Pattern Recognition 2015
Linear Models for Regression

Ad Feelders

Universiteit Utrecht

November 12, 2015

Linear Regression Model

The central assumption of linear regression is

$$\mathbb{E}[t|x] = y(x) = w_0 + w_1 x$$

Or, alternatively

$$t = w_0 + w_1 x + \varepsilon$$

with $\mathbb{E}[\varepsilon|x] = 0$.

Usually, we also assume that $\text{var}[t|x] = \sigma^2$, i.e. t has the same variance for each value of x.
For ML estimation, we typically assume $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

Minimizing empirical loss

Given training data

$$D = \{(x_1, t_1), (x_2, t_2), \ldots, (x_N, t_N)\},$$

find the values of w_0 and w_1 such that the sum of squared errors

$$\begin{aligned} E_D(w_0, w_1) &= \sum_{n=1}^{N} (t_n - \overbrace{(w_0 + w_1 x_n)}^{\text{prediction for } t_n})^2 \\ &= \sum_{n=1}^{N} (t_n - w_0 - w_1 x_n)^2 \end{aligned}$$

is minimized.

General Solution: Calculus

Partial derivative with respect to intercept:

$$\begin{aligned}\frac{\partial E_D}{\partial w_0} &= \sum_{n=1}^{N} 2(t_n - w_0 - w_1 x_n)(-1) \\ &= -2\sum_{n=1}^{N}(t_n - w_0 - w_1 x_n)\end{aligned}$$

Equate to zero

$$\sum_{n=1}^{N}(t_n - w_0 - w_1 x_n) = \sum_{n=1}^{N} e_n = 0$$

General Solution: Calculus

Partial derivative with respect to slope:

$$\frac{\partial E_D}{\partial w_1} = \sum_{n=1}^{N} 2(t_n - w_0 - w_1 x_n)(-x_n)$$
$$= -2\sum_{n=1}^{N} x_n(t_n - w_0 - w_1 x_n)$$

Equate to zero

$$\sum_{n=1}^{N} x_n(t_n - w_0 - w_1 x_n) = \sum_{n=1}^{N} x_n e_n = 0$$

General Solution: Calculus

Expand and collect terms:

$$\sum_{n=1}^{N} t_n = N w_0 + w_1 \sum_{n=1}^{N} x_n \tag{1}$$

$$\sum_{n=1}^{N} x_n t_n = w_0 \sum_{n=1}^{N} x_n + w_1 \sum_{n=1}^{N} x_n^2 \tag{2}$$

To solve for w_0 divide (1) by N:

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

Hence, the least squares fitted line goes through the point of means $(\bar{x}, \bar{t})$.

General Solution: Calculus

To solve for w_1, multiply (1) by $\sum x_n$ and (2) by N

$$\sum x_n \sum t_n = N w_0 \sum x_n + w_1 \left(\sum x_n\right)^2 \quad (3)$$

$$N \sum x_n t_n = N w_0 \sum x_n + N w_1 \sum x_n^2 \quad (4)$$

Subtract (3) from (4) and solve for w_1:

$$w_1 = \frac{N \sum x_n t_n - \sum x_n \sum t_n}{N \sum x_n^2 - (\sum x_n)^2}$$

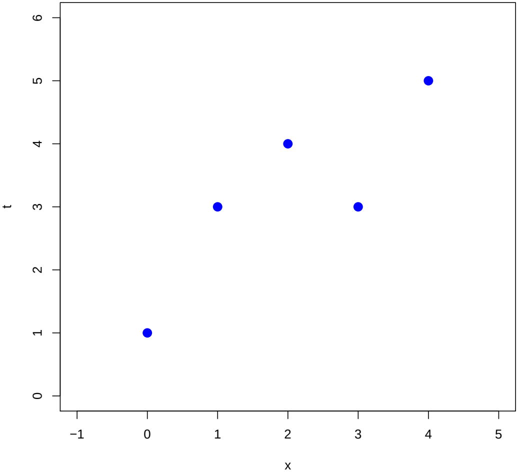
Example

n	x_n	t_n	$x_n t_n$	x_n^2
1	0	1	0	0
2	1	3	3	1
3	2	4	8	4
4	3	3	9	9
5	4	5	20	16
$\sum$	10	16	40	30

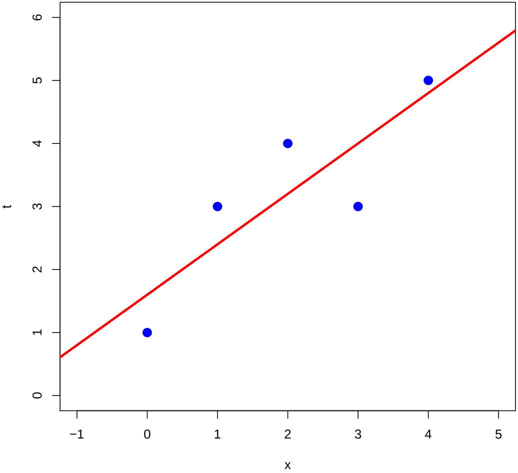
$$
\begin{aligned}
w_1 &= \frac{N \sum x_n t_n - \sum x_n \sum t_n}{N \sum x_n^2 - \left(\sum x_n\right)^2} \\
&= \frac{5 \times 40 - 10 \times 16}{5 \times 30 - 10^2} = \frac{4}{5}
\end{aligned}
$$

$$
w_0 = \bar{t} - w_1 \bar{x} = \frac{16}{5} - \left(\frac{4}{5}\right)\left(\frac{10}{5}\right) = \frac{8}{5}
$$

