第五章 预测方法

预测方法

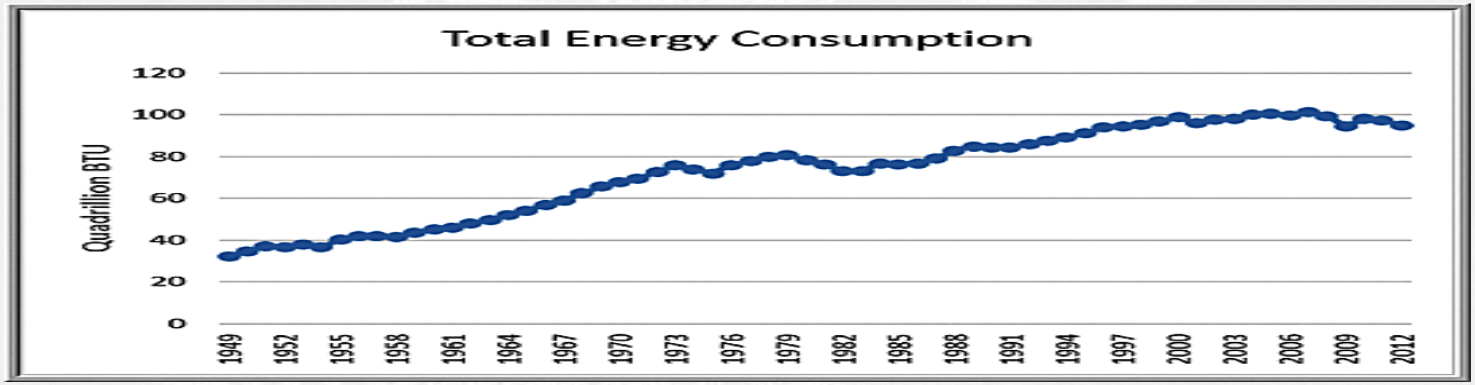
- 经理人需要基于对未来的预测做出决策
- 三大类预测方法：
  - 主观预测
  - 基于时间序列的预测模型
  - 基于特征的预测模型

时间序列预测模型

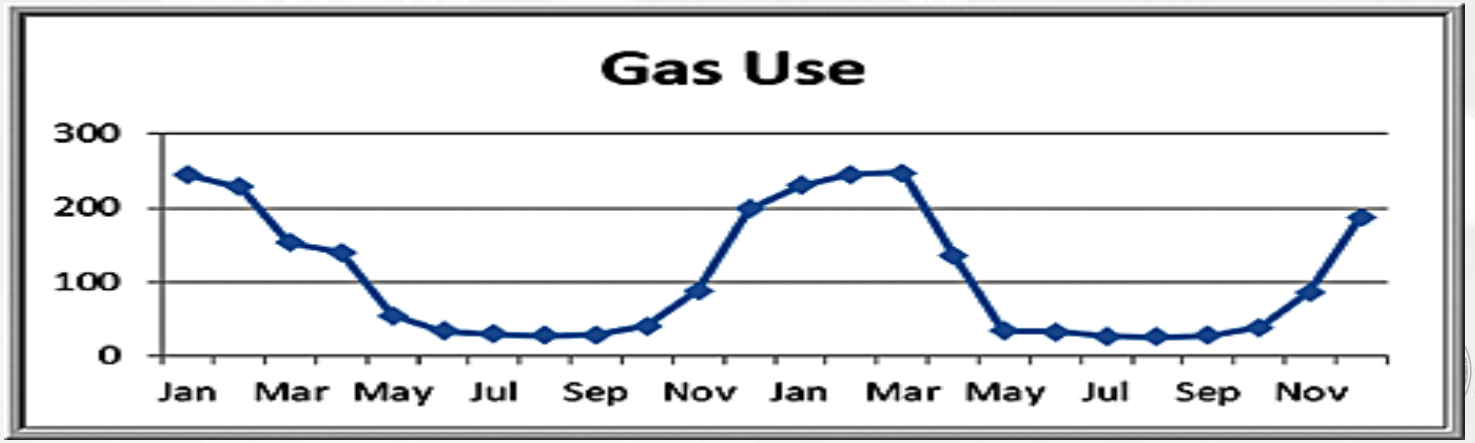
- 时间序列 - 一系列历史数据，例如每周的销售额
  - $t = 1, 2, \ldots, T$
- 时间序列中可能包括趋势性、季节性
- 平稳时间序列：没有趋势性、季节性，只有随机波动

