Biost 517: Applied Biostatistics I
Emerson, Fall 2012

Homework #8 Key
November 30, 2012

Written problems: To be handed in at the beginning of class on Friday, December 7, 2012.

The written problems all refer to the data on inflammatory markers of cardiovascular disease (inflamm.txt stored in the project data file on the class web pages).

In this key, I embed the relevant Stata output inline in order that you can see exactly where I get the numbers. But I did not want you to include such output on your homework.

Questions for Biost 514 and Biost 517:

1. Perform an analysis to compare the mean cholesterol values across groups defined by age, while allowing that each age might have a distinct average cholesterol.

If my model is to allow each age to have a distinct average cholesterol, I need to use a linear regression model with the continuous (not dichotomized) age variable. Parts a, b, and c ask about the regression estimates. To that end, I can perform either classical linear regression analysis or linear regression analysis with the robust standard errors, because those two methods yield the exact same parameter estimates. They only differ in the standard errors, p values, and CI. Parts d and e ask for inference based on the assumption of homoscedasticity, so I need to use classical linear regression:

```
. regress cholest age

      Source |       SS       df       MS              Number of obs =    4953
-------------+------------------------------           F(  1,  4951) =   31.34
       Model |  48076.8381     1  48076.8381           Prob > F      =  0.0000
    Residual |  7595623.96  4951  1534.15956           R-squared     =  0.0063
-------------+------------------------------           Adj R-squared =  0.0061
       Total |   7643700.8  4952  1543.55832           Root MSE      =  39.168

------------------------------------------------------------------------------
     cholest |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.5589903   .0998554    -5.60   0.000    -.7547512   -.3632295
       _cons |   252.3886   7.291604    34.61   0.000     238.0938    266.6834
------------------------------------------------------------------------------
```

a. Provide an interpretation for the estimated intercept. What use would you make of this estimate in this scientific setting?

Ans: The average cholesterol among newborns (age 0) is estimated to be 252 mg/dl. This age group is way outside the range of our data, and we should be very reluctant to try to extrapolate to that age group. I would make no scientific use of this estimate.

b. Provide an interpretation for the estimated slope. What use would you make of this estimate in this scientific setting?

Ans: The mean cholesterol is estimated to differ between two age groups by 0.599 mg/dl (on average) for each year difference in age, with the older group tending toward lower average cholesterol. This is a measure of association between cholesterol and age, interpretable (at least) as a first order trend in the means.

c. Using the estimated regression model, what is the best estimate of the mean cholesterol for 70 year olds? What is the best estimate of the cholesterol for 75 year olds.

Ans: The mean cholesterol in 70 year olds is estimated as 252.3886 – 0.5589903 x 70 = 213 mg/dl. The mean cholesterol in 75 year olds is estimated as 252.3886 – 0.5589903 x 75 = 210 mg/dl. *(Note that I use full precision of my estimates for the calculations, and then only report 3 significant digits in the answer. You might not get the same answer if you round off the numbers in the intermediate calculations.)*

d. Provide full inference (i.e., provide point estimates, confidence intervals, and p values where possible, along with a statement of your scientific/statistical conclusions) when presuming that the variance of cholesterol is equal across all age groups.

Ans: From a linear regression analysis, we estimate that mean cholesterol differs between two age groups by 0.559 mg/dl (on average) for each year difference in age, with the older group tending toward lower average cholesterol. This result is significantly different from 0 (P < 0.0005), with a 95% CI suggesting that such observed results would not be unusual if the true difference in mean cholesterol between age groups were anywhere between 0.363 mg/dl and 0.755 mg/dl for each year difference in age, with the older group tending toward lower average cholesterol. We thus reject the null hypothesis that mean cholesterol does not differ across age groups, in favor of a hypothesis that mean cholesterol tends to be lower for older ages.

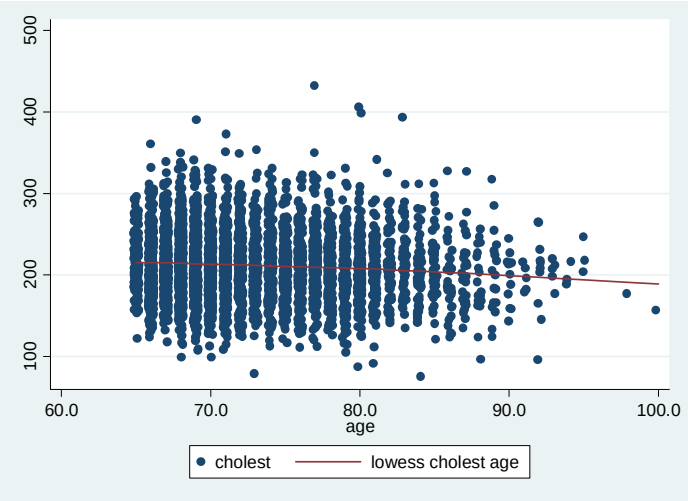
e. Using the analysis in part d, what is your best estimate of the standard deviation of cholesterol in each age group?

Ans: From the root mean squared error, we estimate a within group standard deviation of 39.2 mg/dl.

f. Provide descriptive statistics that would assess (in a *post hoc* fashion) whether you believe that the estimates provided in parts c, d, and e are reliable. Explain the issues that you must consider.

