**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。



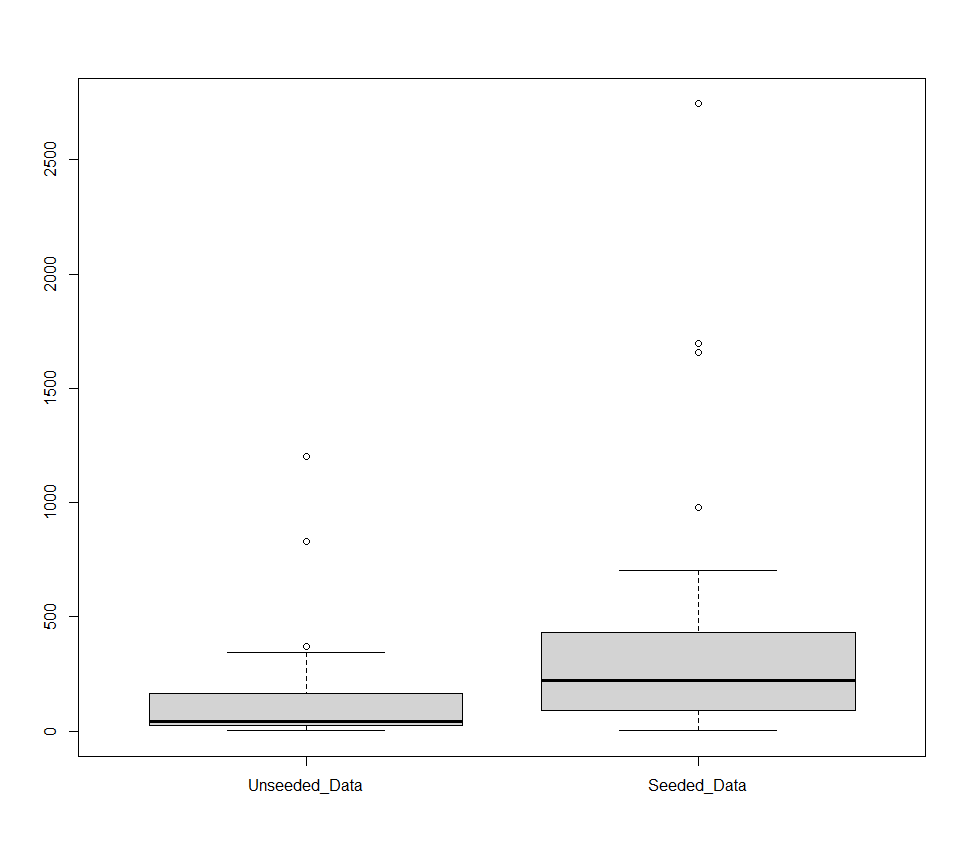


图1

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。



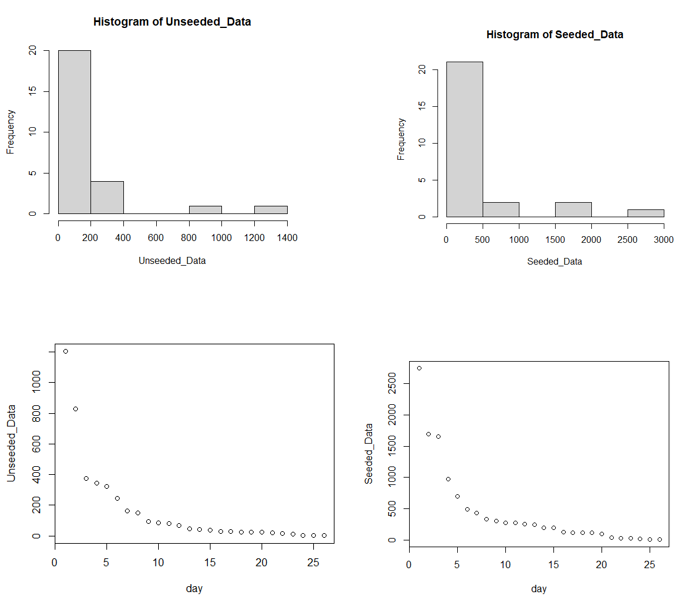


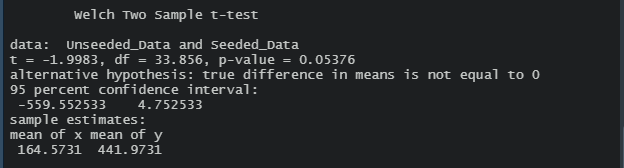
图2

**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。



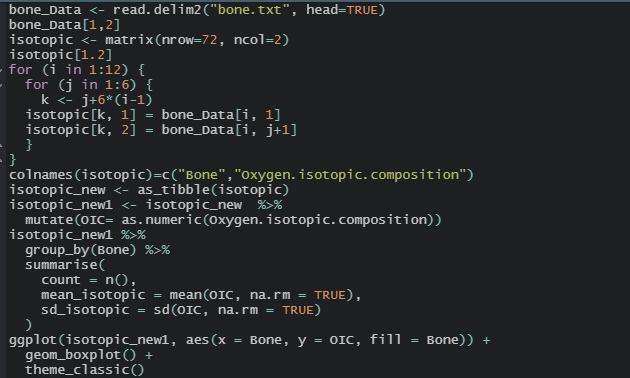


