**STAT GR5205 Group Project**

## **Bird Body Size and Life History (Lislevand et al. 2007)**

Group Name: 925ers

Group Members: Meiyu Chen(mc5261), Tianhong Chen(tc3234), Yuxue Cui(yc4145), Yuchen Wang(yw3890), Hongjun Zeng( hz2815)

**Introduction**

The dataset we researched is “Bird Body Size and Life History (Lislevand et al. 2007).” We are interested in having a deeper understanding on bird body size with the large data set containing a wide range of variables of studies on birds. What’s more, before choosing the data, our group read sources related to birds. For example, in an extremely interesting article called “Why Birds Matter”, the author analyzes that different kinds of birds with diverse bodies will represent various fecundity, mating systems or the behavior. Therefore, we are surprised that just a very tiny organism could make a great effect on their physiology, life cycles, and overall ecology, and finally choose this dataset as our main topic.

The dataset extracted species-specific measurements on male and female bird body mass, wing length, tarsus length, bill length, and tail length from different regions on earth. These measurements were matched with measures of egg and clutch sizes, and scores of mating system, sexual display agility, and the degree of intersexual resource division. In total, 3769 species from 125 of 146 different bird families are included. We have used comparative studies of avian sexual size dimorphism, where we test adaptive hypotheses concerning the influence of sexual selection, fecundity, and the degree of within-pair resource sharing.

We focused on the relationship between male and female bird’s mass and size of tail, wing, tarsus and bill, mating competition, sexual display behavior, and inner sexual resource sharing. We mainly analyze the relationship between three categories of behavioral scores and bird body size.

**Analysis of data**

Before we analyze the data, we check the data validity at first. There is some missing data in the dataset: “if no information is available for a given record, this is indicated by -999.” This will lead to possible errors as the software interprets it as an existing value. So we replaced all “-999” in the data with NA values.

The coefficient of the simple linear regression (lm(M\_tail~F\_tail) is 1.033, and the intercept is 1.227 *(See Figure 1).* Therefore, there is a positive correlation between the tail length of males and the tail length of females *(See Figure 2)*. The tail length of females tends to increase in tandem with the tail length of males. In addition, the p-value is less than the significance level, so the sample data provide sufficient evidence to reject the null hypothesis for the entire population. As we can see from the simple linear regression (lm(M\_tail\_N~F\_tail\_N) *(See Figure 3)* , the coefficient is 1.12, so there is a positive correlation between the sample size for male tail length data and the sample size for female tail length data. However, the majority of data is gathered between 0 and 200. From the plot(M\_tail\_N~M\_tail) *(See Figure 4)* and plot(F\_tail\_N~F\_tail) *(See Figure 5)*, we can see that the length of male and female tails in most samples is between 0 and 300.

From summary of the simple regression (M\_mass~M\_wing) *(See Figure 6)*, male bird’mass is positively correlated with male bird’s wing size with the slope coefficient of 6.55. The r-squared value shows that about 27.29% of the mass change is explained by the change in wing size, which is not a large portion. We assumed the linear relationship between the two variables existed while there were several potential outliers appearing on the scatterplot *(See Figure 7)* detected by the outlierTest() function *(See Figure 8).* From the residual plot *(See Figure 9)* and qqnorm plot *(See Figure 10)*, a problem can be seen that the residuals are not actually dispersed randomly and symmetrically around the mean of 0, rather, it shows a trend to curve upward as the wing size increases. And this indicates that the linear model above may not be very appropriate.

Then, we try to find transformations to make the transformed predictors as close to linearly related as possible. From the summary of using powerTransform() function to the model *(See Figure 11)*, the p-value in the likelihood ratio test that transformation parameter is equal to 0 (log transformation) is 0.19. The value is large and thus we cannot reject the null hypothesis. So it is suggested to transform the predictor to log scale form. Further, we tried to explore whether a log transformation of the response is also needed with invResPlot() function *(See Figure 12)*. From the inverse response plot, we found that the fitted line of 0 power matched the nonlinear least squares estimate much better and therefore the log transformation of the response is also reasonable. Finally, we summarized the new model (log(M\_mass)~log(M\_wing)) and found the r-squared value is very large this time of 88.07% *(See Figure 13)*. Also, the visualization of the model from the scatter plot fits the linear relationship much better *(See Figure 14)*.

From the summary table of our multiple regression model (M\_mass~M\_wing+M\_tail+M\_tarsus+M\_bill) *(See Figure 15)*, the r-squared value increased to 62.06%, which is much larger than the single regression model (M\_mass~M\_wing) before , as we added more predictor variables that we believed would have influence on the response into the model. We further proved that all the regressors in the model are essential based on the small p-value and the obvious slopes in the added-variable plot *(See Figure 16)*, which means that the size of wing, tail, tarsus and bill are all important factors in predicting a male bird’s mass. In the result, the tail is the only factor with a negative coefficient. In other words, the male bird mass is expected to decrease as their tail size increases, which is not reasonable. A potential explanation is that m\_tail is correlated with other factors already included in the function and its influence has already been included in the model. However, we have not further explored this problem and will try to find the true reason if more time is available.

From the boxplot *(See Figure 17)* and the lines in the scatterplot *(See Figure 18)*, it can be seen that male birds in different sexual display groups, especially group 5 (mainly aerial displays, acrobatic), should have different mean functions on the relationship of their wing size and mass. Considering the sexual display group level as a dummy variable, we tried the models with different intercepts and a shared slope first and got a small r-squared value *(See Figure 19)*. Then we added an interaction variable of the mass display group and found that they all have small p-values *(See Figure 20)*. By using F-test to compare the reduced model (M\_wing~M\_mass+factor(Display)) and full model (M\_wing~M\_mass\*factor(Display)) *(See Figure 21)*, a very small p-value is obtained, which indicates that interaction variable is essential in the model and different slopes are necessary for different display groups. In other words, the relationships between male bird mass and their wing size are different in different sexual display groups. However, we noticed the sample size of different groups varies a lot and the size of group 4 and 5 are very small. This may lead to a deviation of the data and so we would like to validate this conclusion when we have more sources of data available.

From the regression of the lm(M\_tarsus~Egg\_mass,data=jan07) *(See Figure 22)*, we could find that there is a strong relationship between Male\_tarsus and the Egg mass. For example, in the plot of the Egg\_Mass and M\_tarsus *(See Figure 23)*, it is clear that the observation data is increasing, and there is a positive relationship between these two variables, which suggests that with the increase of the Egg\_Mass, M\_tarsus will strictly increase and it follows the rules of linear regression. This relationship indicates that body size could make a great effect on the next generation and precisely plays an important characteristic of animals, influencing physiology, life histories, and general ecology. Also, from the summary of the data, the slope equals to 0.217084 and the p-value infinitely close to 0, which means that it is readily to reject the null hypothesis.

Besides, from the regression of log(M\_tarsus)~M\_wing+M\_tail+Egg\_mass+M\_bill *(See Figure 24)*, we firstly utilize the Added Variable Plots *(See Figure 25)* to do the research of the birds. Basically, we could still see that the tarsus of the birds is related to the bird's wing, tail, egg, and bill. And, from the slope, it indicates a slight positive relationship, which suggests the importance of birds’ physiology, life histories, and general ecology with the change of the bird’s body size. However, from the summary, we could detect that the p-value of the M\_tail is greater than 0.05, which suggests the data is not a critical variable in the model. Actually, from the plot, there is no sufficient linear regression relationship that can precisely demonstrate the data. Simultaneously, the p-value of the rest of the variables are extremely small, indicating that it successfully rejects the null hypothesis.Therefore, from the whole regression, it still can strongly demonstrate that the relationship between bird’s wing, egg’s mass and bill depends on their tarsus which affect their life, history, and their own body.

There are five categories of mating systems. According to the score, male-male competition becomes more intense from 1 to 5. The p-value of the length of the female bird wing and mating system is near 0 when the F-test *(See Figure 28)* is used to check the reduced model (m8<-lm(Egg\_mass~F\_wing+factor(Mating\_System)) *(See Figure 26)*​​ and full model (m9<-lm(Egg\_mass~F\_wing\*factor(Mating\_System)) *(See Figure 27)*. As a result, the relationship between egg mass and female wing length varies depending on mating systems. Furthermore, the p-value *(See Figure 31)*of length of female and male bird wings and mating systems is near 0 from the reduced model (m10<-lm(Egg\_massM\_wing+F\_wing+factor(Mating\_System)) *(See Figure 29)* and full model (m11<-lm(Egg\_massM\_wing+F\_wing\*factor(Mating\_System)) *(See Figure 30)*. Therefore, the relationship between egg mass and female and male wing length depends on mating systems. However, the R-squared of m9 is 0.557 and the R-squared of m11 is 0.619. So the relationship between egg mass and both female and male wing length fits the data better.

For checking assumptions, we did scatterplot, outlierTest, avPlots() from the above analysis. Then, we also checked normality, homoscedasticity and multicollinearity using shapiro.test(), ncvTest() and vif()(*See Figure 32*), the result shows all of our models have very low P-value(2.22e-16), which means we reject the null hypothesis, and it means that the model has no normality and non constant error variance. And for the multicollinearity test in our multiple regression model, most of the values are less than 5, so it means there is no multicollinearity problem. But for the m7 model(*See Figure 33*) M-wings, the value is 5.0138, which is larger than 5. So m7 has a multicollinearity problem. Usually, deleting the variables that have the largest value in the test and fitting a new model can help.

**Conclusion**

**From the plots and analysis results, we conclude that**

1)The tail length of females tends to increase in tandem with the tail length of males.

2) There is a trend to curve upward as the wing size increases.

3) The size of wing, tail, tarsus and bill are all important factors in predicting a male bird’s mass (the male bird mass is expected to decrease as their tail size increases, which is not reasonable.)

4) The relationships between male bird mass and their wing size are different in different sexual display groups.

5) There is a strong relationship between Male\_tarsus and the Egg mass. This relationship indicates that body size could make a great effect on the next generation and precisely plays an important characteristic of animals, influencing physiology, life histories, and general ecology.

6) The relationship between bird’s wing, egg’s mass and bill depends on their tarsus which affects their life, history, and their own body.

7) The relationship between egg mass and both female and male wing length fits the data better.

8) The relationship between egg mass and female wing length varies depending on mating systems.

9) There is no multicollinearity in the dataset, except a specific model.

**Discussion**

If there were more time and resources, we would like to expand and improved our analysis by:

1)Standardizing based on the scale or potential range of the data.

2)Adding interaction terms like the multiple of two variables to plot the interactions effect plot about how two or more independent variables together impact the target variable.

3)Add polynomial terms to model the nonlinear relationship between an independent variable and the target variable.

When exploring complex research questions, mixed methods designs can provide practical benefits for combining methods. The process of quantitative qualitative analytical and survey data can be time consuming and expensive. Finally, these data can be appropriate for research that does not require in-depth analysis.

If there is more time available, we would like to collect more videos or images of data from various areas to do the research of avian body sizes in relation to fecundity, mating system, display behavior, and resource sharing. Like the article “Why Birds Matter”, the author analyzes that different kinds of birds with diverse bodies will represent various fecundity, mating systems or the behavior.

**Appendix**

**Appendix1: Individual Contribution**

Meiyu Chen(mc5261): summarize the analysis result, write the intro and conclusion.

Tianhong Chen(tc3234): analyzed the simple and multiple regressions and summarized the result

Yuxue Cui(yc4145): Analyzed and visualized the dataset with the software, explored the reason for some unreasonable data, organized the whole structure

Yuchen Wang(yw3890): Checked the assumption, did the diagnostic for all models, found the possible errors and wrote improvement.

Hongjun Zeng( hz2815):data analysis about the tarsus, analyze the plot, summary, and explain the unreasonable data.

**Appendix2:Link to the dataset**

<https://figshare.com/articles/dataset/Data_Paper_Data_Paper/3527864?file=5599229>

**Appendix3: R Code and Output**

Figure1-m1<-lm(M\_tail~F\_tail,data=jan07)