Scatter plot of Training Data



Fitted Line: $y(x) = 1.6 + 0.8\ x$



Decomposition of total sample variation in t

1. $\sum(t_n - \bar{t})^2$ = total sum of squares = SST
2. $\sum(y_n - \bar{t})^2$ = explained sum of squares = SSR
3. $\sum(t_n - y_n)^2 = \sum e_n^2$ = error sum of squares = SSE

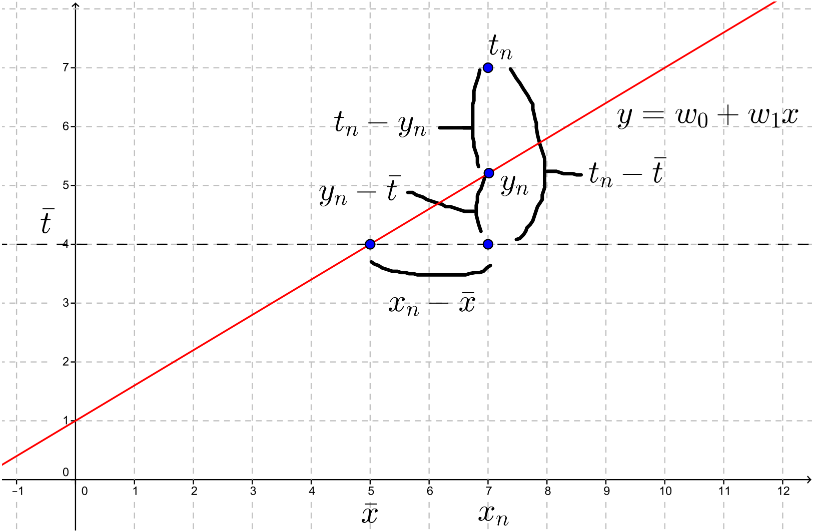
We have

$$\text{SST} = \text{SSR} + \text{SSE}$$

Proportion of variation in t explained by x:

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

Decomposition of variation in t



Example: Computation of R^2

Fitted model

$$y(x) = 1.6 + 0.8x$$

n	x_n	t_n	y_n	e_n	e_n^2	$(t_n - \bar{t})^2$	$(y_n - \bar{t})^2$
1	0	1	8/5	−3/5	9/25	121/25	64/25
2	1	3	12/5	3/5	9/25	1/25	16/25
3	2	4	16/5	4/5	16/25	16/25	0
4	3	3	20/5	−1	25/25	1/25	16/25
5	4	5	24/5	1/5	1/25	81/25	64/25
$\sum$	10	16	16	0	60/25	220/25	160/25

$$\begin{array}{ccccc} 220/25 & = & 60/25 & + & 160/25 \\ (SST) & & (SSE) & & (SSR) \end{array}$$

$$R^2 = \frac{SSR}{SST} = \frac{160}{220} \approx 0.73$$

Linear regression through the origin

Suppose we know that the population regression line goes through the origin, i.e.

$$\mathbb{E}[t|x] = wx$$

Find the value of w such that the sum of squared errors

$$E_D(w) = \sum_{n=1}^{N} (t_n - wx_n)^2$$

is minimized.

Regression through the origin: calculus

Take the derivative

$$\frac{dE_D}{dw} = -2\sum(t_n - wx_n)x_n$$

and equate to zero

$$\sum x_n t_n - w\sum x_n^2 = 0$$

so we get

$$w = \frac{\sum x_n t_n}{\sum x_n^2}$$

Regression through the origin: geometry

Regression through the origin: $y_n = wx_n$

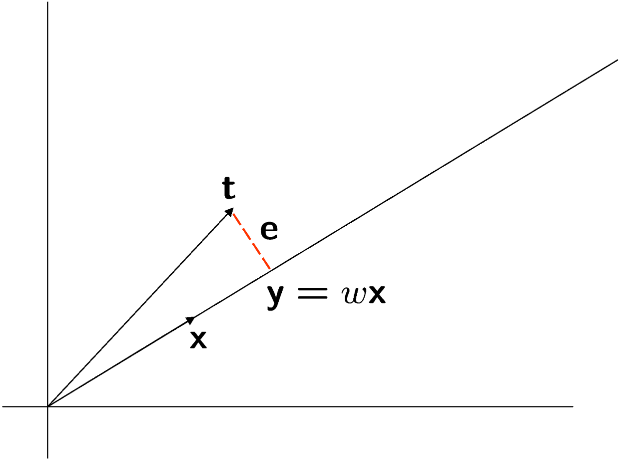
$D = \{(2,5),(1,3)\}$ contains only two observations.