I need to assess the assumption of homoscedasticity necessary for the P values and CI in part d to be correct and the assumption of linearity necessary for the estimation of mean cholesterol in individual age groups. A jittered scatterplot with lowess smooth seems about as good a way as any:

```
. twoway (scatter cholest age, jitter(1)) (lowess cholest age)
```



<u>Ans:</u> A lowess curve superimposed on a scatterplot appears relatively linear, thus suggesting that the estimates provided in part c might be reasonable. The scatterplot shows similar ranges of cholesterol measurements within age groups having comparable sample sizes gives no strong evidence of heteroscedasticity, thus suggesting that the inference provided in part d and the average SD estimated in part e might be reasonable. *(There are 134 subjects who are age 65, and 142 subjects who are age 86 or higher. The spread of the data for the 65 year olds looks to my eye similar to the spread of the data among the subjects 86-100, and the trend by age is not so drastic that it would contribute markedly to the spread of the data in the 86-100 year olds. I further note that there were 314 – 466 subjects in the next few age groups, so I would expect more spread of the data in those age groups. Given the large sample sizes in each age group, we could also look at stratified SD estimates in order to try to judge systematic trends in heteroscedasticity.*

To the extent that you thought that the spread of the data in the highest age groups was less than that in the younger age groups, then we can think about whether we would expect the classical linear regression inference to be conservative or anti-conservative for the mean. Recall that inference based on homoscedasticity is anti-conservative (P values too low, CI too narrow) if the groups with the smaller sample size have the larger variance. If you thought there was less variance in the older groups, the fact that there was also smaller sample sizes in those groups means that the classical linear regression inference would tend to be conservative.

While both "pre-testing" and "post-testing" of assumptions are problematic from the standpoint of inflating your type I error, I do note that in Biost 515/518 you will cover methods that allow you to perform hypothesis tests for linearity and homoscedasticity. For instance, we could perform a regression on both age and age squared. If the term for age squared were significantly different from 0, that would be evidence suggesting nonlinearity. If that term were not significant, that would not prove linearity for two main reasons. First, there are many other forms of nonlinearity beyond parabolas, so there may be a second order term that averages 0, but higher order terms do exist. Second, we have to worry that we just did not have enough precision to detect a nonzero second order trend: Lack of evidence of an effect is not evidence of a lack of effect. Nonetheless, fitting such a regression model (with robust standard errors to allow for the possibility of heteroscedasticity)

```
. g agesqr= age^2
. regress cholest age agesqr, robust
```

```
Linear regression                               Number of obs =    4953
                                                F(  2,  4950) =   15.94
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0065
                                                Root MSE      =  39.168
------------------------------------------------------------------------------
             |               Robust
     cholest |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    1.654227   1.949348     0.85   0.396    -2.16736    5.475815
      agesqr |   -.0146876   .0129304    -1.14   0.256    -.0400369    .0106617
       _cons |    169.5639   73.15376     2.32   0.020     26.15014    312.9777
------------------------------------------------------------------------------
```

finds no statistically significant second order trend as evidenced by a P value of 0.256 for the agesqr term. Looking ahead to next quarter, I will note that the interpretation of the age parameter has changed, because we have the agesqr term in the model as well. We thus do not test for an association between cholesterol and age in this model by looking at either parameter by itself. Instead we would need to test them simultaneously. You will learn how to do this in "post-estimation" commands next quarter. However, because the only variables in this model are both related to age, I can use the F test (statistic F= 15.94 with 2, 4950 degrees of freedom) that tests the entire model. From that test, we see that age and cholesterol are significantly associated (P < 0.0001), but looking at the individual parameter estimates shows that we do not have sufficient precision to be able to decide whether the linear term by itself or the squared term by itself might be the best way to model it.

To test for heteroscedasticity, we can examine the variability of the estimated residuals. Recall that residuals are the difference between the observed value and the fitted values. So we could create the fitted values, calculate the residual values, and then examine how the variance of the residuals varies with age. Also recall that the residuals have mean zero, so the expected value of the squared residuals should be the variance of the residuals. The following code can look for a linear trend in the variance of the residuals (and it is the linear trend in heteroscedasticity that would harm the inference in classical linear regression):

```
. g fit= 252.3886 - 0.5589903 * age
. g resid = cholest - fit
. g residsqr = resid^2
. regress residsqr age, robust

Linear regression                               Number of obs =    4953
                                                F(  1,  4951) =    2.40
                                                Prob > F      =  0.1217
                                                R-squared     =  0.0005
                                                Root MSE      =  2551.5
------------------------------------------------------------------------------
             |               Robust
    residsqr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    10.72631     6.9302     1.55   0.122    -2.859952    24.31258
       _cons |    752.5723   497.1733     1.51   0.130    -222.1078    1727.252
------------------------------------------------------------------------------
```

The p value of 0.122 for the age effect suggest that we do not have sufficient evidence to establish with high confidence that there is a linear trend in heteroscedasticity across the age groups. Again, lack of evidence for heteroscedasticity can not be interpreted as evidence that there is no heteroscedasticity. Were we to regard that the positive slope is indicative of a true trend in heteroscedasticity, then that tendency toward higher variance in the smaller age groups would suggest that classical linear regression based inference is anti-conservative.

Lastly, I note that the Stata "post-estimation" command "predict" could have been used to obtain the estimated residuals from the most recently performed regression. The command

```
. predict rsd, re
```

would have created a variable named rsd that contains the residuals. There are many other such "post-estimation" commands that will be used next quarter.)

g. Provide full inference when allowing that the variance of cholesterol measurements might be unequal across some age groups.

