**Biost 518: Applied Biostatistics II**

**Biost 515: Biostatistics II**

Emerson, Winter 2015

**Homework #2**

January 13, 2015

1. Provide a suitable descriptive statistical analysis for the association between CRP and FIB both overall (a), and separately for groups having no prior history of diagnosed cardiovascular disease or having prior diagnosed CVD (b).
2. Overall CRP/FIB analysis

Methods: The mean value of fibrinogen was compared across participants with low (≤3mg/L) CRP and those with high (>3mg/L) CRP. 101 participants were excluded from the analysis since they were missing data on CRP, FIB, or both. A two-sided t-test allowing for unequal variances was used to analyze differences in mean fibrinogen between the two groups. 95% Confidence intervals were calculated also allowing for unequal variances.

Results: The mean value of fibrinogen given low CRP (n=3732) was 308 mg/dL and the mean value of fibrinogen given high CRP (n=1167) was 373 mg/dL. The difference in mean fibrinogen between the two groups was 65.2 mg/dL, however it would not be unusual if the true value were between 60.2 and 70.2 mg/dL. The t-test (two-sided p<0.0001) allowing for unequal variance showed that this observation is statistically significant at the 0.05 level. Therefore, we can reject the null hypothesis that mean fibrinogen does not differ between groups with high and low CRP. This suggests that mean fibrinogen is higher given high CRP.

1. Separate CRP/FIB analysis, split on prior CVD
2. No history of CHD

Methods: The mean value of fibrinogen was compared across participants with low (≤3mg/L) CRP and those with high (>3mg/L) CRP who did not have a prior history of CVD. 101 participants were excluded from the analysis since they were missing data on CRP, FIB, or both (no participants were missing data on CVD). A two-sided t-test allowing for unequal variances was used to analyze differences in mean fibrinogen between the two groups. 95% Confidence intervals were calculated also allowing for unequal variances.

Results: The mean value of fibrinogen with no history of CVD given low CRP (n=1687) was 305 mg/dL and given high CRP (n=471) was 365 mg/dL. The difference in mean fibrinogen between the two groups was 59.5 mg/dL, however it would not be unusual if the true value were between 52.5 and 66.4 mg/dL. The t-test (two-sided p<0.0001) allowing for unequal variance showed that this observation is statistically significant at the 0.05 level. Therefore, we can reject the null hypothesis that mean fibrinogen does not differ between groups with high and low CRP with no history of CVD. This suggests that mean fibrinogen is higher given high CRP and no history of CVD.

1. History of CHD

Methods: The mean value of fibrinogen was compared across participants with low (≤3mg/L) CRP and those with high (>3mg/L) CRP who had a prior history of CVD. 101 participants were excluded from the analysis since they were missing data on CRP, FIB, or both (no participants were missing data on CVD). A two-sided t-test allowing for unequal variances was used to analyze differences in mean fibrinogen between the two groups. 95% Confidence intervals were calculated also allowing for unequal variances.

Results: The mean value of fibrinogen with a history of CVD given low CRP (n=448) was 310 mg/dL and given high CRP (n=170) was 390 mg/dL. The difference in mean fibrinogen between the two groups was 79.4 mg/dL, however it would not be unusual if the true value were between 64.6 and 94.2 mg/dL. The t-test (two-sided p<0.0001) allowing for unequal variance showed that this observation is statistically significant at the 0.05 level. Therefore, we can reject the null hypothesis that mean fibrinogen does not differ between groups with high and low CRP with a history of CVD. This suggests that mean fibrinogen is higher given high CRP and a history of CVD.

iii) Summary

The difference in mean fibrinogen between low CRP and high CRP groups for those with a history of CVD is 19.9 mg/dL larger than for those without a history of CVD. This increase is almost statistically significant, however, the 95% confidence intervals overlap slightly.

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 mean value of fibrinogen was compared among those participants with a prior history of CVD and those participants without a prior history of CVD. 85 participants were excluded from the analysis since they were missing data on CVD, FIB, or both. A two-sided t-test assuming equal variances was used to analyze differences in the means. 95% Confidence intervals were calculated also assuming equal variances.

Results: The mean value of fibrinogen for those participants with a history of CVD (n=1122) was 334 mg/dL and the mean value of fibrinogen for those participants without a history CVD (n=3777) was 320 mg/dL. The difference in mean fibrinogen between participants who did and participants who did not have a prior history of CVD was 14.9 mg/dL, however it would not be unusual if the true value were between 10.4 and 19.4 mg/dL. The t-test (two-sided p<0.0001) assuming equal variance showed that this observation is statistically significant at the 0.05 level. Therefore, we can reject the null hypothesis that there is no difference in mean fibrinogen between those with and those without a prior history of CVD. This suggests that those with a prior history of CVD have a higher mean value of fibrinogen.

