AI200: APPLIED MACHINE LEARNING

LINEAR REGRESSION

OVERVIEW & LITERATURE OF MACHINE LEARNING

Machine Learning

1. Spectrum of supervision
   - Supervised Learning
   - Unsupervised Learning
   - Reinforcement Learning

	Supervised Learning		Unsupervised Learning			Reinforcement Learning
2. Types of problems	Regression	Classification	Clustering	Association	Dimensionality Reduction	
3. Types of algorithm / models	Linear regression Polynomial regression Neural Networks Decision Tree Random Forest XGBoost	Logistic Regression KNN Naïve Bayes SVM Neural Networks Decision Tree Random Forest XGBoost	Neural Networks K-Means Hierarchical Clustering Hidden Markov Models Gaussian Mixture	Apriori Algorithm FP-Growth Algorithm Eclat Algorithm	Neural Networks SVD PCA	
4. Types of applications	Forecasting Predictions Process optimization New Insights	Fraud Detection Image Classification Customer Retention Diagnostics	Recommendations Targeted Marketing Customer Segmentation	Market Basket Analysis	Feature Elicitation Structure Discovery Meaningful Compression Big Data Visualisation	Real-time Decisions Game AI Learning Tasks Skill Acquisition Robot Navigation

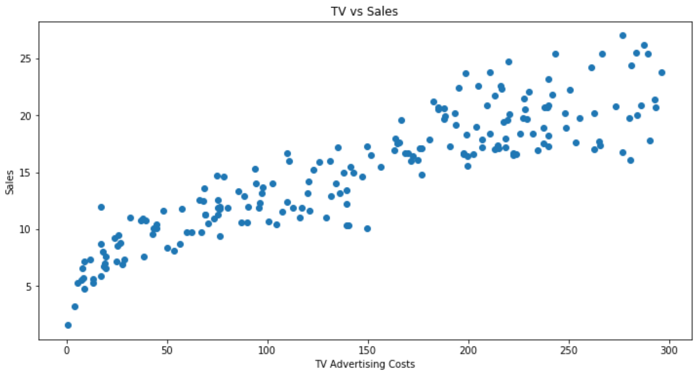
BROAD IDEA OF REGRESSION PROBLEMS

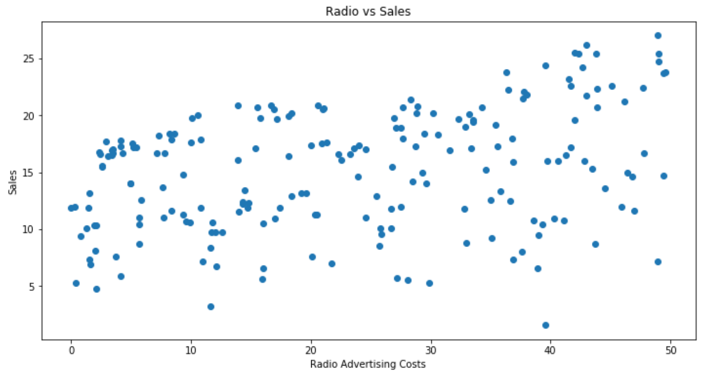
- Before we begin, let's represent a simple dataset in graphical form. We shall use the very widely-used advertising dataset here.
- Here we are trying to study <u>how spending in different advertising mediums (TV, radio & newspaper) affects sales</u> of a company

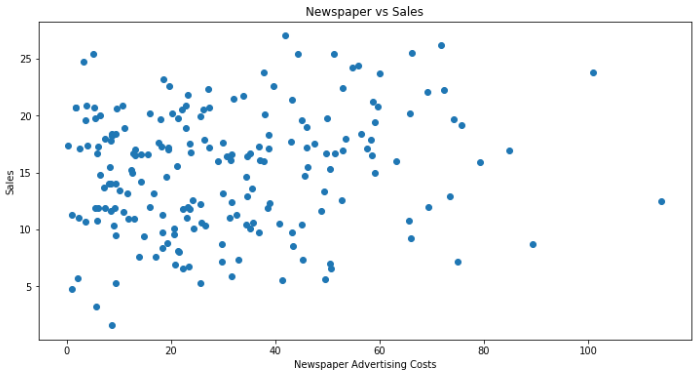
Features **Outcome**

	TV	Radio	Newspaper	Sales
0	230.1	37.8	69.2	22.1
1	44.5	39.3	45.1	10.4
2	17.2	45.9	69.3	12.0
3	151.5	41.3	58.5	16.5
4	180.8	10.8	58.4	17.9
...	...	...	...	...
195	38.2	3.7	13.8	7.6
196	94.2	4.9	8.1	14.0
197	177.0	9.3	6.4	14.8
198	283.6	42.0	66.2	25.5
199	232.1	8.6	8.7	18.4

