Homework 4

Luke Todd

2023-05-26

Loading Packages

```
library(here)
```

```
## here() starts at /Users/luketodd/Desktop/Rprojects/ES 193DS/Homework 4
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.1     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(car)
```

```
## Loading required package: carData
##
## Attaching package: 'car'
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
## The following object is masked from 'package:purrr':
##
##     some
```

```
library(sqldf)
```

```
## Loading required package: gsubfn
## Loading required package: proto
```

```
## Warning in doTryCatch(return(expr), name, parentenv, handler): unable to load shared object '/Library
##   dlopen(/Library/Frameworks/R.framework/Resources/modules//R_X11.so, 6): Library not loaded: /opt/X1
##   Referenced from: /Library/Frameworks/R.framework/Versions/4.2/Resources/modules/R_X11.so
##   Reason: image not found

## tcltk DLL is linked to '/opt/X11/lib/libX11.6.dylib'
## Could not load tcltk.  Will use slower R code instead.
## Loading required package: RSQLite
```

```
library(janitor)
```

```
##
## Attaching package: 'janitor'
##
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
library(ggthemes)
library(naniar)
library(performance)
library(flextable)
```

```
##
## Attaching package: 'flextable'
##
## The following object is masked from 'package:purrr':
##
##     compose
```

```
library(broom)
library(ggeffects)
```

Loading data

```
raw_data <- read_csv(here('ntl6_v12.csv'))
```

```
## Warning: One or more parsing issues, call `problems()` on your data frame for details,
## e.g.:
##   dat <- vroom(...)
##   problems(dat)

## Rows: 349229 Columns: 15
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (8): lakeid, gearid, spname, sampletype, indid, fishpart, spseq, flag
## dbl  (5): year4, depth, rep, length, weight
## lgl  (1): sex
## date (1): sampledate
##
## 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.
```

Problem 1

Mathematical:
H0: $\beta_1 = 0$
H1: β_1 != 0

Biological:
H0: Fish length is not a significant predictor of fish weight for trout perch.
H1: Fish length is a significant predictor of fish weight for trout perch.

Problem 2

```
# cleaning data
fish_data <- sqldf("SELECT year4, spname, length, weight FROM raw_data
                    WHERE spname = 'TROUTPERCH'")

# missing data viz
vis_miss(fish_data)
```

