PS3 Report

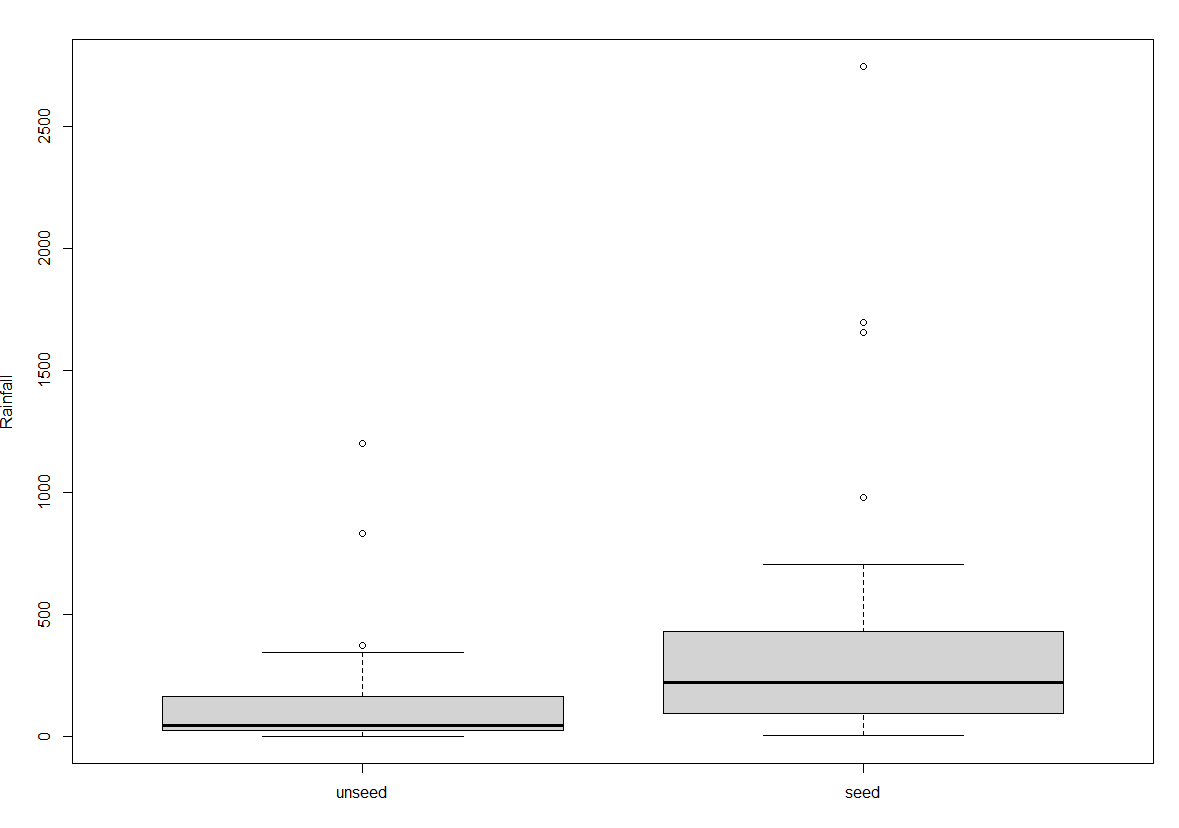
吴文浩 12032567

1. PS3\_1

**1.1：**读入unseed和seed时候降雨量的两组数据，并将其并排画为箱型图

**结果解读：**从图上可以看出，Unseeded Days中降水量整体偏小，降水量离散程度小；Seeded Days中降水量整体有所升高，四分位间距较开，数据离散程度大

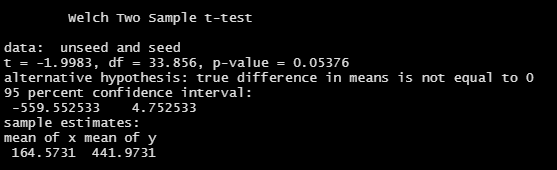
**结果截图：**



**1.2：**由于需要判断seed对于降水量是否有影响，且每个独立样本数目小于30，这里采用t检验的方式对两组数据进行均值检验

**结果解读：**从结果上看，p-value > 0.05，表明两者均值差异不显著，即在该实验中无法确切的证明seeding对于降水量有影响

**结果截图：**

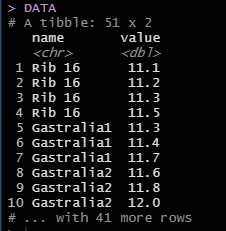


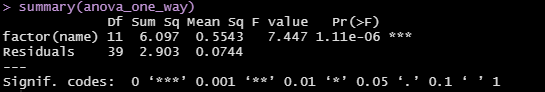
1. PS3\_2

根据题目要求，可以发现需要对多组数据进行均值检验，故采用方差分析的方法解决，在代码开始前，先将网页数据制作成了csv数据集便于读取，而后使用aov函数进行方差分析。

