Zürcher Hochschule
für Angewandte Wissenschaften

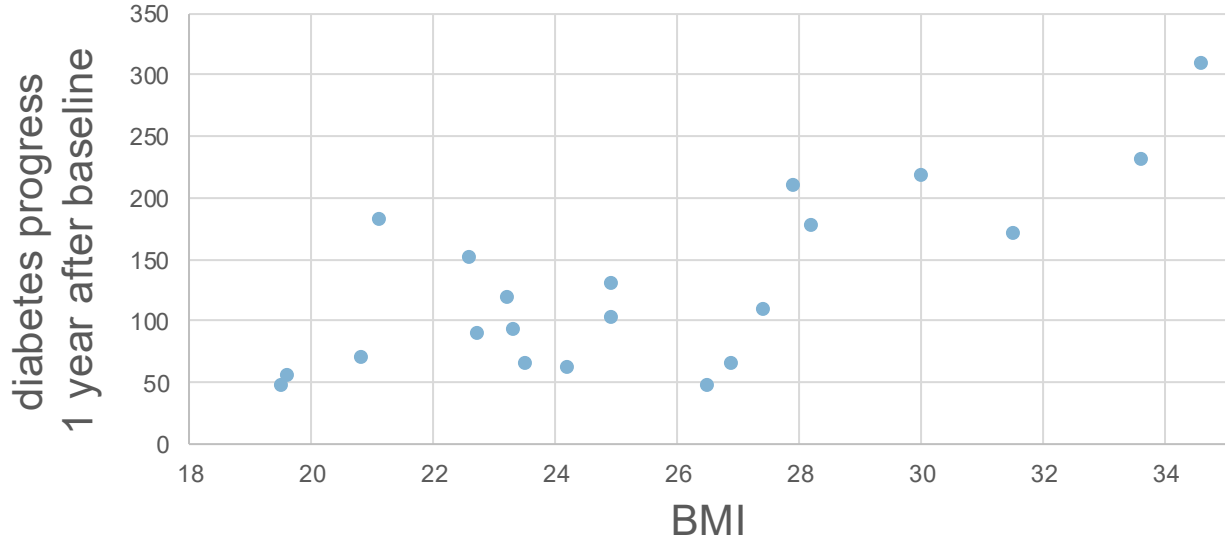
zh aw School of Engineering

Supervised Machine Learning
Regression

Exemplary Regression Problem: Course of diabetes progress depending on patient's BMI

Zürcher Hochschule für Angewandte Wissenschaften
School of Engineering

bmi x	progress y
27.8	201
22.8	40
35	140
34.6	264
⋮	⋮



Univariate: 1 indepentent variable (feature)

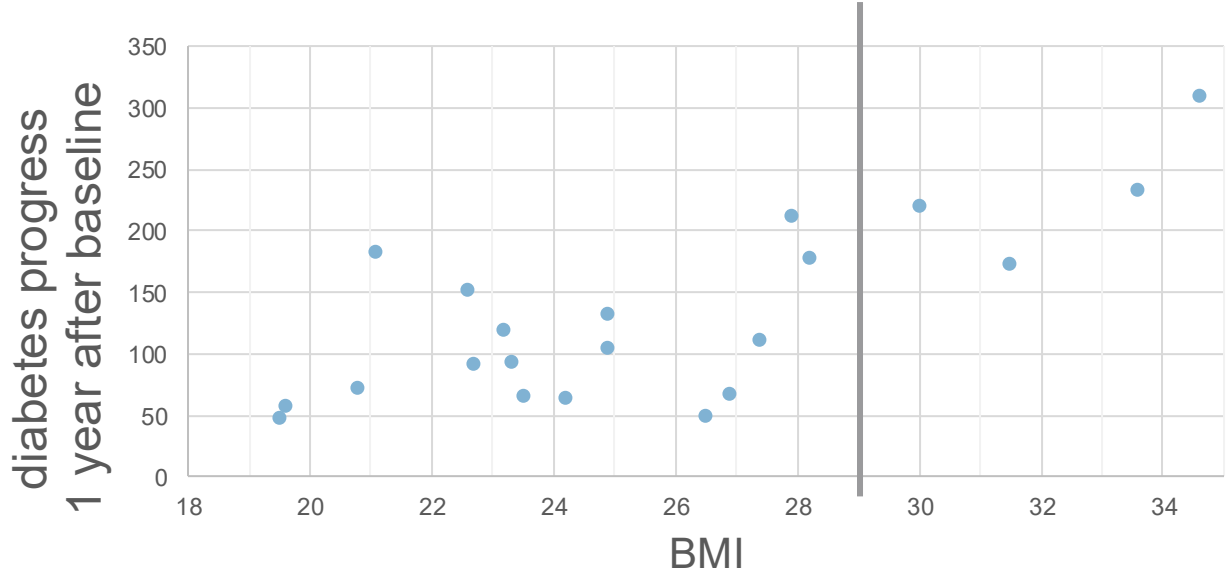
Data: https://www4.stat.ncsu.edu/~boos/var.select/diabetes.html

Illustration of the Regression Problem

Zürcher Hochschule
für Angewandte Wissenschaften

zh aw School of Engineering

What is the expected course of disease with a BMI of 29?



Instance-Based Learning Applied to a Regression Problem
KNN: K-Nearest Neighbour Algorithm

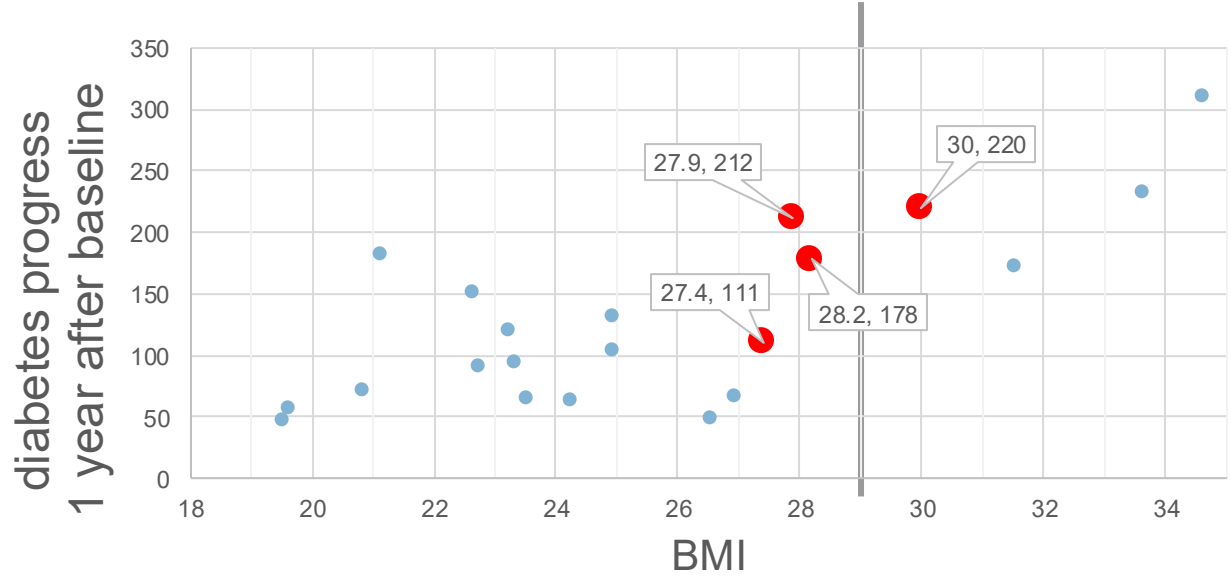
Zürcher Hochschule für Angewandte Wissenschaften
School of Engineering

