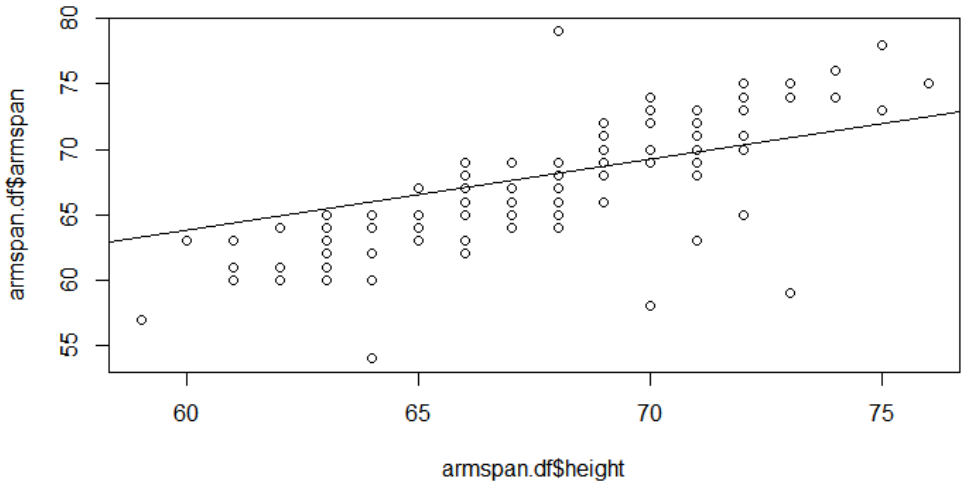
**Kitu Komya** (UID: 404-491-375)
Statistics 101A (Discussion 3A)
Homework 1 (due: 10/07/16)

Please note: I'm concurrently taking Stats 20, so I'm still familiarizing myself with R.

---

1a.



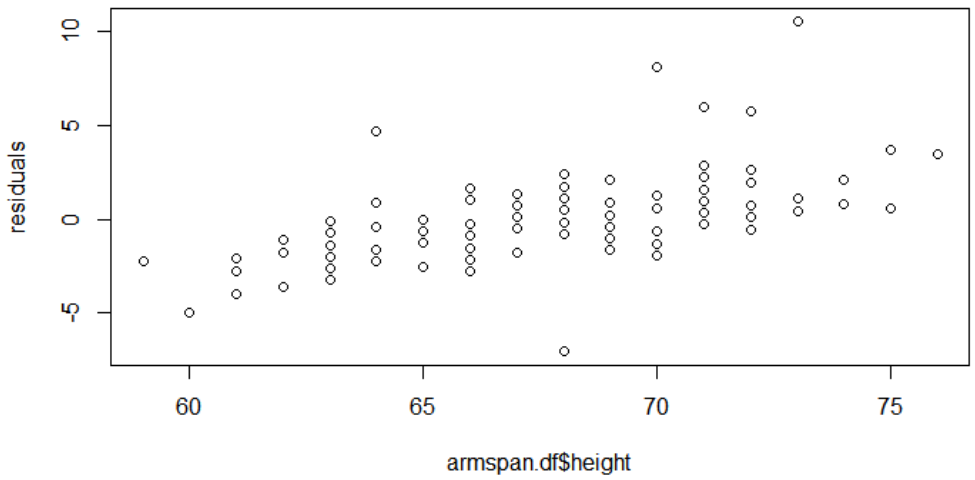

This plot shows a positive, linear, moderately strong association between height and armspan. An unusual feature includes the fact that the plot looks very grid-like since the data are discrete values rounded to the nearest inch. Another unusual feature are the six outliers that clearly do not fit the trend of the plot, but are still relevant and true data to be included.

1b. The linear model best fit trend line has been superimposed in 1a. The equation given in the R console is $\hat{E}(Y|x) = 1.00049x - 0.63014$, where x is the height in inches.

1c. My height is 65 inches. Based on the formula from the linear model, my predicted armspan is $\hat{E}(Y|x) = 1.00049 * (65) - 0.63014$, = 64.40171. However, since my actual armspan is 64 inches, my residual, $e_i = y_i - \hat{y}$ = 64 – 64.40171 = -0.40171.

