**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.

**Methods**: The sample size, mean, standard deviation, median, minimum, maximum, and interquartile range were calculated for the variables CRP and fibrinogen using the inflamm.txt dataset. The same descriptive data was also calculated for the subsets of participants with and without a history of cardiovascular disease (CVD).

**Results**: The results are summarized in the tables below. Overall, the mean CRP level was 3.61 mg/L with a standard deviation of 6.15 mg/L, and the median CRP level was 2 mg/L with an IQR of 1 to 3 mg/L. The overall mean fibrinogen level was 323 mg/dL with a standard deviation of 67.3 mg/dL, and the median fibrinogen level was 311 mg/dL with an IQR of 281 to 367 mg/dL. Among participants without a history of CVD, the mean CRP level was 3.38 mg/L with a standard deviation of 5.90 mg/L, and the median CRP level was 2 mg/L with an IQR of 1 to 3 mg/L. Among those without a history of CVD, the mean fibrinogen level was 320 mg/dL with a standard deviation of 64.8 mg/dL, and the median fibrinogen level was 311 mg/dL with an IQR of 277 to 361 mg/dL. Finally, among participants with a history of CVD, the mean CRP level was 4.40 mg/L with a standard deviation of 6.88 mg/L, and the median CRP level was 2 mg/L with an IQR of 1 to 5 mg/L. Among those with a history of CVD, the mean fibrinogen level was 334 mg/dL with a standard deviation of 74.1 mg/dL, and the median fibrinogen level was 328 mg/dL with an IQR of 285 to 367 mg/dL.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Overall Descriptive Statistics for CRP and Fibrinogen | | | | | | | | |
| Variable | n | Mean | SD | Median | Min | 25% | 75% | Max |
| CRP (mg/L) | 4933 | 3.61 | 6.15 | 2 | 0 | 1 | 3 | 108 |
| Fibrinogen (mg/dL) | 4915 | 323 | 67.3 | 311 | 109 | 281 | 367 | 872 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Descriptive Statistics for CRP & Fibrinogen among Participants without History of CVD | | | | | | | | |
| Variable | n | Mean | SD | Median | Min | 25% | 75% | Max |
| CRP (mg/L) | 3802 | 3.38 | 5.90 | 2 | 0 | 1 | 3 | 108 |
| Fibrinogen (mg/dL) | 3791 | 320 | 64.8 | 311 | 109 | 277 | 361 | 872 |
| Descriptive Statistics for CRP & Fibrinogen among Participants with History of CVD | | | | | | | | |
| Variable | n | Mean | SD | Median | Min | 25% | 75% | Max |
| CRP (mg/L) | 1131 | 4.40 | 6.88 | 2 | 0 | 1 | 5 | 83 |
| Fibrinogen (mg/dL) | 1124 | 334 | 74.1 | 328 | 138 | 285 | 367 | 695 |

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.

**Methods**: The t-test assuming equal variances across groups was used to compare mean fibrinogen levels among participants with and without a history of CVD. It was also used to generate 95% confidence intervals for the difference in means between groups.

**Results**: The mean fibrinogen level among participants with a history of CVD is 335 mg/dL and among those without a history of CVD, it is 320 mg/dL. The finding of a 14.9 mg/dL lower mean fibrinogen level among participants without a history of CVD would not be unusual if the true difference in the means was between a 10.4 mg/dL to 19.3 mg/dL lower mean fibrinogen level among participants without a history of CVD using a 95% confidence interval. Based on the results of the t test, this finding is statistically significant at a 0.05 level of significance (two-sided p <0.00001). This means we can reject the null hypothesis that the mean fibrinogen levels are not different based on history of CVD and support the alternative hypothesis that a history of CVD is associated with higher mean fibrinogen levels.

* 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.

**Methods**: Classical linear regression analysis was performed using history of previous CVD (yes/no) as the predictor variable and mean fibrinogen level as the outcome variable. It was also used to generate 95% confidence intervals for the slope and y-intercept of the regression model.