200 rows × 4 columns

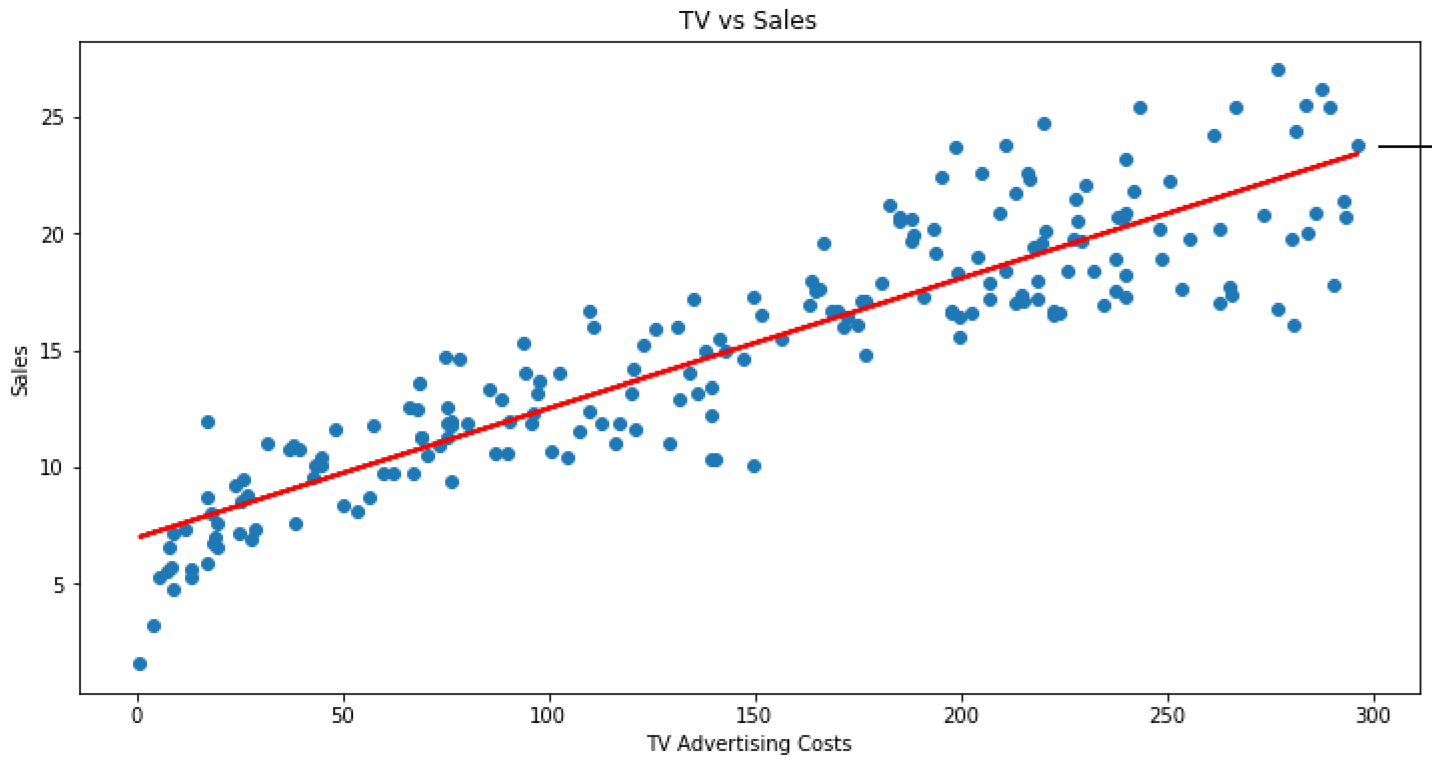






WHAT IS LINEAR REGRESSION: LAYMAN INTUITION

- Linear regression operates under the assumption that the **feature** is linearly related to the **outcome**
- Using some math (which we'll elaborate on later), we draw a line of best fit (straight line that best describes the relationship between our features and outcome) that is represented by a simple algebraic equation. This is your linear regression model.



