Simple Linear Regression

Arjun Mannem

2024-01-21

# Load necessary packages  
library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.3.2

## Warning: package 'ggplot2' was built under R version 4.3.2

## Warning: package 'dplyr' was built under R version 4.3.2

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.3 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.4.4 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(ggthemes)  
library(flextable)

## Warning: package 'flextable' was built under R version 4.3.2

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

library(corrr)

## Warning: package 'corrr' was built under R version 4.3.2

library(ggfortify)  
library(broom)

#### Set default theme settings for plots, and load a function to simplify table customization and creation using the code below.

# Set ggplot theme for visualizations  
theme\_set(ggthemes::theme\_few())  
  
# Set options for flextables  
set\_flextable\_defaults(na\_str = "NA")  
  
# Load function for printing tables nicely  
source("https://raw.githubusercontent.com/dilernia/STA323/main/Functions/make\_flex.R")

## Simple Linear Regression

a.Exploring monthly stock prices based on an economic indicator such as the gross domestic product (GDP) in U.S. dollars 📈

b.Analyzing the relationship between health outcomes of patients, e.g. blood pressure (mmHG), and the dosage level of a prescribed treatment in milligrams for lowering people’s blood-pressure ⚕️💊

**What are the response and predictor variables for examples a. and b. above?**

1. **Response Variable**: monthly stock prices

* **Predictor Variable**: gross domestic product (GDP) in U.S. dollars

1. **Response Variable**: blood pressure (mmHG)

* **Predictor Variable**: the dosage level of a prescribed treatment in milligrams

**For examples a. through d., would we expect the true slope 1 to be negative or positive?**

1. Positive
2. Negative
3. Negative
4. Positive

**Provide a statement of the assumed simple linear regression model for example d.**

The assumed or theoretical simple linear regression model is

y = β0 + β1\*x + ε,

where:

* y is the sales in U.S dollars
* β0 is the intercept
* β1 is the slope
* x is the amount spend on marketing/advertising in U.S. dollars
* ε is error / residual

**Suppose that a sample of data was collected for example d., and an estimated slope was calculated as βHat1=2. Interpret this value in the context of the problem.**

**Correct**: Here slope is 2. So for every 1 dollar increase in advertising, we expect sales to increase by 2 dollars, on average.

**Incorrect**: Here slope is 2. So for every 1 dollar increase in advertising, sales increases by 2 dollars, on average.

**Further suppose for example d. that an estimated intercept was calculated as βHat0=5000. Interpret this value in the context of the problem.**

The expected amount of sales given that we are spending nothing on advertising is $5000.

**Given estimates of βHat0=5000 and βHat1=2 for example d., what would be predict our company’s sales revenue if we invested $500 in advertising on social media platforms?**

Given x = 500, βHat0=5000 and βHat1=2

y = βHat0 + βHat1 \* x

y = 5000 + 2\*500 = 6000

Estimated sales of the company in this scenario would be $6000.

## Example: Scrabble Letters

# Importing data  
scrabble <- read\_csv("https://raw.githubusercontent.com/dilernia/STA323/main/Data/scrabble.csv")

## Rows: 26 Columns: 3  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (1): letter  
## dbl (2): perc\_rf\_english\_writing, scrabble\_value  
##   
## 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.

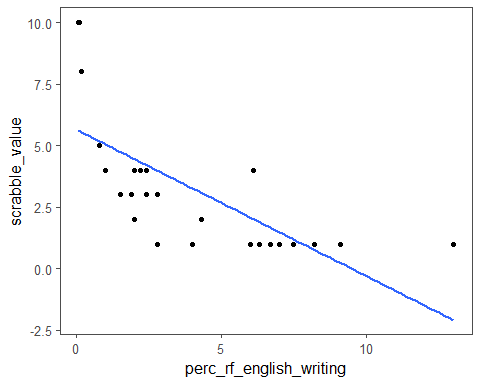
# Printing table of data  
scrabble %>%   
 make\_flex()

