**MGS 8040 Data Mining**

**Review of Regression Analysis**

**Dr. Satish Nargundkar**

**Introduction**: There are many situations in all walks of life where we are aware that some things are related to others, or we simply suspect that they might be a relationship, without necessarily knowing the relationship exactly. For instance, the amount of exercise one does perhaps affects lung capacity, age of a child is related to her height and weight, job satisfaction is related to absenteeism, motivation may be related to the type of reward, a customer’s likelihood of repaying a loan may be related to income as well as past payment behavior, and a person’s overall happiness may be linked to certain genetic material in his DNA. In most cases, knowing what the relationship is like precisely and how strong it is can help with decision making. Regression analysis is a mathematical technique to model the relationship between a dependent variable (Y) and one or more independent variables (Xs). The dependent is also referred to as the predicted variable, and the independent variables as the predictors, since we can predict the value of Y for a given set of Xs, given the knowledge of the Xs and their relationship with Y.

We will begin the discussion with Simple Linear Regression, where a single independent variable and its relationship with a dependent is modeled as a straight line. Both variables are continuous numeric variables.

**Simple Regression**

Consider the following (fictitious) data about the ages of some teens and their shoe sizes.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  | | | | | | |
| **Obs** | **Age** | **Shoe Size** |  |
| 1 | 11 | 5.0 |  |
| 2 | 12 | 6.0 |  |
| 3 | 12 | 5.0 |  |
| 4 | 13 | 7.5 |  |
| 5 | 13 | 6.0 |  |
| 6 | 13 | 8.5 |  |
| 7 | 14 | 8.0 |  |
| 8 | 15 | 10.0 |  |
| 9 | 15 | 7.0 |  |
| 10 | 17 | 8.0 |  |
| 11 | 18 | 11.0 |  |
| 12 | 18 | 8.0 |  |
| 13 | 19 | 11.0 |  |
|  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |

Does it appear that there is a relationship between Age and Shoe Size?

A glance at the scatter plot shows some random variation about a linear relationship. In simple terms, as age goes up, shoe size goes up. It is important to realize that this relationship is meaningful only for the range of numbers in the sample. In other words, we do not expect this growth in shoe sizes to remain the same as age keeps on going up.

Can we determine the mathematical relationship between Shoe Size and Age for our data? A simple linear regression model is of the form

Y = b0 + b1X + error (1),

where b0 and b1 are the estimates for the intercept and slope of the straight line, respectively.

Once again, glancing at the scatter plot shows that one may visualize several plausible straight lines at various angles that represent the relationship between X and Y. Since the points are not all in a straight line, any straight line model, if used to predict Y using X, will result is some error. Practically speaking, one can say that the error is because Shoe Size does not solely depend on Age alone. There are other factors that we have not considered here. The question at this point is, which of the infinite possible straight lines that we might draw through the points will result in the least (squared) error overall? Writing the equation for the sum of squared errors and using elementary calculus, one can find the values of b0 and b1 that minimize the error, and guarantee the “best fitting” line through those points. You can refer to any basic statistics book for the formulas. Here, we will use Excel to draw the line on the chart, give us the equation, as well as a more detailed statistical output.

Using the **Chart Layout/Trendline** option in Excel gives the following result.

The equation to predict y is more correctly written as Ŷ = -1.1759 + 0.612X, where Ŷ represents the predicted value (as opposed to actual value) of Shoe Size, and X is the Age. This line, among all other possible lines, is guaranteed to have the least overall squared error value, and is called the regression line. How is Ŷ different from Y? What is Ŷ when X equals 15?

**Detailed Output**: A more detailed look at the statistical output can be obtained from Excel (2007) under **Data/Data Analysis/ Regression**. If you do not see the Data Analysis option in under Data, you will have to go to add-ins and add the Analysis Tool Pack. The output is shown below.

