**Biost 518: Applied Biostatistics II**

**Biost 515: Biostatistics II**

Emerson, Winter 2015

**Homework #2**

January 13, 2015

**Written problems:** To be submitted as a MS-Word compatible file to the class Catalyst dropbox by noon on Tuesday, January 20, 2015. 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.)*

***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.***

All questions relate to associations between the two biomarkers C-reactive protein (CRP) and fibrinogen (FIB), and how any such association might depend upon prevalence of prior cardiovascular disease (CVD). This homework again uses the subset of information that was collected to examine inflammatory biomarkers and mortality. The data can be found on the class web page (follow the link to Datasets) in the file labeled inflamm.txt. Documentation is in the file inflamm.pdf. See homework #1 for information about reading the data into R and/or Stata.

1. Provide a suitable descriptive statistical analysis for the association between CRP and FIB both overall, and separately for groups having no prior history of diagnosed cardiovascular disease or having prior diagnosed CVD.

**Method: To analyze the association descriptively, I will plot CRP against FIB in a scatter plot of total number of available subjects, and separately for groups having no prior history of diagnosed cardiovascular disease or having prior diagnosed CVD. In addition, the lowess lines are added to the scatter plot to help visualize the trends in between CRP and FIB. Subjects with missing values in CRP, FIB or prior history of diagnosed cardiovascular disease or having prior diagnosed CVD will be excluded from this analysis as well as following analyses.**

**Result: There are 67 subjects with missing values in CRP and 85 subjects with missing values in FIB. 4899 subjects remained in the study after removal these subjects with missing values. Of 4899 subjects, there are 22.98% of subjects having prior diagnosed CVD. From the scatter plots and lowess lines, we notice that CRP and FIB are positively associated with each other for overall subjects and in separate groups. Subjects with prior CVD history seems have stronger positivity trend between CRP and FBI**



1. Perform t test analyses exploring an association between mean fibrinogen and prior history of CVD.
	1. Perform an analysis presuming that the standard deviation of fibrinogen is similar within each group defined by presence of absence of prior history of CVD.

**Method: The two samples t-test assuming equal variance and of non-directional alternative hypothesis is used to compare fibrinogen between groups defined by presence of absence of prior history of CVD.**

**Results: 3777 subjects with no prior history of CVD has mean blood fibrinogen 319.6 mg/L and 1122 subjects with prior history of CVD has mean blood fibrinogen 334.5 mg/L. Based on 95% confidence interval, the blood fibrinogen of subjects with prior CVD history is 14.8 mg/L higher than the blood fibrinogen of subjects with no prior CVD history which would not be unusual if the true difference is between 10.4 and 19.3 mg/L (FIB in subjects with prior CVD history is higher). At a 0.05 level significance (two-side p-value < 0.0001, allowing equal variance), we can reject the null hypothesis that the blood fibrinogen values are not different between subjects with and without prior history of CVD. And the results are in favor of the hypothesis that blood fibrinogen is associated with prior history of CVD.**

* 1. How could the same analysis as presented in part a have been performed with linear regression? Explicitly provide the correspondences between the various statistical output from each of the analyses.

**Simple linear regression assuming homoscedasticity and including FIB as response variable and prior history CVD as exploratory variable will give the same results as in part a. Explicitly, the estimated slope of the linear regression line is 14.8, which corresponds to the point estimate of difference between FIB of two groups. The standard error estimate of the slop is 2.28, from which we can construct 95% CI for slope: 14.8 ± 1.96\*2.28 = (10.4, 19.3), corresponding to the 95% CI of difference between FIB of two groups. The two-side p-value of t-test is the same as the p-value of F test of the simple linear regression.**

* 1. Perform an analysis allowing for the possibility that the standard deviation of fibrinogen might differ across groups defined by presence of absence of prior history of CVD.

