Regression

Regression

Regression analysis is the statistical method you use when both the **response** variable and the **explanatory** variable are continuous variables (i.e. real numbers with decimal places – things like heights, weights, volumes, or temperatures).

Perhaps the easiest way of knowing when regression is the appropriate analysis is to see that a scatterplot is the appropriate graphic (in contrast to analysis of variance, say, where it would have been a box-and-whisker plot or a bar chart).

Here we will cover some important kinds of regression analysis :

linear regression (the simplest, and much the most frequently used);

multiple regression (where there are numerous explanatory variables);

non-linear regression (to fit a specified non-linear model to data);

Regression

The essence of regression analysis is using sample data to estimate parameter values and their standard errors. First, however, we need to select a model which describes the relationship between the response variable and the explanatory variable(s). The simplest of all is the linear model

$$y = a + bx.$$

There are two variables and two parameters. The response variable is y, and x is a single continuous explanatory variable. The parameters are a and b: the intercept is a (the value of y when $x = 0$); and the slope is b (the change in y divided by the change in x which brought it about).

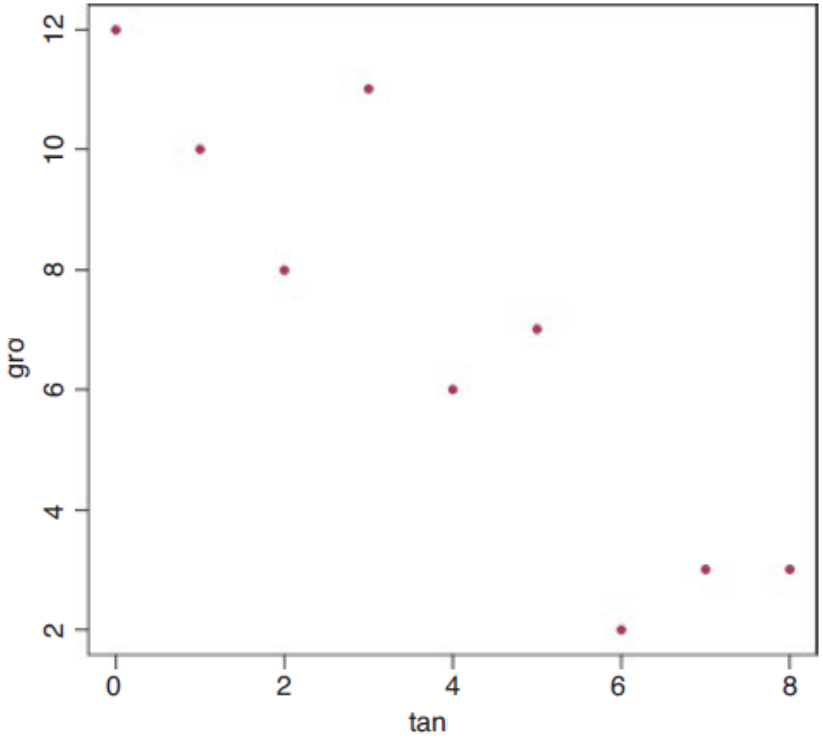
Regression

Let us start with an example which shows the growth of caterpillars fed on experimental diets differing in their tannin content:

```
reg.data <- read.table("f:/regtest.txt",header=T)
attach(reg.data)
names(reg.data)
```

[1] "gro" "tan"

```
plot(tan,gro,pch=21,col="blue",bg="red")
```



The higher the percentage of tannin in the diet, the more slowly the caterpillars grew. You can get a crude estimate of the parameter values by eye. Tannin content increased by 8 units, in response to which growth declined from about 12 units to about 2 units, a change of −10 units of growth. The slope, *b*, is the change in *y* divided by the change in *x*, so

$$b \approx \frac{-10}{8} = -1.25$$

Regression

The intercept, *a*, is the value of *y* when *x* = 0, and we see by inspection of the scatterplot that growth was close to 12 units when tannin was zero. Thus, our rough parameter estimates allow us to write the regression equation as :

$$y \approx 12.0 - 1.25x.$$

Of course, different people would get different parameter estimates by eye. What we want is an objective method of computing parameter estimates from the data that are in some sense the 'best' estimates of the parameters for these data and this particular model.

