Biost 518: Applied Biostatistics II
Biost 515: Biostatistics II
Emerson, Winter 2014

Homework #3 Key
January 27, 2014

***Instructions for grading*: Each part of each problem is worth 3 points, except parts relating to reporting the methods and results of association analyses are worth 10 points. Refer to the grading instructions for Homework #1 to see the requirements for reporting associations: Half the points are awarded for a valid analysis (providing the methods were described precisely enough that you can tell it was valid), and half the points are awarded for reporting all necessary information.**

Please insert comments on to the document indicating the points you have awarded for the problem, commenting on any reasons points were deducted.

***My answer to each question is provided in boldface type. In giving the answers, I sometimes provide alternative approaches in order that you can assess whether the numbers match up. I also provide some discussion of the choices or some additional material that I did not really expect to be provided in the answer. This additional information is provided in** normal type.*

Written problems: To be submitted as a MS-Word compatible file to the class Catalyst dropbox by 9:30 am on Monday, January 27, 2014. See the instructions for peer grading of the homework that are posted on the web pages.

> *On this (as all homeworks) Stata / R code and unedited Stata / R output is **TOTALLY** unacceptable. Instead, prepare a table of statistics gleaned from the Stata output. The table should be appropriate for inclusion in a scientific report, with all statistics rounded to a reasonable number of significant digits. (I am interested in how statistics are used to answer the scientific question.)*
>
> ***Unless explicitly told otherwise in the statement of the problem, in all problems requesting "statistical analyses" (either descriptive or inferential), you should present both***
> - ***Methods:** A brief sentence or paragraph describing the statistical methods you used. This should be using wording suitable for a scientific journal, though it might be a little more detailed. A reader should be able to reproduce your analysis. DO NOT PROVIDE Stata OR R CODE.*
> - ***Inference:** A paragraph providing full statistical inference in answer to the question. Please see the supplementary document relating to "Reporting Associations" for details.*

This homework builds on the analyses performed in homeworks #1 and #2, As such, all questions relate to associations among death from any cause, serum low density lipoprotein (LDL) levels, age, and sex in a population of generally healthy elderly subjects in four U.S. communities. This homework uses the subset of information that was collected to examine MRI changes in the brain. The data can be found on the class web page (follow the link to Datasets) in the file labeled mri.txt. Documentation is in the file mri.pdf. See homework #1 for additional information.

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 $\geq$ 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.)

a. Is this a saturated regression model? Explain your answer.

Ans: In the logistic regression model used to answer this question, two distinct groups (those with high LDL and those with low LDL) are modeled with two regression parameters (the intercept and the slope). It is a saturated model.

b. 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?

Ans: Because this is a saturated logistic regression model, the fitted odds of mortality within the low LDL group must agree exactly with the sample odds and sample proportion of subjects with low LDL who die within 5 years: sample odds = 105 / 513 = 0.205 and sample proportion = 105/ 618 = 0.170 from homework #1. To see this note that the logistic regression resulted in an estimated intercept of -0.15863. This corresponds to the log odds of 5 year mortality within the low LDL group, so the estimated odds would be $e^{-1.5863} = 0.20468$, so the estimated probability of 5 year mortality would be 0.20468 / (1 + 0.20468) = 0.1699.

c. 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?

Ans: Because this is a saturated logistic regression model, the fitted odds of mortality within the high LDL group must agree exactly with the sample odds and sample proportion of subjects with high LDL who die within 5 years: sample odds = 14 / 93 = 0.151 and sample proportion = 14/ 107 = 0.131 from homework #1. To see this note that the logistic regression resulted in an estimated intercept of -0.15863 and an estimated slope of -0.30723. Hence the fitted value for the high LDL group corresponds to a log odds of 5 year mortality of $-1.5863 + 1 \times (-0.30723) = -1.89353$, so the estimated odds would be $e^{-1.89353} = 0.15054$, so the estimated probability of 5 year mortality would be 0.15054 / (1 + 0.15054) = 0.1308.

d. 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?

Ans:

Methods: The odds of subjects dying within 5 years of study enrollment were compared between subjects who had serum LDL greater than or equal to 160 mg/dL and subjects whose serum LDL was measured to be 159 mg/dL or less using a logistic regression model. Statistical inference was based on the Wald statistic computed from the regression slope parameter and its standard error, with two-sided p value and 95% confidence interval computed using the approximate normal distribution for logistic regression parameter estimates. *(I did not use the robust SE, because in a saturated logistic regression model there is no real reason: the model based SE and the robust SE will agree very closely. Given that I was using classical logistic regression, I could have alternatively decided to use the p value from the likelihood ratio test.)*

Results: Of the 618 subjects whose serum LDL was less than or equal to 159 mg/dL, the odds of dying within 5 years from study enrollment was 0.205, while for the subjects with serum LDL greater than or equal to 160 mg/dL the odds of 5 year mortality was 0.151. Based on a 95% confidence interval, this observed odds ratio of 0.735 for the comparison of the high

LDL group to the low LDL group would not be judged unusual if the true odds ratio were anywhere between 0.404 to 1.340. A two-sided p value of 0.315 suggests that we can not with high confidence reject the null hypothesis that the odds of 5 year mortality are not associated with serum LDL levels. *(The likelihood ratio p value was 0.302.)*

The above results agree exactly with the estimated OR reported for problem 6 of homework #1. The 95% CI agrees exactly with the Woolf CI that might have been used in the answer for that problem, but differs somewhat from the exact CI for the OR that I reported in the key to homework #1. The p value from the Wald test (0.315) is very close to the p value from the chi squared test I reported in problem 5 of homework 1 (0.314). The chi squared test corresponds to the score test from logistic regression, which in very large samples will be equivalent to the Wald test, but in small samples can differ. Similarly, both the Wald and score tests are asymptotically equivalent to the likelihood ratio test, but in this sample the likelihood ratio test p value of 0.302 was a little different All of these p values differed from the typically more conservative Fisher's exact test p value of 0.396, though it should be noted that the Wald, score, and LR tests can be anti-conservative depending upon the sample size in each group and the level of confidence. *(In this case, unconditional exact tests based on the chi squared statistic would yield a two-sided p value of 0.373. Had the results corresponded to a p value closer to 0.05, the agreement between the three maximum likelihood based tests and the unconditional exact test would have been closer. With these sample sizes, the chi squared statistic and Wald statistic p values are at most 0.0015 too small for unconditional exact p values between 0.03 and 0.10, and at most 0.008 too small for unconditional exact p values between 0.01 and 0.03 (for instance, we might report chi squared statistic p values of 0.003 and 0.016 instead of more accurate 0.01 and 0.024, respectively).On the other hand, when the true unconditional exact p value is between 0.04 and 0.06, the Fisher's exact test p values tend to be 0.01 to 0.02 too high, and range as high as 0.10 instead of a true unconditional exact p value of 0.045.)*

e. 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?

