Statistical and Application 统计学与应用, 2015, 4(3), 144-154
Published Online September 2015 in Hans. http://www.hanspub.org/journal/sa
http://dx.doi.org/10.12677/sa.2015.43016

Hans汉斯

The Analysis of Influence Factors of Single Box Consumption Based on the PLS Regression

—From the Data of Tobacco Consumption Control in Honghe Cigarette Factory

Lei Xu[1*], Xingxu Li[1], Yan Zhang[2], Wenneng Li[2], Bo Zhang[2]

[1]School of Statistics and Mathematics, Yunnan University of Finance and Economics, Kunming Yunnan
[2]Honghe Cigarette Factory, Hongyun Honghe Group, Honghe Yunnan
Email: [*]xulei-2008@163.com

Received: Aug. 16th, 2015; accepted: Aug. 31th, 2015; published: Sep. 7th, 2015

Abstract

On the basis of some conditions for the application of partial least squares regression analysis and multivariate linear regression analysis in this paper, we can conclude that partial least squares regression (PLS) can effectively improve multicollinearity of variables. When the sample size is less than the number of variables, it also can be used to do regression modeling. Then, from 12 groups of sample data of Tobacco consumption control in Honghe Cigarette Factory, we have analyzed and compared the results of partial least squares regression modeling and multivariate linear regression modeling in the paper. It has shown that the significant factors affecting the single box consumption are single case of Wasting, single case of Running, single case of Packet rejection and single case of Short excluded volume. Therefore, the work of the cigarette factory in the process of reducing the cost should be firstly controlling these four single box loss indicators, so that we will achieve the immediate results.

Keywords

Single Box Consumption, Influence Factors, PLS Regression, Comparative Analysis

[*]通讯作者。

文章引用: 许磊, 李兴绪, 张雁, 李文能, 张波. 基于 PLS 回归的单箱消耗影响因素分析[J]. 统计学与应用, 2015, 4(3): 144-154. http://dx.doi.org/10.12677/sa.2015.43016

基于PLS回归的单箱消耗影响因素分析

—来自红河卷烟厂卷包过程中的烟丝消耗控制数据

许　磊[1*]，李兴绪[1]，张　雁[2]，李文能[2]，张　波[2]

[1]云南财经大学统计与数学学院，云南 昆明
[2]红云红河烟草(集团)有限责任公司红河卷烟厂，云南 红河
Email: *xulei-2008@163.com

收稿日期：2015年8月16日；录用日期：2015年8月31日；发布日期：2015年9月7日

摘　要

本文在分析了偏最小二乘回归分析和多元线性回归分析的适用条件基础上，认为偏最小二乘回归(PLS)可以有效地解决变量间多重共线性的问题，甚至适合在样本量少于变量个数的情况下进行回归建模。然后，依据红河卷烟厂烟丝消耗控制的12组样本数据，本文比较分析了偏最小二乘回归建模和多元线性回归建模的结果，发现影响因变量单箱耗丝的显著性因素为单箱废烟、单箱跑条、单箱小包机剔除量和单箱空头剔除量。因此，卷烟厂在卷包过程中的降耗工作应当首先从控制这四个单箱损耗指标开始实施，才能取得立竿见影的效果。

关键词

单箱消耗，影响因素，PLS回归，比较分析

1. 引言

所谓单箱烟丝消耗，或简称单箱耗丝，是指卷烟企业每制造一箱卷烟而实际使用的烟丝重量。单箱烟丝消耗不仅能直观地反映出卷烟企业的主要物耗水平，而且也可以透视出卷烟企业的生产技术管理水平和经济效益情况。因此，单箱烟丝消耗历来是衡量卷烟企业生产经营水平的一项重要技术经济指标。在实际的卷烟生产过程中，单箱烟丝消耗往往与单箱生产成本是成正比关系，并且与单箱收益是成反比关系。因此，降低单箱烟丝消耗是卷烟企业为降低生产成本和提高经济效益而必须抓好的重要工作之一。然而，单箱烟丝消耗不单单是卷烟生产中的某一项可直接观测的指标，其不仅与卷烟厂制丝线的各项生产指标息息相关，而且也受到卷制和包装的生产环节中诸多因素的影响。可以说，单箱烟丝消耗贯穿于卷烟生产的全过程。

另外，探究烟丝消耗的关键影响因素，一直是致力于烟草行业研究的专家学者们所关注的重要问题。贺万华[1]等在《卷烟制丝和卷制过程中主要质量指标与消耗指标的关系及评价方法》一文中，通过一般线性回归分析阐述了烟丝碎丝率、填充值和含水率是影响烟丝消耗的重要因素，并且提出了对制丝线以及卷制过程消耗的评价方法；汪涛[2]等所著的《利用主成分分析和正交试验解决卷烟加工中的原料消耗问题》，通过运用主成分分析和正交试验研究了卷烟某些物理指标在产品质量和原料消耗两个方面的表现形式，也探索出了通过控制主要影响因素有效降低原料消耗的重要途径。以上研究都是着重分析某些质量指标或者物理指标对烟丝消耗的影响，并提出降低烟丝的消耗要通过控制这些指标因素而实现的观点。

