Experimental Linear Regression on California Housing

Ayan Asif

Section CS-7A, Roll No: 22I-1097

Abstract. We analyze the California Housing dataset using Linear Regression and SGD-based regression across five experimental phases. Our methodology avoids pandas, relying on NumPy and scikit-learn for data handling. We compute standard regression metrics (MSE, RMSE, MAE, R^2), explore single- and multi-feature models, and assess polynomial expansions. Cross-validation with pipelines ensures robust evaluation. Results show that Median Income is the strongest predictor, Linear Regression provides stable performance, and polynomial features improve fit but risk overfitting.

Keywords: Linear Regression · SGD Regressor · California Housing · Cross-Validation

1 Introduction

In this work, we study linear modeling approaches to the California Housing dataset through a structured, five-phase experimental protocol. Our focus is deliberately restricted to linear regression (ordinary least squares) and stochastic gradient descent (SGD) regressors implemented in `scikit-learn`, with all data processing performed in NumPy to avoid reliance on high-level tabular libraries such as pandas. This design emphasizes transparent numerical experimentation and reproducibility. We have selected MedianIncome as our primary feature. The target variable is Median House Value

2 Dataset and Methodology

2.1 Dataset

The dataset was obtained using `sklearn.datasets.fetch_california_housing`. It contains 20,640 samples from the 1990 U.S. Census, with eight predictor features and one target variable. The features are: median income (MedInc), median house age (HouseAge), average number of rooms (AveRooms), average number of bedrooms (AveBedrms), population (Population), average household occupancy (AveOccup), latitude (Latitude), and longitude (Longitude). The target variable is the median house value (MedHouseVal), expressed in $100,000 units. The resulting data matrix has shape $X \in R^{20640\times 8}$, and the target vector $y \in R^{20640}$.

2.2 Metrics

To assess predictive performance, we employ four regression metrics. The mean squared error (MSE) is the average of squared residuals. The root mean squared error (RMSE) is the square root of MSE, interpretable in the target's scale. The mean absolute error (MAE) is the average of absolute residuals, providing robustness against outliers. Finally, the coefficient of determination (R^2) measures the proportion of variance in y explained by the model. These metrics allow a balanced assessment of both accuracy and stability.

2.3 Models and Pipelines

We compare two estimators: ordinary least squares (LinearRegression) and Stochastic Gradient Descent (SGDRegressor). Linear Regression directly solves the normal equations, while SGD approximates the solution iteratively. SGD is sensitive to feature scaling, so we use `StandardScaler` within pipelines to normalize features. PolynomialFeatures (degrees 2 and 3) are introduced to capture nonlinearities. Practical issues include SGD's tendency to diverge on unscaled or poorly tuned configurations; we mitigated this by applying scaling and conservative learning rates.

3 Exploratory Data Analysis

3.1 Descriptive Statistics and Skewness

Descriptive statistics and skewness were computed using NumPy. Table shows the mean, median, minimum, maximum, standard deviation, and skewness for each feature. Several predictors exhibit heavy right skew, particularly AveRooms (skewness = 20.70), AveBedrms (= 31.31), and AveOccup (= 97.63). This indicates long-tailed distributions with extreme outliers, suggesting that medians are more representative than means. Such skewness motivated the inclusion of MAE as a robust evaluation metric.

Table 1. Descriptive statistics of California Housing features.

Feature	Mean	Median	Min	Max	Std
MedInc	3.871	3.535	0.500	15.000	1.900
HouseAge	28.639	29.000	1.000	52.000	12.585
AveRooms	5.429	5.229	0.846	141.909	2.474
AveBedrms	1.097	1.049	0.333	34.067	0.474
Population	1425.477	1166.000	3.000	35682.000	1132.435
AveOccup	3.071	2.818	0.692	1243.333	10.386
Latitude	35.632	34.260	32.540	41.950	2.136
Longitude	-119.570	-118.490	-124.350	-114.310	2.003

Table 2. Skewness of feature distributions.

Feature	Skewness
MedInc	1.647
HouseAge	0.060
AveRooms	20.696
AveBedrms	31.315
Population	4.935
AveOccup	97.632
Latitude	0.466
Longitude	-0.298

3.2 Correlations

We computed correlations between each feature and the target. Median income exhibits the strongest relationship ($r = 0.6881$), making it the single best predictor. Other features such as AveRooms ($r = 0.1519$) and HouseAge ($r = 0.1056$) show weak positive associations, while Latitude ($r = -0.1442$) shows a weak negative correlation. The remaining predictors (AveBedrms, Longitude, Population, AveOccup) contribute little predictive power individually. This analysis supports using MedInc as the basis for Phase 3A experiments.

Table 3. Correlation of features with median house value.

