ML3 – Linear regression

Stefano Rovetta

A.y. 2024-2025

Possible scenarios

1 (useful mostly for reasoning):

- Learner = $\mathbb{H}$ is fixed
- Data are fixed (population)

→ Find correct learners in $\mathbb{H}$ (a *representation* task)

Possible scenarios

1 (useful mostly for reasoning):

- Learner = $\mathbb{H}$ is fixed
- Data are fixed (population)

→ Find correct learners in $\mathbb{H}$ (a *representation* task)

2 (not realizable):

- No data necessary: Probabilities are assumed to be known!

→ Find best hypotesis space $\mathbb{H}$ and optimal learner

Possible scenarios

1 (useful mostly for reasoning):

- Learner = $\mathbb{H}$ is fixed
- Data are fixed (population)

→ Find correct learners in $\mathbb{H}$ (a *representation* task)

2 (not realizable):

- No data necessary: Probabilities are assumed to be known!

→ Find best hypotesis space $\mathbb{H}$ and optimal learner

3 **(your usual situation)**:

- Data will be **stochastic** but probabilities are **not known**
- Hypothesis space $\mathbb{H}$ fixed, chosen in advance

→ Find learners in H which are correct **for any possible realization of the data** (a *generalization* task)

Scenario 3:
Only data are available

Linear regression

See either (or both) of

- Trevor Hastie, Robert Tibshirani, Jerome Friedman
  The Elements of Statistical Learning
  Data Mining, Inference, and Prediction
  Second Edition, Springer 2008

- Christopher M. Bishop
  Pattern Recognition and Machine Learning
  Springer 2006

The regression problem

Regression means:
approximating a functional dependency based on measured data.

The regression problem

Regression means:
approximating a functional dependency based on measured data.

a typical supervised problem

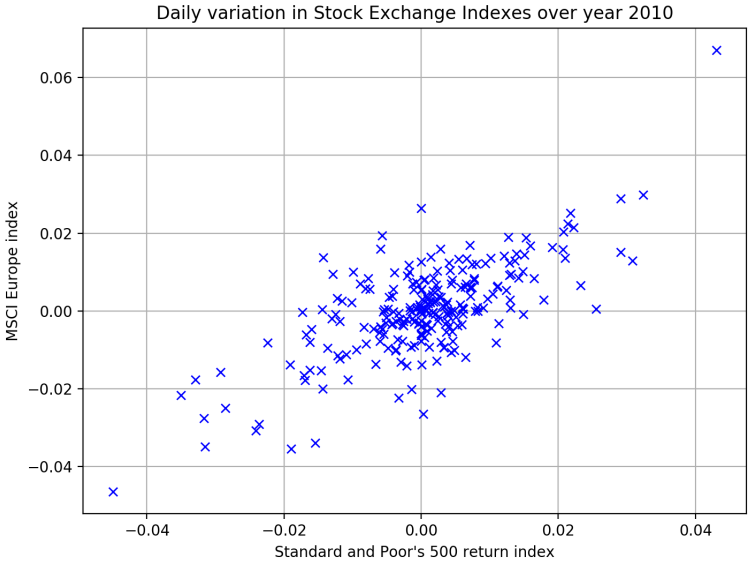
Data

$$\begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \mathbf{x}_3 \\ \vdots \\ \mathbf{x}_N \end{pmatrix} \quad \begin{pmatrix} \mathbf{t}_1 \\ \mathbf{t}_2 \\ \mathbf{t}_3 \\ \vdots \\ \mathbf{t}_N \end{pmatrix}$$

Observations Target

1: One-dimensional linear regression

Example



249 observations in year 2010. (Source: UCI)
We want to predict the variation of the MSCI European index by observing Standard and Poor's 500 return index.

- Observation: x is the value of the variation of Standard and Poor's (SP) 500 return index on a given day.

- Observation: x is the value of the variation of Standard and Poor's (SP) 500 return index on a given day.
- Target: t is the value of the variation of the MSCI European index (MSCI) on the same day.

- Observation: x is the value of the variation of Standard and Poor's (SP) 500 return index on a given day.
- Target: t is the value of the variation of the MSCI European index (MSCI) on the same day.

- Observation: x is the value of the variation of Standard and Poor's (SP) 500 return index on a given day.
- Target: t is the value of the variation of the MSCI European index (MSCI) on the same day.

There is clearly some relationship between the two values.

- Observation: x is the value of the variation of Standard and Poor's (SP) 500 return index on a given day.
- Target: t is the value of the variation of the MSCI European index (MSCI) on the same day.

There is clearly some relationship between the two values.

But not one-to-one

A model for approximating the data

A linear model $y(x)$ that predicts t given x:

$$t \approx y \quad \text{where} \quad y = wx$$

For instance
$t_1 \approx y_1$ where $y_1 = wx_1$
or $t_2 \approx y_2$ where $y_2 = wx_2$

We want $y(x)$ to be similar to $t(x)$ for any x.

Computing the parameter

$x_1 = 0.0159 \quad \Rightarrow \quad y_1 = 0.0159 \times w \quad$ must approximate $\quad t_1 = 0.0167$

So:

$$w = \frac{0.0167}{0.0159} = 1.0503$$

Computing the parameter

$x_1 = 0.0159 \quad \Rightarrow \quad y_1 = 0.0159 \times w \quad$ must approximate $\quad t_1 = 0.0167$

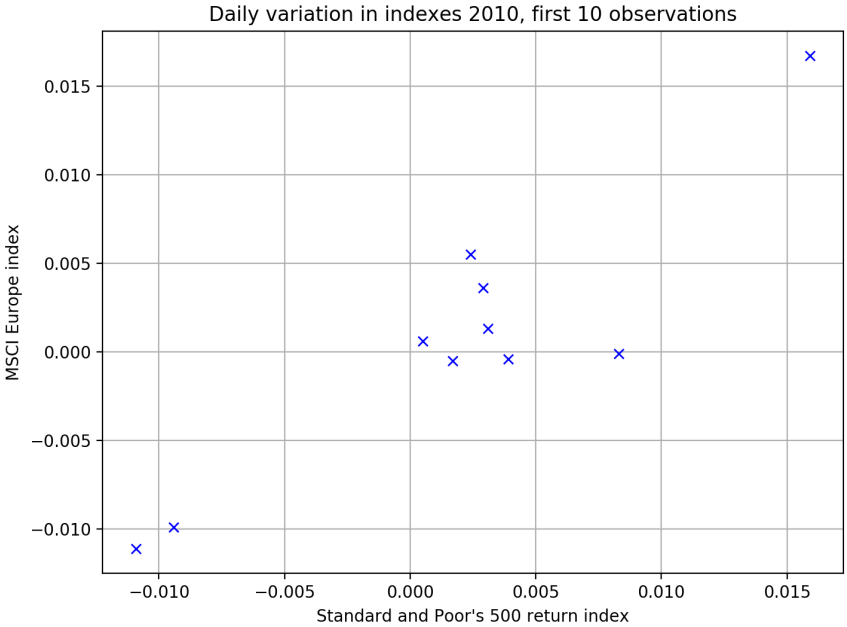
So:

$$w = \frac{0.0167}{0.0159} = 1.0503$$

However,
$x_2 = 0.0031 \quad \Rightarrow \quad y_2 = 0.0031 \times w \quad$ must approximate $\quad t_2 = 0.0013$

