**微观医疗费用预测模型：从线性回归到机器学习**

石菊 王小倩 王熙

**目 录**

[附录Ⅰ 其他预测结果 1](#_Toc150525910)

[附录Ⅱ 机器学习模型与评价指标 1](#_Toc150525908)2

[附录Ⅲ 其他描述性统计 1](#_Toc150525909)4

[附录Ⅳ 医保支付方式和政府补贴方式设计 2](#_Toc150525911)0

# 

# 附录Ⅰ 其他预测结果

**表Ⅰ1 各计量模型样本外预测评价结果**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | R2 | RMSE | MAE | MAPE |
|  | (1) | (2) | (3) | (4) |
| 线性回归模型 | **0.021** | **3,563** | 980 | 137 |
| 广义线性模型(伽玛分布） | - | 3,609 | 997 | 137 |
| 广义线性模型(泊松分布） | 0.015 | 3,573 | **977** | 139 |
| Tobit模型 | - | 3,692 | 1735 | 368 |
| 两部分模型: |  |  |  |  |
| 线性回归模型 | 0.020 | 3,564 | 979 | **136** |
| 广义线性模型(伽玛分布） | 0.014 | 3,574 | 983 | 137 |
| 广义线性模型(泊松分布） | 0.018 | 3,567 | 979 | 137 |

注：广义线性模型的连接函数为log函数，分布分别采用伽玛分布与泊松分布。两部分模型的第一部分模型为Probit模型，用于预测该年产生医疗费用的概率Prob（y＞0|x），其中*y*为医疗费用变量，*x*为个人特征变量；第二部分模型分别为线性回归模型和广义线性模型，用于预测该年有医疗费用条件下的年度医疗费用E（*y*|*x*，*y*＞0）；年度医疗费用预测值为E（*y*|*x*) = Prob（*y*＞0|*x*）E（*y*|*x*，*y*＞0）。根据预测结果，伽玛分布广义线性模型与Tobit模型计算得到的R2无效。加粗数字表示该项模型在该项指标上表现最优。

**表Ⅰ2 各模型样本内预测评价结果**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | R2 | RMSE | MAE | MAPE |
|  | (1) | (2) | (3) | (4) |
| 线性回归 | 0.022 | 3,620 | 901 | 141 |
| 拉索回归 | 0.022 | 3,621 | 904 | 144 |
| (0.00%) | (-0.03%) | (-0.33%) | (-2.13%) |
| 岭回归 | 0.022 | 3,620 | 901 | 141 |
| (0.00%) | (0.00%) | (0.00%) | (0.00%) |
| 弹性网络 | 0.021 | 3,621 | 902 | 142 |
| (-4.55%) | (-0.03%) | (-0.11%) | (-0.71%) |
| 正交匹配追踪 | 0.022 | 3,621 | 903 | 142 |
| (0.00%) | (-0.03%) | (-0.22%) | (-0.71%) |
| 随机森林 | 0.027 | 3,612 | 906 | 156 |
| (22.7%) | (0.22%) | (-0.55%) | (-10.6%) |
| 梯度提升 | **0.030** | **3,605** | 897 | 143 |
| **(36.4%)** | **(0.41%)** | (0.44%) | (-1.42%) |
| 前馈神经网络 | 0.027 | 3,610 | **886** | **138** |
| (22.7%) | (0.28%) | **(1.66%)** | **(2.13%)** |

注：括号中的数字表示该模型的该指标相对于线性回归模型的提升比例。加粗数字表示该项模型在该项指标上表现最优。

**表Ⅰ3** **新增婚姻状况、职业和受教育水平预测变量的各模型样本外预测评价结果**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | R2 | RMSE | MAE | MAPE |
|  | (1) | (2) | (3) | (4) |
| 线性回归 | 0.020 | 3,613 | 998 | 134 |
| 拉索回归 | 0.020 | 3,614 | 996 | 137 |
| (0.00%) | (-0.03%) | (0.20%) | (-2.24%) |
| 岭回归 | 0.020 | 3,613 | 998 | 134 |
| (0.00%) | (0.00%) | (0.00%) | (0.00%) |
| 弹性网络 | 0.020 | 3,614 | 996 | 137 |
| (0.00%) | (-0.03%) | (0.20%) | (-2.24%) |
| 正交匹配追踪 | 0.020 | 3,614 | 998 | 134 |
| (0.00%) | (-0.03%) | (0.00%) | (0.00%) |
| 随机森林 | 0.021 | 3,612 | **985** | 145 |
| (5.00%) | (0.03%) | **(1.30%)** | (-8.21%) |
| 梯度提升 | **0.023** | **3,610** | 991 | 135 |
| **(15.0%)** | **(0.08%)** | (0.70%) | (-0.75%) |
| 前馈神经网络 | 0.021 | 3,613 | 1,005 | 137 |
| (5.00%) | (0.00%) | (-0.70%) | (-2.24%) |

