Analytic Techniques for Public Management and Policy

Analytic Techniques for Public Management and Policy

JIWON N. SPEERS

SALIH BINICH AND RUSSELL ALMOND

Contents

Preface 1

Chapter 1. Correlation 3

Preface

It took 10 years to organize this e-book. I was introduced to general linear models as a teaching assistant in 2012 in the program of measurement and statistics at Florida State University (FSU). At the time I was struck by the power of the analytic techniques – a suite of statistical knowledge and techniques for drawing analytical conclusions from a set of numeric values. At the same time, I was struck by the fundamental limitations (e.g., regression assumptions) of the applicability of the techniques, especially when applied to the field of public management and policy. I left the program with the impression that statistics was attractive but ultimately not very useful for social scientists, vowing to monitor progress in the field.

I returned to the program in 2015, and have learned in-depth analytical techniques since that time. I earned my Ph.D. in measurement and statistics in 2020 (I have initially had my Ph.D. in public management and policy). The past decade has been an exciting time for my academic journey as computer technology has developed rapidly, particularly in the area of psychometrics. These developments have coupled with recent advances in econometrics and the ever-increasing quality of quantitative research in the social sciences. *Analytic Techniques for Public Management and Policy* was written with the hope where the techniques can be used effectively to be evidence-based research and that it might encourage public management and policy researchers to inform more effective governance. This e-book based on ordinary least squares (OLS) regression is mostly based on three resources: Dr. Russell G. Almond's statistics classes, Dr. Salih Binich's measurement classes, and Dr. Tom Cook's quasi-experimental design workshop. I am very grateful to Dr. Almond and Dr. Binich at FSU, and Dr. Cook at Northwestern University.

Chapter 1. Correlation

In the Basic Statistical Analysis Course (e.g., PUAD 628), we dealt with a single variable or univariate data. Another type of important statistical analysis problem is the problem of identifying the relationship between multiple variables. To do so, we need to turn to bivariate data. For instance, economists are often interested to understand the relationship between two variables as follows,

(1) Education and wages,
(2) Salaries and CEO performance, and
(3) Aid and economic growth

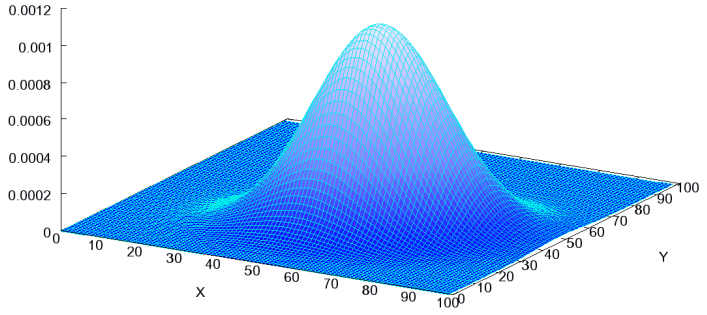
In these problems, we are interested in whether one variable increases accordingly to the other, and whether the relationship is very pronounced or to the extent that there is a trend. If this relationship is identified, it can be appropriately used for business strategy, investment strategy, economic policy, and educational policy establishment.

Correlation analysis and regression analysis are methods of analyzing the relationship between the two variables. Correlation analysis is interested in the degree to which the correlation between the two variables is clear. On the other hand, regression analysis is interested in deriving the relationship between the two variables into a specific equation. Accordingly, in correlation analysis, two variables are treated as two equal random variables, whereas in regression analysis, one of the two variables is regarded as an independent variable, so only the dependent variable is treated as a random variable.

In other words, under the perspective of correlation analysis, the levels of the variables are not under the control of the researcher because variables constitute random samples from the population. However, under the mentality of regression, one variable is clearly an outcome we want to predict or understand. In regression, a dependent variable is treated as a random variable, but independent

variables (predictors) are treated as fixed variables (i.e., predictors constitute the only values of interest in the study so that the levels of the variables are under the control of the researcher).