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \text{ and } \mathbf{t} = \begin{bmatrix} t_1 \\ t_2 \end{bmatrix} = \begin{bmatrix} 5 \\ 3 \end{bmatrix}$$

$$\mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} \text{ and } \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

$$\mathbf{y} = w\mathbf{x} \text{ and } \mathbf{e} = \mathbf{t} - w\mathbf{x}$$

Least Squares Solution (N dimensional space!)



Length of $\mathbf{e} = \sqrt{\mathbf{e} \cdot \mathbf{e}} = \sqrt{e_1^2 + e_2^2}$.

Least Squares Solution

To minimize the length of $\mathbf{e}$, it must be perpendicular to $\mathbf{x}$ so $\mathbf{x} \cdot \mathbf{e} = 0$.

$$\mathbf{x} \cdot \mathbf{e} = \mathbf{x} \cdot (\mathbf{t} - w\mathbf{x}) = \mathbf{x} \cdot \mathbf{t} - w\mathbf{x} \cdot \mathbf{x} = 0$$

Therefore

$$w = \frac{\mathbf{x} \cdot \mathbf{t}}{\mathbf{x} \cdot \mathbf{x}}$$

Matrix notation

$$w = \frac{\mathbf{x}^\top \mathbf{t}}{\mathbf{x}^\top \mathbf{x}} \qquad \text{or} \qquad w = (\mathbf{x}^\top \mathbf{x})^{-1} \mathbf{x}^\top \mathbf{t}$$

Solution of Numerical Example

Solution of the numerical example

$$\mathbf{x}^\top \mathbf{t} = [2 \;\; 1] \begin{bmatrix} 5 \\ 3 \end{bmatrix} = 13$$

and

$$\mathbf{x}^\top \mathbf{x} = [2 \;\; 1] \begin{bmatrix} 2 \\ 1 \end{bmatrix} = 5$$

which yields

$$w = \frac{\mathbf{x}^\top \mathbf{t}}{\mathbf{x}^\top \mathbf{x}} = \frac{13}{5} = 2.6$$

Simple linear regression in matrix terms

We can write the observed t values as

$$t_n = w_0 + w_1 x_n + e_n \qquad\qquad n = 1, \ldots, N$$

which is short for

$$\begin{aligned} t_1 &= w_0 + w_1 x_1 + e_1 \\ t_2 &= w_0 + w_1 x_2 + e_2 \\ &\vdots \\ t_N &= w_0 + w_1 x_N + e_N \end{aligned}$$

Matrix Notation

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_N \end{bmatrix} \mathbf{t} = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{bmatrix} \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix} \mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$

Then we can simply write

$$\mathbf{t} = \mathbf{X}\mathbf{w} + \mathbf{e}$$

Check

$$\mathbf{t} = \mathbf{X}\mathbf{w} + \mathbf{e}$$

$$\begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_N \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix}$$

$$= \begin{bmatrix} w_0 + w_1 x_1 \\ w_0 + w_1 x_2 \\ \vdots \\ w_0 + w_1 x_N \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix}$$

$$= \begin{bmatrix} w_0 + w_1 x_1 + e_1 \\ w_0 + w_1 x_2 + e_2 \\ \vdots \\ w_0 + w_1 x_N + e_N \end{bmatrix}$$

Least Squares Solution

$\mathbf{y}$ is a linear combination of the columns of $\mathbf{X}$:

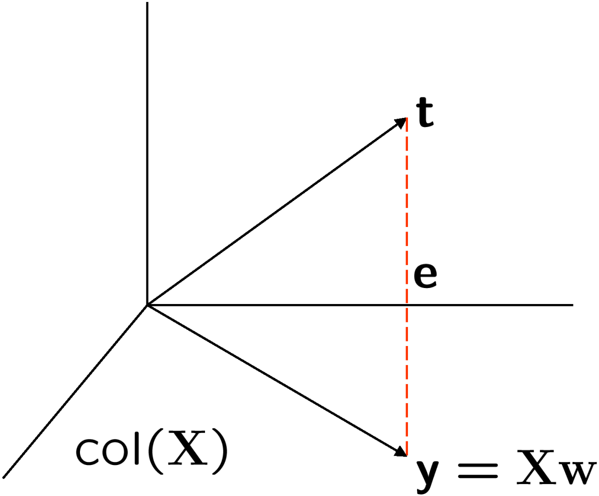
$$\mathbf{y} = \mathbf{X}\mathbf{w}$$

Typically, $\mathbf{t}$ is not in the column space of $\mathbf{X}$. Find the value of $\mathbf{y}$ that is closest to $\mathbf{t}$. For this to be the case, the error vector

$$\mathbf{e} = \mathbf{t} - \mathbf{X}\mathbf{w}$$

must be orthogonal to *all columns* of $\mathbf{X}$.

