本次课首先对上周测验进行了讲解和点评。

1. 对于系数不显著的变量（例如“女司机”），虽然系数为正，但是不应该解读这种变量

2. 逻辑回归是使用极大似然估计，它没有广为接受的残差平方和，但有广为接受的离差

本次课的第一部分：行业案例（位置轨迹）

位置数据的来源：手机、电脑、车联网等等。

1. 案例1:疫情防控与位置轨迹数据

将450万台设备看成种子，对种子进行分类，希望能够将有限的防疫资源投放到重要的种子上。例如去过发热门诊的种子应该是重点防控人群。什么样的土壤最适合种子的生长？即什么是高风险疫情地区。

2. 基于微博数据的景区游客行为特征分析

例如分析在三山五园签到客户的地域分布（外省、北京），外地用户在北京不同城区到达景点、餐饮地点、购物地点、住宿地点的Top5。

3. 基于车联网数据的保险产品创新

有了车联网数据可以区分不同的十公里（例如拥挤路段的十公里和平坦无障碍的十公里），主要数据包括速度、加速度、经纬度、时间、里程、瞬时油耗、发动机转速、轮胎胎压、方向盘转角等等。可以利用以上数据刻画行车轨迹、方向盘转动情况、急刹车行为（t时刻和t+1时刻的速度发生了突变，有点类似异常值）等。基于以上数据可以使用逻辑回归来预测下一年是否出险。可以根据出险率将人群分为高风险、偏高、中等、偏低和低风险人群。

本节课的第二部分：搜索引擎营销

例如在百度中搜索“北京 上海 机票”，搜索引擎会呈现一些付费搜索结果。可供分析的内容包括：广告创意、关键词拓展和关键词竞价。其他可供分析的点：搜索的着陆页面、最优排名/出价、投放地点、投放时间、广告创意等。

本节课的第三部分：点击率预测（在RTB广告投放的应用）

1. 什么是RTB（Real Time Bidding）：实时广告竞拍

2. 一些行业术语：ADX（广告交易平台）、SSP（供给侧平台）、DSP（需求侧平台）

3. 商业模式：帮助广告主向ADX平台竞拍展现机会，向广告主按点击收钱

4. 核心任务：预测点击率。其中X变量包括平台信息、时间信息、广告信息、设备信息。模型采用逻辑回归，发现利于点击的显著因素包括：下午和晚上、竞拍底价高、开屏广告、使用WIFI等；不利于点击的显著因素包括：早上、使用移动数据等。

5. 拍卖规则：价高者得；只需付第二高出价+最小货币单位。

6. 应用模型指导定价：总出价/（点击率\*次数）=单位点击成本

本节课的第四部分：社交网络数据分析

1. 传统统计学假设样本之间是独立同分布的，但是实际世界中的人一定和小伙伴们存在千丝万缕的关系，而网络结构能够描述这种联系。

2. 网络数据=网络拓扑结构+其他相关数据。网络拓扑结构：节点和边的集合。

3. 网络结构的基本特征：稀疏性、传递性、互粉性、幂律分布

4. 基于社交网络分析可以帮助定制个性化广告，例如有标签“旅行”+关注了很多旅游博主的微博用户，可以推荐团游广告；有标签“美食”+关注了烘焙博主的微博用户，可以推荐糕点广告。

5. 移动通信客户流失预警分析，X变量包括：通话人数、人均通话时长、人均通话时长分布等，回归结果发现通话人数是非常显著的变量，说明社交网络是一个值得关注的重要的资产。