

第六届湖南省研究生数学建模竞赛

题 目

黄桃的采摘贮藏分析建模

摘

要：

针对问题 B——黄桃的采摘贮藏分析，研究的目的是建立模型，根据机械指标和人工评分的测评数据，更加科学地评判果实品质，从而应对市民对高品质果品的需求。具体要求是根据某黄桃种植基地 X 品牌黄桃 2018 年、2019 年和 2020 年连续 3 年 2234 条实验数据，构建黄桃品质分析的定量模型，更为精准地为商家推荐优质黄桃。基于此，本研究采用多元线性回归分析、曲线拟合分析、聚类判别分析等方法，利用系统分析、对比分析、数学分析等多种方法，研究构建适合于某黄桃种植基地 X 品牌黄桃的分析模型。

针对问题一，根据 2018 年至 2020 年连续 3 年人工评分和机器指标评分数据作为训练样本集，利用 SPSS 软件，共分析冷柜 8℃共 704 条实验数据，冷库 1℃共 648 条实验数据，以及 882 条人工评分数据，从而确定合适的变量指数，构建定量数学模型对黄桃品质进行分析和判断，运用对比分析的方法，研究模型和专家经验、普遍认知的差异性，并对模型的使用进行了综合性的说明。

针对问题二，在问题 1 所构建模型的基础上，根据 2020 年度样本数据，建立黄桃的品质等级选取的判别模型。采用基于因果关系推断的特征选择方法，通过判别分析（Discriminant Analysis）方法，利用 SPSS 分析样本数据，构建黄桃品质分析的定量模型，较好的改进了两种传统方法的不足，对模型进行优化，并对随机抽取冷柜贮藏方式下 3 个批次采摘的 9 个测试样本确定测评等级。

针对问题三，构建品质随贮藏环境和贮藏时间变化的预测模型，为商家推荐合适的贮藏方式和时间。基本思路是区分冷柜 8℃、冷库 1℃两种贮藏环境进行分别研究，将复杂的多目标线性回归分析转化为单目标回归问题，构建黄桃品质随贮藏环境和贮藏时间变化的预测模型。最后，将目标参数与“普遍认为”进行一致性对比，在分析结果与“普遍认为”具有显著一致性的基础上，给出果农下一年度需要进行有损测试的次数，经测算，在冷库 1℃的贮藏环境下，当间隔天数为 19 天时，黄桃优质率为 0；在冷柜 8℃的贮藏环境下，当间隔天数为 14 天时，黄桃优质率为 0。本研究认为应当以最少的有损测试次数来达到最佳的黄桃品质测试结果，因此冷库 1℃贮藏环境下需要 4 次有损测试，冷柜 8℃贮藏环境下至少需要 2 次有损测试，以保证达到较长的 14 天保鲜期。

综合分析，建议果农于中旬采摘的果实，在 8℃冷柜环境进行贮藏，且贮藏期不多于 6 天；而于月底采摘的果实，于 1℃冷库环境进行贮藏，贮藏时间不高于 14 天。

关键词：多元线性回归 判别分析 曲线拟合 水果采摘贮藏 数学模型

目 录

1. 问题重述.....	3
1.1 问题背景.....	3
1.2 问题介绍.....	3
2. 问题分析.....	4
2.1 问题整体分析.....	4
2.2 问题 1 的分析.....	4
2.3 问题 2 的分析.....	4
2.4 问题 3 的分析.....	4
3. 模型假设.....	5
4. 符号说明.....	5
5. 问题 1 的建模与求解.....	6
5.1 模型的拟合特征选择.....	6
5.2 模型的判别式函数分析.....	7
5.3 品质判别结果分析.....	8
5.3.1 2018 年度品质判别结果.....	8
5.3.2 2019 年度品质判别结果.....	9
5.4 结论.....	9
5.4.1 关于普遍认知.....	9
5.4.2 关于回归模型与“专家经验”的关系.....	9
5.4.3 关于回归模型与“普遍认知”的关系.....	10
5.4.4 综合分析结论.....	11
6. 问题 2 的建模与求解.....	12
6.1 基于判别分类的基本步骤.....	12
6.2 基于因果推断的模型函数.....	12
6.3 基于模型函数的判别分类.....	12
6.4 基于判别分类的品质等级.....	13
7. 问题 3 的建模与求解.....	14
7.1 建立黄桃贮藏预测模型.....	14
7.1.1 冷库 1℃贮藏环境回归模型的判别式函数.....	14
7.1.2 冷柜 8℃贮藏环境回归模型的判别式函数.....	14
7.2 预测模型与“普遍认为”对比分析.....	15
7.3 果农有损测试问题分析.....	15
7.3.1 冷库 1℃贮藏环境果农有损测试问题.....	15
7.3.2 冷柜 8℃贮藏环境果农有损测试问题.....	15
7.3.3 果农有损测试问题综合分析结论.....	16
8. 模型改进与推广.....	17
8.1 模型优点.....	17
8.2 模型缺点.....	17
8.3 模型推广.....	17
参考文献.....	18
附录.....	19

1. 问题重述

1.1 问题背景

黄桃是分布于我国长江流域及北部地区的早熟大果型桃，一般于6月中下旬成熟，果面淡黄至金黄，果肉金黄，果汁丰富。果实成熟后可挂树10天左右，采摘后对比同期其他品种的耐储运能力也较强，是商用价值较高的一款水果产品。由于黄桃比较容易腐烂，需要精准把握其采摘时机和贮藏环境及时间，才能保证在风味最佳时上架各个水果店铺。

黄桃品质的评测与可溶性固形物（TSS）、硬度、色差等诸多因素相关，而对于这些指标的测定，多为有损测试，如利用FTA硬度计、SMS质构仪对果实质地（硬度）进行有损测试，利用手持式折光仪对果实可溶性固形物含量进行有损测试等。

位于S市的某黄桃种植基地通过多种实验手段对基地内的X品牌黄桃进行质量评测，其实验思路主要是在黄桃成熟期的一周左右分三批进行采摘，并设置冷库和冷柜两种储存方式，每隔3~5天取出一部分样本测量各项机械指标与相对应的人工品尝评分。通过这些数据，种植基地希望能够科学刻画黄桃品质与贮藏运输的关系，以帮助农户与商户更好的采摘、存运与售卖，推动S市黄桃产业的数字化转型。

