LAB 1: Bias-Variance Decomposition with Linear and Polynomial Regression

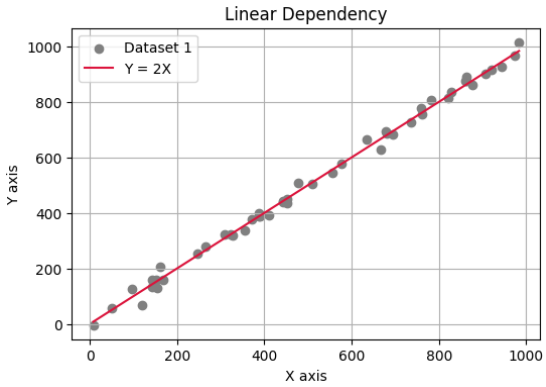
Task 1 - Data Generation

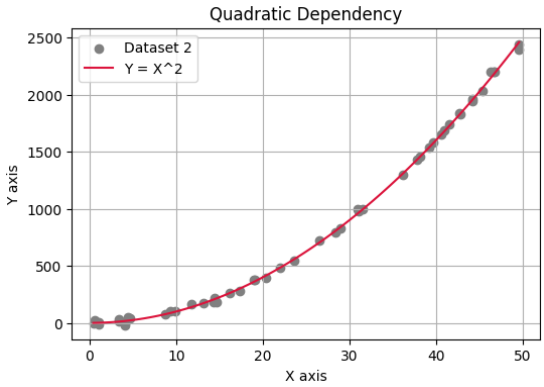
1. The data generation for the first dataset has the next characteristics:
- The quantity of records for dataset 1 is N = 50
- The linear dependency is stated by the function Y = X

2. The data generation for the second dataset has the next characteristics:
- The quantity of records for dataset 2 is N =50
- The quadratic dependency is stated by the function Y = X ^ 2

3. The gaussian noise was added to dataset 1 and 2. It was created using the normal distribution with mean = 0 and standard deviation = 20

Plots generated with datasets:





• What is the purpose of adding Gaussian noise?
The purpose of adding gaussian noise is to make the data more realistic. With noise we are trying to simulate perturbance and uncertainty that can occur on realistic data.

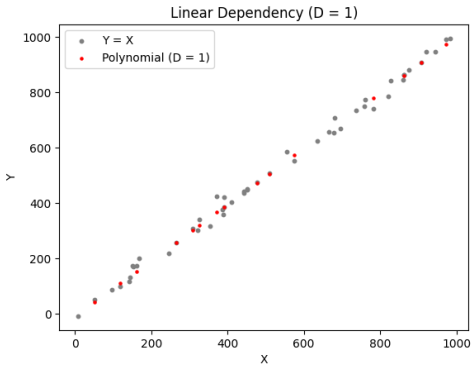
• Why is it important to visualize the data and the true functions?
Visualize the data and the true functions helps to validate that noise was generated properly and follows the relationship stablished by the true functions.

Task 2 - Model Training with Polynomial Regression

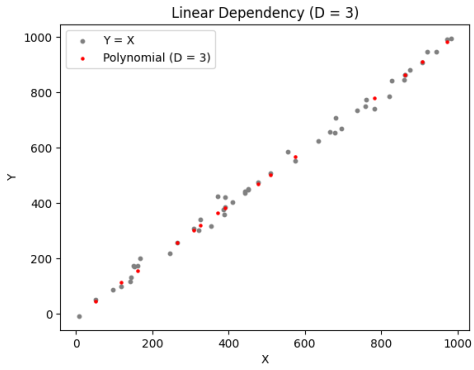
1. The polynomial models created for this tasks are with the next degrees: 3, 5, 7, 9, 12
2. Two Linear Regression models is trained using 70% of synthetic data. One Linear Regression is for the model with linear dependency, and one more for the model with quadratic dependency.
3. One plot was created per model and polynomial degree created.

