# Intro to Machine Learning

Lecture 2

Adil Khan

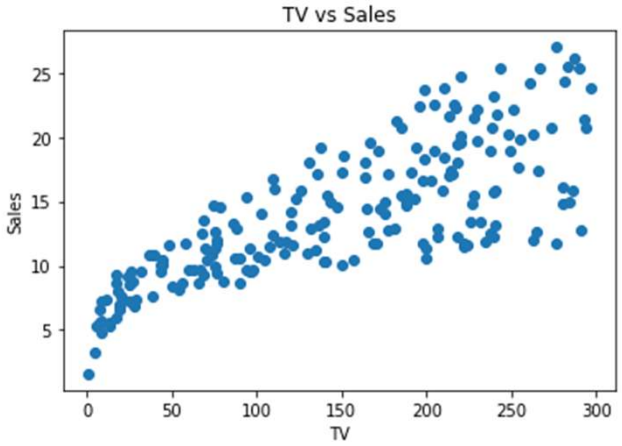
# Recap

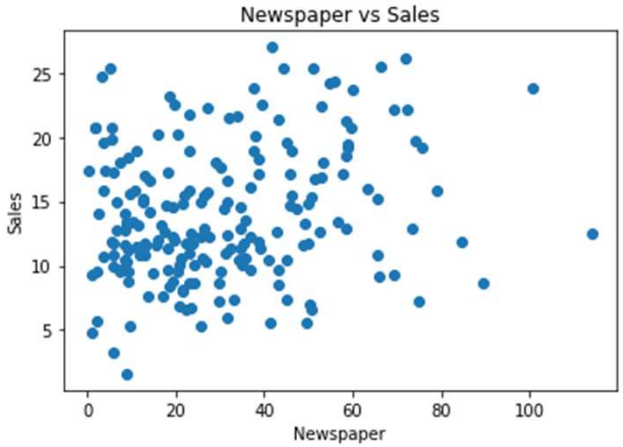
- Machine learning
  - What is machine learning?
  - Why learn/estimate?
  - Predictors and response variables
  - Types of learning
  - Regression and classification
  - Parametric and non-parametric models
  - Bias and variance

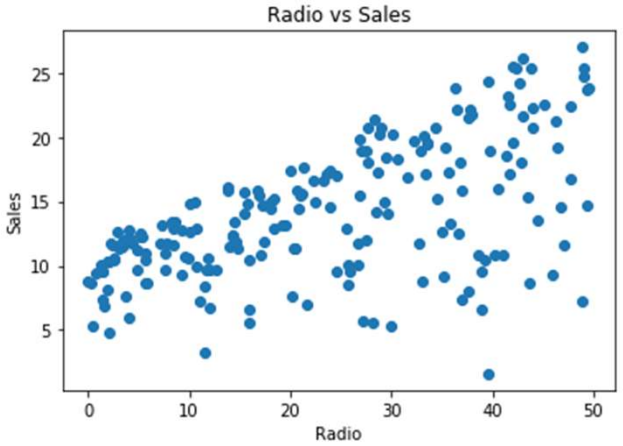
# Objectives

- What is linear regression?
- Why study linear regression?
- What can we use it for?
- How to perform linear regression?
- How to estimate its performance?
- What are its extensions?
- What are dummy variables?

# Example (1)

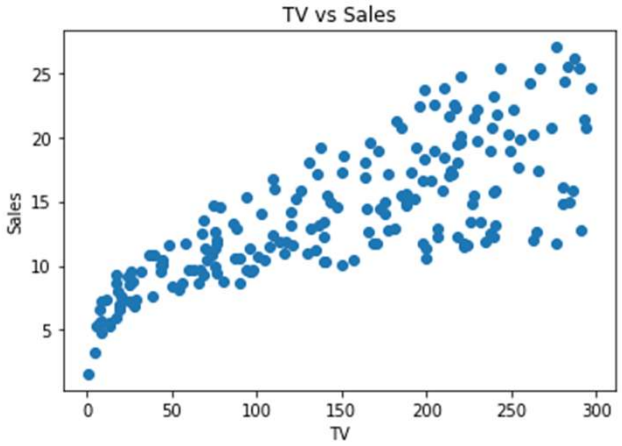


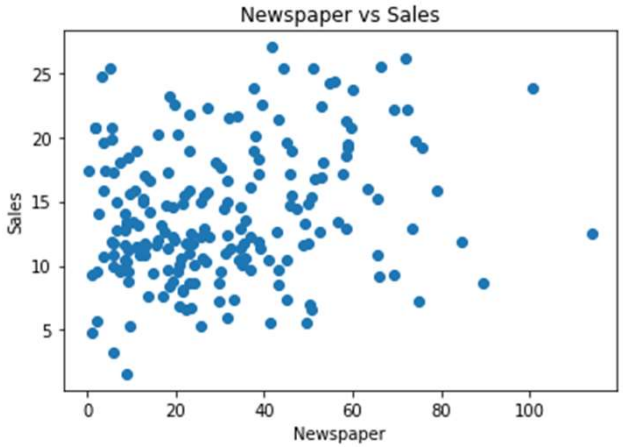



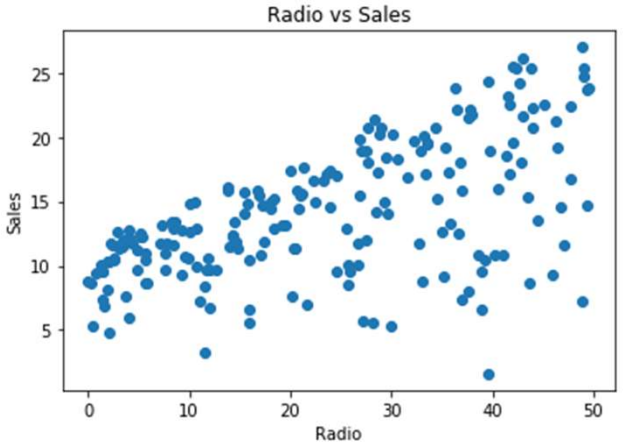




| | Unnamed: 0 | TV | radio | newspaper | sales |
|---|---|---|---|---|---|
| **0** | 1 | 230.1 | 37.8 | 69.2 | 22.1 |
| **1** | 2 | 44.5 | 39.3 | 45.1 | 10.4 |
| **2** | 3 | 17.2 | 45.9 | 69.3 | 9.3 |
| **3** | 4 | 151.5 | 41.3 | 58.5 | 18.5 |
| **4** | 5 | 180.8 | 10.8 | 58.4 | 12.9 |
| **5** | 6 | 8.7 | 48.9 | 75.0 | 7.2 |

# Example (2)


TV vs Sales
Sales
TV
0
50
100
150
200
250
300
5
10
15
20
25


Newspaper vs Sales
Sales
Newspaper
0
20
40
60
80
100
5
10
15
20
25


Radio vs Sales
Sales
Radio
0
10
20
30
40
50
5
10
15
20
25

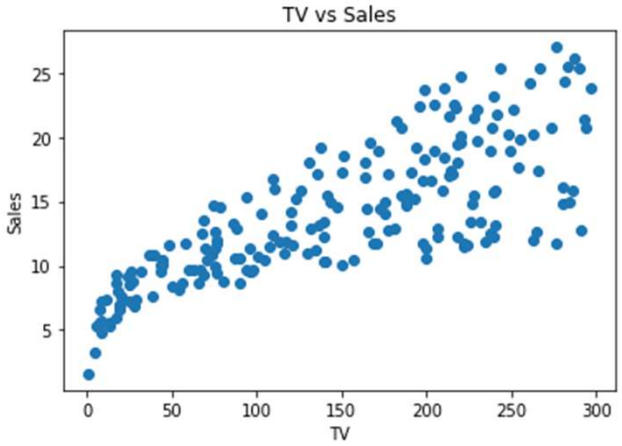
# What we might want to know?

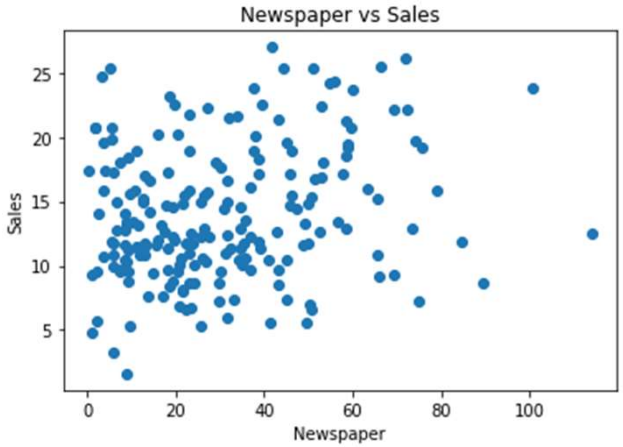
- Is there a relationship between advertising budget and sales?
- How strong is the relationship between advertising budget and sales?
- Which media contribute to sales?
- How accurately can we estimate the effect of each medium on sales?
- How accurately can we predict future sales?
- Is there synergy among the advertising media?