**结果分析：**可以看出p-value < 0.05，有信心说恐龙的各个骨骼样品中存在均值差异，表明霸王龙并不是一种温血动物。

**结果截图：**





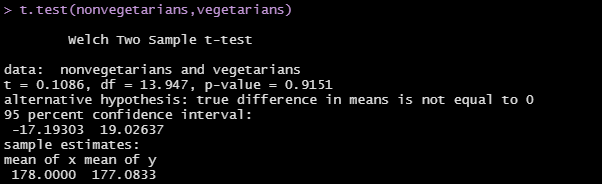
1. PS3\_3

导入表格，由于本题数据中同时有未怀孕素食主义者，怀孕素食主义者，由于暂无确切证据表明怀孕会大量降低女性Zn含量，故此处未怀孕素食主义者算干扰项，则只需分别将怀孕非素食主义者，怀孕素食主义者的数据提取出来，t检验判断即可。

**结果分析：**从结果上来看，p-value>0.05，可以认为怀孕非素食主义者和怀孕素食主义者的女性Zn含量无明显差异，即无确切证据指出素食主义者比非素食主义者的锌含量更低

（额外补充：在计算中还加入了对于素食主义者怀孕及非怀孕期间的检验，发现两者也无很大差别，表明无明显证据指出怀孕会使得女性zn含量大量下降）

**结果截图：**

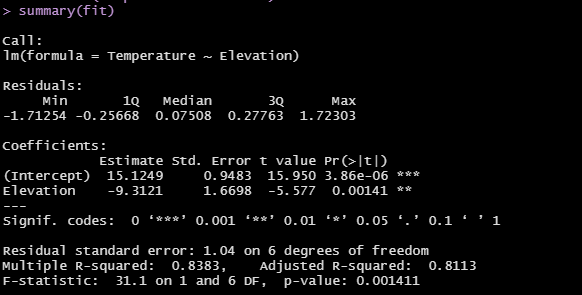


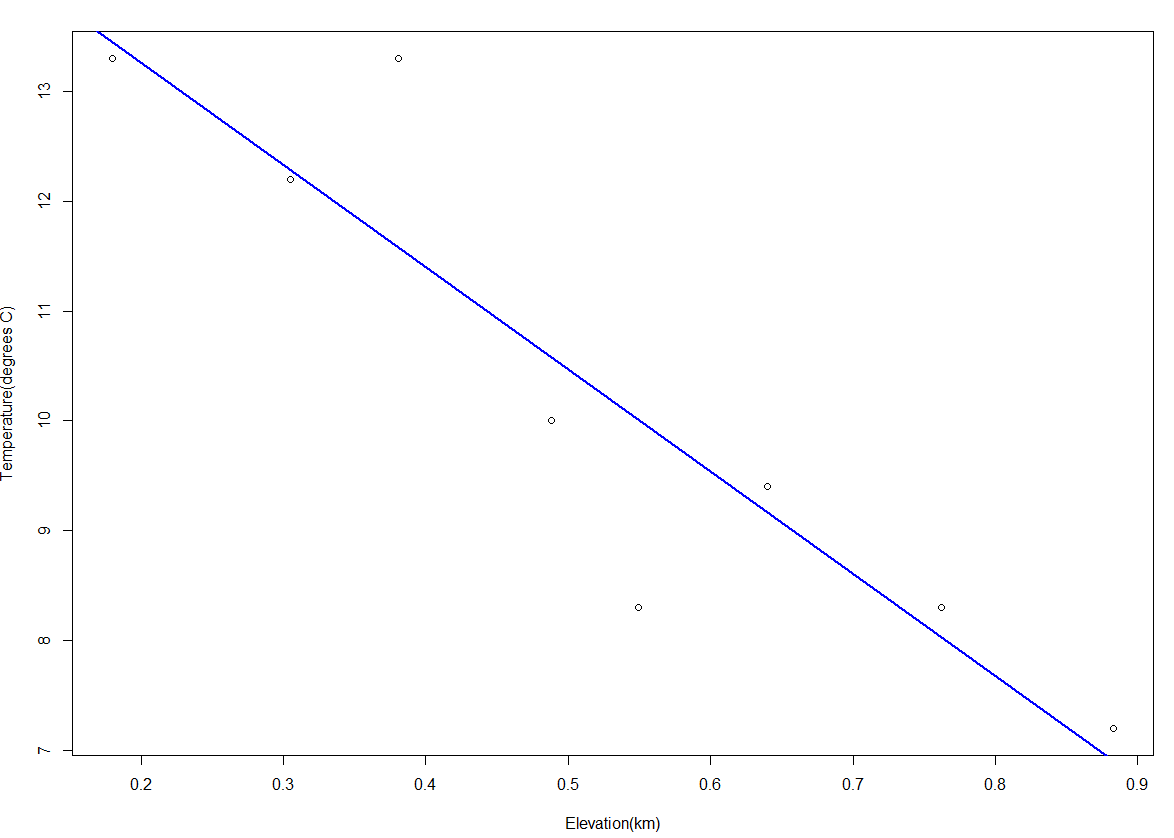
1. PS3\_4

导入数据，同时将高度单位转换为km，绘制Temperature ~ Elevation散点图，并使用lm得到其回归曲线，使用summary函数获得其详细信息（其中斜率即为 lapse rate），最后使用abline绘制回归线

