Example of the Lab Assignment

Assumptions and Diagnostics in Linear Regression

Yulia E

Contents

Setup

Install teh following libraries:

```
library(readr)
library(ggplot2)
library(cowplot)
library(praise)
```

Exercise 1: Conceptual Questions

rubric={autograde:6}

Ex 1.1

Residuals are . . .

1. Data collected from individuals that is not consistent with the rest of the group
2. Variation in the response variable that is explained by the model
3. The difference between the observed response and the values predicted by the model
4. Possible models not explored by the researcher

5. The difference between the values predicted by the model and the observed response

```
answer1_1 <- NULL

# BEGIN SOLUTION
answer1_1 <- 3
answer1_1
```

```
## [1] 3
```

```
# END SOLUTION
```

Ex 1.2

Which of the following is true about Residuals ?

1. Lower is better
2. Higher is better
3. 1 or 2 - depends on the situation
4. None of these

```
answer1_2 <- NULL

# BEGIN SOLUTION
answer1_2 <- 1
answer1_2
```

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

```
# END SOLUTION
```

Ex 1.3

A scientist is graphing data points from an experiment and concludes that the data set is linear. Which of these statements BEST explains why they drew this conclusion?

1. Because the relationship between the dependent and independent variables is linear.
2. Because the variables are independent of each other.
3. Because the variables are related exponentially.
4. Because the residuals are random in magnitude

```
answer1_1 <- NULL

# BEGIN SOLUTION
answer1_1 <- 1
answer1_1
```

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

```
# END SOLUTION
```

Ex 1.4

True or false? before checking assessing a numeric measures of goodness of fit (like R^2), you should evaluate the residuals plot.

```
answer1_4 <- NULL

# BEGIN SOLUTION
```

```
answer1_4 <- "true"
answer1_4
```

```
## [1] "true"
```

```
# END SOLUTION
```

Ex 1.5

Suppose you plotted a scatter plot between the residuals and predicted values in linear regression and you found that there is a relationship between them. Which of the following conclusion do you make about this situation?

1. Since the there is a relationship, it means our model is good
2. Can't really say
3. Since the there is a relationship, it means our model is not good

```
answer1_5 <- NULL

# BEGIN SOLUTION
answer1_5 <- 3
answer1_5
```

```
## [1] 3
```

```
# END SOLUTION
```

Ex 1.6

When the error terms have a constant variance, a plot of the residuals versus the independent variable x has a pattern that

1. fans out
2. funnels in
3. fans out, but then funnels in
4. forms a horizontal band pattern
5. forms a linear pattern that can be positive or negative

```
answer1_6 <- NULL

# BEGIN SOLUTION
answer1_6 <- 4
answer1_6
```

```
## [1] 4
```

```
# END SOLUTION
```

Exercise 2: Age of Abalone Shells

This task is a common example of the activity from the field of zoology. We will be working with the Abalone, a marine mollusk related to snails and whelks. To monitor a population, it is critical to know the age of the mollusc to make sure that it is harvested sustainably. The age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope. However this time is labour and time consuming. However, there are several measurements that can be easily obtained can be use to predict the age of the organism (even without killing the organism!).

We will work with the dataset that can be accessed here. The dataset has the following columns:

- `Sex`: M, F, and I (infant)
- `Length`: Longest shell measurement (feet)
- `Diameter`: Shell diameter (perpendicular to length)
- `Height`: Height of the organism (with meat in shell)
- `Whole weight`: whole abalone weight (pounds)
- `Shucked weight`: weight of meat
- `Viscera weight`: gut weight (after bleeding)
- `Shell weight`: after being dried
- `Rings`: +1.5 gives the age in years

Let's start by loading the data:

```
abalone <- read_csv("data/abalone.csv")
```

```
## Rows: 4177 Columns: 9
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): Sex
## dbl (8): Length, Diameter, Height, Whole weight, Shucked weight, Viscera wei...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(abalone)
```

