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 > 160 mg/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.

Saturated. We have two parameters in the model and predictor is split into two groups.

* 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?

Odds of dead within 5 years =0.205 (intercept from the logistic regression). p = 0.17 (calculated by hand using formula $\frac{p}{1-p}$).

|  |  |  |  |
| --- | --- | --- | --- |
|  | Survive at least 5 yr | Dead within 5 yr | Total |
| LDL (mg/dL) |  |  |  |
| LOW | 513 | 105 | 618 |
| HIGH | 93 | 14 | 107 |
|  | P(Dead within 5 yr)  | Odds(Dead within 5 yr) |
| LOW | 0.17 | 0.205 |
| HIGH | 0.131 | 0.151 |

For low LDL, observed proportion of death with 5 years = 105/618 = 0.1699 = 0.17. So they are the same.

* 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 high LDL dying within 5 years?

Odds of dead within 5 years =0.1505 (since odds ratio is the slope from the logistic regression, we get the odds of high LDL group by multiplying slope by intercept). Probability = 0.131 (calculated by hand using formula $\frac{p}{1-p}$). According to the table for observed value above, for high LDL, observed proportion of death with 5 years = 14/107 = 0.131. So they are the same.

* 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?

We compared two groups with different LDL levels, the odds of dying within 5 years is estimated to be 26.5% lower in the high LDL group (odds ratio 0.735) than the group with low LDL. This observed difference is not statistically different from an odds ratio of 1 (P = 0.316), with 95% confidence interval suggesting that the observed odds ratio is what might be typically observed if the true odds of dying within 5 years for high LDL group is anywhere between 59.6% lower and 34.04% higher comparing to the true odds of death for low LDL group. We thus fail to reject the null hypothesis of no association between 5-year mortality and LDL level.

By looking at the question 5, 6 from homework 1, we can tell that the probability, odds and odds ratio are exactly the same. Confidence Interval is a bit narrower [404, 1.34] and p-value is slightly off (0.316 compare to 0.314). This might be because logistic regression model use score based C.I and p-value while chi-square test computed the cornfield C.I.

* 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?

a-c won’t change at all. If we used an indicator of low LDL as predictor, point estimate of odds ratio would be the reciprocal of the point estimator (slope) we obtained from part d. That’s because after the reparameterization, we flipped the odds in the numerator and denominator. Therefore, part a would be the same since the new model holds the same character as the prior one. b and c won’t change since odds and probability are the state of nature of the data given research interest which won’t depend on the model. We won’t lose any information through reparameterization so we can always compute those values through the new estimators by manipulating the formula.

If we used an indicator of survival for at least 5 years as the response variable, new slope will be the reciprocal of the one in part d but we will have a different intercept (4.89). It’s the odds of survival at least 5 years for low LDL group.

* 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?

Odds ratio wouldn’t change. but the odds changed. If we flip the response variable and explanatory variable, the intercept would be odds of having high LDL given that the subject survived at least 5 years. Thus, we can derive the probability and odds of dying within 5 years for low and high LDL group using Bayes theorem. a-c won’t change.

Note that this only works if we have saturated model which gives us the full information of the data (it can be expressed as the 2 by 2 table).

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 > 160 mg/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.

Yes. We have two parameters in the model and predictor is split into two groups.

* 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?

For low LDL group, estimated probability of dying within 5 years is 17.0% (Intercept from least square regression is 0.1699). Estimated odds is 0.205. They are the same as the observed value.

* 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?

For high LDL group, estimated probability of dying within 5 years is 13.1% (Intercept + slope from least square regression is 0.1308). Estimated odds is about 0.151. They are the same as the observed value.

* 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?

The estimator for difference in probability of death within 5 years across the group (group with low LDL – group with high LDL) is 3.9%. The difference is not significantly different from 0 because the p-value is 0.278 (>0.05). The 95% confidence Interval suggested that the observed difference would be reasonable is the true probability of dying with 5 years of low LDL group is from 3.2% lower to 11% higher comparing to high LDL group. Thus we cannot reject null hypothesis that states no difference of probability of death within 5 years across the groups (no association between LDL level and mortality).

Point estimate are exactly the same comparing to the 5 and 6 in homework 1, but the p-value and 95% C.I are slightly different. Again, this might be caused by the error of using either wald-based SE or robust standard error.

* 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?

Again, this is just parameterizing the model. So a-c will have exactly the same answer.

If we used an indicator of survival for at least 5 years as the response variable, slope will be 0.039 and intercept is 0.83. The baseline values and interpretation is changed but we can always derive b and c. a won’t change because re-parameterization don’t change the character of the model.

* 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?

