CS4220 Machine Learning

Linear Regression

Monday, 20 November 2023

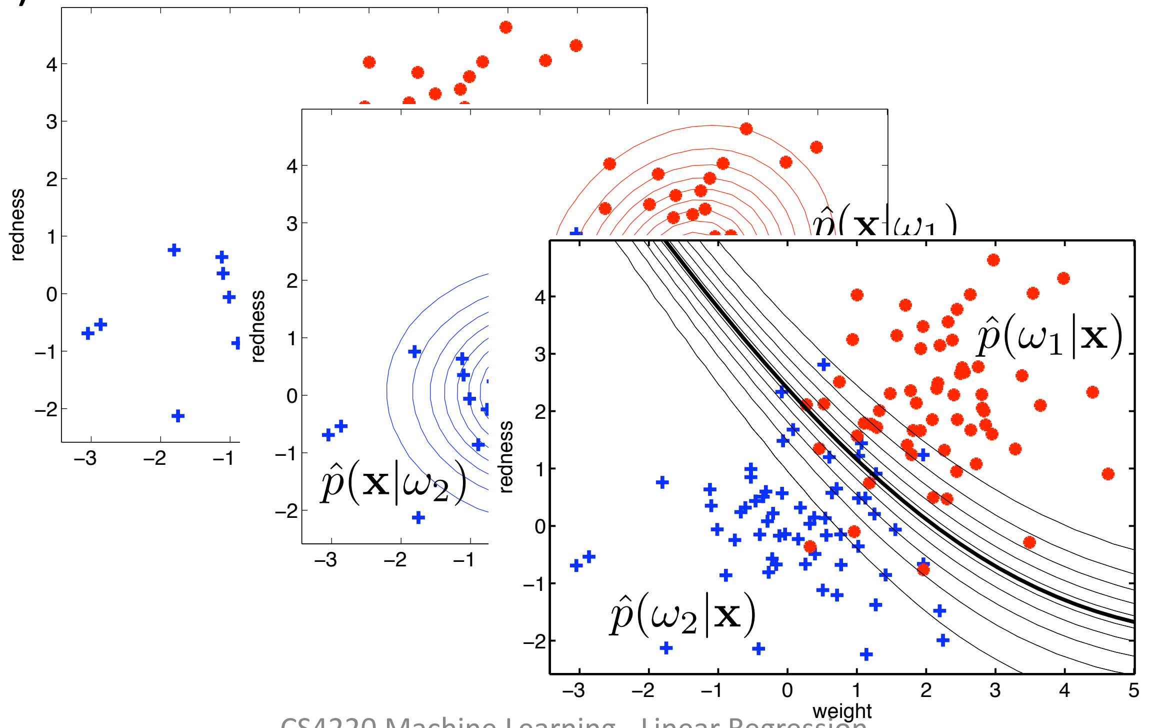
Merve Gürel

TU Delft

Last Week

Classification

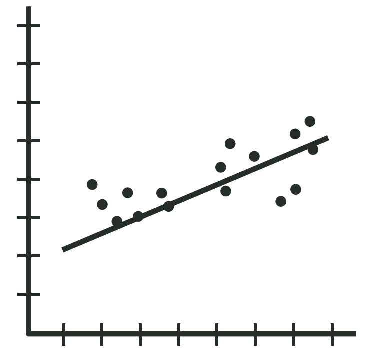
- How to set a decision boundary using Bayes' Rule, Bayes optimal classifier, Misclassification Costs
- Parametric & non-parametric classifiers (QDA, LDA … & Histogram, Parzen classifier…)



This Week

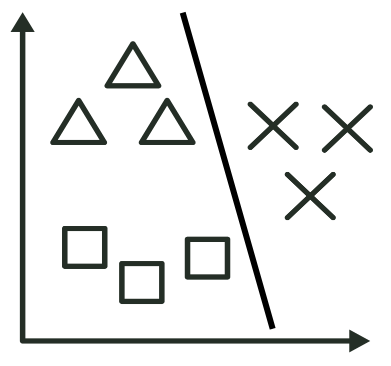
Regression

- Focus on the Linear Regression



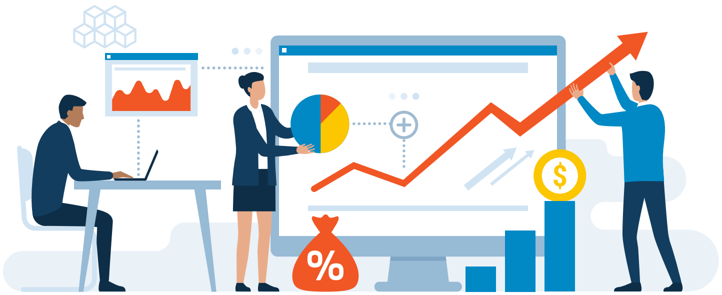
Classification

- Focus on Linear Classifiers
  (more on this tomorrow!)



Regression

- Many application areas from finance and health to agriculture



- Today's focus is on the Linear Regression
  - Function fitting with Ordinary Least Squares
  - Dealing with nonlinearity via feature transformation
  - Gaussian Error
  - Regularization
  - Bayesian Linear Regression

Goals for Today

At the end of this lecture, you should be able to

- ❑ Identify regression problems
- ❑ For linear regression (with or without feature transformation), explain
  - ❑ Least Squares solution and its geometric interpretation
  - ❑ Maximum Likelihood Estimation (MLE) and Maximum a Posteriori (MAP) estimation for Gaussian models
- ❑ Apply regularization
  - ❑ Ridge, Lasso Regression

Regression Approaches

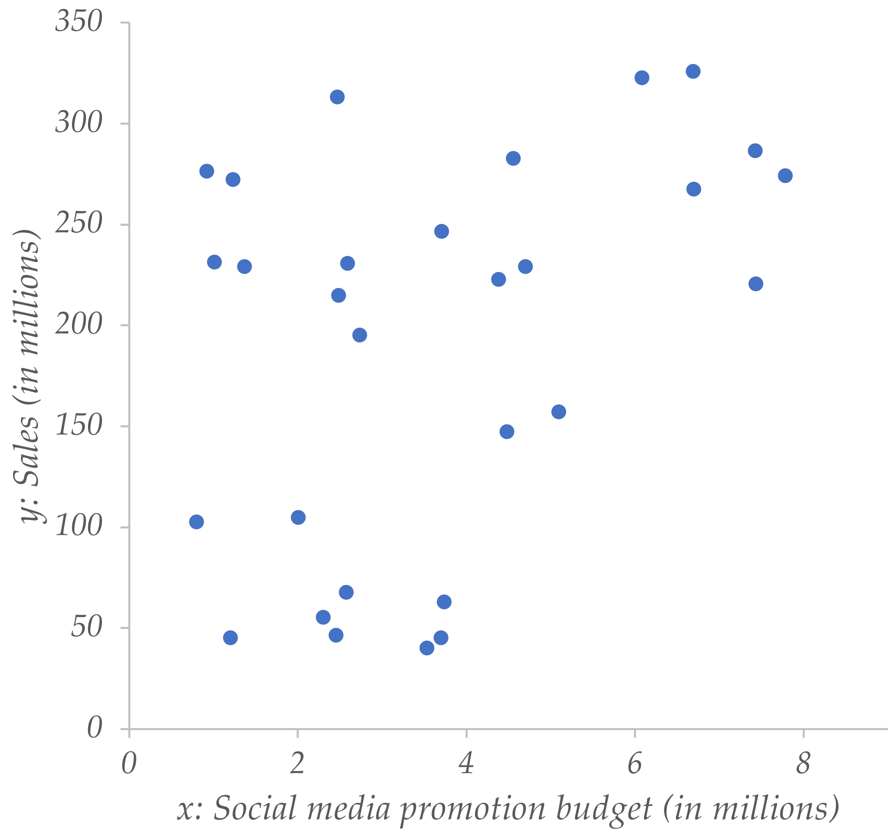
Preliminaries: Given a set of features $x \in \mathbb{R}^d$ *we want to predict a target variable* $y \in \mathbb{R}^m$

