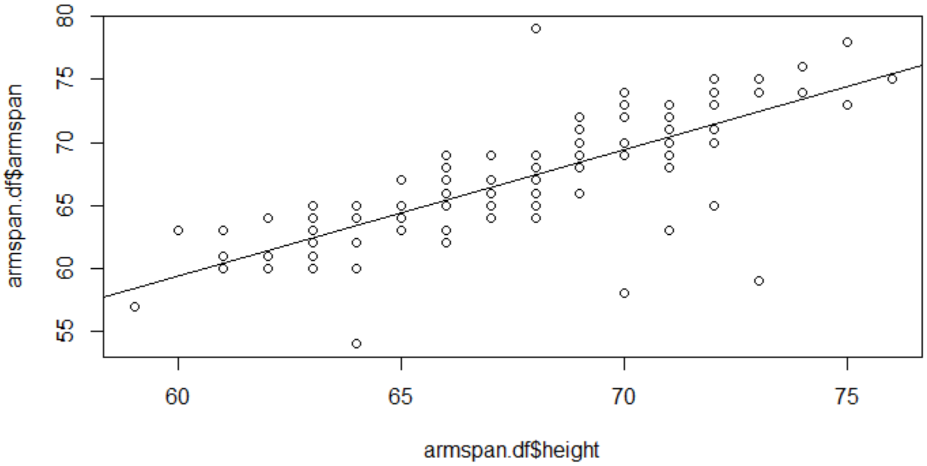
Kitu Komya (UID: 404-491-375)
Statistics 101A (Discussion 3A)
Homework 3 (due: 10/14/16)

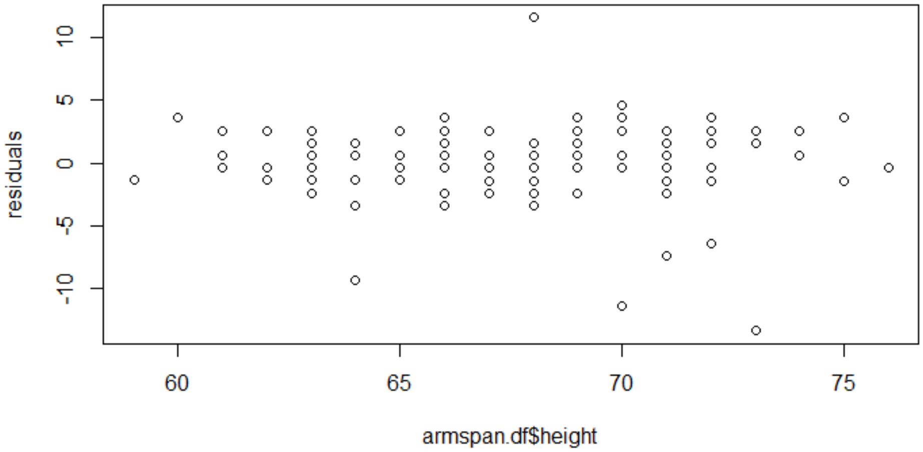
Please note: I'm concurrently taking Stats 20, so I'm still familiarizing myself with R.

1a. From the armspan data set, I have removed outliers (see R code in #1) by only sub-setting those values that have at least 30 inches of an armspan. This will eliminate any bad data, such as unusually low inches that result from typos or from improper units, such as feet instead of inches.



Above is the resulting scatterplot. The summary function in R provides the following equation of the model in the R console: $\hat{E}(Y|x) = 1.00049x - 0.63014$.

1b.



The residuals are randomly scattered with no obvious pattern or trend. For a good linear model fit, the residual plot should have residuals equally balanced above and below the line when residuals = 0, which is clearly the case. Thus, the lack of a pattern means that we can assume linearity and proceed.

1c. The following outputs show a 95% confidence interval found using the confint function.

y-intercept: [-9.7383189, 8.478036]
slope: [0.8654558, 1.135531]

Vitrivius' theory claimed that a person's height is approximately the same as her armspan. This claim essentially translates over to a y-intercept of 0 and a slope of 1. Mathematically, *E*(armspan|height) = 0 + 1 * height. Since both of these values are in the 95% confidence interval range, we fail to reject his claim since our data is consistent with his claim.

1d. Using the linear model fit, Phelps' armspan should be 75.40737 inches. Using the predict function, Phelps' armspan at a 95% prediction interval is between [69.62804, 81.1867] inches.

1e. Using the linear model fit, the mean armspan for all those with a height of 76 inches should be 75.40737 inches, just like Phelps' value. However, using the predict function, the mean armspan for all those with a height of 76 inches at a 95% confidence interval is between [74.13804, 76.6767] inches. We see that this range is much more narrow than Phelps' individual range since Phelps' range takes into account the uncertainty of individual variability.

