Contact: gwmelvin@colby.edu

Writing Assignment[1]

In this writing assignment you will

- learn how to derive the **line of best fit** for a given data set;
- learn about an application of calculus to statistical analysis;
- gain practice communicating technical information;
- perform a statistical analysis on a data set of your choice.

Background

Consider a collection of n items of data of the form

$$(x_1, y_1), \ldots, (x_n, y_n)$$

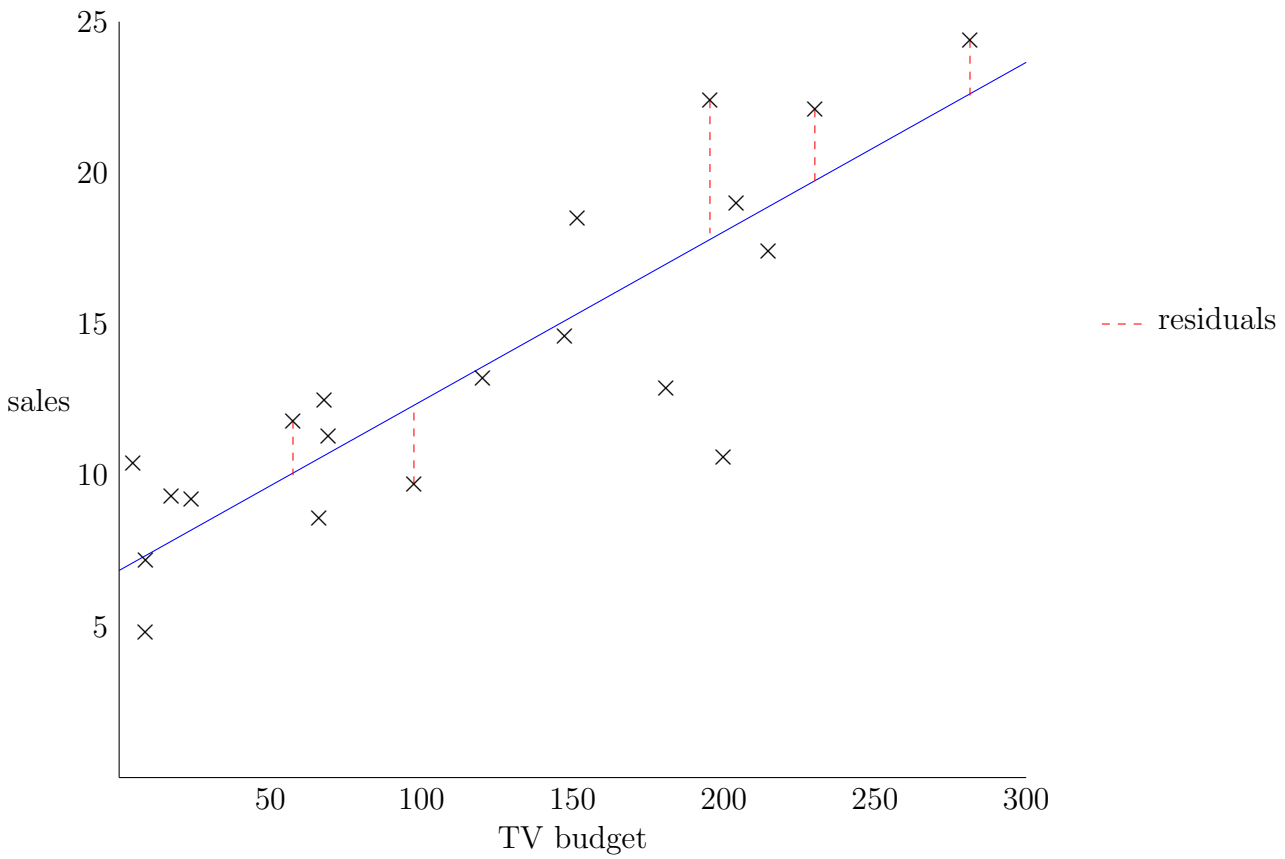
For example, each datum could represent an incoming Colby College student's high school GPA score and SAT score: the pair $(3.83, 1400)$ would denote that a student had a GPA of 3.83 in high school and received a 1400 on the SAT.

