Assignment 3

1. Cloud Seeding

1.1 Plot two box plots side-by-side of data from the two groups. Describe the distributions.

非催化和催化降雨数据（Rainfall from unseeded days and seeded days）分别计入 Unseeded_Data 和 Seeded_Data。首先利用 boxplot 绘制箱图。

```
Unseeded_Data <- c(1202.6, 830.1, 372.4, 345.5, 321.2, 244.3, 163.0, 147
Seeded_Data <- c(2745.6, 1697.1, 1656.4, 978.0, 703.4, 489.1, 430.0, 334
boxplot(cbind(Unseeded_Data,Seeded_Data))
```

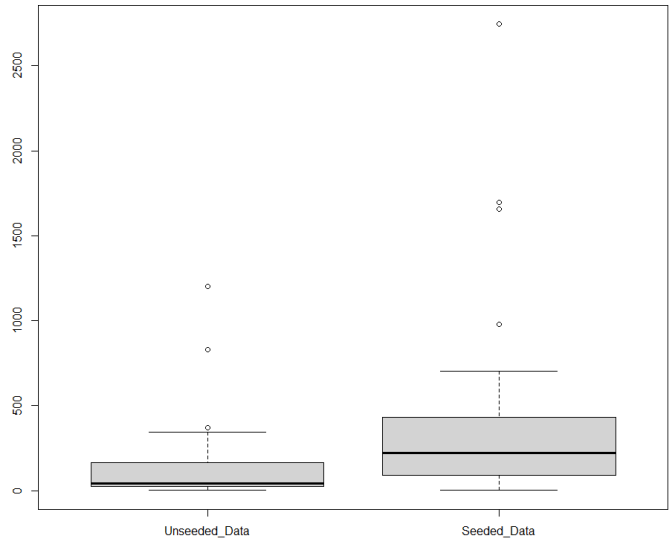


图 1

为了进一步对两组数据的分布进行分析，还另外绘制了柱状图和降雨-时间散点图。

```
hist(Unseeded_Data)
hist(Seeded_Data)
day <- array(data=1:26)
plot(Unseeded_Data~day)
plot(Seeded_Data~day)
```

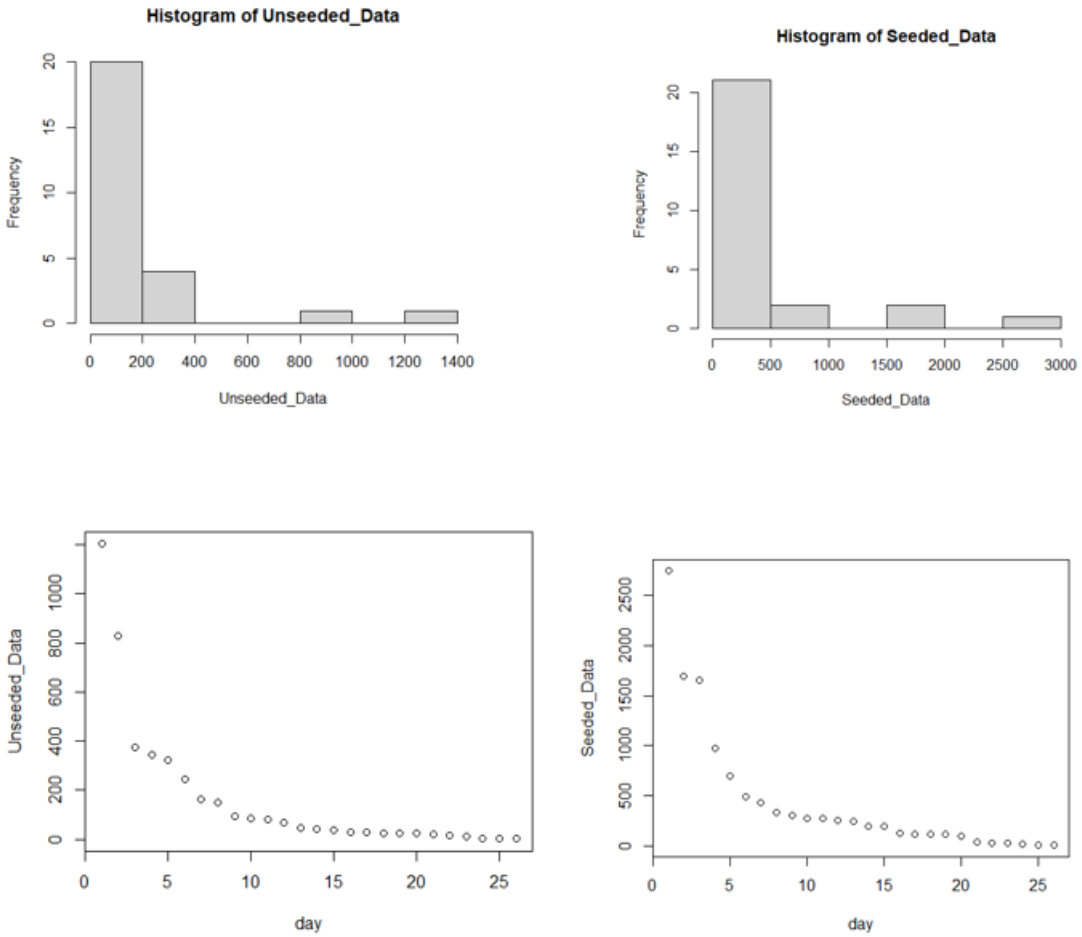


图 2

结果与讨论：从箱图（图 1）可以看出，催化之后的降雨量（Rainfall from seeded days）更大。根据图 2 可是，两组数据首先不服从正态分布，其次降雨量随着时间减小，与时间不呈现明显的线性关系。