So:

$$w = \frac{0.0013}{0.0031} = 0.4194$$


Daily variation in indexes 2010, first 10 observations
MSCI Europe index
Standard and Poor's 500 return index
0.015
0.010
0.005
0.000
−0.005
−0.010
−0.010
−0.005
0.000
0.005
0.010
0.015

The first 10 days of 2011

The first 10 observations:

No.	Date	SP	MSCI
1	2010-01-04	0.0159	0.0167
2	2010-01-05	0.0031	0.0013
3	2010-01-06	0.0005	0.0006
4	2010-01-07	0.0034	−0.0004
5	2010-01-08	0.0029	0.0036
6	2010-01-11	0.0017	−0.0005
7	2010-01-12	−0.0094	−0.0099
8	2010-01-13	0.0083	−0.0001
9	2010-01-14	0.0024	0.0055
10	2010-01-15	−0.0109	−0.0111

The first 10 equations:

$$\begin{cases} 0,0159\, w = 0,0167 \\ 0,0031\, w = 0,0013 \\ 0,0005\, w = 0,0006 \\ 0,0039\, w = -0,0004 \\ 0,0029\, w = 0,0036 \\ 0,0017\, w = -0,0005 \\ -0,0094\, w = -0,0099 \\ 0,0083\, w = -0,0001 \\ 0,0024\, w = 0,0055 \\ -0,0109\, w = -0,0111 \end{cases}$$

2: Linear regression as an optimization problem

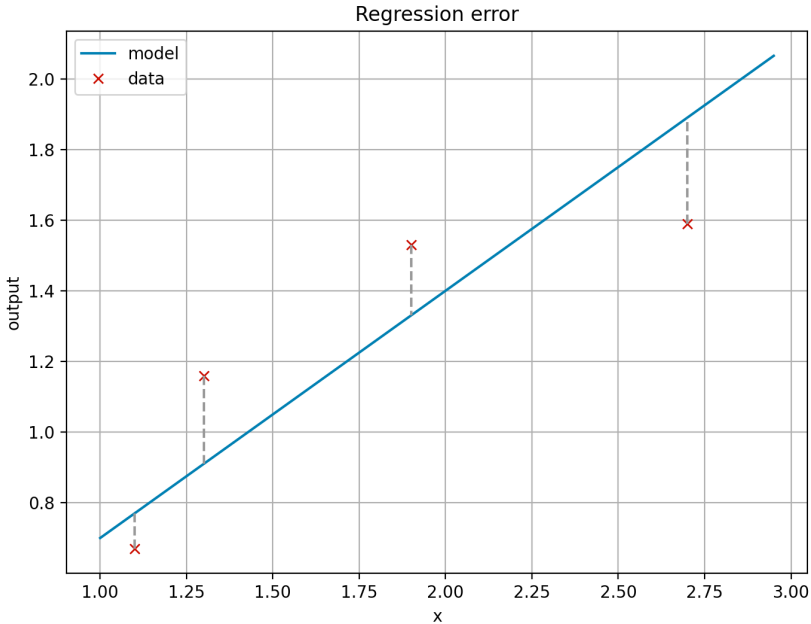
We cannot hope to find a value for w that is good for all points.

We should be satisfied with a value which is *generally not so bad* for *most of the points*

We cannot hope to find a value for w that is good for all points.

We should be satisfied with a value which is *generally not so bad* for *most of the points*

More rigorously: The expected (average) error should be low.


Regression error
model
data
output
x
2.0
1.8
1.6
1.4
1.2
1.0
0.8
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00

A solution based on optimization

Idea:

- quantify **how wrong** is each estimate using some **measure**,
- make this measure as small as possible **on average**.

A solution based on optimization

Idea:

- quantify **how wrong** is each estimate using some **measure**,
- make this measure as small as possible **on average**.

Measure = **loss function**.

Properties of a good loss for regression

Properties of a good loss for regression

- All errors give a positive contribution
  $\Rightarrow \lambda(y,t) = \lambda(-y,-t)$.
- Differentiable

Properties of a good loss for regression

- All errors give a positive contribution
  $\Rightarrow \lambda(y,t) = \lambda(-y,-t)$.
- Differentiable

Some options

- Error $\lambda_{\mathrm{E}}(y,t) = y - t$

Properties of a good loss for regression

- All errors give a positive contribution
  $\Rightarrow \lambda(y,t) = \lambda(-y,-t)$.
- Differentiable

Some options

- Error $\lambda_{\text{E}}(y,t) = y - t$
- Absolute error $\lambda_{\text{AE}}(y,t) = |y - t|$

Properties of a good loss for regression

- All errors give a positive contribution
  $\Rightarrow \lambda(y,t) = \lambda(-y,-t)$.
- Differentiable

Some options

- Error $\lambda_{\text{E}}(y,t) = y - t$
- Absolute error $\lambda_{\text{AE}}(y,t) = |y - t|$
- Square error $\lambda_{\text{SE}}(y,t) = (y - t)^2$

Properties of the square error loss

Square error loss

$$\lambda_{\text{SE}}(y, t) = (y - t)^2$$

- is even: $(t - y)^2 = (y - t)^2$
- grows more than linearly, giving heavier weight to larger errors
- is differentiable with respect to the model output:

$$\frac{d}{dy}\lambda_{\text{SE}}(y, t) = 2(y - t)$$

The objective function

Generic goal: **minimize the mean value of the loss over the whole data set**

$$J = \frac{1}{N} \sum_{l=1}^{N} \lambda(y_l, t_l) \, .$$

$J =$ **Objective function** or cost function

The objective function

Generic goal: **minimize the mean value of the loss over the whole data set**

$$J = \frac{1}{N}\sum_{l=1}^{N}\lambda(y_l, t_l)\,.$$

$J =$ **Objective function** or cost function

In the specific case of the square error loss:

$$J_{\text{MSE}} = \frac{1}{N}\sum_{l=1}^{N}(y_l - t_l)^2\,.$$

$J_{\text{MSE}} =$ mean square error objective.

