**Homework #7**

1. A. Methods: Using large sample approximations and the fact that sample means are approximately normally distributed, we tested to see if mean cholesterol levels are associated with sex in Caucasians. We calculated a point estimate of the effect of sex on cholesterol by taking the difference of sample mean cholesterol levels, subtracting that of the male group from that of the female group. Assuming that the mean cholesterol levels in females and males are independent, we estimated the standard error for the difference in mean cholesterol by taking the square root of the sum of the squares of the individual standard errors for the males and females. We created a Z-score to test the null hypothesis of no difference by dividing the estimate of the difference in sample means by the estimate for the standard error of this difference. This Z-score was approximately normally distributed with mean 0 and variance 1, so we found a two-sided p-value using STATA by multiplying the smallest area to the left or the right of the Z-score under the standard normal distribution by 2. We finally computed a 95% confidence interval for the difference in mean cholesterol levels: the estimate of the difference in sample means plus or minus the product of 1.96 (the 97.5th percentile of the standard normal distribution) and the estimate for the standard error of this difference.Results: Mean cholesterol levels were found to be 25.3 mg/dl higher in Caucasian females than in Caucasisan males. Such a difference was found sufficiently extreme to be able to rule out a null hypothesis of no difference in mean cholesterol across groups of Caucasian patients defined by sex (two-sided p-value less than 10^(-15)). Based on a 95% confidence interval, we find that the observed difference in mean cholesterol is not atypical of settings in which the true difference in mean cholesterol were such that Caucasian females had mean cholesterol anywhere from 22.26 mg/dl to 28.34 mg/dl higher than Caucasian males (computed using an estimated standard error for the difference of 1.55).

B. Methods: Using large sample approximations and the fact that sample means are approximately normally distributed, we tested to see if mean cholesterol levels are associated with sex in Non-Caucasians. We calculated a point estimate of the effect of sex on cholesterol by taking the difference of sample mean cholesterol levels, subtracting that of the male group from that of the female group. Assuming that the mean cholesterol levels in females and males are independent, we estimated the standard error for the difference in mean cholesterol by taking the square root of the sum of the squares of the individual standard errors for the males and females. We created a Z-score to test the null hypothesis of no difference by dividing the estimate of the difference in sample means by the estimate for the standard error of this difference. This Z-score was approximately normally distributed with mean 0 and variance 1, so we found a two-sided p-value using STATA by multiplying the smallest area to the left or the right of the Z-score under the standard normal distribution by 2. We finally computed a 95% confidence interval for the difference in mean cholesterol levels: the estimate of the difference in sample means plus or minus the product of 1.96 (the 97.5th percentile of the standard normal distribution) and the estimate for the standard error of this difference.Results: Mean cholesterol levels were found to be 15.7 mg/dl higher in Non-Caucasian females than in Non-Caucasisan males. Such a difference was found sufficiently extreme to be able to rule out a null hypothesis of no difference in mean cholesterol across groups of Non-Caucasian patients defined by sex (two-sided p-value of 5.458\*10^(-6)). Based on a 95% confidence interval, we find that the observed difference in mean cholesterol is not atypical of settings in which the true difference in mean cholesterol were such that Non-Caucasian females had mean cholesterol anywhere from 8.93 mg/dl to 22.47 mg/dl higher than Non-Caucasian males (computed using an estimated standard error for the difference of 3.45).

