分类号_____ 编 号_____

U D C_____ 密 级_____

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

本科生毕业设计（论文）

题　　目：低出生体重风险因子的回归分析与预测

姓　　名：李林森

学　　号：11510320

系　　别：数学系

专　　业：统计学

指导教师：蒋学军

2019 年 4 月 30 日

诚信承诺书

1.本人郑重承诺所呈交的毕业设计(论文),是在导师的指导下，独立进行研究工作所取得的成果，所有数据、图片资料均真实可靠。

2.除文中已经注明引用的内容外，本论文不包含任何其他人或集体已经发表或撰写过的作品或成果。对本论文的研究作出重要贡献的个人和集体，均已在文中以明确的方式标明。

3.本人承诺在毕业论文（设计）选题和研究内容过程中没有抄袭他人研究成果和伪造相关数据等行为。

4.在毕业论文（设计）中对侵犯任何方面知识产权的行为，由本人承担相应的法律责任。

作者签名：

2019 年 4 月 30 日

低出生体重风险因子的回归分析与预测

李林森

（数学系　指导教师：蒋学军）

[摘要]：研究表明，婴儿期死亡率增加与婴儿出生时体重过轻有关。值得注意的是，相较于正常孩子，低出生体重的婴儿在儿童期和青春期后期，有着更高的的死亡率。除此之外，他们还遭受许多其他问题，比如对婴儿的智力发育也有一定影响。基于这些原因，研究关于哪些因素导致婴儿出生体重低是非常有必要的。尽管有大量研究文献可用，但仍存在许多尚未解决的问题，需要对该主题进行进一步研究，以便更好地估计婴儿低出生体重的性质和原因，从而帮助我们对母亲提出建议，降低低体重婴儿发生的概率。在我们的研究中，主要目标是找出可能导致婴儿低体重出生的关键因素。我们发现了一个非常有价值的数据集，它是一个是纵向的横截面数据，我们有婴儿出生体重的数值和分类值，以及母亲的一些特征比如母亲的年龄，人种等。它为我们提供了一个很好的机会，我们首先对数据的各个变量进行了简要分析，针对不同对数据类型，我们分别采用了线性回归模型和逻辑回归模型。 我们还对结果的有效性进行了深入的诊断，包括常态检验，多重假设检验等。并利用改善后的模型，构造了一个分类器，将数据随机分成训练集和测试集，测试分类器预测的准确性，最后达到高约 80%准确度的标准。这项研究表明，诸如最后一次月经期母亲的

体重，种族，怀孕期间的吸烟习惯，高血压病史以及子宫过敏的存在等变量对婴儿的低出生体重有显着影响。

[关键词]：低体重出生， 逻辑回归，线性回归，分类器

[ABSTRACT] : This paper concludes the research about the risk factors of low birth weight. Studies have shown that increased infant mortality is associated with underweight infants at birth. It is worth noting that low birth weight infants have a higher mortality rate during childhood and later puberty than normal children. In addition, they suffer from many other problems, such as the mental development of the baby. For these reasons, it is necessary to study which factors contribute to the low birth weight of the baby. Although a large body of research literature is available, there are still many unresolved issues that require further research on the subject to better estimate the nature and causes of low birth weight in infants, in order to help us to advise mothers to lower low birth weight infants. In our study, the main goal was to identify the key factors that could lead to low birth weight in infants. We found a very valuable data set, which is a longitudinal cross-sectional data, we have the value and classification of the baby's birth weight, as well as some characteristics of the mother such as the mother's age, ethnicity and so on. It provides us with a good opportunity. We first briefly analyze the various variables of the data. For different pairs of data types, we use linear regression model and logistic regression model. We also conducted in-depth diagnosis of the validity of the results, including normality tests, multiple hypothesis tests, and so on. And using the improved model, a classifier is constructed to randomly divide the data into training sets and

test sets, test the accuracy of the classifier prediction, and finally reach the standard of about 80% accuracy. The study showed that variables such as the weight, ethnicity, smoking habits during pregnancy, history of hypertension, and the presence of uterine allergies in the last menstrual period had a significant effect on the low birth weight of the infant.

[Keywords]: Low birth weight, Logistic regressive, Linear regressive, Classifier

目录

1. 引言

1.1 低出生体重儿概述性介绍

在 2001 年，世界卫生组织（WHO）将婴儿出生时体重小于 2500 克定义为低出生体重（LBW）。据统计，低出生体重（LBW）的全球患病率为 15.5%，每年约有 2000 万低出生体重婴儿，其中 96.5%的情况发生在发展中国家[1]。早在 20 世纪晚期，出生时体重低的婴儿引起了极大的关注，因为根据健康科学家的说法，这些婴儿可能比正常体重的婴儿更有可能在以后的生活中有某些健康状况的隐患，包括糖尿病，心脏病，高血压，代谢综合征。在正常情况下，婴儿的出生体重平均应为 2700 克至 4000 克。然而，在极端情况下，婴儿的体重甚至可能小于 1000 克，这被认为是极低的出生体重。并且这些出生低体重儿的死亡率极高，相较于正常体重范围之内的新生婴儿，体重低于 2500 克的，也就是低出生体重患者，死亡率是这些正常婴儿的 130 倍。所以，这个严峻的问题引起了越来越多的研究者的关注。研究者们希望探究出影响该病症发病的主要原因。事实上，除了先天的基因上的一些缺陷，母亲的一些行为特征也会影响该病症的发病率，比如包括母亲是否吸烟，她的种族，年龄，受教育程度，产前护理以及之前是否有过孩子，甚至婴儿的性别等。如果能找出相关的致病因素，那么就能够在母亲妊娠之前或者妊娠时，提出有效降低患病风险的建议，降低患病率，而这也正是本次研究的主要目的。

1.2 低出生体重儿的在我国现状

中国是世界上人口最多的国家，在 20 世纪 60 年代到 80 年代，经历了三年自然灾害过后，经济发展状况逐渐好转，加上生育政策的宽松，中国人口开始高速增长。这一时期，人口出生率最高达到 43.6‰，平均水平在 36.8‰，但是婴儿

死亡率高达 60.5‰[2]。受时代的限制，当时这个问题没有得到应有的重视。其中，早产以及其伴随的低出生体重是造成婴儿死亡的重要原因，大约占到了死亡婴儿的 70%。所以从 1991 年起，中国政府制定了《中国儿童发展纲要（2001—2010 年）》，婴儿死亡率也从 90 年代初的 5.02%下降至 2000 年的 3.2%，到了 2010 年，婴儿死亡率降至 1.31%[3] 。中国政府的最终目标是将婴儿死亡率降至 1%以下。下图显示了 1996—2013 年中国检测地区婴儿死亡率和早产或者低出生体重死亡率的随时间变化：

