SF2930: Regresion analysis
Lecture 1
Simple linear regression.

Tatjana Pavlenko

17 January 2018

What is regression? Introduction

Regression analysis is a statistical technique for investigating and modeling the relationship between variables. The reason why it is so widely applied is because it provides the answer to an everyday question, namely

how a response variable of special interest depends on several other, explanatory variables.

What is regression? Examples

Applications of regression techniques are numerous; example include engineering, medical, biological and social sciences, physical and chemical sciences, economics, ... Regression is probably is the most widely used statistical methodology. Examples of applied problems and questions in which regression might be helpful:

- how the apartment prise depends on size, location, floor, closeness to subway, ...
- how growing of plants depends on fertilizer, soil quality, ...
- how home insurance premium depends on age of homeowner, value of the home and its contents, region, ...

In various quantitative settings, the regression techniques models the relationship between the *response variable* of special interests (Y) and a (set) $x_1, \dots x_k$ of *explanatory* or *predictor* variables.

KTH Matematik

Regression mathematics

- Linking the response variable to the predictors:

$$Y = \underbrace{f(x_1, \ldots, x_k)}_{\text{deterministic}} + \underbrace{\varepsilon}_{\text{random}},$$

  ε is the error term which is can neither be controlled or predicted.
- The goal is to learn about the function $f(\cdot)$.
- In full generality, finding $f(\cdot)$ without any conditions is very difficult: function space is infinite-dimensional!
- Solution: restrict the form of $f(\cdot)$.
- *Linear modeling:* $Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$.
- Finding f boils down to determining $\beta_0, \beta_1, \ldots, \beta_k$ from the data.

KTH Matematik

Regression and model building

- Why modeling? It is stated once by George Box that *All models are wrong, but some are useful*.
- Certain degree of variability (uncertainty) is present in almost all processes. Most of this variability can not be modeled with deterministic methods of mathematics. Statistics (mathematics under uncertainty) provides a powerful toll of modeling phenomena under uncertainty.
- Larger data collected in real-world applications demand models in order to *extract big knowledge of big data*!
- In this course we start with simple but highly applicable, *linear* regression models. We then turn to more advance subjects.

KTH Matematik

Regression and model building (cont.)

In this course we start with highly applicable, *simple* linear regression models. It is rather simple modes but the tools constructed within the simple linear regression will be naturally extended to the multiple case.

Later in the course we turn to more advance subject such as

- multiple linear regression
- logistic regression
- Generalized linear models containing Poisson and negative binomial regression

KTH Matematik

Example

Figure: The wash water tank on SAS aircrafts. It is impossible to predict exactly the amount of wash water needed, therefore the tank is always filled to 100% at Arlanda airport. The project on minimizing the amount of water is performed with the main goal to lower the weight of the aircraft, and as a result reduce fuel consumption and cost. Goal: to investigate (to model!) the relationship between wash water consumption (target variable) and number of passengers, duration of a flight, time of the flight (night/day), ... (explanatory variables). Statistical approach: multiple linear regression analysis.