而本文所研究的问题是分析单箱烟丝消耗的内在影响因素，例如在生产加工过程中各个工序点的烟丝损耗量。本文通过分析单箱损耗指标和其它重要因素对单箱烟丝消耗的影响并构建实证模型，从统计分析的角度探究出单箱烟丝消耗的关键影响因素，从而有针对性地进行工序改进和生产管理并达到直接并显著降低烟丝的单箱消耗水平的目的。另外，在工业生产的实验测试中常常会出现所采集的数据量较少而需研究的指标变量较多的情况。针对此种样本少、变量多的统计分析，本文将选取偏最小二乘回归分析来建立计量模型。同时，以相同样本数据下的多元线性回归模型为参照面，比较分析说明偏最小二乘回归模型在较少样本量的工业生产数据的统计实证分析中具有较好的普适性。

2. 统计分析方法的选择

2.1. 多元线性回归

在研究烟丝消耗问题的过程中，一般常见的统计分析方法是多元线性回归，其主要适用于模型中存在一个变量受到多个变量影响的情况。例如，单箱烟丝消耗，除了受到成品烟丝含水率的影响之外，还可能会受到单箱废烟、单箱烟末等损耗类指标的影响。这样表现为一般线性回归模型就会有多个解释变量，该模型被称为多元线性回归模型[3]。

多元线性回归模型的一般形式为：

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k+\varepsilon$$

其中，Y是模型的被解释变量(因变量)，$X_1,X_2,\cdots,X_k$表示模型的解释变量(自变量)。$\beta_1,\beta_2,\cdots,\beta_k$称为模型的回归系数，其表示在其他解释变量保持不变的情况下，$X_i\left(i=1,2,\cdots,k\right)$每增加或减少一个单位，$Y$的均值$E(Y)$的变化量。另外，$\varepsilon$是随机误差，对随机误差项我们常需要假定其零均值和同方差，即$E\left(\varepsilon\right)=0,Var\left(\varepsilon\right)=\delta^2$。

对于多元线性回归模型，回归函数的矩阵表达式可为$Y=X\beta+\varepsilon$。采用普通最小二乘法参数进行估计，参数的估计量为$\hat{\beta}=\left(X^TX\right)^{-1}X^TY$，$\hat{\beta}$是$\beta$的最小方差无偏估计量。为了使参数的估计量具有良好的统计性质，对多元线性回归模型还做出了若干基本假设，其基本假设之一是自变量$X_1,X_2,\cdots,X_k$相互独立。如果当两个或者多个自变量之间存在相关关系，那么就认为存在多重共线性。当自变量存在严重的多重相关性，或者自变量个数明显少于样本量时，参数的最小二乘估计量$\hat{\beta}$就会失效。而现代统计领域解决多重共线性常用的方法是逐步回归法，就是找出引起多重共线性的自变量，然后将其剔除。但是在实际的课题中，如果通过采用逐步回归法将影响因变量的重要自变量剔除出去，这样势必会对后续的研究工作产生不利的影响。另外，从统计建模的角度而言，样本容量显然越大越好，但是在研究的过程中收集和整理样本数据是一件较为困难的工作，所以选择合适的样本量是一个重要的问题。从最小二乘的原理出发，一般要得到参数的估计量，样本容量应不少于模型中自变量的个数(包括常数项)，即$n\geq k+1$。

2.2. 偏最小二乘回归(PLS)

偏最小二乘回归方法是近些年来随着统计方法不断革新而产生和发展的一种具有广泛适用性的多元统计分析方法，它于 1983 年由 S. Wold 和 C. Albano 等人[4]首次提出。该方法提出了一种多个因变量对多个自变量的回归建模思路，而且它可以较好地体现模型的整体性。偏最小二乘回归与一般多元线性回归方法的主要区别是在回归建模过程中它采用了信息综合与筛选技术，也就是采用成分提取的研究思路。因此，偏最小二乘回归不再直接考虑单个因变量与单个自变量之间的回归建模，而是在因变量与自变量的集合中提取若干对集合具有最佳解释能力的综合变量(又称成分)，并要求综合变量之间的相关程度达到最大。这表明，一方面新的成分要尽可能的代表原始变量集合的信息，另一方面这些代表自变量的成分

