Python for AI

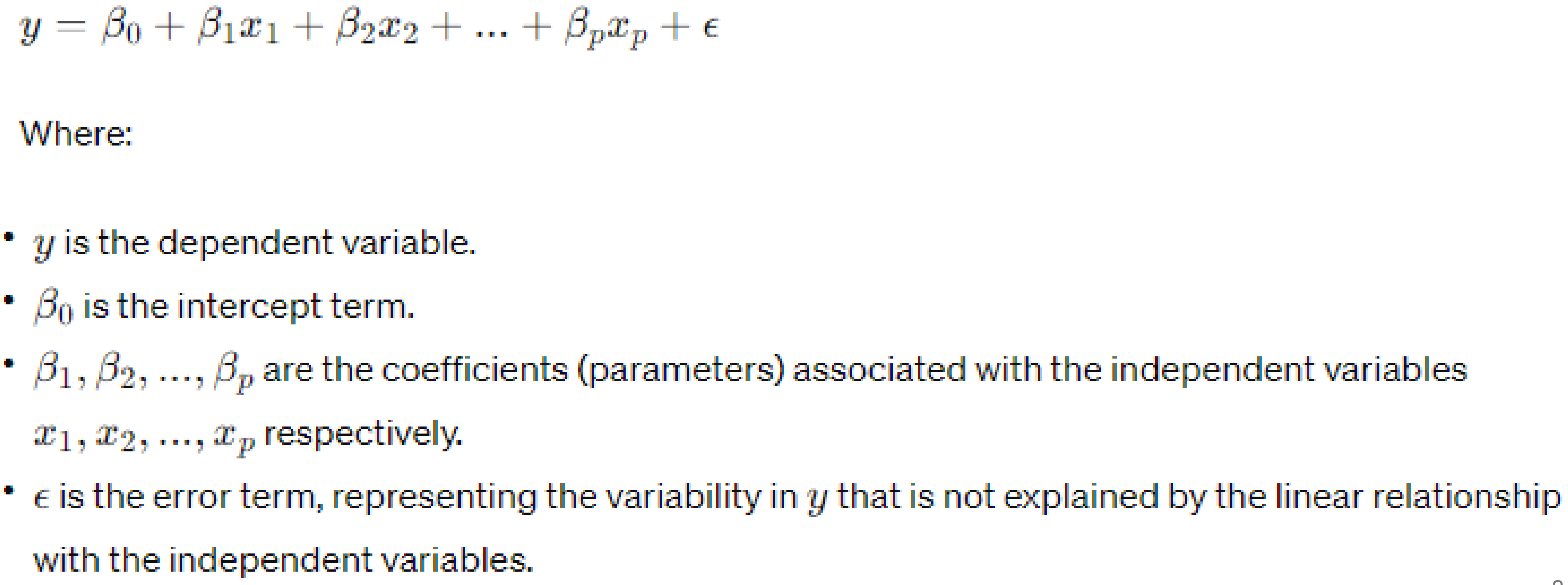
Rina BUOY

AMERICAN UNIVERSITY
OF PHNOM PENH

STUDY LOCALLY. LIVE GLOBALLY.

Linear Regression

- A statistical model used to describe the relationship between one or more independent variables (inputs or predictors) and a dependent variable (output or response).



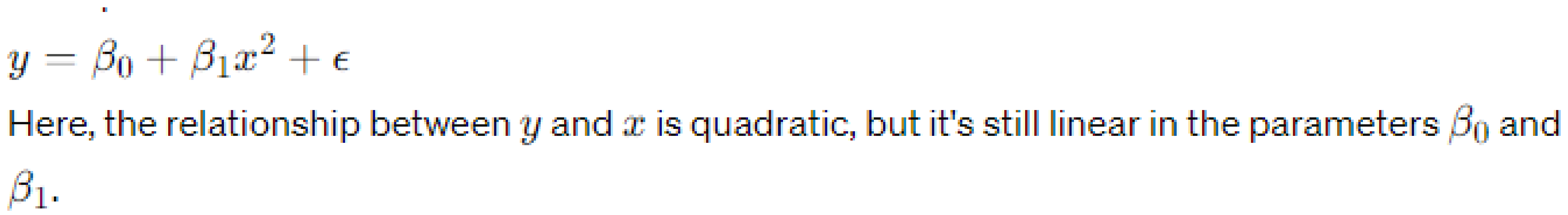
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \epsilon$$

Where:

- y is the dependent variable.
- β_0 is the intercept term.
- $\beta_1, \beta_2, ..., \beta_p$ are the coefficients (parameters) associated with the independent variables $x_1, x_2, ..., x_p$ respectively.
- ϵ is the error term, representing the variability in y that is not explained by the linear relationship with the independent variables.

