Sport's Influence on Athletes' Physical Markers

SUPPLEMENTARY DOCUMENT

Student Number: 2938740
Statistics for Data Science
25th November 2020

1. EXPLORATORY DATA ANALYSIS

1.1. ATHLETE BREAKDOWN

The dataset is sex balanced with 102 male and 100 female athletes however we see from the analysis of athletes that the number of athletes is not uniformly distributed by sport. Three sports are only represented by one sex – gymnastics (female only), netball (female only) and water polo (men only).

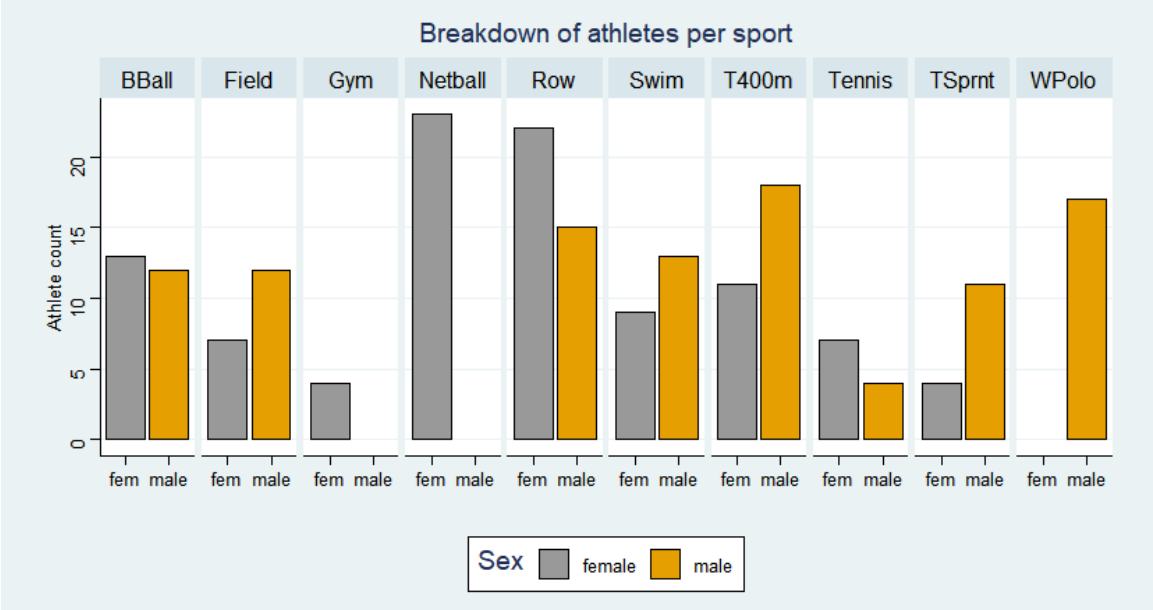


Figure 1: Count of athletes by Sport and Sex

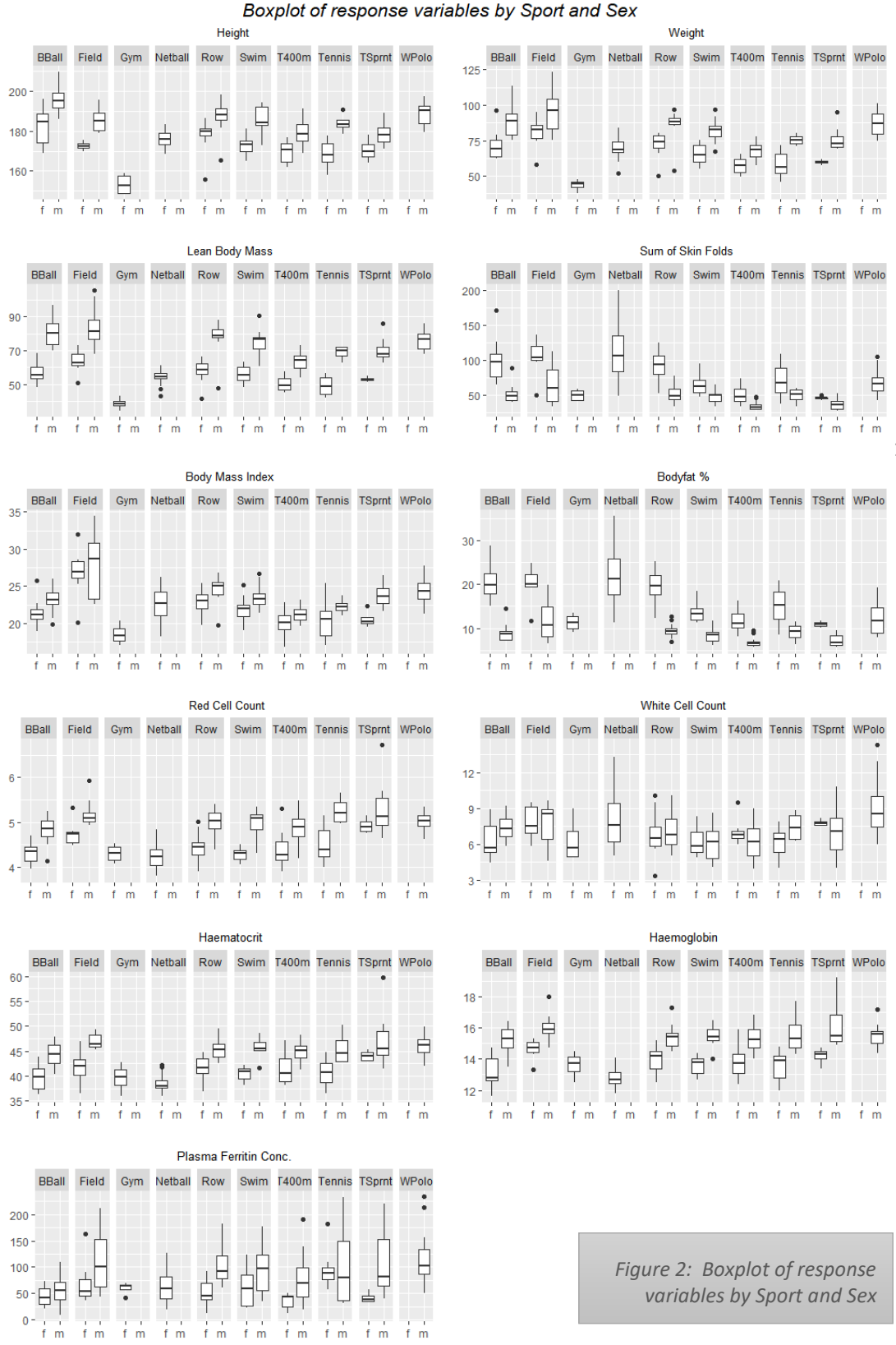
1.2. BOXPLOTS

The boxplots show differing means and variance between the Sex, Sport and Sex&Sport factors. There are several outliers within the data however as we are dealing with athletes one could reasonably argue that extreme values are not unexpected.

The physical markers appear to vary markedly not only between the sexes but also inter-sport. The non-physical markers appear more stable yet show differences between sexes apart from iron ferritin (Ferr) which contains a lot of variability.

The White Cell Counts (WCC) is the only response variable that looks to have any consistency across the dataset.

The boxplots are shown overleaf as figure 2. They can be re-sized for ease of viewing.



