**TOTAL 106/108**

**1.**

We use logistic regression to compare the odds of death within 5 years between high LDL (≥ 160 mg/dL) and low LDL groups. I exclude the 10 observations with missing values of LDL leaving 725 observations for this logistic regression analysis of the odds ratio. For parts a-c, we fit the following logistic regression model

,

where is an indicator of whether the *i*th observation died within 5 years and is an indicator of whether the *i*th individual has high LDL. We do not use robust standard errors. Since my predictor of interest is LDL level, my coefficient of interest is and I will make inference using my estimate, (, which is the estimated odds ratio for mortality in the high LDL group relative to mortality in the low LDL group. We also compute a Wald-based confidence interval for this estimated odds ratio.

**(a)**

Yes, this is a saturated model. There are 2 groups (high LDL and low LDL) and two parameters ().

**(b)**

For individuals with low LDL, logistic regression estimates the odds of dying within 5 years as This corresponds to an estimated probability of 5-year mortality of 0.205 / (1+0.205) = **0.17.**

We observed that 105 of 618 subjects with low LDL died within 5 years. This corresponds to a proportion dying within 5 years of 0.17 and odds of dying within 5 years of 0.205. Our saturated model fit the sample means of these groups exactly.

**(c)**

For individuals with high LDL, logistic regression estimates the probability of dying within 5 years is This gives to an estimated probability of 5-year mortality of 0.151 / (1+0.151) = **0.131.**

We observed that 14 of 107 subjects with high LDL died within 5 years. This corresponds to a proportion dying within 5 years of 0.131 and odds of dying within 5 years of 0.151. Our saturated model fit the sample means of these groups exactly.

**(d)**

**Method:** We use logistic regression to compare the odds of death within 5 years between high LDL (≥ 160 mg/dL) and low LDL groups. I exclude the 10 observations with missing values of LDL leaving 725 observations for this logistic regression analysis of the odds ratio. For parts a-c, we fit the following logistic regression model

,

where is an indicator of whether the *i*th observation died within 5 years and is an indicator of whether the *i*th individual has high LDL. We do not use robust standard errors. Since my predictor of interest is LDL level, my coefficient of interest is and I will make inference using my estimate, (, which is the estimated odds ratio for mortality in the high LDL group relative to mortality in the low LDL group. We also compute a Wald-based confidence interval for this estimated odds ratio. We are interested in testing the hypothesis:

**Inference:** From our logistic regression, we estimate that the odds of 5-year mortality in the high LDL group are 73.5% of the odds of mortality in the low LDL group. This comes from our estimated The corresponding Wald-based 95% CI for this estimate is [.404, 1.340], which suggests that our estimate would not be unusual if the odds or 5-year mortality for elderly individuals with high LDL was between 40.4% and 134% of the mortality odds for elderly individuals with low LDL. We note that 100% is contained within this interval. The two-sided p-value of our estimated odds ratio is 0.315. As a result, we do not have evidence to reject the null hypothesis that the odds of 5-year all-cause mortality are the same in both groups at the .05 level of significance (two-sided p-value = 0.315 > .05, n=725).

Our point estimate of this odds ratio (0.735) is the same as in Homework 1. However, the 95% and p-value are different for this estimate than in Homework 1. This is because we report a Wald-based CI and p-value while HW 1 used Fisher’s exact test.

**(e)**

1. If we had instead fit a regression model with an indicator of death within 5 years as our response and an indicator of low LDL as our predictor variable, our model would still be saturated (it would still have 2 groups and 2 parameters) so our answer to part (a) would not change. Our answers to parts (b) and (c) would change.
2. If we had instead fit a regression model with an indicator of survival within 5 years as our response and an indicator of high LDL as our predictor variable, our model would still be saturated (it would still have 2 groups and 2 parameters) so our answer to part (a) would not change. Our answers to parts (b) and (c) would change, as our new estimated odds ratios would be the inverses of the previous values.