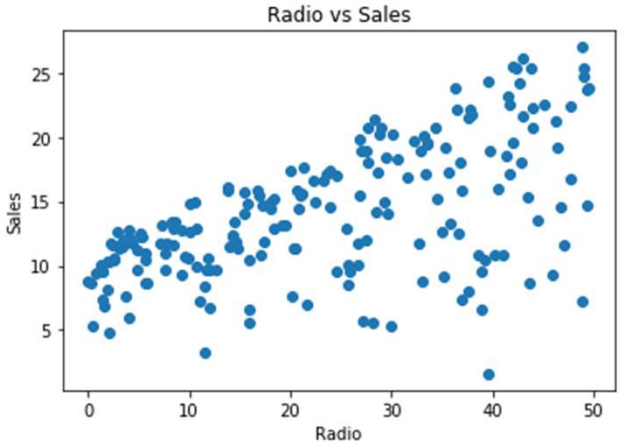
# What we might want to know?

- Is there a relationship between advertising budget and sales?
- How strong is the relationship between advertising budget and sales?
- Which media contribute to sales?
- How accurately can we estimate the effect of each medium on sales?
- How accurately can we predict future sales?
- Is there synergy among the advertising media?

# Example (3)

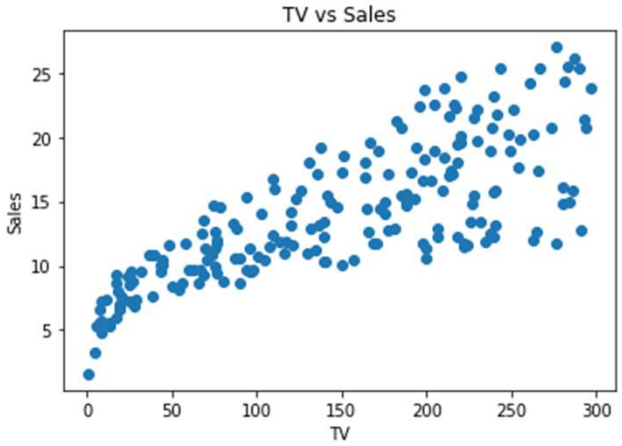


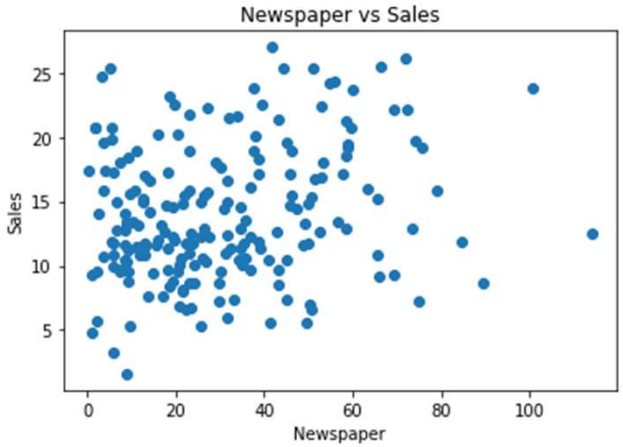



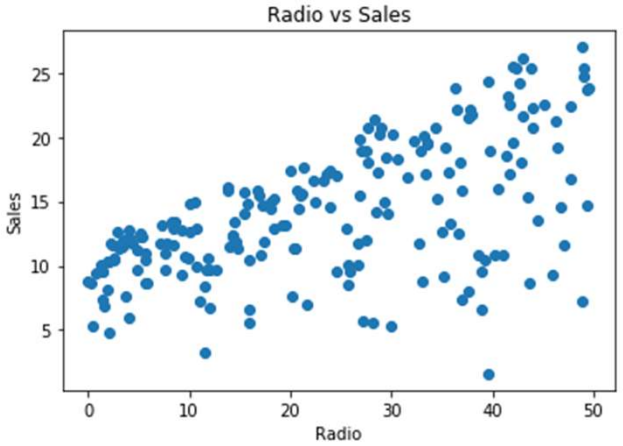




$$Sales = f(TV, Newspaper, Radio) + \epsilon$$

# Example (4)



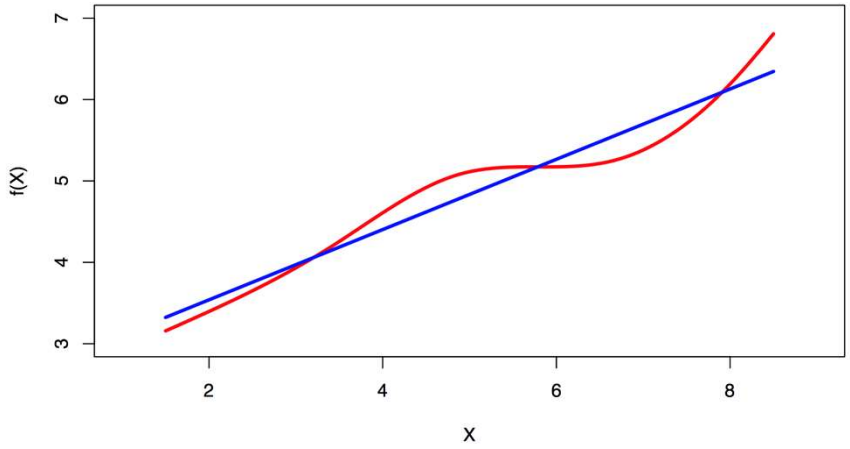

$$Sales = f(TV, Newspaper, Radio) + \epsilon$$

$$\widehat{Sales} \approx \hat{f}(TV, Newspaper, Radio)$$

# What is linear regression?

- A simple supervised learning approach

- Assumes a linear relationship between the predictors and the response




$$Y = \beta_0 + \beta_1 X$$

# Why study linear regression?

- Although it may seem overly simplistic, linear regression is extremely useful both conceptually and practically.

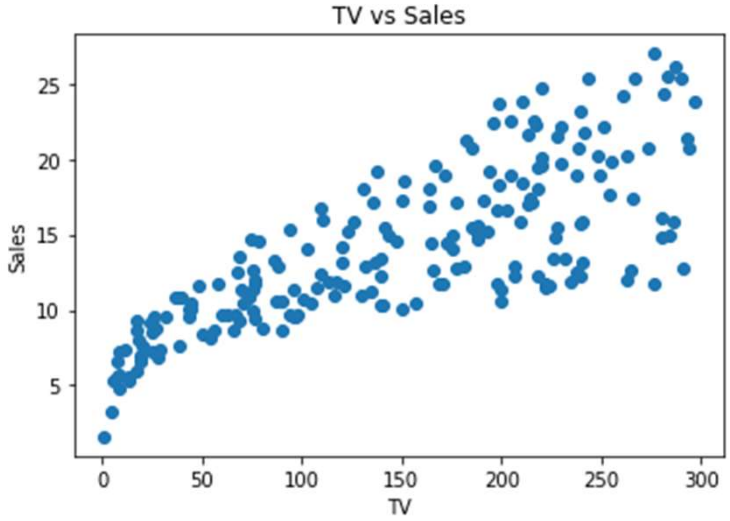
  - It is still a useful and widely used statistical learning method

  - It serves as a good jumping-off point for newer approaches:

# What can we use it for?

- Is there a relationship between predictors and response?
- How strong is the relationship between?
- Which predictors contribute to response?
- How accurately can we estimate the effect of each predictors?
- How accurately can we predict response?
- Is there synergy among the predictors?

