**Homework #3**

Total grade: 68(+6)/108

Grade for Q1: 17/25

1. A. This is a saturated regression model, because two distinct groups (those who have serum LDL greater than or equal to 160 mg/dL and those who have serum LDL less than 160 mg/dL) are modeled with two regression parameters (the intercept and the slope).

Grade: 3/3

B. For subjects with serum LDL less than 160 mg/dL, the estimated odds of dying within 5 years are 0.205, and hence the corresponding estimated probability of dying within 5 years is 0.170. This probability is exactly the same as the observed proportion of subjects with serum LDL less than 160 mg/dL dying within 5 years.

Grade: 2/3. You need to explain why they are exactly the same—the model is saturated.

C. For subjects with serum LDL greater than or equal to 160 mg/dL, the estimated odds of dying within 5 years are 0.151, and hence the corresponding estimated probability of dying within 5 years is 0.131. This probability is exactly the same as the observed proportion of subjects with serum LDL greater than or equal to 160 mg/dL dying within 5 years.

Grade: 2/3. Same as above

D. Methods: The odds of subjects dying within 5 years of study enrollment were compared between subjects with serum LDL greater than or equal to 160 mg/dL and subjects with serum LDL less than 160 mg/dL. Using logistic regression methods with robust standard errors, we estimated the odds ratio, for which the odds of 5 year mortality for subjects with serum LDL greater than or equal to 160 mg/dL are divided by the odds of 5 year mortality for subjects with serum LDL less than 160 mg/dL. We also created a Wald-based 95% confidence interval for this odds ratio point estimate and tested the null hypothesis that the odds ratio was equal to 1, which is equivalent to saying that the odds of 5 year mortality are not associated with serum LDL levels.Results: For subjects (618 subjects) with serum LDL greater than or equal to 160 mg/dL, the estimated odds of 5 year mortality are 0.151, and for subjects with serum LDL less than 160 mg/dL, the estimated odds of 5 year mortality are 0.205. From logistic regression analysis, we estimate that in the group with serum LDL greater than or equal to 160 mg/dL, the odds of 5 year mortality are 26.5% lower than those in the group with serum LDL less than 160 mg/dL, corresponding to an odds ratio point estimate of 0.735. A Wald-based 95% confidence interval suggests that this observation would not be unusual if in the group with serum LDL greater than or equal to 160 mg/dL, the odds of 5 year mortality were anywhere from 59.6% lower to 34.0% higher than those in the group with serum LDL less than 160 mg/dL, corresponding to an odds ratio confidence interval of 0.404 to 1.340. A Wald-based two-sided p value of 0.316 suggests that we cannot with high confidence reject the null hypothesis that the odds of 5 year mortality are not associated with serum LDL levels. The estimated odds of 5 year mortality in each serum LDL group and the point estimate for the odds ratio are identical to those presented in Problem 6 from Homework 1. However, the exact confidence interval of 0.373 to 1.360 for the odds ratio from Homework 1, Problem 6, is wider than the Wald-based confidence interval of 0.404 to 1.340 for the odds ratio (The same with the Woolf CI) presented in this homework. Also, the Fisher’s exact p value of 0.396 from Homework 1, Problem 6 is larger than the Wald-based p value of 0.316 calculated for this homework. (You need to discuss more about different p-values and their similarities and differences as in the key)

Grade: 7/10. Extended discuss about different p-values is needed. Do not use robust option for saturated model.

E. If the logistic regression model had an indicator of 5 year mortality as the response variable, but had an indicator of LDL less than 160 mg/dL as the predictor of interest, this would be a reparameterization of the logistic regression model analyzed in parts A through C of this question. The answer to part A of this question would remain the same, because as two distinct groups (those who have serum LDL less than 160 mg/dL and those who have serum LDL greater than or equal to 160 mg/dL) are modeled with two regression parameters (the intercept and the slope), this is still a saturated regression model. Similarly, the answers to parts B and C of this question would also remain the same. If the logistic regression model had an indicator of 5 year mortality as the response variable, but had an indicator of LDL less than 160 mg/dL as the predictor of interest, this would be a reparameterization of the logistic regression model analyzed in parts A through C of this question. The answer to part A of this question would remain the same, because as two distinct groups (those who have serum LDL greater than or equal to 160 mg/dL and those who have serum LDL less than 160 mg/dL) are modeled with two regression parameters (the intercept and the slope), this is still a saturated regression model. The answers to parts B and C of this question would also remain the same, and in each case, the odds of 5 year mortality would be obtained by taking the multiplicative inverse of the odds of 5 year survival that are given naturally by this recoded logistic regression model.

