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

Introduction to Classification

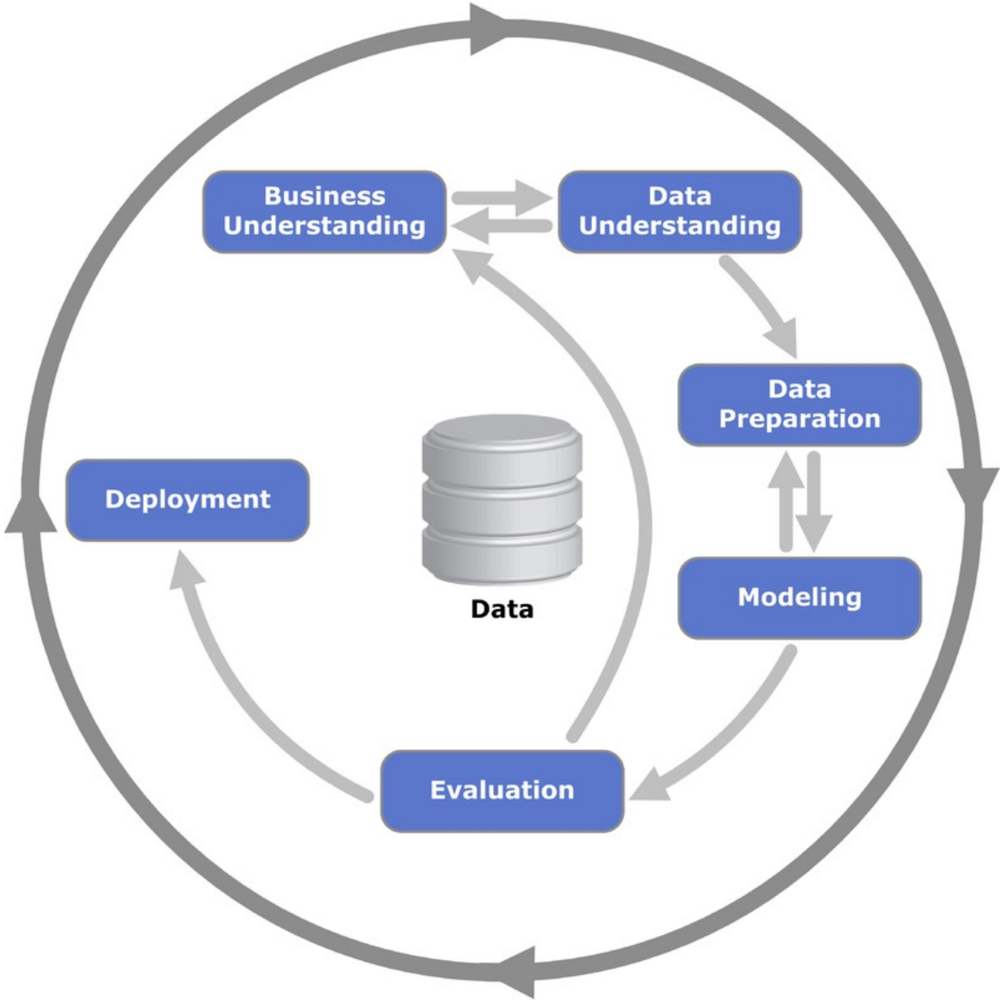
Logistic Regression

CRISP-DM

CRISP-DM → Simple Linear Regression → Multi Linear Regression → Coding Scheme for Categorical Variables → Introduction to Classification → Logistic Regression

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)

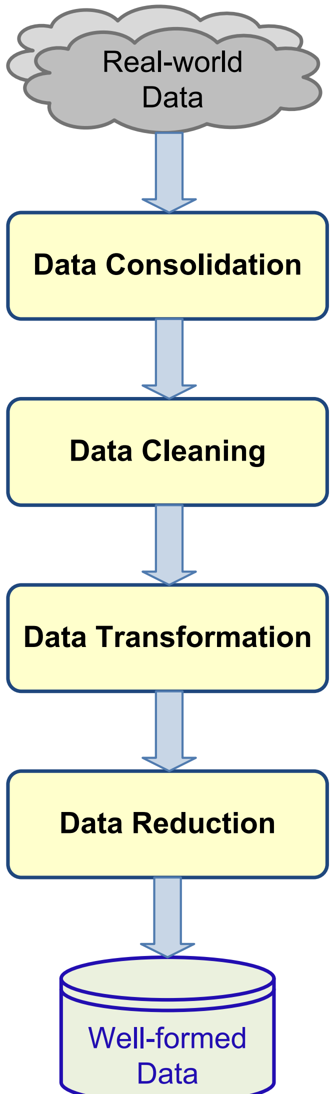
- It does NOT start with the data!

