Lecture 8: An intro to regression models

Dr. Greg Chism

Objectives

- Intro to regression analysis
- Linear regression
- Regularized regression
- Non-linear regression & ensemble models
- Advanced regression techniques
- Robust regression
- Quantile regression
- Model selection

Objectives

- **Intro to regression analysis**
- Linear regression
- Regularized regression
- Non-linear regression & ensemble models
- Advanced regression techniques
- Robust regression
- Quantile regression
- Model selection

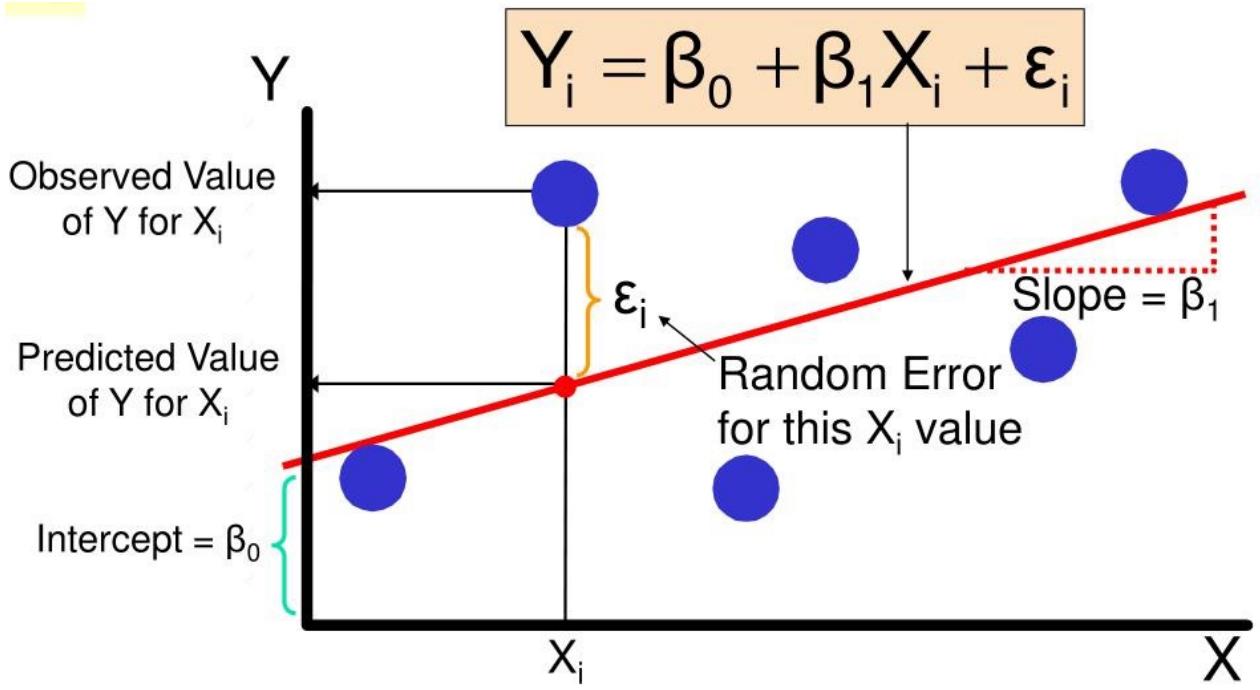
Machine Learning

- **Supervised:** We are given input samples (X) and output samples (y) of a function $y = f(X)$. We would like to "learn" *f* and evaluate it on new data. Types:
  - Classification: y is discrete (class labels).
  - **Regression:** y is continuous, e.g., linear regression

What are regression models?

Set of methods that are used to predict a response variable from one or more predictor variables

Key terms: dependent and independent variables!



Simple linear model

Let's discuss its mathematical structure:

$$\widehat{Y_i} = \widehat{\beta_0} + \widehat{\beta_1} X_{li} + \cdots + \widehat{\beta_k} X_{ki} \quad i = 1 \ldots n$$

$\hat{Y}_i$ is the predicted value of the dependent variable for observation i

$\widehat{\beta_0}$ is the intercept

$\widehat{\beta_k}$ is the regression coefficient for the jth predictor

n is the number of observations

k is the number of predictor variables

Simple linear model (model parameters)

We are going to focus on four main regression parameters

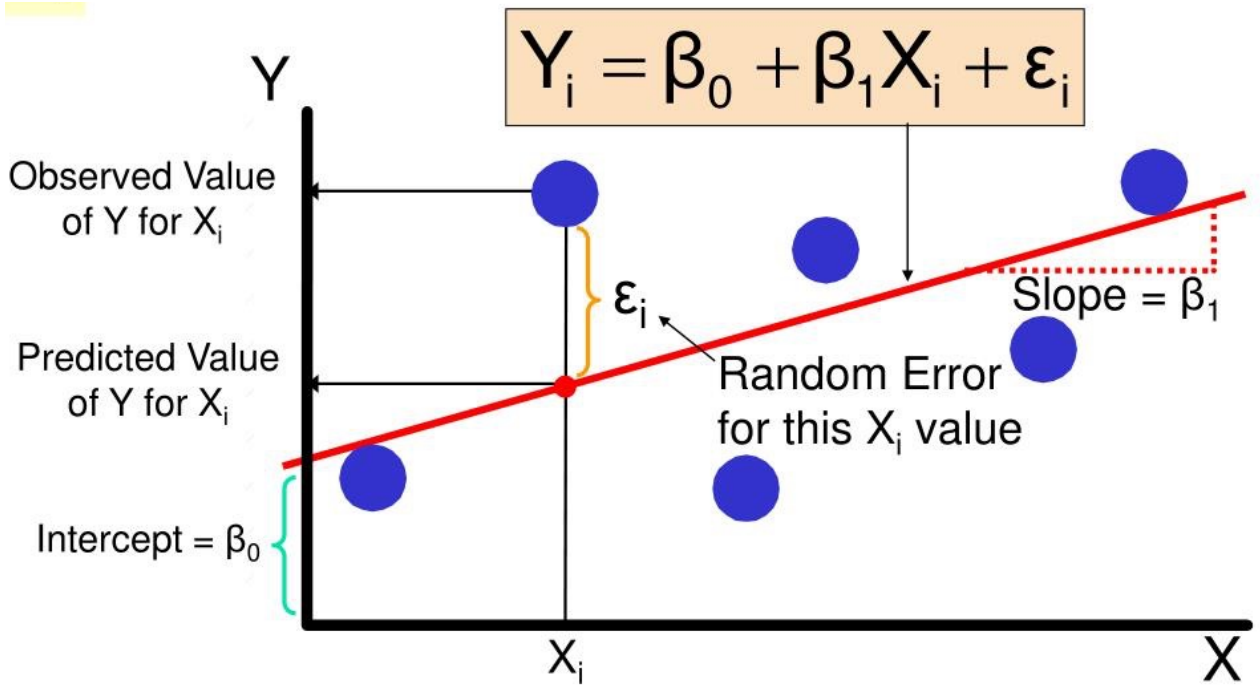
Slope
Intercept
R^2
P-values

We will review these concepts using a simple linear regression

Activity 4. Interpreting simple regression models?

Simple linear model (model parameters)

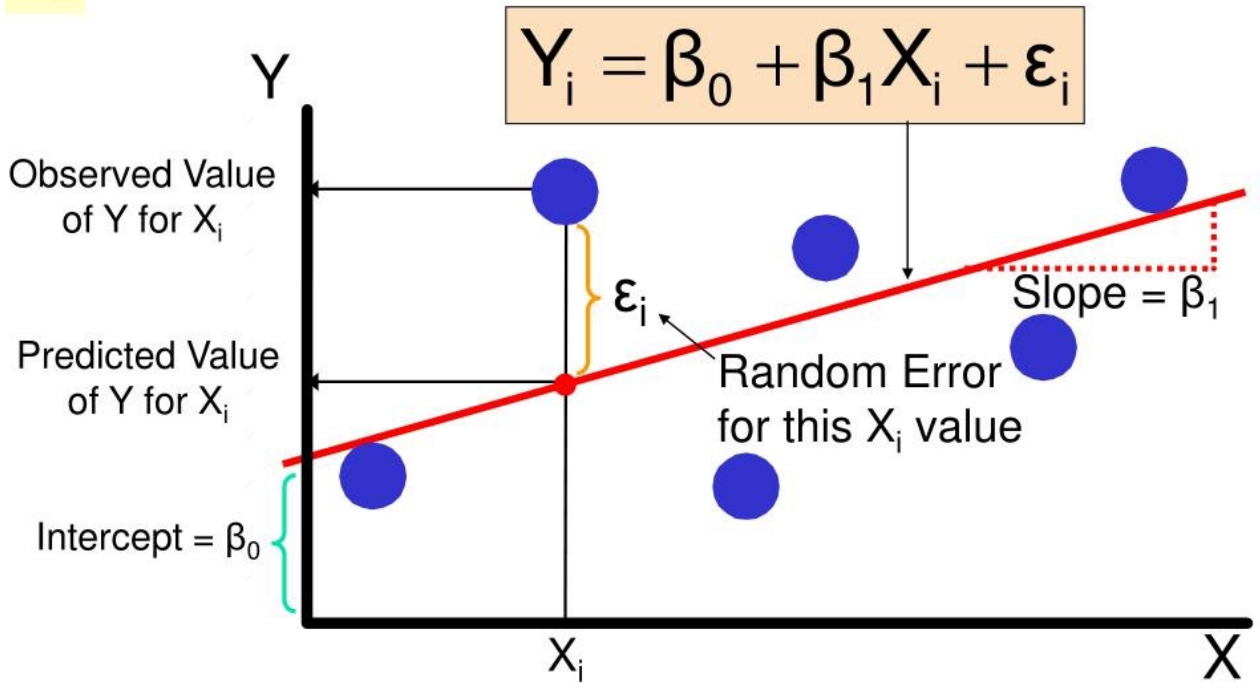
- **Slope:** Change in Y units given a change in X of a single unit.



Activity 4. Interpreting simple regression models?

