Foundations of Data Science

Lecture 2: Statistics

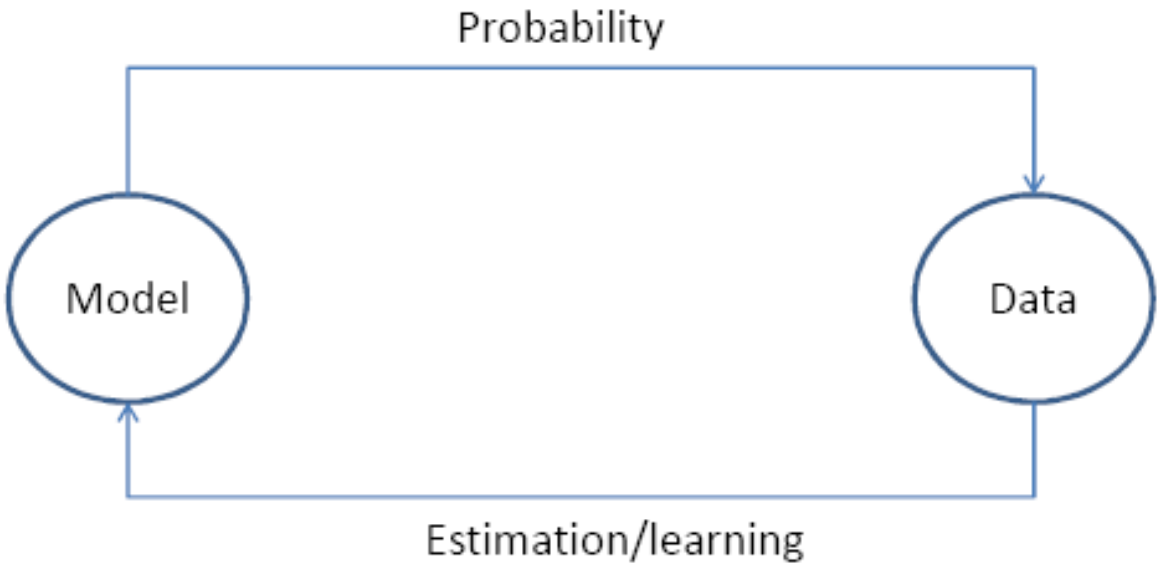
MING GAO

SE & DaSE @ ECNU
(for course related communications)
mgao@sei.ecnu.edu.cn

Sep. 23, 2016

Outline

Big picture of probability theory and statistics



Statistics Review

Overview

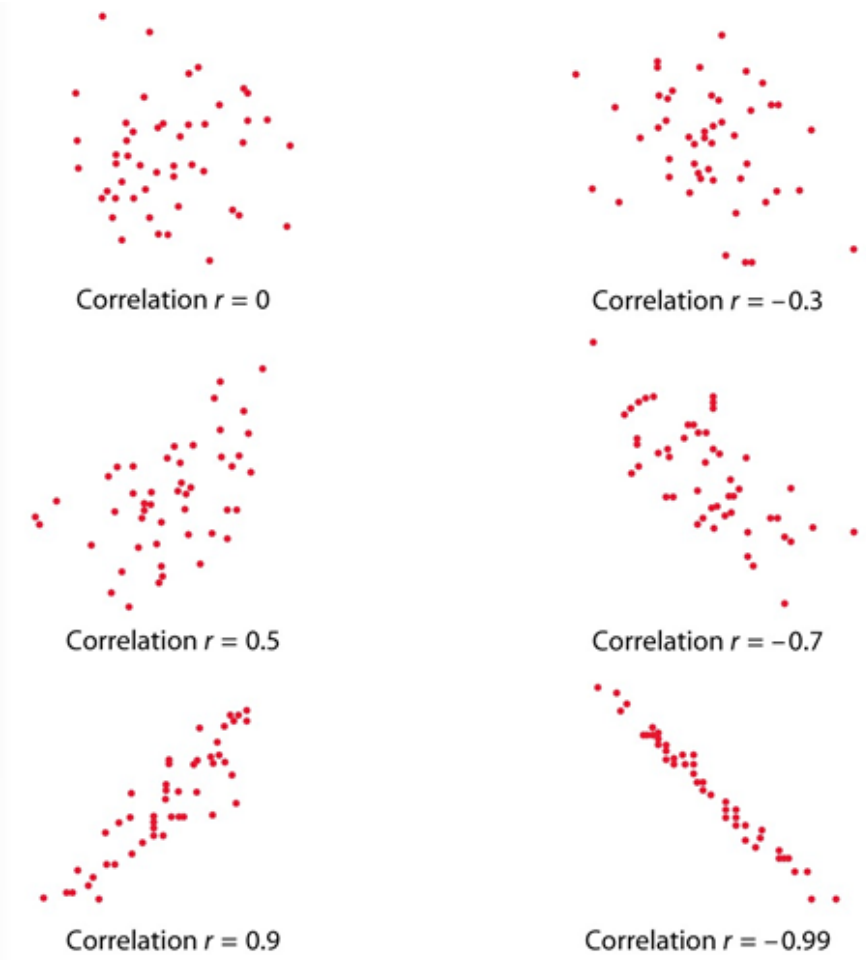
Statistics: Set of methods for collecting/analyzing data (the art and science of learning from data)

