

数学建模讲义

统计模型

——孕妇吸烟与胎儿健康

博雅教育学院

孕妇吸烟与胎儿健康



吸烟有害健康！孕妇吸烟是否会伤害到腹中的胎儿？

- 对于新生儿体重，吸烟比妇女怀孕前身高、体重、受孕历史等因素的影响更为显著——美国公共卫生总署警告



美国儿童保健和发展项目(CHDS)提供的数据(1236个出生后至少存活28天男性单胞胎新生儿体重及其母亲的资料)

1.新生儿体重 (oz)	120	113	128	123	108	...
2.孕妇怀孕期 (天)	284	282	279	999	282	...
3.新生儿胎次(1~第1胎,0~非第1胎)	1	0	1	0	1	...
4.孕妇怀孕时年龄	27	33	28	36	23	...
5.孕妇怀孕前身高(in)	62	64	64	69	67	...
6.孕妇怀孕前体重(lb)	100	135	115	190	125	
7.孕妇吸烟状况(1~吸烟,0 ~不吸烟)	0	0	1	1	1	

研究目的

利用CHDS的数据建立新生儿体重与孕妇怀孕期、吸烟状况等因素的数学模型，定量地讨论：

- 对于新生儿体重来说，**孕妇吸烟**是否是比较孕妇年龄、身高、体重等**更为显著的决定因素**；
- 孕妇吸烟是否会使**早产率增加**，怀孕期长短对新生儿体重有影响吗；
- 对**每个年龄段**来说，孕妇吸烟对新生儿体重和早产率的影响是怎样的。

问题背景及分析

美国公共卫生总署的警告容易受到人们的**质疑**：按照是否吸烟划分人群所做的研究，只能依赖于观测数据，而**无法做人为的实验**，很难确定新生儿体重的差别是因为吸烟，还是其它因素(如怀孕期长短、吸烟孕妇多是体重较轻的年青人等)。

“孕妇吸烟可能导致胎儿受损、早产及新生儿低体重”的警告不如“吸烟导致肺癌”来得强，是由于**对孕妇吸烟与胎儿健康间的生理学关系研究得不够**。

参数估计

参数估计	不吸烟孕妇 ($n=742$)	吸烟孕妇 ($n=484$)
新生儿体重均值的点估计	$\mu_{y0}=123.0472$	$\mu_{y1}=114.1095$
新生儿体重均值的区间估计	[121.7932 124.3011]	[112.4930 115.7260]
新生儿体重低比例的点估计	$r_0=0.0310$	$r_1=0.0826$
怀孕期均值的点估计	$\mu_{x0}=280.1869$ ($n=733$)	$\mu_{x1}=277.9792$
怀孕期均值的区间估计	[278.9812 281.3926]	[276.6273 279.3311]
早产率的点估计	$q_0=0.0764$	$q_1=0.0854$

- 吸烟比不吸烟孕妇新生儿体重**平均低9 oz (250g)**, 新生儿体重低的比例明显高.
- 吸烟比不吸烟孕妇怀孕期**平均短2天**, 早产率差不多.

新生儿体重和怀孕期的差别在统计学上是否显著?

假设检验

假设检验	假设	检验结果($\alpha=0.05$)
新生儿体重均值	$H_0: \mu_{y0}=\mu_{y1}, H_1: \mu_{y0}\neq\mu_{y1}$	拒绝 H_0 , 接受 H_1
怀孕期均值	$H_0: q_0=q_1, H_1: q_0\neq q_1$	接受 H_0 , 拒绝 H_1

- 吸烟孕妇的新生儿体重比不吸烟孕妇的低、且新生儿体重低的比例高，在**统计学上有显著意义**。
- 吸烟与不吸烟孕妇孕期的差别**难以肯定是显著的**(若 $\alpha=0.01$ 将接受怀孕期均值相等的假设)

