talk10 练习与作业

目录

0.1 练习和作业说明

将相关代码填写入以 “‘{r} “‘ 标志的代码框中，运行并看到正确的结果；

完成后，用工具栏里的”Knit” 按键生成 PDF 文档；

将 PDF 文档改为：`姓名-学号-talk10 作业.pdf`，并提交到老师指定的平台/钉群。

0.2 Talk10 内容回顾

- 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
  - groupping & bar/ box/ plots

- statistics
  - parametric tests
    - t-test
    - one way ANNOVA
    - two way ANNOVA
    - linear regression
    - model / prediction / coefficients
  - non-parametric comparison

0.3 练习与作业：用户验证

请运行以下命令，验证你的用户名。

如你当前用户名不能体现你的真实姓名，请改为拼音后再运行本作业！

```
Sys.info()[["user"]]
```

```
## [1] "lucas"
```

```
Sys.getenv("HOME")
```

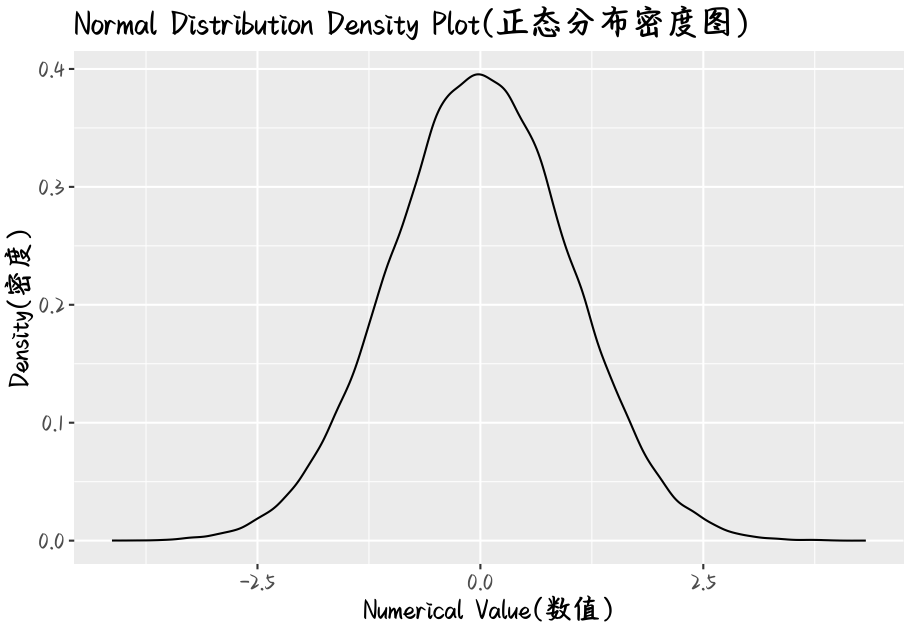
```
## [1] "/Users/lucas"
```

0.4 练习与作业 1：数据查看

- **正态分布**

1. 随机生成一个数字（numberic）组成的 vector，长度为 10 万，其值符合正态分布；
2. 用 ggplot2 的 density plot 画出其分布情况；
3. 检查 mean +- 1 * sd，mean +- 2 * sd 和 mean +- 3 * sd 范围内的取值占总值数量的百分比。

```
## 代码写这里，并运行；
library(ggplot2)

# Setting random seeds to ensure reproducibility
set.seed(123)
data = rnorm(100000)
df = data.frame(value = data)

# Creating density maps with ggplot2
ggplot(df, aes(x = data)) +
  geom_density() +
  # Prevent the GBK character to show as block
  theme(
    text=element_text(
      family="RLQDMSWR",
      size=14)) +
  labs(
    title = "Normal Distribution Density Plot(正态分布密度图)",
    x = "Numerical Value(数值)",
    y = "Density(密度)"
  )
```




```r
# Calculate the percentage of the total number
# of values taken in different ranges:
mean_value = mean(data)
sd_value = sd(data)

# Calculate the percentage in the range mean +- 1 * sd
within_1_sd =
  sum(data >= mean_value - sd_value &
        data <= mean_value + sd_value) / length(data)

# Calculate the percentage in the range mean +- 2 * sd
within_2_sd =
  sum(data >= mean_value - 2 * sd_value &
        data <= mean_value + 2 * sd_value) / length(data)

# Calculate the percentage in the range mean +- 3 * sd
within_3_sd =
  sum(data >= mean_value - 3 * sd_value &
```

```
        data <= mean_value + 3 * sd_value) / length(data)

# Convert results to percentages
within_1_sd * 100
```

```
## [1] 68.149
```

```
within_2_sd * 100
```

```
## [1] 95.463
```

```
within_3_sd * 100
```

