1. Below is a table of inferential statistics obtained from logistic regression analysis

|  |  |  |  |
| --- | --- | --- | --- |
|  | Odds ratio | 95% CI | p-value |
| LDL>=160 mg/dl | 0.7355 | (0.4036, 1.3404) | 0.316 |
| Intercept  | 0.2047 | (0.1659, 0.2525) |  |

(a)

There are two parameter estimates for the two parameters used. Therefore the model is saturated. Even though LDL is a continuous variable, it is dichotomized to high and low status hence it is considered as a binary variable.

(b)

When comparing two groups with different LDL levels, the odds that a subject with low LDL dies within 5 years is estimated to be 0.2047 while the estimated probability of dying within 5 years is 16.99% (p=odds/(1+odds)). Of the 618 subjects who had low LDL, 105 died within 5 years giving us an observed proportion of 16.99%. This proportion estimate is exactly the same as the estimated probability of death within 5 years. In contrast to the observed odds estimate, there was a 3.48% absolute difference between the two.

(c)

When comparing two groups with different LDL levels, the odds that a subject with high LDL dies within 5 years is estimated to be 0.1505 while the estimated probability of death within 5 years is 13.08% (p=odds/(1+odds)). Of the 107 subjects who had high LDL, 14 died within 5 years giving us an observed proportion of 13.08%. This proportion is exactly the same as the probability estimate. In contrast to the odds, there is a 1.97% absolute difference between the two.

(d)

The problem doesn’t specify how LDL is to be modelled and I consider LDL untransformed which estimates the common odds ratio for each additive difference in LDL. It’s also not specified in the question which regression method to use and I proceed with logistic regression with robust standard error estimates.

When comparing two groups with different LDL levels, the odds of dying within 5 years of study enrollment is estimated to be 26.45% (odds ratio 0.7355), with the group having higher LDL tending toward lower odds of death within 5 years. The observed difference is not statistically different from 1 (P=0.316), with a 95% confidence interval suggesting that the observed odds ratio is what might be typically observed if the true odds of dying within 5 years were anywhere between 0.4036 and 1.3404. We thus with high confidence, fail to reject the null hypothesis of no association between survival time and LDL level.

Compared to inference in problems 5 and 6 of homework 1, the observed p-value is almost the same as the one obtained in problem 5 using chi-square test (p=0.314). There is however a relatively significant difference compared to the p-value obtained in problem 6 using Fisher’s exact test (p=0.396). There is also a slight variation in the confidence interval obtained using Fisher’s exact test (0.373, 1.36).

Fisher’s exact test computes the exact statistics as compared to logistic regression which normally relies on approximation (Wald statistics). This explains the variation of observed odds ratio, p-value, and confidence intervals. Compared to the chi-square test used for problem 5 of homework 1, the small variations are as a result of the method of approximation used. Chi-square uses the score statistics and Cornfield CI estimates while logistic regression uses Wald statistics.

(e)

Fitting a logistic regression model with low LDL as the predictor is a re-parameterization of the model used in parts a-c (low LDL=1-high LDL). Therefore, the model will be saturated and the p-value remains the same. In addition, we can use this model to derive the odds used in part a-c. Therefore inference using this model will be similar to the one obtained using the model used in a-c.

Similarly using survival for at least 5 years as the response variable is a re- parameterization of the model used in parts a-c (dead in at least 5=1-dead within 5). The model will be saturated and the p-values remain the same. Thus inference using this model will be similar to the one obtained using model used in a-c.

 (f)

The model obtained in this case will be saturated and therefore no changes on answer to part (a) of this problem. The p-values observed for parts b-c stay the same. In regard to parts b and c, since we only interchanged the roles of the predictor and the response variable, odds ratio are invariant. Bayes rule can be used to estimate the probabilities which can be used to estimate the odds that a subject has high LDL given that they die within 5 years. In this case the odds estimate are 0.1333.