**(f)** If we fit a logistic regression model with high LDL level as the response variable and vital status as the predictor variable, the model would remain saturated (two groups and two parameters), so our answer to part (a) would not change. Our answer to parts (b) and (c) would also be the same, since the odds ratio is invariant under a reversal of outcome and exposure.

**2.** I exclude the 10 observations with missing values of LDL leaving 725 observations for this linear regression analysis of risk difference. We classify as high LDL as being LDL ≥160 mg/dL. For parts a-c, we fit the following linear regression model

,

where is an indicator of whether the *i*th observation died within 5 years and is an indicator of whether the *i*th individual has high LDL. We do not use robust standard errors. Since my predictor of interest is LDL level, my coefficient of interest is and I will make inference using my estimate, , which is the estimated absolute difference in probability of 5-year all-cause mortality between individuals with high LDL relative to that of individuals with low LDL..

**(a)**

Yes, this is a saturated model. There are 2 groups (high LDL and low LDL) and two parameters ().

**(b)**

For individuals with low LDL, linear regression estimates the probability of dying within 5 years as **,** where (set to 0) is an indicator of having high LDL. For these subjects, the estimated odds of dying within 5 years are

We observed that 105 of 618 subjects with low LDL died within 5 years. This corresponds to a proportion dying within 5 years of 0.17 and odds of dying within 5 years of 0.205. Our saturated model fit the sample means of these groups exactly.

**(c)**

For individuals with high LDL, linear regression estimates the probability of dying within 5 years is **,** where (set to 1) is an indicator of having high LDL. For these subjects, the estimated odds of dying within 5 years are

We observed that 14 of 107 subjects with high LDL died within 5 years. This corresponds to a proportion dying within 5 years of 0.131 and odds of dying within 5 years of 0.151. Our saturated model fit the sample means of these groups exactly.

**(d)**

**Method:** I exclude the 10 observations with missing values of LDL leaving 725 observations for this linear regression analysis of differences in 5-year all-cause mortality across high LDL and low LDL groups. We classify as high LDL as being LDL ≥160 mg/dL. We fit the following linear regression model

,

where is an indicator of whether the *i*th observation died within 5 years and is an indicator of whether the *i*th individual has high LDL. We do not use robust standard errors. Since my predictor of interest is LDL level, my coefficient of interest is and I will make inference using my estimate, , which is the estimated absolute difference in probability of 5-year all-cause mortality between individuals with high LDL relative to that of individuals with low LDL. We are interested in testing the hypothesis:

**Inference:** The regression model estimates a difference in 5 year all-cause mortality rate for observations with high LDL relative to observations with high LDL () of -0.039. However, the standard error of this estimate is 0.389 and the corresponding 95% confidence interval is [-0.115, 0.037]. That is, our observed sample would not be surprising if the true absolute difference in 5 year all-cause mortality rate between elderly individuals with high LDL relative to that of elderly individuals with low LDL were between 0.115 lower and 0.037 higher. We note that 0 is contained in this interval. The model also reports a two-sided p-value of 0.315; that is, given that there were no true difference in population 5-year all-cause mortality between elderly individuals with high LDL and low LDL, there is roughly a 31.5% chance of observing a sample as or more extreme than what we actually observe. As a result, we do not have evidence to reject the null hypothesis that 5-year all-cause mortality is not associated with high/low LDL level at the .05 level of significance (two-sided p-value = 0.315 > .05, n=725).

Our point estimates of the absolute difference in 5-year all-cause mortality between high LDL and low LDL groups are the same as in problems 5 and 6 of Homework 1. However, our 95% confidence interval for this inference is [-0.115, 0.037], while it was [-0.109, 0.314] in Homework 1. Our p-value is also slightly different. This is because Homework 1 used a Chi-squared test statistic and here we use a t-statistic to determine the critical value used in our CI p-value.

**(e)**

1. If we had instead fit a regression model with an indicator of death within 5 years as our response and an indicator of low LDL as our predictor variable, we would have estimated as 0.131 and as 0.039. The magnitude of the slope would remain the same, but its sign would flip. Our model is still saturated, so our answer to (a) would be the same.