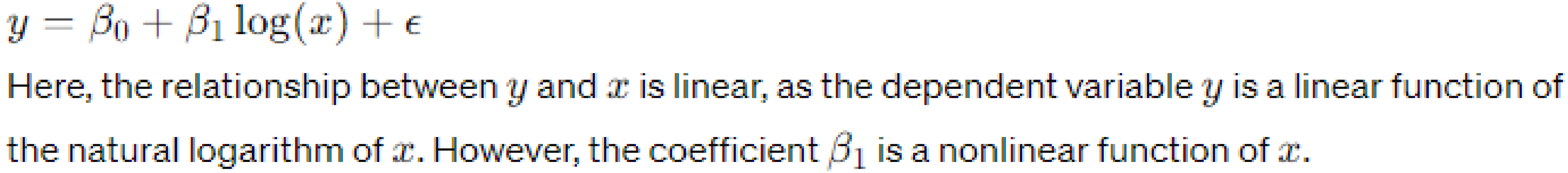
Linear Regression

- The model assumes a linear relationship between the independent variables and the dependent variable in parameters



$$y = \beta_0 + \beta_1 x^2 + \epsilon$$

Here, the relationship between y and x is quadratic, but it's still linear in the parameters β_0 and β_1.



$$y = \beta_0 + \beta_1 \log(x) + \epsilon$$

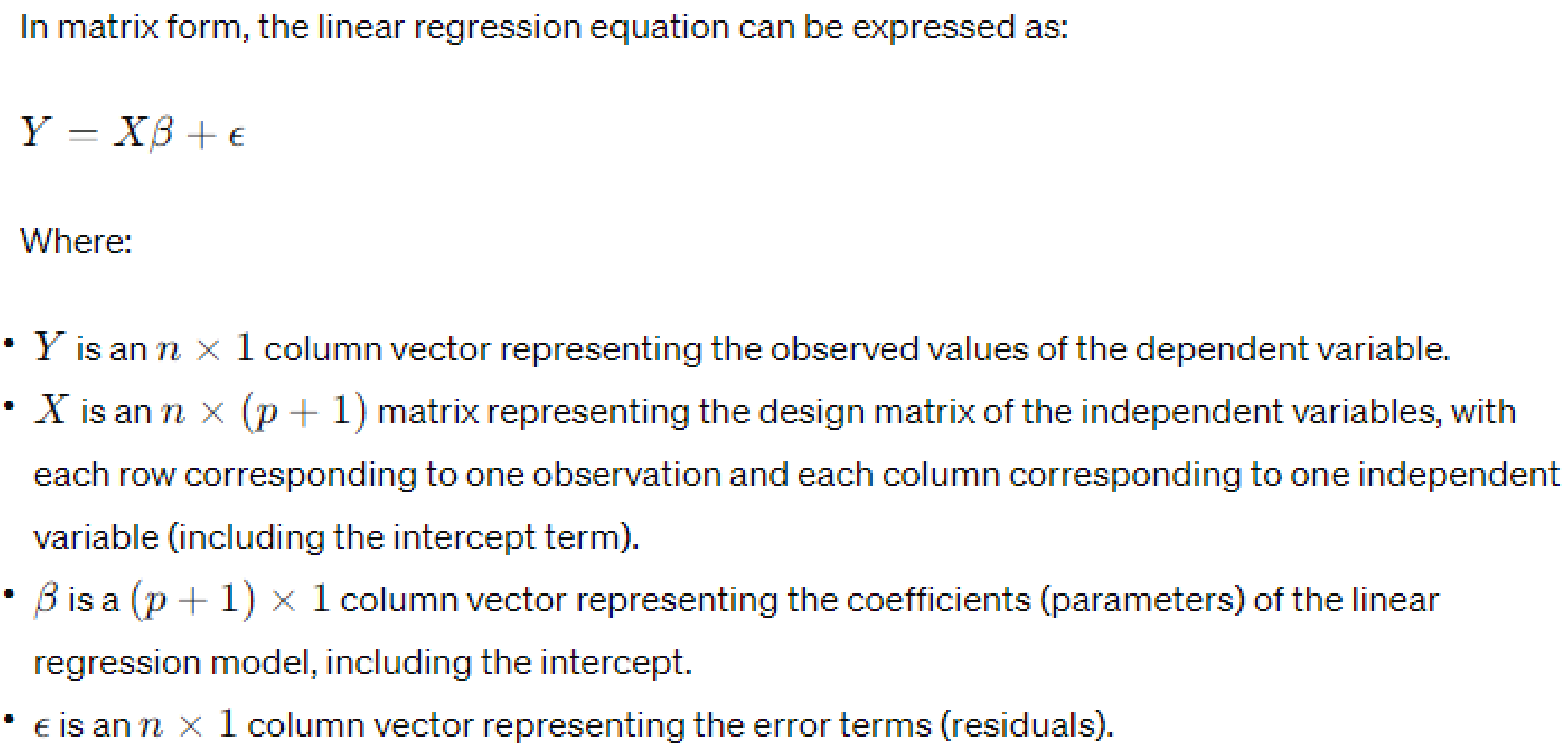
Here, the relationship between y and x is linear, as the dependent variable y is a linear function of the natural logarithm of x. However, the coefficient β_1 is a nonlinear function of x.

- Thus, a polynomial regression is an extension of a LR model.
- With proper feature engineering, a LR model can be extremely powerful and explainable.

Assumptions

1. **Linearity**: The relationship between the independent variables (predictors) and the dependent variable (response) is linear. This means the change in the dependent variable is proportional to the change in the independent variables.
2. **Independence**: The observations in the dataset are independent of each other. This means there should be no autocorrelation among the observations. In other words, the error terms (residuals) are uncorrelated with each other.
3. **Homoscedasticity**: The variance of the error terms is constant across all levels of the independent variables. In simpler terms, the spread of the residuals should be uniform along the range of predicted values.
4. **Normality of Residuals**: The error terms (residuals) are normally distributed. This assumption typically applies when the sample size is large enough due to the Central Limit Theorem. Normality of residuals ensures the accuracy of confidence intervals and hypothesis tests.
5. **No Perfect Multicollinearity**: There should be no exact linear relationship among the independent variables (multicollinearity). High multicollinearity can lead to unstable parameter estimates, making it difficult to interpret the effects of individual predictors.
6. **No Endogeneity**: The independent variables are not correlated with the error term. Violation of this assumption leads to biased parameter estimates.

