# INTRODUCTION TO MACHINE LEARNING

## DATA 602

## Lecture 4

Dr. Tony Diana

tonydian@umbc.edu

**Sir Francis Galton and Regression Modeling**

- Legendre (1805) and Gauss (1809) used the method of least squares to study astronomical phenomena
- Francis Galton, a geneticist, coined the term 'regression'
- In 1889, he published an article in which he demonstrated how **every characteristic of an individual is inherited by their offspring, but on an average to a lesser degree**
  - For example, children with tall parents are also tall, but on an average their height will be comparatively less than that of their parents
- This phenomenon, also graphically described, is called **regression**
- Since then, this term has lived on to define statistical techniques that analyze relationships between two or more variables

# ***Regression Analysis: An Overview***

**Why Regression Analysis?**

- *It is a statistical process for estimating the **relationship** between one or several independent variables (**features,** or **explanatory**) and a dependent variable (**label,** or **response**)*
- It makes it possible to assess the percent variation in the label variable explained by the feature variable(s) measured by the coefficient of determination, $R^2$

**What are the goals of Regression Analysis?**

- Use a sample from the population to estimate the properties of the population and predict the label
- The coefficients in the model are estimations of the actual population parameters

**What are the Key Assumptions of a Linear Regression Model?**

- **Linearity:** the dependent variable is a **linear function** of the independent variables
- **Homoscedasticity:** the error term has a **mean of zero** and **equal finite variance** at all levels of the independent variable(s)
- **Multivariate Normality:** the **residuals** are assumed to be **normally distributed**
- **Independence of Errors:** the **residuals** are **uncorrelated** with each other
- **No Multicollinearity**: **no independent variable** is perfectly **correlated** to another independent variable

# Regression Analysis: An Overview (Cont.)

**What are the Types of Linear Regression Models?**

- **Simple:** $y = \alpha + \beta x + \varepsilon$, with $\alpha$ as the intercept (the value when $\beta = 0$), $\beta$ as the coefficient, and $\varepsilon$ as the error
- **Multiple:** $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon$

**How to write the Regression Model?**

- $\hat{y} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 x_1} + \ldots + \boldsymbol{\beta_n x_n} + \boldsymbol{\epsilon}$ or
- $\hat{y} = h_0(x) = \boldsymbol{\beta}\mathbf{x}$
- $\hat{y} = \boldsymbol{\beta}^T \mathbf{x}$

with $\hat{y}$ as the predicted value, $\boldsymbol{\beta}$ the model's parameter vector including the intercept or bias vector $\boldsymbol{\beta_0}$, the feature weights $\boldsymbol{\beta}_1$ to $\boldsymbol{\beta}_n$, and **x** as the instance's feature vector containing $\boldsymbol{x}_1$ to $\boldsymbol{x}_n$ with $\boldsymbol{x}_0$ always equal to 1. $\boldsymbol{\beta}\mathbf{x}$ is the dot product of the vectors $\boldsymbol{\beta}$***X**. $\mathbf{H_0}$ is the hypothesis function using the model parameter $\boldsymbol{\beta}$*$\boldsymbol{\beta}^T$ is the transpose of $\boldsymbol{\beta}$ (a row vector instead of a column vector) and $\boldsymbol{\beta}^T$ **x** is a matrix multiplication of $\boldsymbol{\beta}^T$ and **x**

- The **cost function** for a **linear regression model** is **the Mean Square Error(X, $h_0$) =** $\frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{\beta}^T\ x^{(i)} - y^{(i)})^2$
- If we are looking at the $\boldsymbol{\beta}$ value that minimizes the cost function, we use the closed-form solution

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

**What are the Types of Variables?**

- **Categorical or Nominal Variable:** refers to a category (Republican, Democrat, Independent) or a label ("1" or "0"). There is **no intrinsic order** or **measure of distance** between categorical variables
- **Ordinal**: ordinal variables have two or more categories that are **ordered** or **ranked** ("unsatisfied," "neutral," "satisfied," or "do not know")
- **Interval**: interval variables are **measures along a continuum**, have a **numerical value**, can be **ranked by virtue of the distance** between each of them ($78^0$ F compared with $80^0$ F)
- Regression models with categorical, ordinal, or interval (also called **qualitative** dependent variables) are not linear regression models. A linear regression model uses **ratio** or **quantitative** variables
- **Ratio**: a ratio variable can be **added**, **subtracted**, **multiplied**, **divided**, and **transformed** (log, for instance). It measures **ranking** and **interval**. A **zero measurement** for a ratio variable indicates the **non-existence** of the variable as opposed to $0^0$ C

