**Biost 518: Applied Biostatistics II**

**Biost 515: Biostatistics II**

Emerson, Winter 2015

**Homework #2**

January 13, 2015

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.

Answer 1. Methods: Descriptive statistics are presented as means and standard deviation for continuous variables or percentages for binary variables. Figure 1 shows the descriptive statistics by groups defined by those with no prior history of cardiovascular disease and those with a prior history of cardiovascular disease.

Inference: Data was available for 5000 subjects. The biomarker CRP was missing for 67 of those subjects and was excluded from the analysis. The overall sample mean CRP was 3.61 mg/L and the same SD was 6.17 mg/L. The biomarker fibrinogen was missing in 85 of the 5000 subjects and was excluded from the analysis. The sample mean for fibrinogen was 323.02 mg/dL and standard deviation was 67.35 mg/dL. The following figure shows the scatterplots of CRP and fibrinogen in those with no prior history of cardiovascular disease (pink) and in those with prior history of cardiovascular disease (blue).

The mean fibrinogen in the 3777 subjects who had no prior history of cardiovascular disease, the sample mean CRP was 3.39mg/L and standard deviation was 5.91 mg/L. The mean fibrinogen was 319.62 mg/dL and standard deviation was 64.83 mg/dL. In those with a prior history of cardiovascular disease, the sample mean CRP was 4.41mg/L and standard deviation was 6.90 mg/L. The mean fibrinogen was 334.46 mg/dL and standard deviation was 74.11 mg/dL.

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.

Answer 2a. Methods: A t test (assuming equal variance) was performed to compare the differences in means of fibrinogen across groups defined by those who had no prior history of CVD compared to those who did have a history of CVD excluding those without measurements of fibrinogen (N=4915). A 95% confidence interval for the difference in means was similarly based on the possibility of unequal variances and p value with alpha of 0.05 was considered significant.

Inference: The mean fibrinogen level was 319.57 mg/dl among the 3791 subjects who had no history of CVD and 334.46mg/dl in the 1124 subjects with a history of CVD. Based on a 95% confidence interval the observed difference of 14.89mg/dl higher mean fibrinogen in those who had a prior history of CVD would not be unusual if the true difference was between 10.42mg/dl and 19.35mg/dl higher among those with a prior history of CVD. Using a t test, this difference was statistically significant at p = <0.001, so we can reject the null hypothesis that the mean fibrinogen levels are not different between those who do and do not have a 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.

Answer 2b. The analysis in part 2a could have been performed with linear regression as well. The use of the robust SE in linear regression leads to the inference that approximates the t test (with equal variance). The point estimates are the same as well as the SE for the estimated differences in means (2.275). The confidence intervals are also almost identical because the SE are the same and the degrees of freedom used to calculate the difference are almost the same (t test DF: 4913 and regression DF: 4914). The p values are the same as well (p<0.0001).

* 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.

Answer 2c. Methods: A t test (assuming unequal variance) was performed to compare the differences in means of fibrinogen across groups defined by those who had no prior history of CVD compared to those who did have a history of CVD. A 95% confidence interval for the difference in means was similarly based on the possibility of unequal variances and p value with alpha of 0.05 was considered significant. Those with missing values of fibrinogen were excluded.

Inference: The mean fibrinogen level was 319.57 mg/dl among the 3791 subjects who had no history of CVD and 334.46mg/dl in the 1124 subjects with a history of CVD. Based on a 95% confidence interval the observed difference of 14.89mg/dl higher mean fibrinogen in those who had a prior history of CVD would not be unusual if the true difference was between 10.09 mg/dl and 19.68 mg/dl higher among those with a prior history of CVD. Using a t test, this difference was statistically significant at p = <0.001, so we can reject the null hypothesis that the mean fibrinogen levels are not different between those who do and do not have a 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.

Answer 2d. The analysis in part 2c could have been performed with linear regression as well. The use of the robust SE in linear regression leads to the inference that approximates the t test (with unequal variance). The point estimates are the same as well as the SE for the estimated differences in means (t test SE: 2.45 compared to robust SE with linear regression: 2.45). The confidence intervals differ slightly because the SE differs but also because of the different degrees of freedom used to calculate the difference (t test DF: 1664 compared to regression DF: 4914). The larger DF leads to a slightly larger confidence interval of 10.08926 mg/dl to 19.68091 mg/dl compared to 10.08409 mg/dl and 10.08607 mg/dl in the t test and the same p value (<0.001).