**Method: The two samples t-test assuming non-equal variance and of non-directional alternative hypothesis is used to compare fibrinogen between groups defined by presence of absence of prior history of CVD. Noticed that Satterthwaite method is used to approximate variances in test and confidence interval construction.**

**Results: 3777 subjects with no prior history of CVD has mean blood fibrinogen 319.6 mg/L and 1122 subjects with prior history of CVD has mean blood fibrinogen 334.5 mg/L. Based on 95% confidence interval, the blood fibrinogen of subjects with prior CVD history is 14.9 mg/L higher than the blood fibrinogen of subjects with no prior CVD history which would not be unusual if the true difference is between 10.0 and 19.7 mg/L (FIB in subjects with prior CVD history is higher). At a 0.05 level significance (two-side p-value < 0.0001, allowing equal variance), we can reject the null hypothesis that the blood fibrinogen values are not different between subjects with and without prior history of CVD. And the results are in favor of the hypothesis that blood fibrinogen is associated with prior history of CVD.**

* 1. How could a similar analysis as presented in part c have been performed with linear regression? Explicitly provide the correspondences between the various statistical output from each of the analyses.

**Simple linear regression without assuming homoscedasticity and including FIB as response variable and prior history CVD as exploratory variable will give the same results as in part a. Explicitly, the estimated slope of the linear regression line is 14.8, which corresponds to the point estimate of difference between FIB of two groups. The robust standard error estimate of the slop is 2.47, from which we can construct 95% CI for slope: 14.8 ± 1.96\*2.47 = (10.0, 19.7), corresponding to the 95% CI of difference between FIB of two groups. The two-side p-value of t-test is the same as the p-value of F test of the simple linear regression.**

* 1. How could you have used the results of the analysis performed in part a to predict whether the analysis in part c would have found a stronger or weaker association (as measured by the magnitude of the t statistic and p value)?

**In part c, t-test without assuming equal variance will have larger (robust) standard error and smaller degree of freedom than t-test assuming equal variance in part a. Larger standard error will results smaller t statistics (6.06 in part c comparing to 6.51 in part a) and larger p-value (1.71\*10^-9 in part c comparing to 8.23\*10^-11 in part a). Hence we could have used the results in part a to predict weaker association in part c as measured by magnitude of the t statistics and p value.**

For problems 3 – 6, we are interested in exploring alternative approaches to the use of simple linear regression to explore associations between CRP and FIB. In each of those problems, I ask you to report fitted values from the regression. **Please always use at least 4 significant figures when making calculations, and report the fitted values to three significant digits**.

1. Perform a statistical analysis evaluating an association between mean fibrinogen across groups defined by CRP, modeling CRP as a continuous, untransformed random variable.

**Method: The simple linear regression is used to evaluate the association between mean fibrinogen across groups defined by CRP. Robust standard error estimates are used to construct confident interval.**

**Results: The point estimate of the intercept is 304.02 mg /dL (two-side P-value < 0.0001). And the point estimate of the slop is 5.2509 (two-side P-value < 0.0001).**

* 1. Provide an interpretation of the estimated intercept from the fitted regression model as it pertains to fibrinogen levels.

**The estimate mean of fibrinogen is 304.02 mg/dL when CRP is 0 mg/L.**

* 1. Provide an interpretation of the estimated slope from the fitted regression model as it pertains to fibrinogen levels.

**We estimate that for one more mg/L difference in CRP, the mean fibrinogen levels increases in 5.2509 mg/L. Note that it's merely association not casual relationship.**

* 1. Provide full statistical inference about the presence of an association between fibrinogen and CRP using this regression analysis.

**From the simple linear regression, we estimate that for one more mg/L difference in CRP, the mean fibrinogen level increases in 5.2509 mg/dL. A 95% CI suggests that this observation is not unusual if the true difference in mean fibrinogen per one mg/ L CRP is between 4.604 mg/dL and 5.598 mg/dL. Note that it's merely association not casual relationship. At a 0.05 level significance (two-side p-value < 0.0001), we can reject the null hypothesis that there is no linear trend in mean fibrinogen across different CRP group.**

* 1. In a table similar to table 1 below, provide estimates of the central tendency for fibrinogen levels within groups having CRP of 1, 2, 3, 4, 6, 8, 9, and 12 mg/L. (Make clear what summary measure is being estimated).

