Assignment 2: Introduction to multiple linear regression

Matthew D. Sherman

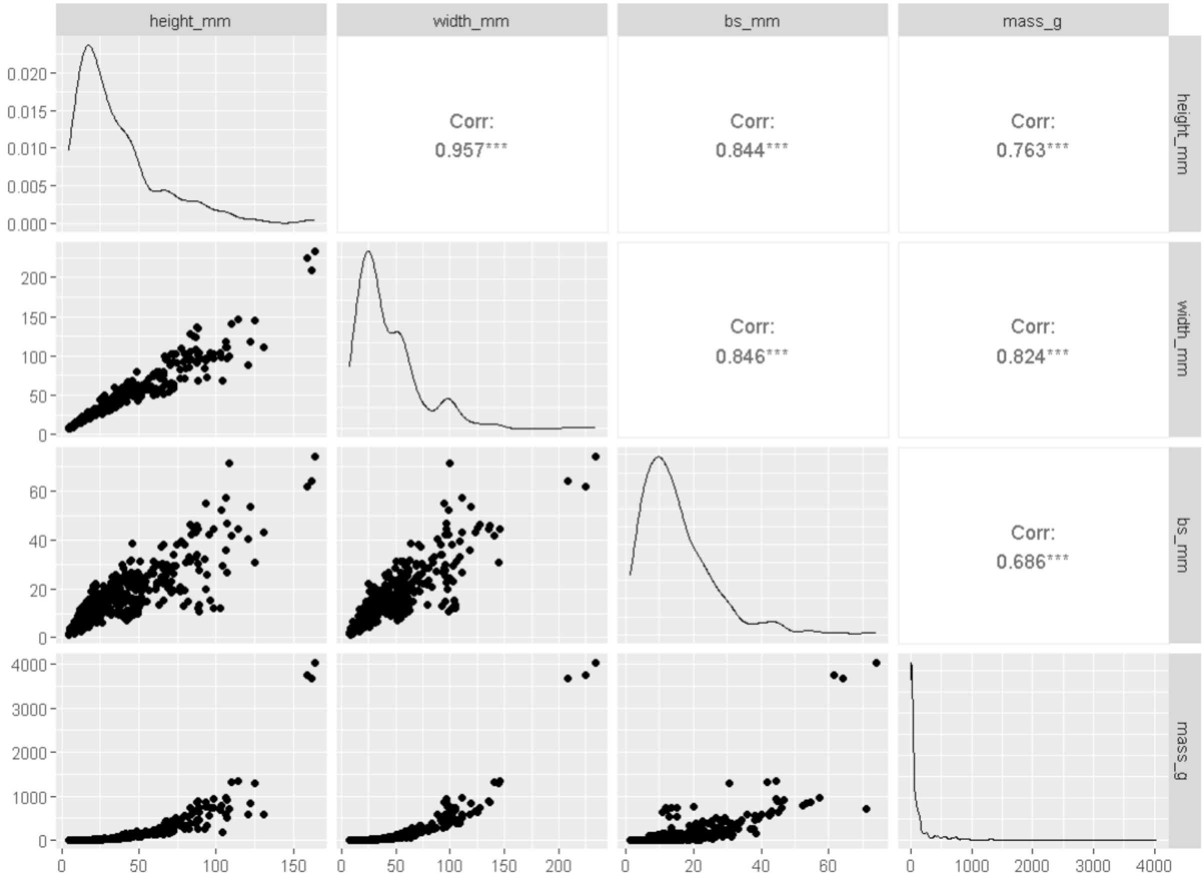
Department of Mathematics and Statistics, University of Canterbury

STAT202: Regression Modelling

August 10, 2023

Question One:

```
ggpairs(my_mammals, columns = c("height_mm", "width_mm", "bs_mm",
"mass_g"))
```



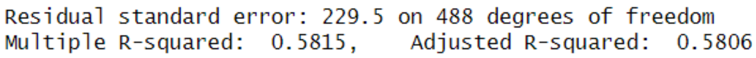
Comment on the relationships observed:

The relationships all appear to have a positive, linear correlation, with the data points roughly forming upward diagonal lines for each. They also all show a strong correlation (and have a correlation value over 0.7) apart from the x=bs_mm, y=mass_g relationship which shows a medium strength correlation (and has a correlation value in the range of 0.3-0.7).