1

Figure 2: Boxplot of response variables by Sport and Sex

1.3. FIVE NUMBER SUMMARY (PLUS MEAN & SD)

The five-number summary plus mean and standard deviation was generated for the full dataset to highlight the variables showing a lot of variation and extreme values for the min and max. The data is to 2 decimal places for all entries.

Biomarker	Min	LQ	Median	UQ	Max	Mean	SD
RCC	3.80	4.37	4.75	5.03	6.72	4.72	0.46
WCC	3.30	5.90	6.85	8.30	14.30	7.11	1.80
Hc	35.90	40.55	43.50	45.60	59.70	43.09	3.66
Hg	11.60	13.50	14.70	15.60	19.20	14.57	1.36
Ferr	8.00	41.00	65.50	97.50	234.00	76.90	47.50
BMI	16.75	21.06	22.72	24.48	34.42	22.96	2.86
SSF	28.00	43.80	58.60	90.60	200.80	69.00	32.60
Bfat	5.63	8.53	11.65	18.09	35.52	13.51	6.19
LBM	34.40	54.60	63.00	75.00	106.00	64.90	13.10
Ht	148.90	174.00	179.70	186.22	209.40	180.10	9.73
Wt	37.80	66.50	74.40	84.30	123.20	75.00	13.9

Figure 3: Five number summary at a dataset level (including mean and Standard deviation)

Using this table in conjunction with the boxplots in figure 2 shows that the iron ferritin (Ferr) levels contain a large amount of variance and it can be seen that the female athletes typically occupy the lower end of the response variable with male athletes showing high degrees of upper tail variance along with a number of outlying data points.

A custom function can be found in the accompanying R cookbook should you wish to compute a deeper analysis of the five-number summary (plus mean and standard deviation) at any level of the data.

Example of how to use the function in isolation:

```
x <- data$Ferr[data$Sex=="male"]
SevenNumSum(x)
```

1.4. OUTLIER ANALYSIS

The previous summaries have identified outlying values that may or may not affect the subsequent hypothesis testing. The parameters for outlying values are given below.

Lower Extreme: *Lower Quartile - 3 * Inter-Quartile Range*

Lower Mild: *Lower Quartile - 1.5 * Inter-Quartile Range*

Upper Mild: *Upper Quartile + 1.5 * Inter-Quartile Range*

Upper Extreme: *Upper Quartile + 3 * Inter-Quartile Range*

This table represents the total dataset but can be recreated for each factor with a custom function in the cookbook by simply filtering the input.

Biomarker	Lower Extreme	Lower Mild	Lower Quartile	IQR	Upper Quartile	Upper Mild	Upper Extreme
RCC	2.37	3.37	4.37	0.67	5.03	6.03	7.03
WCC	-1.30	2.30	5.90	2.40	8.30	11.19	15.5
Hc	25.40	32.98	40.55	5.05	45.60	53.18	60.75
Hg	7.20	10.35	13.50	2.10	15.60	18.75	21.90
Ferr	-128.50	-43.75	41.00	56.50	97.50	182.25	267.00
BMI	10.81	15.94	21.06	3.42	24.48	29.61	34.73
SSF	-96.45	-26.33	43.8	46.75	90.55	160.68	230.80
Bfat	-20.16	-5.81	8.53	9.56	18.10	32.44	46.78
LBM	-6.54	24.04	56.62	20.39	75.00	105.58	136.16
Ht	137.33	155.66	174.00	12.23	186.23	204.56	222.90
Wt	12.93	39.70	66.48	17.85	84.33	111.10	137.88

Figure 4: Table of outlying value boundaries for the total dataset

Outliers based on full dataset: Range (Count)

Biomarker	Low Extreme	Low Mild	Upper Mild	Upper Extreme
RCC	-	-	6.72 (1)	-
WCC	-	-	12.7 – 14.3 (4)	-
Hc	-	-	59.3 (1)	-
Hg	-	-	19.2 (1)	-
Ferr	-	-	183 – 234 (11)	-
BMI	-	-	29.97 – 34.42 (7)	-
SSF	-	-	171.1 – 200.8 (3)	-
Bfat	-	-	35.52 (1)	-
LBM	-	-	106 (1)	-
Ht	-	148.9 – 149.0 (2)	209.4 (1)	-
Wt	-	37.8 (1)	111.3 – 123.2 (3)	-

Figure 5a: Outlying values of the full dataset

Given the large number of outliers it is prudent to break the outliers down by the sex factor to see if any insights could be gained. Some changes to the categorization of the outliers has been noticed.

Outliers for female athletes only: Range (Count)

Biomarker	Low Extreme	Low Mild	Upper Mild	Upper Extreme
RCC	-	-	5.16 – 5.33(4)	-
WCC	-	-	13.3 (1)	-
Hc	-	-	-	-
Hg	-	-	-	-
Ferr	-	-	164 (1)	182 (1)
BMI	-	-	28.57 – 31.93 (2)	-
SSF	-	-	200.8 (1)	-
Bfat	-	-	35.52 (1)	-
LBM	-	34.36 – 39.03 (3)	72.98 (1)	-
Ht	-	148.9 – 156.9 (4)	193.4 – 195.9 (2)	-
Wt	-	37.8 (1)	96.3 (1)	-

Figure 5b: Outlying values for female athletes only

Outliers for male athletes only: Range

Biomarker	Low Extreme	Low Mild	Upper Mild	Upper Extreme
RCC	-	4.13 – 4.32 (3)	5.69 – 5.93 (3)	6.72 (1)
WCC	-	-	12.7 – 14.3 (3)	-
Hc	-	-	-	59.7 (1)
Hg	-	-	17.7 – 18.5 (3)	19.2 (1)
Ferr	-	-	233 – 234 (2)	-
BMI	-	-	29.97 – 33.73 (5)	34.42 (1)
SSF	-	-	96.3 – 113.5 (6)	-
Bfat	-	-	14.98 – 19.17 (6)	19.94 (1)
LBM	-	48 (1)	102 – 106 (2)	-
Ht	-	-	209.4 (1)	-
Wt	-	-	123.2 (1)	-

Figure 5c: Outlying values for male athletes only

10 male athletes are responsible for 24 of the 40 male outlying values with 5 female athletes responsible for 10 of the 23 female outlying values. This leaves 16 outlying male athletes and 18 outlying female athletes, equivalent to 16% and 18% of each sex factor.

2. TWO TAIL, TWO SAMPLE T-TEST ON FACTOR: SEX

2.1. TEST ON FULL DATASET

The full dataset was testing for each response variable and then re-tested with outliers removed. The outliers that were removed were calculated at the full dataset level as the null hypothesis assumes that the data comes from a set of response variables with equal means.

The Central Limit Theorem (CLT) was applied to assume normality in conjunction with the robustness of the t-test. Welch's t-test is also considered more robust and does not assume that the data comes from samples with equal variance.