We obtain a **scatterplot** by plotting this data in the xy-plane. For example, the following table represents (real) data relating twenty company's total television advertising budgets (in thousands of dollars) to sales of their product (in thousands of units)

TV Budget	**Sales**
230.1	22.1
44.5	10.4
17.2	9.3
151.5	18.5
180.8	12.9
8.7	7.2
57.5	11.8
120.2	13.2
8.6	4.8
199.8	10.6
66.1	8.6
214.7	17.4
23.8	9.2
97.5	9.7
204.1	19
195.4	22.4
67.8	12.5
281.4	24.4
69.2	11.3
147.3	14.6

[1]Adapted from *Applications in Calculus*, MAA Notes No. 29, Vol. 3, p. 23-41, and *"An Introduction to Statistical Learning, with applications in R" (Springer, 2013)* with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani.

Let x_i denote the TV budget of company i (in thousands of dollars) and y_i denote the sales of the product made by company i (in thousands of units). The resulting scatterplot is given below:



Each $\times$ represents a data point. Also represented on the scatterplot is the **line of best fit**: this is the unique line that *minimises* statistical error. Understanding what this means requires a detour into **linear regression**. You can learn more about this in the statistics courses SC212 *Intro to Stats and Data Science* and SC321 *Statistical Modelling*. Linear regression is a basic tool used in data analysis and machine learning, among many other fields.

Linear Regression

Regression analysis is a collection of tools for exploring relationships between variables. For example, in the budget/sales example given above we would like to predict sales based on TV advertising budget.

By *relationship* between two variables, we mean that we would like to understand the degree to which values of one variable rise or fall with values of the second variable. For example, we say that there is a positive relationship between TV budget and sales: as TV budget increases so do sales. Due to extraneous factors outside our control our data does not precisely follow a straight line. These '*natural*' deviations of our hypothetical straight line relationship we call **statistical error**: it simply means unknown and random deviations from an underlying model.

In linear regression we try to describe the relationship between two variables x, y using a so-called *model*: this is simply a mathematical description of the relationship between x and y via a linear function

$$y = f(x) + error$$

where $f(x)$ is a linear function and *error* represents unknown, random deviations from $f(x)$.

Given a scatterplot and a line through the plot we need a measure of fit in order to compare the fit of different lines to the data represented by the scatterplot. By *measure of fit* we mean a single number that

summarises how well a given line fits a given scatterplot, so that smaller values indicate a better fit. If the data fall exactly on a straight line then the measure of fit between the line and the data will be zero. Having chosen a measure of fit, we can use it to find the line that minimises this measure for a given scatterplot, thereby providing an objective algorithm for choosing the line of best fit for the scatterplot.

Given any line through a scatterplot, the difference between an actual value and a fitted value is called a **residual**. For example, on the scatterplot above, the residual is the (signed) distance from the data point (represented by $\times$) and the point on the line with the same x-value: this is the (signed) length of the red dashed line. Mathematically, if $f(x) = mx + b$ is a linear function used to give a model of the data then a residual is given by the difference

$$y_i - f(x_i) = y_i - mx_i - b$$

In assessing how well a particular line fits the data we would like for these residuals (ignoring sign) to be collectively as small as possible. We need a measure of fit that summarises this desired property in a single number. There are several ways to do this and we will use the **root mean square error (RMSE)**: the root mean squared error is defined as the average of the square root of the squared residuals

$$\text{RMSE } = \sqrt{\frac{(y_1 - mx_1 - b)^2 + (y_2 - mx_2 - b)^2 + \ldots + (y_n - mx_n - b)^2}{n}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - mx_i - b)^2}$$

Our goal is to find a line (i.e. a linear function $f(x) = mx + b$) that will minimise RMSE. Such a line is called the **line of best fit**, the **least squares line** or the **regression line**.

Description of the line of best fit

Given a data set

$$(x_1, y_1), \ldots, (x_n, y_n)$$

denote the mean averages

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Theorem: *The line of best fit for the data set is the graph $y = f(x) = mx + b$, where*

$$b = \overline{y} - m\overline{x}, \qquad m = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

You will derive the formula for the line of best fit for a given data set using optimisation methods in calculus.

Exercise: Using a calculator, check that the line of best fit for the TV advertising/sales data set given above is

$$y = 0.0557x + 6.8502$$

Derivation of the line of best fit

Consider a data set

$$(x_1, y_1), \ldots, (x_n, y_n)$$

We want to find values m, b that will minimise the RMSE for the data set. It is sufficient[2] to find values for m and b that minimise the sum of squared residuals

$$(y_1 - mx_1 - b)^2 + (y_2 - mx_2 - b)^2 + \ldots + (y_n - mx_n - b)^2 = \sum_{i=1}^{n}(y_i - mx_i - b)^2$$

[2]This is a George-given Truth.

Important Note: in this minimisation problem the x_i's and y_i's are **numerical constants**, the given data. The **variables** are m and b, the quantities we are free to manipulate to minimise the sum of squared residuals

You will derive the equation for the line of best fit in several steps.

