Weight and Sleeping Patterns in the UK

John Karuitha

Monday, August 08, 2022

Contents

Background

With sedentary lifestyles, obesity has become a significant health issue globally. Overweight individuals have a higher risk of developing heart disease, stroke, cancer, kidney disease, among others (Shah et al. 2021). These health issues place additional strain on health facilities and state financial resources. Consequently, much research goes into tracking obesity, mapping out possible health complications associated with obesity, establishing the factors contributing to obesity. Critically, many resources go to the design of measures to minimise the prevalence of obesity (Fruh 2017; Malik, Willett, and Hu 2013; Lopez 2007).

In this project, we explore the link between age, sleep, and body mass index in a sample of individuals from the United Kingdom.

```
## Read in the data
weight_data <- read_csv("dataset.csv") %>%

## Clean names by removing special characters and capital letters
  janitor::clean_names() %>%

  mutate(sleep = factor(sleep,

                        ordered = TRUE,

                        levels = 5:10,

                        labels = c("five", "six", "seven", "eight", "nine", "ten")))
```

Objective

The broad objective of the project is to explore the relationship between age, sleeping patterns and the weight of individuals, as captured using the body mass index (BMI). Specifically, the study seeks answers to the following questions.

Summary of Results

1. Age and sleep are the significant drivers of BMI.
2. Older people tend to have higher BMI and a higher risk of diabetes due to obesity.
3. People who sleep fewer hours tend to have high BMI.
4. People who sleep 5 hours or less or sleep are likely to have higher BMI than average.
5. Older, sleep defficient people are at an even higher risk of diabetes due to obesity.

Data Description and Validation

In this section, we briefly describe the data, including the steps taken to validate the data.

The general structure of the dataset

The data consists of 4 variables. The first variable is the ID which serves as an identifier of the respondents. The second is age in years that has been discretized by rounding off to the nearest whole number. similarly, the third variable, sleep in hours, has been discretized into a whole number. The last variable is the body mass index, the ratio of weight in kilograms to height in metres squared. Further there are 153 observations. The following two tables confirm the structure of the data.

```
##Get the first six rows of the dataset
head(weight_data) %>%

## Convert the row names to upper case
  set_names(names(.) %>% str_to_upper()) %>%

## Make a table
  knitr::kable(booktabs = TRUE,

## Insert table title
```

Table 1: The First Six Rows of the BMI-Sleep-Age Data set from the UK

ID	AGE	SLEEP	BMI
1	24	eight	23.6
2	26	seven	24.1
3	28	eight	25.3
4	29	six	26.7
5	33	eight	26.2
6	28	ten	25.4

```
    caption = "The First Six Rows of the BMI-Sleep-Age Data set from the UK") %>%

  kableExtra::kable_styling(full_width = TRUE,

                          bootstrap_options = "striped")
```

```
str(weight_data)
```

```
## tibble [153 x 4] (S3: tbl_df/tbl/data.frame)
##  $ id   : num [1:153] 1 2 3 4 5 6 7 8 9 10 ...
##  $ age  : num [1:153] 24 26 28 29 33 28 19 27 18 36 ...
##  $ sleep: Ord.factor w/ 6 levels "five"<"six"<"seven"<..: 4 3 4 2 4 6 2 2 2 4 ...
##  $ bmi  : num [1:153] 23.6 24.1 25.3 26.7 26.2 25.4 22.7 25.4 19.7 28.2 ...
```

Exploring the data

In this section we explore the data first by visualizing it and then computing some summary statistics. The central hypothesis in this case is whether there is a relationship on the one side, and age and hours of sleep among the respondents on the other.

Data validation

To validate the data, I check for two issues.

- Missing values.
- Unreasonable observations and extreme values.
- Duplicates or repeated values.

The data has no missing values, as the table below shows.

```
sapply(weight_data, is.na) %>%

  colSums() %>%

  tibble(Variables = names(weight_data), missing_values = .) %>%

  mutate(Variables = (Variables %>% str_to_upper())) %>%
```

Table 2: Missing Values

Variables	missing_values
ID	0
AGE	0
SLEEP	0
BMI	0

```
  knitr::kable(booktabs = TRUE, caption = "Missing Values") %>%

  kableExtra::kable_styling(full_width = TRUE,

                            bootstrap_options = "striped")
```

