Homework 2

Biostatistics 515/518

ID code: 2416

Date: 01/20/15

**Question 1**:

**Method**: In order to first address the question of whether inflammatory biomarkers serum C-reactive protein (CRP) or fibrinogen (FIB) are associated with each other, and whether this association might depend on the prevalence of cardiovascular disease (CVD), I used data collected on 5,000 elderly (age 65+) subjects across four different communities. Of these 5,000 subjects, 16 were missing just CRP measurements, 34 were just missing FIB measurements, and 51 were missing both, leaving 4899 subjects with measurements for both CRP and FIB. All subsequent analyses will be carried out only on these subjects with CRP and FIB measurements. I created a table of some relevant descriptive statistics for these subjects [Table 1].

**Answer**: Of the 4899 subjects with data on both CRP and FIB, 1122 had a history of CVD, as defined by prevalent atherosclerotic disease at study enrollment, including history of previous angina, myocardial infarction, transient ischemic attack, and/or stroke, and 3777 subjects did not have a history of CVD. These two groups had similar mean age, BMI, serum LDL levels, as well as similar percentages of subjects with a history of smoking. The group with a history of CVD had a greater percentage of males (53.0 vs 38.5%) compared with those without a history of CVD. Brachial systolic blood pressure, a risk factor for developing CVD disease was similar between the two groups. The mean ratio of ankle systolic blood pressure to brachial systolic BP, where lower numbers indicate current hardening of the arteries, was lower in those with a history of CVD (1.014 vs 1.079). When looking at the serum markers of interest, those with a history of CVD had higher mean serum CRP levels (4.41 vs 3.39 mg/L) and higher mean FIB levels (334.5 vs 319.6 mg/dL).

|  |  |  |  |
| --- | --- | --- | --- |
| Characteristics | No CVD\*  n = 3777 | History of CVD\*  n = 1122 | All Subjects  n = 4899 |
| Age (yrs) | 72.5 (5.4, 65.0 – 98) | 74.0 (5.8, 65.0 – 100.0) | 72.8 (5.6, 65.0 – 100.0) |
| Male (%) | 38.5 | 53.0 | 41.8 |
| BMI1 | 26.6 (4.7, 14.7 – 58.8) | 27.0 (4.9, 16.7 – 48.0) | 26.7 (4.7, 14.7 – 58.8) |
| Serum LDL (mg/dL) | 212.6 (30.8, 78 – 407) | 208.5 (40.4, 73 – 430) | 211.7 (39.2, 73 -430) |
| Smoking History1 (%) | 12.5 | 11.0 | 12.1 |
| Syst. BP (mmHg) 1 | 136.1 (24.4, 79 – 227) | 137.5 (22.9, 80 – 235) | 136.4 (21.8, 79 – 235) |
| BP ratio‡1 | 1.079 (0.16, 0.278 – 2.385) | 1.014 (0.21, 0.298 – 2.148) | 1.065 (0.17, 0.278 – 2.385) |
| Serum CRP (mg/L) | 3.39 (5.92, 0 – 108) | 4.41 (6.90, 0 – 83) | 3.62 (6.17, 0 – 108) |
| Serum FIB (mg/dL) | 319.6 (64.8, 109 – 872) | 334.5 (74.1, 138 – 695) | 323.0 (67.4, 109 – 872) |

Numbers represent the mean (standard deviation, minimum – maximum), unless stated otherwise.

Abbreviations used: CVD, cardiovascular disease; BMI, Body mass index; LDL, Low density lipoprotein (measure of cholesterol); Syst. BP, Systolic blood pressure, CRP, C-reactive protein; FIB, Fibrinogen

\* History of cardiovascular disease (CVD) determined by prevalent atherosclerotic disease at study enrollment, including history of previous angina, myocardial infarction, transient ischemic attack, and/or stroke.

‡ BP ratio is the ratio of systolic blood pressure in the ankle to that in the arm

1 There were 13 subjects missing data on BMI, 2 subjects missing data on serum LDL, 6 subjects missing data on smoking history, 10 missing systolic blood pressure measurements, and 117 missing the BP ratio.

**Question 2**:

**Part A**.

**Method**: To explore an association between mean fibrinogen and prior history of CVD, I first used a variety of t tests on the subjects dichotomized by presence of CVD as defined in question 1. In part A, I used a t-test that assumed that the distributions of fibrinogen values in each group had equal variances. Without a strong basis to predict the direction of association, I decided to do a two-sided t-test.