**Results**: The regression model gives a mean fibrinogen level of 320 mg/dL when the value of the predictor variable is 0 (i.e. no history of CVD) and an increase in mean fibrinogen level of 14.9 mg/dL when the value of the predictor variable is 1 (i.e. positive history of CVD). This finding would not be unusual if the true increase in mean fibrinogen level was between a 10.4 mg/dL to 19.3 mg/dL higher mean fibrinogen level among participants with a history of CVD using a 95% confidence interval. Based on the results of the classical linear regression, this finding is statistically significant at a 0.05 level of significance (p <0.0001). This means we can reject the null hypothesis that the mean fibrinogen levels are not different based on history of CVD and support the alternative hypothesis that a history of CVD is associated with higher mean fibrinogen levels.

**Correspondence between parts a & b**: The mean fibrinogen level in the group without a history of CVD using the t-test is identical to the mean fibrinogen level in the regression model when the value of the predictor is 0 (i.e. no history of CVD) 🡪 both are 319.574 mg/dL. The difference in the mean fibrinogen level between groups and the 95% confidence interval for the difference in means using the t-test is identical to the slope in the regression model and its 95% confidence interval (which is equivalent to the change in mean fibrinogen level for a 1 unit change in the predictor) 🡪 both are 14.88508 (95% CI: 10.42389, 19.34628).

* 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.

**Methods**: The t-test assuming *unequal* variances across groups was used to compare mean fibrinogen levels among participants with and without a history of CVD. It was also used to generate 95% confidence intervals for the difference in means between groups.

**Results**: The results of this analysis are the same as in part a). The mean fibrinogen level among participants with a history of CVD is 335 mg/dL and among those without a history of CVD, it is 320 mg/dL. The finding of a 14.9 mg/dL lower mean fibrinogen level among participants without a history of CVD would not be unusual if the true difference in the means was between a 10.1 mg/dL to 19.7 mg/dL lower mean fibrinogen level among participants without a history of CVD using a 95% confidence interval. Based on the results of the t test, this finding is statistically significant at a 0.05 level of significance (two-sided p <0.00001). This means we can reject the null hypothesis that the mean fibrinogen levels are not different based on history of CVD and support the alternative hypothesis that a history of CVD is associated with higher mean fibrinogen levels.

* 1. How could a smilar 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.

**Methods**: Linear regression analysis assuming unequal variances was performed using history of previous CVD (yes/no) as the predictor variable and mean fibrinogen level as the outcome variable. It was also used to generate 95% confidence intervals for the slope and y-intercept of the regression model.

**Results**: The regression model gives a mean fibrinogen level of 320 mg/dL when the value of the predictor variable is 0 (i.e. no history of CVD) and an increase in mean fibrinogen level of 14.9 mg/dL when the value of the predictor variable is 1 (i.e. positive history of CVD). This finding would not be unusual if the true increase in mean fibrinogen level was between a 10.1 mg/dL to 19.7 mg/dL higher mean fibrinogen level among participants with a history of CVD using a 95% confidence interval. Based on the results of the linear regression, this finding is statistically significant at a 0.05 level of significance (p <0.0001). This means we can reject the null hypothesis that the mean fibrinogen levels are not different based on history of CVD and support the alternative hypothesis that a history of CVD is associated with higher mean fibrinogen levels.

**Correspondence between parts c & d**: The mean fibrinogen level and 95% confidence interval in the group without a history of CVD using the t-test with unequal variances is identical to the mean fibrinogen level and 95% CI in the regression model when the value of the predictor is 0 (i.e. no history of CVD) 🡪 both are 319.574 mg/dL (95% CI: 317.5117, 321.6362). The difference in the mean fibrinogen level between groups using the t-test with unequal variances is identical to the slope in the regression model (which is equivalent to the change in mean fibrinogen level for a 1 unit change in the predictor) 🡪 both are 14.88508.

* 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)?

The t-test assuming unequal variances generally results in weaker associations (a smaller t statistic and lower p value) than the t-test assuming equal variance. This can be predicted using the 95% confidence intervals of the difference in the means for the two tests. The 95% CI is wider for the t-test that assumes unequal variances.

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.

**Methods**: Linear regression analysis assuming unequal variances was performed using CRP level (continuous, untransformed) as the predictor variable and mean fibrinogen level as the outcome variable. It was also used to generate 95% confidence intervals for the slope and y-intercept of the regression model.

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

The y-intercept of the regression model gives a mean fibrinogen level of 304 mg/dL when the value of the predictor variable is 0 (i.e. CRP = 0 mg/L).

