Homework 3

1. We perform a logistic regression analysis where the response variable is the indicator of death within 5 years and the predictor of interest is the indicator variable of high LDL, where high LDL is classified to be LDL levels greater than or equal to 160 mg/dL. We also use robust standard errors since there is no reason to assume that the group variances are equal.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Logistic regression with robust standard error** | | | | |  |  |
| **deadin5** | **Odds Ratio** | **Std. Err.** | **z** | **P>|z|** | **[95% Conf.** | **Interval]** |
| **LDL > 160** | 0.735 | 0.225 | -1.000 | 0.316 | 0.404 | 1.340 |
| **constant** | 0.205 | 0.022 | -14.810 | 0.000 | 0.166 | 0.253 |

1. This is a saturated model since there are two parameters and two groups. The parameters are the slope and the intercept, and the two groups are the group of people with LDL levels greater than or equal to 160 mg/dL and the group of people with LDL levels lower than 160 mg/dL.
2. We can use the Stata command *logit* to obtain so that inferences will be returned on the log odds scale (as opposed to the odds ratio scale in logistic regression).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Logistic regression with robust standard error (reported in logit scale)** | | | | | | |
| **deadin5** | **Coef.** | **Std. Err.** | **z** | **P>|z|** | **[95% Conf.** | **Interval]** |
| **LDL > 160** | -0.307 | 0.306 | -1.000 | 0.316 | -0.907 | 0.293 |
| **constant** | -1.586 | 0.107 | -14.810 | 0.000 | -1.796 | -1.376 |

For subjects with low LDL (LDL < 160 mg/dL), the estimated odds of dying within 5 years is 0.205. The estimated probability of dying within 5 years is 0.170. The observed proportion of subjects with low LDL dying in 5 years is 105/618 (or 0.170), which is exactly the same as our estimate (accounting for rounding).

**Descriptive Table of LDL level versus 5 year death**

|  |  |  |
| --- | --- | --- |
|  | **deadin5** |  |
| **highldl** | **0** | **1** |
| **0** | 513 | 105 |
| **1** | 93 | 14 |

1. For subjects with high LDL, the estimated odds of dying within 5 years is 0.151. The estimated probability of dying within 5 years is 0.131. The observed proportion of subjects with low LDL dying in 5 years is 14/107 (or 0.131), which is exactly the same as our estimate (seen in table above).
2. When comparing the group of subjects with high LDL (LDL>160 mg/dL) and without high LDL (LDL<160 mg/dL), the odds of dying in 5 years is estimated to be 26.5% lower (odds ratio 0.735) for the group of subjects with high LDL. The 95% confidence interval suggests that the observed odds ratio is what might be typically observed if the true odds of dying within 5 years was anywhere between 59.6% lower and 34.0% higher for subjects with high LDL. Based on the p-value of 0.316 as well as a very wide confidence interval, we cannot reject the null hypothesis that the odds of dying within 5 years is equal for the group of subjects with high LDL and the group of subjects without high LDL.

Comparing our result to problems 5 and 6 from homework 1, we can infer that our results here are nearly identical. Our point estimate of odds ratio is identical to problem 6, and our point estimate of proportions is identical to problem 5. However, our confidence interval for odds ratio and p-value (and thus proportion) are slightly different due to the way standard error is calculated. In logistic regression, Wald-based confidence intervals are used while in chi-squared test (from homework 1) Cornfield confidence intervals are used for inferences in odds ratio.