1.2 问题介绍

本研究的目的是建立模型，根据机械指标和人工评分的测评数据，更加科学地评判果实品质，从而应对市民对高品质果品的需求。具体要求是根据某黄桃种植基地X品牌黄桃2018年、2019年和2020年连续3年的实验数据，构建黄桃品质分析的定量模型，更为精准地为商家推荐优质黄桃。根据以上研究目的和要求，主要研究以下3个问题。

1) 分析机械指标、人工评分2种传统方法的一致性水平，构建黄桃品质分析的定量数学模型，并比较模型与机械指标、人工评分之间的差异。

2) 根据2020年度样本数据，建立数学模型判断黄桃的品质等级。随机抽取冷柜贮藏方式下第批次采摘的3个测试样本，根据模型确定9个样本的测评等级（“优质”或“非优质”）。

3) 根据实验数据，综合考虑贮藏环境（冷库或冷柜）和贮藏时长对黄桃品质的影响，构建黄桃品质随贮藏环境和贮藏时间变化的预测模型，判断贮藏环境及贮藏时间对黄桃品质的影响关系，为商家推荐合适的贮藏方式和时间，并根据预测模型判断下一年度果农需要进行有损测试的次数。

2. 问题分析

2.1 问题整体分析

根据题目要求，建立数学模型区分黄桃品质是典型的分类问题。目前的机械指标评分主要以果实硬度大于 1.5 kg/cm^2 ，TSS（可溶性固形物）含量下降率小于 20% 来度量。人工评分主要是专业测试员以人工品尝的方式，对果实色泽、质量、芳香、味道四项进行百分制独立打分，当色泽、质地和味道三个指标均分大于等于 80 分时认为样品“优质”。现已有某黄桃种植基地 X 品牌黄桃 2018 年、2019 年和 2020 年连续 3 年的实验数据，可以作为模型分析的先验数据，可以运用多元线性回归分析、曲线拟合分析、聚类判别分析等数学方法构建黄桃品质分析的定量模型，更为精准地为商家推荐优质黄桃。

样本果实的品质判别是目标参数，果实的采摘日期、测量日期、间隔天数、温度、硬度、TSS 含量、色差（L,a,b,C 和 h）等信息是特征，目标是将特征对目标参数的重要性进行排序，并给出对目标参数最具有显著影响的特征。常用的特征选择方法有很多种，但这些方法各有其适用性和特点，也有各自的局限性。因此仅用其中某一种方法进行选择必然是片面的，需要一种综合的筛选方法。除此之外，要找到对目标参数有“显著影响”的特征，应当还包含与目标参数之间存在“因果关系”，即所选特征应该是目标参数的“原因”，故需要把因果关系的判定标准和强度衡量指标作为排序的重要准则。有效地整合以上方法，就可以得到较为合理的特征选择和重要性排序结果。

2.2 问题 1 的分析

根据题目要求，问题 1 是典型的分类问题。目前主要是两种方法，根据专家经验，优质黄桃的机械指标评分应以果实硬度 $> 1.5 \text{ kg/cm}^2$ ，TSS 含量下降率 $< 20\%$ （即 TSS 初始指标和贮藏时间的指标对比）来衡量。而人工评分时，当色泽、质地和味道这三个指标均分 ≥ 80 分时认为样品“优质”。目前采用双评分机制，普遍认知时机械指标较容易满足，而人工评分要求更高。根据两种传统方法计算出优质率，结果差异较大，指导意义不强。通过判别分析（Discriminant Analysis）方法，利用 SPSS 分析样本数据，构建黄桃品质分析的定量模型，较好的改进了两种传统方法的不足，更加科学的满足了市民对高品质果品的需求。

2.3 问题 2 的分析

问题 2 是特征选择问题，在问题 1 所构建模型的基础上，随机抽出冷柜贮藏方式下每批次采摘的 3 个测试样本，根据 2020 年度的数据建立模型，判断 9 个样品的人工评分等级（“优质”或“非优质”）。本问的核心问题有 2 个：一是选择什么样的模型；二是如何选择特征值。应该把两个问题结合起来考虑，对样品的各项数据指标进行处理，筛选出对目标参数有显著影响的特征值，根据模型判断样品的等级。

2.4 问题 3 的分析

问题 3 是分类预测问题，构建黄桃品质随贮藏环境和贮藏时间变化的预测模型，为商家推荐合适的贮藏方式和时间。基本思路是区分冷柜 8°C 、冷库 1°C 两种贮藏环境进行分别研究，对不同环境（温度）分开讨论，将复杂的多目标线性回归分析转化为单目标线性回归。然后，根据特征值筛选出目标参数。最后，将目标参数与“普遍认为”进行一致性对比，在分析结果与“普遍认为”具有显著一致性的基础上，给出果农需要进行的有损测试次数。

3. 模型假设

本研究的模型假设包括：

- 1.对黄桃果实的机械指标测评结果服从正态分布；
- 2.机械指标测试结果是连续变量，且观测变量之间的相关性在不同样本类中是一样的；
- 3.2018 年样本数据中硬度值为果皮和果肉的硬度平均值，2020 年样本数据中果皮、果肉色差均为果实左、右侧色差实验的平均值；
- 4.2019 年黄桃机械指标数据（色差）样本中，部分样本的实验数据记录错误，本研究假设按照间隔 2 天的抽取频率为标准，实验数据有误的日期依次前推调整。
- 5.题目中叙述冷柜与冷库的温度差别，本研究以冷柜 8℃、冷库 1℃的标准，在出现贮藏环境和温度指标不一致时，以温度指标为研究基准。

4. 符号说明

符号	含义
D_S	样本的采摘日期
D_T	样本的测量日期
D_d	样本测量的间隔天数
T	温度
X_0	果实的硬度
X_1	果实的带皮硬度
X_2	果实的去皮硬度
X_3	果实的 TSS（可溶性固体物）含量
X_L	果实色差（L,a,b,C 和 h）中的 L 属性
X_a	果实色差（L,a,b,C 和 h）中的 a 属性
X_b	果实色差（L,a,b,C 和 h）中的 b 属性
X_C	果实色差（L,a,b,C 和 h）中的 C 属性
X_h	果实色差（L,a,b,C 和 h）中的 h 属性
F	样品果实品质的判别值
$F(t)$	样品果实品质关于时间的函数关系
$F_{(0)}$	果实品质为 0 的判别表达式
$F_{(1)}$	果实品质为 1 的判别表达式
R	自变量与因变量之间的相关性程度