Simple linear model (model parameters)

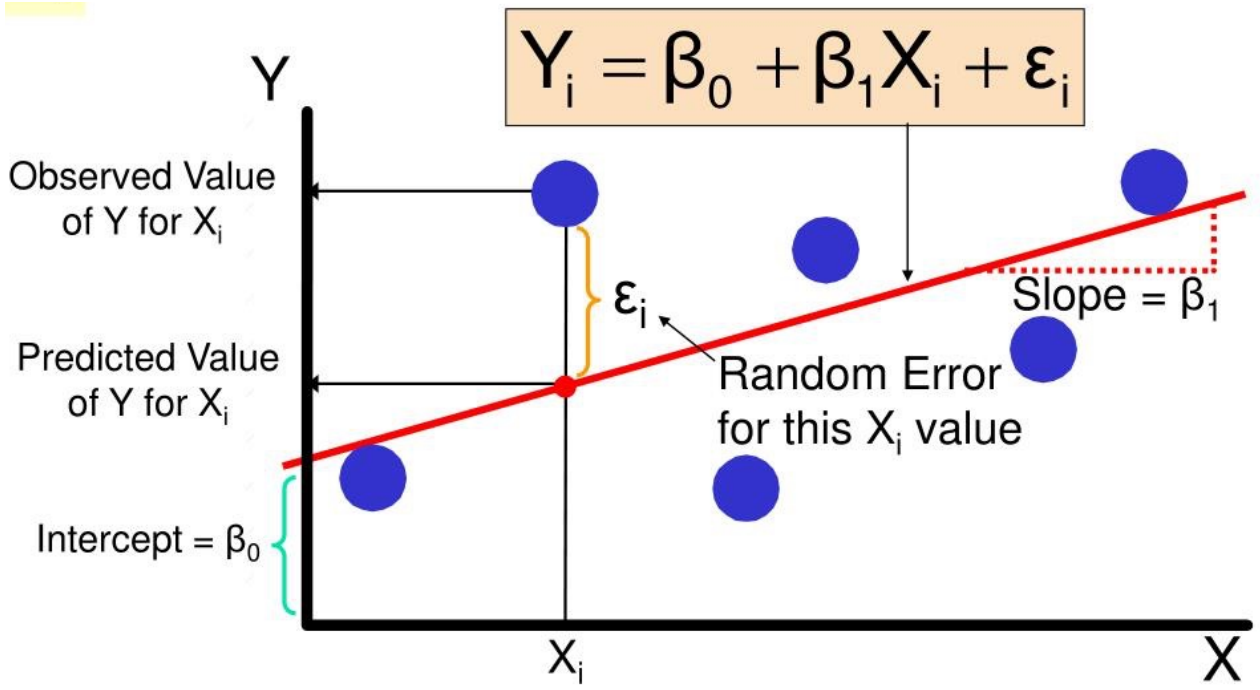
- **Intercept**: Adjustment constant. The Y value when X, the predictor is 0.



Activity 4. Interpreting simple regression models?

Simple linear model (model parameters)

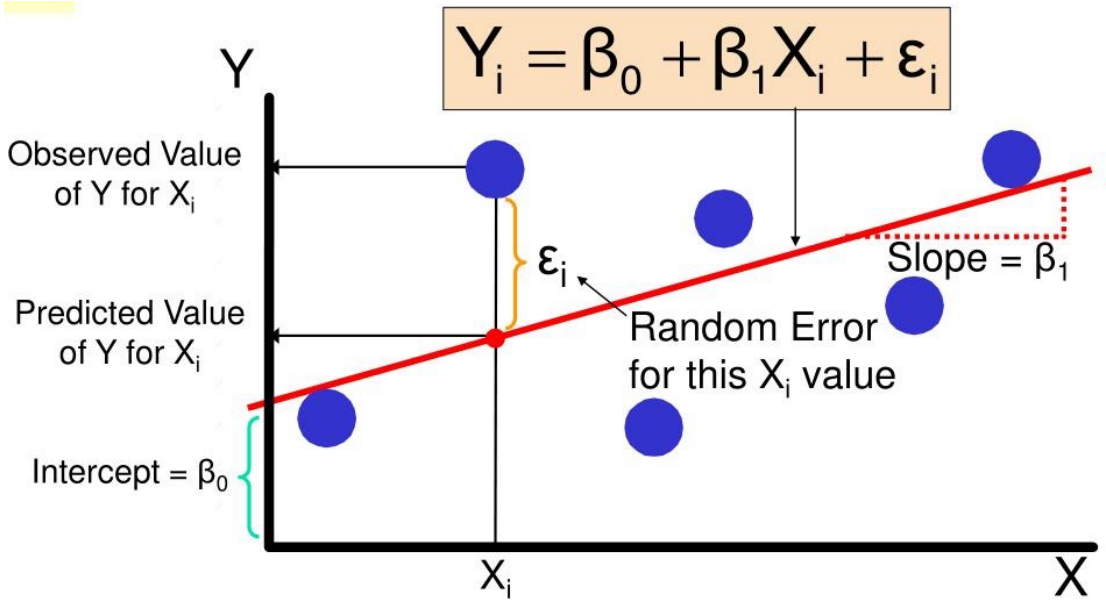
- **R^2:** The fraction of variance in Y that is explained by X.



Activity 4. Interpreting simple regression models?

Simple linear model (model parameters)

- **P-values:** Indicates whether the slope is significantly different from zero.



Activity 4. Interpreting simple regression models?

What are regressions used for?

Today we are going to review specific examples of how to use regression models.

- For now, remember that regressions can be used to:
  identify explanatory variables
  describe the form of the relationships involved
  predicting the response variable from the explanatory variables

Types of regression models?

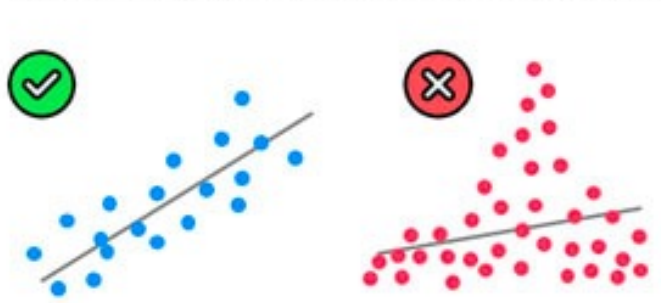
- **Linear Regression:** Predicting a quantitative response variable from a quantitative explanatory variable.
- **Polynomial:** Predicting a quantitative response variable from a quantitative explanatory variable, where the relationship is modeled as an *n*th order polynomial.
- **Ridge Rregression:** adds L2 regularization to linear regression, handles collinearity, prevents overfitting
- **Lasso Regression:** L1 regularization, known for feature selection, pushes less important variables to zero
- **Elastic Net Regression:** Combines L2 and L1 regularization, balancing Ridge and Lasso regression techniques
- **Multiple linear:** Predicting a quantitative response variable from two or more explanatory variables.
- **Multilevel:** Predicting a response variable from data that have a hierarchical structure

Objectives

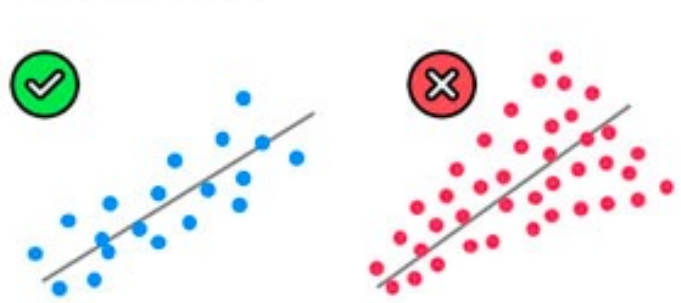
- Intro to regression analysis
- Linear regression
- **Regularized regression**
- Non-linear regression & ensemble models
- Advanced regression techniques
- Robust regression
- Quantile regression
- Model selection

Assumptions of Linear Regression

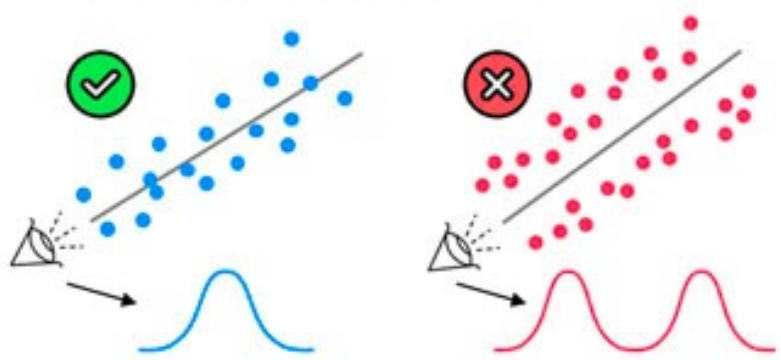
Linearity
(Linear relationship between Y ~ X)



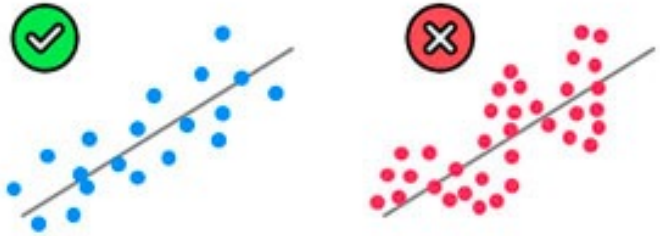
Homoscedasticity
(Equal variance)



Multivariate Normality
(Normally distributed residuals)



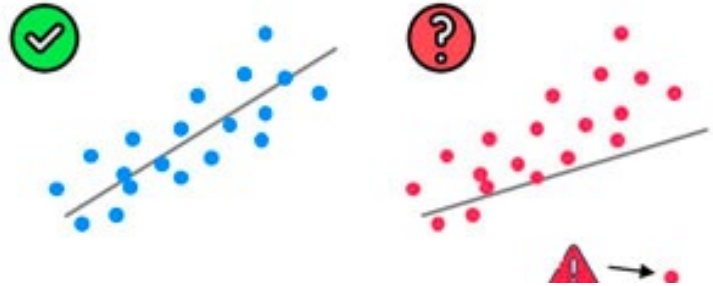
Independence
(of observations)



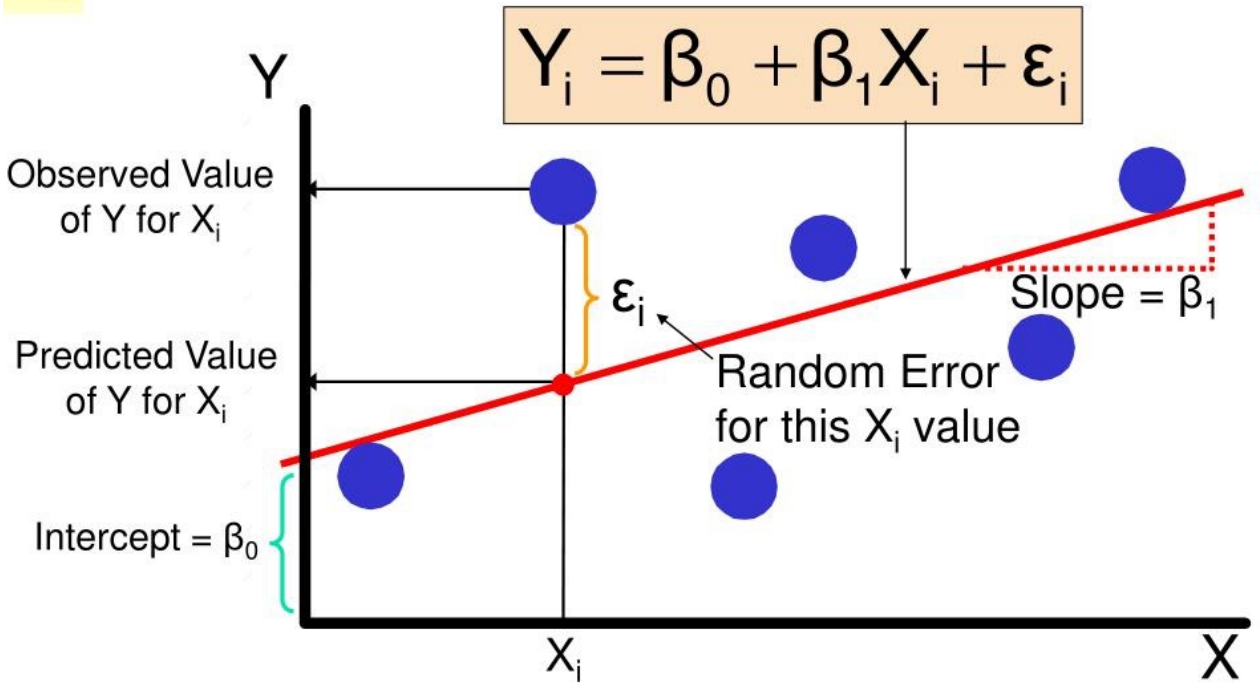
Lack of Multicollinearity
(Predictors are not correlates)

