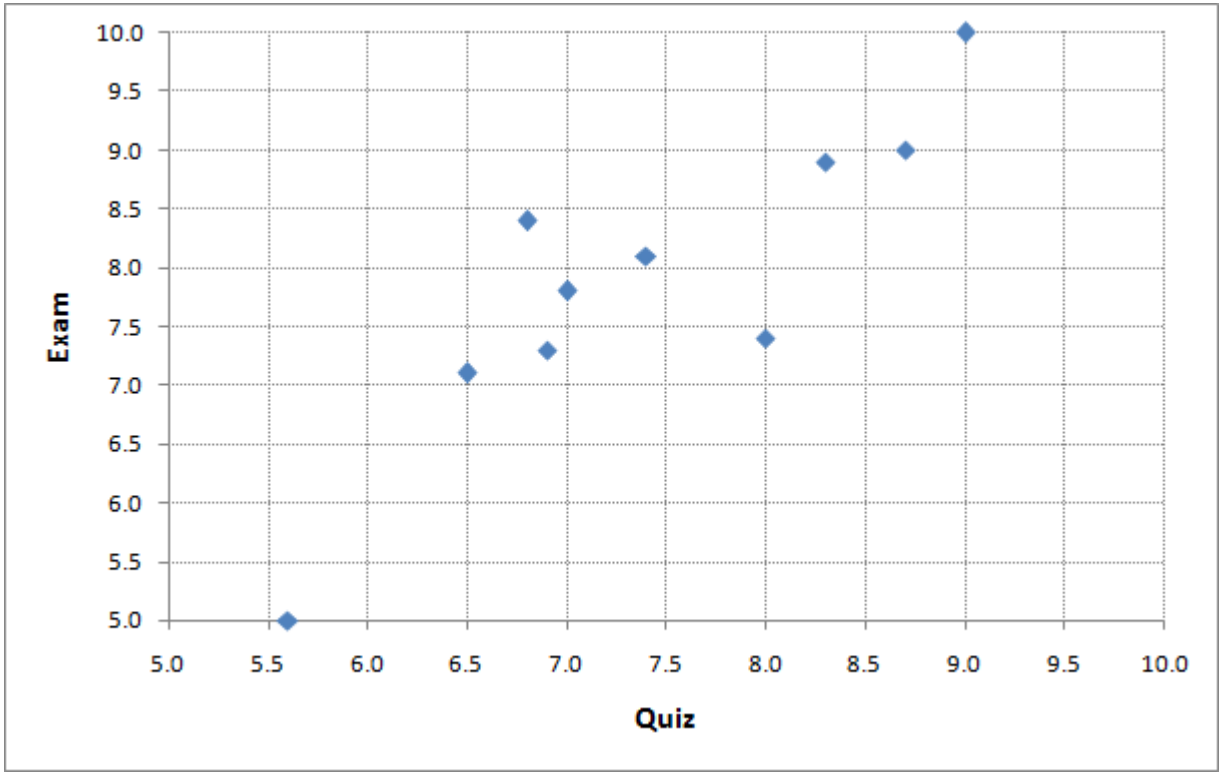
Lecture #2

Notation
Linear Regression
Least Squares Method

Regression

- A reminder: In supervised learning, if the output y is real-valued quantitative, the problem is called regression
- In this lecture we will use the already mentioned example of student's exam mark prediction from his/her quiz mark



Quiz	Exam
5.6	5.0
6.5	7.1
6.8	8.4
6.9	7.3
7.0	7.8
7.4	8.1
8.0	7.4
8.3	8.9
8.7	9.0
9.0	10.0