Ans: These alternative models are all just re-parameterizations of the first model, owing to the linear transformation of the predictor and the fact that the response is a binary variable. Hence, all are saturated. Furthermore, whether the models were saturated or not, we can exactly predict the estimates and inference that would be reached with these other models, because they just represent linear transformations of the predictor variables and probabilities of complementary events sum to 1.

Notationally, suppose we fit the following four linear regression models, where *alivefor5 = 1 – deadin5* and *loLDL = 1 - hiLDL*:

$$\log\left[\frac{\Pr(deadin5=1\,|\,hiLDL)}{\Pr(deadin5=0\,|\,hiLDL)}\right]=\beta_0+\beta_1\times hiLDL \qquad \log\left[\frac{\Pr(alivefor5=1\,|\,hiLDL)}{\Pr(alivefor5=0\,|\,hiLDL)}\right]=\gamma_0+\gamma_1\times hiLDL$$

$$\log\left[\frac{\Pr(deadin5=1\,|\,loLDL)}{\Pr(deadin5=0\,|\,loLDL)}\right]=\alpha_0+\alpha_1\times loLDL \qquad \log\left[\frac{\Pr(alivefor5=1\,|\,loLDL)}{\Pr(alivefor5=0\,|\,loLDL)}\right]=\delta_0+\delta_1\times loLDL$$

Now, using the definition of our transformed variables, we note

$$\log\left[\frac{\Pr(deadin5=1\,|\,hiLDL)}{\Pr(deadin5=0\,|\,hiLDL)}\right]=\log\left[\frac{\Pr(alivefor5=0\,|\,hiLDL)}{\Pr(alivefor5=1\,|\,hiLDL)}\right]=-\log\left[\frac{\Pr(alivefor5=1\,|\,hiLDL)}{\Pr(alivefor5=0\,|\,hiLDL)}\right]$$

$$\Rightarrow \quad \beta_0=-\gamma_0 \quad \beta_1=-\gamma_1$$

$$\log\left[\frac{\Pr(deadin5=1\,|\,loLDL)}{\Pr(deadin5=0\,|\,loLDL)}\right]=\log\left[\frac{\Pr(alivefor5=0\,|\,loLDL)}{\Pr(alivefor5=1\,|\,loLDL)}\right]=-\log\left[\frac{\Pr(alivefor5=1\,|\,loLDL)}{\Pr(alivefor5=0\,|\,loLDL)}\right]$$

$$\Rightarrow \quad \alpha_0=-\delta_0 \quad \alpha_1=-\delta_1$$

and

$$\log\left[\frac{\Pr(deadin5=1\,|\,hiLDL)}{\Pr(deadin5=0\,|\,hiLDL)}\right]=\beta_0+\beta_1\times hiLDL=\beta_0+\beta_1\times(1-loLDL)=\beta_0+\beta_1+(-\beta_1)\times loLDL$$

$$\Rightarrow \quad \beta_0+\beta_1=\alpha_0 \quad \beta_1=-\alpha_1$$

Then correspondences among the intercepts and slopes are found as:

$$\beta_1=-\gamma_1=-\alpha_1=\delta_1 \qquad\qquad \beta_0=-\gamma_0=\alpha_0+\alpha_1=-\delta_0-\delta_1$$

In a fit of redundancy, I have included logistic regression output from all of these models in the Appendix to this homework.

f. 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?

Ans: These alternative models are interchanging the role of response variable and predictor of interest in logistic regression models. In this case, the regression model is fitting two groups (those dying within 5 years and those surviving for at least 5 years) using two regression parameters (intercept and slope), so it is a saturated model.

Owing to the fact that we are using logistic regression to model odds ratios, the invariance properties of the odds ratios guarantees that the slope estimate from a regression of 5 year mortality (response) on LDL category (predictor) will be exactly equal to the slope estimate from a regression of LDL category (response) on 5 year mortality (predictor) and that these slope estimates are each consistently estimating the same population quantity (population odds ratios) with the same precision. On the other hand, the intercepts from the two models are not estimating the same quantities, so the intercepts and fitted values within specific groups (as estimated in parts b and c) are not directly comparable across the models.

Notationally, we can consider two models:

$$\log\left[\frac{\Pr(deadin5=1\,|\,hiLDL)}{\Pr(deadin5=0\,|\,hiLDL)}\right]=\beta_0+\beta_1\times hiLDL \qquad\qquad \log\left[\frac{\Pr(hiLDL=1\,|\,deadin5)}{\Pr(hiLDL=0\,|\,deadin5)}\right]=\gamma_0+\gamma_1\times deadin5$$

Using the definition of conditional probability, we find

$$\log\left[\frac{\Pr(deadin5=1\,|\,hiLDL=h)}{\Pr(deadin5=0\,|\,hiLDL=h)}\right]=\log\left[\frac{\Pr(deadin5=1\,\&\,hiLDL=h)/\Pr(hiLDL=h)}{\Pr(deadin5=0\,\&\,hiLDL=h)/\Pr(hiLDL=h)}\right]$$
$$=\log\left[\frac{\Pr(deadin5=1\,\&\,hiLDL=h)}{\Pr(deadin5=0\,\&\,hiLDL=h)}\right]=\beta_0+\beta_1\times h$$
$$\log\left[\frac{\Pr(hiLDL=1\,|\,deadin5=d)}{\Pr(hiLDL=0\,|\,deadin5=d)}\right]=\log\left[\frac{\Pr(hiLDL=1\,\&\,deadin5=d)/\Pr(deadin5=d)}{\Pr(hiLDL=0*deadin5=d)/\Pr(deadin5=d)}\right]$$
$$=\log\left[\frac{\Pr(hiLDL=1\,\&\,deadin5=d)}{\Pr(hiLDL=0*deadin5=d)}\right]=\gamma_0+\gamma_1\times d$$

