八、生产线、操作人员与产量、合格率的分析

8.1 数据预处理

鉴于赛题提供的十条生产线数据量较大，直接分析导致运算效率低下。因此，对数据进行预处理是必要的。具体地，应以天为颗粒度对数据进行汇总，计算出每天的产量、合格率以及每个故障的发生次数和相应的持续时间。这样的预处理方式不仅有助于减少数据冗余，提高分析效率，还能为后续探究产量、合格率与生产线、操作人员等因素之间的关系提供便利。如图 8-1

生产线编号	日期	产量	合格率(%)	操作人员编号	工龄	故障1	故障2	故障3	故障4	故障5	故障6	故障7	故障8	故障9
M301	1	1416	100.0	A001	1	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
M301	2	1440	100.0	A001	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M301	3	1420	100.0	A001	1	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0
M301	4	1444	100.0	A001	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M301	5	1356	99.34	A001	1	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0

故障1持续时间	故障2持续时间	故障3持续时间	故障4持续时间	故障5持续时间	故障6持续时间	故障7持续时间	故障8持续时间	故障9持续时间
0.0	174.0	0.0	169.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
181.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	350.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

图 8-1 预处理后的数据图

8.2 各因素间的相关性系数分析

为了深入探究特征之间的潜在关系，可以使用皮尔逊相关系数矩阵量化特征之间的线性相关程度。

皮尔逊相关系数矩阵是一种统计分析方法，用于衡量数据集中多个特征之间的线性关系强度。通过计算每个特征对之间的皮尔逊相关系数，我们可以得到一个描述了数据集中特征之间相关性的矩阵，如图 8-2

皮尔逊相关系数的计算公式如下：

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

皮尔逊相关系数绝对值越接近 1，说明这两个特征之间存在很强相关性；如果接近 0，说明几乎没有线性关系[10]。

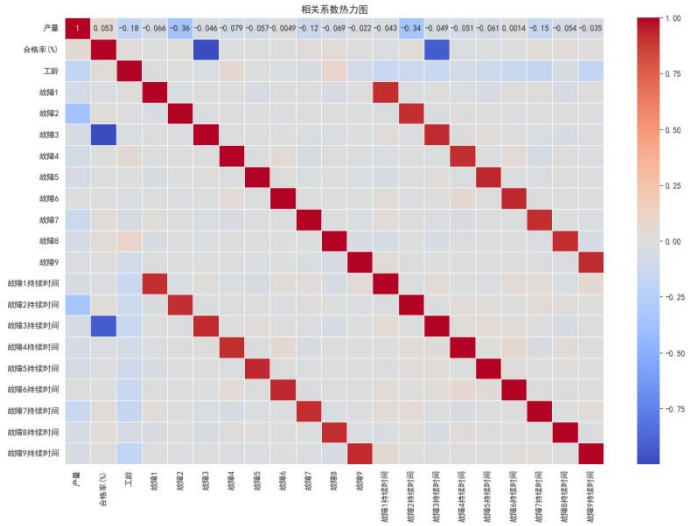

图 8-2 各因素间的相关系数热力图

观察可知，产量与工龄、故障及故障持续时间这几个因素之间存在微弱的相关关系，即随着工龄的增加、故障次数的增多、故障持续时间的增加，产量可能会略有下降，其中故障 2（即物料检测装置故障 2001）和故障 2 持续时间这两个因素对产量的影响较大。

合格率与产量、工龄之间的相关系数为 0.03-0.05，这表示合格率与产量、工龄之间存在微弱的正相关关系。

合格率与故障及故障持续时间等故障因素之间的相关系数有正有负，但大部分数值的绝对值都较小，说明合格率与这些故障因素之间的相关性不强。其中，合格率与故障 3 的相关系数为-1.00，表示它们之间有强烈的负相关关系。这意味着当故障 3 发生时，合格率会显著降低。

