Applied Business Analytics

Week 3: Regression – Descriptive and Predictive Analytics

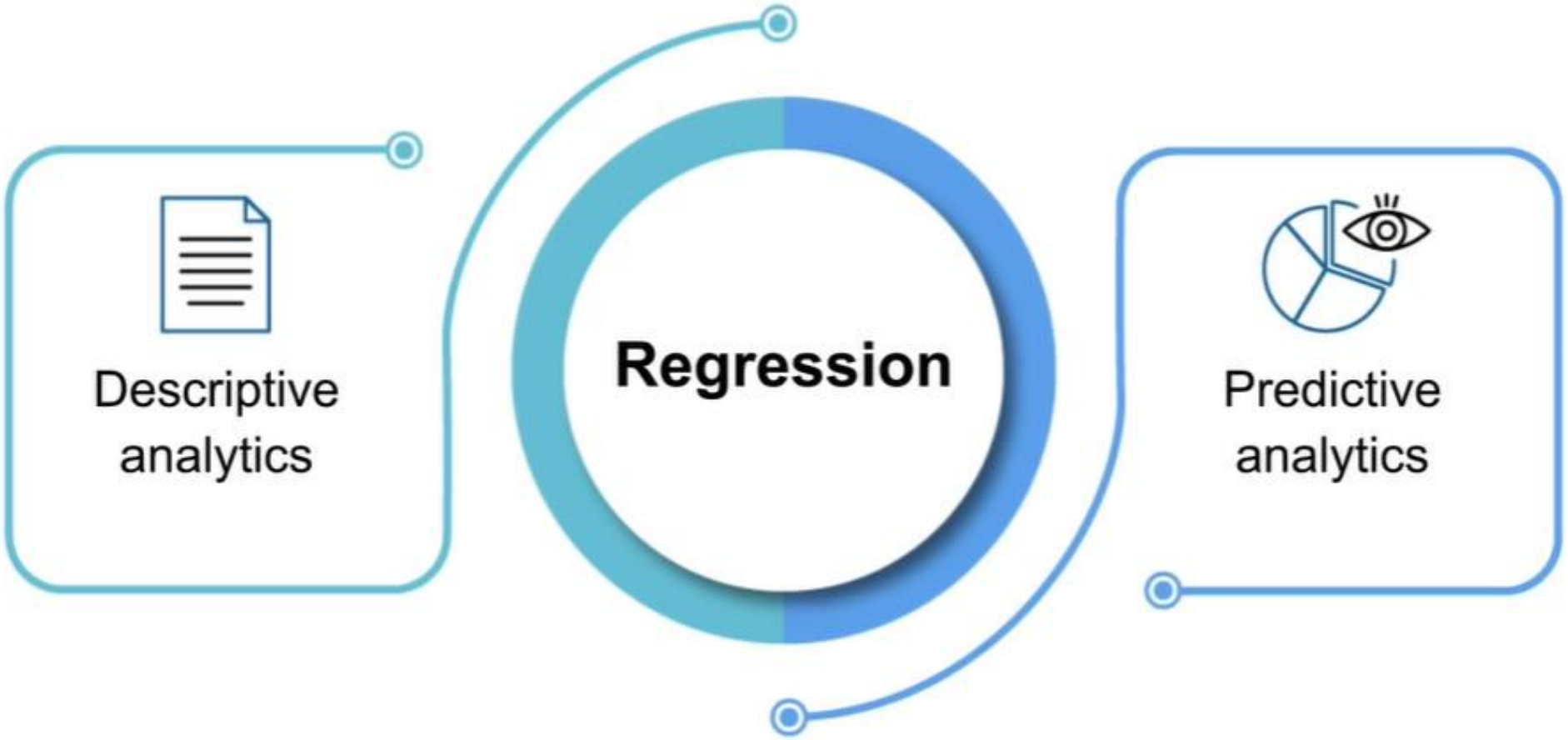
ISB | Executive Education

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



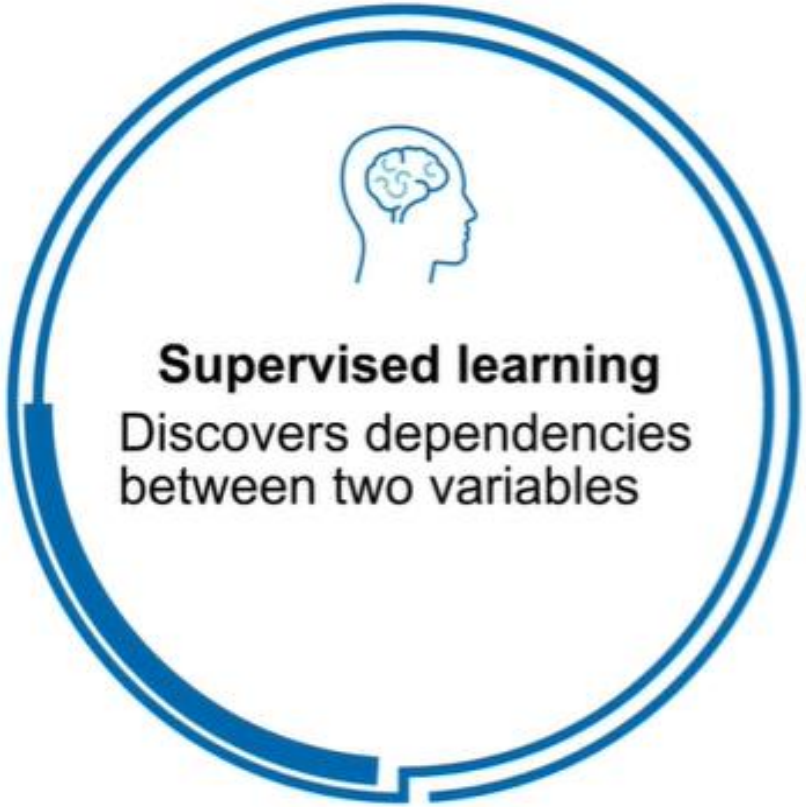
ML Approaches in Business Analytics

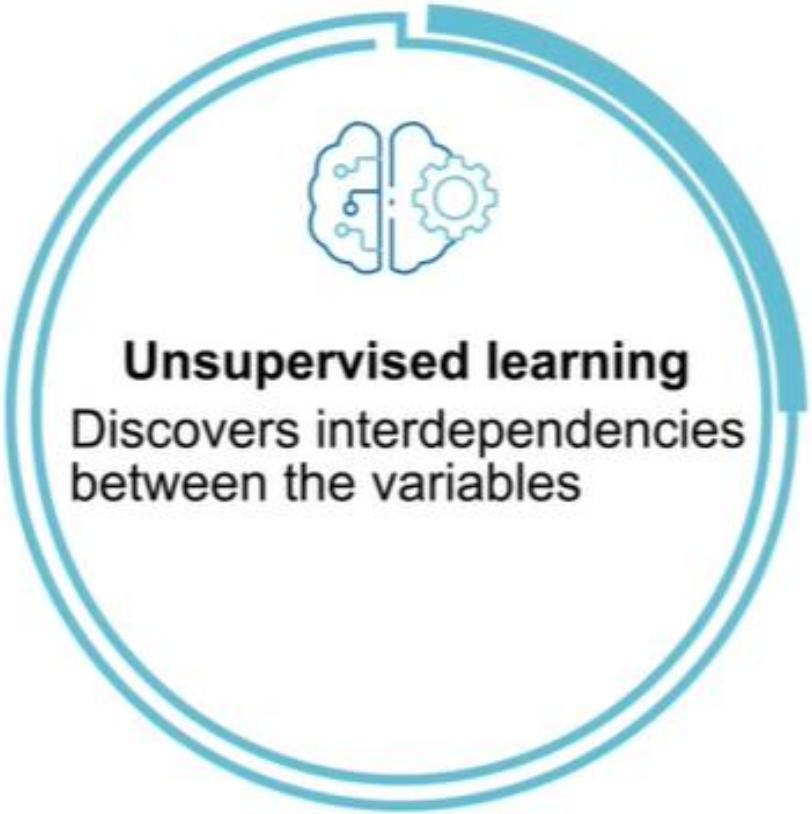
Questions that businesses are interested in:

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

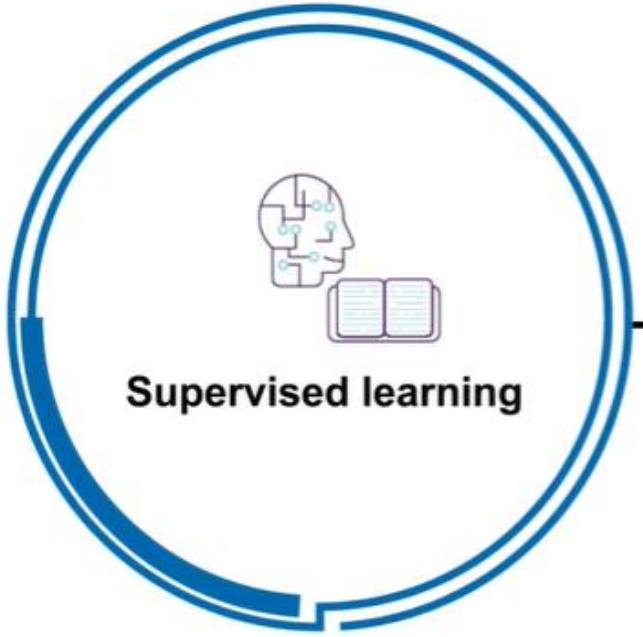
Machine Learning Domain

Pattern recognition





Machine Learning Domain



Goal: To uncover dependencies, relationships between two variables in the presence or absence of other variables

Identifies the mathematical relationship between the variables of interest (Y) and variables affecting the Y variable (X)

Machine Learning

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

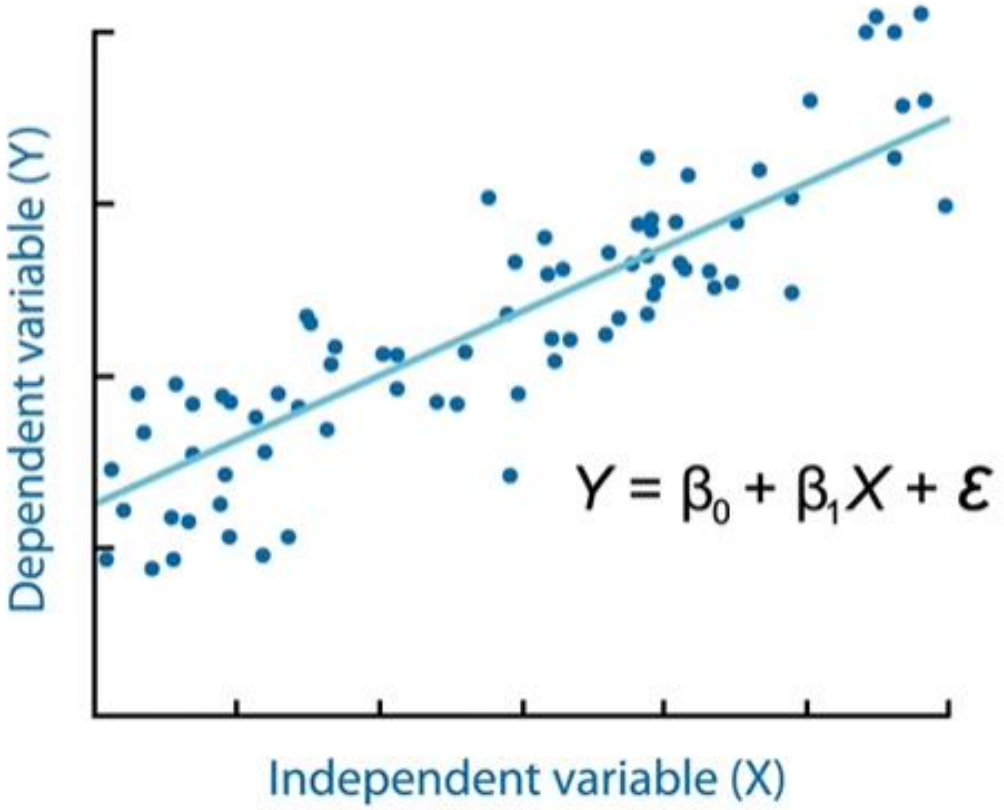
Regression

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

Sales = f(Own Price, Compt_Price, Feature, Display, etc.)

Dependent variable

Modeling Basics: From Concept to Analytical Tools

Regression model

$$Sales = \beta_0 + \beta_1 * Price + \beta_2 * Compt\ Price + \beta_3 * Display + + \beta_4 * Feature + \in$$

Y variable

Affect the sales

Conceptual model

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

Example: Weekly Sales and Marketing Data

Weekly sales data

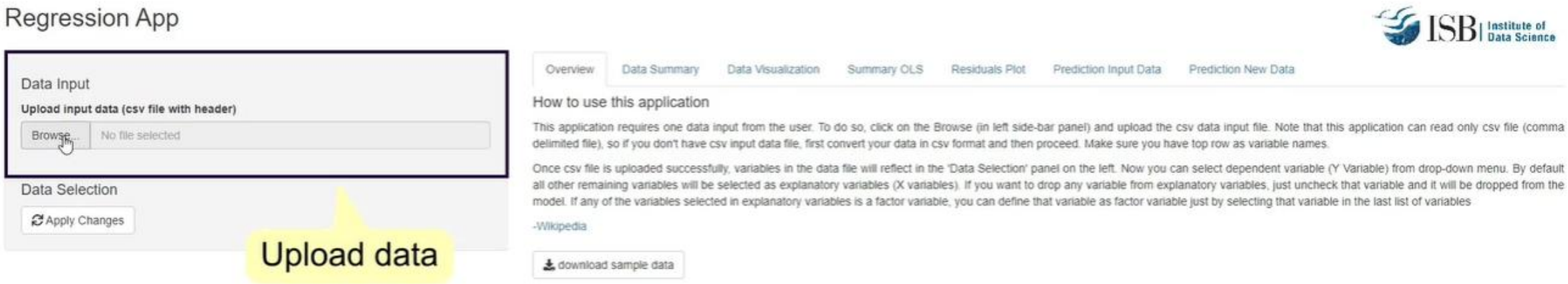
	A	B	C	D	E	F	G
1	Week	Sales Brand A	Price Brand A	Price Brand B	Price Brand C	Feature Brand A	Display Brand A
2	40	6528	3.66	1.89	2.99	1	0
3	43	6016	3.66	2.84	2.99	0	0
4	44	6272	3.66	2.84	2.59	0	0
5	45	6848	3.66	2.39	2.99	0	0
6	46	7424	3.66	2.84	2.99	0	0
7	47	6848	3.66	2.84	2.39	0	0
8	48	7488	3.66	2.39	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