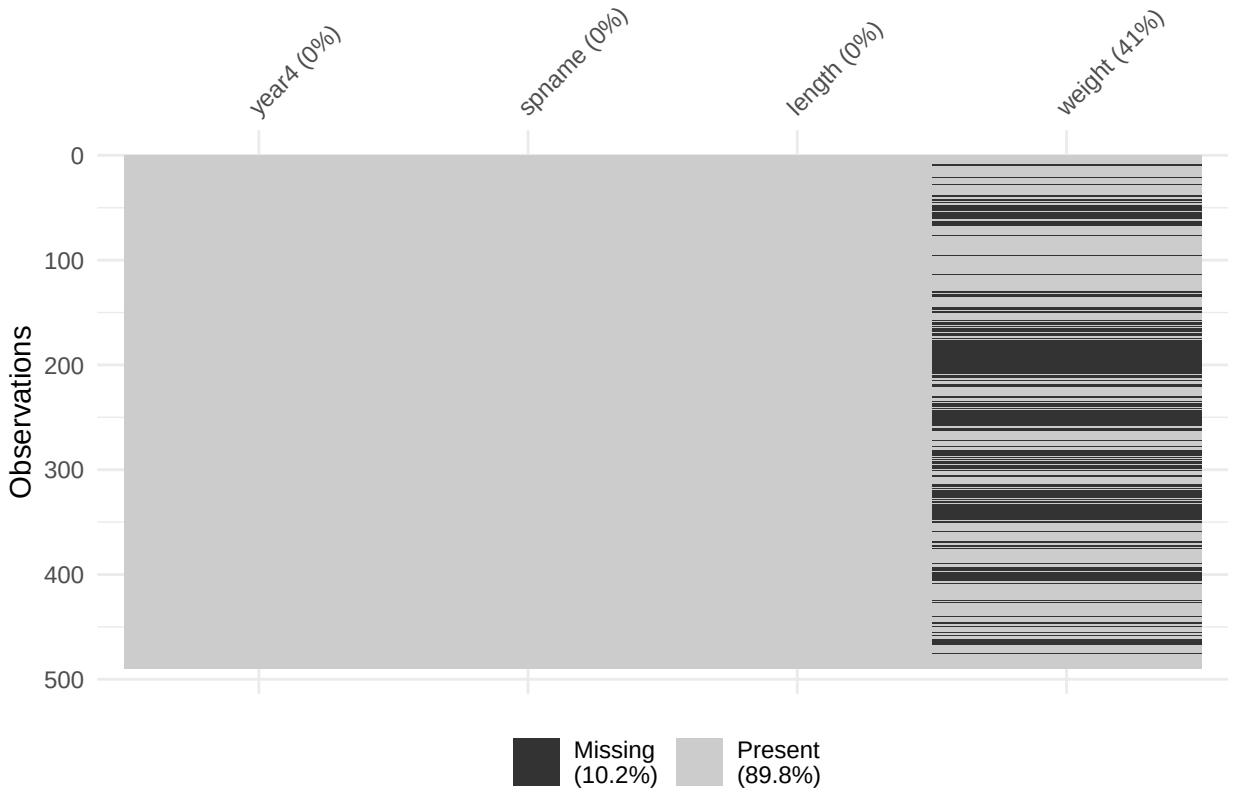

Figure 1: This figure shows that there are no missing values for year4, spname, and length, but 41% of the data is missing for weight.

Problem 3

```
fish_lm <- lm(weight ~ length, data = fish_data) # linear model
fish_res <- fish_lm$residuals # calc residuals
```

Problem 4

```
par(mfrow = c(2,2)) # plots two next to each other
plot(fish_lm) # plotting
```

