This homework builds on the analyses performed in homeworks #1 and #2, As such, all questions relate to associations among death from any cause, serum low density lipoprotein (LDL) levels, age, and sex in a population of generally healthy elderly subjects in four U.S. communities. This homework uses the subset of information that was collected to examine MRI changes in the brain. The data can be found on the class web page (follow the link to Datasets) in the file labeled mri.txt. Documentation is in the file mri.pdf. See homework #1 for additional information.

1. Perform a statistical regression analysis evaluating an association between serum LDL and 5 year all-cause mortality by comparing the odds of death within 5 years across groups defined by whether the subjects have high serum LDL (“high” = LDL > 160mg/dL). In your regression model, use an indicator of death within 5 years as your response variable, and use an indicator of high LDL as your predictor. (Only give a formal report of the inference where asked to.)
	1. Is this a saturated regression model? Explain your answer.

**In this regression model two distinct groups, those who have high serum LDL levels (LDL ≥160mg/dL) and those who have normal/low serum LDL levels (LDL < 160mg/dL), are modeled with two regression parameters, the intercept and the slope. Because the number of modeled groups is equivalent to the number of regression parameters this is a saturated model.**

* 1. For subjects with low LDL, what is the estimated odds of dying within 5 years? What is the estimated probability of dying within 5 years? How do these estimates compare to the observed proportion of subjects with low LDL dying within 5 years?

**The estimated odds of dying within 5 years for a population of individuals with low serum LDL (LDL < 160mg/dL) is given by 0.2047, which is the exponentiated intercept of the logistic regression model. The estimated probability of dying within 5 years for a population of individuals with low serum LDL is given by the following relationship:**

$$prob= \frac{odds}{1+odds}$$

**The estimated probability of death within 5 years for a population of individuals with low serum LDL is 0.1699 (16.99%). The observed proportion of subjects with low LDL dying within 5 years is 0.1699 (16.99%). The observed proportion is equivalent to the estimated probability of death within 5 years for a population of individuals with low serum LDL. However, the estimated odds of death within 5 years for subjects with low LDL is not equivalent to either the estimated or the observed proportion of subjects with low LDL who died within 5 years. This is because the odds is only a good estimate of the probability when the probability is very small.**

* 1. For subjects with high LDL, what is the estimated odds of dying within 5 years? What is the estimated probability of dying within 5 years? How do these estimates compare to the observed proportion of subjects with low LDL dying within 5 years?

**The estimated odds of dying within 5 years for a population of individuals with high serum LDL (LDL ≥ 160mg/dL) is given by 0.1505, which is the exponentiated sum of the slope and the intercept of the logistic regression model. The estimated probability of dying within 5 years for a population of individuals with high serum LDL is given by the following relationship:**

$$prob= \frac{odds}{1+odds}$$

**The estimated probability of death within 5 years for a population of individuals with high serum LDL is 0.1308 (13.08%). The observed proportion of subjects with low LDL (LDL < 160mg/dL) dying within 5 years is 0.1699 (16.99%), which is higher than the estimated probability of subjects with high serum LDL dying within 5 years by a relative difference of roughly 26% or an absolute difference of 3.91%. The estimated odds of death within 5 years for a population of individuals with high serum LDL is different from both the estimated probability of death within 5 years for a population of individuals with high serum LDL and the observed proportion of subjects with low serum LDL dying within 5 years.**

* 1. Give full inference regarding the association between 5 year mortality and high LDL levels. How does this differ from the inference that was made on problems 5 and 6 of homework #1? What is the source of any differences?

**From a logistic regression analysis of 725 subjects, we estimate that for a population of individuals with low serum LDL levels (LDL < 160mg/dL) the odds of 5 year mortality is 0.2046 and for a population of individuals with high serum LDL levels (LDL ≥160mg/dL) the odds of 5 year mortality is 0.1505. From an odds ratio of 0.7355 the odds of death within 5 years is estimated as 26.45% less for a population with high serum LDL levels than a population with low serum LDL levels. A 95% confidence interval suggests that this observation is not unusual if the odds of 5 year mortality for a population with high serum LDL was anywhere from 59.64% lower to 34.04% higher than the odds of 5 year mortality for a population with low serum LDL. Using a two-sided Wald p-value this observation was not statistically significant at a 0.05 level of significance (p = 0.3158), and we fail to reject the null hypothesis that the odds of death within 5 years is not associated with serum LDL levels.**

**The point estimates of the odds and the odds ratio are the same as those given in problem 6 of homework 1. However, the 95% confidence interval and the p-value differ between the two problems, 0.373 to 1.36 and p=0.396 for homework 1 and 0.404 to 1.34 and p=0.3158 for this model. This is because the CI and p-value were obtained using Fisher’s exact test for homework 1 and a Wald test was used in this regression model. However, the inference made that the observations were not statistically significant at a 0.05 level of significance and that we fail to reject the null hypothesis that the odds of death within 5 years is not associated with serum LDL levels was the same for both problems.**

* 1. How would the answers to parts a-c change if I had instead asked you to fit a logistic regression model using the indicator of death within 5 years as your response variable, but using an indicator of low LDL as your predictor? What if we had used an indicator of survival for at least 5 years as the response variable?