Data about the brand, where it was displayed

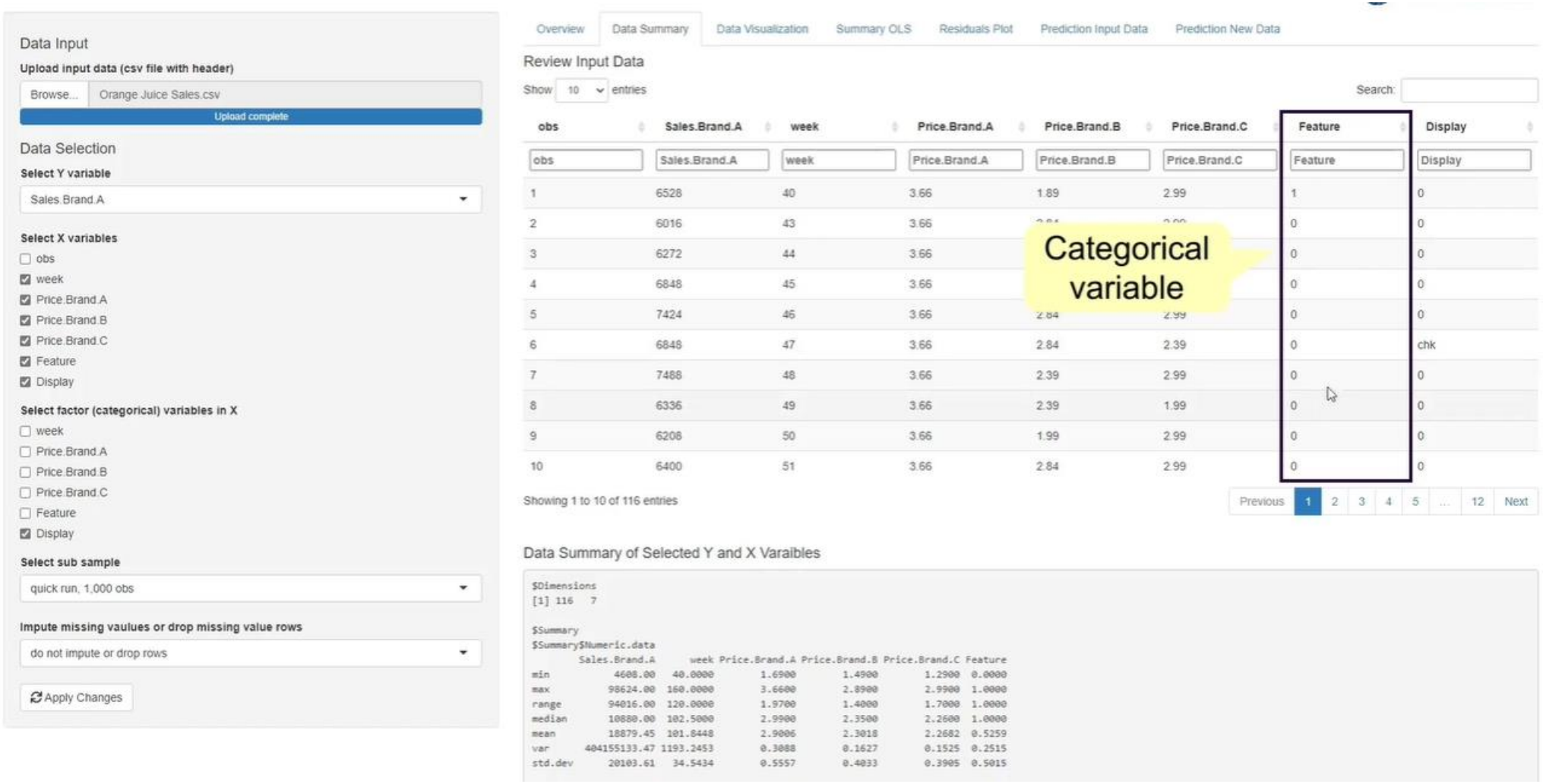
Data about your price and competitor's price