1.2 Did cloud seeding have an effect on rainfall in this experiment? If so, how much?

利用 t.test 对两组数据进行对比分析。

```
#1.2
t.test(Unseeded_Data, Seeded_Data)
```

```
        Welch Two Sample t-test

data:  Unseeded_Data and Seeded_Data
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
```

结果与讨论：t.test 的结果显示，两组数据的平均值不一样，且催化之后的平均降雨量（mean rainfall from seeded days）更大。因此，进行催化确实会对降雨量产生影响，催化之后的平均降雨量（mean rainfall from seeded days）为 491.9731，未催化之后的平均降雨量（mean rainfall from unseeded days）为 164.5731，差异显著。

2. Was Tyrannosaurus Rex Warm-Blooded?

Is there evidence that the means are different for the different bones? Does the dataset support Tyrannosaurus Rex is warm-blooded or not?

首先对数据进行读取，预先绘制了箱图，对数据进行一个简单的了解。

```
bone_Data <- read.delim2("bone.txt", head=TRUE)
bone_Data[1,2]
isotopic <- matrix(nrow=72, ncol=2)
isotopic[1.2]
for (i in 1:12) {
  for (j in 1:6) {
    k <- j+6*(i-1)
  isotopic[k, 1] = bone_Data[i, 1]
  isotopic[k, 2] = bone_Data[i, j+1]
  }
}
colnames(isotopic)=c("Bone","Oxygen.isotopic.composition")
isotopic_new <- as_tibble(isotopic)
isotopic_new1 <- isotopic_new  %>%
  mutate(OIC= as.numeric(Oxygen.isotopic.composition))
isotopic_new1 %>%
  group_by(Bone) %>%
  summarise(
    count = n(),
    mean_isotopic = mean(OIC, na.rm = TRUE),
    sd_isotopic = sd(OIC, na.rm = TRUE)
  )
ggplot(isotopic_new1, aes(x = Bone, y = OIC, fill = Bone)) +
  geom_boxplot() +
  theme_classic()
```

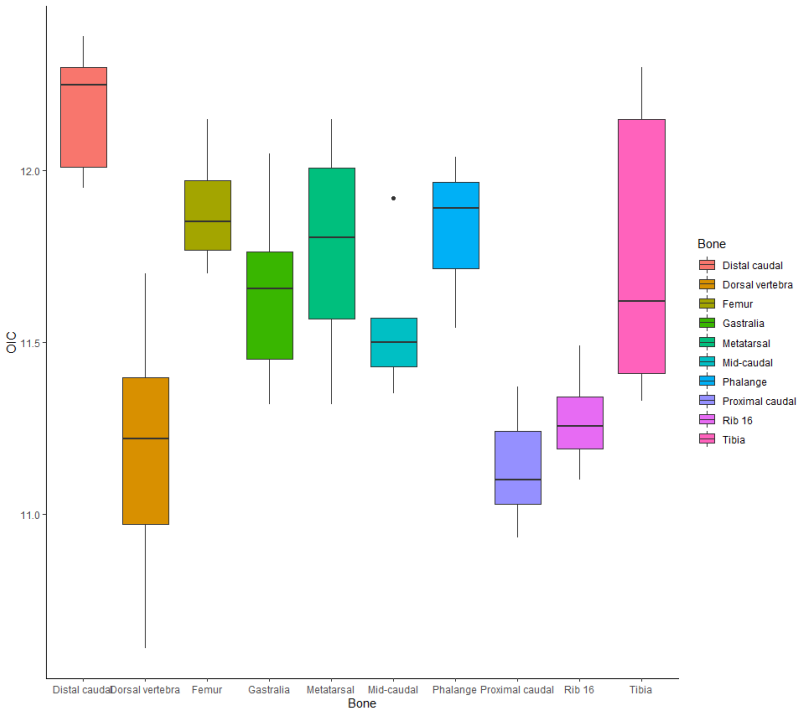


图 3

此处的问题是证明不同骨骼的平均值不一样，这里使用 aov 来实现。

```
anova_one_way <- aov(OIC ~ Bone, data = isotopic_new1)
summary(anova_one_way)
TukeyHSD(anova_one_way)
```

```
> summary(anova_one_way)
            Df Sum Sq Mean Sq F value   Pr(>F)
Bone         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
20 observations deleted due to missingness
```

```
> TukeyHSD(anova_one_way)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = OIC ~ Bone, data = isotopic_new1)

$Bone
                                        diff          lwr         upr
Dorsal vertebra-Distal caudal     -0.97800000 -1.4945276301 -0.46147237
Femur-Distal caudal               -0.29250000 -0.9251145659  0.34011457
Gastralia-Distal caudal           -0.53666667 -1.1077094477  0.03437611
Metatarsal-Distal caudal          -0.41000000 -1.0426145659  0.22261457
Mid-caudal-Distal caudal          -0.62600000 -1.2224347326 -0.02956527
Phalange-Distal caudal            -0.35666667 -1.0453701735  0.33203684
Proximal caudal-Distal caudal     -1.04833333 -1.6193761143 -0.47729055
Rib 16-Distal caudal              -0.90500000 -1.5376145659 -0.27238543
Tibia-Distal caudal               -0.41800000 -1.0144347326  0.17843473
Femur-Dorsal vertebra              0.68550000  0.1275863943  1.24341361
Gastralia-Dorsal vertebra          0.44133333 -0.0456535866  0.92832025
Metatarsal-Dorsal vertebra         0.56800000  0.0100863943  1.12591361
Mid-caudal-Dorsal vertebra         0.35200000 -0.1645276301  0.86852763
Phalange-Dorsal vertebra           0.62133333  0.0005443815  1.24212229
Proximal caudal-Dorsal vertebra -0.07033333 -0.5573202532  0.41665359
Rib 16-Dorsal vertebra             0.07300000 -0.4849136057  0.63091361
```