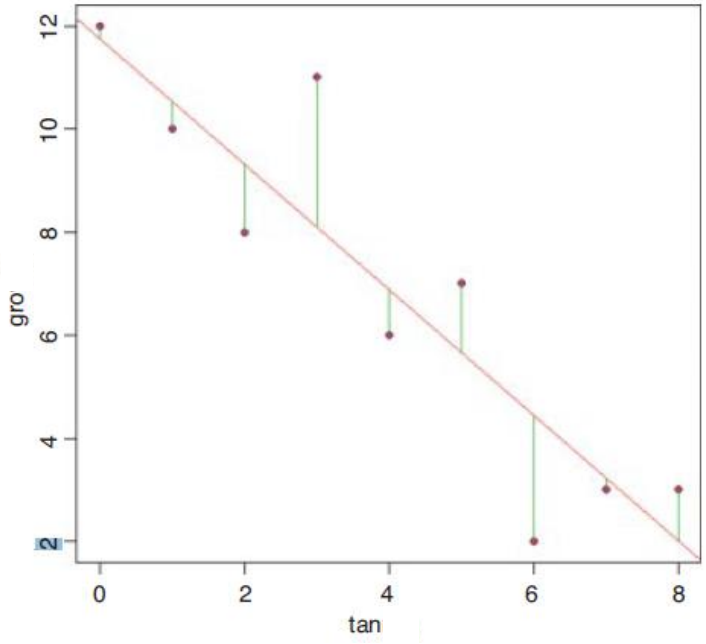
The convention in modern statistics is to use the **maximum likelihood estimates** of the parameters as providing the 'best' estimates. That is to say that, given the data, and having selected a linear model, we want to find the values of the slope and intercept that make the data most likely.

Regression

For the simple kinds of regression models with which we begin, we make several important assumptions:

The variance in *y* is constant (i.e. the variance does not change as *y* gets bigger).
The explanatory variable, *x*, is measured without error.
The difference between a measured value of *y* and the value predicted by the model for the same value of *x* is called a residual.
Residuals are measured on the scale of *y* (i.e. parallel to the *y* axis).
The residuals are normally distributed.

```
model <- lm(gro~tan)
abline(model,col="red")
yhat <- predict(model,tan=tannin)
join <- function(i)
lines(c(tan[i],tan[i]),c(gro[i],yhat[i]),col="green")
sapply(1:9,join)
```



Regression

Under these assumptions, the maximum likelihood is given by the **method of least squares**. The phrase 'least squares' refers to the residuals, as shown in the figure. The residuals are the vertical differences between the data (solid circles) and the fitted model (the straight line).

Regression

Linear Regression in R

Linear regression is a statistical procedure used to predict the value of a dependent variable, also known as **response variable**, on the basis of one or more Independent variables, also known as **predictor variables**.

There are two types of linear regressions:

1. Simple Linear Regression
2. Multiple Linear Regression

Simple Linear Regression - Simple linear regression is used to find relationship between a continuous dependent variable **Y** and a continuous independent variable **X**.

The general simple linear regression model to evaluate the value of Y for a value of X:

$$Y_i = \beta_0 + \beta_1 X + \varepsilon$$

Here, the i^{th} data point, y_i, is determined by the variable x_i; β_0 and β_1 are

regression coefficients; and ε_i is the error in the measurement of the i^{th} value of **x**.

Regression

Multiple Linear Regression in R

Multiple regression is a statistical technique used to obtain the value of a dependent variable from two or more independent or predictor variables.

A multiple linear regression model with two explanatory variables can be given as:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i} + \varepsilon_1$$

Here, the i^{th} data point, y_i, is determined by the levels of the two continuous explanatory variables X_{1i} and $X_{1i'}$ by the three parameters β_0 ,β_1 **and** β_2 of the model, and by the residual ε_1 of point i from the fitted surface.

Regression

Least Square Estimation in R

The sum of squares of residuals (SSR) is calculated for finding out the best fit line for a regression analysis. The formula for calculating SSR is as follows:

$$SSR=\Sigma e^2-\Sigma(y - (bo+b1x))^2$$

Where e is the error, y and x are the variables, and b0 and b1 are the unknown parameters or coefficients.

Regression Assumptions :

The validity of a regression model is ensured by statisticians, while building the model, on the basis of certain assumptions. These are:

Linearity: Assumes a linear relationship between the dependent and independent variables.

Independence: Assumes there is no correlation in the errors of the collected samples.

Homoscedasticity: Assumes constant variation in errors.

Normality: Assumes normal distribution of errors in the collected samples.

Regression

Multicollinearity in R

Multicollinearity is a non-linear relationship between two explanatory variables, leading to inaccurate parameter estimates. Two or more variables representing an exact/approximate linear relationship cause multicollinearity.

Multicollinearity can be detected by calculating VIF with the help of the following formula:

$$\mathbf{VIF = 1/(1-R^2)}$$

Here, Ri is the regression coefficient for the explanatory variable xi, with respect to all other explanatory variables.

Regression

Working with Linear Regression

Two numerical variables, X and Y, having at least a moderate correlation, been established through both correlation and scatterplot, are in a linear relationship.