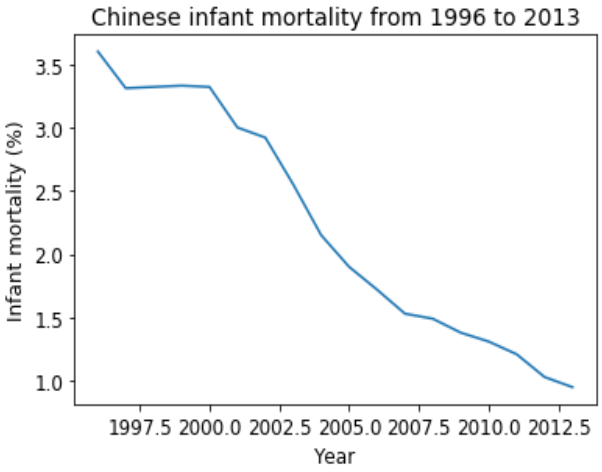

图 1　中国婴儿死亡率随时间变化图

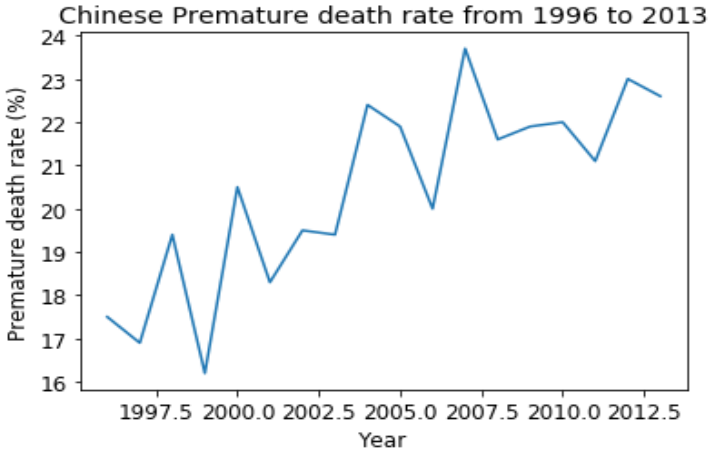

图 2　中国婴儿患低体重出生所占比例随时间变化图

上图中，图 1 是中国检测地区婴儿死亡率 1996—2013 年的变化，图 2 是低出生体重死亡占婴儿死亡比例 1996—2013 年的变化。可以看出，虽然我国的婴儿死亡率在逐年递减，但是其中低出生体重病情所占比例却是有着上升的趋势，这一点值得我们重视。所以在未来一段时间，低出生体重将对我国的婴儿健康产生极大的挑战。研究影响低出生体重发病情况的因素，可以帮助我们降低这种病的发病率，所以这个课题是有必要的。

1.3 影响低出生体重儿出现的风险因子

一些数据表明，婴儿是否患有低出生体重婴儿母亲有关。比如母亲的种族，是否吸烟，母亲年龄，是否有高血压等。然而，并没有一些关于低出生体重与这些因素之间关系的直接证据。而且，这些医院的数据更多是一种经验主义上的判断，描绘的更是患病与这些因素的相关性，仍不能体现因果性。即便如此，这些发现不容忽视，因此也促使对这一主题的进一步研究，以便研究者更好地估计出生体重低的婴儿的性质和原因。基本上所有因素可以总结为以下两类：

1. 内在因素（遗传构成因素与地理区域的因素）

我们知道在达尔文进化论的前提下，生物的生存环境会对物种进行自然选择，所以一个地区的基因频率以及基因种类，可以说大致是一定的，所以这两个因素是相辅相成的。同时，低出生体重儿（ LBW）可能由于婴儿性别，种族或民族血统的变化而导致。比如，相较于白人来说，黑人孕妇更容易生出低出生体重儿，而黑人一般集中在非洲和拉美地区。所以遗传和地理区域是一个重要的因素。

2. 外在因素

这些因素往往是很复杂并且难以量化的，比如社会的人口分布，很大程度上影响了妇女的生育年龄，一般来说，大龄产妇面临的风险比普通产妇要大。并且，产妇的心理健康状况同样也对结果产生影响，一个社会如果没有相关的机构或者官方人员关注产妇的心理健康，无疑会大大增加患病风险。同样，怀孕期间孕妇健康状况有可能也会影响的产妇发病率，可能会导致婴儿的体重状况出现异常，如患有疟疾，尿路感染和生殖道感染的孕妇，曾记录多有发生低出生体重（LBW）。进一步来说，孕妇对有毒化学品浓度高的环境或区域的有毒暴露很可能使他们的孩子患有低出生体重症。一般而言，接触吸烟，饮酒，使用咖啡因和咖啡消费的母亲，在生育低出生体重婴儿方面的可能性最高。从事使用大麻，麻醉药和其他

药物等强效药物的妇女，也有较大可能会生下低出生体重（LBW）婴儿。最后一点就是产妇对产检的重视，一般来说，孕妇如果坚持按照医生的规定产检，并且遵照医生的建议及时补充营养，可以大大降低发病概率，当然这也与孕妇家庭的经济状况有关。

总而言之，本次研究主要研究的是这些外在因素对发病率的影响，特别是给出一些母亲的特征，来预测孩子患有该疾病的风险。

2. 数据来源与描述性统计分析

2.1 数据来源

数据由 Hosmer 和 Lemeshow 于 1986 年在马萨诸塞州斯普林菲尔德的 Baystate 医疗中心收集。低出生体重是医生多年来一直关注的结果，研究者认为，这是因为低出生体重婴儿的婴儿死亡率和出生缺陷率非常高。女性在怀孕期间的行为（包括饮食，吸烟习惯和接受产前护理）可以极大地改变携带婴儿到足月的机会，从而改善正常出生体重的婴儿。故该数据收集了 189 名母亲的数据，其中 59 名母亲生育了低体重婴儿，其中 130 名婴儿出生体重正常。被认为重要的四个变量是年龄，最后一次月经期间受试者的体重，种族，以及怀孕前三个月的医生就诊次数。下图描述了这个数据的变量以及对应的解释：

变量介绍

变量	描述	变量类型（单位）	代号
1	编号（ID）	自然数	ID
2	出生时婴儿体重判断（Low Birth Weight）	1: < 2500g, 2: >2500 g	LOW
3	母亲年龄（AGE）	自然数（年）	AGE
4	母亲在最后一次月经期的体重（Weight of Mother at Last Menstrual Period）	自然数（磅）	LWT
5	母亲肤色（RACE）	1: 白, 2: 黑, 3: 其他	RACE
6	母亲是否吸烟	1: 否, 2: 是	SMOKE
7	早产史（History of premature labor）	1: 无， 2: 1次 ， 3: 2次及以上	PTL
8	高血压史（History of Hypertension）	1: 否, 2: 是	HT
9	是否存在子宫过敏（Presence of Uterine irritability ）	1: 否, 2: 是	UI
10	第一个季度的医生就诊次数（Number of Physician Visits During the First Trimester）	1: 无， 2: 1次 ， 3: 2次及以上	FTV
11	婴儿出生体重（Birth Weight）	自然数（克）	BWT

图 3　数据变量解释图

Lowbwt

	id	low	age	lwt	race	smoke	ptl	ht	ui	ftv	bwt
1	4	< 2500 g	28	120	Other	Yes	One	No	Yes	None	709
2	10	< 2500 g	29	130	White	No	None	No	Yes	Two, etc.	1021
3	11	< 2500 g	34	187	Black	Yes	None	Yes	No	None	1135
4	13	< 2500 g	25	105	Other	No	One	Yes	No	None	1330
5	15	< 2500 g	25	85	Other	No	None	No	Yes	None	1474

图 4　数据的前 5 条事例图

2.2 数据简要分析与预处理

如前面的变量介绍表所示，该数据共涉及母亲的 11 种变量：其中 ID 为母亲的编号；LOW 是 Low birth weight ，实际就是对出生时婴儿体重的一个判断，分为小于 2500g 和大于 2500g；AGE 是母亲年龄；LWT 是母亲最后一次月经期间的体重；RACE 是母亲的肤色，由白色，黑色以及其它颜色三种类型；SMOKE 是母亲是否吸烟；PTL 是早产史，分为 0 次，1 次，2 次及以上共三种类型；HT 是母亲是否有高血压史；UI 是母亲是否存在子宫过敏；FTV 是母亲在怀孕妊娠期第一个季度的医生就诊次数；BWT 是婴儿出生时的体重。以后为了方便起见，全部用代号里的大写字母来代替相应的变量。

首先，ID 肯定不是风险因子，它只是研究人员为了方便给被调查者贴上的编号；LOW 和 BWT 是结果，也就是我们回归模型的自变量，LOW 其实是二值虚拟变量,取值为 0，1，取 0 时代表没有患病，取 1 时代表患病，BTW 则是婴儿出生时的体重，特别是我们做线性回归时，它相当于模型中的 Y。 对于 “SMOKE”,, “HT”, “UI”, 这 3 个变量，因为他们都是虚拟变量，所以我用“1”代替“是”，“0”代替“否” 。对于“PTL”“FTV”这两个变量，它们两个也是虚拟变量， 我用“0” 代表 0 次，“1”代表 1 次，“2”代表 2 次及以上。对于“RACE”，我令 1 = 白, 2 = 黑, 3 = 其它。最后剩下“AGE”和“LWT”，它们本身就是数值所以不用处理。 故最终我得到的处理数据为以下样式：

