Modeling

Nik Bear Brown

In this module, we learn what a model is, and how to create statistical models. We start by building and interpreting simple models with linear regression. We then learn to interpret the diagnostics and quality of our models. We expand our modeling tool box with multivariate analysis. We finish by using logistic regression to predict probabilities and categories.

In the first lesson, we study the theory behind modeling, and learn what a model is.

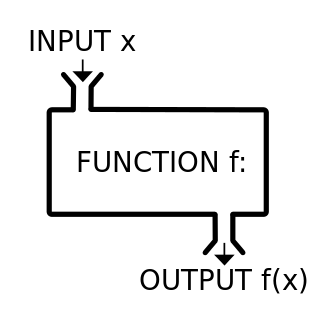
## Modeling Basics

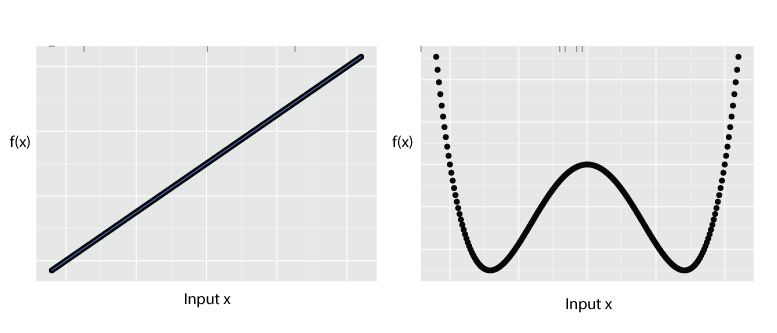
*What is a Model?*

A statistical model embodies a set of assumptions concerning the generation of the observed data, and similar data from a larger population. A model represents, often in considerably idealized form, the data-generating process. The model assumptions describe a set of probability distributions, some of which are assumed to adequately approximate the distribution from which a particular data set is sampled. A model is usually specified by mathematical equations that relate one or more random variables and possibly other non-random variables. As such, "a model is a formal representation of a theory" (Herman Adèr quoting Kenneth Bollen).

<https://en.wikipedia.org/wiki/Statistical_model>

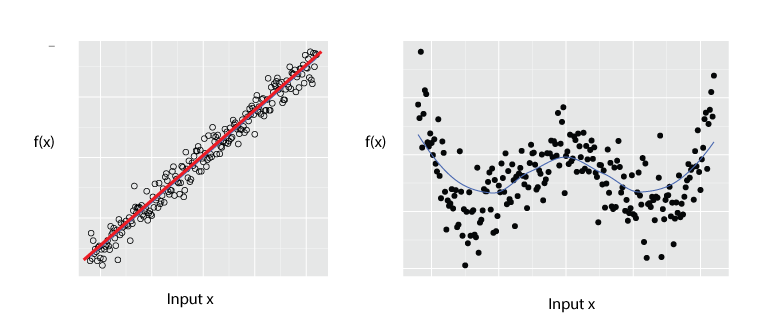
A function is a relationship between two variables. A function can be defined by any mathematical condition relating each argument (input value) to the corresponding output value.





## Regression

Regression fits a model to data. The goal is to find a functional relation between the response variable y and the predictor variable(s) x.



In statistics, regression analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables (or 'predictors').

from [Regression analysis Wikipedia](https://en.wikipedia.org/wiki/Regression_analysis)

Regression Analysis was first developed by Sir Francis Galton, who studied the relation between heights of sons and fathers. Heights of sons of both tall and short fathers appeared to “revert” or “regress” to the mean of the group. Later in this module (lesson 2) we will use R to recreate Galton's original regression analysis.

*Names for f(x)/output and x/input variables*

Some terminology for the f(x)/output and x/input:

|  |  |
| --- | --- |
| f(x) | x |
| response | ~ predictor |
| response | ~ explanatory |
| dependent | ~ independent |
| outcome | ~ predictor |
| forecast | ~ predictor |
| regressand | ~ regressor |
| explained | ~ explanatory |

## The Modeling process

1. EDA - exploritory data analysis.
2. Cleaning the data.
3. Formula (i.e. what amd which )
4. Fit (i.e. Estimate the unknown parameters for the model.)
5. Analysis of fit (i.e. how good is the model)

Usually many models are created so 3,4, & 5 are often repeated. We've discussed exploritory data analysis and data Cleaning in module one.

