Inference and Data Analysis with GAMs

Inference with Generalized Additive Models: Effective Degrees of Freedom (EDF)

1. Introduction to Inference in GAMs

- In this video, the focus is on interpreting the inference output from generalized additive models (GAMs). The key concept is understanding the "Effective Degrees of Freedom" (EDF) reported in the model summaries.
- EDF is crucial in hypothesis testing within the context of GAMs, similar to how degrees of freedom are used in linear regression models for T-tests and F-tests.

2. Degrees of Freedom in Linear Models

- In standard linear regression, the degrees of freedom associated with the residual sum of squares is $n - p + 1$, where n is the number of data points and $p + 1$ is the number of parameters in the model.
- Degrees of freedom are essential because they define the distribution (e.g., T-distribution) used in the inference process.

3. Effective Degrees of Freedom in GAMs

- In GAMs, the concept of EDF serves as an analogy to the degrees of freedom in linear models, though it is more complex due to the smoothing process involved.
- In GAMs, the functions $f_j(x_j)$ are not specified parametrically, so it's not straightforward to count the number of parameters as in linear models.

4. The Hat Matrix and EDF

- The hat matrix in linear regression helps in calculating degrees of freedom. The trace of this matrix equals $p + 1$, which gives the model degrees of freedom.
- For GAMs, a generalized version of the hat matrix is used, and the trace of this matrix provides the EDF.

5. Interpretation of EDF

- **EDF ≈ 1**: Indicates that the term could have been entered linearly rather than as a smooth term. If the EDF for a smooth term is close to 1, it suggests that a linear term might suffice.
- **EDF > 1**: Indicates that a nonlinear (smooth) relationship is more appropriate. Larger EDFs suggest that the relationship between the predictor and the response is more complex and should be modeled nonlinearly.

6. Visual and Numerical Analysis

- It is important to complement the numerical value of EDF with a visual inspection of the smooth plot. If the smooth plot shows a linear trend and the EDF is close to 1, it suggests a linear model might be sufficient.
- Conversely, if the EDF is large and the smooth plot shows significant nonlinearity, the smooth term is likely necessary.

7. Caution with EDF

- While EDF is a useful concept, it should be interpreted with caution. It is not directly analogous to degrees of freedom in linear models, and its reliability for inference is somewhat debated.
- Some argue that EDF can be misleading, and it should not be relied upon exclusively for making decisions about the model.

8. Practical Example

- For example, in a previous video, an EDF of 8.375 was observed for a predictor x_1 in a GAM. This high EDF suggests a complex, nonlinear relationship between x_1 and the response variable.

Conclusion

- Effective Degrees of Freedom (EDF) is a key concept in GAMs for assessing the complexity of the smooth terms. It helps determine whether a predictor should be included as a linear or nonlinear term, but it should be used alongside graphical analysis and interpreted with caution due to its limitations.

使用广义加性模型（GAM）进行推断：有效自由度（EDF）

1. GAM中的推断简介

- 本视频重点讲解如何解读广义加性模型（GAM）的推断输出，特别是理解模型总结中报告的“有效自由度”（Effective Degrees of Freedom, EDF）。
- 在GAM中，EDF在假设检验中起着类似于线性回归模型中自由度的作用。就像线性回归中自由度用于T检验和F检验一样，EDF也是GAM推断过程中的关键概念。

2. 线性模型中的自由度

- 在标准线性回归中，残差平方和的自由度为 $n - p + 1$，其中 n 是数据点的数量，$p + 1$ 是模型中的参数数量。
- 自由度很重要，因为它们定义了推断过程中使用的分布（例如T分布）。

3. GAM中的有效自由度（EDF）

- 在GAM中，EDF的概念类似于线性模型中的自由度，但由于涉及平滑过程，其复杂性更高。
- 在GAM中，函数 $f_j(x_j)$ 不是以参数形式指定的，因此不像在线性模型中那样容易计算参数数量。