$X_1 \not\sim X_2$ $X_1 \sim X_2$

Outlier check
(Technically not an assumption)

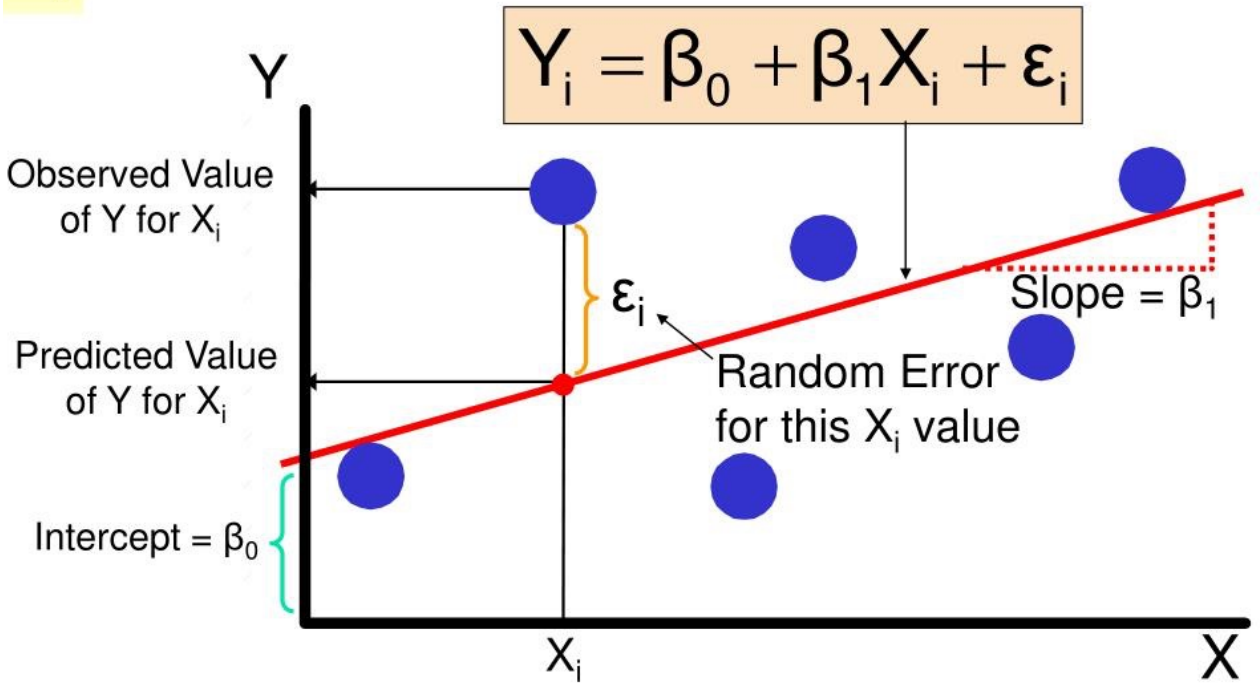


Ordinary Least Squares (OLS)



- Estimate coefficients of linear equation
- Finds the line that minimizes sum of square residuals

Model evaluation



$$R^2 = 1 - \frac{SS_{Regression}}{SS_{Total}}$$

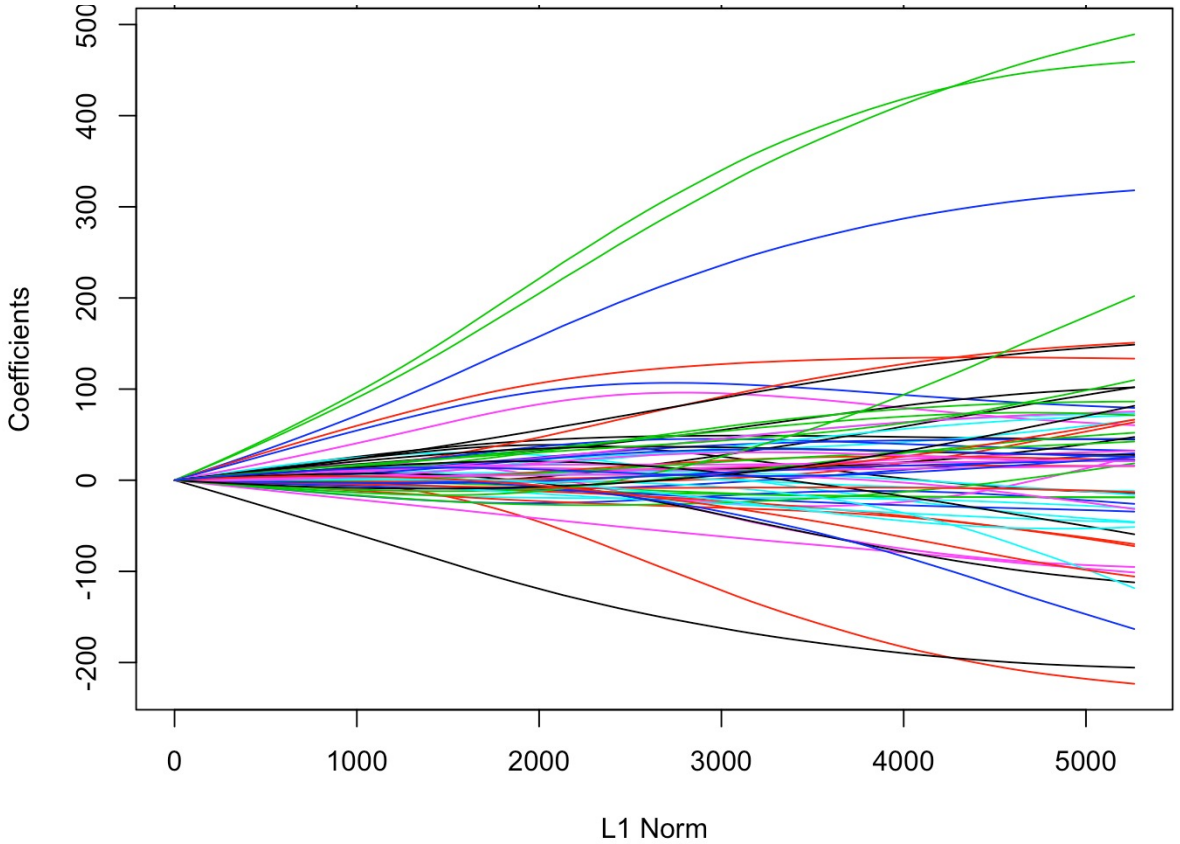
$$MSE = \frac{1}{n} \Sigma \left(y - \hat{y} \right)^2$$

Objectives

- Intro to regression analysis
- Linear regression
- **Regularized regression**
- Non-linear regression & ensemble models
- Advanced regression techniques
- Robust regression
- Quantile regression
- Model selection

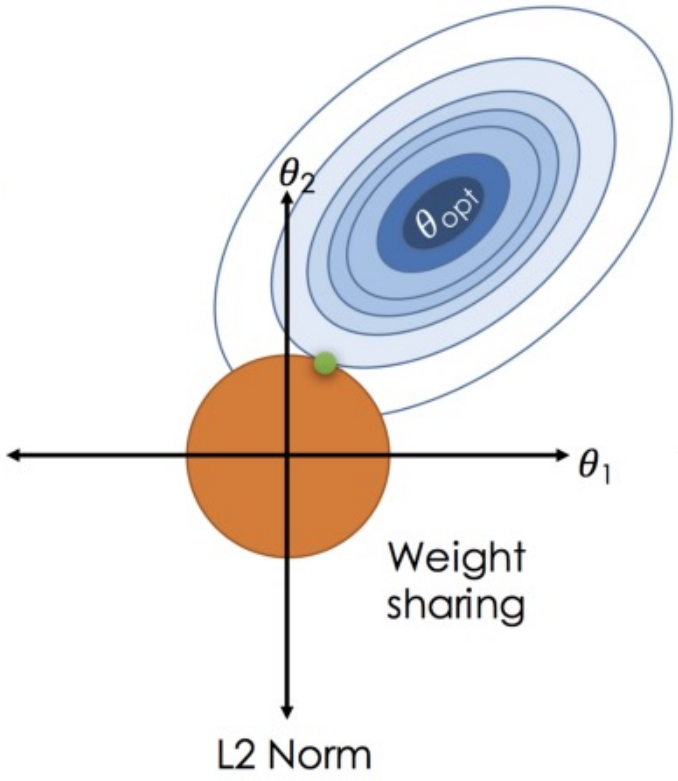
Ridge Regression

Extension of linear regression with L2 regularization to address issues like multicollinearity



L2 Regularization

Penalizes large coefficients, controlled by a hyperparameter λ (lambda)

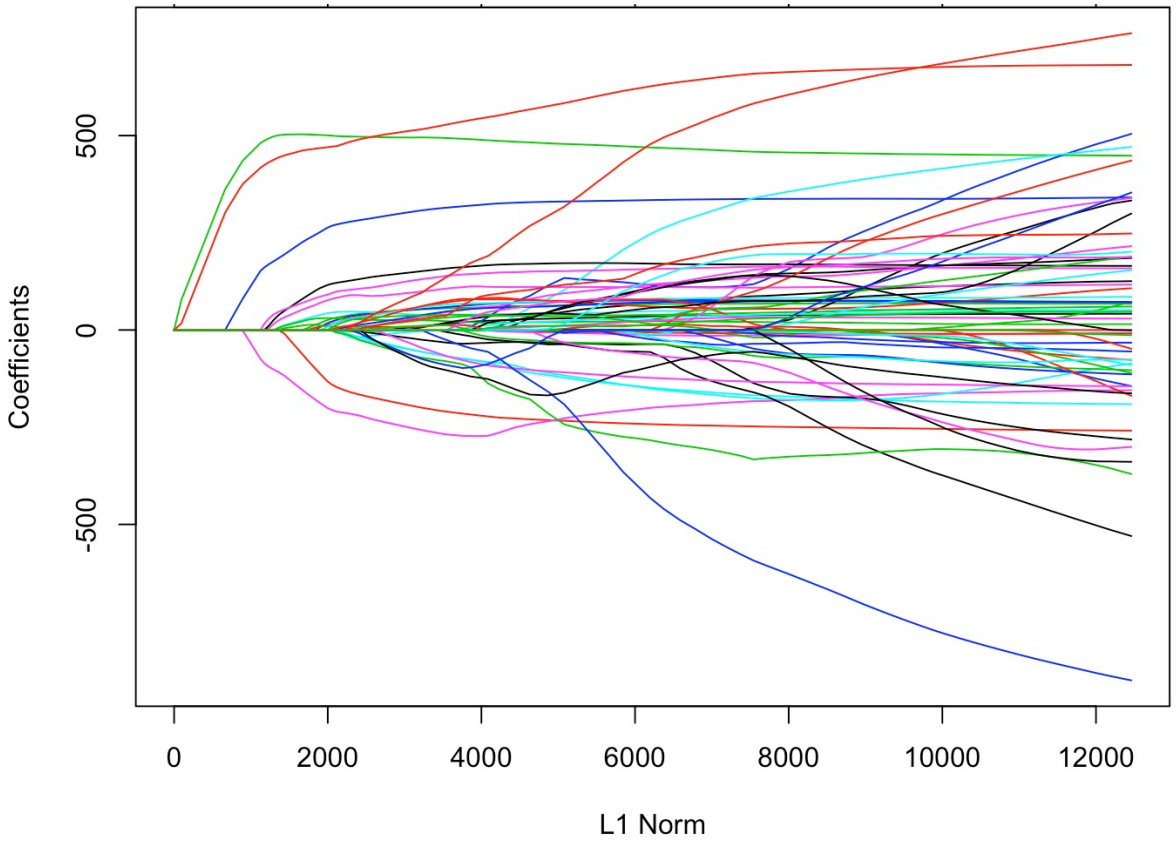


- Stabilizes coefficient estimates, making them less sensitive
- Reduces the magnitude of OLS coefficients, but doesn't set them to 0