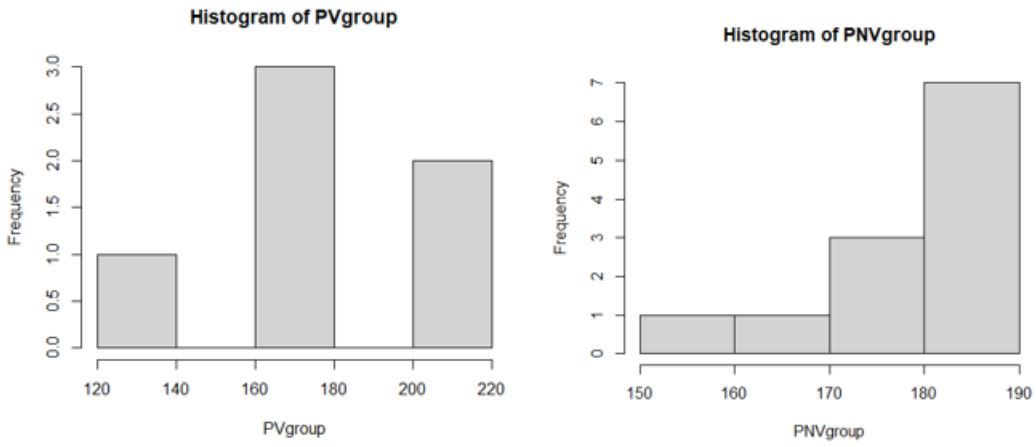
结果与讨论：脊椎动物骨磷酸盐的氧同位素组成与骨形成时的体温有关，同位素差异说明了温度的差异。p 值小于 0.5，表示不同骨骼的同位素值具有统计学差异。但是组间平均值的大小差别不是特别大，根据图 3 可以整体看出来，同位素平均值在 11~13 之间，没有显著差别，符合温血动物（warm-blooded anaimals）的条件，即温血动物骨骼温度差别不大。因此，霸王龙（Tyrannosaurus rex skeleton）应该可以认为是温血动物。

3. Vegetarians and Zinc

What evidence is there that pregnant vegetarians tend to have lower zinc levels than pregnant nonvegetarians?

将素食怀孕者（pregnant vegetarians）的锌（zinc）含量和非素食怀孕者（pregnant nonvegetarians）的锌含量分别记为 PVgroup 和 PNVgroup 两个数据组，通过直方图和箱图对两组数据的分布先进行分析，最后运用 t.test 对两组数据的差异进行研究。

```
vp_Data <- read.delim("problem3.txt", head=TRUE)
PVgroup <- vp_Data$Pregnant.vegetarians
PNVgroup <- vp_Data$Pregnant.nonvegetarians
mean(vp_Data$Pregnant.vegetarians, na.rm=T)
mean(vp_Data$Pregnant.nonvegetarians, na.rm=T)
hist(PVgroup)
hist(PNVgroup)
boxplot(cbind(PVgroup, PNVgroup))
t.test(PVgroup, PNVgroup)
```



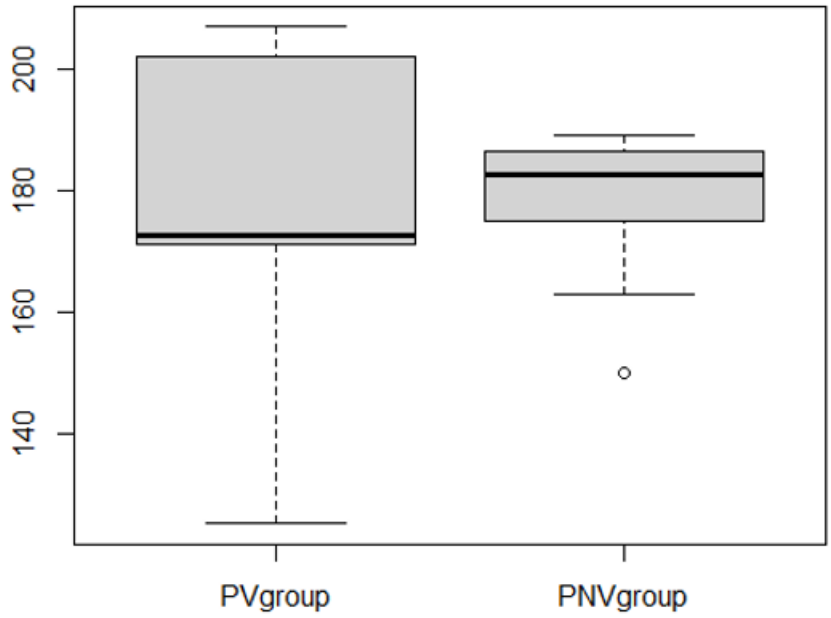


图 4

```
> t.test(PVgroup, PNVgroup)

        Welch Two Sample t-test

data:  PVgroup and PNVgroup
t = -0.28849, df = 5.8141, p-value = 0.783
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -34.21367  27.04700
sample estimates:
mean of x mean of y
 175.0000  178.5833
```