Next, I generate a summary of the data and inspect it for unreasonable values. For instance, we do not expect that an individual sleeps for more than 24 hours in a day. Typically, a value that lies too far off the average sleep hours and age in the dataset is a potential error or otherwise an outlier.

```
summary(weight_data)
```

```
##        id              age          sleep          bmi
##  Min.   :  1.0   Min.   :18.0   five :11   Min.   :18.0
##  1st Qu.: 36.0   1st Qu.:22.0   six  :40   1st Qu.:24.4
##  Median : 71.0   Median :28.0   seven:24   Median :26.1
##  Mean   : 72.1   Mean   :27.7   eight:46   Mean   :26.8
##  3rd Qu.:109.0   3rd Qu.:33.0   nine :24   3rd Qu.:29.6
##  Max.   :147.0   Max.   :39.0   ten  : 8   Max.   :36.5
```

The ID column shows 147 observations, yet there are 153 rows in the dataset. Thus the ID column has a discrepancy. Otherwise, there are repeated individuals in the dataset.

```
weight_data %>%

  filter(duplicated(.)) %>%

  knitr::kable(booktabs = TRUE, caption = "Duplicate Values") %>%

  kableExtra::kable_styling(full_width = TRUE,

                            bootstrap_options = "striped")
```

It appears that individuals coded by IDs 10, 11, 12, 41, 42, and 43 appear twice in the dataset. These 6 extra observations are the source of the discrepancy.

We examine whether the repeated observations are duplicates. It turns out all are duplicates as the table below shows.

```
weight_data %>%

  filter(id %in% c(10, 11, 12, 41, 42, 43)) %>%
```

Table 3: Duplicate Values

id	age	sleep	bmi
10	36	eight	28.2
11	32	six	31.1
12	18	eight	22.0
41	34	nine	24.8
42	33	seven	34.8
43	18	eight	24.7

Table 4: Duplicate Values

id	age	sleep	bmi
10	36	eight	28.2
10	36	eight	28.2
11	32	six	31.1
11	32	six	31.1
12	18	eight	22.0
12	18	eight	22.0
41	34	nine	24.8
41	34	nine	24.8
42	33	seven	34.8
42	33	seven	34.8
43	18	eight	24.7
43	18	eight	24.7

```
  arrange(id) %>%

  knitr::kable(booktabs = TRUE,

               caption = "Duplicate Values") %>%

  kableExtra::kable_styling(full_width = TRUE,

                  bootstrap_options = "striped")
```

We clean the data by removing all the duplicates to get a clean dataset.

```
weight_data <- weight_data %>%

  filter(!duplicated(.))
```

Now we have a clean dataset with 147 observations. In the next section, we visualize the data.

Descriptive Statistics

The table blow shows the summary statistics for the variables, except for the ID.

Table 5: Summary Statistics

Variable	Missing	factor.ordered	factor.n_unique	factor.top_counts	Mean	SD	Min	Q1	Median	Q3	Max
SLEEP	0	TRUE	6	eig: 43, six: 39, sev: 23, nin: 23	NA	NA	NA	NA	NA	NA	NA
AGE	0	NA	NA	NA	27.68	6.215	18	22.00	28.0	33.0	39.0
BMI	0	NA	NA	NA	26.72	3.697	18	24.25	26.1	29.5	36.5

```
weight_data %>%

  select(-id) %>%

  skimr::skim_without_charts() %>%

  select(-complete_rate, -skim_type) %>%

  rename(Variable = skim_variable, Mean = numeric.mean,

         SD = numeric.sd, Missing = n_missing,

         Min = numeric.p0, Q1 = numeric.p25, Median = numeric.p50,

         Q3 = numeric.p75, Max = numeric.p100) %>%

  mutate(Variable = (Variable %>% str_to_upper())) %>%

   knitr::kable(booktabs = TRUE, caption = "Summary Statistics") %>%

  kableExtra::kable_styling(full_width = TRUE,

                            bootstrap_options = "striped")
```

Next, I convert sleep to an ordinal categorical variable.