Question Two:

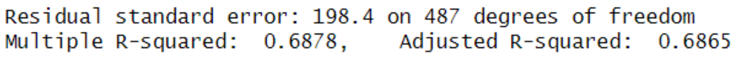
```
m1 <- lm(mass_g ~ height_mm, data = my_mammals)
summary(m1)
m2 <- lm(mass_g ~ height_mm + width_mm, data = my_mammals)
summary(m2)
m3 <- lm(mass_g ~ height_mm + width_mm + bs_mm , data = my_mammals)
summary(m3)
```

Model 1:



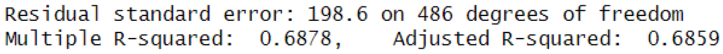
```
Residual standard error: 229.5 on 488 degrees of freedom
Multiple R-squared:  0.5815,    Adjusted R-squared:  0.5806
```

Model 2:



```
Residual standard error: 198.4 on 487 degrees of freedom
Multiple R-squared:  0.6878,    Adjusted R-squared:  0.6865
```

Model 3:



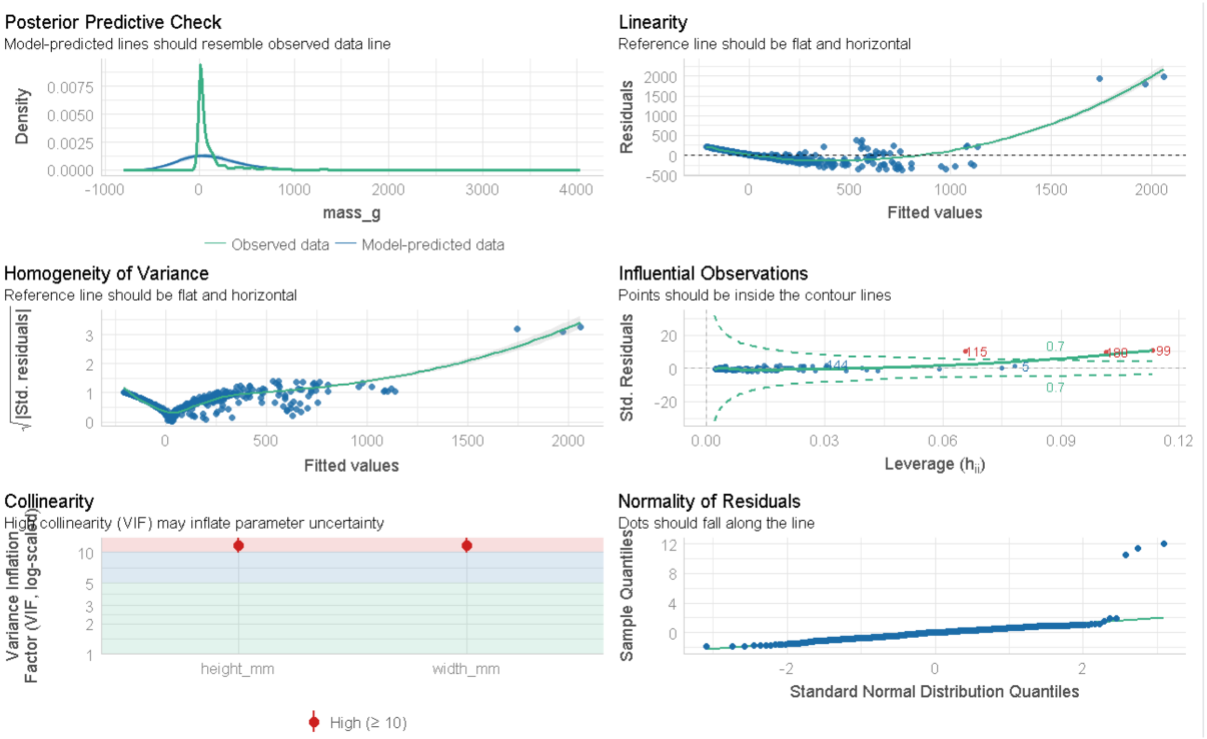
```
Residual standard error: 198.6 on 486 degrees of freedom
Multiple R-squared:  0.6878,    Adjusted R-squared:  0.6859
```

Comment on the improvement of fit when moving from M1 through M3:

The fit improves from M1 to M2 due to higher multiple and adjusted R^2 values, as well as a lower RSE. Then the fit deteriorates very slightly from M2 to M3 as M3 has the same multiple R^2 as M2, but less favourable (in terms of fit) adjusted R^2 and RSE. M3 is still an overall improvement over M1 regardless.