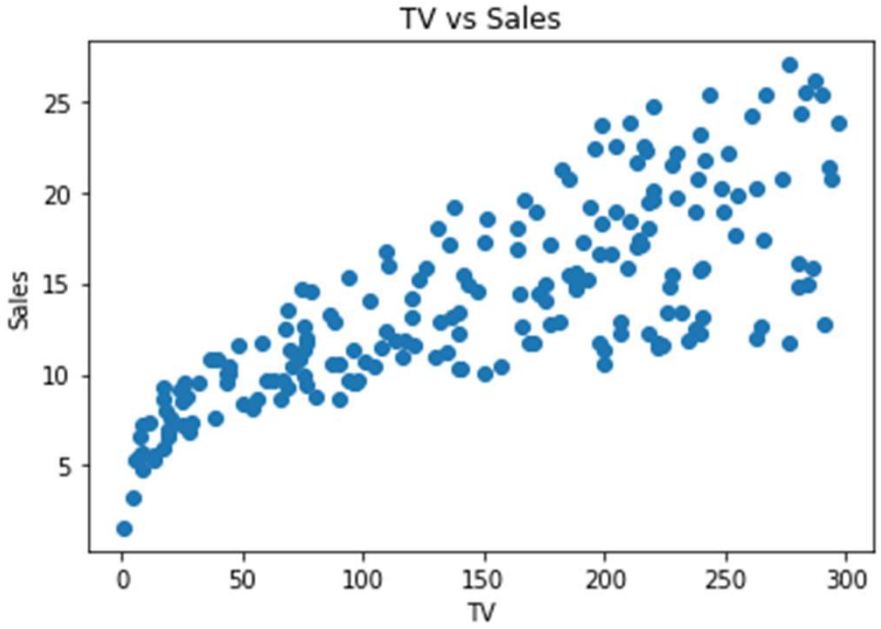
# Example (5)