Notation

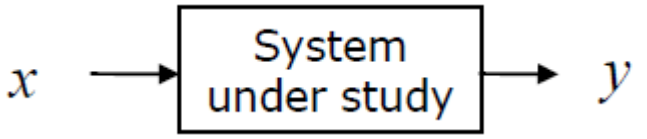
- n – the number of examples in the training dataset (the number of data points)
- m – the number of available features in the dataset (the number of input variables)
- j – used to index features (from 1 to m)
- i – used to index examples (from 1 to n)
- x_j – jth input variable (feature, attribute, independent variable)
- x_i – feature values of the ith training example (scalar or vector)
- y –output variable (response, dependent variable)

m features

n examples

x_1	x_2	y
Quiz	Homework	Exam
5.6	6.0	5.0
6.5	7.0	7.1
6.8	7.2	8.4
6.9	6.8	7.3
7.0	7.2	7.8
7.4	8.5	8.1
8.0	6.5	7.4
8.3	7.9	8.9
8.7	7.3	9.0
9.0	9.1	10.0

Regression problem



- When we are studying a behavior of some system, we observe m different input variables (features) $x = (x_1, x_2, \ldots, x_m)$ and a quantitative output variable y.
- We assume that there is a relationship between x and y, which can be written in the general form:

$$y = h(x) + \varepsilon$$

  where h is some fixed but unknown function of x and ε is a random error term, which is independent of x and has zero mean.
- h represents systematic information that x provides about y. ε is just noise that doesn't give us any useful information.
- We are interested in finding a function f which would estimate h for prediction of y:

$$\hat{y} = f(x)$$

- In machine learning, we are typically not concerned with the exact form of f, provided that it yields accurate predictions for y.

Linear regression

- Linear regression is the very first deeply studied type of regression modeling. And in practical applications, nowadays it's still one of the most popular. It is also used as a part of many other more sophisticated regression modeling approaches.
- Linear regression assumes that the relationship between x and y is liner (simply a sum of the features):

$$y = a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_m x_m + \varepsilon \qquad \text{or} \qquad y = a_0 + \sum_{j=1}^{m} a_j x_j + \varepsilon$$

- Once we have assumed that the relationship is linear, the problem of estimating f becomes greatly simplified. Instead of having to estimate an entirely arbitrary m-dimensional function $f(x)$, one only needs to estimate the $m + 1$ parameters $a_0, a_1, \ldots, a_m$ for this linear model.
- After the model has been selected, we need a procedure that uses the training data to *fit* or *train* the model. For linear models, we need to estimate the parameters $a_0, a_1, \ldots, a_m$ such that it predicts y with the least error.

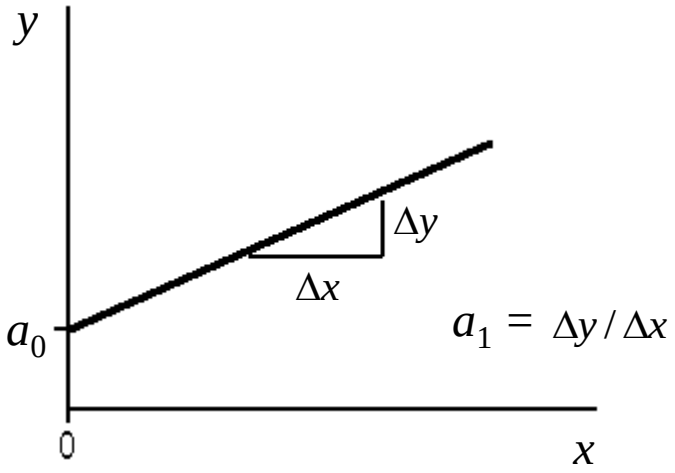
Simplest case

- Simplest case – one feature ($m = 1$)
- In this case the linear model is simply an equation of a line:

$$\hat{y} = a_0 + a_1 x$$

a_0 determines y value when $x = 0$, i.e., when the line crosses the y axis (it's also called y-intercept)