**Discrete v. Continuous Variables:**

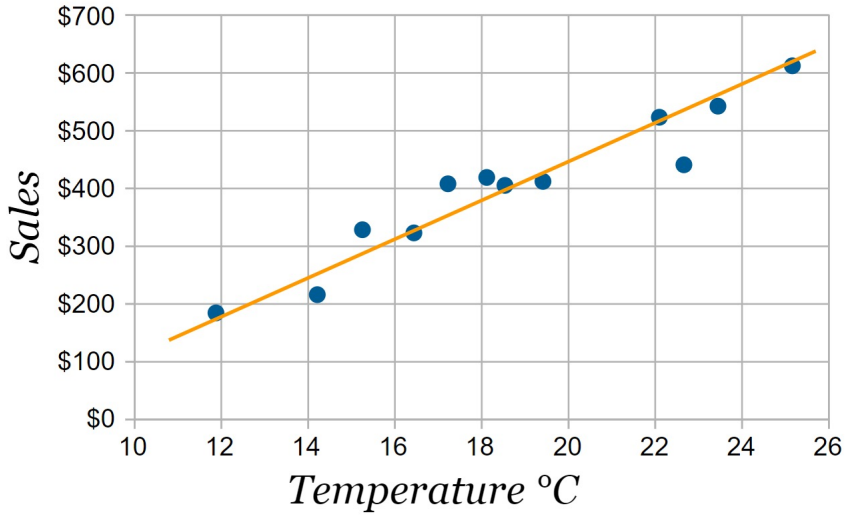
- A **continuous** variable can take any value between value A and value B (2.5 stands between 2 and 3)
- A **discrete** variable is countable and defined (number of heads or tails)

**What is a Polynomial Regression?**

- When independent variable(s) is(are) raised to a specific power: $y = \alpha + \beta_1 x + \beta_2 x^2 \varepsilon$

## *Linear Model Two-Dimensional Plot / Best-Fit Line*

**Curve fitting** is the process to specify the model that provides the **best fit** to specific curves in our data set. Example: the sales of ice-creams in relation to weather temperature




**Step 1**: For each (x,y) point calculate $x^2$ and xy

**Step 2**: Sum all x, y, $x^2$ and xy, which gives us $\Sigma x$, $\Sigma y$, $\Sigma x^2$ and $\Sigma xy$

**Step 3**: Calculate Slope **m**:

$$\mathbf{m} = \frac{N\,\Sigma(xy) - \Sigma x\,\Sigma y}{N\,\Sigma(x^2) - (\Sigma x)^2}$$

(N is the number of points.)

**Step 4**: Calculate Intercept **b**:

$$\mathbf{b} = \frac{\Sigma y - m\,\Sigma x}{N}$$

**Step 5**: Assemble the equation of a line

$$y = mx + b$$

## Log-Transformed Variables

- Logarithmically transforming variables in a regression model is a very common way to handle situations where a **non-linear relationship** exists between the **independent** and **dependent** variables
- Using the logarithm of one or more variables instead of the un-logged form makes the effective relationship non-linear, while still preserving the linear model

| | $X$ | |
|---|---|---|
| $Y$ | $X$ | $\log X$ |
| $Y$ | *linear* $\hat{Y}_i = \alpha + \beta X_i$ | *linear-log* $\hat{Y}_i = \alpha + \beta \log X_i$ |
| $\log Y$ | *log-linear* $\log\hat{Y}_i = \alpha + \beta X_i$ | *log-log* $\log\hat{Y}_i = \alpha + \beta \log X_i$ |

```
df1['log_Price'] =
np.log(df1['Price’])
print(df1)
```

