# Probability and Statistical Inference Coursework

**By Daniel Wilkinson**

# Part 1: Plasma Ferritin Concentration Study

## Introduction

In this part I will be analysing the relationship of Plasma Ferritin Concentration levels in a sample of 202 Australian athletes with multiple other variables such as sport type, BMI and Hemoglobin levels. The relationship between variables will be tested through fitting a regression model and using regression analysis to determine correlations and normality of the data. Ferritin is a protein that stores and releases iron and Plasma Ferritin is the total amount iron stored in the body[1] , a normal amount of iron is healthy , however low levels indicate iron deficiency , which can lead to iron deficiency anemia , which is a decrease in red blood cell count and low Hemoglobin levels[2] , which are both variables in the data.

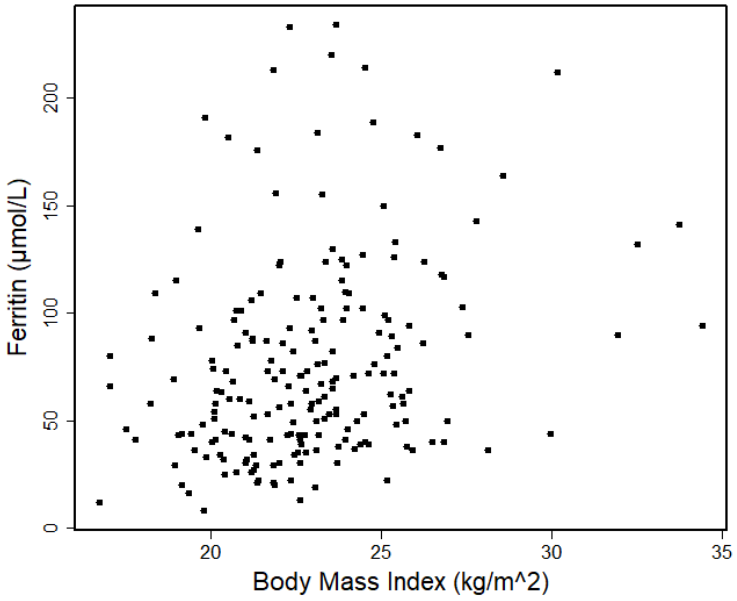
## Task 1:

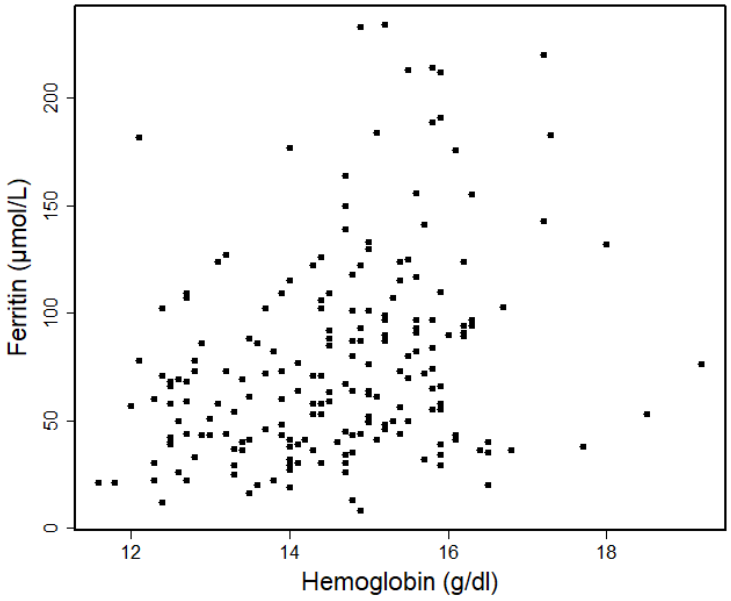
**(a)** Table of summary statistics derived from the sports data csv file (see Task 1 (a) in R script). (N=202)

Sport and Sex are not included in the table as they are not continuous variables. There are 100 female athletes and 102 male athletes.

| Variable | Mean | Standard Deviation | Minimum | Median | Maximum |
|---|---|---|---|---|---|
| **Plasma Ferritin Concentration (Ferr) (µmol/L)** | 76.88 | 47.501 | 8 | 65.50 | 234 |
| **Lean Body Mass (LBM) (%)** | 64.87 | 13.070 | 34.36 | 63.03 | 106.00 |
| **Red cell count (RCC) (cells/µL)** | 4.719 | 0.458 | 3.800 | 4.755 | 6.720 |
| **White cell count (WCC) (cells/µL)** | 7.109 | 1.800 | 3.300 | 6.850 | 14.300 |
| **Hematocrit (Hc) (%)** | 43.09 | 3.663 | 35.90 | 43.50 | 59.70 |
| **Hemoglobin (Hg) (g/dl)** | 14.57 | 1.362 | 11.60 | 14.70 | 19.20 |
| **Body Mass Index (BMI) ($kg/m^2$)** | 22.96 | 2.864 | 16.75 | 22.72 | 34.42 |
| **Sum of Skin Folds (SSF) (mm)** | 69.02 | 32.565 | 28.00 | 58.60 | 200.80 |
| **Body Fat Percentage (Bfat) (%)** | 13.507 | 6.190 | 5.630 | 11.650 | 35.520 |