结果与讨论：从 t.test 结果看出来，素食怀孕者（pregnant vegetarians）的锌（zinc）含量均值为 175，非素食怀孕者（pregnant nonvegetarians）的锌含量均值

为 178.5833，素食怀孕者（pregnant vegetarians）的锌（zinc）含量均值偏低，但两者的差值不大。结合图 4 可知，首先直方图的结果显示，PVgroup 中出现锌含量>200 的值出现，但是出现次数很少，锌含量>200 值的出现导致了素食怀孕者（pregnant vegetarians）的锌（zinc）含量均值偏高；同时从箱图也可以看出，尽管 PVgroup 的最大值高于 PNVgroup，但是 PVgroup 均值反而偏低，也印证了上述的推测。因此，相比非素食怀孕者（pregnant nonvegetarians），素食怀孕者（pregnant vegetarians）的锌（zinc）含量有偏低的趋势。

4. Atmospheric Lapse Rate

Draw a scatter plot with regression line, and investigate if the lapse rate is 9.8 degrees C km^{-1}.

绘制温度-时间散点图，并进行简单的线性回归分析。

```
temp_Data <- read.delim("problem4.txt", head=TRUE)
names(temp_Data)
temp_Data1 <- temp_Data%>%
  mutate(Elevation..km.=Elevation..m./1000)
plot(Temperature..degrees.C. ~ Elevation..km., data=temp_Data1,
     xlab = "Elevation",
     ylab = "Temperature",
     main = "Temperature vs Elevation",
     pch = "+",
     cex = 2,
     col = "navy")
fit <- lm(Temperature..degrees.C. ~ Elevation..km., data=temp_Data1)
summary(fit)
abline(fit, lwd = 5, col = "red")
```

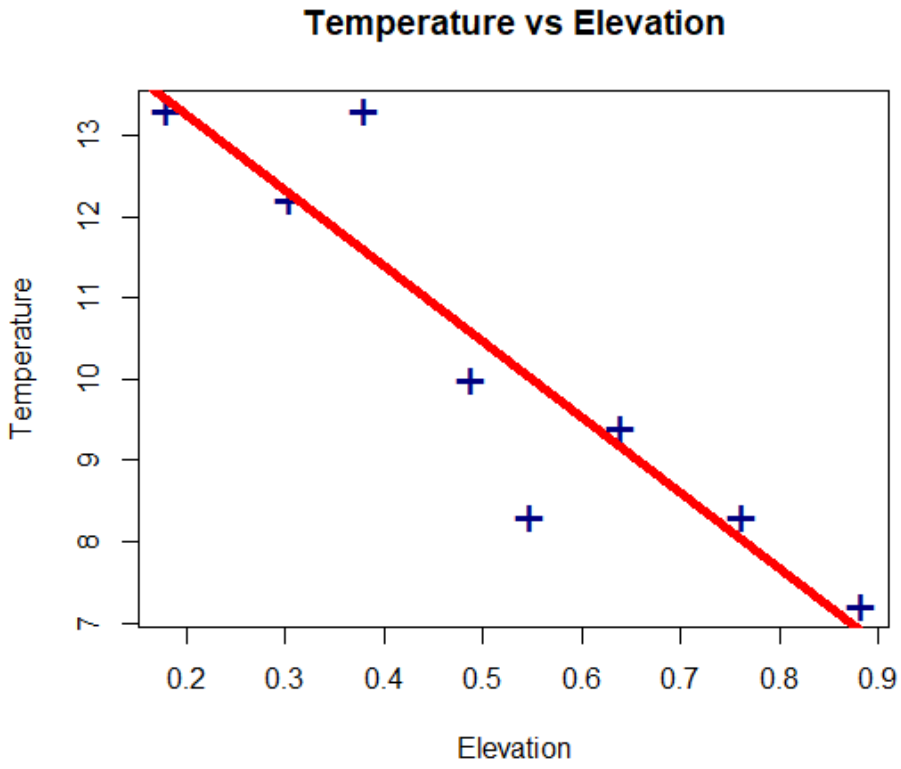


图 5

```
> summary(fit)

Call:
lm(formula = Temperature..degrees.C. ~ Elevation..km., data = temp_Data1)

Residuals:
     Min       1Q   Median       3Q      Max
-1.71254 -0.25668  0.07508  0.27763  1.72303

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      15.1249     0.9483  15.950 3.86e-06 ***
Elevation..km.   -9.3121     1.6698  -5.577  0.00141 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.04 on 6 degrees of freedom
Multiple R-squared:  0.8383,    Adjusted R-squared:  0.8113
F-statistic:  31.1 on 1 and 6 DF,  p-value: 0.001411
```

结果与讨论：根据线性拟合结果可知，温度的下降速率为 9.3121degrees C km^{-1}。

5. The Big Bang Theory

5.1 Make a scatter plot with distance as the Y-axis and recession velocity as the X-

axis. Describe what you see.

