LECTURE 2
CROSS-INDUSTRY STANDARD PROCESS FOR DATA MINING (CRISP-DM)
&
PREDICTIVE ANALYTICS I

LEK HSIANG HUI

OUTLINE

CRISP-DM

Simple Linear Regression

Multi Linear Regression

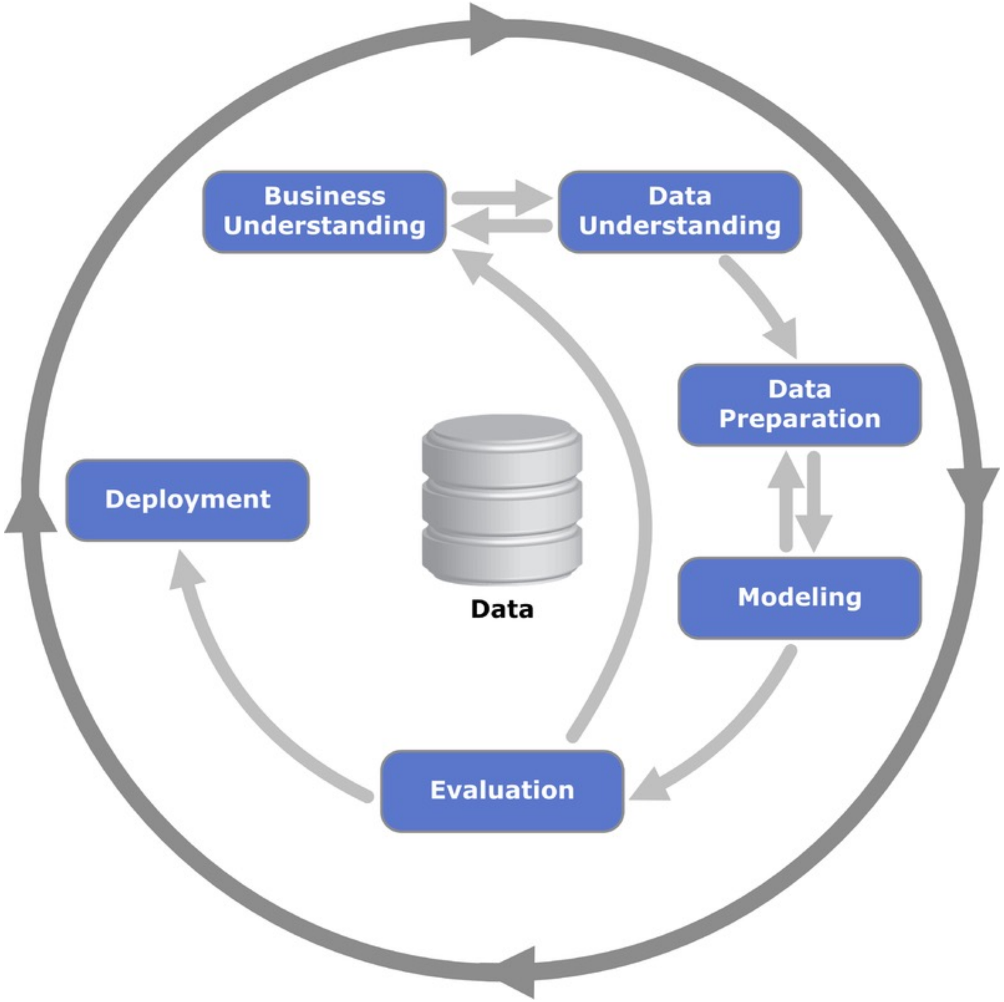
Coding Scheme for Categorical Variables

CRISP-DM

CRISP-DM → Simple Linear Regression → Multi Linear Regression → Coding Scheme for Categorical Variables

CRISP-DM

Cross-industry standard process for data mining (CRISP-DM) breaks the process of data mining into 6 major phases



STEP 1 – BUSINESS UNDERSTANDING

Understand the purpose of the data mining study

- Project objectives
- Requirements of the business
- Rough idea of potential data to use for analysis
- Preliminary plan

Notice that the process starts with the business understanding (i.e. problem)

- It does NOT start with the data!

