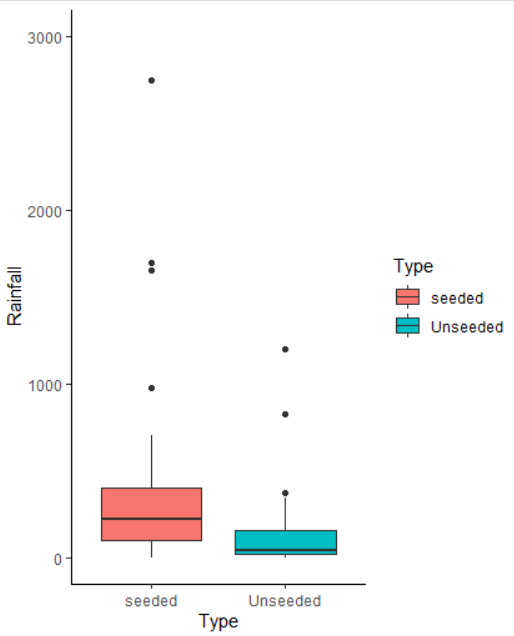
**Problem1**

#1.1

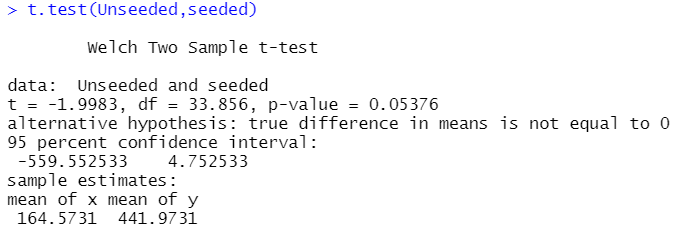
首先设置工作环境，载入程辑包，因为数据量不大，本题手动输入数据，并将播种和未播种的数据按行合并，以tibble储存在变量中。

分别作播种和未播种对应降雨量的箱线图，经过比较箱线图得出，对积云进行播种后，降雨量明显大于未播种时的降雨量。且播种后降雨量的波动范围明显大于未播种时的波动范围。值得注意的是，即使播种后也有降雨量为零的时候，说明播种并不能百分之百引起降雨。



#1.2

通过对两个样本进行T-test，观察p-value=0.05376得出，样本中变量关联有5.4%的可能是由于偶然性造成的，可以说播种对于降雨量有相关性，但是不显著。