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

The slope of the regression model gives a change in mean fibrinogen level of 5.25 mg/dL for every 1 unit increase in the value of the predictor variable CRP level.

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

**Results**: The regression model gives a mean fibrinogen level of 304 mg/dL when the value of the predictor variable is 0 (i.e. CRP = 0) and an increase in mean fibrinogen level of 5.25 mg/dL for every 1 unit increase in the value of the predictor variable (CRP level). This finding would not be unusual if the true increase in mean fibrinogen level was between a 4.60 mg/dL to 5.90 mg/dL higher mean fibrinogen level for every 1 unit increase in CRP level, using a 95% confidence interval. Based on the results of the linear regression, this finding is statistically significant at a 0.05 level of significance (p <0.0001). This means we can reject the null hypothesis that the mean fibrinogen levels are not different based on CRP level and support the alternative hypothesis that a higher CRP level is associated with higher mean fibrinogen levels.

* 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 results in table 1 below.

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.)

**Methods**: Linear regression analysis assuming unequal variances was performed using CRP level (continuous, log transformed) as the predictor variable and mean fibrinogen level as the outcome variable. It was also used to generate 95% confidence intervals for the slope and y-intercept of the regression model.

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

The y-intercept of the regression model gives a mean fibrinogen level of 296 mg/dL when the value of the predictor variable is 0 (i.e. CRP = 1 mg/L).

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

The slope of the regression model gives a change in mean fibrinogen level of 36.8 mg/dL for every 1 unit increase in the value of the predictor variable, ln(CRP).

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

**Results**: The regression model gives a mean fibrinogen level of 296 mg/dL when the value of the predictor variable is 0 (i.e. CRP = 1) and an increase in mean fibrinogen level of 36.8 mg/dL for every 1 unit increase in the value of the predictor variable (ln(CRP)). This finding would not be unusual if the true increase in mean fibrinogen level was between a 34.6 mg/dL to 39.1 mg/dL higher mean fibrinogen level for every 1 unit increase in ln(CRP) level, using a 95% confidence interval. Based on the results of the linear regression, this finding is statistically significant at a 0.05 level of significance (p <0.0001). This means we can reject the null hypothesis that the mean fibrinogen levels are not different based on ln(CRP) level and support the alternative hypothesis that a higher ln(CRP) level is associated with higher mean fibrinogen levels.

* 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 results in table 1 below.

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.

**Methods**: Linear regression analysis assuming unequal variances was performed using CRP level (continuous, untransformed) as the predictor variable and geometric mean fibrinogen level as the outcome variable. It was also used to generate 95% confidence intervals for the slope and y-intercept of the regression model.

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

The y-intercept of the regression model gives a geometric mean fibrinogen level of 301 mg/dL (or e^5.71) when the value of the predictor variable is 0 (i.e. CRP = 0 mg/L).

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

The slope of the regression model gives a change in geometric mean fibrinogen level of 1.01 mg/dL (or e^0.0139) for every 1 unit increase in the value of the predictor variable, CRP level.

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

**Results**: The regression model gives a geometric mean fibrinogen level of 301 mg/dL when the value of the predictor variable is 0 (i.e. CRP = 0) and an increase in geometric mean fibrinogen level of 1.01 mg/dL for every 1 unit increase in the value of the predictor variable (CRP level). This finding would not be unusual if the true increase in geometric mean fibrinogen level was between a 1.01 mg/dL to 1.02 mg/dL higher geometric mean fibrinogen level for every 1 unit increase in CRP level, using a 95% confidence interval. Based on the results of the linear regression, this finding is statistically significant at a 0.05 level of significance (p <0.0001). This means we can reject the null hypothesis that the geometric mean fibrinogen levels are not different based on CRP level and support the alternative hypothesis that a higher CRP level is associated with higher geometric mean fibrinogen levels.

* 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 results in table 1 below.

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.)

**Methods**: Linear regression analysis assuming unequal variances was performed using CRP level (continuous, log transformed) as the predictor variable and geometric mean fibrinogen level as the outcome variable. It was also used to generate 95% confidence intervals for the slope and y-intercept of the regression model.

