# ****Executive summary/abstract****

**在这个作业中，我使用不同食谱下鸡的体重数据来判断哪个食谱对鸡的生长促进作用最好，并且回答一些概率相关的问题。我通过构建ANOVA贝叶斯模型来回答上面的问题，得到的结论是3号食谱对鸡的生长促进作用最强。而其他概率问题的结论见结果部分。**

# ****Introduction****

## Understand the problem

在课程下方提供的链接中，我找到了关于鸡生长的数据。数据包含了共50只鸡在4种不同的食谱下，21天之内的体重变化。通过这个数据，我们可以研究不同食谱对鸡生长的影响，从而选取最佳的食谱。同时，通过课程中学习的贝叶斯模型，我们可以探究不同食谱之下，食谱对鸡体重影响的概率问题。比如使用各个食谱的鸡比使用其他食谱的鸡生长更好（即体重更大）的概率是多少？ 为了解决上述问题，我们会先对数据进行预处理、数据清洗、数据探索等工作。上述工作会在Python种进行。准备好数据之后，我们将在R Studio中利用Jags进行贝叶斯模型的建立。根据数据我们可以判断，我们需要使用ANOVA贝叶斯模型。

# ****Data****

## Plan and properly collect relevant data

在所选的数据中，包含了鸡在不同食谱下的生长体重数据，由此我们可以根据鸡在不同组下的最终或者最大体重来判断食谱对鸡生长的影响，从而选定最佳食谱以及食谱对鸡生长的概率问题。

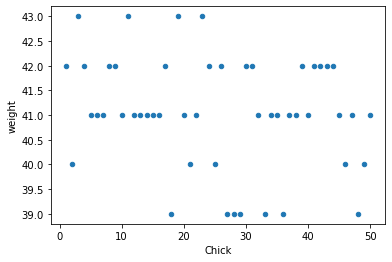
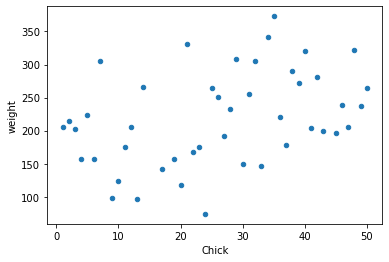
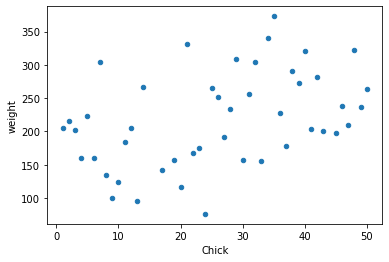
数据来自于课程下方链接，页面中数据详细描述如下： The ChickWeight data frame has 578 rows and 4 columns from an experiment on the effect of diet on early growth of chicks. The body weights of the chicks were measured at birth and every second day thereafter until day 20. They were also measured on day 21. There were four groups on chicks on different protein diets.

在使用页面中所提供的代码进行画图，并且人工大致观察578行数据之后，我们发现大多数鸡有21天的体重数据，少部分鸡只有20天的体重数据，而有几只鸡只有不到20天的体重数据。我们需要对鸡的数据进行筛选，选择使用20或者21天或者历史最终的体重数据作为鸡生长的评判指标。同时我们还要考虑是否将鸡出生的体重作为参考，仅仅将体重的变化量作为鸡生长的评判标准。

## Explore data

首先我们对数据之中的天数进行探究与可视化，观察是否每只鸡的最终天数都相同，如果不同应该选择哪一天作为最终体重，或者将体重的最大值作为最终体重。同时，我们要观察每只鸡的出生体重是否相似，如果相似我们可以忽略鸡最开始体重的影响。如果不相似，我们则需要取最终体重与初始体重的差值来作为鸡生长好坏的评判标准。

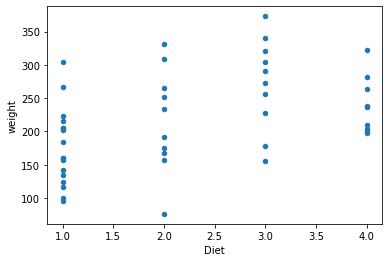
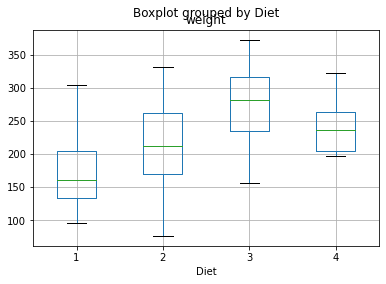
首先我们对鸡的最终天数进行探究，发现有4只鸡的最终测量天数小于20天，有1只鸡的测量天数结束与20天，其他45只鸡的体重测量均结束于21天。