4. Hat矩阵与EDF

- 在线性回归中，Hat矩阵帮助计算自由度。该矩阵的迹（即对角线元素之和）等于 $p + 1$，这给出了模型的自由度。
- 对于GAM，使用了Hat矩阵的广义版本，这个矩阵的迹提供了EDF的值。

5. EDF的解释

- **EDF ≈ 1**: 表示该项可以线性方式加入模型，而不是以平滑项的形式加入。如果一个平滑项的EDF接近1，说明该项可能以线性形式就足够了。
- **EDF > 1**: 表示非线性（平滑）关系更为适合。较大的EDF表明预测变量和响应变量之间的关系更复杂，应该以非线性方式建模。

6. 数值和视觉分析

- 重要的是将EDF的数值与平滑图的视觉检查结合起来。如果平滑图显示线性趋势且EDF接近1，说明线性模型可能就足够了。
- 反之，如果EDF较大且平滑图显示显著的非线性，那么保留该平滑项是合理的。

7. 对EDF的谨慎使用

- 尽管EDF是一个有用的概念，但在解释时应谨慎。它并不完全等同于线性模型中的自由度，其在推断中的可靠性也存在争议。
- 一些研究认为EDF可能具有误导性，因此不应仅依赖它来做出模型决策。

8. 实例

- 例如，在之前的视频中，GAM中的一个预测变量 x_1 的EDF为8.375。这个高EDF值表明 x_1 与响应变量之间存在复杂的非线性关系。

总结

- 有效自由度（EDF）是GAM中评估平滑项复杂性的关键概念。它帮助确定一个预测变量应该以线性还是非线性形式加入模型。然而，由于其局限性，EDF应与图形分析结合使用，并谨慎解释。

使用广义加性模型（GAMs）进行推断：假设检验

1. GAM中的假设检验

- 在分析GAM时，R会在模型的总结输出中提供一些假设检验。这些检验结果是近似的，类似于我们对有效自由度（EDF）的解释，不应被视为最终结论。
- 这些检验可以用作粗略的近似，帮助我们评估某些预测变量是否应该留在模型中。

2. 平滑项的假设检验

- GAM的总结输出中，与平滑项相关的检验通常是近似F检验，这些检验用于检验给定的平滑项是否为零。
  - **原假设（H0）**：平滑项为零（即该项不应作为平滑项进入模型）。
  - **备择假设（H1）**：平滑项非零（即该项应以平滑项的形式存在于模型中）。
- 如果与平滑项相关的p值非常小（例如小于0.1），则这提供了一些证据，表明该平滑项应该保持平滑状态，而不是线性化。

3. 模型的拟合优度

- GAM的总结输出还提供了一些类似于线性模型和广义线性模型中的拟合优度信息，如偏差解释率（Deviance Explained）和调整后的R平方（Adjusted R-squared）。
  - **偏差解释率（Deviance Explained）**：类似于线性模型中的决定系数R平方，表示模型解释了响应变量变异的百分比。

- **调整后的R平方**: 对偏差解释率进行调整，以考虑样本量较小或模型复杂性（如预测变量的数量）。它尝试对更复杂的模型进行一定的惩罚，因此在不同模型之间进行比较时，调整后的R平方更为合理。

4. 检验结果的解释

- 这些检验结果可以帮助我们判断在GAM中是否需要一个平滑项。虽然这些检验是近似的，但它们为我们提供了一种粗略的方式来评估平滑项的必要性。
- 例如，如果我们看到某个平滑项的p值非常小，那么这个项很可能确实需要保持平滑，而不是被简化为线性项。

5. 模型比较

- 由于调整后的R平方考虑了模型的复杂性，因此可以用于不同模型之间的比较。更高的调整后R平方值通常意味着更好的模型拟合，但同时也要注意不要过度拟合。

总结

- 在GAM中，假设检验提供了一种粗略的方式来评估模型中平滑项的必要性。虽然这些检验结果不应被视为绝对的决定因素，但它们仍然可以为模型改进和变量选择提供有价值的指导。此外，偏差解释率和调整后的R平方等拟合优度指标在模型评价和比较中也发挥着重要作用。