**Interpretation of Log Model Coefficients**

- **Linear** ($Y_i = \alpha + \beta X_i + \varepsilon_i$): the coefficient β gives us directly the change in Y for a one-unit change in X
- **Linear-Log** ($Y_i = \alpha + \beta \log X_i + \varepsilon_i$): the literal interpretation of the estimated coefficient $\hat{\beta}$ is that a one-unit increase in logX will produce an expected increase in Y of $\hat{\beta}$ units
- **Log-Linear** ($\log Y_i = \alpha + \beta X_i + \varepsilon_i$): the literal interpretation of the estimated coefficient $\hat{\beta}$ is that a one-unit increase in X will produce an expected increase in log Y of $\hat{\beta}$ units. For instance, for $\hat{\beta} = .06$, $e^{.06} \approx 1.06$, so a 1-unit change in X corresponds to (approximately) an expected increase in Y of 6%
- **Log-Log** ($\log Y_i = \alpha + \beta \log X_i + \varepsilon_i$): the interpretation is given as an expected percentage change in Y when X increases by some percentage. Such relationships, where both Y and X are log-transformed, are commonly referred to as elastic in econometrics, and the coefficient of log X is referred to as an elasticity

# Linear Regression and Scikit-Learn

```
from sklearn.metrics import mean_squared_error, r2_score
Import matplotlib.pyplot as plt
%maplotlib inline
# Load the diabetes dataset
diabetes = datasets.load_diabetes()


# Use only one feature
diabetes_X = diabetes.data[:, np.newaxis, 2]

# Split the data into training/testing sets
diabetes_X_train = diabetes_X[:-20]
diabetes_X_test = diabetes_X[-20:]

# Split the targets into training/testing sets
diabetes_y_train = diabetes.target[:-20]
diabetes_y_test = diabetes.target[-20:]

# Create linear regression object
regr = linear_model.LinearRegression()

# Train the model using the training sets
regr.fit(diabetes_X_train, diabetes_y_train)

# Make predictions using the testing set
diabetes_y_pred = regr.predict(diabetes_X_test)

# The coefficients
print('Coefficients: \n', regr.coef_)
# The mean squared error
print("Mean squared error: %.2f"
    % mean_squared_error(diabetes_y_test, diabetes_y_pred))
# Explained variance score: 1 is perfect prediction
print('Variance score: %.2f' % r2_score(diabetes_y_test, diabetes_y_pred))

# Plot outputs
plt.scatter(diabetes_X_test, diabetes_y_test,  color='black')
plt.plot(diabetes_X_test, diabetes_y_pred, color='blue', linewidth=3)

plt.xticks(())
plt.yticks(())

plt.show()
```

We are import the MSE and $R^2$ functions from the metrics module of the sklearn class

We load the data from Scikit-Learn: https://scikit-learn.org/stable/datasets/index.html#diabetes-dataset

We split the feature data into training and testing datasets

We split the label data into training and testing datasets

We instantiate the model

We train the model

We predict the labels

We measure errors and model performance

We plot the outcomes

# *The statsmodels library for Linear Regression Models*

```
import numpy as np
from sklearn import datasets
import pandas as pd
from pandas import DataFrame
import statsmodels.formula.api as sm
# Load the diabetes dataset
diabetes = datasets.load_diabetes()
# Reset data as pandas DataFrame
diabetes_df = DataFrame(diabetes.data)
# Label columns
diabetes_df.columns = diabetes.feature_names
result = sm.ols(formula="s4 ~ bp + age + bmi",
data=diabetes_df).fit()
print(result.params)
print(result.summary())
```

Using statsmodels

S4 is one of the six blood serum measurement, bp is blood pressure, bmi is body mass index