**Indicator for low LDL as the predictor:**

**This model is a reparametarization of the model used in parts a-c. Therefore there are still two distinct groups, those who have high serum LDL levels (≥160mg/dL) and those who have normal/low serum LDL levels (< 160mg/dL) that are modeled with two regression parameters, the intercept and the slope. Because the number of modeled groups is equivalent to the number of regression parameters this is also a saturated model.**

**The estimated odds of dying within 5 years for a population of individuals with low serum LDL (LDL < 160mg/dL) is given by 0.2047, which is the exponentiated sum of the slope and the intercept of the logistic regression model. The estimated probability of dying within 5 years for a population of individuals with low serum LDL is given by the following relationship:**

$$prob= \frac{odds}{1+odds}$$

**The estimated probability of death within 5 years for a population of individuals with low serum LDL is 0.1699 (16.99%). The observed proportion of subjects with low LDL dying within 5 years is 0.1699 (16.99%). The observed proportion is equivalent to the estimated probability of death within 5 years for a population of individuals with low serum LDL. However, the estimated odds of death within 5 years for subjects with low LDL is not equivalent to either the estimated or the observed proportion of subjects with low LDL who died within 5 years. This is because the odds is only a good estimate of the probability when the probability is very small. This is exactly the answer that was obtained for part b, when an indicator for high LDL was used as the predictor.**

**The estimated odds of dying within 5 years for a population of individuals with high serum LDL (LDL ≥ 160mg/dL) is given by 0.1505, which is the exponentiated intercept of the logistic regression model. The estimated probability of dying within 5 years for a population of individuals with high serum LDL is given by the following relationship:**

$$prob= \frac{odds}{1+odds}$$

**The estimated probability of death within 5 years for a population of individuals with high serum LDL is 0.1308 (13.08%). The observed proportion of subjects with low LDL (LDL < 160mg/dL) dying within 5 years is 0.1699 (16.99%), which is higher than the estimated probability of subjects with high serum LDL dying within 5 years by a relative difference of roughly 26% or an absolute difference of 3.91%. The estimated odds of death within 5 years for a population of individuals with high serum LDL is different from both the estimated probability of death within 5 years for a population of individuals with high serum LDL and the observed proportion of subjects with low serum LDL dying within 5 years. This is exactly the answer that was obtained for part c, when an indicator for high LDL was used as the predictor.**

 **Indicator of Survival for at least 5 years as the response variable:**

**This model uses the same predictor as the model used in parts a-c, only the response variable has been reparameterized. Therefore there are still the same two distinct groups, those who have high serum LDL levels (≥160mg/dL) and those who have normal/low serum LDL levels (< 160mg/dL) that are modeled with two regression parameters, the intercept and the slope. Because the number of modeled groups is equivalent to the number of regression parameters this is also a saturated model.**

**The estimated odds of dying within 5 years for a population of individuals with low serum LDL (LDL < 160mg/dL) is given by 0.2047, which is given by the following relationship:**

$$odds\left(low LDL\right)= \frac{1+odds(survival|low)}{odds(survival|low)}-1$$

**The odds of surviving at least 5 years for a population of individuals with low serum LDL was estimated as 4.886, which is the exponentiated intercept of the logistic regression model. The estimated probability of dying within 5 years for a population of individuals with low serum LDL is given by the following relationship:**

$$prob= \frac{odds}{1+odds}$$

**The estimated probability of death within 5 years for a population of individuals with low serum LDL is 0.1699 (16.99%). The observed proportion of subjects with low LDL dying within 5 years is 0.1699 (16.99%). The observed proportion is equivalent to the estimated probability of death within 5 years for a population of individuals with low serum LDL. However, the estimated odds of death within 5 years for subjects with low LDL is not equivalent to either the estimated or the observed proportion of subjects with low LDL who died within 5 years. This is because the odds is only a good estimate of the probability when the probability is very small. This is exactly the answer that was obtained for part b, when an indicator for high LDL was used as the predictor.**

**The estimated odds of dying within 5 years for a population of individuals with high serum LDL (LDL ≥ 160mg/dL) is given by 0.1505, which is given by the following relationship:**

$$odds\left(high LDL\right)= \frac{1+odds(survival|high)}{odds(survival|high)}-1$$