**See table 1**

1. Repeat problem 3, except perform a statistical analysis evaluating an association between mean fibrinogen across groups defined by CRP, modeling CRP as a continuous, log transformed random variable. (For the purpose of this problem in this homework, replace all observations of CRP=0 with CRP=0.5.)

**Method: The classic simple linear regression is used to evaluate the association between mean fibrinogen across groups defined by log transformed CRP. Note that replace all observations of CRP=0 with CRP=0.5. Robust standard error estimates are used to construct confident interval.**

**Results: The point estimate of the intercept is 295.57 (two-side P-value < 0.0001). And the point estimate of the slop is 36.833(two-side P-value < 0.0001).**

**The estimate mean of fibrinogen is 295.57 mg/dL when log(CRP) is 0.**

**We estimate that for one unit more in log CRP difference, the mean fibrinogen levels increases in 36.833 mg/L. Note that it's merely association not casual relationship.**

**From the simple linear regression, we estimate that for one more mg/L difference in log CRP, the mean fibrinogen level increases in 36.833 mg/L. A 95% CI suggests that this observation is not unusual if the true difference in mean fibrinogen per one unit log CRP is between 34.58 mg /dL and 39.09 mg/dL. Note that it's merely association not casual relationship. At a 0.05 level significance (two-side p-value < 0.0001), we can reject the null hypothesis that there is no linear trend in mean fibrinogen across different log CRP group.**

**See table 1 for estimates.**

1. Repeat problem 3, except perform a statistical analysis evaluating an association between the geometric mean fibrinogen across groups defined by CRP, modeling CRP as a continuous, untransformed random variable.

**Method: The classic simple linear regression is used to estimate log mean fibrinogen across groups defined by CRP. Then, by log back transforming, we can interpret the association between geometric mean fibrinogen across groups defined by CRP. Robust standard error estimates are used to construct confident interval.**

**Results: The point estimate of the intercept is 5.7067 with 0.0030238 as standard error (two-side P-value < 0.0001). And the point estimate of the slop is 0.01392 with 0.0004228 as standard error (two-side P-value < 0.0001).**

**The estimate geometric mean of fibrinogen is exp (5.7067)= 301 mg/dL when CRP is 0.**

**We estimate that ratio of geometric mean of fibrinogen between groups differing in the value of the CRP by 1 mg / L is 1.0140. Note that it's merely association not casual relationship.**

**From the simple linear regression, we estimate that ratio of geometric mean of fibrinogen between groups differing in the value of the CRP by 1 mg / L is 1.0140. A 95% CI suggests that this observation is not unusual if the true ratio of geometric mean of fibrinogen between groups differing in the value of the CRP by 1 mg / L is 1.012 and 1.016. Note that it's merely association not casual relationship. At a 0.05 level significance (two-side p-value < 0.0001), we can reject the null hypothesis that there is no linear trend in log geometric mean fibrinogen across different CRP group.**

**See table 1 for estimates.**

1. Repeat problem 3, except perform a statistical analysis evaluating an association between the geometric mean fibrinogen across groups defined by CRP, modeling CRP as a continuous, log transformed random variable. (For the purpose of this problem in this homework, replace all observations of CRP=0 with CRP=0.5.)