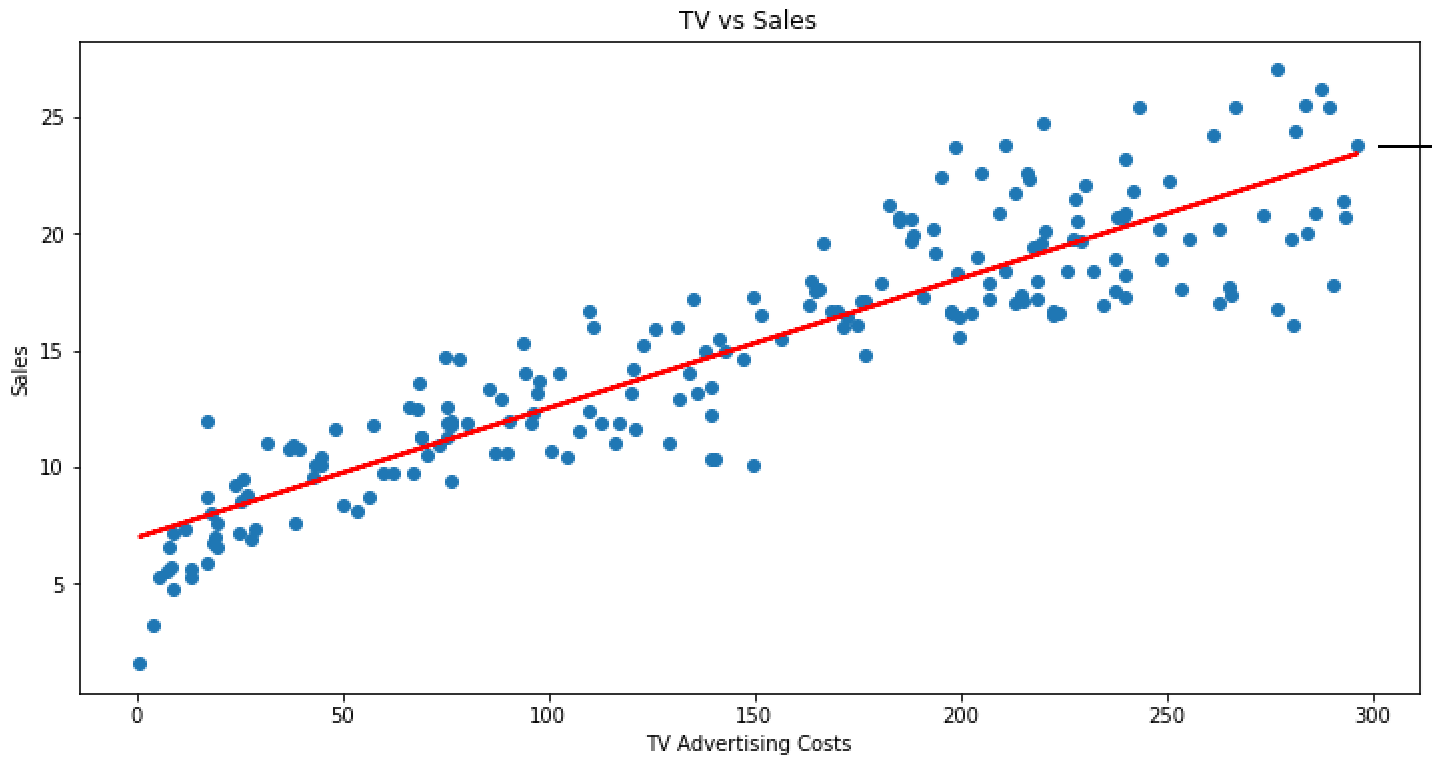
Line of best fit

Y = 6.9748 + 0.0555 x tv

WHAT IS LINEAR REGRESSION: LAYMAN INTUITION

- Linear regression operates under the assumption that the **feature** is linearly related to the **outcome**
- Using some math (which we'll elaborate on later), we draw a line of best fit (straight line that best describes the relationship between our features and outcome) that is represented by a simple algebraic equation. This is your linear regression model.