Grade: 2/3. Saturation is not the key reason. The key reason is linear transformation of the predictor and a binary response variable

F. If we fit a logistic regression model with an indicator of LDL greater than or equal to 160 mg/dL as the response variable and an indicator of 5 year mortality as the predictor of interest, this would reverse the roles these variables had in parts A through C in this question. The answer to part A of this question would remain the same, because as two distinct groups (those who died within 5 years and those who survived greater than 5 years) are modeled with two regression parameters (the intercept and the slope), this is still a **saturated regression model**. This logistic regression model will give us the estimated odds of serum LDL greater than or equal to 160 mg/dL for those who died within 5 years and those who survived greater than 5 years, and we can use these odds to calculate the corresponding estimated probabilities of serum LDL greater than or equal to 160 mg/dL for those who died within 5 years and for those who survived greater than 5 years, respectively. Since this is a saturated model, we can enter these conditional probabilities and the observed proportions of 5 year mortality and serum LDL greater than or equal to 160 mg/dL into Bayes’ Formula to find the conditional estimated probabilities of 5 year mortality given serum LDL greater than or equal to 160 mg/dL or given serum LDL less than 160 mg/dL. From these estimated conditional probabilities, we can then find the estimated odds that are asked for in parts B and C of this question. Upon completion of these steps, we found that the answers to parts B and C of this question using this logistic regression model were exactly the same as those found using the original logistic regression model.

Grade: 1/3. The key is the invariance properties of the odds ratios. odds(outcome|exposure)= odds(exposure|outcome). But the intercept and the fitted values will not be comparable.

Grade for Q2: 15/25

2. A. This is a saturated regression model, because two distinct groups (those who have high serum LDL, greater than or equal to 160 mg/dL, and those who have low serum LDL, less than 160 mg/dL) are modeled with two regression parameters (the intercept and the slope).

Grade: 2/3. You need to point out that it is a linear regression.

B. For subjects with serum LDL less than 160 mg/dL, the estimated probability of dying within 5 years is 0.170, and hence the corresponding estimated odds of dying within 5 years are 0.205. This probability is exactly the same as the observed proportion of subjects with serum LDL less than 160 mg/dL dying within 5 years.

Grade: 2/3. Provide some explanation for the results. Saturated linear regression.

C. For subjects with serum LDL greater than or equal to 160 mg/dL, the estimated probability of dying within 5 years is 0.131, and hence the corresponding estimated odds of dying within 5 years are 0.151. This probability is exactly the same as the observed proportion of subjects with serum LDL greater than or equal to 160 mg/dL dying within 5 years.

Grade: 2/3. Same as above

D. Methods: The probability of subjects dying within 5 years of study enrollment was compared between subjects with serum LDL greater than or equal to 160 mg/dL and subjects with serum LDL less than 160 mg/dL. Using linear regression methods with robust standard errors, we estimated the risk difference, for which the probability of 5 year mortality for subjects with serum LDL greater than or equal to 160 mg/dL is subtracted from the probability of 5 year mortality for subjects with serum LDL less than 160 mg/dL. We also created a Wald-based 95% confidence interval for this risk difference point estimate and tested the null hypothesis that the risk difference was equal to 0, which is equivalent to saying that the probability of 5 year mortality is not associated with serum LDL levels.Results: For subjects with serum LDL greater than or equal to 160 mg/dL, the estimated probability of 5 year mortality is 0.170, and for subjects with serum LDL less than 160 mg/dL, the estimated probability of 5 year mortality is **0.151** (This seems wrong 0.131). From linear regression analysis, we estimate that in the group with serum LDL greater than or equal to 160 mg/dL, the probability of 5 year mortality is an absolute difference of 0.0391 lower than in the group with serum LDL less than 160 mg/dL. A Wald-based 95% confidence interval suggests that this risk difference point estimate would not be unusual if in the group with serum LDL greater than or equal to 160 mg/dL, there was anywhere from a 0.110 absolute lower probability of 5 year mortality to a 0.0316 absolute higher probability of 5 year mortality than in the group with serum LDL less than 160 mg/dL. A Wald-based two-sided p value of 0.278 suggests that we cannot with high confidence reject the null hypothesis that the probability of 5 year mortality is not associated with serum LDL levels. The estimated probabilities of 5 year mortality in each serum LDL group and the point estimate for the risk difference are identical to those presented in Problem 5 from Homework 1. However, the exact confidence interval of anywhere from a 0.109 absolute lower probability of 5 year mortality to a 0.0314 absolute higher probability of 5 year mortality for the risk difference from Homework 1, Problem 6, is narrower than the Wald-based confidence interval of anywhere from a 0.110 absolute lower probability of 5 year mortality to a 0.0316 absolute higher probability of 5 year mortality for the risk difference presented in this homework (Why?). Also, the two-sided exact p value of 0.314 from Homework 1, Problem 6 is larger than the Wald-based p value of 0.278 calculated for this homework.