*Formula (i.e. what amd which )*

The functional form of the regression model (linear or non-linear) must be decided once the data have been collected and analyzed. Selection of independent variables must be reduced whenever we construct models, only a limited number of independent or predictor variables can or should be included in a regression model. Therefore a central problem is that of choosing the most important predictor variables.

Choosing predictors is typically done in 3 ways:

1. Theory / domain knowledge
2. EDA - exploritory data analysis.
3. Fitting various combinations of outcome and predictors and selecting the best fits (r.g. Stepwise regression)

The selection of data is crucial as well as linear regression is sensitive to outliers, so the modeling process often leads to further data cleaning/

*Estimate the unknown parameters for the model*

The R software handles the estimation of the unknown parameters in a model. This is usally just a matter of specifiying the model correctly syntactically.

*Model Syntax in R*

* fx ~ x
* y ~ x

and for multivariate models

* fx ~ x1 + x2

e.g. Income ~ Education + Gender + Race

Note that many, many models in R use this syntax

* m1 <- lm(Income ~ Education + Gender + Race,data=wages)
* m2 <- lm(Income ~ Education + Gender + Race,data=wages)
* m3 <- aov(Income ~ Education,data=wages)

These models are typcially polynomial linear, quadratic, etc. (i.e. linear and non-linear)

*Interaction terms in R*

Interaction terms in R are specified with the following syntax

* fx ~ x1 + x2 + x1:x2

or equivelently

* fx ~ x1\*x2

3rd order interections are specified with the following syntax

( x1 + x2 + x3)^3

or more generally

3rd order interections are specified with the following syntax

( x1 + x2 + x3 ... xn)^n

We will discuss interaction terms more the the next lesson but if you would like to look ahead see [Visual interpretation of interaction terms in linear models with ggplot](http://www.r-bloggers.com/visual-interpretation-of-interaction-terms-in-linear-models-with-ggplot-rstats/)

## Linear Models

Linear regression predicts the response variable assuming it has a linear relationship with predictor variable(s) or .

*Simple* regression use only one predictor variable . *Mulitple* regression uses a set of predictor variables .

The *response variable* is also called the regressand, forecast, dependent or explained variable. The *predictor variable* is also called the regressor, independent or explanatory variable.

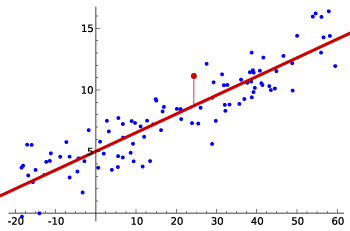
The parameters and determine the intercept and the slope of the line respectively. The intercept represents the predicted value of when . The slope represents the predicted increase in resulting from a one unit increase in .

Note that the regression equation is just our famliar equation for a line with an error term.

The equation for a line:

The equation for a line with an error term:

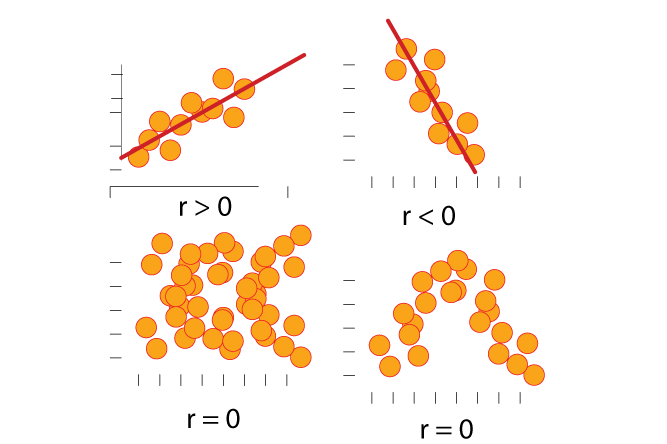
* = = slope
* = = intercept
* = error term