If my model is to allow each age to have a distinct average cholesterol and allow for the possibility of heteroscedasticity, I perform linear regression analysis with the robust standard errors. As noted above, the parameter estimates will not differ from what was obtained with classical linear regression. The only difference will be in the standard errors, p values, and CI.

```
. regress cholest age

Linear regression                               Number of obs =    4953
                                                F(  1,  4951) =   30.73
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0063
                                                Root MSE      =  39.168
------------------------------------------------------------------------------
             |               Robust
     cholest |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.5589903   .1008394    -5.54   0.000    -.7566803   -.3613004
       _cons |   252.3886   7.341587    34.38   0.000     237.9958    266.7813
------------------------------------------------------------------------------
```

Ans: From a linear regression analysis using standard errors calculated with the Huber-White sandwich estimator, we estimate that mean cholesterol differs between two age groups by 0.559 mg/dl (on average) for each year difference in age, with the older group tending toward lower average cholesterol. This result is significantly different from 0 (P < 0.0005), with a 95% CI suggesting that such observed results would not be unusual if the true difference in mean cholesterol between age groups were anywhere between 0.361 mg/dl and 0.757 mg/dl for each year difference in age, with the older group tending toward lower average cholesterol. We thus reject the null hypothesis that mean cholesterol does not differ across age groups, in favor of a hypothesis that mean cholesterol tends to be lower for older ages. *(The negligibly higher SE and wider CI agree with the fact that there was not strong evidence of heteroscedasticity, but any trend that there was showed higher variance in the groups with smaller sample sizes.)*

h. Of the analyses considered in parts d and g, which would you prefer *a priori*.

Ans: As I would be primarily interested in how the mean cholesterol might vary across ages, I would want to allow for the possibility that there might be heteroscedasticity: If I am unsure of how the mean varies across ages, I certainly don't know how the variance might differ across ages. If I cannot know that the groups are homoscedastic, then all that classical linear regression can tell me is that the strong null hypothesis of exact equality of all aspects of the distribution is not true. I would not be able to state with high confidence that I knew that the mean differed across groups.

i. Using the analysis in part g, provide an estimate and confidence interval for the difference in mean cholesterol measurements that might be expected between two groups that differ in age by 5 years.

Ans: **From a linear regression analysis using standard errors calculated with the Huber-White sandwich estimator, we estimate that mean cholesterol differs between two age groups by 2.79 mg/dl (on average) for each 5 year difference in age, with the older group tending toward lower average cholesterol. This result is significantly different from 0 (P < 0.0005), with a 95% CI suggesting that such observed results would not be unusual if the true difference in mean cholesterol between age groups were anywhere between 1.81 mg/dl and 3.78 mg/dl for each 5 year difference in age, with the older group tending toward lower average cholesterol. We thus reject the null hypothesis that mean cholesterol does not differ across age groups, in favor of a hypothesis that mean cholesterol tends to be lower for older ages.** *(All I had to do was multiply the slope estimates and CI by 5.)*

2. Perform an analysis to assess the correlation between age and cholesterol. What is the estimated correlation? Is this estimate significantly different from 0? How does the P value from this analysis compare to the results of your analysis in problem 1?

I use the command pwcorr to get the statistical significance..

```
. pwcorr cholest age, sig

             |  cholest      age
     cholest |   1.0000
             |
         age |  -0.0793   1.0000
             |   0.0000
```

Ans: **We estimate a correlation of -0.0793, which is statistically significantly different from 0 (P < 0.0001). This estimate agrees exactly with the signed square root of the R squared reported in a simple linear regression of cholesterol on age or a simple linear regression of age on cholesterol. The P value will agree exactly with the statistical significance of the age parameter in a classical simple linear regression of cholesterol on age or with the statistical significance of the cholesterol parameter in a classical simple linear regression of age on cholesterol. It will not necessarily agree with the P value testing for association in a simple linear regression that used robust standard error estimates.** *(Note also that the correlation estimate will not have any correspondence with the R squared reported in a multiple regression model having more than one modeled predictor, nor will the P value testing the correlation match the p value for an adjusted regression analysis.)*

3. Perform an analysis to compare the geometric mean cholesterol values across groups defined by age, while allowing that each age might have a distinct geometric mean.

If my model is to allow each age to have a distinct geometric mean cholesterol, I need to use a linear regression model on log transformed cholesterol with the continuous (not dichotomized) age variable. The problem did not specify whether I should use classical linear regression or linear regression with robust standard error estimates. As I believe the latter is better, I use it.

```
. g logchol= log(cholest)
. regress logchol age, robust
Linear regression                               Number of obs =    4953
                                                F(  1,  4951) =   34.16
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0073
                                                Root MSE      =  .18805
             |               Robust
```

```
     logchol |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   -.002895   .0004953    -5.85   0.000     -.003866    -.001924
       _cons |   5.548433   .0359619   154.29   0.000     5.477932    5.618934
------------------------------------------------------------------------------

. display exp(5.548433)
256.83478

. display exp(-.002895), exp(-.003866), exp(-.001924)
.99710919 .99614146 .99807785

. display exp(-.002895*10), exp(-.003866*10), exp(-.001924*10)
.97146504 .96207776 .98094391
```

a. Provide an interpretation for the estimated intercept. What use would you make of this estimate in this scientific setting?

Ans: The geometric cholesterol among newborns (age 0) is estimated to be 257 mg/dl. This age group is way outside the range of our data, and we should be very reluctant to try to extrapolate to that age group. I would make no scientific use of this estimate. *(Note that I exponentiate the parameter estimates to obtain interpretations on the original measurement scale.)*

b. Provide an interpretation for the estimated slope. What use would you make of this estimate in this scientific setting?

Ans: The geometric mean cholesterol is estimated to differ between two age groups by 0.289% mg/dl (on average) lower geometric mean for each year difference in age, with the older group tending toward lower geometric mean cholesterol. This is a measure of association between cholesterol and age, interpretable (at least) as some sort of a first order trend in the geometric means. *(Again I exponentiate the parameter estimates to obtain interpretations on the original measurement scale.)*

c. Provide full inference to address the question of an association between age and cholesterol based on the geometric mean.

