Statistical Evaluation Project

02445 Project for Statistical Evaluation for Artificial Intelligence and Data

Albert Frisch Møller s214610 & Mark Andrawes s214654

Summary

Lameness is a problem affecting many horses, and it is a difficult condition to detect - even for experts. The aim of this project was to successfully classify lameness using a data-driven approach by applying machine learning models. In addition, the aim was also to use statistical models to evaluate model performances. Our results show that horse type only had a significant effect on the fourth principal component. Additionally, we concluded that there was an interaction effect between horse type and lameness.

Multinomial logistic regression models were used to successfully classify lameness. After performing non-parametric bootstrapping to compute confidence intervals, and pairwise McNemar tests, we were able to conclude that a model with symmetry scores as explanatory variables had a significantly different performance, when compared to a model with principal component scores. When comparing models on collapsed categories, we concluded that there was no significant performance difference between models trained using symmetry scores, models trained using principal components, and models trained using both. It was concluded that there was no significant gain in the performance when the models were trained on 5 categories and 3 categories. Lastly, different methods for determining if the models contained bias were described and discussed.

Contents

1 Introduction

Horse lameness refers to the impaired ability of a horse to trot and move around normally, and is a condition that is difficult to spot - even for experts. The purpose of this project is to take a data-driven approach to detect lameness in horses by creating machine learning models that can classify whether a horse is lame or not, where the focus will be on statistically evaluating the performances of these models.

The dataset that is used for this project contains information about different horses who have artificially induced lameness, including symmetry scores and principal components. This will be described further in Section 2. The machine learning models that will be used to classify lameness are multinomial logistic regression models, where three models will be trained and tested - each with a different set of explanatory variables and with lameness as the response variable. The first goal of this project is to investigate significance between variables. For the second and third goals, we will compare the performances of the three machine learning models when there are 5 classes and 3 classes respectively, as well as investigate if there is a significant gain by collapsing categories. The final goal is to discuss the models in relation to fairness and bias from a statistical perspective, and evaluate fairness towards horses of different heights.

2 Description of Data

The dataset used in this project contains a total of 85 observations and 11 attributes. The observations are composed of trials using eight different horses, and the attributes consist of the symmetry scores (S, A, and W), horsetype, lameness (lameLeg, lameSide, lameForeHind), and the first four principal components of the full dataset. Note that symmetry scores and principal components are continuous, whereas lameness and horsetype are categorical.

	B1	B2	B3	B4	B5	B6	B7	B9	**Total**
none	5	5	4	5	1	1	1	1	**23**
Left fore	2	2	2	2	2	2	2	2	**16**
Right fore	2	2	2	2	2	2	2	2	**16**
Left hind	2	2	2	2	2	2	2	2	**16**
Right hind	2	2	0	2	2	2	2	2	**14**
Total	**13**	**13**	**10**	**13**	**9**	**9**	**9**	**9**	**85**

Table 1: Table with an overview of the number of trials belonging to each combination of horse type and lameness type.

From the table above, we see that there is some class imbalance, for instance, there are 5 observations belonging to B1 that have no lameness, whereas there is only 1 observation with no lameness in B5, B6, B7, and B9 respectively. We also see that B3 contains no observation

with right hind lameness. This could perhaps be due to a lack of data. We would expect this to skew our classification results, specifically within the fold where B3 is our test set. We also see that the data set contains more information about the lameness of horse types B1, B2, and B4. Note that there is a class imbalance in terms of the lameness distribution within each horse type, but also within the distribution of lameness categories across the horse types. Next, some of the relationships between the variables will be investigated statistically.

3 Goal 1: Significance of Horse type on Variables

The first goal of this project is to investigate whether or not the horse type has a significant effect on the different variables - namely the symmetry scores, S, A, and W, as well as principal components 3 ($PC3$) and 4 ($PC4$), as only these are expected to be associated with lameness.

