**Biost 518: Applied Biostatistics II**

**Biost 515: Biostatistics II**

Emerson, Winter 2015

**Grade: 151/195**

**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: Firstly, we exclude those observations with CRP, FIB and prior history of cardiovascular disease not available from all of our analysis. Scatter plot for CRP and FIB were obtained and stratified by the history of CVD. Lowess lines were also fitted overall and for groups stratified by the history of CVD. In the plot, blue circle stands for subject without CVD history, red triangle stands for subject with CVD. Lowess lines for each CVD history status were marked by different color: blue for subject without CVD history and red for subject with CVD, as well as for the combined sample (black) were also displayed. The descriptive statistics of FIB for subjects with different level of CRP (subjects with CRP level less than 1 mg/l, between 1mg/l to 3 mg/l and greater than 3 mg/l) were also compared between groups stratifying by whether the subjects had prior history of CVD or not.**

**Result: Totally, there are 4899 subjects included in our analysis. The plot below is a scatterplot of CRP level and FIB level, and their scatterplot stratified by CVD. Overall, the scatter plot suggests a linear trend of higher FIB level for subjects with higher CRP level; this trend still exists after stratifying the subjects into two group by their CVD history. The slopes of the lowess lines of two groups defined by CVD history are similar, suggesting the history of CVD has little effect modification. Our finding is further illustrated by the following table, which shows that mean FIB levels tend to be higher for subjects with higher CRP level both overall and in each group (with prior CVD history or without prior CVD history).**

****

|  |  |  |  |
| --- | --- | --- | --- |
|  | Mean FIB(mean(sd,max-min))(mg/dl) | | |
| CRP (mg/l) | NO CVD (n=3777) | CVD (n=1122) | Overall (n=4899) |
| <1 mg/l | 277.5(48.5,172.0-436.0, n=348) | 290.2(57.9,180.0-540.0, n=78) | 279.8(50.5,172.0-540.0, n=426) |
| 1 mg/l ~ 3mg/l | 310.0(52.5,109.0-562.0,n=2597) | 314.8(55.6,138.0-592.0, n=709) | 311.1(53.2,109.0-592.0, n=3306) |
| 3 mg/l | 367.2(78.9,132.0-872.0,n=832) | 386.3(84.5,175.0-695.0,n=335) | 372.7(81.0,132.0-872.0, n=167) |

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: Here we compare the mean fibrinogen levels between subjects with prior history of CVD and subjects without prior history of CVD. A t test that assumes equal variance for these two groups was used to test the differences between the means. 95% confidence interval for the difference in population means was based on the similar t test.**

**Result: The estimated mean level of fibrinogen for the 1122 subjects with prior history of CVD is 334.46 mg/dl. The estimated mean level of fibrinogen for the 3777 subjects without prior history of CVD is 319.62 mg/dl. With 95% confidence, this observed difference of 14.85 mg/dl higher mean fibrinogen level for subjects with prior history of CVD would not be unusual if the true difference of mean were between 10.38 mg/dl and 19.32 mg/dl. By a t test assuming equal variance, at the significance level of 0.05, this observation is statistically significant (two-side P value< 0.00001), hence with high confidence we conclude that the mean level of fibrinogen differs between subjects with prior history of CVD and subjects without 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.

**Method: Classical linear regression model that assumes homoscedasticity were applied to analyze the association between mean fibrinogen and prior history of CVD. In the model, the history of CVD is treated as predictor and the fibrinogen level is treated as response.**

**Result: The estimated mean level of fibrinogen for the 1122 subjects with prior history of CVD is 334.46 mg/dl. The estimated mean level of fibrinogen for the 3777 subjects without prior history of CVD is 319.62 mg/dl. With 95% confidence, this observed difference of 14.85 mg/dl higher mean fibrinogen level for subjects with prior history of CVD would not be unusual if the true difference mean were anywhere between 10.38 mg/dl and 19.32 mg/dl. Based on a two-side P value< 0.00001, we could with high confidence conclude that the mean level of fibrinogen differs between subjects with prior history of CVD and subjects without prior history of CVD. This is exactly the same inference that was reported in part a.**

* 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: Mean fibrinogen levels were compared between subjects with prior history of CVD and subjects without prior history of CVD. Differences between the means were tested by a t test that does not assume equal variance between these two groups. 95% confidence interval for the difference in population means was based on the similar t test.**

**Result: The estimated mean level of fibrinogen for the 1122 subjects with prior history of CVD is 334.46 mg/dl. The estimated mean level of fibrinogen for the 3777 subjects without prior history of CVD is 319.62 mg/dl. With 95% confidence, this observed tendency of 14.85 mg/dl higher mean fibrinogen level for subjects with prior history of CVD would not be unusual if the true difference mean were anywhere between 10.04 mg/dl and 19.65 mg/dl. By a t test without assuming equal variance, at the significance level of 0.05, this observation is statistically significant (two-side P value< 0.00001), hence with high confidence we conclude that the mean level of fibrinogen differs between subjects with prior history of CVD and subjects without prior history of CVD.**

* 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.

