第七章：多元线性回归
Multiple Linear Regression

王树佳 | 深圳大学经济学院
sjwang123@163.com

例1：意大利餐厅的定价策略

纽约两位年轻律师Tim和Nina Zagat计划在曼哈顿的第五大道开一家新的意大利餐厅。

他们的目标定位是：

A. 在食物方面，要“提供最高质量的食品”；
B. 在服务方面，要树立“本地区意大利餐厅服务质量新标准”。

为了给餐厅菜单定价，他们在目标地区组织了一次抽样调查，调查了168位意大利餐厅顾客，得到了消费价格、顾客对食品、装饰、服务的评价等方面的数据。

例1：意大利餐厅的定价策略

Y =Price: 正餐价格（包括饮料）（美元）

X_1=Food: 顾客对食物的评价（总分30分）

X_2=Décor: 顾客对装饰的评价（总分30分）

X_3=Service: 顾客对服务的评价（总分30分）

X_4=East: 餐厅地址（1=在第五大道东边，0=西边）

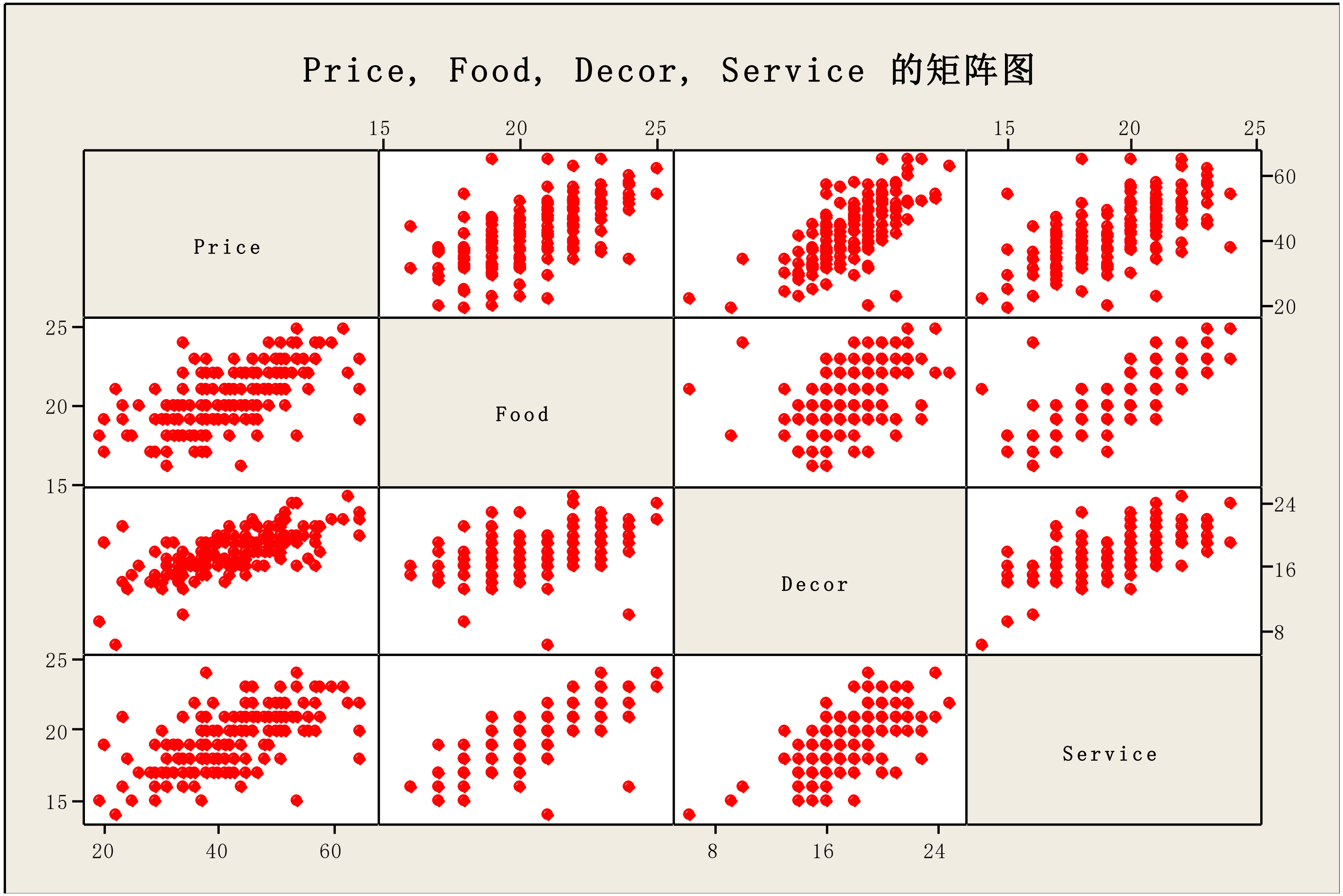
建立模型：研究价格（Price）与Food、Décor、Service和East之间的关系。