- Description: Graphical and numerical methods for summarizing the data
- Estimation: Learning parameters or distribution characteristics from data
- Inference: Methods for making predictions about a population (total set of subjects of interest), based on a sample (subset of the sample on which study collects data)

Numerical descriptions

- Let X denote a quantitative variable, with observations $x_1, x_2 \cdots, x_n$
- Describing the center
  - Mean: $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - Median: middle measurement of ordered sample
  - Mode: the value that appears most often in a set of sample
- Describing variability
  - Range: difference between largest and smallest observations
  - Variance: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$ (not n, due to technical reasons)
  - Standard deviation: $s = \sqrt{s^2}$
- Measures of position: p% quartile
  - p percent of observations below it, (100 - p)% above it
  - 50% quartile = median
  - Box plot: Minimum, 25% Q, Median, 50%, Maximum

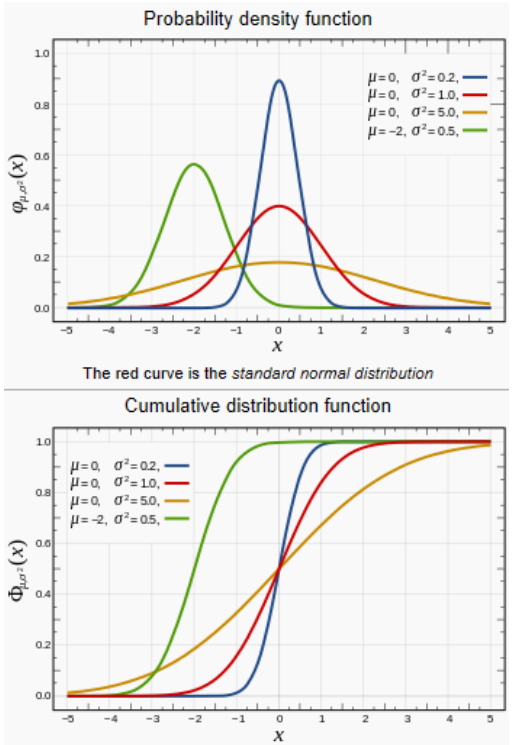
Correlation



Definition

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2 \sum_{i=1}^{n}(y_i-\overline{y})^2}}$$

Normal distribution



Normal distribution

Most important probability distribution

$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} exp^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Symmetric, bell-shaped
- Characterized by mean (μ) and standard deviation (σ), representing center and spread
- An individual observation from an approximately normal distribution has probability
  - $P(\mu - \sigma \leq \mu + \sigma) \approx 0.68$
  - $P(\mu - 2\sigma \leq \mu + 2\sigma) \approx 0.95$
  - $P(\mu - 3\sigma \leq \mu + 3\sigma) \approx 0.997$

Central limit theorem

Theorem

For random sampling with "large" n, the sampling distribution of the sample mean $\overline{x}$ is approximately a normal distribution.

- $E(\overline{x}) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = E(X)$
- $Var(\overline{x}) = \frac{1}{n^2}\sum_{i=1}^{n} Var(x_i) = \frac{Var(X)}{n}$
- How "large" n needs to be depends on skew of population distribution, but usually $n \geq 30$ sufficient
- For example, you plan to randomly sample 100 students to estimate population proportion who have selected a course. $P(x = 1) = p$, $P(x = 0) = 1 - p$, $\overline{x} = \sum_{i=1}^{100} x_i$, then $\overline{x} \sim N(p, \frac{p(1-p)}{100})$.