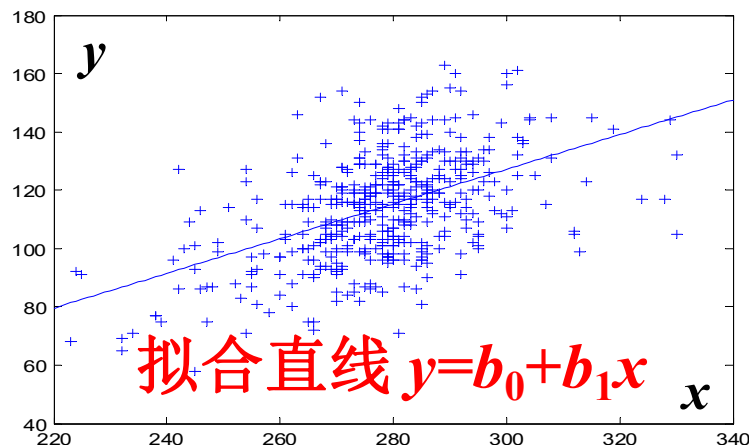


一元线性回归分析

假设检验结果：孕妇吸烟状况对新生儿体重大小有显著影响，但是对怀孕期长短的影响难以确定。

- 新生儿体重与怀孕期的关系如何？

480位吸烟孕妇的怀孕期 x 和新生儿体重 y



直线 $y=b_0+b_1x$ 描述了数据的变化趋势，但是拟合得不好

- 怎样衡量由拟合得到的模型的有效性？
- 模型系数精确度和模型预测的数值范围多大？

一元线性回归模型 $y=b_0+b_1x+\varepsilon$ 怀孕期 x , 新生儿体重 y

随机变量 $\varepsilon \sim$ 除 x 外, 影响 y 的随机因素的总和,
对于不同的 x , ε 相互独立且服从 $N(0,\sigma^2)$ 分布.

模型求解
480位吸烟孕妇数据 x, y

系数	系数估计值	系数置信区间
b_0	-51.2983	[-77.5110 -25.0856]
b_1	0.5949	[0.5008 0.6891]
$R^2=0.2438, F=154, p<0.0001, s^2=249$		

模型检验

- b_1 置信区间不含零点, $F=154 \gg F_{(1,n-2)}=3.8610$ ($\alpha=0.05$), 应拒绝 $H_0: b_1=0$ 的假设, 模型有效。
- b_1 置信区间较长, 决定系数 R^2 较小(y 的24.38%由 x 决定), 剩余方差 s^2 较大, 模型的精度不高。

一元线性回归模型 $y=b_0+b_1x+\varepsilon$

怀孕期 x , 新生儿体重 y

模型
解释

$$\widehat{b}_1=0.5949$$

- 吸烟孕妇怀孕期增加一天, 新生儿体重平均增加约0.6 oz.

$$\widehat{b}_0=-51.2983$$

- 不是 $x=0$ 时 y 的估计, 只能在数据范围内($x=220\sim 340$ 天) 估计.

模型
预测

$$\widehat{y} = \widehat{b}_0 + \widehat{b}_1 x = -51.2983 + 0.5949x$$

若怀孕期 $x=280$ 天, 新生儿体重 $\widehat{y}=114.5937$ oz, 预测区间为[88.0949, 141.0925]

- 模型精度不高导致预测区间如此之大!

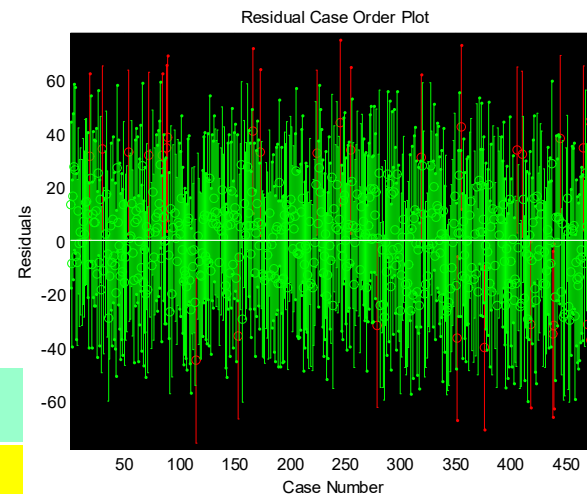
一元线性回归模型 $y=b_0+b_1x+\varepsilon$

怀孕期 x , 新生儿体重 y

模型残差 $e=y-\hat{y}$ ~误差 ε 的估计值(均值为0的正态分布)