1f. Although Phelps' actual armspan of 79 inches is beyond the 95% confidence interval of the sampling distribution of the means, that is not unusual since the confidence interval is narrower than the prediction interval anyway because the prediction interval has the added uncertainty of predicting a single response versus the mean response. That is because the mean of a sampling distribution can be measured more accurately, and so most of the data points may actually also fall outside the means' range.

Since Phelps' armspan does lie in the prediction interval which measures a 95% prediction interval for two standard deviations for an individual (while for the sampling distribution of the means you must divide 2σ with $\sqrt{n}$, giving a smaller range), his individual armspan is not unusual. Had it been over 81.1867 inches, we would consider it unusual since that would imply that his armspan is outside the prediction interval which says that 95% of the time the individual armspan should be in the interval.

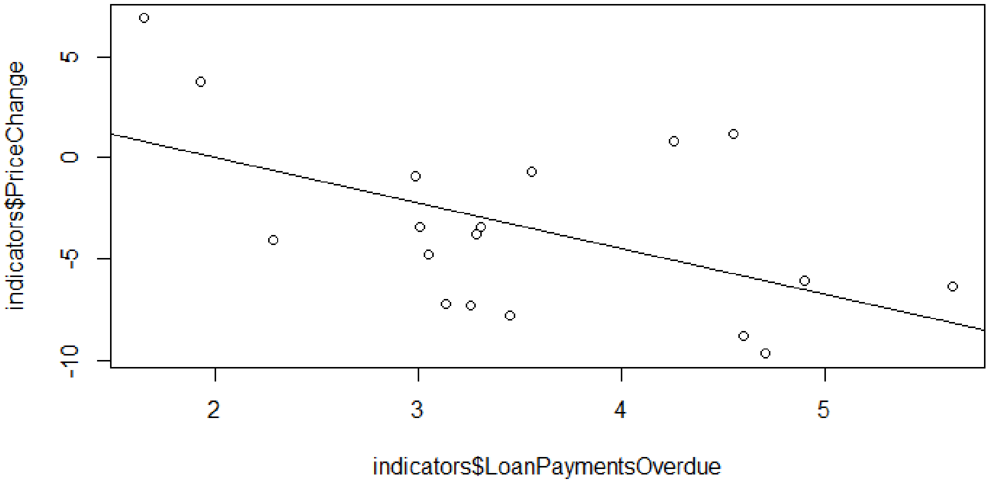
2a. Using the confint function, a 95% confidence interval for the slope of the regression model β_1 is [0.9514971, 1.012666]. Since the value 1 is within the interval range, we fail to reject the claim that 1 is a plausible value for β_1 since our data is consistent with the claim.

2b. Using the pt() function to calculate the two-tailed probability that there is no difference between the expected and calculated y-intercepts where H_0: β_0 = $10,000, we reach a p-value of 0.6242306, indicating that the expected y-intercept of $10,000 is not statistically different from the calculated y-intercept of 6805 (see R code). Thus, we fail to reject the null hypothesis, which means that $10,000 is a plausible value of the y-intercept since our data is consistent with the claim. [source for R code help: www.tinyurl.com/stats101A-ttest.com, pages 5-6]

2c. Using the linear model fit, the gross box office returns for the current week should be $399,637.50. Using the predict function, the gross box office returns for the current week at a 95% prediction interval is between [$359,832.80, $439,442.20]. $450,000 is not a feasible value for the gross box office returns for the current week since the value is beyond the range for an individual box office returns. Earning $450,000 would imply that this value is outside the interval that says that 95% of the time the individual box office returns for the current week should be in the interval, which is, unfortunately, unlikely.

2d. Using such an assumption would mean that every week would output the same gross box office returns. Moreover, such a claim would assume that the slope, H_0: $\beta_1 = 1$ against the H_A: $\beta_1 \neq 1$. We can test this claim using the t.test function. We get a p-value of 0.9656, indicating that the expected slope of 1 is not statistically different from the calculated slope of 0.9821 (see R code). Thus, we fail to reject the null hypothesis, which means that 1 is a possible value of the slope since our data is consistent with our claim. In other words, yes, it is significantly appropriate to predict the next week's gross box office returns by using the current week's gross box office returns. Note that a similar result was found in 2a.

3a.



Although the question did not ask for it, I've attached the scatterplot of the data with a best fit linear model trendline so that it may be of use while answering the questions. Anyway, using the confint function, a 95% confidence interval for the slope of the regression model, β_1, is [-4.163454%, -0.3335853%]. Since the entire range of this 95% confidence interval belongs in the negative values, there is evidence for a significant negative linear association.

