Impact of Dimensionality Reduction and Optimization Methods on Linear Regression Performance:

A Comprehensive Analysis Using Life Expectancy Data

Mathematical Foundation for AI

Group Members:

Muhammad Azhar (24k-7606)

Hamza Mughal (25k-7623)

November 23, 2025

Abstract

This report presents a comprehensive comparison of linear regression methods for predicting life expectancy using WHO Global Health Observatory data (2000-2015, 2,928 observations from 193 countries). We implement and evaluate four approaches: Ordinary Least Squares (OLS), Singular Value Decomposition (SVD), Gradient Descent (GD), and Principal Component Analysis (PCA) for dimensionality reduction. Results show that OLS and SVD produce numerically identical solutions (coefficient difference $< 10^{-14}$) with excellent performance ($R^2 = 0.819$, RMSE = 3.95 years). Gradient Descent converges to 99.9% optimal performance in 400 iterations, while PCA achieves 40% dimensionality reduction (20→12 components) with less than 1% performance loss. Key predictors include child mortality rates (coefficient: -11.34), schooling (+2.13), and HIV/AIDS prevalence (-2.48). The condition number analysis ($\kappa \approx 1,962$) confirms numerical stability. Our findings demonstrate that SVD offers superior stability without computational penalty, GD scales efficiently for large datasets, and PCA enables effective regularization. All implementations use from-scratch NumPy code with comprehensive visualizations and reproducible workflows.

Linear Regression Project Report

Implementation of OLS, SVD, and Gradient Descent

Mathematical Foundation for AI

November 23, 2025

Dataset: Life Expectancy Data (WHO & United Nations)
Available at Kaggle:
https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who/data

Contents

1 Introduction

1.1 Dataset Description and Problem Formulation

Dataset: WHO Global Health Observatory & UN (2000-2015), 193 countries, 2,928 obs $\times$ 20 features.

Target: Life Expectancy (years).

Features: Mortality (adult/infant/under-five deaths, HIV/AIDS), Immunization (Hepatitis B, Polio, Diphtheria), Economic (GDP, health expenditure, HDI), Social (schooling, alcohol, BMI, development status).

Split: 80/20 train/test (2,342/586), seed=42.

Objectives: Compare OLS, SVD, and Gradient Descent for life expectancy prediction. Analyze numerical stability, dimensionality reduction via PCA, and identify key predictors.

Preprocessing: Median/mode imputation, label encoding, StandardScaler ($\mu = 0$, $\sigma = 1$).

2 Methodology

2.1 Implementation of Each Method and Theoretical Background

OLS: Closed-form $\hat{\beta} = (X^T X)^{-1} X^T y$. Fast (0.9 ms), condition number 1.96×10^3 (well-conditioned).

SVD: Pseudoinverse via $X = U\Sigma V^T$, $\hat{\beta} = V\Sigma^+ U^T y$. Stable (condition 9.98×10^2), handles rank deficiency, 3.2 ms. Coefficients identical to OLS (diff $< 10^{-14}$).

Gradient Descent: Iterative $\beta_{k+1} = \beta_k - \eta \nabla L(\beta_k)$ where $L(\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - X_i^T \beta)^2$. Variants: Batch, Mini-batch (best, 32 samples, $\eta = 0.01$), Stochastic, Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$). Converges at $\sim$400 iterations, 450 ms.

PCA: SVD on centered data: $X_c = U\Sigma V^T$, $\text{Var}_k = \sigma_k^2 / \sum \sigma_i^2$. Optimal: k = 12-15 captures 95% variance, 99.7% performance, 8.5 ms.

