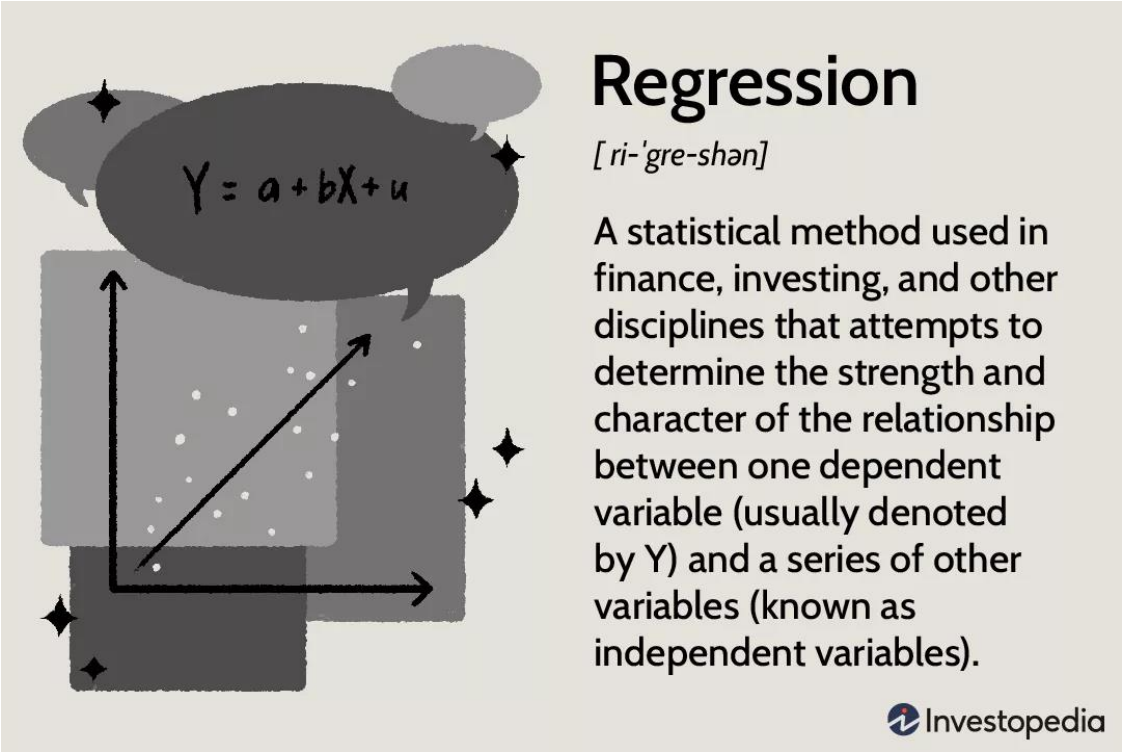
第三章 线性回归

任浩

材料物理系

renh@upc.edu.cn

线性回归（Linear Regression）

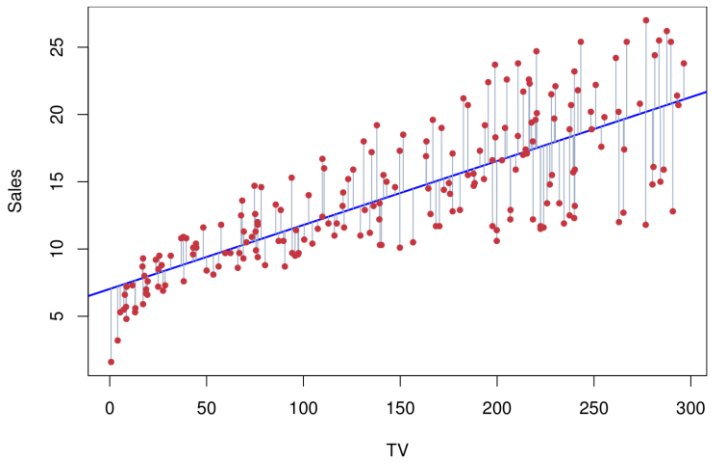


- 19世纪，Francis Galton第一次用于描述生物学现象：身高较高的父母所生的后代，其身高往往会向人群平均值回归。
- 最小二乘法（least squares）：勒让德（Legendre，1805），高斯（Gauss，1809）用于确定行星轨道；1821，高斯-马尔科夫定理
- Udny Yule （1897），Karl Pearson （1903）扩展了统计学意义
- Fisher （1922,1925）弱化了Yule和Pearson的假设
- 二十世纪50-60年代，应用于经济学

