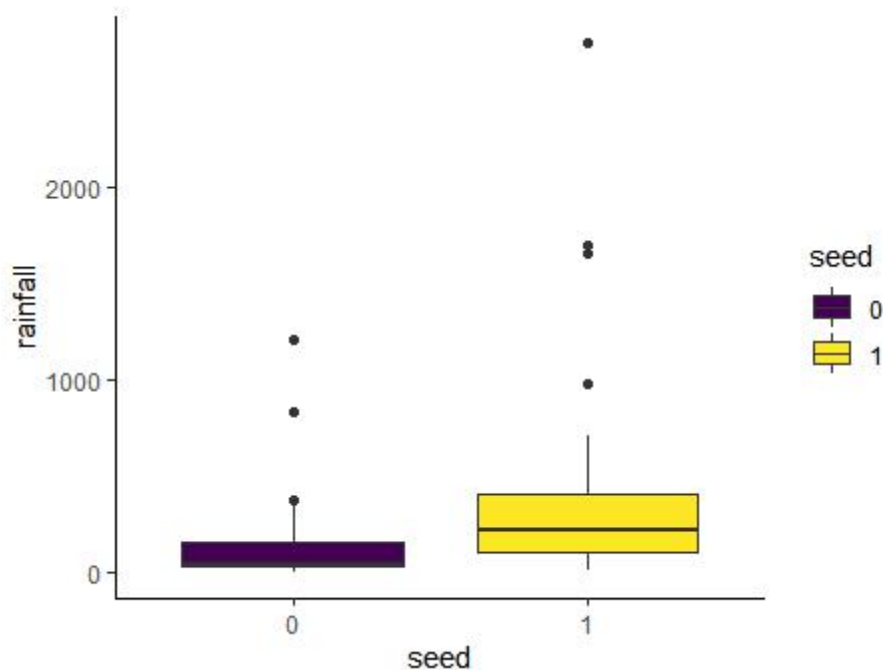


Assignment 03

1. Cloud Seeding

1.1 [5 points] Plot two box plots side-by-side of data from the two groups. Describe the distributions.

Answer: The second box has more dispersion degree than first box.



1.2 [5 points] Did cloud seeding have an effect on rainfall in this experiment? If so, how much?

Answer: According the result of anova, we find cloud seeing don't have significant effect on rainfall in this experiment (Pvalue>0.05)

```

      Df  Sum Sq Mean Sq F value Pr(>F)
seed    1 1000360 1000360   3.993 0.0511 .
Residuals 50 12525457  250509
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

