
Data Analytics

Yong Zheng

Illinois Institute of Technology
Chicago, IL, 60616, USA



School of Applied Technology
ILLINOIS INSTITUTE OF TECHNOLOGY

Midterm Exam

- Live Students take exams in the class.
- Remote students take exams in India
- Online students take exams at your selected locations when you enroll in online sections



Midterm Exam

- Time: Mar 26, 08:30 AM – 09:50 AM
- Location: SB 111
- Closed Note, Closed Book, Closed Devices
- **You can bring a calculator.** You can NOT share calculator with others.
- The questions in the exam will be the similar ones in your assignments; but you do not need to produce the R outputs, the outputs will be given in the exam papers.



Questions in Midterm Exam

The type of questions (similar to your assignments)

- **Concepts**
Use your own language to explain something
- **Descriptive and Inferential Statistics**
Describe quantitative and qualitative data
- **Hypothesis Testing/ANOVA**
One sample vs two sample hypothesis testing
- **Multiple Linear Regressions**
Read the outputs and interpret results or models



Topics Covered in Midterm Exam

Basics in Data Analytics

- Statistics Basics
- Hypothesis Testing
- Linear Regressions by R
- ANOVA



data statistics, distributions and visualizations, hypothesis testing, statistical inference, predictive models

----- Midterm Exam -----

More Topics in Data Analytics

- ~~Using SAS~~
- ~~Classification Models~~
- ~~Extended topics~~



More predictive models

Statistic Basics



Statistical Basics

- **Data Types and Descriptive Statistics**
 - Especially, how to interpret visualizations
 - How to calculate descriptive statistics
- **Inferential Statistics**
 - Using sample to estimate population
 - Hypothesis testing
 - One sample
 - Two independent vs dependent samples
 - ANOVA

Describe Qualitative Data

- Describe qualitative data Numerically
 - By class frequency
 - By class relative frequency
- Describe qualitative data by visualizations
 - By bar graph
 - By pie chart

Describe Quantitative Data

- Describe quantitative data Numerically
 - By range, min, max, mean, median, mode, Q1, Q3
 - By variance, standard deviation
- Describe quantitative data by visualizations
 - By histogram
 - By box plot
 - By probability distribution

Describe Quantitative Data

- Population Statistics

μ = Population mean

σ^2 = Population variance

σ = Population standard deviation

$$\sigma^2 = E[(x_i - \mu)^2] = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n}$$

It will underestimate the value if the sample variance is divided by n.

Note: usually population mean & var are unknown

- Sample Statistics

\bar{x} = Sample mean

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

Sample Variance:

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

sample standard deviation.

$$s = \sqrt{s^2}$$

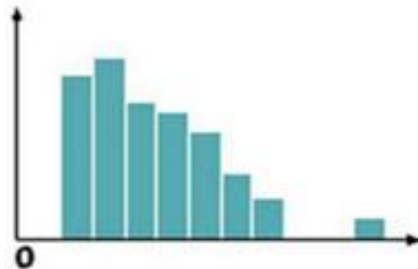
Describe Quantitative Data

- Describe quantitative data by visualizations

- By histogram

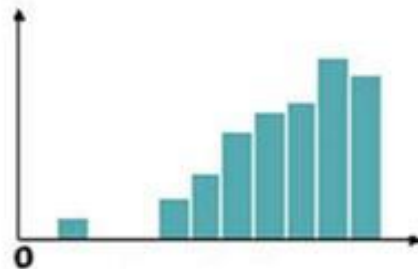
How to interpret histogram? (skewness and outlier)

Analyzing Shape:



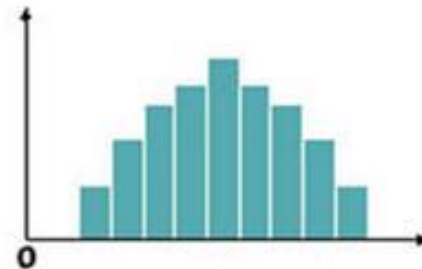
Positive Skew

Data is skewed to the right. The long tail of the data is on the right side of the peak.



Negative Skew

Data is skewed to the left. The long tail of the data is on the left side of the peak.

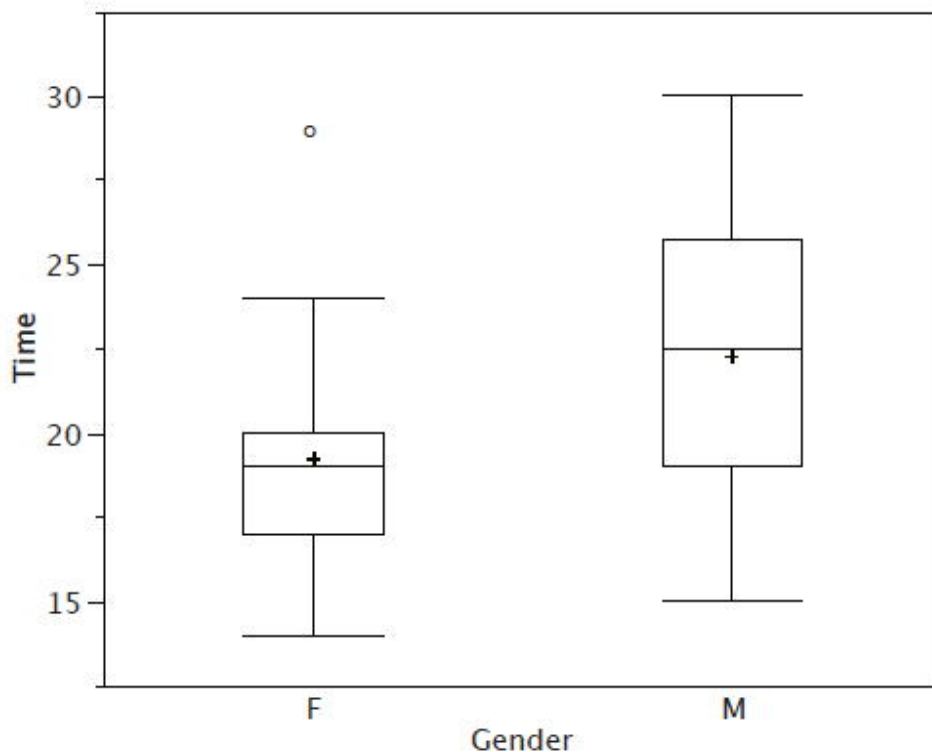


Normal Distribution

Data is not skewed to the right or left. The data is evenly distributed on both sides of the peak.