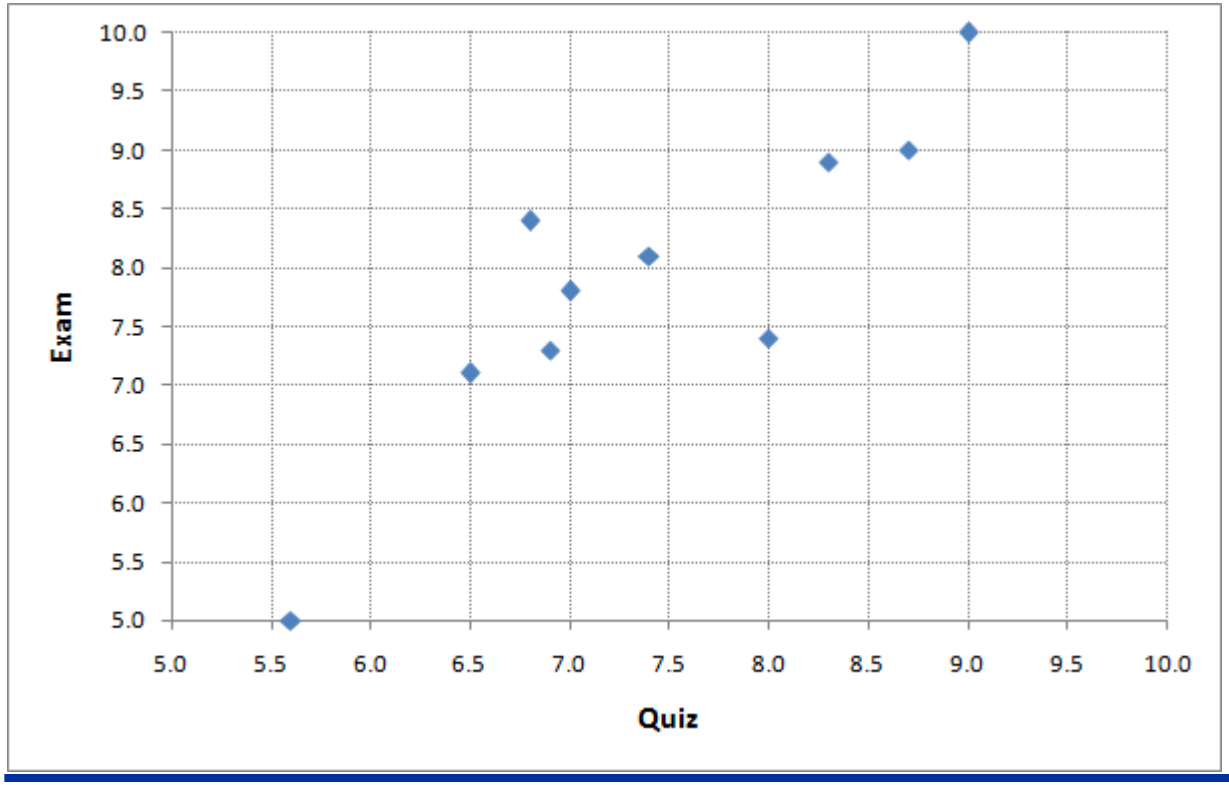
a_1 determines the slope of the line



Our dataset

- Let's say that from previous year we have a dataset with quiz and exam marks for 10 students in a specific lecture course. $n = 10$, $m = 1$
- Our task – create a regression model that would allow predicting a student's exam mark from his/her quiz mark in the lecture course.

  (of course, the model would stay valid only if the contents and the requirements of the lecture course wouldn't change)



x	y
Quiz	**Exam**
5.6	5.0
6.5	7.1
6.8	8.4
6.9	7.3
7.0	7.8
7.4	8.1
8.0	7.4
8.3	8.9
8.7	9.0
9.0	10.0

Problem

- Line equation

$$\hat{y} = a_0 + a_1 x$$

- We need to train the model – we need a procedure, that uses the training data to estimate model's parameters

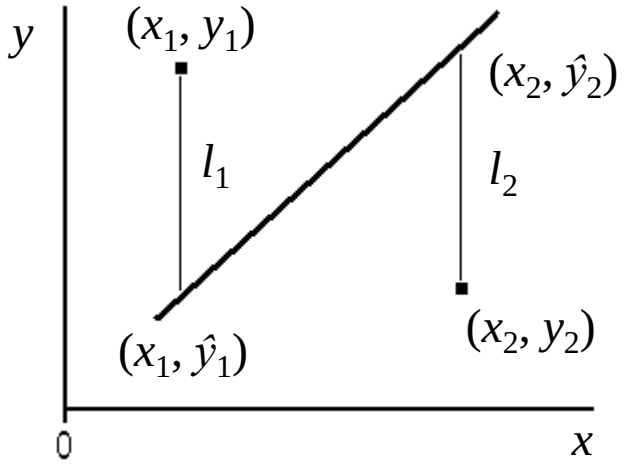
$a_0 = ?$

$a_1 = ?$

- Informally: The line should be in minimal distance from all the data points at the same time.
- But how do we measure the distance?