除了故障 4、故障 8 与工龄之间存在微弱的正相关关系外，其他故障与工龄之间的线性关系不明显，这意味着工龄不是大多数故障发生的主要影响因素。

大部分故障之间并没有显著的线性关系，说明这些故障可能是独立的事件，或者它们之间的关系复杂且非线性。

故障持续时间之间虽然存在一定的相关性，但并不强烈，说明不同故障的持续时间可能受到一些共同因素的影响，但这些因素的影响并不显著。

8.3 生产线与产量的关系分析

为探究不同生产线的产量是否存在显著差异，以生产线为分组变量，产量为因变量进行分析，构造箱线图和折线图。见图 8-3、8-4、8-5

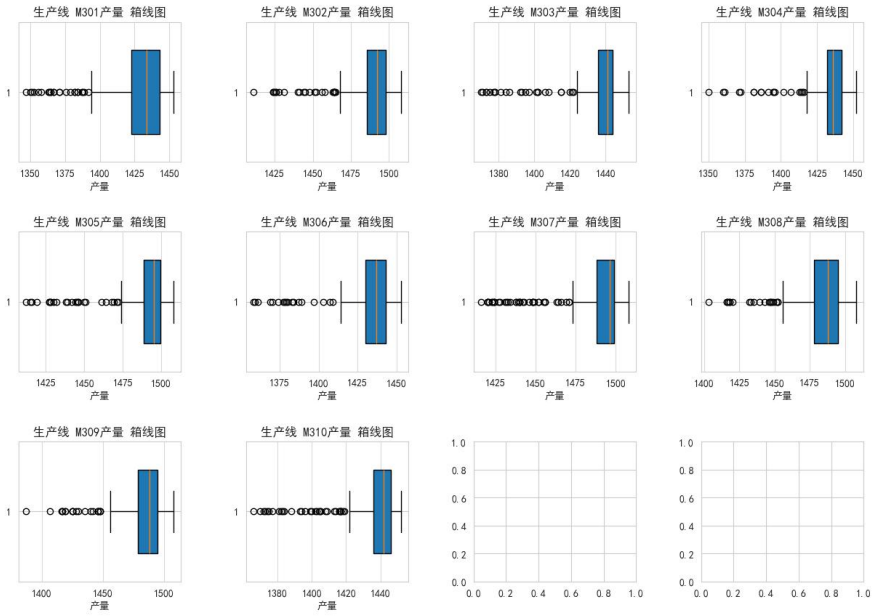

图 8-3 各生产线产量分布箱线图

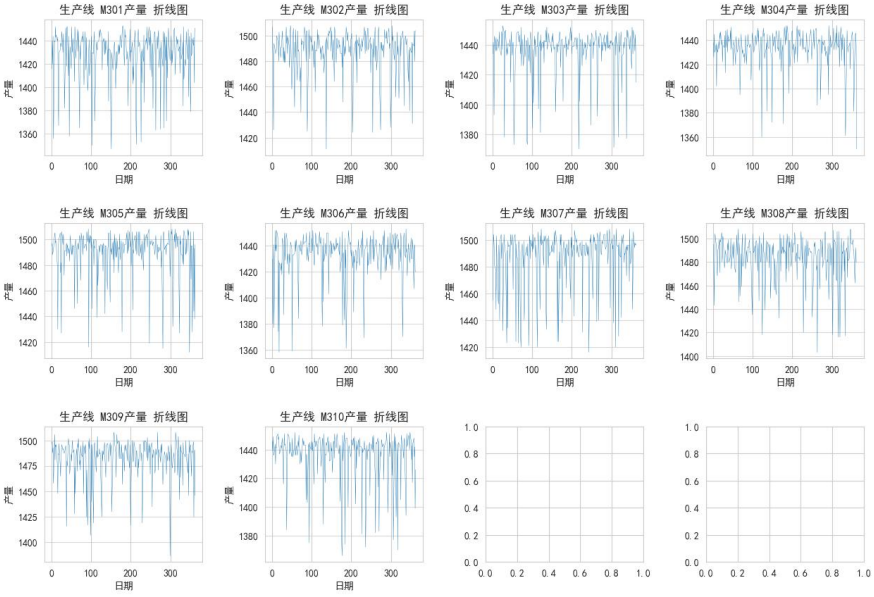

图 8-4 各生产线产量折线图

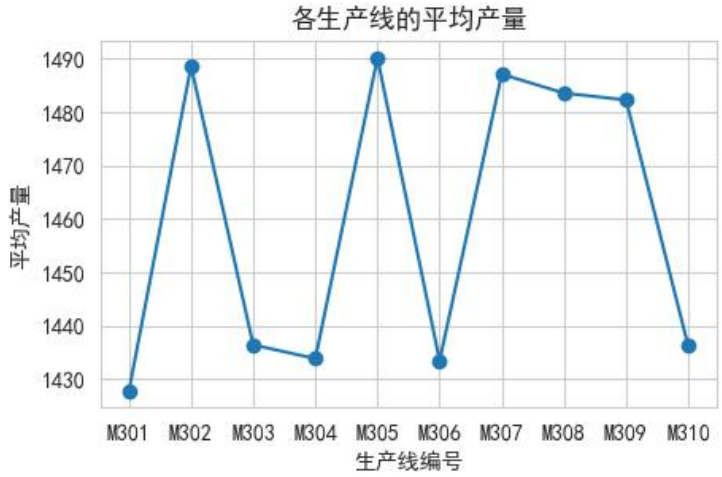

图 8-5 各生产线平均产量折线图

经过对生产线的产量数据进行分析，M301、M303、M304、M306、M310 这五条生产线的产量主要分布在 1425-1450 的范围内，平均日产量约为 1425-1440。相比之下，M302、M305、M307、M308、M309 这五条生产线的产量则集中在 1480-1500 的区间，平均日产量大致为 1480-1490。尽管最高平均产量与最低平均产量之间存在约 62 的差距，但这一差异仅占日产量的 4.1%，因此可以初步判断生产线对产量的影响并不显著。

为探究生产线的产量与故障发生次数及故障持续时间是否存在相关性，以生产线为分组变量，构造故障发生次数和持续时间的折线图。如图 8-6、8-7

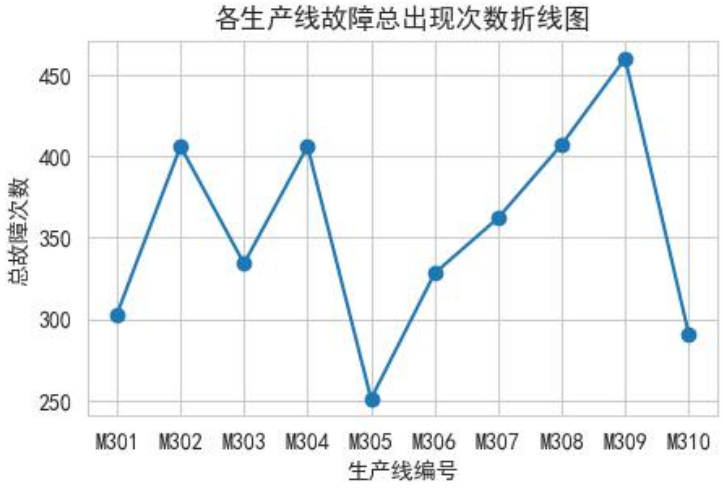

图 8-6 各生产线故障总出现次数折线图

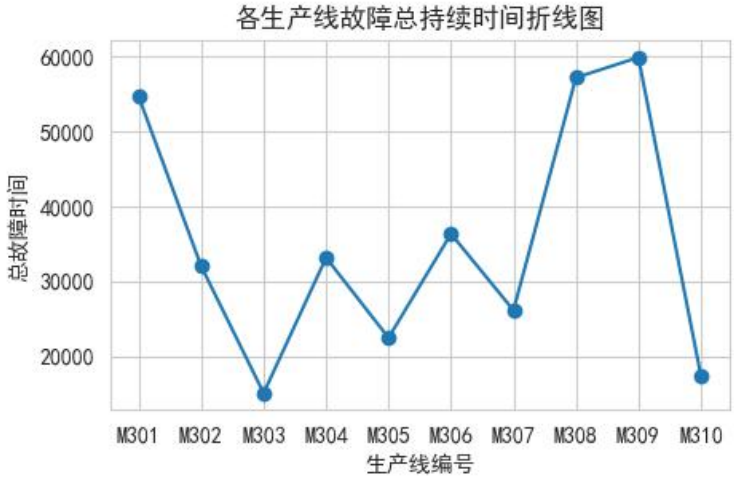

图 8-7 各生产线故障总持续时间折线图

通过对比这两张图表和图 8-5 的趋势，发现它们之间并未呈现出明显的相关性。例如，M305 和 M310 两条生产线故障次数均较少且故障持续时间较短，但 M305 的平均产量比 M310 高出近 55。

