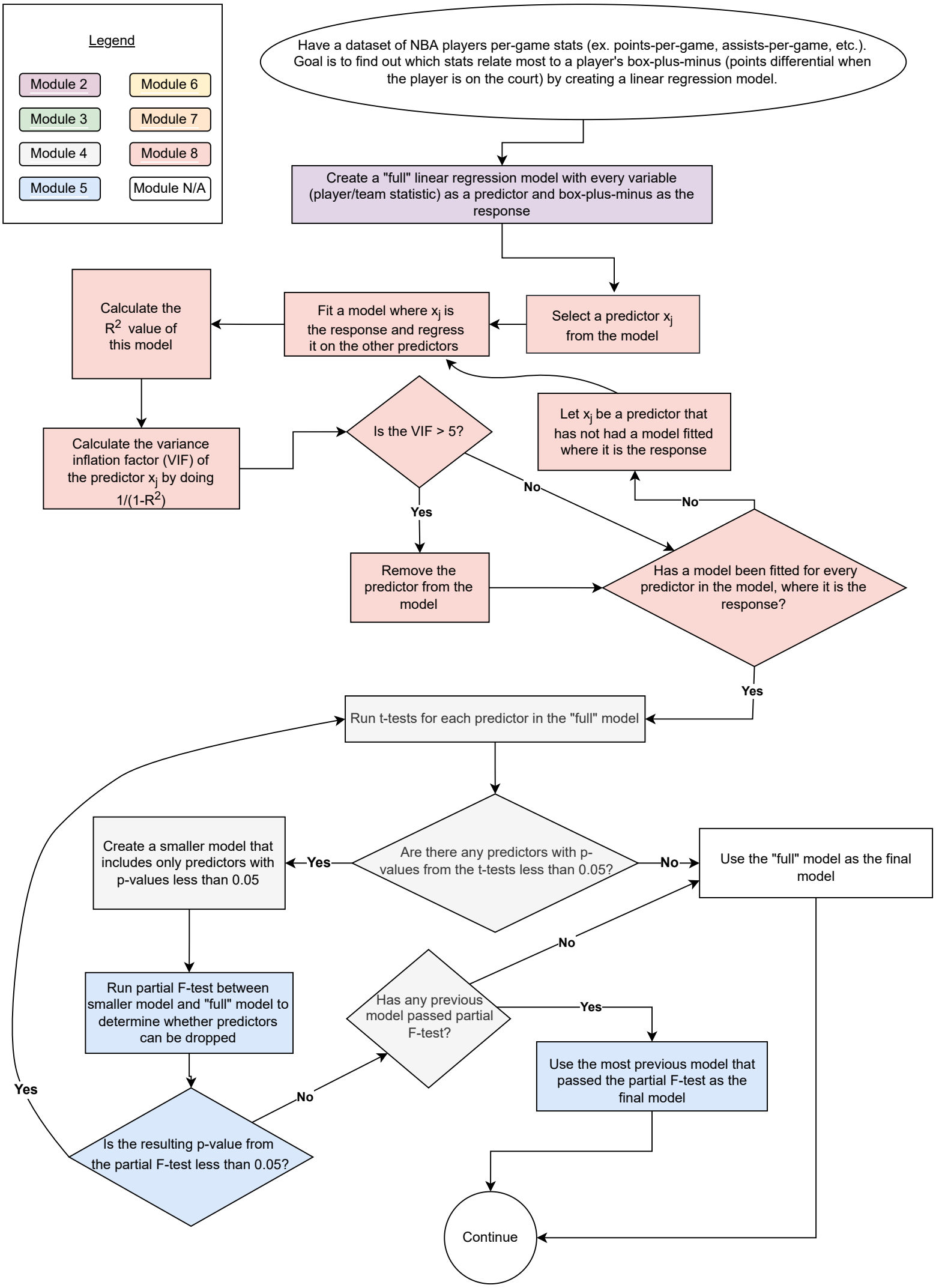
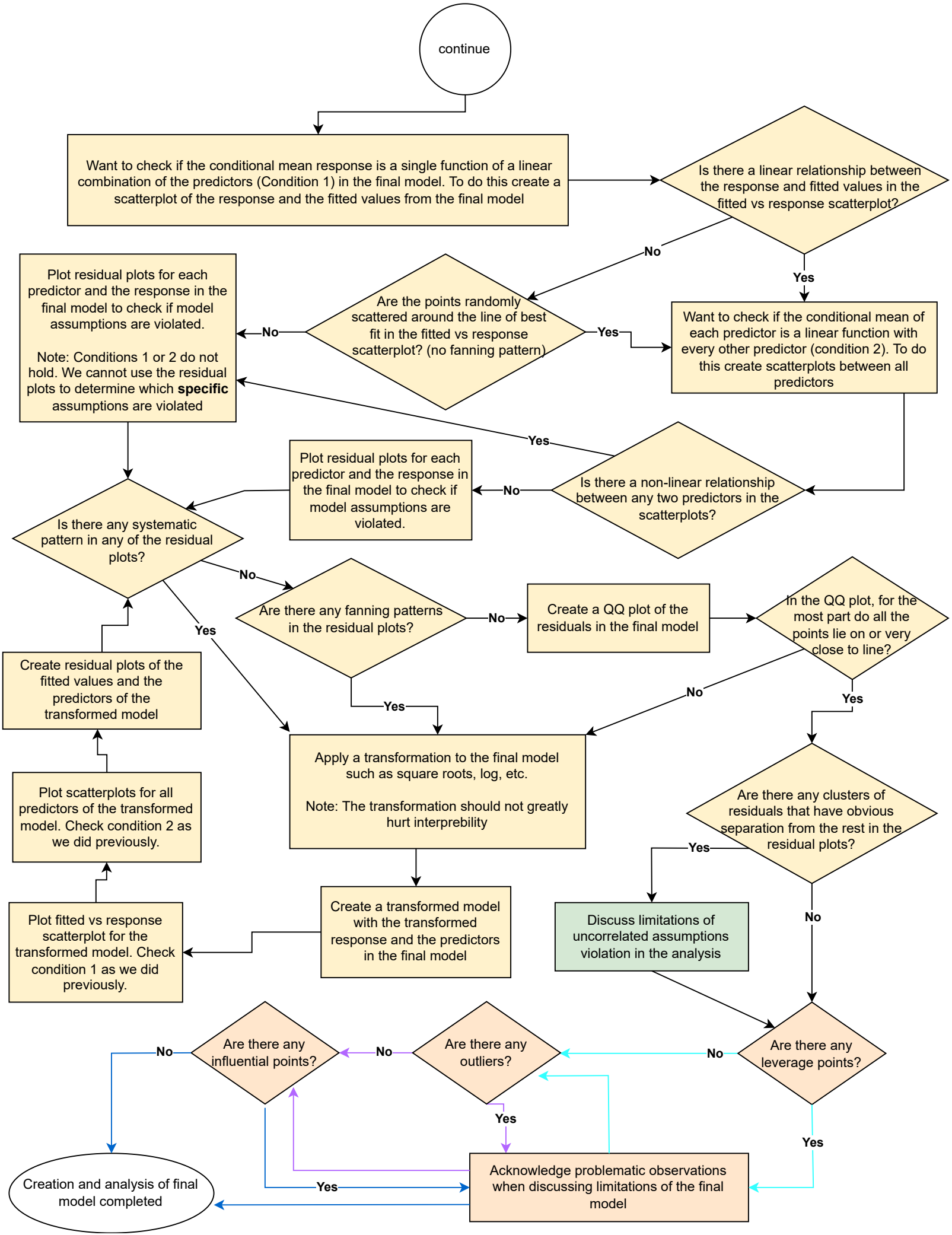

Legend
Module 2
Module 3
Module 4
Module 5
Module 6
Module 7
Module 8
Module N/A
Have a dataset of NBA players per-game stats (ex. points-per-game, assists-per-game, etc.). Goal is to find out which stats relate most to a player's box-plus-minus (points differential when the player is on the court) by creating a linear regression model.
Create a "full" linear regression model with every variable (player/team statistic) as a predictor and box-plus-minus as the response
Select a predictor x_j from the model
Fit a model where x_j is the response and regress it on the other predictors
Calculate the R^2 value of this model
Calculate the variance inflation factor (VIF) of the predictor x_j by doing $1/(1-R^2)$
Is the VIF > 5?
Yes
No
Remove the predictor from the model
Let x_j be a predictor that has not had a model fitted where it is the response
No
Has a model been fitted for every predictor in the model, where it is the response?
Yes
Run t-tests for each predictor in the "full" model
Are there any predictors with p-values from the t-tests less than 0.05?
Yes
No
Create a smaller model that includes only predictors with p-values less than 0.05
Use the "full" model as the final model
Run partial F-test between smaller model and "full" model to determine whether predictors can be dropped
Has any previous model passed partial F-test?
No
Yes
Use the most previous model that passed the partial F-test as the final model
Is the resulting p-value from the partial F-test less than 0.05?
No
Yes
Continue


continue
Want to check if the conditional mean response is a single function of a linear combination of the predictors (Condition 1) in the final model. To do this create a scatterplot of the response and the fitted values from the final model
Is there a linear relationship between the response and fitted values in the fitted vs response scatterplot?
No
Yes