Criterion

- We need a formal quantitative criterion which would allow us to calculate the distance from the line to all the data points.
- (x_i, y_i) are coordinates of ith data point
- $(x_i, \hat{y}_i)$ are coordinates of our model at the same x_i



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

Criterion

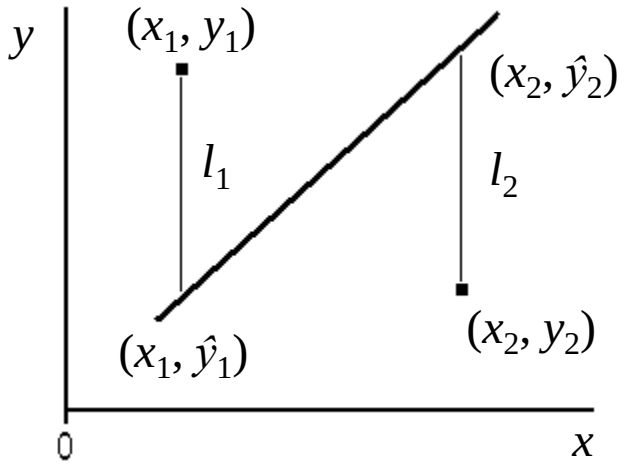
- How this criterion for minimizing distance from the line to the data points could look like?
  - Minimize absolute residuals

$$S = \sum_{i=1}^{n} | l_i | \rightarrow \min \qquad l_i = y_i - \hat{y}_i$$

  - The most popular one – minimize residual sum of squares

$$S = \sum_{i=1}^{n} (l_i)^2 \rightarrow \min$$

(*Residual Sum of Squares, RSS*)



Minimize residual sum of squares

- Minimize residual sum of squares

$$S = \sum_{i=1}^{n} (l_i)^2 \rightarrow \min$$

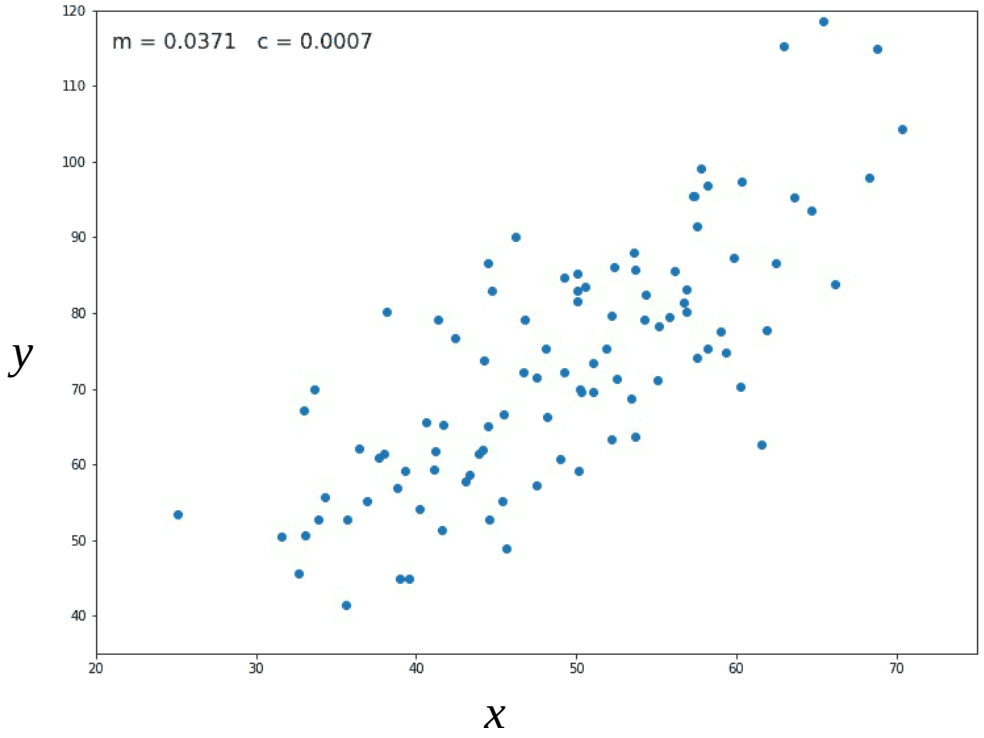
- We must estimate parameters a_0 and a_1 which would minimize S

