From the Expert: Simple Linear Regression

Week 2: Simple Linear Regression

Regression was originally developed in 18th century to solve astronomy problems. Adrien-Marie Legendre; a French mathematician, formulated the least squares method in 1805. Later, in 1875, Francis Galton coined the term regression to explain the situation where the heights of descendants of tall ancestors tend to regress towards a normal average. For example, sons of tall fathers tend for their heights to be closer to the average (shorter) while sons of short fathers tend to also have their heights be closer to average (taller). This effect is referred as “regression to mediocrity”. This is an originating of the term “regression analysis”.

Regression analysis is a statistical process for estimating linear dependence relationships among variables. The main goal is to predict the response from one or more variables. Regression analysis can be used in prediction and forecasting. These usages, however, are overlap with the field of machine learning. Regression analysis can be used to answer the following questions:

- (**Descriptive**) What is the relationship between the dependent and independent variables?
- (**Inferences**) Which independent variables are the most important?
- (**Prediction**) What is the value of the response variable given one or more observation values?

In general, regression analysis is used when both independent and dependent variables are continuous. Nevertheless, regression can also apply to categorical independent variables and dichotomous dependent variables with some modifications.

The dependent variable is also referred to as the response variable or outcome variable, whereas, independent variable is also known as predictor variable.

Examples of regression analysis applied in business applications are:

- The effect/relationship of interest rates and stock prices
- The wage and retention rate of the employees
- The advertisement budget and corporate sales
- The subscription rate and the membership cost
- Which promotions generate the most sales?

Regression analysis consists of:

- A single response variable *Y* (must be continuous variable)

- One or more predictor variables:
  Where *n*=1 is simple regression
  n 2 is multiple regression (or multivariate in public health)

- The unknown parameters β (scalar or vector)

1. Simple Linear Regression

Simple linear regression is used for examining the relationship between two quantitative variables by linear equations that best summarize the relation, for instance, advertisement budget and the revenue. Typically, the dependent variable or response variable (*y*) measures an outcome of a study, whereas, the independent variable or explanatory variable (*x*) cause the change in the response variable. Simple linear regression involves only a ***single*** quantitative ***explanatory variable***.

Can you specify the possible explanatory variable and response variable in following problems?

- The number of clicks and the amount of revenue

- The membership fees and number of subscriptions

- The yield of produce and inches of rain

The (population) relationship between *y* and *x* can be formulated as:

Where , independeent
The unknown parameters include:
(Intercept) is the point where the line intercept *y*-axis
(Slope) is the slope of the line or the increase in *y* per unit change in *x*. Note that, is positive when *y* (linearly) increases as *x* increases. is negative when *y* (linearly) decreases as *x* increases. This straight line often used to describe the trend of the data set.

Linear Regression (population)

Source

Line Type	Regression Equation	Intercept	Slope
Population			
Sample	=		

In most settings, we cannot determine the population parameters directly. Thus, the values are estimated from a sample. The sample regression line is an estimation of the population regression. The goal is to find the equation of line that fits the data the best. In other words, we need to find and such that the observed value () and the fitted value or predicted value () is minimized. The fitted value is given by

$$=$$

The difference between the observed value () and the fitted value is known as residual. Therefore, the residual is . If most of the residuals are small, it usually indicates that the model is good at explaining the response variable or the model has a good fit.

2. Fitting Regression Lines: (Ordinary) Least Squares or OLS

There are many methods for fitting a line such as minimize sum of prediction errors (Unfortunately, there are many lines that satisfy this equation criterion. Thus, a better method is needed. The OLS is the most common method for estimating unknown parameters in linear regression models by minimizing the sum of squared residuals (SSE) or residual sum squares (RSS), which are the sum squared deviations (vertical distance) between each data point and the regression line. Subject to constraint that total error is 0.

$$=$$

$$=$$

The relationship between 2 variables can be simply explored using scatter plots and the correlation coefficient. For regression analysis, however, there is additional step that is a straight line is superimposed (overlaid) on a scatter plot to clarify the relationship.

