UNIVERSITÉ ÉVRY PARIS-SACLAY

M2 SMART AEROSPACE AND AUTONOMOUS SYSTEMS

TRAVAUX PRATIQUES OF AI AND AEROSPACE SYSTEMS
LINEAR REGRESSION REPORT
LINEAR MODELS, OUTLIERS, AND CLASSIFICATION

DIYARI FARIQ M SALIH
(20256924@ETUD.UNIV-EVRY.FR)
13-*011-2025*

Introduction:

This report summarizes the experiments conducted in the TP1 Machine Learning notebook. Different contamination and noise values were used for testing; the goals were to:

1. Explore logistic regression decision boundaries using the Iris dataset.
2. Analyze the sensitivity of Ordinary Least Squares (OLS) to outliers.
3. Compare OLS with robust alternatives such as **RANSAC** and **Ridge Regression**.
4. Study how the number of outliers affects the OLS slope estimate.
5. Compare batch OLS with gradient-based methods such as **Stochastic Gradient Descent (SGD)**.

All results are supported by the visualizations included throughout this report.

2. LOGISTIC REGRESSION ON IRIS DATASET

2.1 Objective

Classify Iris flower samples into:

- **Versicolor (0)**
- **Virginica (1)**
- **Setosa (2)**

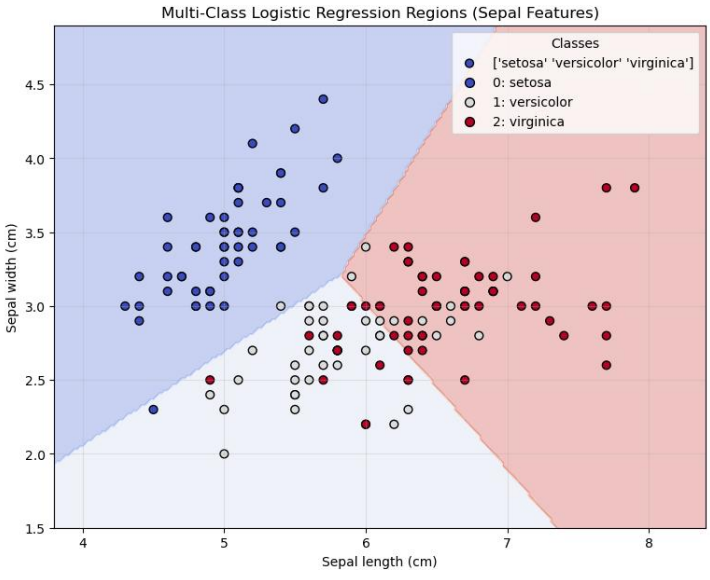
using logistic regression with two input features:

- Petal length
- Petal width

2.2 Result

The model successfully learns a linear decision boundary separating the two classes.

Figure 1 – Logistic Regression Decision Boundary:



This figure to the right shows:

- Blue points: Versicolor
- Red points: Virginica
- Grey points: Setosa

The model performs well due to the near-linear separability of these two classes in petal dimensions. While there are some classification errors between class 0 and 1.

3. SENSITIVITY OF OLS TO OUTLIERS

3.1 Motivation

OLS minimizes squared error, making it highly sensitive to extreme values.
This experiment compares:

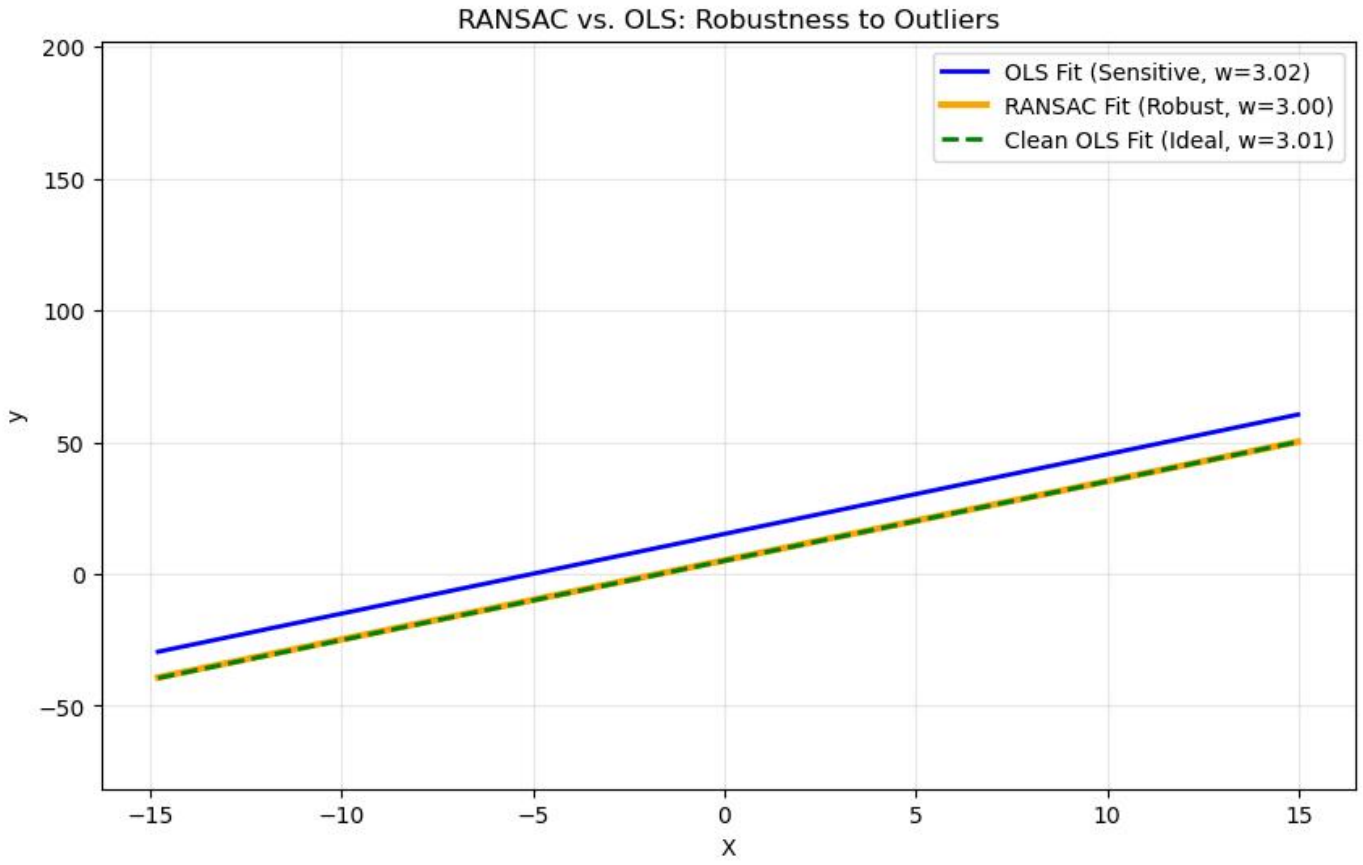
- **OLS** (sensitive)
- **RANSAC** (robust)

3.2 Result

Key observations:

- OLS is pulled strongly by outliers, causing a distorted slope.
- RANSAC identifies inlier points and fits a line closer to the true underlying trend.
- Clean OLS (with no outliers) aligns closely with RANSAC

Figure 2 – RANSAC vs OLS:



4. COMPARING CLEAN OLS, CONTAMINATED OLS, AND RIDGE REGRESSION

4.1 Objective

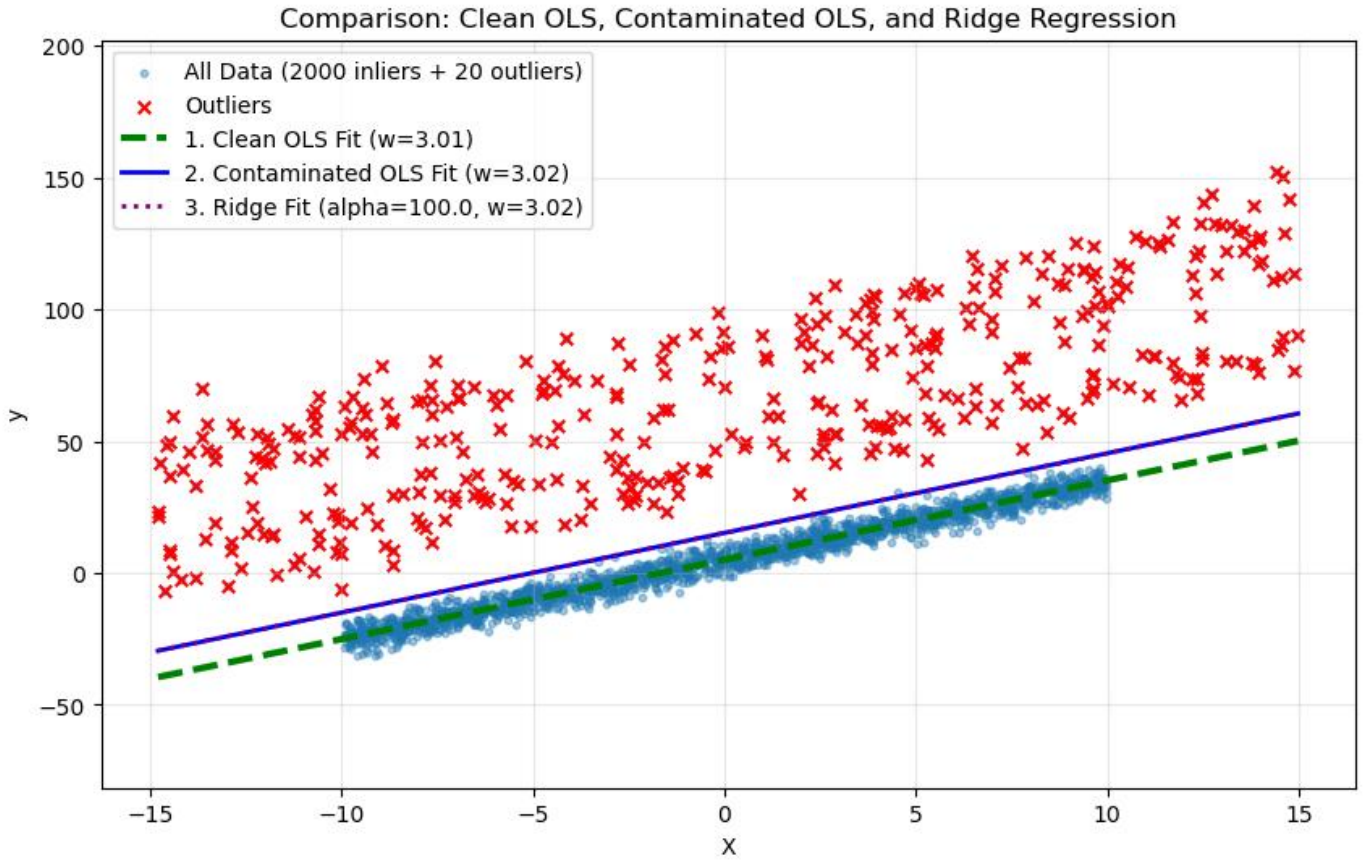
Analyze how OLS changes when outliers are added, and examine whether **L2-regularization** (Ridge) mitigates the issue.

4.2 Result

Observations:

- Contaminated OLS is biased upward due to outliers.
- Ridge regression slightly reduces the effect but **cannot fully solve** outlier sensitivity.
- For outlier robustness, RANSAC performs better.

Figure 3 – Clean OLS vs Contaminated OLS vs Ridge:



5. EFFECT OF OUTLIERS ON OLS SLOPE

5.1 Purpose

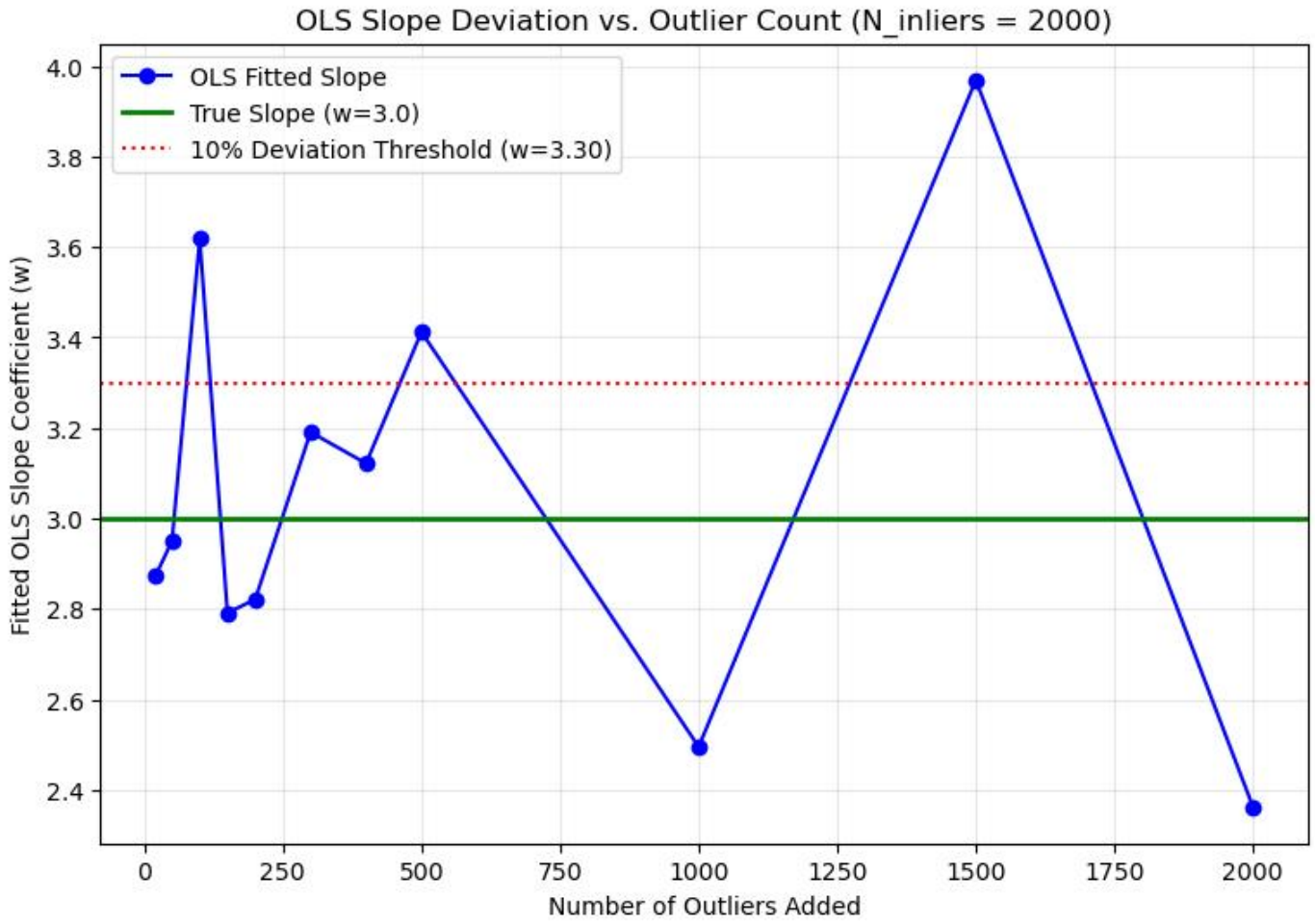
Measure how the estimated slope changes as more outliers are added.

5.2 Result

Takeaways:

- Even relatively small numbers of outliers cause substantial changes in slope.
- Past a threshold, OLS becomes highly unstable.
- A 10% deviation band illustrates the rapid performance deterioration.

Figure 4 – OLS Slope Deviation vs Outlier Count:



6. VISUAL EXAMPLES OF OLS DISTORTION

Figures 5 and 6 illustrate how contaminated data shifts the regression line.

Figure 5 – Moderate Outliers, single Cloud:

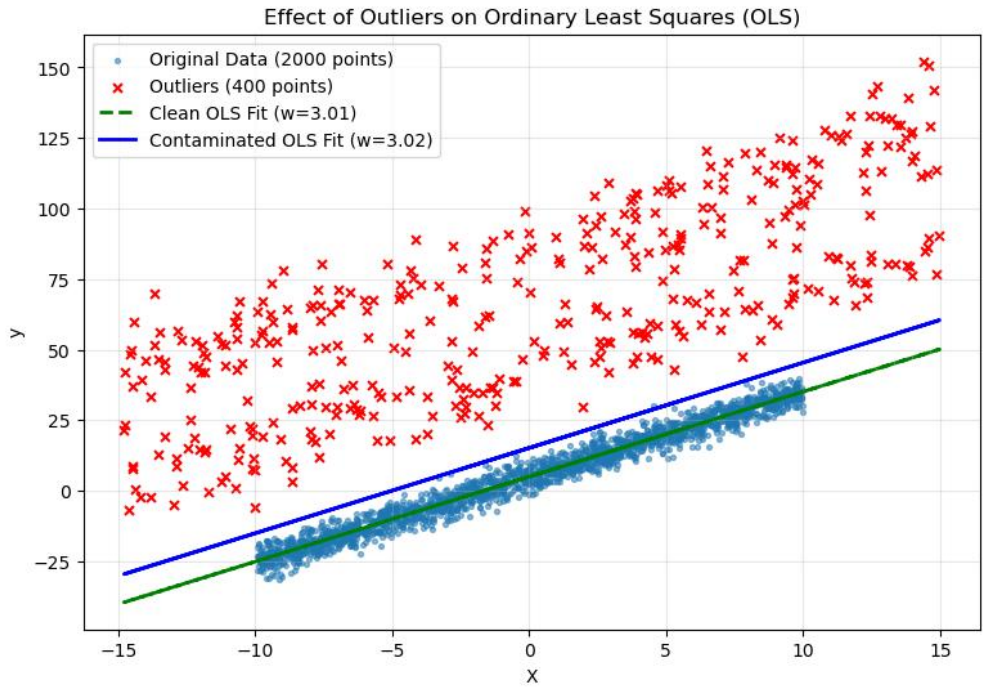
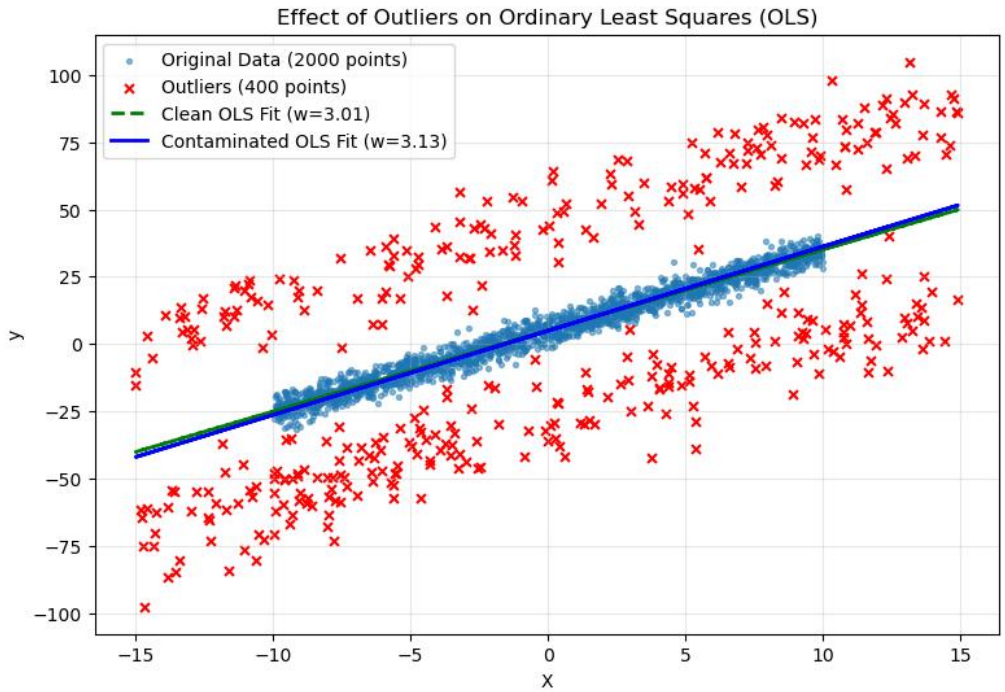


Figure 6 – Extreme Outliers, two Clouds:



As outlier magnitude increases:

- The slope deviates further from the true slope w=3.0w = 3.0w=3.0
- OLS is dragged towards the outlier cloud
- Inliers become less influential on the regression line if the distribution is on both sides