已处理数据

ID	LOW	AGE	LWT	RACE	SMOKE	PTL	HT	UI	FTV	BWT
4	1	28	120	3	1	1	0	1	0	709
10	1	29	130	1	0	0	0	1	2	1021
11	1	34	187	2	1	0	1	0	0	1135
13	1	25	105	3	0	1	1	0	0	1330

图 5　已处理数据示意图

2.3 数据的描述性统计分析

2.3.1 描述性分析

首先对所有变量进行描述性统计分析，得到以下结果：

```
> summary(data_new)
       ID              LOW              AGE             LWT             RACE           SMOKE       
 Min.   :  4.0   Min.   :0.0000   Min.   :14.00   Min.   : 80.0   Min.   :1.000   Min.   :0.0000  
 1st Qu.: 68.0   1st Qu.:0.0000   1st Qu.:19.00   1st Qu.:110.0   1st Qu.:1.000   1st Qu.:0.0000  
 Median :123.0   Median :0.0000   Median :23.00   Median :121.0   Median :1.000   Median :0.0000  
 Mean   :121.1   Mean   :0.3122   Mean   :23.24   Mean   :129.8   Mean   :1.847   Mean   :0.3915  
 3rd Qu.:176.0   3rd Qu.:1.0000   3rd Qu.:26.00   3rd Qu.:140.0   3rd Qu.:3.000   3rd Qu.:1.0000  
 Max.   :226.0   Max.   :1.0000   Max.   :45.00   Max.   :250.0   Max.   :3.000   Max.   :1.0000  
      PTL               HT               UI              FTV              BWT        
 Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   : 709  
 1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:2414  
 Median :0.0000   Median :0.00000   Median :0.0000   Median :0.0000   Median :2977  
 Mean   :0.1905   Mean   :0.06349   Mean   :0.1481   Mean   :0.6931   Mean   :2945  
 3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:3475  
 Max.   :2.0000   Max.   :1.00000   Max.   :1.0000   Max.   :2.0000   Max.   :4990  
```

图 6 所有变量描述性统计示意图

从这个结果我们并不能得到太多有效的信息，因为其中除了“AGE”，“LWT”，“BWT” 这三个变量，其他的都是虚拟变量。

2.3.2 各变量相关性分析

为了研究各个变量的相关性，我们做出各个变量的散点图矩阵：

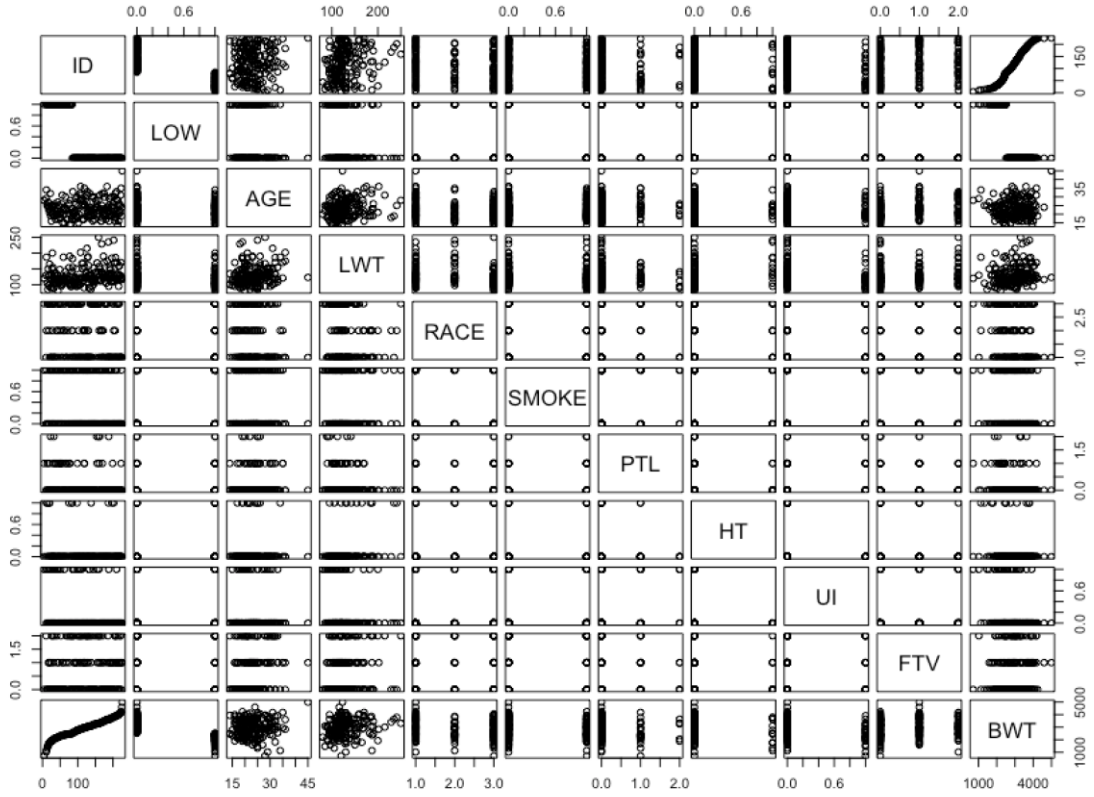

图 7 各变量散点图矩阵

与描述性统计分析类似，我们很难从这个散点图矩阵（图 7）得到太多有效信息，原因也是因为变量中的隐性变量太多，无法直接从散点图中看出变量变化的趋势。故现在考虑先研究“AGE”，“LWT”， “BWT” 这三个变量。

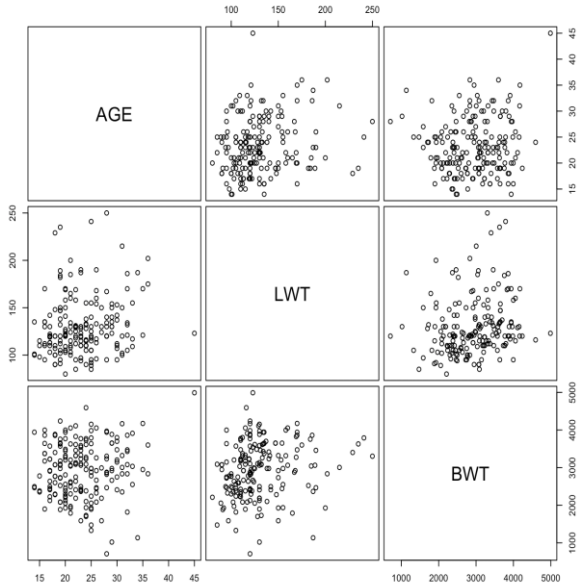

图 8 “AGE”，“LWT”， “BWT”三个变量的散点图矩阵

其中，“BWT” 是被解释变量，“AGE”，“LWT”是解释变量。从这个散点图矩阵（图 8）中，我们很难发现这三个变量存在线性关系，这是可以理解的，因为决定“BWT”可能是很多的因素，如果只关注“AGE”，“LWT”这两个变量，相当于缺失了很多关键变量。所以这个结果并不能说明“AGE”，“LWT”不是风险因子。因为在医学常识上来说，母亲年龄越高，孩子的发病率更高，故我们需要对“AGE”，“LWT”进行进一步的研究。

2.3.3 基于 A/B test 的研究

A/B test 主要研究两个样本是否来自于同一个总体。首先将样本分为 A 组和 B 组，这里样本 A 是患病的母亲（LOW=1），样本 B 是没有患病母亲（LOW=0），注意这里研究的是“LOW”这个被解释变量了，这是一个二之变量，LOW=1 表示患病，LOW=0 表示没有患病。在这里分别研究“AGE”和“LWT”这两种情况，根本研究的问题就是：患病组与非患病组的“AGE”，“LWT”是否有显著差异。

