Plasma Ferritin Concentration Study

Which factors affect plasma ferritin concentration (Ferr) among Australian athletes?

John Karuitha; diakingathia2005@gmail.com

Sunday October 23, 2022

Contents

1 Background **2**

2 Objective **2**

3 Summary of Results **2**

4 Is there a Significant Difference in Average Ferritin Concentration Between Male and Female Athletes in Australia? **2**

4.1 Checking Assumptions of the Independent Sample T-test . . . 4

4.1.1 Continous Dependent Variable . . . 4

4.1.2 Independence of observations . . . 4

4.1.3 Normality of Independent variable . . . 4

4.1.4 Outliers . . . 6

5 Regression Analysis: Drivers of Ferritin Concentration Among Athletes in Australia **6**

5.1 Simple Linear Regression . . . 6

5.2 Choosing a Model Using Backward Stepwise Regression . . . 7

5.3 Regression Diagnostics . . . 10

5.3.1 Linear Relationship . . . 10

5.3.2 Multivariate normality. . . . 12

5.3.3 No or little multicollinearity. . . . 12

5.3.4 No auto-correlation . . . 12

5.3.5 Homoscedasticity . . . 12

6 Conclusion **12**

References **12**

1 Background

In this article, we assess the effect of a collection of explanatory variables on the plasma ferritin concentration (Ferr) in 202 Australian athletes. The file `Sports Data CW 2021.csv` contains the data on the plasma ferritin concentration as well as a selection of demographic variables of 202 male and female athletes. In particular, the data set comprises observations on the following 11 variables: Sex, Sport, LBM, RCC, WCC, Hc, Hg, BMI, SSF, X.Bfat, Ferr. Table 1 shows the desciption of the variables.

Table 1: Description of Variables

Variable	Description
Sport	Type of Sport.
Sex	Sex of athlete; male or female.
LBM	Lean body mass of athlete.
RCC	Red blood cells count.
WCC	White blood cells count.
Hc (%)	Hematocrit (Hc) is the volume percentage (vol%) of red blood cells in blood. It is normally 47% ± 5% for men and 42% ± 5% for women.
Hg (g/dl)	Hemoglobin (Hg) is the protein contained in red blood cells that is responsible for delivery of oxygen to the tissues. The normal Hg level for ***males is 14 to 18 g/dl; that for females is 12 to 16 g/dl.***
BMI	Body mass index = weight/height^2.
SSF (mm)	Sum of skin folds.
% Bfat	% Body fat.
Ferr (mol/L)	Plasma ferritin concentration.

2 Objective

The broad objective of the study is to uncover factors that are related to the level of ferritin concentration among Australian athletes.

3 Summary of Results

1. There is a significant difference in mean ferritin concentration between male and female athletes in Australia.
2. Sex, lean body mass (LBM) and body mass index (BMI) are the significant drivers of ferritin concentration among athletes in Australia.
3. The regression model using the raw variables has outliers, is not homoscedastic, and fails the multivariate normality test. A transformation of the variables could make the model a better fit.

4 Is there a Significant Difference in Average Ferritin Concentration Between Male and Female Athletes in Australia?

In this section, we test if plasma ferritin concentration differs between male and female athletes while making sure that the assumptions of the test are satisfied. I start by visualizing the ferritin concentration levels between male and female athletes. Indeed, Figure 1 (below) shows a substantial gap in the median ferritin concentration between the genders with males on the higher side. But is the observed difference significant? To answer this question we have to conduct a hypothesis test.

```
boxplot(Ferr ~ Sex, data = ferritin_data,
        main = "Ferritin Concentration for Male and Female Athletes",
        col = c('#0000FF', '#CCCCFF'))
```

