Statistics Lab 5 Notes

Jon Ashbrock

February 21, 2018

1 introduction

After this week's lab a student will:

1. Understand the idea of regression modelling
2. Know how the least squares linear regression model is computed
3. Know the meaning of each parameter in the least squares method and what the R^2 value means
4. Be able to set up and plot linear regressions in **R**

2 Regression Modeling

There are two types of statistical analysis one can generally do. The first is determining whether or not some statement about data is true or false. This is hypothesis testing and will be covered later in this course. The second is using a sample of past events to make a prediction about things which we do not have data for. This called regression modeling.

In this type of modeling we try to determine the relationship between two variables. As an example from the previous lab, we might be interested in predicting the military size of the countries where that information was missing by using the population of those countries. To do this prediction, we compute the relationship between the military and population variables for the countries that have available information.

This prediction is done by writing the the unknown quantity (called the dependent variable) as a function of the measured quantity (called the independent variable). There are many types of functions we can use to model data, but in this lab we will focus only on linear functions. That is, we are investigating how to find a "line of best fit" for our data.. In future labs we will explore how to build models out of different types of functions.

In general, we could try to find a relationship between any two variables. However the relationship is not always as strong as we might like it to be. For instance, we would expect population to be a much stronger predictor of military size than a country's average high temperature. So, in addition to computing the functional relationship, we also want to compute how good this relationship is.

3 The least squares regression model

In this section we are interested in modeling the dependent variable, y, as a linear function of the independent variables, x. To do this we must determine the slope and the $y-$intercept of the line. Intuitively we will do this by minimizing the distance (or error) between our line and our data points. To formally describe this, we need to define the mathematical notation we will use.

Let $\{(x_i, y_i)\}_{i=1}^n$ be a set of points in the cartesian plane (which may be interpreted as a sample of points on a scatter plot). Now suppose we have a line $f(x) = \beta_1 x + \beta_0$. For each point x_i in our sample, the point on the line at the value of x_i is given by $\beta_1 x_i + \beta_0$. However, we know the true value of our function at this point to be y_i. Therefore, we can define the difference between this true value y_i and the predicted value $f(x_i)$. Intuitively, we wish to minimize the total error of all points. For the ease of mathematical analysis, we actually choose to minimize the squared error $(y_i - (\beta_1 x + \beta_0))^2$ for each point. This is exactly why the method introduced today is called the least squares regression model. The problem can then be stated formally as

$$\text{Choose the values of } \beta_0, \beta_1 \text{ so that the quantity}$$
$$\sum_{i=1}^{n}(y_i - f(x_i))^2 = \sum_{i=1}^{n}(y_i - (\beta_1 x_i + \beta_0)^2 \text{ is minimized}$$

There are a number of mathematical results telling us that there always exists exactly one line which minimizes this error as we desire. Thus we can always be sure we get an answer and that every time we do this we get the same answer!

4 Interpretation of the Least Squares Model

Suppose we have a sample $\{(x_i, y_i)\}_{i=1}^n$ of points in the plane and a least squares model

$$y = f(x) = \beta_1 x + \beta_0$$

Typically along with such a model, the analysis will provide an r^2 value associated to the model. In this section we will explain what the parameters β_1, β_0, and r^2 mean in relation to the model.

4.1 The r^2 statistic

The r^2 statistic is the key value in determining how good our model is at prediction. r^2 is also referred to as the coefficient of determination. The r^2 statistic can take any value between 0 and 1. Values near 1 indicate a strong linear relationship while values close to 0 indicate lack of a linear relationship between two variables. On an intuitive level, the r^2 statistic tells you how good your regression line is at predicting new data.