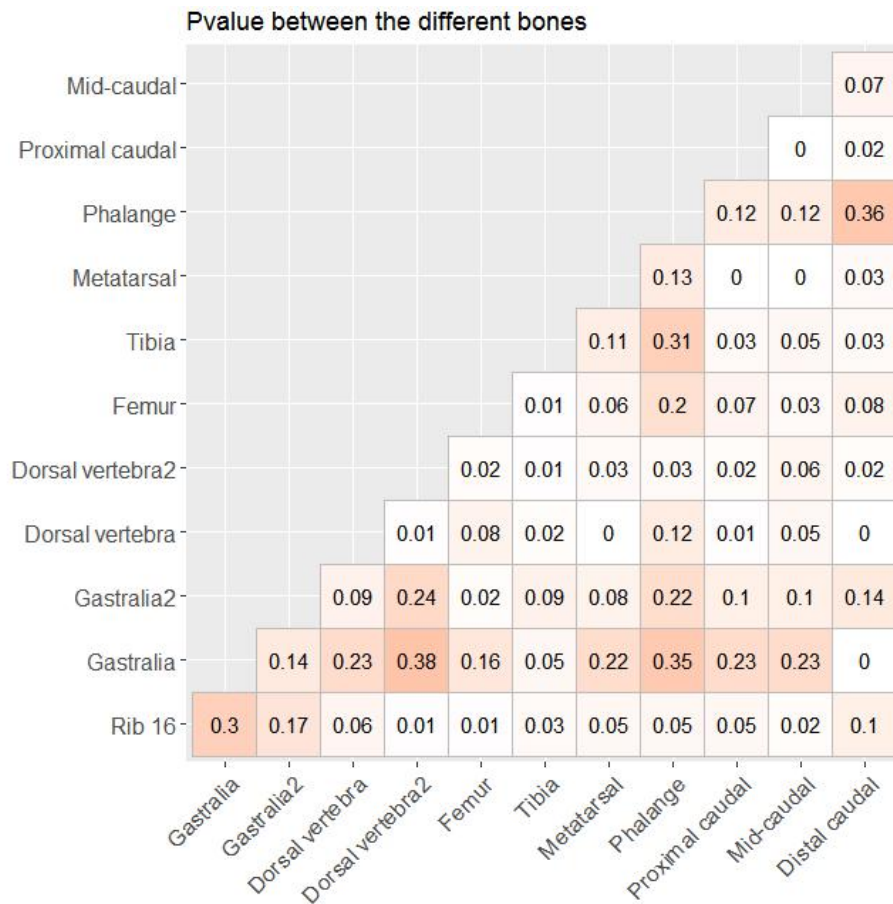
2. Was Tyrannosaurus Rex Warm-Blooded?

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

Answer2:

1, Yes, we can see the Pvalue from the Figure, the Pvalue between most two bone are less than 0.05.

2, No, From the figure, we can draw a conclusion that different bones still have difference.



3. Vegetarians and Zinc

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

Answer3: No, according the pvalue of anova with pregnant vegetarians and pregnant nonvegetarians, they don't have significant(Pvalue=0.584), so we think there no evidence can prove that pregnant vegetarians tend to have lower zinc levels than pregnant nonvegetarians.