5. 问题 1 的建模与求解

本问的解题思路：种植基地于 2018 年至 2020 年连续 3 年的实验数据，均为黄桃进入成熟期时在一周左右分三批采摘。2018 年共 850 个数据样本，其中机械指标评分 480 个、人工评分 370 个。区分温度为 1℃和 8℃，机械指标评分样本数各为 240 个，人工评分样本数分别为 190 个和 180 个。按照黄桃品质分析的评分标准，筛选机械指标评分和人工评分结果，如表 1 所示。

表 1 2018 年度机械指标和人工评分结果

方法	样本量	小计	优质数	不优质数	优质率
人工评分	小计	370	242	128	65.4%
	8 度	180	124	56	68.9%
	1 度	190	118	72	62.1%
机械指标评分	小计	480	397	83	82.7%
	8 度	240	173	67	72.1%
	1 度	240	224	16	93.3%

考虑到黄桃果实品质评判受到多种因素影响，要进行客观准确的分析和预测，必须首先将无关或关联很小的特征删除。因此，在第一轮变量筛选时，应采用最可靠的筛选准则。在进一步的筛选和排序过程中，由于不同特征选择和重要性计算的方法各自有一定的局限性，故需要综合考虑多种方法进行排序和筛选。完成特征选择后，要对目标参数产生“显著影响”的特征进行排序。对于“显著影响”，应理解为与目标参数有“因果关系”的特征，因此在最后筛选和排序时必须把因果关系的判定和度量标准作为评判准则。

5.1 模型的拟合特征选择

对于问题一，本研究采取基于曲线拟合特征筛选方法，以及基于因果关系的特征选择和排序方法。在题目明确果实硬度大于 1.5 kg/cm^2 ，TSS（可溶性固形物）含量下降率小于 20% 来度量的基础上，利用曲线拟合回归的方法观察样本品质特征值与硬度、TSS、间隔天数等因素的函数表达式，拟合出多种关于硬度、TSS、间隔时间的数学模型。

回归分析是通过提供变量之间的数学表达式来定量描述变量间相关关系的数学过程。回归分析还可以找出哪些自变量对因变量的影响是显著的，哪些是不显著的，比较适合用于此研究。回归分析常用的方法包括线性回归、多项式回归、决策树回归、支持向量机（SVM）回归、K 近邻（KNN）回归、随机森林回归、XGBoost、Lightgbm 和神经网络回归等多种。

①线性回归模型速度快，比较适用于建模有线性关系的数据，命题对异常值比较敏感，可以用于本问，但需要通过加入正则项以减少共线性的影响。

②多项式回归可以拟合非线性的数据，可以考虑根据人工评分和机器指标评分数据作为训练样本集，从而拟合确定合适的变量指数，可以用于本问。

③支持向量机回归是一种较好的小样本学习方法，可以避免维数灾难，具有较强的鲁棒性，可以用于本问。

④KNN 回归方法模型简单，不用太多调节，即可获得较好效果，并且在小数据集上速度比较快，但对于特征多的数据集不适用，可能会造成维数灾难，因此不适用于本问。

⑤随机森林回归能够处理高维度的数据，不需要作出特征选择，模型的泛化能力比较强，训练速度比较快，即使在部分特征遗失的情况下依然能保持较高的准确度，可以适用于本问。

⑥决策树回归相比随机森林回归，模型的泛化性能比较一般，因此本问不作选择。

⑦神经网络回归是一种非参数非线性方法，不需要进行降维处理，模型泛化能力强，可以适用于本问。

⑧XGBoost 和 Lightgbm 方法都借鉴了随机森林的方法，并在其基础上进行了改进，计算复杂度更低，能有效防止过拟合，也可以适用于本问，但相比随机森林需要的参数更多。

综合对以上一些方法的分析，本问利用 IBM SPSS Statistics 23 数学工具进行曲线拟合及多元线性回归模型分别进行回归分析，通过在训练数据集上选取部分验证数据进行模型训练，然后比较不同模型的自变量和因变量的相关性程度，从而作出最优的模型选择，如图 1 所示。

