Market Forecasting Assignment: Multiple Linear Regression

Teacher: Dr. Stefan Groesser, Dr. Christoph Imboden

Autors: Benedech Rodolpho, 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 $\Theta = (X^T X)^{-1} X^T y$.

Dataset

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

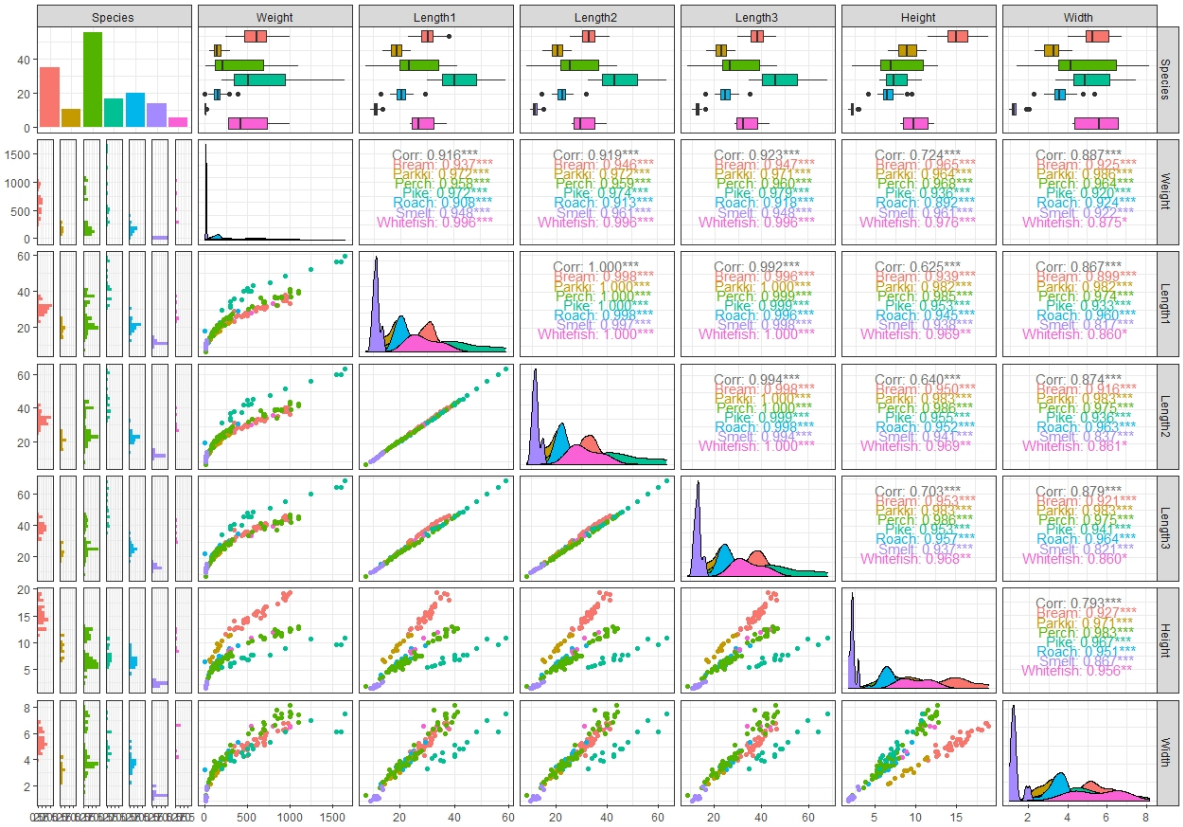

Figure 1: pairwise plot of the fish dataset

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)                               #  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 coloring 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 scatterplots a nonlinear relation between weight and the other numerical variables is observed, therefore the correlation *cannot* be correctly interpreted with these variablepairs. 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 helps see how the variables themselves are distributed. For example the width of Smelt tends to have a bidmodal distribution, altough this has to be taken with a grain of salt due to the sparse data, see on the inverted histograms on the left. Also the pairwise boxplots gives a good overview for comparison between the different variables. For example it can be seen that the variance of the species vary 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 set is split into a train and test set. We split the dataset using Simple Random Sampling without Replacement (SRSWR) in a ratio of 80% to 20%. 80% of the data is used for training and the remaining 20% is used to test prediction accuracy. The corresponding R code is given below.

```
set.seed(42)                                          # Reproducable code
Perch_dat=fish_dat%>%dplyr::filter(Species=="Perch") # Use only Perch
samp= sample(1:56,size = 11,replace = F)              # sample size 80%
Perch_train=Perch_dat[-samp,]                         # train sample
Perch_test =Perch_dat[samp,]                          #test sample
```