时间序列数据

- 趋势性



- 季节性



平稳时间序列预测

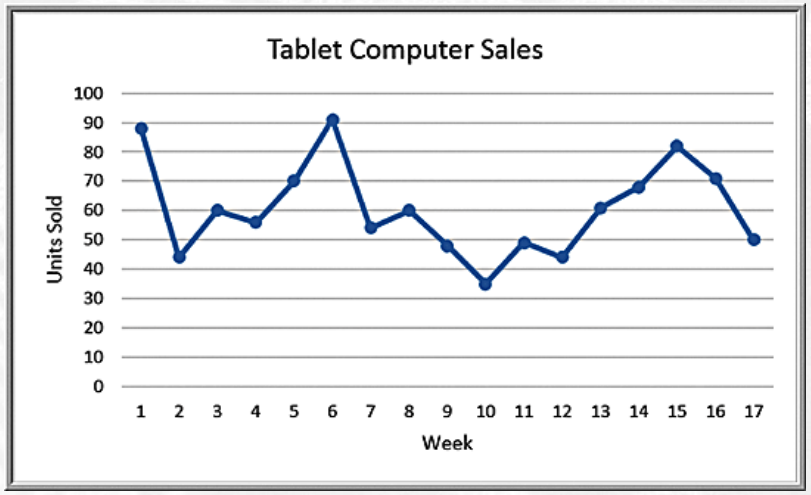
- 移动平均法
- 指数平滑法
  - 时间序列没有明显的趋势和季节性
  - 短期预测

移动平均法

- 移动平均法：对下一期的预测等于最近k期观测值的平均
  - k越大，预测越“平滑”，受极端值影响越小
  - 移动平均法的思想：通过多期的加总，把随机波动的影响去除，得到数据真正的水平

例5.1：移动平均法预测

- *Tablet Computer Sales* 数据集展示了过去17周的平板电脑销量.



- 预测第18周的销量，k=3:

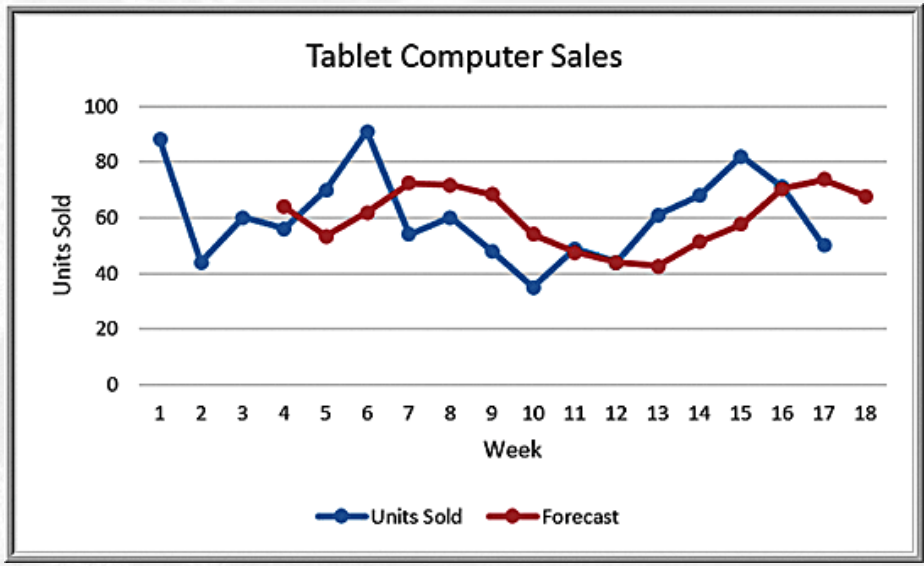
$$(82 + 71 + 50)/3 = 67.67$$

例5.1 续

	A	B	C	D	E	F
1	Tablet Computer Sales					
2			Moving Average			
3	Week	Units Sold	Forecast			
4	1	88				
5	2	44				
6	3	60				
7	4	56	64.00			
8	5	70	53.33			
9	6	91	62.00			
10	7	54	72.33			
11	8	60	71.67			
12	9	48	68.33			
13	10	35	54.00			
14	11	49	47.67			
15	12	44	44.00			
16	13	61	42.67			
17	14	68	51.33			
18	15	82	57.67			
19	16	71	70.33			
20	17	50	73.67			
21	18		67.67			
22						