```

      Df Sum Sq Mean Sq F value Pr(>F)
Pregnant_vegetarians 1  85.1  85.12  0.354  0.584
Residuals           4 962.9 240.72
6 observations deleted due to missingness

```

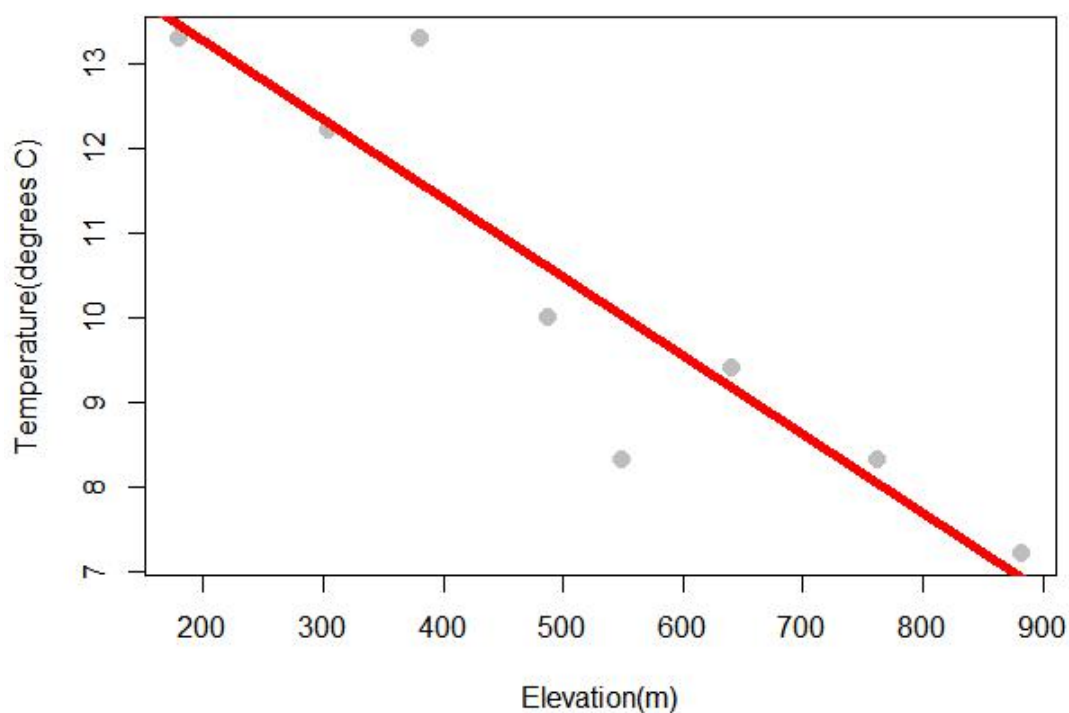
4. Atmospheric Lapse Rate

[15 points] Draw a scatter plot with regression line, and investigate if the lapse rate is 9.8 degrees C km⁻¹.

Answer4:

From the function "summary(fit)\$coefficients", we find that : The lapse rate is 9.312degrees C km⁻¹, there is a bit different with 9.8 degrees C km⁻¹

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	15.124886623	0.948282001	15.949777	3.856494e-06
Elevation	-0.009312104	0.001669811	-5.576742	1.410783e-03

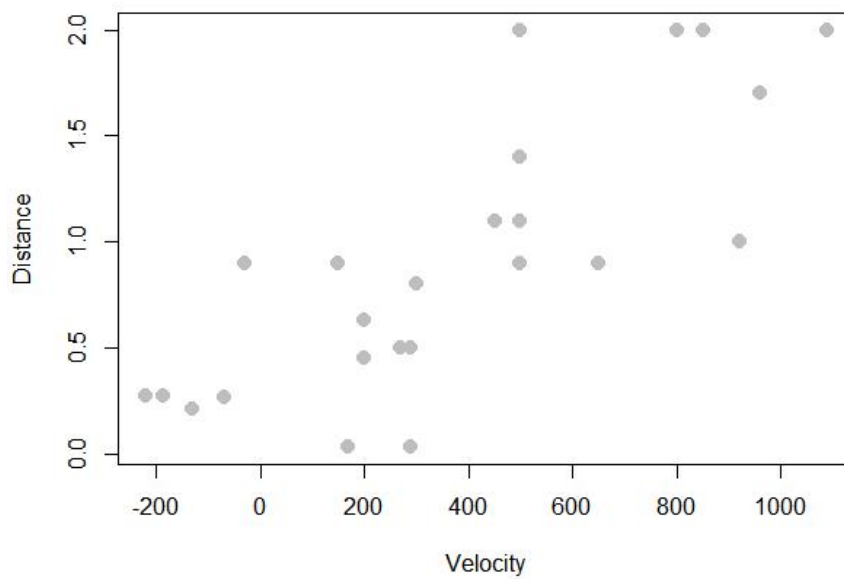


5. The Big Bang Theory

5.1 [5 points] Make a scatter plot with distance as the Y-axis and recession velocity as the X-axis. Describe what you see.