Using these expressions to isolate the slopes as log odds ratios, we find they are equal to each other:

$$\beta_1=\log\left[\frac{\Pr(deadin5=1\,\&\,hiLDL=1)}{\Pr(deadin5=0\,\&\,hiLDL=1)}\right]-\log\left[\frac{\Pr(deadin5=1\,\&\,hiLDL=0)}{\Pr(deadin5=0\,\&\,hiLDL=0)}\right]$$
$$=\log\left[\frac{\Pr(deadin5=1\,\&\,hiLDL=1)}{\Pr(deadin5=0\,\&\,hiLDL=1)}\frac{\Pr(deadin5=0\,\&\,hiLDL=0)}{\Pr(deadin5=1\,\&\,hiLDL=0)}\right]$$
$$\gamma_1=\log\left[\frac{\Pr(hiLDL=1\,\&\,deadin5=1)}{\Pr(hiLDL=0\,\&\,deadin5=1)}\right]-\log\left[\frac{\Pr(hiLDL=1\,\&\,deadin5=0)}{\Pr(hiLDL=0\,\&\,deadin5=0)}\right]$$
$$=\log\left[\frac{\Pr(hiLDL=1\,\&\,deadin5=1)}{\Pr(hiLDL=0\,\&\,deadin5=1)}\frac{\Pr(hiLDL=0\,\&\,deadin5=0)}{\Pr(hiLDL=1\,\&\,deadin5=0)}\right]$$

In these saturated models, then, it is clear the estimated slopes are equal. In fact, even when adjusting for other covariates, the estimated slopes are equal. Furthermore, this holds whether the sampling is cross-sectional (where we can estimate both *Odds[deadin5 | hiLDL]* and *Odds[hiLDL | deadin5]*), the sampling is within LDL category cohorts (where we can estimate *Odds[deadin5 | hiLDL]*, but not *Odds[hiLDL | deadin5]*), or the sampling is case-control by vital status at 5 years (where we can estimate *Odds[hiLDL | deadin5]*, but not *Odds[deadin5 | hiLDL]*). However, because of the noted restrictions on estimating some odds under cohort or case-control sampling, the intercepts and fitted values for individual groups from the logistic regression models might not be scientifically interpretable.

I have included logistic regression output from all of this "reversed" model in the Appendix to this homework.

2. 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 $\geq$ 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.)

    a. Is this a saturated regression model? Explain your answer.

Ans: In the linear regression model used to answer this question, two distinct groups (those with high LDL and those with low LDL) are modeled with two regression parameters (the intercept and the slope). It is a saturated model.

    b. 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?

Ans: Because this is a saturated linear regression model, the fitted probability of mortality within the low LDL group must agree exactly with the sample proportion and sample odds of subjects with low LDL who die within 5 years: sample proportion = 105/ 618 = 0.170 and sample odds = 105 / 513 = 0.205 from homework #1. To see this note that the linear regression resulted in an estimated intercept of 0.1699. This corresponds to the proportion of 5 year mortality within the low LDL group, so the estimated odds would be 0.1699 / (1 – 0.1699) = 0.20467.

c. 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?

Ans: Because this is a saturated linear regression model, the fitted probability of mortality within the high LDL group must agree exactly with the sample proportion and sample odds of subjects with high LDL who die within 5 years: sample odds = 14 / 93 = 0.151 and sample proportion = 14/ 107 = 0.131 from homework #1. To see this note that the linear regression resulted in an estimated intercept of 0.169903 and an estimated slope of -0.039062. Hence the fitted value for the high LDL group corresponds to a proportion of 5 year mortality of 0.169903 + 1 × (-0.039062) = 0.1308, so the estimated odds would be 0.1308 / (1 – 0.1308) = 0.15054.

d. 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?

Ans:

Methods: The probabilities of subjects dying within 5 years of study enrollment were compared between subjects who had serum LDL greater than or equal to 160 mg/dL and subjects whose serum LDL was measured to be 159 mg/dL or less using a linear regression model. Statistical inference on the difference in probabilities of death was based on the Wald statistic computed from the regression slope parameter and its standard error as estimated using the Huber-White sandwich estimator, with two-sided p value and 95% confidence interval computed using the approximate normal distribution for linear regression parameter estimates. *(I did use the robust SE, because a difference in probabilities dictates their will be a difference in variances across groups. Of course, for testing purposes, it would not be incorrect to use the classical standard error estimates, because under the null hypothesis of no difference in survival probabilities, the variances would be equal across groups in this simple regression model using a binary predictor. For the purposes of compute CI, the variances would not be equal for every alternative, but the robust SE does not really handle the mean-variance relationship that much better than the classical methods when computing the CI for a single binary predictor. It would make much more of a difference when adjusting for other covariates..)*

Results: Of the 618 subjects whose serum LDL was less than or equal to 159 mg/dL, the proportion dying within 5 years from study enrollment was 0.170, while for the subjects with serum LDL greater than or equal to 160 mg/dL the proportion for 5 year mortality was 0.131. Based on a 95% confidence interval, this observed difference of proportions of 3.91% suggesting higher survivor probabilities for the high LDL group compared to the low LDL group would not be judged unusual if the true difference in probabilities were anywhere from a 11.0% higher survival probability for the high LDL group to a 3.16% lower survival probability for the high LDL group.. A two-sided p value of 0.278 suggests that we can not with high confidence reject the null hypothesis that the probability of 5 year mortality is not associated with serum LDL levels. *(Using classical linear regression SE, the p value was 0.315.)*

The above results agree exactly with the estimated difference in proportions reported for problem 5 of homework #1. The 95% CI agrees is narrower than the Woolf CI that might

have been used in the answer for that problem, because of the way the mean-variance relationship is handled. Similarly the p value computed using robust SE (0.278) is lower than that from the chi squared test I reported in problem 5 of homework 1 (0.314). *(See the comments in the answer to problem 1d about the unconditional exact test. When using a binary response with a binary predictor, that is what I would really use.)*

e. 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?

Ans: These alternative models are all in some sense just re-parameterizations of the first model, because we are allowed to perform arbitrary linear transformations of both the response and predictor in linear regression. Hence, all are saturated. Furthermore, whether the models were saturated or not, we can exactly predict the estimates and inference that would be reached with these other models, because they just represent linear transformations of the variables.

Notationally, suppose we fit the following four linear regression models, where *alivefor5 = 1 – deadin5* and *loLDL = 1 - hiLDL*:

$$\Pr(deadin5 = 1 \mid hiLDL) = \beta_0 + \beta_1 \times hiLDL \qquad \Pr(alivefor5 = 1 \mid hiLDL) = \gamma_0 + \gamma_1 \times hiLDL$$
$$\Pr(deadin5 = 1 \mid loLDL) = \alpha_0 + \alpha_1 \times loLDL \qquad \Pr(alivefor5 = 1 \mid loLDL) = \delta_0 + \delta_1 \times loLDL$$