绘制距离-速率散点图。

```
#5.1
RV_Data <- read.delim("problem5.txt", head=TRUE)
plot(Distance ~ Velocity, data=RV_Data,
     xlab = "Velocity",
     ylab = "Distance",
     main = "Distance vs Velocity",
     pch = "+",
     cex = 2,
     col = "navy")
```

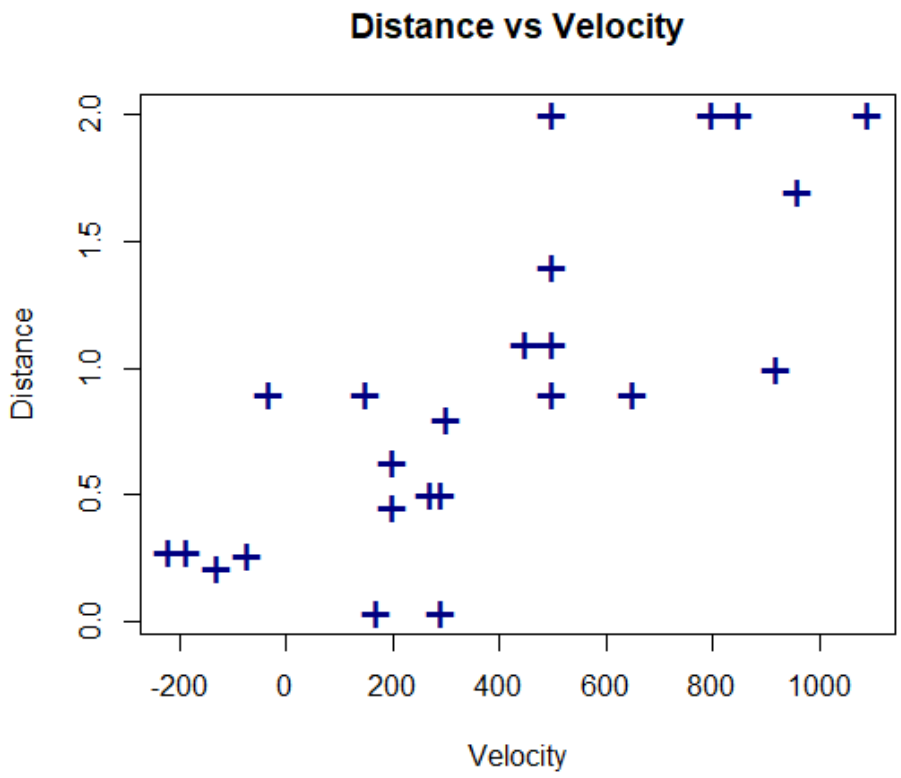


图 6

结果与讨论：随着距离变远，速度变大，可尝试做简单的线性回归分析。

5.2 Add a simple linear regression line to the above scatter plot.

在 5.1 的基础之上进行简单的线性回归分析。

```
#5.2
fit <- lm(Distance ~ Velocity, data=RV_Data)
summary(fit)
abline(fit, lwd = 5, col = "red")
```

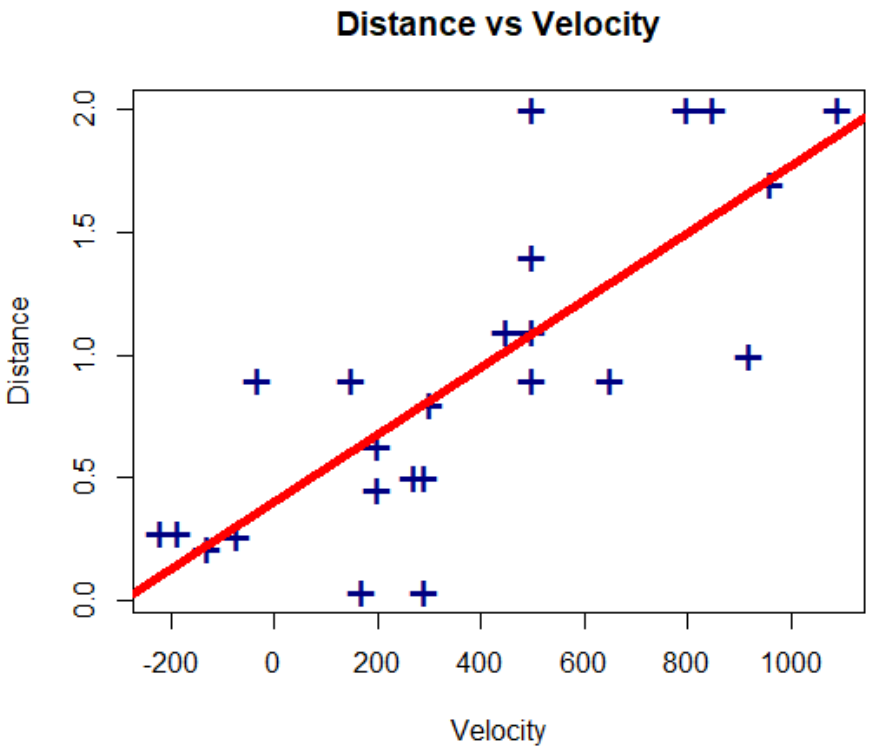


图 7

```
> summary(fit)

Call:
lm(formula = Distance ~ Velocity, data = RV_Data)

Residuals:
    Min      1Q  Median      3Q     Max
-0.7632 -0.2352 -0.0088  0.2072  0.9144

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.3990982  0.1184697   3.369  0.00277 **
Velocity    0.0013729  0.0002274   6.036 4.48e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.405 on 22 degrees of freedom
Multiple R-squared:  0.6235,    Adjusted R-squared:  0.6064
F-statistic: 36.44 on 1 and 22 DF,  p-value: 4.477e-06
```

结果与讨论：线性拟合结果的 R^2 为 0.6235，说明线性拟合结果不是特别好。

5.3 If Hubble's* Big Bang Theory *is correct, explain why the following two assumptions about the regression line you made in 5.2 need to be true:

- ***The intercept should be zero***
- ***And the slope is the age of the universe***

Address the first assumption with your regression results; and estimate the age of the universe.