STEP 2 – DATA UNDERSTANDING

Identify the relevant data from the many sources

- Normally: download and use datasets off internet
- Now: learn how to mine the datasets yourself
- Then, perform **Exploratory Data Analysis (EDA)**
  - Perform statistical analysis
  - Perform various types of visualizations

HANDS-ON: EDA - Q1

Download and access: **Regression.ipynb**

jupyter Regression (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Code

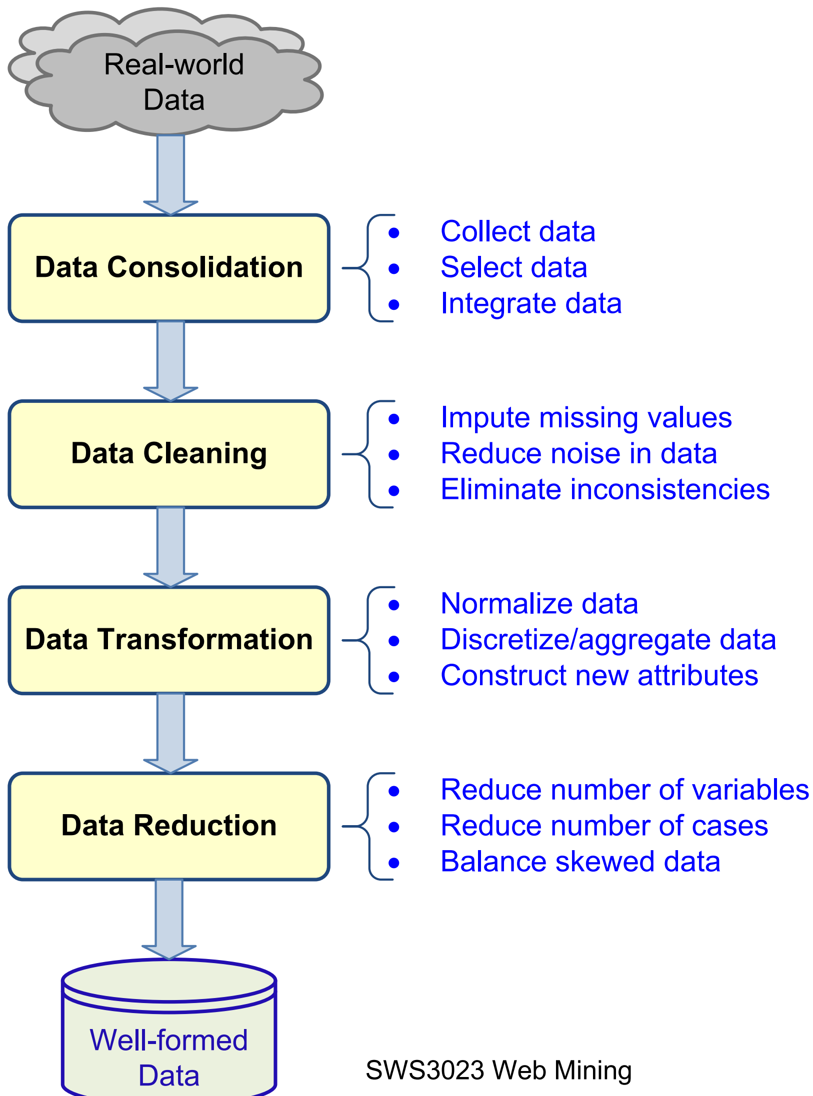
In [1]:
```
import statsmodels.api as sm
import pandas as pd
import matplotlib.pyplot as plt
```

In [2]:
```
data = pd.read_csv("Advertising.csv")
data
```

Out[2]:

	TV	Radio	Newspaper	Sales
0	230.1	37.8	69.2	22.1
1	44.5	39.3	45.1	10.4
2	17.2	45.9	69.3	9.3
3	151.5	41.3	58.5	18.5
4	180.8	10.8	58.4	12.9

STEP 3 – DATA PREPARATION



STEP 4 – MODEL BUILDING

Apply and compare various data mining techniques

- Some techniques have specific requirements on the form of data (e.g. need to be numeric)
- Most techniques can only be applied to one type of problem (e.g. classification) while others can be applied for both regression and classification