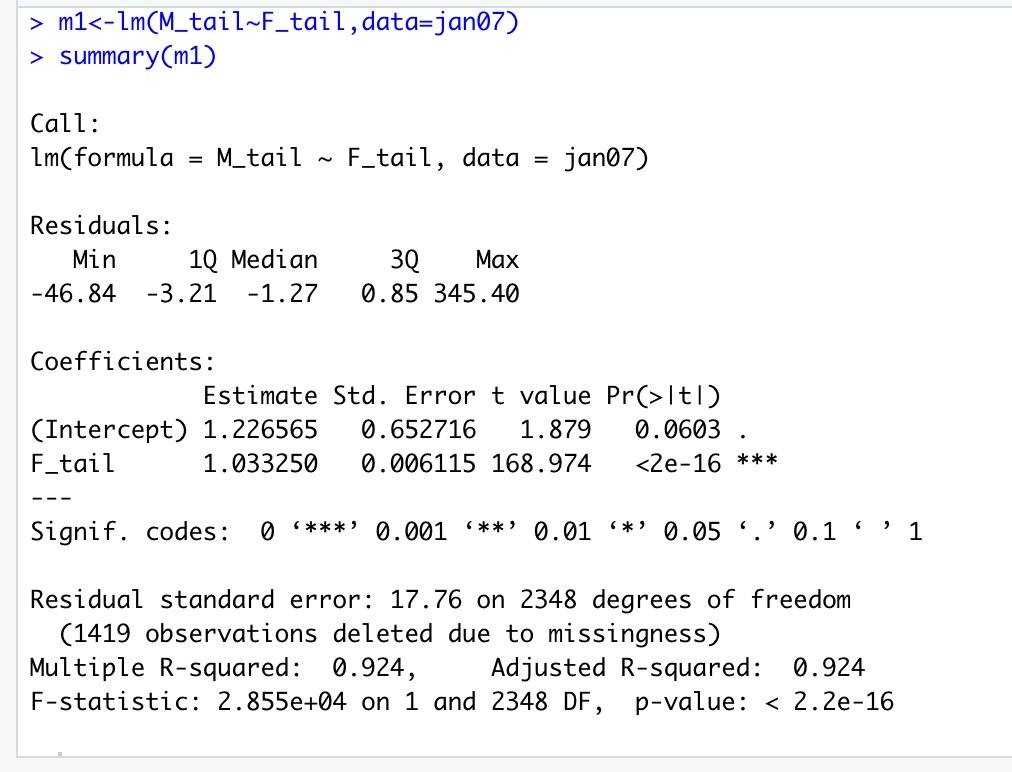


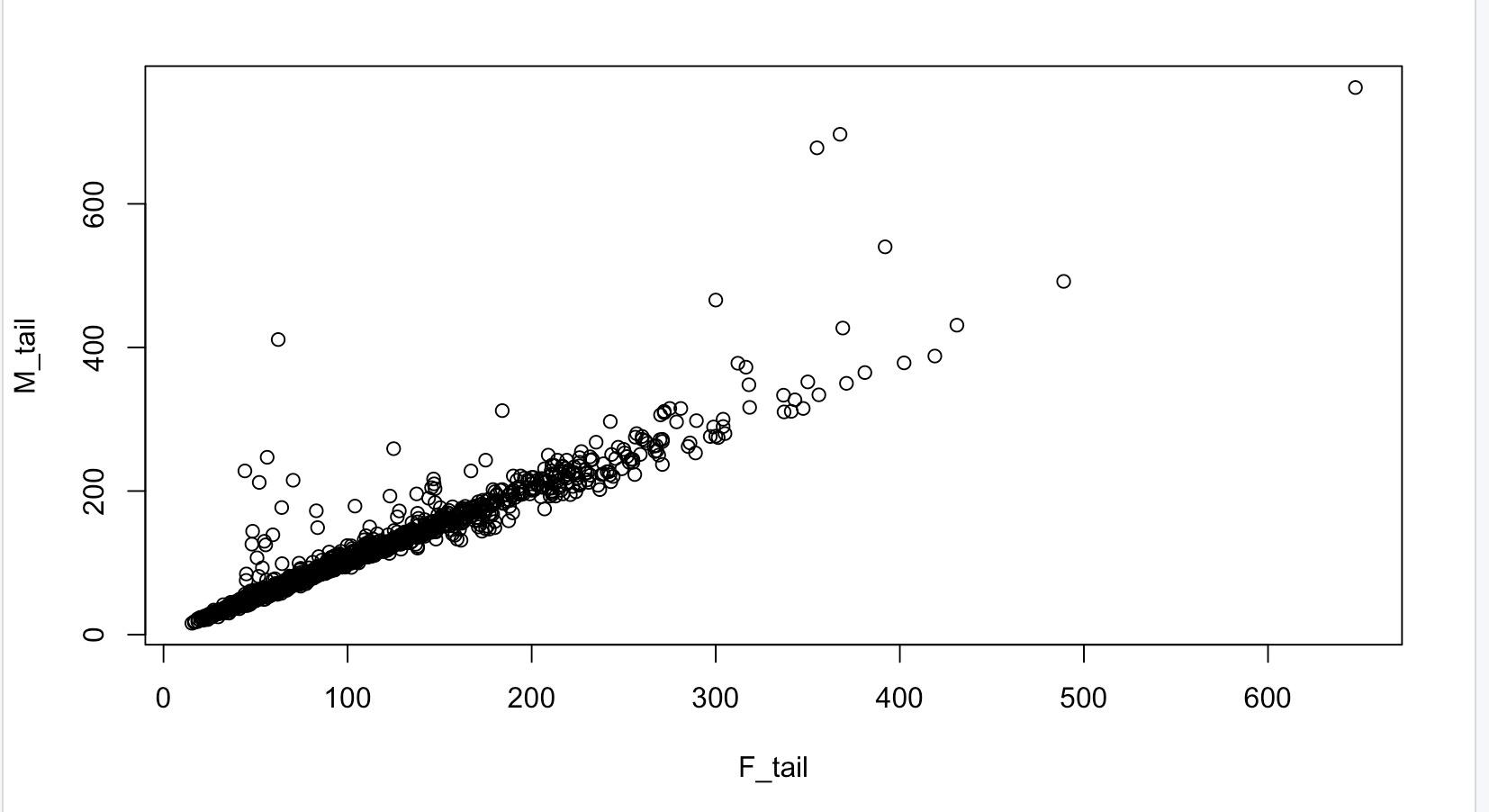
Figure2-plot(M\_tail~F\_tail,data=jan07)

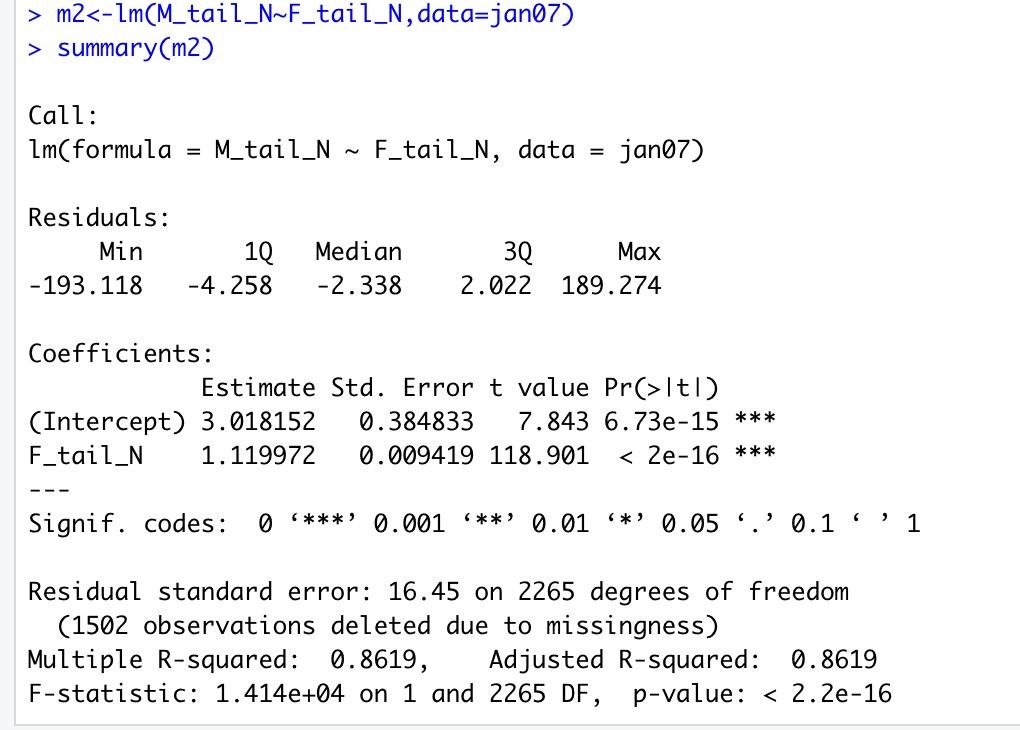
Figure3-m2<-lm(M\_tail\_N~F\_tail\_N,data=jan07)

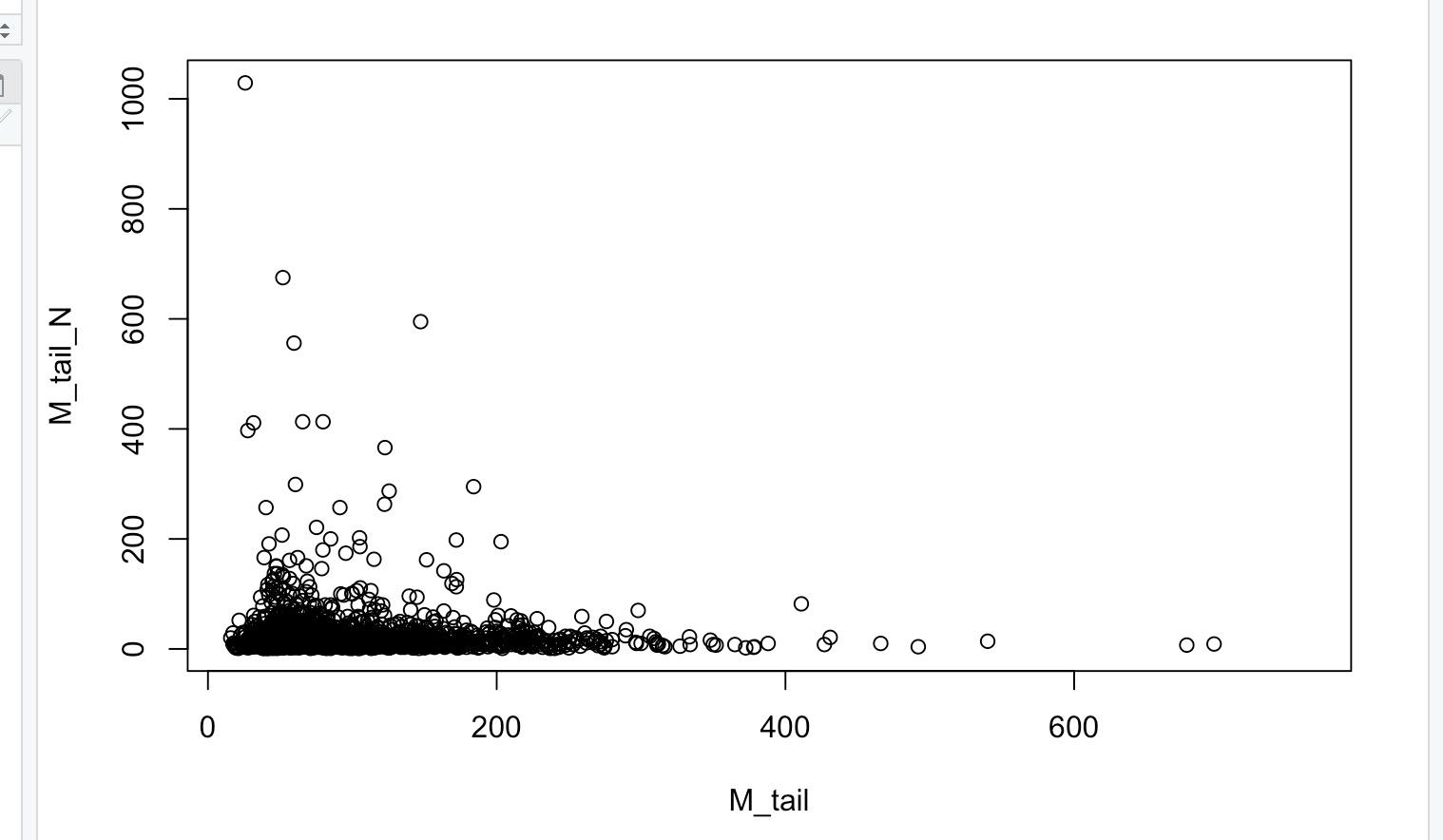
Figure4-plot(M\_tail\_N~M\_tail,data=jan07)

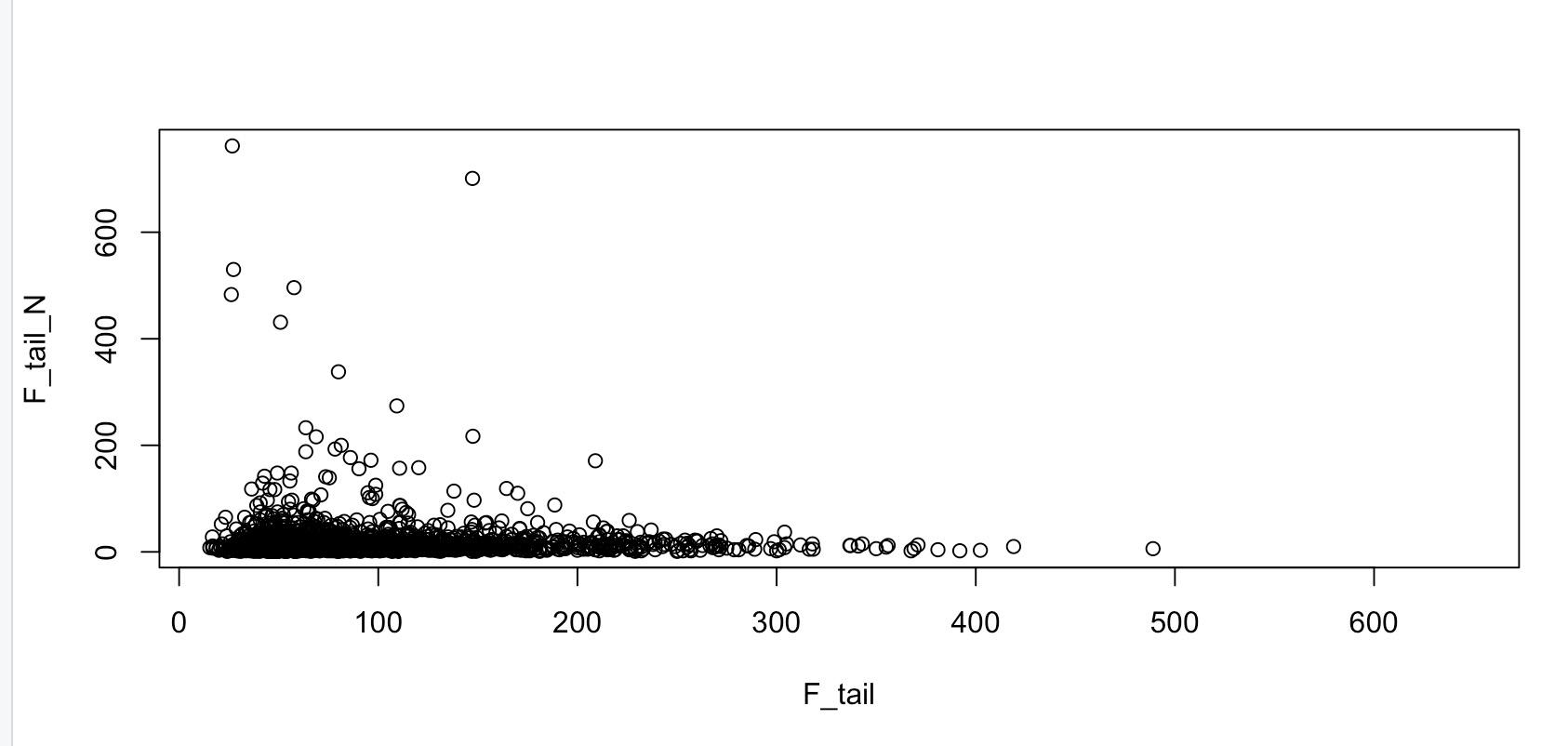
Figure5-plot(F\_tail\_N~F\_tail,data=jan07)

