Homework 2

STT 461 - Danny Buglak

Contents

1 Material of homework 2 **2**

2 Task 1 - Prediction of fish weight **2**

2.1 Load the fish data set remove the *obs* variable and remove the rows of the data set for which there is a missing value for the variable sex. Also remove fish of species=2, because there will be only one observation there. 2

2.2 Convert the variables *sex* and *species* to factor variables 2

2.3 We want to build a linear regression model that will predict the weight of fish from other variables, as an alterantive of measuring the weight of fish with weight scales. Notice that length1, length2, width_percent and height_percent are all related to length3. Fit the linear regression model weight~species+sex+length3 and name it model1. Using the estimated coefficients write the equation of the linear regression line. 3

2.4 In order to do inference for the coefficients of the linear regression model we need the errors of the model to be independent and normally distributed with mean zero and constant variance. Plot the residual plot and the qq plot of the residuals and comment on wether the conditions for the model errors are met. 3

2.5 Provide an interpretation for the coefficient of determination R^2 of model1. 5

2.6 Based on the summary of model1 provide the test statistic and the p-value when testing Ho: $\beta_{length3} = 0$ vs Ha: $\beta_{length3} \neq 0$. Using the results of this test can we conclude that there is a linear relationship between length3 and weight? 5

2.7 Similarly, based on the summary of model1 provide the test statistic and the p-value when testing Ho: $\beta_{sex} = 0$ vs Ha: $\beta_{sex} \neq 0$. Using the results of this test should we remove the variable sex from model1? 6

2.8 Make the scatterplot of weight vs length3 and color it by the variable species. Also overlay the linear regrssion lines of weight vs length3 for each of the species. You will observe that that different species have a different relationship between their weight and lenght3. Hint: This type of scatterplot can be found in inclass assignment 25. 7

2.9 Based on the previous scatterplot there might be an interaction between lenght3 and species. Fit the model that will predict weight from species, length3 and the interaction term between length3 and species. Call this model, model2. 8

2.10 Using the ANOVA test, test if the interaction term should be included in the model2 or not. Write the null and alternative hypothesis conduct the test and make a conclusion. 8

2.11 Using model2 provide a prediction for the weight of fish of species 5 (Smelt) with legth3 (from the nose to the end of the tail) equal to 14cm. In addition provide a prediction interval for the weight of fish with those characteristics. Use confidence level equal to 0.95. 8

1 Material of homework 2

For homework 2 you will need to use methods covered in in-class assignments: 19 through 22 and also 25.

2 Task 1 - Prediction of fish weight

The *fish* data set has fish measurements from Laengelmavesi Lake, near Tampere in Finland. It has 159 observations of the following 8 variables:

- obs : Observation number
- species: One of 1 = Bream, 2 = Whitefish, 3 = Roach, 4 = Parkki, 5 = Smelt, 6 = Pike or 7 = Perch
- weight: Weight of fish (g)
- length1: Length from nose to the beginning of the tail (cm)
- length2: Length from nose to the notch of the tail (cm)
- length3: Length from nose to end of tail (cm)
- height_percent: Maximal height as percentage of length3
- width_percent: Maximal width as percentage of length3
- sex: 1 = male, 0 = female, NA = unknown

2.1 Load the fish data set remove the *obs* variable and remove the rows of the data set for which there is a missing value for the variable sex. Also remove fish of species=2, because there will be only one observation there.

```
fish  <- fosdata::fish [,-1]
id <- which(is.na(fish$sex))
fish <- fish[-id,]
species_2 <- which(fish$species == 2)
fish <- fish[-species_2,]
```

2.2 Convert the variables *sex* and *species* to factor variables

```
fish$sex <- as.factor(fish$sex)
fish$species <- as.factor(fish$species)
```

2.3 We want to build a linear regression model that will predict the weight of fish from other variables, as an alterantive of measuring the weight of fish with weight scales. Notice that length1, length2, width_percent and height_percent are all related to length3. Fit the linear regression model weight~species+sex+length3 and name it model1. Using the estimated coefficients write the equation of the linear regression line.