**Answer**: The group with a history of CVD had a mean serum fibrinogen level of 334.5 mg/dL that was 14.9 mg/dL higher than the mean serum fibinogen level of 319.6 mg/dL in the group without history of CVD. The 95% confidence interval suggests that this difference is not surprising if the true difference were between 10.4 and 19.3 mg/dL higher in the group with a history of CVD. Based on the two-sided p value <0.0001, we can reject the null hypothesis that the mean serum fibrinogen levels are the same in the two groups.

**Part B.**

**Method**: In part B, I still assumed equal variances in the distribution of fibrinogen values in the linear regression model that I applied. I this model, the predictor of interest was history of CVD (dichotomized as in question 1), the response variable was serum fibrinogen, and I was modeling the mean.

**Answer**: Based on linear regression analysis, we estimate that the difference between the mean serum fibrinogen level in the two groups dichotomized by history of CVD is 14.8 mg/dL. The 95% confidence interval suggests that this difference is not surprising if the true difference were between 10.4 and 19.3 mg/dL higher in the group with a history of CVD. Based on the two-sided p value <0.0001, we can reject the null hypothesis that there is no difference in the mean serum fibrinogen levels between the two groups. Explicitly, the estimate of the slope of the linear regression line equals the point estimate for the difference in mean fibrinogen in the two groups (with some differences due to rounding). The lower and upper bound of the 95% CI for the slope corresponds to the lower and upper bounds of the 95% CI of the estimate in the difference of means in the two sided two sample t test in A. The p-value calculated on the F statistic for the slope corresponds to the p-value of the two-sided t test.

**Part C.**

**Method:** In part C, I used a two-sided t-test that **did not** assume that the distributions of fibrinogen values in each group had equal variances.

**Answer:** The group with a history of CVD had a mean serum fibrinogen level of 334.5 mg/dL that was 14.9 mg/dL higher than the mean serum fibinogen level of 319.6 mg/dL in the group without history of CVD. The 95% confidence interval suggests that this difference is not surprising if the true difference were between 10.0 and 19.7 mg/dL higher in the group with a history of CVD. Based on the two-sided p value <0.0001, we can reject the null hypothesis that the mean serum fibrinogen levels are the same in the two groups.

**Part D.**

**Method:** In part D, I **did not** assume equal variances in the distribution of fibrinogen values in the linear regression model that I applied (i.e. used Huber-White sandwich estimator). In this model, the predictor of interest was still history of CVD (dichotomized as in question 1), the response variable was serum fibrinogen, and I was modeling the mean.

**Answer:** Based on linear regression analysis, we estimate that the difference between the mean serum fibrinogen level in the two groups dichotomized by history of CVD is 14.8 mg/dL. The 95% confidence interval suggests that this difference is not surprising if the true difference in means were between 10.0 and 19.7 mg/dL higher in the group with a history of CVD. Based on the two-sided p value <0.0001, we can reject the null hypothesis that there is no difference in the mean serum fibrinogen levels between the two groups. Explicitly, the estimate of the slope of the linear regression line equals the point estimate for the difference in mean fibrinogen in the two groups (with some differences due to rounding). The lower and upper bound of the 95% CI for the slope corresponds to the lower and upper bounds of the 95% CI of the estimate in the difference of means in the two sided two sample t test in A. The p-value calculated on the F statistic for the slope corresponds to the p-value of the two-sided t test.

**Part E.**

**Answer:** Because we have the descriptive statistics from question 1, we can see that the group with a smaller number of subjects (with a history of CVD) had a larger SD (and thus variance), which indicates that analyses that assume homoscedasticity will tend to be anti-conservative, or estimate a confidence interval that is more narrow than it should be. This is also reflected in a p value that is smaller than that in the test allowing for heteroscedasticity.

**Question 3**:

**Method**: I used linear regression with Huber-White sandwich estimator to model the predictor of interest, untransformed continuous serum CRP levels on the mean of the response variable, the serum fibrinogen levels.

**Answer:**

**a)** The estimated intercept from the fitted regression model indicates that we estimate subjects with a serum CRP of 0 mg/L to have a mean serum fibrinogen level of 304.0 mg/dL (95% CI 301.5 – 306.5 mg/dL).

**b)** The estimated slope of the model suggests that as serum CRP levels increase by each 1 mg/L, we estimate the mean serum fibrinogenlevel to increase by 5.251 mg/dL (95% CI 4.604 – 5.898 mg/dL).

**c)** Based on this linear regression model, we estimate that when the serum CRP level is 0 mg/L the mean serum fibrinogen level is 304.0 mg/dL and, as the serum CRP level increases by 1 mg/L, the mean serum fibrinogen level increases by 5.351 mg/dL. The 95% confidence intervals indicate that we would not be surprised if the true mean fibrinogen level at a CRP of 0 mg/L were between 301.5 and 306.5 mg/dL and if the true difference of mean serum fibrinogen across groups differing by 1 mg/L were between 4.604 and 5.898 mg/dL. The p-value of <0.0001 indicates that we can reject the null hypothesis that there is not a linear association between serum CRP and serum fibrinogen levels.