线性回归（Linear Regression）

- （可能是）最简单的统计学习工具，也是最为广泛使用的统计学习方法之一
- 以此为基础发展了多个更加精巧的方法，有助于学习其他统计学习算法
- 本章任务：
  - 深入理解线性回归
  - 回顾最小二乘法

以广告数据为例



Advertising数据中，销量对电视广告预算作图

任务：基于销售数据，制定广告投送方案

- 广告预算与销量是否相关?
- 若存在相关，相关性有多强?
- 哪类媒体能够促进销售?
- 不同媒体广告如何影响销量?
- 对未来销量的预测能达到何种精度?
- 得到的关联是否是线性的?
- 不同的媒体广告之间是否有协同效应?

简单线性回归（Simple linear regression）

根据单一预测变量X预测响应值Y的方法：

- 假定X和Y之间存在线性关系

$$Y \approx \beta_0 + \beta_1 X$$

- 上式称为Y对X的回归（regressing Y on X）
- 以广告数据为例，X可能为电视广告投入，Y为销量增长

$$\text{sales} \approx \beta_0 + \beta_1 \times \text{TV}$$

- β_0和β_1为未知，对应线性模型中的截距和斜率，称为模型中的系数（coefficient）或参数（parameter）
- 训练结束后，可根据上式预测（一定电视广告投放预算下）未来的销量

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

hat符号“^”表示对未知参数的估计

估算系数

实际问题中，β_0和β_1都未知，需基于已有数据集进行估计。

令$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$表示n组观测数据，即对$i = 1, 2, \cdots, n$，有

$$y_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i$$

即：寻找截距β_0和斜率β_1使得上式对应的直线尽量接近广告数据集中的200组数据。

需要某种方法定量描述接近程度（closeness）：残差平方和最小化准则（最小二乘，least squares）

残差平方和（residue sum of squares）

残差：基于变量X中的第i个值x_i来估计

$$\hat{y}_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i$$

则对应的第i个残差（residual）为 $e_i = y_i - \hat{y}_i$

残差平方和则为：

$$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2 = \sum_i e_i^2$$

或写为：

$$\mathrm{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \cdots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

最小二乘估计

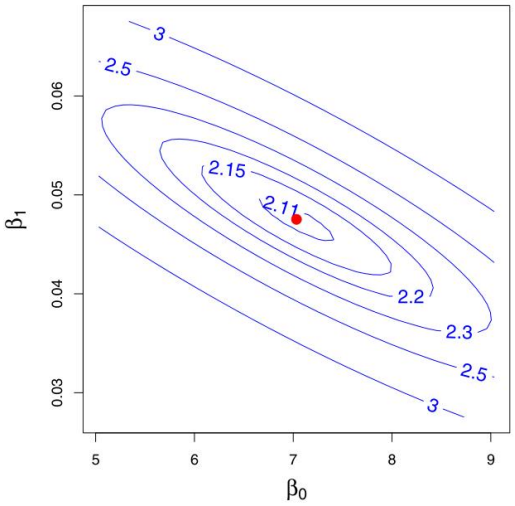
寻找使残差平方和最小的β_0和β_1值，得到：

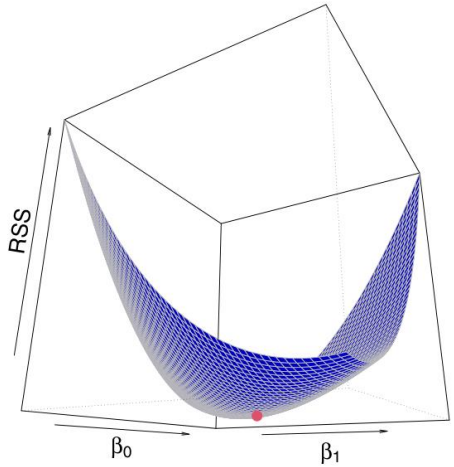
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