STEP 5 – TESTING AND EVALUATION

Evaluate the models developed in step 4 (depending on the problem)

- Regression – how far is the prediction from the actual values
- Classification – classification error rates
- Could also have other evaluation methods for other tasks

We usually divide the labeled data into training and testing data and perform K-Fold Cross Validation

STEP 6 – DEPLOYMENT

Development and assessment of model is usually not the end of the project

Depending on the requirements, the deployment phase can be:

- As simple as generating a report
- Or as complex as implementing a system that uses the model for daily operations

Monitoring and maintenance of models

- Over time, the models built may be become obsolete

SIMPLE LINEAR REGRESSION

CRISP-DM → **Simple Linear Regression** → Multi Linear Regression → Coding Scheme for Categorical Variables

ADVERTISING EXAMPLE

Suppose we hypothesize that there is a relationship between <u>Sales</u> and amount spend on <u>TV</u> advertisement

SIMPLE LINEAR REGRESSION

Simple linear regression assumes that there is a single predictor variable X and the relationship between the response Y and X is linear

This model contains 2 unknown constants that we aim to find

$$\underbrace{\beta_0}_{\text{intercept}} \quad \underbrace{\beta_1}_{\text{Slope}}$$

$$Y \approx \beta_0 + \beta_1 X$$

intercept

Slope

ADVERTISING EXAMPLE

Assume that there is a <u>linear relationship</u> between <u>Sales</u> and amount spend on <u>TV</u> advertisement

$$Sales \approx \beta_0 + \beta_1 TV$$

- Want to see how the spending on TV advertisement can affect Sales
- How to estimate β_0 and β_1?
    - Using training data (supervised learning)

TRAINING DATA

Advertising.csv

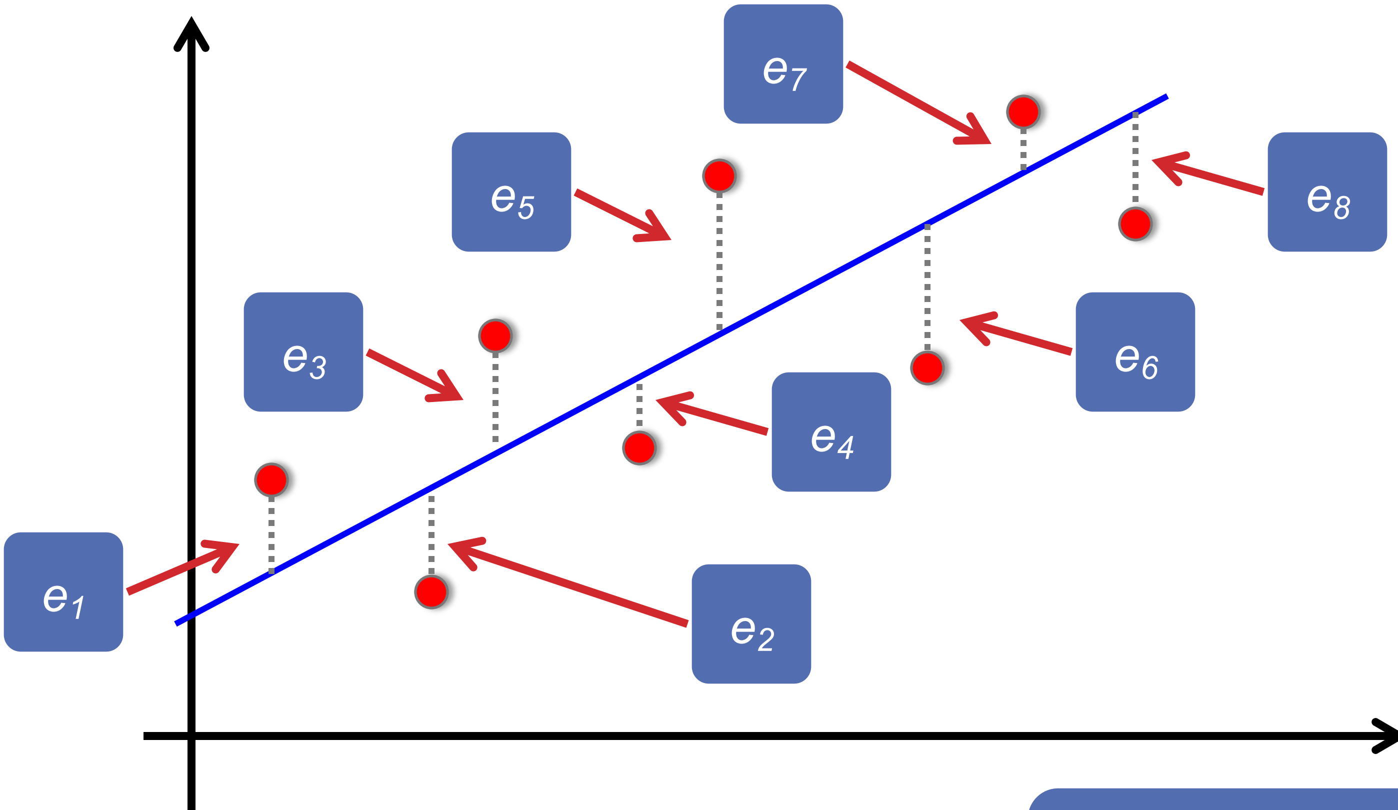
	TV	Sales
1	230.1	22.1
2	44.5	10.4
3	17.2	9.3
4	151.5	18.5
5	180.8	12.9
6	8.7	7.2
7	57.5	11.8
8	120.2	13.2
9	8.6	4.8
10	199.8	10.6
11	66.1	8.6
12	214.7	17.4
13	23.8	9.2
14	97.5	9.7
15	204.1	19
16	195.4	22.4
17	67.8	12.5
18	281.4	24.4
19	69.2	11.3
20	147.3	14.6
21	218.4	18
22	237.4	12.5