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

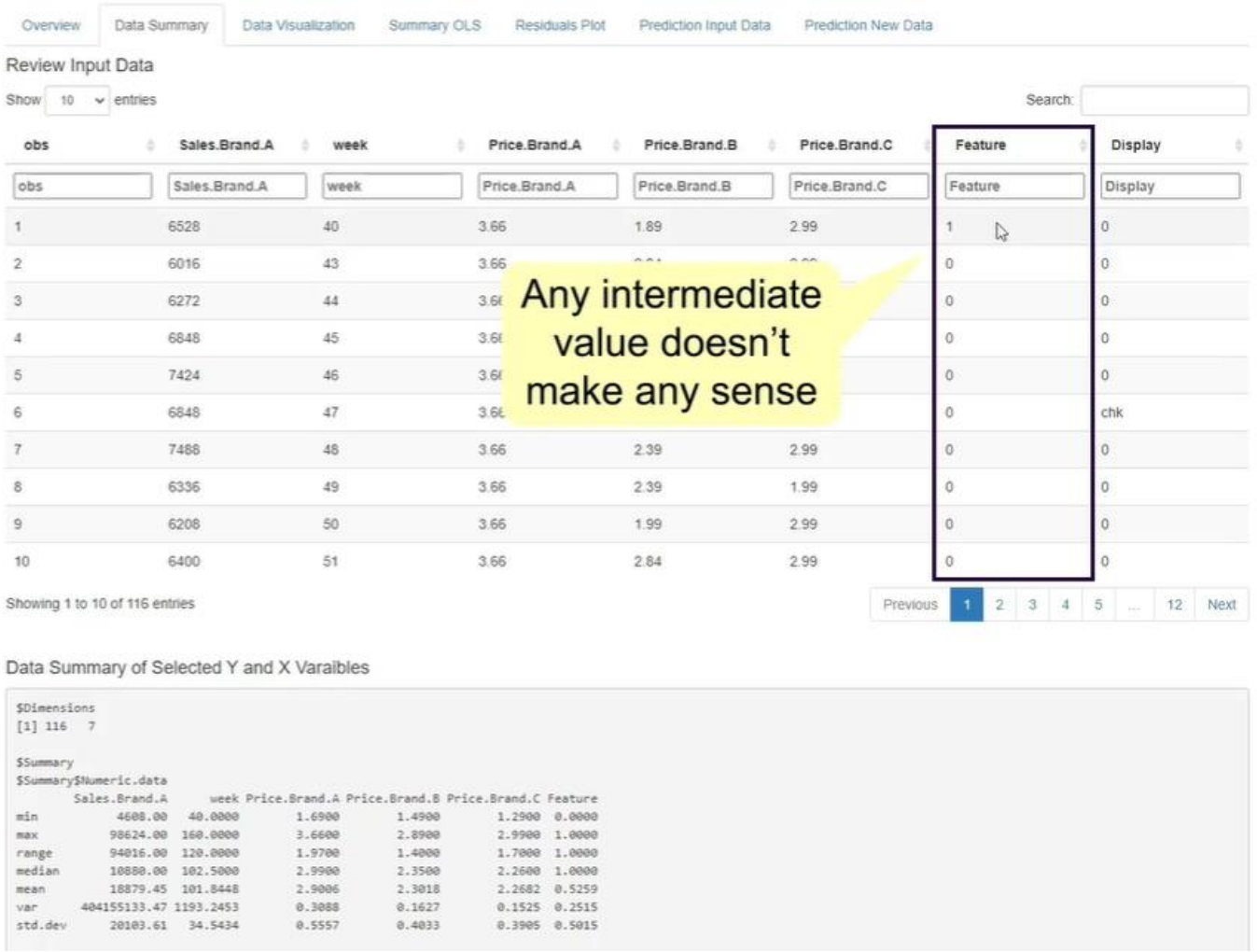
Regression App



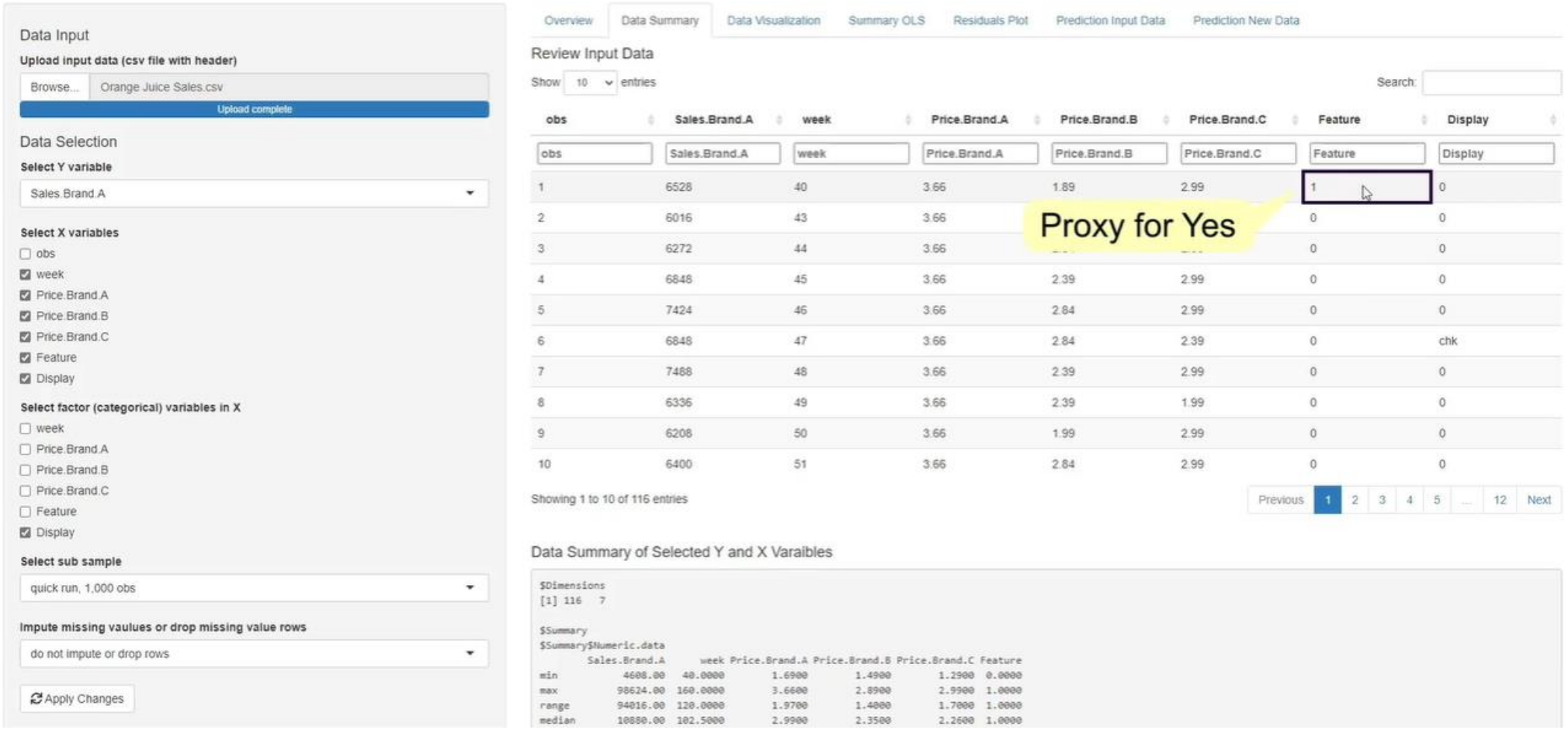
Regression App



Regression App

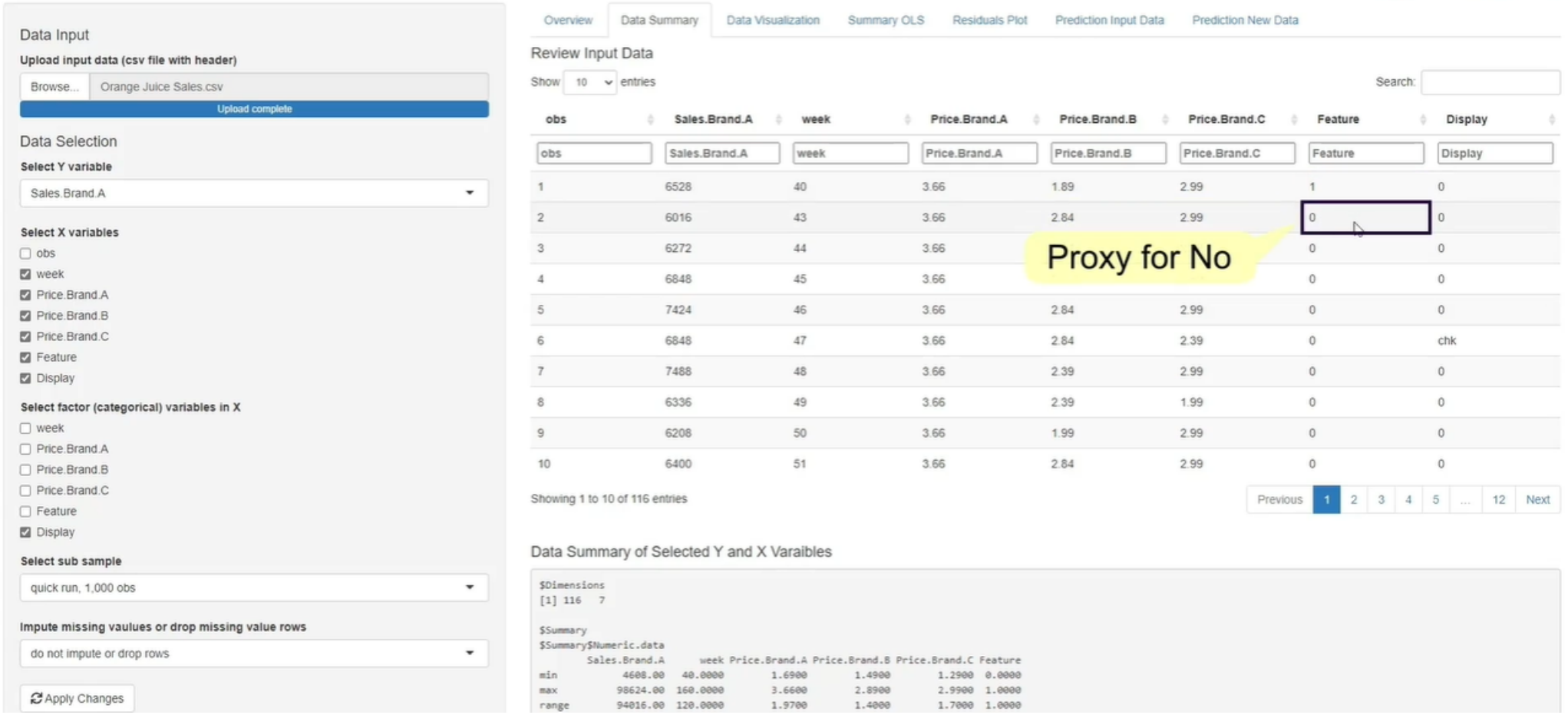


Regression App

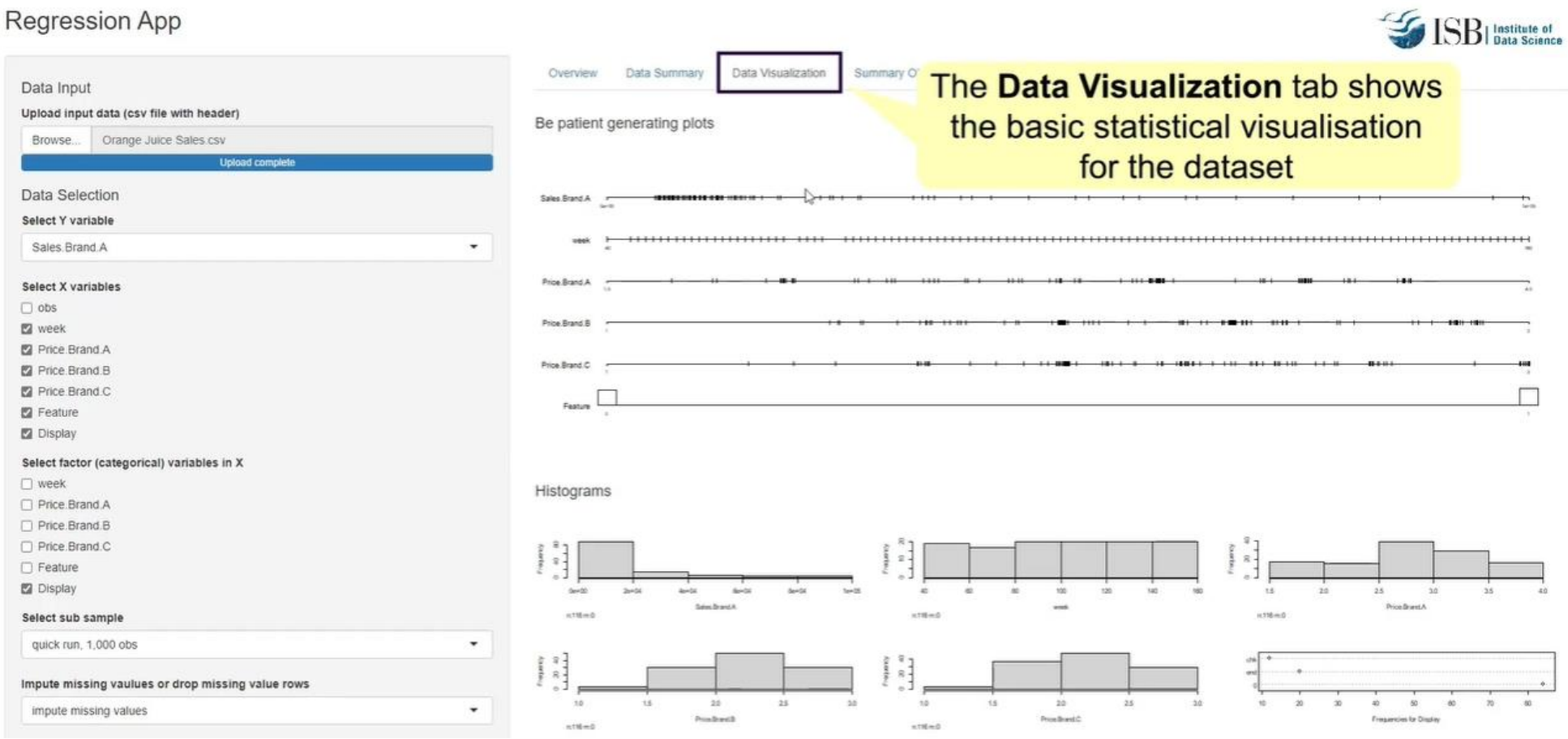


ISB Executive Education

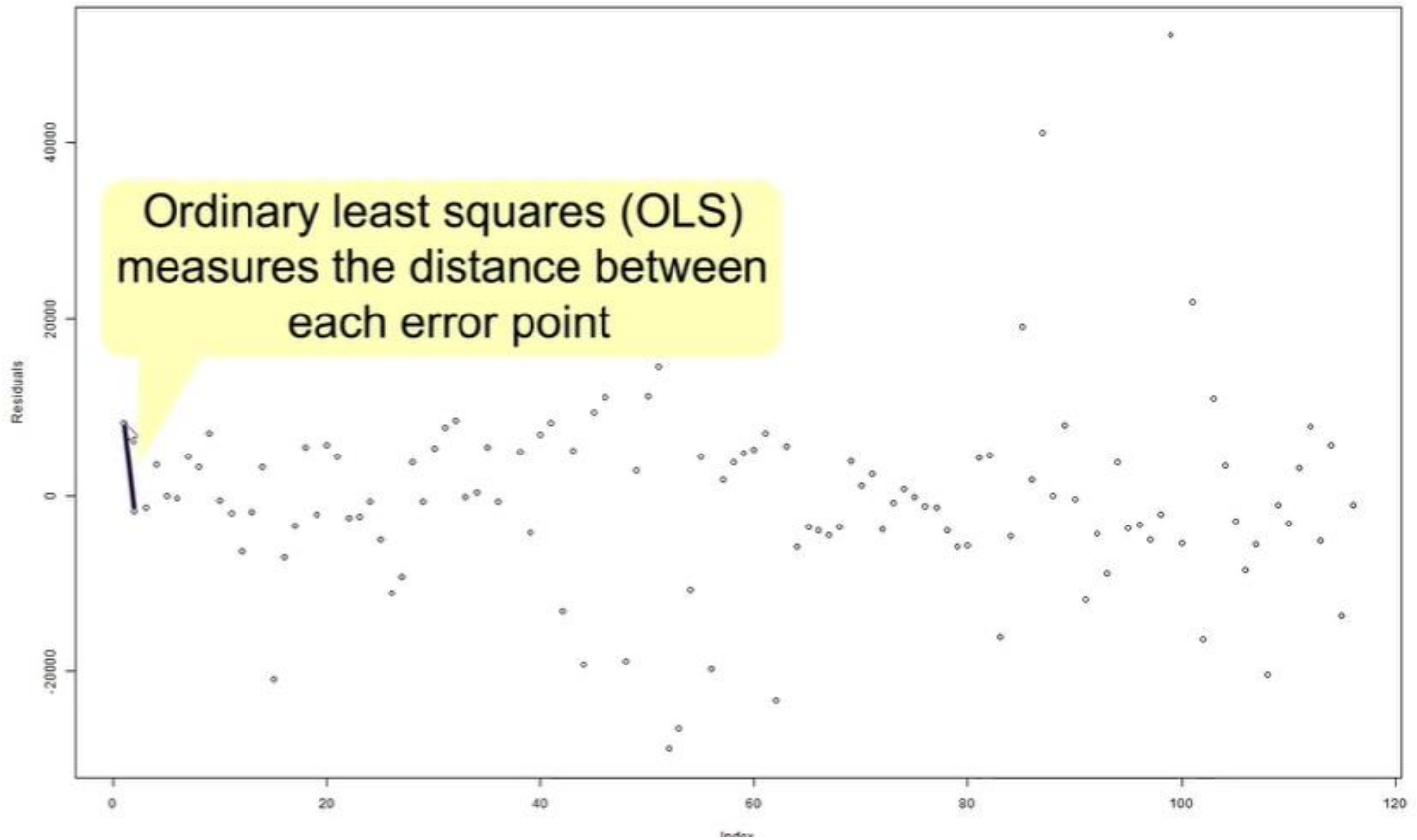
Regression App



Regression App



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
- [x] week
- [x] Price.Brand.A
- [x] Price.Brand.B
- [x] Price.Brand.C
- [x] Feature
- [x] Display

Select factor (categorical) variables in X

- [ ] week
- [ ] Price.Brand.A
- [ ] Price.Brand.B
- [ ] Price.Brand.C
- [x] Feature
- [x] Display