**结果与讨论：**t.test的结果显示，两组数据的平均值不一样，且催化之后的平均降雨量（mean rainfall from seeded days）更大。因此，进行催化确实会对降雨量产生影响，催化之后的平均降雨量（mean rainfall from seeded days）为491.9731，未催化之后的平均降雨量（mean rainfall from unseeded days）为164.5731，差异显著。

# 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。



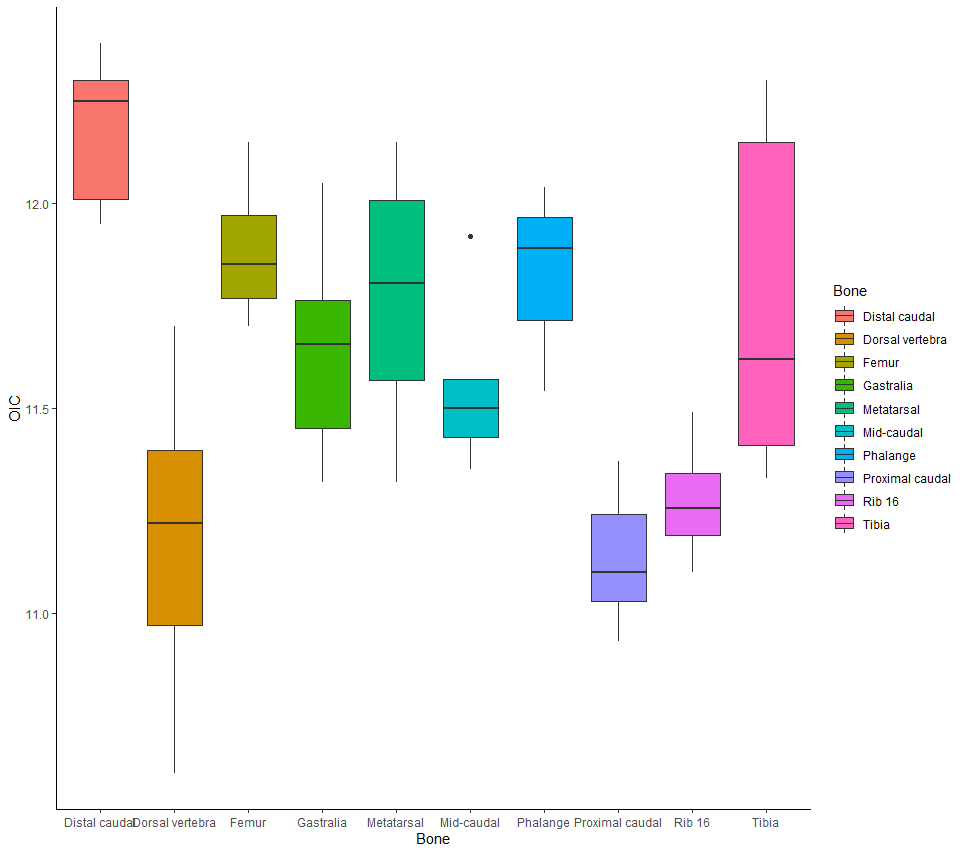
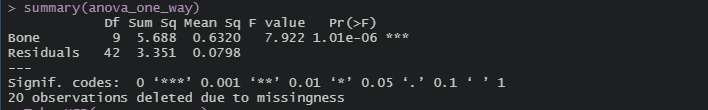
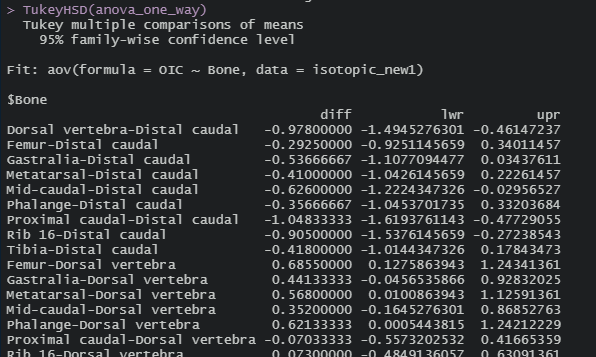


