今天是第三次课程，以下为今天的课堂纪要，希望对大家课后复习总结有帮助。今天的课程主要有三部分内容。

第一部分：课后测验和上周作业进行反馈

1. 接受原假设只能说明没有足够多的证据推翻原假设，但并不意味原假设一定正确。

2. 对变量进行对数变换主要是为了减小异常值对结果的影响，使模型结果更加稳定；如果对变量进行对数变换以后仍存在大量异常值，建议将数据离散化。

3. 在构建指标时，可以对连续型变量进行离散化处理。离散化虽然会损失掉部分信息，但是可以降低原始变量中异常值的影响。

4. 对于作业报告，摘要要简明扼要，阐明主要的工作和发现，切忌冗余；背景介绍可以以三段式展开：行业背景、业务痛点和分析计划；数据介绍要写清数据来源、数据范围和包含字段；模型结果以三线表呈现，不要出现科学计数法和中英文混写；对于描述性分析，主要是进行单因素分析和相关性分析（e.g., 分组箱线图），注意每张图和表都要加以文字表述、标题、编号和引用，不要盲目堆积，不要横跨多页；模型诊断、数学公式等技术细节建议放在附录中；最后可以简要展现一下自己的产品化实现。

第二部分：模型选择

1. 什么是模型选择？在线性模型的框架下选出重要变量。

2. 为什么要做模型选择？理论上，加入不重要的变量会影响模型的精度；实际中，指标的采集有成本，最好将资源投放到重要变量上。

3. 怎样进行模型选择？

（1）第一步：确定全模型。全模型确定变量的搜索范围，其中可能存在变量之间的交互项，也可能存在变量的非线性变换。

（2）第二步：确定选择标准（AIC、BIC）。什么是一个好的模型？可以根据模型选择的标准来挑选好的模型。模型选择标准之一为AIC，它假设真模型不在待选模型中而去近似真模型，它同时考虑模型的拟合优度和模型的复杂度（参数个数），具有损失有效性。AIC的优点：预测精度不错。另一个模型选择标准为BIC，它假设最好的模型就在待选集中，接着从待选集中挑出这个模型。BIC对复杂模型的惩罚力度更强，具有选择相合性。AIC和BIC的根本区别在于是否认为真模型在待选模型中。如何选择？如果要求预测精度，选择AIC；如果要求可解释性和稳定性，选择BIC。

（3）第三步：选择计算方法：最好子集法（计算成本太高，几乎无法实现），前进法（从空模型出发逐渐加入重要变量）；后退法（从全模型出发逐渐删去不重要的变量）、逐步回归法（前进法与后退法的结合）、LASSO+SCAD等等。在实际数据分析中一般选择AIC或BIC准则后使用逐步回归法即可。

（4）第四步：小心解读，谨慎使用：模型选择可能不稳定，采用模型平均法可以提高稳定性。

第三部分：模型诊断

1. 什么是模型诊断？简单来说就是给模型看病，再挑出模型的毛病。

2. 什么是多重共线性？某个变量能被其他多个变量线性表出。含有多重共线性变量的模型存在可识别性问题，即模型系数不能唯一确定。

3. 怎样诊断多重共线性？使用VIF（相当于给每个变量打分），VIF\_j对第j个变量进行诊断，第j个变量作为因变量对其他解释变量进行线性回归。VIF越大，该变量能够被其他变量代替的程度越高,多重共线性影响越严重，一般VIF>10的变量被认为是存在多重共线性的变量。

4. 怎样找异常值：使用Cook距离（相当于给每个样本打分），Cook距离对每个样本进行诊断，删去第i个样本得到的估计值与全样本估计值进行比较。Cook距离越大，该样本对最后回归分析的结果影响也越大。查看Cook距离图时，如果只存在极个别样本的Cook距离与其他样本差别巨大，可以删去这些样本；若差别不大，要允许不完美的存在。

5. 绘制残差图也可以帮助寻找异常值，残差最好与X变量不相关；当存在多个X变量时，可以选择X变量的线性组合——因变量y的估计值来绘制残差图。若个别样本的残差与其他样本的残差相差过大，可以认为它们为异常值从而删去。

6. 在实际数据分析中，如果出现回归模型的系数与预期不相符（例如正负号不对），可以进行模型诊断，删去存在多重共线性的变量和异常值的样本，再进行回归。