7. COMPARISON OF OLS AND SGD REGRESSION

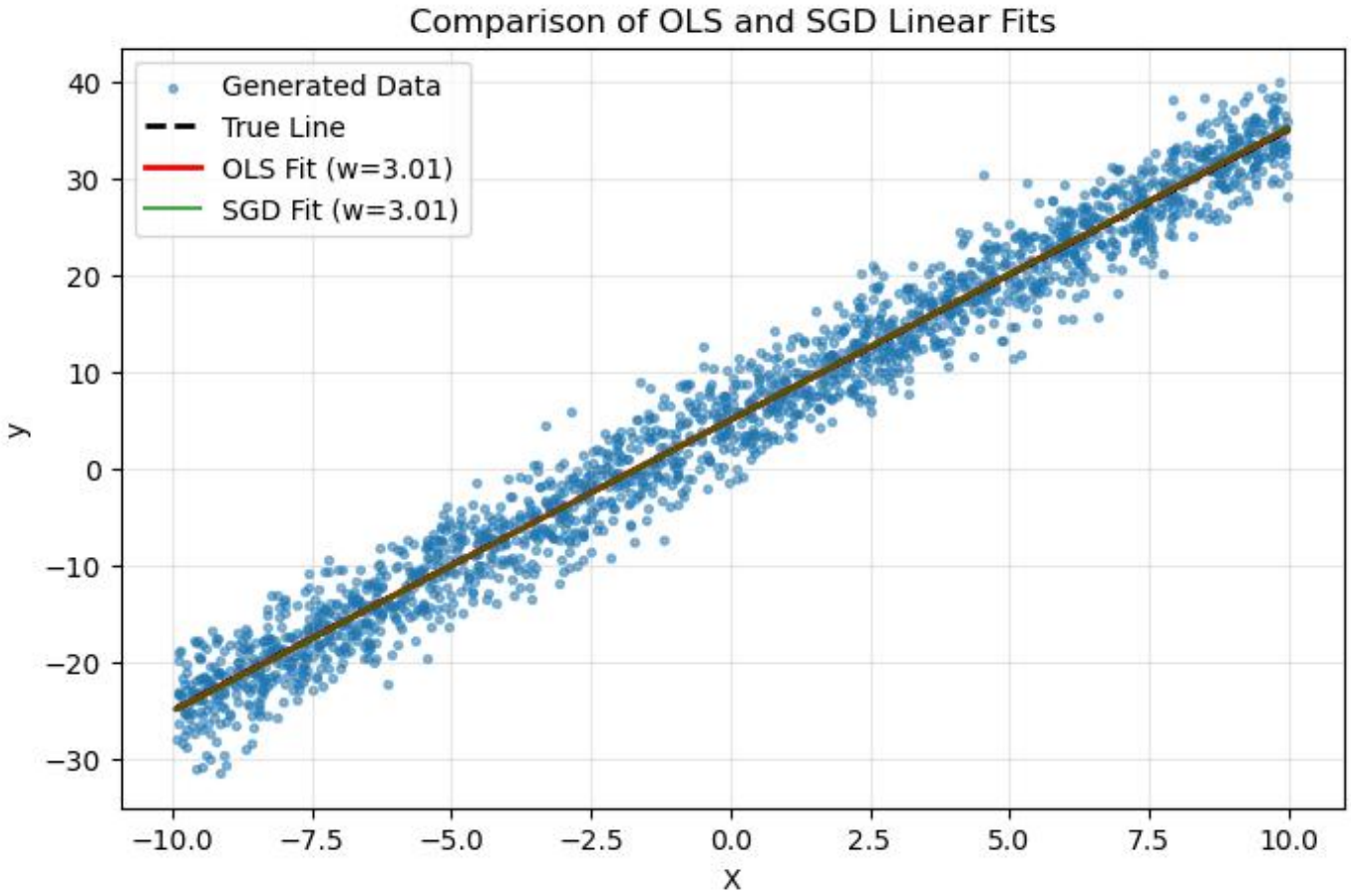
7.1 Objective

Compare:

- Closed-form OLS
- Gradient-based learning (SGD)

7.2 Result

Figure 7 – OLS vs SGD Fit:



Observations:

- Both methods converge to nearly identical slopes: w≈3.01w \approx 3.01w≈3.01
- This confirms that SGD can reliably approximate the optimal OLS solution
- SGD is more scalable for large datasets

8. CONCLUSIONS

From the experiments:

8.1 Classification

- Logistic regression effectively separates Versicolor and Virginica classes using petal measurements.

8.2 Regression

- OLS is extremely **non-robust to outliers**.
- RANSAC provides a significantly more stable and accurate fit under contamination.
- Ridge regression provides limited robustness but cannot fix the core issue.
- Outliers can cause large deviations in OLS slope even when the dataset is mostly clean.
- SGD performs similarly to OLS on clean datasets and is computationally efficient.

8.3 General Recommendation

For datasets with possible outliers or heavy-tailed noise use **RANSAC**, **Huber regression**, or **quantile regression** instead of plain OLS.

8.4 Notes

The full python notebook is attached with this report, all done under supervision and available to review and to adjust as per the needs of the observer, to conduct several tests, it is recommended to simply change the seed number variable to test different data sets.