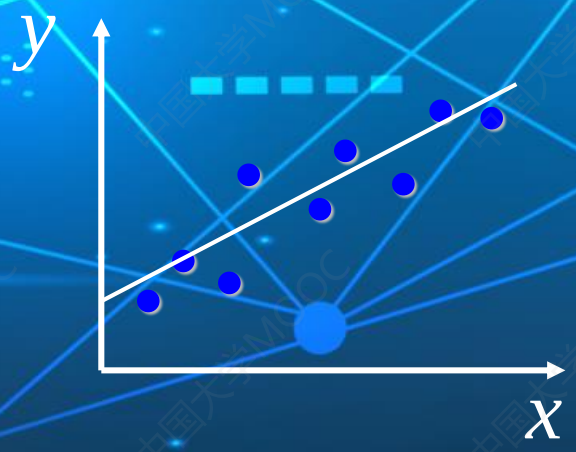
数据科学基础 I
（Matlab）

东北大学

一元线性回归

问题提出

若可控变量x与随机变量 y之间有线性相关关系，其 n 对观测值记为 $(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$,



试找出 y 与 x 间近似函数关系。

一元线性回归

回归模型

若两个变量x, y之间有线性相关关系，其回归模型为

$$y_i = a + bx_i + \varepsilon_i$$

y 称为因变量，x 称为自变量， 称为随机扰动，a, b 称为待估计的回归参数，下标 i 表示第 i 个观测值。

一元线性回归

对于回归模型，我们假设：

$$\varepsilon_i \sim N(0,\sigma^2), i=1,2,\cdots,n$$

$$E(\varepsilon_i \varepsilon_j)=0, i \neq j$$

可得到：$y_i \sim N(a+bx_i,\sigma^2)$

一元线性回归

回归方程

- 去掉回归模型中的扰动项，得理论回归方程为：

$$y_i = a + bx_i$$

- 如果给出a 和b 的估计量分别为 $\hat{a},\hat{b}$ ，则经验回归方程为：

$$\hat{y}_i = \hat{a} + \hat{b}x_i$$

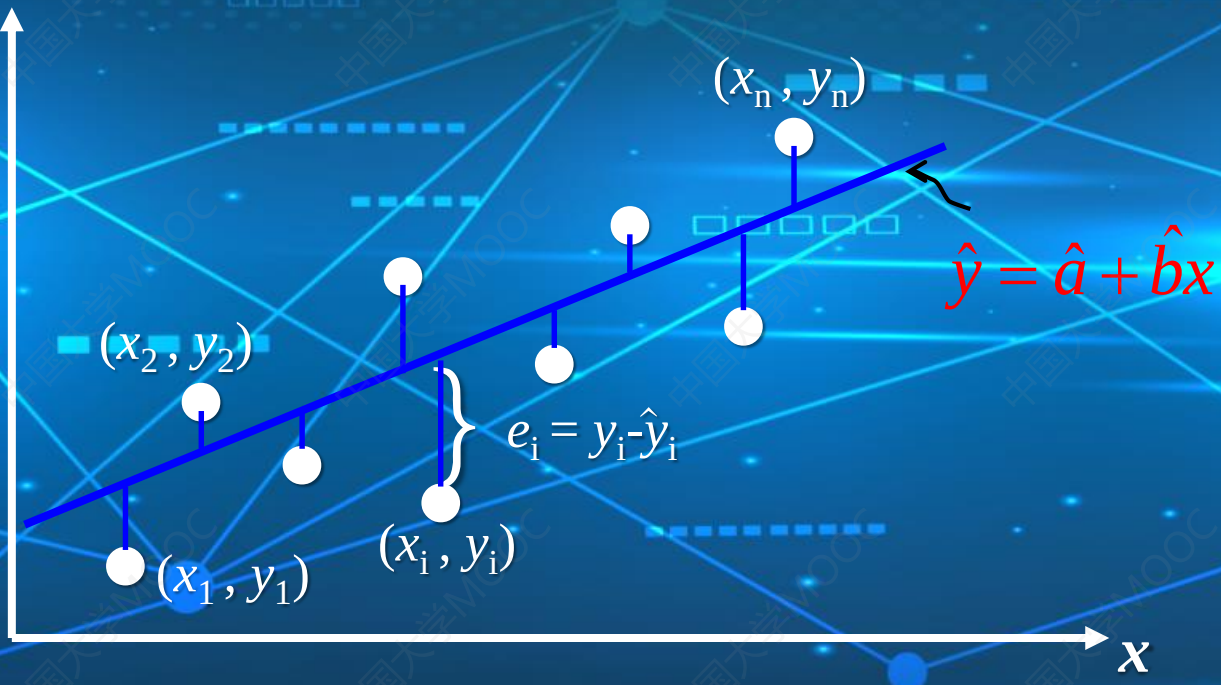
一元线性回归

回归方程

一般地， $e_i = y_i - \hat{y}_i$ 称为残差，

残差 e_i 可视为扰动 ε_i 的“估计量”。

一元线性回归



一元线性回归

最小二乘估计

记 $Q(a,b)=\sum_{i=1}^{n}\varepsilon_i^{\,2}=\sum_{i=1}^{n}[y_i-(a+bx_i)]^2$

二元函数 $Q(a,b)$ 的最小值点 $(\hat{a},\hat{b})$ **称为*a*, *b*的最小二乘估计**

