**Biost 518: Applied Biostatistics II**
**Biost 515: Biostatistics II**
Emerson, Winter 2015

**Homework #2 Key**
January 13, 2015

***Instructions for grading*: Prior to the answer for each problem, I provide the maximum points to be given for each problem, and the way that points should be distributed. Please insert comments on to the document indicating the points you have awarded for the problem, commenting on any reasons points were deducted.**
***My answer to each question is provided in boldface type. In giving the answers, I sometimes provide alternative approaches in order that you can assess whether the numbers match up. I also provide some discussion of the choices or some additional material that I did not really expect to be provided in the answer. This additional information is provided in** normal type.*

> *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.

***Instructions for grading*: This problem is worth 15 points. There must be suitable descriptive statistics and a discussion of what they would mean for effect modification and confounding. Below I give a rather lengthy discussion of the thinking that should go into the selection of appropriate descriptive statistics for this problem. You are urged to consider the extent to which the student's presentation took such issues into account. In any case, you are responsible for knowing this material.**

- ***For this homework's question and data, I think it highly preferable that both a graph (5 points) and table (5 points) are presented, because the density of the data makes the plots only partially***

***useful. However, if the student presents only the table or only the plot deduct 3 points and assign the other 2 points proportionately to whichever modality (table or graph) they did present.*** *(If only a graph were presented, we have little ability to assess the variability of the measurements and the separation of the strata defined by prior CVD. But the graph does allow us to quickly assess any curvilinearity in the association between fibrinogen and CRP When it comes to publishing results, we might not be able to get both the graph and the table published, but often we can..)*

- ***Points to consider are how well the student addressed the goals of descriptive statistics. In this problem, we want to present statistics relating to the investigation of the association between fibrinogen and CRP both with and without prior CVD. There are multiple goals in providing these descriptive statistics, and we will want to find the combination of statistics that can succinctly address all of (or nearly all of) these goals.***
    - ***A portion of our goal will be to "describe the materials and methods". Our ability to find associations will depend in part on the patients included in the analysis, so we would want to know the range, central tendency, and variability of both the CRP and fibrinogen measurements, both overall and within groups defined by CVD. Clearly ability to communicate our impressions of what we found in the sample is important here.*** *(Note that in a randomized trial, having descriptive statistics in the overall sample is not as necessary, because the distribution should be the same in each treatment arm. But in this observational study, the distributions can be quite different by CVD status, so we want to see each group separately, as well as combined.)*
    - ***A portion of our goal will be to "assess the validity of assumptions". To the extent that most of the questions are asking about unadjusted associations between fibrinogen and CRP, we are in some sense relying on the absence of effect modification and / or confounding. To a lesser extent, we might want to assess "linearity" of the association and patterns of variability (e.g., any tendency toward heteroscedasticity is of interest in linear regression). The descriptive statistics appropriate for this goal will depend heavily on how you would analyze the data:***
        - ***For potential effect modification, we will just want to assess the association of interest within subgroups defined by the third variable.***
        - ***For potential confounding we want to be able to judge two things:***
            - ***We want to show evidence related to an association between the potential confounder and the response variable within groups that are homogeneous with respect to the predictor of interest. That evidence will be most relevant if it is the association is assessed on the same scale as used in our primary analysis.*** *(We are only interested in such associations that are truly causal, but we will have to just go with what we believe to be causal.)*
            - ***We want to show evidence related to an association in the sample between the potential confounder and the predictor of interest. That evidence will be most relevant if it reflects the difference in the mean of the covariate that would be used in adjustment. Hence, if you would adjust for a variable using a continuous untransformed variable, it is the sample mean that we care the most about. If you would have adjusted for logarithmically transformed variables, it is truly the sample geometric means we would care the most about.***

- ***In assessing any departures from linearity we want to examine how the summary measure used in the primary analysis (e.g., mean or geometric mean) varies as the predictor of interest varies.***
- ***In assessing heteroscedasticity we want to examine how the standard deviation of the modeled response variable differs as the predictor of interest varies.***

- ***The major goal of the descriptive statistics will be to "provide preliminary straightforward estimates in support of the association".***
  - ***In some cases, a graph will be able to present useful description. This will very often be the case when we are using linear regression analysis with continuous response and continuous predictor of interest.***
  - ***In all cases, we can use stratified descriptive statistics for the response variable, where the strata are based on categorizations of the predictor of interest.***
    - ***This will be most relevant if you provide descriptive statistics based on the summary measure (e.g., mean or geometric mean) that is used in the primary analysis.***
    - ***We will also want to provide this information within groups that would be judged most homogenous, in order that we could compute the relative contrasts (either differences or ratios). For discrete predictors of interest, this is simple. But for continuous covariates (such as the CRP we have here), we will have to categorize the measurements in order to provide the most similar statistics.***
      - ***So, if we will model our association using an untransformed continuous variable, categorizations should be based as much as possible on equal length intervals.***
      - ***On the other hand, if logarithmic transformations of the predictor of interest would be used in the primary analysis, we should categorize on a logarithmic scale (e.g., based on a doubling or 10-fold of covariate values.)***

- ***Based on the above discussion, we clearly have to consider both what we anticipated to be the best primary analysis of the association (this should be pre-specified to avoid the multiple comparison issue) and what we learned when we analyzed the data (we want to communicate our impressions of the sample to the reader).***
  - ***In terms of what we would have anticipated, we reflect on what is known about the measures of inflammation:***
    - ***Fibrinogen is a normal constituent of blood measured in mg/dL. It is a pre-cursor protein to fibrin, and fibrin is a key protein in forming blood clots. Because our body is constantly subjected to small bleeds, we need a constant supply of fibrinogen. In healthy people, this will generally be tightly controlled. However, in the presence of inflammation it will be elevated somewhat beyond those normal limits, but not necessarily at a level that will greatly overwhelm the levels associated with normal "homeostasis" (the body's ability to control its balance of chemicals and biochemical reactions).***
    - ***C reactive protein is synthesized in the liver in response to stimulation of the immune system. It is typically at very low levels, except in the presence of inflammation. It is generally measured in units of mg/L, and its typical range***