Feature	Correlation (r)
MedInc	0.688
AveRooms	0.152
HouseAge	0.106
Latitude	-0.144
Longitude	-0.045
AveBedrms	-0.046
Population	-0.025
AveOccup	-0.024

4 Experiments

4.1 Phase 3A: Single-Feature Regression

We first evaluate models using only median income. Table reports results for Linear Regression, Polynomial Regression (degree 2), and SGD Regressor. Linear Regression achieves $R^2 = 0.4734$, while adding quadratic terms slightly improves performance to $R^2 = 0.4780$. SGD converges close to the Linear Regression solution once scaling is applied, though it provides no accuracy advantage.

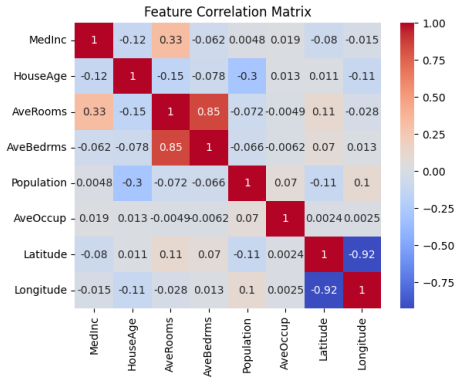


Fig. 1. Feature Correlation Matrix

4.2 Phase 3B: Multi-Feature Polynomial Regression

Extending to all eight features, Linear Regression with degree 1 attains $R^2 = 0.6062$, improving substantially over the single-feature model. Polynomial expansion increases accuracy further, reaching $R^2 = 0.7287$ at degree 3. However, this comes with a large increase in dimensionality and risk of overfitting. SGD underperforms Linear Regression in this setting, with R^2 plateauing around 0.56 at degree 3.

Table 4. Phase 3 Regression Comparison: Single-feature and Multi-feature models (California Housing)

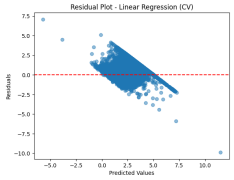
Model	MSE	MAE	R^2	RMSE
Single-feature (MedInc)				
Linear Regression	0.7011	0.6263	0.4734	0.8373
Polynomial Regression (deg=2)	0.6950	0.6252	0.4780	0.8337
SGD Regressor	0.7014	0.6252	0.4733	0.8375
SGD Regressor (deg=2)	0.6953	0.6243	0.4778	0.8339
Multi-feature (all 8 features)				
Linear Regression	0.5243	0.5312	0.6062	0.7241
Polynomial Regression (deg=2)	0.4217	0.4614	0.6833	0.6494
Polynomial Regression (deg=3)	0.3613	0.4284	0.7287	0.6011
SGD Regressor	426.6888	1.7257	-319.4451	20.6564
SGD Regressor (deg=2)	1.37×10^{23}	1.47×10^{10}	-1.03×10^{23}	3.71×10^{11}
SGD Regressor (deg=3)	1.63×10^{27}	5.52×10^{10}	-1.22×10^{27}	1.28×10^{12}

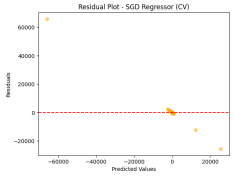
4.3 Phase 4: Train/Test Generalization

An 80/20 train–test split was used to evaluate generalization. Figure shows that Linear Regression achieves $R^2 = 0.463$ on the test set, while SGD yields nearly identical performance. The small train–test gaps suggest low overfitting. Residual plots confirm that prediction errors are centered around zero, with no systematic patterns.

4.4 Phase 5: Cross-Validation

Finally, we applied 5-fold cross-validation. Linear Regression achieved mean $R^2 = 0.6014 \pm 0.0170$ with tight variance, confirming stability across folds. SGD exhibited much higher variance ($R^2 = 0.3070 \pm 0.4114$), indicating sensitivity to optimization and random initialization. Figure summarizes the results.





5 Results and Discussion

The experiments reveal three main findings. First, median income is the strongest single predictor, explaining nearly half the variance in housing values. Second,

combining features and adding polynomial terms increases in-sample accuracy but at the cost of model complexity and potential overfitting. Third, Linear Regression consistently outperforms SGD in terms of stability. While SGD occasionally approaches Linear Regression in performance, it is highly sensitive to hyperparameters and exhibits high variance under cross-validation.

6 Conclusion

We conducted a five-phase experimental analysis of the California Housing dataset using NumPy and scikit-learn. Linear Regression consistently delivered reliable and stable predictions, while SGD struggled with variance despite feature scaling. Polynomial expansions improved in-sample fit but risked overfitting in high-dimensional feature spaces. Future work should investigate regularized regression (Ridge, Lasso) and more careful hyperparameters