Fit / Training a model



Line of best fit

Y = 6.9748 + 0.0555 x tv

WHAT IS LINEAR REGRESSION: LAYMAN INTUITION

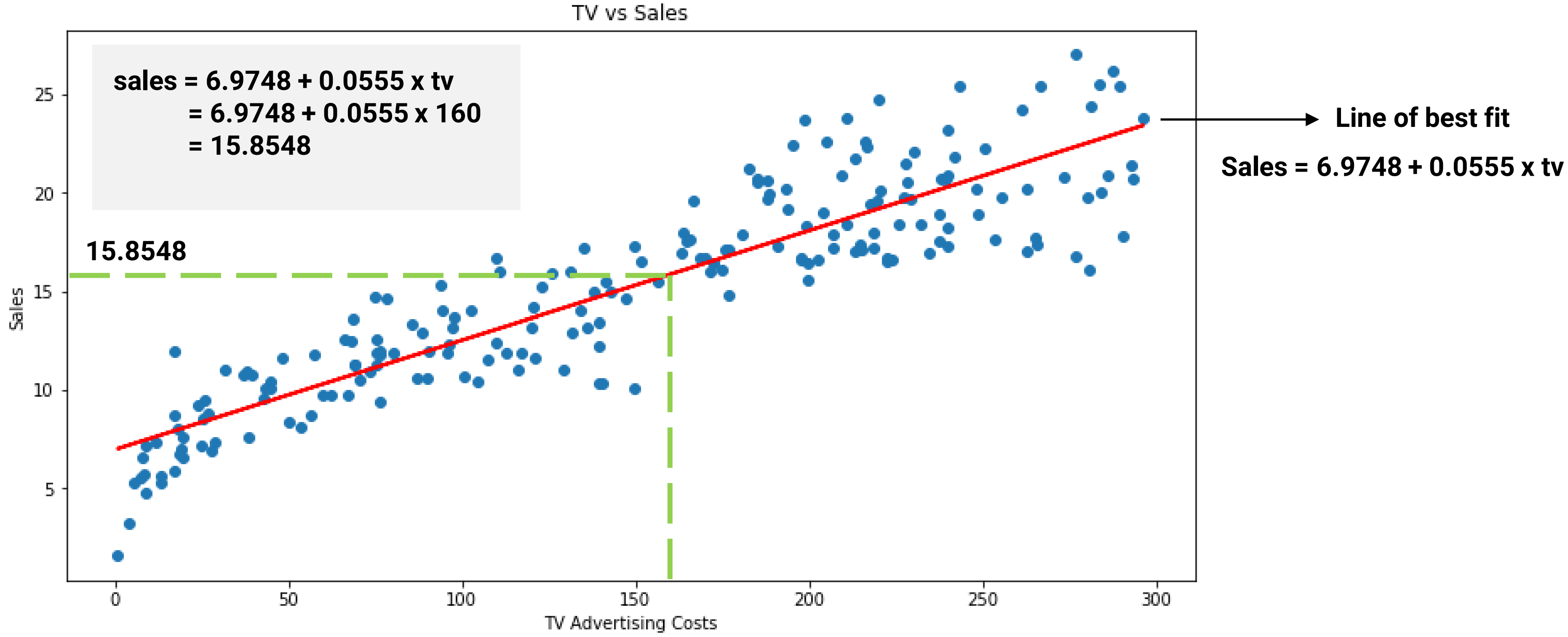
- Linear regression operates under the assumption that the **feature** is linearly related to the **outcome**
- Using some math (which we'll elaborate on later), we draw a line of best fit (straight line that best describes the relationship between our features and outcome) that is represented by a simple algebraic equation. This is your linear regression model.

Fit / Training a model

Let's say in future, Company X is keen on spending 160 on TV advertising, but it wants to first forecast or ***predict*** the resulting sales to evaluate whether the investment is worthwhile or not.