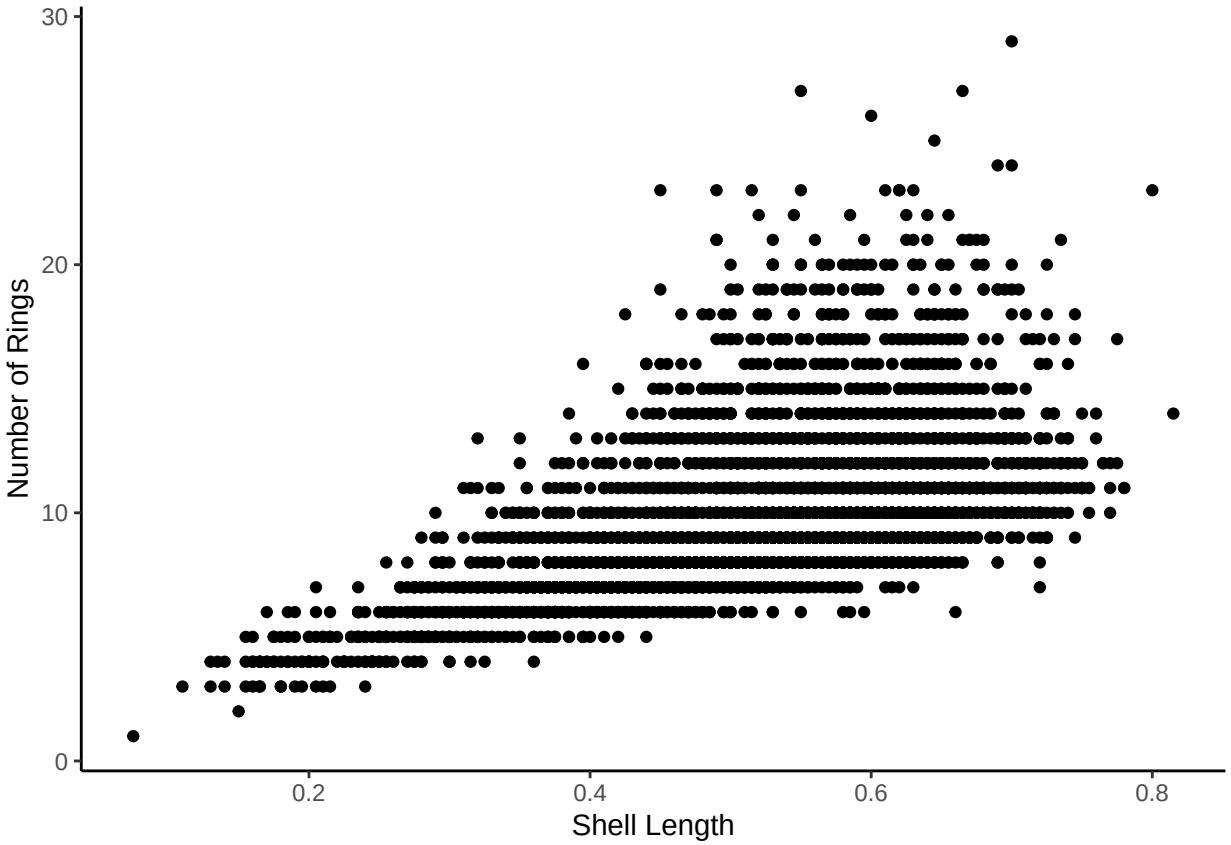
```
## # A tibble: 6 x 9
##   Sex   Length Diameter Height `Whole weight` `Shucked weight` `Viscera weight`
##   <chr>  <dbl>    <dbl>  <dbl>          <dbl>            <dbl>             <dbl>
## 1 M      0.455    0.365  0.095          0.514           0.224             0.101
## 2 M      0.35     0.265  0.09           0.226           0.0995            0.0485
## 3 F      0.53     0.42   0.135          0.677           0.256             0.142
## 4 M      0.44     0.365  0.125          0.516           0.216             0.114
## 5 I      0.33     0.255  0.08           0.205           0.0895            0.0395
## 6 I      0.425    0.3    0.095          0.352           0.141             0.0775
## # ... with 2 more variables: `Shell weight` <dbl>, Rings <dbl>
```

Ex 2.1

rubric={reasoning:2,viz:3}

Abalone researcher desires to devise a method that will help them to determine the age of the organism, without killing it. Thus, they propose to use the length of the shell as a proxy of the abalone's age. They ask you to produce a scatterplot `ring_len.plt` showing the relationship between two variables.

```
# BEGIN SOLUTION
ring_len.plt <- ggplot(abalone, aes(y=Rings, x=Length))+geom_point()+
  theme_classic()+
  labs(x="Shell Length", y="Number of Rings")
ring_len.plt
```




```
# END SOLUTION
```

In one or two sentences, explain Which variable would go on the y-axis? Why?)

Shell length is a independent variable while number of rings is the dependent variables. Traditionally, we would plot the dependent variable (`Rings`) on y-axis and the `Length` on the axis.