$$\lambda \sum_{j=1}^{p} \beta_j^2$$

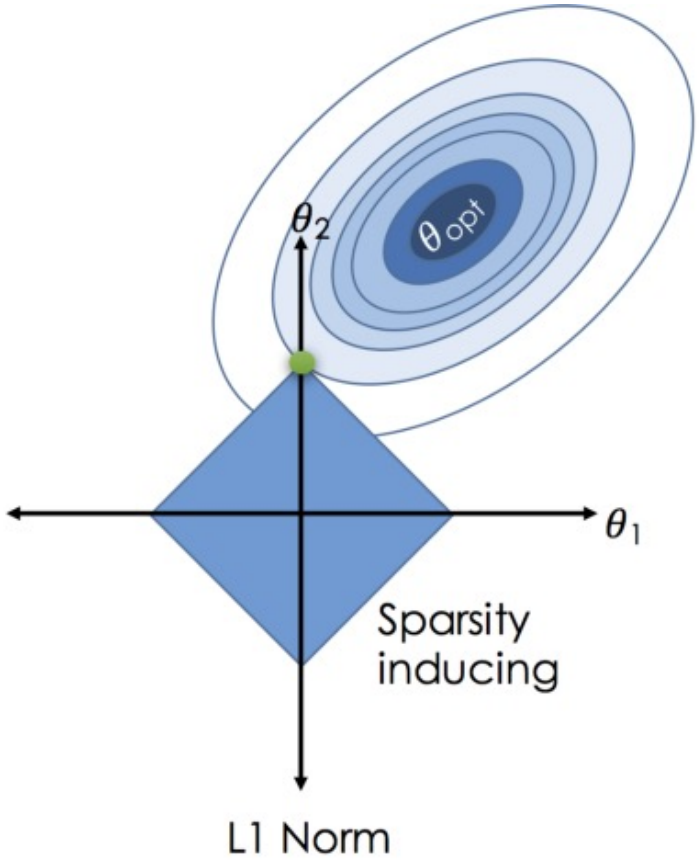
Lasso Regression

Least Absolute Shrinkage and Select Operator: uses L1 regularization penalty



L1 Regularization

Adds a penalty term to OLS regression, encouraging sparsity by pushing some coefficients to 0

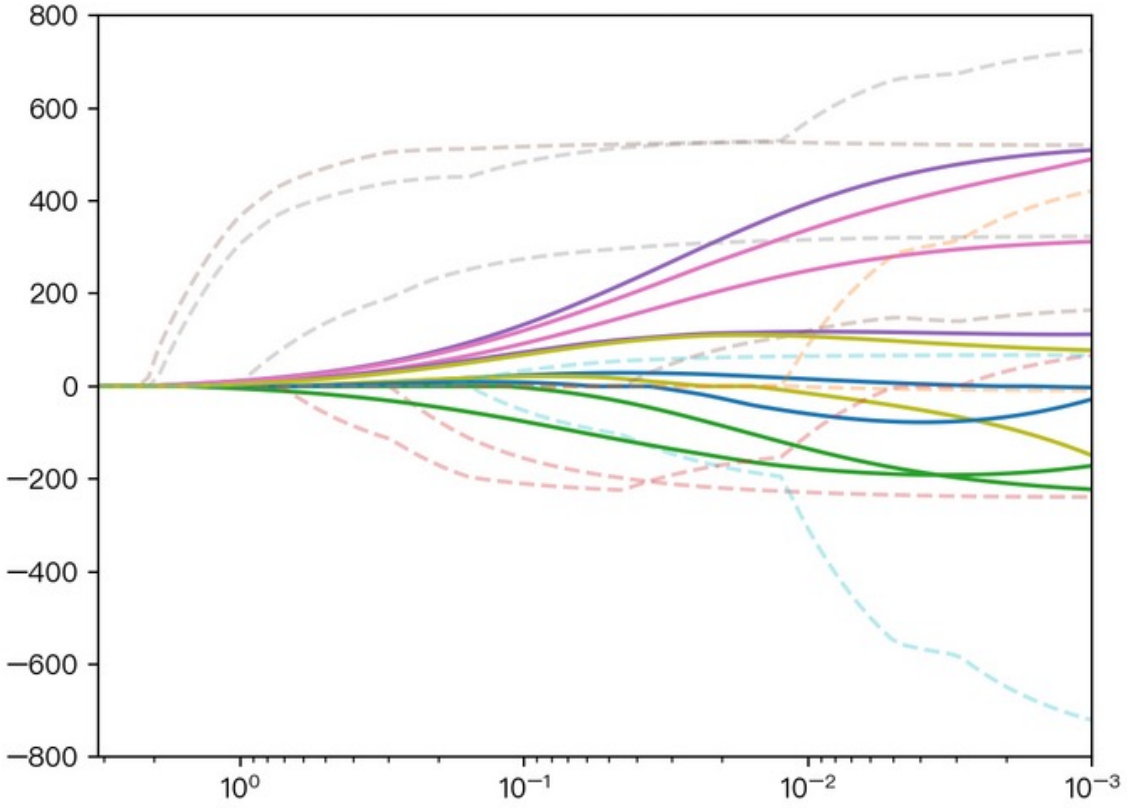


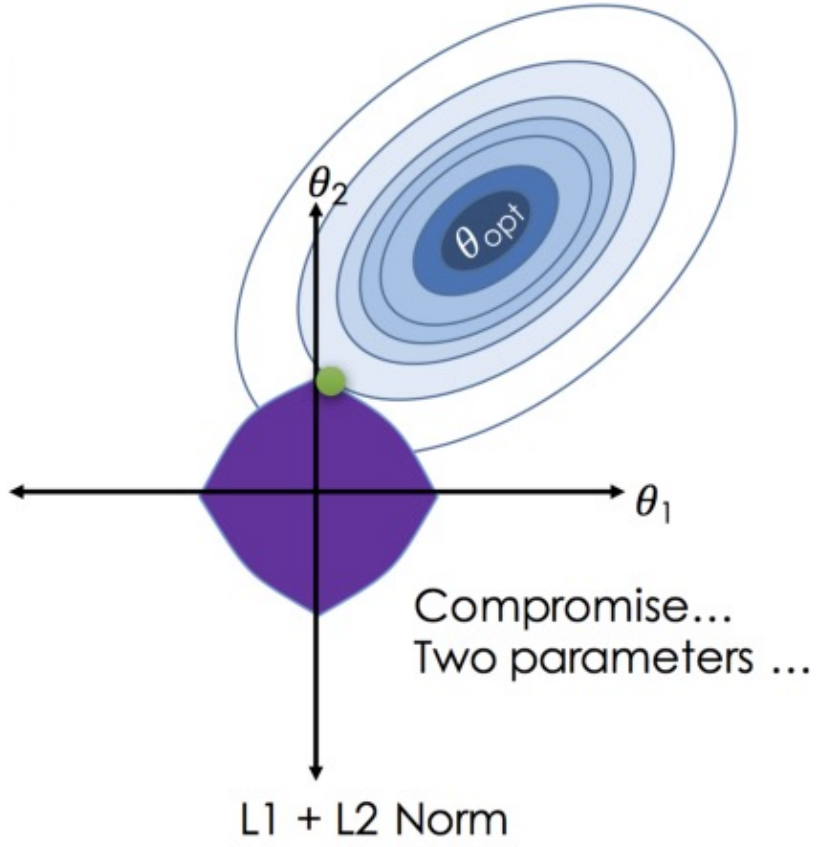
- Eliminated less important variables
- Sensitive to outliers

$$\lambda \sum_{j=1}^{p} |\beta_j|$$

Elastic Net Regression

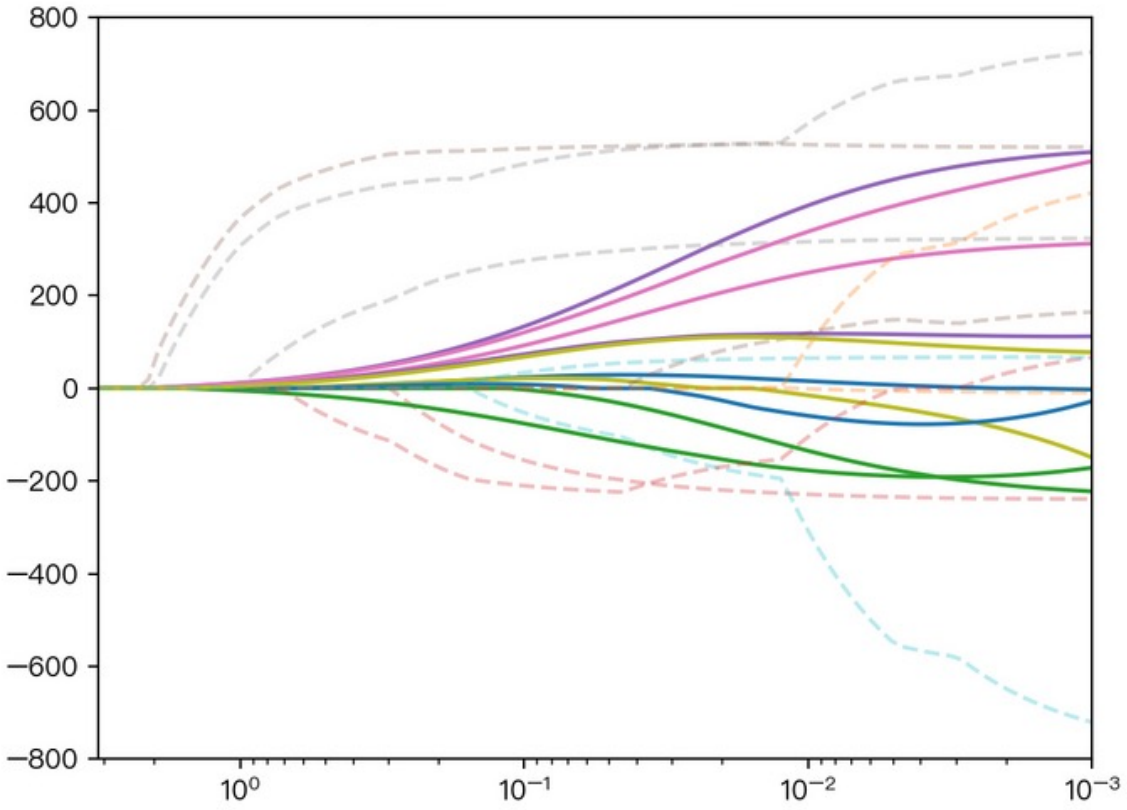
Combines L2 (Ridge) and L1 (Lasso) regularization techniques (balance)