- How?

- ~~1st option – try to guess the values~~
- 2nd option – search iteratively
- 3rd option – in linear regression we can do it directly using the least squares method

Iterative search

- 2nd option – iterative search
  - For example using gradient descend method
  - *(Other methods exist as well)*



In this animation

$\hat{y} = a_0 + a_1 x$

is

$\hat{y} = c + mx$

With each iteration, the line is nearer to the data points and the residual sum of squares becomes smaller

Animation taken from: https://towardsdatascience.com/linear-regression-using-gradient-descent-97a6c8700931

Least Squares Method

Model

$$\hat{y} = a_0 + a_1 x$$

Criterion

$$S = \sum_{i=1}^{n} (l_i)^2 \rightarrow \min$$

Data

Quiz	Exam
5.6	5.0
6.5	7.1
6.8	8.4
6.9	7.3
7.0	7.8
7.4	8.1
8.0	7.4
8.3	8.9
8.7	9.0
9.0	10.0

- Because $l_i = y_i - \hat{y}_i$, replace l_i with $y_i - \hat{y}_i$

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

- Replace $\hat{y}_i$ with our equation's left side

$$S = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2$$

- So we have to minimize this:

$$S = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2 \rightarrow \min$$

Partial derivatives

$$S = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2 \rightarrow \min$$

- We must find minimum. Differentiation will help.
- Partial derivatives: separately for a_0 and for a_1

$$\frac{\partial S}{\partial a_0} = -2 \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)$$

$$\frac{\partial S}{\partial a_1} = -2 \sum_{i=1}^{n} x_i (y_i - a_0 - a_1 x_i)$$

- We are interested in the position (values for a_0 and a_1) where the derivative is 0 because there is the minimum of our original function S

$$\sum_{i=1}^{n} (y_i - a_0 - a_1 x_i) = 0$$

$$\sum_{i=1}^{n} x_i (y_i - a_0 - a_1 x_i) = 0$$

System of equations

$$\sum_{i=1}^{n}(y_i - a_0 - a_1 x_i) = 0$$

$$\sum_{i=1}^{n} x_i(y_i - a_0 - a_1 x_i) = 0$$

- We get linear system of equations:

$$\begin{cases} na_0 + a_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ a_0 \sum_{i=1}^{n} x_i + a_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases}$$

- Good news: in practice the differentiation step is skipped, i.e., we can go directly to the system of equations

Solving the linear system of equations

Model

$$\hat{y} = a_0 + a_1 x$$

Criterion

$$S = \sum_{i=1}^{n} (l_i)^2 \rightarrow \min$$

Data

Quiz	Exam
5.6	5.0
6.5	7.1
6.8	8.4
6.9	7.3
7.0	7.8
7.4	8.1
8.0	7.4
8.3	8.9
8.7	9.0
9.0	10.0

- Linear system of equations

$$\begin{cases} na_0 + a_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ a_0 \sum_{i=1}^{n} x_i + a_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases}$$