***below 3 mg/L argues that it's presence is several orders of magnitude less than fibrinogen (which averages around 3,000 mg/L). Because there is no real "background level" for CRP, we might expect its distribution to be markedly elevated (in a multiplicative fashion) with inflammation. (And, indeed, this has been observed.)***

- ***The end result of the above observations is that we might expect multiplicative differences (e.g., every doubling) of CRP to be reasonable surrogates for level of subclinical inflammation, but levels of fibrinogen will not be as greatly affected by subclinical inflammation, as the need for "normal levels" in blood clotting will predominate.***
- ***Thus a priori, I would be very interested in comparing patients having constant ratios of CRP. In that setting, I will anticipate analyses based on a logarithmic transformation.***
- ***A priori, it will not probably matter as much whether we look at absolute differences in mean fibrinogen levels or relative ratios of geometric mean fibrinogen, because the log function will likely be approximately linear over the range of fibrinogen. Relative advantages include***
    - ***For descriptive purposes, most people are more familiar with the behavior of means as a measure of central tendency.***
    - ***For inferential purposes, because we are wanting to regard how these two measures of inflammation might behave similarly, putting them both on the multiplicative scale might seem attractive.***
- ***After looking at the data, we would certainly notice that CRP is markedly skewed (as we anticipated). So we do want to let that observation shine through in our presentation of descriptive statistics***
    - ***We can consider histograms or boxplots, but those would contribute little information for the other purposes of descriptive statistics: Histograms can depict outliers well, but really do not do a very good job at means or SD. Boxplots (as usually implemented) focus on a too liberal definition of "outliers" (if you ask me), and provide medians and interquartile ranges instead of the means / geometric means and standard deviations that we would really want***
    - ***Scatterplots can give ideas of the distribution of the response and the POI, as well as the trends in the response as the POI differs across groups, especially if smoothed curves are superimposed. But the scatterplots can be too dense to assess distributions, and they are also not really showing the standard deviations: Instead we tend to use the range of measurements.***
        - ***In any case, when we use scatterplots, we do have the capability to consider using log scales on either the y or x axis, or both. We should make such choices when we anticipate multiplicative effects more than additive effects.***
    - ***Standard descriptive statistics can go a long way to highlighting skewed data. With high numbers of "outliers" in the distribution, we will typically observe***

- ***A mean that is not in the center of the range of the data.***
- ***A mean that is not in the center of the interquartile range.***
- ***A mean that is not similar to the median,***
- ***For positive measurements, a standard deviation that is large relative to the mean. (A standard rule of thumb that I use is that I start to worry about outliers when the standard deviation of a positive random variable is more than about half the mean.)***

- ***Bottom line (for me) from all of this:***
  - ***I present a scatterplot stratified by prior CVD, but I present the plot on a log-log scale.***
  - ***I present univariate statistics for both CRP and fibrinogen in the entire population, highlighting the patterns of missing data. I choose mean, SD, minimum and maximum, because such allow me to demonstrate the marked skewness of CRP while sticking with the more broadly understood mean.***
  - ***I present univariate statistics for fibrinogen within strata defined by CRP.***
  - ***When stratifying CRP, I use intervals corresponding roughly to a doubling of the measurements, rather than constant width intervals.***
- ***Assign 5 points to a scatterplot. Criteria to examine the labeling of the graph, axes (including units of measurement), labeling of points by strata, and display of lowess curves.***
- ***Assign 5 points to the table. Criteria to examine are the layout and labeling of columns, rows, and descriptive statistics; choice of descriptive statistics, and categorization of CRP within the table. Best would be to use categories that are logarithmically based (I used approximate doubling of the CRP levels). Students do not have to have presented things exactly as I did, but you should consider whether you can obtain information about the points I raised above.***
  - ***In particular consider how well you could ascertain patterns of sample sizes and missing data across strata.***
- ***Assign 5 points to the discussion of the findings. Criteria to examine are to comment on the univariate distributions of the data, the first order (linear) trends in the association (on which scale- additive or multiplicative), possibility of effect modification by prior CVD, and possible confounding by prior CVD.***

**Ans:** ***Methods:*** **Summary statistics (mean, standard deviation, minimum, and maximum) are presented for C-reactive protein (CRP) and fibrinogen in strata defined by prior history of cardiovascular disease (CVD) and in the combined sample. Descriptive statistics for fibrinogen are further detailed within strata defined by CRP level, with intervals chosen to represent approximate doubling of CRP levels. A separate stratum is included for those subjects missing CRP data but having fibrinogen measurements. Also presented is a scatterplot of fibrinogen level by CRP level, with status of prior CVD marked by color. Observations were jittered slightly in order to better display tied observations. Superimposed on the scatterplot are smoothed lowess curves for each CVD stratum and for the combined population. The plot is displayed on a log-log scale, with subjects having CRP reported at 0 plotted as if they were truly one-half the lowest positive measurement of 1 mg/L. Subjects missing data were omitted only if they were missing data for a variable needed for the specific analysis.**

***Results:*** **Of the 5,000 subjects in the dataset, CRP measurements are available on 3,802 of the 3,851 subjects without prior history of CVD, and available on 1,131 of the 1,149 subjects with prior**

**history of CVD. As depicted in Table 0 below, there is a trend toward higher average CRP in patients with prior CVD (4.40 mg/L vs 3.38 mg/L) with marked skewness of the measurements: approximately 62% of CRP measurements are less than or equal to 2 mg/L, while the range of measurements is from 0 – 108 mg/L. This skewness is further evidenced by the high standard deviation (on the order of 6 mg/L) relative to the sample mean (3.6 mg/L) that was observed in these positive measurements.**

**Fibrinogen measurements are subject to only slightly higher rates of missingness than CRP, with measurements available on 3,791 of the 3,851 subjects without prior CVD and 1,124 of the 1,149 subjects with prior CVD. As with CRP, there is a trend toward higher average fibrinogen in patients with prior CVD (334 mg/dL vs 320 mg/dL), with the measurements ranging between a minimum of 109 mg/dL and 872 mg/dL without marked skewness.**

