FUNDAMENTALS OF MACHINE LEARNING

Review on Statistics

CSCI3320

Prof. John C.S. Lui, CSE Department, CUHK
Introduction to Machine Learning

STATISTICAL SAMPLING

John C.S. Lui, CSE Department, CUHK

Sampling Theory

- A population is considered to be known if we know the probability distribution function (pdf) $f(x)$
- For a given probability distribution (e.g., normal, exponential), there are some intrinsic *parameters* we need to know.
- Define:
  - X_i as the random variable for the i^{th} sample *(give example)*
  - x_i as the value of X_i *(give example)*
- Sample of size *n* (**with replacement**), we have

$$P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = f(x_1)f(x_2) \dots f(x_n)$$

- Sample statistics: $g(X_1, \dots, X_n)$

Sampling Distribution

- A sample statistics that is computed from $X_1, \ldots, X_n$ is a function of these random variables, and therefore itself is also a random variable: $g(X_1, \ldots, X_n)$
- The probability distribution of a sample statistics is a *sampling distribution* of the statistics
- Some well-known statistics $g()$ of sampling distribution
  - Sample mean
  - Sample variance
- Let $X_1, \ldots, X_n$ be *iid* for a random sample of size *n*. The ***sample mean*** is a random variable: $\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$
- Mean of that sample: $\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$

Sampling Distribution of Mean

- Let $f(x)$ be the probability distribution of some given population from which we draw a sample of size *n*.
- We examine the probability distribution of the sample statistics, ***sampling distribution for the sample mean***, $\bar{X}$

- **Theorem 1:** The mean of the sampling distribution of means, denoted as $\mu_{\bar{X}}$, is given by

$$E(\bar{X}) = \mu_{\bar{X}} = \mu$$

Pictorial illustration

where μ is the mean of the population

Sampling Distribution of Mean

- **Theorem 2:** If a population is ***infinite*** and the sampling is random, or if the population is ***finite*** and sampling is with replacement, then the variance of the sampling distribution of means, denoted by $\sigma_{\bar{X}}^2$, is given by

$$E[(\bar{X} - \mu)^2] = \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$$

where σ^2 is the variance of the population

Pictorial illustration in class

Sampling Distribution of Mean

- **Theorem 3:** If a population is of size N, if sampling is without replacement, and if the sample size is $n \leq N$, then the variance of the sample means is

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

while σ^2 is from Theorem 2

Sampling Distribution of Mean

- **Theorem 4:** If the population from which samples are taken is ***normally distributed*** with mean μ and variance σ^2, then the sample mena is also ***normally distributed*** with mean μ and variance σ^2/n.

Similar to Theorem 2

- **Theorem 5:** Suppose that the population from which samples are taken has a probability distribution with mean μ and variance σ^2, and that it is ***not necessary*** a normal distribution. Then the ***standardized variable*** associated with $\bar{X}$, given by

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

is asymptotically normal, i.e.,

$$\lim_{n\to\infty} P(Z \le z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\mu^2/2}\, du$$

Standard normal distribution

Sampling Distribution of Mean

- Note that **Theorem 5** is a consequence of the ***central limit theorem (CLT).*** It is assumed here that the population is infinite, or that sampling is with replacement.

- Otherwise, **Theorem 5** is still correct if we replace $\sigma/\sqrt{n}$ in **Theorem 5** by $\sigma_{\bar{X}}^2$ as given in **Theorem 3.**

Examples

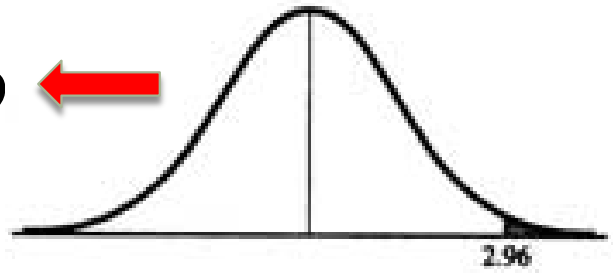
Example 6.2. Five hundred ball bearings have a mean weight of 5.02 oz and a standard deviation of 0.30 oz. Find the probability that a random sample of 100 ball bearings chosen from this group will have a combined weight of more than 510 oz.

For the sampling distribution of means, $\mu_{\bar{X}} = \mu = 5.02$ oz, and

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{0.30}{\sqrt{100}}\sqrt{\frac{500-100}{500-1}} = 0.027 .$$

The combined weight will exceed 510 oz if the mean weight of the 100 bearings exceeds 5.10 oz.

$$5.10 \text{ in standard units} = \frac{5.10-5.02}{0.027} = 2.9$$



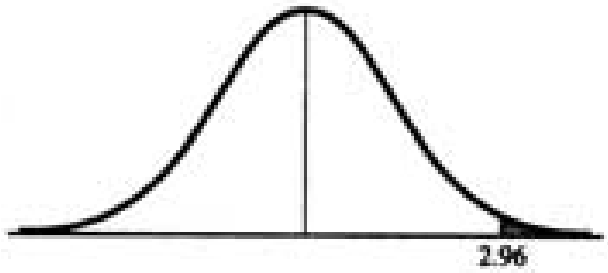
Examples

Required Probability = (area to the right of $z = 2.96$)

= (area to the right of $z = 0$) –
(area between $z = 0$ and $z = 2.96$)

= $0.5 - 0.4985 = 0.0015$

Therefore, there are only 3 chances in 2000 of picking a sample of 100 ball bearings with a combined weight exceeding 510 oz.



Key Points

- Assume existence of distribution for the underlying population
- By construction (via sampling), we can create the "sampling distribution" (*note that each sampling experiment will generate a random outcome, so we have a "sampling distribution"*)
- Given the existence of sampling distribution, we can now draw one sample from the population, then use the sampling distribution to quantify the probability of our random outcome

Sampling Distribution of Proportions

- Suppose that a population is infinite and binomially distributed with *p* and *q=1-p* being the respective probabilities that any given member exhibits or does not exhibit of a certain property *(give example)*
- Consider all possible samples of size *n* drawn from the above population and each sample determines the statistics that is the proportion *P* of success (e.g., for a coin, the proportion of heads turning up in *n* tosses)
- We obtain a *sampling distribution* with mean μ_P and standard deviation σ_P

$$\mu_P = p \qquad \sigma_P = \sqrt{\frac{pq}{n}} = \sqrt{\frac{p(1-p)}{n}}$$

derive in class

Sampling Distribution of Proportions

- The result can be obtained by Theorem 1 and 2 by replacing $\mu = p, \sigma = \sqrt{pq}$

- For large values of n $(n \geq 30)$, the sampling distribution is nearly a normally distribution (as seen from Theorem 5).

- For finite population in which sampling is without replacement, the equation for σ_p given above, is replaced by $\sigma_{\bar{X}}$ as given in Theorem 3 with $\sigma = \sqrt{pq}$.

The Sampling Variance

- Let $X_1, X_2, \ldots, X_n$ denote the random variables for a sample of size n, then the random variable giving the ***variance of the sample*** or the ***sample variance*** is

$$S^2 = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2}{n}$$

- The ***unbiased estimator*** for the above sample variance is

$$\hat{S}^2 = \frac{n}{n-1} S^2 = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2}{n-1}$$

ESTIMATION THEORY

John C.S. Lui, CSE Department, CUHK

Estimation Theory

- **Unbiased Estimator:** a statistics of a population parameter *if the expectation of the statistics is equal to the parameter*
- If the sampling distribution of two statistics have the same mean, the statistics with the smaller variance is called a ***more efficient estimator*.**
- **Point Estimate:** an estimate of a population parameter given by a *single number*
- **Interval Estimate:** an estimate of a population parameter given by two numbers between which the parameter may be considered
- A statement of the error or precision of an estimate is called the **reliability**

Confidence Interval Estimates of Population Parameters

Let μ_S and σ_S be the mean and standard deviation (standard error) of the sampling distribution of a statistic S. Then, if the sampling distribution of S is approximately normal (which as we have seen is true for many statistics if the sample size $n \geq 30$), we can expect to find S lying in the interval $\mu_S - \sigma_S$ to $\mu_S + \sigma_S$, $\mu_S - 2\sigma_S$ to $\mu_S + 2\sigma_S$ or $\mu_S - 3\sigma_S$ to $\mu_S + 3\sigma_S$ about 68.27%, 95.45%, and 99.73% of the time, respectively.

Similarly, $S \pm 1.96\sigma_S$ and $S \pm 2.58\sigma_S$ are 95% and 99% (or 0.95 and 0.99) confidence limits for μ_S. The percentage confidence is often called the *confidence level*. The numbers 1.96, 2.58, etc., in the confidence limits are called *critical values*, and are denoted by z_C. From confidence levels we can find critical values.