Testing is two tailed, significant at the 5% level with 95% confidence intervals shown:

Biomarker	Degrees Freedom	t-statistic	p-value	Lower CI	Upper CI	Mean Male	Mean Female
RCC	199.06	13.156	2.2e-16	0.529	0.715	5.027	4.405
WCC	**198.27**	**0.897**	**0.3709**	**-0.272**	**0.726**	**7.221**	**6.994**
Hc	199.66	14.141	2.2e-16	4.447	5.889	45.650	40.482
Hg	199.99	15.248	2.2e-16	1.735	2.251	15.553	13.560
Ferr	163.96	6.504	9.0e-10	27.468	51.416	96.402	56.960
BMI	199.85	5.031	1.08e-6	1.164	2.665	23.904	21.989
SSF	154.38	-9.196	2.4e-16	-43.187	-27.914	51.423	86.943
Bfat	158.87	-13.65	2.2e-16	-9.842	-7.354	9.251	17.849
LBM	181.07	16.482	2.2e-16	17.396	22.128	74.657	54.895
Ht	199.24	9.601	2.2e-16	8.671	13.153	185.506	174.594
Wt	197.71	9.238	2.2e-16	11.940	18.421	82.524	67.343

Figure 6a: Results of Welch's t-test on the sex factor

The tests find that for the WCC response variable we do not see sufficient evidence to reject the null hypothesis of the alternate hypothesis at the 5% significance level with P(t(0.897) > 0.05, 198.27df)) = 0.3709.

For all other response variables, we find sufficient evidence to reject the null hypothesis in favour of the alternate hypothesis at the 5% significance level with significant p-values as shown in figure 6 above.

Given the overwhelming rejection of the null hypothesis for 12/13 response variables, the test will be re-run with outlying data removed to see if this affects the results.

2.2. TEST ON REDUCED DATASET

The outliers have been removed from the dataset. Testing is two tailed as before and significant at the 5% level with 95% confidence intervals shown:

Biomarker	Degrees Freedom	t-statistic	p-value	Lower CI	Upper CI	Mean Male	Mean Female
RCC	171.31	11.883	2.2e-16	0.484	0.676	4.981	4.401
WCC	**170.14**	**-0.129**	**0.897**	**-0.504**	**0.442**	**6.897**	**6.928**
Hc	172.13	13.293	2.2e-16	4.089	5.516	45.276	40.473
Hg	172.69	13.937	2.2e-16	1.602	2.131	15.416	13.549
Ferr	162.36	4.659	6.58e-6	13.778	34.049	80.634	56.720
BMI	172.49	4.618	7.56e-6	0.862	2.150	23.365	21.858
SSF	145.5	-10.319	2.2e-16	-42.787	-29.031	48.632	84.541
Bfat	148.82	-14.918	2.2e-16	-9.979	-7.645	8.754	17.566
LBM	144.17	16.156	2.2e-16	16.442	21.026	73.720	54.985
Ht	167.94	9.4931	2.2e-16	8.618	13.144	185.795	174.914
Wt	164.71	9.0454	4.0e-16	10.827	16.874	80.944	67.093

Figure 6b: Results of Welch's t-test on the sex factor with outlying data removed

The tests find that for the WCC response variable we do not see sufficient evidence to reject the null hypothesis of the alternate hypothesis at the 5% significance level with P(t(0.897) > 0.05, 198.27df)) = 0.3709.

For all other response variables, we find sufficient evidence to reject the null hypothesis in favour of the alternate hypothesis at the 5% significance level with significant p-values as shown in figure 6 above.

Following the removal of outliers, the means of the WCC response variable for each sex are now very close and this is reflected by the very high p-value. The iron ferritin (Ferr) response factor also improved slightly but is still very firmly in favour of rejecting the null hypothesis.

3. ONE-WAY ANOVA TEST ON FACTOR: SPORT

3.1. TEST FOR HOMOGENEITY OF VARIANCE

The assumptions of the one-way ANOVA were checked before starting the test and it was found that all bar one response variables failed the test of homogeneity of variance for both the full dataset and with outliers removed. It was also noted that transforming the data had no effect. The results shown are for Levene's Test on the full untransformed dataset:

Biomarker	Df1	Df2	Statistic	p-value
RCC	9	192	2.36	0.0152
WCC	9	192	2.20	0.0238
Hc	**9**	**192**	**1.28**	**0.252**
Hg	9	192	2.29	0.0182
Ferr	9	192	2.42	0.0128
BMI	9	192	4.70	1.2e-5
SSF	9	192	7.02	9.3e-9
Bfat	9	192	5.84	3.4e-7
LBM	9	192	5.39	1.4e-6
Ht	9	192	1.96	0.0455
Wt	9	192	2.30	0.0176

Figure 7: Levene's Test for homogeneity of variance assumption by sport factor

The null hypothesis that there is no difference in variances across the various sports is rejected in favour of the alternative hypothesis at the 5% significance level. We conclude that significant differences in the variance of the response variables is present for all but one response variable (Hc). For Hc, there is insufficient evidence to reject the null hypothesis.

Now that a key assumption has been violated, we must consider an alternative method for performing the one-way ANOVA. Welch's one-way ANOVA provides the perfect choice as there is no assumption of equal variance.

3.2. QQ PLOT FOR TEST OF NORMALITY

The following QQ Plots show excellent normality for proceeding with testing of the data against Welch's ANOVA. The iron ferritin (Ferr) and Sum of Skin Fold (SSF) response variables were square rooted to correct for wayward upper tails. The square roots shall be carried forward into the ANOVA analysis.

The QQ Plot was preferred to the more formal Shapiro-Wilk test due to the large sample size (n > 50) where the Shapiro test is known to be highly sensitive to slight deviations from the norm.

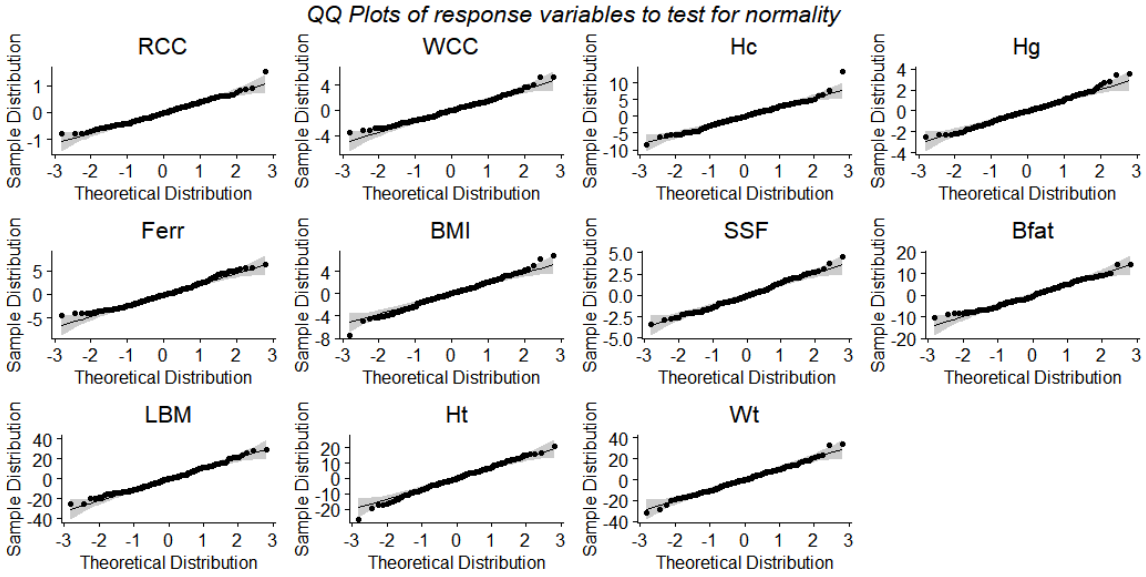


Figure 8: QQ Plot of response variables to test assumption of normality for ANOVA

The QQ Plot shows that the response variables are now satisfying the assumption of normality for the one-way ANOVA.

3.3. WELCH'S ONE-WAY ANOVA

Welch's one-way ANOVA was carried out against the full dataset.

