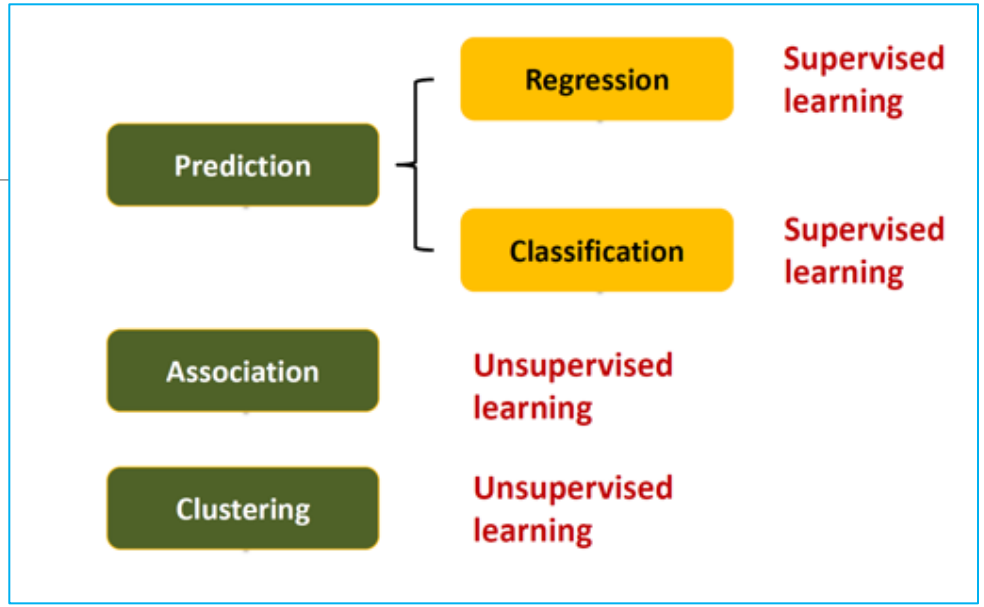
MM5425 数据分析

WEEK 4 LECTURE – LINEAR REGRESSION

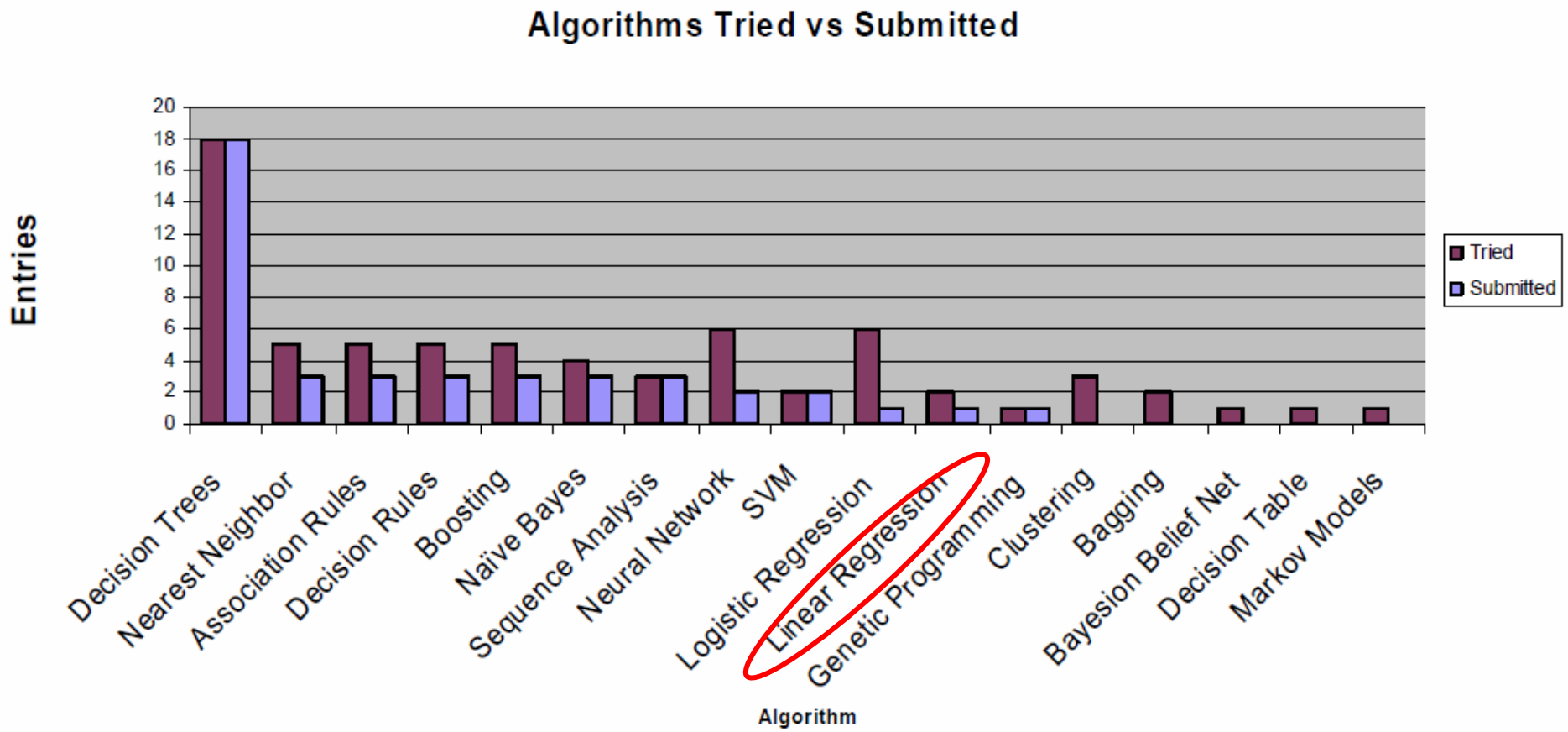
第四课 内容



回归分析 Regression

- 线性回归 Linear Regression
  - 斜率 slope
  - 截距 intercept
- 多变量线性回归 Multi-variate Linear Regression
- 线性回归分析 Linear Regression Evaluation

常用各类算法


Algorithms Tried vs Submitted
Entries
20
18
16
14
12
10
8
6
4
2
0
Tried
Submitted
Decision Trees
Nearest Neighbor
Association Rules
Decision Rules
Boosting
Naïve Bayes
Sequence Analysis
Neural Network
SVM
Logistic Regression
Linear Regression
Genetic Programming
Clustering
Bagging
Bayesion Belief Net
Decision Table
Markov Models
Algorithm

例子: 酒店大堂员工管理

大堂管理一直是许多企业关注的重点。银行、医疗机构，尤其是酒店，都希望通过提供优质服务，缩短顾客办理入住的等待时间，从而提升顾客体验。

为了确保提供优质的服务，优化大堂人员配置计划并准确预估客人入住情况将大有裨益.

设定目标变量 y = 每日签到，哪些属性可能影响 y？

什么是回归分析?

- 回归是一种统计技术，它将一个因变量（目标变量）与一个或多个独立（解释）变量联系起来.

- 回归分析的目的是找出哪些变量确实有影响。例如：

  哪些因素最重要？

  哪些因素可以忽略？等等.