Describe Quantitative Data

- Describe quantitative data by visualizations
 - By box plot



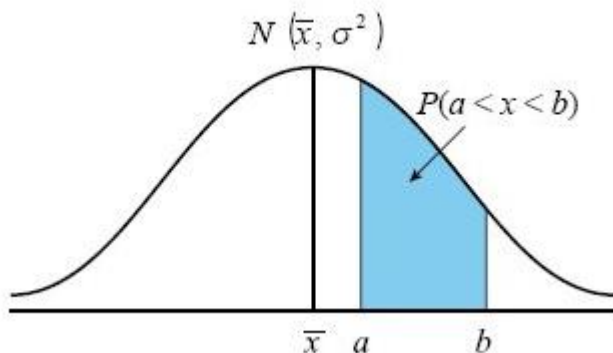
Normal Distribution

- Normal Probability Distribution

- ☐ It is a symmetric distribution.
- ☐ It is centered by **mean μ**
- ☐ Its spread is determined by **STD σ**

- Notes

- Variable X follows normal distribution, $X \sim N(\mu, \sigma^2)$



The normal curve area between a and b is the area under the normal distribution curve, and it is equal to the probability that x falls into the range $[a, b]$

Statistical Inference



Statistical Inference

- There are two ways for us to estimate or infer the population parameter, such as population mean:
 - 1) By estimating its value
For example: estimate the age of people in USA
 - 2) By testing hypothesis about its value
For example:
Method-1 is better than method 2.
Students in 527(04) are better than 527(01).
The average of working hours/day is no more than 8

Summary: Statistical Inference by Estimating Population Mean

- You can follow these steps
 - 1) Collect sample statistics, such as sample mean
 - 2) Sample is larger ($n \geq 30$), we assume sample mean follows normal distribution; otherwise, we assume it follows t distribution
 - 3) The standard error of the sampling distribution is expected to be as small as possible. Note: usually it becomes smaller if your n is larger
 - 4) Finally, make a conclusion by using statistical statements with confidence intervals
sample estimate \pm margin of error
margin of error = z value or t value \times standard error

Summary: Statistical Inference by Estimating Population Mean

- How to calculate z value or t value

1) If $n \geq 30$, normal distribution, z value

$$\bar{y} \pm z_{\alpha/2} \sigma_{\bar{y}} \approx \bar{y} \pm z_{\alpha/2} \left(\frac{s}{\sqrt{n}} \right), \alpha = 1 - \text{confidence level}$$

2) Otherwise, t distribution, t value

$$\bar{y} \pm t_{\alpha/2} s_{\bar{y}} = \bar{y} \pm t_{\alpha/2} \left(\frac{s}{\sqrt{n}} \right), \alpha = 1 - \text{confidence level}$$

Statistical Inference by Estimating Population Mean

- In general a confidence interval has the form:
sample estimate \pm margin of error

90% Confidence Interval	1.64	$\frac{sd}{\sqrt{n}}$	<div>Margin of error</div>
95% Confidence Interval	1.96	$\frac{sd}{\sqrt{n}}$	
99% Confidence Interval	2.57	$\frac{sd}{\sqrt{n}}$	

- How to calculate?
z value, TextBook section 1.8, Page 34

$$\bar{y} \pm z_{\alpha/2} \sigma_{\bar{y}} \approx \bar{y} \pm z_{\alpha/2} \left(\frac{s}{\sqrt{n}} \right), \alpha = 1 - \text{confidence level}$$

Elements in Hypothesis Testing

- Null Hypothesis, H_0

This is the hypothesis we have doubts

- Alternative Hypothesis, H_a

This is the hypothesis which is counter to the null hypothesis. Usually it is what we want to support

- Test Statistics

It is used to make the conclusions

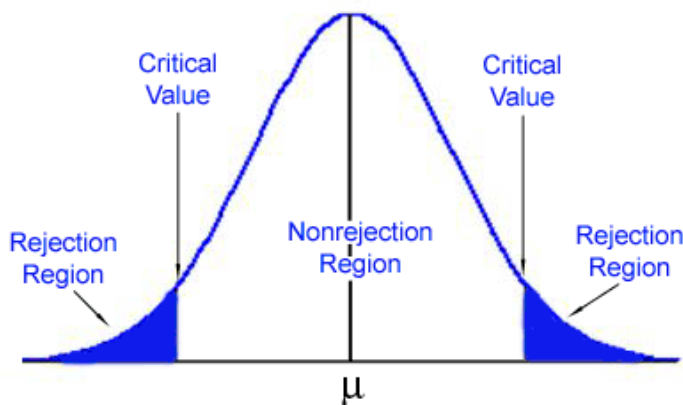
- Level of significance, α

The probability of rejecting H_0 giving H_0 is true



Elements in Hypothesis Testing

- Rejection Region



If our test statistics fall into rejection region, we reject null hypothesis and accept the alternative hypothesis.

- P-value

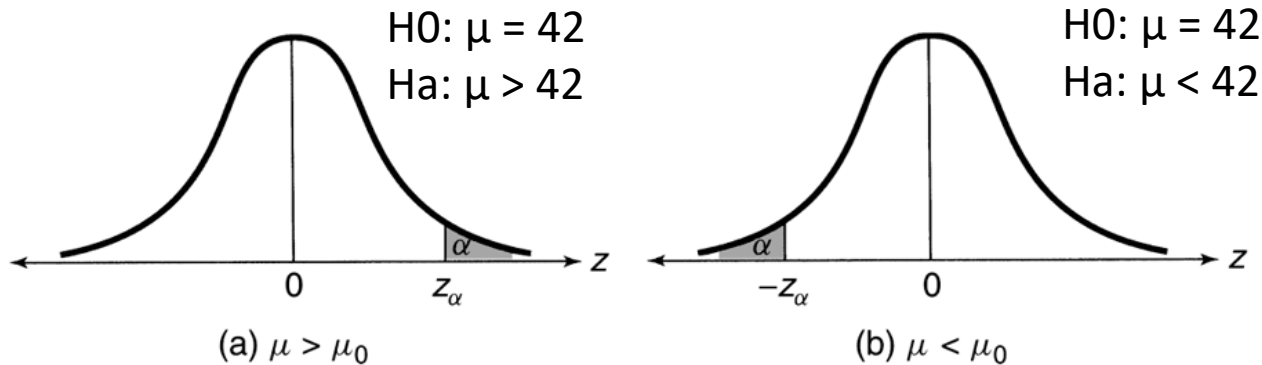
It is a probability value between 0 and 1 as evidence to reject the null hypothesis.

95% confidence level, we reject H_0 if $p\text{-value} < 0.05$

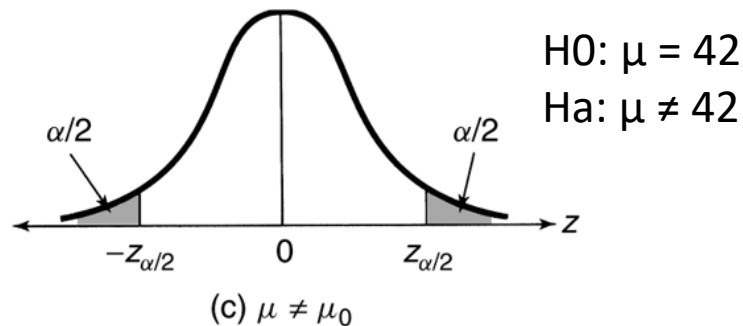
P-value = area under normal curve based on the test statistics

