**Homework #2**

1. A. There are 614 patients who survived at least 5 years, and 606 of these patients have non-missing LDL levels. The sample mean LDL level for patients who survived at least 5 years is 127.2 mg/dL, and the sample standard deviation for these patients is 32.93. There are 121 patients who died within 5 years, and 119 of these patients have non-missing LDL levels. The sample mean LDL level for patients who died within 5 years is 118.7 mg/dL, and the sample standard deviation for these patients is 36.16. The sample means are dissimilar in magnitude, since the sample mean for patients who survived at least 5 years is 8.5 mg/dL greater than the sample mean for patients who died within 5 years. The sample standard deviations are also dissimilar in magnitude, since the sample standard deviation for patients who died within 5 years is 9.8% greater than the sample standard deviation for patients who survived at least 5 years.

B. The point estimate for the true mean LDL in a population of similar subjects who would survive at least 5 years is 127.2 mg/dL, with a standard error of 1.338. A 95% confidence interval suggests this point estimate would not be unusual if the true in mean LDL level for the subjects who would survive at least 5 years were between 124.6 and 129.8 mg/dL. The point estimate for the true mean LDL in a population of similar subjects who would die within 5 years is 118.7 mg/dL, with a standard error of 3.315. A 95% confidence interval suggests this point estimate would not be unusual if the true in mean LDL level for the subjects who would survive at least 5 years were between 112.1 and 125.3 mg/dL. The point estimates are dissimilar in magnitude, since the point estimate of the mean for patients who would survive at least 5 years is 8.5 mg/dL greater than the point estimate of the mean for patients who would die within 5 years. The standard errors are also dissimilar in magnitude, since the standard error for patients who would die within 5 years is 147.8% greater than the standard error for patients who would survive at least 5 years. Our answer regarding the sample means is exactly the same as that regarding the point estimate of the mean. However, our answers regarding the sample standard deviations and the standard errors are different. Although the patients who died within 5 years have greater sample standard deviations than those who survived greater than 5 years, and the patients who would die within 5 years had greater standard errors than those who would survive greater than 5 years, the percent increase from low to high standard error (147.8%) is much greater than the percent increase from low to high sample standard deviation (9.8%). This is because within-group standard error is equal to the sample standard deviation divided by the square root of the group sample size, so for the group of patients surviving greater than 5 years, their smaller sample standard deviation (32.93) is divided by the square root of their greater sample size (606), whereas for the group of patients surviving less than 5 years, their larger sample standard deviation (36.16) is divided by the square root of their smaller sample size (119), so the percent difference in standard errors is much greater than between standard deviations.

C. The confidence interval for the mean LDL in a population surviving 5 years does overlap with the confidence interval for the mean LDL in a population dying within 5 years, but since neither confidence interval contains the point estimate for the other population, we cannot reach any conclusions about the statistical significance of an estimated difference in the estimated means at a 0.05 level of significance.

D. If we presume the variances are equal in the two populations, but we want to allow for the possibility that the means might be different, the best estimate for the standard deviation of LDL measurements in each group is the pooled standard deviation, which is also known as the root mean squared error, with the equation given on slide 60 of the Lecture 3. The root mean squared error is similar to the square root of a weighted average of the two sample variances, where the sample variance for the group with the greatest sample size has the greater weight.

E. The point estimate for the true difference in mean LDL levels between a population that survives at least 5 years and a population that dies within 5 years is 8.5 mg/dL greater for the population that survives at least 5 years, with a standard error of 3.357. A 95% confidence interval suggests this point estimate would not be unusual if the true difference in mean LDL levels were between 1.9 and 15.1 mg/dL greater for the population that survives at least 5 years. The two-sided p-value testing the hypothesis that the two populations have the same mean LDL is 0.0115. Since this two-sided p-value is less than 0.05, we can with 95% confidence reject the null hypothesis that there is no association between serum LDL levels and 5-year all-cause mortality in favor of the alternative hypothesis that there is a statistically significant association between serum LDL levels and 5-year all-cause mortality.

2. A. Both regression model A and regression model B are saturated, since each involves two parameters (the intercept and the slope) and predictors of interest (indicator variables for death within 5 years and survival greater than or equal to 5 years, respectively) with just two possible values. Since the number of parameters and possible values for the predictor of interest in each model are equal, each group mean can be fit exactly, without the need to borrow information across groups to make inference about group means.

