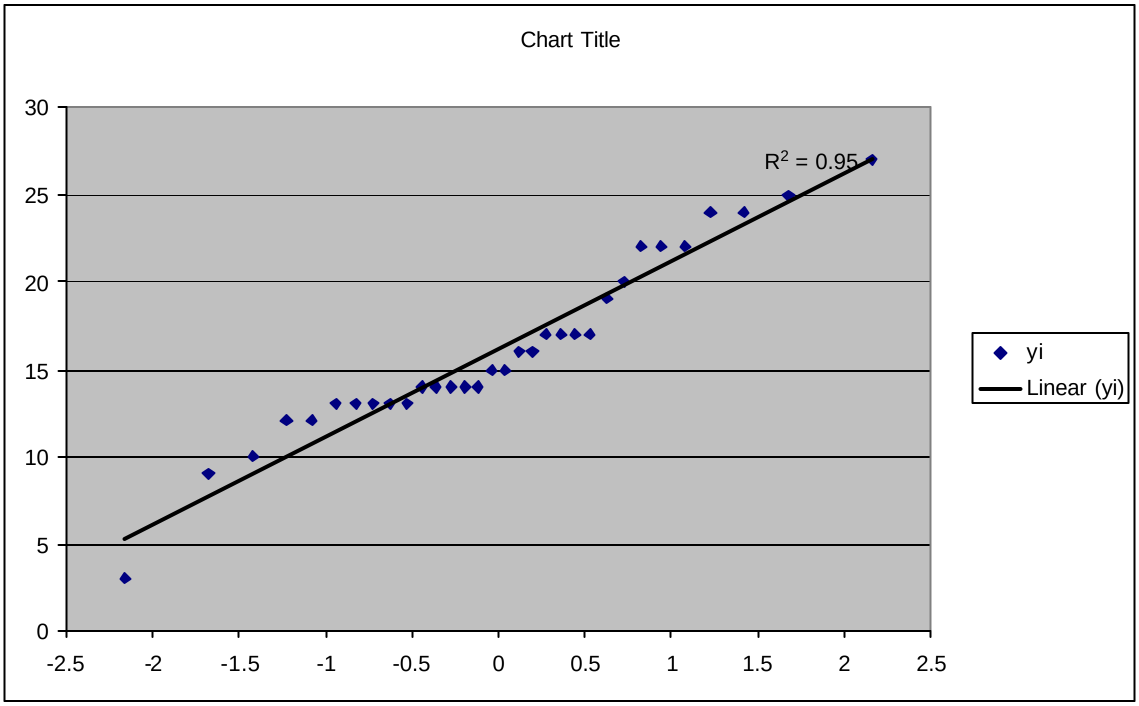
Answers to Homework #1, CS 239, Section 1, Spring 2007

1. Afflalo’s scoring average was 16.125 points per game.
The standard deviation of his scoring average was 5.14

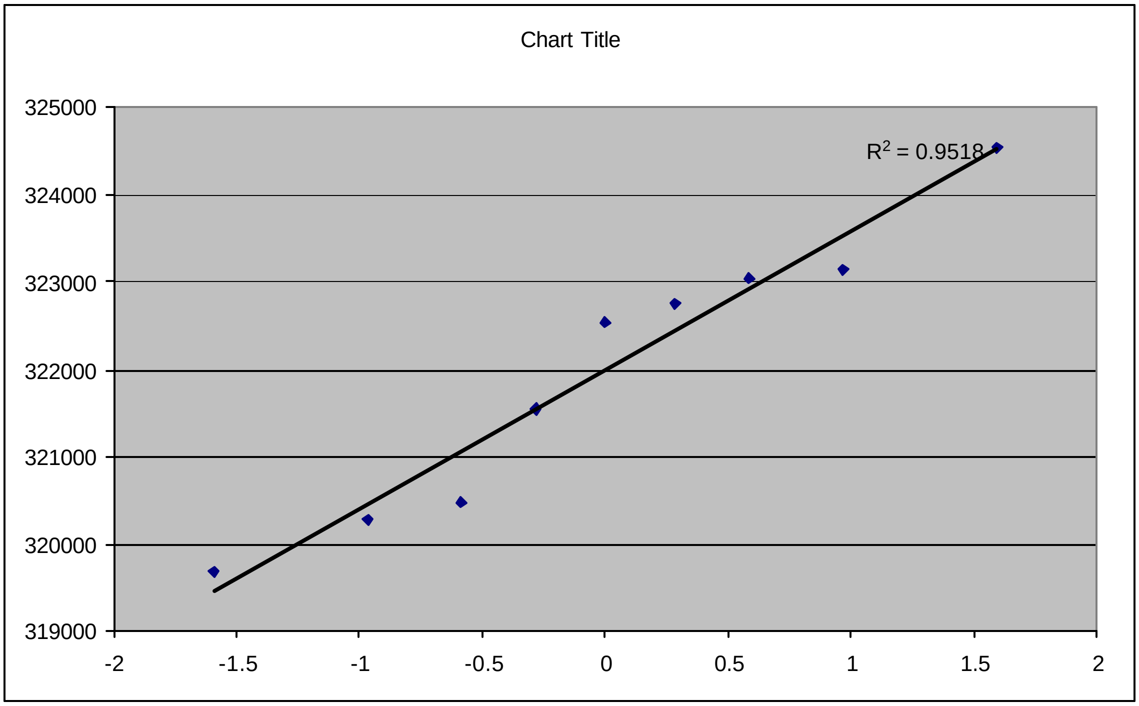
One could test for normal distribution in various ways. One way is to plot a quantile-quantile plot against the normal distribution. Doing so gives the following graph. While not perfectly linear, it is reasonably close. An R2 of .95 indicates that a linear regression is a good fit. One could do further tests to determine if the distribution is normal.



