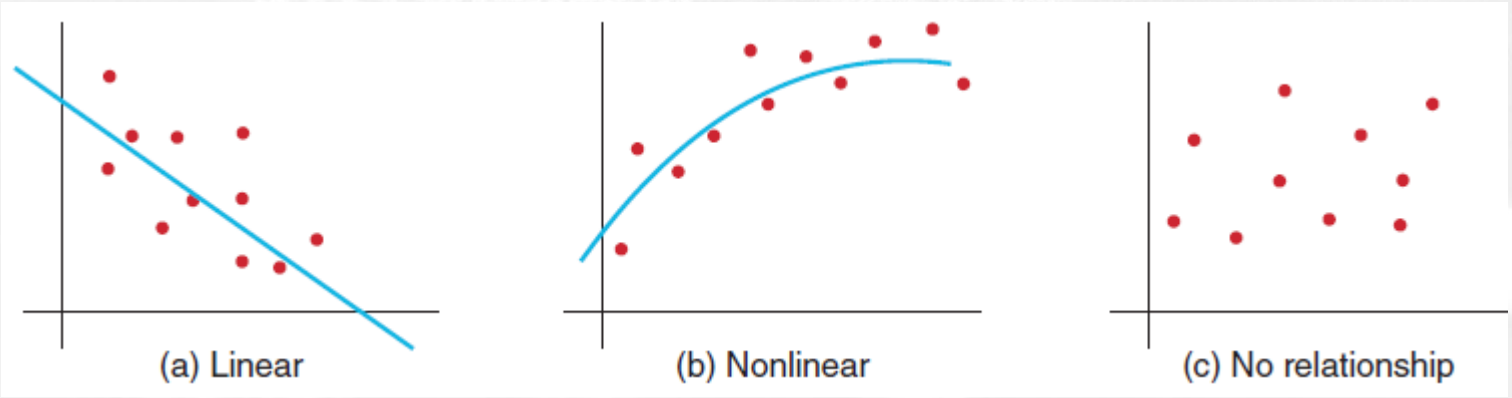
基于特征的预测模型：回归分析

- **回归分析**：一种建立一个因变量（被解释变量，Y）和一个或若干个自变量（解释变量，X）关系的统计模型
- **简单线性回归：只有一个自变量**
- **多元线性回归：有多个自变量**

简单线性回归

- 建立以下变量的关系：
  - 一个自变量X
  - 一个自变量Y
- 首先绘制X和Y的散点图，确认数据存在线性关系
  - 如果数据明显不存在线性关系，应当用其他工具建立变量之间的关系



例5.4: *Home Market Value*

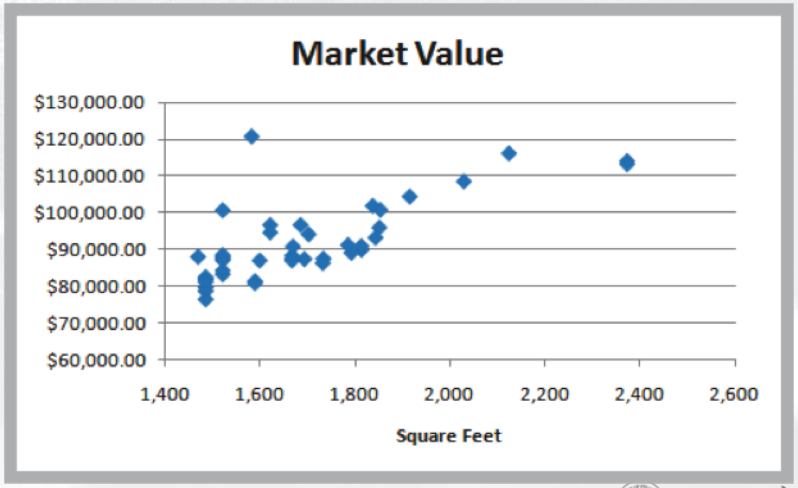
房屋面积与房屋市场价相关:

X = 房屋面积

Y = 市场价 ($)

