Simple Linear Regression Incomne/Happiness

Karl Mbouombouo

2023-08-21

Data

```
income_data <- read_csv("income.data.csv")
```

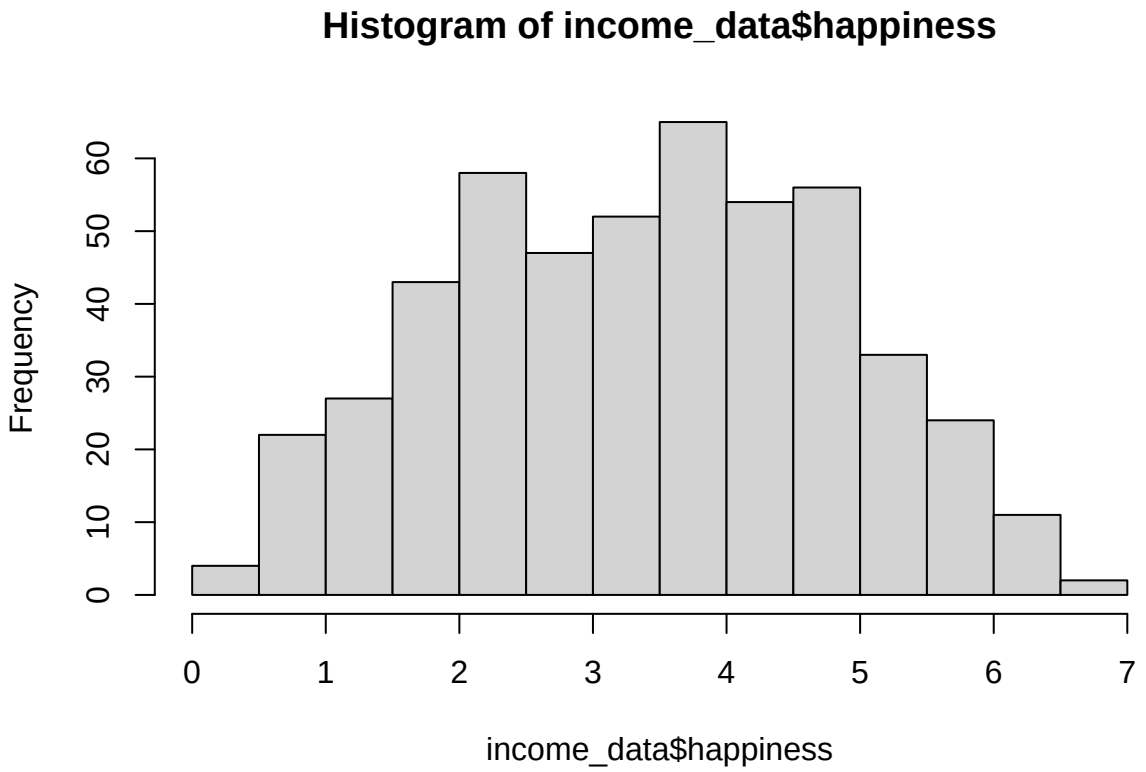
```
## New names:
## Rows: 498 Columns: 3
## -- Column specification
## -------------------------------------------------------- Delimiter: "," dbl
## (3): ...1, income, happiness
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```
summary(income_data)
```

```
##       ...1           income        happiness
##  Min.   :  1.0   Min.   :1.506   Min.   :0.266
##  1st Qu.:125.2   1st Qu.:3.006   1st Qu.:2.266
##  Median :249.5   Median :4.424   Median :3.473
##  Mean   :249.5   Mean   :4.467   Mean   :3.393
##  3rd Qu.:373.8   3rd Qu.:5.992   3rd Qu.:4.503
##  Max.   :498.0   Max.   :7.482   Max.   :6.863
```