Elements in Hypothesis Testing

- One-sided or one-tailed statistical test



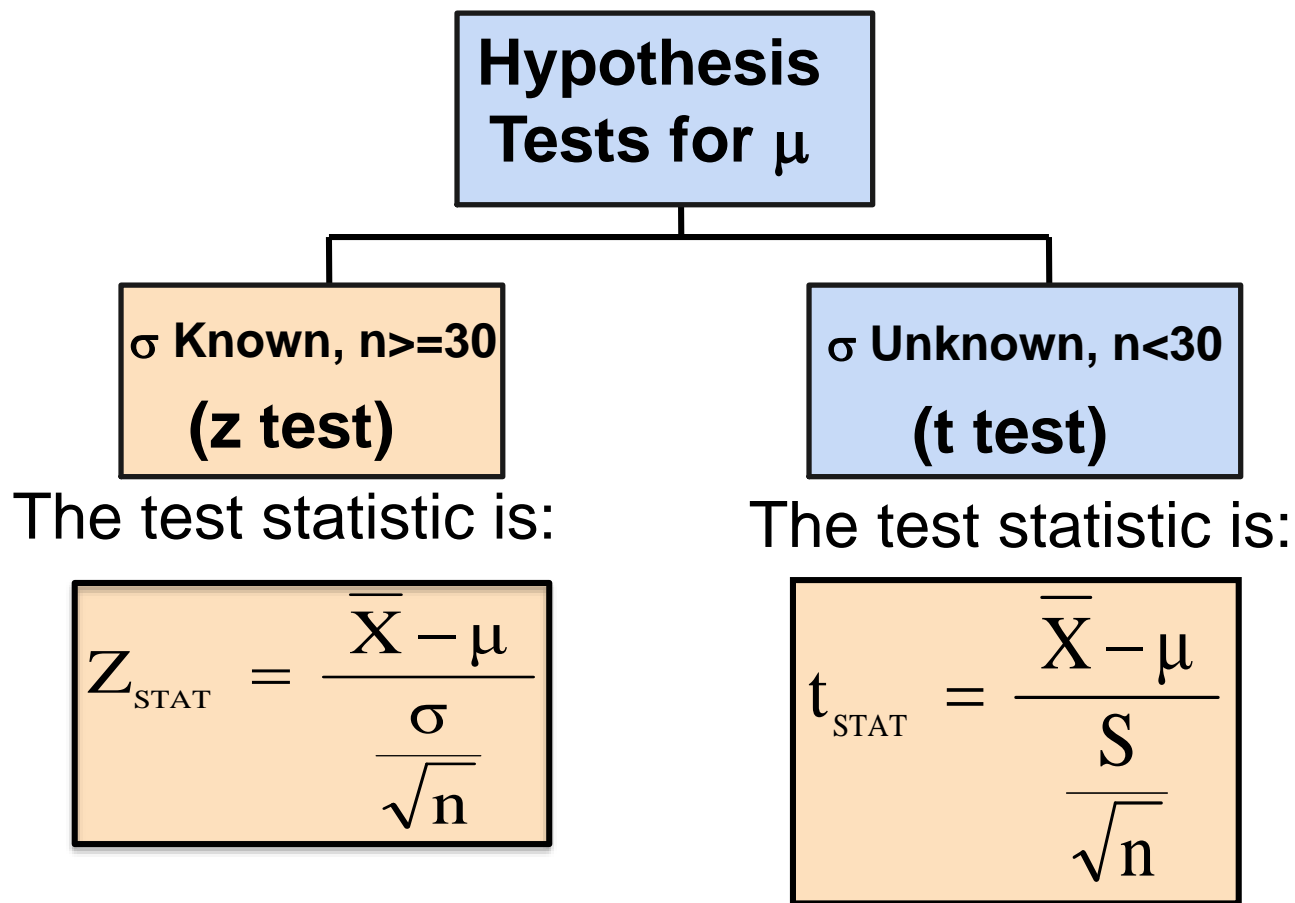
- Two-sided or two-tailed statistical test



Hypothesis testing on one sample mean



Hypothesis testing on one sample mean



Steps in Hypothesis Testing

1. State the null hypothesis, H_0 and the alternative hypothesis, H_1
 2. Choose the level of significance, α , and the sample size, n . Or, you can claim statistical confidence level, $\alpha = 1 - \text{confidence level}$
 3. Determine the appropriate test statistic and sampling distribution
 4. Determine the critical values that divide the rejection and non-rejection regions
-

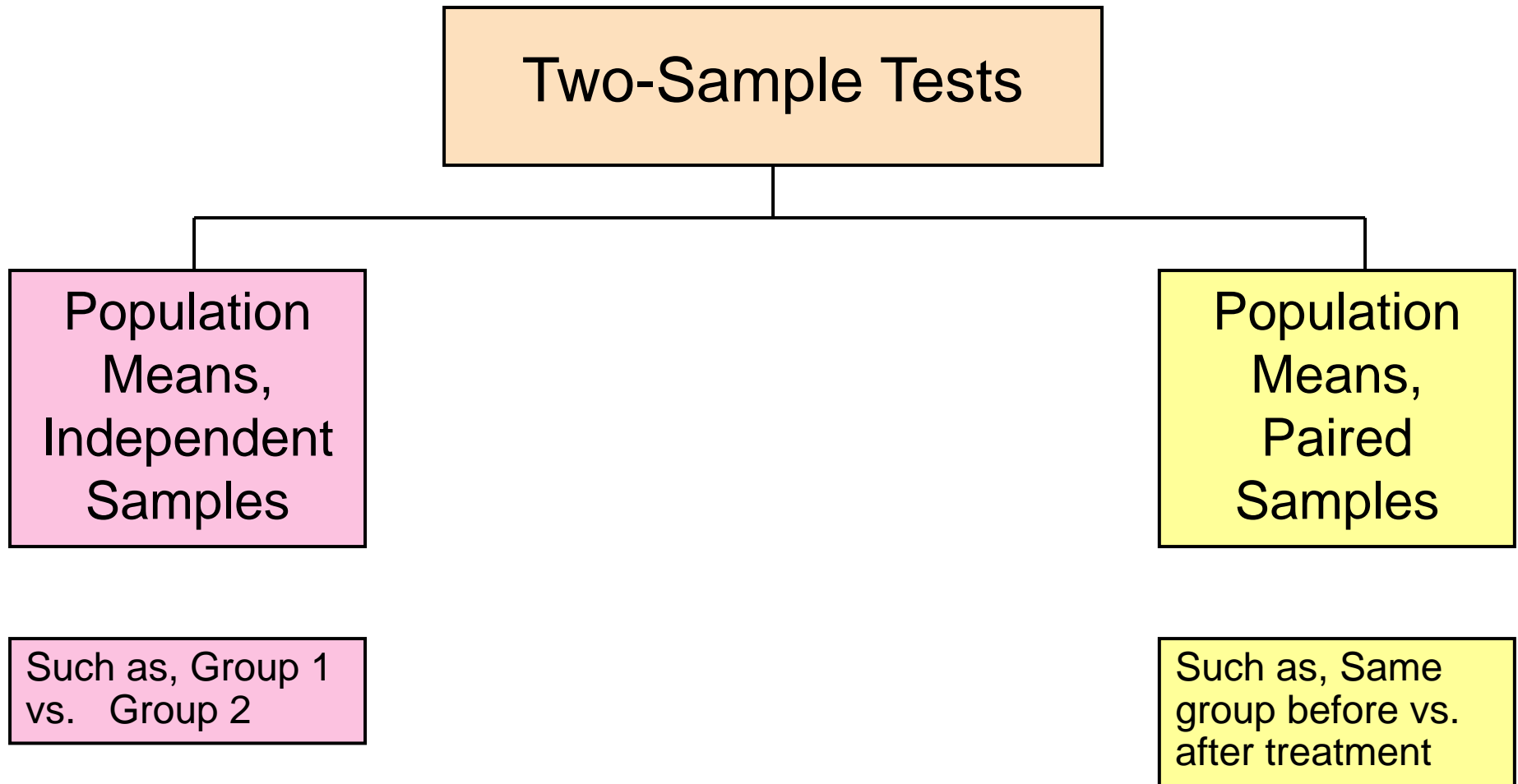
Steps in Hypothesis Testing

5. Collect sample and obtain the test statistic
 6. Make the statistical decision and state the managerial conclusion.
 - ☐ By using test statistics
If it falls in the rejection area, we reject H_0
 - ☐ By using p-value
If the p-value $< \alpha$, we reject H_0 and accept H_a
 - ☐ By using confidence interval
-

