



# Applied Business Analytics

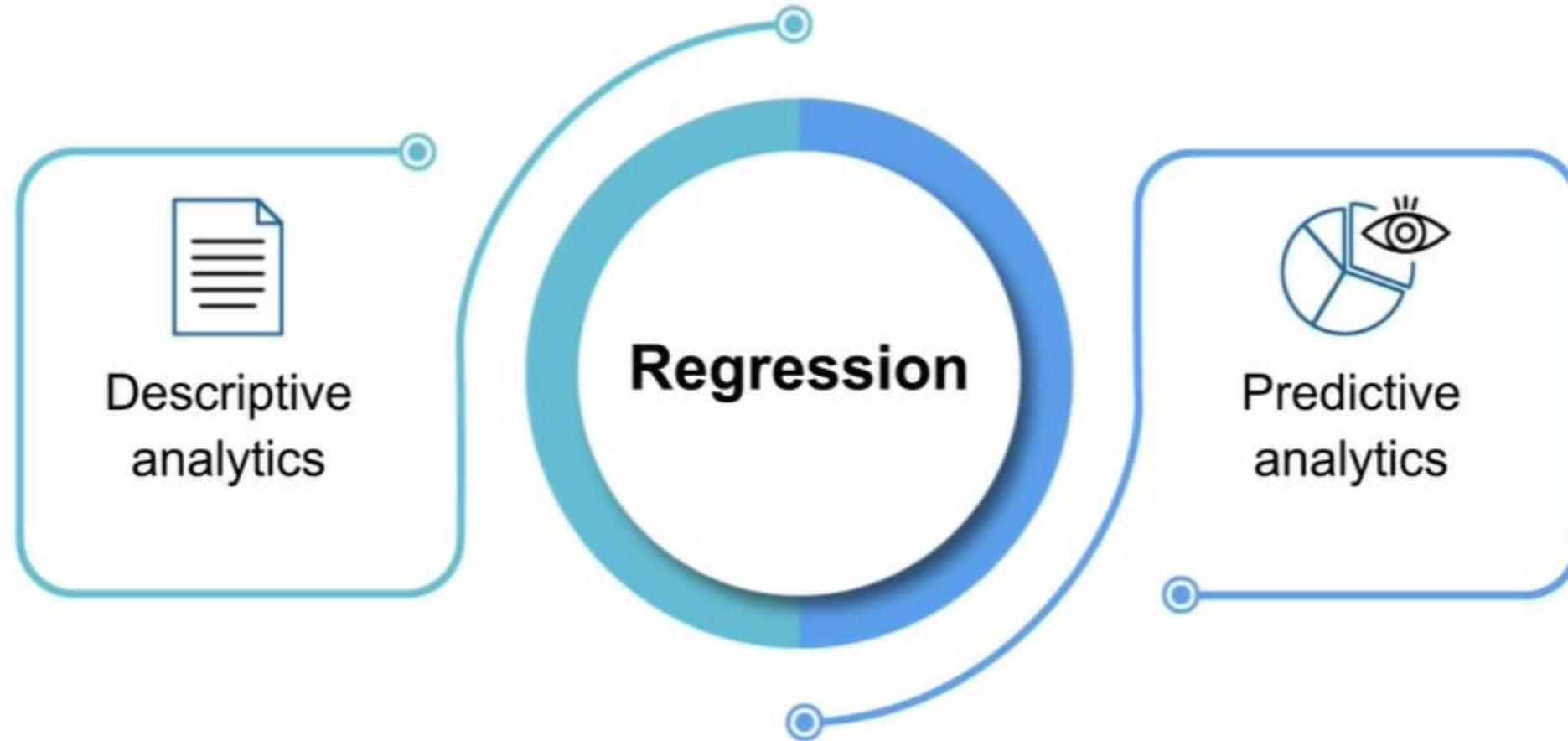
## Week 3: Regression – Descriptive and Predictive Analytics

# Recap

In the previous module, we looked at:

- Data pre-processing steps
- Data types used in business analytics
- Data cleaning and wrangling
- Missing values imputation
- Dummy variable coding for non-metric data
- Basics of statistics

# Regression Analysis



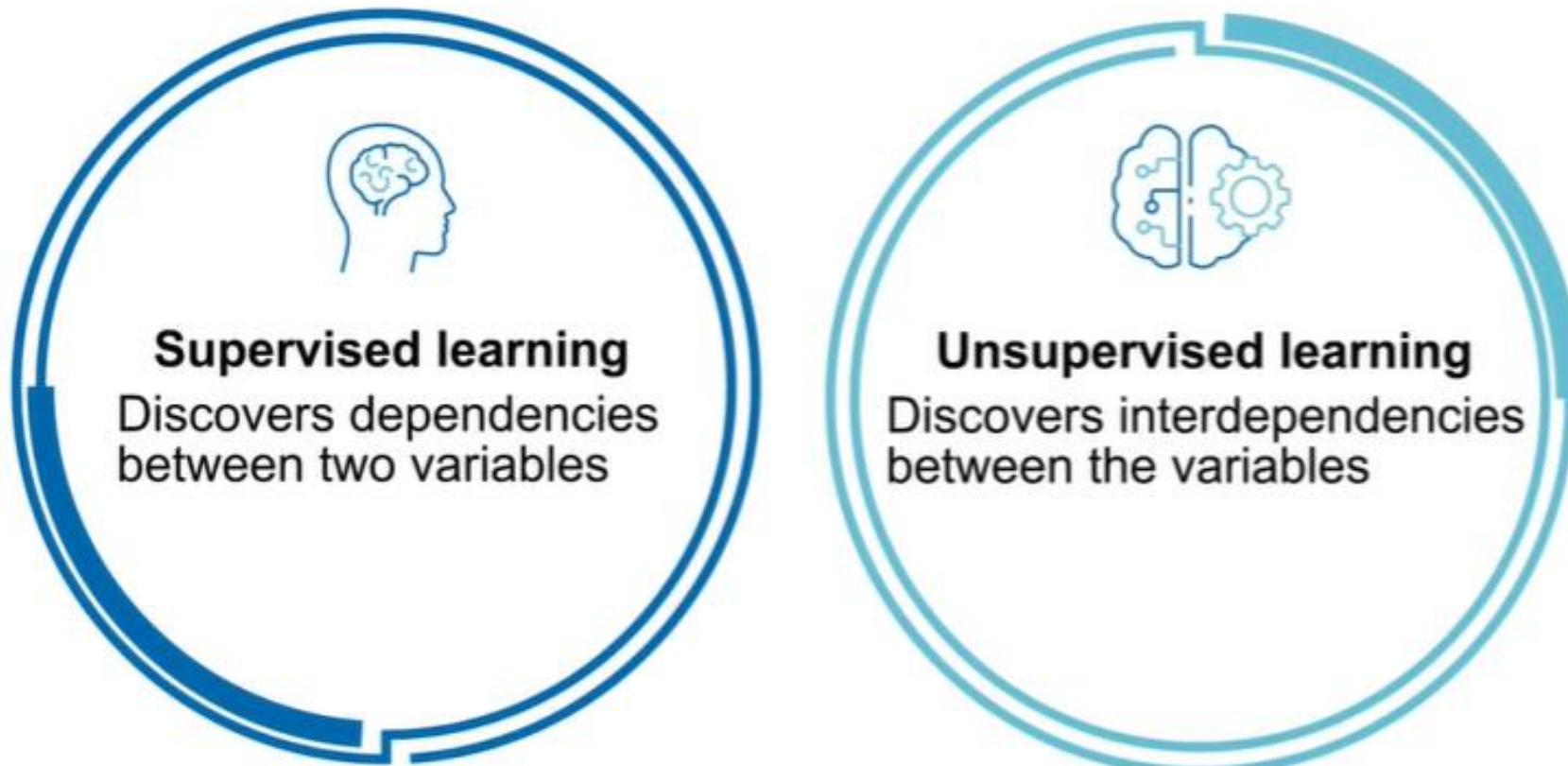
# ML Approaches in Business Analytics

Questions that businesses are interested in:

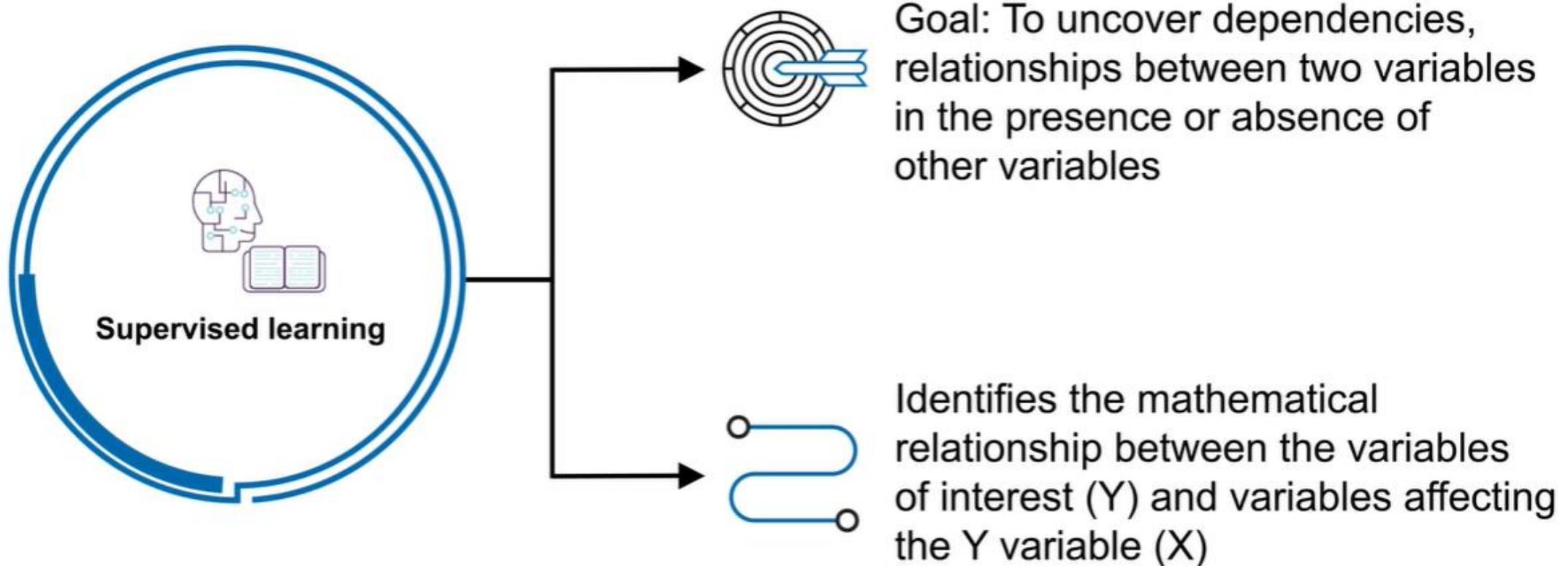
- How do variables affect each other?  
(Supervised learning)
- How to identify novel patterns in data?