注：括号中的数字表示该模型的该指标相对于线性回归模型的提升比例。加粗数字表示该项模型在该项指标上表现最优。

**表Ⅰ4 新增所属村居预测变量的各模型样本外预测评价结果**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | R2 | RMSE | MAE | MAPE |
|  | (1) | (2) | (3) | (4) |
| 线性回归 | 0.021 | 3,563 | 981 | 137 |
| 拉索回归 | 0.020 | 3,565 | 983 | 139 |
| (-4.76%) | (-0.06%) | (-0.20%) | (-1.46%) |
| 岭回归 | 0.019 | 3,565 | 987 | 139 |
| (-9.52%) | (-0.06%) | (-0.61%) | (-1.46%) |
| 弹性网络 | 0.020 | 3,565 | 986 | 139 |
| (-4.76%) | (-0.06%) | (-0.51%) | (-1.46%) |
| 正交匹配追踪 | 0.019 | 3,565 | 985 | 138 |
| (-9.52%) | (-0.06%) | (-0.41%) | (-0.73%) |
| 随机森林 | 0.022 | 3,560 | 973 | 152 |
| (4.76%) | (0.08%) | (0.82%) | (-10.95%) |
| 梯度提升 | **0.023** | **3,558** | 972 | 139 |
| **(9.52%)** | **(0.14%)** | (0.92%) | (-1.46%) |
| 前馈神经网络 | 0.021 | 3,562 | **963** | **135** |
| (0.00%) | (0.03%) | **(1.83%)** | **(1.46%)** |

注：括号中的数字表示该模型的该指标相对于线性回归模型的提升比例。加粗数字表示该项模型在该项指标上表现最优。

**表Ⅰ5 新增婚姻状况、职业、受教育水平和所属村居预测变量的各模型样本外预测评价结果**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | R2 | RMSE | MAE | MAPE |
|  | (1) | (2) | (3) | (4) |
| 线性回归 | 0.020 | 3,614 | 999 | 134 |
| 拉索回归 | 0.020 | 3,615 | 997 | 137 |
| (0.00%) | (-0.03%) | (0.20%) | (-2.24%) |
| 岭回归 | 0.020 | 3,614 | 1001 | 135 |
| (0.00%) | (0.00%) | (-0.20%) | (-0.75%) |
| 弹性网络 | 0.020 | 3,615 | 997 | 137 |
| (0.00%) | (-0.03%) | (0.20%) | (-2.24%) |
| 正交匹配追踪 | 0.020 | 3,614 | 997 | 135 |
| (0.00%) | (0.00%) | (0.20%) | (-0.75%) |
| 随机森林 | 0.021 | 3,612 | **989** | 148 |
| (5.00%) | (0.06%) | **(1.00%)** | (-10.45%) |
| 梯度提升 | **0.024** | **3,607** | 990 | 136 |
| **(20.00%)** | **(0.19%)** | (0.90%) | (-1.49%) |
| 前馈神经网络 | 0.020 | 3,614 | 990 | **130** |
| (0.00%) | (0.00%) | (0.90%) | **(2.99%)** |

注：括号中的数字表示该模型的该指标相对于线性回归模型的提升比例。加粗数字表示该项模型在该项指标上表现最优。

**表Ⅰ6 使用前三年病种的各模型样本外预测评价结果**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | R2 | RMSE | MAE | MAPE |
|  | (1) | (2) | (3) | (4) |
| 线性回归 | 0.018 | 3,754 | 1,082 | 131 |
| 拉索回归 | 0.017 | 3,756 | 1,082 | 134 |
| (-5.56%) | (-0.05%) | (0.00%) | (-2.29%) |
| 岭回归 | 0.019 | 3,754 | 1,082 | 131 |
| (5.56%) | (0.00%) | (0.00%) | (0.00%) |
| 弹性网络 | 0.017 | 3,756 | 1,082 | 134 |
| (-5.56%) | (-0.05%) | (0.00%) | (-2.29%) |
| 正交匹配追踪 | 0.018 | 3,754 | 1,080 | 131 |
| (0.00%) | (0.00%) | (0.18%) | (0.00%) |
| 随机森林 | 0.017 | 3,757 | 1,068 | 144 |
| (-5.56%) | (-0.08%) | (1.29%) | (-9.92%) |
| 梯度提升 | **0.021** | **3,750** | **1,074** | 132 |
| **(16.7%)** | **(0.11%)** | **(0.74%)** | (-0.76%) |
| 前馈神经网络 | 0.019 | 3,753 | 1,111 | 139 |
| (5.56%) | (0.03%) | (-2.68%) | (-6.11%) |

注：括号中的数字表示该模型的指标相对于线性回归的提升比例。加粗数字表示该项指标最优。

**表Ⅰ7 使用2011-2013年样本的各模型样本外预测评价结果**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | R2 | RMSE | MAE | MAPE |
|  | (1) | (2) | (3) | (4) |
| 线性回归 | 0.020 | 3,623 | 842 | 119 |
| 拉索回归 | 0.020 | 3,624 | 842 | 121 |
| (0.00%) | (-0.03%) | (0.00%) | (-1.68%) |
| 岭回归 | 0.020 | 3,623 | 842 | 119 |
| (0.00%) | (0.00%) | (0.00%) | (0.00%) |
| 弹性网络 | 0.020 | 3,624 | 842 | 121 |
| (0.00%) | (-0.03%) | (0.00%) | (-1.68%) |
| 正交匹配追踪 | 0.020 | 3,623 | 841 | 120 |
| (0.00%) | (0.00%) | (0.12%) | (-0.84%) |
| 随机森林 | 0.022 | 3,620 | **831** | 128 |
| (10.0%) | (0.08%) | **(1.31%)** | (-7.56%) |
| 梯度提升 | **0.024** | **3,616** | 837 | 121 |
| **(20.0%)** | **(0.19%)** | (0.59%) | (-1.68%) |
| 前馈神经网络 | 0.023 | 3,619 | 870 | 130 |
| (15.0%) | (0.11%) | (-3.33%) | (-9.24%) |