Forecast for week 4
=AVERAGE(B4:B6)

Forecast for week 18
=AVERAGE(B18:B20)



尝试$k = 5, 10$

预测准确性衡量指标

- Mean absolute deviation (MAD)（平均绝对偏差）

$$MAD = \frac{\sum_{t=1}^{n} |A_t - F_t|}{n}$$

- Mean square error (MSE)（均方误差）

$$MSE = \frac{\sum_{t=1}^{n} (A_t - F_t)^2}{n}$$

- Root mean square error (RMSE)（均方误差根）

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n} (A_t - F_t)^2}{n}}$$

- Mean absolute percentage error (MAPE)（平均绝对百分比误差）

$$MAPE = \frac{\sum_{i=1}^{n} |\frac{A_t - F_t}{A_t}|}{n} \times 100\%$$

例5.2：比较不同预测模型预测准确度

- *Tablet Computer Sales* 数据集
- K=2，3，4
- K=2 预测结果最准确

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Tablet Computer Sales																
2			k = 2					k = 3					k = 4				
3	Week	Units Sold	Forecast	Error	Absolute Deviation	Squared Error	Absolute % Error	Forecast	Error	Absolute Deviation	Squared Error	Absolute % Error	Forecast	Error	Absolute Deviation	Squared Error	Absolute % Error
4	1	88															
5	2	44															
6	3	60	66.00	-6.00	6.00	36.00	10.00										
7	4	56	52.00	4.00	4.00	16.00	7.14	64.00	-8.00	8.00	64.00	14.29					
8	5	70	58.00	12.00	12.00	144.00	17.14	53.33	16.67	16.67	277.78	23.81	62.00	8.00	8.00	64.00	11.43
9	6	91	63.00	28.00	28.00	784.00	30.77	62.00	29.00	29.00	841.00	31.87	57.50	33.50	33.50	1122.25	36.81
10	7	54	80.50	-26.50	26.50	702.25	49.07	72.33	-18.33	18.33	336.11	33.95	69.25	-15.25	15.25	232.56	28.24
11	8	60	72.50	-12.50	12.50	156.25	20.83	71.67	-11.67	11.67	136.11	19.44	67.75	-7.75	7.75	60.06	12.92
12	9	48	57.00	-9.00	9.00	81.00	18.75	68.33	-20.33	20.33	413.44	42.36	68.75	-20.75	20.75	430.56	43.23
13	10	35	54.00	-19.00	19.00	361.00	54.29	54.00	-19.00	19.00	361.00	54.29	63.25	-28.25	28.25	798.06	80.71
14	11	49	41.50	7.50	7.50	56.25	15.31	47.67	1.33	1.33	1.78	2.72	49.25	-0.25	0.25	0.06	0.51
15	12	44	42.00	2.00	2.00	4.00	4.55	44.00	0.00	0.00	0.00	0.00	48.00	-4.00	4.00	16.00	9.09
16	13	61	46.50	14.50	14.50	210.25	23.77	42.67	18.33	18.33	336.11	30.05	44.00	17.00	17.00	289.00	27.87
17	14	68	52.50	15.50	15.50	240.25	22.79	51.33	16.67	16.67	277.78	24.51	47.25	20.75	20.75	430.56	30.51
18	15	82	64.50	17.50	17.50	306.25	21.34	57.67	24.33	24.33	592.11	29.67	55.50	26.50	26.50	702.25	32.32
19	16	71	75.00	-4.00	4.00	16.00	5.63	70.33	0.67	0.67	0.44	0.94	63.75	7.25	7.25	52.56	10.21
20	17	50	76.50	-26.50	26.50	702.25	53.00	73.67	-23.67	23.67	560.11	47.33	70.50	-20.50	20.50	420.25	41.00
21	18		60.50		13.63	254.38	23.63	67.67		14.86	299.84	25.37	67.75		16.13	355.25	28.07
22					MAD	MSE	MAPE			MAD	MSE	MAPE			MAD	MSE	MAPE