Biomarker	n	Statistic	DFn	DFd	p-value
RCC	202	17.5	9	47	3.0e-12
WCC	202	3.64	9	45.3	0.002
Hc	202	21.9	9	45.5	9.5e-14
Hg	202	25.5	9	45.7	5.8e-15
Sqrt(Ferr)	202	4.3	9	49.1	3.7e-4
BMI	202	13.6	9	46.0	2.7e-10
Sqrt(SSF)	202	21.4	9	47.6	6.9e-14
Bfat	202	16.6	9	48.1	5.36e-12
LBM	202	37.8	9	48.1	7.8e-19
Ht	202	19.2	9	46.2	7.8e13
Wt	202	32.1	9	48.4	1.9e-17

Figure 9: Welch's one-way ANOVA against the sport factor

The null hypothesis of equal means is rejected for all response variables in favour of the alternative hypothesis at the 5% significance level, meaning that there is at least one difference between within the sports factor. A pairwise multiple comparison was carried using Bonferroni's method at the 5% significance level to identify where the difference are located.

The results are also significant down to the 1% level.

A summary of significant differences for the Haemoglobin (Hg) response variable is given, but the full results are very lengthy and can be found by running the code in the accompanying cookbook. **Results are significant to the 5% level**:

Biomarker	Df	Statistic	Sport 1	Sport 2	Lower CI	Upper CI	P-adjusted
Hg	42.0	-3.97	BBall	Field	-2.13	-0.69	0.0130
Hg	32.8	4.33	BBall	Netball	0.68	1.88	0.0060
Hg	38.2	-4.42	BBall	WPolo	-2.07	-0.77	0.0040
Hg	27.0	10.3	Field	Netball	2.16	3.23	3.26e-9
Hg	57.8	-8.80	Netball	Row	-2.17	-1.36	1.35e-10
Hg	30.3	-6.95	Netball	Swim	-2.45	-1.34	4.3e-6
Hg	42.3	-7.71	Netball	T400m	-2.41	-1.41	6.08e-8
Hg	16.4	-6.66	Netball	TSprnt	-3.74	-1.93	2.2e-4
Hg	29.6	-12.8	Netball	WPolo	-3.13	-2.27	5.76e-12
Hg	41.8	-3.94	Row	WPolo	-1.42	-0.46	0.0140

Figure 10: Summary of differences Haemoglobin (Hg) under Bonferroni's method

Other response variables contain up to 20 differences. For the sake of brevity, only Hg is reported here.

4. TWO WAY ANOVA TEST ON FACTORS: SEX & SPORT

4.1. TEST FOR NORMALITY

The data was checked for normality using the Shapiro-Wilk test and QQ Plot. As the data is now stratified by sex and sport the sample size within each pot is smaller ($n < 30$) for all response variables, thus making Shapiro-Wilk relevant. The data was tested with and without outliers and neither yielded perfect results:

Biomarker	Full Data Statistic	Full Data p-value		No Outliers Statistic	No Outliers p-value
RCC	0.9700	0.0003		**0.9894**	**0.2179**
WCC	0.9818	0.0103		**0.9863**	**0.0854**
Hc	0.9613	2.5e-5		**0.9956**	**0.8866**
Hg	0.9826	0.0132		**0.9897**	**0.2372**
Ferr	0.9495	1.5e-6		0.9759	0.0039
BMI	0.9852	0.0323		**0.9921**	**0.4506**
SSF	0.94603	7.1e-7		0.9776	0.0063
Bodyfat	0.9708	0.0003		0.9791	0.0098
LBM	0.9634	4.3e-5		0.9556	2.5e-5
Ht	0.9794	0.0045		0.9707	0.0009
Wt	0.9748	0.0011		0.9671	0.0004

Figure 11: Shapiro-Wilk test for normality for two-way ANOVA

The data was transformed to try and find a better fit for ANOVA. The data without outliers was the only one that responded significantly better improving from 5 to 8 response variables satisfying normality with the 3 remaining finding sufficient evidence to reject the null hypothesis of normality at the 5% level ($p < 0.05$).

Biomarker	Transform		No Outliers Statistic	No Outliers p-value
RCC	-		**0.9894**	**0.2179**
WCC	-		**0.9863**	**0.0854**
Hc	-		**0.9956**	**0.8866**
Hg	-		**0.9897**	**0.2372**
Ferr	Sqrt		**0.99541**	**0.8708**
BMI			**0.9921**	**0.4506**
SSF	Sqrt		**0.99075**	**0.3189**
Bodyfat	Sqrt		**0.9874**	**0.1193**
LBM	None found		0.9556	2.5e-5
Ht	None found		0.9707	0.0009
Wt	None found		0.9671	0.0004

Figure 12: Shapiro-Wilk test for normality for two-way ANOVA on transformed data

4.2. QQ PLOTS TO TEST NORMALITY

The QQ Plots show that most of the data is normally distributed with some curving at the lower tail within the BMI, Lean Body Mass, Heigh and Weight response variables.

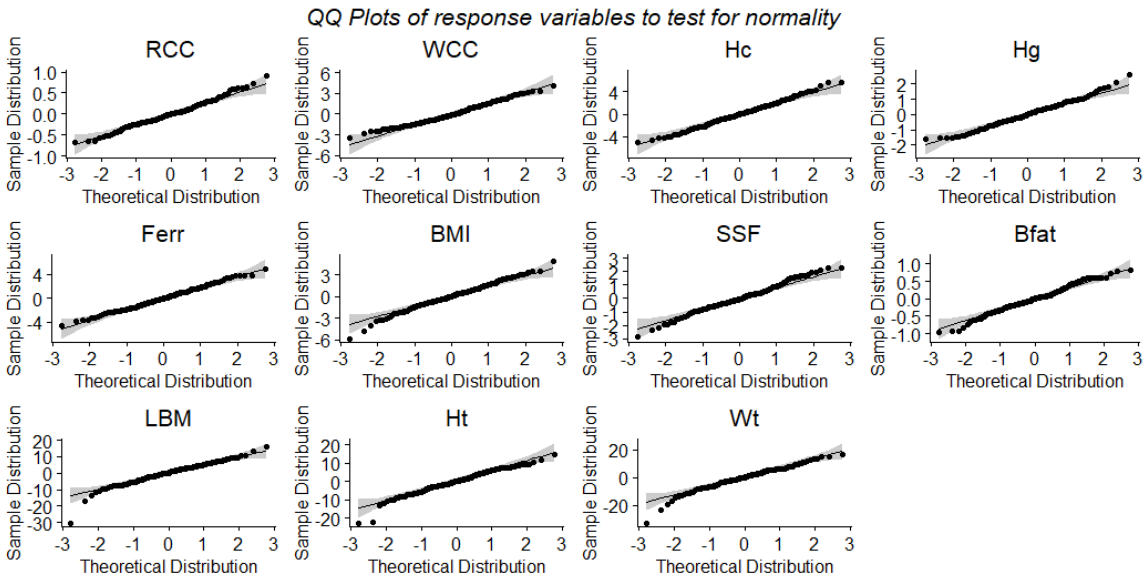


Figure 13: QQ Plot to show normality of data within the sex and sport factor analysis

Overall, the data looks good to proceed albeit with some caution shown when interpreting the results.

4.3. TEST FOR HOMOGENEITY

Using the data with outlying values removed, and the transformation listed in the previous section, the data was tested for homogeneity of variance.