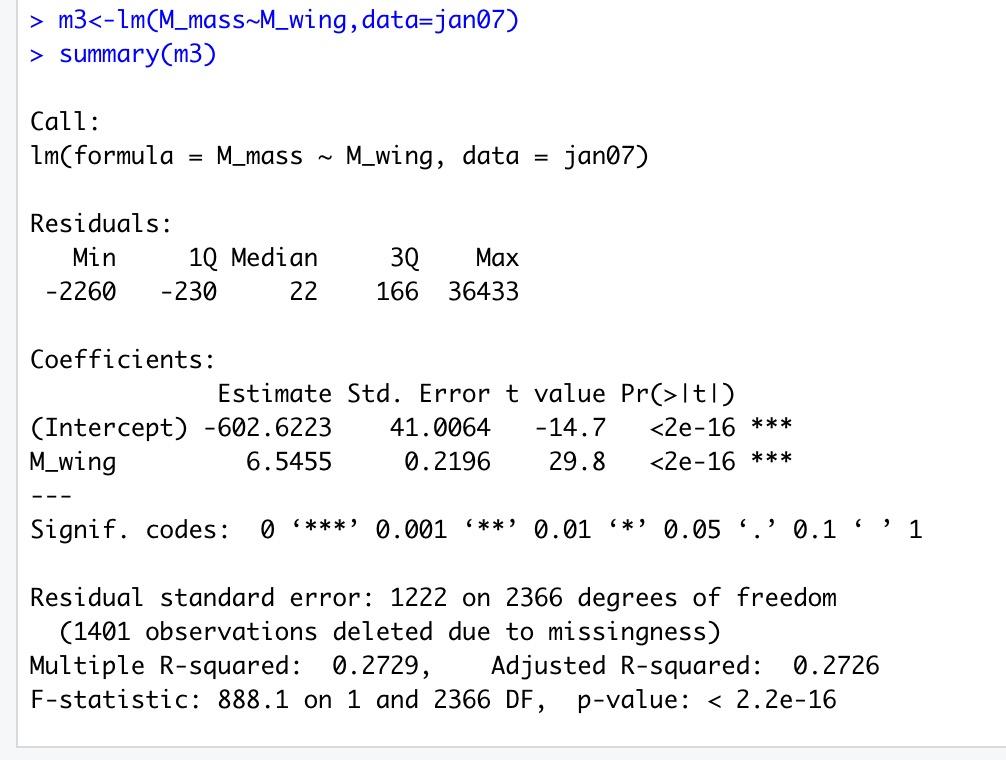
Figure6-m3<-lm(M\_mass~M\_wing,data=jan07)

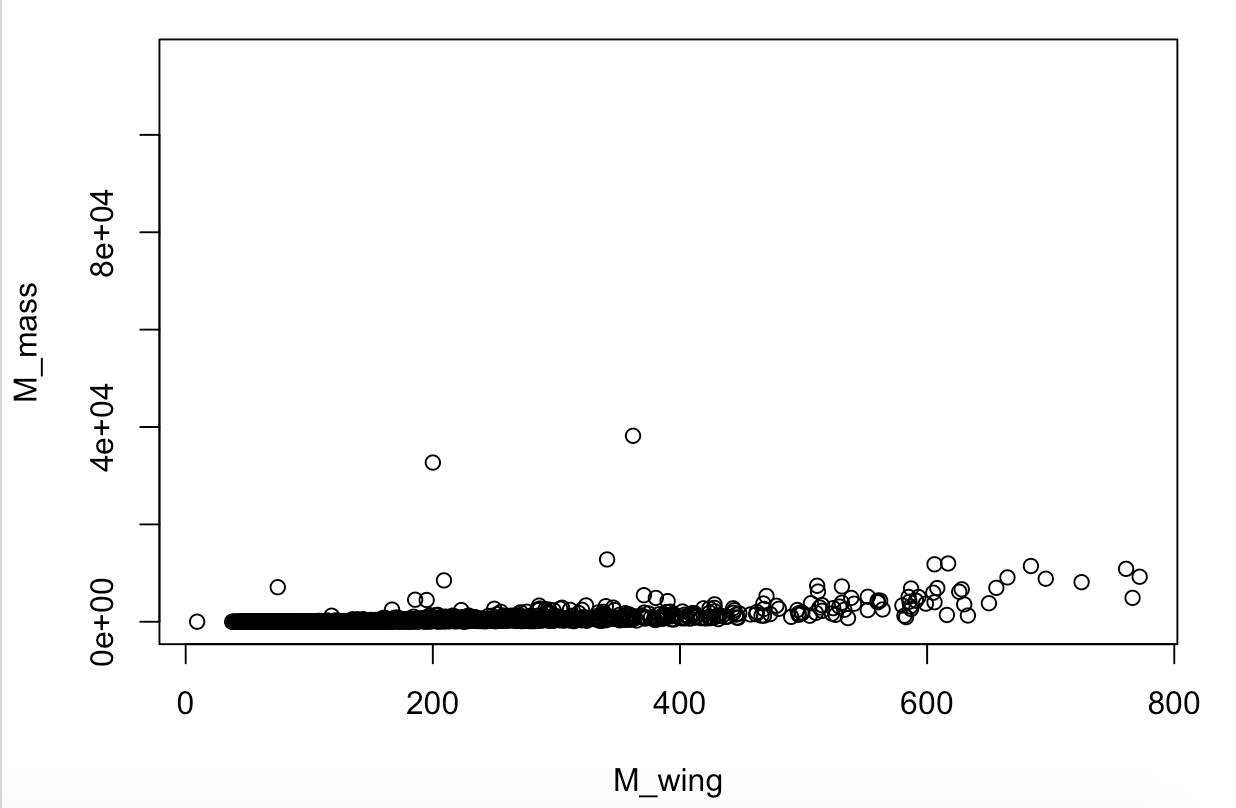
Figure7-plot(M\_mass~M\_wing,data=jan07)

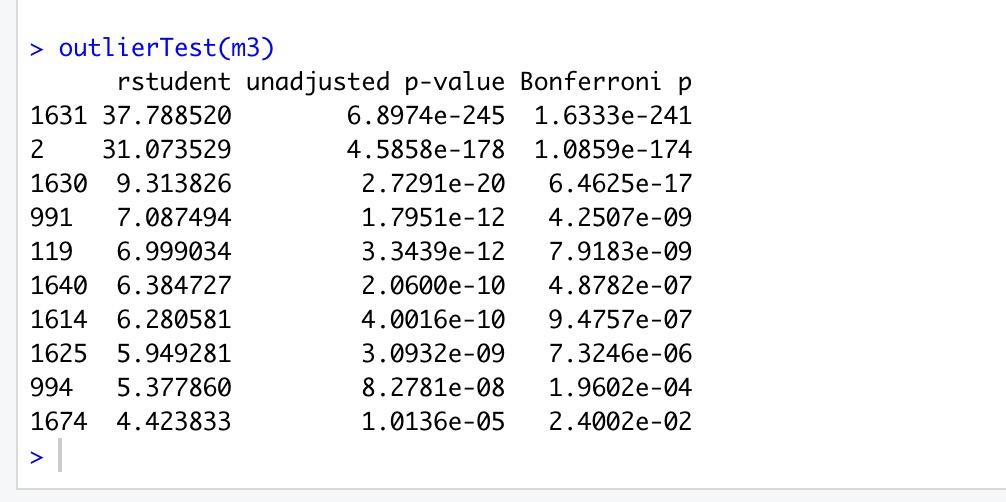
Figure8-outlierTest(m3)

Figure9-plot(predict(m3),residuals(m3),data=jan07)

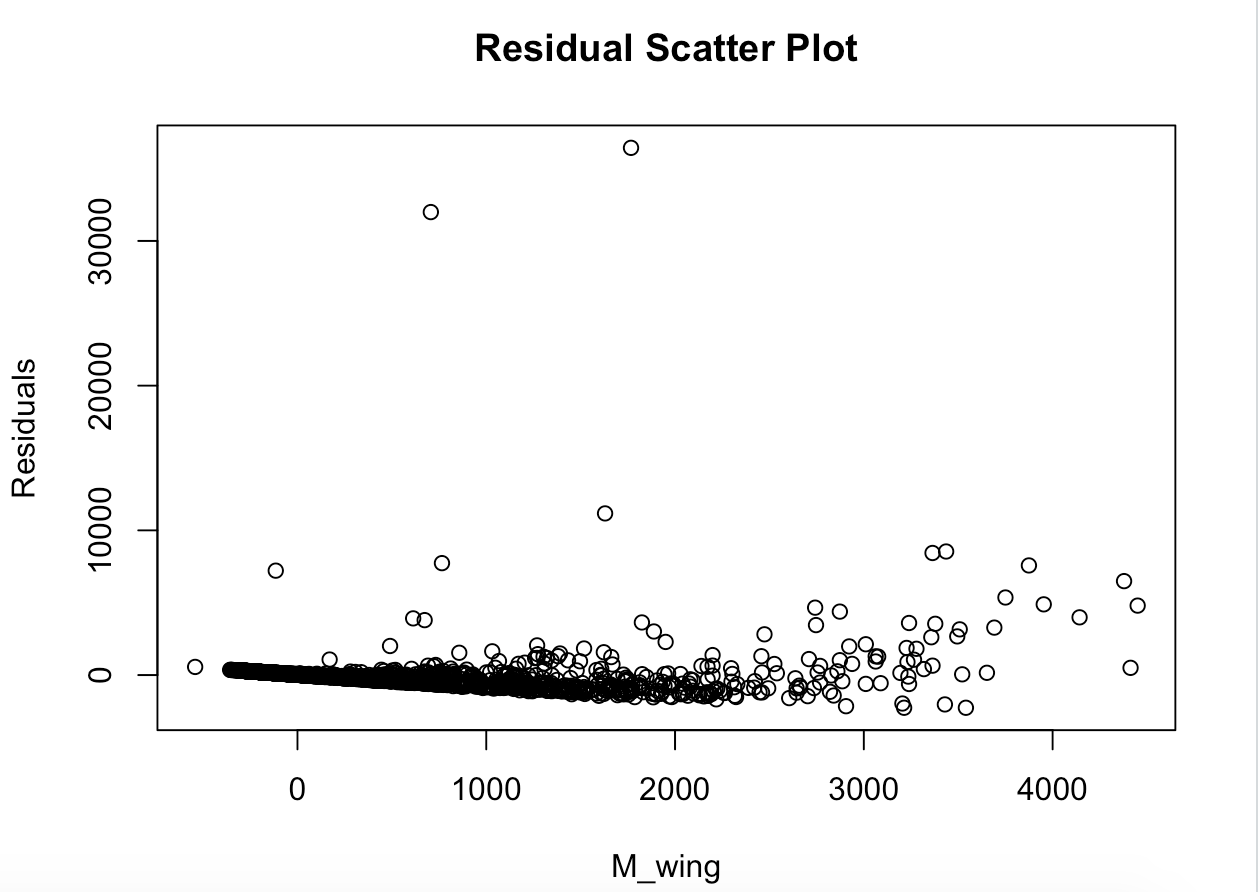


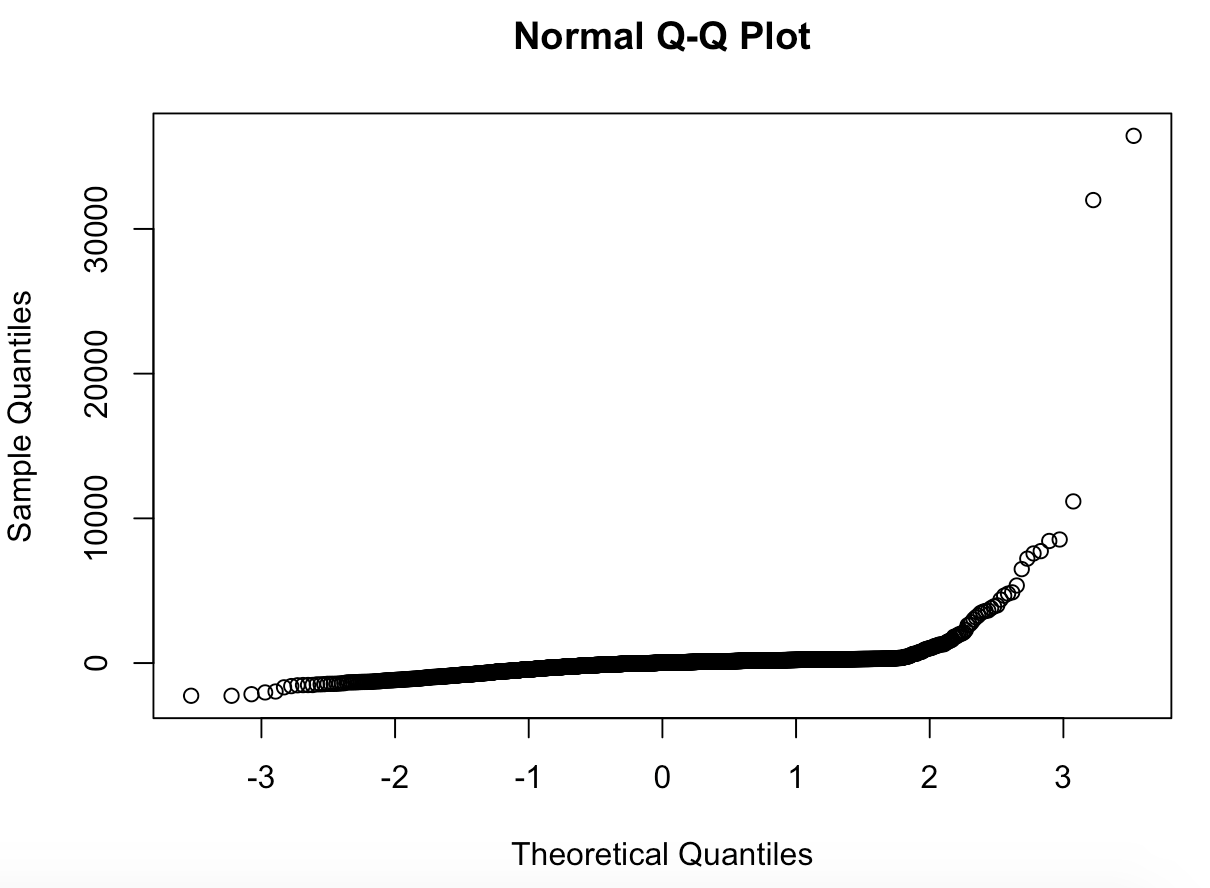
Figure10-qqnorm(residuals(m3))

