

R for bioinformatics, data summarisation and statistics

HUST Bioinformatics course series

Wei-Hua Chen (CC BY-NC 4.0)

06 November, 2023

section 1: TOC

前情提要

- basic plot functions
- basic ggplot2
- special letters
- equations
- advanced ggplot2

本次提要

- data summarisation functions (vector data)
 - median, mean, sd, quantile, summary
- 图形化的 data summarisation (two-D data/ tibble/ table)
 - dot plot
 - smooth
 - linear regression
 - correlation & variance explained
 - grouping & bar/ box/ plots
- statistics
 - parametric tests
 - t-test
 - one way ANNOVA
 - two way ANNOVA
 - linear regression
 - model / prediction / coefficients
 - non-parametric comparison

section 2: vector summarisation

vector data

1 distribution

```
library(tidyverse);  
ggplot( swiss, aes( x = Infant.Mortality ) ) + geom_density() +  
  ggtitle("Swiss Fertility and Socioeconomic Indicators (1888) Data")
```

Swiss Fertility and Socioeconomic Indicators (1888) Data



describe normal distributions

可以用 mean 和 sd 来描述

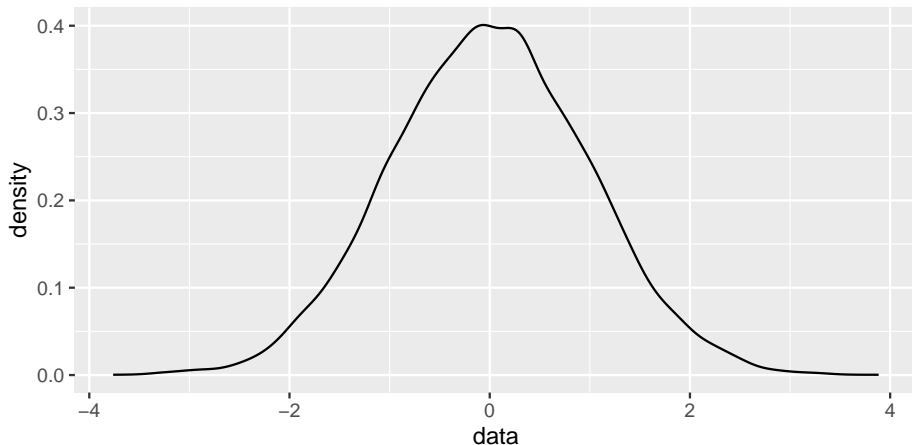
- Ⓐ It's symmetrical.
- Ⓑ Mean and median are the same.
- Ⓒ Most common values are near the mean; less common values are farther from it.
- Ⓓ Standard deviation marks the distance from the mean to the inflection point.

$(\text{mean} + 1 * \text{sd}) \geq 68\%$

$(\text{mean} + 2 * \text{sd}) \geq 95\%$ 的数据

functions to generate random normal distributions

```
# 生成 10000 个随机数字, 使其 mean = 0, sd = 1, 且为 normal distribution ...
x <- rnorm(10000, mean = 0, sd = 1);
ggplot( data.frame( data = x ), aes( data ) ) + geom_density( );
```



More to read: <http://uc-r.github.io/generating-random-numbers/>

other functions to generate random normal distributions

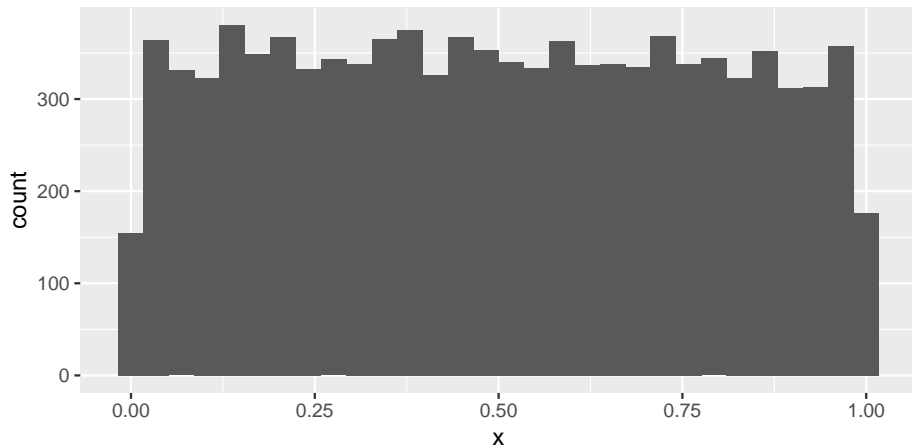
注意，以下函数中的 q , p , x 需要自行提供

```
# generate CDF probabilities for value(s) in vector q  
pnorm(q, mean = 0, sd = 1)  
  
# generate quantile for probabilities in vector p  
qnorm(p, mean = 0, sd = 1)  
  
# generate density function probabilities for value(s) in vector x  
dnorm(x, mean = 0, sd = 1)
```

其它规律的 distributions

① uniform distributions

```
x <- runif( 10000 ); ## random numbers of uniform distributions between 0 and 1
ggplot( data.frame( dat = x ), aes( x ) ) + geom_histogram();
```



uniform distribution 的各种函数

注：以下函数中的 n 需要自行决定

```
# generate n random numbers between 0 and 25  
runif(n, min = 0, max = 25)
```

```
# generate n random numbers between 0 and 25 (with replacement)  
sample(0:25, n, replace = TRUE)
```

```
# generate n random numbers between 0 and 25 (without replacement)  
sample(0:25, n, replace = FALSE)
```

other distributions, cont.

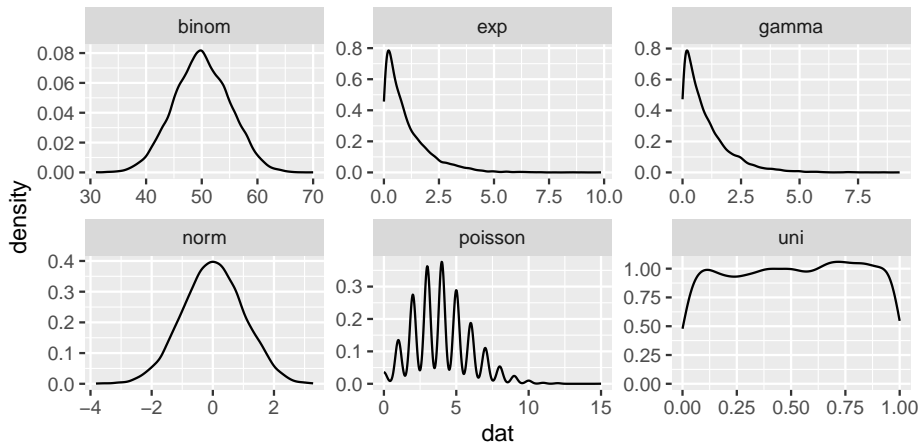
```
n <- 10000;
uni <- tibble( dat = runif(n), type = "uni" );
norm <- tibble( dat = rnorm(n), type = "norm" );
binom <- tibble( dat = rbinom(n, size = 100, prob = 0.5), type = "binom" );
poisson <- tibble( dat = rpois(n, lambda = 4), type = "poisson" );
exp <- tibble( dat = rexp(n, rate = 1) , type = "exp");
gamma <- tibble( dat = rgamma(n, shape = 1) , type = "gamma");

combined <- bind_rows( uni, norm, binom, poisson, exp, gamma );

plot1 <-
  ggplot( combined , aes( dat ) ) + geom_density() +
  facet_wrap( ~type, ncol = 3, scales = "free" );
```