**Method: The classic simple linear regression is used to estimate log mean fibrinogen across groups defined by log CRP. Then, by log back transforming, we can interpret the association between geometric mean fibrinogen across groups defined by log CRP. Robust standard error estimates are used to construct confident interval.**

**Results: The point estimate of the intercept is 5.6786 with 0.003194 as standard error (two-side P-value < 0.0001). And the point estimate of the slop is 0.10539 with 0.002655 as standard error (two-side P-value < 0.0001).**

**The estimate geometric mean of fibrinogen is exp (5.6786)= 293 mg/dL when log CRP is 0.**

**We estimate that ratio of geometric mean of fibrinogen between groups differing in the value of the log CRP by unit is 1.111. Note that it's merely association not casual relationship.**

**From the simple linear regression, we estimate that ratio of geometric mean of fibrinogen between groups differing in the value of log CRP by 1 mg / L is 1.11. A 95% CI suggests that this observation is not unusual if the true ratio of geometric mean of fibrinogen between groups differing in the value of log CRP by 1 unit is 1.105 and 1.118. Note that it's merely association not casual relationship. At a 0.05 level significance (two-side p-value < 0.0001), we can reject the null hypothesis that there is no linear trend in log geometric mean fibrinogen across different log CRP group.**

**See table 1 for estimates.**

**Table 1**: Example of possible display of fitted values. You should indicate the summary measure of the fibrinogen distribution that is being estimated in each column.

|  |  |
| --- | --- |
|  | **Fitted Values for Fibrinogen (mg/dL)** |
| **CRP level** | **Problem 3:** **Mean**  | **Problem 4:** **Mean**  | **Problem 5:** **Geometric Mean** | **Problem 6:****Geometric Mean**  |
| **1 mg/L** | 309 | 296 | 305 | 293 |
| **2 mg/L** | 315 | 321 | 309 | 315 |
| **3 mg/L** | 320 | 336 | 314 | 328 |
| **4 mg/L** | 325 | 347 | 318 | 339 |
| **6 mg/L** | 336 | 362 | 327 | 353 |
| **8 mg/L** | 346 | 372 | 336 | 364 |
| **9 mg/L** | 351 | 377 | 341 | 369 |
| **12 mg/L** | 367 | 387 | 356 | 380 |

1. Complete the following table that makes comparisons (differences or ratios) of the fitted values for each of the models.

**Table 2**: Example of possible display of comparisons of fitted values.

|  |  |
| --- | --- |
|  | **Fitted Values for Fibrinogen (mg/dL)** |
| **Comparisons across CRP level** | **Problem 3:** **Mean** | **Problem 4:** **Mean** | **Problem 5:** **Geometric Mean** | **Problem 6:** **Geometric Mean** |
| ***Differences*** |
| **2 mg/L – 1 mg/L** | 5.25 | 25.5 | 4.28 | 22.2 |
| **3 mg/L – 2 mg/L** | 5.25 | 14.9 | 4.34 | 13.7 |
| **4 mg/L – 1 mg/L** | 15.8 | 51.1 | 13.0 | 46.0 |
| **4 mg/L – 2 mg/L** | 10.5 | 25.5 | 8.73 | 23.9 |
| **6 mg/L – 3 mg/L** | 15.8 | 25.5 | 13.4 | 24.9 |
| **8 mg/L – 4 mg/L** | 21.0 | 25.5 | 18.2 | 25.7 |
| **9 mg/L – 6 mg/L** | 15.8 | 14.9 | 13.9 | 15.4 |
| **9 mg/L – 8 mg/L** | 5.25 | 4.34 | 4.71 | 4.55 |
| **12 mg/L – 6 mg/L** | 31.5 | 25.5  | 28.5 | 26.8 |
| ***Ratios*** |
| **2 mg/L / 1 mg/L** | 1.017 | 1.09 | 1.01 | 1.08 |
| **3 mg/L / 2 mg/L** | 1.017 | 1.05 | 1.01 | 1.04 |
| **4 mg/L / 1 mg/L** | 1.052 | 1.17 | 1.04 | 1.16 |
| **4 mg/L / 2 mg/L** | 1.033 | 1.08 | 1.03 | 1.08 |
| **6 mg/L / 3 mg/L** | 1.049 | 1.08 | 1.04 | 1.08 |
| **8 mg/L / 4 mg/L** | 1.065 | 1.07 | 1.06 | 1.08 |
| **9 mg/L / 6 mg/L** | 1.047 | 1.04 | 1.04 | 1.04 |
| **9 mg/L / 8 mg/L** | 1.015 | 1.01 | 1.01 | 1.01  |
| **12 mg/L / 6 mg/L** | 1.094 | 1.07 | 1.09 | 1.08 |