为进一步挖掘其是否存在相关性，在以生产线为分组变量的基础上再按故障类型细分，绘制各故障发生次数和持续时间图，如图 8-8、8-9

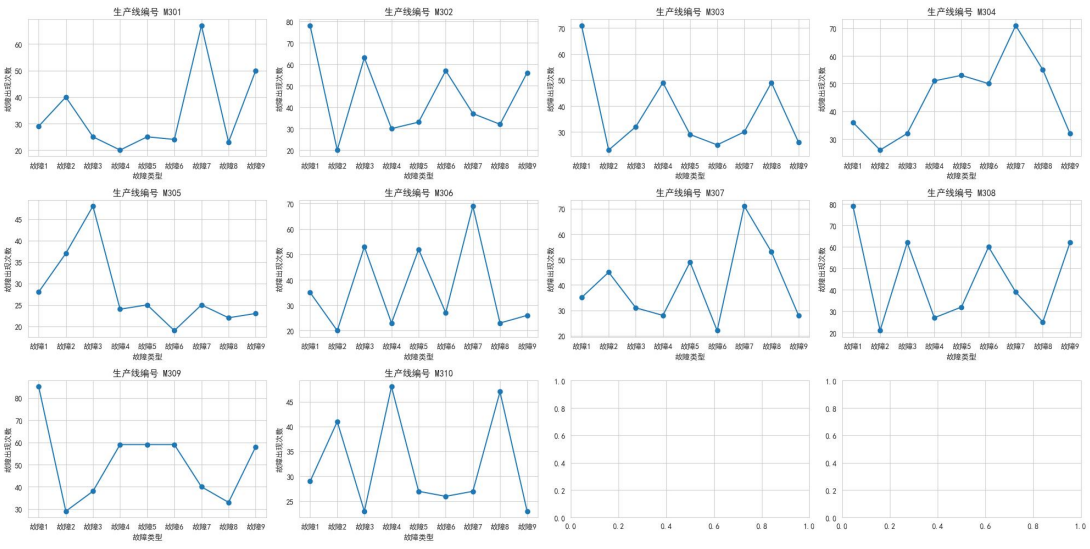

图 8-8 各生产线各故障发生次数图

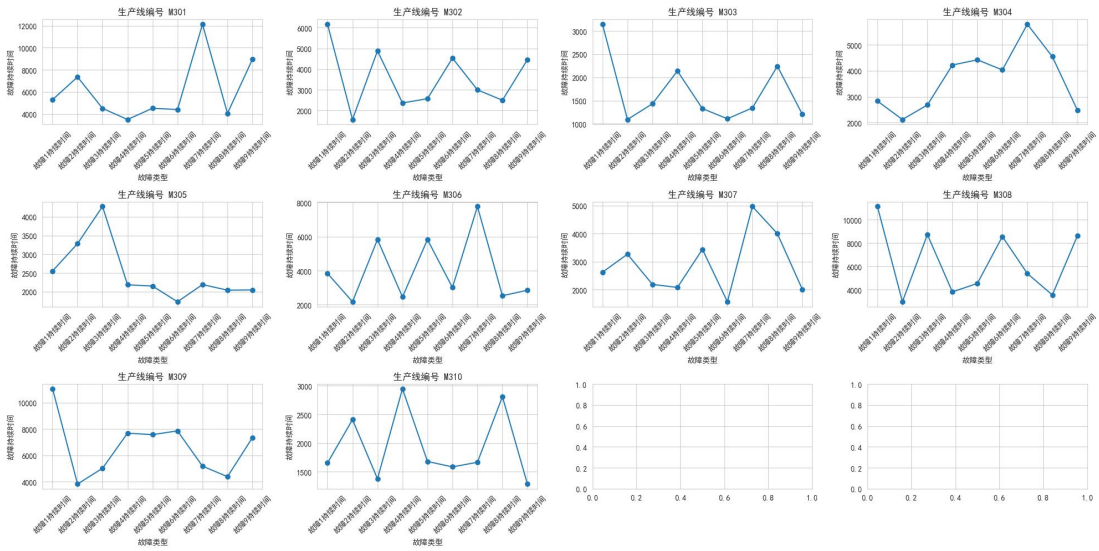

图 8-9 各生产线各故障持续时间图

故障发生次数和故障持续时间表现出正相关性，然而这种正相关性并未能明显反映出与产量之间的因果联系。通过热力图所展示的相关系数，故障 2（即物料检测装置故障 2001）及其持续时间与产量的相关性相对明显，故绘制故障 2 相关图表，如图 8-10、8-11