**结果分析：**通过线性回归结果，可以看出斜率为-9.3121，即 lapse rate为9.3121 degrees C km-1

**结果截图：**



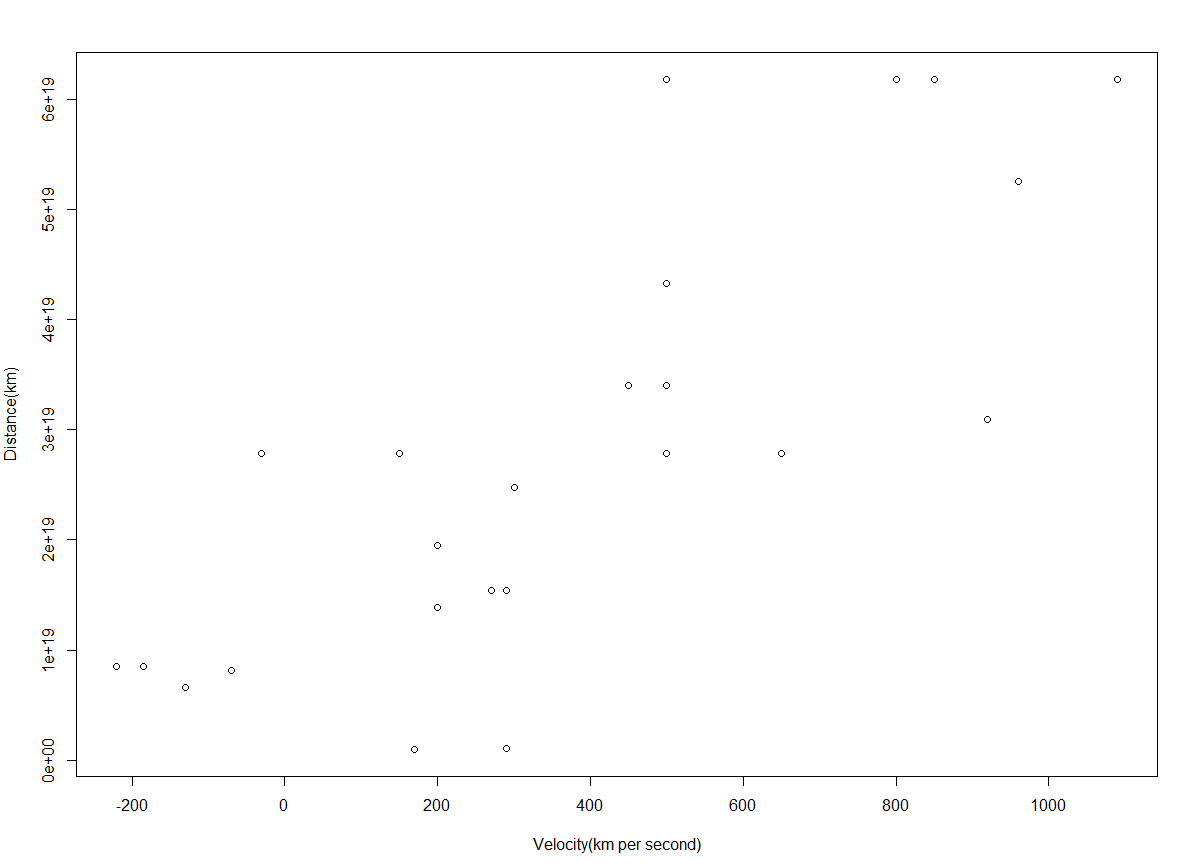


1. PS3\_5

**5.1：**将表格导入，提取出向量，并将Distance中的单位换算成km，绘制散点图

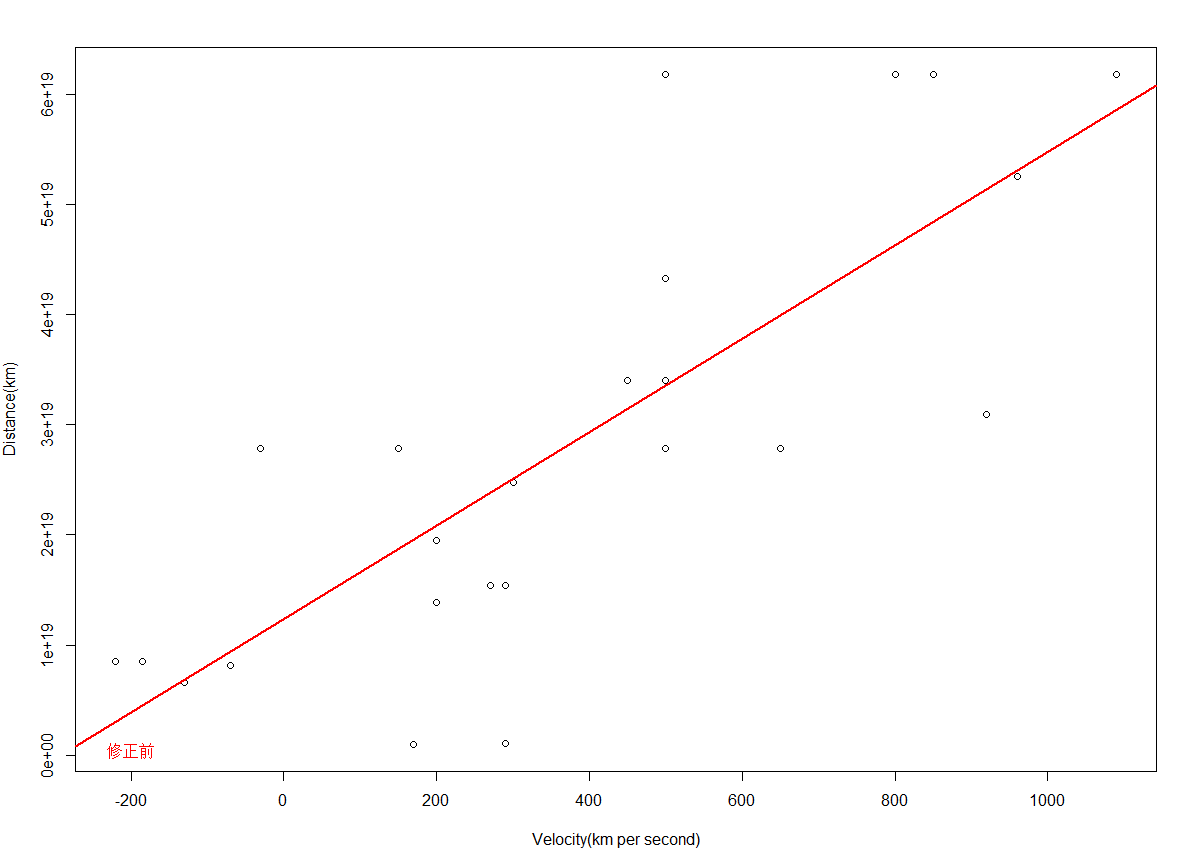
**结果分析：**从图中可以看出，recession velocity越快的星云，距离地球的距离越远，这可能暗示了宇宙目前是在一个加速膨胀的过程，这才导致大爆炸后距离远的星云加速远离

**结果截图：**



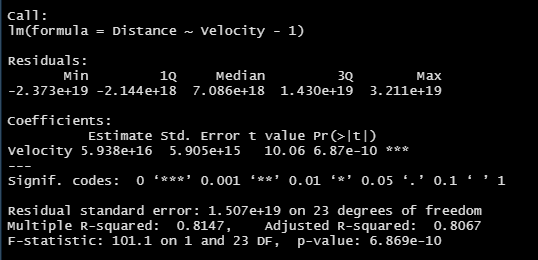
**5.2：**直接使用lm函数进行简单的线性回归拟合，此时的线性回归函数带有截距项，并绘制该函数曲线