Linear Regression in Matrix Form



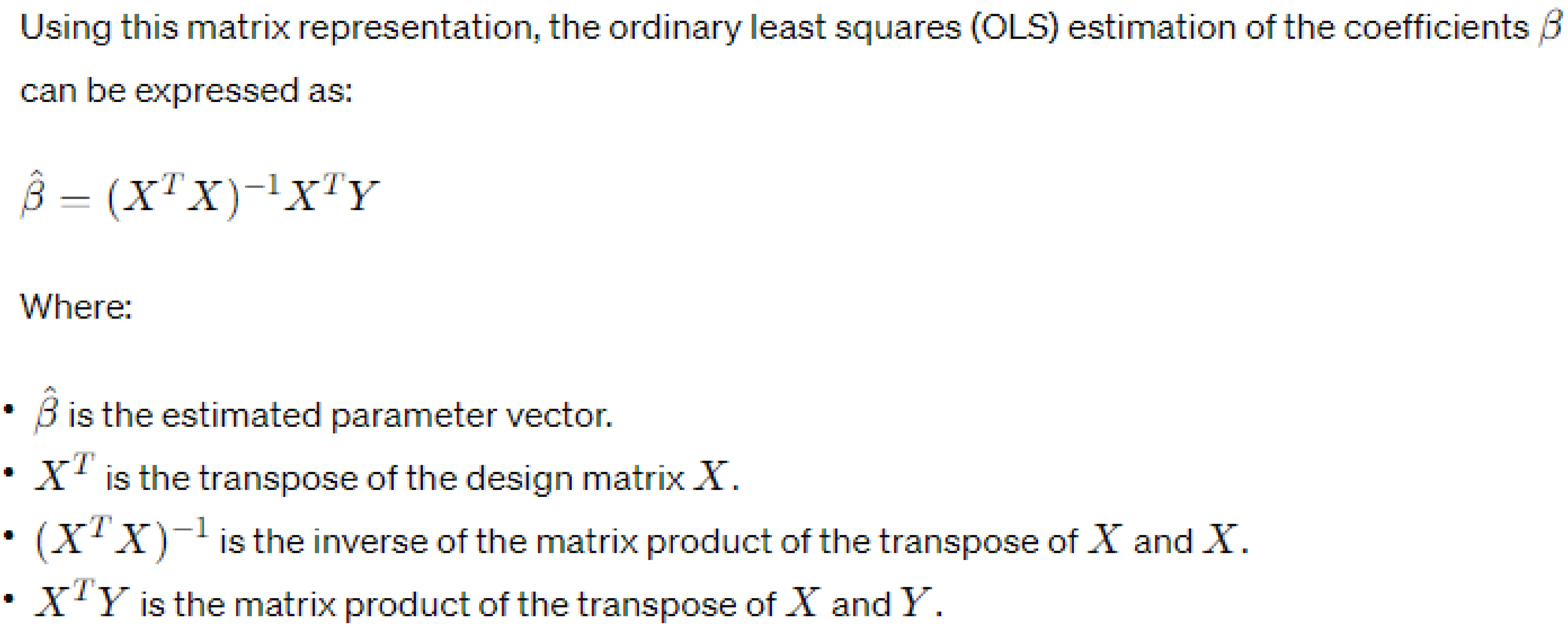
In matrix form, the linear regression equation can be expressed as:

$$Y = X\beta + \epsilon$$

Where:

- Y is an $n \times 1$ column vector representing the observed values of the dependent variable.
- X is an $n \times (p+1)$ matrix representing the design matrix of the independent variables, with each row corresponding to one observation and each column corresponding to one independent variable (including the intercept term).
- β is a $(p+1) \times 1$ column vector representing the coefficients (parameters) of the linear regression model, including the intercept.
- ϵ is an $n \times 1$ column vector representing the error terms (residuals).

Ordinary Least Squares



Using this matrix representation, the ordinary least squares (OLS) estimation of the coefficients β can be expressed as:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

