**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:** Patients are divided into two groups based on whether they have prior history of diagnosed cardiovascular disease or not. Descriptive statistics are presented in the table below of each group as well as the entire sample population. For serum fibrinogen level, the mean, standard deviation and the data range are included. A scatter plot with lowess lines of each group and total sample population is also presented below.

**Results:** This set of data includes a total of 5000 subjects, but there are 67 subjects are missing data on their CRP levels and 34 subjects are missing data on their fibrinogen levels. These subjects are omitted in the data analysis. The rest of the subjects (n=4899) are divided into two groups based on whether they have prior history of diagnosed cardiovascular disease or not.

The descriptive statistics within two groups are demonstrated in the table below. There is a trend of increasing mean fibrinogen level with the increasing CRP level in both groups, however, the effect is stronger in the group with prior history of CVD.



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Prior history of CVD | | No prior history of CVD | |
| CRP level  (mg/L) | N | Fib  (mg/dL) | N | Fib  (mg/dL) |
| 0-20 | 1087 | 330  (69.1, 138-662) | 3710 | 317  (59.9, 109-624) |
| 20-40 | 28 | 444  (75.7, 270-584) | 49 | 457  (123, 235-872) |
| 40-60 | 4 | 532  (69.1, 445-596) | 13 | 509  (114, 274-741) |
| 60-80 | 2 | 627  (96.2, 559-695) | 2 | 528  (47.4, 494-561) |
| 80-100 | 1 | 674  (n/a, 674-674) | 1 | 518  (n/a, 518-518) |
| 100-120 |  |  | 2 | 556  (129, 464-647) |
| Total | 1122 | 335  (74.1, 138-695) | 3777 | 320  (64.8, 109-872) |

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:** A t test (with the assumption of equal variances) was performed to compare mean serum fibrinogen levels between subjects who have prior history of CVD and those who don’t have prior history of CVD. 95% confidence intervals for the difference in population means were estimated based on the same assumption.

**Results:** For the subjects who don’t have the prior history of CVD (n=3777), the mean serum fibrinogen level is 319.6 mg/dL. For the subjects who have the prior history of CVD (n=1122), the mean serum fibrinogen level is 334.5 mg/dL. On average, the serum fibrinogen level of the subjects who have the prior history of CVD is 14.85 mg/dL higher than that of the subjects who don’t have the prior history of CVD. With 95% confidence, this observed result would not be unusual if the true mean serum fibrinogen level of the subjects who have the prior history of CVD is higher by 10.38 to 19.32mg/dL than that of the subjects who don’t have the prior history of CVD. This observation is statistically significant at a 0.05 level of significance based on a t test (two-sided P< 0.0001). Therefore, we can reject the null hypothesis that the mean serum fibrinogen levels are the same between the subjects who have the prior history of CVD and the subjects who don’t have the prior history of CVD with high confidence. We conclude that the mean serum fibrinogen level is associated with the prior history of CVD of the subjects

* 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:** A linear regression analysis (with the assumption of equal variances) was performed on mean serum fibrinogen across the groups with or without prior history of CVD in a saturated model.

**Results:**

|  |  |  |  |
| --- | --- | --- | --- |
| t test | | Linear regression | |
| Mean FIB level  in the group with no prior CVD | 319.6 | Intercept | 319.6 |
| Difference in FIB level between two groups | -14.85 | Slope | 14.85 |
| Two side P value | 0.000 | P value | 0.000 |
| 95% CI in difference | (-19.32- -10.38) | 95% CI | (10.38-19.32) |

* 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:** A t test (with the assumption of unequal variances) was performed to compare mean serum fibrinogen levels between subjects who have prior history of CVD and those who don’t have prior history of CVD. 95% confidence intervals for the difference in population means were estimated based on the same assumption.

**Results:** For the subjects who don’t have the prior history of CVD (n=3777), the mean serum fibrinogen level is 319.6 mg/dL. For the subjects who have the prior history of CVD (n=1122), the mean serum fibrinogen level is 334.5 mg/dL. On average, the serum fibrinogen level of the subjects who have the prior history of CVD is 14.85 mg/dL higher than that of the subjects who don’t have the prior history of CVD. With 95% confidence, this observed result would not be unusual if the true mean serum fibrinogen level of the subjects who have the prior history of CVD is higher by 10.04 to 19.65mg/dL than that of the subjects who don’t have the prior history of CVD. This observation is statistically significant at a 0.05 level of significance based on a t test (two-sided P< 0.0001). Therefore, we can reject the null hypothesis that the mean serum fibrinogen levels are the same between the subjects who have the prior history of CVD and the subjects who don’t have the prior history of CVD with high confidence. We conclude that the mean serum fibrinogen level is associated with the prior history of CVD of the subjects

* 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:** A linear regression analysis was performed on mean serum fibrinogen across the groups with or without prior history of CVD in a saturated model using Huber-White estimates of the standard error.