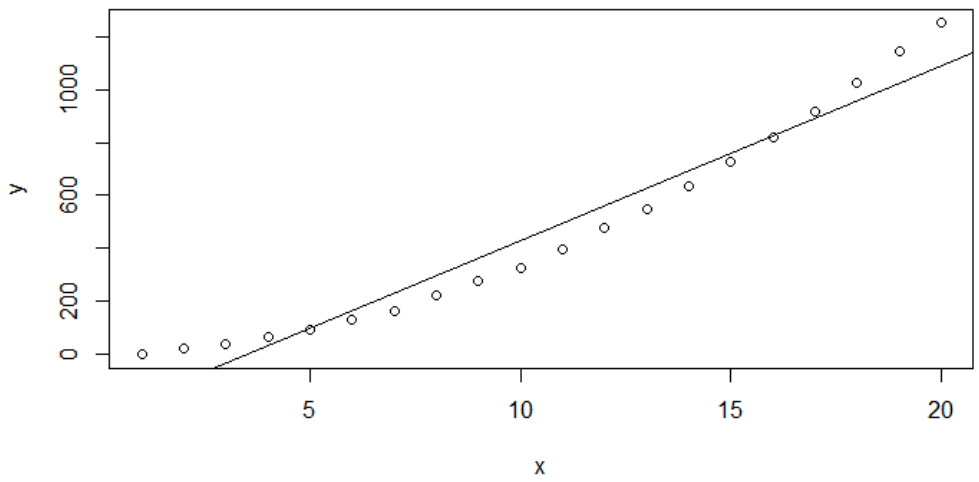
1d. Using the quartiles (1Q and 3Q) found in the residuals summary in the console, the median/middle 50% of residuals from our sample lie between -1.4014 and 1.5976. Michael Phelps' armspan has a residual of $e_i = y_i - \hat{y}$ = 79 – [$1.00049 * (76) - 0.63014$] = 3.5929, which is clearly much higher than the median/middle 50% of the residuals from our sample, so yes this seems unusual to me. Good.

1e.




The residuals are not as randomly scattered as I had hoped. In fact, the plot of the residuals looks a bit positively linear, suggesting that perhaps our data is invalid for a linear model fit. Or perhaps, we should exclude the six outliers for a better linear model fit. For a good linear model fit, the residual plot should have residuals equally balanced above and below the line when residual = 0. This clearly is not the case in our situation, since most of the data points in the first half are below 0, while the next half are mostly above 0.

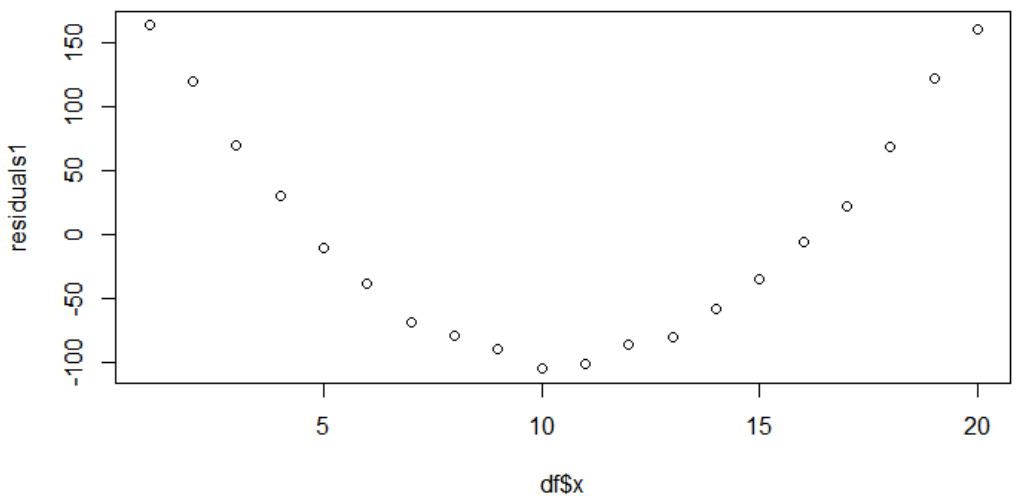
2a.




This model is invalid for a linear regression. As the plot above shows, the scatterplot is exponential, while the linear model trendline is linear. One of the assumptions made with linear regression is that the data is linear, which is clearly not the case here, so this is a bad model.