图3

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。





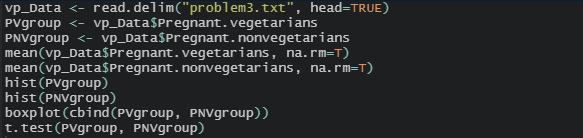


**结果与讨论：**脊椎动物骨磷酸盐的氧同位素组成与骨形成时的体温有关，同位素差异说明了温度的差异。p值小于0.5，表示不同骨骼的同位素值具有统计学差异。但是组间平均值的大小差别不是特别大，根据图3可以整体看出来，同位素平均值在11~13之间，没有显著差别，符合温血动物（warm-blooded anaimals）的条件，即温血动物骨骼温度差别不大。因此，霸王龙（Tyrannosaurus rex skeleton）应该可以认为是温血动物。

# 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。



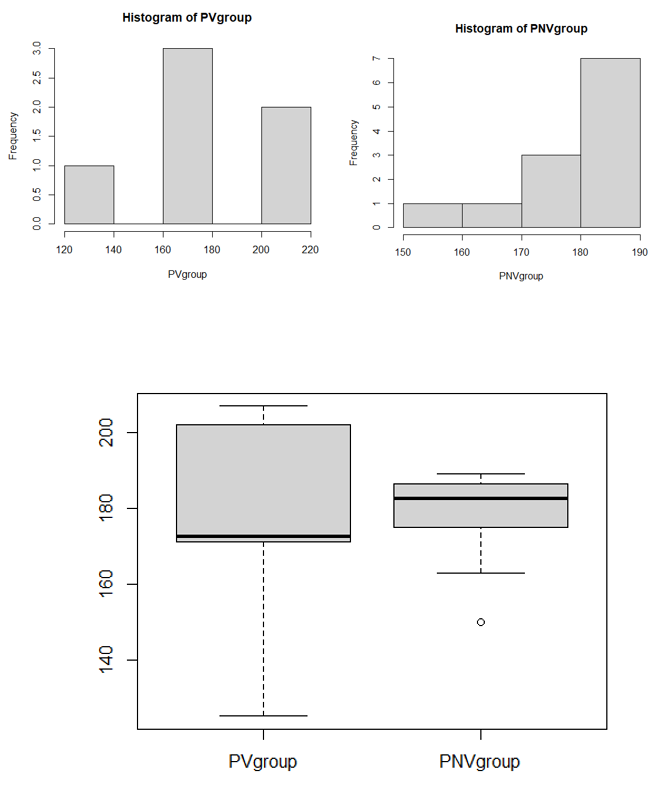
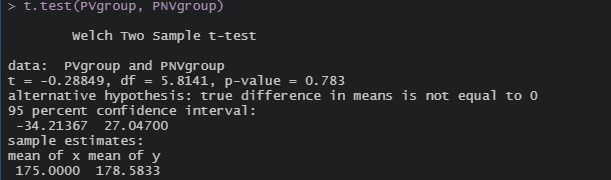