Ex 2.2

rubric={reasoning:2,accuracy=2,viz:2}

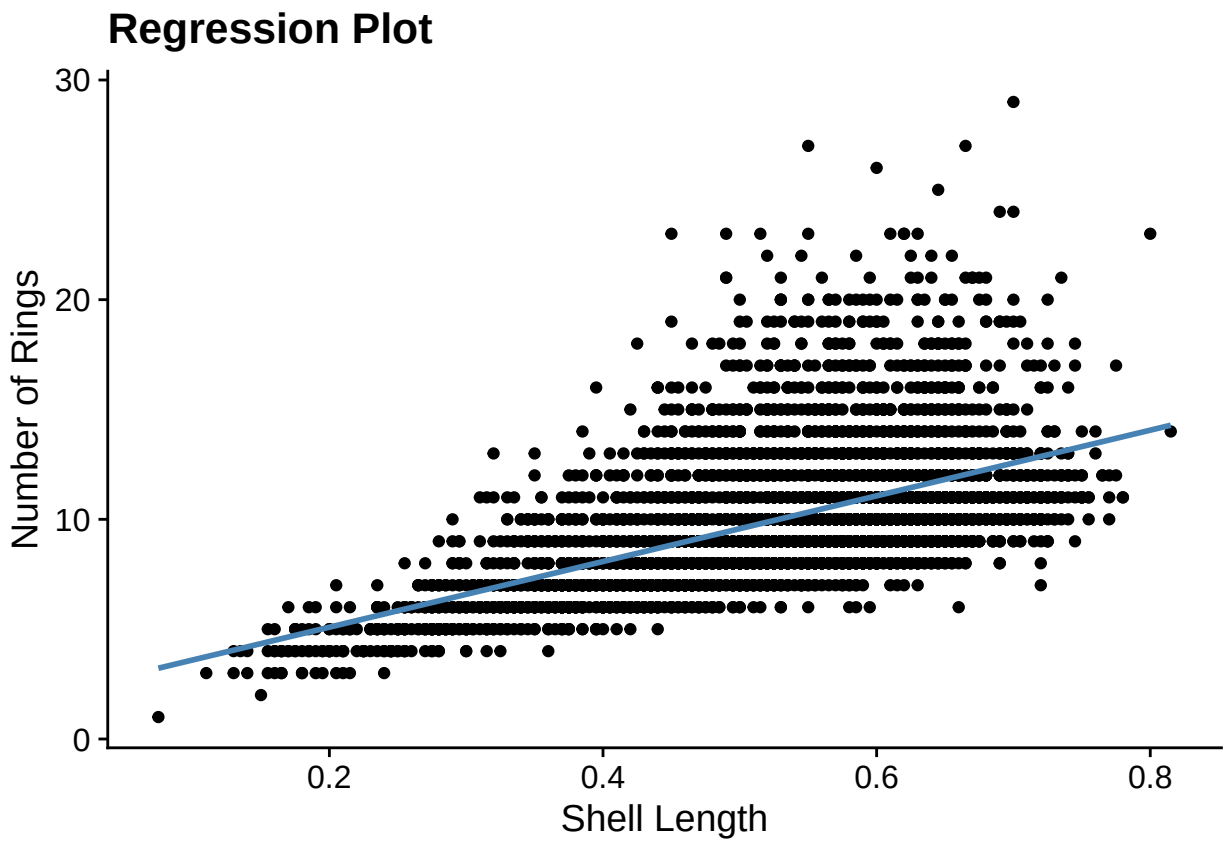
Create a simple linear regression model called `model` to help determine the associations the response `Rings` and `Height` as the predictor x.

```
# BEGIN SOLUTION
model <- lm(Rings~ Length, data=abalone)
# END SOLUTION
```

Add a regression line to the scatterplot `ring_len.plt`

```
# BEGIN SOLUTION
ring_len.plt +
  geom_smooth(method = lm, se = FALSE, color="steelblue")+
  labs(title="Regression Plot")+
  theme_cowplot()
```

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




```
# END SOLUTION
```

In 1-2 sentences, describe whether there appear to be a linear relationship between the variables?