（1） 先对“AGE”研究

首先做出 A 组和 B 组的“AGE”的频率分布直方图：

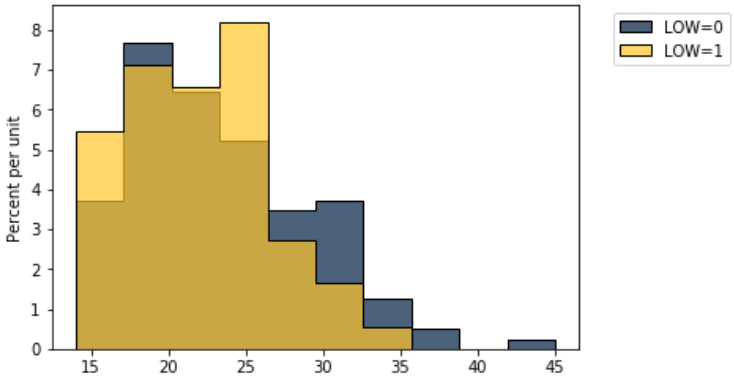

图 9 患病组与非患病组“AGE”频率分布直方图

从图中发现两个样本并没有明显差异。下一步进行 A/B test：

Ho：A 组和 B 组的“AGE”来自于同一个总体，两者的均值没有差别。

H1：A 组和 B 组的“AGE”来不自于同一个总体，两者的均值有差别

统计量：A 组的均值与 B 组均值的差

结果如下：

观测值： -1.35645371577575
p值： 0.047

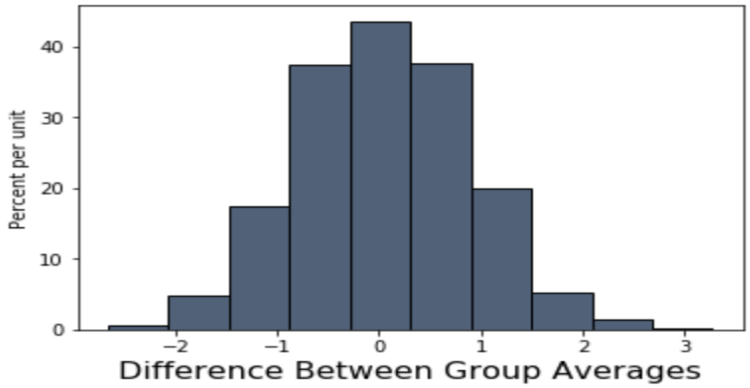

图 10　A/B 假设检验统计量频率分布直方图

因为 p 值为 0.047，所以我们在 95%的置信水平上拒绝原假设，也就是说，我们拒绝患病与非患病的母亲年龄上没有差异这一假设，换句话说，患病与非患病的母亲在年龄分布上是有差异的，不是来自于同一个总体的。这与我们的预判一致，因为医学上的普遍认知是高龄产妇一般生产有更大的风险，伴随的婴儿患病风险也更大一些。

（2）对“LWT”研究

类似于前面对“AGE”研究，这里直接给出对“LWT”研究的结果：

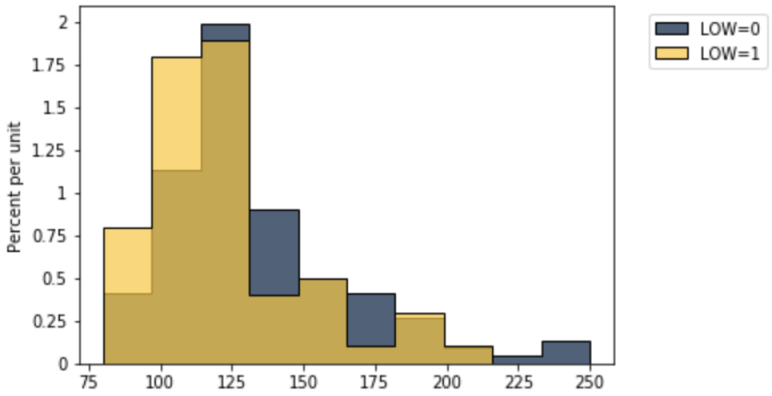

图 11　患病组与非患病组“LWT”频率分布直方图

```
观测值： -11.164406779661022
p值： 0.004
```

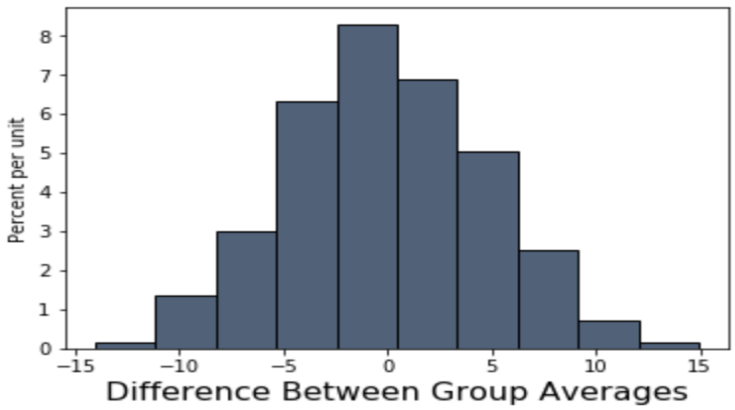

图 12　A/B 假设检验统计量频率分布直方图

因为 p 值为 0.004，所以我们在 99%的置信水平上拒绝原假设，这个显著性更大，也就是说，我们拒绝患病与非患病的母亲在最后一次月经期的体重上没有差异这一假设，换句话说，患病与非患病的母亲在最后一次月经期的体重上分布上是有显著差异的。

分析了“AGE”，“LWT”这两个变量，由 A/B test 我们发现这两个变量应该是与是否患病有关的。下面，我们将用统计的一些回归模型对剩下的虚拟变量进行进一步分析。

3. 实证分析与比较

3.1 基于线性统计模型的诊断

3.1.1 模型构建

多元回归模型有以下形式：

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

其中 Y 为因变量，X 为解释变量。这里我们取“BWT”为因变量，也就是婴儿出生时的体重，剩下的变量为解释变量“SMOKE”，“HT”，“UI”，“PTL”，“FTV”，“AGE”，“RACE”，“LWT”，来进行模型拟合，目的是求出每个变量对模型的显著性，从而判断各个因子对婴儿出生时的体重的影响。考虑到解释变量里面有 5 个是虚拟变量，而 R 可以直接对虚拟变量进行转换，所以使用的是原始数据。

3.1.2 回归结果与解释分析

回归结果如下：

```
Call:
lm(formula = bwt ~ age + lwt + race + smoke + ptl + ht + ui +
    ftv, data = Lowbwt)

Residuals:
     Min       1Q   Median       3Q      Max
-1779.89  -428.59    59.96   438.73  1591.26

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2381.928    325.279   7.323 8.23e-12 ***
age            -2.123      9.629  -0.220 0.825785
lwt             4.461      1.718   2.597 0.010185 *
raceOther     142.244    158.787   0.896 0.371566
raceWhite     445.607    149.460   2.981 0.003273 **
smokeYes     -296.662    109.143  -2.718 0.007219 **
ptlOne       -323.552    149.062  -2.171 0.031294 *
ptlTwo, etc.  152.118    277.229   0.549 0.583897
htYes        -580.663    200.376  -2.898 0.004232 **
uiYes        -502.513    137.319  -3.659 0.000333 ***
ftvOne        118.018    122.486   0.964 0.336600
ftvTwo, etc.  -42.513    122.906  -0.346 0.729829
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 643.1 on 177 degrees of freedom
Multiple R-squared:  0.2673,    Adjusted R-squared:  0.2217
F-statistic:  5.87 on 11 and 177 DF,  p-value: 4.289e-08
```

图 13 线性回归模型总结图

最终模型为：

BWT = 2381.9 - 2.1*AGE + 4.5*LWT + 142.2*RACE_other + 445.6*RACE_white - 296.7*SOMKE_yes - 323.6*PTL_one + 152.1*PTL_two.etc - 580.7*HT_yes - 502.5*UI_yes + 118.0*FTV_one - 42.5*FTV_two.etc + e