图4

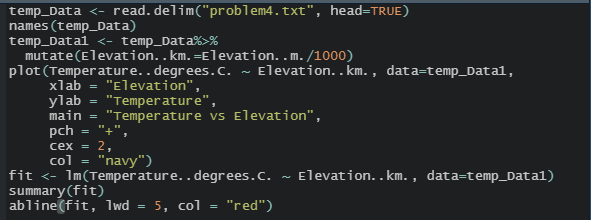


**结果与讨论：**从t.test结果看出来，素食怀孕者（pregnant vegetarians）的锌（zinc）含量均值为175，非素食怀孕者（pregnant nonvegetarians）的锌含量均值为178.5833，素食怀孕者（pregnant vegetarians）的锌（zinc）含量均值偏低，但两者的差值不大。结合图4可知，首先直方图的结果显示，PVgroup中出现锌含量>200的值出现，但是出现次数很少，锌含量>200值的出现导致了素食怀孕者（pregnant vegetarians）的锌（zinc）含量均值偏高；同时从箱图也可以看出，尽管PVgroup的最大值高于PNVgroup，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。



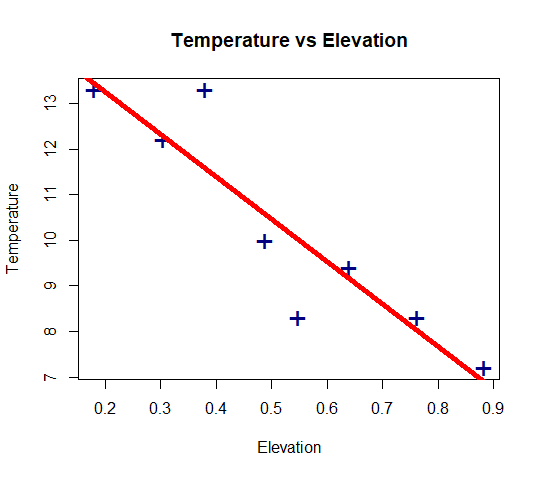
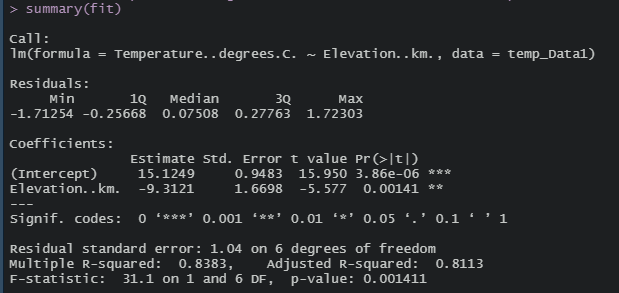