- Function fitting: Assume $y = f(x)$ and learn
$$f(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}^m$$

- Probabilistic Approach
  - Maximum Likelihood estimation: Model $p(\mathrm{data}|\mathrm{parameter})$
  - Maximum a Posteriori Estimation: Model $p(\mathrm{data}, \mathrm{parameter})$

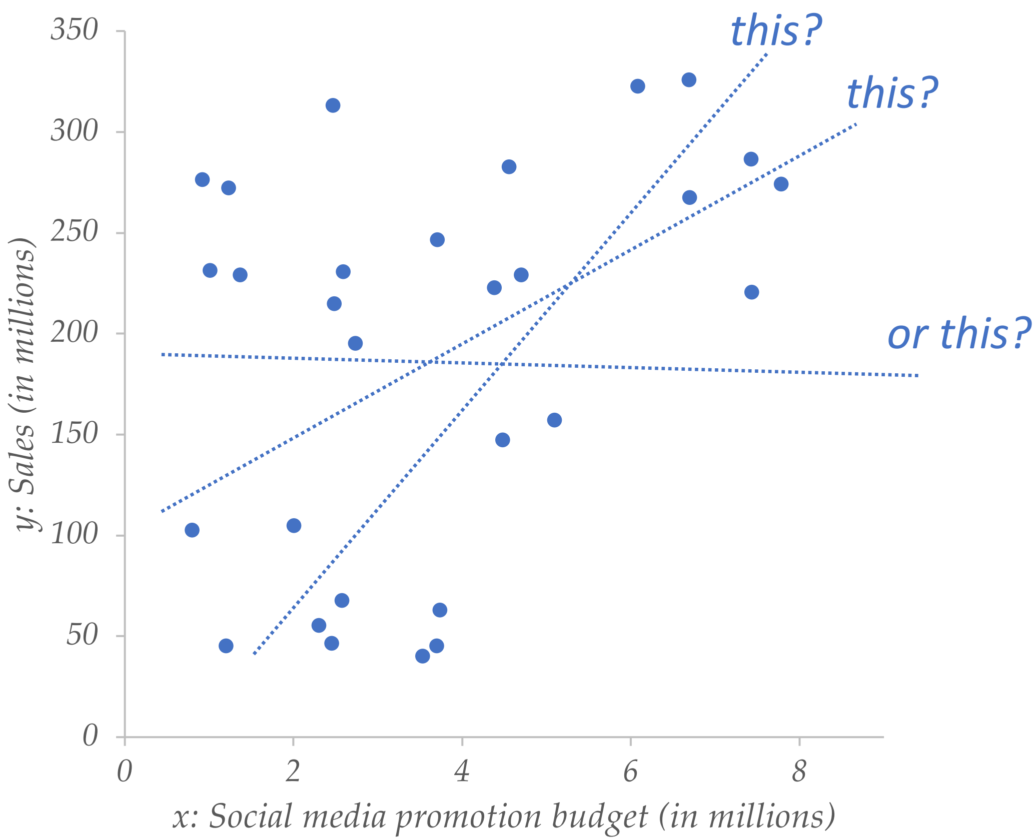
Linear Regression, Function fitting and Ordinary Least Squares

Linear Regression



- We are given input-output observations $d = 1, m = 1$
- How to predict an unseen x^*
- Fit a linear function $f(x)$ to y using the observed data

Linear Regression



- Assume that the function $f(x)$ is linear in x

$$y = \beta x + \beta_0$$

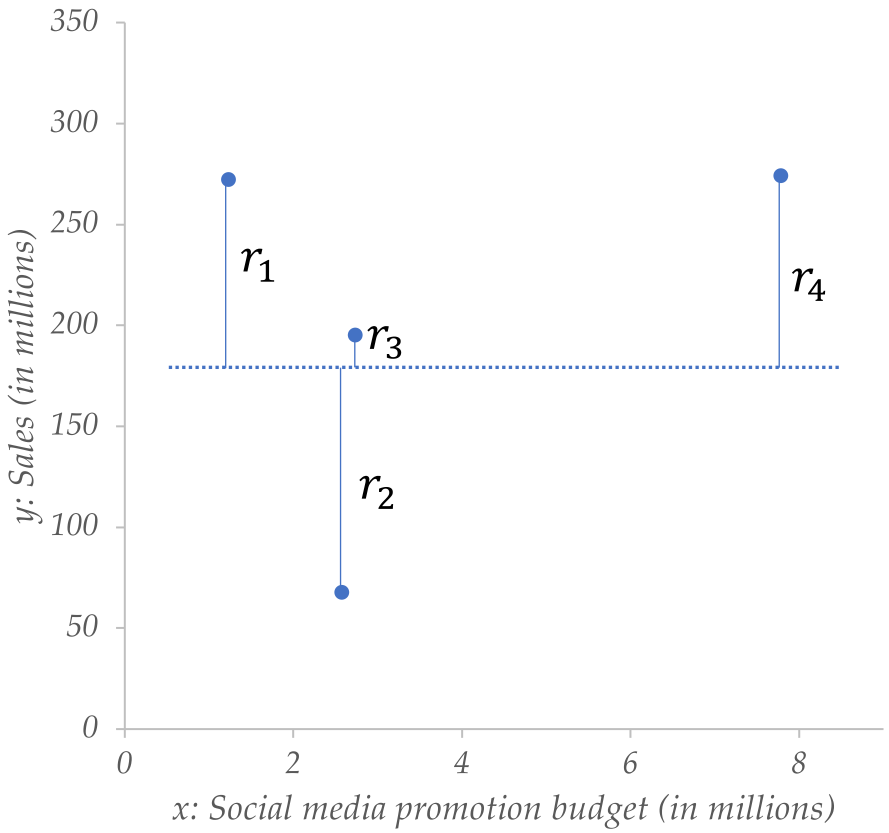
slope *intercept*

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_d \end{pmatrix} \quad x = \begin{pmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix}$$

$$y = x^T \beta$$

How do we estimate the model parameters β?