$$Sales = f(TV) + \epsilon$$

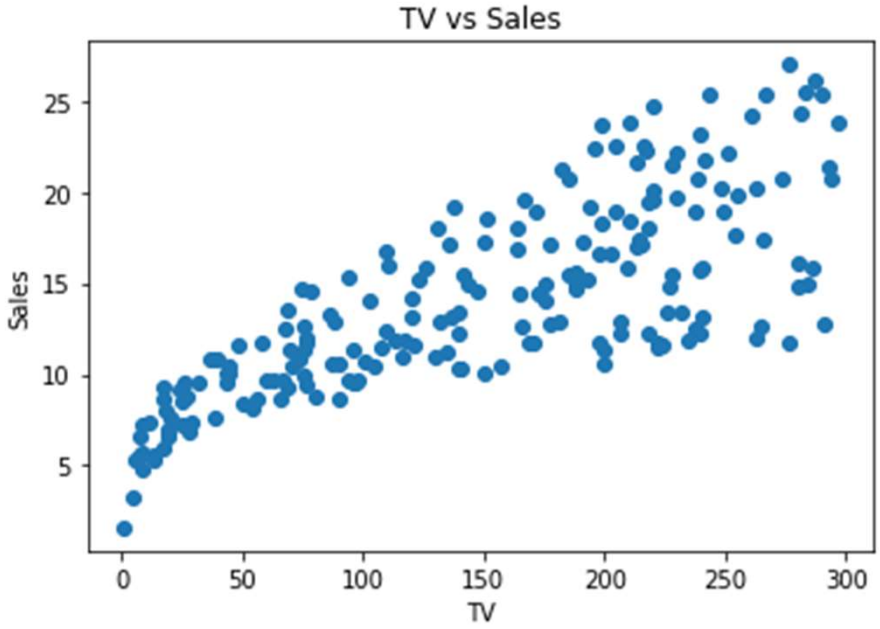
$$\widehat{Sales} \approx \hat{f}(TV)$$

# Example (6)



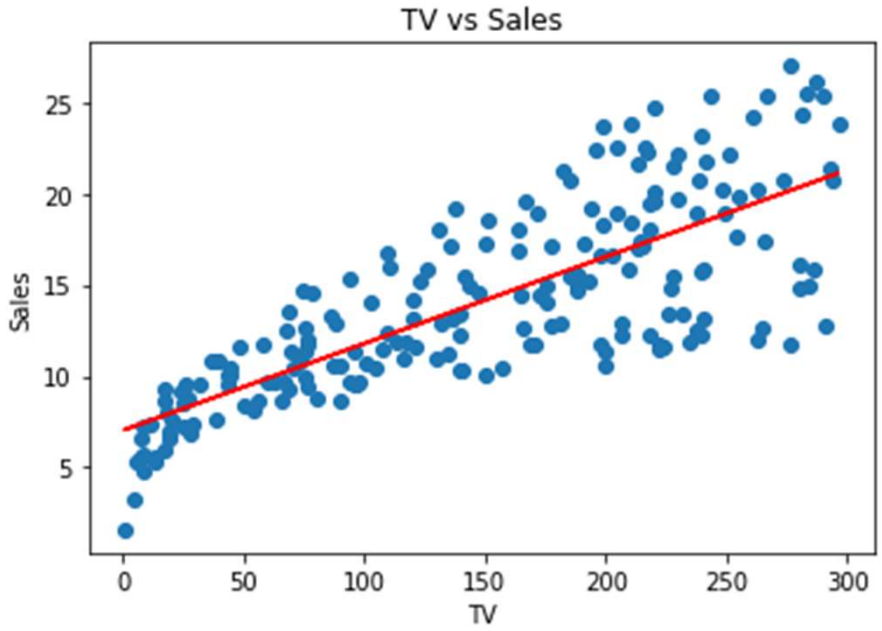

$$Sales = \beta_0 + \beta_1\, TV + \epsilon$$

# Example (7)



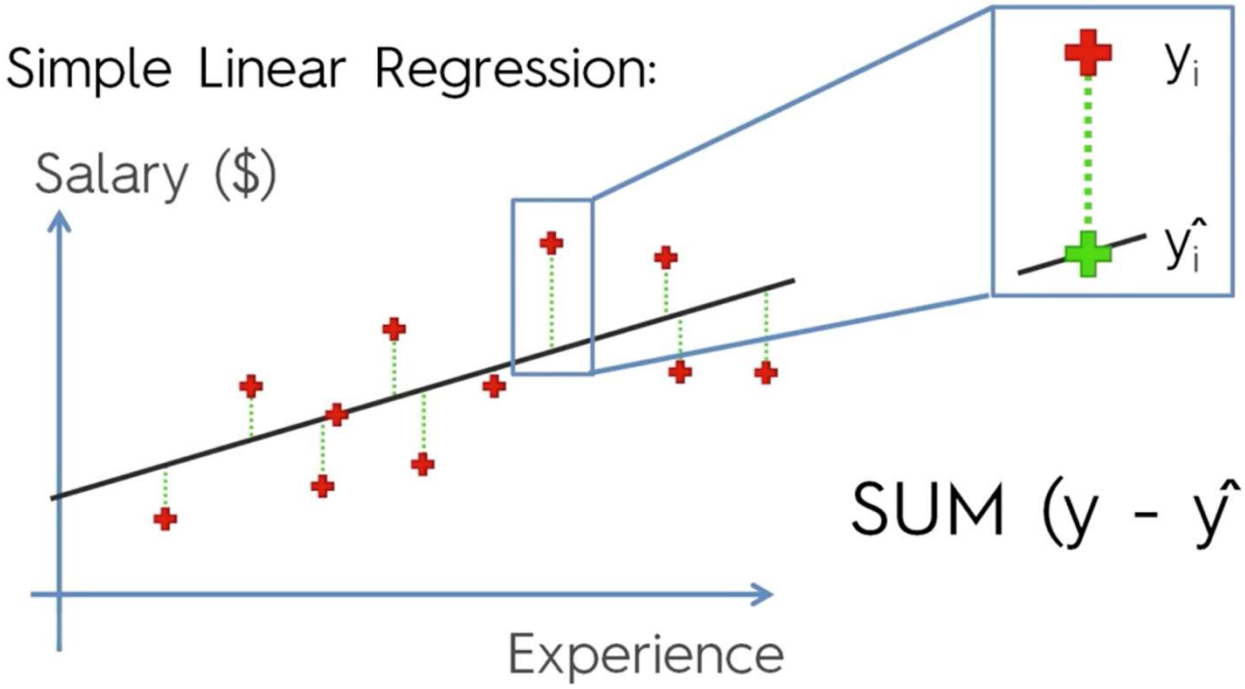

$$\widehat{Sales} \approx \hat{\beta}_0 + \hat{\beta}_1 TV$$

# Example (8)



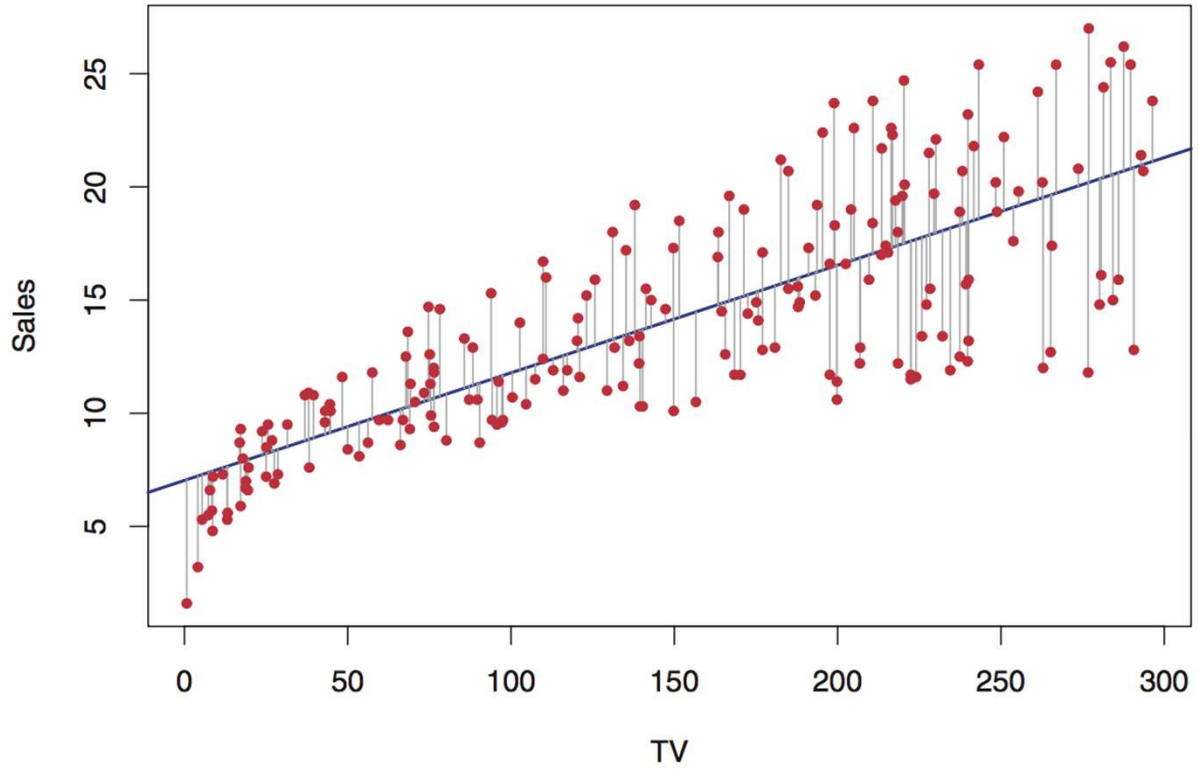

$$\widehat{Sales} \approx \hat{\beta}_0 + \hat{\beta}_1 TV$$

# How to find the optimum fit?

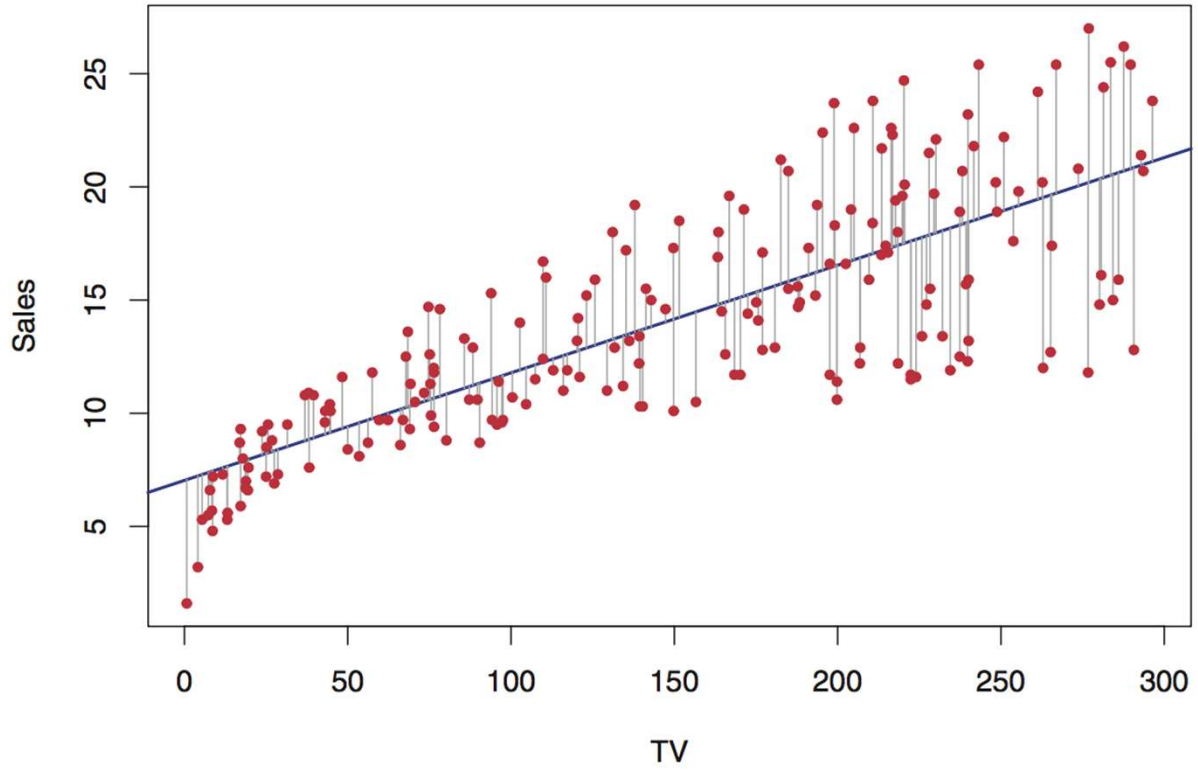



$$SUM\ (y - \hat{y})^2 \rightarrow min$$

# For our example



# More formally – Least Squares



# Estimating Parameters by Least Squares (1)

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$$e_i = y_i - \hat{y}_i$$

- Residual sum of squares

$$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2$$

$$\mathrm{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

# Estimating Parameters by Least Squares (2)

- The least squares approach chooses the parameters that minimize the RSS

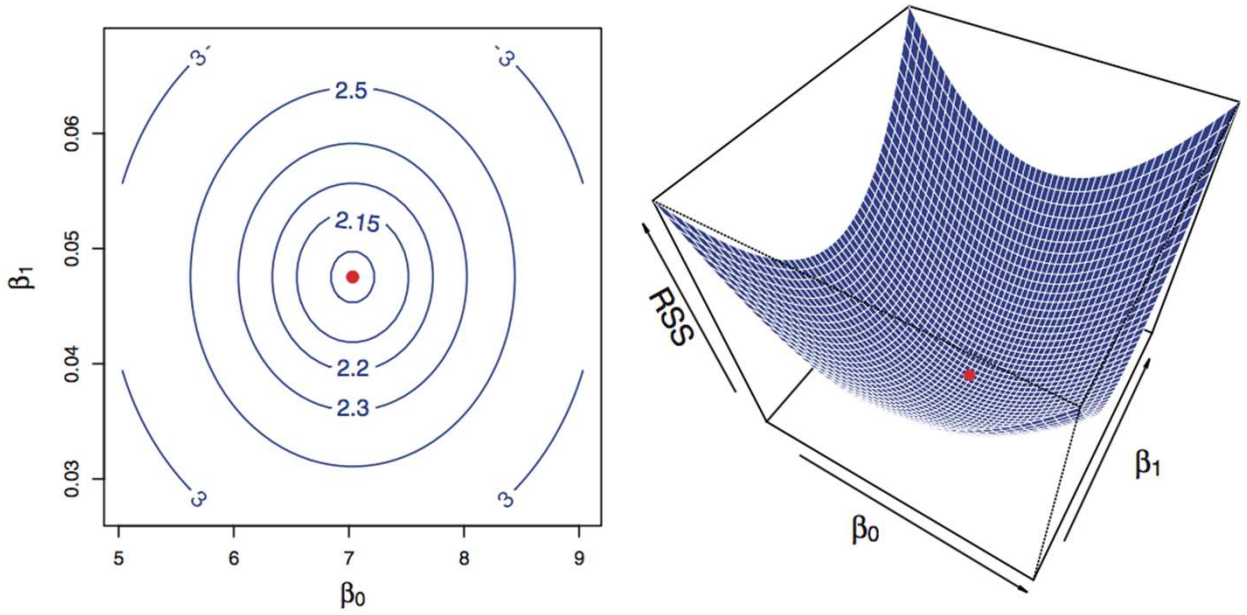
- The minimizing values can be found as

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

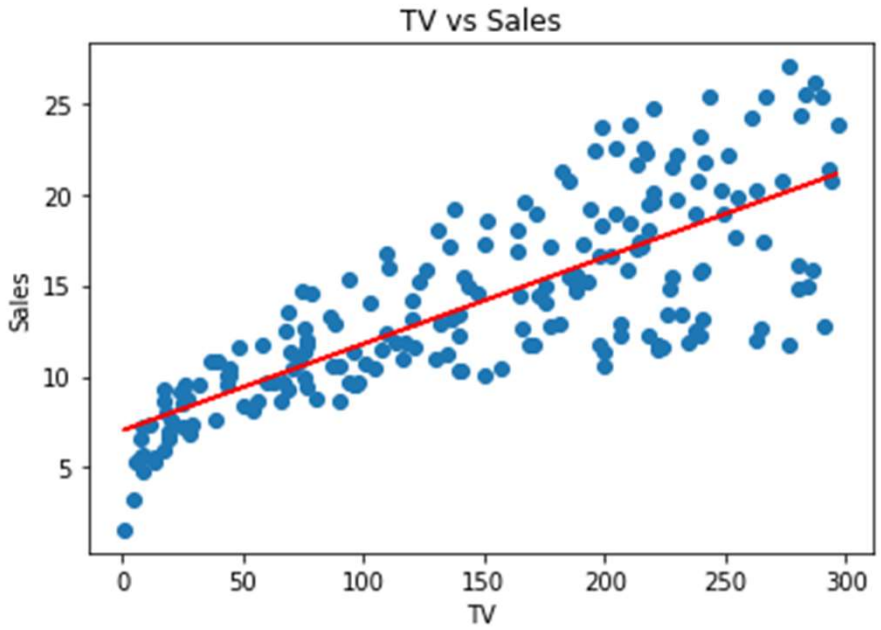
where $\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$ are the sample means.