图5

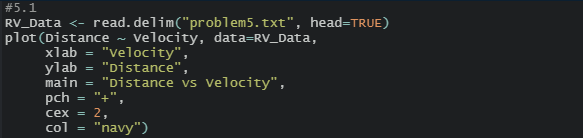


**结果与讨论：**根据线性拟合结果可知，温度的下降速率为9.3121degrees C km-1。

# 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。



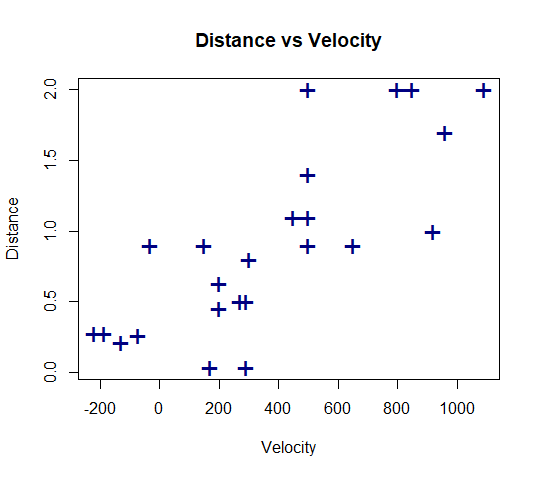


图6

**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。



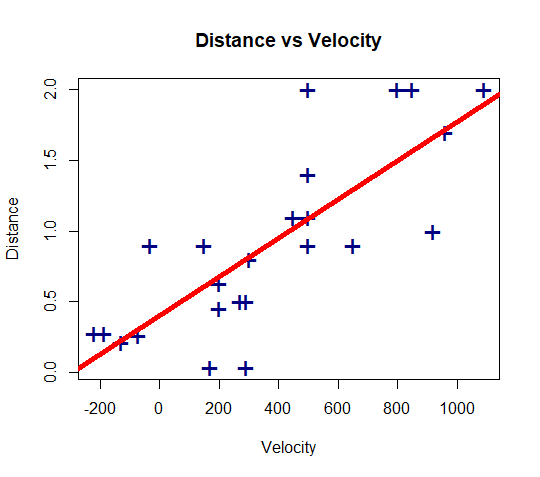
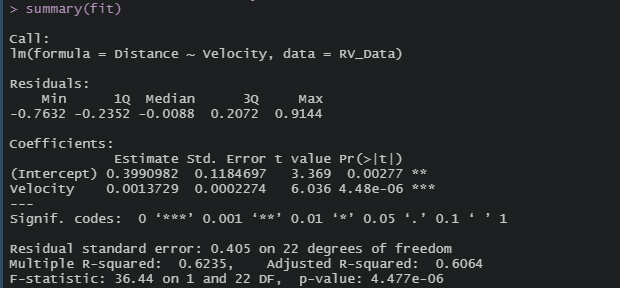


图7



**结果与讨论：**线性拟合结果的R2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。





**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\*1019km，速率的单位是km/s，age=3.09\*1019\*0.001372936/（60\*60\*24\*365）=1.35\*109 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.***

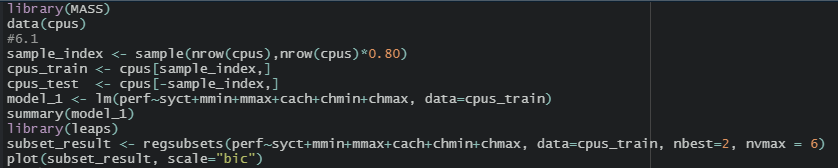
**结果与讨论：**提高距离的测试精度，使得距离的结果更加接近真实值，那么线性拟合的结果就更加接近于真实值，所以，回归参数的结果就更加准确。

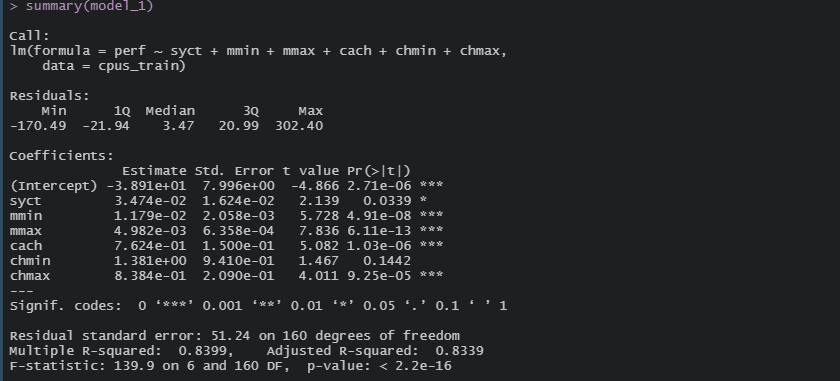
注：5.3和5，4请教李元同学解释了一下题意，帮助解题。

# 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。





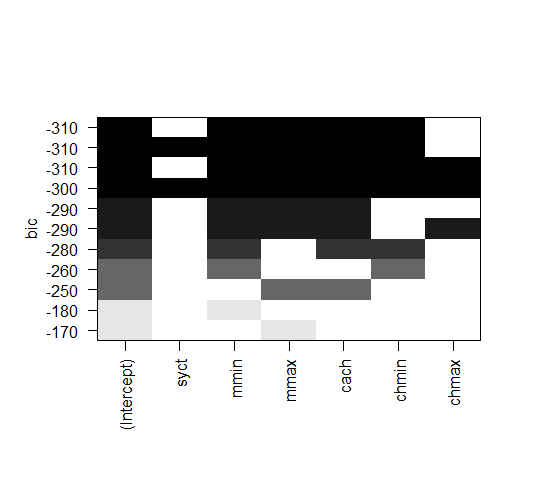
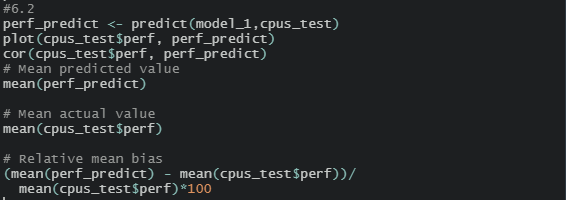