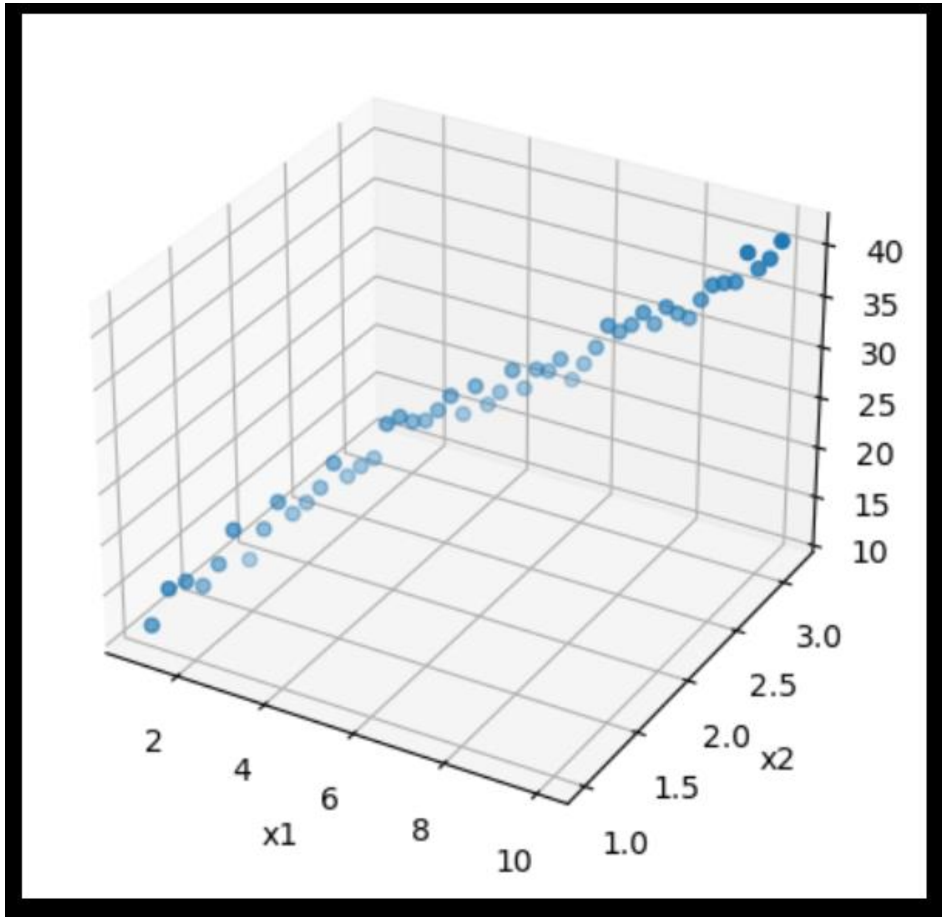
Where:

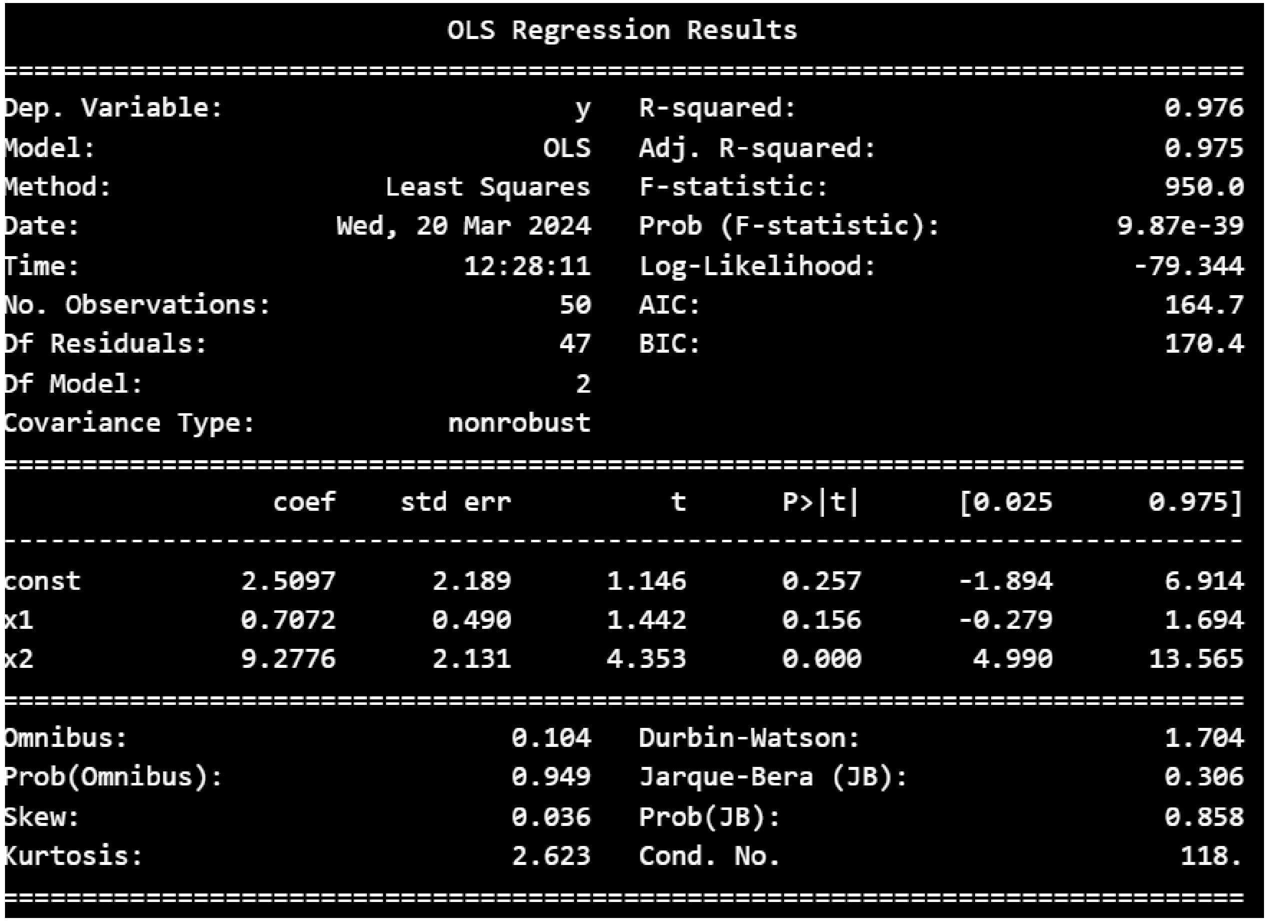
- $\hat{\beta}$ is the estimated parameter vector.
- X^T is the transpose of the design matrix X.
- $(X^T X)^{-1}$ is the inverse of the matrix product of the transpose of X and X.
- $X^T Y$ is the matrix product of the transpose of X and Y.

