Introduction to Machine Learning

Linear Regression Models

Department of Statistics – LMU Munich

LINEAR REGRESSION: HYPOTHESIS SPACE

We want to predict a numerical target variable by a *linear transformation* of the features $\mathbf{x} \in \mathbb{R}^p$.

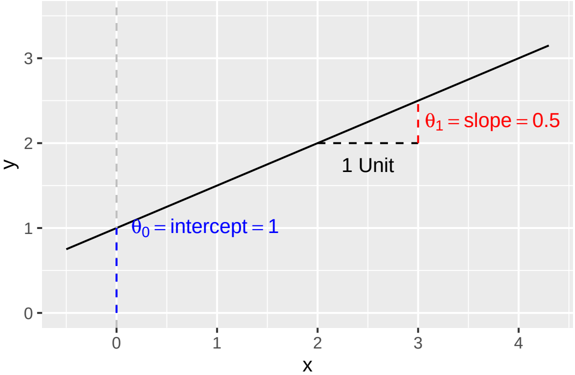
So with $\boldsymbol{\theta} \in \mathbb{R}^p$ this mapping can be written as:

$$\begin{aligned} y = f(\mathbf{x}) &= \theta_0 + \boldsymbol{\theta}^T \mathbf{x} \\ &= \theta_0 + \theta_1 x_1 + \cdots + \theta_p x_p \end{aligned}$$

This defines the hypothesis space $\mathcal{H}$ as the set of all linear functions in θ:

$$\mathcal{H} = \{\theta_0 + \boldsymbol{\theta}^T \mathbf{x} \mid (\theta_0, \boldsymbol{\theta}) \in \mathbb{R}^{p+1}\}$$

LINEAR REGRESSION: HYPOTHESIS SPACE



$$y = \theta_0 + \theta \cdot x$$

LINEAR REGRESSION: HYPOTHESIS SPACE

Given observed labeled data $\mathcal{D}$, how to find $(\theta_0, \boldsymbol{\theta})$?

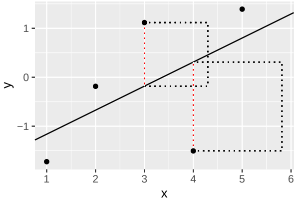
This is **learning** or parameter estimation, the learner does exactly this by **empirical risk minimization**.

NB: We assume from now on that θ_0 is included in $\boldsymbol{\theta}$.

LINEAR REGRESSION: RISK

We could measure training error as the sum of squared prediction errors (SSE). This is the risk that corresponds to **L2 loss**:

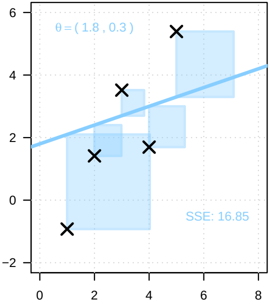
$$\mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) = \text{SSE}(\boldsymbol{\theta}) = \sum_{i=1}^{n} L\left(y^{(i)}, f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right) = \sum_{i=1}^{n} \left(y^{(i)} - \boldsymbol{\theta}^T \mathbf{x}^{(i)}\right)^2$$



Minimizing the squared error is computationally much simpler than minimizing the absolute differences (**L1 loss**).

LINEAR MODEL: OPTIMIZATION

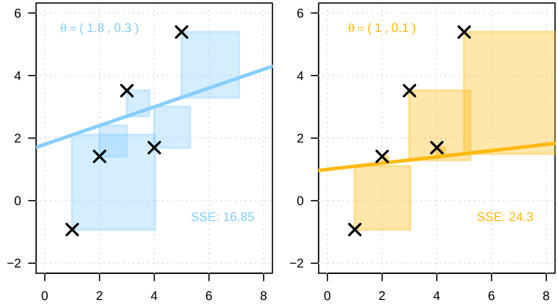
We want to find the parameters θ of the linear model, i.e., an element of the hypothesis space $\mathcal{H}$ that fits the data optimally.
So we evaluate different candidates for $\boldsymbol{\theta}$.
A first (random) try yields a rather large SSE: (**Evaluation**).



LINEAR MODEL: OPTIMIZATION

We want to find the parameters of the LM / an element of the hypothesis space $\mathcal{H}$ that best suits the data. So we evaluate different candidates for $\boldsymbol{\theta}$.

