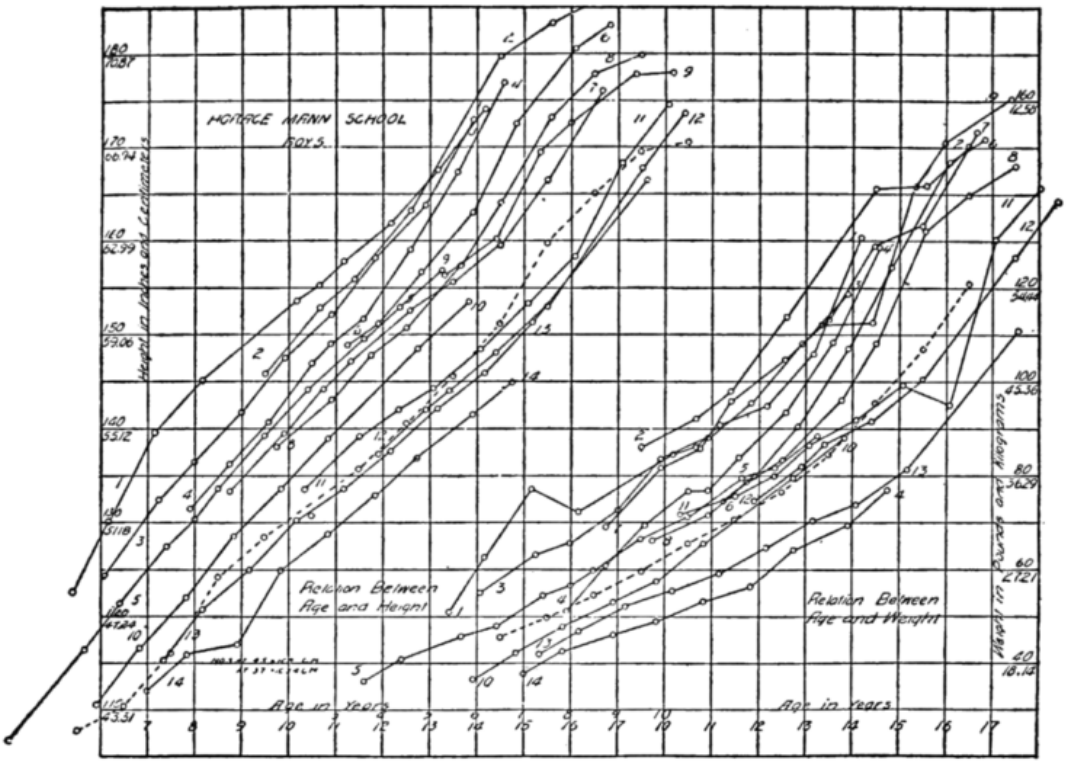
Linear Regression

See online here

This article deals with linear regression as a means to describe the relationship between two variables using a line. You will learn about assumptions, residuals, extrapolation, variation and the appropriateness of linear regression in different cases.



The Linear Model (Describing Relationship with Lines)

It is the most widely used technique of data analysis and measurement in statistics. This model is aimed at explaining the relationship between two quantitative variables with a line. Of these two quantitative variables, one variable is of independent nature, whereas the other is dependent in nature. Normally, in a linear model,“Y” is considered as the dependent variable and “X” as the independent variable. The equation of straight line or linear model comprises of the following factors:

Y = Dependent variable

X = Independent variable

t = Time period

b = Coefficient of variable

a = Slope of intercept

Assumptions of linear regression

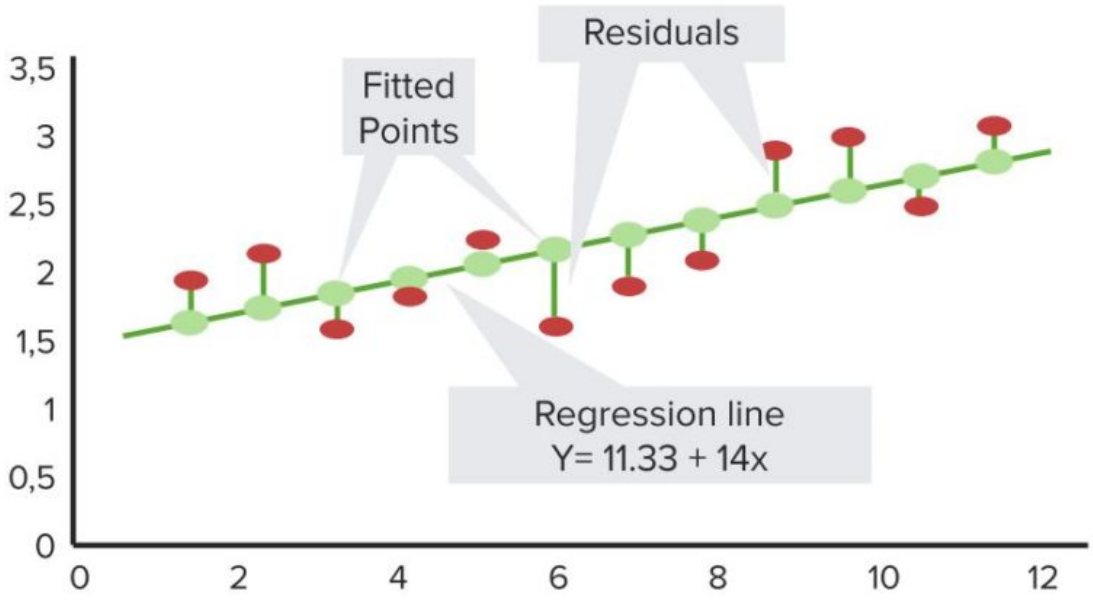
The assumptions that the relationship between dependent and independent variables should be linear are based on the following considered facts:

- The linear relationships between variables are **non-trivial relationships** which are imaginary.
- The true relationship between variables is often and at least **linear** over the range of values.
- Despite the fact that the relationship between variables is not linear, we can linearize it accordingly by **transformation** of variables.

Residuals

This model of linear regression helps in measuring **imperfections** in data. In case of linear model, the straight line may deviate from data to the line. These deviations create imperfections in linear regression analysis.

In order to find out the required linear relationship between variables, we predict value and observe it later using linear regression formulas. The observations which are observed using linear models are denoted by “y”. The values which are predicted before the data analysis process are denoted by “ŷ”. The **difference between observed values y and predicted values ŷ** are termed as residuals.



“Residuals” Image created by Lecturio

The regression line

This line helps in measuring line which fits best with available data. Larger a residual, it shows that a line poorly fits the available data. The regression line creates a line which fits best with given data with small residual values. The sum of positive and negative

residuals is always zero.

The regression line helps in finding out the line which minimizes the sum of all residuals. Eventually, positive residuals cancel the effect of negative residuals and the ultimate effect of residual is zero.

A regression line is the one which minimizes the sum of the squares of all residuals of available data. When we find out the line which plays the role of the regression line, it is considered as line of best fit or least square line.

The formula of regression line is given as follows:

$$\mathbf{Y' = bx + a}$$

Calculating the Y-intercept

Y-intercept is the value which is measured when the regression line hits Y axis. It tells about the expected value of y or dependent variable when value of independent value x = 0. It is simply the value which is measured at the point regression line intersects Y-axis.

The formula of Y-intercept is the same like linear regression i.e. y = bx + a.

Suppose the value of b = 4, if x value = -1 and y value = -6. Putting the values in Y-intercept formula, the value of b comes out:

(–6) = (4) (–1) + a

–6 = –4 + a

–2 = **a (y intercept)**

Using the slope to find the intercept

The slope of the data can be measured by analyzing data which is the steepest. The slope and intercept of an equation indicates the relationship between independent and dependent variable. It finds out the average rate of change of variable. In case the magnitude has a huge slope, the regression line becomes steeper which indicates the higher rate of change.

Extrapolation

This is the process of **making predictions outside the values available of a given data**. In case a prediction about response variable is huge outside the range of given data; it becomes riskier and difficult to predict the continuation of a linear relationship between independent and dependent variables.

Extrapolation uses regression equation to make a **prediction about the correspondence of explanatory and response variables**. Response variables are dependent (Y) whereas explanatory variables are independent variables (X) in a data set.

Extrapolation involves regressions equation for prediction outside the data range despite the fact that it does not indicate what is happening outside the range of given data. **It should be avoided to make predictions outside the range of given data**.

Using Regression as a Crystal Ball

A regression line helps in predicting the value of dependent variable “Y” by observing the impact of independent variable “X”. In **moderate correlation** between two variables through scatterplot and correlation coefficient, indications are there that some kind of linear relationship exists.

While using regression as a crystal ball, it is relevant to find out which variable will play the role of a dependent value and which one will be taken as X or dependent variable. In order to find the best fitting lines for valuable predictions, the choice of X and Y makes a difference. In order to make a correct prediction regarding Y, the following conditions are required to be met:

- The scatterplot of given data should create a linear pattern.
- The correlations coefficient i.e. r should be moderate or lying between +0.50 or -0.50.

Revisiting residuals

In order to find the appropriateness of linear regression model for making prediction of response and explanatory variables, it is necessary to look at the **distribution of residuals**. If residuals show a normal distribution, the prediction about response variable becomes easy and clear.