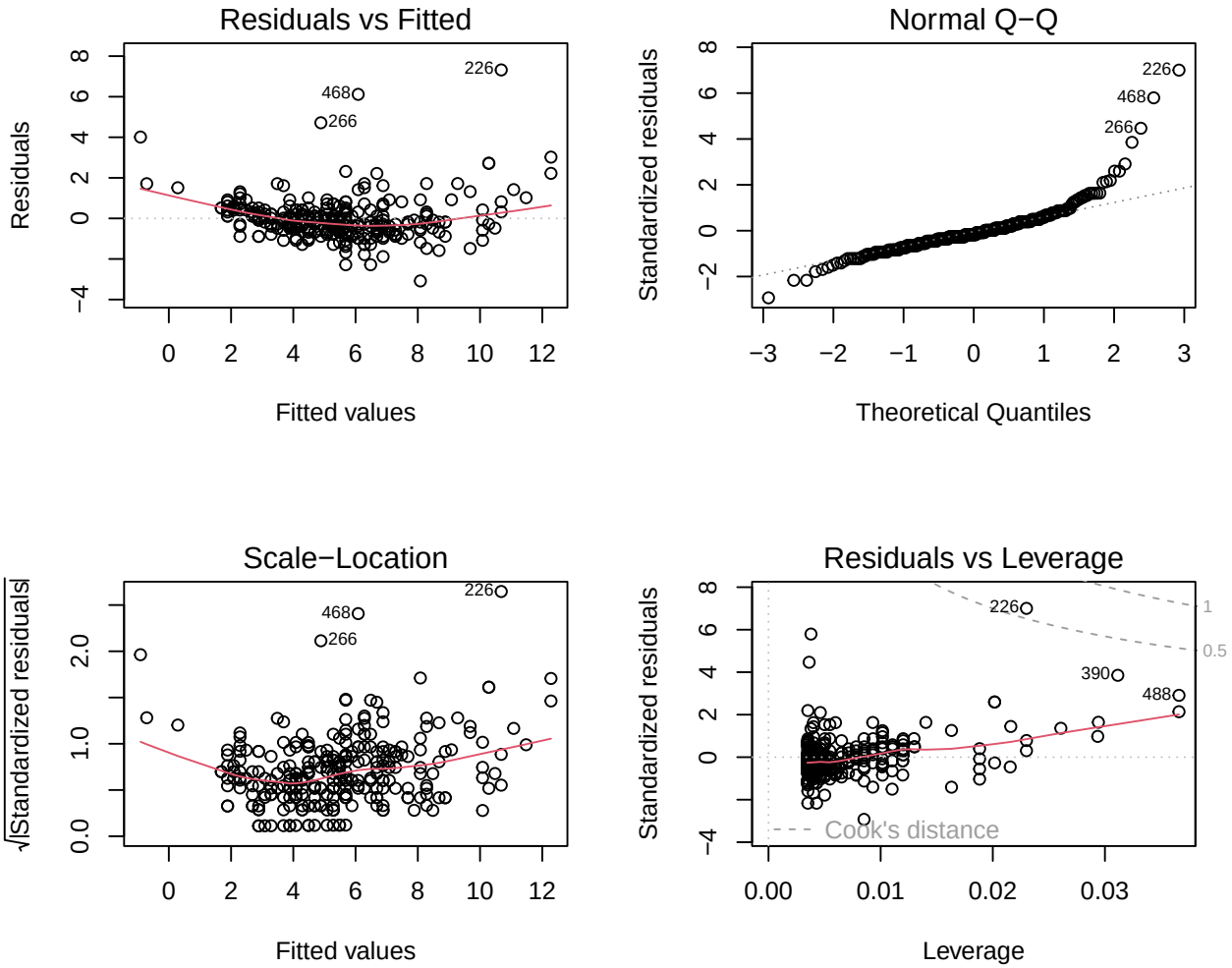

Figure 2: This figure contains four different diagnostic plots.

Problem 5

The residuals vs. fitted plot shows linearity and constant variance. Based on the plot, the points seem to be randomly distributed.

The QQ norm plot is used to test if a dataset is normally distributed. Based on the plot, it appears to be normally distributed.

The scale location plot is used to show homoscedasticity of variance. Based on the plot, the data seems to be randomly distributed.

The residuals vs. leverage shows influential data points. Based on the plot, there appear to be some high leverage points.

Problem 6

```
summary(fish_lm)
```

```
##
## 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
```

Problem 7

```
fish_anova <- anova(fish_lm) # anova table

# pulled from workshop 7
fish_table <- tidy(fish_anova) %>%
                      #round the sum of squares and mean squares columns to have 5 digits
                      mutate(across(sumsq:meansq, ~ round(.x, digits = 5))) %>%
                      #round the F-statistic to have 1 digit
                      mutate(statistic = round(statistic, digits = 1)) %>%
                      #replace the very very very small p value with < 0.001
                      mutate(p.value = case_when(
                             p.value < 0.001 ~ "< 0.001")) %>%
                      #rename the stem_length cell to be meaningful
                      mutate(term = case_when( term == "length" ~ "Fish Length (mm)",
                             TRUE ~ term)) %>%
```

```
                    #make the data frame a flextable object
                    flextable() %>%
                    #change the header labels to be meaningful
                    set_header_labels(df = "Degrees of Freedom",
                                      sumsq = "Sum of squares",
                                      meansq = "Mean squares",
                                      statistic = "F-statistic",
                                      p.value = "p-value")
fish_table
```

```
## Warning: fonts used in ‘flextable‘ are ignored because the ‘pdflatex‘ engine is
## used and not ‘xelatex‘ or ‘lualatex‘. You can avoid this warning by using the
## ‘set_flextable_defaults(fonts_ignore=TRUE)‘ command or use a compatible engine
## by defining ‘latex_engine: xelatex‘ in the YAML header of the R Markdown
## document.
```

term	Degrees of Freedom	Sum of squares	Mean squares	F-statisticp-value
Fish Length (mm)	1	1,432.28771,432.28769		1,280.8< 0.001
Residuals	288	322.0525	1.11824	

Problem 8

This table shares a lot of the same information as the summary() including the f-statistic, p-value, and degrees of freedom.

Problem 9

With the data passing all diagnostic checks in problem 4, I was able to run a linear regression model to investigate the relationship between length and weight in trout. This linear regression, as shown in the table in problem 7, calculated a p-value of < 0.001, showing that length is a significant predictor of weight in trout.