Because of multicolinearity, the first thing we want to do is fit a simple linear regression model. We use the weight of a fish as a response variable. We use the explanatory variable width, 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 fit 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 Squared Error (MSE).

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.

For a first approach we could just focus on the species Perch (in swiss-german it's called Egli) since we have the most data there. Lets find a model for predicting the weight of this particular species with the variables width. Since the weight hasn't got a linear relation to the other variables, we'll transform it properly.

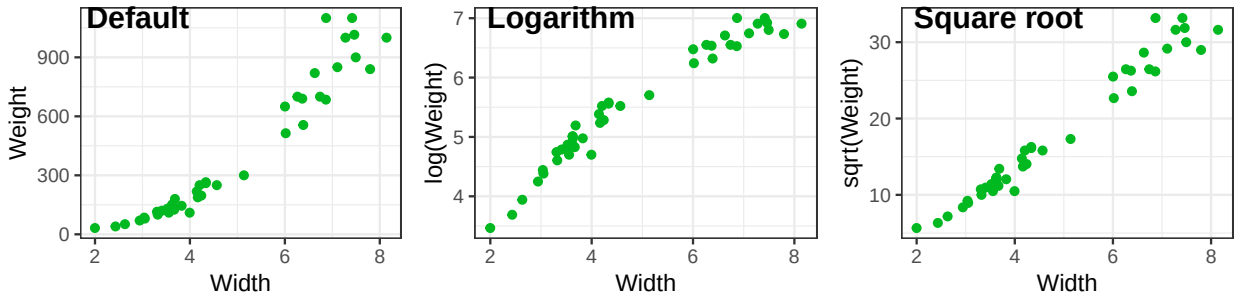

Figure 2: transformations of weight for Perch

If we transform the weight with the logarithm it looks a bit better but still isn't linear see Figure 2. The squareroot transformation seems to be the right one. Now the data seems much more linearly dependent. Therefore a first model can be set up

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

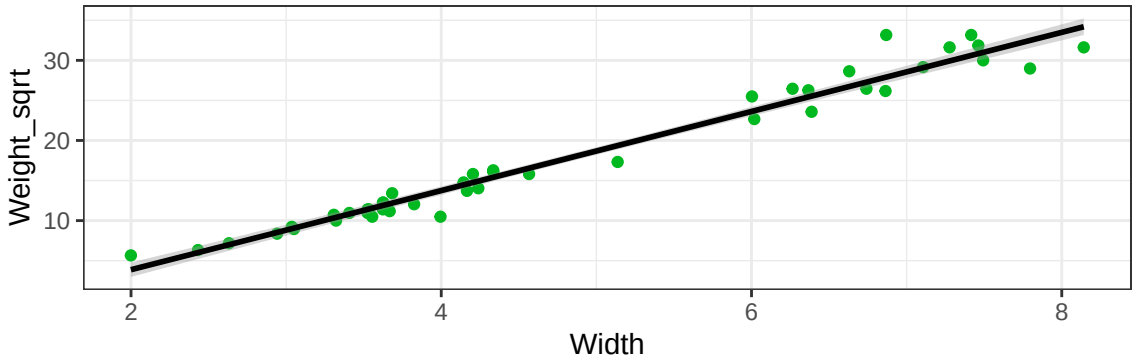

Figure 3: linear regression model for Perch weigth vs width

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 are also significant. By inspecting the residuals in Figure 4 we can see that the model assumptions of the . The residuals seem to follow a normal distribution, have constant mean and there are no points which are crucial to shift the model.