  - To predict the sales, we can use the same line of best fit in the following way:

Predicting an outcome with the trained model



WHAT IS LINEAR REGRESSION: LAYMAN INTUITION

- If you look at your equation carefully, you'd realize it takes the general form of an equation that most of us would be pretty familiar with from our secondary school education:

$$y = 6.9748 + 0.0555 \text{ x tv} \longrightarrow y = mx + c$$

- However in machine learning we like to be fancier (and it helps command that 100k/year salary ☺), and so we represent it with:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

LINEAR REGRESSION

INTUITION BEHIND ORDINARY LEAST SQUARES

FINDING THE BEST FIT: ORDINARY LEAST SQUARES

- It is evident that the line of best fit & the algebraic equation representing this line is what allows us to generate prediction for **outcomes (e.g sales)** based on given **features (e.g tv)**

- And given the formula, we basically just need to find the values of α and β

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

- To do this, we use the Ordinary least squares method to find α and β and create your line of best fit: (math as shown below)

$$\hat{\alpha} = \min_{\alpha} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 = \min_{\alpha} \sum_{i=1}^{n} \varepsilon_i^2$$

$$\hat{\beta} = \min_{\beta} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 = \min_{\beta} \sum_{i=1}^{n} \varepsilon_i^2$$

Isn’t there an easier way to understand regression?!?!

FINDING THE BEST FIT: ORDINARY LEAST SQUARES

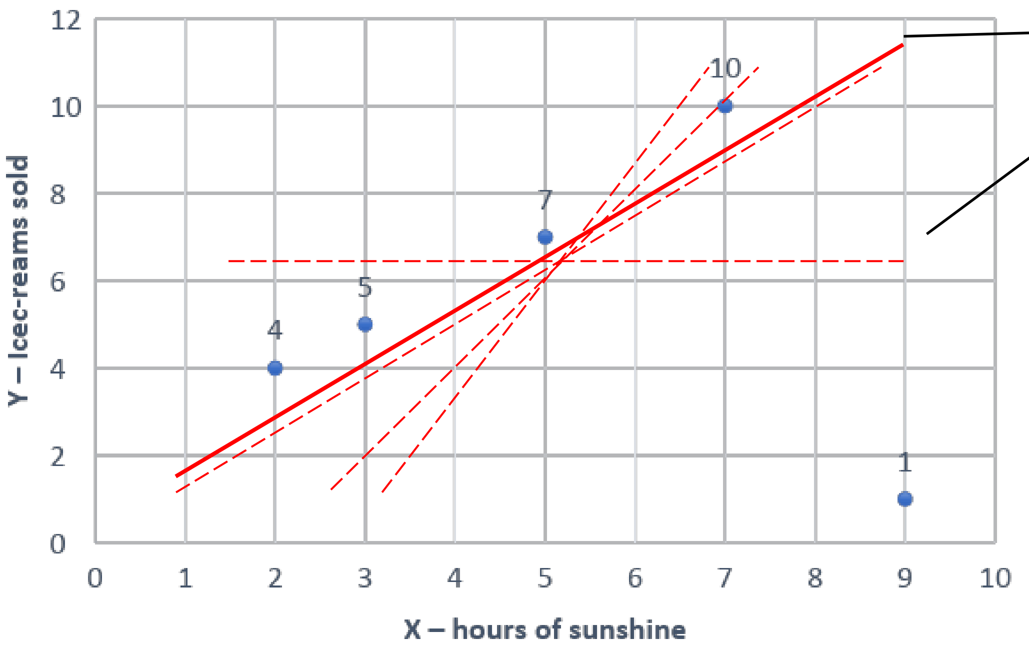
- Let's chuck the fancy math, and understand its intuition from a graphical point of view.

- We draw many possible lines of best fit (in red) which represent different set of **α** and **β** values
- Then, we calculate the sum of residuals (RSS) of each point.
- Repeat multiple times to find the optimal value for **α** and **β** which has the lowest aggregate loss

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

Example, α=0 and β=1.5 and RSS = ??

"x" Hours of Sunshine	"y" Ice Creams Sold
2	4
3	5
5	7
7	10
9	1



We will repeatedly calculate the RSS for each set of α and β values

FINDING THE BEST FIT: ORDINARY LEAST SQUARES

- You can perform ordinary least squares with just 2 lines of code

```
lm = LinearRegression()
model = lm.fit(X, y)
```

- And with this you would have derive at the optimal **α** and **β** with the lowest aggregate error, to construct your line of best fit. Now, you are ready to do some prediction!