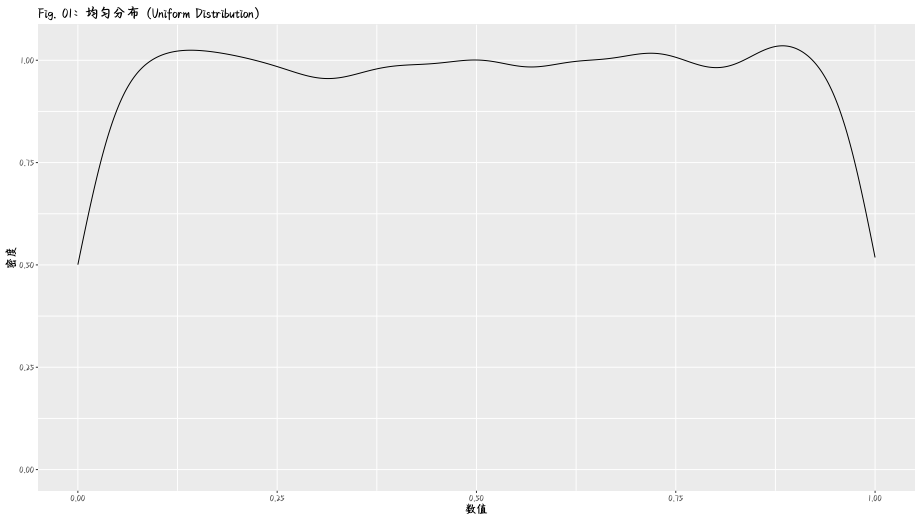
```
## [1] 99.735
```

- **用函数生成符合以下分布的数值，并做图：**

另外，在英文名后给出对应的中文名：

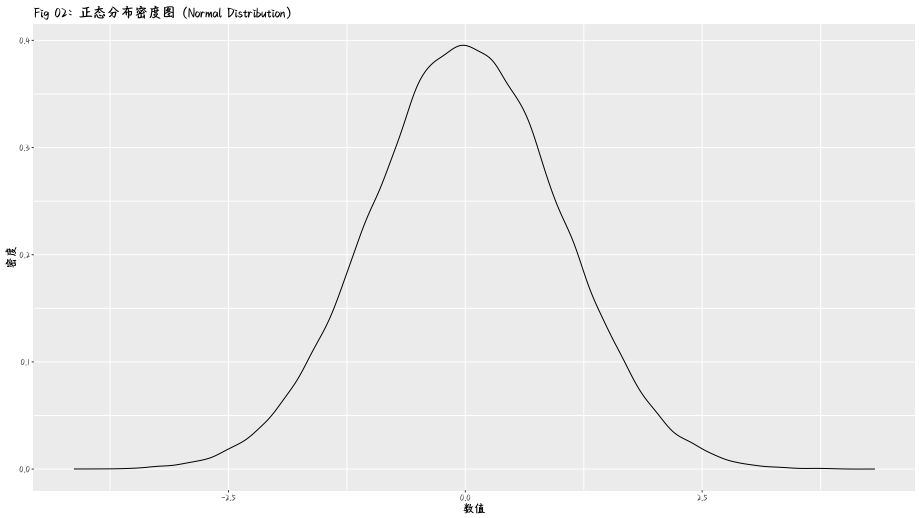
- Uniform Distribution
- Normal Distribution
- Binomial Distribution
- Poisson Distribution
- Exponential Distribution
- Gamma Distribution

```
## 代码写这里，并运行;
# Task 01: Uniform Distribution
# Generate uniformly distributed random values
uniform_data =
  runif(10000,
        min = 0,
        max = 1)

# Creating Density Maps
ggplot(
  data.frame(
    x = uniform_data),
  aes(x)) +
  geom_density() +
  # Prevent the GBK character to show as block
  theme(
    text=element_text(
      family="RLQDMSWR",
      size=14)) +
  labs(
    title = "Fig. 01: 均匀分布 (Uniform Distribution)",
    x = " 数值",
    y = " 密度")
```



