HW-4

Jesse Grigolite

2023-05-23

The repository is linked here: here.

Problem #1

1.) Null Hypothesis: Fish length does not predict fish weight in trout perch. H0: beta sub one = 0 Alternative Hypothesis: Fish length does predict fish weight in trout perch. Ha: beta sub one does not = 0

2.)

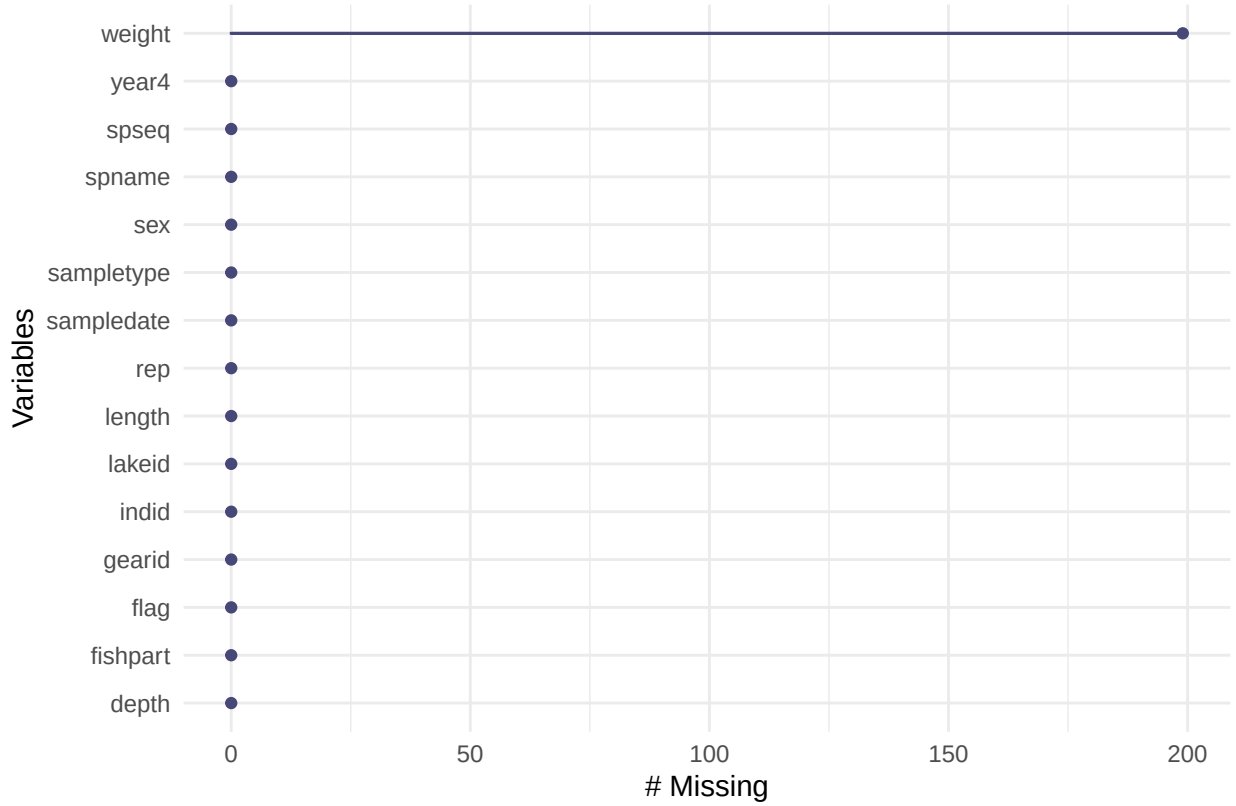
```
library(tidyverse)
library(here)
library(lterdatasampler)
library(performance)
library(broom)
library(flextable)
library(ggeffects)
library(car)
library(naniar)
```

```
#read in data set
fish <- read.csv(here("data", "ntl6_v12.csv"))
```

```
#only focus in on the species trout perch
fish_data <- fish %>%
  filter(spname == "TROUTPERCH")
```

b.)

```
#now visualzing missing data
gg_miss_var(fish_data) +
  #adding a caption
  labs(caption = "The above figure illustrates that there are 200 missing values for trout perch weight
```



re illustrates that there are 200 missing values for trout perch weight which will influence the sample size of our data.

3.)