Confidence Intervals for Means (for large samples of n >= 30)

- If the statistics *S* is the sample mean $\overline{X}$, then the 95% and 99% confidence limits for estimating the population mean μ are given by $\bar{X} \pm 1.96\sigma_{\bar{X}}$ and $\bar{X} \pm 2.58\sigma_{\bar{X}}$
- For infinite population or if sampling with replacement for finite population:

$$\bar{X} \pm z_C \frac{\sigma}{\sqrt{n}}$$

- Finite population size *N* and sampling without replacement

$$\bar{X} \pm z_C \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

- When the population standard deviation σ is unknown, we use estimator $\hat{S}$

Example

Example 7.2: Find a 95% confidence interval estimating the mean height of the 1,546 male students at the XYZ University by taking a sample of size 100. Assume the mean of the sample, $\bar{X} = 67.45$ and that the standard deviation of the population σ, is 2.93 inches

The 95% confidence limits are $\bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}$.

Using $\bar{x} = 67.45$ inches and $\hat{s} = 2.93$ inches as an estimate of σ, the confidence limits are

$$67.45 \pm 1.96\left(\frac{2.93}{\sqrt{100}}\right) \text{ inches} \qquad \text{or} \qquad 67.45 \pm 0.57 \text{ inches}$$

Confidence Intervals for Means (for samples of n < 30)

- In this case, we use the *t*-distribution to obtain confidence intervals
- If $-t_{0.975}$ and $t_{0.975}$ are values of *T* for with 2.5% of the area lies in each tail of the *t* distribution
- The 95% confidence interval for *T* is

$$-t_{0.975} < \frac{(\bar{X} - \mu)\sqrt{n}}{\hat{S}} < t_{0.975}$$

- We can estimate μ with 95% confidence via:

$$\bar{X} - t_{0.975}\frac{\hat{S}}{\sqrt{n}} < \mu < \bar{X} + t_{0.975}\frac{\hat{S}}{\sqrt{n}}$$

Confidence Interval for Proportions

Suppose that the statistic S is the proportion of "successes" in a sample of size $n \geq 30$ drawn from a binomial population in which p is the proportion of successes (i.e., the probability of success). Then the confidence limits for p are given by $P \pm z_c \sigma_P$, where P denotes the proportion of success in the sample of size n. Using the values of σ_P obtained

- If infinite population or finite population & sampling with replacement

$$P \pm z_c \sqrt{\frac{P(1-P)}{n}}$$

- If finite population N, sampling without replacement

$$P \pm z_c \sqrt{\frac{P(1-P)}{n}} \sqrt{\frac{N-n}{N-1}}$$

Example

Example 7.3. A sample poll of 100 voters chosen at random from all voters in a given district indicate that 55% of them were in favor of a particular candidate. Find the 99% confidence limits for the proportion of all voters in favor of this candidate.

$$P \pm 2.58\sqrt{\frac{p(1-p)}{n}} = 0.55 \pm 2.58\sqrt{\frac{(0.55)(0.45)}{100}}$$

$$= 0.55 \pm 0.13$$

HYPOTHESIS TESTING

John C.S. Lui, CSE Department, CUHK

Test of Hypothesis & Significance

- We make *statistical hypotheses*, which are statements about the population distributions of the populations
- It is also the theoretical basis for **A/B test**
  - https://conversionsciences.com/ab-testing-statistics/
- **Example:** if we want to decide whether a given coin is loaded, we formulate the hypothesis that the coin is fair, i.e., *p=0.5*
- **Example:** if we want to decide whether one procedure is better than another, we formulate the hypothesis that there is *no difference* between the two procedures.
- The above hypotheses are called the **null-hypothesis,** which we denote has H_0
- A hypothesis alternative to H_0 is called H_1, which is the **alternate hypothesis**

Test of Hypothesis & Significance

- **Possible outcome of our hypothesis test:**
  - Reject H_0 (or accept H_1)
  - Fail to reject H_0, so we retain H_0
- **How do we test? :** *Statistical test via sampling*
  - *Example:*
    - *Flip the coin 50 times*
    - *Calculate the probability of seeing a head*
    - *Several possibilities:*
      - *The sample average is: 0.51 (**fail to reject** H_0)*
      - *The sample average is 0.92 (**reject** H_0)*
      - *The sample average is 0.59 (**should we accept/reject** H_0 **?**)*
- ***What we need is a formal procedure to decide/test.***

Type I and Type II Errors

- **Type I Error:** If we reject the hypothesis H_0 when it happens to be true (e.g., false negative)
- **Type II Error:** If we accept the hypothesis H_0 when it happens to be false (e.g., false positive)
- In testing a given hypothesis, the maximum probability with which we would be willing to risk a Type I error is called the **level of significance** *(the complement of it is level of confidence)*
- In practice, a level of significance of 0.05 or 0.01 is customary (of the level of confidence is 0.95 or 0.99)
- **Example:** whenever the null hypothesis is true, we are about 95% confident that we would make the right decision (*if we set the level of significance is 0.05*)

Test Involving the Normal Distribution

To illustrate the ideas presented above, suppose that under a given hypothesis, the sampling distribution of a statistic S is a normal distribution with mean μ_S and standard deviation σ_S. The distribution of that standard variable $Z = (S - \mu_S)/\sigma_S$ is the standard normal distribution (mean 0, variance 1) shown in Figure 8-1, and extreme values of Z would lead to the rejection of the hypothesis.

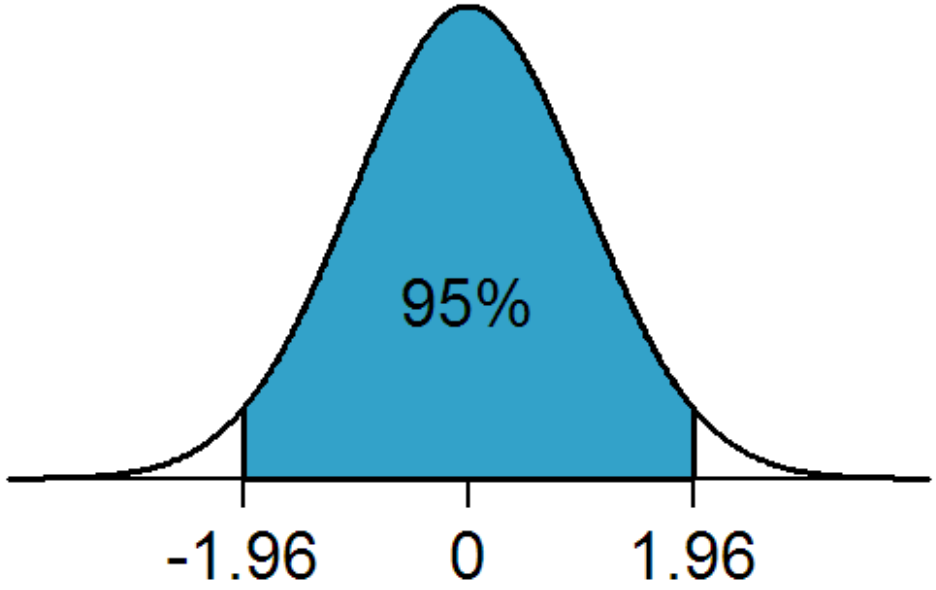


Figure 8-1

Test Involving the Normal Distribution

- We can be 95% confident that if the **hypothesis is true**, the *z* score on a sample statistics *S* is between -1.96 and 1.96
- However, if in a sample we find that the *z* score of its statistics lies ***outside*** the range [-1.96,1.96], we conclude the event could happen with only 0.05 probability if the given hypothesis is true. So we ***inclined to reject the hypothesis ('cause of such a low probability).***
- **Decision Rule:**
  - Reject the hypothesis at a 0.05 level of significance if the *z* score of the statistics *S* is outside [-1.96, 1.96]
  - Accept the hypothesis (or , if desired, make no decision at all), otherwise

One-Tailed and Two-Tailed Tests

- Above descriptions (or tests) are called the ***two-tailed tests*** or ***two-sided tests.***
- Many times, we are only interested with ***one-tailed tests***
- Table shows values of *z* for both one-sided/two-sided tests

Level of significance	0.10	0.05	0.01	0.005
Critical values of *z* for One-Tailed Tests	-1.28 or 1.28	-1.645 or 1.645	-2.33 or 2.33	-2.58 or 2.58
Critical values of *z* for Two-Tailed Tests	-1.645 and 1.645	-1.96 and 1.96	-2.58 and 2.58	-2.81 and 2.81

P-value

- The p-value is the probability of observing a sample statistic as extreme (or more extreme) than the one observed under the assumption that the null hypothesis is true.

Example of *P* value

For example, suppose the standard deviation σ of a normal population is known to be 3, and H_0 asserts that the mean μ is equal to 12. A random sample of size 36 drawn from the population yields a sample mean $\bar{x} = 12.95$. The test statistic is chosen to be

$$Z = \frac{\overline{X} - 12}{\sigma / \sqrt{n}} = \frac{\overline{X} - 12}{0.5},$$

which, if H_0 is true, is the standard normal variable. The test value of Z is the following:

$$Z = \frac{12.95 - 12}{0.5} = 1.9.$$

Example of *P* value

The P value for the test then depends on the alternative hypothesis H_1 as follows:

(i) For H_1: $\mu > 12$ [case (*i*) above], the P value is the probability that a random sample of size 36 would yield a sample mean of 12.95 or more if the true mean were 12, i.e., $P(Z \geq 1.9) = 0.029$. In other words, the chances are about 3 in 100 that $\bar{x} \geq 12.95$ if $\mu = 12$.