**The odds of surviving at least 5 years for a population of individuals with high serum LDL was estimated as 6.6429, which is the exponentiated sum of the slope and the intercept of the logistic regression model. The estimated probability of dying within 5 years for a population of individuals with high serum LDL is given by the following relationship:**

$$prob= \frac{odds}{1+odds}$$

**The estimated probability of death within 5 years for a population of individuals with high serum LDL is 0.1308 (13.08%). The observed proportion of subjects with low LDL (LDL < 160mg/dL) dying within 5 years is 0.1699 (16.99%), which is higher than the estimated probability of subjects with high serum LDL dying within 5 years by a relative difference of roughly 26% or an absolute difference of 3.91%. The estimated odds of death within 5 years for a population of individuals with high serum LDL is different from both the estimated probability of death within 5 years for a population of individuals with high serum LDL and the observed proportion of subjects with low serum LDL dying within 5 years. This is exactly the answer that was obtained for part c, when an indicator for high LDL was used as the predictor.**

* 1. In parts a-d of this problem, we described the distribution of death within 5 years across groups defined by LDL level. What if we fit a logistic regression model mimicking the approach used in problems 1 – 4 of homework #2, where we described the distribution of LDL across groups defined by vital status? How would our answers to parts a-c change?

**Because the same sample is used for this regression model and the model used in parts a-c the estimated odds of death within 5 years for a population with low serum LDL will be the same as will the estimated odds of death within 5 years for a population with high serum LDL. The observed proportion of subjects with low serum LDL (LDL < 160mg/dL) will also be the same for both models. This model uses an indicator of death within 5 years as the predictor and an indicator of high LDL as the response variable. This means that the interpretation of the outputs as the odds of LDL level given vital status instead of the odds of vital status given LDL level. It is difficult to relate those quantities, so while I expect the answers to parts a-c to stay the same directly observing this is complicated using these model parameters.**

1. Perform a statistical regression analysis evaluating an association between serum LDL and 5 year all-cause mortality by comparing the differences in the probability of death within 5 years across groups defined by whether the subjects have high serum LDL (“high” = LDL > 160mg/dL). In your regression model, use an indicator of death within 5 years as your response variable, and use an indicator of high LDL as your predictor. (Only give a formal report of the inference where asked to.)
	1. Is this a saturated regression model? Explain your answer.

**In this regression model two distinct groups, those who have high serum LDL levels (LDL ≥160mg/dL) and those who have normal/low serum LDL levels (LDL < 160mg/dL), are modeled with two regression parameters, the intercept and the slope. Because the number of modeled groups is equivalent to the number of regression parameters this is a saturated model.**

* 1. For subjects with low LDL, what is the estimated probability of dying within 5 years? What is the estimated odds of dying within 5 years? How do these estimates compare to the observed proportion of subjects with low LDL dying within 5 years?

**The estimated probability of dying within 5 years for a population of individuals with low serum LDL (LDL < 160mg/dL) is given by 0.1699 (16.99%), which is the intercept of the linear regression model. The estimated odds of dying within 5 years for a population of individuals with low serum LDL is given by the following relationship:**

$$odds= \frac{prob}{1-prob}$$

**The estimated odds of death within 5 years for a population of individuals with low serum LDL is 0.2046. The observed proportion of subjects with low LDL dying within 5 years is 0.1699 (16.99%). The observed proportion is equivalent to the estimated probability of death within 5 years for a population of individuals with low serum LDL. However, the estimated odds of death within 5 years for subjects with low LDL is not equivalent to either estimated or the observed proportion of subjects with low LDL who died within 5 years. This is because the odds is only a good estimate of the probability when the probability is very small.**

* 1. For subjects with high LDL, what is the estimated probability of dying within 5 years? What is the estimated odds of dying within 5 years? How do these estimates compare to the observed proportion of subjects with low LDL dying within 5 years?

**The estimated probability of dying within 5 years for a population of individuals with high serum LDL (LDL ≥ 160mg/dL) is given by 0.1308 (13.08%), which is the sum of the slope and the intercept of the linear regression model. The estimated odds of dying within 5 years for a population of individuals with high serum LDL is given by the following relationship:**

$$odds= \frac{prob}{1-prob}$$

**The estimated odds of death within 5 years for a population of individuals with high serum LDL is 0.1505. The observed proportion of subjects with low LDL (LDL < 160mg/dL) dying within 5 years is 0.1699 (16.99%), which is higher than the estimated probability of subjects with high serum LDL dying within 5 years by a relative difference of roughly 26% or an absolute difference of 3.91%. The estimated odds of death within 5 years for a population of individuals with high serum LDL is different from both the estimated probability of death within 5 years for a population of individuals with high serum LDL and the observed proportion of subjects with low serum LDL dying within 5 years.**

* 1. Give full inference regarding the association between 5 year mortality and high LDL levels. How does this differ from the inference that was made on problems 5 and 6 of homework #1? What is the source of any differences?