These two analysis methods are used complementarily to identify the relationship between variables. In this chapter, we first look at correlation analysis, and then we look at regression analysis in the next chapter. To measure the association between two variables, a joint distribution is used as follows:



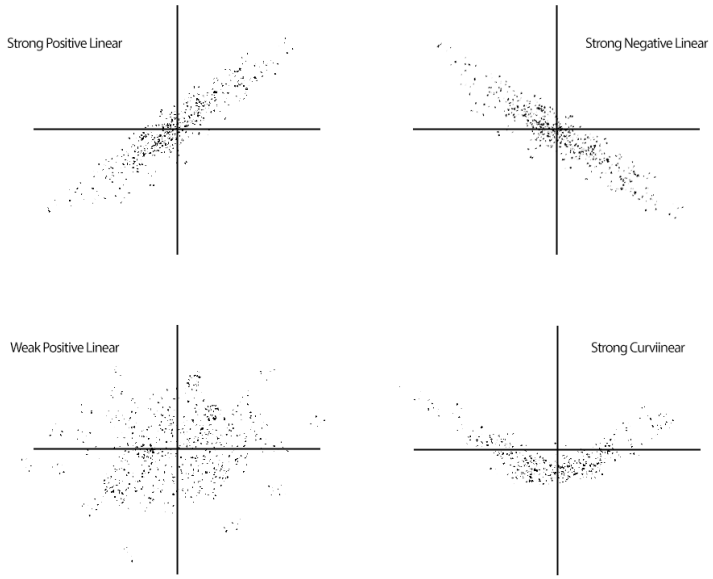
Here, an independent variable is located on the x-axis and the dependent variable is depicted on the y-axis. The independent variable is labeled X and is usually placed on the horizontal axis, while the other, dependent variable, Y, is mapped to the vertical axis. The height is seen as the frequency of observations (or cases).

The above joint distribution displays a normal distribution, and the normal distribution consists of three elements:

(1) Bell-shaped,
(2) Symmetric, and
(3) Unimodal.

1. Scatterplot

To explore relationships between two variables, we often employ a scatterplot, which plots two variables against one another.



We typically begin depicting relationships between two variables (i.e., X-Y Relationship) using a scatterplot—a bivariate plot that depicts three key characteristics of the relationship between two variables.

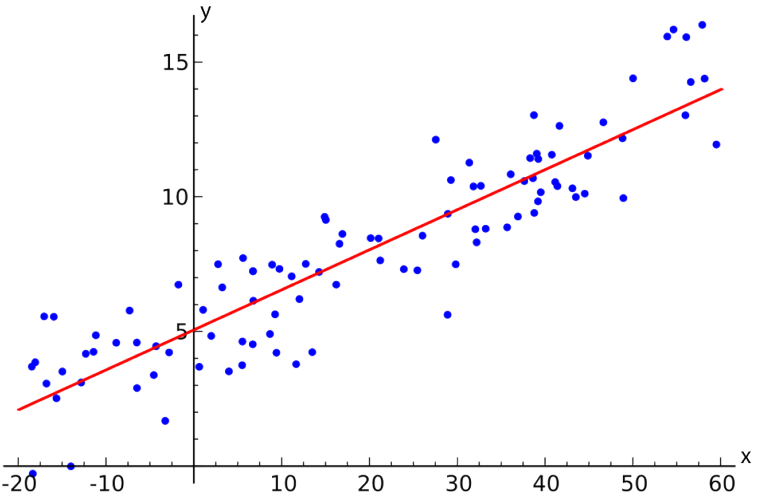
(1) **Strength**: How closely related are the two variables? (Weak vs. strong)

(2) **Direction**: Which values of each variable are associated with the values of the other variable? (Positive vs. negative)

(3) **Shape**: What is the general structure of the relationship? (Linear vs. curvilinear or some other form)

By convention, when we intend to use one variable as a predictor of the other variable (called the criterion or outcome variable), we put the predictor on the x-axis and the criterion or outcome on the y-axis.

When we want to show that a certain function (here, a line) can describe the relationship and that that function is useful as a predictor of the Y variable based on X, we include a regression line—the line that best fits the observed data.