What is the expected course of disease with a BMI of 29?

k=4

Dsitance metric: difference in BMI

BMI=29 -> average of the 4 closest points = (111+212+178+220)/4 = 180.3

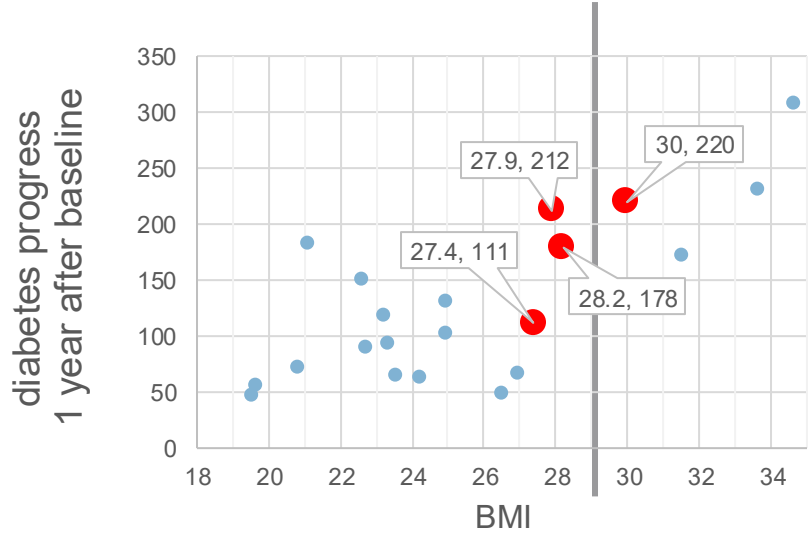


The Nearest Neighbours-Algorithmus Applied to a Regression Problem

Zürcher Hochschule für Angewandte Wissenschaften
School of Engineering

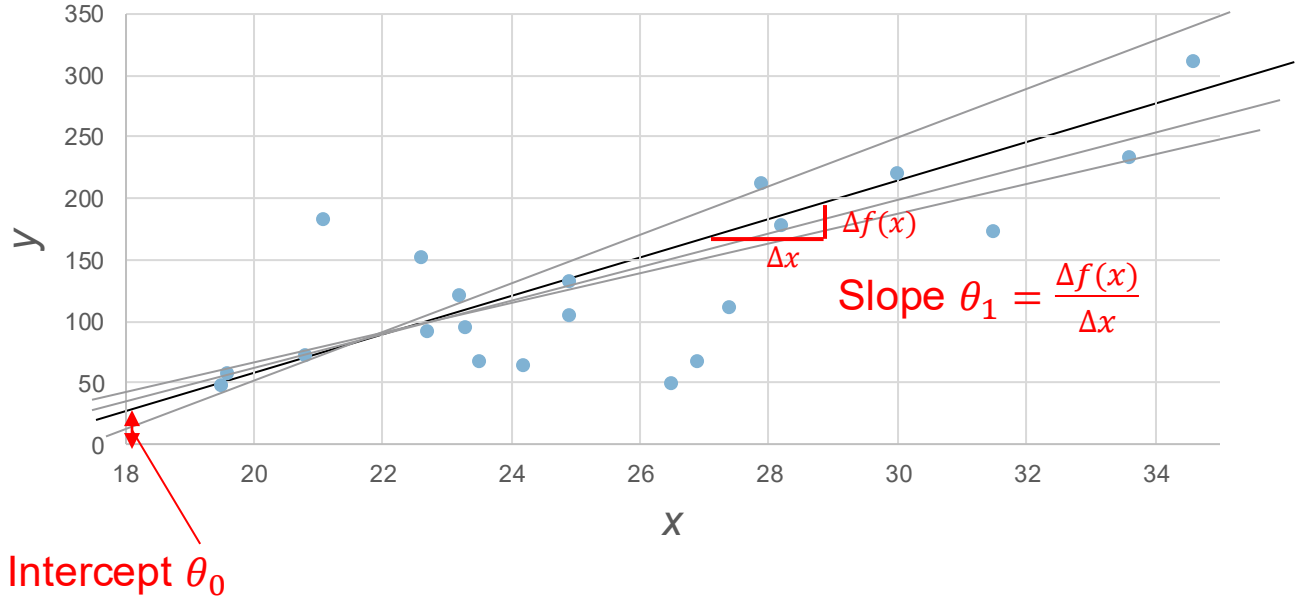
Disadvantages:

- Inference on a new data point **always requires to consider explicitly all training data** (select k closest points and compute the output)
- Is **sensitive to outliers** and noise
- Training samples further away than the k neighbours are ignored for inference → **loss of information** contained in the training data



Univariate linear regression

Mathematical expression which consistently models the relationship between the variables.



Hypothesis: Linear function with parameters θ_0, θ_1

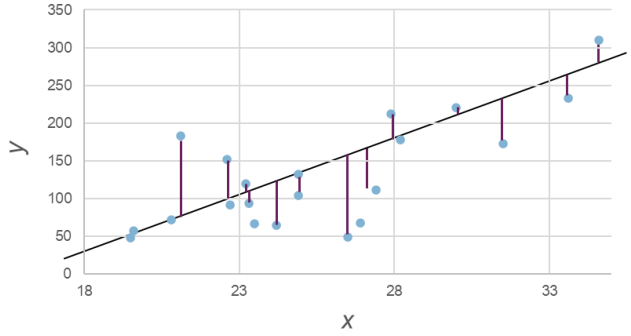
$$h(x; \theta_0, \theta_1) = \theta_0 + \theta_1 x$$

Goal of the Training/Learning: Parameter values, which describe the relationship best.

«Training» of the univariate linear regression model

Zürcher Hochschule
für Angewandte Wissenschaften

zh aw School of Engineering



Training data: M samples $\{(x_m, y_m)\}$

The outputs y are explained by the hypothesis h plus

the stochastic residuals e: $y_m = h(x_m; \theta_0, \theta_1) + e_m$

The model predicts for the sample x_m: $\hat{y}_m = \theta_0 + \theta_1 x_m$

Loss function: Sum of squared residuals $\mathcal{L}_{RSS}(\theta_0, \theta_1) = \sum_{m=1}^{M} e_m^2 = \sum_{m=1}^{M} (y_m - \hat{y}_m)^2$

Learning: Minimise the cost function

$$J(\theta_0, \theta_1) = \frac{1}{2M} \sum_{m=1}^{M} (y_m - \hat{y}_m)^2$$

Parameters $\hat{\theta}_0, \hat{\theta}_1$ minimise the cost function

and can then be used for inference on new data samples $\hat{y}_i = h(x_i; \hat{\theta}_0, \hat{\theta}_1)$

