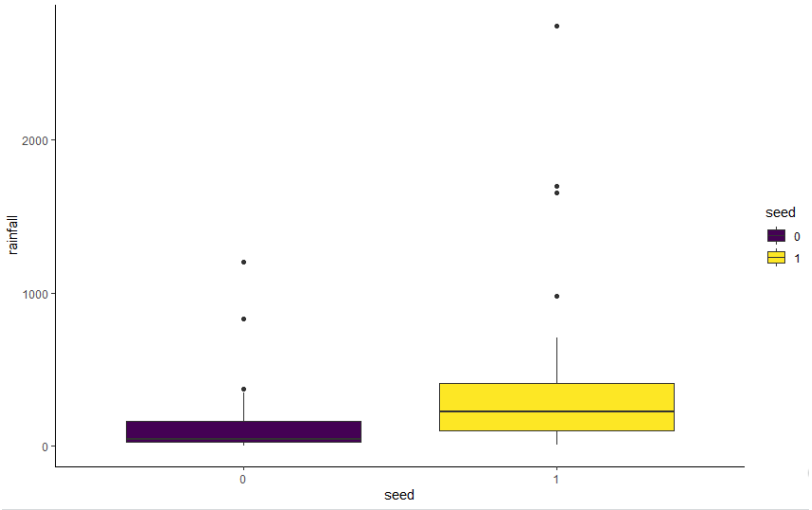
Name:Liu Yiwen

SID:12032364

The problem of this assignment mainly reviews the mathematical statistics related methods mentioned in the previous several classes, and applies them in the data analysis. From the completion process of the assignment, the mastery of some mathematical concepts still needs to be improved.

1. Cloud Seeding

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



1.2.

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

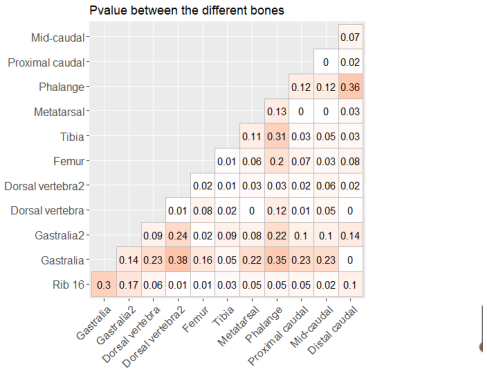
```
            Df   Sum Sq Mean Sq F value Pr(>F)
seed         1  1000360 1000360   3.993 0.0511 .
Residuals   50 12525457  250509
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> |
```

2. Was Tyrannosaurus Rex Warm-Blooded?

Answer:

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

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



3. Vegetarians and Zinc

Answer:

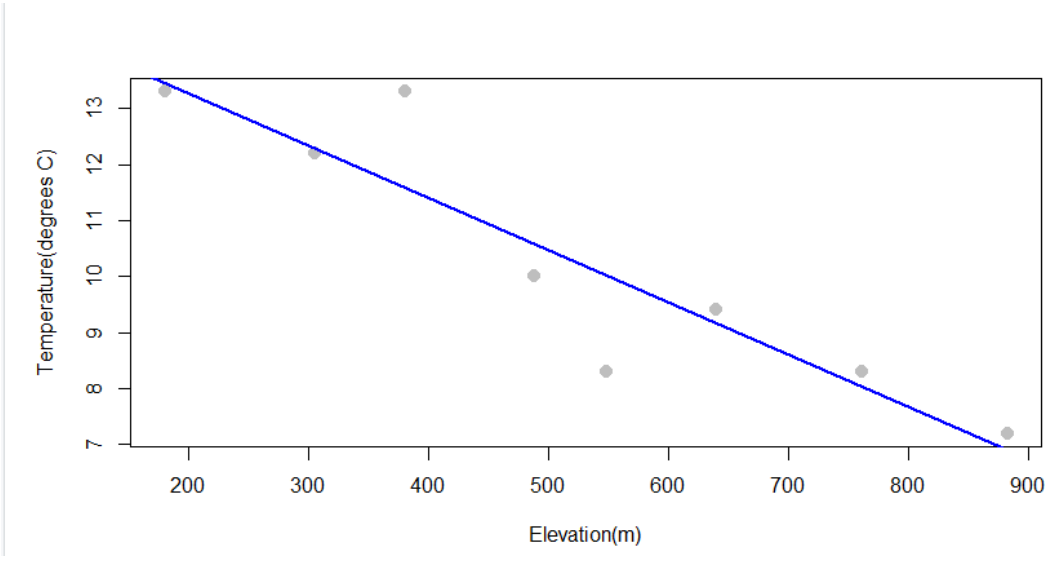
No, according the p value of anova with pregnant vegetarians and pregnant non vegetarians, they don't have significant(P value=0.584), so I think there no evidence can prove that pregnant vegetarians tend to have lower zinc levels than pregnant non-vegetarians.

```
                  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

Answer:

From the function "summary(fit)$coefficients", we found that : the lapse rate is 9.312degrees C km^{-1}, there is a little difference with 9.8 degrees C km^{-1}

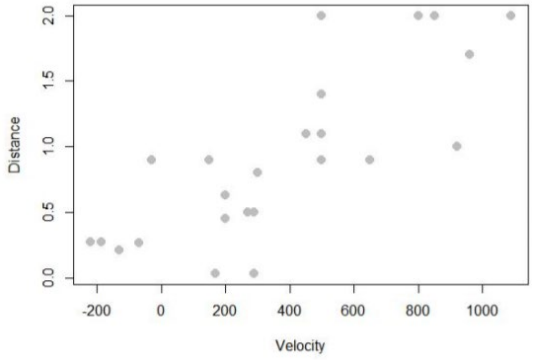