**d)** See Table 1

**Question 4**:

**Method**: I first converted all 0 measurements of CRP to 0.5 mg/L, which is half the lowest nonzero measurement (i.e. 1 mg/L) and then log-transformed CRP measurements. I used the log transformed CRP as the predictor of interest in a linear regression model looking at the mean of serum fibrinogen levels as the response variable. Huber-White sandwich estimators were used in the regression analysis.

**Answer:**

**a)** The estimated intercept from the fitted regression model indicates that we estimate subjects with a serum CRP of 0 mg/L to have a mean serum fibrinogen level of 295.6 mg/dL (95% CI 293.6 – 297.5 mg/dL).

**b)** The estimated slope of the model suggests that as serum CRP levels doubles, we estimate the mean serum fibrinogenlevel to increase by 25.55 mg/dL (95% CI 23.97 – 27.09 mg/dL).

**c)** Based on this linear regression model, we estimate that when the serum CRP level doubles, the mean serum fibrinogen level increases by 25.55 mg/dL. The 95% confidence intervals indicate that we would not be surprised if the true difference of mean serum fibrinogen increased by between 23.97 and 27.09 mg/dL as the serum CRP doubles. The p-value of <0.0001 indicates that we can reject the null hypothesis that there is not a linear association between log-transformed serum CRP and mean serum fibrinogen levels.

**d)** See Table 1

**Question 5**:

**Method**: I used non-transformed CRP as the predictor of interest in a linear regression model looking at the geometric mean of serum fibrinogen levels as the response variable. Huber-White sandwich estimators were used in the regression analysis.

**Answer:**

**a)** The estimated intercept from the fitted regression model indicates that we estimate subjects with a serum CRP of 0 mg/L to have a geometric mean serum fibrinogen level of 300.9 mg/dL (95% CI 298.6 – 303.2 mg/dL).

**b)** The estimated slope of the model suggests that as serum CRP levels increase by each 1 mg/L, we estimate the geometric mean serum fibrinogenlevel will be 1.4% higher, with a 95% confidence interval indicating that we would not be surprised by this result if subjects with CRP 1 mg/L higher had a true geometric mean of serum fibrinogen 1.2 to 1.6% higher.

**c)** Based on this regression analysis, we estimate that when the serum CRP level is 0 mg/L the mean serum fibrinogen level is 300.9 mg/dL and, as the serum CRP level increases by 1 mg/L, the geometric mean serum fibrinogen level will be 1.4% higher. The 95% confidence intervals indicate that we would not be surprised by these results if the true geometric mean fibrinogen level at a CRP of 0 mg/L were between 298.6 and 303.2 mg/dL and if the true geometric mean serum fibrinogen were between 1.2 and 1.6% higher in groups with serum CRP measurements 1 mg/L higher. The p-value of <0.0001 indicates that we can reject the null hypothesis that there is not a linear association between serum CRP and the geometric mean of serum fibrinogen levels.

**d)** See Table 1

**Question 6**:

**Method**: I first converted all 0 measurements of CRP to 0.5 mg/L, which is half the lowest nonzero measurement (i.e. 1 mg/L) and then log-transformed CRP measurements. I used the log transformed CRP as the predictor of interest in a linear regression model looking at the geometric mean of serum fibrinogen levels as the response variable. Huber-White sandwich estimators were used in the regression analysis.

**Answer:**

**a)** The estimated intercept from the fitted regression model indicates that we estimate subjects with a serum CRP of 0 mg/L to have a geometric mean serum fibrinogen level of 292.5 mg/dL (95% CI 290.7 – 294.4 mg/dL).

**b)** The estimated slope of the model suggests that as serum CRP levels doubles, we estimate the geometric mean serum fibrinogenlevel to be 7.6% higher, with the 95% confidence interval indicating that this result would not be unusual if truly the geometric mean was between 7.1 and 8.0% higher as the serum CRP doubled.

**c)** Based on this regression analysis, we estimate that when the serum CRP is 0 mg/L, the geometric mean serum fibrinogen will be 292.5 mg/dL and as the CRP level doubles, the geometric mean serum fibrinogen level will be 7.6% higher. The 95% confidence intervals indicate that these results would not be unusual if truly the geometric mean of serum fibrinogen was between 290.7 and 294.4 mg/dL when serum CRP is 0 and as CRP doubles, the geometric mean fibrinogen was between 7.1 and 8.0% higher. The p-value of <0.0001 indicates that we can reject the null hypothesis that there is not a linear association between log-transformed serum CRP and the geometric mean od serum fibrinogen levels.