**Results:**

|  |  |  |  |
| --- | --- | --- | --- |
| t test | | Linear regression | |
| Mean FIB level  in the group with no prior CVD | 319.6 | Intercept | 319.6 |
| Difference in FIB level between two groups | -14.85 | Slope | 14.85 |
| Two side P value | 0.000 | P value | 0.000 |
| 95% CI in difference | (-19.65- -10.04) | 95% CI | (10.04- 19.65) |

* 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:** As shown in the table in Question 1, the standard deviation of serum FIB levels for the group of smaller sample size (the group with prior history of CVD) is larger than that of the group without prior history of CVD. Therefore, the result reported in part a with an assumption of equal variances would be anti-conservative inference. Reported p values are too small and the confidence interval is too narrow.

Compared to part a, t test in part c with an assumption of unequal variances would have a larger p-value, smaller t statistic and wider CI, which indicates a weaker association.

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:** The intercept is 304 mg/dL. It is an estimate that the mean serum fibrinogen level would be 304 mg/dL if the CRP level of the subject is 0 mg/L.

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

**Answer:** The slope is 5.25mg/dL. It is an estimate that for every 1 mg/L difference in serum CRP level between two groups of subjects, the mean fibrinogen level is 5.25mg/dL higher in the higher population.

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

**Methods:** A linear regression analysis was performed on an association between mean serum fibrinogen level and CRP level using Huber-White estimates of the standard error. The slop of the regression model and the 95% confidence interval were obtained.

**Results:** Based on the linear regression analysis, we estimate that for every 1 mg/L difference in serum CRP level between two groups of subjects, the mean fibrinogen level is 5.25mg/dL higher in the higher population. With 95% confidence, it would not be unusual if the true relationship between the means were such that the higher group’s mean fibrinogen levels were between 4.60mg/dL and 5.90 mg/dL higher for each 1 mg/L difference in serum CRP level. With high confidence (the two sided P value is P < .0005), we reject the null hypothesis that there is no linear trend in the average FIB across CRP groups.

* 1. See table 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.)
   1. The intercept is 296 mg/dL. It is an estimate that the mean serum fibrinogen level would be 296 mg/dL if the CRP level of the subject is 1 mg/L.
   2. The slop is 36.8 mg/dL. This indicates that for every 1mg/L difference in log transformed serum CRP level between two groups of subjects, the mean serum fibrinogen level would be 36.8 mg/dL higher for the group with higher CRP level.
   3. **Methods:** A linear regression analysis was performed on an association between mean serum fibrinogen level and the log transformed CRP level using Huber-White estimates of the standard error. The slop of the regression model and the 95% confidence interval were obtained. The subjects with the serum CRP level equal to 0 mg/L have been changed to 0.5mg/L (half of the detection limit in this study).

**Results:** This set of data includes a total of 5000 subjects, but there are 67 subjects are missing data on their CRP levels and 34 subjects are missing data on their fibrinogen levels. These subjects are omitted in the data analysis. There were 426 subjects that had 0 mg/L of the serum CRP level, which had been replaced by 0.5 mg/L. Based on the linear regression analysis, we estimate that for every 10% difference in serum CRP level between two groups of subjects, the mean fibrinogen level is 3.51 mg/dL higher in the higher population. With 95% confidence, it would not be unusual if the true relationship between the means were such that the higher group’s mean fibrinogen levels were between 3.30 mg/dL and 3.73 mg/dL higher for each 10% difference in serum CRP level. With high confidence (the two sided P value is P < .0005), we reject the null hypothesis that there is no linear trend in the average FIB across CRP groups.

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.
   1. The intercept is 5.71 mg/dL. It is an estimate that the geometric mean serum fibrinogen level would be 301 mg/dL (calculated by e5.71) if the CRP level of the subject is 0 mg/L.
   2. The slop is 0.0139 mg/dL. This indicates that for every 1 mg/L difference in serum CRP levels between two groups of subjects, the geometric mean fibrinogen level is 1.40% higher in the higher population.
   3. **Methods:** A linear regression analysis was performed on an association between geometric mean serum fibrinogen level and CRP level using Huber-White estimates of the standard error. The slop of the regression model and the 95% confidence interval were obtained.

**Results:** Based on the linear regression analysis, we estimate that for every 1 mg/L difference in serum CRP level between two groups of subjects, the mean fibrinogen level is 1.40% higher in the higher population. With 95% confidence, it would not be unusual if the true relationship between the means were such that the higher group’s mean fibrinogen levels were between 1.22% and 1.58% higher for each 1 mg/L difference in serum CRP level. With high confidence (the two sided P value is P < .0005), we reject the null hypothesis that there is no linear trend in the average FIB across CRP groups.

