Assignment_One

Oluwanifemi

2024-09-19

```r
He <- "C:/Users/modim/Downloads/Regression Methods/RegressionMethods/Health_Sleep_Statistics.csv"
data <- read.csv(He)
#View(Health)

#library(ggplot2)
#library(stats)
#file <- "C:/Users/modim/Downloads/RegressionMethods/Health_Sleep_Statistis.csv"
#Health_Sleep_Statistics <- read.csv("Health_Sleep_Statistis.csv", header=T)
```

Question One (A)

2.2.1 The main difference is that the points above the line y=x are cities where the price of rice in the year 2009 was higher than that of 2003, while the points below the line are cities where the price of rice in 2009 are lower than that of 2003.

2.2.2 a) Vilnius, it had the highest increase b) Mumbai, it had the lowest decrease

2.2.3 It does not necessarily show that prices are lower in 2009 in comparison to 2003 due to the fact most of the points lie above the y=x line suggesting a price increase 2009 compared to 2003. If $\beta_1 = 1$, this means that a 1% increase in 2003 prices leads to exactly a 1% increase in 2009 prices. but $\beta_1 < 1$, which implies that the effect is less than proportional—i.e., a 1% increase in the 2003 price leads to less than a 1% increase in the 2009 price as we can see in the graph.

2.2.4 Secondly, 2003 rice price is not the only predictor for 2009's price of rice, as they are other factors like recession, the absence of these variables will make the regression equation erroneous.

Question One (B)

2.3.1 Looking at the graph, it is evident that using log-scale forces the data to be more clustered more tightly and evenly around the fitted ols line as it gives a better fit in comparison to the original scale. Log-scale brings ols line and y=x line much closer to each other.

2.3.2 β_0 is the expected value of the log of price in 2009 (y) when the log of price in 2003 (x) is zero.

β_1 is interpreted as: for every 1% increase in the price of rice in 2003, the average increase in the price of rice in 2009 is β_1%.

Question Three

URL:

https://www.kaggle.com/datasets/hanaksoy/health-and-sleep-statistics

This data set contains various information about individuals' sleep habits and physical activities. The data provides important indicators of individuals' overall health and quality of life. Below is detailed information about the columns in the data set and their contents:

User ID: An individual's unique identification number.

Age: The age of the individual.

Gender: The sex of the individual ('f' female, 'm' male)

Sleep Quality: The quality of an individual's sleep (a scale of 1-10, with 10 indicating the highest quality)

Bedtime: The individual's bedtime (in 24-hour format)

Wake-up Time: The individual's wake-up time (in 24-hour format)

Daily Steps: Number of steps per day

Calories Burned: The amount of calories burned per day Physical Activity Level: The individual's physical activity level (low, medium, high)

