602Project_part2

2024-02-13

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(mosaic)
```

```
## Registered S3 method overwritten by 'mosaic':
##   method                           from
##   fortify.SpatialPolygonsDataFrame ggplot2

##
## The 'mosaic' package masks several functions from core packages in order to add
## additional features.  The original behavior of these functions should not be affected by this.

##
## Attaching package: 'mosaic'

## The following object is masked from 'package:Matrix':
##
##     mean

## The following objects are masked from 'package:dplyr':
##
##     count, do, tally

## The following object is masked from 'package:ggplot2':
##
##     stat

## The following objects are masked from 'package:stats':
##
##     binom.test, cor, cor.test, cov, fivenum, IQR, median, prop.test,
##     quantile, sd, t.test, var
```

```
## The following objects are masked from 'package:base':
##
##     max, mean, min, prod, range, sample, sum
```

```
library(resampledata)
```

```
##
## Attaching package: 'resampledata'
```

```
## The following object is masked from 'package:datasets':
##
##     Titanic
```

```
library(olsrr)
```

```
##
## Attaching package: 'olsrr'
```

```
## The following object is masked from 'package:datasets':
##
##     rivers
```

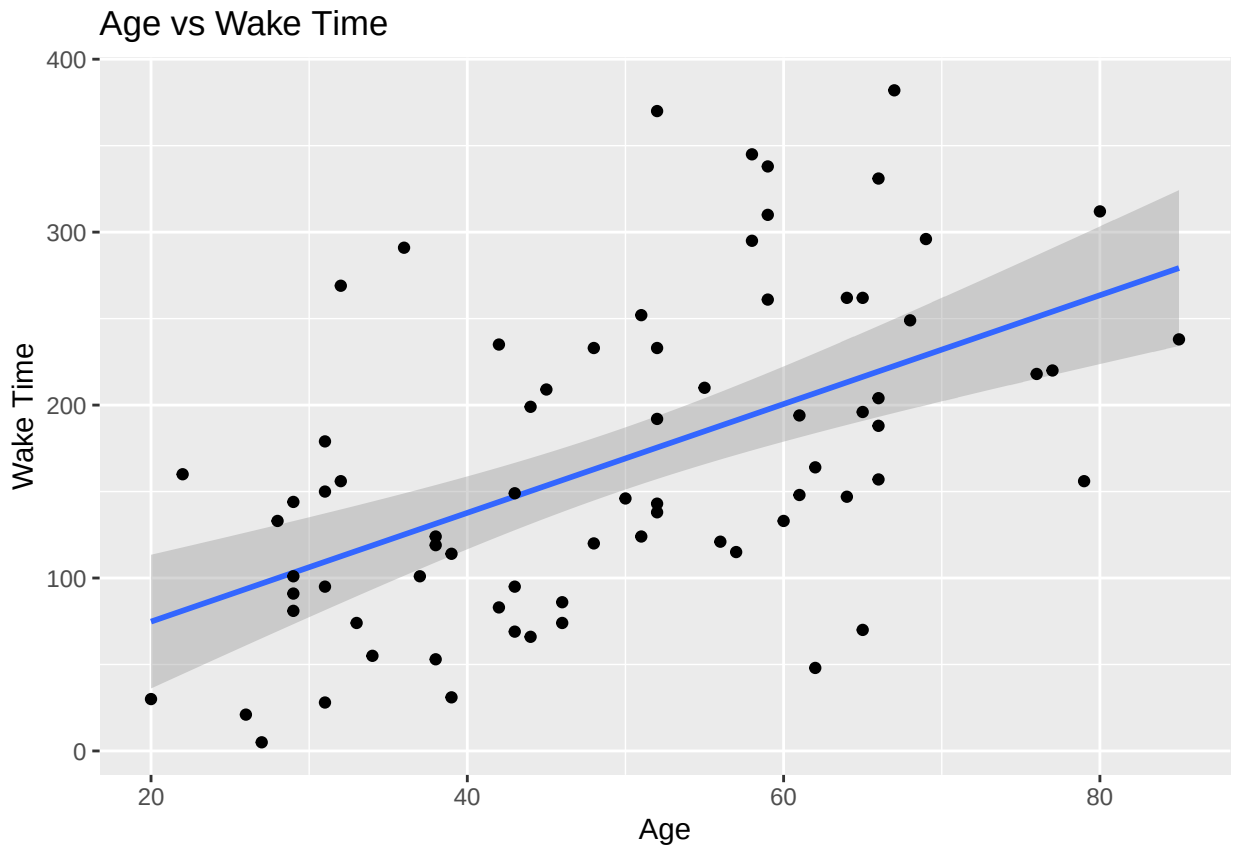
Part1: ***Read data***:

```
# Read data
data_s1 = read.csv("summary_data.csv")
data_stages = read.csv("sleep_stage_output.csv")

data = c(data_s1, data_stages)

data = data.frame(data)

data = filter(data, Age>0)
data = filter(data, W < 400)
```

Build Model:

```
ggplot(data, aes(x = Age, y = W)) + geom_smooth(method = "lm") + geom_point() + xlab("Age") + ylab("Wak
```

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




```
predicted_rate = lm(W ~ Age, data=data)
predicted_rate$coef
```

```
## (Intercept)         Age
##   11.884170    3.145238
```

The linear regression equation representing the model is $\hat{W}_i = 11.884170 + 3.145238 * Age_i$. ***Correlation Coefficient Check***:

```
age <- data$Age
wake <- data$W
cor(age, wake)
```

```
## [1] 0.5364623
```

Correlation Coefficient shows that there is a strong positive correlation between the age and amount of time spent awake during sleep.

Check significance of coefficient estimates Null hypothesis: $H_0 : \beta_1 = 0$

Alternative hypothesis: $H_A : \beta_1 \neq 0$

We will set the alpha value to 0.05.

We can use a t test to check our claim.

```r
summary(predicted_rate)
```

```
##
## Call:
## lm(formula = W ~ Age, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -158.89  -59.67  -17.90   48.80  194.56
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.8842    30.2154   0.393    0.695
## Age           3.1452     0.5831   5.394 8.37e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 77.52 on 72 degrees of freedom
## Multiple R-squared:  0.2878, Adjusted R-squared:  0.2779
## F-statistic: 29.09 on 1 and 72 DF,  p-value: 8.367e-07
```

From our t-test, we get test statistics of β_1 is 5.394. P-values of β_1 is $8.37 * 10^-7$. The p-value of β_1 is smaller than the set alpha value of 0.05, so we reject our null hypothesis that the linear regression coefficient is 0. Therefore, we can conclude that the time spent awake during sleep can be expressed as a linear function of age, and since $\beta_1 > 0$, we can say it is also positive.