**Also depicted in Table 0 are descriptive statistics for fibrinogen levels within strata defined by CRP level. Within the CVD categories, there was a clear trend toward increasing average fibrinogen with higher levels of CRP, with a suggestion of trends toward greater differences between successive CRP strata with higher CRP. Because the CRP strata were defined based on approximate doubling of CRP, this is suggestive that the effect of a proportionate difference in CRP might similarly be associated with a proportionate difference in fibrinogen. In each CRP stratum, the subjects with prior CVD had higher sample average fibrinogen, though the differences of those within stratum means ranged from approximately 3 mg/dL to 24 mg/dL. There was perhaps a slight trend toward larger differences at higher levels of CRP.**

**Figure 1 is a graphical display of fibrinogen versus CRP (on a log-log scale) in the subjects having both fibrinogen and CRP measurements available. For the purposes of this plot measurements of CRP equal to 0 mg/L were recoded as 0.5 mg/L (the lowest nonzero measurement of CRP was 1 mg/L). Clearly evident from the smooth curves displayed on this plot is a trend toward higher fibrinogen measurements in groups having higher CRP measurements. Furthermore, the smooth curves show similar, fairly linear trends for both CVD strata with little suggestion of marked vertical separation between the curves. Hence, from this log-log plot it would appear that constant proportionate diffierences in CRP levels will be associated with tendencies toward constant proportionate differences in fibrinogen levels.**

**Table 0: Descriptive statistics for CRP and fibrinogen with strata defined by prior history of CVD. Fibrinogen levels are further described within strata defined by approximate doubling of CRP levels. Statistics presented are the sample mean (sample SD; minimum – maximum; number of observations with available measurements / number of subjects in the stratum).**

| | **Prior History of Cardiovascular Disease (CVD)** | | |
|---|---|---|---|
| | **No Prior CVD** | **Prior CVD** | **All Subjects** |
| | | | |
| **C reactive protein (mg/L)** | 3.38 (5.90; 0 - 108; n=3,802 / 3,851) | 4.40 (6.88; 0 - 83; n=1,131 / 1,149) | 3.61 (6.15; 0 - 108; n=4,933 / 5,000) |
| | | | |
| **Fibrinogen (mg/dL)** | | | |
| **CRP: 0 mg/L** | 277 (48.5; 172 - 436; n=348 / 350) | 290 (57.9; 180 - 540; n=78 / 78) | 280 (50.5; 172 - 540; n=426 / 428) |
| **CRP: 1 mg/L** | 298 (48.5; 109 - 482; n=1,238 / 1,246) | 304 (52.5; 171 - 532; n=292 / 295) | 299 (49.3; 109 - 532; n=1,530 / 1,541) |
| **CRP: 2 mg/L** | 314 (51.2; 183 - 482; n=835 / 841) | 317 (52.5; 138 - 470; n=246 / 247) | 314 (51.5; 138 - 482; n=1,081 / 1,088) |
| **CRP: 3-4 mg/L** | 335 (56.2; 199 - 578; n=711 / 716) | 337 (64.2; 204 - 592; n=222 / 224) | 336 (58.1; 199 - 592; n=933 / 940) |
| **CRP: 5-8 mg/L** | 353 (62.6; 132 - 584; n=330 / 333) | 365 (70.0; 235 - 662; n=126 / 128) | 356 (64.9; 132 - 662; n=456 / 461) |
| **CRP: 9-16 mg/L** | 377 (70.9; 190 - 624; n=222 / 223) | 391 (81.5; 175 - 614; n=110 / 111) | 382 (74.7; 175 - 624; n=332 / 334) |
| **CRP: 17-32 mg/L** | 419 (109.2; 232 - 872; n=59 / 59) | 442 (83.0; 270 - 584; n=36 / 36) | 428 (100.3; 232 - 872; n=95 / 95) |
| **CRP: > 33 mg/L** | 498 (115.4; 274 - 741; n=34 / 34) | 522 (102.3; 367 - 695; n=12 / 12) | 504 (111.5; 274 - 741; n=46 / 46) |

| | | | |
|---|---|---|---|
| **CRP: Missing** | 308 (41.7; 238 - 395; n=14 / 49) | 332 (50.2; 296 - 367; n=2 / 18) | 311 (41.7; 238 - 395; n=16 / 67) |
| **All Subjects** | 320 (64.8; 109 - 872; n=3,791 / 3,851) | 334 (74.1; 138 - 695; n=1,124 / 1,149) | 323 (67.3; 109 - 872; n=4,915 / 5,000) |