```
> summary(fit)$coefficients
                Estimate  Std. Error   t value     Pr(>|t|)
(Intercept) 15.124886623 0.948282001 15.949777 3.856494e-06
Elevation   -0.009312104 0.001669811 -5.576742 1.410783e-03
> |
```

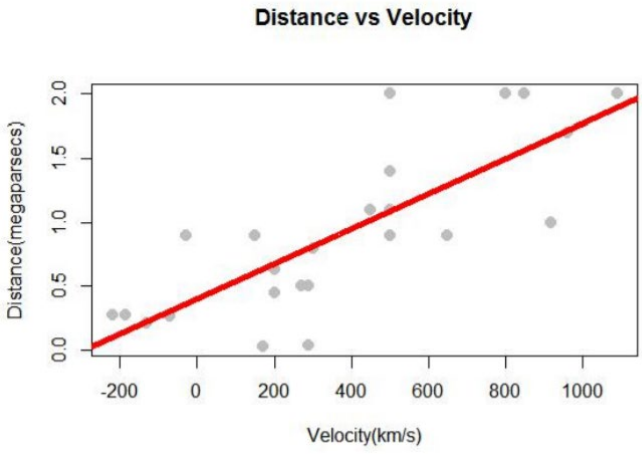
5.The Big Bang Theory

Answer:

(5.1)There many outliers distribute around the regression line,and I haven't found the obvious tendency of distance and velocity.



(5.2)



(5.3)

Answer:

(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:

30.9 * 10 6 *10 12 = 3.09 * 10 19 (S) , 3.09 * 10 19 / (60 * 60 * 24 * 365) * 0.001372936 =~ 1.35 billion years.
The age of the universe is about : 1.35 billion years