Next, I then Produce make a frequency table for the variable “Sleep”.

```
table(weight_data$sleep) %>%

  knitr::kable(booktabs = TRUE, caption = "Frequency Table for Sleep") %>%

  kableExtra::kable_styling(full_width = TRUE,

                            bootstrap_options = "striped")
```

Table 6: Frequency Table for Sleep

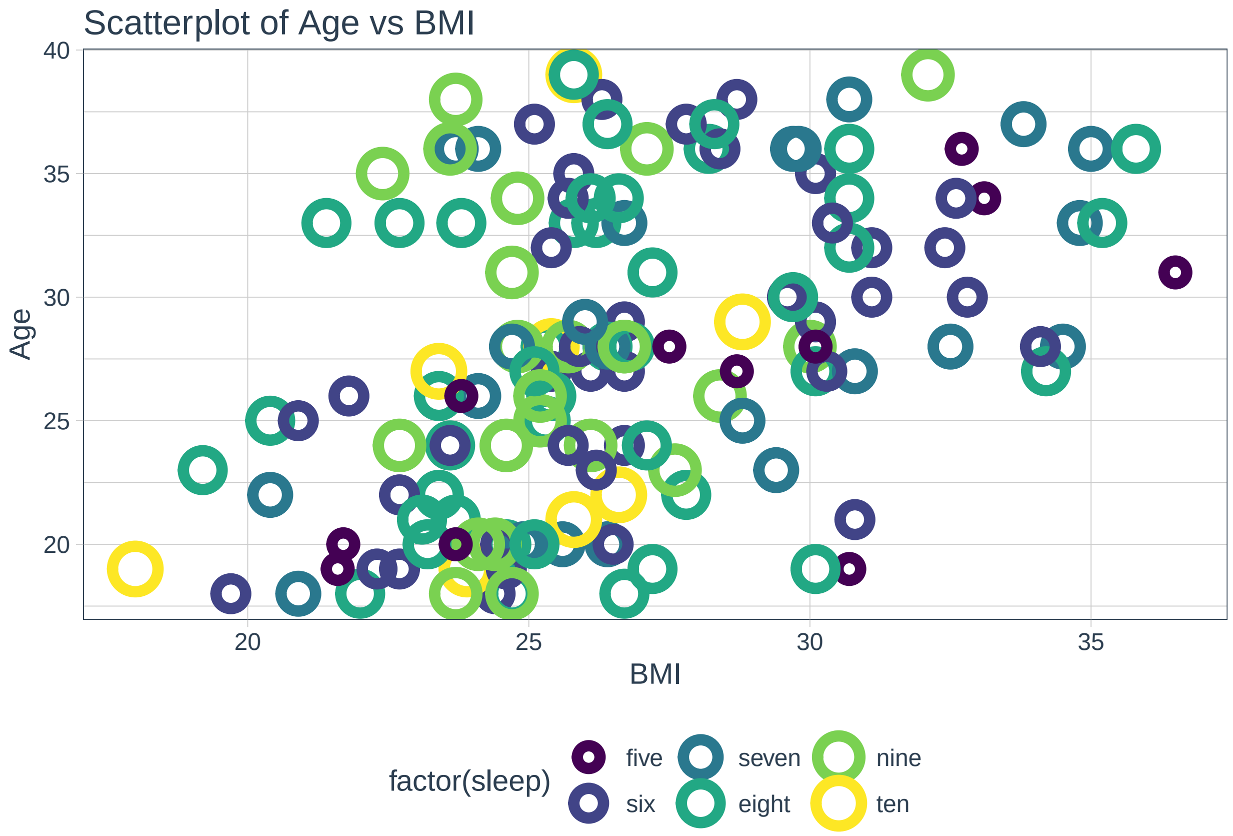
Var1	Freq
five	11
six	39
seven	23
eight	43
nine	23
ten	8

Data Visualization

Age vs BMI Produce a scatter plot using ggplot2 package for the following:

- Age vs. BMI, adjust the transparency (alpha) for the points to 0.5.
- Use white background theme (theme_bw).
- Add title for the graph and coordinates.
- Describe the relationship between the two variables.

```
weight_data %>%

  ggplot(aes(x = bmi, y = age,

             col = factor(sleep), size = factor(sleep))) +

  geom_point(shape = 1, stroke = 3) +

  labs(x = "BMI", y = "Age",

       title = "Scatterplot of Age vs BMI") +

  scale_color_viridis_d()
```