200 observations

Thousands $ spent

Thousands Units sold

LEAST SQUARES CRITERION



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

$$E = e_1^2 + e_2^2 + \ldots + e_8^2$$

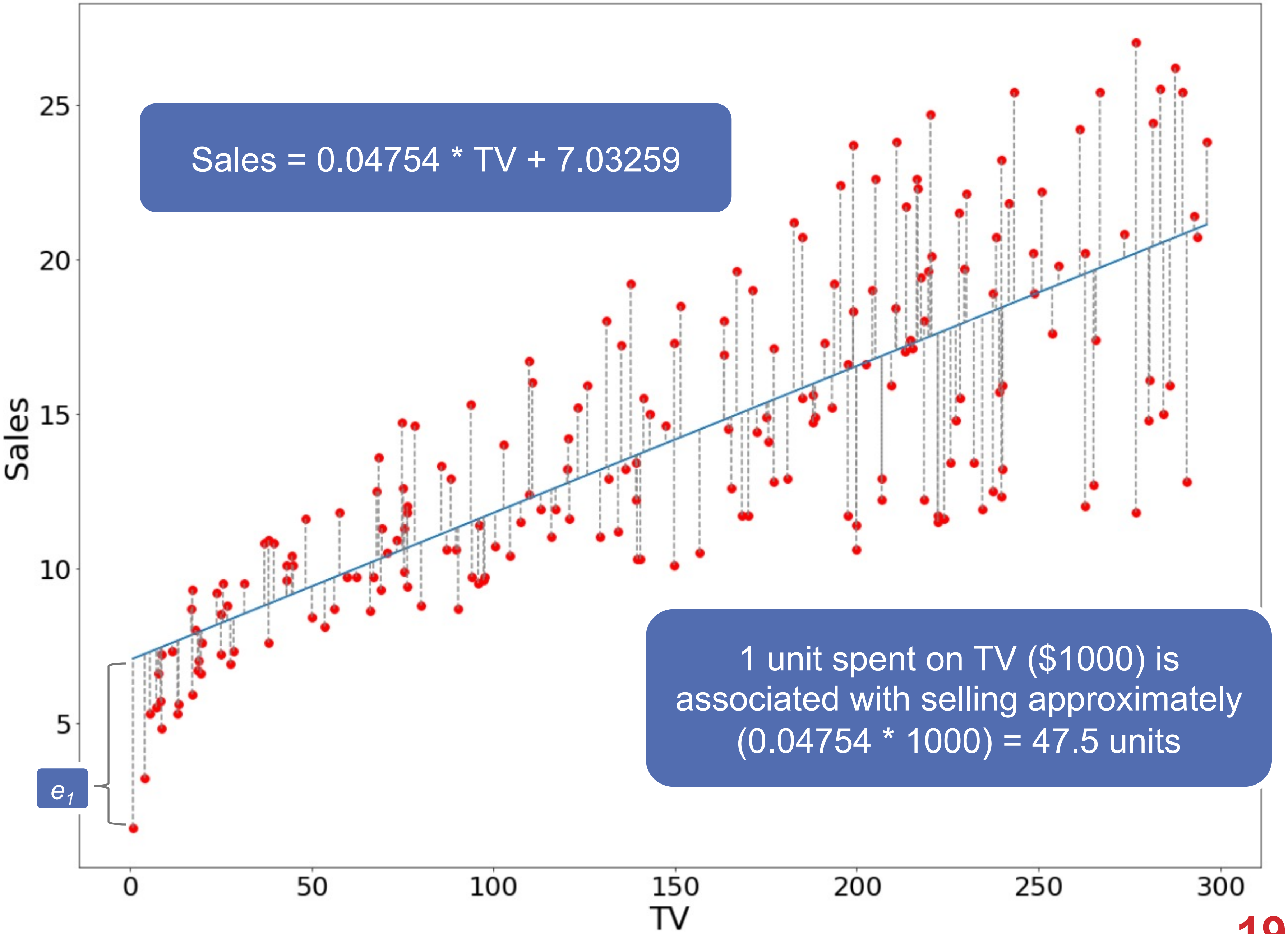
β_0 and β_1 estimate by minimizing the least squares criterion