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.)
   1. The intercept is 5.68 mg/dL. It is an estimate that the geometric mean serum fibrinogen level would be 293 mg/dL (calculated by e5.68) if the CRP level of the subject is 1 mg/L.
   2. The slop is 0.105 mg/dL. This indicates that for every 10% difference in serum CRP level between two groups of subjects, the mean fibrinogen level is 1.01% higher in the higher population.
   3. **Methods:** A linear regression analysis was performed on an association between geometric mean serum fibrinogen level and log transformed CRP level using Huber-White estimates of the standard error. The slop of the regression model and the 95% confidence interval were obtained. The subjects with the serum CRP level equal to 0 mg/L have been changed to 0.5mg/L (half of the detection limit in this study).

**Results:** This set of data includes a total of 5000 subjects, but there are 67 subjects are missing data on their CRP levels and 34 subjects are missing data on their fibrinogen levels. These subjects are omitted in the data analysis. There were 426 subjects that had 0 mg/L of the serum CRP level, which had been replaced by 0.5 mg/L. Based on the linear regression analysis, we estimate that for every 10% difference in serum CRP level between two groups of subjects, the mean fibrinogen level is 1.01% higher in the higher population. With 95% confidence, it would not be unusual if the true relationship between the means were such that the higher group’s mean fibrinogen levels were between 0.953% and 1.07% mg/dL higher for each 10% difference in serum CRP level. With high confidence (the two sided P value is P < .0005), we reject the null hypothesis that there is no linear trend in the average FIB across CRP groups.

**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 | 376 | 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 | 25.53 | 4.28 | 22.17 |
| **3 mg/L – 2 mg/L** | 5.25 | 14.93 | 4.34 | 13.74 |
| **4 mg/L – 1 mg/L** | 15.75 | 51.06 | 13.01 | 46.02 |
| **4 mg/L – 2 mg/L** | 10.50 | 25.53 | 8.73 | 23.85 |
| **6 mg/L – 3 mg/L** | 15.75 | 25.53 | 13.38 | 24.89 |
| **8 mg/L – 4 mg/L** | 21.00 | 25.53 | 18.21 | 25.66 |
| **9 mg/L – 6 mg/L** | 15.75 | 14.93 | 13.95 | 15.43 |
| **9 mg/L – 8 mg/L** | 5.25 | 4.34 | 4.71 | 4.55 |
| **12 mg/L – 6 mg/L** | 31.51 | 25.53 | 28.49 | 26.78 |
| ***Ratios*** | | | | |
| **2 mg/L / 1 mg/L** | 1.02 | 1.09 | 1.01 | 1.08 |
| **3 mg/L / 2 mg/L** | 1.02 | 1.05 | 1.01 | 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 | 1.08 |
| **9 mg/L / 6 mg/L** | 1.05 | 1.04 | 1.04 | 1.04 |
| **9 mg/L / 8 mg/L** | 1.02 | 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.

**Answer:** The linear regression analysis that was used to assess the association between mean FIB across groups defined by CRP (as performed in Question 3). The differences are constant among pairs (2 mg/L – 1mg/L, 3 mg/L – 2 mg/L, 9 mg/L – 8 mg/L) and among pairs (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..

**Answer:** The linear regression analysis that was used to assess the association between the mean of log transformed FIB levels across groups defined by CRP (as performed in Question 5). The ratios are constant among pairs (2 mg/L – 1mg/L, 3 mg/L – 2 mg/L, 9 mg/L – 8 mg/L) and among pairs (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.

**Answer:** The linear regression analysis that was used to assess the association between mean FIB levels across groups defined by log transformed CRP levels (as performed in Question 4). The differences are constant among pairs (2 mg/L /1mg/L, 4 mg/L / 2 mg/L, 6 mg/L / 3 mg/L, 8 mg/L / 4 mg/L, 12 mg/L / 6 mg/L) and among pairs (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.

**Answer:** The linear regression analysis that was used to assess the association between the mean of log transformed FIB levels across groups defined by log transformed CRP levels (as performed in Question 6). The ratios are constant among pairs (2 mg/L / 1mg/L, 4 mg/L / 2 mg/L, 6 mg/L / 3 mg/L, 8 mg/L / 4 mg/L, 12 mg/L / 6 mg/L) and among pairs (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?

**Answer:** Following are the potential analyses that I would prefer:

1. The linear regression analysis was performed on an association between mean serum fibrinogen level and CRP level (as performed in Question 3).
2. The linear regression analysis was performed on an association between geometric mean serum fibrinogen level and log transformed CRP level (as performed in Question 6).

These two analysis focus on different aspects of population (the first one on the absolute change in CRP value and the second one is on the ratio change in CRP level). They are both straight forward and easy to be understood and can be used in different circumstances.

At last, I do notice that the wide range of CRP level (it can be different by orders of magnitude across individuals). One benefit of the second analysis (as performed in Question 6) is that it works on log transformed CRP levels. Therefore, the CRP level will be in a closer range, which also downweight the outliers.