Ans: From a linear regression analysis using standard errors calculated with the Huber-White sandwich estimator, we estimate that geometric mean cholesterol is estimated to differ between two age groups by 0.289% mg/dl (on average) lower geometric mean for each year difference in age, with the older group tending toward lower geometric mean cholesterol. This result is significantly different from 0 (P < 0.0005), with a 95% CI suggesting that such observed results would not be unusual if the true difference in geometric mean cholesterol between age groups were anywhere between 0.192% lower and 0.386% lower for each year difference in age, with the older group tending toward lower average cholesterol. We thus reject the null hypothesis that geometric mean cholesterol does not differ across age groups, in favor of a hypothesis that geometric mean cholesterol tends to be lower for older ages.

(The very small effects estimated for 1 year differences would often lead to a decision to present estimates according to 10 year differences. I am always nervous when I report something like 0.3%, because casual readers sometimes do not recognize the difference between 0.3 = 30% and 0.003 = 0.3%. So I would merely multiply the slope parameter estimates and CI by 10 prior to exponentiation (see above), to obtain a description:

From a linear regression analysis using standard errors calculated with the Huber-White sandwich estimator, we estimate that geometric mean cholesterol is estimated to differ between two age groups by 2.85% mg/dl (on average) lower geometric mean for each 10 year difference in age, with the older group tending toward lower geometric mean cholesterol. This result is significantly different from 0 (P < 0.0005), with a 95% CI suggesting that such observed results would not be unusual if the true difference in geometric mean cholesterol between age groups were anywhere between 1.91% lower and 3.79% lower for each 10 year difference in age, with the older group tending toward lower average cholesterol. We thus reject the null hypothesis that geometric mean cholesterol does not differ across age groups, in favor of a hypothesis that geometric mean cholesterol tends to be lower for older ages..)

4. Analyze the data to assess whether there is an association between time to death and cholesterol level.

    a. Problems b-e below consider analyses comparing groups defined by whether the patient died within 4 years or not. Why is this a valid analysis in this dataset containing censored observations of time to death?

Even when the measurement of time to an event is subject to censoring, I can perform more standard analyses based on observations less than the earliest censoring time. So I examine the minimum censoring time in the sample.

```
. g yrtodth= ttodth / 365.25
. summ yrtodth if death==0

    Variable |       Obs        Mean    Std. Dev.       Min        Max
     yrtodth |      3879    7.128573    1.132354   4.052019   8.054757
```

Ans: The earliest censoring time was at 4.05 years. Hence analyses based on dichotomizing subjects with respect to 4 year survival or not are valid.

Ans: The geometric mean cholesterol is estimated to differ between two age groups by 0.289% mg/dl (on average) lower geometric mean for each year difference in age, with the older group tending toward lower geometric mean cholesterol. This is a measure of association between cholesterol and age, interpretable (at least) as some sort of a first order trend in the geometric means. *(Again I exponentiate the parameter estimates to obtain interpretations on the original measurement scale.)*

    b. Base your analysis on a comparison of mean cholesterol across groups defined by whether the patient died within 4 years or not.

The standard analysis to compare means of a continuous random variable across two groups is a t test. The problem did not specify whether I should presume equal variances for the two groups or allow for the possibility that the variances might differ across groups. As I believe the latter is better, I use it.

```
. g deadin4= yrtodth
. recode deadin4 0/4=1 4/max=0
. ttest cholest, by(deadin4) unequal

Two-sample t test with unequal variances
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
```

```
       0 |    4467     212.518    .5830693     38.9698    211.3749    213.6611
       1 |     486     204.072    1.876692    41.37245    200.3846    207.7595
combined |    4953    211.6893    .5582482    39.28814    210.5949    212.7837
    diff |           8.446005    1.965183                4.586298    12.30571
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   4.2978
Ho: diff = 0                     Satterthwaite's degrees of freedom =  582.562

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

Ans: **The mean cholesterol is estimated to be 204 mg/dl among subjects who die within 4 years of study entry and 213 mg/dl among subjects who survive at least 4 years. A comparison of the two groups thus estimates that the mean cholesterol is 8.45 mg/dl higher among subjects who survive 4 years relative to those who die within 4 years. This observed difference is statistically different from 0 (P < 0.0001), with a 95% confidence interval suggesting that the observed difference is what might be typically observed if the true difference between survivors and nonsurvivors was anywhere between 4.59 mg/dl and 12.3 mg/dl, with the survivors averaging higher cholesterol. We thus reject the null hypothesis of no association between survival time and cholesterol at study entry in favor of a trend toward higher mean cholesterol among subjects surviving the longer period of time.**

c. Perform a regression analysis of cholesterol (response variable) on an indicator of death within 4 years. How does this analysis compare with the analysis you performed in part b?

The problem did not specify whether I should use classical linear regression or linear regression with robust standard error estimates. As I believe the latter is better, I use it.