显示线性拟合结果的截距和斜率数值。

```
#5.3
summary(fit)$coefficients
#explained by Yuan Li
```

```
> summary(fit)$coefficients
               Estimate   Std. Error  t value     Pr(>|t|)
(Intercept) 0.399098216 0.1184697343 3.368778 2.770039e-03
velocity    0.001372936 0.0002274443 6.036362 4.477491e-06
```

结果与讨论：根据 Edwin Hubble 的大爆炸理论来说，宇宙起源于大爆炸，因此，最初的距离理论上是 0，也就是线性回归中的截距。根据第二个假设，线性回归的斜率是宇宙的年龄，斜率结果是 0.001372936，1 megaparsec=$3.09*10^{19}$km，速率的单位是 km/s，age=$3.09*10^{19}*0.001372936/(60*60*24*365)=1.35*10^{9}$ year。宇宙的年龄为 13.5 亿年。

5.4 Explain why improved measurement of distance would lead to more precise estimates of the regression coefficients.

结果与讨论：提高距离的测试精度，使得距离的结果更加接近真实值，那么线性拟合的结果就更加接近于真实值，所以，回归参数的结果就更加准确。

注：5.3 和 5，4 请教李元同学解释了一下题意，帮助解题。

6. CPU Performance

6.1 For the train set, fit the best subset regression between predictor variable perf and response variables including syct, mmin, mmax, cach, chmin, and chmax.

根据题干，先将数据分为 train 和 test 两组，然后利用 train 的数据进行回归分析。

```r
library(MASS)
data(cpus)
#6.1
sample_index <- sample(nrow(cpus),nrow(cpus)*0.80)
cpus_train <- cpus[sample_index,]
cpus_test  <- cpus[-sample_index,]
model_1 <- lm(perf~syct+mmin+mmax+cach+chmin+chmax, data=cpus_train)
summary(model_1)
library(leaps)
subset_result <- regsubsets(perf~syct+mmin+mmax+cach+chmin+chmax, data=cpus_train, nbest=2, nvmax = 6)
plot(subset_result, scale="bic")
```

```
> summary(model_1)

Call:
lm(formula = perf ~ syct + mmin + mmax + cach + chmin + chmax,
    data = cpus_train)

Residuals:
    Min      1Q  Median      3Q     Max
-170.49  -21.94    3.47   20.99  302.40

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.891e+01  7.996e+00  -4.866 2.71e-06 ***
syct         3.474e-02  1.624e-02   2.139   0.0339 *
mmin         1.179e-02  2.058e-03   5.728 4.91e-08 ***
mmax         4.982e-03  6.358e-04   7.836 6.11e-13 ***
cach         7.624e-01  1.500e-01   5.082 1.03e-06 ***
chmin        1.381e+00  9.410e-01   1.467   0.1442
chmax        8.384e-01  2.090e-01   4.011 9.25e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 51.24 on 160 degrees of freedom
Multiple R-squared:  0.8399,    Adjusted R-squared:  0.8339
F-statistic: 139.9 on 6 and 160 DF,  p-value: < 2.2e-16
```

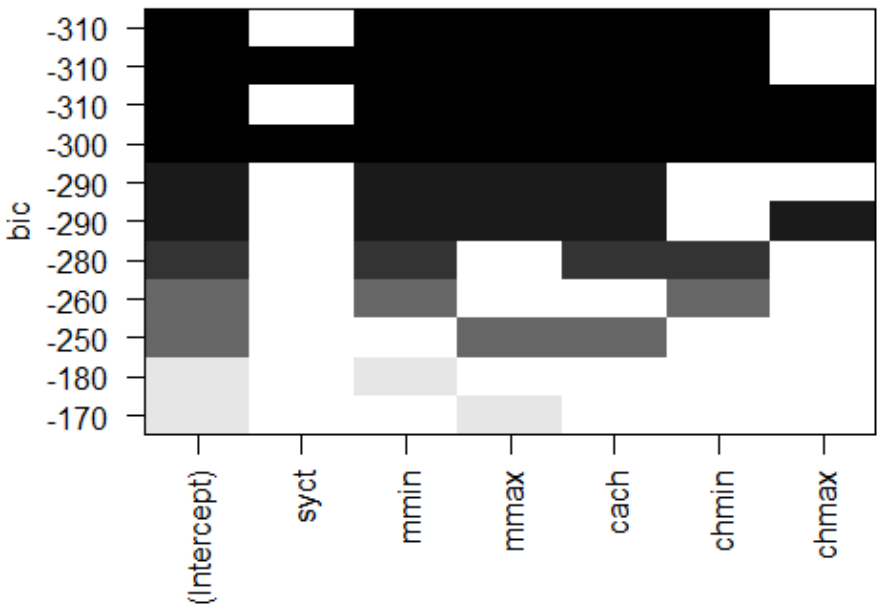


图 8

结果与讨论：回归结果的 R^2 为 0.8399，拟合结果较好。

6.2 Apply the best regression model to the test set, and compare your predicted perf values with the actual values that provided in the test set. Quantify the mean bias between predicted perf values and provided perf values.