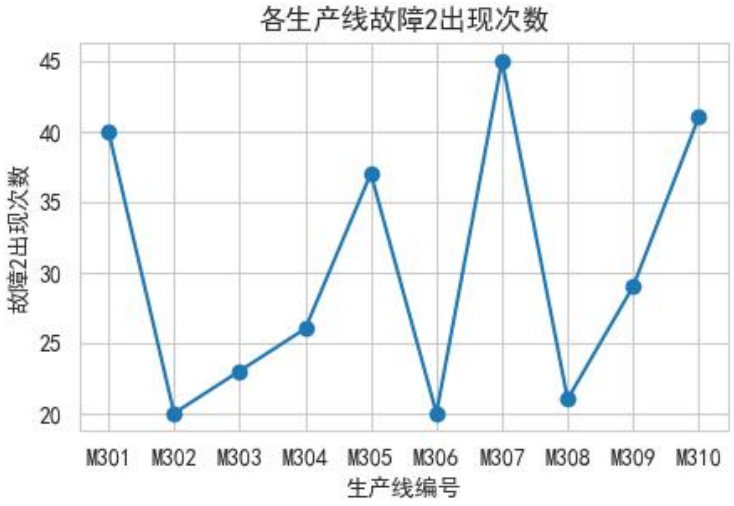

图 8-10 各生产线故障 2 出现次数图

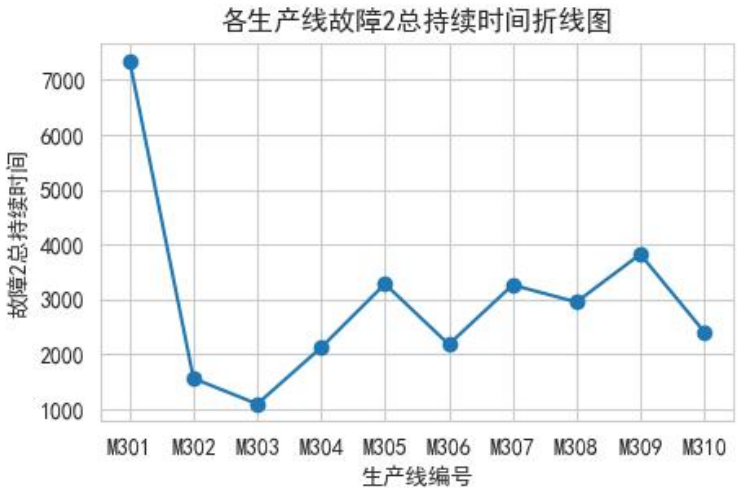

图 8-11 各生产线故障 2 总持续时间图

进一步绘制故障 2 的相关图表后，确实观察到故障 2 对产量存在负相关性，如在 M301 和 M302 生产线中表现显著。但值得注意的是，这种负相关性并非适用于所有生产线，如 M310 生产线并未呈现明显的相关性。因此，可以得出结论，产量与故障及故障持续时间之间存在微弱的负相关关系，但影响较小。

8.4 生产线与合格率的关系分析

为探究不同生产线的合格率是否存在显著差异，以生产线为分组变量，合格率为因变量进行分析，构造箱线图和折线图，如图 8-12、8-13、8-14

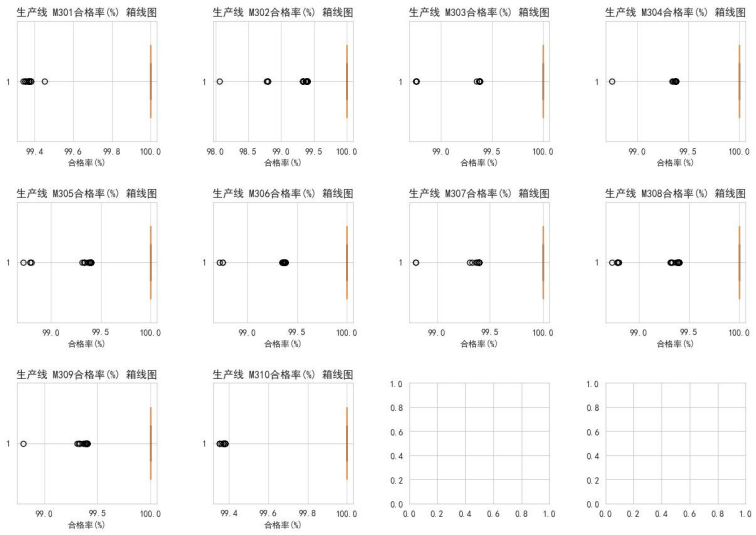

图 8-12 各生产线合格率分布箱线图

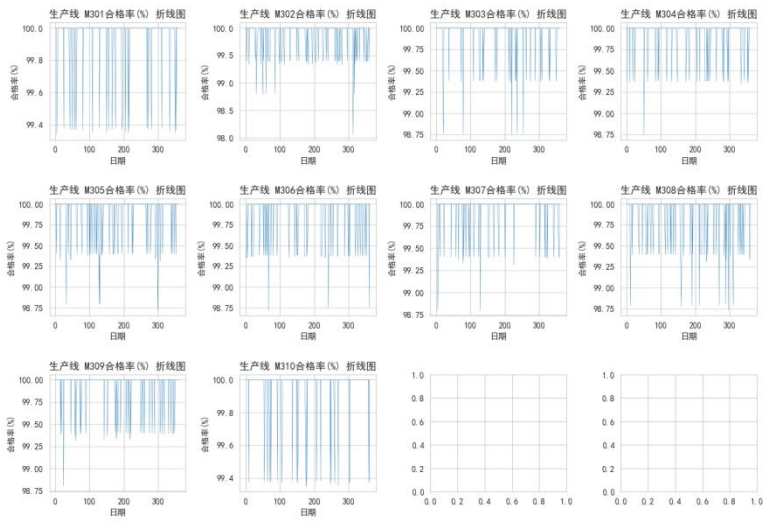

图 8-13 各生产线合格率折线图

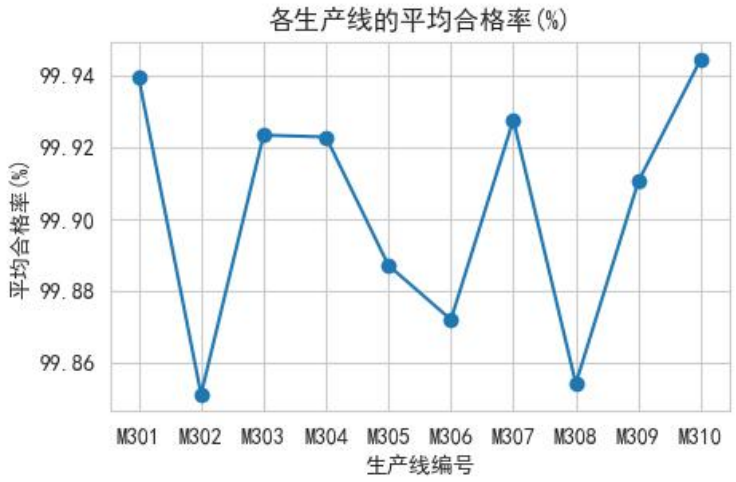

图 8-14 各生产线平均合格率折线图

分析结果显示，各条生产线每天的合格率主要维持在 100%，少数离群点出现在 98.75%-99.5%的范围内。平均合格率最高与最低之间的差异仅为 0.09%，这表明生产线对合格率的影响并不显著。