```
. regress cholest deadin4, robust
Linear regression                               Number of obs =    4953
                                                F(  1,  4951) =   18.50
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0041
                                                Root MSE      =  39.212
------------------------------------------------------------------------------
             |               Robust
     cholest |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     deadin4 |  -8.446005   1.963715    -4.30   0.000    -12.29576   -4.596252
       _cons |    212.518   .5831218   364.45   0.000     211.3748    213.6612
------------------------------------------------------------------------------
```

Ans: **The mean cholesterol is estimated to be 213 mg/dl among subjects who survive at least 4 years to be on average 8.45 mg/dl lower among subjects who die within 4 years. This observed difference is statistically different from 0 (P < 0.0001), with a 95% confidence interval suggesting that the observed difference is what might be typically observed if the true difference between survivors and nonsurvivors was anywhere between 4.60 mg/dl and 12.3 mg/dl, with the survivors averaging higher cholesterol. We thus reject the null hypothesis of no association between survival time and cholesterol at study entry in favor of a trend toward higher mean cholesterol among subjects surviving the longer period of time.**

This linear regression analysis with robust standard errors provides nearly the same inference as does the t test that allows for the possibility of unequal variances. The estimated means are exactly the same (and they agree with what would have been obtained using the t test that presumes equal variances and or classical linear regression), and the standard errors (1.963715 vs 1.965183) and CI are negligibly different. The primary

difference is in the use of n vs n-2 in estimating the standard errors and in determining the number of degrees of freedom.

(Had I used the t test that presumes equal variances in part b and classical linear regression in part c, the correspondences would have been exact. Note the highlighted output in the two analyses.

```
. ttest cholest, by(deadin4)

Two-sample t test with equal variances
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |    4467     212.518    .5830693    38.9698    211.3749    213.6611
       1 |     486     204.072    1.876692    41.37245    200.3846    207.7595
---------+--------------------------------------------------------------------
combined |    4953    211.6893    .5582482    39.28814    210.5949    212.7837
---------+--------------------------------------------------------------------
    diff |            8.446005    1.872938                4.774216    12.11779
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   4.5095
Ho: diff = 0                     degrees of freedom =     4951

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000

. regress cholest deadin4

      Source |       SS       df       MS              Number of obs =    4953
-------------+------------------------------           F(  1,  4951) =   20.34
       Model |   31267.022     1   31267.022           Prob > F      =  0.0000
    Residual |  7612433.78  4951  1537.55479           R-squared     =  0.0041
-------------+------------------------------           Adj R-squared =  0.0039
       Total |  7643700.8   4952  1543.55832           Root MSE      =  39.212
------------------------------------------------------------------------------
     cholest |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     deadin4 |  -8.446005   1.872938    -4.51   0.000    -12.11779   -4.774216
       _cons |    212.518   .5866882   362.23   0.000     211.3679    213.6682
------------------------------------------------------------------------------
```

Note that the use of the robust SE does not make a huge difference here, though it does tend to have a less extreme T statistic (-4.30 vs -4.51) and a wider CI. This is as we would expect, given that the group with the smaller sample size has the larger standard deviation.)

d. Base your analysis on a comparison of geometric mean cholesterol across groups defined by whether the patient died within 4 years or not.

The standard analysis to compare geometric means of a positive continuous random variable across two groups is a t test on log transformed data, with back transformation of the resulting estimates. The problem did not specify whether I should presume equal variances for the two groups or allow for the possibility that the variances might differ across groups. As I believe the latter is better, I use it. (I could of course have just used linear regression with robust SE on the log transformed data.)

```
. ttest logchol, by(deadin4) unequal

Two-sample t test with unequal variances
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |    4467    5.342002    .0027856    .1861762    5.336541    5.347463
       1 |     486    5.297667    .0093744    .2066627    5.279248    5.316087
---------+--------------------------------------------------------------------
combined |    4953    5.337652    .0026816    .1887241    5.332395    5.342909
---------+--------------------------------------------------------------------
    diff |            .0443349    .0097795                .0251269     .063543
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   4.5334
Ho: diff = 0                     Satterthwaite's degrees of freedom =  573.943
```

```
    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

```
. di exp(5.342002), exp(5.297667)
208.93057 199.86997

. di exp(.0443349), exp( .0251269), exp( .063543)
1.0453324 1.0254452 1.0656053
```

Ans: The geometric mean cholesterol is estimated to be 200 mg/dl among subjects who die within 4 years of study entry and 209 mg/dl among subjects who survive at least 4 years. A comparison of the two groups thus estimates that the geometric mean cholesterol is 4.53% higher among subjects who survive 4 years relative to those who die within 4 years. This observed difference is statistically different from 0 (P < 0.0001), with a 95% confidence interval suggesting that the observed difference is what might be typically observed if the true difference between survivors and nonsurvivors was such that the geometric mean for survivors was anywhere between 2.54% and 6.56% higher than that for nonsurvivors. We thus reject the null hypothesis of no association between survival time and cholesterol at study entry in favor of a trend toward higher geometric mean cholesterol among subjects surviving the longer period of time.

e. Base your analysis on a comparison of the odds of dying within 4 years as a function of cholesterol levels. Allow for the possibility that each cholesterol level might have a distinct odds of death within 4 years.

If my model is to allow each cholesterol level to have a distinct odds of surviving for 4 years, I need to use a logistic regression model of the indicator of early death on some continuous form of cholesterol. The problem did not specify how I should model cholesterol, however. For instance, I could consider modeling the cholesterol untransformed, or I could consider taking the logarithm of cholesterol (many others are possible). The model with untransformed cholesterol estimates a common odds ratio for each additive difference in cholesterol, while the model with log transformed cholesterol estimates a common odds ratio for every multiplicative difference in cholesterol (e.g., every doubling). As a general rule, I tend to put in the untransformed variable unless I know that the variable is prone to act multiplicatively (this is often, but not always, the case with highly skewed predictor variables) in a highly diseased population. As this population consisted of more or less healthy elderly Americans, I opted for the untransformed cholesterol as my first choice. I also note that the question did not state whether I should use classical logistic regression or logistic regression with robust standard error estimates. As I believe the latter is better owing to the possibility of handling model misspecification, I use it. I do note, however, that there is not a compelling reason to use the robust SE with logistic regression, because the impact of nonlinearity on the SE is typically small.

```
. logistic deadin4 cholest, robust

Logistic regression                               Number of obs   =       4953
                                                  Wald chi2(1)    =      17.25
                                                  Prob > chi2     =     0.0000
Log pseudolikelihood = -1579.2394                 Pseudo R2       =     0.0065

             |               Robust
     deadin4 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
     cholest |   .9943485   .0013569    -4.15   0.000     .9916926    .9970114

. di .9943485^10,     .9916926^10,    .9970114^10
.94490083 .91996377 .97051274
```