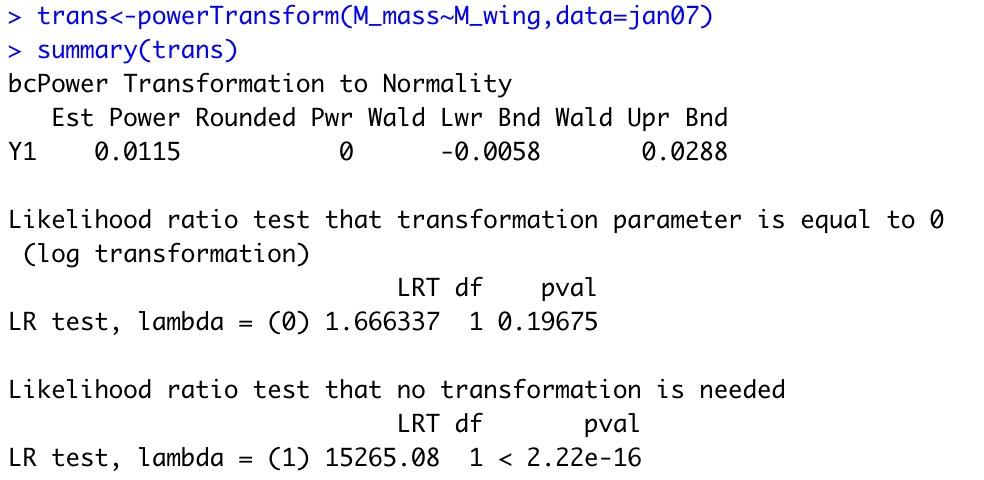
Figure11-trans<-powerTransform(M\_mass~M\_wing,data=jan07)

Figure12-invResPlot(lm(M\_mass~log(M\_wing),data=jan07))

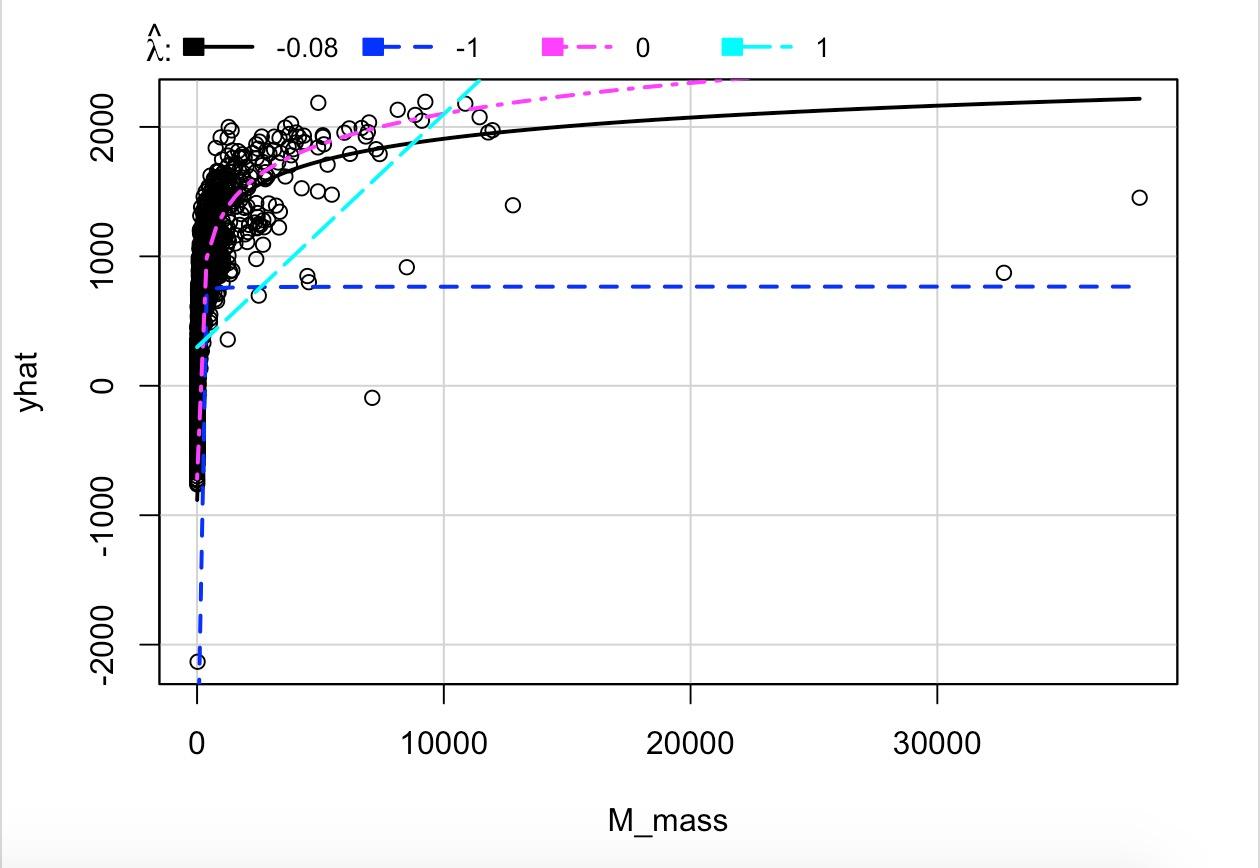


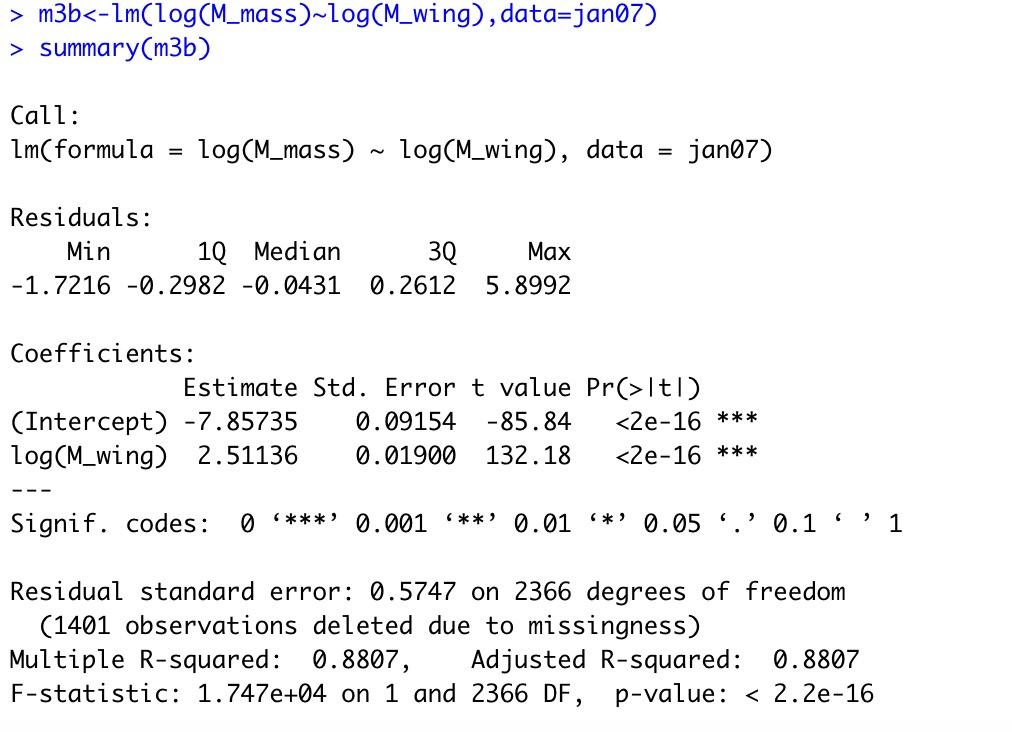
Figure13-m3b<-(log(M\_mass)~log(M\_wing),data=jan07)

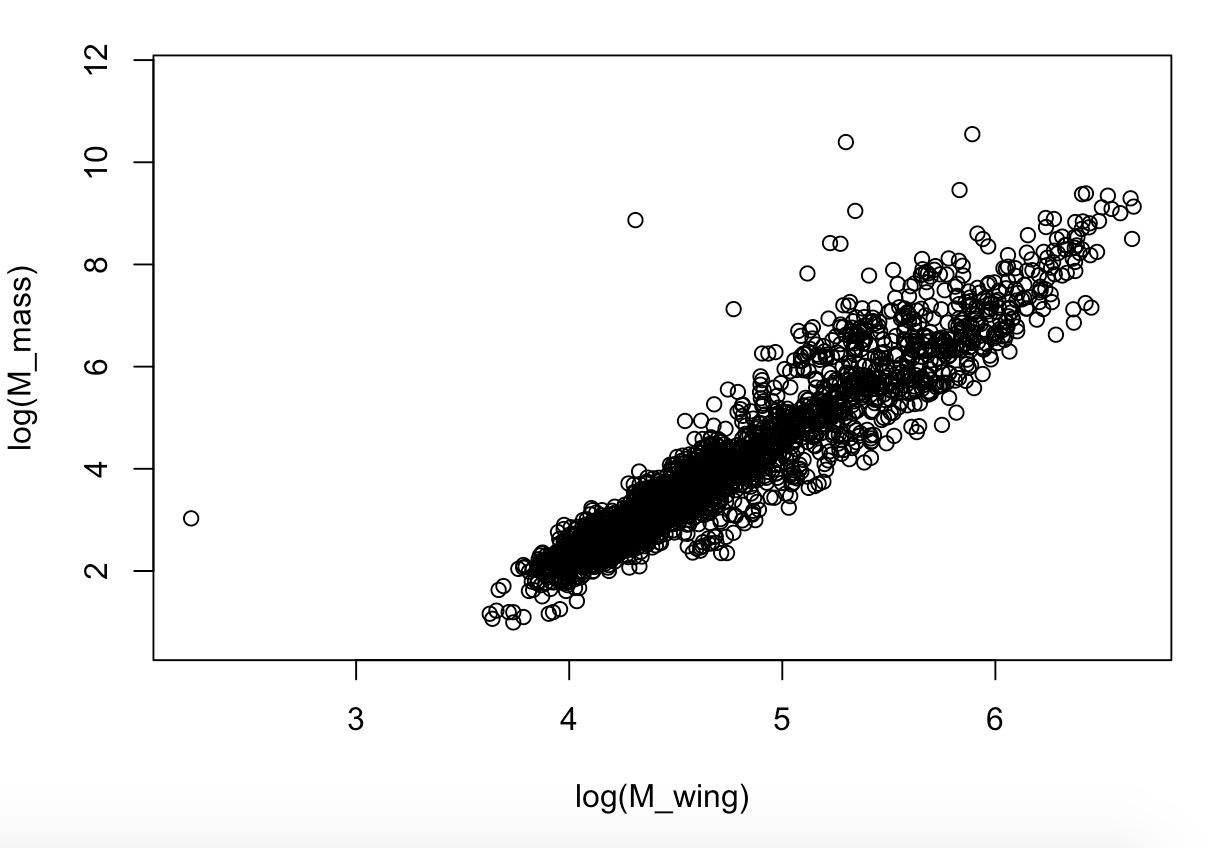
Figure14-plot(log(M\_mass)~log(M\_wing),data=jan07)