例1：意大利餐厅的定价策略

希望回答如下问题：

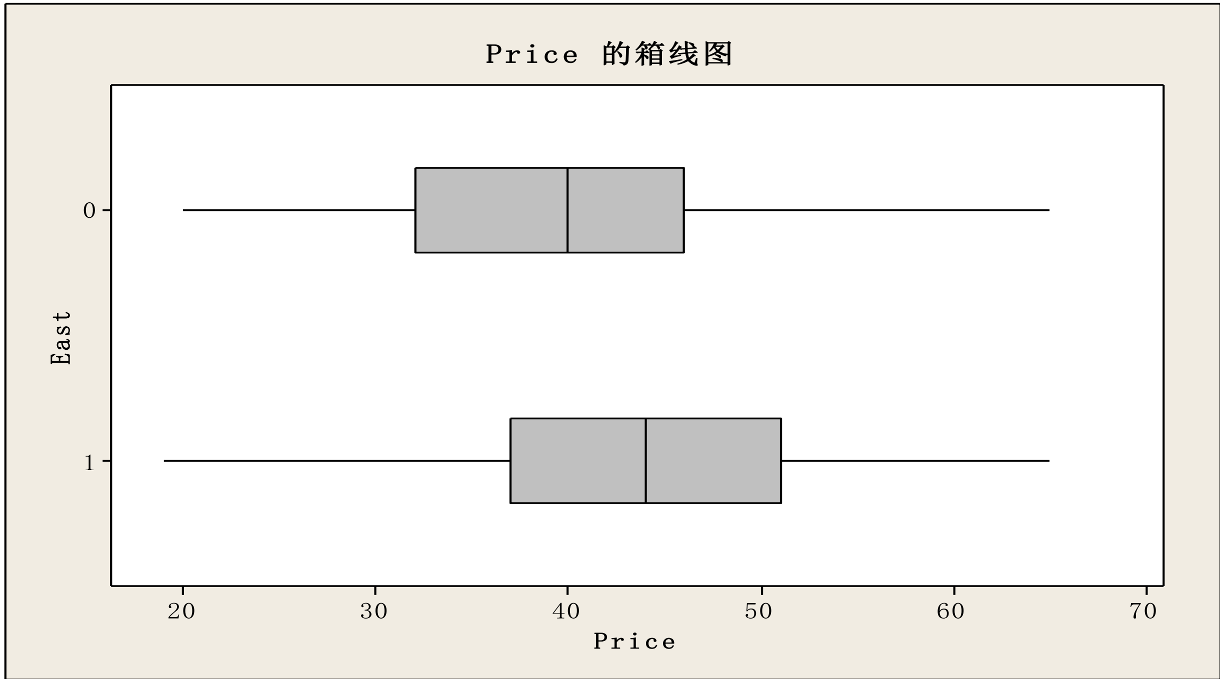
1. 基于调查数据，如何为菜单确定食品价格？
2. 食物、装饰和服务，哪项对价格影响最大？
3. 为了获得高价，新餐厅应该选在第五大道的东边还是西边？
4. 在服务方面，树立“曼哈顿地区意大利餐厅服务质量的新标准”是否能获得相应的溢价？