LEAST SQUARES FIT

- Let $\hat{y}_{\hat{i}} = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y based on the i^{th} value of X
- **R**esidual **S**um of **S**quares (RSS)

$$RSS = e_1^2 + e_2^2 + \ldots + e_n^2$$

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



Sales = 0.04754 * TV + 7.03259
1 unit spent on TV ($1000) is associated with selling approximately (0.04754 * 1000) = 47.5 units
e_1
Sales
TV
25
20
15
10
5
0
50
100
150
200
250
300

HANDS-ON: REGRESSION

Download and access: **Regression.ipynb**

jupyter Regression (autosaved)

In [1]:
```python
import statsmodels.api as sm
import pandas as pd
import matplotlib.pyplot as plt
```

In [2]:
```python
data = pd.read_csv("Advertising.csv")
data
```

Out[2]:

	TV	Radio	Newspaper	Sales
0	230.1	37.8	69.2	22.1
1	44.5	39.3	45.1	10.4
2	17.2	45.9	69.3	9.3
3	151.5	41.3	58.5	18.5
4	180.8	10.8	58.4	12.9

USEFUL PREDICTORS

To determine whether a predictor is useful:

- We check whether the p-value of the coefficient estimate is < 0.05
- Low p-value → coefficient estimate is statistically significant

MODEL SUMMARY

OLS Regression Results

Dep. Variable:	Sales	**R-squared:**	0.612
Model:	OLS	**Adj. R-squared:**	0.610
Method:	Least Squares	**F-statistic:**	312.1
Date:	Sat, 12 Jun 2021	**Prob (F-statistic):**	1.47e-42
Time:	12:49:18	**Log-Likelihood:**	-519.05
No. Observations:	200	**AIC:**	1042.
Df Residuals:	198	**BIC:**	1049.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	7.0326	0.458	15.360	0.000	6.130	7.935
TV	0.0475	0.003	17.668	0.000	0.042	0.053

MEASURE MODEL PERFORMANCE

To measure the quality of fit (of the entire model), we can use:

- R^2
- F-statistics
- Mean Square Error (MSE)