Statsmodels

1. Import necessary libraries.
2. Load your dataset.
3. Define your independent and dependent variables.
4. Add a constant term to the independent variables to account for the intercept. (sm.add_constant(X))
5. Create and fit the linear regression model. (sm.OLS(y, X_with_const).fit())
6. Analyze the results (Beta, R-squared, P-values etc.). (model.summary())

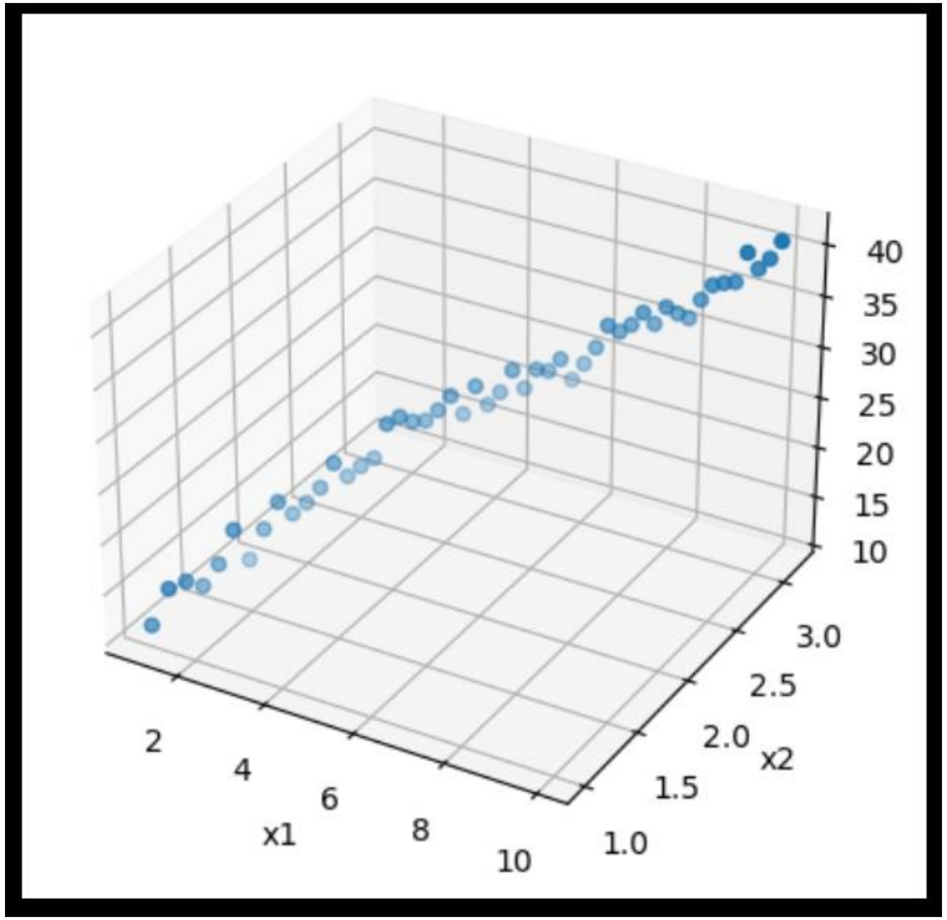
Example, $y = \beta_0 + \beta_1 x_1 + \beta_2 \sqrt{x_1}$