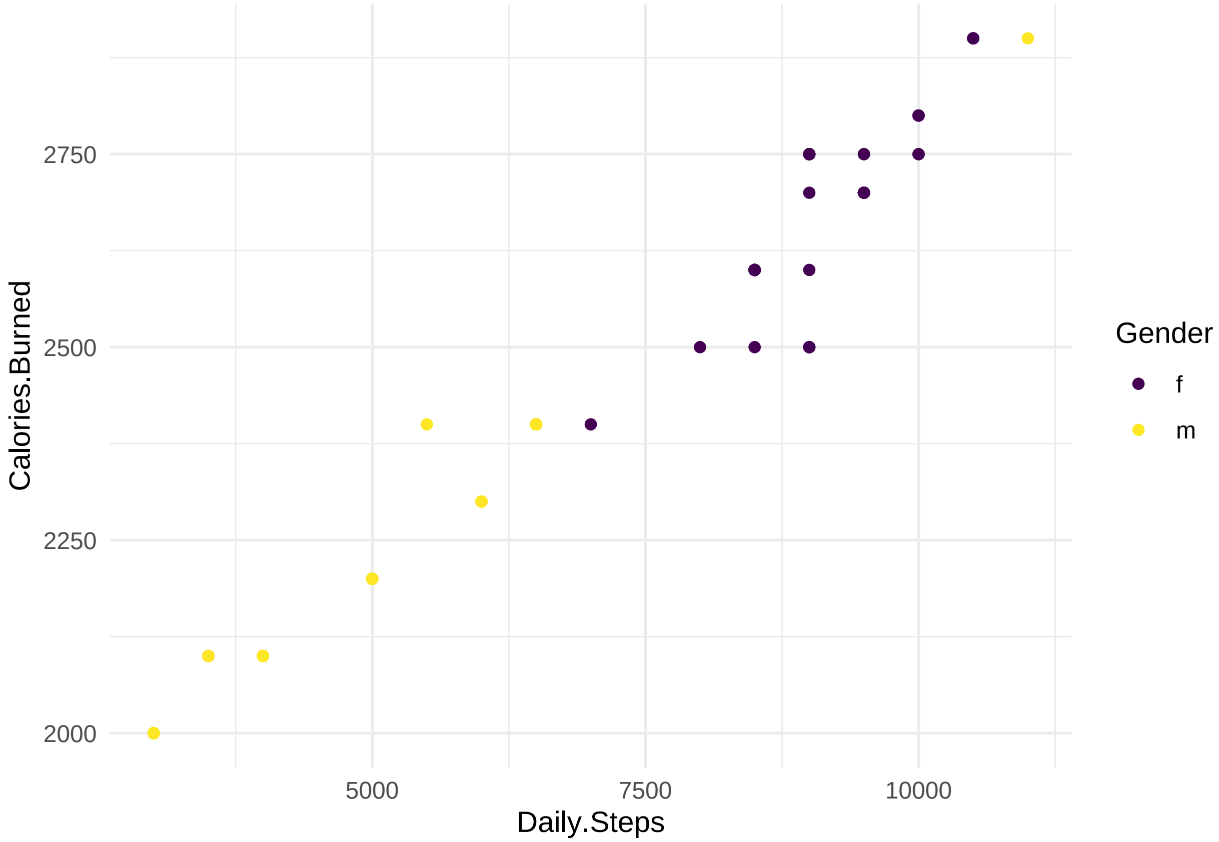
Dietary Habits: Dietary habits of the individual (healthy, medium, unhealthy)

Sleep Disorders: Whether the individual has sleep disorders (yes, no)

Medication Usage: Whether the individual uses medication for sleep disorders (yes, no)

Explanation: These data are imaginary data. It was created entirely for the purpose of improving users, it has nothing to do with reality.

```r
#This data set looks at the relationship of sleep quality with other variables
#like Calories burned, Daily Steps, etc
ggplot(data, aes(x = Daily.Steps, y = Calories.Burned)) +
  geom_point(aes(color = Gender))+
  scale_color_viridis_d() +theme_minimal()
```



```
#There is a linear relationship between
#both quantitative variables,
#it is a positive linear relationship
```

There appears to be a positive relationship between Daily Steps and Calories Burned. As the number of daily steps increases, the calories burned also increase for both genders. The data points suggest that the relationship could be linear since the points follow an upward trend, especially for larger numbers of steps. A linear model seems appropriate, as it looks semi-linear, but it seems like gender also plays a factor so making it a multiple regression with gender might make it more appropriate.

A random question:

Given that an individual takes 8000 steps daily, how many calories would they burn?

```
lin<- lm(formula = Daily.Steps~Calories.Burned, data = data)

summary(lin)
```

```
##
## Call:
## lm(formula = Daily.Steps ~ Calories.Burned, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1145.78  -145.78    42.83   108.67  1476.99
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1.441e+04  3.551e+02  -40.57   <2e-16 ***
## Calories.Burned  8.772e+00  1.457e-01   60.21   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 407.5 on 98 degrees of freedom
## Multiple R-squared:  0.9737, Adjusted R-squared:  0.9734
## F-statistic:  3625 on 1 and 98 DF,  p-value: < 2.2e-16
```

From the linear regression model, the estimated coefficients are:

- **Intercept** (β_0): -14,410
- **Slope** (β_1) **for Calories Burned**: 8.772

Reporting the Linear Regression Results

Interpretation of the Intercept:

The intercept (β_0) is the estimated value of the Daily Steps when the Calories Burned is 0. This means that if 0 calories were burned, the model predicts that the daily steps would be around -14,410. This is not a meaningful interpretation since negative steps are impossible and burning zero calories is unlikely. The large negative intercept might indicate extrapolation beyond the data's range.

Interpretation of the Slope:

The slope (β_1) of 8.772 indicates that for every additional calorie burned, the model predicts an increase of approximately 8.772 steps. This suggests a positive linear relationship between the number of calories burned and daily steps.

Statistical Significance:

The slope is statistically significant with a p-value of less than 2×10^{-16}, suggesting that the relationship between calories burned and daily steps is not due to random chance.Please Note that I am setting the significance value used in this analysis as $\alpha = 0.05$.