Machine Learning I

Lecture 2 : Regression

Issam Falih
Department of Computer Science

September 8, 2025

Overview

1. Introduction
2. Simple Linear regression
3. Extension to multiple linear regression
4. Logistic regression
5. Validating regression results

A data ?

A **data** *(datum) is a basic description of a thing, an individual, a fact, an instruction or a phenomenon.*

- A data is made of one or several descriptive criteria called **variables** or **features**:
  - A single criterion: Univariate data
  - Several criteria: Multivariate data (bivariate data for 2 criteria)

A data ?

A **data** *(datum) is a basic description of a thing, an individual, a fact, an instruction or a phenomenon.*

- A data is made of one or several descriptive criteria called **variables** or **features**:
  - A single criterion: Univariate data
  - Several criteria: Multivariate data (bivariate data for 2 criteria)

A **data set** *is a set containing several data, usually identified by an id*. The id is not considered a variable.

- Other names for a data: individual, object, data object, observation, point.
- Other names for a data set: population
- Other names for a variable: feature, descriptor, criterion

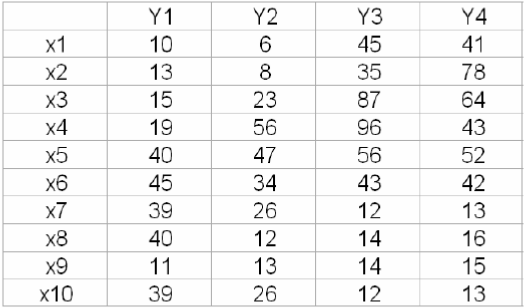
A data matrix

A data set can usually be represented in the form of a matrix with N lines (for N objects) and D columns (for D variables):

$$X = \{x_1, ...x_N\} = \begin{pmatrix} x_{1,1} & x_{1,2} & & \cdots & x_{1,D} \\ x_{2,1} & x_{2,2} & & & \\ & & \ddots & & \\ \vdots & & & x_{i,j} & \vdots \\ & & & \ddots & \\ x_{N,1} & & \cdots & & x_{N,D} \end{pmatrix}$$

A data array

A data set can also be represented as an array:



	Y1	Y2	Y3	Y4
x1	10	6	45	41
x2	13	8	35	78
x3	15	23	87	64
x4	19	56	96	43
x5	40	47	56	52
x6	45	34	43	42
x7	39	26	12	13
x8	40	12	14	16
x9	11	13	14	15
x10	39	26	12	13

- Objects: x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9 and x_{10}.
- Variables: $Y1$, $Y2$, $Y3$ and $Y4$.

Types of variables

Quantitative variables

- **Continuous variables**: Size, weight, time, volume, speed, etc.
- **Discrete variables**: Counting the number of item in a room, number of items, etc.

Qualitative variables

Categorical variables: Binary data, colors, gender, having a credit or not, labels, etc.

- **Ordered**: survey result (not satisfied, satisfied, very satisfied), nominal sizes (small, medium, tall, very tall), etc.
- **Unordered**: eye color

Others

Text, images, videos, etc.

Introduction

A **regression** *between several variables calculates the equation that best represents the link between these variables.*

We distinguish two types of variables:

- The **target variable** that we want to model.
- The **explanatory variables** that we will use to build the model for the target variable.

Introduction

- Given n observations of a random variable Y and several random variables $\mathrm{X}_1, \cdots, \mathrm{X}_\mathrm{i}$, the goal of a regression is to find a function $f()$ that best matches Y. In the linear model we asume that:
  - $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
  - $\epsilon \sim N(0, \sigma^2 I_{nxn})$, where I_{nxn} is the identity matrix
- Depending on what type of function $f()$ is, a regression can be linear, multi-linear, or logistic.