```
> summary(fit)$coefficients
               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
> |
```

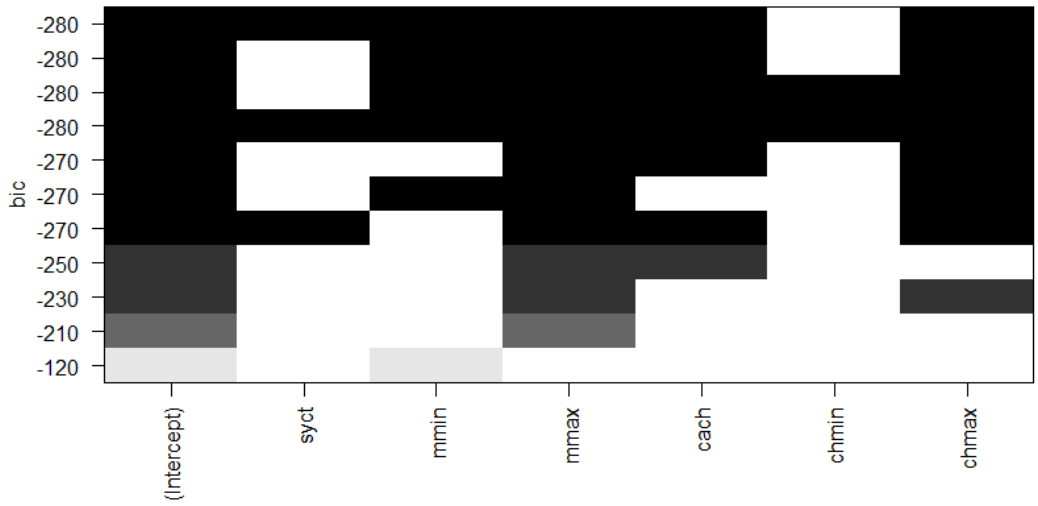
(5.4)

Answer:

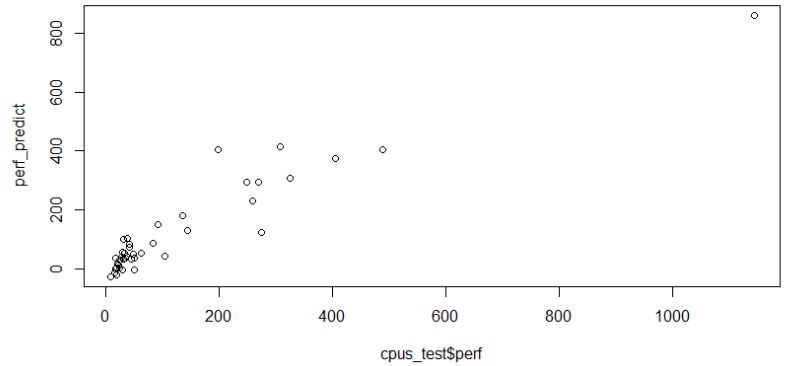
Because the improved measurement of distance are closed to the true value.

6. CPU Performance

(6.1)



(6.2)



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

Residuals:
    Min      1Q  Median      3Q     Max
-211.76  -26.24    7.89   27.80  360.90

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.264e+01  9.186e+00  -5.731 4.80e-08 ***
syct         4.664e-02  2.046e-02   2.279    0.024 *
mmin         1.034e-02  2.281e-03   4.534 1.13e-05 ***
mmax         5.821e-03  7.031e-04   8.278 4.54e-14 ***
cach         8.184e-01  1.762e-01   4.644 7.06e-06 ***
chmax        1.593e+00  2.284e-01   6.975 7.52e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 59.54 on 161 degrees of freedom
Multiple R-squared:  0.8461,    Adjusted R-squared:  0.8413
F-statistic:   177 on 5 and 161 DF,  p-value: < 2.2e-16
```

```
> cor(cpus_test$perf, perf_predict) #0.97
[1] 0.9405514
>
> # Mean predicted value
> mean(perf_predict) #70.68673
[1] 123.3808
>
> # Mean actual value
> mean(cpus_test$perf) #77.5
[1] 126.881
>
> # Relative mean bias
> (mean(perf_predict) - mean(cpus_test$perf))/
+   mean(cpus_test$perf)*100 #-8.79132
[1] -2.758575
>
```

7.

Note:Data from Li yuan,and the he inspired me about the linear regression model

```
# t-test
hist(x = sample1[,-1]$Q9HBB8 )
hist(x = sample2[,-1]$Q9HBB8 )
t.test(sample1$Q9HBB8,sample2$Q9HBB8) #0.8365


# one-way anova
ggplot(data_tibble, aes(x = label, y = Q9HBB8 , fill = label)) +
  geom_boxplot() +
  theme_classic()

anova_one_way <- aov(Q9HBB8 ~ label, data = data_tibble)
summary(anova_one_way)


# linear regression model

library(leaps)
subset_result <- regsubsets(label ~ ., data=data_tibble, nbest=2, nvmax = 6,really.big=T)
plot(subset_result, scale="bic")

data_tibble$label
label_group<-c(rep(0,15),rep(1,15),rep(2,15))
model_log <- lm(label_group ~ O60613+P02746+Q99944+P10619+Q99102+P30511+P48664, data=data_tibble )
summary(model_log) #O60613       P02746         P48664
```