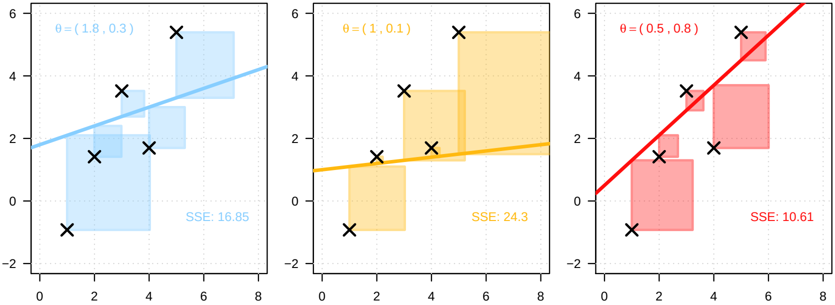
Another line yields an even bigger SSE (**Evaluation**). Therefore, this one is even worse in terms of empirical risk.



LINEAR MODEL: OPTIMIZATION

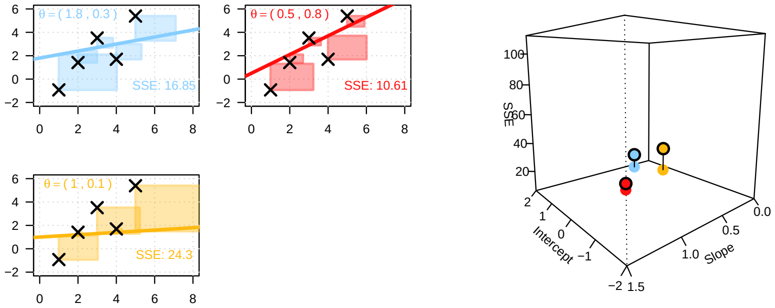
We want to find the parameters of the LM / an element of the hypothesis space $\mathcal{H}$ that best suits the data. So we evaluate different candidates for θ.

Another line yields an even bigger SSE (**Evaluation**). Therefore, this one is even worse in terms of empirical risk. Let's try again:



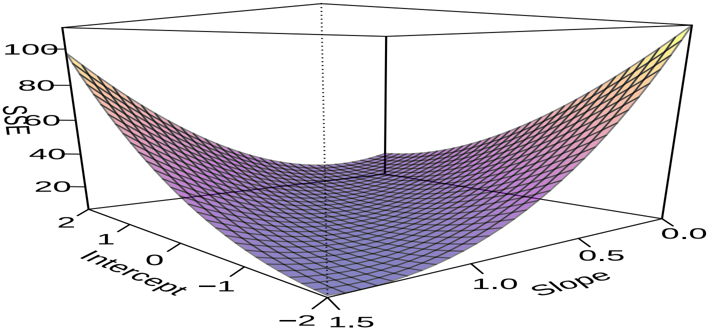
LINEAR MODEL: OPTIMIZATION

Since every θ results in a specific value of $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$, and we try to find $\arg\min_{\theta} \mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$, let's look at what we have so far:



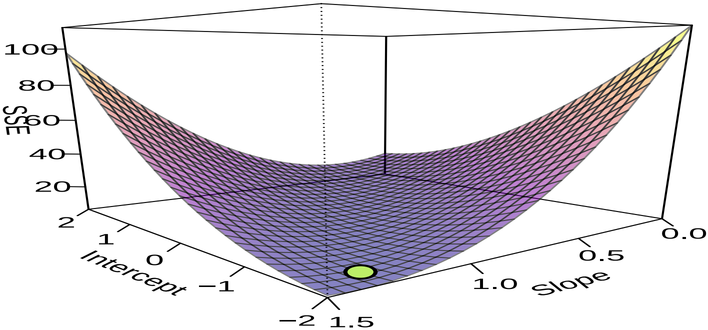
LINEAR MODEL: OPTIMIZATION

Instead of guessing, we use **optimization** to find the best θ:



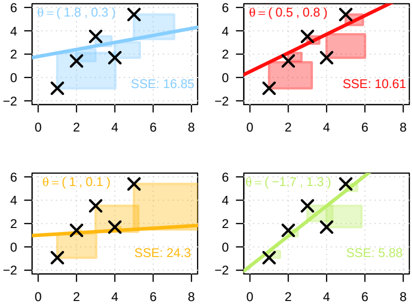
LINEAR MODEL: OPTIMIZATION

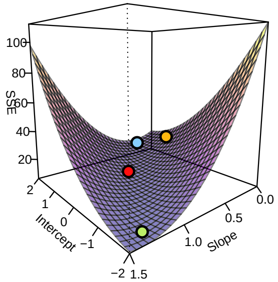
Instead of guessing, we use **optimization** to find the best $\boldsymbol{\theta}$:



LINEAR MODEL: OPTIMIZATION

Instead of guessing, we use **optimization** to find the best θ:





LINEAR MODEL: OPTIMIZATION

For L2 regression, we can find this optimal value analytically:

$$\begin{aligned}\hat{\theta} &= \arg\min_{\theta} \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left(y^{(i)} - \theta^T \mathbf{x}^{(i)}\right)^2 \\ &= \arg\min_{\theta} \|\mathbf{y} - X\boldsymbol{\theta}\|_2^2\end{aligned}$$

where $X = \begin{pmatrix} 1 & x_1^{(1)} & \ldots & x_p^{(1)} \\ 1 & x_1^{(2)} & \ldots & x_p^{(2)} \\ \vdots & \vdots & & \vdots \\ 1 & x_1^{(n)} & \ldots & x_p^{(n)} \end{pmatrix}$ is the $n \times (p+1)$-**design matrix**.

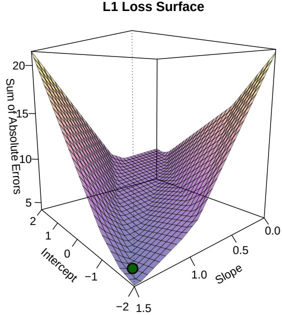
This yields the so called normal equations for the LM:

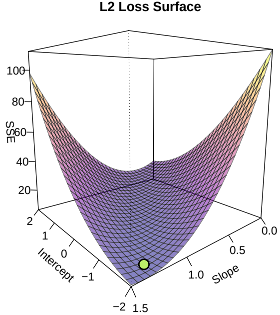
$$\frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) = 0 \quad \Longrightarrow \quad \hat{\boldsymbol{\theta}} = \left(X^T X\right)^{-1} X^T \mathbf{y}$$

EXAMPLE: REGRESSION WITH L1 VS L2 LOSS

We could also minimize the L1 loss. This changes the risk and optimization steps:

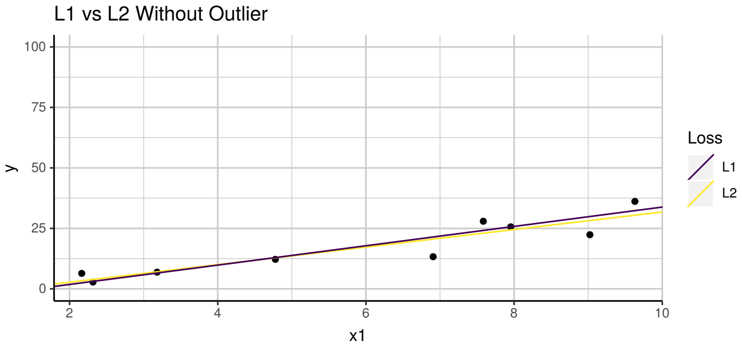
$$\mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} L\left(y^{(i)}, f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right) = \sum_{i=1}^{n} \left|y^{(i)} - \theta^T \mathbf{x}^{(i)}\right| \qquad \text{(Risk)}$$





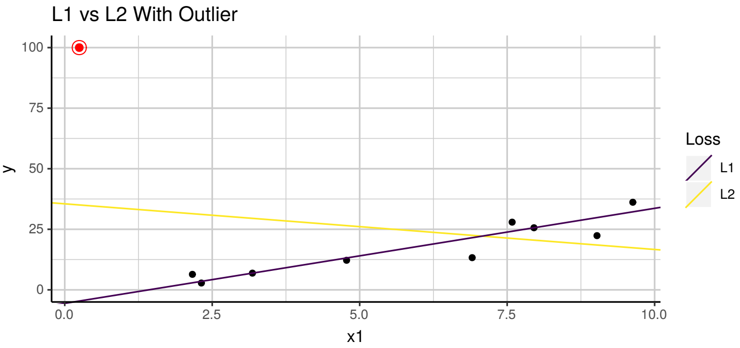
L1 loss is harder to optimize, but the model is less sensitive to outliers.

EXAMPLE: REGRESSION WITH L1 VS L2 LOSS



EXAMPLE: REGRESSION WITH L1 VS L2 LOSS

Adding an outlier (highlighted red) pulls the line fitted with L2 into the direction of the outlier:



LINEAR REGRESSION

Hypothesis Space: Linear functions $\mathbf{x}^T\boldsymbol{\theta}$ of features $\in \mathcal{X}$.

Risk: Any regression loss function.

Optimization: Direct analytic solution for L2 loss, numerical optimization for L1 and others.