Least Squares Solution (N dimensional space)



Least Squares Solution

In other words, we should have

$$\mathbf{X}^\top \mathbf{e} = \mathbf{0}.$$

Since

$$\mathbf{e} = (\mathbf{t} - \mathbf{X}\mathbf{w})$$

we should have

$$\mathbf{X}^\top(\mathbf{t} - \mathbf{X}\mathbf{w}) = \mathbf{X}^\top\mathbf{t} - \mathbf{X}^\top\mathbf{X}\mathbf{w} = \mathbf{0}.$$

It follows that

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

Least Squares Solution

So we have

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

Premultiply both sides by the inverse of $\mathbf{X}^\top \mathbf{X}$:

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

We then find, since $(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{X} = \mathbf{I}$ and $\mathbf{I}\mathbf{w} = \mathbf{w}$:

$$\mathbf{w} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{t} \tag{3.15}$$

Numeric example

$$D = \{(0,1),(1,1),(2,2),(3,2)\}$$

$$\mathbf{X} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \quad \mathbf{t} = \begin{bmatrix} 1 \\ 1 \\ 2 \\ 2 \end{bmatrix}$$

$$\mathbf{X}^\top \mathbf{X} = \begin{bmatrix} 4 & 6 \\ 6 & 14 \end{bmatrix} \quad \mathbf{X}^\top \mathbf{t} = \begin{bmatrix} 6 \\ 11 \end{bmatrix}$$

Numeric Example

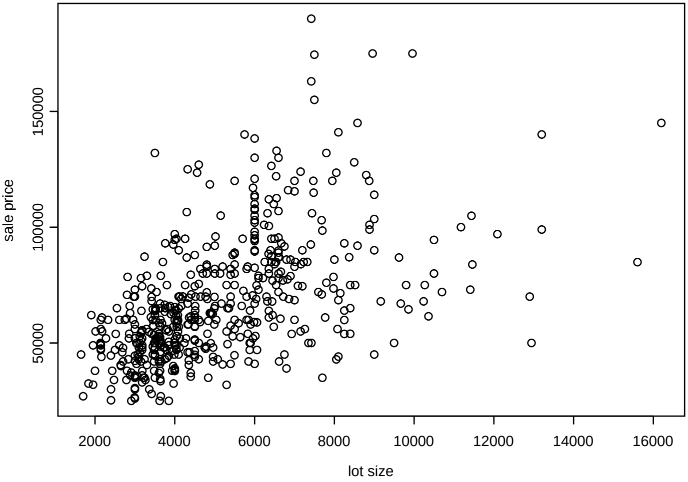
Now, since

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

we get

$$\begin{aligned} \mathbf{w} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{t} &= \frac{1}{20}\begin{bmatrix} 14 & -6 \\ -6 & 4 \end{bmatrix}\begin{bmatrix} 6 \\ 11 \end{bmatrix} \\ &= \frac{1}{20}\begin{bmatrix} 18 \\ 8 \end{bmatrix} = \begin{bmatrix} 9/10 \\ 4/10 \end{bmatrix} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \end{aligned}$$

Scatter plot of lot size and sale price



Least Squares fitted line

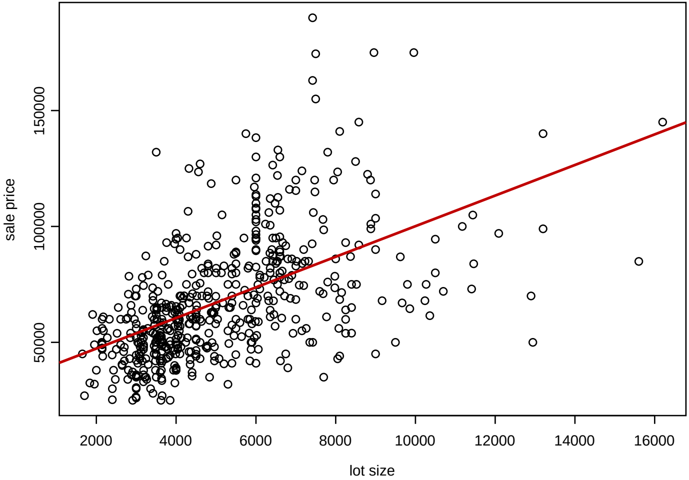
Using R we find:

$$\text{sale price} = 34136.1916 + 6.5988 \times \text{lot size}$$

$R^2 = 0.2871$

Model explains only about 30% of variation in sale price.

Least Squares fitted line



Multiple Linear Regression

Usually, you want to use more than one input variable to predict t.