Explicit solution of univariate linear regression

Zürcher Hochschule
für Angewandte Wissenschaften
School of Engineering

Setting $\frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1) = 0$ and $\frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1) = 0$

and solving these for θ_0 and θ_1 yields (without derivation):

$$\hat{\theta}_0 = \mu_y - \theta_1 \mu_x \qquad \text{and} \qquad \hat{\theta}_1 = \frac{\sum_{m=1}^{M} (x_m - \mu_x)(y_m - \mu_y)}{\sum_{m=1}^{M} (x_m - \mu_x)^2}$$

where μ_x is the mean of all $\{x_m\}$ and μ_y is the mean of all $\{y_m\}$ of the training data

Normal Equations

Write the hypothesis in matrix form: $\boldsymbol{y} = \boldsymbol{X\theta} + \boldsymbol{e}$

For the (univariate) diabetes problem:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & \vdots \\ 1 & x_M \end{pmatrix} \begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_M \end{pmatrix}$$

The vector of residuals is $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X\theta}$

Loss: The sum of squared residuals is

$$\begin{aligned} \boldsymbol{e}^T\boldsymbol{e} &= (\boldsymbol{y} - \boldsymbol{X\theta})^T(\boldsymbol{y} - \boldsymbol{X\theta}) \\ &= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{\theta}^T\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X\theta} + \boldsymbol{\theta}^T\boldsymbol{X}^T\boldsymbol{X\theta} \\ &= \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{\theta}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{\theta}^T\boldsymbol{X}^T\boldsymbol{X\theta} \end{aligned}$$

Normal Equations

The gradient of the squared residuals (without derivation):

$$\frac{\partial \boldsymbol{e}^T \boldsymbol{e}}{\partial \boldsymbol{\theta}} = -2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\theta}$$

Equating the gradient with zero produces the so-called normal equations:

$$(\boldsymbol{X}^T\boldsymbol{X})\hat{\boldsymbol{\theta}} = \boldsymbol{X}^T\boldsymbol{y}$$

which leads to a closed form-expression for the optimal parameters:

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

Multivariate linear regression with the normal equations

Zürcher Hochschule für Angewandte Wissenschaften

School of Engineering

$$h(\boldsymbol{x}_m; \theta_0, \theta_1, \ldots, \theta_N) = \theta_0 x_{m0} + \theta_1 x_{m1} + \theta_2 x_{m2} + \ldots + \theta_N x_{mN} = \boldsymbol{\theta}^T \boldsymbol{X}_{m,:}$$

with $x_{m0} = 1$ for all $m = 1, \ldots, M$

Diabetes example

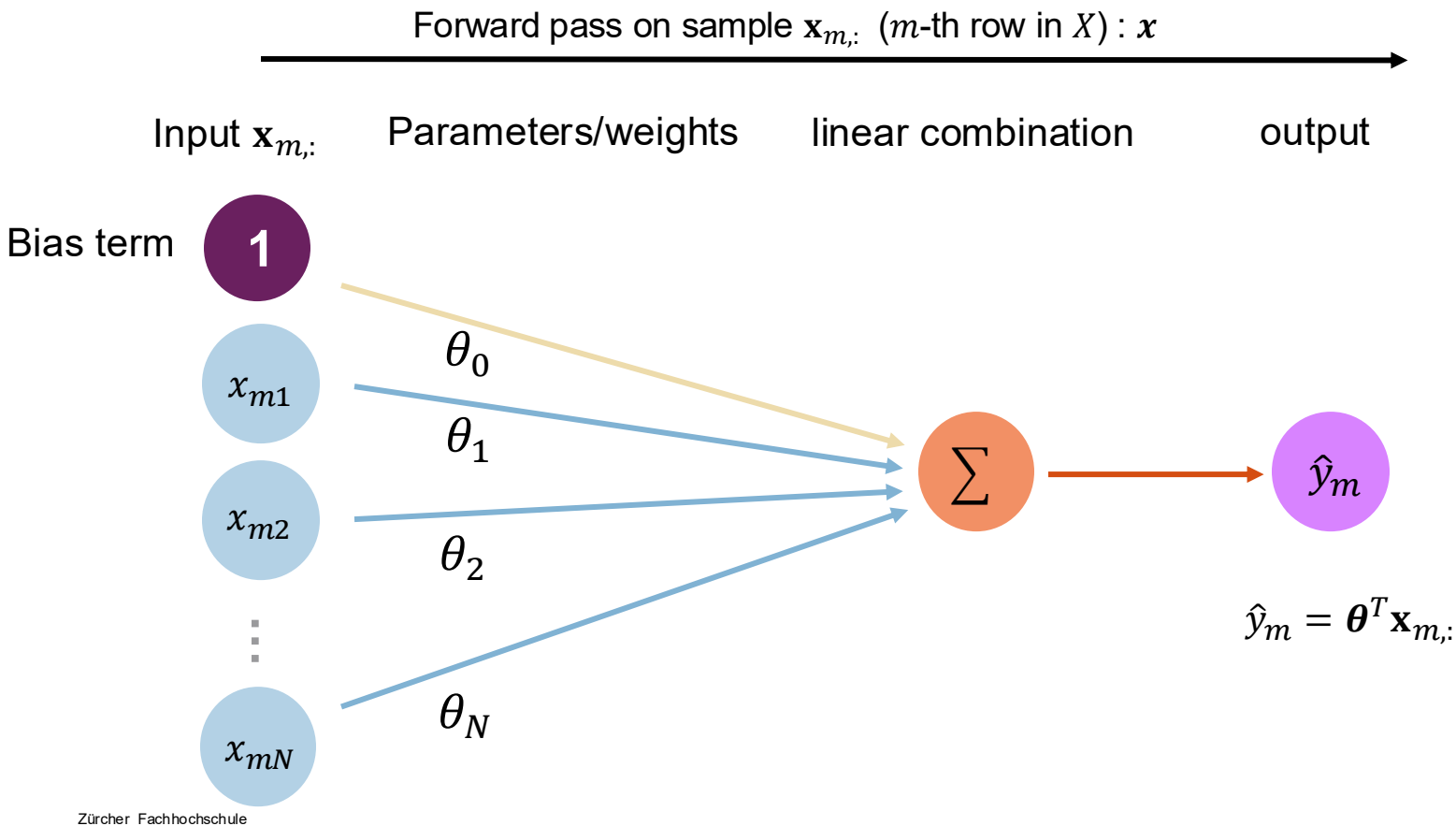
with $M = 4$, $N = 3$:

	bmi	**bp**	**glu**	**progress**
x_0	x_1	x_2	x_3	y
1	27.8	73	73	201
1	22.8	101	97	40
1	35	79.33	96	140
1	34.6	115	109	264

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1N} \\ 1 & x_{21} & x_{22} & \cdots & x_{2N} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{M1} & x_{M2} & \cdots & x_{MN} \end{pmatrix} \quad \boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{pmatrix} \quad \boldsymbol{\theta} = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_N \end{pmatrix}$$

Dimensions: $M \times (N+1)$ $\quad M \times 1$ $\quad N+1$

Linear regression for one sample, N features



Zürcher Hochschule
für Angewandte Wissenschaften

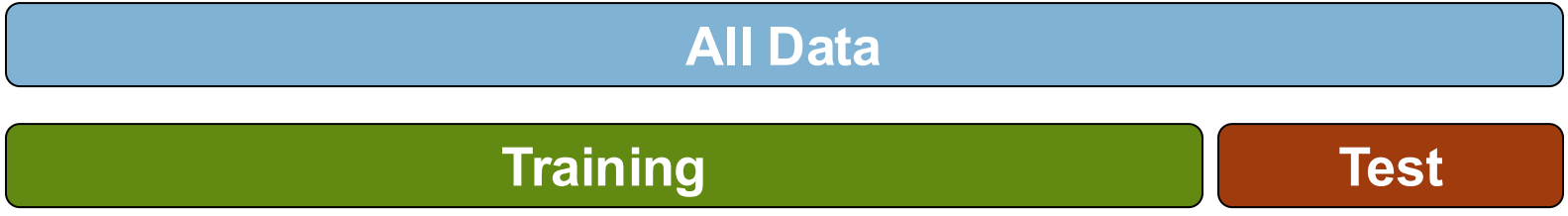
zh aw School of Engineering

Evaluating Regression Models

Can we estimate quantitatively how well the model generalises beyond the training data?