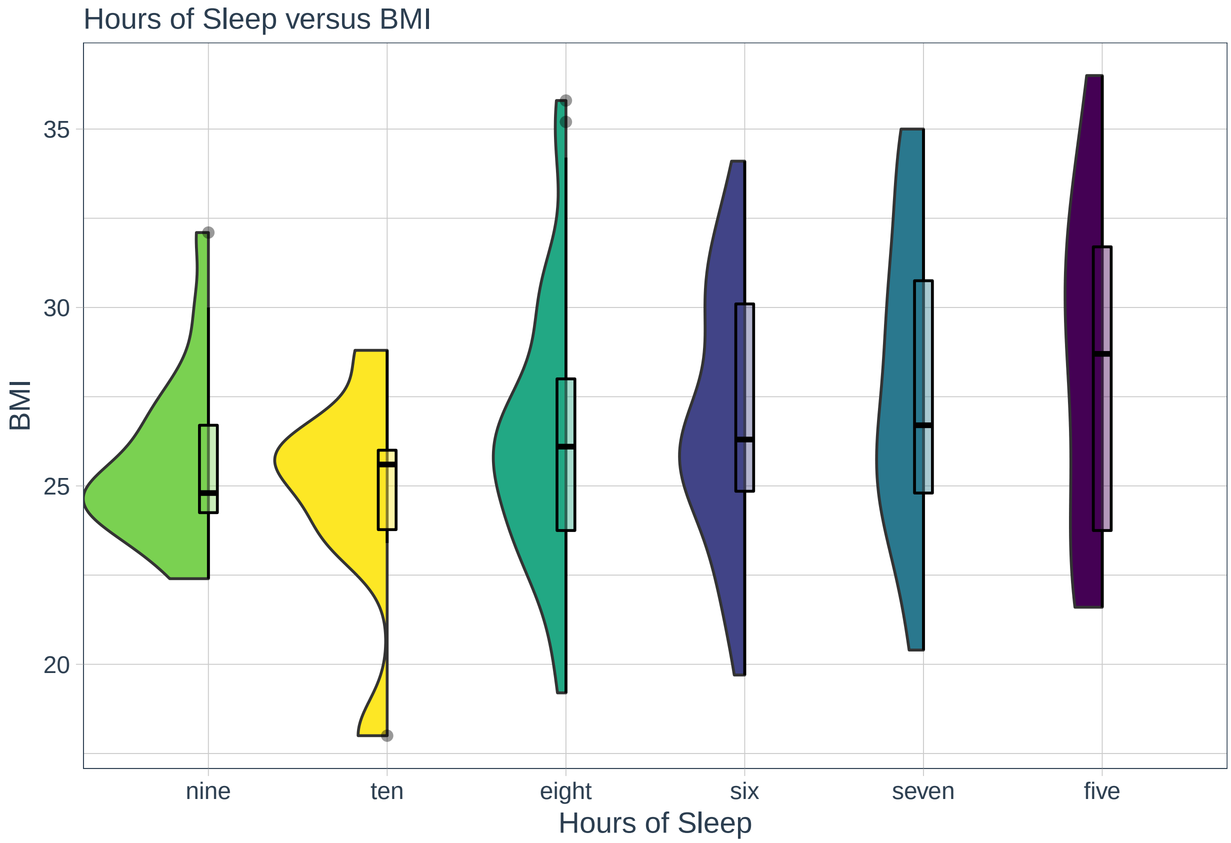


There appears to be a substantial positive correlation between age and BMI, with older people more likely to have higher BMI.

BMI vs Sleep Produce a boxplot using ggplot2 package for the following:

- BMI factored over the variable Sleep.
- Use white background theme.
- Add title for the graph and coordinates.
- Describe the graph provided.

```
weight_data %>%

  ggplot( aes(x = fct_reorder(sleep, bmi, median), y = bmi, fill = sleep)) +

    geom_half_violin(width=1.4) +

    geom_boxplot(width=0.1, color="black", alpha=0.4) +

    scale_fill_viridis_d() +

    theme(
      legend.position="none",
      plot.title = element_text(size=11)
    ) +
    ggtitle("Hours of Sleep versus BMI") +

    xlab("Hours of Sleep") +

    ylab("BMI")
```