Let us think back on our military size example. Even when knowing the population of a country, we cannot exactly predict the size of their military. That is, there is still variance even amongst countries of similar population. However knowing the population of a country allows us to more accurately estimate the distribution of the military sizes. The r^2 value measures exactly how much of the variance of the $y-$values is explain by controlling for the predicted $y-$value by our model. Said another way, it is the amount of the variance in y which can be attributed to the variance of x. For the sake of completeness, given a linear model $f(x) = \beta_1 x + \beta_0$ fit to a data set $\{(x_i, y_i)\}_{i=1}^n$ the value of r^2 is given by:

$$r^2 = 1 - \frac{\sum_{i=1}^n (\beta_1 x_i + \beta_0 - y_i)^2}{\sum_{i=1}^n (y_i - \overline{y})^2}$$

where $\overline{y}$ is the average value of the y. One will notice the denominator is the variance of the y values and the numerator is the squared error in prediction.

4.2 β_1, β_0

The parameter β_1 is the computed slope of the least squares regression line. While the r^2 value tells us how *good* the model is, the parameter β_1 tells us how much y increases when x increases.

The parameter β_0 is the $y-$intercept of the model. Statistically, this number is the expected $y-$value when the $x-$ value is 0. Often times this does not have a meaningful physical interpretation, but sometimes it does. This will become clearer when we do an example in the following subsection.

4.3 An example of LSRL with Interpretation

As we saw with the previous assignment, there is not a very strong relationship between the population and the military size of the countries for which the data is available. In fact, if you run an LSRL (least squares regression line) analysis on the data set, the coefficient of determination (r^2) is only 0.203. This tells us that if we look at the distribution of military sizes and the variance of this sample, only about 20% of the variance can be explained as being due to the different sizes of the country. If I remove some outliers, the coefficient of determination increase to 0.389, which is still quite a small value. From that we can conclude that the relationship between military size and population is not in a linear relationship.

If we look at the relationship between the GDP of a country and the amount of electricity (in kWh/person) that the country uses, there is a stronger linear relationship. The equation for the LSRL is $y = 1020.25 + 0.1892x$ with $r^2 = 0.700$. The interpretation for this model is as follows. About 70% of the variation in how much electricity a country uses can be explained away by looking at the countries GDP. Only 30% of the total variance in electricity usage varies once we control for GDP. The slope of the regression line is 0.1892 which tells us that an increase by 1 USD in GDP corresponds to an expected increase in 0.1892 kWh of electricity usage per person per year. The y-intercept is fairly meaningless in this model, it tells us that if a country would have 0 USD GDP then their expected electricity usage would be 1020 kwh per person per year.

5 LSRL in R

There are three functions which are used in computing and plotting least squares regression lines in **R**. The first is the function "lm" which we will be using over the next few weeks as it allows us to fit more than just regression lines to our data. The second is the "cor" function which outputs the coefficient of determination between two lists. The third is "abline" which lets us plot lines on a scatter plot.

5.1 lm

The lm function accepts many parameters as input, however for our cases we only need to use two of them. In general, when we pass inputs (arguments) to functions, we just need to specify them in the correct order. However, if we know the name of the parameter we can specify that value by "$parameter = value$" where parameter is the name of the input and value is what we want to set it equal to. The two parameters we need to specify in the *lm* function are "formula" and "data". The "formula" parameters tells us what form we want our output to look like. The second, "data", is a data frame containing the data (remember you can turn two lists into a data frame using data.frame(list1,list2)). "formula" is an object that explains what kind of relationship we want. If the name of the columns in our data frame are list1 and list2 and we want a formula explaining list1 as a linear function of list2 we specify that "formula = list1 $\sim$ list2". All together this code could look like: and the function will output the slope and intercept of our

```
dataset <- data.frame(list1,list2)
lm(formula = list1 ~ list2, data = dataset)
```

regression line.

5.2 cor

The cor function accepts two lists as inputs and outputs the coefficient of determination when a least squares model is fit on the two lists. The first input is the independent variable an the second is the dependent.

5.3 abline

After plotting the scatterplot of our data we can use the "abline" function to draw the regression line. We must specify two parameters to this function. The first is the y-intercept of the line and the second is the slope of the line. These numbers will be computed by the lm function.