Now, using the definition of our transformed variables, we note

$$\Pr(deadin5 = 1 \mid hiLDL) = \beta_0 + \beta_1 \times hiLDL = 1 - \Pr(alivefor5 = 1 \mid hiLDL) = 1 - \gamma_0 - \gamma_1 \times hiLDL$$
$$\Rightarrow \quad \beta_0 = 1 - \gamma_0 \quad \beta_1 = -\gamma_1$$
$$\Pr(deadin5 = 1 \mid hiLDL) = \alpha_0 + \alpha_1 \times hiLDL = 1 - \Pr(alivefor5 = 1 \mid hiLDL) = 1 - \delta_0 - \delta_1 \times hiLDL$$
$$\Rightarrow \quad \alpha_0 = 1 - \delta_0 \quad \alpha_1 = -\delta_1$$

and

$$\Pr(deadin5 = 1 \mid hiLDL) = \beta_0 + \beta_1 \times hiLDL = \beta_0 + \beta_1 \times (1 - loLDL) = \beta_0 + \beta_1 + (-\beta_1) \times loLDL$$
$$\Rightarrow \quad \beta_0 + \beta_1 = \alpha_0 \quad \beta_1 = -\alpha_1$$

Then correspondences among the intercepts and slopes are found as:

$$\beta_1 = -\gamma_1 = -\alpha_1 = \delta_1 \qquad \beta_0 = 1 - \gamma_0 = \alpha_0 + \alpha_1 = 1 - \delta_0 - \delta_1$$

In a fit of redundancy, I have included linear regression output from all of these models in the Appendix to this homework.

f. 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?

Ans: This alternative model is interchanging the role of response variable and predictor of interest in linear regression models. In this case, the regression model is fitting two groups (those dying within 5 years and those surviving for at least 5 years) using two regression parameters (intercept and slope), so it is a saturated model.

Owing to the fact that we are using linear regression to model difference in proportions, the interpretation of the intercept and slope parameters will differ markedly between the model used in parts a-c and this "reversed" model. Hence, the intercepts and slopes from the two models are not

estimating the same quantities, so the fitted values within specific groups (as estimated in parts b and c) are not directly comparable across the models.

However, any statistical significance of the association should be approximately the same for the two models owing to the relationship between tests for nonzero slopes and test for nonzero correlation.

This can be easily seen by noting first that the classical regression t test for a statistically significant slope is exactly the same as the t test for a nonzero correlation. Because correlation treats the variables symmetrically, we have to get the same exact p value for the slope for a linear regression of Y on X as we would get for a linear regression of X on Y. The same property holds approximately when using the Huber-White sandwich estimator, though there will be some numerical differences due to the way the estimates are computed.

Notationally, we can consider two linear regression models:

$$E(Y \mid X = x) = \beta_0 + \beta_1 \times x \qquad E(X \mid Y = y) = \gamma_0 + \gamma_1 \times y$$

Now using a property of least squares estimates we know that for population correlation ρ_{XY} and populations standard deviations *SD(Y)* and *SD(X)*, the slopes are estimating:

$$\beta_1 = \rho_{XY} \frac{SD(Y)}{SD(X)} \qquad \gamma_1 = \rho_{XY} \frac{SD(X)}{SD(Y)}$$

and the least squares estimates are found using the sample correlation r_{XY} and sample standard deviations s_X and s_Y:

$$\hat{\beta}_1 = r_{XY} \frac{s_Y}{s_X} \qquad \hat{\gamma}_1 = r_{XY} \frac{s_X}{s_Y}$$

Clearly, the slopes will not be estimating the same numbers, but the statistical significance of the estimates will be similar.

Note that this same sort of reasoning will hold when we are adjusting for covariates W:

$$\beta_1 = \rho_{XY} \frac{SD(Y \mid W)}{SD(X \mid W)} \qquad \gamma_1 = \rho_{XY} \frac{SD(X \mid W)}{SD(Y \mid W)}$$

I have included linear regression output from the "reverse" model in the Appendix to this homework.

3. 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 $\geq$ 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.)

    a. Is this a saturated regression model? Explain your answer.

Ans: In the Poisson regression model used to answer this question, two distinct groups (those with high LDL and those with low LDL) are modeled with two regression parameters (the intercept and the slope). It is a saturated model.

    b. 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?

Ans: Because this is a saturated Poisson regression model, the fitted probability of mortality within the low LDL group must agree exactly with the sample proportion and sample odds of subjects with low LDL who die within 5 years: sample proportion = 105/ 618 = 0.170 and sample odds = 105 / 513 = 0.205 from homework #1. To see this note that the Poisson regression resulted in an estimated intercept of -1.77253, This corresponds to a mortality of $e^{-1.77253}$ = 0.1699 within the low LDL group, so the estimated odds would be 0.1699 / (1 – 0.1699) = 0.20467.

c. 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?

Ans: Because this is a saturated Poisson regression model, the fitted probability of mortality within the high LDL group must agree exactly with the sample proportion and sample odds of subjects with high LDL who die within 5 years: sample odds = 14 / 93 = 0.151 and sample proportion = 14/ 107 = 0.131 from homework #1. To see this note that the Poisson regression resulted in an estimated intercept of -1.77253 and an estimated slope of -0.2612434. Hence the fitted value for the high LDL group corresponds to a log rate of 5 year mortality of -1.77253 + 1 × (-0.2612434) = -2.03377, so the estimated mortality probability would be $e^{-2.03377}$ = 0.13084 and the estimated odds would be 0.1308 / (1 – 0.1308) = 0.15054.

d. 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?

Ans:

Methods: The probabilities of subjects dying within 5 years of study enrollment were compared between subjects who had serum LDL greater than or equal to 160 mg/dL and subjects whose serum LDL was measured to be 159 mg/dL or less using a Poisson regression model. Statistical inference on the ratio of the probabilities of death was based on the Wald statistic computed from the regression slope parameter and its standard error as estimated using the Huber-White sandwich estimator, with two-sided p value and 95% confidence interval computed using the approximate normal distribution for Poisson regression parameter estimates. *(I did use the robust SE, because a difference in probabilities dictates their will be a difference in variances across groups. Of course, for testing purposes, it would not be incorrect to use the classical standard error estimates, because under the null hypothesis of no difference in survival probabilities, the variances would be equal across groups in this simple regression model using a binary predictor. For the purposes of compute CI, the variances would not be equal for every alternative, but the robust SE does not really handle the mean-variance relationship that much better than the classical methods when computing the CI for a single binary predictor. It would make much more of a difference when adjusting for other covariates.)*

Results: Of the 618 subjects whose serum LDL was less than or equal to 159 mg/dL, the proportion dying within 5 years from study enrollment was 0.170, while for the subjects with serum LDL greater than or equal to 160 mg/dL the proportion for 5 year mortality was 0.131. Based on a 95% confidence interval, this observed ratio of probability of death of 0.770 suggesting 23.0% relative decrease in the probability of death for the high LDL group compared to the low LDL group would not be judged unusual if the true ratio of probabilities were anywhere from a 54.2% relative decrease (rate ratio = 0.458) in the survival probability for the high LDL group to a 29.4% relative increase (rate ratio = 1.294) in survival probability for the high LDL group. A two-sided p value of 0.321 suggests that we can not with high confidence reject the null hypothesis that the probability of 5 year mortality is not associated with serum LDL levels. *(Using classical Poisson regression SE, the p value was 0.359.)*

The above results agree are not particularly comparable to the results in homework #1, because we did not consider the rate ratio. Nonetheless, the results for the odds ratio analysis will be somewhat similar, though we cannot really consider event probabilities of 13% - 17% rare. The p value for the test of association is in the same ballpark as the other analyses.

e. 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?