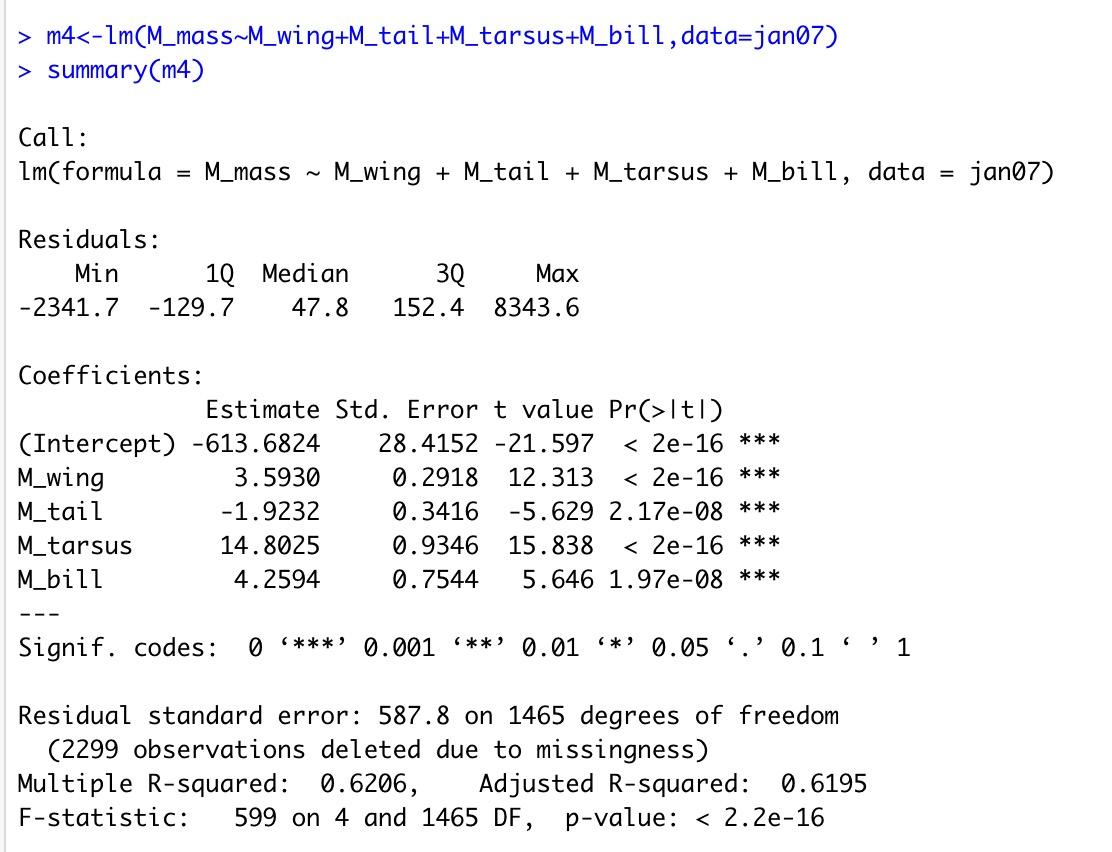
Figure 15-m4<-lm(M\_mass~M\_wing+M\_tail+M\_tarsus+M\_bill,data=jan07)

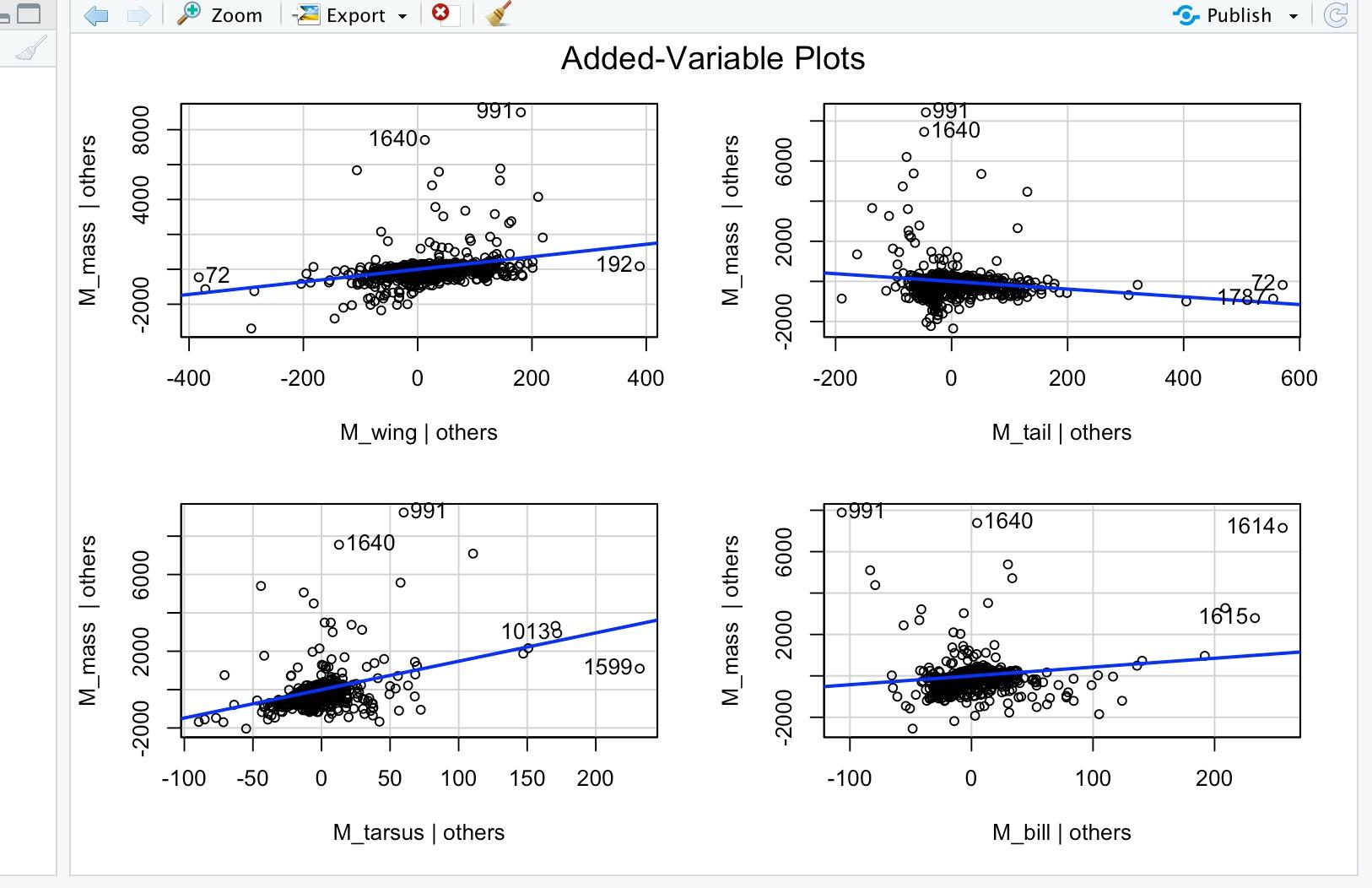
Figure 16-avPlots(m4)

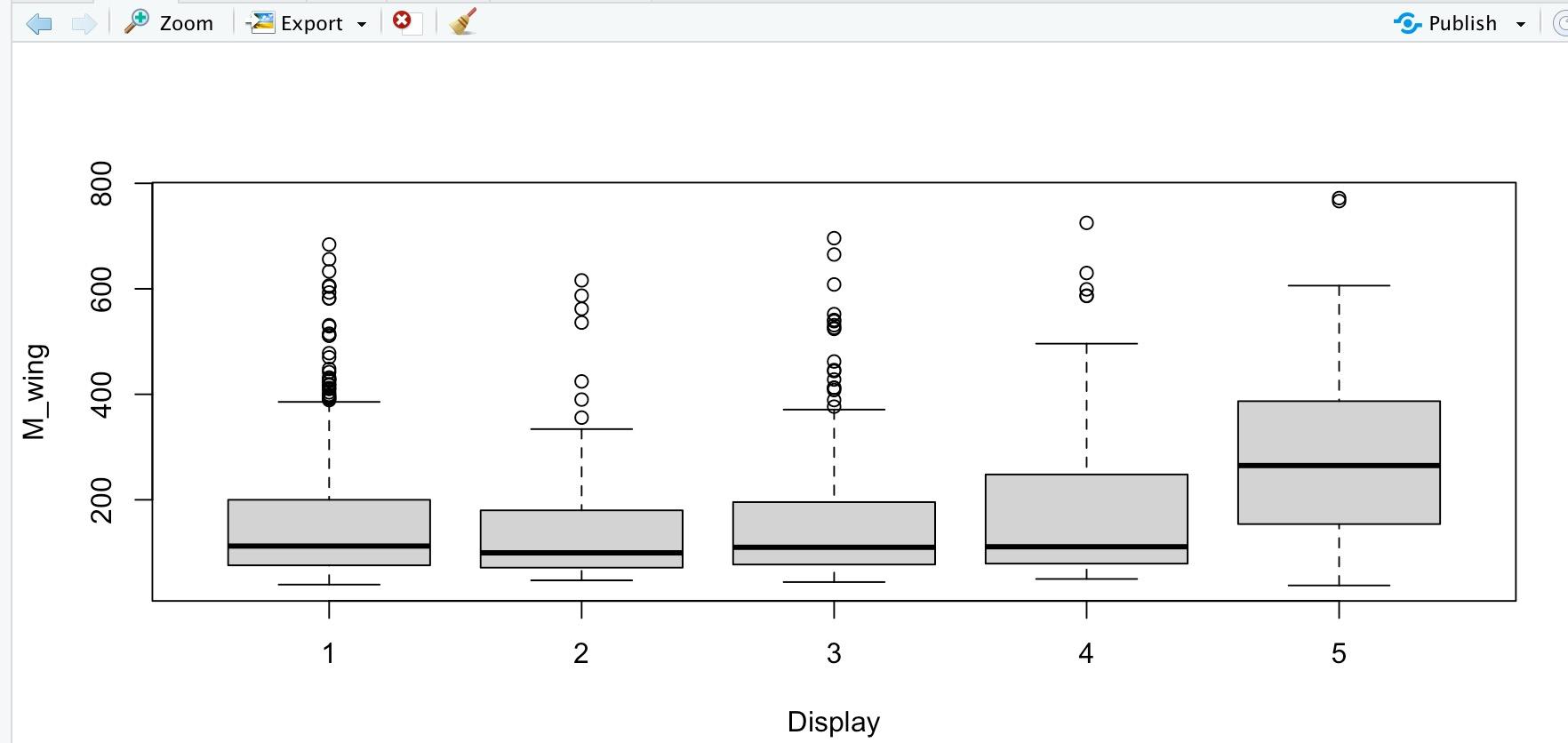
Figure 17-boxplot(M\_wing~Display,data=jan07)

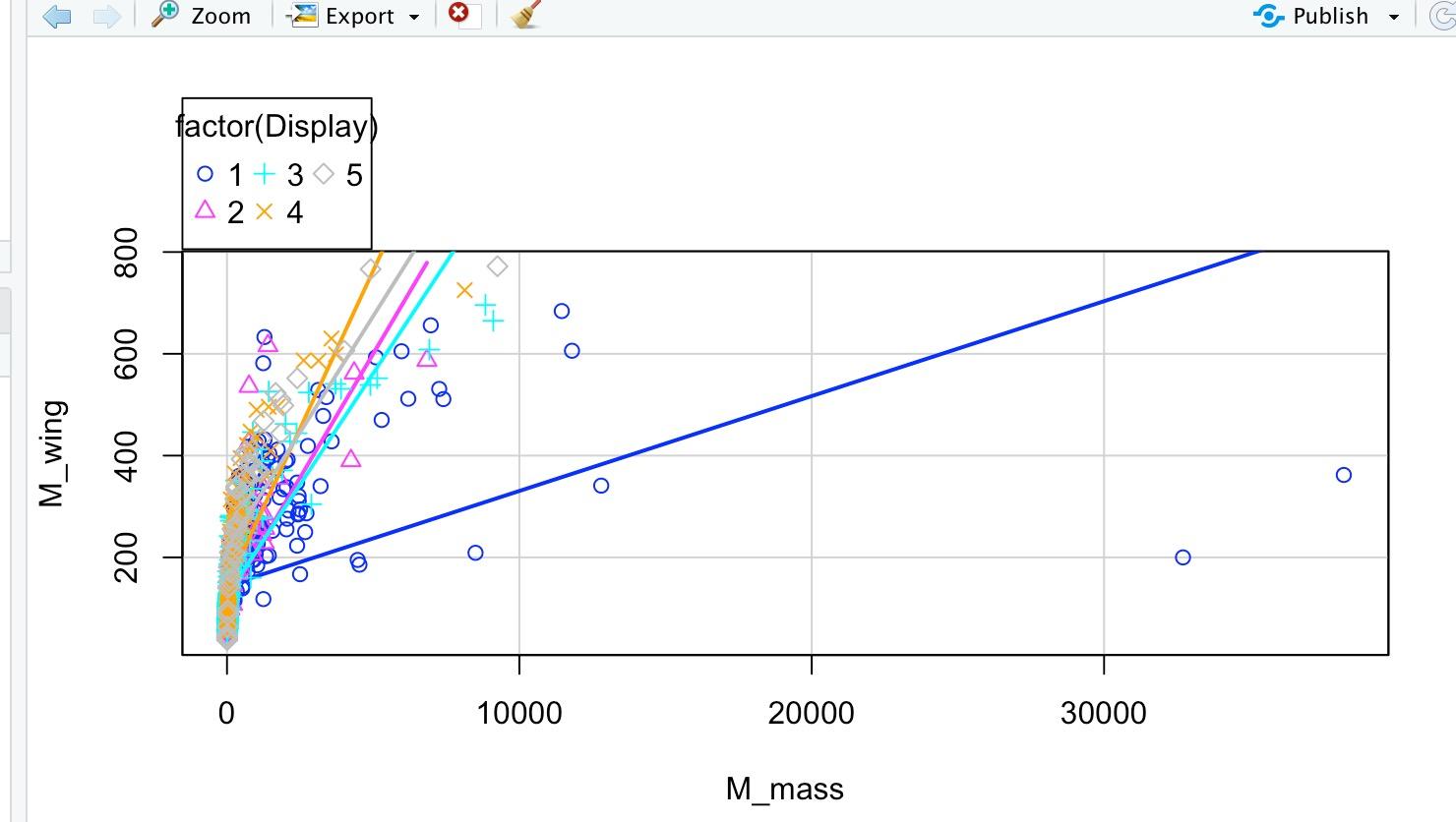
Figure 18-scatterplot(M\_wing~M\_mass|factor(Display),data=jan07,smooth=F)