Remark

$$J = \frac{1}{N}\sum_{l=1}^{N}\lambda(y_l, t_l) \approx \int_{-\infty}^{+\infty} \lambda\big(\, y(x), t(x)\,\big)\, p(x)\, dx \;=\; \mathrm{E}\{\lambda\big(\, y(x), t(x)\,\big)\}$$

We use an average because we cannot compute the expectation

The objective function

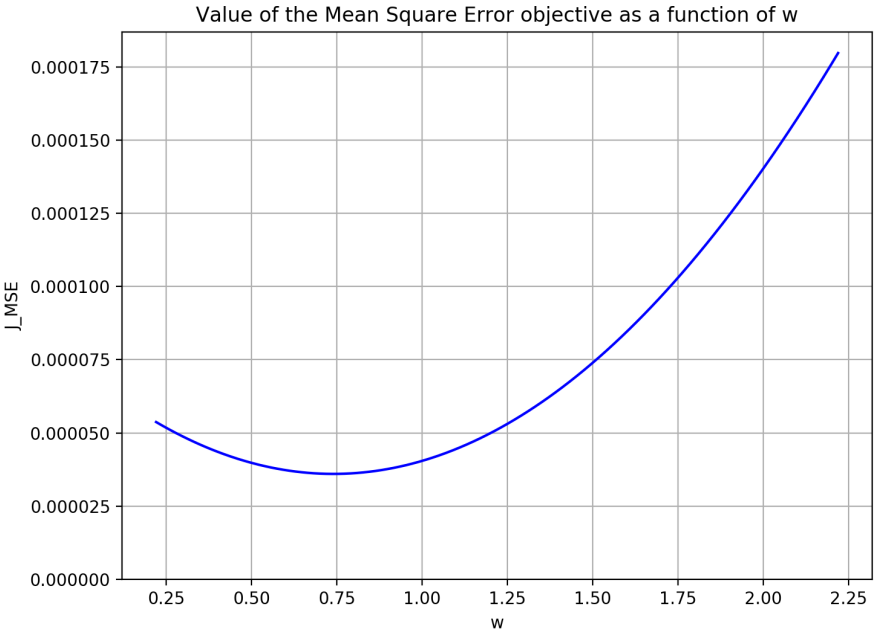
A loss function $\lambda(y, t)$ is a function of the two arguments y and t.

We can look at the objective in two complementary situations:

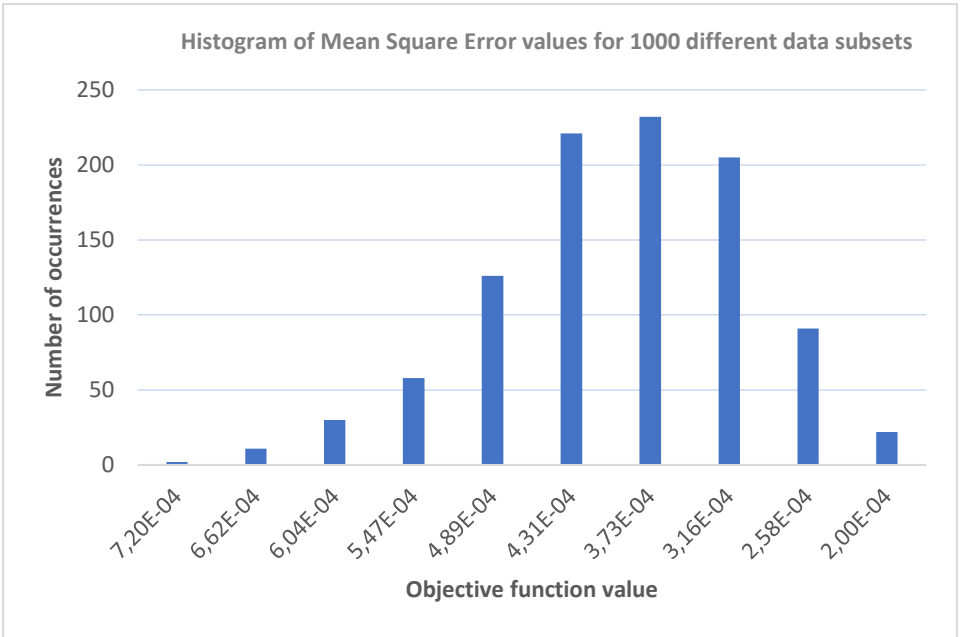
- Given the data set, the targets $t_1, \ldots, t_N$ are fixed
  $\longrightarrow$ **The objective depends only on the model parameter(s)**

- Given a model, the parameters are fixed.
  However we can apply this model to various data sets.
  $\longrightarrow$ **The objective depends only on the data.**

The objective as a function of the model parameter(s)



The objective as a function of the data



Summing up

- When **building a model (= training)**, the objective is a function of the parameters in the model and the data are fixed
- When **using a model (= inference)**, the objective is a function of the data, which uses the (now fixed) model parameters.

3: Solution of the one-dimensional linear regression problem

The least squares method

Problem:
minimize J_{MSE} with respect to the parameters with fixed data

The least squares method

Problem:
minimize J_{MSE} with respect to the parameters with fixed data

we know that J_{MSE} is a parabola

The least squares method

Problem:
minimize J_{MSE} with respect to the parameters with fixed data

we know that J_{MSE} is a parabola

Unique solution: w such that

$$\frac{d}{dw} J_{\text{MSE}} = 0$$

Computing $\frac{d}{dw} J_{\text{MSE}}$

$$\frac{d}{dw}\lambda_{\text{SE}}(y,t) = \frac{d}{dy}\lambda_{\text{SE}}(y,t) \times \frac{d}{dw} y$$

Here we have used the chain rule of differentiation to write the derivative:

$$\frac{df(g(x))}{dx} = \frac{df(y)}{dy}\bigg|_{y=g(x)} \frac{dg(x)}{dx}$$

Computing $\frac{d}{dw} J_{\text{MSE}}$

$$\frac{d}{dw}\lambda_{\text{SE}}(y,t) \quad = \quad \frac{d}{dy}\lambda_{\text{SE}}(y,t) \times \frac{d}{dw} y$$

Here we have used the chain rule of differentiation to write the derivative:

$$\frac{df(g(x))}{dx} = \left.\frac{df(y)}{dy}\right|_{y=g(x)} \quad \frac{dg(x)}{dx}$$

$$\downarrow \qquad\qquad \downarrow$$

$$\left.\frac{d}{dy}(y-t)^2\right|_{y=wx} \quad \frac{d}{dw}(wx)$$

Computing $\frac{d}{dw} J_{\text{MSE}}$

$$\frac{d}{dw}\lambda_{\text{SE}}(y,t) \quad = \quad \frac{d}{dy}\lambda_{\text{SE}}(y,t) \times \frac{d}{dw} y$$

Here we have used the chain rule of differentiation to write the derivative:

$$\frac{df(g(x))}{dx} = \left.\frac{df(y)}{dy}\right|_{y=g(x)} \quad \frac{dg(x)}{dx}$$

$$\downarrow \qquad\qquad \downarrow$$

$$\left.\frac{d}{dy}(y-t)^2\right|_{y=wx} \quad \frac{d}{dw}(wx)$$

$$\downarrow \qquad\qquad \downarrow$$

$$2(xw-t) \qquad\qquad x$$

Derivative of the loss

This is for one observation

$$\frac{d}{dw}\lambda_{\text{SE}}(t,y) = 2(xw-t)(x)$$
$$= 2x^2 w - 2xt\ .$$

Derivative of the objective

This is for the average over all observations

$$\frac{d}{dw} J_{\text{MSE}} = \frac{d}{dw} \frac{1}{N} \sum_{l=1}^{N} \lambda_{\text{SE}}(y_l, t_l)$$