Note: The p-value spit out by R was 6.73\*10^-11. While a p-value this small is absurdly inaccurate it is useful to compare to the p-value that R provides for part B, classical linear regression. Also for comparison, the t-statistic for this analysis was -6.54.

* 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.

The analysis in part A (a t-test that assumes equal variances) could have been performed with classical linear regression. Classical linear regression is exactly the t-test that assumes equal variance. When we run a classical linear regression on the data we find an intercept of 320 mg/dL which is exactly the same as the mean value of fibrinogen for those participants without a history of CVD. The slope, 14.9, is the same as the difference in mean fibrinogen between the two groups. The t-value for those without a history of CVD, 6.54, is the same as the t-value found in part A. The p-value for those without a history of CVD, 6.73\*10^-11, is the same as the p-value found in part A. The standard errors are slightly different which leads to different confidence intervals; the SEs and CIs for the regression are slightly larger.

|  |  |  |
| --- | --- | --- |
| **Statistic** | **t-test Value** | **Regression Value** |
| Intercept | 319.574 mg/dL(mean fibrinogen with no history of CVD) | 319.574 mg/dL |
| Slope | 14.885 (difference in mean fibrinogen for the two groups) | 14.885 |
| t-value | 6.54 | 6.54 (no history of CVD) |
| p-value | 6.73\*10^-11 | 6.73\*10^-11 (no history of CVD) |
| Degrees of freedom | 4913 | 4913 |
| Standard error (history of CVD) | 2.21 | 2.276 |
| Standard error (no history of CVD) | 1.05 | 1.088 |
| 95% Confidence intervals (history of CVD) | 317.5, 321.6 | 317.441, 321.707 |
| 95% Confidence intervals (difference) | 10.4, 19.3 | 10.424, 19.346 |

* 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 mean value of fibrinogen was compared among those participants with a prior history of CVD and those participants without a prior history of CVD. 85 participants were excluded from the analysis since they were missing data on CVD, FIB, or both. A two-sided t-test allowing for unequal variances was used to analyze differences in the means. 95% Confidence intervals were calculated also allowing for unequal variances.

Results: The mean value of fibrinogen for those participants with a history of CVD (n=1122) was 334 mg/dL and the mean value of fibrinogen for those participants without a history CVD (n=3777) was 320 mg/dL. The difference in mean fibrinogen between participants who did and participants who did not have a prior history of CVD was 14.9 mg/dL, however it would not be unusual if the true value were between 12.1 and 19.7 mg/dL. The t-test (two-sided p<0.0001) allowing for unequal variance showed that this observation is statistically significant at the 0.05 level. Therefore, we can reject the null hypothesis that there is no difference in mean fibrinogen between those with and those without a prior history of CVD. This suggests that those with a prior history of CVD have a higher mean value of fibrinogen.

Note: The p-value spit out by R was 1.46\*10^-9. While a p-value this small is absurdly inaccurate it is useful to compare to the p-value that R provides for part D, robust linear regression. Also for comparison, the t-statistic for this analysis was -6.08.

* 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.

The analysis in part C (a t-test that allows for unequal variances) could have been performed with robust linear regression. Robust linear regression is similar to the t-test that allows for unequal variance. When we run a robust linear regression on the data we find an intercept of 320 mg/dL which is exactly the same as the mean value of fibrinogen for those participants without a history of CVD. The slope, 14.9, is the same as the difference in mean fibrinogen between the two groups. The t-value for those without a history of CVD, 6.0847, is close to but not exactly the same as the absolute value of the t-value found in part C, 6.0836. The standard errors are also slightly different with the regression values larger than the t-test values. The degrees of freedom are also greater for the regression model. This leads to slightly smaller 95% confidence intervals for the regression value and a larger p-value.

|  |  |  |
| --- | --- | --- |
| **Statistic** | **t-test Value** | **Regression Value** |
| Intercept | 319.574 mg/dL (mean fibrinogen with no history of CVD) | 319.574 mg/dL |
| Slope | 14.885 (difference in mean fibrinogen for the two groups) | 14.885 |
| t-value | -6.0836 | 6.0847 (no history of CVD) |
| p-value | 1.455\*10^-9 | 1.26\*10^-9 (no history of CVD) |
| Degrees of freedom | 1665 | 4913 |
| Standard error (history of CVD) | 2.21 | 2.276 |
| Standard error (no history of CVD) | 1.05 | 1.088 |
| 95% Confidence intervals (history of CVD) | 317.5, 321.6 | 317.51, 321.636 |
| 95% Confidence intervals (difference) | 10.086, 19.684 | 10.089, 19.681 |

* 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)?

