**MGS 8040 Data Mining**

**Review of Regression Analysis – Multiple Regression**

**Dr. Satish Nargundkar**

Consider the Simple Regression example of predicting the Shoe Size of children based on their age. We saw that while Age was shown to be related to shoe size beyond any reasonable doubt, the model that predicted shoe size still had enough error that one might want to improve on it. What can be done to reduce prediction errors? One typical thought is that we take a larger sample size. Remember, though, that no matter how large a sample we take, we cannot alter the relationship between two variables. We can merely be more confident about it being there. Since in the simple regression we already were very confident that the relationship was true, increasing sample size will not help improve prediction. Age only accounted for about 64% of the variation in Shoe Size. What is important is to discover what explains the remaining variation in Y. In other words, we need other independent variables that might be related to shoe size. What might those be?

Consider the revised dataset in the table below with information on other variables for the same children. Their age and shoe size information is the same as before.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Shoe Size** | **Age** | **Weight** | **Sex** | **IQ Score** |
| 5 | 11 | 75 | 0 | 100 |
| 6 | 12 | 85 | 1 | 80 |
| 5 | 12 | 88 | 0 | 50 |
| 7.5 | 13 | 135 | 1 | 120 |
| 6 | 13 | 80 | 0 | 115 |
| 8.5 | 13 | 180 | 0 | 106 |
| 8 | 14 | 140 | 0 | 96 |
| 10 | 15 | 200 | 0 | 88 |
| 7 | 15 | 110 | 0 | 78 |
| 8 | 17 | 120 | 0 | 65 |
| 11 | 18 | 150 | 1 | 101 |
| 8 | 18 | 125 | 0 | 105 |
| 11 | 19 | 165 | 1 | 130 |
|  |  |  | Female=0 |  |
|  |  |  | Male=1 |  |

Is it reasonable to suppose that these other variables might have something to do with Shoe Size? Who decides what data to collect? How does one make that decision?

For learning purposes, it made sense to show simple regression first, with only one independent variable, age, in order to keep it easy to understand. However, in reality it would make sense to think of all possible things that might possibly explain the variation in Y, and proceed to collect data on all variables that we can.

Given the data above, what is the next step? In general, when you have many variables, it is important before looking at their effect on Y simultaneously to first examine each variable individually to check whether the distributions are as expected, look for outliers, missing data, or other possible errors. Second, it may help to draw scatter plots of Y against each X variable to see if there appears to be a relationship in each case. This helps make sure that we have clean data. The final step is to perform *Multiple Regression*, where we model the relationship of Y with all the X variables taken together. Consider the scatter plot of Shoe Size and Weight below:

As expected, the relationship is a positive one, and the range of weighs is plausible, indicating that no obvious mistakes were made in recording the data. Note that one of the other variables, Sex, is a categorical one, and has been coded as 0 for Females and 1 for Males. This sort of variable is called an **Indicator** or **Dummy** variable. This will be discussed more a little later. Here, it is important to note that a scatter plot will not make sense for Shoe Size and Sex. What kind of chart would you draw?

**Multiple Regression**: To proceed with Multiple Regression, the process in Microsoft Excel is the same (Data/Data Analysis/Regression), except that when selecting the range for X, select the range that includes all the X variables. The output of Multiple Regression for the above data with four Xs is shown below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT |  |  |  |  |  |
| *Regression Statistics* |  |  |  |  |  |
| Multiple R | 0.9805340 |  |  |  |  |  |
| R Square | **0.9614469** |  |  |  |  |  |
| Adjusted R Square | 0.9421703 |  |  | This is in scientific notation. 1.07054E-05 = 0.0000107054 |  |  |
| Standard Error | **0.4849857** |  |  |  |  |  |
| Observations | 13 |  |  |  |  |  |
|  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |
| Regression | 4 | 46.926004 | 11.731501 | 49.876479 | **1.07054E-05** |  |
| Residual | 8 | 1.881689 | 0.235211 |  |  |  |
| Total | 12 | 48.807692 |   |   |   |  |
|  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
| Intercept | **-1.774343** | 0.929792 | -1.908321 | 0.092771 | -3.918449 | 0.369764 |
| Age | **0.346261** | 0.062326 | 5.555598 | **0.000537** | 0.202536 | 0.489986 |
| Weight | **0.030380** | 0.004180 | 7.267505 | **0.000087** | 0.020741 | 0.040020 |
| Sex | **0.791891** | 0.322696 | 2.453982 | **0.039689** | 0.047751 | 1.536030 |
| IQ Score | **0.003963** | 0.007147 | 0.554522 | **0.594381** | -0.012518 | 0.020445 |