The basic assumption is

$$\mathbb{E}[t|\mathbf{x}] = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_{M-1} x_{M-1}$$

Multiple Linear Regression

We can write the observed t values as

$$t_n = w_0 + w_1 x_{n,1} + w_2 x_{n,2} + \ldots + w_{M-1} x_{n,M-1} + e_n$$

which is short for

$$\begin{aligned}
t_1 &= w_0 + w_1 x_{1,1} + w_2 x_{1,2} + \ldots + w_{M-1} x_{1,M-1} + e_1 \\
t_2 &= w_0 + w_1 x_{2,1} + w_2 x_{2,2} + \ldots + w_{M-1} x_{2,M-1} + e_2 \\
&\vdots \\
t_N &= w_0 + w_1 x_{N,1} + w_2 x_{N,2} + \ldots + w_{M-1} x_{N,M-1} + e_N
\end{aligned}$$

Notation and Least Squares Solution

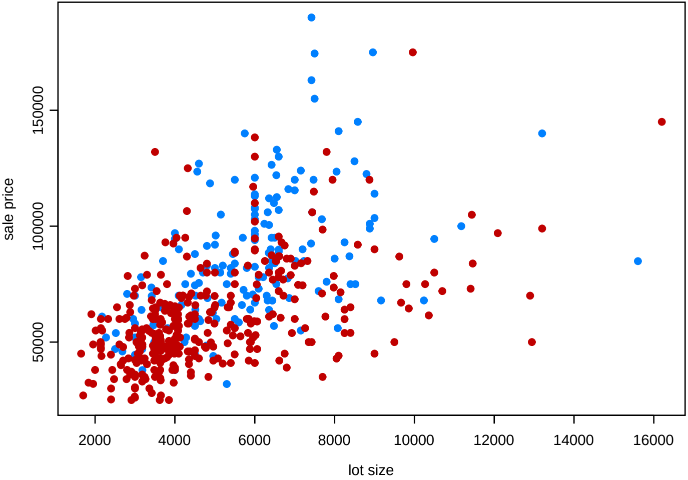
$$\mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,M-1} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,M-1} \\ \vdots & & & & \\ 1 & x_{N,1} & x_{N,2} & \dots & x_{N,M-1} \end{bmatrix}$$

$$\mathbf{t} = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{bmatrix} \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix} \mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_{M-1} \end{bmatrix}$$

Then we can write

$$\mathbf{t} = \mathbf{X}\mathbf{w} + \mathbf{e}, \quad \mathbf{w} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{t}$$

Scatter plot of lot size, airco and sale price



Fitted Equation

$$\text{sale.price} = 32692.9 + 5.6 \times \text{lot.size} + 20174.5 \times \text{air.cond}$$

Or, since air.cond is binary:

$$\text{sale.price} = 32692.9 + 5.6 \times \text{lot.size}$$

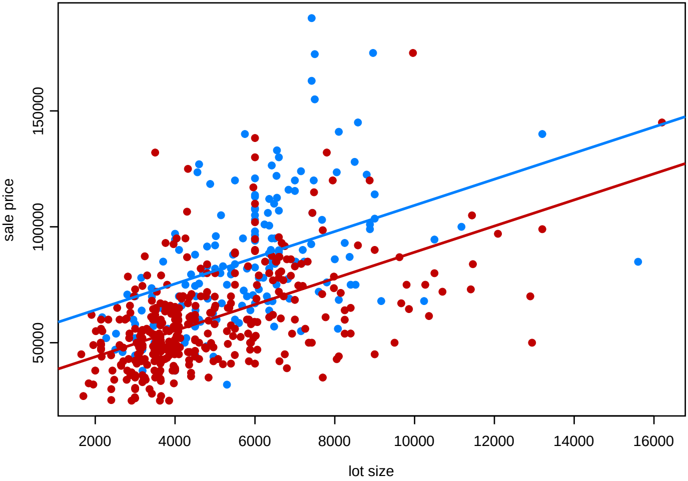
when air.cond=0.

$$\text{sale.price} = (32692.9 + 20174.5) + 5.6 \times \text{lot.size}$$

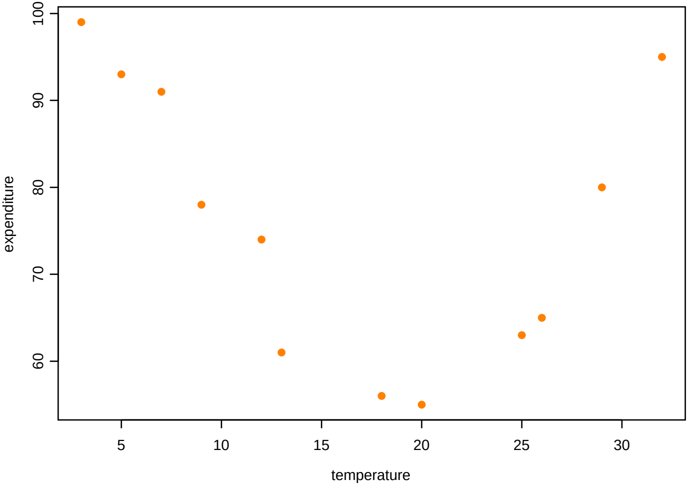
when air.cond=1.