2.

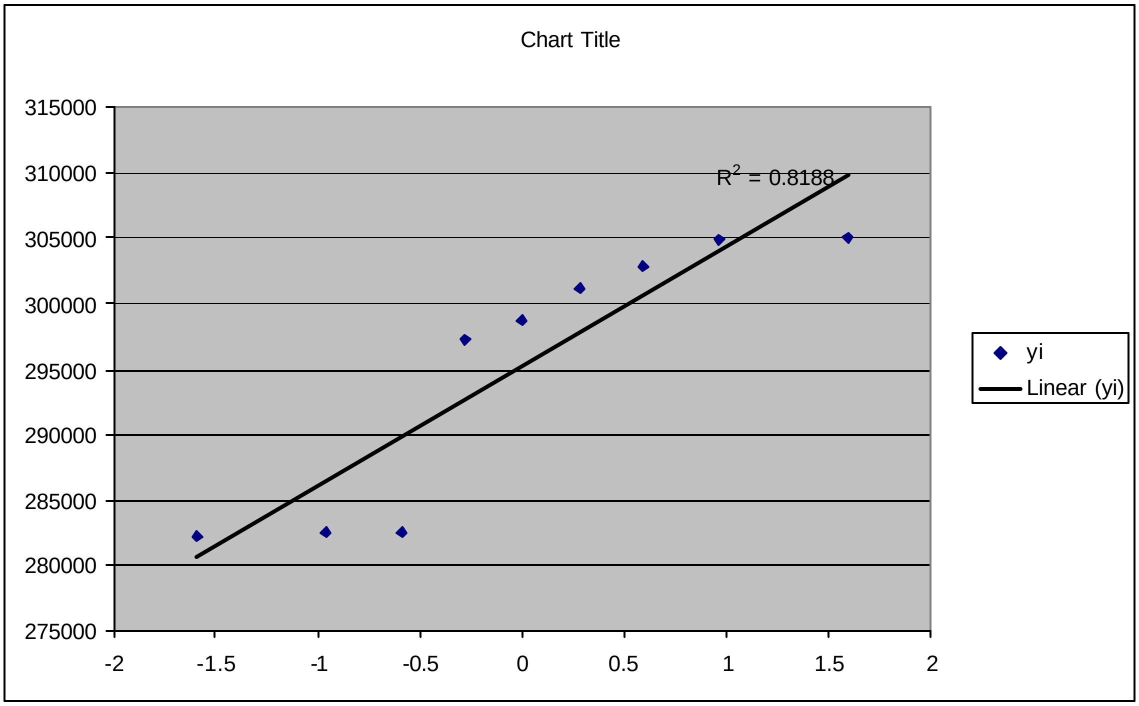
	CFS	RFS
Mean	321996	295260
Stdev	1601	9937.21

Plotting a quantile-quantile plot of CFS against the normal distribution gives the following graph:



The points are approximately linear, and the R2 of the linear regression fitting the points is .95, a good fit.