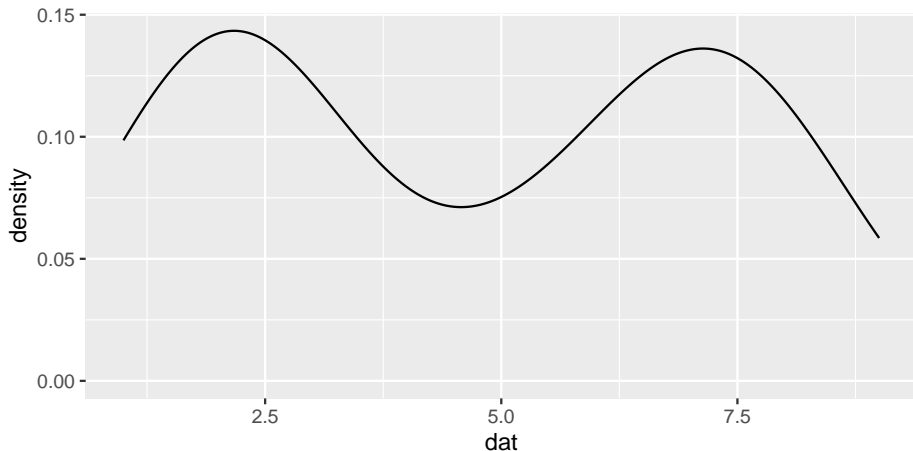
other distributions, plot

```
plot1;
```



non-parametric distribution

```
## votes on people's desire to visit  
bi <- c(7, 3, 2, 1, 7, 3, 4, 5, 7, 6, 2, 2, 1, 3, 7, 2, 6, 8, 2, 7, 2, 2, 1,  
3, 5, 8, 2, 6, 7, 8, 6, 2, 8, 7, 9, 2, 7, 5, 1, 8, 8, 2, 3, 7, 3, 8);  
ggplot( data.frame( dat = bi ), aes(dat)) + geom_density();
```



量化描述数据

使用以下同名函数

mean: aka average, is the sum of all of the numbers in the data set divided by the size of the data set.

median: The median is the value that is in the middle when the numbers in a data set are sorted in increasing order.

sd: standard deviation

var: measures how far a set of numbers are spread out

range: 取值范围

note: from: <https://www.ai-therapy.com/psychology-statistics/descriptive/mean-mode-median>

量化描述函数

```
mean( norm$dat );  
median( norm$dat );  
## mode( norm$dat ); ## ???  
  
sd(norm$dat);  
var(norm$dat);  
range(norm$dat);
```


quantile and summary

```
quantile( norm$dat );
```

```
##           0%           25%           50%           75%           100%
## -3.82084584 -0.66214402  0.01223215  0.68831782  3.28013446
```

`quantile` 还接受其它参数

```
quantile( norm$dat, probs = seq(0, 1, length = 11));
```

```
##           0%           10%           20%           30%           40%           50%
## -3.82084584 -1.26664692 -0.82366292 -0.50565948 -0.24010092  0.01223215
##           60%           70%           80%           90%           100%
##  0.26349063  0.53313804  0.85244470  1.30843229  3.28013446
```

`summary` ...

```
summary( norm$dat );
```

```
##      Min.   1st Qu.   Median     Mean   3rd Qu.    Max.
## -3.82085 -0.66214  0.01223  0.01152  0.68832  3.28013
```

`summary` 也可应用于非数值

```
summary( combined$type );
```

```
##      Length      Class      Mode
##      60000 character character
```

summary, cont.

```
## summary 可应用于整个表格；相当于对每列进行 summary ...
summary( combined );
```

```
##      dat      type
## Min.   :-3.8209 Length:60000
## 1st Qu.: 0.3184 Class :character
## Median : 0.9389 Mode  :character
## Mean    : 9.4180
## 3rd Qu.: 4.0000
## Max.    :70.0000
```

table 函数

返回 vector 当中 unique 值和它们的出现次数

```
table( combined$type );
```

```
##
##   binom      exp   gamma   norm poisson    uni
##  10000  10000  10000  10000   10000  10000
```

**** 注 **** : table 还接受 data.frame 作为输入, 比如 table(combined)。
请自行尝试并理解结果

count in dplyr

```
combined %>% dplyr::count( type );
```

```
## # A tibble: 6 x 2
##   type      n
##   <chr>   <int>
## 1 binom  10000
## 2 exp    10000
## 3 gamma  10000
## 4 norm   10000
## 5 poisson 10000
## 6 uni    10000
```

which is similar to:

```
combined %>% group_by(type) %>% count( name = "my_count" );
```

```
## # A tibble: 6 x 2
## # Groups:   type [6]
##   type      my_count
##   <chr>     <int>
## 1 binom    10000
## 2 exp      10000
## 3 gamma    10000
## 4 norm     10000
## 5 poisson  10000
## 6 uni      10000
```

dplyr::count cont.

For table()-like output with two factors:

```
iris %>%
  group_by(Species) %>%
  count(Petal.Width)
```

```
## # A tibble: 27 x 3
## # Groups:   Species [3]
##   Species    Petal.Width     n
##   <fct>         <dbl> <int>
## 1 setosa         0.1      5
## 2 setosa         0.2     29
## 3 setosa         0.3      7
## 4 setosa         0.4      7
## 5 setosa         0.5      1
## 6 setosa         0.6      1
## 7 versicolor     1       7
## 8 versicolor     1.1      3
## 9 versicolor     1.2      5
## 10 versicolor    1.3     13
## # i 17 more rows
```

which is similar to

```
iris %>% count(Species, Petal.Width);
```

```
##   Species Petal.Width  n
## 1 setosa         0.1    5
```

section 3: two column data: part 1

数据介绍: a numeric vector and a factorial vector

此类数据，通常一列是数值，另一列是分组信息，如下例：

```
data.fig3a <- read_csv( file = "data/talk10/nc2015_data_for_fig3a.csv" );
head( data.fig3a[ c("tai", "trans.at") ] ); ## 只显示有用的两列
```

```
## # A tibble: 6 x 2
##   tai trans.at
##   <dbl>    <dbl>
## 1     1    0.195
## 2     1   -0.320
## 3     1   -0.327
## 4     1   -0.304
## 5     1   -0.275
## 6     1   -0.167
```