分析 p 值，发现有较高显著水平的变量为：“LWT”，“RACE_white”，“SMOKE_yes”，“PTL_one”，“HT_yes”，“UI_yes”以及截距。其余的变量 p 值都至少超过了 0.5，所以无法拒绝原假设，也就是说在模型中这些变量对自变量“BWT”对影响不显著。

“LWT” 的 p 值为 0.01，拟合系数为 4.461，意思是母亲在最后一次月经期的体重每增加 1 磅，那么出生婴儿的体重将增加 4.461 克；“RACE_white”的 p 值是 0.003，拟合系数为 445.6，意思是白人母亲相较于其他肤色的母亲，她们的孩子出生体重要重 445.6 克，直观上来讲，就是白人母亲的孩子体重往往更重，也就是说患低出生体重的风险要更小一些；“SMOKE_yes”的 p 值是 0.007，拟合系数为-296.7 克，意思是吸烟的母亲比不吸烟的母亲，她们的孩子要轻 296.7 克，就是说吸烟的母亲生下的孩子体重往往更低，也就是说患低出生体重的风险要更大一些；“PTL_one”的 p 值是 0.03，拟合系数为-323.6，意思是有一次早产史的孕妇，她们的婴儿体重比正常的要轻 323.6 克，也就是说有早产史的孕妇，她们孩子患有低出生体重的风险要更大一些；“HT_yes”的 p 值是 0.004，拟合系数为-580.7，意思是有高血压史的母亲，她们的孩子相对于没有高血压史的母亲体重要轻 580.7 克，也就是说高血压史的母亲生下来的孩子患低出生体重的风险要更大一些；“UI_yes”的 p 值是 0.0003，拟合系数为-502.5，意思是存在子宫过敏的母亲，她们的孩子比正常母亲要轻 502.5 克，也就是说子宫过敏的母亲的孩子患有低出生体重的风险要大。

3.1.3 模型评价

一般线性模型，我们选择分析拟合优度，就是 R^2。这里 R^2 为 0.27，意味着这个模型只能解释数据的 27%的方差。说明从拟合优度的角度来说，这个模型拟合效果并不好。

残差分析：

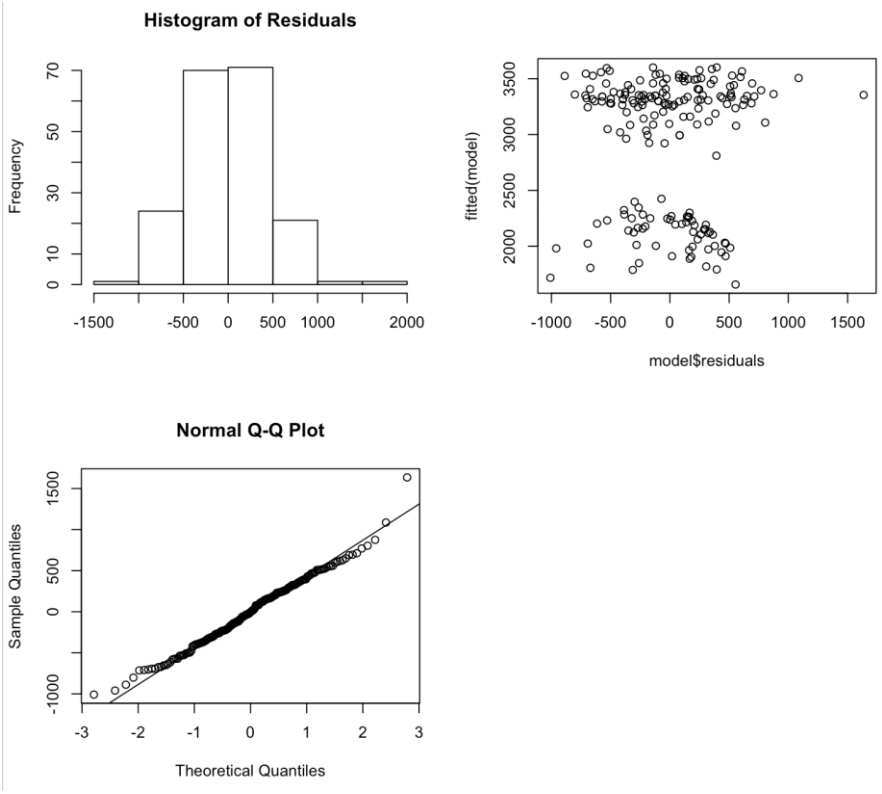

图 14　线性统计模型残差图

可以看出残差比较类似正态分布，Q-Q plot 确认了我们能用这个模型进行预测。

上一节我们算出了 8 个解释变量的 p 值，其中有 2 个变量的拟合系数是不显著的，把它们删去。如果我们选择只对“LWT”，“RACE”，“SMOKE”，“PTL”，“HT”，“UI”这 6 个显著的变量进行回归的话，结果如下：

```
Call:
lm(formula = bwt ~ lwt + race + smoke + ptl + ht + ui, data = Lowbwt)

Residuals:
     Min       1Q   Median       3Q      Max
-1858.88  -424.36    53.36   486.96  1626.24

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2401.187    280.608   8.557 4.97e-15 ***
lwt             4.180      1.668   2.506  0.01310 *
raceOther     124.183    156.319   0.794  0.42800
raceWhite     448.478    144.893   3.095  0.00228 **
smokeYes     -323.487    104.481  -3.096  0.00227 **
ptlOne       -299.869    144.285  -2.078  0.03910 *
ptlTwo, etc.  182.040    274.274   0.664  0.50772
htYes        -566.221    198.114  -2.858  0.00477 **
uiYes        -513.139    136.062  -3.771  0.00022 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 640.4 on 180 degrees of freedom
Multiple R-squared:  0.2612,    Adjusted R-squared:  0.2284
F-statistic: 7.956 on 8 and 180 DF,  p-value: 3.729e-09
```

图 15 受约束模型线性回归模型总结图

模型则变为：

BTW = 2401.2 + 4.2*LWT + 124.2*RACE_other + 448.5*RACE_white - 323.5*SOMKE_Yes - 300*PTL_one + 182*PTL_two.etc - 566.2*HT_yes - 513.1*UI_yes

比较两个模型发现，拟合系数的估计值相差不大，R^2 也只相差 0.006 ，说明原模型拟合效果较好。这个模型相当于设立了排除性约束的联合假设检验：

Ho：“AGE”和“FTV”的拟合系数都为 0

H1：“AGE”和“FTV”的拟合系数不同时为 0

统计量：

$$\mathbf{F} = \frac{(\boldsymbol{SSR_R} - \boldsymbol{SSR_{UR}})/\boldsymbol{q}}{\boldsymbol{SSR_{UR}}/(\mathbf{n} - \mathbf{k} - \mathbf{1})}$$

其中SSR_R是受约束模型残差平方和，SSR_{UR}是非受约束模型残差平方和，q 是约束系数的个数，这里 q 是 2，n 是样本量，这里 n 是 189，k 是自变量数量，这里 k 是 6。

由此，算出 F 统计量为 0.749，故无法显著的拒绝“AGE”和“FTV”的拟合系数都为 0，也就是说“AGE”和“FTV”这两个变量是联合对自变量“BTW”产生影响的。进一步分析，说明“AGE”和“FTV”这两个变量存在相关性，它们可能是共同对出生婴儿体重产生影响的。

3.2 基于逻辑回归模型的诊断

前面的一般线性模型研究的因变量 Y“BWT”是一个常规变量，含义是婴儿出生时的体重时的体重，事实上我们得到模型后，能够拟合出婴儿的出生体重，但是我们还需要将其与 2500 克比较来判断是否患病。而且对拟合的参数，我们得到的也仅仅是一个解释变量的改变量，能引起婴儿的出生体重的改变量，不能告诉我们婴儿的患病概率会增加多少。因此，我们现在研究因变量“LOW”，这是一个二值虚拟变量,取值为 0，1，取 0 时代表没有患病，取 1 时代表患病。其它解释变量不变。当我们在考虑 Y 是一个二值虚拟变量时，往往会使用逻辑回归。

