Machine Learning (ML)

Chapter 4:

Multiple Linear Regression, and

Multivariate Multiple Regression

Saeed Saeedvand, Ph.D.

Outline

In this Chapter:

- ✓ Multiple Linear regression
- ✓ Feature Selection
  - All subsets
  - Best subsets
  - Forward selection
  - Backward elimination
  - Univariate feature selection
  - Embedded methods
  - Wrapper methods
- ✓ Qualitative and Quantitate independent variables
- ✓ Interactions Between independent variables
- ✓ Multivariate Multiple Regression

Aim of this chapter:

- ✓ Understanding the Multiple Linear regression concept besides the Interactions Between independent variables. Different Feature Selection techniques and Qualitative and Quantitate independent variables.

What was the Linear regression?

- ✓ A statistical method that models the **relationship between two variables** (X and Y).
- ✓ We assuming there is a **linear relationship** between variables (for now).
- ✓ Find the best-fit line that describes this relationship.
- ✓ Can be used to make predictions about Y for a given value of X.

$$\hat{y} = \hat{\theta}_0 + \hat{\theta}_1 X$$

Multiple Linear Regression

Model

- ✓ A statistical method that analyzes the relationship between a dependent variable and **two or more** independent variables.
- ✓ Extension of simple linear regression.
- ✓ The goal is to find the best fitting line that predicts the dependent variable based on the values of the independent variables.

$$\hat{y} = \hat{\theta}_0 + \hat{\theta}_1 X_1 + \hat{\theta}_2 X_2 + \cdots + \hat{\theta}_p X_p + \varepsilon$$

Coefficients of the independent variables

Multiple Linear Regression

- ✓ The coefficients represent the change in the dependent variable (for a one-unit increase in the corresponding independent variable), **while holding all** other independent variables constant.
- ✓ Multiple linear regression is a widely used technique in machine learning and artificial intelligence for predictive modeling.
- ✓ Multiple linear regression **still has its place** in the field of data science and machine learning.
- ✓ **Advantage of multiple linear regression:**
  - ➢ Simplicity
  - ➢ Interpretability
  - ➢ Computationally efficient

Multiple Linear Regression

For example:

- Stock market prediction
- Customer behavior prediction
- Medical diagnosis
- Energy consumption prediction
- Environmental modeling

Note: In some cases, the simpler model may perform just as well as, or even better than, the more complex model.

Challenges

✓ If two or more independent variables are highly correlated, it can cause challenges in the regression analysis.

Multiple Linear Regression

Multicollinearity:

- ✓ Situation where two or more independent variables are highly correlated.
- ✓ Difficult to determine their individual effects on the dependent variable.
- ✓ Result in unstable or unreliable estimates of the coefficients.
- ✓ May lead to incorrect conclusions about the relationships between the variables

In multicollinearity high correlation between independent variables **does not necessarily mean** that change in one variable **causes the other change**, but it does indicate that there is high relationship.

Claims of causality

- ✓ The relationship between variables and whether changes in one variable directly cause changes in another variable.
- ✓ Different from Multicollinearity.

Multiple Linear Regression

What we should consider? (ideal)

- ✓ Choose independent variables that are as independent as from each other as possible.
  - ➢ It is to ensure that each variable's unique contribution to the dependent variable can be accurately estimated.

- ✓ **Interpretations possibility:**
  - ✓ By keeping all other variables fixed we can interpret one variable changes on Y.

Multiple Linear regression

Example - Correlated inputs:

Housing prices, with **two** independent variables:

- ✓ Square footage and number of bedrooms, may be **highly correlated**.
- ✓ If footage increases, the number of bedrooms increase also tends to increase price.
- ✓ **Multicollinearity can occur**.
- ✓ **Not Claims of causality** because changes directly may not leads to a change in another variable. .

Example - Uncorrelated inputs:

- ✓ **Customer satisfaction** at a restaurant with **two** independent variables:
  - **Waiting time for queue** and **food quality**, most likely are uncorrelated.
    - ➢ Waiting 20 minutes instead of 5 minutes will not make food more delicious!