数据介绍, cont.

tai: 表达量的一种计算方式, 1 == lowest, 5 == highest

trans.at: A - T 碱基使用偏好;

假说:

- (a) de novo synthesis cost of A is higher than T,
- (b) therefore highly expressed genes will tend to use T than A when possible.
- (c) 因此, 在高表达的基因当中, A - T 的差值会变大 (更负)。

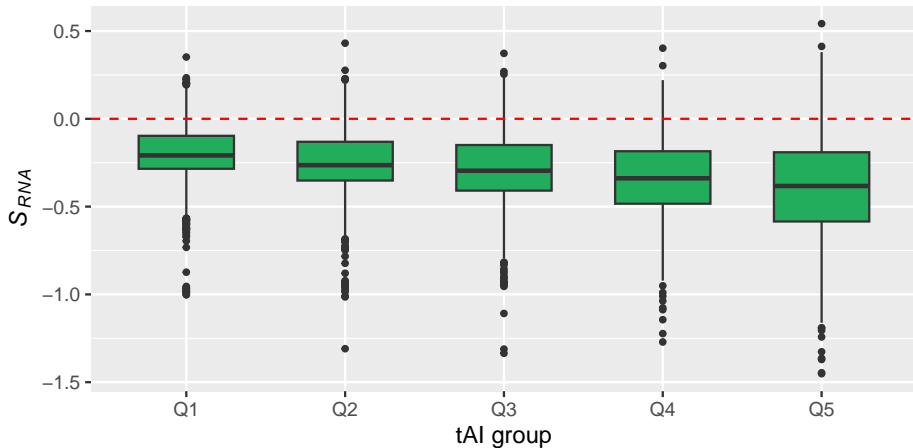
data source: Chen et al, Nature Communications, 2016

boxplot

```
fig3a <-
  ggplot( data.fig3a, aes( factor(tai), trans.at ) ) +
    geom_boxplot( fill = "#22AD5C", linetype = 1 ,outlier.size = 1, width = 0.6) +
    xlab( "tAI group" ) +
    ylab( expression( paste( italic(S[RNA]) ) ) ) +
    scale_x_discrete(breaks= 1:5 , labels= paste("Q", 1:5, sep = "")) +
    geom_hline( yintercept = 0, colour = "red", linetype = 2);
```

show the plot

```
fig3a;
```



说明：

- ① 此种情况下，我们通常只看 median 值的趋势；
- ② 也可增加 Q1 与 Q5 之间的差异分析 (Wilcoxon Rank-sum test): p-value = 2.242528e-07

another example plot

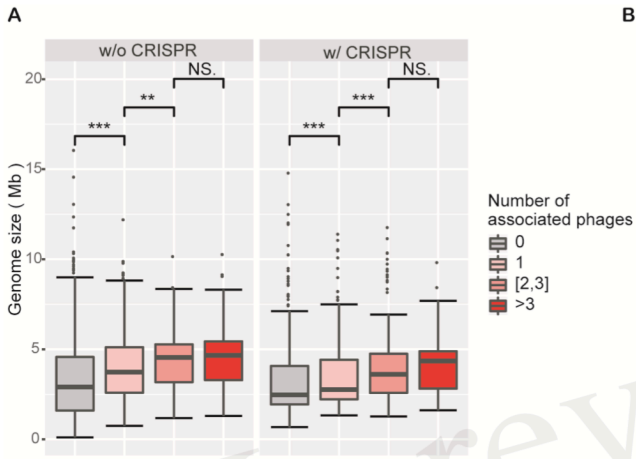
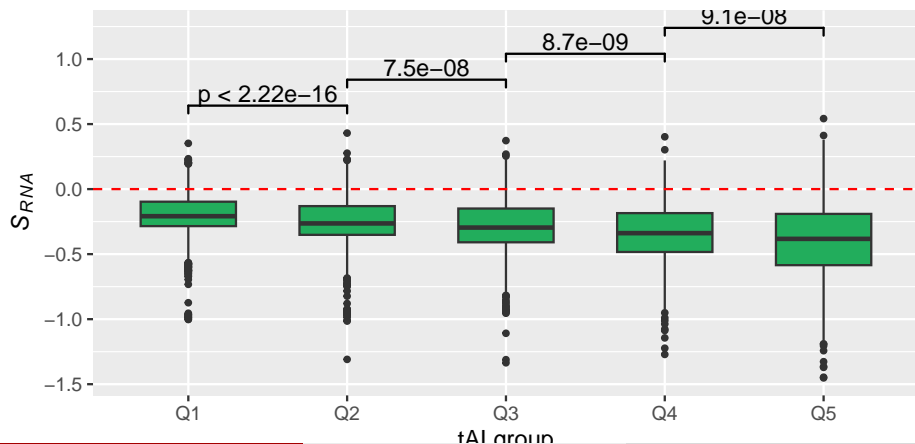


Figure 1: a figure from an in-press manuscript

how to add significance indicators to plot??

ggplot2 的扩展包, `geom_signif`; 如果第一次使用, 请先安装。

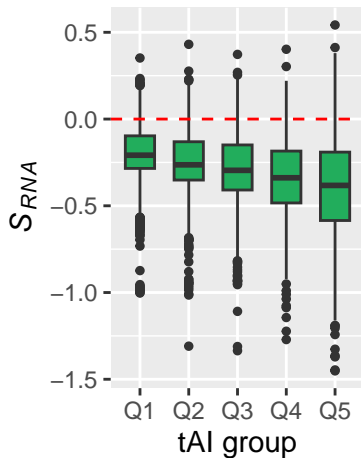
```
library(ggsignif);
fig3a + geom_signif( comparisons = list(1:2, 2:3, 3:4, 4:5), test = wilcox.test,
                     step_increase = 0.1 );
```



boxplot 画图注意事项

正确的画法为：要高瘦，不要矮胖!!!!