Residuals

The difference between a predicted value $\hat{y}$ and the observed one y is called a **residual** and is usually denoted e.

$$e_i = \hat{y}_i - y_i$$

Estimation

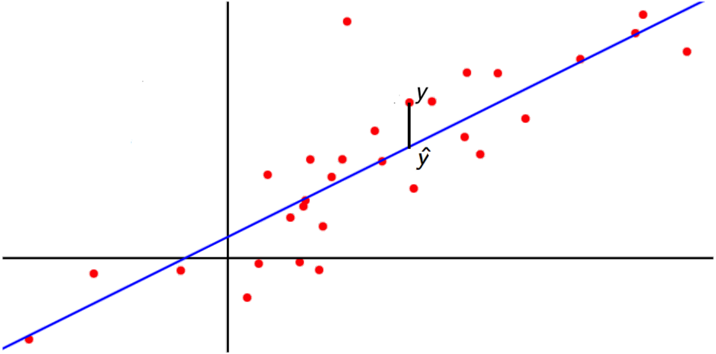
Let us consider two random variables X and Y that are strongly correlated. Let us suppose we have n observations of these variables (y_i, x_i)

- We want to obtain the "best" **linear equation** $\hat{\mathrm{Y}} = \hat{\beta}_1 \mathrm{X} + \hat{\beta}_0$ from a cloud of n points the coordinates of which are X and Y.
- We want a line that passes close to the centroid $G = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}$, so that

  $\hat{\beta}_0 = \mu_Y - \hat{\beta}_1 \mu_X$.
- We also want each item to be as close as possible from the regression line., i.e. minimizing the residual sum of squares:

$$RSS = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

Optimization criterion



Ordinary Mean square error

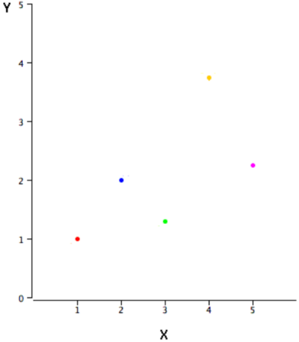
The values that minimize the RSS are:

$$\hat{\beta}_1 = \frac{COV(\mathrm{X},\mathrm{Y})}{VAR(\mathrm{X})} = r \times \frac{\sigma_Y}{\sigma_X} \qquad \text{and} \qquad \hat{\beta}_0 = \mu_Y - \hat{\beta}_1 \mu_X$$

They minimize the errors between the predicted values and the observed values.

Example

X	Y
1.0	1.0
2.0	2.0
3.0	1.3
4.0	3.75
5.0	2.25



- Let's compute the best regression factors for $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$

Example

X	Y
1.0	1.0
2.0	2.0
3.0	1.3
4.0	3.75
5.0	2.25

Example

X	Y
1.0	1.0
2.0	2.0
3.0	1.3
4.0	3.75
5.0	2.25

μ_X	3.0
μ_Y	2.06
σ_X	1.414
σ_Y	0.959
r	0.627

Example

X	Y
1.0	1.0
2.0	2.0
3.0	1.3
4.0	3.75
5.0	2.25

μ_X	3.0
μ_Y	2.06
σ_X	1.414
σ_Y	0.959
r	0.627

Computing the regression factors

Let's compute the best regression factors for $\hat{\mathrm{Y}} = \hat{\beta}_1 \mathrm{X} + \hat{\beta}_0$

$$\hat{\beta}_1 = \frac{COV(X,Y)}{VAR(X)} = r \times \frac{\sigma_Y}{\sigma_X}$$

$$\hat{\beta}_0 = \mu_Y - \hat{\beta}_1 \times \mu_X$$

Example

Computing the regression factors

$$\hat{\beta}_1 = 0.627 \times \frac{0.959}{1.414} = 0.425$$

$$\hat{\beta}_0 = 2.06 - 0.425 \times 3.0 = 0.785$$

$$\hat{\mathrm{Y}} = 0.425\mathrm{X} + 0.785$$

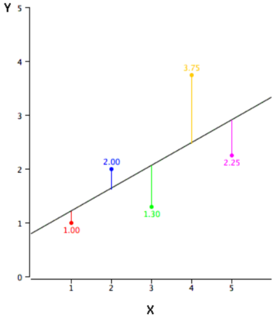
Example

Computing the regression factors

$$\hat{\beta}_1 = 0.627 \times \frac{0.959}{1.414} = 0.425 \qquad \hat{\beta}_0 = 2.06 - 0.425 \times 3.0 = 0.785$$