The way we have chosen is starting with a sparse model. Lets make now a full model and test if a larger model is more beneficial. The multicollinearity is checked with Variance Inflation Factor and a F test checks the importance of the variables

```
Perch_fit_full=lm(sqrt(Weight)~.,data = Perch_dat[,-1]) #  Multiple linear regression model
summary(Perch_fit_full)
```

```
##
```

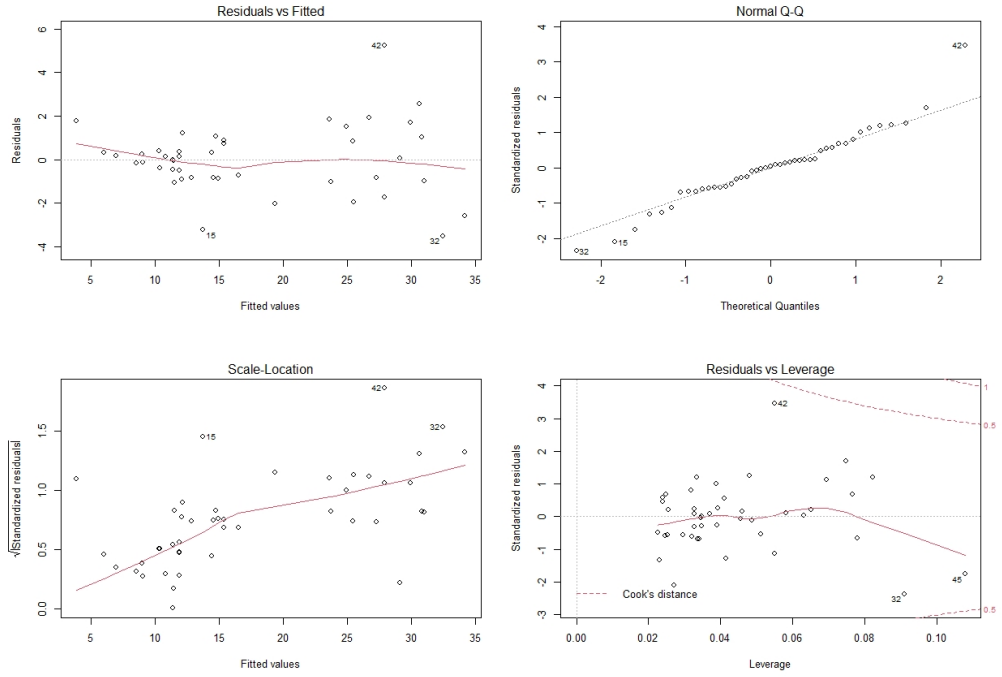

Figure 4: Residuals for the Perchfit

```
## Call:
## lm(formula = sqrt(Weight) ~ ., data = Perch_dat[, -1])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5930 -0.6264 -0.0330  0.5929  2.5313
##
## Coefficients:
##             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 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9372 on 50 degrees of freedom
## Multiple R-squared:  0.9896, Adjusted R-squared:  0.9885
## F-statistic: 947.4 on 5 and 50 DF,  p-value: < 2.2e-16
```

```
car::vif(Perch_fit_full)
```

```
##    Length1    Length2    Length3     Height      Width
## 1780.07795 4626.41486 2376.77263   54.03761   30.85782
```

```
drop1(Perch_fit_full)
```

```
## Single term deletions
##
## Model:
## sqrt(Weight) ~ Length1 + Length2 + Length3 + Height + Width
##         Df Sum of Sq    RSS     AIC
## <none>               43.916 -1.6126
## Length1  1    0.4445 44.360 -3.0487
## Length2  1    0.2739 44.190 -3.2644
## Length3  1    0.3563 44.272 -3.1601
## Height   1   11.7162 55.632  9.6306
## Width    1   10.6182 54.534  8.5143
```

Vif tells us that all the variables have very high kollinearity (as expected) whereas the vif of the lengths is massively hiigher than of the other varibles.