Step 1. (*Warm-up*) Consider a line having fixed slope $m = 1$. We want to find the value of b that minimises the function

$$g(b) = (y_1 - x_1 - b)^2 + (y_2 - x_2 - b)^2 + \ldots + (y_n - x_n - b)^2$$

(a) Explain why $g(b)$ is a quadratic polynomial in b. Deduce that $g(b)$ must have a unique global minimum.

(b) Show that

$$g'(b) = -2\left(\sum_{i=1}^{n} y_i - nb - \sum_{i=1}^{n} x_i\right)$$

(c) Deduce that there is a unique critical point when

$$b = \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{1}{n}\sum_{i=1}^{n} x_i$$

(d) Show that this critical point is a minimum of $g(b)$.

Step 2. Now, consider a line having fixed slope m and consider the function

$$g(b) = (y_1 - mx_1 - b)^2 + (y_2 - mx_2 - b)^2 + \ldots + (y_n - mx_n - b)^2$$

(a) Explain why $g(b)$ is a quadratic polynomial in b. Deduce that $g(b)$ must have a unique global minimum.

(b) Show that

$$g'(b) = -2\left(\sum_{i=1}^{n} y_i - nb - m\sum_{i=1}^{n} x_i\right)$$

(c) Deduce that there is a unique critical point when

$$b = \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{m}{n}\sum_{i=1}^{n} x_i = \overline{y} - m\overline{x}$$

Here $\overline{x}, \overline{y}$ denote the mean average of x_i's and y_i's respectively.

(d) Show that this critical point is a minimum of $g(b)$.

(e) Deduce that lines of best fit must be of the form

$$y = mx + (\overline{y} - m\overline{x}) = \overline{y} + m(x - \overline{x}) \qquad (*)$$

Step 3. Now consider all lines of the form $(*)$. Among these lines you will find the unique value of m that minimises the RMSE. Consider the function

$$f(m) = (y_1 - \overline{y} - m(x_1 - \overline{x}))^2 + (y_2 - \overline{y} - m(x_2 - \overline{x}))^2 + \ldots + (y_n - \overline{y} - m(x_n - \overline{x}))^2$$

Remember that this is a function of the variable m and that $x_i, y_i, \overline{x}, \overline{y}$ are all numerical constants.

(a) Explain why $f(m)$ is a quadratic polynomial in m. Deduce that $f(m)$ must have a unique global minimum.

(b) Show that

$$f'(m) = \sum_{i=1}^{n} 2(y_i - \overline{y} - m(x_i - \overline{x}))(-(x_i - \overline{x}))$$

(c) Show that there is a unique critical point m satisfying

$$\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x}) = m\sum_{i=1}^{n}(x_i - \overline{x})^2$$

and deduce that

$$m = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Step 4. Use your solutions to the above problems to show that the line of best fit is

$$y = mx + b$$

where

$$b = \overline{y} - m\overline{x}, \qquad m = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Writing Assignment

Submission Date: Wednesday December 5, 4pm

Write a 3-5 page report outlining your derivation of the line of best fit. You write-up should be typed and single-spaced. You should also include in your report the following:

1. A brief summary of linear regression and some of its applications in the real world (e.g. the natural sciences, psychology, machine learning etc.)

2. Determine the line of best fit on a data set that you have generated yourself. Your data set must contain at least twenty items of data. An example of how you could generate a data set could be: ask some of your friends for their height and shoe size. Let x_i = height of person i, y_i = shoe size of person i. The data set you generate should be more interesting than this example, however.

3. Your report should include at least three distinct references (i.e. not just Wikipedia). This document does not constitute a reference (though you may use all information presented here without reference).

Some tips and suggestions:

- **Do not leave this assignment until the weekend before submission.** You are expected to devote at least 10 hours of your time towards this assignment.

- Your report should be written in a professional tone (i.e. don't use colloquial language, don't begin sentences with *So,* etc.). Interesting personal touches and **WOW** factors are encouraged, however.

- When including equations in your report it is a good idea to leave some blank space and write them in by hand. You will not be penalised for doing this. If you have the time/inclination you could try to use LaTeX - this is the typesetting software I used to create this document. You can create LaTeX documents online at overleaf.com (sign-up required). LaTeX is not required, however.

- Diagrams and visuals are an excellent way to keep the reader interested.
- Feel free to skip several steps of algebra and/or computation in your write-up. You should be aiming to convince the reader that the computation you are performing is valid, not drowning them in formulae.
- You should include an introduction, a conclusion and a list of references.