3b. Using the linear model fit, the percentage change in average price from July 2006 to July 2007 is -4.479585%. Using the predict function, the percentage change in average price from July 2006 to July 2007 at a 95% confidence interval is between [-6.648849%, -2.310322%]. Based on the confidence interval, 0% is not a feasible value of the mean percentage change in average price since the value 0% is beyond between the range and this interval says that 95% of the time the mean is in this interval. It's possible that 0% is a feasible value for a prediction interval in which individual observations are measured since its range is wider. In this case though, we reject the null hypothesis that 0% is an acceptable value since our data is inconsistent with the claim. This question asked to test the claim of no association between the variables (in other words, the slope is 0). This was further refuted in 3a, when the 95% confidence interval of the slope of the regression model did not include 0. It only included negative values, indicating an association between the two variables, and that too, in the negative direction.

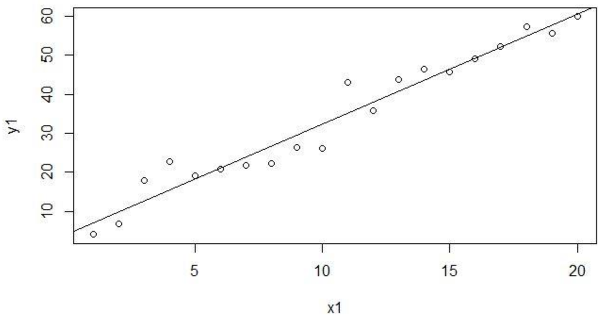
4a. Using the confint function, a 95% confidence interval for the y-intercept of the regression model, β_0, is between [0.391249620, 0.89217014] hours for the start-up time.

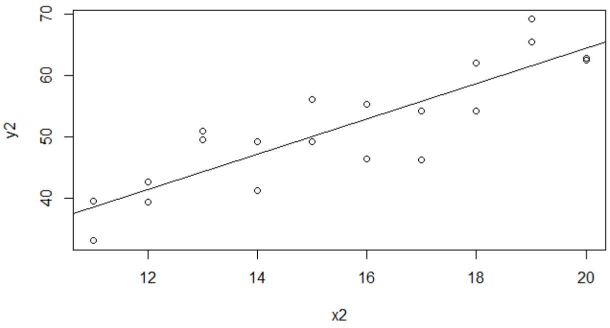
4b. Using the pt() function to calculate the two-tailed probability that there is no difference between the expected and calculated slopes where H_0: $\beta_1 = 0.01$ hours against the H_A: $\beta_1 \neq 0.01$ hours, we reach a p-value of 0.9373167, indicating that the expected slope of 0.01 hours is not statistically different from the calculated slope of 0.0112916 hours (see R code). Thus, we fail to reject the null hypothesis, which means that 0.01 hours is a plausible value of the slope since our data is consistent with the claim. [source for R code help: www.tinyurl.com/stats101A-ttest.com, pages 5-6]

4c. Using the linear model fit, the point estimate for the time to process 130 invoices is 2.109624 hours. Using the predict function, the time to process 130 invoices at a 95% prediction interval is between [1.422947, 2.7963] hours.

5a. The second situation will have a larger confidence interval for the slope. It too has 20 values, but half of them are repeated x-values, and with the addition of the random epsilon, there is variability in the y-axis for each pair of x values. This additional variability on top of epsilon creates a larger interval for the confidence interval. The first situation only has one y-value for each x-value, and thus its confidence interval for the slope has a smaller margin of error and is a more precise estimate of the slope.

Below are the graphs with their best fit linear model trendline that verify these statements. The leftmost graph is the first situation, and the rightmost is the second situation. Again, it is visibly clear that the second situation produces more variability from a best fit linear model trendline since there are multiple y-values for each x-value, making the scatterplot genuinely "scattered."





5b. Using the confint function, a 95% confidence interval for the slope of the regression model for the first situation is between [2.410785, 3.320485]; for the second situation, it is between [2.041785, 3.515933]. By definition, 95% of the confidence intervals will capture the population parameter's value, so although we cannot confirm for certain, we can feel confident that both intervals capture the population parameter's value for both situations.

Moreover, although the interval for the second situation is larger, since its interval as well as the first situation's interval only encompass positive values, we can feel confident that there is a positive, linear association between the two variables, x and y, in both situations. In other words, for both of the situations we reject the null hypothesis that the slope is 0 at the 5% significance level, since that value is not included in either of the 95% confidence intervals.