偏差解释率（Deviance Explained） 是广义线性模型（GLM）和广义加性模型（GAM）中用来衡量模型拟合优度的一个指标。它类似于线性回归中的 R^2（决定系数），表示模型解释了响应变量变异的百分比。

偏差（Deviance）概念

- **偏差（Deviance）** 是一个衡量模型拟合效果的统计量，用于比较一个模型与最理想模型（即“完全拟合”的模型）的拟合差异。
- 偏差可以看作是一个模型的“残差平方和”的推广，它基于对数似然函数，尤其适用于非正态分布的数据（例如二项分布或泊松分布）。

偏差解释率的计算

- **残差偏差（Residual Deviance, D_{res}）**：模型拟合的偏差，即模型未能解释的部分。
- **空模型偏差（Null Deviance, D_{null}）**：基线模型（通常只包含截距项）的偏差，表示没有预测变量时模型的偏差。

偏差解释率的计算公式为：

$$\text{Deviance Explained} = 1 - \frac{D_{res}}{D_{null}}$$

其中：

- D_{res}：当前模型的残差偏差。
- D_{null}：空模型的偏差。

解读偏差解释率

- **偏差解释率值接近1**：表示模型解释了大部分的响应变量变异，拟合优度较好。
- **偏差解释率值接近0**：表示模型对响应变量变异的解释能力较差，拟合效果不好。

例如，如果某个GAM模型的偏差解释率为0.83（即83%），这意味着该模型解释了响应变量83%的变异，其拟合效果是较为优异的。

偏差解释率的应用

偏差解释率在广义线性模型和广义加性模型的结果中用于评估模型的整体表现。当选择模型或比较不同模型时，偏差解释率是一个重要的参考指标。较高的偏差解释率通常表示模型的预测能力更强，拟合更精确。

调整后的 R^2（Adjusted R-squared）是在线性回归和广义加性模型（GAM）中用来衡量模型拟合优度的指标之一。它在普通 R^2 的基础上进行了调整，以便更好地反映模型的复杂性和解释能力。

普通 R^2 的局限性

- R^2 是一个常用的指标，用于衡量模型解释了响应变量总变异的百分比。
- R^2 的取值范围是 0 到 1，值越接近 1 表示模型的解释能力越强。
- 但是，普通的 R^2 随着模型中变量的增加总是会变大，即使这些变量可能没有实际的解释力。因此，当增加不相关的预测变量时， R^2 也可能虚高。

调整后的 R^2 的计算

调整后的 R^2 对模型的复杂性进行了惩罚，以避免简单地通过增加预测变量来提升 R^2 值。它的计算公式为：

$$\text{Adjusted } R^2 = 1 - \left(\frac{(1 - R^2)(n - 1)}{n - p - 1} \right)$$

其中：

- R^2：普通的 R^2 值。
- n：样本的数量。
- p：模型中的预测变量的数量。

调整后的 R^2 的特点

- **惩罚模型复杂性**：调整后的 R^2 对于增加的预测变量进行了惩罚，因此它不会因为添加无关变量而提高。如果增加的变量并未显著提高模型的拟合效果，调整后的 R^2 可能会下降。
- **模型比较**：调整后的 R^2 更适合用来比较具有不同数量预测变量的模型。一个模型的调整后的 R^2 较高，通常表明它在考虑了模型复杂性的情况下，解释了更多的响应变量变异。

调整后的 R^2 的解读

- **值接近1**：表示模型在解释响应变量变异方面表现优异，并且模型的复杂性得到了合理的控制。
- **值较低或负值**：如果调整后的 R^2 值非常低甚至为负，可能表明模型的预测能力很差，或者增加的变量没有提供额外的解释力，反而增加了噪声。

应用场景

调整后的 R^2 通常用于模型选择和比较。尤其是在比较不同复杂度的模型时，它能够帮助你确定增加的预测变量是否真正有助于提高模型的解释力，而不是仅仅因为变量数量的增加而提高 R^2 值。

Inference with Generalized Additive Models (GAMs): Hypothesis Tests

1. Hypothesis Testing in GAMs

- When analyzing GAMs in R, the model summary includes some hypothesis tests. These tests are approximate, similar to our interpretation of Effective Degrees of Freedom (EDF), and should not be taken as definitive.
- These tests serve as rough approximations to assess whether certain predictors should remain in the model.