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

The y-intercept of the regression model gives a geometric mean fibrinogen level of 293 mg/dL (or e^5.68) when the value of the predictor variable is 0 (i.e. CRP = 1 mg/L).

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

The slope of the regression model gives a change in geometric mean fibrinogen level of 1.11 mg/dL (or e^0.105) for every 1 unit increase in the value of the predictor variable, ln(CRP).

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

**Results**: The regression model gives a geometric mean fibrinogen level of 293 mg/dL when the value of the predictor variable is 0 (i.e. CRP = 1) and an increase in geometric mean fibrinogen level of 1.11 mg/dL for every 1 unit increase in the value of the predictor variable (ln(CRP)). This finding would not be unusual if the true increase in geometric mean fibrinogen level was between a 1.10 mg/dL to 1.12 mg/dL higher geometric mean fibrinogen level for every 1 unit increase in ln(CRP) level, using a 95% confidence interval. Based on the results of the linear regression, this finding is statistically significant at a 0.05 level of significance (p <0.0001). This means we can reject the null hypothesis that the geometric mean fibrinogen levels are not different based on ln(CRP) level and support the alternative hypothesis that a higher ln(CRP) level is associated with higher geometric mean fibrinogen levels.

* 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 results in table 1 below.

**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 fibrinogen level)** | **Problem 4: (Mean fibrinogen level)** | **Problem 5: (Geometric mean fibrinogen level)** | **Problem 6: (Geometric mean fibrinogen level)** |
| **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 fibrinogen level)** | **Problem 4: (Mean fibrinogen level)** | **Problem 5: (Geometric mean fibrinogen level)** | **Problem 6: (Geometric mean fibrinogen level)** |
| ***Differences*** | | | | |
| **2 mg/L – 1 mg/L** | 6 | 25 | 4 | 22 |
| **3 mg/L – 2 mg/L** | 5 | 15 | 5 | 13 |
| **4 mg/L – 1 mg/L** | 16 | 51 | 13 | 46 |
| **4 mg/L – 2 mg/L** | 10 | 26 | 9 | 24 |
| **6 mg/L – 3 mg/L** | 16 | 26 | 13 | 25 |
| **8 mg/L – 4 mg/L** | 21 | 25 | 18 | 25 |
| **9 mg/L – 6 mg/L** | 15 | 15 | 14 | 16 |
| **9 mg/L – 8 mg/L** | 5 | 5 | 5 | 5 |
| **12 mg/L – 6 mg/L** | 31 | 25 | 29 | 27 |
| ***Ratios*** | | | | |
| **2 mg/L / 1 mg/L** | 1.02 | 1.08 | 1.01 | 1.08 |
| **3 mg/L / 2 mg/L** | 1.02 | 1.05 | 1.02 | 1.04 |
| **4 mg/L / 1 mg/L** | 1.05 | 1.17 | 1.04 | 1.16 |
| **4 mg/L / 2 mg/L** | 1.03 | 1.08 | 1.03 | 1.08 |
| **6 mg/L / 3 mg/L** | 1.05 | 1.08 | 1.04 | 1.08 |
| **8 mg/L / 4 mg/L** | 1.06 | 1.07 | 1.06 | 1.07 |
| **9 mg/L / 6 mg/L** | 1.04 | 1.04 | 1.04 | 1.05 |
| **9 mg/L / 8 mg/L** | 1.01 | 1.01 | 1.01 | 1.01 |
| **12 mg/L / 6 mg/L** | 1.09 | 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 regression analyses from Problem 3 and Problem 5 (both using CRP as an untransformed continuous random variable) gave constant differences in the fitted values for fibrinogen when comparing groups that differed by an absolute increase of *c* units in CRP levels. For every 1 unit increase in CRP level, the difference in fibrinogen level was approximately 5 mg/dL. This was seen in the 2 -1 mg/L CRP row, the 3 – 2 mg/L CRP row and the 9 – 8 mg/L CRP row (all highlighted in yellow above). For every 3 unit increase in CRP level, the difference in fibrinogen level was approximately 15 mg/dL. This was seen in the 4 – 1 mg/L CRP row, the 6 – 3 mg/L CRP row, and the 9 – 6 mg/L CRP row (all highlighted in green above).

* 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.
  2. 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.
  3. 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.

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