Bonnie Turek

Eco 602 – Week 12 Reading Questions

11/16/2021

Model Comparison

Q1. In the context of a dataset (real or made up), describe the inherent conflict between using a complicated model that minimizes the unexplained variation and using a simple model that is easy to communicate. Consider the trade-off between model complexity and interpretability.

There are most certainly trade offs between model complexity and accuracy and interpretability. Typically, more complex models have greater accuracies and might be based on unseen data or data that is harder to visualize. Whereas, you could have a much simpler model, like a simple linear regression, that is easy for most people to understand and visualize, however the accuracy of the model might be poor. For example, some of the simplest models are linear regressions or decision trees, while the more complex models include machine learning such as support vector machines and neural networks.

Interpretability of your model is certainly important because it encourages a better understanding of the problem and the data. By keeping your model simple and easily interpretable, it also helps to ensure there is not too much bias in your model. Generally, there may be more social acceptance and response to your model and your research as well. On the other hand, model accuracy is also very important. So many people are conducting research on a vast number of topics today, and in order to really advance the research in a particular field, you'll probably have to dive into the data and at least explore more complex models. Some may refer to complex models as a "black box." This is somewhat concerning because at some point, you aren't fully aware of what is going on behind the data or software you might be using. Trying to explain that to your audience would be even more challenging. A model with higher accuracy, however, can lead more opportunities, benefits, time or money for a company or university. But unfortunately, the predictive models that are most powerful are usually the least interpretable.

As an example, I'll use my own research data related to modeling sediment organic carbon in tidal marshes using various spatial metrics. In the simplest sense, I could create a simple linear regression of Carbon Content versus Distance to the tidal outlet. This is simple to understand, and we might see a general correlation between these two variables: that carbon content increases with increasing distance from the tidal outlet. I can explain this simply, considering that the tidal outlet, which is closest to the ocean provides the most mineral (and least organic and carbon sediments). There is more organic and carbon sediment in the marsh as you move away from the tidal outlet and into more vegetated, organic-rich uplands. That's a simple model. In considering a more complex but accurate modeling of soil organic carbon spatial variability, I would want to incorporate more parameters. For instance, elevation, mean high water, the vegetation density and health all also affect carbon density in the sediment. I would create a more complex multi-variate regression model that becomes more complicated for me to explain to my audience, but the accuracy greatly improves. The more complicated model fills in the gaps of the simple single parameter model.

Interpreting a Coefficient Table

Q2. Which of the following predictor variables had slope coefficients that were significantly different from zero at a 95% confidence level? Select the correct answer(s)

A. water
B. nitrogen
C. phosphorus
D. None

You can see this based on the Pr(>|t|) values in the model coefficient table. Water and nitrogen only have significant p-values of 0.021 and 0.007. Therefore, it means these predictor variables had slope coefficients significantly different than zero at the 95% confidence interval.

Q3. Biomass Calculation 1

Using the information in the model coefficient table above, calculate the expected biomass for a plant given:

- 0 mL water per week
- 0 mg nitrogen per week
- 0 mg phosphorus per week

Explain how you made the calculation.

-1.7 grams of biomass growth.

-1.7 + (0*0.043) + (0*0.192) + (0*-0.027) = -1.7 g

We multiply 0 by all of the coefficient estimate values since this scenario says we give 0 treatments of water, nitrogen and phosphorous. We start with the base case/intercept.

Q4. Biomass Calculation 2

Using the information in the model coefficient table above, what is the expected biomass for a plant given:

- 10 mL water per week
- 30 mg nitrogen per week
- 20 mg phosphorus per week

Explain how you made the calculation.

-1.7 + (10*0.043) + (30*0.192) + (20*-0.027)

= -1.7 + 0.43 + 5.76 – 0.54 = **3.95 grams of biomass growth**

We multiply 10*the coefficient estimate of water, 30*nitrogen’s coefficient and 20*phosphorous’ coefficient based on the scenario given.

Q5. Describe the key difference between a simple linear regression and a 1-way analysis of variance. Consider the data types/scales of the predictor and response variables.

Simple Linear Regression vs. 1-way ANOVA:

In both simple linear regression and 1-way ANOVA, the dependent variable (or the response variable) is a continuous one, but in the ANOVA analysis the independent variable (predictor variable) can only be a categorical variable. In the simple linear regression, either categorical or continuous predictors can be used as independent variables. Thus, ANOVA can be considered as a case of a simple linear regression in which all predictors are strictly categorical.

We often present the equation for a simple linear regression model as: **$y_i = \alpha + \beta_1 x_i + \epsilon$**

Q6. Identify the deterministic component(s) of the model equation.

The deterministic component is a linear function of the unknown regression coefficients which need to be estimated so that the model "best" describes the data.

$y_i = \alpha + \beta_1 x_i$ is the deterministic component then. α is the y-intercept, β_1 is the slope, x_i is the predictor/independent variable, and y_i is the response variable.

Q7. Identify the stochastic component(s) of the model equation.

ϵ is the stochastic component of the model equation. This relates to the errors, or the residual terms represented by the difference between the predicted values and the observed values of the data.