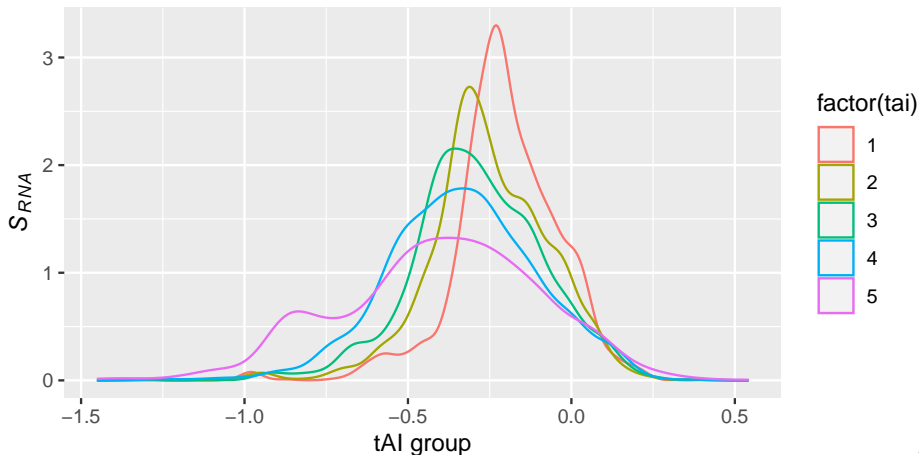
```
fig3a;
```



此类数据的另一种可视化方式

density plot; 但在此例中，不如 boxplot 好。

```
ggplot( data.fig3a, aes( trans.at, colour = factor(tai) ) ) + geom_density( ) +  
  xlab( "tAI group" ) + ylab( expression( paste( italic(S[RNA]) ) ) );
```



section 4: two column data: part 2

数据介绍: two numerical vectors

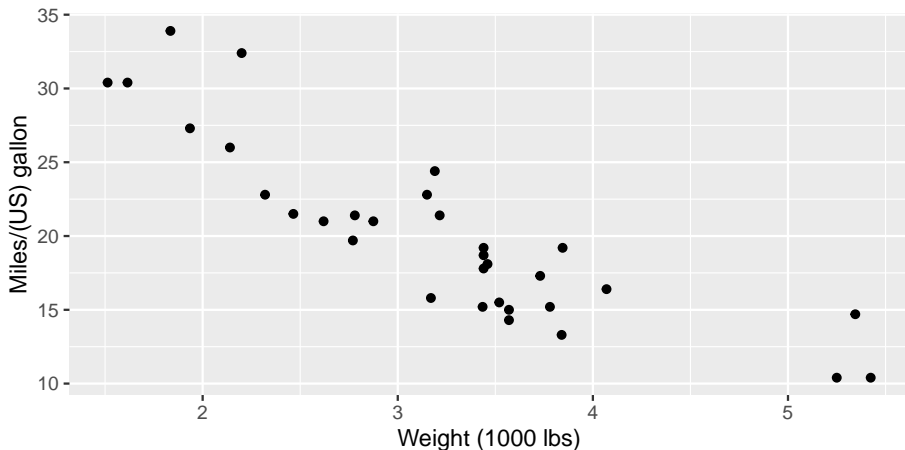
多用于描述两组（量化）数据之间的关系；
以 mtcars 为例：

```
head(mtcars);
```

```
##           mpg cyl  disp  hp  drat    wt  qsec vs am gear carb
## Mazda RX4      21.0   6  160 110  3.90  2.620 16.46  0  1    4    4
## Mazda RX4 Wag  21.0   6  160 110  3.90  2.875 17.02  0  1    4    4
## Datsun 710     22.8   4  108  93  3.85  2.320 18.61  1  1    4    1
## Hornet 4 Drive  21.4   6  258 110  3.08  3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175  3.15  3.440 17.02  0  0    3    2
## Valiant        18.1   6  225 105  2.76  3.460 20.22  1  0    3    1
```


查看重量与燃油效率之间的关系

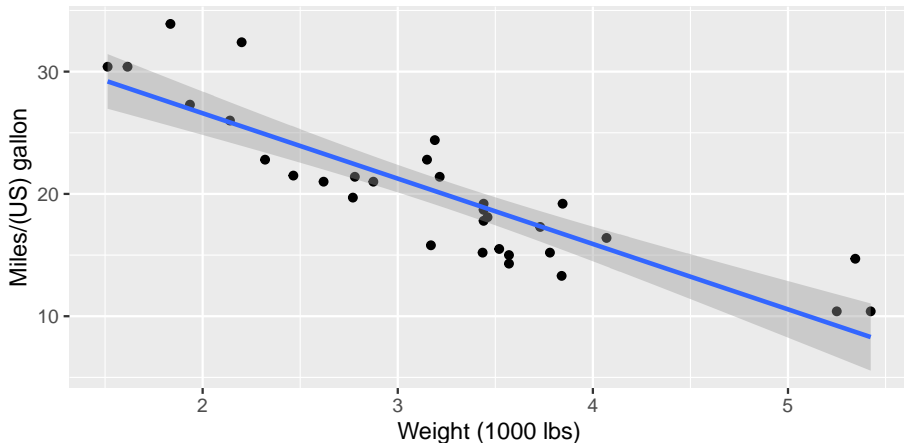
```
plotcars <-
  ggplot( mtcars, aes( x = wt, y = mpg ) ) +
    geom_point() + xlab( "Weight (1000 lbs)" ) + ylab( "Miles/(US) gallon" );
plotcars;
```



smooth, 减少噪音

```
plotcars + geom_smooth( method = "lm" ); ## default is lowess
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



expression(R^2) variance of x explained by y ...

```
( r <- with( mtcars, cor.test( mpg, wt )$estimate ) );
```

```
##          cor
## -0.8676594
```

```
## variance of mpg can be explained by weight
r^2;
```

```
##          cor
## 0.7528328
```

当趋势不明显时，可以按另一组数据分组

这里还以 mtcars 为例。

两种分组 (binning) 方法 equal-distance, equal-size binning

举例：

```
mtcars2 <- mtcars %>%
  mutate( group1 = ntile( wt, 4 ), ## equal-size binning
           group2 = cut( wt,
                         breaks = seq( from = min(wt), to = max(wt),
                                       by = (max(wt) - min(wt)) / 4 ),
                         include.lowest = T ) ## equal-distance ...
  ) ;
```

ntile 函数的参数

... *tile 函数都是 equal size

```
## ntile 的结果
table( mtcars2$group1 );
```

```
##
## 1 2 3 4
## 8 8 8 8
```

cut 函数

按指定的间隔 (breaks) 对数据进行分割。

```
table( mtcars2$group2);
```

```
##
## [1.51,2.49] (2.49,3.47] (3.47,4.45] (4.45,5.42]
##           8           13           8           3
```

使用方法：

```
cut(x, ...)
# S3 method for default
cut(x, breaks, labels = NULL,
    include.lowest = FALSE, right = TRUE, dig.lab = 3,
    ordered_result = FALSE, ...)
```

cut 示例

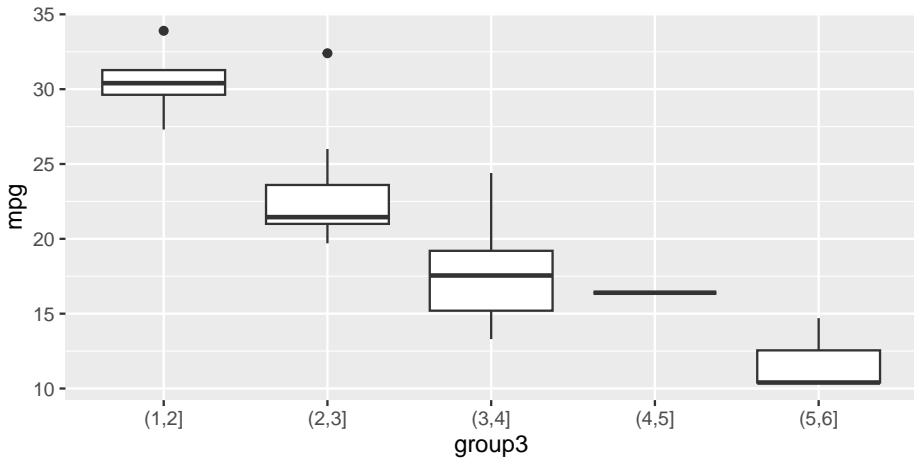
不仅可用于 equal distance, 还可以用于任意间距