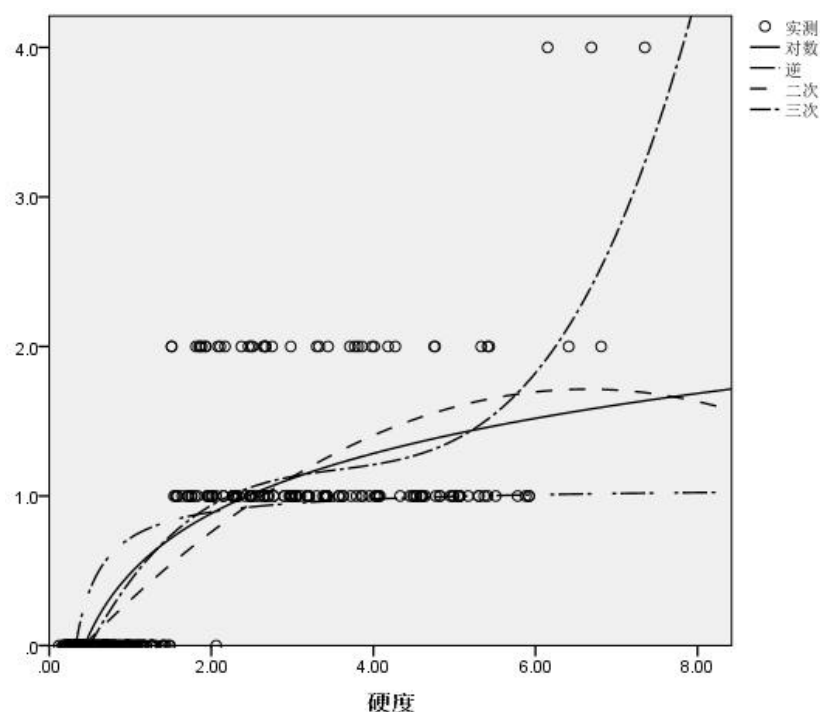


图 1 模型的曲线拟合回归分析

5.2 模型的判别式函数分析

通过计算，得到：

(1) 样本果实品质关于硬度值的对数判别式函数：

$$F = 0.579 \ln x + 0.482, R = 0.763$$

(2) 样本果实品质关于硬度值的二次判别式函数：

$$F = 0.589x - 0.044x^2 - 0.241, R = 0.781$$

(3) 样本果实品质关于硬度值的三次判别式函数：

$$F = 1.337x - 0.355x^2 + 0.33x^3 - 0.583, R = 0.814$$

(4) 样本果实品质关于机械指标的 Fisher 的线性判别式函数：

$$\hat{F}_{(1)} = -0.572D_d - 0.747x_1 + 1.568x_2 + 0.334x_3, R = 0.837$$

$$\hat{F}_{(8)} = 0.119D_d - 0.583x_1 + 1.462x_2 + 0.016x_3, R = 0.882$$

其中 R 表示自变量与因变量之间的相关性程度，其值越大，说明模型越接近四个模型都是在经过超参数优化后，选择其中分析性能最好的参数进行建模分析，然后将分析结果进行比较。通过比较，多元线性回归明显要好于其他方法，因此选择多元线性回归模型作为最终的数据分析模型。经计算，得到样本品质判别数据交叉分析结果，如表 2 所示。

表 2 样本品质判别数据的交叉分析

2018 年度 1℃回归模型的样本品质 * 1 度机械指标的样本品质 交叉表				
区 分		1 度机械指标的样本品质		总计
		.0	1.0	
1 度回归模型	.0	14	2	16
的样本品质	1.0	60	164	224
总 计		74	166	240
2018 年度 8℃回归模型的样本品质 * 8 度机械指标的样本品质 交叉表				
区 分		1 度机械指标的样本品质		总计
		.0	1.0	
1 度回归模型	.0	67	0	67
的样本品质	1.0	42	131	173
总 计		109	131	240
2019 年度 1℃回归模型的样本品质 * 1 度机械指标的样本品质 交叉表				
区 分		1 度机械指标的样本品质		总计
		.0	1.0	
1 度回归模型	.0	40	15	55
的样本品质	1.0	12	107	119
总 计		52	122	174
2019 年度 8℃回归模型的样本品质 * 8 度机械指标的样本品质 交叉表				
区 分		1 度机械指标的样本品质		总计
		.0	1.0	
1 度回归模型	.0	79	9	88
的样本品质	1.0	17	48	65
总 计		96	57	153

5.3 品质判别结果分析

5.3.1 2018 年度品质判别结果

2018 年度共有 480 个实验样本数据。由于 2018 年样本实验数据相互独立，可以考虑采取优质率对比对样本品质结果进行一致性分析。通过对 2018 年 1℃冷库贮藏条件下 240 个样本的机械检测结果进行统计分析，将机械指标评判结果与构建的回归模型分析结果比较发现：机械检测结果与回归模型预测均为不优质的样本数量为 14 个，占总样本数的 5.80%；机械检测结果为不优质、回归模型预测为优质的样本数量为 2 个，占总样本数的 2.24%；机械检测结果为优质，回归模型预测为不优质的数量为 60 个，占总样本数的 25%；机械检测结果与回归模型预测均为优质的数量为 164，占总样本数的 68%。通过对 2018 年 8℃冷柜贮

藏条件下 240 个样本的机械检测结果进行统计分析,将机械指标评判结果与构建的回归模型分析结果比较发现:机械检测结果与回归模型分析均为不优质的样本数量为 53 个,占总样本数的 22.1%;机械检测结果为优质,回归模型分析为不优质的数量为 18 个,占总样本数的 7.5%;机械检测结果与回归模型分析均为优质的数量为 169,占总样本数的 70.4%;无机械检测结果为不优质、回归模型分析为优质的样本。对比机械指标评定和回归模型评定的品质结果,区分 1 度和 8 度两种贮藏环境,对其皮尔逊相关系数进行计算。经计算,2018 年样本在 1 度贮藏环境下相关系数为 0.328,相关程度为弱相关,8℃贮藏环境下相关系数为 0.811,相关程度为极强相关,显著性(双尾)水平均小于 0.01,可以判断回归模型与实测结果的显著性差水平为极显著。

5.3.2 2019 年度品质判别结果

2019 年度共有 332 个实验样本数据。通过对 2019 年 1℃冷库贮藏条件下 174 个样本的机械检测结果进行统计分析,将机械指标评判结果与构建的回归模型分析结果比较发现:机械检测结果与回归模型预测均为不优质的样本数量为 43 个,占总样本数的 24.7%;机械检测结果为不优质、回归模型预测为优质的样本数量为 9 个,占总样本数的 5.2%;机械检测结果为优质,回归模型预测为不优质的数量为 32 个,占总样本数的 18.4%;机械检测结果与回归模型预测均为优质的数量为 90,占总样本数的 51.7%。通过对 2019 年 8℃冷柜贮藏条件下 158 个样本的机械检测结果进行统计分析,将机械指标评判结果与构建的回归模型分析结果比较发现:机械检测结果与回归模型预测均为不优质的样本数量为 79 个,占总样本数的 51.6%;机械检测结果为不优质、回归模型预测为优质的样本数量为 17 个,占总样本数的 11%;机械检测结果为优质,回归模型预测为不优质的数量为 9 个,占总样本数的 5.9%;机械检测结果与回归模型预测均为优质的数量为 40,占总样本数的 26.1%。

5.4 结论

5.4.1 关于普遍认知

1.2018 年人工评判 1℃贮藏环境中样本优质数量为 118,占总人工评判量比重为 62.1%,相较于机械评判 93.3%的优质率,说明机械评判标准相对人工评分更加容易满足果实的优质标准,与“普遍认知”相符合。

2.2018 年人工评判 8℃贮藏环境中样本优质数量为 124,占总人工评判量比重为 68.9%,相较于机械评判 72.1%的优质率,说明机械评判标准相对人工评分更加容易满足果实的优质标准,与“普遍认知”相符合。

3.2019 年人工评判 1℃贮藏环境中样本优质数量为 177,占总人工评判量比重为 65.1%,相较于机械评判 70.1%的优质率,说明机械评判标准相对人工评分更加容易满足果实的优质标准,与“普遍认知”相符合。