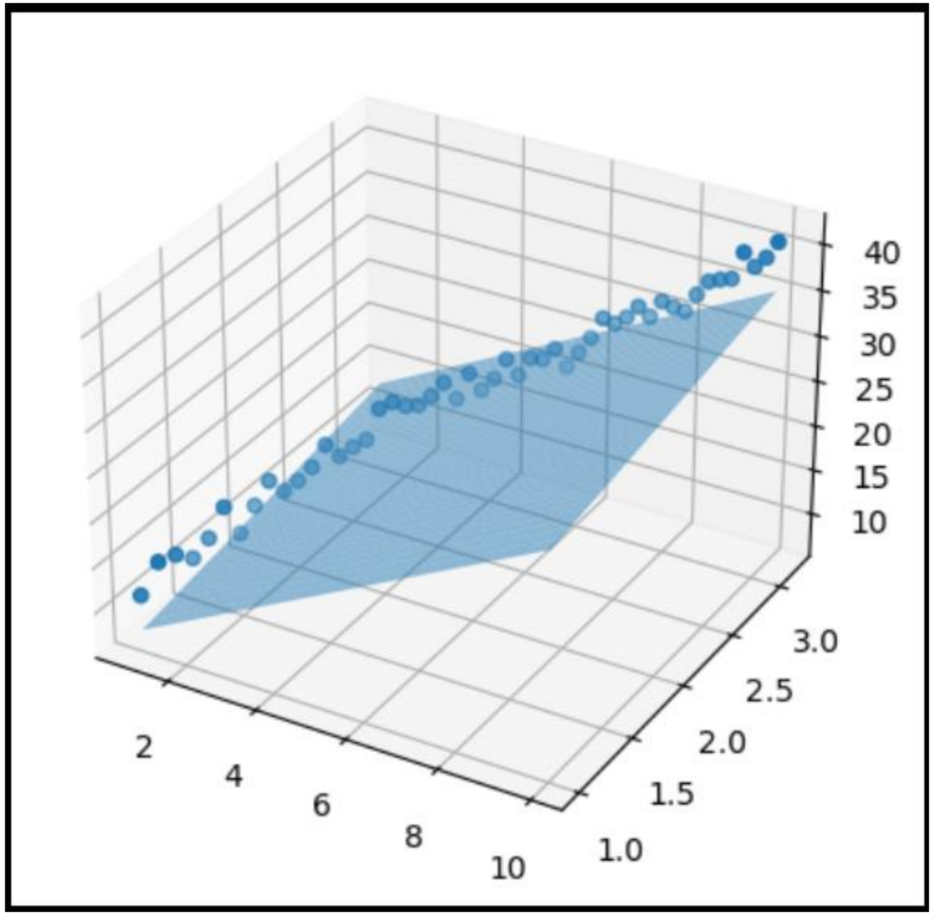




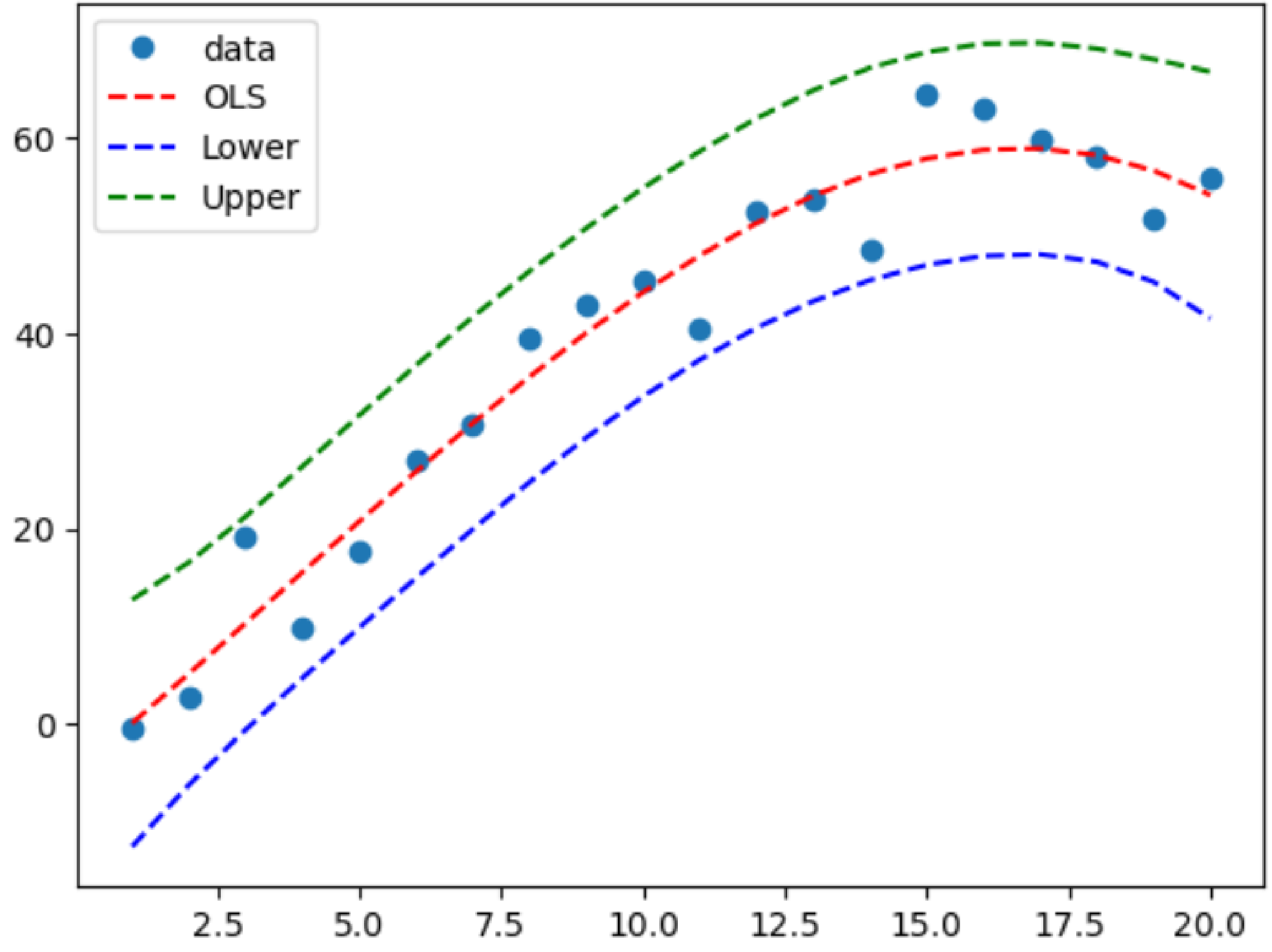
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.976
Model:                            OLS   Adj. R-squared:                  0.975
Method:                 Least Squares   F-statistic:                     950.0
Date:                Wed, 20 Mar 2024   Prob (F-statistic):           9.87e-39
Time:                        12:28:11   Log-Likelihood:                -79.344
No. Observations:                  50   AIC:                             164.7
Df Residuals:                      47   BIC:                             170.4
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          2.5097      2.189      1.146      0.257      -1.894       6.914
x1             0.7072      0.490      1.442      0.156      -0.279       1.694
x2             9.2776      2.131      4.353      0.000       4.990      13.565
==============================================================================