# Machine Learning Domain

## Pattern recognition



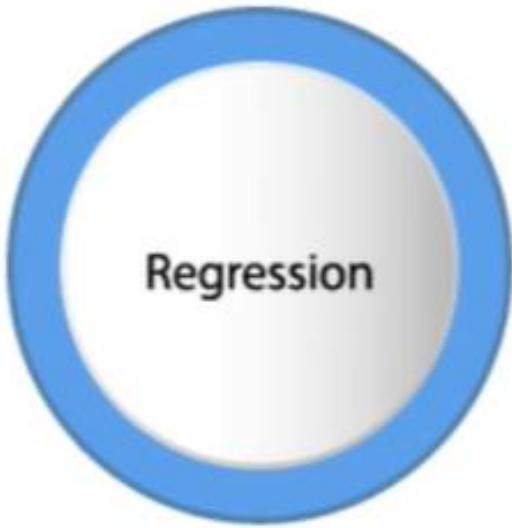
# Machine Learning Domain



# Machine Learning

- Machines identify the relationship between variables
- Machines are provided with past decisions to identify the decision-making models and processes

# Regression

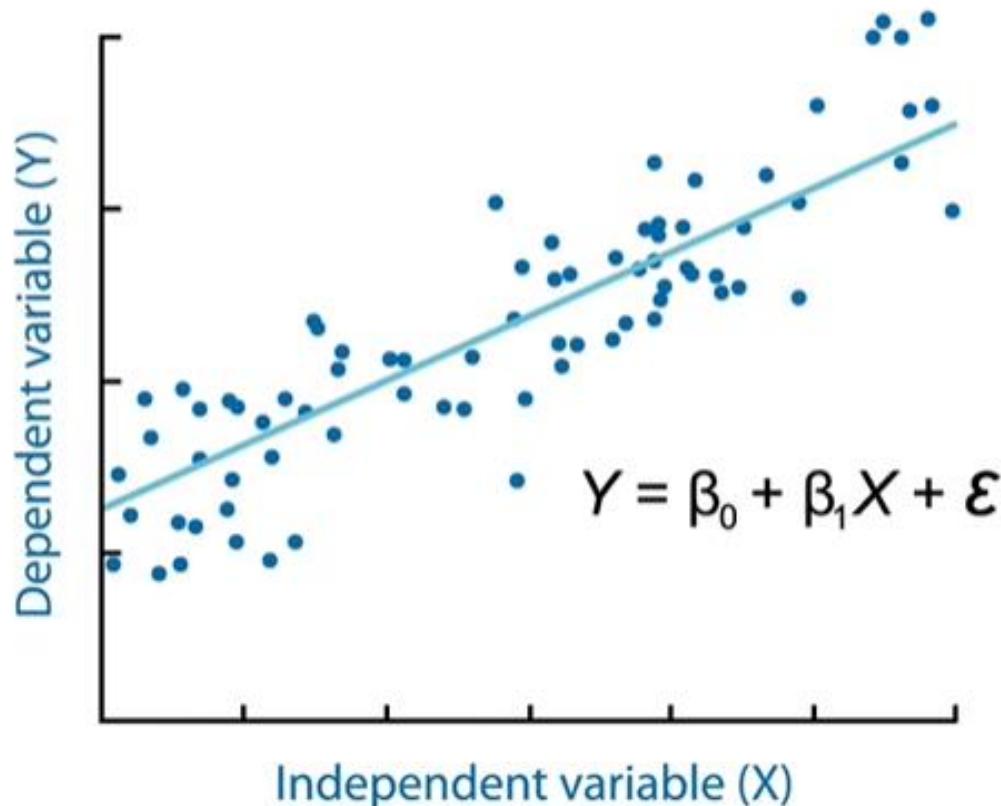


Dependent variable  
(Y)

Independent variable  
(X)

# Regression Analysis

The goal of regression is to identify the pattern and to find the equation that best fits the pattern.



# Managerial Questions: Descriptive Analysis

1

What is the impact of featuring a brand  
in a flyer on unit sales of a brand?

2

What is the impact of display on unit  
sales of a brand?

# Managerial Questions: Predictive Analysis

- 1 How does a brand's price affect its sales?
- 2 Does it affect the brand if a competitor changes the price? By how much?
- 3 If the store brand's price is changed, how does it affect the sales of Brand A?
- 4 Are there synergistic effects between two marketing activities, for example, feature and display marketing?
- 5 Can we predict the sales for the next period?

# Data-Driven Decision Making

01



- What is the **business challenge** to be addressed?
- What to achieve and why?

02



- What kind of **information** is available?
- Can we address the problem with the current information?
- Do we need additional information and what kind?

03



- What kind of **tools** can be applied to get to that information?
- What kind of **data** would be needed to gain insight?

04



- How to interpret the results, make predictions and **gain insights** using the tool?

# Modeling Basics: From Concept to Analytical Tools

Independent variables

$$\text{Sales} = f(\text{Own Price, Compt\_Price, Feature, Display, etc.})$$

Dependent variable

# Modeling Basics: From Concept to Analytical Tools

## Regression model

$$\text{Sales} = \beta_0 + \beta_1 * \text{Price} + \beta_2 * \text{Compt Price} + \beta_3 * \text{Display} + \beta_4 * \text{Feature} + \epsilon$$

Y variable

Affect the sales

## Conceptual model

$$\text{Sales} = f(\text{Own Price}, \text{Compt Price}, \text{Feature}, \text{Display}, \text{etc.})$$

# Example: Weekly Sales and Marketing Data

A	B	C	D	E	F	G	
1	Week	Sales Brand A	Price Brand A	Price Brand B	Price Brand C	Feature Brand A	Display Brand A
2	40	6528	3.66	1.89	2.99	1	0
3	43	6016	3.66	2.84	2.99	0	0
4	44	6272	3.66	2.84	2.59	0	0
5	45	6848	3.66	2.39	2.99	0	0
6	46	7424	3.66	2.84	2.99	0	0
7	47	6848	3.66	2.84	2.39	0	0
8	48	7488	3.66	2.39	2.99	0	0
9	49	6336	3.66	2.39	1.99	0	0
10	50	6208	3.66	1.99	2.99	0	0
11	51	6400	3.66	2.84	2.99	0	0
12	52	13056	3.29	2.84	2.99	1	0
13	53	8704	3.29	2.84	2.19	1	0
14	54	8832	3.29	2.39	2.99	1	0
15	55	5696	3.66	2.39	2.99	0	0
16	56	14208	3.29	2.84	2.99	1	1
17	57	7616					0

Weekly sales data

Data about the brand, where it was displayed

Data about your price and competitor's price

Data about the brand, if it was featured in a flyer

# Regression App

## Regression App



How to use this application

This application requires one data input from the user. To do so, click on the Browse (in left side-bar panel) and upload the csv data input file. Note that this application can read only csv file (comma delimited file), so if you don't have csv input data file, first convert your data in csv format and then proceed. Make sure you have top row as variable names.

Once csv file is uploaded successfully, variables in the data file will reflect in the 'Data Selection' panel on the left. Now you can select dependent variable (Y Variable) from drop-down menu. By default all other remaining variables will be selected as explanatory variables (X variables). If you want to drop any variable from explanatory variables, just uncheck that variable and it will be dropped from the model. If any of the variables selected in explanatory variables is a factor variable, you can define that variable as factor variable just by selecting that variable in the last list of variables

-Wikipedia

Upload data

Download sample data

Apply Changes

Browse... No file selected

Upload input data (csv file with header)

Data Input

Overview Data Summary Data Visualization Summary OLS Residuals Plot Prediction Input Data Prediction New Data

Data Selection

# Regression App

### Data Input

Upload input data (csv file with header)

Browse... Orange Juice Sales.csv  
Upload complete

### Data Selection

Select Y variable

Sales.Brand.A

Select X variables

obs  
 week  
 Price.Brand.A  
 Price.Brand.B  
 Price.Brand.C  
 Feature  
 Display

Select factor (categorical) variables in X

week  
 Price.Brand.A  
 Price.Brand.B  
 Price.Brand.C  
 Feature  
 Display

Select sub sample

quick run, 1,000 obs

Impute missing values or drop missing value rows

do not impute or drop rows

Apply Changes

Overview Data Summary Data Visualization Summary OLS Residuals Plot Prediction Input Data Prediction New Data

