A Peek into the Life of a Statistician

Dootika Vats

Abstract

This paper aims at introducing to the reader basic statistical analysis done by most applied statisticians. The paper introduces simple linear regression with the help of an example. The objective is to build a linear model between two variables and test the significance of the model.

Introduction

Statistics is generally understood to be a branch of mathematics concerned with collecting and interpreting data. This however, is largely untrue. Statistics, by itself should not be understood as a branch of mathematics. Mathematics is definitive, and in most cases deterministic. A problem in mathematics has one correct answer. In statistical analysis, there is no correct answer. This is what makes statistics the most flexible and sought after tool in understanding the workings of the world.

This is not to say that mathematics renders itself useless in the field of statistics. Indeed, understanding most topics in statistics requires a sound knowledge of mathematics, which is why most statisticians have an undergraduate degree in mathematics. Today, we live in a world where everything turns into data at the end of the day, and there are not enough qualified statisticians to make sense of the data. In 2009, an article in the New York Times [1] elaborated very articulately, on why being a statistician is one of the most sought after professions.

One of the most important aspect of statistics is to find out if certain things are related to each other. For example, the fact that smoking causes lung cancer was proved largely due to statistical analysis. Similarly, we might want to analyze if a certain variable, Y (response) depends on a set of variables $X_1, X_2, \ldots, X_p$ (predictors). And if there is such a dependence, then we want to find the f such that $Y = f(X_1, X_2, \ldots X_p)$. This function f is found by using Regression.

This paper focuses on Simple Linear Regression, i.e., when we have only one predictor, X, and f is a linear function, giving the relation, $Y = \beta_0 + \beta_1 X$. The concept of simple linear regression will be explained step by step with the help of an example dataset.

Dataset

A dataset is a collection of data, usually presented in tabular form, where each column represents a variable of interest. In simple linear regression, there are only two columns, one for the response variable, Y and one for the predictor variable, X. This paper uses the example of one of the most basic and famous datasets.

Karl Pearson organized the collection of data of over 1100 families in England in the period 1893 - 1898. This particular data set gives the heights in inches of mothers and their daughters. All daughters are at least age 18, and all mothers are younger than 65. The objective is to find out whether there is a relation between the height of mothers and their daughters.

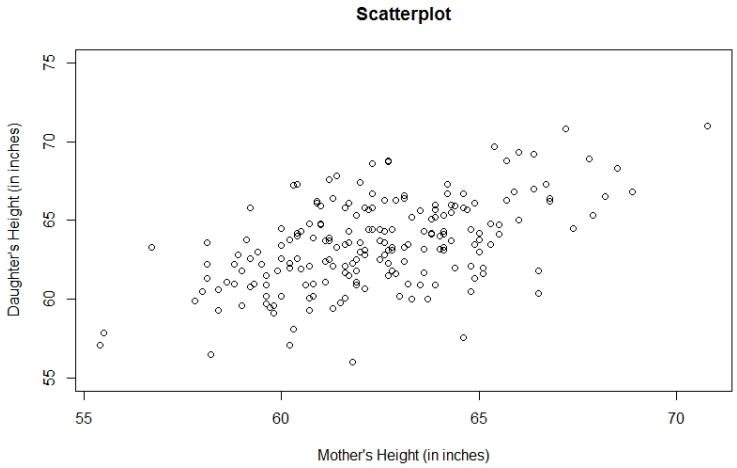
Notice how the background of the dataset is as important as the numbers in the dataset. For example, it is important to know that the daughters are atleast age 18, so we can assume that they have attained their full height.

Now, the original dataset has 1100 observations, but for the purpose of this paper, I have chosen a random subset of 200 observations from the dataset. This is just to ensure that the graphs produced at not messy. The numbers in the dataset are given in the table below.

Of course, looking at the numbers does not really help us, specially when we have 200 such pairs. This is where graphical tools prove to be much more useful.

X	Y
Mother's Height	Daughter's Height
63.5	66.0
63.5	63.2
62.7	63.0
⋮	⋮

Each of the rows above, corresponds to a **datapoint**, which is to say that each point can be written as (x_i, y_i), and this would correspond to a point on a graph of Y vs X. When all the datapoints are plotted together, we get something called a **scatterplot**. This is essentially the first tool in understanding whether there is any relation between X and Y.



From the scatterplot above, we notice that as mother's height increases, we see some increase in the daughter's height. Note that this is not individually true, but the **trend** indicates that it is generally true.

As mentioned earlier, the objective is to find a linear relationship between X and Y.