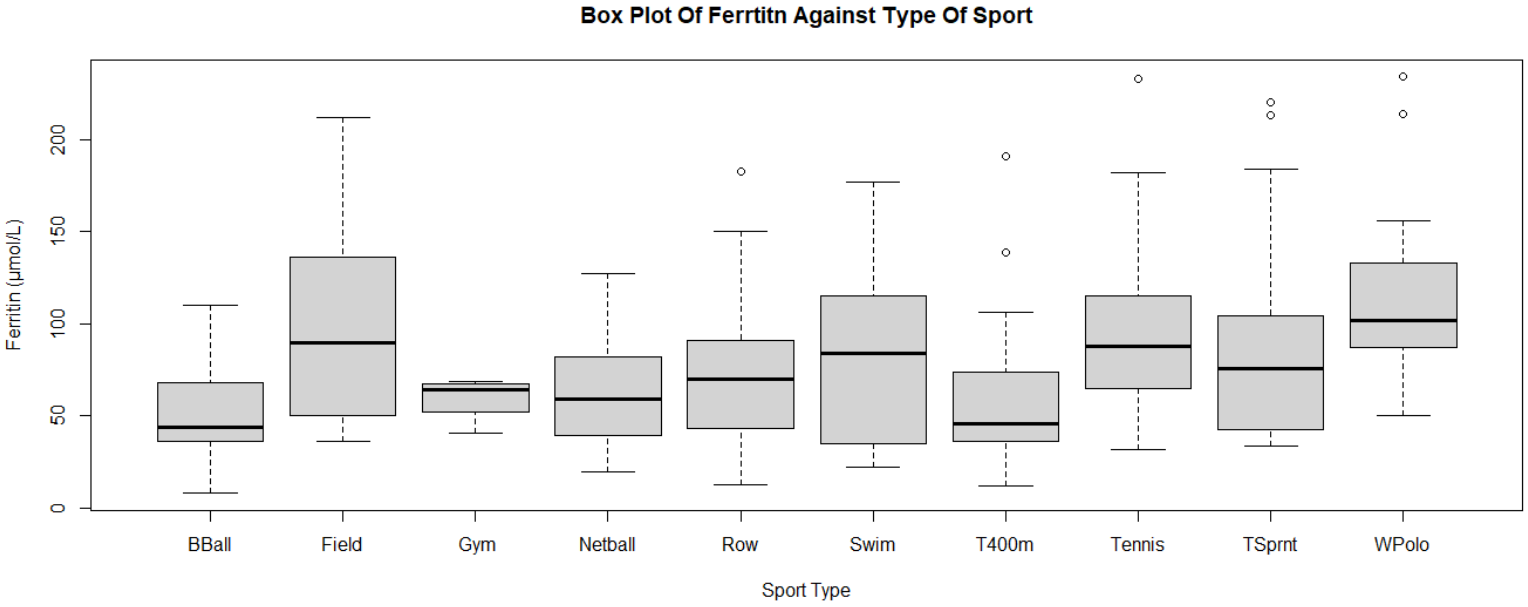
The LBM mean of 64.87% is in the standard range of 60%-90%.[3] which shows that the average athlete in this sample has a healthy mass to body weight ratio, and the standard deviation is relatively small meaning most athletes LBM are centred around the mean. RCC, WCC, Hc and Hg are in the standard range for both male and female [4] , which shows that most of these athletes are healthy , also the mean BMI of 22.96 lies in the healthy range of 18-25 which also supports this .Also these variables mentioned so far have small standard deviations showing the possibility of little to no outliers in the data to be analysed as the majority of data lies close to the mean. Also, the medians are close to the mean which shows that the data distribution is not skewed, this is evidence supporting the distribution of data to be normal, however other tests must be done to confirm this. SSF has the largest standard deviation showing the data is more spread out than the others.