R^2

R^2 measures the proportion of variability in Y that can be explained using X

- Takes value between 0 and 1
- Value close to 0 → regression did not explain much of the variability in the response (linear model likely to be wrong)
- In the Advertising dataset, $R^2 \approx 0.61$ → 0.61 of the variability in <u>Sales</u> is explained by a linear regression on <u>TV</u>
- What is a good R^2 value depends on the application

ADJUSTED R^2

R^2 will always increase with more variables

- Thus, not really a good way to evaluate the effectiveness of the predictors

Adjusted R^2 factors into the number of predictors in the calculation of R^2. (Penalize cases where many irrelevant predictors are added)

- Adjusted R^2 is always lesser than R^2
- This is often used instead

MODEL SUMMARY

OLS Regression Results

Dep. Variable:	Sales	**R-squared:**	$\boxed{0.612}$
Model:	OLS	**Adj. R-squared:**	$\boxed{0.610}$
Method:	Least Squares	**F-statistic:**	312.1
Date:	Sat, 12 Jun 2021	**Prob (F-statistic):**	1.47e-42
Time:	12:49:18	**Log-Likelihood:**	-519.05
No. Observations:	200	**AIC:**	1042.
Df Residuals:	198	**BIC:**	1049.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	7.0326	0.458	15.360	0.000	6.130	7.935
TV	0.0475	0.003	17.668	0.000	0.042	0.053

F STATISTICS

F-Statistics is another test to determine whether there is a relationship between the response and the predictors

- Value close to 1 → no relationship between the response and predictors
- Value much larger than 1 → likely to find relationship between the response and predictors
- More importantly to look at the p-value, whether the F-statistics is significant

MODEL SUMMARY

OLS Regression Results

Dep. Variable:	Sales	**R-squared:**	0.612
Model:	OLS	**Adj. R-squared:**	0.610
Method:	Least Squares	**F-statistic:**	312.1
Date:	Sat, 12 Jun 2021	**Prob (F-statistic):**	1.47e-42
Time:	12:49:18	**Log-Likelihood:**	-519.05
No. Observations:	200	**AIC:**	1042.
Df Residuals:	198	**BIC:**	1049.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	7.0326	0.458	15.360	0.000	6.130	7.935
TV	0.0475	0.003	17.668	0.000	0.042	0.053

MSE

While R^2 and F-statistics gives a rough idea of how effective is the regression model, it does not tell how much is the error

- The prediction error is sometimes more important

Mean Squared Error (MSE) is able to measure the prediction accuracy/error

$$MSE = \frac{1}{\text{degrees\_of\_freedom}} \sum_{i=1}^{n} (y_i - \underline{\hat{f}(x_i)})^2$$