**Method: Linear regression model that allows heteroscedasticity were applied to analyze the association between mean fibrinogen and prior history of CVD. In the model, the history of CVD is treated as predictor and the fibrinogen level is treated as response.**

**Result: The estimated mean level of fibrinogen for the 1122 subjects with prior history of CVD is 334.46 mg/dl. The estimated mean level of fibrinogen for the 3777 subjects without prior history of CVD is 319.62 mg/dl. With 95% confidence, this observed tendency of 14.85 mg/dl higher mean fibrinogen level for subjects with prior history of CVD would not be unusual if the true difference mean were anywhere between 10.04 mg/dl and 19.65 mg/dl. Based on a two-side P value< 0.00001, we could with high confidence conclude that the mean level of fibrinogen differs between subjects with prior history of CVD and subjects without prior history of CVD. This is exactly the same inference that was reported in part a.**

* 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: Since the t test that assumes equal variance has larger degrees of freedom than t test without assuming equal variance, thus this would lead to a smaller critical value for test assuming equal variance. Also note that when assuming equal variance, the standard would be smaller than that when not assuming equal variance, which would lead to smaller critical value for test assuming equal variance. Thus we can predict that the critical value of t test assuming equal variance would have smaller t statistic and thus lead to smaller p value. Thus we can predict analysis in part c would find 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.

**Method: A linear regression model of mean fibrinogen level as a function of CRP was fitted to access linear trends. Standard errors were computed using the Huber-White sandwich estimator, which allows the possible heterscedasticity.**

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

**Answer: The estimated intercept of the fitted model is 304.0 mg/dl. This is the estimated mean fibrinogen level of subjects with CRP level 0 mg/l.**

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

**Answer: The estimated slope is 5.25 mg/dl. This is the estimated value of increase every time when the mean level of CRP increased by 1 mg/l.**

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

**Answer: By fitting a linear regression model of sample size 4899, the estimated difference in mean fibrinogen level is 5.25 mg/dl for each unit of mg/l difference in CRP level, with higher mean fibrinogen for subjects with higher CRP level. With 95% confidence, this observation is not unusual if the true difference of mean FIB level was between 4.60 mg/dl to 5.90 mg/dl higher per mg/l higher in CRP level. Based on a two side P value<0.00001, this result is significant thus we could conclude with high confidence that the fibrinogen level is associated with CRP level.**

* 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: In 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:** **Linear regression model assuming heteroscedasticity were fitted for FIB level and log transformed CRP level. All observations of CRP=0 were replaced with CRP=0.5. 95% confidence interval was obtained based on the Huber-White estimates of the standard error.**

**Result: The estimate intercept of the fitted model is 295.6 mg/dl, which means when CRP=1 mg/l the estimated mean FIB level is 296.5 mg/dl. The estimate slope is 3.68 mg/dl, which means when increasing the level of log transformed CRP level by one unit, the mean level of FIB would increase 3.68 mg/dl. The estimated difference in mean fibrinogen level is 36.8 mg/dl for each unit of mg/l difference in log transformed CRP level, with higher mean fibrinogen for subjects with higher CRP level. With 95% confidence, this observation is not unusual if the true difference in mean FIB level were between 34.6 mg/dl to 39.1 mg/dl higher per mg/l higher in log transformed CRP level. Based on a two side P value<0.00001, this result is significant thus we could conclude with high confidence that the fibrinogen level is associated with CRP level.**

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:** **Linear regression model assuming heteroscedasticity were fitted to test the association between the geometric mean of FIB level across the group with different CRP level. 95% confidence interval was obtained based on the Huber-White estimates of the standard error.**

**Result: The estimated intercept from the fitted regression model was 301 mg/dl after exponentiation, which means when CRP=0 mg/l, the estimated geometric mean FIB value is 301 mg/dl. The estimated slope was 1.014 after exponentiation. It means that the estimated geometric mean FIB value would increase by 1.4% if we increase 1 mg/L in CRP. When comparing two groups of subjects differing in CRP level by 1 mg/l, the estimated geometric mean of fibrinogen level is 1.4% higher for subjects with higher CRP level. With 95% confidence, this observation is not unusual if the true difference between geometric means of FIB levels were between 1.2% to 1.6% higher for subjects with a unit CRP level higher. Based on a two side P value<0.00001, this result is significant thus we could conclude with high confidence that the fibrinogen level is associated with CRP level.**

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 model assuming were applied to test the association between the geometric mean of FIB levels across groups differed by different CRP level. All observations with CRP=0 were replaced by CRP=0.5. 95% confidence interval was obtained based on the Huber-White estimates of the standard error.**