Exchange sum and derivative:

$$= \frac{1}{N} \sum_{l=1}^{N} \frac{d}{dw} \lambda_{\text{SE}}(y_l, t_l)$$

$$= \frac{1}{N} \sum_{l=1}^{N} 2\left(x_l^2 w - x_l t_l \right) ,$$

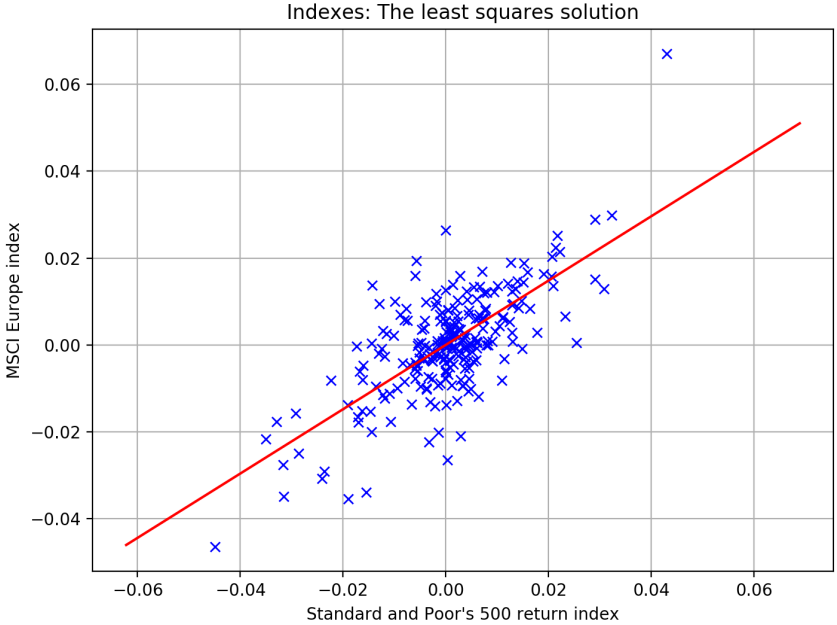
where the constant coefficient $(2/N)$ is irrelevant and can be disregarded.

This equation is solved by bringing w outside the sum, since it does not depend on l:

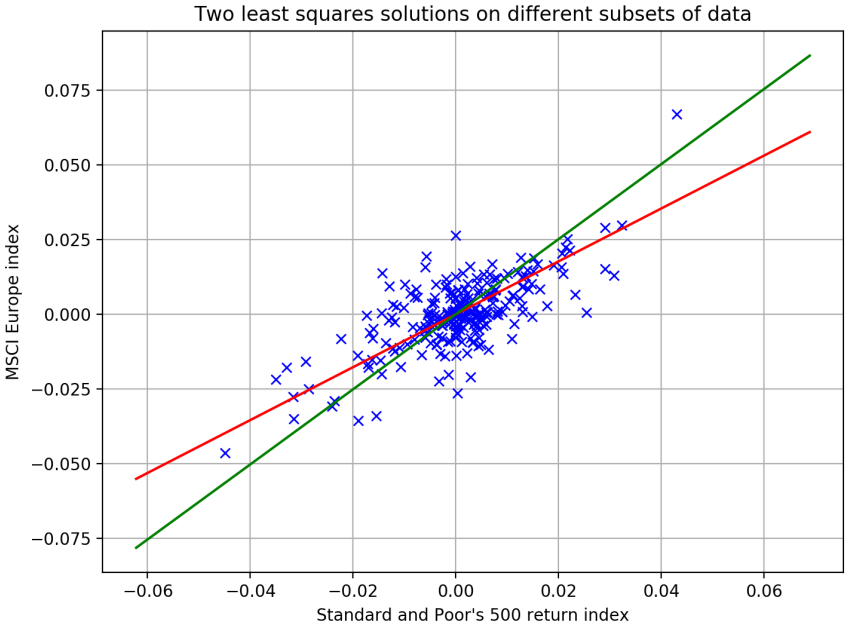
$$w \sum_{l=1}^{N} x_l^2 - \sum_{l=1}^{N} x_l t_l = 0$$

$$w = \frac{\sum_{l=1}^{N} x_l t_l}{\sum_{l=1}^{N} x_l^2}$$

This is the least squares solution to the linear regression problem.


Indexes: The least squares solution
MSCI Europe index
Standard and Poor's 500 return index
0.06
0.04
0.02
0.00
−0.02
−0.04
−0.06
−0.04
−0.02
0.00
0.02
0.04
0.06

Two different subsets of the data



4: One-dimensional linear regression problem with offset

1974 “Motor Trends” car data (four columns)

Model	mpg	disp	hp	weight
Mazda RX4	21	160	110	2.62
Mazda RX4 Wag	21	160	110	2.875
Datsun 710	22.8	108	93	2.32
Hornet 4 Drive	21.4	258	110	3.215
Hornet Sportabout	18.7	360	175	3.44
Valiant	18.1	225	105	3.46
Duster 360	14.3	360	245	3.57
Merc 240D	24.4	146.7	62	3.19
Merc 230	22.8	140.8	95	3.15
Merc 280	19.2	167.6	123	3.44
Merc 280C	17.8	167.6	123	3.44
Merc 450SE	16.4	275.8	180	4.07
Merc 450SL	17.3	275.8	180	3.73
Merc 450SLC	15.2	275.8	180	3.78
Cadillac Fleetwood	10.4	472	205	5.25
Lincoln Continental	10.4	460	215	5.424
Chrysler Imperial	14.7	440	230	5.345
Fiat 128	32.4	78.7	66	2.2
Honda Civic	30.4	75.7	52	1.615
Toyota Corolla	33.9	71.1	65	1.835
Toyota Corona	21.5	120.1	97	2.465
Dodge Challenger	15.5	318	150	3.52
AMC Javelin	15.2	304	150	3.435
Camaro Z28	13.3	350	245	3.84
Pontiac Firebird	19.2	400	175	3.845
Fiat X1-9	27.3	79	66	1.935
Porsche 914-2	26	120.3	91	2.14
Lotus Europa	30.4	95.1	113	1.513
Ford Pantera L	15.8	351	264	3.17
Ferrari Dino	19.7	145	175	2.77
Maserati Bora	15	301	335	3.57
Volvo 142E	21.4	121	109	2.78