又要对因变量具有最强的解释能力。当然，偏最小二乘回归提取的综合变量之间也会保持正交关系，因而在使用它们进行回归分析时将可以避免使用普通最小二乘回归所遇到的问题。所以，偏最小二乘回归不仅能有效解决变量间的多重共线性问题，并且适合在样本容量小于变量个数的情况下进行回归建模。

在前人研究的基础上，偏最小二乘回归的建模思路终于得以完善。其基本原理[5]如下：假设有 p 个自变量 $\{x_1,x_2,\cdots,x_P\}$ 和 q 个因变量 $\{y_1,y_2,\cdots,y_q\}$，为了研究这些变量之间的相关关系，假设我们获取了 n 个样本点数据，由此构成变量的数据表 $X=\{x_1,x_2,\cdots,x_p\}_{n\times p}$ 和 $Y=\{y_1,y_2,\cdots,y_q\}_{n\times q}$。

偏最小二乘回归提出要分别在 X 与 Y 中提取出成分 τ_1 和 υ_1，也就是说 τ_1 是自变量的线性组合，υ_1 是因变量的线性组合。在提取 τ_1 和 υ_1 这两个成分时，为了统计建模的需要，有以下两个要求：

① τ_1 和 υ_1 应尽可能体现所对应的自变量和因变量数据表中的变异信息；

② τ_1 和 υ_1 的相关程度能够达到最大。

然后，当第一组成分 τ_1 和 υ_1 被提取后，偏最小二乘回归提出了分别拟合 X 对 τ_1 的回归以及 Y 对 τ_1 的回归。如果这两个回归方程可以达到满意的精度，那么算法终止。否则，将继续利用被解释后的残余信息进行第二轮的成分提取。如此往复，直到能达到一个较满意的精度为止。最终，若对自变量共提取了 m 个成分 $\tau_1,\tau_2,\cdots,\tau_m$，那么偏最小二乘回归将拟合出因变量 $Y_k\ (k=1,2,\cdots,q)$ 对 $\tau_1,\tau_2,\cdots,\tau_m$ 的回归方程，并进一步转化表达成 Y_k 关于原自变量 $X_1,X_2,\cdots,X_p$ 的回归方程。

3. 实证分析和结果的比较

3.1. 数据的来源及变量的选取

为了有效地降低生产过程中的烟丝消耗，云南省红河卷烟厂对所生产的云烟(紫)、红河(硬 88)、红河(软甲)和红河(硬)这四个主要牌号在生产过程中的烟丝消耗情况进行实验调查，并记录了相应批次的数据。采用的调查方法为跟踪调查，具体步骤如图 1 所示。

依据图 1 所示烟支生产的流程，卷烟厂的统计人员在制丝过程和卷包过程的每一工序点上记录烟丝的投入量和产出量，从而得出各个工序的损耗数据。基于以上由跟踪调查采集所得的实验数据，本文重点关注和分析卷包这一生产过程的烟丝消耗情况。由烟支的生产理论可知，烟支的卷包生产过程包括卷制和包装两个阶段，而诸多主客观因素会导致烟丝在卷制和包装的过程中产生一些不必要的损耗。在卷制环节，在生产工序点上度量烟丝损耗的指标主要有：废烟、跑条、梗签剔除量和烟末；在包装环节，在生产工序点上度量烟丝损耗的指标主要有：空头剔除量、小包机剔除量、废烟和烟末。显然，以上烟丝损耗指标的数值越大，那么相应工序点所消耗的烟丝也就会越多。所以，此六项损耗类指标也就构成了影响烟丝单箱消耗的直接因素。另外，供卷包的成品烟丝的含水率也是影响卷包过程烟丝消耗水平的重要因素。因为烟丝的含水率越低，那么生产过程中产生的烟丝碎片就会越多以及形成的烟末也就越多，相应的单箱烟丝消耗也就越多。

为了消除产量这一因素的影响，根据调查所得的数据合理推断出四个牌号在卷包过程中每一批次的各个单箱损耗指标以及单箱烟丝消耗量。并以此作为本文研究的样本数据，详见表 1。

在不考虑牌号这一因素影响的前提假设下，本文研究选取 7 个变量为影响单箱耗丝 (Y) 的主要因素：单箱废烟 (X_1)、单箱烟末 (X_2)、单箱跑条 (X_3)、单箱梗签剔除量 (X_4)、单箱空头剔除量 (X_5)、单箱小包剔除量 (X_6) 和成品丝含水率 (X_7)。

3.2. 相关性分析

为了保证拟合模型的有效性，我们首先需要考虑单箱耗丝与各个单箱损耗指标及含水率的相关性。本文运用软件 Minitab 计算出以上 8 个变量之间的相关关系及其检验的 P 值，所得相关系数矩阵见表 2。

