MGS 3701 / MKT 3950: Data Mining

Chapter6: Multiple Linear Regression

Chungil Chae

温州肯恩大学
WENZHOU-KEAN UNIVERSITY

2022/01/01 (updated: 2022-05-09)

Chapter Objectives

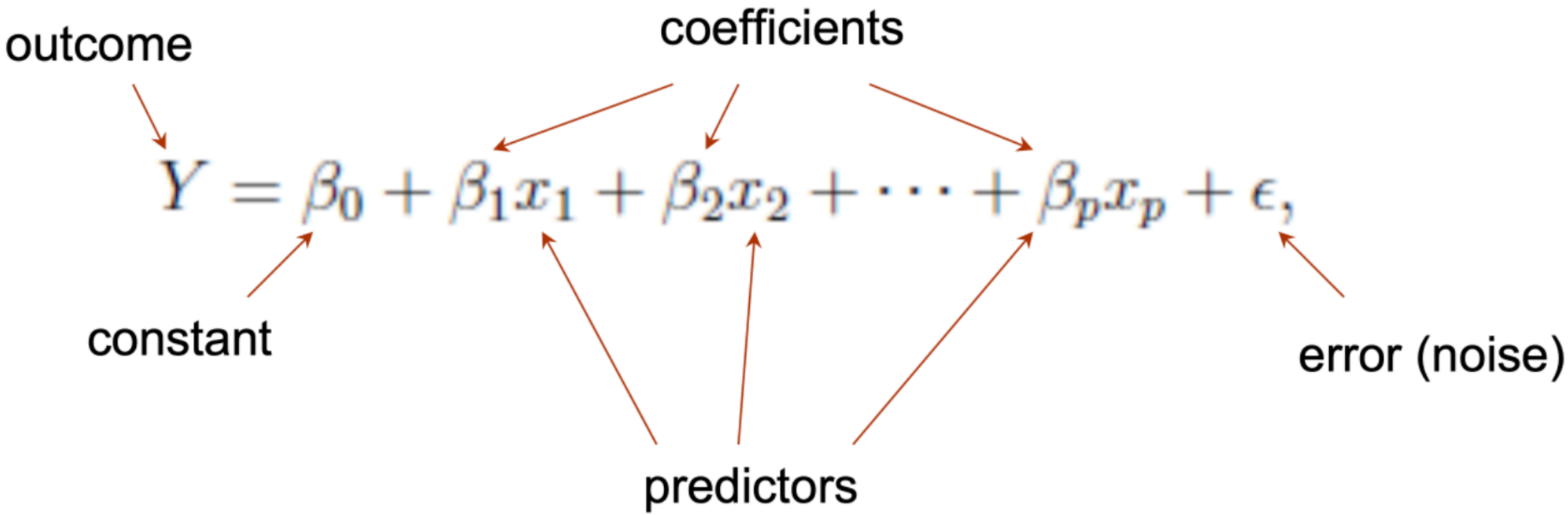
In this chapter

- Linear regression models for the purpose of prediction.
- The differences between fitting and using regression models for the purpose of inference (as in classical statistics) and for prediction.
- A predictive goal calls for evaluating model performance on a validation set, and for using predictive metrics.
- The challenges of using many predictors and describe variable selection algorithms .

Introduction

- This model is used to fit a relationship between
    - a numerical outcome variable Y (also called the response, target, or dependent variable) and
    - a set of predictors X1,X2,...,Xp (also referred to as independent variables, input variables, regressors, or covariates)
- The assumption is that the following function approximates the relationship between the predictors and outcome variable:

$$\boldsymbol{Y} = \beta_0 + \beta_1 \mathbf{1} \boldsymbol{x}_1 + \beta_2 \boldsymbol{x}_2 + \cdots + \beta_p \boldsymbol{x}_p + \varepsilon$$



- $\beta 0, \ldots, \beta p$ are coefficients
- ε is the noise or unexplained part

- Regression modeling means not only **estimating the coefficients** but also **choosing which predictors** to include and in what form.
  - For example, a numerical predictor can be
    - included as is, or
    - in logarithmic form [log(X)], or
    - in a binned form (e.g., age group).
  - Choosing the right form depends on
    - domain knowledge, data availability, and needed predictive power.

> Multiple linear regression is applicable to **numerous predictive modeling** situations.

Examples are predicting:

- Customer activity on credit cards from their demographics and historical activity patterns
- Predicting expenditures on vacation travel based on historical frequent flyer data
- Predicting staffing requirements at help desks based on historical data and product and sales information
- Predicting sales from cross-selling of products from historical information
- Predicting the impact of discounts on sales in retail outlets

Explanatory vs. Predictive Modeling

The two popular but different objectives behind fitting a regression model are:

1. Explaining or quantifying the average effect of inputs on an outcome (explanatory or descriptive task, respectively)
2. Predicting the outcome value for new records, given their input values (predictive task)

The classical statistical approach is focused on the first objective.