Regression line is used to predict the value of Y for a given value of X.

X and Y are called explanatory and response variables or independent and dependent variables, respectively. The condition of linearity is checked by creating scatterplot that must form a linear pattern.

The regression line is calculated by using the following formula:

$$Y = mx + b$$

Where, **m** is the slope of the line and **b** is the y-intercept. The best fit line is obtained using the method of **least squares**, which takes the line with least possible sum of squares of errors (**SSE**).

During the estimation of the value of Y, the slope can be calculated by multiplying the correlation between X and Y with the division of the standard deviation of y-values by the standard deviation of x-values.

Regression

Working with Linear Regression

The y-intercept, b, of the best fit line is obtained by subtracting the product of slope and mean of x-values from the mean of y-values.

In the context of regression, the slope is interpreted from the change in y-values with respect to change in x-values.

The y-intercept, which is sometimes meaningful and sometimes not, is the place where the regression line crosses the Y-axis, where x=0.

Regression

Simple Linear Regression in R

Simple calculators are not fit for large and complex calculations involved in linear regression, therefore we require a tool for the purpose. R is a popular tool used in these cases. There are five famous functions in R, which are mentioned as follows:

Famous Five Functions in R	Explanation
sum(x)	Calculates the sum of all x values.
sum(y)	Calculates the sum of all y values.
sum(x^2)	Calculates the sum of the squares of all the values of x.
sum(y^2)	Calculates the sum of the squares of all the values of y.
sum(xy)	Calculates the sum of the product of each respective values of x and y.

Regression

The famous five functions are used in various calculations that are performed in regression.

One of the calculations in regression is calculating of **corrected sum of squares**. The formula for calculating sum of squares of x is:

$$SSX=\Sigma x^2-((\Sigma x)^2/n)$$

The other sums of squares can be calculated in a manner similar to above.

The calculation of sum of products uses the following formula:

$$SSXY=\Sigma xy-((\Sigma x)(\Sigma y)/n)$$

The degree of scatter is calculated as the sum of squares of errors (SSE) by using a formula as the following:

$$SSE=\Sigma(y-a-bx)^2$$

The other calculations in regression are the analysis of variance and unreliability estimates for parameters. After calculating the values, predictions can be made.

Regression

Linear Model Results Objects :

Simple linear regression is similar to correlation with one predictor and a response variable. We use **lm()** function for this kind of linear modeling in R.

A dataset, named **grades**, having two columns that can be correlated, is taken to implement the **lm()** and **summary()** functions:

Using the **lm()** Function
Performs linear regression analysis for the grade and gpa data.

```
> grades.f = lm(grade ~ gpa, data = grades)
>
> grades.f

Call:
lm(formula = grade ~ gpa, data = grades)

Coefficients:
(Intercept)          gpa
   10.48807      -0.07855
```

Regression

Linear Model Results Objects :

Using the **summary()** Function
Stores result in the grades.f object

```
> summary(grades.f)

Call:
lm(formula = grade ~ gpa, data = grades)

Residuals:
      1       2       3       4       5       6
 0.6565 -0.3867 -0.3695 -0.3530 -0.2022  0.6549

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.48807    0.64152  16.349 8.19e-05 ***
gpa         -0.07855    0.30326  -0.259    0.808
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5726 on 4 degrees of freedom
Multiple R-squared:  0.01649, Adjusted R-squared:  -0.2294
F-statistic: 0.06708 on 1 and 4 DF,  p-value: 0.8084
```

Using the **names()** Command

```
> names(grades.f)
 [1] "coefficients"  "residuals"     "effects"       "rank"          "fitted.values" "assign"
 [7] "qr"            "df.residual"   "xlevels"       "call"          "terms"         "model"
>
```

Regression

The components can be extracted by using the **$** syntax, as follows:

Using the $ Syntax to Extract Components

Finds the coefficients in the regression analysis.

```
> grades.f$coefficients
(Intercept)         gpa
10.48806852 -0.07854578
>
```

Regression

Nonlinear Regression in R

Introduction to Nonlinear Regression Analysis :

Nonlinear Regression and **Generalized Linear Models** :

The **nonlinear regression analysis** is the process of building a nonlinear function for predicting the outcome of a dependent variable on the basis of independent variables with the help of model parameters that depend on the degree of relationship among variables.

Generalized Linear Models are commonly applied in the following types of regressions:

- Count data expressed as proportions (e.g. logistic regressions)
- Count data that are not proportions (e.g. log linear models of counts)
- Binary response variables (e.g. "yes/no", "day/night", "sleep/awake", buy/not buy)
- Data showing a constant coefficient of variation (e.g. time data with gamma errors)