Regression

Lelys Bravo de Guenni

April 24th, 2015
Outline

Regression

Simple Linear Regression
Prediction of an individual value
Estimate Percentile Ranks
Regression

Simple Linear Regression

The idea behind the calculation of the coefficient of correlation is that the scatter plot of the data corresponds to a cloud that follows a straight line. This idea can be formalized by regression methods.

In this chapter we will:

- Consider the definition of simple linear regression
- Find a method to predict an individual value
- Use the normal curve to estimate the percentile rank
- Describe the regression effect
Correlation and Regression

- Compute the regression errors and its RMS
- Study the behavior of regression errors
The regression method describes how one variable depends on another.

The Northern California temperature data have average altitude of 3,524 feet and a SD of 1,839 feet; average temperature of 70.3 degrees and SD 6.5 degrees. The correlation between temperature and altitude is -0.76
The cloud of points shows a mild negative association between the two variables, as does the value of $r$. Can we use the values of altitude to estimate the average values of temperature?

The regression line for $y$ on $x$ estimates the average value of $y$ corresponding to each value of $x$. 

April 24th, 2015 | UCSC — AMS5
Associated with an increase of one SD in $x$ there is an increase of $r \times$ SDs in $y$ on average.
Clearly, if the correlation coefficient is negative, then the average value of \( y \) decreases as \( x \) increases.

In the temperature and altitude example, an increase of height of 1,839 feet produces an increase of \( -0.76 \times 6.5 = -4.95 \) degrees in the average temperature.
Prediction of an individual value

How do we use the method to predict an individual value?

If we consider two variables $x$ and $y$ and we want to predict the value of $y$ for a specific value of $x$, we use the average value of $y$ that corresponds to the value of $x$ according to the regression method.
Example:

The first year GPAs and the Math SAT for the students of a university produce the following data

\[
\begin{align*}
\text{average SAT score} & = 550 \quad SD = 80 \\
\text{average 1st-year GPA} & = 2.6 \quad SD = 0.6 \\
\end{align*}
\]

\[ r = 0.40 \]

We want to predict the 1st-year GPA of a student with a SAT score of 650.
The student’s SAT score in standard units is

$$\frac{650 - 550}{80} = 1.25$$

so the score is 1.25 SDs above average. An increase of one SD above the average SAT score produces an increase of $0.4 \times 0.6$ GPA points. This implies that our student will have an increase of

$$1.25 \times 0.4 \times 0.6 = 0.3$$

points of GPA above average. Since the average GPA is 2.6, the predicted GPA is

$$2.6 + 0.3 = 2.9$$

This is the average GPA that we expect for students with SAT scores around 650.
**WARNING:** You can use the regression method on new subjects provided that they are similar to the ones that were used to produce the averages, SDs and $r$ used in the regression method.

In the previous example the method will not be valid for students of a different institution.
Estimate Percentile Ranks

We can use the regression method and the normal curve to produce estimates of the percentile ranks.

Example: In the previous example suppose a student has a percentile rank of 90% for the SAT scores. That is, only 10% of the scores are higher than his. What is the predicted percentile rank for the 1st-year GPA of this student?

Using the normal curve we have that a 90% probability corresponds to $z$ score of 1.3. This means that the student’s SAT score is 1.3 SDs above average.

This corresponds to being

$$0.4 \times 1.3 \approx 0.5 \text{ SDs above the average GPA}$$
and this corresponds to an accumulated probability, under the normal curve, of approximately 69%.
So the percentile rank on 1st-year GPA of a student with a percentile rank on SAT score of 90% is predicted to be 69%.

In solving this problem, the averages and SDs of the two variables are not used. The whole problem is worked in standard units.

Notice that the student with a SAT percentile rank of 90% was ‘pulled down’ to only 69% by the regression method. Why is that?

Suppose the correlation was perfect, $r = 1$, then 90% will convert to 90%. The other extreme is that there is no correlation, so, in the absence of any information, the best guess is the median or 50% percentile. The regression method produces a rank that is somewhere between these two extremes.
Prediction example

Shoe sizes: The shoe size and the heights of 14 men are recorded. The shoe size average is 10.46 with a SD of 1.21. The average height is 70.45 inches with a SD of 2.45 inches. The correlation is 0.93. What is the average height of a man that uses shoes of size 11.5?

We convert 11.5 to standard units

\[
\frac{11.5 - 10.46}{1.21} = 0.859
\]

so the shoe size is 0.859 SDs above average. This means that the height will be

\[
0.859 \times 0.93 \times 2.45 = 1.95
\]
inches above average. So the average height of a man with shoe size 11.5 will be

\[ 70.45 + 1.95 = 72.40 \]

inches.