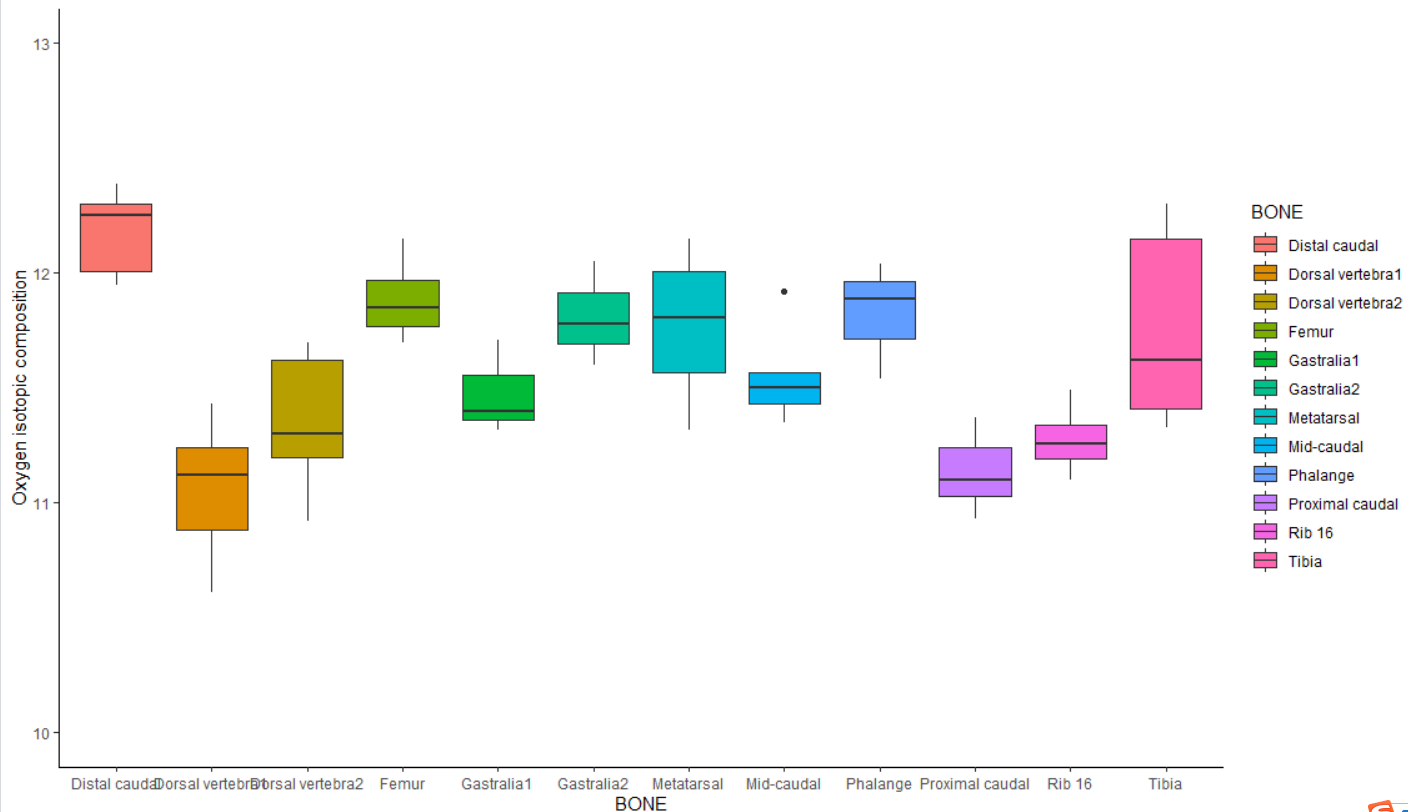


**Problem2**

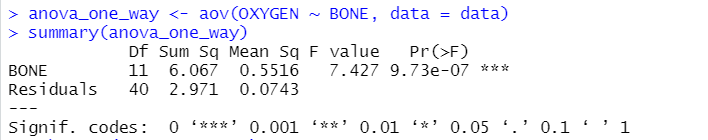
#2.1

本题数据较多，采用Excel表输入数据，并以tibble型读取

以骨头类型聚类，画箱线图，大致可见，各部位骨头的氧同位素组成存在差异，但是差异不大。



接着对样本进行单向方差分析，Pr（>F）=9.37e-07=0.0085>0.001,所以接受原假设，各类骨头磷酸盐氧同位素组分的平均值有99.9%的把握是相同的。即霸王龙的各部分骨头炭形成温度不存在很大差异。已知不同骨骼部位的均值差异表明整个身体的温度不恒定,在温血动物中，预期会有较小的温差。所以，可以认为霸王龙是温血动物。



