Chapter II.3

1. Hypothesis testing
2. Linear regression
    2.1. Regularizers
    2.2. Model selection
3. Logistic regression
4. Summary

Hypothesis testing

- Suppose we throw a coin n times and we want to estimate if the coin is fair, i.e. if Pr(heads) = Pr(tails).
- Let $X_1, X_2, \ldots, X_n \sim$ Bernoulli(p) be the i.i.d. coin flips
  - Coin is fair $\Leftrightarrow p = 1/2$
- Let the **null hypothesis** H_0 be "coin is fair".
- The **alternative hypothesis** H_1 is then "coin is not fair"
- Intuitively, if $|n^{-1}\sum_i X_i - 1/2|$ is large, we should reject the null hypothesis
- *But can we formalize this?*

Hypothesis testing terminology

- $\theta = \theta_0$ is called **simple hypothesis**
- $\theta > \theta_0$ or $\theta < \theta_0$ is called **composite hypothesis**
- H_0: $\theta = \theta_0$ vs. H_1: $\theta \neq \theta_0$ is called **two-sided test**
- H_0: $\theta \leq \theta_0$ vs. H_1: $\theta > \theta_0$ and H_0: $\theta \geq \theta_0$ vs. H_1: $\theta < \theta_0$ are called **one-sided tests**
- **Rejection region** R: if $X \in R$, reject H_0 o/w retain H_0
  - Typically $R = \{x : T(x) > c\}$ where T is a **test statistic** and c is a **critical value**
- **Error types:**

	Retain H_0	Reject H_0
H_0 true	✓	type I error
H_1 true	type II error	✓

The *p*-values

- The *p*-value is the *probability that* ***if*** *H_0* ***holds****, we observe values at least as extreme as the test statistic*
  - It is *not* the probability that H_0 holds
  - If *p*-value is small enough, we can reject H_0
  - How small is small enough depends on application
- Typical *p*-value scale:

p-value	evidence
< 0.01	very strong evidence against H_0
0.01–0.05	strong evidence against H_0
0.05–0.1	weak evidence against H_0
> 0.1	little or no evidence against H_0

The Wald test

For two-sided test H_0: $\theta = \theta_0$ vs. H_1: $\theta \neq \theta_0$

Test statistic $W = \dfrac{\hat{\theta} - \theta_0}{\hat{\mathsf{se}}}$, where $\hat{\theta}$ is the sample estimate and

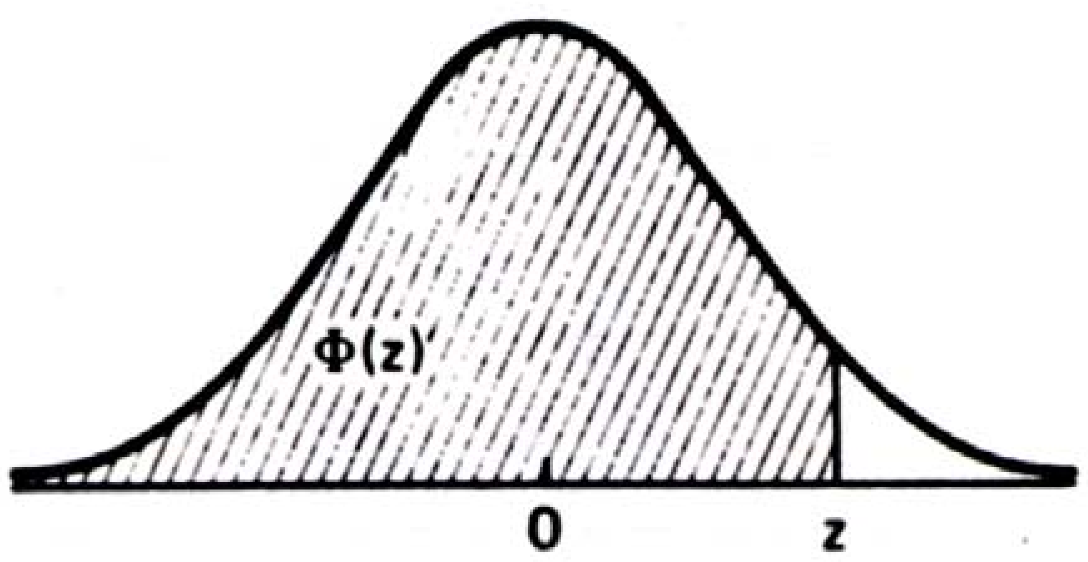
$\hat{\mathsf{se}} = \mathsf{se}(\hat{\theta}) = \sqrt{\mathrm{Var}[\hat{\theta}]}$ is the standard error.

W converges in probability to N(0,1).

If w is the observed value of Wald statistic, the p-value is $2\Phi(-|w|)$.

The coin-tossing example revisited

Using Wald test we can test if our coin is fair. Suppose the observed average is 0.6 with estimated standard error 0.049. The observed Wald statistic w is now $w = (0.6 - 0.5)/0.049 \approx 2.04$. Therefore the p-value is $2\Phi(-2.04) \approx 0.041$, and we have strong evidence to reject the null hypothesis.



The χ^2 distribution

Let $X_1, X_2, \ldots, X_n$ be i.i.d. N(0,1) distributed random variables.

The random variable $\chi_n^2 = \sum_{i=1}^{n} X_i^2$

is χ^2-distributed with n degrees of freedom.

$$f(x) = \frac{x^{(n/2)-1} e^{-x/2}}{2^{n/2}\Gamma(n/2)} \quad \text{for } x > 0$$

$E[x] = n$

$Var[x] = 2n$

Pearson's χ^2 test for multinomial data

If $X = (X_1, X_2, \ldots, X_k)$ has Multinomial(n, $\boldsymbol{p}$) distribution, then MLE of $\boldsymbol{p}$ is $(X_1/n, X_2/n, \ldots, X_k/n)$. Let $\boldsymbol{p}_0 = (p_{01}, p_{02}, \ldots, p_{0k})$ and we want to test H_0: $\boldsymbol{p} = \boldsymbol{p}_0$ vs. H_1: $\boldsymbol{p} \neq \boldsymbol{p}_0$.

Pearson's χ^2 statistic is

$$T = \sum_{j=1}^{k} \frac{(X_j - np_{0j})^2}{np_{0j}} = \sum_{j=1}^{k} \frac{(X_j - E_j)^2}{E_j}$$

where $E_j = \mathrm{E}[X_j] = np_{0j}$ is the expected value of X_j under H_0.

The p-value is $\Pr(\chi^2_{k-1} > t)$ where t is the observed value of T.

Extending Pearson to non-multinomial

- Pearson's χ^2 can be used to test the fitness of sample to *any* distribution (goodness-of-fit test)
- Let $X_1, X_2, \ldots, X_n$ be the sample and $f(x; \boldsymbol{\theta})$ some probability distribution with parameters $\boldsymbol{\theta}$
- Divide the possible values of X_is (under the null hypothesis) into k disjoint intervals and let Oj be the number of times we see value in interval I_j
- Compute the theoretical interval frequencies $p_j(\theta) = \int_{I_j} f(x; \theta)dx$
- Obtain estimates $\tilde{\theta}$ by maximizing

$$Q(\theta) = \prod_{j=1}^{k} p_i(\theta)^{O_j}$$

- Now the multinomial χ^2 test applies with $k-1-s$ degrees of freedom, where s is the number of parameters in $\boldsymbol{\theta}$

Extending Pearson to test of independence

- Pearson's χ^2 can also be used to test the independence of two variables
- Let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ be two samples
- Divide the outcomes into r (for X_is) and c disjoint intervals and compute the frequencies
- Populate r-by-c table O with the frequencies (O_{lk} tells how many (X_i, Y_i) pairs have values from rth and cth interval, respectively)
- Assuming independency, the expected value for O_{lk} is

$$E_{lk} = \frac{\sum_{j=1}^{c} O_{lj} \sum_{j=1}^{r} O_{jk}}{\sum_{i=1}^{r} \sum_{j=1}^{c} O_{ij}}$$

- The value of the test statistic is $\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$
- There are $(r-1)(c-1)$ degrees of freedom

χ^2 distribution table

1.1.2.10. Obere 100α-prozentige Werte χ^2_α der χ^2-Verteilung (s. 5.2.3.)

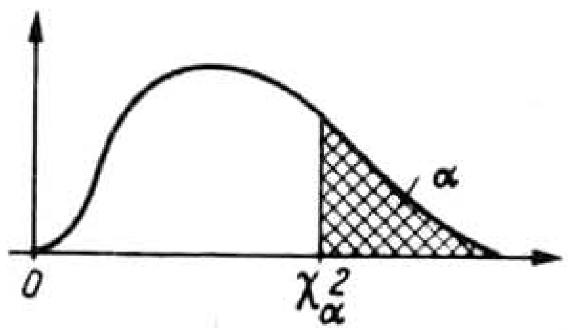


Abb. 1.4

Anzahl der Freiheits-grade m	Wahrscheinlichkeit $p = \alpha$															
	0,99	0,98	0,95	0,90	0,80	0,70	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,005	0,002	0,001
1	0,00016	0,0006	0,0039	0,016	0,064	0,148	0,455	1,07	1,64	2,7	3,8	5,4	6,6	7,9	9,5	10,83
2	0,020	0,040	0,103	0,211	0,446	0,713	1,386	2,41	3,22	4,6	6,0	7,8	9,2	10,6	12,4	13,8
3	0,115	0,185	0,352	0,584	1,005	1,424	2,366	3,67	4,64	6,3	7,8	9,8	11,3	12,8	14,8	16,3
4	0,30	0,43	0,71	1,06	1,65	2,19	3,36	4,9	6,0	7,8	9,5	11,7	13,3	14,9	16,9	18,5
5	0,55	0,75	1,14	1,61	2,34	3,00	4,35	6,1	7,3	9,2	11,1	13,4	15,1	16,8	18,9	20,5
6	0,87	1,13	1,63	2,20	3,07	3,83	5,35	7,2	8,6	10,6	12,6	15,0	16,8	18,5	20,7	22,5
7	1,24	1,56	2,17	2,83	3,82	4,67	6,35	8,4	9,8	12,0	14,1	16,6	18,5	20,3	22,6	24,3
8	1,65	2,03	2,73	3,49	4,59	5,53	7,34	9,5	11,0	13,4	15,5	18,2	20,1	22,0	24,3	26,1
9	2,09	2,53	3,32	4,17	5,38	6,39	8,34	10,7	12,2	14,7	16,9	19,7	21,7	23,6	26,1	27,9
10	2,56	3,06	3,94	4,86	6,18	7,27	9,34	11,8	13,4	16,0	18,3	21,2	23,2	25,2	27,7	29,6
11	3,1	3,6	4,6	5,6	7,0	8,1	10,3	12,9	14,6	17,3	19,7	22,6	24,7	26,8	29,4	31,3
12	3,6	4,2	5,2	6,3	7,8	9,0	11,3	14,0	15,8	18,5	21,0	24,1	26,2	28,3	30,9	32,9
13	4,1	4,8	5,9	7,0	8,6	9,9	12,3	15,1	17,0	19,8	22,4	25,5	27,7	29,8	32,5	34,5
14	4,7	5,4	6,6	7,8	9,5	10,8	13,3	16,2	18,2	21,1	23,7	26,9	29,1	31,3	34,0	36,1
15	5,2	6,0	7,3	8,5	10,3	11,7	14,3	17,3	19,3	22,3	25,0	28,3	30,6	32,8	35,6	37,7
16	5,8	6,6	8,0	9,3	11,2	12,6	15,3	18,4	20,5	23,5	26,3	29,6	32,0	34,3	37,1	39,3
17	6,4	7,3	8,7	10,1	12,0	13,5	16,3	19,5	21,6	24,8	27,6	31,0	33,4	35,7	38,6	40,8
18	7,0	7,9	9,4	10,9	12,9	14,4	17,3	20,6	22,8	26,0	28,9	32,3	34,8	37,2	40,1	42,3
19	7,6	8,6	10,1	11,7	13,7	15,4	18,3	21,7	23,9	27,2	30,1	33,7	36,2	38,6	41,6	43,8
20	8,3	9,2	10,9	12,4	14,6	16,3	19,3	22,8	25,0	28,4	31,4	35,0	37,6	40,0	43,0	45,3
21	8,9	9,9	11,6	13,2	15,4	17,2	20,3	23,9	26,2	29,6	32,7	36,3	38,9	41,4	44,5	46,8
22	9,5	10,6	12,3	14,0	16,3	18,1	21,3	24,9	27,3	30,8	33,9	37,7	40,3	42,8	45,9	48,3
23	10,2	11,3	13,1	14,8	17,2	19,0	22,3	26,0	28,4	32,0	35,2	39,0	41,6	44,2	47,3	49,7
24	10,9	12,0	13,8	15,7	18,1	19,9	23,3	27,1	29,6	33,2	36,4	40,3	43,0	45,6	48,7	51,2
25	11,5	12,7	14,6	16,5	18,9	20,9	24,3	28,2	30,7	34,4	37,7	41,6	44,3	46,9	50,1	52,6
26	12,2	13,4	15,4	17,3	19,8	21,8	25,3	29,2	31,8	35,6	38,9	42,9	45,6	48,3	51,6	54,1
27	12,9	14,1	16,2	18,1	20,7	22,7	26,3	30,3	32,9	36,7	40,1	44,1	47,0	49,6	52,9	55,5
28	13,6	14,8	16,9	18,9	21,6	23,6	27,3	31,4	34,0	37,9	41,3	45,4	48,3	51,0	54,4	56,9