```
mtcars3 <- mtcars2 %>%
  mutate( group3 = cut( mtcars$wt, breaks = c(0,1,2,3,4,5,6) ) ) ;
table(mtcars3$group3);
```

```
##
## (0,1] (1,2] (2,3] (3,4] (4,5] (5,6]
##      0      4      8     16      1      3
```

分组后的数据适合用 boxplot

```
ggplot( mtcars3, aes( group3, mpg ) ) +  
  geom_boxplot();
```



小结

目前讲述以下内容:

一维数据

table, summary, range, quantile, mean, median ...

二维数据

- boxplot
- point plot
- correlation
- 分组: equal distance, equal size binning ...

section 5: parametric tests

parametric tests

- ① 包括：
 - t-test
 - analysis of variance
 - linear regression
- ② 数据有较明确的分布 (e.g. normal distribution), 或假设数据有明确的分布; 当假设不成立时, 检测会无效;
- ③ 更灵敏 (相比 nonparametric test), p-value 更低

more to read: http://rcompanion.org/handbook/I_01.html

适用性

适用于

- 数量化性状，比如：身高、体重、产量、污染值
- 整数值：成绩、年龄、每天步数

不适用于

- 其它 count data 或者 discrete data;
- 或者有太多趋向于 min 或 max 的值
- 百分比或比例

详见：<http://rcompanion.org/handbook/index.html>

需要的 packages

需要的 packages

```
## chooseCRANmirror()
if(!require(psych)) {
  install.packages("psych");
}
if( !require(rcompanion) ) {
  install.packages("rcompanion");
}

library(psych);
library(rcompanion)
```

数据

注意 source() 函数的用法

```
source("data/talk10/input_data1.R"); ## 装入 Data data.frame ...

str(Data);
```

```
## 'data.frame':    26 obs. of  5 variables:
## $ Student: chr   "a" "b" "c" "d" ...
## $ Sex      : chr   "female" "female" "female" "female" ...
## $ Teacher: chr   "Catbus" "Catbus" "Catbus" "Catbus" ...
## $ Steps    : int   8000 9000 10000 7000 6000 8000 7000 5000 9000 7000 ...
## $ Rating   : int    7 10 9 5 4 8 6 5 10 8 ...
```

检查数据

```
library(psych)
```

```
##
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':
##
##      %+%, alpha
```

```
headTail(Data); ## psych 包提供的函数
```

```
##      Student      Sex Teacher Steps Rating
## 1      a female Catbus  8000      7
## 2      b female Catbus  9000     10
## 3      c female Catbus 10000      9
## 4      d female Catbus  7000      5
## ...    <NA>    <NA>    <NA>    ...    ...
## 23     w  male Totoro  6000      8
## 24     x  male Totoro  8000     10
## 25     y  male Totoro  7000      7
## 26     z  male Totoro  7000      7
```

查看数据, cont.

```
## 其它常用函数
str(Data)
```

```
## 'data.frame':    26 obs. of  5 variables:
## $ Student: chr   "a" "b" "c" "d" ...
## $ Sex     : chr   "female" "female" "female" "female" ...
## $ Teacher: chr   "Catbus" "Catbus" "Catbus" "Catbus" ...
## $ Steps   : int   8000 9000 10000 7000 6000 8000 7000 5000 9000 7000 ...
## $ Rating  : int    7 10 9 5 4 8 6 5 10 8 ...
```

```
summary(Data)
```

```
##      Student           Sex           Teacher           Steps
## Length:26          Length:26          Length:26          Min.   : 5000
## Class :character    Class :character    Class :character    1st Qu.: 7000
## Mode  :character    Mode  :character    Mode  :character    Median : 8000
##                                     Mean   : 7692
##                                     3rd Qu.: 8750
##                                     Max.   :10000
##
##      Rating
## Min.   : 4.000
## 1st Qu.: 7.000
## Median : 8.000
## Mean   : 7.615
## 3rd Qu.: 9.000
## Max.   :10.000
```

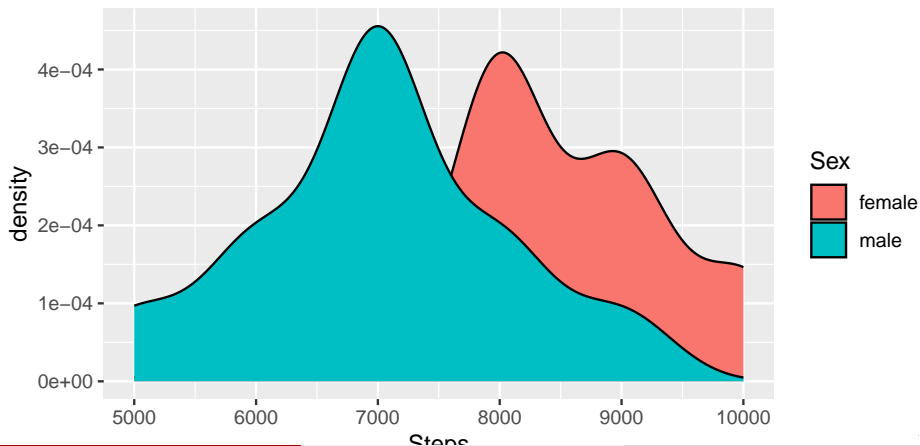

parametric test 的要求

- 1 随机取样
- 2 值或 residuals 为正态分布; residues 是指观察值与预测值 (mean) 之差

数据的分布

```
ggplot(Data, aes(Steps, fill = Sex)) +  
  geom_density(position="dodge", alpha = 1)
```

```
## Warning: Width not defined  
## i Set with `position_dodge(width = ...)`
```



parametric test 的要求, cont.

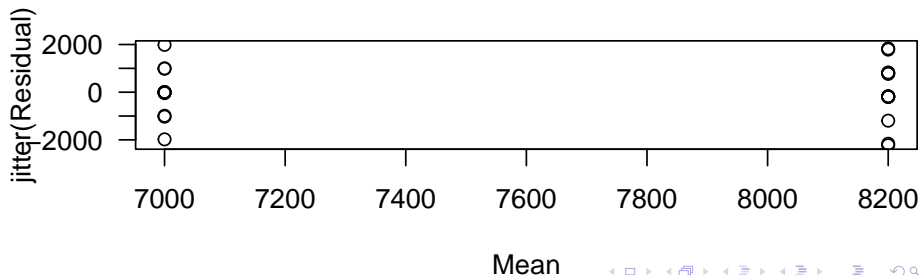
3 有相同的 variance

