Tuesday: Linear Regression Optimization - Best Practices

Agenda

5 min: Overview

45 min: Suggested Readings

90 min: Exercises

Specific Learning Outcomes

- I can incorporate categorical independent variables into their models.
- I can check for multicollinearity, and understand the circumstances when it is relevant to do so.
- I can create residual plots for their models, and assess their heteroskedasticity using Barlett's test.

Overview

Encoding categorical features

Linear regression is best suited to handle continuous, numerical variables. If you want to include categorical independent variables in your model, then you will have to encode them as numerical ones.

This can be done very simply as you will see in the example. For a variable with two categories, we can encode them as binary values of either 0 or 1. For variables with multiple categories, there are a few approaches you can explore by following this **tutorial. (https://towardsdatascience.com/encoding-categorical-features-21a2651a065c)**

Multilinearity

One of our fundamental assumptions when performing multivariate linear regressions is that our independent variables are truly independent: They must not be strongly related to each other.

In instances when there is a relationship between your independent variables, you may notice oddities when exploring the coefficients your model assigned to each independent variable:

- An independent variable that in theory should strongly influence the dependent variable is given an extremely low coefficient.

- An independent variable that is positively related to the dependent variable is given a negative coefficient or vice versa.
- When you add or delete an independent variable from the model, the coefficients change drastically.

In the demo below, we will show how to compute the **Variance Inflation Factor** (VIF). This is a measure of how much the *variance* of a regression coefficient in your model increases if your independent variables are correlated. If no independent variables are correlated, you'll expect the VIF for each to be 1.

Typically, a VIF value around 5 is a potential problem, and value around 10 is considered seriously problematic and suggests that the related variable should be dropped from the model.

Now one important thing to keep in mind is that **high multicollinearity does not make for a bad predictive model,** rather it poses the following challenges:

- The model may be trained inefficiently, you may get similar performance faster by not using some of the highly correlated variables. This is more relevant the more your dataset grows
- The **coefficients** of your model may not be useful to interpret. If you were interested in not only predicting your dependent variable but also **understanding how the various independent variables contribute to it,** then you should perform a multicollinearity check **before** assessing the coefficients.

Residual plots and the Heteroskedasticity test

Residual plots are a powerful tool for assessing the correctness of your model. The graph is straightforward:

- On the x-axis, you will have the predicted values of your model.
- On the y-axis, you will have the difference between the actual values and said predicted values - also known as the **residual.**

Fundamentally, there will always be randomness to our predictions, and the residual plot will help us determine that our errors are indeed **due to chance:** There should be **no predictability, no correlation between any variable, and the residual.** If there is, that means your model is not capturing all the elements that deterministically influence your dependent variable.

This can be assessed by performing a **heteroskedasticity test.**

While this is probably one of the most complex words we've seen so far, the concept behind it is simple: Is the variability of a given dependent variable **unequal** (heteroscedasticity) or **even** (homoscedasticity) across the range of values of an independent variable that predicts it.

Here is an example: Imagine trying to predict **income** based on **age:** At ages under ~20 years old, we should have fairly consistent variability: While in school, many teenagers do not work, and if you do at this stage your salaries will tend to be within a small range.

As you move to your mid 20's and start your first jobs, your choice of industry influences salaries much more drastically, leading to a higher variability. This keeps on increasing over time.

Therefore, performing a heteroskedasticity test on our residual plot would help us determine if the variability of a given dependent variable unequal.

Exercises

- Examples and Practice [link (https://colab.research.google.com/drive/19yG1_hBDnVv0dZ2yvwx4Lvonckq3tSHg?usp=sharing)]

Suggested Readings

- Encoding categorical features [link (https://towardsdatascience.com/encoding-categorical-features-21a2651a065c)]
- Tackling multicollinearity [link (https://stattrek.com/multiple-regression/multicollinearity.aspx)]
- Introducing the concept of residual plots [link (https://statisticsbyjim.com/regression/check-residual-plots-regression-analysis/)]
- Heteroskedasticity for residual plots [link (https://statisticsbyjim.com/regression/heteroscedasticity-regression/)]

"Machine learning will automate jobs that most people thought could only be done by people." ~ Dave Waters