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**
    - Perform statistical analysis
    - Perform various types of visualizations

STEP 3 – DATA PREPARATION



- Collect data
- Select data
- Integrate data

- Impute missing values
- Reduce noise in data
- Eliminate inconsistencies

- Normalize data
- Discretize/aggregate data
- Construct new attributes

- Reduce number of variables
- Reduce number of cases
- Balance skewed data

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

Simple Linear Regression → Multi Linear Regression → Coding Scheme for Categorical Variables → Introduction to Classification → Logistic Regression

ADVERTISING EXAMPLE

Suppose we hypothesize that there is a relationship between Sales and amount spend on TV 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

$$Y \approx \underbrace{\beta_0}_{\text{intercept}} + \underbrace{\beta_1}_{\text{Slope}} X$$

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

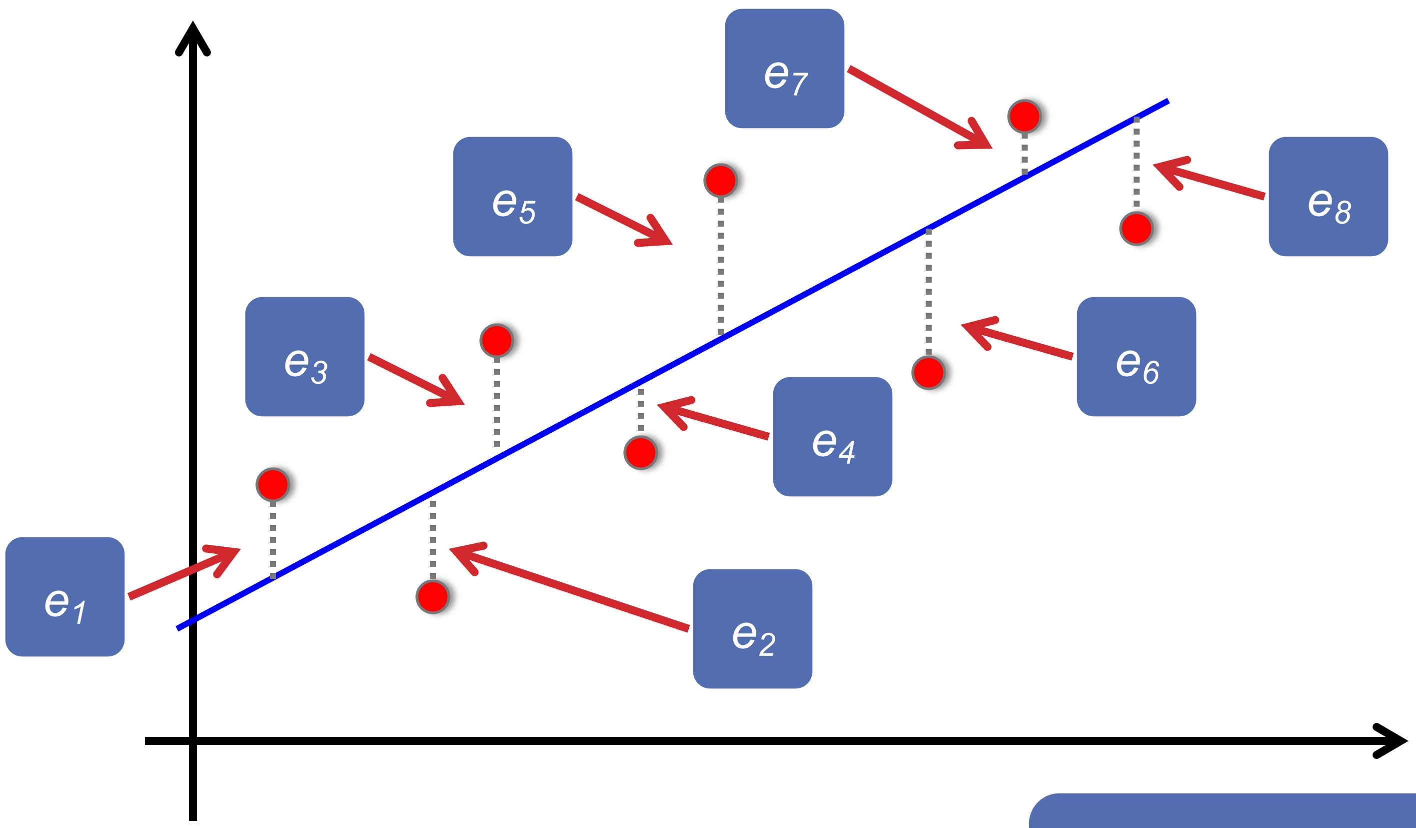
	TV	Sales
1	230.1	22.1
2	44.5	10.4
3	17.2	9.3
4	151.5	18.5
5	180.8	12.9
6	8.7	7.2
7	57.5	11.8
8	120.2	13.2
9	8.6	4.8
10	199.8	10.6
11	66.1	8.6
12	214.7	17.4
13	23.8	9.2
14	97.5	9.7
15	204.1	19
16	195.4	22.4
17	67.8	12.5
18	281.4	24.4
19	69.2	11.3
20	147.3	14.6
21	218.4	18
22	237.4	12.5