Below is a quantile-quantile plot of the data for RAMFS:



The points are not really linear, with two poor fits at the lower end and one at the upper end. The R2 of the linear regression is .82, not a very strong fit.

The 90% confidence interval for CFS is (321003,322989). The 90% confidence interval for RAMFS is (289099,301421). The confidence intervals do not overlap, so, overlooking the poor fit of the RAMFS data to the normal distribution, the two systems are different at the 90% confidence level. One could also treat the observations as paired and analyze the results that way. In this case, one would get a mean difference of 26,736, with a 90% confidence interval of +-1.86*(9539/sqrt(9)), or 26738+-5914. Since the confidence interval does not include 0, the systems are different at this level of significance.

The 95% confidence interval for CFS is (320766,323227). The 95% confidence interval for RAMFS is (287622,302899). Again, the confidence intervals do not overlap, so the performance of the two systems is different at the 95% confidence level, as well, again overlooking issues of normality. One could also treat the observations as paired and analyze the results that way. In this case, one would get a mean difference of 26,736, with a 95% confidence interval of +-2.306*(9539/sqrt(9)), or 26738+-7332. Since the confidence interval does not include 0, the systems are different at this level of significance.

3.
On some of these, I accepted other answers if I felt the answer had been argued well.

A. Mean is the best choice to express the central tendency of this metric. Depending on issues not fully specified in the question, one should use either arithmetic mean or harmonic mean. The former would be appropriate if, in each test, the same number of legitimate packets were sent. The latter would be appropriate if, in each test, a different number of legitimate packets were sent. If the number was approximately the same in all cases, using arithmetic mean would probably be OK, anyway.

To appropriately choose the index of variability, we would need to know more about the characteristics of the measured values. If we assume the distribution of those values is unimodal symmetric, we would use variance. If not, we would use semi-interquartile range. We would not use the range, because the range of a percentage is uniformly bounded by 0 and 1. Chances are that range would tell us little about the variability of the measurements.

B. Mode is the proper choice of index of central tendency. Fractional disk drives have no real-world meaning. On the other hand, knowing what number of disks is most likely to be on tells you something important about system behavior. Median is not too helpful, since we can’t really tell from the median if the number expressed is something that happens regularly or rarely.

To express variability, range won’t work well. We can be pretty sure, regardless of the actual experimental data we gathered, that the minimum is one and the maximum is N. To make our other choices, again we would want to know a bit more about the actual distribution of the data observed. Also, the value of N might matter. If N is small, then something like interquartile range is unlikely to be too helpful. For example, with an N of 4, there are only four possible measures. It is highly likely, without seeing the data, that the low quartile value is 1 and the high quartile value is 4, or, if the system rarely uses lots of disks, 3. A value like variance is more likely to provide insight.

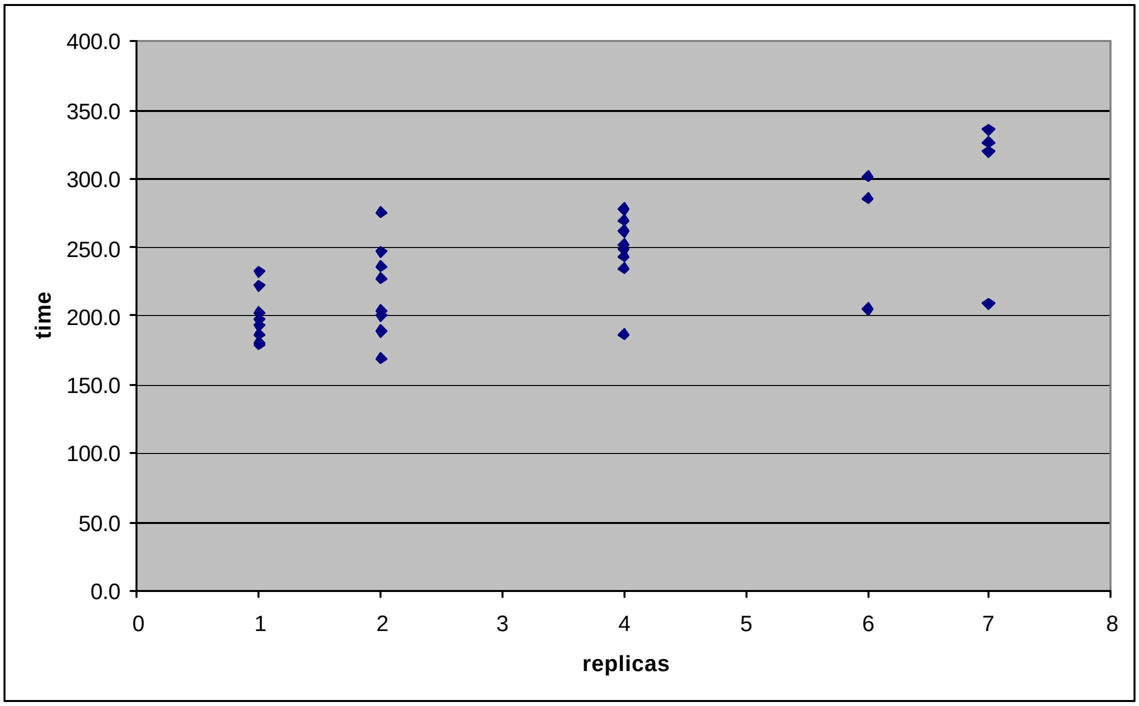
C. For this algorithm, using the mean is probably the best choice. The data is not categorical, and a sum of times is a meaningful value. Arithmetic mean will be fine, here.