此处$\bar{y} = \frac{1}{n}\sum_i y_i, \bar{x} = \frac{1}{n}\sum_i x_i$。

即为简单线性回归系数的最小二乘估计（least squares coefficient estimate）





准确性评估

若X和Y之间的真实关系为 $Y = f(X) + \epsilon$，其中ϵ为均值为零的随机误差，f为某未知函数。则在线性回归中，可表示为：

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

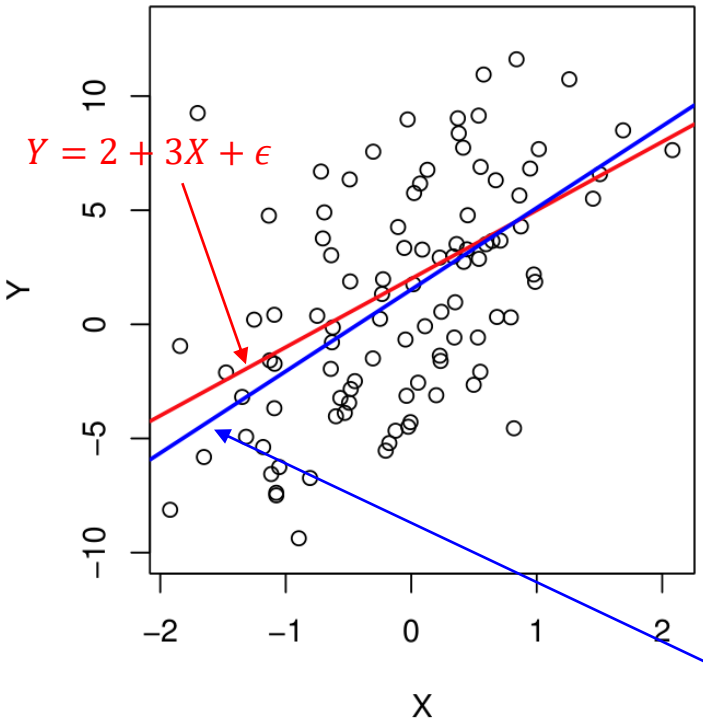
其中的误差项包含了该简单模型所未考虑的因素，如：

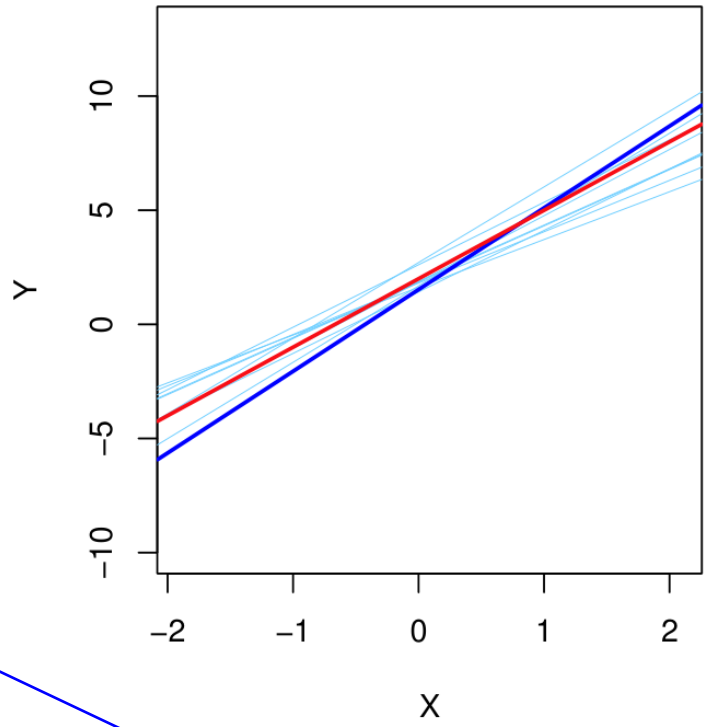
- 真实关系并非线性
- 除X外，还有其他因素影响Y
- 测量误差

上式为总体回归直线（population regression line），是对X与Y之间真实关系的最佳线性估计。

准确性评估

使用样本信息估计总体特征





总体无偏
unbiased

红线为真实关系（实际问题中难以得到），蓝线为基于数据得到的最小二乘估计
淡蓝色线：基于不同的随机数据（子集）的到的最小二乘线，各自不同，但其均值接近总体回归直线

准确性评估：与均值估计类比

只有一个数据集，为何会有不同的线性回归直线？它们分别揭示了特征和响应之间的什么关系？

例：估计总体平均值。对一个随机数据集Y，在无法获取全部y_i时，估算其均值μ。

所取数据集不同，会得到不同的均值估计$\hat{\mu}$，可能高估，也可能低估，但总体无偏（unbiased）。随训练集增大，$\hat{\mu}$总体趋向真实均值。