| letter | perc\_rf\_english\_writing | scrabble\_value |
| --- | --- | --- |
| A | 8.20 | 1.00 |
| B | 1.50 | 3.00 |
| C | 2.80 | 3.00 |
| D | 4.30 | 2.00 |
| E | 13.00 | 1.00 |
| F | 2.20 | 4.00 |
| G | 2.00 | 2.00 |
| H | 6.10 | 4.00 |
| I | 7.00 | 1.00 |
| J | 0.15 | 8.00 |
| K | 0.77 | 5.00 |
| L | 4.00 | 1.00 |
| M | 2.40 | 3.00 |
| N | 6.70 | 1.00 |
| O | 7.50 | 1.00 |
| P | 1.90 | 3.00 |
| Q | 0.10 | 10.00 |
| R | 6.00 | 1.00 |
| S | 6.30 | 1.00 |
| T | 9.10 | 1.00 |
| U | 2.80 | 1.00 |
| V | 0.98 | 4.00 |
| W | 2.40 | 4.00 |
| X | 0.15 | 8.00 |
| Y | 2.00 | 4.00 |
| Z | 0.07 | 10.00 |

#### Let’s look at a scatter plot and the correlations for the Scrabble data:

# Creating a scatter plot  
scrabble %>%   
 ggplot(aes(x = perc\_rf\_english\_writing,   
 y = scrabble\_value)) +   
 geom\_point() +   
 geom\_smooth(method = "lm", se = FALSE)

## `geom\_smooth()` using formula = 'y ~ x'



# Calculating correlations  
corTable <- scrabble %>%   
 corrr::correlate(diagonal = 1)

## Non-numeric variables removed from input: `letter`  
## Correlation computed with  
## • Method: 'pearson'  
## • Missing treated using: 'pairwise.complete.obs'

# Printing table of correlations  
corTable %>%   
 make\_flex(caption = "Table of pairwise correlations.")

Table 2: Table of pairwise correlations.

| term | perc\_rf\_english\_writing | scrabble\_value |
| --- | --- | --- |
| perc\_rf\_english\_writing | 1.00 | -0.71 |
| scrabble\_value | -0.71 | 1.00 |

**Does there appear to be a positive or negative relationship between the relative frequency of a letter in English writing (perc\_rf\_english\_writing), and its Scrabble tile point value (scrabble\_value)?**

Negative

**Does the sign (being positive or negative) of the correlation coefficient seem sensible in this context? Why or why not?**

Yes, the correlation being negative is sensible since a letter having higher percent relative frequency means it is easier to play in scrabble, so it should be worth less points.

**Does there appear to be a linear relationship between the relative frequency of a letter in English writing, and its Scrabble tile point value?**

No, the relationship is non-linear, so truly simple linear regression is inappropriate, but we will proceed for the sake of the problem.

**Provide a statement of the assumed simple linear regression model for the Scrabble data.**

The assumed or theoretical simple linear regression model is

y = β0 + β1\*x + ε,

where:

* y is the scrabble tile point value
* β0 is the intercept
* β1 is the slope
* x is the percent relative frequency of the letter in English writing
* ε is error / residual

#### Fit a simple linear regression model for the Scrabble data and print the model output using the code below.

# Fitting a simple linear regression model  
slrModel <- lm(scrabble\_value ~ perc\_rf\_english\_writing,  
 data = scrabble)  
  
# Printing model output  
slrModel %>%   
 broom::tidy() %>%   
 make\_flex()

| term | estimate | std.error | statistic | p.value |
| --- | --- | --- | --- | --- |
| (Intercept) | 5.64 | 0.61 | 9.18 | 2.5e-09 |
| perc\_rf\_english\_writing | -0.59 | 0.12 | -4.89 | 5.5e-05 |

**Provide a statement of the estimated simple linear regression model for the Scrabble data.**

A statement of the estimated regression equation is

yHat = 5.64 - 0.59\*x,

where:

yHat is the predicted Scrabble tile point value

βHat0 = 5.64 is the estimated intercept

βHat1 = -0.59 is the estimated slope

x is the percent relative frequency of a letter in English writing

**Provide and interpret the value of the coefficient of determination in context.**

r2 = (-0.71)2 = 0.5041

**General Blueprint Interpretation**

The proportion of the variability in y explained by x is *r2*;

**Scrabble example**

About 50.41% of the variability in the scrabble tile point values is explained by the percent relative frequencies of the letters in English writing.

or

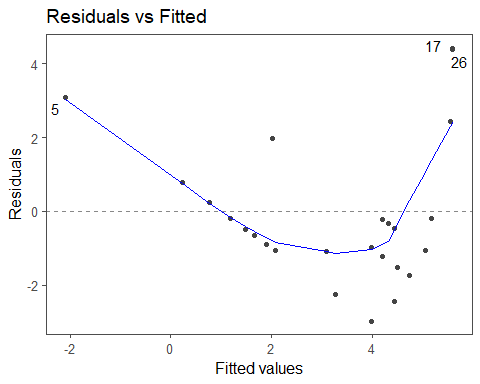
The proportion of the variability in the scrabble tile point values is explained by the percent relative frequencies of the letters in English writing is 0.5041.