```
M1 = mean(Data$Steps[Data$Sex=="female"])
M2 = mean(Data$Steps[Data$Sex=="male"])

Data$Mean[Data$Sex=="female"] = M1
Data$Mean[Data$Sex=="male"]   = M2

Data$Residual = Data$Steps - Data$Mean

plot(jitter(Residual) ~ Mean, data = Data, las = 1);
```



how to detect outlier ??

一个很模糊的定义：Outliers are extreme values that fall a long way outside of the other observations. For example, in a normal distribution, outliers may be values on the tails of the distribution.

对于 normal distribution, 通常 $\text{mean} \pm 2 \text{ or } 3 * \text{sd}$

对于 non-parametric distribution (注：IQR 计算可使用同名函数：IQR)：

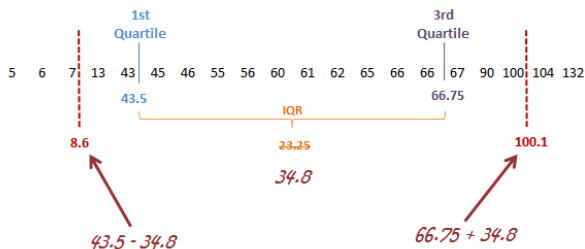
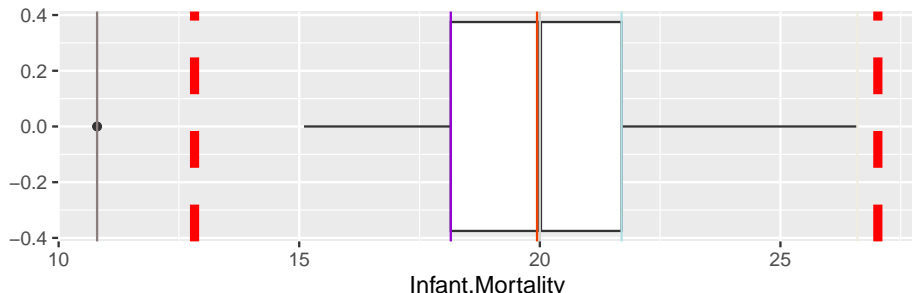


Figure 2: Tukey's method for outlier detection

an example of outlier values

```
s <- summary( swiss$Infant.Mortality );
iqr <- IQR(swiss$Infant.Mortality);
ggplot( swiss, aes( y = Infant.Mortality ) ) + geom_boxplot() + coord_flip() +
  geom_hline( yintercept = s, colour = sample( colors(), length(s) ) ) +
  geom_hline( yintercept = c( s["1st Qu."] - 1.5 * iqr, s["3rd Qu."] + 1.5 * iqr ),
    colour = "red", size = 2, linetype = 2);
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



one sample t-test

检测分布是否与预期一致；比如：男生每天的步数是否显著区别于 1 万

```
with( Data, t.test( Steps[ Sex == "male" ], mu = 10000 ) );
```

```
##
##  One Sample t-test
##
## data:  Steps[Sex == "male"]
## t = -9.083, df = 10, p-value = 3.81e-06
## alternative hypothesis: true mean is not equal to 10000
## 95 percent confidence interval:
##  6264.07 7735.93
## sample estimates:
## mean of x
##      7000
```

two samples t-test

比较 sd 和 mean , 可应用于正态分布。几种使用方法:

```
with( Data, t.test( Steps ~ Sex ) )
```

```
##
##  Welch Two Sample t-test
##
## data:  Steps by Sex
## t = 2.6424, df = 22.816, p-value = 0.01461
## alternative hypothesis: true difference in means between group female and group male is not
## 95 percent confidence interval:
##  260.1421 2139.8579
## sample estimates:
## mean in group female    mean in group male
##           8200           7000
```

two sample t-test 使用方法 2

```
with( Data, t.test( Steps[ Sex == "male" ], Steps[ Sex == "female" ] ) );
```

```
##
##  Welch Two Sample t-test
##
## data:  Steps[Sex == "male"] and Steps[Sex == "female"]
## t = -2.6424, df = 22.816, p-value = 0.01461
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2139.8579  -260.1421
## sample estimates:
## mean of x mean of y
##      7000      8200
```


two sample t test 检测结果

```
res <- with( Data, t.test( Steps ~ Sex ) );
str(res);
```

```
## List of 10
##  $ statistic   : Named num 2.64
##  .. attr(*, "names")= chr "t"
##  $ parameter   : Named num 22.8
##  .. attr(*, "names")= chr "df"
##  $ p.value      : num 0.0146
##  $ conf.int     : num [1:2] 260 2140
##  .. attr(*, "conf.level")= num 0.95
##  $ estimate     : Named num [1:2] 8200 7000
##  .. attr(*, "names")= chr [1:2] "mean in group female" "mean in group male"
##  $ null.value   : Named num 0
##  .. attr(*, "names")= chr "difference in means between group female and group male"
##  $ stderr       : num 454
##  $ alternative: chr "two.sided"
##  $ method       : chr "Welch Two Sample t-test"
##  $ data.name    : chr "Steps by Sex"
##  - attr(*, "class")= chr "htest"
```

paired two sample t test

例如：辅导前后的学生成绩：

```
source("data/talk10/input_data2.R");
```

```
head(scores);
```

```
##      Time Student Score
## 1 Before      a     65
## 2 Before      b     75
## 3 Before      c     86
## 4 Before      d     69
## 5 Before      e     60
## 6 Before      f     81
```

paired two sample t test

```
scores.wide <- scores %>% spread( Time, Score );
head(scores.wide, n = 3);
```

```
##      Student After Before
## 1         a      77      65
## 2         b      98      75
## 3         c      92      86
```

```
with( scores.wide, t.test( After, Before, paired = T ) );
```

```
##
## Paired t-test
##
## data:  After and Before
## t = 3.8084, df = 9, p-value = 0.004163
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  4.141247 16.258753
## sample estimates:
## mean difference
##          10.2
```

one way ANOVA

ANOVA: similar to independent t-test, but can be applied to multiple groups

主要分析一个变量（多组）对另一变量的贡献度/解释度/影响度。

比如：年级对学生的体重是否有影响？

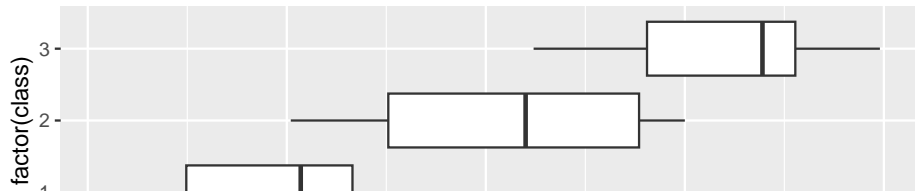
准备：3 个班学生的体重

```

wts <- bind_rows( tibble( class = 1, wt = sample( seq(50, 60, by = 0.1), 20 ) ),
                  tibble( class = 2, wt = sample( seq(55, 65, by = 0.1), 20 ) ),
                  tibble( class = 3, wt = sample( seq(60, 70, by = 0.1), 20 ) )
                );

ggplot(wts, aes( factor( class ), wt ) ) + geom_boxplot() + coord_flip();