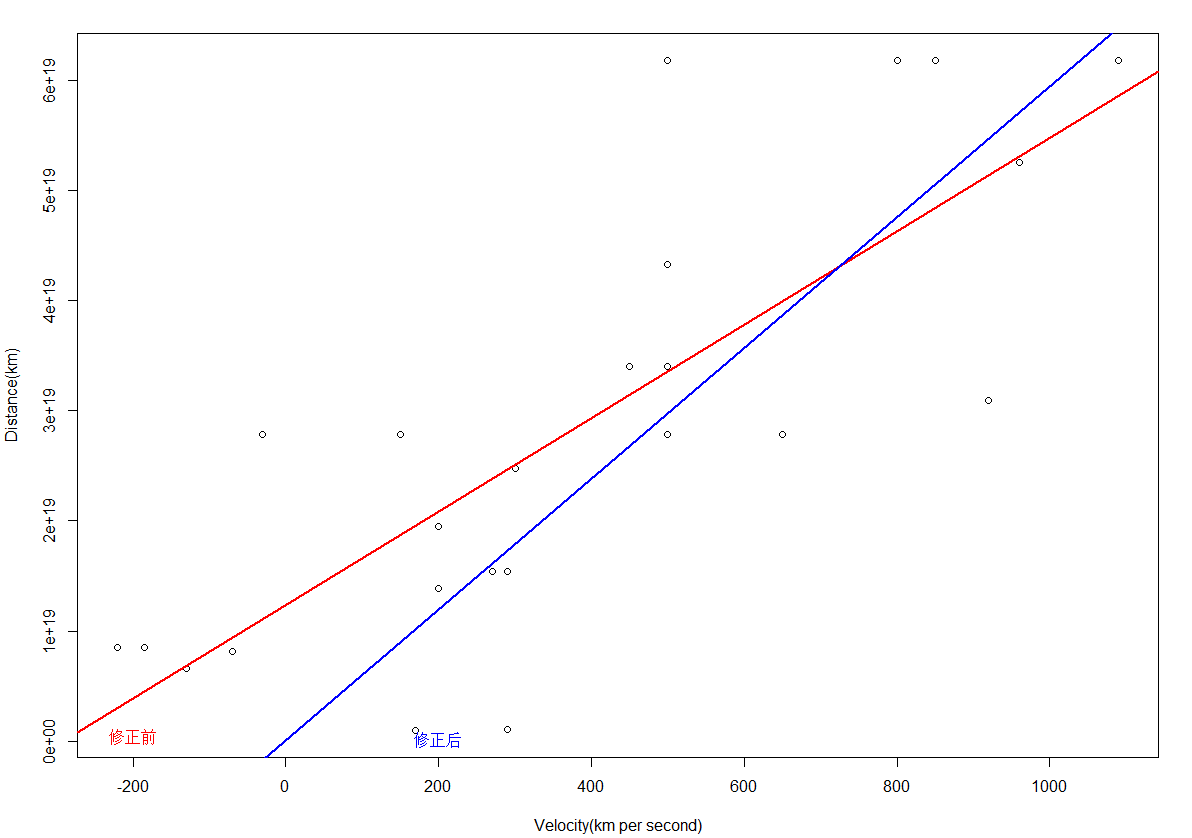
**结果截图：**



**5.3：**因为按照大爆炸理论，所有的星云都是从大爆炸的起点开始飞速的向外逃逸，即不存在初始距离，所以需要去掉截距项，因此对lm函数进行修改，关系式改为Distance~Velocity-1从而使截距为0 。线性回归后绘制出该曲线，并得到斜率为5.938e+16（s），即估计宇宙的年龄为1882016779年

**结果截图：**



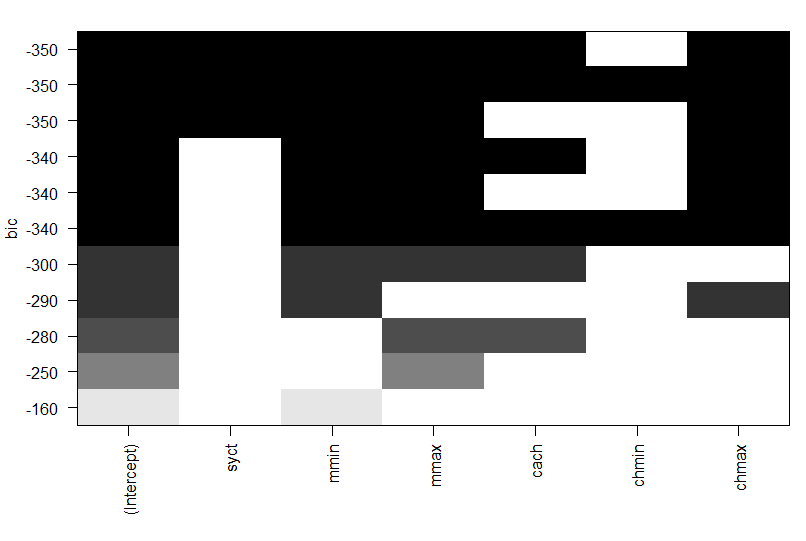


**5.4：**因从5.3的结果中可以看出，这条回归曲线的p值很小，p-value<0.05，即这条回归曲线的可信度很低，观察散点图也可以发现点的离散程度很高，可以断定在修正距离测算的方法后可以提升回归系数的估计的精度。

1. PS3\_6

**6.1：**读入cpus数据，并使用sample函数将其划分为80%的训练集和剩下20%的测试集，使用regsubsets函数进行全子集回归，并按照bic的方法绘制出图表，可以看出在选择syct+mmin+mmax+cach+chmax作为影响因子时，bic结果最小，将他们作为影响因子用于回归