Testing with implicit distribution

- Suppose we have found association rule "diapers" ⇒ "beer" with confidence 0.9
  - I.e. **E**["x buys beer" | "x buys diapers"] = 0.9 in the sample
- Possible explanation: everybody buys beer
  - Result is not interesting
    - also "vegetables" ⇒ "beer" has high confidence, etc.
  - Null hypothesis: "Result is due to the fact that (almost) everybody buys beer"
- How can we test that?

Testing “diapers” ⇒ “beer”, part 1

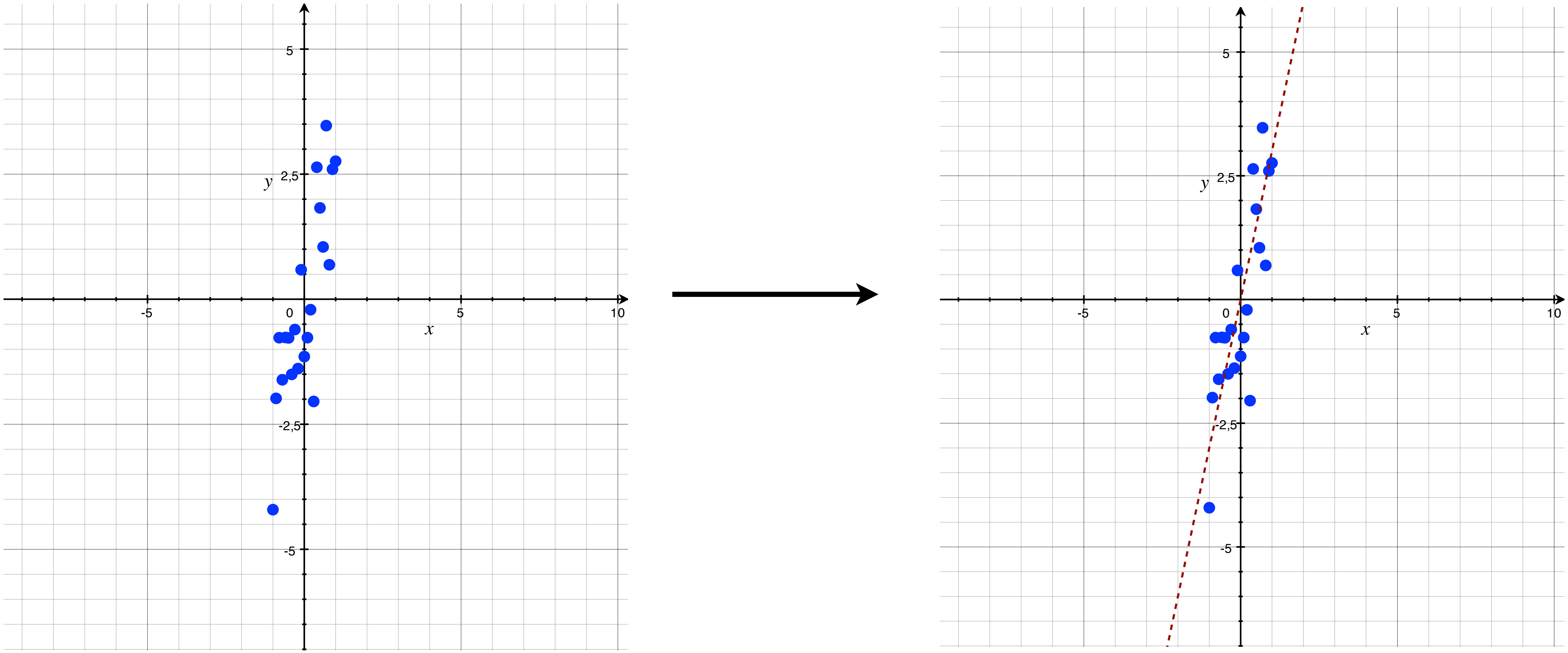
- The idea: generate data sets that have similar properties to the real data, but are random
  - See how good your result is in these random data sets
  - Let N be the number of data sets and M the number of times the result is at least as good in random data than it is in the real data
  - The empirical p-value is then $(M + 1)/(N + 1)$
- Independent random data:
  - Data is n-by-m (m items) binary matrix
  - Let $\boldsymbol{c}$ be m-dimensional vector of column margins
  - Make random matrix (a_{ij}) by sampling $a_{ij} \sim \text{Bernoulli}(c_j)$

Testing "diapers" ⇒ "beer", part 2

- Independent random samples have estimated column margins $\boldsymbol{c}$
- They do not take into account that some people buy many different things while others buy only few
  - Compute also row margins $\boldsymbol{r}$
- Let $\mathcal{M}(\boldsymbol{r}, \boldsymbol{c})$ be a family of 0/1 matrices with row margin $\boldsymbol{r}$ and column margin $\boldsymbol{c}$
  - Sample u.a.r. from this family and test in that sample
- Problem: how to sample efficiently
  - In this case solution is known (so-called swap randomization)

Linear Regression

- Fit a line to a set of observation points



Intermission: basic linear algebra

A *linear combination* of n vectors $\boldsymbol{v}_i$ is $\boldsymbol{w} = \sum_{i=1}^{n} \alpha_i \boldsymbol{v}_i$

A set of vectors $V = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n\}$ is *linearly independent* if no vector $\boldsymbol{v} \in V$ can be written as a linear combination of vectors of $V \setminus \{\boldsymbol{v}\}$. Otherwise V is *linearly dependent.*

The vector *inner product* of two vectors is $\langle \boldsymbol{v}, \boldsymbol{w} \rangle = \sum_{i=1}^{n} v_i w_i$

The vector *outer product* of n- and m-dimensional (row) vectors $\boldsymbol{v}$ and $\boldsymbol{w}$ is n-by-m matrix $\boldsymbol{v}^T\boldsymbol{w} = (a_{ij})$ where $a_{ij} = v_i w_j$.

Intermission: basic linear algebra

The product of *n*-by-*k* matrix $\boldsymbol{A}$ and *k*-by-*m* matrix $\boldsymbol{B}$ is the *n*-by-*m* matrix (c_{ij}) with $c_{ij} = \sum_{l=1}^{k} a_{il} b_{lj}$.

The column rank of matrix $\boldsymbol{M}$ is the number of linearly independent columns of $\boldsymbol{M}$. The row rank is the number of linearly independent rows.

Fact. The row and column rank of *n*-by-*m* real matrix $\boldsymbol{M}$ are the same and called the *rank* of $\boldsymbol{M}$. Hence $\text{rank}(\boldsymbol{M}) \leq \min(n, m)$.

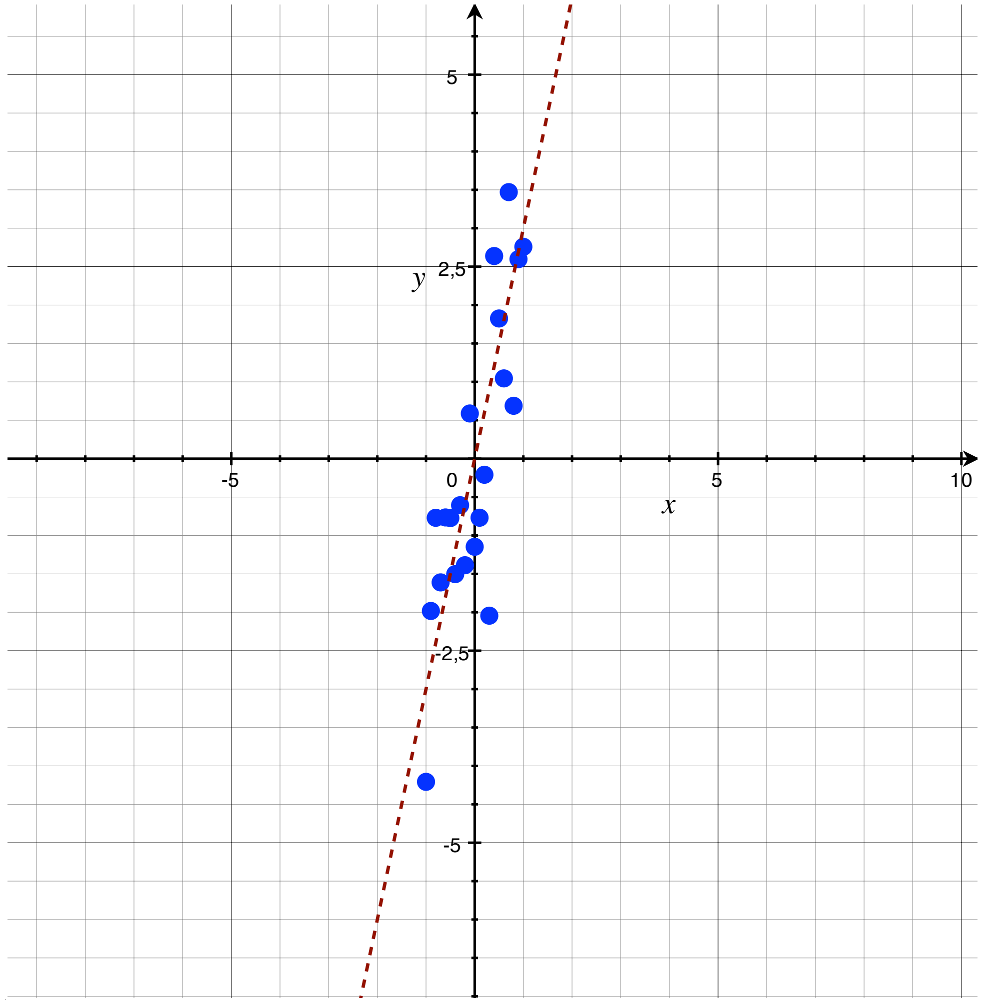
The *inverse* of an *n*-by-*n* square matrix $\boldsymbol{A}$, if exists, is the unique *n*-by-*n* matrix $\boldsymbol{B}$ for which $\boldsymbol{AB} = \boldsymbol{I}$, where $\boldsymbol{I}$ is the *n*-by-*n* identity matrix. The inverse of $\boldsymbol{A}$ is denoted by $\boldsymbol{A}^{-1}$.