描述性分析（散点图矩阵）



Price, Food, Decor, Service 的矩阵图
Price
Food
Decor
Service
15
20
25
60
40
20
24
16
8
8
16
24

描述性分析（箱线图）



描述性分析能否回答餐厅希望知道的四个问题？

Contents

1. 多元线性回归模型：概念
2. 多元线性回归模型：估计
3. 多元线性回归模型：推断

Contents

模型概念

多元回归模型的目的：研究多个自变量与因变量之间的关系（因变量如何受其它因素的影响）。

- □ 因变量Y：响应变量，被解释变量。
  - ✓ 因变量根据研究目的决定；
- □ 自变量X_1，X_2，…，X_n：解释变量，预测变量。
  - ✓ 自变量是影响因变量Y的因素。
- □ **做法**：根据观察数据（y_i,x_i), i=1,2,...,n，估计Y与自变量x_i之间的函数关系。
  - ✓ 如何估计？
  - ✓ 误差如何？

模型概念

多元线性回归模型：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

- β_0：截距
- β_i：回归系数，表示在其它因素不变条件下，自变量x_i对因变量的影响大小
- ϵ：随机误差，表示除自变量 $x_1, x_2, \ldots x_p$以外，其它所有可能的因素对Y的综合影响
- 线性：指对参数是线性函数
  - ✓ 如$x_2 = x_1^2$或$x_2 = \ln x_1$仍为线性模型

回归模型的函数形式

A. 线性模型

wage = - 4.474 + 1.281 education

- Wage：时薪; Education：受教育年限
- 回归系数的解释：受教育年限每增加1年，工人的时薪平均增加1.281美元
- 一般：$y = \beta_0 + \beta_1 x$

解释：x增加一个单位，y增加β_1个单位。

推导：$dy = d(\beta_o + \beta_1 x) = \beta_1 dx$

- 不合理之处：受教育年限为10和20时是一样的

回归模型的函数形式

B. 半对数模型

$$\ln(wage) = 1.116 + 0.093education$$

- 回归系数的解释：受教育年限每增加1年，工人的时薪平均增加9.3%美元
- 一般：$\ln(y) = \beta_0 + \beta_1 x$

解释：x增加一个单位，y增加100β_1%。

推导：$d\ln(y) = d(\beta_o + \beta_1 x),\ \dfrac{dy}{y} = \beta_1 dx$

即 $\%dy = (100\beta_1)dx$

回归模型的函数形式

C. Lin-Log模型

Engel = 0.93 - 0.8 ln(Expend)

- 回归系数的解释：家庭总支出每增加1%，恩格尔系数将减小0.008.
- 一般：$y = \beta_0 + \beta_1 \ln(x)$

解释：x增加1%，y增加 0.01β_1个单位。

推导：

$$dy = d(\beta_o + \beta_1 \ln(x)) = \beta_1 \frac{dx}{x} = (0.01\beta_1)\%dx$$

回归模型的函数形式

D. 对数模型

著名的Cobb-Douglas生产函数：$Q = AL^a K^b$

其中Q为产出，L为劳动投入，K为资本，A为常数。

化为线性模型：$\ln Q = C + a \ln L + b \ln K$

- 一般：$\ln(y) = \beta_0 + \beta_1 \ln(x)$

解释：β_1为y对x的弹性(elasticity)，即：x增加1%，y增加 β_1%。

推导：$d \ln y = d(\beta_o + \beta_1 \ln x)$

$$\frac{dy}{y} = \beta_1 \frac{dx}{x}$$

Contents

模型假设

1. 线性(Linearity)：因变量与自变量之间存在线性关系
2. 零均值：误差项ϵ的平均值（数学期望）为0
3. 同方差性，或方差齐性（Homoscedasticity)：误差项ϵ_i的方差为常数（不同的观察值）
4. 独立性：不同数值下，误差项ϵ_i相互独立
5. 正态性：误差项ϵ服从正态分布
6. 自变量之间不存在共线性(Collinearity)：自变量中没有常数变量，且自变量之间不存在线性关系