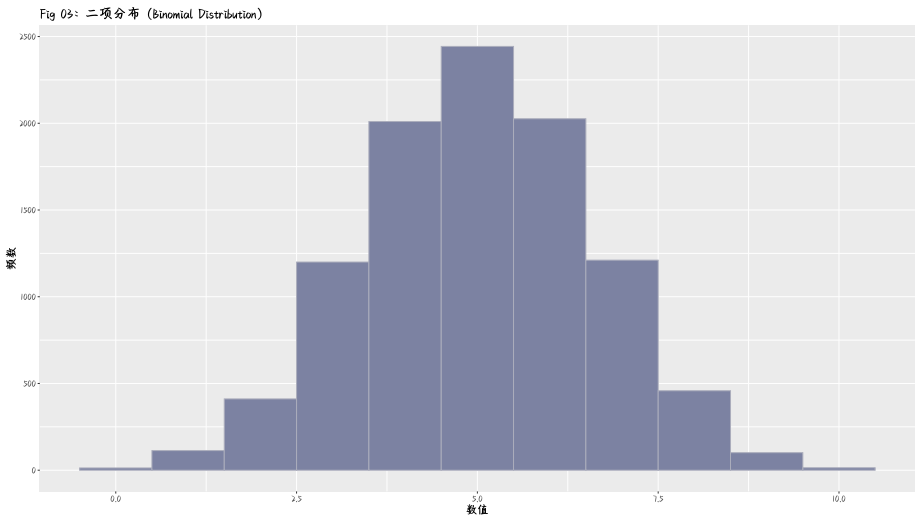

```r
# Task 02: Normal Distribution
# Setting random seeds to ensure reproducibility
set.seed(123)

# Generate normally distributed random values
normal_data =
  rnorm(100000)

# Creating data.frame
normal_df =
  data.frame(value = normal_data)

# Creating Density Plots with ggplot2
library(ggplot2)

ggplot(normal_df,
       aes(x = value)) +
  geom_density() +
  labs(
    title = "Fig 02: 正态分布密度图 (Normal Distribution)",
```

```
    x = " 数值",
    y = " 密度") +
  # Prevent the GBK character to show as block
  theme(
    text = element_text(
      family = "RLQDMSWR",
      size = 14))
```



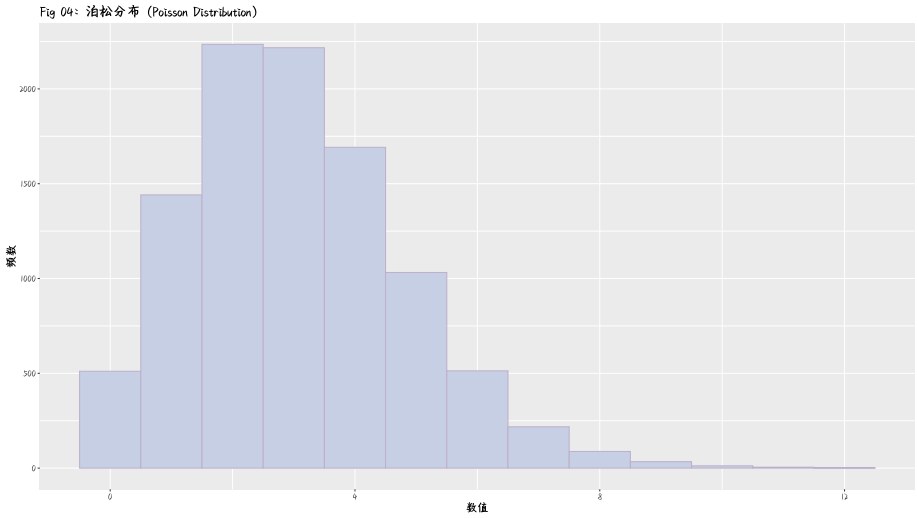

```
# Task 03: Binomial Distribution
# Generate random values for binomial distribution
binomial_data =
  rbinom(10000, size = 10, prob = 0.5)

# Creating Histograms
ggplot(
  data.frame(
    x = binomial_data),
  aes(x)) +
  geom_histogram(
    binwidth = 1,
```

```
    fill = "#7C82A2",
    color = "#ABADBC") +
  labs(
    title = "Fig 03: 二项分布 (Binomial Distribution)",
    x = " 数值",
    y = " 频数") +
  # Prevent the GBK character to show as block
  theme(
    text = element_text(
      family = "RLQDMSWR",
      size = 14))
```