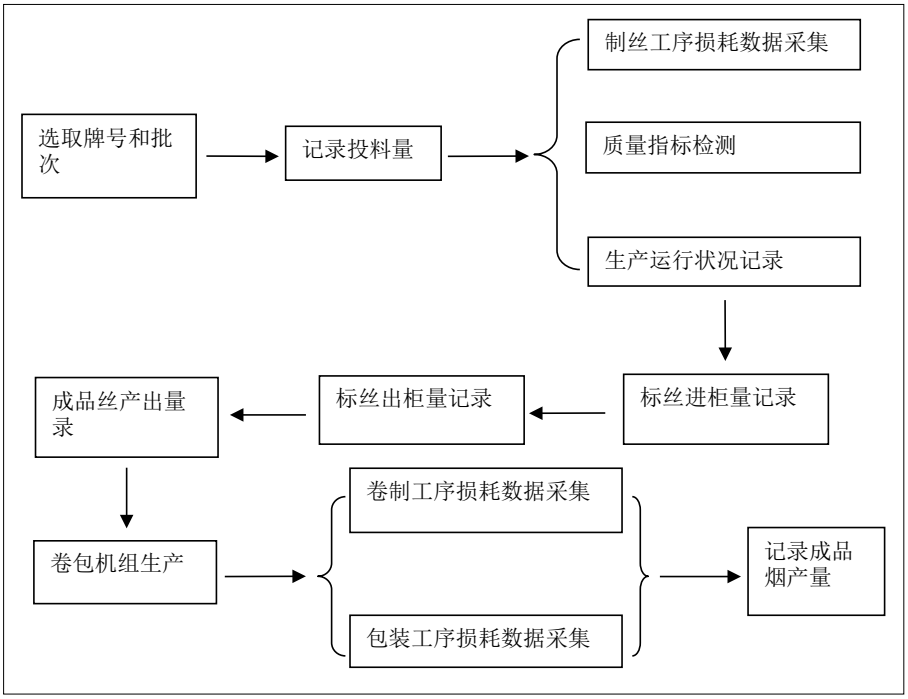

Figure 1. Chart: production consumption tracking survey
图 1. 生产消耗跟踪调查图示

Table 1. The data table of single box consumption in the process of winding
表 1. 卷包过程中的单箱消耗数据表

牌号	批次	成品烟(箱)	单箱耗丝(kg/箱)	卷包过程中的损耗指标						供卷包的成品丝含水率(%)
				单箱废烟(kg/箱)	单箱烟末(kg/箱)	单箱跑条(kg/箱)	单箱梗签剔除(kg/箱)	单箱空头剔除(包/箱)	单箱小包机剔除(包/箱)	
云烟(紫)	一	267.774	34.13	0.289	0.087	0.150	0.623	0.960	5.605	12.77
	二	259.876	35.16	0.337	0.151	0.135	0.693	2.420	7.500	12.57
	三	263.436	34.55	0.339	0.154	0.147	0.576	1.906	5.060	12.67
	四	271.956	33.57	0.281	0.145	0.072	0.451	2.335	7.038	12.84
	五	265.032	34.39	0.317	0.145	0.032	0.396	3.132	7.105	12.78
红河(硬 88)	一	249.176	35.03	0.564	0.150	0.342	0.297	4.302	13.392	12.43
	二	245.668	35.45	0.656	0.155	0.188	0.227	4.457	15.008	12.27
	三	243.472	35.99	0.455	0.093	0.208	0.764	4.592	12.412	12.63
红河(软甲)	一	285.5	34.00	0.129	0.142	0.081	0.937	2.396	5.135	12.23
	二	272.82	34.63	0.163	0.182	0.087	1.059	3.188	4.984	12.23
	三	273.384	34.21	0.179	0.123	0.100	1.143	2.835	5.403	12.4
红河(硬)	一	258.06	34.42	0.396	0.078	0.148	0.410	1.414	4.239	12.51

注：单箱废烟含卷烟机和包装机的单箱废烟量，单箱烟末含卷烟机和包装机的单箱废烟量。

Table 2. Correlation coefficient matrix of single box indexes
表 2. 单箱耗丝与单箱损耗指标及含水率的相关系数矩阵

