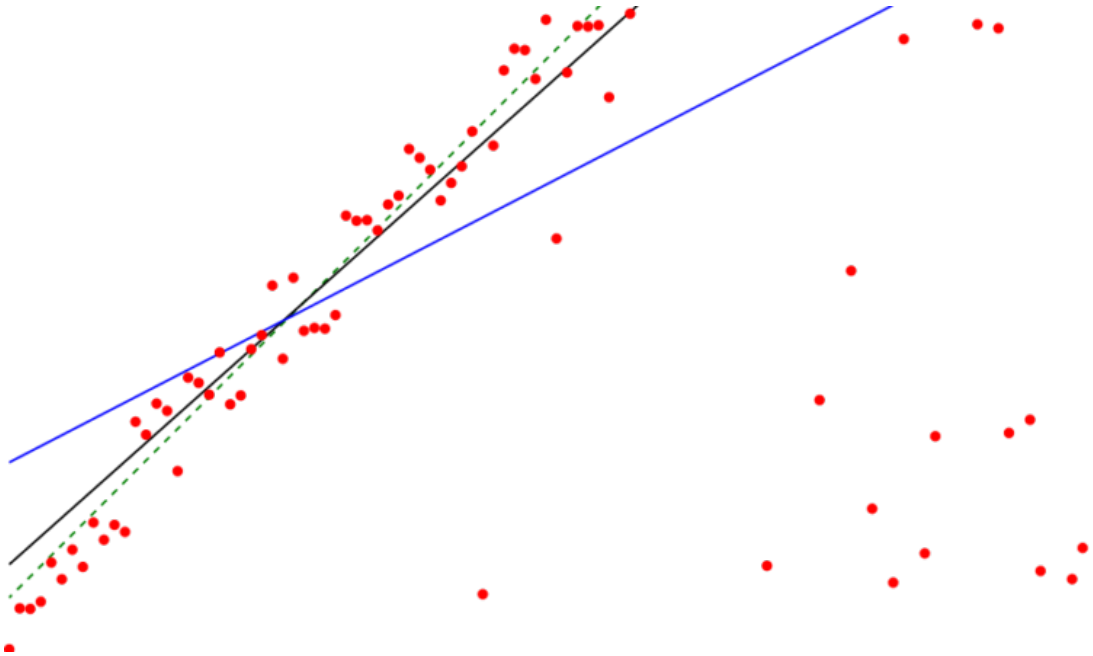
Statistics: Addressing Issues with Regression Assumptions

See online here

This article discusses residuals plots explaining the dependent and independent variable on random and non-random pattern. The article has explained the impact of the violation of regression assumptions. Groups and subsets mentioned in this article explain problems with multiple groups and separate regression. This article has also discussed outliers, leverage, influential points, lurking variables, and causation. Transformation of data explained in this article stated goals of transformation, appropriate transformations and types of Logarithmic transformations.



Residual Plots

This plot shows residuals on the vertical axis and independent or explanatory variable on the horizontal axis. In case the points scattered on the residual plot are aligned around a straight line, it indicates the linear model is suitable for such data. If residual points are scattered and show deviation from a straight line, it indicates that the non-linear model is more suitable for such data set. The residual plot containing fitted and dispersed values are shown below:

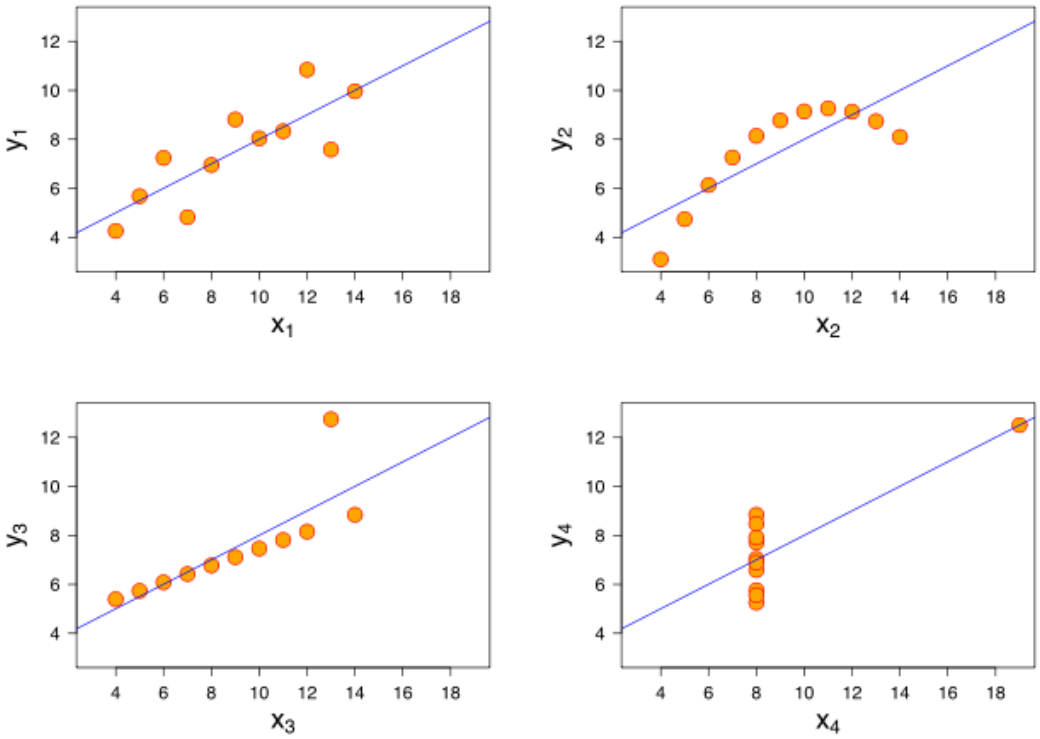


Image: “The data sets in the Anscombe’s quartet are designed to have the same linear regression line but are graphically very different.” by Anscombe.svg: Schutz Derivative works of this file: (label using subscripts): Avenue – Anscombe.svg. License: CC BY-SA 3.0

Random scatter around 0 (reasonableness of linear model)

In order to check whether the assumptions taken for the linear regression model are reasonable or not, the pattern and spread of residuals have to be taken into account. If residuals show a random scatter value = 0 or around 0, it should linear regression assumptions are taken for data are reasonable.

Violation of regression assumptions

In case the following assumptions are not properly taken into account, there is the possibility of a lack of a best fitted and strongly linear relationship between variables.

An **outlier** may occur after completion of regression calculation and at the stage of drawing scatterplot. The fitted values should be checked by drawing a plot of the residuals. In the case of any **bend in the plot**, there is an indication of non-alignment of the explanatory variable in a straight line. Checking for any **vertical spread** of values from one part of the plot to the other.

If any of the assumptions of the linear model are not properly accounted for, the residual plot depicts violation of underlying assumptions of the linear model which has to be rectified.

Groups and Subsets (Problems with Multiple Groups)

The presence of **small clusters** of data or residuals in different areas of a residual Vs fitted value plot indicates that there are more than one group of data. In case there is more than one group, the regression of all separate groups is required.

Example: If we are observing the relationship between petrol sales and prices of cars in a country since the prices of cars have a direct impact on sales of petrol of different brands, there will be several groups of brands according to the price of petrol.

Separate regressions

Each group of petrol brands will have to be analyzed by separate regression models. For each group of petrol brand, different linear and non-linear models will be suitable. These models of data analysis for separate groups will be different from models of an entire set of data.

Rule of regression

All sets of data should belong to a single and homogeneous population. In the case of separate groups (petrol brand example), each individual group should be analyzed separately.