Residuals



Given $(x_n, y_n) \quad n = 1, 2, \ldots, N$ we want to estimate β such that y and $\hat{y}$ are as close as possible

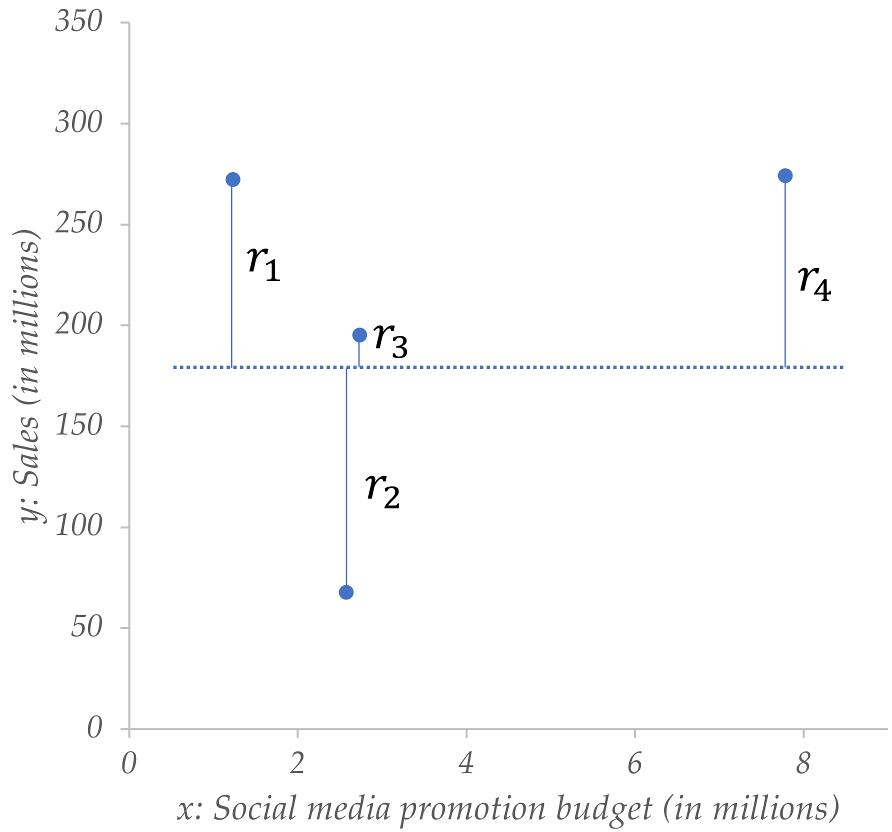
Consider the residuals:

$$r_n = y_n - \hat{y}_n = y_n - x_n^T \hat{\beta}$$

Initial idea is to make sure the residuals are small

Should we minimize $\sum_{n=1}^{N} |r_n|$?

Ordinary Least Squares Solution to Linear Regression



- Minimize the sum of residual squares!
- Estimate β such that $\sum_{n=1}^{N} (r_n)^2$ is minimized

Ordinary Least Squares Solution to Linear Regression

- Let's combine all the data for simplicity: $X = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} & \dots & x_{1d} \\ 1 & x_{21} & x_{22} & x_{23} & \dots & x_{2d} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & x_{N2} & x_{N3} & \dots & x_{Nd} \end{pmatrix}$ $Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{pmatrix}$

- Parameter estimation:

$$\hat{\beta}_{OLS} = \arg \min_{\beta \in \mathbb{R}^{d+1}} \sum_{n=1}^{N} (y_n - x_n^T \beta)^2 = \arg \min_{\beta \in \mathbb{R}^{d+1}} (Y - X\beta)^T (Y - X\beta)$$

- Solution given at $\frac{\partial}{\partial \beta} (Y - X\beta)^T (Y - X\beta) = 0$

- Therefore: $\hat{\beta}_{OLS} = (X^T X)^{-1} X^T Y$

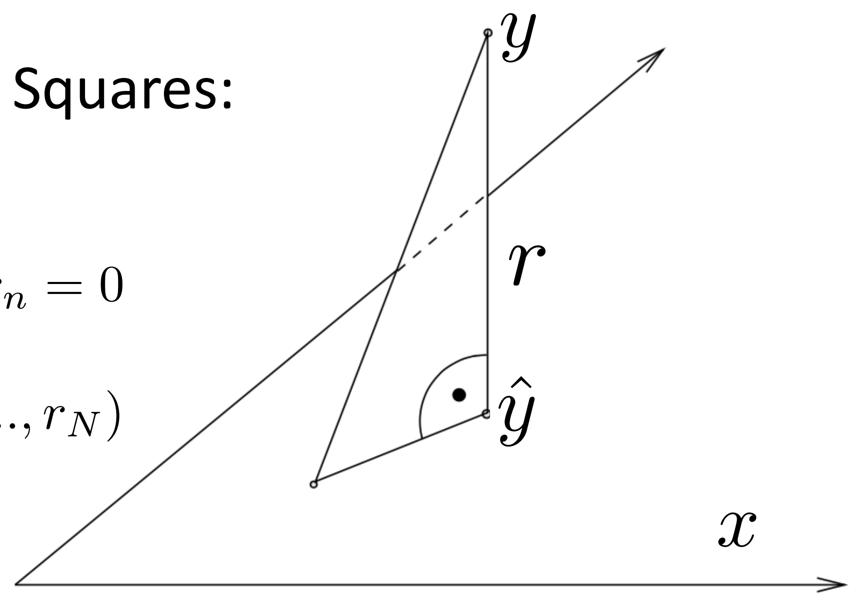
Inverse exists if X is full column rank. In other words, if the features are linearly independent. Otherwise, either use regularization

Geometric Interpretation

- Closer look at the Least Squares:
  (let $p = 1$)

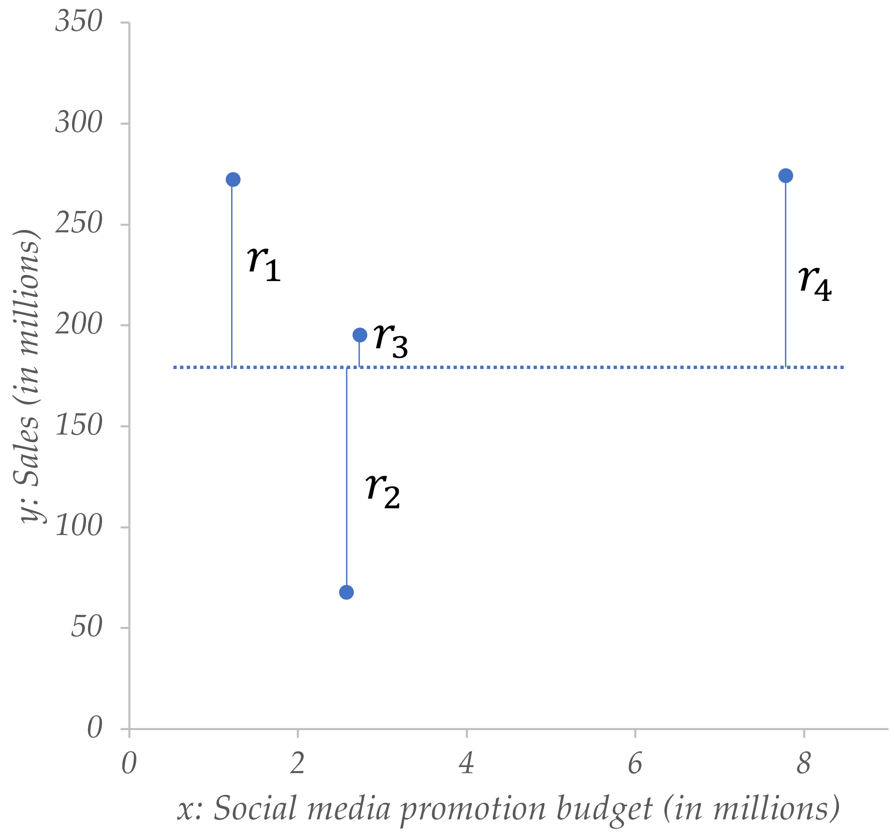
$$\sum_n x_n(y_n - x_n^T\beta) = \sum_n x_n r_n = 0$$

