1. This question uses data from this paper:

<https://science.sciencemag.org/content/347/6217/78>

Variation in cancer risk among tissues can be explained by the number of stem cell divisions.

Science 02 Jan 2015: Vol. 347, Issue 6217, pp. 78-81

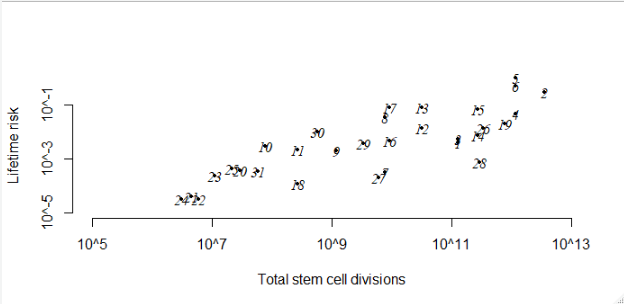
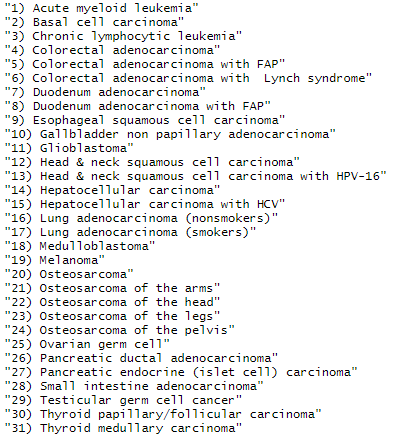
**The R code for questions 1A, 1B, 1C & 1D is located at this link:** <https://github.com/jyoung67/advstatistics-labs/blob/master/labs/lab09/processCancerRisk.R>

**(1A)**: Download the data from here examining the relationship between the number of cell divisions and cancer risk: <https://fodorclasses.github.io/classes/stats2020/cancerRisk.txt>

**On a log10-log10 scale graph Lifetime\_cancer\_risk (on the y-axis) vs. CumulativeCellDivisions (on the x-axis). (This reproduces Fig. 1 from the paper).**

(You can read in the file with read.table("cancerRisk.txt", header=TRUE, sep="\t")

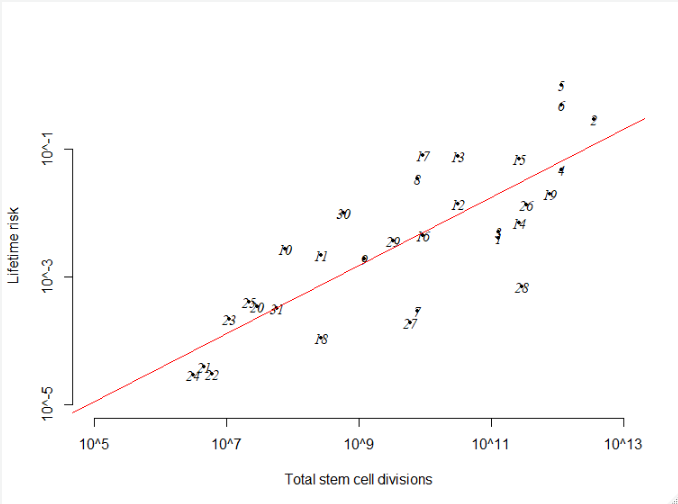
See figure 1 for log10-log10 scale graph of Lifetime cancer\_risk vs. cumulative cell divisions.

**Figure 1:** log10-log10 scale graph of Lifetime cancer risk vs. total stem cell divisions

**(1B):**  **Using the lm function, fit a linear model with Lifetime\_cancer\_risk as the Y variable and CumulativeCellDivisions as the x-data. Add the regression line to the plot using the function abline(myLm) (where myLm is the linear model you created).**

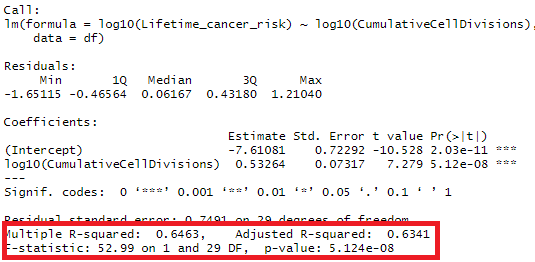
**See figure 2.**



**Figure 2:** Log10-log10 scale graph of Lifetime cancer\_risk vs. CumulativeCellDivisions with added linear model regression line.

