Regression Analysis

The primary objective of regression analysis is to estimate the value of a random variable (the dependent variable) given that the value of an associated variable (the independent variable) is known.

The dependent variable is also called the response variable, while the independent variable is also called the predictor variable. The regression equation is the algebraic formula by which the estimated value of the dependent, or response, variable is determined

Simple and Multiple Linear Regression

The term simple regression analysis indicates that the value of a dependent variable is estimated on the basis of one independent variable. Multiple regression analysis is concerned with estimating the value of a dependent variable on the basis of two or more independent variables.

The general assumptions underlying the regression analysis model are that

(1) the dependent variable is a random variable,
(2) the independent and dependent variables are linearly associated.

Assumption (1) indicates that although the values of the independent variable **may be controlled**, the values of the dependent variable must be obtained through the process of random sampling.

When hypothesis testing is done in the regression analysis, three additional required assumptions are that

(3) the variances of the conditional distributions of the dependent variable, given different values for the independent variable, are all equal,

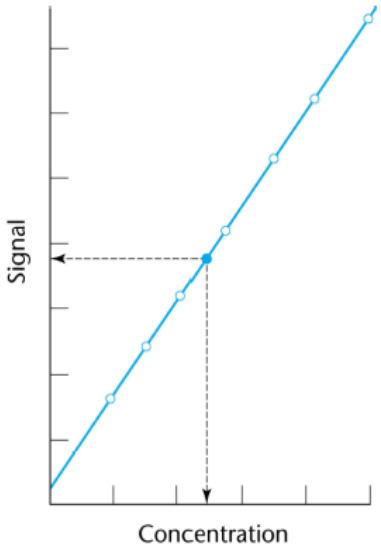
(4) the conditional distributions of the dependent variable, given different values for the independent variable, are all normally distributed in the population of values,

(5) the observed values of the dependent variable are independent of each other.

Regression Line

A regression line is a line drawn through the points on a scatterplot to summarise the relationship between the variables being studied. When it slopes down (from top left to bottom right), this indicates a negative or inverse relationship between the variables; when it slopes up (from bottom right to top left), a positive or direct relationship is indicated.

The regression line often represents the regression equation on a scatterplot.



Regression Equation

A regression equation allows us to express the relationship between two (or more) variables algebraically. It indicates the nature of the relationship between two (or more) variables. In particular, it indicates the extent to which you can predict some variables by knowing others, or the extent to which some are associated with others.

Recall that we class one variable as the response or dependent variable, usually denoted Y, and the other as the predictor, or dependent variable, usually denoted X. (X can be said to "cause" changes in Y.)

If a linear relationship , the relationship between X and Y is formulated as follows:

$$Y = \beta_o + \beta_1 X + \varepsilon$$

- is the dependent variable
- β_0 is the intercept coefficient
- β_1 is the slope coefficient
- X is the independent variable
- ε is the residual term (i.e random error)

Both β_0 and β_1 are almost always unknown (population) values. However these are the key terms in the model. From a sample of data, estimates for the intercept and slope coefficient can be derived.

A fitted line to model the data as a linear regression mode (i.e. a regression equation) is usually written as

$$Y^* = b_o + b_1 X$$

where

- Y* is the predicted value for the dependent variable
- b_0 is the intercept estimate
- b_1 is the slope estimate
- X is the independent variable

Simple linear regression is from a family of models known as Linear Models. The ***R*** command used to implement such models is `lm()` **.**

The regression model is specified in the following form: `lm(Y ~ X)`

The operator "~" (the tilde sign) is taken to mean "is explained by" or "is predicted by".

For our previous example (used in the last class), the simple linear model can be implemented as follows:

```
> lm(Fluo~Conc)

Call:
lm(formula = Fluo ~ Conc)

Coefficients:
(Intercept)         Conc
      1.518        1.930
```

Using the coefficients given in the computer output, the regression equation is therefore

= 1.52 + 1.93X

Where

= Fluoresence (i.e. predicted value for Fluoresence)