In addition, we would have an estimated probability of 5-year all-cause mortality for low LDL individuals as **,** where (set to 0) is an indicator of having low LDL. For these subjects, the estimated odds of dying within 5 years are Our saturated model fit the sample means of these groups exactly. These are the same answers we reported in part (b).

We would also have an estimated probability of 5-year all-cause mortality for low LDL individuals as **,** where (set to 1) is an indicator of having low LDL. For these subjects, the estimated odds of dying within 5 years are Our saturated model fit the sample means of these groups exactly. These are the same answers we reported in part (b).

1. If we had instead fit a regression model with an indicator of survival of for at least 5 years as our response and an indicator of high LDL as our predictor variable, we would have estimated as 1 – 0.17 = 0.83 and as 0.039. That is, the magnitude of the slope would remain the same but its sign would flip. Our model is still saturated, so our answer to (a) would be the same.

In addition, we would have an estimated probability of surviving at least 5-years for low LDL individuals as **,** where (set to 0) is an indicator of having low LDL. This is the complement of our proportion estimate in part (b). Our saturated model fit the sample means of these groups exactly. The odds of surviving at least 5 years would be .

We would also have an estimated probability of 5-year survival for high LDL individuals as where (set to 1) is an indicator of having high LDL. This is the complement of our proportion estimate from part (c). Our saturated model fit the sample means of these groups exactly. The odds of surviving at least 5 years would be .

**(f)**

If we fit a linear regression model with high/LDL as the response variable and vital status as the predictor variable, the model would remain saturated (two groups and two parameters), so our answer to part (a) would not change. However, our answers to parts (b) and (c) would change because risk differences are not invariant under a reversal of response and predictor.

**3.**

I exclude the 10 observations with missing values of LDL leaving 725 observations for comparing risk ratios of 5-year all-cause mortality across high LDL and low LDL groups. We classify as high LDL as being LDL ≥160 mg/dL. We fit the following Poisson regression model

,

where is an indicator of whether the *i*th observation died within 5 years and is an indicator of whether the *i*th individual has high LDL. We do not use robust standard errors. Since my predictor of interest is LDL level, my coefficient of interest is and I will make inference using my estimate, , which is the estimated 5-year all-cause mortality risk ratio of individuals with high LDL relative to the risk of individuals with low LDL. We will use Wald-based confidence intervals for these risk ratios. We are interested in testing the hypothesis:

which is equivalent to testing

**(a)**

Yes, this is a saturated model. There are 2 groups (high LDL and low LDL) and two parameters () in our model.

**(b)**

For individuals with low LDL, Poisson regression estimates the probability of dying within 5 years as **,** where (set to 0) is an indicator of having high LDLFor these subjects, the estimated odds of dying within 5 years are We observed that 105 of 618 subjects with low LDL died within 5 years. This corresponds to a proportion dying within 5 years of 0.17 and odds of dying within 5 years of 0.205. Our saturated Poisson regression fit the sample means of these groups exactly.

**(c)**

For individuals with high LDL, Poisson regression estimates the probability of dying within 5 years as **,** where (set to 1) is an indicator of having high LDL. For these subjects, the estimated odds of dying within 5 years are

We observed that 14 of 107 subjects with high LDL died within 5 years. This corresponds to a proportion dying within 5 years of 0.131 and odds of dying within 5 years of 0.151. Our saturated model fit the sample means of these groups exactly.

**(d)**

**Method:** I exclude the 10 observations with missing values of LDL leaving 725 observations for comparing risk ratios of 5-year all-cause mortality across high LDL and low LDL groups. We classify as high LDL as being LDL ≥160 mg/dL. We fit the following Poisson regression model