Grade: 7/10. Wrong estimated risk for group LDL less than 160 mg/d. Should give sources of the difference observed.

E. If the linear regression model had an indicator of 5 year mortality as the response variable, but had an indicator of LDL less than 160 mg/dL as the predictor of interest, this would be a reparameterization of the linear regression model analyzed in parts A through C of this question. The answer to part A of this question would remain the same, because as two distinct groups (those who have serum LDL less than 160 mg/dL and those who have serum LDL greater than or equal to 160 mg/dL) are modeled with two regression parameters (the intercept and the slope), this is still a saturated regression model. Similarly, the answers to parts B and C of this question would also remain the same. If the linear regression model had an indicator of 5 year mortality as the response variable, but had an indicator of LDL less than 160 mg/dL as the predictor of interest, this would be a reparameterization of the linear regression model analyzed in parts A through C of this question. The answer to part A of this question would remain the same, because as two distinct groups (those who have serum LDL greater than or equal to 160 mg/dL and those who have serum LDL less than 160 mg/dL) are modeled with two regression parameters (the intercept and the slope), this is still a saturated regression model. The answers to parts B and C of this question would also remain the same, and in each case, the probabilities of 5 year mortality would be obtained by taking 1 minus the additive inverse of the probabilities of 5 year survival that are given naturally by this recoded linear regression model.

Grade: 2/3. The key reason is linear transformation of the predictor.

F. If we fit a linear regression model with an indicator of LDL greater than or equal to 160 mg/dL as the response variable and an indicator of 5 year mortality as the predictor of interest, this would reverse the roles these variables had in parts A through C in this question. The answer to part A of this question **would remain the same** (this is not true), because as two distinct groups (those who died within 5 years and those who survived greater than 5 years) are modeled with two regression parameters (the intercept and the slope), this is still a saturated regression model. This linear regression model will give us the estimated probabilities of serum LDL greater than or equal to 160 mg/dL for those who died within 5 years and those who survived greater than 5 years. Since this is a saturated model, we can enter these conditional probabilities and the observed proportions of 5 year mortality and serum LDL greater than or equal to 160 mg/dL into Bayes’ Formula to find the conditional estimated probabilities of 5 year mortality given serum LDL greater than or equal to 160 mg/dL or given serum LDL less than 160 mg/dL, which are asked for in parts B and C of this question. **Upon completion of these steps, we found that the answers to parts B and C of this question using this linear regression model were exactly the same as those found using the original linear regression model**. (wrong)

Grade: 0/3. The answer is wrong, the results of the two models are no longer the same.

Grade for Q3: 12/25

3. A. This is a saturated regression model, because two distinct groups (those who have high serum LDL, greater than or equal to 160 mg/dL, and those who have low serum LDL, less than 160 mg/dL) are modeled with two regression parameters (the intercept and the slope).

Grade: 2/3. You need to point out what regression is this

B. For subjects with serum LDL less than 160 mg/dL, the estimated probability of dying within 5 years is 0.170, and hence the corresponding estimated odds of dying within 5 years are 0.205. This probability is exactly the same as the observed proportion of subjects with serum LDL less than 160 mg/dL dying within 5 years.