3.1 Method & Analysis

We will approach this problem by creating appropriate boxplots to explore the possibility of there being a significant effect of horse type on the symmetry scores and principal component scores. These boxplots are shown below:

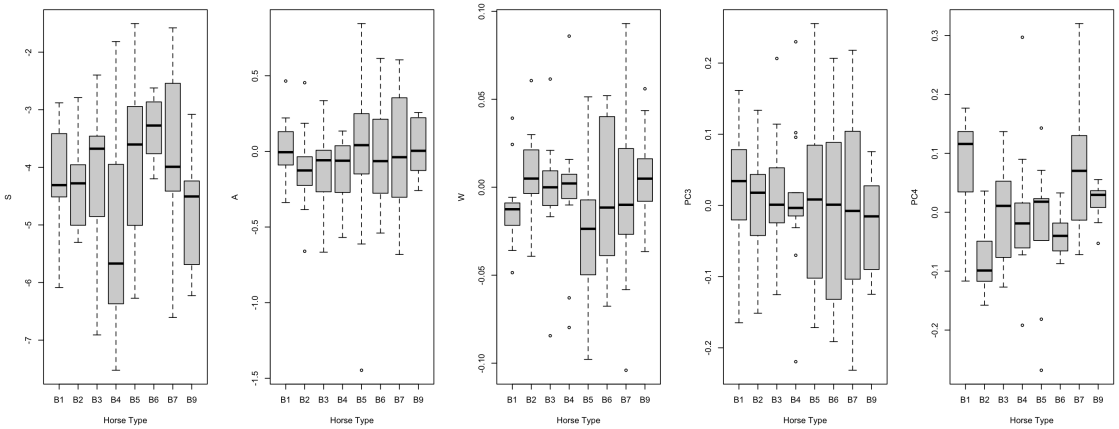


Figure 1: Boxplots of all horse types and their values for S, A, W, PC3, and PC4 respectively.

From the boxplots above, it is evident that there is a large variation in the values for S, W, and PC4 for the different horse types, whereas the values for A and PC3 do not vary as much. There is therefore a basis for exploring the effect by using linear regression models.

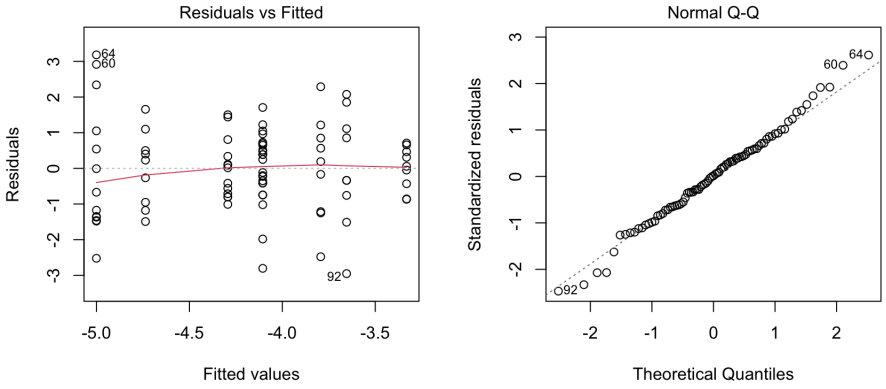


Figure 2: Fitted residual plots and QQ-norm residuals plot for the linear regression model with horse type as explanatory variable and S as response variable.

Seeing as the data points on the residual vs. fitted plot do not follow a specific pattern, and the points on the QQ norm plot approximately fit the line, the residuals of the linear model of S as a function of horse type seem to be normally distributed. Similar observations were made for the plots of the other symmetry scores as a function of horse type. These can be found in the appendix.

