**About p value calculation**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **自变量类型** | **因变量类型** | **检验目的/情景** | **推荐检验方法** | **R 中常用函数** |
| 分类（2组） | 分类（2组） | 检验两个分类变量是否有关联（如性别与疾病） | 卡方检验（Chi-square test） | chisq.test() |
| 分类（2组） | 分类（2组，小样本） | 分类变量频数太小（如某一组 < 5） | Fisher 精确检验 | fisher.test() |
| 分类（2组） | 连续变量 | 比较两组的均值差异（如男女 BMI 是否不同） | 独立样本 t 检验 | t.test() |
| 分类（2组，非正态） | 连续变量 | 连续变量不服从正态分布 | Mann–Whitney U 检验 | wilcox.test() |
| 分类（≥3组） | 连续变量 | 比较多组间的均值（如不同民族的 BMI 差异） | 单因素方差分析（One-way ANOVA） | aov() 或 anova() |
| 连续变量 | 连续变量 | 检验两个连续变量之间的相关性 | Pearson 相关系数（正态分布） | cor.test(method="pearson") |
| 连续变量 | 连续变量 | 检验等级相关性或非正态分布 | Spearman 秩相关系数 | cor.test(method="spearman") |
| 分类/连续变量 | 分类变量（0/1） | 建立分类预测模型（如营养是否不良） | 逻辑回归（Logistic Regression） | glm(family = binomial) |
| 分类/连续变量 | 连续变量 | 建立连续型结果的预测模型（如营养评分） | 线性回归（Linear Regression） | lm() |
| 多个分类/连续变量 | 分类变量 | 多个因素联合对分类结果影响 | 多因素 Logistic 回归 | glm(family = binomial) |
| 多个分类/连续变量 | 连续变量 | 多个因素联合对连续结果影响 | 多元线性回归 | lm() |

* 连续变量需判断是否服从正态分布（可用 shapiro.test()），再决定使用 t 检验或非参数检验。
* 分类变量若为多分类，可以考虑多项 Logistic 回归（multinom()）。
* 多组比较如显著，还需做事后检验（如 TukeyHSD）。

比较内容 检验方法

两组均值比较（正态分布） 独立样本 t 检验

三组及以上均值比较（正态） 单因素方差分析 ANOVA

非正态连续变量的两组比较 Mann-Whitney U 检验（wilcox.test()）

分类变量之间的独立性 卡方检验 / Fisher检验

连续变量间线性相关性 Pearson 相关

连续变量间非线性相关性 Spearman 相关

**About linear regression**

1. 残差 vs 拟合值图（Residuals vs Fitted plot）

* 目的：用于检查模型的线性假设和异方差性（即残差是否随拟合值的变化而变化）。
* 如何解读：
  + 没有模式：理想情况下，残差应该随机分布，且没有系统性模式。如果图上没有明显的弯曲或趋势线，说明线性假设成立，模型适合数据。
  + 有模式：如果残差显示出某种模式（如弯曲或趋势），这通常意味着模型存在未捕获的非线性关系，可能需要转换自变量（例如，使用多项式回归或对数转换）。
  + 异方差性：如果残差在某些拟合值范围内的变动比其他范围大，说明数据可能存在异方差性，即误差的方差不恒定。

理想情况：残差应均匀分布，且没有明显模式。

2. QQ 图（Quantile-Quantile plot）

* 目的：用于检查残差是否符合正态分布。QQ 图比较数据的实际分位数与理论正态分布的分位数。
* 如何解读：
  + 理想情况：如果数据点几乎沿着对角线分布，表示残差接近正态分布。
  + 偏离对角线：如果点偏离对角线，尤其是两端，表明残差可能不是正态分布，这可能影响模型的准确性。常见问题包括：
    - 左右尾部的偏离：可能表明数据中存在重尾分布，或存在极端值。
    - 中间的偏离：可能表示数据中的非正态性。