最小二乘估计法是多种求解回归模型方法中最为基础的一种。

一元线性回归

MATLAB方法

```
%生成一元线性回归测试数据
X = randn(100, 1);
y = 2 * X + 3 + randn(100,1);   %带扰动
%建立回归模型
Mdl = fitlm(X, y)
%模型拟合效果图
Mdl.plot;
%预测
newx = 0.5;
newy = predict(Mdl, newx)
```

一元线性回归

MATLAB方法

```
Mdl = fitlm(X,    y);
```

自变量

因变量

一元线性回归

MATLAB方法

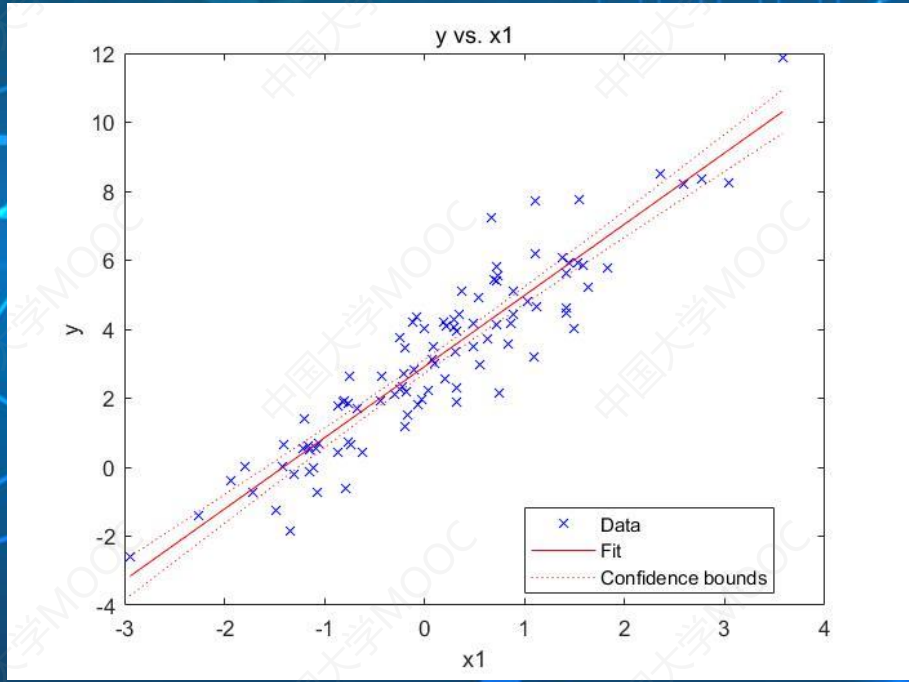
```
Mdl =
Linear regression model:
    y ~ 1 + x1
Estimated Coefficients:
                   Estimate      SE         tStat       pValue
                   ________    ________    _______    _____________

    (Intercept)     2.9193     0.1013        28.819    1.1939e-49
    x1              2.0652     0.087093      23.713    2.1985e-42
Number of observations: 100, Error degrees of freedom: 98
Root Mean Squared Error: 1.01
R-squared: 0.852,  Adjusted R-Squared 0.85
F-statistic vs. constant model: 562, p-value = 2.2e-42
```