We also get an R-squared value of 0.2878, meaning our independent variable explains approx. 28.78% of the variance in the dependent variable. A likely cause of the majority o this variance is the presence of various sleep-related illnesses. The data set used in this analysis was from subjects that had at least one sleep-related illness. Given that these illnesses are categorical data where each individual has at least one of several different illnesses, it is beyond the scope of this analysis to control for these factors.

Residual Analysis:

There are two conditions that must be met for our linear regression model to be valid.

1. Normality of residuals: The dependent variable (wake time) must be normally distributed with a mean of μ and standard deviations of σ. To check this we will plot a stat_qq plot of the residuals since $e_i = y_i - \hat{y}_i$, if y is normally distributed, so will the residuals. We will also do a Shapiro test to test for normality.

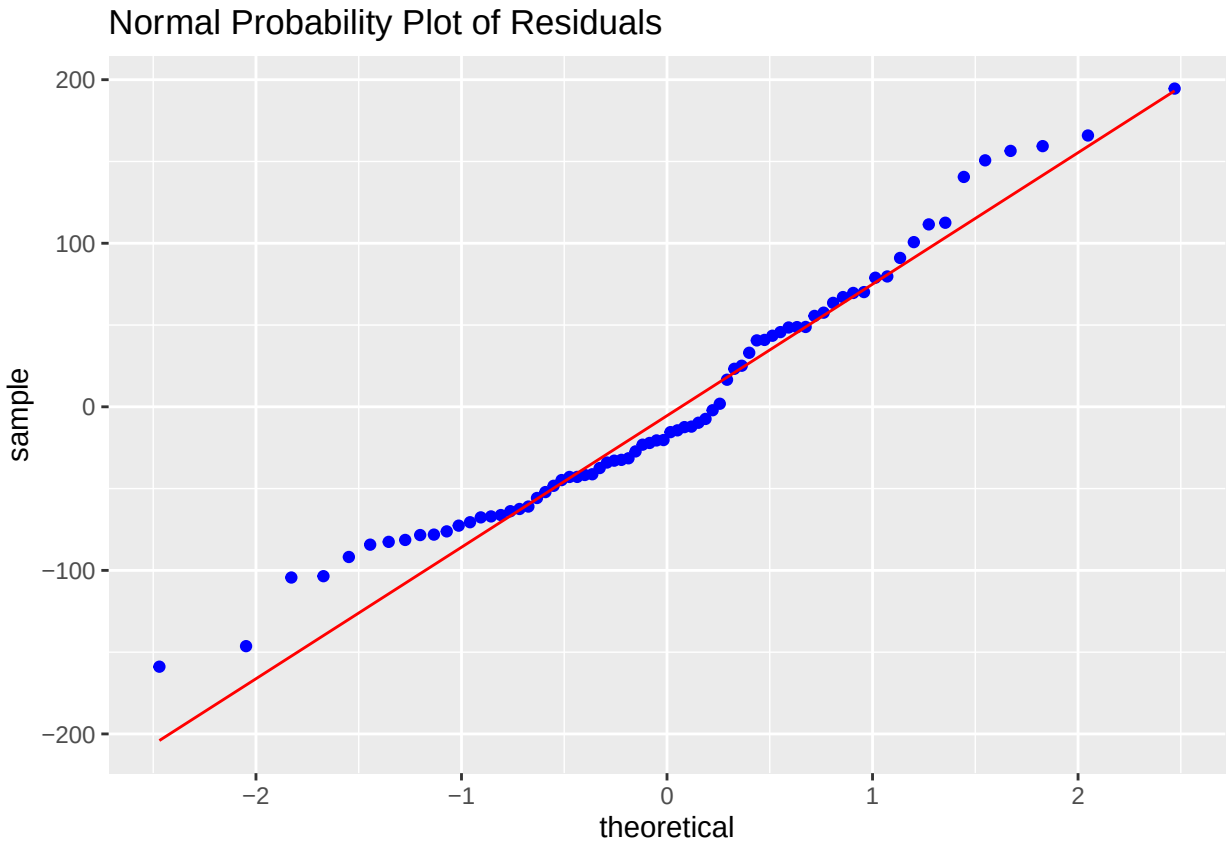
2. Homoscedasticity: For each distinct value of the independent variable (age), the dependent variable (wake time) has the same standard deviation σ. To check this, we will plot a scatter plot of the fitted values and the residuals. We will also perform a Breusch pagan test to test for homoscedasticity.

```r
# Get the and residuals fitted values
predicted.rate = predicted_rate$fitted.values
ei_hrat = predicted_rate$residuals
data.df = data.frame(predicted.rate, ei_hrat)
```

Normality of Residuals Plot:

```r
ggplot(data.df, aes(sample = ei_hrat)) +  stat_qq(col='blue') + stat_qqline(col='red') + ggtitle("Normal
```

```
## Warning: The following aesthetics were dropped during statistical transformation: sample
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a 'group' aesthetic or to convert a numerical
##   variable into a factor?
```



Looking the normality probability plot, the residuals do seem to be approximately normally distributed and therefore so is the dependent variable (wake time). The normality of residuals condition holds.