$R^2 = 0.4048$

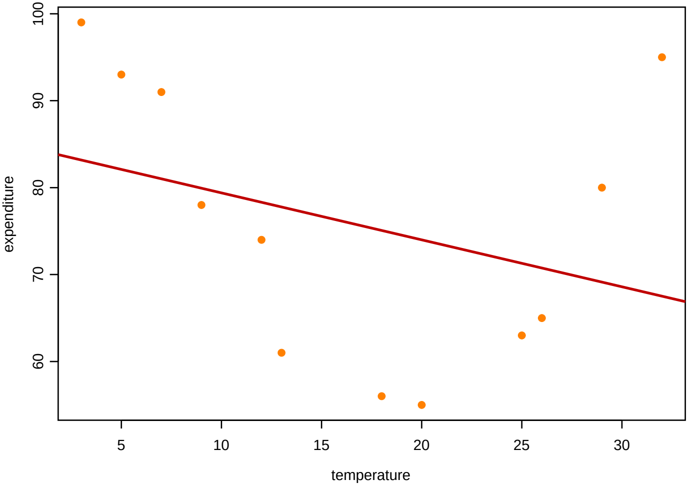
Fitted Equation



Scatter plot of Temperature and Energy Use



Fitting a Linear Function



Linear Equation

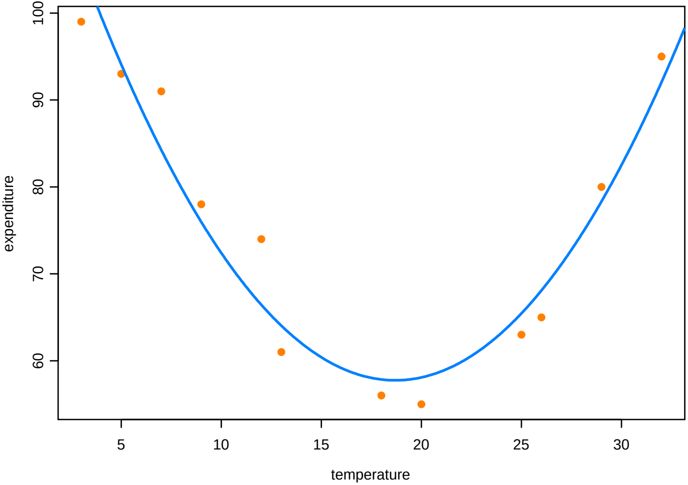
Fitted equation:

$$\text{expenditure} = 84.78 - 0.54 \times \text{temperature}$$

$R^2 \approx 0.11$

Bad fit!

Fitting a Quadratic Function



Quadratic Equation

Fitted equation:

$$\text{expenditure} = 125.44 - 7.24 \times \text{temp} + 0.19 \times \text{temp}^2$$

$R^2 \approx 0.93$

Spectacular improvement for only one extra parameter!

General Linear Model (OK, so I lied ...)

The term *linear* in linear regression means linear in the *parameters*, not linear in the *input variables*!

For example:

$$t = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1^2 + w_4 x_1 x_2 + \varepsilon$$

is still a linear regression model.

General Linear Model

Linear in the features, but not necessarily linear in the input variables that generate them.

$$\phi(t_n) = w_0\phi_0(\mathbf{x}_n) + w_1\phi_1(\mathbf{x}_n) + \ldots + w_{M-1}\phi_{M-1}(\mathbf{x}_n) + \varepsilon_n$$

- $\mathbf{x}_n = (x_{n,1}, \ldots, x_{n,D})$
- $w_0, w_1, \ldots, w_{M-1}$: unknown parameters to be estimated;
- $\phi(\cdot), \phi_0(\cdot), \ldots, \phi_{M-1}(\cdot)$: functions that do not involve unknown parameters; "basis functions".

Modeling Pizza Expenditure

Linear Model

$$\mathbb{E}(\text{pizza}) = w_0 + w_1 \times \text{inc} + w_2 \times \text{age}$$

$$\frac{\partial \mathbb{E}(\text{pizza})}{\partial \text{inc}} = w_1 \qquad \frac{\partial \mathbb{E}(\text{pizza})}{\partial \text{age}} = w_2$$

Fitted equation:

$$\text{pizza} = 342.88 + 0.0024 \times \text{inc} - 7.58 \times \text{age}$$

$R^2 \approx 0.33$

Modeling Pizza Expenditure

Model with interaction between income and age:

$$\mathbb{E}(\text{pizza}) = w_0 + w_1 \times \text{inc} + w_2 \times \text{age} + w_3 \times (\text{age} \times \text{inc})$$

Effects of income and age:

$$\frac{\partial \mathbb{E}(\text{pizza})}{\partial \text{inc}} = w_1 + w_3 \times \text{age}$$

$$\frac{\partial \mathbb{E}(\text{pizza})}{\partial \text{age}} = w_2 + w_3 \times \text{inc}$$

Modeling Pizza Expenditure

Fitted equation:

$$\begin{aligned} \text{pizza} &= 161.47 + 0.01 \times \text{inc} - 2.98 \times \text{age} \\ &- 0.0002 \times (\text{age} \times \text{inc}) \end{aligned}$$

$R^2 \approx 0.39$

Effect of income on pizza expenditure

$$\begin{aligned}\frac{\partial \mathbb{E}(\text{pizza})}{\partial \text{inc}} &= w_1 + w_3 \times \text{age} = 0.01 - 0.0002 \times \text{age} \\ &= \begin{cases} 0.006 & \text{for age} = 20 \\ 0 & \text{for age} = 50 \end{cases}\end{aligned}$$

So on average, a 20 year old will spend 60 cents on pizza of every 100 dollar of extra income.