Multiple Linear regression

Update for Calculation of the RSS:

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$\hat{y}_i = \hat{\theta}_0 + \hat{\theta}_1 x_{i\_1} + \hat{\theta}_2 x_{i\_2} + \cdots + \hat{\theta}_p x_{i\_p}$$

i^{th} **value** of x

First variable

Multiple Linear Regression

Question

- ✓ **Is Y** explained by all independent variables or only a useful subset of them?

Feature Selection

- ✓ Deciding on the important variables in a regression model and ML models is always **curtail**.
- ✓ Feature selection is process of identifying the most relevant independent variables (i.e., features) that are most useful for predicting the dependent variable.
- ✓ Feature selection is not limited to multiple linear regression models and it is valid for all ML algorithms.

Feature Selection

Common methods for feature selection

- All subsets
- Best subsets
- Forward selection
- Backward elimination
- Univariate feature selection
- Embedded methods
- Wrapper methods

Feature Selection

All subsets

- ✓ Fitting a regression model for every possible combination of the available independent variables.
- ✓ Then selecting the best subset of variables based on some evaluation metrics (e.g. MSE).

Problem?

- **Computationally expensive** (We most of the time cannot test all possible combinations and generate all models).
- **Suffer from overfitting** if the number of independent variables is much larger than the number of observations.

Feature Selection

Best Subsets

- ✓ Only evaluates a small number of promising subsets based on some criterion **without** considering every possible subset.
- ✓ We use a heuristic approach to search for the most promising subsets of variables, such as a greedy algorithm,. (again it may use MSE, …)

Heuristic: Problem-solving method that involves using practical techniques to quickly find an approximate solution like Binary search, Dijkstra's algorithm, … (**greedy approaches** are **subset of heuristic**)

Problem?

- Computational complexity
- Overfitting
- Missing variables
- Multicollinearity

Feature Selection

Univariate feature selection

- ✓ Selecting the most significant features based on their individual statistical significance, such as p-values or correlation coefficients.
- ✓ Simple and computationally efficient.

Problem?

- It does not consider the interactions or correlations between the features
- May miss important variables.

Feature Selection

Forward selection

✓ Starting with an empty model ($\hat{y}_i = \hat{\theta}_0$) and sequentially adding the most significant variable (e.g. $\hat{y}_i = \hat{\theta}_0 + \hat{\theta}_1 x_3$), until no significant variables remain.

How to detect Significant Variable:

- Fit a simple linear regression model for each independent variable and select the one that has the highest correlation or the lowest p-value or the F-statistic with the dependent variable:

Problem?

- Multicollinearity
- Computational complexity
- Overfitting
- **Sequentially bias:** the order of adding variables may affect the final model (especially if some variables are highly correlated or interact with each other).

Feature Selection

Backward Elimination

✓ Starting with a full model that includes all variables and sequentially removing the least significant variable, until all remaining variables are significant.

Variables are significant definition (how to stop approaches)

✓ **Significance level**: the algorithm can stop when all remaining variables have p-values below a predetermined significance level, typically 0.05 or 0.01.

✓ **Model fit**: The algorithm can stop when removing additional variables does not significantly decrease the goodness-of-fit (e.g. R^2). (This approach **ensures that the model is not overfitting**).

✓ **Expert judgment:** stop based on expert judgment or domain knowledge.

Problem?

- Similar to forward selection.

Feature Selection

Embedded methods

- ✓ Integrating feature selection into the model training process itself (objective function can change).
- ✓ For instance regularization techniques that penalize the complexity of the model and **shrink the coefficients** of less important features to zero.
  - For example we can use, decision trees, random forests, L1 regularization.
- ✓ Embedded methods can select a subset of features that are relevant, while avoiding the overfitting, (may arise from backward elimination or forward selection).

Problem?

- Parameter tuning challenges.
- Non-robustness: can be sensitive to outliers
- Interpretability: may produce complex models that are difficult to interpret or explain

Feature Selection

Wrapper methods

- ✓ Using a specific machine learning algorithm to evaluate subsets of features and select the ones for the best performance.
- ✓ Iteratively building models with different subsets of features, evaluating their performance.
- ✓ Wrapper methods can potentially lead to better performance.
- ✓ Similar to embedded methods decision trees, random forests, or support vector machines (SVMs).

