

油气管道投资建模新方法

段善宁¹, 赵伏锐², 汪玉春³

(1. 中国石化西南分公司川西采输处, 四川 德阳 618000;

2. 中国石油工程设计有限公司西南分公司, 四川 成都 610017; 3. 西南石油大学, 四川 成都 610500;)

摘 要: 油气管道投资计算的准确性对管道优化设计和工程计算是否正确具有举足轻重的作用。在进行建模时, 若采用线性最小二乘法对基期投资费用进行线性拟合, 有时要遇到管道参数之间的多重相关性、样本容量小等困难。针对这些问题, 提出采用全新建模方法——偏最小二乘法建立管道基期投资计算模型。偏最小二乘法适用于参数多重相关, 样本容量小的情况。在只有十个样本数据的条件下, 将样本数据按照三组小范围管径进行划分, 再用偏最小二乘法分别建立三元线性回归模型。从对模型进行了精度分析看, 采用偏最小二乘法所建模型的累计解释能力都在 97% 以上, 克服了采用线性最小二乘法建立投资费回归模型计算结果误差大和管道参数之间的多重相关性问题。

关键词: 管道; 投资模型; 线性回归; 偏最小二乘法; 精度

文章编号: 1006-5539(2006)06-0014-04

文献标识码: A

0 前言

油气管道投资模型是油气管道优化设计模型的

重要组成部分, 模型的准确性对优化设计和工程计算结果具有举足轻重的影响, 实际应用中通常采用投资费随管径、壁厚变化的数学模型。但要建立准确的数学模型则必须根据数据的特点来选取建模的

基金项目: 四川省重点学科建设资助项目(编号: SZD0416)

收稿日期: 2006-02-14

作者简介: 段善宁(1979-), 男, 四川南充人, 硕士, 毕业于西南石油大学, 目前主要从事天然气储运研究与管理工。

电话: 13541715258

方法。

在文献^[1~2]中管道基期投资费用被认为随管道直径和壁厚成线性变化,采用二元线性模型求出管道的投资回归方程,可得到公式:

$$K = \alpha_0 + \alpha_1 D + \alpha_2 \delta + \epsilon \quad \epsilon \sim N(0, \sigma^2) \quad (1)$$

式中 $\alpha_1, \alpha_2, \alpha_0$ ——回归系数;
 ϵ ——随机偏差;
 D ——管道直径;
 δ ——管道壁厚;
 K ——管道投资。

从文献^[1~2]可知,式(1)是基于线性最小二乘法的多元回归。采用普通多元回归分析时,样本点数目不宜太少,样本容量要求超过 50 个才能得到保证。分析文献^[1~2]的算例,该模型采用线性最小二乘法拟合效果好的原因是因为该算例的样本容量达到 19×5,样本属于大容量样本。而油气管道工程中由于时间、不同管径类型的项目限制,所能采集到的样本点很少,甚至样本点个数少于管道参数个数。例如某地区管道投资与各个参数的关系(见表 1),若采用线性最小二乘法建立 K 对 $D, \hat{q}_g, \hat{q}_b$ 的回归方程,回归结果如下所示:

$$K = 0.183\,7\,D + 6.057\,2\,\hat{q}_g - 1.398\,9\,\hat{q}_b + 1.137\,9 \quad (2)$$

式中 K ——管道投资费;
 D ——管道直径;
 $\hat{q}_g$ ——管道壁厚;
 $\hat{q}_b$ ——保温层厚。

表 1 管道投资一览表

管径 d	外径 D	壁厚 $\hat{q}_g$	保温层厚度 $\hat{q}_b$	投资费 K
/mm	/mm	/mm	/mm	/10 ⁴ 元·km ⁻¹
150	168.0	4.0	30.0	21.50
200	219.0	5.0	30.0	35.13
250	273.0	5.0	40.0	37.52
300	323.9	5.0	40.0	43.66
350	355.6	6.0	40.0	51.79
400	406.4	7.0	50.0	64.04
500	508.0	8.0	50.0	85.60
600	610.0	10.0	50.0	115.60
700	711.0	11.0	50.0	140.52
750	762.0	12.0	60.0	182.13

