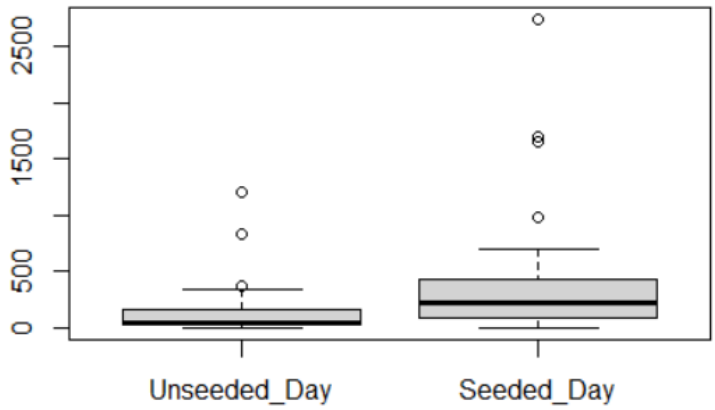
PS3

曾海翔 12032760

```r
library(tidyr)
library(dplyr)
library(ggplot2)

#1
#1.1
#set the array
Unseeded_Day <- c(1202.6, 830.1, 372.4, 345.5, 321.2, 244.3, 163.0, 147.8, 95.0, 87.0, 81.2, 68.5, 47.3, 41.1, 36.
Seeded_Day <- c(2745.6, 1697.1, 1656.4, 978.0, 703.4, 489.1, 430.0, 334.1, 302.8, 274.7, 274.7, 255.0, 242.5, 200.
Rainfall_day <- cbind(Unseeded_Day,Seeded_Day)
boxplot(Rainfall_day)
#1.2
Rainfall_day_tbl <- as_tibble(Rainfall_day)
Rainfall_day_tbl %>%
  summarise(
    mean_Seeded = mean(Rainfall_day_tbl$Seeded_Day),
    mean_Unseeded = mean(Rainfall_day_tbl$Unseeded_Day),
    increse_rate = mean(((Rainfall_day_tbl$Seeded_Day)-mean(Rainfall_day_tbl$Unseeded_Day))/mean(Rainfall_day_tbl$
  )
```

#1



通过 cbind 把接种和未接种的天气降雨量组成一个矩阵，使用 boxplot 画图

```
# A tibble: 1 x 3
  mean_Seeded mean_Unseeded increse_rate
        <dbl>         <dbl>        <dbl>
1         442.          165.         1.69
> t.test(Unseeded_Day, Seeded_Day)

        Welch Two Sample t-test

data:  Unseeded_Day and Seeded_Day
t = -1.9983, df = 33.856, p-value = 0.05376
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -559.552533    4.752533
sample estimates:
mean of x mean of y
 164.5731  441.9731
```