```



one way ANOVA, cont.

```
library(FSA); ## 如果没有这个包，请先安装 ...
```

```
## ## FSA v0.9.3. See citation('FSA') if used in publication.
## ## Run fishR() for related website and fishR('IFAR') for related book.
```

```
##
## Attaching package: 'FSA'
```

```
## The following object is masked from 'package:psych':
##
##      headtail
```

```
with( wts, Summarize( wt ~ class, digits = 3 ) );
```

```
##      class  n   mean    sd  min    Q1 median    Q3   max
## 1      1  20 55.070 2.862 50.3 52.475 55.35 56.650 59.9
## 2      2  20 60.390 3.520 55.1 57.550 61.00 63.850 65.0
## 3      3  20 66.195 2.531 61.2 64.050 66.95 67.775 69.9
```

linear model

两个问题：

① 组间有显著区别吗？

```
model <- lm( wt ~ class, data = wts );
```

```
anova( model );
```

```
## Analysis of Variance Table
```

```
##
```

```
## Response: wt
```

```
##          Df Sum Sq Mean Sq F value    Pr(>F)
## class      1 1237.66  1237.66   139.78 < 2.2e-16 ***
```

```
## Residuals 58   513.55     8.85
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA

② 分组对变量的贡献 (r-square, aka. variance explained)

```
summary( model );
```

```
##
## Call:
## lm(formula = wt ~ class, data = wts)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5.4517 -2.4767  0.4796  2.3827  4.9108
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  49.4267      1.0164   48.63  <2e-16 ***
## class        5.5625      0.4705   11.82  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.976 on 58 degrees of freedom
## Multiple R-squared:  0.7067, Adjusted R-squared:  0.7017
## F-statistic: 139.8 on 1 and 58 DF,  p-value: < 2.2e-16
```

其中的值都是什么意思 ???

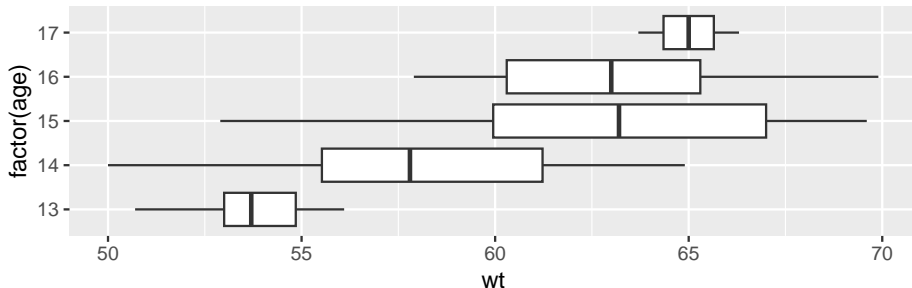
更多内容见: http://rcompanion.org/handbook/T_05.html

one way ANOVA with blocks

同时有多个因素影响体重时，哪些才是主要的？

```
wts2 <- bind_rows(
  tibble( class = 1, age = sample( 13:15, 20, replace = T ), wt = sample( seq(50, 60, by = 0.1), 20, replace = T ),
  tibble( class = 2, age = sample( 14:16, 20, replace = T ), wt = sample( seq(55, 65, by = 0.1), 20, replace = T ),
  tibble( class = 3, age = sample( 15:17, 20, replace = T ), wt = sample( seq(60, 70, by = 0.1), 20, replace = T )
);

ggplot(wts2, aes( factor( age ), wt )) + geom_boxplot() + coord_flip();
```



one way ANOVA with blocks, cont.

```
model2 <- lm( wt ~ class + age, data = wts2);
anova( model2 );
```

```
## Analysis of Variance Table
```

```
##
```

```
## Response: wt
```

```
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## class      1 1128.91 1128.91 123.4477 6.88e-16 ***
## age        1    1.09    1.09   0.1188  0.7316
## Residuals 57   521.25    9.14
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

如何获得 r.squre value ???

one way ANOVA with blocks, 各个 factor 的重要性 ??

```
library(relaimpo);
```

```
res3 <- calc.relimp( wt ~ factor(class) + age, data = wts2 );
res3$R2; ## 总 R2
```

```
## [1] 0.6901301
```

```
res3$lmg; ## 每个因素的贡献;
```

```
## factor(class)          age
##      0.5162359      0.1738942
```

```
## 测试 rela 参数:
```

```
res4 <- calc.relimp( wt ~ factor(class) + age, data = wts2, rela = T);
res4$R2; ## 总 R2
```

```
## [1] 0.6901301
```

```
res4$lmg; ## 每个因素的贡献;
```

```
## factor(class)          age
##      0.748027      0.251973
```

更多请见: http://rcompanion.org/handbook/I_06.html

two way ANOVA

一个变量受另外两个因素影响的分析；比如上例中 体重受 年级和 年龄的影响。

年级和 年龄至少有 4 个 unique combinations .

实际上，上面的 block test 可以认为是 two-way ANOVA 分析

```
Summarize(wt ~ age + class, data = wts2, digits=3);
```

##	age	class	n	mean	sd	min	Q1	median	Q3	max
## 1	13	1	7	53.743	1.778	50.7	53.000	53.70	54.85	56.1
## 2	14	1	7	54.800	3.496	50.0	51.950	55.60	57.45	59.2
## 3	15	1	6	57.217	3.177	52.9	54.650	58.35	59.80	60.0
## 4	14	2	9	60.833	3.466	55.3	58.100	61.20	64.30	64.9
## 5	15	2	8	62.625	1.845	59.1	61.925	62.90	63.90	64.6
## 6	16	2	3	59.133	1.201	57.9	58.550	59.20	59.75	60.3
## 7	15	3	9	67.044	2.895	60.0	66.500	67.60	68.70	69.6
## 8	16	3	9	64.689	3.170	60.3	62.700	64.40	65.60	69.9
## 9	17	3	2	65.000	1.838	63.7	64.350	65.00	65.65	66.3

two way ANOVA, cont.

```
model3 <- lm( wt ~ class + age + class:age, data = wts2);
anova( model3 );
```

```
## Analysis of Variance Table
##
## Response: wt
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## class      1 1128.91  1128.91  137.8878 < 2.2e-16 ***
## age        1    1.09    1.09    0.1327  0.716985
## class:age   1    62.77    62.77    7.6675  0.007611 **
## Residuals 56   458.48     8.19
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

其中: class 和 age 称为 **main effects**, class:age 称为 **interaction effects**

relative importance of interactions