注：括号中的数字表示该模型的指标相对于线性回归的提升比例。加粗数字表示该项指标最优。

**表Ⅰ8 各模型平均样本外预测评价结果**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | R2 | RMSE | MAE | MAPE |
|  | (1) | (2) | (3) | (4) |
| 线性回归 | 0.021 | 3,563 | 980 | 137 |
| 随机森林+梯度提升 | 0.023 | 3,559 | 970 | 143 |
| (9.52%) | (0.11%) | (1.02%) | (-4.38%) |
| 随机森林+前馈神经网络 | 0.022 | 3,560 | 955 | 137 |
| (4.76%) | (0.08%) | (2.55%) | (0.00%) |
| 梯度提升+前馈神经网络 | 0.023 | 3,559 | 958 | 132 |
| (9.52%) | (0.11%) | (2.24%) | (3.65%) |
| 随机森林+梯度提升+前馈神经网络 | 0.023 | 3,559 | 961 | 137 |
| (9.52%) | (0.11%) | (1.94%) | (0.00%) |

注：括号中的数字表示该模型的该指标相对于线性回归模型的提升比例。

**表Ⅰ9 第一种医保支付方式下以梯度提升模型预测结果为支付依据的结果**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | 无风险走廊 | |  | 单边风险走廊 | |  | 双边风险走廊 | |
| 乡镇 | 成本 | 收入 | 收入成本比率 |  | 收入 | 收入成本比率 |  | 收入 | 收入成本比率 |
|  | (1) | (2) | (3)=(2)/(1) |  | (4) | (5)=(4)/(1) |  | (6) | (7)=(6)/(1) |
| 1 | 6,113,787 | 5,661,917 | 0.926 |  | 5,661,917 | 0.926 |  | 5,661,917 | 0.926 |
| 2 | 10,125,700 | 10,278,938 | 1.015 |  | 10,278,938 | 1.015 |  | 10,278,938 | 1.015 |
| 3 | 12,775,556 | 12,732,441 | 0.997 |  | 12,732,441 | 0.997 |  | 12,732,441 | 0.997 |
| 4 | 11,525,740 | 11,974,936 | 1.039 |  | 11,974,936 | 1.039 |  | 11,974,936 | 1.039 |
| 5 | 10,818,740 | 9,901,470 | 0.915 |  | 9,901,470 | 0.915 |  | 9,901,470 | 0.915 |
| 6 | 8,140,266 | 7,709,576 | 0.947 |  | 7,709,576 | 0.947 |  | 7,709,576 | 0.947 |
| 7 | 13,266,642 | 14,444,435 | 1.089 |  | 14,444,435 | 1.089 |  | 14,444,435 | 1.089 |
| 8 | 12,926,489 | 12,504,842 | 0.967 |  | 12,504,842 | 0.967 |  | 12,504,842 | 0.967 |
| 9 | 11,891,490 | 13,549,697 | 1.139 |  | 13,549,697 | 1.139 |  | 13,549,697 | 1.139 |
| 10 | 890,593 | 1,138,413 | 1.278 |  | 1,138,413 | 1.278 |  | 1,001,918 | 1.125 |
| 11 | 4,631,023 | 4,196,296 | 0.906 |  | 4,196,296 | 0.906 |  | 4,196,296 | 0.906 |
| 12 | 7,791,352 | 8,224,794 | 1.056 |  | 8,224,794 | 1.056 |  | 8,224,794 | 1.056 |
| 13 | 11,492,682 | 11,529,872 | 1.003 |  | 11,529,872 | 1.003 |  | 11,529,872 | 1.003 |
| 14 | 16,203,138 | 14,745,581 | 0.910 |  | 14,745,581 | 0.910 |  | 14,745,581 | 0.910 |
| 整体财务风险（） | | | 1.051  (4.19%) |  |  | 1.051  (4.19%) |  |  | 0.898  (2.60%) |

注：括号中的数字表示该模型下的整体财务风险相对于线性回归模型下的降低比例。

**表Ⅰ10 第二种医保支付方式下以线性回归模型预测结果为支付依据的结果**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | 无风险走廊 | |  | 单边风险走廊 | |  | 双边风险走廊 | |
| 乡镇 | 成本 | 收入 | 收入成本比率 |  | 收入 | 收入成本比率 |  | 收入 | 收入成本比率 |
|  | (1) | (2) | (3)=(2)/(1) |  | (4) | (5)=(4)/(1) |  | (6) | (7)=(6)/(1) |
| 1 | 6,113,787 | 6,368,322 | 1.042 |  | 6,368,322 | 1.042 |  | 6,368,322 | 1.042 |
| 2 | 10,125,700 | 10,751,546 | 1.062 |  | 10,751,546 | 1.062 |  | 10,751,546 | 1.062 |
| 3 | 12,775,556 | 12,694,628 | 0.994 |  | 12,694,628 | 0.994 |  | 12,694,628 | 0.994 |
| 4 | 11,525,740 | 11,068,469 | 0.960 |  | 11,068,469 | 0.960 |  | 11,068,469 | 0.960 |
| 5 | 10,818,740 | 9,550,147 | 0.883 |  | 9,643,507 | 0.891 |  | 9,643,507 | 0.891 |
| 6 | 8,140,266 | 7,702,058 | 0.946 |  | 7,702,058 | 0.946 |  | 7,702,058 | 0.946 |
| 7 | 13,266,642 | 15,009,489 | 1.131 |  | 15,009,489 | 1.131 |  | 15,009,489 | 1.131 |
| 8 | 12,926,489 | 9,864,473 | 0.763 |  | 10,749,156 | 0.832 |  | 10,749,156 | 0.832 |
| 9 | 11,891,490 | 18,811,964 | 1.582 |  | 18,811,964 | 1.582 |  | 13,377,926 | 1.125 |
| 10 | 890,593 | 1,606,779 | 1.804 |  | 1,606,779 | 1.804 |  | 1,001,918 | 1.125 |
| 11 | 4,631,023 | 3,574,745 | 0.772 |  | 3,871,333 | 0.836 |  | 3,871,333 | 0.836 |
| 12 | 7,791,352 | 7,173,054 | 0.921 |  | 7,173,054 | 0.921 |  | 7,173,054 | 0.921 |
| 13 | 11,492,682 | 10,019,764 | 0.872 |  | 10,181,589 | 0.886 |  | 10,181,589 | 0.886 |
| 14 | 16,203,138 | 14,397,759 | 0.889 |  | 14,490,292 | 0.894 |  | 14,490,292 | 0.894 |
| 整体财务风险（） | | | 2.621 |  |  | 2.461 |  |  | 1.325 |