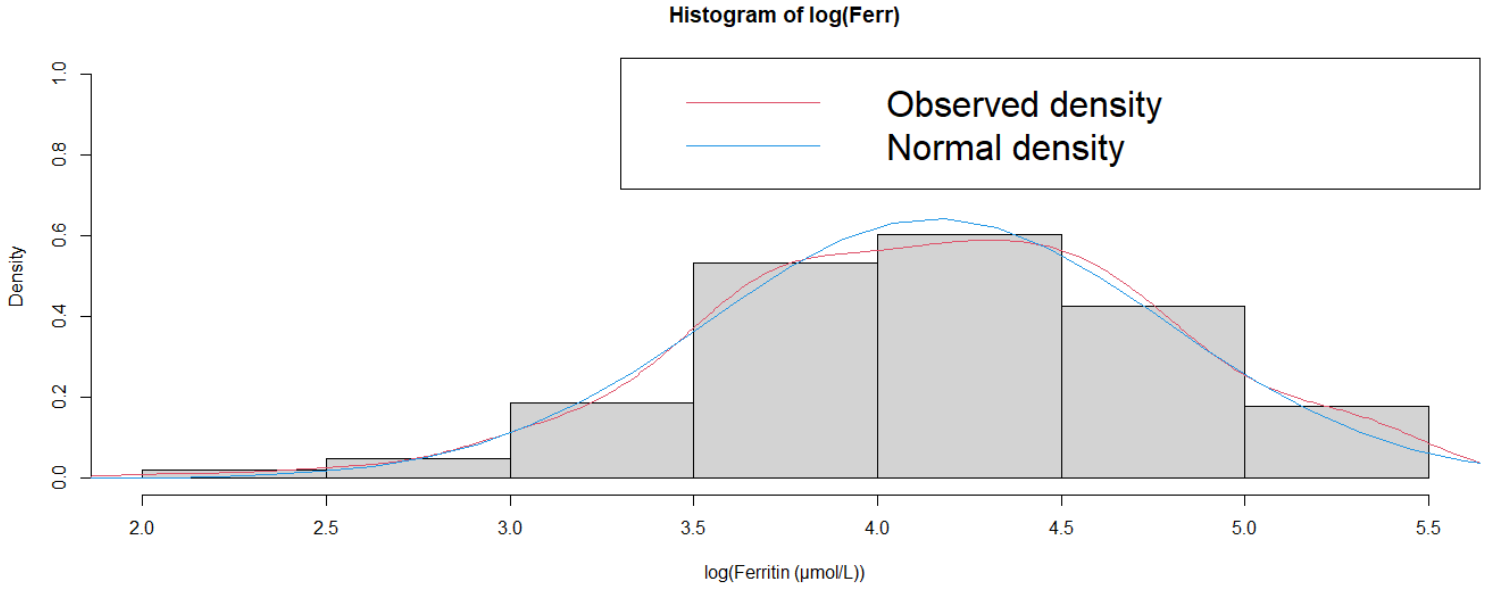

These two scatter plots show Ferritin Concentration against the predictors BMI and Hg. The distribution of data from looking at the plots seems to be very similar with most data in the lower left corner around a Ferritin concentration of 55 µmol/L. There are some areas where data is sparse for example the right-hand side, however

there is still enough data there to not consider them as outliers. These plots suggest a strong positive correlation between Ferritin concentration and BMI/Hg.

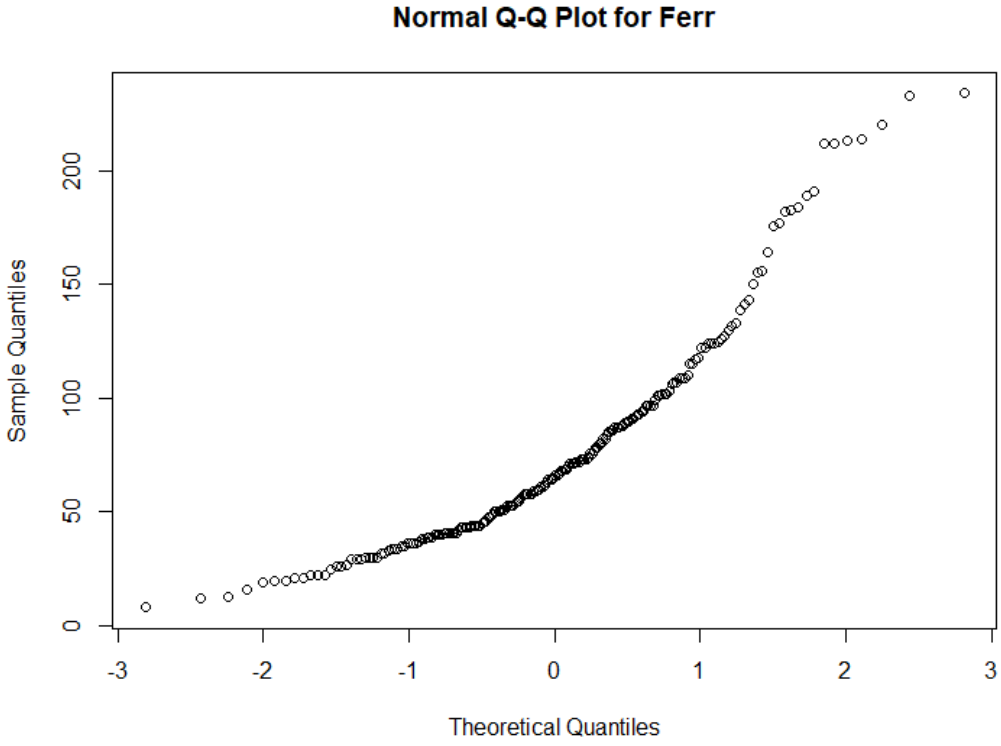



From the box plot we can see that athletes who play water polo have the highest average Ferritin concentration and BBall and T400m have the lowest, this suggests that iron in the system is depleted at a slower rate for water polo and there could be less risk of iron deficiency. Field has the largest standard deviation/spread and Row and Tennis have outlying data for high Ferritin concentration.