Estimation

Goal

How can we use sample data to estimate values of population parameters?

- Point estimate: A single statistic value that is the best guess for the parameter value
  - Sample mean estimates population mean μ, $\widehat{\mu} = \overline{x}$
  - Sample std. dev. estimates population std. dev. σ, $\widehat{\sigma} = s$
  - Properties of good estimators: unbiased and efficient (smallest possible standard error)
- Interval estimate: an interval of numbers around the point estimate, that has a fixed "confidence level" of containing the parameter value. Called a confidence interval.
  - A confidence interval (CI) is an interval of numbers believed to contain the parameter value in form **point estimate $\pm$ margin of error**
  - For example, margin of error 2(standard error) for 95% confidence
  - $P(\mu - 1.96\widehat{\sigma} \leq \overline{x} \leq \mu + 1.96\widehat{\sigma}) \approx 0.95$, thus greater sample size gives narrower CI (we can further infer the sample size given margin of error)

Maximum likelihood estimation

Goal

Finding the parameter values that maximizes the likelihood

- Suppose there is a sample $x_1, x_2, \cdots, x_n$ of n *i.i.d.* observations coming from a distribution with an unknown probability density function $f(\cdot)$.
- We first specifies the joint density function for all observations as $\mathcal{L}(\theta|x_1, x_2, \cdots, x_n) = f(x_1, x_2, \cdots, x_n|\theta) = \Pi_{i=1}^{n} f(x_i|\theta)$
- In practice it is often more convenient to work with the natural logarithm of the likelihood function, called the log-likelihood $\ln \mathcal{L} = \sum_{i=1}^{n} \ln f(x_i|\theta)$
- Maximum likelihood estimator (MLE) $\widehat{\theta} = \arg\max_{\theta \in \Theta} \ln \mathcal{L}(\theta|x_1, x_2, \cdots, x_n)$

MLE Cont.

Example

For a sample observing from a normal distribution

$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} exp^{-\frac{(x-\mu)^2}{2\sigma^2}}$

- $\ln \mathcal{L}(\mu,\sigma) = -\frac{n}{2}\ln 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$
-

$$\frac{\partial}{\partial\mu}\ln \mathcal{L}(\mu,\sigma) = 0$$
$$\frac{\partial}{\partial\sigma}\ln \mathcal{L}(\mu,\sigma) = 0$$

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

Expectation maximization algorithm

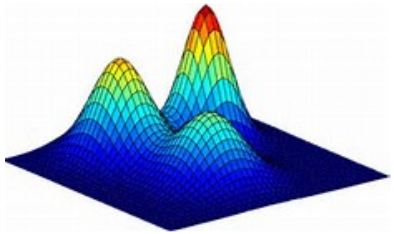
EM algorithm

The EM algorithm is used to find (locally) maximum likelihood parameters of a statistical model in cases where the equations cannot be solved directly

- Missing value
- Likelihood contain latent variables

- It consists of two steps: E-step (expectation) and M-step (maximization)
- It works in an iterative manner
- For example: Gaussian mixed model
  - Let $x = \{x_1, x_2, \cdots, x_n\}$ be a sample of n independent observations from a mixture of two multivariate normal distribution
  - Let $z = \{z_1, z_2, \cdots, z_n\}$ be the latent variables that determine the component from which the observation originates
  - $X_i|(Z_i = 1) \sim N(\mu_1, \Sigma_1)$ and $X_i|(Z_i = 2) \sim N(\mu_2, \Sigma_2)$

EM Cont.



GMM

- $P(Z_i = 1) = \tau_1,\ P(Z_i = 2) = \tau_2(1 - \tau_1)$, and $\theta = (\tau, \mu_1, \mu_2, \Sigma_1, \Sigma_2)$
- $f(x|\theta) = \sum_{i=1}^{2} \mathbb{I}_{z=j}\tau_i f(x|\mu_i, \Sigma_i)$
- $\ln \mathcal{L}(\theta; x, z) = \Pi_{i=1}^{n} \sum_{j=1}^{2} \mathbb{I}_{z_i=j}\tau_j f(x_i|\mu_j, \Sigma_j)$

E-step

$$T_{j,i}^{(k+1)} = P(Z_i = j|x_i; \theta^(k)) = \frac{P(Z_i = j, x_i|\theta^(k))}{P(x_i|\theta^(k))}$$

$$= \frac{\tau_j^{(k)} f(x_i|\mu_j^{(k)}, \Sigma_j^{(k)})}{\sum_{j=1}^{2} \tau_j^{(k)} f(x_i|\mu_j^{(k)}, \Sigma_j^{(k)})}$$

EM Cont.