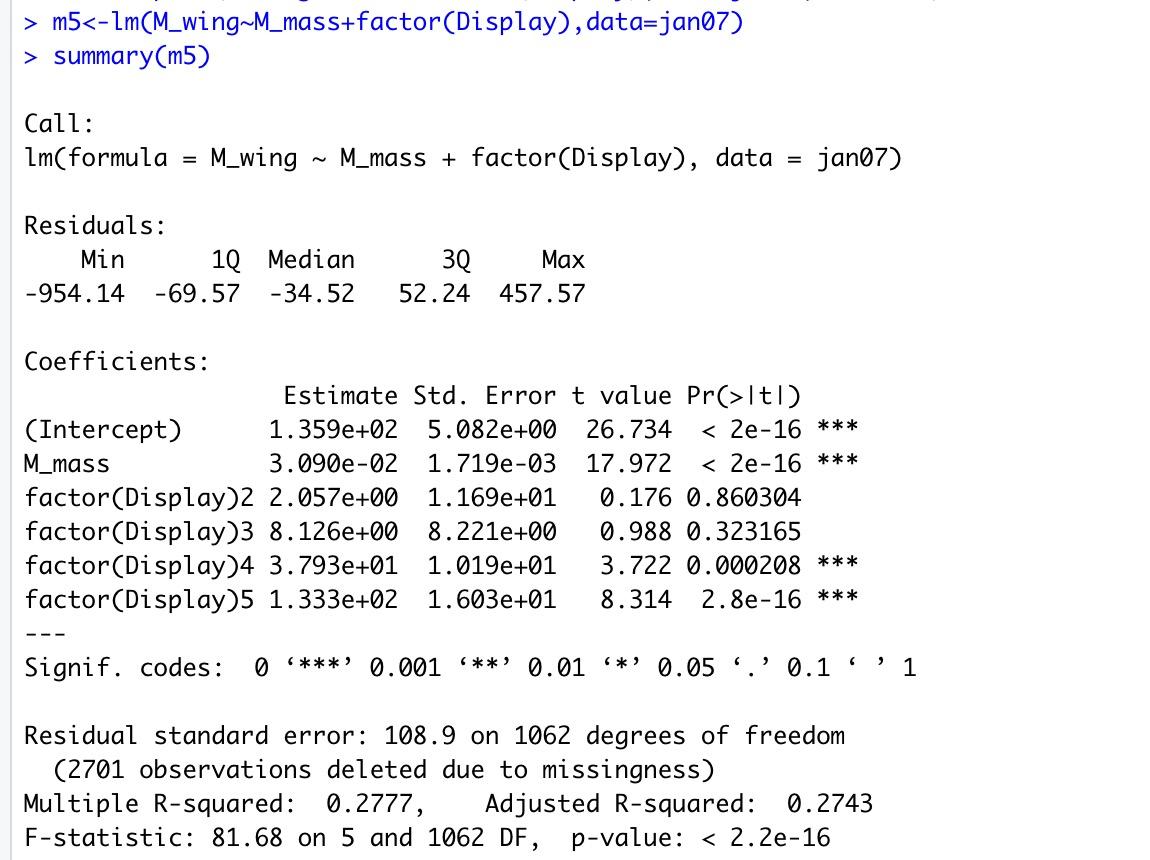
Figure 19-m5<-lm(M\_wing~M\_mass+factor(Display),data=jan07)

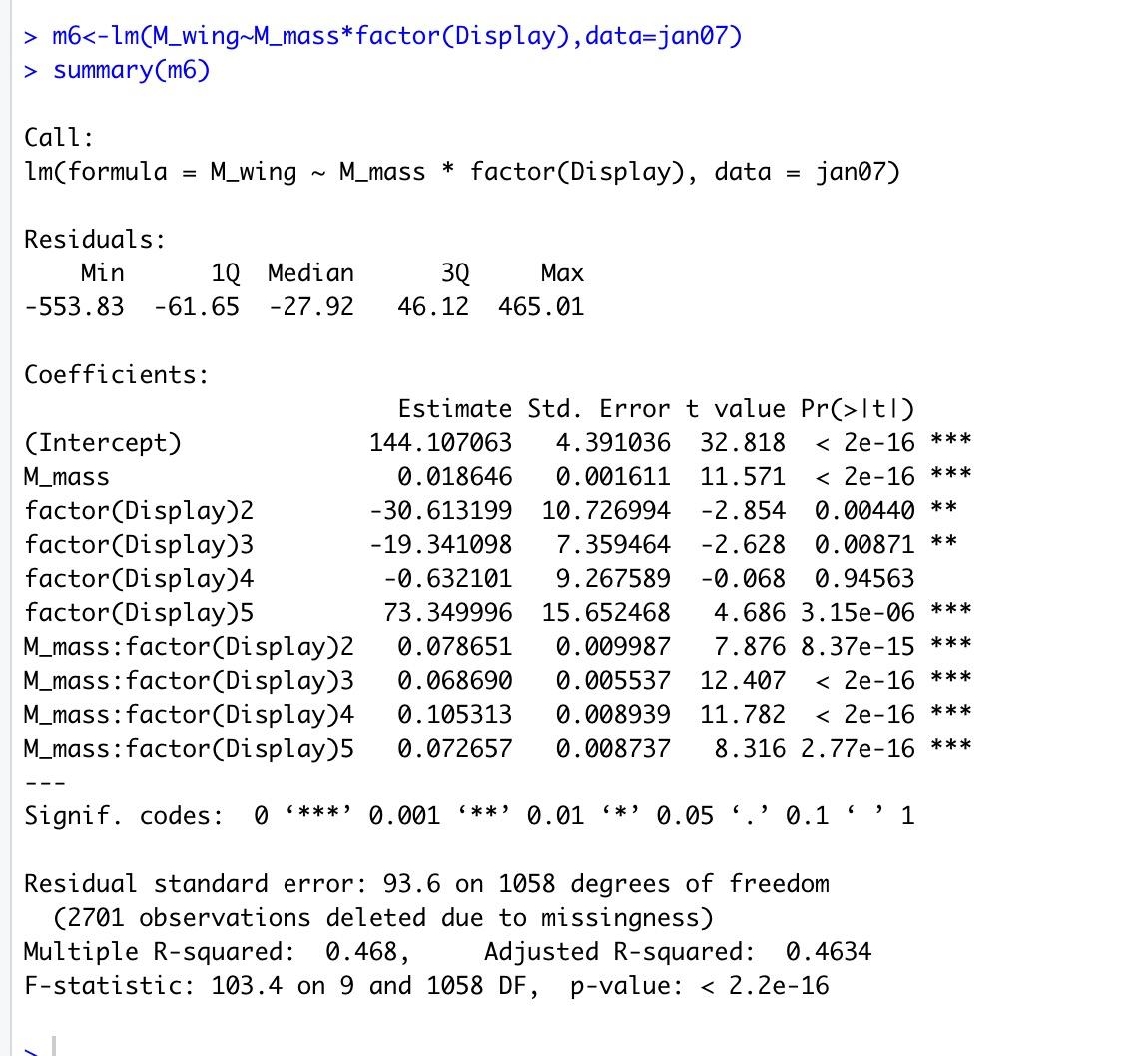
Figure 20-m6<-lm(M\_wing~M\_mass\*factor(Display),data=jan07)

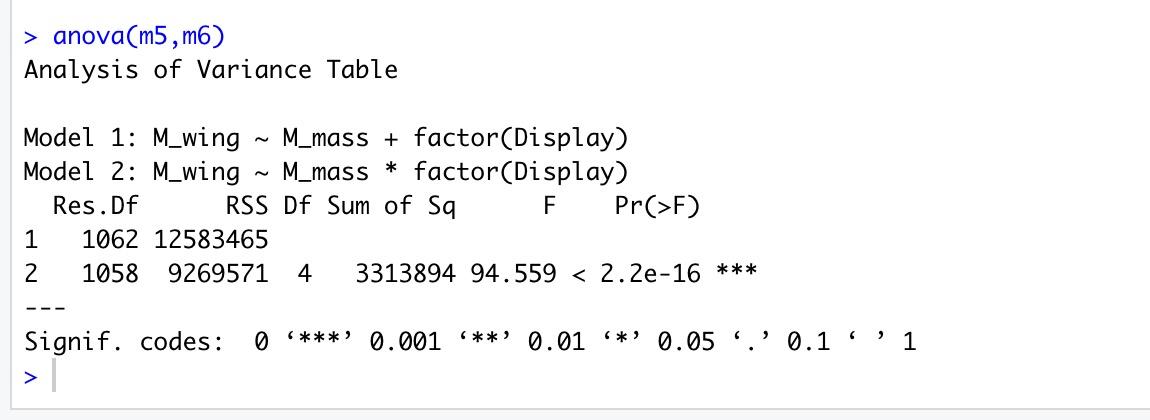
Figure 21-anova(m5,m6)

Figure 22-model(7\_1)<-lm(log(M\_tarsus)~Egg\_mass,data=jan07)

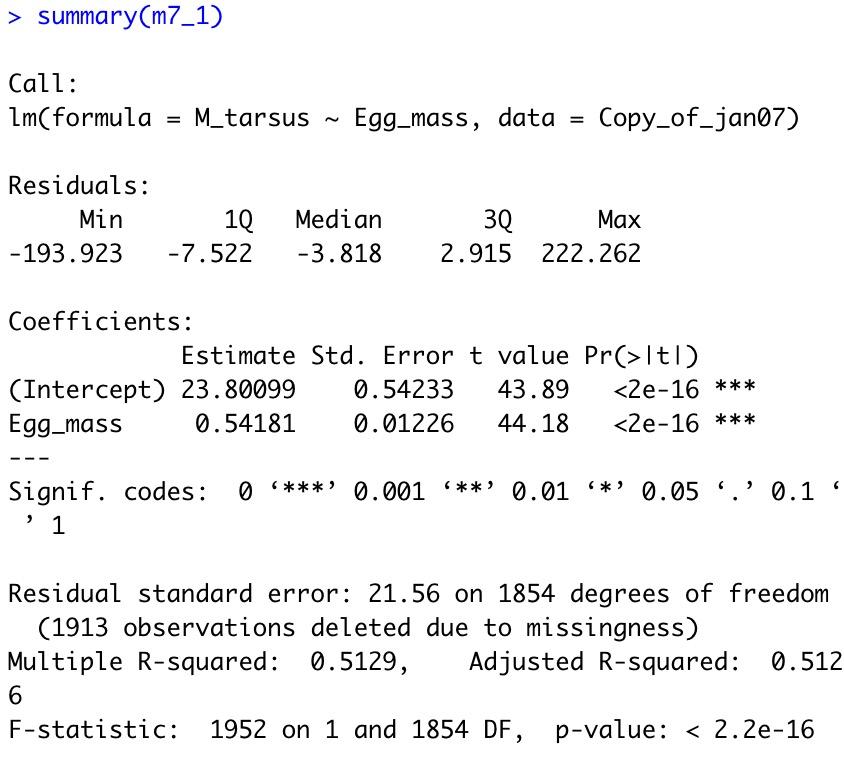


Figure 23-plot(M\_tarsus~Egg\_mass,data=jan07)