As is true for many other characteristics of distributions that we wish to describe, parameters and statistics describe the association between two variables. The most commonly used statistic is the Pearson Product Moment Correlation (r, which estimates a population parameter of ρ – rho). The correlation coefficient captures the three aspects of the relationship depicted in the scatterplot.

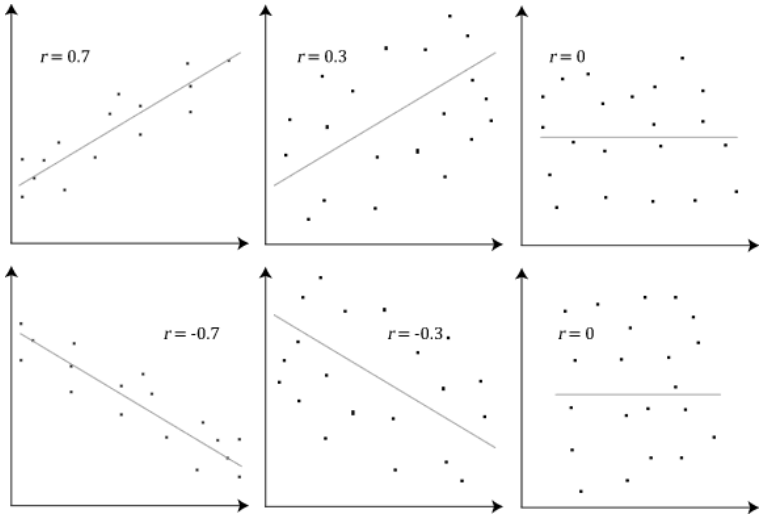
(1) **Strength**: How closely related are the two variables? The absolute value of r ranges from 1 (*positive or negative*) for a perfect relationship to 0 for no relationship at all.

(2) **Direction**: Which values of each variable are associated with the values of the other variable? A positive sign, or no sign, in front of r indicates a positive relationship while a negative sign indicates a negative relationship.

(3) **Shape**: What is the general structure of the

relationship? Correlation (r) always depicts the fit of the observed data to the best-fitting straight line.

Note that if a scatterplot does not show a linear relationship, we do not take it as a correlation because if a relationship is not linear. In other words, even though a statistical software program generates a numeric value for a correlation once you input data that represent two variables, it does not mean it is an actual correlation (r) because it is not always linear between the two variables. If the actual relation is nonlinear, then the correlation value generated by the statistics tool should be nullified.



The magnitude of correlation (r) is between -1 and +1. A no correlation represents $r = 0$. Both -1 and +1 are the maximum correlation, whereas the signs are opposite. According to Cohen's rules of thumbs, a small correlation ranges $0 < r < .1$, a medium correlation is $.1 < r < .3$, and a large correlation is $.3 < r < .5$, respectively.

II. Covariance

An important concept relating to correlation is the covariance of two variables (S_{XY}–note that the covariance is a measure of dispersion between X and Y). The covariance reflects that degree to which two variables vary together or covary. The equation for the covariance is very similar to the equation for the variance, only the covariance has two variables.

$$S_{XY} = \frac{\Sigma_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n-1},$$

where $\bar{X}$ is mean of X_i, $\bar{Y}$ is mean of Y_i, n is number of sample size, and individual is i. Note the denominator is $n-1$, not just n. In general, when the covariance is a large, positive number, Y tends to be large when X tends to be large (both are positive). When the covariance is a large, negative number, Y tends to be large and positive when X tends to be large but negative. When the covariance is near zero, there is no clear pattern like this–positive values tend to be canceled by negative values of the product.

However, there is one problem with the covariance–it is in raw score units, so we cannot tell much about whether the covariance is indeed large enough to be important by looking at it. The solution to this problem is the same solution applied in the realm of comparing two means–we standardize the statistic by dividing by a measure of the spread of the relevant distributions. Thus, the correlation coefficient is defined as:

$$r_{XY} = \frac{s_{XY}}{s_X s_Y},$$

where S_X and S_Y are standard deviations of the X and Y scores and S_{XY} is the covariance. That is, correlation is standardized covariance.