- In that scenario, the data are treated as a random sample from a larger population of interest.
- The regression model estimated from this sample is an attempt to capture the average relationship in the larger population.

In predictive analytics, the focus is predicting new individual records

- Here we are not interested in the coefficients themselves, nor in the “average record,” but rather in the predictions that this model can generate for new records.
- In this scenario, the model is used for micro-decision-making at the record level.

The modeling steps and performance assessment differ in the two cases, usually leading to different final models.

> - Therefore, the choice of model is closely tied to whether the goal is explanatory or predictive.
> - In explanatory and descriptive modeling, where the focus is on modeling the average record
> - we try to fit the best model to the data in an attempt to learn about the underlying relationship in the population.

In predictive modeling (data mining), the goal is to find a regression model that best predicts new individual records.

- A regression model that fits the existing data too well is not likely to perform well with new data.
- Hence, we look for a model that has the highest predictive power by evaluating it on a holdout set and using predictive metrics (see Chapter 5).

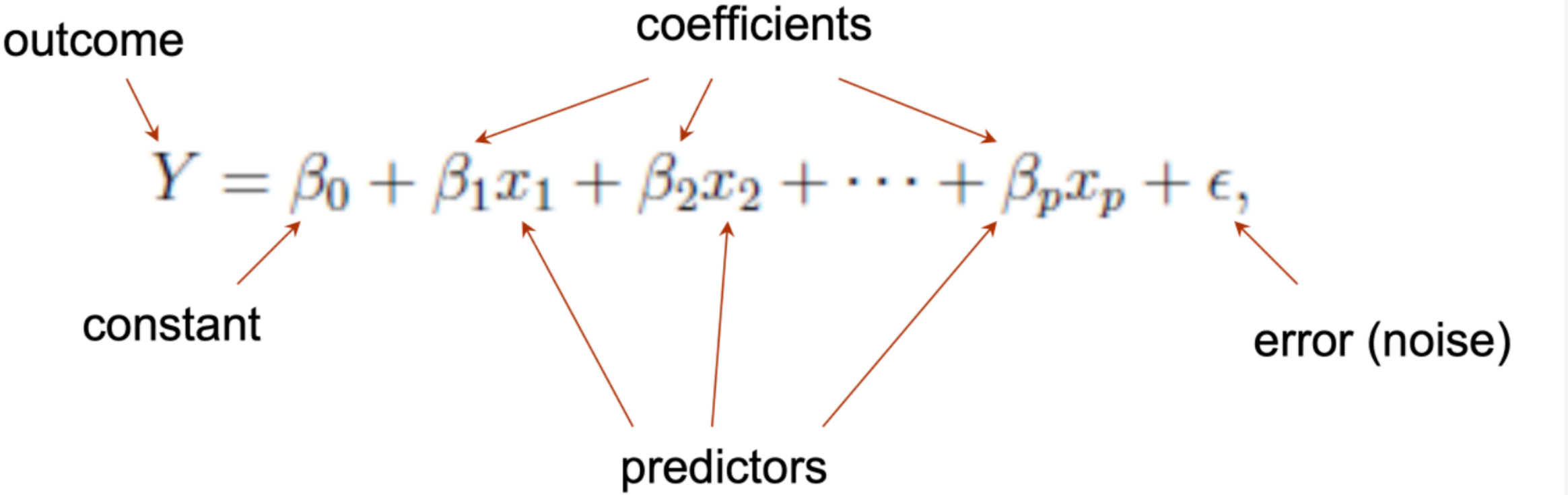
The main differences in using a linear regression in the two scenarios:

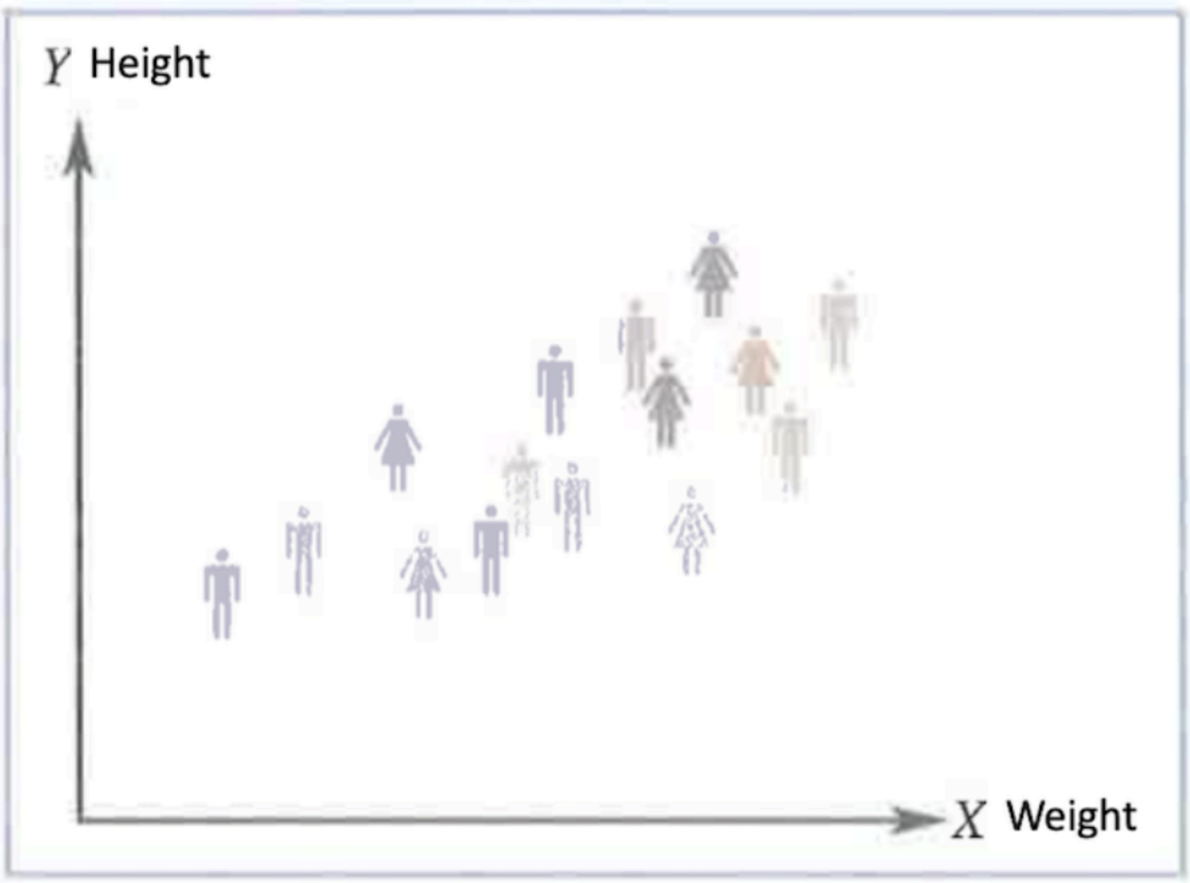
1. Fit
    - A good explanatory model is one that fits the data closely,
    - A good predictive model is one that predicts new records accurately.
2. Dataset
    - In explanatory models, the entire dataset is used for estimating the best-fit model
    - When the goal is to predict outcomes of new individual records, the data are typically split into a training set and a validation set. The training set is used to estimate the model, and the validation or holdout set is used to assess this model's predictive performance on new, unobserved data.
3. Performance measure
    - Performance measures for explanatory models measure how close the data fit the model (how well the model approximates the data) and how strong the average relationship is,
    - In predictive models performance is measured by predictive accuracy (how well the model predicts new individual records).
4. Focus
    - In explanatory models the focus is on the coefficients (β),
    - In predictive models the focus is on the predictions (y).