经检验,该模型的平均相对误差为 8.15%,最大相对误差为 15.81%。因此可以看出,在样本容量较小的情况下,采用线性最小二乘法建立管道基

期投资模型效果较差,不宜用于工程计算。

另外,从三个变量 $D, \hat{q}_g, \hat{q}_b$ 的相关系数矩阵(见表 2)中,我们可以看出变量之间存在严重的相关性,这将影响所选取的自变量对因变量投资费的解释能力。

表 2 相关系数矩阵

$r(^{\circ}, ^{\circ})$	$\hat{q}_g$	$\hat{q}_b$	K
D	0.989 8	0.857 0	0.978 7
$\hat{q}_g$		0.815 4	0.986 3
$\hat{q}_b$			0.763 9

针对以上提到的问题,本文提出采用一种全新的数值拟合方法——偏最小二乘法,来建立管道基期投资模型,从而可以避免变量之间的多重相关性以及样本点较少的困难。与采用线性最小二乘法进行建模的方法完全不同,偏最小二乘法在提取新的综合变量时,要求自变量与因变量的相关程度达到最大,即新的自变量尽可能代表原始变量集合的信息,同时又要对因变量有最强的解释能力。并且,该算法还可以实现数据结构简化以及两组变量间的相关分析,这给多维复杂系统的分析带来了极大的便利^[3~4]。

1 用偏最小二乘法建立油气管道基期投资模型

将为管道基期投资模型选取的自变量(应为列向量)组合成为自变量集,在本文中选取外径 D ,壁厚 $\hat{q}_g$,保温层厚 $\hat{q}_b$,即为 $E=[D\ \hat{q}_g\ \hat{q}_b]$ 。由于该模型为单因变量模型系统,可以直接令管道基期投资费 K 为目标函数,根据文献^[4]确定模型形式如下:

$$K = \alpha_1 D + \alpha_2 \hat{q}_g + \alpha_3 \hat{q}_b + \alpha_0 \quad (3)$$

式中 α_0 ——回归系数, $\alpha_0 = \frac{1}{n} \sum_{i=1}^n F_{mi}$, m 为该模型中提取的成分个数, F_{mi} 为该模型第 m 个残差, n 为样本容量;

α_j —— $D, \hat{q}_g, \hat{q}_b$ 对应的回归系数, $\alpha_j =$

$$\sum_{h=1}^m r_h w_{hj}^*, j = 1, 2, 3, r_h = \frac{F_{h-1}^T t_h}{\|t_h\|^2},$$

F_{h-1}^T 为该模型的第 $h-1$ 次残差的转置矩阵;

w_{hj}^* —— w_h^* 的第 j 个分量, $w_h^* = t_h / E_0, E_0$

是自变量集合 E 的标准化矩阵;

t_h ——该偏最小二乘模型提取的第 h 个成分;

$$t_h = \frac{1}{\sqrt{r^2(D, K) + r^2(\hat{Q}_g, K) + r^2(\hat{Q}_b, K) + [r(D, K)E_{h-1,1} + r(\hat{Q}_g, K)E_{h-1,2} + r(\hat{Q}_b, K)E_{h-1,3}]}}$$

$r(\cdot, \cdot)$ ——任意两个变量的相关系数。

因在建立基期投资模型以前将数据进行了标准化,所以一定要将式(3)按照标准化的逆运算过程还原。从式(3)中 α_j 的构造可以看出,如果 $D, \hat{Q}_g, \hat{Q}_b$ 在构造成分 t_h 时贡献越大(即 w_{hj}^* 越大), t_h 在解释 K 时的作用就越大(即 r_h 越大),即 $D, \hat{Q}_g, \hat{Q}_b$ 在偏最小二乘回归模型中的回归系数就越大。