200 observations

Thousands $ spent

Thousands Units sold

LEAST SQUARES CRITERION



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

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

β_0 and β_1 estimate by minimizing the least squares criterion

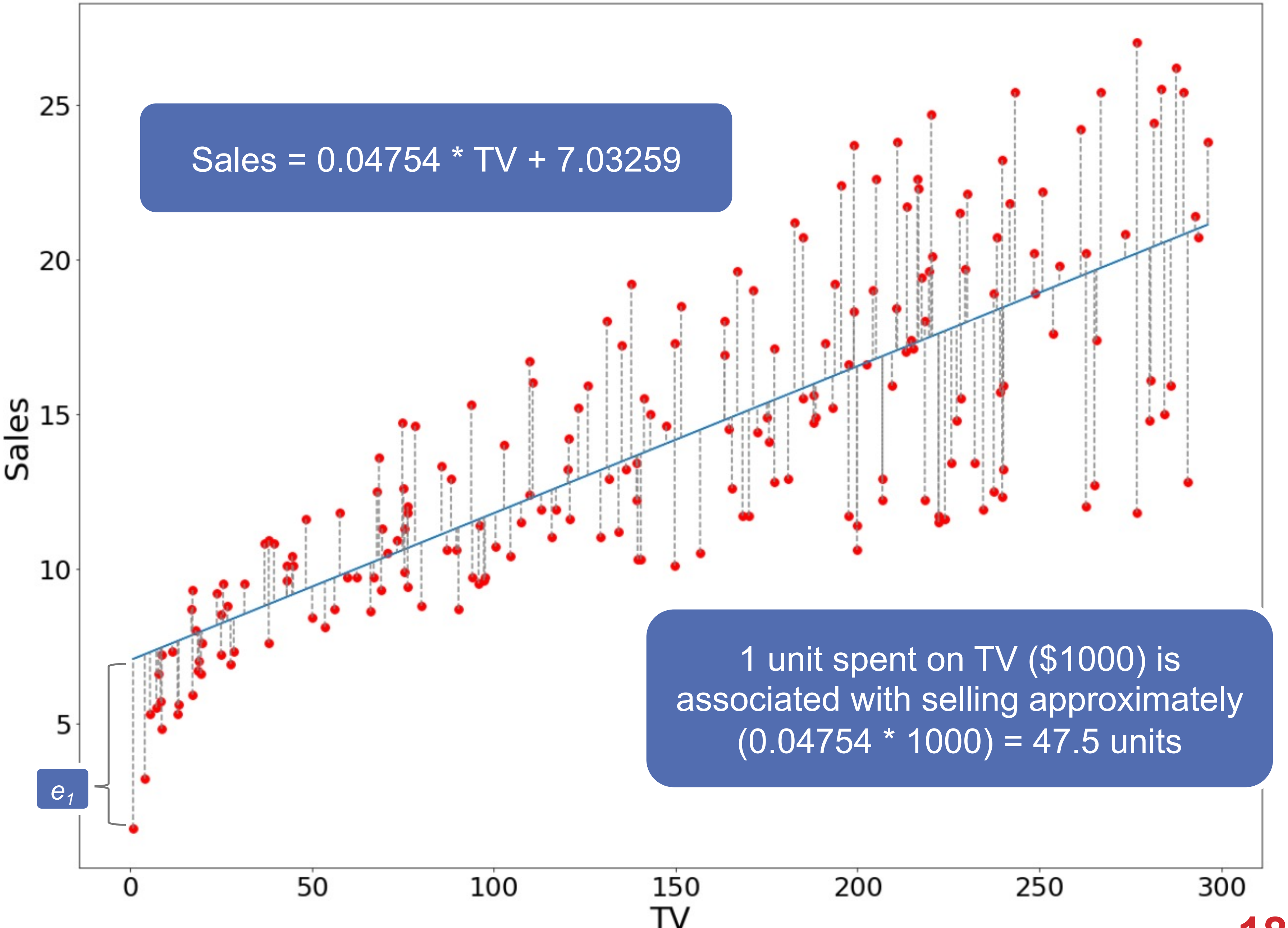
LEAST SQUARES FIT

- 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)

Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Code

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

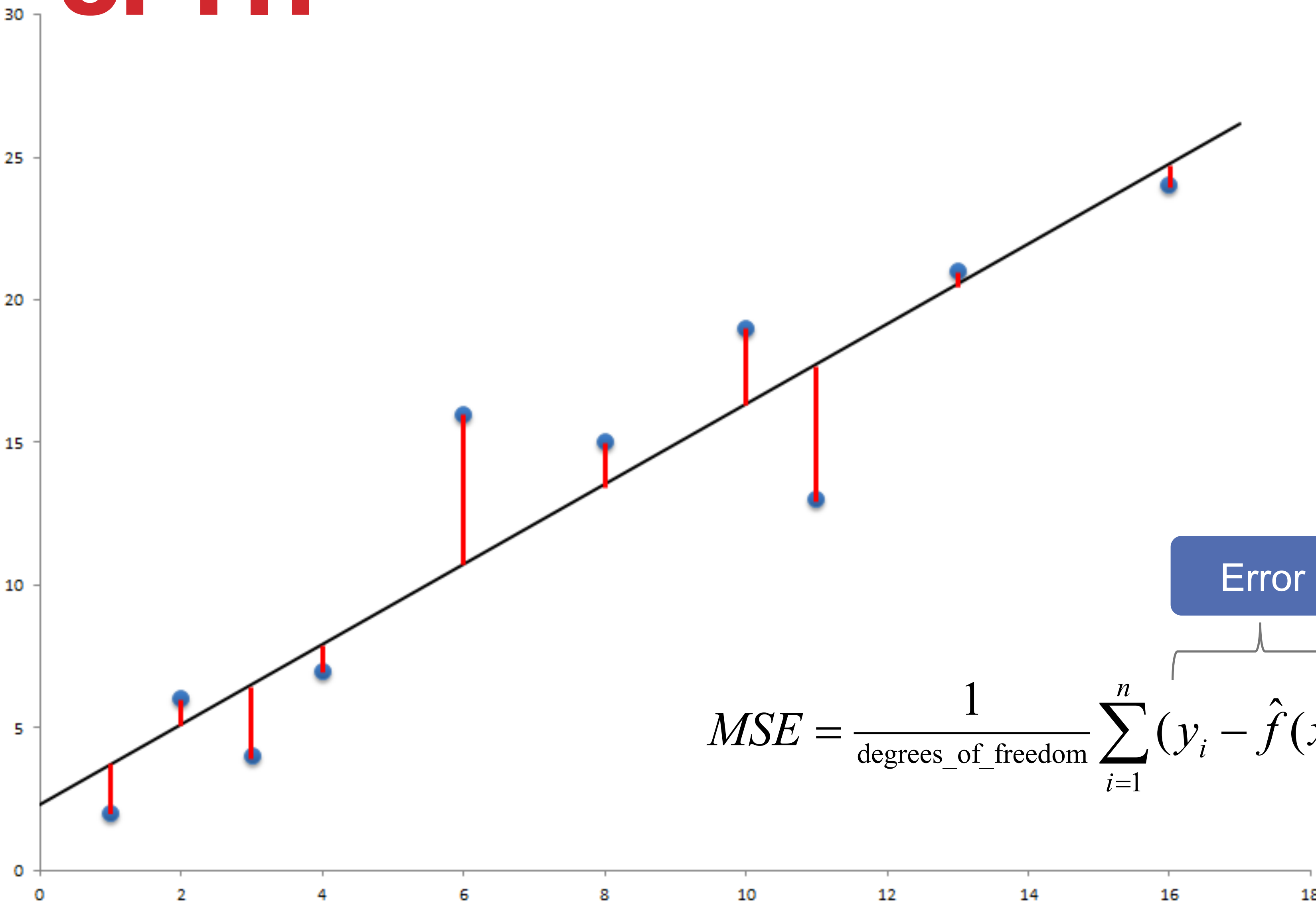
(n-2) = degrees of freedom
(Lost 2 degrees of freedom because we estimate β_0 and β_1)

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



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

MULTI LINEAR REGRESSION

Simple Linear Regression → **Multi Linear Regression** → Coding Scheme for Categorical Variables → Introduction to Classification → Logistic Regression

MULTI LINEAR REGRESSION

In practice, we would have more than 1 predictor

MULTI LINEAR REGRESSION

How do we consider these 3 predictors (TV, Radio, Newspaper)?

- One approach: run 3 separate simple linear regressions
- What's the problem with such an approach?
  - Unclear how to make a single prediction of sales based on the different advertising media budget
  - Each of the 3 regression equations are isolated from the others which might result in unexpected observations

MULTI LINEAR REGRESSION

Multi linear regression

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon$$

Generalization of Simple Linear Regression

Instead of 1, ***p*** predictors

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

LEAST SQUARES FIT

- β_0 , β_1, β_2, and β_3 are still estimated by minimizing the least squares criterion
- **R**esidual **S**um of **S**quares (RSS)

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

- where

$$e_i = y_i - (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)_i$$

MULTIPLE SIMPLE LINEAR REGRESSIONS

Coefficients are all significant

Simple regression of sales on radio

	Coefficient	Std. error	t-statistic	p-value
Intercept	9.312	0.563	16.54	< 0.0001
radio	0.203	0.020	9.92	< 0.0001

Simple regression of sales on newspaper

	Coefficient	Std. error	t-statistic	p-value
Intercept	12.351	0.621	19.88	< 0.0001
newspaper	0.055	0.017	3.30	< 0.0001

$1000 spending on radio adv → 203 units increase in sales

$1000 spending on newspaper adv → 55 units increase in sales

MULTIPLE LINEAR REGRESSION

Not significant

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	~~−0.001~~	0.0059	−0.18	0.8599

Simple regression of sales on newspaper

	Coefficient	Std. error	t-statistic	p-value
Intercept	12.351	0.621	19.88	< 0.0001
newspaper	$\boxed{0.055}$	0.017	3.30	< 0.0001

MULTI LINEAR REGRESSION (ADVERTISING EXAMPLE)

Individual simple linear regression each suggests relationship with Sales

But multi linear regression shows no significant relationship between Newspaper Adv spending and Sales

Why the conflicting observation?

- This is due to one predictor might be correlated with another

MULTI LINEAR REGRESSION (ADVERTISING EXAMPLE)

```
In [31]: data.corr()
```

Out[31]:

	TV	Radio	Newspaper	Sales
TV	1.000000	0.054809	0.056648	0.782224
Radio	0.054809	1.000000	0.354104	0.576223
Newspaper	0.056648	0.354104	1.000000	0.228299
Sales	0.782224	0.576223	0.228299	1.000000

MULTI LINEAR REGRESSION (ADVERTISING EXAMPLE)

Explanation:

- Tendency to spend more on newspaper adv on markets where we spend more on radio adv
- Supposed the model is correct, radio adv spending does increases sales
- Then, in markets where we spend more on newspaper adv, the radio adv spending is also higher, thus resulting in higher sales
- But the results of this phenomenon is because of radio adv spending (not because of the spending on newspaper adv)

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

HOW WELL DOES THE MODEL FIT THE DATA?

