Math151 – Analysis of the data in the paper "Variation in cancer risk among tissues can be explained by the number of stem cell divisions" by Cristian Tomasetti and Bert Vogelstein

1. In Figure 1 of this paper there is a graph of data for various human tissues on
CR = Cancer risk over lifetime
SD = Total stem cell divisions over lifetime
The data appear to be strongly correlated when plotted on a log scale for both variables, with SD on the horizontal axis.

2. Suppose you wish to find an equation that summarizes the relationship between these two variables. If we assume it is linear on the log scale, then an appropriate equation would be

$$\log(CR) = m\log(SD) + b$$

where m and b are the slope and intercept on the log scales for these variables. So we wish to find m and b.

3. We could use Matlab to find these using the complete dataset that is in the supplementary material for this paper (e.g. it gives the data values for each point in Figure 1). If you'd like some extra credit for this course (worth one point out of 100 for the entire course grade), find the supplementary material and use Matlab to calculate the regression line, compile a report about what you've done and discuss how the results differ from what we describe below.

4. Instead of using Matlab, let us illustrate how to find an equation describing the data by using two points that might be "close" (using an eyeball guess) to a line summarizing the data. In this case, we will use the data for Head osteosarcoma and for HCV hepatocellular carcinoma (you should look at these on the graph to see where they are and whether you agree that these are reasonable choices). The supplemental material for the paper gives the data values – that is the (SD, CR) values as

Head osteosarcoma $(6 \times 10^{6}, 3 \times 10^{-5})$
HCV hepatocellular carcinoma $(2.7 \times 10^{11}, 7 \times 10^{-2})$

5. To place these points on a log scale, take the logarithms so the points are now
$(6+\log(6), -5+\log(3)) = (6.7, -4.5)$ and $(11+\log(2.7), -2+\log(7)) = (11.4, -1.2)$

and using these two points in the equation above for the line (I assume you remember how to do this to find the ma and b) you can calculate that m=.7 and b=-9.3 so

$$\log(CR) = .7\log(SD) - 9.3$$

and taking 10 to each side of this equation we get

$$CR = 5 \times 10^{-10} SD^{.7}$$

6. One way to check this to make sure we haven't made a mistake is to choose another value from the data set and see how its value on the graph in the paper differs from that using this equation. You can choose a different point, but lets use the one for Esaphageal cancer which is at (SD, CR) value (1.2×10^{9}, 1.9×10^{-3})
and when we use the above formula with this value for SD we get 1.1×10^{-3} which is not too terribly far from the actual value for CR.