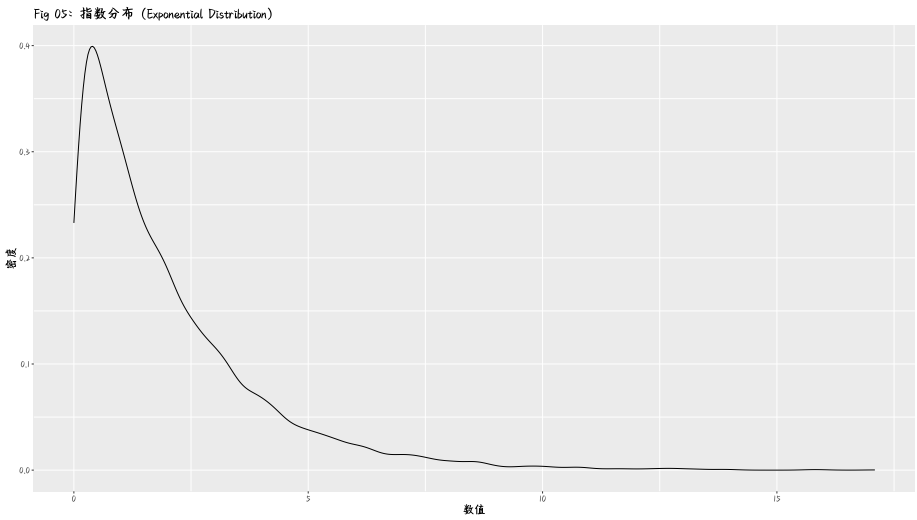

```
# Task 04: Poisson Distribution
# Generate random values for poisson distribution
poisson_data =
  rpois(10000, lambda = 3)

# Creating Histograms
ggplot(
  data.frame(
    x = poisson_data),
```

```
  aes(x)) +
  geom_histogram(
    binwidth = 1,
    fill = "#C7CFE4",
    color = "#BDAFCA") +
  labs(
    title = "Fig 04: 泊松分布 (Poisson Distribution)",
    x = " 数值",
    y = " 频数") +
  # Prevent the GBK character to show as block
  theme(
    text = element_text(
      family = "RLQDMSWR",
      size = 14))
```



