**Discussion 8: Regression**

1. What does the dependent (y) variable measure? What do the independent (x) variables measure? What does the regression constant measure?

The independent or x-variable(s) are the variables that explain or predict the value of the dependent variable depending on the values they contain.

The dependent or y-variable measures the effects that the inputted x-variables have on it. The y-variable’s value “depends” on the x-variables’ values.

1. What does the R^2 (r-squared) coefficient measure? Why is it important?

The r-squared coefficient measures amount of the variation of the independent variable that can be attributed to the dependent variable(s). This value is important because provides a measure of how well the model’s predictions replicates observed values.

1. What is a p-value, the most common p-value of interest, and why does it matter if a variable’s p-value is below that common value of interest?

The p-value is the probability of obtaining the observed results provided that the null hypothesis (that the variable of interest’s coefficient = 0) is true. The most common p-value of interest is 0.05 – namely, when the p-value is less than 0.05 we can feel safe enough to reject the null hypothesis and assume that our variable we are examining has a non-zero coefficient.

1. How do we know when variables are insignificant? If insignificant, what does that mean for our interpretation of the variable’s coefficients?

We know that a variable is insignificant when the variable’s p-value is >= .05. When this happens, we interpret the variable as having no effect on the dependent variable and thus assume that we do not have enough evidence to overturn our assumption that the variable’s true coefficient is 0.

1. Interpret the coefficients (marked Coef.) being careful to consider their p-values as well. If you are unfamiliar with this regression output; the y-variable is “lprice”, the x-variables are “y81, ldist, y81ldist”, and the regression coefficient is “\_cons”.

Table

Description automatically generated

*(source: https://i.stack.imgur.com/phgkm.png)*

“y81” has a p-value of .989 so we assume that the “y81” has a true coefficient of 0 and thus no effect on the dependent variable “lprice”. Similiarly, “y81ldist” has a p-value of 0.556 and thus we cannot reject the null hypothesis that the true coefficient of “y81ldist” is 0. “ldist” has a p-value of 0, so we reject the null hypothesis and assume that “ldist” has a coefficient of 0.317 (rounded) and thus a one unit increase of “ldist” indicates that “lprice” goes up by .317. The regression coefficient also has a p-value of 0, so we can reject the null hypothesis that its true value is 0 and thus assume the value given by the model which is 8.06 (rounded). This means that when all of the significant dependent variables are set to 0 (which in this case is just “ldist”), that the value of “lprice” is 8.06. So essentially, the model suggests that our true model given the data we have can be measured as: lprice = .317\*ldist + 8.06

1. What is overfitting? How do we know if our model is overfitting (hint: think of training and testing accuracy)?

Overfitting is a common error when a trained model too closely aligns itself to a limited amount of features and/or data points and becomes too sensitive to the random variation that occurs from that collection of data. An easy way to see if your model has succumbed to this error is by checking if the training accuracy is much higher than the testing accuracy, as your model has not been tuned to the testing data and can thus be used as sort of a blind test on how the model works.

1. Which of the following models would you feel the most comfortable using and why?

|  |  |  |
| --- | --- | --- |
| Model | Training Accuracy | Testing Accuracy |
| A | 55% | 85% |
| B | 70% | 65% |
| C | 95% | 55% |
| D | 50% | 50% |

Model B provides the best tradeoff between training and testing accuracy (they are virtually identical) while also having both training and testing accuracy above Model D. Model C - while having a great training accuracy – is incredibly overfitted to the training data (training accuracy is much higher than testing accuracy) . Model A despite having a high testing accuracy on the unknown data to the model, has a very low training accuracy so the likely explanation is that the testing accuracy occurred by chance.

1. What is the difference between Lasso Regression and Linear Regression?

Lasso regression is a modified version of linear regression which penalizes model complexity (having too many dependent variables as explanation for the independent variable) by adding a cost for every feature and often simplifying the model by having the unnecessary features have coefficients of 0, or put simply – having no impact on the dependent variable. In short, lasso regression prefers a good but simple model whereas a linear regression may be more partial to a more complex model that is more accurate. Having more parameters can lead to overfitting so using lasso regression can be a way to try to avoid overfitting.

1. Interpret the regression output below including the R^2 coefficient as well as the dependent variable’s and regression constant’s coefficients and p-values.

Table

Description automatically generated

*(source: https://www.academicianhelp.com/ckfinder/userfiles/files/fig1(2).png)*

Initially, we see that our model has an r-squared value of .975 which indicates that our model can explain 97.5% of the variation of y if given the explanatory variables of x1, x2 and x3. However, we can see that “x3” does not have significant evidence to overturn the null hypothesis that x3 has a true coefficient of 0 (p-value of .173 >= .05) and thus x3 does not show significant evidence of being predictive of the value of y. x1, x2 and the regression coefficient have p-values < .05 and thus we can assume they have a non-zero impact on y with our model’s suggested coefficients. If x1 and x2 are zero (the significant variables), our model suggests that y takes on the value of 25.746 because of the regression coefficient. Ultimately, we can explain 97.5% of the variation of y given x1 and x2 with the regression equation of: y = .929\*x1 – 2.337\*x2 + 25.746