**表Ⅰ11 第二种医保支付方式下以梯度提升模型预测结果为支付依据的结果**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | 无风险走廊 | |  | 单边风险走廊 | |  | 双边风险走廊 | |
| 乡镇 | 成本 | 收入 | 收入成本比率 |  | 收入 | 收入成本比率 |  | 收入 | 收入成本比率 |
|  | (1) | (2) | (3)=(2)/(1) |  | (4) | (5)=(4)/(1) |  | (6) | (7)=(6)/(1) |
| 1 | 6,113,787 | 6,387,089 | 1.045 |  | 6,387,089 | 1.045 |  | 6,387,089 | 1.045 |
| 2 | 10,125,700 | 10,705,821 | 1.057 |  | 10,705,821 | 1.057 |  | 10,705,821 | 1.057 |
| 3 | 12,775,556 | 12,627,370 | 0.988 |  | 12,627,370 | 0.988 |  | 12,627,370 | 0.988 |
| 4 | 11,525,740 | 11,026,150 | 0.957 |  | 11,026,150 | 0.957 |  | 11,026,150 | 0.957 |
| 5 | 10,818,740 | 9,607,991 | 0.888 |  | 9,672,428 | 0.894 |  | 9,672,428 | 0.894 |
| 6 | 8,140,266 | 7,749,796 | 0.952 |  | 7,749,796 | 0.952 |  | 7,749,796 | 0.952 |
| 7 | 13,266,642 | 14,970,407 | 1.128 |  | 14,970,407 | 1.128 |  | 14,970,407 | 1.128 |
| 8 | 12,926,489 | 9,925,415 | 0.768 |  | 10,779,627 | 0.834 |  | 10,779,627 | 0.834 |
| 9 | 11,891,490 | 18,740,721 | 1.576 |  | 18,740,721 | 1.576 |  | 13,377,926 | 1.125 |
| 10 | 890,593 | 1,584,995 | 1.780 |  | 1,584,995 | 1.780 |  | 1,001,918 | 1.125 |
| 11 | 4,631,023 | 3,557,917 | 0.768 |  | 3,862,919 | 0.834 |  | 3,862,919 | 0.834 |
| 12 | 7,791,352 | 7,174,890 | 0.921 |  | 7,174,890 | 0.921 |  | 7,174,890 | 0.921 |
| 13 | 11,492,682 | 10,203,086 | 0.888 |  | 10,273,250 | 0.894 |  | 10,273,250 | 0.894 |
| 14 | 16,203,138 | 14,331,550 | 0.884 |  | 14,457,187 | 0.892 |  | 14,457,187 | 0.892 |
| 整体财务风险（） | | | 2.572  (1.87%) |  |  | 2.420  (1.67%) |  |  | 1.314  (0.83%) |

注：括号中的数字表示该模型下的整体财务风险相对于线性回归模型下的降低比例。

# 附录Ⅱ 机器学习模型与评价指标

我们所使用的七种机器学习模型可以被分为三个类别。第一类为变量选择类模型，包含了拉索回归（Tibshirani，1996）、岭回归（Hoerl and Kennard，1970）、弹性网络（Zhao and Yu，2006）和正交匹配追踪（Mallat and Zhang，1993）。拉索回归和岭回归在目标函数中加入了对于模型参数模的一阶惩罚项和二阶惩罚项；弹性网络是综合了两种惩罚方式，它选择一个一阶惩罚项与二阶惩罚项的组合进行优化，已有统计文献中将他们视为变量选择的方法。相似地，正交匹配追踪可以提取具有显著预测能力的特征变量，即非零参数，并删除多余的变量，直接限制了模型中非零参数的个数。第二类为基于决策树的模型，包含随机森林和梯度提升（Breiman et al.，1984）。决策树模型类似于计量经济学中的非参估计（Friedman，1999），该类模型通过将样本分割成数个子样本进行预测。在考虑到过拟合时，不同模型会对于模型估计过程做出不同限制。最后一类为神经网络模型，我们主要考察了前馈神经网络。该类模型是当下深度学习的基础建模方法（Goodfellow et al.，2016），采用多层隐藏层神经元与非线性变化以拟合各变量的非线性关系，并自动估计变量的交互影响。在我们训练前馈神经网络的过程中，不仅采用了系数惩罚项，还采用了随机丢弃法以提高模型的样本外预测能力。

