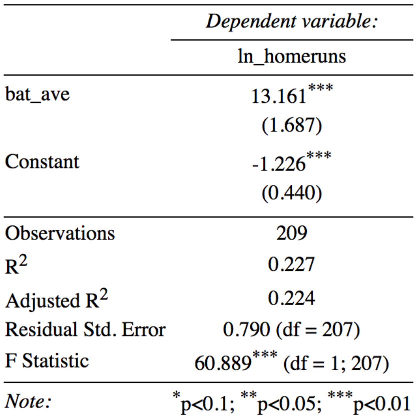
PART I:

1. ***Take the natural log of home runs, and perform a regression of ln home runs on batting average. Write down the regression equation and interpret the meaning of β_0 and β_1.***

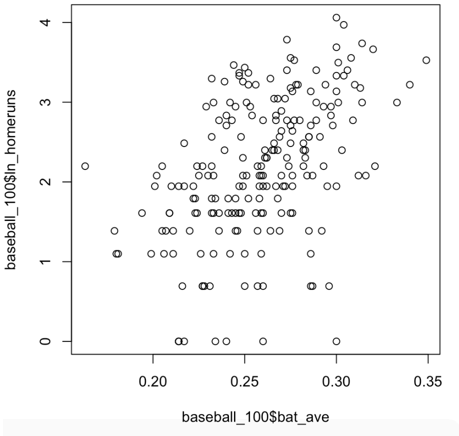


	Dependent variable:
	ln_homeruns
bat_ave	13.161***
	(1.687)
Constant	-1.226***
	(0.440)
Observations	209
R^2	0.227
Adjusted R^2	0.224
Residual Std. Error	0.790 (df = 207)
F Statistic	60.889*** (df = 1; 207)
Note:	*p<0.1; **p<0.05; ***p<0.01

The regression used is: *ln (homerun + 1) = -1.226 + 13.161(bat_ave)*. The reason why we add the 1 to the ln is because ln 0 is not determined.

This means that for players that have at least 100 bats, when the batting average increases by 1, the ln + 1 would increase by 13.161. Furthermore, if the batting average is 0 then the ln homeruns + 1 is equal to -1.226.

2. ***Does a linear relationship between ln home runs and batting average exist? Why?***



When looking of the outcome of the regression we can observe that the p-value of the regression is 2.94 x 10-13, a value small enough to reject the null hypothesis which by default is that there is no linear relationship between the variables. By rejecting it we can assume that there is enough evidence to point towards a linear regression.

Furthermore, if we look at the plot we can observe a linear correlation, even though it has a high variance it seems linear.

3. ***Based on the regression result above, estimate the number of home runs for a player with batting average of 0.45 in 2002.***

If the batting average is 0.45,

ln (homerun + 1) = -1.226 + 13.161(0.45) = 4.6964

$$e^{\ln(homeruns+1)} = e^{4.6964}$$

$$homeruns = e^{4.6964} - 1 = 108.55$$

Therefore, by rounding up, the number of homeruns will be approximately 109 if the batting average is 0.45. However, one must be careful when using this result as it is fruit of extrapolation since the highest batting in the dataset is 0.349, and we do not know what happens after. The type of regression might change and this model might not be useful.

PART II:

4. ***Build the best multiple regression model you can for the purpose of estimating calories. You can use as many variables as the predictors. You may want to compare and contrast the results from the forward selection, backward elimination, stepwise variable selection, and best subsets procedures.***

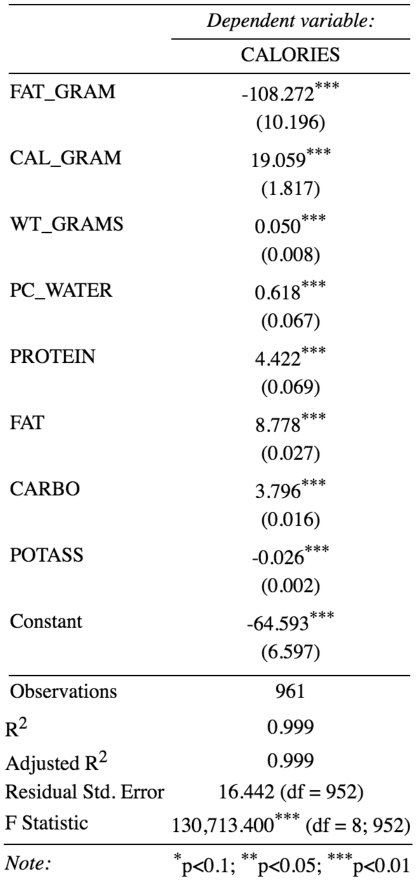
To determine what model is the best, we built 4, one for each procedure:

- CALORIES ~ CARBO + FAT + PROTEIN + POTASS + WT_GRAMS + THIAMIN + IRON + PHOSPHOR + RIBOFLAV + CALCIUM + POLUNSAT (Forward Selection)
- CALORIES ~ WT_GRAMS + PC_WATER + PROTEIN + FAT + CARBO + CALCIUM + PHOSPHOR + IRON + POTASS + SODIUM + THIAMIN + CAL_GRAM + IRN_GRAM + PRO_GRAM + FAT_GRAM (Backward elimination)
- CALORIES ~ WT_GRAMS + PROTEIN + FAT + CARBO + PHOSPHOR + IRON + POTASS + SODIUM + THIAMIN + RIBOFLAV + CALCIUM (Stepwise variable)
- CALORIES ~ FAT_GRAM + CAL_GRAM + WT_GRAMS + PC_WATER + PROTEIN + FAT + CARBO + POTASS (Best subset)

After having built the models, their p-values were compared and the smallest one would have been interpreted as the best one but as they all had the same one, I picked the best subset model as it has the least number of variables, thus less chance of overfitting but is as statistically significant.