**(b)**



I logged the Ferritin values so that the density curve could be analysed more efficiently. The observed density follows the normal density curve very well but has a lower peak shifted to the right. This is strong evidence that the Ferritin data is normally distributed.




The normal Q-Q plot data somewhat follows a straight line but has a right skew, this is also evidence for the Ferritin data being normally distributed but will need to be transformed later.

## Task 2:

The dataset has now been randomly split into two sets: Training and Testing, containing 141 and 61 athletes respectively (see Task 2 in R script), this task uses the Training Dataset.

### (a)

The regression equation is:

$$Ferr = \beta_0 + Sex\beta_1 + Sport\beta_2 + LBM\beta_3 + RCC\beta_4 + WCC\beta_5 + Hc\beta_6 + Hg\beta_7 + BMI\beta_8 + SSF\beta_9 + Bfat\beta_{10} + \epsilon$$

### (b)

Most of the work for this was done in R (see Task 2 (b)), the regression equation in part (a) was fitted in R and the summary shows only Sex and one sports type as significant predictors(P-value<0.1). To tidy this up I identified insignificant sport types and merged them together to obtain a model that wasn't as overfitted. The new model gives three significant predictors: Sex, Lean Body Mass and BMI.

I then refit the model using only these three predictors as variables as the rest were insignificant. These variables all have a p-value of <0.01, which is below the significance level of 0.05.

To test if the full model is better than this smaller model I will perform an F-test in R.

We test the hypothesis:

$$H_0: \beta_0 = \beta_1 = \beta_2 = \cdots = \beta_{10} = 0 \ (All\ coefficients\ equal\ 0)$$

$$H_1: \beta_0 \neq 0\ or\ \beta_1 \neq 0\ or\ \beta_2 \neq 0 \ldots (Some\ coefficients\ not\ 0)$$

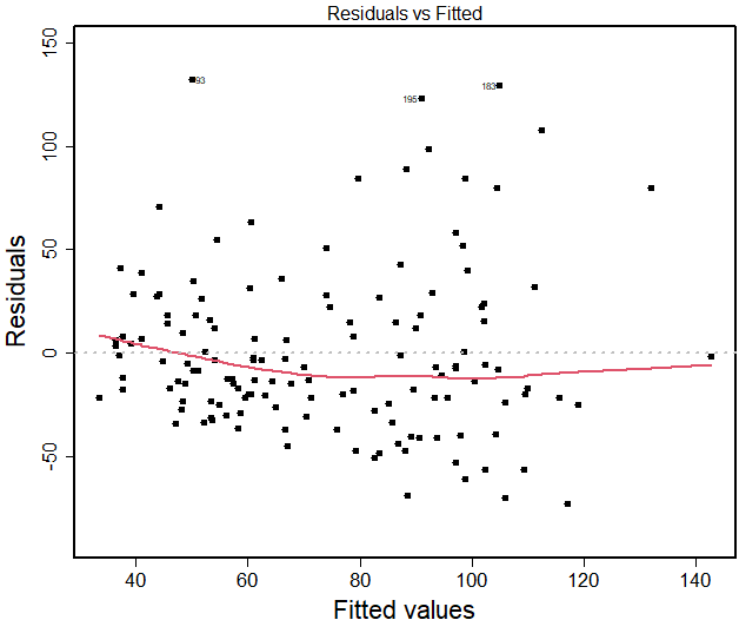
Using analysis of variance between the two models we obtain an $F_{calc}$ value of 0.7418. The critical value $F_{0.05,10,127}$ is roughly 3.9. $F_{calc} < F_{0.05,10,127}$ , therefore, we accept the null hypothesis $H_0$ that the coefficients are 0 showing that the smaller model is better than the original model.

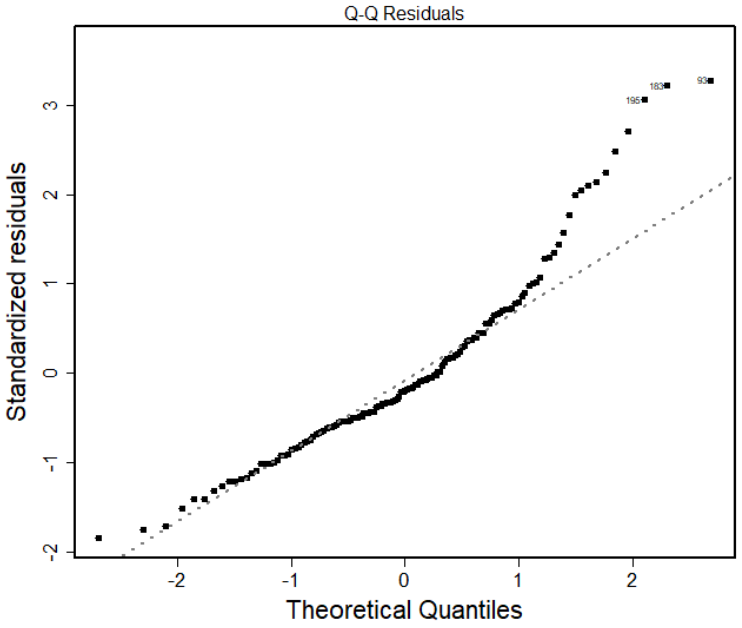
To further support this claim, I used the Akaike and Bayesian Information Criterion tests in R to receive scores for each model. The original model has an AIC score of 1462.276 and BIC score of 1506.508. The smaller model has an AIC score of 1450.281 and BIC score of 1465.025. Both values for the smaller model are lower than the values for the original model, which shows that the smaller model has a better fit and is the one to be used.