Review Input Data

Show 10 entries Search:

obs	Sales.Brand.A	week	Price.Brand.A	Price.Brand.B	Price.Brand.C	Feature	Display
1	6528	40	3.66	1.89	2.99	1	0
2	6016	43	3.66	2.04	2.00	0	0
3	6272	44	3.66			0	0
4	6848	45	3.66			0	0
5	7424	46	3.66	2.84	2.99	0	0
6	6848	47	3.66	2.84	2.39	0	chk
7	7488	48	3.66	2.39	2.99	0	0
8	6336	49	3.66	2.39	1.99	0	0
9	6208	50	3.66	1.99	2.99	0	0
10	6400	51	3.66	2.84	2.99	0	0

Showing 1 to 10 of 116 entries

Previous 1 2 3 4 5 ... 12 Next

Categorical variable

### Data Summary of Selected Y and X Variables

```
$Dimensions  
[1] 116 7  
  
$Summary  
$Summary$Numeric.data  
Sales.Brand.A      week Price.Brand.A Price.Brand.B Price.Brand.C Feature  
min   4668.00 40.0000  1.6900  1.4900  1.2900  0.0000  
max   98624.00 160.0000 3.6600  2.8900  2.9900  1.0000  
range  94016.00 120.0000 1.9700  1.4000  1.7000  1.0000  
median 10880.00 102.5000 2.9900  2.3500  2.2600  1.0000  
mean   18879.45 101.0448 2.9006  2.3018  2.2682  0.5259  
var    404155133.47 1193.2453 0.3088  0.1627  0.1525  0.2515  
std.dev 20103.61 34.5434 0.5557  0.4833  0.3905  0.5015
```

ISB | Executive Education

# Regression App

Data Input

Upload input data (csv file with header)

Browse... Orange Juice Sales.csv

Upload complete

Data Selection

Select Y variable

Sales.Brand.A

Select X variables

obs

week

Price.Brand.A

Price.Brand.B

Price.Brand.C

Feature

Display

Select factor (categorical) variables in X

week

Price.Brand.A

Price.Brand.B

Price.Brand.C

Feature

Display

Select sub sample

quick run, 1,000 obs

Impute missing values or drop missing value rows

do not impute or drop rows

Apply Changes

Overview Data Summary Data Visualization Summary OLS Residuals Plot Prediction Input Data Prediction New Data

Review Input Data

Show 10 entries

Search:

obs	Sales.Brand.A	week	Price.Brand.A	Price.Brand.B	Price.Brand.C	Feature	Display
1	6528	40	3.66	1.89	2.99	1	0
2	6016	43	3.66	1.89	2.99	0	0
3	6272	44	3.66	1.89	2.99	0	0
4	6848	45	3.66	1.89	2.99	0	0
5	7424	46	3.66	1.89	2.99	0	0
6	6848	47	3.66	1.89	2.99	0	chk
7	7488	48	3.66	2.39	2.99	0	0
8	6336	49	3.66	2.39	1.99	0	0
9	6208	50	3.66	1.99	2.99	0	0
10	6400	51	3.66	2.84	2.99	0	0

Any intermediate value doesn't make any sense

Showing 1 to 10 of 116 entries

Previous 1 2 3 4 5 ... 12 Next

Data Summary of Selected Y and X Variables

\$Dimensions  
[1] 116 7

\$Summary  
\$Summary\$Numeric.data

	Sales.Brand.A	week	Price.Brand.A	Price.Brand.B	Price.Brand.C	Feature
min	4608.00	40.0000	1.6900	1.4900	1.2900	0.0000
max	98624.00	160.0000	3.6600	2.8900	2.9900	1.0000
range	94016.00	120.0000	1.9700	1.4000	1.7000	1.0000
median	10800.00	102.5000	2.9900	2.3500	2.2600	1.0000
mean	18879.45	101.8448	2.9006	2.3018	2.2682	0.5259
var	404155133.47	1193.2453	0.3088	0.1627	0.1525	0.2515
std.dev	20103.61	34.5434	0.5557	0.4033	0.3905	0.5015

# Regression App

**Data Input**

Upload input data (csv file with header)

Browse... Orange Juice Sales.csv

Upload complete

**Data Selection**

Select Y variable

Sales.Brand.A

Select X variables

obs

week

Price.Brand.A

Price.Brand.B

Price.Brand.C

Feature

Display

Select factor (categorical) variables in X

week

Price.Brand.A

Price.Brand.B

Price.Brand.C

Feature

Display

Select sub sample

quick run, 1,000 obs

Imppute missing values or drop missing value rows

do not impute or drop rows

Apply Changes

Overview Data Summary Data Visualization Summary OLS Residuals Plot Prediction Input Data Prediction New Data

Review Input Data

Show 10 entries Search:

obs	Sales.Brand.A	week	Price.Brand.A	Price.Brand.B	Price.Brand.C	Feature	Display
1	6528	40	3.66	1.89	2.99	1	0
2	6016	43	3.66			0	0
3	6272	44	3.66			0	0
4	6848	45	3.66	2.39	2.99	0	0
5	7424	46	3.66	2.84	2.99	0	0
6	6848	47	3.66	2.84	2.39	0	chk
7	7468	48	3.66	2.39	2.99	0	0
8	6336	49	3.66	2.39	1.99	0	0
9	6208	50	3.66	1.99	2.99	0	0
10	6400	51	3.66	2.84	2.99	0	0

Showing 1 to 10 of 116 entries

Previous 1 2 3 4 5 ... 12 Next

**Proxy for Yes**

**Data Summary of Selected Y and X Variables**

```
$Dimensions  
[1] 116 7  
  
$Summary  
$Summary$Numeric.data  
Sales.Brand.A week Price.Brand.A Price.Brand.B Price.Brand.C Feature  
min 4688.00 40.0000 1.6900 1.4900 1.2900 0.0000  
max 98624.00 160.0000 3.6600 2.8900 2.9900 1.0000  
range 94016.00 128.0000 1.9700 1.4000 1.7000 1.0000  
median 10880.00 102.5000 2.9900 2.3500 2.2600 1.0000
```

# Regression App

**Data Input**

Upload input data (csv file with header)

Browse... Orange Juice Sales.csv

Upload complete

**Data Selection**

Select Y variable

Sales.Brand.A

Select X variables

obs

week

Price.Brand.A

Price.Brand.B

Price.Brand.C

Feature

Display

Select factor (categorical) variables in X

week

Price.Brand.A

Price.Brand.B

Price.Brand.C

Feature

Display

Select sub sample

quick run, 1,000 obs

Impute missing values or drop missing value rows

do not impute or drop rows

Apply Changes

Overview Data Summary Data Visualization Summary OLS Residuals Plot Prediction Input Data Prediction New Data

Review Input Data

Show 10 entries Search:

obs	Sales.Brand.A	week	Price.Brand.A	Price.Brand.B	Price.Brand.C	Feature	Display
obs	Sales.Brand.A	week	Price.Brand.A	Price.Brand.B	Price.Brand.C	Feature	Display
1	6528	40	3.66	1.89	2.99	1	0
2	6016	43	3.66	2.84	2.99	0	0
3	6272	44	3.66			0	0
4	6848	45	3.66			0	0
5	7424	46	3.66	2.84	2.99	0	0
6	6848	47	3.66	2.84	2.39	0	chk
7	7488	48	3.66	2.39	2.99	0	0
8	6336	49	3.66	2.39	1.99	0	0
9	6208	50	3.66	1.99	2.99	0	0
10	6400	51	3.66	2.84	2.99	0	0

Showing 1 to 10 of 116 entries

Previous 1 2 3 4 5 ... 12 Next

Proxy for No

**Data Summary of Selected Y and X Variables**

```
$Dimensions [1] 116 7
$Summary
$Summary$Numeric.data
Sales.Brand.A    week Price.Brand.A Price.Brand.B Price.Brand.C Feature
min      4688.00  40.0000   1.6900    1.4900    1.2900  0.0000
max     98624.00 160.0000   3.6600    2.8900    2.9900  1.0000
range    94816.00 120.0000   1.9700    1.4000    1.7000  1.0000
```

# Regression App

## Regression App

Data Input

Upload input data (csv file with header)