变量	单箱耗丝	单箱废烟	单箱烟末	单箱跑条	单箱梗签剔除	单箱空头	单箱小包机剔除
单箱废烟	**0.669 (0.017)**	1	−0.084 (0.794)	**0.73 (0.007)**	**−0.785 (0.002)**	0.547 (0.065)	**0.851 (0)**
单箱烟末	−0.018 (0.956)	−0.084 (0.794)	1	−0.139 (0.667)	0.061 (0.851)	0.337 (0.285)	0.113 (0.726)
单箱跑条	**0.589 (0.044)**	**0.73 (0.007)**	−0.139 (0.667)	1	−0.392 (0.207)	0.449 (0.143	**0.681 (0.015)**
单箱梗签剔除	−0.162 (0.615)	**−0.785 (0.002)**	0.061 (0.851)	−0.392 (0.207)	1	−0.137 (0.672)	−0.509 (0.091)
单箱空头	**0.705 (0.011)**	0.547 (0.065)	0.337 (0.285)	0.449 (0.143)	−0.137 (0.672)	1	**0.841 (0.001)**
单箱小包机剔除	**0.725 (0.008)**	**0.851 (0)**	0.113 (0.726)	**0.681 (0.015)**	−0.509 (0.091)	**0.841 (0.001)**	1
成品丝含水率	−0.227 (0.477)	0.018 (0.956)	−0.349 (0.266)	−0.166 (0.606)	−0.353 (0.26)	−0.362 (0.248)	−0.144 (0.656)

注：括号内为检验的 P 值

由上表可知，单箱废烟、单箱跑条、单箱空头、单箱小包机剔除这四个损耗指标与单箱耗丝的相关系数依次为 0.669 (p = 0.017)、0.589 (p = 0.044)、0.705 (p = 0.011)和 0.725 (p = 0.008)，这说明在 5%的显著性水平下单箱废烟、单箱跑条、单箱空头、单箱小包机剔除与单箱耗丝之间存在统计意义上的显著正向相关关系。而对于含水率这一因素，其不仅与单箱耗丝存在一定的负相关关系，并且与除了单箱废烟的其他单箱损耗指标之间也存在一定的负相关关系。

另外，单箱跑条与单箱废烟的相关系数为 0.73 (p = 0.007)，单箱小包机剔除与单箱废烟的相关关系为 0.851 (p = 0)，这说明单箱废烟与单箱跑条、单箱梗签剔除和单箱小包机剔除之间存在统计意义上的显著相关关系。单箱小包机剔除与单箱跑条的相关系数为 0.681 (p = 0.015)，单箱小包机剔除与单箱空头的相关系数为 0.841 (p = 0.001)，这也说明单箱小包机剔除与单箱跑条和单箱空头也存在统计意义上的显著相关关系。以上自变量之间存在的显著相关关系表明采用常用的多元线性回归分析建模一定会存在严重的多重共线性问题，因此我们应当选择合适的统计模型来研究卷包过程中的单箱烟丝消耗问题。

3.3. 偏最小二乘回归建模

由以上偏最小二乘回归的建模理论，本文将拟合的模型方程为 $Y=\beta_0^*+\beta_1^*X_1+\beta_2^*X_2+\beta_3^*X_3+\beta_4^*X_4+\beta_5^*X_5+\beta_6^*X_6+\beta_7^*X_7+\varepsilon$。此模型的因变量只有一个，即单箱烟丝消耗$(Y)$。而自变量 $X_1,X_2,\cdots,X_k$ 表示为卷包过程中的各个单箱损耗指标和其它影响因素。根据表 1 所示的样本数据，运用统计软件 Minitab 和软件 R 所得偏最小二乘回归的建模结果如下：

3.3.1. 选取主成分

根据偏最小二乘回归分析提取成分的原则与做法，本文使用逐一剔除法这一交叉验证的方法选择出能够使模型预测能力最大化的主成分。

由图 2 可知，当模型自变量选择一个主成分时，交叉验证下的该模型的预测 R^2 达到最大值。也就是说，采用 τ_1 这一成分做偏最小二乘回归分析，模型的效果最好。