无偏：模型不会对真实关联进行系统性的高估或低估。

准确性评估：与均值估计类比

$\hat{\mu}$ 对真实均值 μ 估计的准确性如何？

标准差：

$$\mathrm{Var}(\hat{\mu}) = \mathrm{SE}(\hat{\mu})^2 = \frac{\sigma^2}{n}$$

真实数据的标准差

样本量增大，标准差变小

线性回归（当ϵ_i之间不相关时）：

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right]$$

$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

x分布越广，精度越高

置信区间（confidence intervals）

95%置信区间：

- A range of values such that with 95% probability, the range will contain the true unknown value of the parameter.
- 重复选取不同样本，并基于每个样本构造置信区间。则有95%的区间含有模型参数的真值。

线性回归：

- β_1的95%置信区间

$$\hat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_1)$$

- β_0的95%置信区间

$$\hat{\beta}_0 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_0)$$

置信区间：广告数据为例

β_0的95%置信区间为 [6.130, 7.935]

- 若取消掉所有广告投入，销量将下降至6130~7935之间

β_1的95%置信区间为 [0.042, 0.053]

- 每增加$1000电视广告投入，销量将增长42~53单位

假设检验（hypothesis test）

零假设（null hypothesis）：

- X和Y之间并无关联
- $\beta_1 = 0$

$$Y = \beta_0 + \epsilon$$

备择假设（alternative hypothesis）：

- X和Y之间存在关联
- $\beta_1 \neq 0$

需验证β_1是否为零（β_1的值与0相比是否差距足够大）？多大算大？
依赖于我们对β_1估计的准确度。。。$SE(\hat{\beta}_1)$

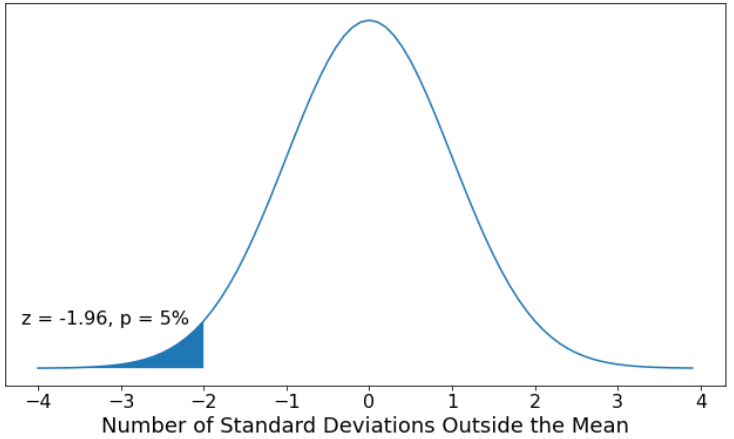
z统计量（z-statistic, z-test）

z统计量用于度量样本属性偏离总体属性有多远（相差多少个标准差）

- 确立零假设：样本均值与总体均值相等
- 确立备泽假设：样本均值与总体均值不等
- 选取一个临界值，如α，度量我们对零假设/备择假设的接受程度，如$z = \pm 1.96$
- 计算z统计量

$$z = \frac{\hat{\mu} - \mu}{\sigma/\sqrt{n}}$$

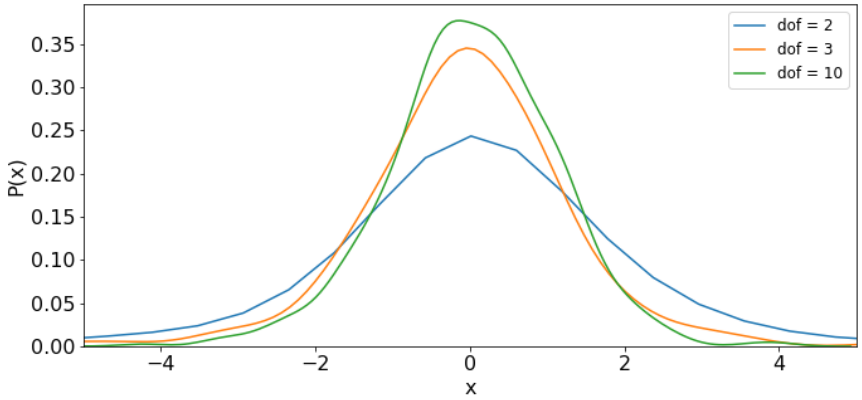
- 若z统计量大于临界值，则备择假设为真。



t统计量（t-statistic，t-test）

t统计量通常用于度量两个样本之间的统计偏差

- William Gosset, Student's T-test
- 总体偏差和总体标准差未知
- 样本数大于30，样本之间独立
- T-分布，类似于正态分布，但更宽更尖
- 样本数（自由度）增大，收敛于正态分布。



t统计量（t-statistic，t-test）

可用t统计量度量$\hat{\beta}_1$偏离0的幅度（相差多少个标准差）

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}$$

若X与Y无关，则服从t分布；

若样本数大于30，则近似于正态分布，可计算偏离0的幅度大于某个临界值（如$|t|$）的概率，即为p值（p-value）.