(ii) For H_1: $\mu < 12$ [case (*ii*) above], the P value is the probability that a random sample of size 36 would yield a sample mean of 12.95 or less if the true mean were 12, i.e., $P(Z \leq 1.9) = 0.971$. In other words, the chances are about 97 in 100 that $\bar{x} \leq 12.95$ if $\mu = 12$.

Example of *P* value

(iii) For H_1: $\mu \neq 12$ [case (*iii*) above], the *P* value is the probability that a random sample mean 0.95 or more units away from 12, i.e., $\bar{x} \geq 12.95$ or $\bar{x} \leq 11.05$, if the true mean were 12. Here the *P* value is $P(Z \geq 1.9) + P(Z \leq -1.9) = 0.057$, which says the chances are about 6 in 100 that $|\bar{x} - 12| \geq 0.095$ if $\mu = 12$.

Small *P* values provide evidence for rejecting the null hypothesis in favor of the alternative hypothesis, and large *P* values provide evidence for not rejecting the null hypothesis in favor of the alternative hypothesis. In case (*i*) of the above example, the small *P* value 0.029 is a fairly strong indicator that the population mean is greater than 12, whereas in case (*ii*), the large *P* value 0.971 strongly suggests that $H_0 : \mu = 12$ should not be rejected in favor of H_1: $\mu < 12$. In case (*iii*), the *P* value 0.057 provides evidence for rejecting H_0 in favor of H_1: $\mu \neq 12$ but not as much evidence as is provided for rejecting H_0 in favor of H_1: $\mu > 12$.

Example

Example 8.1. The mean lifetime of a sample of 100 fluorescent light bulbs produced by a company is computed to be 1570 hours with a standard deviation of 120 hours. If μ is the mean lifetime of all the bulbs produced by the company, test the hypothesis $\mu = 1600$ hours against the alternative hypothesis $\mu \neq 1600$ hours. Use a significance level of 0.05 and find the *P* value of the test.

$$H_0\colon\ \mu = 1600 \text{ hours}$$

$$H_1\colon\ \mu \neq 1600 \text{ hours}$$

Example

A two-tailed test should be used here since $\mu \neq 1600$ includes both values large and smaller than 1600.

For a two-tailed test at a level of significance of 0.05, we have the following decision rule:

1. Reject H_0 if the z score of the sample mean is outside the range –1.96 to 1.96.
2. Accept H_0 (or withhold any decision) otherwise.

Example

The statistic under consideration is the sample mean $\overline{X}$. The sampling distribution of X has a mean $\mu_{\overline{X}} = \mu$ and standard deviation $\sigma_{\overline{X}} = \sigma / \sqrt{n}$, where μ and σ are the mean and standard deviation of the population of all bulbs produced by the company.

Under the hypothesis H_0, we have $\mu = 1600$ and $\sigma_{\overline{X}} = \sigma / \sqrt{n} = 120 / \sqrt{100} = 12$, using the sample standard deviation as an estimate of σ. Since $Z = (\overline{X} - 1600)/12 = (1570 - 1600)/12 = -2.50$ lies outside the range –1.96 to 1.96, we reject H_0 at a 0.05 level of significance.

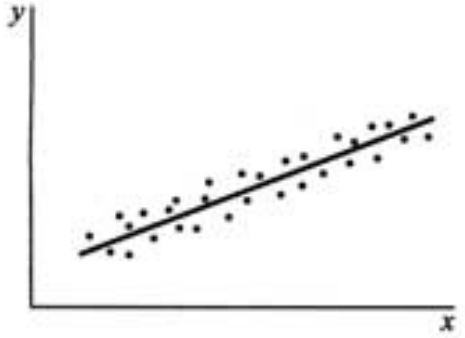
The P value of the two tailed test is $P(Z \leq -2.50) + P(Z \geq 2.50) = 0.0124$, which is the probability that a mean lifetime of less than 1570 hours or more than 1630 hours would occur by chance if H_0 were true.

REGRESSION AND CORRELATION

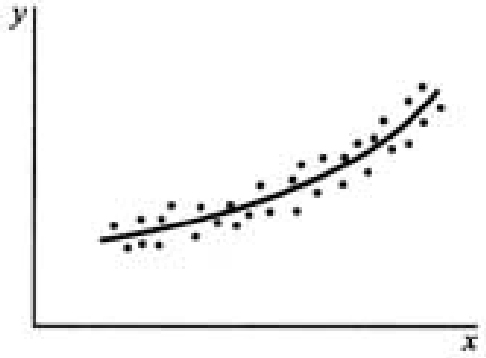
John C.S. Lui, CSE Department, CUHK

Curve Fitting

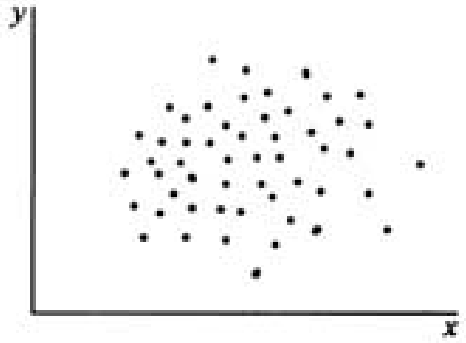
- In practice, a relationship is found to exist between two (or more) variables, and we wish to express this relationship in mathematical form by determining an equation "connecting" the variables
- For example, x and y be the height and weight of an adult male. We have a sample of n of x_i and $\mathrm{y_i}$, $i = 1,2,\ldots,n$
- We plot them in a scatter diagram



Linear



Non-linear



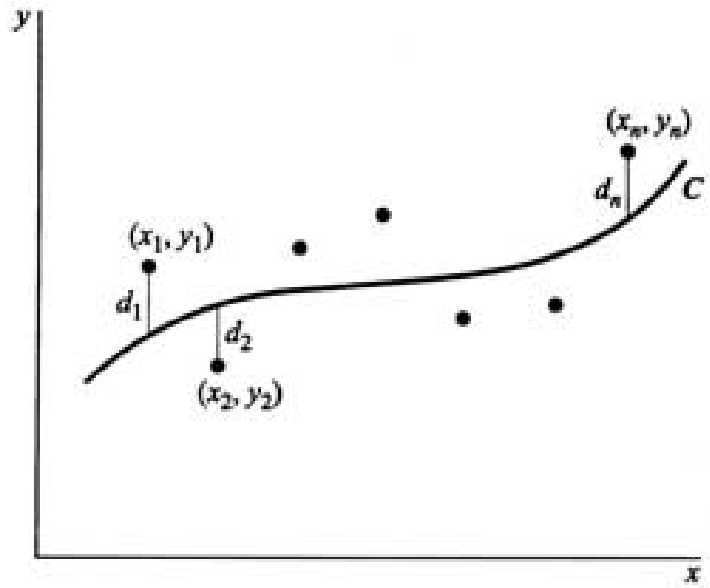
No relationship

Curve Fitting

- **Curve fitting**: finding equations approximating curves that fit given sets of data
- **Linear approximation:** $y = a + bx$
- **Quadratic approximation:** $y = a + bx + cx^2$
- Sometimes, it helps to plot scatter diagrams in terms of *transformed variables*, e.g., $\log y \ \ vs. \log x$

Regression

- **Regression:** do curve fitting so to estimate one of the variables (the *dependent variable*) from the other (the *independent variables*)
- One popular regression method is the "**Method of Least Squares**"
- Define deviation errors d_i, $i=1,\ldots,n$



Method of Least Squares

- How good is curve C? If $d_1^2 + \ ... + d_n^2$ is small, it is good

Definition Of all curves in a given family of curves approximating a set of n data points, a curve having the property that

$$d_1^2 + d_2^2 + \cdots d_n^2 = \text{a minimum}$$

is called a *best-fitting curve* in the family.

The Linear Least-Squares Line

- The linear least-squares line approximating the set of points has the equation form of: $y = a + bx$
- The constant *a* and *b* are determine by solving:

$$\sum_{j=1}^{n} y_j = an + b\sum_{j=1}^{n} x_j$$

$$\sum_{j=1}^{n} x_j y_j = a\sum_{j=1}^{n} x_j + b\sum_{j=1}^{n} x_j^2$$

- Solving for *a* and *b*, we have:

$$a = \frac{\left(\sum y\right)\left(\sum x^2\right) - \left(\sum x\right)\left(\sum xy\right)}{n\sum x^2 - \left(\sum x\right)^2} \qquad b = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2}$$

Example

Example 9.1. Table 9-1 shows the respective heights x and y of a sample of 12 fathers and their oldest sons. Find the least-squares regression line of y on x.