Ans: These alternative models are not all just re-parameterizations of the first model, though all are saturated. When we use the same response variable, whether we use an indicator of high LDL or an indicator of low LDL as the predictor is just a re-parameterization: it is just a linear transformation of our predictor. However, a linear transformation of the response may lead to very different answers in Poisson regression: it is okay to re-scale the response (i.e., multiply by some constant), but it is not okay to "shift" the response (i.e., it is not okay to add some nonzero constant to all measurements). So we will in general get noncomparable answers when we analyze probability of death vs probability of survival. In the case of "shift" transformations in the setting of a saturated model, however, the fitted values will all be the same. The measures of association will not.

Notationally, suppose we fit the following four linear regression models, where *alivefor5 = 1 – deadin5* and *loLDL = 1 - hiLDL*:

$$\log\left[\Pr\left(deadin5=1\,|\,hiLDL\right)\right]=\beta_0+\beta_1\times hiLDL \qquad \log\left[\Pr\left(alivefor5=1\,|\,hiLDL\right)\right]=\gamma_0+\gamma_1\times hiLDL$$
$$\log\left[\Pr\left(deadin5=1\,|\,loLDL\right)\right]=\alpha_0+\alpha_1\times loLDL \qquad \log\left[\Pr\left(alivefor5=1\,|\,loLDL\right)\right]=\delta_0+\delta_1\times loLDL$$

Now, using the definition of our transformed variables, we note

$$\Pr\left(deadin5=1\,|\,hiLDL\right)=\beta_0+\beta_1\times hiLDL=\beta_0+\beta_1\times\left(1-loLDL\right)=\beta_0+\beta_1+\left(-\beta_1\right)\times loLDL$$
$$\Rightarrow \quad \beta_0+\beta_1=\alpha_0 \quad \beta_1=-\alpha_1$$
$$\Pr\left(alivefor5=1\,|\,hiLDL\right)=\gamma_0+\gamma_1\times hiLDL=\gamma_0+\gamma_1\times\left(1-loLDL\right)=\gamma_0+\gamma_1+\left(-\gamma_1\right)\times loLDL$$
$$\Rightarrow \quad \gamma_0+\gamma_1=\delta_0 \quad \gamma_1=-\delta_1$$

However,

$$\log\left[\Pr\left(deadin5=1\,|\,hiLDL\right)\right]=\beta_0+\beta_1\times hiLDL=\log\left[1-\Pr\left(alivefor5=1\,|\,hiLDL\right)\right]$$
$$=\log\left[1-\exp\left(\gamma_0-\gamma_1\times hiLDL\right)\right]$$

is not easily solved to obtain general correspondences between the regression parameters for the two models. If we are in a setting of saturated models with two binary predictors, then I can derive some rather complicated relationships: correspondences among the intercepts and slopes are found as:

$$\exp\left(\beta_0+\beta_1\right)=1-\exp\left(\gamma_0+\gamma_1\right) \qquad \exp\left(\beta_0\right)=1-\exp\left(\gamma_0\right)$$

In a fit of redundancy, I have included Poisson regression output from all of these models in the Appendix to this homework.

f. 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?

Ans: This alternative model is interchanging the role of response variable and predictor of interest in Poisson regression models. In this case, the regression model is fitting two groups (those dying

within 5 years and those surviving for at least 5 years) using two regression parameters (intercept and slope), so it is a saturated model.

Owing to the fact that we are using Poisson regression to model ratios of proportions, the interpretation of the intercept and slope parameters will differ markedly between the model used in parts a-c and this "reversed" model. Hence, the intercepts and slopes from the two models are not estimating the same quantities, so the fitted values within specific groups (as estimated in parts b and c) are not directly comparable across the models.

Nonetheless, I will not find it too surprising if any statistical significance of the association should be approximately the same for the two models, though deriving exact correspondences is difficult.

4. 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.)

    a. Evaluate associations between 5 year mortality and LDL using risk difference (RD: difference in probabilities).

Ans:

Methods: **The probabilities of subjects dying within 5 years of study enrollment were compared between subjects who differed in serum LDL using a linear regression model. Statistical inference on the difference in probabilities of death as a function of serum LDL modeled as a linear continuous variable was based on the Wald statistic computed from the regression slope parameter and its standard error as estimated using the Huber-White sandwich estimator. Two-sided p value and 95% confidence interval were computed using the approximate normal distribution for linear regression parameter estimates.**

Results: **Data was available on 725 subjects having mean serum LDL of 126 mg/dL (SD 33.6 mg/dL; range 11 – 247 mg/dL). Within 5 years of study accrual, 119 (16.4%) of the patients were observed to die, with 608 patients (83.6%) surviving at least 5 years. From a linear regression analysis, we estimate that the probability of death within 5 years is an absolute 1.03% lower for each 10 mg/dL higher serum LDL level at baseline. Based on a 95% confidence interval, this observed difference of proportions suggesting higher survivor probabilities for groups of patients with higher LDL levels would not be judged unusual if the true difference in probabilities were anywhere from 0.185% to 1.88% higher survival probability in a group having baseline serum LDL 10 mg/dL higher than another group. A two-sided p value of 0.017 suggests that we can with high confidence reject the null hypothesis that the probability of 5 year mortality is not associated with serum LDL levels in favor of a tendency for lower mortality with higher serum LDL levels.**

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

Ans:

Methods: **The probabilities of subjects dying within 5 years of study enrollment were compared between subjects who differed in serum LDL using a Poisson regression model. Statistical inference on the ratio of probabilities of death as a function of serum LDL modeled as a linear continuous variable was based on the Wald statistic computed from the regression slope parameter and its standard error as estimated using the Huber-White sandwich estimator. Two-sided p value and 95% confidence interval were computed using the approximate normal distribution for Poisson regression parameter estimates.**

Results: **Data was available on 725 subjects having mean serum LDL of 126 mg/dL (SD 33.6 mg/dL; range 11 – 247 mg/dL). Within 5 years of study accrual, 119 (16.4%) of the patients**

were observed to die, with 608 patients (83.6%) surviving at least 5 years. From a Poisson regression analysis, we estimate that the probability of death within 5 years is a relative 6.26% lower for each 10 mg/dL higher serum LDL level at baseline (mortality ratio 0.937). Based on a 95% confidence interval, this observed rate ratio suggesting higher survivor probabilities for groups of patients with higher LDL levels would not be judged unusual if the true ratio of survival probabilities were anywhere from 1.11% to 11.1% higher survival probability in a group having baseline serum LDL 10 mg/dL higher than another group (95% CI for mortality ratio 0.889 to 0.989). A two-sided p value of 0.018 suggests that we can with high confidence reject the null hypothesis that the probability of 5 year mortality is not associated with serum LDL levels in favor of a tendency for lower mortality with higher serum LDL levels.