KTH Matematik

EXAMPLE: LINEAR REGRESSION FOR ADVERTISING DATA

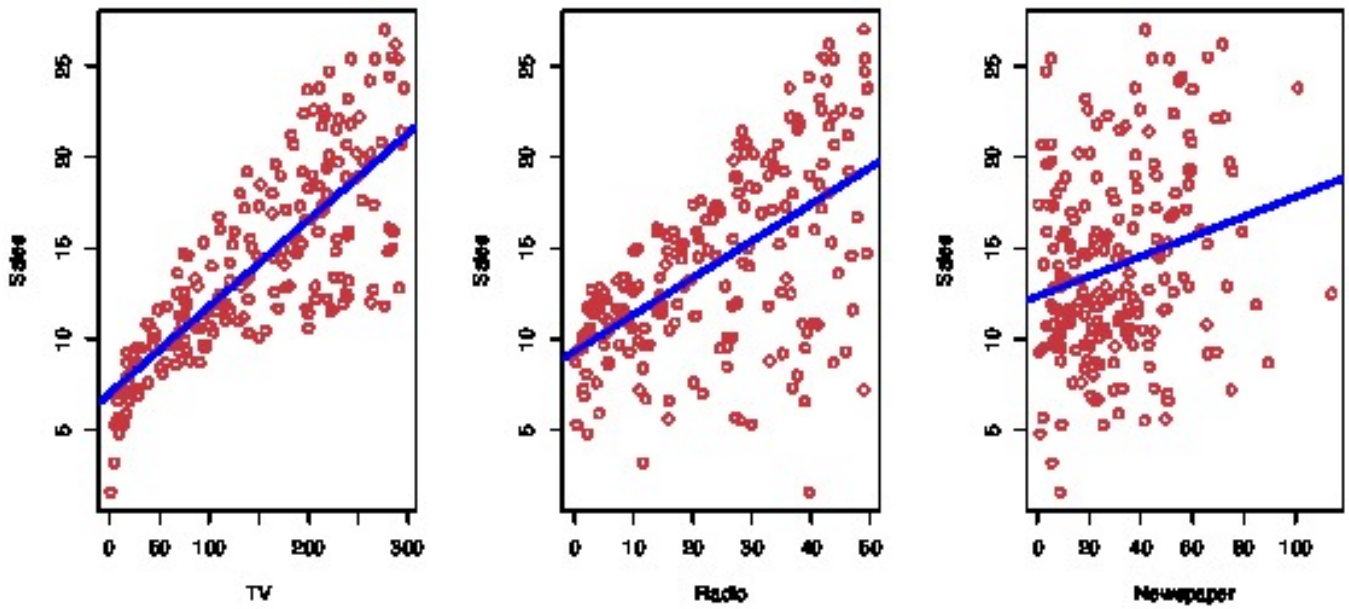


FIGURE: The `Advertising` data set. The plots displays `sales` for a particular product as a function of advertising budget for `TV`, `radio` and `newspaper`, for 200 different markes. Each blue line represents the simple model that can be used to predict `sales` from each feature variable.

KTH VETENSKAP OCH KONST
KTH Matematik

Linear regression for advertising data.

`Advertising` data, see Ch. 2 in ISL. The goal is using the data, to design a marketing plan (for the next year) that will result in high product sales. Specific questions are (see Intro to Ch. 3 in ISL):

- Q1: Is there a (linear) relationship between advertising budget and sales?
- Q2: How strong is the relationship between advertising budget and sales?
- Q3: Which media contribute to sales (Which predictors are associated with response)?
- Q4: How to estimate effect of each variable on sales? How accurate are these estimates?
- Q5: How can we predict future sales?
- Q6: Is the synergy among the advertising media?

KTH Matematik

The goals of such regression analysis are usually two-fold:

- *to model* the relation between output Y and input vector $\mathbf{x} = (x_1, \ldots, x_k)$, and to specify which of the predictor variables have effect on the output variable Y, as well as to quantify the strength of this relation (The regression equation, inference).
- Using the regression equation, *to predict* the expected response value of Y for an arbitrary (new) configuration of $x_1, \ldots, x_k$.
- The dependence of Y on $\mathbf{x} = (x_1, \ldots, x_k)$ is assumed to be *linear*, i.e. the *model* formula is

$$Y = \beta_0 + \beta_1 x_1 + \cdots \beta_k x_k + \varepsilon,$$

where $\beta_0, \ldots, \beta_k$ are *unknown* coefficients or parameters, and ε is the random error, which accounts for measurement error and effect of other variables not explicitly considered in the model.