**From a linear regression analysis of 725 subjects, we estimate that for a population of individuals with low serum LDL levels (LDL < 160mg/dL) the risk of 5 year mortality is 16.99% and for a population of individuals with high serum LDL (LDL ≥ 160mg/dL) the risk of 5 year mortality is 13.08%. From an absolute risk difference of 3.91% the risk of death within 5 years is estimated as being less for a population with high serum LDL levels than a population with low serum LDL levels. A 95% confidence interval suggests that this observation is not unusual if the true absolute risk difference for 5 year mortality was anywhere between 10.97% lower to 3.16% higher for a population that has high serum LDL than for a population with low serum LDL. Using a two-sided F-test p-value this observation was not statistically significant at a 0.05 level of significance (p = 0.2780), and we fail to reject the null hypothesis that the risk of death within 5 years is not associated with serum LDL levels.**

**The point estimates of the probabilities and the risk difference are the same as those given in problem 5 of homework 1. However, the 95% confidence interval and the p-value differ between the two problems, 10.9% lower to 3.14% higher and p=0.314 for homework 1 and 10.97% lower to 3.16% higher and p=0.2780 for this model. This is because the CI and p-value were obtained using a chi squared test in homework 1 and an F-test, which is equivalent to a squared t-statistic for this univariate model, was used in this problem. However, the inference made that the observations were not statistically significant at a 0.05 level of significance and that we fail to reject the null hypothesis that the risk of death within 5 years is not associated with serum LDL levels was the same for both problems.**

* 1. How would the answers to parts a-c change if I had instead asked you to fit a regression model using the indicator of death within 5 years as your response variable, but using an indicator of low LDL as your predictor? What if we had used an indicator of survival for at least 5 years as the response variable?

**Indicator for low LDL as the predictor:**

**This model is a reparametarization of the model used in parts a-c. Therefore there are still two distinct groups, those who have high serum LDL levels (≥160mg/dL) and those who have normal/low serum LDL levels (< 160mg/dL) that are modeled with two regression parameters, the intercept and the slope. Because the number of modeled groups is equivalent to the number of regression parameters this is also a saturated model.**

**The estimated probability of dying within 5 years for a population of individuals with low serum LDL (LDL < 160mg/dL) is given by 0.1699 (16.99%), which is the sum of the slope and the intercept of the linear regression model. The estimated odds of dying within 5 years for a population of individuals with low serum LDL is given by the following relationship:**

$$odds= \frac{prob}{1-prob}$$

**The estimated odds of death within 5 years for a population of individuals with low serum LDL is 0.2046. The observed proportion of subjects with low LDL dying within 5 years is 0.1699 (16.99%). The observed proportion is equivalent to the estimated probability of death within 5 years for a population of individuals with low serum LDL. However, the estimated odds of death within 5 years for subjects with low LDL is not equivalent to either estimated or the observed proportion of subjects with low LDL who died within 5 years. This is because the odds is only a good estimate of the probability when the probability is very small. This is exactly the answer that was obtained for part b, when an indicator for high LDL was used as the predictor.**

**The estimated probability of dying within 5 years for a population of individuals with high serum LDL (LDL ≥ 160mg/dL) is given by 0.1308 (13.08%), which is the intercept of the linear regression model. The estimated odds of dying within 5 years for a population of individuals with high serum LDL is given by the following relationship:**

$$odds= \frac{prob}{1-prob}$$

**The estimated odds of death within 5 years for a population of individuals with high serum LDL is 0.1505. The observed proportion of subjects with low LDL (LDL < 160mg/dL) dying within 5 years is 0.1699 (16.99%), which is higher than the estimated probability of subjects with high serum LDL dying within 5 years by a relative difference of roughly 26% or an absolute difference of 3.91%. The estimated odds of death within 5 years for a population of individuals with high serum LDL is different from both the estimated probability of death within 5 years for a population of individuals with high serum LDL and the observed proportion of subjects with low serum LDL dying within 5 years. This is exactly the answer that was obtained for part c, when an indicator for high LDL was used as the predictor.**

 **Indicator of Survival for at least 5 years as the response variable:**

**This model uses the same predictor as the model used in parts a-c, only the response variable has been reparameterized. Therefore there are still the same two distinct groups, those who have high serum LDL levels (≥160mg/dL) and those who have normal/low serum LDL levels (< 160mg/dL) that are modeled with two regression parameters, the intercept and the slope. Because the number of modeled groups is equivalent to the number of regression parameters this is also a saturated model.**

**The estimated probability of dying within 5 years for a population of individuals with low serum LDL (LDL < 160mg/dL) is given by 0.1699 (16.99%), which is given by the following relationship:**

$$prob\left(low LDL\right)=1 -prob(surval|lowLDL)$$

**The probability of surviving at least 5 years for a population of individuals with low serum LDL was estimated as 0.8301 (83.01%) which is the intercept of the linear regression model. The estimated odds of dying within 5 years for a population of individuals with low serum LDL is given by the following relationship:**

$$odds= \frac{prob}{1-prob}$$

**The estimated odds of death within 5 years for a population of individuals with low serum LDL is 0.2047. The observed proportion of subjects with low LDL dying within 5 years is 0.1699 (16.99%). The observed proportion is equivalent to the estimated probability of death within 5 years for a population of individuals with low serum LDL. However, the estimated odds of death within 5 years for subjects with low LDL is not equivalent to either the estimated or the observed proportion of subjects with low LDL who died within 5 years. This is because the odds is only a good estimate of the probability when the probability is very small. This is exactly the answer that was obtained for part b, when an indicator for high LDL was used as the predictor.**