Table 9-1

Height x of father (inches)	65	63	67	64	68	62	70	66	68	67	69	71
Height y of son (inches)	68	66	68	65	69	66	68	65	71	67	68	70

$a = 35.82$ and $b = 0.476$, so that $y = 35.82 + 0.476x$

Least-Squared Regression Line in terms of Sample Variances and Co-variances

- We define sample averages, sample variances, sample covariances as:

$$\bar{x} = \frac{\sum_{j=1}^{n} x_j}{n}$$

$$\bar{y} = \frac{\sum_{j=1}^{n} y_j}{n}$$

$$s_x^2 = \frac{\sum_{j=1}^{n} (x_j - \bar{x})^2}{n}$$

$$s_y^2 = \frac{\sum_{j=1}^{n} (y_j - \bar{y})^2}{n}$$

$$s_{xy} = \frac{\sum_{j=1}^{n} (x_j - \bar{x})(y_j - \bar{y})}{n}$$

Least-Squared Regression Line in terms of Sample Variances and Co-variances

- Least-squared regression lines of *y on x* and *x on y*

$$y-\bar{y}=\frac{s_{xy}}{s_x^2}(x-\bar{x}) \quad \text{and} \quad x-\bar{x}=\frac{s_{xy}}{s_y^2}(y-\bar{y})$$

- We define **sample correlation coefficient** as: $r=\dfrac{s_{xy}}{s_x s_y}$
- The least-squared regression lines are now:

$$\frac{y-\bar{y}}{s_y}=r\frac{(x-\bar{x})}{s_x} \quad \text{and} \quad \frac{x-\bar{x}}{s_x}=r\frac{(y-\bar{y})}{s_y}$$

- Interesting to note that if the two regression lines are written as $y=ax+b$, $x=c+dy$. Then we have

$$bd=r^2$$

Standard Error of Estimate

- **Standard Error of estimate** *y* on *x*

$$S_{y,x} = \sqrt{\frac{\sum (y - y_{est})^2}{n}}$$

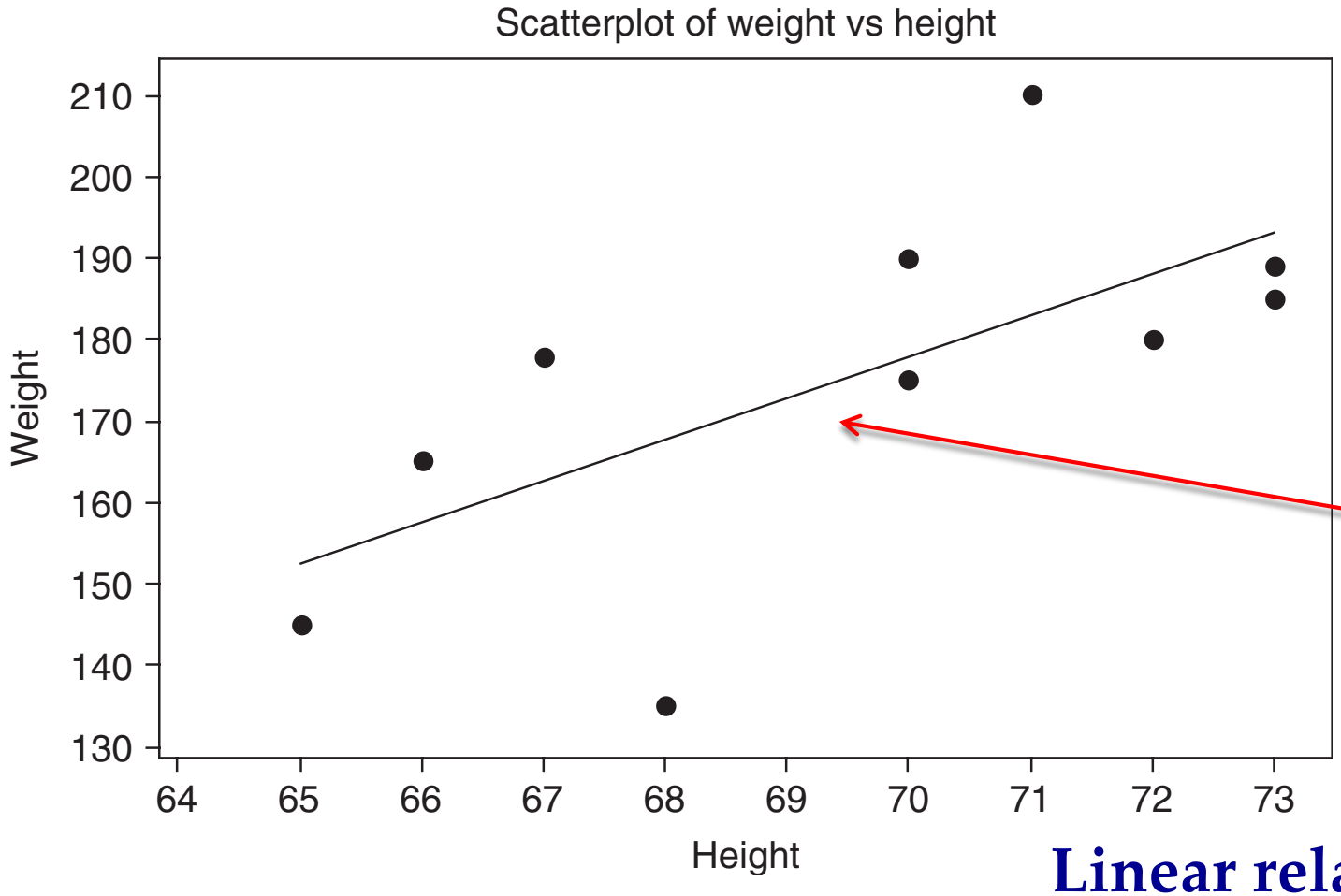
physical meaning

MORE ON CURVE FITTING & METHOD OF LEAST SQUARES

John C.S. Lui, CSE Department, CUHK

Curve Fitting

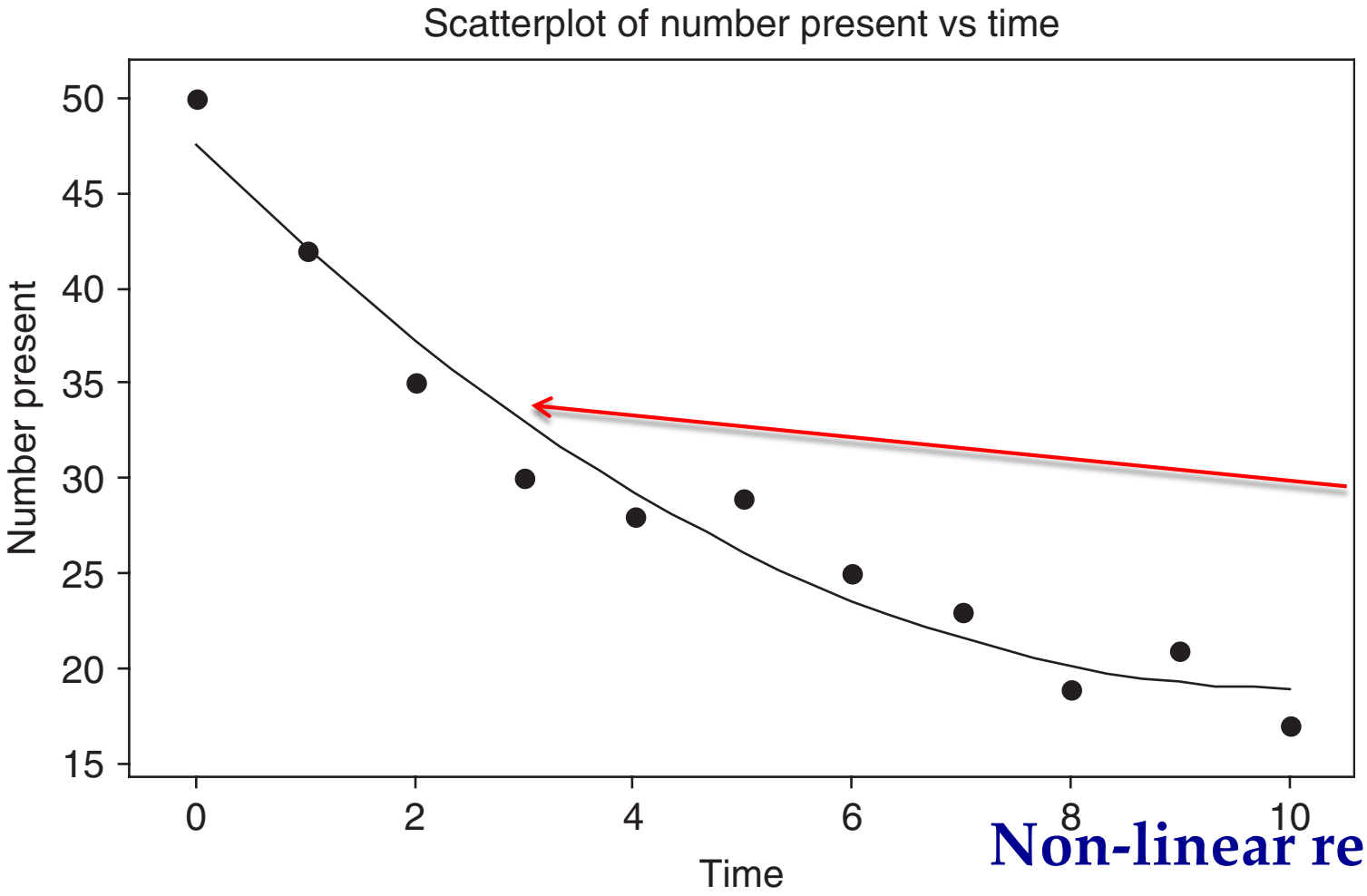
- We have collection of *n* points: $(x_1,y_1),\ldots, (x_n, y_n)$
- Scatter plot to determine *relationship*