R-square indicates the portion of variation in Y that is explained by X. Here, 63.7% of the variations in shoe size can be attributed to the variations in ages of the children.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT |  |  |  |  |  |  |
| *Regression Statistics* | |  |  |  |  |  |
| Multiple R | 0.798498 |  |  |  |  |  |
| R Square | **0.637599** |  |  |  |  |  |
| Adjusted R Square | 0.604653 | A standardized measure of the overall error in this regression. |  |  |  |  |
| Standard Error | **1.268068** |  |  |  |  |  |
| Observations | 13 |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |
| Regression | 1 | 31.11973 | 31.11973 | 19.35311 | **0.00106** |  |
| Residual | 11 | 17.68796 | 1.60800 |  |  |  |
| Total | 12 | 48.80769 |  |  |  |  |
|  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
| Intercept | **-1.1759** | 2.0635 | -0.5699 | 0.5802 | -5.7178 | 3.3659 |
| Age | **0.6120** | 0.1391 | 4.3992 | 0.0011 | 0.3058 | 0.9182 |
|  |  |  |  |  |  |  |
| RESIDUAL OUTPUT |  |  | The estimates b0 and b1 for intercept and slope, which form the prediction equation. |  |  |  |
|  |  |  |  |  |  |  |
| *Observation* | *Predicted Shoe Size* | *Residuals* |  |  |  |  |
| 1 | 5.55648 | -0.55648 |  |  |  |  |
| 2 | 6.16852 | -0.16852 |  |  |  |  |
| 3 | 6.16852 | -1.16852 |  |  |  |  |
| 4 | 6.78056 | 0.71944 |  | This is the p-value for the regression. If we assume that X and Y are unrelated in the population (regression coefficient of 0), what is the probability that we would pick a sample from such a population which would show a relationship like the one we see here (coefficient of 0.612 or bigger)? The answer is 0.00106. |  |  |
| 5 | 6.78056 | -0.78056 |  |  |  |  |
| 6 | 6.78056 | 1.71944 |  |  |  |  |
| 7 | 7.39259 | 0.60741 |  |  |  |  |
| 8 | 8.00463 | 1.99537 |  |  |  |  |
| 9 | 8.00463 | -1.00463 |  |  |  |  |
| 10 | 9.22870 | -1.22870 |  |  |  |  |
| 11 | 9.84074 | 1.15926 |  |  |  |  |
| 12 | 9.84074 | -1.84074 |  |  |  |  |
| 13 | 10.45278 | 0.54722 |  |  |  |  |

**Interpretation**

Three basic questions must be answered for each regression.

1. Is the regression meaningful at all? Are X and Y really related to each other?
2. What is the model that describes the relationship?
3. How good is the model for predicting Y?

Each of these questions can be answered by looking at different parts of the output. Each is described below.

**Checking if relationship exists**: The first question is answered in the regression output by the **Significance F** value. **The smaller it is the better**. It indicates the likelihood of getting a sample like the one we got from a population where there is in fact no relationship between X and Y. Imagine that we live in a universe where shoe size has nothing to do with the age of a child. If that were true for the entire population of children, would it still be possible to pick a random sample of 13 children like we did that shows the linear relationship? The answer is of course YES, since by sheer luck, the 13 points in the sample may arrange themselves in a line even though the larger population may not. However, that would be very unlikely to happen. How unlikely? For this regression, the answer is that the probability is 0.00106, or about a tenth of a percent.

So we can conclude that there is a high probability that what we see in our sample did not simply happen by chance, that instead it reflects a true relationship between X and Y in the population. The larger the Significance F gets, the more likely that the relationship does not really exist. **How small should the Significance F be** for us to confidently proclaim a relationship between X and Y? That depends on our tolerance for error in our conclusion. The typical **default values used by statisticians tend to be 0.05 or 0.01 (Sig. F less than these numbers)**. In other words, we want to be at least 95% or in some cases at least 99% sure about it in order to conclude that X and Y are related, and the regression is meaningful.