Because S_{XY} cannot exceed $|S_X S_Y|$, the limit of $|r|$ is 1.00. Hence, one way to interpret r is as a measure of the degree to which

the covariance reaches its maximum possible value—when the two variables covary as much as they possibly could, the correlation coefficient equals 1.00. Note that we typically do not interpret r as a proportion, however. Therefore, the correlation coefficient tells us the strength of the relationship between the two variables. If this relationship is strong, then we can use knowledge about the values of one variable to predict the values of the other variable.

Recall that the shape of the relationship being modeled by the correlation coefficient is linear. Hence, r describes the degree to which a straight line describes the values of the Y variable across the range of X values. If the absolute value of r is close to 1, then the observed Y points all lie close to the best-fitting line. As a result, we can use the best-fitting line to predict what the values of the Y variable will be for any given value of X. To make such a prediction, we obviously need to know how to create the best-fitting (i.e., regression) line.

III. Principles of Regression

Recall that the equation for a line takes the form $Y = mX + b$. However, it is common to use two symbols that are b's with subscripts. I will use the notation

$$Y = b_0 + b_1X$$

We need to show whether Y is an actual score or our estimate of a score. We will put a hat (^) over the Y to indicate that we are using the linear equation to estimate Y. Also, we subscript the Y and X with i to index the scores for the i^{th} case. The line is

$$\hat{Y}_i = b_0 + b_1X_i$$

This is called a "fitted or estimated regression line." The components or parameters in the equation are defined as follows:

$\hat{Y}_i$ is the value of Y predicted by the linear model for case i.

b_1 is the slope of the regression line (the change in $\hat{Y}_i$ associated with a one-unit difference in X).

b_0 is the intercept (the value of $\hat{Y}_i$ when $X = 0$).

X_i is the value of the predictor variable for case i.

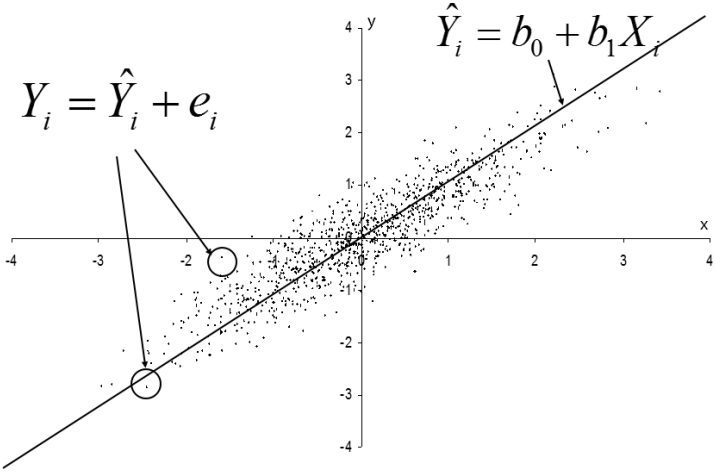
There are several other versions of the model. The one above represents the predicted scores but we can also write the model in terms of the observed scores:

$$Y_i = b_0 + b_1 X_i + e_i.$$

Note that we've added an error term e_i and now Y does not have a hat. This is also equivalent to the model showing the predicted score plus an error or residual:

$$Y_i = \hat{Y}_i + e_i.$$

Note that the first model is an equation for the line and the other two are equations for the points that fall around the line. Therefore, the equation for the line describes the points right along the line and the other equations describe the points:



We will have a variety of notations for regression and different books do not all use the same notation. I use the hat (^) over Y to indicate an estimated score of $\hat{Y}_i$.

We also use the hat over Greek symbols such as β to indicate estimates of population parameters. One confusion we will need to deal with (later) concerns "beta weights" or "standardized coefficients" which some books denote using Greek letters even though they are sample estimates. I will call these b^*.

Our task is to identify the values of b_0 and b_1 that produce the best-fitting linear function. That is, we use the observed data to identify the values of b_0 and b_1 that minimize the distances between the observed values (Y) and the predicted values ($\hat{Y}_i$). However, we can't simply minimize the sum of differences between Y and $\hat{Y}_i$ (recall that $e_i = Y_i - \hat{Y}_i$ is the residual from the linear model) because any line that intersects $(\overline{X}, \overline{Y})$ on the coordinate plane will result in an average residual equal to 0.