2. Hypothesis Tests for Smooth Terms

- The tests associated with smooth terms in a GAM summary are typically approximate F-tests. These tests evaluate the hypothesis that the given smooth term is zero.

    - **Null Hypothesis (H0):** The smooth term is zero (i.e., it should not be included as a smooth term in the model).
    - **Alternative Hypothesis (H1):** The smooth term is non-zero (i.e., it should be included as a smooth term in the model).
- If the p-value associated with a smooth term is very small (e.g., below 0.1), it suggests that the term should remain as a smooth rather than being linearized.

3. Model Goodness-of-Fit

- The GAM summary also provides goodness-of-fit information, similar to what we see in linear and generalized linear models. Two key metrics are Deviance Explained and Adjusted R-squared.
    - **Deviance Explained:** This is analogous to the R-squared (coefficient of determination) in linear models, indicating the percentage of variability in the response explained by the model.
    - **Adjusted R-squared:** This is a version of deviance explained that adjusts for sample size or model complexity (e.g., the number of predictors). It penalizes more complex models, making it useful for comparing models.

4. Interpreting the Test Results

- These tests help determine whether a smooth term is necessary in the GAM. Although approximate, they offer a rough indication of the term's significance.
- For example, if a smooth term has a very small p-value, it likely needs to remain as a smooth term rather than being simplified to a linear term.

5. Model Comparison

- Adjusted R-squared is particularly useful for comparing models, as it accounts for model complexity. A higher adjusted R-squared generally indicates a better fit, but care must be taken to avoid overfitting.

Conclusion

- In GAMs, hypothesis tests provide a rough method to evaluate the necessity of smooth terms in the model. While these tests are not definitive, they offer valuable guidance for model refinement and variable selection. Additionally, goodness-of-fit metrics like Deviance Explained and Adjusted R-squared play crucial roles in evaluating and comparing models.

Generalized Additive Models in R: Inference and Interpretation

1. Overview of the Example

- The video describes principles for interpreting generalized additive models (GAMs) and how to perform inference using them.
- The example involves simulated data with three predictors: two continuous variables and one three-level factor. The relationship between the response and the predictors is modeled using an additive model with a normality assumption.

2. Fitting the Additive Model

- The `gam` function from the `mgcv` package is used, similar to how `lm` and `glm` functions are used for linear and generalized linear models, respectively.
- The key difference is that one of the predictors, x_1, is wrapped in an `s()` function, indicating that it should be smoothed.

3. Interpretation of Parametric Terms

- **Parametric terms** (e.g., x_2 and x_3) that enter the model linearly can be interpreted similarly to how they are in standard regression.
  - For example, moving from level A to level B of the factor x_3 results in a mean increase of approximately 1.028 in the response, adjusting for other predictors like x_1 and x_2.
  - A one-unit increase in x_2 results in a mean increase of about 2.814 in the response, adjusting for the levels of x_3 and the non-linear effect of x_1.

4. Interpretation of the Smoothed Term

- The interpretation of the smoothed term x_1 is more complex because it does not enter the model linearly.
- Interpretation can be recovered by plotting the marginal relationship between x_1 and the response, adjusting for other predictors.
  - For example, between -3 and -1 on the x_1 axis, there is a sharp decrease in the response, followed by an increase from -1 to 1, and then a decrease from 1 to 3.