**The estimated probability of dying within 5 years for a population of individuals with high serum LDL (LDL ≥ 160mg/dL) is given by 0.1308 (13.08%), which is given by the following relationship:**

$$prob\left(high LDL\right)=1 -prob(surval|high LDL)$$

**The probability of surviving at least 5 years for a population of individuals with high serum LDL was estimated as 0.8692 (86.92%) which is the sum of the slope and the intercept of the linear regression model. The estimated odds of dying within 5 years for a population of individuals with high serum LDL is given by the following relationship:**

$$odds= \frac{prob}{1-prob}$$

**The estimated odds of death within 5 years for a population of individuals with high serum LDL is 0.1505. The observed proportion of subjects with low LDL (LDL < 160mg/dL) dying within 5 years is 0.1699 (16.99%), which is higher than the estimated probability of subjects with high serum LDL dying within 5 years by a relative difference of roughly 26% or an absolute difference of 3.91%. The estimated odds of death within 5 years for a population of individuals with high serum LDL is different from both the estimated probability of death within 5 years for a population of individuals with high serum LDL and the observed proportion of subjects with low serum LDL dying within 5 years. This is exactly the answer that was obtained for part c, when an indicator for high LDL was used as the predictor.**

* 1. In parts a-d of this problem, we described the distribution of death within 5 years across groups defined by LDL level. What if we fit a regression model mimicking the approach used in problems 1 – 4 of homework #2, where we described the distribution of LDL across groups defined by vital status? How would our answers to parts a-c change?

**Because the same sample is used for this regression model and the model used in parts a-c the estimated probabilities of death within 5 years for a population with low serum LDL will be the same as will the estimated probabilities of death within 5 years for a population with high serum LDL. The observed proportion of subjects with low serum LDL (LDL < 160mg/dL) will also be the same for both models. This model uses an indicator of death within 5 years as the predictor and an indicator of high LDL as the response variable. This means that the interpretation of the outputs is as the probability of LDL level given vital status instead of the probability of vital status given LDL level. It is difficult to relate those quantities, so while I expect the answers to parts a-c to stay the same directly observing this is complicated using these model parameters.**

1. Perform a statistical regression analysis evaluating an association between serum LDL and 5 year all-cause mortality by comparing the ratios of the probability of death within 5 years across groups defined by whether the subjects have high serum LDL (“high” = LDL > 160mg/dL). In your regression model, use an indicator of death within 5 years as your response variable, and use an indicator of high LDL as your predictor. (Only give a formal report of the inference where asked to.)
	1. Is this a saturated regression model? Explain your answer.

**In this regression model two distinct groups, those who have high serum LDL levels (LDL ≥160mg/dL) and those who have normal/low serum LDL levels (LDL < 160mg/dL), are modeled with two regression parameters, the intercept and the slope. Because the number of modeled groups is equivalent to the number of regression parameters this is a saturated model.**

* 1. For subjects with low LDL, what is the estimated probability of dying within 5 years? What is the estimated odds of dying within 5 years? How do these estimates compare to the observed proportion of subjects with low LDL dying within 5 years?

**The estimated probability of dying within 5 years for a population of individuals with low serum LDL (LDL < 160mg/dL) is given by 0.1699 (16.99%), which is the exponentiated intercept of the Poisson regression model. The estimated odds of dying within 5 years for a population of individuals with low serum LDL is given by the following relationship:**

$$odds= \frac{prob}{1-prob}$$

**The estimated odds of death within 5 years for a population of individuals with low serum LDL is 0.2046. The observed proportion of subjects with low LDL dying within 5 years is 0.1699 (16.99%). The observed proportion is equivalent to the estimated probability of death within 5 years for a population of individuals with low serum LDL. However, the estimated odds of death within 5 years for subjects with low LDL is not equivalent to either estimated or the observed proportion of subjects with low LDL who died within 5 years. This is because the odds is only a good estimate of the probability when the probability is very small.**

* 1. For subjects with high LDL, what is the estimated probability of dying within 5 years? What is the estimated odds of dying within 5 years? How do these estimates compare to the observed proportion of subjects with low LDL dying within 5 years?