We know that the confidence intervals produced assuming equal variance will be either equally or less conservative than the CIs produced allowing for unequal variance. This means that the p-values for tests allowing for unequal variance will be larger and the t-statistics smaller in magnitude, leading to a weaker association for tests allowing for unequal variance compared to those that assume equal variance.

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.

The intercept (304 mg/dL) is the estimated fibrinogen level in somebody with 0 mg/L CRP.

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

The slope (5.25) is the change in fibrinogen across a 1mg/L difference in CRP.

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

Methods: The relationship between mean fibrinogen across CRP levels was examined using robust linear regression. 101 participants were excluded from the analysis since they were missing data on CRP, FIB, or both. 95% Confidence intervals were calculated using robust standard errors.

Results: For every 1 mg/L increase in CRP a 5.25 mg/dL increase in fibrinogen is estimated. 95% confidence intervals suggest that it would not be unusual if the true value were between 4.60 and 5.90 mg/dL for every 1 mg/L increase in CRP. A two-sided p value (p<0.0001) shows that this observation is statistically significant at the 0.05 level. Therefore, we can reject the null hypothesis that there is no linear trend between fibrinogen and CRP.

* 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).

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

The intercept (296 mg/dL) is the estimated fibrinogen level in somebody with 0 mg/L log CRP.

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

The slope (36.8) is the change in fibrinogen across a 1mg/L difference in log CRP, which is a ~2.72 (e) fold increase in CRP.

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

Methods: The relationship between mean fibrinogen and log CRP was examined using robust linear regression. 101 participants were excluded from the analysis since they were missing data on CRP, FIB, or both. CRP values that were 0 were set to 0.5. 95% Confidence intervals were calculated using robust standard errors.

Results: For every 1 mg/L increase in log CRP a 36.8 mg/dL increase in fibrinogen is estimated. 95% confidence intervals suggest that it would not be unusual if the true value were between 34.6 and 39.1 mg/dL for every 1 mg/L increase in log CRP. A two-sided p value (p<0.0001) shows that this observation is statistically significant at the 0.05 level. Therefore, we can reject the null hypothesis that there is no linear trend between fibrinogen and log CRP.

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

The intercept (5.71 mg/dL) is the estimated log fibrinogen level in somebody with 0 mg/L CRP.

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

The slope (0.0139) is the change in log fibrinogen across a 1mg/L difference in CRP. This is not easily deconvolved to give the change in fibrinogen per 1mg/L difference in CRP. However it can be said, thinking about log fibrinogen as the geometric mean, that the ratio of the geometric mean fibrinogen between two groups that differ by 1 mg/L CRP is e^0.0139.

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

Methods: The relationship between geometric mean fibrinogen (log fibrinogen) and CRP was examined using robust linear regression. 101 participants were excluded from the analysis since they were missing data on CRP, FIB, or both. 95% Confidence intervals were calculated using robust standard errors.

Results: For every 1 mg/L increase in CRP a 0.0139 mg/dL increase in log fibrinogen is estimated. 95% confidence intervals suggest that it would not be unusual if the true value were between 0.0122 and 0.0157 mg/dL for every 1 mg/L increase in CRP. A two-sided p value (p<0.0001) shows that this observation is statistically significant at the 0.05 level. Therefore, we can reject the null hypothesis that there is no linear trend between log fibrinogen and CRP.

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

The intercept (5.68 mg/dL) is the estimated log fibrinogen level in somebody with 0 mg/L log CRP.

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

The slope (0.105) is the change in log fibrinogen across a 1mg/L difference in log CRP. It can also be said that the geometric mean ratio of fibrinogen is *x*^0.105 in two populations of which one has *x* times higher CRP than the other.

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

Methods: The relationship between geometric mean fibrinogen (log fibrinogen) and log CRP was examined using robust linear regression. 101 participants were excluded from the analysis since they were missing data on CRP, FIB, or both. 95% Confidence intervals were calculated using robust standard errors.