**d)** See Table 1

**Table 1**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **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** | 309.266 | 295.566 | 305.113 | 292.536 |
| **2 mg/L** | 314.517 | 313.263 | 309.390 | 307.730 |
| **3 mg/L** | 319.768 | 323.615 | 313.726 | 316.981 |
| **4 mg/L** | 325.019 | 330.959 | 318.123 | 323.713 |
| **6 mg/L** | 335.520 | 341.311 | 327.103 | 333.445 |
| **8 mg/L** | 346.022 | 348.656 | 336.337 | 340.527 |
| **9 mg/L** | 351.273 | 351.663 | 341.051 | 343.469 |
| **12 mg/L** | 367.025 | 359.008 | 355.593 | 350.764 |

**Table 2**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **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** | 5.251 | 17.697 | 4.276 | 15.194 |
| **3 mg/L – 2 mg/L** | 5.251 | 10.352 | 4.336 | 9.251 |
| **4 mg/L – 1 mg/L** | 15.753 | 35.393 | 13.010 | 31.178 |
| **4 mg/L – 2 mg/L** | 10.502 | 17.697 | 8.734 | 15.983 |
| **6 mg/L – 3 mg/L** | 15.753 | 17.697 | 13.377 | 16.464 |
| **8 mg/L – 4 mg/L** | 21.003 | 17.697 | 18.214 | 16.814 |
| **9 mg/L – 6 mg/L** | 15.753 | 10.352 | 13.948 | 10.024 |
| **9 mg/L – 8 mg/L** | 5.251 | 3.007 | 4.714 | 2.943 |
| **12 mg/L – 6 mg/L** | 31.505 | 17.697 | 28.490 | 17.319 |
| ***Ratios*** | | | | |
| **2 mg/L / 1 mg/L** | 1.017 | 1.060 | 1.014 | 1.052 |
| **3 mg/L / 2 mg/L** | 1.017 | 1.033 | 1.014 | 1.030 |
| **4 mg/L / 1 mg/L** | 1.051 | 1.120 | 1.043 | 1.107 |
| **4 mg/L / 2 mg/L** | 1.033 | 1.056 | 1.028 | 1.052 |
| **6 mg/L / 3 mg/L** | 1.049 | 1.055 | 1.043 | 1.052 |
| **8 mg/L / 4 mg/L** | 1.065 | 1.053 | 1.057 | 1.052 |
| **9 mg/L / 6 mg/L** | 1.047 | 1.030 | 1.043 | 1.030 |
| **9 mg/L / 8 mg/L** | 1.015 | 1.009 | 1.014 | 1.009 |
| **12 mg/L / 6 mg/L** | 1.094 | 1.052 | 1.087 | 1.052 |

**Question 8:**

**a)** The only analysis that gave constant differences in the fitted values of this type was that in problem 3. Examples:

x + c; c = 1; (2-1, 3-2, 9-8); difference = 5.251;

x + c; c = 3; (4-1, 6-3, 9-6); difference = 15.753;

**b)** The only analysis that gave constant ratios in the fitted values with an absolute increase of c units was that in Problem 5. Examples:

x + c; c = 1; (2/1, 3/2, 9/8); ratio = 1.014;

x + c; c = 3; (4/1, 6/3, 9/6); ratio = 1.043;

**c)** The only analysis that gave constant differences in the fitted values when comparing two groups with a relative c-fold increase was that in Problem 4. Examples:

x \* c; c = 2; (2-1, 4-2, 6-3, 8-4,12-6); difference = 17.697;

x \* c; c = 1.5; (3-2, 9-6); difference = 10.352;

**d)** The only analysis that gave constant ratio when comparing two groups that were c-fold different from each other was that in Problem 6. Examples:

x \* c; c = 2; (2/1, 4/2, 6/3, 8/4,12/6); ratio = 1.052;

x \* c; c = 1.5; (3/2, 9/6); ratio = 1.030;

**Question 9:**

I would decide which analysis to do based primarily on the scientific question and the scientific understanding of what a possible association might be. Since we are looking at two markers of inflammation and different markers of inflammation most likely scale with the immune response (i.e. a multiplicative effect), I think it would make sense to look at the geometric mean and the log-transformed data. I would expect that if the inflammatory response was “twice as large” (whatever that means) some markers of inflammation would be some percentage larger (where 2 times some unknown factor for each marker = “some percentage”), not some units larger.