下面通过一个算例进一步说明采用偏最小二乘法建立管道投资回归模型。

2 实例计算

2.1 计算结果

从表 1 可以看出,管道直径变化较宽(从 DN 150~750),根据管径范围的划分,本文采用两个

模型计算该算例。

方案 1:将表 1 全部数据用一个回归模型拟合,管径范围从 DN 150~750。

$$K = 0.048\ 9\ D + 11.734\ 9\ \hat{Q}_g + 0.387\ 5\ \hat{Q}_b - 42.265 \tag{4}$$

方案 1 的最大相对误差为 15.793 9%,平均相对误差为 4.735 3%,该模型的误差较大,不宜用于实际。

方案 2:将表 1 数据分为三组:DN 150~300、DN 350~500、DN 600~750,用三个回归模型进行回归。

a. DN 150~300

$$K = 5.412\ 6 \times 10^{-2}\ D + 9.441\ 3\ \hat{Q}_g + 0.191\ 3\ \hat{Q}_b - 30.963\ 5 \tag{5}$$

b. DN 350~500

$$K = 0.113\ 4\ D + 6.921\ 4\ \hat{Q}_g + 1.358\ 1\ \hat{Q}_b - 28.706\ 8 \tag{6}$$

c. DN 600~750

$$K = -0.185\ 9\ D + 22.849\ 8\ \hat{Q}_g + 2.285\ 0\ \hat{Q}_b - 87.955\ 6 \tag{7}$$

表 3 方案 2 的计算结果

管径 / mm	外径 / mm	壁厚 / mm	保温层厚度 / mm	投资实际值 / 10^4 元 $\cdot$ km $^{-1}$	投资拟合值 / 10^4 元 $\cdot$ km $^{-1}$	相对误差 / (%)
DN 150	168.0	4	30	21.50	21.63	0.62
DN 200	219.0	5	30	35.13	33.84	3.69
DN 250	273.0	5	40	37.52	38.67	3.07
DN 300	323.0	5	40	43.66	41.43	5.12
DN 350	355.6	6	40	51.79	53.68	3.65
DN 400	406.4	7	50	64.04	66.50	3.84
DN 500	508.0	8	50	85.60	84.94	0.77
DN 600	610.0	10	50	115.60	118.53	2.53
DN 700	711.0	11	50	140.52	145.45	3.50
DN 750	762.0	12	60	182.13	181.67	0.25

由表 3 可知,方案 2 的平均误差最大为 5.12%,可见方案 2 拟合的准确性比采用线性最小二乘法建立的模型和方案 1 中的模型高。

2.2 模型分析

根据模型的精度分析结果,我们可以判断该模型对原始数据解释正确的程度,进而确认该模型是否可以用于工程实际计算。方案 2 的误差远小于方案 1,因此仅对方案 2 的模型进行分析。根据文献^[4]中的精度估计定义,成分 t_h 对 $D, \hat{Q}_g, \hat{Q}_b$ (模型

自变量集 $E=[D, \hat{Q}_g, \hat{Q}_b]$ 和 K 解释能力为:

$$Rd(e_j; t_h) = r^2(e_j; t_h), Rd(K; t_h) = r^2(K; t_h) \tag{8}$$

式中 e_j —— E 的第 j 个向量,在该模型中 e_j 为 $D, \hat{Q}_g$ 和 $\hat{Q}_b, j=1, 2, 3$ 。

而成分 t_h 对整个数据表 E 的综合解释能力定义为:

$$Rd(E; t_h) = \frac{1}{3} \sum_{j=1}^3 Rd(e_j; t_h) \tag{9}$$

t_h 对 e_j, E 和 K 的累计解释能力为:

$$Rd(e_j;t_1\cdots t_m)=\sum_{h=1}^mRd(e_j;t_h),$$
$$Rd(E;t_1\cdots t_m)=\sum_{h=1}^mRd(E;t_h),$$
$$Rd(K;t_1\cdots t_m)=\sum_{h=1}^mRd(K;t_h)\tag{10}$$

在该例中,从自变量集 E 中提取了两个成分 t_1, t_2 , 这些成分对 $D, \hat{q}_g, \hat{q}_h, E$ 及 K 的解释能力见表 4。