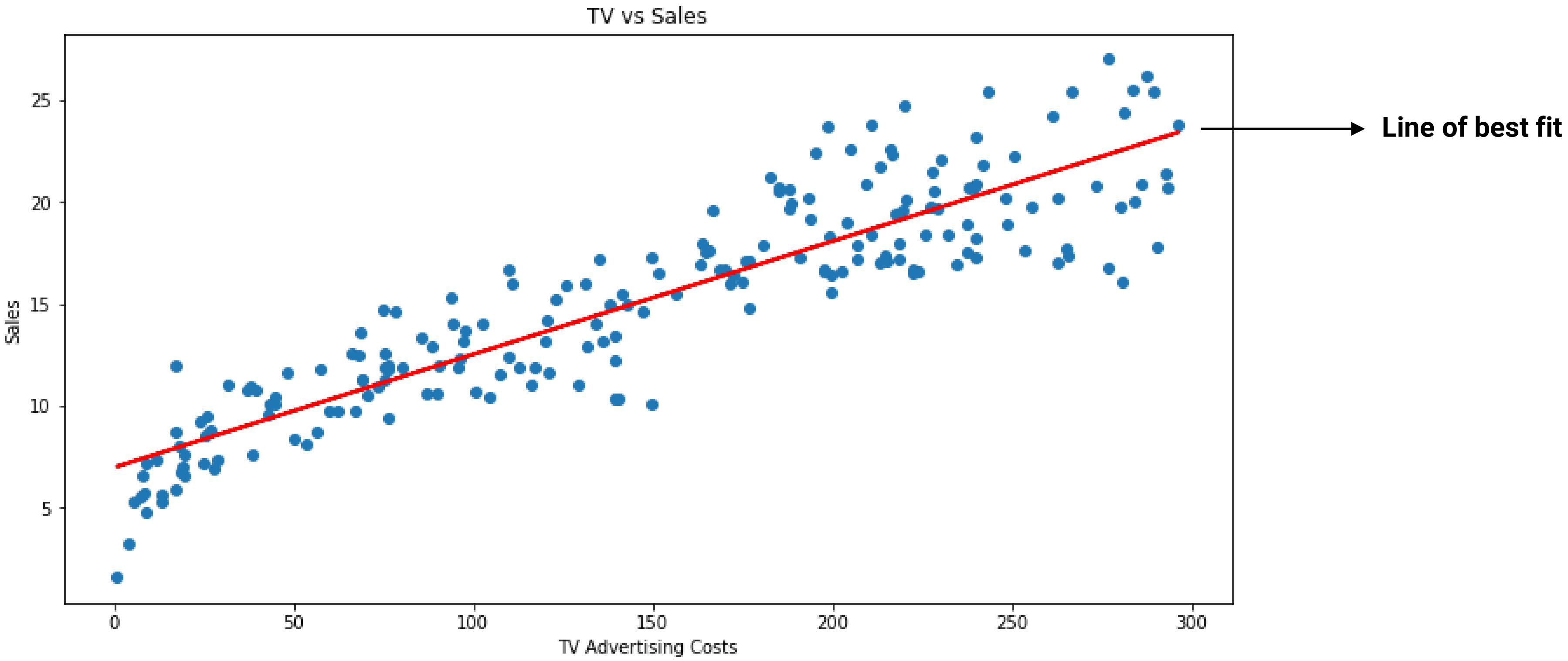
LINEAR REGRESSION

MULTIPLE LINEAR REGRESSION

MULTIPLE LINEAR REGRESSION

- Earlier, we established a linear regression model for the TV advertising and derived sales. Let's rewrite the equation for the line of best fit **Y = 6.9748 + 0.0555 * tv** into something more machine-learning friendly:

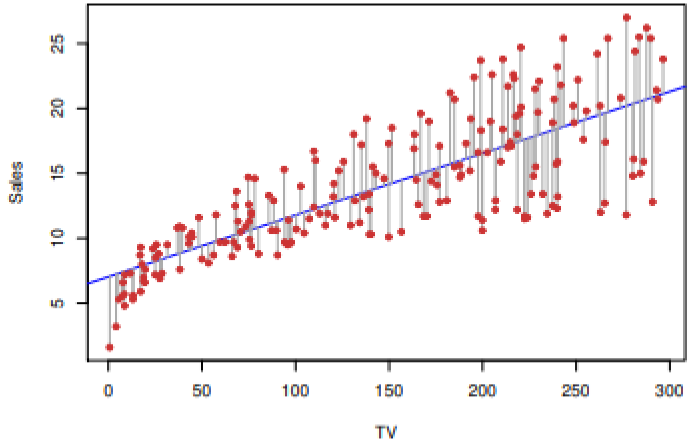
  - Sales = $\alpha + \beta \times$ TV $+ \mathcal{E}$