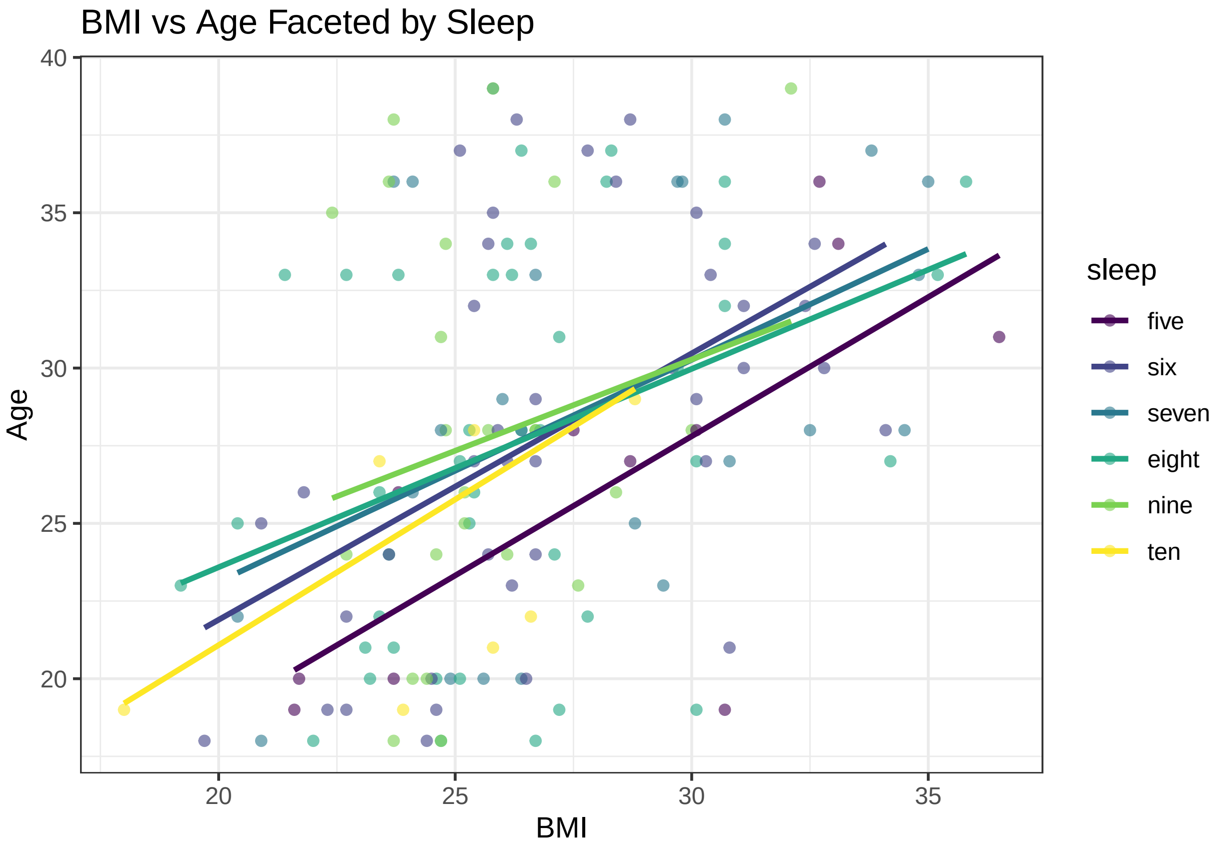


More hours of sleep appear to correlate with lower levels of BMI. Individuals who sleep 5 to 7 hours of sleep have a higher likelihood of high BMI.

Age vs BMI Produce a scatter plot using ggplot2 using the following:

- Age vs. BMI factored over the variable Sleep.
- Adjust transparency to 0.6.
- Add linear models to graphs without the standard error area.
- Use white background.
- Add title for the graph and coordinates.
- Describe the graph produced.

```
weight_data %>%

  ggplot(aes(x = bmi, y = age, col = sleep)) +

  geom_point(alpha = 0.6) +
  geom_smooth(se = FALSE, method = "lm") +
  labs(x = "BMI", y = "Age",
       title = "BMI vs Age Faceted by Sleep") +
  theme_bw()
```



Although there is a relationship between age and BMI, the effect varies by sleep habits. For people who sleep 10 hours and those that sleep 5 hours, there is a notably higher chance of having a greater BMI.