E-step Cont.

$$Q(\theta|\theta^k) = E_{Z|X,\theta^k} \ln \mathcal{L}(\theta; x, z) = \sum_{i=1}^{n} E_{Z|X,\theta^k} \ln \mathcal{L}(\theta; x_i, z_i)$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{2} P(Z_i = j|x_i; \theta^{(k)}) \ln \mathcal{L}(\theta_j; x_i, z_i) = \sum_{i=1}^{n}\sum_{j=1}^{2} T_{j,i}^k \ln \mathcal{L}(\theta_j; x_i, z_i)$$

M-step (for j = 1, 2)

$$\tau_j^{k+1} = \frac{\sum_{i=1}^{n} T_{j,i}^k}{\sum_{i=1}^{n}\sum_{j=1}^{2} T_{j,i}^k}$$

$$\mu_j^{k+1} = \frac{\sum_{i=1}^{n} T_{j,i}^k x_i}{\sum_{i=1}^{n} T_{j,i}^k}$$

$$\Sigma_j^{k+1} = \frac{\sum_{i=1}^{n} T_{j,i}^k (x_i - \mu_j^{k+1})(x_i - \mu_j^{k+1})^T}{\sum_{i=1}^{n} T_{j,i}^k}$$

Significance tests

Goal

We answer a question such as, "If the hypothesis were true, would it be unlikely to get data such as we obtained?"

- Spending money on other people has a more positive impact on happiness than spending money on oneself.
- Mental health tends to be better at higher levels of socioeconomic status (SES)

Definition

A significance test uses data to evaluate a hypothesis by comparing sample point estimates of parameters to values predicted by the hypothesis

- Null hypothesis (H_0): A statement that parameter(s) take specific value(s) (Usually: no effect)
- Alternative hypothesis (H_1): states that parameter value(s) falls in some alternative range of values (an effect)

Significance tests

Goal

We answer a question such as, "If the hypothesis were true, would it be unlikely to get data such as we obtained?"

- Spending money on other people has a more positive impact on happiness than spending money on oneself.
- Mental health tends to be better at higher levels of socioeconomic status (SES)

Definition

A significance test uses data to evaluate a hypothesis by comparing sample point estimates of parameters to values predicted by the hypothesis

- Null hypothesis (H_0): A statement that parameter(s) take specific value(s) (Usually: no effect)
- Alternative hypothesis (H_1): states that parameter value(s) falls in some alternative range of values (an effect)

How to do a test?

Steps

- Test statistic: compares data to what null hypo. H_0 predicts, often by finding the number of standard errors between sample point estimate and H_0 value of parameter
- P-value (P): a probability measure of evidence about H_0. The probability (under presumption that H_0 true) the test statistic equals observed value or value even more extreme in direction predicted by H_1.
- Conclusion:
  - If no decision needed, report and interpret P-value
  - If decision needed, select a cutoff point (such as 0.05 or 0.01, namely confidence level) and reject H_0 if P-value $P \leq$ that value
  - If the P-value is not sufficiently small, we fail to reject H_0

Significance test for mean

Steps

- Assumptions: randomization, quantitative variable, normal population distribution (robustness?)
- Null Hypothesis: $H_0 : \mu = \mu_0$ where μ_0 is particular value for population mean (typically no effect or no change from a standard)
- Alternative hypothesis: $H_1 : \mu \neq \mu_0$ (2-sided alternative)
- Test statistic: the number of standard errors that the sample mean falls from the H_0 value $t = \frac{\bar{x} - \mu_0}{se}$, where $se = \frac{s}{\sqrt{n}}$
- When H_0 is true, the T-test statistic is a T-distribution with $df = n - 1$. Under presumption that H_0 true, probability the T-test statistic equals observed value or even more extreme (larger in absolute value), providing stronger evidence against H_0
- Conclusion: report and interpret P-value. If needed, make decision about H_0.