# Estimating Parameters by Least Squares (3)




Contour and three-dimensional plots of the RSS on the Advertising data, using sales as the response and TV as the predictor. The red dots correspond to the least squares estimates

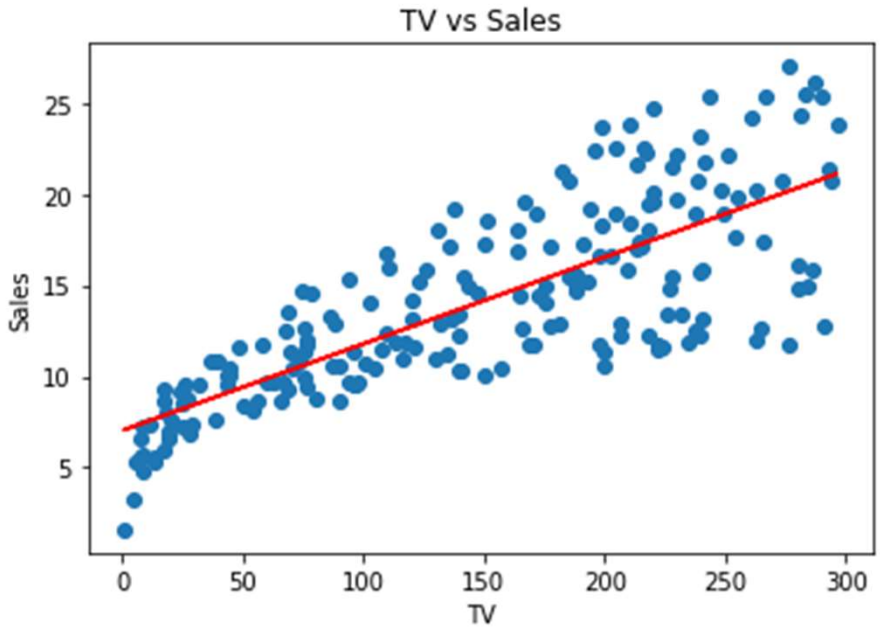
# Example (9)




$$\widehat{Sales} \approx \hat{\beta}_0 + \hat{\beta}_1 TV$$

| Parameters | Values |
|---|---|
| Intercept | 7.0326 |
| TV | 0.0475 |

# Example (10)




As per this estimation, an additional $1,000 spent on TV advertising is associated with selling approximately 47.5 additional units of the product.

| Parameters | Values |
|---|---|
| Intercept | 7.0326 |
| TV | 0.0475 |

# Answering questions with LR (1)

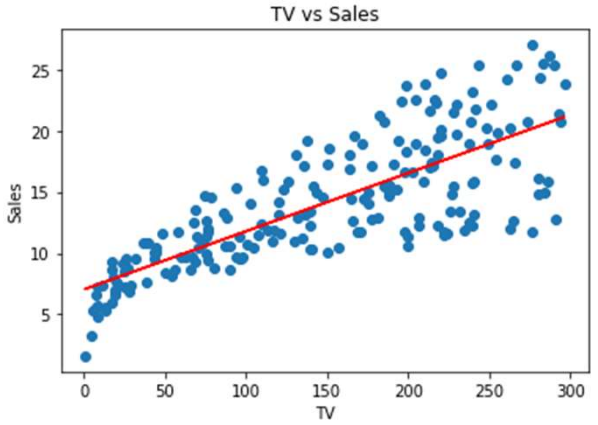
- 1. Is there a relationship between sales (response) and TV (predictor)?

- Mathematically this corresponds to

$$H_0: \beta_1 = 0$$

  - verses

$$H_a: \beta_1 \neq 0$$




$$\widehat{Sales} \approx \hat{\beta}_0 + \hat{\beta}_1 TV$$

| Parameters | Values |
|---|---|
| Intercept | 7.0326 |
| TV | 0.0475 |

# Answering questions with LR (2)

- 1. Is there a relationship between sales (response) and TV (predictor)?

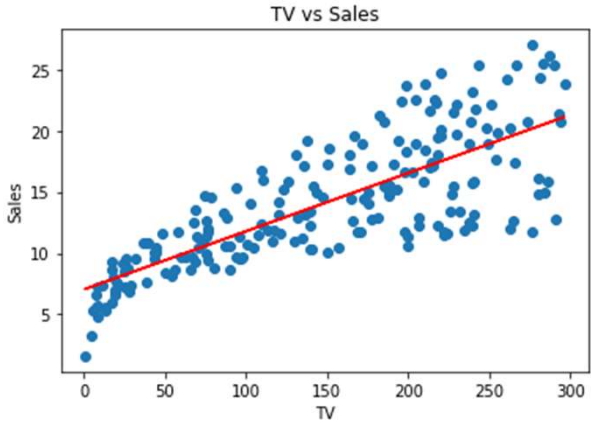
- Mathematically this corresponds to

$$H_0: \beta_1 = 0$$

  - verses

$$H_a: \beta_1 \neq 0$$

But we do not have true values for these!!!!




$$\widehat{Sales} \approx \hat{\beta}_0 + \hat{\beta}_1 TV$$

| Parameters | Values |
|---|---|
| Intercept | 7.0326 |
| TV | 0.0475 |

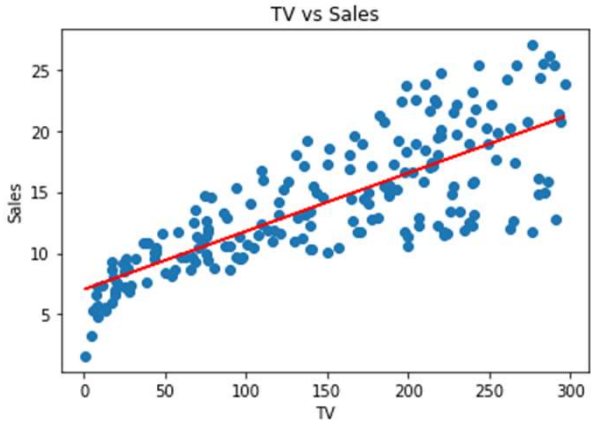
# Answering questions with LR (3)

- 1. Is there a relationship between sales (response) and TV (predictor)?

- Therefore, we calculate t-statistics

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}$$

- Where SE is an estimate of how close the estimated parameter value is to its true value