Browse... Orange Juice Sales.csv

Upload complete

Data Selection

Select Y variable

Sales.Brand.A

Select X variables

obs

week

Price.Brand.A

Price.Brand.B

Price.Brand.C

Feature

Display

Select factor (categorical) variables in X

week

Price.Brand.A

Price.Brand.B

Price.Brand.C

Feature

Display

Select sub sample

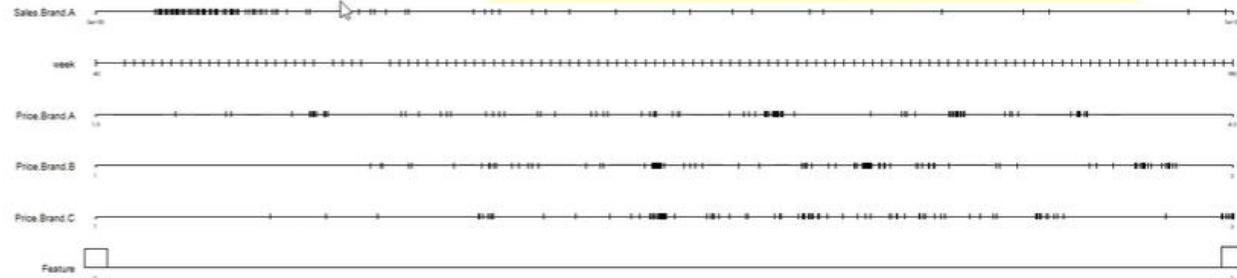
quick run, 1,000 obs

Impute missing values or drop missing value rows

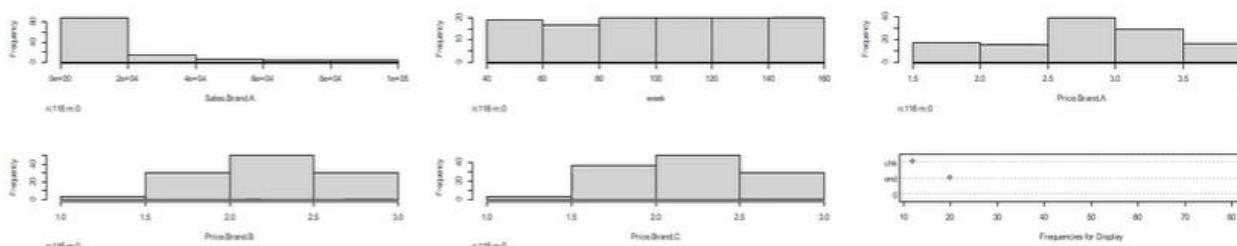
impute missing values

Overview Data Summary Data Visualization Summary

Be patient generating plots

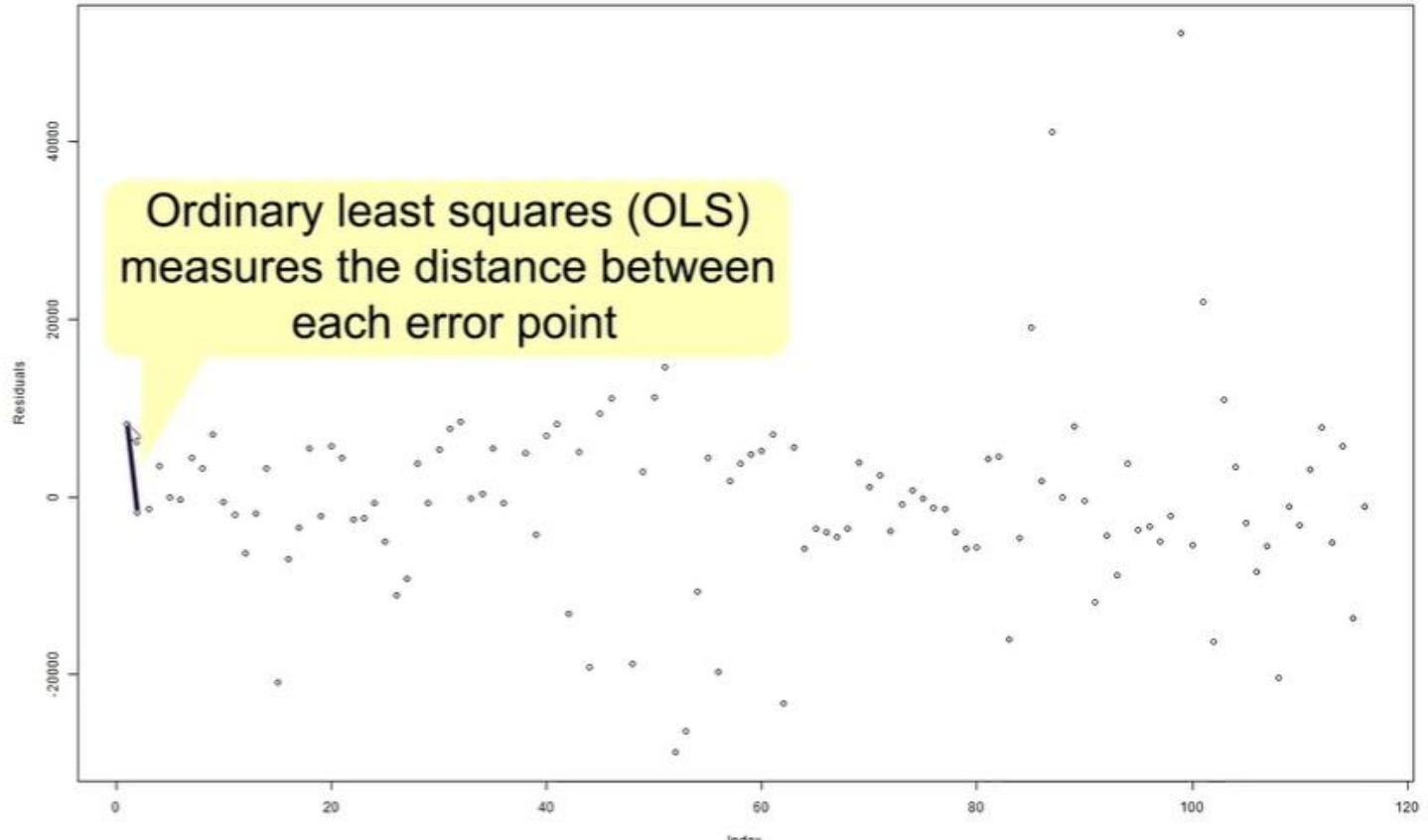


Histograms



The **Data Visualization** tab shows  
the basic statistical visualisation  
for the dataset

# Regression App



# Regression App

Data Input

Upload input data (csv file with header)

Browse... Orange Juice Sales.csv Upload complete

Data Selection

Select Y variable

Sales.Brand.A

Select X variables

obs  
 week  
 Price.Brand.A  
 Price.Brand.B  
 Price.Brand.C  
 Feature  
 Display

Select factor (categorical) variables in X

week  
 Price.Brand.A  
 Price.Brand.B  
 Price.Brand.C  
 Feature  
 Display

Select sub sample

quick run, 1,000 obs

Impute missing values or drop missing value rows

do not impute or drop rows

Apply Changes

Overview Data Summary Data Visualization Summary OLS Residuals Plot Prediction Input Data Prediction New Data

Mean Square Error of Input Data

```
$Number_of_Observations [1] 116
$Mean_Square_Error [1] 0.3371458
```

Predicted Values \ Y is Sales.Brand.A

Lower the mean squared error value, lower the error

# Regression Demo: Input Data

- When we select display as a factor (categorical) variable, software automatically does the dummy variable coding (one hot coding) of the categorical variable before running a regression
- Essentially, it creates one column for each level, where 1 represents the presence of a level and zero represents absence of a level for each row. e.g., 3 levels of display are represented by 3 columns

If observation 1 is coded as 0 in the dummy variable regression, the display column will be coded as 1 and other two variables as 0

Display	Display0	Displaychk	displayend
0	1	0	0
0	1	0	0
chk	0	1	0
0	1	0	0
end	0	0	1
end	0	0	1
0	0	0	0

# Regression App: Data Selection

- When we select display as a factor (categorical) variable, software automatically does the dummy variable coding (one hot coding) of the categorical variable before running a regression.
- Essentially, it creates one column for each level, where 1 represents the presence of a level and 0 represents absence of a level for each row. e.g., 3 levels of display are represented by 3 columns.

Multiple beta coefficients are associated with every categorical variable

	n	missing	distinct
116	0	3	
Value	0	chk	end
Frequency	84	12	20
Proportion	0.724	0.103	0.172

Display	Display0	Displaychk	displayend
0	1	0	0
0	1	0	0
chk	0	1	0
0	1	0	0
end	0	0	1
end	0	0	1
0	0	0	0

# Regression App: Data Selection

- When we select display as a factor (categorical) variable, software automatically does the dummy variable coding (one hot coding) of the categorical variable before running a regression
- Essentially, it creates one column for each level, where 1 represents the presence of a level and zero represents absence of a level for each row. e.g., 3 levels of display are represented by 3 columns

	n	missing	distinct
Value	116	0	3
Frequency	84	12	20
Proportion	0.724	0.103	0.172

Display	Display0	Displaychk	displayend
0	1	0	0
0	1	0	0
chk	0	1	0
0	1	0	0
end	0	0	1
end	0	0	1
0	0	0	0

- For technical reasons and identification purpose, once the categorical variable is converted into the dummy variable, the first column is dropped

# Regression App: Data Selection

- When we select display as a factor (categorical) variable, software automatically does the dummy variable coding (one hot coding) of the categorical variable before running a regression
- Essentially, it creates one column for each level, where 1 represents the presence of a level and zero represents absence of a level for each row. e.g., 3 levels of display are represented by 3 columns

	n	missing	distinct
116	0	3	
Value	0	chk	end
Frequency	84	12	20
Proportion	0.724	0.103	0.172

Display	Display0	Displaychk	displayend
0	1	0	0
0	1	0	0
chk	0	1	0
0	1	0	0
end	0	0	1
end	0	0	1
0	0	0	0

Data coefficients

- For technical reasons and identification purpose, once the categorical variable is converted into the dummy variable, the first column is dropped

# Regression: Summary OLS Tab

## Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	75656.19	12963.43	5.836	5.67e-08	***
week	-95.31	37.11	-2.569	0.01157	*
Price.Brand.A	-24561.45	2865.94	-8.570	8.08e-14	***
Price.Brand.B	9238.26	2807.66	3.290	0.00135	**
Price.Brand.C	-52.17	3015.06	-0.017	0.98623	
Feature	-904.65	2694.14	-0.336	0.73768	
Displaychk	-234.20	3629.76	-0.065	0.94867	
Displayend	20450.15	3176.36	6.438	3.43e-09	***
---					

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

# Understanding Price Change

## Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	75656.19	12963.43	5.836	5.67e-08	***
week	-95.31	37.11	-2.569	0.01157	*
Price.Brand.A	-24561.45	2865.94	-8.570	8.08e-14	***
Price.Brand.'3	9238.26	2807.66	3.290	0.00135	**
How does the price of Brand A change its sales?	-52.17	3015.06	-0.017	0.98623	
Displaycirk	-904.65	2694.14	-0.336	0.73768	
Displayend	-234.20	3629.76	-0.065	0.94867	
---					
Signif. codes:	0 ‘***’	0.001 ‘**’	0.01 ‘*’	0.05 ‘.’	0.1 ‘ ’ 1