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

Ans:

Methods: The odds of subjects dying within 5 years of study enrollment were compared between subjects who differed in serum LDL using a logistic regression model. Statistical inference on the ratio of odds of death as a function of serum LDL modeled as a linear continuous variable was based on the Wald statistic computed from the regression slope parameter and its standard error as estimated using maximum likelihood. Two-sided p value and 95% confidence interval were computed using the approximate normal distribution for logistic regression parameter estimates.

Results: Data was available on 725 subjects having mean serum LDL of 126 mg/dL (SD 33.6 mg/dL; range 11 – 247 mg/dL). Within 5 years of study accrual, 119 (16.4%) of the patients were observed to die, with 608 patients (83.6%) surviving at least 5 years. From a logistic regression analysis, we estimate that the odds of death within 5 years is a relative 7.48% lower for each 10 mg/dL higher serum LDL level at baseline (mortality odds ratio 0.925). Based on a 95% confidence interval, this observed odds ratio suggesting higher survivor probabilities for groups of patients with higher LDL levels would not be judged unusual if the true ratio of survival odds were anywhere from 1.70% to 12.9% higher odds of survival in a group having baseline serum LDL 10 mg/dL higher than another group (95% CI for mortality odds ratio 0.871 to 0.983). A two-sided p value of 0.012 suggests that we can with high confidence reject the null hypothesis that the odds of 5 year mortality is not associated with serum LDL levels in favor of a tendency for lower mortality with higher serum LDL levels.

d. 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*.?

Ans: *A priori* the analyses in problems 1-3 of this homework suffered from the imprecision that typically results from dichotomization of a variable—in this case LDL. For reasons of statistical precision, in homework 2 I preferred an analysis based on the geometric mean of LDL across groups defined by vital status at 5 years. None of the analyses in problems 1-3 of this homework would have been more attractive. However, any of the analyses in problem 4 of this homework treat LDL continuously, and thus do not suffer as much from the loss of precision. Furthermore, the analyses in problem 4 are conditioning on the variable we might think of as a "cause" (serum LDL) and consider the distribution of the putative "effect" (mortality within 4 years). This seems scientifically more pleasing. When choosing among the three regressions (RD, RR, or OR), the advantage of the RD regression is that it is expressed as a difference in absolute survival probabilities, rather than as a relative difference. If all that is reported is a relative difference, we

have difficulty assessing the public health impact. Hence, I might prefer the RD regression for the unadjusted analyses. I note that I would probably have preferred using a logarithmic transformation of LDL (see below for example report). I note that if I needed to adjust for covariates, I might instead consider logistic regression, depending on the strength of association anticipated in the confounders / precision variables I would be using.

An analysis based on a logarithmically transformed LDL:

> Methods: The probabilities of subjects dying within 5 years of study enrollment were compared between subjects who differed in serum LDL using a linear regression model. Statistical inference on the difference in probabilities of death as a function of serum LDL modeled as a logarithmically transformed continuous variable was based on the Wald statistic computed from the regression slope parameter and its standard error as estimated using the Huber-White sandwich estimator. Two-sided p value and 95% confidence interval were computed using the approximate normal distribution for Poisson regression parameter estimates.
>
> Results: Data was available on 725 subjects having mean serum LDL of 126 mg/dL (SD 33.6 mg/dL; range 11 – 247 mg/dL). Within 5 years of study accrual, 119 (16.4%) of the patients were observed to die, with 608 patients (83.6%) surviving at least 5 years. From a linear regression analysis, we estimate that the probability of death within 5 years is an absolute 1.41% lower for each 10% higher serum LDL level at baseline. Based on a 95% confidence interval, this observed difference of proportions suggesting higher survivor probabilities for groups of patients with higher LDL levels would not be judged unusual if the true difference in probabilities were anywhere from 0.439% to 2.38% higher survival probability in a group having baseline serum LDL 10% higher than another group. A two-sided p value of 0.004 suggests that we can with high confidence reject the null hypothesis that the probability of 5 year mortality is not associated with serum LDL levels in favor of a tendency for lower mortality with higher serum LDL levels.

PREPARING THE DATA

```
. quietly:                                                                    ///
> infile ptid mridate age male race weight height packyrs yrsquit alcoh physact   ///
>        chf chd stroke diabetes genhlth ldl alb crt plt sbp aai fev dsst atrophy ///
>        whgrd numinf volinf obstime death                                    ///
>        using http://www.emersonstatistics.com/datasets/mri.txt

.
. g deadin5 = 0

. replace deadin5 = 1 if obstime <= 5 * 365.25
(121 real changes made)

. g alivefor5 = 1 - deadin5

. recode ldl 160/max=1 min/160=0, gen(hiLDL)
(725 differences between ldl and hiLDL)

. g loLDL = 1 - hiLDL
(11 missing values generated)
```

PROBLEM #1

```
. logit deadin5 hiLDL

Logistic regression                               Number of obs   =        725
                                                  LR chi2(1)      =       1.07
                                                  Prob > chi2     =     0.3019
Log likelihood = -323.15665                       Pseudo R2       =     0.0016

------------------------------------------------------------------------------
     deadin5 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hiLDL |  -.3072267     .30603    -1.00   0.315    -.9070345    .2925812
```

```
       _cons |  -1.586315   .1071127   -14.81   0.000    -1.796252   -1.376378
------------------------------------------------------------------------------
```

. logit deadin5 hiLDL, robust

```
Logistic regression                                     Number of obs   =        725
                                                        Wald chi2(1)    =       1.01
                                                        Prob > chi2     =     0.3158
Log pseudolikelihood = -323.15665                       Pseudo R2       =     0.0016

------------------------------------------------------------------------------
             |               Robust
     deadin5 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hiLDL |  -.3072267   .3062413    -1.00   0.316    -.9074486    .2929953
       _cons |  -1.586315   .1071866   -14.80   0.000    -1.796397   -1.376234
------------------------------------------------------------------------------
```