4.2019 年人工评判 8℃贮藏环境中样本优质数量为 119,占总人工评判量比重为 49.6%,相较于机械评判 36.1%的优质率,说明机械评判标准相对人工评分并不是更加容易满足果实的优质标准,因此,该条件下的结果与“普遍认知”并不相符合,分析原因如下。一是在 8℃的环境下,黄桃应代谢较好,硬度及 TSS 两项指标下降较快,在检测后期大部分已不满足这两项机械指标的优质标准,因此总体优质率相对较低。

5.4.2 关于回归模型与“专家经验”的关系

“专家经验”主要以果实硬度及 TSS 含量下降率两个指标为判断依据,说明机械评判结果即为“专家经验”评判结果。

1.通过对比 2018 年度 1℃贮藏环境下样本的优质率,人工评判的为 62.1%,机械评判的为 93.3%,回归模型评判为 69.2%。机械检测结果与回归模型预测结果一致(即同时评判为优质或不优质)数量占比为 75%。另外,机械检测结果不优质为 16 个,回归模型检测结果不优质为 60 个,交叉不优质样本有 14 个,说明回归模型对机械评判出的不优质样本具有比较高的一致性检测水平。即所建立的回归模型与“专家经验”具有一定的一致性,同时对黄桃品质评判标准比“专家经验”更高。

2.通过对比 2018 年度 8℃贮藏环境下样本的优质率,人工评判的为 68.9%,机械评判的为 72.1%,回归模型评判为 70.4%。机械检测结果与回归模型预测结果一致(即同时评判为优质或不优质)数量占比为 91.3%。另外,机械检测结果不优质为 53 个,回归模型检测结果不优质为 71 个,交叉不优质样本有 53 个,说明回归模型对机械评判出的不优质样本具有比较高的一致性检测水平。即所建立的回归模型与“专家经验”具有一定的一致性,同时对黄桃品质评判标准比“专家经验”更高。

3.通过对比 2019 年度 1℃贮藏环境下样本的优质率,人工评判的为 65.1%,机械评判的为 70.1%,回归模型评判为 68.4%。机械检测结果与回归模型预测结果一致(即同时评判为优质或不优质)数量占比为 84.5%。另外,机械检测结果不优质为 52 个,回归模型检测结果不优质为 55 个,交叉不优质样本有 40 个,说明回归模型对机械评判出的不优质样本具有比较高的一致性检测水平。即所建立的回归模型与“专家经验”具有一定的一致性,同时对黄桃品质评判标准比“专家经验”更高。

4.通过对比 2019 年度 8℃贮藏环境下样本的优质率,人工评判的为 49.6%,机械评判的为 36.1%,回归模型评判为 42.5%。机械检测结果与回归模型预测结果一致(即同时评判为优质或不优质)数量占比为 83.0%。另外,机械检测结果不优质为 96 个,回归模型检测结果不优质为 88 个,交叉不优质样本有 79 个,说明回归模型对机械评判出的不优质样本具有比较高的一致性检测水平。即所建立的回归模型与“专家经验”具有一定的一致性,同时对黄桃品质评判标准比“专家经验”受限于指标参考维度单一的问题具有更好的评判价值。

5.4.3 关于回归模型与“普遍认知”的关系

1.通过对比 2018 年度 1℃贮藏环境下样本的优质率,人工评判的为 62.1%,机械评判的为 93.3%,回归模型评判为 69.2%。回归模型优质率与人工评判较为接近,且远小于机械指标评判的优质率,说明专家经验设定的硬度和 TSS 含量标准评判“优质”果实还需改进,而回归模型在利用硬度和 TSS 含量等指标确定的品质分析中,结果与人工评判优质率结果接近,表明回归模型可以更加科学地评判果实品质,更好地满足市民对高品质果品的需要。

2.通过对比 2018 年度 8℃贮藏环境下样本的优质率,人工评判的为 68.9%,机械评判的为 72.1%,回归模型评判为 70.4%。回归模型优质率与人工评判及机械检测结果均较为接近,且小于机械指标评判的优质率,说明专家经验设定的硬度和 TSS 含量标准评判“优质”果实具有一定的参考价值,而回归模型在利用硬度和 TSS 含量等指标确定的品质分析中,结果与人工评判优质率结果更为接近,表明回归模型可更加科学地评判果实品质,更好地满足市民对高品质果品的需要。

3.通过对比 2019 年度 1℃贮藏环境下样本的优质率,人工评判的为 65.1%,机械评判的为 70.1%,回归模型评判为 68.4%。回归模型优质率与人工评判、机械评判均较为接近,且小于机械指标评判的优质率,说明专家经验设定的硬度和

TSS 含量标准评判“优质”果实具有较好的参考价值，而回归模型在利用硬度和 TSS 含量等指标确定的品质分析中，结果与人工评判优质率结果接近，表明回归模型可以更加科学地评判果实品质，更好地满足市民对高品质果品的需要。

4.通过对比 2019 年度 8℃贮藏环境下样本的优质率，人工评判的为 49.6%，机械评判的为 36.1%，回归模型评判的为 42.5%。回归模型优质率与人工评判、机械评判均较为接近，由于在 8℃环境下机械评判与人工评分差距较大，与普遍认知不符，而回归模型在利用硬度和 TSS 含量等指标确定的品质分析中，结果与人工评判优质率结果接近，表明回归模型可以更加科学地评判果实品质，更好地满足市民对高品质果品的需要。

5.4.4 综合分析结论

综合分析 2018、2019 年度 1℃、8℃贮藏环境下机械、人工、回归模型评判数据，可得以下结论：

1.机械指标、人工评分之间的关系与两类方法的评判一致性。通过对在 2018 年 1℃、8℃和 2019 年 1℃的样本机械指标和人工评分结果进行分析，两种评判方式均能对黄桃品质做出较为合理和一致的评判，且人工评判的标准比机械指标评判标准更为严格，而在 2019 年 8℃环境中，机械指标评判优质率比人工评判更低，我们认为主要是由于相对于在 8 摄氏度环境中检测间隔时间过长，黄桃机械指标下降较快，所以导致机械评判优质率低于人工评判。

2.回归模型与“专家经验”的差异。通过对回归模型分析结果与“专家经验”方法中采用机械评判的结果进行比较。可以发现回归模型对机械评判出的不优质样本具有较高的一致性检测水平。即所建立的回归模型与“专家经验”具有一定的一致性，同时对黄桃品质评判标准比“专家经验”更高。

