DATA SCIENCE WITH R

REGRESSION

★ Regression Analysis ★

REGRESSION ANALYSIS

➔ **Overview**

Simple Linear Regression

Multiple Linear Regression

Regression Assumptions

Implementation in SAS

Regression

A regression is used to understand and quantify cause-effect relationships

Regression

A regression is used to understand and quantify cause-effect relationships

For example, what happens to sales of a brand of shampoo if there is a discount of 15% offered in a particular week?

Regression

A regression is used to understand and quantify cause-effect relationships

For example, what happens to sales of a brand of shampoo if there is a discount of 15% offered in a particular week?

We expect sales to go up

Regression

A regression is used to understand and quantify cause-effect relationships

For example, what happens to sales of a brand of shampoo if there is a discount of 15% offered in a particular week?

We expect sales to go up

Here:

the cause: → a reduction in price

Regression

A regression is used to understand and quantify cause-effect relationships

For example, what happens to sales of a brand of shampoo if there is a discount of 15% offered in a particular week?

We expect sales to go up

Here:

the cause: → a reduction in price

the effect: → an increase in sales

Regression

We know that the effect of a decrease in price is an increase in sales -

What if we also want to know, by how much? What is the increase in sales because of a 15% discount in price?

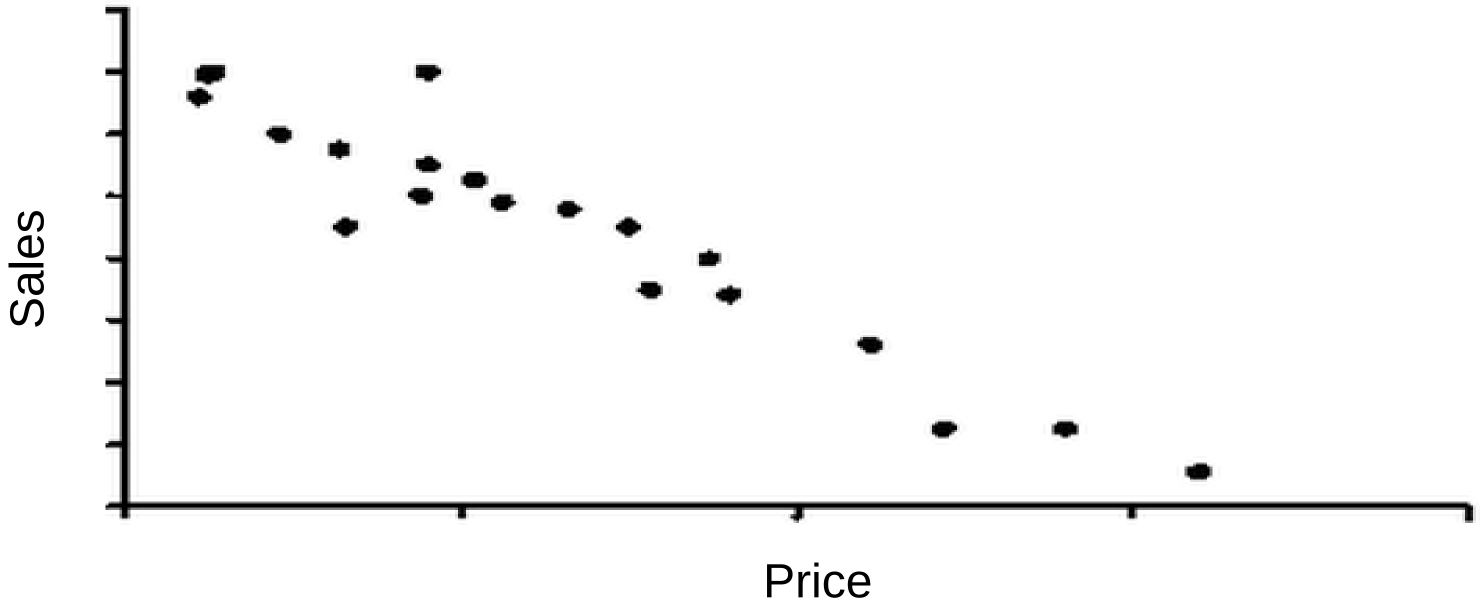
That is the quantification of the impact

Regression

Regression analysis is a statistical technique used to infer the magnitude and direction of a possible causal relationship between an observed pattern and variables assumed to have an impact on the observed pattern

Regression

Regression analysis is a statistical technique used to infer the magnitude and direction of a possible causal relationship between an observed pattern and variables assumed to have an impact on the observed pattern



Regression

Statistical

Magnitude

Direction

Causal

Observed Pattern

Regression

Statistical – a mathematical approach that assumes that the pattern of interest and the variables that impact the pattern are all random samples from underlying population

Magnitude

Direction

Causal

Observed Pattern

Regression

Statistical – a mathematical approach that assumes that the pattern of interest and the variables that impact the pattern are all random samples from underlying population

Magnitude – Size of impact (2 times, 10 times)

Direction

Causal

Observed Pattern

Regression

Statistical – a mathematical approach that assumes that the pattern of interest and the variables that impact the pattern are all random samples from underlying population

Magnitude – Size of impact (2 times, 10 times)

Direction – Positive or Negative

Causal

Observed Pattern

Regression

Statistical – a mathematical approach that assumes that the pattern of interest and the variables that impact the pattern are all random samples from underlying population

Magnitude – Size of impact (2 times, 10 times)

Direction – Positive or Negative

Causal – Magnitude of rainfall has impact on crop yield, but crop yield does not influence

Observed Pattern

Regression

Statistical – a mathematical approach that assumes that the pattern of interest and the variables that impact the pattern are all random samples from underlying population

Magnitude – Size of impact (2 times, 10 times)

Direction – Positive or Negative

Causal – Magnitude of rainfall has impact on crop yield, but crop yield does not influence

Observed Pattern – Dependent variable, Distribution

REGRESSION ANALYSIS

Overview

➔ **Simple Linear Regression**

Multiple Linear Regression

Regression Assumptions

Business Application

Regression

Example:

You work for a hospital and are looking to understand factors that may influence the birth weight of a baby

Regression

Example:

You work for a hospital and are looking to understand factors that may influence the birth weight of a baby

What factors would you think of?