Model

If there is a linear relationship between X and Y, all points (x_i, y_i) should lie on a common line. We also know that they do not exactly lie on a common line and there is some deviation(as demonstrated in the scatterplot above). This is represented by the **model** below

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where β_0 is the y-intercept, β_1 is the slope of the line, and ϵ_i is known as the **error**. When we write this model, we make the following very important **assumptions:**

- A linear model is appropriate
- All the observations, $y_1, \ldots y_n$ are independent of each other. In this dataset, we assumed that none of the daughters are related.
- The errors $\epsilon_i \sim N(0, \sigma^2)$. This means that the errors follow a normal distribution[3] with mean 0 and variance σ^2.
- All observations have a constant variance (σ^2, as opposed to σ_i^2).

Whenever a model is fit to the dataset, it is the duty of the statistician to ensure that the model assumptions stated above hold true.

One important point to note in the model, is that the unknown quantities are β_0 and β_1, and the points (x_i, y_i) are all known, since the line is fit only after we have the data. Thus, the term "linear" refers to the equation being linear

in the βs, and not in the xs. If instead we had the equation

$$y_i = \beta_0 + \beta_1 x_i^2 + \epsilon_i$$

This would still be a linear regression model.

The next step is to try and estimate β_0 and β_1 from the data. Notice how we use the word "estimate", because every time the experiment is done, we get a different dataset, and every dataset gives a new "estimate" of these two parameters. These estimates are denoted by $\hat{\beta_0}$ and $\hat{\beta_1}$, and are calculated by using a method known as Ordinary Least Squares (OLS).

OLS - Ordinary Least Squares

Clearly, we can not fit a line by joining all the points. This is where mathematics, gives way to statistics. We fit a line to the data, in such a way that the overall deviation of the datapoints from the line is minimized. This is done by a method known as Ordinary Least Squares, or OLS.

Notice that in the model, the error ϵ_i is nothing but the deviation of each point from the line. Now let us assume that we have already fit the line, and have found $\hat{\beta_0}, \hat{\beta_1}$ and for each x_i we have a $\hat{y_i}$. Thus each $(x_i, \hat{y_i})$ lies on the line:

$$\hat{y_i} = \hat{\beta_0} + \hat{\beta_1} x_i$$

The estimated errors are known as **residuals** $= y_i - \hat{y_i}$, giving :

$$\hat{\epsilon_i} = y_i - (\hat{\beta_0} + \hat{\beta_1} x_i)$$

The OLS method estimates $\hat{\beta_0}$ and $\hat{\beta_1}$ are obtained by minimizing the sum of the squared residuals over all observations, i.e.

$$(\hat{\beta_0}, \hat{\beta_1}) = \min_{(\beta_0, \beta_1)} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

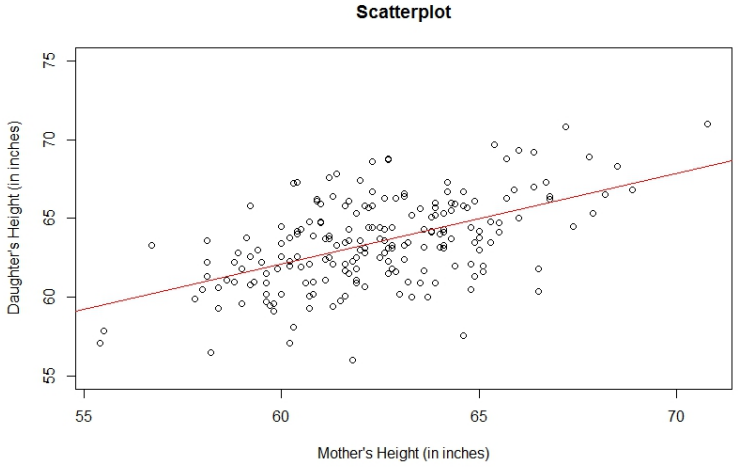
This turns out to be an exercise in basic calculus, the proof for which can be found in most statistics books [2]. The OLS estimates we get are:

$$\hat{\beta_1} = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad \hat{\beta_0} = \bar{y} - \hat{\beta_1} \bar{x}$$

where $\bar{y}$ and $\bar{x}$ are the means of y_is and x_is. $\hat{\beta_0}$ and $\hat{\beta_1}$ represent the actual y-intercept and slope of the line.

In our dataset, we get $\hat{\beta_0} = 25.05$ and $\hat{\beta_1} = 0.62$ giving the equation

$$Daughter_{height} = 25.04 + 0.58 Mother_{height}$$



Interpretation:

$\hat{\beta}_0 = 25.05$ implies that when the mother's height is 0 inches, the daughters height on average is 25.05 inches. This of course does not make sense, and that is alright. In most cases $\hat{\beta}_0$ need not be interpreted. We are more interested in $\hat{\beta}_1$. $\hat{\beta}_1 = 0.58$ implies that with one inch increase (decrease) in the mother's height, we expect the daughters height to increase(decrease) by 0.58 inches. This is what gives us the relation between mother's height and daughter's height. We need to analyze this more carefully.