$$\hat{\mathrm{Y}} = 0.425\mathrm{X} + 0.785$$

X	Y	$\hat{\mathrm{Y}}$	$\hat{\mathrm{Y}} - \mathrm{Y}$	$(\hat{\mathrm{Y}} - \mathrm{Y})^2$
1.0	1.0	1.21	0.21	0.044
2.0	2.0	1.635	-0.365	1.333
3.0	1.3	2.06	0.76	0.578
4.0	3.75	2.485	-1.265	1.6
5.0	2.25	2.91	0.66	0.436



Exercise: Missing Values in Blood Pressure

Data sample:

Age x_i	BP y_i
36	12
42	13.5
48	?
54	13.6
56	14.3
60	15.4

1. Suppose a linear dependence exists. Which imputation method would you use?
2. Test the hypothesis with listwise deletion:

$$y = ax + b$$

$$a = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}, \quad b = \bar{y} - a\bar{x}$$

   with error $e_i = y_i - \hat{y}_i$.

3. Conclude on the relevance of the dependence.
4. Impute the missing value.

Exercise: Regression-based Single Imputation

Data (Age x, BP y):

x_i	y_i
36	12.0
42	13.5
48	?
54	13.6
56	14.3
60	15.4

Goal. Treat y at $x = 48$ as missing and impute it using the *fitted linear regression* on observed pairs.

Model. Fit OLS on observed data (listwise deletion):

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim (0, \sigma^2).$$

$$\bar{x} = \frac{1}{n} \sum x_i, \quad \bar{y} = \frac{1}{n} \sum y_i, \qquad S_{xx} = \sum (x_i - \bar{x})^2,$$

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}), \qquad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Imputation at $x_0 = 48$:

$$\hat{y}(48) = \hat{\beta}_0 + \hat{\beta}_1 \cdot 48.$$

Extension to multiple linear regression

This technique can be extended to multi-variate data.

- A **multiple linear regression** can be applied to find the best equation describing one variable (target) from several others (features).

Let us denote **Y** of n observations of the target variable and **X** the $n \times p$ matrix which columns contain the features. We want to find the $\hat{\beta}_0 ... \hat{\beta}_{p-1}$ that fit the best the following model :

$$\mathrm{Y} = \hat{\beta}_0 + \hat{\beta}_1 \mathrm{X}_1 + \hat{\beta}_{p-1} \mathrm{X}_{p-1} + \epsilon = \hat{\beta}_0 + \sum_{j=1}^{p-1} \hat{\beta}_i \mathrm{X}_i + \epsilon$$

Extension to multiple linear regression

This technique can be extended to multi-variate data.

- A **multiple linear regression** can be applied to find the best equation describing one variable (target) from several others (features).

Let us denote $\mathbf{Y}$ of n observations of the target variable and $\mathbf{X}$ the $n \times p$ matrix which columns contain the features. We want to find the $\hat{\beta}_0 ... \hat{\beta}_{p-1}$ that fit the best the following model :

$$\mathrm{Y} = \hat{\beta}_0 + \hat{\beta}_1 \mathrm{X}_1 + \hat{\beta}_{p-1} \mathrm{X}_{p-1} + \epsilon = \hat{\beta}_0 + \sum_{j=1}^{p-1} \hat{\beta}_i \mathrm{X}_i + \epsilon$$

Remarks :

- This model is also used with linear combinations of known functions of X_j.

Extension to multiple linear regression

This technique can be extended to multi-variate data.

- A **multiple linear regression** can be applied to find the best equation describing one variable (target) from several others (features).