As with simple regression, we start by looking at the Significance F, and note that it is very small, indicating that we can be very confident that one or more of these X variables is in fact related to Y. The next step is to check whether each of the variables is significant in the model, and to do that, we look at the p-values for each coefficient estimate. The p-value for IQ Score is 0.5943, which is rather large, indicating that the coefficient for that variable cannot be clearly distinguished from 0. In other words, IQ Score shows no significant relationship with Shoe Size. At this point, instead of writing the model with the other numbers, it is necessary to drop the variable IQ Score, and rerun the regression with the remaining variables.

**Revised Multiple Regression**: The output of the regression with IQ Score removed is shown below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT |  |  |  |  |  |
| *Regression Statistics* |  |  |  |  |
| Multiple R | 0.9797780 |  |  |  |  |
| R Square | **0.9599650** |  |  |  |  |
| Adjusted R Square | 0.9466200 |  |  |  |  |
| Standard Error | **0.4659536** |  |  |  |  |
| Observations | 13 |  |  |  |  |
|  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |
| Regression | 3 | 46.853678 | 15.617893 | 71.934479 | **1.30775E-06** |
| Residual | 9 | 1.954015 | 0.217113 |  |  |
| Total | 12 | 48.807692 |   |   |   |
|  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* |
| Intercept | **-1.5104494** | 0.7674279 | -1.9681973 | 0.0805754 | -3.2464931 |
| Age | **0.3474510** | 0.0598451 | 5.8058406 | **0.0002576** | 0.2120719 |
| Weight | **0.0309663** | 0.0038858 | 7.9690334 | **0.0000228** | 0.0221759 |
| Sex | **0.8582171** | 0.2879489 | 2.9804490 | **0.0154383** | 0.2068309 |
|  |  |  |  |  |  |
| RESIDUAL OUTPUT |  |  |
|  |  |  |
| *Observation* | *Predicted Shoe Size* | *Residuals* |
| 1 | 4.6339836 | 0.3660164 |
| 2 | 6.1493147 | -0.1493147 |
| 3 | 5.3839964 | -0.3839964 |
| 4 | 8.0450804 | -0.5450804 |
| 5 | 5.4837171 | 0.5162829 |
| 6 | 8.5803466 | -0.0803466 |
| 7 | 7.6891458 | 0.3108542 |
| 8 | 9.8945745 | 0.1054255 |
| 9 | 7.1076079 | -0.1076079 |
| 10 | 8.1121728 | -0.1121728 |
| 11 | 10.2468298 | 0.7531702 |
| 12 | 8.6144553 | -0.6144553 |
| 13 | 11.0587752 | -0.0587752 |

Is the Regression significant? What is the model? What do the coefficients mean? How much of the variation in Y does it explain?

**Interpretation**: It is once again clear from the Significance F value that the regression overall is meaningful. The p-values for each variable’s coefficients are also small (<0.05), indicating that each is significantly related to Y in the presence of the other variables.

The model is written as an additive function as follows:

ŷ = b0 + b1X1 + b1X1 + b2X2 + b3X3.