How to in R

```
# read data (put header=T if first row in data file contains
# names of variables)

> pizza.dat <- read.table("C:/PR/pizza.txt", header=T)

# show first 5 rows

> pizza.dat[1:5,]

  pizza sex edu1 edu2 edu3 income age
1   109   1    0    0    0  15000  25
2     0   1    0    0    0  30000  45
3     0   1    0    0    0  12000  20
4   108   1    0    0    0  20000  28
5   220   1    1    0    0  15000  25
```

How to in R

```
# fit model with interaction between age and income

> pizza.model1 <- lm(pizza ~ age + income + age:income,
data=pizza.dat)

# show results (stuff deleted)

> summary(pizza.model1)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.615e+02  1.207e+02   1.338   0.1892
age         -2.977e+00  3.352e+00  -0.888   0.3803
income       9.074e-03  3.670e-03   2.473   0.0183 *
age:income  -1.602e-04  8.673e-05  -1.847   0.0730 .

Multiple R-Squared: 0.3873
```

Interaction with a binary variable: house prices

$$\mathbb{E}(\text{sale.price}) = w_0 + w_1 \times \text{lot.size} + w_2 \times (\text{air.cond} \times \text{lot.size})$$

Price per square foot depends on presence of airco:

$$\mathbb{E}(\text{sale.price}) = w_0 + w_1 \times \text{lot.size}$$

if no airco (air.cond=0), and

$$\mathbb{E}(\text{sale.price}) = w_0 + (w_1 + w_2) \times \text{lot.size}$$

if air.cond=1.

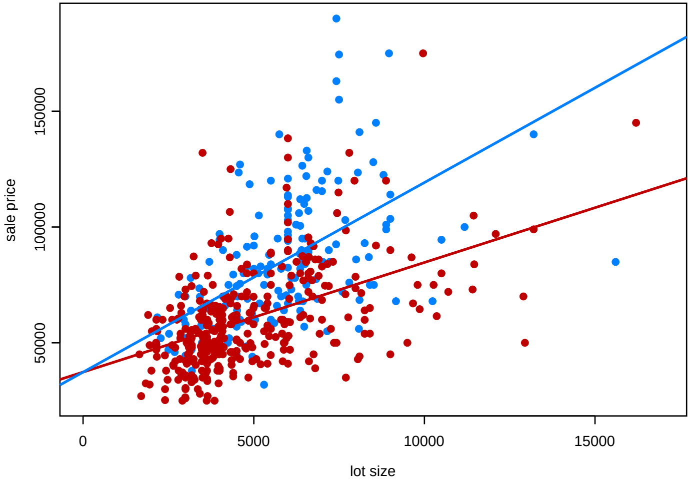
Fitted Equation

Fitted equation:

$$\begin{aligned}\text{sale.price} &= 37341.69 + 4.73 \times \text{lot.size} + \\ &\quad 3.45 \times (\text{air.cond} \times \text{lot.size})\end{aligned}$$

$R^2 \approx 0.41$.

Graph of fitted equation



Detecting influenza epidemics using search engine query data

1. Weekly counts for 50 million of the most common search queries in the US between 2003 and 2008.
2. Separate aggregate weekly counts were kept for every query in each state.
3. Divide by total number of queries during that week to obtain a query fraction.
4. CDC publishes national and regional data on influenza-like illness (ILI) physician visits on a weekly basis, typically with a 1-2 week reporting lag.

Detecting influenza epidemics using search engine query data

Fit linear regression model:

$$\text{logit}(P) = w_0 + w_1 \text{logit}(Q) + \varepsilon$$

where P is the percentage of ILI physician visits, Q is the ILI-related query fraction, and

$$\text{logit}(P) = \ln \left\{ \frac{P}{1-P} \right\}$$

Note that P and Q are fractions, i.e. numbers between 0 and 1.

Detecting influenza epidemics using search engine query data

Given fitted values for w_0 and w_1, and a value for Q, predict

$$\text{logit}(P) = w_0 + w_1 \text{logit}(Q),$$

or

$$P = \frac{e^{w_0 + w_1 \text{logit}(Q)}}{1 + e^{w_0 + w_1 \text{logit}(Q)}}$$

1. First this model is fitted for 50 million candidate queries.
2. Sort queries on correlation between prediction and true target values.
3. Compare models including top N queries ($N = 1, \ldots, 100$).