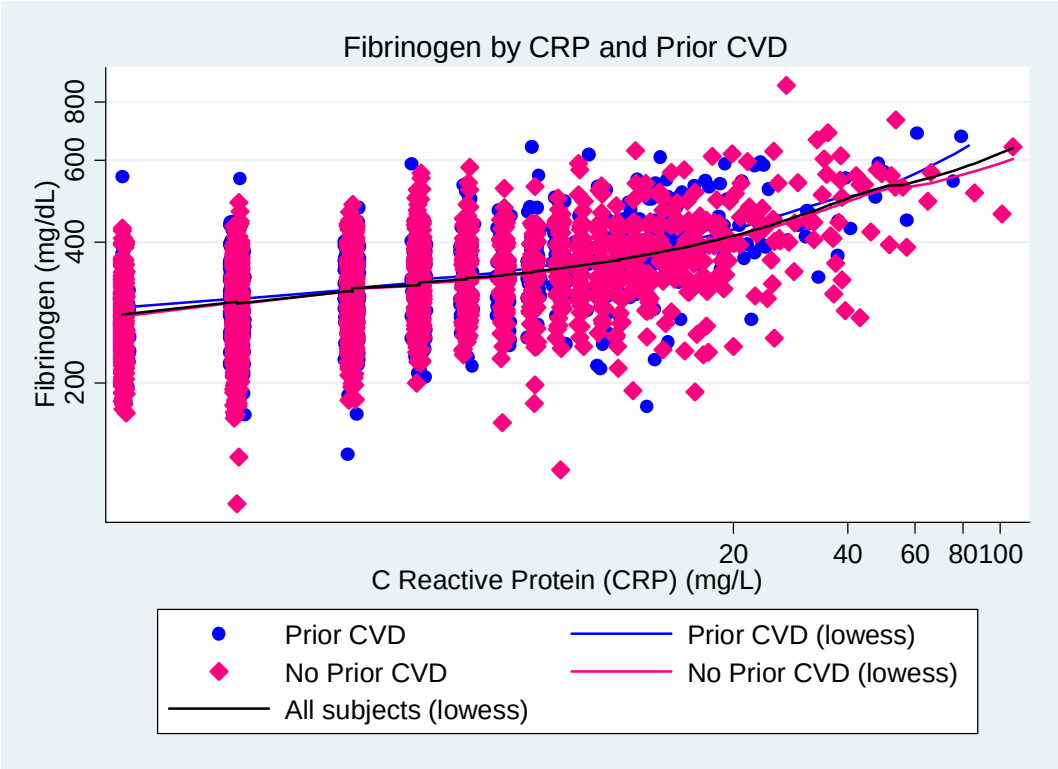
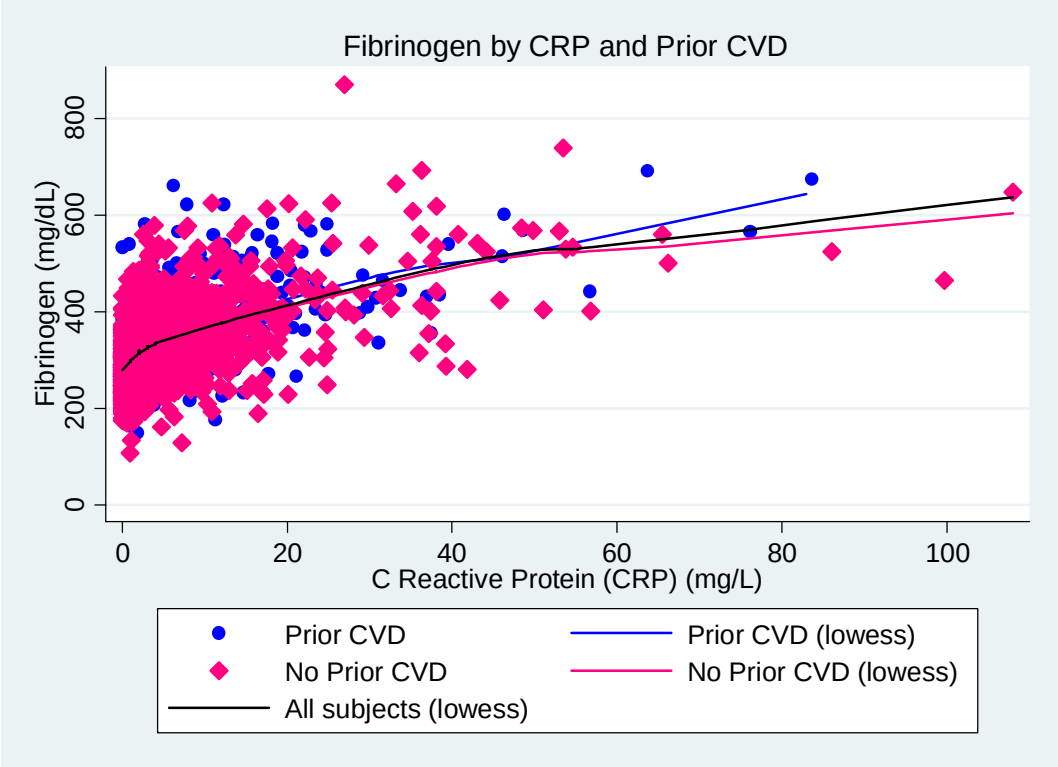



Figure 1: Scatterplot of fibrinogen levels versus CRP levels for 4,899 subjects (3,777 without prior CVD and 1,122 with prior CVD) having both measurements available. Data are plotted on the log-log scale, with subjects having CRP levels reported as 0 mg/L recoded to 0.5 mg/L. Data are plotted in different colors according to the prevalence of prior cardiovascular disease (CVD), though density of data means that some points are obscured. Superimposed on the plot are lowess smoothed curves describing the general trend in the association between fibrinogen and CRP levels.

*(Just for comparison, I present below the scatterplot that would have been obtained if we did not use the log-log scale. In this plot we would need to have commented on the nonlinear trend. Note that the curvilinear pattern we see here is what we would tend to expect if the true linear relationship existed on the log-log scale (i.e. if the true linear relationship were that constant proportionate differences in one variable are associate with constant proportionate differences in the geometric mean of the other variable.)*




2. Perform t test analyses exploring an association between mean fibrinogen and prior history of CVD.

*(In providing answers to this problem, I put the primary differences between the answers to parts a and c in blue font just to emphasize the learning points.)*

a. 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.

**Ans: (10 points total)**

***Methods:*** **The t test that presumes equal variances was used to compare mean fibrinogen across groups defined by the prevalence of prior diagnosis of cardiovascular disease. Estimates of the association were based on the difference in sample means, and 95% confidence intervals and a two sided p value were computed using a pooled variance estimate. Subjects missing data for fibrinogen were omitted from the analysis.**

***Results:*** **Among subjects with available fibrinogen levels, the mean fibrinogen level was 320 mg/dL in the 3,791 subjects without prior diagnosed CVD and 334 mg/dL in the 1,124 subjects who had a prior diagnosis of CVD. Based on a t test that presumes equal variances across groups, the two sided p value P < .0001 allows us to reject the null hypothesis of no difference in the distributions of fibrinogen across CVD groups ~~in favor of the hypothesis that the mean fibrinogen is higher in patients with prior CVD~~. A 95% confidence interval suggests that if variances were truly equal for the two groups, this observed tendency of patients with CVD to have a mean fibrinogen level 14.9 mg/dL higher than patients with no history of CVD is a reasonably typical observation if the true difference in means were anywhere between 10.4 mg/dL to 19.3 mg/dL higher in the prior CVD group.** *(I am being a purist here and pointing out that the t test that presumes equal variances can only be regarded as a test of means, rather than just a test of differences in the distribution, if the group variances were known to be equal.)*

b. 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.