Question Three:

```
check_model(m2)
```



Comment:

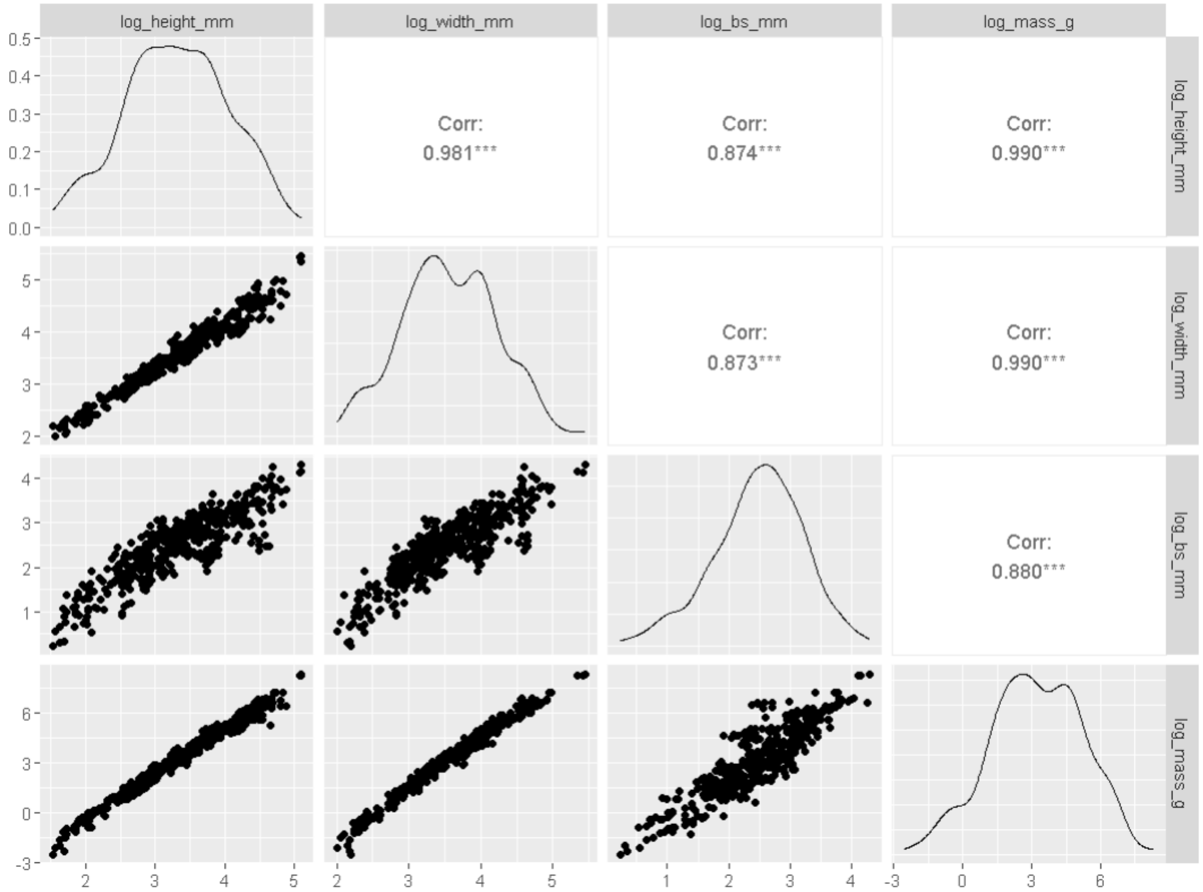
In the IO model, there are three points outside the contour lines which indicates that they have high influence and could be considered potential outliers. The linearity, HoV, collinearity, and PPC models all show unfavourable results (indicating poor residual pattern, spread, and distribution). The model should be refined to achieve more favourable results.

Question Four:

```
my_mammals_log <- my_mammals %>% mutate_if(is.numeric, log)
```

Question Five:

```
ggpairs(my_mammals_log, columns = c("height_mm", "width_mm", "bs_mm", "mass_g"))
```



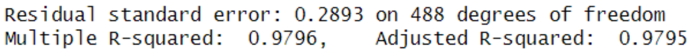
Comment on the change in relationships:

The correlation strength for every relationship has improved and now all relationships are strongly correlated. All relationships have improved in terms of linearity and all still all have a positive direction.