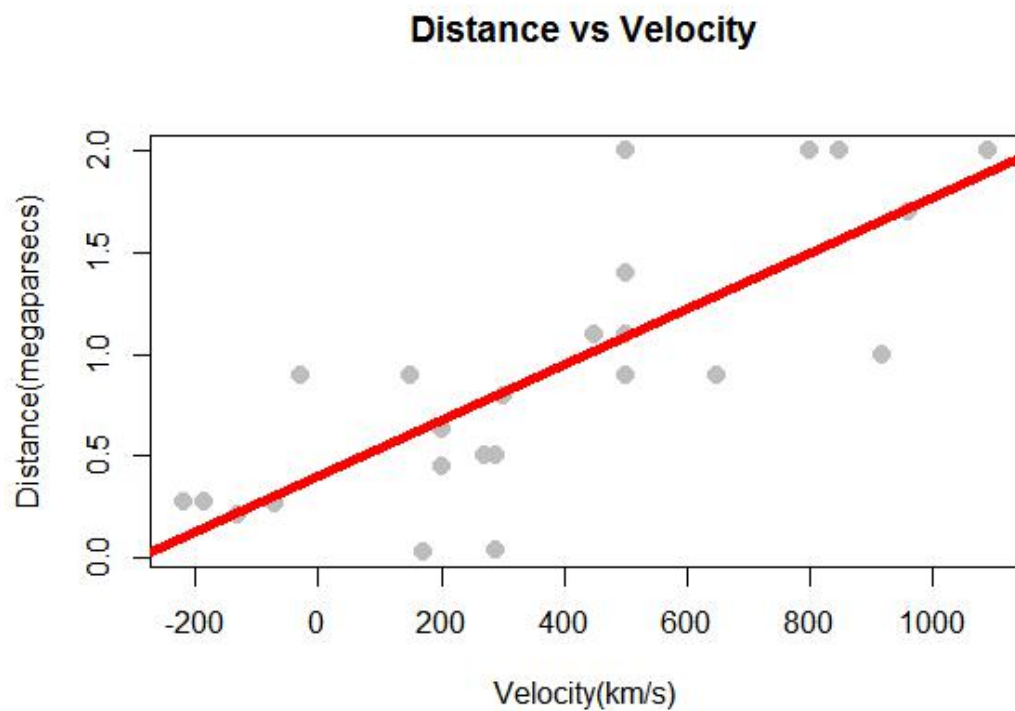
Answer5.1:

There many outliers distribute around the regression line, so I can't find the obvious tendency of distance and velocity.



5.2 [5 points] Add a simple linear regression line to the above scatter plot.

Answer5.2:



5.3 [15 points] 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.

Answer5.3:

- 1) Because universe is come from the exploded of singular point according to Hubble's Big Bang Theory, so the Distance must be zero at the beginning, which is the intercept.
- 2) According to the assumption of "And the slope is the age of the universe", so the age of universe equal to the slope, which can be calculated by: $0.001372936 * 30.9 * 10^4 * 10^8 * 10^3 / 10^6 = 6.18 * 10^{17} \text{ (S)}$, $6.18 * 10^{17} / (60 * 60 * 24 * 365) \approx 19.6 \text{ billion years}$. The age of the universe is about : 19.6 billion years