若p值足够小，即偏离0足够远，接受备择假设（或拒绝零假设，reject the null hypothesis）

t统计量（t-statistic，t-test）

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

Advertising数据集中，将销量对TV广告做线性回归所得结果。

- β_1估计值为0.0475，远大于其标准差（0.0027）
- T统计量为17.67，距0值差距较大
- p值小于0.0001，表明TV广告和销量之间确实存在关联
- 可以确定

$$\beta_0 \neq 0, \text{and } \beta_1 \neq 0$$

模型准确性评价

量化模型对数据的描述有多好（或多差）？

- 标准化残差（residual standard error， RSE）

$$\mathrm{RSE}=\sqrt{\frac{\mathrm{RSS}}{n-2}}=\sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$$

- R^2 统计量

$$R^2=\frac{\mathrm{TSS}-\mathrm{RSS}}{\mathrm{TSS}}=1-\frac{\mathrm{RSS}}{\mathrm{TSS}}$$

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

模型准确性评价：RSE

Advertising 数据集中销量对TV广告的线性回归估计的标准残差和R^2

Quantity	Value
Residual standard error	3.26
R^2	0.612
F-statistic	312.1

- 由于存在随机误差，不可能基于X对Y进行完美预测
- RSE是对标准偏差 (ϵ) 的估计，每个样本实际销量偏离真正回归直线约3260个单位
- OR：即使模型正确，且系数的真实值已知，基于TV广告的预测将有3260单位的偏离
- 该偏离是否是一个可接受的结果依赖于具体应用场景

模型准确性评价：R^2

- RSE是模型对真实关联拟合程度的一个绝对测度方法，但以Y的单位衡量，通常难以确定RSE的可接受范围（多小的RSE算好？）
- R2统计量为一比例形式，在0~1间取值，与Y的量级无关。

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} \qquad \mathrm{TSS} = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2$$

- TSS为总方差，是响应变量Y中故有偏差（variance，即便没做回归也存在）
- RSS为回归分析之后仍存留的偏差，无法被当前模型解释的部分
- R^2为数据偏差中当前回归模型能够解释的部分
- R^2➔0：回归效果较差（数据本身非线性、数据本身偏差大）

多个特征的情况

- 实际场景中，数据存在不止一个特征（Advertising，三个特征）
- 简单线性回归可描述响应与单个特征的关联
- 如何描述响应与多个特征的关联?
  - 建立三个简单线性回归模型，分别对应响应值与三个特征的关联
  - 若给定总预算，如何分配?
  - 每个回归模型都忽略了其他特征的影响

Simple regression of `sales` on `radio`

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	9.312	0.563	16.54	< 0.0001
`radio`	0.203	0.020	9.92	< 0.0001

Simple regression of `sales` on `newspaper`

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	12.351	0.621	19.88	< 0.0001
`newspaper`	0.055	0.017	3.30	0.00115

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	7.0325	0.4578	15.36	< 0.0001
`TV`	0.0475	0.0027	17.67	< 0.0001

多元线性回归

- 更好的方法：建立一个统一的模型，包含三个特征与响应的关联
- 利用一个模型预测每个特征导致的响应变化
- 若有p个特征，多元线性回归可写作：

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

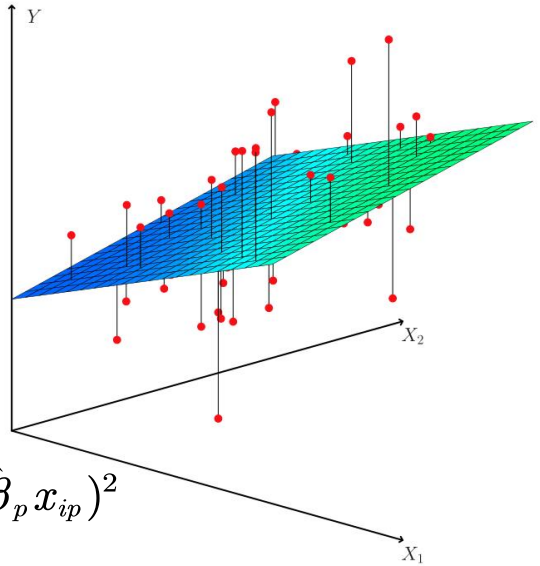
- β_j为第j个特征与Y之间的关联，即在所有其他特征不变情况下，X_j改变一个单位对Y产生的平均效果。
- 对于Advertising数据，

$$\text{sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times \text{newpaper} + \epsilon$$