C. Methods: Using large sample approximations and the fact that sample means are approximately normally distributed, we tested to see if mean cholesterol levels are associated with sex after adjustment for race. We calculated a point estimate of the effect of sex on cholesterol after adjustment for race by taking a weighted average of the point estimates for the difference of sample mean cholesterol levels found in parts A and B of this problem, where we subtracted the sample mean cholesterol of the male group from that of the female group. Assuming that the mean cholesterol levels in females and males are independent, we estimated the standard error for the effect of sex on cholesterol after adjustment for race by taking the square root of a weighted sum of the squares of the estimated standard errors of the difference in mean cholesterol levels for the males and females by race found in parts A and B of this problem. The weights used in this estimation of the standard error were the respective squares of the weights used in calculating the point estimates. We created a Z-score to test the null hypothesis of no difference between females and males by dividing the estimate of the effect of sex on cholesterol after adjustment for race by the estimate for the standard error of this effect. This Z-score was approximately normally distributed with mean 0 and variance 1, so we found a two-sided p-value using STATA by multiplying the smallest area to the left or the right of the Z-score under the standard normal distribution by 2. We finally computed a 95% confidence interval for the difference in mean cholesterol levels after adjustment for race: the estimate of the effect of sex on cholesterol after adjustment for race plus or minus the product of 1.96 (the 97.5th percentile of the standard normal distribution) and the estimate for the standard error of this effect.

Results: We first used importance weights which were proportional to the distribution of race within our sample, where the weight for the Caucasian estimates was 0.8405 and the weight for the Non-Caucasian estimates was 0.1595. After adjustment for race, mean cholesterol levels were found to be 23.77 mg/dl higher in females than in males of the same race. Such a difference was sufficiently extreme to be able to rule out a null hypothesis of no difference in mean cholesterol across groups defined by sex (two-sided p-value less than 10^(-15)). Based on a 95% confidence interval, we find that the observed difference in mean cholesterol is not atypical of settings in which the true difference in mean cholesterol were such that females had mean cholesterol anywhere from 20.99 mg/dl to 26.54 mg/dl higher than males of the same race (computed using an estimated standard error for the difference of 1.42). We next used importance weights which were proportional to the distribution of race according to the US Census, where the weight for the Caucasian estimates was 0.8637 and the weight for the Non-Caucasian estimates was 0.1363. After adjustment for race, mean cholesterol levels were found to be 23.99 mg/dl higher in females than in males of the same race. Such a difference was sufficiently extreme to be able to rule out a null hypothesis of no difference in mean cholesterol across groups defined by sex (two-sided p-value less than 10^(-15)). Based on a 95% confidence interval, we find that the observed difference in mean cholesterol is not atypical of settings in which the true difference in mean cholesterol were such that females had mean cholesterol anywhere from 21.21 mg/dl to 26.78 mg/dl higher than males of the same race (computed using an estimated standard error for the difference of 1.42). We finally used efficiency weights which were proportional to the inverse of the square of the estimated standard errors for the difference in mean cholesterol between females and males by race (as calculated in parts A and B of this question), where the weight for the Caucasian estimates was 0.8319372128 and the weight for the Non-Caucasian estimates was 0.1680627872. After adjustment for race, mean cholesterol levels were found to be 23.69 mg/dl higher in females than in males of the same race. Such a difference was sufficiently extreme to be able to rule out a null hypothesis of no difference in mean cholesterol across groups defined by sex (two-sided p-value less than 10^(-15)). Based on a 95% confidence interval, we find that the observed difference in mean cholesterol is not atypical of settings in which the true difference in mean cholesterol were such that females had mean cholesterol anywhere from 20.91 mg/dl higher to 26.46 mg/dl higher than males of the same race (computed using an estimated standard error for the difference of 1.42).

