Bonnie Turek

11/1/2021

Eco 602 – Week 11 Reading Questions

Frequentist Linear Models Nov 1 – 5

Q1. Model Selection: Why would we want a model selection criterion to penalize the number of parameters in a model?

The AIC (Akaike information criterion) is a model selection criterion that chooses the best-fit model based on the maximum likelihood estimate (how well the model reproduces the data) and models that have the fewest number of independent variables/parameters. AIC scores with fewer parameters and an overall lower AIC score typically perform better. This is because we would rather require less information (in the form of independent variable parameters) to predict in our model. Using more parameters has potential to slightly increase the precision of a particular model, however this slight increase in precision could also have happened by chance. The formula for AIC is as follows:

$AIC = 2K - 2\ln(L)$ where K is the number of parameters and L is the log likelihood estimate.

Adding more and more parameters may actually be redundant and unnecessary, plus it could just make the model more complicated to understand and explain. With fewer parameters, you would also have less measurement error and stochastic uncertainties to worry about. When you add more parameters, you inherently can be introducing more error into your model.

Q2. Interpreting a Slope: Consider the regression equation for a simple linear regression:

$y_i=\alpha+\beta 1 x_i+\epsilon$

In 2 - 3 paragraphs, describe the meaning of the slope parameter β1 in the context of the relationship between the predictor variable, x, and the response variable y. Your answer must be in plain non-technical language. Your explanation will be most effective if you use a narrative approach, using a concrete example to illustrate the concept.

For each 1 – unit change in the predictor variable, x, we would expect a β1 change in the value of y, on average. As an example, let's say data were collected on the depth of a dive of penguins and the duration of the dive. The following linear model is a fairly good summary of the data, where t is the duration of the dive in minutes and d is the depth of the dive in yards. The equation for the model is:

$d = 0.015 + 2.915t$

Here, the predictor variable is t, the duration of the dive, and the response variable is d, the depth of the dive of the penguins. According to the above model equation, for each one unit change in the predictor, (duration of dive in seconds) so adding 1 additional second to the dive time, we can expect a 2.915 feet change in the depth of the dive. The 2.915 value represents the β1 slope parameter. The β1 parameter helps us understand the magnitude and direction of the relationship between the predictor variable and

the response variable. In this case, the relationship between the predictive duration of the dive and the response of depth of the dive has a magnitude of 2.915. On average, we would expect a penguin adding 1 additional second to the dive time would result in a 2.915 ft deeper dive.

Interpreting a Coefficient Table

Q3. Base Case: What is the base case water treatment?

Low water level treatment is the base case and the intercept here. The intercept coefficient corresponds to the base case.

Q4. Low Water Treatment: What is the average plant mass, in grams, for the low water treatment? How did you calculate this quantity?

2.4 grams is the mean plant mass. This value was pulled directly from the coefficient table since the low water treatment level is the intercept. We do not have to do any calculations with the base case.

Q5. Medium Water Treatment: What is the average plant mass, in grams, for the medium water treatment? How did you calculate this quantity?

2.4 + (1 * 1.3) + (0 * 13.6) = 3.7 grams is the average plant mass for the medium water treatment level. We calculate this based off of the intercept or base case mean of 2.4 for the low water treatment level. We use the estimate values in the coefficient table to obtain the means of the other treatment levels.

Q6. Coefficient Interpretation: Which of the following questions cannot be addressed with the model coefficient table? Select the correct answer or answers:

A. Is there a positive relationship between increased water availability and plant biomass accumulation?
B. Is water availability a significant predictor for plant biomass accumulation?
C. What is the average biomass of plants in the high water treatment?

Model coefficient tables characterize the strength and significance of individual intercept and slope coefficients. It does not tell us about the overall significance of the categorical predictor. Therefore, B cannot be addressed. Neither the model coefficient table nor the ANOVA table tell us whether a particular pair of factor levels are significantly different form one another. This is where post-hoc testing comes into play in order to answer B.

*I worked independently on these Reading Questions