STA 302 H1F / 1001 HF - Assignment 1 Comments

2. Most students got this right. There were a few who had numbers that were slightly off, which I did not deduct marks for. If the numbers were very much off, or if many numbers were wrong, a mark was deducted.
3. Quality of responses were mixed. Most people got the right idea about the data points being further from the mean, but some went about it in an indirect way, ie: citing the bimodal nature of the distribution. A few resorted to $R^2 = 1 - RSS/SST$, and argued that SST increases but failed to mention that RSS is relatively the same after considering sample size. You need to talk about the ratio RSS/SST, not just the denominator (what if RSS also increased in proportion to SST?)
A few tried heuristic arguments, including comments like "the data varies around the residual line more in the two regressions than overall", referring me to the figure. Keep in mind the default figures are deceptive as the axis scales are smaller when you plot the individual clusters (short and long). Figures can lie, even when done by software.
Examples:

- `http://www.datavis.ca/gallery/goosed-up.php`
- `http://www.datavis.ca/gallery/lie-factor.php`

Further, simply plugging numbers into the formula for R^2 and showing that is it bigger for the whole data set is not explaining why.
Finally, keep in mind that R^2 is a measure of variation explained by the **linear** relationship, not any arbitrary relationship. It also doesn't necessarily generalize to prediction (as claimed by some students). Look at Anscombe's Quartet: `http://en.wikipedia.org/wiki/Anscombe's_quartet`.
Some suggested that it was bigger for the larger data set by the "law of large numbers". Keep in mind that this law talks about estimates converging to the true value. There is no reason a priori to believe the true value of R^2 is large (what if there was absolutely no relationship? Then R^2 would go to zero as the sample size goes to infinity).
4. This was generally well done. Some people wrote down the formula without numbers and then gave an (incorrect) test statistic. Next time, write down intermediate steps so I can give part marks. That way, I'll know if it is conceptual error (did you not understand the SAS output?) or did you hit the wrong key on your calculator?
5. This was a two part question

- Might this create any problems...
- Why or why not?

Almost all answer the first (whether it was right or wrong). Many failed to explain why, or if they did, they didn't justify the response sufficiently well. For instance, a common response was "yes, because missing data might cause bias". Missing data isn't always bad. If it can be assumed to be "missing (completely) at random", then there might not be that big of a problem. When it is not, then there are problems. Many didn't check for this.
Some responded that there would be no problems because SAS removes the missing observations. Yes, it does, because it does not contribute information. However, that omission can still create problems in prediction. Don't trust the software to do the right thing.
6. Like the previous question, almost all gave an answer, but some failed to justify it. In particular, some said that there might be correlation, but didn't say why.
Some others mixed up statements about ϵ_i and e_i, claiming that $\text{Var}(e_i) < \infty$ as there is finite data. Remember the difference between the data and the model.
7. This was overall well done. Some recurring issues included people citing arguments about R^2 or not justifying the decision (or in one case, opting for saying "neither, get more data").