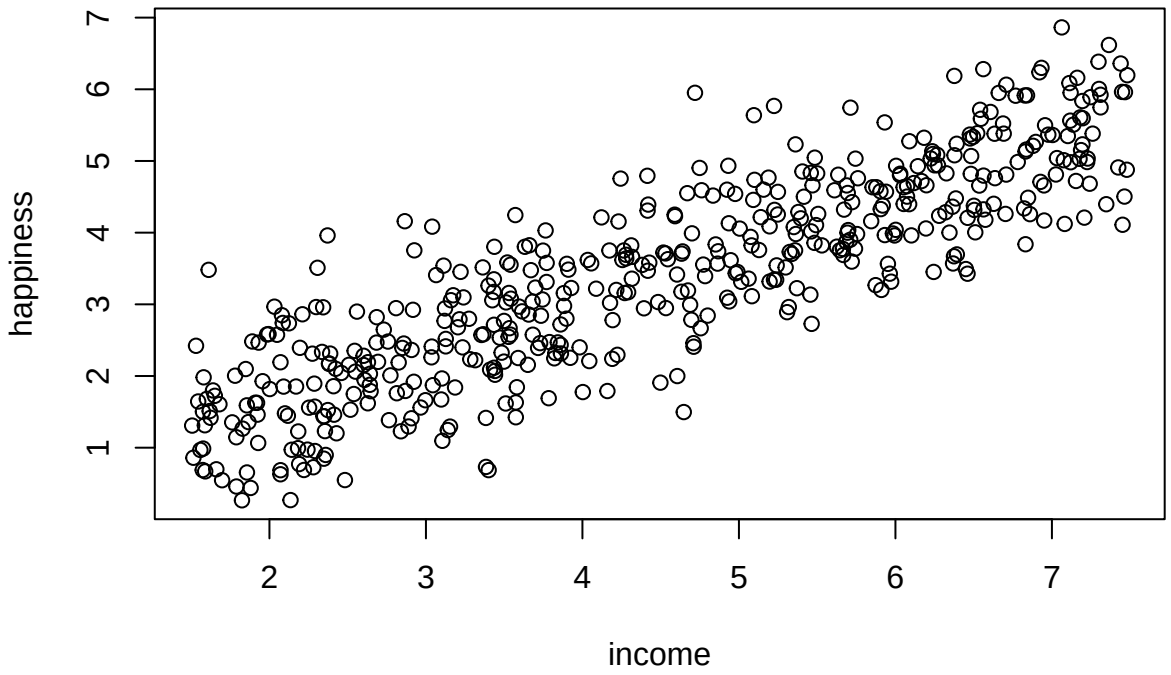
Plots

```
hist(income_data$happiness)
```



The observations are "bell-shaped", this is a good to proceed a linear model

```
plot(happiness ~ income, data = income_data)
```



The relationship between income and happiness look pretty linear, so we can use the linear model

Modeling

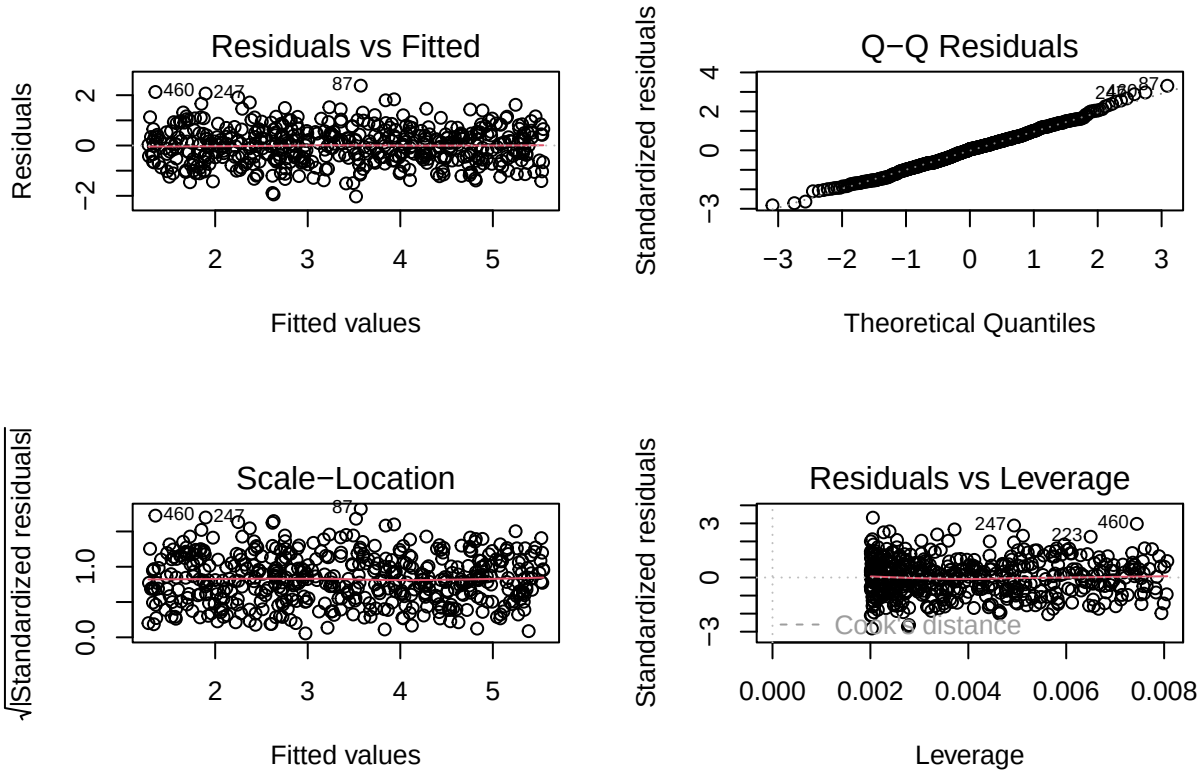
```
income.lm <- lm(happiness ~ income, data = income_data)
summary(income.lm)
```

```
##
## Call:
## lm(formula = happiness ~ income, data = income_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.02479 -0.48526  0.04078  0.45898  2.37805
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.20427    0.08884   2.299   0.0219 *
## income       0.71383    0.01854  38.505   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7181 on 496 degrees of freedom
## Multiple R-squared:  0.7493, Adjusted R-squared:  0.7488
## F-statistic:  1483 on 1 and 496 DF,  p-value: < 2.2e-16
```

From these results, we can say that there is a significant positive relationship between income and happiness (p value < 0.001), with a 0.71383 unit increase in happiness for every unit increase in income.