**从左到右分别是鸡在第0天的重量，鸡在第21天的重量，鸡的最大重量**

随后，我们对鸡的初始体重作散点图，并作鸡第21天体重以及最大体重的散点图。发现鸡的初始体重的差别与最终体重的差别比起来相差不大，由此我们可以直接将最终体重最为鸡的生长好坏标准。 同时，我们通过统计可以发现，44只测量结束于21天的鸡中，共有35只在第21天达到体重的最大值，其他9只的体重最大值均不在第21天。由此，我们可以选择鸡体重的最大值作为生长好坏标准。

最后，我们做出每一组鸡最大体重的箱型图，可以看出第三组的鸡的最大体重明显高于其他组。但是，我们还不能仅仅凭此就做出概率的判断。我们还需要对数据进行模型的拟合。同时，我们做出了不同组别之间的鸡体重散点图，通过散点图可以认为组内鸡体重大致成正态分布。



**不同食谱下鸡的体重箱型图与散点图**

# ****Model****

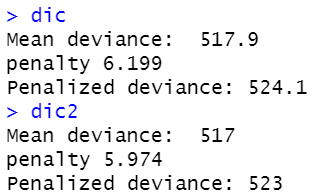
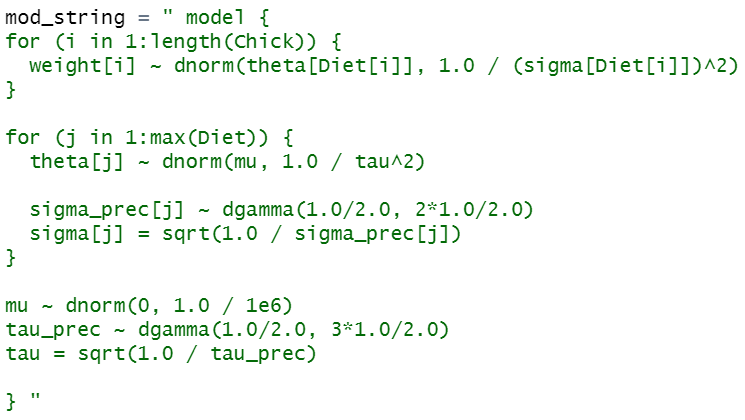
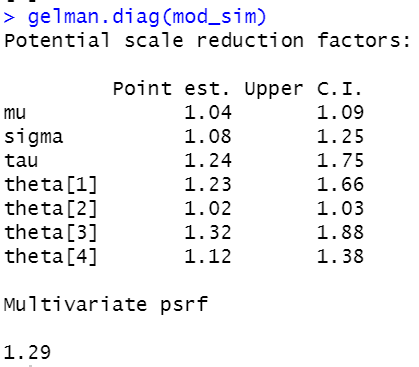
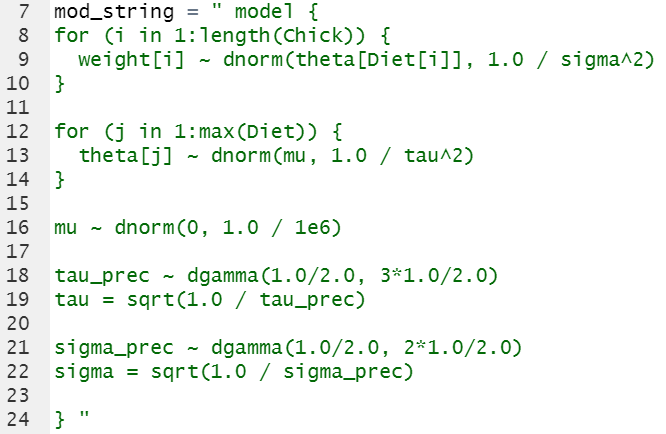
## Postulate a model

