**total score 83/94**

1. 1. Yes, this model is saturated. This is because we know the exact proportions for both the predictor of interest and the outcome (we know the exact proportion of subjects with high LDL levels, and the exact proportion of individuals that died within five years). Thus we do not use any ‘borrowed’ information in the model, and the two-parameter model is fitted to data with two groups.
	2.
* The observed *proportion* of subjects with low LDL that died within 5 years is 0.1699.
* For subjects with low LDL, the estimated *odds* of dying within 5 years is 0.205.
* For subjects with low LDL the *probability* of dying within 5 years is 16.99%.
* Comparison: The probability of dying is the same as the proportion of individuals dying (this is because we have a binary variable, and thus the amount of times that we see one outcome is the probability of that outcome occurring). The odds of dying within 5 years is the same as (p/(1-p)), where p is the proportion of subjects with low LDL that died within 5 years.

	1.
* The observed proportion of subjects with high LDL that died within 5 years is 0.1368.
* For subjects with high LDL the estimated odds of dying within 5 years is 0.158.
* For subjects with high LDL the probability of dying within 5 years is 13.68%.
* Comparison: The proportion of subjects who died, the probability of dying, and the odds of dying are all lower for subjects with high LDL concentrations. The high LDL group had an odds of dying within five years that was 0.047 lower than the low LDL group. The probability of death was 3.31% lower in the high LDL group.

	1. Based on observations for 107 subjects with high LDL (“high” = LDL > 160 mg/dL), the probability that high LDL individuals died within 5 years was 13.7%. The estimated odds ratio for high LDL individuals’ risk of dying was 0.735, with a standard error of 0.225, thus demonstrating a slight protective effect of having high LDL. Based on our 95% confidence interval, it would not be unusual if the true population odds ratio were found between .404 and 1.37 (p= 0.315). Our analysis demonstrates that observing this odds ratio of dying within five years for individuals with high LDL would not be unusual under the null hypothesis. Therefore we fail to reject the null hypothesis that there is no association in the odds of dying within five years for individuals with high LDL versus low LDL.

The inference obtained from logistic regression as discussed above is highly similar to that from homework 1 problem 6. The obtained odds ratios are exactly the same, although the p-value obtained above (0.315) is slightly lower than that obtained in problem 6 of homework 1 (0.396). When Wald-based confidence intervals are obtained from the Fisher’s Exact test from problem 1, the 95% CI’s are exactly the same between logistic regression and the Fisher’s Exact test. The difference in the p-values is likely due to the increased precision of fitting a regression line using a saturated model as opposed to the Fisher’s exact test.

* 1. I would expect to see no changes in the answers to parts a-c change if I had used the different variables for predictors of interest and different response variables. This is because we are merely using the ‘other part’ of the proportion. What we label as p or 1-p is completely arbitrary, and thus the values obtained from the model will not differ if we parameterize with p or with 1-p.
	2. *Model Saturation:* The model would still be saturated because the distribution of death within 5 years is a binary variable for which we would be fitting a 2 parameter model.

*Odds, proportions and probabilities:* With logistic regression, we could condition on what was previously our response variable and obtain an odds ratio that would be accurate and the same as the odds ratio that would be obtained using logistic regression where we condition on the variable we sampled by. This is because the odds ratio can use the properties of multiplication and division to change the input values in the denominator and numerator of the ratio without causing changes in the overall odds ratio. The proportions and probabilities would also not change if we condition upon death status instead of LDL status.

1. 1. Yes, this model is saturated. This is because we know the exact proportions for both the predictor of interest and the outcome (we know the exact proportion of subjects with high LDL levels, and the exact proportion of individuals that died within five years). Thus we do not use any ‘borrowed’ information in the model, and the two-parameter model is fitted to data with two groups.
	2.
* The observed *proportion* of subjects with low LDL that died within 5 years is 0.1699.
* For subjects with low LDL, the estimated *odds* of dying within 5 years is 0.205.
* For subjects with low LDL the *probability* of dying within 5 years is 16.99%.
* Comparison: The probability of dying is the same as the proportion of individuals dying (this is because we have a binary variable, and thus the amount of times that we see one outcome is the probability of that outcome occurring). The odds of dying within 5 years is the same as (p/(1-p)), where p is the proportion of subjects with low LDL that died within 5 years.

	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 observed proportion of subjects with high LDL that died within 5 years is 0.1368.
* For subjects with high LDL the estimated odds of dying within 5 years is 0.158.
* For subjects with high LDL the probability of dying within 5 years is 13.68%.
* Comparison: The proportion of subjects who died, the probability of dying, and the odds of dying are all lower for subjects with high LDL concentrations. The high LDL group had an odds of dying within five years that was 0.047 lower than the low LDL group. The probability of death was 3.31% lower in the high LDL group.