Without going into calculation details, the coefficients and from least squares equation can be formulated as:

- Intercept () is the value of the (predicted) response value where the value of explanatory variable is equal to zero.
- Slope () is the amount of change in the predicted response variable where the explanatory variable is changed by one unit.
- The regression line can be used to predict the value of the response variable for a given value of explanatory variable . Interpolation is the prediction ***in the range*** of the observed value . While, extrapolation is the prediction ***outside*** the range of the observed value. Pay attention in prediction for any values when *x* is further away from the observed range since the linear relationship between and may not valid outside this range.

These are many statistical packages such as R, SPSS, SAS, and Excel that can be used to compute these coefficients and other regression measures as part of regression analysis.
Least squares regression consists of these properties:

- The sum of the residuals of the least squares regression line is equal to zero
- The sum of squared residuals is minimized, that is
  = 0
- The simple regression line always pass through)
- The least square coefficients are unbiased estimations of and

3. Inference for Regression Parameters

Since there is only one predictor variable in simple linear regression, therefore, the main focus in on the slope . The slope indicates a change in the response *y* for a unit change of *x* of *x*.

T-test is used for testing the slope of the population to see whether there is any linear relationship between 2 variables.

3.1 Hypothesis for slope testing:

(there is no linear relationship between tested variables)
(there is a linear relationship)

Note that, we can also use one tail for the alternative hypothesis.

Test statistic: with DF (degree of freedom) = n-2
is the slope coefficient of sample regression
is the hypothesized slope
is the standard error estimator of the slope

will be rejected if Test statistic t fall in the critical region.
In this case, or , when is the level of confidence.

3.2 Confidence Intervals for Regression Coefficients

Estimation of confidence interval for :
Estimation of confidence interval for :

4. Underlying Assumptions for linear regression include:

- The sample is representative of the population for the inference prediction.
- Independence: the value of each outcome variable is independent from each other (need to know how data were collected)
- Linearity: the relationship between predictor and outcome variable is a reasonably straight line. This can be detected by plotting the data between observed values and predicted values. The points should be distributed along a diagonal line.

Scatter plot to test linear relationship

Source

- Normality: for a fixed value of *x*,the response *y* varies according to a normal distribution. Non-normally distributed (e.g. highly skewed, kurtosis) can distort relationship. This can be examined by a histogram, which should be close to normal distribution. The Kolmogorov-Smirnov, Anderson-Darling, and Shapiro-Wilk test provides inference statistics test on normality.

Data transformation such as inverse, square root or log can improve normality. Note, that transformation can be used for correcting model assumption violations and improving the fit. However, the interpretation could be complicated.

- Homoscedasticity: the prediction error should be spread with the same degree (or constant) for the entire data range. In other words, probability distribution of the errors has constant variance. Violation of homoscedasticity is known as Heteroscedasticity, where error spread with different degree and with many shapes such as fan or bow-tie. This assumption can be verified by plotting the residual against the predicted values, the residuals should randomly scattered around the horizontal line and distribute relatively consistent.

Homoscedasticity (left) and Heteroscedasticity (right) Examples

Source

- Independence and normality of error: the prediction error should be independent from each other error and should be normally distributed. Independence can be detected by plotting residuals against the predicted value (). The error distributed normality can be checked by normal quantile plot of residuals (theoretical standardized error and actual standardized error). The points should line close to the diagonal reference line.

Errors are independent Errors are normally distributed

Source

Note that it is important to check the validity of the assumptions before continuing with the inference or prediction. The first two assumptions are fulfilled for the proper design study. The last 4 assumptions should not be violated otherwise the results may not be reliable. The violation consequences include Type I or Type II error, over or under-estimation of significance and/or effect sizes.