,

where is an indicator of whether the *i*th observation died within 5 years and is an indicator of whether the *i*th individual has high LDL. We do not use robust standard errors. Since my predictor of interest is LDL level, my coefficient of interest is and I will make inference using my estimate, , which is the estimated 5-year all-cause mortality risk ratio of individuals with high LDL relative to the risk of individuals with low LDL. We will use Wald-based confidence intervals for these risk ratios. We are interested in testing the hypothesis:

which is equivalent to testing

**Inference:** The Poisson regression model estimates a ratio of the probability of 5-year all-cause mortality in the high LDL group relative to the low LDL group with or . The corresponding Wald-based 95% CI of is [0.441, 1.345]. That is, our observed sample would not be surprising if the true ratio of 5-year all-cause mortality risk between elderly individuals with high LDL relative to that of elderly individuals with low LDL were between 0.441 and 1.345. We note that 1 is contained in this interval. The model also reports a Wald-based two-sided p-value of 0.359; that is, given that there were no true difference in the 5-year all-cause mortality risk between elderly individuals with high LDL and low LDL, there is roughly a 35.9% chance of observing a sample as or more extreme than what we actually observe. As a result, we do not have evidence to reject the null hypothesis of that the probability of 5-year all-cause mortality is not associated with high/low LDL level at the .05 level of significance (two-sided p-value = 0.359 > .05, n=725).

**(e)**

1. If we had instead fit a regression model with an indicator of death within 5 years as our response and an indicator of low LDL as our predictor variable, we would have estimated as -2.033 and as 0.261. Under this reparameterization, the magnitude of the slope would remain the same, but its sign would flip. Our model is still saturated, so our answer to (a) would be the same.

In addition, we would have an estimated probability of 5-year all-cause mortality for low LDL individuals as **,** where (set to 1) is an indicator of having low LDL. For these subjects, the estimated odds of dying within 5 years is We observed that 105 of 618 subjects with low LDL died within 5 years. This corresponds to a proportion dying within 5 years of 0.17 and odds of dying within 5 years of 0.205. Our saturated Poisson regression fit the sample means of these groups exactly. These are the same answers we reported in part (b).

In addition, we would have an estimated probability of 5-year all-cause mortality for high LDL individuals as **,** where (set to 0) is an indicator of having low LDL. For these subjects, the estimated odds of dying within 5 years are We observed that 14 of 107 subjects with high LDL died within 5 years. This corresponds to a proportion dying within 5 years of 0.131 and odds of dying within 5 years of 0.151. Our saturated model fit the sample means of these groups exactly. These are the same answers we reported in part (c).

1. If we had instead fit a regression model with survival of at least 5 years as the response variable and high/low LDL as the predictor of interest, the model would remain saturated (two groups and two parameters), so our answer to part (a) would not change. However, our answers to parts (b) and (c) would be change; we would now estimate survival instead of death at 5 years, so our estimated probability in parts (b) and (c) would be equal to the complements of the events we report above.

 **(f)**

If we fit a Poisson regression model with high/low LDL as the response variable and vital status as the predictor variable, the model would remain saturated (two groups and two parameters), so our answer to part (a) would not change. However, our answers to parts (b) and (c) would change because risk ratios are not invariant under a reversal of response and predictor.

**4.** In all cases I exclude the 10 observations with missing values of LDL, leaving 725 observations.

**(a)**

**Method:** I estimate a linear regression (not assuming homoscedasticity) under the following model:

,

where is an indicator of whether the *i*th observation died within 5 years and is the level of serum LDL (in mg/dL) of the *i*th observation. We use robust standard errors. Since my predictor of interest is LDL, my coefficient of interest is and I will make inference using my estimate, , which is the estimated absolute difference in probability of 5-year all-cause mortality between individuals whose LDL levels differ by 1 mg/dL. We are interested in testing the hypothesis:

**Inference:** The linear regression estimates that an individual whose LDL level is 1 mg/dL higher than another’s LDL level would have a risk of 5-year mortality that is 0.10% lower ( = -.0010). The corresponding 95% CI of this estimated risk difference is [-0.0019, -0.0002], with a two-sided p-value of 0.017. As a result, we have evidence to reject the null hypothesis of there being no true risk difference between 5-year mortality and a continuous measure of LDL at the .05 level of significance (p-value = 0.017 < .05, n=725).