热力图分析显示，生产线发生故障时其合格率会有所降低。为了深入剖析故障与次品出现之间的关系，绘制了各生产线故障与次品出现时间的散点图，其中红色点表示次品出现时间，蓝色点表示故障出现时间。如图 8-14--8-17

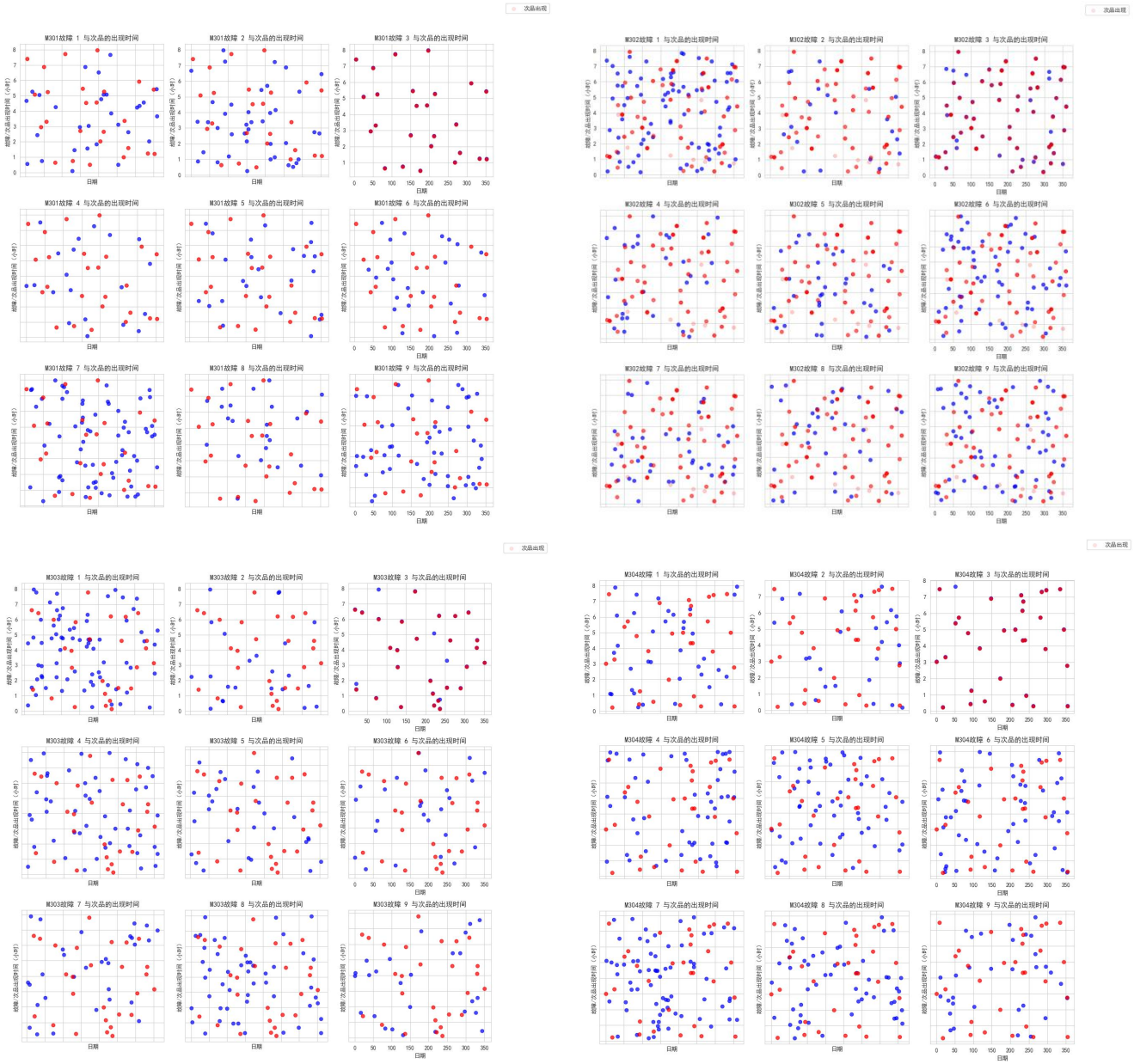

图 8-14--8-17 各生产线故障与次品出现时间的散点图（仅展示 4 条）

通过观察散点图，发现故障 3（即填装装置检测故障 4001）的出现时间与次品出现时间存在高度重合，这强烈表明故障 3 与合格率的降低之间存在显著的负相关关系。然而，其他故障的出现时间与次品出现时间的重合度并不明显，因此合格率与这些故障因素之间的相关性相对较弱。