**Does our model explain significantly more of the variability in the response than an intercept only model? Why or why not?**

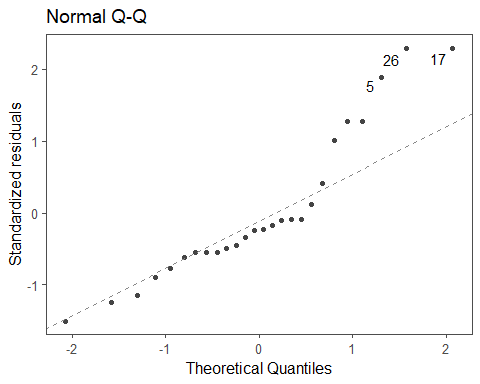
Since the p-value for the slope (5.5e-05) is less than 0.05, our model explain significantly more of the variability in the response than an intercept only model.

### Model Assumptions

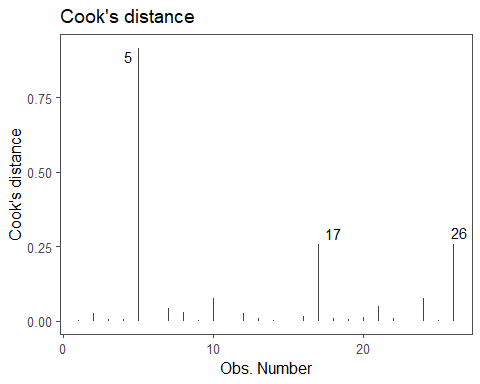
# Creating diagnostic plots  
  
# Residual by fitted value plot  
autoplot(slrModel, which = 1, label.repel = TRUE)@plots[[1]]



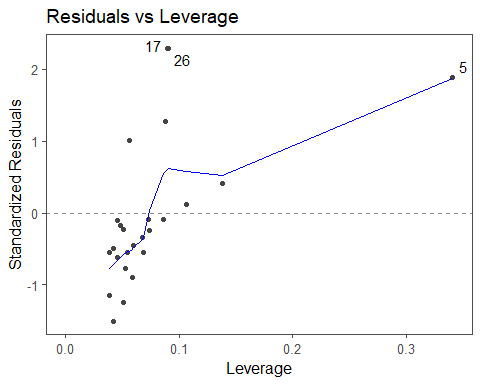
# Quantile-quantile (QQ) plot  
autoplot(slrModel, which = 2, label.repel = TRUE)@plots[[1]]



# Cook's Distance plot  
autoplot(slrModel, which = 4, label.repel = TRUE)@plots[[1]]



# Standardized residuals by leverage  
autoplot(slrModel, which = 5, label.repel = TRUE)@plots[[1]]



**Indicate if each assumption for fitting a simple linear regression model is met, providing specific evidence to support your conclusions from the output provided.**

We must assume that we have a random sample.

**Normality:** the residuals must be normally distributed. Checked using a quantile-quantile (QQ) plot of the residuals. Since there are several points that stray far from the diagonal line through the origin, the normality assumption is violated.

We will proceed with checking the other assumptions for the sake of the problem.

**Homoskedasticity:** the residuals must have constant variance / spread across all predicted values Checked using the residual by predicted value plot.

There is an outward fanning pattern in the residual by fitted value plot, so the homoskedasticity assumption is violated.

**Independence:** the residuals must be independent of the predicted values. Checked using the residual by predicted value plot.

There is a quadratic pattern in the residual by fitted value plot, so the independence assumption is violated.

**Linearity:** there must be a linear relationship between each predictor and Y, our response variable. Checked using the residual by predicted value plot or a scatter plot between X and Y.