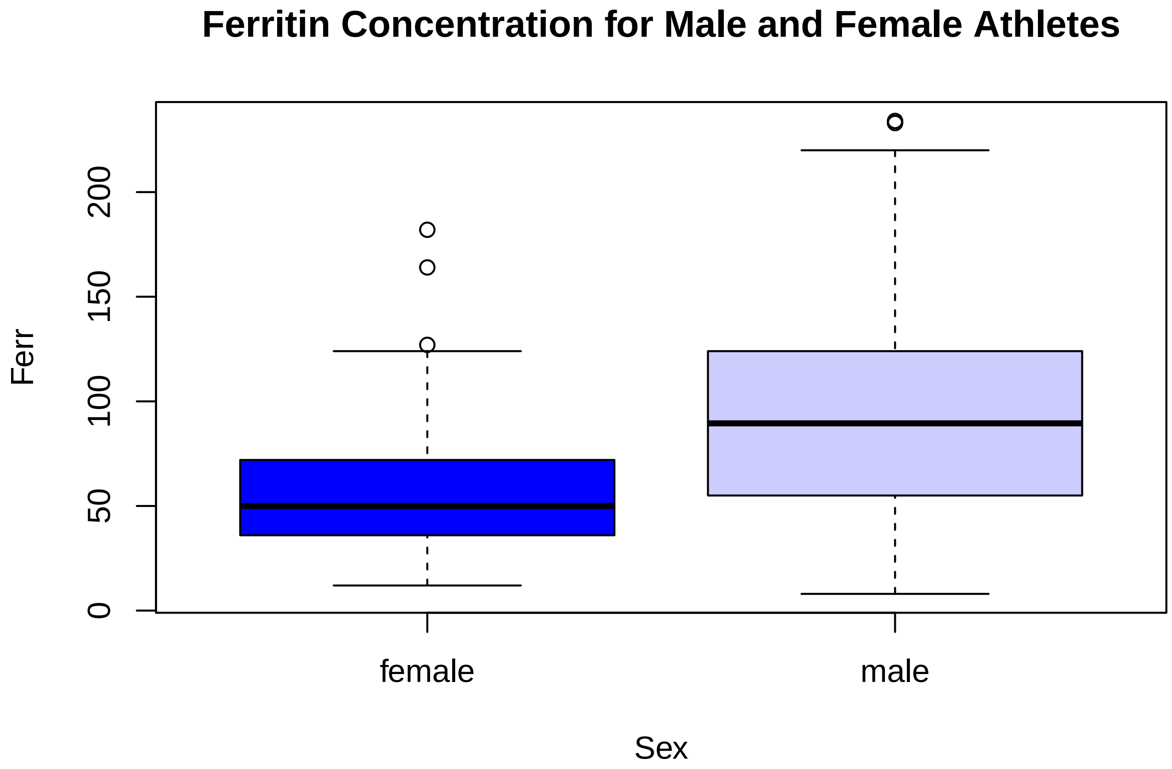


Figure 1: Ferritin Concentration for Male and Female Athletes

We conduct a independent sample t-test for the mean difference in ferritin concentration between male and female athletes (Xu et al. 2017). Note that this is a two-tailed test.

The hypotheses for the t-test are as follows;

H0: The true difference in mean ferritin concentration between group female and group male is equal to 0.
HA: The true difference in mean ferritin concentration between group female and group male is **NOT** equal to 0.

We then conduct the t-test.

```
t.test(Ferr ~ Sex, data = ferritin_data)
```

```
##
##  Welch Two Sample t-test
##
## data:  Ferr by Sex
## t = -6.5043, df = 163.96, p-value = 9.033e-10
## alternative hypothesis: true difference in means between group female and group male is not equal to
## 95 percent confidence interval:
##  -51.41560 -27.46832
## sample estimates:
## mean in group female   mean in group male
##             56.96000             96.40196
```

The results of the t-test shows a significant difference in mean ferritin levels between male and female athletes (*t = -6.5043, df = 163.96, p-value = 9.033e-10*). Hence we reject the NULL hypothesis and conclude that the difference in mean ferritin concentration between male and female athletes is NOT equal to zero.

4.1 Checking Assumptions of the Independent Sample T-test

The independent samples t-test has several assumptions:

- The dependent variable must be continuous (interval/ratio).
- The observations are independent of one another.
- The dependent variable should be approximately normally distributed.
- The dependent variable should not contain any outliers.