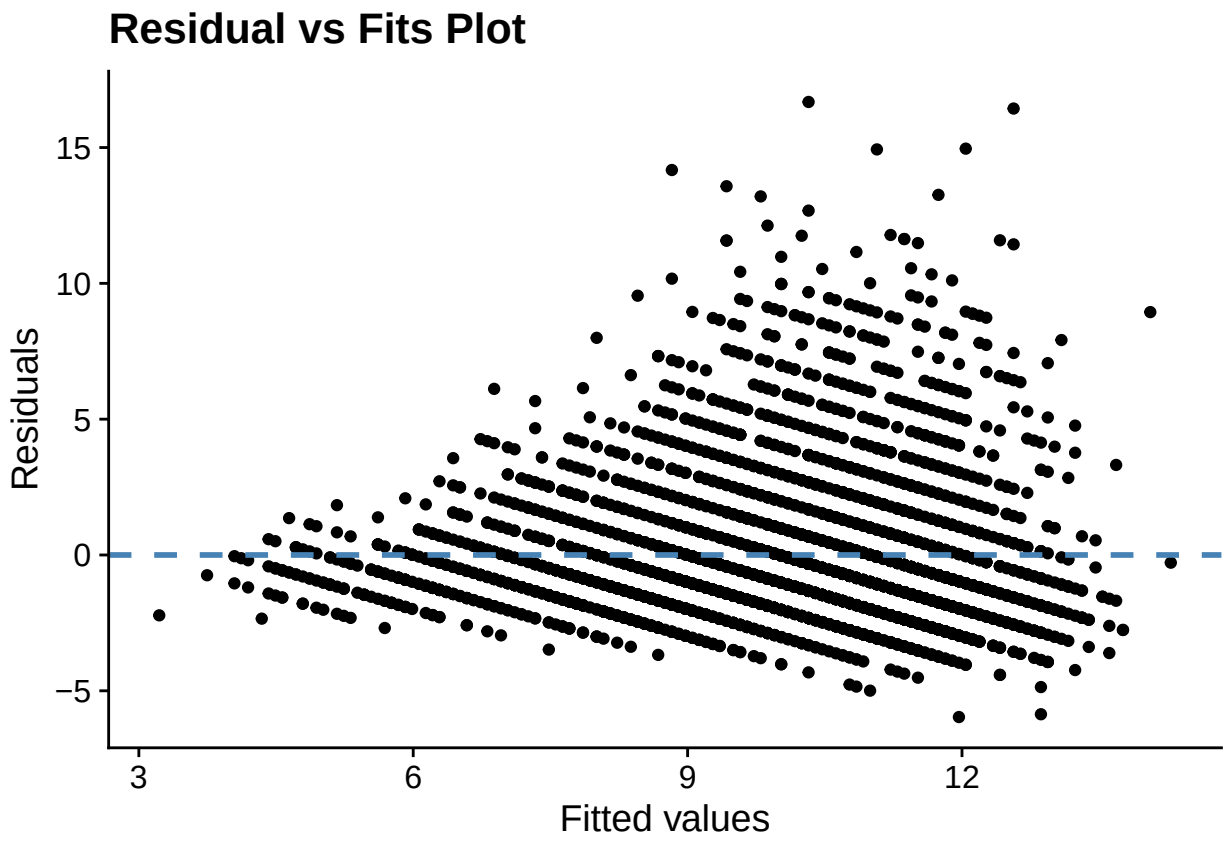
Originally, the plot looked quite linear. However after adding a regression line, we can observe that the number of rings increase more sharper with the shell length as the linear model propose. In addition, we can see that the older shells (shels with larger number of rings) have more variable shell length compare to the younger shells.

Ex 2.3

rubric={reasoning:3,accuracy=2,viz:2}

Built a residuals vs. fit plot.

```
# BEGIN SOLUTION
ggplot(model , aes(.fitted, .resid))+geom_point()+
  geom_hline(yintercept=0, col="steelblue", linetype="dashed", lwd=1)+
  xlab("Fitted values")+ylab("Residuals")+
  ggtitle("Residual vs Fits Plot")+theme_cowplot()
```




```
# END SOLUTION
```

In 3-5 sentences, interpret the output of the residual plot, state whether assumptions were violated or not?

we can observe a “fanning” effect of the residuals where residuals are more spread when fitted values are larger than when the fitted values are small.residuals “fan out” from left to right rather than exhibiting a consistent spread around the residual = 0 line. The residual vs. fits plot suggests that the error variances are not equal.

Thinking back to what type of data can be modelled with simple linear regression, do you think it is reasonable to use this model to answer this research question?

SLR is used to model continuous dependent variable, while in this case we have we are modelling count data, so other regression models will be more appropriate (eg Poisson Regression)

Writing

rubric={writing:2}

To get the marks for this writing component, you should:

- Use proper English, spelling, and grammar throughout your submission (the non-coding parts).
- Be succinct. Please dont go above the sugested sentence count

Submission

rubric={mechanics:5}

Congratulations! You are done the lab! Run the following command to receive a small praise from R:)

```
praise()
```

```
## [1] "You are brilliant!"
```

The final (but important steps):

- Knit the assignment to generate `.pdf` file . Make sure that all the text, figures and tables are properly displayed in your `.pdf` and `.Rmd`.
- Push everything to your Github repo. You need to have a minimum of 3 commits.
- Paste a link to your GitHub Repo:

`URL to your GitHub repo: **INSERT YOUR GITHUB REPO URL HERE**`

- Upload the `.Rmd` AND the `.pdf` files to Gradescope.