3.2.1 模型构建

当我们在考虑 Y 是一个二值虚拟变量时，往往会使用逻辑回归。

公式为：

$$P(Y = 1) = 1 / (1 + \exp[-(B_0 + B_1x_1 + B_2X_2 + B_3X_3 + \ldots + B_pX_p)])$$

更加简化来说，就是 f(z) = 1 / (1 + exp(-z))，z 是我们能见到的常规线性回归的预测。 f(z) 的表现为：

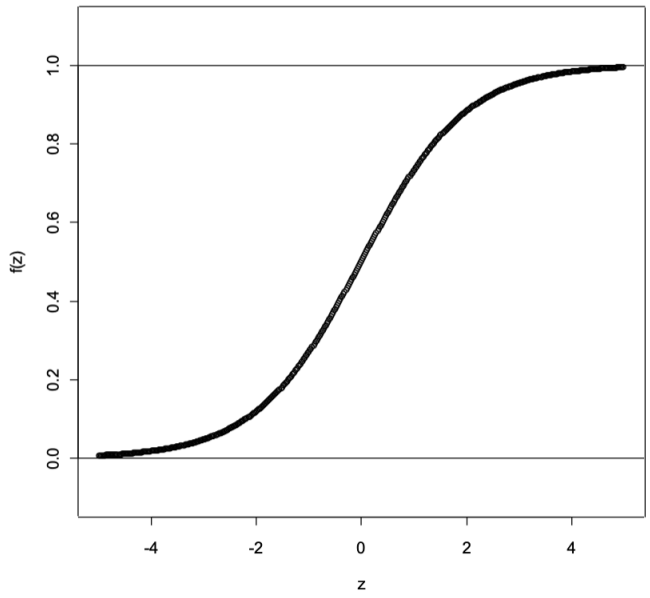

图 15　逻辑回归示意图

这里的 Y 我们取“LOW”，其取值为 0，1，取 0 时代表没有患病，取 1 时代表患病，X 取变量“SMOKE”，“HT”，“UI”，“PTL”，“FTV”，“AGE”，“RACE”，“LWT” 来进行逻辑回归。目的是求出各个解释变量对应的拟合系数，对患病风险进行进一步评估。

3.2.2 回归结果与解释分析

回归结果如下：

```
Call:
glm(formula = LOW ~ age + lwt + race + smoke + ptl + ht + ui +
    ftv, family = binomial, data = logit_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7722  -0.8040  -0.5045   0.8746   2.2231

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.15887    1.31940   1.636  0.10179
age          -0.04079    0.03922  -1.040  0.29832
lwt          -0.01636    0.00721  -2.270  0.02323 *
raceOther    -0.42864    0.56012  -0.765  0.44412
raceWhite    -1.12251    0.54311  -2.067  0.03875 *
smokeYes      0.75024    0.43167   1.738  0.08221 .
ptlOne        1.71542    0.54301   3.159  0.00158 **
ptlTwo, etc. -0.02002    0.96939  -0.021  0.98352
htYes         1.90929    0.72963   2.617  0.00888 **
uiYes         0.75203    0.47273   1.591  0.11165
ftvOne       -0.48603    0.48814  -0.996  0.31941
ftvTwo, etc.  0.11418    0.46233   0.247  0.80494
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 192.54  on 177  degrees of freedom
AIC: 216.54

Number of Fisher Scoring iterations: 4
```

图 16　逻辑回归结果示意图

在这个逻辑回归模型中， 较为显著的有“LWT”，“RACE_white”，“PTL_one”，“HT_yes”，“SMOKE_yes”，都是 p 值至少小于 0.1 的，也就是说至少在 90%的置信水平上显著。还有一个“UI_yes”，p 值为 0.11 。

为了理解每个系数的含义，我们需要对每个系数取自然指数（exp(拟合系数)），因为每个系数是仅与该变量相关联的对数几率。 截距是每个人开始的基本对数几率（很像线性回归中的截距）， 负对数赔率会降低风险，正对数赔率会增加风险。 0 是中性的。所以我们对每个变量的拟合参数取对数：

```
> exp(bwt$coef)
 (Intercept)          age          lwt    raceOther    raceWhite     smokeYes       ptlOne
   8.6612974    0.9600292    0.9837702    0.6513961    0.3254610    2.1175091    5.5589861
ptlTwo, etc.        htYes        uiYes       ftvOne ftvTwo, etc.
   0.9801773    6.7483180    2.1213104    0.6150656    1.1209489
```

图 17　拟合参数取对数示意图

这里我们引入比值比（odds ratio）这个概念，比值比是暴露与结果之间关联的度量

$$odds = P(y=1)/P(y=0)$$

比代表在给定特定暴露时结果将发生的几率，与在没有暴露的情况下发生的结果的几率相比。在这种情况下，我们使用每个变量系数的指数来生成比值比我们可以看到每单位变量的增加或减少对获得低出生体重的几率的影响。

例如：“LWT”的比值比（odds ratio）是 0.98，意思是母亲在最后一次月经期的体重每增加一磅，那么婴儿的患病可能将下降 2%；“RACE_white” 的比值比（odds ratio）为 0.65，意思是如果母亲是白人，那么相较于其它人种，婴儿的患病可能将下降 35%；“PTL_one” 的比值比（odds ratio）是 5.55，意思是如果母亲有过一次早产史，那么相对于其它情况婴儿患病可能将上升 450%；“HT_yes”的比值比（odds ratio）为 6.75，意思是如果母亲是有高血压史的，那么她的孩子相对于别人，患病可能将增大 575%；“SMOKE_yes”的比值比（odds ratio）为 2.12，意思是如果母亲吸烟，那么相对于不吸烟的母亲，她孩子患病风险将增大 112%。“UI_yes”的比值比（odds ratio）是 2.12，意思是如果母亲存在子宫过敏，那么相较于正常母亲，她孩子的患病概率将增大 112% 。

置信区间：

	est	upper.ci	lower.ci
(Intercept)	8.6612974	114.9962146	0.6523525
age	0.9600292	1.0367408	0.8889937
lwt	0.9837702	0.9977703	0.9699666
raceOther	0.6513961	1.9526851	0.2172992
raceWhite	0.3254610	0.9436353	0.1122519
smokeYes	2.1175091	4.9347830	0.9086205
ptlOne	5.5589861	16.1143225	1.9176932
ptlTwo, etc.	0.9801773	6.5533219	0.1466046
htYes	6.7483180	28.2012917	1.6148124
uiYes	2.1213104	5.3580085	0.8398564
ftvOne	0.6150656	1.6011597	0.2362698
ftvTwo, etc.	1.1209489	2.7741244	0.4529453

图 18　拟合参数的置信区间

3.2.2 模型评价

（1）基于 AIC 选择解释变量

评价一个逻辑回归模型的好坏，首先是分析这个模型的 AIC 信息准则，AIC 是衡量统计模型拟合优良性(Goodness of fit)的一种标准，一般来说，AIC 的值越小，说明模型的拟合效果越好。我们用 R 的 stepwise 逻辑回归函数，这个函数可以选择出预测低出生体重最好的解释变量组合：

```
> bwt.step<-step(bwt)
Start:  AIC=216.54
LOW ~ age + lwt + race + smoke + ptl + ht + ui + ftv

        Df Deviance    AIC
- ftv    2   193.91 213.91
- age    1   193.64 215.64
<none>       192.54 216.54
- ui     1   195.03 217.03
- race   2   197.45 217.45
- smoke  1   195.59 217.59
- lwt    1   198.34 220.34
- ht     1   199.78 221.78
- ptl    2   203.58 223.58

Step:  AIC=213.91
LOW ~ age + lwt + race + smoke + ptl + ht + ui

        Df Deviance    AIC
- age    1   195.16 213.16
<none>       193.91 213.91
- ui     1   196.68 214.68
- race   2   199.46 215.46
- smoke  1   198.23 216.23
- lwt    1   199.23 217.23
- ht     1   200.93 218.93
- ptl    2   203.95 219.95

Step:  AIC=213.16
LOW ~ lwt + race + smoke + ptl + ht + ui

        Df Deviance    AIC
<none>       195.16 213.16
- ui     1   198.27 214.27
- smoke  1   199.85 215.85
- race   2   202.03 216.03
- lwt    1   201.61 217.61
- ptl    2   204.22 218.22
- ht     1   202.35 218.35
```