# *The 'statsmodels' Linear Regression Outputs*

```
Intercept    3.876633e-16
bp           7.699470e-02
age          1.108580e-01
bmi          3.628436e-01
dtype: float64
                            OLS Regression Results
==============================================================================
Dep. Variable:                     s4   R-squared:                       0.193
Model:                            OLS   Adj. R-squared:                  0.187
Method:                 Least Squares   F-statistic:                     34.82
Date:                Wed, 14 Aug 2019   Prob (F-statistic):           3.34e-20
Time:                        14:28:16   Log-Likelihood:                 766.28
No. Observations:                 442   AIC:                            -1525.
Df Residuals:                     438   BIC:                            -1508.
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   3.877e-16      0.002    1.9e-13      1.000      -0.004       0.004
bp             0.0770      0.049      1.576      0.116      -0.019       0.173
age            0.1109      0.046      2.428      0.016       0.021       0.201
bmi            0.3628      0.047      7.748      0.000       0.271       0.455
==============================================================================
Omnibus:                       40.779   Durbin-Watson:                   1.967
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               49.573
Skew:                           0.773   Prob(JB):                     1.72e-11
Kurtosis:                       3.549   Cond. No.                         28.0
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

- What can we say about this model?
- How do you interpret the coefficient and intercept?
- Are all the variables significant at a 95% level? Why?

# *The 'statsmodels' Linear Regression Outputs (Cont.)*

**Key Regression Metrics**

- **Skewness**: a symmetrical bell curve has a value of 0
  - It is symmetrical between -0.5 and 0.5
  - In this case, the right tail is longer (positive skewness)
- **Kurtosis**: it is a measure of outliers in the distribution
  - The normal distribution has a value of 3
  - Kurtosis > 3 → tall distribution
  - Kurtosis < 3 → flat distribution
- **Durbin-Watson**: it is a test of autocorrelation in the residuals
  - A value of 2 means there is no autocorrelation
  - Values from 0 to 2 mean positive autocorrelation and values from 2 to 4 indicate negative autocorrelation
- **Jarque-Bera**: it is a goodness of fit test
  - A value of zero means the data is normally distributed. We reject the null hypothesis that the disturbance term is normally distributed at a 95% level.
- **AIC**: Akaike Information Criterion. When working with different models, the one with the lowest AIC is preferred
- **BIC**: Bayesian Information Criterion.
  - AIC = 2k – 2ln(L) where k is the number of estimated parameters in the model and L the maximized likelihood function of the estimated model
  - BIC = -2ln(L) + k*ln(n) with n as the number of observations and k the number of regressors plus intercept

# How Good is a Regression Model?

**There are four metrics widely used for evaluating a linear model performance**

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

$$\frac{Total\ Sum\ of\ Square\ Residuals}{Sum\ of\ Square\ Total}$$

$$Adjusted\ R^2 = 1 - \frac{(1 - R^2)(N-1)}{N-p-1}$$

The **adjusted R-squared** increases only if the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected by chance

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

**Root Mean Squared error** indicates how close the predicted values are to the actual values

$$MAE = \frac{\sum_{i=1}^{n} |y_{pred,i} - y_i|}{n}$$