ANOVA also assumes that the variances of each group in the test are equal. From observing the boxplots in Figure 1 in combination with sample variance calculations, it is evident that this assumption is violated. Thus, we cannot use the ANOVA test and must turn to the Kruskal-Wallis test. As each observation is assumed to be independent and the data is continuous, the Kruskal-Wallis test can be used. We can therefore investigate the effect of horse type on the symmetry scores and principal component scores by conducting Kruskal-Wallis tests. Moreover, we wanted to investigate whether there is an interaction effect between horse type and lameness on the symmetry scores and principal components. For this, we must use an ANOVA test - despite the violation of the homoscedasticity assumption. The null-hypotheses state that there is no interaction effect, meaning that the effect of lameness on the score in question is the same for all horse types.

3.2 Results

We conducted Kruskal-Wallis tests for each linear model, where the null hypothesis stated that the explanatory variable has no significant effect on the response variable.

Response variable	Explanatory variable	p-value	Hypothesis interpretation
S	Horsetype	0.0766	Not rejected
A	Horsetype	0.721	Not rejected
W	Horsetype	0.334	Not rejected
PC3	Horsetype	0.983	Not rejected
PC4	Horsetype	0.000479	Rejected

Table 2: Table with the p-values of Kruskal-Wallis tests and hypothesis interpretations for each linear model.

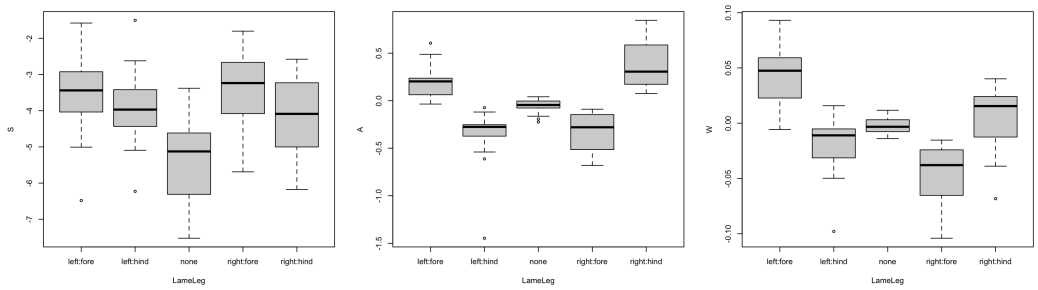
The table above shows that horsetype only has a significant effect on the fourth principal component. Additionally, from the ANOVA tests we obtained that there was an interaction effect between horse and lameness on the symmetry scores S and A, and the principal component score, $PC4$. As this is the case, when training a classifier, we need to use leave-one-horsetype-out cross-validation, where in each fold we leave one horse type out of the training set. This is because we want to remove the effect of horse type.

4 Goal 2: Classifying Lameness & Assessing Performance

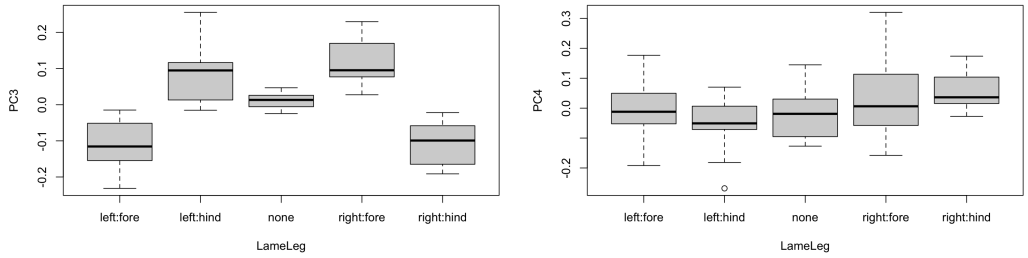
The second goal of this project is to investigate the performance when classifying lameness using a machine learning model. As we wish to classify lameness, we will use the variable $lameLeg$ as the response variable, as this variable contains the most information concerning the lameness of horses (side and fore/hind). We also wish to assess model performance when using only A/W symmetry scores, only $PC3/PC4$ scores, and when using both pairs of scores.