**Ans: (10 points total) We perform a classical linear regression of response fibrinogen on a predictor indicating the prevalence of prior CVD.**

- **The estimated intercept from the linear regression will be exactly equal to the sample mean reported for the non-CVD group in the t test ouput.**
- **The standard error for the intercept from the linear regression will not be equal to the standard error for the sample mean reported in the t test output, because the regression output will use an estimate of the variance from pooling both the CVD and non-CVD groups.**
- **The estimated slope from the linear regression will be exactly equal in absolute value to the difference in sample means reported in the t test output.**
- **The standard error for the linear regression will be exactly equal to the standard error reported for the difference in means in the t test output.**
- **The t statistics from the two analyses will be equal in absolute value.**
- **The two-sided p value from the t test will be exactly equal to the p value for the test of the slope in linear regression analysis.**
- **In absolute value, the bounds of the 95% confidence intervals for the slope will be exactly equal to the bounds of the 95% confidence intervals for the difference in means from the t test.**

c. 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.

**Ans: (10 points total)**

***Methods:*** **The t test that allows for the possibility of unequal variances (Satterthwaite approximation) was used to compare mean fibrinogen across groups defined by the prevalence of prior diagnosis of cardiovascular disease. Estimates of the association were based on the difference in sample means, and 95% confidence intervals and a two sided p value were computed using a sample variance estimates from each group. Subjects missing data for fibrinogen were omitted from the analysis.**

***Results:*** **Among subjects with available fibrinogen levels, the mean fibrinogen level was 320 mg/dL in the 3,791 subjects without prior diagnosed CVD and 334 mg/dL in the 1,124 subjects who had a prior diagnosis of CVD. Based on a t test that allows unequal variances across groups, the two sided p value P < .0001 allows us to reject the null hypothesis of no difference in the mean fibrinogen across CVD groups, in favor of the hypothesis that the mean fibrinogen is higher in patients with prior CVD. A 95% confidence interval suggests that ~~if variances were truly equal for the two groups,~~ this observed tendency of patients with CVD to have a mean fibrinogen level 14.9 mg/dL higher than patients with no history of CVD is a reasonably typical observation if the true difference in means were anywhere between 10.1 mg/dL to 19.7 mg/dL higher in the prior CVD group.**

d. 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.

**Ans: (10 points total) We perform a linear regression of response fibrinogen on a predictor indicating the prevalence of prior CVD using the "robust standard errors" computed using the Huber-White sandwich estimator.**

- **The estimated intercept from the linear regression will be exactly equal to the sample mean reported for the non-CVD group in the t test ouput when allowing unequal variances.**
- **The standard error for the intercept from the linear regression will be more approximately equal to the standard error for the sample mean reported in the t test output when allowing unequal variances, because the regression output will use an estimate of the variance closer to that of the sample variance for the non-CVD groups. (There is still a slight difference in how the sample size is handled in computing the standard error.)**
- **The estimated slope from the linear regression will be exactly equal in absolute value to the difference in sample means reported in the t test output when allowing unequal variances.**
- **The standard error for the linear regression will be approximately equal to the standard error reported for the difference in means in the t test output when allowing unequal variances. (There are differences in the way the sample sizes are handled in the SE, though those differences are negligible when sample sizes are large.)**
- **The t statistics from the two analyses will be only approximately equal in absolute value. (There are differences in the way the sample sizes are handled in the SE, though those differences are negligible when sample sizes are large.)**
- **The two-sided p value from the t test when allowing unequal variances will be only approximately equal to the p value for the test of the slope in linear regression analysis. (There are differences in the way the sample size is handled in the SE and in the way that degrees of freedom are computed, though those differences are neglible when sample sizes are large.)**
- **In absolute value, the bounds of the 95% confidence intervals for the slope will be approximately equal to the bounds of the 95% confidence intervals for the difference in means from the t test. (Again slight differences in small samples due to the handling of the sample sizes and the degrees of freedom.)**

e. 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)?

**Ans: (5 points total) From the output from the t test that presumes equal variances, we see that the group with the smaller sample sizes has a larger estimated standard deviation. In such a setting, the t test that presumes equal variances will report a t statistic that is more extreme than that for the t test that allows the possibility of unequal variances. This is borne out in the result reported. The t statistic was -6.54 in the results in part a and -6.08 in the results in part c.** *(Note that if the variances are truly unequal and the sample sizes are unequal, this will be the general trend across analyses: The t test that presumes equal variances will tend to be anti-conservative when the smaller group has larger variances. On the other hand, the te test that allows for the possibility of unequal variances will always be valid even if the variances are equal. And this is why we use different wording: the t test that presumes equal variances versus the t test that allows for the possibility of unequal variances. That latter test does not "presume" unequal variances.)*

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**.

3. Perform a statistical analysis evaluating an association between mean fibrinogen across groups defined by CRP, modeling CRP as a continuous, untransformed random variable.

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

**Ans: (5 points total) The mean fibrinogen level in a population having a CRP of 0 mg/L is estimated to be 304 mg/dL.**

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

**Ans: (5 points total) When comparing two populations who differ in their CRP levels, we estimate that the mean fibrinogen level will be 5.25 mg/dL higher for each 1 mg/dL absolute difference in CRP levels, with the group having the higher CRP level having higher mean fibrinogen.**

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

**Ans: (10 points total)** *(I think using robust SE is very important in these data.)*

***Methods:*** **A linear regression model of (untransformed) fibrinogen as the response and (untransformed) CRP as predictor was fit to describe the linear trend in mean fibrinogen as a function of absolute differences in CRP levels. The Huber-White sandwich estimator was used to compute standard errors, thereby relaxing assumptions of constant variance across groups. Point estimates of the association were based on the (untransformed) slope parameter from the linear regression analysis, and 95% confidence intervals and a two sided p value were computed using the approximately normally distributed Wald statistics computed using the regression slope estimate and its standard error. Subjects missing data for fibrinogen or CRP were omitted from the analysis.**

***Results:*** **From a linear regression analysis of the 4,899 subjects having both fibrinogen and CRP levels available, we estimate that when comparing two populations having two different CRP levels, the mean fibrinogen level is 5.25 mg/dL higher for each absolute 1 mg/L difference between the two groups in their CRP levels, with the higher CRP group tending toward higher fibrinogen levels. Such a result is highly statistically significant (P < .0001), thus allowing us to with high confidence reject the null hypothesis of no difference in mean fibrinogen across CRP groups in favor of the hypothesis that mean fibrinogen tends to be higher with higher levels of CRP. A 95% confidence interval suggests that true linear trends of 4.60 mg/dL to 5.90 mg/dL higher mean fibrinogen level per 1 mg/L absolute difference in CRP levels might reasonably result in the estimated trends we observed.**

    d. 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).