Biomarker	Transform		Df1	Df2	Statistic	p-value
RCC	-		16	158	0.659	0.830
WCC	-		16	158	1.46	0.120
Hc	-		16	158	1.31	0.196
Hg	-		16	158	0.924	0.544
Ferr	Sqrt		16	158	0.993	0.467
BMI			16	158	1.13	0.329
SSF	Sqrt		16	158	2.22	0.006
Bodyfat	Sqrt		16	158	3.05	0.0002
LBM	None found		16	158	0.754	0.735
Ht	None found		16	158	0.728	0.763
Wt	None found		16	158	1.07	0.392

Figure 14: Levene's Test for equal variance of response variables

Levene's test for equality of variance reveals that 9 of the 11 response variables possess a p-value greater than 0.05 indicating that those variables share equal variance across the factors. Two factors however, Sum of Skin Folds (SSF) and Bodyfat (Bfat) are significant at the 5% level and therefore the null hypothesis of equal variance across factors is rejected in favour of the alternative hypothesis at the 5% significance level ($P < 0.05$).

4.4. TWO-WAY ANOVA

There is some concern in the data where the Shapiro statistic does not agree with the QQ Plot and two response variables that are significant in Levene's Test however the two-way ANOVA is considered robust and the data comes from human athletes, which is generally considered to be normally distributed. Care must be taken when interpreting the results of the two-way ANOVA test when deciding whether to accept or reject the null hypothesis.

The two-way ANOVA will look at the Sex and Sport factors as well as any potential interaction effect between them. The results are summarized at the 5% level using the dataset with outlying values removed:

Response Variable: RCC

	Df	SS	Mean SS	F value	Pr(>F)	Sig. Level
Sex	1	14.659	14.659	171.383	2e-16	**0.001**
Sport	9	3.761	0.418	7.885	8.87e-6	**0.001**
Sex:Sport	6	0.748	0.125	1.458	0.196	-
Residuals	158	13.514	0.086			

Figure 15.1: Two-way ANOVA table for response variable RCC

Response Variable: WCC

	Df	SS	Mean SS	F value	Pr(>F)	Sig. Level
Sex	1	0.0	0.042	0.018	0.8931	-
Sport	9	53.2	53.2	2.558	0.009	**0.01**
Sex:Sport	6	13.8	13.8	0.994	0.4315	-
Residuals	158	365.3	365.3			

Figure 15.2: Two-way ANOVA table for response variable WCC

Response Variable: Hc

	Df	SS	Mean SS	F value	Pr(>F)	Sig. Level
Sex	1	1005.1	1005.1	213.457	2e-16	**0.001**
Sport	9	217.3	24.1	5.128	4.3e-6	**0.001**
Sex:Sport	6	47.4	7.9	1.677	0.13	-
Residuals	158	743.9	4.7			

Figure 15.3: Two-way ANOVA table for response variable Hc

Response Variable: Hg

	Df	SS	Mean SS	F value	Pr(>F)	Sig. Level
Sex	1	151.80	151.80	234.468	2e-16	**0.001**
Sport	9	30.19	3.35	5.182	3.66e-6	**0.001**
Sex:Sport	6	4.16	0.69	1.070	0.383	-
Residuals	158	102.29	0.65			

Figure 15.4: Two-way ANOVA table for response variable Hg

Response Variable: sqrt(Ferr)

	Df	SS	Mean SS	F value	Pr(>F)	Sig. Level
Sex	1	94.4	94.41	27.094	5.95e-7	**0.001**
Sport	9	93.8	10.42	2.991	0.0026	**0.01**
Sex:Sport	6	53.4	8.90	2.555	0.0217	**0.05**
Residuals	158	550.5	3.48			

Figure 15.5: Two-way ANOVA table for response variable Ferr

Response Variable: BMI

	Df	SS	Mean SS	F value	Pr(>F)	Sig. Level
Sex	1	98.9	98.85	32.46	5.79e-8	**0.001**
Sport	9	293.9	32.65	10.72	6.6e-13	**0.001**
Sex:Sport	6	45.5	7.58	2.49	0.025	**0.05**
Residuals	158	481.1	3.04			

Figure 15.6: Two-way ANOVA table for response variable BMI

Response Variable: sqrt(SSF)

	Df	SS	Mean SS	F value	Pr(>F)	Sig. Level
Sex	1	203.73	203.73	214.696	2e-16	**0.001**
Sport	9	156.10	17.34	18.278	2e-16	**0.001**
Sex:Sport	6	20.36	3.39	3.576	0.0024	**0.01**
Residuals	158	149.93	0.95			

Figure 15.7: Two-way ANOVA table for response variable SSF

Response Variable: sqrt(Bfat)

	Df	SS	Mean SS	F value	Pr(>F)	Sig. Level
Sex	1	64.94	64.94	480.044	2e-16	**0.001**
Sport	9	22.93	2.55	18.836	2e-16	**0.001**
Sex:Sport	6	3.32	0.55	4.095	0.0008	**0.001**
Residuals	158	21.37	0.14			

Figure 15.8: Two-way ANOVA table for response variable Bfat

Response Variable: LBM (care when interpreting)

	Df	SS	Mean SS	F value	Pr(>F)	Sig. Level
Sex	1	15295	15295	420.079	2e-16	**0.001**
Sport	9	3537	393	10.793	5.51e-13	**0.001**
Sex:Sport	6	437	73	1.999	0.0689	-
Residuals	158	5753	36			

Figure 15.9: Two-way ANOVA table for response variable LBM

Response Variable: Ht (care when interpreting)

	Df	SS	Mean SS	F value	Pr(>F)	Sig. Level
Sex	1	5160	5160	141.862	2e-16	**0.001**
Sport	9	3919	435	11.972	2.73e-14	**0.001**
Sex:Sport	6	180	30	0.827	0.551	-
Residuals	158	5746	36			

Figure 15.10: Two-way ANOVA table for response variable Ht

Response Variable: Wt (care when interpreting)

	Df	SS	Mean SS	F value	Pr(>F)	Sig. Level
Sex	1	8360	8360	146.761	2e-16	**0.001**
Sport	9	8016	891	15.635	2e-16	**0.001**
Sex:Sport	6	442	74	1.293	0.263	-
Residuals	158	9000	57			

Figure 15.11: Two-way ANOVA table for response variable Wt

The results of the two-way ANOVA test are summarized below:

Biomarker	Sex Signif. Lvl	Sport Signif. Lvl	Sex:Sport Signif. Lvl
RCC	0.001	0.001	-
WCC	-	0.01	-
Hc	0.001	0.001	-
Hg	0.001	0.001	-
Ferr	0.001	0.01	0.05
BMI	0.001	0.001	0.05
SSF	0.001	0.001	0.01
Bfat	0.001	0.001	0.001
LBM	0.001	0.001	-
Ht	0.001	0.001	-
Wt	0.001	0.001	-

Figure 16: Summary of two-way ANOVA for sex and sport factors

Except for the WCC variable we can categorically reject the null hypothesis in favour of the alternative hypothesis at the 5% significance level. There is overwhelming evidence to suggest at all levels (excluding WCC by Sex) that the biomarkers are affected by athlete sex and their chosen sport in at least one category of the factor.

Four variables produced a significant result for the interaction of sport and sex at the 5% significance level therefore we find sufficient evidence to also reject the null hypothesis in at least one category of the factor for the interaction of a given sport on the sex of the athlete.

Multiple comparison testing will be performed on the four variables where an interaction effect was noticed. To perform it also on the sex and sport factors you can refer to the cookbook and amend the settings as required.

