Market Forecasting Assignment: Multiple Linear Regression

Teacher: Dr. Stefan Groesser, Dr. Christoph Imboden

Autors: Benedech Rodolfo, Buehler Pascal, Geeler Ken, 04.12.2021

Short Theory Recap Multiple Linear Regression (MLR)

The model equation for MLR has the form $y = X\Theta + \epsilon$,with y being the target vector, X the explanatory Matrix,Θ the parameter vector and the residuals with the assumption that they are normally distributed $\epsilon \sim N(0, \sigma^2)$. The estimation of Θ is done with the Maxium likelihood $L(\theta) = P(X|\theta) = \frac{1}{\sqrt{(2\pi)}\sigma} * e^{-\frac{1}{2\sigma^2} * ||y - \theta X||^2}$ by solving the equation $\Theta = (X^T X)^{-1} X^T y$ *(1)*.

Dataset

The data from 7 common fish species for fishmarket is provided from Kaggle.

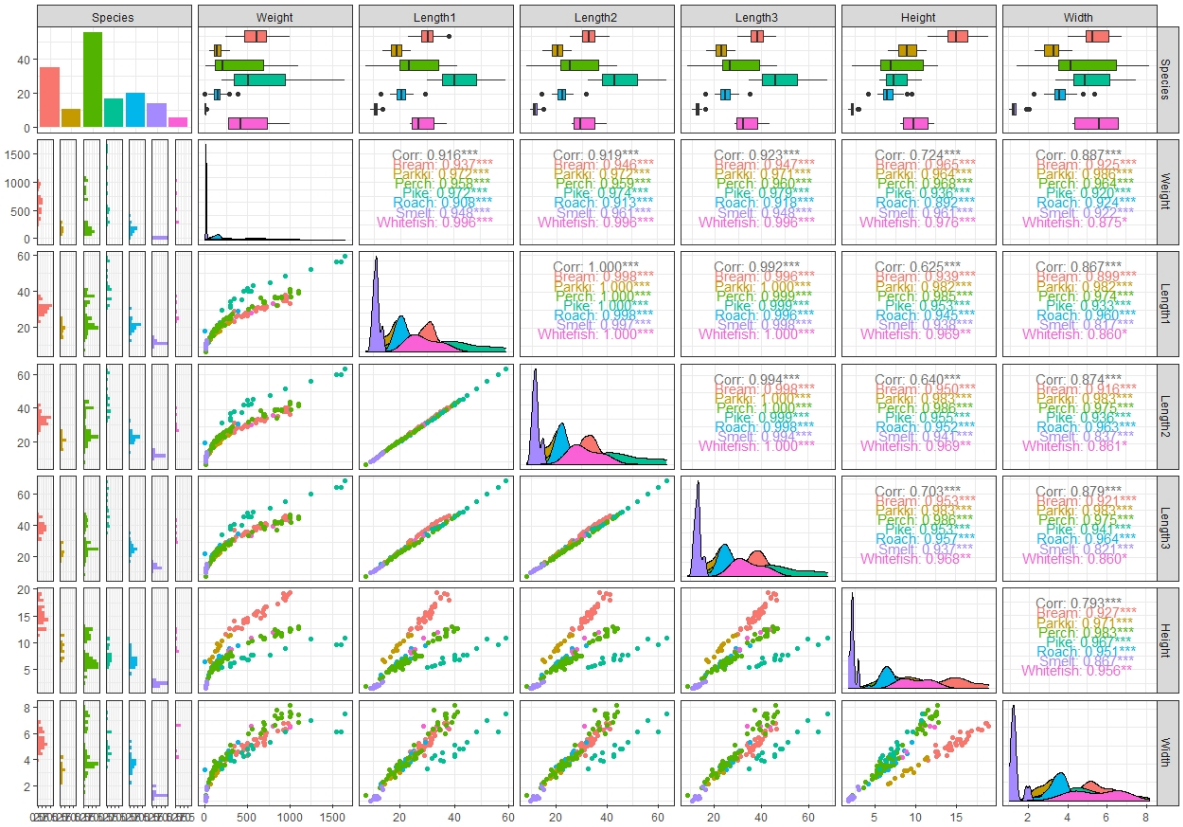


Figure 1: pairwise plot of the fish dataset

Data exploration

Information of the Dataset is gained with the following commands.

```
fish_dat=readr::read_csv("Fish.csv")                          #  reading as tibble
GGally::ggpairs(fish_dat,aes(color = Species)) + theme_bw() #  pairwise plot
summary(fish_dat);str(fish_dat);View(fish_dat)               #  statistics and structure
```

The pairwise plot Figure 1 provides a visual overview of the dataset. It consists of 159 observations measured for 6 numerical variables which measure properties of the fish and one categorical variable to describe the Species. The colouring in the pairwise plot allows a distinction of the Species, one can immediately observe that the species Perch has the most observations.