Zürcher Hochschule
für Angewandte Wissenschaften
zh aw School of Engineering

General for supervised machine learning: Split the available data into a training and independent test set



Estimate the generalisation error based on the independent test set

Evaluation Metrics for Regression

Zürcher Hochschule
für Angewandte Wissenschaften
zh aw School of Engineering

Mean Absolute Error: $MAE = \frac{\sum_{i=1}^{I} |y_i - \hat{y}_i|}{I}$

Mean Squared Error (MSE): $MSE = \frac{\sum_{i=1}^{I} (y_i - \hat{y}_i)^2}{I}$

Root Mean Squared Deviation: $RMSD = \sqrt{MSE} = \sqrt{\frac{\sum_{i=1}^{I} (y_i - \hat{y}_i)^2}{I}}$

with I number of samples in the independent test set

Evaluation Metrics for Linear Regression

Zürcher Hochschule für Angewandte Wissenschaften

zh aw School of Engineering

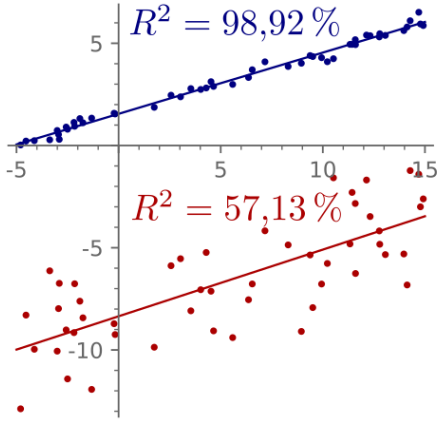
Coefficient of Determination $R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_{i=1}^{I}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{M}(y_i - \mu_y)^2}$

Sum of squares of residuals: unexplained variance (variance of the model's errors)

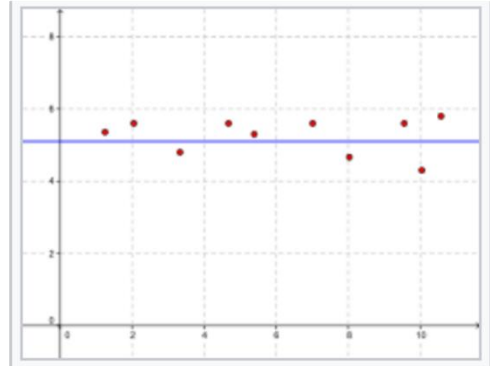
Total sum of squares (proportional to the variance of the data)

Measures fraction of the variance of the data, which can be explained by the model.

In the best case: $y_i = \hat{y}_i \rightarrow SS_{res} = 0 \rightarrow \mathrm{R}^2 = 1$



A baseline model, which always predicts $\mu_y \rightarrow \mathrm{R}^2 = 0$



Models which perform worse than the baseline $\rightarrow \mathrm{R}^2 < 0$

https://de.wikipedia.org/wiki/Bestimmtheitsma%C3%9F

Zürcher Hochschule
für Angewandte Wissenschaften

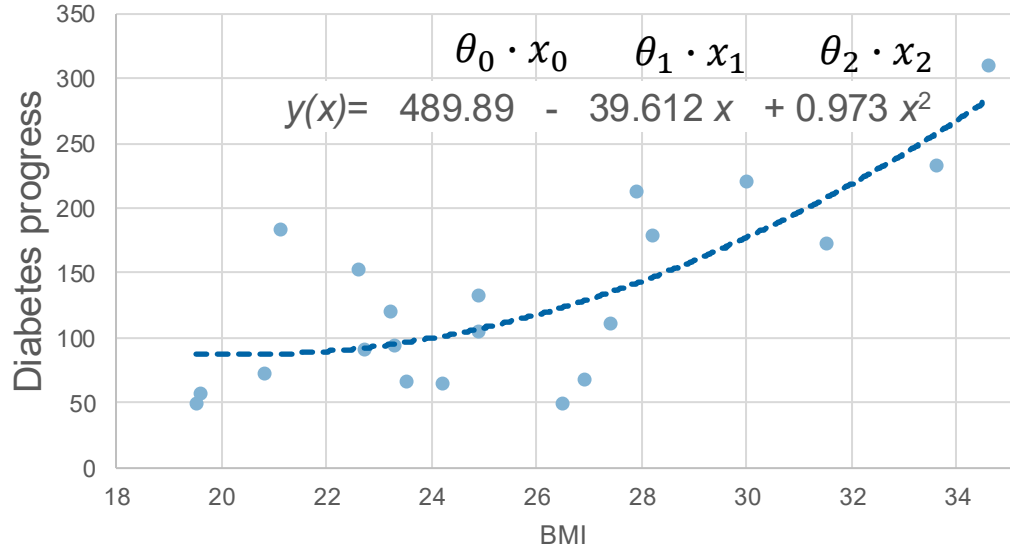
zh aw School of Engineering

Polynomial Regression

Modelling non-linear relationships using higher order polynomials

Zürcher Hochschule für Angewandte Wissenschaften
School of Engineering

e.g. quadratic function:



→ linear combination of higher order features.

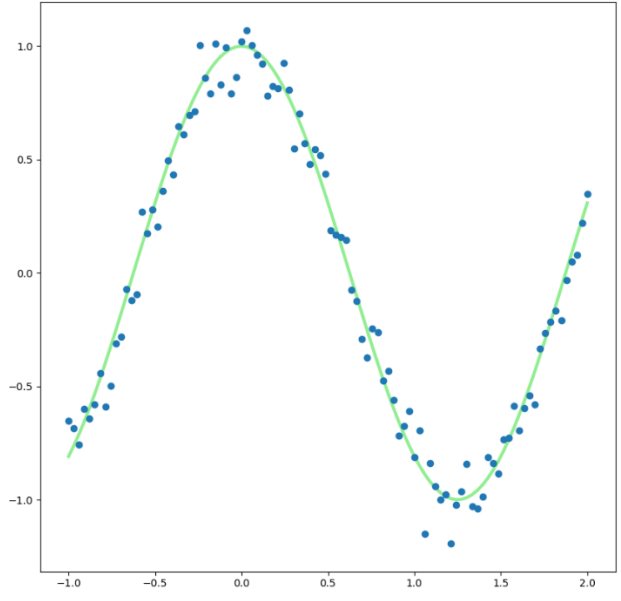
→ treat $x, x^2, \ldots, x^K$ as distinct independent features in a multivariate regression model and use the same procedure to solve for the optimal parameters.