3.回归模型与“普遍认知”的差异。通过对回归模型分析结果与“普遍认知”中机械指标和人工评分的结果进行比较分析。可以发现回归模型的分析结果优质率低于机械评判结果，更接近于人工评分优质率，一定程度上能够修正机械指标满足标准较低的问题。

6. 问题 2 的建模与求解

针对问题 2，常用的特征选择方法有很多种，但这些方法各有其适用性和特点，也有各自的局限性。本研究考虑可以在问题 1 所构建曲线拟合和多元线性回归模型的基础上，按照以下步骤进行样本的判别分类（Discriminant Analysis）。

6.1 基于判别分类的基本步骤

第 1 步 计算特征值。计算需要用到的一些反映样本的特征值，比如均值、协方差矩阵等。

第 2 步 建立判别函数。 $F = a_1x_1 + a_2x_2 + a_3x_3 + \cdots + a_nx_n$

第 3 步 确定判别准则。本研究以 Fisher 准则进行判别。

第 4 步 检验判别效果。验证判别函数用来进行判别时的准确度。

第 5 步 分类。

6.2 基于因果推断的模型函数

分类判别的基本原理是对确定性样本做出分类标记后，利用已知分类样本的特征信息作为未分类样本的参考，从而对未分类样本进行函数拟合的分类。对问题 2 冷柜（8℃）贮藏方式下的 306 个数据样本的采摘日期、测量日期、间隔天数、温度、硬度、TSS 含量、色差（L,a,b,C 和 h）等信息，利用问题一中的多元线性回归模型综合判断，得到初始判断结果。通过计算，建立在 2020 年 8℃贮藏环境下的样本品质的判别函数为：

$$F = 0.765D_S - 1.255D_T + 0.096x_1 + 0.188x_2 - 0.106x_3 - 0.310x_L + 0.432x_a + 0.440x_b - 0.240x_C + 0.770x_h + 0.130x_L + 0.670x_a + 0.167x_b + 0.165x_h$$

其中，样本品质特征值为 3.310，表明模型的判别函数具有较强的区别力；威尔克 Lambda 值为 0.232，显著性水平为极显著，表明模型的判别函数比较稳定，差异性不大；典则判别函数的典型相关性值为 0.876，说明模型用来进行判别果实品质分类的准确度可以达到 87.6%。

6.3 基于模型函数的判别分类

抽取需要判断的 9 个随机样品，将测试的各变量值代入品质判别函数，得出差别得分，从而确定 9 个样品的品质情况，并计算判别结果属于各类别的概率，从而精确判断出随机样品的具体类别，如表 3 所示。

表 3 随机样品的判别分类

年份	温度	采摘日期	测试日期	间隔天数	样本	品质	不优质的概率	优质的概率
2020	8℃	6 月 14 日	6 月 14 日	0	样本 A	1	0.00001	0.99999
2020	8℃	6 月 14 日	6 月 22 日	8	样本 B	0	0.85311	0.14689
2020	8℃	6 月 14 日	6 月 30 日	16	样本 C	0	0.99998	0.00002
2020	8℃	6 月 18 日	6 月 18 日	0	样本 D	1	0.00003	0.99997
2020	8℃	6 月 18 日	6 月 28 日	10	样本 E	0	0.97957	0.02043
2020	8℃	6 月 18 日	7 月 6 日	18	样本 F	0	1	0
2020	8℃	6 月 22 日	6 月 22 日	0	样本 G	1	0.00241	0.99759
2020	8℃	6 月 22 日	6 月 30 日	8	样本 H	0	0.75328	0.24672
2020	8℃	6 月 22 日	7 月 8 日	16	样本 I	0	0.99997	0.00003

6.4 基于判别分类的品质等级

按照题目要求，根据 2020 年度的数据建立模型，判断这 9 个样本的评分等级（“优质”或“非优质”）的结果，如表所示。

样本	品质	样本	品质	样本	品质
样本 A	优质	样本 D	优质	样本 G	优质
样本 B	不优质	样本 E	不优质	样本 H	不优质
样本 C	不优质	样本 F	不优质	样本 I	不优质

7. 问题 3 的建模与求解

针对问题 3，其建模求解与问题 1、问题 2 的思路整体相似，但又略有不同。第三问的根本目的在于根据模型预测不同环境和贮藏时间对果实品质的影响。因此我们在特征选择时，应尽量以所选模型的预测性能为依据。同时特征的数量不受限制，因此可以通过优化方法，从所有备选特征中选择能使模型最优化的特征，作为最终输入。而特征作为模型输入时，部分特征中可能存在冗余特征，导致模型的性能下降。为了从中选择最合适的特征，本研究重点对测量日期、间隔天数、硬度及 TSS 等特征进行筛选求解。

7.1 建立黄桃贮藏预测模型

考虑到问题 3 是分类预测问题，构建黄桃品质随贮藏环境和贮藏时间变化的预测模型，为商家推荐合适的贮藏方式和时间，基本思路是区分冷柜 8℃、冷库 1℃两种贮藏环境进行分别研究，对不同环境（温度）分开讨论，将复杂的多目标线性回归分析转化为单目标回归问题。