**Ans: (5 points total)** *(Make certain that the students provide their answers in units of mg/dL and identify whether the summary measure was the mean or the geometric mean.)*

**We estimated the fitted mean fibrinogen from the equation:**

***FittedMeanFIB = 304.0152 + 5.250855 * CRP***

**The fitted value results for the specified values of CRP are shown in Table 1 below.** *(I actually used Excel to produce my estimated central tendency and contrasts.)*

4. 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.)

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

**Ans: (5 points total) The mean fibrinogen level in a population having a CRP of 1 mg/L is estimated to be 296 mg/dL.** *(I transformed CRP using the base 2 logarithm in order that I could easily talk about a "doubling" of CRP level. This answer should not depend at all on which transformation was used because $\log_b(1) = 0$ for every base.)*

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

**Ans: (5 points total) When comparing two populations who differ in their CRP levels, we estimate that the mean fibrinogen level will be 25.5 mg/dL higher for each two-fold relative difference in CRP levels, with the group having the higher CRP level having higher mean fibrinogen.** *(Again, I used a base 2 logarithm to transform CRP, so the software would get me decent units. The student who instead used some other base, might report a different effect for some other relative difference in CRP. If they used a base 10 logarithm, they might have obtained 25.5 * $\log_2(10)$ = 25.5 * log(10) / log(2) = 84.7 mg/dL as the difference in mean fibrinogen per 10-fold difference in CRP. Similarly, if they used the natural logarithm, they might have obtained 25.5 * $\log_2(e)$ = 25.5 * 1 / log(2) = 36.8 mg/dL as the difference in mean fibrinogen per 2.718-fold difference in CRP. Personally, I think it highly suboptimal to talk about a 2.718-fold difference when talking to a human, so I think the student should have chosen some natural unit for their interpretation. In any case, check that they got the answer right. Ask me or the TAs if you have problems interpreting their result.)*

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

**Ans: (10 points total)** *(I think using robust SE is very important in these data. Note that it is not so important what base logarithm was used in the transformation, so long as the interpretations are correct.)*

***Methods:*** **A linear regression model of (untransformed) fibrinogen as the response and logarithmically transformed (base 2) CRP as predictor was fit to describe the linear trend in mean fibrinogen as a function of proportionate differences in CRP levels. CRP measurements for subjects with recorded measurements of 0mg/L were recoded to 0.5 mg/L for the purposes of the regression analysis (1 mg/L was the lowest positive measurement observed in the sample). The Huber-White sandwich estimator was used to compute standard errors, thereby relaxing assumptions of constant variance across groups. Point estimates of the association were based on the (untransformed) slope parameter from the linear regression analysis, and 95% confidence intervals and a two sided p value were computed using the approximately normally distributed Wald statistics computed using the regression slope estimate and its standard error. Subjects missing data for fibrinogen or CRP were omitted from the analysis.**

***Results:*** **From a linear regression analysis of the 4,899 subjects having both fibrinogen and CRP levels available, we estimate that when comparing two populations having two different CRP levels, the mean fibrinogen level is 25.5 mg/dL higher for each two-fold proportionate difference between the two groups in their CRP levels, with the higher CRP group tending toward higher fibrinogen levels. Such a result is highly statistically significant (P < .0001), thus allowing us to with high confidence reject the null hypothesis of no difference in mean fibrinogen across CRP groups in favor of the hypothesis that mean fibrinogen tends to be higher with higher levels of CRP. A 95% confidence interval suggests that true linear trends of 24.0 mg/dL to 27.1 mg/dL higher mean**

**fibrinogen level per two-fold relative difference in CRP levels might reasonably result in the estimated trends we observed.**

d. 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).

**Ans: (5 points total)** *(Make certain that the students provide their answers in units of mg/dL and identify whether the summary measure was the mean or the geometric mean.)*

**We estimated the fitted mean fibrinogen from the equation:**

$$FittedMeanFIB = 295.5663 + 25.5308 * \log_2(CRP)$$

**The fitted value results for the specified values of CRP are shown in Table 1 below.** *(I actually used Excel to produce my estimated central tendency and contrasts.)*

5. 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.

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

**Ans: (5 points total) The geometric mean fibrinogen level in a population having a CRP of 0 mg/L is estimated to be $e^{5.706764} = 301$ mg/dL.**

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

**Ans: (5 points total) When comparing two populations who differ in their CRP levels, we estimate that the geometric mean fibrinogen level will be 1.40% higher for each 1 mg/dL absolute difference in CRP levels, with the group having the higher CRP level having higher mean fibrinogen.**

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

**Ans: (10 points total)** *(I think using robust SE is very important in these data. Again, the base used in the logarithmic transformation is unimportant, so long as the parameters are later interpreted appropriately. I used the natural log in order that Stata would back transform the parameter estimates using the eform() option.)*

***Methods:*** **A linear regression model of logarithmically transformed (natural log) fibrinogen as the response and (untransformed) CRP as predictor was fit to describe the linear trend in geometric mean fibrinogen as a function of absolute differences in CRP levels. The Huber-White sandwich estimator was used to compute standard errors, thereby relaxing assumptions of constant variance across groups. Point estimates of the association were based on the exponentiated slope parameter from the linear regression analysis, and 95% confidence intervals and a two sided p value were computed using the approximately normally distributed Wald statistics computed using the regression slope estimate and its standard error. Subjects missing data for fibrinogen or CRP were omitted from the analysis.**

***Results:*** **From a linear regression analysis of the 4,899 subjects having both fibrinogen and CRP levels available, we estimate that when comparing two populations having two different CRP levels, the geometric mean fibrinogen level is 1.40% higher for each absolute 1 mg/L difference between the two groups in their CRP levels, with the higher CRP group tending toward higher fibrinogen levels. Such a result is highly statistically significant (P < .0001), thus allowing us to with high confidence reject the null hypothesis of no difference in mean fibrinogen across CRP groups in**

**favor of the hypothesis that geometric mean fibrinogen tends to be higher with higher levels of CRP. A 95% confidence interval suggests that true linear trends of 1.22% to 1.58% higher geometric mean fibrinogen level per 1 mg/L absolute difference in CRP levels might reasonably result in the estimated trends we observed.**

d. 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).