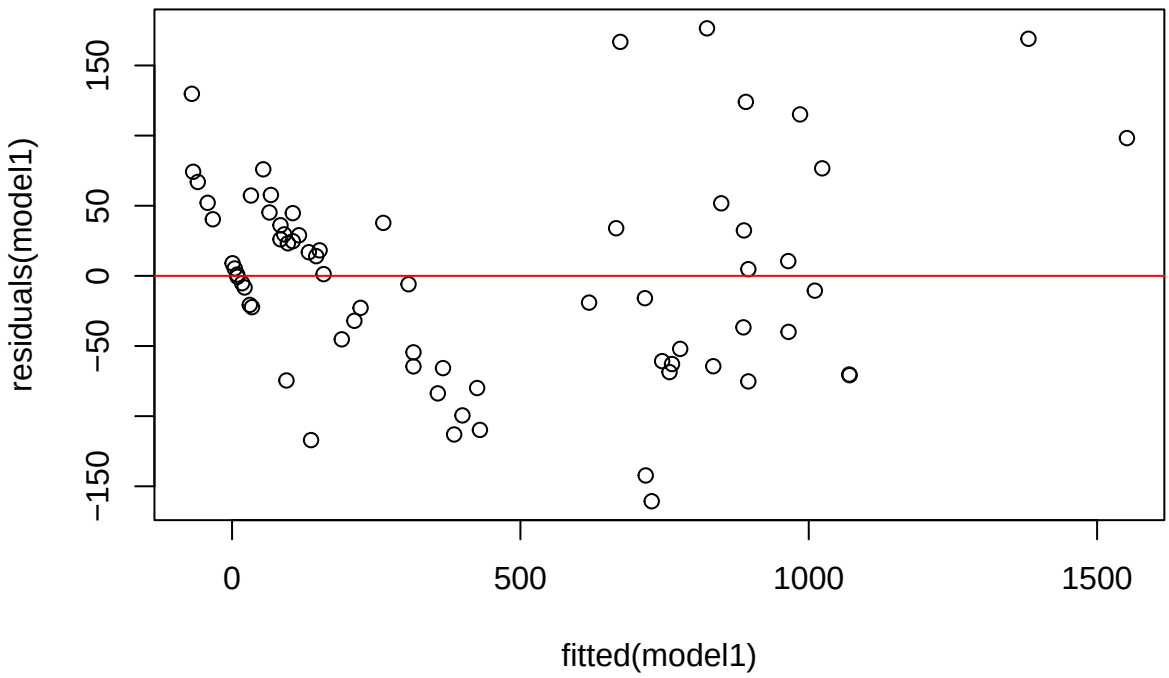
```
model1 <- lm(weight ~ species + sex + length3, data = fish)
model1
```

```
##
## Call:
## lm(formula = weight ~ species + sex + length3, data = fish)
##
## Coefficients:
## (Intercept)     species3     species4     species5     species6     species7
##     -995.48        73.92       156.64       440.71      -356.37        75.96
##        sex1      length3
##       26.10        42.70
```

$model1 = -995.48 + 73.92 * species3 + 156.64 * species4 + 440.71 * species5 - 356.37 * species6 + 75.96 * species7 + 26.1 * sex1 + 42.70 * length3$

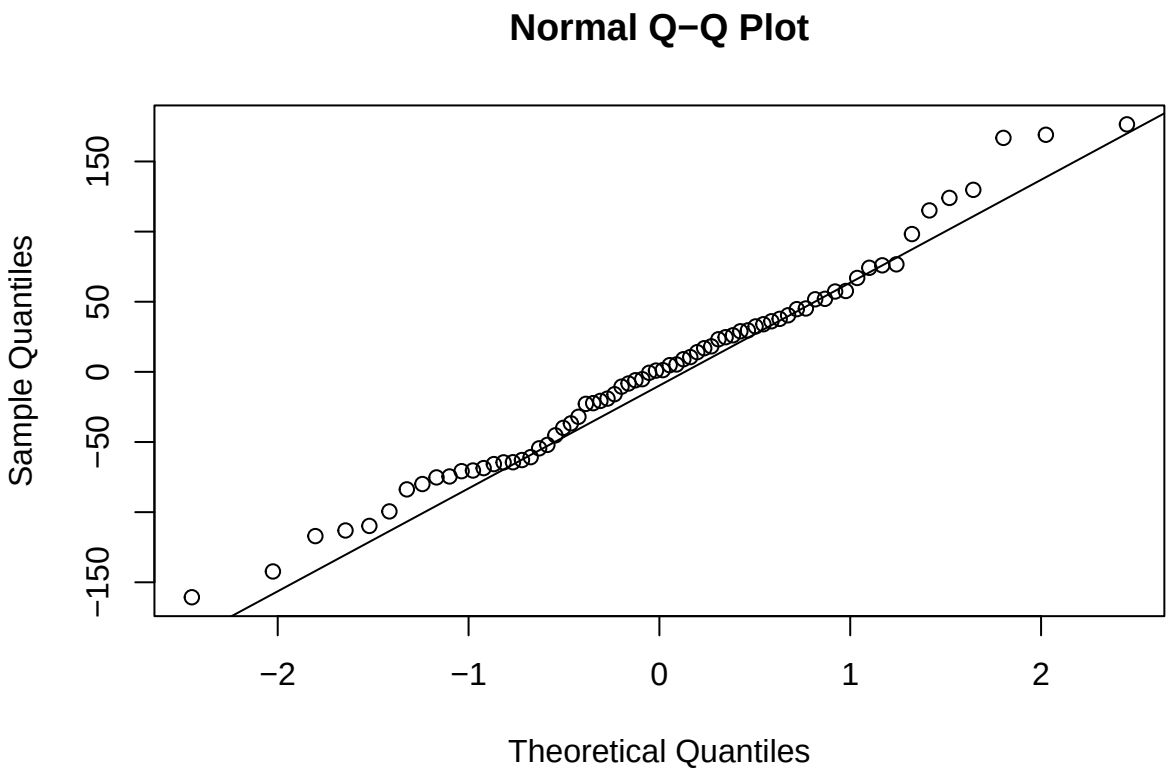
2.4 In order to do inference for the coefficients of the linear regression model we need the errors of the model to be independent and normally distributed with mean zero and constant variance. Plot the residual plot and the qq plot of the residuals and comment on wether the conditions for the model errors are met.

```
plot(residuals(model1) ~ fitted(model1))
abline(h = 0, col = 'red')
```