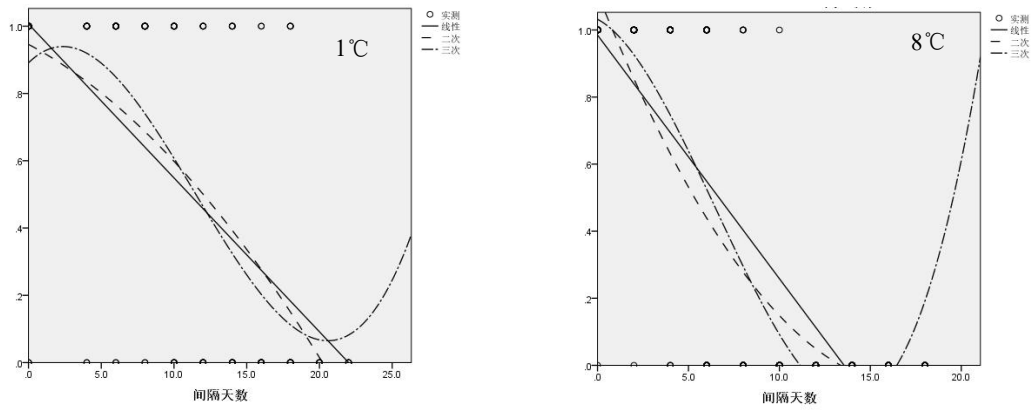


图 2 贮藏环境回归模型优质率随贮藏时间的判别式函数

通过 IBM SPSS Statistics 23 回归分析，曲线拟合出果实品质与间隔天数的线性、二次和三次 3 种回归模型，如图 2 所示。其中：

7.1.1 冷库 1℃贮藏环境回归模型的判别式函数

(1) 样本果实品质关于间隔天数的线性判别式函数：

$$F = -0.073x + 0.985, R = 0.832$$

(2) 样本果实品质关于间隔天数的二次判别式函数：

$$F = -0.135x + 0.004x^2 + 1.109, R = 0.859$$

(3) 样本果实品质关于间隔天数的三次判别式函数：

$$F = -0.029x - 0.013x^2 + 0.001x^3 + 1.032, R = 0.877$$

7.1.2 冷柜 8℃贮藏环境回归模型的判别式函数

(1) 样本果实品质关于间隔天数的线性判别式函数：

$$F = -0.051x + 1.143, R = 0.805$$

(2) 样本果实品质关于间隔天数的二次判别式函数：

$$F = -0.039x - 0.001x^2 + 1.103, R = 0.807$$

(3) 样本果实品质关于间隔天数的三次判别式函数：

$$F = 0.063x - 0.012x^2 + 0.001x^3 + 0.990, R = 0.845$$

其中 R 表示自变量与因变量之间的相关性程度，其值越大，说明模型越接近间隔时间对品质的影响程度。

7.2 预测模型与“普遍认为”对比分析

目前普遍认为：在冷库贮藏情况下，低温会影响黄桃代谢，风味有所下降，但保质期较长；在冷柜贮藏情况下，黄桃代谢情况较好，但保鲜期较短。

结合问题 1 和问题 2 的回归分析模型，曲线拟合建立的黄桃贮藏预测模型判断，从左图品质最高值小于右图品质最高值可分析得出诸多结果，比如，在冷库 1℃贮藏情况下，低温会影响黄桃的代谢，黄桃的品质值有了明显下降，但保质期更长，直到 20 天时品质降至最低值；而在冷柜 8℃贮藏情况下，黄桃的代谢情况较好，但保质期较短，仅 10 天时品质降至最低值。

因此，综合分析得到结论：预测模型的数据结果与“普遍认为”是一致的。

7.3 果农有损测试问题分析

通过对模型数据进行聚类分析，得出 2019 年度贮藏环境回归模型优质率随贮藏时间的变化情况，如表 4 所示。根据预测模型得到的结果，该表格反映了黄桃品质随贮藏环境和贮藏时间变化的情况。该聚类结果表明，本研究建立的回归分析模型具有较高的稳定性和一致性，因此，可以通过回归模型对果农进行有损测试的次数作出判断。

表 4 2019 年度贮藏环境回归模型优质率随贮藏时间的变化情况

	间隔天 数	不优质 数	优质 数	总 数	优质率	间隔天 数	不优质 数	优质 数	总 数	优质 率
1℃ 冷库	0	0	28	28	100.0%	14	7	8	15	53.3%
	4	0	15	15	100.0%	16	13	2	15	13.3%
	6	0	15	15	100.0%	18	13	2	15	13.3%
	8	0	15	15	100.0%	20	15	0	15	0.0%
	10	6	9	15	60.0%	22	5	0	5	0.0%
	12	10	5	15	33.3%	24	6	0	6	0.0%
8℃ 冷柜	0	0	28	28	100.0%	12	15	0	15	0.0%
	2	1	14	15	93.3%	14	15	0	15	0.0%
	4	0	15	15	100.0%	16	15	0	15	0.0%
	6	8	7	15	46.7%	18	15	0	5	0.0%
	8	14	1	15	6.7%	12	15	0	15	0.0%

7.3.1 冷库 1℃贮藏环境果农有损测试问题

通过预测模型求解，当间隔天数为 19 天时，黄桃优质率为 0。通过分析，在贮藏环境为冷库（1℃）的条件下，间隔天数小于 8 天时，该黄桃的优质率始终处于较高水平（可达到 90%以上），从第 10 天开始，黄桃的优质率出现明显降低的情况。当间隔时间大于 16 天时，黄桃优质率接近低于 10%，本研究认为此时已基本失去市场价值。因此，最好需要 4 次有损测试，具体为间隔天数为第 8 天、第 10 天、第 12 天、第 14 天。

7.3.2 冷柜 8℃贮藏环境果农有损测试问题

通过预测模型求解，当间隔天数为 14 天时，黄桃优质率为 0。通过分析，在贮藏环境为冷柜（8℃）的条件下，间隔天数小于 4 天时，该黄桃的优质率始终处于较高水平（可达到 90%以上），从第 6 天开始，黄桃的优质率出现明显降

低的情况。当间隔时间大于 8 天时，黄桃优质率接近低于 10%，本研究认为此时已基本失去市场价值。因此，**最好需要 2 次有损测试**，具体为间隔天数为第 4 天、第 6 天。

7.3.3 果农有损测试问题综合分析结论

考虑该黄桃一般于 6 月中下旬成熟，果实成熟后可挂树 10 天左右，成熟期之后，黄桃的品质便快速下降，而做有损测试对黄桃的品质影响较大，本研究认为应当以最少的有损测试次数来达到最佳的黄桃品质测试结果，因此冷柜 8℃贮藏环境下至少需要 2 次有损测试，以保证保鲜期可以达到较长的 6 天，而冷库 1℃贮藏环境下需要 4 次有损测试，以保证保质期可以达到较长的 14 天。

综合上述分析，建议果农于中旬采摘的果实，在 8℃冷柜环境进行贮藏，且贮藏期不多于 6 天；而于月底采摘的果实，建议于 1℃冷库环境进行贮藏，贮藏时间不超过 14 天。

8. 模型改进与推广

8.1 模型优点

1.在特征选择的基础上，并不单独采用某种方法进行筛选，而是通过多种方法进行综合分析，进而确定关键指标，筛选出优质黄桃，客观性和准确性都更强。

2.在特征选择时，并不只用“相关”或“关联”作为特征对目标参数是否有显著影响的评判标准，而是对每个有可能影响黄桃品质的变量均进行了回归分析，根据结果的显著性确定关键指标，使用该方法建立的模型更具有可解释性。