若数据残差的置信区间不含
零点, 称为**异常点**(偏离整体
数据的变化趋势), **应剔除**。

系数	系数估计值	系数置信区间
b_0	-53.6126	[-77.0606 -30.1645]
b_1	0.6007	[0.5164 0.6850]
$R^2= 0.3040$ $F=196$ $p<0.0001$ $s^2= 182$		



虽然 b_0 和 b_1 的估计值变化不大, 但置信区间变短,
且 R^2 和 F 变大, s^2 减小, 说明**模型精度得到提高**。

一元线性回归模型 $y=b_0+b_1x+\varepsilon$

怀孕期 x , 新生儿体重 y

690位不吸烟孕妇数据 x,y (剔除异常点后)

系数	系数估计值	系数置信区间
b_0	33.5330	[14.9989 52.0671]
b_1	0.3201	[0.2541 0.3860]
$R^2=0.1165$ $F=90$ $p<0.0001$ $s^2=181$		

$$\widehat{b_1}=0.3201$$

- 不吸烟孕妇怀孕期增加一天，新生儿体重平均只增加0.32oz.

- 对吸烟孕妇是增加约0.6oz，二者相差很大！

将吸烟状况作为另一自变量，建立新生儿体重与2个自变量的回归模型，利用全体孕妇数据进行分析



多元线性回归分析

$$\text{模型 } y = b_0 + b_1 x_1 + b_2 x_2 + \varepsilon$$

y ~新生儿体重, x_1 ~孕妇怀孕期, $x_2=0,1$ ~不吸烟, 吸烟

1145位全部孕妇数据 (剔除异常点后)

$$\hat{y} = 0.7698 + 0.4365x_1 - 8.7610x_2$$

$$\widehat{b_1} = 0.4365$$

- 对于吸烟状况 x_2 相同的孕妇,
 x_1 增加一天 y 平均增加0.44oz.

在吸烟孕妇的0.6与不吸烟孕妇的0.32oz之间.

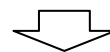
$$\widehat{b_2} = -8.7610$$

- x_1 相同时, 吸烟比不吸烟孕妇
的新生儿体重平均约低8.8oz.

与参数估计的数值相同, 但增加了 x_1 相同的条件.

多元线性回归分析

$$\text{模型 } y=b_0+b_1x_1+b_2x_2+\varepsilon$$



增加乘积项 $x_1x_2 \sim x_1$
和 x_2 对 y 的综合影响

$$y=b_0+b_1x_1+b_2x_2+b_3x_1x_2+\varepsilon$$

系数	系数估计值	系数置信区间
b_0	34.0925	[15.4605 52.7244]
b_1	0.3181	[0.2517 0.3844]
b_2	-87.0738	[-116.9656 -57.1820]
b_3	0.2804	[0.1734 0.3875]
$R^2=0.2766 \quad F=145 \quad p<0.0001 \quad s^2=183$		

模型有效,但是
 R^2 较小, s^2 较大,
仍有改进余地.

$$\hat{y} = 34.0925 + 0.3181x_1 - 87.0738x_2 + 0.2804x_1x_2$$

$$x_2=0 \quad \Downarrow$$

$$\hat{y} = 34.0925 + 0.3181x_1$$

不吸烟孕妇的一元模型

$$\Downarrow x_2=1$$

$$\hat{y} = -52.9813 + 0.5985x_1$$

吸烟孕妇的一元模型



变量选择与逐步回归

- CHDS提供的数据中除孕妇怀孕期和吸烟状况外, 还有孕妇怀孕时的年龄、体重、身高和胎次状况.
- 新生儿体重模型中是否应该加入其他的自变量?