. logistic deadin5 hiLDL

```
Logistic regression                                     Number of obs   =        725
                                                        LR chi2(1)      =       1.07
                                                        Prob > chi2     =     0.3019
Log likelihood = -323.15665                             Pseudo R2       =     0.0016

------------------------------------------------------------------------------
     deadin5 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hiLDL |   .7354839   .2250802    -1.00   0.315     .4037197    1.339882
------------------------------------------------------------------------------
```

. logit deadin5 loLDL

```
Logistic regression                                     Number of obs   =        725
                                                        LR chi2(1)      =       1.07
                                                        Prob > chi2     =     0.3019
```

```
Log likelihood = -323.15665                     Pseudo R2       =     0.0016

------------------------------------------------------------------------------
     deadin5 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       loLDL |   .3072267     .30603     1.00   0.315    -.2925812    .9070345
       _cons |  -1.893542   .2866727    -6.61   0.000     -2.45541   -1.331674
------------------------------------------------------------------------------
```

```
. logit alivefor5 hiLDL

Logistic regression                             Number of obs   =        725
                                                LR chi2(1)      =       1.07
                                                Prob > chi2     =     0.3019
Log likelihood = -323.15665                     Pseudo R2       =     0.0016

------------------------------------------------------------------------------
   alivefor5 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hiLDL |   .3072267     .30603     1.00   0.315    -.2925812    .9070345
       _cons |   1.586315   .1071127    14.81   0.000     1.376378    1.796252
------------------------------------------------------------------------------
```

```
. logit alivefor5 loLDL

Logistic regression                             Number of obs   =        725
                                                LR chi2(1)      =       1.07
                                                Prob > chi2     =     0.3019
Log likelihood = -323.15665                     Pseudo R2       =     0.0016

------------------------------------------------------------------------------
   alivefor5 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       loLDL |  -.3072267     .30603    -1.00   0.315    -.9070345    .2925812
       _cons |   1.893542   .2866727     6.61   0.000     1.331674     2.45541
------------------------------------------------------------------------------
```

```
. logit hiLDL deadin5

Logistic regression                                     Number of obs   =        725
                                                        LR chi2(1)      =       1.07
                                                        Prob > chi2     =     0.3019
Log likelihood = -302.87907                             Pseudo R2       =     0.0018

------------------------------------------------------------------------------
       hiLDL |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     deadin5 |  -.3072267     .30603    -1.00   0.315    -.9070345    .2925812
       _cons |  -1.707676   .1127032   -15.15   0.000    -1.928571   -1.486782
------------------------------------------------------------------------------
```

PROBLEM #2

```
. regress deadin5 hiLDL, robust

Linear regression                                       Number of obs =     725
                                                        F(  1,   723) =    1.18
                                                        Prob > F      =  0.2780
                                                        R-squared     =  0.0014
                                                        Root MSE      =  .37065

------------------------------------------------------------------------------
             |               Robust
     deadin5 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hiLDL |  -.0390618   .0359806    -1.09   0.278    -.1097008    .0315772
       _cons |   .1699029   .0151276    11.23   0.000     .1402036    .1996022
------------------------------------------------------------------------------

. regress deadin5 hiLDL

      Source |       SS       df       MS              Number of obs =     725
-------------+------------------------------           F(  1,   723) =    1.01
       Model |  .139167733     1  .139167733           Prob > F      =  0.3145
```

```
    Residual |  99.3284185   723  .137383705           R-squared     =  0.0014
-------------+------------------------------           Adj R-squared =  0.0000
       Total |  99.4675862   724  .137386169           Root MSE      =  .37065

------------------------------------------------------------------------------
     deadin5 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hiLDL |  -.0390618   .0388106    -1.01   0.315    -.1152567    .0371332
       _cons |   .1699029   .0149099    11.40   0.000     .1406311    .1991747
------------------------------------------------------------------------------
```

. regress deadin5 loLDL, robust

```
Linear regression                                       Number of obs =     725
                                                        F(  1,   723) =    1.18
                                                        Prob > F      =  0.2780
                                                        R-squared     =  0.0014
                                                        Root MSE      =  .37065

------------------------------------------------------------------------------
             |               Robust
     deadin5 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       loLDL |   .0390618   .0359806     1.09   0.278    -.0315772    .1097008
       _cons |   .1308411    .032646     4.01   0.000     .0667489    .1949334
------------------------------------------------------------------------------
```

. regress alivefor5 hiLDL, robust

```
Linear regression                                       Number of obs =     725
                                                        F(  1,   723) =    1.18
                                                        Prob > F      =  0.2780
                                                        R-squared     =  0.0014
                                                        Root MSE      =  .37065

------------------------------------------------------------------------------
             |               Robust
```

```
   alivefor5 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hiLDL |   .0390618   .0359806     1.09   0.278    -.0315772    .1097008
       _cons |   .8300971   .0151276    54.87   0.000     .8003978    .8597964
------------------------------------------------------------------------------
```

```
. regress alivefor5 loLDL, robust

Linear regression                                      Number of obs =     725
                                                       F(  1,   723) =    1.18
                                                       Prob > F      =  0.2780
                                                       R-squared     =  0.0014
                                                       Root MSE      =  .37065

------------------------------------------------------------------------------
             |               Robust
   alivefor5 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       loLDL |  -.0390618   .0359806    -1.09   0.278    -.1097008    .0315772
       _cons |   .8691589    .032646    26.62   0.000     .8050666    .9332511
------------------------------------------------------------------------------
```

```
.
. regress hiLDL deadin5, robust

Linear regression                                      Number of obs =     725
                                                       F(  1,   723) =    1.18
                                                       Prob > F      =  0.2783
                                                       R-squared     =  0.0014
                                                       Root MSE      =  .35493

------------------------------------------------------------------------------
             |               Robust
       hiLDL |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     deadin5 |  -.0358183   .0330107    -1.09   0.278    -.1006266      .02899
       _cons |   .1534653   .0146619    10.47   0.000     .1246803    .1822504
```