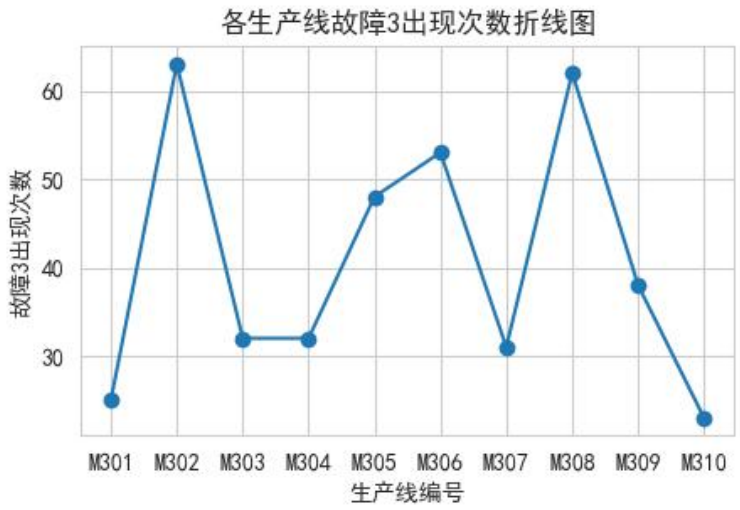

图 8-18 各生产线故障 3 出现次数折线图

为深入探究故障 3 与合格率之间的关系，进一步绘制了各生产线故障 3 出现次数的折线图（8-18）。通过对比图 8-14，发现两者趋势截然相反，故障 3 出现次数增多时，平均合格率则呈现下降趋势，更加明确了故障 3 对合格率产生的负面影响。

8.5 工龄与产量的关系分析

为探究不同工龄的操作人员的产量是否存在显著差异，以工龄为分组变量，产量为因变量进行分析，构造折线图，如图 8-19

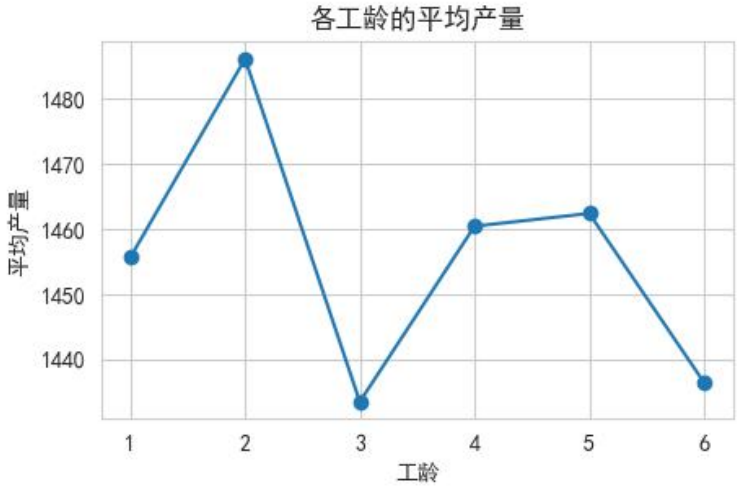

图 8-19 各工龄平均产量折线图

可以看出，工龄与产量的关系并非简单的线性关系。工龄为两年的操作人员平均产量最高，而工龄为三年的操作人员平均产量最低，随后是工龄为六年的操作人员。这一现象表明，工龄并非是影响产量的唯一或主要因素。

为了深入探究操作人员工龄与产量的关系，以工龄为分组变量，绘制各个故障的持续时间，如图 8-20

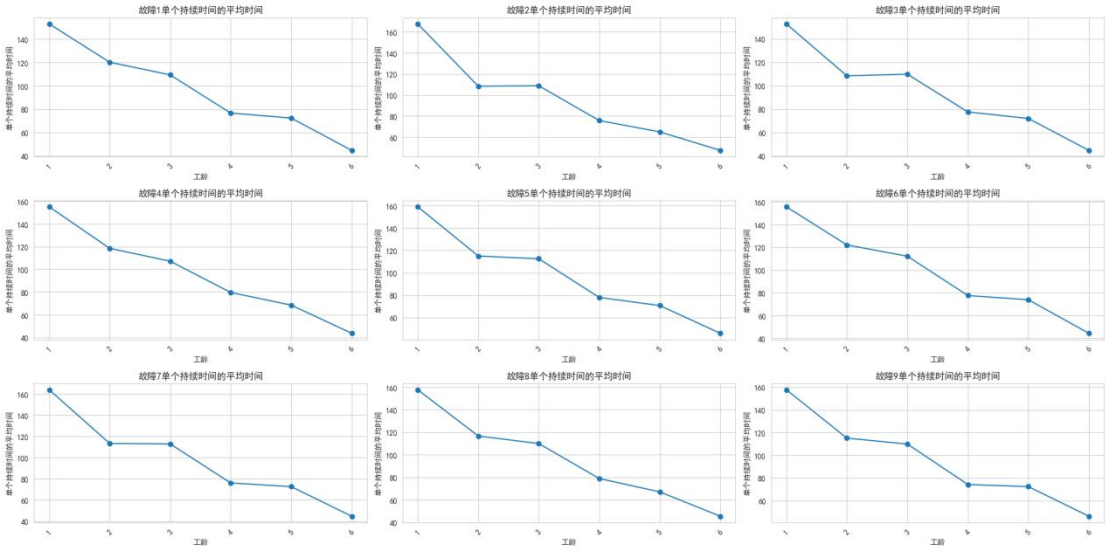

图 8-20 各工龄工作人员各个故障持续时间图

可以看出，工龄与故障持续时间之间存在明显的负相关性。随着工龄的增加，每个故障的持续时间都在减少，这表明工龄较大的操作人员在解决生产线故障方面具备更高的效率。然而，考虑到每条生产线的工作时间大致相同，如果按照上述结论推断，工龄越大的操作人员在相同时间内应该产量越多，但实际情况并不符合这一预期。因此，进一步绘制了以工龄为分组变量的每天各个故障持续时间堆叠图，如图 8-21--8-26