变量选择 ~ 从应用的角度希望将所有影响显著的自变量都纳入模型, 又希望最终的模型尽量简单.

逐步回归 ~ 迭代式的变量选择方法.

- 利用CHDS数据提供的全部信息, 通过逐步回归方法选择变量, 建立新生儿体重的线性回归模型.

用逐步回归方法建立新生儿体重 y 的线性回归模型

x_1 (孕妇怀孕期), x_2 (胎次状况), x_3 (年龄), x_4 (身高), x_5 (体重), x_6 (吸烟状况) 组成候选变量集合 S .

- 选取 x_1, x_6 为初始子集 S_0
- 从 S_0 外的 S 中引入一个对 y 影响最大的 x , $S_0 \rightarrow S_1$.
- 对 S_1 中的 x 进行检验, 移出一个影响最小的, $S_1 \rightarrow S_2$
- 继续进行, 直到不能引入和移出为止.
- 引入和移出都以给定的显著性水平为标准.

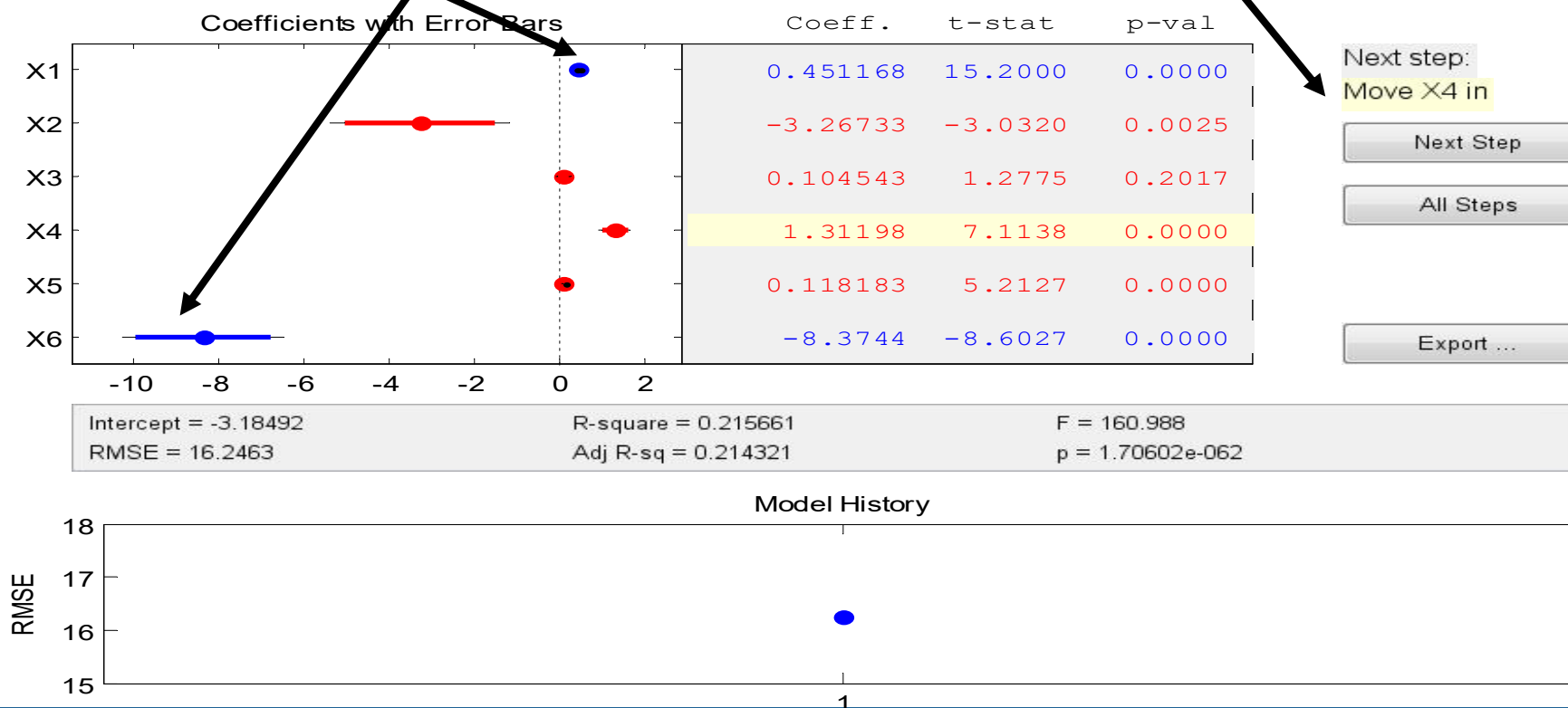
显著性水平取缺省值(引入 $\alpha=0.05$, 移出 $\alpha=0.10$)