With the plot between *y* and *x*, we can investigate the followings:

Residual Analysis is a diagnostic method based mainly on the residuals. The model requires that . Thus, the standardized residuals should follow a standard normal distribution. Residual analysis is often done graphically using

- Quantile plots – to examine normality
- Scatter plots – to assess model assumptions such as linearity, constant variance and potential outliers
- Histograms, stem, boxplot, and leaf diagram

Source

Source

5. Outliers:

- Outliers are not typical observations. They usually appear outside the pattern of other observations.
- To identify outliers, do the scatter plot.
- Problems of outliers
- Including outliers in analysis may result in changing the conclusions.
- Excluding outliers that influence (correct operation of) the system can mislead the conclusion.
- Values that are different from others should be examined for experimental error. After eliminating experimental error, one can decide whether to use or not use these values.

6. Regression Analysis general procedure:

- Define research questions, questions of interest (including theory and hypothesis to be tested)

- Review the study design (including data availability, error corrections, and assumptions)
- Explore the data and check the assumptions such as
- linearity assumption can check by scatter plot. Transformation may be needed if not linear trend
- normality can be checked by histogram and statistical tests
- Perform regression analysis
- Evaluate the model, check the fit of the model
- Examine to see how much variance in the response is explained by the model
- To make sure that the results are valid, we need to perform residual analysis such as
- check homoscedasticity assumption
- check error normality assumption and error independency
- Interpret results (e.g. test statistics, confidence interval, prediction values)
- Presentation of results

7. Correlation coefficient (r) and coefficient of determination ()

Correlation coefficient is a standard measurement of association or relationship between 2 variables. Typically, the symbol ρ denotes the population correlation (from the population data) and r is the sample correlation. Please keep in mind that correlation or association is not causation. As a reminder, regression is used to predict Y from X using a linear rule. Correlation describes how good the relationship is.

The Pearson (product-moment) correlation or simply called correlation coefficient (r) is a typical numerical measure of the strength and direction between two variables relationship. It can be calculated using the following formula:

Here are important properties of r:

- The value is between [-1,1]
- $r = +1$ when there is a prefect linear relationship between Y and X (with positive slope)
- $r = -1$ when there is a prefect linear relationship between Y and X (with negative slope)
- For positive correlation ($r > 0$), Y tends to increase linearly with X
- For negative correlation ($r < 0$), Y tends to decrease linearly with X
- Size of r suggests strength of the linear relationship

When there is a strong linear relationship (r is close to +1 or -1), this suggests that Y can be accurately predicted. A value of r that is close to 0 indicates A weak correlation or the linear equation is not so helpful in Y prediction.

Different **linear** correlation coefficient values

Source

Source

The coefficient of determination indicates how much of the variation in one variable can be accounted for by the other variable or total variation in the dependent variable this is explained by variation in the independent variable. In general, the higher the , the better the model fits the data. The value of is between 0 and 1 (no negative value). If there is no linear relationship between outcome and response variables, is 0. If there is a perfect linear relationship between outcome and response variables, is 1.
Regression is used to predict *y* from *x* using a linear rule. Correlation describes how good the relationship is.

8. Measuring Goodness of Fit

Coefficient of determination is the proportion of the total variability explained by the regression model, and it indicates how well the model fits the data.

Read the following examples on linear regression using R:

Example 1

Example 2

Example 3

Model diagnostics for regression can be found here.

References:

Lindquist, A.M. (2009) Introduction to Statistics, course Notes. Columbia University.
Retrieved from: https://web.archive.org/web/20181223071756/http://www.stat.columbia.edu/~martin/W2024/R7.pdf

Mellor-Crummey, J (2005) Linear Regression Models, Rice University.
Retrieved from: https://web.archive.org/web/20190925233315/https://www.cs.rice.edu/~johnmc/comp528/lecture-notes/Lecture9.pdf

Torres-Reyna, O (2010) Getting Started in Linear Regression using R. Princeton University. Retrieved from:
https://web.archive.org/web/20190925233640/http://www.princeton.edu/~otorres/Regression101R.pdf

Shanker, M (2006), Fundamentals of Business Statistics, Kent State University,
Retrieved from: https://web.archive.org/web/20190925233203/http://www.personal.kent.edu/~mshanker/personal/Classes/f06/ch13_F06.pdf

Reflect in ePortfolio Download Print ‹ ›

Activity Details

Task: View this topic