**Problem3**

#3

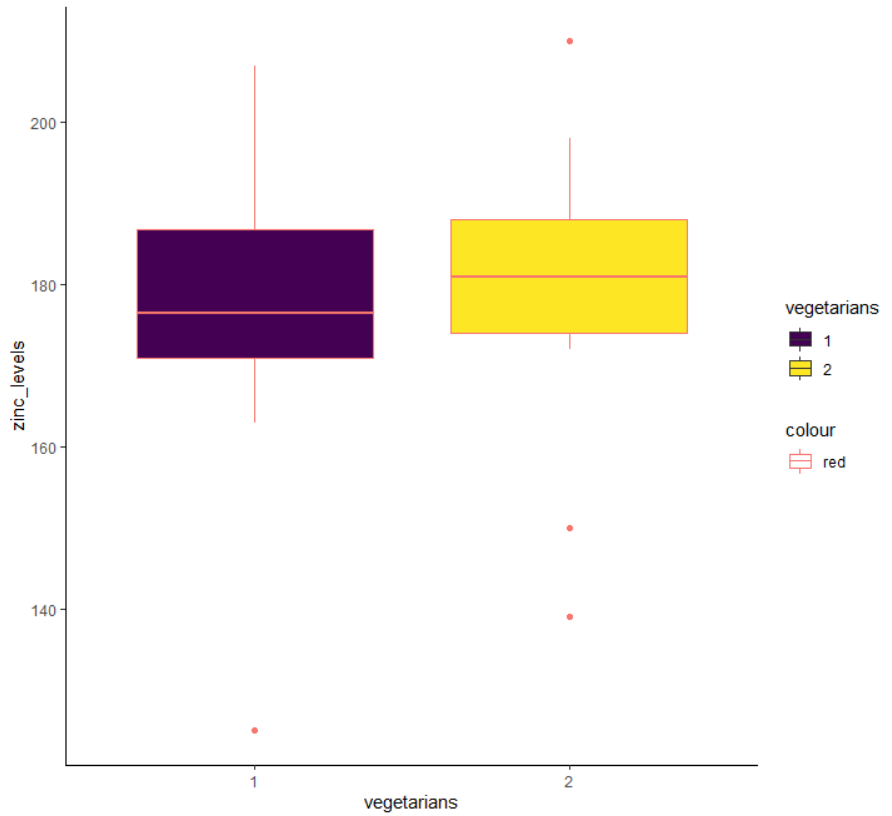
为方便数据输入，

定义pregnant:怀孕状况：因子水平：1和2(1为怀孕)

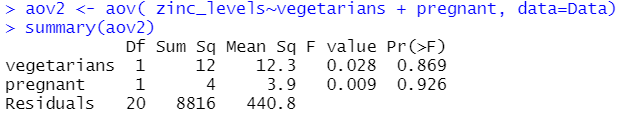
定义vegetarians：是否素食者；因子水平：1和2(1为素食者)

首先设置工作环境，载入程辑包，本题手动输入数据，并将怀孕情况、素食者情况、锌含量的数据按列合并，以tibble型储存在变量中。

做素食者和非素食者身体内，锌含量的箱线图，观察结果，非素食者体内的平均锌含量相比素食者更高一些，但差异不明显。



接着做双因子ANOVA 分析，



解释结果：

是否为素食者之间差异额显著性为：F=0.028。这个值比查表值低，并且p-value>0.05。因此我们接受null hypothesis，即根据是否素质这来评估的锌含量均值都相等，所以变量“素食情况”在锌含量上没有影响。

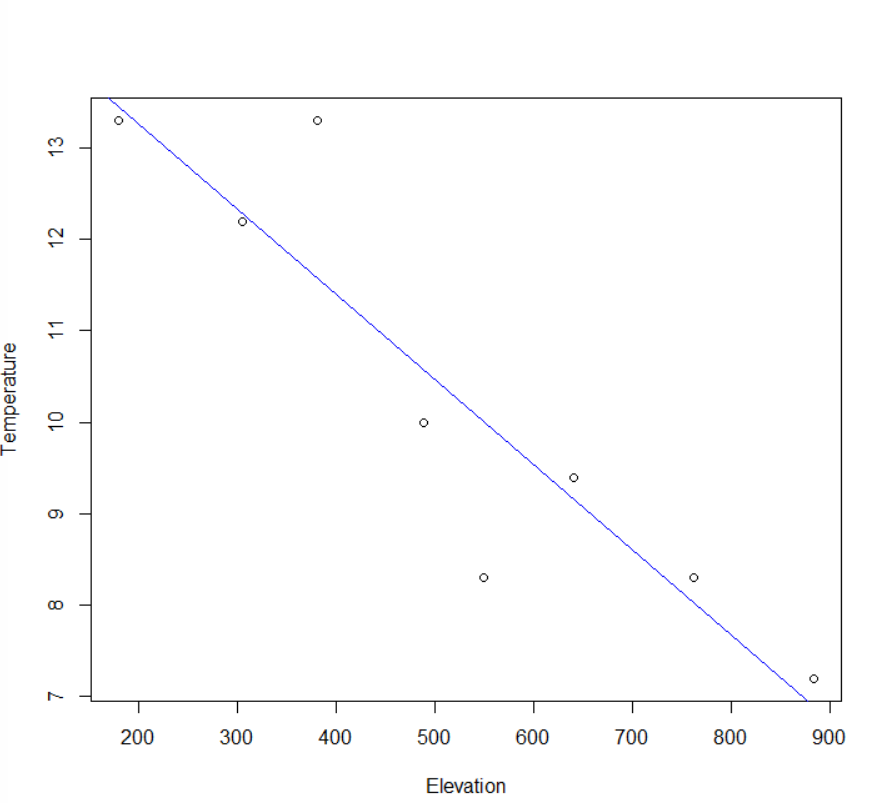
是否怀孕之间差异额显著性为：F=0.009。这个值比查表值低，并且p-value>0.05。因此我们接受null hypothesis，即根据是否怀孕来评估的收入均值都相等，所以变量“怀孕情况”在锌含量上没有影响。