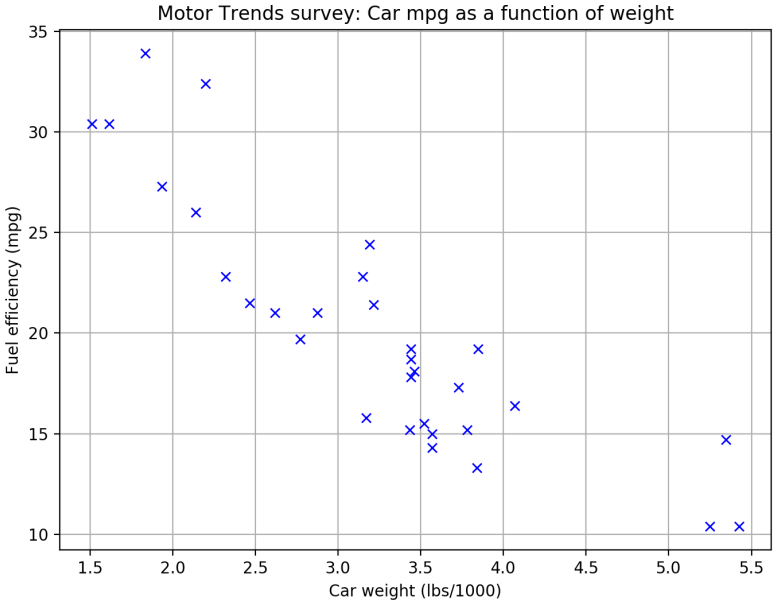
1974 survey on some car models.

Among other variables:

- mpg or miles-per-gallon
- disp or displacement (in cu.in)
- hp or horse-power
- weight, total (in lbs/1000)

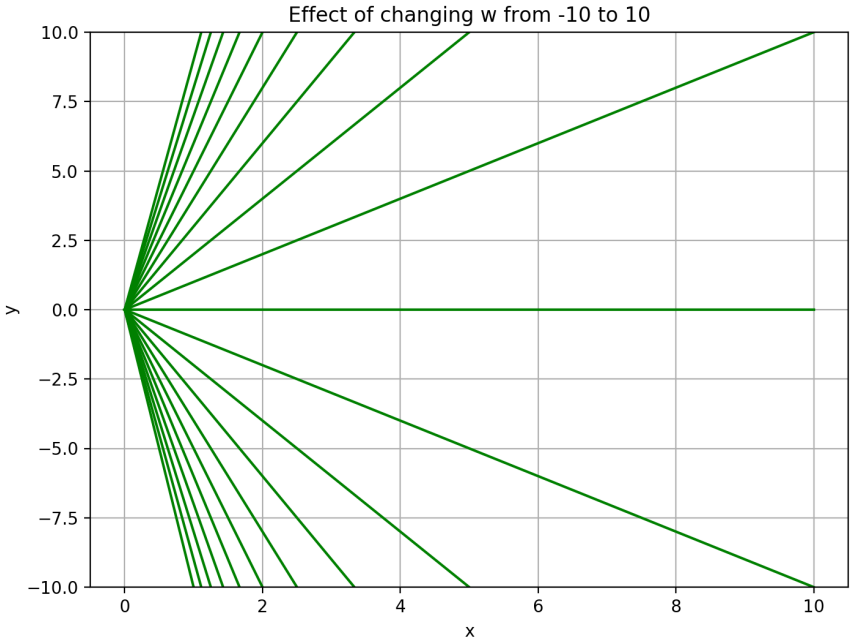
(Note that these are USA units.)

Forecasting mpg with weight



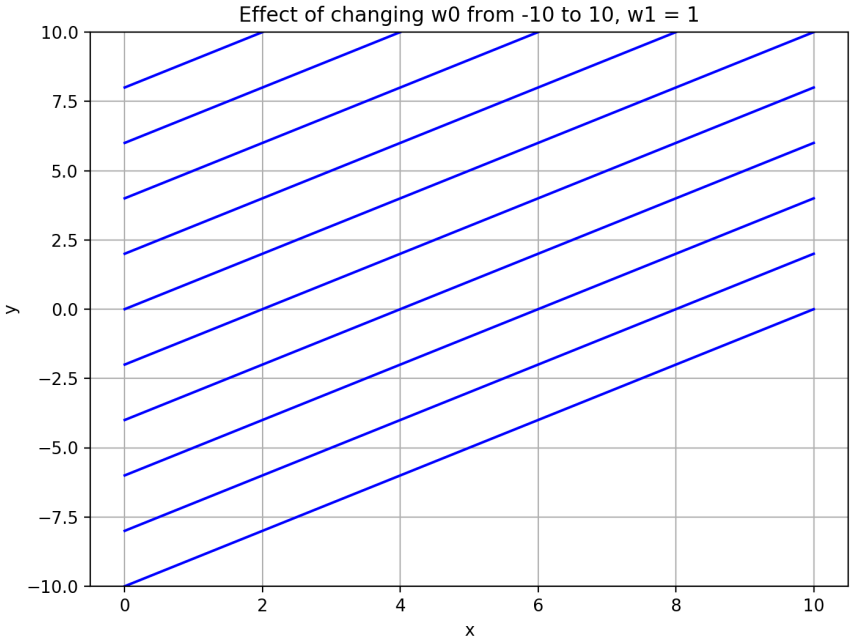
Our current model

$$y = wx$$



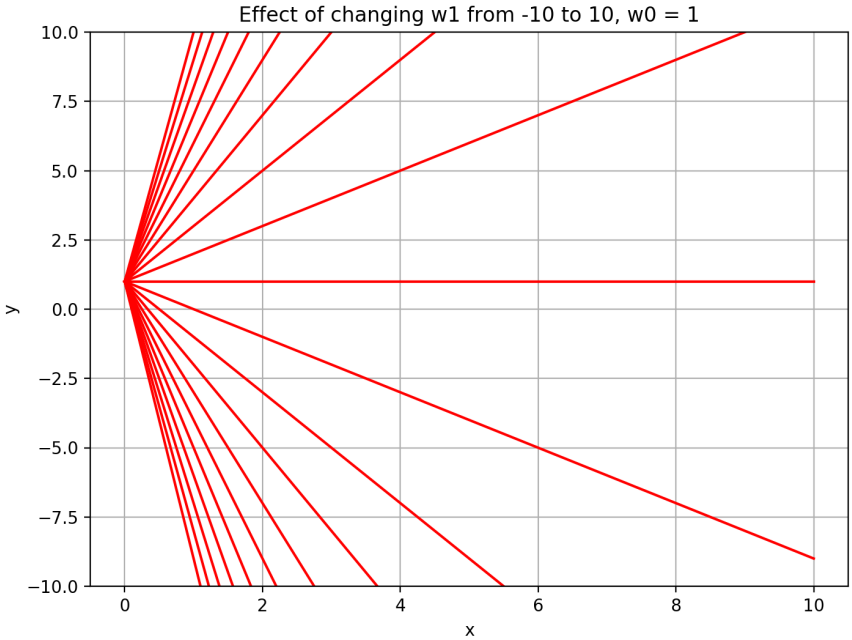
A more flexible model

$$y = w_1 x + w_0$$



A more flexible model

$$y = w_1 x + w_0$$



Solving

The solution in this case can be found by **centring** around the mean $\overline{x}$ of x and $\overline{t}$ of t

$$\overline{x} = \frac{1}{N}\sum_{l=1}^{N} x_l \qquad \overline{t} = \frac{1}{N}\sum_{l=1}^{N} t_l$$