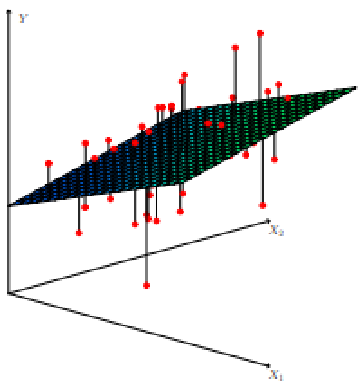
MULTIPLE LINEAR REGRESSION

- We can use more than one feature to predict an outcome in linear regression. This is called a multiple linear regression model
- Essentially you are only adding additional features to the equation:
  - Sales = α + β x TV + Ɛ
- For instance, if we wish to include the features radio and newspaper into predicting the outcome:
  - Sales = α + β_1 x TV + β_2 x Radio + β_3 x Newspaper + Ɛ
  - *Where, (β_1, β_2, β_3) are the parameters / weight for each of the features.*
- The ordinary least squares method can be generalised even beyond 2-dimensions. Here's how it might look like in 3D:

Least squares picture in 1-dimension



Least squares picture in 2-dimensions



The 2-dimensional plane in the 3D picture is the least squares fit of Y onto the predictors x_1 and x_2.

If you tilt this plane in any way, you would get a larger sum of squared vertical distances between the plane and the observed data.

MULTIPLE LINEAR REGRESSION

- After fitting the model using ordinary least squares:

  - Sales = $\alpha + \beta_1$ x TV + β_2 x Radio + β_3 x Newspaper + Ɛ

	Coefficient	Std. Error
Intercept	2.939	0.3119
TV	0.046	0.0014
radio	0.189	0.0086
newspaper	-0.001	0.0059

Results of parameter weights after fitting model using ordinary least squares

- The coefficient β_1 tells us the expected change in sales per unit change of the TV budget, with all other predictors held fixed.

- What the above table tells us is that:
  - Holding the other budgets fixed, for every \$1000 spent on TV advertising, sales on average increase by (1000 × 0.046) = 46 units sold

MULTIPLE LINEAR REGRESSION

- A regression coefficient β_j estimates the expected change in Y per unit change in X_j , ***assuming all other predictors are held fixed.***

- But:
  - **Predictors are often not independent of each other**. For example, a firm may reap non-linear economies of scale by ramping up budgets on multiple advertising channels at the same time.
  - **Predictors typically change together**. For example, a firm might not be able to increase the TV ad budget without reallocating funds from the newspaper or radio budgets.
  - We assume here that the relationship between the features and outcome is linear

- So, how do we know if the multiple linear regression model is fit for purpose or not?
  - Visualise the input features. Are they highly correlated?
  - Evaluate model by training & testing, and comparing against other models

LINEAR REGRESSION

ADDITIONAL CONSIDERATIONS

ADDITIONAL CONSIDERATIONS

In the words of a famous statistician…

"Essentially, all models are wrong, but some are useful."

—George Box

So how do we make regression models less wrong (or make it great again) ?

- Feature selection to avoid multicollinearity
- Feature engineering
  - Polynomials
  - Step functions
  - Splines
  - Local regression
  - Generalized additive models

ADDITIONAL CONSIDERATIONS

- There is more to linear regression than what was covered. But now, you already know enough to:

    - Build your own linear regression model and understand what it does
    - Learn more about the other intricacies of linear regression on your own

- Some recommended topics for you to read up on your own:

    - Feature selection:
        - Stepwise regression
        - Forward selection
        - Backward elimination

    - Feature engineering
        - Polynomials
        - Step functions
        - Splines
        - Local regression
        - Generalized additive models

    - Regularization (another approach to addressing overfitting)
        - Lasso Regression
        - Ridge Regression