一元线性回归

MATLAB方法

Mdl.plot;



回归诊断

- 回归诊断是对回归分析中的假设以及数据的检验与分析。
- 从数据的角度，回归诊断的主要任务是查找异常点并做相应处理。
- 通过MATLAB回归模型对象的Residuals属性可以查看残差，找到异常点。

回归诊断

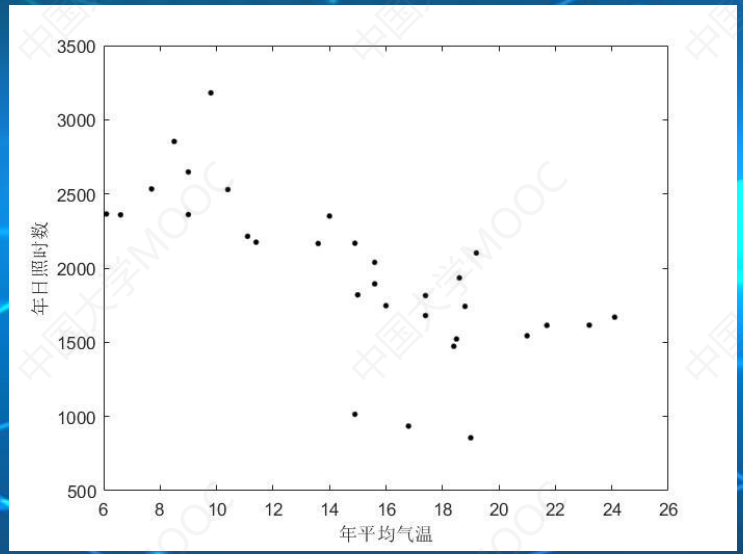
案例：各城市年平均气温与年日照时数关系

数据

	A	B	C	D	E	F
1	城 市	年平均气温 单位：℃	年极端最高气温 单位：℃	年极端最低气温 单位：℃	年均相对湿度 单位：%	全年日照时数 单位：小时
2	北 京	14	37.3	-11.7	54	2351.1
3	天 津	13.6	38.5	-10.6	61	2165.4
4	石家庄	14.9	39.7	-7.4	59	2167.7
5	太 原	11.4	35.8	-13.2	55	2174.6
6	呼和浩特	9	35.6	-17.6	47	2647.8
7	沈 阳	9	33.9	-23.1	68	2360.9
8	长 春	7.7	35.8	-21.7	58	2533.6
9	哈尔滨	6.6	35.8	-22.6	58	2359.2
10	上 海	18.5	39.6	-1.1	73	1522.2
11	南 京	17.4	38.2	-4.5	70	1680.3
12	杭 州	18.4	39.5	-1.9	71	1472.9
13	合 肥	17.4	37.2	-3.5	79	1814.6
14	福 州	21	39.8	3.6	68	1543.8
15	南 昌	19.2	38.5	0.5	68	2102
16	济 南	15	38.5	-7.9	61	1819.8
17	郑 州	16	39.7	-5	60	1747.2
18	武 汉	18.6	37.2	-1.5	67	1934.2
19	长 沙	18.8	38.8	-0.5	70	1742.2
20	广 州	23.2	37.4	5.7	71	1616
21	南 宁	21.7	37.7	0.7	76	1614
22	海 口	24.1	37.9	10.7	80	1669.1
23	重 庆	19	37.9	3	81	856.2
24	成 都	16.8	34.9	-1.6	77	935.6
25	贵 阳	14.9	31	-1.7	75	1014.8
26	昆 明	15.6	30	0.7	72	2038.6
27	拉 萨	9.8	29	-9.8	34	3181
28	西 安	15.6	39.8	-5.9	58	1893.6
29	兰 州	11.1	34.3	-11.9	53	2214.1
30	西 宁	6.1	30.7	-21.8	57	2364.7
31	银 川	10.4	35	-15.4	52	2529.8
32	乌鲁木齐	8.5	37.6	-24	56	2853.4