Ans: When comparing two groups with different cholesterol levels, the odds of dying within 4 years is estimated to be 5.51% lower (odds ratio 0.9449) for each 10 mg/dl difference in cholesterol level, with the group having the higher level of cholesterol tending toward a lower odds of death within 4 years. This observed difference is statistically different from an odds ratio of 1 (P < 0.0005), with a 95% confidence interval suggesting that the observed odds ratio is what might be typically observed if the true odds of dying within 4 years was anywhere between 2.95% and 8.00% lower for each 10/mg/dl higher cholesterol level. We thus reject the null hypothesis of no association between survival time and cholesterol at study entry in favor of a trend toward higher odds of survival among subjects with higher cholesterol levels. *(The OR estimated per 1 mg/dl difference in cholesterol is exceedingly small, but statistically significant. Of course, a 1mg/dl difference in cholesterol is not clinically important, though a larger difference would be. I found it useful to talk about a 10 mg/dl difference, and I found the estimates by exponentiating the OR based on a 1 mg/dl by 10. I note that in real life, I might just use the statistical jargon "odds ratio". This term is standard in scientific reports. While I do not believe that everyone truly understands how odds and odds ratios behave, there is not really anything in my verbiage above that addresses the major problem. It would take looking at the odds within specific groups to decide whether a given odds ratio was clinically important (see problem 5).*

I could have compared the odds of dying within 4 years across cholesterol groups by modeling the log cholesterol level. Had this study been in patients with familial hypercholesterolemia, that likely would have been my first choice even with no prior knowledge on the subject. I present here how such an analysis might be presented.

```
. logistic deadin4 logchol, robust

Logistic regression                               Number of obs   =       4953
                                                  Wald chi2(1)    =      22.05
                                                  Prob > chi2     =     0.0000
Log pseudolikelihood = -1577.7383                 Pseudo R2       =     0.0075
------------------------------------------------------------------------------
             |               Robust
     deadin4 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     logchol |   .2979607   .0768285    -4.70   0.000      .179754    .4939003
------------------------------------------------------------------------------

. di .2979607^log(1.1),     .179754^log(1.1),    .4939003^log(1.1)
.89100881 .84910868 .93497652
```

> *Ans: When comparing two groups with different cholesterol levels, the odds of dying within 4 years is estimated to be 10.9% lower (odds ratio 0.891) for each 10% difference in cholesterol level, with the group having the higher level of cholesterol tending toward a lower odds of death within 4 years. This observed difference is statistically different from an odds ratio of 1 (P < 0.0005), with a 95% confidence interval suggesting that the observed odds ratio is what might be typically observed if the true odds of dying within 4 years was anywhere between 6.50% and 15.1% lower for each 10% higher cholesterol level. We thus reject the null hypothesis of no association between survival time and cholesterol at study entry in favor of a trend toward higher odds of survival among subjects with higher cholesterol levels.*

The OR estimated per 1 unit of log cholesterol is not of much clinical interest, as it would be comparing two groups in which on group had a cholesterol level 2.72-fold that of the other group. While such a comparison was possible in our data range (the minimum of 73 mg/dl and maximum 430 mg/dl correspond to a 5.89-fold comparison), I felt that a 10% difference would be

of greater interest. Note the way that such is calculated when you are already given the OR: You exponentiate the OR by the logarithm of 1.1.)

f. Base your analysis on a comparison of the instantaneous risk of death as a function of cholesterol levels. Allow for the possibility that each cholesterol level might have a distinct hazard rate for death.

If my model is to allow each cholesterol level to have a distinct instantaneous risk (hazard) of death, I need to use a proportional hazards regression model of the censored time to death on some continuous cholesterol variable. The problem did not specify how I should model cholesterol, however. For instance, I could consider modeling the cholesterol untransformed, or I could consider taking the logarithm of cholesterol (many others are possible). The model with untransformed cholesterol estimates a common hazard ratio for each additive difference in cholesterol, while the model with log transformed cholesterol estimates a common hazard ratio for every multiplicative difference in cholesterol (e.g., every doubling). As a general rule, I tend to put in the untransformed variable unless I know that the variable is prone to act multiplicatively (this is often, but not always, the case with highly skewed predictor variables) in a highly diseased population. As this population consisted of more or less healthy elderly Americans, I opted for the untransformed cholesterol as my first choice. I also note that the question did not state whether I should use classical proportional hazards regression or proportional hazards regression with robust standard error estimates. As I believe the latter is better owing to the possibility of handling model misspecification and nonproportional hazards, I use it. Unlike with logistic regression, there can be more of a difference between the two approaches in proportional hazards regression.

```
. stset yrtodth death

     failure event:  death != 0 & death < .
obs. time interval:  (0, yrtodth]

. stcox cholest, robust

         failure _d:  death
   analysis time _t:  yrtodth

Cox regression -- Breslow method for ties

No. of subjects      =         4953                 Number of obs    =      4953
No. of failures      =         1111
Time at risk         =  32191.17589
                                                    Wald chi2(1)     =     41.53
Log pseudolikelihood =   -9173.1039                 Prob > chi2      =    0.0000
------------------------------------------------------------------------------
             |               Robust
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     cholest |   .9945688   .0008405    -6.44   0.000     .9929228    .9962175
------------------------------------------------------------------------------

. di  .9945688^10,  .9930136^10,    .9961265^10
.94699636 .93229202 .96193325
```