D. Methods: Using large sample approximations and the fact that sample means are approximately normally distributed, we tested for evidence of effect modification by race in the association between cholesterol level and sex. We calculated a point estimate of the effect of race on the association between cholesterol level and sex by taking the difference, Caucasian estimate minus Non-Caucasian estimate, of the differences in mean cholesterol levels by sex that were calculated in parts A and B of this problem, female estimate minus male estimate within each race category. Assuming that the mean cholesterol levels in females and males are independent and that the differences in mean cholesterol by race are independent, we estimated the standard error for the effect of race on the association between cholesterol level and sex by taking the square root of the sum of the squares of the estimated standard errors of the difference in mean cholesterol levels for the males and females by race found in parts A and B of this problem. We created a Z-score to test the null hypothesis of no effect modification by race in the association between cholesterol level and sex by dividing the estimate of the effect of race on the association between cholesterol level and sex by the estimate for the standard error of this effect. This Z-score was approximately normally distributed with mean 0 and variance 1, so we found a two-sided p-value using STATA by multiplying the smallest area to the left or the right of the Z-score under the standard normal distribution by 2. We finally computed a 95% confidence interval for the effect of race on the association between cholesterol level and sex: the estimate of the effect of race on the association between cholesterol level and sex plus or minus the product of 1.96 (the 97.5th percentile of the standard normal distribution) and the estimate for the standard error of this effect.Results: The difference in mean cholesterol across groups defined by sex was found to be 9.6 mg/dl lower in Non-Caucasians than in Caucasians. Such a difference was sufficiently extreme to be able to rule out with high confidence a null hypothesis of no effect modification by race in the association between cholesterol level and sex (two-sided p-value of 0.01122524). Based on a 95% confidence interval, we find that the observed difference in the association between cholesterol and sex across the race groups is not atypical of settings in which the true difference in effect were such that Non-Caucasians had a mean difference in cholesterol across sex groups anywhere from 2.18 mg/dl to 17.02 mg/dl lower than that in Caucasians (computed using an estimated standard error of 3.79 for the estimated interaction contrast).

2. A. Methods: Using large sample approximations and the fact that sample means are approximately normally distributed, we tested to see if mean fibrinogen levels are associated with sex in Caucasians. We calculated a point estimate of the effect of sex on fibrinogen by taking the difference of sample mean fibrinogen levels, subtracting that of the male group from that of the female group. Assuming that the mean fibrinogen levels in females and males are independent, we estimated the standard error for the difference in mean fibrinogen by taking the square root of the sum of the squares of the individual standard errors for the males and females. We created a Z-score to test the null hypothesis of no difference by dividing the estimate of the difference in sample means by the estimate for the standard error of this difference. This Z-score was approximately normally distributed with mean 0 and variance 1, so we found a two-sided p-value using STATA by multiplying the smallest area to the left or the right of the Z-score under the standard normal distribution by 2. We finally computed a 95% confidence interval for the difference in mean fibrinogen levels: the estimate of the difference in sample means plus or minus the product of 1.96 (the 97.5th percentile of the standard normal distribution) and the estimate for the standard error of this difference.Results: Mean fibrinogen levels were found to be 2.9 mg/dl higher in Caucasian females than in Caucasisan males. Such a difference was insufficiently extreme to be able to rule out a null hypothesis of no difference in mean fibrinogen across groups of Caucasian patients defined by sex (two-sided p-value of 0.27869698). Based on a 95% confidence interval, we find that the observed difference in mean fibrinogen is not atypical of settings in which the true difference in mean fibrinogen were such that Caucasian females had mean fibrinogen anywhere from -2.35 mg/dl lower to 8.15 mg/dl higher than Caucasian males (computed using an estimated standard error for the difference of 2.68).

B. Methods: Using large sample approximations and the fact that sample means are approximately normally distributed, we tested to see if mean fibrinogen levels are associated with sex in Non-Caucasians. We calculated a point estimate of the effect of sex on fibrinogen by taking the difference of sample mean fibrinogen levels, subtracting that of the male group from that of the female group. Assuming that the mean fibrinogen levels in females and males are independent, we estimated the standard error for the difference in mean fibrinogen by taking the square root of the sum of the squares of the individual standard errors for the males and females. We created a Z-score to test the null hypothesis of no difference by dividing the estimate of the difference in sample means by the estimate for the standard error of this difference. This Z-score was approximately normally distributed with mean 0 and variance 1, so we found a two-sided p-value using STATA by multiplying the smallest area to the left or the right of the Z-score under the standard normal distribution by 2. We finally computed a 95% confidence interval for the difference in mean fibrinogen levels: the estimate of the difference in sample means plus or minus the product of 1.96 (the 97.5th percentile of the standard normal distribution) and the estimate for the standard error of this difference.Results: Mean fibrinogen levels were found to be 15.7 mg/dl higher in Non-Caucasian females than in Non-Caucasisan males. Such a difference was found sufficiently extreme to be able to rule out a null hypothesis of no difference in mean fibrinogen across groups of Non-Caucasian patients defined by sex (two-sided p-value of 0.03140841). Based on a 95% confidence interval, we find that the observed difference in mean fibrinogen is not atypical of settings in which the true difference in mean fibrinogen were such that Non-Caucasian females had mean fibrinogen anywhere from 1.40 mg/dl to 30.00 mg/dl higher than Non-Caucasian males (computed using an estimated standard error for the difference of 7.30).