Grade: 2/3. Provide some explanation for the results. Saturated linear regression.

C. For subjects with serum LDL greater than or equal to 160 mg/dL, the estimated probability of dying within 5 years is 0.131, and hence the corresponding estimated odds of dying within 5 years are 0.151. This probability is exactly the same as the observed proportion of subjects with serum LDL greater than or equal to 160 mg/dL dying within 5 years.

Grade: 2/3. Same as above

D. Methods: The probability of subjects dying within 5 years of study enrollment was compared between subjects with serum LDL greater than or equal to 160 mg/dL and subjects with serum LDL less than 160 mg/dL. Using Poisson regression methods with robust standard errors, we estimated the risk ratio, for which the probability of 5 year mortality for subjects with serum LDL greater than or equal to 160 mg/dL is divided by the probability of 5 year mortality for subjects with serum LDL less than 160 mg/dL. We also created a Wald-based 95% confidence interval for this risk ratio point estimate and tested the null hypothesis that the risk ratio was equal to 1, which is equivalent to saying that the probability of 5 year mortality is not associated with serum LDL levels.Results: For subjects with serum LDL greater than or equal to 160 mg/dL, the estimated probability of 5 year mortality is 0.170, and for subjects with serum LDL less than 160 mg/dL, the estimated probability of 5 year mortality is **0.151** (wrong). From Poisson regression analysis, we estimate that in the group with serum LDL greater than or equal to 160 mg/dL, the probability of 5 year mortality is 23.0% lower than in the group with serum LDL less than 160 mg/dL, corresponding to a risk ratio point estimate of 0.770. A Wald-based 95% confidence interval calculated with Poisson regression suggests that this risk ratio point estimate would not be unusual if the probability of 5 year mortality in the group with serum LDL greater than or equal to 160 mg/dL was anywhere from 54.1% lower to 29.4% higher than in the group with serum LDL less than 160 mg/dL, corresponding to a risk ratio confidence interval of 0.458 to 1.294. A Wald-based two-sided p value of 0.324 (P-value seems wrong) suggests that we cannot with high confidence reject the null hypothesis that the probability of 5 year mortality is not associated with serum LDL levels. The estimated probabilities of 5 year mortality in each serum LDL group and the point estimate for the risk ratio are identical to those calculated in Problem 5 from Homework 1. However, the exact confidence interval for the risk ratio of 0.459 to 1.293 calculated in Problem 5 from Homework 1 is narrower than the Wald-based confidence interval for the risk ratio of 0.458 to 1.294 presented in this homework. Also, the two-sided exact p value of 0.314 calculated in Homework 1, Problem 5 is smaller than the Wald-based p value of 0.324 calculated for this homework.

Grade: 6/10. Wrong estimated risk of death and P-value. Should give sources of the difference observed.

E. If the Poisson regression model had an indicator of 5 year mortality as the response variable, but had an indicator of LDL less than 160 mg/dL as the predictor of interest, this would be a reparameterization of the Poisson regression model analyzed in parts A through C of this question. The answer to part A of this question would remain the same, because as two distinct groups (those who have serum LDL less than 160 mg/dL and those who have serum LDL greater than or equal to 160 mg/dL) are modeled with two regression parameters (the intercept and the slope), this is still a saturated regression model. Similarly, the answers to parts B and C of this question would also remain the same. If the Poisson regression model had an indicator of 5 year mortality as the response variable, but had an indicator of LDL less than 160 mg/dL as the predictor of interest, this would be a reparameterization of the Poisson regression model analyzed in parts A through C of this question. The answer to part A of this question would remain the same, because as two distinct groups (those who have serum LDL greater than or equal to 160 mg/dL and those who have serum LDL less than 160 mg/dL) are modeled with two regression parameters (the intercept and the slope), this is still a saturated regression model. The answers to parts B and C of this question would also remain the same, and in each case, the probabilities of 5 year mortality would be obtained by taking 1 minus the additive inverse of the probabilities of 5 year survival that are given naturally by this recoded linear regression model.

Grade: 0/3. Look at the key, for Poisson regression, the results are not comparable any more.