Problem 10

```
# extract model predictions
predictions <- ggpredict(fish_lm, terms = "length")
predictions
```

```
## # Predicted values of weight
##
## length | Predicted |         95% CI
## ----------------------------------
```

```
##     50 |     -1.71 | [-2.12, -1.30]
##     60 |      0.29 | [-0.02,  0.59]
##     65 |      1.29 | [ 1.03,  1.54]
##     75 |      3.29 | [ 3.12,  3.45]
##     85 |      5.28 | [ 5.16,  5.41]
##     95 |      7.28 | [ 7.12,  7.44]
##    105 |      9.28 | [ 9.04,  9.53]
##    120 |     12.28 | [11.88, 12.68]
```

```
# visualization code
plot_predictions <- ggplot(data = fish_data, aes(x = length, y = weight)) +
    geom_point() + # add points
    geom_line(data = predictions,
              aes(x = x, y = predicted),
              color = 'lightblue',
              linewidth = 1) + # add regression line
    geom_ribbon(data = predictions, aes(x = x, # add confid. interval
                                        y = predicted,
                                        ymin = conf.low,
                                        ymax = conf.high),
                alpha = 0.2) +
    theme_classic() + # add theme
    labs(x = 'Fish Length (mm)', y = 'Fish Weight (g)', # add labels and caption
         title = 'Fish Length as a Predictor of Fish Weight',
         caption = "Figure 3: Fish lengths and weights against predicted values, shown with confidence i
    theme(plot.caption = element_text(hjust = 0), # adjustments
          text = element_text(family = 'Helvetica'))

plot_predictions
```

```
## Warning: Removed 199 rows containing missing values (‘geom_point()‘).
```

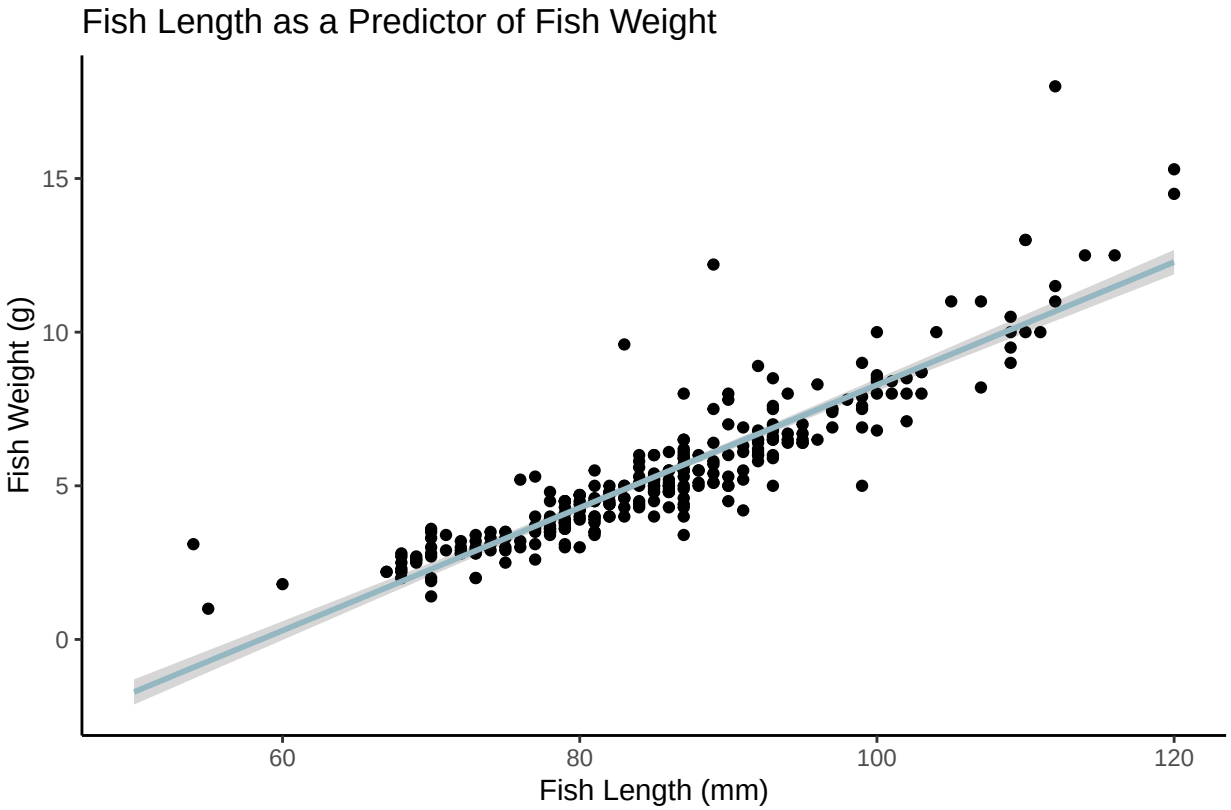

Figure 3: Fish lengths and weights against predicted values, shown with confidence interval.

Link to repo:
https://github.com/lukegtodd/ENVS-193DS_homework-04_todd-luke