4.1.1 Continous Dependent Variable

The dependent variable meets this assumption as it is numeric and continous.

4.1.2 Independence of observations

The independence of observations will depend upon the data collection process. If the selection of the subjects was random without replacement, then we can safely assume independence between observations.

4.1.3 Normality of Independent variable

To check for the normality of residuals, we plot the dependent variable. The plot is in figure 2.

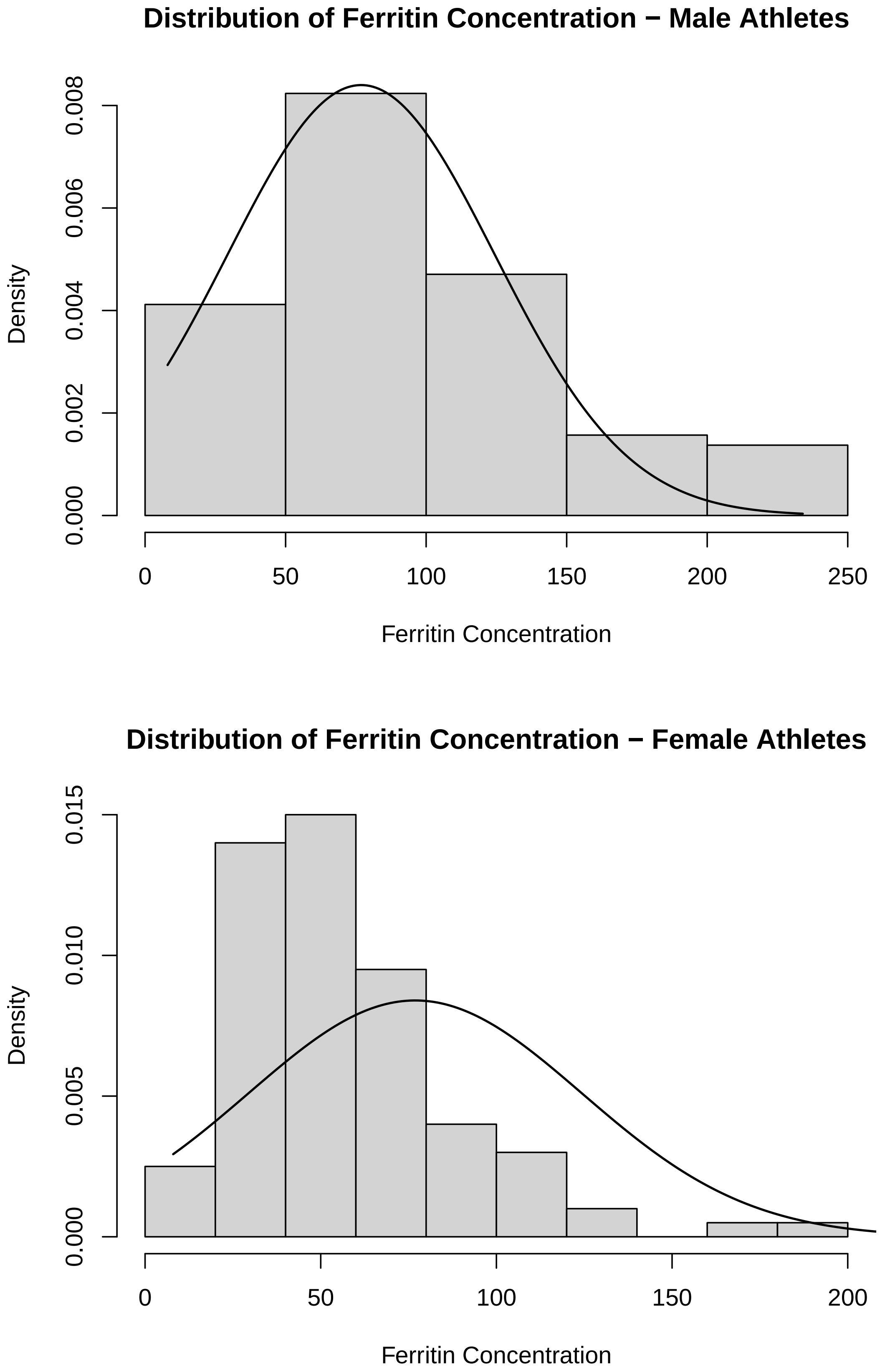
```r
par(mfrow = c(2, 1))
x_values <- seq(min(ferritin_data$Ferr), max(ferritin_data$Ferr),
                length = nrow(ferritin_data))

fun <- dnorm(x_values, mean = mean(ferritin_data$Ferr), sd = sd(ferritin_data$Ferr))
####################################
hist(ferritin_data %>% filter(Sex == "male") %>% pull(Ferr),
     main = "Distribution of Ferritin Concentration - Male Athletes",
     xlab = "Ferritin Concentration", prob = TRUE)

lines(x_values, fun, lwd = 1.5)

#############################################
hist(ferritin_data %>% filter(Sex == "female") %>% pull(Ferr),
     main = "Distribution of Ferritin Concentration - Female Athletes",
     xlab = "Ferritin Concentration", prob = TRUE)

lines(x_values, fun, lwd = 1.5)
```