**Mean Absolute error** is the average of the absolute value of the errors (predicted minus actual values)

## ***Problems to Watch in Regression Models***

- **Non-Linearity**: A change in one entity does not correspond with constant change in the other entity

- **Multicollinearity**: It refers to a situation in which two or more explanatory variables in a multiple regression model are highly linearly related
  - A variance inflation factor (VIF) provides a measure of multicollinearity among the independent variables in a multiple regression model
  - A large VIF on an independent variable indicates a highly collinear relationship to the other variables that should be considered or adjusted for in the structure of the model and selection of independent variables

- **Heteroscedasticity**: we talk about this problem in the context of the residuals or error terms
  - The estimated **standard error** is wrong. Because of this, confidence intervals and hypotheses tests cannot be relied on. In addition, the OLS estimator is no longer **BLUE**
  - One way to **correct** for **heteroscedasticity** is to compute the weighted least squares (WLS) estimator

## ***Problems to Watch in Regression Models (Cont.)***

- **Non-Normally-Distributed Error Terms**: This may indicate the presence of outliers
  - We assume that the error term in the linear relationship between the dependent variable $Y$ and some feature vector $X$ is normally distributed with mean zero and constant variance $\sigma^2$, that is, $Y|X \sim \boldsymbol{X}^\top \boldsymbol{\beta} + \varepsilon$ with $\varepsilon \in \mathcal{N}(0, \sigma^2)$ and $\beta$ being a set of coefficients
  - You could try (1) **robust regression**, which deals with outliers; (2) **quantile regression**, which makes no assumptions about the distribution of the residuals; or (3) **Huber regression**

# *Special Types of Regression Models*

# Polynomial Regression Models

$$y = a + b_1x + b_2x^2 + \ldots + b_nx^n$$

**Justifications**

- The straight **OLS line is unable to capture the patterns** in the data. This is an example of **underfitting**
- The polynomial model is still considered to be **linear model** as the coefficients/weights associated with the features are still linear
- $x^2$ is only a feature. The coefficients are not squared or raised to a higher power
- However, the curve that we are fitting is **quadratic** in nature

**How Do We Know We Should Use a Polynomial Model?**

- We hypothesize the **relationship** is not **curvilinear** (experience in modeling)
- A visual inspection of the data shows a **curvilinear relationship**
- We **cannot assume** that all the **independent variables are independent**
- Inspection of residuals:
  - If you try to fit a linear model to curved data, a scatter plot of residuals (Y axis) on the predictor (X axis) will have patches of many positive residuals in the middle, but patches of negative residuals at either end (or vice versa)

## *Polynomial Regression Model (Cont.)*

**Benefits**

- Broad range of functions can be fit under polynomial regression model assumptions
- Polynomial basically fits wide range of curvature
- Polynomial provides the best approximation of the relationship between dependent and independent variable

**Disadvantages**

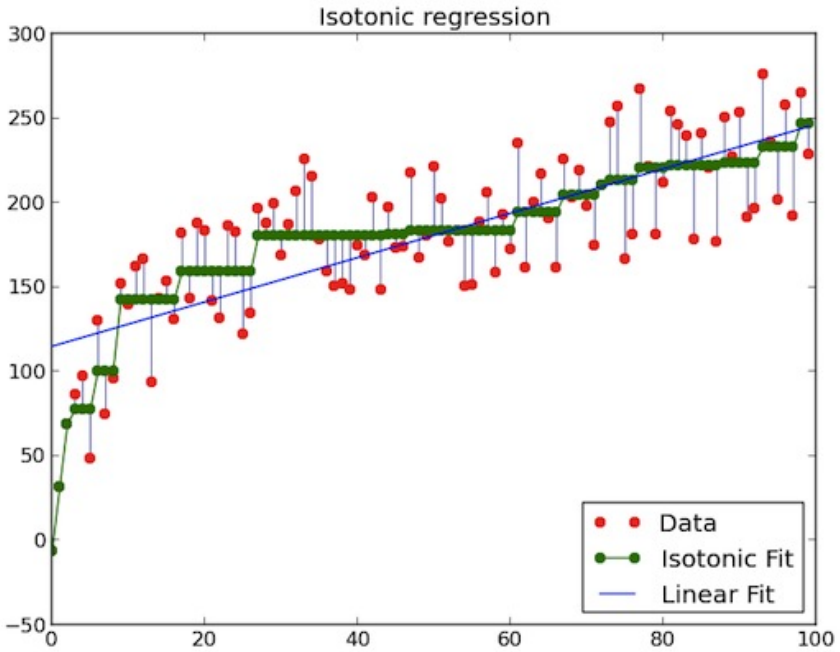
- These are **too sensitive** to the **outliers**
- The presence of one or two outliers in the data can seriously affect the results of a nonlinear analysis
- In addition, there are unfortunately **fewer model validation tools** for the detection of outliers in nonlinear regression than there are for linear regression