**The estimated probability of dying within 5 years for a population of individuals with high serum LDL (LDL ≥ 160mg/dL) is given by 0.1308 (13.08%), which is the exponentiated sum of the slope and the intercept of the Poisson regression model. The estimated odds of dying within 5 years for a population of individuals with high serum LDL is given by the following relationship:**

$$odds= \frac{prob}{1-prob}$$

**The estimated odds of death within 5 years for a population of individuals with high serum LDL is 0.1505. The observed proportion of subjects with low LDL (LDL < 160mg/dL) dying within 5 years is 0.1699 (16.99%), which is higher than the estimated probability of subjects with high serum LDL dying within 5 years by a relative difference of roughly 26% or an absolute difference of 3.91%. The estimated odds of death within 5 years for a population of individuals with high serum LDL is different from both the estimated probability of death within 5 years for a population of individuals with high serum LDL and the observed proportion of subjects with low serum LDL dying within 5 years.**

* 1. Give full inference regarding the association between 5 year mortality and high LDL levels. How does this differ from the inference that was made on problems 5 and 6 of homework #1? What is the source of any differences?

**From a Poisson regression analysis of 725 subjects, we estimate that for a population of individuals with low serum LDL levels (LDL < 160mg/dL) the risk of 5 year mortality is 16.99% and for a population of individuals with high serum LDL (LDL ≥ 160mg/dL) the risk of 5 year mortality is 13.08%. From a risk ratio of 0.7701 the risk of death within 5 years is estimated as being 22.99% less for a population with high serum LDL levels than a population with low serum LDL levels. A 95% confidence interval suggests that this observation is not unusual if the true risk ratio for 5 year mortality was anywhere between 54.16% lower to 29.39% higher for a population that has high serum LDL than for a population with low serum LDL. Using a two-sided Wald p-value this observation was not statistically significant at a 0.05 level of significance (p = 0.3237), and we fail to reject the null hypothesis that the risk of death within 5 years is not associated with serum LDL levels.**

**The point estimates of the probabilities and the risk ratio are the same as those given in problem 5 of homework 1. The risk ratio was not given directly but the ratio of the point estimates can be used to calculate the ratio. However, the 95% confidence interval and the p-value would differ between the two problems. This is because the CI and p-value would have been obtained using a Fisher’s exact test in homework 1 and a Wald test was used in this problem. However, the inference made that the observations were not statistically significant at a 0.05 level of significance and that we fail to reject the null hypothesis that the risk of death within 5 years is not associated with serum LDL levels was the same for both problems.**

* 1. How would the answers to parts a-c change if I had instead asked you to fit a regression model using the indicator of death within 5 years as your response variable, but using an indicator of low LDL as your predictor? What if we had used an indicator of survival for at least 5 years as the response variable?

**Indicator for low LDL as the predictor:**

**This model is a reparametarization of the model used in parts a-c. Therefore there are still two distinct groups, those who have high serum LDL levels (≥160mg/dL) and those who have normal/low serum LDL levels (< 160mg/dL) that are modeled with two regression parameters, the intercept and the slope. Because the number of modeled groups is equivalent to the number of regression parameters this is also a saturated model.**

**The estimated probability of dying within 5 years for a population of individuals with low serum LDL (LDL < 160mg/dL) is given by 0.1699 (16.99%), which is the exponentiated sum of the slope and the intercept of the Poisson regression model. The estimated odds of dying within 5 years for a population of individuals with low serum LDL is given by the following relationship:**

$$odds= \frac{prob}{1-prob}$$

**The estimated odds of death within 5 years for a population of individuals with low serum LDL is 0.2046. The observed proportion of subjects with low LDL dying within 5 years is 0.1699 (16.99%). The observed proportion is equivalent to the estimated probability of death within 5 years for a population of individuals with low serum LDL. However, the estimated odds of death within 5 years for subjects with low LDL is not equivalent to either estimated or the observed proportion of subjects with low LDL who died within 5 years. This is because the odds is only a good estimate of the probability when the probability is very small. This is exactly the answer that was obtained for part b, when an indicator for high LDL was used as the predictor.**

**The estimated probability of dying within 5 years for a population of individuals with high serum LDL (LDL ≥ 160mg/dL) is given by 0.1308 (13.08%), which is the exponentiated intercept of the Poisson regression model. The estimated odds of dying within 5 years for a population of individuals with high serum LDL is given by the following relationship:**

$$odds= \frac{prob}{1-prob}$$

**The estimated odds of death within 5 years for a population of individuals with high serum LDL is 0.1505. The observed proportion of subjects with low LDL (LDL < 160mg/dL) dying within 5 years is 0.1699 (16.99%), which is higher than the estimated probability of subjects with high serum LDL dying within 5 years by a relative difference of roughly 26% or an absolute difference of 3.91%. The estimated odds of death within 5 years for a population of individuals with high serum LDL is different from both the estimated probability of death within 5 years for a population of individuals with high serum LDL and the observed proportion of subjects with low serum LDL dying within 5 years. This is exactly the answer that was obtained for part c, when an indicator for high LDL was used as the predictor.**