In case the residuals are not normally distributed, it indicates **deviation** from linearity. Values are scattered and deviate from a straight line showing that the line is not best fit.

Variation in Regression Model

The variation accounted for the relationship between response variable (Y) and explanatory variable (X) can be measured by using **R-squared quantity** denoted by R^2. R-squared quantity indicates the percentage of variation between values of X and Y. The simplest way to calculate R-squared quantity is by squaring the correlation.

R^2 (rule of thumb)

There is no rule of thumb for good value of R^2 in dataset. It **varies from data to data**. Scientific experimental data normally has R^2 between 80 and 90%. Observations have shown lower R^2 value is useful if lies between 30 and 50%.

Appropriateness of Linear Regression Model

In order to get effective results by use of linear regression model, the following four conditions should be met:

1. Quantitative variable condition

Both dependent and independent variables of a data set, neither of the variables should be categorical. A **categorical variable** is the one which can take a fixed number or limited value, further assigning the other variable a specific category based on its quantitative property. In case any of the variables are categorical, the linear regression model should be stopped immediately.

2. Straight on scatterplot condition

A scatterplot helps in finding out whether the regression line fits best in a straight line or not. If the scatterplot shows **dispersion of residuals**, it should be stopped.

3. Outlier condition

A scatterplot should be used to identify outliers. In case outliers are identified, linear regression model does not work best with such data. Outliers are observations which show larger value than predictor values, response or dependent variable i.e. Y in this case.

4. Consistency of explanatory variable with straight line

The straight line of linear regression graph should be aligned with values of explanatory or independent variables (X).

In case any of the above given four conditions is missing, the data set is not suitable for linear regression model.

Checking Assumptions of Regression Model

- The fitted values should be checked by **drawing a plot** of the residuals.
- In case of any **bend in the plot**, there is indication of non-alignment of explanatory variable in a straight line.
- Find any **outlier** it may occur after completion of regression calculation and at the stage of drawing scatterplot.
- Checking for any **vertical spread of values** from one part of the plot to the other.
- Ideally random scatter “0” is considered best fit of regression line on graph which is practically hard to achieve.

Residuals vs. fitted values plot results

In case the following observations are noted from the residual plot, it indicates the **violation of straight line condition** which is a worrisome factor:

- The plot has shown a huge and steep bend of values.
- The thickness of plot increased when fitted value lies 0.50 to 0.60.

Common Regression Mistakes

Normally, the following mistakes are practised by researchers while using the linear regression model which should be avoided to get the desired results.

- The linear regression model **should not be used for non-linear relationship** between two or more variables.
- **Outliers are ignored** which further create a violation of the straight line relationship between variables.
- It is considered that independent variable X causes dependent variable Y to occur which is incorrect. **X only influences Y** due to the strong linear relationship between these two.
- The **choice of X and Y** is not made at the initial stage of the linear regression

process.
- It may be possible that the **regression line is used to predict X from Y**.

References

Arendacká & S.Puntanen, 2014. Further remarks on the connection between fixed linear model and mixed linear model. Statistical Papers, 56(4), pp. 1235-1247.

Eldar, Y., 2006. Comparing between estimation approaches: admissible and dominating linear estimators. IEEE Transactions on Signal Processing, 54(5), pp. 1689-1702.

Greene, M., 2017. Interpret the slope and intercept of a regression line. [Online] Available at: https://learnzillion.com/lesson_plans/78 [Accessed 8 March 2017].

Groß, J. & Markiewicz, A., 2004. Characterizations of admissible linear estimators in the linear model. Linear Algebra and its Applications, Volume 388, pp. 239-248.

Lane, D. M., 2016. Introduction to Linear Regression. [Online] Available at: http://onlinestatbook.com/2/regression/intro.html [Accessed 5 March 2016].

Nayland College, 2016. Residuals (Extension). [Online] Available at: http://maths.nayland.school.nz/Year_13_Maths/3.9_Bivariate_data/13_Residuals.html [Accessed 5 March 2017].

PennState Eberly College of Science , 2016. Normal Probability Plot of Residuals. [Online] Available at: https://onlinecourses.science.psu.edu/stat501/node/281 [Accessed 5 March 2017].

Rumsey, D. J., 2017. USING LINEAR REGRESSION TO PREDICT AN OUTCOME. [Online] Available at: http://www.dummies.com/education/math/statistics/using-linear-regression-to-predict-an-outcome/ [Accessed 28 Feb 2017].

Synówka-Bejenka, E. & Zontek, S., 2007. A characterization of admissible linear estimators of fixed and random effects in linear models. Metrika, 68(2), pp. 157-172.

Notes