Hypothesis testing on two sample means



Two-Sample Test



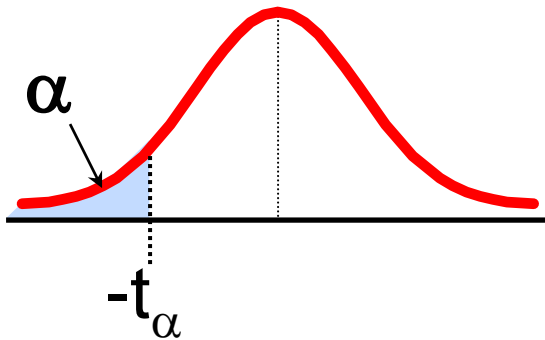
Hypothesis in Two-Sample Test

Two Population Means, Independent Samples

Lower-tail test:

$$H_0: \mu_1 - \mu_2 \geq 0$$

$$H_1: \mu_1 - \mu_2 < 0$$

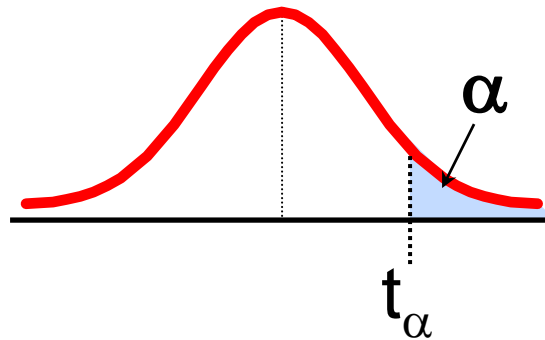


Reject H_0 if $t_{\text{STAT}} < -t_\alpha$

Upper-tail test:

$$H_0: \mu_1 - \mu_2 \leq 0$$

$$H_1: \mu_1 - \mu_2 > 0$$

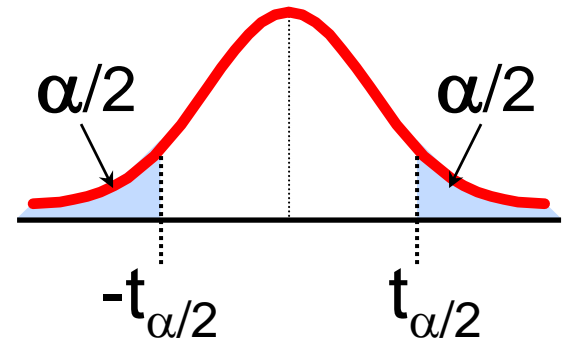


Reject H_0 if $t_{\text{STAT}} > t_\alpha$

Two-tail test:

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 - \mu_2 \neq 0$$



Reject H_0 if $t_{\text{STAT}} < -t_{\alpha/2}$
or $t_{\text{STAT}} > t_{\alpha/2}$

Two-Sample Test: Paired Samples

Two Population Means, Paired Samples

Paired Difference Confidence Interval for $\mu_d = \mu_1 - \mu_2$

Large Sample

$$\bar{y}_d \pm z_{\alpha/2} \frac{\sigma_d}{\sqrt{n_d}} \approx \bar{y}_d \pm z_{\alpha/2} \frac{s_d}{\sqrt{n_d}}$$

Assumption: Sample differences are randomly selected from the population.

Small Sample

$$\bar{y}_d \pm t_{\alpha/2} \frac{s_d}{\sqrt{n_d}}$$

where $t_{\alpha/2}$ is based on $(n_d - 1)$ degrees of freedom

Assumptions:

1. Population of differences has a normal distribution.
2. Sample differences are randomly selected from the population.



Two-Sample Test: Paired Samples

Paired Difference Test of Hypothesis for $\mu_d = \mu_1 - \mu_2$

ONE-TAILED TESTS

TWO-TAILED TEST

$$H_0: \mu_d = D_0 \quad H_0: \mu_d = D_0 \quad H_0: \mu_d = D_0$$

$$H_a: \mu_d < D_0 \quad H_a: \mu_d > D_0 \quad H_a: \mu_d \neq D_0$$

Large Sample

$$\text{Test statistic: } z = \frac{\bar{y}_d - D_0}{\sigma_d / \sqrt{n_d}} \approx \frac{\bar{y}_d - D_0}{s_d / \sqrt{n_d}}$$

<i>Rejection Region:</i>	$z < -z_\alpha$	$z > z_\alpha$	$ z > z_{\alpha/2}$
<i>p-value:</i>	$P(z < z_c)$	$P(z > z_c)$	$2P(z > z_c)$ if z_c positive $2P(z < z_c)$ if z_c negative

Assumption: The differences are randomly selected from the population of differences.

Small Sample

$$\text{Test statistic: } t = \frac{\bar{y}_d - D_0}{s_d / \sqrt{n_d}}$$

<i>Rejection region:</i>	$t < -t_\alpha$	$t > t_\alpha$	$ t > t_{\alpha/2}$
<i>p-value:</i>	$P(t < t_c)$	$P(t > t_c)$	$2P(t > t_c)$ if t_c is positive $2P(t < t_c)$ if t_c is negative

Assumptions:

1. The relative frequency distribution of the population of differences is normal.
2. The differences are randomly selected from the population of differences.

ANOVA

Analysis of variance (ANOVA) is an approach to compare statistics (such as means) among more than two groups.

From the name, we can see it is not the analysis on the group means, but the variance. Why????????? ➔ because the variance matters!!!! See the box plots

The goal in ANOVA: **compare group means among more than two groups by analyze the variances!!** The two-sample t-test can be considered as a special case of ANOVA, when there are only two groups.

ANOVA, is also an application of linear regression models.



Comparing more than two groups

Problem:

- We have K **independent** simple random samples from each of K populations.
- Each population is **normal** with unknown average μ_t
- All the populations have the same standard deviation σ

Question: *Are the observed differences among the sample means statistically significant? (or just due to chance variability?)*

Answer: The ANOVA F-test tests the hypotheses:

H₀: $\mu_1 = \mu_2 = \dots = \mu_K$ - *the averages are all equal*

H_a: *not all the μ_t are equal*



Steps for comparing K groups

1. Be sure that the observations arise from independent groups!
2. Draw side-by-side box plots for the groups, to visualize the differences among the groups and the within-group variation
3. Estimate the ANOVA regression model for $t=1,\dots,K$

where the errors e_{it} are normally distributed and with constant standard deviation σ . Use the regression F-test to check the hypothesis that the averages are equal.