Results: For every 1 mg/L increase in log CRP a 0.105 mg/dL increase in log fibrinogen is estimated. 95% confidence intervals suggest that it would not be unusual if the true value were between 0.0995 and 0.111 mg/dL for every 1 mg/L increase in log CRP. A two-sided p value (p<0.0001) shows that this observation is statistically significant at the 0.05 level. Therefore, we can reject the null hypothesis that there is no linear trend between log fibrinogen and log CRP.

**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 CRP)** | **Problem 4: (Mean of log(CRP))** | **Problem 5: (geometric mean)** | **Problem 6: (GM of log(CRP))** |
| **1 mg/L** | 309 | 296 | 305 | 293 |
| **2 mg/L** | 613 | 321 | 309 | 315 |
| **3 mg/L** | 917 | 336 | 314 | 328 |
| **4 mg/L** | 1221 | 347 | 318 | 339 |
| **6 mg/L** | 1829 | 362 | 327 | 353 |
| **8 mg/L** | 2437 | 372 | 336 | 364 |
| **9 mg/L** | 2741 | 376 | 341 | 369 |
| **12 mg/L** | 3653 | 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: (Difference in Mean CRP)** | **Problem 4: (Difference in Mean log(CRP))** | **Problem 5: (Difference in GM of CRP)** | **Problem 6: (Difference in GM of log(CRP))** |
| ***Differences*** |
| **2 mg/L – 1 mg/L** | 304 | 25.5 | 4.28 | 22.2 |
| **3 mg/L – 2 mg/L** | 304 | 14.9 | 4.34 | 13.7 |
| **4 mg/L – 1 mg/L** | 912 | 51.1 | 13.0 | 46.0 |
| **4 mg/L – 2 mg/L** | 608 | 25.5 | 8.73 | 23.9 |
| **6 mg/L – 3 mg/L** | 912 | 25.5 | 13.4 | 24.9 |
| **8 mg/L – 4 mg/L** | 1216 | 25.5 | 18.2 | 25.7 |
| **9 mg/L – 6 mg/L** | 912 | 14.9 | 13.9 | 15.4 |
| **9 mg/L – 8 mg/L** | 304 | 4.34 | 4.71 | 4.55 |
| **12 mg/L – 6 mg/L** | 1824 | 25.5 | 28.5 | 26.8 |
| ***Ratios*** |
| **2 mg/L / 1 mg/L** | 1.98 | 1.09 | 1.01 | 1.08 |
| **3 mg/L / 2 mg/L** | 1.50 | 1.05 | 1.01 | 1.04 |
| **4 mg/L / 1 mg/L** | 3.95 | 1.17 | 1.04 | 1.16 |
| **4 mg/L / 2 mg/L** | 1.99 | 1.08 | 1.03 | 1.08 |
| **6 mg/L / 3 mg/L** | 1.99 | 1.08 | 1.04 | 1.08 |
| **8 mg/L / 4 mg/L** | 2.00 | 1.07 | 1.06 | 1.08 |
| **9 mg/L / 6 mg/L** | 1.50 | 1.04 | 1.04 | 1.04 |
| **9 mg/L / 8 mg/L** | 1.13 | 1.01 | 1.01 | 1.01 |
| **12 mg/L / 6 mg/L** | 2.00 | 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 basic regression (#3): 2mg/L – 1mg/L = 3mg/L - 2mg/L = 304mg/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..

The regression with a log transformed response (#5): 2mg/L / 1mg/L = 3mg/L / 2mg/L = 1.01

* 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.

The regression with a log transformed predictor (#4): 12mg/L - 6mg/L = 6mg/L - 3mg/L = 8mg/L - 4mg/L = 25.5mg/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.

The regression with a log transformed response and log transformed predictor (#6): 12mg/L / 6mg/L = 6mg/L / 3mg/L = 8mg/L / 4mg/L = 1.08

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

Before looking at the data I would consider the relationship between fibrinogen and CRP using what is already known about the relationship between the two substances. If we think about what they represent biologically it makes sense to use a log scale. Both CRP and fibrinogen are inflammatory markers and when disease processes occur in the body they often exert their effects in a multiplicative manner. Dealing with multiplicative effects is easier on the log scale. We can also take a look at the log vs. untransformed graphs.

The log fibrinogen/fibrinogen graph is relatively straight. This means that using the log value for fibrinogen will have only a moderate effect. The log CRP/CRP graph is quite curved. This means that using the log value for CRP will have a strong effect.

Therefore I would choose to use either the fibrinogen/log CRP analysis or the log fibrinogen/log CRP analysis.