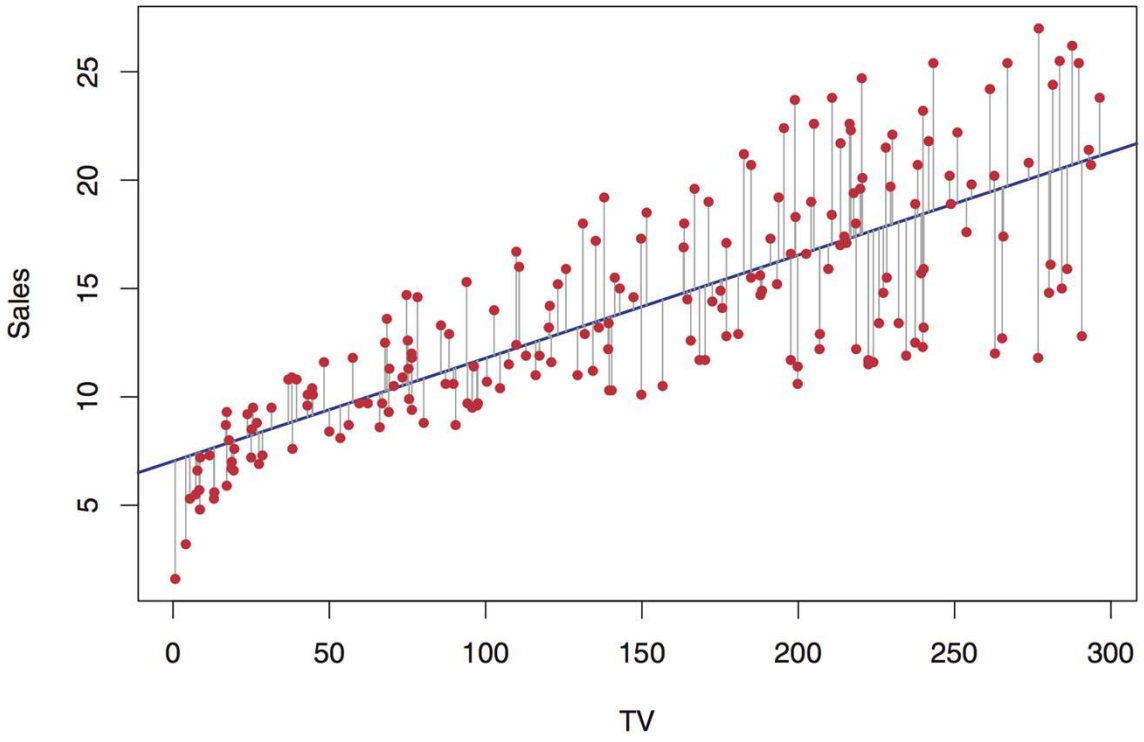



$$\widehat{Sales} \approx \hat{\beta}_0 + \hat{\beta}_1 TV$$

| Parameters | Values |
|---|---|
| Intercept | 7.0326 |
| TV | 0.0475 |

# Aside: SE

$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

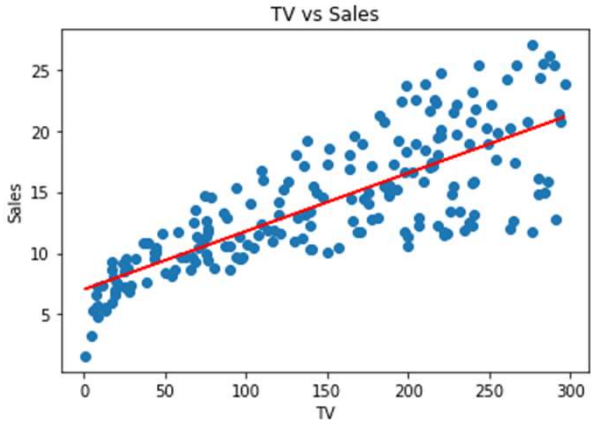


# Answering questions with LR (4)

- 1. Is there a relationship between sales (response) and TV (predictor)?

- Therefore, we calculate t-statistics

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}$$




$$\widehat{Sales} \approx \hat{\beta}_0 + \hat{\beta}_1 TV$$

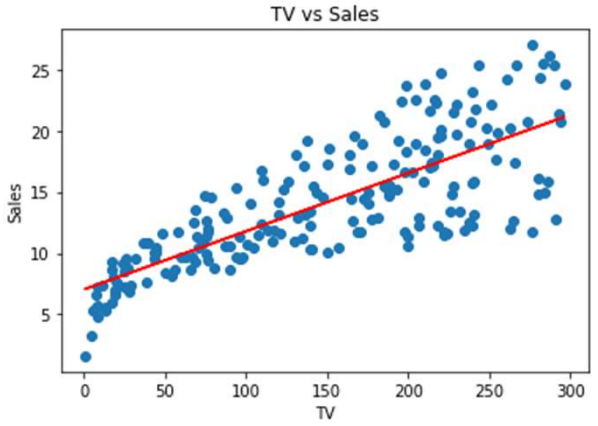
| Parameters | Values | t-value |
|---|---|---|
| Intercept | 7.0326 | 15.360 |
| TV | 0.0475 | 17.668 |

# Answering questions with LR (5)

- 1. Is there a relationship between sales (response) and TV (predictor)?

- Finally, we calculate *p-value*

  - Probability of getting $|t|$ assuming $\beta_1$ was 0




$$\widehat{Sales} \approx \hat{\beta}_0 + \hat{\beta}_1 TV$$

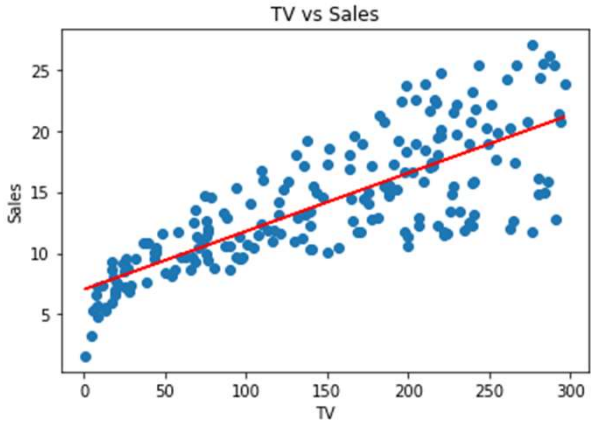
| Parameters | Values | t-value |
|---|---|---|
| Intercept | 7.0326 | 15.360 |
| TV | 0.0475 | 17.668 |

# Answering questions with LR (6)

- 1. Is there a relationship between sales (response) and TV (predictor)?

- Finally, we calculate *p-value*

  - Probability of getting $|t|$ assuming $\beta_1$ was 0

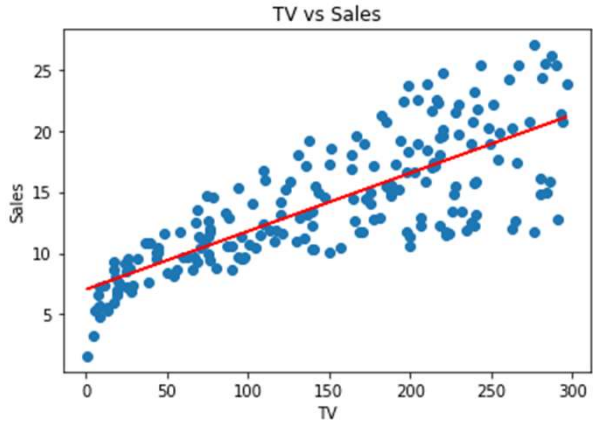



$$\widehat{Sales} \approx \hat{\beta}_0 + \hat{\beta}_1 TV$$

| Parameters | Values | t-value | p-value |
|---|---|---|---|
| Intercept | 7.0326 | 15.360 | < 0.0001 |
| TV | 0.0475 | 17.668 | < 0.0001 |

# Answering questions with LR (7)

- 2. What is the extent to which the model fits the data?



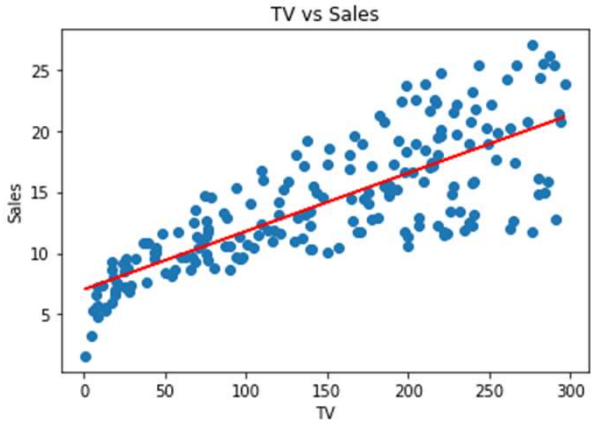

$$\widehat{Sales} \approx \hat{\beta}_0 + \hat{\beta}_1 TV$$

| Parameters | Values | t-value | p-value |
|---|---|---|---|
| Intercept | 7.0326 | 15.360 | < 0.0001 |
| TV | 0.0475 | 17.668 | < 0.0001 |

# Answering questions with LR (8)

- 2. What is the extent to which the model fits the data?