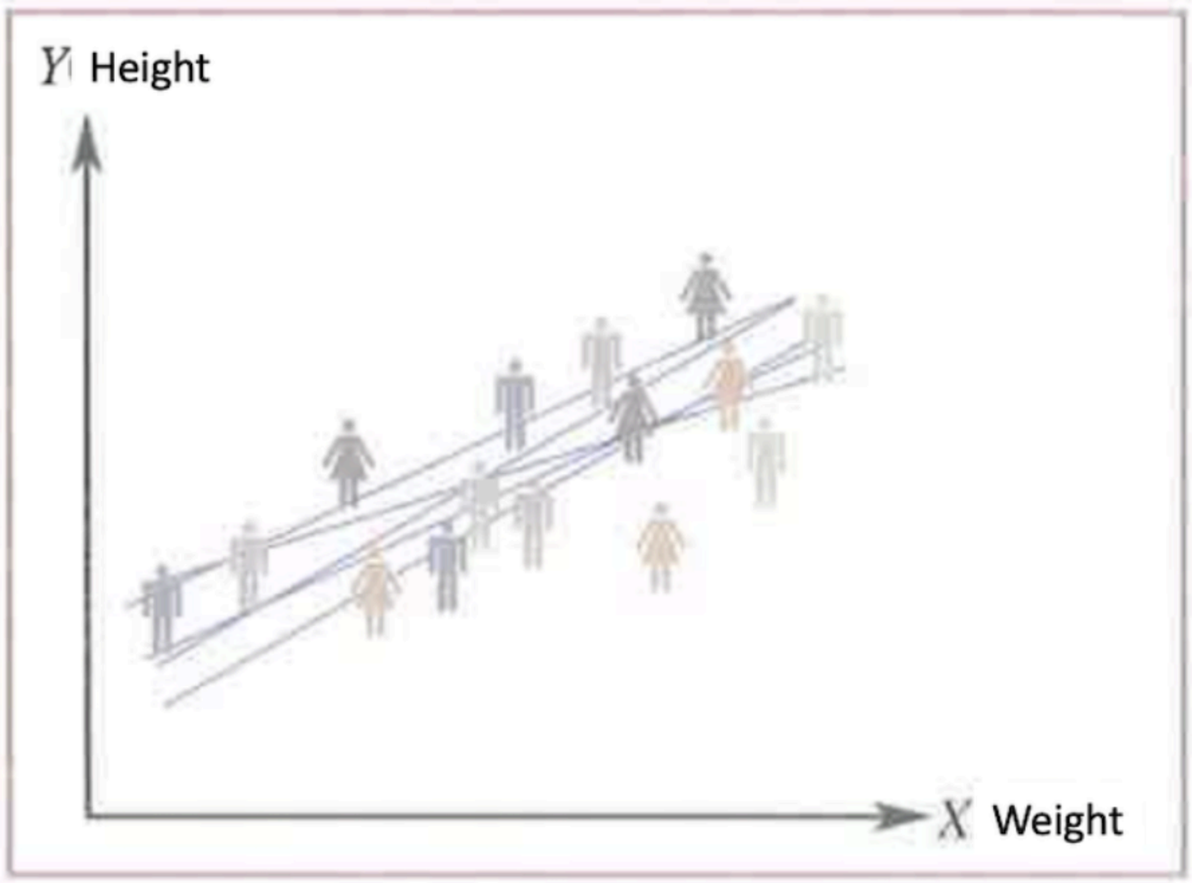
It is extremely important to know the goal of the analysis before beginning the modeling process.

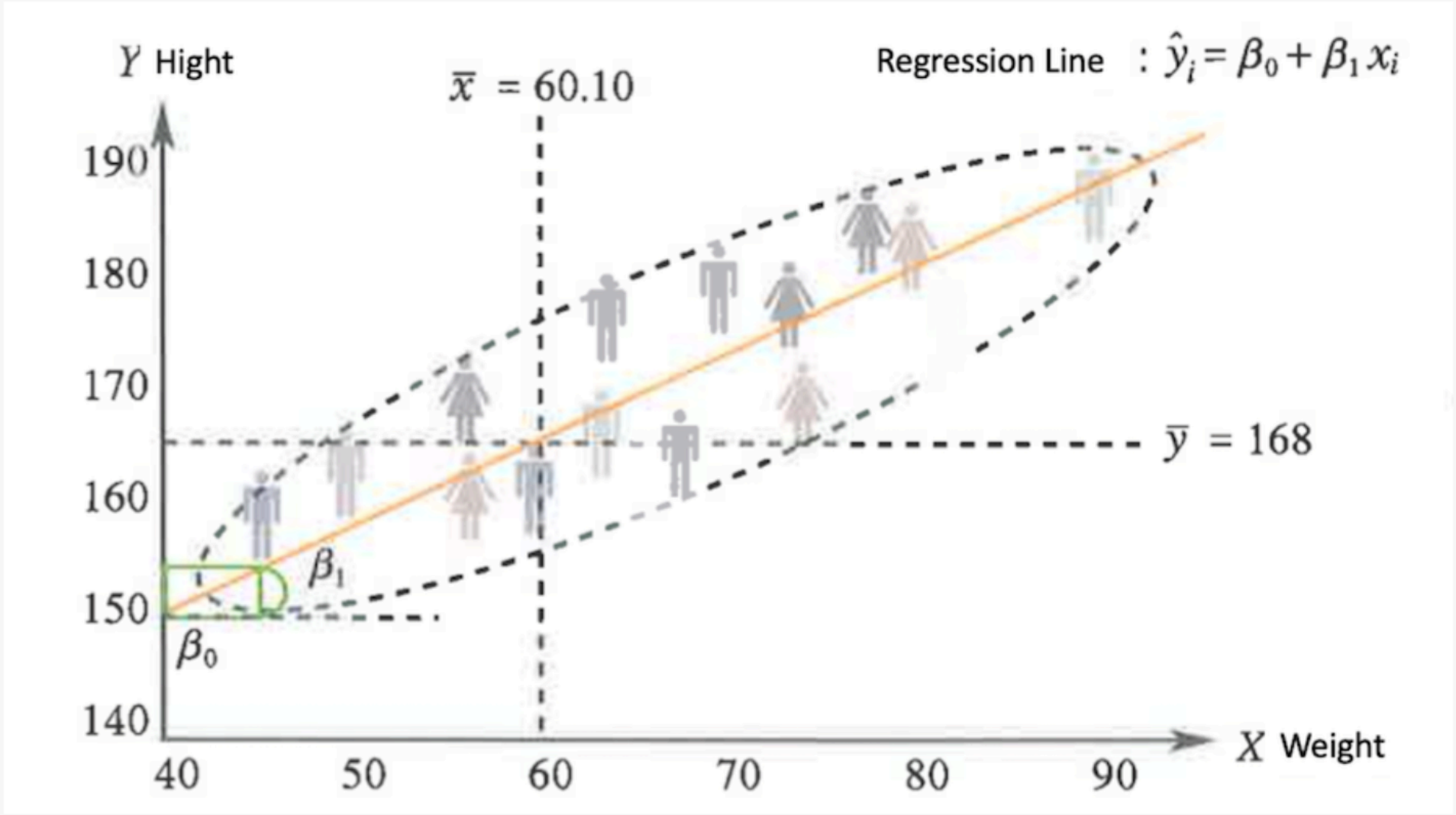
- A good predictive model can have a looser fit to the data on which it is based
- A good explanatory model can have low prediction accuracy