图 19 寻找最低的 AIC 示意图

原始模型的 AIC 为 216.54，经过解释变量筛选之后 AIC 变为 213.16.我们发现解释变量比原模型少了“AGE”和“FTV”。生下的变量为“LWT”，“RACE”，“SMOKE”，“PTL”，“HT”，“UI”。这是符合预期的，因为在上一节的分析中，只有这 6 个变量 p 值较低，也就是显著性较高，而另外 2 个解释变量的 p

值较高。故如果考虑新的模型的话，也是先删减掉这两个解释变量。于是模型变成了 LOW = lwt + race + smoke + ptl + ht + ui。 此时的 AIC 最小，为 213.16 。

（2）逻辑回归模型的预测与评价

现在，我们选用这个 AIC 最小的模型进行进一步分析。首先我们引入一个叫混淆矩阵（confusion matrix）的概念，它通常用于描述已知真值的一组测试数据的分类模型（或“分类器”）的性能。例如：

表 1　混淆矩阵事例

N 样本总量	预测值为 0	预测值为 1
真实值为 0	N1	N2
真实值为 1	N3	N4

其中 N = N1 + N2 + N3 + N4

现在定义 4 个值：TP,TN,FP,FN 。其中，TP 指预测与真实值都是 1 的样本，也就是 N4；TN 指预测与真实值都是 0，也就是 N1；FP 指预测为 1 但是实际是 0 的，也就是 N2；FN 是预测为 0 但是实际是 1 的，也就是 N1 。

定义：

敏感度（Sensitivity ）= TP/(TP + FN)
特异性（Specificity）= TN/(TN + FP)
准确度（Accuracy）= (TP+TN) / N

首先我们要确定一个基准模型，我们知道在这 189 个数据中，N3 = 130，所以准确率 = 130/189 =0.69 。所以这个准确率也就是我们企图通过我们的逻辑回归模型击败的准确率。

现在我们把原始 189 条数据随机分为训练集和测试集，分离的比例为 0.69，所以训练集为 131 名母亲的数据，测试集为 58 名母亲的数据 。然后我们用训练集来训练我们的逻辑回归模型。

模型拟合结果如下：

```
Call:
glm(formula = LOW ~ lwt + race + smoke + ptl + ht + ui, family = binomial,
    data = qualityTrain)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5949  -0.8156  -0.6016   0.9455   2.0396

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.702359   1.188619   0.591   0.5546
lwt          -0.012886   0.007554  -1.706   0.0880 .
raceOther    -0.258223   0.614147  -0.420   0.6742
raceWhite    -1.334554   0.610409  -2.186   0.0288 *
smokeYes      1.201961   0.502218   2.393   0.0167 *
ptlOne        0.516865   0.647406   0.798   0.4247
ptlTwo, etc. -0.147230   0.999807  -0.147   0.8829
htYes         1.723035   0.775889   2.221   0.0264 *
uiYes         0.754238   0.558997   1.349   0.1773
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 162.82  on 130  degrees of freedom
Residual deviance: 140.73  on 122  degrees of freedom
AIC: 158.73

Number of Fisher Scoring iterations: 4
```

图 20 训练集数据拟合结果

模型的 AIC 值为 158.73，确实比原来的模型的 216.54 要小很多。说明这个新的模型拟合效果要更好。

对训练集的概率预测：

```
> predictTrain = predict(QualityLog, type="response")
> summary(predictTrain)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.03786 0.19406 0.27422 0.31298 0.41388 0.79752
```

图 21 训练集数据概率预测分位数

我们可以使用所谓的阈值 t 将概率转换为预测。 如果患病的可能性大于此阈值 t，我们预测该婴儿患有低出生体重疾病。 但如果患病的可能性小于阈值 t，那么我们就会预测该婴儿没有患病。为了探求不同阈值对敏感度和特异性的影响，我们分别对阈值为 0.2，0.5，0.7 进行了测试：

```
> # Confusion matrix for threshold of 0.2
> table(qualityTrain$LOW, predictTrain > 0.2)

    FALSE TRUE
  0    29   61
  1     4   37
> # Sensitivity
> 37/(4+37)
[1] 0.902439
> # Specificity
> 29/(29+61)
[1] 0.3222222
```

图 22 阈值为 0.2 的混淆矩阵

```
> table(qualityTrain$LOW, predictTrain > 0.5)

    FALSE TRUE
  0    82    8
  1    27   14
> #Sensitivity
> 14/(27+14)
[1] 0.3414634
> # Specificity
> 82/(82+8)
[1] 0.9111111
```

图 23 阈值为 0.5 的混淆矩阵

```
> # Confusion matrix for threshold of 0.7
> table(qualityTrain$LOW, predictTrain > 0.7)

    FALSE TRUE
  0    88    2
  1    36    5
> # Sensitivity
> 5/(36+5)
[1] 0.1219512
> # Specificity
> 88/90
[1] 0.9777778
```

图 24 阈值为 0.7 的混淆矩阵

我们发现随着阈值的升高，敏感度在下降，而特异性在上升。下面，我们使用 ROC 曲线图来选择阈值。该模型的 ROC 曲线图如下所示：

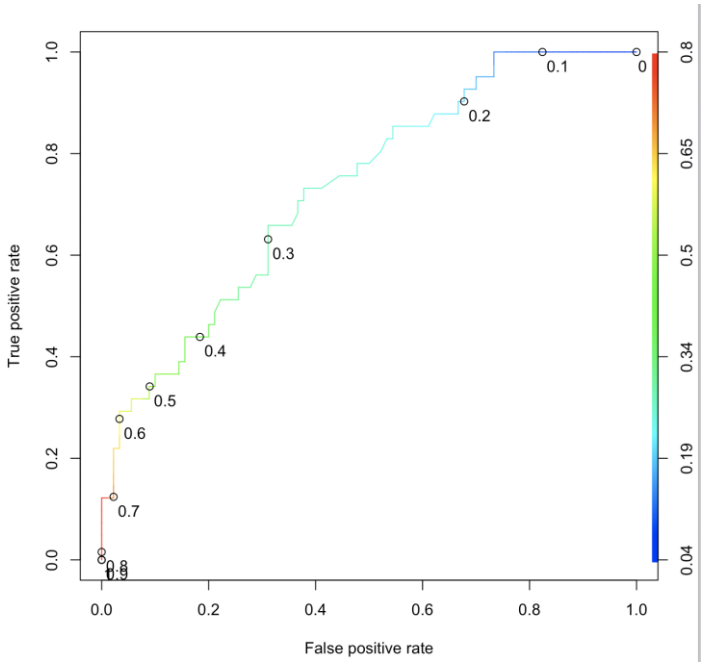

图 25 该模型的 ROC 曲线

我们肯定希望 TP 尽量大，意思就是如果一个人真的有病，那么我们更希望能尽大的可能预测出他有病，反而对 TN 我们不需要太关注；而对于两种误判，我们肯定更关心的是 FN，这种错误是指一个人实际上是有病的，但是被误判为没有病，这样往往会耽误治疗，所以我们希望 FN 越小越好；同样 FP 我们也会关心，FP 的也是越小越好。故从 ROC 图中，阈值取 0.4 是较为理想的。也就是说，在测试集中，当预测值大于 0.4，我们就判定为患病，当预测值小于 0.4，则判定为没有患病。