接下来我们介绍机器学习模型的超参数设定。在带惩罚项的模型中，我们采用格点法，将惩罚项参数选取范围定为（0.001，0.01，0.1，0.5，1.0，3.0，5，7，10，20，100），以测试在各种可能情况下惩罚项对于模型回归结果的影响。拉索回归与岭回归模型分别通过调整一阶惩罚项与二阶惩罚项的参数以提升模型的预测能力。在弹性网络模型中，我们以相同的参数选取范围，同时调整两种惩罚项的权重。正交匹配追踪模型中非零参数个数调整范围为20—45个。基于决策树的模型的每个子节点至少需要50个样本。对于前馈神经网络模型，我们设计了三层网络构架，并调整每层神经元节点个数。模型在（10，9，8）中挑选第一层神经元数量，在（7，6，5，4）中挑选第二层神经元数量，在（3，2，1）中挑选第三层数量。同时，我们在30-200次间调整模型的最大迭代次数，通过K折交叉验证方法（K=10）选择最优超参数。[[1]](#footnote-0) 表Ⅱ1总结了我们最终采用的各类模型所对应的超参数设定。

**表Ⅱ1 预测模型及其超参数**

|  |  |  |
| --- | --- | --- |
| 名称 | 模型英文名称 | 超参数 |
| 线性回归 | Linear Regression，LR | - |
| 拉索回归 | Least Absolute Shrinkage and Selection Operator，Lasso | 一阶模惩罚系数=1 |
| 岭回归 | Ridge Regression | 二阶模惩罚系数=20 |
| 弹性网络 | Elastic Net | 一阶模惩罚系数=0.001；二阶模惩罚系数=0.001 |
| 正交匹配追踪 | Orthogonal Matching Pursuit，OMP | 参数个数=30 |
| 随机森林 | Random Forest | 子决策树个数=200；最大深度=5 |
| 梯度提升 | Gradient Boosting Decision Tree，GBDT | 子决策树个数=100；最大深度=3 |
| 前馈神经网络 | Feedforward Neural Network，FNN | 三层隐藏层；神经元（10,7,3） |

我们采用了多个指标用以更加全面地衡量不同预测模型在不同维度上的性能。除了最为常见的评价指标R2（拟合优度），我们还采用了RMSE（Root mean square error，均方根误差）、MAE（Mean absolute error，平均绝对误差）以及MAPE（Mean absolute percentage error，平均绝对百分误差）。上述各项对于模型的评价指标的定义如下：

其中，表示个体()，是模型预测值，是真实值，表示真实值的平均值。R2中分子为预测值与真实值的平方差之和，分母为均值与真实值的平方差之和。它表示我们感兴趣的变量偏离均值的全部变异通过模型所能解释的比例，可以近似看作模型预测值与真实值的Pearson相关系数之平方。R2越接近1表示模型的拟合能力越好。RMSE是模型预测值相较于真实值之差平方的期望值的算数平方根，该值越小表示模型拟合误差越小，拟合越优。MAE代表模型预测值与真实值之差的绝对值的平均，该指标越小表示模型的误差越小。基于高费用人群本身可能存在更大不确定性的特征，我们还采用了MAPE指标。MAPE反映了模型预测值与真实值之间的偏离相对于真实值的比例。与前述指标类似，MAPE越小表示模型的预测更为准确。由于真实值取值可能为0，因此我们在计算MAPE的过程中对于所有医疗费用以真实取值加1取代。

# 附录Ⅲ 其他描述性统计



**图Ⅲ1 2012年人口金字塔**



**图Ⅲ2 2012年各年龄分性别平均个人年度医疗费用**



**图Ⅲ3 2012年按疾病分组平均个人年度医疗费用**

**表Ⅲ1 2012-2014年分类别年度医疗费用描述性统计**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | 2012年 | 2013年 | 2014年 |
|  |  | (1) | (2) | (3) |
| *门诊费用* |  |  |  |  |
| 零值比例（%） |  | 38.30 | 32.91 | 30.74 |
| 非零费用（元） |  |  |  |  |
| 均值 |  | 163 | 191 | 176 |
| 25%分位 |  | 39 | 51 | 54 |
| 50%分位 |  | 97 | 128 | 129 |
| 75%分位 |  | 219 | 271 | 248 |
| 95%分位 |  | 537 | 574 | 495 |
| 99%分位 |  | 870 | 874 | 741 |
| 最大值 |  | 3,782 | 2,768 | 2,935 |
| *住院费用* |  |  |  |  |
| 零值比例（%） |  | 90.54 | 89.76 | 88.89 |
| 非零费用（元） |  |  |  |  |
| 均值 |  | 4,301 | 4,980 | 5,096 |
| 25%分位 |  | 931 | 1,074 | 1,167 |
| 50%分位 |  | 2,032 | 2,242 | 2,392 |
| 75%分位 |  | 4,468 | 4,957 | 5,144 |
| 95%分位 |  | 15,434 | 17,098 | 17,890 |
| 99%分位 |  | 39,237 | 48,961 | 48,299 |
| 最大值 |  | 186,686 | 433,970 | 220,381 |
| *慢病费用* |  |  |  |  |
| 零值比例（%） |  | 98.89 | 98.72 | 98.50 |
| 非零费用（元） |  |  |  |  |
| 均值 |  | 566 | 633 | 690 |
| 25%分位 |  | 97 | 129 | 169 |
| 50%分位 |  | 284 | 378 | 435 |
| 75%分位 |  | 807 | 918 | 1,031 |
| 95%分位 |  | 2,072 | 2,097 | 2,138 |
| 99%分位 |  | 2,847 | 2,855 | 2,827 |
| 最大值 |  | 5,561 | 4,301 | 4,930 |