Select sub sample

quick run, 1,000 obs

Impute missing vaulues 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

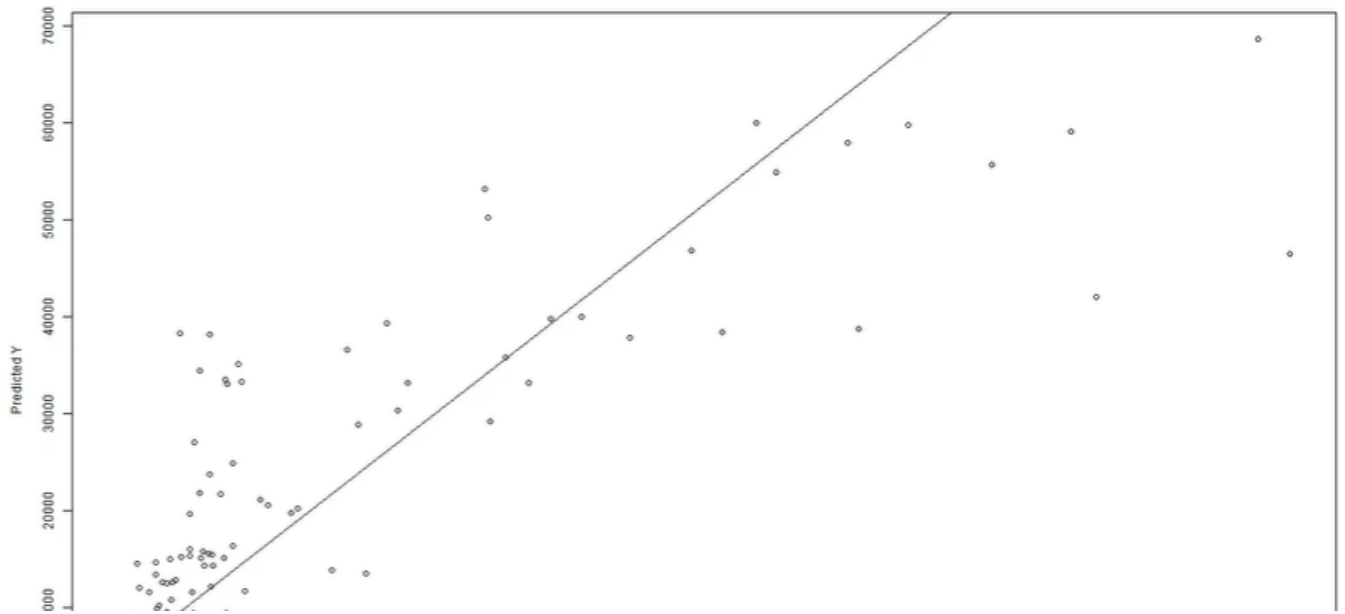
```
$Number_of_Observations
[1] 116

$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 automaticlly 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
0
0
chk
0
end
end
0

Display0	Displaychk	displayend
1	0	0
1	0	0
0	1	0
1	0	0
0	0	1
0	0	1
0	0	0

Regression App: Data Selection

- When we select display as a factor (categorical) variable, software automaticlly does the dummy variable coding (one hot coding) of the categorical variable before running a r
- Essentially, it creates one column for each level, where 1 represents th represents absence of a level for each row. e.g., 3 levels of display are

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 automaticlly 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

Dropped

- 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 automaticlly 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

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.B   9238.26    2807.66   3.290  0.00135 **
                 -52.17    3015.06  -0.017  0.98623
                -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
```

How does the price of Brand A change its sales?

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    9138.26    2807.66   3.290  0.00135 **
                            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 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

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

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 **
                  -52.17    3015.06  -0.017  0.98623
                 -904.65    2694.14  -0.336  0.73768
                 -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
```

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

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   9?38.26    2807.66   3.290  0.00135 **
Pri...          -53.17     3015.06  -0.017  0.98623
Fe...            ...55     2694.14  -0.336  0.73768
Di...           ...20.20   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
```

What is the confidence level for this value?

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 ' ' 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 ' ' 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 ' ' 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 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