4.1 Method & Analysis

We approached this problem by firstly creating a boxplot of the symmetry scores for each category of lameness and a boxplot of the principal component scores for each category of lameness. This was done in order to investigate if there is a basis for including the scores in the proposed logistic regression classification model. The boxplots are given below:



(a) Boxplots of *lameLeg* for S, A, and W scores.



(b) boxplots of *lameLeg* for PC3 and PC4.

Figure 3: Boxplots for seeing if classification of *lameLeg* is significantly affected by the parameters in question.

After inspecting the above boxplots, it is clear to see that boxplot of the S symmetry score shows a lot of variation by lameness category. Similiarly, the boxplots of W and $PC3$ also display variation within each category. However, the boxplots of $PC4$ and A show little variation, and so thereby their effect on the predicted class will be smaller than the other variables in question. Hence it was found that there was a basis for building machine learning models with the symmetry and principal component scores as explanatory variables.

In order to achieve the goal of classifying lameness, we constructed three logistic regression models trained on different sets of variables, namely a model which used A/W scores, a model which used $PC3/PC4$ scores, and a model which used both pairs of scores. Afterwards, leave-one-horsetype-out cross-validation was performed, where in each fold a horse type was excluded in the training dataset. This was done as we ultimately want to be able to classify the lameness of an unseen horse, i.e. where the horse type is unknown. Seen from a statistical point of view, we concluded that horse type only had a significant effect on $PC4$, and that there was an interaction effect. Therefore we want to remove the effect of horse type when classifying lameness.

To explore the difference in the performances of the models, their classification accuracies were investigated and contingency tables were constructed for each pair of models. The confidence interval for the mean classification accuracy for each model was computed using non-parametric

bootstrapping, whereby 1000 samples were drawn with replacement from the set of respective classification errors found within each fold. Note that non-parametric bootstrapping was applied, due to small a sample size and not being able to assume normality.

Before being able to perform multiple McNemar tests comparing each model with the other models, we had to check if the appropriate assumptions were satisfied. The observations are assumed to be independent, and so the independent assumption is satisfied. Each model was trained and tested within each of the folds, and so the same train data sets and test data sets were used. The observations present within the data set came from the same population and so had the same distribution. The required assumptions for McNemar seem to be satisfied. Seeing as we are conducting more than one comparison between classifiers that use the same data, we have to account for the multiple comparisons problem. This entails that our probability of committing a type I error increases beyond the significance level used. We therefore must adjust the p-values of the McNemar tests using the Benjamin-Hochberg correction.

4.2 Results

Model A is the multinomial logistic regression model with A/W as explanatory variables, model B has PC3/PC4 as explanatory variables, and model C uses both pairs as explanatory variables. The mean classification accuracy of each model was computed on the entire dataset, meaning that predictions within each fold were used. Table 2 shows the mean classification accuracies for the three models, as well as their confidence intervals at 95% significance.

Model	Mean accuracy	Confidence interval
model A	0.704	[0.561, 0.833]
model B	0.541	[0.480, 0.610]
model C	0.674	[0.551, 0.787]

Table 3: Table with the mean classification accuracy, as well as the confidence interval found using non-parametric bootstrapping.

Table 3 shows that the confidence intervals for the mean accuracies of the models overlap with each other. For this reason, we cannot conclude whether or not there is a significant difference in their performances. We, therefore, conduct pairwise McNemar tests. Table 4 shows the results of the pairwise McNemar tests, including the interpretations of the null hypothesis which states that the two classification models in question have equal classification performances.

Pairwise McNemar Test	p-value	Adjusted p-value	Hypothesis interpretation
model A and model B	0.0123	0.0370	Rejected
model A and model C	0.547	0.547	Not rejected
model B and model C	0.0442	0.0663	Not rejected

Table 4: Table with the p-values, adjusted p-values, and hypothesis interpretations for each pairwise McNemar test of the models.

From the p-values in the table above, we see that there is a significant difference between the performances of model A and model B. On the other hand, the difference in the performances of model A and model C and that of model B and model C were not significant. Although we know that the performances of model A and model B are different, we do not have a basis for concluding which model performed better.