5. Inference on Parametric and Non-Parametric Terms

- **Parametric Terms:** The p-values for parametric terms (e.g., x_2) are interpreted in the same way as in linear models. Small p-values suggest that the coefficients are significantly different from zero.
- **Smoothed Terms:** A small p-value for a smoothed term (e.g., x_1) suggests that the term should remain in the model as a smooth rather than being flattened to zero.
  - The effective degrees of freedom (EDF) for x_1 is above 8, indicating a need for a non-linear fit.

6. Model Comparison and Interpretation

- A second model is fitted where both x_1 and x_2 are smoothed. The EDF for x_2 is much smaller than for x_1, suggesting that x_2 may not need to be smoothed and could enter the model linearly.
- This interpretation is supported by plotting the marginal relationship for x_2, which shows only slight curvature, implying that a linear fit could be sufficient.

7. Model Performance Evaluation

- The Mean Squared Error (MSE) is used to compare models on a test set:
  - The original model with only x_1 smoothed has an MSE similar to the model with both x_1 and x_2 smoothed, indicating that smoothing x_2 did not improve predictive performance significantly.
  - Both GAM models outperform a simple linear model, which has a much higher MSE, especially because x_1 is incorrectly assumed to be linear in the simple model.

Conclusion

- Generalized Additive Models (GAMs) allow for a mix of parametric and non-parametric terms, offering both interpretability and flexibility.
- Inference on parametric terms is similar to linear models, while inference on smoothed terms relies on p-values and EDF to determine the necessity of smoothing.
- Model comparison using MSE on test data is a practical approach to evaluate the effectiveness of different model specifications.

R中的广义加性模型：推断与解释

1. 示例概述

- 本视频讲解了如何解释广义加性模型（GAM）以及如何使用这些模型进行推断。
- 示例数据包含三个预测变量：两个连续变量和一个三水平因子。模型假设响应变量与这些预测变量之间的关系是加性模型，并假设数据符合正态分布。

2. 拟合加性模型

- 使用 `mgcv` 包中的 `gam` 函数来拟合模型，这类似于使用 `lm` 和 `glm` 函数来拟合线性和广义线性模型。
- 关键区别在于，某些预测变量（例如 x_1）使用 `s()` 函数包裹，表示该变量需要进行平滑处理。

3. 参数项的解释

- **参数项**（如 x_2 和 x_3）以线性方式进入模型，可以类似于标准回归模型进行解释。
  - 例如，从因子 x_3 的水平A转换到水平B时，响应变量的平均增加约为1.028，调整了其他预测变量（如 x_1 和 x_2）。
  - x_2 增加一个单位时，响应变量的平均增加约为2.814，调整了 x_3 的水平和 x_1 的非线性效应。

4. 平滑项的解释

- x_1 的平滑项的解释较为复杂，因为它不是以线性方式进入模型的。
- 可以通过绘制 x_1 和响应变量之间的边际关系图来解释，图中调整了其他预测变量的影响。
  - 例如，在 x_1 轴的-3到-1之间，响应变量显著下降；在-1到1之间，响应变量显著上升；而在1到3之间，响应变量再次下降。

5. 参数项和非参数项的推断

- **参数项**：参数项（如 x_2）的p值可以像在线性模型中那样解释。较小的p值表明这些系数显著不同于零。
- **平滑项**：平滑项（如 x_1）的p值较小，表明该项应以平滑方式保留在模型中，而不是被简化为零。
  - x_1 的有效自由度（EDF）大于8，表明需要进行非线性拟合。

6. 模型比较与解释

- 拟合了另一个模型，其中对 x_1 和 x_2 都进行了平滑处理。x_2 的EDF比 x_1 小得多，表明 x_2 可能不需要平滑处理，可以线性地进入模型。
- 通过绘制 x_2 的边际关系图，发现其曲率很小，说明线性拟合可能已经足够。

7. 模型性能评估

- 使用均方误差（MSE）来比较测试集上的模型表现：
  - 仅对 x_1 进行平滑处理的原始模型的MSE与对 x_1 和 x_2 都进行平滑处理的模型的MSE基本相同，表明对 x_2 的平滑处理并未显著提高预测性能。
  - 两个GAM模型的表现都优于简单线性模型，后者的MSE高得多，特别是因为在简单模型中错误地假设 x_1 与响应变量之间是线性关系。