8.2 模型缺点

1.在问题 1 中构建模型对黄桃品质进行评判时，仅考虑了硬度、TSS 等对目标参数的影响，各自变量之间的自相关和偏相关考虑还不够全面。

2.不管什么预测模型，都会因为初始值、搜索等过程的随机性，导致其结果存在一定的波动，将预测模型与实际应用相结合，可能会导致预测误差被放大和实际结果的不稳定性。

8.3 模型推广

1.线性回归分析方法已经在统计学、经济学等领域都有了广泛的运用，其在农业、医疗、工业等领域的应用值得在未来进行深入研究。

2.对于问题 2、问题 3 的模型，分别使用因果推断和预测模型，对影响指标进行筛选，确定关键指标，以优化模型性能，该思路在高维数据建模预测中具有一定的参考和借鉴价值。

参考文献

- [1]闫蕊,赵守梅,张馨心,吕雨梅.决策树与 Logistic 回归模型在老年人社区养老意愿影响因素分析中的应用研究[J/OL].中国全科医学:1-7[2021-11-21].<http://kns.cnki.net/kcms/detail/13.1222.R.20211118.2028.008.html>.
- [2]邓维斌,唐兴艳,胡大权,周玉敏.SPSS19 统计分析实用教程[M].电子工业出版社,2012.
- [3]林彬.多元线性回归分析及其应用[J].中国科技信息,2010(09):60-61.
- [4]孙振宇.多元回归分析与 Logistic 回归分析的应用研究[D].南京信息工程大学,2008.
- [5]尹建杰. Logistic 回归模型分析综述及应用研究[D].黑龙江大学,2011.
- [6]Shiran Gholamreza; Imaninasab Reza; Khayamim Razieh. Crash Severity Analysis of Highways Based on Multinomial Logistic Regression Model, Decision Tree Techniques, and Artificial Neural Network: A Modeling Comparison[J].Sustainability, 2021, 13(10): 5670-5670.
- [7]Golhar A.R,Choudhari N.K,Patil A.K.Prediction of aluminium content in a metal using SPSS based linear regression analysis.[J].Journal of Physics:Conference Series,2021,1913(1)
- [8]Hu JiaYu et al.When to consider logistic LASSO regression in multivariate analysis?[J]. European journal of surgical oncology :the journal of the European Society of Surgical Oncology and the British Association of Surgical Oncology, 2021, 47(8):2206-.
- [9]Yuan Lu.Empirical Analysis on the Relationship Between Tertiary Industry Structure and Economic Growth Based on Multiple Regression — A Case Study of Henan Province[J].E3S Web of Conferences,2021,235:02019-.
- [10]刘东升,谢朝武.基于 SPSS 的胶结充填配比及多元线性回归分析试验研究[J].黄金,2019,40(09):32-36.
- [11]颜鲁林.利用 SPSS 对大学生学习注意力集中程度进行多元线性回归分析[D].兰州大学,2012.
- [12]祁辉,李玉武,刘咸德,董亮,封跃鹏,吴忠祥.用多元线性回归分析法定量判别 PCBs 污染物类型[J].环境科学研究,2005(04):91-95.
- [13]李洪成,张茂军,马广斌. SPSS 数据分析实用教程[M].人民邮电出版社,2017.
- [14]汤后林.SPSS 轻松学: 实战案例[M].人民卫生电子音像出版社有限公司,2019.
- [15]尹剑,陆程敏,杨贵军.判别分析与 Logistic 回归组合分类[J].数理统计与管理,2014,33(02):256-265.

附录

附件 1 问题 1 机械评分和回归模型评分结果相关性和一致性检验

2018 年 1 度机械指标和回归指标品质相关性			
		2018 年 1 度机械指标 评定品质结果	2018 年 1 度回归模型 评定品质结果
2018 年 1 度机械指标 评定品质结果	皮尔逊相关性	1	.328**
	显著性（双尾）		.000
	个案数	240	240
2018 年 1 度回归模型 评定品质结果	皮尔逊相关性	.328**	1
	显著性（双尾）	.000	
	个案数	240	240
**. 在 0.01 级别（双尾），相关性显著。			

2018 年 8 度机械指标和回归指标品质相关性			
		2018 年 8 度机械指标 评定品质结果	2018 年 8 度回归模型 评定品质结果
2018 年 8 度机械指标 评定品质结果	皮尔逊相关性	1	.811**
	显著性（双尾）		.000
	个案数	240	240
2018 年 8 度回归模型 评定品质结果	皮尔逊相关性	.811**	1
	显著性（双尾）	.000	
	个案数	240	240

2019 年 1 度机械指标和回归指标品质相关性			
		1 度回归模型的样本 品质	1 度机械指标的样本 品质
1 度回归模型的样本品质	皮尔逊相关性	1	.522**
	显著性（双尾）		.000
	个案数	174	174
1 度机械指标的样本品质	皮尔逊相关性	.522**	1
	显著性（双尾）	.000	
	个案数	174	174

2019 年 8 度机械指标和回归指标品质相关性				
		8 度回归模型的样本品质		8 度机械指标的样本品质
8 度回归模型的样本品质	皮尔逊相关性		1	.650**
	显著性（双尾）			.000
	个案数		153	153
8 度机械指标的样本品质	皮尔逊相关性		.650**	1
	显著性（双尾）		.000	
	个案数		153	158
2019 年 1 度回归模型的样本品质 * 1 度机械指标的样本品质 交叉表				
计数				
		1 度机械指标的样本品质		总计
		.0	1.0	
1 度回归模型的样本品质	.0	43	32	75
	1.0	9	90	99
总计		52	122	174

2019 年 8 度回归模型的样本品质 * 8 度机械指标的样本品质 交叉表				
计数				
		8 度机械指标的样本品质		总计
		.0	1.0	
8 度回归模型的样本品质	.0	79	9	88
	1.0	17	48	65
总计		96	57	153

附件 2 问题 2 判别分类结果输出

具体计算过程见附件：《附件—第 2 问样本 A-I 品质分类计算过程》

最终计算结果见附件：《附件—第 2 问样本 A-I 品质分类结果(0-1) - 含概率统计》