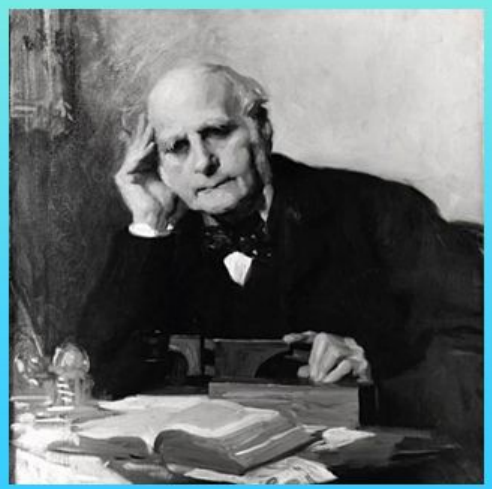
Goals in regression analysis

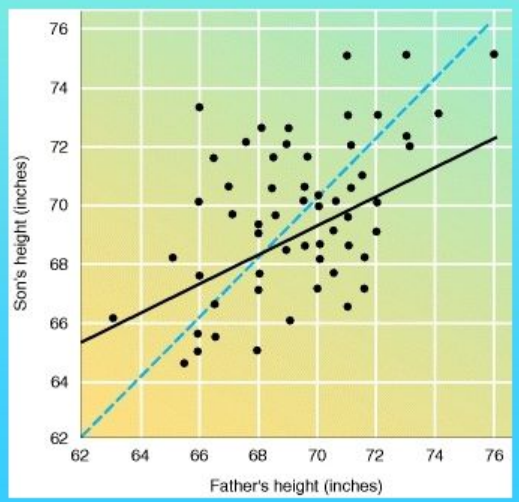
- **A "good" fit.** Estimating a (hyper)-plane over $x_1, \ldots, x_k$ to explain the response such that the errors are "small". The standard tool is **L**east **S**quares estimates, LS.
- **Good parameter estimates** are useful to describe the change of the response when varying some some predictor variable(s).
- **Good prediction** is useful to predict a new response as a function of new predictor variables.
- Uncertainties and significance with the three goals above.
- Development of a **good model**: In an interactive process, with the help of methods for the above goals, we may change parts of the model to come up with a better mode.

Why call regression?

History

- Developed by Sir Francis Galton (1822-1911) in his article "Regression towards mediocrity in hereditary structure"





KTH Matematik

How regression got its name?

- In the late 1800's, a scientist Sir Francis Galton was working with large observational studies on humans, in part with the data on the heights of fathers and first sons.
- In terms of simple regression model x was the height of the father and Y was the height of the first, fully grown, son. Goal: to predict Y in terms of x.
- Galton concluded: *For father's whose heights were taller than the average, the LS line predicts the son to be shorter than the father. Likewise, for father's whose heights were shorter than the average, the line predicts the son to be taller than the father.*

There is a **regression** towards the mean effect!

KTH Matematik

Simple linear regression

We begin with a mathematically most simple way of describing the relation between the variables: the *linear* relationship between a continuous response variable Y in dependence on

- a *single* explanatory variable x (*simple* linear regression model)
- *several* explanatory variables $x_1, \ldots, x_k$ (*multiple* regression model)

Later in the course we will consider

- modeling of the nonlinear relationship between a binary response Y and a set of explanatory variables $x_1, \ldots, x_k$ (logistic regression).
- additional models when Y is discrete.

Simple linear regression

The simple linear regression model is $Y = \beta_0 + \beta_1 x + \varepsilon$, where the *intercept* β_0 and the *slope* β_1 are unknown constants, and ε is a random error component.

Model, interpretation of variables and parameters:

- Y is a **continuous** dependent variable, assumed to be random and called for *response* or *outcome*.
- x is independent variable assumed to be **non-random**, i.e. we focus on a *fixed x-case*. (Random X-case will be considered later in the course).
- For the random error ε, assume that $E(\varepsilon) = 0$ and $V(\varepsilon) = \sigma^2$ (σ^2 is unknown)
- β_0 and β_1 are called *regression coefficients*, assumed non-random.
- $E(Y) = \beta_0 + \beta_1 x$ and $V(Y) = \sigma^2$, because β_0, β_1 and x are non-random.