Linear relationship

Curve Fitting

- We have collection of *n* points: $(x_1,y_1),\ldots, (x_n, y_n)$
- Scatter plot to determine *relationship*



Equations of Approximating Curves

- Polynomial of the *n*th degrees

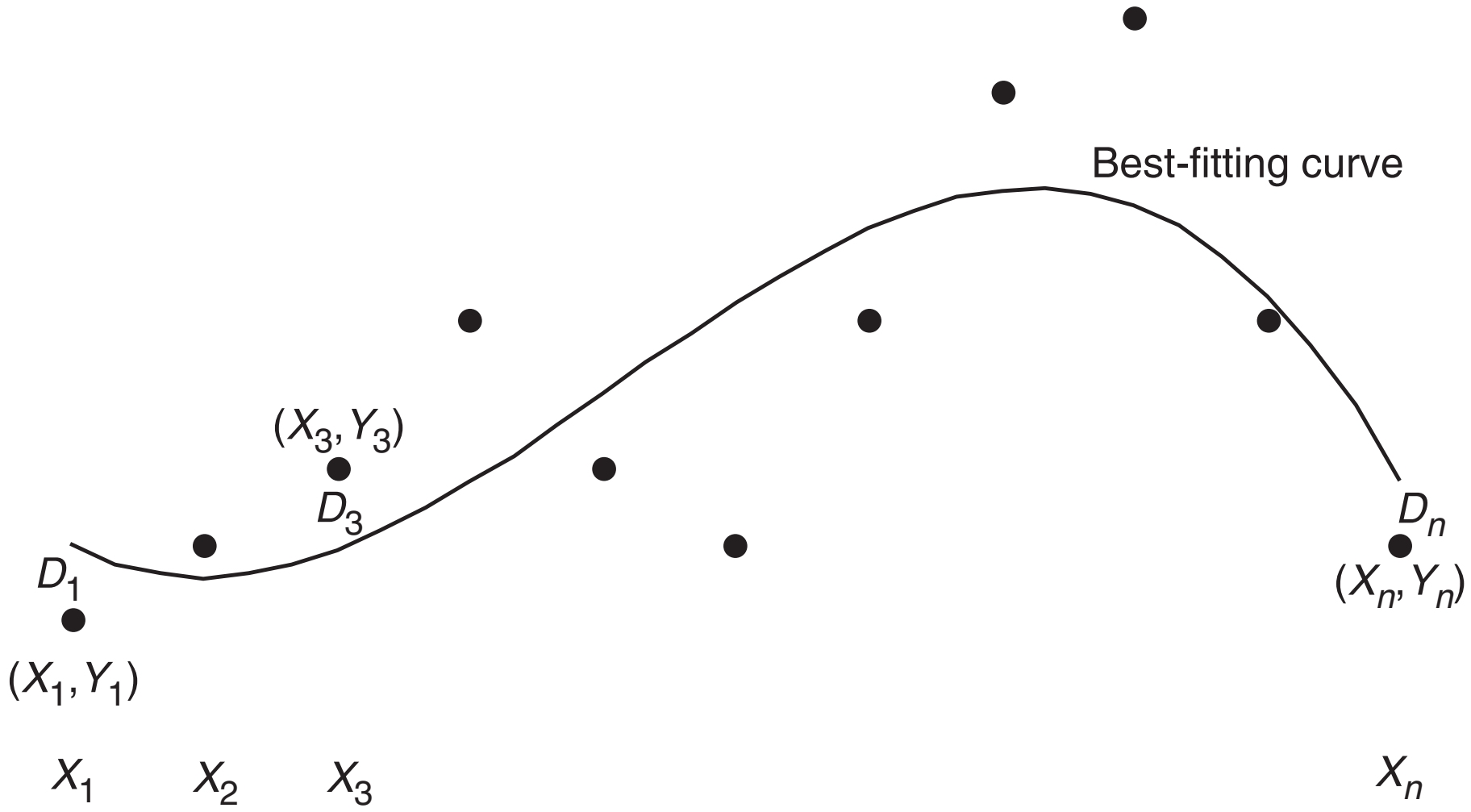
Straight line	$Y = a_0 + a_1 X$	**linear**
Parabola, or quadratic curve	$Y = a_0 + a_1 X + a_2 X^2$	
Cubic curve	$Y = a_0 + a_1 X + a_2 X^2 + a_3 X^3$	
Quartic curve	$Y = a_0 + a_1 X + a_2 X^2 + a_3 X^3 + a_4 X^4$	
*n*th-Degree curve	$Y = a_0 + a_1 X + a_2 X^2 + \cdots + a_n X^n$	

Equations of Approximating Curves

- Other possible curves

Hyperbola	$Y = \dfrac{1}{a_0 + a_1 X}$ or $\dfrac{1}{Y} = a_0 + a_1 X$
Exponential curve	$Y = ab^X$ or $\log Y = \log a + (\log b)X = a_0 + a_1 X$
Geometric curve	$Y = aX^b$ or $\log Y = \log a + b(\log X)$
Modified exponential curve	$Y = ab^X + g$
Modified geometric curve	$Y = aX^b + g$
Gompertz curve	$Y = pq^{b^X}$ or $\log Y = \log p + b^X(\log q) = ab^X + g$
Modified Gompertz curve	$Y = pq^{b^X} + h$
Logistic curve	$Y = \dfrac{1}{ab^X + g}$ or $\dfrac{1}{Y} = ab^X + g$
	$Y = a_0 + a_1(\log X) + a_2(\log X)^2$

Illustration of "best fit" curve



The Least-Square Line

- Use ***linear function*** $Y=a_0 + a_1X$ to approximate points
- As shown previous, we have

$$a_0 = \frac{(\sum Y)(\sum X^2) - (\sum X)(\sum XY)}{N\sum X^2 - (\sum X)^2} \qquad a_1 = \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum X^2 - (\sum X)^2}$$

Nonlinear Equations Reduced to Linear Form

Table 13.13 gives experimental values of the pressure P of a given mass of gas corresponding to various values of the volume V. According to thermodynamic principles, a relationship having the form $PV^{\gamma} = C$, where γ and C are constants, should exist between the variables.

(*a*) Find the values of γ and C.

(*b*) Write the equation connecting P and V.

(*c*) Estimate P when $V = 100.0\,\text{in}^3$.

Table 13.13

Volume V in cubic inches (in^3)	54.3	61.8	72.4	88.7	118.6	194.0
Pressure P in pounds per square inch (lb/in^2)	61.2	49.2	37.6	28.4	19.2	10.1

Solution

Since $PV^{\gamma} = C$, we have

$$\log P + \gamma \log V = \log C \qquad \text{or} \qquad \log P = \log C - \gamma \log V$$

Calling $\log V = X$ and $\log P = Y$, the last equation can be written

$$Y = a_0 + a_1 X$$

where $a_0 = \log C$ and $a_1 = -\gamma$.

Table 13.14

$X = \log V$	$Y = \log P$	X^2	XY
1.7348	1.7868	3.0095	3.0997
1.7910	1.6946	3.2077	3.0350
1.8597	1.5752	3.4585	2.9294
1.9479	1.4533	3.7943	2.8309
2.0741	1.2833	4.3019	2.6617
2.2878	1.0043	5.2340	2.2976
$\sum X = 11.6953$	$\sum Y = 8.7975$	$\sum X^2 = 23.0059$	$\sum XY = 16.8543$

Solution

$$\sum Y = a_0 N + a_1 \sum X \qquad \text{and} \qquad \sum XY = a_0 \sum X + a_1 \sum X^2$$

from which

$$a_0 = \frac{(\sum Y)(\sum X^2) - (\sum X)(\sum XY)}{N \sum X^2 - (\sum X)^2} = 4.20 \qquad a_1 = \frac{N \sum XY - (\sum X)(\sum Y)}{N \sum X^2 - (\sum X)^2} = -1.40$$

$$\text{Thus } Y = 4.20 - 1.40X.$$

(*a*) Since $a_0 = 4.20 = \log C$ and $a_1 = -1.40 = -\gamma$, $C = 1.60 \times 10^4$ and $\gamma = 1.40$.