 **Indicator of Survival for at least 5 years as the response variable:**

**This model uses the same predictor as the model used in parts a-c, only the response variable has been reparameterized. Therefore there are still the same two distinct groups, those who have high serum LDL levels (≥160mg/dL) and those who have normal/low serum LDL levels (< 160mg/dL) that are modeled with two regression parameters, the intercept and the slope. Because the number of modeled groups is equivalent to the number of regression parameters this is also a saturated model.**

**The estimated probability of dying within 5 years for a population of individuals with low serum LDL (LDL < 160mg/dL) is given by 0.1699 (16.99%), which is given by the following relationship:**

$$prob\left(low LDL\right)=1 -prob(surval|lowLDL)$$

**The probability of surviving at least 5 years for a population of individuals with low serum LDL was estimated as 0.8301 (83.01%) which is the exponentiated intercept of the Poisson regression model. The estimated odds of dying within 5 years for a population of individuals with low serum LDL is given by the following relationship:**

$$odds= \frac{prob}{1-prob}$$

**The estimated odds of death within 5 years for a population of individuals with low serum LDL is 0.2047. The observed proportion of subjects with low LDL dying within 5 years is 0.1699 (16.99%). The observed proportion is equivalent to the estimated probability of death within 5 years for a population of individuals with low serum LDL. However, the estimated odds of death within 5 years for subjects with low LDL is not equivalent to either the estimated or the observed proportion of subjects with low LDL who died within 5 years. This is because the odds is only a good estimate of the probability when the probability is very small. This is exactly the answer that was obtained for part b, when an indicator for high LDL was used as the predictor.**

**The estimated probability of dying within 5 years for a population of individuals with high serum LDL (LDL ≥ 160mg/dL) is given by 0.1308 (13.08%), which is given by the following relationship:**

$$prob\left(high LDL\right)=1 -prob(surval|high LDL)$$

**The probability of surviving at least 5 years for a population of individuals with high serum LDL was estimated as 0.8692 (86.92%) which is the exponentiated sum of the slope and the intercept of the Poisson regression model. The estimated odds of dying within 5 years for a population of individuals with high serum LDL is given by the following relationship:**

$$odds= \frac{prob}{1-prob}$$

**The estimated odds of death within 5 years for a population of individuals with high serum LDL is 0.1505. The observed proportion of subjects with low LDL (LDL < 160mg/dL) dying within 5 years is 0.1699 (16.99%), which is higher than the estimated probability of subjects with high serum LDL dying within 5 years by a relative difference of roughly 26% or an absolute difference of 3.91%. The estimated odds of death within 5 years for a population of individuals with high serum LDL is different from both the estimated probability of death within 5 years for a population of individuals with high serum LDL and the observed proportion of subjects with low serum LDL dying within 5 years. This is exactly the answer that was obtained for part c, when an indicator for high LDL was used as the predictor.**

* 1. In parts a-d of this problem, we described the distribution of death within 5 years across groups defined by LDL level. What if we fit a regression model mimicking the approach used in problems 1 – 4 of homework #2, where we described the distribution of LDL across groups defined by vital status? How would our answers to parts a-c change?

**Because the same sample is used for this regression model and the model used in parts a-c the estimated probabilities of death within 5 years for a population with low serum LDL will be the same as will the estimated probabilities of death within 5 years for a population with high serum LDL. The observed proportion of subjects with low serum LDL (LDL < 160mg/dL) will also be the same for both models. This model uses an indicator of death within 5 years as the predictor and an indicator of high LDL as the response variable. This means that the interpretation of the outputs is as the probability of LDL level given vital status instead of the probability of vital status given LDL level. It is difficult to relate those quantities, so while I expect the answers to parts a-c to stay the same directly observing this is complicated using these model parameters.**

1. Perform a regression analysis of the distribution of death within 5 years across groups defined by the continuous measure of LDL. (In all cases we want formal inference.)
	1. Evaluate associations between 5 year mortality and LDL using risk difference (RD: difference in probabilities).

**Methods: Serum LDL levels were selected as the predictor of interest and vital status as the response variable. A linear regression model allowing for heteroscedasticity was used to determine if there was a linear trend in probability of 5 year mortality across groups defined by serum LDL levels. An alpha level of 0.05 was used and the 95% confidence interval was calculated, the Huber-White sandwich estimator for robust standard errors was used. The null hypothesis was that there is no linear trend in probability of 5 year mortality across groups defined by serum LDL. The alternative hypothesis is that there is a linear trend in probability of 5 year mortality across groups defined by serum LDL.**

**Inference: From a linear regression analysis of 725 observations we estimate a mean absolute difference of 0.103% in risk of death within 5 years per mg/dL increase in serum LDL, with higher LDL groups tending to have a decreased risk of death within 5 years. A 95% confidence interval suggests that this observation is not unusual if the true absolute risk difference for 5 year mortality was anywhere between 0.188% and 0.018% lower per mg/dL increase in serum LDL. Using a two-sided F-test p-value this observation is statistically significant at a 0.05 level of significance (p = 0.0171) and we reject the null hypothesis that there is no linear trend in 5 year mortality across groups defined by serum LDL levels.**

* 1. Evaluate associations between 5 year mortality and LDL using risk ratio (RR: ratios of probabilities).