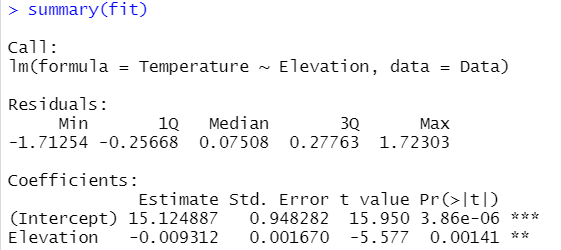
**Problem4**

#4

首先设置工作环境，载入程辑包，因为数据量不大，本题手动输入数据，以tibble型储存在变量中。

画散点图，并进行简单地线性拟合。作图结果如下：





拟合直线的斜率为-0.009312，即温度随海拔的下降速度9.31摄氏度km -1

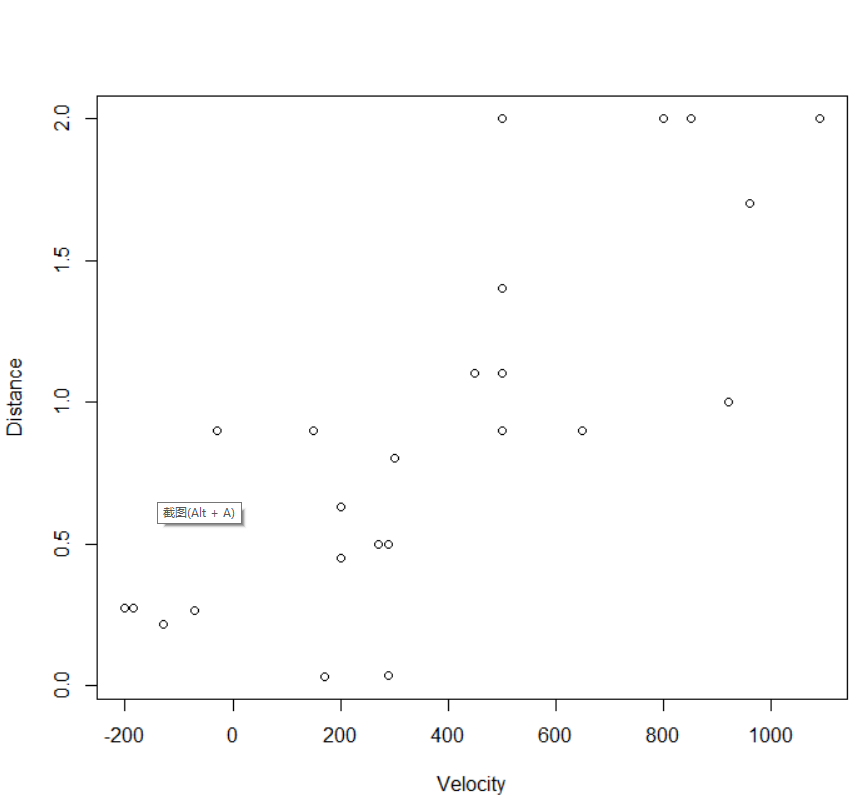
不符合旅行适宜的温度下降率9.8摄氏度km -1

**Problem5**

#5.1

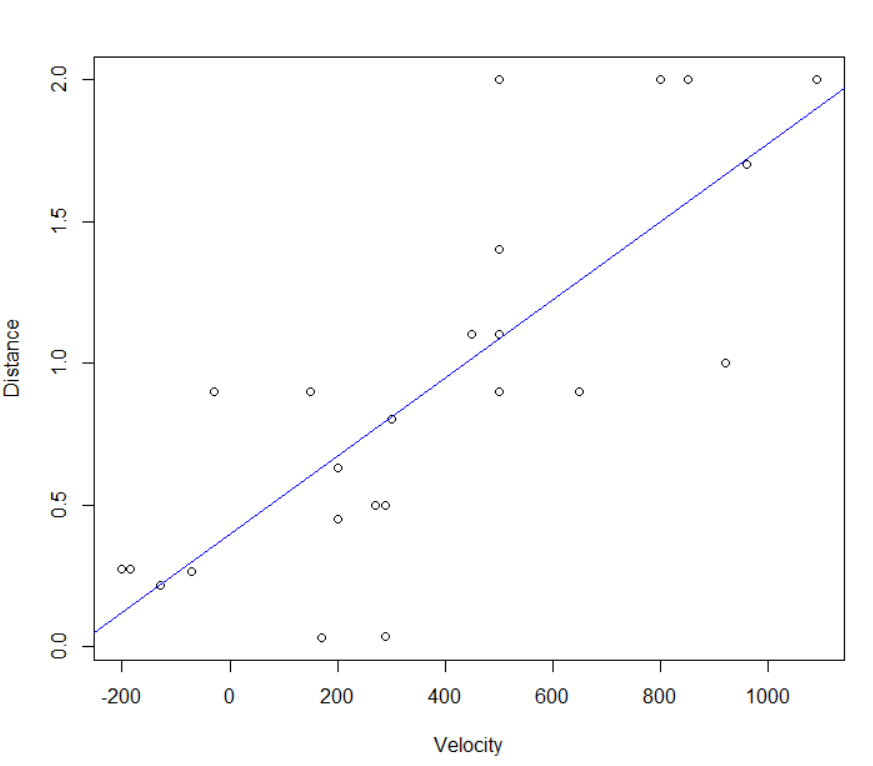
首先设置工作环境，载入程辑包，因为数据量不大，本题手动输入数据，以tibble型储存在变量中。

