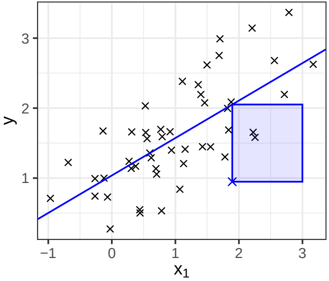
Introduction to Machine Learning

Supervised Regression: Linear Models with $L2$ Loss



Learning goals

- Grasp the overall concept of linear regression
- Understand how $L2$ loss optimization results in SSE-minimal model
- Understand this as a general template for ERM in ML

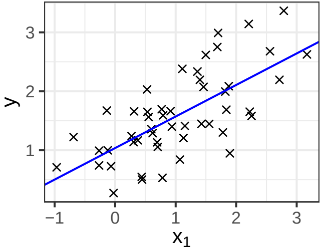
LINEAR REGRESSION

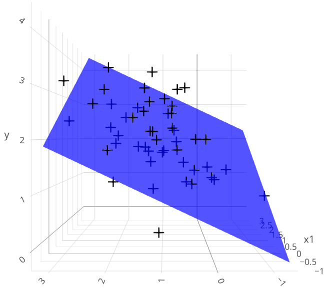
- Idea: predict $y \in \mathbb{R}$ as **linear** combination of features[1]:

$$\hat{y} = f(\mathbf{x}) = \boldsymbol{\theta}^\top \mathbf{x} = \theta_0 + \theta_1 x_1 + \dots + \theta_p x_p$$

$\rightsquigarrow$ find loss-optimal params to describe relation $y|\mathbf{x}$

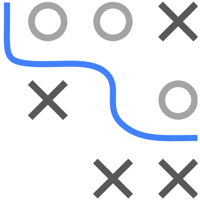
- Hypothesis space: $\mathcal{H} = \{f(\mathbf{x}) = \boldsymbol{\theta}^\top \mathbf{x} \mid \boldsymbol{\theta} \in \mathbb{R}^{p+1}\}$





[1] Actually, special case of linear model, which is linear combo of *basis functions* of features $\rightsquigarrow$ Polynomial Regression Models

DESIGN MATRIX

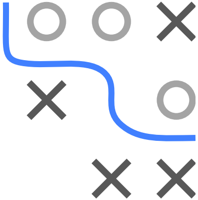


- Mismatch: $\boldsymbol{\theta} \in \mathbb{R}^{p+1}$ vs $\mathbf{x} \in \mathbb{R}^p$ due to intercept term
- Trick: pad feature vectors with leading 1, s.t.
  - $\mathbf{x} \mapsto \mathbf{x} = (1, x_1, \ldots, x_p)^\top$, and
  - $\boldsymbol{\theta}^\top \mathbf{x} = \theta_0 \cdot 1 + \theta_1 x_1 + \cdots + \theta_p x_p$
- Collect all observations in **design matrix** $\mathbf{X} \in \mathbb{R}^{n \times (p+1)}$
  $\rightsquigarrow$ more compact: single param vector incl. intercept
- Resulting linear model:

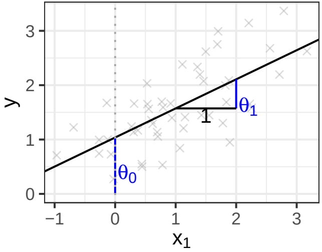
$$\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\theta} = \begin{pmatrix} 1 & x_1^{(1)} & \ldots & x_p^{(1)} \\ 1 & x_1^{(2)} & \ldots & x_p^{(2)} \\ \vdots & \vdots & & \vdots \\ 1 & x_1^{(n)} & \ldots & x_p^{(n)} \end{pmatrix} \begin{pmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_p \end{pmatrix} = \begin{pmatrix} \theta_0 + \theta_1 x_1^{(1)} + \cdots + \theta_p x_p^{(1)} \\ \theta_0 + \theta_1 x_1^{(2)} + \cdots + \theta_p x_p^{(2)} \\ \vdots \\ \theta_0 + \theta_1 x_1^{(n)} + \cdots + \theta_p x_p^{(n)} \end{pmatrix}$$

- We will make use of this notation in other contexts

EFFECT INTERPRETATION



- Big plus of LM: immediately **interpretable** feature effects
- "Marginally increasing x_j by 1 unit increases y by θ_j units"
  $\rightsquigarrow$ *ceteris paribus* assumption: $x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p$ fixed