An *n*-by-*n* matrix $\boldsymbol{A}$ is *invertible* (i.e. has inverse) iff $\text{rank}(\boldsymbol{A}) = n$.

Single-variable case

- A simple case with one variable
  - vector $\boldsymbol{y}$ is called the **response** variables (or *regressands*)
  - vector $\boldsymbol{x}$ is called the **predictor** variables (or *regressors*)
  - constant β is called the **parameter**
  - random variable ε is called the **error**

$$\boldsymbol{y} = \beta \boldsymbol{x} + \varepsilon$$

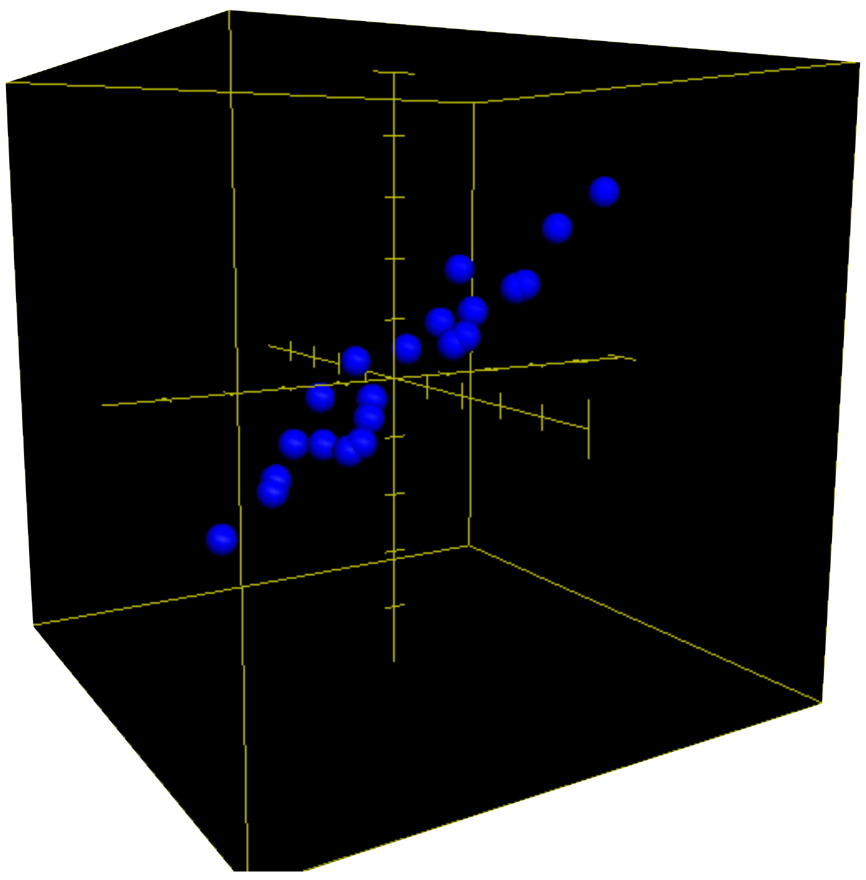


Multi-dimensional case

- The regressors are multi-dimensional
- Each regressor is a row of **design matrix *X***
- Parameters form a vector $\boldsymbol{\beta}$, and errors form a vector $\boldsymbol{\varepsilon}$
  - n respond variables and errors, k parameters, $\boldsymbol{X}$ is n-by-k

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$y_i = \sum_{j=1}^{k} x_{ij}\beta_j + \varepsilon_i$$

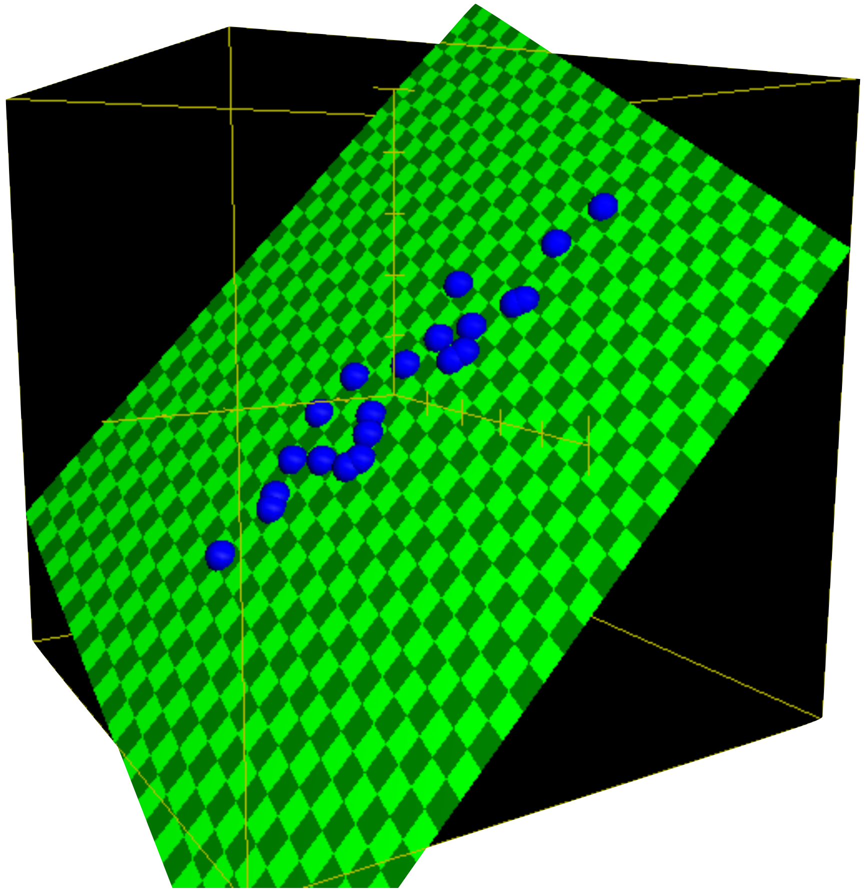


Multi-dimensional case

- The regressors are multi-dimensional
- Each regressor is a row of **design matrix** ***X***
- Parameters form a vector $\boldsymbol{\beta}$, and errors form a vector $\boldsymbol{\varepsilon}$
  - n respond variables and errors, k parameters, ***X*** is n-by-k

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$y_i = \sum_{j=1}^{k} x_{ij}\beta_j + \varepsilon_i$$



Important assumptions

- The design matrix must have full column rank
  - rank($\boldsymbol{X}$) $\geq k$
  - $n \geq k$ is a necessary but not sufficient condition
  - "There has to be enough data per parameter"
- The i.i.d. errors ε_i are $\mathbf{N}(0,\sigma^2)$ distributed
  - With this assumption ordinary least squares matches maximum likelihood estimation
- The assumptions on errors can weakened
  - Uncorrelated only conditional to regressors
  - Mean and variance only conditional to regressors

Ordinary least squares linear regression

Problem. Find $\boldsymbol{\beta}$ that minimize

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 = \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{k} x_{ij}\beta_j\right)^2$$

Solution. Estimate $\boldsymbol{\beta}$ with

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

The fitted values of $\boldsymbol{y}$ are

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

Some comments on OLS

- The matrix $\boldsymbol{X}^{\dagger} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}$ is the *Moore–Penrose pseudo-inverse* of $\boldsymbol{X}$
  - The full column rank of $\boldsymbol{X}$ is required for $(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})$ to be invertible (HW)
  - Alternatively, the full column rank guarantees unique solutions
- **Fact:** The Moore–Penrose pseudo-inverse is the least-squares solution to linear program $\boldsymbol{y} = \boldsymbol{X\beta}$
  - I.e. setting $\boldsymbol{\beta} = \boldsymbol{X}^{\dagger}\boldsymbol{y}$ minimizes the squared error, as supposed
  - If $\boldsymbol{X}$ is invertible, $\boldsymbol{X}^{\dagger} = \boldsymbol{X}^{-1}$, as supposed (HW)

The intercept

- So far we have considered through-the-origin regression
  - The fitted line crosses the origin
- Usually we add an *intercept* β_0

$$y_i = \sum_{j=1}^{k} x_{ij}\beta_j + \beta_0 + \varepsilon_i$$

- To simplify notation, this is done by adding an extra column full of 1s to $\boldsymbol{X}$

$$y_i = \sum_{j=0}^{k} x_{ij}\beta_j + \varepsilon_i \quad \text{where } \boldsymbol{x}_{i0} = 1 \text{ for all } i$$