然后直接绘制散点图即可



#5.2

在此基础上添加一条简单地线性回归线，此线以蓝色标识



#5.3

根据宇宙大爆炸理论，宇宙诞生之初是一个奇点，那时，观测时各空间点的距离和相对速度也为零，即回归线满足假设一，通过原点。哈勃定律表明，宇宙正在膨胀变大，而非一直处于稳恒态。根据哈勃定律可以计算出宇宙的诞生时间：

V=H·D

上式中V表示星系远离银河系的速度，D表示星系与银河系的距离，H表示哈勃常数。通过测量各种星系的远离速度及其对应距离，可以得到比例系数哈勃常数。

对于一个距离银河系非常遥远的星系，假设它从宇宙诞生之初（星系在很早就已经形成）就以恒定的速度在远离银河系，那么，它的退行速度又可以由下式计算出来：

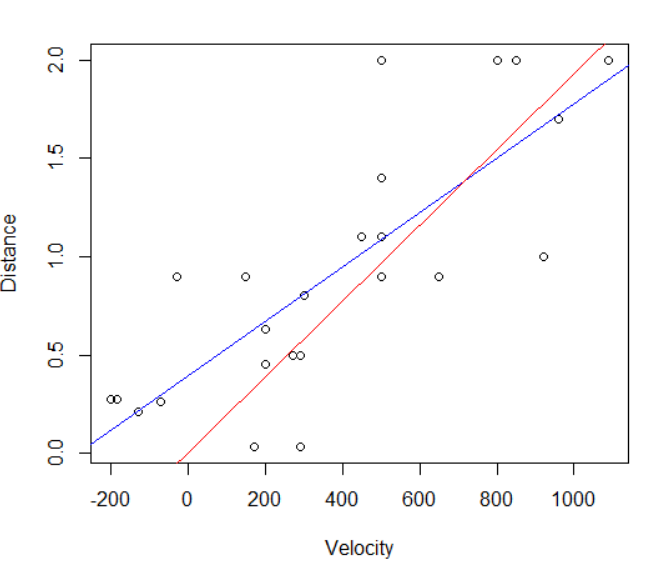
V=D/t

变化一下形式：

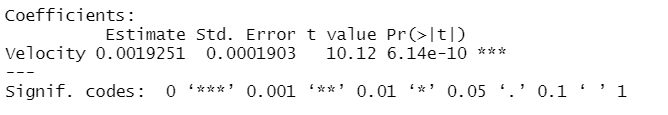
t=D/V

上式中的t可近似认为是宇宙的年龄。然后再把哈勃定律代入上式可得：

t=1/H



在本题所画的图中，红色线为假设一处理后的回归线，宇宙寿命等于回归线坡度，符合假设二。



经过单位换算，输出计算得出的宇宙年龄。



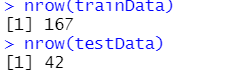
#5.4

改进的距离测量将带来与实值更小的误差，这在线性拟合中将带来更接近1的决定系数，提升回归方程的拟合优度。