Again, we flipped the response and predictor. So we would get the probability of a subject being in the high LDL group given he/she died within 5 years or not. Using bayes theorem, we can derive odds and probability of dying with 5 years for high/low LDL group so answer a-c won’t change.

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 > 160 mg/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.

Yes. We have two parameters in the model and predictor is split into two groups.

* 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?

Estimated p = $Exp\left[Intercept from possion regression\right]=e^{-1.773}=0.169 ≈17\%.$ Thus the estimated odds would be 0.205. It’s the same as the observed proportion.

* 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?

Estimated p = $Exp\left[Intercept from possion regression+slope estimate from the poisson regrssion\right]=e^{-1.773-0.261}=0.1308 ≈13.1\%$. Thus the estimated odds would be 0.151. And they are the same as the values we observed.

* 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?

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Death in 5 year | Coef. | Robust SE | p value | 95% C.I |  |
| slope | -0.261 | 0.265 | 0.324 | [-0.780, | 0.258] |
| intercept | -1.773 | 0.089 | 0 | [-1.947, | -1.598] |

The estimate of risk ratio or ratio of probability of death within 5 years across groups [$ \frac{p(death in high LDL group)}{p(death in low LDL group)}]$ is 0.77 (slope of poison regression is -0.261). This is the same as saying the probability of death within 5 years for high LDL group is 23% lower than that probability for low LDL group. The ratio is not significantly different from 1 according to the p-value (p-value = 0.324 > 0.05). According to the 95% confidence interval, the probability ratio is reasonable if the true probability of death within 5 years for high LDL group is anywhere from 54.2% lower to 29.3% higher comparing to the true probability of death for low LDL group.

Maximum likelihood estimation is used to calculate the parameter estimator. But as a saturated model, they are the same as the value from prior question (after transformation back). Also P values and CI that are calculated for each parameter estimate are Wald- based estimates, which might be the source causing the difference.

* 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?

Flip the dummy variables (indicators) would only change our reference as the baseline value of the model but won’t give us any more or less information. Thus answer to a-c will stay the same.

* 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?

We won’t lose any information by flipping the response and explanatory. So a-c won’t change. The reason is explained in 1 (f) and 2 (f).

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).

Apply least square regression model with assumption of unequal variance. The point estimate for intercept is 0.294 but here it’s meaningless (We don’t care the probability of death with 5 years for subjects with LDL level = 0. We are going to ignore this for b and c). The slope estimate is -0.001. This is saying when the LDL level goes up by 1 mg/dL, the probability of subjects dying within 5 years goes down by 0.1%. The slope estimate is significantly different than 0 (p-value = 0.017 less than 0.05). And the 95% confidence interval tells us the slope estimate observed is reasonable if the probability of subjects dying within 5 years goes down in the range from 0.19% to 0.018% when LDL level goes up by 1 mg/dL. Thus, we conclude that we reject null hypotheses and in favor of the alternative hypotheses which suggesting there’s association between mortality and LDL level.

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

Apply poisson regression model with assumption of unequal variance. The slope estimate is -0.00647. Risk ratio would the exp(-0.00647) = 0.9936. This is saying when LDL level increased by 1 mg/dL, the probability of death in 5 years decreases by 0.645%. The observation is significant since p-value = 0.018 < 0.05. 95% confidence interval suggested that observed risk ratio is reasonable if the true probability of subjects dying within 5 years decrease in the range from 1.17% to 0.11% when LDL level increased by 1 mg/dL. Thus, we conclude that we reject null hypotheses and in favor of the alternative hypotheses which suggesting there’s association between mortality and LDL level.

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

Apply logistic regression model with assumption of unequal variance. The slope estimate is 0.992 (odds ratio). This is saying when LDL level increased by 1 mg/dL, the odds of death in 5 years decreases by 0.8%. The observation is significant since p-value = 0.019 < 0.05. 95% confidence interval suggested that observed odds ratio is reasonable if the true odds of subjects dying within 5 years decrease in the range from 0.125% to 1.42% when LDL level increased by 1 mg/dL. Thus, we conclude that we reject null hypotheses and in favor of the alternative hypotheses which suggesting there’s association between mortality and LDL level.

* 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*.?

First comparing to 1-3, models in part 4 is better because using continuous LDL level won’t let us lose any information collected whereas dichotomizing LDL definitely excluded some information although it’s convenient.

Second, comparing to homework 2, we prefer the model in this homework because we tend to be more interested in probability of death given having high or low LDL comparing to the probability of being in the high/low LDL group given death/survive.

Finally, I think logistic regression model best fits our research interest. Logistic regress allows the true relationship between variables to be non-linear.

So overall, I prefer logistic regression.