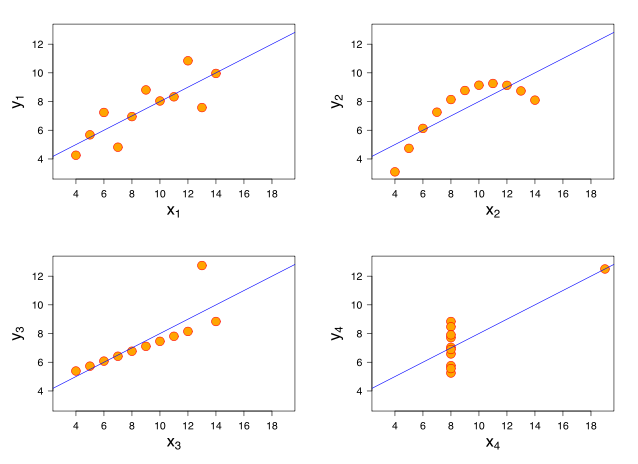
We can think of each observation consisting of the systematic or explained part of the model, , and the random *error*, .

## The slope

*Linear Relationships*



We can have a *positive linear relationship* (r>0), *negative linear relationship* (r<0), or *no linear relationship* (r=0) (Note that no linear relationship doesn't mean no relationship.)



Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

from [Anscombe's quartet - Wikipedia](https://en.wikipedia.org/wiki/Anscombe%27s_quartet)

*Zero Slope*

Note that when then response does not change as the predictor changes.

## The error

The error term is a catch-all for anything that may affect other than . We assume that these errors:

* have mean zero; otherwise the forecasts will be systematically biased.
* statistical independence of the errors (in particular, no correlation between consecutive errors in the case of time series data).
* homoscedasticity (constant variance) of the errors.
* normality of the error distribution.

If any of these assumptions is violated then the robustness of the model to be taken with a grain of salt.

## Least squares estimation

In a linear model, the values of and . These need to be estimated from the data. We call this *fitting a model*.

The least squares method iis the most common way of estimating and by minimizing the sum of the squared errors. The values of and are chosen so that that minimize

Using mathematical calculus, it can be shown that the resulting **least squares estimators** are

and

where is the average of the observations and is the average of the observations. The estimated line is known as the *regression line*.

## Fitted values and residuals

The response values of obtained from the observed values are called *fitted values*: , for . Each is the point on the regression line corresponding to .

The difference between the observed values and the corresponding fitted values are the *residuals*:

The residuals have some useful properties including the following two:

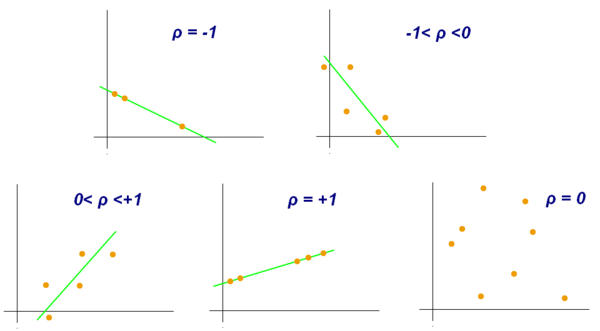
Residuals are the errors that we cannot predict.Residuals are highly useful for studying whether a given regression model is an appropriate statistical technique for analyzing the relationship.

## Regression and correlation

The correlation coefficient measures the strength and the direction of the linear relationship between the two variables. The stronger the linear relationship, the closer the observed data points will cluster around a straight line.

The *Pearson product-moment correlation coefficient* is the most widely used of all correlation coefficients. In statistics, the Pearson product-moment correlation coefficient (/ˈpɪərsɨn/) (sometimes referred to as the PPMCC or PCC or Pearson's r) is a measure of the linear correlation (dependence) between two variables X and Y, giving a value between +1 and −1 inclusive, where 1 is total positive correlation, 0 is no correlation, and −1 is total negative correlation. It is widely used in the sciences as a measure of the degree of linear dependence between two variables. It was developed by Karl Pearson from a related idea introduced by Francis Galton in the 1880s. Early work on the distribution of the sample correlation coefficient was carried out by Anil Kumar Gain and R. A. Fisher from the University of Cambridge.

from [Pearson product-moment correlation coefficient](https://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient)



Examples of scatter diagrams with different values of correlation coefficient (ρ)

The value of r is such that -1 < r < +1.

Strong positive correlation r is close to +1. Strong negative correlation r is close to -1. No correlation r is close to 0.

The advantage of a regression model over correlation is that it asserts a predictive relationship between the two variables ( predicts ) and quantifies this in a useful way for forecasting.

