Statistics 410: Regression
Spring 2002

Regression: An Introduction

(January 15, 2002)

This introduction follows chapters 8-12 from the introductory statistics book by Freedman, Pisani, and Purves (*Statistics*, Third edition, 1998, Norton). It is arguably the best introduction to simple linear regression, without calculus or geometry. The emphasis is on basic concepts and application. Some of this may be review for you, but perhaps from a different perspective.

Correlation

Karl Pearson (England, 1857-1936) was interested in studies considering the resemblances among family members. For example, he was interested in knowing how the height of a son is related to the height of his father. To this end, he developed what is now known as the *Pearson correlation coefficient* (r), which numerically summarizes the linear association between two quantitative variables. More specifically, r measures how tightly bivariate data (x,y) cluster about a line. (Which line?) The tighter the clustering, the better one variable can be predicted from the other. (Later in the course we will study another correlation coefficient, the *intra*class correlation coefficient. Sometimes r is called the *inter*class correlation coefficient.)

Some facts of the Pearson correlation coefficient.

1. $-1 \le r \le 1$ (r is unitless) [r > 0, positive association. r < 0, negative association.] Nevertheless, even when |r| is close to 1, the scatterplot will still show a fair amount of spread around the line of clustering: A typical point will be above or below the line by an amount approximatley equal to $\sqrt{2(1-|r|)} \times SD(y)$.

2. Mathematical definition: $r = \dfrac{\mathrm{cov}(X,Y)}{SD(X) \times SD(Y)}$

3. Calculation: $r = \dfrac{\mathrm{cov}(x,y)}{SD(x) \times SD(y)} = \dfrac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2}\sqrt{\sum (y-y)^2}}$

4. The points in a scatterplot tend to cluster around the SD line, which goes through the point of averages and has slope ± SD(y)/SD(x).

5. R is not affected by (a) interchanging the two variables and (b) affine transformations, defined by $x' = a + x$ and $x' = bx$ for b > 0. (Why?)

Additional Remarks:

How r works as a measure of (linear) association: See figure in class.

Correlations based on averages or rates can be misleading as they tend to overstate the strength of an association. Such *ecological correlations* are often used in political science and sociology. Example: Current Population Survey for 1993 data show that the correlation between income and education (men, age 25-34) in the U.S. is r = 0.44. However, the correlation between average income and average education calculated for each state and Washington D.C. is r = 0.64. (Why does this happen?)

Correlation measure association, but association is not the same as causation.

Problem 1

In a study of 1993 Math SAT scores, the Educational Testing Service computed the average score for each of the 51 states (D.C. included), and the percentage of the high-school seniors in that state who took the test. The correlation between these two variables was -0.86. (a) True or false: test scores tend to be lower in the states where a higher percentage of the students take the test. If true, how do you explain this? If false, what accounts for the negative correlation? (b) In New York, the average score math score was only 471, but in Wyoming the average was 507. True or false, and explain: the data show that on average, the teachers in Wyoming are doing a better job at math than the teachers in New York. (c) The average verbal score for each state was also computed for each state. The correlation between the 51 pairs of average math and verbal scores was 0.97. Anecdotal evidence seems to indicate that many students do well on one of the two sections of the SAT, but rarely both. Does the reported correlation of 0.97 address this issue? Explain. .■

Simple Linear Regression

In simple linear regression we have data (x,y) and we wish to describe the association between x and y by a statistical model, which takes into account the stochastic nature of the context. Following standard notation, x is the independent variable and y is the dependent variable. In a deterministic setting, y is a function of x and for each (fixed) value of x we will observe the same y-value. In a stochastic setting, we observe an association between y and x through empirical data, but for each fixed value of x we may observe different values of y - because of chance variation. And, the distribution of y-values may be different from one x-value to the next x-value. Although this is the case, it may very well be that the average values of the different y-distributions follows a nice functional form, which we call the *regression function*: $E(Y|X=x)$. Thus, regression deals with conditional expectation in that the regression line for y on x estimates the average value for y corresponding to each value of x.

Graph of Averages: The regression line is a smoothed version of the graph of averages. If the graph of averages follows a straight line, that line is the regression line.

Simple linear regression: $E(Y|X=x) = a + bx$.

Question: Does the correlation between x and y play a role in regression?

Regression method: Associated with each increase of one SD in x there is an increase of only r SDs in y, on average.

Problem 2

A university has made a statistical analysis of the relationship between math SAT scores (ranging from 200-800) and first-year GPAs (ranging from 0-4.0), for students who complete the first year. The average math SAT is 550 with an SD of 80, while the average GPA is 2.6 with an SD of 0.6. The correlation between the math SAT and GPA is r = 0.4. The scatterplot is football-shaped and math SAT and GPA are each normally distributed. (a) A student is chosen at random, and has a math SAT of 650. Predict his first-year GPA. (b) The percentile rank of one student on the math SAT is 90%, among the fist-year students. Predict his percentile rank on first-year GPA. (c) Repeat part b for a student at the 10^{th} percentile of the SAT.■

Question: How does the regression line differ from the SD line?

Prediction: The regression method gives sensible results when the association between x and y is not non-linear. Be careful.

The *root-mean-square (r.m.s.) error for regression* says how far typical points are above or below the regression line. When the scatterplot is football-shaped, the r.m.s error for the regression of y on x can be found as $\sqrt{(1-r^2)} \times SD(y)$.

Problem 3

Using the results from Problem 1: (a) About what percentage of the students had first-year GPAs over 3.3? (b) Of the students who scored 600 on math SAT, about what percentage had first-year GPAs over 3.3? .■

Closing Remarks

Regression is much more complicated than what is indicated in this brief introduction. However, the ideas presented here should be kept in mind for the rest of your statistical lives. Much of the discussion found in Freedman et al. will not be found in more advanced textbooks covering regression and I encourage you to follow-up with a complete reading of cited chapters. We will cover regression via matrix algebra, least squares, analysis of variance, multiple testing, and so, but it is important that you clearly understand what regression is about at its most basic level. Throughout the course we will learn about modeling, estimation, inference, model diagnostics and remedial measures. The underlying mathematics and formulas are certainly worth knowing, but it is actually more important (at this stage) for you to understand the role of (stochastic) assumptions, basis for application, and interpretation of results. Next to t-tests, regression and ANOVA methods are the most widely used statistical procedures. It is important to learn how to properly apply these methods.