* 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)?

Answer 2e. The results from part a could have been used to predict the analysis in part c because the use of a t test that assumes equal variances will have the same point estimate for the difference in means but the SE for the estimated difference in means will vary between the two (t test with equal variance: 2.27 compared to t test with unequal variance SE: 2.44). The confidence intervals differ as well because the SE differs but also because of the different degrees of freedom used to calculate the difference (t test DF: 4915 compared to unequal DF: 1664). The larger DF leads to a slightly wider confidence interval but the same 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.
	1. Provide an interpretation of the estimated intercept from the fitted regression model as it pertains to fibrinogen levels.

Answer a. The intercept of 5.25mg/dL is the estimated mean fibrinogen level for a subject with a CRP of 0 mg/L.

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

Answer b. The slope of 0.33mg/dL is the estimated difference in mean fibrinogen level for a subject with a CRP change of 1 mg/L.

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

Answer c. The intercept estimates a mean fibrinogen level of 5.25mg/dl among subjects with a CRP of 0 mg/L. The slope of 0.33mg/dL is the estimated difference in mean fibrinogen level for a subject with a CRP change of 1 mg/L. The 95% confidence interval for the intercept estimates the mean fibrinogen level is 4.60 to 5.90 mg/dl. The p value is <0.001 so we can reject the null hypothesis that there is no relationship between fibrinogen levels and CRP 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).

Answer d. See 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.)

Answer 4. The intercept estimates a mean fibrinogen level of 36.83mg/dl among subjects with a CRP of 0 mg/L. The slope of 1.15 mg/dL is the estimated difference in mean fibrinogen level for a subject with a log CRP change of 1 mg/L. The 95% confidence interval for the intercept estimates the mean fibrinogen level is 34.58 to 39.09 mg/dl. The p value is <0.001 so we can reject the null hypothesis that there is no relationship between fibrinogen levels and log CRP levels. See table 1 below for estimates of central tendency.

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.

Answer 5. The intercept estimates a mean log fibrinogen level of 0.014mg/dl among subjects with a CRP of 0 mg/L. The slope of 0.0009 mg/dL is the estimated difference in mean log fibrinogen level for a subject with a CRP change of 1 mg/L. The 95% confidence interval for the intercept estimates the mean log fibrinogen level is 0.012 to 0.016 mg/dl. The p value is <0.001 so we can reject the null hypothesis that there is no relationship between log fibrinogen levels and CRP levels. See table 1 below for estimates of central tendency.

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.)

Answer 6. The intercept estimates a mean log fibrinogen level of 0.105 mg/dl among subjects with a log CRP of 0 mg/L. The slope of 0.003 mg/dL is the estimated difference in mean log fibrinogen level for a subject with a log CRP change of 1 mg/L. The 95% confidence interval for the intercept estimates the mean log fibrinogen level is 0.100 to 0.111 mg/dl. The p value is <0.001 so we can reject the null hypothesis that there is no relationship between log fibrinogen levels and log CRP levels. See table 1 below for estimates of central tendency.

**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: (mean log fibrinogen level)** | **Problem 6: (mean log fibrinogen level)** |
| **1 mg/L** | 5.58 | 37.98 | 0.0148 | 0.1080 |
| **2 mg/L** | 5.91 | 39.13 | 0.0157 | 0.1110 |
| **3 mg/L** | 6.24 | 40.28 | 0.0166 | 0.1140 |
| **4 mg/L** | 6.57 | 41.43 | 0.0175 | 0.1170 |
| **6 mg/L** | 7.23 | 43.73 | 0.0193 | 0.1229 |
| **8 mg/L** | 7.89 | 46.03 | 0.0211 | 0.1289 |
| **9 mg/L** | 8.22 | 47.18 | 0.0220 | 0.1319 |
| **12 mg/L** | 9.21 | 50.63 | 0.0247 | 0.1409 |