表 4 方案 2 的 t_1, t_2 对 $D, \hat{q}_g, \hat{q}_h, E$ 集合及 K 的解释能力

$r(^{\circ},^{\circ})$	DN150~300		DN350~500		DN600~750	
	t_1	t_2	t_1	t_2	t_1	t_2
D	0.949 7	0.011 7	0.931 2	0.068 9	0.983 9	0.016 1
$\hat{q}_g$	0.757 9	0.225 5	0.994 3	0.005 7	0.996 3	0.003 7
$\hat{q}_h$	0.795 5	0.196 6	0.812 2	0.187 8	0.996 3	0.003 7
E	0.834 4	0.144 6	0.912 6	0.087 4	0.992 2	0.007 8
K	0.936 2	0.043 6	0.946 7	0.053 3	0.958 8	0.041 2

以第一组管道直径数据为例,从拟合的角度看,只需要取一个成分 t_1 即可以构成模型,增加第二个成分对管道参数的解释能力有提高。根据式(9)和表 4 数据,当只提取了成分 t_1 时,在 $D, \hat{q}_g, \hat{q}_h$ 组成的自变量集 E 中有 83.44% 的信息, K 中有 93.62% 的信息被用于分析过程。又因为 $Rd(E, t_2)=0.144\ 6, Rd(K, t_2)=0.043\ 6$, 说明提取成分 t_2 后, E 中又有 14.46% 的信息, 投资费 K 中又有 4.36% 的信息被用于参与了对模型的解释。因此在提取两个成分后,采用偏最小二乘法建立的模型可以解释自变量集合 E , 即模型参数中 97.896% 的信息, 投资费 K 中的 97.986% 的信息。

有关 t_1, t_2 的累计解释能力的数值见表 5。根据式(10)和表 5 数据,在提取了两个成分后,该偏最小二乘模型对投资模型参数的解释能力分别为 97.90%、97.78% 和 99.99%, 而对投资费 K 的累计解释能力分别是 97.99%、99.89% 和 99.98%。因此在该实例中,提取两个成分建立的偏最小二乘回归模型可以解释原始数据集 97% 以上的信息。

由此可见,方案 2 中所建模型对表 1 的数据有较好的解释能力,因此完全可以将其用于油气管道基期投资计算。

表 5 方案 2 对 $D, \hat{q}_g, \hat{q}_h, E$ 集合及 K 的累计解释能力

管径	DN150~300		DN350~500		DN 600~750	
成分个数 n	1	2	1	2	1	2
D	0.949 7	0.961 4	0.931 2	0.999 8	0.983 9	0.999 8
$\hat{q}_g$	0.757 9	0.983 4	0.994 3	0.999 7	0.996 3	0.999 9
$\hat{q}_h$	0.795 5	0.992 1	0.812 2	0.998 9	0.996 3	0.999 7
E	0.834 4	0.979 0	0.912 6	0.998 7	0.992 2	0.999 9
K	0.936 2	0.979 9	0.946 7	0.998 9	0.958 8	0.999 8

3 结束语

综上所述,本文采用偏最小二乘法分组建立的管道基期投资回归模型能够解决管径、壁厚、保温层厚度三个自变量存在严重多重相关性问题,且允许样本点个数少于自变量个数,所得模型能够最大限度地提取与基期投资费相关的管道参数信息,排除管道参数集合中多余信息,模型精度也能满足工程计算的要求。

将偏最小二乘回归用于建立管道基期费用投资模型时要注意,如果所收集的基期管道投资费,其直径、壁厚、保温层厚与投资费不满足线性的变化关系,可采用多元非线性方程来描述管道基期投资模型形式,其它的建模过程与采用线性模型完全相同。

参考文献:

[1] 汪玉春,赵 胜,陈 娟,等.通累积法预测油气管道投资[J].天然气工业,2002,22(4):102-105.

[2] 汪玉春,刘玉峰.油气管道投资的灰色预测与计算[J].天然气工业,1998.18(6):68-71.

[3] 付凌晖,王惠文.多项式回归的建模方法比较研究[J].数理统计与管理,2004,23(1):48-52.

[4] 王惠文.偏最小二乘回归方法及其应用[M].北京:国防工业出版社,1999.