Prediction for observation *i* based on our model

MODEL SUMMARY

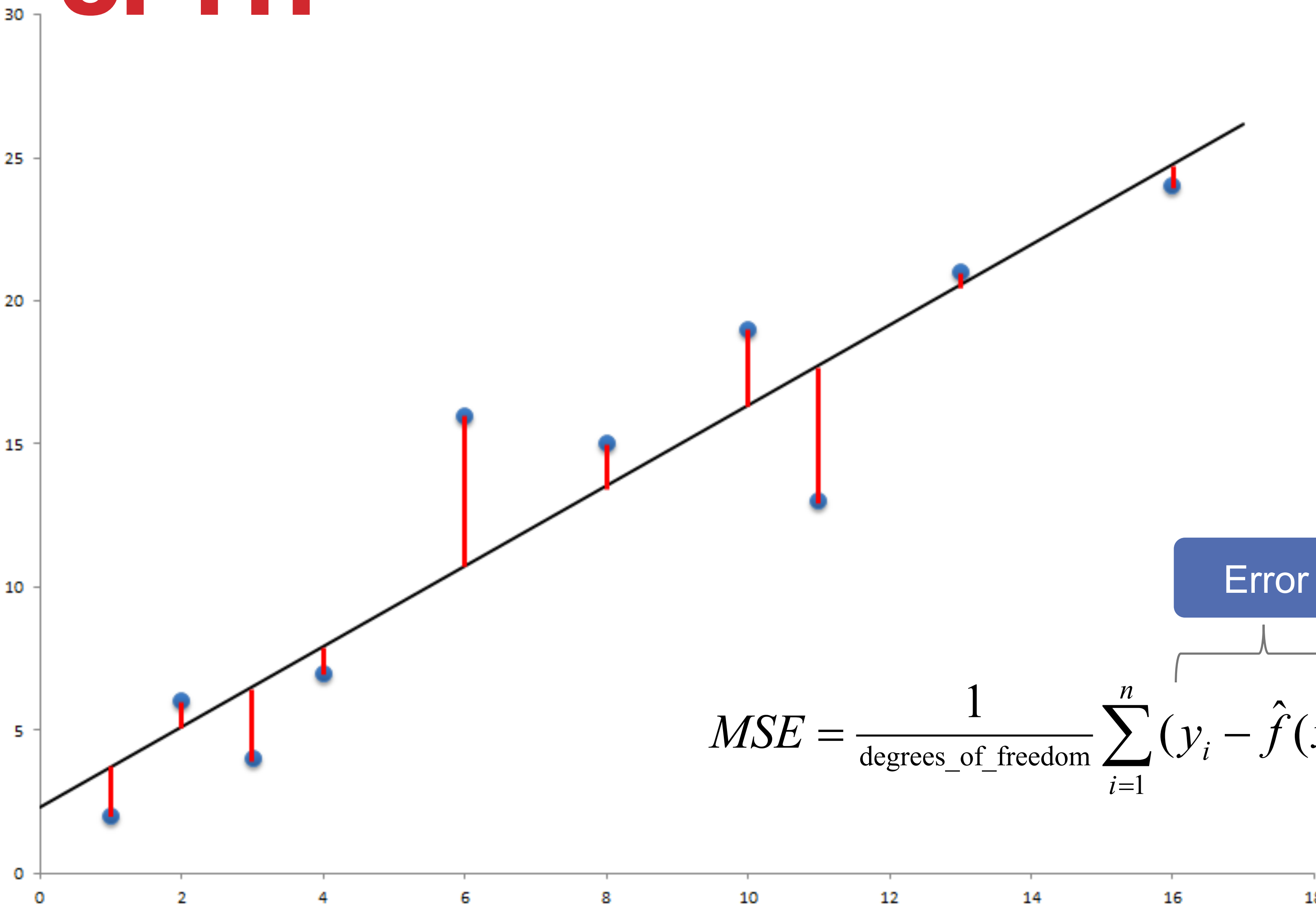
(n-2) = degrees of freedom
(Lost 2 degrees of freedom because we estimate β_0 and β_1)

OLS Regression Results

Dep. Variable:	Sales	**R-squared:**	0.612
Model:	OLS	**Adj. R-squared:**	0.610
Method:	Least Squares	**F-statistic:**	312.1
Date:	Sat, 12 Jun 2021	**Prob (F-statistic):**	1.47e-42
Time:	12:49:18	**Log-Likelihood:**	-519.05
No. Observations:	200	**AIC:**	1042.
Df Residuals:	198	**BIC:**	1049.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	7.0326	0.458	15.360	0.000	6.130	7.935
TV	0.0475	0.003	17.668	0.000	0.042	0.053

MEASURING QUALITY OF FIT



$$MSE = \frac{1}{\text{degrees\_of\_freedom}} \sum_{i=1}^{n} \underbrace{(y_i - \hat{f}(x_i))^2}_{\text{Error}}$$

MULTI LINEAR REGRESSION

CRISP-DM → Simple Linear Regression → **Multi Linear Regression** → Coding Scheme for Categorical Variables

MULTI LINEAR REGRESSION

In practice, we would have more than 1 predictor

ADVERTISING EXAMPLE

Suppose we hypothesize that there might be a linear relationship between <u>Sales</u> and amount spend on <u>TV</u> , <u>Radio</u> , <u>Newspaper</u> advertisement

$$Sales \approx \beta_0 + \beta_1 TV + \beta_2 Radio + \beta_3 Newspaper$$

- β_1, β_2, β_3 are the coefficients that quantifies the association between TV, Radio, Newspaper spending on the Sales (response)
- β_i is the average effect on Y for one unit increase in X_i while keeping the other predictors fixed