In this example, would you say that you are at least 95% confident that Age is related to Shoe Size? How about 99% confident?

**The model:** If the regression is meaningful based on the Significance F, the next step is to write down the regression model. The parameter estimates b0 and b1 are shown as the **coefficients for the intercept and the X variable in the output**. As from the graph, we can look at this output and write out the equation as

Ŷ = -1.1759 + 0.612X.

The intercept of -1.1759 indicates the point at which the regression line crosses the Y axis (when X is 0). If the scatter plot above were rescaled to show the 0 on the X axis, you would see the intercept.

What is the practical interpretation of the intercept? It could be interpreted as the shoe size that a child of age 0 would wear. However, children ages 0 to 10 are not part of the sample, so the line cannot truly be extended to that point for prediction purposes, since the growth rate for children from 0 to 10 could be very different from 10 to 20. So **in this case, the intercept has no practical interpretation**. It is simply a mathematical artifact that is necessary for prediction within the range of Xs that the model is valid for.

(A further point in this regression is that the p-value for the intercept is 0.5802, which indicates that it is not significantly different from 0. The P-values are interpreted just like the Significance F. One could run the regression again and force the intercept to be 0, if desired. In any case, even if it were significant, it would have no practical interpretation.)

What about the coefficient for Age (the slope)? The value 0.612 in the equation indicates that **on average, for each year of age, shoe sizes grow by 0.612 units**. Test is out as follows – what is the predicted shoe size for a 14 year old according to our equation? How about for a 15 year old? What is the difference between the two?

**Evaluation of the model**: We now know from Significance F that X and Y are related, and we also know the equation that shows the relationship and can use it to predict Y. How good is the prediction, though? Just because a relationship exists does not mean it is a very strong one and will provide good predictions. We need a measure of how well the model will perform.

**Computing Standard Error:** Note that the last part of the output shows the predicted values for each of the X values in the sample, and the error in prediction, computed as the difference between the actual Y in our sample and the value predicted by the equation.

Error = Y – Ŷ

On the graph, Error is the vertical distance between each of the original data points and the regression line. The sum of all the errors would be 0, since positive and negative errors cancel each other. This is always true with regression, since the regression line creates unbiased errors (they are equal in the positive and negative direction). Squared errors are used instead to evaluate the model.

|  |  |  |  |
| --- | --- | --- | --- |
| *Obs* | *Predicted*  *Shoe Size* | *Residuals* | *Squared*  *Residuals* |
| 1 | 5.55648 | -0.55648 | 0.30967 |
| 2 | 6.16852 | -0.16852 | 0.02840 |
| 3 | 6.16852 | -1.16852 | 1.36544  This is the *Sum of Squared Errors* (residuals), shown under the **SS** **column** and **Residual** **Row** of the ANOVA table in the output. |
| 4 | 6.78056 | 0.71944 | 0.51760 |
| 5 | 6.78056 | -0.78056 | 0.60927 |
| 6 | 6.78056 | 1.71944 | 2.95649 |
| 7 | 7.39259 | 0.60741 | 0.36894 |
| 8 | 8.00463 | 1.99537 | 3.98150 |
| 9 | 8.00463 | -1.00463 | 1.00928 |
| 10 | 9.22870 | -1.22870 | 1.50971 |
| 11 | 9.84074 | 1.15926 | 1.34388 |
| 12 | 9.84074 | -1.84074 | 3.38833 |
| 13 | 10.45278 | 0.54722 | 0.29945 |
|  |  |  | **17.68796** |

The Sum of the squared residuals or errors (SSE) is can be used to compute the mean squared error. Rather than divide the number by 13 for the 13 data points we have, we divide it by the degrees of freedom for error, which in this case is 11. The reason is that if we had only 2 points, the regression line would have had no errors, no matter what two points we take, since the line will go through them both. So only 11 of the 13 points are responsible for creating all the error we see. Thus, 17.68796/11 gives us a Mean Squared Error (MSE) of 1.608. This is the average error in our predictions in squared terms, and we can bring it back to normal units by taking the square root. The **square root of MSE is called the Standard Error, and is equal to 1.268 in this case**. Note that both MSE and Standard Error are also shown in the Excel output.