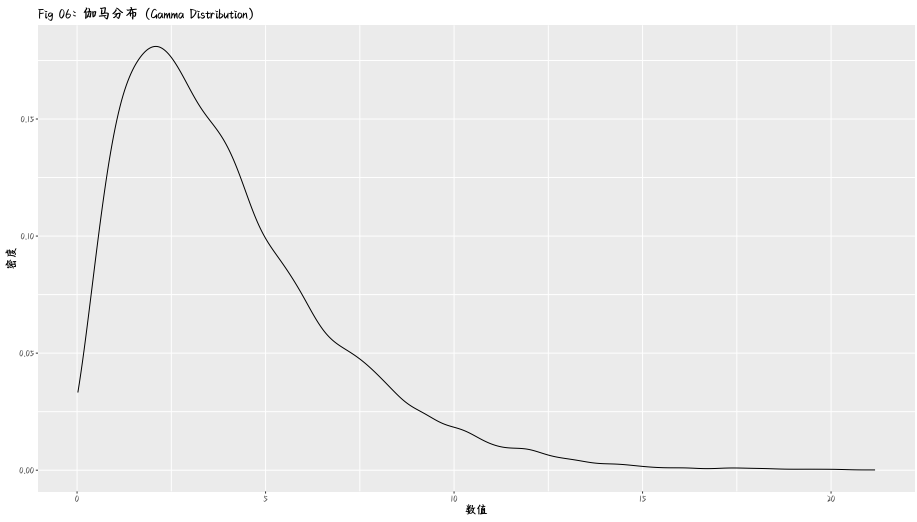

```
# Task 05: Exponential Distribution
# Generate random values for exponential distribution
exponential_data =
  rexp(10000, rate = 0.5)

# Generate density plot
```

```
ggplot(
  data.frame(
    x = exponential_data),
  aes(x)) +
  geom_density() +
  labs(
    title = "Fig 05: 指数分布 (Exponential Distribution)",
    x = " 数值",
    y = " 密度") +
  # Prevent the GBK character to show as block
  theme(
    text = element_text(
      family = "RLQDMSWR",
      size = 14))
```




```
# Task 06: Gamma Distribution
# Generate random values for gamma distribution
gamma_data =
  rgamma(10000, shape = 2, rate = 0.5)

# Generate density plot
```

```r
ggplot(
  data.frame(
    x = gamma_data),
  aes(x)) +
  geom_density() +
  labs(
    title = "Fig 06: 伽马分布 (Gamma Distribution)",
    x = " 数值",
    y = " 密度") +
  # Prevent the GBK character to show as block
  theme(text = element_text(
    family = "RLQDMSWR",
    size = 14))
```



- **分组的问题**
  - 什么是 `equal-sized bin` 和 `equal-distance bin`? 以 `mtcars` 为例，将 `wt` 列按两种方法分组，并显示结果。

```
## 代码写这里，并运行;
```

- **boxplot 中 outlier 值的鉴定**
  - 以 swiss$Infant.Mortality 为例，找到它的 outlier 并打印出来;

```
## 代码写这里，并运行;
```

- **以男女生步数数据为例，进行以下计算:**

首先用以下代码装入 Data:

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

```
##   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
## 5       e female  Catbus  6000      4
## 6       f female  Catbus  8000      8
```

- 分别用`t.test`和`wilcox.test`比较男女生步数是否有显著差异; 打印出`p.value`

```
## 代码写这里，并运行;
```

- 两种检测方法的`p.value`哪个更显著? 为什么?

答:

- **以下是学生参加辅导班前后的成绩情况，请计算同学们的成绩是否有普遍提高?**

注: 先用以下代码装入数据:

```
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
```

注: 计算时请使用 `paired = T` 参数;

```
## 代码写这里，并运行;
```

0.5 练习与作业 2: 作图

- **利用 talk10 中的 `data.fig3a` 作图**
  - 首先用以下命令装入数据:

```
library(tidyverse);
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v lubridate 1.9.2     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts
```

```
data.fig3a <- read_csv( file = "../data/talk10/nc2015_data_for_fig3a.csv" );
```

```
## Rows: 7109 Columns: 8
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): acc
## dbl (7): tai, trans.at, trans.gc, zAA2.at, zAA2.gc, zAA1.at, zAA1.gc
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

- 利用两列数据：`tai` `zAA1.at` 做`talk10`中的`boxplot`（详见：`fig3a`的制作）；

- 用`ggsignif`为相邻的两组做统计分析（如用 `wilcox.test` 函数），并画出`p.value`；

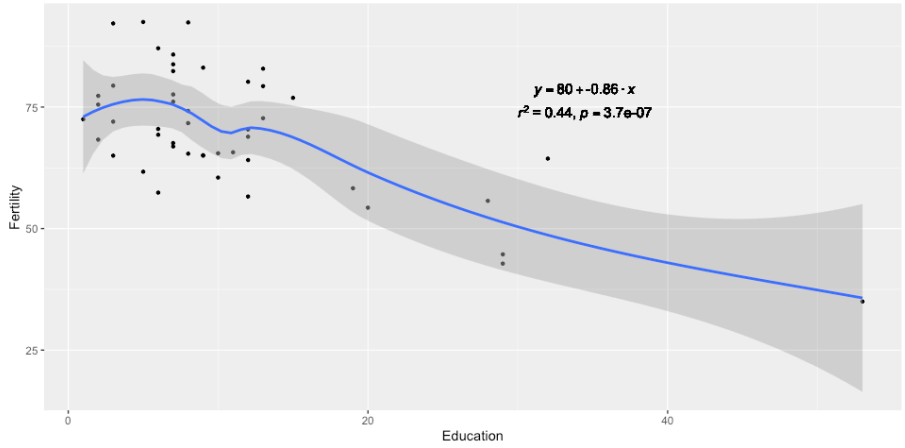
```
## 代码写这里，并运行；
```

问：这组数据可以用 t.test 吗？为什么？

答：

```
## 代码写这里，并运行；
```

- **用系统自带变量 `mtcars` 做图**
  - 用散点图表示 `wt`（x-轴）与 `mpg`（y-轴）的关系
  - 添加线性回归直线图层
  - 计算 `wt`）与 `mpg` 的相关性，并将结果以公式添加到图上。其最终效果如下图所示（注：相关代码可在 `talk09` 中找到）：

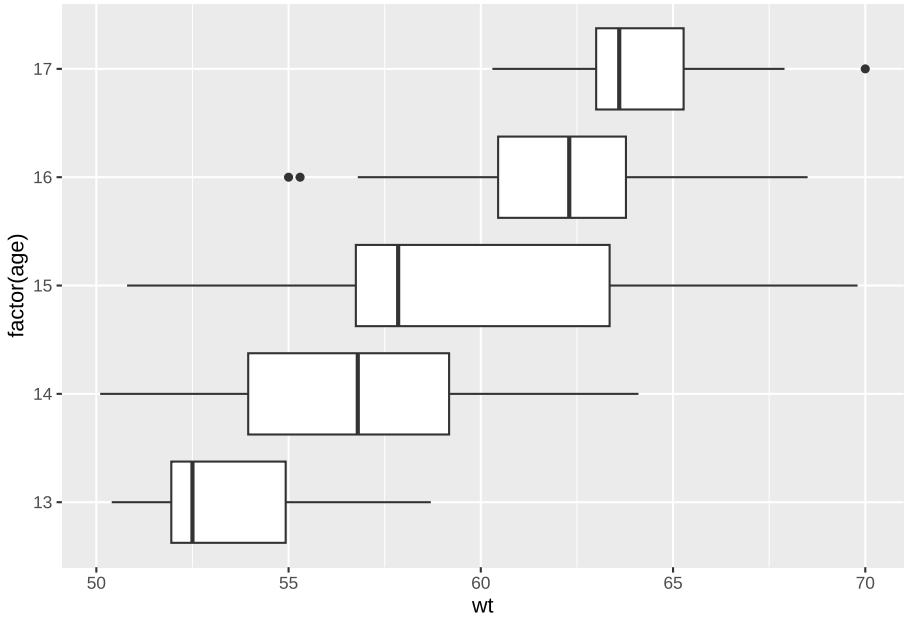



```
## 代码写这里，并运行；
```

0.6 练习与作业 3：线性模型与预测

- **使用以下代码产生数据进行分析**

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

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



- 用线性回归检查`age`, `class` 与 `wt` 的关系，构建线性回归模型；
- 以`age`, `class`为输入，用得到的模型预测`wt`；
- 计算预测的`wt`和实际`wt`的相关性；
- 用线性公式显示如何用`age`, `class`计算`wt`的值。

```
## 代码写这里，并运行;
```