Here, the model can be written out as

Ŷ = -1.51 + 0.347(Age) + 0.031(Weight) + 0.858(Sex)

How do we interpret the coefficient 0.347 for Age? Recall that when Age alone was used in our simple regression, the coefficient was 0.612, meaning that for each year of age, shoe size went up on average by 0.612 units. How can the impact of the same variable on the same data be different now? The answer is that now, the coefficient of 0.347 is the impact of Age on Shoe Size given that Weight and Sex are held constant. In other words, if we were to only compare Males of different ages, all weighing 100 pounds, then for each year of age, shoe size would increase by 0.347 units on average.

How does one interpret the coefficient for Sex, the categorical variable? Assume you are predicting the shoe size for a 15 year old Male weighing 100 pounds, and a 15 year old Female weighing 100 pounds. If you substitute 15 for Age, 100 for weight, and either 1 or 0 respectively for Sex, you will see that the predicted value for Males is higher than for Females by 0.858. Thus, the coefficient for Sex tells us that Males on average wear a shoe size that is 0.858 units larger than Females, assuming age and weight are held constant.

Notice that the R-square value is about 0.96, indicating that 96% of the variability in Y is explained by the three variables in our model. This is a big improvement over the 64% or so explained by age alone. Also, Standard Error is down to 0.46 from 1.26 in the Simple Regression with Age. This means that the predictions will be much more accurate.

**More on Dummy Variables**: Whenever we want to use a categorical independent variable in our regression model, it must be converted into dummy variables. We need one less dummy variable than the number of categories. Thus, for Sex, with two categories, we only need one dummy with the values 0 and 1. It does not matter which category is coded as 1 and which one is 0. Try running the same regression analysis as above, with the Sex variable coded in reverse, and look at the model and its predictions. If there are 3 categories, say for a variable like hair color (assume Black, Brown, and Blonde are the only colors), we will need two dummy variables, say one for Black Hair and one for Brown. The dummy for Black hair has a value of 1 if the person has Black Hair, 0 otherwise. Similarly, the dummy for Brown Hair has a value of 1 if the person has Brown hair, 0 otherwise. The person with Blond hair, therefore, will have a value of 0 for both dummies, and becomes the baseline against which to compare the results.

Consider the example in the table below:

|  |  |  |  |
| --- | --- | --- | --- |
| Obs | Color of Hair  | Black\_Dummy | Brown\_Dummy |
| 1 | Black | 1 | 0 |
| 2 | Blonde | 0 | 0 |
| 3 | Black | 1 | 0 |
| 4 | Brown | 0 | 1 |
| 5 | Brown | 0 | 1 |
| 6 | Blonde | 0 | 0 |
| 7 | …. | …. | …. |

As we see in the table above, the original categorical variable with 3 values is converted into 2 numerical variables, each a dummy with only 0 and 1 as the possible values. These two dummies can now be used in regression as independent variables to model the impact of color on the dependent variable y. It is important to note that it is not correct to simply code the 3 colors as 1, 2, and 3 in a single variable. It will give misleading results. As for the dummies, however, you can choose to make dummies for any 2 of the 3 categories. The other category becomes the baseline or neutral group against which results are interpreted.

Assume that a regression was performed, and the model was as follows:

Ŷ = 10 + 5 (Black\_Dummy) + 3 (Brown\_Dummy)

How do we interpret the coefficients? The coefficient 5 implies that the predicted value of y is 5 units greater on average for Black haired people compared with Blond haired ones (the neutral group). Similarly, the predicted value of y is 3 units greater on average for Brown haired people compared with Blond haired ones (the neutral group). In both cases, the interpretation is with respect to the neutral group. What if the coefficient for Brown\_Dummy turned out to be not significant, and we had to drop that variable from the model? In that case, it would mean that Brown haired people were no different from Blond haired ones, but the two groups together are still different from the Black haired group, in terms of the impact on y.

**Limitations of Regression**

Regression Analysis assumes that the model is linear, and that the errors are normally distributed, independent, and that they have a constant variance over the time and over the range of values of the independent variables. These assumptions may be violated in reality leading to less accurate or misleading predictions.