Question Six:

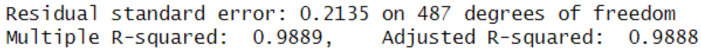
```
m4 <- lm(mass_g ~ height_mm, data = my_mammals_log)
summary(m4)
m5 <- lm(mass_g ~ height_mm + width_mm, data = my_mammals_log)
summary(m5)
m6 <- lm(mass_g ~ height_mm + width_mm + bs_mm , data = my_mammals_log)
summary(m6)
```

Model 4:



```
Residual standard error: 0.2893 on 488 degrees of freedom
Multiple R-squared:  0.9796,	Adjusted R-squared:  0.9795
```

Model 5:



```
Residual standard error: 0.2135 on 487 degrees of freedom
Multiple R-squared:  0.9889,	Adjusted R-squared:  0.9888
```

Model 6:

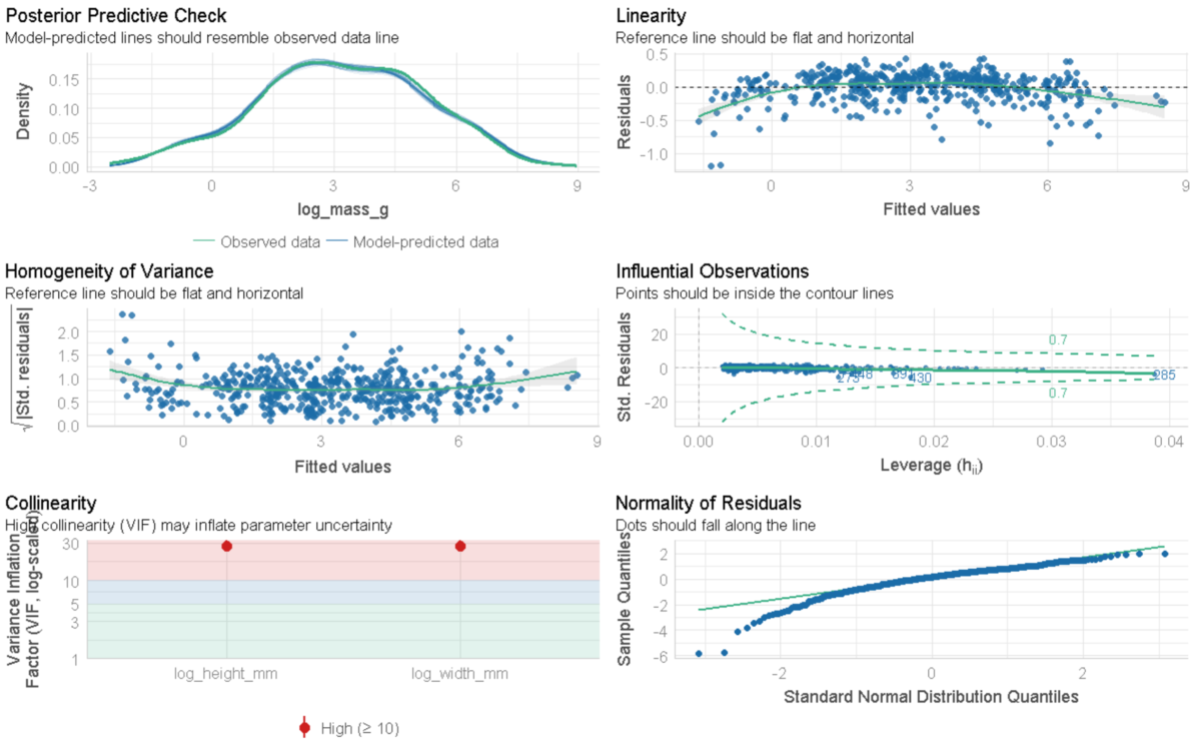
```
Residual standard error: 0.2114 on 486 degrees of freedom
Multiple R-squared:  0.9891,	Adjusted R-squared:  0.9891
```

Comment:

There is a consistent improvement in fit moving from model 4 to model 6. The fit improves from M4 to M5 due to M5 having a lower RSE and higher multiple R^2 and adjusted R^2 values. The fit improves from M5 to M6 due to M6 having a lower RSE and higher multiple R^2 and adjusted R^2 values.