多元线性回归：求解

多元线性回归问题：

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$



通过最小化RSS

$$\begin{aligned} \mathrm{RSS} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2 \end{aligned}$$

得到回归系数$\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_p$。

实际场景中可利用统计学习软件或模块计算。

多元线性回归：求解

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	−0.001	0.0059	−0.18	0.8599

多元
线性回归

	Coeff.	Std	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	<0.0001
TV	0.0475	0.0027	17.67	<0.0001
Intercept	9.312	0.563	16.54	<0.0001
Radio	0.203	0.020	9.92	<0.0001
Intercept	12.351	0.621	19.88	<0.0001
newspaper	0.055	0.017	3.30	0.00115

三个简单
线性回归

特征间相关性

两组数据X和Y之间的相关性：

$$\mathrm{Cor}(X,Y)=\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}}$$

	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

进行多元线性回归需考虑的几个问题

- 特征$X_1, X_2, \cdots, X_p$中是否至少有一个与响应有相关性？（可用来进行预测）
- 是否所有特征都有助于预测响应？（或许只有一部分特征可关联至响应）
- 模型对数据拟合得有多好？或多差？
- 给出一组特征值，预测的响应值是多少？这个预测/估计有多准确？

问题1：响应和特征之间有无关联?

- 简单线性回归：可用t-统计量、p值等判断X与Y是否相关（$\beta_1 = 0$?）
- 多元线性回归：是否 $\beta_1 = \beta_2 = \cdots = \beta_p = 0$?
  - 零假设

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

  - 备择假设

    H_a: 至少有1个β_j不为零
- 使用F-统计量描述

F-统计量（F-statistic）

$$F = \frac{(\mathrm{TSS} - \mathrm{RSS})/p}{\mathrm{RSS}/(n-p-1)}$$

其中：

- $\mathrm{TSS} = \sum(y_i - \bar{y})^2$
- $\mathrm{RSS} = \sum(y_i - \hat{y}_i)^2$

若线性模型正确：

- $\mathrm{E}\{RSS/(n-p-1)\} = \sigma^2$

若零假设为真：

- $E\{(\mathrm{TSS} - \mathrm{RSS})/p\} = \sigma^2$

Quantity	Value
Residual standard error	1.69
R^2	0.897
F-statistic	570

至少有一个特征与响应有强关联

若X和Y无关联，则F~1
若备择假设为真，则F>1

F-统计量（F-statistic）

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	−0.001	0.0059	−0.18	0.8599

- 每个特征都有对应的t统计量和p值，提供了该特征是否与响应相关的证据。
- 每个t统计量等价于排除掉该特征后的F统计量，反映了该特征对模型的局部影响（partial effect）
- TV和radio都与销量强相关，但无证据表明报纸广告投入与销量相关。

问题2：识别重要的特征

- 若已经确定（如通过F统计量和p值）至少有一个特征与响应强相关，如何确定哪个特征相关性更强？
- 单个特征的p值（不太靠谱，尤其是特征数很大时）
- 通常响应与部分（而非全部）特征相关性更强
- ➔ 特征（变量）选取 （variable selection）
  - 遍历不同特征的组合，分别训练模型，选取最优者（特征数大时不现实，2^p个模型）
  - 需要某种自动、高效的方法选择重要特征

问题2：识别重要的特征—特征选取

三种经典方法：

- 向前选择（forward selection）
  - 由零假设开始，尝试不使用任何特征来解释数据
  - 训练p个简单线性回归模型，分别包含每个特征，将RSS最小者加入第一个模型
  - 将RSS第二小者加入前面模型，构建二元模型
  - 以此往复，直到停止条件满足
- 向后选择（backward selection）
- 混合选择（mixed selection）

问题2：识别重要的特征—特征选取

三种经典方法：

- 向前选择（forward selection）
- 向后选择（backward selection）
  - 构建包含所有特征的多元模型
  - 删除p值最大的特征，重新训练
  - 以此往复，直至停止条件
- 混合选择（mixed selection）

问题2：识别重要的特征—特征选取

三种经典方法：

- 向前选择（forward selection）
- 向后选择（backward selection）
- 混合选择（mixed selection）
  - 与向前选择类似，逐步增加变量
  - 考察新模型各特征的p值，若大于某一阈值，删除
  - 以此往复，直至停止条件

