Regression Analysis

Chapter 0. Introduction

Prof. dr. Thomas Neyens

Course notes: Prof. dr. Mia Hubert & Prof. dr. Stefan Van Aelst

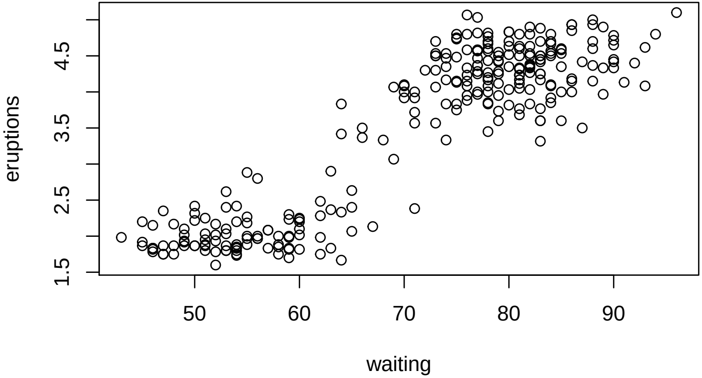
Practical information

- Classes and computer sessions (3, possibly a 4th will be added)
- PC sessions will focus on the use of R in regression analysis
- Additional exercises will be provided to do at home (not graded)
- Material will be provided on Toledo
- Suggested reference: Applied Linear Statistical Models, 5th Edition, Kutner et al. (2005)

Practical information

- Evaluation
  - Open book exam: course notes + slides
  - Focus on correct application of regression analysis
  - Individual written project at the end of semester involving data analysis tasks
  - Oral exam with written preparation + possibly questions related to the project
  - Final grade: project grade * 0.35 + exam result * 0.65
  - Second chance exam: opportunity to write a new report (new data and questions)

How can one best capture the relationship between variables?



Simple regression

- A method to model the **relation** between an input variable X and an output or *response* variable Y
  - **Asymmetric**
  - To what extent does the outcome Y **change** due to a change in the value of X?
  - **Predict** Y from X
  - X is the *independent* variable, or *regressor*, Y is the *dependent* variable.

Regression: a more general definition

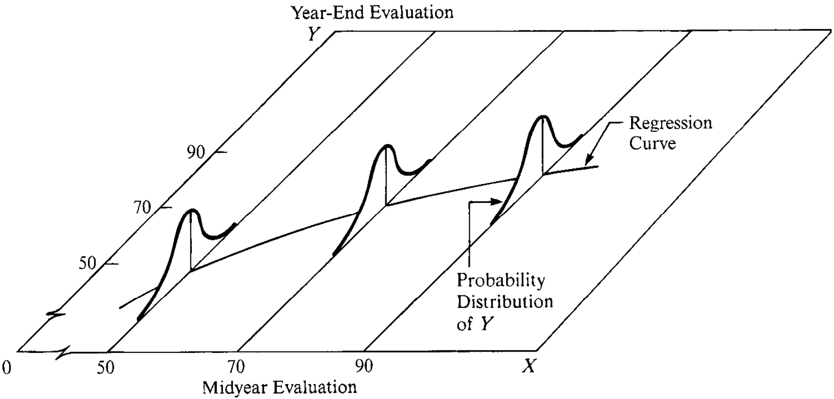
- Regression analysis models the relationship between a set of predictor variables $X_1, X_2, \ldots, X_{l-1}$ and a response variable Y that are measured on n observations.
- Find a **relation** between the X_j ($j = 1, \ldots, l-1$) and Y, which reveals the joint influence of the X-variables on Y.
- Predict the dependent variable Y from the independent variables $X_1, \ldots, X_{l-1}$
- In a very general form, we seek real functions g, f and a parameter vector $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})^t$ such that $g(Y)$ can be well described by $f(X_1, \ldots, X_{l-1}, \boldsymbol{\beta})$.
- Unless otherwise stated, we will assume that the response variable is *continuous*.

Regression: a more general definition

- The observations will in general not satisfy this functional relation exactly
- A **stochastic component** ϵ **expresses the variation of the data points around the regression curve**.
- A regression model thus postulates that:

1. There is a probability distribution of Y for each level of $X = (X_1, \dots, X_{I-1})$.
2. The means of these probability distributions vary in some systematic fashion with X.

FIGURE 1.4 **Pictorial Representation of Regression Model.**



General linear regression Model:

$$g(y_i) = \beta_0 + \beta_1 f_1(x_{i1}, \ldots, x_{i,l-1}) + \ldots + \beta_{p-1} f_{p-1}(x_{i1}, \ldots, x_{i,l-1}) + \epsilon_i$$

for $i = 1, \ldots, n$ and for certain choices of $g, f_1, \ldots, f_{p-1}$. The error terms ϵ_i represent the random variation of the data points around the regression curve. We assume that the standard Gauss-Markov conditions are satisfied:

$$\begin{aligned} \mathrm{E}[\epsilon_i] &= 0 \\ \mathrm{Var}[\epsilon_i] &= \sigma^2 \\ \mathrm{E}[\epsilon_i \epsilon_j] &= 0 \text{ for all } i \neq j. \end{aligned}$$

This general linear model includes:

1. The first-order regression model:
$l = p, g(y_i) = y_i, f_j(x_{i1}, \ldots, x_{i,p-1}) = x_{ij} \;\; (j = 1, \ldots, p-1)$
$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{p-1} x_{i,p-1} + \epsilon_i.$$
2. Simple regression: the first-order regression model with $p = 2$.
3. Polynomial regression: $g, f_1, \ldots, f_{l-1}$ as in the first-order regression model, and e.g. additionally
$f_l = X_1^2, f_{l+1} = X_3^2, f_{l+2} = X_1 X_2$.
4. Variable selection: $g, f_1, \ldots, f_{l-3}$ as in the first-order regression model, all other $f_j = 0$.
5. Transformations in X or Y:
$g(Y) = \log(Y), g(Y) = \frac{y^\lambda - 1}{\lambda}, f_j = \log(X_j)$.

- Linear models are linear in $\boldsymbol{\beta}$ and not necessarily in the independent variables X_j!
- An example of a nonlinear model is

$$y_i = \beta_0 + \beta_1 e^{\beta_2 x_i} + \epsilon_i.$$

Course contents

1. Simple regression model
2. The general linear model
3. Statistical inference
4. Polynomial regression
5. Categorical predictors
6. Transformations
7. Variable selection methods
8. Multicollinearity
9. Influential observations and outliers
10. Nonlinear regression
11. Nonparametric regression