<u>Ans:</u> When comparing two groups with different cholesterol levels, the instantaneous risk of death is estimated to be 5.30% lower (hazard ratio 0.9449) for each 10 mg/dl difference in cholesterol level, with the group having the higher level of cholesterol tending toward a lower instantaneous risk of death. This observed difference is statistically different from an hazard ratio of 1 (P < 0.0005), with a 95% confidence interval suggesting that the observed hazard ratio is what might be typically observed if the true instantaneous risk of dying was

anywhere between 3.81% and 6.77% lower for each 10/mg/dl higher cholesterol level. We thus reject the null hypothesis of no association between survival time and cholesterol at study entry in favor of a trend toward lower risk of death among subjects with higher cholesterol levels. *(The HR estimated per 1 mg/dl difference in cholesterol is exceedingly small, but statistically significant. Of course, a 1mg/dl difference in cholesterol is not clinically important, though a larger difference would be. I found it useful to talk about a 10 mg/dl difference, and I found the estimates by exponentiating the HR based on a 1 mg/dl by 10. I note that in real life, I might just use the statistical jargon "hazard ratio". This term is standard in scientific reports. While I do not believe that everyone truly understands how hazard ratios as estimated in the Cox model behave, there is not really anything in my verbiage above that addresses the major problem. It would take looking at the way that the hazard functions might be nonproportional over the support of the censoring distribution to understand possible pitfalls. This is beyond the scope of this course (but I did show you one graph illustrating this problem earlier in the course).*

I could have compared the hazards across cholesterol groups by modeling the log cholesterol level. Had this study been in patients with familial hypercholesterolemia, that likely would have been my first choice even with no prior knowledge on the subject. I present here how such an analysis might be presented.

```
. stcox logchol, robust

         failure _d:  death
   analysis time _t:  yrtodth

Cox regression -- Breslow method for ties

No. of subjects      =         4953                 Number of obs    =      4953
No. of failures      =         1111
Time at risk         =  32191.17589
                                                    Wald chi2(1)     =     51.29
Log pseudolikelihood =   -9170.4763                 Prob > chi2      =    0.0000
------------------------------------------------------------------------------
             |               Robust
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     logchol |    .318471   .0508824    -7.16   0.000      .232848    .4355792
------------------------------------------------------------------------------

. di .318471^log(1.1),  .232848^log(1.1),    .4355792^log(1.1)
.89668004 .87031319 .92384566
```

> Ans: *When comparing two groups with different cholesterol levels, the instantaneous risk of dying is estimated to be 10.3% lower (hazard ratio 0.897) for each 10% difference in cholesterol level, with the group having the higher level of cholesterol tending toward a lower instantaneous risk of death. This observed difference is statistically different from an odds ratio of 1 ($P < 0.0005$), with a 95% confidence interval suggesting that the observed hazard ratio is what might be typically observed if the true instantaneous risk of death was anywhere between 7.62% and 13.0% lower for each 10% higher cholesterol level. We thus reject the null hypothesis of no association between survival time and cholesterol at study entry in favor of a trend toward risk of death among subjects with higher cholesterol levels.*

The HR estimated per 1 unit of log cholesterol is not of much clinical interest, as it would be comparing two groups in which on group had a cholesterol level 2.72-fold that of the other group. While such a comparison was possible in our data range (the minimum of 73 mg/dl and maximum 430 mg/dl correspond to a 5.89-fold comparison), I felt that a 10% difference would be

of greater interest. Note the way that such is calculated when you are already given the HR: You exponentiate the OR by the logarithm of 1.1.)

g. How similar are the decisions you make about associations in parts b – e? Which analyses would you have preferred *a priori*? How do these results agree with your prior notions about mortality and serum cholesterol?

Ans: Parts b – e all dichotomized the survival time distribution at 4 years and modeled the cholesterol distribution continuously. As we have highly statistically significant results, looking at the P values is difficult. But looking at the test statistics (which measure the number of SE we are away from the null hypothesis) we find generally similar values-certainly close enough that we do not care about the differences. We do find stronger associations (greater statistical significance, smaller P values) when we model the time to event distribution continuously in a proportional hazards regression. A summary of the Z (or T) statistics from the above analyses (plus a couple more) are given below

Summary Measure	Grouping Variable	Inference	
		Classical	Robust SE
Chol: Pr > 210	Surv: 4 Yr (dich)	4.19	---
Chol: Mean	Surv: 4 Yr (dich)	4.51	4.30
Chol: Geom mean	Surv: 4 Yr (dich)	4.93	4.53
Surv: 4 Yr Odds (dich)	Chol: additive (linear)	4.50	4.13
Surv: 4 Yr Odds (dich)	Chol: multiplicative (log)	4.91	4.70
Surv: Hazard (cts)	Chol: additive (linear)	6.82	6.44
Surv: Hazard (cts)	Chol: multiplicative (log)	7.37	7.10

We see that in this data, the use of robust SE tended to suggest slightly less strong evidence of an association. This would vary according to the evidence for departures from the assumptions about variance in the classical approaches:

- **t test and classical linear regression: homoscedasticity**
- **logistic regression: correct mean-variance relationship (so linearity of the model)**
- **proportional hazards regression: correct mean-variance relationship (so linearity and proportional hazards assumption)**

We also see that modeling the variables continuously tended to provide greater precision. This will be true unless there are extremely striking departures from linearity.