5. ***Write down the regression equation and interpret the results. Make sure to explain if each coefficient has a linear relationship with calories.***

The chosen model is the following: CALORIES ~ FAT_GRAM + CAL_GRAM + WT_GRAMS + PC_WATER + PROTEIN + FAT + CARBO + POTASS. It was determined with the best subset procedures. The results of the regression are the following:



	Dependent variable:
	CALORIES
FAT_GRAM	-108.272***
	(10.196)
CAL_GRAM	19.059***
	(1.817)
WT_GRAMS	0.050***
	(0.008)
PC_WATER	0.618***
	(0.067)
PROTEIN	4.422***
	(0.069)
FAT	8.778***
	(0.027)
CARBO	3.796***
	(0.016)
POTASS	-0.026***
	(0.002)
Constant	-64.593***
	(6.597)
Observations	961
R^2	0.999
Adjusted R^2	0.999
Residual Std. Error	16.442 (df = 952)
F Statistic	130,713.400*** (df = 8; 952)
Note:	*p<0.1; **p<0.05; ***p<0.01

From this we can see that FAT_GRAMS and POTASS have a negative correlation. This means that as the number of grams of those two factors goes up, then the number of calories goes down. However, all the other factors are positively correlated. This means that as they go up then so does calories. Finally, the constant tells us that if there is nothing present the number of calories will probably be 0 as the number of calories cannot be negative and the constant or y-intersection is -64.593

Furthermore, all variables are statistically significant as their p-values allow us to reject the null hypothesis.

6. ***Does multicollinearity exist when building the model for question 4? If it exists, what should we do?***

Multicollinearity occurs when one variable can be lineal predicted using another variable. It creates a big problem as two predictors are included but they do not all value to the model and can lead to overfitting. This is why if multicollinearity exists we must remove one of the variables.

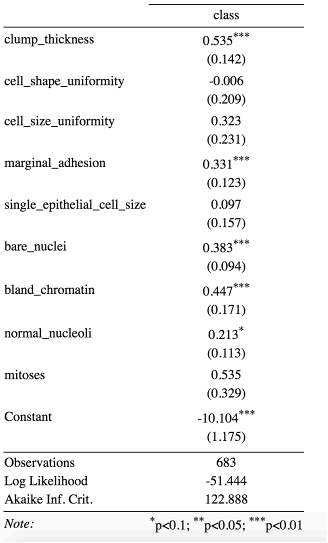
In this dataset, there is a huge amount of multicollinearity. When checking for the variance inflation factor, (if VIF is greater than 4 then there is multicollinearity) with all variables included in the model, only one variable, cholesterol passes with a VIF of 2.92.

Focusing more specifically on our model, the VIF values are better but still not good. PROTEIN, FAT and POTASS have no multicollinearity but the other values have VIF > than 4. Some by more than other as it ranges from 43.9 with CAL_GRAMS to 5.91 with CARBO.

To make the model better, we would have to analyse where this multicollinearity comes from using a panels and see if it all comes from one variable and remove or if it come from various variables and then decided which ones not to include in the model.

PART III:

7. Build a regression model that estimate the class of breast cancer based on all predictors available in the dataset (9 in total).



	class
clump_thickness	0.535***
	(0.142)
cell_shape_uniformity	-0.006
	(0.209)
cell_size_uniformity	0.323
	(0.231)
marginal_adhesion	0.331***
	(0.123)
single_epithelial_cell_size	0.097
	(0.157)
bare_nuclei	0.383***
	(0.094)
bland_chromatin	0.447***
	(0.171)
normal_nucleoli	0.213*
	(0.113)
mitoses	0.535
	(0.329)
Constant	-10.104***
	(1.175)
Observations	683
Log Likelihood	-51.444
Akaike Inf. Crit.	122.888
Note:	*p<0.1; **p<0.05; ***p<0.01

The above result is the result of the regression. The values represent the coefficient of the variable and the first one is the intercept.

8. Is the overall regression model built in question 7 statistically significant? How can you tell?

Using the lrtest function in R we tried to determine whether or not the model is significant. The p-value obtained being < 2.2e^{-16} we can reject the null hypothesis that states that the model is not significant, and we have enough evidence to assume our model is significant.

9. Which variables are significant predictors of breast tumor class? Why?

From the image above, one can see that the significant variables are clump thickness, marginal addition, bare nuclei and bland chromatid. All of these values have a p-value under 0.05 which means that for all of them we can reject the null hypothesis and assume they are significant.

10. Should we include variables that are not significant predictors in the model? Why/Why not?

As most questions, it can be answered with it depends. On one hand, we must determine what is significant. A p-value of 0.05 is a value we chose. It is not a fixed rule. Furthermore, as this sample size is relatively small, p-value will tend to be a little higher. We could therefore accept some variables such as normal nucleoli or mitosis.

However, we must be careful, when including extra variables. Too many variables could lead to overfitting and therefore the model could not be applied to another data set and ruins the purpose of the model.