Omnibus:                        0.104   Durbin-Watson:                   1.704
Prob(Omnibus):                  0.949   Jarque-Bera (JB):                0.306
Skew:                           0.036   Prob(JB):                        0.858
Kurtosis:                       2.623   Cond. No.                         118.
==============================================================================
```

Example, $y = \beta_0 + \beta_1 x_1 + \beta_2 \sqrt{x_1}$

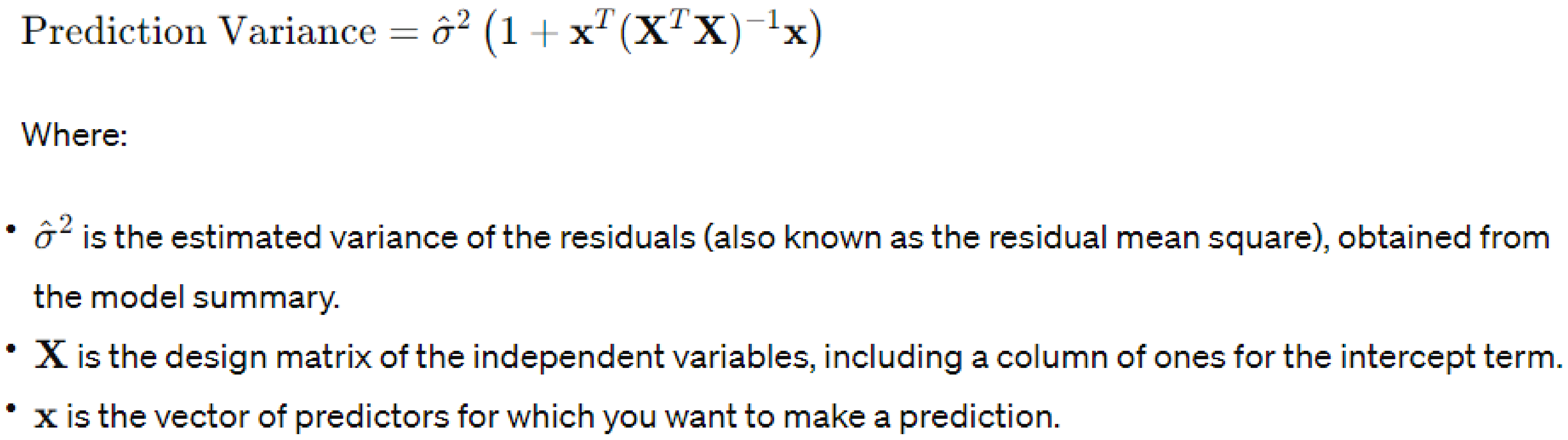




Prediction Bands



Prediction Bands



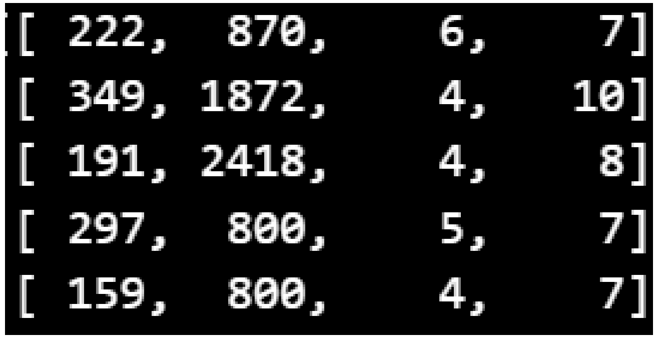
$$\text{Prediction Variance} = \hat{\sigma}^2\left(1 + \mathbf{x}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}\right)$$

Where:

- $\hat{\sigma}^2$ is the estimated variance of the residuals (also known as the residual mean square), obtained from the model summary.