4. Examine the residuals to verify that the model assumptions are satisfied.

Understand why ANOVA problem can be solved by regressions



Predictive Models

Linear Regression Model



Regression Analysis

- **Regression analysis** is used to:
 - Predict the value of a dependent variable based on the value of at least one independent variable
 - Explain the impact of changes in an independent variable on the dependent variable
- **Dependent variable:** the variable we wish to predict or explain
- **Independent variable:** the variable used to predict or explain the dependent variable

Types of Regression Analysis

- Simple Linear Regression Analysis

- $y = f(x) = \beta_0 + \beta_1 x + e$

- Multiple Linear Regression Analysis

- $y = f(x_1, x_2, x_3, \dots, x_n) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + e$

- Multivariate Regression Analysis

- $y_1, y_2, y_3, \dots, y_m = f(x_1, x_2, x_3, \dots, x_n)$

Data Splits for Evaluations

1). Hold-out Evaluation



If your data is large enough

Color	Weight (lbs)	Stripes	Tiger?
Orange	300	no	no
White	50	yes	no
Orange	490	yes	yes
White	510	yes	yes
Orange	490	no	no
White	450	no	no
Orange	40	no	no
Orange	200	yes	no
White	500	yes	yes
Green	560	yes	no
Orange	500	yes	?
White	50	yes	?

Training Data Set

Validation Data Set



Unseen data set

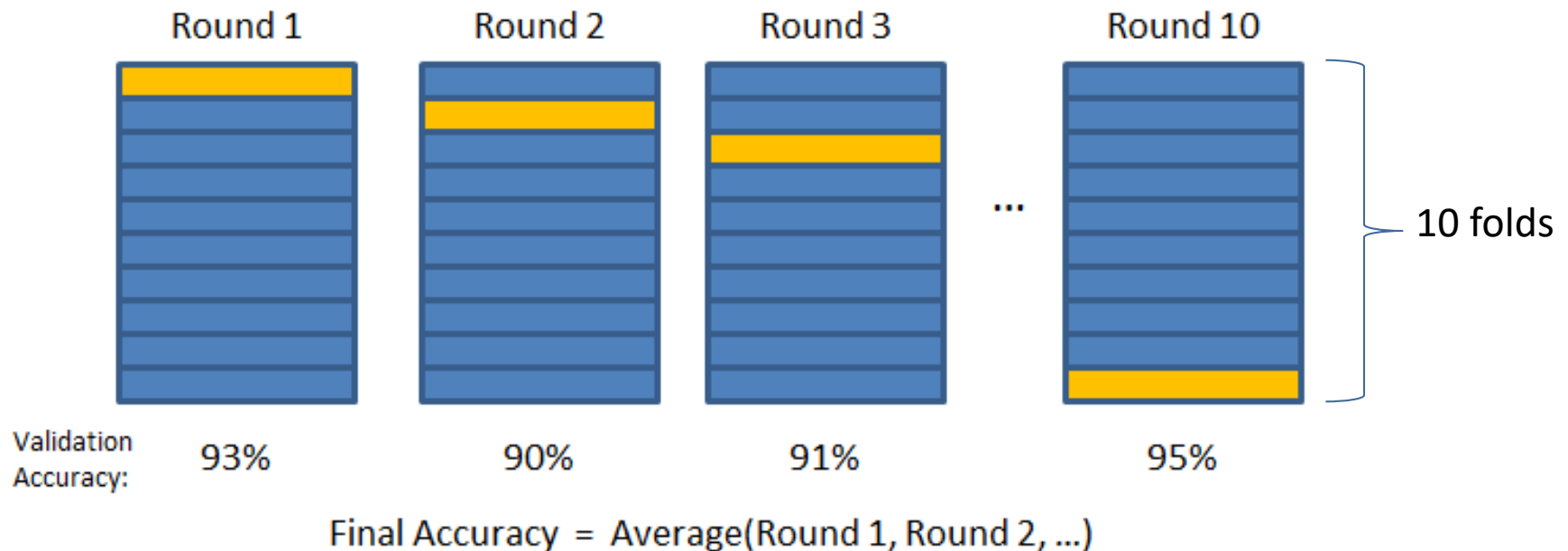
Data Splits for Evaluations

2). N-folds Cross Evaluation



If your data is relatively small

 Validation Set
 Training Set



Evaluate Models based on Test Sets

Root Mean Square Error :

Best model minimizes RMSE

$$RMSE = \sqrt{\frac{\sum_{i=1}^m (y_i - \hat{y}_i)^2}{m}}$$

Mean Absolute Error

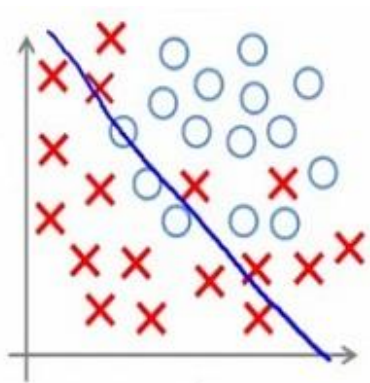
Best model minimizes MAE

$$MAE = \frac{\sum_{i=1}^m |y_i - \hat{y}_i|}{m}$$



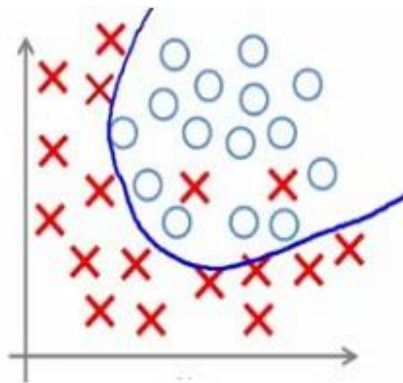
Overfitting Problem

- Overfitting is a general problem in learning algorithms
- It refers to that the predictive model is learned too much from the training set that it cannot be used on unseen data
- How to identify overfitting? ➔ you model can obtain high accuracy or low error in the training set, but it results in low accuracy or higher error in your testing data set

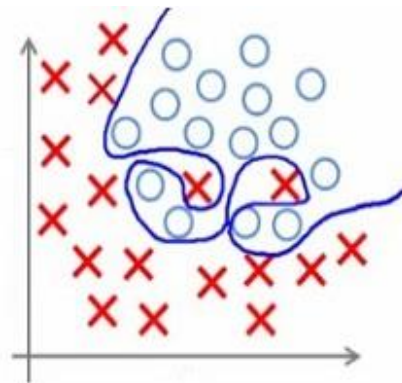


Under-fitting

(too simple to
explain the
variance)



Appropriate-fitting



Over-fitting

(forcefitting -- too
good to be true)