图8

**结果与讨论：**回归结果的R2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。



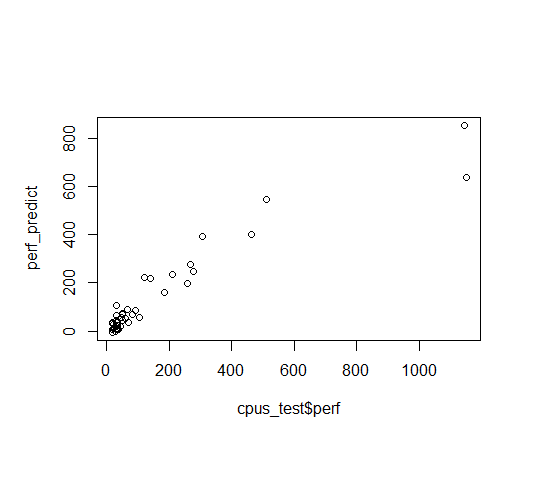
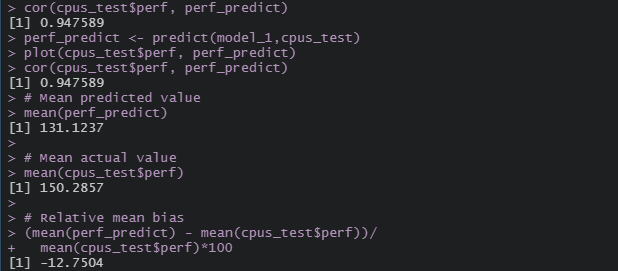


图9



**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 NO3- concentration on denitrification rate and NO3- removal efficiency in the low-permeability media, a set of column experiments with different flow rates and injected NO3- concentrations were conducted.

**Table1** Summary of five column experiments, in terms of flow rates, influent NO3- concentrations and durations

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Experiment No. | Flow rate (m/d) | Influent NO3- concentration (mmol/L) | NO3- 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 |

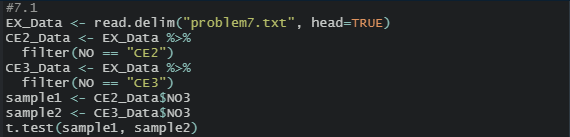
**Table2** The calibration results of the maximum rate constants for denitrification ()

|  |  |  |
| --- | --- | --- |
| Experiment No. | Initial ( mol/L/s) | ( mol/L/s ) |
| CRslow | 8🞨10-10 | 1.8🞨10-8 |
| CRmid/CChigh | 1.9🞨10-8 |
| CRfast | 2.0🞨10-8 |
| CCmid | 1.29🞨10-8 |
| CClow | 4.5🞨10-9 |

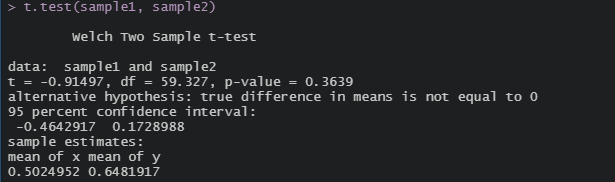
***7.1 Define a simple research question that can be tested with the t-test. Test your question with R, and describe your findings.***

|  |  |
| --- | --- |
| 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.

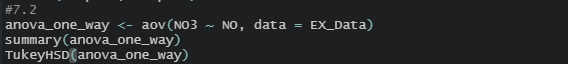


**A:** With the increase of velocity, the mean of nitrate concentrations increased.

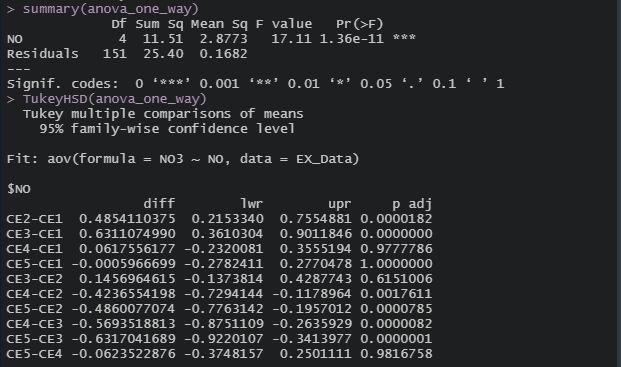


***7.2 Define a simple research question that can be tested with the ANOVA. Test your question with R, and describe your findings.***

