Comments on Biostatistics 208 Lab #1

1. NUMERICAL DESCRIPTION

One notable feature in the `summarize` output for `logvl0` is that the four smallest values of baseline viral load are the same. If you computed the measured values using the command

```
gen vl0 = 10^logvl0
```

and then listed or tabulated the values, you would see that these are approximately equivalent to 500 copies per ml – presumably the lower limit of detection of the assay that was being used at the time of this trial. There do not appear to be any large outliers.

2. SINGLE VARIABLE GRAPHS

2.1 Histograms

The default number of bins selected by Stata for the log-10 viral load measures is eight, and seems to be a good selection, providing a useful compromise between summarization and detail. Decreasing the number of bins to, say, 5, makes little sense, and increasing it even just to 10 does not add useful information. The distribution for the baseline measure is notably short tailed and does not appear greatly skewed. As discussed in Chapter 2 of the textbook, this is relatively good news, since it is long-tailed distributions and outliers that are more problematic, especially in small samples, for statistical procedures involving the Normality assumption. The distribution of the change measure is fairly similar, with some indication of a longer tail on the left.

2.2 Boxplots

Reading off the percentiles from the boxplot is usually difficult; it's usually easier to get these using the `summarize` command with the `detail` option. You can see that there is slight left skewness and a few mild outliers at the lower end of the distribution, corresponding to reductions in viral load of between 1 and 2 logs, presumably due to treatment. With a sample of 76 there would likely be little reason for concern in using ANOVA to compare the changes in log-10 viral load across treatment groups, or the *t*-test to assess the effects of mutations in the protease gene at codons 10, 71, or 90. We'll try applying the simple linear regression model below to look at the effect of presence of a mutation at codon 90 as an exercise.

2.3 Q-Q Normal Plots

The Q-Q plot of baseline viral load has the S-curvature typical of the short tails that were also evident in the histogram. Again, it is worth recalling that this is a relatively benign violation of the Normality assumption. The corresponding plot for the change in log viral load measure looks similar on the right, but on the left reveals the presence of a heavier tail.

What will probably be evident from running Q-Q plots on Stata-generated observations from a Normal distribution is that with small samples of 25, 50, or even 76 observations (as in the ACTG 333 dataset), there is noise in these plots, particularly at the left and right extremes. Thus the "art" in making a good judgment about whether the Normality assumption is seriously enough violated to warrant transforming the variable or using a procedure that avoids this assumption. A few other things to keep in mind:

1. In Chapter 2 we note that the Normality-based procedures are surprisingly robust,

especially in larger samples. This is important, since non-parametric multi-predictor regression methods are not nearly so well developed as the parametric methods we will cover in this course.

2. The Normality assumption involves the outcome in the analysis; we do not have to make this assumption about predictors, although outliers in the predictors can cause trouble of a different sort, and transformation of predictors can resolve various other problems. This will be discussed further in Lecture 6.

3. Sensitivity analyses are useful for checking whether you get qualitatively the same answers to your research questions using non-parametric methods or after Normalizing transformation of the variable.

2.4 Normal Density Plots

The *density* function for a random variable is the probability of occurrence of possible values for that variable. It can be graphically represented with the values on the horizontal axis, and the corresponding probability on the vertical axis. A familiar density is that from the standard normal distribution (mean=zero, SD=1), which is symmetric about 0, and appears bell-shaped.

The `kdensity` command in Stata produces a plot of the empirical density for sampled values of a continuous variable. This can be interpreted as a smoothed version of a histogram, with the vertical scale representing probability of occurrence (or relative frequency) of particular values of the variable. By supplying the `normal` option, this produces a plot, along with a corresponding normal density, selected to have the same mean and variance as those observed for the chosen variable. Looking at these two curves together reveals the nature of any discrepancy between the distribution of the observed variable, and an idealized normal counterpart. Looking at the resulting plot and a corresponding Q-Q plot provides a lot of useful information about how the observed variable agrees with a normal distribution.

Question 1: The plot reveals the heaviness in the left tail picked up in the corresponding QQ plot. The overall impression is that the distribution of the change measure in not drastically non-normal (i.e. no major skewness), but it's not particularly symmetrical either. Based on the comments in lab, the relatively short tail might make us feel OK to assume normality for regression purposes. But some sensitivity analyses might be in order.