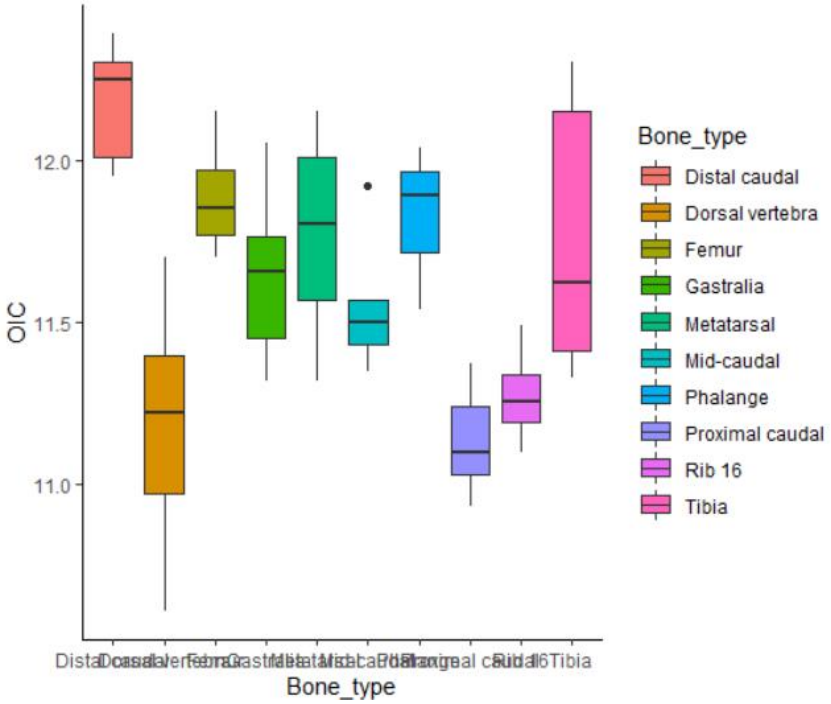
分别的平均数，降雨增加率以及 t 检验结果

#2

```r
#2
TR_bone <- read.csv("PS3_data.csv",header = T)
TR_bone_tbl <- as_tibble(TR_bone)
TR_bone_tbl %>%
  group_by(Bone_type) %>%
  summarise(
    count = n(),
    mean_OIC = mean(OIC),
    sd_OIC = sd(OIC)
  )
#use boxplot to check if there is a different between the distribution
ggplot(TR_bone_tbl, aes(x=Bone_type, y=OIC, fill=Bone_type))+
  geom_boxplot()+
  theme_classic()
#anova one way test
AOW_TR <- aov(OIC~Bone_type, data = TR_bone_tbl)
summary(AOW_TR)
```

把数据复制进 excel 表格中，另存为 csv 格式（见附件）。使用 ggplot 中的 boxplot 画图

用 aov() 进行单因素显著性差异分析，组间有显著性差异性，所以霸王龙可能为变温动物。




```
             Df Sum Sq Mean Sq F value   Pr(>F)
Bone_type     9  5.688  0.6320   7.922 1.01e-06 ***
Residuals    42  3.351  0.0798
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

#3

```
#3
V_Z <- read.csv("PS3_data.csv",header = T)
V_Z_tbl <- as_tibble(V_Z)
names(V_Z_tbl)
V_Z_tbl %>%
  filter(people_type != "") %>%
  select(people_type,zine_level) %>%
  group_by(people_type) %>%
  summarise(
    count = n(),
    mean_Zine = mean(zine_level),
    sd_Zine = sd(zine_level)
  )
AOW_Zine <- aov(zine_level~people_type, data = V_Z_tbl)
summary(AOW_Zine)
```

在管道里用 filter 筛选出有效的数据，按实验对象类型分类，分析 Zn 平均值。

```
`summarise()` ungrouping output (override with `.groups` argument)
# A tibble: 3 x 4
  people_type             count mean_Zine sd_Zine
  <chr>                   <int>     <dbl>   <dbl>
1 Nonpregnant vegetarians     5      179.    27.3
2 Pregnant nonvegetarians     6      178     14.5
3 Pregnant vegetarians       12      177.    20.9
> AOW_Zine <- aov(zine_level~people_type, data = V_Z_tbl)
> summary(AOW_Zine)
             Df Sum Sq Mean Sq F value Pr(>F)
people_type   2     16     8.1   0.018  0.982
Residuals    20   8816   440.8
29 observations deleted due to missingness
>
```

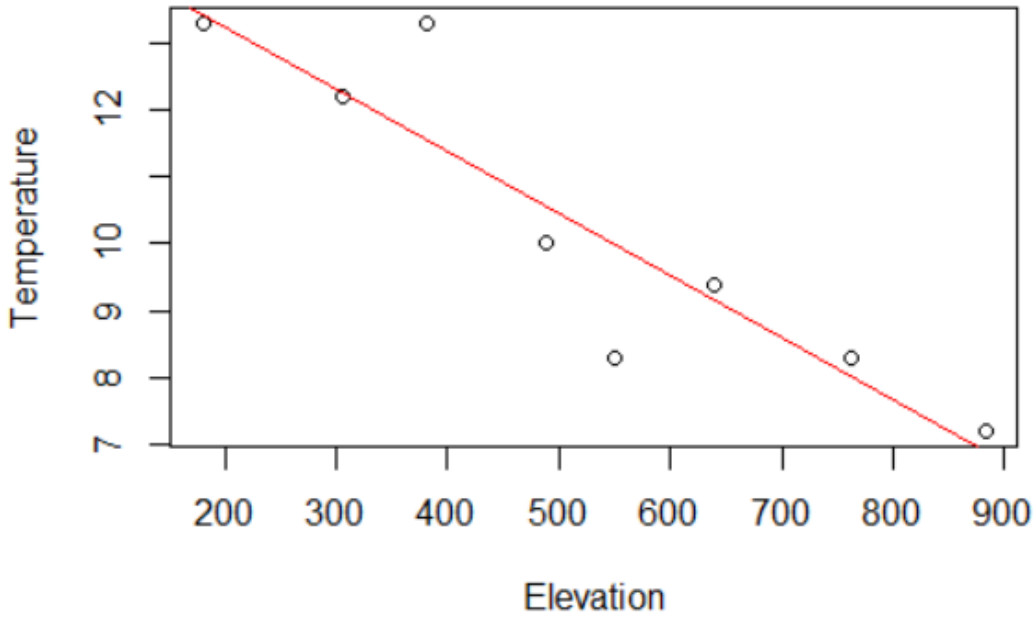
经分析发现组间无显著性差异

#4