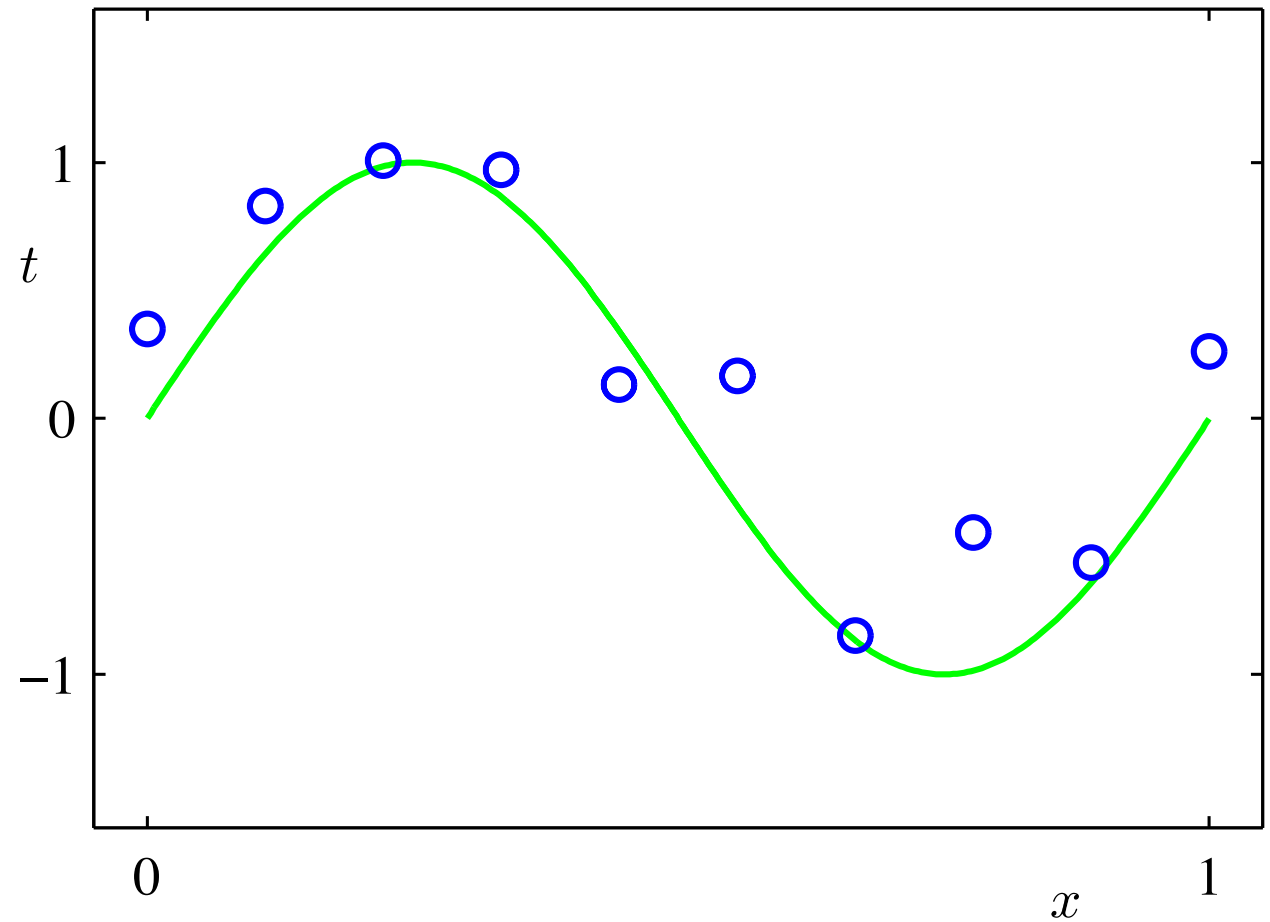
Non-linear regressors

- The all-linear model is very restrictive
- The regressors ***x*** can be non-linear
  - But the response variables ***y*** must be linear combination of regressors
  - An example: polynomial of degree *M*

$$y_i = \sum_{d=0}^{M} x_i^d \beta_d + \varepsilon_i$$

Example: fitting $\sin(2\pi x)$

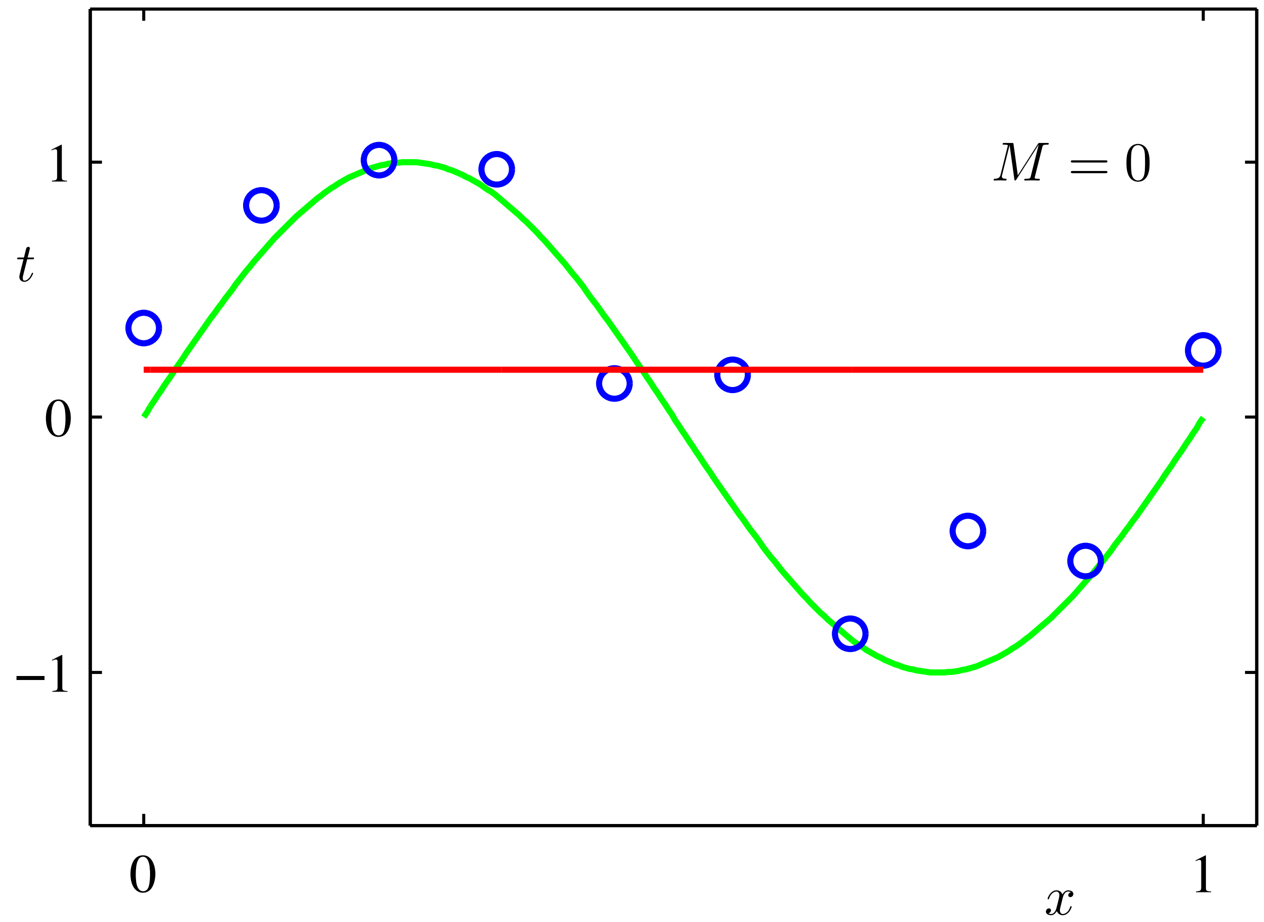
Example and images by Bishop (Chapter 1)



N=10 data points and $\sin(2\pi x)$

Example: fitting sin(2πx)

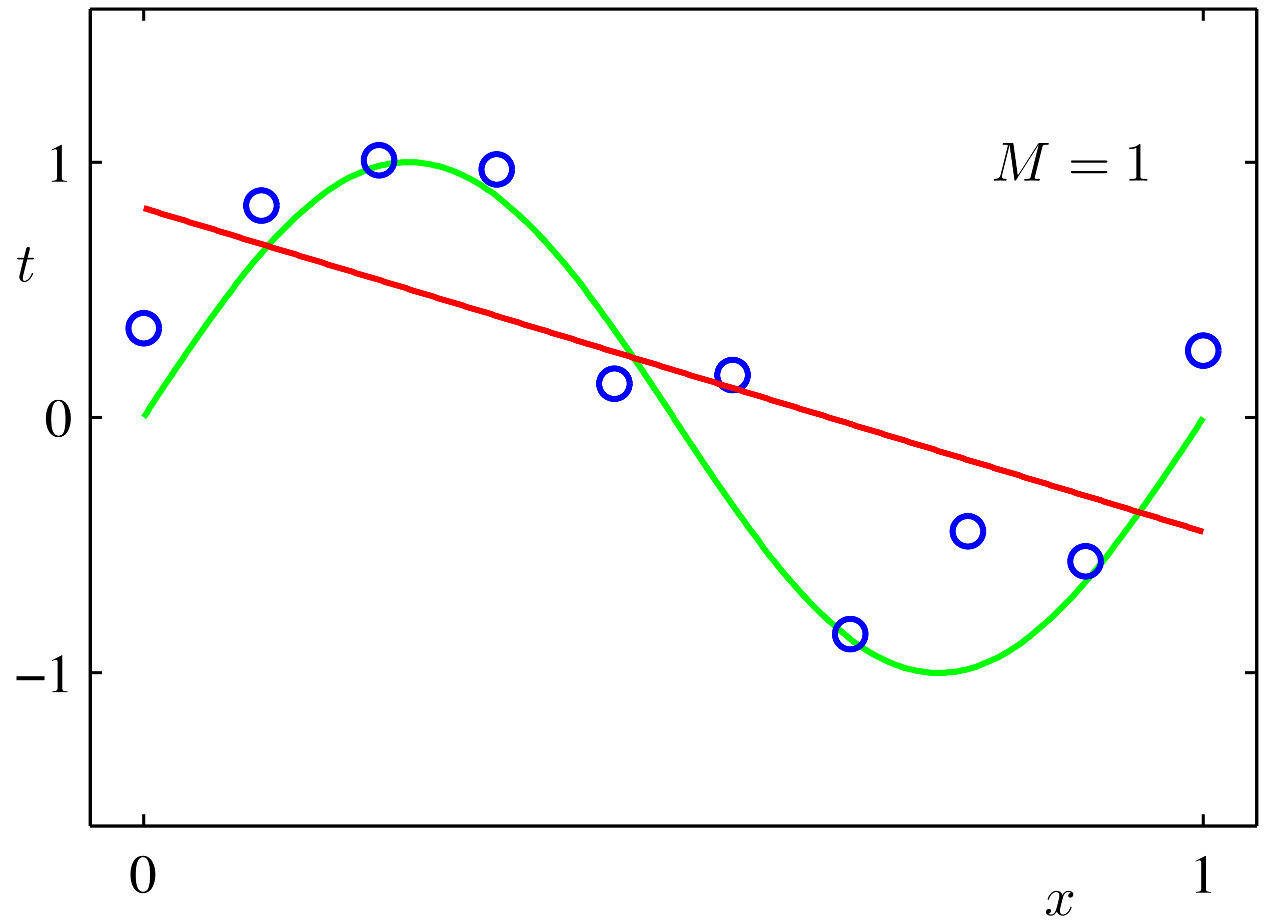
Example and images by Bishop (Chapter 1)



$$y = \beta_0$$

Example: fitting sin(2πx)