Solving

The solution in this case can be found by **centring** around the mean $\overline{x}$ of x and $\overline{t}$ of t

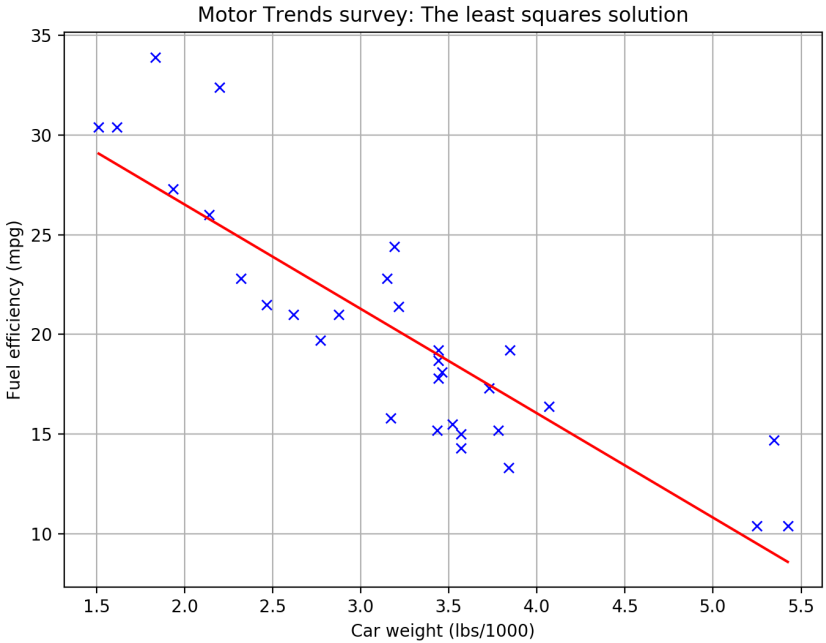
$$\overline{x} = \frac{1}{N}\sum_{l=1}^{N} x_l \qquad \overline{t} = \frac{1}{N}\sum_{l=1}^{N} t_l$$

$$w_1 = \frac{\sum_{l=1}^{N}(x_l - \overline{x})(t_l - \overline{t})}{\sum_{l=1}^{N}(x_l - \overline{x})^2}$$

$$w_0 = \overline{t} - w_1\overline{x}$$

$w_1 =$ slope; (EE) gain

$w_0 =$ intercept, offset; (EE) bias


Motor Trends survey: The least squares solution
Fuel efficiency (mpg)
Car weight (lbs/1000)
35
30
25
20
15
10
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5

Forecasting mpg with one variable

We have switched from a *linear* to an *affine* model

$$y = w_0 + x w_1$$

5: The multi-dimensional linear regression problem

Multidimensional inputs and solution in matrix form

The data is now composed of d-dimensional vectors:

$$\begin{aligned}\mathbf{x}_1 &= [\, x_{1,1}, x_{1,2}, \ldots, x_{1,d}\,] \\ \mathbf{x}_2 &= [\, x_{2,1}, x_{2,2}, \ldots, x_{2,d}\,] \\ &\ldots \\ \mathbf{x}_N &= [\, x_{N,1}, x_{N,2}, \ldots, x_{N,d}\,]\end{aligned}$$

Multidimensional inputs and solution in matrix form

The data is now composed of d-dimensional vectors:

$$\begin{aligned}\mathbf{x}_1 &= [\, x_{1,1}, x_{1,2}, \ldots, x_{1,d}\,] \\ \mathbf{x}_2 &= [\, x_{2,1}, x_{2,2}, \ldots, x_{2,d}\,] \\ &\ldots \\ \mathbf{x}_N &= [\, x_{N,1}, x_{N,2}, \ldots, x_{N,d}\,]\end{aligned}$$

so we can organize them into a $N \times d$ matrix:

$$X = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \mathbf{x}_3 \\ \vdots \\ \mathbf{x}_N \end{pmatrix} = \begin{pmatrix} x_{1,1} & x_{1,2} & \ldots & x_{1,d} \\ x_{2,1} & x_{2,2} & \ldots & x_{2,d} \\ x_{3,1} & x_{3,2} & \ldots & x_{3,d} \\ & & \vdots & \\ x_{N,1} & x_{N,2} & \ldots & x_{N,d} \end{pmatrix}$$

Since the data are now d-dimensional, we have d parameters in a d-dimensional vector:

$$\mathbf{w} = \begin{pmatrix} w_1 \\ w_2 \\ w_3 \\ \vdots \\ w_d \end{pmatrix}$$

The linear model takes the d inputs of each observation and combines them by using the d parameters to produce one output:

$$
\begin{aligned}
y_1 &= x_{1,1}w_1 + x_{1,2}w_2 + \ldots + x_{1,d}w_d \\
y_2 &= x_{2,1}w_1 + x_{2,2}w_2 + \ldots + x_{2,d}w_d \\
&\ldots \\
y_N &= x_{N,1}w_1 + x_{N,2}w_2 + \ldots + x_{N,d}w_d
\end{aligned}
$$

This can be expressed as a matrix-vector multiplication between the data matrix X and the parameter vector $\mathbf{w}$:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,d} \\ x_{2,1} & x_{2,2} & \dots & x_{2,d} \\ x_{3,1} & x_{3,2} & \dots & x_{3,d} \\ & & \vdots & \\ x_{N,1} & x_{N,2} & \dots & x_{N,d} \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ w_3 \\ \vdots \\ w_d \end{pmatrix}$$

$$= \begin{pmatrix} x_{1,1}w_1 + x_{1,2}w_2 + & \dots & +x_{1,d}w_d \\ x_{2,1}w_1 + x_{2,2}w_2 + & \dots & +x_{2,d}w_d \\ x_{3,1}w_1 + x_{3,2}w_2 + & \dots & +x_{3,d}w_d \\ & & \vdots \\ x_{N,1}w_1 + x_{N,2}w_2 + & \dots & +x_{N,d}w_d \end{pmatrix} = X\mathbf{w}$$

The affine case

We can incorporate the additive parameter w_0 by adding one constant column to the data matrix. We have $d+1$ parameters and the data become $(d+1)$-dimensional