- 一些回归类型：
  线性回归
  逻辑回归
  多项式回归
  逐步回归
  套索回归

更多商业实例

- 需求预测

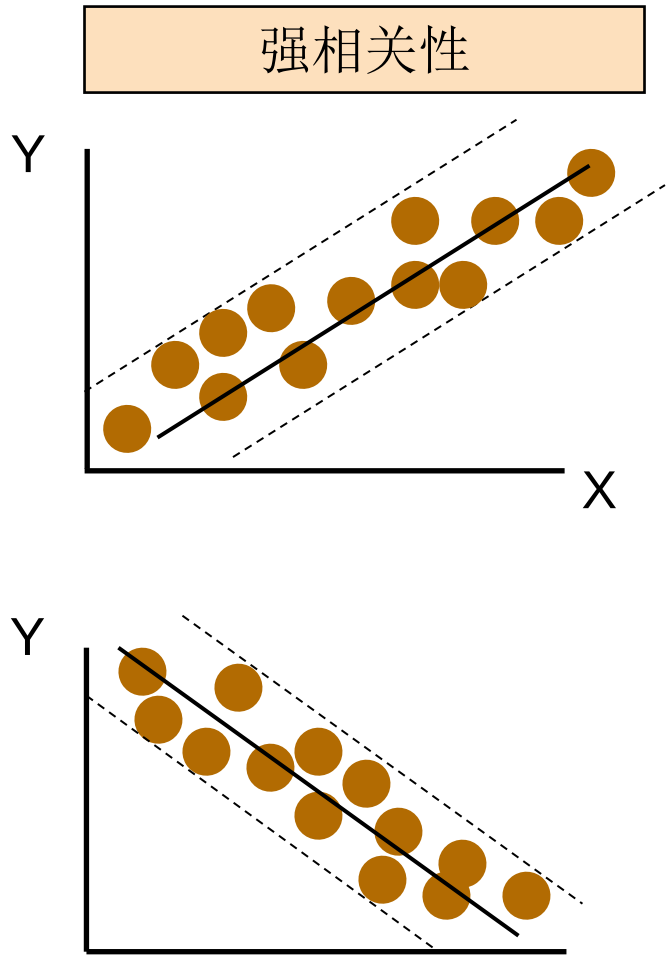
商店经理尝试预测新价格促销（自变量）下产品的需求（因变量）

- 评估出勤率对学业成绩的影响

学校/辅导中心想知道出勤率（自变量）是否有助于学生获得更高的成绩（因变量）

- 评估风险

线性相关



相关性

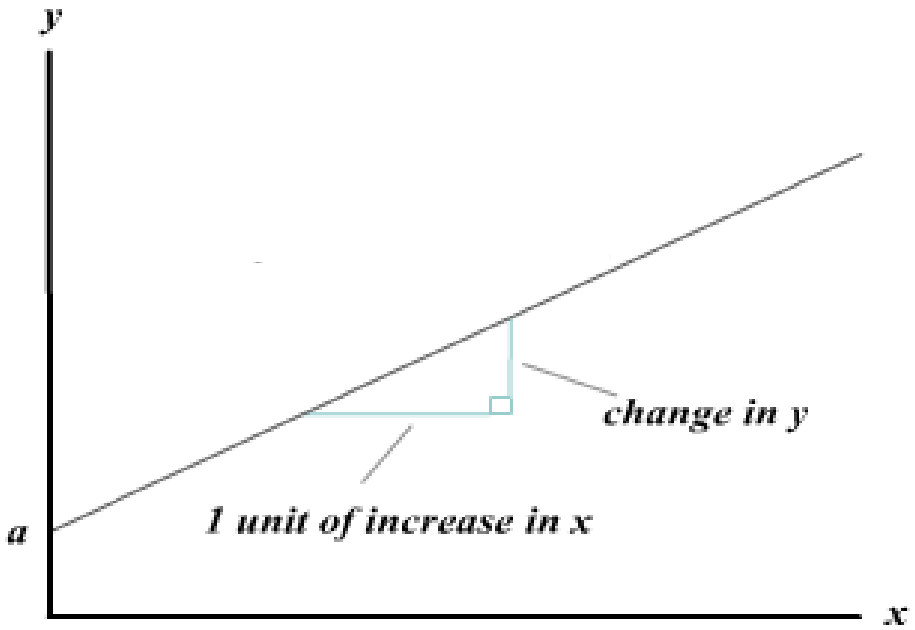
- 衡量两个变量之间线性关系的相对强度
- 取值范围在 –1 到 1 之间
- 越接近 –1，负线性关系越强
- 越接近 1，正线性关系越强
- 越接近 0，线性关系越弱

$$\operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\mathrm{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