Check Homoscedasticity

```
par(mfrow=c(2,2))
plot(income.lm)
```




```
par(mfrow=c(1,1))
```

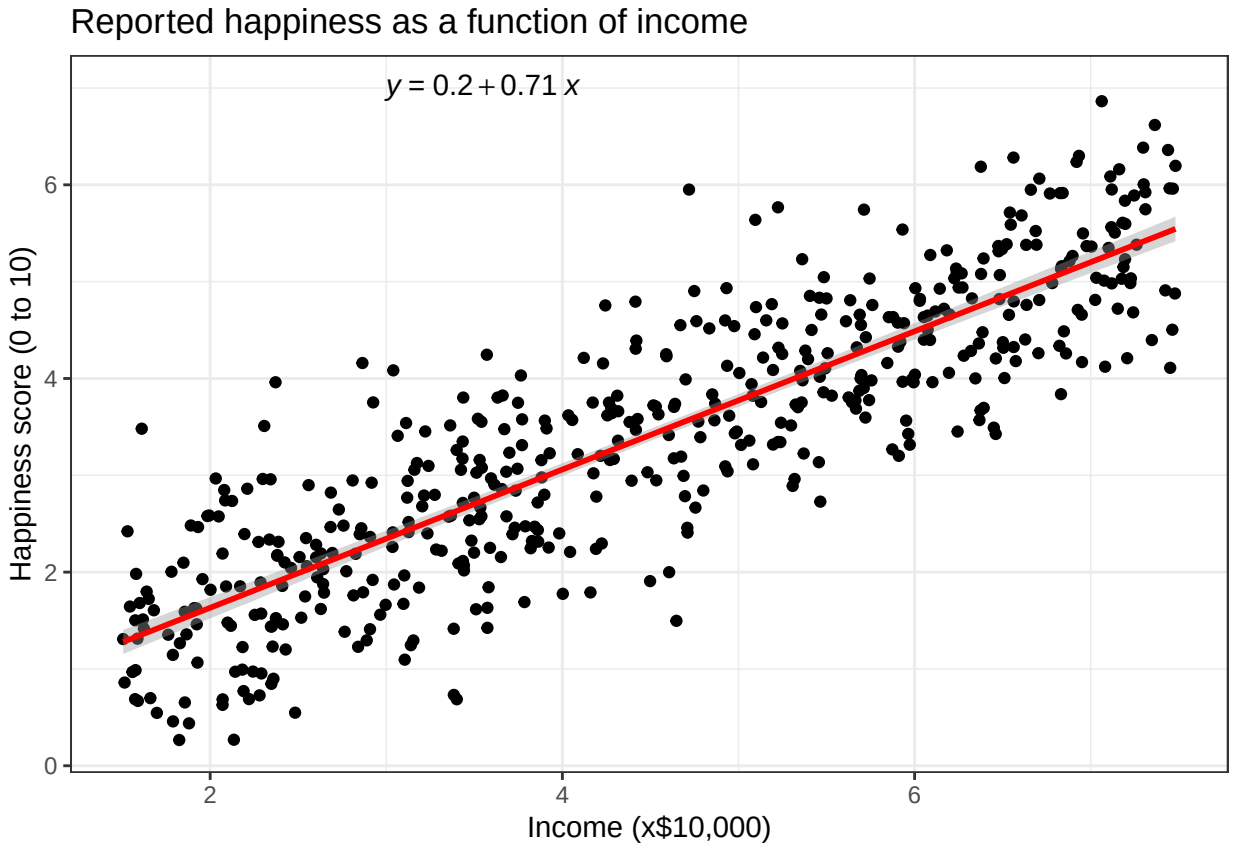
The residuals are for most of them all horizontal and centered to zero, this is mean there is no biases or outliers in the data that would make the linear regression invalid

Visualization of result

```
income.graph<-ggplot(income_data, aes(x=income, y=happiness))+
                      geom_point() +
                      geom_smooth(method="lm", col="red") +
                      stat_regline_equation(label.x = 3, label.y = 7) +
                      theme_bw() +
                      labs(title = "Reported happiness as a function of income",
                      x = "Income (x$10,000)",
                      y = "Happiness score (0 to 10)")


income.graph
```

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



Conclusion

There is a significant relationship between income and happiness ($p < 0.001$, $R2 = 0.73 \pm 0.0193$), with a 0.73 unit increase in reported happiness for every $10,000 increase in income.