指数平滑法

- **简单指数平滑模型：**
- $F_{t+1} = (1-\alpha)F_t + \alpha A_t = F_t + \alpha(A_t - F_t)$
  - $0 \leq \alpha \leq 1$ 平滑系数
  - $F_1 = A_1$

例5.3 指数平滑法

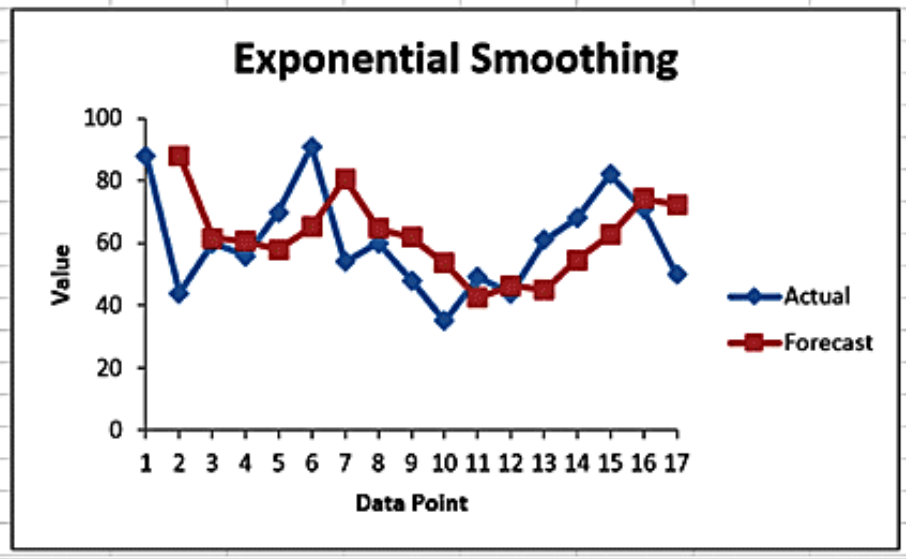
	A	B	C	D	E	F	G	H	I	J	K
1	Tablet Computer Sales										
2			Smoothing Constant								
3	Week	Units Sold	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
4	1	88	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00
5	2	44	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00
6	3	60	83.60	79.20	74.80	70.40	66.00	61.60	57.20	52.80	48.40
7	4	56	81.24	75.36	70.36	66.24	63.00	60.64	59.16	58.56	58.84
8	5	70	78.72	71.49	66.05	62.14	59.50	57.86	56.95	56.51	56.28
9	6	91	77.84	71.19	67.24	65.29	64.75	65.14	66.08	67.30	68.63
10	7	54	79.16	75.15	74.37	75.57	77.88	80.66	83.53	86.26	88.76
11	8	60	76.64	70.92	68.26	66.94	65.94	64.66	62.86	60.45	57.48
12	9	48	74.98	68.74	65.78	64.17	62.97	61.87	60.86	60.09	59.75
13	10	35	72.28	64.59	60.45	57.70	55.48	53.55	51.86	50.42	49.17
14	11	49	68.55	58.67	52.81	48.62	45.24	42.42	40.06	38.08	36.42
15	12	44	66.60	56.74	51.67	48.77	47.12	46.37	46.32	46.82	47.74
16	13	61	64.34	54.19	49.37	46.86	45.56	44.95	44.70	44.56	44.37
17	14	68	64.00	55.55	52.86	52.52	53.28	54.58	56.11	57.71	59.34
18	15	82	64.40	58.04	57.40	58.71	60.64	62.63	64.43	65.94	67.13
19	16	71	66.16	62.83	64.78	68.03	71.32	74.25	76.73	78.79	80.51
20	17	50	66.65	64.47	66.65	69.22	71.16	72.30	72.72	72.56	71.95
21	18		64.98	61.57	61.65	61.53	60.58	58.92	56.82	54.51	52.20
22		MAD	19.33	17.16	16.15	15.36	14.93	14.71	14.72	14.88	15.36
23		MSE	496.07	390.84	359.18	346.56	340.77	338.41	339.03	343.32	352.38
24		MAPE	38.28%	32.71%	30.12%	28.36%	27.54%	27.09%	27.09%	27.38%	28.23%