Let us denote **Y** of n observations of the target variable and **X** the $n \times p$ matrix which columns contain the features. We want to find the $\hat{\beta}_0 ... \hat{\beta}_{p-1}$ that fit the best the following model :

$$\mathrm{Y} = \hat{\beta}_0 + \hat{\beta}_1 \mathrm{X}_1 + \hat{\beta}_{p-1} \mathrm{X}_{p-1} + \epsilon = \hat{\beta}_0 + \sum_{j=1}^{p-1} \hat{\beta}_i \mathrm{X}_i + \epsilon$$

Remarks :

- This model is also used with linear combinations of known functions of X_j.
- There are formulas to compute the regression factors manually, but it is best and faster to use an algorithm.

Extension to multiple linear regression

Like for the simple linear regressions, we want to minimize the residual sum of squares (RSS) between the observed and fitted values):

$$MSE = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

The estimator that minimizes the MSE is given by $\hat{\beta} = (X^T X)^{-1} X^T Y$

Extension to multiple linear regression

Like for the simple linear regressions, we want to minimize the residual sum of squares (RSS) between the observed and fitted values):

$$MSE = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

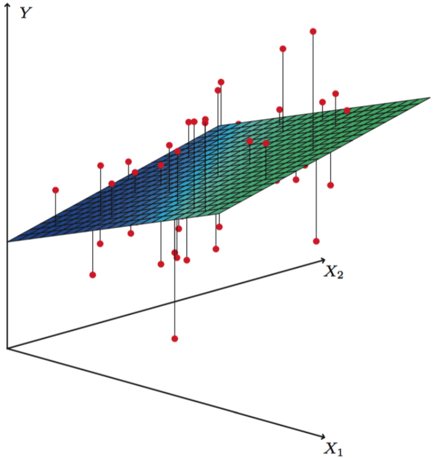
The estimator that minimizes the MSE is given by $\hat{\beta} = (X^T X)^{-1} X^T Y$

Things to do before running a multiple-linear regression

- Checking whether or not all variables seem relevant to the model: Do we need all of them ?
- Checking that the explanatory variables are not too correlated:
  - Removing correlated variables from the model can simplify it a lot.
  - Keeping correlated variables makes the regression model unstable and can lead to regression factors that are difficult to interpret.

Accuracy of the Coefficient Estimates

With two predictors and one response, the least squares regression line becomes a plane. The plane is chosen to minimize the sum of the squared vertical distances between each observation (shown in red) and the plane.



Accuracy of the Coefficient Estimates for the simple linear model

For the simple linear model

- The coefficients estimates are unbiased and follow a normal law with parameters: $E(\hat{\beta}) = \beta$ and variances $var(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\mu_x^2}{(n-1)\sigma_X^2} \right)$ and $var(\hat{\beta}_1) = \sigma^2 \left(\frac{1}{(n-1)\sigma_X^2} \right)$
- It is possible to estimate the variance σ^2 from the residual:

  $\hat{s}^2 = \frac{RSS}{n-2}$ whose distribution is $\dfrac{(n-2)\hat{s}^2}{\sigma^2} \sim \chi^2_{n-2}$

These formula can be generalized for the general linear model

- Law of the coefficients estimates : $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2 (X^T X)^{-1})$
- the estimator of the variance : $\hat{s}^2 = \frac{RSS}{n-p}$ whose distribution is $\dfrac{(n-p)\hat{s}^2}{\sigma^2} \sim \chi^2_{p-2}$

From these formula is possible to calculate confidence intervals.

Significance of the variables in a multiple linear regression

When using multiple linear regressions, it is important to check the significance of the variables in the model to ensure that the slope of the regression differs significantly from zero.

Student T test

Let a_i be a parameter estimate, and s_{a_i} its standard deviation:

$$T_i = \frac{a_i}{s_{a_j}}$$

- A p-value must then be calculated from T_i using tables or a graph of the Student law with $(n - p - 1)$ degrees of freedom. (remember the χ^2 test)

Assessing the Accuracy of the Model

The quality of a linear regression fit can be assessed using two related quantities: the residual standard error (RSE) and the Coefficient of determination R^2.

- The RSE is an estimate of the standard deviation of ϵ

$$RSE = \sqrt{\frac{RSS}{n-2}}$$

- The coefficient of determination R^2 is the proportion of variance explained. So it always takes on a value between 0 and 1, and is independent of the scale of Y.