Distribution of Ferritin Concentration − Male Athletes

Distribution of Ferritin Concentration − Female Athletes

Figure 2: Checking normality of Dependent Variable- Ferritin Concentration

```
par(mfrow = c(1, 1))
```

The dependent variable is not normally distributed based on the shape of the histogram. Transformation of the variable would make the variable normally distributed. In this case, I take the logatithm of the variable and rerun the t-test.

```
t.test(log(Ferr) ~ Sex, data = ferritin_data)
```

```
##
##  Welch Two Sample t-test
##
## data:  log(Ferr) by Sex
## t = -6.4009, df = 199.03, p-value = 1.086e-09
## alternative hypothesis: true difference in means between group female and group male is not equal to
## 95 percent confidence interval:
##  -0.6661013 -0.3523453
## sample estimates:
## mean in group female   mean in group male
##             3.903397             4.412621
```

The results of the t-test remain the same even after transforming the variable. Hence, we conclude that there is a significant difference in ferritin concentration between male and female athletes in Australia.

4.1.4 Outliers

The histogram shows that the dependent variable has outliers.

5 Regression Analysis: Drivers of Ferritin Concentration Among Athletes in Australia

We start by randomly divide the dataset into two sets, training (n1 = 141) and testing (n2 = 61).

```
set.seed(123) ## To allow for reproducibility
index <- sample(nrow(ferritin_data), size = 141, replace = FALSE)

training_set <- ferritin_data[index, ]
testing_set <- ferritin_data[-index, ]
```

5.1 Simple Linear Regression

We regress Ferririn Concentration (Ferr) against other variables as predictors except for the Sport variable (Draper and Smith 1998).

```
## Specifying regression model of ferritin against all variables except sport
my_regression_model <- lm(Ferr ~ . - Sport, data = training_set)
```

We can view the summary of the model.

```
summary(my_regression_model)
```

```
##
## Call:
## lm(formula = Ferr ~ . - Sport, data = training_set)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -85.04 -28.20  -9.65  20.76 131.22
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  52.6138    60.0586   0.876  0.38261
## Sexmale      81.3109    17.0984   4.755 5.14e-06 ***
## LBM          -2.1571     0.6511  -3.313  0.00119 **
## RCC         -11.1982    20.7055  -0.541  0.58954
## WCC           3.1449     2.0045   1.569  0.11907
## Hc           -4.6780     3.9536  -1.183  0.23886
## Hg           12.1649     9.6241   1.264  0.20847
## BMI           7.8100     2.6814   2.913  0.00421 **
## SSF          -0.4361     0.5634  -0.774  0.44024
## X.Bfat        2.2730     3.3869   0.671  0.50333
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42.86 on 131 degrees of freedom
## Multiple R-squared:  0.3001, Adjusted R-squared:  0.252
## F-statistic:  6.24 on 9 and 131 DF,  p-value: 2.621e-07
```

The model shows that Sex, lean body mass (LBM) and body mass index (BMI) are the significant associated with ferritin levels. Specifically, all other factors remaining constant, male athletes have higher average ferritin levels than females. Again, all other factors remaining constant, athletes with higher LBM have lower ferritin levels, and vice versa. Higher BMI corresponds to higher ferritin levels. Finally, the model is significant foing by the F-statistic.