例5.3续 确定最优的平滑系数

	A	B	C	D	E	F	G	H	I	J	K
1	Tablet Computer Sales										
2			Smoothing Constant								
3	Week	Units Sold	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.9(
4	1	88	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.0(
5	2	44	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.0(
6	3	60	83.60	79.20	74.80	70.40	66.00	61.60	57.20	52.80	48.4(
7	4	56	81.24	75.36	70.36	66.24	63.00	60.64	59.16	58.56	58.84
8	5	70	78.72	71.49	66.05	62.14	59.50	57.86	56.95	56.51	56.2E
9	6	91	77.84	71.19	67.24	65.29	64.75	65.14	66.08	67.30	68.63
10	7	54	79.16	75.15	74.37	75.57	77.88	80.66	83.53	86.26	88.7E
11	8	60	76.64	70.92	68.26	66.94	65.94	64.66	62.86	60.45	57.4E
12	9	48	74.98	68.74	65.78	64.17	62.97	61.87	60.86	60.09	59.75
13	10	35	72.28	64.59	60.45	57.70	55.48	53.55	51.86	50.42	49.17
14	11	49	68.55	58.67	52.81	48.62	45.24	42.42	40.06	38.08	36.42
15	12	44	66.60	56.74	51.67	48.77	47.12	46.37	46.32	46.82	47.74
16	13	61	64.34	54.19	49.37	46.86	45.56	44.95	44.70	44.56	44.37
17	14	68	64.00	55.55	52.86	52.52	53.28	54.58	56.11	57.71	59.34
18	15	82	64.40	58.04	57.40	58.71	60.64	62.63	64.43	65.94	67.13
19	16	71	66.16	62.83	64.78	68.03	71.32	74.25	76.73	78.79	80.51
20	17	50	66.65	64.47	66.65	69.22	71.16	72.30	72.72	72.56	71.95
21	18		64.98	61.57	61.65	61.53	60.58	58.92	56.82	54.51	52.2(
22		MAD	19.33	17.16	16.15	15.36	14.93	14.71	14.72	14.88	15.3E
23		MSE	496.07	390.84	359.18	346.56	340.77	338.41	339.03	343.32	352.3E
24		MAPE	38.28%	32.71%	30.12%	28.36%	27.54%	27.09%	27.09%	27.38%	28.23%

例5.3 续：指数平滑法

	A	B	C
1	Tablet Computer Sales		
2			Exponential Smoothing
3	Week	Units Sold	Forecast
4	1	88	#N/A
5	2	44	88.00
6	3	60	61.60
7	4	56	60.64
8	5	70	57.86
9	6	91	65.14
10	7	54	80.66
11	8	60	64.66
12	9	48	61.87
13	10	35	53.55
14	11	49	42.42
15	12	44	46.37
16	13	61	44.95
17	14	68	54.58
18	15	82	62.63
19	16	71	74.25
20	17	50	72.30