Recall: R^2 provides a measure of fit of the model

- Measures the proportion of variability in Y that can be explained using X
- As we add in more predictors, the R^2 will always increase
- For the advertising dataset:
  - R^2 for 1 predictors (tv) = 0.6118751
  - R^2 for 2 predictors (tv+radio) = 0.8971943
  - R^2 for 3 predictors (tv+radio+newspaper) = 0.8972106

Only small increase

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

	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

CODING SCHEME

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

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

INTRODUCTION TO CLASSIFICATION

Simple Linear Regression → Multi Linear Regression → Coding Scheme for Categorical Variables → **Introduction to Classification** → Logistic Regression

WHAT IS CLASSIFICATION?

No.	1: outlook Nominal	2: temperature Nominal	3: humidity Nominal	4: windy Nominal	5: **play** Nominal
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

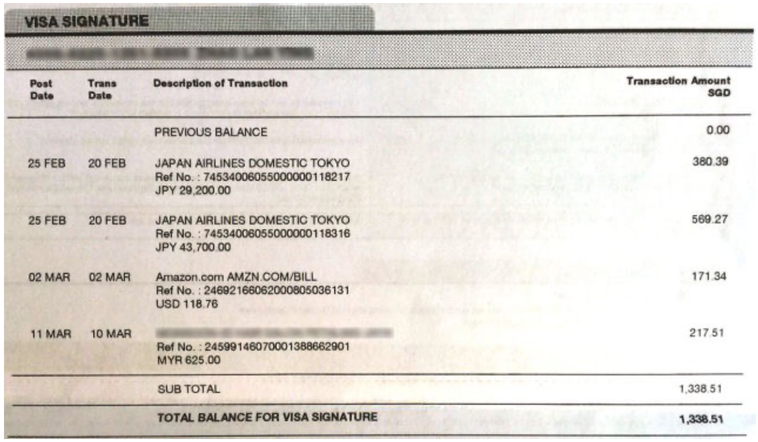
Based on a set of predictors, decide whether to play?

How is it different from Regression?

CREDIT CARD DEFAULT EXAMPLE

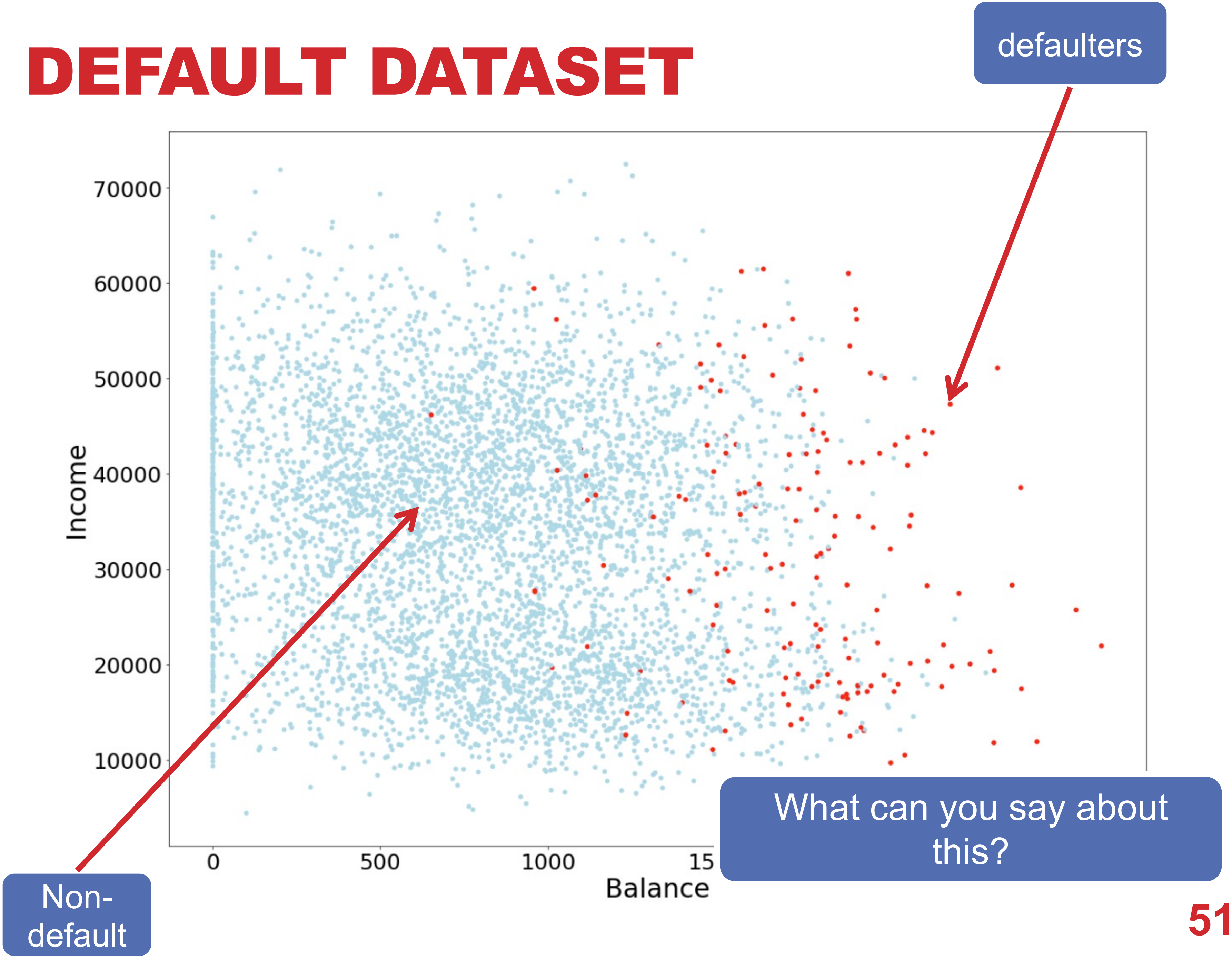
Predict whether an individual will <u>default</u> on his/her credit card payment