1. If we use the indicator of low LDL as our predictor, our point estimates are exactly the reciprocals of when we used high LDL as our predictor. The intercept would be based on the odds of death for subjects with high LDL (whereas before it was based on the odds of death for subjects with low LDL). Our confidence interval would also be calculated based on the reciprocals of our CI obtained from using high LDL as our predictor. If we used an indicator of survival for at least 5 years as the response variable, our inferences would be based on odds of survival instead of odds of death. Thus, our odds ratio would remain the same, but our inferences of odds would be based on survival and not death (thus our odds will be the reciprocals of odds obtained when using probability of death). The confidence intervals are also computed in the same way. We can think of this as a reparametrization of our regression model.
2. If we mimic the approach used in problems 1-4 of homework #2 and describe the distribution of LDL across groups defined by vital status, our answers to parts a-c will not change. We would still have a saturated model with two parameters and two groups, and we would still be able to make the same inferences. Our odds ratio is still the same, and the probability that we can infer using odds are still the same since our model is saturated (we are not borrowing any information from the model). Since we are changing the roles of our response variable and our predictor of interest, we can use the Bayes rule to obtain the same answers as we did in homework 2.
3. We perform a linear regression analysis where the response variable is the indicator of death within 5 years and the predictor of interest is the indicator variable of high LDL, where high LDL is classified to be LDL levels greater than or equal to 160 mg/dL. We also use robust standard errors since there is no reason to assume that the group variances are equal.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Linear regression with robust standard error** | | | |  |  |  |
|  |  | **Robust** |  |  |  |  |
| **deadin5** | **Coef.** | **Std. Err.** | **t** | **P>|t|** | **[95% Conf.** | **Interval]** |
| **LDL > 160** | -0.039 | 0.036 | -1.090 | 0.278 | -0.110 | 0.032 |
| **constant** | 0.170 | 0.015 | 11.230 | 0.000 | 0.140 | 0.200 |

1. This is a saturated model since there are two parameters and two groups. The parameters are the slope and the intercept, and the two groups are the group of people with LDL levels greater than or equal to 160 mg/dL and the group of people with LDL levels lower than 160 mg/dL.
2. For subjects with low LDL, the estimated probability of dying within 5 years is 0.170. The estimated odds of dying within 5 years is 0.205. The observed proportion of subjects with low LDL dying in 5 years is 105/618 (or 0.170), which is exactly the same as our estimate (accounting for rounding).

**Indicator of High LDL (LDL>160 mg/dL) versus Indicator of Death in 5 Years**

|  |  |  |
| --- | --- | --- |
|  | **deadin5** |  |
| **highldl** | **0** | **1** |
| **0** | 513 | 105 |
| **1** | 93 | 14 |

1. For subjects with high LDL, the estimated probability of dying within 5 years is 0.131. The estimated odds of dying within 5 years is 0.151. The observed proportion of subjects with high LDL dying in 5 years is 14/107 (or 0.131), which is exactly the same as our estimate (accounting for rounding).
2. From a linear regression analysis of 725 subjects using robust standard errors, we estimate a difference in probability of death within 5 years across groups defined by whether the subjects have high LDL or low LDL of 3.9%, where the group with subjects who have high LDL is associated with lower probability of death. The 95% CI suggests that the observed difference would not be surprising if the true difference in probability of death was between 11.0% lower and 3.2% higher for subjects with high LDL. Based on the p-value of 0.278 as well as the wide confidence interval, we cannot reject the null hypothesis that the difference in probability of death for subjects with high LDL and subjects with low LDL is zero.

The point estimate that we obtain from this analysis are equal to the point estimate we obtained in problem 5 of homework 1, but the standard error and confidence interval are calculated differently. The chi-square test in homework 1 uses exact confidence intervals (Wald Based) for the risk difference, while here we used linear regression with robust standard errors to compute the standard error and confidence intervals (thus different p-values as well).

1. If we fit a regression model using the indicator of death within 5 years as our response variable but using an indicator of low LDL as our predictor, our answers to parts a-c would not change since we could still make the same inferences. This regression model would be the same model as if we used indicator of high LDL, just that it would be a reparametrization of our parameters. If we used an indicator of survival for at least 5 years as the response variable, we would still be able to make the same inferences but the results of our test would be based on the probability of survival and not probability of death.
2. If we mimic the approach used in problems 1-4 of homework #2 and describe the distribution of LDL across groups defined by vital status, our answers to parts a-c will not change. We would still have a saturated model with two parameters and two groups, and we would still be able to make the same inferences. The probabilities that we can infer from our model are still the same since we are working with a binary response variable (again, we are not borrowing any information from the model from doing regression). Since we are changing the roles of our response variable and our predictor of interest, we can use the Bayes rule to obtain the same answers as we did in homework 2.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Poisson regression with robust standard error** | | | | |  |  |
| **deadin5** | **Coef.** | **Std. Err.** | **z** | **P>|z|** | **[95% Conf.** | **Interval]** |
| **LDL>160** | -0.261 | 0.265 | -0.990 | 0.324 | -0.780 | 0.258 |
| **constant** | -1.773 | 0.089 | -19.920 | 0.000 | -1.947 | -1.598 |