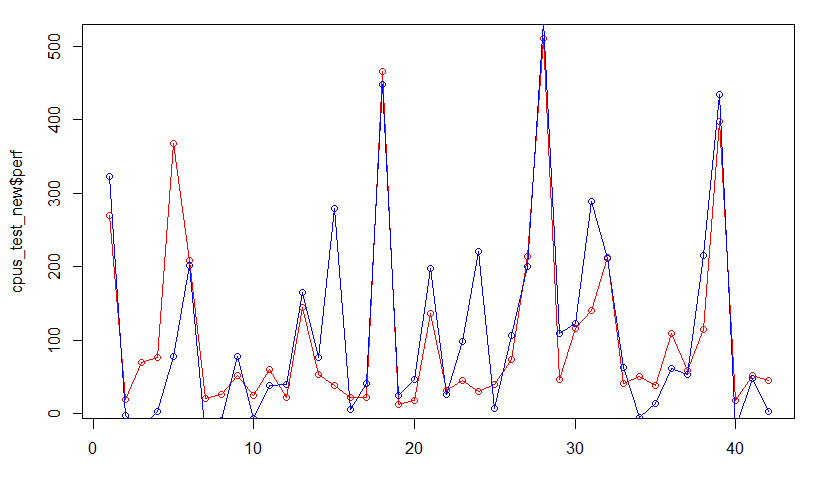
**结果截图：**

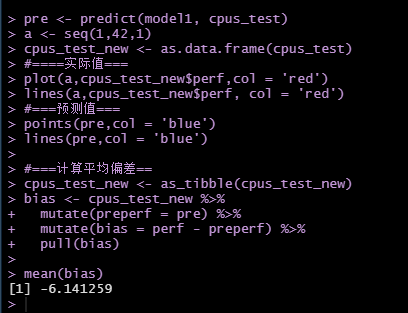


**6.2：**将上述因子用于线性回归得到模型model，用该模型对测试集进行预测，并将其绘制成图表，图上蓝色点为预测值，红色点为实际值，为了方便对比，这里将点连接起来。最后计算预测值与实际值之间的差异平均值

**结果解读：**从结果上看，多元线性回归得到的模型依然存在问题，尤其在预测时甚至出现了负值，这是不应该的，需要考虑进一步优化回归模型。

**结果截图：**





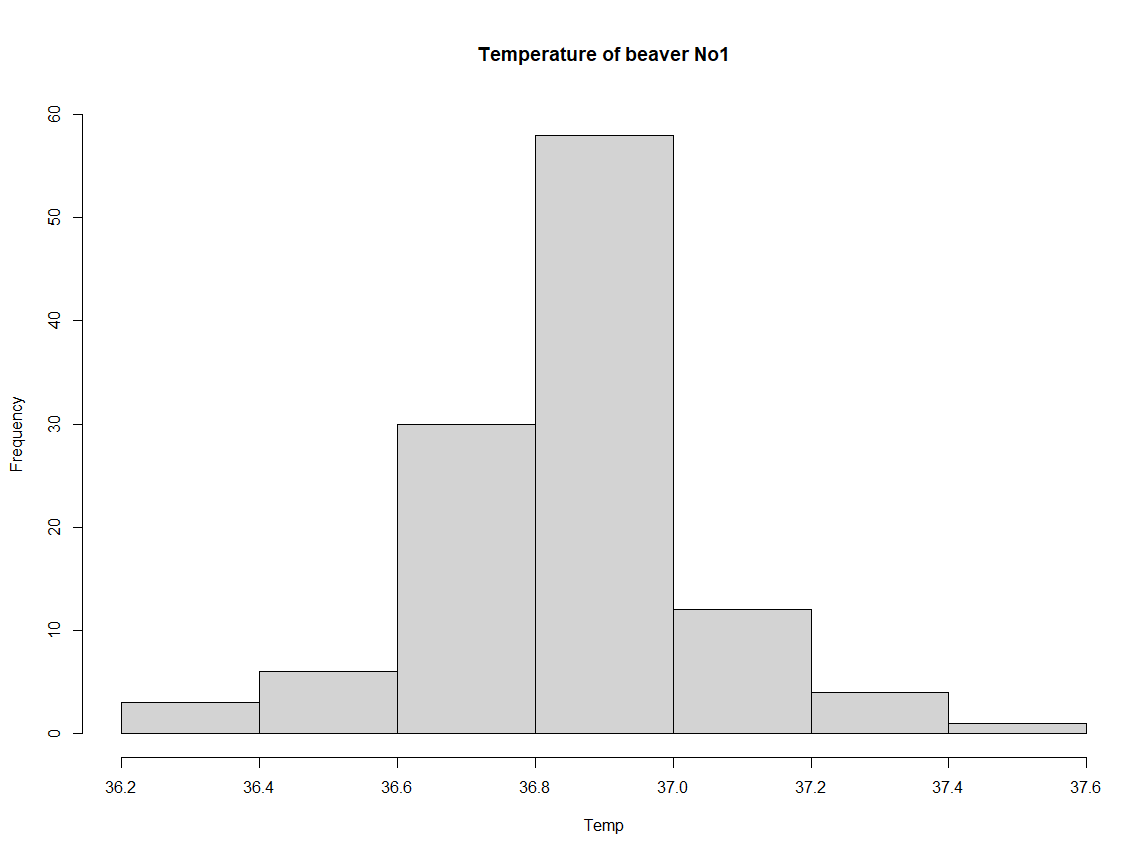
1. PS3\_7

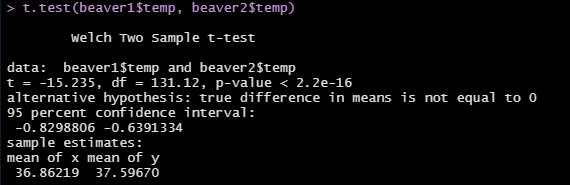
本题由于研究方向为无人机设计，没有相关观测性的数据，故选用R语言自带数据集进行代码书写