线性回归

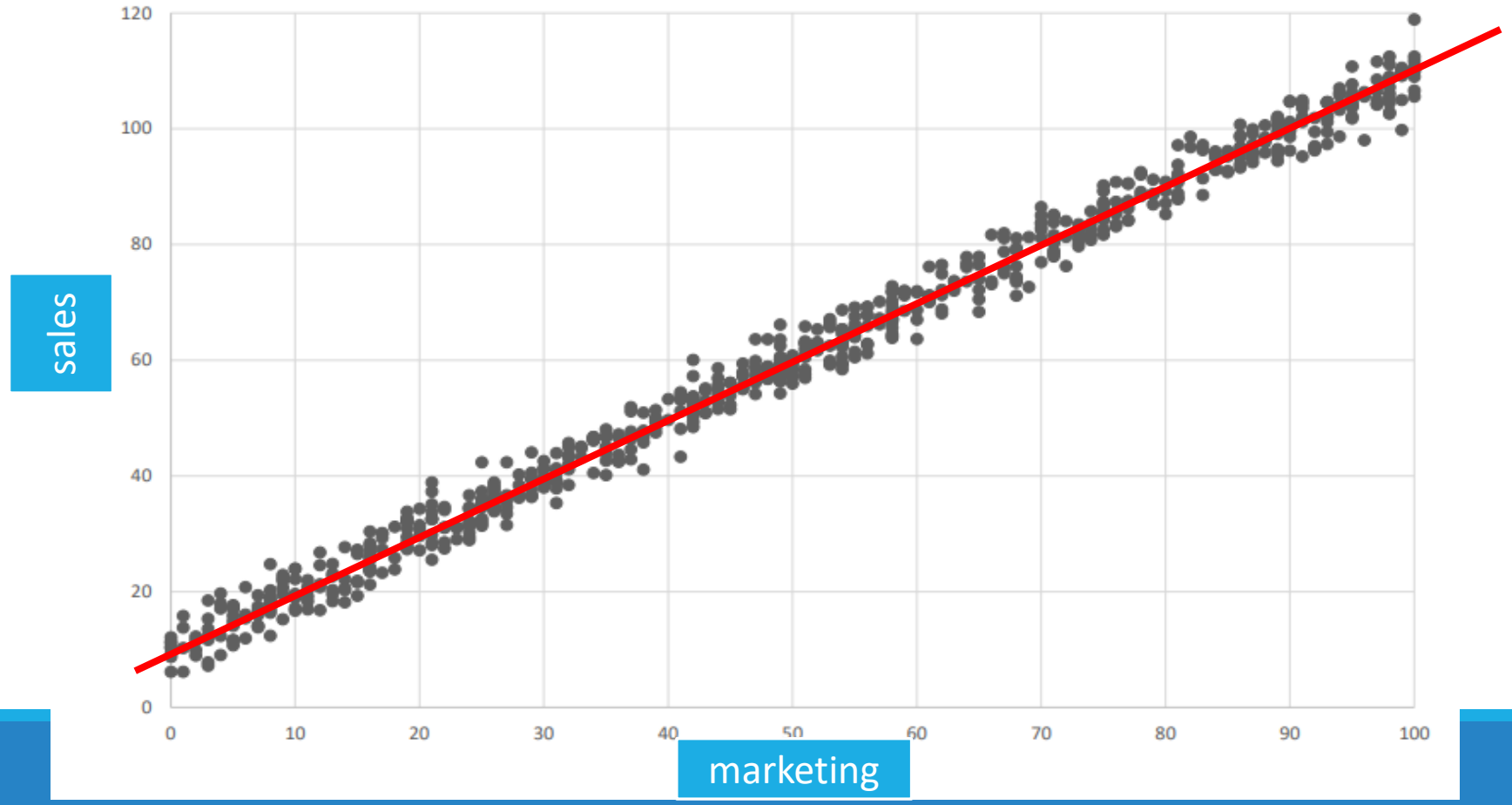
假设***Y***（因变量）和***X***（自变量）之间存在线性关系的线性模型