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,d} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,d} \\ 1 & x_{3,1} & x_{3,2} & \dots & x_{3,d} \\ & & & \vdots & \\ 1 & x_{N,1} & x_{N,2} & \dots & x_{N,d} \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \\ \vdots \\ w_d \end{pmatrix}$$

$$= \begin{pmatrix} w_0 + x_{1,1}w_1 + x_{1,2}w_2 + & \dots & +x_{1,d}w_d \\ w_0 + x_{2,1}w_1 + x_{2,2}w_2 + & \dots & +x_{2,d}w_d \\ w_0 + x_{3,1}w_1 + x_{3,2}w_2 + & \dots & +x_{3,d}w_d \\ & \vdots & \\ w_0 + x_{N,1}w_1 + x_{N,2}w_2 + & \dots & +x_{N,d}w_d \end{pmatrix} = X\mathbf{w}$$

Finally, our goal is to make this model's prediction $\mathbf{y}$ as similar as possible to the measured outputs for each observation, which again can be organized as a vector, this time N-dimensional:

$$\mathbf{t} = \begin{pmatrix} t_1 \\ t_2 \\ t_3 \\ \vdots \\ t_N \end{pmatrix}$$

The square error objective in matrix-vector form

$$
\begin{aligned}
J_{\text{MSE}} &= \frac{1}{2}\|\mathbf{y}-\mathbf{t}\|^2 \\
&= \frac{1}{2}\|X\mathbf{w}-\mathbf{t}\|^2 \\
&= \frac{1}{2}(X\mathbf{w}-\mathbf{t})^\intercal(X\mathbf{w}-\mathbf{t}) \\
&= \frac{1}{2}(\mathbf{w}^\intercal X^\intercal-\mathbf{t}^\intercal)(X\mathbf{w}-\mathbf{t}) \\
&= \frac{1}{2}(\mathbf{w}^\intercal X^\intercal X\mathbf{w}-\mathbf{w}^\intercal X^\intercal\mathbf{t}-\mathbf{t}^\intercal X\mathbf{w}+\mathbf{t}^\intercal\mathbf{t}) \\
&= \frac{1}{2}\|X\mathbf{w}\|^2-\mathbf{w}^\intercal X^\intercal\mathbf{t}+\frac{1}{2}\|\mathbf{t}\|^2
\end{aligned}
$$

We can simplify the objective by disregarding the constant term $\frac{1}{2}\|\mathbf{t}\|^2$:

$$
\begin{aligned}
J_{\text{MSE}} &= \frac{1}{2}\left(\mathbf{w}^\intercal X^\intercal X\mathbf{w} - 2\mathbf{w}^\intercal X^\intercal \mathbf{t}\right) \\
&= \frac{1}{2}\|X\mathbf{w}\|^2 - \mathbf{w}^\intercal X^\intercal \mathbf{t}
\end{aligned}
$$

We can simplify the objective by disregarding the constant term $\frac{1}{2}\|\mathbf{t}\|^2$:

$$
\begin{aligned}
J_{\text{MSE}} &= \frac{1}{2}\left(\mathbf{w}^\intercal X^\intercal X\mathbf{w} - 2\mathbf{w}^\intercal X^\intercal \mathbf{t}\right) \\
&= \frac{1}{2}\|X\mathbf{w}\|^2 - \mathbf{w}^\intercal X^\intercal \mathbf{t}
\end{aligned}
$$

This is a paraboloid, a d-dimensional parabola, in the variables $\mathbf{w}$. It has a minimum when

$$
\begin{cases}
\partial J_{\text{MSE}}/\partial w_0 = 0 \\
\partial J_{\text{MSE}}/\partial w_1 = 0 \\
\partial J_{\text{MSE}}/\partial w_2 = 0 \\
\partial J_{\text{MSE}}/\partial w_3 = 0 \\
\vdots \\
\partial J_{\text{MSE}}/\partial w_d = 0
\end{cases}
\quad \text{or} \quad \nabla J_{\text{MSE}} = \mathbf{0}
$$

(⟵Note: this "zero" is a vector)

where $\mathbf{0} = \left(0, 0, \ldots\ 0\right)^\intercal$, and we can still solve this equation in closed form

Closed-form solution

It can be proven that we can write:

$$\begin{aligned}\nabla J_{\text{MSE}} &= \frac{\partial}{\partial \mathbf{w}} J_{\text{MSE}} \\ &= X^{\intercal} X \mathbf{w} - X^{\intercal} \mathbf{t}\end{aligned}$$

Closed-form solution

It can be proven that we can write:

$$\begin{aligned}\nabla J_{\text{MSE}} &= \frac{\partial}{\partial \mathbf{w}} J_{\text{MSE}} \\ &= X^{\intercal} X \mathbf{w} - X^{\intercal} \mathbf{t}\end{aligned}$$

Setting $\nabla J_{\text{MSE}} = \mathbf{0}$ we get

$$X^{\intercal} X \mathbf{w} = X^{\intercal} \mathbf{t}$$

By premultiplying both sides by $(X^{\intercal} X)^{-1}$, we obtain the closed-form solution:

$$\mathbf{w} = (X^{\intercal} X)^{-1} X^{\intercal} \mathbf{t}$$

The matrix form of the **normal equations** for the least squares problem.

Remarks

- One matrix equation = a system of $d+1$ simultaneous equations in $d+1$ unknowns

Remarks

- One matrix equation = a system of $d+1$ simultaneous equations in $d+1$ unknowns
- Checking the dimensions, everything works

Remarks

- One matrix equation = a system of $d+1$ simultaneous equations in $d+1$ unknowns
- Checking the dimensions, everything works
- X is $N \times (d+1)$ and $X^\intercal X$ is square $(d+1) \times (d+1)$,

Remarks

- One matrix equation = a system of $d+1$ simultaneous equations in $d+1$ unknowns
- Checking the dimensions, everything works
- X is $N \times (d+1)$ and $X^\intercal X$ is square $(d+1) \times (d+1)$,
- $(X^\intercal X)^{-1} X^\intercal =$ **Moore-Penrose pseudoinverse** of $X = X^\dagger$

$$\mathbf{w} = X^\dagger \mathbf{t}$$

Q: I have solved a linear regression problem.

Now I have received a new point $\mathbf{x}^*$ and I want to obtain the value y^* that estimates the most probable value of the output.

How do I proceed?

Q: I have solved a linear regression problem.

Now I have received a new point $\mathbf{x}^*$ and I want to obtain the value y^* that estimates the most probable value of the output.

How do I proceed?

A: This is **inference** and works simply like this:

$$y^* = \mathbf{w}^* \cdot \mathbf{x}^*$$

I.e. you apply the function representing your model to a new input, and you obtain the corresponding output...

6: Numerical issues and discussion

Invertibility and stability

- $X^\intercal X$ might not be invertible. This is when the data have exactly linearly dependent components

Invertibility and stability

- $X^\intercal X$ might not be invertible. This is when the data have exactly linearly dependent components
- What if the variables are **correlated**?