5 Goal 3: Collapsing Categories

We are given that it is difficult to separate "diagonal" lameness (right hind vs. left fore and left hind vs. right fore), which is likely causing the models to incorrectly classify the lameness when using five categories. Thus, for Goal 3 we will collapse these two pairs of categories into one, meaning that we have a total of three categories - namely 'right hind or left fore', 'left hind or right fore', and 'none'.

5.1 Method & Analysis

We will first investigate whether there is overlap between the diagonals for the input variables of the models. This is shown in the scatter plots below.

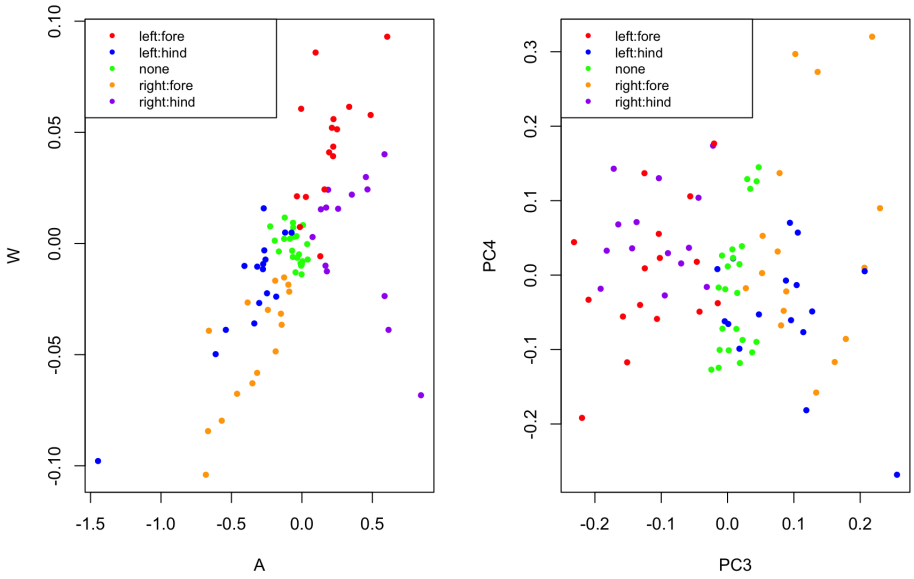


Figure 4: Scatter plots of W as a function of A, and $PC4$ as a function of $PC3$

As seen from the scatter plots for A/W and $PC3/PC4$, there is some overlap between the diagonal observations - specifically between the left fore and right hind data points and between the right fore and left hind data points. Therefore, the models find it difficult to distinguish between the diagonal lamenesses, and there is basis for reducing the classification categories in order to improve the model performances. This can be done by collapsing the categories from five to three categories. Having collapsed the categories, the multinomial logistic regression models were trained and tested using the same method as in Goal 2 using the new set of categories.

5.2 Results

The table below shows the mean classification accuracies for each model with collapsed categories, as well as their confidence intervals.

Model	Mean accuracy	Confidence interval
model A	0.905	[0.790, 0.981]
model B	0.859	[0.789, 0.934]
model C	0.850	[0.752, 0.938]

Table 5: Table with the mean classification accuracy, as well as the confidence interval found using non-parametric bootstrapping.

We see from the table above that there is, once again, overlap between all of the confidence intervals. We therefore must conduct pairwise McNemar tests:

Table 6 shows the p-values and hypothesis interpretations of the pairwise McNemar tests.

Pairwise McNemar Test	p-value	Adjusted p-value	Hypothesis interpretation
model A and model B	0.387	0.580	Not rejected
model A and model C	0.131	0.392	Not rejected
model B and model C	1.00	1.00	Not rejected

Table 6: Table with the p-values and adjusted p-values, as well as hypothesis interpretations for each pairwise McNemar test of the models with collapsed categories.

As shown by the table, we cannot reject any of the null-hypotheses. Therefore, we conclude that there is no significant difference in the performances of the models.