对于该数据，我选择使用ANOVA贝叶斯模型。因为我们的问题是探究不同的食谱是否会对鸡的生长产生影响，使用不同食谱的鸡属于不同的组别，所以ANOVA模型是合适的。 通过该模型，我们可以得到不同组别的最终体重平均值以及其后验概率分布，由此可以判断哪一组对鸡生长效果最好以及回答相关的概率问题。 正如上所说，得到的概率分布均值以及后验概率分布可以回答哪个食谱对鸡生长最好以及不同组别间生长更好的概率问题。

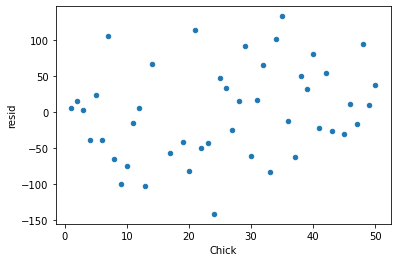
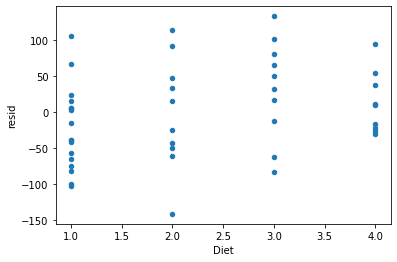
对于该模型，我们可以认为鸡的体重在组内呈现正态分布，不同的组间存在不同的期望。我们可以先假设不同组的组间方差相同，随后假设不同组间的方差不同，然后计算不同模型的DIC值，从而判断哪个模型更好。