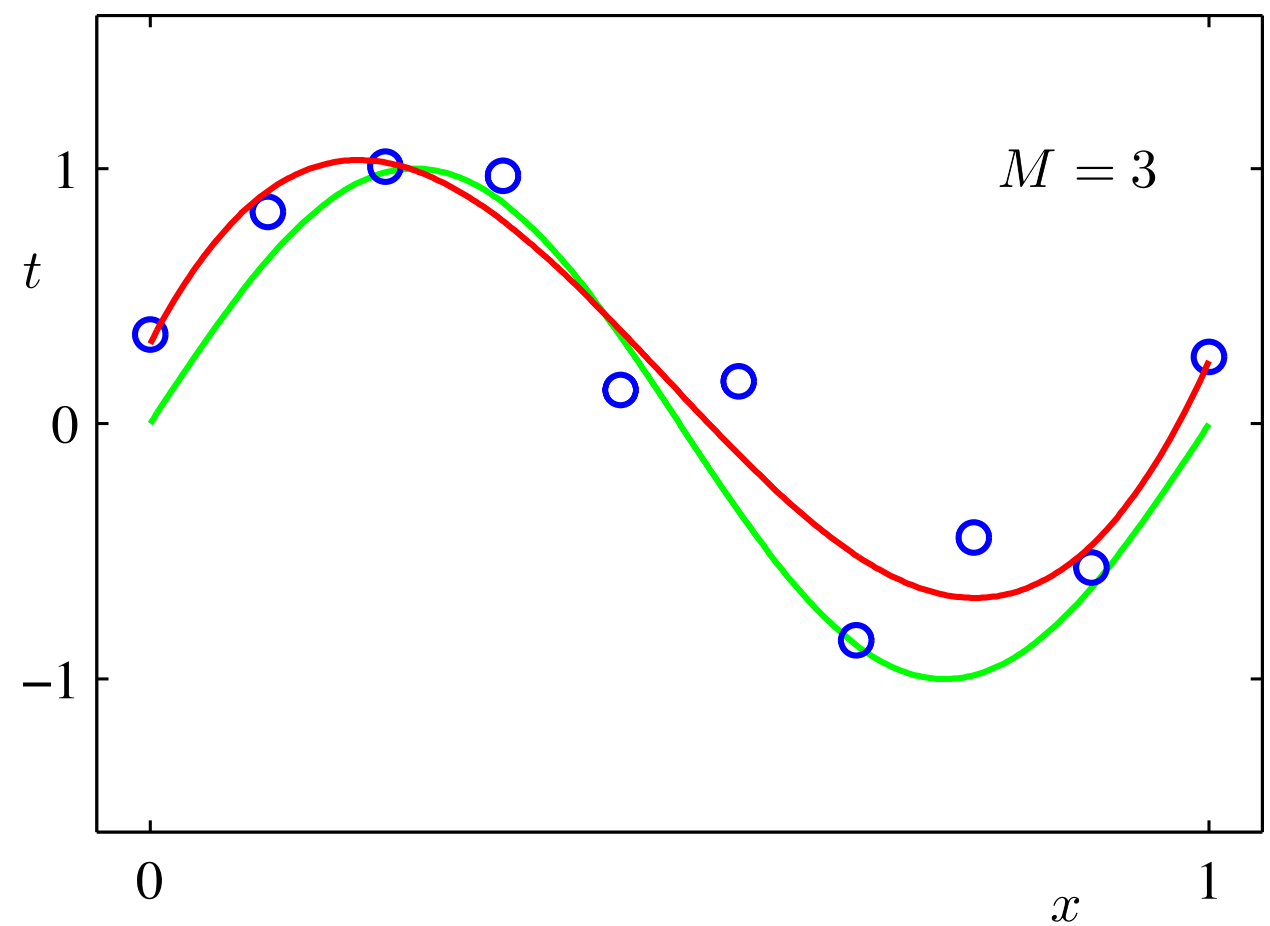
Example and images by Bishop (Chapter 1)



$$y = \beta_1 x + \beta_0$$

Example: fitting sin(2πx)

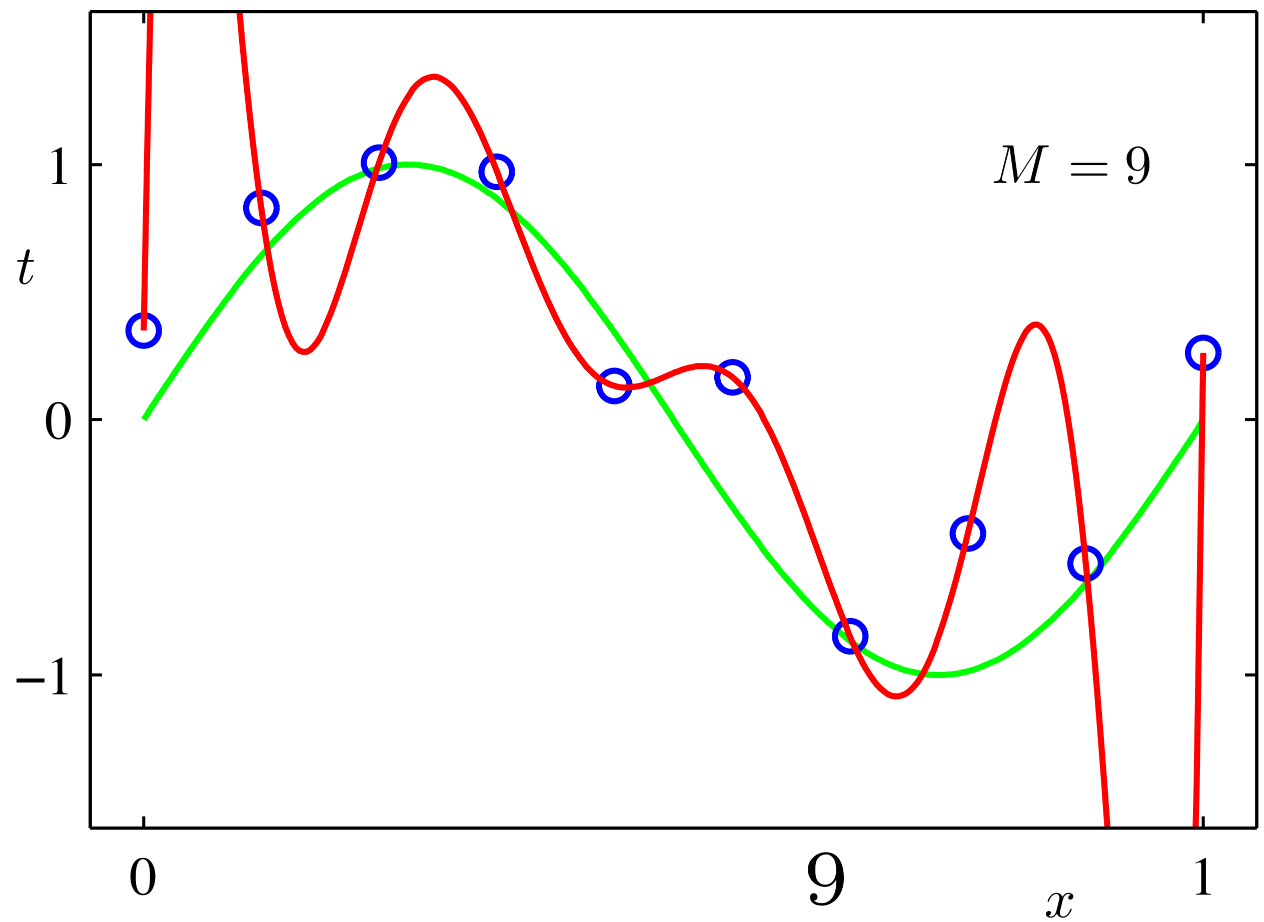
Example and images by Bishop (Chapter 1)



$$y = \beta_3 x^3 + \beta_2 x^2 + \beta_1 x + \beta_0$$

Example: fitting sin(2πx)

Example and images by Bishop (Chapter 1)



$$y = \sum_{d=0}^{9} \beta_d x^d$$

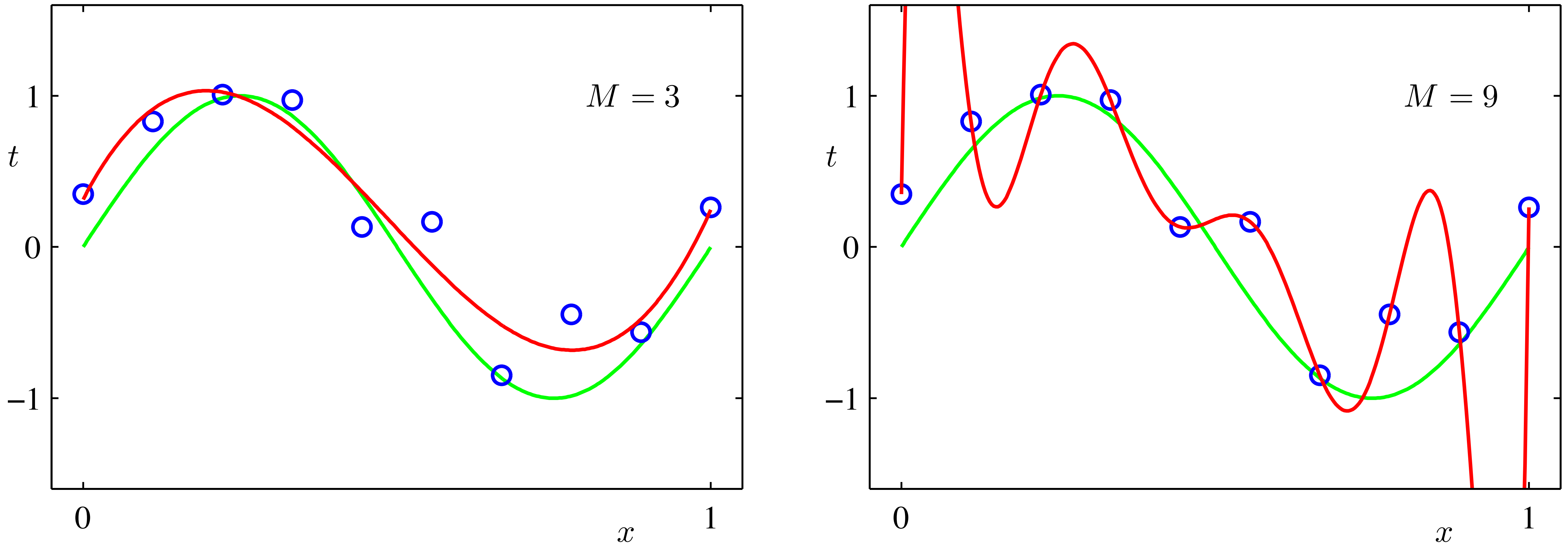
Non-linear regressors (cont'd)

- In general we have $k+1$ **basis functions** $\varphi_j(x)$
  - φ_0 is constant ($\varphi_0(x) = 1$) for the intercept
  - In the previous example, $\varphi_j(x) = x^j$
  - Other basis functions are possible
- The design matrix $\boldsymbol{X}$ is replaced with $\boldsymbol{\Phi}$:

$$\boldsymbol{\Phi} = \begin{pmatrix} \varphi_0(x_1) & \varphi_1(x_1) & \cdots & \varphi_k(x_1) \\ \varphi_0(x_2) & \varphi_1(x_2) & \cdots & \varphi_k(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_0(x_n) & \varphi_1(x_n) & \cdots & \varphi_k(x_n) \end{pmatrix}$$