5.2 Choosing a Model Using Backward Stepwise Regression

Next, we run a backward stepwise regression model- gradually removing insignificant predictors until all the variables in the model are statistically significant.

```
final_model <- step(my_regression_model, direction = 'backward')
```

```
## Start:  AIC=1069.36
## Ferr ~ (Sex + Sport + LBM + RCC + WCC + Hc + Hg + BMI + SSF +
##     X.Bfat) - Sport
##
##          Df Sum of Sq    RSS    AIC
## - RCC     1       537 241169 1067.7
## - X.Bfat  1       827 241459 1067.8
## - SSF     1      1101 241732 1068.0
## - Hc      1      2572 243203 1068.9
## - Hg      1      2935 243566 1069.1
## <none>                240631 1069.4
## - WCC     1      4522 245153 1070.0
## - BMI     1     15584 256215 1076.2
## - LBM     1     20161 260792 1078.7
## - Sex     1     41540 282171 1089.8
##
## Step:  AIC=1067.67
## Ferr ~ Sex + LBM + WCC + Hc + Hg + BMI + SSF + X.Bfat
##
##          Df Sum of Sq    RSS    AIC
## - X.Bfat  1       870 242039 1066.2
## - SSF     1      1214 242382 1066.4
## - Hg      1      2743 243912 1067.3
## <none>                241169 1067.7
## - WCC     1      4582 245751 1068.3
## - Hc      1      5647 246815 1068.9
## - BMI     1     15738 256907 1074.6
## - LBM     1     19734 260903 1076.8
## - Sex     1     41003 282171 1087.8
##
## Step:  AIC=1066.18
## Ferr ~ Sex + LBM + WCC + Hc + Hg + BMI + SSF
##
##        Df Sum of Sq    RSS    AIC
## - SSF   1       467 242506 1064.5
## - Hg    1      2440 244479 1065.6
## <none>              242039 1066.2
## - WCC   1      4544 246582 1066.8
## - Hc    1      5184 247223 1067.2
## - BMI   1     16467 258506 1073.5
## - LBM   1     21095 263133 1076.0
## - Sex   1     48315 290353 1089.8
##
## Step:  AIC=1064.45
## Ferr ~ Sex + LBM + WCC + Hc + Hg + BMI
##
##        Df Sum of Sq    RSS    AIC
## - Hg    1      2809 245314 1064.1
## <none>              242506 1064.5
## - WCC   1      4189 246695 1064.9
## - Hc    1      5207 247712 1065.5
## - BMI   1     20506 263012 1073.9
## - LBM   1     20817 263323 1074.1
## - Sex   1     54050 296556 1090.8
##
## Step:  AIC=1064.08
## Ferr ~ Sex + LBM + WCC + Hc + BMI
##
##        Df Sum of Sq    RSS    AIC
## - WCC   1      3150 248464 1063.9
## - Hc    1      3396 248711 1064.0
## <none>              245314 1064.1
## - LBM   1     23340 268655 1074.9
## - BMI    1     27815 273130 1077.2
## - Sex    1     63927 309241 1094.7
##
## Step:  AIC=1063.88
## Ferr ~ Sex + LBM + Hc + BMI
##
##         Df Sum of Sq    RSS    AIC
## - Hc    1      2524 250988 1063.3
## <none>              248464 1063.9
## - LBM   1     23887 272351 1074.8
## - BMI   1     30509 278973 1078.2
## - Sex   1     62863 311327 1093.7
##
## Step:  AIC=1063.3
## Ferr ~ Sex + LBM + BMI
##
##         Df Sum of Sq    RSS    AIC
## <none>              250988 1063.3
## - LBM   1     24140 275128 1074.2
## - BMI   1     29442 280430 1076.9
## - Sex   1     65462 316449 1094.0
```

