今天是第二次课。首先，大家可以在课堂派的资料板块查看课程的相关资料。还有每堂课后会有测试题，便于对大家学习情况的了解，大家记得完成。

复习上周的内容：

（1）数据分析核心目标：为了创造商业价值。找到好的场景往往非常重要，而含有不确定性的场景通常具有商业价值。

（2）参照系非常重要，有了参照系后可能业务上1%的提升也是一个巨大的提升。

（3）对业务的深入了解可以让我们将笼统的业务问题转化为明确的X和Y的问题。

接下来，是这一节课的主要内容，有三部分：

第一部分：线性回归（案例介绍）

（1）案例：移动通信网络的客户价值分析（校园网计划）

（2）痛点：不同的种子用户能为运营商贡献不同的利润，需在有限的预算中，将更多的资源投放到优秀的种子用户身上，因此需要找到优质的种子用户。

（3）如何选择模型：主要是看因变量Y的数据类型，若Y为连续型数据，优先考虑线性回归。现实中一般具有代数意义且取值丰富的数据就可以被当作连续数据。

（4）因变量Y：被推荐者对运营商贡献的前后变化率。

（5）解释变量X：大网占比、小网占比、通话总量等。

（6）要善于提炼业务问题的目标（Y）和构建合适的（二级）指标体系（X），例如可以从用量、位置和时间等角度更加深入挖掘更多的X变量。

（7）报告的写作：推荐采用三线表绘制表格；数据可视化中，如果连续型变量对应的散点图无明显规律，可以对其进行离散化处理再绘制分组箱线图

第二部分：线性回归（参数估计）

（1）回归分析是研究X与Y的相关关系（不是因果关系）。线性回归方程不等同于线性函数，线性回归中的“线性”是指Y对于未知的回归系数Beta是线性的。 线性模型中的error term体现了不确定性。

（2）有很多估计线性回归系数的方法，OLS只是一种最常用的方法。

（3）注意区分统计量和参数：统计量是不包含未知参数的样本函数

（4）判决系数R-squared的大小可以衡量X对Y的拟合情况（反映模型的拟合优度）。其中R-squared = (1-SSE/SST), SST反映的是Y自身的变异性，不随模型变化；SSE反映的是由random noise产生的变异性，随模型变化而变化。因此SST-SSE表示由解释变量X产生的变异性，

（1-SSE/SST）反映了在因变量Y的变异性中，由解释变量产生的变异性所占的比例。

（5）Adjusted R-squared 和外样本判决系数可以弥补R-squared随变量个数增加而增加的不足，可能会出现负值。

第三部分：线性回归（假设检验）

（1）大数定律是指样本均值收敛到总体均值，中心极限定理是指当样本量变大时，样本均值的分布会收敛到正态分布，与样本所属的总体的分布类型无关

（2） 假设检验是人们在信息不充分的条件下， 面对两种策略（其中一个是保守策略）进行二选其一的决策； 通常选择较为保守的决策作为原假设。假设检验会产生两种错误，一般情况下两种错误带来的损失是不一样的

（3）当考虑H0: β= β0时如何构造统计量？对于参数β有一估计值hat(β)和真值β0，在H0成立时，希望hat(β)-β0 的绝对值应该足够小。但是很难判断绝对值到底多大才叫大。因此需要一个合理的参照系去看待它的大小。hat(SE)是一个很合理的参照系，因为它反映了 hat(β) 的估计误

差。此时构造了t统计量: |hat(β)-β0|/hat(SE)。

（4）对多重检验问题，用模型的全局检验：F-检验。F检验的原假设是认为所有的回归系数为0，从而比较全模型和空模型的残差平方和大小