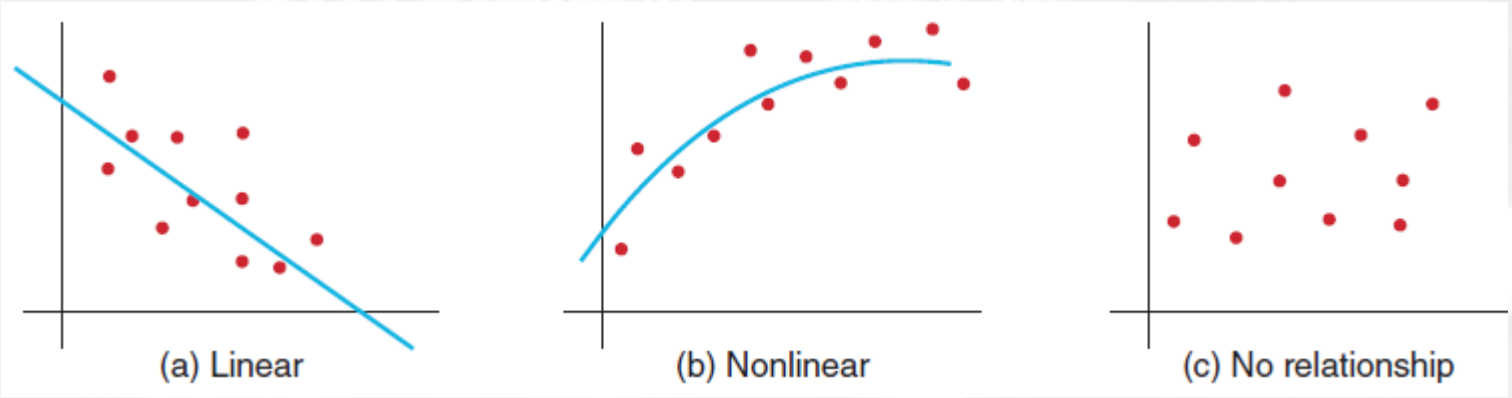


基于特征的预测模型：回归分析

- **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*

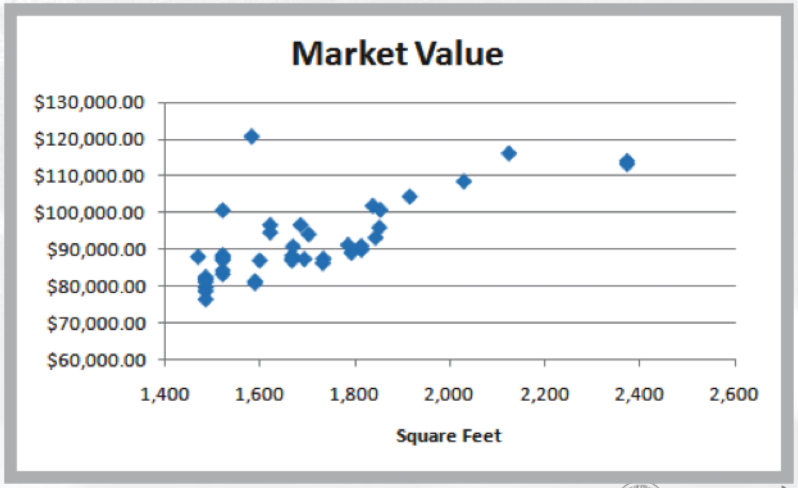
房屋面积与房屋市场价相关:

X = 房屋面积

Y = 市场价 ($)

42 个房子的散点图显示线性趋势

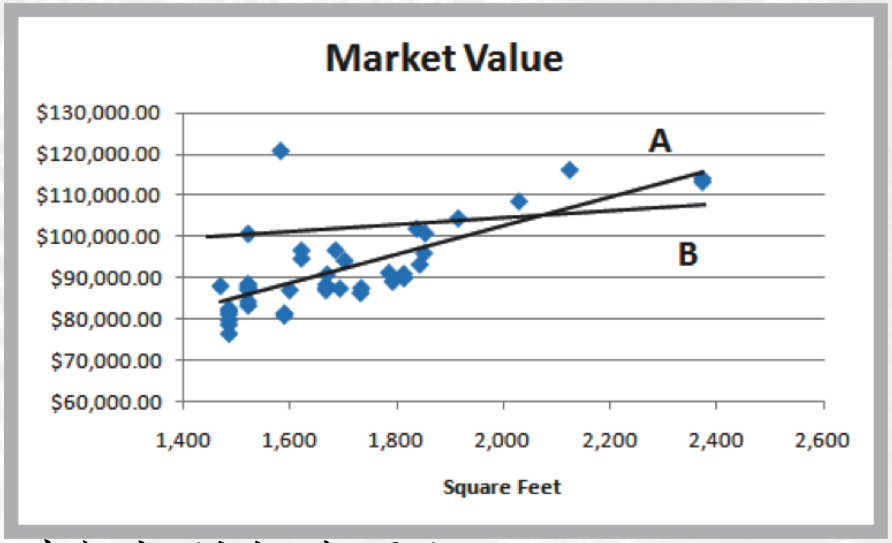
	A	B	C
1	Home Market Value		
2			
3	House Age	Square Feet	Market Value
4	33	1,812	$90,000.00
5	32	1,914	$104,400.00
6	32	1,842	$93,300.00
7	33	1,812	$91,000.00
8	32	1,836	$101,900.00
9	33	2,028	$108,500.00
10	32	1,732	$87,600.00