Estimating the Regression Equation and Prediction




Y Height
X Weight

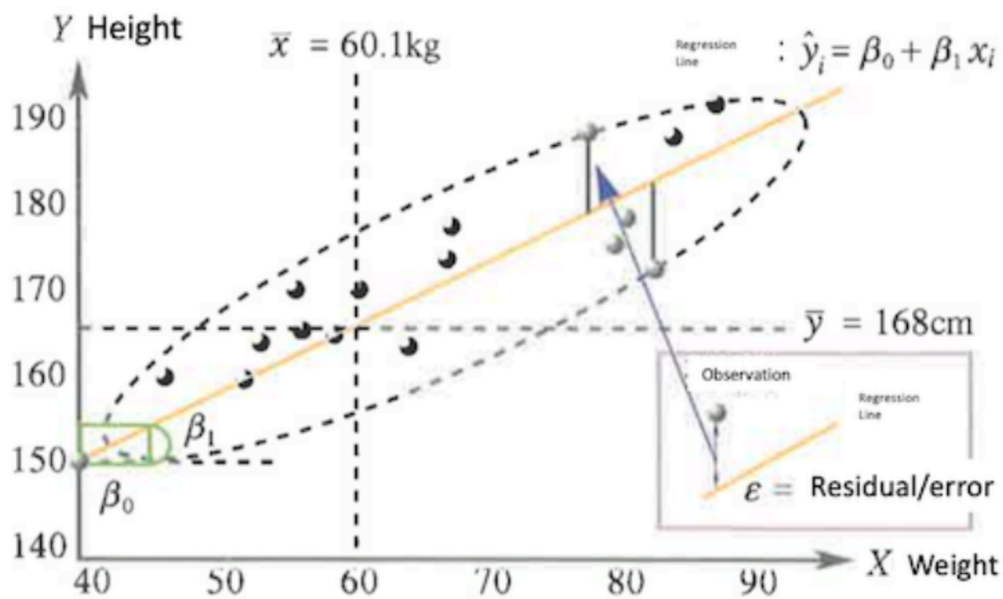

Y Height
X Weight


Y Hight
Regression Line : $\hat{y}_i = \beta_0 + \beta_1 x_i$
$\overline{x}$ = 60.10
190
180
170
160
150
140
$\overline{y}$ = 168
β_1
β_0
40
50
60
70
80
90
X Weight