**Methods: Serum LDL levels were selected as the predictor of interest and vital status as the response variable. A Poisson regression model that did not assume a mean-variance relationship was used to determine if there was a linear trend in risk of 5 year mortality across groups defined by serum LDL levels. An alpha level of 0.05 was used and the 95% confidence interval was calculated, robust standard errors were used. The null hypothesis was that there is no linear trend in 5 year mortality rates across groups defined by serum LDL. The alternative hypothesis is that there is a linear trend in 5 year mortality rates across groups defined by serum LDL.**

**Inference: From a Poisson regression analysis of 725 observations we estimate a mean ratio of 0.9936 for risk of death within 5 years per mg/dL increase in serum LDL, with higher LDL groups tending to have a 0.645% decreased risk of death within 5 years per mg/dL increase in LDL. A 95% confidence interval suggests that this observation is not unusual if the true risk ratio for 5 year mortality was anywhere between 0.9883 and 0.9989 per mg/dL increase in serum LDL, with higher LDL groups having between a 0.112% and 1.175% decrease in risk of death per mg/dL increase in LDL. Using a two-sided Wald test p-value this observation is statistically significant at a 0.05 level of significance (p = 0.0177) and we reject the null hypothesis that there is no linear trend in 5 year mortality rates across groups defined by serum LDL levels.**

* 1. Evaluate associations between 5 year mortality and LDL using odds ratio (OR: ratios of odds)

**Methods: Serum LDL levels were selected as the predictor of interest and vital status as the response variable. A logistic regression model was used to determine if there was a linear trend in the log odds of 5 year mortality across groups defined by serum LDL levels. An alpha level of 0.05 was used and the 95% confidence interval was calculated using the Wald test, robust standard errors were used. The null hypothesis was that there is no linear trend in the log odds of 5 year mortality across groups defined by serum LDL. The alternative hypothesis is that there is a linear trend in the log odds of 5 year mortality rates across groups defined by serum LDL.**

**Inference: From a logistic regression analysis of 725 observations we estimate a mean odds ratio of 0.9923 for odds of death within 5 years per mg/dL increase in serum LDL, with higher LDL groups tending to have a 0.774% decreased odds of death within 5 years per mg/dL increase in LDL. A 95% confidence interval suggests that this observation is not unusual if the true odds ratio for 5 year mortality was anywhere between 0.9858 and 0.9987 per mg/dL increase in serum LDL, with higher LDL groups having between a 0.125% and 1.419% decrease in odds of death per mg/dL increase in LDL. Using a two-sided Wald test p-value this observation is statistically significant at a 0.05 level of significance (p = 0.0194) and we reject the null hypothesis that there is no linear trend in log odds of 5 year mortality across groups defined by serum LDL levels.**

* 1. How do your conclusions about such an association from this model compare to your conclusions reached in problems 1-3 of this homework and problems 2 and 4 of homework #2? Which analyses would you prefer *a priori*?

**The conclusion reached using these models is that there is a statistically significant difference (α=0.05) in odds or probability of 5 year mortality per mg/dL increase in serum LDL with higher LDL groups tending to have lower odds or probability of 5 year mortality. The conclusion reached in problems 1-3 of this assignment is that there is no statistically significant difference (α=0.05) in odds or probability of 5 year mortality between high serum LDL groups (LDL ≥ 160mg/dL) and low serum LDL groups (LDL < 160mg/dL). The conclusions reached in problems 2 and 4 of homework 2 was that there was a statistically significant difference (α=0.05) in serum LDL between groups defined by vital status at 5 years. The lack of statistical significance seen in problems 1-3 of this assignment was likely due to the loss of precision due to dichotomizing serum LDL rather than treating it as a continuous measurement. The results obtained for questions 2 and 4 in homework two were reparameterizations of the models used in this problem, where the response variable and predictor of interest were switched.**

***A priori* I would have preferred to use the method in part a of this problem:**

**Serum LDL levels were selected as the predictor of interest and vital status as the response variable. A linear regression model allowing for heteroscedasticity was used to determine if there was a linear trend in probability of 5 year mortality across groups defined by serum LDL levels. An alpha level of 0.05 was used and the 95% confidence interval was calculated, the Huber-White sandwich estimator for robust standard errors was used. The null hypothesis was that there is no linear trend in probability of 5 year mortality across groups defined by serum LDL. The alternative hypothesis is that there is a linear trend in probability of 5 year mortality across groups defined by serum LDL.**

**This is because this model keeps serum LDL as a continuous measure so there is no loss of precision due to categorization and the results give the risk difference which is easily interpretable, more so than the odds or risk ratios.**