384M practice problems 2

Practice with plots in R and residuals

We recorded our heights and shoe sizes in class. Let's get some practice working with that data.

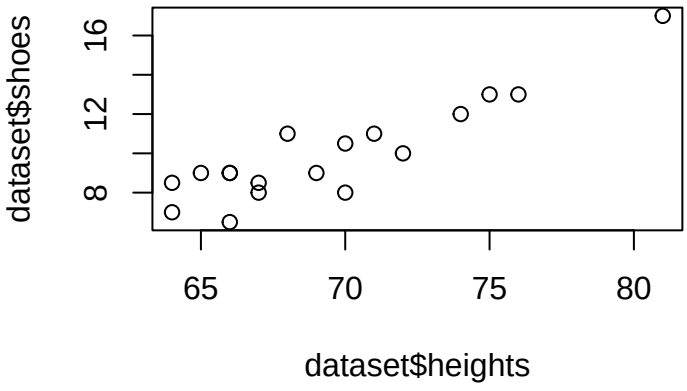
1. Download the heightShoes.csv dataset from canvas if you don't have the data already. You can read it into R with the "import dataset" button in the environment (you can also read it with `read.csv()` if you give it the file's path in quotations). Name it something informative, e.g. `dataset`. (Double check: This should create a data frame in your environment with that name)

2. How can we determine our sample size (in terms of number of people responding) from properties of this data frame?

3. Find the mean of each of our numeric columns. (Bonus review: How would we write this in summation notation? Call the heights x_i and the shoe sizes y_i)

4. Bonus: Make histograms of our heights and shoe sizes by giving those columns of the data frame to the `hist()` function (run each one separately). How might you describe these distributions?

5. Recreate the plot of shoes by height that we made in class. Are the default choices R makes for the axes sensible? Try changing them using `xlim =` and `ylim =` in the plot call (see lecture code for refresher). What seem like the best ones to you for displaying this information? The default one should look like this:



6. Add the centroid to the plot using the `points()` function by giving it the proper coordinates. (remember, there is a help function in R if you run `?points`). Make it your favorite color using the `col=` argument (see this handy website for some color names). Try changing the symbol to something besides the default circle using the `pch=` argument (see this other handy website)

7. Find the best fit line using the `lm()` function. For `lm()` we specify a formula for our model (our $f(x_i)$!) using the variable names and `~` (tilde symbol). The outcome comes first, then the `~` then the predictor. E.g. `lm(y~x, data=my_data)` would correspond to a dataset called `my_data` with columns `y` and `x`, with a model written $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. Give it the relevant variable names and dataset you are using. Save this model as an object in the environment by giving it a name. What are our $\hat{\beta}_0$ and $\hat{\beta}_1$ values?

8. Add the best fit line to your plot using `abline()` giving it the `coefficients` from the model (review: what are these estimates of?). Make the line your second favorite color. (Remember, check out `?abline` for a help file)

The trend line we use doesn't capture the data pattern perfectly. Residuals are the parts of the data that the model doesn't capture, and since $y_i = \hat{f}(x_i) + e_i$ we can write them as $e_i = y_i - \hat{f}(x_i)$. Let's look at the residuals for different candidate models and see how they compare.

9. Add a new column to the data frame with the residuals from a candidate model that says $\hat{f}(x_i) = \bar{Y}$ (i.e. ignores x_i completely; how would you add this line to your scatterplot?). You can add a new column by naming it and defining it, e.g. `dataset$e1 <- dataset$shoes - ybar`. Find the sum of the squared residuals for this model with `sum(dataset$e1^2)`. Which points have the largest/smallest residuals?

10. Now try it for a different candidate model that says $\hat{f}(x_i) = -22 + .47x_i$. You could name this new column of residuals `e2` with `dataset$e2 <- ...`. What is the sum of squared residuals for this model? How does it compare to the value from the model that ignores x_i? (Bonus: Which points have smaller residuals with the second model? Which points have larger ones?)

11. Now find the residuals for a candidate model that uses the best fit line. You could name this column with `dataset$e3 <-...`. What is the sum of squared residuals for this model? How does the result compare to the previous models? (Bonus: Which points are explained best by this model vs the others?)