Bayes rule: $p(ldl\geq 160|deadin5)=\frac{p\left(ldl\geq 160\right)\*p(ldl\geq 160)}{p(deadin5)}$

1. Below is a table of inferential statistics from the linear regression analysis.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Coefficient estimates | 95% CI | p-value |
| LDL>=160 mg/dl | -0.03906 | (-.1097, .03158) | 0.278 |
| Intercept  | 0.1699 | (.1402, .1996) |  |

(a)

The model is saturated since there are two parameter estimates corresponding to each of the two groups (high and low LDL).

(b)

For subjects with low LDL, the estimated probability of dying within 5 years is 16.99% while the estimated odds is 0.2047. Of the 618 subjects who had low LDL, 105 died within 5 years giving us an observed proportion of 16.99%. This proportion estimate is exactly the same as the estimated probability of death within 5 years. In contrast to the observed odds estimate, there is a 3.48% absolute difference between the two.

(c)

For subjects with high LDL, the estimated probability of death within 5 years is 13.08% while the estimated odds are 0.1505. Of the 107 subjects who had high LDL, 14 died within 5 years giving us an observed proportion of 13.08%. Comparing the odds to probability and proportion, we observe an absolute difference of 2.97%.

(d)

From the robust linear regression analysis, the risk difference for those subjects dying within 5 years compared to those that are not is 0.0391 (on average), for each mg/dl difference in LDL, with the subjects having high LDL tending towards lower risk of death. This result is not significantly different from 0 (two sided p=0.278), with 95% CI suggesting that such results would not be unusual if the true risk difference were anywhere between 10.97% lower to 3.158% higher for each mg/dl difference in LDL, with subjects having high LDL tending towards lower risk. We thus with high confidence fail to reject the null hypothesis of no association between risk of death and LDL levels.

Compared to problem 5 and 6 homework 1, there are some differences in the parameter estimates. In this problem Huber sandwich estimator is used in the robust t test while as mentioned in problem 1 d, Chi-square uses score test and Cornfield CI’s while Fisher’s exact uses the exact estimates.

(e)

Fitting a linear regression model with low LDL as the predictor is a re-parameterization of the model used in parts a-c (low LDL=1-high LDL). Therefore, the model will be saturated and the p-value remains the same. In addition, we can use this model to derive the odds used in part a-c. Therefore inference using this model will be similar to the one obtained using model used in a-c.

Similarly using survival for at least 5 years as the response variable is a re- parameterization of the model used in parts a-c (dead in at least 5=1-dead within 5). The model will be saturated and the p-values remain the same. Thus inference using this model will be similar to the one obtained using model used in a-c.

(f)

The model obtained in this case will be saturated and therefore no changes on answer to part (a) of this problem. The p-values observed for parts b-c stay the same. In regard to parts b and c, since we only interchanged the roles of the predictor and the response variable, odds ratio are invariant. Bayes rule can be used to estimate the probabilities which can be used to estimate the odds that a subject has high LDL given that they die within 5 years.

3.

Below is a table of inferential statistics obtained from Poisson regression analysis

|  |  |  |  |
| --- | --- | --- | --- |
|  | Incidence rate ratios | 95% CI | p-value |
| LDL>=160 mg/dl | 0.7701 | (0.4584, 1.2939) | 0.324 |
| Intercept  | 0.1699 | (0.1427, 0.2023) |  |

(a)

This is a saturated model since there are two parameter estimates corresponding to the two LDL groups.

(b)

The estimated probability of dying within 5 years for subjects with low LDL is 0.1699 (exp -1.772528) and the estimated odds are 0.2047. This probability estimate is exactly the same as the proportion estimate.

(c)

The estimated probability of dying within 5 years for subjects with high LDL is 0.1308 (exp (-1.772528 -.2612434)) and the estimated odds are 0.1505. This probability estimate is the same as the proportion estimate.

(d)

From the robust Poisson regression analysis comparing the risk of death of subjects with different LDL levels, the estimated risk of dying 5 years into study entry are 22.99% (risk ratio 0.7701) for each mg/dl difference in LDL with the subjects having high LDL, tending towards lower risk of death within 5 years. This observed difference is not statistically different from 1 (two sided p-value 0.324), with a 95% CI suggesting that the observed risk ratio is what might typically be observed if the true risk ratio of death within 5 years were anywhere between 54.16% lower and 23.39% higher for each mg/dl higher LDL level. We therefore fail to reject the null hypothesis of no association between survival time and LDL at study entry.