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

  where $TSS = \sum(y_i - \mu_y)^2$ is the total sum of squares.
  The R^2 measures the proportion of variability in Y that is explained (or removed) by performing the regression.

  In the case of a simple regression $R^2 = \rho^2$

Multiple linear regression: example

“Healthy breakfast” data set

77 cereals described by: name, brand, type, calories, protein, fat, sodium, fiber, carbo, sugars, potassium, shelf of display, weight and a rating.

We want to do a regression to predict the rating from the other variables.

- We first remove the variables that are obviously irrelevant or inconvenient to use in a regression: name of the cereal, brand of the cereal, and the shelf in which they were displayed.
- We then run simple linear regressions to see which variables have the best individual correlation to predict the rating.

Multiple linear regression: example

Best linear model: Sugar

- Linear model: Rating$= 59.3 - 2.40$Sugars
- $R^2 = 0.577$ between “Sugar” and “Rating”

The dataset includes several other variables such as the gram of fat and dietary fiber per serving. Can we significantly improve the model by adding these variables ?

Adding the “Fat variable” ?

- The correlation between “Fat” and “Rating” is -0.409.
- The correlation between “Fat” and “Sugar” is 0.271.

The ”Fat” variable therefore seems to be a good candidate to improve the model.

Multiple linear regression: example

Sugars and Fat model

- Linear model: Rating$= 61.1 - 3.07$Fat -2.21Sugars
- $R^2 = 0.621$

Predictor	Coeff	Std dev	T	p
Constant	61.089	1.953	31.28	0.000
Fat	-3.066	1.036	-2.96	0.004
Sugars	-2.2128	0.2347	-9.43	0.000

The model improved and all variables are significant.

Multiple linear regression: example

Sugars, Fiber and Fat model

- Linear model: Rating$= 53.4 - 3.48$Fat $+2.95$Fiber -1.96Sugars
- $R^2 = 0.861$

Predictor	Coeff	Std dev	T	p
Constant	53.437	1.342	39.82	0.000
Fat	-3.4802	0.6209	-5.61	0.000
Fiber	2.9503	0.2549	11.57	0.000
Sugars	-1.9640	0.1420	-13.83	0.000

The correlation augmented significantly and the variables are still significant.

Multiple linear regression: example

Sugars, Fiber and Fat model

- Linear model: Rating$= 53.4 - 3.48$Fat $+2.95$Fiber -1.96Sugars
- $R^2 = 0.861$

Predictor	Coeff	Std dev	T	p
Constant	53.437	1.342	39.82	0.000
Fat	-3.4802	0.6209	-5.61	0.000
Fiber	2.9503	0.2549	11.57	0.000
Sugars	-1.9640	0.1420	-13.83	0.000

The correlation augmented significantly and the variables are still significant.
Remark: The R^2 automatically increases when extra explanatory variables are added to the model. It is possible to calculate the Adjusted R-squared $\bar{R}^2$ to penalize the number of variables.

Logistic regression

The goal of a logistic regression is to predict the outcome of a categorical variable depending on the value of other numerical variables.

The binary logistic regression

We want to find the β_i that best matches:

$$y = \begin{cases} 1, & \text{if } f(\beta, X) > 0 \\ 0, & \text{otherwise} \end{cases}$$

- We want to find a function $f()$ that is the probability that $y = 1$.
- We would like $f()$ to be a function of the linear model that we already know: $\beta_0 + \sum_{i=1}^{p} \beta_i x_i$.