To solve this problem, we take the same approach used in the computation of the variance–we find the values of b_0 and b_1 that minimize the squared residuals. This solution is called the (ordinary) least-squares solution (i.e., OLS regression).

Fortunately, the least-squares solution is simple to find, given statistics that you already know how to compute.

$$b_0 = \overline{Y} - b_1\overline{X}$$

$$b_1 = \frac{S_{XY}}{S_X^2} = r_{XY}\frac{S_Y}{S_X}$$

These values minimize $\Sigma_{i=1}^{n}(Y_i - \hat{Y}_i)^2$, the sum of the squared residuals.

[Exercise 1]

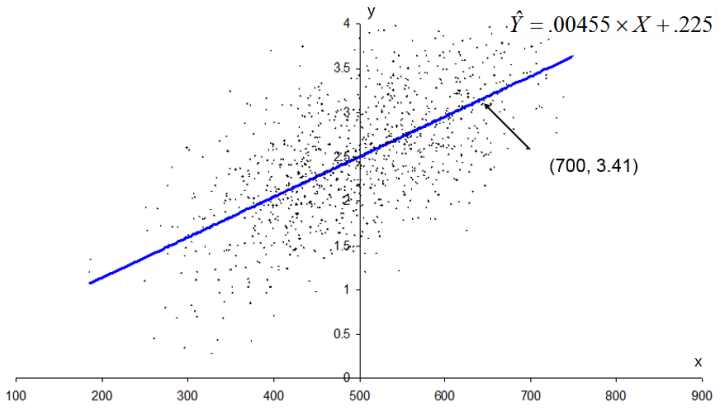
As an exercise example, consider the data below. We are interested in determining whether wages (X) would be useful in predicting first-quarter productivity (Y) for factory workers. So, we decide the wages for a group of workers, allow all of them to work, and then obtain each worker's productivity after one-quarter of work. We get the following descriptive statistics.

$\overline{X} = 500$
$\overline{Y} = 2.5$
$s_X = 100$
$s_Y = .70$
$r_{xy} = .65$

Based on the given conditions, provide the estimated regression equation.

Let's plot that regression line (i.e., a statistics software program will

plot this line for you if you have raw data). The line will always pass through the point ($\overline{X}$, $\overline{Y}$) which is (500, 2.5) for our data.



We may compute $\hat{Y}_i$ for another X value (say, 700) to get a second point on the line:

$$\hat{Y}_i = .00455(700) + .225 = 3.41$$

Therefore, what do this regression line and its parameters tell us?

The intercept tells us that the best guess at productivity when wages = 0 equals .225–a situation that is conceptually impossible because wages cannot be as low as zero. This points out an important point, sometimes the model will predict impossible values.

What is $\hat{Y}_i$ for $X = 900$?

The slope tells us that, for every 1-point increase in wages, we get an increase in productivity of .00455. The covariance and correlation (as well as the slope) tell us that the relationship between wages and productivity is positive. That is, productivity tends to increase when wages increase.

Note, however, that it is incorrect to ascribe a causal relationship between wages and productivity in this context. There are several other conditions that need to be met in order to confidently state

that interventions that change wages will also change productivity. Do you know what those are?

We will now spend the next two months or so learning all of the steps in regression analysis. This is where we are headed, but there are many pieces of this process to learn. Today we saw how to "estimate model" for one X (one independent variable).

(1) Preliminary analyses

- Inspect scatterplots.
- Conduct case analysis.
- If no problems, continue with regression analyses.

(2) Regression analyses

- Estimate model.
- Check possible violations of assumptions for this model.
- Test overall relationship.
- If the overall relationship is significant, continue with the description of the effects of independent variable (IV)'s (or if not, try other models).
- For each interval and dichotomous IV, test coefficient, compute interval, assess the importance, and compute a unique contribution to R^2.
- For each categorical IV, test global effect and, if significant, follow up with test, interval, and assessment of importance for each comparison.
- If the equation will be used for prediction, assess the precision of prediction.

Sources: Modified from the class notes of Salih Binich (2011) and Russell G. Almond (2011).