The one-way ANOVA produced an overwhelming amount of data for just one factor.

4.5. MULTIPLE COMPARISONS TEST FOR FACTOR INTERACTION

Post-hoc test to see where the main effect can be found.

Sex	Effect	DFn	F	p-value
female	Sport	8	158	**8e-5**
male	Sport	7	158	**0.018**

Figure 17: Discovering the main effect on the sex factor

As expected, both sexes are affected by sport. The effect of sport on females is much greater than the effect of sport on males for differentiating the athletes.

The Bonferroni method reveals where the differences are between the sex and sport factors. The differences below are shown for the **FERR and BMI variables only**. To view the differences for SSF and BFAT please refer to the cookbook. These are all significant at the 5% level.

Biomarker	Sex	Sport1	Sport2	Df	Statistic	p-adj.
Ferr	female	BBall	Tennis	158	7.58e-4	0.0341
Ferr	female	Row	Tennis	158	8.27e-4	0.0372
Ferr	female	400m	Tennis	158	5.95e-5	0.0027
BMI	female	BBall	Field	158	-5.65	3.23e-6
BMI	female	Field	Gym	158	5.35	1.37e-5
BMI	female	Field	Netball	158	4.59	4.05e-4
BMI	female	Field	Row	158	4.02	0.0041
BMI	female	Field	Swim	158	4.44	7.44e-4
BMI	female	Field	T400m	158	6.79	9.65e-9
BMI	female	Field	Tennis	158	5.72	2.3e-6
BMI	female	Field	TSprnt	158	4.79	1.74e-4
BMI	female	Gym	Row	158	-3.41	0.0374
BMI	female	Netball	T400m	158	3.49	0.0283
BMI	female	Row	T400m	158	4.32	0.0012
BMI	male	Row	T400m	158	5.07	5.03e-5
BMI	male	Swim	T400m	158	3.68	0.0145
BMI	male	T400m	TSprnt	158	-3.41	0.0377
BMI	male	T400m	WPolo	158	-5.25	2.22e-5

Figure 18: Multiple comparison for two of the significant interaction variables

Figure 18 is pre-dominantly female for the two chosen response variables. This appear to tally with figure 17 in suggesting that the sex:sport interaction is more significant in female athletes.

5. MULTIPLE LINEAR REGRESSION MODEL

The data was partitioned into a training and test set in preparation for the model prediction. The training set consists of 80% of the dataset, with the remaining 20% assigned to the test set. The data was chosen at random with the sample() function.

5.1. CORRELATION AND COLINEARITY

The data was check for correlation and collinearity. Some of the response variables exhibited high degrees of correlation however the iron ferritin appears to be uncorrelated with the other response variables with maximum correlation with BMI (r = 0.35) and a minimum correlation with SSF (r = -0.06).

Plot of pairs:

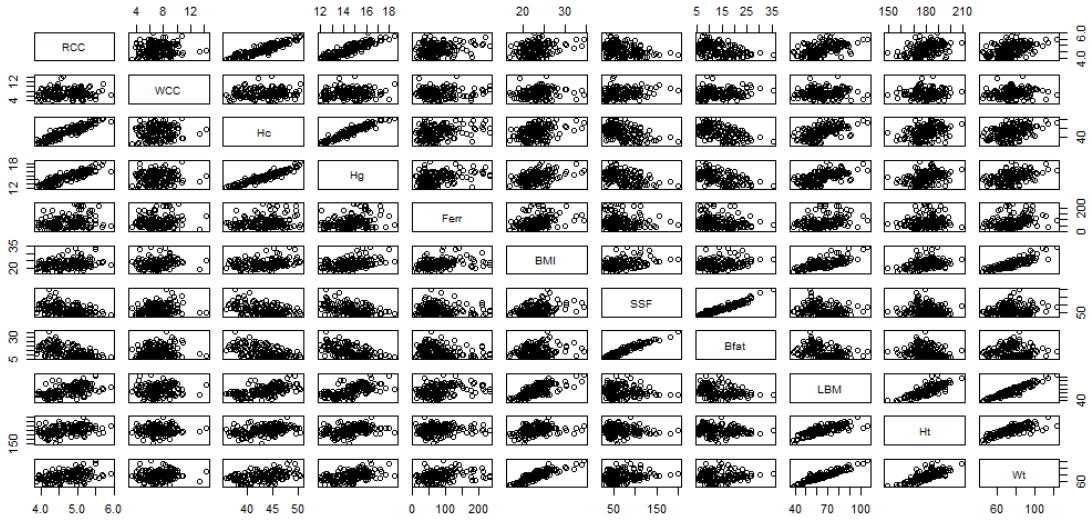


Figure 19: Plot of pairs showing no relationships between FERR and the other variables

Correlation of iron ferritin Variable to other response variables:

RCC	WCC	Hc	Hg	BMI
0.2617	0.1436	0.2756	0.3400	0.3247
SSF	**Bfat**	**LBM**	**Ht**	**Wt**
-0.1257	-0.1958	0.3532	0.1345	0.3049

Figure 20: Correlation values of the iron ferritin response variable

5.2. BASELINE MODEL

Summary of the baseline model as calculated in R:

Residuals:

Min	LQ	Median	UQ	Max
-82.55	-27.98	-10.70	19.03	143.41

Figure 21.1: Summary of baseline model residuals

Coefficients:

	Estimate	Std. Error	t-value	Pr(>|t|)
{Intercept)	432.8984	467.6694	0.926	0.356
RCC	2.0498	21.2373	0.097	0.923
WCC	2.6518	1.9991	1.326	0.187
Hc	-5.4506	4.0140	-1.358	0.177
Hg	16.9019	8.9880	1.880	0.062
BMI	-5.1416	10.9734	-0.469	0.640
SSF	0.4743	0.5204	0.911	0.363
Bfat	-3.7278	4.4654	-0.835	0.405
LBM	0.5947	54934	0.108	0.914
Ht	-2.7296	2.6560	-1.028	0.306
Wt	2.5919	5.8507	0.443	0.658

Figure 21.2: Summary of baseline model coefficients

Summary:

Summary
Residual Standard error: 44.53 on 151 degrees of freedom
Multiple R Squared: 0.2233
Adjust R Squared: 0.1718
F-Statistic = 4.34 on 10 and 151 degrees of freedom, p-value = 2.406-05

Figure 21.3: Summary of baseline model statistics

5.3. PLOTS OF BASELINE MODEL

The QQ Plot shows a deviation from the normal distribution at the upper tail and the plot of ordered observed residuals shows increasing deviation towards end of the dataset.