KTH Matematik

BRING THE *data* INTO CONSIDERATION

Given is a set of paired data points (y_i, x_i) obtained for $i = 1, \ldots, n$ observational units; each y_i is the observed value of a r.v. Y_i, i.e. a generic observation i is modeled as

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

Besides linearity of the model, we assume that

- The linear regression equation is **correct**: $E(\varepsilon_i) = 0$ for all i.
- All x's are **exact**, i.e we can observe x_i's perfectly.
- The variance of the errors is **constant** (*homoscedasticity*), i.e. $V(\varepsilon_i) = \sigma^2$ for all i.
- ε_i are pairwise **uncorrelated**, i.e $Cov(\varepsilon_i, \varepsilon_j) = 0$ for all $i \neq j$.
- Since the only random element in the model is ε_i, Y_i's have also $E(Y_i) = \beta_0 + \beta_1 x_i$ and **common variance** σ^2.
- **For purpose of making tests of significance**, we assume that ε_i are iid $N(0, \sigma^2)$. Then Y_i are iid and $Y_i \in N(\beta_0 + \beta_1 x_i, \sigma^2)$.

KTH Matematik

A LINEAR FIT

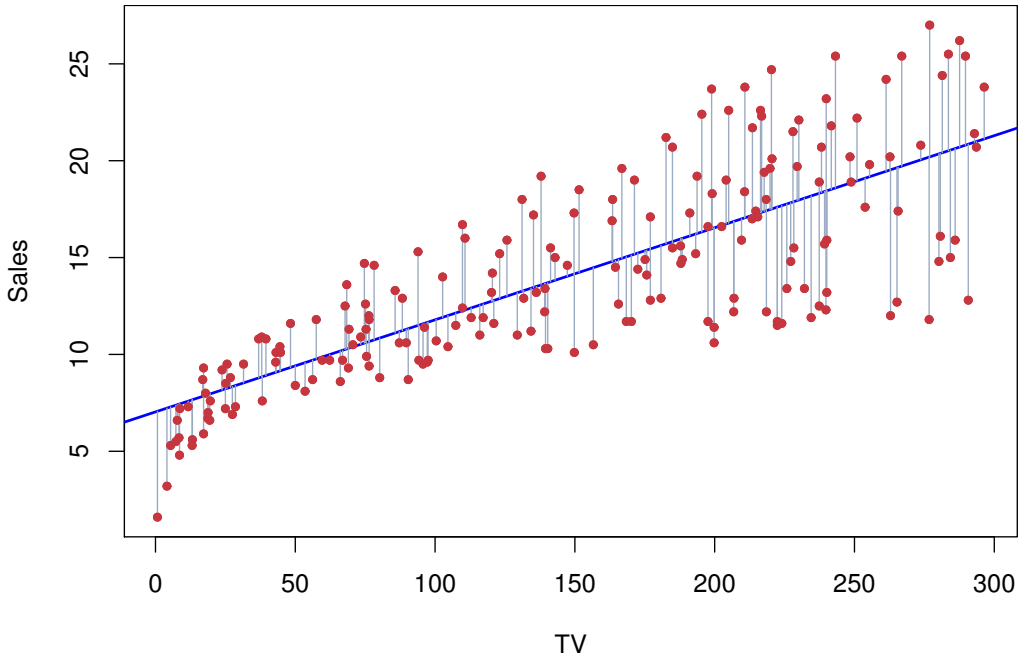


FIGURE: Drawing the linear regression line for `Advertising+` data (see Ex 1 in Ch 3 of ISL). The plot displays the regression of `sales+` onto `TV+`. The linear fit is obtained by minimizing the sum of squared errors. Each gray vertical line represents an error and the fit make a *compromise by averaging* the squares of errors.

Least Squares fit

- Example with `Advertising+` above.
- Q: How to fit a straight line that fits the data well?
- Our strategy is to fit the line in such a way such that the *squared errors are minimal*. This is called Least Squares (LS) fitting.
- Residuals vs. errors. The *residual* $e_i = y_i - \hat{y}_i$ is the the difference between the observed and the fitted y-value for the ith observation. Residuals are numerical realizations of random errors ε_i.
- Illustration of residuals: see white board.