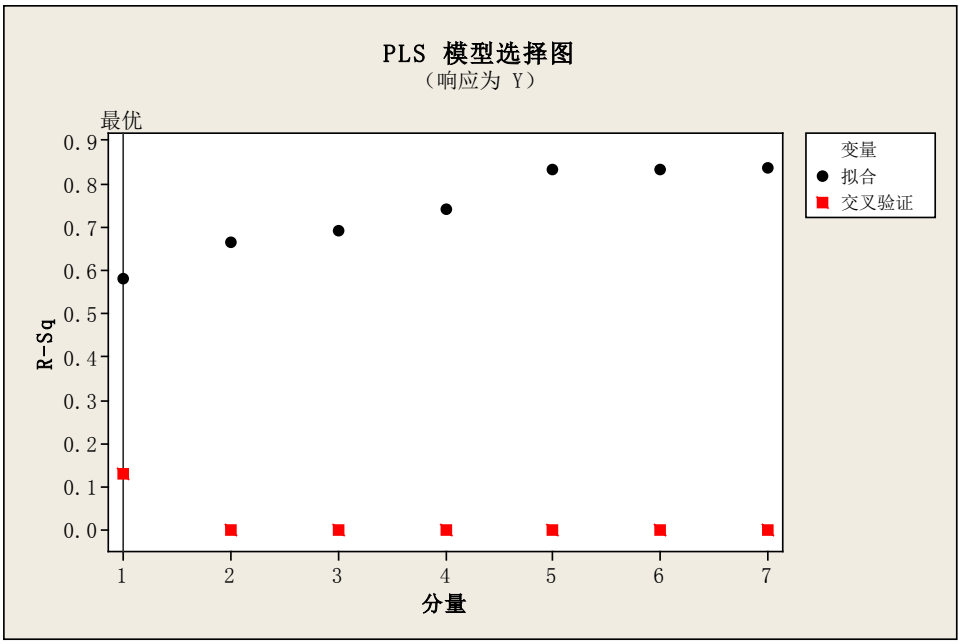

Figure 2. Chart: selecting the principal component of independent variables
图 2. 自变量主成份选择图

3.3.2. 拟合偏最小二乘回归模型

由整体方差分析表 3 可知，F = 13.85，P 值等于 0.004，在 5%显著水平下模型整体通过检验，且计算出模型的可决系数 R^2 为 0.58，表明偏最小二乘回归方程在整体上显著。由表 4 可知，在 10%的显著性水平下，单箱废烟(X_1)、单箱跑条(X_3)和单箱小包机剔除(X_6)这三个变量对因变量单箱耗丝(Y)具有显著影响，而在 11%的显著性水平下单箱空头剔除(X_5)对因变量单箱耗丝(Y)具有显著影响。

另外，标准化回归系数是无量纲的系数，可以反映解释变量(以标准差衡量)的单位变化所能引起的被解释变量的变化幅度，从而也能比较各个自变量对因变量的相对影响程度。由以上拟合模型的结果可知，单箱小包机剔除(X_6)、单箱空头剔除(X_5)、单箱废烟(X_1)和单箱跑条(X_3)的标准化回归系数位列 7 个自变量中的前四位，分别为 0.2221、0.2158、0.2050、0.1806，这说明单箱小包机剔除(X_6)、单箱空头剔除(X_5)、单箱废烟(X_1)和单箱跑条(X_3)这四个变量对因变量的影响相对较大。而成品丝含水率(X_7)的标准化回归系数为−0.0696，也表明其对因变量单箱耗丝有一定的负向影响。

3.3.3. 检验所建模型的残差

对模型残差的 Shapiro-Wilk 正态性检验的 P 值为 0.7912，在 5%的显著性水平下接受残差来自正态总体的假定。且由图 3 可知，在 95%的置信区间，所建立的偏最小二乘回归模型的残差服从正态分布，说明模型具有无偏性和有效性。

因此，采用偏最小二乘回归建模所得的数量方程为：

$$Y = 36.2414 + 0.8762X_1 - 0.1154X_2 + 1.506X_3 - 0.1121X_4 + 0.1244X_5 + 0.0404X_6 - 0.2159X_7$$

3.4. 多元线性回归建模

由以上多元线性回归的建模理论，本文将拟合的模型方程为 $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 + \beta_5X_5 + \beta_6X_6 + \beta_7X_7 + \varepsilon$。因变量 Y 表示单箱烟丝消耗，自变量 $X_1, X_2, \cdots, X_k$ 表示卷包生产过程中的各个单箱损耗指标和其它影响因素。另外，$\beta_1, \beta_2, \cdots, \beta_k$ 为模型方程

Table 3. The table of ANOVA
表 3. 模型整体方差分析表

来源	自由度	SS	MS	F	P
回归	1	2.9137	2.9137	13.85	0.004
残差误差	10	2.1042	0.21042		
合计	11	5.0178			

Table 4. The table of coefficient of model independent variable and standard coefficient
表 4. 模型自变量的系数及标准化系数表

自变量	回归系数	标准化回归系数	系数标准误差	自由度	t	P
常量	36.2414	——	——	——	——	——
单箱废烟(X_1)	0.8762	0.205	0.0814	11	2.52	0.0286
单箱烟末(X_2)	−0.1154	−0.0054	0.0999	11	−0.05	0.9576
单箱跑条(X_3)	1.506	0.1806	0.0861	11	2.1	0.0598
单箱梗签剔除(X_4)	−0.1121	−0.0494	0.076	11	−0.65	0.5287
单箱空头(X_5)	0.1244	0.2158	0.1221	11	1.77	0.1049
单箱小包机剔除(X_6)	0.0404	0.2221	0.0989	11	2.24	0.0463
含水率(X_7)	−0.2159	−0.0696	0.0993	11	−0.7	0.4975