现在我们用测试集就行检验：

```
> table(qualityTest$LOW,predictTest >= 0.4)

   FALSE TRUE
0     36    4
1      8   10
```

图 26 测试集检验的混淆矩阵

所以准确率为(36+10) / 58 = 79.31%

综上所述，我们的逻辑回归模型预测患病的准确率为 79.31% ，大于了基准模型的 69%，也就是说我们的逻辑回归模型是有效的。

3.3 两种模型的比较

逻辑回归模型与线性回归模型都是广义线性回归模型。它们最关键的区别就是因变量 Y 的数据类型不同。逻辑回归主要处理因变量 Y 是二值虚拟变量，比如这里的婴儿是否患病，而一般线性回归模型主要处理的变量 Y 是连续变量，比如这里的婴儿出生时的体重。因为这两个模型的因变量 Y 不同，所以我们无法直接比较这两个模型，比如比较他们 AIC 值等，是没有意义的。

两个模型的最终目的都是达到一种分类器的效果，而一个分类器，最重要的是它的准确性。所以我们可以从这个角度来比较这两个模型。

3.3.1 解释变量的比较

两个回归模型一开始使用的解释变量都是一样的，虽然因变量 Y 不同，但是两者的 Y 都是由解释变量“SMOKE”，“HT”，“UI”，“PTL”，“FTV”，“AGE”，“RACE”，“LWT”进行拟合的。经过第一次拟合之后，我们删去了显著性较低的解释变量，两者的模型的效果都有提高。值得一提的是，两个模型中，都是“FTV”，“AGE”这两个变量的显著性较低并且被删去，这说明了原始数据的真实性和有效性。也说明了确实这两个变量与婴儿是否患病或者与婴儿出生时的体重没有太大的关系。

3.3.2 分类器准确率的比较

在前面我们已经计算了逻辑回归模型的准确率，现在我们计算线性模型的准确率，这里线性模型的预测值(predict value)是婴儿出生时的体重，所以我们要进一步判断婴儿是否有得病。若预测值小于 2500 克，则为患病，反之则是没有患病。

为了保持公平，同样把原始 189 条数据随机分为训练集和测试集，分离的比例为 0.69，所以训练集为 131 名母亲的数据，测试集为 58 名母亲的数据。在训练集上完成模型构建之后，计算该模型在测试集上的预测值：（共有 58 个数据）

```
> Pred
       1        2        3        4        5        6        7        8        9       10       11       12       13       14       15       16
2413.739 3164.730 2843.405 2625.063 3008.728 2546.506 2434.078 2811.368 3520.479 2319.274 2840.817 2318.802 2289.401 2985.494 2761.581 3427.541
      17       18       19       20       21       22       23       24       25       26       27       28       29       30       31       32
2788.134 2362.682 2420.801 3008.728 2982.174 2929.491 2929.491 2859.106 3065.154 3560.309 2968.167 3576.905 3062.259 3009.152 3311.198 2742.289
      33       34       35       36       37       38       39       40       41       42       43       44       45       46       47       48
3500.564 3444.137 2940.558 3048.558 2031.860 3174.687 3061.835 3031.962 3048.558 3162.582 2699.871 3045.663 3353.192 3527.117 2867.794 3497.244
      49       50       51       52       53       54       55       56       57       58
3626.693 2600.296 3493.925 3477.329 3503.883 3111.623 3623.373 3029.067 3161.834 3470.691
```

图 27 测试集在模型上的预测值

再将这些预测值进行进一步判断，也就是是否患病的判断：

```
> Pred < 2500
    1     2     3     4     5     6     7     8     9    10    11    12    13    14    15    16    17    18    19    20    21    22    23    24
 TRUE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE
   25    26    27    28    29    30    31    32    33    34    35    36    37    38    39    40    41    42    43    44    45    46    47    48
FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
   49    50    51    52    53    54    55    56    57    58
FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

图 28 测试集的预测值转化为布尔变量

而测试集的真实值为：

```
> Test$low == '< 2500 g'
 [1]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE
[24] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[47] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

图 29 测试集上布尔变量的真实值

比对上面两个列向量，若相同指数对应的布尔变量相同，则算预测成功算出准确率为 44/58 = 75.86% 。这个准确率要高于基准模型的准确率 69%，说明线性模型是有效的，但是这个准确率低于逻辑回归模型的 79.31%，说明在这个数据集上，逻辑回归模型的预测结果要比线性回归模型的效果要更好。

4. 结论与建议

前面我们利用一组有效的横截面数据，研究了 8 个风险因子对低体重出生的影响，这 8 个风险因子分别是：母亲年龄（AGE），母亲在最后一次月经期的体重（LWT），母亲肤色（RACE），母亲是否吸烟（SMOKE），早产史（PTL），高血压史（HT），是否存在子宫过敏（UI），第一个季度的医生就诊次数（FTV）。因变量有两个，一个是是否患病（LOW），另一个是婴儿出生时的体重（BWT）。根据因变量不同的性质，我们对因变量(LOW)使用逻辑回归模型，对因变量(BWT)使用线性回归模型。两个模型的回归结果没有较大的出入，说明这个数据是准确可靠的，也说明该数据集收集的变量是有效的。这对我们研究低体重出生有着极大的意义。

4.1 主要结论

根据逻辑回归和线性回归的结果。我们发现母亲年龄（AGE）和第一个季度的医生就诊次数（FTV）这两个变量对是否患病影响不大，因为这两个变量的在模型中的显著性都不强，而且在删去这这两个变量之后，模型的拟合效果都有不同程度的提高。

剩下的 6 个风险因子中，我们发现能使患病风险升高的因素是母亲抽烟（SMOKE），有早产史（PTL），有高血压史（HT），存在子宫过敏（UI），从升高风险的程度来看，是“HT” > “PTL” > “SMOKE” > “UI”，也就是说患有高血压的母亲孩子患病的风险最大，其次是有早产史的母亲，母亲吸烟与存在子宫过敏的两种情况，对患病的风险概率的贡献是差不多的；患病风险较低的为白人肤色的母亲（RACE），以及母亲在最后一次月经期有较高的体重（LWT）。

从模型的角度，线性回归模型与逻辑回归模型都有有效的预测效果，即可以根据母亲的以上变量的参数，来预测孩子的患病可能。线性回归虽然有一些不寻常的数据点，但它们的影响并不是很显着，因此我们可以依赖于线性拟合的推论。当我们采用逻辑回归模型时，我们得到完全相同的结论。比较准确率来说，逻辑回归预测的准确率要高于线性回归模型，逻辑回归预测的准确率甚至高达 80%。说明在这种有虚拟变量的回归情况下，逻辑回归模型可能是更加适合的。

4.2 相关建议

低出生体重确实是危害婴儿以及其后续发育的较为严重的疾病，所以未来对该疾病的防范与预防显得尤为重要。从本文的角度，我们建议有高血压以及有早产史的母亲一定要做好产前检查，时刻关注腹中婴儿的健康状况，另外不提倡母亲吸烟以及在妊娠期间控制体重，因为要保证在最后一次月经期时母亲的体重不能过低。

当然，以上几点只是从这个数据集中得来的经验，事实上可能还有很多重要因素没有被考虑到，因为线性回归的拟合优度只有 0.26。所以，这些东西还远远不够，所以未来对该疾病的探索仍然任重而道远。

参考文献

[1] WTO. Guidelines on Optimal feeding of low birth-weight infants in low-and middle-income countries，WTO，2011: 5-6

[2] 李鸿斌，顾建明，丁晓丽，等. 上世纪八十年代中国婴儿死亡率的调整与矫正. 中国卫生统计，2015，32（3）：刊用

[3] 冯江，袁秀琴，朱军，等. 中国 2000—2010 年 5 岁以下儿童死亡抽样调查[J]. 中华儿科杂志，1994，32（3）：149-152

附录

附录 A

致谢

在这里衷心感谢我的毕业论文指导老师蒋学军老师对我在毕业设计期间提供的帮助。