Example: Anorexia study

Weight measured before and after period of treatment with "family therapy"

Procedure

- y = weight change (end - beginning) ($\{$11.4, 11.0, 5.5, 9.4, 13.6, -2.9, -0.1, 7.4, 21.5, -5.3, -3.8, 13.4, 13.1, 9.0, 3.9, 5.7, 10.7$\}$)
- Let μ = population mean weight change, $H_0 : \mu = 0$ (no effect) against where μ_0 is particular value for population mean (typically "no effect" or "no change" from a standard)
- T-test result in R: t.test(y,mu=0) or t.test(y1, y2, paired = true)
  $t = 4.1849$, $df = 16$, p-value = 0.0007003
  95% confidence interval: 3.58470 10.94471
  sample estimates: mean of y : 7.264706
- Conclusion: very strong evidence that the population mean differs from 0 (i.e., "family therapy" is helpful)

Example: Anorexia study Cont.

Equivalence between result of significance test and result of CI

- When P-value < 0.05 in two-sided test, 95% CI for μ does not contain H_0 value of μ_0 (such as 0)
- Example: $P = 0.0007$, 95% CI was $(3.6, 10.9)$
- When P-value > 0.05 in two-sided test, 95% CI necessarily contains H_0 value of μ
- CI has more information about actual value of μ
- Suppose sample mean $= 7.265$, $s = 7.16$, but based on $n = 4$ (instead of $n = 17$), $se = \frac{s}{\sqrt{4}} = 3.58$ and $t = \frac{7.265-0}{3.58} = 2.0$, T-statistic has two-sided P-value $= 0.14$ when $df = 3$
- 95% CI is (-4.1, 18.7) which contains 0
- One-sided test about mean: for example, if study predicts "family therapy" has positive effect, could use $H_1 : \mu > 0$. P-value: $P(t > 2.0) = 0.07$

Effect of sample size on tests

Suppose anorexia study for weight change had

- $\overline{y} = 1.0, s = 2.0$, for $n = 400$, then $se = \frac{2.0}{\sqrt{400}} = 0.1$, $t = \frac{1.0-0}{0.1} = 10.0$, $P - value = 0.000000\cdots$
- 95% CI is $1.0 \pm 1.96 \times 0.1$ (0.8, 1.2)
- For a given observed sample mean and standard deviation, the larger the sample size n, the larger the test statistic (because se in denominator is smaller) and the smaller the P-value. (i.e., we have more evidence with more data)
- We're more likely to reject a false H_0 when we have a larger sample size (the test then has more "power")
- This shows there is a positive effect, but it is very small in practical terms. (There is statistical significance, but not practical significance.)

Significance test for a proportion π

Assumptions: categorical variable, randomization, large sample

Hypotheses

- Null hypothesis: $H_0 : \pi = \pi_0$
- Alternative hypothesis: $H_0 : \pi \neq \pi_0$ (2-sided) ($H_0 : \pi > \pi_0$ or $H_0 : \pi < \pi_0$ for 1-sided)
- Test statistic: $t = \frac{\pi - \pi_0}{se/n} = \frac{\pi - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}$
- Conclusion: As in test for mean (e.g., reject H_0 if P-value $\leq \alpha$)

Example: can dogs smell bladder cancer? (British Medical Journal, 2004)

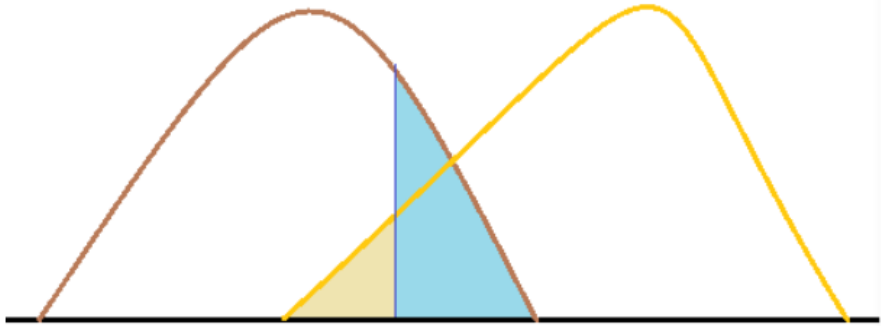
Each trial, one bladder cancer urine sample placed among six control urine samples. Do dogs make the correct selection better than with random guessing? In 54 trials, dogs made correct selection 22 times.