特征数量大于样本数时，向后选择无法使用，可用向前选择；
向前选择是一种贪心算法，可能将后期p值较大的特征纳入模型；
混合选择可修正以上问题

问题3：模型拟合质量

通常可用RSE和R^2来度量模型对数据的拟合质量

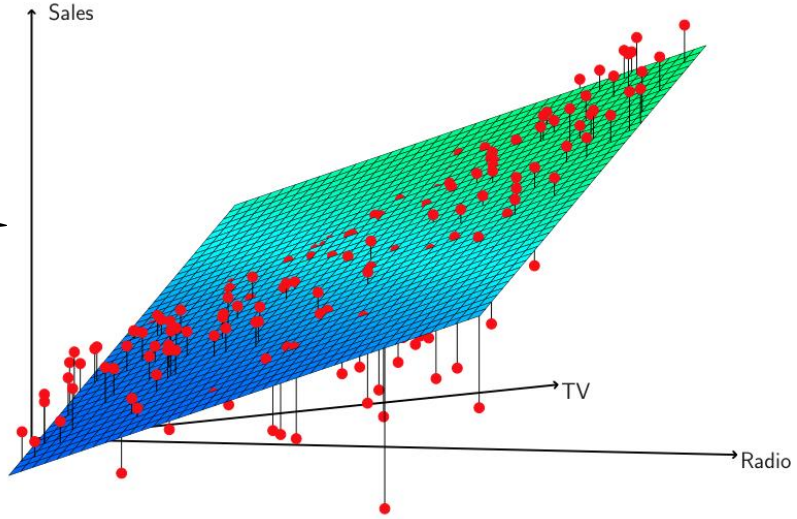
- R^2对应于拟合的响应值与真实响应值的相关系数$\mathrm{Cor}(Y,\hat{Y})^2$
- 使R^2最大是线性模型的特征
- R^2接近于1表明模型可解释响应变量的大部分偏差
- 例：Advertising数据
  - $R^2 = 0.8972$ 仅含TV：0.61
  - 仅包含TV和radio时，$R^2 = 0.89719$
  - ➔ 加入newspaper几无提升
  - newspaper无用的又一证据

Quantity	Value
Residual standard error	1.69
R^2	0.897
F-statistic	570

问题3：模型拟合质量

通过RSE考察拟合质量：

- 仅含TV： 3.26
- TV + radio：1.681
- TV + radio + newspaper：1.686
- ➔newspaper广告投入对销量几无效果



可视化提供直观趋势：

- 大部分预算投入某一类广告形式时，模型高估
- 预算分别投入两个渠道时，模型低估
- ➔ TV和radio广告之间可能有协同效应（互相促进）

问题4：预测精度

训练完毕的模型可用来对任一组特征进行预测，得到对应的响应值。该预测包含三类不确定性：

- 模型中的系数是对真实关联的估计，其中含有的不确定性属于可约误差范畴，可通过计算置信区间考察模型准确性
- 模型偏差（model bias）：将真实关联近似为线性模型导致的偏差
- 随机误差导致的不可约误差，可用预测区间（prediction intervals）量化
  - 如通过大量地区作为样本，估计销量均值，95%置信区间为[10985, 11528] (TV 10万；radio 2万)
  - 模型估计，95%预测区间 [7930, 14580]