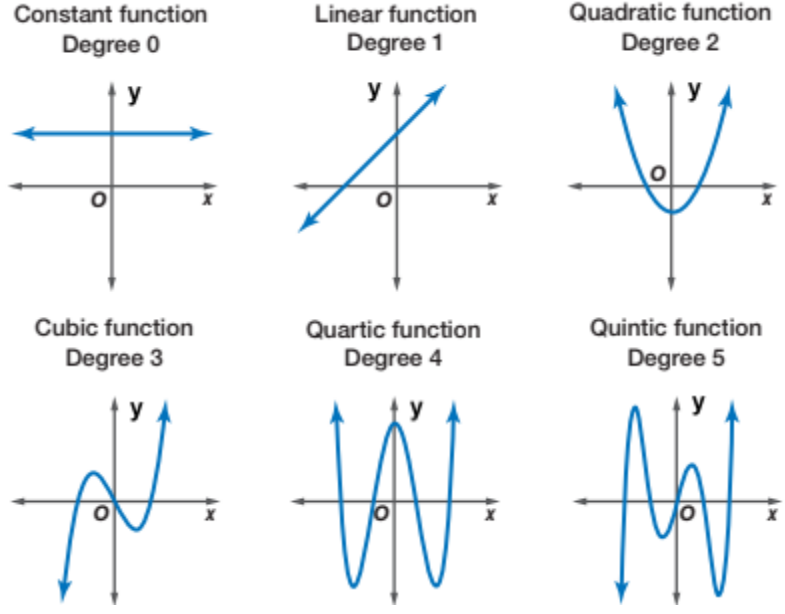
Higher Order Polynomials

Zürcher Hochschule
für Angewandte Wissenschaften

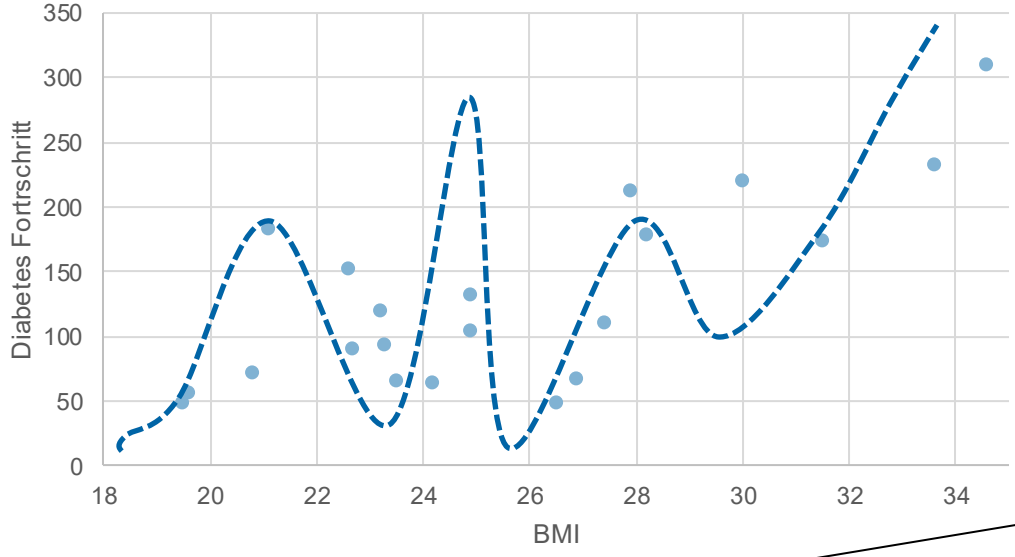
zh aw School of Engineering

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 (x_2)^3 + \theta_3 {x_2}^3 x_3 x_4$$





Overfitting



(here the order of the polynome)

Increasing the flexibility of the model (e.g. polynomial order) can, to a certain extent, lead to a better fit with the training data.

Overfitting occurs when the flexibility of the model leads to fitting the noise or outliers rather than representing the generalizing pattern of the data.

Other Aspects of Model Flexibility/Complexity

Zürcher Hochschule für Angewandte Wissenschaften
School of Engineering

	$K = 0$	$K = 1$	$K = 3$	$K = 9$
$\hat{w}_0$	-0.06	1.04	-0.11	0.38
$\hat{w}_1$		-2.04	11.28	-15.58
$\hat{w}_2$			-33.18	450.84
$\hat{w}_3$			22.13	-4228.63
$\hat{w}_4$				20509.34
$\hat{w}_5$				-57747.54
$\hat{w}_6$				97233.05
$\hat{w}_7$				-96374.54
$\hat{w}_8$				51856.86
$\hat{w}_9$				-11684.80

Values of the parameters learned during training of polynomial regression model with varying degree on the synthetic dataset

Overfitting is manifested by large oscillations, i.e. large absolute values of the coefficients

Besides restricting the number of parameters, the model's flexibility/complexity (and therefore overfitting) can be controlled by preventing larger values of the coefficients during model training → **Regularisation**

Zürcher Hochschule
für Angewandte Wissenschaften

zh aw School of Engineering

Regularisation in polynomial regression

Regularization

Zürcher Hochschule
für Angewandte Wissenschaften
zh aw School of Engineering

Hypothesis: $h(x_m; \theta_0, \theta_1, \dots, \theta_N) = \boldsymbol{\theta}^T \boldsymbol{x} = \theta_0 x_{m0} + \theta_1 x_{m1} + \dots + \theta_N x_{mN}$

Add an additional term to the cost function that penalises large values of the model parameters.

Cost Function of Ridge Regression:

$$J(\{(x_m, y_m)\}, \theta_0, \theta_1, \dots, \theta_N) = \frac{1}{2M}\left[\sum_{m=1}^{M} \left(y_m - h(x_m; \theta_0, \theta_1, \dots, \theta_N)\right)^2 + \boldsymbol{\lambda} \sum_{\boldsymbol{n=1}}^{\boldsymbol{N}} \boldsymbol{\theta_n}^2\right]$$