1. We perform a poisson regression analysis where the response variable is an indicator of death within 5 years and the predictor of interest is high LDL. We also use robust standard errors since there is no reason to assume that the variances across groups are equal.
2. This is a saturated model since there are two parameters and two groups. The parameters are the slope and the intercept, and the two groups are the group of people with LDL levels greater than or equal to 160 mg/dL and the group of people with LDL levels lower than 160 mg/dL.
3. For subjects with low LDL, the estimated probability of dying within 5 years is 0.170. The estimated odds of dying within 5 years is 0.205. The observed proportion of subjects with low LDL dying in 5 years is 105/618 (or 0.170), which is exactly the same as our estimate (accounting for rounding).
4. For subjects with high LDL, the estimated probability of dying within 5 years is 0.131. The estimated odds of dying within 5 years is 0.151. The observed proportion of subjects with high LDL dying in 5 years is 14/107 (or 0.131), which is exactly the same as our estimate (accounting for rounding).
5. From this poisson regression analysis, we estimate that the probability of death for subjects with low LDL is 0.170 (based on exponentiating the intercept parameter). The probability of death for subjects with high LDL is 23.0% lower (13.1%), based on the inference that the risk ratio of death between the subjects with high LDL and the subjects with low LDL is 0.770. A 95% CI suggests that this observation is not unusual if the probability of death for subjects with high LDL was between 54.2% lower and 29.4% higher than for subjects with low LDL. Based on the statistically insignificant p-value of 0.324, we are unable to reject the null hypothesis that the risk ratio between those two groups is equal to 1.

Comparing our inferences to problems 5 and 6 of homework 1, we can see that the results obtained from this analysis are almost the same. The source of discrepancy in our point estimates, standard errors (and thus the 95% CI) comes from the fact that the Poisson regression uses the method of Maximum Likelihood Estimation to estimate the parameters. The standard error and confidence intervals are based on Wald-based estimates, as opposed to the exact confidence intervals used by the chi-square test. For large samples, the two estimates are almost equal, but we do notice a slight discrepancy based on our sample size.

1. If we fit a regression model using the indicator of low LDL as our predictor variable, our answers to parts a-c would not change. We would be able to make exactly the same inferences as if we used the indicator of high LDL as our predictor variable, but our parameters would change so that the exponentiation of the intercept would give us the probability of death for subjects with high LDL (instead of low LDL) and the exponentiation of our slope would give us the risk ratio between the subjects with low LDL and the subjects with high LDL. We can think of this as a reparametrization of our regression model.
2. If we mimic the approach used in problems 1-4 of homework #2 and describe the distribution of LDL across groups defined by vital status, our answers to parts a-c will not change. We would still have a saturated model with two parameters and two groups, and we would still be able to make the same inferences. The probabilities that we can infer from our model are still the same since we are working with a binary response variable (again, we are not borrowing any information from the model from doing regression). Since we are changing the roles of our response variable and our predictor of interest, we can use the Bayes rule to obtain the same answers as we did in homework 2.
3. Regression analysis of the distribution of death within 5 years across groups defined by continuous measure of LDL:
4. We perform a linear regression analysis where the response variable is an indicator of death in 5 years and the predictor of interest is continuous measure of LDL.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Linear regression with robust standard error** | | | |  |  |  |
| **deadin5** | **Coef.** | **Std. Err.** | **t** | **P>|t|** | **[95% Conf.** | **Interval]** |
| **ldl** | -0.00103 | 0.00043 | -2.39000 | 0.01700 | -0.00188 | -0.00018 |
| **constant** | 0.29426 | 0.05803 | 5.07000 | 0.00000 | 0.18033 | 0.40819 |

