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.*

This model is saturated. In each regression model, there are two distinct groups (those who died in 5 years and those who survived past 5 years). These are modeled with two regression parameters, the slope and the intercept.

* 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?*

Using a regression model on a log odds scale as described in the question stem, we estimate that:

Log odds dying within 5 years = -1.586 -0.307(“high LDL”). As we are interested in the intercept of this model, where the indicator variable “high LDL” is equal to zero, we exponentiate the values. The odds of dying within 5 years for subjects with low LDL is equal to e^-1.586 = 0.2047. The estimated probability of dying within 5 years for subjects with low LDL is 0.2047/1+0.2047 = 16.99%

The 2x2 table below compares the group of subjects with “high LDL” and “low LDL” with those who died within 5 years and those who survived beyond 5 years. Among those with low LDL, (105/618) 16.99% died within 5 years. These estimates are the same as the estimated probability of dying using our regression model.

|  |  |  |  |
| --- | --- | --- | --- |
|   | High LDL (>160mg/dL) | Low LDL (<160mg/dL) |  Total |
| Dies <5 years, n  | 14 | 105 | 119 |
| Survives >5 years, n | 93 | 513 | 606 |
| Total | 107 | 606 | 725 |

* 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?*

Again, using a regression model on a log odds scale as described in the question stem, we estimate:

Log odds dying within 5 years = -1.586 + (-0.307)(“high LDL”).

After exponentiating the values in our model, the odds of dying within 5 years = e^-1.586 \*e^(-0.307\*1) = 0.2047 \* 0.7356 = 0.1506. The estimated probability is 0.1506/1.1506 = 13.08%. The observed proportion of subjects with high LDL who die within 5 years is 14/107 = 13.08%. These estimates are equivalent.

* 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?*

From logistic regression analysis of 725 available observations of a sample of 735 elderly subjects aged 65-99, we estimate that patients who did not survive to 5 years and a serum LDL > 160mg/dl had an odds ratio of 0.7355 compared with patients with a serum LDL <160mg/dL. This means that the group with the “high LDL” had an odds of response 26.45% lower than subjects who had a “low LDL.” This finding is not statistically significant (p=0.316) with a 95% confidence interval that suggests this observation would not be unusual if the true odds of death among subjects with high LDL was between 59.64% less to up to 34.04% higher than subjects with low LDL. We cannot reject the null hypothesis that the odds of survival beyond 5 years is associated with serum LDL levels, classified as either high or low.

Compared to the inference from Homework 1, the proportions of subjects dying within 5 years who had LDL<160mg/dL were equivalent (17% in both models) and of subjects dying within 5 years who had LDL>160mg/dL (13.1% in both models). Also, the odds of dying for each group was equivalent, 0.205 for subjects with “low LDL” and 0.1506 for subjects with “high LDL.” The odds ratios were also the same, 0.735 in Homework #1 and 0.735 (0.1506/0.2047) .

* 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?*

This is reparameterization, where the predictor variable of “low LDL” is the same as “high LDL,” as the predictor variable is binary. The answers would be the same as long as we were explicit about which group was being compared relative to the other group. For example,

Log odds of dying within 5 years = intercept + slope (low LDL). In this model, the exponentiated intercept estimates the odds of dying within 5 years if the among those subjects with HDL, which would be the same as the answer in 1c. Likewise, the exponentiated intercept \* (slope\*1) gives us the estimate from 1b.

* 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?*

This changes the study design from a cohort study to a case-control study. The model remains saturated with two groups (low LDL and high LDL) and 2 parameters, slope and intercept. We can still estimate the odds ratio based on the slope or the exponentiated value of the slope, depending on whether we use a logit or logistic model in Stata. However, we cannot estimate the odds (and ignore the intercept value in this regression analysis) and cannot estimate the probability, which we calculated using the odds.

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, the model is saturated. There are two distinct groups in the response variable (those who died in 5 years and those who survived past 5 years). These are modeled with two regression parameters, the slope and the intercept.

* 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 probability of dying within 5 years = - 0.0391 \*(high LDL) + 0.1699 = 0.1699.

The odds of dying in 5 years for subjects with low LDL is 0.1699/(1-0.1699) = 0.2047. These estimates of the observed proportion of subjects with low LDL dying within 5 years is 105/606, which equals 0.1733. This value is 0.34 less than what our linear regression model estimated.