In this case, we would choose the model `ferritin = Sex + LBM + BMI + error_term` as it has a lower AIC(1063.3) compared to the full model with an AIC of 1069.36. See the output summary for the chosen model below.

```
lm(Ferr ~ Sex + LBM + BMI, data = training_set) |> summary()
```

```
##
## Call:
## lm(formula = Ferr ~ Sex + LBM + BMI, data = training_set)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -79.61 -27.44 -10.22  23.19 129.95
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.6637    30.4826   0.022 0.982661
## Sexmale      72.1522    12.0704   5.978 1.86e-08 ***
## LBM          -2.2423     0.6177  -3.630 0.000399 ***
## BMI           8.1551     2.0343   4.009 9.98e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 42.8 on 137 degrees of freedom
## Multiple R-squared:  0.2699, Adjusted R-squared:  0.254
## F-statistic: 16.89 on 3 and 137 DF,  p-value: 2.17e-09
```

The model can be summarised as follows:

$Ferr = 0.6637 + 72.1522 Sex - 2.2423 LBM + 8.1551 BMI + error$

5.3 Regression Diagnostics

We then check the linear regression assumptions for the model fitted above.

Linear regression models draw from five key assumptions:

- Linear relationship.
- Multivariate normality.
- No or little multicollinearity.
- No auto-correlation.
- Homoscedasticity.

The model fails to meet the multivariate normality and homoscedasticity assumptions. The data also has extreme values.

The easiest way to check whether the model violates these assumptions is by plotting the model. Figure 3 below shows the diagnostics plots.

5.3.1 Linear Relationship

The relationship between the dependent variable (Ferritin concentration) and the dependent variables should be approximately linear. The `residual vs fitted` plot shows a roughly horizontal trend with no distinct pattern. Hence we conclude that the model meets this assumption.