**表Ⅲ2 2012-2013年新增预测变量分布（%）**

|  |  |  |
| --- | --- | --- |
|  | 2012年 | 2013年 |
|  | (1) | (2) |
| 婚姻状况 |  |  |
| 已婚 | 30.1 | 29.9 |
| 未婚、离婚或丧偶 | 69.9 | 70.1 |
| 职业状况 |  |  |
| 从事于国家机关、党群组织、企业、事业单位的人员 | 0.09 | 0.09 |
| 从事于科学研究和专业技术工作的人员 | 0.19 | 0.17 |
| 从事于商、服务、农、林、牧、渔、水利、生产和运输业的人员 | 97.4 | 97.4 |
| 其他从业人员 | 0.21 | 0.22 |
| 退休或无业 | 2.14 | 2.09 |
| 受教育水平（最高学历） |  |  |
| 文盲或半文盲 | 17.5 | 17.4 |
| 小学 | 34.5 | 34.9 |
| 初中 | 38.5 | 38.5 |
| 高中、中技或中专 | 8.29 | 8.13 |
| 大学专科及以上 | 0.43 | 0.38 |

**表Ⅲ3 2012-2013年前三年病种分布（%）**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | 2012年 | 2013年 |
|  | | (1) | (2) |
| A00-B99 | 某些传染病和寄生虫病 | 4.33 | 6.17 |
| C00-D48 | 肿瘤 | 1.20 | 1.29 |
| D50-D89 | 血液和造血器官疾病以及某些涉及免疫机能的异常 | 1.88 | 1.01 |
| E00-E90 | 内分泌、营养和代谢疾病 | 2.97 | 2.70 |
| F00-F99 | 精神和行为障碍 | 0.54 | 0.51 |
| G00-G99 | 神经系统疾病 | 1.76 | 2.78 |
| H00-H59 | 眼和附器疾病 | 2.92 | 4.08 |
| H60-H95 | 耳和乳突疾病 | 2.19 | 2.80 |
| I00-I99 | 循环系统疾病 | 9.33 | 12.2 |
| J00-J99 | 呼吸系统疾病 | 80.7 | 79.7 |
| K00-K93 | 消化系统疾病 | 47.2 | 52.3 |
| L00-L99 | 皮肤和皮下组织疾病 | 12.0 | 19.3 |
| M00-M99 | 肌肉骨骼系统和结缔组织疾病 | 25.6 | 35.8 |
| N00-N99 | 泌尿生殖系统疾病 | 17.4 | 19.7 |
| O00-O99 | 妊娠、分娩和产褥期 | 0.67 | 0.83 |
| P00-P96 | 起源于围生期的某些疾病 | 0.01 | 0.06 |
| Q00-Q99 | 先天性畸形、变形和染色体异常 | 0.12 | 0.18 |
| R00-R99 | 症状、体征和异常的临床和化验结果 | 31.0 | 32.2 |
| S00-T98 | 损伤、中毒和外因的某些其它结果 | 14.1 | 20.4 |
| V01-Y98 | 发病和死亡的外因 | 4.80 | 5.23 |
| Z00-Z99 | 影响健康状况和接触健康服务的因素 | 2.66 | 4.13 |

# 附录Ⅳ 医保支付方式和政府补贴方式设计

第一种医保支付方式下新农合对乡镇卫生院的支付额的具体计算过程如下。首先，根据线性回归模型或梯度提升模型预测得到居民的当期医疗费用，分别记为和。其次，个体的预测费用除以总体预测费用均值，得到医疗费用风险评分，具体见公式（5）：

其中，风险评分代表了居民在所有参保居民中医疗费用风险的相对水平，表示以线性回归模型预测结果为依据，表示以梯度提升模型预测结果为依据。再次，居民的人头费由当年新农合的预算总额和个体的风险评分确定。具体见公式（6）：

其中表示了平均支付水平。预算越高，支付水平越高。风险评分体现了对个体的风险调整。居民费用风险越高，新农合支付给为该居民服务的乡镇卫生院的人头费就越多。最后，乡镇卫生院收到的支付总额为其服务的参与新农合的居民的人头费之和，具体见公式（7）：