回归诊断

```
climatedata = xlsread('climate.xls');
x = climatedata(:,1);   %年平均气温
y = climatedata(:,5);   %年日照时数
plot(x, y, 'k.', 'Markersize', 10);
xlabel('年平均气温');
ylabel('年日照时数');
```



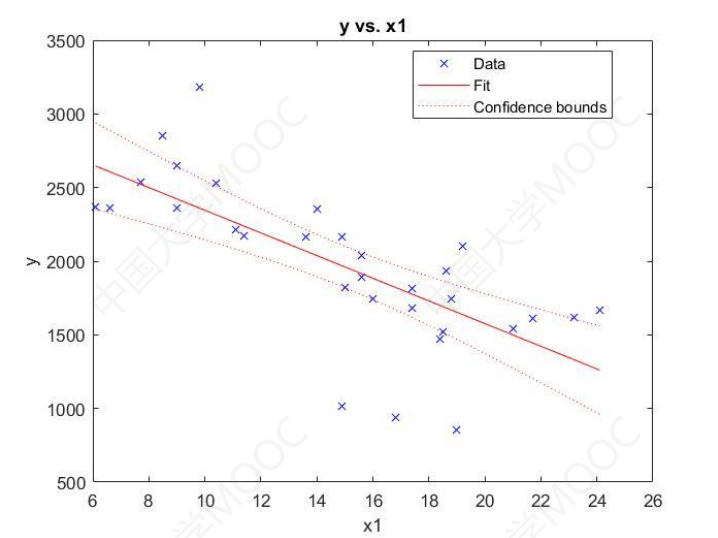
回归诊断

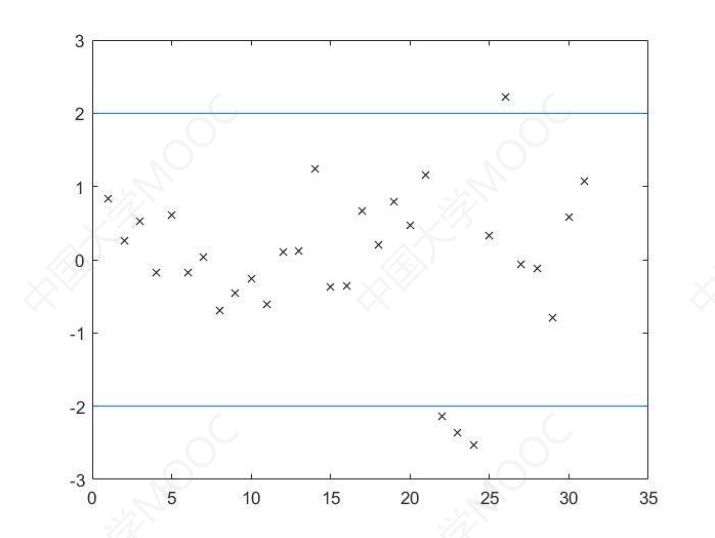
案例：各城市年平均气温与年日照时数关系

```
%回归
mdl = fitlm(x, y);
figure;
mdl.plot;
%诊断
Res = mdl.Residuals;
Res_stan = Res.Standardized;    %标准化残差
figure;
plot(Res_stan, 'kx');
refline(0, 2);
refline(0, -2);
```

回归诊断

案例：各城市年平均气温与年日照时数关系





回归诊断

案例：各城市年平均气温与年日照时数关系

```
%剔除异常值
id = find(abs(Res_stan)>2);
mdl2 = fitlm(x, y, 'Exclude', id);
figure;
mdl2.plot;
```

回归诊断

案例：各城市年平均气温与年日照时数关系