今天首先是回顾上周的测验：

1. 假设检验和AIC/BIC其实都在进行变量选择，它们的区别是什么：假设检验的原假设是这个变量不重要，是一个偏保守的假设，可以防治决策过于偏激，但如果关注预测精度，删除一个变量可能会带来精度上很大的降低，因此可以使用AIC/BIC来选择变量。AIC的信仰是真模型不在待选模型中，继而在待选模型中找一个最像的（具有损失有效性）；BIC的信仰是真模型在待选模型之中（具有选择相合性），AIC和BIC两者不等价。AIC选择出来的变量通常要多于BIC选择出来的变量，实际中可以两个都做，如果AIC模型和BIC模型在某些变量上意见一致，则可以认为这些变量是非常重要的变量。

2. 残差图我们希望残差与X之间呈现杂乱无章的关系，一个表现良好的残差图和预测精度没有关系，只能说明模型没有毛病。

3.根据中心极限定理，如果所有数据是独立分布的，当样本量非常大时几乎都可以用正态分布来处理。

接下来跟大家探讨一个新的内容――方差分析（ANOVA），ANOVA其实就是线性回归分析的一个特例，其特殊之处在于存在离散的X变量。在线性回归的基础上，方差分析学习的核心要点是：哑变量（dummy variable）。

第一部分：北京市二手房交易案例

1、背景问题：什么因素决定了房价？

2、因变量：单位面积的房价Y（做对数变换：稳定性变好，绝对变化率变为相对变化率，但回归系数解读变难了）

3、解释变量：如何设置X指标体系？按生活常识分几个大类（如：距离市中心距离、商超、医疗、教育、交通、安全） 。通过本案例的讨论，希望跟大家一起体验一个二级指标体系的简历过程。

第二部分：建立模型

1、注意因素与水平的区别：一个因素（离散型X变量）可以有多个水平。而建模可以看做是将分组求均值转化为线性回归，最关键的一步是将离散型X变量转化为哑变量，其中哑变量个数=水平数-1 。

2、单因素方差分析：会将因素某一个水平设为baseline, 其他水平对应的模型系数表示的是该水平相对于baseline的差异；截距项表示的是baseline的平均响应

3、双因素模型： （1）可加模型：简单地叠加各个因素的效应，需要的参数个数少，易于解读，但可能与实际情况相悖 （2）交互作用模型：两因素相乘，可以分析因素之间的交互作用，但需要的参数更多（多消耗（因素A的水平个数-1）\*（因素B的水平个数-1）个参数），也加大了解读的难度。

4、实际数据分析中，如果考虑多因素模型，先构建简单可加模型。慎选交互作用。

第三部分：假设检验

方差分析的假设检验理论同线性回归基本一致，包括模型的F检验。但是，有一个核心区别就是关于：因素水平的检验。关于因素显著性检验的基本思想是：比较全模型和把需检验的因素剔除后的模型的RSS，通过构建F统计量，判断一个新因素的显著性。注意两个模型的自由度之差等于两个模型囊括的哑变量个数之差。如果某一个因素既有主效应又有对应的交互作用，可以先检验交互作用是否显著，如果交互作用显著，则主效应也显著