```
#4
Elevation <- c(180,305,381,488,549,640,762,883)
Temperature <- c(13.3,12.2,13.3,10.0,8.3,9.4,8.3,7.2)
matrix1 <- data.frame(Elevation,Temperature)
plot(Temperature~Elevation, data = matrix1)
regression_line <- lm(Temperature~Elevation, data = matrix1)
abline(regression_line, col='red')
summary(regression_line)$coefficients[2,1]
#summary(regression_line)$coefficients[,1] or coef(regression_line) 第一个是截距，第二个是斜率
#learn from https://blog.csdn.net/dingchenxixi/article/details/50543822
lapse_rate <- summary(regression_line)$coefficients[2,1]*1000
lapse_rate
```

用 lm() 函数生成拟合直线，用 abline 把直线添加进已画好的图表里。

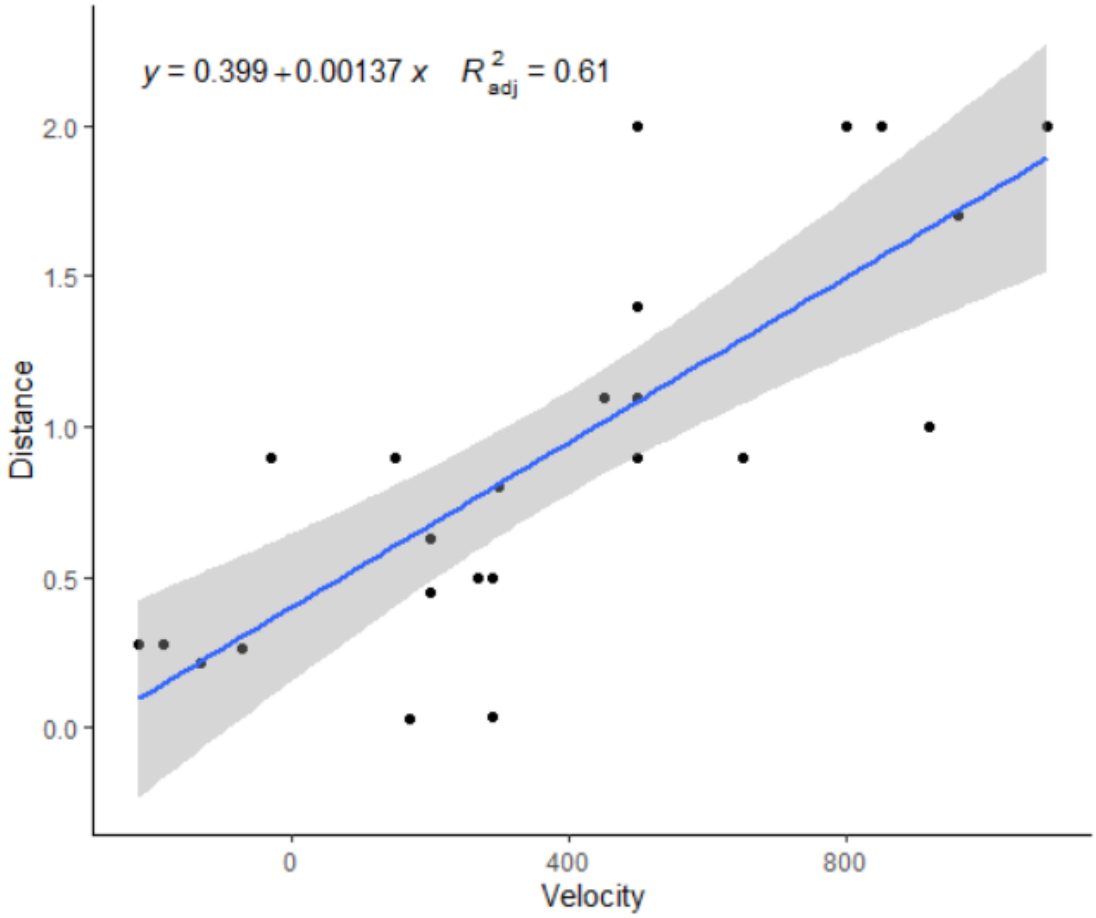
在 summary 中寻找斜率的对应值，经过调整单位运算后得出结果。计算的斜率为-9.312104