	1. Robust linear regression was performed using high LDL concentration as the predictor of interest and death within five years as the outcome of interest. Based on 107 subjects with high LDL concentrations, the probability of dying was 3.9% lower for individuals with high LDL than for individuals with low LDL concentrations. Based on our 95% CI, it would not be unusual if the true population difference in the probability of dying within five years was anywhere between 11.0% *lower* in the high LDL group to 3.16% *higher* in the high LDL group than in the low LDL groups (p = 0.278). Given the observed p-value, there is insufficient evidence to demonstrate that the observed difference in probabilities would be unusual under the null hypothesis, and thus we fail to reject the null hypothesis that there is no difference in the probability of dying between high LDL individuals and low LDL individuals.

The inference between this problem and numbers 5 and 6 from homework 1 are highly similar. The risk difference of – 0.039 is very similar to the risk difference reported in problem 6 of homework #1, and the 95% confidence intervals are also essentially the same between the two inferences. The p-value is lower when robust linear regression is used (p-value is 0.278 in the above inference, whereas it is 0.314 when the chi-square test is used as in homework 1). This difference in the p value is likely due to the increased precision of linear regression over the chi-square test.

* 1. If I had fit a regression model using an indicator of low LDL as the predictor, the difference in the probabilities would be a positive value instead of a negative value since the probability of death is higher in the low LDL group. The actual value of the difference would not change, but the sign would change (ie from a negative value to a positive value). The intercept of the regression line would likely change, but the slope of the line would not.

If we had used an indicator of survival for at least 5 years as the response variable then the difference in the probabilities (indicated as the slope of the regression line) would continue to be the same.

* 1. *Model Saturation:* The model would still be saturated because the distribution of death within 5 years is a binary variable for which we would be fitting a 2 parameter model.

*Odds, proportions, and probabilities:* When using a linear regression model we cannot condition upon LDL concentrations (which is the variable we sampled on). If we were to do this, the risk difference would be incorrect since our sampling on LDL essentially equates to our pre-determining the distribution of LDL levels. Thus the odds should not be calculated in this instance. However, the proportions and probabilities would not change if you condition on survival status, since these values represent counts of the number of individuals displaying combinations of our binary ‘traits’ (high LDL, and death within five years, of high LDL, alive after 5 years etc.) which does not change depending on which variable you condition on.

1. 1. Yes, this model is saturated. This is because we know the exact proportions for both the predictor of interest and the outcome (we know the exact proportion of subjects with high LDL levels, and the exact proportion of individuals that died within five years). Thus we do not use any ‘borrowed’ information in the model, and the two-parameter model is fitted to data with two groups.
	2.
* The observed *proportion* of subjects with low LDL that died within 5 years is 0.1699.
* For subjects with low LDL, the estimated *odds* of dying within 5 years is 0.205.
* For subjects with low LDL the *probability* of dying within 5 years is 16.99%.
* Comparison: The probability of dying is the same as the proportion of individuals dying (this is because we have a binary variable, and thus the amount of times that we see one outcome is the probability of that outcome occurring). The odds of dying within 5 years is the same as (p/(1-p)), where p is the proportion of subjects with low LDL that died within 5 years.

	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 observed proportion of subjects with high LDL that died within 5 years is 0.1368.
* For subjects with high LDL the estimated odds of dying within 5 years is 0.158.
* For subjects with high LDL the probability of dying within 5 years is 13.68%.
* Comparison: The proportion of subjects who died, the probability of dying, and the odds of dying are all lower for subjects with high LDL concentrations. The high LDL group had an odds of dying within five years that was 0.047 lower than the low LDL group. The probability of death was 3.31% lower in the high LDL group.

	1. Robust poisson regression was performed using high LDL concentration as the predictor of interest and death within five years as the outcome of interest. Based on 107 subjects with high LDL concentrations (high = LDL > 160 mg/dL), the risk ratio for high LDL individuals dying within five years was 0.77, indicating a protective effect from high LDL. Based on our 95% CI (calculated using Wald methods), it would not be unusual if the true population risk ratio was any value between 0.46 and 1.29 (p = 0.324). Given the observed p-value, there is insufficient evidence to demonstrate that the observed relative risk would be unusual under the null hypothesis, and thus we fail to reject the null hypothesis that there is no difference in the risk of dying within five years between high LDL individuals and low LDL individuals.

The inference between this problem and numbers 5 and 6 from homework 1 are highly similar. The relative risk of 0.77 is very similar to odds ratio reported in problem 6 of homework #1. The main difference is that the 95% confidence intervals are narrower when poisson regression is performed as opposed to when using the Fisher’s Exact test. These differences are likely due to the method in which 95% CI’s are formed (in problem 6 of homework 1, the 95% CI’s are calculated using exact methods, whereas in this problem the 95% CI’s were calculated using Wald methods).