Regression Analysis

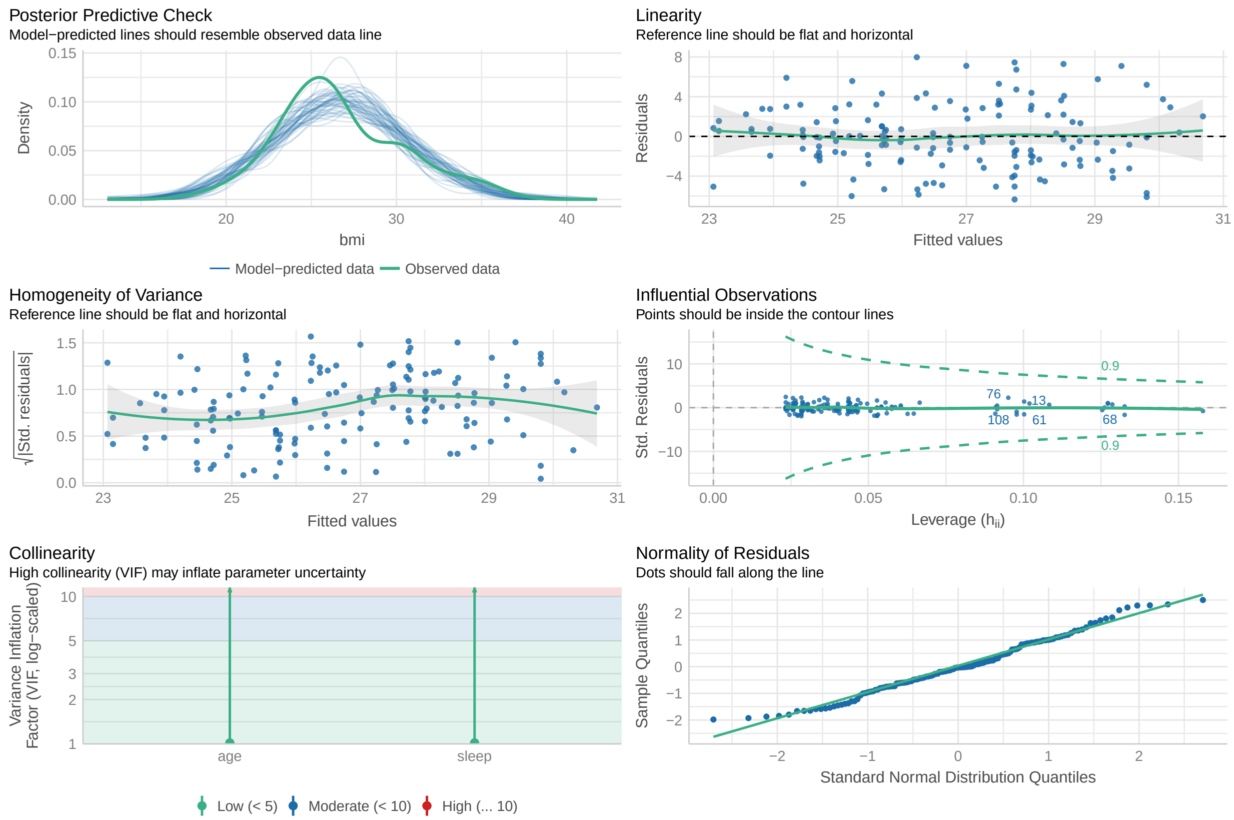
```
weight_lm <- lm(bmi ~ age + sleep, data = weight_data)
summary(weight_lm)
```

```
##
## Call:
## lm(formula = bmi ~ age + sleep, data = weight_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.350 -2.024 -0.106  2.209  7.971
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  19.7244     1.2472   15.82  < 2e-16 ***
## age           0.2536     0.0441    5.75  5.5e-08 ***
## sleep.L      -2.6011     0.9692   -2.68   0.0082 **
## sleep.Q       0.0291     0.9232    0.03   0.9749
## sleep.C      -0.1606     0.7687   -0.21   0.8348
## sleep^4       0.8435     0.6529    1.29   0.1985
## sleep^5      -0.6103     0.6084   -1.00   0.3175
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.29 on 140 degrees of freedom
## Multiple R-squared:  0.243,  Adjusted R-squared:  0.21
## F-statistic: 7.48 on 6 and 140 DF,  p-value: 5.76e-07
```

Regression Diagnostics

The regression diagnostics results do not indicate a poor fit. Hence, we can rely on the model for fairly accurate predictions. However, there is not enough data to do an out of sample check of accuracy.

```
performance::check_model(weight_lm)
```



References

Fruh, Sharon M. 2017. "Obesity: Risk Factors, Complications, and Strategies for Sustainable Long-Term Weight Management." *Journal of the American Association of Nurse Practitioners* 29 (S1): S3–14.

Lopez, Russ P. 2007. "Neighborhood Risk Factors for Obesity." *Obesity* 15 (8): 2111–19.

Malik, Vasanti S, Walter C Willett, and Frank B Hu. 2013. "Global Obesity: Trends, Risk Factors and Policy Implications." *Nature Reviews Endocrinology* 9 (1): 13–27.

Shah, Paras P, Tammy M Brady, Kevin EC Meyers, Michelle M O'Shaughnessy, Keisha L Gibson, Tarak Srivastava, Jarcy Zee, et al. 2021. "Association of Obesity with Cardiovascular Risk Factors and Kidney Disease Outcomes in Primary Proteinuric Glomerulopathies." *Nephron* 145 (3): 245–55.