From the pairwise scatter plots a nonlinear relation between weight and the other numerical variables is observed, therefore the correlation *cannot* be correctly interpreted with these variable pairs. Further the other numerical variables are linearly correlated (the correlation statistic can be interpreted), interestingly the 3 length measurements have a very strong correlation to each other. The length of the fish is measured in different ways, which is why there are three different $Length1$ (vertical length), $Length2$ (diagonal length) and $Length3$ (cross length).

The density plots help see how the variables themselves are distributed. For example, the width of *Smelt* tends to have a bimodal distribution, although this must be taken with a grain of salt due to the sparse data, see on the inverted histograms on the left. Also, the pairwise boxplots give a good overview for comparison between the different variables. For example, it can be seen that the variance of the species varies amongst the levels.

The distribution of the target variables is actually very important since MLR assumes multivariate normality of the data, for other distributions we might choose another model such as a Generalized-Linear-Model to improve the model validity, but for this analysis we stick to MLR.

Building a Model

To evaluate the accuracy of the model created in this chapter, the data is split into a train and test set, using Simple Random Sampling without Replacement ($SRSWR$). 80% of the data is used for training and the remaining 20% is for testing prediction accuracy. The corresponding R code is given below.

```
set.seed(42);samp= sample(1:56,size = 11,replace = F)       # Seed & sample size 80%
Perch_train=Perch_dat[-samp,];Perch_test=Perch_dat[samp,] # train / test sample
```

The dataset provides many different options to do linear regression, nevertheless for MLR there is a problem with multicollinearity since a lot of the variables are strongly correlated, meaning that if multiple independent variables would be in the model it wouldn't be clear which one explains the effect the best for the target variable.

Therefore, the first thing we do is fitting a simple linear regression model. We use the weight of a fish as a response variable. The explanatory variable width is used, since this is intuitively a good indicator for the weight of a fish from "fishing experience". In a next step, we will also check whether a MLR model can represent the data even better.

Finally, the models are compared quantitatively with the residual sum of squares (RSS), R-squared and the Akaike information criterion (AIC). Subsequently, the models will be used with the test set for predictions, whose performance will be measured with the Mean Absolute Deviation (MAD) and the MSE. The reason why we not only use Mean Squared Error (MSE) is because of the small test set, one far point could massively influence the statistic due to the square.

Simple Linear Regression

For the sake of simplicity we just focus on the species Perch (in Swiss-German it's called Egli) since we have the most data there. Since the weight hasn't got a linear relation to the other variables, we'll have to transform it properly.

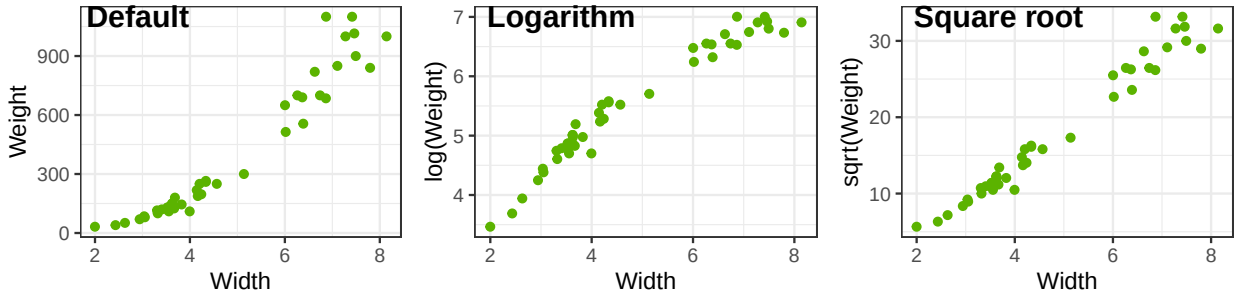


Figure 2: Transformations of Weight for Perch

If we transform the weight with the logarithm it looks a bit better but still isn't linear see Figure 2. The square root transformation looks better, the data seems much more linearly dependent. One can observe that the variance towards the right side of the model increases even after the transformations.