**Problem6**

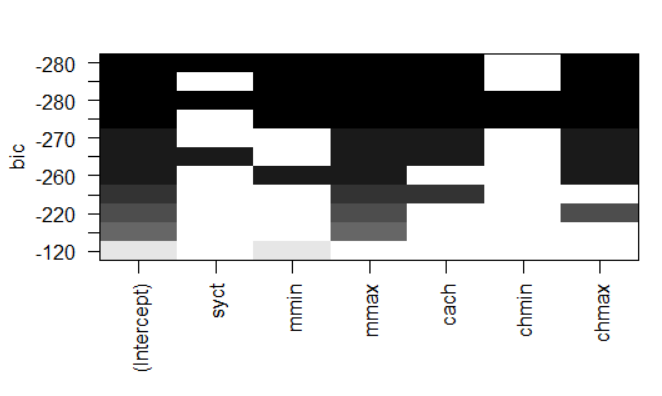
首先载入MASS数据包，获得有关cpu的数据。

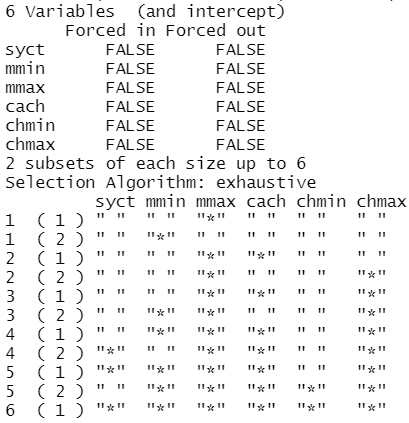
根据syct，mmin，mmax，cach，chmin，和chmax，拟合最佳子集回归以预测cpu性能。我们将随机不放回地抽取80%的数据点作为训练集，以及20%的数据点作为测试集。简单地验证一下结果



#6.1

使用regsubsets（）函数对训练集进行最佳子集回归，结果如下



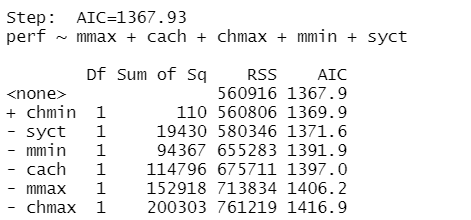


#6.2

将最佳回归模型应用于测试集：

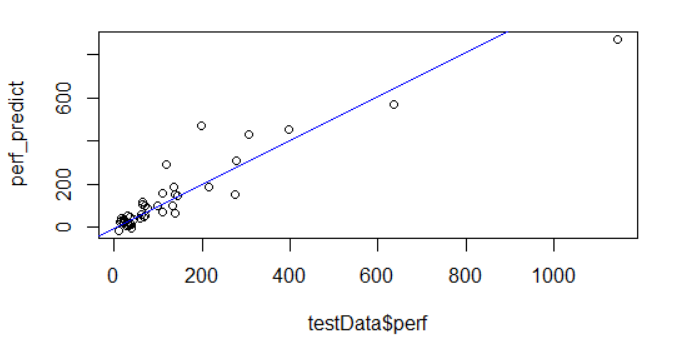
首先建立训练集的线性回归模型，然后分别使用向前逐步回归（forward stepwise）、

向后逐步回归（backward stepwise）、向前向后逐步回归，比较后获得最佳回归模型如下：

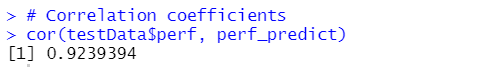


随后将得到的最佳回归模型应用到测试集

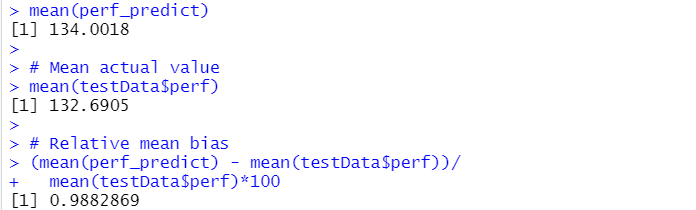
以预测值为x轴，测试集实际值为y轴，进行简单线性拟合，得到的直线斜率为1.01726，说明预测值可信。