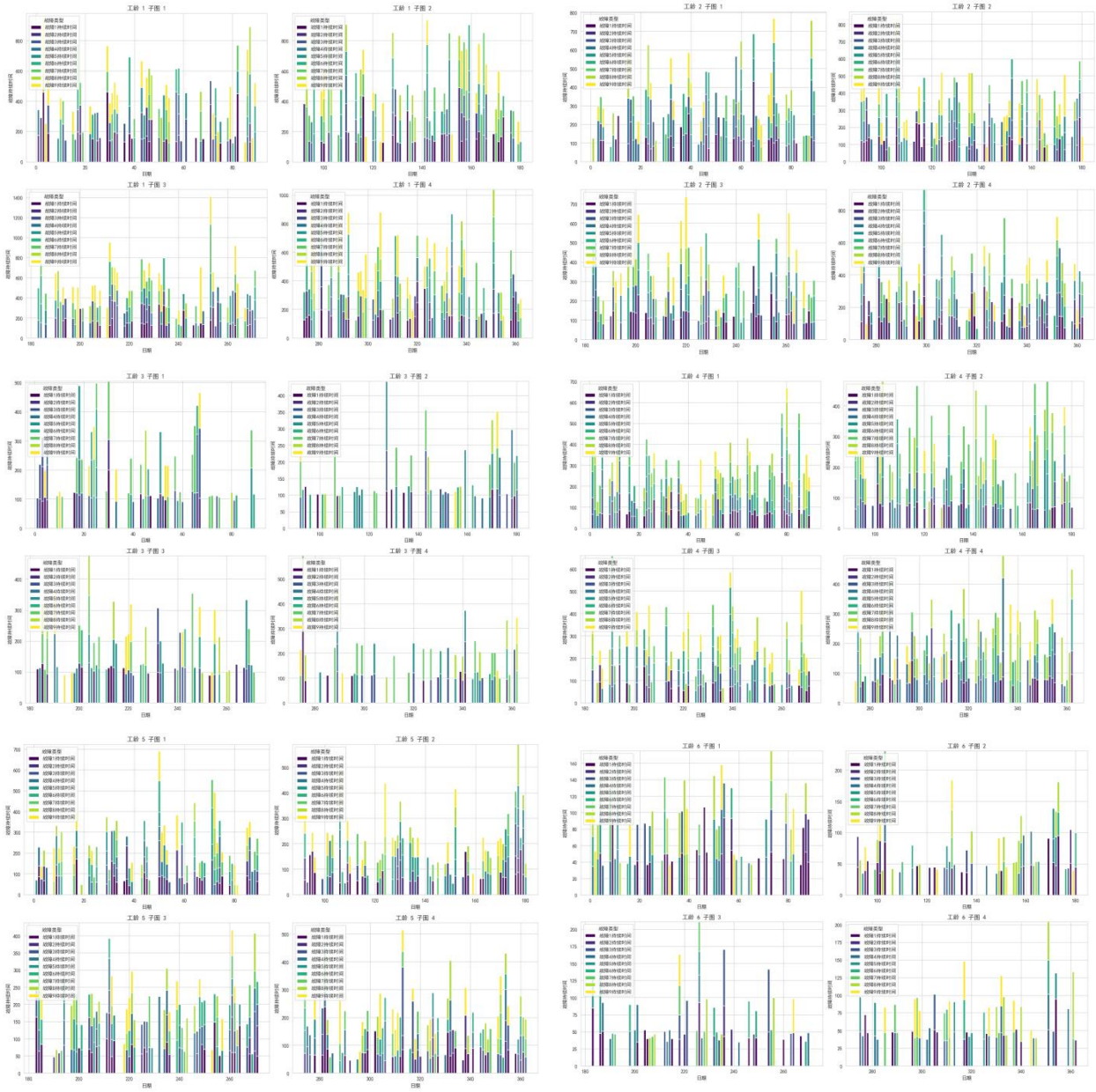

图 8-21--8-26 各工龄各故障持续时间堆叠图

分析结果显示，尽管工龄增加导致故障持续时间减少，但由于各生产线故障发生次数的不同，工龄与产量之间并未呈现出显著的正相关关系。假设各生产线故障发生次数相同，那么工龄越大的操作人员产量确实可能更高。

8.6 工龄与合格率的关系分析

为探究不同工龄的操作人员的合格率是否存在显著差异，以工龄为分组变量，合格率为因变量进行分析，构造箱线图和折线图，如图 8-27、8-28

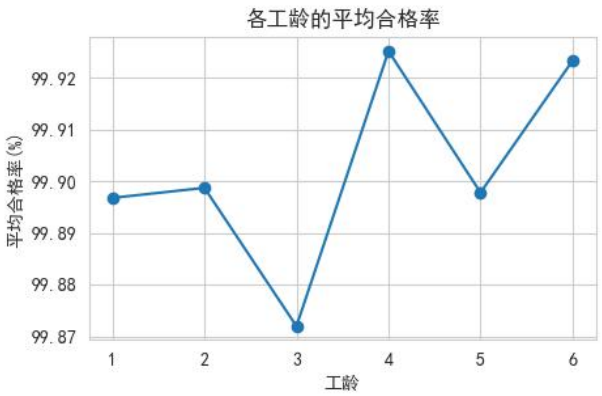

图 8-27 各工龄平均合格率折线图

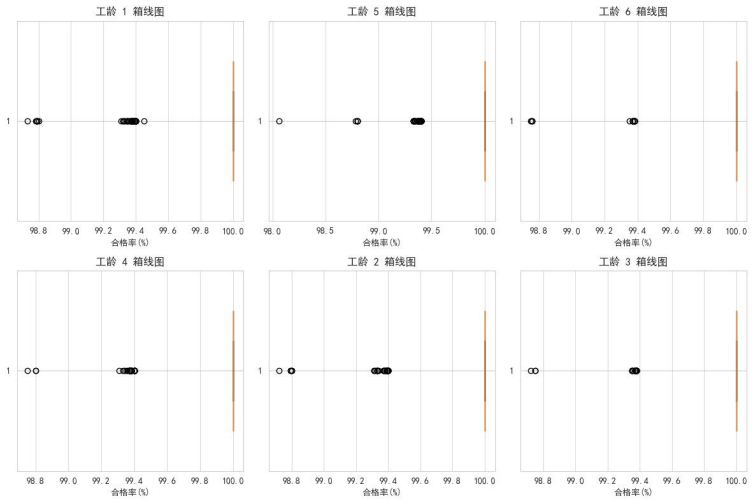

图 8-28 各工龄合格率分布箱线图

对于合格率的分析表明，每位操作人员每天的合格率大体维持在 100%，少数离群点出现在 98.75%-99.5%的范围内。平均合格率最高与最低之间的差异仅为 0.05%左右，且两者之间的相关系数仅为 0.03，说明不同工龄对合格率的影响并不显著。