```
res5 <- calc.relimp( wt ~ factor(class) + age + factor(class):age, data = wts2);
res5$R2; ## 总 R2
```

```
## [1] 0.7245423
```

```
res5$lm; ## 每个因素的贡献;
```

```
##      factor(class) factor(class):age      age
##      0.51623593      0.03441224      0.17389415
```

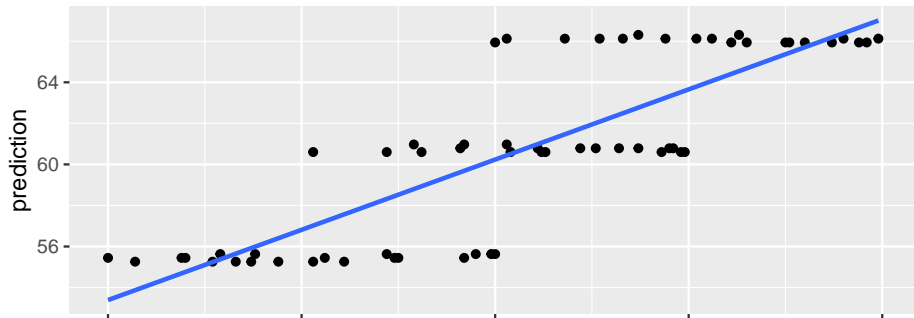
用模型进行预测 predict

```
model2 = lm(formula = wt ~ class + age, data = wts2)
```

```
newdata <- wts2 %>% dplyr::select( class, age );  
wt.predicted <- predict( model2, newdata );
```

```
dat <- data.frame( reference = wts2$wt, prediction = wt.predicted );  
ggplot( dat , aes( x = reference, y = prediction ) ) + geom_point() +  
  geom_smooth( method = "lm", se = F );
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



prediction 与 original data 的 correlation 是多少 ??

```
with( dat, cor.test( prediction, reference ) )$estimate;
```

```
##          cor
## 0.8272405
```

```
## R ^ 2
with( dat, cor.test( prediction, reference ) )$estimate ^2;
```

```
##          cor
## 0.6843269
```

```
## 正好是 model2 的 r.squared ...
summary( model2 )$r.squared;
```

```
## [1] 0.6843269
```

手动计算 prediction

在一个 linear model 中, $wt = \text{intercept} + a * \text{class} + b * \text{age}$
而 intercept , a, b 的值分别为:

```
( paras <- coef( model2 ) );
```

```
## (Intercept)      class      age
##  47.7040785    5.1554503  0.1847644
```

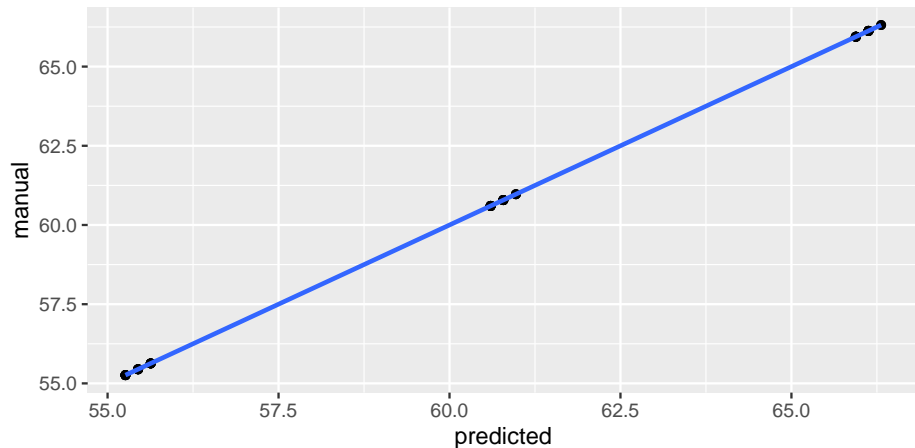
```
predicted2 <-
  paras[1] + paras["age"] * wts2$age + paras["class"] * wts2$class;

plot <-
  ggplot( data.frame( predicted = wt.predicted, manual = predicted2 ),
    aes( predicted, manual ) ) +
    geom_point() + geom_smooth( method = "lm", se = F );
```


show the plot

```
plot;
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



更多更方便的函数

以下函数也可以用于 multivariable analysis / multiple regression

```
fit <- lm(y ~ x1 + x2 + x3, data=mydata)
summary(fit) # show results

# Other useful functions
coefficients(fit) # model coefficients
confint(fit, level=0.95) # CIs for model parameters
fitted(fit) # predicted values
residuals(fit) # residuals
anova(fit) # anova table
vcov(fit) # covariance matrix for model parameters
influence(fit) # regression diagnostics
```

linear regression 注意事项

- ① 是 parametric test
- ② 假设变量之间独立（比如：年龄和班级之间没有关联）
- ③ homogeneity of variance

但实际上 ...

multivariable analysis

```
# Multiple Linear Regression Example
fit <- lm(y ~ x1 + x2 + x3, data=mydata)
summary(fit) # show results

# compare models
fit1 <- lm(y ~ x1 + x2 + x3 + x4, data=mydata)
fit2 <- lm(y ~ x1 + x2)
anova(fit1, fit2)

# K-fold cross-validation
library(DAAG)
cv.lm(df=mydata, fit, m=3) # 3 fold cross-validation

# Stepwise Regression; feature selection
library(MASS)
fit <- lm(y~x1+x2+x3,data=mydata)
step <- stepAIC(fit, direction="both")
step$anova # display results
```

extended reading

- 1 repeated measures ANOVA , 同一变量、不同时间段的重复测量（对上例中学生的体重进行多次测量）
- 2 correlation and linear regression
- 3 non-linear regression

section 6: non-parametric test

wilcox.test and kruskal.test

```
# independent 2-group Mann-Whitney U Test
with( Data, wilcox.test( Steps ~ Sex ) );
```

```
## Warning in wilcox.test.default(x = DATA[[1L]], y = DATA[[2L]], ...): cannot
## compute exact p-value with ties
```

```
##
## Wilcoxon rank sum test with continuity correction
##
## data: Steps by Sex
## W = 127.5, p-value = 0.01773
## alternative hypothesis: true location shift is not equal to 0
```

```
# Kruskal Wallis Test One Way Anova by Ranks
with( Data, kruskal.test( Steps ~ Sex ) );
```

```
##
## Kruskal-Wallis rank sum test
##
## data: Steps by Sex
## Kruskal-Wallis chi-squared = 5.7494, df = 1, p-value = 0.01649
```

作业与练习

作业与练习

- Exercises and homework 目录下 talk10-homework.Rmd 文件
- 完成时间：见钉群的要求