Question Seven:

```
check_model(m5)
```



Comment:

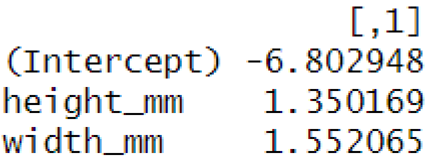
There are no longer any potential outliers as shown by all the data points fitting between the contour lines of the IO model. The M5 PPC, IO, linearity, and HoV models are generally satisfactory, and all show a significant improvement compared to the equivalent M3 models (indicating superior residual pattern, spread, and distribution). Overall, M5 is an improvement over M3.

Question Eight:

```
X <- model.matrix(m5)
y <- my_mammals_log$mass_g
```

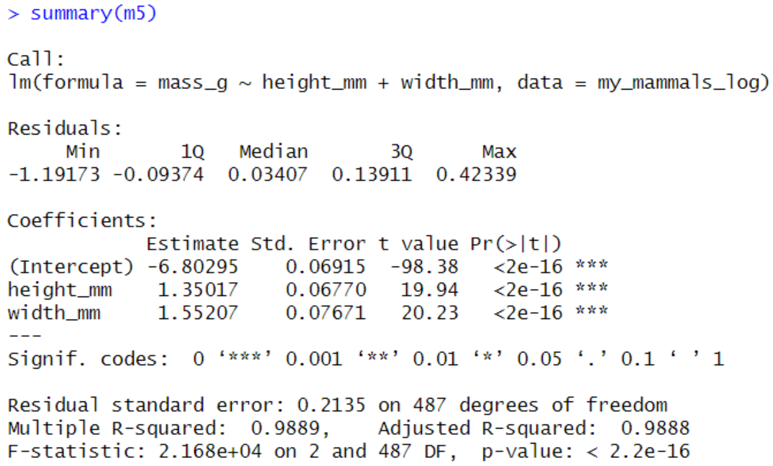
```
XtX <- t(X) %*% X
Xty <- t(X) %*% y
b <- solve(XtX) %*% Xty

b
```



```
                  [,1]
(Intercept) -6.802948
height_mm    1.350169
width_mm     1.552065
```

Reproduces M5 coefficients:



```
> summary(m5)

Call:
lm(formula = mass_g ~ height_mm + width_mm, data = my_mammals_log)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.19173 -0.09374  0.03407  0.13911  0.42339 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -6.80295    0.06915  -98.38   <2e-16 ***
height_mm    1.35017    0.06770   19.94   <2e-16 ***
width_mm     1.55207    0.07671   20.23   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2135 on 487 degrees of freedom
Multiple R-squared:  0.9889,	Adjusted R-squared:  0.9888 
F-statistic: 2.168e+04 on 2 and 487 DF,  p-value: < 2.2e-16
```

Full Code:

```
library(tidyverse)
library(performance)
library(GGally)

set.seed(16645573)
mammals <- read_csv('endocranial_volume.csv')
my_mammals <- mammals %>% sample_n(490)

ggpairs(my_mammals, columns = c("height_mm", "width_mm", "bs_mm", "mass_g"))

m1 <- lm(mass_g ~ height_mm, data = my_mammals)
summary(m1)
m2 <- lm(mass_g ~ height_mm + width_mm, data = my_mammals)
summary(m2)
m3 <- lm(mass_g ~ height_mm + width_mm + bs_mm , data = my_mammals)
summary(m3)

check_model(m2)

my_mammals_log <- my_mammals %>% mutate_if(is.numeric, log)

ggpairs(my_mammals_log, columns = c("height_mm", "width_mm", "bs_mm", "mass_g"))

m4 <- lm(mass_g ~ height_mm, data = my_mammals_log)
summary(m4)
m5 <- lm(mass_g ~ height_mm + width_mm, data = my_mammals_log)
summary(m5)
m6 <- lm(mass_g ~ height_mm + width_mm + bs_mm , data = my_mammals_log)
summary(m6)

check_model(m5)

X <- model.matrix(m5)
y <- my_mammals_log$mass_g
```

```
XtX <- t(X) %*% X
Xty <- t(X) %*% y
b <- solve(XtX) %*% Xty

b
```