```
> 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
```

5.4 [5 points] Explain why improved measurement of distance would lead to more precise estimates of the regression coefficients.

Answer5.4:

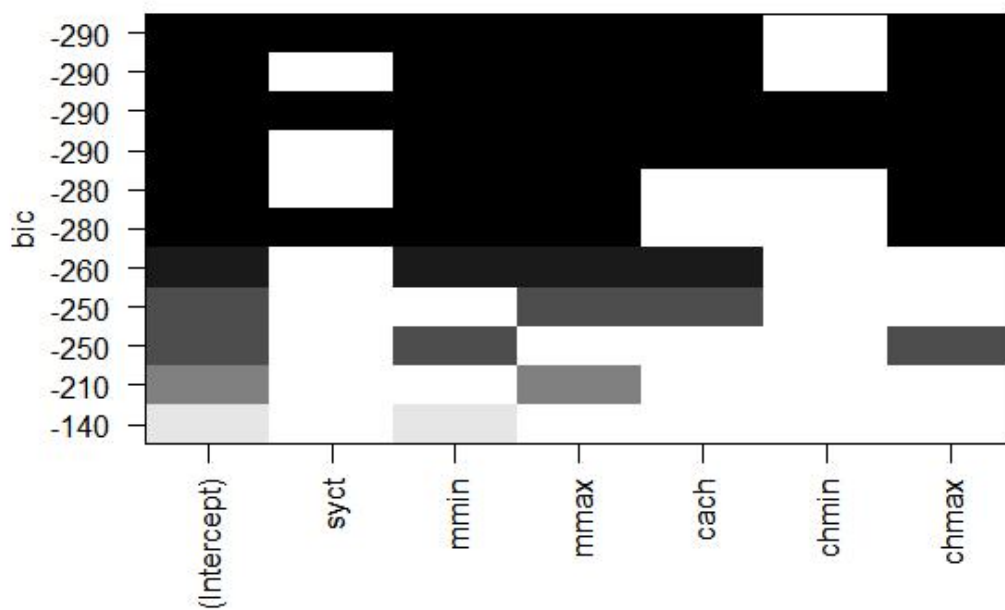
Because the improved measurement of distance are closed to the true value, so the model , which used the improved measurement of distance will fit better than before, so lead to more precise estimates of the regression coefficients.

6. CPU Performance

6.1 [5 points] For the train set, fit the best subset regression between predictor variable `perf` and response variables including `syct`, `mmin`, `mmax`, `cach`, `chmin`, and `chmax`.

Answer6.1:

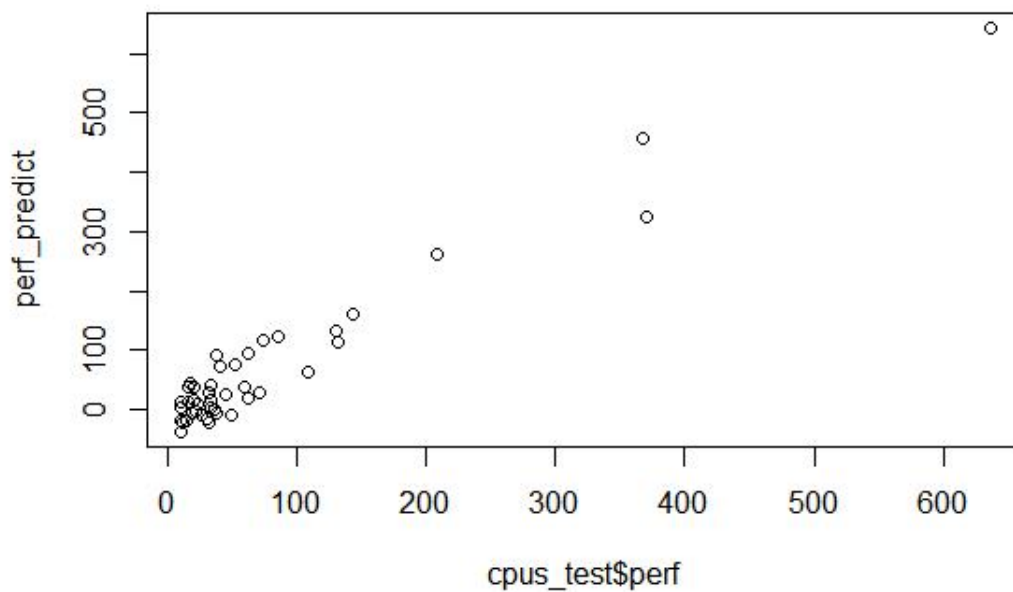
The best model is " `model_log <- lm(perf ~ syct+mmin+mmax+cach+chmax, data=cpus_train)`", which is remove "chmin" in this model.



6.2 [5 points] 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.

Answer6.2:

```
cor(cpus_test$perf, perf_predict)
[1] 0.9672987
> mean(perf_predict)
[1] 70.68673
> mean(cpus_test$perf)
[1] 77.5
> # Relative mean bias
> (mean(perf_predict) - mean(cpus_test$perf))/
+ mean(cpus_test$perf)*100
[1] -8.79132
```



7. Analysis of Data Sets from Your Group

7.1 [5 points] Define a simple research question that can be tested with the t-test. Test your question with R, and describe your findings.

Answer7.1:

Question: Can gene "Q9HBB8" be potential biomarker in detection of cancer of "PDAC"

Answer: No, it can't. Because the p-value between Normal and PDAC is more than 0.05, so there not significant of the gene.

```
data: sample1$Q9HBB8 and sample2$Q9HBB8
t = 0.20847, df = 25.621, p-value = 0.8365
```

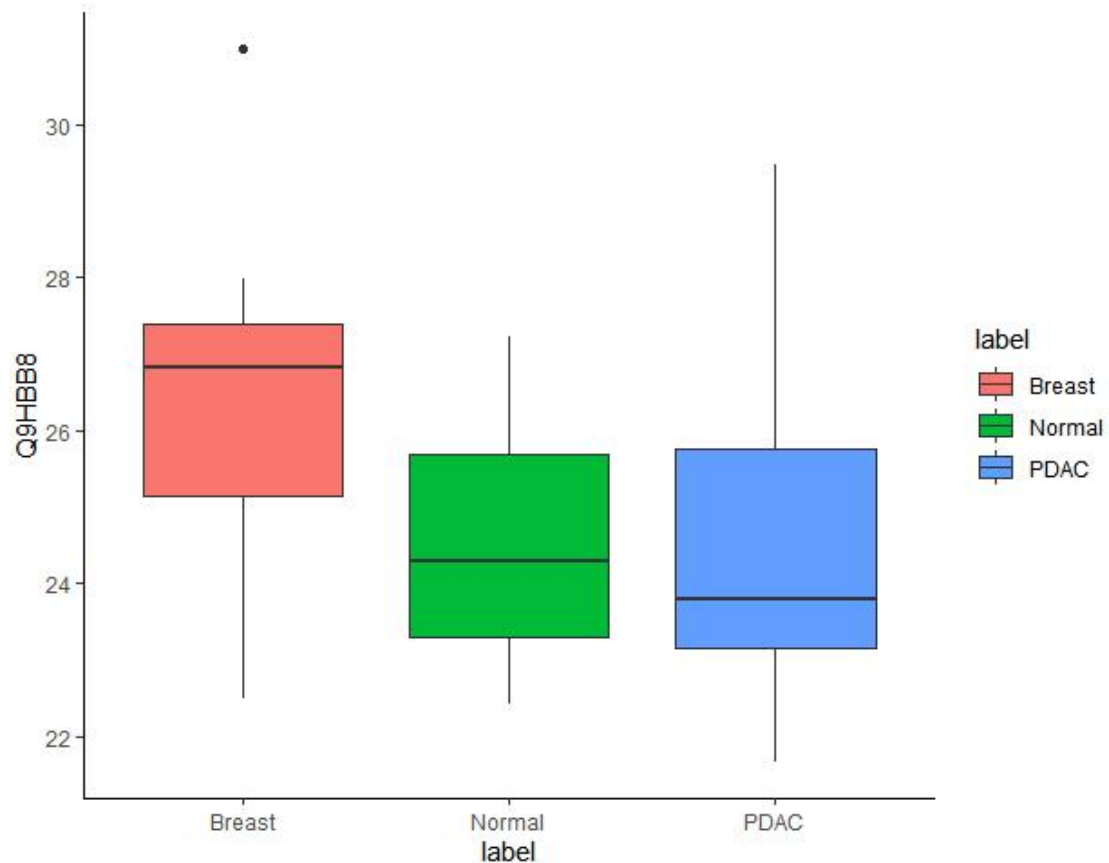
7.2 [5 points] Define a simple research question that can be tested with the ANOVA. Test your question with R, and describe your findings.

Answer7.2:

Question: Can gene " Q9HBB8" used to detect cancers of "PDAC" or "Breast cancer"?

Yes, it can. Because there significant between group of "Normal", "PDAC" or "Breast cancer"

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
label	2	36.66	18.328	4.829	0.0129 *
Residuals	42	159.39	3.795		



7.3 [5 points] Define a simple research question that can be tested with a simple linear regression model. Test your question with R, and describe your findings.

Answer7.3:

Question: Which genes can we select to detect the cancer?

Answer: We can select "O60613 ; P02746; P48664" to detect the development of cancer.

```

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -14.50985    2.47047  -5.873 9.32e-07 ***
O60613       0.19916     0.07552   2.637 0.012155 *
P02746       0.40532     0.07757   5.225 7.02e-06 ***
Q99944       0.10492     0.07168   1.464 0.151748
P10619      -0.24211     0.12322  -1.965 0.056980 .
Q99102      -0.01024     0.02863  -0.358 0.722683
P30511      -0.08487     0.07174  -1.183 0.244387
P48664       0.19971     0.04949   4.035 0.000263 ***
  
```