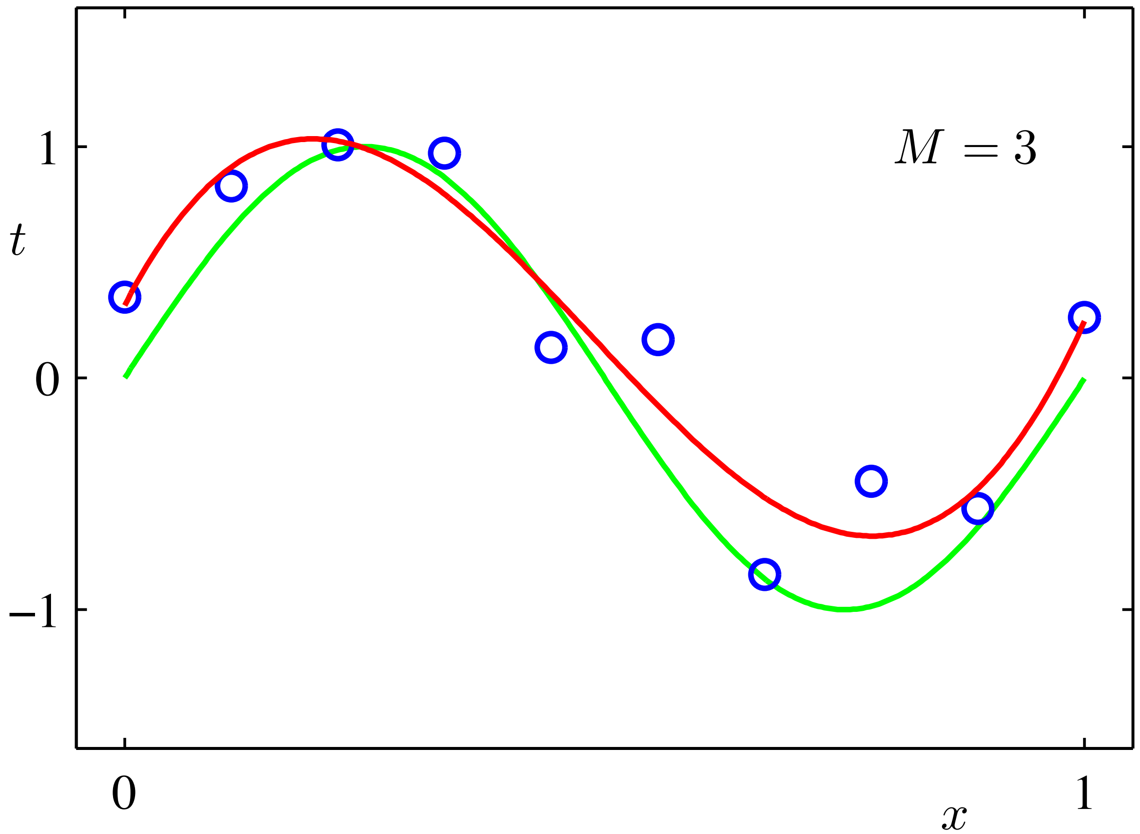
Regularization

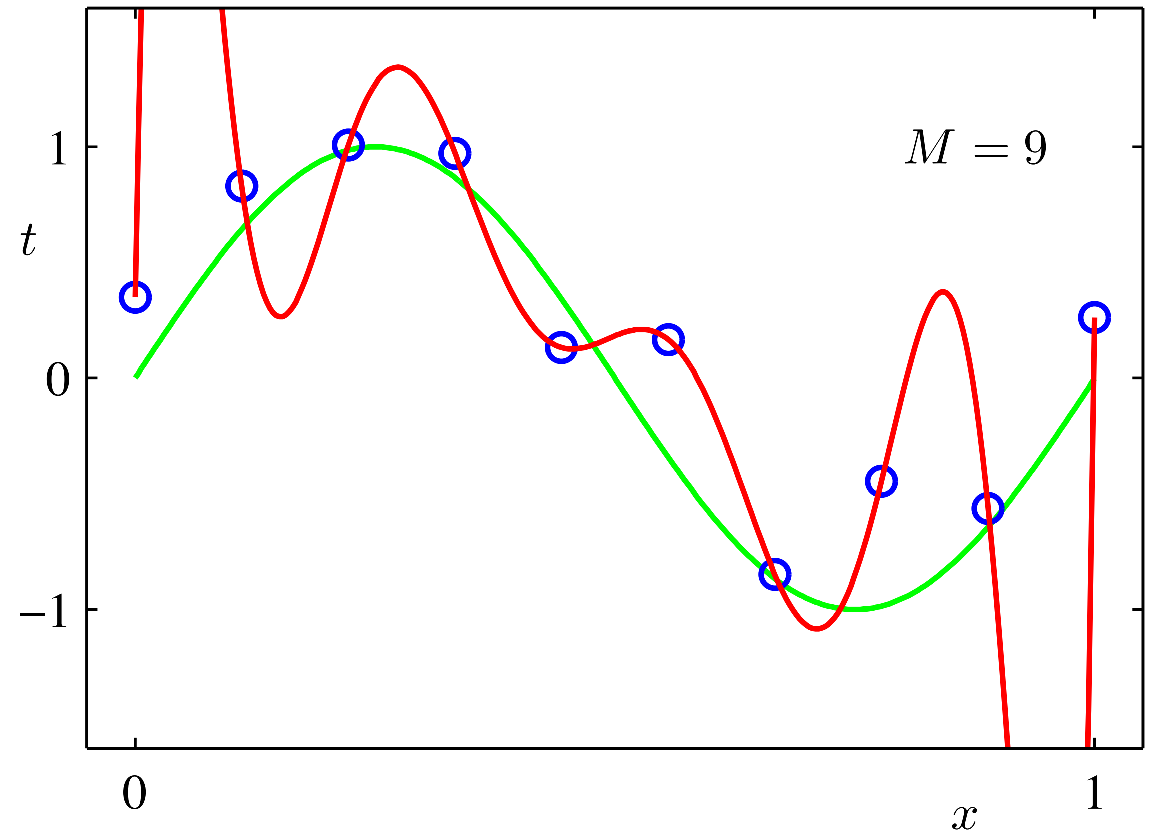
- Which of the two models fit the data better?



Regularization

- Which of the two models fit the data better?

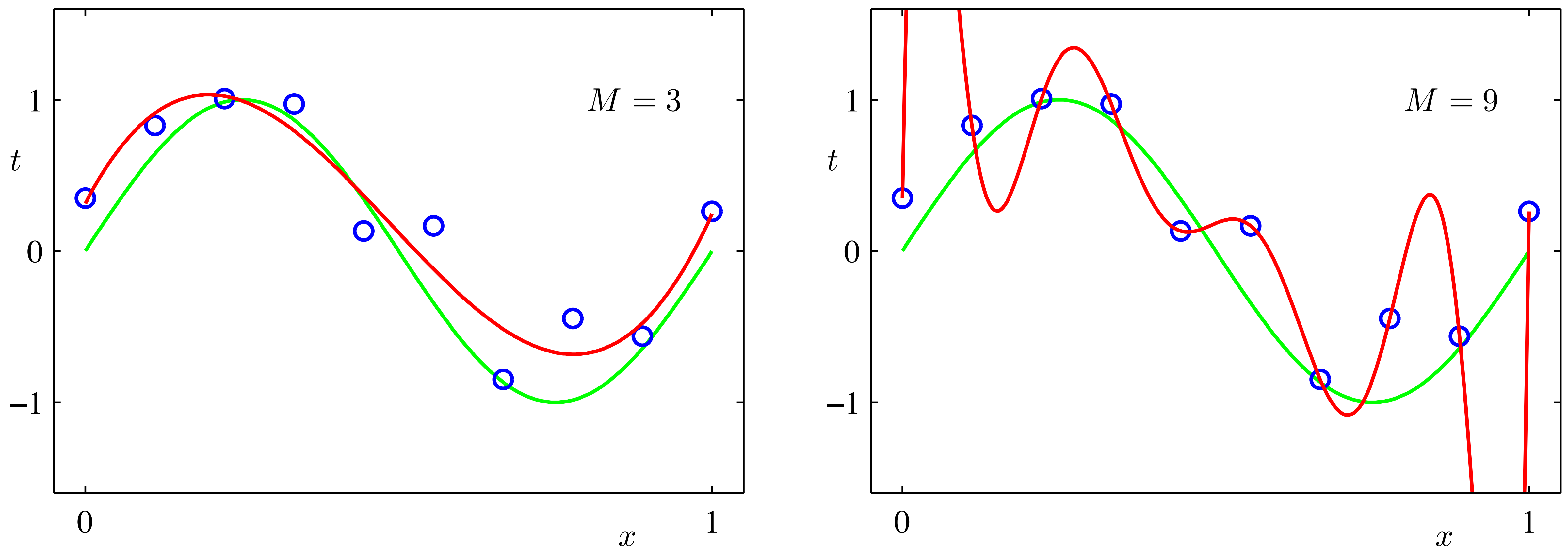




This looks better

Regularization

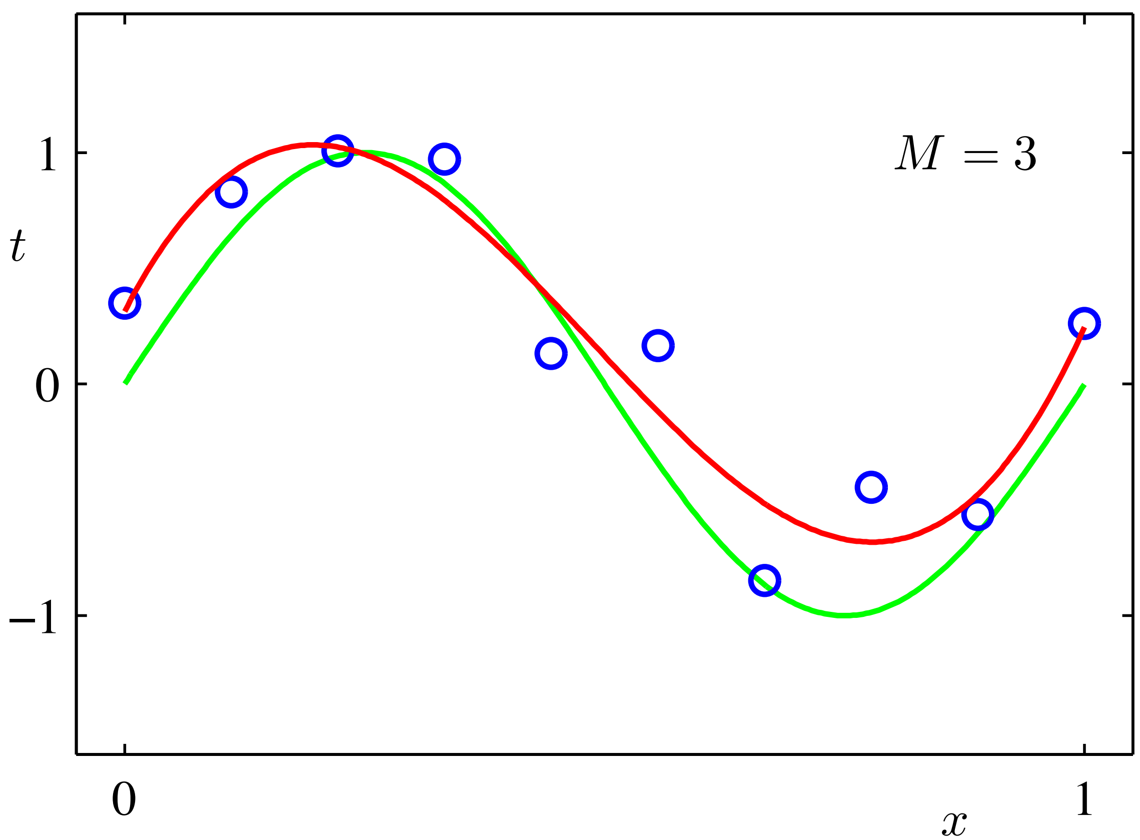
- Which of the two models fit the data better?