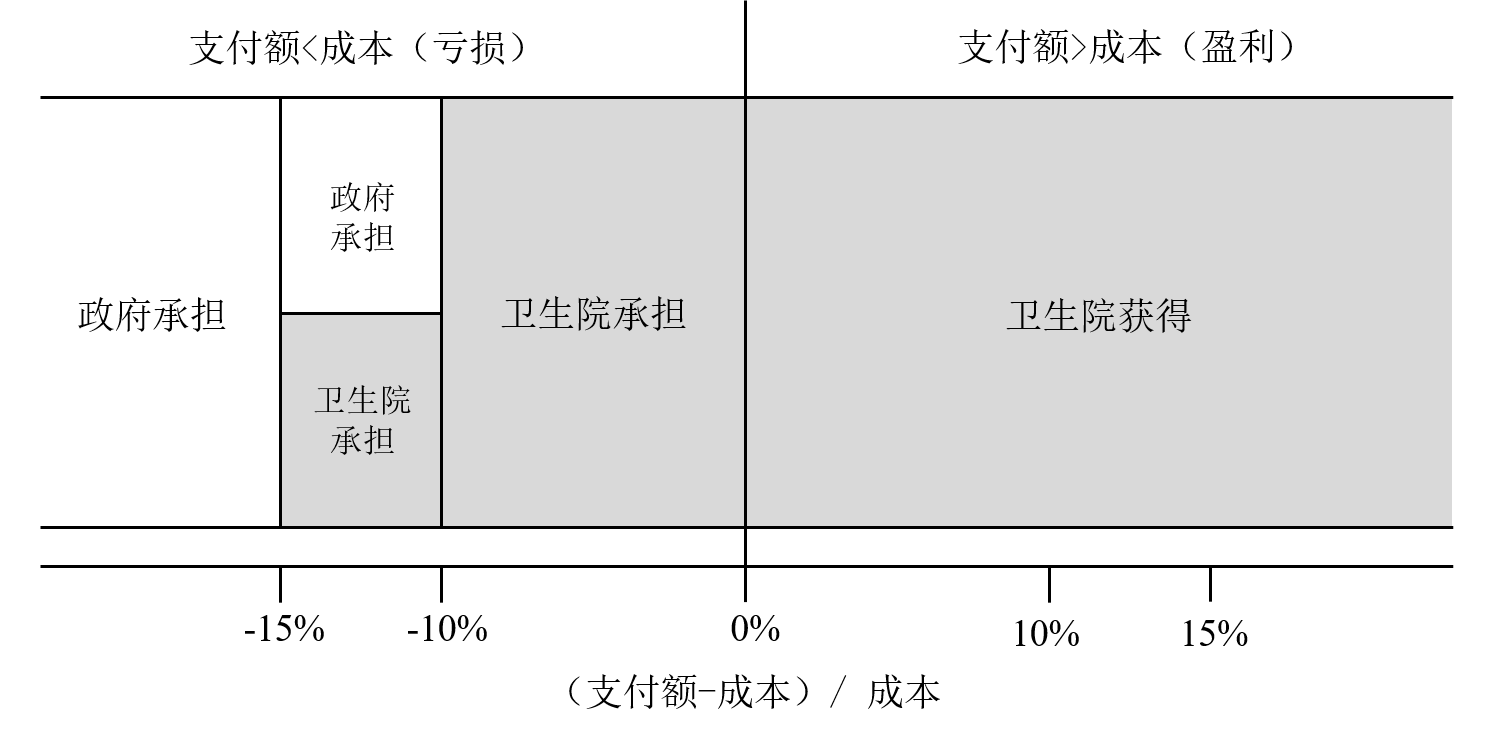
其中，为乡镇卫生院服务的参与新农合的居民的集合。

第二种医保支付方式由按人头付费与按服务付费相结合。与第一种方式类似，对每个居民的支付是经医疗费用风险调整后的按人头付费，而不同之处在于，第二种方式允许居民选择多个乡镇卫生院就诊，居民的人头费根据居民的医疗服务使用情况分配给其就诊的乡镇卫生院。因此，就诊人次更多的乡镇卫生院和医疗服务价格更高的乡镇卫生院将获得更高的支付额。第二种支付方式下新农合对乡镇卫生院的支付额的具体计算过程如下。设乡镇卫生院的医疗服务价格为（实际计算中为乡镇卫生院的平均就诊费用），居民在乡镇卫生院的就诊次数为。乡镇卫生院为居民服务所得到的支付由居民在乡镇卫生院就诊的权重和居民本身的人头费决定。具体见公式（8）：

其中的定义与计算方式与前文一致，和分别表示以线性回归模型和梯度提升模型预测结果为依据计算出的居民的人头费。乡镇卫生院的医疗服务价格越高，居民去该乡镇卫生院的就诊次数越多，该乡镇卫生院的权重就越大，获得的支付就越多。类似地，居民的风险水平越高，人头费越高，乡镇卫生院的获得的支付就越多。乡镇卫生院收到的支付总额由该乡镇卫生院所服务的所有居民的支付之和。具体见公式（9）：

根据以上两种以按人头付费为基础设计的体系，新农合支付并不依赖于每次就诊的实际支出，因此都有助于纠正医疗服务提供者过度使用医疗服务的动机。为了降低医疗机构的收入不确定性，我们引入政府干预来分担乡镇卫生院的财务风险。在这里，我们提供两种政府分担方式。第一种是单边风险走廊（one-sided risk corridor），政府与乡镇卫生院共担亏损。具体的分担方式如下。当乡镇卫生院亏缺率小于10%时，政府不做任何补贴。当乡镇卫生院的亏缺率在10%到15%之间时，乡镇卫生院承担10%的亏损，余下亏损由政府与乡镇卫生院平均分担。当乡镇卫生院的亏缺率大于15%时，乡镇卫生院只承担12.5%的亏损，余下亏损由政府承担。政府不参与分担乡镇卫生院盈利，乡镇卫生院获得所有盈利。具体如图Ⅳ1所示。在这种模式下，乡镇卫生院收到的政府补贴由公式（10）确定：