Hypotheses

- Let π = probability of correct guess, for particular trial
- Null hypothesis: $H_0 : \pi = \frac{1}{7}$ VS. $H_1 : \pi > \frac{1}{7}$
- Sample proportion $= 22/54 = 0.407$, Standard error $se_0 = \sqrt{\pi_0(1-\pi_0)/n} = 0.0476$
- $t = \frac{\pi_0 - 0.14}{se_0} = 5.6$, thus p-value $= 0.00000001$
- Conclusion, there is extremely strong evidence that dogs' selections are better than random guessing (for the conceptual population this sample represents)

Error type

Test result	*Reject* H_0	*Accept* H_0
Reality H_0 True	Type I	Correct
Reality H_0 False	Correct	Type II



Error analysis

- Suppose $\alpha-$level $= 0.05$. Then, $P(\text{Type I error}) = P(\text{Reject } H_0, \text{ given it is true}) = 0.05$ (i.e., the $\alpha-$level is the $P(\text{Type I error})$). Type I error is also called false positive
- $P(\text{Type II error}) = \beta$ depends on the true value of the parameter. The farther the true parameter value falls from the null value, the easier it is to reject null, and β value goes down. Type II error is also called false negative
- Power of test $= 1 - \beta = P(\text{Reject } H_0, \text{ given it is false})$

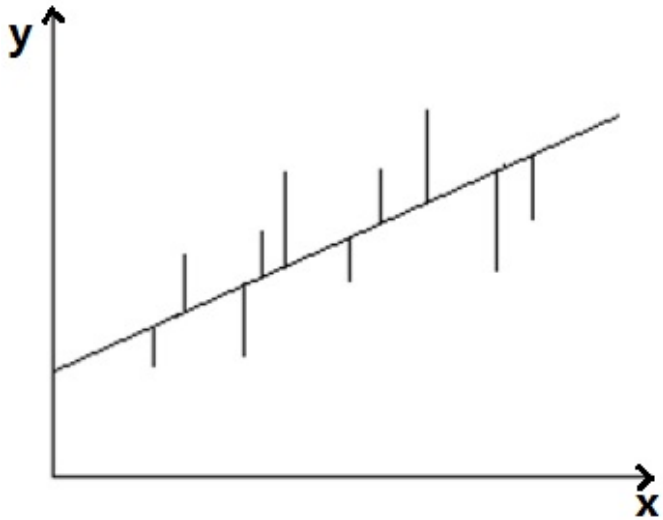
More on mean comparison

Analysis of variance, ANOVA

- One-way analysis of variance: analyzes relationship between quantitative response y and single categorical explanatory factor
- Multiple-way analysis of variance: analyzes relationship between quantitative response y and multiple categorical explanatory factors
- Let #group be g associated with means $\mu_1, \mu_2, \cdots, \mu_g$, the analysis of variance (ANOVA) is an F test of
  $H_0 : \mu_1 = \mu_2 = \cdots = \mu_g$
  H_1 : The means are not all identical
- The F test statistic is large (and P-value is small) if variability between groups is large relative to variability within groups
  $$F = \frac{\text{inter-group estimate of variance}\sigma^2}{\text{intra-group estimate of variance}\sigma^2}$$

Linear regression

Given a set of n points (x_i, y_i) on a scatterplot, find the relationship between x and y: $\widehat{y_i} = \alpha + \beta x_i + \epsilon_i$ such that the sum of squared residuals (SSR) $\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2$



Parameter estimation

- $SSR = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$
-

$$\frac{\partial SSR}{\partial \alpha} = 0, \frac{\partial SSR}{\partial \beta} = 0$$

$$\alpha = \overline{y} - \beta \overline{x}, \beta = \frac{\sum_{i=1}^{n} (x_i y_i - x_i \overline{y})}{\sum_{i=1}^{n} (x_i^2 - x_i \overline{x})} = \frac{Cov(x, y)}{Var(x)}$$

Linear regression Cont.