Logistic regression

Since we want a probability, it needs to be positive and between 0 and 1. We use the logistic function to ensure positivity and normalization:

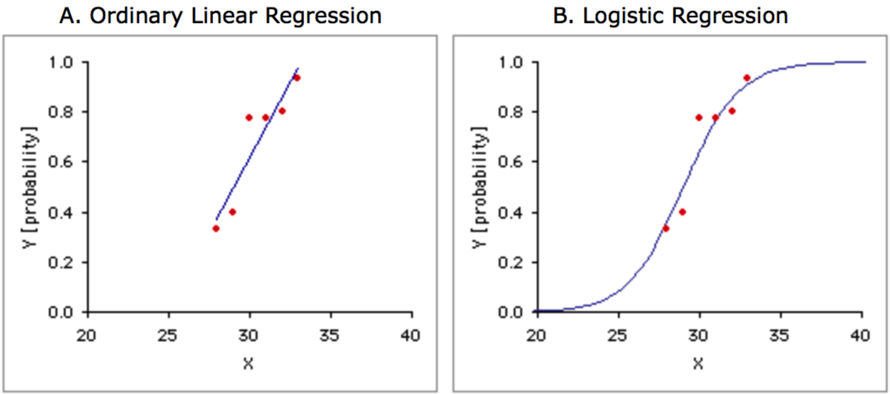
$$p(X) = \frac{\exp(\beta_0 + \sum_{i=1}^{d} \beta_i x_i)}{1 + \exp(\beta_0 + \sum_{i=1}^{d} \beta_i x_i)} \tag{1}$$

After a bit of manipulation, we obtain

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \sum_{i=1}^{d} \beta_i x_i$$

The left-hand side is called the log-odds or logit.
The rule of prediction is the following: if $\hat{p} > 0.5$ the predicted value is 1 and 0 otherwise.

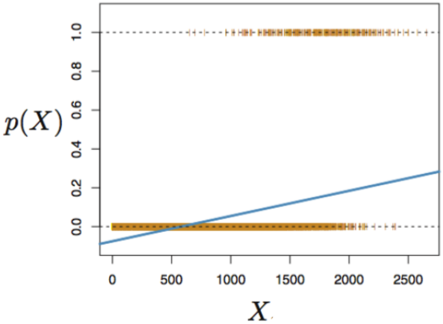
Linear regression VS logistic regression (1/2)



Linear regression VS logistic regression (2/2)

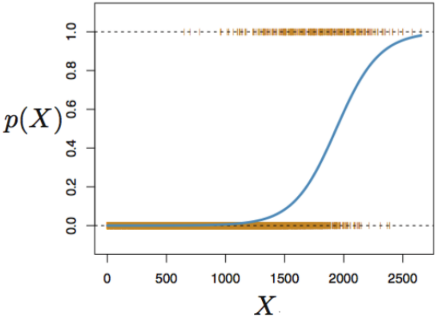
Linear regression

$$p(X) = \beta_0 + \beta_1 X$$



Logistic regression

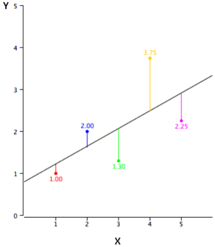
$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$



Validating regression results

So far, we have seen two ways of assessing the quality of regression results:

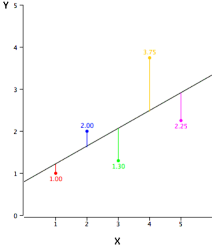
- Plotting the observed and fitted values and check that it looks fine
- Checking the correlation and p-value on the student test for all estimates.



Validating regression results

So far, we have seen two ways of assessing the quality of regression results:

- Plotting the observed and fitted values and check that it looks fine
- Checking the correlation and p-value on the student test for all estimates.



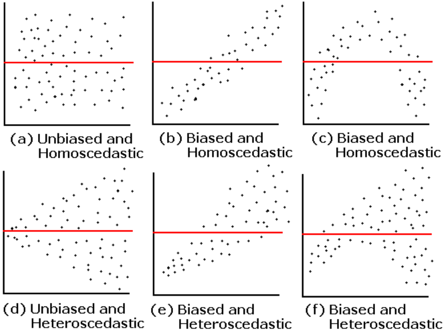
A deeper analysis will require checking the residuals: $e_i = \hat{y} - y$

- The idea is to detect patterns in the residuals that may reveal a bias.

Residuals VS fitted values graphs

The residuals (ordinates) vs fitted values (abscissa) is a good tool to validate the pertinence of a model.