利用 test 的数据进行检验分析，比较拟合结果与 test 中的真实值，量化平均偏差。

```
#6.2
perf_predict <- predict(model_1,cpus_test)
plot(cpus_test$perf, perf_predict)
cor(cpus_test$perf, perf_predict)
# Mean predicted value
mean(perf_predict)

# Mean actual value
mean(cpus_test$perf)

# Relative mean bias
(mean(perf_predict) - mean(cpus_test$perf))/
  mean(cpus_test$perf)*100
```

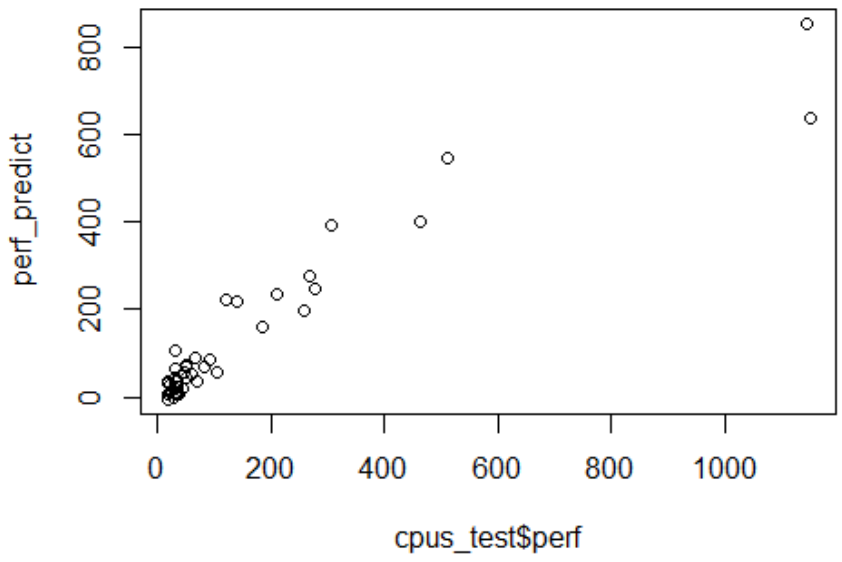


图 9

```
> cor(cpus_test$perf, perf_predict)
[1] 0.947589
> perf_predict <- predict(model_1,cpus_test)
> plot(cpus_test$perf, perf_predict)
> cor(cpus_test$perf, perf_predict)
[1] 0.947589
> # Mean predicted value
> mean(perf_predict)
[1] 131.1237
>
> # Mean actual value
> mean(cpus_test$perf)
[1] 150.2857
>
> # Relative mean bias
> (mean(perf_predict) - mean(cpus_test$perf))/
+   mean(cpus_test$perf)*100
[1] -12.7504
```

结果与分析：根据图 9 可知，除了最边上的两个点，其他位置拟合值与实测值的差别不大，拟合结果较好。拟合值和实测值的相关性较高，平均偏差为-12.7534%。

7. Analysis of Data Sets from Your Group

To better understand the combined effects of flow rate and NO_3^- concentration on denitrification rate and NO_3^- removal efficiency in the low-permeability media, a set of column experiments with different flow rates and injected NO_3^- concentrations were conducted.

Table1 Summary of five column experiments, in terms of flow rates, influent NO_3^- concentrations and durations

Experiment No.	Flow rate (m/d)	Influent NO_3^- concentration (mmol/L)	NO_3^- injected time (h)	Time of experiment(h)
CRslow	0.023	3.23	133.5	628.21
CRmid/CChigh	0.044	3.23	58	530.49
CRfast	0.070	3.23	42.2	532.36

CCmid	0.044	1.29	60.32	323.10
CClow	0.044	0.24	59.82	360.23

Table2 The calibration results of the maximum rate constants for denitrification (γ_{nit})

Experiment No.	Initial γ_{nit} (mol/L/s)	γ_{nit} (mol/L/s)
CRslow		1.8×10^{-8}
CRmid/CChigh		1.9×10^{-8}
CRfast	8×10^{-10}	2.0×10^{-8}
CCmid		1.29×10^{-8}
CClow		4.5×10^{-9}

7.1 Define a simple research question that can be tested with the t-test. Test your question with R, and describe your findings.

Experiment No.	NO in problem7.txt
CRslow	CE1
CRmid/CChigh	CE2
CRfast	CE3
CCmid	CE4
CClow	CE5

Q: Compare the nitrate concentrations in CRmid and CRfast.

```r
#7.1
EX_Data <- read.delim("problem7.txt", head=TRUE)
CE2_Data <- EX_Data %>%
  filter(NO == "CE2")
CE3_Data <- EX_Data %>%
  filter(NO == "CE3")
sample1 <- CE2_Data$NO3
sample2 <- CE3_Data$NO3
t.test(sample1, sample2)
```

A: With the increase of velocity, the mean of nitrate concentrations increased.

```
> t.test(sample1, sample2)

        Welch Two Sample t-test

data:  sample1 and sample2
t = -0.91497, df = 59.327, p-value = 0.3639
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.4642917  0.1728988
sample estimates:
mean of x mean of y
0.5024952 0.6481917
```

7.2 Define a simple research question that can be tested with the ANOVA. Test your question with R, and describe your findings.