$$y = \alpha + \beta X$$



假设***Y***（因变量）和***X***（自变量）之间存在线性关系的线性模型

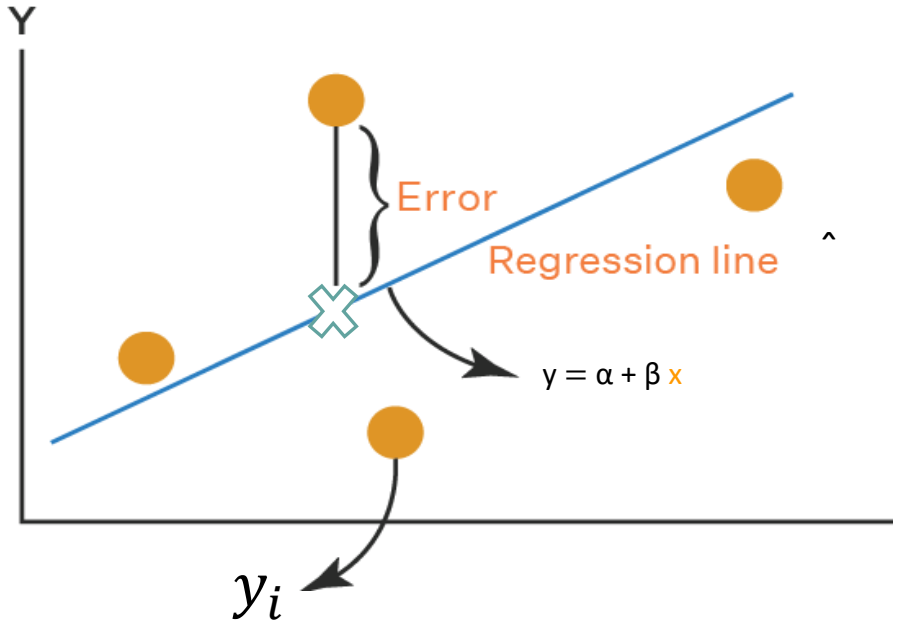
线性回归示例


sales
120
100
80
60
40
20
0
0
10
20
30
40
50
60
70
80
90
100
marketing

如何得到回归直线?

最小二乘法:

寻找 α 和 β，使得 $\sum_{i=1}^{n}(y_i - \hat{y})^2$ 最小



$$\hat{\beta} = \frac{Cov(x,y)}{Var(x)}$$

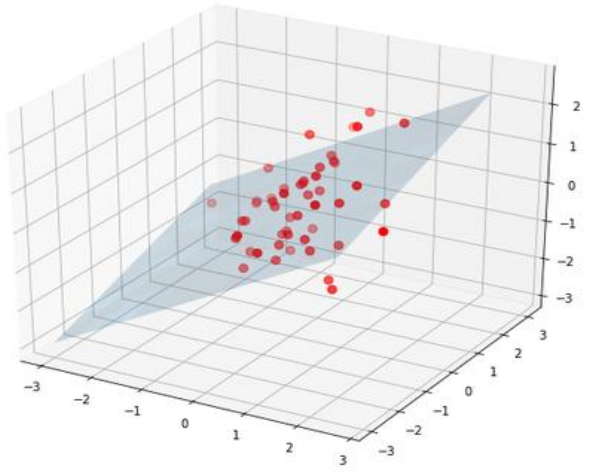
$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

多元线性回归

$$y = \alpha + \beta_1 * X_1 + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3 * X_3 + \ldots + \beta_n * X_n$$

例子:

$$\text{salary} = \alpha + \beta_1 * age + \beta_2 * experience + \beta_3 * gender + \beta_4 * level$$



线性回归评估

LINEARITY, R^2 , STANDARD ERROR

线性回归分析结果

OLS Regression Results

Dep. Variable:	y	R-squared:	0.767
Model:	OLS	Adj. R-squared:	0.708
Method:	Least Squares	F-statistic:	13.15
Date:	Fri, 01 Apr 2022	Prob (F-statistic):	0.00296
Time:	11:10:16	Log-Likelihood:	-31.191
No. Observations:	11	AIC:	68.38
Df Residuals:	8	BIC:	69.57
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	70.4828	3.749	18.803	0.000	61.839	79.127
x1	5.7945	1.132	5.120	0.001	3.185	8.404
x2	-1.1576	1.065	-1.087	0.309	-3.613	1.298

Omnibus:	0.198	Durbin-Watson:	1.240
Prob(Omnibus):	0.906	Jarque-Bera (JB):	0.296
Skew:	-0.242	Prob(JB):	0.862
Kurtosis:	2.359	Cond. No.	10.7