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: (mean log fibrinogen level)** | **Problem 6: (mean log fibrinogen level)** |
| ***Differences*** |
| **2 mg/L – 1 mg/L** | 0.330 | 1.150 | 0.000900 | 0.002988 |
| **3 mg/L – 2 mg/L** | 0.330 | 1.150 | 0.000900 | 0.002989 |
| **4 mg/L – 1 mg/L** | 0.990 | 3.450 | 0.002699 | 0.008966 |
| **4 mg/L – 2 mg/L** | 0.660 | 2.300 | 0.001800 | 0.005977 |
| **6 mg/L – 3 mg/L** | 0.990 | 3.450 | 0.002699 | 0.008966 |
| **8 mg/L – 4 mg/L** | 1.320 | 4.600 | 0.003599 | 0.011954 |
| **9 mg/L – 6 mg/L** | 0.990 | 3.450 | 0.002699 | 0.008966 |
| **9 mg/L – 8 mg/L** | 0.330 | 1.150 | 0.000900 | 0.002989 |
| **12 mg/L – 6 mg/L** | 1.980 | 6.900 | 0.005399 | 0.017931 |
| ***Ratios*** |
| **2 mg/L / 1 mg/L** | 1.059 | 1.030 | 1.061 | 1.028 |
| **3 mg/L / 2 mg/L** | 1.056 | 1.029 | 1.057 | 1.027 |
| **4 mg/L / 1 mg/L** | 1.177 | 1.091 | 1.182 | 1.083 |
| **4 mg/L / 2 mg/L** | 1.112 | 1.059 | 1.115 | 1.054 |
| **6 mg/L / 3 mg/L** | 1.159 | 1.086 | 1.163 | 1.079 |
| **8 mg/L / 4 mg/L** | 1.201 | 1.111 | 1.206 | 1.102 |
| **9 mg/L / 6 mg/L** | 1.137 | 1.079 | 1.140 | 1.073 |
| **9 mg/L / 8 mg/L** | 1.042 | 1.025 | 1.043 | 1.023 |
| **12 mg/L / 6 mg/L** | 1.274 | 1.158 | 1.280 | 1.146 |

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.

Answer a. All of the analyses had constant differences in the fitted values in the differences in the mean fibrinogen level when comparing two groups of CRP levels. The differences between CRP 1 and 2 mg/L, 2 and 3 mg/L, and 8 and 9 mg/L was constant for mean CRP and mean fibrinogen, log-CRP and mean fibrinogen, and mean CRP and log-fibrinogen, and log-CRP and log-fibrinogen (0.33mg/dL, and 1.15 mg/dL, 0.009mg/dL, and 0.002989 mg/dL, respectively).

* 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.

Answer b. None of the analyses had constant ratios in the fitted values in the differences in the mean fibrinogen level when comparing two groups of CRP levels that differed by an absolute increase in c units. The differences between the ratios with CRP 2/1 mg/L and 3/2 mg/L and 9/8mg/L were all different for all four groups.

* 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.

Answer c. None of the analyses had constant differences in the fitted values in the differences in the mean fibrinogen level when comparing two groups of CRP levels that differed by a relative c-fold increase. The differences between CRP 1 and 2 mg/L and 2 and 4 mg/L and 3 and 6 mg/L increased for mean CRP and mean fibrinogen (0.33 mg/dL, 0.66 mg/dL, 0.99 mg/dL, respectively). The log-CRP and mean fibrinogen performed similarly (1.15mg/dL, 2.30mg/dL, 3.45mg/dL, respectively), as well as the mean CRP and log-fibrinogen (0.0009, 0.0018, 0.0027, respectively) and the log-CRP and log-fibrinogen (0.003mg/dL, and 0.006 mg/dL, and 0.012mg/dL, respectively).

* 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.

Answer d. None of the analyses had constant ratios in the fitted values in the differences in the mean fibrinogen level when comparing two groups of CRP levels that differed by a c-fold increase. The differences between the ratios with CRP 2/1 mg/L and 4/2 mg/L and 8/4mg/L and 12/6 mg/L were all different.

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

Answer 9. The best analysis to investigate associations between fibrinogen and CRP should be the one that uses log-CRP and mean fibrinogen. The histograms from these variables show that the CRP values are heavily skewed toward 0 and the mean fibrinogen is relatively normally distributed. The mean differences in the fitted values is similar between the four analyses and the t test has similar SE and 95% confidence intervals and they were all statistically significant so the differences are not that great. Interpreting the values is more difficult when both variables are log transformed.