**Result: The estimated intercept was 292.5 mg/dl after exponentiation, which means when CRP=1, the estimated geometric mean FIB value is 292.5 mg/dl. The estimated slope was 1.111 after exponentiation, which means that the estimated geometric mean FIB value would increase by 11.1% if we increase 1 unit in log transformed CRP. When comparing two groups of subjects differing in log transformed CRP level by 1 mg/l, the estimated geometric mean of fibrinogen level is 11.1% higher for subjects with higher CRP level. With 95% confidence, this observation is not unusual if the true difference between geometric means were between 10.5% to 11.8% higher for subject with one unit log transformed CRP level higher. Based on a two side P value<0.00001, this result is significant thus we could conclude with high confidence that the fibrinogen level is associated with CRP level.**

**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** | 309.3 | 295.6 | 305.1 | 292.5 |
| **2 mg/L** | 314.5 | 321.1 | 309.4 | 314.7 |
| **3 mg/L** | 319.8 | 336.0 | 313.7 | 328.4 |
| **4 mg/L** | 325.0 | 346.6 | 318.1 | 338.5 |
| **6 mg/L** | 335.5 | 361.5 | 327.1 | 353.3 |
| **8 mg/L** | 346.0 | 372.1 | 336.3 | 364.2 |
| **9 mg/L** | 351.3 | 376.5 | 341.1 | 368.7 |
| **12 mg/L** | 367.0 | 387.0 | 355.6 | 380.1 |

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** | 5.3 | 25.5 | 4.3 | 22.2 |
| **3 mg/L – 2 mg/L** | 5.3 | 15 | 4.3 | 13.7 |
| **4 mg/L – 1 mg/L** | 15.8 | 51 | 13 | 46 |
| **4 mg/L – 2 mg/L** | 10.5 | 25.5 | 8.7 | 23.8 |
| **6 mg/L – 3 mg/L** | 15.8 | 25.5 | 13.4 | 24.9 |
| **8 mg/L – 4 mg/L** | 21 | 25.5 | 18.2 | 25.7 |
| **9 mg/L – 6 mg/L** | 15.8 | 15 | 13.9 | 15.4 |
| **9 mg/L – 8 mg/L** | 5.3 | 4.4 | 4.8 | 4.5 |
| **12 mg/L – 6 mg/L** | 31.5 | 25.5 | 28.5 | 26.8 |
| ***Ratios*** | | | | |
| **2 mg/L / 1 mg/L** | 1.012 | 1.086 | 1.014 | 1.076 |
| **3 mg/L / 2 mg/L** | 1.017 | 1.046 | 1.014 | 1.043 |
| **4 mg/L / 1 mg/L** | 1.051 | 1.173 | 1.043 | 1.157 |
| **4 mg/L / 2 mg/L** | 1033 | 1.079 | 1.028 | 1.076 |
| **6 mg/L / 3 mg/L** | 1.049 | 1.076 | 1.043 | 1.076 |
| **8 mg/L / 4 mg/L** | 1.065 | 1.074 | 1.057 | 1.076 |
| **9 mg/L / 6 mg/L** | 1.047 | 1.041 | 1.043 | 1.044 |
| **9 mg/L / 8 mg/L** | 1.015 | 1.012 | 1.014 | 1.012 |
| **12 mg/L / 6 mg/L** | 1.094 | 1.071 | 1.087 | 1.076 |

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 analysis in problem 3 did so. The absolute CRP increase for “2-1”, “3-2”, “9-8” are all 1, and the difference in the mean of FIB are all 5.3 mg/dl. The absolute CRP increase for “4-1”, “6-3”, “9-6” are all 3, and the difference in the mean of FIB are all 15.8 mg/dl.**

* 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 analysis in problem 5 did so. The absolute CRP increase for “2-1”, “3-2”, “9-8” are all 1, and the ratio in the geometric mean of FIB are all 1.014. The absolute CRP increase for “4-1”, “6-3”, “9-6” are all 3, and the ratio in the geometric mean of FIB are all around 1.043.**

* 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 analysis in problem 4 did so. The CRP increase for “2/1”, “4/2”, “6/3”, “12/6” are all 2-fold, and the difference in the mean of FIB are all 25.5 mg/dl. The CRP increase for “3/2”, “9/6” are all 1.5-fold, and the difference in the mean of FIB are all 15 mg/dl.**

* 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 analysis in problem 6 did so. The CRP increase for “2/1”, “4/2”, “6/3”, “12/6” are all 2-fold, and the ratios in the geometric mean of FIB are all 1.076. The CRP increase for “3/2”, “9/6” are all 1.5-fold, and the ratios in the geometric mean of FIB are all 1.043.**

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

**Answer: According to the biological characteristic of fibrinogen, it might be better to use geometric mean to fit the model. And notice when fitting model, we don’t necessarily need to log transform our predictor, and when doing so have to deal with those CRP with value 0, which would bring errors to our analysis. Thus I would use the analysis in question 5.**