* 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 high LDL dying within 5 years?*

Estimated probability of dying within 5 years= - 0.0391 (high LDL) + 0.1308 = 0.1308

The estimated odds of dying within 5 years for subjects with high LDL is 0.1308/(1-0.1308) = 0.1505

Based on our 2x2 table in Problem 1, we observed 14/107 subjects with high LDL who did not survive to 5 years, or 13.08% of the sample.

* 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?*

From linear regression analysis of 725 available observations of a sample of 735 elderly subjects aged 65-99, we estimate the difference in probability of 5-year-all cause mortality is approximately 0.0391, or 3.91% less for subjects with high LDL (>160mg/dL) when compared with subjects with low LDL (<160mg/dL. This finding is not statistically significant (p=0.278) with a 95% confidence interval that suggests this observation would not be unusual if the true difference in probability of death was between 10.97% less to 3.16% higher for subjects with high LDL when compared with subjects with low LDL. We fail to reject the null hypothesis that the probability difference of survival beyond 5 years is associated with serum LDL levels, classified as either high or low.

Compared to the inference from Homework 1, the proportions of subjects dying within 5 years who had LDL<160mg/dL were equivalent (17% in both models) and of subjects dying within 5 years who had LDL>160mg/dL (13.1% in both models). Also, the odds of dying for each group was equivalent, 0.205 for subjects with “low LDL” and 0.1506 for subjects with “high LDL.” The odds ratios were also the same, 0.735 in Homework #1 and 0.735 (0.1506/0.2047) .

* 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?*

This is an example of reparameterization, where the predictor variable is binary. The answers would be the same as long as we were explicit about which group was being compared relative to the other group. For example, the difference in the probability of dying within 5 years = intercept + slope (low LDL). In this model, the only difference would be a change in the sign of the slope. If low LDL is the predictor variable, the absolute value would stay the same but it would now have a positive slope.

* 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?*

This question asks us to change our study design from a cohort study, to a case-control study which is grouped by survival status rather than the exposure, or serum LDL group. We would not be able to calculate the individual odds of mortality within 5 years given low LDL, but we could still calculate an odds ratio. We could also still calculate the probability using linear regression, but we cannot make estimates about the population because our groups were selected on mortality status first and are not representative. In this case, we cannot estimate the probability of mortality in 5 years given LDL status.

Instead, we can calculate the observed proportion of high LDL given mortality in 5 years using the 2x2 table. This value is 14/119 and represents an 11.76% chance of having a high LDL among those who died in the first 5 years.

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, the model is saturated. The response variable is binary and again has two groups in the response variable (those who died in 5 years and those who survived past 5 years). These are modeled with two regression parameters, the slope and the intercept.

* 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?*

Log rate = -1.7725 – 0.2612 (high LDL)

Estimated ratio of rates = e^-1.7725 = 0.1699.

The odds of dying within 5 years for these subjects is probability / (1-probabilty) = 0.1699/(1-0.1699) = 0.2047. These estimates are 0.34 less than the observed proportion of subjects with low LDL dying within 5 years, or 105/606, or 0.1733.

* 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 high LDL dying within 5 years?*

Log rate = -1.7725 + -0.2612 (high LDL)

Estimated probability of death in 5 years if high LDL = e^-1.7725 \* e^-0.2612 = (0.1699) (0.7701) = 0.1308. The observed proportion of subjects with high LDL who die in 5 years is 14/107, or 0.1308 of the sample which is equal to our estimate.

* 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?*

From Poisson regression analysis of 725 available observations of a sample of 735 elderly subjects aged 65-99, we estimate that for subjects with a serum LDL >160mg/dL, the probability of 5-year-all cause mortality decreases by approximately 22.99% (or is 77.01% as large) when compared with subjects with serum LDL <160mg/dL. This observation is not statistically significant with a p-value of 0.324 and 95% confidence interval that suggests this observation would not be unusual if the true probability of death for subjects with high LDL to be between 0.4584 and 1.2938. We cannot reject the null hypothesis that the probability of survival beyond 5 years is associated with serum LDL levels, classified as either high or low.

Compared to the inference from Homework 1, the proportions of subjects dying within 5 years who had LDL<160mg/dL were equivalent (17% in both models) and of subjects dying within 5 years who had LDL>160mg/dL (13.1% in both models). Also, the odds of dying for each group was equivalent, 0.205 for subjects with “low LDL” and 0.1506 for subjects with “high LDL.” The odds ratios were also the same, 0.735 in Homework #1 and 0.735 (0.1506/0.2047) .