总结

- 广义加性模型（GAM）允许混合参数和非参数项的使用，提供了可解释性和灵活性。
- 对参数项的推断类似于线性模型，而对平滑项的推断则依赖于p值和EDF来确定是否需要平滑处理。
- 通过在测试数据上的MSE比较模型，是评估不同模型规格有效性的实际方法。

广义加性模型：使用真实数据的完整示例

1. 引言与数据概述

- 在本课中，我们使用广义加性模型（GAMs）在R中分析真实数据，特别是再次分析之前在学习泊松回归时使用的纽约市自行车数据。
- 我们的目标是通过广义加性模型来获得比之前使用的广义线性模型（GLM，特别是泊松回归）更好的拟合效果和预测能力。
- 数据的目标是记录通过东河桥（如曼哈顿大桥）进出皇后区、曼哈顿和布鲁克林的自行车骑行者数量。每条记录表示24小时内通过曼哈顿大桥的骑行者总数。我们将尝试使用天气数据（如当天的最高温度、最低温度和降水量）来预测骑行者数量。

2. 数据处理与初步分析

- 数据读取和清理之后，数据被分为训练集和测试集。我们将在训练集上拟合模型，并使用测试集来评估模型的预测性能。
- 我们首先探索了响应变量（即骑行者数量的对数）与几个预测变量（最高温度、最低温度和降水量）之间的边际关系。
  - 初步分析显示，骑行者数量的对数与最低温度之间的关系接近线性，但存在一些非线性特征。

- 类似地，骑行者数量的对数与最高温度之间也显示出某些非线性趋势。
- 降水量与骑行者数量之间的关系更加复杂，存在一些非线性趋势，特别是在降水量较高时。

3. 拟合广义加性模型

- 使用 `mgcv` 包中的 `gam` 函数来拟合模型。这里，我们将天气变量（最高温度、最低温度和降水量）以平滑项的形式加入模型，而星期几作为因子变量线性地进入模型。
- 模型的家族设置为泊松分布，这是因为我们处理的是计数数据，而泊松回归是适合这类数据的模型。

4. 平滑项和参数项的解释

- **平滑项解释**：
  - 有效自由度（EDF）较高，表明这些项应以平滑形式进入模型，而不是线性地进入模型。
  - 通过 `plot.gam` 函数绘制边际关系图，可以看到调整了其他预测变量后，最低温度与响应变量之间的关系存在显著的非线性趋势。
  - 类似地，最高温度与响应变量之间也显示出非线性关系，特别是在温度达到某个临界点时，骑行者数量的增加速度开始减缓。
  - 降水量的边际关系显示出更复杂的非线性趋势，在降水量增加时，骑行者数量总体呈下降趋势。
- **参数项解释**：
  - 星期几作为因子变量进入模型，有七个水平（星期一到星期日），P值表明所有的T检验都非常显著。这表明星期几对骑行者数量有统计上的显著影响。
  - 例如，模型显示在调整了天气条件后，星期六和星期日的骑行者数量显著低于星期五。

5. 模型比较与预测性能

- 与之前拟合的广义线性模型（泊松回归）进行比较：
  - GAM模型允许非线性关系，而GLM模型假设所有关系都是线性的。
  - 虽然GLM模型的伪R平方（McFadden's R-squared）接近0.7，但GAM模型的偏差解释率（Deviance Explained）达到了83%，显示出更好的拟合效果。
  - 在测试集上的均方误差（MSE）进一步支持了GAM模型的优势：GLM的MSE为808,855，而GAM的MSE则显著降低至662,778。这表明GAM在预测性能方面优于GLM。

总结

- 广义加性模型结合了非线性拟合的灵活性和广义线性模型处理不同类型响应变量的能力。在这个真实数据的示例中，GAM模型比传统的GLM模型提供了更好的拟合效果和更强的预测能力，特别是在数据中存在非线性关系时。