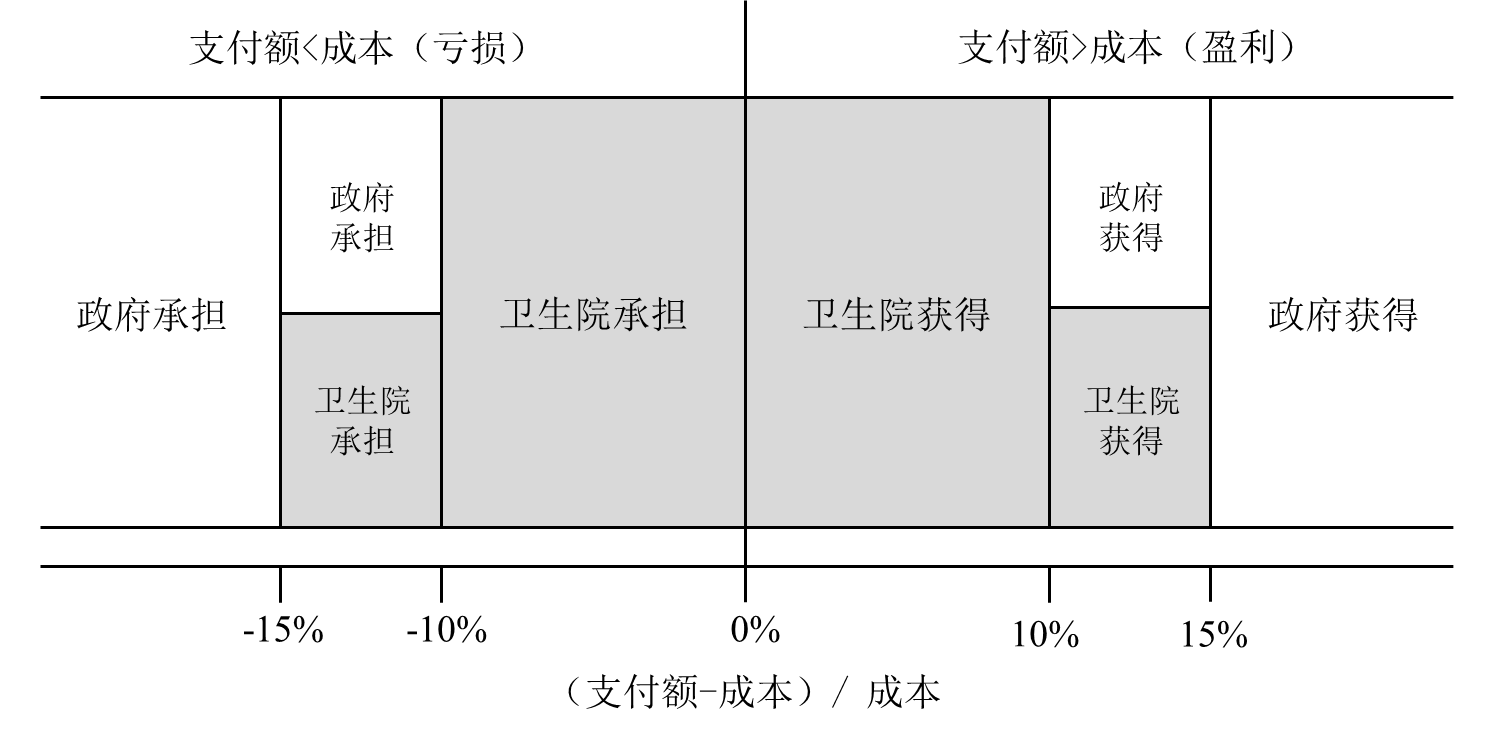
其中是乡镇卫生院的成本，是乡镇卫生院在政府补偿前应获得的新农合支付总额。



**图Ⅳ1 单边风险走廊**

第二种是双边风险走廊（two-sided risk corridor），政府不仅与乡镇卫生院分担亏损，也与乡镇卫生院分担盈利。图Ⅳ2展示了双边风险走廊下乡镇卫生院的盈利或亏损的分配。具体而言，政府分担乡镇卫生院亏损的方式与单边风险走廊相同。同时，政府以类似的方式分担乡镇卫生院的盈利，因此乡镇卫生院最多获得12.5%盈利。双边风险走廊下乡镇卫生院收到的政府补贴（或者抽成）由公式（11）确定：

其中和的定义与前文一致。



**图Ⅳ2 双边风险走廊**

# 参考文献

1. Bertsimas, D., M. V. Bjarnadóttir, M. A. Kane, J. C. Kryder, R. Pandey, S. Vempala, and G. Wang,“Algorithmic Prediction of Health–Care Costs”, *Operations Research*, 2008, 56(6), 1382–1392.
2. Friedman, J. H.,“Stochastic Gradient Boosting”, *Computational Statistics & Data Analysis*, 2002, 38(4), 367-378.
3. Goodfellow, I., Y. Bengio, and A. Courville,“Deep Learning”, MIT Press, 2016.
4. Hoerl, A. E., and R. W. Kennard,“Ridge Regression: Biased Estimation for Nonorthogonal Problems”, *Technometrics*, 1970, 12(1), 55-67.
5. Mallat, S. G., and Z. Zhang,“Matching Pursuits with Time-Frequency Dictionaries”, *IEEE Transactions on Signal Processing*, 1993, 41(12), 3397-3415.
6. Tibshirani, R.,“Regression Shrinkage and Selection via the Lasso”, *Journal of the Royal Statistical Society: Series B (Methodological)*, 1996, 58(1), 267-288.
7. Zhao, P., and B. Yu,“On Model Selection Consistency of Lasso”, *Journal of Machine Learning Research*, 2006, 7(2), 2541–2563.

**注：该附录是期刊所发表论文的组成部分，同样视为作者公开发表的内容。如研究中使用该附录中的内容，请务必在研究成果上注明附录下载出处**。

1. K折交叉验证（K-Fold cross-validation）是指将数据集划分为K个大小相似的互斥子集，并在划分时保持数据分布的一致性，每次用K-1个子集的并集作为训练集，剩余的做测试集，进行K次训练，最后取K次结果的均值。 [↑](#footnote-ref-0)