$$(x_1, .., x_N) \perp (r_1, ..., r_N)$$



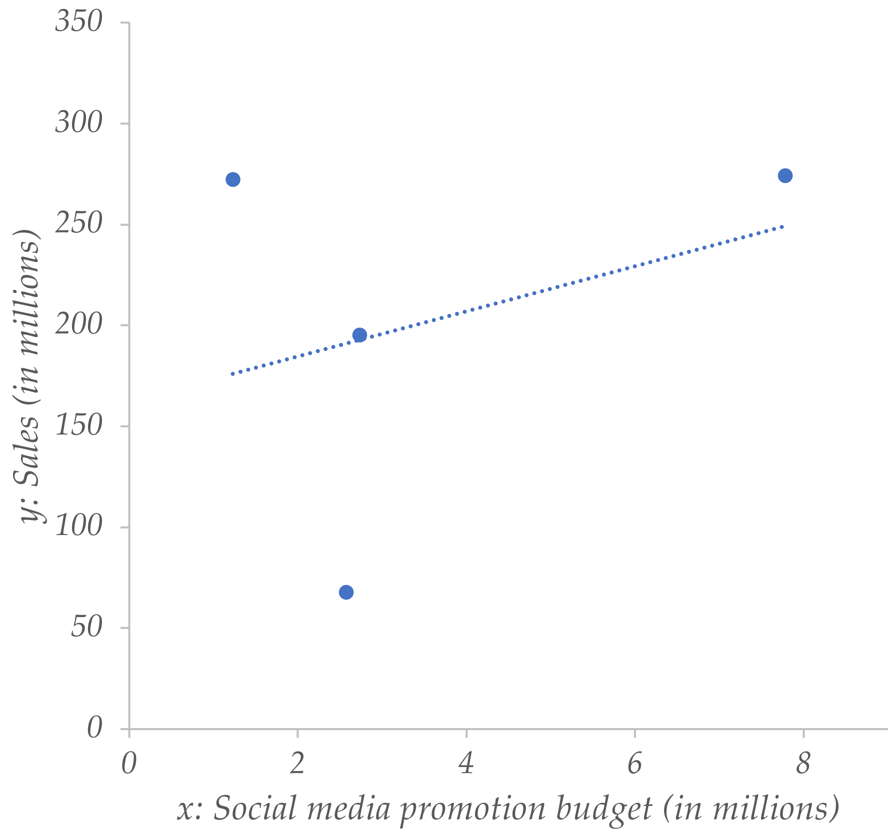
- Residuals are projection of y onto a d-dimensional subspace in $\mathbb{R}^n$

Least Squares Regression



- Minimize the sum of residual squares!
- Estimate β such that $\sum_{n=1}^{N}(r_n)^2$ is minimized

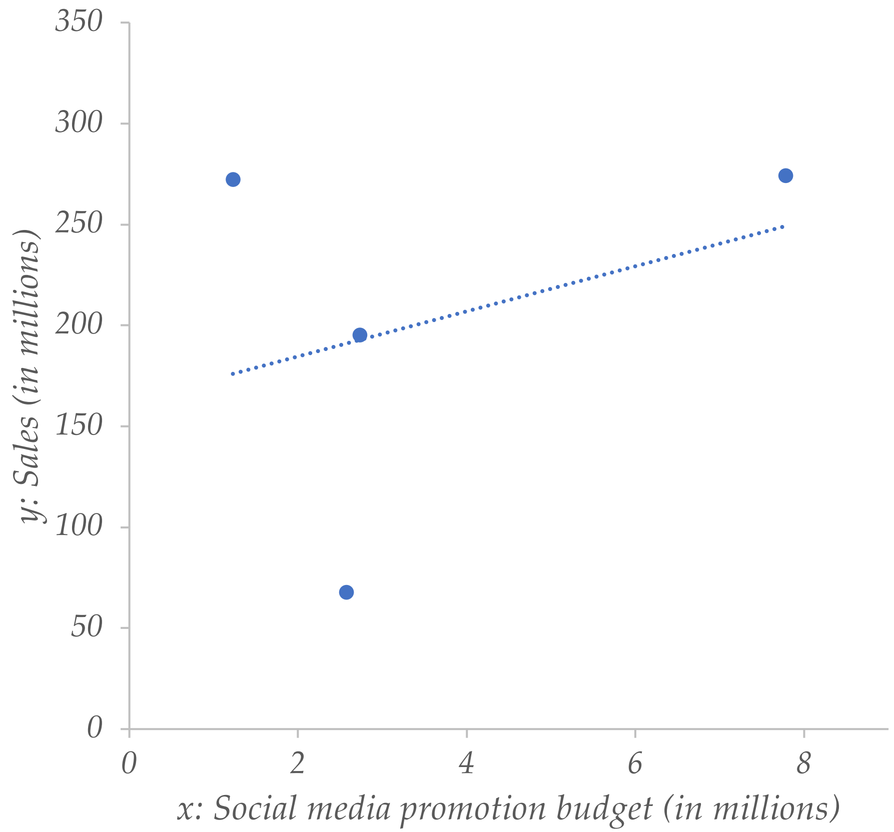
Least Squares Regression



- Least Squares solution where

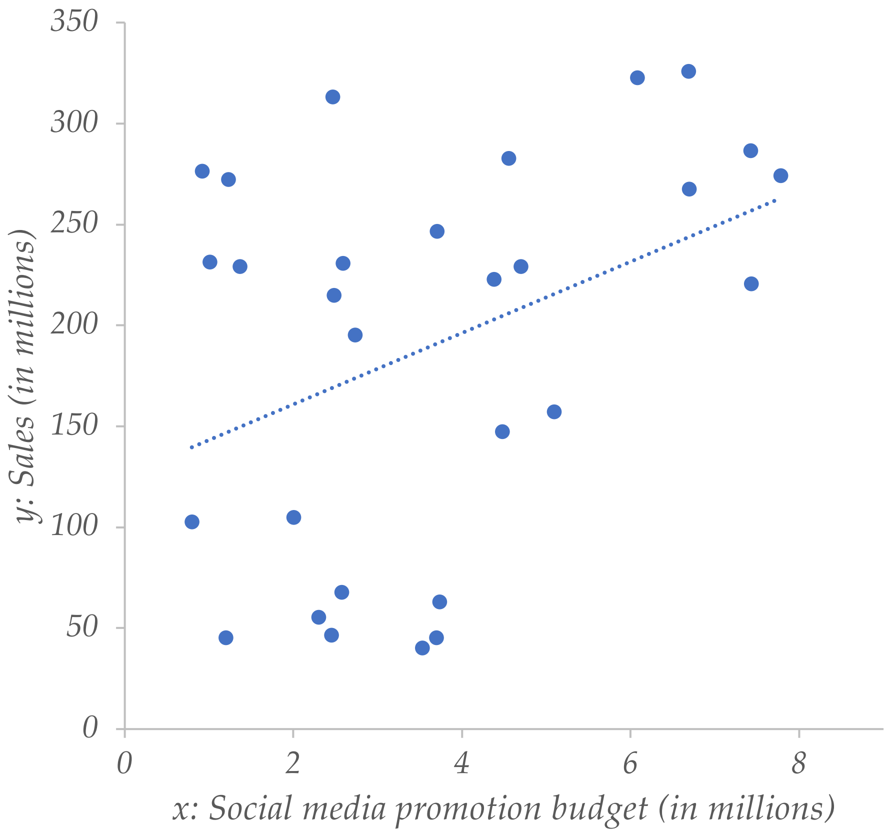
$$\sum_{n=1}^{N}(r_n)^2$$ is minimized

Goodness of the Fit



- What data says?