```
shapiro.test(data$Age)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$Age
## W = 0.97276, p-value = 0.11
```

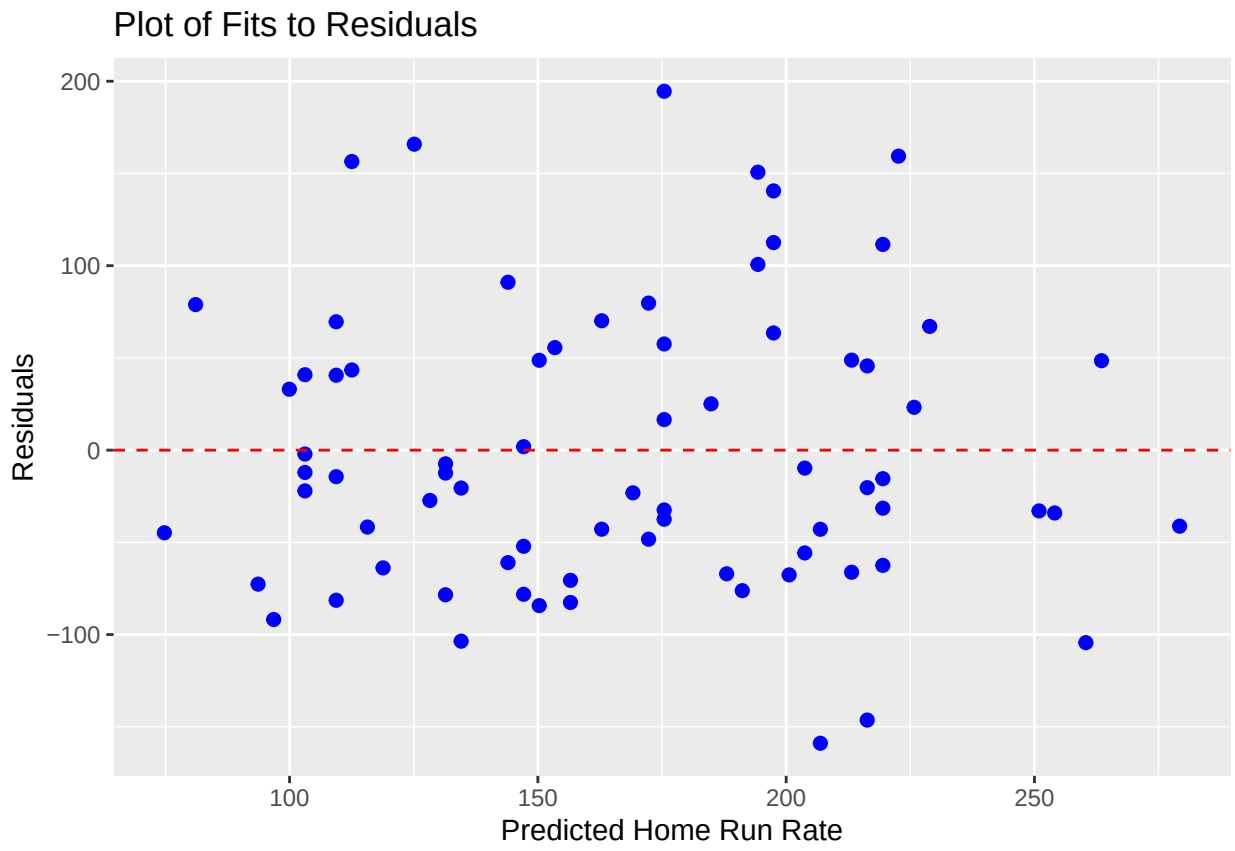
```
shapiro.test(data$W)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$W
## W = 0.97199, p-value = 0.09909
```

Given that the p-value for both age and wake time are greater than 0.05, we fail to reject the null hypothesis that the data is normally distributed. Therefore, the data passes the normality test.

Homoscedasticity:

```
ggplot(data.df, aes(x = predicted.rate, y = ei_hrat)) +  geom_point(size=2, col='blue', position="jitte
```



Looking at the plot of fits to residuals, the residuals do seem to be evenly distributed over wake time. We can they say that the condition of homoscedasticity holds.

```
ols_test_breusch_pagan(predicted_rate)
```

```
##
##  Breusch Pagan Test for Heteroskedasticity
##  -----------------------------------------
##  Ho: the variance is constant
##  Ha: the variance is not constant
##
##             Data
##  ----------------------------
##  Response : W
##  Variables: fitted values of W
##
##         Test Summary
##  ---------------------------
##  DF            =    1
```

```
##  Chi2          =    0.5701916
##  Prob > Chi2   =    0.4501828
```

Given the p-value (0.4501828) is greater than the threshold of 0.05, we fail to reject the null hypothesis that the data has constant variance. Therefore, the data passes the homoscedasticity test.

Since both conditions hold, our linear regression model is valid.