数据

样本序号	x_1	x_2	$\cdots$	x_p	y
1	x_{11}	x_{21}	$\cdots$	x_{p1}	y_1
2	x_{12}	x_{22}	$\cdots$	x_{p2}	y_2
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
n	x_{1n}	$x_{2\mathrm{n}}$	$\cdots$	x_{pn}	y_n

数据模型

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$

其中 $\epsilon_i \sim iidN(0, \sigma^2), i = 1,2, \ldots, n$

多元线性回归模型：估计

1. 系数估计
   - ✓ 最小二乘法：LSE
   - ✓ 最大似然法：MLE
2. 估计精度
   - ✓ 误差的方差
   - ✓ 标准误
3. 拟合优度
   - ✓ 判定系数R^2

估计：最小二乘法（OLS）

最小二乘法：使误差平方和$\sum(y_i - \hat{y}_i)^2$最小

$$\text{Min} \sum\left[y_i - \left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}\right)\right]^2$$

- 需估计系数$\beta_0, \beta_1, \dots, \beta_p$和方差$\sigma^2$
- 系数估计量记为：$b_0, b_1, \dots, b_p$
- $\hat{y}_i = b_0 + b_1 x_{i1} + \cdots + b_p x_{ip}$称为回归方程
- $e_i = y_i - \hat{y}_i$称为残差(Residual)
- 统计软件（Minitab，SPSS，SAS...)
- OLS是一种估计方法，不是OLS模型！

估计：最大似然法（MLE）

最大似然估计：使似然函数最大

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$

$$Y_i \sim N(\mu_i, \sigma^2)$$

其中 $\mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$

似然函数

$$L(\beta, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i - \mu_i)^2}{2\sigma^2}}$$

例1：意大利餐厅的定价策略

回归分析: Price 与 Food, Decor, Service, East

回归方程为

Price = - 24.2 + 1.55 Food + 1.90 Decor + 0.005 Service + 2.00 East

自变量	系数	系数标准误	T	P
常量	-24.244	4.758	-5.10	0.000
Food	1.5539	0.3727	4.17	0.000
Decor	1.8979	0.2191	8.66	0.000
Service	0.0053	0.3995	0.01	0.989
East	2.0034	0.9618	2.08	0.039

S = 5.77649　R-Sq = 62.7%　R-Sq（调整） = 61.8%

方差分析

来源	自由度	SS	MS	F	P
回归	4	8991.8	2248.0	67.37	0.000
残差误差	160	5338.9	33.4		
合计	164	14330.7			

例1：意大利餐厅的定价策略

- ✓ 因为顾客对食品、装饰、服务的评价都是30分制，所以系数大小可以对比；
- ✓ 装饰（Décor）对价格影响最大：系数为1.90，表示在固定其他变量条件下，顾客对装饰的评价每提高1分，食品价格平均提高1.90美元；
- ✓ 服务（Service）对价格影响最小：系数仅为0.005，表示在固定其他变量条件下，顾客对服务的评价每提高1分，食品价格平均仅提高0.005美元；
- ☐ 问题是：此回归方程是否能真正用于决策？

精度：误差的方差和标准误

- 系数的估计：估计了均值 $E(Y|x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$

- 最近很时髦的词汇：
  精准医疗，精准扶贫。。。
- 现在，我们要精准估计，精准预测，看什么？

- 精度：看$\epsilon \sim N(0, \sigma^2)$的方差$\sigma^2$！

精度：误差的方差和标准误

y的总平方和(SST) = 被回归方程所解释的变异(SSR)