Again, the best selection of index of variability will depend on distribution of the actual data, but variance is likely to be the right choice.

4.

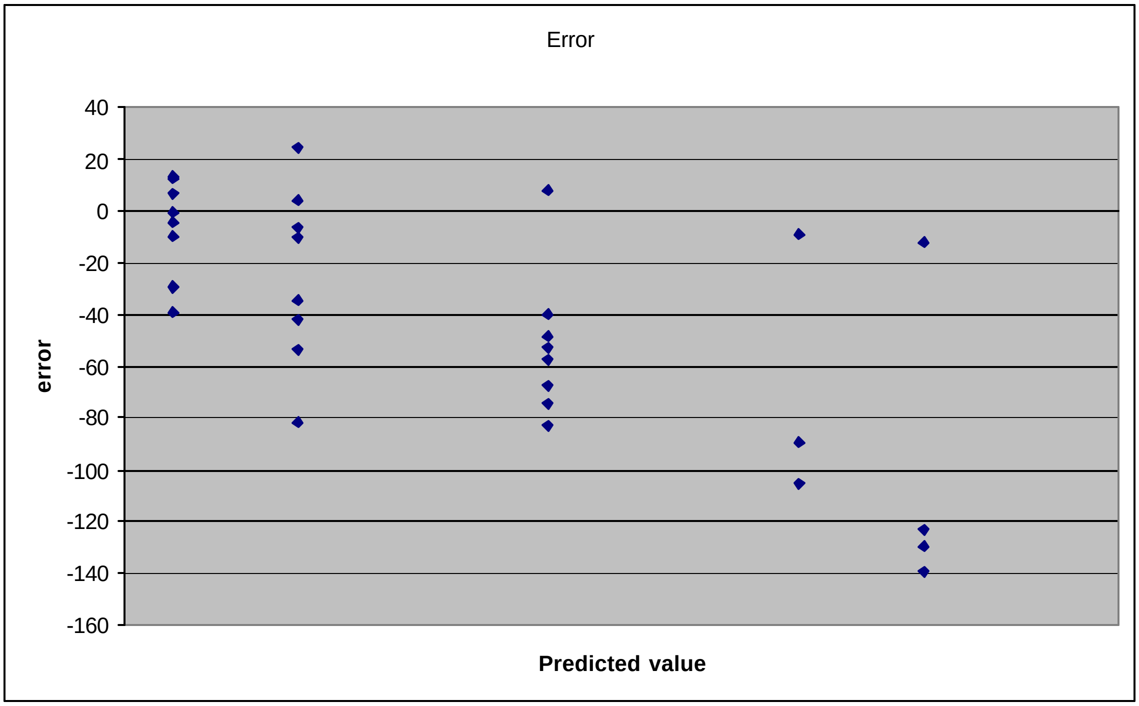
A. Is the data a good candidate for linear regression?

Plotting the simple x vs. y values gives you this chart:



There appears to be a basically linear trend, but the spread of values, and particularly the outliers at high numbers of replicas, are troubling.