B. Using the estimate for the constant parameter in regression model A, the estimate for the true mean LDL level among a population of subjects who survive at least 5 years is 127.2 mg/dL, exactly the same as the corresponding estimate from problem 1.

C. Using the 95% confidence interval for the estimate of the constant parameter in regression model A, the estimate for the true mean LDL level given in question 2 part B would not be unusual if the true mean LDL level among the population of subjects who survive at least five years were between 124.5 and 129.9 mg/dL. This 95% confidence interval is wider than the corresponding 95% confidence interval given in question 1 part B, which was 124.6 mg/dL to 129.8 mg/dL. This is because in regression models, confidence intervals are constructed using root mean squared error in place of the standard deviation for the specified group. Equations illustrating the differences between within-group standard deviation and root mean squared error can be found on Lecture 3, slide 60.

D. Using the estimate for the constant parameter in regression model B, the estimate for the true mean LDL level among a population of subjects who die within 5 years is 118.7 mg/dL, exactly the same as the corresponding estimate from problem 1.

E. Using the 95% confidence interval for the estimate of the constant parameter in regression model B, the estimate for the true mean LDL level given in question 2 part D would not be unusual if the true mean LDL level among the population of subjects who die within five years were between 112.7 and 124.7 mg/dL. This 95% confidence interval is narrower than the corresponding 95% confidence interval given in question 1 part B, which was 112.1 mg/dL to 125.3 mg/dL. This is because in regression models, confidence intervals are constructed using root mean squared error in place of the standard deviation for the specified group. Equations illustrating the differences between within-group standard deviation and root mean squared error can be found on Lecture 3, slide 60.

F. If we presume the variances are equal in the two populations, the regression-based estimate of the standard deviation within each group for each model is the root mean squared error, also known as the pooled standard deviation, which is exactly the same as the corresponding estimate from problem 1. The equation for the root mean squared error is given on slide 60 of the Lecture 3. The root mean squared error is similar to the square root of a weighted average of the two sample variances, where the sample variance for the group with the greatest sample size has the greater weight.

G. Models A and B are essentially the same model, but with the predictor of interest variables defined differently. The indicator that the subject died within 5 years in model A is equal to 1 if and only if the indicator that the subject survived at least 5 years in model B is equal to 0. Similarly, the indicator that the subject died within 5 years in model A is equal to 0 if and only if the indicator that the subject survived at least 5 years in model B is equal to 1. The scatterplot graph for model A is the same as the scatterplot graph for model B, flipped horizontally across the line where the predictor of interest indicator variable is equal to 0.5.

H. Since the intercept from regression model A is equal to 127.2, and since model A is a saturated model, the sample mean LDL level for the group of subjects that survived at least 5 years is exactly equal to 127.2 mg/dL.

I. Since the slope from regression model A is equal to -8.5, and since model A is a saturated model, the sample mean LDL level for the group of subjects that died within 5 years is exactly equal to 127.2 - 8.5 = 118.7 mg/dL.

J. Using the estimated confidence interval for the slope in regression model B, the point estimate of the difference in mean LDL levels between a population that survives at least 5 years and a population that dies within 5 years is 8.5 mg/dL greater for the population that survives at least 5 years, with an estimated standard error of 3.357. Using the 95% confidence interval for the estimate of the slope parameter in regression model B, this point estimate for the true difference in mean LDL levels would not be unusual if the true difference in mean LDL levels between a population that survives at least 5 years and a population that dies within 5 years were between 1.9 and 15.1 mg/dL greater for the population that survives at least 5 years. This 95% confidence interval is equivalent to the corresponding 95% confidence interval given in question 1 part E.