+ 未能被回归方程解释的变异(SSE)

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- σ^2的无偏估计：$\hat{\sigma}^2 = \frac{SSE}{n-p-1}$
- $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$ 称为估计的标准误(Standard Error)

拟合优度：判定系数

判定系数(Coefficient of Determination)：

$$R^2 = \frac{SSR}{SST}$$

- ✌ R^2反映因变量的总变动中由自变量所解释部分所占的百分比
- ✌ R^2反映线性模型拟合的好坏程度
- ✌ R^2反映线性模型的预测能力

拟合优度：判定系数

调整的判定系数(Adjusted R²)：

$$R_{adj}^2 = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$$

$$R_{adj}^2 = 1-(1-R^2)\times\frac{n-1}{n-p-1}$$

- ✌ 增加与Y无关的自变量，R_{adj}^2不变
- ✌ 多元线性回归要看R_{adj}^2

例1：意大利餐厅的定价策略

回归分析: Price 与 Food, Decor, Service, East

回归方程为

Price = - 24.2 + 1.55 Food + 1.90 Decor + 0.005 Service + 2.00 East

自变量	系数	系数标准误	T	P
常量	-24.244	4.758	-5.10	0.000
Food	1.5539	0.3727	4.17	0.000
Decor	1.8979	0.2191	8.66	0.000
Service	0.0053	0.3995	0.01	0.989
East	2.0034	0.9618	2.08	0.039

S = 5.77649 R-Sq = 62.7% R-Sq（调整） = 61.8%

方差分析

来源	自由度	SS	MS	F	P
回归	4	8991.8	2248.0	67.37	0.000
残差误差	160	5338.9	33.4		
合计	164	14330.7			

Contents

1. 多元线性回归模型：概念
2. 多元线性回归模型：估计
3. 多元线性回归模型：推断

多元线性回归模型：推断

A. 显著性检验
   1. 回归方程的显著性检验：F-检验
      - ✓ 自变量整体上对Y是否有显著解释力
      - ✓ 回归系数不能全部等于0
   2. 回归系数的显著性检验：t-检验
      - ✓ 检验每个自变量对Y是否有显著影响
      - ✓ 即检验它们的系数是否等于0

B. 预测

多元线性回归模型：推断

A. 显著性检验

1. 回归方程的显著性检验：F-检验
   - ✓ 自变量整体上对Y是否有显著解释力
   - ✓ 回归系数不能全部等于0
2. 回归系数的显著性检验：t-检验
   - ✓ 检验每个自变量对Y是否有显著影响
   - ✓ 即检验它们的系数是否等于0

B. 预测

回归方程的显著性检验：F-检验

回归方程检验或总体显著性检验

(test for overall significance)

$$y=\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p+\varepsilon$$

$\mathrm{H}_0：\beta_1=\beta_2=\cdots=\beta_p=0$

$\mathrm{H}_1：\beta_1,\beta_2,\ldots,\beta_p$至少有一个不等于0

回归方程的显著性检验：F-检验

在多元线性回归模型中，一般需要给出方差分析表 (Analysis of Variance, ANOVA):

来源	自由度	SS	MS	F
回归	p	SSR	$MSR=SSR/p$	$F=MSR/MSE$
误差	n-p-1	SSE	$MSE=SSE/(n$-p-1)	
合计	n-1	SST		

回归方程的显著性检验：F-检验

F-检验（总体显著性检验）的关键步骤：

1、假设：要检验整个回归方程是否显著

H_0：$\beta_1 = \beta_2 = \cdots = \beta_p = 0$

H_1：$\beta_1, \beta_2, \ldots, \beta_p$至少有一个不等于0

2、检验统计量：$F = MSR/MSE$

3、判断规则：对于给定显著性水平α，(0.05，0.01)