Since the points in the scatter-plot follow a curved pattern, there is not a linear relationship between the percent relative frequency of a letter in English writing and the Scrabble tile point values.

**Are there any influential points in the data set? Provide specific evidence to support your answer.**

Yes, there is one influential point with a Cook’s D value of about 0.9(the letter E) which is greater than 0.5.

**Are there any potential outliers in the data set? Provide specific evidence to support your answer.**

Yes, there are two outliers since they have standardized residuals that are larger than 2 in magnitude(observations 17 and 26).

### Inference of the SLR model

**Regardless of whether or not assumptions are met, formally state the hypotheses for testing if the regression slope is 0 or not.**

Our hypotheses are H0:β1=0 vs. Ha:β1≠0

**Provide the test statistic, p-value, and decision for this hypothesis test.**

**test statistic**: t = -4.89

**p-value**: 5.5e-05

**decision**: Reject the null hypotheses, since the p-value is less than 0.5.

**Interpret the results of the hypothesis test in the context of the problem.**

We have sufficient evidence that there is a negative linear relationship between the percent relative frequency of a letter in English writing and the Scrabble tile point value at 5% significance level.

**If the p-value had been larger than 0.05(not the case here though)**

We have insufficient evidence that there is a linear relationship between the percent relative frequency of a letter in English writing and the Scrabble tile point value at 5% significance level.

**Regardless of whether or not assumptions are met, provide and interpret the estimated slope in context, rounding to two decimal places.**

βHat1 = -0.59

*Blueprint Interpretation*

For every 1 unit increase in x, we expect y to increase/decrease by |βHat1|.

*Interpretation in context*

For every 1 percent increase in the percent relative frequency of a letter in English writing, we expect the Scrabble tile point value of a letter to decrease by 0.59.

**Provide and interpret the estimated intercept in context, rounding to two decimal places. Is our interpretation sensible in this setting? Why or why not?**

βHat0 = 5.64

*Blueprint Interpretation*

The expected value of y given that x=0 is βHat0.

*Interpretation in context*

(WE DO NOT INTERPRET THIS KIND OF CASE)The expected Scrabble tile point value of a letter given that is not used in English writing is 5.64.

This is inappropriate since a letter being used in a Scrabble that doesn’t exist in English writing is not possible.

It is also inappropriate since it is considered **extrapolation**, since a percent relative frequency of 0 is outside the range of the observed data set.

**Using the model output, what is the expected Scrabble tile point value for a new letter, λ, that has a percent relative frequency in English writing of 1.5?**

yHat = 5.64 - 0.59 \* x = 5.64 - 0.59 \* (1.5) = 4.755 points

**Based on this predicted value, what is the corresponding residual value for the letter B?**

residual = observed - predicted

e = y - yHat = 3.00 - 4.755 = -1.755

**for the letter C**

e = y - yHat = 3.00 - (5.64 - 0.59 \* (2.8)) = 3.00 - 3.988 = -0.988

**Using the model estimates, what would be the expected Scrabble tile point value for a new letter that has a percent relative frequency in English writing of 20? Would this be a reliable prediction in this context?**

Since 20 is outside the scope of the percent relative frequency in the data set, this would be extrapolation, so a predicted value is inappropriate here. Therefore, no, the prediction is not reliable in this context.

**A 95% prediction interval for the Scrabble tile point value of a new letter, λ, that has a percent relative frequency in English writing of 1.5 can be obtained using the code below. Calculate and interpret this prediction interval in context.**

# Obtaining predicted value & 95% prediciton interval  
predData <- tibble(letter = "λ",  
 perc\_rf\_english\_writing = 1.5)  
   
predict(slrModel, newdata = predData,   
 interval = "prediction", level = 0.95) %>%   
 as\_tibble() %>%   
 make\_flex(ndigits = 2)

| fit | lwr | upr |
| --- | --- | --- |
| 4.75 | 0.46 | 9.04 |

(Lower Bound, Upper Bound) = (0.46, 9.04)

*Blueprint Interpretation*

When X = x\*, we expect y to be between *Lower Bound* and *Upper Bound* units, 95% of the time.

*For this context*

When a letter has a percent relative frequency in English writing of 1.5, we expect its Scrabble tile point value to be between 0.46 and 9.04 points, 95% of the time.