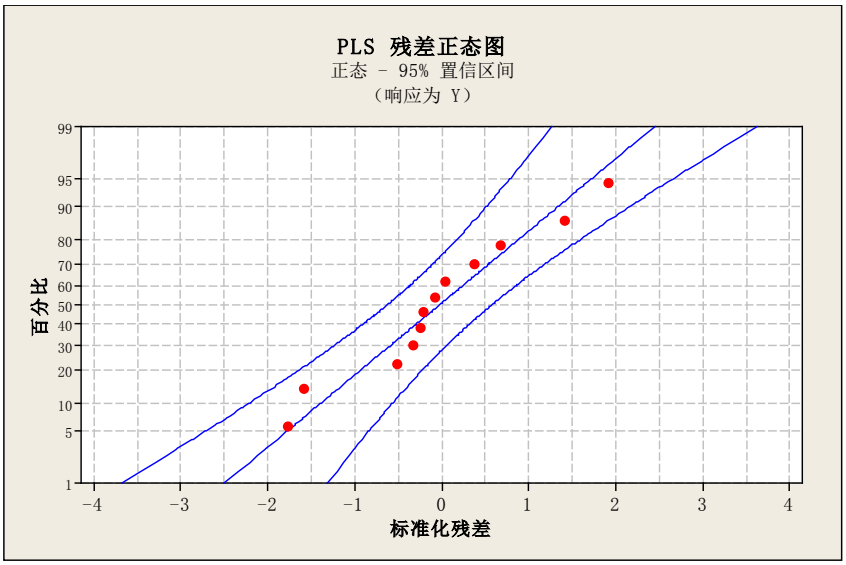

Figure 3. Chart: normality of residuals
图 3. 残差的正态图

的回归系数，其反映卷包过程中显著的自变量指标对单箱烟丝消耗这一因变量的影响。根据表 1 所示的样本数据，运用统计软件 Minitab 和软件 R 所得多元线性回归建模结果如下：

3.4.1. 拟合多元线性回归模型

由前文的相关性分析可知，自变量之间存在显著的相关关系，这说明采用多元线性回归分析建模会存在多重共线性的问题。但是本文所研究的样本数据只有 12 组，如果为了解决多重共线问题性而使用逐步回归可能会将重要影响因素剔除出模型。因此，作为偏最小二乘回归模型的参照面，本文将同样选取

7 个自变量建立多元线性回归模型。

由表 5 可知，F = 2.89，P 值等于 0.151。由于样本数据量太少和模型存在多重共线性，在 5%显著水平下多元线性回归模型整体不能通过显著性检验。由表 6 可知，对模型各个自变量进行 t 检验，在 10%的显著性水平下，只有单箱废烟(X_1)这一个变量对因变量单箱耗丝(Y)具有显著影响。而在 12%的显著性水平下单箱梗签剔除量(X_4)对因变量单箱耗丝(Y)具有显著影响。

3.4.2. 检验模型的残差

模型残差的 Shapiro-Wilk 正态性检验的 P 值为 0.8752，在 5%的显著性水平下接受残差来自正态总体的假定。多元线性回归模型的残差服从正态分布，这表明所建模型具有无偏性和有效性。

因此，采用多元线性回归建模而得的数量方程为：

$$Y = 23.47 + 7.539X_1 + 1.099X_2 - 0.93X_3 + 2.455X_4 + 0.1825X_5 - 0.0719X_6 + 0.5633X_7$$

3.5. 两种建模结果的比较

3.5.1. 对比模型方程

通过偏最小二乘回归和多元线性回归这两种统计建模方法，对影响单箱烟丝消耗的因素进行实证分析所得的计量模型方程如下：

偏最小二乘回归建模方程①：

$$Y = 36.2414 + 0.8762X_1 - 0.1154X_2 + 1.506X_3 - 0.1121X_4 + 0.1244X_5 + 0.0404X_6 - 0.2159X_7$$

其中，显著自变量为：单箱废烟(X_1)、单箱跑条(X_3)、单箱小包机剔除(X_6)和单箱空头剔除(X_5)。

多元线性回归建模方程②：

$$Y = 23.47 + 7.539X_1 + 1.099X_2 - 0.93X_3 + 2.455X_4 + 0.1825X_5 - 0.0719X_6 + 0.5633X_7$$

其中，显著自变量为：单箱废烟(X_1)、单箱梗签剔除量(X_4)。

Table 5. The table of ANOVA
表 5. 模型整体方差分析表

来源	自由度	SS	MS	F	P
回归	7	4.1886	0.5984	2.89	0.161
残差误差	4	0.8293	0.2073		
合计	11	5.0178			

Table 6. The table of coefficient of model independent variable
表 6. 模型自变量的系数及标准化系数表

自变量	回归系数	系数标准误差	t	P
常量	23.47	11.73	2	0.116
X_1	7.539	3.427	2.2	0.093
X_2	1.099	5.716	0.19	0.857
X_3	−0.93	2.892	−0.32	0.764
X_4	2.455	1.21	2.03	0.112
X_5	0.1825	0.3282	0.56	0.608
X_6	−0.0719	0.1403	−0.51	0.635
X_7	0.5633	0.854	0.66	0.546