X = Concentration

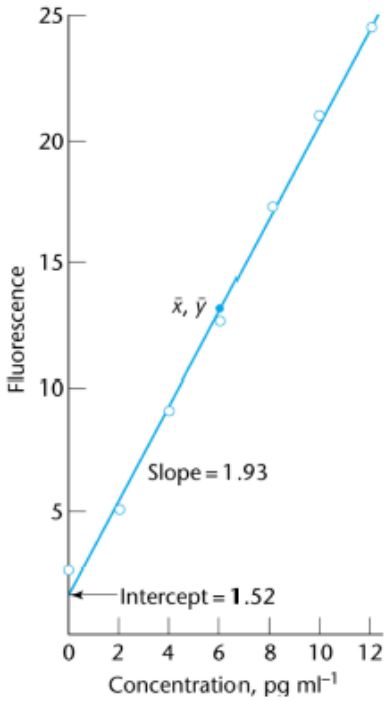
The equation will specify the average magnitude of the expected change in Y given a change in X. The regression equation is often represented on a scatterplot by a regression line.

Once the regression equation is formulated, then this equation can be used to estimate the value of the dependent variable given the value of the independent variable. However, such estimation should be done only within the range of the values of the independent variable originally sampled, since there is no statistical basis to assume that the regression line is appropriate outside these limits.

Question: What is the predicted fluoresence, given that the concentration is 5 units.

= 1.52 + 1.93X
= 1.52 + 1.93(5)
= 11.17

Further, it should be determined whether the relationship expressed by the regression equation is real or could have occurred in the sample data purely by chance

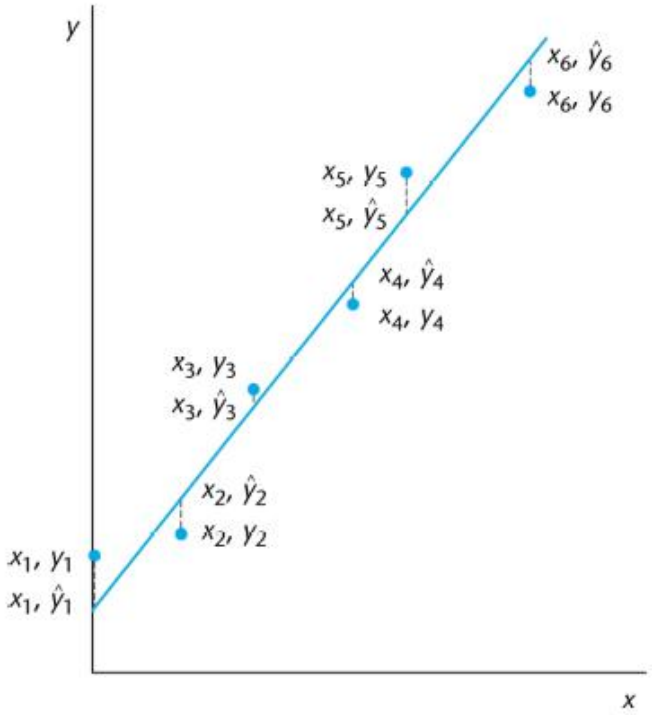


RESIDUALS AND RESIDUAL PLOTS

For a given value X of the independent variable, the regression line value ***Y****often is called the ***fitted value*** of the dependent variable. The difference between the observed value Y and the fitted value Y is called the ***residual*** for that observation and is denoted by e:

$$e = Y - Y^*$$

(Important for later: Residuals represent unexplained (or residual) variation after fitting a regression model.)