42 个房子的散点图显示线性趋势

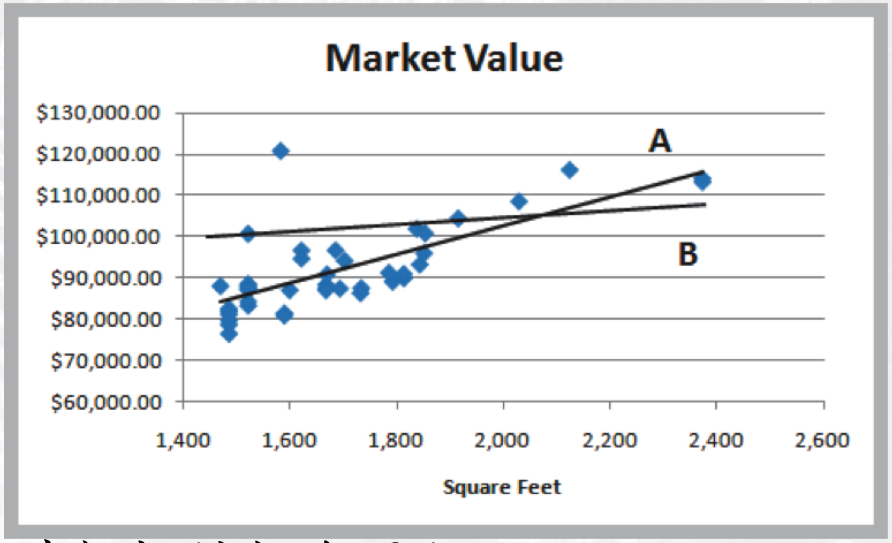
	A	B	C
1	Home Market Value		
2			
3	House Age	Square Feet	Market Value
4	33	1,812	$90,000.00
5	32	1,914	$104,400.00
6	32	1,842	$93,300.00
7	33	1,812	$91,000.00
8	32	1,836	$101,900.00
9	33	2,028	$108,500.00
10	32	1,732	$87,600.00



找到最优的拟合直线

$$Market\ value = a + b \times square\ feet$$

- 两条可能的拟合线



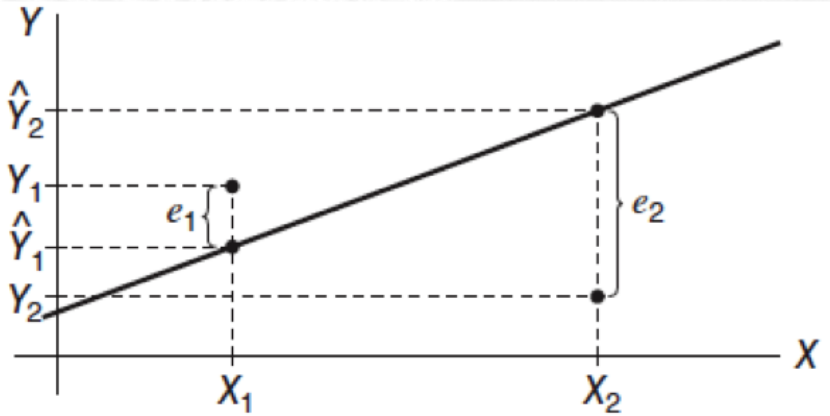
  - A线比B线对数据的拟合更好
  - 我们希望找到最优的拟合线

最小二乘法

- 简单线性回归模型：
- $Y = \beta_0 + \beta_1 X + \varepsilon$
  - 通过对样本数据的估计得到参数的估计值：
  - 真实的β_0和β_1不知道，基于样本数据估计
  - $\hat{Y} = b_0 + b_1 X$

残差

- 残差：真实值与根据拟合线的估计值之差：
- $e_i = Y_i - \hat{Y}_i$



Errors associated with individual observations

最小二乘法

- 最优的拟合线是最小化所有残差平方和的拟合线

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - [b_0 + b_1 X_i])^2$$

- Excel 函数:
  - b_0=INTERCEPT(*known_y's, known_x's*)
  - b_1 =SLOPE(*known_y's, known_x's*)

例5.4：利用Excel 函数估计参数

- 斜率= b_1 = 35.036

  =SLOPE(C4:C45, B4:B45)

- 截距 = b_0 = 32,673

  =INTERCEPT(C4:C45, B4:B45)