理想情况：点应该大致沿对角线分布。

3. 尺度位置图（Scale-Location plot，又称Homoscedasticity plot）

* 目的：用于检查异方差性，即误差的方差是否恒定。通过绘制标准化残差的平方根对拟合值的关系来检查。
* 如何解读：
  + 理想情况：标准化残差的平方根应该均匀分布，显示出没有系统性的模式，且随着拟合值的增大或减小，残差的变动幅度保持一致（即没有异方差性）。
  + 不理想情况：如果图中显示出某种趋势（例如，随着拟合值增大或减小，残差的变动幅度增大或减小），则说明模型可能存在异方差性。此时可能需要对数据进行转换（例如，使用加权最小二乘法或对数转换）来解决。

理想情况：数据点应该均匀分布，且没有系统性趋势。

4. 杠杆值图（Leverage plot）

* 目的：用于检查数据点在模型拟合中的影响力，尤其是看哪些数据点可能对模型的拟合结果有过大的影响。
* 如何解读：
  + 杠杆值衡量数据点在预测变量空间中位置的“远离”程度。杠杆值较大的数据点（即距离其他点较远的点）可能对模型拟合有较大影响。
  + 高杠杆值点：如果一个点的杠杆值很高，可能是一个异常值或杠杆点，影响模型的拟合。需要警惕这些点，因为它们可能不代表一般数据，可能会影响模型的稳定性和准确性。
  + Cook's 距离：通常与杠杆值图一起使用，计算每个数据点的影响力。高杠杆值和大Cook’s 距离的点可能需要被进一步调查和处理。

理想情况：大多数数据点应集中在杠杆值较低的区域，只有少数点有较高的杠杆值。极端的杠杆点需要进行额外检查，看看是否应该从数据中删除或进行特殊处理。

1. Estimate (回归系数)

* 含义：回归系数表示自变量对因变量的影响大小。在回归方程中，Estimate 是自变量的系数，表示自变量每增加一个单位，因变量预计变化的数量。
* 例如：如果 Estimate 为 2，意味着自变量每增加 1 单位，因变量的值将增加 2 单位。

2. Std.Error (标准误差)

* 含义：标准误差是回归系数估计值的标准差，表示该估计值的精确程度。标准误差越小，回归系数估计值的精度越高。
* 例如：如果 Std.Error 为 0.5，意味着回归系数估计的精确度是 0.5。

3. t.value (t 值)

* 含义：t 值是用来检验回归系数是否显著不为零的统计量。计算公式为：t.value = Estimate / Std.Error。它表示回归系数与标准误的比值，用于判断回归系数的显著性。
* 例如：如果 t 值为 4，且 p 值小于 0.05，说明该回归系数在统计学上是显著的。
* 判断依据：通常，t 值的绝对值越大，回归系数的显著性越强。

4. P.value (P 值)

* 含义：P 值用于检验回归系数是否显著不为零。它表示观察到的数据在原假设（回归系数为 0）下发生的概率。如果 P 值小于某个显著性水平（通常是 0.05），我们拒绝原假设，认为回归系数是显著的。
* 例如：如果 P 值为 0.03，意味着在原假设下，观察到回归系数为该值的概率是 3%，因此可以认为该系数在统计上显著。
* 显著性水平：
  + p < 0.001：非常显著（\*\*\*）
  + 0.001 ≤ p < 0.01：显著（\*\*）
  + 0.01 ≤ p < 0.05：略显著（\*）
  + p ≥ 0.05：不显著（无星号）

5. CI (置信区间)

* 含义：置信区间是回归系数估计值的区间范围，通常使用 95% 置信区间，意味着我们有 95% 的把握认为回归系数真实值落在这个区间内。置信区间越窄，回归系数的估计越精确。
* 例如：如果 CI 为 [1.5, 3.5]，意味着回归系数的真实值在 95% 的置信水平下位于这个区间内。
* 判断依据：如果置信区间不包含零，则回归系数显著；如果包含零，则说明回归系数可能不显著。

**About logistic regression:**

1. 共线性检测（VIF）

* VIF（方差膨胀因子，Variance Inflation Factor）：用于检测回归模型中的多重共线性问题。当自变量之间存在高度相关性时，回归系数的估计会变得不稳定，VIF 就是用来衡量这种不稳定性的。
* 如何解读：
  + VIF 值：VIF 越大，表示变量之间的共线性越强，可能会影响模型估计的稳定性。一般来说：
    - VIF > 10：表示严重的多重共线性，需要考虑处理共线性问题（例如，删除变量或进行主成分分析等）。
    - VIF > 5：表明可能有共线性问题，值得关注。
  + 在这段代码中，vif(multi\_model) 用来计算多元逻辑回归模型中各个变量的 VIF 值。