Elastic Net Regression

Combines L2 (Ridge) and L1 (Lasso) regularization techniques (balance)



- Robust and flexible
- Feature selection & multicollinearity
- Must balance two hyperparameters (randomization)

Use cases

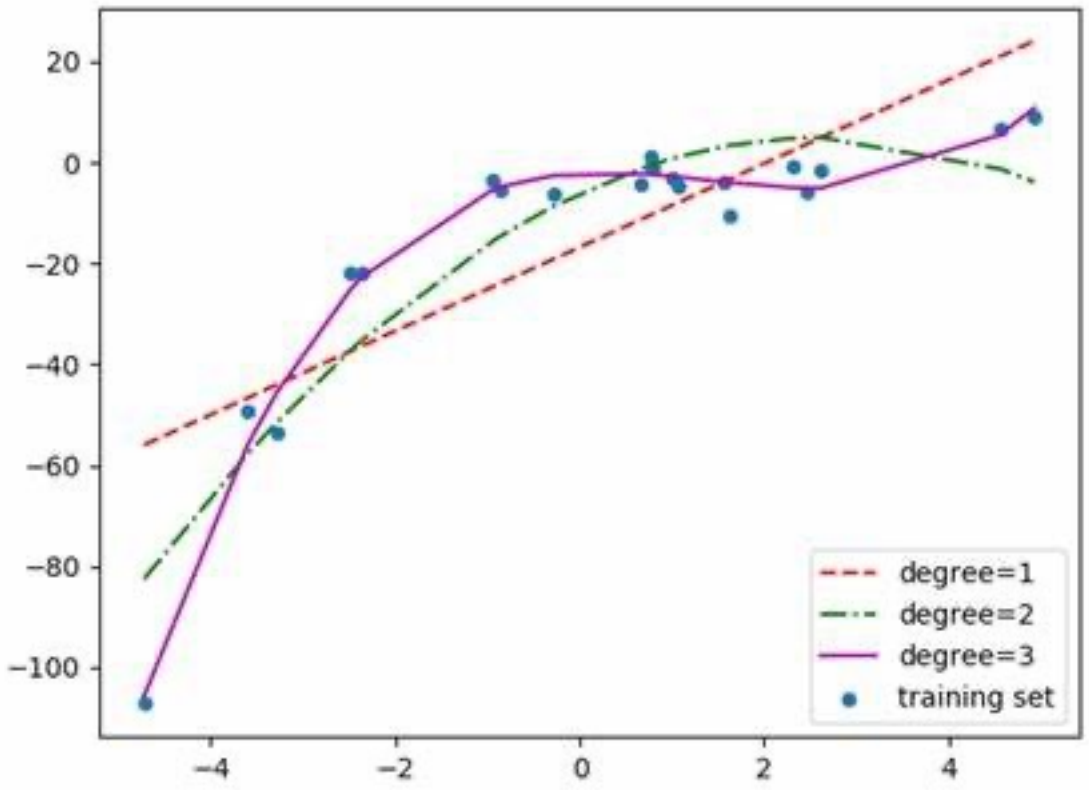
- **Risk Assessment**: Predicting credit risk or financial market volatility.
- **Marketing Spend Optimization**: Allocating marketing resources to maximize return on investment.
- **Healthcare Cost Prediction**: Estimating medical treatment costs based on patient characteristics.
- **Customer Churn Prediction**: Identifying factors influencing customer churn in a subscription-based service.

Objectives

- Intro to regression analysis
- Linear regression
- Regularized regression
- **Non-linear regression & ensemble models**
- Advanced regression techniques
- Robust regression
- Quantile regression
- Model selection

Polynomial Regression

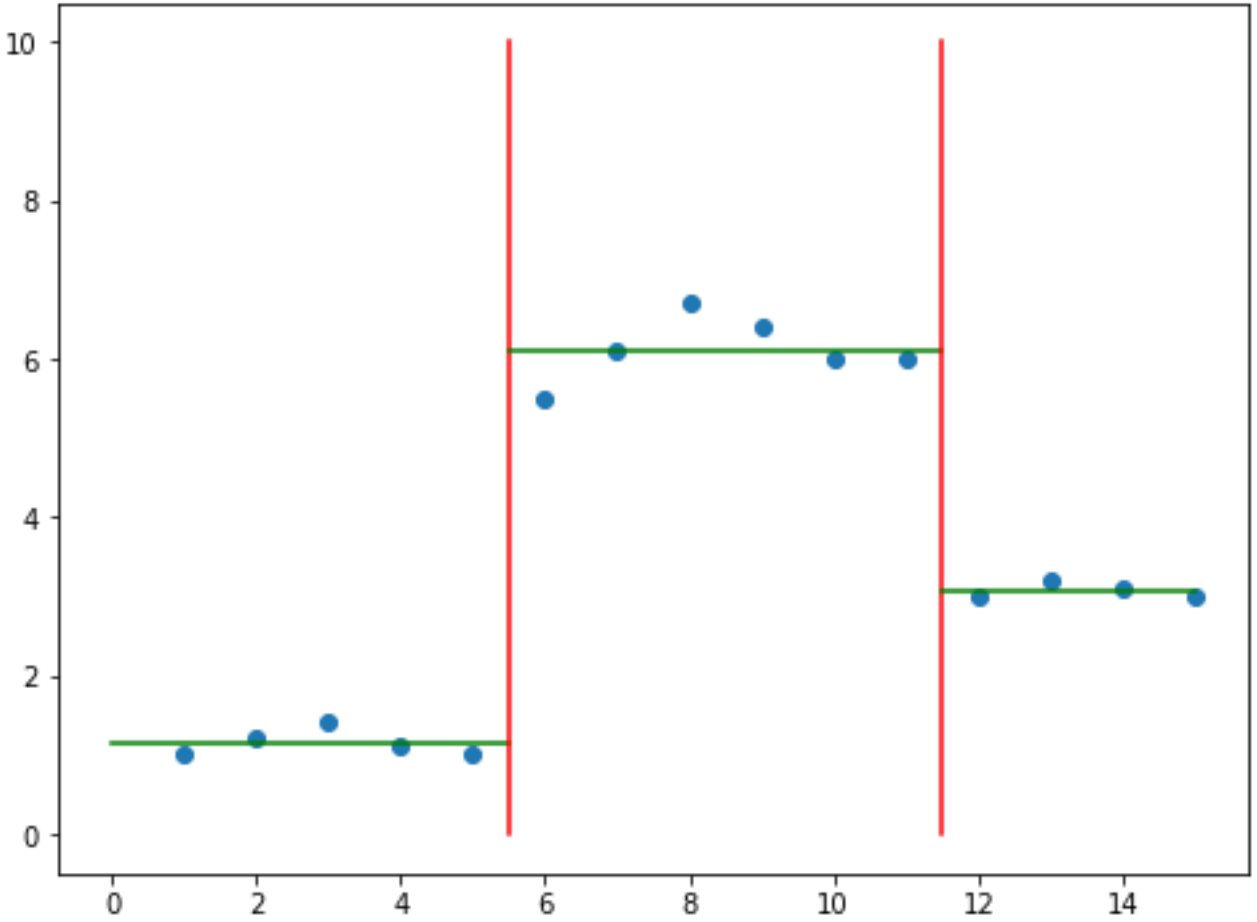
Extends linear regression to capture non-linear relationships



- Models curved relationships (squared, cubed, etc.)
- High-degree polynomials can overfit
- Need to balance complexity and overfitting

Decision Trees

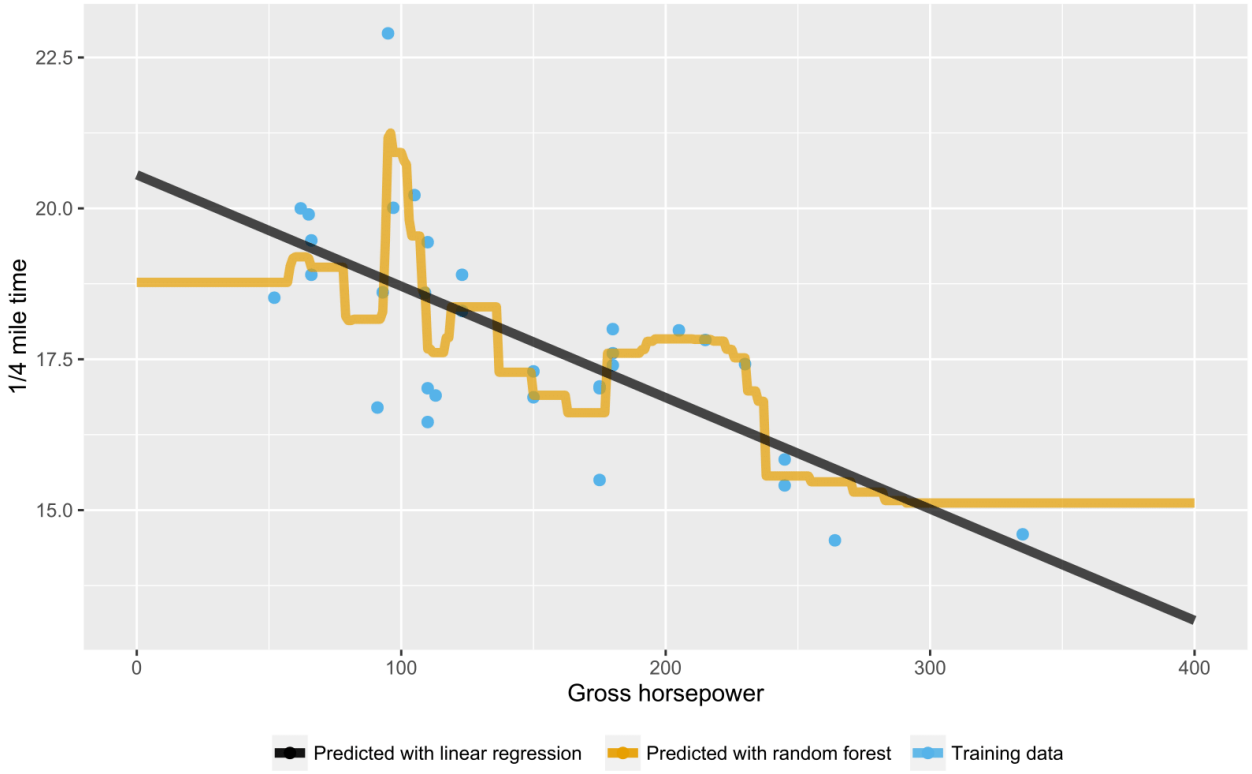
Provides tree-like structure to recursively partition the data



- Partition data based on independent variables, minimizing variance
- Trees can be overly complex (hence pruning)
- Sensitive to small changes in data -> different trees