- ✓ 如果p-值≤α，拒绝H_0，回归方程显著
- ✓ 如果p-值>α，接受H_0，回归方程不显著

例1：意大利餐厅的定价策略

回归分析: Price 与 Food, Decor, Service, East

回归方程为

Price = - 24.2 + 1.55 Food + 1.90 Decor + 0.005 Service + 2.00 East

自变量	系数	系数标准误	T	P
常量	-24.244	4.758	-5.10	0.000
Food	1.5539	0.3727	4.17	0.000
Decor	1.8979	0.2191	8.66	0.000
Service	0.0053	0.3995	0.01	0.989
East	2.0034	0.9618	2.08	0.039

S = 5.77649　R-Sq = 62.7%　R-Sq（调整） = 61.8%

方差分析

来源	自由度	SS	MS	F	P
回归	4	8991.8	2248.0	67.37	0.000
残差误差	160	5338.9	33.4		
合计	164	14330.7			

回归系数的显著性检验：t-检验

t-检验（单个自变量显著性检验）的关键步骤：

1、假设：要检验第i个自变量是否显著

$H_0：\beta_i = 0$　　　　$H_1：\beta_i \neq 0$

2、检验统计量：$t = \dfrac{b_i}{S_{b_i}} \sim t(n-p-1)$

- ✓ b_i为回归方程中第i个自变量系数β_i的估计量
- ✓ S_{b_i}为b_i的标准差的估计量（也叫系数标准误）

3、判断规则：对于给定显著性水平α，(0.05，0.01)

- ✓ 如果p-值$\leq\alpha$ ，拒绝H_0 ，该自变量显著
- ✓ 如果p-值$>\alpha$ ，接受H_0 ，该自变量不显著

例1：意大利餐厅的定价策略

回归分析: Price 与 Food, Decor, Service, East

回归方程为

Price = - 24.2 + 1.55 Food + 1.90 Decor + 0.005 Service + 2.00 East

自变量	系数	系数标准误	T	P
常量	-24.244	4.758	-5.10	0.000
Food	1.5539	0.3727	4.17	0.000
Decor	1.8979	0.2191	8.66	0.000
Service	0.0053	0.3995	0.01	0.989
East	2.0034	0.9618	2.08	0.039

S = 5.77649   R-Sq = 62.7%   R-Sq（调整） = 61.8%

方差分析

来源	自由度	SS	MS	F	P
回归	4	8991.8	2248.0	67.37	0.000
残差误差	160	5338.9	33.4		
合计	164	14330.7			

例1：意大利餐厅的定价策略

1. 基于调查数据，如何为菜单确定食品价格？

回归方程：

Price=- 24.2+1.55Food+1.90Decor+0.005Service+2.00East

- ✌ 能否用于预测食品价格？
- ✌ 答：还不能用于预测食品价格，因为没有经过检验

例1：意大利餐厅的定价策略

2. 食物、装饰和服务，哪项对价格影响最大？

✌ 答：装饰（Décor）对价格影响最大，因为其系数最大，t-检验的p-值表明其高度显著

例1：意大利餐厅的定价策略

3.为了获得高价，新餐厅应该选在第五大道的东边还是西边？

✌ 答：East的t-检验p-值=0.039<0.05，表明其对食品价格有显著影响。系数=2，表明在第五大道的东边(East=1)平均比西边价格高2美元。

例1：意大利餐厅的定价策略

4. 在服务方面，树立“曼哈顿地区意大利餐厅服务质量的新标准”是否能获得相应的溢价？

✌ 答：想通过提高服务标准获得溢价，这个目标恐怕难以实现。因为Service对价格的影响很小且不显著

多元线性回归模型：推断

A. 显著性检验

1. 回归方程的显著性检验：F-检验
   - ✓ 自变量整体上对Y是否有显著解释力
   - ✓ 回归系数不能全部等于0
2. 回归系数的显著性检验：t-检验
   - ✓ 检验每个自变量对Y是否有显著影响
   - ✓ 即检验它们的系数是否等于0

