HW for week 10

The red parts are directly quoted, either from a website or from chatbots.

Graded Problems:

Problem 4

Note that R\*\*2 focuses on how much of the overall variation in HP is explained by the model.

According to Wikipedia, “R\*\*2 is a measure of the goodness of fit of a model.”

“*R*2 does not indicate whether: the correct regression was used;”

“The most appropriate set of independent variables has been chosen;”

In this case (pokeaman), the high coefficients and low P-values do imply that the outcome is affected by the predictor variables. Yet, this does not mean that outcome is **ONLY** explained by the predictors in the model. The predictors do have a linear relation with the outcome, but some other variables (like the attack or class, etc) might be ignored in the model, which leads to the model’s failed in explaining 80%+ of the variance of the data (as R\*\*2 tells us)

On the other hand, small R\*\*2 might also suggest that the outcome is highly random, so that it is impossible to capture the pattern even when we have found most of the related variables.

Problem 7

Problem 7 shows an approach to improve the pokeaman codes.

In model 5, to save the model from **multicollinearity,** all the interaction terms are removed from the model.

In model 6, those variables proved to be significant in model 5 are kept while the rest are removed. This should slightly lower the performance of R\*\*2 of the model, but further reduce the risk of multicollinearity.

In model 7, after removing the unrelated variables, some interaction terms are introduced back again to improve the performance.

The late cell calculated the “real” Cond. No. of model 7 after centering and scaling. Though the number is still very large, the model did not lose its generalizability, because the irrelevant variables has been removed, avoiding the problem of overfitting.

Problem 9

Two sequential data splits are used in the illustration:

* **Generation 1 model** (model\_gen1\_predict\_future): Trained on Generation 1 Pokémon only and tested on Pokémon from later generations.
* **Generations 1–5 model** (model\_gen1to5\_predict\_future): Trained on Generations 1–5 Pokémon and tested on Generation 6 Pokémon.

Model simplicity, generalizability, and interpretability are crucial in predictive modeling. The comparison of model6\_fit and model7\_fit demonstrates these principles.

As the result, Model7\_fit shows better predictive performance (higher out-of-sample R²) compared to model6\_fit but features higher-order interaction terms, leading to concerns about overfitting and interpretability. Models are validated by training on early Pokémon generations and testing on future generations to simulate real-world prediction.

Model7\_fit performs well on training data but struggles to generalize to new data, indicating overfitting to training-specific patterns. Model6\_fit, with its simpler structure, performs more consistently and generalizes better across new generations.

Interpretability also holds significant importance. Model6\_fit, with fewer and simpler terms, proves easier to understand and apply in practice, supporting transparency and stakeholder trust. Complex interactions in model7\_fit, such as four-way interactions, reduce interpretability and make practical application challenging.

Simplicity enhances generalizability while complexity risks overfitting. Model6\_fit **balances interpretability and generalizability,** making it a **more reliable choice for practical use**. Complex models should only be favored if they consistently provide substantial gains in performance across different conditions. In this scenario, model6\_fit achieves comparable performance with better generalizability and interpretability, emphasizing the benefits of a simpler approach.

Not Graded Problems:

Problem 1

Chatbot is doing well on this so I have nothing to add about it.

**1. Difference between Simple and Multiple Linear Regression:**

* **Simple Linear Regression** models the relationship between a dependent variable and a single independent (predictor) variable.
* **Multiple Linear Regression** models the relationship between a dependent variable and multiple independent variables.

**Continuous vs. Indicator Variables in Simple Linear Regression:**

* **Continuous Variable**: Takes on a range of numerical values, allowing the model to estimate a gradient or slope in the relationship.
* **Indicator Variable**: A binary variable (usually 0 or 1) representing categories; it shifts the baseline of the dependent variable based on the category.

**3. Introducing an Indicator Variable in Multiple Linear Regression:**

* **Model Behavior**: By adding an indicator variable alongside a continuous variable, the model can now estimate separate intercepts (i.e., baseline differences) for the indicator's categories, in addition to the effect of the continuous variable.
* **Linear Forms**: Simple Linear Regression provides a single line, while Multiple Linear Regression with an indicator and a continuous variable produces separate parallel lines for each category of the indicator.

**4. Effect of Adding an Interaction between a Continuous and an Indicator Variable:**

* **Effect**: Interaction allows the continuous variable’s effect to vary across categories of the indicator variable, creating different slopes for each category.
* **Linear Form**: This model results in non-parallel lines for each category, allowing for category-specific slopes that reflect a unique relationship between the continuous predictor and the outcome in each category.

**5. Multiple Linear Regression with Indicator Variables for Non-Binary Categorical Variables:**

* **Model Behavior**: When using only indicator variables derived from a categorical variable with multiple levels, the model estimates separate intercepts for each category.
* **Linear Form**: This form represents each category’s effect on the outcome as a unique intercept, leading to parallel lines for each category (if no continuous variables or interactions are present).
* **Binary Encodings**: It uses dummy (or one-hot) encoding to represent each category with binary variables (0 or 1), allowing the model to distinguish between categories.

Problem 2

**Model without Interaction**:

Sales=β0+β1⋅ Online (High)

Where Online(High) is a binary indicator variable that is 1 only when the online budget is high and is 0 when it’s not.

**Model with Interaction**:

Sales=β0+β1⋅TV (High)+β2⋅Online (High)+β3⋅(TV (High)×Online (High)

Here an interaction term β3⋅(TV (High)×Online (High) is introduced to represent the interaction of the Online budget on TV AD’s sale (and vice versa.)

If the two predictor variables are continuous, the model still holds the same, except the indicators are now replaced by real continuous variables.

Problem 3

Please check the “course project” responsory

Problem 5

See the notes on the “Problem 5” cells.

Problem 6

In model 4, Attack \* Defense \* Speed \* Legendary \* Q("Sp. Def") \* Q("Sp. Atk") are considered as input variable and they are highly correlated.

The problem of this approach is that many unrelated factors are taken into consideration. And since they are correlated, multicollinearity is created, making the model struggling in separating the individual contribution of each variable (and thus allows us to remove those unrelated variables)

What’s worse, since the variables are highly corelated, the small, random patterns in the training dataset are amplified. (since they are “counted for multiple times”) Leading to the model’s failed to capture the general patterns of the data.

Problem 8

The loops are runner to create different training and testing samples to see the model’s robustness across different samples.