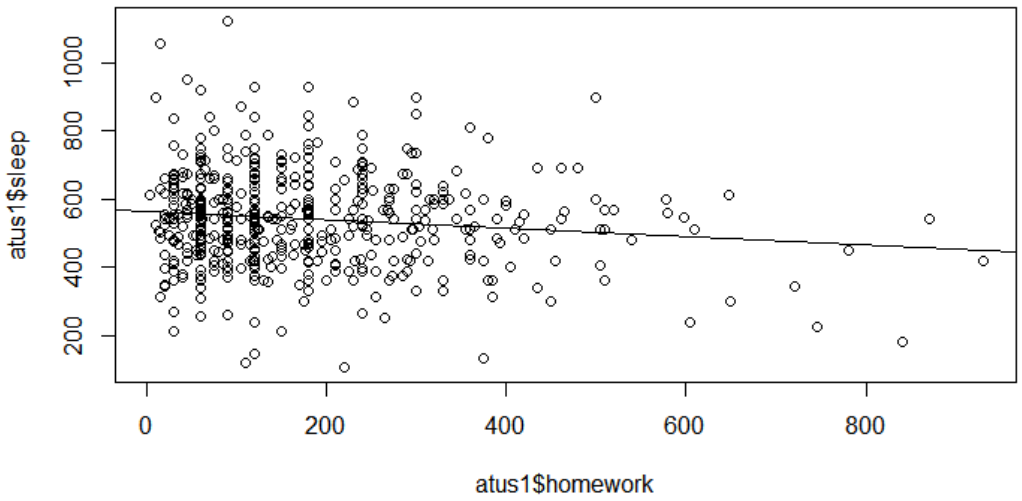
2b. The residual plot will look parabolic since the data are above the trendline for the first few values of x, and then scoops down below the trendline for the majority of the x values, until once again the data rise above the trendline near the end of the x-value.

2c.



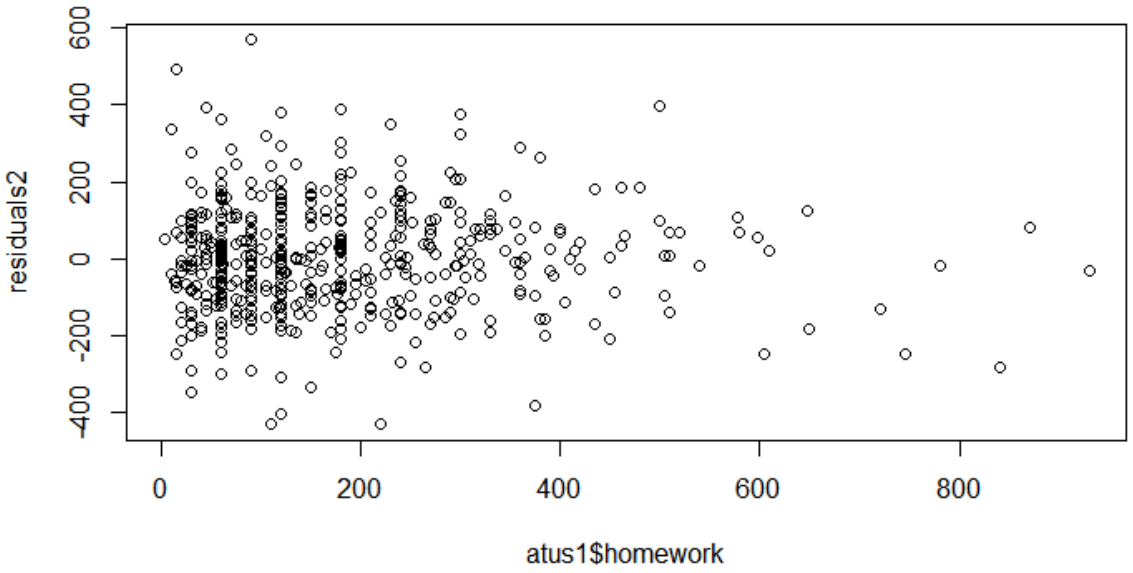

As confirmed, the residual plot is parabolic, indicative of a bad fit for the linear model, since the residuals show a pattern when they should be randomly distributed for a linear model. An exponential fit will perhaps be better.

3a.




This plot shows a negative, linear, weak association between time spent on homework and time spent sleeping. The correlation seems so weak because the majority of the points are not near the linear model trendline. In fact, they seem so randomly scattered that it almost looks like there is no association. Moreover, the slope is nearly horizontal (a slope of zero) which indicates no association at all. There is high variability in sleeping times on the majority of x-values of homework time, particularly before the 400 mark.

3b.




The residual plot looks more randomly scattered than previous residual plots, indicating that it is a good fit for a linear model. However, upon looking at the scatterplot and the linear model trendline, one can easily deduce that although there is a linear association, the association is so weak that it should be neglected. Thus, in this situation, both the scatterplot and the residual plot are needed to make a fair conclusion that the association between time spent on homework and time sleeping is negatively and linearly correlated, but with a weak association.