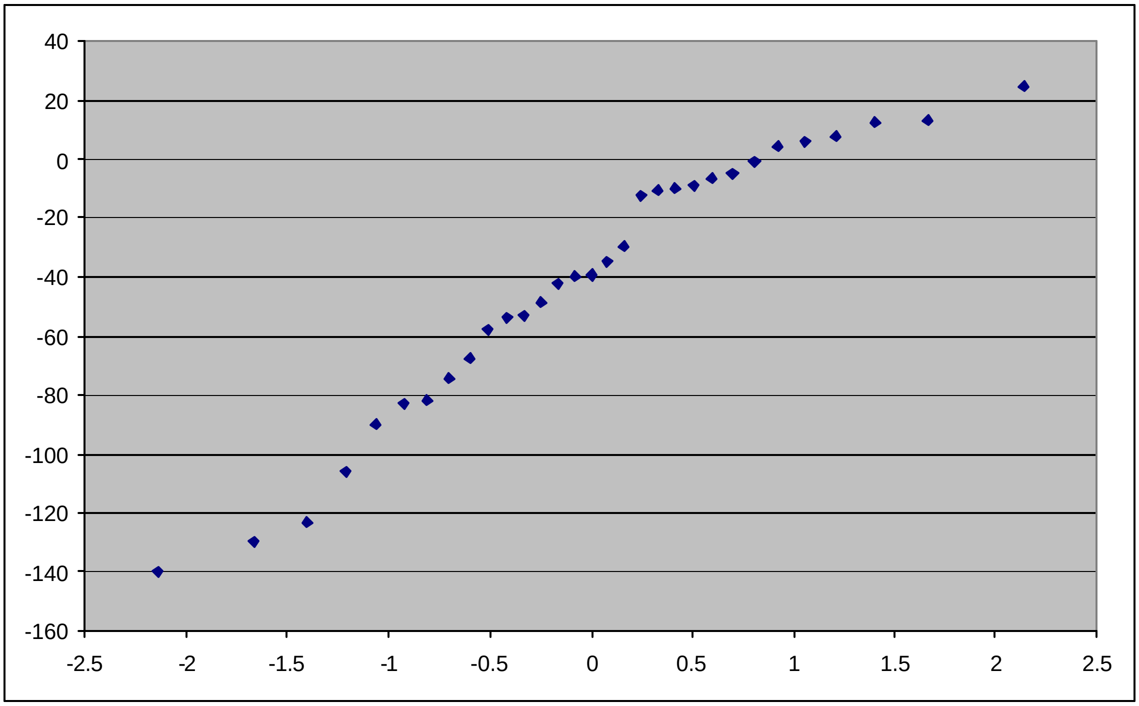
Plotting the errors vs. predicted responses to check for independence of errors, we get the following chart:



Are there visible trends? Trends can be in the eye of the beholder, but the errors seem to be both smaller and more likely to be positive, rather than negative, for low numbers of replicas. This doesn't look so good. It seems likely there's a dependence of the errors on the predictor variable, number of replicas.

We can also use this chart to look for evidence for or against homoscedasticity, a constant standard deviation of errors, which is another property required for linear regression to be valid. Is the spread of values through this chart roughly constant? Not really. For the low points representing one replica, the spread is relatively small. The spread generally (though not uniformly) increases as the number of replicas increase. More evidence that the distribution of errors is not independent of the predictor variable value.

Next, we test for normal distribution of errors, using a quantile-quantile plot of the errors vs. the normal distribution. The chart is below:



As usual on analysis of a quantile-quantile chart, the question is whether the plot is generally linear. While not perfect, this one is fairly linear. The assumption of normal distribution of the errors isn't that bad.

On the whole, the evidence suggests that this data is not a good candidate for a linear regression. The R2 value of .4851 does not indicate a good model, either.

B.

Bulling ahead regardless, the linear regression performed on this data produces the following equation: y = 185.56 + 15.118 x. So b0 is 185.56 and b1 is 15.118. Now to calculate confidence intervals for these parameters.

sb0 = 11.29
sb1 = 2.89

The t distribution value for a 95% confidence interval with 29 degrees of freedom is 2.045. So the 95% CI for b0 is 185.56 +- 23.08 = (162.48,208.64) and the 95% CI for b1 is 15.118 +- 5.91 = (9.1,21.03).

C. The R2 value for the regression is .4851, so the regression describes less than half the variability in the data.

D. Plugging 5 replicas into the regression model, the predicted mean time is 261.15 seconds. Using a 90% confidence interval on the this predicted mean of a single test is (202.24,320.06). One could perform this calculation using the unit normal distribution instead of the t distribution, since there are more than 30 observations. The confidence interval then becomes (204.1,318.2)