Ensemble Methods (Random Forest)

Combines multiple decision trees for increased accuracy



- Bootstrap samples (bagging) of data, selects random subsets of features for each tree
- Known for robustness, ability to handle high-dimensional data, resistance to outliers

Use cases

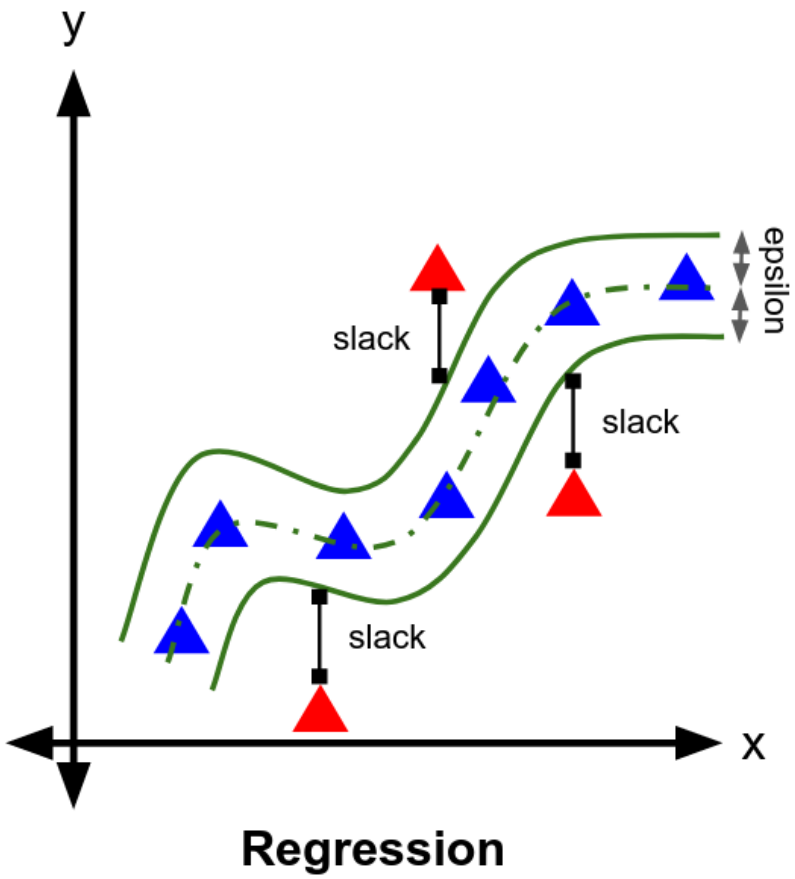
- **Stock Price Prediction:** Capturing complex stock price movements that aren't linear.
- **Customer Lifetime Value:** Estimating the long-term value of customers for business strategy.
- **Demand Forecasting:** Predicting demand for products with non-linear trends.
- **Environmental Modeling:** Modeling environmental factors with complex interactions.

Objectives

- Intro to regression analysis
- Linear regression
- Regularized regression
- Non-linear regression & ensemble models
- **Advanced regression techniques**
- Robust regression
- Quantile regression
- Model selection

Support Vector Machine Regression

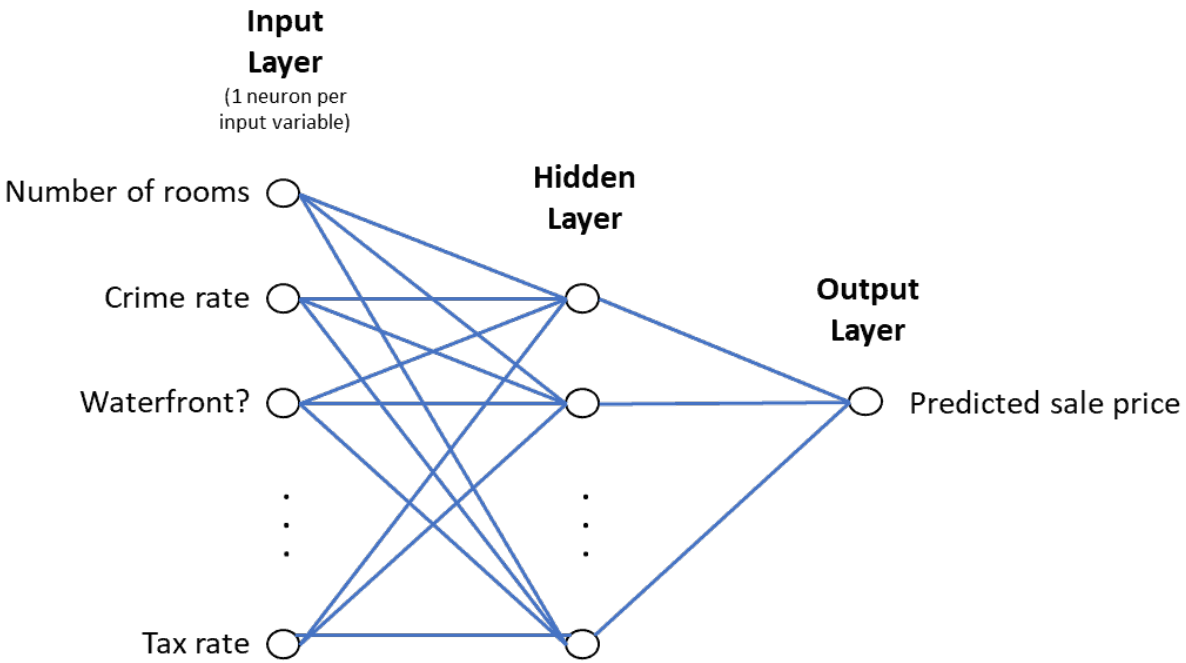
Minimize the vector space around points



- Effective in high-dimensional spaces
- Kernel functions to map data
- Sensitive to Kernel choices
- Requires careful hyperparameter tuning

Support Vector Machine Regression

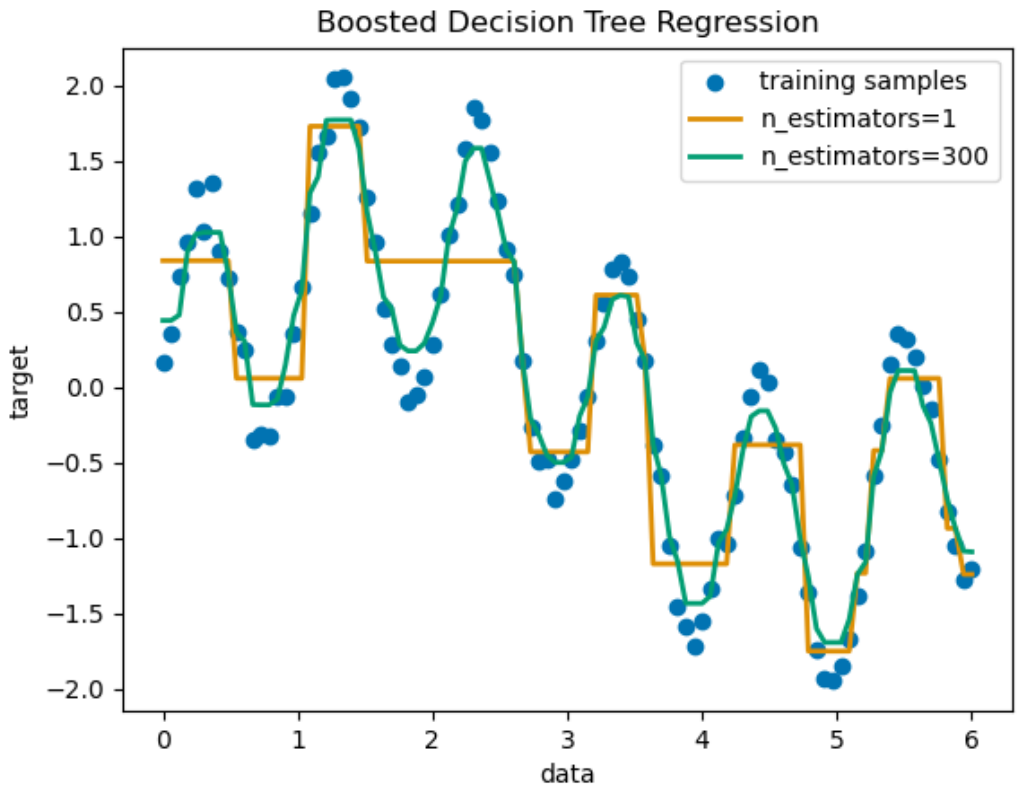
Deals with complex data well through multiple hidden layers



- **Deep Learning** enables models to learn complex non-linear relationships
- Need to understand architecture and training algorithms
- Requires large amounts of data
- Challenging to interpret (*black box*)

Gradient Boosting for Regression

Combines predictions of multiple weak models → strong regression model



- Iteratively adds weak learners (trees) to improve accuracy
- Effective hyperparameter tuning is crucial for optimal performance
- E.g., AdaBoost, XGBoost, LightGBM

Use cases

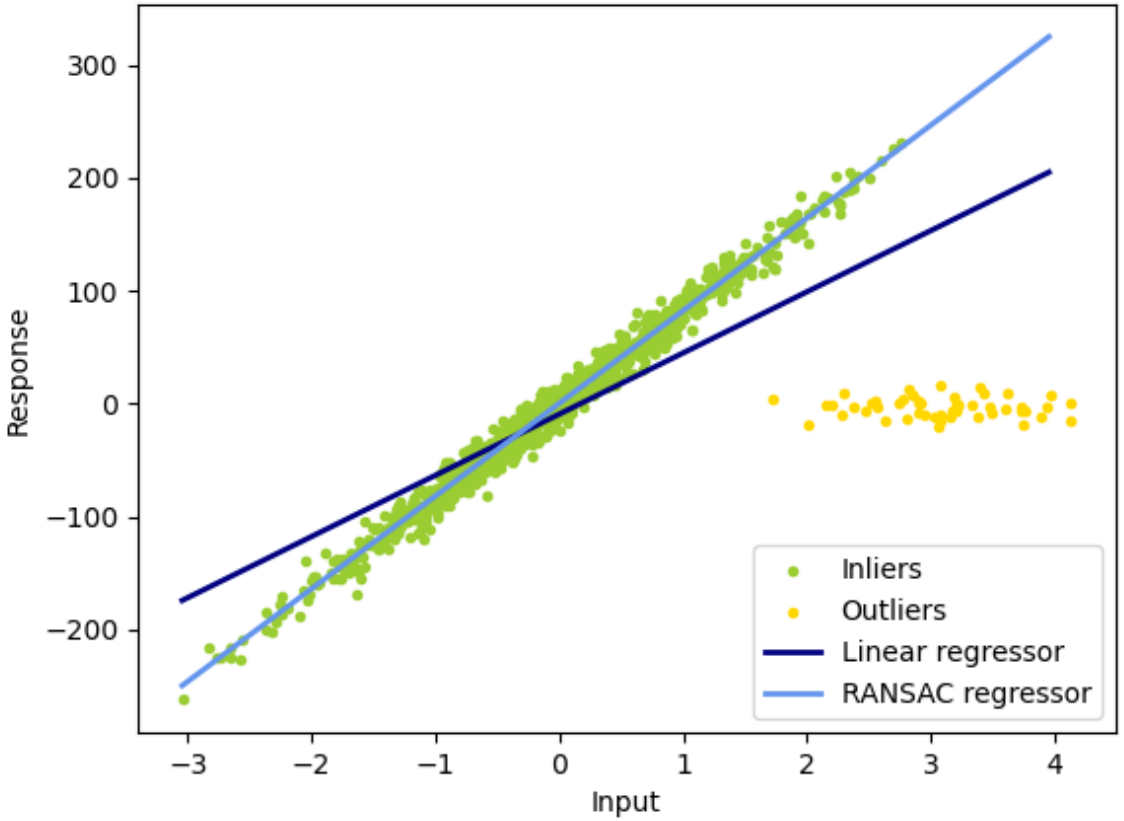
- **Financial Market Forecasting:** SVR for predicting stock prices or volatility.
- **Image Analysis:** Neural networks for image denoising or object recognition.
- **Click-through Rate Prediction:** Gradient boosting for optimizing online ad campaigns.
- **Housing Market Forecasting:** Predicting housing prices in dynamic real estate markets.