Problem?

- Computationally expensive (require building and evaluating a large number of models).
- Wrapper are more computationally expensive than Embedded methods.

Qualitative or Categorical Predictors

What to do with **qualitative** variables that taking a discrete set of values?

- ✓ **Gender:** Male or Female
- ✓ **Education level:** High school, Bachelor's degree, Master's degree, etc.
- ✓ **Marital status:** Married, Single, Divorced, etc.
- ✓ **Occupation:** Managerial, Professional, Sales, Service, etc.
- ✓ **Region:** Northeast, South, Midwest, West, etc.
- ✓ **Type of car:** Sedan (common), SUV, Truck, etc.
- ✓ **Brand of phone:** Apple, Samsung, Google, etc.

Qualitative or Categorical Predictors

✓ For credit card balance between males and females we can use dummy variable:

$$x_i = \begin{cases} 0 & if\ female \\ 1 & if\ male \end{cases}$$

gender

$$y_i = \hat{\theta}_0 + \hat{\theta}_1 x_1 + \varepsilon_i = \begin{cases} \hat{\theta}_0 + \varepsilon_i & if\ female \\ \hat{\theta}_0 + \hat{\theta}_1 + \varepsilon_i & if\ male \end{cases}$$

Qualitative or Categorical Predictors

Average credit card balance for females

$$y_i = \hat{\theta}_0 + \hat{\theta}_1 x_1 + \varepsilon_i = \begin{cases} \hat{\theta}_0 + \varepsilon_i & \text{if female} \\ \hat{\theta}_0 + \hat{\theta}_1 + \varepsilon_i & \text{if male} \end{cases}$$

Difference in credit card balance between males and females

- ✓ **Now we can estimate** the coefficients $\hat{\theta}_0$ and $\hat{\theta}_1$ by fitting a linear regression model to the data.

Qualitative or Categorical Predictors

$$y_i = \hat{\theta}_0 + \hat{\theta}_1 x_1 + \varepsilon = \begin{cases} \hat{\theta}_0 + \varepsilon_i & if\ female \\ \hat{\theta}_0 + \hat{\theta}_1 + \varepsilon_i & if\ male \end{cases}$$

Important note:

- In this case, the sign of the coefficient indicates the direction of the relationship between **gender** and **credit card balance**.

If find that $\hat{\theta}_1$ is positive and statistically significant:

- We can conclude that, on average, males have a higher credit card balance than females.

If find that $\hat{\theta}_1$ is negative and statistically significant:

- We can conclude that, on average, females have a higher credit card balance than females.

Qualitative or Categorical Predictors

Question

What if find that $\hat{\theta}_1$ is negative or positive but and not statistically significant?

- ✓ We have not have enough evidence to conclude that there is a significant difference in credit card balance between the two groups.

Question

Does it means between the two groups there is no significant?

- ✓ To make sure we need to consider other possible variables in the model too.

Qualitative or Categorical Predictors

If more than two category for one variable

- ✓ We use multiple dummy variables, one for each category
- ✓ For example educational college level: Bachelor, Master, Ph.D.

Independent variable number

$$x_{i\_1} = \begin{cases} 1 & if \ \text{Bachelor} \\ 0 & if\ not\ \text{Bachelor} \end{cases}$$

$$x_{i\_2} = \begin{cases} 1 & if \ \text{Master} \\ 0 & if\ not\ \text{Master} \end{cases}$$

Note: one fewer dummy variable than the number of levels.

Qualitative or Categorical Predictors

✓ To obtain the model we can use these variables in the regression equation.

$$x_{i1} = \begin{cases} 1 & if \text{ Bachelor} \\ 0 & if\ not \text{ Bachelor} \end{cases} \qquad x_{i2} = \begin{cases} 1 & if \text{ Master} \\ 0 & if\ not \text{ Master} \end{cases}$$

Income

$$y_i = \hat{\theta}_0 + \hat{\theta}_1 x_{i\_1} + \hat{\theta}_2 x_{i\_2} + \varepsilon_i = \begin{cases} \hat{\theta}_0 + \varepsilon_i & Ph.D. \\ \hat{\theta}_0 + \hat{\theta}_2 + \varepsilon_i & if\ Master \\ \hat{\theta}_0 + \hat{\theta}_1 + \varepsilon_i & if \text{ Bachelor} \end{cases}$$