- $SS_x = \sum_{i=1}^{n} x^2 = (n-1)Var(x) = \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2$
- $SS_y = \sum_{i=1}^{n} y^2 = (n-1)Var(y) = \sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n} y_i^2 - n\overline{y}^2$
- $SS_{xy} = \sum_{i=1}^{n} xy = (n-1)Cov(x,y) = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - n\overline{xy}$
- The third parameter: $r^2 = \frac{Var(\widehat{y})}{Var(y)} = \frac{\beta^2 Var(x)}{Var(y)} = \frac{Cov(x,y)^2}{Var(x)Var(y)} = \frac{SS_{xy}^2}{SS_x SS_y}$
- Uncertainty in regression parameters
  - $t = \frac{\beta}{s_\beta} = r\frac{\sqrt{n-2}}{\sqrt{1-r^2}} \approx t(n-2)$
  - ANOVA approach to test the significance of regression: $\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \overline{y}_i)^2 + \sum_{i=1}^{n}(\overline{y}_i - \widehat{y}_i)^2 = SS_R + SS_E$
  - $F = \frac{SS_R/1}{SS_E/(n-2)} \sim F(1, n-2)$
  - Statistically significant of linear relationship: $H_0 : \beta = 0$ VS. $H_1 : \beta \neq 0$

Case study

Example for linear regression

x: 1065, 1254, 1300, 1577, 1600, 1750, 1800, 1870, 1935, 1948, 2254, 2600, 2800, 3000

y: 199.9, 228.0, 235.0, 285.0, 239.0, 293.0, 285.0, 365.0, 295.0, 290.0, 385.0, 505.0, 425.0, 415.0

$> lm.r = lm(y \sim x)$

$> summary(lm.r)$

Call: lm(formula = y ~ x)

Residuals:

Min 1Q Median 3Q Max

-53.602 -23.650 -1.192 10.948 91.898

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	52.350	37.285	1.404	0.186
x	0.138	0.018	7.407	8.2e-06 ***

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

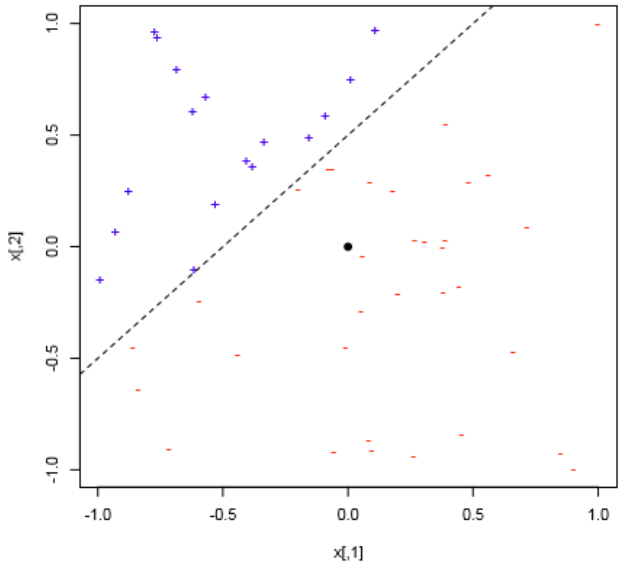
Residual standard error: 39.02 on 12 degrees of freedom

Multiple R-squared: 0.8205, Adjusted R-squared: 0.8056

F-statistic: 54.86 on 1 and 12 DF, p-value: 8.199e-06 >

Logistic regression

Given a set of n points (x_i, y_i) (where the output y_i is a binary variable), and we would like to model the conditional probability $P(Y = 1|X = x)$ as a function of x. How can we use linear regression to solve this?