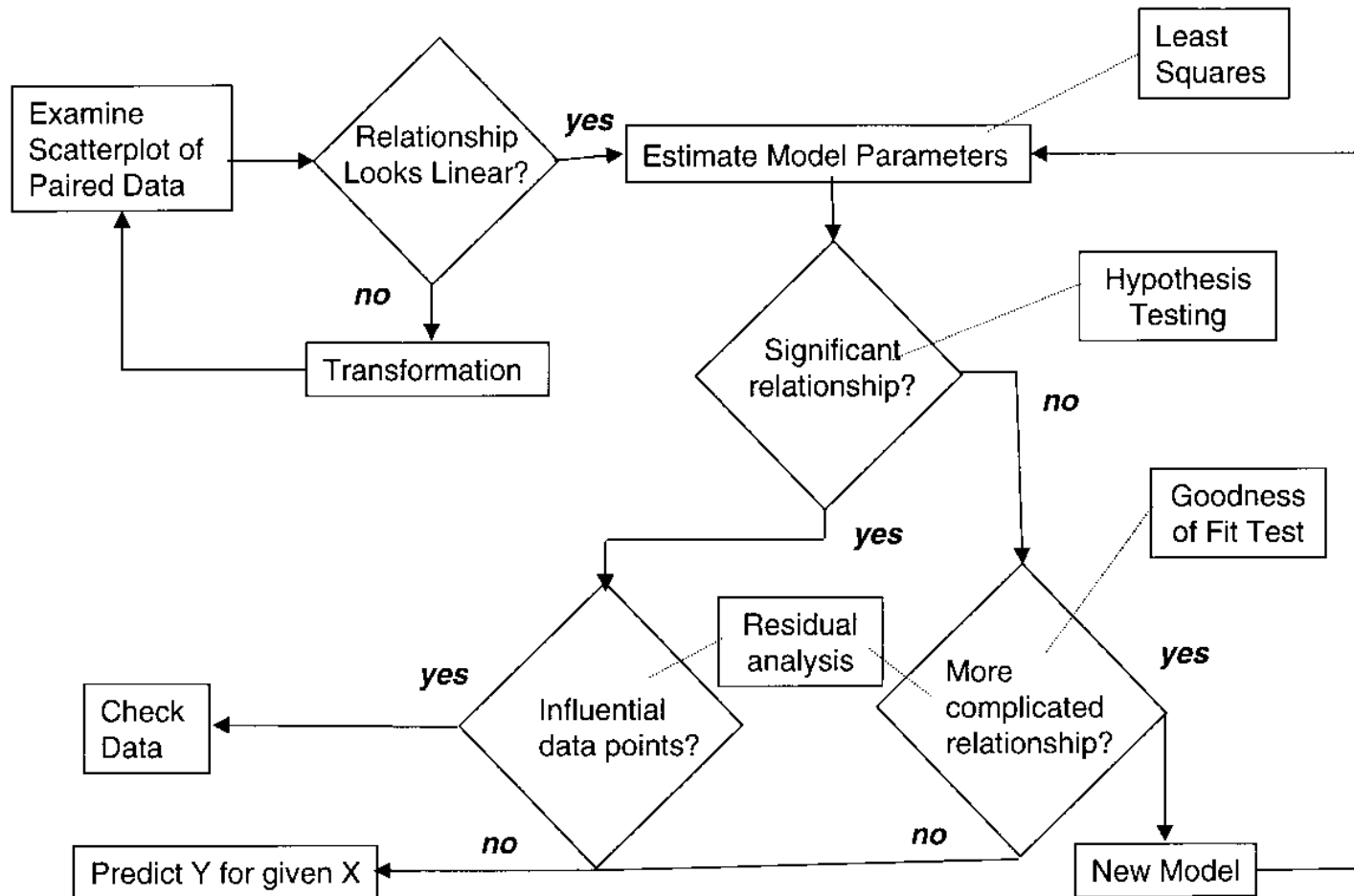
Overfitting Problem

How to alleviate overfitting problems in regression models?

- Use N-fold cross validation if data is small
- Add regularization terms in your objective functions in the regression models
 - Ridge regression terms
 - Lasso regression terms



Multiple Linear Regression



11-16

Steps for Linear Regression Models

- ❑ Understand your data, figure out x and y variables
- ❑ According to the size of data, make a decision about which evaluation strategy you are going to use, hold-out or N-fold cross validation
- ❑ Examine the relationship between y and x. Apply transformations to x or y variables, if necessary
- ❑ Split the data if necessary
 - ❑ Hold-out evaluation: split data to train and test sets, build models based on train, and test it over test set
 - ❑ N-fold: use full data or a sample of the data to build models first, and evaluate models by N-fold



Steps for Linear Regression Models

- ☐ Build Models by different feature selections
- ☐ For each model, validate they are qualified or not
 - ☐ Vif function to examine multi-collinearity problem
 - ☐ F-test to examine at least one x variable is influential
 - ☐ Residual analysis to examine the assumptions of residual
- ☐ Evaluate your models
 - ☐ Hold-out: evaluate models based on the testing set
 - ☐ N-fold: using `cv.glm()`
 - ☐ Metrics: MAE, RMSE, MSE



Steps for Linear Regression Models

☐ Improve your models

- ☐ Using nominal data?
- ☐ Try higher-order terms, if necessary
- ☐ Try interaction terms, especially dummy vs numerical variable
- ☐ Identify and remove influential points?

☐ Final Steps

- ☐ Write down the model
- ☐ Try to explain it
- ☐ Use the best model to make predictions



Assumptions on the regression model

1. The relationship between the Y-variable and the X-variables is linear
2. The error terms e_i (measured by the residuals)
 - have zero mean ($E(e_i)=0$)
 - have the same standard deviation σ_e for each fixed x
 - are close to normal distribution- Typically true for large samples!
 - are independent (true if sample is from simple random sampling)

Such assumptions are necessary to derive the inferential methods for testing and prediction (to be examined by residual analysis)

WARNING: if the sample size is small ($n < 30$) and residuals are not normal, you can't use regression methods!

Goals in Residual Analysis

1. Validate the constant variance
 2. Validate the linearity relationship
 3. Validate normal distribution of residuals
 4. Identify potential outliers (Optional)
-

How do I know I have a problem?

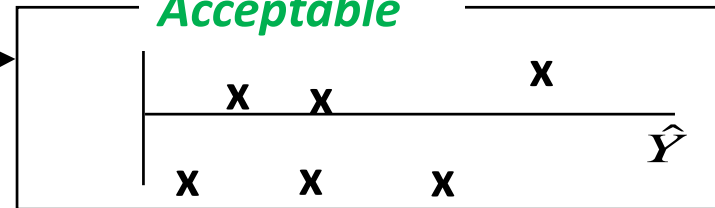
Assumption of constant variance

- 1) Plot residuals vs predicted values
- 2) Plot residuals vs each x-variable

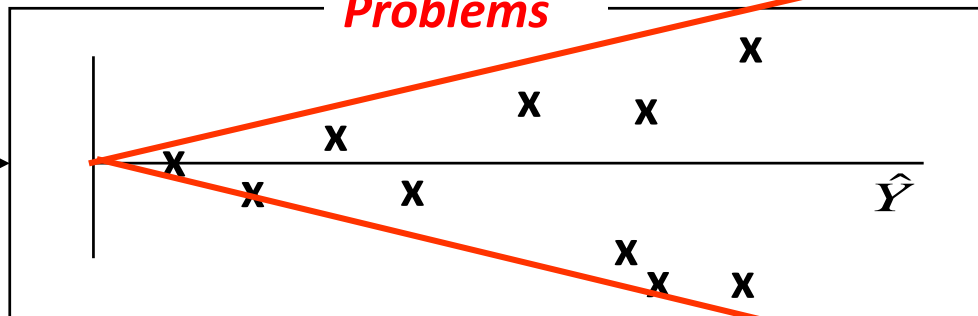
What is the pattern of the spread in the residuals as the predicted values increase?