- This can judged using $R^2$ *statistics*




$$\widehat{Sales} \approx \hat{\beta}_0 + \hat{\beta}_1 TV$$

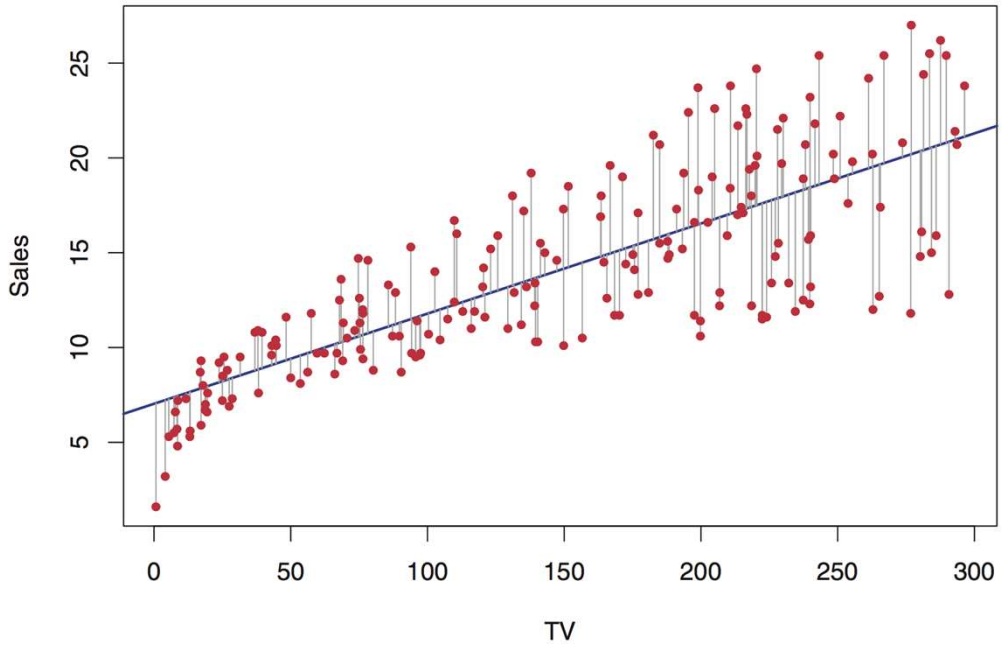
| Parameters | Values | t-value | p-value |
|---|---|---|---|
| Intercept | 7.0326 | 15.360 | < 0.0001 |
| TV | 0.0475 | 17.668 | < 0.0001 |

# Calculating $R^2$ *statistics*

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

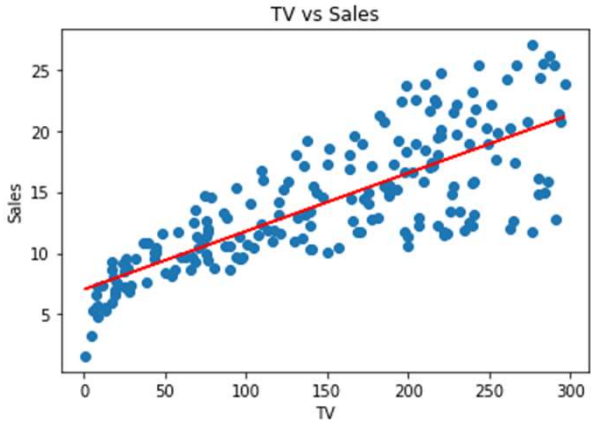
$$\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

$$\text{TSS} = \sum(y_i - \bar{y})^2$$



# Answering questions with LR (9)

- 2. What is the extent to which the model fits the data?
- This can judged using $R^2$ *statistics*
- In this case, it is 0.612



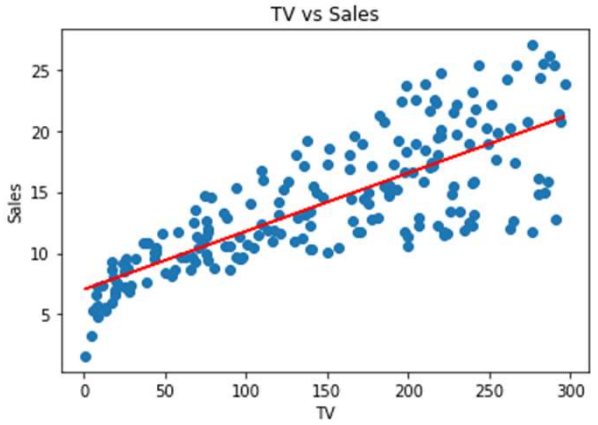

$$\widehat{Sales} \approx \hat{\beta}_0 + \hat{\beta}_1 TV$$

| Parameters | Values | t-value | p-value |
|---|---|---|---|
| Intercept | 7.0326 | 15.360 | < 0.0001 |
| TV | 0.0475 | 17.668 | < 0.0001 |

# Answering questions with LR (10)

- 3. Increasing the budget for TV will cause how much increase/decrease in sales?

- This can judged the value of the coefficient




$$\widehat{Sales} \approx \hat{\beta}_0 + \hat{\beta}_1 TV$$

| Parameters | Values | t-value | p-value |
|---|---|---|---|
| Intercept | 7.0326 | 15.360 | < 0.0001 |
| TV | 0.0475 | 17.668 | < 0.0001 |

# Multiple Linear Regression (1)

- Simple linear regression is a useful approach for predicting a response on the basis of a single predictor variable.

- However, in practice we often have more than one predictor

  - Sales (TV, Radio, Newspaper)
  - Income (Years of education, Years of experience, Age, Gender)

# Multiple Linear Regression (2)

- Options

1. Fit $p$ separate linear regressions (where $p$ is the number of predictors)

2. Extend the simple linear regression model, so that it can directly accommodate multiple predictors

# Multiple Linear Regression (3)

- Options

1. Fit $p$ separate linear regressions (where $p$ is the number of predictors)

2. Extend the simple linear regression model, so that it can directly accommodate multiple predictors

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

# Multiple Linear Regression (4)

- For $p$ predictors,

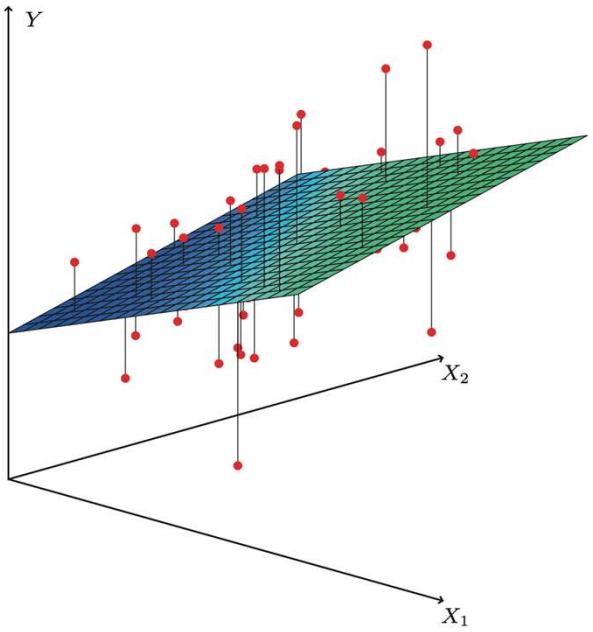
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

- The parameters are estimated using the same least squares approach that we saw in the context of simple linear regression

- The coefficients can be calculated using statistical packages

# Multiple Linear Regression (5)

- For two predictors, the regression might look as follows



# Multiple Linear Regression (6)

| Parameters | Values | t-value | p-value |
|---|---|---|---|
| Intercept | 2.939 | 9.42 | < 0.0001 |
| TV | 0.46 | 32.81 | < 0.0001 |
| Radio | 0.189 | 21.89 | < 0.0001 |
| Newspaper | -0.001 | -0.18 | < 0.8599 |

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper} + \epsilon.$$

For the Advertising data, least squares coefficient estimates of the multiple linear regression of number of units sold on radio, TV, and newspaper advertising budgets.

# Multiple Linear Regression (7)

| Parameters | Values | t-value | p-value |
|---|---|---|---|
| Intercept | 2.939 | 9.42 | < 0.0001 |
| TV | 0.46 | 32.81 | < 0.0001 |
| Radio | 0.189 | 21.89 | < 0.0001 |
| Newspaper | -0.001 | -0.18 | < 0.8599 |

Compare the results for 'Newspaper' of **multiple regression (above)** to that of **linear regression** (above)

| Parameters | Values | t-value | p-value |
|---|---|---|---|
| Intercept | 12.351 | 19.88 | < 0.0001 |
| Newspaper | 0.055 | 3.30 | 0.00115 |

# Multiple Linear Regression (7)

| Parameters | Values | t-value | p-value |
|---|---|---|---|
| Intercept | 2.939 | 9.42 | < 0.0001 |
| TV | 0.46 | 32.81 | < 0.0001 |
| Radio | 0.189 | 21.89 | < 0.0001 |
| Newspaper | -0.001 | -0.18 | < 0.8599 |

Correlation matrix for TV, radio, newspaper, and sales for the Advertising data

| | TV | radio | newspaper | sales |
|---|---|---|---|---|
| TV | 1.0000 | 0.0548 | 0.0567 | 0.7822 |
| radio | | 1.0000 | 0.3541 | 0.5762 |
| newspaper | | | 1.0000 | 0.2283 |
| sales | | | | 1.0000 |