3.1 Simple linear regression with a single categorical predictor

The output of the model is given below.

```
. regress logvlch i.codon90

      Source |       SS       df       MS              Number of obs =      76
-------------+------------------------------           F(  1,    74) =    3.85
       Model |  1.21130234     1  1.21130234           Prob > F      =  0.0535
    Residual |  23.2794674    74  .314587397           R-squared     =  0.0495
-------------+------------------------------           Adj R-squared =  0.0366
       Total |  24.4907697    75  .326543596           Root MSE      =  .56088

------------------------------------------------------------------------------
     logvlch |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   1.codon90 |   .2557006   .1303095     1.96   0.053    -.0039468     .515348
       _cons |  -.3902997   .0991507    -3.94   0.000    -.5878617   -.1927377
------------------------------------------------------------------------------
```

Note that the "`i.`" prefix for the binary predictor isn't required here since it's coded as 0/1.

Question 2: The constant term (`_cons`) gives the estimated mean change in log 10 viral load observed in individuals without a mutation at codon 90. The *t*-test result ($P<0.001$) and 95% confidence interval (-0.588, -0.193) indicate that this is significantly different than zero. The coefficient for codon 90 measures the difference in mean change in log 10 viral load between individuals with a codon 10 mutation and those without a mutation. The coefficient (0.256) and accompanying *t*-test result ($P=0.053$) evaluating the null hypothesis that the true coefficient equals zero indicate a marginally non-significant increase (i.e. not quite significant at the 5% level) in observed change in individuals with a mutation (95% CI: -0.003, 0.515), compared to those without a mutation.

Question 3: The *R*-squared statistic indicates that the model accounts for approximately 5% of the observed variation in change in log 10 VL. The default *F* test compares the model including the codon 10 indicator variable as a predictor to one with no predictors (i.e. using only the overall mean to predict log 10 change in VL). The result is the same as the *t*-test, since in this case, both tests address the same hypothesis.

Question 4: Independence and linearity are clearly not of concern here. Based on our graphical assessments, there is some evidence for non-normality of the distribution of the outcome variable, but this doesn't appear to be severe. The equal variance assumption could be examined using a side-by-side box plot for example (`graph box logvlch, over(codon90)`). Based on these results, we may have some concern about these last two assumptions.

Question 5: In this case, a Wilcoxon rank-sum test would be an easy way to evaluate these results avoiding the assumptions made by the linear model:

```
. ranksum logvlch, by(codon90)

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

     codon90 |      obs    rank sum    expected
-------------+---------------------------------
      absent |       32        1068        1232
     present |       44        1858        1694
-------------+---------------------------------
    combined |       76        2926        2926

unadjusted variance     9034.67
adjustment for ties       -0.12
                     ----------
adjusted variance       9034.54

Ho: logvlch(codon90==absent) = logvlch(codon90==present)
             z =  -1.725
    Prob > |z| =   0.0845
```

The result indicates a non-significant difference in average observed changes between mutation groups. Note that he *P*-value is somewhat larger than the *t*-test result, possibly dampening enthusiasm for a "marginally significant" finding.

Note that we could also consider the bootstrap as discussed in lecture. This wouldn't really be necessary in this simple example given the availability of the Wilcoxon test. Here are the results anyway:

```
. bootstrap `"regress logvlch codon90"' _b, reps(1000)

command:      regress logvlch codon90
statistics:   b_codon90  = _b[codon90]
              b_cons     = _b[_cons]

Bootstrap statistics                              Number of obs    =        76
                                                  Replications     =      1000

------------------------------------------------------------------------------
Variable     |  Reps  Observed      Bias  Std. Err. [95% Conf. Interval]
-------------+----------------------------------------------------------------
   b_codon90 |  1000  .2557006  .0026882  .1419083  -.0227719   .5341731   (N)
             |                                     -.0159343   .5399283   (P)
             |                                     -.0287454   .5351689  (BC)
      b_cons |  1000 -.3902997 -.0015908  .1218157  -.6293437  -.1512557   (N)
             |                                     -.6349715  -.1526274   (P)
             |                                     -.6298484  -.1471314  (BC)
------------------------------------------------------------------------------
Note:  N   = normal
       P   = percentile
       BC  = bias-corrected
```

Looking at the preferred bias-corrected (BC) bootstrap CI for the predictor, the conclusions remain unchanged.