From this linear regression analysis, we estimate that on average the difference in probability of death for two groups differ by 0.103% for each 1 mg/dL difference in LDL, with higher levels of LDL being associated with lower probability of death. The 95% CI suggests that this difference is what might typically be observed if the true difference in mean death rate for 5 years is between 0.188% lower and 0.018% lower for each 1 mg/dL increase in LDL. Based on the p-value of 0.017, we can reject the null hypothesis that there is no association between LDL levels and death in 5 years in favor of higher probability of survival for those with higher LDL levels. Also, we note that since 1 mg/dL may not be very clinically significant, it could be more advantageous to consider increases or decreases in intervals of 5 mg/dL or 10 mg/dL for more clinical reference.

1. We perform a poisson regression analysis where the response variable is an indicator of death in 5 years and the predictor of interest is continuous measure of LDL.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Poisson regression with robust standard error** | | | | |  |  |
| **deadin5** | **Coef.** | **Std. Err.** | **z** | **P>|z|** | **[95% Conf.** | **Interval]** |
| **ldl** | -0.00647 | 0.00273 | -2.37000 | 0.01800 | -0.01182 | -0.00112 |
| **constant** | -1.01637 | 0.32954 | -3.08000 | 0.00200 | -1.66226 | -0.37049 |

From this poisson regression analysis, we estimate that for each 1 mg/dL increase in LDL level, the probability of death decreases by 0.64%. Based on the 95% CI, this observation is not unusual if each 1 mg/dL increase in LDL level was associated with anywhere from a 1.18% to 0.11% decrease in mortality rate within 5 years. Based on the p-value of 0.018 which is significant at the 0.05 level, we can reject the null hypothesis that there is no association between LDL levels and death in 5 years in favor of higher probability of survival for those with higher LDL levels. Also, we note that since 1 mg/dL may not be very clinically significant, it could be more advantageous to consider increases or decreases in intervals of 5 mg/dL or 10 mg/dL for more clinical reference.

1. We perform a logistic regression analysis where the response variable is an indicator of death in 5 years and the predictor of interest is continuous measure of LDL.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Logistic regression with robust standard error** | | | | |  |  |
| **deadin5** | **Odds Ratio** | **Std. Err.** | **z** | **P>|z|** | **[95% Conf.** | **Interval]** |
| **ldl** | 0.99226 | 0.00330 | -2.34000 | 0.01900 | 0.98581 | 0.99875 |
| **constant** | 0.51056 | 0.20936 | -1.64000 | 0.10100 | 0.22855 | 1.14051 |

When comparing two groups with different LDL levels, the odds of dying within 5 years is estimated to be 0.774% lower for each 1 mg/dL difference in LDL level, where higher levels of LDL is associated with lower odds of death. Based on the 95% CI, this observed difference is what might typically be observed if the true odds of dying within 5 years was anywhere between 0.125% and 1.419% lower for each 1 mg/dL increase in LDL level. Based on the p-value of 0.019 which is significant at the 0.05, we can reject the null hypothesis that there is no association between LDL levels and death in 5 years in favor of higher odds of survival for those with higher LDL levels. Also, we note that since 1 mg/dL may not be very clinically significant, it could be more advantageous to consider increases or decreases in intervals of 5 mg/dL or 10 mg/dL for more clinical reference.

1. Our conclusion is consistent with the conclusions from our previous analyses that higher levels of LDL are associated with lower probabilities of death. Although we did not get significant results in problems 1-3 of this homework (results were significant in problems 2 and 4 of homework #2), our analyses are consistent with the idea that higher levels of LDL are linked to higher probabilities of survival.

The conclusions reached from this model are also more reliable. In problems 1-3 we lost information from dichotomizing LDL levels, and in problems 2 and 4 from homework 2 we were modeling LDL levels based on the probability of death. Ideally, predicting the probability of death based on differing levels of LDL is more clinically significant to us than predicting LDL levels based on the probability of death.

A priori, I would have preferred logistic regression since we have a binary response variable. Also, using logistic regression we have greater possibility of avoiding major nonlinearities and avoiding effect modification due to the fact that the probabilities range from 0 to 1 while odds range from 0 to infinite.