- Insert values from the table

$$\begin{cases} 10a_0 + a_1 74.2 = 79.0 \\ a_0 74.2 + a_1 560.8 = 597.55 \end{cases}$$

$$\sum_{i=1}^{n} x_i = 74.2$$

$$\sum_{i=1}^{n} y_i = 79.0$$

$$\sum_{i=1}^{n} x_i y_i = 597.55$$

$$\sum_{i=1}^{n} x_i^2 = 560.8$$

Solving the linear system of equations

Model

$$\hat{y} = a_0 + a_1 x$$

Criterion

$$S = \sum_{i=1}^{n} (l_i)^2 \to \min$$

Data

Quiz	Exam
5.6	5.0
6.5	7.1
6.8	8.4
6.9	7.3
7.0	7.8
7.4	8.1
8.0	7.4
8.3	8.9
8.7	9.0
9.0	10.0

- Linear system of equations

$$\begin{cases} 10a_0 + a_1 74.2 = 79.0 \\ a_0 74.2 + a_1 560.8 = 597.55 \end{cases}$$

- Solve

$$10a_0 = -a_1 74.2 + 79.0 \quad a_0 = (-a_1 74.2 + 79.0)/10$$

$$((-a_1 74.2 + 79.0)/10) \cdot 74.2 + a_1 560.8 = 597.55$$

$$a_1 = 1.11 \qquad a_0 = (-a_1 74.2 + 79.0)/10 = -0.342$$

This can be solved using Gaussian elimination or other methods.

But in the case $m = 1$ it's even simpler.

Estimated parameters

Model

$$\hat{y} = a_0 + a_1 x$$

Criterion

$$S = \sum_{i=1}^{n} (l_i)^2 \rightarrow \min$$

Data

Quiz	Exam
5.6	5.0
6.5	7.1
6.8	8.4
6.9	7.3
7.0	7.8
7.4	8.1
8.0	7.4
8.3	8.9
8.7	9.0
9.0	10.0