Goodness of the Fit



- What data says?
- R^2: Coefficient of Determination

R^2: Coefficient of Determination

Proportion of total variation of Y around m_Y which is explained by the regression

$$R^2 = \frac{\|\hat{Y} - m_Y\|^2}{\|Y - m_Y\|^2} \qquad \text{where} \quad m_Y = \frac{1}{N}\sum_{n=1}^{N} y_n$$

We measure how much of the variance in data is explained by the fit

$$R^2 = \frac{\text{variation(data)-variation(fit)}}{\text{variation(data)}}$$

We will show later that this is prune to overfitting

Ordinary Least Squares: Known issues so far

- Assumes linear relationship!
- If some features are co-linear, pseudo inverse is problematic: apply dimensionality reduction or regularization!
- Sensitive to outliers: use regularization

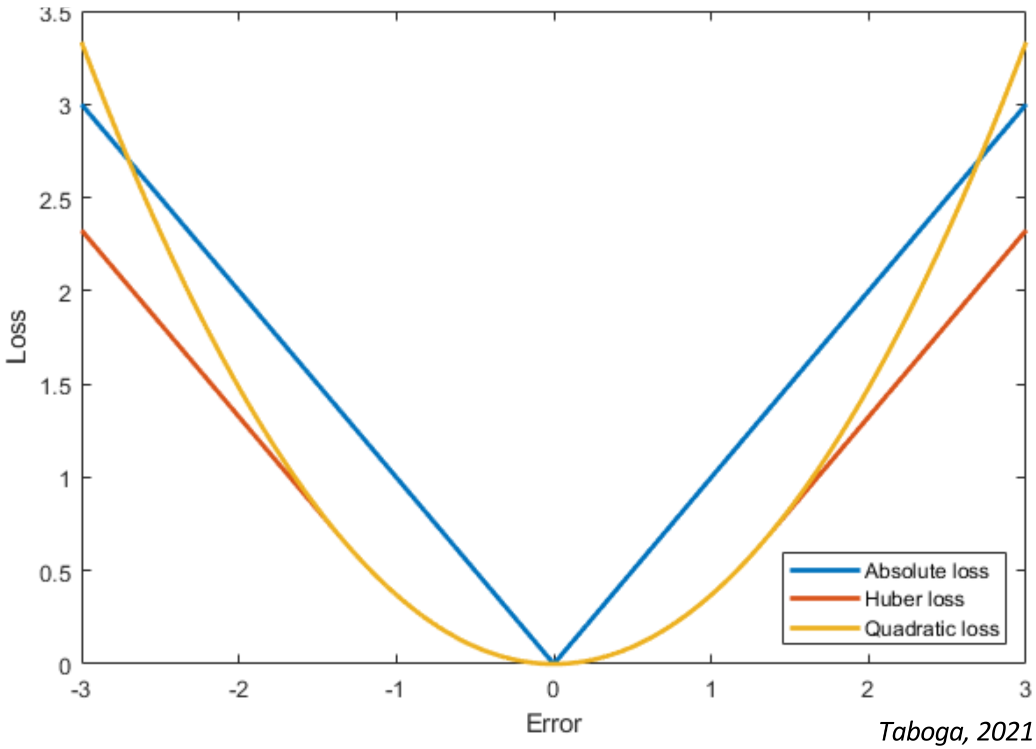
Other loss functions

Absolute loss: $|\text{Error}|$

Huber loss (Smooth absolute loss):

$$\begin{cases} \frac{1}{2}\text{Error}^2 & \text{if Error} < \delta \\ \delta(\text{Error} - \frac{\delta}{2}) & \text{otherwise} \end{cases}$$

Quadratic loss (OLS): Error^2



Taboga, 2021

How to deal with nonlinearity?

Nonlinear Feature Transformation

- Feature transformation example:

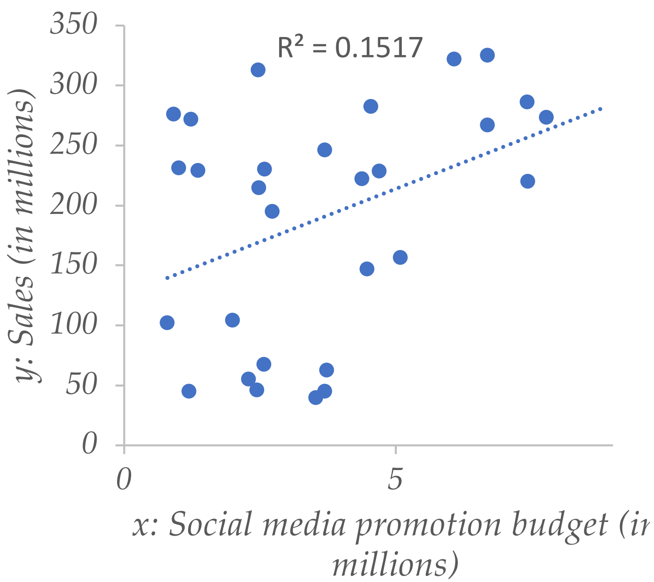
$$(x_1, x_2) \mapsto (z_1, z_2, z_3) = (x_1^2,\ \sqrt{2x_1}x_2,\ \sin(x_2))$$

- Determine a transformation $\phi(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}^D$ where you can model $y = \phi(x)\beta$
- Transform the features with non-linear basis functions:

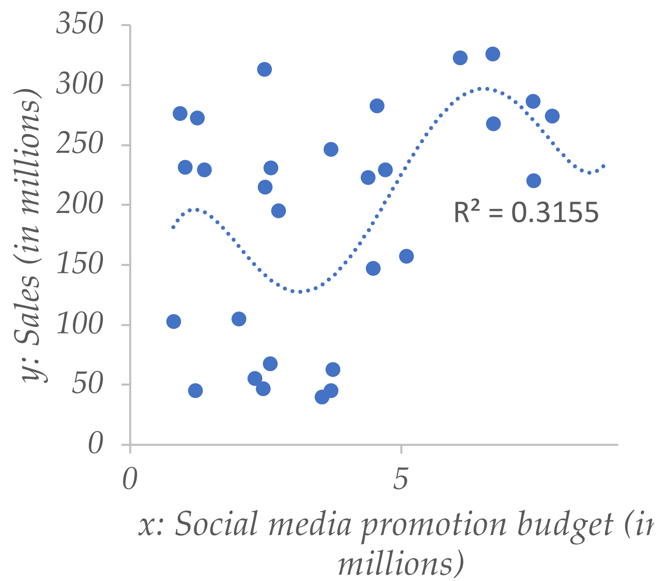
$$\phi(X) = \begin{pmatrix} z_{11} & z_{12} & z_{13} & \cdots & z_{1D} \\ z_{21} & z_{22} & z_{23} & \cdots & z_{2D} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ z_{N1} & z_{N2} & z_{N3} & \cdots & z_{ND} \end{pmatrix} \qquad z_{nj} = \phi_{i \mapsto j}(x_{ni})$$

Nonlinear Feature Transformation

Linear regression over the raw feature



Linear regression over the polynomial transformation of degree 5



- You can replace and apply all that we have seen today!
- Which concept becomes more important than before now?