对于该模型的先验概率分布，我们可以选择一个信息很少的先验概率分布，从而让模型自己选择最佳的后验概率分布。

**为了探究组内方差对结果的影响，我使用了两个模型。第一个模型中所有的组的方差都相同，第二个模型中所有组内的方差都不同。通过DIC值判断出第一个模型更好，即认为组内方差相同。**

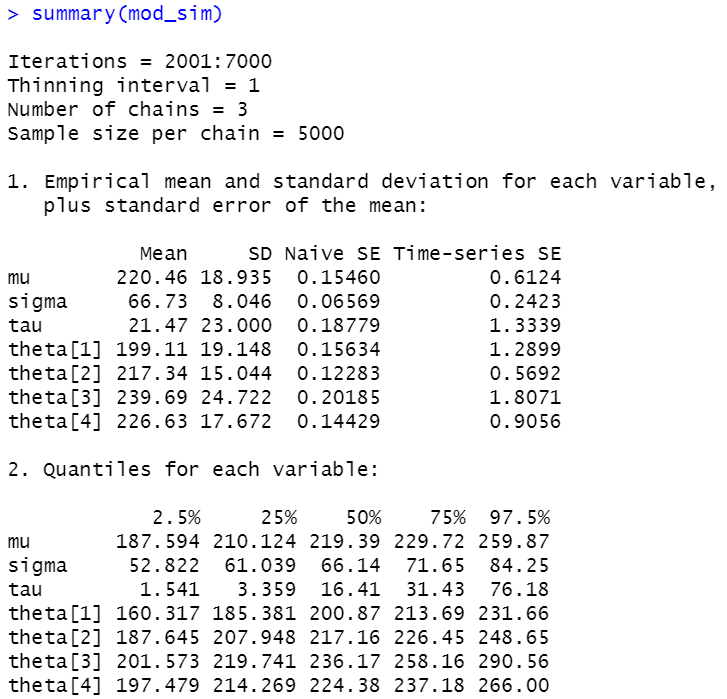


**从左到右分别是模型1的结构，模型1的拟合结果，模型2的结构，模型1、2的DIC值**

** **

**从左到右分别是模型1的残差以及各个组的残差散点图**

# **Results & Conclusions**



**模型参数的后验概率总结**

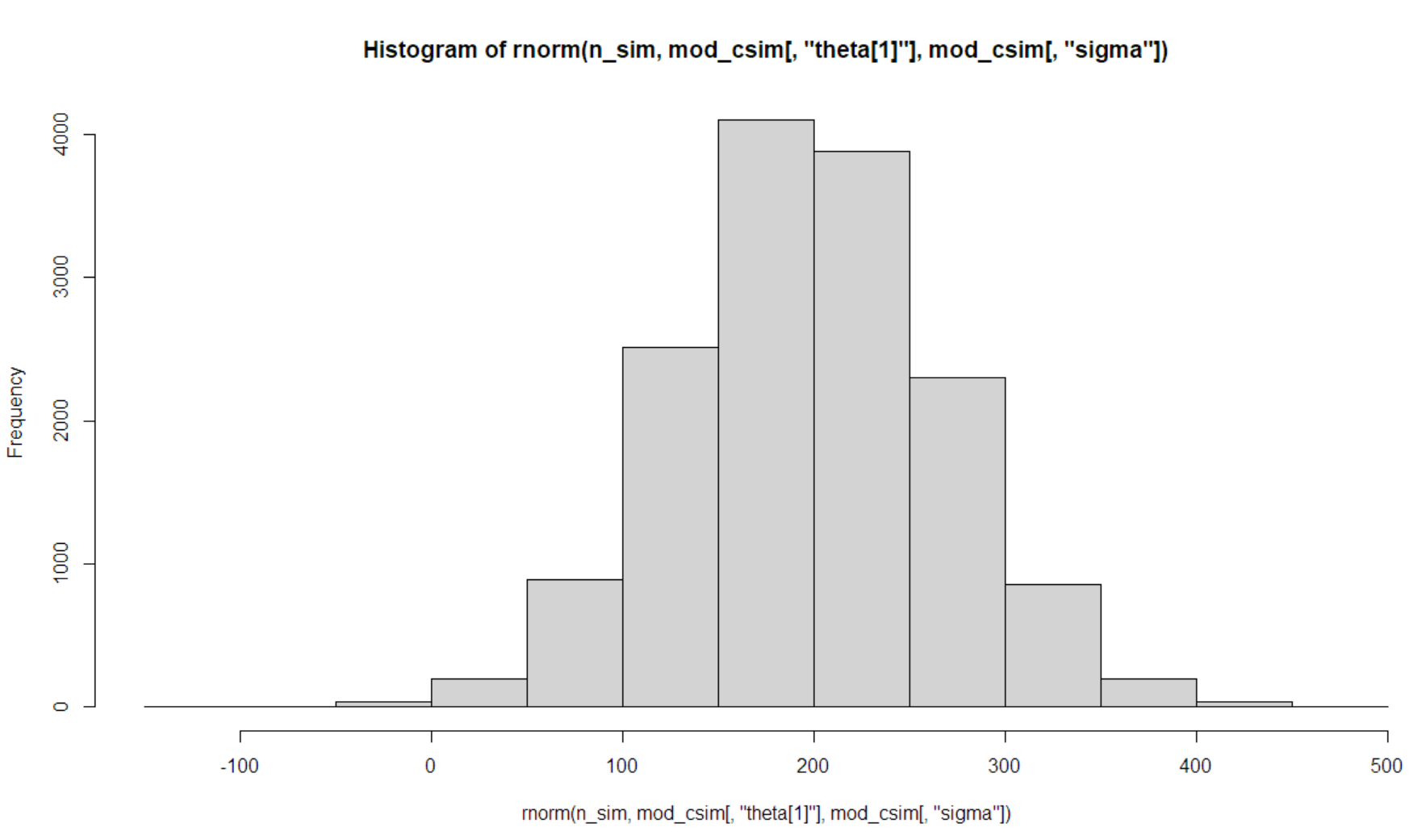
**Theta1~4代表4个食谱的鸡体重平均值，sigma代表鸡体重的方差，mu和tau代表控制theta分布的正态分布的均值和方差。**

**# 一号食谱鸡的体重分布**

**hist(**

**rnorm(n\_sim, mod\_csim[, 'theta[1]'], mod\_csim[, 'sigma'])**

**)**



**# 二号食谱鸡平均体重比一号食谱鸡平均体重高的概率**

**mean(mod\_csim[, 'theta[2]'] > mod\_csim[, 'theta[1]'])**

**0.7944**

**# 一只三号食谱的鸡比一只一号食谱的鸡体重高的概率**

**mean(rnorm(n\_sim, mod\_csim[, 'theta[3]'], mod\_csim[, 'sigma']) > rnorm(n\_sim, mod\_csim[, 'theta[1]'], mod\_csim[, 'sigma']))**