- **Income** and **Balance** = 2 predictor variables
  (Similar to the regression case)
- **Default** = response (2 possible categories: Yes or No)

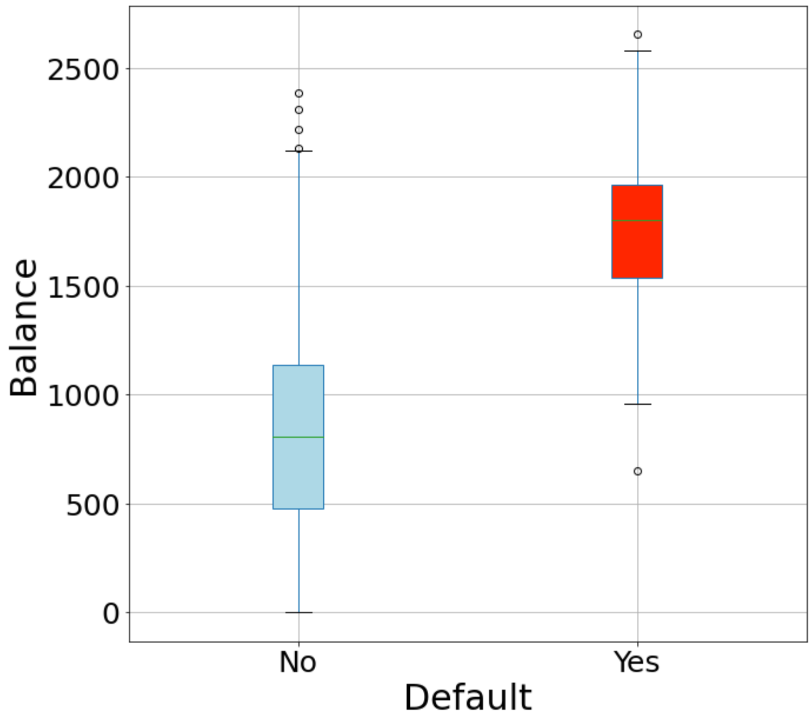

VISA SIGNATURE

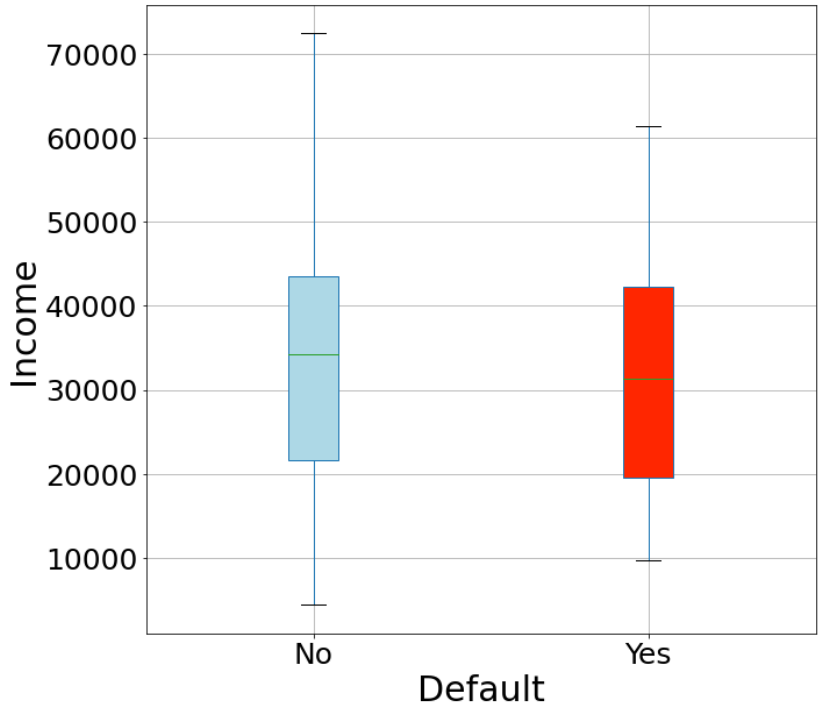
Post Date	Trans Date	Description of Transaction	Transaction Amount SGD
		PREVIOUS BALANCE	0.00
25 FEB	20 FEB	JAPAN AIRLINES DOMESTIC TOKYO Ref No. : 74534006055000000118217 JPY 29,200.00	380.39
25 FEB	20 FEB	JAPAN AIRLINES DOMESTIC TOKYO Ref No. : 74534006055000000118316 JPY 43,700.00	569.27
02 MAR	02 MAR	Amazon.com AMZN.COM/BILL Ref No. : 24692166062000805036131 USD 118.76	171.34
11 MAR	10 MAR	Ref No. : 24599146070001388662901 MYR 625.00	217.51
		SUB TOTAL	1,338.51
		TOTAL BALANCE FOR VISA SIGNATURE	**1,338.51**