# Understanding Price Change

## Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	75656.19	12963.43	5.836	5.67e-08	***
week	-95.31	37.11	-2.569	0.01157	*
Price.Brand.A	<b>-24561.45</b>	2865.94	-8.570	8.08e-14	***
Price.Brand.B	9138.26	2807.66	3.290	0.00135	**
Price.Brand.C	52.17	3015.06	-0.017	0.98623	
		2694.14	-0.336	0.73768	
		3629.76	-0.065	0.94867	
		3176.36	6.438	3.43e-09	***
---					

If the price of Brand A is changed by one dollar, the impact on the outcome variable, which is the sales of Brand A, is 24561 units

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

# Understanding Price Change

### Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	75656.19	12963.43	5.836	5.67e-08	***
week	-95.31	37.11	-2.569	0.01157	*
Price.Brand.A	24561.45	2865.94	-8.570	8.08e-14	***
Price.Brand.B	9238.26	2807.66	3.290	0.00135	**
minus sign indicates that the price of the brand ases, its sales decreases	-2.17	3015.06	-0.017	0.98623	
displays	4.65	2694.14	-0.336	0.73768	
	4.20	3629.76	-0.065	0.94867	
	2840.15	3176.36	6.438	3.43e-09	***

The minus sign indicates that if the price of the brand increases, its sales decreases

- 3 -

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

# Understanding Price Change

## Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	75656.19	12963.43	5.836	5.67e-08	***
week	-95.31	37.11	-2.569	0.01157	*
Price.Brand.A	<b>-24561.45</b>	2865.94	-8.570	8.08e-14	***
Price.Brand.B	9238.26	2807.66	3.290	0.00135	**
Price.Brand.C	-52.17	3015.06	-0.017	0.98623	
Fee	What is the confidence level for this value?	55	2694.14	-0.336	0.73768
Display....	.....20	3629.76	-0.065	0.94867	
Displayend	20450.15	3176.36	6.438	3.43e-09	***
---					
Signif. codes:	0 ****	0.001 ***	0.01 **	0.05 *	0.1 . 1

# Understanding Price Change

## Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	75656.19	12963.43	5.836	5.67e-08	***
week	-95.31	37.11	-2.569	0.01157	*
Price.Brand.A	-24561.45	2865.94	-8.570	8.08e-14	***
Price.Brand.B	9238.26	2807.66	3.290	0.00135	**
Price.Brand.C	-52.17	3015.06	-0.017	0.98623	
Feature	-904.65	2694.14	-0.336	0.73768	
Displaychk	-234.20	3629.76	-0.065	0.94867	
Displayend	20450.15	3176.36	6.438	3.43e-09	***
---					

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Very confident

The probability of this number being 0 is 0.001 OR 99.9% sure

# Understanding Price Change

## Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	75656.19	12963.43	5.836	5.67e-08	***
week	-95.31	37.11	-2.569	0.01157	*
Price.Brand.A	-24561.45	2865.94	-8.570	8.08e-14	***
Price.Brand.B	9238.26	2807.66	3.290	0.00135	**
Price.Brand.C	-52.17	3015.06	-0.017	0.98623	
Feature	-904.65	2694.14	-0.336	0.73768	
Displaychk	-234.20	3629.76	-0.065	0.94867	
Displayend	20450.15	3176.36	6.438	3.43e-09	***
---					
Signif. codes:	0 ****	0.001 ***	0.01 **	0.05 *	0.1 ' '

99% sure

# Understanding Price Change

## Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	75656.19	12963.43	5.836	5.67e-08	***
week	-95.31	37.11	-2.569	0.01157	*
Price.Brand.A	-24561.45	2865.94	-8.570	8.08e-14	***
Price.Brand.B	9238.26	2807.66	3.290	0.00135	**
Price.Brand.C	-52.17	3015.06	-0.017	0.98623	
Feature	-904.65	2694.14	-0.336	0.73768	
Displaychk	-234.20	3629.76	-0.065	0.94867	
Displayend	20450.15	3176.36	6.438	3.43e-09	***
---					
Signif. codes:	0 ****	0.001 ***	0.01 **	0.05 *	. 0.1 ' '

95% sure

# Understanding Price Change

## Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	75656.19	12963.43	5.836	5.67e-08	***
week	-95.31	37.11	-2.569	0.01157	*
Price.Brand.A	-24561.45	2865.94	-8.570	8.08e-14	***
Price.Brand.B	9238.26	2807.66	3.290	0.00135	**
Price.Brand.C	-52.17	3015.06	-0.017	0.98623	
Feature	-904.65	2694.14	-0.336	0.73768	
Displaychk	-234.20	3629.76	-0.065	0.94867	
Displayend	20450.15	3176.36	6.438	3.43e-09	***
---					
Signif. codes:	0 ****	0.001 ***	0.01 **	0.05 *	. 0.1 ' '

90% sure

# How Does Display Affect Sales?

## Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	75656.19	12963.43	5.836	5.67e-08	***
week	-95.31	37.11	-2.569	0.01157	*
Price.Brand.A	-24561.45	2865.94	-8.570	8.08e-14	***
Price.Brand.B	9238.26	2807.66	3.290	0.00135	**
Price.Brand.C	-52.17	3015.06	-0.017	0.98623	
Feature	-904.65	2694.14	-0.336	0.73768	
Displaychk	-234.20	3629.76	-0.065	0.94867	
Displayend	20450.15	3176.36	6.438	3.43e-09	***

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

# Coefficient of Factor Variable: Interpretation

When there are no stars against a coefficient, then it is no different from 0

Coefficients:						
	Estimate	Std. Error	t value	Pr(> t )		
(Intercept)	75656.19	12963.43	5.836	5.67e-08	***	
week	-95.31	37.11	-2.569	0.01157	*	
Price.Brand.A	-24561.45	2865.94	-8.570	8.08e-14	***	
Price.Brand.B	9238.26	2807.66	3.290	0.00135	**	
Price.Brand.C	-52.17	3015.06	-0.017	0.98623		
Feature	-904.65	2694.14	-0.336	0.73768		
Displaychk	-234.20	3629.76	-0.065	0.94867		
Displayend	20450.15	3176.36	6.438	3.43e-09	***	
---						
Signif. codes: 0 '****' 0.001 '***' 0.01 '**' 0.05 '*' 0.1 '.' 1						

# Coefficient of Factor Variable: Interpretation

## Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	75656.19	12963.43	5.836	5.67e-08	***
week	-95.31	37.11	-2.569	0.01157	*
Price.Brand.A	-24561.45	2865.94	-8.570	8.08e-14	***
Price.Brand.B	9238.26	2807.66	3.290	0.00135	**
Price.Brand.C	-52.17	3015.06	-0.017	0.98623	
Feature	-904.65	2694.14	-0.336	0.73768	
Displaychk	-234.20	3629.76	-0.065	0.94867	
Displayend	20450.15	3176.36	6.438	3.43e-09	***
---					99.9% confident
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1					

# Coefficient of Factor Variable: Interpretation

Base is the column that is dropped.

## Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	75656.19	12963.43	5.836	5.67e-08	***
week	-95.31	37.11	-2.569	0.01157	*
These effects are relative to the base column, which indicated that the brand was not on display	.45 .26 .17	2865.94 2807.66 3015.06	-8.570 3.290 -0.017	8.08e-14 0.00135 0.98623	*** **
feature	-504.65	2694.14	-0.336	0.73768	
Displaychk	-234.20	3629.76	-0.065	0.94867	
Displayend	20450.15	3176.36	6.438	3.43e-09	***
---					
Signif. codes:	0 ****	0.001 ***	0.01 **	0.05 *	0.1 . 1

# Coefficient of Factor Variable: Interpretation

Displaying a brand next to the aisle has a significant effect, increasing the sales by 20450 units.

## Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	75656.19	12963.43	5.836	5.67e-08	***
week	-95.31	37.11	-2.569	0.01157	*
Price.Brand.A	-24561.45	2865.94	-8.570	8.08e-14	***
Price.Brand.B	9238.26	2807.66	3.290	0.00135	**
Price.Brand.C	-52.17	3015.06	-0.017	0.98623	
Feature	-904.65	2694.14	-0.336	0.73768	
Displaychk	-234.20	3629.76	-0.065	0.94867	
Displayend	20450.15	3176.36	6.438	3.43e-09	***
---					
Signif. codes:	0 ****	0.001 ***	0.01 **	0.05 *	0.1 ' '
	1				

# Regression App: Interpretation

The columns that are relevant for interpretation are:

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	75656.19	12963.43	5.836	5.67e-08	***
week	-95.31	37.11	-2.569	0.01157	*
Price.Brand.A	-24561.45	2865.94	-8.570	8.08e-14	***
Price.Brand.B	9238.26	2807.66	3.290	0.00135	**
Price.Brand.C	-52.17	3015.06	-0.017	0.98623	
Feature	-904.65	2694.14	-0.336	0.73768	
Displaychk	-234.20	3629.76	-0.065	0.94867	
Displayend	20450.15	3176.36	6.438	3.43e-09	***
---					

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

# Regression App: Interpretation

The stars are generated by the t values.

## Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	75656.19	12963.43	5.836	5.67e-08	***
week	-95.31	37.11	-2.569	0.01157	*
Price.Brand.A	-24561.45	2865.94	-8.570	8.08e-14	***
Price.Brand.B	9238.26	2807.66	3.290	0.00135	**
Price.Brand.C	-52.17	3015.06	-0.017	0.98623	
Feature	-904.65	2694.14	-0.336	0.73768	
Displaychk	-234.20	3629.76	-0.065	0.94867	
Displayend	20450.15	3176.36	6.438	3.43e-09	***
---					
Signif. codes:	0 ****	0.001 ***	0.01 **	0.05 *	0.1 . 1

If the absolute number of the t-value is greater than two, it will have at least one star

# Regression for Predictions

- Helps understand the relationships between variables -> Descriptive analytics
- Also helps make predictions -> Predictive analytics

# Regression: Evaluating the Model

[Overview](#)[Summary Stats](#)[Summary OLS](#)[Input Data with Predictions](#)

Download input data with predicted Y

 Download data (works only in browser)

The 'Input Data with Prediction' tab shows the actual sales value (Y) and the predicted sales value

Y.hat	Sales.Brand A	week	Price.Brand.A	Price.Brand.B	Price
-1651.53	6528	40	3.66	1.89	
7743.53	6016	43	3.66	2.84	
7669.08	6272	44	3.66	2.84	
3395.69	6848	45	3.66	2.39	
7457.59	7424	46	3.66	2.84	
7169.38	6848	47	3.66	2.84	
3109.76	7488	48	3.66	2.39	

# Regression Demo: Evaluating the Model

[Overview](#)[Summary Stats](#)[Summary OLS](#)[Input Data with Predictions](#)

Download input data with predicted Y

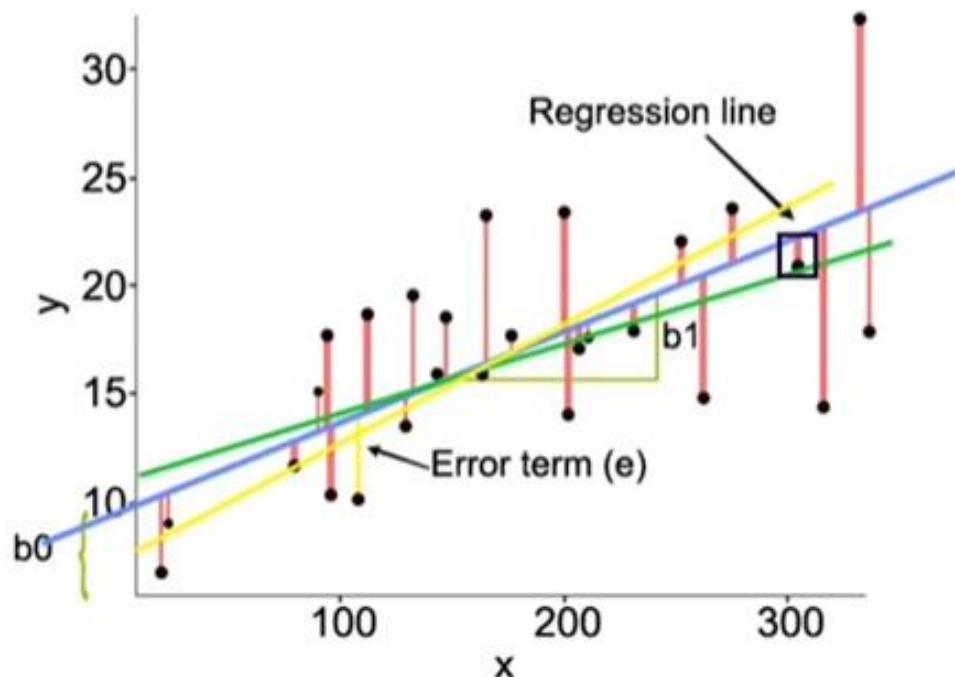
 Download data (works only in browser)

Predicts all X variables

Y.hat	Sales.Brand A	week	Price.Brand.A	Price.Brand.B	Price
-1651.53	6528	40	3.66	1.89	
7743.53	6016	43	3.66	2.84	
7669.08	6272	44	3.66	2.84	
3395.69	6848	45	3.66	2.39	
7457.59	7424	46	3.66	2.84	
7169.38	6848	47	3.66	2.84	
3109.76	7488	48	3.66	2.39	

# Regression Analysis: Model Estimation

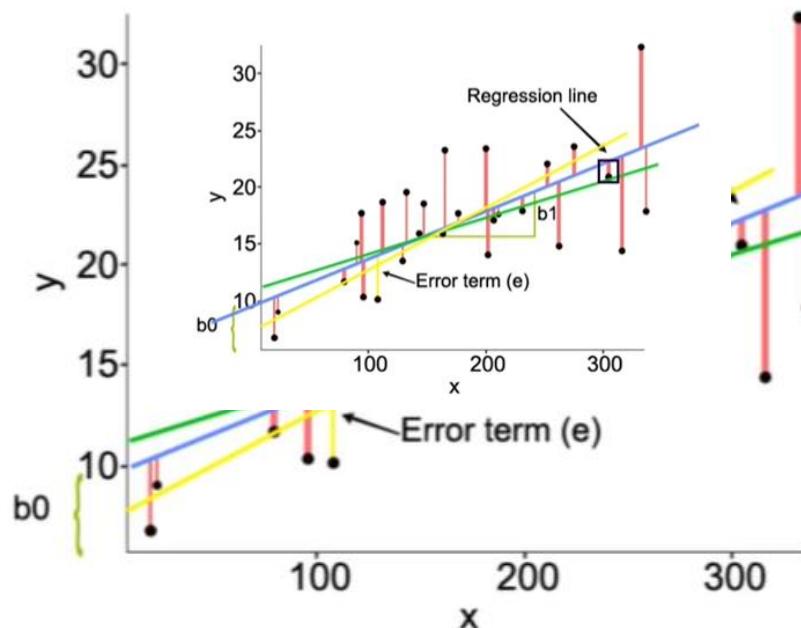
Regression analysis identifies the pattern in the different data points and fits the best line.



The model measures the distance between each point and the distance from the line, squares the values and sums the value

# Regression Analysis: Model Estimation

Regression analysis identifies the pattern in the different data points and fits the best line.



Error is the deviation between the actual data point and the predicted data point

Regression minimises the error between predicted\_Y and actual\_Y

# Regression Analysis: Evaluating the Model

## How good is the model?

Coefficients:

	Estimate	Std.Error	t value	Pr(>  t  )
(Intercept)	75656.19	12963.43	5.836	5.67e-08
week	-95.31	37.11	-2.569	0.01157
Price .Brand . A	-24561.45	2865.94	-8.570	8.08e-14
Price .Brand . B	9238.26	2807.66	3.290	0.00135
Price .Brand . C	-52.17	3015.06	-0.017	0.98623
Feature	-904.65	2694.14	-0.336	0.73768
Displaychk	-234.20	3629.76	-0.065	0.94867
Displayend	20450.15	3176.36	6.438	3.43e-09
---				
Signif. codes :	0	0.001	0.01	0.05
			0.1	1

Residual standard error: 11570 on 108 degrees of freedom

Multiple R-squared: 0.6888 ,

Adjusted R-squared: 0.6687,

F-statistic: 34.15 on 7 and 108 DF , p-value: < 2.2e-16

R-squared is a measure of how good a regression model is.

# Multiple R-squared vs Adjusted R-squared

- Multiple R-squared: Raw number
- Adjusted R-squared: Adjusts for the number of parameters in the model

# Regression Analysis: Evaluating the Model

## How good is the model?

Coefficients:

	Estimate	Std.Error	t value	Pr(>  t  )
(Intercept)	75656.19	12963.43	5.836	5.67e-08
week	-95.31	37.11	-2.569	0.01157
Price .Brand . A	-24561.45	2865.94	-8.570	8.08e-14
Price .Brand . B	9238.26	2807.66	3.290	0.00135
Price .Brand . C	-52.17	3015.06	-0.017	0.98623
Feature	-904.65	2694.14	-0.336	0.73768
Displaychk	-234.20	3629.76	-0.065	0.94867
Displayend	20450.15	3176.36	6.438	3.43e-09
---				

Signif. codes : 0 0.001 0.01 0.05

Indicates the probability that the model is not good

Residual standard error: 11570 on 108 degrees of freedom

Multiple R-squared: 0.6888 , Adjusted R-squared: 0.6687,

F-statistic: 34.15 on 7 and 108 DF , p-value: < 2.2e-16

The probability that this model is not good can be considered 0 since the value is small

# Regression Analysis: Evaluating the Model

## How good is the model?

### Coefficients:

	Estimate	Std.Error	t value	Pr(>  t  )
(Intercept)	75656.19	12963.43	5.836	5.67e-08
week	-95.31	37.11	-2.569	0.01157
Price .Brand . A	-24561.45	2865.94	-8.570	8.08e-14
Price .Brand . B	9238.26	2807.66	3.290	0.00135
Price .Brand . C	-52.17	3015.06	-0.017	0.98623
	-904.65	2694.14	-0.336	0.73768
	-234.20	3629.76	-0.065	0.94867
	20450.15	3176.36	6.438	3.43e-09

R-squared can fall between 0 and 1. If it is 0, then the model is not good, and if it is closer to 1, the model is good

Residual standard error: 11570 on 108 degrees of freedom

Multiple R-squared: 0.6888 , Adjusted R-squared: 0.6687,

F-statistic: 34.15 on 7 and 108 DF , p-value: < 2.2e-16

# Multiple R-Squared of 0.688: Interpretation

From the analysis, the sales of the brand is well explained by the five variables selected:

- Price.Brand.A
- Price.Brand.B
- Price.Brand.C
- Feature
- Displaychk / Displayend

# Using the Regression App for Prediction

The screenshot shows a user interface for a regression analysis application. On the left, under 'Data Input', a file 'Orange Juice Sales.csv' has been uploaded. Under 'Data Selection', 'Sales.Brand.A' is selected as the Y variable, and several X variables are checked: week, Price.Brand.A, Price.Brand.B, Price.Brand.C, Feature, Display, In.salesA., In.priceA., In.priceB., and In.priceC. In the main panel, the 'Prediction New Data' tab is active, showing a placeholder for new data upload and a download button for predictions. A yellow callout points to the 'Upload data' button with the text 'Upload data'. Another yellow callout points to the 'Download predictions' button with the text 'Download predictions'. A third yellow callout points to the text 'The structure of the data should be similar to the structure of the data used for analysis'.