## (c)

## i.

The multiple R squared value for the model is 0.2642, which shows only roughly 26% variation of ferritin values can be explained by the model, however the r value obtained by square rooting this number gives 0.514, which shows a moderate positive correlation between Ferr and the predictor variables. I plotted the model in R to obtain the following graphs:



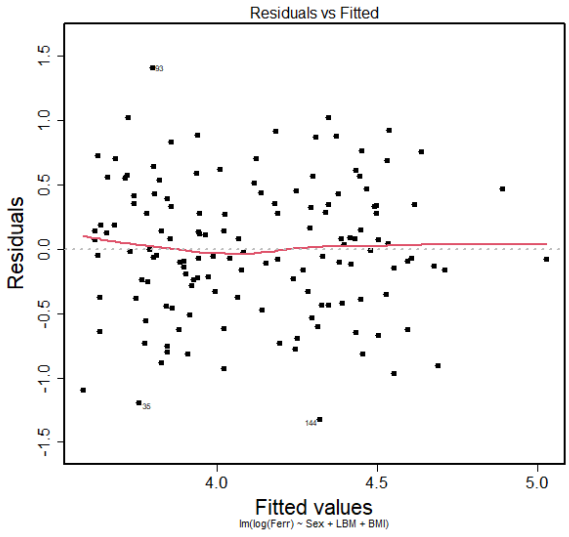



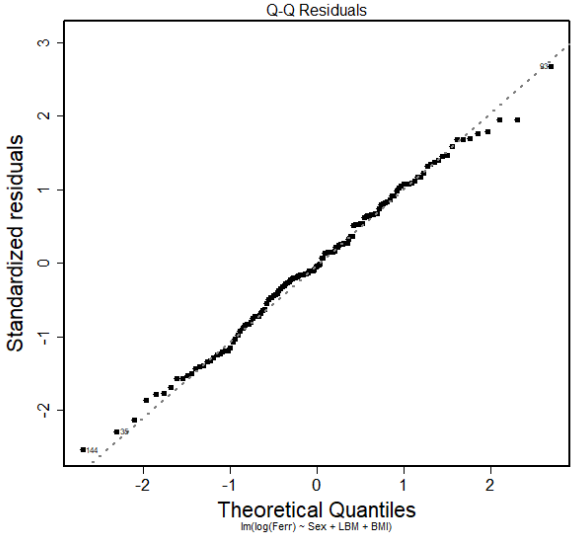

The residual graph on the left does not show any pattern in the residuals themselves, showing that there is a constant variance within the model, however the red line indicates a non-linear relationship. The Q-Q plot on the right does not fit the line y=x, therefore shows a non-normal structure. To further test the normality of the data I performed the Shapiro-Wilk test in R with the null hypothesis that the residuals are normally distributed. The p-value obtained was $1.492 \times 10^{-6}$, which is much lower than the significance level of 0.05. Therefore, the hypothesis is rejected showing the residuals are not normally distributed.

To improve the model, I transformed the response variable by logging the Plasma Ferritin values. Hence the regression equation is now:

$$\ln(Ferr) = \beta_0 + Sex\beta_1 + LBM\beta_2 + BMI\beta_3 + \epsilon$$

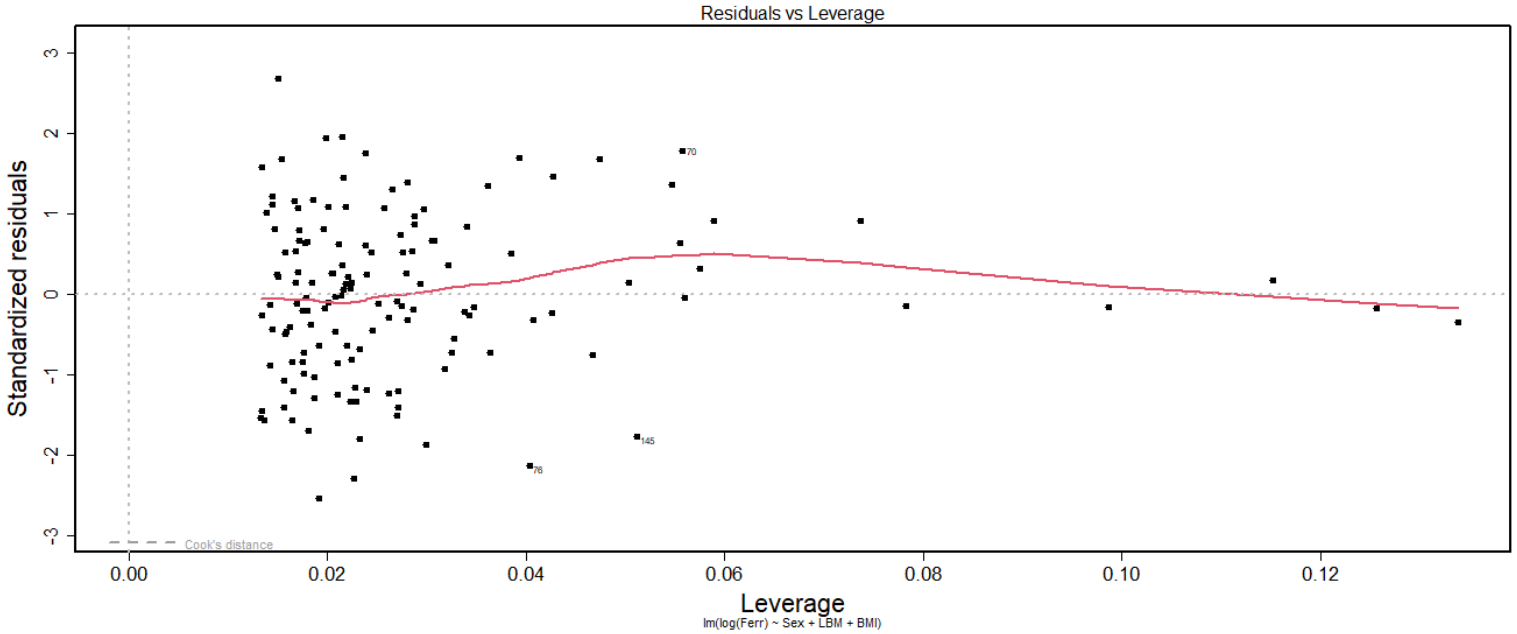
The multiple R squared, and r values are very similar to the previous model, however the graphs are different:



The residual graph on the left shows a much more linear relationship and shows a constant variance across the residuals, the Q-Q plot residuals follow the line y=x much more closely now , suggesting the data is now normally distributed , I performed a Shapiro-Wilk test and this time I obtained a p-value of 0.8925 , which is way above the significance level therefore the null hypothesis is accepted showing the data is normally distributed.

**ii.**




In the graph above there are three flagged observations identifying them as outliers in the leverage range of 0.04 to 0.06, for observations 70, 76 and 145. There are also roughly 6 observations that have large leverage; however, these are close to the line y=0, therefore these data have low potential influence on the data. The three identified outliers are far away from y=0 and have moderate leverage, therefore these data can be considered influential points. I would deal with this by removing the three outliers from the data.

## (d)

Looking at the summary of the TransformTraining.model in R, the predictor with largest effect on Ferr values is Sex as the p-value ($8.21 \times 10^{-7}$) is much smaller than the p-values for LBM and BMI. This shows that gender is a key factor in influencing plasma ferritin concentration , a possible reason for this is that female athletes may have been subject to menstrual blood loss at the time , and with a loss of blood would also lower the plasma ferritin concentration , meaning women would have a significantly lower plasma ferritin concentration than men.

The lean body mass of a person is equal to the body weight of a person minus their body fat expressed as a percentage, therefore the larger the percentage the healthier the person is as it shows they have only small amounts of body fat. The estimated coefficient for lean

body mass is -0.022677, which is negative, and is therefore negatively correlated. This shows that the lower the LBM, the greater the Ferr levels of the person.

Finally the effect of BMI on Ferr is quite similar to LBM, as these variables are positively correlated , the higher BMI values (overweight/obese) indicates a higher Ferr level.

## Task 3:

In R (see task 3) I created a new column in the testing dataset for ln (Ferr), as when I predict the Ferritin values in the testing dataset the model I used has had Ferr transformed into ln (Ferr) already. The values in the new column can then be compared against the predicted values in the fit column output in the R console. From observation we can see that the values are within the lower and upper bounds of the prediction and are less than ~1.5 units off the actual values, this shows the model is quite accurate at predicting values from other data sources. I then found the Root Mean Square Error (RMSE) between the data and obtained a value of 1.038518, showing the error deviation from the actual values, this shows the model is accurate and is suitable for future use.

## Part 2: Bayesian Inference

### (a)

From the question we obtain the likelihood and prior distribution functions. The likelihood function is $f(x|\theta) = x \sim N(\theta, 4)$ and the prior is $f(\theta) \sim N(12,9)$ which is a conjugate prior also following a normal distribution. And the value transmitted is x=13.25.