Linear Relationship

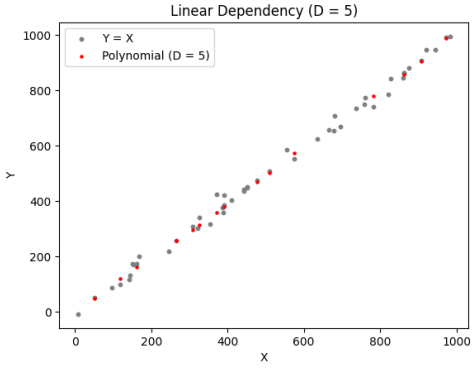
Degree: 1



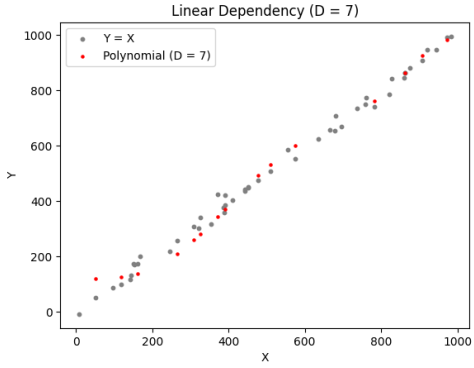
Degree: 3



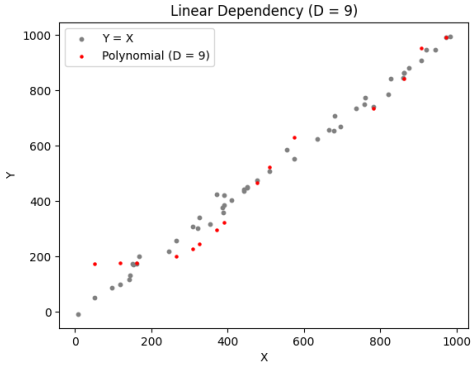
Degree: 5



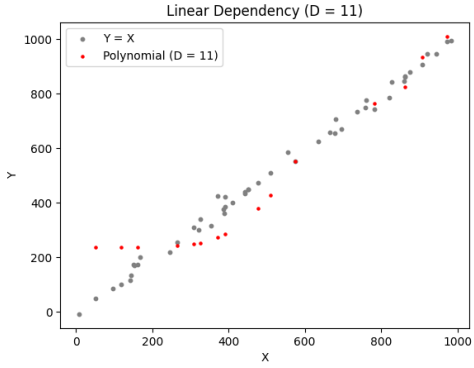
Degree: 7



Degree: 9

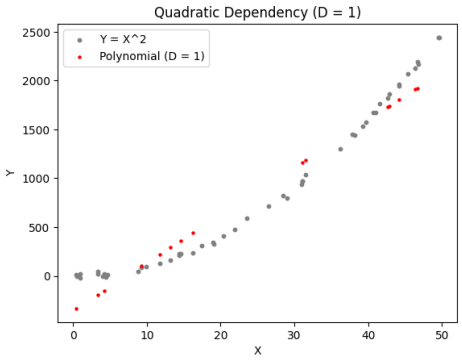


Degree: 11

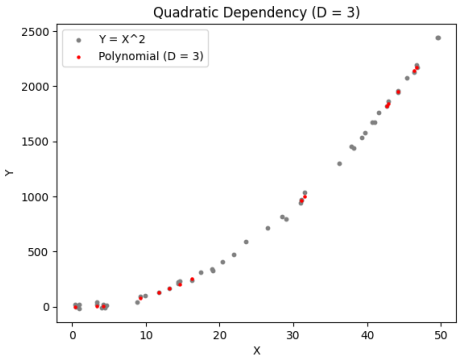


Quadratic Relationship

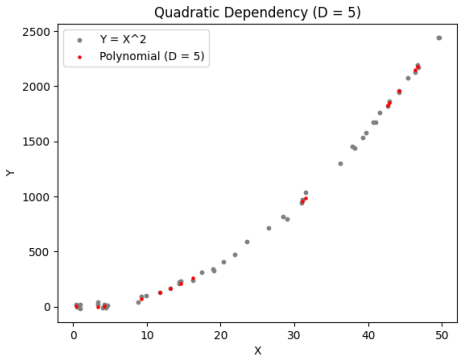
Degree: 1



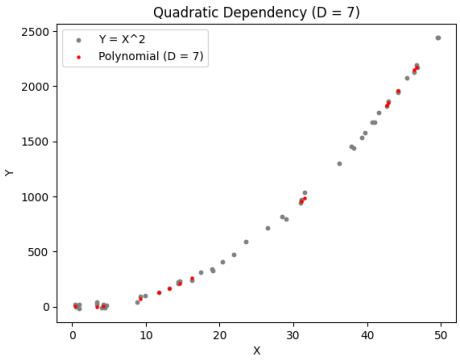
Degree: 3



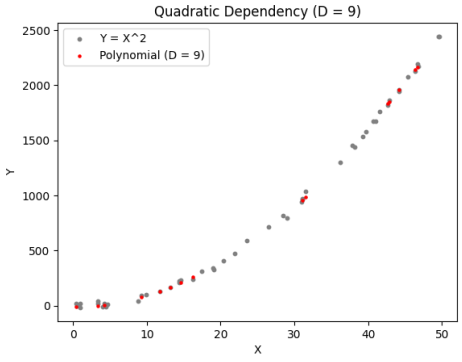
Degree: 5



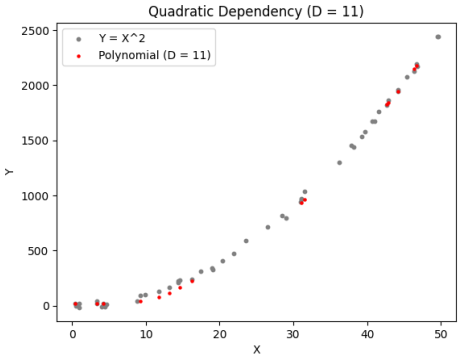
Degree: 7



Degree: 9



Degree: 11



In the model with linear relationship, it's evident that the model fits well when the degree is 1 and it becomes worse as the complexity increase. This tells us that is probable that bias increase as the complexity get higher.
In contrast, in the model with quadratic relationship, it's evident that the fits improves as the degree increases. This suggests that there could be an overfitting if the trained model fit's is not able to generalize the real model using the training data.

- What is a polynomial regression model and how is it different from a linear regression model?

The polynomial regression is a model where the relationship between the dependent variable and the independent variable is defined by a polynomial function. By comparison, the linear regression captures a straight line relationship between the dependent and independent variable.

- What do you expect to happen with the bias and the variance of polynomial regression models when the polynomial degree increases?