**7.1：**beaver1 & beaver2是两只海狸每10分钟的体温数据，现需要探究两只海狸的体温数据是否相同，先检查海狸体温数据是否符合正态分布，从图中可以看出基本满足正态分布条件，故可使用双样本T检验对两只海狸体温数据进行检验。

**结果分析：**从T检验结果可以看出，p-value <0.05，可以确认两只海狸的体温存在差异。

**结果截图：**

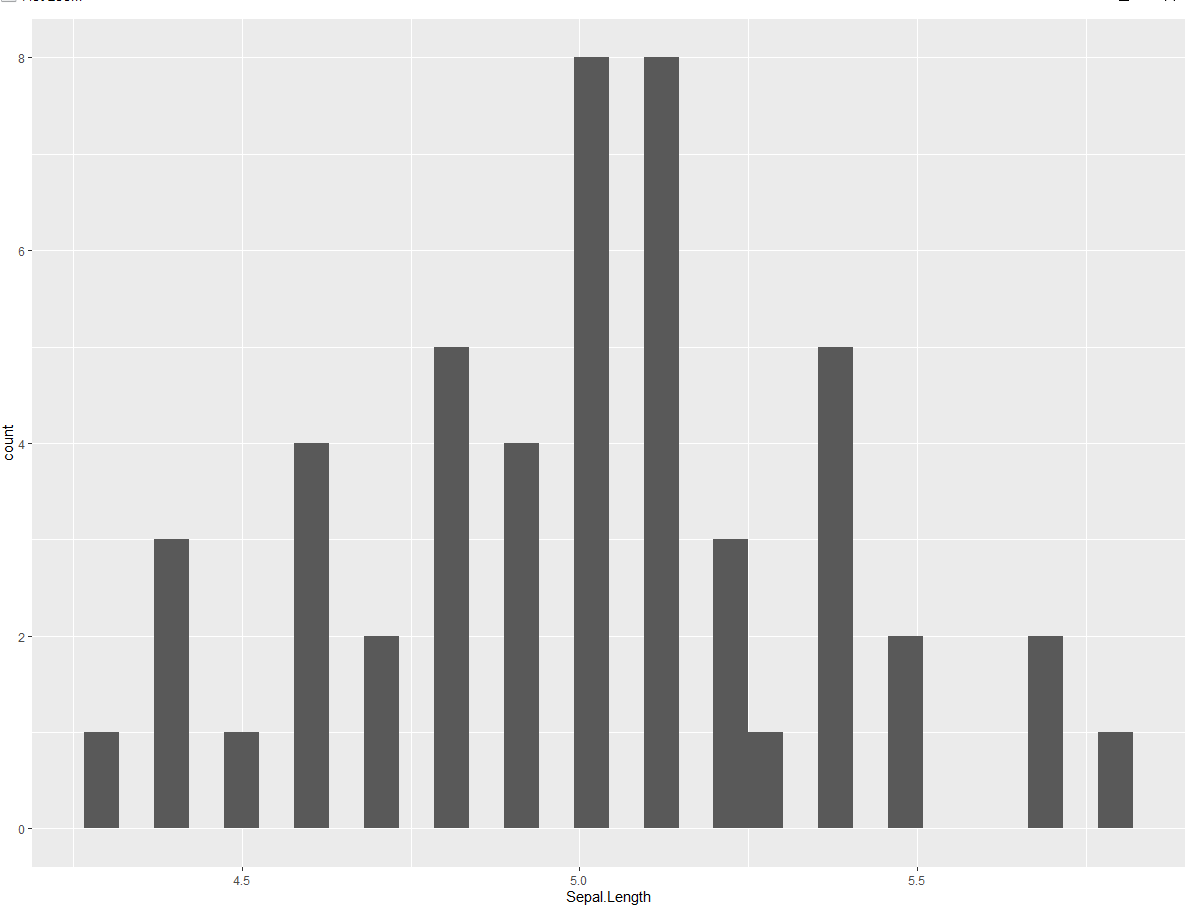


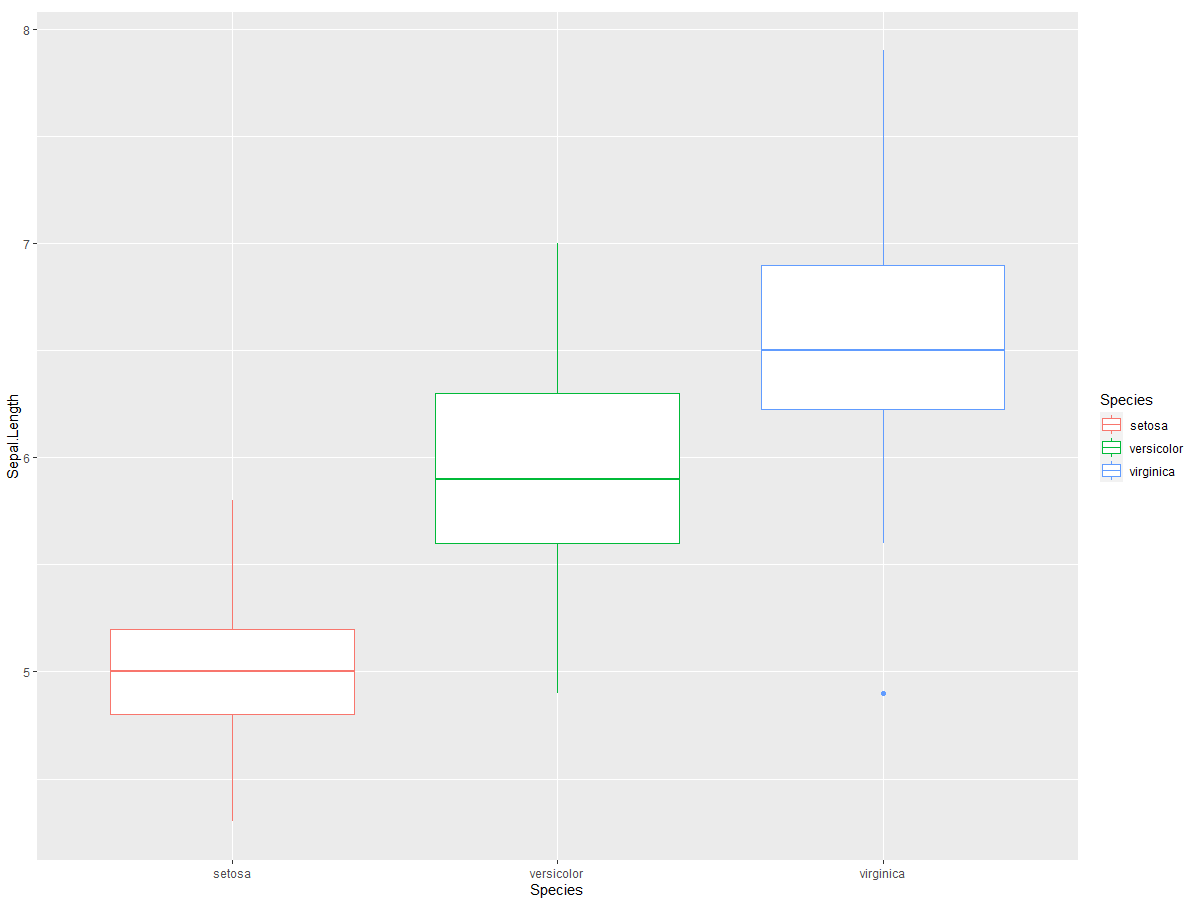


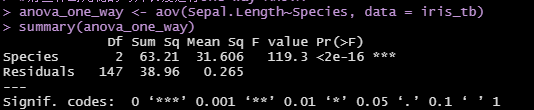
**7.2：**iris是3种鸢尾花形态数据，目前需要检测三种鸢尾花，其萼片长度是否一致。首先对萼片长度的数据绘制条形图，确认其符合正态分布，然后绘制三个种类的箱型图。之后采用单边anova进行方差检验

**结果分析：**从方差检验的结果来看，p-value <0.05，确认不同品种的鸢尾花，其萼片长度存在差异

**结果截图：**







**7.3：**swiss是瑞士的生育率以及其社会经济指标，目前需要探究生育率与社会经济指标之间的相关性，并能够用社会经济指标去估算各个地区的生育率。由于是一个多元回归问题，首先要使用最佳子集回归，bic作为模型来确认最佳的线性自变量组合，从bic模型中可以看出Agriculture + Education + Catholic + Infant.Mortality的组合就可以进行线性回归。故使用这个组合作为线性回归的自变量，由于生育率Fertility，并不完全和经济指标相关，包含截距项，此处保留截距进行线性回归。由于数据量小，直接使用回归模型作用于原数据集获得预测值，通过图表对比预测情况，其中蓝色为预测点，红色为实际值，通过连线便于两者对比，并最后计算平均偏差.

**结果分析：**从对比散点图情况可以看出，整体预测效果比较接近，所计算的平均偏差也比较小，拟合较好。但这不够完善，因为数据集过小，结果还是不具有普适性。

**结果截图：**