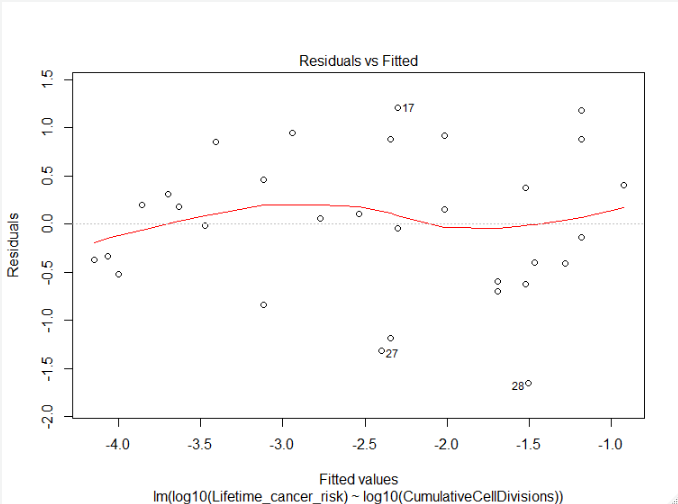
**(1C): What is the p-value for the null hypothesis that the slope of the regression between these two variables is zero? What is the r-squared value of the model?**

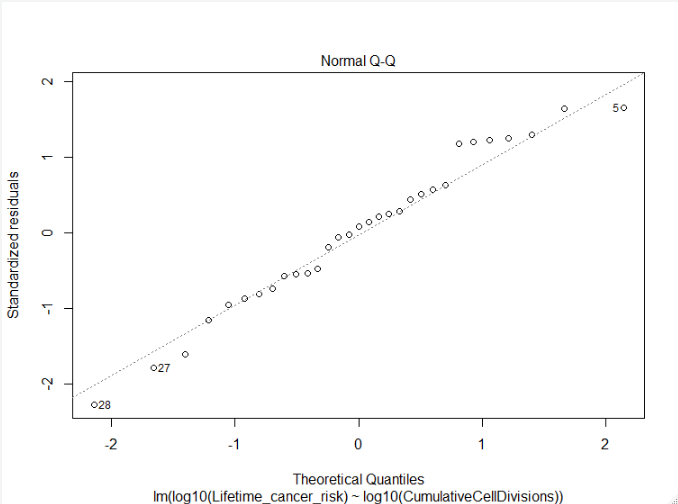
* The p-value is 5.124e-08 (figure 3)
* The multiple r-squared and adjusted r-squared values are 0.6463 and 0.6341, respectively (figure 3).

  
**Figure 3:** lm function results for Lifetime cancer\_risk vs. CumulativeCellDivisions

**(1D): Are the assumptions of constant variance and normal distribution of the residues reasonable for this model? Justify your answer.**

Yes, the assumptions of constant variance and normal distribution of the residues are reasonable for this model. In the residuals-vs-fitted plot in figure 4, the residuals mostly bounce randomly around the zero line, suggesting that a linear relationship is reasonable. Also, the residuals roughly form a “horizonal band” around the zero line in figure 4, suggesting that the variances of the residual terms are equal. Finally, the normal Q-Q plot depict a reasonably liner relationship of standardized residuals vs theoretical quantiles, suggesting normal distribution in figure 5.

  
**Figure 4:** Residuals vs Fitted plot

  
**Figure 5:** Normal Q-Q plot of standardized residuals vs theoretical quantiles

(2) Consider the case-control file for the colorectal adenomas data set that is here:

<http://afodor.github.io/classes/stats2015/caseControlData.txt>

A separate file gives obesity (BMI) data for these same subjects:

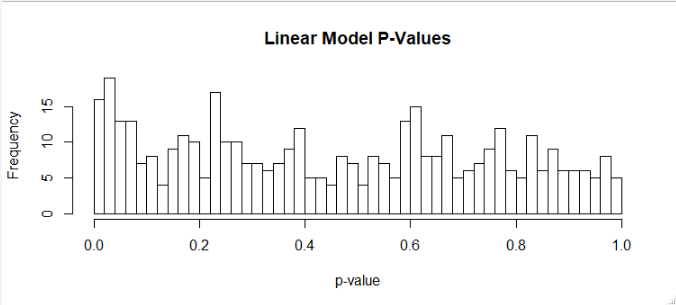
<http://afodor.github.io/classes/stats2015/BMI_Data.txt>

**For each OTU in the spreadsheet, generate a p-value from linear regression comparing BMI to the relative abundance of each OTU.**

**The R code at the following URL was utilized to generate the results for the following questions:** <https://github.com/jyoung67/advstatistics-labs/blob/master/labs/lab09/processMouseData.R>

**Graph out all the p-values.**

See **figure 6** for generated histogram of p-values.

  
**Figure 6:** Histogram of linear model p-values

**Do they appear uniformly distributed?**

Per figure 6, the p-values are not normally distributed.

**Does the microbial community appear to be influencing body weight in this cohort?**

Since the distribution of p-values is not normally distributed, the microbial community appears to be influencing body weight in this cohort.

**Are any of these associations significant at a 10% false discovery rate?**

Using the raw p-values, there were 68 significant values. However, there were no significant values using Bonferroni and BH/FDR adjustment methods (figure 7).

  
**Figure 7:** # of significant p-values using three different threshold methods