99.9% confident

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 *
               .45     2865.94  -8.570 8.08e-14 ***
               .26     2807.66   3.290  0.00135 **
               .17     3015.06  -0.017  0.98623
feature        -964.65     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
```

These effects are relative to the base column, which indicated that the brand was not on display

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

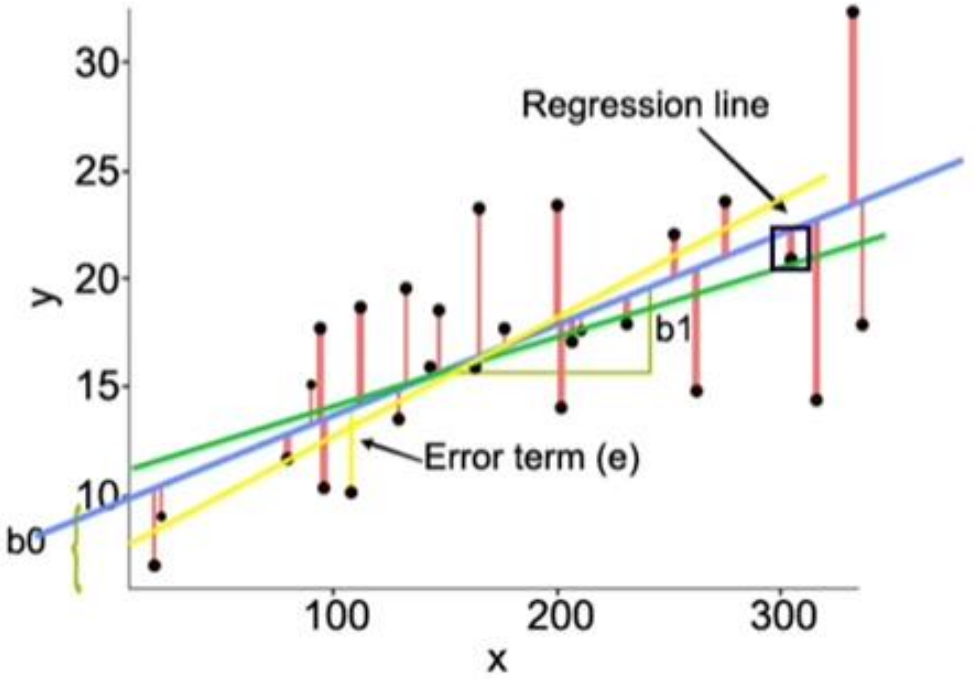
Download data (works only in browser)

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	

Predicts all X variables

Regression Analysis: Model Estimation

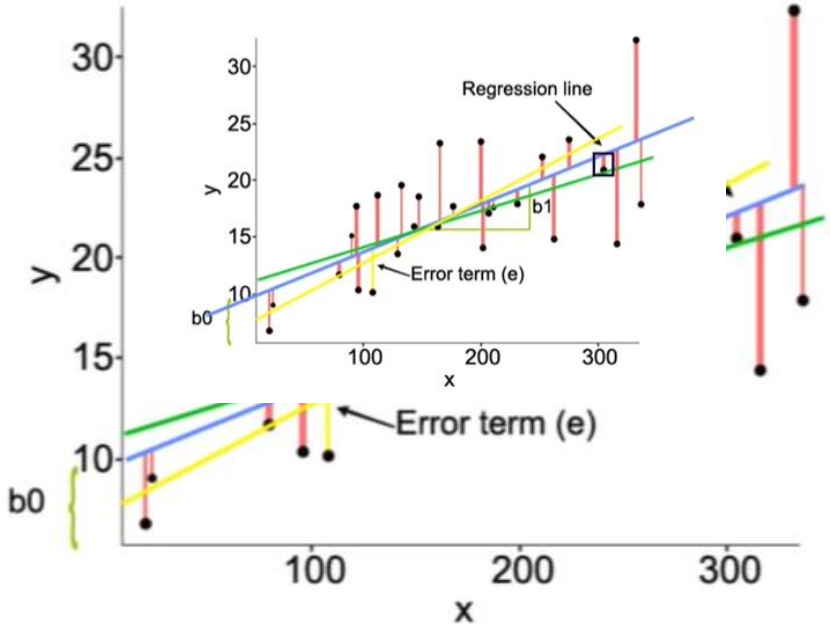
Regression analysis identifies the pattern in the different data points and fits the best line.



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

Residual standard error: 11570 on 108 degrees of f[illegible]om

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

Indicates the probability that the model is not good

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

...es : 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

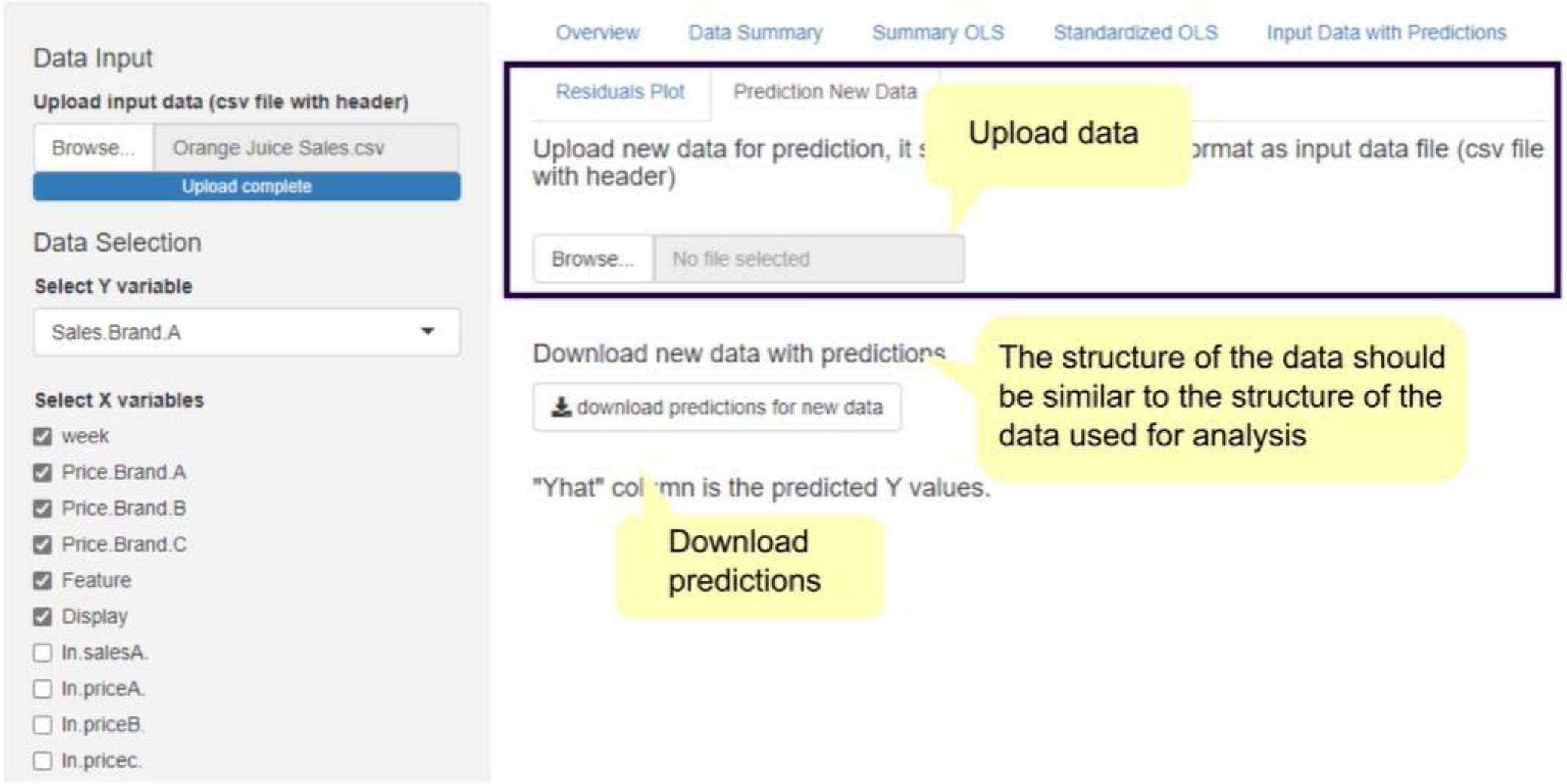
R-squared can fall between 0 and 1. If it is 0, then the model is not good, and if it is closer to 1, the model is good

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



Prediction with the Regression App

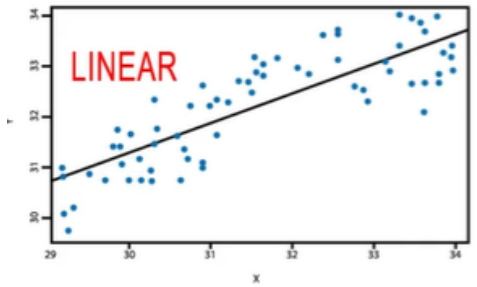
Data for prediction will:

- 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

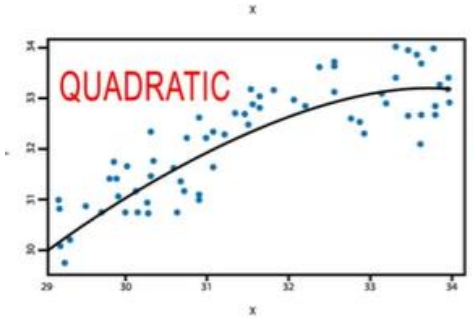


Multiple R-squared: 0.7044

Y = 30.53 + 3.05*X

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

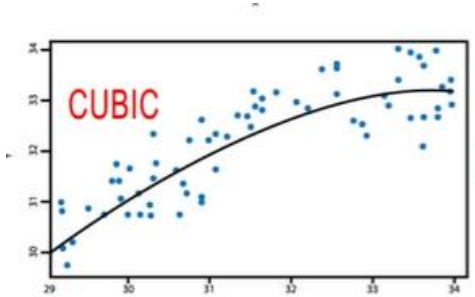


Multiple R-squared: 0.7559

$Y = 29.90 + 6.48*X - 3.22*X^2$

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

Improving the Model: Adjusted R-squared



Multiple R-squared: 0.7559

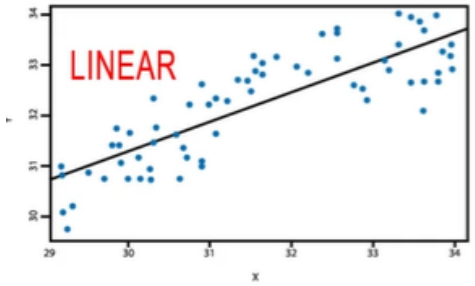
$Y = 30.17 + 3.61*X - 3.71*X^2 - 4.48*X^3$

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

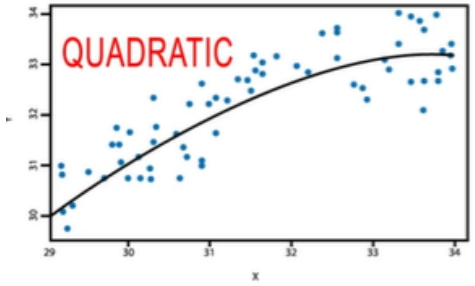
Using higher order terms will fit the model better and better; however, the goal is not for the model to closely fit the data but to infer the relationship

Improving the Model: Adjusted R-squared



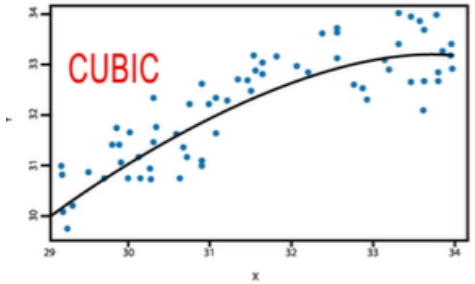
Multiple R-squared: 0.7044

$Y = 30.53 + 3.05*X$



Multiple R-squared: 0.7559

$Y = 29.90 + 6.48*X - 3.22*X^2$



Multiple R-squared: 0.7559

$Y = 30.17 + 3.61*X - 3.71*X^2 - 4.48*X^3$

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

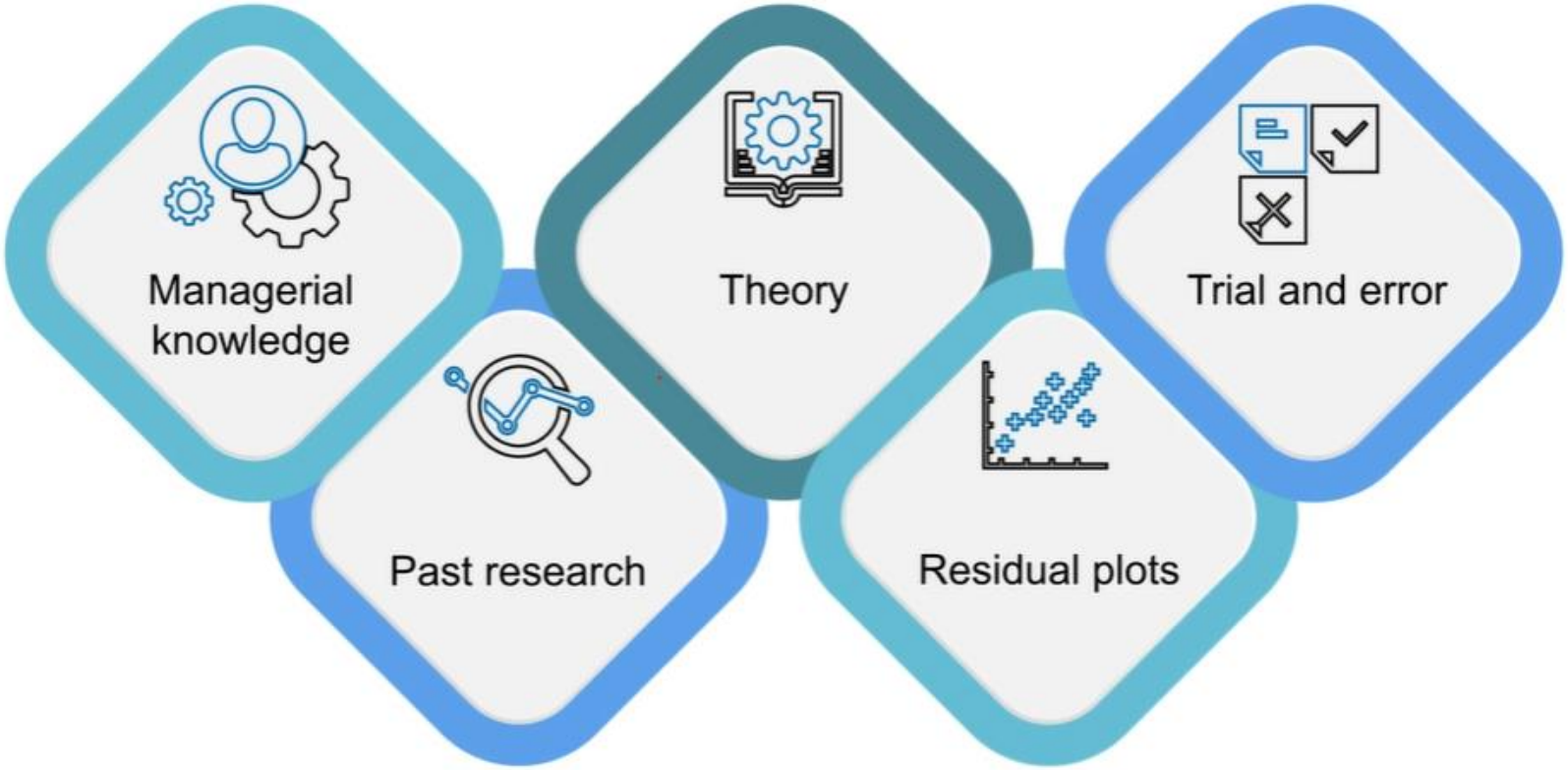
Capturing the non-linear effects in the data:

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 Regrssion Equation

Conceptual model

Sales = f (Own Price, Compt_Price, Feature, Display, 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

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

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(PriceA), log(PriceB), etc.
  - A column for log(Sales)
- Upload the new data into the app
- Select Y variable as log(Sales)
- Select the X variables as log(Prices), log(Feature) and log(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 ‘ ’ 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 ***
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

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 ‘.’ 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
```