从模型方程进行比较，方程①中影响因变量单箱烟丝消耗的显著自变量有 4 个，而方程②中影响因变量单箱烟丝消耗的显著自变量有 2 个，这说明对于采用偏最小二乘建模可以有效克服由于变量间的多重相关性所带来的影响，得到较为理想的回归模型。另外，成品丝含水率(X_7)这一自变量虽然在两个模型方程中表现为不显著，但是在实际的卷烟生产过程中它对烟丝消耗是具有负向的影响，并且这一特征已经在相关性分析得以体现。方程①中因变量 X_7 的回归系数为负数，而方程②中因变量 X_7 的回归系数为正数，这说明偏最小二乘回归分析在样本量较少的情况下建模更能契合生产实际中的变量关系。

3.5.2. 预测误差平方和(PRESS)

PRESS 类似于误差平方和(SSE)，是预测误差的平方和，用于评估模型的预测能力。一般而言，PRESS 值越小，模型的预测能力越强。

$$\mathrm{PRESS}=\sum_{i=1}^{n}\left(y_i-\hat{y}_{i,-i}\right)^2$$

其中，y_i 表示第 i 个因变量的观测值，$\hat{y}_{i,-i}$ 表示通过从数据集中排除第 i 组自变量观测值后拟合回归模型获得第 i 个因变量观测值的预测值。

使用软件 Minitab 对样本数据建模的过程中，以上多元线性回归模型的预测误差平方和(PRESS)等于 14.5924，偏最小二乘回归模型的预测误差平方和(PRESS)等于 4.3699。这表明所建偏最小二乘回归模型的预测能力比多元线性回归模型的预测能力要强大。

因此，无论是通过比较这两种建模方法的回归方程，还是比较能代表它们预测能力的预测误差平方和，这都说明了偏最小二乘回归模型更适合于分析类似本文的样本数据(样本少、变量多)，尤其是存在所研究的自变量个数多于样本量的情况。

4. 结论

本文依据卷烟厂测试实验所采集的 12 组样本数据，分别采用偏最小二乘回归模型和多元线性回归模型对卷包过程中影响单箱烟丝消耗的因素进行了实证分析，得出如下结论：

1) 不仅各个单箱损耗指标和成品丝含水率与单箱烟丝消耗存在统计意义上的显著相关关系，并且这几个损耗指标之间也存在统计意义上显著的相关关系。因此，卷烟厂要想降低其单箱烟丝消耗水平不能只管控某一项单箱损耗指标，应当从卷包生产过程的全局进行考虑。

2) 成品丝含水率不仅对单箱烟丝消耗具有一定的负向影响，而且与大部分单箱损耗指标之间存在一定的负相关关系。这说明成品丝含水率在卷包生产过程中的降耗工作中有着重要的作用，应当给予重点关注。

3) 由偏最小二乘回归模型的结果可知，单箱废烟(X_1)、单箱跑条(X_3)、单箱小包机剔除(X_6)和单箱空头剔除(X_5)这四个变量对因变量单箱耗丝(Y)具有显著性影响。因此，卷烟厂在卷包过程中的降耗工作应当首先从控制这四个单箱损耗指标开始实施，才能取得立竿见影的效果。

4) 通过比较偏最小二乘模型和多元线性回归模型所得的结果，可以认为偏最小二乘回归模型更适合于分析这种样本少、变量多的样本数据，尤其是所研究自变量个数多于样本量的情况。而在工业生产的过程中，因各种原因导致测试所采集的数据量较少和研究的指标变量较多的情况屡见不鲜，因此采用偏最小二乘回归分析建模更符合实际工作和研究的需要，值得深入学习和推广。

致　谢

感谢红云红河烟草(集团)有限责任公司红河卷烟厂烟叶消耗控制项目对本次研究工作的大力支持，并使得本论文的撰写得以顺利完成。

参考文献 (References)

[1] 贺万华, 曹兴洪, 等 (2007) 卷烟制丝和卷制过程中主要质量指标与消耗指标的关系及评价方法. *中国烟草学报*, **5**, 17-22.

[2] 汪涛, 张琦 (2011) 利用主成分分析和正交试验解决卷烟加工中的原料消耗问题. *黑龙江科技信息*, **5**, 50.

[3] 何晓群, 刘文卿 (2007) 应用回归分析. 中国人民大学出版社, 北京.

[4] Wold, H. (1975) Soft modelling by latent variables: The non-linear iterative partial least squares (NIPALS) approach. Perspectives in probability and statistics. Papers in Honour of M. S. Bartlett. Academic Press, London, 117-142.

[5] 王惠文 (1999) 偏最小二乘回归方法及应用. 国防工业出版社, 北京.