```
> Elevation <- c(180,305,381,488,549,640,762,883)
> Temperature <- c(13.3,12.2,13.3,10.0,8.3,9.4,8.3,7.2)
> matrix1 <- data.frame(Elevation,Temperature)
> plot(Temperature~Elevation, data = matrix1)
> regression_line <- lm(Temperature~Elevation, data = matrix1)
> abline(regression_line, col='red')
> summary(regression_line)$coefficients[2,1]
[1] -0.009312104
> #summary(regression_line)$coefficients[,1] or coef(regression_line) 第一个是截距，第二个是斜率
> #learn from https://blog.csdn.net/dingchenxixi/article/details/50543822
> lapse_rate <- summary(regression_line)$coefficients[2,1]*1000
> lapse_rate
[1] -9.312104
>
```

#5

```
#5
#5.1
install.packages("ggpmisc")
library(ggpmisc)
BBT <- read.csv("PS3_data.csv",header = T)
BBT_tbl <- as_tibble(BBT)
names(BBT_tbl)
BBT_tbl %>%
  filter(Nebula != "") %>%
  select(Nebula,Velocity,Distance) %>%
  ggplot(aes(y = Distance,x = Velocity))+
  geom_point()+
  #5.2
  geom_smooth(method="lm", formula = y ~ x)+ # learn from https://blog.csdn.net/weixin_42933967/article/details/96
  stat_poly_eq(aes(label = paste(..eq.label.., ..adj.rr.label.., sep = '~~~~')), formula = y ~ x, parse = T) +
  theme_classic()
#5.3
#the first assumption sounds reasonable because the universe
#was come from a single point( big bang theory)
#5.4
```



调用了新的包：ggpmise 用于计算回归方程。经计算得宇宙的年龄为 990 亿年。

#5.4

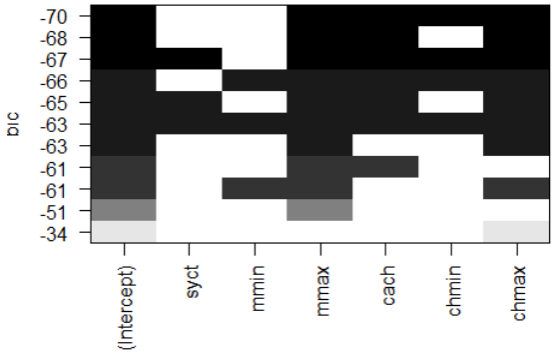
如果测量的点能更准确，得到的回归曲线也当然更精确。

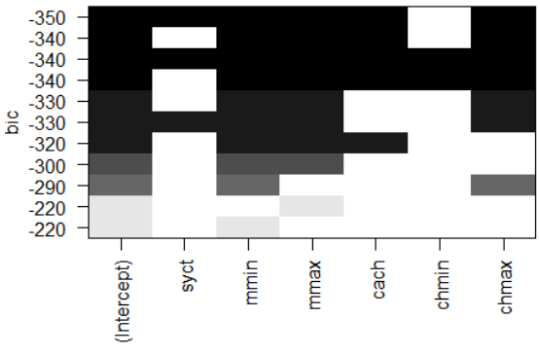
#6

```
#6
install.packages('leaps')
library(tidyr)
library(dplyr)
library(ggplot2)
library(leaps)
library(MASS)
data(cpus)
head(cpus)
nrow(cpus)
#make the order random
random_selcet <- runif(209,0,100)
cpu1 <- cbind(cpus,random_selcet)
cpu_random <- cpu1 %>%
  arrange(desc(random_selcet))
#seperate it into two group
cpu_train_set <- cpu_random[1:167,]
cpu_test_set <- cpu_random[168:209,]
#6.1
train_subset_result <- regsubsets(perf ~ syct+mmin+mmax+cach+chmin+chmax, data=cpu_train_set, nbest=2, nvmax = 6)
plot(train_subset_result)