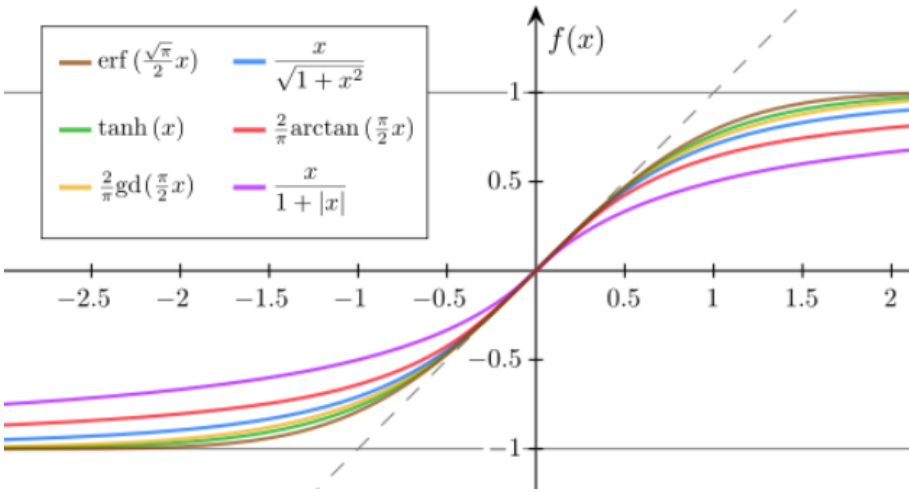


Logistic regression

- Formally, the logistic regression model is that $\log \frac{p(x)}{1-p(x)} = \beta_0 + x \cdot \beta$
- Solving for p, this gives $P(x; \beta_0, \beta) = \frac{e^{\beta_0 + x\cdot\beta}}{1+e^{\beta_0 + x\cdot\beta}} = \frac{1}{1+e^{-(\beta_0 + x\cdot\beta)}}$
- To minimize the mis-classification rate, we should predict $Y = 1$ when $p \geq 0.5$ and $Y = 0$ otherwise

Logistic regression Cont.

A sigmoid function is a mathematical function having an "S" shaped curve (sigmoid curve). Often, sigmoid function refers to the special case of the logistic function defined by the formula $S(t) = \frac{1}{1+e^{-t}}$



Parameter estimation

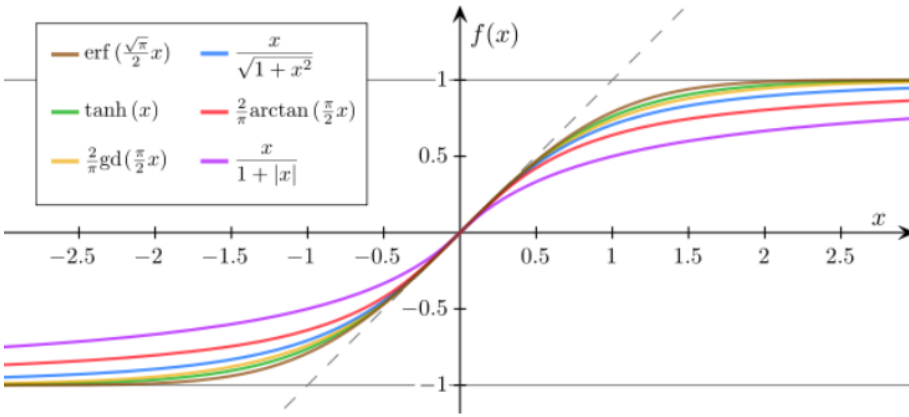
- $L(\beta_0, \beta) = \Pi_{i=1}^{n} P(x_i)^{y_i}(1 - p(x_i))^{1-y_i}$
-

$$\ln L(\beta_0, \beta) = \mathcal{L}(\beta_0, \beta) = \sum_{i=1}^{n}[y_i \ln P(x_i) + (1 - y_i) \ln 1 - P(x_i)]$$

$$= \sum_{i=1}^{n} - \ln 1 + e^{\beta_0 + x_i \cdot \beta} + \sum_{i=1}^{n} y_i(\beta_0 + x_i \cdot \beta)$$

Parameter estimation for logistic regression Cont.

A sigmoid function is a mathematical function having an "S" shaped curve (sigmoid curve). Often, it refers to the special case of the logistic function defined by the formula $S(t) = \frac{1}{1+e^{-t}}$



Parameter estimation

- $$L(\beta_0, \beta) = \Pi_{i=1}^{n} P(x_i)^{y_i}(1 - p(x_i))^{1-y_i}$$

$$\ln L(\beta_0, \beta) = \mathcal{L}(\beta_0, \beta) = \sum_{i=1}^{n}[y_i \ln P(x_i) + (1 - y_i) \ln 1 - P(x_i)]$$

$$= \sum_{i=1}^{n} - \ln 1 + e^{\beta_0 + x_i \cdot \beta} + \sum_{i=1}^{n} y_i(\beta_0 + x_i \cdot \beta)$$

- It's a transcendental equ. & no closed-form solu. (Newtons method)

Take-aways

- Probability review
  - r.v.
  - Probability operations
  - Bayes rule
- Statistics review
  - Descriptive statistics
  - Estimation
  - Hypothesis

Acknowledgement

Many slides are copied or adapted from:

- Prof. Anthony D. Joseph's slides for course cs109 (2006) at EECS@UCBerkeley (`http://cs109.github.io/2015/pages/videos.html`)