Compared to problem 5 and 6 homework 1, there are some differences in the parameter estimates. In this problem Wald estimates are used while as mentioned in problem 1 d, Chi-square uses score test and Cornfield CI’s while Fisher’s exact uses the exact estimates.

(e)

Fitting a Poisson regression model with low LDL as the predictor is a re-parameterization of the model used in parts a-c (low LDL=1-high LDL). Therefore, the model will be saturated and the p-value remains the same. In addition, we can use this model to derive the odds used in part a-c. Therefore inference using this model will be similar to the one obtained using model used in a-c.

Similarly using survival for at least 5 years as the response variable is a re- parameterization of the model used in parts a-c (dead in at least 5=1-dead within 5). The model will be saturated and the p-values remain the same. Thus inference using this model will be similar to the one obtained using model used in a-c.

(f)

The model obtained in this case will be saturated and therefore no changes on answer to part (a) of this problem. The p-values observed for parts b-c stay the same. In regard to parts b and c, since we only interchanged the roles of the predictor and the response variable, odds ratio are invariant. Bayes rule can be used to estimate the probabilities which can be used to estimate the odds that a subject has high LDL given that they die within 5 years.

4.

(a)

From the robust linear regression analysis, the risk difference for those subjects dying within 5 years compared to those that are not is 0.001034 (on average), for each mg/dl difference in LDL, with the subjects having high LDL tending towards lower risk of death. This result is significantly different from 0 (two sided p=0.017),

with 95% CI suggesting that such results would not be unusual if the true risk difference were anywhere between 0.188% lower to 0.0185% lower for each mg/dl difference in LDL, with subjects having high LDL tending towards lower risk. We thus with high confidence reject the null hypothesis of no association between risk of death and LDL levels, in favor of a hypothesis that the risk of death is lower for subjects with high LDL .

(b)

From the robust Poisson regression comparing the risk of death of subjects with different LDL levels, the estimated risk ratio of dying 5 years within the study entry are 0.6448% (risk ratio .9935516) for each mg/dl difference in LDL with the subjects having high LDL, tending towards lower risk of death within 5 years. This observed difference is statistically different from 1 (two sided p-value 0.018), with a 95% CI suggesting that the observed risk ratio is what might typically be observed if the true risk ratio of death within 5 years were anywhere between 1.174% and 0.1123% lower for each mg/dl higher LDL level. We therefore reject the null hypothesis of no association between survival time and LDL at study entry in favor of a trend towards lower risk of death among subjects with high LDL levels.

(c)

From the robust logistic regression, comparing the subjects with different LDL levels, the odds of dying 5 years within the study entry are 0.7744% lower (odds ratio 0.9923) for each mg/dl difference in LDL with the subjects having high LDL, tending towards lower odds of death within 5 years. This observed difference is statistically different from 1 (two sided p-value 0.019), with a 95% CI suggesting that the observed odds ratio is what might typically be observed if the true odds of death within 5 years were anywhere between 1.42% lower and 0.13% lower for each mg/dl higher LDL level. We thus reject the null hypothesis of no association between survival time and LDL at study entry in favor of a trend toward higher odds of survival among subjects with high LDL levels.

(d)

The conclusion in this problem is different from the conclusions reached in problems 1-3 of this homework. For problems 1-3 of this homework, we failed to reject the null but we do reject the null in this part of the problem. The lack of significance to parts 1-3 of this problem may be attributed to loss of information due to dichotomization of the LDL which is a continuous variable. Compared to problems 2 and 4 of homework 2, the conclusion is the same.

I would prefer fitting logistic regression model as a priori in this analyses since the response variable is binary. In addition, I would also prefer fitting LDL as a continuous variable to avoid any loss of information that may arise from dichotomizing this variable. Since it would also be more clinically important to understand how death of subjects is associated with LDL rather than how LDL levels are associated with death of subject, I would further consider fitting a regression analysis model of the distribution of death within 5 years across groups defined by the continuous measure of LDL.