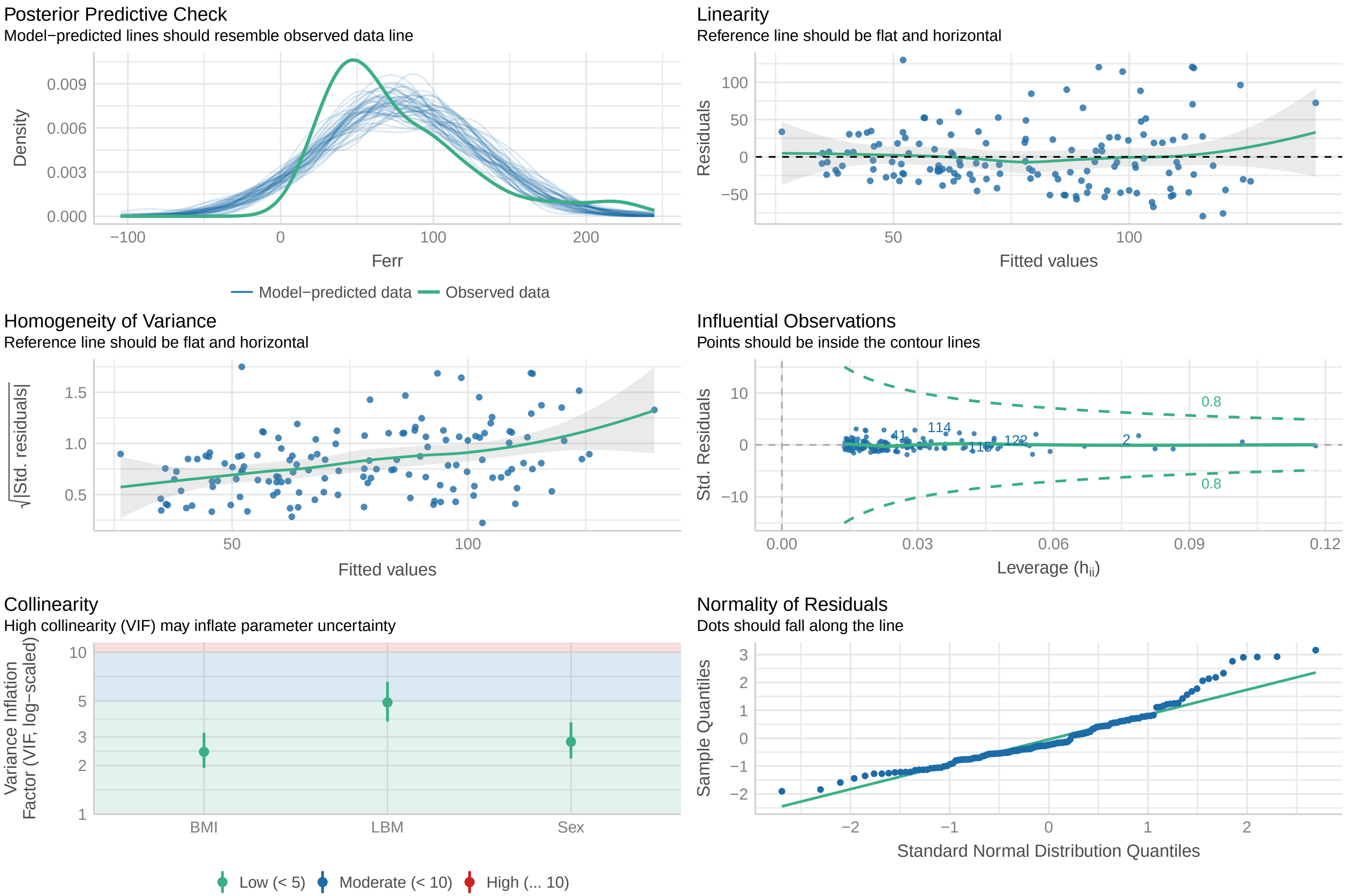


Figure 3: Checking Regression Model Assumptions

5.3.2 Multivariate normality.

The second plot `Normal Q-Q` shows the normality of residuals. In an ideal situation, the residuals should fall along the straight line. The figure shows a departure from this, more so at the edges. The model does not meet this assumption and cannot reliably be applied for forecasting. Indeed, the `Residuals vs Leverage` plot shows the existence of outliers.

5.3.3 No or little multicollinearity.

The variance inflation factor shows that multicollinearity is not an issue in the model. The rule of thumb is a VIF of less than 5 as the threshold for multicollinearity.

```r
vif(lm(Ferr ~ Sex + LBM + BMI, data = training_set))
```

```
##      Sex      LBM      BMI
## 2.799790 4.898196 2.427108
```

5.3.4 No auto-correlation

This assumption is not relevant as our data has no time series dimension.

5.3.5 Homoscedasticity

The `scale-location` plot shows a rising trend instead of a horizontal pattern with equally spread points. Again, the regression model does not meet this assumption. The model is heteroscedastic.

6 Conclusion

In this article, we assess the effect of a collection of explanatory variables on the plasma ferritin concentration (Ferr) among athletes in Australia. Sex, lean body mass (LBM) and body mass index (BMI) are the significant drivers of ferritin concentration among athletes. The difference in ferritin concentration is especially notable between male and female athletes. However, the regression model is is neither multivariate normal nor homoscedastic with notable outliers. Transforming the independent variables could improve the model.

References

Draper, Norman R, and Harry Smith. 1998. *Applied Regression Analysis*. Vol. 326. John Wiley & Sons.

Xu, Manfei, Drew Fralick, Julia Z Zheng, Bokai Wang, FENG Changyong, et al. 2017. “The Differences and Similarities Between Two-Sample t-Test and Paired t-Test.” *Shanghai Archives of Psychiatry* 29 (3): 184.