post02-nomin-amar.html

Predicting the future?

Introduction

Do you ever wanted to predict the future? Do you ever want to predict the outcome of something unknown given some known information? This is where a very powerful statistical tool called linear regression comes in.

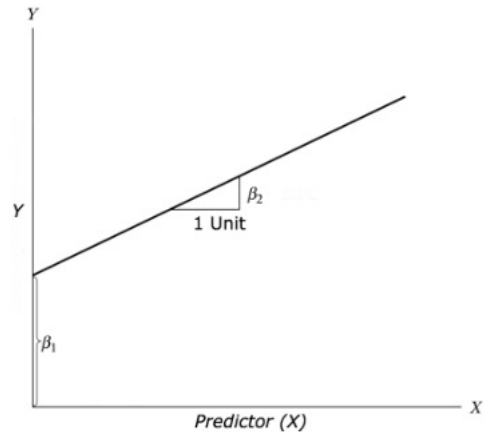
Motivation

Say your company's sales have went up drastically every month for the past few years, by using linear regression on the sales data of month sales, the company has the power to forecast sales in the future months.

Background

Linear regression lets us predict the value (say Y) of something based on one or multiple independent variables (say X). The goal is to create a linear relationship between the Y and the X. Furthermore, using this linear relationship to estimate Y given only the X. It is amazing how it can predict Y only when X values are known to a certain error. The basic layout of a predictor is as follows:

$Y = \beta_1 + \beta_2 X + \epsilon$,where b1 and b2 are coefficient which we will find using R, X correspond to the independent variable, and e is the error term. Since linear regression is a statistical method, R has nice built-in functions that makes linear regression user friendly and relatively simple than other programs.



Graph

Examples/Walk-through

The aim of this blog post is to build a simple linear regression to predict Distance given Speed. We will be using the already built-in data in R called 'cars.' Readers can simply access it by typing in cars in their R console. Before jumping into analyzing anything, few things need to be done in order to ensure that linear regression is the best method of estimation for this type of data.

A good habit to form before doing linear regression is to create a scatterplot with X as the independent variable and Y as the dependent (predicted) value to see how all the points are spread out. With our data, X would be the Speed, and Y would be the Distance. Here we have only one independent variable, but if we had more than one independent variable then, one can create scatterplot for each pair.

```
scatter.smooth(x=cars$speed, y=cars$dist, main="Dist ~ Speed")  # scatterplot
```

From looking at the scatterplot, it seems that Distance and Speed have positive strong correlation. It is now seen that X and Y have a linear relationship; hence it is suitable to use linear regression as a predictor for the Y values.

Another good habit to form is to calculate the correlation between the variables both independent and dependent. Correlation tells us information about the linearity of two variables. It can take values between -1 and 1. -1 corresponding to strong negative linear relationship. 1 corresponding to strong positive linear relationship. Values closer to 0 shows very little linear correlation; hence it would not be the best data to use linear regression on. On the other hand, when there is a strong linear correlation, it would be suitable to continue to use linear regression as predictor.

```
cor(cars$speed, cars$dist)  # calculate correlation between speed and distance
#> [1] 0.8068949
```

Calculating the correlation between Distance and Speed, we see that the correlation is 0.806, which is a strong positive linear relationship. Hence, let's continue with our linear regression model.

Building Linear Model

R has a nice built-in function called lm() that gives us a simple way to figure out the coefficients b1 and b2. This function has two input parameters, formula and data. Formula looks something like dist~speed or more generally Y value~X value(s). Data corresponds to the data frame you're working on. For our case, data=cars.

```
linearMod <- lm(dist ~ speed, data=cars)  # build linear regression model on full data
print(linearMod)
#> Call:
#> lm(formula = dist ~ speed, data = cars)
#>
#> Coefficients:
#> (Intercept)        speed
#>     -17.579        3.932
```

After running this, we find that intercept is -17.579 and speed is 3.932, where the speed corresponds to b2 and -17.579 corresponding to b1. Now that we have found our coefficients, b1 and b2, we put it back into our original formula to obtain:

$$dist = Intercept + (\beta * speed)$$

$$\Rightarrow dist = -17.579 + 3.932 * speed$$

General Equation

Now that we have a linear model, can we just simply use it to predict Distance given Speed? NO! We must make sure it is statistically significant by using a hypothesis test. How do we do this?

One can type in summary(linearMod) to obtain more information on the linear model. One of the items returned by this command is p-value. This value is very important since we can decide whether a linear model is statistically significant or not by comparing this value to significance value, which is usually 0.05.

```
summary(linearMod)  # model summary
#> Call:
#> lm(formula = dist ~ speed, data = cars)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -29.069  -9.525  -2.272   9.215  43.201
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept) -17.5791     6.7584  -2.601   0.0123 *
#> speed         3.9324     0.4155   9.464 1.49e-12 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 15.38 on 48 degrees of freedom
#> Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
#> F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

To run this hypothesis test, we need two hypotheses, a null and alternate hypothesis. Null simply states that there exists no relationship between the X and Y. Alternate simply states that there exists a relationship between the X and Y. When the p-value is less than the significance value 0.05. We can say reject the Null and say that the model is in fact statistically significant.

Your Turn

Now that we have a statistically significant model, predict distance of car with speeds, 30,35,40,45,50 using the obtained linear regression predictor

Take Away

What are some cool and useful applications of this statistical tool? How can it be revolutionary?

References

http://r-statistics.co/Linear-Regression.html#Predicting%20Linear%20Models

https://www.r-bloggers.com/r-tutorial-series-simple-linear-regression

http://analyticstraining.com/2014/popular-applications-of-linear-regression-for-businesses/

https://www.r-bloggers.com/simple-linear-regression-2/

https://onlinecourses.science.psu.edu/stat501/node/253

https://www.statmethods.net/stats/regression.html

http://www.r-tutor.com/elementary-statistics/simple-linear-regression/significance-test-linear-regression