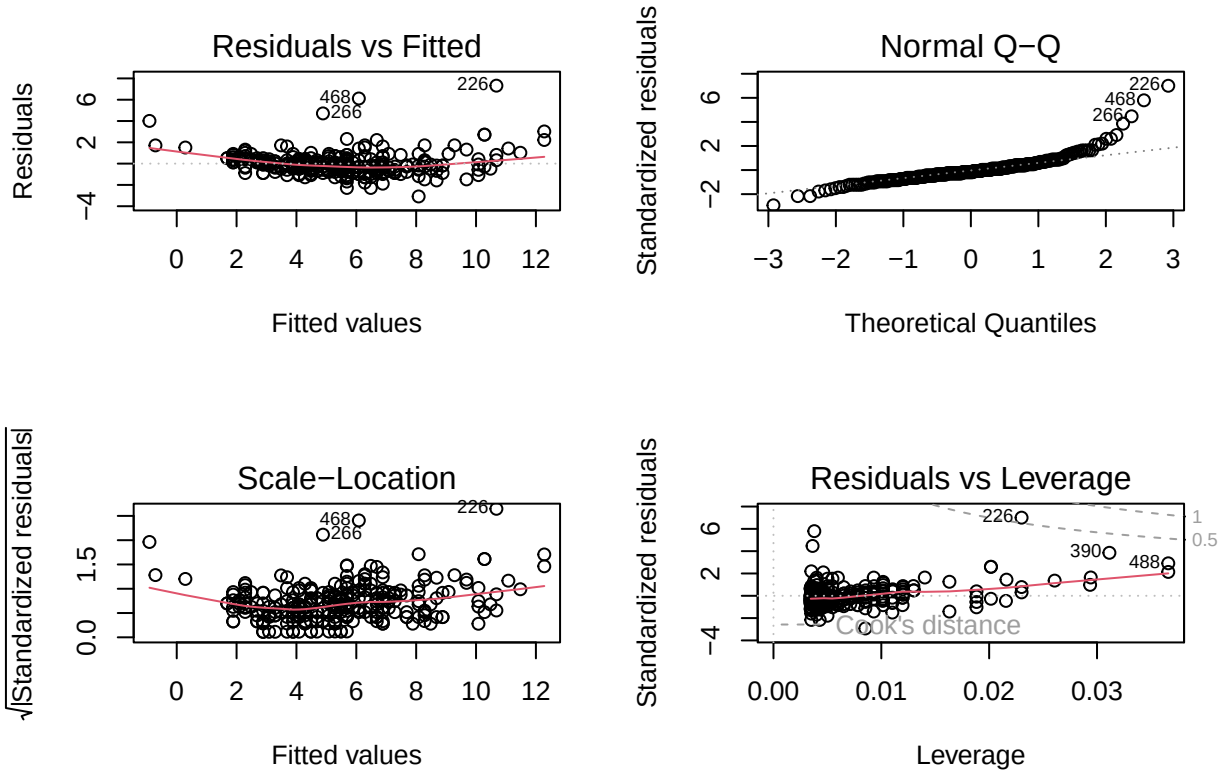
```
#creating a linear model between predictor and response variables using the filtered data set
fish_model <- lm(weight ~ length, data = fish_data)

fish_model
```

```
##
## Call:
## lm(formula = weight ~ length, data = fish_data)
##
## Coefficients:
## (Intercept)       length
##    -11.7025       0.1999
```

4.)

```
#first create a 2 by 2 grid and then visually check your assumptions
par(mfrow = c(2, 2))
plot(fish_model)
```



5.) Residuals vs Fitted: This plot shows the constant variance of residuals, and appears to be randomly distributed about the red line and is fairly homoscedastic. Normal Q-Q: This plot displays how normally distributed the residuals are. Based on the majority of the points following a linear trend this does appear to be normally distributed. Scale Location: This plot also demonstrates that the square root of residuals are randomly placed around the red line and appear to be homoscedastic. Cooks Distance: This plot demonstrates that there is only a single out liar in the data set, however, since this is only a single point I do not think it will significantly influence the linear model.

```
#turning off the 2 by 2 grid
dev.off()
```

```
## null device
##           1
```

6.)

```
#creating a summary and naming that object
model_summary <- summary(fish_model)
model_summary
```

```
##
## Call:
## lm(formula = weight ~ length, data = fish_data)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -3.0828 -0.4862 -0.1830  0.4128  7.3191
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -11.702476   0.481564  -24.30   <2e-16 ***
## length        0.199852   0.005584   35.79   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.057 on 288 degrees of freedom
##   (199 observations deleted due to missingness)
## Multiple R-squared:  0.8164, Adjusted R-squared:  0.8158
## F-statistic:  1281 on 1 and 288 DF,  p-value: < 2.2e-16
```

7.)

```
#creating anova table, first by naming object
model_squares <- anova(fish_model)

#now making sure to add easy to read names and tidying up labels
model_squares_table <- tidy(model_squares) %>%
  mutate(p.value = case_when(
    p.value < 0.001 ~ "< 0.001"
  )) %>%
  flextable() %>%
  set_header_labels(df = "Degrees of Freedom", sumsq = "Sum of Squares", meansq = "Mean Squares", statis
model_squares_table
```

term	Degrees of Freedom	Sum of Squares	Mean Squares	F-statisticp-value
length	1	1,432.28771	1,432.287687	1,280.844< 0.001
Residuals	288	322.0525	1.118238	

8.) The ANOVA table relates to the previous results of the "summary()" function by providing data on how well the predictor variable, in this case length, predicts the response variable, in this case weight, and illustrates this by providing an F-statistic with a corresponding p-value. Similarly, the "summary()" function provides data on the model's estimated coefficients (i.e. slope and intercept) and gives a t statistic to show whether they are significant predictors within the model.

9.) We hypothesized that fish length would predict fish weight in trout perch, with our null stating that length would not predict weight. We analyzed 289 trout perch and found that length does significantly predict weight in a linear regression model via an F-test (F = 1, 280.8, DF = 288, R-squared = 0.82, alpha = 0.05,p = < 0.001). For each gram increase in trout perch weight, we expect a 0.2 (SE +/- 0/006) millimeter increase in trout perch length.

10.)

```
#creating an object for predictions
predictions <- ggpredict(fish_model, terms = "length")

#naming the object and first plotting the observations of length and weight in a point plot
```

```r
plot_predictions <- ggplot(data = fish_data,
                           aes(x = length, y = weight)) +
  geom_point() +
  # then plot the predictions with the linear regression line
  geom_line(data = predictions,
            aes(x = x, y = predicted),
            color = "red", linewidth = 1) +
  # then plot the 95% confidence interval from ggpredict which will be displayed with the thicker band
  geom_ribbon(data = predictions,
              aes(x = x, y = predicted, ymin = conf.low, ymax = conf.high),
              alpha = 0.2) +
  # designating theme and meaningful labels and title
  theme_bw() +
  labs(x = "Trout Perch Length (mm)",
       y = "Trout Perch Weight (g)",
       title = "Predicted Relationship of Trout Perch Length to Weight Over Recorded Observations")

plot_predictions
```