Data

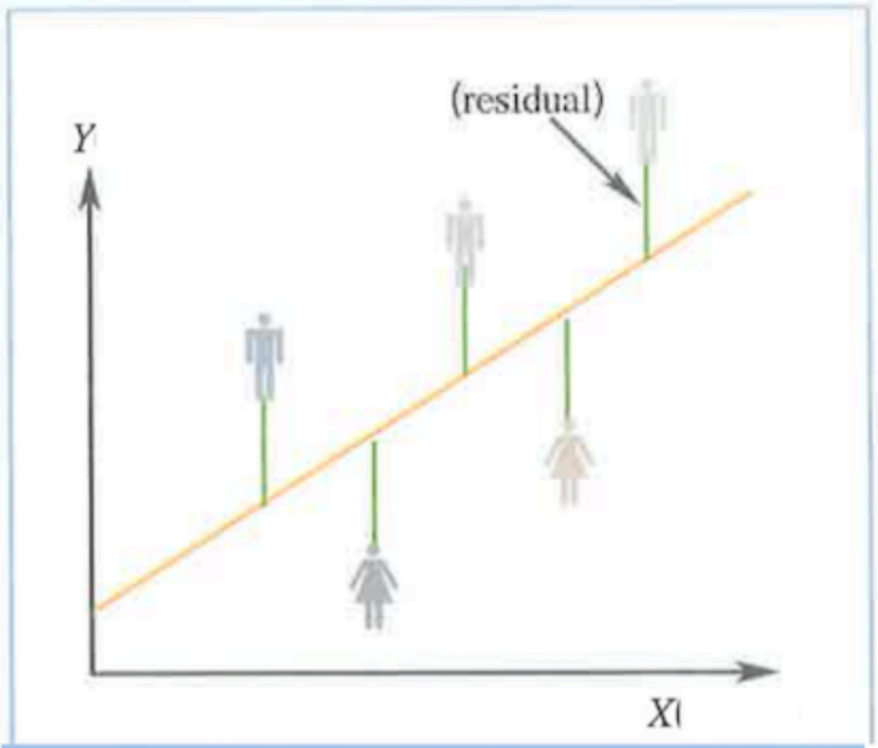
Participants	Variable X	Variable Y
	Weight	Height
1	72	176
2	72	172
3	70	182
4	43	160
5	48	163
6	54	165
7	51	168
8	52	163
9	73	182
10	45	148
11	60	170
12	62	166
13	64	172
14	47	160
15	51	163
16	74	170
17	88	182
18	64	174
19	56	164
20	56	160
Mean	60.1	168

Graph

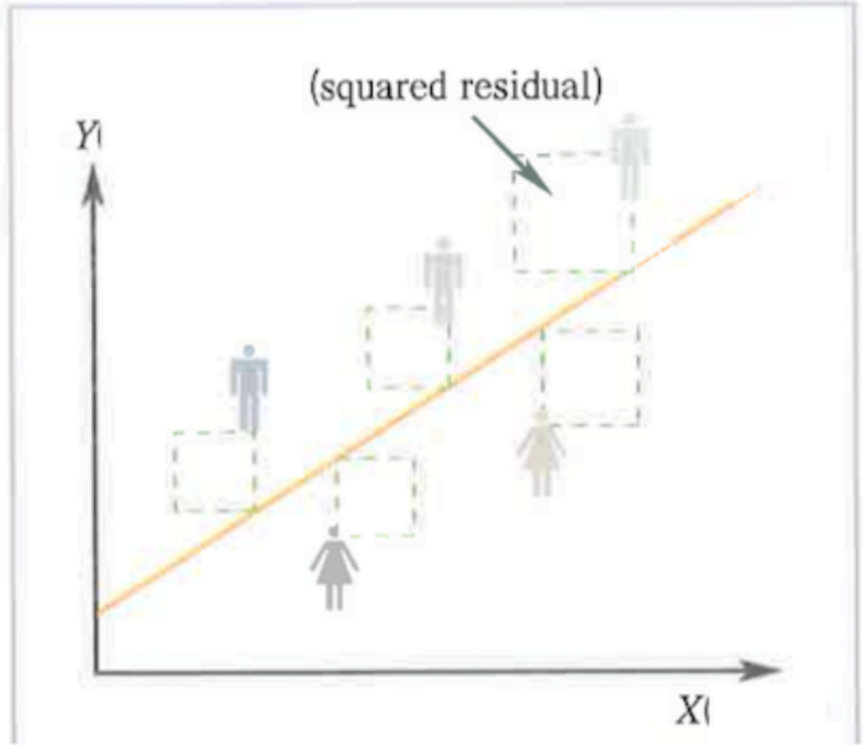


Propose

- Estimating regression coefficient using data
- Predict height using regression equation
- Estimated regression line pass through mean of two variables (weight & height)



$$Min \sum_{i=1}^{n} |\varepsilon_i| = Min \sum_{i=1}^{n} |y_i - \hat{y}_i|$$



$$Min \sum_{i=1}^{n} \varepsilon_i^2 = Min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$
$$= Min \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

Sum of squared residuals (SSR)

$$E = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

Minimizing sum of squared residuals (SSR)

$$Min\,[E] = Min\left[\sum_{i=1}^{n} \varepsilon_i^2\right] = Min\left[\sum_{i=1}^{n} (y_i - \hat{y}_i)^2\right] = Min\left[\sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2\right]$$

Partial derivative of intercept and slope

$$\frac{\partial E}{\partial \beta_0} = 2\sum_{i=1}^{n} (-1)[y_i - (\beta_0 + \beta_1 x_i)] = 0$$

$$\frac{\partial E}{\partial \beta_1} = 2\sum_{i=1}^{n} (-x_i)[y_i - (\beta_0 + \beta_1 x_i)] = 0$$