- 当 X = 1750 时估计Y

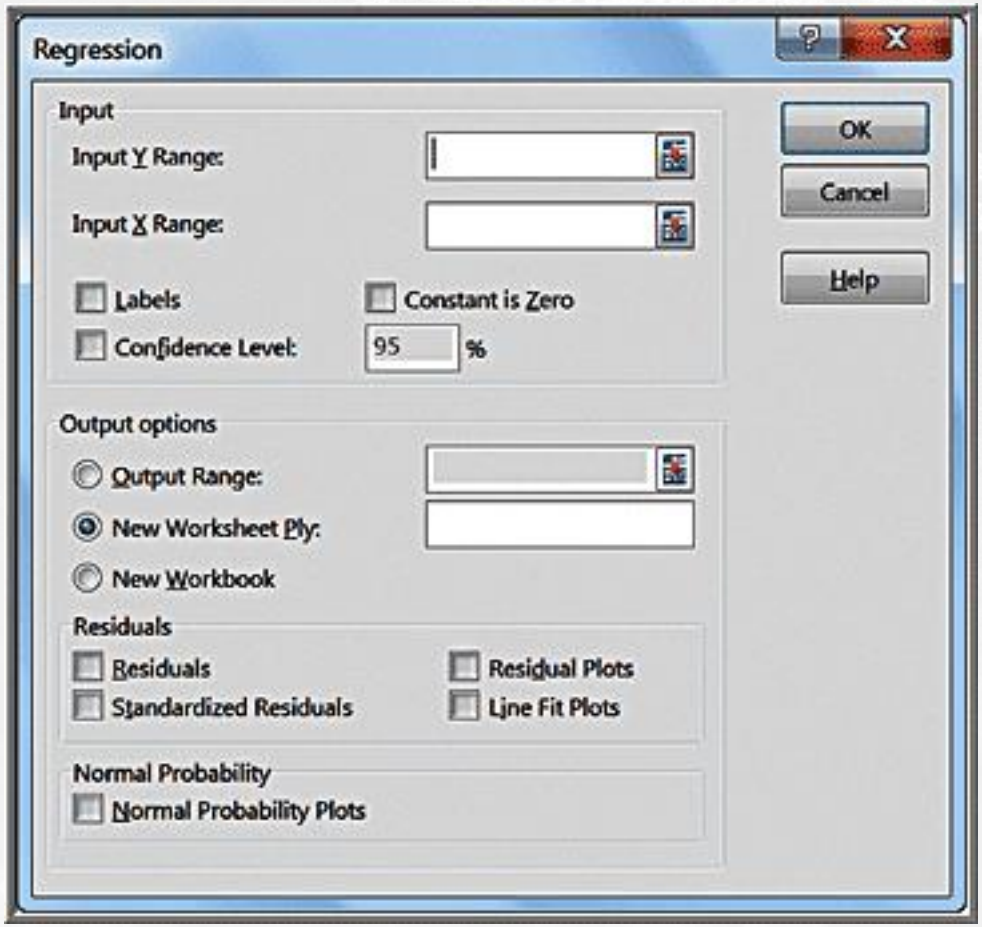
  $\hat{Y}$ = 32,673 + 35.036(1750) = $93,986

  =TREND(C4:C45, B4:B45, 1750)

	A	B	C
1	Home Market Value		
2			
3	House Age	Square Feet	Market Value
4	33	1,812	$90,000.00
5	32	1,914	$104,400.00
6	32	1,842	$93,300.00
7	33	1,812	$91,000.00
8	32	1,836	$101,900.00
9	33	2,028	$108,500.00
10	32	1,732	$87,600.00

数据分析中的回归

数据> 数据分析>回归



Home Market Value 回归结果

SUMMARY OUTPUT

回归统计	
Multiple R	0.731255
R Square	0.534734
Adjusted R Square	0.523103
标准误差	7287.723
观测值	42

方差分析

	df	SS	MS	F	Significance F
回归分析	1	2.44E+09	2.44E+09	45.97236	3.8E-08
残差	40	2.12E+09	53110902		
总计	41	4.57E+09			

	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	32673.22	8831.951	3.699434	0.00065	14823.18	50523.26	14823.18	50523.26
Square Feet	35.03637	5.167384	6.780292	3.8E-08	24.5927	45.48004	24.5927	45.48004