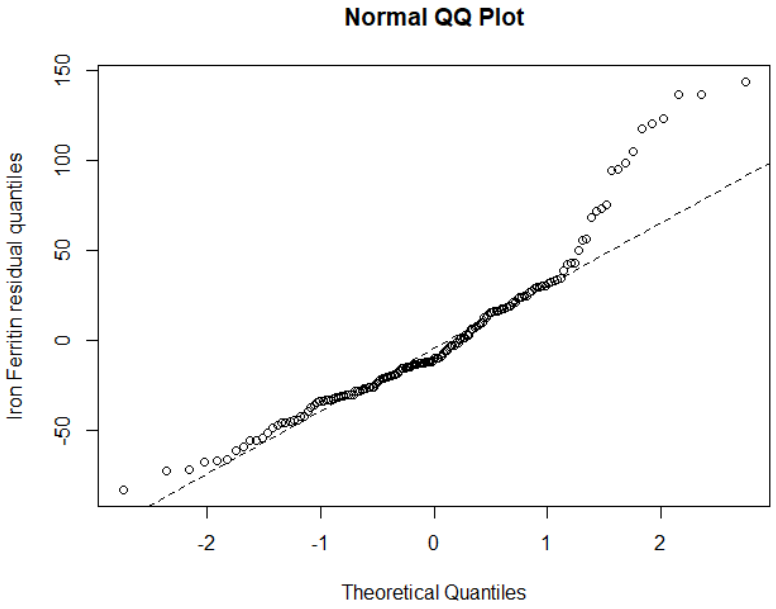


Figure 22: QQ Plot of baseline model for normality

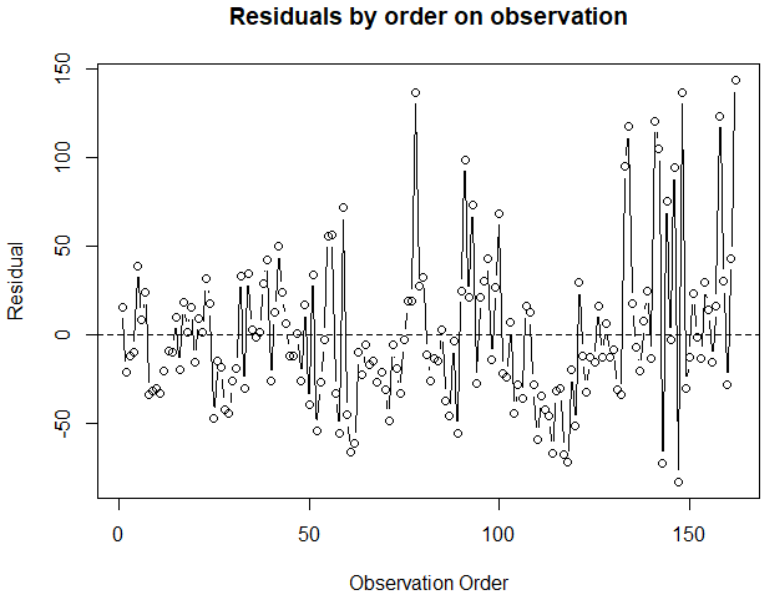


Figure 23: Residual plot of observation order

The plot of the baseline model seems to suggest that the upper tail values will be problematic as they deviate substantially from the main cluster along the XY line.

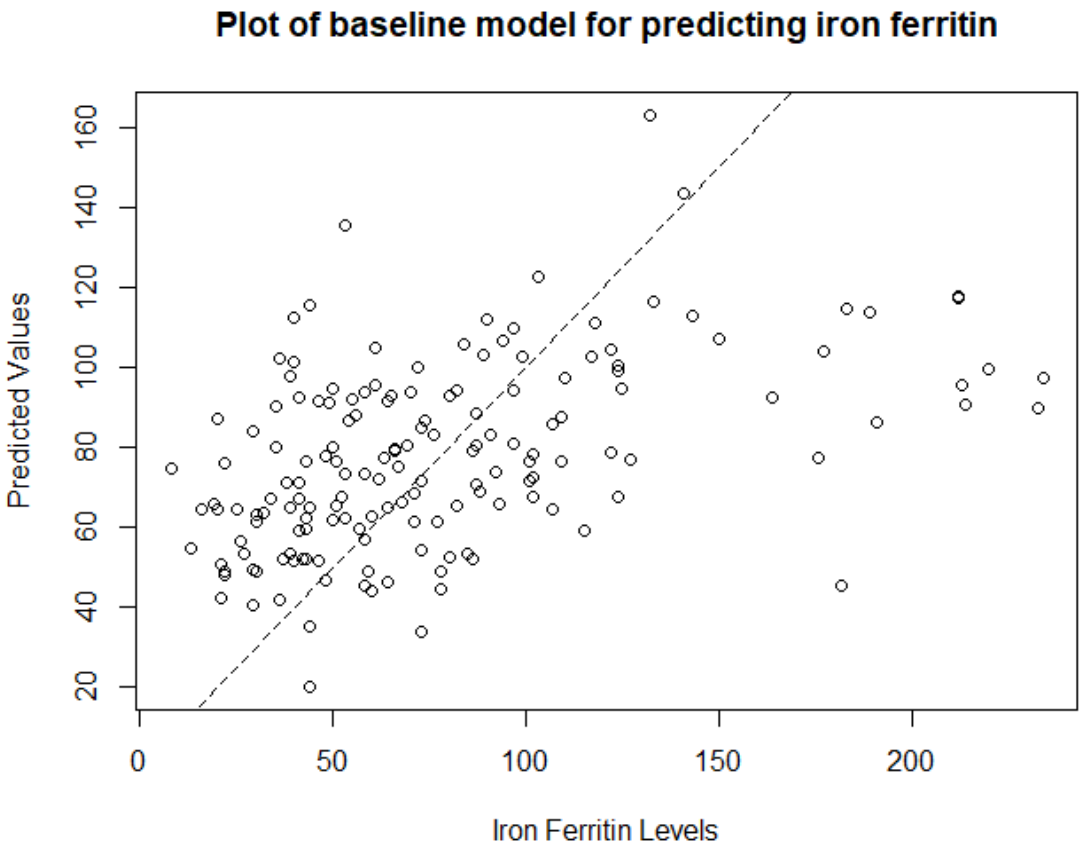


Figure 24: Plot of baseline model for predicting iron ferritin levels

5.4. VARIANCE INFLATED FACTOR (VIF)

The VIF for this model was calculated as:

RCC	WCC	Hc	Hg	BMI
7.7764	1.0926	18.1911	12.4108	81.5559
SSF	**Bfat**	**LBM**	**Ht**	**Wt**
29.5910	65.6905	397.8385	51.0856	506.8559

Figure 25: VIF data for baseline model indicating high levels of multicollinearity

Any value over 10 is undesirable and shows a high degree of multicollinearity in the baseline model. The drop1 method will help refine and eliminate this issue.

5.5. ITERATE THROUGH DROP1 METHOD FOR REFINEMENT

The model was refined through the following code:

```
basedrop <- lm(data=train, Ferr ~ RCC+WCC+Hc+Hg+BMI+SSF+Bfat+LBM+Ht+Wt)
        with (train, { drop1(basedrop, test="F") })  # Remove RCC
basedrop <- lm(data=train, Ferr ~ WCC+Hc+Hg+BMI+SSF+Bfat+LBM+Ht+Wt)
        with (train, { drop1(basedrop, test="F") })  # Remove LBM
basedrop <- lm(data=train, Ferr ~ WCC+Hc+Hg+BMI+SSF+Bfat+Ht+Wt)
        with (train, { drop1(basedrop, test="F") })  # Remove BMI
basedrop <- lm(data=train, Ferr ~ WCC+Hc+Hg+SSF+Bfat+Ht+Wt)
        with (train, { drop1(basedrop, test="F") })  # Remove SSF
basedrop <- lm(data=train, Ferr ~ WCC+Hc+Hg+Bfat+Ht+Wt)
        with (train, { drop1(basedrop, test="F") })  # Remove WCC
basedrop <- lm(data=train, Ferr ~ Hc+Hg+Bfat+Ht+Wt)
        with (train, { drop1(basedrop, test="F") })  # Remove Hc
basedrop <- lm(data=train, Ferr ~ Hg+Bfat+Ht+Wt)
        with (train, { drop1(basedrop, test="F") })  # Remove WCC
basedrop <- lm(data=train, Ferr ~ Hg+Bfat+Ht+Wt)
        with (train, { drop1(basedrop, test="F") })  # Remove Hg
basedrop <- lm(data=train, Ferr ~ Bfat+Ht+Wt)
      with (train, { drop1(basedrop, test="F") })


summary(basedrop)
predmodel <- lm(data=train, Ferr ~ Bfat + Ht + Wt
```

This section of code can be run from the cookbook. Results may differ due to random sampling of the training dataset (another indicator of data not providing a great model).