The points seem to scatter around the horizontal red line. This suggests that, on average, the residuals are close to zero.

```
qqnorm(residuals(model1))
qqline(residuals(model1))
```



The points generated from qqnorm are nearly on the line generated from qqline. Ideally, all points would be lying on the qqline. This suggests that the residuals are approximately normally distributed. The beginning and end of the plot have points that deviate from the line more than others, suggesting that there are some residuals that are outliers.

2.5 Provide an interpretation for the coefficient of determination R^2 of model1.

```
summary(model1)$r.squared
```

```
## [1] 0.9699016
```

This result from the coefficient of determination $ R^2 $ of approximately 0.9699 indicates that 96.99% of the variance in weight can be predicted by species, sex, and length3.

2.6 Based on the summary of model1 provide the test statistic and the p-value when testing Ho: $\beta_{length3} = 0$ vs Ha: $\beta_{length3} \neq 0$. Using the results of this test can we conclude that there is a linear relationship between length3 and weight?

```
summary_output <- summary(model1)
summary_output
```

```
##
## Call:
## lm(formula = weight ~ species + sex + length3, data = fish)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -160.657  -59.218    1.087   39.706  176.449
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -995.484     70.922 -14.036  < 2e-16 ***
## species3      73.919     49.153   1.504  0.13770
## species4     156.639     48.466   3.232  0.00197 **
## species5     440.712     55.033   8.008 3.83e-11 ***
## species6    -356.366     44.327  -8.039 3.38e-11 ***
## species7      75.956     35.837   2.120  0.03806 *
## sex1          26.098     25.318   1.031  0.30662
## length3       42.700      1.446  29.521  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 76.15 on 62 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.9699, Adjusted R-squared:  0.9665
## F-statistic: 285.4 on 7 and 62 DF,  p-value: < 2.2e-16
```

```
length3_estimate <- coef(summary_output)["length3", ]
length3_estimate
```

```
##     Estimate   Std. Error      t value     Pr(>|t|)
## 4.270036e+01 1.446461e+00 2.952057e+01 3.232624e-38
```

The length3 estimate has a p-value of less than 0.05, thus indicating that changes in the length of fish are associated with changes in weight which supports a linear relationship between the two variables.

2.7 Similarly, based on the summary of model1 provide the test statistic and the p-value when testing Ho: $\beta_{sex} = 0$ vs Ha: $\beta_{sex} \neq 0$. Using the results of this test should we remove the variable sex from model1?

```
sex_estimate <- coef(summary_output)["sex1", ]
sex_estimate
```

```
##   Estimate Std. Error    t value   Pr(>|t|)
## 26.0984886 25.3176633  1.0308411  0.3066199
```

The sex estimate has a p-value of approximately 0.307, which is greater than 0.05. This indicates that we could remove the variable sex from model1 and it would not have a large effect on the model accuracy.