5.3 Investigation of significant gain by collapsing

We wanted to investigate if there was a significant gain by collapsing the categories from 5 to 3. This was done by collapsing the predicted labels for each model (trained on 5 categories) according to the diagonals, e.g. for each predicted label made by a model, if the label was equal to either "left:hind" or "right:fore" it was replaced by "left:diagonal". Afterwards, a contingency table was made for the comparison between the models when trained on 5 categories, and when trained on 3 categories. The table below shows the results of the pairwise McNemar tests between the models when trained on 3 categories and 5 categories. Please note that 'model A3' refers to model A when trained on 3 categories, and 'model A5' refers to model A when trained on 5 categories.

Pairwise McNemar Test	p-value	Adjusted p-value	Hypothesis interpretation
model A3 and model A5	1.00	1.00	Not rejected
model B3 and model B5	0.480	1.00	Not rejected
model C3 and model C5	1.00	1.00	Not rejected

Table 7: Table with the p-values and adjusted p-values, as well as hypothesis interpretations for each pairwise McNemar test of the models when trained on 3 and 5 categories.

Table 7 shows that there is no significant gain when collapsing the categories and training a model, as opposed to training a model on 5 categories and thereafter collapsing. This implies that the increase in classification accuracies when collapsing, as shown by Table 3 and Table 5, was majorly due to the reduction in the number of classes rather than the model itself being better. To conclude, there is no statistical difference in training the machine learning model on 5 categories in comparison to training the same model on 3 categories.

6 Goal 4: Discussion of Fairness & Bias in AI

Bias and fairness towards horses of different heights can be viewed from several perspectives, one of which is equalized odds. In relation to equalized odds, one could train a multinomial logistic regression model where lameness is the response variable, and the symmetry scores, principal components and heights as explanatory variables. Afterwards, the classifier could be tested on two testsets, one which contains extreme observations, i.e. very short or very tall horses, and one which contains horses of typical height. One could then investigate the true positive rate and false negative rate across both testssets. If there is a significant difference, this would imply that the machine learning model is unfair towards horses of abnormal heights.

In relation to Simpson's paradox, one could first investigate if the inclusion of an explanatory variable reverses the statistical effect of previous explanatory variables. This could be done by creating several linear regression models, accounting for the different combinations of explanatory variables, and then performing ANOVA tests. The result of the ANOVA tests would indicate whether Simpsons's paradox is in play. The consequence of Simpson's paradox would be that when evaluating the dataset as a whole there might not appear to be any significant bias, whereas when you explore subgroups of the data, then there is bias and unfairness towards the subgroups.

In relation to reference bias, if it was determined that the data set was biased towards certain horse height categories, then the machine learning model would also be biased towards the categories. Hence, bias carries over from the dataset to the model.

7 Discussion & Conclusion

The aim of this project was to build machine learning models which could detect lameness in horses and classify whether a horse was lame or not, and to evaluate their performances. All in all, this aim was achieved. Based on our findings, we propose that a multinomial logistic regression classifier should be used to classify lameness. Our findings show that, by collapsing the categories from 5 to 3, we did not obtain a significant performance increase. In future studies, one could perhaps introduce a regularization parameter, and perform 2-layer cross-validation, where an optimal regularization parameter is chosen, which optimizes the bias-variance trade off.

Although we reached the aim of this project, there are some limitations to be aware of. Firstly, the dataset has a small sample size, meaning that there is limited data for each combination of horse type and lameness type. This in turn limited the accuracy of the models. Another limitation of this project is that the assumptions of the tests that were utilized were not completely satisfied, as we cannot test them with complete certainty.

It was interesting to see that, once the categories were collapsed, that all three models performed equally according to the McNemar tests. This implies that the input variables of the models had little effect on their performances in this case.