模型拟合度的假设检验

1. State the hypothesis
2. Set the criteria for a decision
3. Compute test statistic
4. Make a decision

H0: β = 0 ~ X 和 Y 之间没有关系
H1: β ≠ 0 ~ X 和 Y 之间存在关系

t检验 t test

计算p值班 Calculate p-value

p 值 < 0.05 的样本具有显著性

标准估计误差

误差平方和为

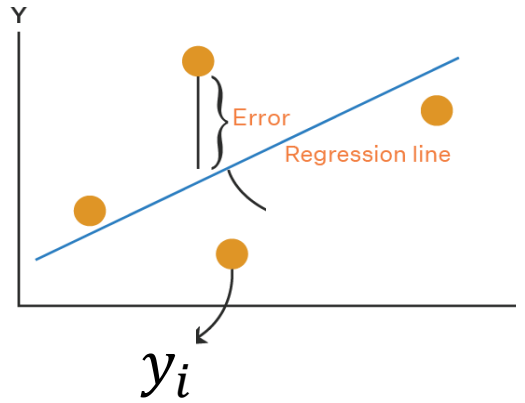
$$SS_E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

可以证明，误差平方和的期望值为 $E(SS_E) = (n-2)\sigma^2$.

回归模型的充分性

判定系数(R^2)

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$$



R2 这个量被称为判定系数，通常用来判断回归模型的充分性。

- $0 \leq R^2 \leq 1$;
- R^2 表示方差中可由自变量解释的部分有多大。方差越大，可解释性越高，回归模型越好。

调整后的 R^2

调整后的决定系数 (R^2)

$$\text{调整后 Adjusted-}R^2 = 1 - (1 - R^2)\frac{(N-1)}{N-k-1}$$

其中 k 是独立变量的数量

调整后的 R^2

- 75%: 非常好 very good
- 50-75%: 好 good
- 25-50%: 还好 fair
- Below 25%: 差 poor

例子 1: 糖果需求预测

- 业务目标：根据之前的销售数据预测糖果销量。
- 模型选择: 线性回归:

$$\#\text{销售数量} = \alpha + \beta * \text{糖果价格 Candy Price}$$

- 利用历史数据通过最小二乘法/最大似然估计得到α和β

检验结果

系数 Coefficients:

	Estimate	p-value
α	153,200.5	4.9e-8
β	-490.7	0.0012

如果价格上涨 1 美元，每周销售量将减少 490.7 箱

p-value:小于0.05，在95%的置信水平下拒绝H0（β=0），即价格与销量之间存在线性关系

特别季节?

万圣节!!

美通社公布的 2021 年万圣节糖果销售额:

- 零售额达 3.24 亿美元（较 2020 年同期增长 48%；较 2019 年同期增长 59.8%）。
- 每家门店万圣节特供商品销量较 2020 年同期增长 26.9%。

完善模型以考虑万圣节效应:

$$\#\text{销售数量} = \alpha + \beta_1 * \text{糖果价格} + \beta * \text{万圣节}$$

需求预测

我们进一步将万圣节变量分为三个：万圣节、万圣节前和万圣节后

将模型修改为多变量模型。

$$\#\text{销售数量} = \alpha + \beta_1 * \text{糖果价格} + \beta_2 * \text{万圣节} + \beta_3 * \text{万圣节前} + \beta_4 * \text{万圣节后}$$

检验结果

系数 Coefficients:

	Estimate	p-value
α	103,20.1	2e-9
β1: price	-190.7	0.09
β2: Halloween	21732.3	2.46e-8
β3: PreHalloween	7502	2e-12
β4: PostHalloween	-32105.2	4e-4

p-value of β1 为 0.09，大于0.05。糖果价格对于季节性销售不再显著

例子 2: 学习成绩表现预测

目标 Goal:预测学生的表现，并检查上课和读书是否有助于提高成绩

可分析数据 Data available: 历史学生数据，包括出勤记录、阅读书籍和考试成绩

Books read	Class Attended	Test Score
3	13	91
4	12	90
2	7	67
0	8	60
1	7	47
4	12	86
0	10	60
2	5	44

多变量模型

- 成绩取决于学生的出勤率和阅读的书籍数量

- 线性回归:

$$成绩 = a + \beta_1 * 上课出勤率 + \beta_2 * 读了几本书$$

分析结果

Coefficients:	Estimate	Pr(>|t|)
Intercept	37.379	0.000***
Attend	1.283	0.027*
Books	4.037	0.035*

Grades = 37.379 + 1.283 x Attend + 4.037 x Books

- 如果学生多上一门课，在其他变量保持不变的情况下，他的成绩会提高1.283分。
- 在95%的置信水平下，上课次数和阅读的书籍数量会影响最终的考试成绩