**Data Input**

Upload input data (csv file with header)

Browse... Orange Juice Sales.csv  
Upload complete

**Data Selection**

Select Y variable

Sales.Brand.A

Select X variables

week  
 Price.Brand.A  
 Price.Brand.B  
 Price.Brand.C  
 Feature  
 Display  
 In.salesA.  
 In.priceA.  
 In.priceB.  
 In.priceC.

Overview Data Summary Summary OLS Standardized OLS Input Data with Predictions

Residuals Plot Prediction New Data

Upload new data for prediction, it s with header)

Browse... No file selected

Download new data with predictions

download predictions for new data

"Yhat" col 'mn is the predicted Y values.

Upload data

format as input data file (csv file

Download predictions

The structure of the data should be similar to the structure of the data used for analysis

# Prediction with the Regression App

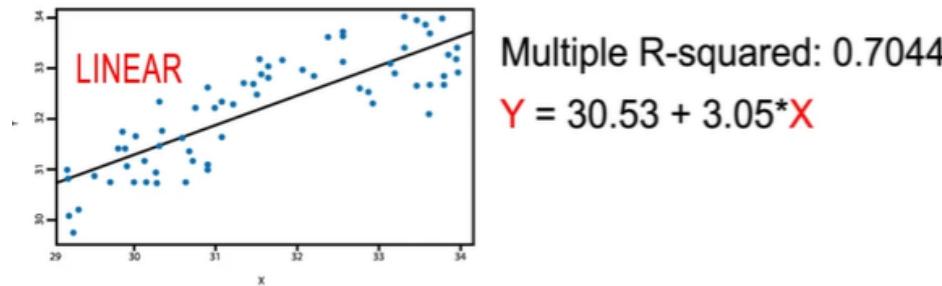
Data for prediction will:

- Not have the Y variable
- Have all the X variables

With this data, the regression model will:

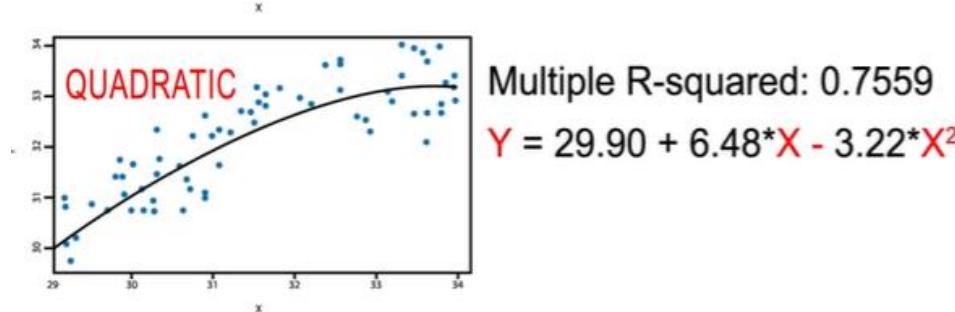
- Compute all the X variables
- Multiply the beta coefficients (known already) with the right X variable
- Predict the Y variable

# Improving the Model: Adjusted R-squared



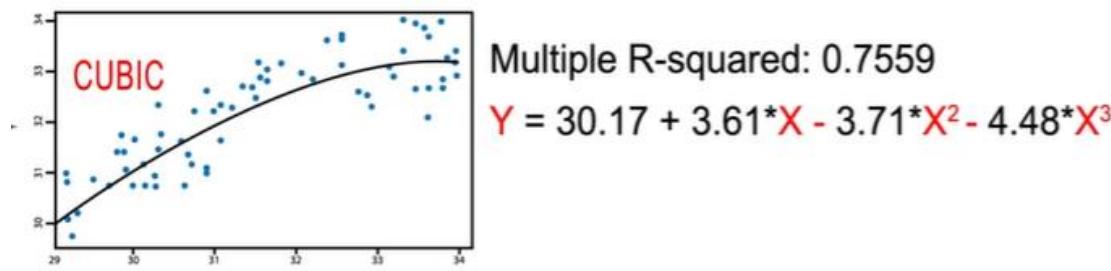
A linear regression model does not mean that the X variables have to be linear;  $X^2$  can be used

# Improving the Model: Adjusted R-squared



The quadrature or curvature  
could be captured by using the  $X^2$  term

# Improving the Model: Adjusted R-squared

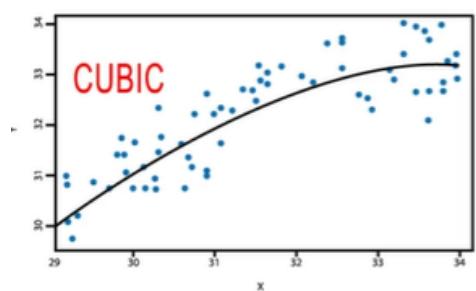
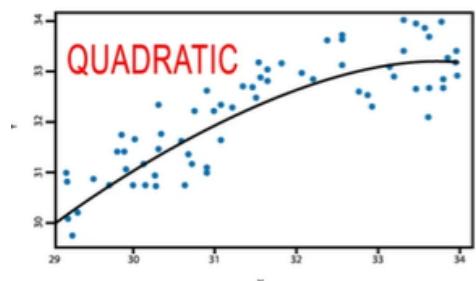
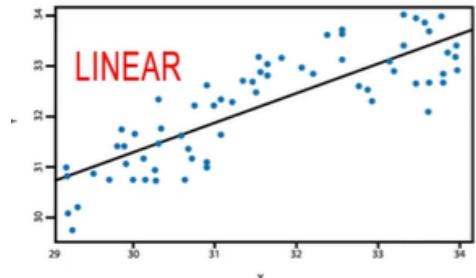


If the second order term does not capture the non-linearity well, then the cubic term could be used

# Improving the Model: Adjusted R-squared

Using higher order terms will fit the model better and better; however, the goal is not for the model to closely fit the data but to infer the relationship

# Improving the Model: Adjusted R-squared



- **Adjusted R-squared:**
  - Captures the effect of the number of variables in the model
  - Decreases with increase in the number of variables

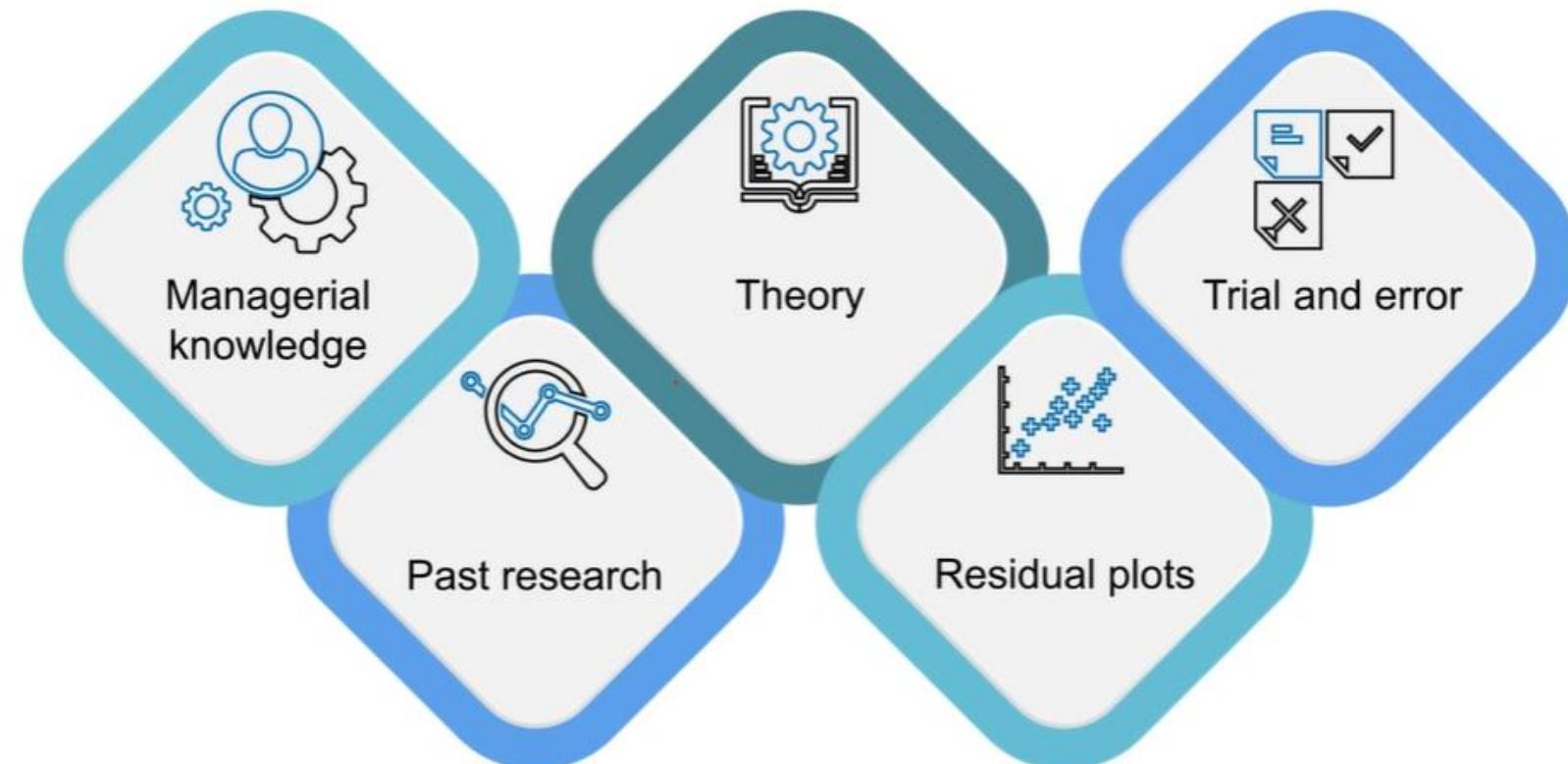
# Capturing the non-linear effects in the data:

To capture the non-linear effects in the data:

- Use higher-order terms up to second order ( $X$  and  $X^2$ )
- Transform  $Y$  and  $X$  variables using the  $\log(X)$  and  $\log(Y)$ ,  
i.e. the log-log model

# Model Improvement through Feature Engineering

**To decide when to use the quadratic term or the log-log model, use:**



# Conceptual Model and Linear Regression Equation

## Conceptual model

$\text{Sales} = f(\text{Own Price}, \text{Compt\_Price}, \text{Feature}, \text{Display}, \text{etc.})$

## Model equation for linear regression

$\text{Sales} = \beta_0 + \beta_1 * \text{Price} + \beta_2 * \text{Compt\_Price} + \beta_3 * \text{Display} + \beta_4 * \text{Feature} + \varepsilon$

# The Log-Log Model to Estimate Elasticity

**This model captures the relationship between sales and price better:**

$$\text{Log}(Sales) = \beta_0 + \beta_1 * \text{Log}(Price) + \beta_2 * \text{log}(Comp.Price) + \beta_3 * \text{feature...})$$

$$\boxed{\text{Log}(Sales)} = \beta_0 + \beta_1 * \boxed{\text{Log}(Price)} + \beta_2 * \text{log}(Comp.Price) + \beta_3 * \text{feature...})$$

Interpret those beta coefficients onto  
the regression directly as elasticity

# What is Elasticity?

A measure of how the percentage change in one variable (e.g., price) affects the percentage in another variable (e.g. sales)

# Elasticity

To run this model and find the elasticity:

- Run the regression with  $\log(Y)$  and  $\log(X)$
- Interpret the beta coefficients, which directly give the elasticity

# The Log-Log Model to Estimate Elasticity

$$\text{Log}(Sales) = \beta_0 + \boxed{\beta_1} * \text{Log}(Price) + \boxed{\beta_2} * \text{log(Comp.Price)} + \boxed{\beta_3 * \text{feature...}}$$

The elasticity of prices with  
respective to the sales

Cross  
elasticity

Feature and display are  
categorical variables. Log  
cannot be applied to  
categorical variables as  
they are nonmetric

# Running Log-Log Regression

## To run a Log-Log regression:

- Open CSV file, which has the original data and transform or create two new columns
  - A column for  $\log(\text{PriceA})$ ,  $\log(\text{PriceB})$ , etc.
  - A column for  $\log(\text{Sales})$
- Upload the new data into the app
- Select Y variable as  $\log(\text{Sales})$
- Select the X variables as  $\log(\text{Prices})$ ,  $\log(\text{Feature})$  and  $\log(\text{Display})$
- Interpret the beta coefficient

# Interpretation

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	11.56146	0.27360	42.257	< 2e-16	***
Ln.Price_A.	-2.51144	0.20371	-12.328	< 2e-16	***
Ln.Price_B.	0.55301	0.18213	3.036	0.00299	**
Ln.Price_C.	-0.04501	0.19674	-0.229	0.81945	
Feature	0.06550	0.07831	0.836	0.40473	
Display	0.63220	0.09439	6.698	9.34e-10	***
---					
Signif. codes:	0 ****	0.001 ***	0.01 **	0.05 *	0.1 ' '

Residual standard error: 0.349 on 110 degrees of freedom

Multiple R-squared: 0.7947, Adjusted R-squared: 0.7853

F-statistic: 85.15 on 5 and 110 DF, p-value: < 2.2e-16

# Interpretation

This is the percentage change in the prices and its effect on the percentage change in the sales

## Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	11.56146	0.27360	42.257	< 2e-16 ***
'~ Price_A.	-2.51144	0.20371	-12.328	< 2e-16 ***
Ln.Price_B.	0.55301	0.18213	3.036	0.00299 **
Ln.Price_C.	-0.04501	0.19674	-0.229	0.81945
Feature	0.06550	0.07831	0.836	0.40473
Display	0.63220	0.09439	6.698	9.34e-10 ***
---				

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.349 on 110 degrees of freedom

Multiple R-squared: 0.7947, Adjusted R-squared: 0.7853

F-statistic: 85.15 on 5 and 110 DF, p-value: < 2.2e-16

# Interpretation

On an average,  
1% change  
in the price of  
brand A,  
decreases  
the sales by  
2.51%

## Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	11.56146	0.27360	42.257	< 2e-16 ***
Ln.Price_A.	-2.51144	0.20371	-12.328	< 2e-16 ***
Ln.Price_B.	0.55301	0.18213	3.036	0.00299 **
Ln.Price_C.	-0.04501	0.19674	-0.229	0.81945
Feature	0.06550	0.07831	0.836	0.40473
Display	0.63220	0.09439	6.698	9.34e-10 ***
---				
Signif. codes:	0 ****	0.001 ***	0.01 **	0.05 *
	.	0.1	'	1

Residual standard error: 0.349 on 110 degrees of freedom

Multiple R-squared: 0.7947, Adjusted R-squared: 0.7853

F-statistic: 85.15 on 5 and 110 DF, p-value: < 2.2e-16

# Interpretation

## Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	11.56146	0.27360	42.257	< 2e-16 ***
Ln.Price_A.	-2.51144	0.20371	-12.328	< 2e-16 ***
Ln.Price_B.	0.55301	0.18213	3.036	0.00299 **
Ln.Price_C.	-0.04501	0.19674	-0.229	0.81945
Feature	0.06550	0.07831	0.836	0.40473
Display	0.63220	0.09439	6.698	9.34e-10 ***
---				
Signif. codes:	0 ****	0.001 ***	0.01 **	0.05 *
	.	.	.	1

99.9%  
confident

Residual standard error: 0.349 on 110 degrees of freedom

Multiple R-squared: 0.7947, Adjusted R-squared: 0.7853

F-statistic: 85.15 on 5 and 110 DF, p-value: < 2.2e-16

# Interpretation

## Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	11.56146	0.27360	42.257	< 2e-16	***
Ln.Price_A.	-2.51144	0.20371	-12.328	< 2e-16	***
Ln.Price_B.	0.55301	0.18213	3.036	0.001299	**
Ln.Price_C.	-0.04501	0.19674	-0.229	0.8145	
Feature	0.06550	0.07831	0.830	0.4106	
Display	0.63220	0.09439	6.630	1.11e-10	***
---					
Signif. codes:	0 ****	0.001 ***	(		

Pr(>|t|) value is very small, i.e.  
only 2e-16 probability that the  
coefficient of Ln.Price\_A can  
have a zero effect

Residual standard error: 0.349 on 110 degrees of freedom

Multiple R-squared: 0.7947, Adjusted R-squared: 0.7853

F-statistic: 85.15 on 5 and 110 DF, p-value: < 2.2e-16

# Regression Variants

In the retailing context, managers often think that multiple promotions are more effective than individual promotions

## Hypothesis

If a brand were featured as well as displayed, the total effect is higher than the individual effect

We can test this hypothesis by including an interaction term in regression

# Regression with Interaction Effects

## Conceptual Model

Sales =  $f(\text{Own Price}, \text{Compt Price}, \text{Feature}, \text{Display}, \text{Feature} \& \text{Display})$

Creates a synergic effect  
and together, the sum of  
these two is greater than the  
individual components

# Regression with Interaction Effects

## Conceptual Model

$\text{Sales} = f(\text{Own Price}, \text{Compt Price}, \text{Feature}, \text{Display}, \text{Feature}\&\text{Display})$

## Modelling Equation

$$\text{Log(sales)} = \beta_0 + \beta_1 * \text{log(price)} + \beta_2 * \text{feat} + \beta_3 * \text{dsply} + \beta_4 * \underline{\text{feat}* \text{dsply}} + \epsilon$$

**By interpreting the coefficient with the new variable, you will know whether:**

- The interaction effects are positive or negative
- There is a synergetic effect or they are negating each other

# Regression: Interaction Effects

To add interaction effects in regression model:

- Open the CSV file
- Create a new column
- Multiply display and feature
- Save the dataset
- Go back to the app and refresh it
- Upload the new data
- Choose the new column with display x feature
- Run the analysis

# Recap: Linear Regression

Regression is a workhorse model, a supervised learning method that identifies linear relationships(i.e., linear  $\beta$  in coefficients)

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2^2 + \dots + \varepsilon$$

Where,

Y: Dependent variable

X: Independent variables

$\beta$ : Coefficients

$\varepsilon$ : Error term

# Regression Analysis: Areas of Focus

## Model fit

- Look at the  $R^2$  and particularly, the adjusted  $R^2$

## Beta coefficients

- Understand whether the variable matters by looking at the statistical significance of the beta coefficients and its values

## Non-linearity

- Consider whether to include quadratic terms or transform the variables

## Interaction effects

- Explore the synergistic effect between two variables by creating a new variable, which is the multiplier of the two variables

## Categorical variables

- Pick the categorical variables carefully as there's a fundamental difference between how they are handled in regression versus in the continuous rating

# Ordinary Least Squares (OLS) Regression

A workhorse and an analytical tool for the real-world business problems and widely used in business and management and social sciences

# OLS Regression Assumption

OLS regression assumes that:

- The independent variables are uncorrelated

This is not a concern for making predictions but for interpreting relationships, in which case, you may check:

- Multicollinearity
- Autocorrelation
- Homoscedasticity

# Regression Is a Business Analytic Workhorse

- We can frame most business and social science questions as regression problems
- We are interested in the relationship between independent and dependent variables, example
  - How promotions and price affect sales
  - How R&D spending affects patents output
  - How gender composition of a company board affects the financial performance
  - How the cultural orientation affects brand extensions
  - How satisfaction of different aspects affects overall satisfaction and churn



**ISB**

**Executive  
Education**