1. With respect to the results presented in Table 2, answer the following questions:
	1. Which analysis gave constant differences in the fitted values when comparing two groups that differed by an absolute increase in *c* units in CRP levels (i.e., comparing CRP=x to CRP = x+c)? Explicitly provide all those similar paired comparisons from the table.

**The analysis in problem 3 gave constant differences in the fitted values when comparing two groups that differed by an absolute increase in *c* units in CRP levels. Those pairs are highlighted in red. Similar paired comparisons are (2 mg/L – 1 mg/L, 3 mg/L – 2 mg/L, 9 mg/L – 8 mg/L) and (4 mg/L – 1 mg/L, 6 mg/L – 3 mg/L, 9 mg/L – 6 mg/L)**

* 1. Which analysis gave constant ratios of the fitted values when comparing two groups that differed by an absolute increase in *c* units in CRP levels (i.e., comparing CRP=x to CRP = x+c)? Explicitly provide all those similar paired comparisons from the table.

**The analysis in problem 5 gave constant ratios of the fitted values when comparing two groups that differed by an absolute increase in *c* units in CRP levels. Those pairs are highlighted in yellow. Similar paired comparisons are (2 mg/L / 1 mg/L, 3 mg/L / 2 mg/L, 9 mg/L / 8 mg/L) and (4 mg/L / 1 mg/L, 6 mg/L / 3 mg/L, 9 mg/L / 6 mg/L)**

* 1. Which analysis gave constant differences in the fitted values when comparing two groups that differed by a relative *c*-fold increase in CRP levels (i.e., comparing CRP=x to CRP = c \* x )? Explicitly provide all those similar paired comparisons from the table.

 **The analysis in problem 4 gave constant differences in the fitted values when comparing two groups that differed by a relative *c*-fold increase in CRP levels. Those pairs are highlighted in green. Similar paired comparisons are (2 mg/L – 1 mg/L, 4 mg/L – 2 mg/L, 6 mg/L – 3 mg/L, 8 mg/L – 4 mg/L) and (3 mg/L – 2 mg/L, 9 mg/L – 6 mg/L)**

* 1. Which analysis gave constant ratios in the fitted values when comparing two groups that differed by a relative *c*-fold increase in CRP levels (i.e., comparing CRP=x to CRP = c \* x )? Explicitly provide all those similar paired comparisons from the table.

**The analysis in problem 6 gave constant ratios in the fitted values when comparing two groups that differed by a relative *c*-fold increase in CRP levels. Those pairs are highlighted in purple. Similar paired comparisons are (2 mg/L / 1 mg/L, 4 mg/L / 2 mg/L, 6 mg/L / 3 mg/L, 8 mg/L / 4 mg/L) and (3 mg/L / 2 mg/L, 9 mg/L / 6 mg/L)**

1. How would you decide which of the four potential analyses should be used to investigate associations between fibrinogen and CRP?

**First of all, based on the last homework key and Scott, because the biochemistry property of CRP, multiplicative level for CRP levels is preferred in the analysis. Hence, log transformed CRP is preferred. Secondly, without prior information about property of fibrinogen (such as multiplicative model, the standard deviation of response in a group is proportional to the mean), the difference in mean is easier to interpret. In addition, we don't have desire to down weight outliers. Hence, we don't prefer geometric mean of fibrinogen.**

**In short, I decide to use the analysis in problem 4.**