MATLAB统计工具箱中的逐步回归

逐步回归命令`stepwise`第1个输出图形

x_1, x_6 在模型中, 给出系数估计值和置信区间

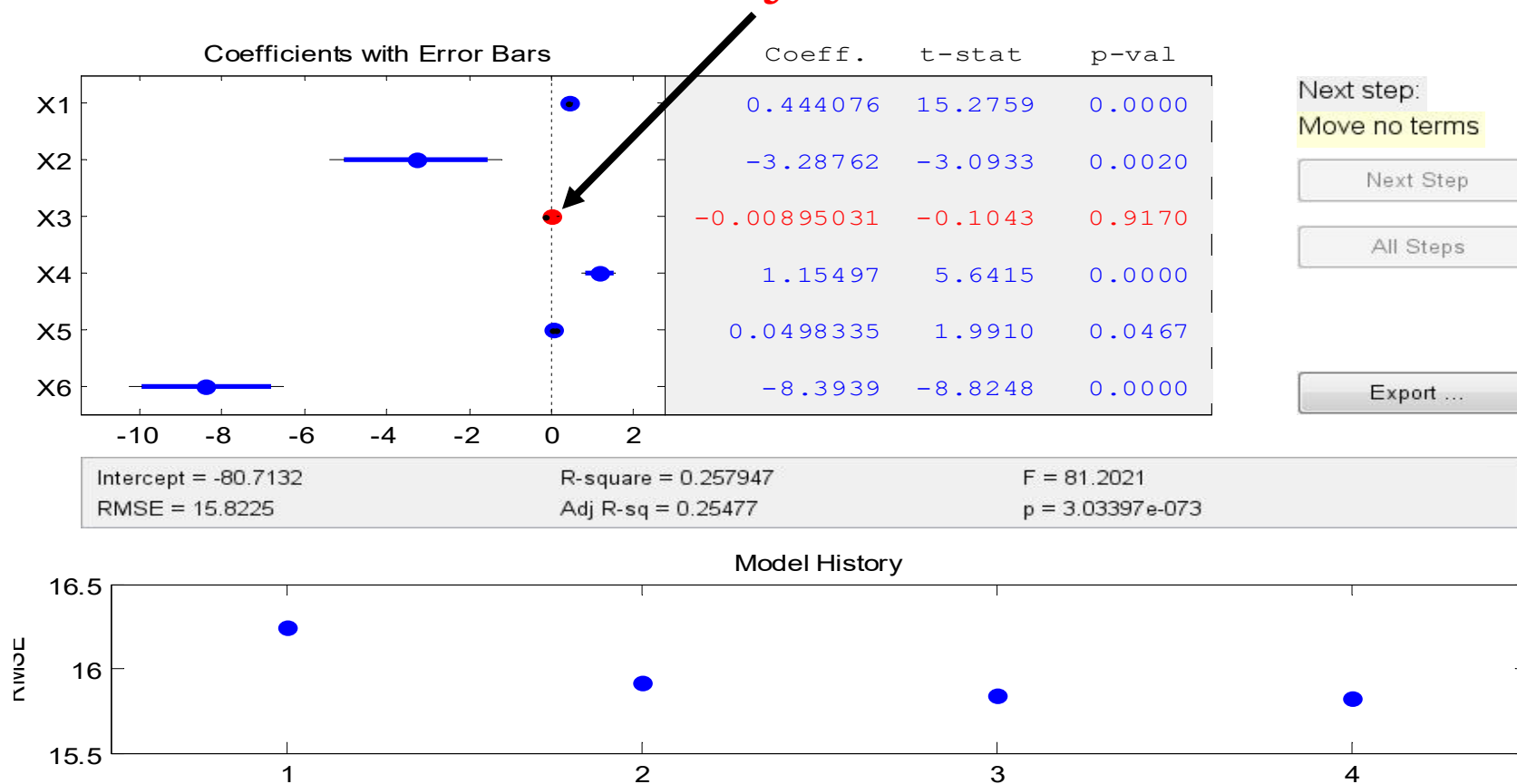
按照提示点击, 引入 x_4



MATLAB统计工具箱中的逐步回归

按照提示点击，依次引入 x_4, x_2, x_5

最终模型包含除 x_3 外的所有自变量



用逐步回归方法建立新生儿体重 y 的线性回归模型

$$\hat{y} = -80.7132 + 0.4441x_1 - 3.2876x_2 + 1.1550x_4 + 0.0498x_5 - 8.3939x_6$$

x_1 (怀孕期), x_2 (胎次状况), x_4 (身高), x_5 (体重), x_6 (吸烟状况).

$$\widehat{b_6} = -8.3939$$

- x_1, x_2, x_4, x_5 相同时, 吸烟比不吸烟孕妇的新生儿体重平均低8.4 oz.

$$\widehat{b_1}, \widehat{b_4}, \widehat{b_5} > 0$$

- 孕妇的怀孕期、身高、体重对新生儿体重的影响是正面的.

$$\widehat{b_2} = -3.2876$$

- 第1胎新生儿体重比非第1胎平均约低3.3 oz (第1胎 $x_2=1$).



相关分析

y 和各自变量的相关系数矩阵

	y	x_1	x_2	x_3	x_4	x_5	x_6
y	1.0000	0.4075	-0.0439	0.0270	0.2037	0.1559	-0.2468
x_1		1.0000	0.0809	-0.0534	0.0705	0.0237	-0.0603
x_2			1.0000	-0.3510	0.0435	-0.0964	-0.0096
x_3				1.0000	-0.0065	0.1473	-0.0678