## Coefficient of determination

In statistics, the coefficient of determination, denoted and pronounced R squared, is a number that indicates how well data fit a statistical model – sometimes simply a line or curve. It is a statistic used in the context of statistical models whose main purpose is either the prediction of future outcomes or the testing of hypotheses, on the basis of other related information. It provides a measure of how well observed outcomes are replicated by the model, as the proportion of total variation of outcomes explained by the model.

from [Coefficient of determination -Wikipedia](https://en.wikipedia.org/wiki/Coefficient_of_determination)

The slope coefficient can also be expressed as

where is the standard deviation of the observations and is the standard deviation of the observations.

The coefficient of determination is not to be confused with but is related to the Pearson product-moment correlation coefficient . With simple linear regression (i.e., only 1 covariate), the slope is the same as Pearson's r if .

The regression slope measures the "steepness" of the linear relationship between two variables and can take any value from .

Note that there are several definitions of that are only sometimes equivalent. Sometime "little" or is simply the square of the sample correlation coefficient (i.e., ).

A common definition for the coefficient of determination or :

where the summations are over all observations.

can be thought of as \*the proportion of variation in the forecast variable that is accounted for (or explained) by the regression model\*\*

In the definition of , as is similar to he value of (the square of the pearson correlation between and .

## AIC, BIC, Mallows's Cp

Beside's there are many tests used to assess the fit of a regression model that has been estimated using ordinary least squares. always get smaller as more variables are added to a model, which is a form of overfitting. Some analysts used other statistical tests to assess the fit of a regression model. The most common being the *Akaike information criterion (AIC)*, the *Bayesian information criterion (BIC)*, and

*Akaike information criterion*

The Akaike information criterion (AIC) is a measure of the relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Hence, AIC provides a means for model selection.

AIC is founded on information theory: it offers a relative estimate of the information lost when a given model is used to represent the process that generates the data. In doing so, it deals with the trade-off between the goodness of fit of the model and the complexity of the model.

* from [Akaike information criterion - Wikipedia](https://en.wikipedia.org/wiki/Akaike_information_criterion)

*Bayesian information criterion (BIC)*

In statistics, the Bayesian information criterion (BIC) or Schwarz criterion (also SBC, SBIC) is a criterion for model selection among a finite set of models; the model with the lowest BIC is preferred. It is based, in part, on the likelihood function and it is closely related to the Akaike information criterion (AIC).

When fitting models, it is possible to increase the likelihood by adding parameters, but doing so may result in overfitting. Both BIC and AIC resolve this problem by introducing a penalty term for the number of parameters in the model; the penalty term is larger in BIC than in AIC. The BIC was developed by Gideon E. Schwarz and published in a 1978 paper, where he gave a Bayesian argument for adopting it.

* from [Bayesian information criterion (BIC) - Wikipedia](https://en.wikipedia.org/wiki/Bayesian_information_criterion)

*Mallows's Cp*

In statistics, Mallows's Cp, named for Colin Lingwood Mallows, is used to assess the fit of a regression model that has been estimated using ordinary least squares. It is applied in the context of model selection, where a number of predictor variables are available for predicting some outcome, and the goal is to find the best model involving a subset of these predictors.

* from [Mallows's Cp- Wikipedia](https://en.wikipedia.org/wiki/Mallows%27s_Cp)

We won't discuss which to use (although R makes it easy to use them all!). Rather just point out that they are all used to assess the fit of a regression model that has been estimated using ordinary least squares. Also note that high (i.e close to 1) is "good" for and low (i.e close to 0) is "good" for AIC, BIC, Mallows's Cp.

|  |  |
| --- | --- |
| Test | "Good" |
|  | High |
| AIC | Low |
| BIC | Low |
| Mallows's Cp | Low |

## Evaluating the regression model

To evaluate a regression model we ask the following questions:

A. Does it make sense?  
B. Is the "true" significantly differnet from ?  
C. Are any assumptions of the model violated?  
D. How tightly the parameter estimation fits the residuals?

## Hypothesis testing: Is he "true" ?

Recall that when the slope we have no relationship between the outcome and predictors.