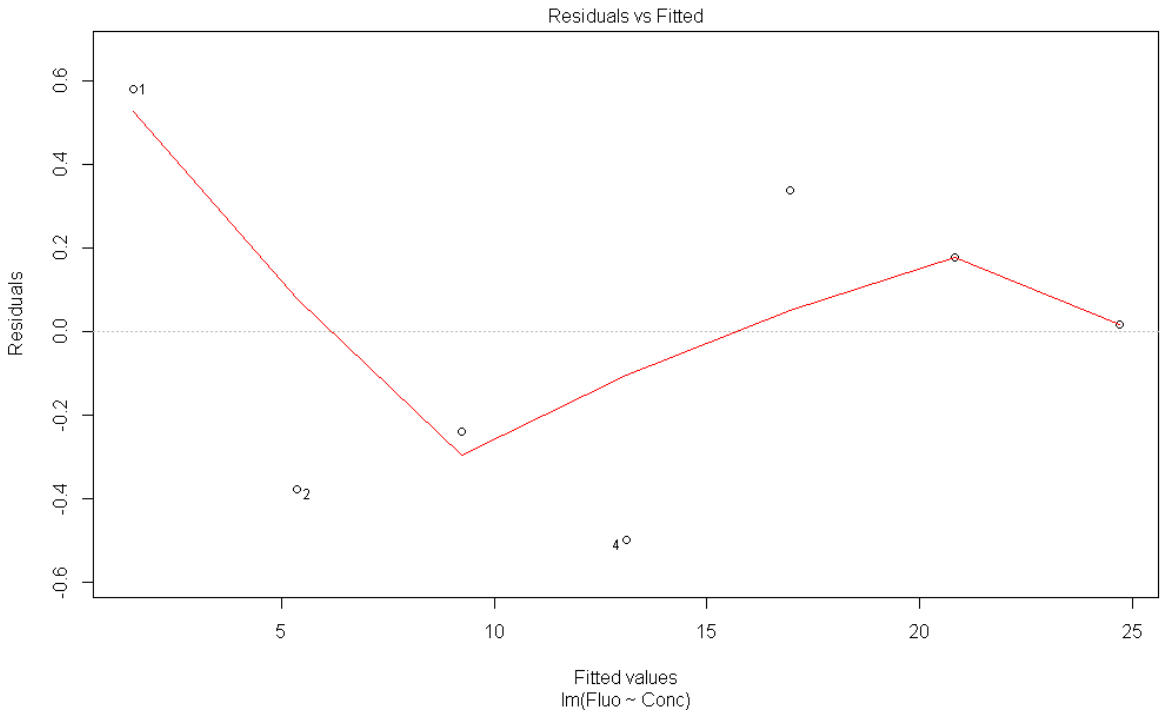


For the example used in this class, the residuals are very small.

Examining the Residuals

Recall: The mean value of the residuals is zero,
The variance of residuals are constant across the range of measurements,
The residuals are normally distributed,
Residuals are independent.

A residual plot is obtained by plotting the residuals ***e*** with respect to the independent variable ***X*** or, alternatively with respect to the fitted regression line values ***Y****. Such a plot can be used to investigate whether the assumptions concerning the residuals appear to be satisfied.



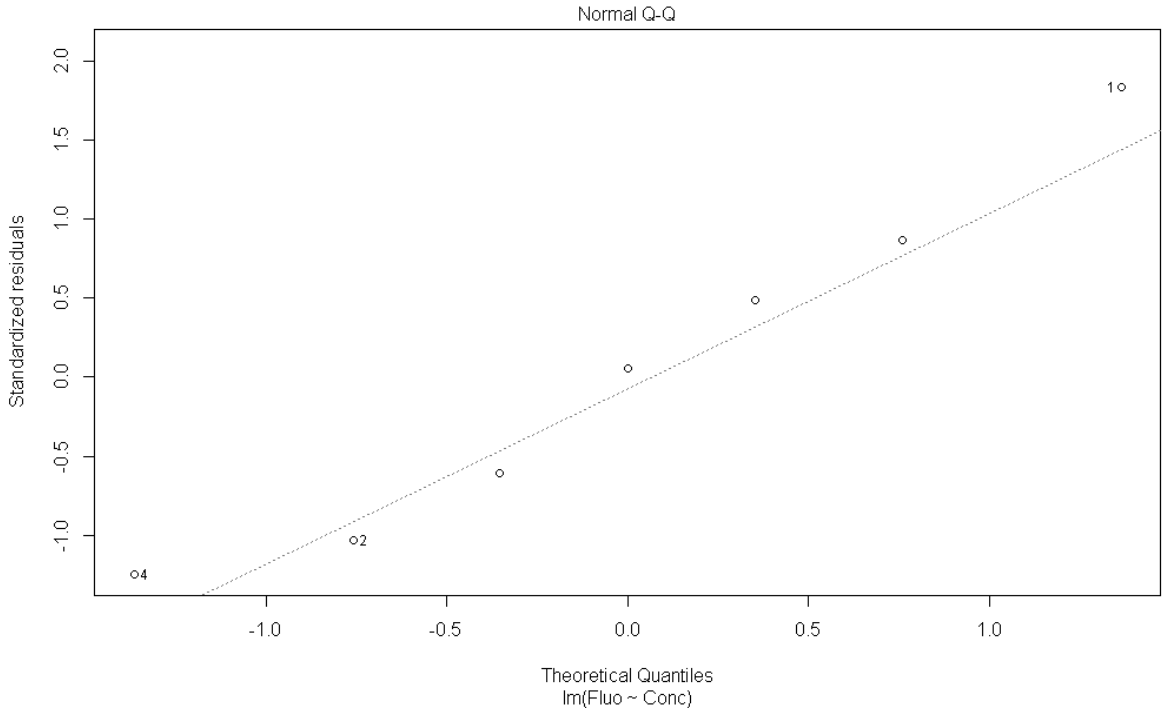
Homoscedascity (also known as constant variance) is one of the assumptions required in a regression analysis in order to make valid statistical inferences about population relationships.