根据热力图已知故障 3 会对合格率产生显著影响，绘制不同工龄的操作人员遇到故障 3 的次数图 8-23，分析操作人员合格率变化的原因。

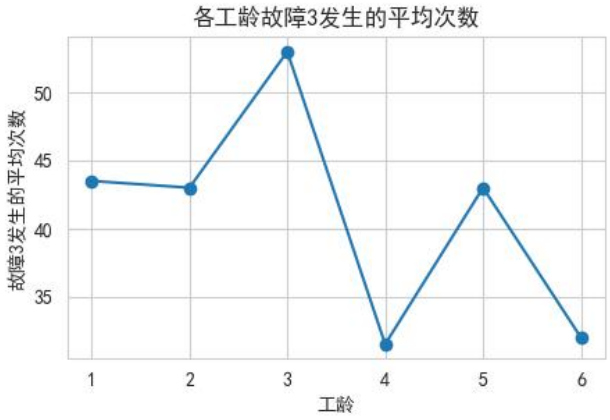

图 8-29 各工龄故障 3 发生的平均次数图

对比 8-27 和 8-29 两张图可知，合格率的变化主要受到故障 3 的影响，而非不同工龄的操作人员之间的差异。

8.7 多元线性回归模型预测

多元线性回归分析的基本任务包括：根据因变量与多个自变量的实际观测值建立因变量对多个自变量的多元线性回归方程；分析各个自变量对因自变量的综合线性影响的显著性；评定各个自变量对因变量影响相对重要性等[11]。

多元线性回归模型的基本形式如下：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon$$

其中，y 是因变量的值，x_1、$x_2 \ldots x_n$ 是自变量的值，β_1、$\beta_2 \ldots \beta_n$ 是模型的参数，ε 表示误差项。

构建各因素与产量、合格率的多元线性回归模型，模型输出的系数（Coefficients）提供了量化关系的衡量标准，如图 8-30、8-31

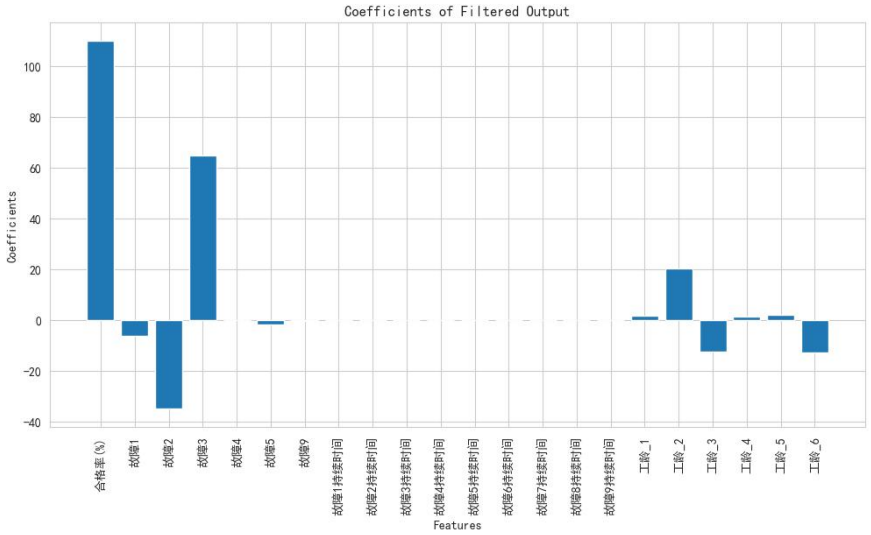

图 8-30 各因素与产量的关系图

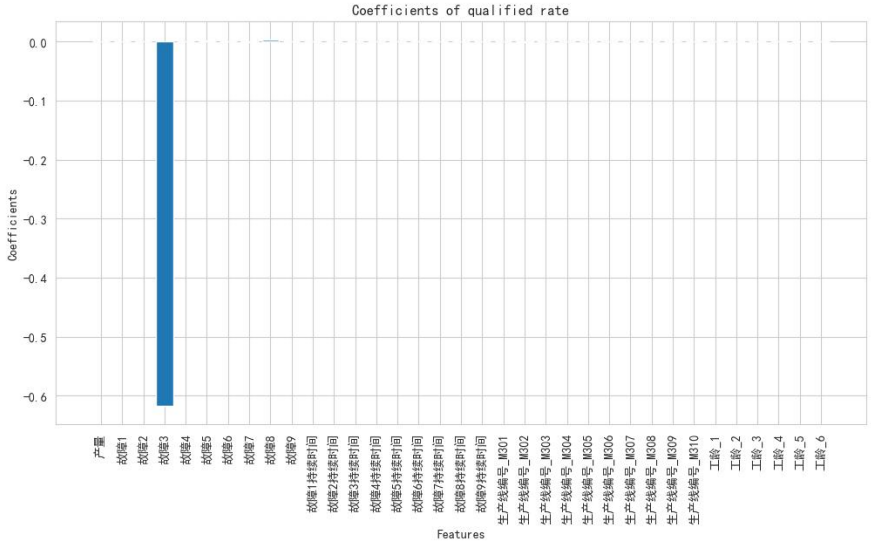

图 8-31 各因素与合格率的关系图

图表清晰地展示了多元线性回归模型中各个特征对产量与合格率预测的贡献程度。从图表数据中可以看出，故障 1 和故障 2 与产量呈现出负相关关系，这意味着当故障发生频率增加时，产品产量会相应减少。此外，工龄为 2 的操作人员所负责的产量相对较高，而工龄为 3 和 6 的操作人员所负责的产量则相对较低，这表明操作人员的经验水平在一定程度上影响了产品的产量。

对于合格率而言，仅有故障 3 对其产生了显著影响，而其他因素几乎没有显著作用。这表明在提升产品合格率方面，应重点关注故障 3 的预防和解决。