With the transformation done the first model is created:

```
Perch_fit=lm(sqrt(Weight)~Width,data = Perch_train) #  Simple linear regression model
```

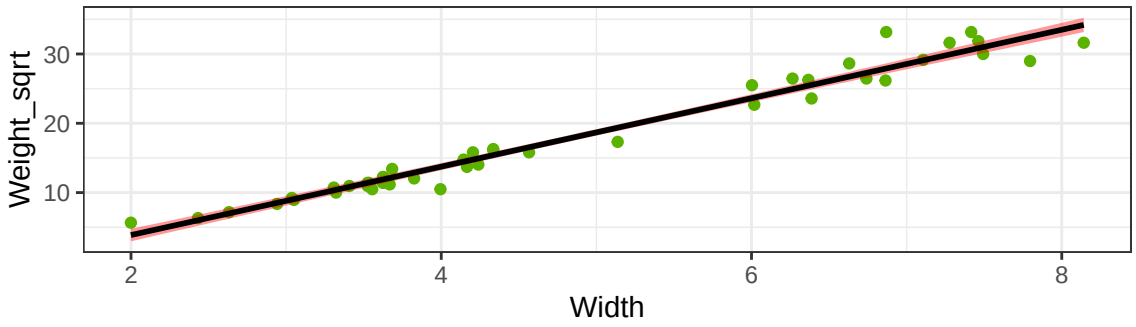


Figure 3: Simple linear regression model for Species Perch weigth vs width

From the summary output we get the following model $Weight_i = (-6.000 + 4.9368 * Width_i)^2$. The explained Variance measurement R^2 is with 0.9683 very high. the AIC for this model is 171.6. The p-values for the both parameters hypothesis tests are also significant. By inspecting the residuals in Figure 4 we can check the model assumptions [2]. The residuals seem to follow a normal distribution, have constant mean and there are no points which are crucial to shift the model. Nevertheless, the variance is not constant and increases towards the right side.

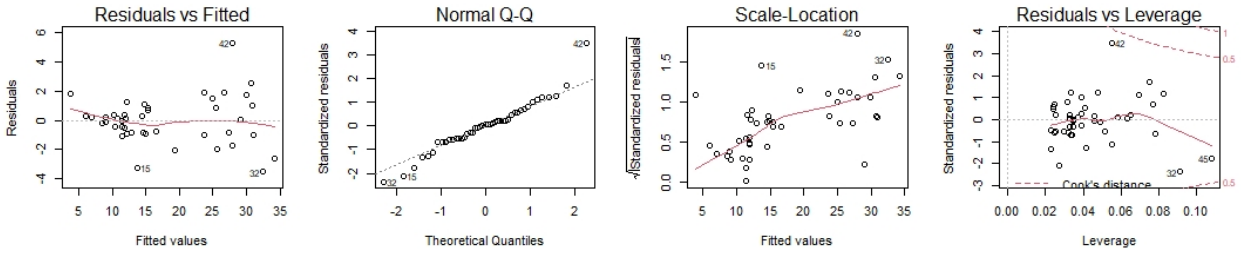


Figure 4: Residualanalysis for the SLR

Multiple Linear Regression and the Problem of Multicollinearity

Lets make now a full model and test if it is more beneficial.

```
Perch_fit_full=lm(sqrt(Weight)~.,data = Perch_dat[,-1])                  #  MLR
car::vif(Perch_fit_full);drop1(Perch_fit_full);summary(Perch_fit_full) #  VIF, drop1, output
```

VIF

Length1	Length2	Length3	Height	Width
1780.07795	4626.41486	2376.77263	54.03761	30.85782

MLR

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-7.0051	0.6527	-10.733	1.41e-14	***
Length1	0.4430	0.6227	0.711	0.480156	
Length2	-0.5321	0.9527	-0.558	0.579014	
Length3	0.4118	0.6465	0.637	0.527082	
Height	1.1788	0.3228	3.652	0.000622	***
Width	1.3754	0.3956	3.477	0.001059	**

Drop1

Df	Sum of Sq	RSS	AIC
		43.916	-1.6126
1	0.4445	44.360	-3.0487
1	0.2739	44.190	-3.2644
1	0.3563	44.272	-3.1601
1	11.7162	55.632	9.6306
1	10.6182	54.534	8.5143

Figure 5: VIF, MLR Output and Drop1 Global F-test, for the Full Model

The VIF helps us to identify multicollinearity and indicates that the variables have a very high collinearity as expected. Especially the three variables of the lengths result in massively higher VIF values than for the other variables.