Homoscedasticity requires that the variance of the residuals are constant for all fitted values, indicated by a uniform scatter or dispersion of data points about the trend line (i.e. "The Zero Line").

From the above plot, we can conclude that the constant variance assumption is valid. We can also see that the mean value of the residuals is close to zero. (Actually it is precisely zero)

Normality of Residuals

Additionally, the linear model approach requires this assumption that the residuals are normally distributed. We can use a Q-Q plot to assess the normality of residuals.



Inferences Concerning the Slope

Before a regression equation is used for the purpose of estimation or prediction, we should first determine if a relationship appears to exist between the two variables in the population, or whether the observed relationship in the sample could have occurred by chance.

In the absence of any relationship in the population, the slope of the population regression line would, by definition, be zero: $\boldsymbol{\beta_1 = 0}$.

Therefore the usual null and alternative hypotheses tested is

H_0: $\boldsymbol{\beta_1 = 0}$.
H_1: $\boldsymbol{\beta_1 \neq 0}$.

Such a test is equivalent to a formal test of the linear relationship between the two variables.
If we fail to reject the null hypothesis, we must conclude that the independent variable has no bearing on the value of the dependent variable.

The null hypothesis can also be formulated as a one-tail test, in which case the alternative hypothesis is not simply that the two variables are related, but that the relationship is of a specific type (direct or inverse).

A hypothesized value of the slope is tested by computing a t statistic and using n - 2 degrees of freedom. Two degrees of freedom are lost in the process of inference because two parameter estimates, $\boldsymbol{b_0}$ and $\boldsymbol{b_1}$, are included in the regression equation.

The standard formula is

$$t = \frac{b_1 - (\beta_1)_0}{s_{b_1}}$$

However, when the null hypothesis is that the slope is zero, which generally is the hypothesis, then the formula is simplified and is stated as:

$$t = \frac{b_1}{s_{b_1}}$$

Inferences Concerning the Intercept

The same approach to formal testing is equally applicable to the Intercept.

Therefore the usual null and alternative hypotheses tested is

H_0: $\boldsymbol{\beta_0 = 0}$.
H_1: $\boldsymbol{\beta_0 \neq 0}$.

We will discuss the use of such a test in future classes. It is not usually given as much attention, in general. However it is quite a useful test for chemists.

Performing Inference procedures for intercept and slope.

Recall the example from the last class: Concentration and Fluoresence.

```
> Fit

Call:
lm(formula = Fluo ~ Conc)

Coefficients:
(Intercept)         Conc
      1.518        1.930
```

To compute the p-values for inferences on the intercept and slope, we use the summary command, specifying the regression model we have chosen to use.

The p-value is written as `Pr(>|t|)`. Additionally there is a useful visual aid : the number of asterisks beside the p-value, if the p-value is sufficiently small.

A guide to reading the significance codes is provided in the output.

- Three asterisks indicate a p-value of less than 0.001
- Two asterisks indicate a p-value of less than 0.01
- One asterisk indicate a p-value of less than 0.05

(Remark: we have chosen 0.01 as a threshold for rejecting the null hypothesis. This is an arbitrary level)

```
> summary(Fit)

Call:
lm(formula = Fluo ~ Conc)

Residuals:
       1        2        3        4        5        6        7
 0.58214 -0.37857 -0.23929 -0.50000  0.33929  0.17857  0.01786

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.5179     0.2949   5.146  0.00363 **
Conc          1.9304     0.0409  47.197 8.07e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4328 on 5 degrees of freedom
Multiple R-squared: 0.9978,     Adjusted R-squared: 0.9973
F-statistic:  2228 on 1 and 5 DF,  p-value: 8.066e-08
```

We reject the null hypotheses for both the slope and intercept.