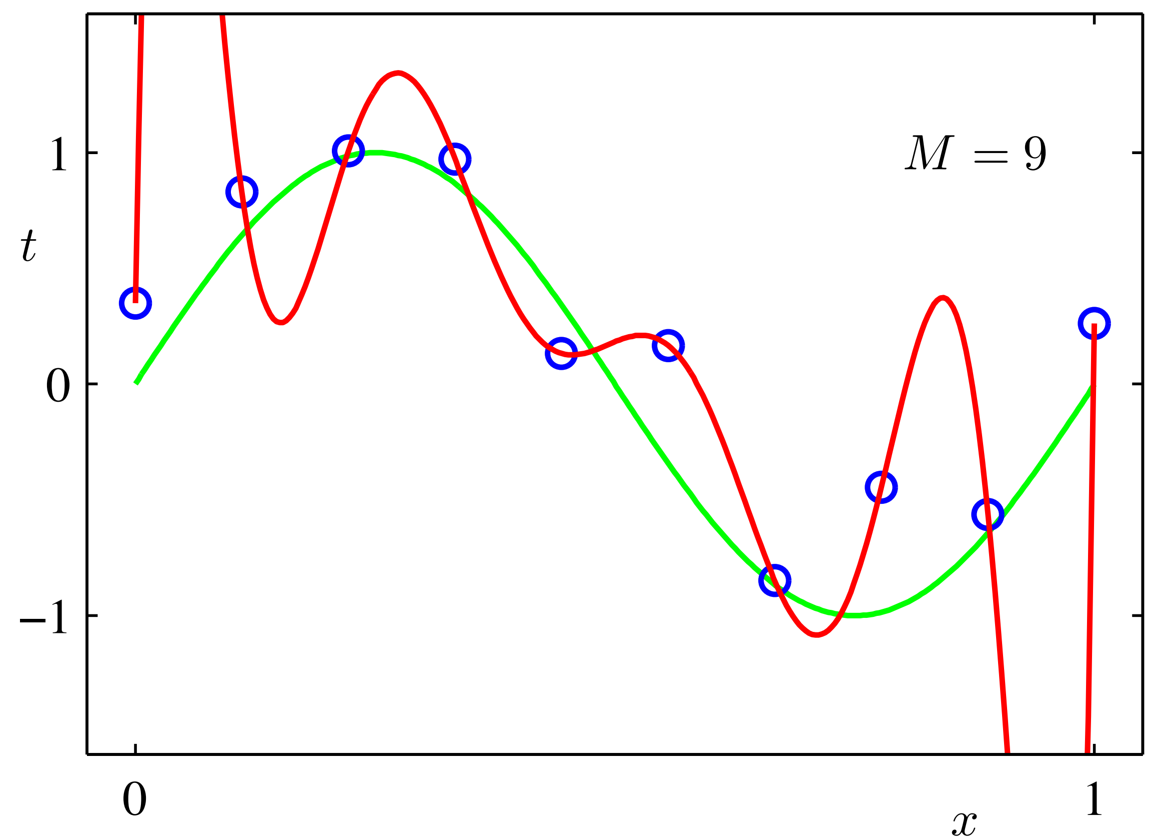
This looks better **This has no error**

Regularization

- Which of the two models fit the data better?



This looks better



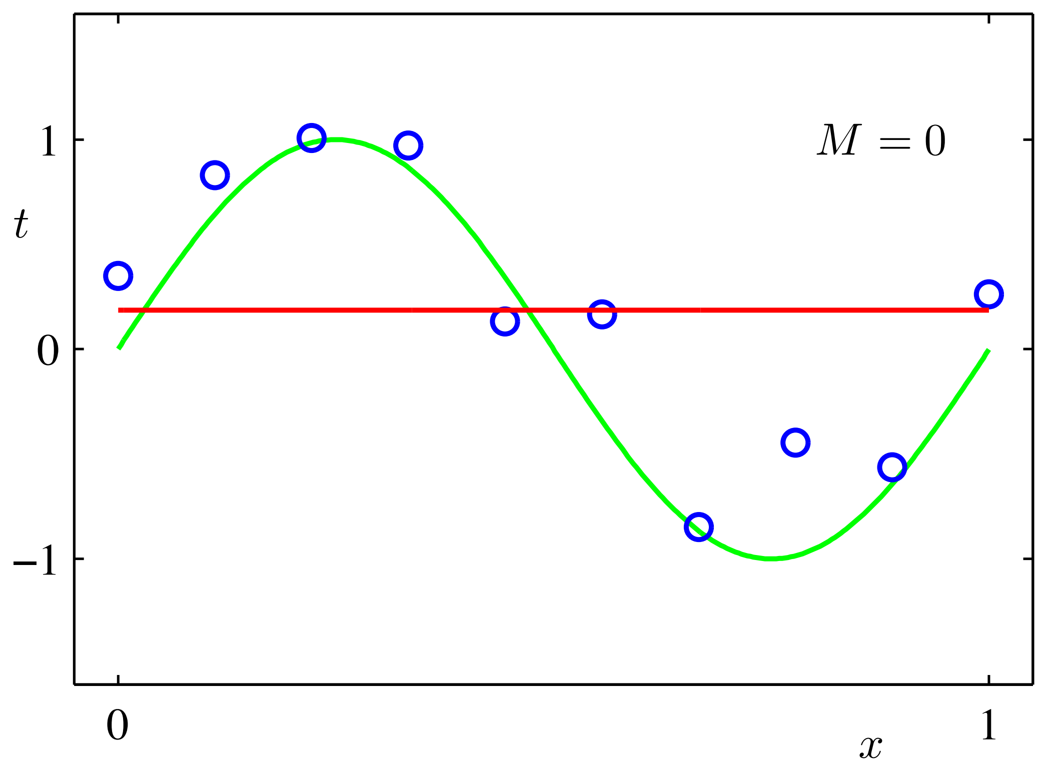
This has no error

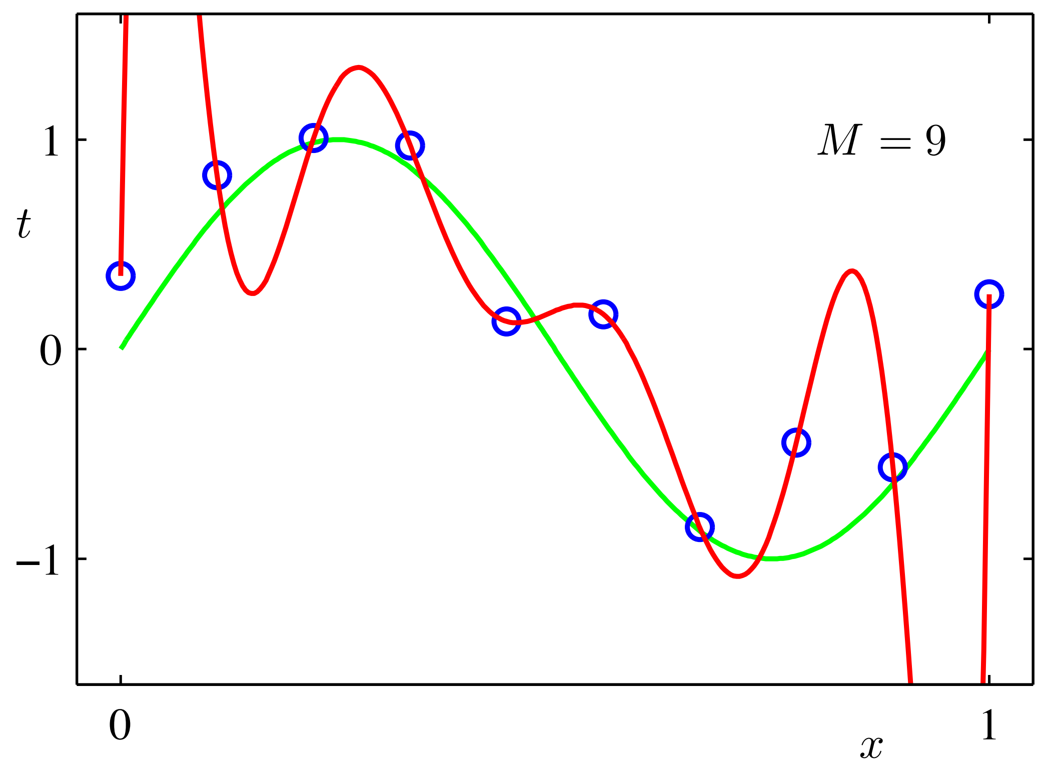
- Can we formalize why we think left is better?

Two roles of regression

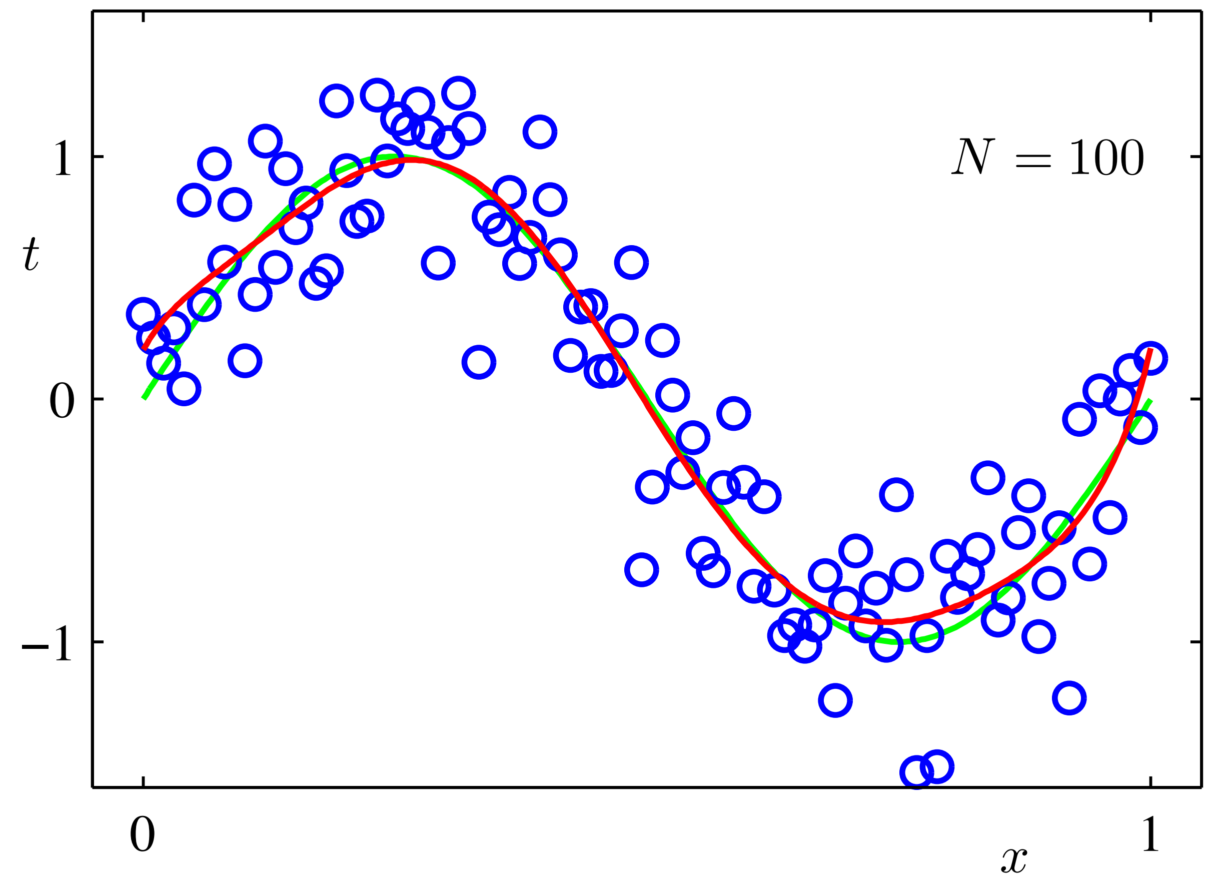
- We can approach regression either as
  - descriptive method explaining the data
  - predictive method allowing us to make predictions of future data
- For predicting, we need to combat against **under-fitting** and **over-fitting**
  - Under-fit model gives poor predictions because it doesn’t model the process well
  - Over-fit model gives poor predictions because it models also the error