#6.2
test_subset_result <- regsubsets(perf ~ syct+mmin+mmax+cach+chmin+chmax, data=cpu_test_set, nbest=2, nvmax = 6)
plot(test_subset_result)
```

在 cpu 组后加上一列从 1 到 100 的随机数，按大小排列。用 nrow 读出矩阵长度，按 80%和 20%的数据量划分为两组。用 regsubset 进行回归拟合（虽然还看不懂）





左图为学习（train）组的，右图为测试（test）组的

#7

```
#7
data_Lab <- read.csv("CMMNDHgfinal.csv",header = T)
data_tbl <- as_tibble(data_Lab)
names(data_tbl)
#which kind of bird have the higest Hg concentration
data_tbl %>%
  filter(Species != "") %>%
  group_by(Species) %>%
  #summarise(Hg_mean = mean(Hgppm)) %>%
  #arrange(desc(Hg_mean)) %>%
  ggplot(aes(x=Species,y=Hgppm),head = T, fill=Species)+
  theme_classic()+
  geom_boxplot()

#t test
MYWA <- data_tbl %>%
  filter(Species == "MYWA")%>%
  select(Hgppm)
NOWA <- data_tbl %>%
  filter(Species == "NOWA")%>%
  select(Hgppm)
t.test(MYWA,NOWA)

#ANOVA test
anova_bird <- aov(Hgppm~Species,data=data_tbl)
summary(anova_bird)

#linear regression model
NOWA_LG <- data_tbl %>%
  filter(Species == "NOWA")
plot(Deterium~Hgppm, data = NOWA_LG)
regression_line <- lm(Deterium~Hgppm, data = NOWA_LG)
abline(regression_line, col='red')
```

用了组里马艳菊师姐的鸟类——重金属（Hg）的数据作为分析对象。对汞浓度最高的两种鸟类进行 t 检验和显著性差异分析。

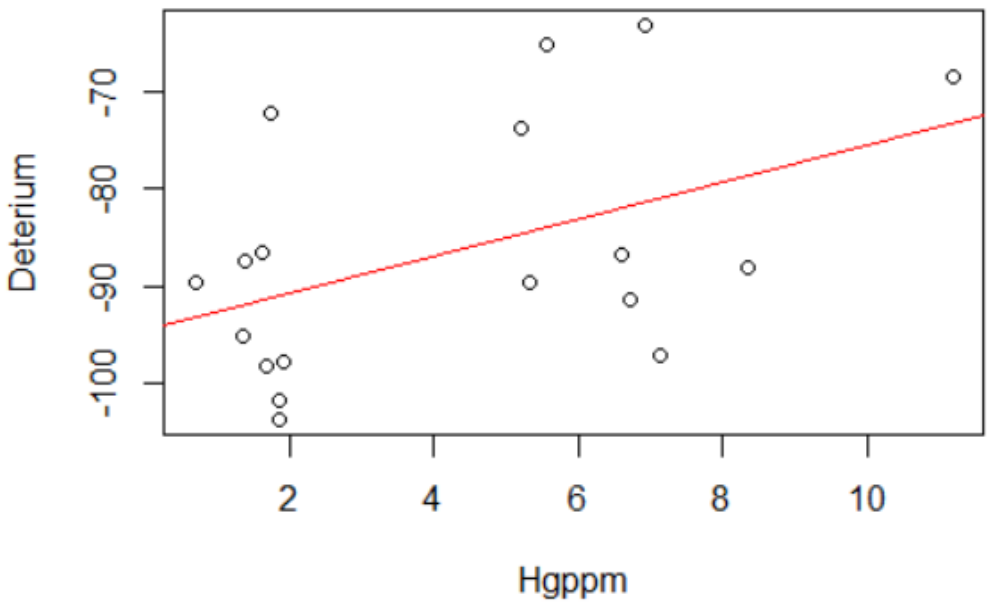
回归曲线用了汞浓度和实验分析所需材料？进行分析（好像不是很好看）

```
> t.test(MYWA,NOWA)

        Welch Two Sample t-test

data:  MYWA and NOWA
t = 0.20376, df = 28.834, p-value = 0.84
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.656957  3.244789
sample estimates:
mean of x mean of y
 4.572010  4.278094
```

```
> summary(anova_bird)
             Df Sum Sq Mean Sq F value Pr(>F)
Species       7  430.8   61.55   21.66 <2e-16 ***
Residuals   477 1355.3    2.84
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
9 observations deleted due to missingness
```


Deterium
-70
-80
-90
-100
2
4
6
8
10
Hgppm