**Ans: (5 points total)** *(Make certain that the students provide their answers in units of mg/dL and identify whether the summary measure was the mean or the geometric mean.)*

**We estimated the fitted mean fibrinogen from the equation:**

$$FittedGeomMeanFIB = \exp( 5.706764 + 0.013919 * CRP)$$

**The fitted value results for the specified values of CRP are shown in Table 1 below.** *(I actually used Excel to produce my estimated central tendency and contrasts.)*

6. 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.)

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

**Ans: (5 points total) The geometric mean fibrinogen level in a population having a CRP of 1 mg/L is estimated to be $e^{5.678587}$ = 293 mg/dL.** *(I transformed CRP using the base 2 logarithm in order that I could easily talk about a "doubling" of CRP level. This answer should not depend at all on which transformation was used because $\log_b(1) = 0$ for every base.)*

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

**Ans: (5 points total) When comparing two populations who differ in their CRP levels, we estimate that the geometric mean fibrinogen level will be 7.58% higher for each two-fold relative difference in CRP levels, with the group having the higher CRP level having higher mean fibrinogen.** *(Again, I used a base 2 logarithm to transform CRP, so the software would get me decent units. The student who instead used some other base, might report a different effect for some other relative difference in CRP. If they used a base 10 logarithm, they might have obtained 1.0758 ^ log2(10) = 1.0758 ^ [log(10) / log(2)] = 1.275 as the proportionate difference in mean fibrinogen per 10-fold difference in CRP. Similarly, if they used the natural logarithm, they might have obtained 1.0758 ^ log2(e) = 1.0758 ^ [1 / log(2)] = 1.111 mg/dL as the proportionate difference in mean fibrinogen per 2.718-fold difference in CRP. Personally, I think it highly suboptimal to talk about a 2.718-fold difference when talking to a human, so I think the student should have chosen some natural unit for their interpretation. In any case, check that they got the answer right. Ask me or the TAs if you have problems interpreting their result.)*

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

**Ans: (10 points total)** *(I think using robust SE is very important in these data. Note that it is not so important what base logarithm was used in the transformation, so long as the interpretations are correct.)*

***Methods:* A linear regression model of logarithmically transformed (natural log) fibrinogen as the response and logarithmically transformed (base 2) CRP as predictor was fit to describe the linear**

**trend in geometric mean fibrinogen as a function of proportionate differences in CRP levels. CRP measurements for subjects with recorded measurements of 0mg/L were recoded to 0.5 mg/L for the purposes of the regression analysis (1 mg/L was the lowest positive measurement observed in the sample). The Huber-White sandwich estimator was used to compute standard errors, thereby relaxing assumptions of constant variance across groups. Point estimates of the association were based on the exponentiated slope parameter from the linear regression analysis, and 95% confidence intervals and a two sided p value were computed using the approximately normally distributed Wald statistics computed using the regression slope estimate and its standard error. Subjects missing data for fibrinogen or CRP were omitted from the analysis.**

***Results:* From a linear regression analysis of the 4,899 subjects having both fibrinogen and CRP levels available, we estimate that when comparing two populations having two different CRP levels, the geometric mean fibrinogen level is proportionately 7.58% higher for each two-fold proportionate difference between the two groups in their CRP levels, with the higher CRP group tending toward higher fibrinogen levels. Such a result is highly statistically significant (P < .0001), thus allowing us to with high confidence reject the null hypothesis of no difference in geometric mean fibrinogen across CRP groups in favor of the hypothesis that geometric mean fibrinogen tends to be higher with higher levels of CRP. A 95% confidence interval suggests that true linear trends of 7.14% to 8.02% proportionately higher geometric mean fibrinogen level per two-fold relative difference in CRP levels might reasonably result in the estimated trends we observed.**

d. 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).

**Ans: (5 points total)** *(Make certain that the students provide their answers in units of mg/dL and identify whether the summary measure was the mean or the geometric mean.)*

**We estimated the fitted mean fibrinogen from the equation:**

$$FittedGeomMeanFIB = \exp( 5.678587 + 0.073052 * \log_2 (CRP) )$$

**The fitted value results for the specified values of CRP are shown in Table 1 below.** *(I actually used Excel to produce my estimated central tendency and contrasts.)*

**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** | 309.27 | 295.57 | 305.11 | 292.54 |
| **2 mg/L** | 314.52 | 321.10 | 309.39 | 314.71 |
| **3 mg/L** | 319.77 | 336.03 | 313.73 | 328.45 |
| **4 mg/L** | 325.02 | 346.63 | 318.12 | 338.56 |
| **6 mg/L** | 335.52 | 361.56 | 327.10 | 353.34 |
| **8 mg/L** | 346.02 | 372.16 | 336.34 | 364.21 |
| **9 mg/L** | 351.27 | 376.50 | 341.05 | 368.76 |

| | | | | |
|---|---|---|---|---|
| **12 mg/L** | 367.03 | 387.09 | 355.59 | 380.12 |

7. Complete the following table that makes comparisons (differences or ratios) of the fitted values for each of the models.

**Ans: (10 points total)** ***Grade this problem according to whether they got the right numbers for each problem, with each problem counting equally. (It is really problem 8 that we care about.)***

**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** | 1.017 | 1.086 | 1.014 | 1.076 |
| **3 mg/L / 2 mg/L** | 1.017 | 1.047 | 1.014 | 1.044 |
| **4 mg/L / 1 mg/L** | 1.051 | 1.173 | 1.043 | 1.157 |
| **4 mg/L / 2 mg/L** | 1.033 | 1.080 | 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 |

8. With respect to the results presented in Table 2, answer the following questions:

   a. 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.