回归结果解读

- **Multiple R** - $|r|$，相关系数
- **R Square**，R^2，拟合优度
- **P值**
  - $H_0: \beta_1 = 1$,房屋面积对市场价的影响不显著
  - $H_1: \beta_1 \neq 1$
  - 房屋面积对市场价的影响显著

多元线性回归

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots\ldots + \beta_k X_k + \varepsilon$
- 例5.5 预测学校的毕业率

	A	B	C	D	E	F	G
1	Colleges and Universities						
2							
3	School	Type	Median SAT	Acceptance Rate	Expenditures/Student	Top 10% HS	Graduation %
4	Amherst	Lib Arts	1315	22%	$ 26,636	85	93
5	Barnard	Lib Arts	1220	53%	$ 17,653	69	80
6	Bates	Lib Arts	1240	36%	$ 17,554	58	88
7	Berkeley	University	1176	37%	$ 23,665	95	68
8	Bowdoin	Lib Arts	1300	24%	$ 25,703	78	90
9	Brown	University	1281	24%	$ 24,201	80	90

多元线性回归

SUMMARY OUTPUT

回归统计	
Multiple R	0.731044
R Square	0.534426
Adjusted R	0.492101
标准误差	5.308338
观测值	49

方差分析

	df	SS	MS	F	gnificance F
回归分析	4	1423.209	355.8023	12.62675	6.33E-07
残差	44	1239.852	28.17845		
总计	48	2663.061			

	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	17.92096	24.55722	0.729763	0.469402	-31.5709	67.41279	-31.5709	67.41279
Median SA	0.072006	0.017984	4.003927	0.000236	0.035762	0.10825	0.035762	0.10825
Acceptanc	-24.8592	8.315185	-2.98962	0.00456	-41.6174	-8.10108	-41.6174	-8.10108
Expenditur	-0.00014	6.59E-05	-2.05744	0.0456	-0.00027	-2.8E-06	-0.00027	-2.8E-06
Top 10% H	-0.16276	0.079345	-2.05136	0.046214	-0.32267	-0.00286	-0.32267	-0.00286

$$\text{Graduation\%} = 17.92 + 0.072\,\text{SAT} - 24.859\,\text{ACCEPTANCE} - 0.000136\,\text{EXPENDITURES} - 0.163\,\text{TOP10\% HS}$$

自变量中的名义变量

- 回归分析要求自变量为数量型变量
- 名义变量要编码为虚拟变量（亚变量、零一变量）
- 例5.6 *Employee Salaries* 通过年龄和是否有MBA学历预测工资
  - MBA：Yes=1,No=0
  - IF(D4=“Yes”,1,0)

	A	B	C	D
1	**Employee Salary Data**			
2				
3	**Employee**	**Salary**	**Age**	**MBA**
4	1	$ 28,260	25	No
5	2	$ 43,392	28	Yes
6	3	$ 56,322	37	Yes
7	4	$ 26,086	23	No
8	5	$ 36,807	32	No

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

where

Y = salary

X_1 = age

X_2 = MBA indicator (0 or 1)

自变量中的名义变量

- Salary = 893.59 + 1044.15 × Age + 14767.23 × MBA
  - If MBA = 0, salary = 893.59 + 1044 × Age
  - If MBA = 1, salary =15,660.82 + 1044 × Age

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.976118476					
5	R Square	0.952807278					
6	Adjusted R Square	0.949857733					
7	Standard Error	2941.914352					
8	Observations	35					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	2	591651177	2795825589	323.0353318	6.05341E-22	
13	Residual	32	276955521.7	8654860.054			
14	Total	34	5868606699				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	893.5875971	1824.575283	0.489751015	0.627650922	-2822.950634	4610.125828
18	Age	1044.146043	42.14128238	24.77727265	1.8878E-22	958.3070599	1129.985026
19	MBA	14767.23159	1351.801764	10.92411031	2.49752E-12	12013.7015	17520.76168

思考题

1. Colleges and Universities 把大学类型添加为变量，预测毕业率？
2. 如果一个名义变量有3种观测值，需要添加几个虚拟变量？