Objectives

- Intro to regression analysis
- Linear regression
- Regularized regression
- Non-linear regression & ensemble models
- Advanced regression techniques
- **Robust regression**
- Quantile regression
- Model selection

RANSAC Regression

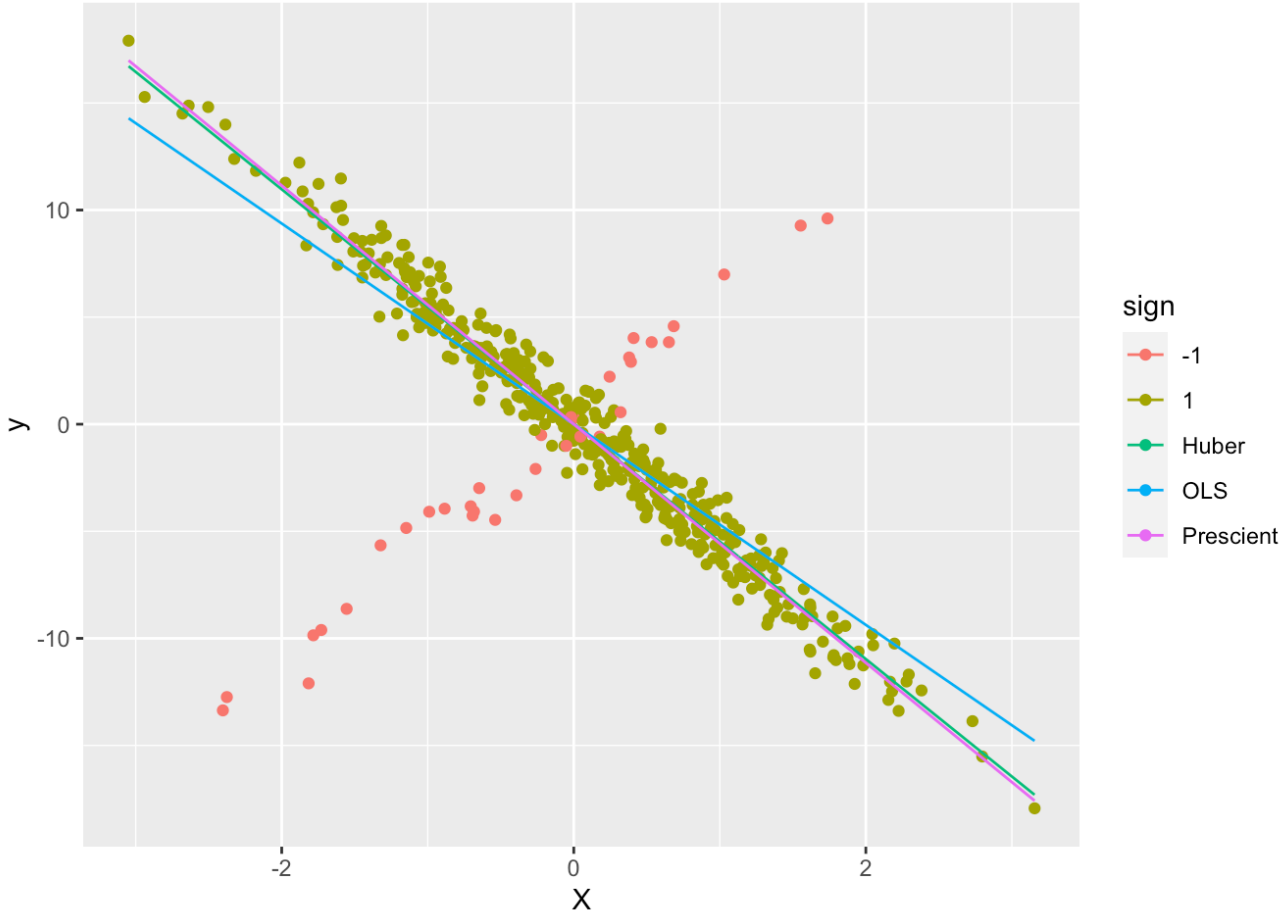
Random Sample Concensus: used for handling data with outliers



- Follows an iterative process, identifying outliers and inliers
- Fits models until best inlier fit is achieved

Huber Regression

Balances Mean Squared Error (MSE) and Mean Absolute Error (MAE)



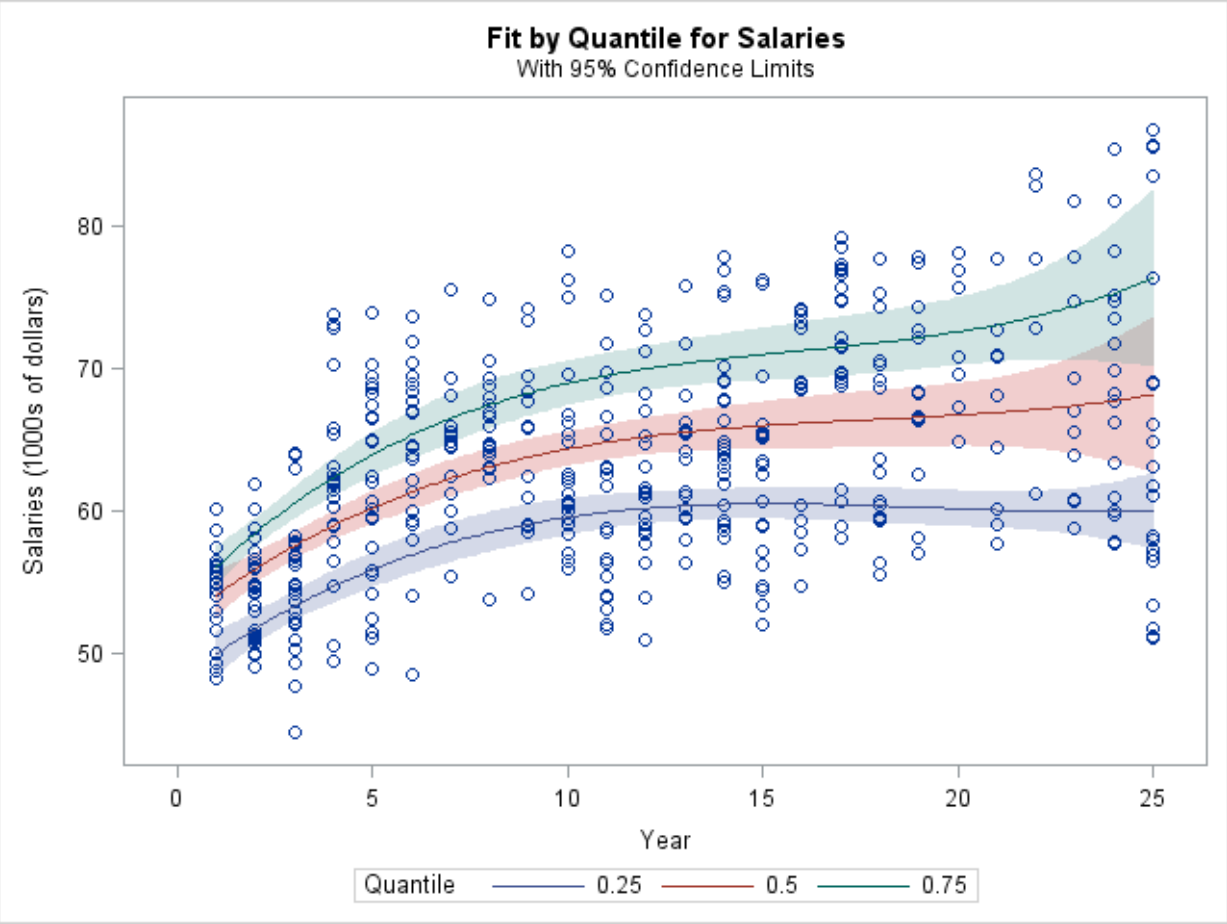
- Robust to outliers (MSE is sensitive, MAE is robust)
- Hyperparameter (δ) controls the balance between MSE and MAE
- Useful with normally distributed data with outliers

Objectives

- Intro to regression analysis
- Linear regression
- Regularized regression
- Non-linear regression & ensemble models
- Advanced regression techniques
- Robust regression
- **Quantile regression**
- Model selection

Quantile Regression

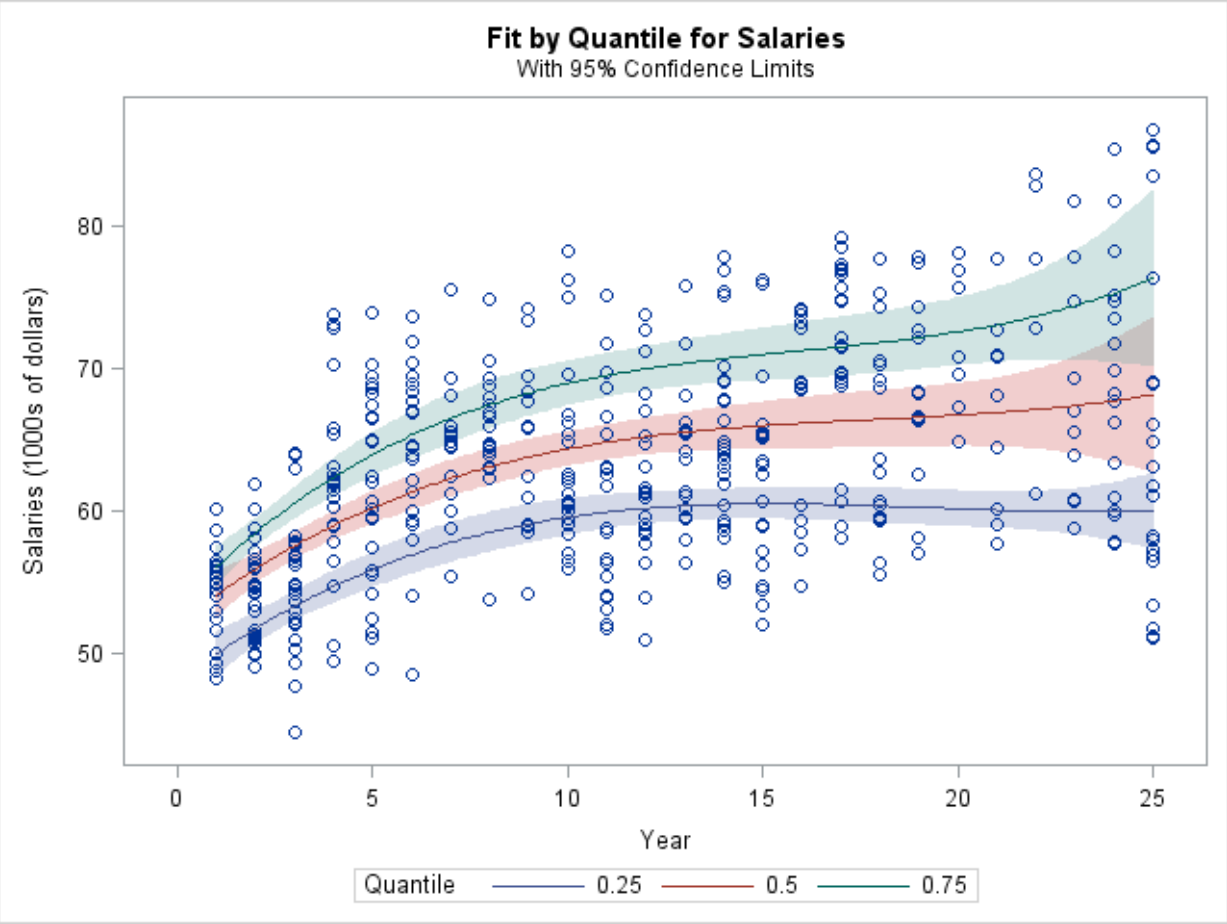
Estimates conditional quantiles of the dependent variable