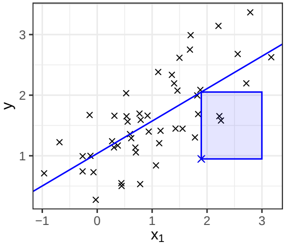



```
Call:
lm(formula = y ~ x_1, data = dt_univ)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.10346 -0.34727 -0.00766  0.31500  1.04284 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.03727    0.11360   9.131 4.55e-12 ***
x_1          0.53521    0.08219   6.512 4.13e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5327 on 48 degrees of freedom
Multiple R-squared:  0.469,	Adjusted R-squared:  0.458 
F-statistic:  42.4 on 1 and 48 DF,  p-value: 4.129e-08
```

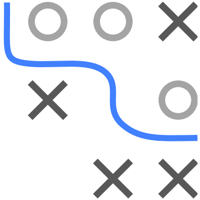
MODEL FIT

- How to determine LM fit? $\rightsquigarrow$ define risk & optimize
- Popular: *L*2 **loss** / **quadratic loss** / **squared error**

$$L(y, f(\mathbf{x})) = (y - f(\mathbf{x}))^2 \text{ or } L(y, f(\mathbf{x})) = 0.5 \cdot (y - f(\mathbf{x}))^2$$

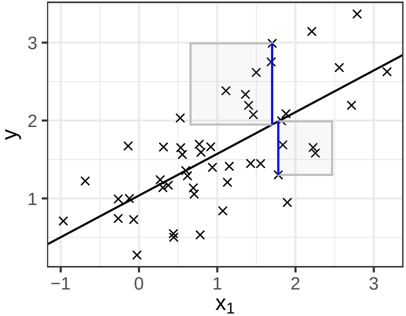


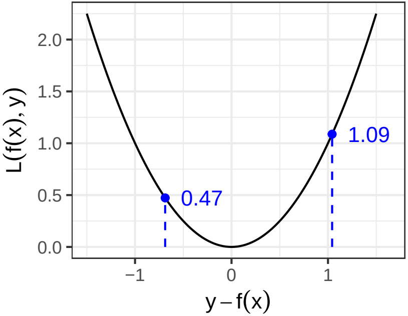
- Why penalize **residuals** $r = y - f(\mathbf{x})$ quadratically?
  - Easy to optimize (convex, differentiable)
  - Theoretically appealing (connection to classical stats LM)



LOSS PLOTS

We will often visualize loss effects like this:





- Data as $y \sim x_1$
- Prediction hypersurface
  $\rightsquigarrow$ here: line
- Residuals $r = y - f(\mathbf{x})$
  $\rightsquigarrow$ squares to illustrate loss

- Loss as function of residuals
  $\rightsquigarrow$ strength of penalty?
  $\rightsquigarrow$ symmetric?
- Highlighted: loss for residuals shown on LHS

OPTIMIZATION

- Resulting risk equivalent to **sum of squared errors (SSE)**:

$$\mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left(y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)} \right)^2$$

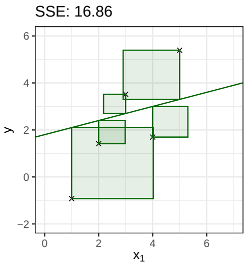
- Consider example with $n = 5 \rightsquigarrow$ different models with varying SSE

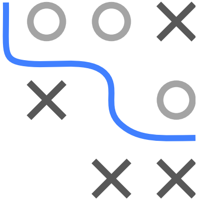
OPTIMIZATION

- Resulting risk equivalent to **sum of squared errors (SSE)**:

$$\mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left(y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)} \right)^2$$

- Consider example with $n = 5 \rightsquigarrow$ different models with varying SSE



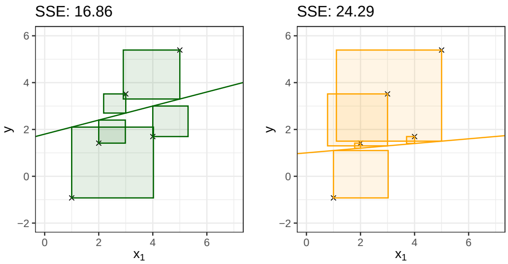


OPTIMIZATION

- Resulting risk equivalent to **sum of squared errors (SSE)**:

$$\mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left(y^{(i)} - \boldsymbol{\theta}^{\top} \mathbf{x}^{(i)} \right)^2$$

- Consider example with $n = 5 \rightsquigarrow$ different models with varying SSE