x_4					1.0000	0.4353	0.0175
x_5						1.0000	-0.0603
x_6							1.0000

- 与 y 相关性较强的是怀孕期 x_1 , 吸烟状况 x_6 , 身高 x_4 .
- 自变量间相关性较强的有: 孕妇体重 x_5 与身高 x_4 的正相关; 年龄 x_3 与胎次状况 x_2 的负相关(年龄越大第1胎 $x_2=1$ 越少).

当几个自变量间有较强相关性时, 删除多余的只保留一个不会对模型有效性和精确度有多大影响

不同年龄段孕妇吸烟对新生儿体重的影响

孕妇按年龄分组建立 y 与 x_1, x_2, x_4, x_5, x_6 的回归模型

	小于25岁	25~30岁	30~35岁	大于35岁
b_0	-66.3893	-39.1296	-157.1307	-130.1740
b_1 (怀孕期)	0.3972	0.3521	0.5951	0.6728
b_2	-0.9978	-7.4124	-0.0932	-4.1835
b_4	1.2144	0.8409	1.6828	0.8747
b_5	-0.0021	0.0959	0.0557	0.0732
b_6 (吸烟状况)	-8.4119	-8.2656	-10.5411	-6.4008
R^2	0.2549	0.2330	0.3394	0.3136
s^2	211.6359	239.7201	272.6021	304.7208
n	444	362	211	157

对于 x_1 和 x_6 两个影响 y 的主要因素, 30岁以下两组结果差别不大, 而与30岁以上两组则有一定差异.



练习

- 运行本PPT中备注里面的代码，并进行结果解释。
- 利用预测值的区间估计的计算公式，对480位吸烟孕妇数据，若怀孕期 $x=280$ 天，计算新生儿体重95%的置信区间。