Estimating intercept and slope

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

$$\beta_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

ID	X	Y	X deviation $(x_i-\bar{x})$	Y deviation $(y_i-\bar{y})$	Product of deviation $(x_i-\bar{x})(y_i-\bar{y})$	Deviation squared $(x_i-\bar{x})^2$
1	72	176	11.9	8.0	95.2	141.6
2	72	172	11.9	4.0	47.6	141.6
3	70	182	9.9	14.0	138.6	98.0
4	43	160	−17.1	−8.0	136.8	292.4
5	48	163	−12.1	−5.0	60.5	146.4
6	54	165	−6.1	−3.0	18.3	37.2
7	51	168	−9.1	0.0	0.0	82.8
8	52	163	−8.1	−5.0	40.5	65.6
9	73	182	12.9	14.0	180.6	166.4
10	45	148	−15.1	−20.0	302.0	228.0
11	60	170	−0.1	2.0	−0.2	0.0
12	62	166	1.9	−2.0	−3.8	3.6
13	64	172	3.9	4.0	15.6	15.2
14	47	160	−13.1	−8.0	104.8	171.6
15	51	163	−9.1	−5.0	45.5	82.8
16	74	170	13.9	2.0	27.8	193.2
17	88	182	27.9	14.0	390.6	778.4
18	64	174	3.9	6.0	23.4	15.2
19	56	164	−4.1	−4.0	16.4	16.8
20	56	160	−4.1	−8.0	32.8	16.8
Sum	1202	3360	0	0	1673	2693.8
Mean	60.1	168				

$\bar{x} = 60.10 \quad \bar{y} = 168$

$\Sigma(x_i-\bar{x})(y_i-\bar{y}) = 1673$

$\Sigma(x_i-\bar{x})^2 = 2693.80$

- Slope

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{1673}{2693.8} = 0.621$$

- Intercept

$\beta_0 = \bar{y} - \beta_1 \bar{x} = 168 - (0.621 \times 60.1)$
$= 130.678$

Usages

- equation $\hat{y}_i = 130.678 + 0.621x_i$
- The height of a student weighing 67 kg can be estimated to be 172.258

$\hat{y}_i = 130.678 + 0.621 \times 67 = 172.285\text{cm}$

Assumtions

Predictions based on this equation are the best predictions possible in the sense that they will be unbiased (equal to the true values on average) and will have the smallest mean squared error compared to any unbiased estimates if we make the following assumptions:

1. The noise ε (or equivalently, Y) follows a normal distribution.
2. The choice of predictors and their form is correct (linearity).
3. The records are independent of each other.
4. The variability in the outcome values for a given set of predictors is the same regardless of the values of the predictors (homoskedasticity).

Variable Selection in Linear Regression

- The last two points mean that there is a trade-off between too few and too many predictors.
- In general, accepting some bias can reduce the variance in predictions.
- This bias–variance trade-off is particularly important for large numbers of predictors, because in that case, it is very likely that there are variables in the model that have small coefficients relative to the standard deviation of the noise and also exhibit at least moderate correlation with other variables.
- Dropping such variables will improve the predictions, as it reduces the prediction variance.
- This type of bias–variance trade-off is a basic aspect of most data mining procedures for prediction and classification.
- In light of this, methods for reducing the number of predictors p to a smaller set are often used.

Reducing the Number of Predictors

- **Goal**: Find parsimonious model (the simplest model that performs sufficiently well)
  - More robust
  - Higher predictive accuracy

Exhaustive Search

- Exhastive search = "best subset"
  - All possible subsets of predictors assessed (single, pairs, triplets, etc)
  - Computationally intensive not feasible for big data
  - Judge by "adjusted R square"
  - Use regsubsets() in package leaps

$$R^2_{adj} = 1 - \frac{n-1}{n-p-1}(1-R^2)$$

Popular Subset Selection Algorithms

- Partial search algorithms
  - Forward
  - Backward
  - Stepwise

Forward selection

- Start with no predictors
- Add them one by one (add the one with largest contribution)
- Stop when the addition is not statistically significant