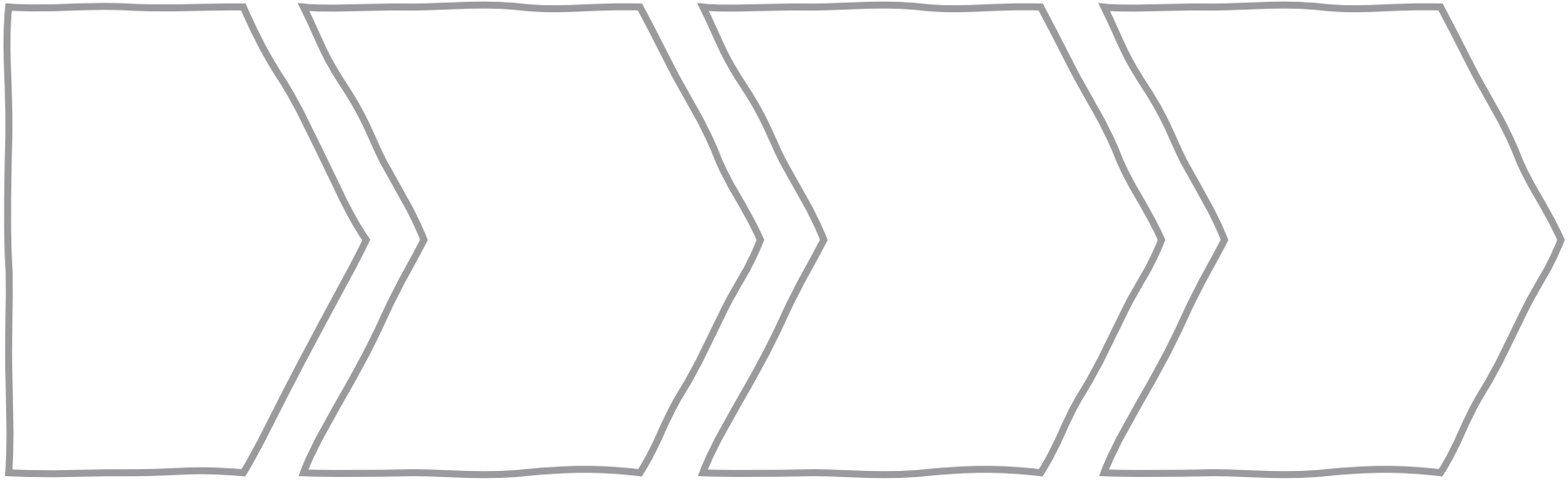
The hyperparameter λ is not optimised during training → model selection

Appendix

Structure of the Learning Problem

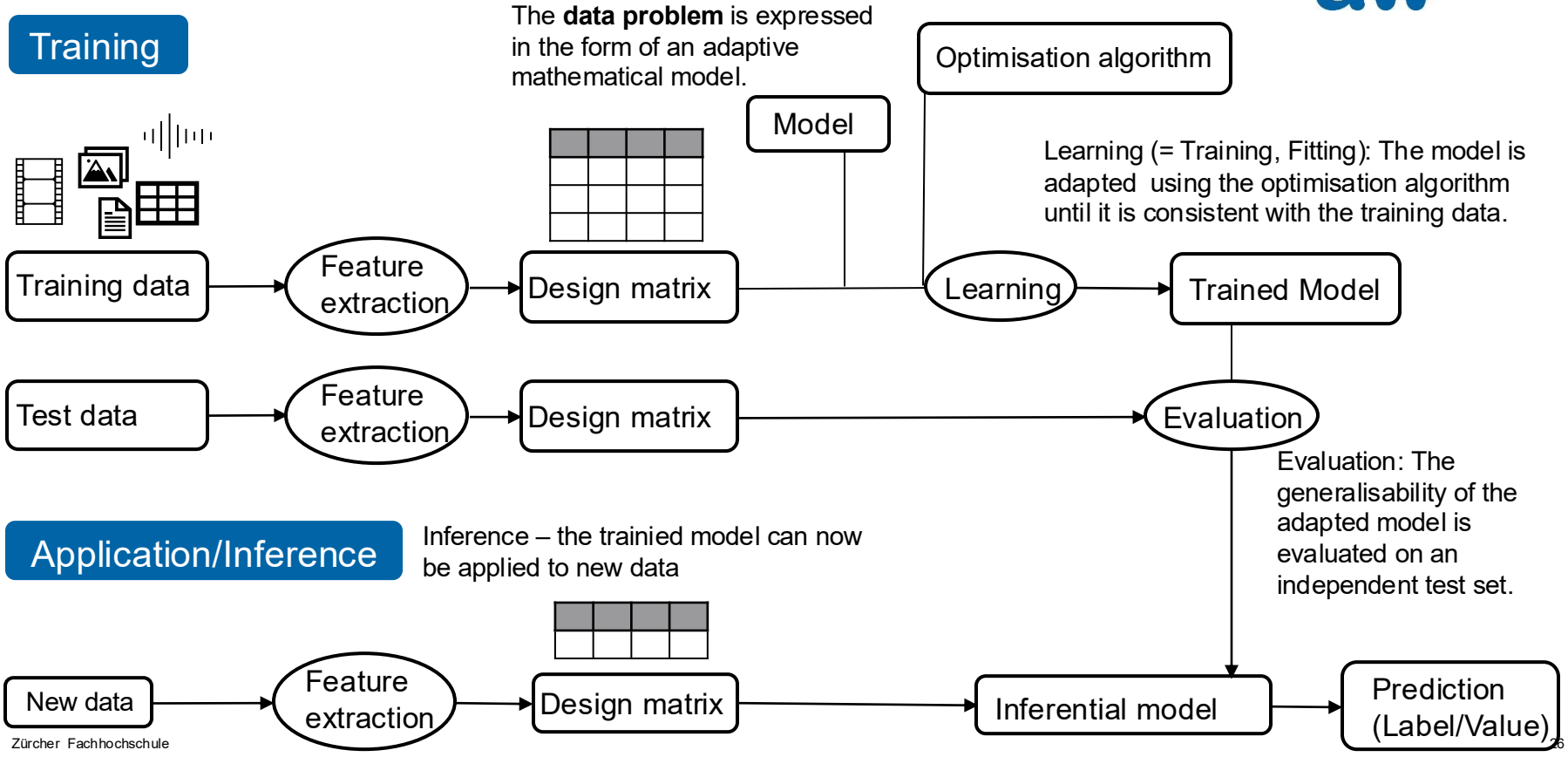
Zürcher Hochschule
für Angewandte Wissenschaften

School of
Engineering



Elements of a «Machine Learning Pipeline»

Zürcher Hochschule für Angewandte Wissenschaften
School of Engineering

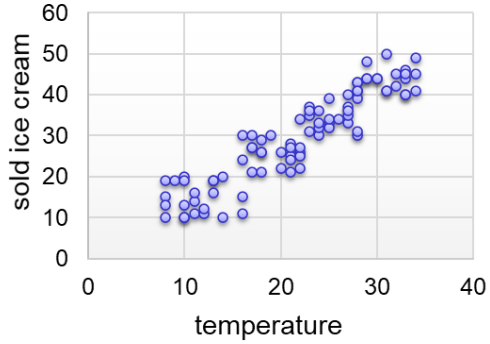


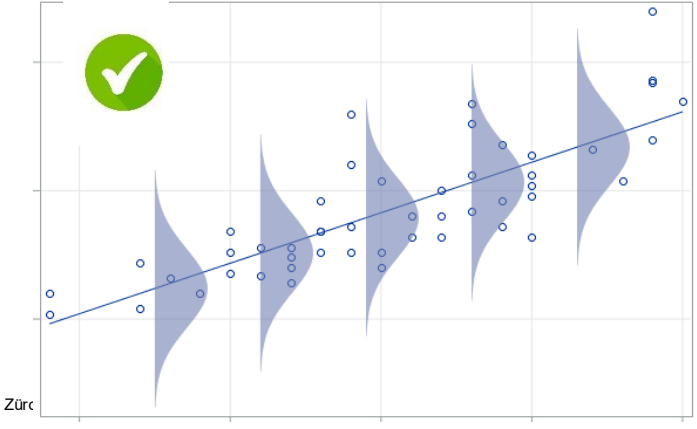
Basic Assumptions of Linear Regression

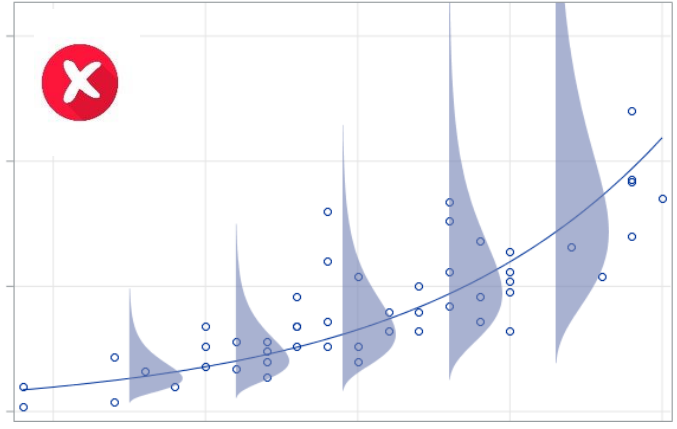
Zürcher Hochschule für Angewandte Wissenschaften