A probabilistic treatment to Linear Regression

A Probabilistic Treatment to Linear Regression

- We now express the uncertainty on the prediction variable using a probability distribution.

- Assume a Gaussian curve fitting (preserving the linear relation between y and x):

$$p(y|x,\beta,\sigma) = \mathcal{N}(y|x^t\beta,\sigma^2)$$

- Note that, underlying this relation, we have:

$$y_n = x_n^T\beta + \epsilon_n \quad \text{where } \epsilon_n \text{ i.i.d with } \mathcal{N}(0,\sigma^2)$$

A Probabilistic Treatment

- Estimate parameters where $p(y|x,\beta,\sigma)$ is maximized:

$$\hat{\beta}_{ML} = \arg\max_{\beta} \prod_n p(y_n|x_n,\beta,\sigma^2)$$

$$\hat{\sigma}_{ML}{}^2 = \arg\max_{\sigma^2} \prod_n p(y_n|x_n,\beta,\sigma^2)$$

- This is achieved when

$$\hat{\beta}_{ML} = (X^T X)^{-1} X^T Y \qquad \hat{\sigma}^2_{ML} = \frac{1}{N}\sum_{n=1}^{N}(r_n)^2$$

You can derive these yourselves by maximizing the logarithm of $p(y|x,\beta,\sigma)$ over β and σ. Please also refer to Chapter 3.1.1 on Bishop's book for full derivation.

Have you noticed?

$$\hat{\beta}_{OLS} = \hat{\beta}_{ML}$$

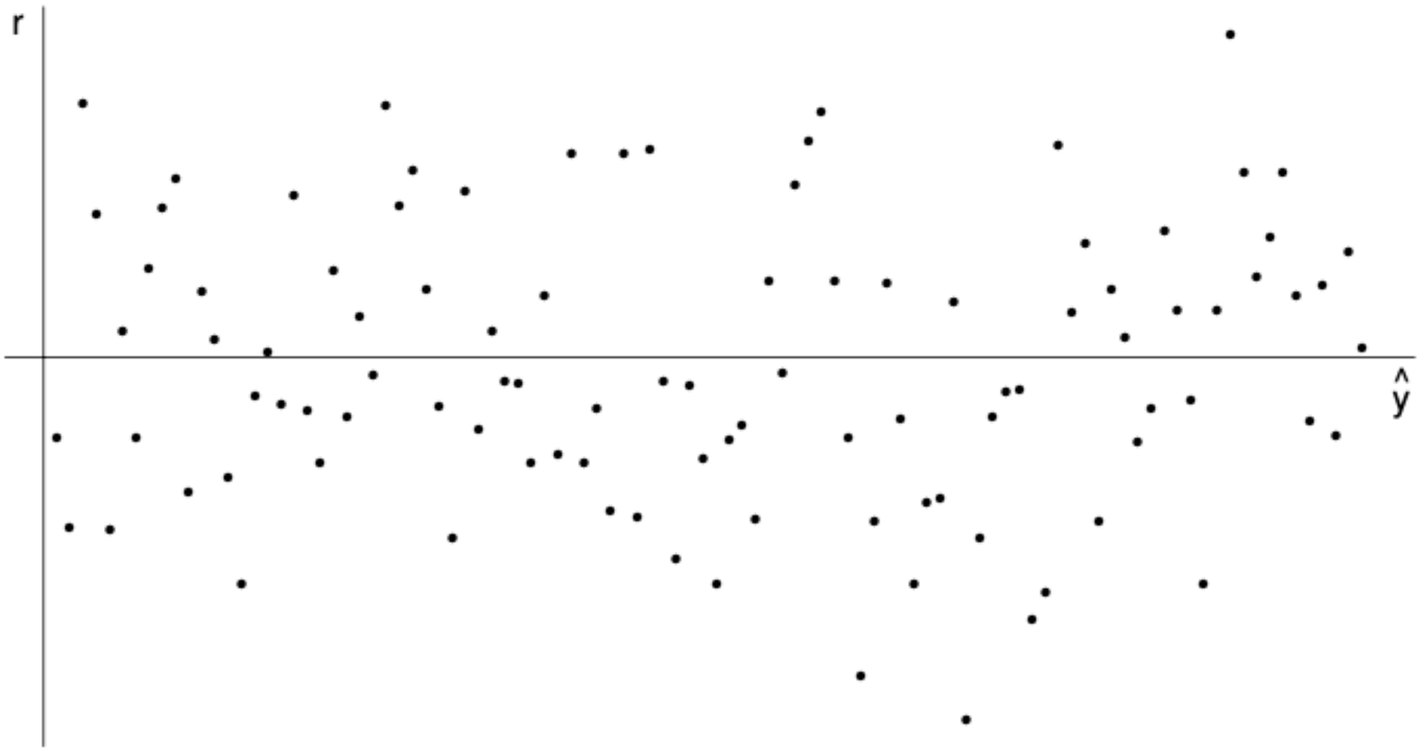
- Point estimate in MLE is equivalent to OLS for Normal error
- The two regressors are not entirely equivalent though!

Model Validation through Residual Analysis

- A brief note on R^2 *Look up: F-test, t-test in regresion, hypothesis testing* ***WEEK 7***
- The Tukey-Anscombe Plot
- The Normal Plot
- Mallows C_p statistic *Optional: Look up for it*