As the polynomial degree increases, the bias generally decreases because the model can represent more complex relationships. However, the variance increases, since a high degree polynomial can overfit the training data and become highly sensitive to small changes.

Task 3 - Bias Variance Decomposition

1. In order to calculate bias, variance, irreducible error and total error, it was used bootstrap with replacement to create multiple training datasets from the synthetic data. The number of bootstraps samples created is 200. These are the different results.

Linear Model

	Bias	Variance	Irreducible Error	Total Error
Degree: 1	29.7178	22.7199	501.0596	553.4973
Degree: 3	42.0205	46.3112	504.9486	593.2802
Degree: 5	109.3959	142.8910	542.5047	794.7916
Degree: 7	1039.0762	559.1025	837.1400	2435.3187
Degree 9	4920.1486	13169.1741	1713.5483	19802.8709
Degree 11	33777.3415	459704.4080	3152.7345	496634.4839

Quadratic Model

	Bias	Variance	Irreducible Error	Total Error
Degree: 1	33187.9615	3167.3365	1813.5240	38168.8220
Degree: 3	50.1865	40.4821	381.2410	471.9096
Degree: 5	61.3398	67.0461	434.2762	562.6621
Degree: 7	76.3146	285.3563	435.7597	797.4305
Degree 9	97.2389	768.6085	421.2653	1287.1127
Degree 11	1121.8834	12950.7369	296.6843	14369.3046

- Is it necessary to train multiple models to estimate the bias, variance and irreducible error?

Yes, it's necessary because you need to calculate the expected predictions using the multiple trained models. This also help with the accuracy of the results.

- Is it possible to estimate the model's bias without an access to the true regression function?

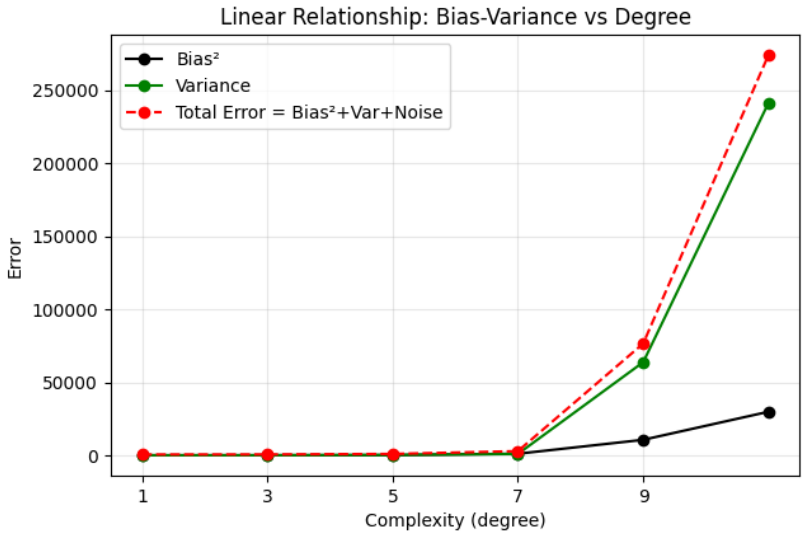
Yes, as an estimation, using some resampling techniques. But strictly, it's not possible to calculate it if the true function is unknown.