* 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?*

If the indicator of low LDL served as a predictor rather than high LDL, the answers to a-c would not change much. This is an example of reparameterization, where the predictor variable is binary. The indicator low LDL = 1- high LDL. However, if we had changed the response variable to survival greater than 5 years, all values including odds, odds ratio, and probability would change. The model is estimated by:

Log rate = -0.1862 + 0.0459 (high LDL).

Compared with our previous estimate of: Log rate = -1.7725 + -0.2612 (high LDL)

* 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?*

If we change our study design to a case-control study, we cannot estimate odds or probability. We could still estimate the rate ratio.

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).*

METHODS:

We used linear regression analysis with robust standard error estimates to compare the association between 5 year mortality and serum LDL values. 725 available observations of a sample of 735 elderly subjects aged 65-99 were grouped by mortality status at 5 years, either “dead” or “alive.” We estimated the probability of 5 year mortality based on serum LDL values. LDL values were not available for 10 subjects and they were excluded from analysis.

RESULTS:

We estimate that the difference in the risk of 5 year mortality is 0. 1034% lower per unit of serum LDL, with risk decreasing as LDL rises. Based on a 95% confidence interval, we find that observing such an estimated difference is no unusual if the true association between LDL and 5 year mortality were such that the risk difference was between 0.1884% less per unit of cholesterol to 0.0185% higher per unit of cholesterol. This is statistically significant when alpha is 0.05 with a p-value of 0.017 and we can reject the null hypothesis that there is no association between 5-year mortality and serum LDL.

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

METHODS:

We used Poisson regression analysis with robust standard error estimates to compare the risk ratio, or ratio of the probability of 5 year mortality based on serum LDL values. 725 available observations of a sample of 735 elderly subjects aged 65-99, with 10 subjects excluded from analysis for missing LDL values.

RESULTS:

We estimate that for each 1 unit increase in LDL, the probability of 5 year mortality decreases by 0.6% with an estimated risk ratio was 0.994. For every 10-fold increase in LDL, the 5 year mortality decreases by 1.47%. The 95% confidence interval suggests that it would not be unusual if a subject experienced a 10-fold increase in serum LDL and could have a 5 year mortality rate that decreased between 0.26% to 2.68%. This observation is statistically significant when alpha is 0.05 with a p-value of 0.018, and we are able to reject the null hypothesis that there is no association between serum LDL and the risk ratio of 5 year mortality.

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

METHODS:

We used logistic regression analysis with robust standard error estimates to compare the odds of 5 year mortality by serum LDL levels. There were 725 available observations from a sample of 735 elderly subjects aged 65-99. 10 subjects were excluded from analysis as LDL data was unavailable.

RESULTS:

We estimate that for each unit increase in serum LDL, the odds of 5 year mortality decreases by 0.77%. This estimate is statistically significant when alpha <0.05 with a p-value of 0.019. A 95% confidence interval suggests that it would not be unusual for the odds of mortality to decrease between 0.13% and 1.42%. We reject the null hypothesis that there is no difference in odds of mortality by serum LDL.

* 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.?*

They fit the overall trend that higher serum LDL is associated with decreased risk for mortality. In Problems 1-3, none of these inferences reached statistical significance. These models benefit from the added power to detect differences we gain by making LDL as a continuous variable vs. dichotomizing it into a binary variable. Problems 2 and 4 from Homework 2 estimated a statistically significant difference in the mean LDL was 8.5mg/dL higher in the group that died within 5 years. This too in consistent with the overall trend that increasing serum LDL was associated with a decreased probability of 5 year mortality.

I would prefer to use linear regression to estimate the public health impact through risk difference for this analysis. The risk ratio estimated using Poisson regression is less useful because our event (5 year mortality) is not particularly rare and occurs > 10% of the time. Logistic regression would also be an appropriate choice because the association between mortality and LDL may be nonlinear, particularly as there is a relative range of possible values for LDL (we would not expect it to be very low, such as <10mg/dL or it would also generally be <300mg/dL). In fact, our sample has a range of LDL from 11mg/dL to 247 mg/dL. While using LDL as a continuous variable gives us more power, the practical utility of knowing that a 1 mg/dL increase in LDL decreases mortality by 0.6% is low. It is more useful and clinically relevant to consider LDL as a categorical variable. Instead, I would create three or four categories of LDL to increase the power of the analysis.

**Discussion Sections: January 22 – 14, 2014**

We continue to discuss the dataset regarding FEV and smoking in children. Come do discussion section prepared to describe the approach to the scientific question posed in the documentation file fev.doc.