B. 预测

多元线性回归模型：推断

A. 显著性检验

1. 回归方程的显著性检验：F-检验
   - ✓ 自变量整体上对Y是否有显著解释力
   - ✓ 回归系数不能全部等于0
2. 回归系数的显著性检验：t-检验
   - ✓ 检验每个自变量对Y是否有显著影响
   - ✓ 即检验它们的系数是否等于0

B. 预测

例1：意大利餐厅的定价策略

由于Service不显著，回归方程中应予以剔除。

回归分析：Price 与 Food, Decor, East

回归方程为

Price = - 24.2 + 1.56 Food + 1.90 Decor + 2.01 East

自变量	系数	系数标准误	T	P
常量	-24.238	4.723	-5.13	0.000
Food	1.5574	0.2662	5.85	0.000
Decor	1.8993	0.1917	9.91	0.000
East	2.0055	0.9458	2.12	0.035

S = 5.75853　R-Sq = 62.7%　R-Sq（调整） = 62.1%

方差分析

来源	自由度	SS	MS	F	P
回归	3	8991.8	2997.3	90.39	0.000
残差误差	161	5338.9	33.2		
合计	164	14330.7			

例1：意大利餐厅的定价策略

Food和Décor的最大得分都是25，在第五大道东边开店的话，预测定价的回归方程为：

Price = - 24.2 + 1.56 Food + 1.90 Decor + 2.01 East

例1：意大利餐厅的定价策略

新观测值的预测值

新观测值	拟合值	拟合值标准误	95% 置信区间	95% 预测区间
1	64.185	1.384	(61.452, 66.918)	(52.489, 75.881)

新观测值的自变量值

新观测值	Food	Decor	East
1	25.0	25.0	1.00

1. 概念
   A. 目的？
   B. 如何确定因变量、自变量？
   C. 不同函数形式：系数的意义？

2. 估计
   A. 模型假设
   B. 回归系数的估计方法
   C. 误差估计
   D. 拟合优度

3. 推断
   A. 显著性检验
      a) 各个自变量的系数的检验
      b) 回归方程的检验
   B. 预测

讨论课问题

问题一：如何报告多元回归模型的结果？

在实证研究中，要撰写多元线性回归的分析报告，需要报告哪些内容？

问题二：一个二元线性回归方程为 $y = 30 + 1.8x_1 - 3.2x_2$

1. 解释回归系数的意义；
2. 假如因变量y变为$ln(y)$，解释回归系数的意义；
3. 假如因变量y变为$ln(y)$，自变量x_1变为$ln(x_1)$，x_2变为$ln(x_2)$，解释回归系数的意义。

实践课问题

一家电影院希望掌握每周票房收入与广告投入之间的关系。过去8周的数据如下：

表1：电影院过去8周的经营数据（单位：万元）

时间（周）	票房收入	电视广告	报纸广告
1	96	5.0	1.5
2	90	2.0	2.0
3	95	4.0	1.5
4	92	2.5	2.5
5	95	3.0	3.3
6	94	3.5	2.3
7	94	2.5	4.2
8	94	3.0	2.5

实践课问题

1. 用软件Minitab建立每周票房收入与电视广告投入之间的线性方程，指出模型的拟合优度。
2. 建立每周票房收入与电视广告、报纸广告投入之间的线性回归方程，此时电视广告投入的回归系数与（1）是否相同？分别解释其意义；
3. 检验（2）所建立的回归方程及回归系数的显著性（显著性水平取0.05）。
4. 假设该电影院计划在第9周加大广告投入，准备投入8万元，其中5万元投入电视广告，3万元投入报纸广告。请预测该电影院第9周的营业票房收入，并以95%的置信水平，求出第9周票房收入的预测区间。