CODING SCHEME FOR CATEGORICAL VARIABLES

CRISP-DM → Simple Linear Regression → Multi Linear Regression → Coding Scheme for Categorical Variables

QUALITATIVE PREDICTORS

Regression requires the attributes to be quantitative (i.e. numerical)

Need to specially handle qualitative predictors

	Income	Limit	Rating	Cards	Age	Education	Gender	Student	Married	Ethnicity	Balance
1	14.891	3606	283	2	34	11	Male	No	Yes	Caucasian	333
2	106.025	6645	483	3	82	15	Female	Yes	Yes	Asian	903
3	104.593	7075	514	4	71	11	Male	No	No	Asian	580
4	148.924	9504	681	3	36	11	Female	No	No	Asian	964
5	55.882	4897	357	2	68	16	Male	No	Yes	Caucasian	331
6	80.18	8047	569	4	77	10	Male	No	No	Caucasian	1151
7	20.996	3388	259	2	37	12	Female	No	No	African American	203
8	71.408	7114	512	2	87	9	Male	No	No	Asian	872
9	15.125	3300	266	5	66	13	Female	No	No	Caucasian	279
10	71.061	6819	491	3	41	19	Female	Yes	Yes	African American	1350

Credit.csv

Qualitative predictors

average credit card debt balance

CODING SCHEME

How to include the gender variable?

2 values: male and female

$$Gender_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male} \end{cases}$$

Supposed we want to include income and gender:

$$Balance_i \approx \beta_0 + \beta_1 \text{Income}_i + \beta_2 Gender_i = \begin{cases} \beta_0 + \beta_1 \text{Income}_i + \beta_2 & \text{if female} \\ \beta_0 + \beta_1 \text{Income}_i & \text{if male} \end{cases}$$

INTERPRETATION

$$Balance_i \approx \beta_0 + \beta_1 \text{Income}_i + \beta_2 Gender_i = \begin{cases} \beta_0 + \beta_1 \text{Income}_i + \beta_2 & \text{if female} \\ \beta_0 + \beta_1 \text{Income}_i & \text{if male} \end{cases}$$

β_2 is the average difference in credit card balance between females and males for a given income level

- Treat males are the “baseline”
- The coding scheme (whether male should be 1 or female should be 1) will not affect the interpretation of the regression

CODING SCHEME

	Income	Limit	Rating	Cards	Age	Education	Gender	Student	Married	Ethnicity	Balance
1	14.891	3606	283	2	34	11	Male	No	Yes	Caucasian	333
2	106.025	6645	483	3	82	15	Female	Yes	Yes	Asian	903
3	104.593	7075	514	4	71	11	Male	No	No	Asian	580
4	148.924	9504	681	3	36	11	Female	No	No	Asian	964
5	55.882	4897	357	2	68	16	Male	No	Yes	Caucasian	331
6	80.18	8047	569	4	77	10	Male	No	No	Caucasian	1151
7	20.996	3388	259	2	37	12	Female	No	No	African American	203
8	71.408	7114	512	2	87	9	Male	No	No	Asian	872
9	15.125	3300	266	5	66	13	Female	No	No	Caucasian	279
10	71.061	6819	491	3	41	19	Female	Yes	Yes	African American	1350

If there is k (k ≥ 3) values, create k-1 dummy variables

$$Ethnicity_{ia} = \begin{cases} 1 & \text{if } i\text{th person is Asian} \\ 0 & \text{if } i\text{th person is not Asian} \end{cases}$$

$$Ethnicity_{ic} = \begin{cases} 1 & \text{if } i\text{th person is Caucasian} \\ 0 & \text{if } i\text{th person is not Caucasian} \end{cases}$$

How about African American?

If person is neither Asian nor Caucasian → person is African American

$$Balance_i \approx \beta_0 + \beta_1 Ethnicity_{ia} + \beta_2 Ethnicity_{ic} = \begin{cases} \beta_0 + \beta_1 & \text{if asian} \\ \beta_0 + \beta_2 & \text{if Caucasian} \\ \beta_0 & \text{if African American} \end{cases}$$

WHAT'S NEXT?

Predictive Analytics II