

请利用 CHDS 的数据建立新生儿体重与孕妇怀孕期、吸烟状况等诸因素的数学模型（可用线性回归模型），定量地讨论以下问题：

（1）对于新生儿体重来说，孕妇吸烟是否是比孕妇年龄、身高、体重等更为显著的决定因素；

ANOVA^a

模型		平方和	自由度	均方	F	显著性
1	回归	11412.609	4	2853.152	8.796	.000 ^b
	残差	399309.089	1231	324.378		
	总计	410721.698	1235			

a. 因变量：新生儿体重
b. 预测变量：(常量), 孕妇体重, 吸烟状况, 孕妇年龄, 孕妇身高

系数^a

模型		未标准化系数		标准化系数	t	显著性
		B	标准错误	Beta		
1	(常量)	84.828	7.790		10.889	.000
	吸烟状况	-2.014	.562	-.101	-3.583	.000
	孕妇年龄	.069	.080	.025	.872	.384
	孕妇身高	.536	.122	.155	4.397	.000
	孕妇体重	-.006	.004	-.046	-1.320	.187

a. 因变量：新生儿体重

在系数估计表格中查看每个自变量的显著性水平（通常使用的是 p 值）。显著性水平低于通常的阈值（如 0.05）表示该自变量对因变量有显著影响。

根据系数估计和显著性水平，判断孕妇吸烟状况是否比孕妇年龄、身高、体重等其他因素更为显著地影响新生儿体重：

孕妇吸烟状况的系数为负数，P 值为 0.000，说明孕妇吸烟状况对新生儿体重有显著的负向影响。其他自变量如孕妇身高等也具有显著性，相比孕妇年龄和孕妇体重，孕妇吸烟状况的影响显得更为突出。

（2）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；

计算变量

目标变量(T):

早产儿

类型和标签(L)...

序号

新生儿体重

数字表达式(E):

怀孕天数<259

该题中的早产率需要自己计算，需要创建一个新的变量来表示是否早产，比如命名为“早产儿”：

如果怀孕天数 < 259 天（即 37 周 * 7 天），则为 1（表示早产）。

如果怀孕天数 >= 259 天，则设定为 0（表示非早产）

方程中的变量							
		B	标准误差	瓦尔德	自由度	显著性	Exp(B)
步骤 1 ^a	吸烟状况	-.064	.136	.219	1	.640	.938
	常量	-2.435	.121	408.051	1	.000	.088

a. 在步骤 1 输入的变量：吸烟状况。

该表中孕妇吸烟状况的系数为负数，P 值为 0.000，说明孕妇吸烟状况对新生儿是否早产有的显著影响。

(3)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。

系数 ^a						
模型		未标准化系数		标准化系数	t	显著性
		B	标准错误	Beta		
1	(常量)	115.226	2.045		56.356	.000
	怀孕天数	.015	.007	.063	2.200	.028

a. 因变量：新生儿体重

若体重小于 2500 克则认为是低体重新生儿；否则认为是非低体重新生儿。

- (1)请用 spss 建立 Logistic 回归模型考虑吸烟和是否为低 体重新生儿的关系；
- (2) 请筛选出新生儿是否为 低体重的影响因素，进一步建立 Logistic 回归模型考 虑这些影响因素和是否低体重新生儿的关系。



该题中的低体重新生儿需要自己计算，需要创建一个新的变量来表示是否早产，比如命名为“低体重新生儿”，结果是为 1，不是为 0

		方程中的变量					
		B	标准误差	瓦尔德	自由度	显著性	Exp(B)
步骤 1 ^a	吸烟状况	.143	.097	2.182	1	.140	1.153
	常量	-3.000	.142	444.517	1	.000	.050

a. 在步骤 1 输入的变量：吸烟状况。

根据 Logistic 回归模型可知吸烟和是否为低体重新生儿有关

		方程中的变量					
		B	标准误差	瓦尔德	自由度	显著性	Exp(B)
步骤 0	常量	-2.924	.129	511.246	1	.000	.054

			未包括在方程中的变量		
			得分	自由度	显著性
步骤 0	变量	吸烟状况	2.322	1	.128
		孕妇年龄	.481	1	.488
		孕妇身高	2.493	1	.114
		孕妇体重	2.300	1	.129
	总体统计		6.075	4	.194

根据系数估计和显著性水平，判断其他因素是否低体重新生儿体重。

由表可知，孕妇身高、孕妇体重和孕妇吸烟状况和是否为低体重新生儿的联系更为密切。

2. 请根据大学某届学生成绩相关数据集，建立分位数回归模型来分析不同的学院、任课教师职称，以及高数成绩对概率论成绩分布的影响。

任课教师	q=0.5	学院名称	q=0.5
	d		d
副高1	90.472	材料与冶金学院	87.801
副高2	98.145	城市建设学院	105.000
副高3	73.447	管理学院	90.472
副高4	81.891	国际学院	108.676
讲师2	78.162	化学与化工学院	100.069
讲师3	75.550	机械自动化学院	67.430
讲师4	90.040	计算机科学与技术学院	84.924
讲师5	96.648	理学院	105.611
讲师6	111.727	汽车与交通工程学院	81.047
讲师7	77.570	文法与经济学院	89.426
教授1	92.139	信息科学与工程学院	98.065
教授2	76.478	医学院	109.529
教授3	85.741	资源与环境工程学院	105.408

a. 因变量：卷面得分

b. 模型：（截距），任课教师, 学院名称, 所在班级

c. 按下列值对模型的预测变量进行求值：学院名称=管理学院, 所在班级=软件工程2016级2班

d. 此效应不可估计

a. 因变量：卷面得分

b. 模型：（截距），任课教师, 学院名称, 所在班级

c. 按下列值对模型的预测变量进行求值：任课教师=副高1, 所在班级=软件工程2016级2班

d. 此效应不可估计

由表中数据分析可知，任课教师的职称高低与学生的概率论成绩相关性不大，在高低职称中都有低分和高分。而学生所属的学院学生的概率论成绩有一定的相关性，可知，理工科的概率论成绩相对较高，而文科类学院的学生概率论成绩相对较低。