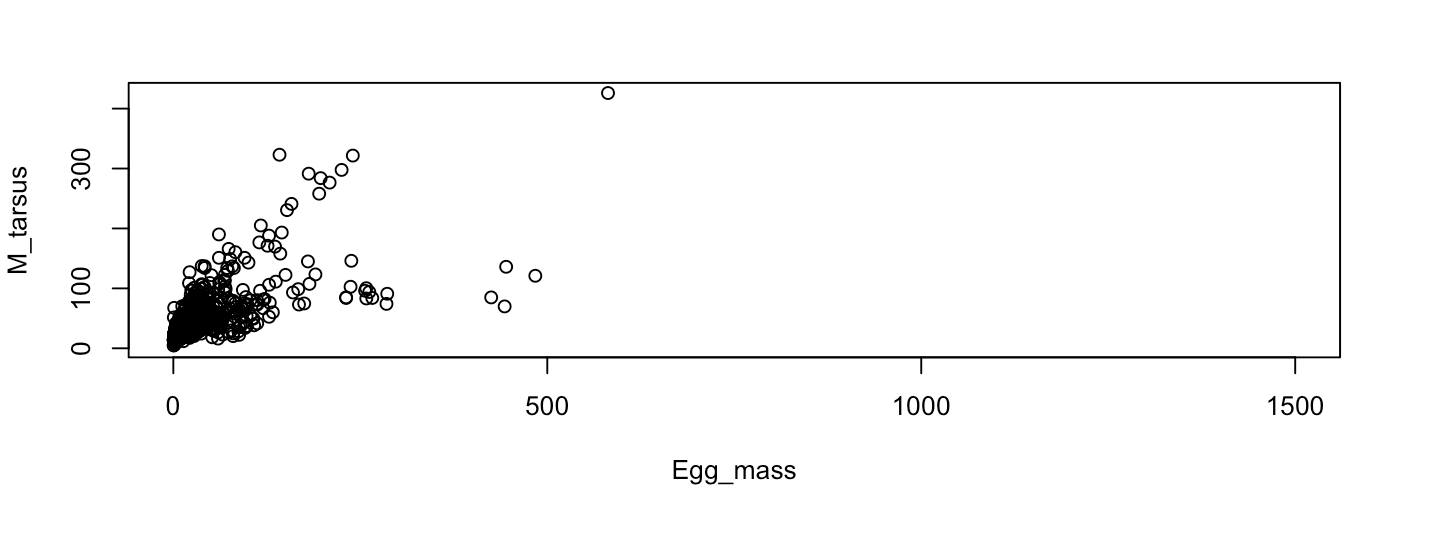


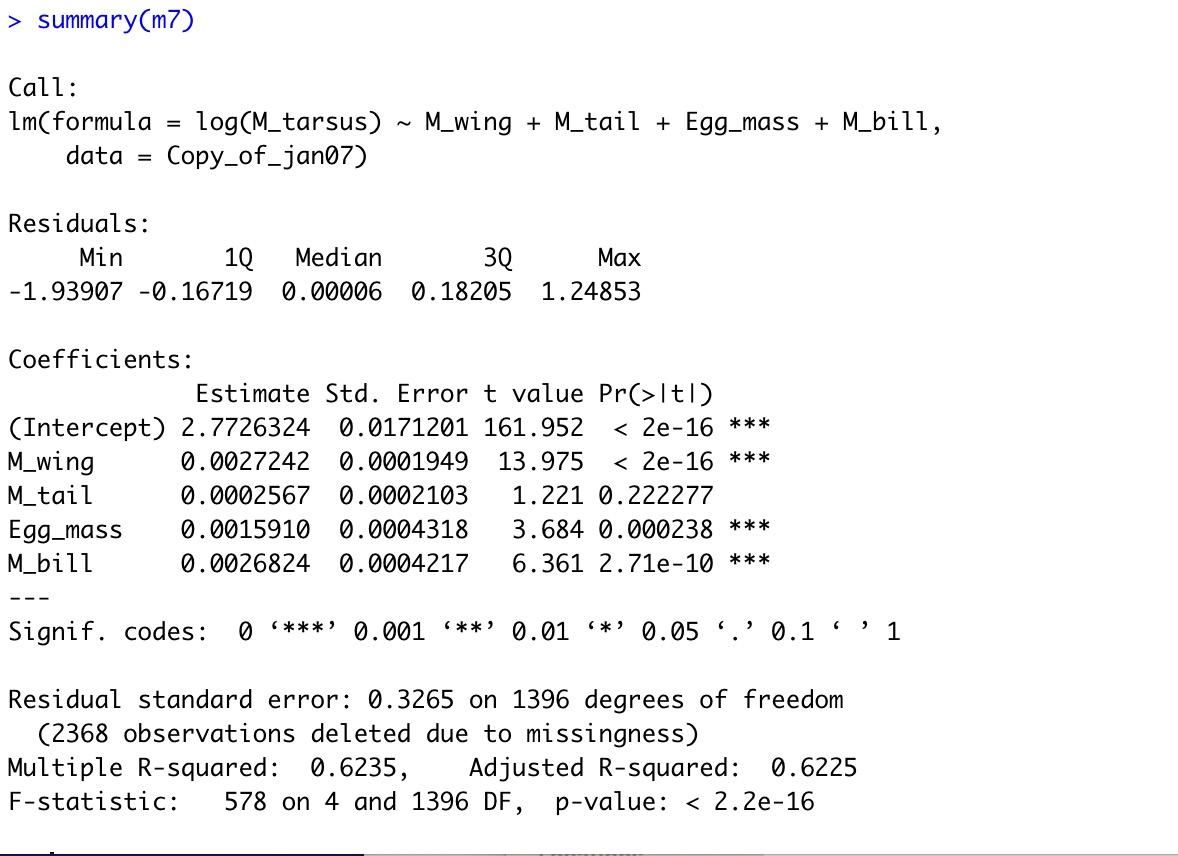
Figure 24-m7<-lm(log(M\_tarsus)~M\_wing+M\_tail+Egg\_mass,data=jan07)

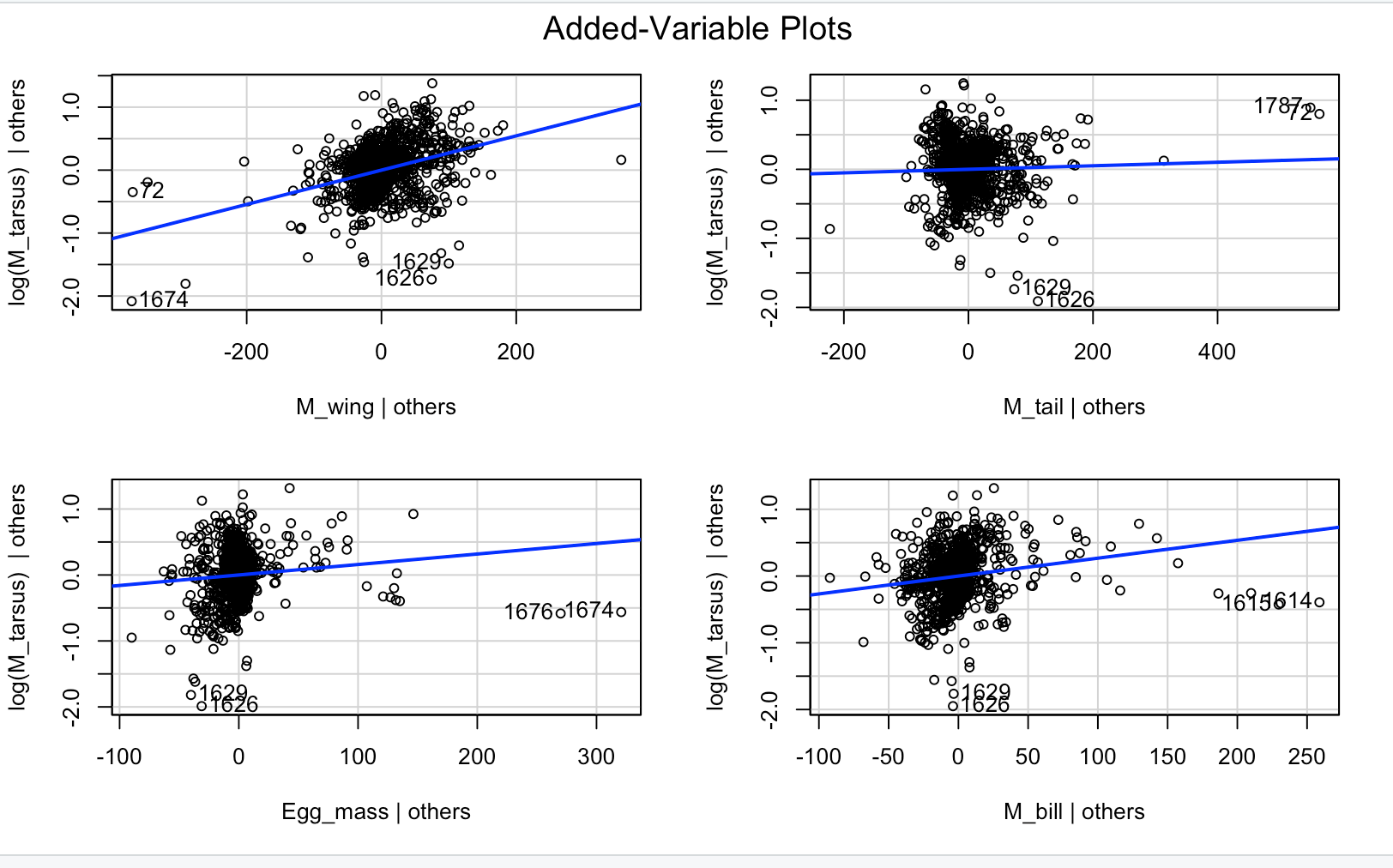
Figure 25-Added-Variable Plots(m7)

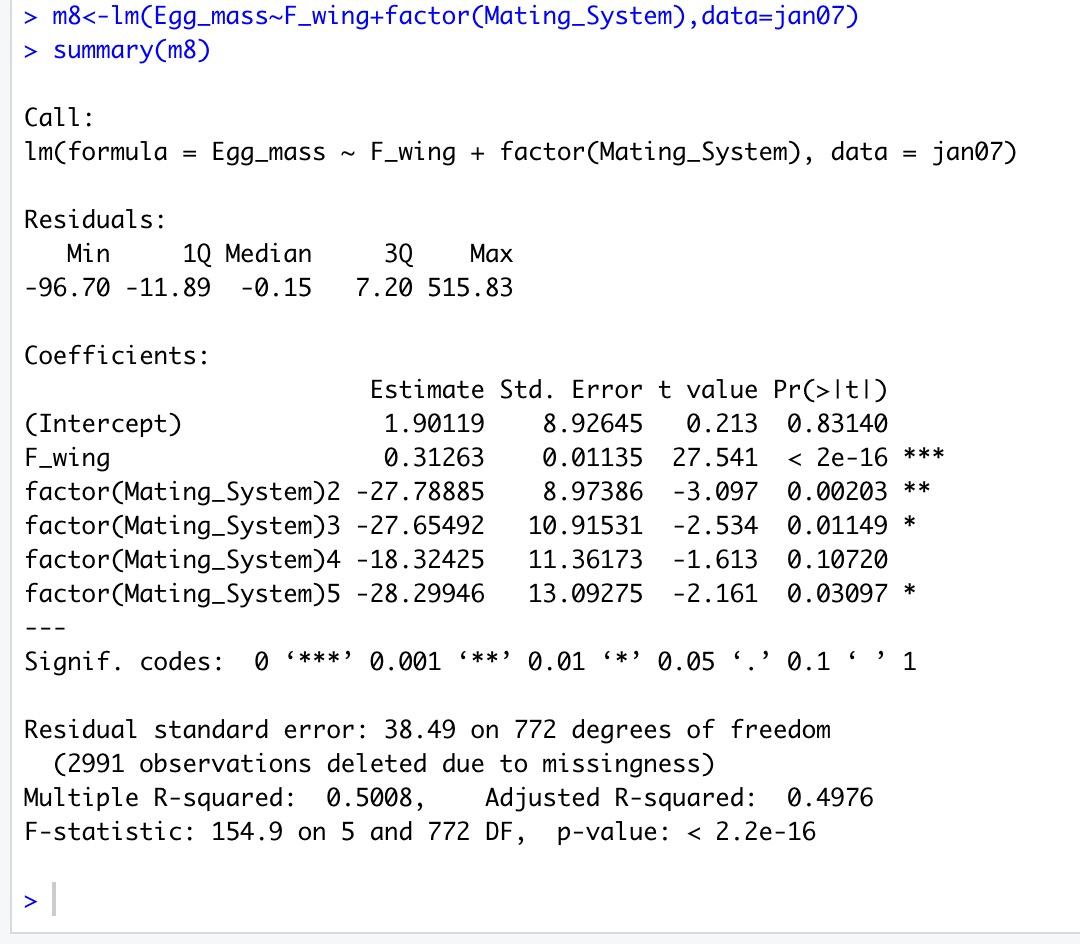
Figure 26-m8<-lm(Egg\_mass~F\_wing+factor(Mating\_System),data=jan07)

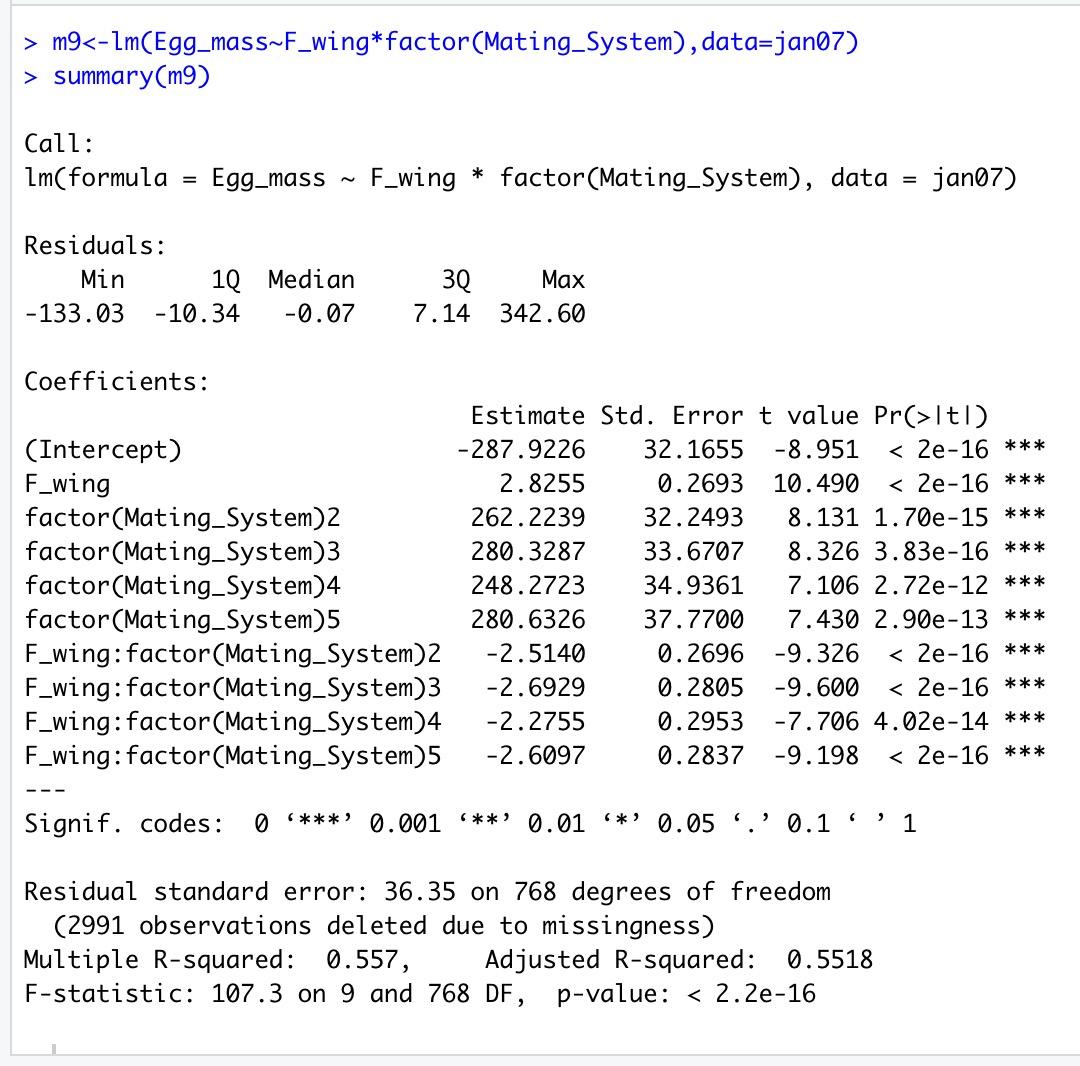
Figure27-m9<-lm(Egg\_mass~F\_wing\*factor(Mating\_System),data=jan07)