3 Results

3.1 Tables/Plots Comparing OLS, SVD, Gradient Descent, PCA, and Ridge

3.1.1 Performance Comparison

Table 1: Performance Summary of All Methods

Method	Train MSE	Test MSE	Train R^2	Test R^2	Train RMSE	Test RMSE	Time
OLS	16.48	15.62	0.820	0.819	4.06	3.95	0.9 ms
SVD	16.48	15.62	0.820	0.819	4.06	3.95	3.2 ms
Gradient Descent	16.55	15.70	0.819	0.819	4.07	3.96	450 ms
PCA (k=12)	16.89	15.89	0.816	0.816	4.11	3.99	8.5 ms

Key Findings:

- OLS and SVD produce identical results (coefficient diff $< 10^{-14}$)
- Gradient Descent achieves 99.9% of optimal performance
- PCA offers 40% dimensionality reduction with $< 1\%$ performance loss

3.1.2 PCA Dimensionality Analysis

Table 2: PCA Performance vs Number of Components

Components	Cumulative Variance	Test R^2	Dimensionality Reduction
1	31.2%	0.655	95.0%
5	71.5%	0.789	75.0%
12	94.8%	0.816	40.0%
20	100.0%	0.819	0.0%

3.1.3 Feature Importance (Top 10)

Table 3: Top 10 Most Important Features

Rank	Feature	Coefficient	Impact
1	Under-five deaths	-11.34	Strong negative
2	Infant deaths	+11.19	Multicollinearity
3	Adult Mortality	-2.55	Negative
4	HIV/AIDS	-2.48	Negative
5	Schooling	+2.13	Positive
6	Income composition	+1.08	Positive
7	Diphtheria coverage	+0.97	Positive
8	BMI	+0.84	Positive
9	Polio coverage	+0.72	Positive
10	GDP	+0.55	Positive

3.1.4 Visualizations

All visualizations are available in the Notebook.

4 Discussion

4.1 Insights, Limitations, and Takeaways

Key Insights: (1) OLS/SVD numerically identical for well-conditioned data; SVD more stable without speed penalty. (2) GD scales better (O(d) vs $O(d^2)$ memory), achieves 99.9% optimal performance. (3) PCA: 40% reduction, $< 1\%$ performance loss. (4) Mortality factors (infant/child deaths), education (schooling +2.13), and socioeconomic factors dominate predictions. R^2=0.82, RMSE=3.95 years shows strong fit without overfitting.

Limitations: (1) Linearity assumption may miss non-linear patterns. (2) Limited feature engineering; interactions unexplored. (3) Imputation bias from 10 removed samples. (4) Temporal autocorrelation ignored. (5) Correlation $\neq$ causation (reverse causality possible).

Key Takeaways: Use SVD as default for stability. Use GD for large-scale/streaming data. Apply PCA when overfitting suspected. Always check condition numbers, validate on test set, and scale features.

5 Conclusion

5.1 Summary of Findings and Possible Extensions

Summary: All methods achieved $R^2 \approx 0.82$. OLS/SVD identical (diff$< 10^{-14}$), GD converged to near-optimal, PCA enabled 40% reduction retaining 99% performance. Top predictors: mortality (infant/child deaths), education (schooling), socioeconomic factors.

Learning Outcomes:

- ✓ OLS with stability analysis
- ✓ SVD robustness comparison
- ✓ Multiple GD variants
- ✓ PCA dimensionality reduction
- ✓ Analytical vs iterative methods comparison

Extensions: *Methodological:* Ridge/Lasso regularization, polynomial features, ensemble methods, k-fold CV. *Analytical:* Feature selection, interaction terms, residual analysis, time series. *Domain:* Causal inference, panel data models, spatial analysis, policy simulation.

Technical Appendix

Software: Python 3.13, NumPy 1.26+, Pandas 1.3+, Scikit-learn 1.0+, Matplotlib/Seaborn, TensorBoard

Reproducibility: Fixed random seed (42), automated execution via `main.py`, modular structure

Execution:

```
pip install -r requirements.txt
python main.py
tensorboard --logdir=runs
```

Performance: Total execution time: 6.24 seconds

Test Set Performance: $R^2 = 0.819$, RMSE = 3.95 years
Dataset: 2,928 observations $\times$ 20 features

References

[1] World Health Organization (WHO). (2015). *Global Health Observatory data repository: Life expectancy and mortality data*. Retrieved from https://www.who.int/data/gho

[2] United Nations. (2015). *World Population Prospects*. Department of Economic and Social Affairs, Population Division.

[3] Kumar, A. (2018). *Life Expectancy (WHO) Dataset*. Kaggle. https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who/data

[4] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.

[5] Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.