(*b*) The required equation in terms of P and V can be written $PV^{1.40} = 16{,}000$.

(*c*) When $V = 100$, $X = \log V = 2$ and $Y = \log P = 4.20 - 1.40(2) = 1.40$.
Then $P = \text{antilog}\, 1.40 = 25.1\ \text{lb/in}^2$.

The Least-Squares Parabola

- Given *N* points, we use equation: $Y = a_0 + a_1 X + a_2 X^2$
- To find these 3 unknowns, we do:

$$\begin{aligned}
\sum Y &= a_0 N + a_1 \sum X + a_2 \sum X^2 \\
\sum XY &= a_0 \sum X + a_1 \sum X^2 + a_2 \sum X^3 \\
\sum X^2 Y &= a_0 \sum X^2 + a_1 \sum X^3 + a_2 \sum X^4
\end{aligned}$$

- **Remember**: multiply equation by *1*, *X*, X^2, and sum all *N* equations.
- We can use similar technique for
  - Least-squares cubic curves
  - Least-squares quartic curves

Curving Fitting for More than Two Variables

- *N* points of: $(X_1, Y_1, Z_1),\ (X_2, Y_2, Z_2), \ldots, (X_N, Y_N, Z_N)$
- Linear equation: $Z = a_0 + a_1 X + a_2 Y$
- Use *least-squares plane* approximating the data
- The equations we need to solve are:

$$\sum Z = a_0 N + a_1 \sum X + a_2 \sum Y$$

$$\sum XZ = a_0 \sum X + a_1 \sum X^2 + a_2 \sum XY$$

$$\sum YZ = a_0 \sum Y + a_1 \sum XY + a_2 \sum Y^2$$

- **Remember**: multiply by *1*, *X* and *Y*, *then do summation*

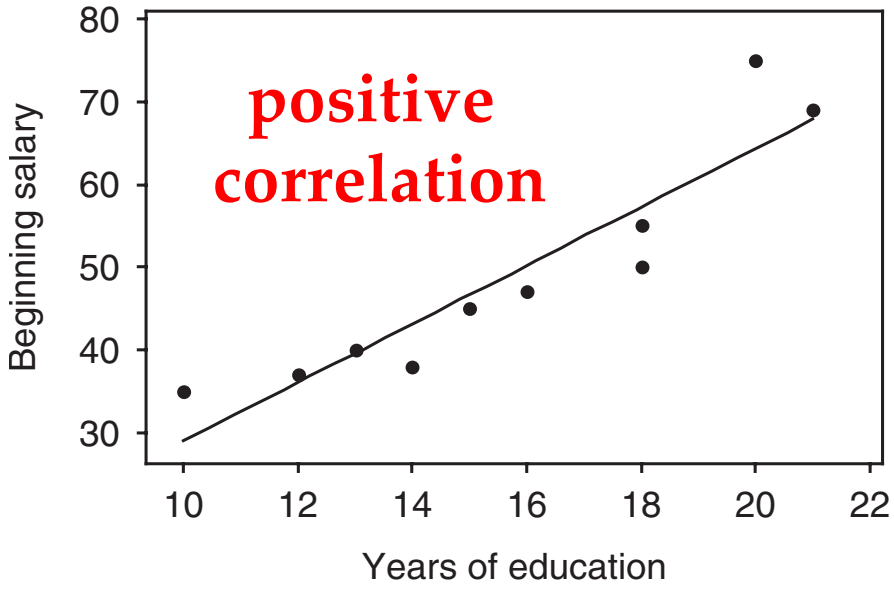
CORRELATION THEORY

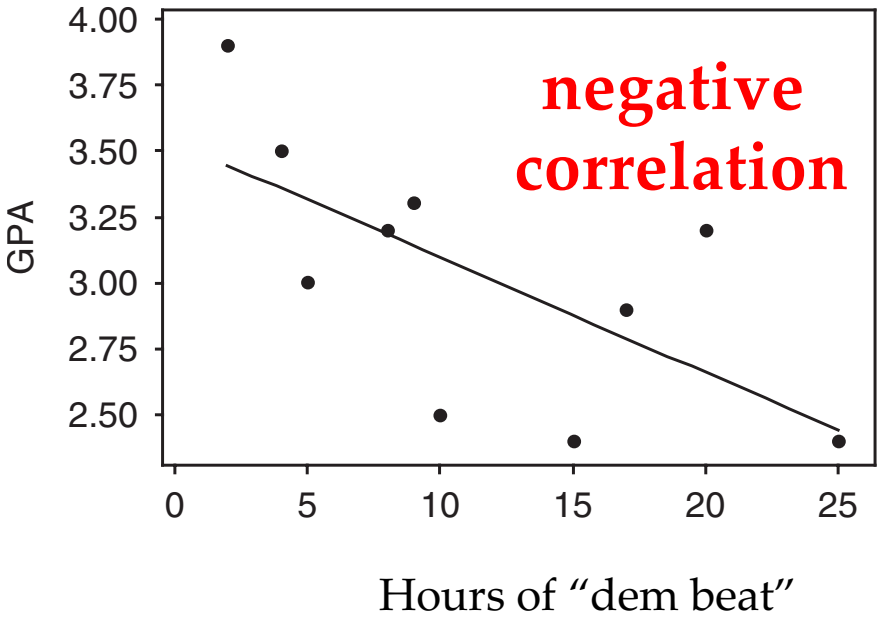
John C.S. Lui, CSE Department, CUHK

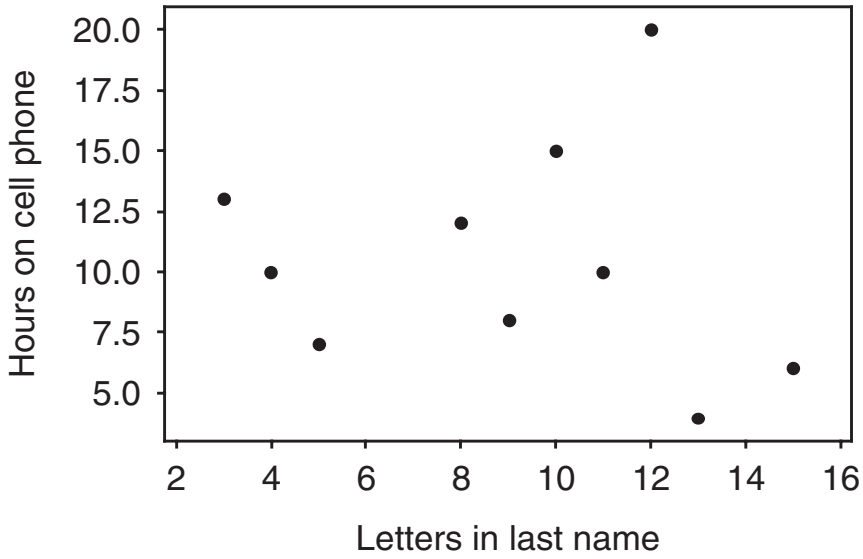
Correlation and Regression

- Previously, we considered ***regression***, or *estimation*, of one dependent variable on one or more independent variables.
- Now we consider ***correlation***, or *the degree of relationship* between variables, so we know ***how well*** a linear or other equation describes or explains the relationship between variables

Linear Correlation







no correlation

Give examples

Measure of Correlation

- Given N points $(X_1, Y_1), \ldots, (X_N, Y_N)$. We consider two linear regression lines