Outliers, Leverage and Influential Points

Outliers

It is the value observed in data which have large residual value. An outlier is far away from values of data set plotted on a scatterplot. Outliers have a **significant impact on a regression model**. A data can be considered outliers in following four ways:

- It could have an extreme X value, compared to other data points.
- It could have an extreme Y value, compared to other data points.
- It could have extreme X and Y values.
- It might be distant from the rest of the data, even without extreme X or Y values.

Influential point

It is the type of outlier which specifically impacts the slope of the linear regression model. In order to make an estimate of the influence of an outlier, the regression equation has to be calculated with and without outlier value. When an outlier is present on a plot, the slope is flatter comparatively.

In the case of **influential point analysis,** the following things should be taken into account:

- Influential point is the representation of bad data. It indicates a measurement error which requires investigation of the validity of data point.
- Comparison of decisions taken after computation of regression equations with and without influential points in a residual plot. In case equation leads to

deviating decisions, researchers should be cautious for using a linear regression model.

Lurking Variable and Causation

Researchers should not assume that an independent variable “X”, says the price of cars in the above-mentioned example causes the dependent variable “Y” i.e. price of petrol of different brands. It does not matter how high the correlation or perfectly linear relationship between two variables, it can’t be inferred that one variable causes the other one. Each variable has its own occurrence conditions independent of each other.

Transforming Data

In order to deal with problems of outlier and thickness on scatterplot or in the residual vs. fitted values plot, the transformation of data is a helpful method.

Goals of transformations

The goals of the transformation of data include:

- It aims at making the distribution of a variable more symmetric and more linear. It helps in achieving normality of a data set. A histogram can be used to assess the linearity of data.
- It aims at creating uniformity in the spread of several groups, despite the difference between their centers. Side by side box plots can be used for this assessment.
- To make the form of scatterplot more linear.
- To avoid thickening around the line in a plot by spreading scatterplot evenly.

Appropriate transformations

Different types of data transformation serve the purpose of a data modification to eliminate residuals.

Ladder of powers

- In the case of unimodal distributions which are skewed left, the dependent variable values “y” should be squared “y^2”.
- For count data transformation, the square root of dependent variable helps in the elimination of errors.
- **Log transformation ln (y):** Log of values help in transformation for the values which can’t be negative and grow by percentage.
- **Negative reciprocal:** It helps in transforming the measuring ratio of response values. It also aids in altering the direction of a relationship.

The logarithmic transformation

In some cases, the transformation of data through the ladder of powers does not aid properly in fixing the curvature of the scatterplot. Logarithmic transformations are helping to sort such issues.

Types of logarithmic transformation

- **X-axis:** x y-axis: **ln(y) – exponential transformation** is suitable for data values which tend to increase by percentage.
- **X-axis:** ln (x), y-axis: **y-Logarithmic model** is helpful when scatter plot decline both at the left and right side of plot.
- **X-axis:** Y-axis: ln (y) – power transformation- when the above-mentioned types of logarithmic transformations are not helpful, this transformation aids.

Common Issues with Regression Assumptions

It should be ensured that the relationship between two variables is straight. Different groups should also be identified in regression analysis. Extrapolation should be avoided. High leverage and influential points have to be identified. In order to examine the unusual impact of points on the linear model, two regressions should be compared.

If the data set has multiple modes, it is an indication of several groups. One should be beware of lurking variables. Use of regression to imply causation should be avoided. It means one variable (independent) does not cause another variable (dependent).

The linear model is not perfect, it is an ideal situation which is hard to achieve. It should not be expected. Don’t stray too far when data transformation is done through a ladder of power. R-squared quantity (R2) should be avoided while choosing a model.

References

Interpreting residual plots to improve your regression. via docs.statwing.com

FENG, C. et al., 2014. Log-transformation and its implications for data analysis. Shanghai Arch Psychiatry, 26(2), p. 105–109.

Linear regression models. via people.duke.edu

Influential Points in Regression. via stattrek.com

Yang, Y. & Liang, X., 2013. Confirmatory factor analysis under violations of distributional and structural assumptions. International Journal of Quantitative Research in Education, 1(1), pp. 61-64.

Notes