* 1. I would expect to see no changes in the answers to parts a-c change if I had used the different variables for predictors of interest and different response variables. This is because we are merely using the ‘other part’ of the proportion. What we label as p or 1-p is completely arbitrary, and thus the values obtained from the model will not differ if we parameterize with p or with (1-p).
	2. *Model Saturation:* The model would still be saturated because the distribution of death within 5 years is a binary variable for which we would be fitting a 2 parameter model.

*Odds, proportions, and probabilities:* When using a poisson regression model we cannot condition upon LDL concentrations (which is the variable we sampled on). If we were to do this, the risk difference would be incorrect since our sampling on LDL essentially equates to our pre-determining the distribution of LDL levels. Thus the odds should not be calculated in this instance. However, the proportions and probabilities would not change if you condition on survival status, since these values represent counts of the number of individuals displaying combinations of our binary ‘traits’ (high LDL, and death within five years, of high LDL, alive after 5 years etc.) which does not change depending on which variable you condition on.

1. 1. METHODS: Robust linear regression was used to evaluate the difference in risk of death conferred by different levels of serum LDL. Serum LDL measurements (our predictor of interest) were kept continuous, however observation times (the response variable) were dichotomized to represent individuals that died within five years and those that were still surviving at five years time. 95% confidence intervals were inferred using Wald-based methods.

INFERENCE: Based on LDL observations for 725 subjects, the difference in risk of death per unit increase in serum LDL concentration was -0.00103 with a standard error of 0.000433. Based on our 95% confidence interval, it would not be unusual if the true population difference in risk was found between -0.00188 and -0.000185 (p= 0.017). Our analysis demonstrates that observing this large of a risk difference per unit increase in serum LDL concentration would be unusual under the null hypothesis, therefore we reject the null hypothesis that there is no association between risk of death and LDL concentration.

* 1. METHODS: Poisson regression was used to evaluate the relative risk of death conferred by different levels of serum LDL. Serum LDL measurements (our predictor of interest) were kept continuous, however observation times (the response variable) were dichotomized to represent individuals that died within five years and those that were still surviving at five years time. 95% confidence intervals were inferred using Wald-based methods.

INFERENCE: Based on LDL observations for 725 subjects, the relative risk of death per unit increase in serum LDL concentration was 0.994 with a standard error of 0.00271. Based on our 95% confidence interval, it would not be unusual if the true population relative risk were found between 0.988 and 0.999 (p= 0.018). Our analysis demonstrates that observing this size of a relative risk per unit increase in serum LDL concentration would be unusual under the null hypothesis, therefore we reject the null hypothesis that there is no association between the relative risk of death and LDL concentration.

* 1. METHODS: Logistic regression to obtain odds ratios was used to evaluate associations between 5 year mortality and LDL. Serum LDL measurements (our predictor of interest) were kept continuous, however observation times (the response variable) were dichotomized to represent individuals that died within five years and those that were still surviving at five years time. 95% confidence intervals were inferred using Wald-based methods.

INFERENCE: Based on LDL observations for 725 subjects, the ratio of the odds of death per unit increase in serum LDL concentration was 0.992 with a standard error of .00307. Based on our 95% confidence interval, it would not be unusual if the true population odds ratio were found between 0.986 and 0.998 (p= 0.012). Our analysis demonstrates that observing this reduction in the odds of death per unit increase in serum LDL concentration would be unusual under the null hypothesis, therefore we reject the null hypothesis that there is no association between the ratio of the odds of death and LDL concentration.

* 1. The conclusions about an association between LDL concentration and 5 year all-cause mortality are much stronger, and are actually statistically significant, when LDL is treated as a continuous variable (as in this question) than when LDL is dichotomized (as in problems 1-3 in this assignment). This is likely due to the increased level of precision that is gained when using LDL levels as a continuous variable rather than a binary variable. Although precision was increased in problem 4, the overall trend of high LDL acting protectively against the risk of death was consistent across all analyses.

The conclusions from this analysis demonstrating decreasing risk of death within five years as LDL concentration increases are consistent with the results seen in problems 2 and 4 of homework #2.

*A priori* I think that the best analysis would be to evaluate the association between 5 year mortality and LDL using the odds ratio (analysis using logistic regression) because logistic regression does not assume a linear relationship between the POI and the outcome, nor does it assume homoscedasticity. From the outset, its probably likely that the relationship between LDL and five year mortality will not be linear, and since regression using the risk difference (linear regression) and regression using the ratios of probabilities (poisson regression) require a linear relationship, these may not be the best choices. In addition, using the logistic regression could allow us to condition upon either LDL concentration or mortality, even though we had sampled on serum LDL measurements, which allows increased flexibility in the ways we can analyze our data. Poisson regression would not be the best method because our outcome of interest (death) is not particularly rare, and therefore the relative risk is likely of less interest.