$$Y = a_0 + a_1 X \quad (1)$$

$$a_0 = \frac{(\sum Y)(\sum X^2) - (\sum X)(\sum XY)}{N \sum X^2 - (\sum X)^2}$$

$$a_1 = \frac{N \sum XY - (\sum X)(\sum Y)}{N \sum X^2 - (\sum X)^2}$$

$$X = b_0 + b_1 Y \quad (2)$$

$$b_0 = \frac{(\sum X)(\sum Y^2) - (\sum Y)(\sum XY)}{N \sum Y^2 - (\sum Y)^2}$$

$$b_1 = \frac{N \sum XY - (\sum X)(\sum Y)}{N \sum Y^2 - (\sum Y)^2}$$

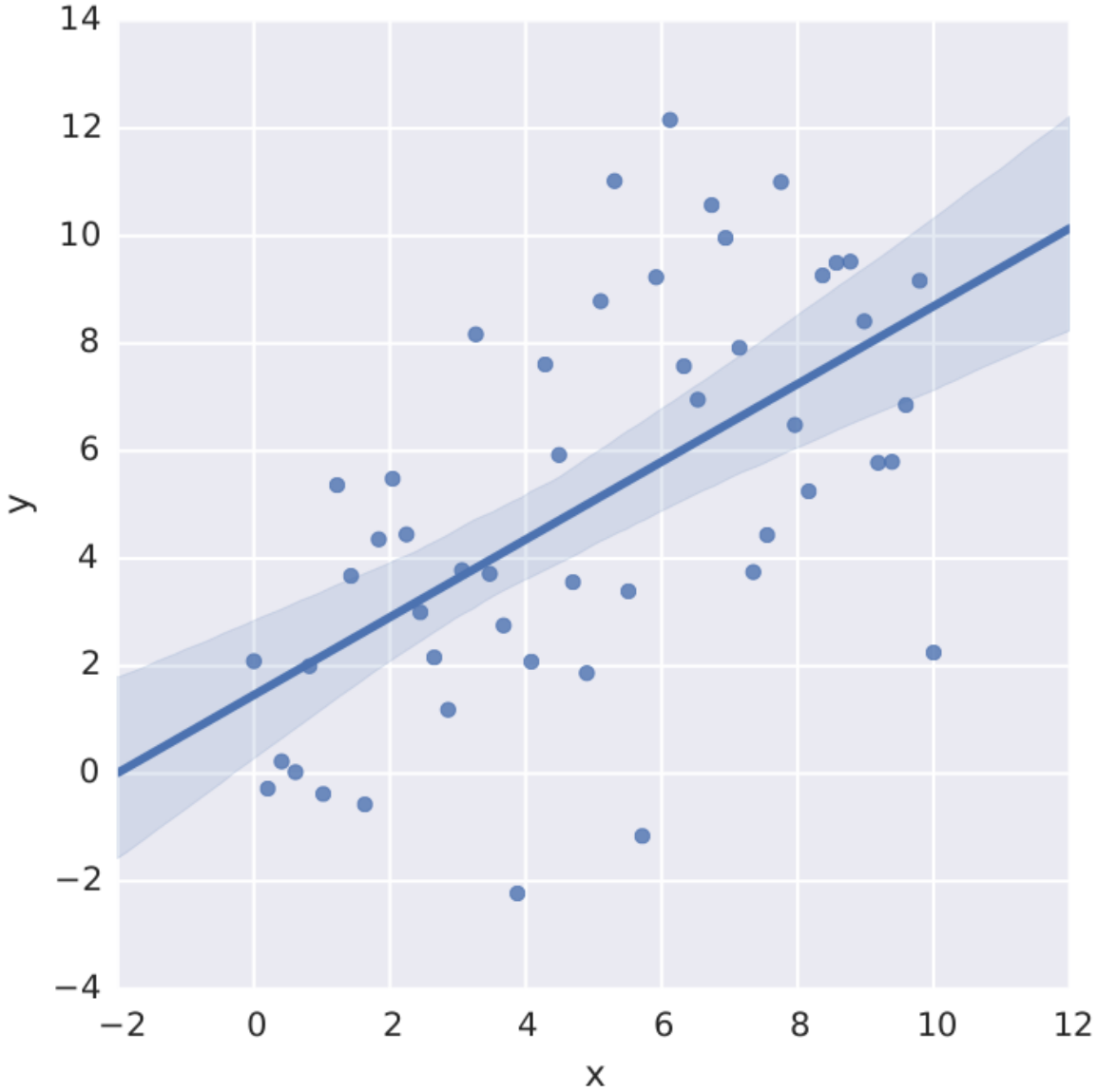
- Let Y_{est} (X_{est}) be the value of Y (X) according to Eq. (1) (2), the **standard error of estimate** **are:**

$$s_{Y.X} = \sqrt{\frac{\sum (Y - Y_{\text{est}})^2}{N}} \qquad s_{X.Y} = \sqrt{\frac{\sum (X - X_{\text{est}})^2}{N}}$$

Property of Standard Error of Estimate

- The standard error of estimate has properties analogous to those of the standard deviation.
- For example, if we construct lines **parallel** to the regression line of *Y* on *X* at respective vertical distances, $s_{Y,X}$, $2s_{Y,X}$, $3s_{Y,X}$ from it, we should find (**if *N* is large enough**), that there would be included between these lines about 68%, 95%, 99.7% of the sample points.

Property of Standard Error of Estimate



Explained and Unexplained Variation

- The ***total variation*** of Y is defined as $\sum(Y-\bar{Y})^2$
- It can be written as

$$\sum(Y-\bar{Y})^2 = \sum(Y-Y_{est})^2 + \sum(Y_{est}-\bar{Y})^2$$

- The first term is the *unexplained variation*
- The 2^{nd} term is the *explained variation* because $(Y_{est}-\bar{Y})$ have a definite pattern, but $(Y-Y_{est})$ behave in a random or unpredictable manner

Coefficient of Correlation

- **Coefficient of determination** is the ratio of the ***explained variation*** to ***total variation***
- When explained variation is zero, the ratio is *0*
- When unexplained variation is zero, the ratio is *1*
- For all other cases, the ratio is: $0 < r < 1$.
- Since the ratio is non-negative, we denote it as r^2
- **Correlation coefficient** is

$$r = \pm\sqrt{\frac{\text{explained variation}}{\text{total variation}}} = \pm\sqrt{\frac{\sum (Y_{\text{est}} - \bar{Y})^2}{\sum (Y - \bar{Y})^2}}$$

- When $r = +1$ *(-1)*, it means positive (negative) linear correlation

Coefficient of Correlation

- The standard deviation of Y is $$s_Y = \sqrt{\frac{\sum (Y - \bar{Y})^2}{N}}$$
- We can express r (disregarding the sign) as $$r = \sqrt{1 - \frac{s_{Y.X}^2}{s_Y^2}}$$
- The metric r is a very good measure of the linear correlation between two variables X and Y

Remark about Correlation Coefficient

- The correlation coefficient *r* is in fact, very general can can be used for nonlinear relationships as well.
- The only difference is that Y_{est} is computed from a nonlinear regression equation, and the + and – signs are omitted.
- If we have: $Y = a_0 + a_1 X + a_2 X^2 + \cdots + a_{n-1} X^{n-1}$

  then we have: $s_{Y.X}^2 = \dfrac{\sum Y^2 - a_0 \sum Y - a_1 \sum XY - \cdots - a_{n-1} \sum X^{n-1} Y}{N}$
- The standard error of estimate is $\hat{s}_{Y.X} = \sqrt{\dfrac{N}{N-n}}\, s_{Y.X}$
- We can express *r*:

$$r = \sqrt{1 - \frac{s_{Y.X}^2}{s_Y^2}}$$

Remark about Correlation Coefficient

- Note that *r* measures the degree of the relationship.
- **If linear equation is assumed and *r* is near zero**, it means *no linear correlation* between variable, but it does not mean there is no correlation (possibly non-linear correlation)
- In other words, *r* measures the "goodness of fit" between the equation and the data
- High value of *r* does not necessary indicate a direct dependence of the variables (e.g., # of books published per year and # of thunderstorms each year). This is called ***spurious correlations***.

What is the use of *r* ?

- With all these formula for the correlation coefficient, what is its use?
- In a nutshell, we can see how an independent variable (*X*) may "impact" (or correlated) with the dependent variable (*Y*).
- **For example:**
  - if $Y = f(X_1, X_2, X_3, \ldots, X_{10})$
  - After computation, if $r_{X1,Y}$, $r_{X2,Y}$,…$r_{X3,Y}$ are high values but others are close to zero, we should focus on X_1, X_2, X_3 instead
  - It also helps us to do "***features selection***"

MULTIPLE AND PARTIAL CORRELATION THEORY

John C.S. Lui, CSE Department, CUHK

Multiple & Partial Correlation

- The degree of relationship existing between three or more variables is called ***multiple correlation***
- We use the following notations: X_1, X_2,... are variables under consideration. X_{11}, X_{12}, X_{13},... denote values of X_1, and X_{21}, X_{22}, X_{23},... are values of X_2
- We write:

$$\sum_{j=1}^{N} X_{2j},\ \sum_{j} X_{2j},\ \text{or simply}\ \sum X_2$$

Regression Equations & Regression Planes

- Regression equation is an estimation of dependent variable, say X_1, from independent variables, X_2, X_3,..
- We have: $X_1 = F(X_2, X_3, \ldots)$
- Example of 3 variables: $X_1 = b_{1.23} + b_{12.3}X_2 + b_{13.2}X_3$
- **Partial regression coefficients:**
  - X_1 on X_2: $b_{12.3}$
  - X_1 on X_3: $b_{13.2}$
- The regression in the above example is the ***linear regression equation*** of X_1 on X_2 and X_3. It represents a plane called a *regression plane*