99.9% confident

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  0.0[illegible]299 **
Ln.Price_C. -0.04501    0.19674  -0.229  0.[illegible]45
Feature      0.06550    0.07831   0.
Display      0.63220    0.09439   6.
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ (

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
```

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

Regression Variants

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

Hypothesis
If a brand werefeatured 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(Own Price, Compt Price, Feature, Display, Feature&Display)

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

Regression with Interaction Effects

Conceptual Model

Sales = f(Own Price, Compt Price, Feature, Display, Feature&Display)

Modelling Equation

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

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

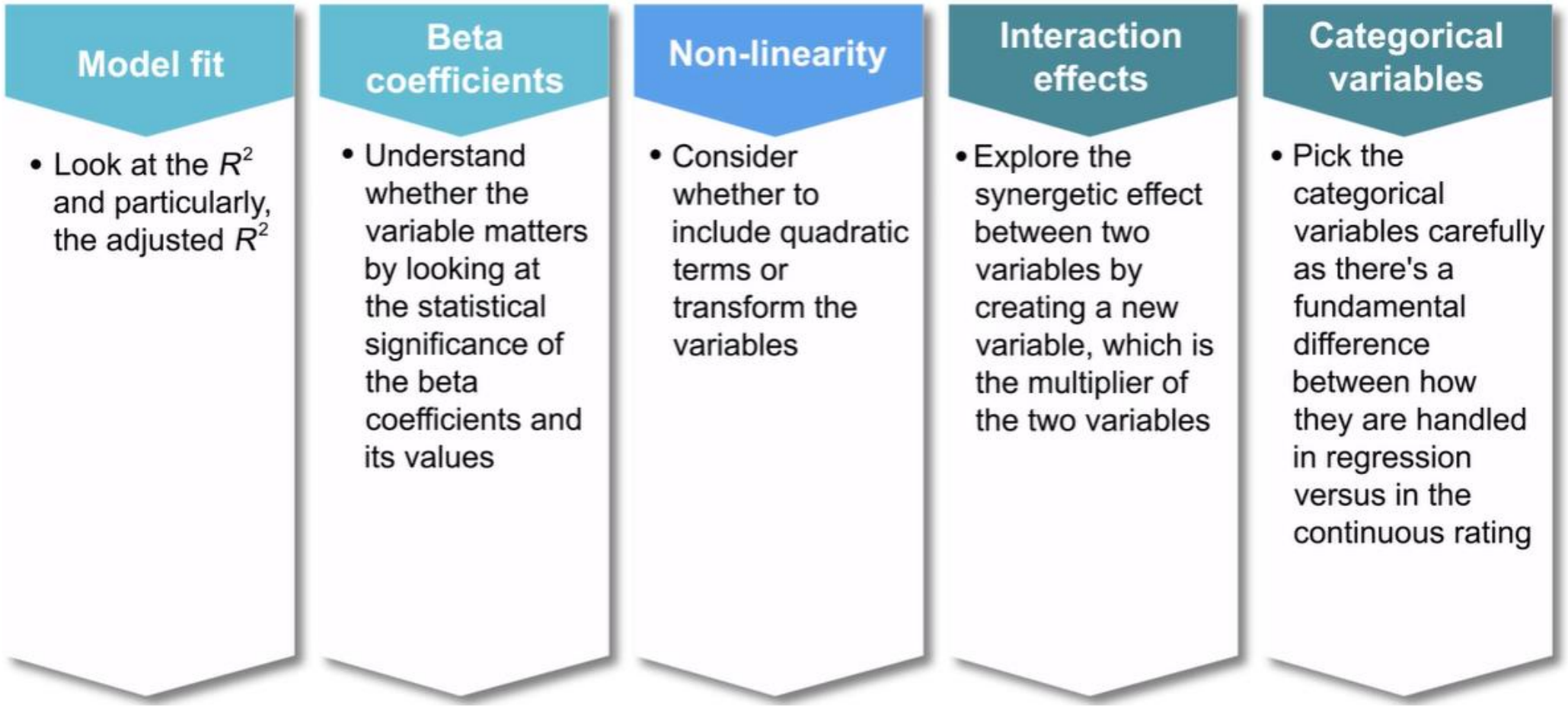
Recap: Linear Regression

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

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

Where,
Y: Dependent variable
X: Independent variables
β: Coefficients
ε: Error term

Regression Analysis: Areas of Focus



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:

- Multicolinearity
- 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