Tukey-Anscombe Plot

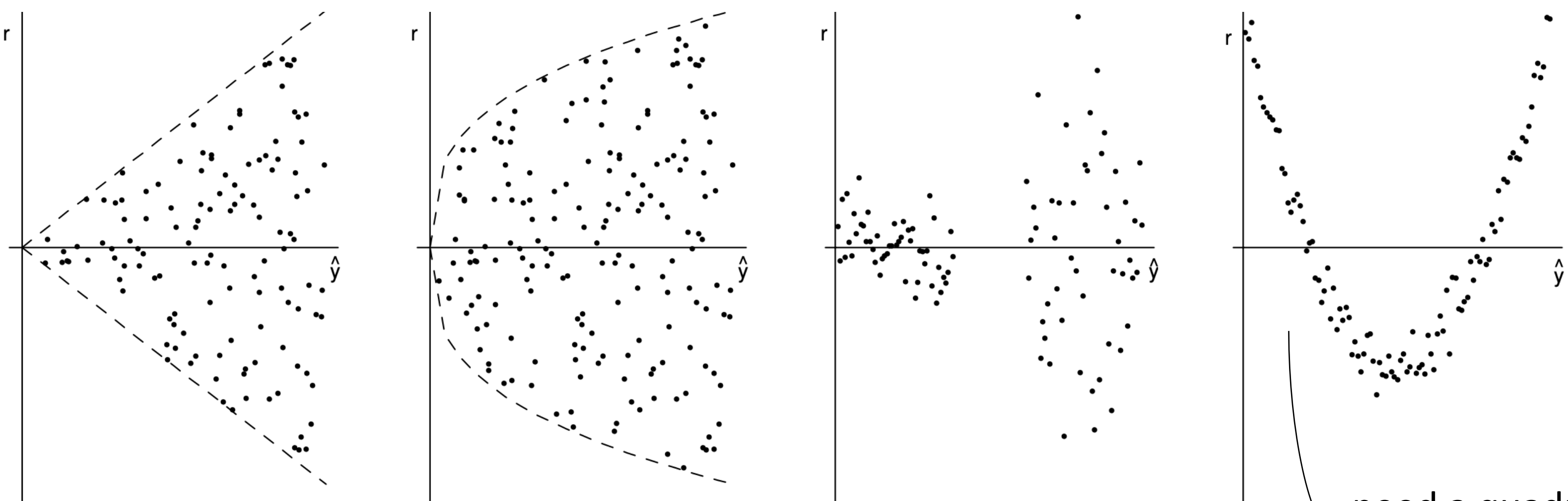
Plot the fitted values against the residual to observe lack of correlation



Mächler, 2022

Tukey-Anscombe Plot

Unwanted cases: transform your data!!

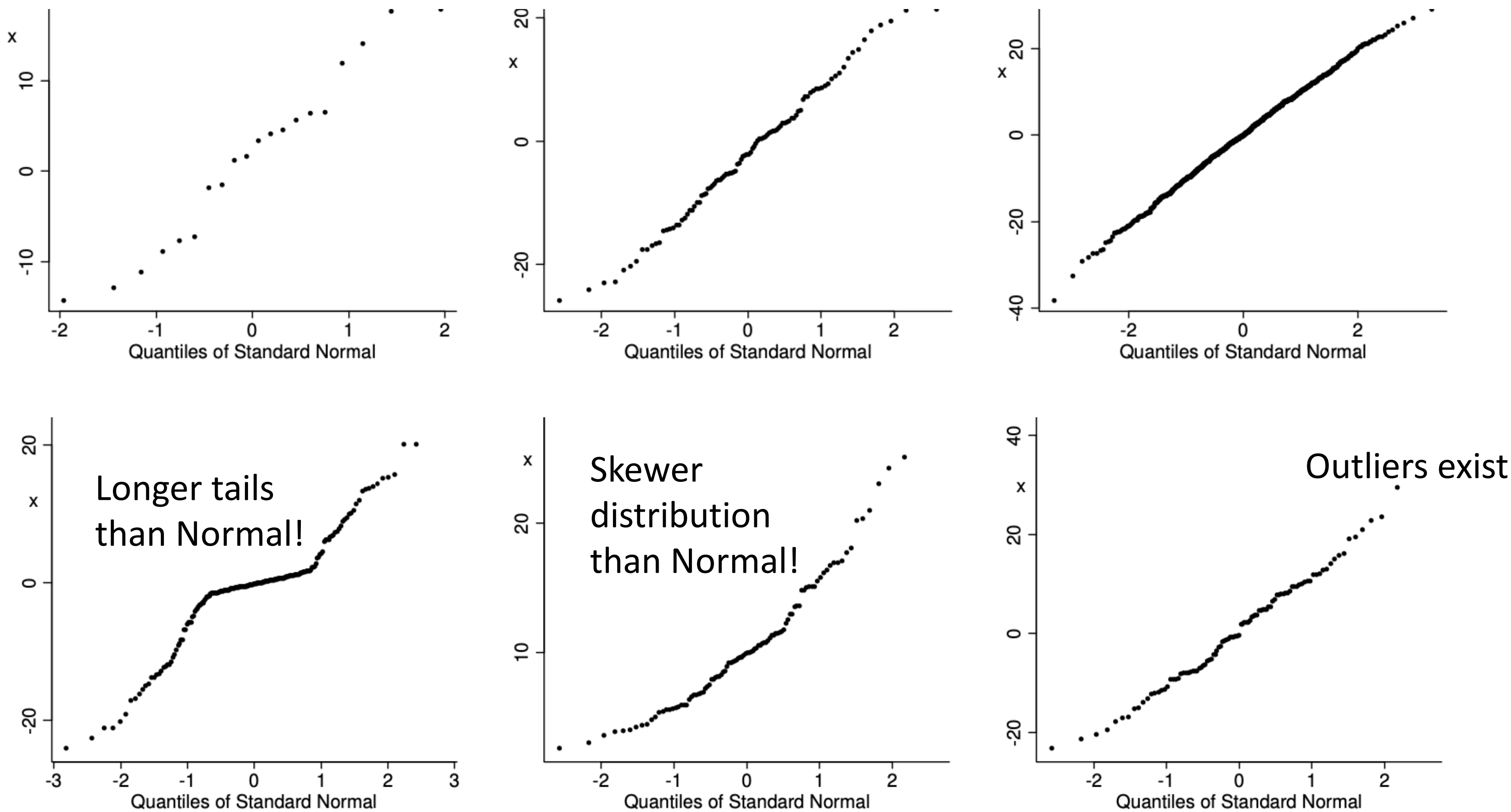


need a quadratic term in the model

Mächler, 2022

Q-Q Plot

First row: expected match between residuals and normal quantiles;
Second row: unwanted cases



Mächler, 2022

OLS Re-visited

Pros

- Simple and efficient
- Unbiased
  … and the best unbiased one
- Suitable for confidence intervals and hypothesis testing

Cons

- Assumes linear relationship!
- Assumes multicollinearity
- Sensitive to outliers: use regularization
- "Bestness" is under homoscedasticity assumption

Regularization to combat overfitting

Regularization: Ridge Regression

- Avoid overfitting to the observed data (especially when you don't have enough data!)
- Worsen your predictions by punishing your model:

$$\hat{\beta}^{OLS}_{Ridge} = \arg\min_{\beta} \sum_{n=1}^{N} (y_n - x_n^T \beta)^2 + \lambda \beta^T \beta$$

$$= \arg\min_{\beta} (Y - X\beta)^T (Y - X\beta) + \lambda \beta^T \beta$$

- Ridge regression has a closed form solution:

$$\hat{\beta}^{OLS}_{Ridge} = (\lambda \mathbf{I} + X^T X)^{-1} X^T Y$$

Regularization: Lasso Regression

Tibshirani, 1996

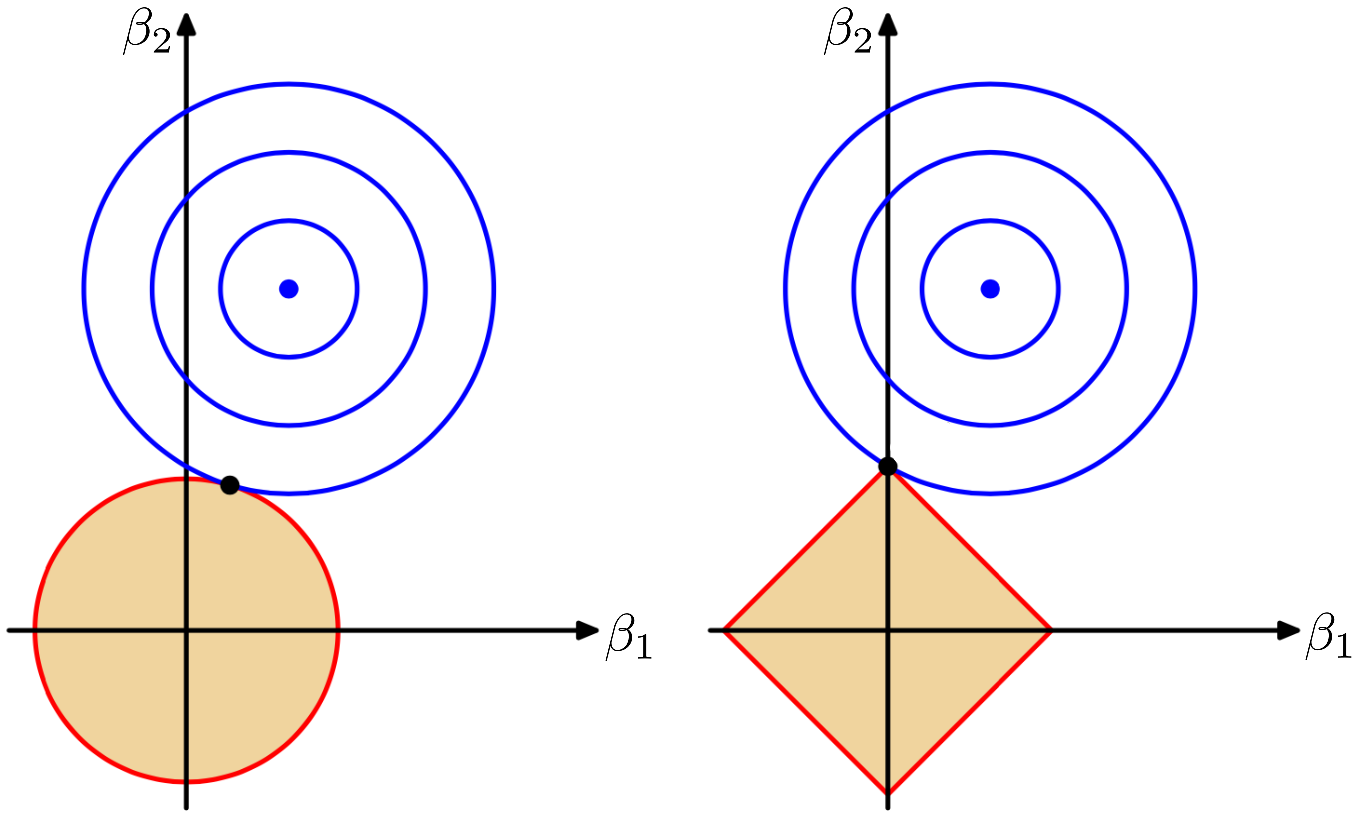
- This time impose sparsity to your parameter space

$$\hat{\beta}^{OLS}_{Lasso} = \arg\min_{\beta} \sum_{n=1}^{N} (y_n - x_n^T \beta)^2 + \lambda \|\beta\|_1$$
$$= \arg\min_{\beta} (Y - X\beta)^T (Y - X\beta) + \lambda \|\beta\|_1$$

- Solution to Lasso: No closed form solution!
  (Check out Subgradient Descent, Coordinate Descent)

Ridge vs. Lasso Regression

Lasso can handle redundant parameters better (if there's any)



Bishop, 2006

A probabilistic Treatment to (Linear Regression) Regularization

Bayesian Linear Regression

- Earlier, we focused on the conditional likelihood

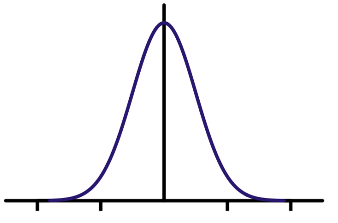
$$p(y|\beta, x) = \mathcal{N}(y|x\beta, \sigma_\epsilon^2)$$

following our linear model: $y = x\beta + \epsilon$

- This time, using this linear model, we will apply a "Bayesian treatment" to the problem and focus on

$$p(\beta|x, y)$$

which we assume to be Gaussian



Maximum a Posteriori (MAP) Estimation

- We maximized $p(y|\beta, x)$ using Maximum (Conditional) Likelihood (MLE) estimation
- Now we want to maximize the posterior $p(\beta|x, y)$
- This is called Maximum a Posteriori Estimation

$$\hat{\beta}_{MAP} = \arg\max_{\beta} p(\beta|x, y)$$
$$= \arg\max_{\beta} \underbrace{\boxed{p(y|\beta, x)}}_{\text{likelihood}} \underbrace{\boxed{p(\beta)}}_{\text{prior}}$$

- MAP estimate imposes a cost on the model parameters too!

MAP Estimation for Bayesian Linear Regression

- Recall our model $p(y|\beta, x) = \mathcal{N}(y|x\beta, \sigma_\epsilon^2)$

  (suppose that we know the noise variance and it's no longer a parameter)

- The conjugate prior of the parameters is also Gaussian $p(\beta) = \mathcal{N}(\beta|0, \sigma_\beta^2)$

- The MAP Estimate is obtained via

$$\hat{\beta}_{MAP} = \arg\max_\beta p(\beta|x, y) = \arg\max_\beta p(y|\beta, x)p(\beta)$$

- Which is given by $\hat{\beta}_{MAP} = \left(\frac{\sigma_\epsilon^2}{\sigma_\beta^2}\mathbf{I} + X^T X \right)^{-1} X^T Y$

Derive this at home!

Goals for Today Re-visited

At the end of this class, you should be able to

- ☑ Identify regression problems
- ☑ For linear regression (with or without feature transformation), write out
  - ☑ Least Squares solution and its geometric interpretation
  - ☑ Maximum Likelihood Estimation (MLE) and Maximum a Posteriori (MAP) estimation for Gaussian models
- ☑ Apply regularization
  - ☑ Ridge, Lasso Regression

Wrap-up

- Linear Regression where feature transformation is possible
- Least squares have similarities with Normal error model
- Statistical validation methods for OLS
- Regularization helps with overfitting and multicollinearity
- MAP can act like a regularization

Reading and References

Chapter 1.2.5-1.26, 3.1 and 3.3 on Bishop, 2006

References

- Taboga, Marco. "Loss function", Lectures on probability theory and mathematical statistics. Kindle Direct Publishing. Online appendix. https://www.statlect.com/glossary/loss-function, 2021
- Bishop, Christopher M., and Nasser M. Nasrabadi. “*Pattern recognition and machine learning*”. Vol. 4. No. 4. New York: springer, 2006.
- Mächler, Martin. “Computational Statistics”, Lecture notes. https://stat.ethz.ch/lectures/ss21/comp-stats.php 2022.