It also appears that the log transformation of cholesterol led to stronger measures of association. It would not surprise me to find that some of the skewness in the cholesterol distribution was due to something of a multiplicative association with survival. But this could just be overfitting the data. I would want to see this replicated in several studies.

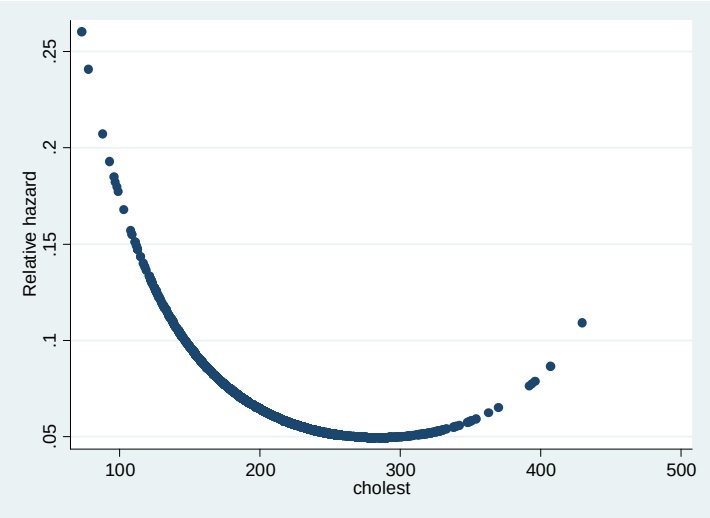
Now, as to the question of whether this is what we would expect. Certainly we are told in the popular press (and many scientific papers as well) that high cholesterol is bad. Current recommendations for cholesterol levels is to treat if cholesterol is above 200 mg/dl (there are more complicated rules dealing with other CVD risk factors).

But in this data we find that higher cholesterol is statistically significantly associated with better survival. That is at least the first order trend. We can explore the second order trend by fitting a proportional hazards model with both cholesterol and cholesterol squared:

```
. g cholsqr= cholest^2
. stcox cholest cholsqr

         failure _d:  death
   analysis time _t:  yrtodth

Cox regression -- Breslow method for ties

No. of subjects =         4953                     Number of obs   =      4953
No. of failures =         1111
Time at risk    =  32191.17589
                                                   LR chi2(2)      =     57.16
Log likelihood  =   -9168.2642                     Prob > chi2     =    0.0000
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     cholest |   .9790614   .0046753    -4.43   0.000     .9699409    .9882678
     cholsqr |   1.000037   .0000111     3.36   0.001     1.000015    1.000059
------------------------------------------------------------------------------
```

Looking at the P value for the cholsqr term, we find a statistically significant association (P= 0.001), by which we can with confidence conclude that we have evidence of a nonlinear association between survival and cholesterol level. To determine what this quadratic function predicts, I ask Stata to compute the estimated hazard ratios for each group relative to a cholesterol level of 0 mg/dl (so, okay, this reference group is silly, but I just want to see which groups that are in my data have the lowest hazard and which have the highest hazard as predicted by this parabolic function). I then plot those estimated hazard ratios against cholesterol. As the following curve shows, we predict a very high hazard for death among subjects with very low cholesterol. As you need good nutritional status, good intestinal absorption, and a functioning liver to make your cholesterol, we may just be picking out those individuals whose organ systems are not functioning well in their old age. That is, increasing infirmity may be causing the low cholesterol, rather than the low cholesterol leading to death. But in any case, we are seeing associations different from the popular view, and it is interesting (at least to me) that the minimal hazard is estimated by this model to be at a cholesterol of 280 mg/dl. (Of course, it is quite unlikely that a parabolic function is the true relationship.)

```
. predict fithr
(option hr assumed; relative hazard)
(47 missing values generated)

. scatter fithr cholest
```



5. Using the analysis in part e of problem 4, provide an estimate of the probability of a subject with a cholesterol of 280 dying within 4 years. *(Note: The Stata commands* `logistic` *and* `logit` *can be used to perform logistic regression in this setting. The* `logistic` *regression output will provide information about the odds ratio. The* `logit` *regression output will present the untransformed intercept and the slope.)*

I perform a logistic regression, estimate the log odds from the regression model, exponentiate that to obtain the odds, and then use the fact that prob = odds / (1 + odds).

```
. logit deadin4 cholest

Logistic regression                               Number of obs   =       4953
                                                  LR chi2(1)      =      20.73
                                                  Prob > chi2     =     0.0000
Log likelihood = -1579.2394                       Pseudo R2       =     0.0065

------------------------------------------------------------------------------
     deadin4 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     cholest |  -.0056675   .0012604    -4.50   0.000    -.0081379   -.0031972
       _cons |  -1.038027   .2626314    -3.95   0.000    -1.552775   -.5232787
------------------------------------------------------------------------------

. di exp(-1.038027 - .0056675 * 280)
.07244505

. di exp(-1.038027 - .0056675 * 280) / (1 + exp(-1.038027 - .0056675 * 280))
.06755129
```

Ans: The logistic regression model of the indicator of death in 4 years on cholesterol estimates that subjects having a serum cholesterol of 280 have an odds of death within 4 years of 0.0724, leading to an estimated probability of death within 4 years of 6.76%. *(Note that the extent to which you might regard the odds and probability to be materially different numbers here would be indicative of whether you felt that the odds ratio was approximating the risk ratio well.)*