**Q:** Compare the nitrate concentrations in different operational conditions.



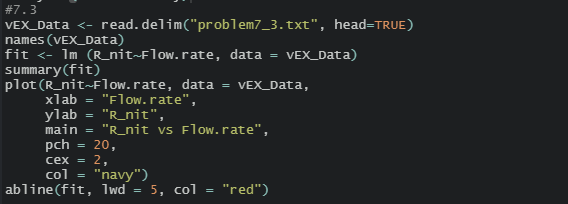
**A:** The p-value is lower than the usual threshold of 0.05. there is a statistical difference between the groups.



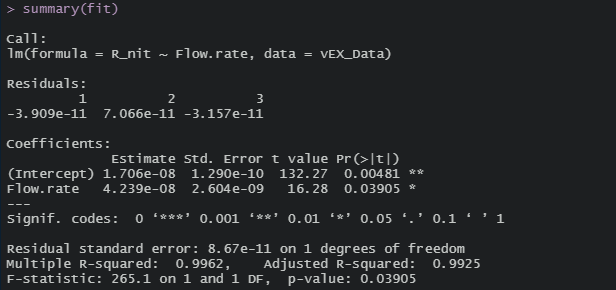
***7.3 Define a simple research question that can be tested with a simple linear regression model. Test your question with R, and describe your findings.***

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

|  |  |  |
| --- | --- | --- |
| Experiment No. | Flow rate (m/d) | ( 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 |



**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.



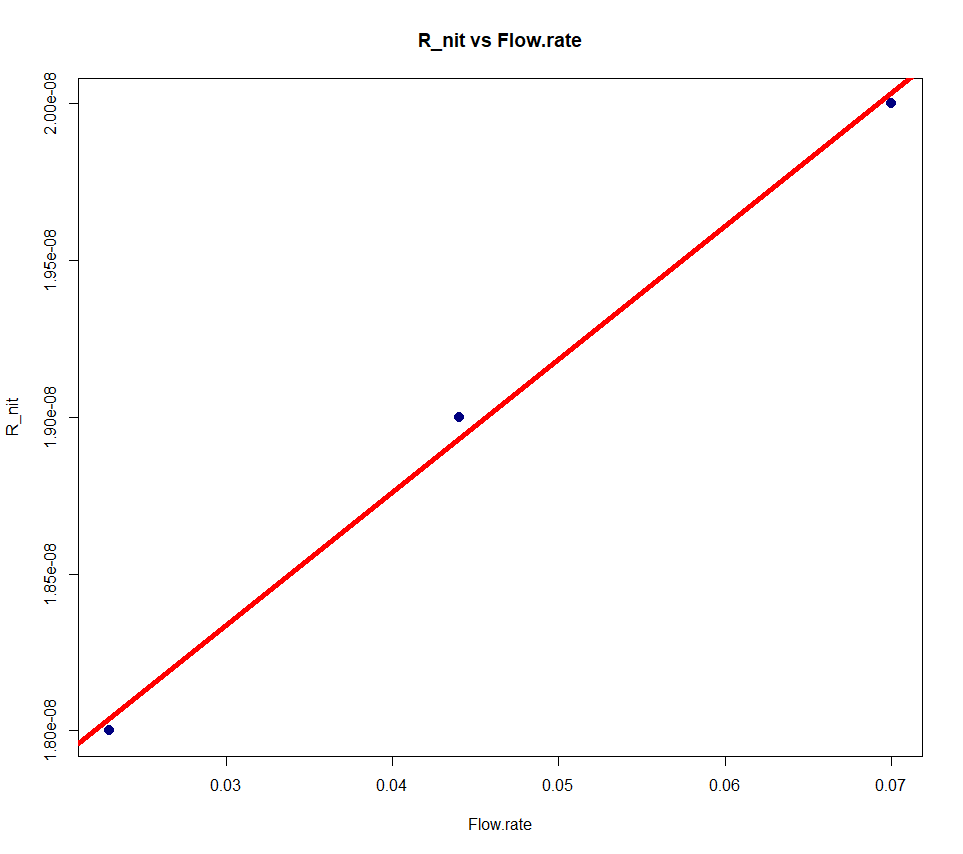


图10