 **(b)**

**Method:** I conduct Poisson regression (not assuming homoscedasticity), with the model

,

where is an indicator of whether the *i*th observation died within 5 years and is a continuous measure of LDL of the *i*th observation. We use robust standard errors to account for the possibility of heteroskedasticity. Since my predictor of interest is LDL level, my coefficient of interest is and I will make inference using my estimate, , which is the estimated 5-year all-cause mortality probability ratio of individuals whose LDL levels differed by 1 mg/dL. We will use Wald-based confidence intervals for these risk ratios. We are interested in testing the hypothesis:

which is equivalent to testing

**Inference:** The Poisson regression model estimates a risk ratio for 5-year all-cause mortality of That is, a group of elderly individuals would be expected to have a 5-year mortality risk that is 99.4% of the 5-year mortality risk for another group whose LDL is 1 mg/dL lower. The Wald-based 95% CI associated with this estimate is [0.988, 0.999]. We note that 1 is not contained within this interval. Our estimate would not be surprising if the true 5-year mortality risk of a group of elderly individuals were between 98.8% and 99.9% of the risk for a group of elderly individuals whose LDL is 1 mg/dL lower. The two-sided Wald-based p-value associated with this estimate is 0.018. As a result, we have evidence to reject the null hypothesis of the true 5-year mortality risk not being associated with serum LDL the .05 level of significance (p-value = 0.018 < .05, n=725).

(c)

**Method:** We use logistic regression to examine the ratio of 5-year mortality odds between individuals with different levels of serum LDL (in mg/dL). We fit the following logistic regression model

,

where is an indicator of whether the *i*th observation died within 5 years and is the serum LDL level (in mg/dL) of the *i*th individual. We use robust standard errors since we do not assume homoskedasticity. Since my predictor of interest is LDL level, my coefficient of interest is and I will make inference using my estimate, (, which is the estimated odds ratio for 5-year mortality odds between individuals whose LDL levels differ by 1 mg/dL. We also compute a Wald-based confidence interval for this estimated odds ratio. We are interested in testing the hypothesis:

**Inference:** From our logistic regression, we estimate that the ratio of 5-year mortality odds between individuals whose LDL levels differ by 1 mg/dL is That is, the estimated odds of 5-year mortality for a given individual are 99.2% of the mortality odds of an individual whose LDL is 1 mg/dL lower. The Wald-based 95% CI associated with this estimate is [0.985, 0.999]. We note that 1 is not contained in the CI. Our estimate would not be surprising if the true 5-year mortality odds of a group of elderly individuals were between 98.5% and 99.9% of the odds for a group of elderly individuals whose LDL is 1 mg/dL lower. The two-sided Wald-based p-value associated with this estimate is 0.019. As a result, we have evidence to reject the null hypothesis of the true 5-year mortality odds not being associated with serum LDL the .05 level of significance (p-value = 0.019 < .05, n=725).

**(d)**

Our conclusions reached in problems 1-3 of this homework were different than the conclusions we reached in problems 2 and 4 of Homework 2. Here, we failed to reject the null hypotheses of there being no association between 5-year mortality (our response variable) and an indicator of LDL ≥ 160 mg/dL (our predictor if interest) under logistic, linear, and Poisson regression. In Homework 2, on the other hand, we had sufficient evidence to reject the null hypothesis of there being no association between high LDL (our response variable instead of our predictor of interest) and 5-year mortality (our predictor instead of our response).

*A priori,* I would prefer the analyses in this homework. I think classifying 5-year mortality as an outcome of interest (or response variable) and serum LDL level as a predictor of interest would allow us to address a meaningful scientific question. I believe that treating LDL as an outcome of interest can make sense in some settings (for example if we are testing the effectiveness of a proposed cholesterol-lowering drug), but with the data and study design we are working with, it makes more sense to think of LDL as a predictor. I would also prefer to keep LDL as a continuous measurement so we do not lose information about its distribution as we would by dichotomizing it and to look at odds ratios, so, *a priori,* I would have preferred the analysis in problem 4, part (c) of this homework.