F. If we fit a Poisson regression model with an indicator of LDL greater than or equal to 160 mg/dL as the response variable and an indicator of 5 year mortality as the predictor of interest, this would reverse the roles these variables had in parts A through C in this question. The answer to part A of this question would remain the same, because as two distinct groups (those who died within 5 years and those who survived greater than 5 years) are modeled with two regression parameters (the intercept and the slope), this is still a saturated regression model. This Poisson regression model will give us the estimated probabilities of serum LDL greater than or equal to 160 mg/dL for those who died within 5 years and those who survived greater than 5 years. Since this is a saturated model, we can enter these conditional probabilities and the observed proportions of 5 year mortality and serum LDL greater than or equal to 160 mg/dL into Bayes’ Formula to find the conditional estimated probabilities of 5 year mortality given serum LDL greater than or equal to 160 mg/dL or given serum LDL less than 160 mg/dL, which are asked for in parts B and C of this question. Upon completion of these steps, we found that the answers to parts B and C of this question using this Poisson regression model were exactly the same as those found using the original Poisson regression model.

Grade: 0/3. Look at the key, for Poisson regression, the results are not comparable any more.

Grade for Q4: 24+6/33

4. A. Methods: The probability of subjects dying within 5 years of study enrollment was compared across groups of subjects defined by a continuous measure of serum LDL level. Using **linear regression methods with robust standard errors**, we estimated the average risk difference, for which the probability of 5 year mortality for subjects with a higher serum LDL level is subtracted from the probability of 5 year mortality for subjects with a serum LDL level 1 mg/dL less than that in the higher group. We also created a Wald-based 95% confidence interval for this risk difference point estimate and tested the null hypothesis that the risk difference was equal to 0, which is equivalent to saying that the probability of 5 year mortality is not associated with serum LDL levels.Results: From linear regression analysis, we estimate that in the group with a higher serum LDL level, the probability of 5 year mortality is an average absolute difference of **0.1034%** lower than in the group with a serum LDL level 1 mg/dL less than that in the higher group. A Wald-based 95% confidence interval suggests that this risk difference point estimate would not be unusual if in the group with a higher serum LDL level, the probability of 5 year mortality was an average absolute difference of **0.1884% to 0.0185%** lower than in the group with a serum LDL level 1 mg/dL less than that in the higher group. A Wald-based two-sided p value of 0.017 suggests that at the 0.05 significance level, we can with high confidence reject the null hypothesis that the probability of 5 year mortality is not associated with serum LDL levels in favor of the alternative hypothesis that the probability of 5 year mortality is associated with serum LDL levels.

Grade: 9/10. Great. Please also provide some descriptive statistics on the sample.

B. Methods: The probability of subjects dying within 5 years of study enrollment was compared across groups of subjects defined by a continuous measure of serum LDL level. Using Poisson regression methods with robust standard errors, we estimated the average risk ratio, for which the probability of 5 year mortality for subjects with a higher serum LDL level is divided by the probability of 5 year mortality for subjects with a serum LDL level 1 mg/dL less than that in the higher group. We also created a Wald-based 95% confidence interval for this risk ratio point estimate and tested the null hypothesis that the risk ratio was equal to 1, which is equivalent to saying that the probability of 5 year mortality is not associated with serum LDL levels.Results: From Poisson regression analysis, we estimate that in the group with a higher serum LDL level, the probability of 5 year mortality is an average of **0.6448% (0.626%?)** lower than in the group with a serum LDL level 1 mg/dL less than that in the higher group, corresponding to a risk ratio point estimate of **0.9936 (0.937?).** A Wald-based 95% confidence interval suggests that this risk ratio point estimate would not be unusual if in the group with a higher serum LDL level, the probability of 5 year mortality was an average of **1.175% (1.11%) to 0.1123% (0.11%)** lower than in the group with a serum LDL level 1 mg/dL less than that in the higher group, corresponding to a risk ratio confidence interval of 0.9883 to 0.9989. A Wald-based two-sided p value of 0.018 suggests that at the 0.05 significance level, we can with high confidence reject the null hypothesis that the probability of 5 year mortality is not associated with serum LDL levels in favor of the alternative hypothesis that the probability of 5 year mortality is associated with serum LDL levels.