To conclude, it was determined that horse type only had a significant effect on the fourth principal component. It was also found that there was an interaction effect between horse type and lameness type on some of the explanatory variable. We were also able to successfully classify lameness and conclude that collapsing the categories did not provide a significant gain in performance, and that the jump in classification accuracy was majorly due to the reduction of classes, rather than the model having improved. Seeing as the classification accuracy reached up to 90%, we believe that there is basis for developing this model into an AI tool for classifying lameness in horses - especially if it is possible to gather more data, which will ultimately boost the models classification performance. However, it is important to note that the models performance was only good enough to be developed into an AI tool when the labels were collapsed. Therefore, the model that we proposed in this project will be of great use for classifying diagonal lameness. On the other hand, it is evident that more data would be necessary for classifying leg lameness.

8 Appendix

8.1 Additional plots

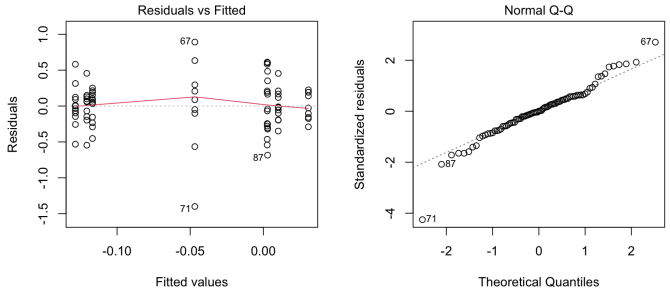


Figure 5: Fitted residual plots and QQ-norm residuals plot for the linear regression model with horse type as explanatory variable and A as response variable

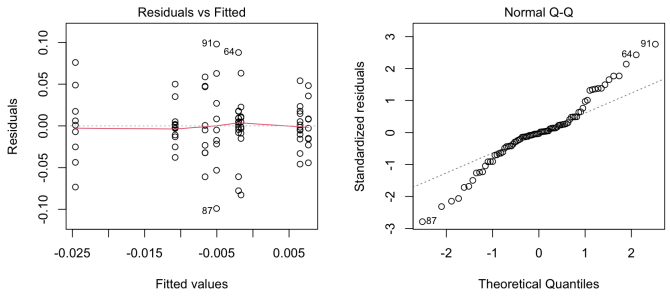


Figure 6: Fitted residual plots and QQ-norm residuals plot for the linear regression model with horse type as explanatory variable and W as response variable

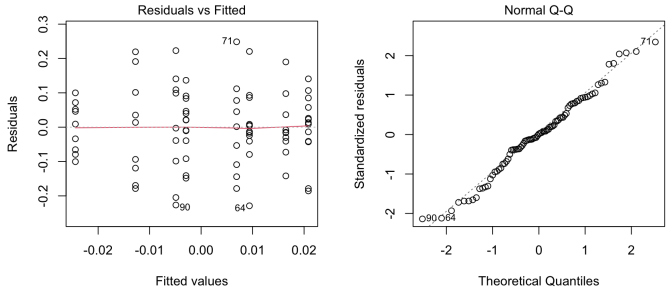


Figure 7: Fitted residual plots and QQ-norm residuals plot for the linear regression model with horse type as explanatory variable and PC3 as response variable

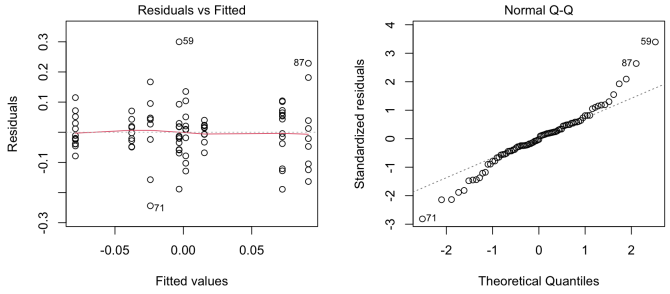


Figure 8: Fitted residual plots and QQ-norm residuals plot for the linear regression model with horse type as explanatory variable and PC4 as response variable

8.2 R code