Hypothesis tests assume the thing you want to disprove, and then to look for evidence that the assumption is wrong. In this case, we assume that there is no relationship between and . This is called the *null hypothesis* and is stated as

Evidence against this hypothesis is provided by the value of , the slope estimated from the data. If is very different from zero, we conclude that the null hypothesis is incorrect and that the evidence suggests there really is a relationship between and .

There are many hypothesis tests that can be used to test whether the "true" :

* Student’s T-Tests
* One-Sample T-Test
* Two-Sample T-Test
* Paired T-Test
* Wilcoxon Rank-Sum Test
* Analysis of Variance (ANOVA)
* Kruskal-Wallis Test
* Multiple Comparison Methods
* Tukey’s HSD Test
* Other Pairwise T-Tests
* Pairwise Wilcoxon Rank-Sum Tests
* Hypothesis Tests for Variance
* F-Test
* Bartlett’s Test

We will discuss these more in the module on hypothesis testing. As R's lm() function gives p-values by default we will focus on them.

## P-value

To determine how big the difference between (the "true" ) and must be before we would reject the null hypothesis, we calculate the probability of obtaining a value of as large as we have calculated if the null hypothesis were true. This probability is known as the *P-value*.

In statistics, the p-value is a function of the observed sample results (a statistic) that is used for testing a statistical hypothesis. Before the test is performed, a threshold value is chosen, called the significance level of the test, traditionally 5% or 1% and denoted as .

If the p-value is equal to or smaller than the significance level (), it suggests that the observed data are inconsistent with the assumption that the null hypothesis is true and thus that hypothesis must be rejected (but this does not automatically mean the alternative hypothesis can be accepted as true). When the p-value is calculated correctly, such a test is guaranteed to control the Type I error rate to be no greater than .

from [P-value](https://en.wikipedia.org/wiki/P-value)

## Confidence intervals

In statistics, a confidence interval (CI) is a type of interval estimate of a population parameter. It provides an interval estimate for lower or upper confidence bounds. For , usually referred to as a *confidence interval* and is typically +/-0.5% (a 99% confidence interval),+/-1% (a 98% confidence interval),+/-2.5% (a 95% confidence interval) or +/-5% (a 90% confidence interval). The lower and upper confidence bounds need not be equal, and they can be any number such that the confidence interval not exceed 100%.

## Residual plots

The error term has the following assumptions:

* have mean zero; otherwise the forecasts will be systematically biased.
* statistical independence of the errors (in particular, no correlation between consecutive errors in the case of time series data).
* homoscedasticity (constant variance) of the errors.
* normality of the error distribution.

Plotting the residuals can asses whether (or how much) these assumptions were violated. We will use R to generate residual plots in lesson 2.

## Outliers

Observations that take on extreme values compared to the majority can strongky effect the least squares estimators:

and

Plotting and occasionally removing outliers and refitting is part of the modeling process.

## Standard error of the regression

How well the model has fitted the data can be thought of as how "tightly" the date fit the regression line. That is, the spread, variance or standard deviation of the residuals.This spread between fitted and actual values is usually known as the *standard error of the regression*:

Here, we divide by because we have estimated two parameters (the intercept and slope) in computing the residuals. Normally, we only need to estimate the mean (i.e., one parameter) when computing a standard deviation. The divisor is always minus the number of parameters estimated in the calculation.

Note that we can (and should) visualize the predicted vs actual values as this gives more information about the homoscedasticity (constant variance) of the errors.

## Forecasting with regression

Forecasts from a simple linear model are obtained using the equation

where is the value of the predictor for which we require a forecast.

Given a new unobserved x we can easily predict given our estimates

## Assingments

There is no assingment with this lesson. It is theory to prepare for the next lesson.

## Resources

* [Regression in R Part I : Simple Linear Regression - UCLA](http://scc.stat.ucla.edu/page_attachments/0000/0139/reg_1.pdf)
* [Simple Linear Regression](http://www.r-tutor.com/elementary-statistics/simple-linear-regression)
* [Simple Linear Regression | R Tutorial](http://ww2.coastal.edu/kingw/statistics/R-tutorials/simplelinear.html)
* [R Tutorials--Multiple Regression](http://ww2.coastal.edu/kingw/statistics/R-tutorials/multregr.html)
* [Getting Started in Linear Regression using R - Princeton](http://www.princeton.edu/~otorres/Regression101R.pdf)