Normal Equations For Least-Squares Regression Plane

- **Least-squares regression planes** of *N* data points (X_1, X_2, X_3):

$$\sum X_1 = b_{1.23}N + b_{12.3}\sum X_2 + b_{13.2}\sum X_3$$

$$\sum X_1X_2 = b_{1.23}\sum X_2 + b_{12.3}\sum X_2^2 + b_{13.2}\sum X_2X_3$$

$$\sum X_1X_3 = b_{1.23}\sum X_3 + b_{12.3}\sum X_2X_3 + b_{13.2}\sum X_3^2$$

- We obtain by multiplying by *1*, X_2 and X_3
- Define $x_1 = X_1 - X_1$, $x_2 = X_2 - X_2$, and $x_3 = X_3 - X_3$, the regression equation is: $x_1 = b_{12.3}x_2 + b_{13.2}x_3$

where $b_{12.3}$ and $b_{13.2}$

$$\sum x_1x_2 = b_{12.3}\sum x_2^2 + b_{13.2}\sum x_2x_3$$

$$\sum x_1x_3 = b_{12.3}\sum x_2x_3 + b_{13.2}\sum x_3^2$$

Regression Planes and Correlation Coefficients

- Define linear correlation coefficient r_{12}, r_{13}, r_{23} as before. For example, r_{12} can be defined when we treat X_3 as constant
- Standard Error of Estimate:

$$s_{1.23} = \sqrt{\frac{\sum (X_1 - X_{1,\text{est}})^2}{N}} = s_1 \sqrt{\frac{1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}}$$

- Coefficient of Multiple Correlation:

$$R_{1.23} = \sqrt{1 - \frac{s_{1.23}^2}{s_1^2}} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}}$$

Q-Q PLOT

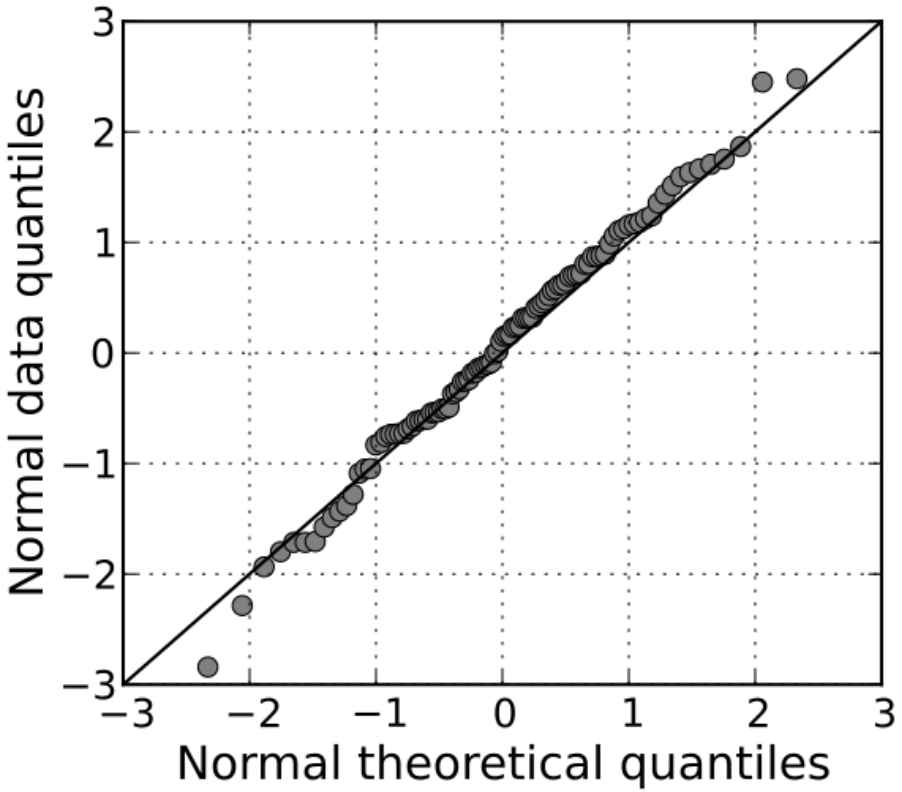
John C.S. Lui, CSE Department, CUHK

Q-Q Plot

- A graphical method for comparing two probability distributions by plotting their quantiles against each other
- If two distributions (e.g., data vs. model) being compared are ***similar***, the Q-Q plot will line on $y=x$
- If the distributions are *linearly related*, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line $y = x$
- **Construction of a Q-Q Plot**
  - The set of intervals for the quantiles is chosen
  - A point (x,y) on the plot corresponds to one of the quantiles of the second distribution (y-coordinate) plotted against the same quantile of the first distribution (x-coordinate)

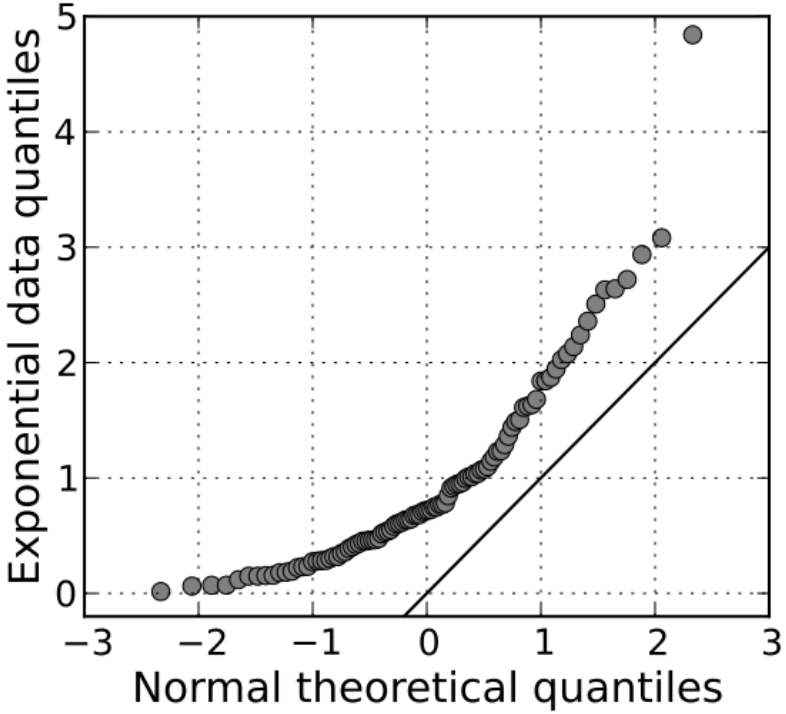
Example of Q-Q Plot

- A normal Q–Q plot comparing randomly generated, independent standard normal data on the vertical axis to a standard normal population on the horizontal axis.
- The linearity of the points suggests that the data are **normally distributed**



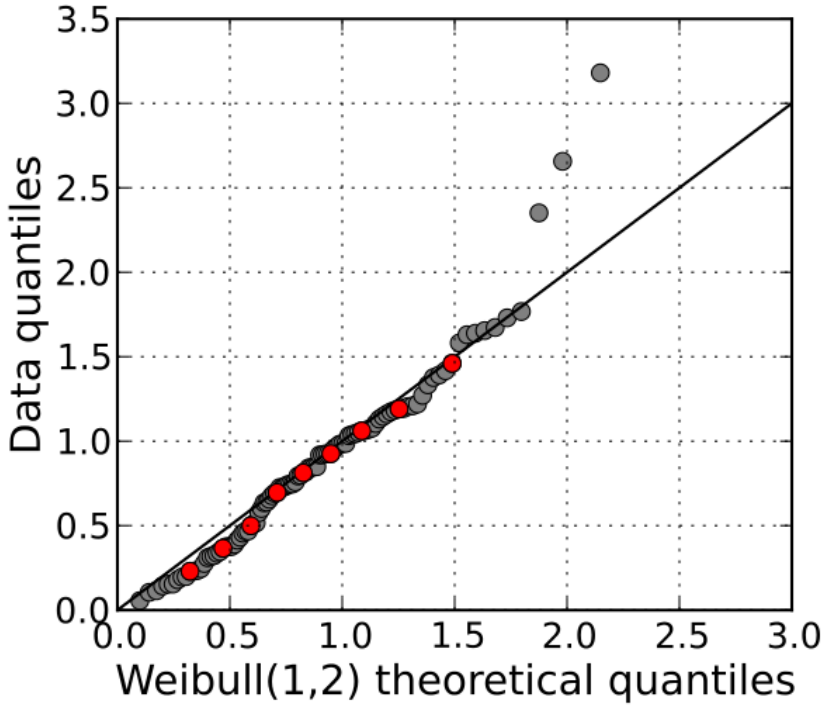
Example of Q-Q Plot

- A normal Q–Q plot of randomly generated, independent standard exponential data, (X ~ Exp(1)). This Q–Q plot compares a sample of data on the vertical axis to a statistical population on the horizontal axis.
- The points follow a strongly nonlinear pattern, suggesting that the data are ***not distributed as a standard normal*** *(X ~ N(0,1))*
- The offset between the line and the points suggests that the mean of the data is not 0.
- The median of the points can be determined to be near 0.7



Example of Q-Q Plot

- A Q–Q plot of a sample of data versus a Weibull distribution
- The deciles of the distributions are shown in red
- Three outliers are evident at the high end of the range
- Otherwise, the data fit the Weibull(1,2) model well



$$F(x) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

k = shape parameter
λ = scale parameter

More on Q-Q plot:

https://www.youtube.com/watch?v=X9_ISJ0YpGw