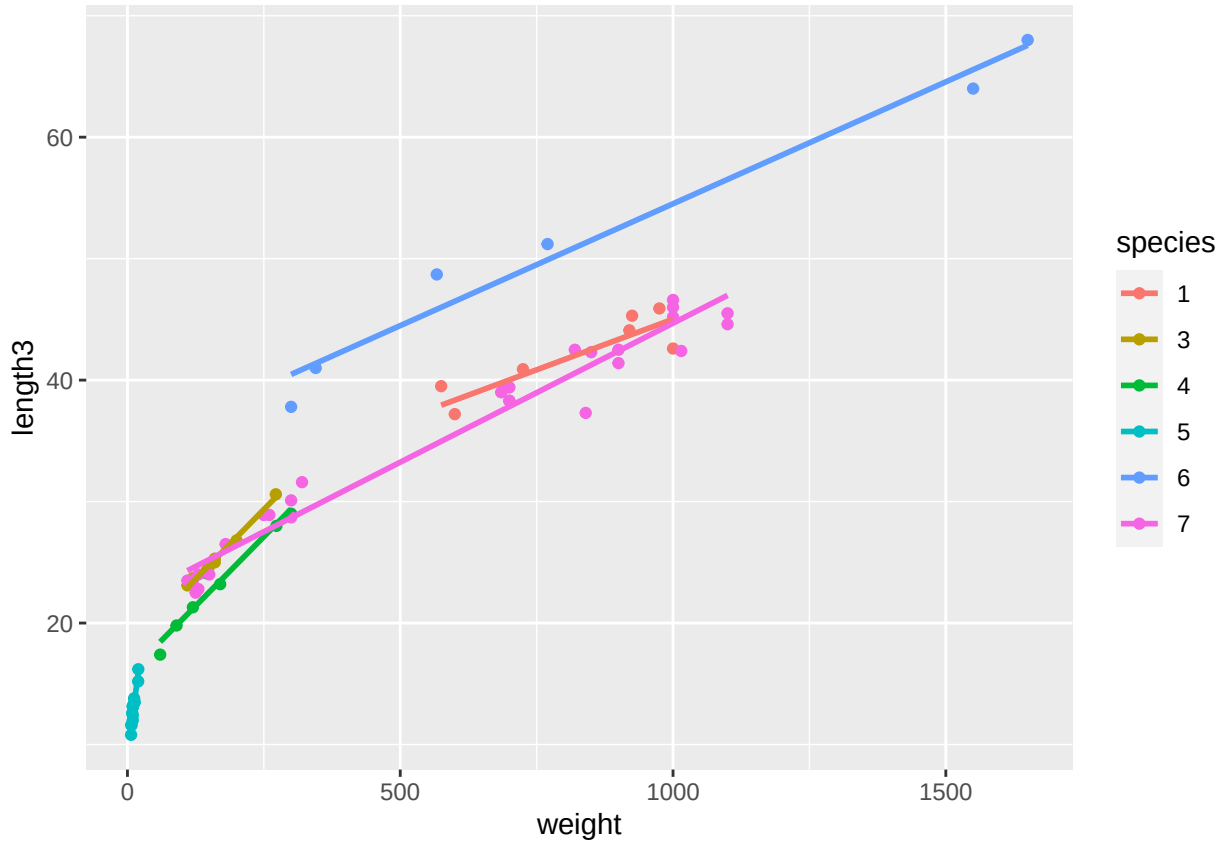
2.8 **Make the scatterplot of weight vs length3 and color it by the variable species. Also overlay the linear regrssion lines of weight vs length3 for each of the species. You will observe that that different species have a different relationship between their weight and lenght3. Hint: This type of scatterplot can be found in inclass assignment 25.**

```
library(ggplot2)
ggplot(fish, aes(x = weight, y = length3, color = species)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 1 rows containing non-finite values (`stat_smooth()`).
```

```
## Warning: Removed 1 rows containing missing values (`geom_point()`).
```



2.9 Based on the previous scatterplot there might be an interaction between lenght3 and species. Fit the model that will predict weight from species, length3 and the interaction term between length3 and species. Call this model, model2.

```
model2 <- lm(weight ~ species * length3, data = fish)
```

2.10 Using the ANOVA test, test if the interaction term should be included in the model2 or not. Write the null and alternative hypothesis conduct the test and make a conclusion.

```
anova(model1, model2)
```

```
## Analysis of Variance Table
##
## Model 1: weight ~ species + sex + length3
## Model 2: weight ~ species * length3
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     62 359544
## 2     58 230075  4    129469 8.1595 2.729e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ho: There is no improvement in the model by adding the interaction term. Ha: There is a significant improvement in the model by adding the interaction term.

The output from the anova function comparing the two models indicates that model2 is better at predicting the weight than model1 is. This can be seen because the p-value calculated is incredibly small and is less than 0.05. This can lead us to reject the null hypothesis. The interaction term found in model2 significantly improves the model.

2.11 Using model2 provide a prediction for the weight of fish of species 5 (Smelt) with legth3 (from the nose to the end of the tail) equal to 14cm. In addition provide a prediction interval for the weight of fish with those characteristics. Use confidence level equal to 0.95.

```
testData <- data.frame(species = factor(5, levels = levels(fish$species)), length3 = 14)

predict(model2, newdata = testData, interval = "prediction", level = 0.95)
```

```
##        fit       lwr      upr
## 1 13.82589 -118.7959 146.4477
```

The predicted value for the weight of a fish of species 5 with length3 14cm is approximately 13.826. The lower bound for the 95% confidence interval is -118.80 and the upper bound is 146.45.