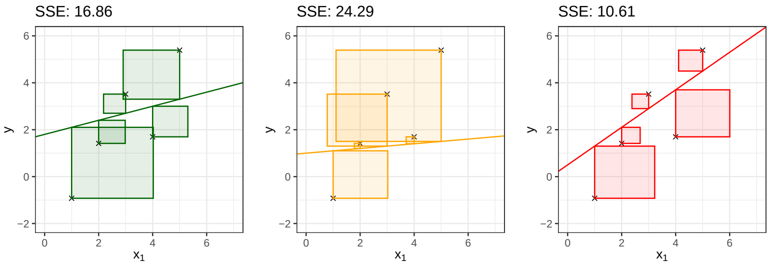


OPTIMIZATION

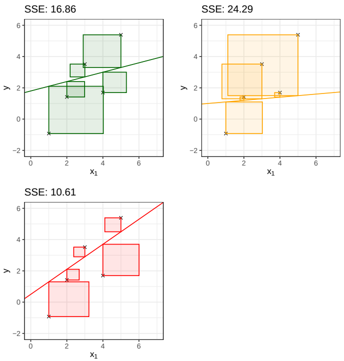
- Resulting risk equivalent to **sum of squared errors (SSE)**:

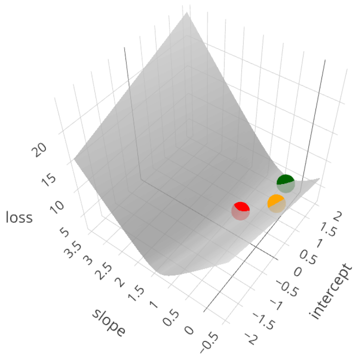
$$\mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left(y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)} \right)^2$$

- Consider example with $n = 5 \rightsquigarrow$ different models with varying SSE



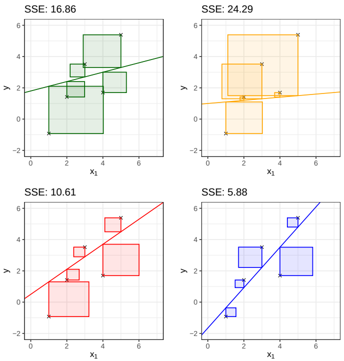
OPTIMIZATION

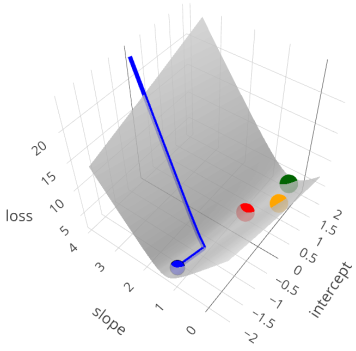




Intercept θ_0	Slope θ_1	SSE
1.80	0.30	16.86
1.00	0.10	24.29
0.50	0.80	10.61

OPTIMIZATION

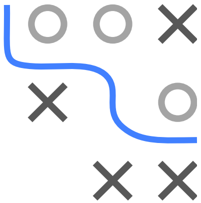




Intercept θ_0	Slope θ_1	SSE
1.80	0.30	16.86
1.00	0.10	24.29
0.50	0.80	10.61
-1.65	1.29	5.88

Instead of guessing, of course, use **optimization**!

ANALYTICAL OPTIMIZATION



- Special property of LM with $L2$ loss: **analytical solution** available

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

- Find via **normal equations**

$$\frac{\partial \mathcal{R}_{\text{emp}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0$$

- Solution: **ordinary-least-squares (OLS)** estimator

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

STATISTICAL PROPERTIES

- LM with *L*2 loss intimately related to classical stats LM
- Assumptions
  - $\mathbf{x}^{(i)}$ **iid** for $i \in \{1, \dots, n\}$
  - **Homoskedastic** (equivariant) **Gaussian** errors

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}, \ \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I})$$

    $\rightsquigarrow$ y_i conditionally independent & normal: $\mathbf{y}|\mathbf{X} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\theta}, \sigma^2 \boldsymbol{I})$
  - Uncorrelated features

    $\rightsquigarrow$ multicollinearity destabilizes effect estimation
- If assumptions hold: statistical **inference** applicable
  - Hypothesis tests on significance of effects, incl. *p*-values
  - Confidence & prediction intervals via student-*t* distribution
  - Goodness-of-fit measure $R^2 = 1 - \text{SSE} \ / \ \underbrace{\text{SST}}_{\sum_{i=1}^{n} (y^{(i)} - \bar{y})^2}$

    $\rightsquigarrow$ SSE = part of data variance *not* explained by model