C. Methods: Using large sample approximations and the fact that sample means are approximately normally distributed, we tested to see if mean fibrinogen levels are associated with sex after adjustment for race. We calculated a point estimate of the effect of sex on fibrinogen after adjustment for race by taking a weighted average of the point estimates for the difference of sample mean fibrinogen levels found in parts A and B of this problem, where we subtracted the sample mean fibrinogen of the male group from that of the female group. Assuming that the mean fibrinogen levels in females and males are independent, we estimated the standard error for the effect of sex on fibrinogen after adjustment for race by taking the square root of a weighted sum of the squares of the estimated standard errors of the difference in mean fibrinogen levels for the males and females by race found in parts A and B of this problem. The weights used in this estimation of the standard error were the respective squares of the weights used in calculating the point estimates. We created a Z-score to test the null hypothesis of no difference between females and males by dividing the estimate of the effect of sex on fibrinogen after adjustment for race by the estimate for the standard error of this effect. This Z-score was approximately normally distributed with mean 0 and variance 1, so we found a two-sided p-value using STATA by multiplying the smallest area to the left or the right of the Z-score under the standard normal distribution by 2. We finally computed a 95% confidence interval for the difference in mean fibrinogen levels after adjustment for race: the estimate of the effect of sex on fibrinogen after adjustment for race plus or minus the product of 1.96 (the 97.5th percentile of the standard normal distribution) and the estimate for the standard error of this effect.Results: We first used importance weights which were proportional to the distribution of race within our sample, where the weight for the Caucasian estimates was 0.8405 and the weight for the Non-Caucasian estimates was 0.1595. After adjustment for race, mean fibrinogen levels were found to be 4.94 mg/dl higher in females than in males of the same race. Such a difference was insufficiently extreme to be able to rule out a null hypothesis of no difference in mean fibrinogen across groups defined by sex (two-sided p-value of 0.0510921). Based on a 95% confidence interval, we find that the observed difference in mean fibrinogen is not atypical of settings in which the true difference in mean fibrinogen were such that females had mean fibrinogen anywhere from 0.02 mg/dl lower to 9.91 mg/dl higher than males of the same race (computed using an estimated standard error for the difference of 2.53). We next used importance weights which were proportional to the distribution of race according to the US Census, where the weight for the Caucasian estimates was 0.8637 and the weight for the Non-Caucasian estimates was 0.1363. After adjustment for race, mean fibrinogen levels were found to be 4.64 mg/dl higher in females than in males of the same race. Such a difference was insufficiently extreme to be able to rule out a null hypothesis of no difference in mean fibrinogen across groups defined by sex (two-sided p-value of 0.06499406). Based on a 95% confidence interval, we find that the observed difference in mean fibrinogen is not atypical of settings in which the true difference in mean fibrinogen were such that females had mean fibrinogen anywhere from 0.29 mg/dl lower to 9.58 mg/dl higher than males of the same race (computed using an estimated standard error for the difference of 2.52). We finally used efficiency weights which were proportional to the inverse of the square of the estimated standard errors for the difference in mean fibrinogen between females and males by race (as calculated in parts A and B of this question), where the weight for the Caucasian estimates was 0.8813386939 and the weight for the Non-Caucasian estimates was 0.1186613061. After adjustment for race, mean fibrinogen levels were found to be 4.42 mg/dl higher in females than in males of the same race. Such a difference was insufficiently extreme to be able to rule out a null hypothesis of no difference in mean fibrinogen across groups defined by sex (two-sided p-value of 0.078712). Based on a 95% confidence interval, we find that the observed difference in mean fibrinogen is not atypical of settings in which the true difference in mean fibrinogen were such that females had mean fibrinogen anywhere from 0.51 mg/dl lower to 9.34 mg/dl higher than males of the same race (computed using an estimated standard error for the difference of 2.51).