找到最优的拟合直线

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

- 两条可能的拟合线



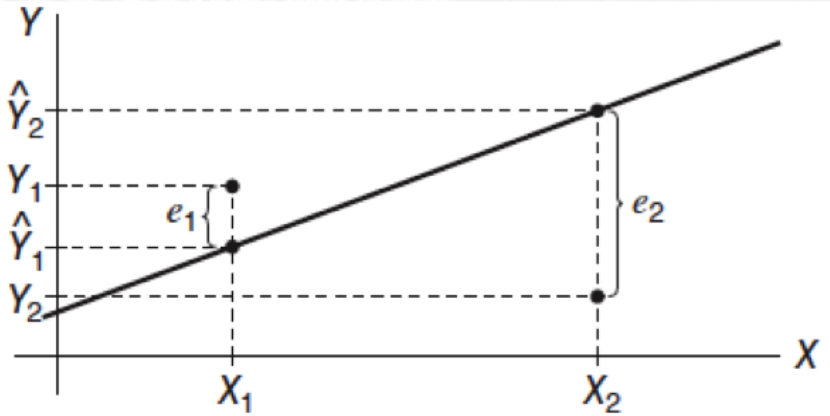
  - A线比B线对数据的拟合更好
  - 我们希望找到最优的拟合线

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

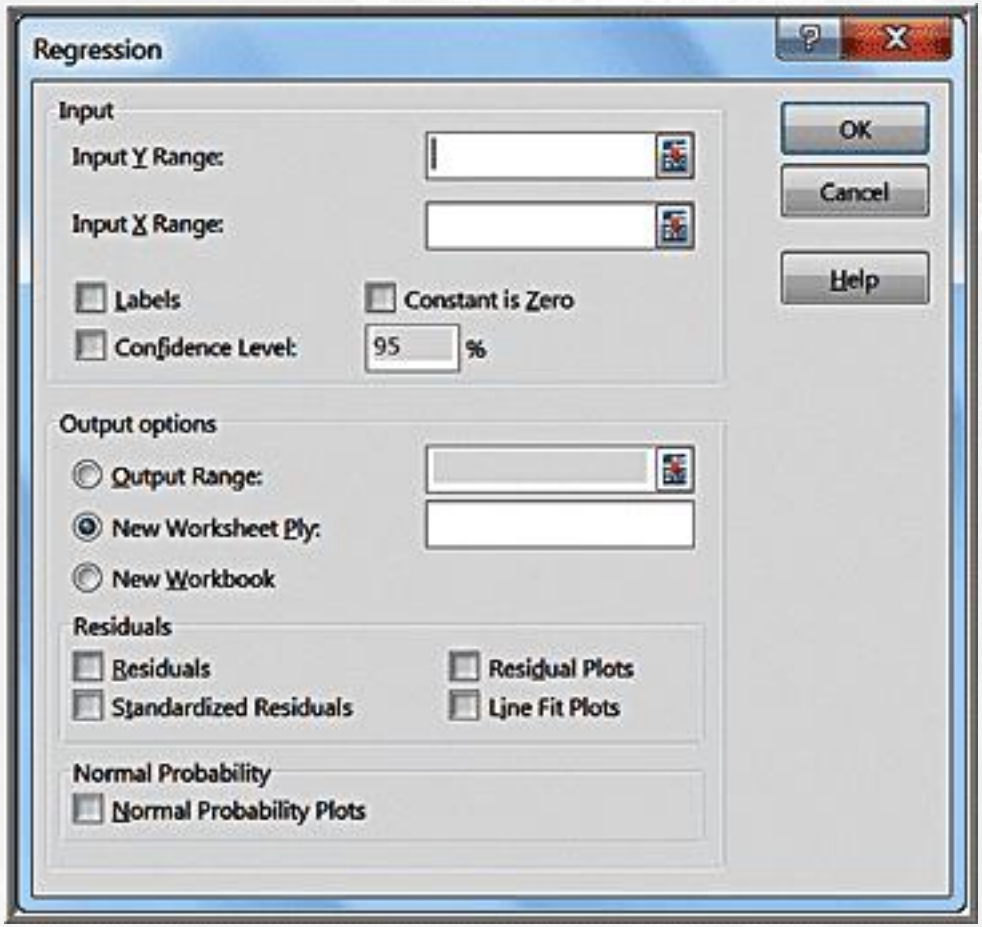
  $\hat{Y}$ = 32,673 + 35.036(1750) = $93,986

  =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
- **Adjusted R Square** – 校正的拟合优度
- **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？