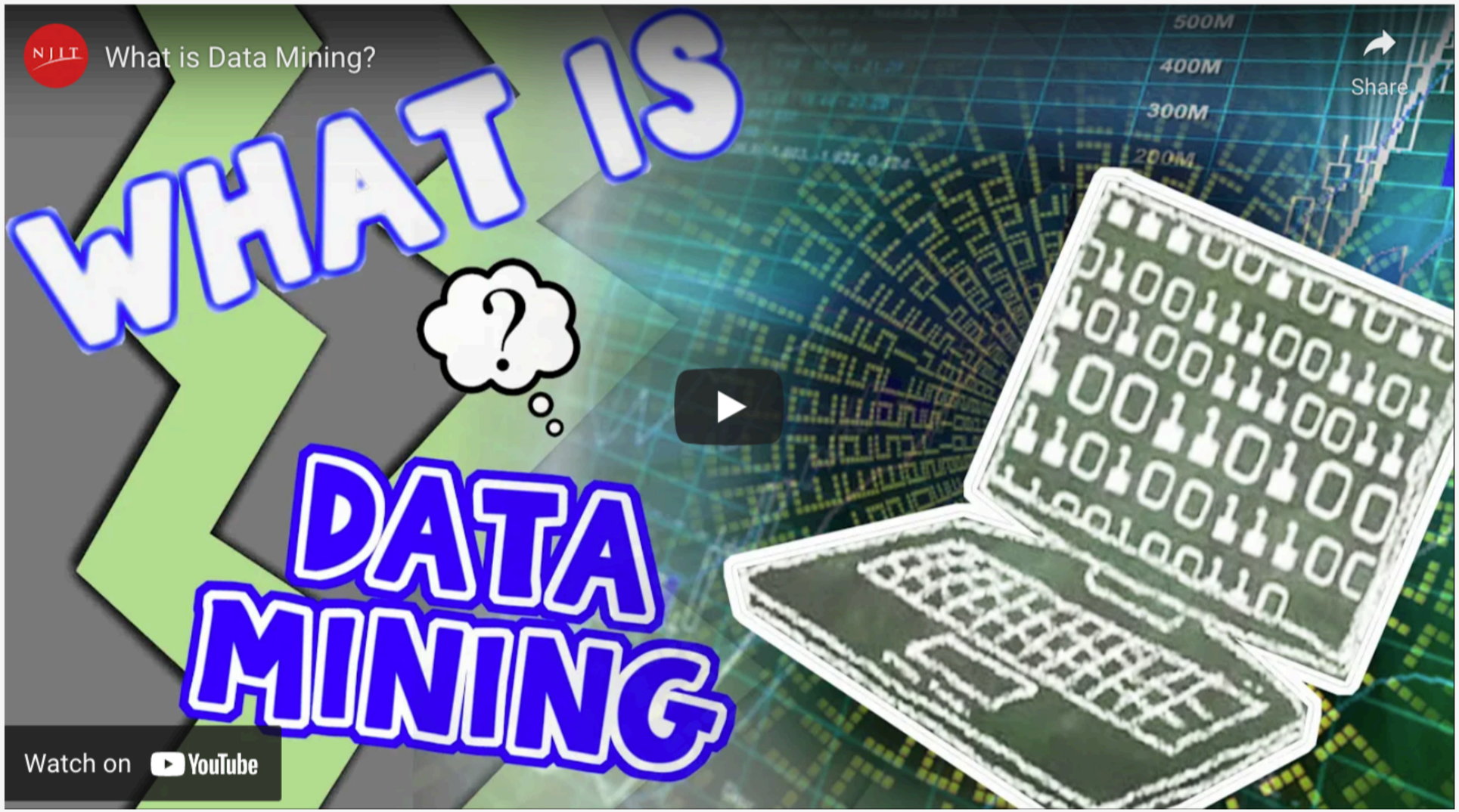
Backward Elimination

- Start with all predictors
- Successively eliminate least useful predictors one by one
- Stop when all remaining predictors have statistically significant contribution

Stepwise

- Like forward selection
- Except at each step, also consider dropping non-significant predictors

Chapter Video



23 / 25 00:59

References

Shmueli, G., P. C. Bruce, P. Gedeck, and N. R. Patel (2019). Data mining for business analytics: concepts, techniques and applications in Python. John Wiley & Sons.

Shmueli, G., P. C. Bruce, I. Yahav, N. R. Patel, and K. C. Lichtendahl Jr (2017). Data mining for business analytics: concepts, techniques, and applications in R. John Wiley & Sons.

Assignment

Class Plan & Final Exam

- Class plan
  - May 10 - 13: Chapter6 - Regression
  - May 16 - 20: Chapter20 - Text Mining
- Range: Chapter 5, 6, 20
- May 24 (MKT)
  - (MKT Grouop1), 5:30 - 6:10
  - (MKT Grouop2), 6:15 - 6:45
- May 25 (MGS)
  - (MGS Grouop1), 2:30 - 3:10
  - (MGS Grouop2), 3:15 - 4:45
- 1 learning sheet (2 page, front and back, hand-writing only)
- No class on 26, 27, 31 (Final week)

Final Project (Project Proposal)

- Due date: May 27 (Friday 23:55pm, Beijing Time)
- To do:
  - Individual Assignment
  - 5 page (up to 7 page, no more than 7 page) project plan
  - Double space, 11 point, Times new roman
  - No reference page
  - Project proposal should includes your project plan guide by data mining process (chapter2)
  - PDF only (both R markdown generated or MS WORD)
  - Including your name and student ID and title
  - Submit to LMS (Submit button will be disappear 5 mins before midnight)