**Ans:** **(5 points total) There were three comparisons of groups that differed in their CRP levels by 1 mg/dL: 2 vs 1; 3 vs 2 and 9 vs 8. In all three of those comparisons, the differences in fitted values for the model fit in Problem 3 (the regression of FIB on CRP) showed a constant difference of 5.25 mg/dL in the estimated mean fibrinogen levels.** *(This difference is of course just the estimated slope.)*

**Similarly, there were three comparisons of groups that differed in their CRP levels by 3 mg/dL: 4 vs 1; 6 vs 3; and 9 vs 6. In all three of those comparisons, the differences in fitted values for the model fit in Problem 3 (the regression of FIB on CRP) showed a constant difference of 15.75 mg/dL in the estimated mean fibrinogen levels.** *(This difference is of course just 3 times the estimated slope.)*

**None of the other three models estimated a constant contrast in the difference of the summary measures (i.e., difference in means for Problem 4 and difference in geometric means for Problems 5 and 6) for a constant additive difference in CRP levels.**

b. 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..

**Ans:** **(5 points total) There were three comparisons of groups that differed in their CRP levels by 1 mg/dL: 2 vs 1; 3 vs 2 and 9 vs 8. In all three of those comparisons, the ratios of fitted values for the model fit in Problem 5 (the regression of log(FIB) on CRP) showed a constant ratio of 1.014 in the estimated geometric mean fibrinogen levels.** *(This difference is of course just the exponentiated estimated slope.)*

**Similarly, there were three comparisons of groups that differed in their CRP levels by 3 mg/dL: 4 vs 1; 6 vs 3; and 9 vs 6. In all three of those comparisons, the ratios of fitted values for the model fit in Problem 5 (the regression of log(FIB) on CRP) showed a constant ratio of 1.043 in the estimated mean fibrinogen levels.** *(This difference is of course just the exponentiated estimated slope that is then raised to the 3$^{rd}$ power.)*

**None of the other three models estimated a constant contrast in the ratio of the summary measures (i.e., ratios of means for Problem 3 and 4 and ratios of geometric means for Problem 6) for a constant additive difference in CRP levels.**

c. 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.

**Ans:** **(5 points total) There were five comparisons of groups that differed in their CRP levels by a multiplicative factor of 2: 2 vs 1; 4 vs 2; 6 vs 3; 8 vs 4; and 12 vs 6. In all five of those comparisons, the differences in fitted values for the model fit in Problem 4 (the regression of FIB on log(CRP)) showed a constant difference of 25.53 mg/dL in the estimated mean fibrinogen levels.** *(Because I used a base 2 logarithm when transforming the CRP level, this difference is of course just the estimated slope.)*

**Similarly, there were two comparisons of groups that differed in their CRP levels by a multiplicative factor of 1.5: 3 vs 2; and 9 vs 6. In both of those comparisons, the differences in fitted values for the model fit in Problem 4 (the regression of FIB on log(CRP)) showed a constant difference of 14.93 mg/dL in the estimated mean fibrinogen levels.** *(Because I used a base 2 logarithm when transforming the CRP level, this difference is of course just the base 2 logarithm of 1.5 (which is log(1.5) / log(2) = 0.5849625) times the estimated slope.)*

**None of the other three models estimated a constant contrast in the difference of the summary measures (i.e., difference in means for Problem 3 and difference in geometric means for Problems 5 and 6) for a constant multiplicatively higher CRP levels.**

d. 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.

**Ans: (5 points total) There were five comparisons of groups that differed in their CRP levels by a multiplicative factor of 2: 2 vs 1; 4 vs 2; 6 vs 3; 8 vs 4; and 12 vs 6. In all five of those comparisons, the ratios of fitted values for the model fit in Problem 6 (the regression of log(FIB) on log(CRP)) showed a constant ratio of 1.076 in the estimated geometric mean fibrinogen levels.** *(Because I used a base 2 logarithm when transforming the CRP level, this difference is of course just the exponentiated estimated slope.)*

**Similarly, there were two comparisons of groups that differed in their CRP levels by a multiplicative factor of 1.5: 3 vs 2; and 9 vs 6. In both of those comparisons, the ratios of fitted values for the model fit in Problem 6 (the regression of log(FIB) on log(CRP)) showed a constant ratio of 1.044 in the estimated geometric mean fibrinogen levels.** *(Because I used a base 2 logarithm when transforming the CRP level, this difference is of course just the exponentiated slope that is then raised to the power of the base 2 logarithm of 1.5 (which is log(1.5) / log(2) = 0.5849625))*

**None of the other three models estimated a constant contrast in the ratio of the summary measures (i.e., ratio of means for Problems 3 and 4 and ratio of geometric means for Problem 5) for a constant multiplicatively higher CRP levels.**

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

**Ans: (5 points total) As discussed above:**

- **We expect inflammation (a latent variable that we are really interested in) to act multiplicatively on CRP: Every increased stage in inflammation (whatever that means) will tend to act to multiply the CRP levels. Hence, we would tend to gain precision if we model the logarithmically transformed CRP. This is probably due both to there being less heteroscedasticity as well as having a more linear model.**
- **It is logical to similarly logarithmically transform fibrinogen so they are more or less on the same scale.**
- **So, while there is nothing inherently wrong in modeling the means, in settings such as this there will likely be more precision in modeling the geometric mean across proportionate differences in CRP levels.**
- *(While it is absolutely wrong to answer this question based on the strength of associations we found (that would lead to inflation of the type I error by multiple comparisons), it is illustrative to look at the strength of associations we observed: The t statistics from modeling the geometric mean were on the order of 30, with the highest test statistic for the model in problem 6. The models based on the mean had t statistics on the order of 15. Clearly we had enough precision in either case, but I will claim that what we observed in this one sample is quite likely to generalize to other datasets with CRP and fibrinogen: The biochemical mechanisms will generally lead to multiplicative models. While we will learn that we can consider multiplicative models for the mean (I will recommend using Poisson regression), the heteroscedasticity is usually better addressed by looking at geometric means.)*