- Spread constant.
- Spread increases.
- Spread decreases then increases.

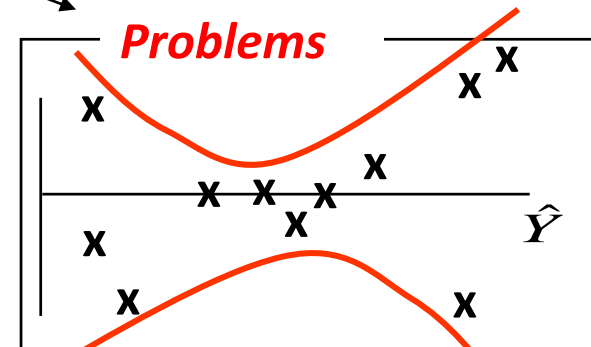
Acceptable



Problems

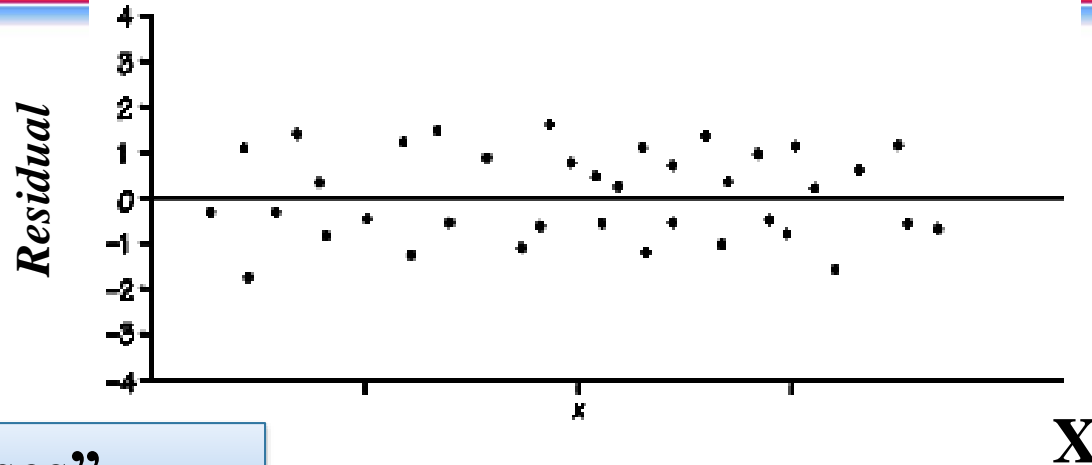


Problems



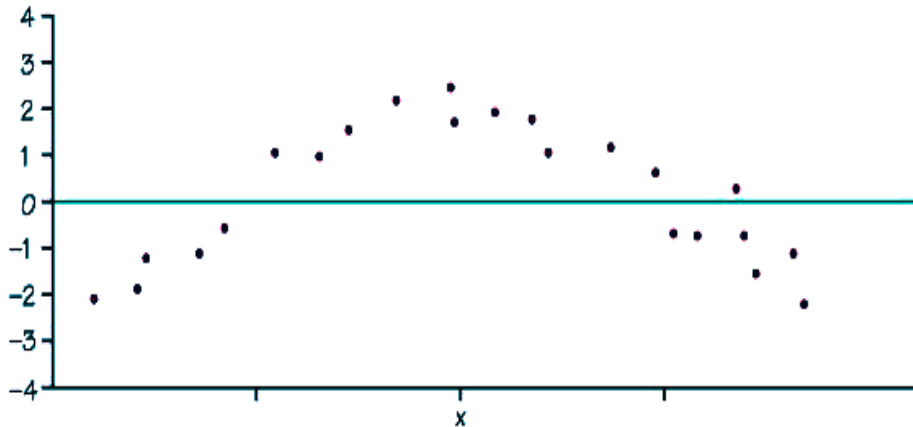
“Good case”

Points are randomly scattered around the zero line

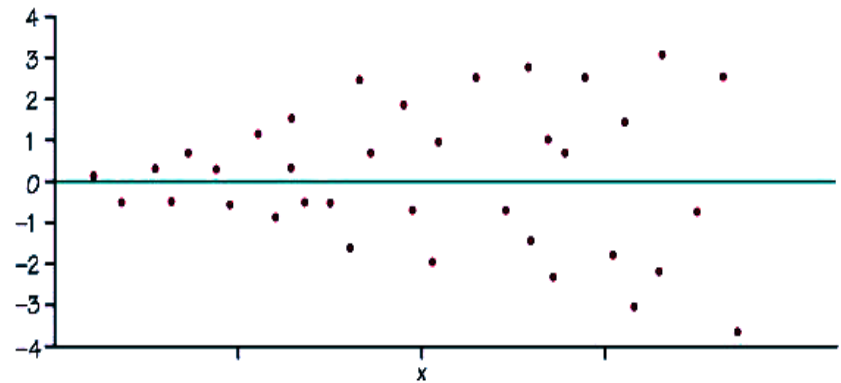


“Bad cases”

Non linear relationship



Non constant variance

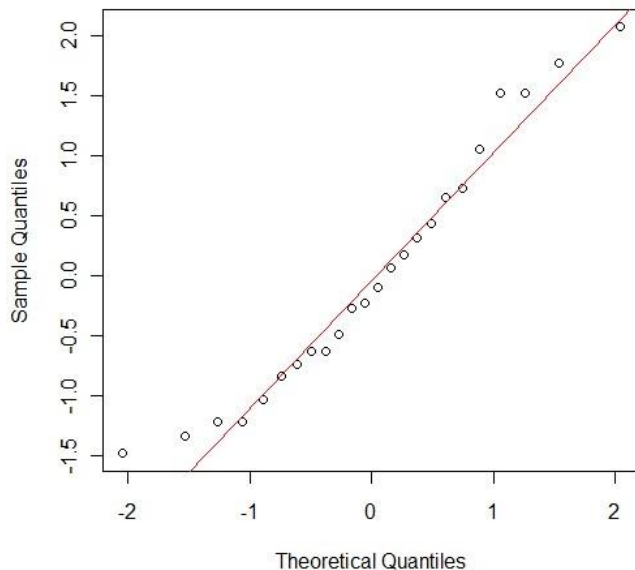


Normal probability plot of residuals

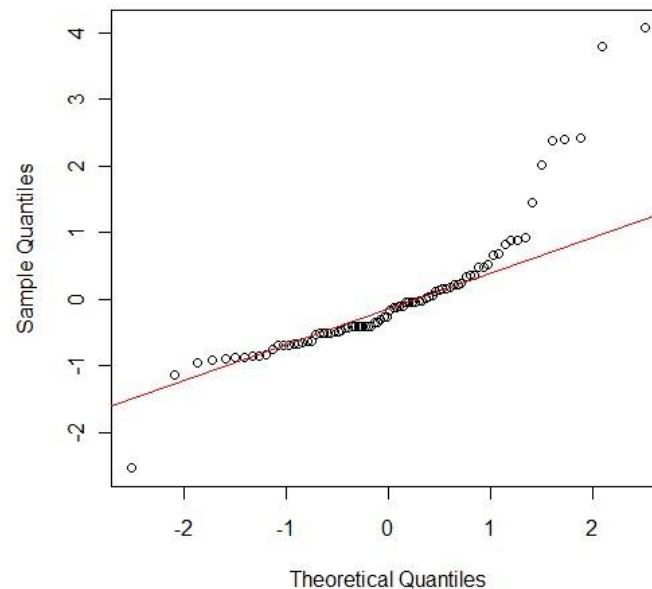
There are two ways to check whether a variable (including residual) follows normal distribution or not.