Standard Error is a measure for overall strength of the model in predicting Y. The smaller the standard error, the more accurate the model is at predicting Y. The standard error value can be used to compute prediction intervals around the predicted Y for a given X.

**Computing R-squared**: Imagine that we only knew the Shoe Sizes (Y) of the 13 children and had no information about the corresponding age of each child. If we were asked to predict the shoe size of a child based on that alone, the best guess we could make would be the mean of the 13 numbers. What would be the Sum of Squared Errors then? This can be computed easily by looking at the deviations of each number from the mean, as shown in the table below.

|  |  |  |
| --- | --- | --- |
|  | **Deviation** | **Squared** |
| **Shoe Size (Y)** | **From Mean** | **Deviations** |
| 5 | -2.76923 | 7.66864 |
| 6 | -1.76923 | 3.13018 |
| 5 | -2.76923 | 7.66864 |
| 7.5 | -0.26923 | 0.07249 |
| 6 | -1.76923 | 3.13018  This is the *Sum of Squared Total* , shown under the **SS** **column** and **Total** **Row** of the ANOVA table in the output. |
| 8.5 | 0.73077 | 0.53402 |
| 8 | 0.23077 | 0.05325 |
| 10 | 2.23077 | 4.97633 |
| 7 | -0.76923 | 0.59172 |
| 8 | 0.23077 | 0.05325 |
| 11 | 3.23077 | 10.43787 |
| 8 | 0.23077 | 0.05325 |
| 11 | 3.23077 | 10.43787 |
| **7.76923** |  | **48.80769** |

The *Sum of Squares Total* value of 48.80769 shows the squared error that we would have if every prediction was equal to the mean value, since there was no information about the age. However, we do know from the *Sum of Squared Error* value what the error would be with the regression line (17.68796). This means that using age to predict shoe size reduced the error in predictions quite a bit over simply using the mean shoe size. Specifically, the difference in *Sum of Squares due to Regression (SSR)* is 48.80769 – 17.68796 = 31.11973. Thus we can say that regression helped us reduce the sum of squares by 31.11973 from the overall value of 48.80769. This reduction can be expressed as a proportion of the overall value, computed as 31.11973 / 48.80769 = **0.637599.** This proportion is called the **R-square** for the regression. Practically speaking, we can say that 0.637599, or about 63.76% of the variation in shoe sizes can be explained as a result of the variation in ages of the children. Put another way, ***Age* accounts for 63.76% of the variations in *Shoe Sizes* of children**.

Other Details of the ANOVA table: The ANOVA table in the output shows the SSR value along with SSE and SST. The next column shows the Mean Squared Regression (MSR) and Mean Squared Error (residual) MSE. Since all the reduction of error (31.119) has been achieved with a single variable (Age), the average explanation per independent variable is the same as the sum (Sum divided by 1 degree of freedom). The ratio MSR/MSE is called the F statistic. MSR is an indicator of the average variation in Y explained by our model, while MSE is the average error that still exists (variation due to causes still unknown to us, or unexplained by regression). Thus the bigger the F, the more the model has explained of Y, which means greater confidence that X is in fact truly related to Y. In other words, bigger the F, the less likely that it came about through random chance from a population with no relationship between X and Y. This likelihood is shown as Significance F value of 0.00106.

**Questions**

1. What possible values can R-square take on?
2. Draw a sketch to show what the data might look like when R-square is 0 and when it is 1.
3. What is the 95% prediction interval for the shoe size of a 15 year old?
4. What if the above answer is unsatisfactory? In other words, Age is related to Shoe Size, but the predictions are not accurate enough for practical use? What would you do to improve predictions?