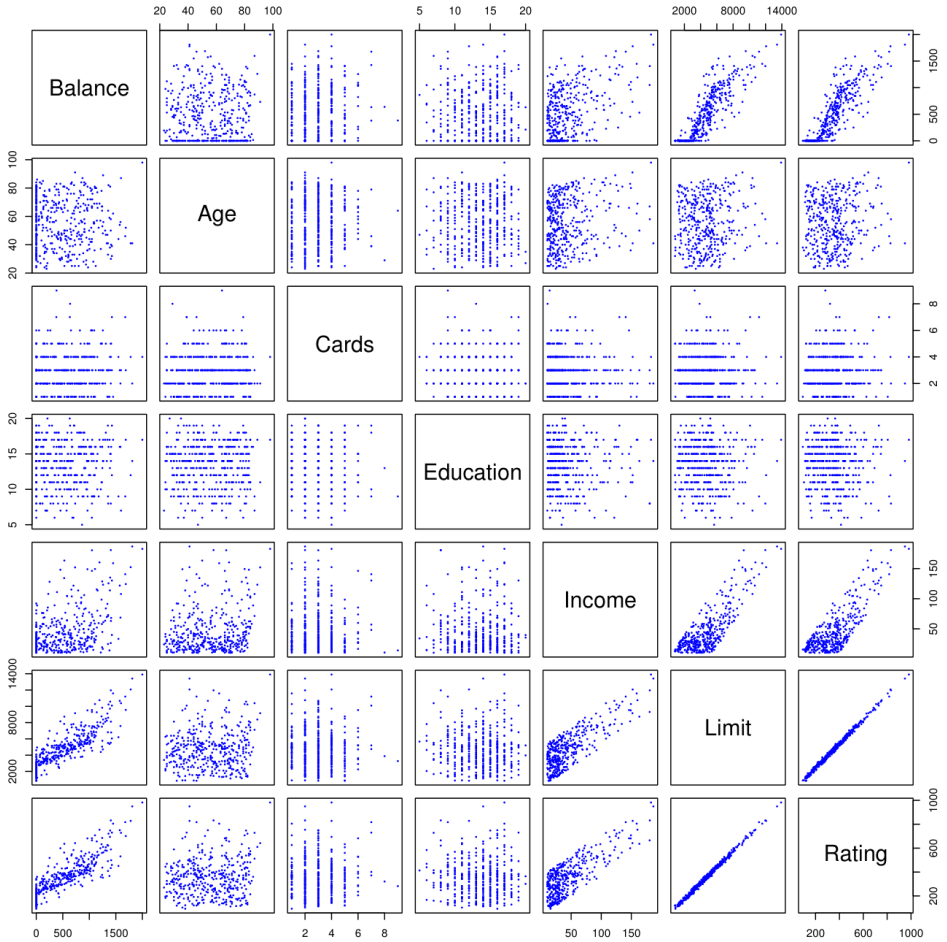
其它需要注意的问题

离散特征值：

- 以Credit为例
- 部分特征为离散值，如婚姻状态、性别、有无房子等

$$x_i = \begin{cases} 1 & \text{if } i\text{th person owns a house} \\ 0 & \text{if } i\text{th person does not own a house} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i \\ \beta_0 + \epsilon_i \end{cases}$$



其它需要注意的问题

多于两个可能取值的离散特征：

- 如种族，宗教信仰，区域等
- 不能用单个哑变量描述，可增加哑变量

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is from the South} \\ 0 & \text{if } i\text{th person is not from the South,} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is from the West} \\ 0 & \text{if } i\text{th person is not from the West.} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is from the South} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is from the West} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is from the East.} \end{cases}$$

其它需要注意的问题

扩展线性模型：线性回归中，假设特征和响应之间的关系是线性（linear）可加（additive）的：

- 线性：响应与特征的线性项相关
- 可加：总的响应变化可分解为单个特征的贡献；不同特征互相独立，其贡献加和
- 可能存在协同（synergy）或者交互（interaction）作用
- 如广告数据中，TV和radio的广告投入互相促进

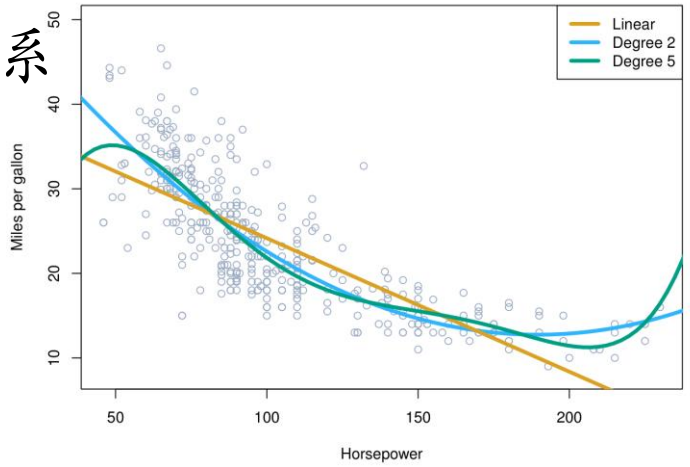
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

其它需要注意的问题

扩展线性模型：线性回归中，假设特征和响应之间的关系是线性（linear）可加（additive）的：

- 线性：响应与特征的线性项相关
- 可加：总的响应变化可分解为单个特征的贡献；不同特征互相独立，其贡献加和
- 响应和特征之间可能存在非线性关系

$$\texttt{mpg} = \beta_0 + \beta_1 \times \texttt{horsepower} + \beta_2 \times \texttt{horsepower}^2 + \epsilon$$



其它需要注意的问题

- 非线性的响应—特征关联
- 误差项之间存在相关
- 误差项偏差非恒定
- 离群点（outliers）
- 高杠杆点（high-leverage point）
- 共线性（collinearity）