Q: Compare the nitrate concentrations in different operational conditions.

```r
#7.2
anova_one_way <- aov(NO3 ~ NO, data = EX_Data)
summary(anova_one_way)
TukeyHSD(anova_one_way)
```

A: The p-value is lower than the usual threshold of 0.05. there is a statistical difference between the groups.

```
> summary(anova_one_way)
             Df Sum Sq Mean Sq F value   Pr(>F)
NO            4  11.51  2.8773   17.11 1.36e-11 ***
Residuals   151  25.40  0.1682
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(anova_one_way)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = NO3 ~ NO, data = EX_Data)

$NO
                 diff        lwr        upr     p adj
CE2-CE1  0.4854110375  0.2153340  0.7554881 0.0000182
CE3-CE1  0.6311074990  0.3610304  0.9011846 0.0000000
CE4-CE1  0.0617556177 -0.2320081  0.3555194 0.9777786
CE5-CE1 -0.0005966699 -0.2782411  0.2770478 1.0000000
CE3-CE2  0.1456964615 -0.1373814  0.4287743 0.6151006
CE4-CE2 -0.4236554198 -0.7294144 -0.1178964 0.0017611
CE5-CE2 -0.4860077074 -0.7763142 -0.1957012 0.0000785
CE4-CE3 -0.5693518813 -0.8751109 -0.2635929 0.0000082
CE5-CE3 -0.6317041689 -0.9220107 -0.3413977 0.0000001
CE5-CE4 -0.0623522876 -0.3748157  0.2501111 0.9816758
```

7.3 Define a simple research question that can be tested with a simple linear regression model. Test your question with R, and describe your findings.

Q: Explore the relationship between the denitrification rate and the velocity.

Experiment No.	Flow rate (m/d)	γ_{nit} (mol/L/s)
CRslow	0.023	1.8×10^{-8}
CRmid/CChigh	0.044	1.9×10^{-8}
CRfast	0.070	2.0×10^{-8}

```
#7.3
vEX_Data <- read.delim("problem7_3.txt", head=TRUE)
names(vEX_Data)
fit <- lm (R_nit~Flow.rate, data = vEX_Data)
summary(fit)
plot(R_nit~Flow.rate, data = vEX_Data,
     xlab = "Flow.rate",
     ylab = "R_nit",
     main = "R_nit vs Flow.rate",
     pch = 20,
     cex = 2,
     col = "navy")
abline(fit, lwd = 5, col = "red")
```

A: R-squared is 0.9962. It can be assumed that the relationship between the denitrification rate and the velocity corresponds to a simple linear regression model.

```
> summary(fit)

Call:
lm(formula = R_nit ~ Flow.rate, data = vEX_Data)

Residuals:
         1          2          3
-3.909e-11  7.066e-11 -3.157e-11

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.706e-08  1.290e-10  132.27  0.00481 **
Flow.rate   4.239e-08  2.604e-09   16.28  0.03905 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.67e-11 on 1 degrees of freedom
Multiple R-squared:  0.9962,    Adjusted R-squared:  0.9925
F-statistic: 265.1 on 1 and 1 DF,  p-value: 0.03905
```

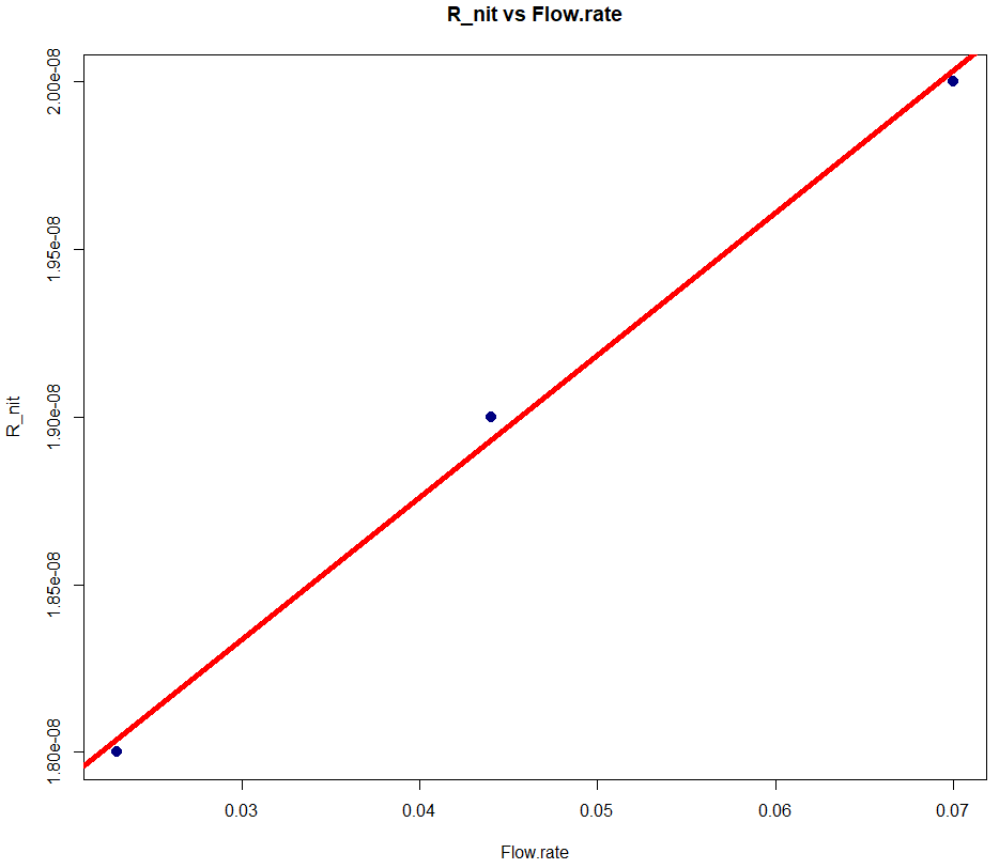

R_nit vs Flow.rate
R_nit
2.00e-08
1.95e-08
1.90e-08
1.85e-08
1.80e-08
0.03
0.04
0.05
0.06
0.07
Flow.rate

图 10