## *Isotonic Regression*

- **Isotonic regression** is a free-form linear model that can be fit to predict sequences of observations
- An **Isotonic Regression** model finds the best least squares fit to a set of points, given the constraint that the fit must be a non-decreasing function
- **Isotonic Regression** is based on **Weighted Least-Square Regression**
- However, there are two major differences between **Isotonic Regression** and a similar model like **Weighted Least Squares**
  - An **isotonic function** must not be **non-decreasing**. This is because an isotonic function is a **monotonic function**, which means it is a function that preserves, or reverses a given order
- The algorithm sweeps through the data looking for violations of the monotonicity constraint
  - When it finds one, it adjusts the estimate to the best possible fit with constraints
  - Sometimes it also needs to modify previous points to make sure the new estimate does not violate the constraints

## *Isotonic Regression (Cont.)*

- **Isotonic regressors** use an interesting concept called "**Order Theory**"
  - **Order Theory** deals with the idea of partially ordered and pre-ordered sets as a generalization of real numbers
  - **Order theory** is a branch of mathematics which investigates the intuitive notion of **order** using binary relations. It provides a formal framework for describing statements such as "this is less than that" or "this precedes that"



## ***Huber Regression***

- **Huber Regression** is a generalized version of the Huber loss function which can be incorporated with **Generalized Linear Models** (GLM). Huber developed the model in 1964
- It is well-suited for heteroscedastic regression problems
- The idea is to use a different loss function rather than the traditional least-squares
- The **Huber loss** is a robust loss function for regression problems defined as

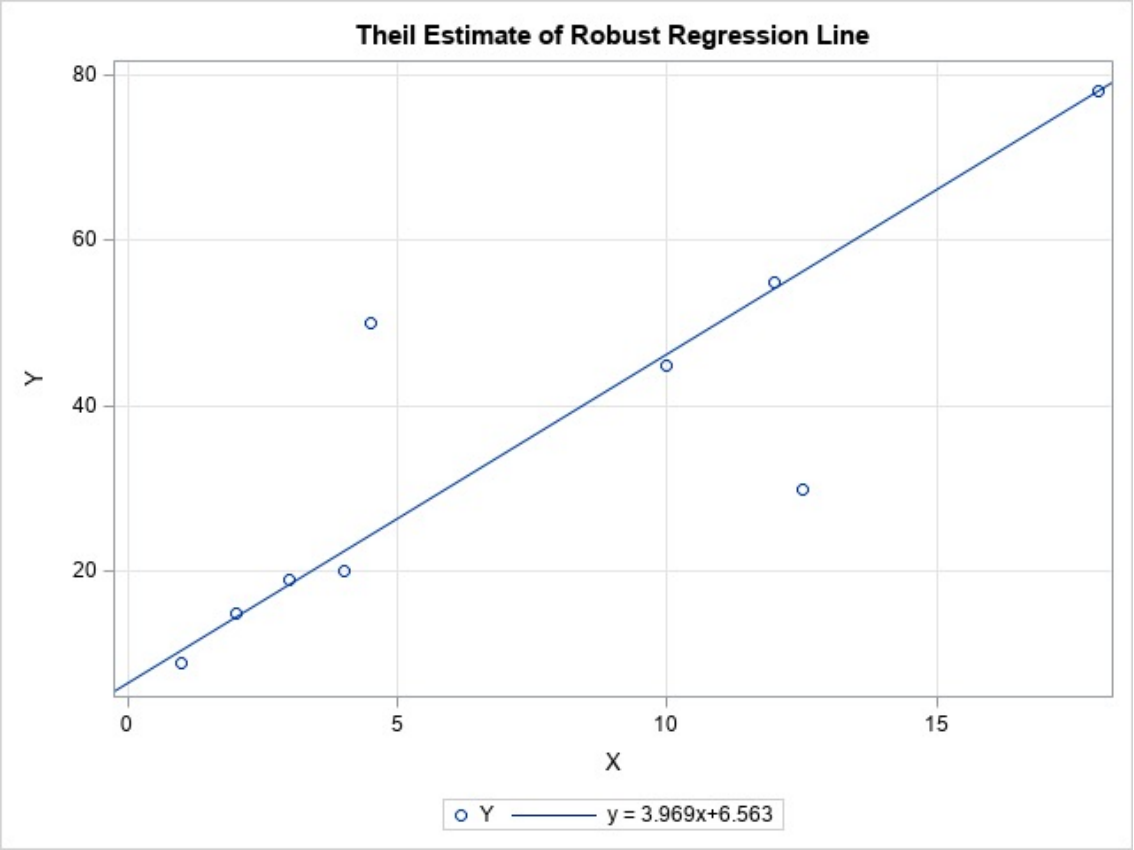
$$\mathcal{L}(y, \hat{y}) = \begin{cases} (y - \hat{y})^2 & \dots \quad |y - \hat{y}| \leq \alpha \\ |y - \hat{y}| & \dots \quad |y - \hat{y}| > \alpha \end{cases}$$

where $y$ is the target variable, $\hat{y}$ are the corresponding predictions and $\alpha \in \mathbb{R}^+$ is a hyperparameter

- For absolute errors smaller than $\alpha$, the corresponding distribution resembles the **normal distribution**, outside this region it coincides with the more heavy-tailed **Laplace distribution**. This makes the model robust against outliers

## ***Theil-Sen Regression***

- **Theil-Sen robust regression** method applies only to one-variable regression
- If the data contains N pairs of (x, y) values, compute all the slopes between pairs of points and choose the median as the estimate of the regression slope
- Using that slope, pass a line through each pair of (x,y) values to obtain N intercepts
- Choose the median of the intercepts as the estimate of the regression intercept
- You compute (N*(N-1)/2) slopes and take the median
- Then compute N intercepts and take the median
- The slope estimate is unbiased, and the process is resistant to outliers

# *Theil-Sen Regression (Cont.)*

- The scatter plot shows the **Theil-Sen regression line** for nine data points
- The seven data points that appear to fall along the regression line were used by Sen (1968)
- Two outliers were added
- The plot shows that the **Theil-Sen regression** line ignores the outliers and passes close to the other data points
- The slope of the **Theil-Sen** line is slightly less than 4
- In contrast, the least squares line through these data has a slope of only 2.4 because of the influence of the two outliers