether the age coefficient is significantly different from zero.

**Inference:** A one-year increase in age corresponded to a 0.09 mg/dL decrease in mean LDL. Based on a 95% confidence interval, this decrease of 0.09 mg/dL would not be judged unusual if the true difference of LDL was anywhere between a decrease of 0.54 mg/dL or increase of 0.36 mg/dL for an increase of age by one year. Using a t-test that presumes equal variance (homoscedastic regression), this observation is not statistically significant at a 0.05 level of significance (two-sided p-value=0.69), suggesting that we cannot with high confidence reject the null hypothesis that age is not linearly associated with LDL level.

5k. The estimate would be the estimate of one year difference multiplied by 5, so -0.09\*5=-0.45. The standard error of the CI can be found for five years by multiplying by the standard error by the square root of 5, so the confidence interval will be -0.45±1.96\*0.23\*√5 (95% CI: -1.46,0.56).

5l. **Method**: Pearson correlation analysis was conducted on LDL and age, both of which were continuous variables. 5% level of significance was used for determining the significance of correlation.

**Inference:** There was a weak correlation between LDL and age (-0.015). This was not statistically significant at a 0.05 level of significance (two-sided p-value=0.69), suggesting that we cannot with high confidence reject the null hypothesis that the correlation is zero. The p-value (and conclusion of association) was the same using correlations and regression.