The problem is also visible as the summary output in Figure 5 is examined. The variables $Length1$ $Length2$ and $Length3$ aren't even significant although they are highly correlated with the target. This is because the collinearity inflates the standard error which leads to insignificant t-statistics *[1]*.

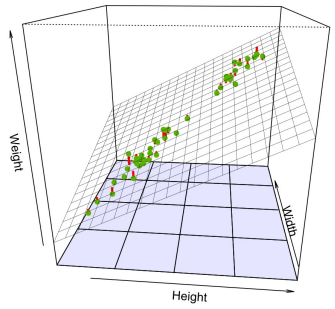
According to that research paper we can't reliably interpret these statistics. Further if there would be a full correlation between 2 variables meaning that X_1 would be a linear combination of X_2, the Covariance $\frac{1}{n}X^TX$ wouldn't have an inverse and *(1)* cannot be solved, although in the computer the estimation is done by minimizing the $RSS(\theta)$ and therefore we get results. With PCA*[3]* we could check which of the variables explain the most variance and then remove the "unimportant" ones.

Due to the task limitation, we just keep this consideration in mind and simply remove the Length variables and fit a model with two explanatory variables for a final comparison.

```
Perch_fit_two=lm(sqrt(Weight)~Width+Height,data = Perch_dat[,-1])
summary(Perch_fit_two)#  MLR
```

The regression model, which is now a 2 dimensional hyperplane is visualized in Figure 6 (left side).

$$Weight_i = (-6.646 + 1.579 * Width_i + 2.115 * Heigth_i)^2$$



Comparison of the models

Parameters		Full	Two	One
Model	RSS	31.39	43.41	104.45
	AIC	125.5	134.1	171.6
	R^2	0.9905	0.9868	0.9683
Predictions	MAD	0.433	0.457	0.469
	MSE	1.56	1.52	1.54

Figure 6: Left: Visualization 3d, Right: Results from the model comparison

Results

In the end we can compare the 3 models in Figure 6 (right table) with the earlier mentioned measurements. The distinction is made for measurements concerning the model itself, how well it fits the data and the predictions which were made with the model for the test data.

The RSS is the best for the full model, that was to expect since a model with more parameters always leads to smaller scores. The AIC is also better for the bigger model. The behavior of the AIC is decreasing and after a minimum it increases again therefore the AIC probably isn't at a minimum.

The R^2 isn't that much different from each other, whereas here is also to add: the more variables are in the model, the closer gets the measurement to 1

MSE and MAD look quite the same for all of the three models.

Conclusion

Einstein once said quote: *"Make everything as simple as possible, but not simpler"*. This is wonderfully consistent with modelling, we can see that the small model still leads to good predictions and we don't have to go to higher dimensions if a relation can be explained very well with two variables.

Appendix

This work is generated in R-Studio 2021.09.0 with R-4.1 with Rmarkdown, Full Code and Project can be found on https://github.com/buehlpa/TSM_MarkFor It is a graded assignment for the module "MarkFor" Market Forecasting from the Master of Science in Engineering MSE 2021 at Zurich University of Applied Science

Literature

[1] Multicollinearity and Regression Analysis, December 2017, Journal of Physics Conference Series 949(1):012009, Jamal Daoud, https://www.researchgate.net/publication/322212939_Multicollinearity_and_Regression_Analysis

[2] Linear Regression and its assumptions, January 24,2020, Manish Sharma, https://towardsdatascience.com/linear-regression-and-its-assumptions-ef6e8db4904d

[3] Principal Component Analysis, January 2017, International Journal of Livestock Research, Sidharth Mishra et. al, https://www.researchgate.net/publication/316652806_Principal_Component_Analysis