D. Methods: Using large sample approximations and the fact that sample means are approximately normally distributed, we tested for evidence of effect modification by race in the association between fibrinogen level and sex. We calculated a point estimate of the effect of race on the association between fibrinogen level and sex by taking the difference, Caucasian estimate minus Non-Caucasian estimate, of the differences in mean fibrinogen levels by sex that were calculated in parts A and B of this problem, female estimate minus male estimate within each race category. Assuming that the mean fibrinogen levels in females and males are independent and that the differences in mean fibrinogen by race are independent, we estimated the standard error for the effect of race on the association between fibrinogen level and sex by taking the square root of the sum of the squares of the estimated standard errors of the difference in mean fibrinogen levels for the males and females by race found in parts A and B of this problem. We created a Z-score to test the null hypothesis of no effect modification by race in the association between fibrinogen level and sex by dividing the estimate of the effect of race on the association between fibrinogen level and sex by the estimate for the standard error of this effect. This Z-score was approximately normally distributed with mean 0 and variance 1, so we found a two-sided p-value using STATA by multiplying the smallest area to the left or the right of the Z-score under the standard normal distribution by 2. We finally computed a 95% confidence interval for the effect of race on the association between fibrinogen level and sex: the estimate of the effect of race on the association between fibrinogen level and sex plus or minus the product of 1.96 (the 97.5th percentile of the standard normal distribution) and the estimate for the standard error of this effect.Results: The difference in mean fibrinogen across groups defined by sex was found to be 12.8 mg/dl higher in Non-Caucasians than in Caucasians. Such a difference was insufficiently extreme to be able to rule out with high confidence a null hypothesis of no effect modification by race in the association between fibrinogen level and sex (two-sided p-value of 0.09955643). Based on a 95% confidence interval, we find that the observed difference in the association between fibrinogen and sex across the race groups is not atypical of settings in which the true difference in effect were such that Non-Caucasians had a mean difference in fibrinogen across sex groups anywhere from 2.43 mg/dl lower to 28.03 mg/dl higher than that in Caucasians (computed using an estimated standard error of 7.77 for the estimated interaction contrast).

3. A. Note that of the 5000 patients enrolled in this study, 47 have missing cholesterol levels. Among the 4953 patients with non-missing cholesterol levels, the standard deviation for cholesterol is 39.28814, so this is our best estimate of the standard deviation within the sample, both at baseline and after two years of treatment, since we only have access to this single cross-sectional measurement.

B. According to the equation from Lecture 8 Slide 40, the best estimate of the standard deviation for the difference of these two paired cholesterol measurements is:sqrt((39.28814^2)+(39.28814^2)-(2\*0.4\* 39.28814\* 39.28814)) = sqrt(2\*(39.28814^2)\*(1-0.4)) = 39.28814 \* sqrt(1.2) = 43.03800104.

C. From the RMSE produced by either classical or robust linear regression with cholesterol as the response variable and sex and a continuously coded age variable as the predictors of interest, the best estimate for standard deviation of cholesterol within groups that have constant age and sex is 37.492.

4. A. If N denotes the necessary sample size providing 80% power to detect the design alternative, sigma = 37.492, rho = 0.4, alpha = 0.025, beta = 0.8, delta = -10, zbeta = 0.84162123, and z(1-alpha) = 1.959964, we calculate N as follows:N = ceiling[((zbeta+z(1-alpha))^2)\*(8\*(sigma^2)\*(1-rho))/((delta)^2)] = ceiling[((0.84162123 + 1.959964)^2)\*(8\*(37.492 ^2)\*(1-0.4))/((-10)^2)] = 530.