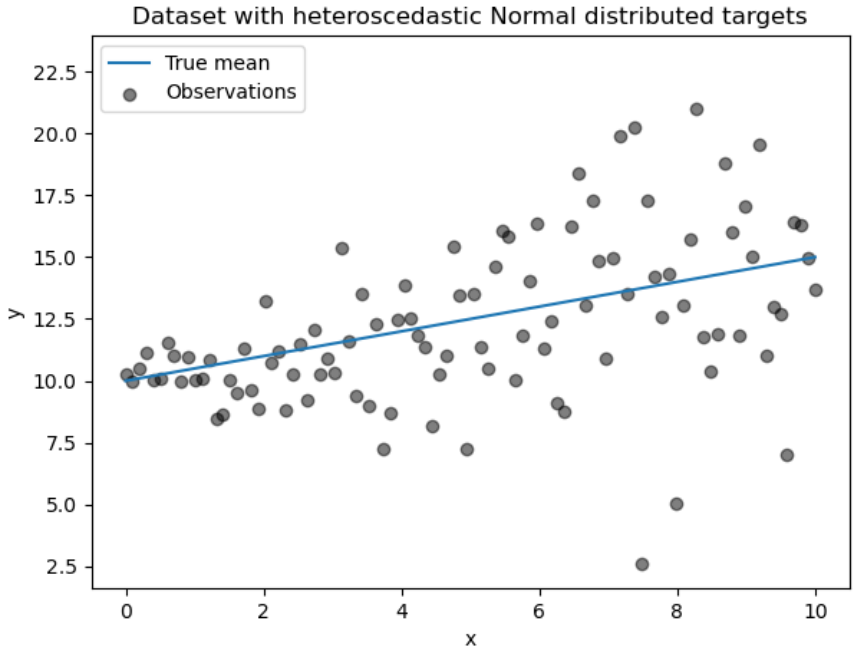
- More comprehensive view of variable relationships (median, 25th percentile, 75th percentile)
- Useful when the relationship varies at different points of a distribution

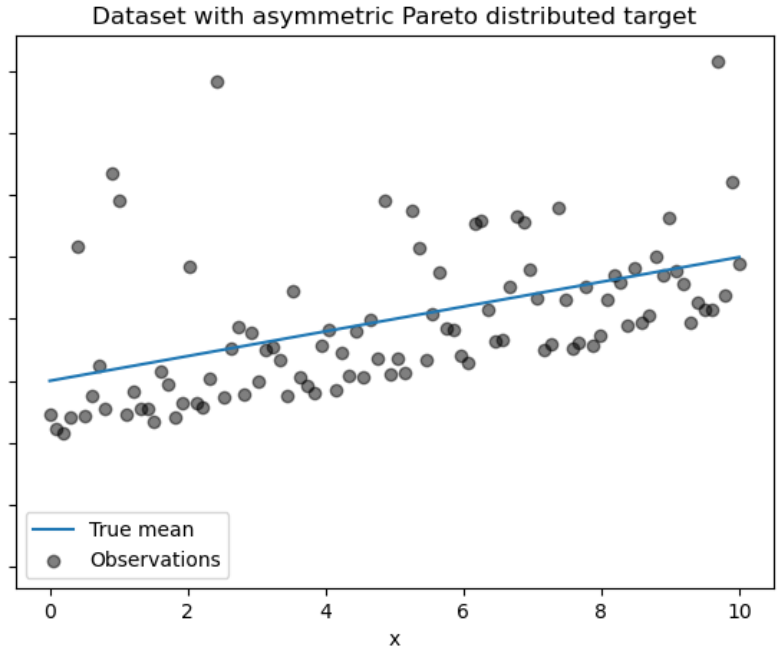
Quantile Regression

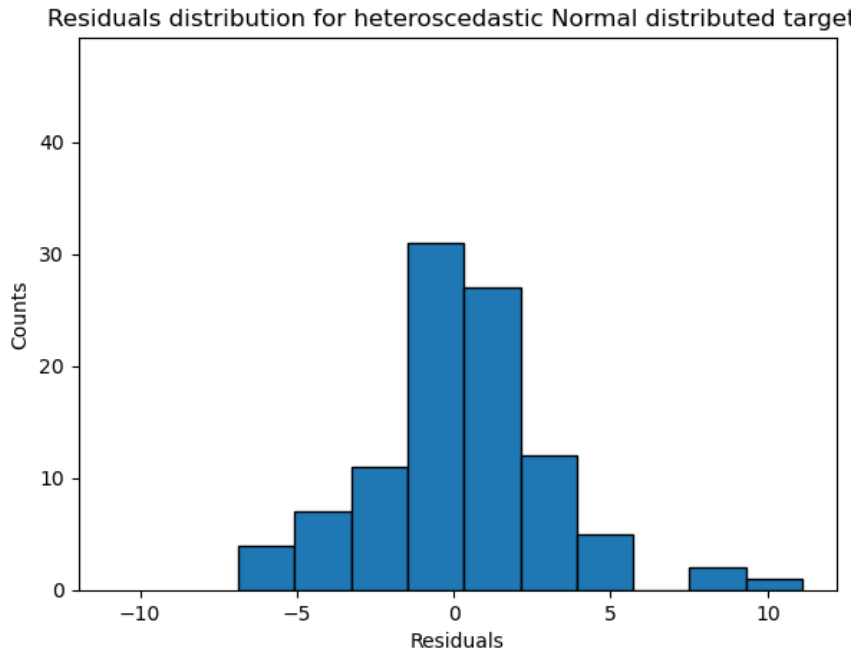
Estimates conditional quantiles of the dependent variable

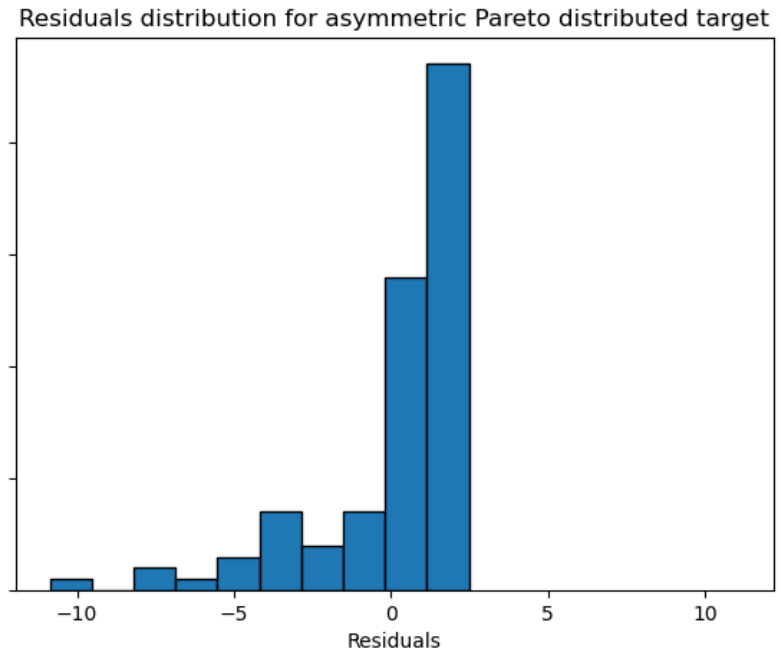


- Separate coefficients for each quantile of interest.
- Quantile-specific coefficients show how changes in independent variables affect conditional distribution of dependent variable at different points


Dataset with heteroscedastic Normal distributed targets
True mean
Observations
y
x


Dataset with asymmetric Pareto distributed target
True mean
Observations
x


Residuals distribution for heteroscedastic Normal distributed targets
Counts
Residuals


Residuals distribution for asymmetric Pareto distributed target
Residuals

Use cases

- **Income Prediction:** Quantile regression estimates income quantiles, revealing income disparities and aiding financial decisions.

- **Risk Assessment:** Quantile regression is valuable in insurance for estimating claim quantiles, assessing risk exposure, and setting premiums or reserves.

Advantages of Quantile Regression

- **Robustness:** Quantile regression is robust to outliers because it estimates conditional quantiles rather than means. Outliers have a more limited impact on quantile estimates.

- **Comprehensive Insights:** It provides a complete picture of the relationship between variables, including extreme cases, which can be crucial for decision-making.

- **Handling Heteroscedasticity:** Quantile regression can be more suitable than linear regression when the variance of the dependent variable varies across levels of the independent variables.

Challenges & Considerations

- **Interpretation:** Interpreting quantile-specific coefficients may require a deeper understanding of the data and domain knowledge.
- **Computational Complexity:** Estimating multiple quantiles can be computationally intensive, especially with large datasets.
- **Model Selection:** Choosing the right quantiles to estimate depends on the problem and goals.

Objectives

- Intro to regression analysis
- Linear regression
- Regularized regression
- Non-linear regression & ensemble models
- Advanced regression techniques
- Robust regression
- Quantile regression
- **Model selection**

Selecting among competing models

We will review a few commonly used methods!

AIC
Stepwise regression
All subset regression

Selecting among competing models

AIC (Akaike Information Criterion)

$AIC = 2k - 2\ln(L)$, where k is the number of parameters and L the likelihood.

Index that takes into account model complexity (number of parameters) and fit. The lowest the AIC score, the better!

Selecting among competing models

Stepwise regression

Y might depend on many variables! Which combination of variables explains better Y?

Two alternative approaches. First, in a "forward" approach, variables are added until no improvement is noted. In a "backward" approach, variables reducing model quality are deleted from a full model.

Selecting among competing models

All subset regression

Exhaustive approach that is likely only useful when the number of variable combinations is reduced. All the possible models are examined.

Similar to stepwise regressions but instead of examining predictor combinations in an alternative fashion, this approach examines *all* possible combinations!