5.6. SUMMARY OF PREDICTIVE MODEL

Summary of the predictive model as calculated in R:

Residuals:

Min	LQ	Median	UQ	Max
-83.893	-30.223	-9.729	21.856	155.499

Figure 26.1: Summary of baseline model residuals

Coefficients:

	Estimate	Std. Error	t-value	Pr(>|t|)
{Intercept)	307.5102	91.0004	3.379	0.000916
Bfat	-2.1874	0.5868	-3.728	0.000269
Ht	-2.0229	0.6144	-3.293	0.001225
Wt	2.1946	0.4195	5.232	5.26e-7

Figure 26.2: Summary of baseline model coefficients

Summary:

Summary
Residual Standard error: 44.51 on 158 degrees of freedom
Multiple R Squared: 0.1879
Adjust R Squared: 0.1725
F-Statistic = 12.19 on 3 and 158 degrees of freedom, p-value = 3.238e-7

Figure 26.3: Summary of baseline model statistics

5.7. OBSERVED VERSUS PREDICTED PLOT

The following plot shows the observed values from the training dataset versus the predicted values from the linear model. Accepting that the model might not be the best for the data provided, it does seem to scatter well when laid over the actual observed data.

The model appears to over-predict at 50/60ng/ml or less and under-predict for levels > 100ng/ml. The predictions follow the fitted values from the observed training model.

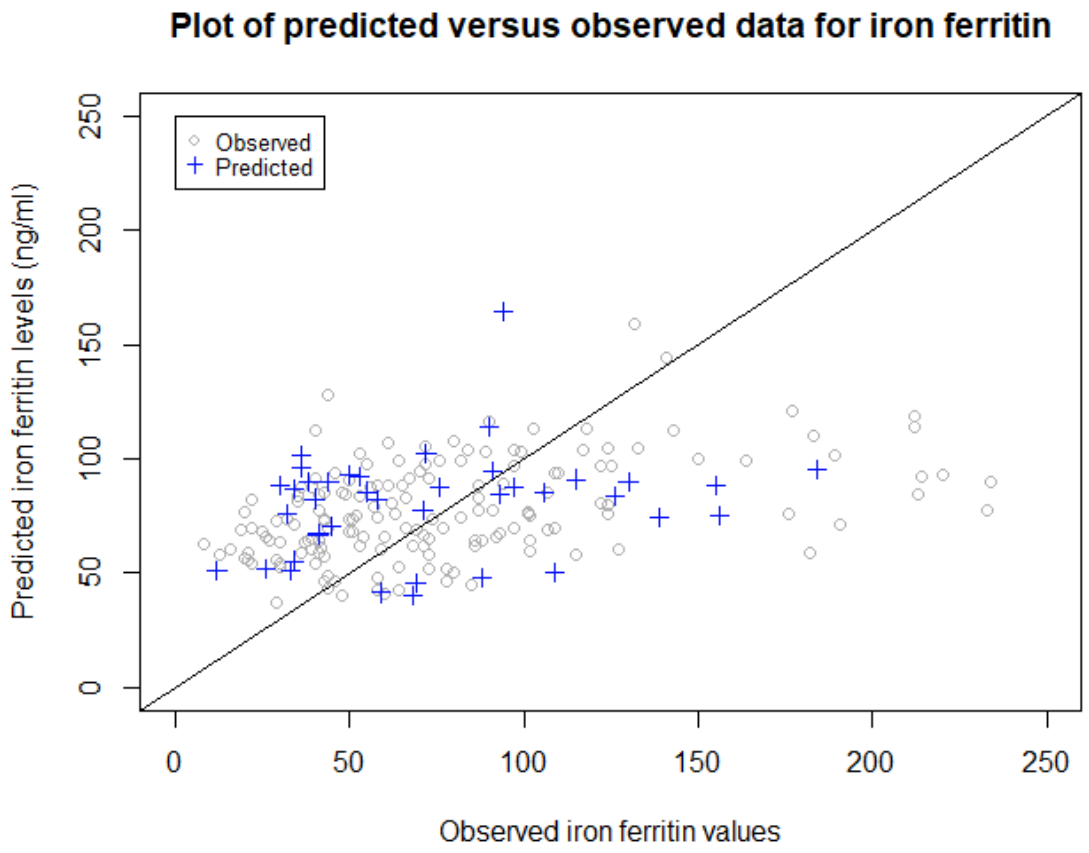


Figure 27: Observed iron ferritin levels overlaid with the predictive model results

6. ADDITIONAL INSIGHTS

6.1. AVERAGE WHITE CELL COUNTS FOR ATHLETES

Does the data support Horn et al in proposing that WCC might be lower in aerobic endurance athletes? The means have been calculated by sex as per Horn's study.

WCC female mean: 6.99 **WCC female SD: 1.70**
WCC male mean: 7.22 **WCC male SD: 1.90**

Sport	Female mean	Female sd		Male mean	Male sd
BBall	6.31	1.53		7.33	1.11
Field	7.86	1.42		7.86	1.58
Gym	6.32	1.91		-	-
Netball	7.93	2.20		-	-
Row	6.76	1.44		7.09	1.50
Swim	6.21	1.28		6.08	1.38
T400m	6.93	0.947		6.19	1.36
Tennis	6.1	1.47		7.45	1.29
TSprnt	7.8	0.283		6.92	2.06
WPolo	-	-		8.92	2.48

Figure 28: Mean white cell count levels of athletes by sport factor

The cells highlighted in green are of interest to this comparison. This seems to hold true for female athletes and male swimmers. The male 400m athletes are a surprise inclusion however they may well carry out sprint endurance training to sustain top speeds over 400m.

6.2. ATHLETES WITH UNUSUAL IRON FERRITIN LEVELS

Extreme iron ferritin levels can indicate issues at both ends of the distribution. For excessively high levels it can be a sign of an underlying cancer or more commonly haemochromatosis, and for excessively low levels, very much prevalent amongst female athletes, it can indicate sports anaemia. The chart below gives some idea of how many athletes in the dataset are at risk of anaemia and of those who are approaching the recommended upper limits for normal people.

(N/A indicates range does not apply to this Sex).

Iron Ferritin Levels	Female athletes	Percentage	[redacted]	Male athletes	Percentage
< 20 ng/ml	4	4%	[redacted]	1	1%
< 40 ng/ml	30	30%	[redacted]	11	11%
> 150 ng/ml	2	2%	[redacted]	N/A	N/A
> 200ng/ml	N/A	N/A	[redacted]	7	7%

Figure 28: Mean white cell count levels of athletes by sport factor

We see a small number of athletes who should seek help for sports anaemia (< 20 ng/ml) and a large number who fall below Team USA's definition cut-off for early detection of sports anemia (< 40 ng/ml). For athletes, it is recommended to stay above 50 ng/ml to maintain peak performance levels.

**** END OF DOCUMENT ****