Grade: 6 (+3)/10. The estimates are slightly off comparing to the keys (But I think maybe the keys are wrong…I am not sure… if the keys are wrong, you get 9 points for this question). Provide some descriptive statistics on the sample

C. Methods: The odds of subjects dying within 5 years of study enrollment were compared across groups of subjects defined by a continuous measure of serum LDL level. Using logistic regression methods with robust standard errors, we estimated the average odds ratio, for which the odds of 5 year mortality for subjects with a higher serum LDL level are divided by the odds of 5 year mortality for subjects with a serum LDL level 1 mg/dL less than that in the higher group. We also created a Wald-based 95% confidence interval for this odds ratio point estimate and tested the null hypothesis that the odds ratio was equal to 1, which is equivalent to saying that the odds of 5 year mortality are not associated with serum LDL levels.Results: From logistic regression analysis, we estimate that in the group with a higher serum LDL level, the odds of 5 year mortality are an average of **0.7744%** lower than in the group with a serum LDL level 1 mg/dL less than that in the higher group, corresponding to an odds ratio point estimate of **0.9923**. A Wald-based 95% confidence interval suggests that this odds ratio point estimate would not be unusual if in the group with a higher serum LDL level, the odds of 5 year mortality were an average **of 1.419% to 0.1255%** lower than in the group with a serum LDL level 1 mg/dL less than that in the higher group, corresponding to an odds ratio confidence interval of 0.9858 to 0.9987. A Wald-based two-sided p value of 0.019 suggests that at the 0.05 significance level, we can with high confidence reject the null hypothesis that the odds of 5 year mortality are not associated with serum LDL levels in favor of the alternative hypothesis that the odds of 5 year mortality are associated with serum LDL levels.

Grade: 6(+3)/10. Same as above…

D. Each of the models presented in parts A through C in this question treat 5 year mortality as the response variable and a continuous measure of serum LDL levels as the predictor of interest, and each finds that with high confidence, in the group with higher serum LDL levels, the probability or odds of 5 year mortality are lower than in the group with serum LDL levels 1mg/dL less than the higher group. Each of the primary models in questions 1 through 3 of this homework treat 5 year mortality as the response variable and a dichotomized measure of serum LDL levels at 160 mg/dL as the predictor of interest, and each finds that we cannot with high confidence reject the null hypothesis that the probability or odds of 5 year mortality are not associated with serum LDL levels. Questions 2 and 4 from Homework 2 each treat a continuous measure of serum LDL levels as the response variable and 5 year mortality as the predictor of interest, and each find that with high confidence, patients who survive at least 5 years have a higher mean LDL level than those who do not. A priori, I would prefer one of the models presented in parts A through C in this question from this homework. One major reason is that, unlike the primary models from questions 1 through 3 of this homework, the models in parts A through C in this question from this homework, as well as the models from questions 2 and 4 of Homework 2, treat serum LDL level as a continuous variable instead of dichotomizing it at 160 mg/dL. This dichotomization, though scientifically based since it is based on the recommendations from the Mayo Clinic website, still causes the analysis to be less informative by forcing serum LDL measurements to lose their inherently continuous quality. I also would prefer to treat 5 year mortality as the response variable and a continuous measure of serum LDL levels, unlike the analyses in questions 2 and 4 of Homework 2, because this approach is more clinically relevant and more informative in helping patients understand how probability or odds of 5 year mortality are influenced by serum LDL levels. Of the analyses presented in parts A through C of this problem on this homework, I would most likely choose the analysis approach presented in part A. This is because 5 year mortality is not typically viewed as a rare event in the age range of 65 to 99 that this study samples from, and hence the benefits of interpretation of the risk ratio and odds ratio are not as great. Also, since this is effectively a cross-sectional study with respect to 5 year mortality and the continuous measure of serum LDL levels, we need not concern ourselves about whether the study participants were sampled by exposure or outcome, so the benefits of the odds ratio for case-control studies are small. Additionally, the risk difference is inherently simple to interpret, and thus is better suited to improve public health, as this study is attempting to do. However, if I were concerned about reducing the amount of effect modification that is present in my analysis, I would most likely choose the analysis approach presented in part C, using the odds ratio to evaluate an association between 5 year mortality and serum LDL levels.

Grade: 3/3. Great!