Regression analysis

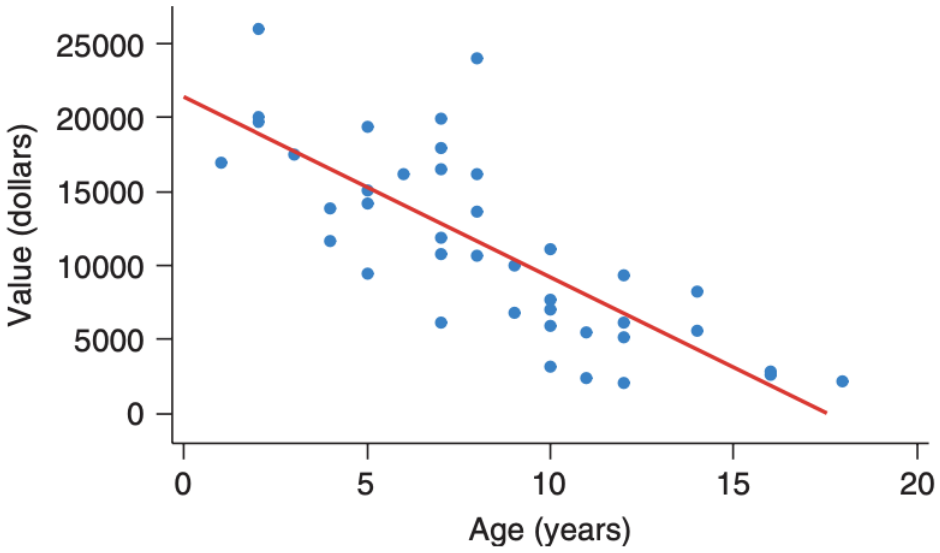
1 Linear regression:

Once we established the relationship between explanatory variable (x) and response variable (y) in a linear relationship by plotting a scatterplot, we started our analysis with correlation. Correlation measures the direction and strength of a straight-line relationship. The next question becomes what next in the analysis?

The general motivation is "We have a linear relationship. can draw a line which closely fits the data points in the scatterplot, and give us what is called a 'line of best fit'?"

A line of best fit summarizes the relationship between an explanatory variable and a response variable. But the question might still linger "Why do we want to find an equation of a line that closely fits our data?" The idea is to use a regression line to predict the value of y for a given value of x when we believe the relationship between y and x is linear. Now we aren't restricted with only the data we have. It broadens our horizon from data analysis to making predictions based on an existing dataset.

Example: The scatterplot below shows the linear relationship between the age of car (in years) and value of car (in dollars). Now the next step is to draw a regression line (shown in red in the scatterplot below) which summarizes this relationship between age of car and value of car. Furthermore, it helps us to make predictions about the value of cars for different years beyond our data based on the existing data.



2 Importance of linear regression:

Since we can make predictions beyond the data we have, linear regression is one of the most fundamental and powerful statistical tools. ChatGPT, or AI models like it, uses linear regression as a part of the training process to learn and understand relationships between input data and outputs. Linear regression is a fundamental statistical method used for predictive analysis, making it a critical component in all machine learning models.

3 Line of best fit:

The natural follow-up question can be "Since we are making predictions, what about the error factor? Predictions are rarely 100% accurate." The method that is developed to write the line of best fit is called least squares regression. In nutshell, least squares regression is used to write a line of best fit so the line gives the best possible prediction on average. We want to minimize the total error because it means that the data points are collectively as close to the model's values as possible. Now the next step is to actually write the equation of this least square regression line.

Now we know the line of best fit is also called least squares regression line. Commonly, it is also referred as trendline. Like any equation of line, the the equation of a least square regression line is

$$y = a + bx$$

where a is y-intercept and b is slope of the line.
It is beyond the scope of the course, but if interested refer to additional reading on how to write thew line of best fit.

The slope of a line describes the change in y for each unit increase in x. The y-intercept of a line is the y-value of the line when $x = 0$. The value of $y-$ intercept may or may not make sense in real life. For this course, we will just look at the regression line and start our analysis.