Solution-1: by OOPlot ("Q" stands for quantile)

Normal Q-Q Plot



Normal Q-Q Plot



Solution-2: Normality Test



How to identify outliers and influential Points

Outliers: residual plot, observations out of $[-3, 3]$ range

Influential Points: you can use `influence.measures` function to produce results based on different metrics, such as cook's distance, hat value, etc

Steps for Linear Regression Models

☐ Improve your models

- ☐ Using nominal data?
- ☐ Try higher-order terms, if necessary
- ☐ Try interaction terms, especially dummy vs numerical variable
- ☐ Identify and remove influential points? By cook's distance, cut off: $4/n$

☐ Final Steps

- ☐ Write down the model
- ☐ Try to explain it
- ☐ Use the best model to make predictions



About Final Projects

- General Idea
- Requirements
- Where to find the data
- Steps to do



About Final Projects

- General Idea
- Requirements
- Where to find the data
- Steps to do



About Final Projects

- **Goals by the Final Project**

Examine your practical skills

Train your research and experimental capabilities

Train your presentation skills

Train your paper or report writing skills

Encourage you to learn going beyond the class

Encourage you to solve problems by yourself

Encourage you to work individually

Encourage you to work in a team



About Final Projects

- **Work In Team or Individually**

You can choose to work individually

You can work together in a team

A team can have up to 3 students



About Final Projects

- **Grading Details**

I will grade your final projects, not TA
It is different from grading assignments.

- In terms of grading assignments
You get 100 as long as your answers are correct
- In terms of grading final projects
Your project will be compared with other groups
For example, you get 80. It does not mean your project is not good, but there are better ones
- You cannot argue once I give you the grades
At the end of your presentation, I will give you feedbacks
Finally, you will see a score (without comments) on blackboard system



About Final Projects

- General Idea
- Requirements
- Where to find the data
- Steps to do



About Final Projects

- Requirements (The minimal requirements)
 - Well-defined problems
 - At least one hypothesis testing – this one can NOT be the one in the process of modeling building. For example, F-test and individual parameter tests are not counted
 - We learnt three types of the predictive models: linear regression, classification and time-series. You must utilize at least one of these models to solve the problems in your data
 - Appropriate solutions
 - Correct evaluations among multiple models
 - By using R or SAS only



About Final Projects

- Two Predictive Tasks
 - Linear regression → y is numerical variable
 - Classification → y is categorical variable
- You should perform at least one task in this list.
More than one are also welcome



About Final Projects

- **Notes: how to find data and define research problems**
 - Where to find the data? See the following slides
 - You must find a data you are interested in
 - You may need more than one data and integrate them together
 - You need to understand the data first, and then think about what are the problems you want to solve
 - Then figure out whether you can use the techniques or models you learnt to solve the problem
 - If the current data is not ideal, you need to find another data and go through the steps again until you are satisfied with one data



About Final Projects

- General Idea
- Requirements
- Where to find the data
- Steps to do



About Final Projects

- You should find data by yourself

Kaggle: <https://www.kaggle.com/>

UCI: <https://archive.ics.uci.edu/ml/datasets.html>

<https://yongzhengme.wordpress.com/research/>

Different challenges or competitions



Example: Data on Kaggle.com

kaggle



Competitions

Datasets

Kernels

Discussion

Jobs



Welcome to Kaggle Competitions

Challenge yourself with real-world machine learning problems



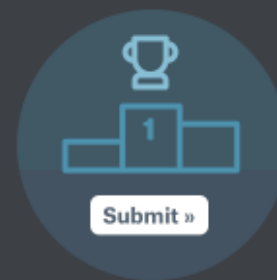
New to Data Science?

Get started with a tutorial on our most popular competition for beginners, [Titanic: Machine Learning from Disaster](#).



Build a Model

Get the data & use whatever tools or methods you prefer to make predictions.



Make a Submission

Upload your prediction file for real-time scoring & a spot on the leaderboard.

Example: Data on Kaggle.com

Datasets

Learn More

New Dataset

445 featured datasets

Sort By Hotness 

Featured All Mine Upvoted



▲
32



Pima Indians Diabetes Database

Predict the onset of diabetes based on diagnostic measures

UCI Machine Learning · updated 5 months ago

2,301 downloads
80 kernels
14 comments

▲
30



Medical Appointment No Shows

Why do 30% of patients miss their scheduled appointments?

JoniHoppen · updated 20 days ago

1,344 downloads
64 kernels
13 comments

Raise some questions

- Do they make sense?
- Can we use data analytics techniques to solve them
- How to evaluate them?
- Any interpretations or explanations?
- How useful your outcomes are? Or, how can you use them in the future?



Rules

- You cannot use the following data
 - Weather data
 - Crime data
 - Housing data
 - Hotel data
 - Airline data
 - Support data
 - Game data



About Final Projects

- General Idea
- Requirements
- Where to find the data
- Steps to do



About Final Projects

- Step 1: fill in a survey
- Step 2: decide to work individually or by team
- Step 3: write your project proposal
- Step 4: once your proposal is qualified, you can start working. Otherwise, you need to revise and resubmit the proposal
- Step 5: final project presentations and final report



About Final Projects

- **Step 1: User Survey**

https://depaul.qualtrics.com/jfe/form/SV_4Nmci4o9fFYAvu5

- **Benefits**

- Give you a list of potential data/topics
- Note: the data you can use is not limited to this list
- Collect your tastes on topics of final projects
- Build better predictive or recommendation models to help students find the appropriate topics for projects



About Final Projects

- Step 2: decide to work individually or by team
 - There should be no more than 3 students in a team



About Final Projects

- Step 3: write the proposal, Due: March 31
 - The template has been uploaded
 - Each team should ONLY submit one copy. For example, if there are 3 students in your team, it is enough for one student to submit the proposal.
Do NOT submit multiple copies by different students!
 - Note: you may start working on it as soon as possible. Every semester, there are many teams whose proposal is disqualified.



About Final Projects

- Step 1: fill in a survey
- Step 2: decide to work individually or by team
- Step 3: write your project proposal
- Step 4: once your proposal is qualified, you can start working. Otherwise, you need to revise and resubmit the proposal
- Step 5: final project presentations and final report