PROBLEM #3

```
. poisson deadin5 hiLDL, robust

Poisson regression                                Number of obs   =        725
                                                  Wald chi2(1)    =       0.97
                                                  Prob > chi2     =     0.3237
Log pseudolikelihood = -333.58825                 Pseudo R2       =     0.0013

------------------------------------------------------------------------------
             |               Robust
     deadin5 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hiLDL |  -.2612434    .264736    -0.99   0.324    -.7801163    .2576296
       _cons |  -1.772528   .0889753   -19.92   0.000    -1.946917    -1.59814
------------------------------------------------------------------------------

. poisson deadin5 hiLDL

Poisson regression                                Number of obs   =        725
                                                  LR chi2(1)      =       0.90
                                                  Prob > chi2     =     0.3425
Log likelihood = -333.58825                       Pseudo R2       =     0.0013

------------------------------------------------------------------------------
     deadin5 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hiLDL |  -.2612434   .2845213    -0.92   0.359    -.8188949    .2964082
       _cons |  -1.772528     .09759   -18.16   0.000    -1.963801   -1.581255
------------------------------------------------------------------------------

. poisson deadin5 loLDL, robust

Poisson regression                                Number of obs   =        725
                                                  Wald chi2(1)    =       0.97
```

```
                                                Prob > chi2     =     0.3237
Log pseudolikelihood = -333.58825               Pseudo R2       =     0.0013

------------------------------------------------------------------------------
             |               Robust
     deadin5 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       loLDL |   .2612434    .264736     0.99   0.324    -.2576296    .7801163
       _cons |  -2.033771   .2493362    -8.16   0.000    -2.522461   -1.545082
------------------------------------------------------------------------------
```

. poisson alivefor5 hiLDL, robust

```
Poisson regression                              Number of obs   =        725
                                                Wald chi2(1)    =       1.21
                                                Prob > chi2     =     0.2704
Log pseudolikelihood =  -714.5684               Pseudo R2       =     0.0001

------------------------------------------------------------------------------
             |               Robust
   alivefor5 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hiLDL |   .0459833   .0417192     1.10   0.270    -.0357848    .1277513
       _cons |  -.1862126   .0182113   -10.23   0.000    -.2219061   -.1505191
------------------------------------------------------------------------------
```

. poisson alivefor5 loLDL, robust

```
Poisson regression                              Number of obs   =        725
                                                Wald chi2(1)    =       1.21
                                                Prob > chi2     =     0.2704
Log pseudolikelihood =  -714.5684               Pseudo R2       =     0.0001

------------------------------------------------------------------------------
             |               Robust
   alivefor5 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
```

```
       loLDL |  -.0459832   .0417192    -1.10   0.270    -.1277513    .0357848
       _cons |  -.1402294   .0375345    -3.74   0.000    -.2137956   -.0666631
------------------------------------------------------------------------------
```

. poisson hiLDL deadin5, robust

```
Poisson regression                                Number of obs   =        725
                                                  Wald chi2(1)    =       0.98
                                                  Prob > chi2     =     0.3227
Log pseudolikelihood = -311.26901                 Pseudo R2       =     0.0015

------------------------------------------------------------------------------
             |               Robust
      hiLDL |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     deadin5 |  -.2657857   .2687515    -0.99   0.323    -.7925289    .2609576
       _cons |   -1.87428    .095473   -19.63   0.000    -2.061404   -1.687157
------------------------------------------------------------------------------
```

PROBLEM #4

Note that I first provide statistics for 1 unit differences in LDL. But to make the differences more meaningful, I consider rescaling LDL to be 10 mg/dL.

. regress deadin5 ldl, robust

```
Linear regression                                 Number of obs =     725
                                                  F(  1,   723) =    5.71
                                                  Prob > F      =  0.0171
                                                  R-squared     =  0.0088
                                                  Root MSE      =  .36928

------------------------------------------------------------------------------
             |               Robust
     deadin5 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ldl |  -.0010343   .0004328    -2.39   0.017    -.0018839   -.0001847
```

```
       _cons |   .2942598   .0580314     5.07   0.000     .1803296      .40819
------------------------------------------------------------------------------
```

```
. poisson deadin5 ldl, robust irr

Poisson regression                                Number of obs   =        725
                                                  Wald chi2(1)    =       5.63
                                                  Prob > chi2     =     0.0177
Log pseudolikelihood = -331.32723                 Pseudo R2       =     0.0081

------------------------------------------------------------------------------
             |               Robust
     deadin5 |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ldl |   .9935516   .0027101    -2.37   0.018     .9882541    .9988775
------------------------------------------------------------------------------
```

```
. logistic deadin5 ldl

Logistic regression                               Number of obs   =        725
                                                  LR chi2(1)      =       6.50
                                                  Prob > chi2     =     0.0108
Log likelihood = -320.44119                       Pseudo R2       =     0.0100

------------------------------------------------------------------------------
     deadin5 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ldl |   .9922557   .0030687    -2.51   0.012     .9862595    .9982885
------------------------------------------------------------------------------
```

```
. g ldl10 = ldl/10
(11 missing values generated)

. regress deadin5 ldl10, robust

Linear regression                                 Number of obs =     725
                                                  F(  1,   723) =    5.71
```

```
                                                Prob > F      =  0.0171
                                                R-squared     =  0.0088
                                                Root MSE      =  .36928

------------------------------------------------------------------------------
             |               Robust
     deadin5 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       ldl10 |  -.0103433   .0043276    -2.39   0.017    -.0188394   -.0018472
       _cons |   .2942598   .0580314     5.07   0.000     .1803296      .40819
------------------------------------------------------------------------------
```

. poisson deadin5 ldl10, robust irr

```
Iteration 0:   log pseudolikelihood = -331.32723
Iteration 1:   log pseudolikelihood = -331.32723

Poisson regression                              Number of obs   =        725
                                                Wald chi2(1)    =       5.63
                                                Prob > chi2     =     0.0177
Log pseudolikelihood = -331.32723               Pseudo R2       =     0.0081

------------------------------------------------------------------------------
             |               Robust
     deadin5 |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       ldl10 |   .9373552   .0255679    -2.37   0.018     .8885589    .9888311
------------------------------------------------------------------------------
```

. logistic deadin5 ldl10

```
Logistic regression                             Number of obs   =        725
                                                LR chi2(1)      =       6.50
                                                Prob > chi2     =     0.0108
Log likelihood = -320.44119                     Pseudo R2       =     0.0100

------------------------------------------------------------------------------
```

```
     deadin5 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+------------------------------------------------------------------
       ldl10 |   .9252012   .0286129    -2.51   0.012     .8707867     .983016
------------------------------------------------------------------------------
```

An auxiliary analysis based on logarithmically transformed LDL. I use base 1.1 logarithm to be able to talk about groups that differ by 10% in their LDL levels.

```
. g logldl = log(ldl) / log(1.1)
(11 missing values generated)

. regress deadin5 logldl, robust

Linear regression                               Number of obs =     725
                                                F(  1,   723) =    8.14
                                                Prob > F      =  0.0045
                                                R-squared     =  0.0136
                                                Root MSE      =  .36838

------------------------------------------------------------------------------
             |               Robust
     deadin5 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      logldl |  -.0140802   .0049357    -2.85   0.004    -.0237702   -.0043902
       _cons |   .8725977   .2506017     3.48   0.001     .3806037    1.364592
------------------------------------------------------------------------------
```