2. OR森林图

这是一个用于展示多因素逻辑回归结果（如每个变量的 OR 值及其 95% 置信区间）的图表。OR（胜算比，Odds Ratio）表示在给定自变量值变化的情况下，因变量发生的相对机会增加的倍数。以下是步骤解释：

3. ROC 曲线和 AUC

* ROC 曲线（Receiver Operating Characteristic curve）：用于评估分类模型的性能。它展示了不同阈值下，模型的\*\*真正率（TPR）和假正率（FPR）\*\*之间的关系。
  + TPR：真正率，表示模型正确预测为正类的比例。
  + FPR：假正率，表示模型错误预测为正类的比例。
* AUC（Area Under Curve）：ROC 曲线下的面积，AUC 值介于 0 和 1 之间，AUC 值越接近 1，模型越优秀。常见的 AUC 解释：
  + AUC = 0.5：模型没有区分度，相当于随机猜测。
  + AUC > 0.7：表示模型具有一定的预测能力。
  + AUC > 0.8：表示模型具有较好的预测能力。

1. Estimate（回归系数，β值）

* 含义：回归系数 Estimate 表示自变量（例如，gender、age 等）对因变量（nul\_all\_bin）的影响程度。它说明自变量每增加一个单位，因变量的对数几率变化多少。
* 解释：在逻辑回归中，Estimate 是对数几率比的估计值（Log odds），表示自变量的一个单位变化对因变量发生某个事件的对数几率的影响。

例如：如果 Estimate = 0.5，意味着自变量每增加一个单位，因变量发生某事件的对数几率增加 0.5。

2. Std.Error（标准误差）

* 含义：标准误差 Std.Error 表示回归系数估计值的精度。标准误差越小，回归系数的估计就越精确。它是回归系数的不确定性的度量。
* 解释：它用于衡量回归系数估计值的波动性，标准误差越小，回归系数的估计值更可信。

例如：如果 Std.Error = 0.2，表示回归系数的标准误差为 0.2。

3. OR（优势比，Odds Ratio）

* 含义：OR 表示自变量每增加一个单位时，因变量发生某事件的相对风险。OR 是回归系数的指数化，表示某个自变量对因变量发生某事件的影响倍数。
* 解释：
  + OR = 1：表示自变量对因变量的发生概率没有影响。
  + OR > 1：表示自变量每增加一个单位时，因变量发生某事件的概率增加。
  + OR < 1：表示自变量每增加一个单位时，因变量发生某事件的概率降低。

例如：如果 OR = 2，表示自变量每增加一个单位，因变量发生某事件的概率是原来的 2 倍。

4. Wald\_Chi2（Wald χ²统计量）

* 含义：Wald 统计量用于检验回归系数是否显著不为零。
* 解释：Wald 统计量越大，回归系数越显著不为零。通常，Wald 统计量与 P 值一起使用来判断回归系数是否显著。

例如：如果 Wald\_Chi2 = 5，表示回归系数的显著性较强。

5. CI（置信区间，Confidence Interval）

* 含义：置信区间 CI 用于估计回归系数的范围，通常是 95% 置信区间，表示我们有 95% 的信心认为回归系数的真实值位于该区间内。
* 解释：置信区间越窄，表示估计值的精度越高。通常，如果置信区间不包括 1（对于 OR）或 0（对于 Estimate），则说明回归系数显著。

例如：CI = [1.2, 2.5] 表示回归系数的 95% 置信区间在 1.2 到 2.5 之间。

6. P.value（P 值）

* 含义：P 值用于检验回归系数是否显著不为零。它表示回归系数在原假设（系数为零）下观测到的概率。如果 P 值小于显著性水平（通常为 0.05），则拒绝原假设，认为回归系数显著不为零。
* 解释：P 值越小，回归系数显著不为零的可能性越大。

例如：如果 P.value = 0.02，表示回归系数在 2% 的概率下为零，可以拒绝原假设，认为回归系数显著。