- **Parameters:**

  $a_0 = -0.342$

  $a_1 = 1.11$

- **Model:**

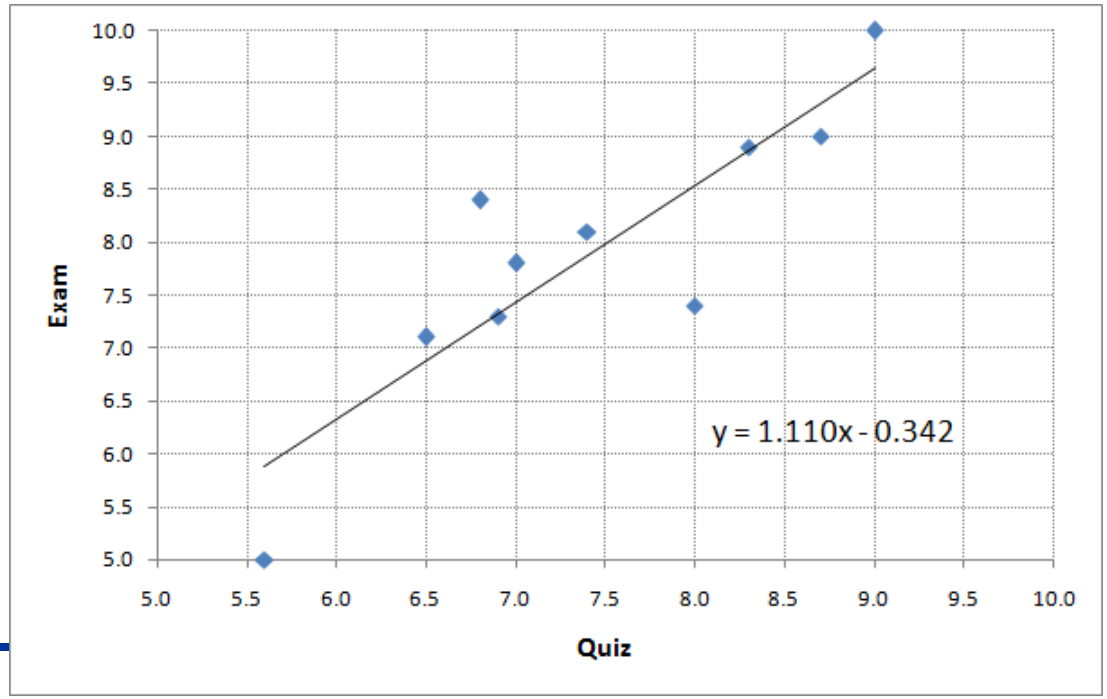
  $\hat{y} = -0.342 + 1.11x$

- **Usage:**

  If $x = 6$, then

  $\hat{y} = -0.342 + 1.11 * 6 =$

  $= 6.3$



Example from previous lecture

- Example of dataset (hundreds of lines):

y	x1	x2	x3	x4	x5	x6
44528352.2	0.74	0.035090909	0.005045455	0.0036	1.827272727	2.654545455
24715277.64	1.694545455	0.061636364	0.003463636	0.005563636	2.018181818	2.190909091
20326865.79	1.007272727	0.068545455	0.003981818	0.003709091	2.672727273	2.781818182
137000567.3	1.707272727	0.031818182	0.005181818	0.004309091	2.754545455	2.263636364
28781035.62	1.478181818	0.061272727	0.005427273	0.003572727	1.677272727	2.2
53326365.5	1.923636364	0.046363636	0.004118182	0.0057	2.659090909	2.727272727
17071631.67	0.981818182	0.063454545	0.005536364	0.005263636	2.713636364	2.218181818
29042876.47	0.816363636	0.042	0.003545455	0.004254545	1.690909091	2.154545455
47433129.96	0.88	0.036181818	0.003681818	0.005018182	2.877272727	2.690909091
49955387.24	1.745454545	0.057272727	0.003627273	0.003109091	2.004545455	2.736363636
42728959.03	1.465454545	0.038363636	0.0051	0.005781818	1.663636364	2.290909091
91624928.58	1.681818182	0.038	0.005836364	0.0039	2.195454545	2.836363636
29472772.8	1.185454545	0.045272727	0.003272727	0.005454545	1.8	2.872727273
14632842.74	0.994545455	0.062363636	0.005645455	0.004663636	1.595454545	2.7
133436389.4	1.630909091	0.034363636	0.003654545	0.0033	2.304545455	2.236363636
15395689.69	0.930909091	0.064545455	0.003518182	0.0054	2.577272727	2.209090909
35502643.92	1.821818182	0.062	0.005372727	0.004854545	2.890909091	2.427272727
14718512.23	0.918181818	0.063818182	0.005263636	0.005318182	2.413636364	2.918181818

- Linear regression model example for this dataset:

$\hat{y}$ = -25.2757 + 24953.4751*x_1*x_4 + 32527.2204*x_1*x_1*x_3 - 3417.2160*x_4*x_6 - 23244.3027*x_1*x_1*x_3*x_5 - 18504.1503*x_1*x_2*x_4 + 7440.6421*x_1*x_1*x_3*x_5*x_5 + 785.9807*x_1*x_3 + 42.8613550433136*x_2 - 5.5812*x_1*x_1*x_3*x_5*x_5*x_5*x_6 + 18.5222*x_1*x_3*x_5*x_5*x_6*x_6 - 1.4918*x_1*x_5 - 4389.5744*x_1*x_1*x_3*x_4*x_5*x_5 - 871.7860*x_1*x_1*x_3*x_5*x_5*x_5 - 6527.5472*x_1*x_4*x_6 + 622.8030*x_1*x_4*x_6*x_6 + 1874.7267*x_1*x_1*x_4 - 2416.2249*x_1*x_1*x_2*x_4 - 384.8857*x_1*x_1*x_4*x_5 + 7.1707*x_6 - 3356.9264*x_1*x_2*x_3 + 32761.9964*x_1*x_2*x_2*x_3 - 400.774338*x_1*x_1*x_2*x_3 + 5249.8395*x_1*x_1*x_2*x_2*x_3 - 1722.6871*x_1*x_3*x_5 - 1091.3215*x_1*x_2*x_4*x_6 + 263.5026*x_2*x_2