同时求出测试值和实际值的相关系数，为



量化预测perf值和提供perf值之间的平均偏差，结果为



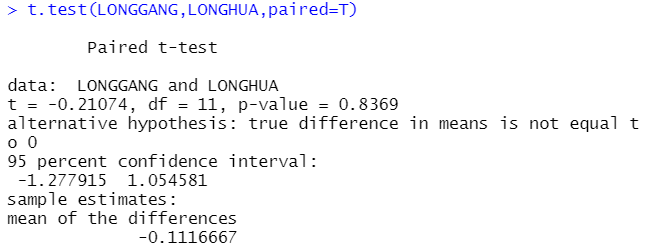
**Problem7**

从课题组中找到了几组数据集，以用于本题

#7.1

数据集一为在同等条件下检验龙华区某地和龙岗区某地的浅层地下水水样测得的碳酸钙含量，每组共12个测量值，共两组。

对两组进行T检验，以探究两不同地域地下水中碳酸钙含量是否不同？



结果解读：p-value = 0.837>0.05，所以两组不同地域的测量结果没有显著差异。

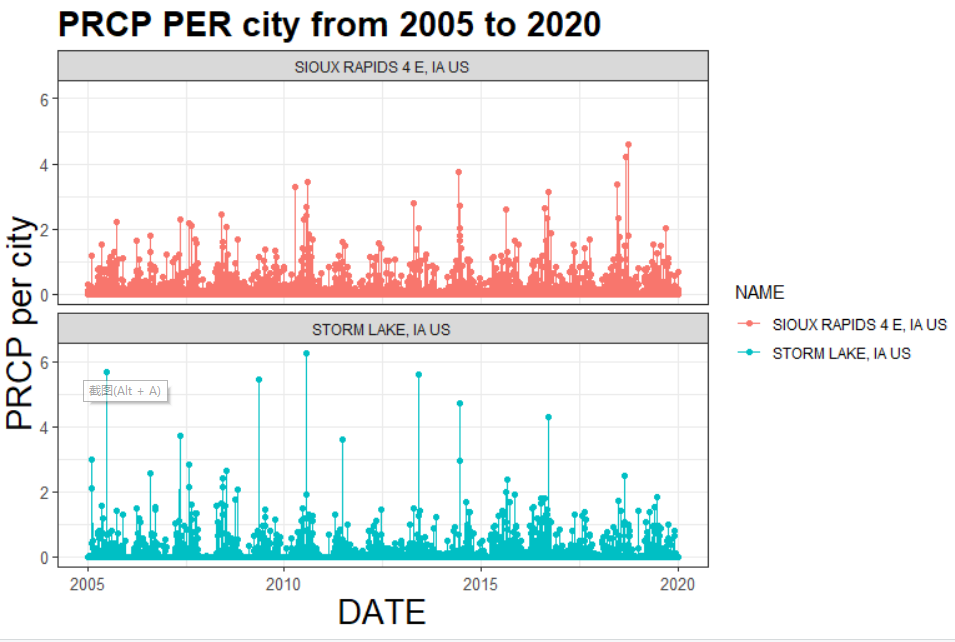
#7.2

根据研究，水体对气候的影响程度主要取决于水体面积、体积，以及水体的形状和周围地形等。一般说来，水体的面积和体积越大，对周围地区气候的影响也越大。水体对其周围气候的影响，涉及气温、风、温度等气象要素和降水、雾等天气现象。本题通过分析美国爱荷华州Storm Lake和远湖山地地区Sioux Rapids4E的降雨量数据，得出水体对降雨量是否有影响，影响效应是怎样的，影响多大。取样点的地理位置如下图所示：



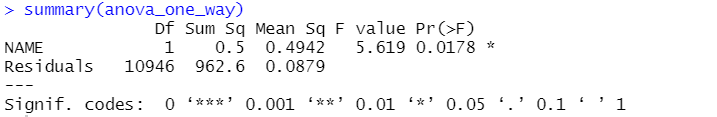
SIOUX RAPIDS 4 E

首先将2015年初至2019年末的两地的降雨量数据读入到脚本中，并将降雨量缺失值赋值为0.因为降雨量本身随季节性变化，所以做两地降雨量的时间序列图，结果如下：



由图可初步得出，两地降雨量随时间变化的趋势大致相同，但是Storm Lake的降雨总量相比同期的Sioux Rapids4E要更大，且极端降雨量也比Sioux Rapids4E要更大。

为进一步说明两地降雨量的偏差，进行单向方差分析



注意到p值为0.0178<0.05,则可说明两组数据显著相关也说两者间确实有明显关系。即地域因素对降雨量有影响，但影响不显著。

#7.3

本题研究CAS City 的降雨量和河道流量的关系，进行简单地线性回归

常规的输入数据，并做降雨量和流量的散点图，接着进行线性回归分析，并将拟合结果画到图中。

