# 第一部分：读书报告

## 第13章

1. y观测值、均值、估计值（由样本回归方程得出）
2. 误差平方和SSE（y与差值平方和）、离差平方和SST（y与差值平方和）、

回归平方和SSR（与差值平方和）

1. SST = SSE + SSR（SSE为0，拟合程度最好都在线上；SSR为0，SST = SSE，SSE

最大，拟合程度最差）

1. 样本决定系数r2 表示模型拟合程度。r表示两变量相关关系的程度。
2.  = （样本相关系数与符号相同）
3. 显著性检验：根据回归方程，如果等于0，则E（y）= ，表明y与x无

线性关系。

1. 回归分析可以用于确定变量之间的关联关系，不能用于表明变量间存在因果

关系。

1. 残差分析是用于验证回归模型的假定是否有效的一种主要方法。

## 第14章

多元回归分析是简单线性回归的延伸，表示y与x1、x2、x3…、xp值相关。多元回归样本决定系数可以衡量拟合程度。F检验用于检验变量间关系是否显著，t检验判断各个参数（各个自变量）的显著性。

残差图检测是否违背模型假定，既要检测误差项，也要检测模型的函数形式。

## 第15章

1. 建立模型的主要问题在于找到变量关系的合适的函数，还有模型选择中该包括的自变量。
2. 模型有线性、曲线、交互作用。
3. 变量选择方法：逐步回归分析、前向选择、后向消去、最优子集回归分析。
4. 在模型建立的初步阶段，变量的选择方法虽然有用，但是不能代替分析中经验和判断。
5. 解决多变量问题的重点是找到自变量的最优子集。
6. DW检验自相关。

# 第二部分：数据分析

（模拟数据不会做，只做了现有数据的分析）

一个完整的数据分析方案需要包含:

1. Orders的描述分析

在order表中：

Orderid订单编号，数值型变量，由七位数字组成，唯一标识。

Customerid用户编号，数值型变量，存在缺失值，表中用0表示，唯一标识。

Campaignid，活动编号，数值型变量，由四位数字组成。

Orderdate，订单时间，名义变量，包含年月日。

City，城市，名义变量。

State，州，名义变量，城市属于州。

Zipcode，数值变量，邮政编码，有部分邮政编码不是数字，可能是邮编描述方式不同，可能是异常值。

Payment，名义变量，支付方式。

Totalprice总价，数值变量，小数精确到小数点后两位，有部分缺失值，表中用0表示，最大值3603.6，最小值2.47。

1. 用其他数据(外部数据)可以做的解决方案

对每条记录，订单对应着活动编号，因此该数据一定是指活动期间客户订单

业务情况。

可以猜测为电子商务交易订单，如：天猫交易订单。

图1 活动促发的订单量

由图可知，2236号活动引发的订单数最多，其次为2141号、2173号、2010

号，其余活动订单和归为其他活动，而其他活动没有超过订单数第四位（2010号）。

因此在以后交易中可通过多举办2236、2141、2173、2010号活动增加订单数

量。

图2 2011-2014年销售额前十的州排名

从图2中可以得知2011-2014年活动期间，订单销售额最高的州为NY，销售额总额达到167867元，销售额是第二名CA州的两倍，说明NY州对产品需求远大于其他州。

因此，NY、CA商家可根据当前订单扩大库存，避免库存不足。

图3 totalprice分布直方图

图3纵向比较各年度及合计totalprice的频数分布。图3横向对各组左图频数最高的密集区域进行细分，分析其范围的频数分布。

从图3可以看出，2011-2014年活动期间销售额大部分在100元内，2013、2014年及合计年度的销售额100元内，销售额大多为0到40。因此，举办活动，商家可以采取薄利多销方式，在保证利润基础上，将商品降至0-40范围。

图4 2011-2014年订单统计与销售额的条形图

从图4 可以看出：2011年活动期间订单数最少，整个年度只有1单，2013年活动期间订单数最多，总额达到7931单。2011-2013年，订单数逐年增多，而2014年订单小于2013年。2013活动期间年销售额最高，从2011年-2013年，销售额逐年增加，2012-2013年变化幅度很大，而2014年比2013年销售额低。

可以猜测有可能是活动增多引发订单增多。

1. 模拟数据

不太会，希望老师说讲解下。

1. 基于模拟数据做的分析工作