Regression

Example:

You work for a hospital and are looking to understand factors that may influence the birth weight of a baby

What factors would you think of?

- Maternal Diet

Regression

Example:

You work for a hospital and are looking to understand factors that may influence the birth weight of a baby

What factors would you think of?

- Maternal Diet
- Gestation period

Regression

Example:

You work for a hospital and are looking to understand factors that may influence the birth weight of a baby

What factors would you think of?

- Maternal Diet
- Gestation period
- Maternal health issues

Regression

Example:

You work for a hospital and are looking to understand factors that may influence the birth weight of a baby

What factors would you think of?

- Maternal Diet
- Gestation period
- Maternal health issues
- Ethnicity

Regression

Example:

You work for a hospital and are looking to understand factors that may influence the birth weight of a baby

What factors would you think of?

- Maternal Diet
- Gestation period
- Maternal health issues
- Ethnicity
- Age of the mother

Regression

Is it possible to analyze the population?

Regression

Is it possible to analyze the population?

We can analyze a sample, and make inferences about the population based on the sample

Regression

Is it possible to analyze the population?

We can analyze a sample, and make inferences about the population based on the sample

Sample:

1115 observations from a hospital in the US

Birthweight (grams)	Weeks of gestation	Mother's education in years	Race	Smoked during pregnancy
2898	40	0	0	1
994	26	0	1	1
3977	38	2	0	0
3040	37	2	0	1
3523	38	2	0	0
3100	40	5	0	1
3670	40	6	1	0
3097	41	7	1	0
3040	39	7	1	1
3239	39	7	1	1
2955	38	8	0	0
2200	38	8	0	0
3182	40	8	1	0
3510	40	8	0	0
3381	39	8	1	1
3530	40	8	1	0
2985	38	8	1	1
3374	39	8	1	0
3765	42	8	0	0
2715	39	8	1	0
3640	39	8	1	0
3040	42	8	1	0

Regression

Is it possible to analyze the population?

We can analyze a sample, and make inferences about the population based on the sample

Sample:

1115 observations from a hospital in the US

Are these the only factors that impact birth weight of a baby?

Birthweight (grams)	Weeks of gestation	Mother's education in years	Race	Smoked during pregnancy
2898	40	0	0	1
994	26	0	1	1
3977	38	2	0	0
3040	37	2	0	1
3523	38	2	0	0
3100	40	5	0	1
3670	40	6	1	0
3097	41	7	1	0
3040	39	7	1	1
3239	39	7	1	1
2955	38	8	0	0
2200	38	8	0	0
3182	40	8	1	0
3510	40	8	0	0
3381	39	8	1	1
3530	40	8	1	0
2985	38	8	1	1
3374	39	8	1	0
3765	42	8	0	0
2715	39	8	1	0
3640	39	8	1	0
3040	42	8	1	0

Regression

We want to analyze the relationship between the variables available & the birth weight to see causes of variation in baby birth weights:

Regression

We want to analyze the relationship between the variables available & the birth weight to see causes of variation in baby birth weights:

- The effect is baby birth weight

Regression

We want to analyze the relationship between the variables available & the birth weight to see causes of variation in baby birth weights:

- The effect is baby birth weight
- The possible causes of baby birth weight in this dataset are gestation weeks, mother’s education, race, and smoking during pregnancy

Regression

We want to analyze the relationship between the variables available & the birth weight to see causes of variation in baby birth weights:

- The effect is baby birth weight
- The possible causes of baby birth weight in this dataset are gestation weeks, mother’s education, race, and smoking during pregnancy

What are possible ways of assessing these relationships?

Regression

We want to analyze the relationship between the variables available & the birth weight to see causes of variation in baby birth weights:

- The effect is baby birth weight
- The possible causes of baby birth weight in this dataset are gestation weeks, mother’s education, race, and smoking during pregnancy

What are possible ways of assessing these relationships?

- **Graphical visualization**

Regression

We want to analyze the relationship between the variables available & the birth weight to see causes of variation in baby birth weights:

- The effect is baby birth weight
- The possible causes of baby birth weight in this dataset are gestation weeks, mother’s education, race, and smoking during pregnancy

What are possible ways of assessing these relationships?

- **Graphical visualization**
- **Correlations**

Regression

We want to analyze the relationship between the variables available & the birth weight to see causes of variation in baby birth weights:

- The effect is baby birth weight
- The possible causes of baby birth weight in this dataset are gestation weeks, mother’s education, race, and smoking during pregnancy

What are possible ways of assessing these relationships?

- **Graphical visualization**
- **Correlations**
- **Run regression model**

Regression

A better way to identify the relationship between these variables is to use a regression technique

Regression

A better way to identify the relationship between these variables is to use a regression technique

Why regression?

Regression

A better way to identify the relationship between these variables is to use a regression technique

Why regression?

- **Multiple factor impact on the effect**

Regression

A better way to identify the relationship between these variables is to use a regression technique

Why regression?

- **Multiple factor impact on the effect**
- **Statistical Significance of the impact**

Regression

A better way to identify the relationship between these variables is to use a regression technique

Why regression?

- **Multiple factor impact on the effect**
- **Statistical Significance of the impact**

We will review the simplest type of regression, linear regression

To Be Continued

Regression Analysis

Simple Linear Regression

THANK YOU