B. If N denotes the necessary sample size providing 90% power to detect the design alternative, sigma = 37.492, rho = 0.4, alpha = 0.025, beta = 0.9, delta = -10, zbeta = 1.2815516, and z(1-alpha) = 1.959964, we calculate N as follows:N = ceiling[((zbeta+z(1-alpha))^2)\*(8\*(sigma^2)\*(1-rho))/((delta)^2)] = ceiling[((1.2815516 + 1.959964)^2)\*(8\*(37.492 ^2)\*(1-0.4))/((-10)^2)] = 709.

C. If N denotes the necessary sample size providing 90% power to detect the design alternative, sigma = 39.28814, rho = 0.4, alpha = 0.025, beta = 0.9, delta = -10, zbeta = 1.2815516, and z(1-alpha) = 1.959964, we calculate N as follows:N = ceiling[((zbeta+z(1-alpha))^2)\*(8\*(sigma^2)\*(1-rho))/((delta)^2)] = ceiling[((1.2815516 + 1.959964)^2)\*(8\*(39.28814 ^2)\*(1-0.4))/((-10)^2)] = 779.

D. By Lecture 8 slides 50 and 52, the sample size computation would give a smaller required sample size if we used the final measurements only (throwing away the baseline measurements) because the correlation of 0.4 is less than 0.5. This is because, while the other terms in the sample size calculation formula would remain the same, using only the final measurements would give V = 4\*(sigma^2) = 4\*(39.28814^2) = 6174.23, while using the change would give V = 8\*(sigma^2)\*(1-rho) = 8\*(39.28814^2)\*(1-0.4) = 7409.08, so the sample size for using the change would be larger than the sample size for using the final measurements only.

E. By Lecture 8 slides 51 and 52, the sample size computation would give a smaller required sample size if we used an ANCOVA model that adjusted for age, sex, and the baseline cholesterol level than if we used either the final measurements only or the change. This is because, while the other terms in the sample size calculation formula would remain the same, using the ANCOVA model that adjusted for age, sex, and the baseline cholesterol level would give V = 4\*(sigma^2)\*(1-(rho^2)) = 4\*(39.28814^2)\*(1-(0.4^2)) = 5186.35, while, by part D of this exercise, using only the final measurements would give V = 6174.23 and using the change would give V = 7409.08. Hence, the sample size for the ANCOVA model that adjusted for age, sex, and the baseline cholesterol level would be smaller than for the model including change or final measurements only.

5. A. The best estimate of the proportion of patients on the control arm who have serum cholesterol below 200 mg/dL after 2 years of treatment is 0.3957198, the proportion of patients in the CHS inflamm.txt data set who have serum cholesterol below 200 mg/dL.

B. The best estimate of the proportion of patients on the treatment arm who have serum cholesterol below 200 mg/dL after 2 years of treatment is 0.4942459, the proportion of patients in the CHS inflamm.txt data set who have serum cholesterol below 210 mg/dL.

C. If N denotes the necessary sample size providing 90% power to detect the design alternative, alpha = 0.025, beta = 0.9, delta = 0.3957198-0.4942459 = -0.0985261, zbeta = 1.2815516, z(1-alpha) = 1.959964, Pt = 0.4942459, and Pc = 0.3957198, we calculate N as follows:N = ceiling[((zbeta+z(1-alpha))^2)\*(2\*(Pt\*(1-Pt)+Pc\*(1-Pc)))/((delta)^2)] = ceiling[((1.2815516 + 1.959964)^2)\*(2\*(0.4942459\*(1-0.4942459) + 0.3957198\*(1-0.3957198)))/((0.4942459-0.3957198)^2)] = 1059.

D. To the extent that it is clinically most important to lower serum cholesterol levels below 200 mg/dL, this study design answers the most relevant scientific question. However, if such a scientific threshold did not exist, then we have clearly lost information about how the new treatment might tend to lower serum cholesterol levels across the population. This loss of information is reflected in the higher sample size requirements when dichotomizing the data: 1059 versus 709.