Invertibility and stability

- $X^\intercal X$ might not be invertible. This is when the data have exactly linearly dependent components
- What if the variables are **correlated**?
- Even if $X^\intercal X$ has full rank, in the case of correlated variables it will have a high **condition number** (largest eigenvalue)/(smallest eigenvalue).

Invertibility and stability

- $X^\intercal X$ might not be invertible. This is when the data have exactly linearly dependent components
- What if the variables are **correlated**?
- Even if $X^\intercal X$ has full rank, in the case of correlated variables it will have a high **condition number** (largest eigenvalue)/(smallest eigenvalue).
- DIFFICULT TO INVERT (numerical precision must be too high)

Invertibility and stability

- $X^\intercal X$ might not be invertible. This is when the data have exactly linearly dependent components
- What if the variables are **correlated**?
- Even if $X^\intercal X$ has full rank, in the case of correlated variables it will have a high **condition number** (largest eigenvalue)/(smallest eigenvalue).
- DIFFICULT TO INVERT (numerical precision must be too high)
- This is a problem of **numeric instability**: even very small numeric errors are amplified by the condition number and cause large errors on the result

Invertibility and stability

- $X^\intercal X$ might not be invertible. This is when the data have exactly linearly dependent components
- What if the variables are **correlated**?
- Even if $X^\intercal X$ has full rank, in the case of correlated variables it will have a high **condition number** (largest eigenvalue)/(smallest eigenvalue).
- DIFFICULT TO INVERT (numerical precision must be too high)
- This is a problem of **numeric instability**: even very small numeric errors are amplified by the condition number and cause large errors on the result

Invertibility and stability

- $X^\intercal X$ might not be invertible. This is when the data have exactly linearly dependent components
- What if the variables are **correlated**?
- Even if $X^\intercal X$ has full rank, in the case of correlated variables it will have a high **condition number** (largest eigenvalue)/(smallest eigenvalue).
- DIFFICULT TO INVERT (numerical precision must be too high)
- This is a problem of **numeric instability**: even very small numeric errors are amplified by the condition number and cause large errors on the result

SOLUTION: Iterative computation by successive approximations

Summary (I)

- Linear regression – An example of simple learning problem:
  Forecast one continuous variable using observations (= one or more other variables) related to it

Summary (I)

- Linear regression – An example of simple learning problem:
  Forecast one continuous variable using observations (= one or more other variables) related to it
- Univariate linear regression:
  Forecasting the most likely value of the target variable t as a linear function $y = w_1 x + w_0$ of the observed variable x

Summary (I)

- Linear regression – An example of simple learning problem:
  Forecast one continuous variable using observations (= one or more other variables) related to it
- Univariate linear regression:
  Forecasting the most likely value of the target variable t as a linear function $y = w_1 x + w_0$ of the observed variable x
- The problem is a system of N linear equations in two unknowns w_1, w_0
  – in general no solution

Summary (I)

- Linear regression – An example of simple learning problem:
  Forecast one continuous variable using observations (= one or more other variables) related to it
- Univariate linear regression:
  Forecasting the most likely value of the target variable t as a linear function $y = w_1 x + w_0$ of the observed variable x
- The problem is a system of N linear equations in two unknowns w_1, w_0
  – in general no solution
- Solved by minimizing an **objective function** that is the expectation of a chosen **loss function**

Summary (I)

- Linear regression – An example of simple learning problem:
  Forecast one continuous variable using observations (= one or more other variables) related to it
- Univariate linear regression:
  Forecasting the most likely value of the target variable t as a linear function $y = w_1 x + w_0$ of the observed variable x
- The problem is a system of N linear equations in two unknowns w_1, w_0
  – in general no solution
- Solved by minimizing an **objective function** that is the expectation of a chosen **loss function**
- We used the **square error loss** obtaining the **mean square error** objective

Summary (I)

- Linear regression – An example of simple learning problem:
  Forecast one continuous variable using observations (= one or more other variables) related to it
- Univariate linear regression:
  Forecasting the most likely value of the target variable t as a linear function $y = w_1 x + w_0$ of the observed variable x
- The problem is a system of N linear equations in two unknowns w_1, w_0
  – in general no solution
- Solved by minimizing an **objective function** that is the expectation of a chosen **loss function**
- We used the **square error loss** obtaining the **mean square error** objective
- The problem admits a closed-form solution

Summary (II)

- Multivariate linear regression:
  Forecasting the most likely value of the target variable t as a linear function $y = \mathbf{w} \cdot \mathbf{x}$ of the observed vector variable $\mathbf{x}$

Summary (II)

- Multivariate linear regression:
  Forecasting the most likely value of the target variable t as a linear function $y = \mathbf{w} \cdot \mathbf{x}$ of the observed vector variable $\mathbf{x}$
- The problem is a $N \times (d+1)$ system of linear equations (N equations in $d+1$ unknowns $\mathbf{w}$ - in general no solution if $d+1 < N$)

Summary (II)

- Multivariate linear regression:
  Forecasting the most likely value of the target variable t as a linear function $y = \mathbf{w} \cdot \mathbf{x}$ of the observed vector variable $\mathbf{x}$
- The problem is a $N \times (d+1)$ system of linear equations (N equations in $d+1$ unknowns $\mathbf{w}$ - in general no solution if $d+1 < N$)
- Again solved by minimizing the **mean square error** objective function

Summary (II)

- Multivariate linear regression:
  Forecasting the most likely value of the target variable t as a linear function $y = \mathbf{w} \cdot \mathbf{x}$ of the observed vector variable $\mathbf{x}$
- The problem is a $N \times (d+1)$ system of linear equations (N equations in $d+1$ unknowns $\mathbf{w}$ - in general no solution if $d+1 < N$)
- Again solved by minimizing the **mean square error** objective function
- The problem admits a closed-form solution using the **Moore-Penrose pseudo-inverse**

Summary (II)

- Multivariate linear regression:
  Forecasting the most likely value of the target variable t as a linear function $y = \mathbf{w} \cdot \mathbf{x}$ of the observed vector variable $\mathbf{x}$
- The problem is a $N \times (d+1)$ system of linear equations (N equations in $d+1$ unknowns $\mathbf{w}$ - in general no solution if $d+1 < N$)
- Again solved by minimizing the **mean square error** objective function
- The problem admits a closed-form solution using the **Moore-Penrose pseudo-inverse**
- However this solution may not exist or be low-quality due to numeric instability

Summary (II)

- Multivariate linear regression:
  Forecasting the most likely value of the target variable t as a linear function $y = \mathbf{w} \cdot \mathbf{x}$ of the observed vector variable $\mathbf{x}$
- The problem is a $N \times (d+1)$ system of linear equations (N equations in $d+1$ unknowns $\mathbf{w}$ - in general no solution if $d+1 < N$)
- Again solved by minimizing the **mean square error** objective function
- The problem admits a closed-form solution using the **Moore-Penrose pseudo-inverse**
- However this solution may not exist or be low-quality due to numeric instability
- Iterative approximation methods exist (e.g., by gradient descent)