Significance

The most important aspect of simple linear regression is to make sense of the $\hat{\beta}_1$. Since $\hat{\beta}_1$ is the slope of the line, a value of 0 would mean a horizontal line. If the line was horizontal, it would imply that there is no relationship between X and Y. Thus, our objective is to always check whether $\hat{\beta}_1 = 0$ or not.

The data that we collect is known as a **sample** and is a representation of the whole **population**. Clearly, it is impossible for us to collect the heights of mothers and their daughters all over the world. So we collect heights from a sample that represents the population. Every time, we collect data from a sample, we will get different estimates of β_0 and β_1.

Remember how in our dataset we had chosen 200 observations at random from 1100 observations so that the scatterplot was not messy. If we take different sets of 200 observations again, we will get different estimates of β_1.

No.	$\hat{\beta}_1$
1	0.61
2	0.59
3	0.51
4	0.54

Notice, from the table that the $\hat{\beta}_1$ values are close to 0.58, but not exactly 0.58. Thus, for each sample we get different estimates. And so we need to check if for our sample, the value of $\hat{\beta}_1$ is different enough from 0 for us to be confident that it is in fact, not 0.

This is done by using a method known as **Hypothesis Testing**. A step by step explanation on this can be found in the references [2] [3]. I present briefly, how we decide whether $\hat{\beta}_1$ is different enough from 0.

We first construct a **Null Hypothesis**, H_0, and an **Alternate Hypothesis**, H_a. H_0 is assumed to be true, and from the data, we want to gather enough evidence to reject the H_0 and accept H_a.

$H_0 : \beta_1 = 0$
$H_a : \beta_1 \neq 0$

This structure makes sense, because we want to be sure that there is in fact a relationship between X and Y, and for that to be true, we want the data to give us enough evidence to reject the null hypothesis.

Next, we calculate something called the **Test Statistic**, t which in this case is

$$t = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} \qquad \text{where } se(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma^2}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

This t follows a t-distribution with $n - 2$ degrees of freedom [3]. Intuitively, this t scales $\hat{\beta}_1$ down by its standard deviation, and gives us an idea on whether the $\hat{\beta}_1$ is different from 0.

If $|t| > 1.96$ (approx) this means, that we can reject the null hypothesis, H_0, and we say the the variable X is significant. That is to say, that we are confident that X and Y are related, and β_1 is in fact different from 0.

In our dataset, $\hat{\beta}_1 = 0.58$, $se(\hat{\beta}_1) = 0.069$ and $t = 7.79$. Since $|t| = 7.79 > 1.96$, we can reject H_0, and claim that there is a significant relationship between mother's height and daughter's height.

Additional Comments

We have, at this point succeeded in analyzing the dataset. We have found the relationship between X and Y, and shown that this relationship is significant. This is what most statisticians have to do when they are given a dataset. However there are some roadblocks, and most datasets are trickier than this one. Following are some other important aspects of statistical analysis:

- Once the model has been fit, it is important to check the assumptions. In a lot of cases, the constant variance assumption is not valid, in which case we need to transform our data [2].
- In a lot of cases, a linear relationship is not adequate. Higher order regression models should then be tried [2].
- We generally have more than one predictor variable, $X_1, X_2, \ldots, X_p$. In that case, we fit the linear regression model

  $$y_i = \beta_0 + \beta_1 x_{1i} + \ldots x_{pi} + \epsilon_i$$

  In our dataset for example, we could also introduce father's height, weight of the daughter, and time of menarche as potential predictor variables. However, in that case, we move from 2 dimensions to $p + 1$ dimensions. The regression model is then built by using matrices [2].
- Sometimes, the data does not come from a Normal distribution, in which case we fit Generalized Linear Regression Models [4].

Another important aspect of statistical analysis is computer programming. Since most datasets are large, and computations are complicated, it is impractical to do regression on paper. Statistical programming languages make life a million times easier and also provide us with some excellent graphical tools. R and SAS are the two most famous languages used. SAS is used mostly in industry settings and biostatistics work. R is used extensively in academic settings. I used R to do the analysis in this paper.

Conclusion

Statistical analysis at its core is about quantifying uncertainty. It is about making inferences from raw data, figuring out trends and concluding with confidence that the results obtained are not coincidental.

In the end, the model obtained would probably not be the exact model. For example, if we were able to get data on all mothers and daughters in the world, then β_1 might be very different from 0.58. But, the idea is to get as much information as possible from the model. There is one statement most statisticians live by, "All models are wrong, but some are useful".

References

[1] `http://www.nytimes.com/2009/08/06/technology/06stats.html?_r=0`

[2] Sanford Weisberg, *Applied Linear Regression*, 3rd edition

[3] Jay L. Devore, *Probability and Statistics for Engineering and the Sciences*, 7th Edition

[4] Julian J. Faraway, *Extending the Linear Model with R: Generalized Linear, Mixed Effects, and Nonparametric Regression Models*, 1st Edition