- Ideally the points should be perfectly randomly distributed.
- Any other distribution may reveal a problem with the model, heterogeneous groups in the data, or outlier values.



(a) Unbiased and Homoscedastic

(b) Biased and Homoscedastic

(c) Biased and Homoscedastic

(d) Unbiased and Heteroscedastic

(e) Biased and Heteroscedastic

(f) Biased and Heteroscedastic

Leverage of an observation

The **leverage** *measures the amount by which the predicted value would change if the observation was shifted one unit in the y-direction.*

- The leverage of an obervation is usualy denoted h_i, or h_{ii}.
- High-leverage observations are extreme or outlying values such that the lack of neighboring observations means that the fitted regression model will pass close to that particular observation.

Leverage of an observation

$$h_{ii} = \frac{\partial \hat{y}_i}{\partial y_i}, \qquad 0 \leq h_{ii} \leq 1$$

Normalized residuals residuals

Standardizing the residuals can be a good idea when their values range on a very large scale.

Student normalization of the residuals

$$r_i = \frac{\hat{y}_i - y_i}{\hat{\sigma}\sqrt{1 - h_i}} = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_i}}$$

All diagrams that we will present thereafter can be interpreted with either regular or standardized residuals.

Residuals VS leverage graphs

Cook's distance *is a commonly used estimate of the influence of a data point when performing a least-squares regression analysis.*

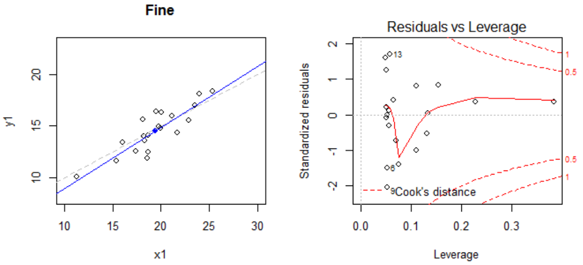
- Cook's distance is also called Cook's D and if most often denoted D.
- Cook's distance measures the effect of deleting a given observation.
- Points with a large Cook's distance are considered to merit closer examination in the analysis.

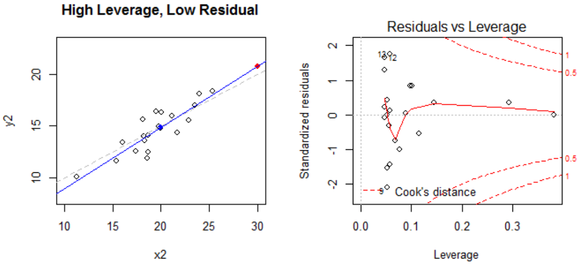
Cook's Distance

$$D_i = \frac{e_i^2}{MSE}\left[\frac{h_i}{(1-h_i)^2}\right]$$

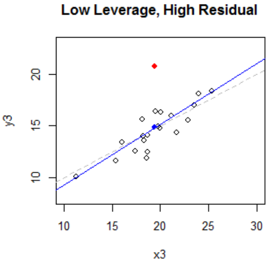
- Cook's distance is often displayed on "Residual VS leverage" to detect outliers.

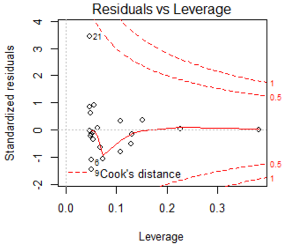
Residuals VS leverage graphs

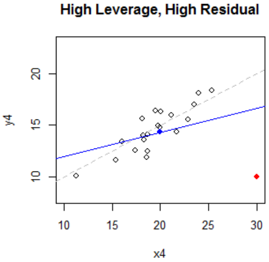


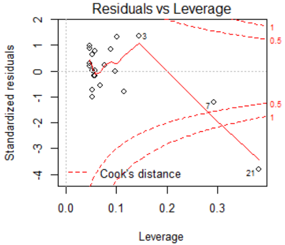


Residuals VS leverage graphs


Low Leverage, High Residual


Residuals vs Leverage


High Leverage, High Residual


Residuals vs Leverage

Questions ?