# Answering questions with MLR (1)

- 1. Is at least one of the predictors useful in predicting the response?
  - We might think that (just like LR) we can use p-value for this, but **we are wrong**

| Parameters | Values | t-value | p-value |
|---|---|---|---|
| Intercept | 2.939 | 9.42 | < 0.0001 |
| TV | 0.46 | 32.81 | < 0.0001 |
| Radio | 0.189 | 21.89 | < 0.0001 |
| Newspaper | -0.001 | -0.18 | < 0.8599 |

# Answering questions with MLR (2)

- 1. Is at least one of the predictors useful in predicting the response?
  - Thus we use another measure called F-statistics

$$F = \frac{(\mathrm{TSS} - \mathrm{RSS})/p}{\mathrm{RSS}/(n-p-1)}$$

# Answering questions with MLR (3)

- 1. Is at least one of the predictors useful in predicting the response?
  - Thus we use another measure called F-statistics

| Parameters | Values | t-value | p-value |
|---|---|---|---|
| Intercept | 2.939 | 9.42 | < 0.0001 |
| TV | 0.46 | 32.81 | < 0.0001 |
| Radio | 0.189 | 21.89 | < 0.0001 |
| Newspaper | -0.001 | -0.18 | < 0.8599 |

| F-statistics | 570 |
|---|---|

Since this is far larger than 1, it provides compelling evidence against the null hypothesis H0.
In other words, the large F-statistic suggests that at least one of the advertising media must be related to sales

# Answering questions with MLR (4)

- 1. Is at least one of the predictors useful in predicting the response?
  - But how far away from 0 F-statistics has to be?

# Answering questions with MLR (5)

- 2. Do all the predictors help explain the response or is only a subset of them useful?
  - Forward selection
  - Backward selection
  - Mixed selection

# Answering questions with MLR (6)

- 2. How well does the model fit the data?

  - Same as LR (R-squared)

# Example (1) – Dummy Variable

| Profit | R&D Spend | Admin | Marketing | State |
|---|---|---|---|---|
| 192,261.83 | 165,349.20 | 136,897.80 | 471,784.10 | New York |
| 191,792.06 | 162,597.70 | 151,377.59 | 443,898.53 | California |
| 191,050.39 | 153,441.51 | 101,145.55 | 407,934.54 | California |
| 182,901.99 | 144,372.41 | 118,671.85 | 383,199.62 | New York |
| 166,187.94 | 142,107.34 | 91,391.77 | 366,168.42 | California |

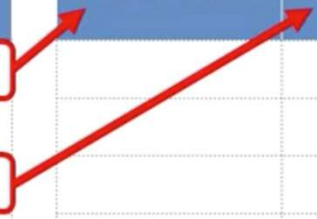
# Example (2) – Dummy Variable

| Profit | R&D Spend | Admin | Marketing | State |
|---|---|---|---|---|
| 192,261.83 | 165,349.20 | 136,897.80 | 471,784.10 | New York |
| 191,792.06 | 162,597.70 | 151,377.59 | 443,898.53 | California |
| 191,050.39 | 153,441.51 | 101,145.55 | 407,934.54 | California |
| 182,901.99 | 144,372.41 | 118,671.85 | 383,199.62 | New York |
| 166,187.94 | 142,107.34 | 91,391.77 | 366,168.42 | California |

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + ???$$

# Example (3) – Dummy Variable

| Profit | R&D Spend | Admin | Marketing | State | New York | California |
|---|---|---|---|---|---|---|
| 192,261.83 | 165,349.20 | 136,897.80 | 471,784.10 | New York | | |
| 191,792.06 | 162,597.70 | 151,377.59 | 443,898.53 | California | | |
| 191,050.39 | 153,441.51 | 101,145.55 | 407,934.54 | California | | |
| 182,901.99 | 144,372.41 | 118,671.85 | 383,199.62 | New York | | |
| 166,187.94 | 142,107.34 | 91,391.77 | 366,168.42 | California | | |



# Example (4) – Dummy Variable

| Profit | R&D Spend | Admin | Marketing | State | New York | California |
|---|---|---|---|---|---|---|
| 192,261.83 | 165,349.20 | 136,897.80 | 471,784.10 | New York | 1 | |
| 191,792.06 | 162,597.70 | 151,377.59 | 443,898.53 | California | 0 | |
| 191,050.39 | 153,441.51 | 101,145.55 | 407,934.54 | California | 0 | |
| 182,901.99 | 144,372.41 | 118,671.85 | 383,199.62 | New York | 1 | |
| 166,187.94 | 142,107.34 | 91,391.77 | 366,168.42 | California | 0 | |

# Example (5) – Dummy Variable

| Profit | R&D Spend | Admin | Marketing | State | New York | California |
|---|---|---|---|---|---|---|
| 192,261.83 | 165,349.20 | 136,897.80 | 471,784.10 | New York | 1 | 0 |
| 191,792.06 | 162,597.70 | 151,377.59 | 443,898.53 | California | 0 | 1 |
| 191,050.39 | 153,441.51 | 101,145.55 | 407,934.54 | California | 0 | 1 |
| 182,901.99 | 144,372.41 | 118,671.85 | 383,199.62 | New York | 1 | 0 |
| 166,187.94 | 142,107.34 | 91,391.77 | 366,168.42 | California | 0 | 1 |

# Example (6) – Dummy Variable

| Profit | R&D Spend | Admin | Marketing | State | Dummy Variables: New York | Dummy Variables: California |
|---|---|---|---|---|---|---|
| 192,261.83 | 165,349.20 | 136,897.80 | 471,784.10 | New York | 1 | 0 |
| 191,792.06 | 162,597.70 | 151,377.59 | 443,898.53 | California | 0 | 1 |
| 191,050.39 | 153,441.51 | 101,145.55 | 407,934.54 | California | 0 | 1 |
| 182,901.99 | 144,372.41 | 118,671.85 | 383,199.62 | New York | 1 | 0 |
| 166,187.94 | 142,107.34 | 91,391.77 | 366,168.42 | California | 0 | 1 |

# Example (7) – Dummy Variable

| Profit | R&D Spend | Admin | Marketing | State | Dummy Variables: New York | Dummy Variables: California |
|---|---|---|---|---|---|---|
| 192,261.83 | 165,349.20 | 136,897.80 | 471,784.10 | New York | 1 | 0 |
| 191,792.06 | 162,597.70 | 151,377.59 | 443,898.53 | California | 0 | 1 |
| 191,050.39 | 153,441.51 | 101,145.55 | 407,934.54 | California | 0 | 1 |
| 182,901.99 | 144,372.41 | 118,671.85 | 383,199.62 | New York | 1 | 0 |
| 166,187.94 | 142,107.34 | 91,391.77 | 366,168.42 | California | 0 | 1 |

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + ???$$

# Example (8) – Dummy Variable

| Profit | R&D Spend | Admin | Marketing | State | Dummy Variables: New York | Dummy Variables: California |
|---|---|---|---|---|---|---|
| 192,261.83 | 165,349.20 | 136,897.80 | 471,784.10 | New York | 1 | 0 |
| 191,792.06 | 162,597.70 | 151,377.59 | 443,898.53 | California | 0 | 1 |
| 191,050.39 | 153,441.51 | 101,145.55 | 407,934.54 | California | 0 | 1 |
| 182,901.99 | 144,372.41 | 118,671.85 | 383,199.62 | New York | 1 | 0 |
| 166,187.94 | 142,107.34 | 91,391.77 | 366,168.42 | California | 0 | 1 |

$$y = b_0 + b_1*x_1 + b_2*x_2 + b_3*x_3 + b_4*D_1 + b_5*D_2$$

# Dummy Variable Trap - Multicollinearity


$$y = b_0 + b_1*x_1 + b_2*x_2 + b_3*x_3 + b_4*D_1 + b_5*D_2$$

# Dummy Variable Trap - Multicollinearity

| Profit | R&D Spend | Admin | Marketing | State | New York | California |
|---|---|---|---|---|---|---|
| 192,261.83 | 165,349.20 | 136,897.80 | 471,784.10 | New York | 1 | 0 |
| 191,792.06 | 162,597.70 | 151,377.59 | 443,898.53 | California | 0 | 1 |
| 191,050.39 | 153,441.51 | 101,145.55 | 407,934.54 | California | 0 | 1 |
| 182,901.99 | 144,372.41 | 118,671.85 | 383,199.62 | New York | 1 | 0 |
| 166,187.94 | 142,107.34 | 91,391.77 | 366,168.42 | California | 0 | 1 |

**Dummy Variables**

$$y = b_0 + b_1*x_1 + b_2*x_2 + b_3*x_3 + b_4*D_1 + b_5*D_2$$

**Always omit one dummy variable**

# Did we achieve today's objectives objectives?

- What is linear regression?
- Why study linear regression?
- What can we use it for?
- How to perform linear regression?
- How to estimate its performance?
- What are its extensions?
- What are dummy variables?