Example: A linear relationship has been found between average SAT score and college GPA in students calculated as $y = 1.52 + 0.0017x$ or GPA $= 1.52 + 0.0017\times$SAT score.

Slope: When SAT score increases by 1, college GPA increases by 0.0017.

$y-$ intercept: For SAT score of 0, college GPA is 1.52. It doesn't make sense.

Example: Suppose a linear regression equation predicts the value of car (in dollars) based on how many years old the car is using the equation $y = 21375 - 1215x$. Interpret the slope and $y-$ intercept.

Here explanatory variable is the age of car and response variable is value of car. The equation can be written as: Value of car = 21375 - 1215 $\times$ Age of car.

Slope: When the age of car increases by 1, the value of car decreases by \$1215. In simpler words, the value of car decreases by \$1215 every year.

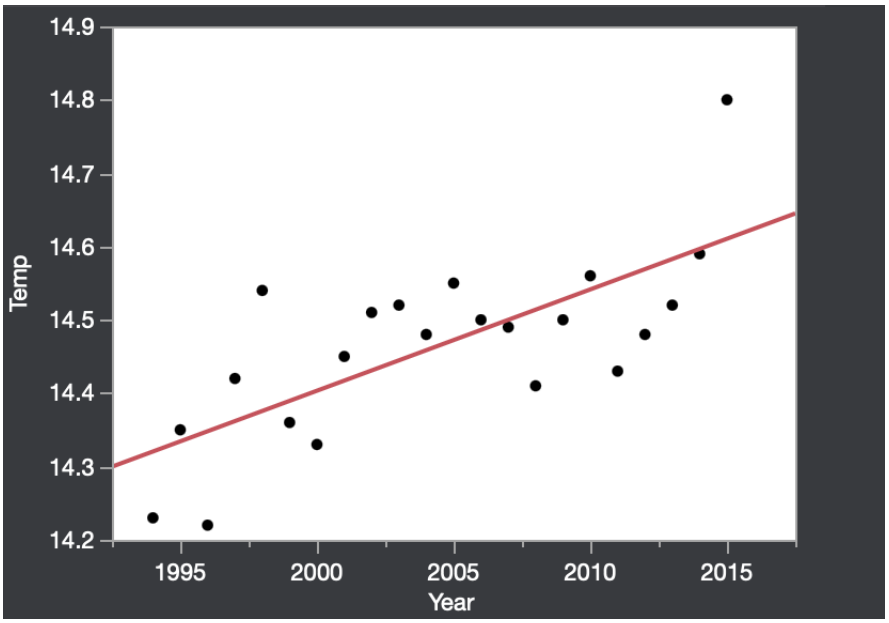
$y-$ intercept: The value of a brand new car is \$21375.

4 Cautionary notes about least squares regression:

1. Unlike in correlation, order matters. If x and y are switched, the regression equation will change.
2. Interpolation is the way of predicting the value inside the range of data whereas extrapolation is the way of predicting the value outside the range of data.
3. Don't extrapolate! Don't make predictions beyond the range of the data, because we are not sure that the linear trend will continue beyond the range of the data.

5 An extended example:

The following scatterplot shows the data on annual average summer temperature in Helsinki, Finland from 1994 to 2018, in degrees Celsius and the regression Line is given as: Temperature = -13.25 + 0.014 × Year



Slope: The average summer temperature increases by 0.014 degree Celsius every year.

$y-$intercept: In 1994, the average summer temperature was -13.25 degree Celsius which doesn't make sense.

Predict the average summer temperature in Helsinki in 2000. Is it a good prediction?

$$\text{Temperature} = -13.25 + 0.014 \times \text{Year} = -13.25 + 0.014 \times 2000 = 14.75$$

Yes, it is a good prediction because we are inside the range of our data i.e. interpolation.

An environmentalist uses the line to predict average summer temperature in Helsinki in 2050. What is the prediction? Do you think this prediction is reliable? Why or why not?

$$\text{Temperature} = -13.25 + 0.014 \times \text{Year} = -13.25 + 0.014 \times 2050 = 15.45$$

No, it is an unreliable prediction because it is an extrapolation.