Interpretation

✓ Assuming one unit increase in x_{i_1}, while holding x_{i_2} is constant:

- If $\hat{\theta}_1$ for x_{i_1} positive and statistically significant:
  - Conclude that there is a positive relationship between x_{i_1} and y and **increase of x_{i_1} results in increase of y** (in x_{i_1} can mean Bachelor is has higher effect then not having Bachelor)

Qualitative and Quantitate together

Combination of both

- ✓ We can consider Qualitative and Quantitate together.
- ✓ For example predict **income** based on working hours (x_{i1}), and gender (x_{i2}).

$$y_i = \hat{\theta}_0 + \boxed{\hat{\theta}_1 x_{i\_1}} + \boxed{\begin{cases} \hat{\theta}_2 & if\ female \\ 0 & if\ male \end{cases}} + \varepsilon_i$$

$$x_{i\_1} = \begin{cases} 1 & if\ female \\ 0 & if\ male \end{cases}$$

Income

Practice

Combine Qualitative and Quantitate approaches and A) compute p-value and t-statistics B) interpret results
Note: You can use chapter 3's implementations.

Interactions Between independent variables:

Interactions Between variables

- ✓ We mostly assumed that the effect of the independent variables are **independent** of **each other** as was in following equation:

Independent effect

$$y_i = \hat{\theta}_0 + \hat{\theta}_1 x_{i\_1} + \hat{\theta}_2 x_{i\_2} + \varepsilon$$

The **effect on $\boldsymbol{y_i}$** if one unit increase in x_{i_1} is always $\hat{\theta}_1$, (regardless of the changes in x_{i_2}).

For Example:

x_{i_1} = advertisement on Social media (its slope $\hat{\theta}_1$)
x_{i_2} = advertisement on TV (its slope $\hat{\theta}_2$)
y_i = sales

Interactions has not been considered!

Interactions Between independent variables:

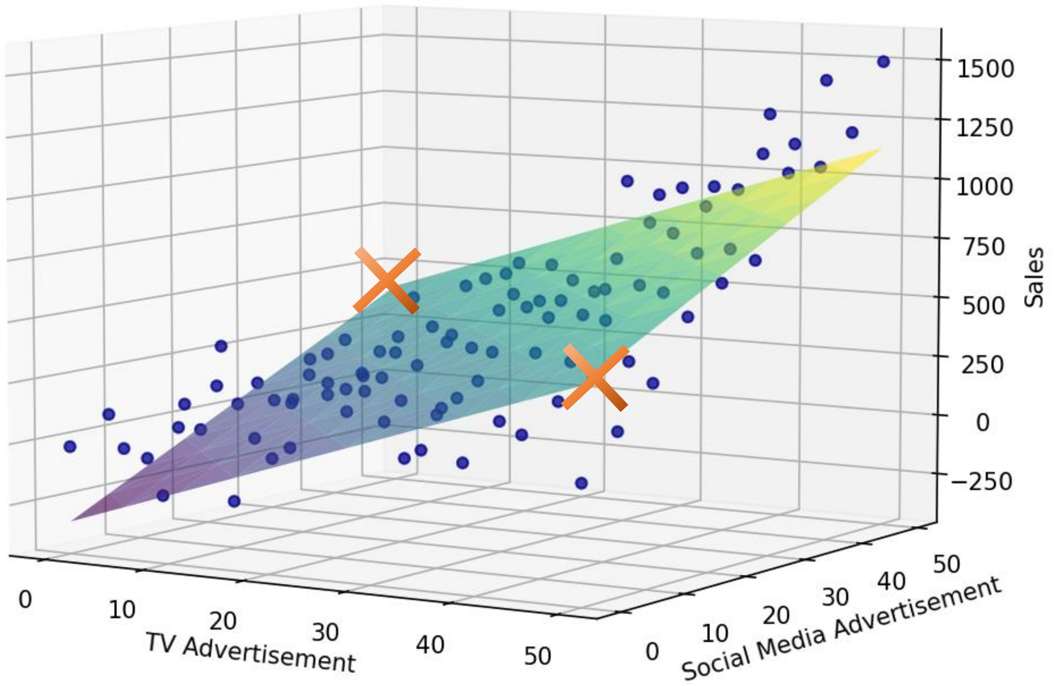
Interactions Between variables

- ✓ Assume spending money on Social media advertising increases the effectiveness of TV advertising.
- ✓ If Social media's has **higher impact** on sale solely we cannot spend all budget on social media to get maximum sale!
- ✓ Maybe %50 each results in higher sale (this known as synergy in marketing).

Interactions Between independent variables:

Interactions Between variables

✓ Then **Social media's** slope $\hat{\theta}_1$ should **change TV** too.



Interactions Between independent variables:

Interactions Between variables

✓ Updating model to consider interactions:

$$y_i = \hat{\theta}_0 + \hat{\theta}_1 x_{i\_1} + \hat{\theta}_2 x_{i\_2} + \boxed{\hat{\theta}_3 (x_{i\_1} \times x_{i\_2})} + \varepsilon$$

x_{i_1} = advertisement on Social media (its slope $\hat{\theta}_1$)
x_{i_2} = advertisement on TV (its slope $\hat{\theta}_2$)
y_i = sales

Note: It is recommended that if we include an interaction in a model, we need to also include the main effects of that independent variable (even if significant is low)