School of Engineering

- **Linearity:** The input and output values have a linear relationship
- **Independence:** The outcome of one sample does not affect the others
- **Normality**: Errors should be normally distributed, i.e. larger deviations from mean should be less likely
- **Equality of Variance ("*Homoscedasticity*")**: Error distribution should be the same for all input values





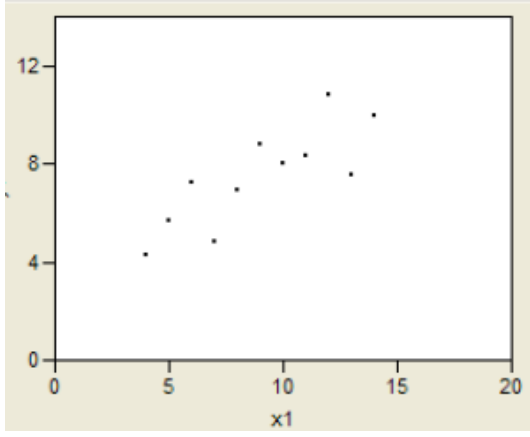


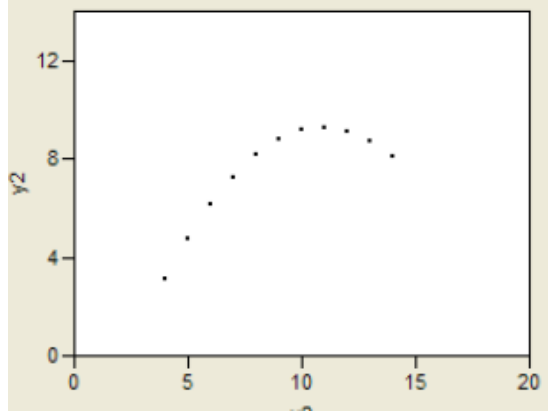
Evaluation Through Visual Inspection

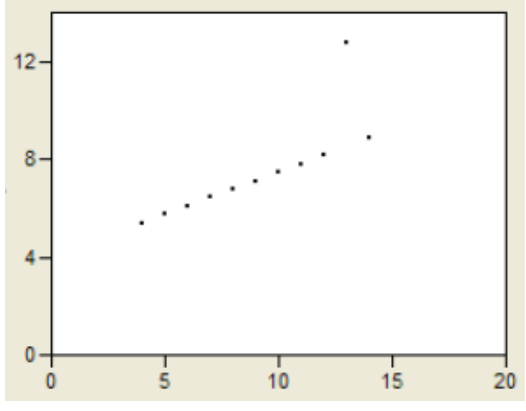
Zürcher Hochschule
für Angewandte Wissenschaften

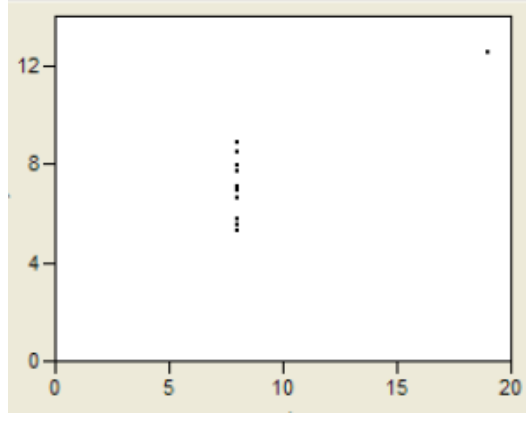
zh aw School of Engineering

Ancombe's Quartet:







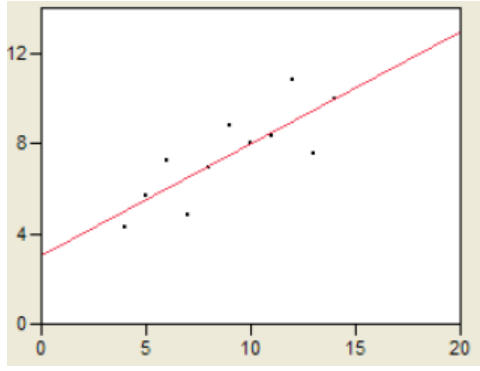


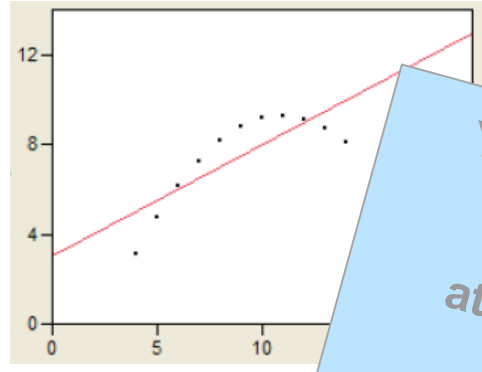
Evaluation Through Visual Inspection

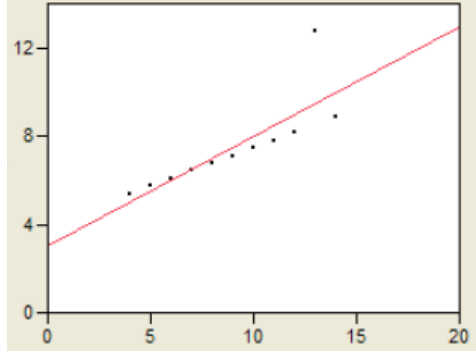
Zürcher Hochschule
für Angewandte Wissenschaften
zh aw School of Engineering

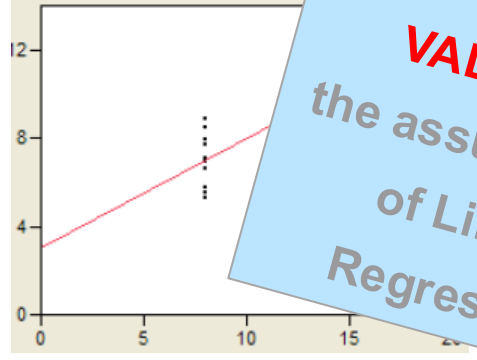
Linear Regression yields same fit line and same mean square error for all examples

Ancombe's Quartet:







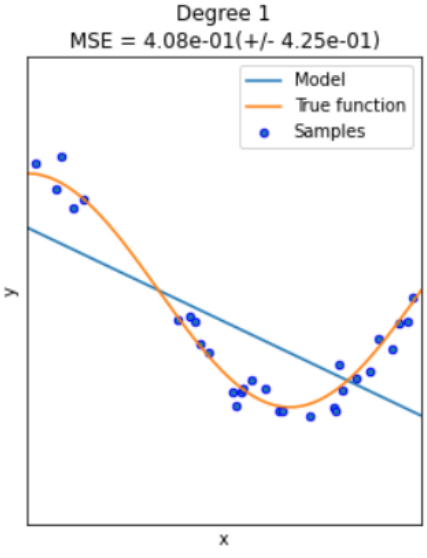


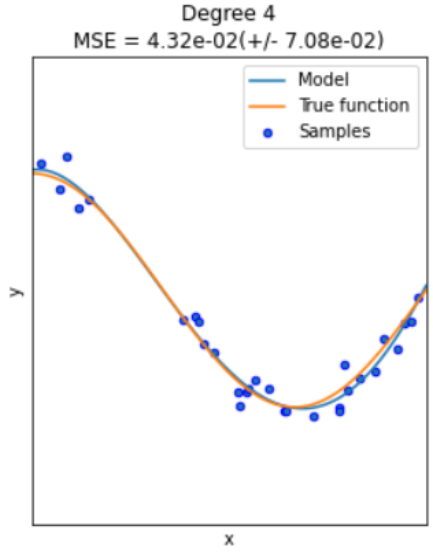
You have to
LOOK
at the data
and
VALIDATE
the assumptions
of Linear
Regression

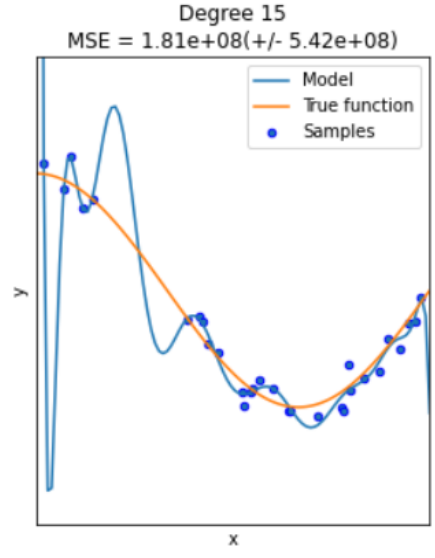
Higher Order Polynomials might Overfit

Zürcher Hochschule
für Angewandte Wissenschaften

zh aw School of Engineering



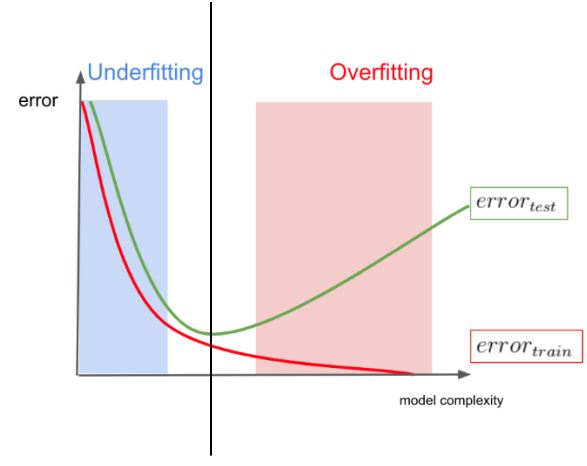




Underfitting

Overfitting

optimal model

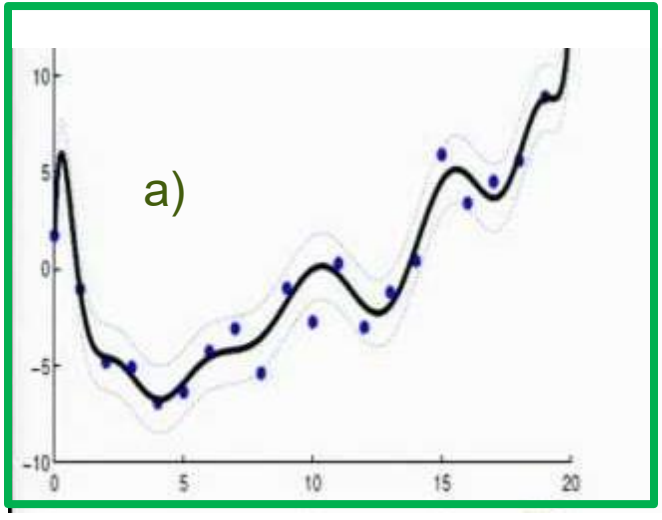


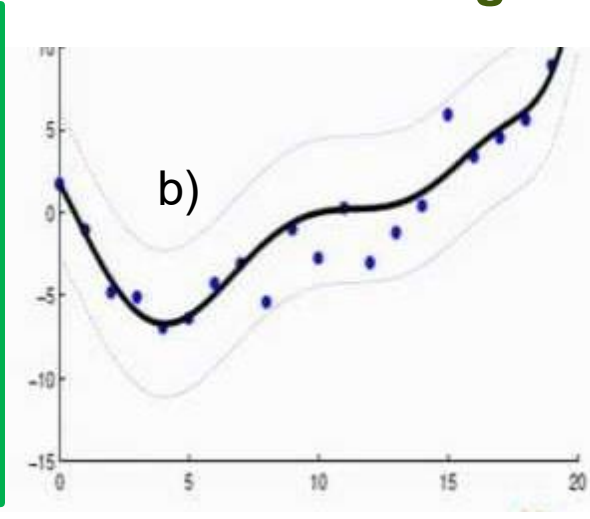
Solution: Regularization Parameter

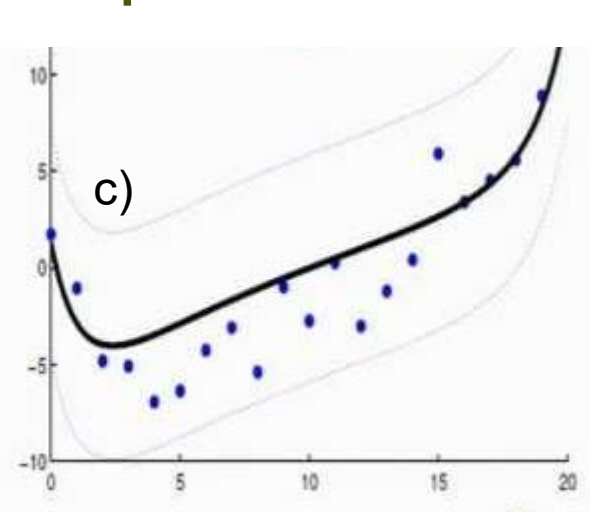
Zürcher Hochschule
für Angewandte Wissenschaften

zh aw School of Engineering

The images below show regression models with polynomials of degree 14. Which of them has the *lowest* value of regularization parameter λ?







Cost Function: $J(\theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2 + \lambda \sum_{j=1}^{n} \theta_j{}^2\right]$

SOLUTION: Large Regularization Parameter

Zürcher Hochschule für Angewandte Wissenschaften
School of Engineering

What happens if we set the regularization parameter λ to a very large value, e.g. λ = 10'000'000?

Cost Function: $J(\theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2 + \boldsymbol{\lambda \sum_{j=1}^{n} \theta_j^2}\right]$

Then all parameters θ_j will be close to or equal to zero, except for θ_0
Thus, we obtain a horizontal line.