Stat 215b (Spring 2004): Comments on Lab 1

B. M. Bolstad
bolstad@stat.berkeley.edu

Feb 19, 2004

For the most part the labs write-up submitted were good. But there seem to a few things that could have been done better.

1. Mention the Gauss-Markov Theorem in association with your linear model assumptions. Let our model be $Y = X\beta + \epsilon$. Remember that the G-M assumptions are that $\mathrm{E}[\epsilon_i] = 0$, $\mathrm{Var}(\epsilon_i) = \sigma^2$ and that the ϵ_i are independent. If these hold then

   $$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y$$

   is the BLUE of β and

   $$\mathrm{Var}(\beta) = \left(X^T X\right)^{-1} \sigma^2$$

2. Make it clear you understand what you gain by introducing the additional assumption of normality. In particular how this related to the t-statistics and F statistics

3. Rather than just saying "the P-value is small so BLAH is not significant" try to make clear what test the P-values correspond to. For example that you are making tests of the form $H_0 : \beta_i = 0$ versus $H_1 : \beta_i \neq 0$, or whatever it is. In addition clearly spell out what the test statistic is (eg a formula). Make sure you do this at least once to make clear you understand what test is being carried out (you need not do it every time you do the same style test).

4. Please note that the `anova()` command give sequential sums of squares so any resultant tests should also be interpreted in that frame work (eg the F test is for the effect of variable C, taking into account the variability already explained by A and B)

5. When looking at residual plots the words "homoskedastic" and "hetroskedastic" are nice to use. If you look at plots of residuals versus index make clear that you are looking for auto-correlation. In the case of this particular dataset I am not sure there is much reason to suspect such a pattern.

6. If you are going to look at outliers in your residual plots also look for influential outliers. Do this by looking at Cook's distances or alternatively look at $h_i i$ choosing values that are higher than $2p/n$. In general practice you would always make an attempt to explain why a particular observation is an outlier and only if you could give a good explanation (eg data recorded incorrectly) would you actually remove it from your dataset.

7. For the simulation of qqplots, it would have been better to show a number of simulated plots (9 is a convenient number) rather than just 1. Note also that your conclusion should be that the residuals are "approximately normal" from such a plot.

8. Several presentation issues

   (a) Don't use computer output. ie don't follow the style of VR and put computer output in a typewriter font. Please go to the effort of putting things in more nicely formatted tables.

   (b) In your tables, use no more than 4 decimal places or 4 significant figures (keep the same standard across your entire report).

   (c) Try to place a slightly more descriptive caption under each plot and always be sure to reference things in the body of your report.

   (d) 10 point is a little hard to read. Use the *fullpage* latex package if you want to use fewer pages.

   (e) Use a spell checker.

9. Remember that you have a great deal of flexibility in how you write-up each lab and what analysis you try. Never be afraid of trying new things, but be sure to explain clearly what it is you are doing.