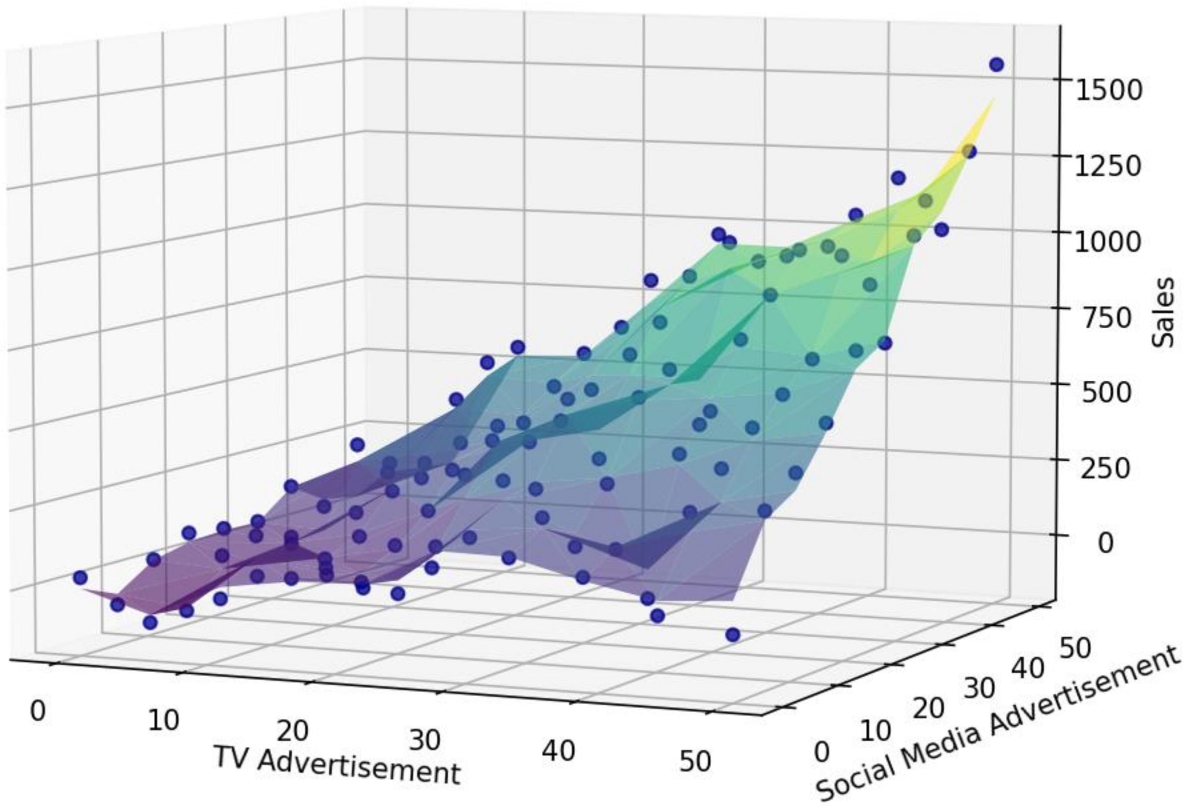
Nonlinear Regression function

Non-linear Models

- ✓ In linear regression models, we assumed a linear relationship between the dependent variable and the independent variables.
- ✓ Nonlinear regression models can capture more complex relationships when the **relationship is not linear**.
- ✓ Nonlinear regression functions can **include** polynomial functions, exponential functions, logarithmic functions, power functions, and many others.
- ✓ The **choice** of the specific nonlinear function **depends on the nature of the data**.

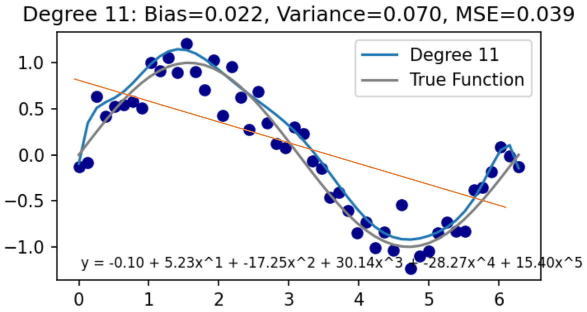
Nonlinear Regression function

Non-linear Models



Nonlinear Regression function

Non-linear Models



- ✓ Polynomial regression model with a single predictor (x_{i1}).

$$y_i = \hat{\theta}_0 + \hat{\theta}_1 x_{i\_1} + \hat{\theta}_2 x_{i\_1}{}^{2} + \cdots + \hat{\theta}_1 x_{i\_1}{}^{n} + \varepsilon$$

- ✓ Polynomial regression model with a multiple predictor (x_{i1} and x_{i2}).

$$\begin{gathered} y_i = \hat{\theta}_0 + (\hat{\theta}_{1\_1} x_{i\_1} + \hat{\theta}_{1\_2} x_{i\_1}{}^{2} + \cdots + \hat{\theta}_{1\_n} x_{i\_1}{}^{n}) \\ \left(\hat{\theta}_{2\_1} x_{i\_2} + \hat{\theta}_{2\_2} x_{i\_2}{}^{2} + \cdots + \hat{\theta}_{2\_n} x_{i\_2}{}^{n}\right) + \cdots \\ \left(\hat{\theta}_{p\_1} x_{i\_2} + \hat{\theta}_{p\_2} x_{i\_2}{}^{2} + \cdots + \hat{\theta}_{p\_n} x_{i\_p}{}^{n}\right) + \varepsilon \end{gathered}$$

Multivariate Multiple Regression

Definition

- ✓ If we have **multiple dependent variables** and multiple **independent variables**.
- ✓ In the **Multivariate Multiple Regression** model, each dependent variable can have its own set of regression coefficients for the predictors (we are not limited).

$$y_{i\_1} = \hat{\theta}_{0\_1} + \hat{\theta}_1 x_{i\_1} + \hat{\theta}_1 x_{i\_2} + \varepsilon_1$$
$$y_{i\_2} = \hat{\theta}_{0\_2} + \hat{\theta}_1 x_{i\_2} + \hat{\theta}_1 x_{i\_2} + \varepsilon_2$$

For instance: Sale for product A and sale for product B based on advertising on Social media and TV.

- ✓ This does not consider correlations between the dependent variables but can still work well **in practice**.

Assignment

Combine Qualitative and Quantitate with Multivariate Multiple approaches in simple python code and A) compute p-value and t-statistics B) Explain if we can define correlations between the dependent variables too, if yes how.

Summery

- ✓ We understood Multiple Linear regression in details.
- ✓ We described different Feature Selection approaches.
- ✓ We saw how to model Qualitative and Quantitate independent variables.
- ✓ We went deep to see modeling of Interactions Between independent variables.
- ✓ We discussed Multivariate Multiple Regression.