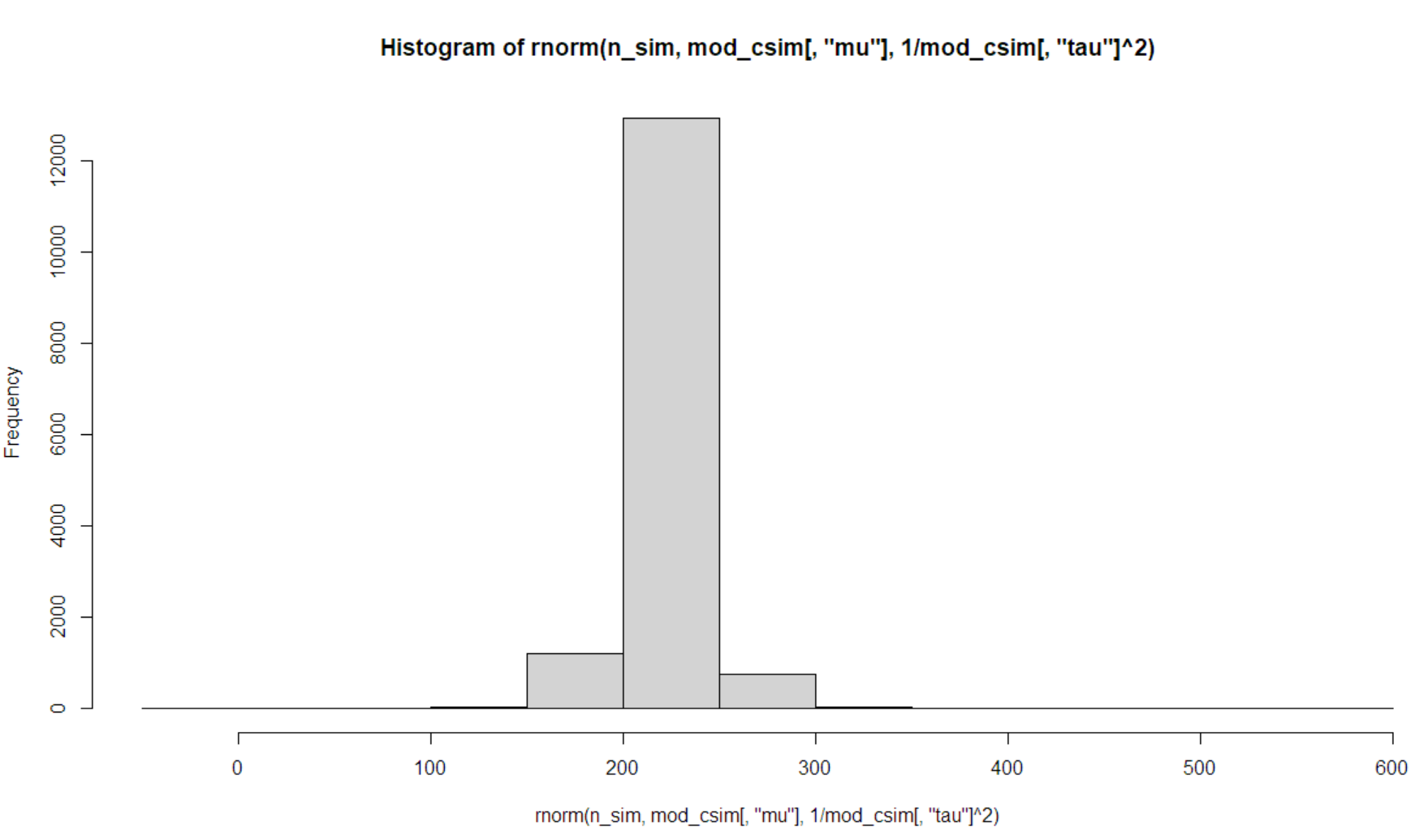
**0.652**

**# 采用一个新的食谱，鸡的平均体重分布以及一只鸡的体重分布**

**hist(**

**rnorm(n\_sim, mod\_csim[, 'mu'], 1 / mod\_csim[, 'tau']^2)**

**)**



**hist(**

**rnorm(n\_sim,**

**rnorm(n\_sim, mod\_csim[, 'mu'], 1 / mod\_csim[, 'tau']^2),**

**mod\_csim[, 'sigma']**

**)**

**)**