6 Coefficient of Determination: r^2

The square of r, the correlation coefficient measures how much variation in the response variable is explained by the explanatory variable. It is usually converted to a percentage, so it always falls between 0% and 100% The larger r^2, the smaller the amount of variation or scatter about the regression line.

Example: For the predicted value and age of car example from before, $r = -0.778$. Compute and interpret r^2.

$$r^2 = (-0.778)^2 = .605, 60.5\%.$$

Thus, car age explains about 60.5% of the variation in car value.

A properly displayed regression line always has the equation of regression line and r^2 value.

7 An interesting example:

The following tweet was posted by the twitter account Universal Life Church in 2022. It claimed that everyone in America will be gay by 2047 if the current trend (trend in 2022) held.

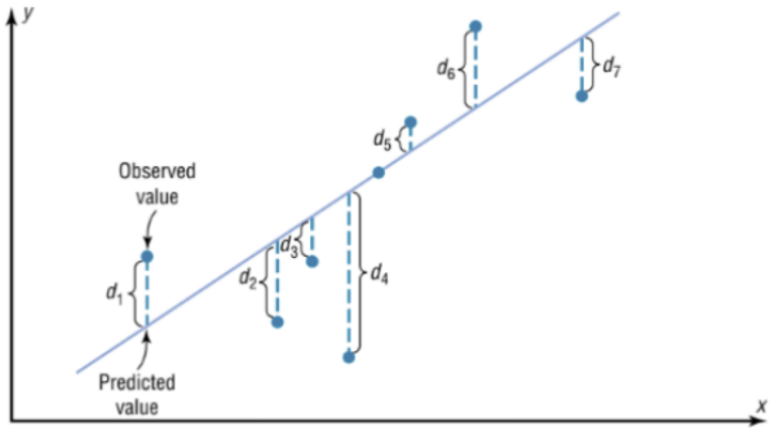
Even though there is no linear regression involved here, we see a classic mistake of incorrectly extrapolating the data. A lot of people have recently came out as gay due to the society becoming more welcoming of the LGBTQ people than ever before. Hence, we have seen a growing trend in LGBTQ population now. However, this does not mean the trend holds and everyone in America will be gay by 2047.

8 Further reading on Line of best fit:

The method that is developed to write the line of best fit is called least squares regression.

Statisticians call it "least squares" because it minimizes the sum of the squared residuals. Residuals are the differences between the observed data values and the least squares regression line. Hence, a residual is the difference between the observed value and the model's predicted value. There is one residual per data point, and they collectively indicate the degree to which the model is wrong. Data points above the line have positive residuals, while those below are negative.

To calculate the residual mathematically, Residual = Observed value – Model value.



The points $d1$, $d5$, and $d6$ are positive residuals whereas $d2$, $d3$, $d4$, and $d7$ are negative residuals.

Residuals represent the error in a least squares model. You want to minimize the total error because it means that the data points are collectively as close to the model's values as possible. Now the next step is to actually write the equation of this least square regression line.

Now we know the line of best fit is also called least squares regression line. Commonly, it is also referred as trendline.

Like any equation of a line, we need a slope and $y-$intercept. It is no different in a least square regression line. Let's build the equation of a least square regression line.

So far we have an explanatory variable x and response variable y. We can then calculate the mean and standard deviation of x denoted by $\bar{x}$ and s_x respectively. Let $\bar{y}$ be the mean and s_y be the standard deviation of y. Let r be the correlation coefficient of x and y. Then we can write our slope as

$$b = r\frac{s_y}{s_x}$$

and $y-$intercept as

$$a = \bar{y} - b\bar{x}$$

Then the equation of a least square regression line is

$$y = a + bx$$

where a is y-intercept and b is slope of the line.

The slope of a line describes the change in y for each unit increase in x. The y-intercept of a line is the y-value of the line when $x = 0$. The value of $y-$ intercept may or may not make sense in real life.