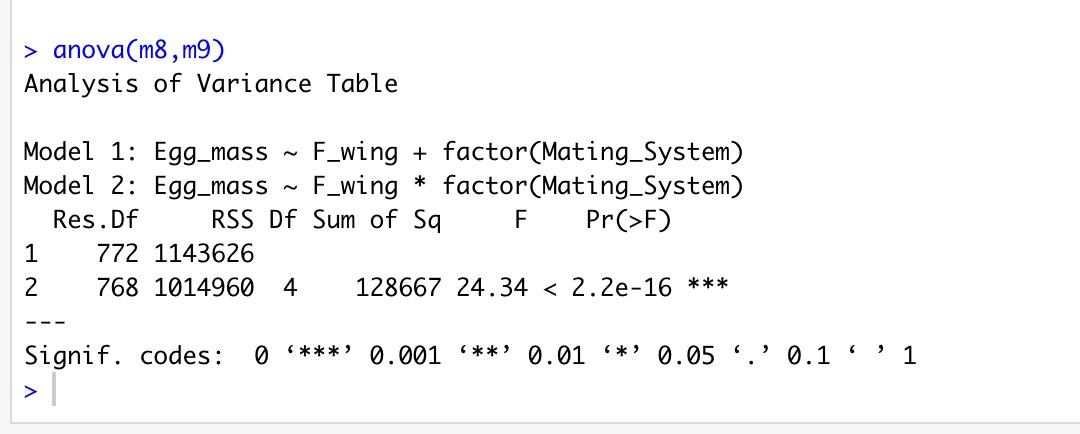
Figure28-anova(m8,m9)

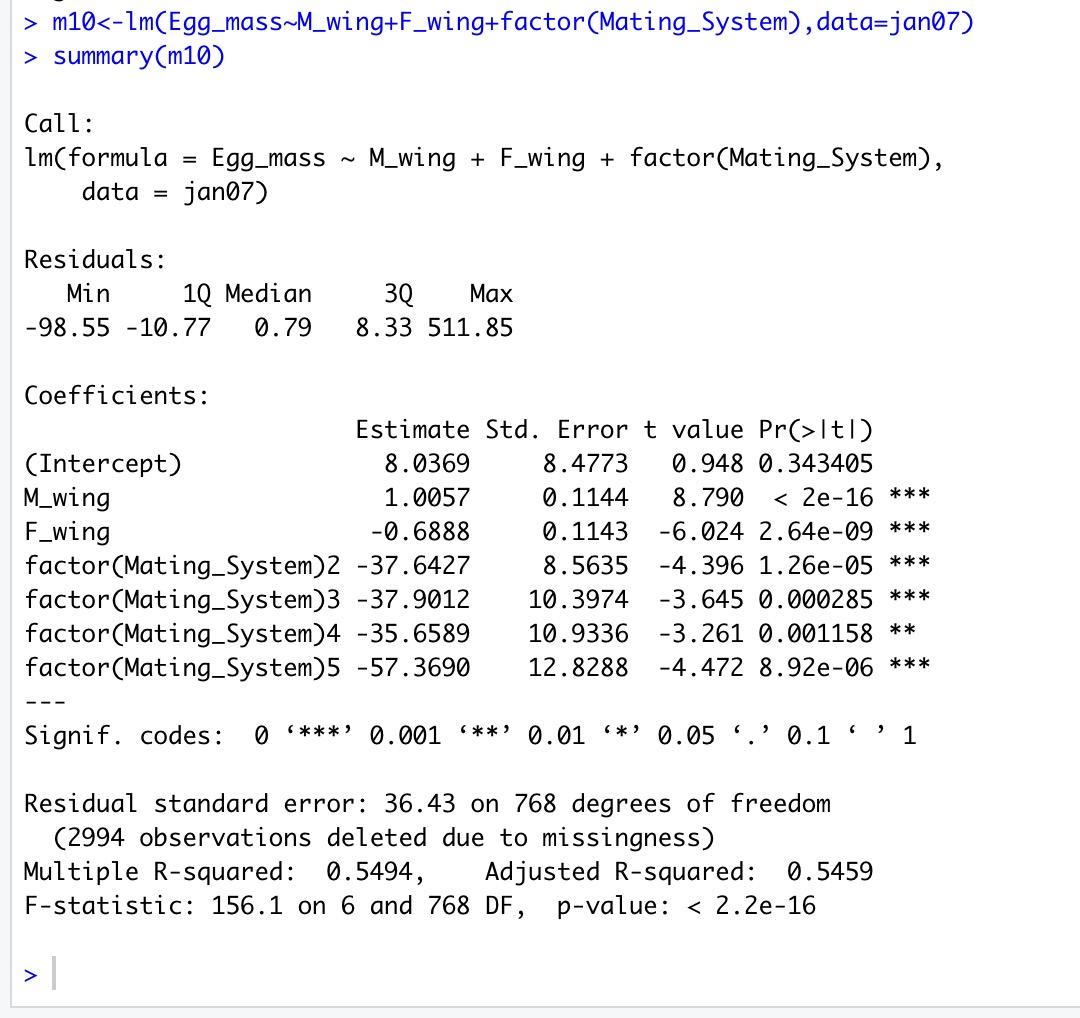
Figure29-m10<-lm(Egg\_mass~M\_wing+F\_wing+factor(Mating\_System),data=jan07)

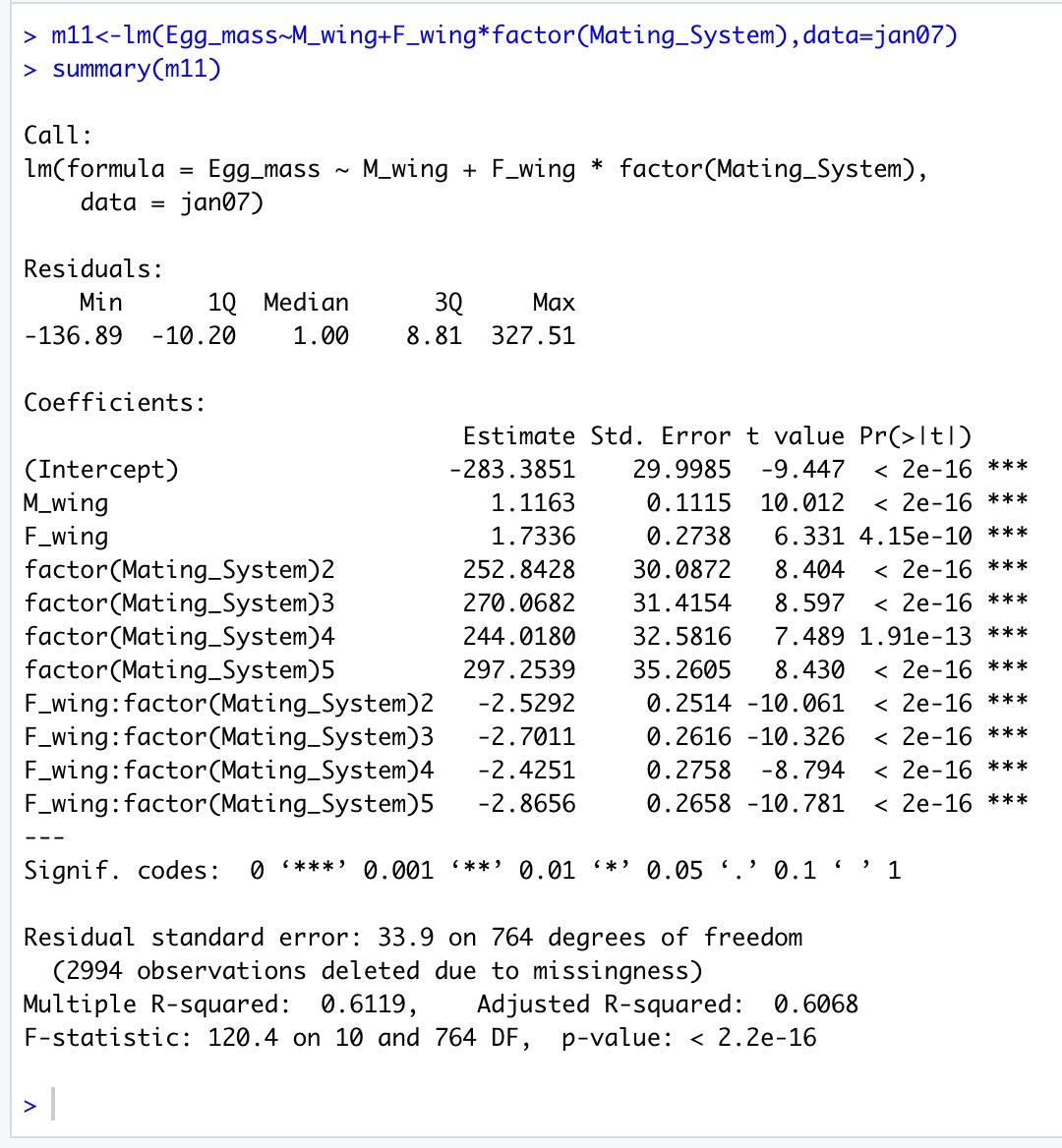
Figure30-m11<-lm(Egg\_mass~M\_wing+F\_wing\*factor(Mating\_System),data=jan07)

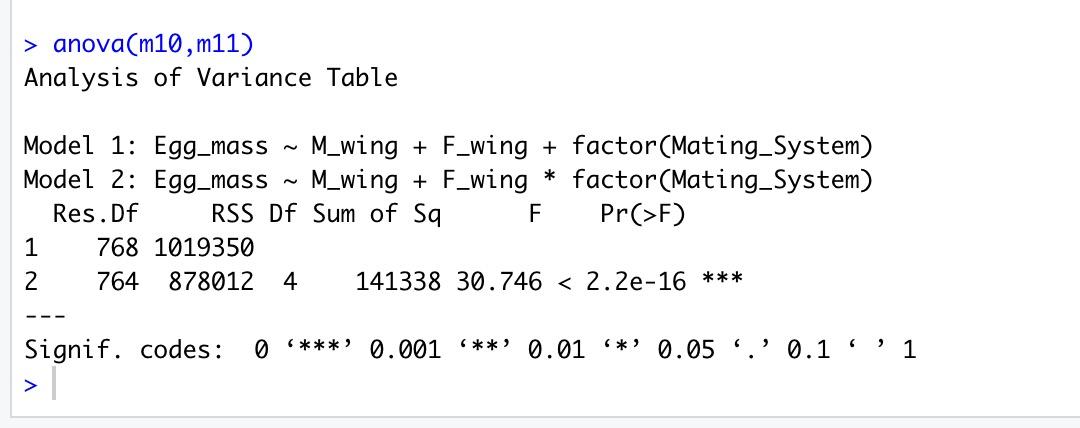
Figure31-anova(m10,m11)

Figure 32-Shapiro-Wilk test, ncvTest and muticollinearity test for m4 model

shapiro.test(m4$residuals)

ncvTest(m4)

vif(m4)

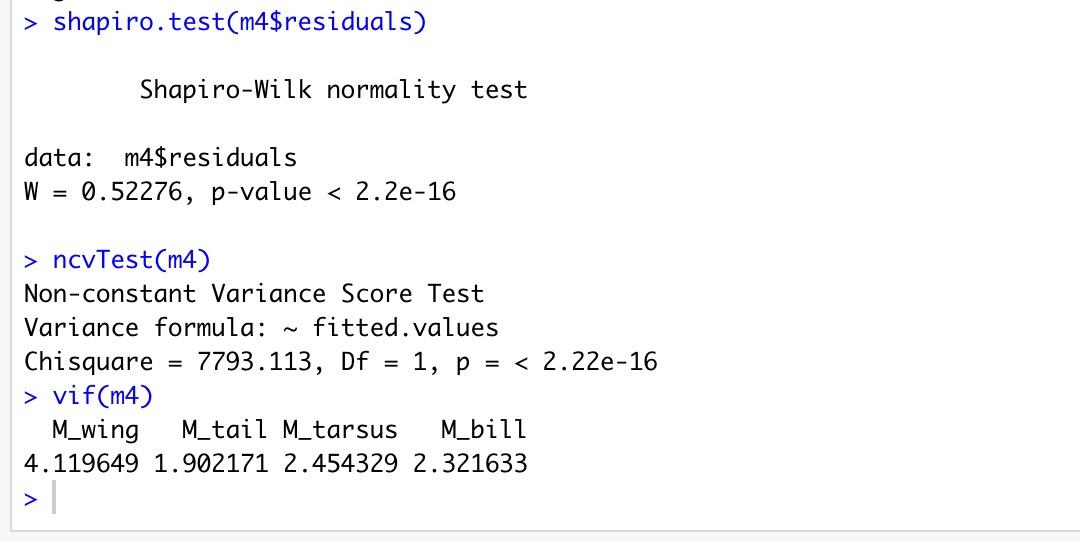


Figure 33-multicollinearity test for m7 model

vif(m7)