Detecting influenza epidemics using search engine query data

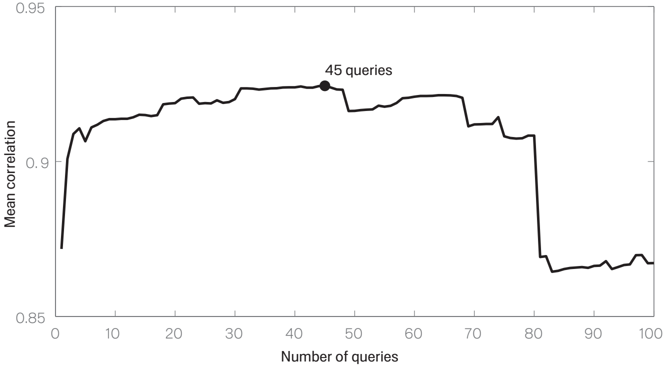


Figure 1: An evaluation of how many top-scoring queries to include in the ILI-related query fraction. Maximal performance at estimating out-of-sample points during cross-validation was obtained by summing the top 45 search queries. A steep drop in model performance occurs after adding query 81, which is “oscar nominations”.

Detecting influenza epidemics using search engine query data

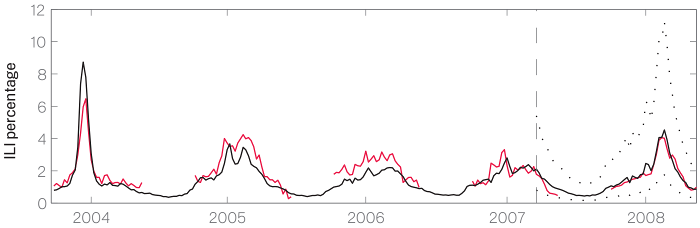


Figure 2: A comparison of model estimates for the Mid-Atlantic Region (black) against CDC-reported ILI percentages (red), including points over which the model was fit and validated. A correlation of 0.85 was obtained over 128 points from this region to which the model was fit, while a correlation of 0.96 was obtained over 42 validation points. 95% prediction intervals are indicated.

Detecting influenza epidemics using search engine query data

Search Query Topic	Top 45 Queries N	Top 45 Queries Weighted	Next 55 Queries N	Next 55 Queries Weighted
Influenza Complication	11	18.15	5	3.40
Cold/Flu Remedy	8	5.05	6	5.03
General Influenza Symptoms	5	2.60	1	0.07
Term for Influenza	4	3.74	6	0.30
Specific Influenza Symptom	4	2.54	6	3.74
Symptoms of an Influenza Complication	4	2.21	2	0.92
Antibiotic Medication	3	6.23	3	3.17
General Influenza Remedies	2	0.18	1	0.32
Symptoms of a Related Disease	2	1.66	2	0.77
Antiviral Medication	1	0.39	1	0.74
Related Disease	1	6.66	3	3.77
Unrelated to Influenza	0	0.00	19	28.37
	45	**49.40**	**55**	**50.60**

Table 1: Topics found in search queries which were found to be most correlated with CDC ILI data. The top 45 queries were used in our final model; the next 55 queries are presented for comparison purposes. The number of queries in each topic is indicated, as well as query volume-weighted counts, reflecting the relative frequency of queries in each topic.

Regularized Least Squares

Add regularization term to control overfitting

$$E(\mathbf{w}) = E_D(\mathbf{w}) + \lambda E_W(\mathbf{w}) \tag{3.24}$$

Ridge regression

$$E_W(\mathbf{w}) = \sum_i w_i^2 = \mathbf{w}^\top \mathbf{w} \tag{3.25}$$

The ridge regression error function is minimized by:

$$\mathbf{w} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{t} \tag{3.28}$$

Regularization

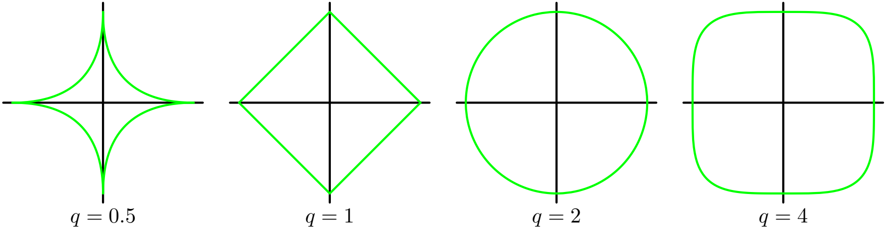
A more general regularizer

$$\sum_{n=1}^{N}\{t_n - \mathbf{w}^\top \mathbf{x}_n\}^2 + \lambda \sum_{j=0}^{M-1} |w_j|^q \tag{3.29}$$

where $q = 2$ corresponds the ridge regression.

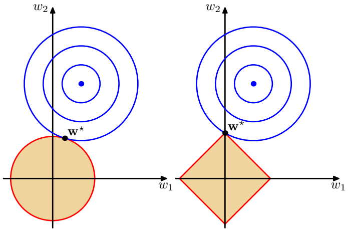
The case of $q = 1$ is known as the LASSO.

Contours of Regularization Term



LASSO RIDGE

LASSO gives sparse solution



Minimize

$$E_D(\mathbf{w}) = \sum_{n=1}^{N} \{t_n - \mathbf{w}^\top \mathbf{x}_n\}^2 \tag{3.12}$$

subject to

$$\sum_{j=0}^{M-1} |w_j|^q \leq \eta \tag{3.30}$$

for an appropriate value of the parameter η.