Are the points randomly scattered around the line of best fit in the fitted vs response scatterplot? (no fanning pattern)
No
Yes
Plot residual plots for each predictor and the response in the final model to check if model assumptions are violated.
Note: Conditions 1 or 2 do not hold. We cannot use the residual plots to determine which specific assumptions are violated
Want to check if the conditional mean of each predictor is a linear function with every other predictor (condition 2). To do this create scatterplots between all predictors
Yes
Is there a non-linear relationship between any two predictors in the scatterplots?
No
Plot residual plots for each predictor and the response in the final model to check if model assumptions are violated.
Is there any systematic pattern in any of the residual plots?
No
Are there any fanning patterns in the residual plots?
No
Create a QQ plot of the residuals in the final model
In the QQ plot, for the most part do all the points lie on or very close to line?
No
Yes
Yes
Yes
Create residual plots of the fitted values and the predictors of the transformed model
Apply a transformation to the final model such as square roots, log, etc.
Note: The transformation should not greatly hurt interpretability
Plot scatterplots for all predictors of the transformed model. Check condition 2 as we did previously.
Are there any clusters of residuals that have obvious separation from the rest in the residual plots?
Yes
No
Plot fitted vs response scatterplot for the transformed model. Check condition 1 as we did previously.
Create a transformed model with the transformed response and the predictors in the final model
Discuss limitations of uncorrelated assumptions violation in the analysis
No
Are there any influential points?
No
Are there any outliers?
No
Are there any leverage points?
Yes
Yes
Yes
Creation and analysis of final model completed
Yes
Acknowledge problematic observations when discussing limitations of the final model