We set the values:

$$\mu = \theta\ , \sigma^2 = 4$$

$$\mu_0 = 12\ , \sigma_0^2 = 9$$

The posterior function can be found by the following equation:

$$f(\theta|x) \propto f(\theta) f(x|\theta)$$

Which gives:

$$f(\theta|x) \propto \exp\left\{-\frac{1}{2\sigma_0^2}(\theta - \mu_0)^2\right\} \exp\left\{-\frac{1}{2\sigma^2}(x - \theta)^2\right\}$$

This can be simplified so that all terms are within one exponential bracket, comparing these with the normal probability distribution function we obtain formulas for the mean $(\mu_1)$ and variance$(\sigma_1^2)$ of the posterior function:

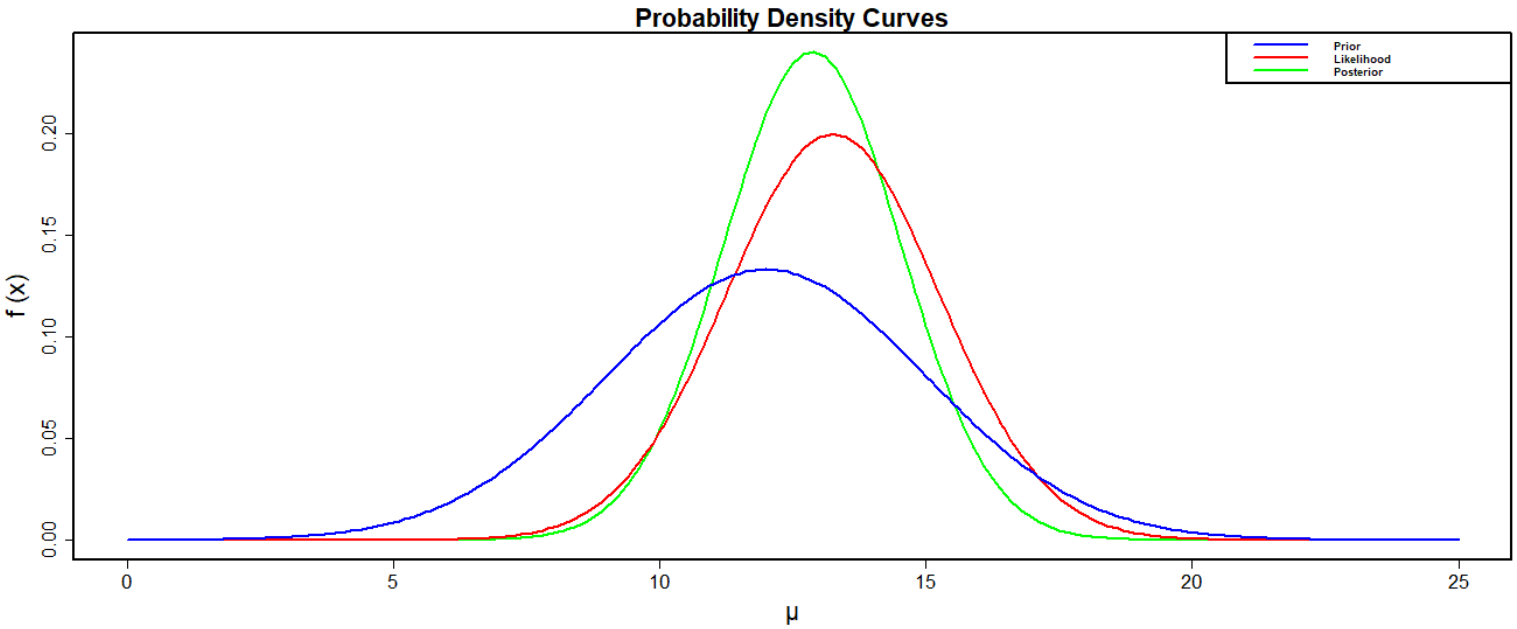
$$\mu_1 = \frac{\mu_0 \sigma^2 + x\sigma_0^2}{\sigma^2 + \sigma_0^2}$$

$$\sigma_1^2 = \frac{\sigma^2 \sigma_0^2}{\sigma^2 + \sigma_0^2}$$

Substituting these values in we get $\mu_1$ = 12.865 and $\sigma_1^2$ = 2.769 both to 3dp. The posterior distribution function therefore is:

$$f(\theta|x) \sim N(12.865, 2.769)$$

## (b)




I plotted the distributions curves in R (see part 2 (b)) , my belief about theta is that the number should have a greater value as the peak of the prior distribution curve is shifted to the left of the other curves.

## (c)

Now we receive values $x_1, \ldots, x_n$ n times with sample mean $\bar{X}$. The equation for the posterior function now changes to:

$$f(\theta|x) \propto \exp\left\{-\frac{1}{2\sigma_0^2}(\theta-\mu_0)^2\right\}\exp\left\{-\frac{1}{2}\left(\frac{\bar{X}-\theta}{\sigma/\sqrt{n}}\right)^2\right\}$$

This is simplified to find the general formulas for mean and variance for the posterior function of the mean in a normal population:

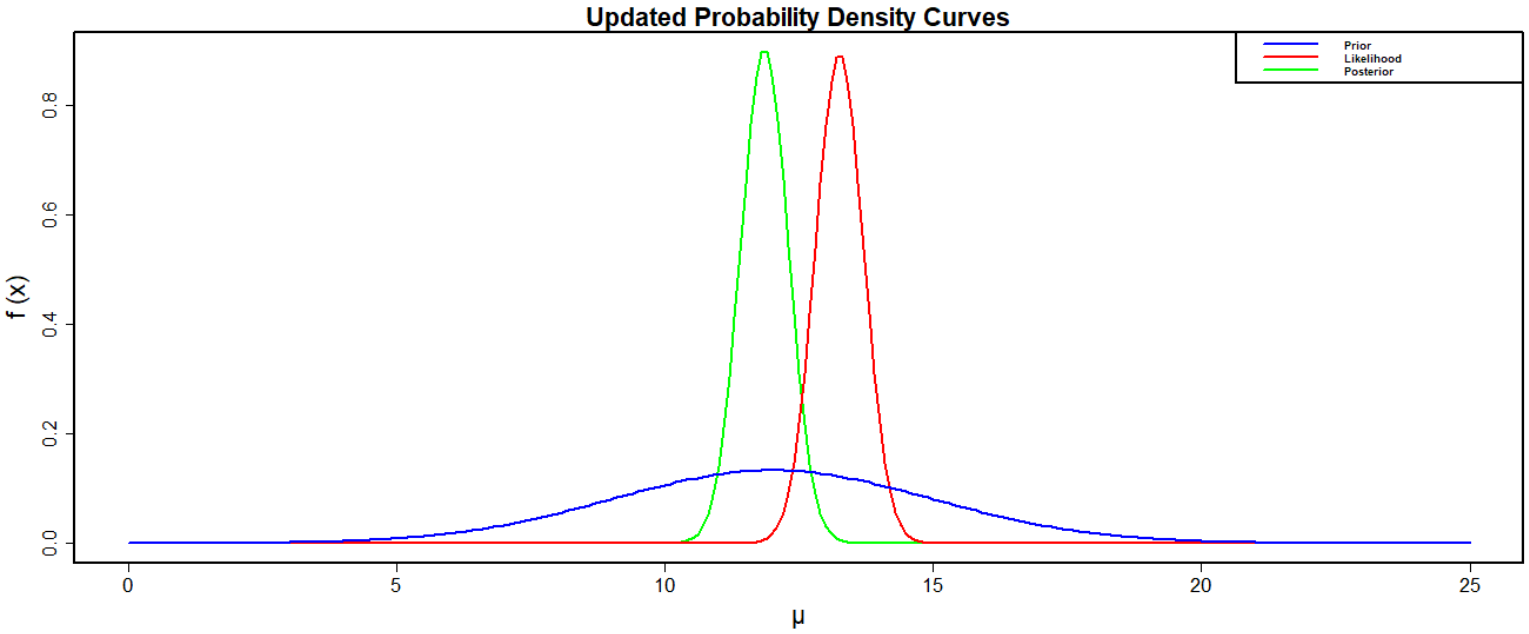
$$\mu_1 = \frac{\mu_0\sigma^2 + n\sigma_0^2\bar{X}}{\sigma^2 + n\sigma_0^2}$$

$$\sigma_1^2 = \frac{\sigma^2\sigma_0^2}{\sigma^2 + n\sigma_0^2}$$

## (d)

Given n =20 and $\overline{X}$ = 11.85, we can use the formulas in part c with the same $\sigma^2, \sigma_0^2$ and $\mu_0$ to obtain an updated posterior distribution for $\theta$ . Therefore, we get the values $\mu_1$ = 11.853 and $\sigma_1^2$ = 0.196 both to 3dp. The posterior distribution function therefore is:

$$f(\theta|x) \sim N(11.853, 0.196)$$




The distributions were replotted in R with the new posterior function and new likelihood function as the standard deviation value became $\frac{\sigma}{\sqrt{n}}$ . The new posterior distribution has a lower mean than the previous and is very close to the sample mean and a very small standard deviation with most values cantered around the mean. Therefore, the true value of theta is roughly 12 as the prior distribution predicted.

## (e)

Take the formulae for $\mu_1$ and $\sigma_1^2$ in part c. If more data is received i.e. n is greater, whilst keeping the sample mean $\overline{X}$ the same, the value of $\mu_1$ becomes closer to the value of $\overline{X}$ and the value of $\sigma_1^2$ becomes lower. Therefore, the more signals received gives a greater approximation of the true value of theta as the variance becomes incredibly small.

# References


[1] "Ferritin," [Online]. Available: https://en.wikipedia.org/wiki/Ferritin. [Accessed 25 March 2019].

[2] "Anemia," [Online]. Available: https://en.wikipedia.org/wiki/Iron-deficiency_anemia. [Accessed 25 March 2019].

[3] "Lean Body Mass," [Online]. Available: https://www.livestrong.com/article/175858-the-average-lean-body-mass/. [Accessed 22 March 2019].

[4] "Test Procedures," [Online]. Available: https://www.mayoclinic.org/tests-procedures/complete-blood-count/about/pac-20384919. [Accessed 22 March 2019].