- $\mathbf{X}$ is the design matrix of the independent variables, including a column of ones for the intercept term.
- $\mathbf{x}$ is the vector of predictors for which you want to make a prediction.

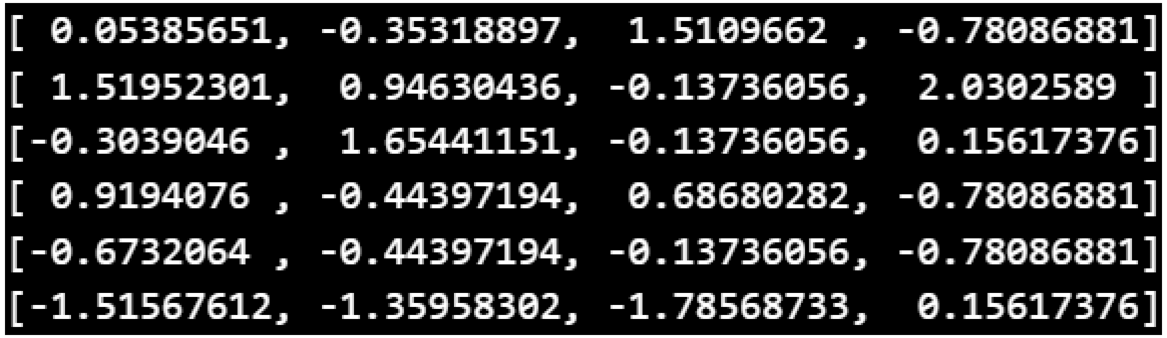
Standardization of Feature Matrix

Unstandardized Feature Matrix

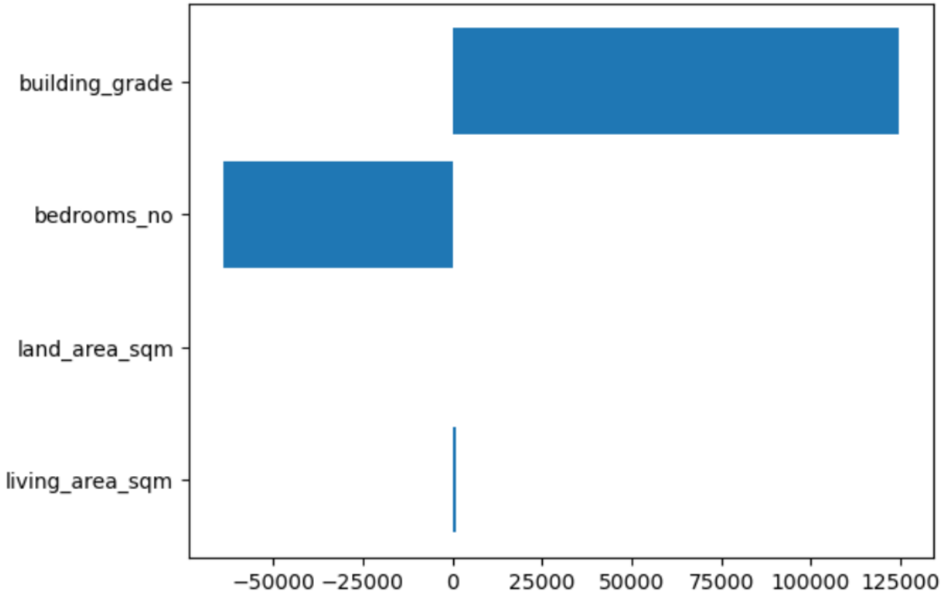


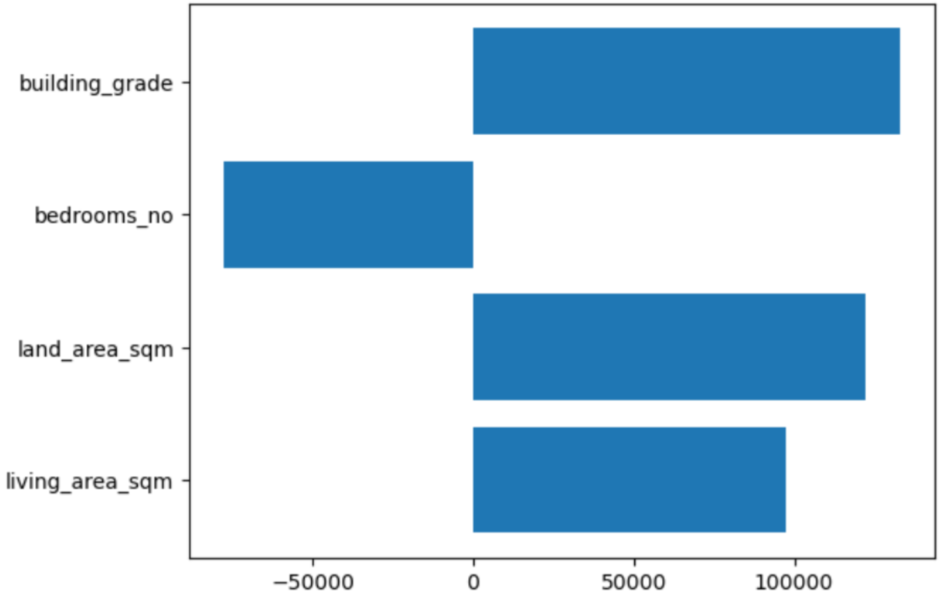
```
[[ 222,  870,    6,    7]
 [ 349, 1872,    4,   10]
 [ 191, 2418,    4,    8]
 [ 297,  800,    5,    7]
 [ 159,  800,    4,    7]
```

Standardized Feature Matrix



```
[ 0.05385651, -0.35318897,  1.5109662 , -0.78086881]
[ 1.51952301,  0.94630436, -0.13736056,  2.0302589 ]
[-0.3039046 ,  1.65441151, -0.13736056,  0.15617376]
[ 0.9194076 , -0.44397194,  0.68680282, -0.78086881]
[-0.6732064 , -0.44397194, -0.13736056, -0.78086881]
[-1.51567612, -1.35958302, -1.78568733,  0.15617376]
```


building_grade
bedrooms_no
land_area_sqm
living_area_sqm


building_grade
bedrooms_no
land_area_sqm
living_area_sqm