DEFAULT DATASET



DEFAULT DATASET





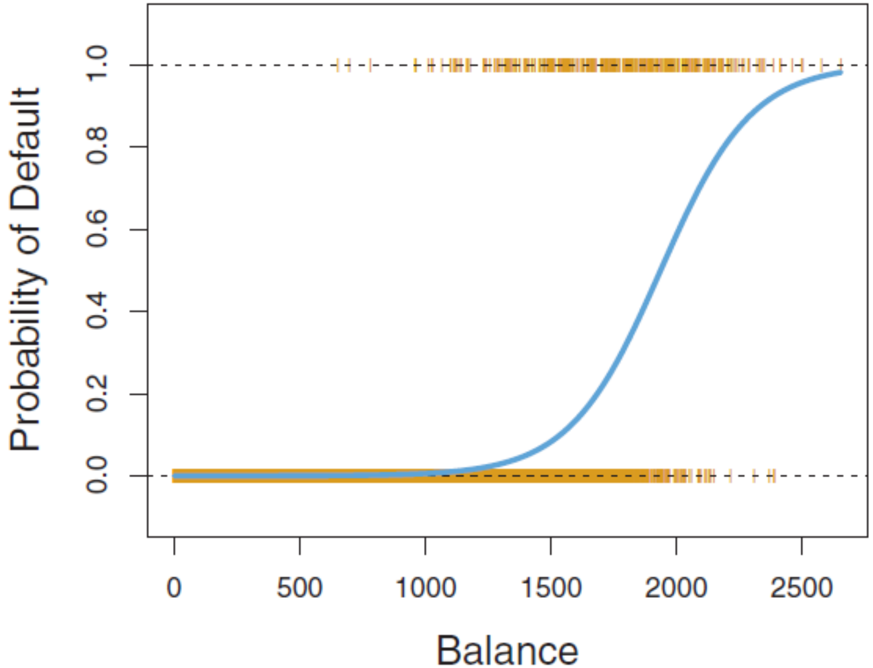
What can you say about this?

LOGISTIC REGRESSION

Simple Linear Regression → Multi Linear Regression → Coding Scheme for Categorical Variables → Introduction to Classification → **Logistic Regression**

LOGISTIC REGRESSION

Unlike linear regression which finds the value of the response (Y) directly, logistic regression finds the <u>probability that Y</u> belongs to a particular category



MODELING BY PROBABILITY OF Y

Consider the Default example, we model the problem as such:

- Denote $Pr(default = Yes \mid balance)$ as $p(\text{balance})$
- We might predict default = Yes for any individual using:

$$p(\text{balance}) > 0.5$$

- Or if we are more conservative, we can use a lower threshold:

$$p(\text{balance}) > 0.1$$

MODELING BY PROBABILITY OF Y

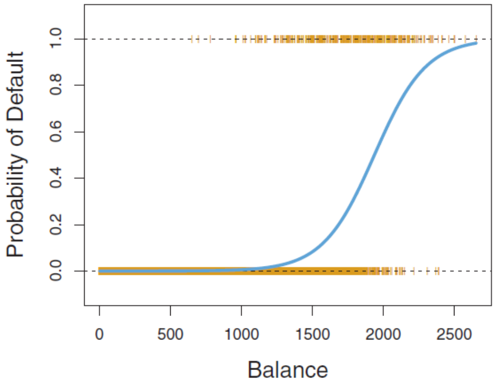
We can generalize the equations to :

$$\Pr(Y = 1 \mid X) = p(X)$$

Response taking a specific category

Predictor

LOGISTIC REGRESSION



The logistic function is given as follows:

$$\Pr(Y=1|X)=p(X)=\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

$$\frac{p(X)}{1-p(X)}=e^{\beta_0+\beta_1 X}$$

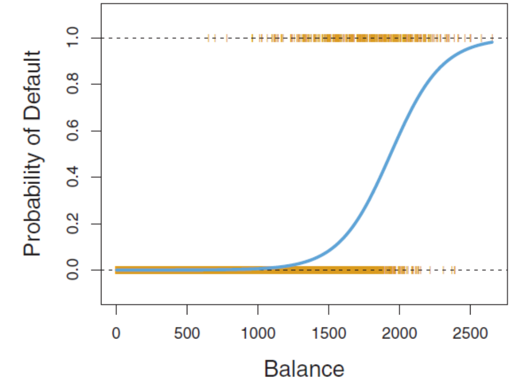
odds

log-odds
or
logit

$$\log\left(\frac{p(X)}{1-p(X)}\right)=\beta_0+\beta_1 X$$

Logistic Regression has a logit that is linear in X

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$



INTERPRETING B$_1$

In linear regression:

- β_1 : average change in Y for 1 unit increase in X

In logistic regression:

- increasing X by 1 unit changes the log-odds by β_1

- If β_1 is positive: $X \uparrow \quad p(X)\uparrow, \qquad X \downarrow \quad p(X)\downarrow$
- If β_1 is negative: $X \uparrow \quad p(X)\downarrow, \qquad X \downarrow \quad p(X)\uparrow$

log-odds
or
logit

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X$$

ESTIMATING THE REGRESSION COEFFICIENTS

The next step is then to estimate β_0 and β_1 using the training data

Intuition: The probability of the response being 1 is either $p(x_i)$ if $y_i = 1$, or $(1 - p(x_i))$ if $y_i = 0$

The likelihood function can then be formulated as:

$$\boxed{\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'}))}$$

- Choose values of β_0 and β_1 to maximize this likelihood function

DEFAULT DATASET EXAMPLE

For Default dataset, using <u>balance</u> as the predictor variable

The β_0 and β_1 coefficients estimates are as follows:

	coef	std err	z	P>|z|	[0.025	0.975]
const	-10.6513	0.361	-29.491	0.000	-11.359	-9.943
balance	0.0055	0.000	24.952	0.000	0.005	0.006

$B_1 > 0 \rightarrow$ increase in balance is associated with an increase in the probability of <u>default</u>

$B_1 = 0.0055 \rightarrow$ 1 unit increase in balance is associated with an increase in log odds of default by 0.0055 units

Coefficients are significant

MAKING PREDICTIONS

Suppose a person has a balance of \$1000, the probability of default is:

$$\hat{p}(balance) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 balance}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 balance}} = \frac{e^{-10.6513 + 0.0055 \times 1000}}{1 + e^{-10.6513 + 0.0055 \times 1000}} = 0.00576$$

- The predicted probability of default for an individual with a balance of \$1000 is < 1%
- On the other hand, the predicted probability of default for an individual with a balance of \$2000 = 0.586 (or 58.6%)

$$\hat{p}(student) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 student}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 student}}$$

QUALITATIVE PREDICTORS IN LOGISTIC REGRESSION

Similar to the linear regression case, we can also use dummy variables for incorporating categorical variables

- For example, the Default dataset contains a categorical variable (student)
- We can create a dummy variable with values:
  1 – student, 0 - non-student

	Coefficient	Std. error	Z-statistic	P-value
Intercept	−3.5041	0.0707	−49.55	<0.0001
student[Yes]	0.4049	0.1150	3.52	0.0004

$$\widehat{\Pr}(\texttt{default=Yes}|\texttt{student=Yes}) = \frac{e^{-3.5041+0.4049\times 1}}{1+e^{-3.5041+0.4049\times 1}} = 0.0431$$

$$\widehat{\Pr}(\texttt{default=Yes}|\texttt{student=No}) = \frac{e^{-3.5041+0.4049\times 0}}{1+e^{-3.5041+0.4049\times 0}} = 0.0292$$

MULTIPLE LOGISTIC REGRESSION

Similar to linear regression, we can generalize the model for multiple predictors:

$$p(X) = \frac{e^{\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X + \ldots + \beta_p X_p}}$$

- where $X = (X_1, \ldots, X_p)$
- the coefficients $(\beta_0, \beta_1, \ldots, \beta_p)$ are also estimated using the maximum likelihood method

$$p(X)=\frac{e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X+...+\beta_p X_p}}$$

$$p(X)=\frac{e^{-10.8690+0.0057 balance+0.0030 income-0.6468 student}}{1+e^{-10.8690+0.0057 balance+0.0030 income-0.6468 student}}$$

DEFAULT DATASET EXAMPLE

Using 3 predictor variables:

- balance (quantitative)
- income (quantitative)
- student status (qualitative)

	coef	std err	z	P>|z|	[0.025	0.975]
const	-10.8690	0.492	-22.079	0.000	-11.834	-9.904
balance	0.0057	0.000	24.737	0.000	0.005	0.006
income	3.033e-06	8.2e-06	0.370	0.712	-1.3e-05	1.91e-05
studentYes	-0.6468	0.236	-2.738	0.006	-1.110	-0.184

$$\hat{p}(X) = \frac{e^{-10.8690 + 0.0057\,balance + 0.0030\,income - 0.6468\,student}}{1 + e^{-10.8690 + 0.0057\,balance + 0.0030\,income - 0.6468\,student}}$$

MAKING PREDICTIONS

Suppose a student has a balance of $1500 and an income of $40000, the probability of default is:

$$\hat{p}(X) = \frac{e^{-10.869 + 0.00574 \times 1{,}500 + 0.003 \times 40 - 0.6468 \times 1}}{1 + e^{-10.869 + 0.00574 \times 1{,}500 + 0.003 \times 40 - 0.6468 \times 1}} = 0.058$$

WHAT'S NEXT?

Predictive Analytics II