LS fitting strategy: *to fit the regression line for which the sum of squared residuals is minimized, i. e.* $\sum_{i=1}^{n} e_i^2 \to \min$.

KTH Matematik

Least Squares fit (cont.)

More precisely, given a set of observations (y_i, x_i), $i = 1, \ldots, n$ the goal is to obtain estimators of β_0 and β_1, (say $\hat{\beta}_0$ and $\hat{\beta}_1$) which minimize the *LS objective function*

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2.$$

Solution strategy: taking partial derivatives on $S(\beta_0, \beta_1)$ with respect to β_0 and β_1, and setting them to zero.

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0}\Big|_{\hat{\beta}_0, \hat{\beta}_1} = 0, \quad \frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1}\Big|_{\hat{\beta}_0, \hat{\beta}_1} = 0.$$

This results in a system of linear equations, called for *normal equations*. These can be solved simultaneously to yield computing formulas for the *LS estimates*, $\hat{\beta}_0$ and $\hat{\beta}_1$ in terms of the observed data (y_i, x_i), $i = 1, \ldots, n$.

KTH Matematik

Least Squares solution

The LS estimates of β_0 and β_1 are computed as

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \quad \hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}.$$

The fitted model: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Residuals: $e_i = y_i - \hat{y}_i =$ observed $-$ predicted.

$$SS_{\text{Res}} = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

A convenient working formula: $SS_{\text{Res}} = SS_{\text{T}} - \hat{\beta}_1 S_{xy}$, where $SS_{\text{T}} = \sum_{i=1}^n (y_i - \bar{y})^2$. See mathematics on the board.

We also obtain the **estimate of the variance of the error term** (residual mean square – measure of residuals variability):

$$\hat{\sigma}^2 = s^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y})^2 = \frac{1}{n-2} \sum_{i=1}^n e_i^2 = \frac{SS_{\text{Res}}}{n-2} = MS_{\text{Res}}.$$

Properties of LS estimators

The LS estimators of β_0 and β_1 have several important properties:

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are **linear combinations** of the observations y_i:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \sum_{i=1}^{n} c_i y_i, \; c_i = \frac{x_i - \bar{x}}{S_{xx}}, \; i = 1, \ldots, n.$$

- Assuming that the model is correct i.e. $E(Y_i) = \beta_0 + \beta_1 x_i$, $\hat{\beta}_0$ and $\hat{\beta}_1$ are **unbiased**:

$$E(\hat{\beta}_0) = \beta_0, \quad E(\hat{\beta}_1) = \beta_1$$

- The variances of $\hat{\beta}_0$ and $\hat{\beta}_1$ are found as (see math and interpretation of $V(\hat{\beta}_1)$ on the board)

$$V(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right), \quad V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}.$$

KTH Matematik

Optimality of LS estimators

Gauss-Markov theorem states the result on the mathematical optimality of the LS: assuming that the following conditions are met,

- $E(\varepsilon_i) = 0$, (i.e. the relation is a straight line),
- $V(\varepsilon_i) = \sigma$ for $i = 1, \ldots, n$,
- $Cov(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j = 1, \ldots, n$, (the errors are uncorrelated),

the LS estimators are unbiased and have minimum variance when compared with all other unbiased estimators that are linear combinations of y_i.

- By Gauss-Markov thm, LS estimators are Best Linear Unbiased Estimators, (BLUE), where *best* means minimum variance.
- Later on in the course: Gauss-Markov theorem for the more general *multiple* regression, of which simple linear regression is a special case.

Benefits of LS approach

Mathematically ...

- The fitted line goes through the center of gravity $(\bar{x}, \bar{y})$ of the data.
- The LS technique is simple in the sense that the solutions, $\hat{\beta}_0$ and $\hat{\beta}_1$, are obtained in closed form as functions of (x_i, y_i).
- $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators
- Large sample properties of the LS estimates have some deeper mathematical advantages (approximate normality of coefficient estimators), and exact properties assuming Gaussian errors. Will be presented later in the course.

KTH Matematik