3. To evaluate a possible association between serum LDL and 5-year all-cause mortality, we compared mean LDL levels across groups defined by vital status at 5 years with a t test that allows for the possibility of unequal variances across groups. We found that the point estimate of the difference in mean LDL levels between the population that survived at least 5 years and the population that died within 5 years remained the same as those from question 1. However, the standard error for the point estimate of the difference in mean LDL levels between these groups is 3.574 without the assumption of equal variances across groups, as opposed to a standard error of 3.357 when across-group variance is presumed to be equal. Given the same point estimate of the true difference in mean LDL levels between these two groups as is given in question 1 part E, a 95% confidence interval suggests this point estimate would not be unusual if the true difference in mean LDL levels between the two groups were between 1.4 and 15.6 mg/dL greater for the group of patients which survived greater than 5 years without the assumption of equal across-group variance. This is a wider 95% confidence interval than that produced in question 1 part E under the presumption of equal across-group variance, which was from 1.9 to 15.1 mg/dL. Also, the two-sided p-value testing the null hypothesis that there is no association between LDL levels and 5-year all-cause mortality is 0.0186, which is still small enough to be statistically significant at the 0.05 significance level, and reject the null hypothesis in favor of the alternative hypothesis that there is an association between LDL levels and 5-year all-cause mortality. This two-sided p-value is larger than that which is produced by the t test presuming equal across-group variances, which is 0.0115.These results of the statistical analyses differ from those produced in question 1 under the presumption of equal across-group variance because the t test that does not presume equal across-group variance attempts to correct for heteroscedasticity. If the variance in LDL levels among the population of all similar subjects who would die within 5 years was greater than the variance in LDL levels among the population of similar subjects who would survive greater than 5 years, as is the case in the sample, then since the group of patients who died within 5 years has a smaller sample size than the group of patients who survived at least 5 years, the t test which presumes equal across-group variance would give anti-conservative inference (see Lecture 3, slide 39). Hence, to correct for this anti-conservative inference, the t test which does not presume equal variance produces a greater standard error for the point estimate of the difference in mean LDL levels between the two groups, as well as a wider confidence interval for the difference in mean LDL levels between the two groups and a larger two-sided p-value testing association between LDL levels and 5-year all-cause mortality.

4. A. To evaluate a possible association between serum LDL and 5-year all-cause mortality, we compared mean LDL levels across groups defined by vital status at 5 years with a linear regression model that allows for the possibility of unequal variances across groups. We found that the point estimate for the true difference in mean LDL levels between the two groups remained the same as the point estimate produced in question 3. However, the standard error for this point estimate of the true difference in mean LDL levels is 3.566, different from the standard error produced in question 3, which is 3.574. Given the point estimate of the true difference in mean LDL levels between these two groups, a 95% confidence interval suggests this point estimate would not be unusual if the true difference in mean LDL levels between the two groups were between 1.5 and 15.5 mg/dL greater for the group of patients which survived greater than 5 years. This confidence interval is narrower than that produced by the t test which allows for the possibility of unequal across-group variances, which is from 1.4 to 15.6 mg/dL. The two-sided p-value testing the null hypothesis that there is no association between LDL levels and 5-year all-cause mortality is 0.017, which is small enough to be statistically significant at the 0.05 significance level, and reject the null hypothesis in favor of the alternative hypothesis that there is an association between LDL levels and 5-year all-cause mortality. This two-sided p-value is smaller than that which is produced by the t test allowing for the possibility of unequal across-group variances, which is 0.0186. These results are different from those produced in question 3 because with linear regression on a binary predictor of interest variable, the use of robust standard errors approximates, but is not equivalent to, the t test which allows for the possibility of unequal variances by using the Huber-White sandwich estimator. This is different from the case of classical linear regression on a binary predictor of interest variable, which is exactly equivalent to the t test that presumes equal variances.

5. A.



Given the scatterplot and lines of best fit for male and female as shown above, we can see that the line of best fit for men is always lower than the line of best fit for women, which signifies that on average in the sample across all ages, the LDL levels of women higher than those of men, up to at most 25 mg/dL higher. Based off the lines of best fit, we can also see that as age increases, on average, LDL levels for women in the sample increase slightly, while LDL levels for men decrease slightly. Due to the fact that these lines of best fit do not intersect and have slopes with different signs, there seems to be a markedly different association between age and LDL level for men and for women in the sample, hence, sex appears to be an effect modifier for the causal pathway with age as the predictor of interest and LDL level as the response. However, since sex is associated with LDL levels in the sample, even if sex is causally associated with serum LDL in the population, we cannot treat sex as a confounder.

B. To address the question of an association between LDL and age, we fit a linear regression model with robust standard error estimates to LDL level and age, where LDL level is our response and age is our predictor of interest.

C. This is not a saturated model, because the linear regression model only has two parameters (the intercept and the slope), but the predictor of interest can take on any of the 35 integer values from 65 to 99.

D. Based on our regression model, the estimated mean LDL level among a population of 70 year old subjects is 126.2 mg/dL. This answer is equal to the intercept plus 70 times the slope, since the slope is the average difference in LDL level for each one year difference in age. It is important to note that since question 5 part C stated that this is not a saturated model, we are borrowing information across groups to estimate the group mean for 70 year old subjects, and therefore not obtaining the exact sample mean for 70 year old subjects.