Example of under- and over-fitting





More data allows complex models



Polynomial of degree 9 fitted to $N = 100$ data points

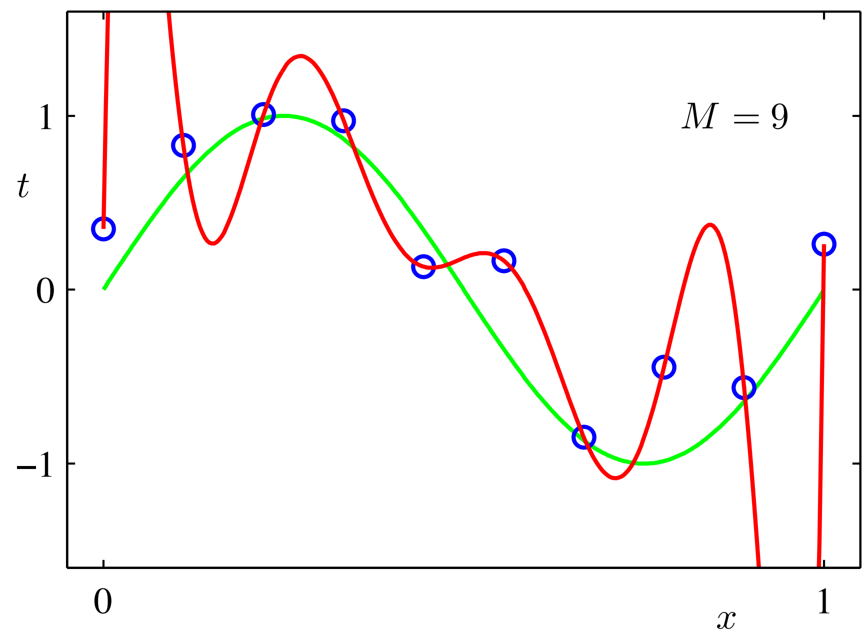
Regularizers

- Selecting the model based on data size does not sound good
- A **regularizer** penalizes on too complex models

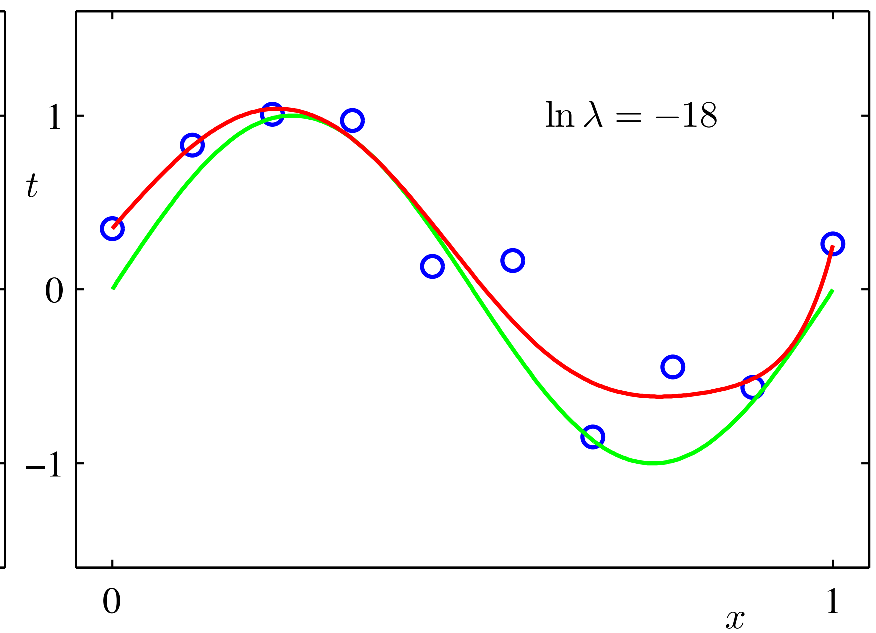
$$\|\mathbf{y} - \mathbf{\Phi\beta}\|^2 + \lambda \left\| (\beta_j)_{j=1}^{k} \right\|^2$$
$$= \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{k} \varphi_j(x_i)\beta_j \right)^2 + \lambda \sum_{j=1}^{k} \beta_j^2$$

- Variable λ is called *regularization parameter*
- Intercept is not included in regularization

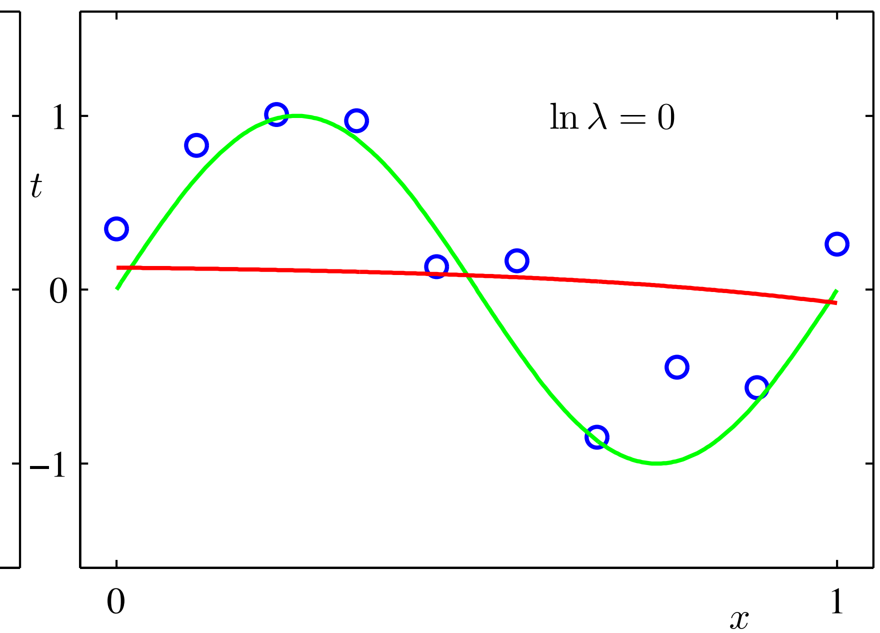
An example



$\lambda = 0$



$\lambda = e^{-18}$



$\lambda = 1$

More on regularizers

- In statistics, regularizers are called *shrinkage methods*
- Regression with quadratic regularizer is also known as *ridge regression*
- Quadratic (L^2) regularizer keeps the loss function quadratic
- The sum-of-absolute-values regularizer $\lambda\sum|\beta_i|$ is known as *lasso* or L^1 regularizer
  - With sufficiently large λ this forces some β_is to 0

Model selection

- **How do we select λ?**
- The goal is prediction, so test which λ predicts best
  - Divide data to training data and test data
    - E.g. y_i and x_i for $i = 1..n-1$ are training data and y_n and x_n are test data
  - Learn βs with training data
  - Measure the error with training and test data
  - Repeat with other values of λ and select the one with least over-all error

S-fold cross validation

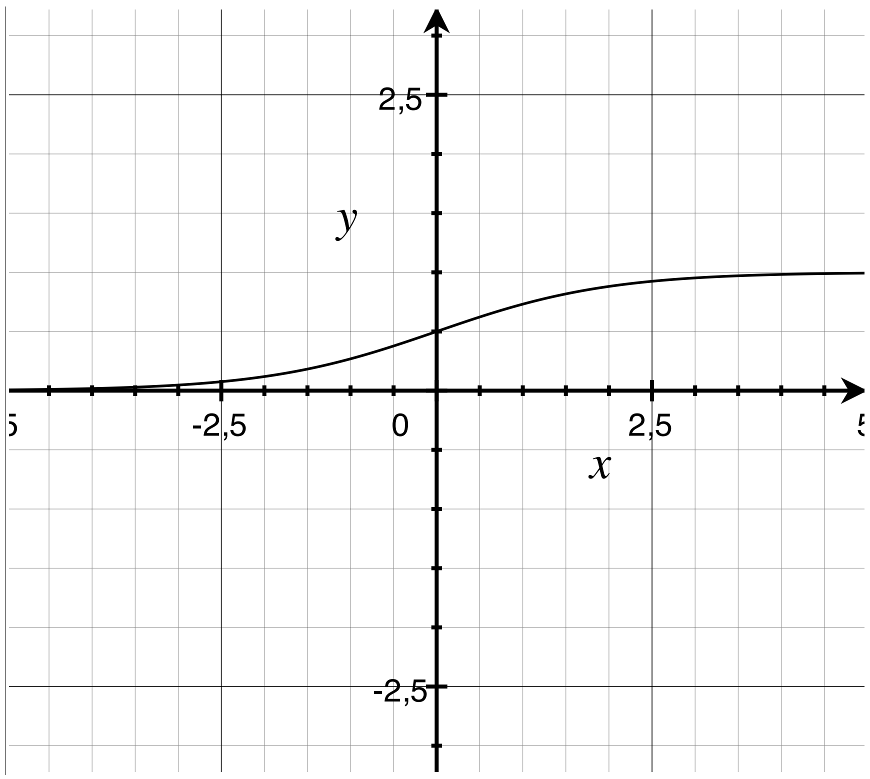
- Divide data to *S* subsets
- Use *S*-1 of these subsets as training data and the last subset as test data
- Repeat *S* times with different subset being the test data
- Average errors over different runs and select the best

Logistic Regression

- Actually classification
- Response variables $y_i \in \{0,1\}$
- Name comes from the *logistic function*

$$f(x) = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}$$

  - The logistic function maps values from $(-\infty,\infty)$ to $(0,1)$



Logistic regression

Given k-dimensional regressors X_i, we estimate y_i as

$$\hat{y}_i = \frac{e^{\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}}}$$

or, equivalently

$$\operatorname{logit}(\hat{y}_i) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$$

where

$$\operatorname{logit}(x) = \ln\left(\frac{x}{1-x}\right)$$

Notes on logistic regression

- No analytic solution to **β**
- Finding **β** needs to use numerical methods
  - Fast method called Iterative Re-Weighted Least Squares is often used
- Logistic function is also known as *sigmoid function*
- Similar to linear regression, we can apply fixed non-linear basis functions **ϕ** to **X**
- Other classification methods will be discussed later in the course

Summary of Chapter 2.3

- Hypothesis testing can be used to test if sample has certain properties
  - same mean, same parameters, goodness-of-fit, etc.
- Linear regression fits linear function of regressors to response variables
- We can combat over-fitting using regularizers
  - Regularizer parameter needs to be selected
- Logistic regression takes the logistic function of linear combination of regressors to classify response variables