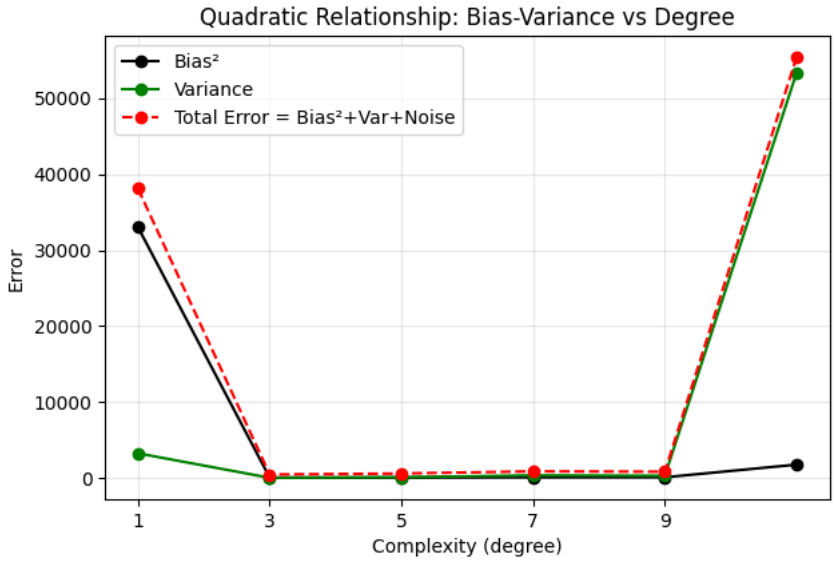
- Is it possible to estimate the model's variance without an access to the true regression function?

Yes, because the variance is calculated over the different predictions in the test dataset.

Task 4 - Visualization and Analysis

After training multiple models with different degrees for the linear and quadratic model, this plot helps to confirm how bias and variance are affected by increasing the complexity of the polynomial.





- Can you describe the trade-offs between the variance for polynomial regression models on the linear and quadratic datasets?

For the linear dataset, a low degree polynomial fits better the relationship between X and Y, so bias and variance are low. As the degree increases, the bias remains low but variance become high because it starts to overfit the noisy data.

For the quadratic dataset, a polynomial model of degree 1 produces a high bias because the model is not able to capture the curvature of the relationship. However, as we continue increasing the degree beyond 2, the bias remains low and the variance become high, again, because it overfits noisy data.

- How and why do the bias-variance trade-offs differ between linear and quadratic datasets?

For the linear dataset, with a degree of 1 the true relationship is captured by the trained model. As the degree increases the bias remains low but the variance increases because it starts to overfit due to noisy data.
For the quadratic dataset, with a degree of 1 the model is too simple and the true relationship is not captured, the bias is high and variance is low, here the model underfit the true relationship. When the degree is 3, the model fits better the true relationship, here the bias and variance are low. As degree goes beyond 2 or 3 the variance increases and the bias remains low which means that the trained model overfit the noisy data.

- Using the bias-variance trade-offs analysis (based on the previous two questions), which polynomial degree would you choose for the linear and quadratic datasets to best balance the bias and the variance?

For the linear dataset, the model with a polynomial degree of 1 is the best choice since captures the true relationship with low bias and variance.
For the quadratic dataset, the best model would be with a polynomial degree of 2 (or 3 since no polynomial model with degree 2 was built) because fits the true relationship with low bias and variance.

- Discuss the concept of overfitting/underfitting in the context of polynomial regression. How can bias-variance analysis help in identifying overfitting/underfitting?

Underfitting occur when the trained model is too simple and is not able to capture the true relationship between the dependent and independent variables.
Overfitting occur when the trained model became to sensitive to the trained data and doesn't generalize well the true relationship between the variables.
In the bias-variance analysis, underfitting can be detected when bias is high and variance is low. In contrast, overfitting can be detected when variance is high and bias is low.