E. Based on our regression model, the estimated mean LDL level among a population of 71 year old subjects is 126.1 mg/dL. This answer is equal to the answer to question 5 part D plus the slope, since the slope is the average difference in LDL level for each one year difference in age. It is important to note that since question 5 part C stated that this is not a saturated model, we are borrowing information across groups to estimate the group mean for 71 year old subjects, and therefore not obtaining the exact sample mean for 71 year old subjects.

F. Based on our regression model, the estimated mean LDL level among a population of 75 year old subjects is 125.8 mg/dL. This answer is equal to the answer to question 5 plus 5 times the slope, since the slope is the average difference in LDL level for each one year difference in age. It is important to note that since question 5 part C stated that this is not a saturated model, we are borrowing information across groups to estimate the group mean for 75 year old subjects, and therefore not obtaining the exact sample mean for 75 year old subjects.

G. In this regression model, we can interpret root mean squared error to be approximately the square root of a weighted average of the all within-group sample variances, where the sample variance for groups with the larger sample sizes have the greater weights. Root mean error can also be interpreted as the sample standard deviation of the residuals, which are estimates for the error of the regression model in predicting LDL levels.

H. The intercept is equal to 132.5, which means that when a subject’s age is 0, that is, for newborns, the serum LDL level is 132.5 mg/dL. However, this is not scientifically relevant, since the study which provided the data which was fit to this linear regression model only included subjects aged 65 and older. This means that there are no nearby data points for this inference about newborns to borrow information from, so any inference about the LDL levels of newborns (or anyone outside the range of ages from which data was collected) is not very useful.

I. The slope is equal to -0.1, which means that, in the sample age range of 65 to 99, for each one year difference in age, the subjects in the sample who were one year older would have, on average, mean LDL levels 0.1 mg/dL less than the subjects in the sample who were one year younger.

J. Our model for linear regression with robust standard errors between LDL levels and age produces a point estimate for the average difference in LDL levels between groups that differ in age by one year. This point estimate indicates that for each one year difference in age, the subjects in the sample who were one year older would have, on average, mean LDL levels 0.1 mg/dL less than the subjects in the sample who were one year younger. A 95% confidence interval for the average difference in LDL levels between groups that differ in age by one year suggests that this point estimate would not be unusual if the true difference in mean LDL levels between these groups was between 0.5 mg/dL less and 0.4 mg/dL greater for the group that was one year older. The two-sided p-value testing the null hypothesis that there is no association between LDL levels and age is 0.698, which is not small enough to be statistically significant at the 0.05 significance level. Thus, we cannot reject the null hypothesis in favor of the alternative hypothesis that there is an association between LDL levels and age.

K. If we wanted a point estimate and confidence interval for the difference in mean LDL across groups that differ by 5 years in age, we would report that our point estimate indicates that for each five year difference in age, the subjects in the sample who were five years older would have, on average, mean LDL levels 0.5 mg/dL less than the subjects in the sample who were five years younger. A 95% confidence interval for the average difference in LDL levels between groups that differ in age by five years suggests that this point estimate would not be unusual if the true difference in mean LDL levels between these groups was between 2.7 mg/dL less and 1.8 mg/dL greater for the group that was five years older.We can produce this point estimate by multiplying the point estimate for the difference in LDL levels for the one year difference in age by 5. This holds because if a new random variable is the same as another random variable multiplied by 5, its mean is 5 times the original mean. We can obtain the confidence interval by taking the difference of the original point estimate and one of the endpoints of the original 95% confidence interval, and multiplying this difference by 5. This new difference must be added and subtracted from the new point estimate described earlier to obtain the new 95% confidence interval. This holds because the difference between a point estimate and the endpoints of its 95% confidence interval is a product of the sample standard deviation. If a new random variable is the same as another random variable multiplied by 5, the variance of the new random variable is 25 times the original variance, and its standard deviation is 5 times that of the original variable.

L. Using the R-squared coefficient produced by our linear regression model, and from the evidence that on average, subjects with higher ages have lower LDL levels, we find that our Pearson’s R coefficient for the correlation between age and LDL levels is -0.014. This indicates a slight negative correlation between age and LDL levels for subjects in our sample age range. The two-sided p-value testing the null hypothesis that there is no correlation between LDL levels and age is 0.698, which is not small enough to be statistically significant at the 0.05 significance level. Thus, we cannot reject the null hypothesis in favor of the alternative hypothesis that there is correlation between LDL levels and age.