GEOMETRICAL APPROACH TO RANDOM VARIABLES

Farida Kachapova
Auckland University of Technology
Ilias Kachapov

Abstract

In this paper we suggest a geometrical approach in teaching random variables and regression. Many statistics textbooks (e.g. Wild and Seber, 2000; Chatterjee, 2000; Moore and McCabe, 2006) concentrate on samples instead of population models. The authors believe that important statistical concepts and ideas should be explained in terms of population first before developing techniques for samples. Below we formulate a population model of linear regression in geometrical terms - as an orthogonal projection. This provides the students with conceptual knowledge that can be further applied to the case of samples.

EUCLIDEAN SPACE OF RANDOM VARIABLES

We will fix a probability space (Ω, F, P) for the rest of the paper. Here Ω is a sample space of elementary events (outcomes), F is a σ-field of events (subsets of Ω) and P is a probability measure on the pair (Ω, F). Denote R the set of all real numbers.

For random variables X and Y, denote E(X) the expected value of X and denote Cov(X, Y) the covariance of X and Y.

Define
$H_0 = \{ X \mid X$ is a random variable on (Ω, F, P) and $E(X^2) < \infty \}$.
Thus, H_0 is the set of all random variables with finite expected values. It is easy to show that H0 is closed under the operations of addition and multiplication by a real number.

Denote $X \sim Y$ if $P\{\omega \mid X(\omega) \neq Y(\omega)\} = 0$. Thus, $X \sim Y$ means that the random variables X and Y are equal with probability 1. Apparently $\sim$ is an equivalence relation on H0. Denote [X] the equivalence class of the random variable X.

Define
$H = \{ [X] \mid X \in H_0 \}$.

Since the equivalence ~ is consistent with the operations on H0, we can define the operations of addition and multiplication by a real number on H:
$[X] + [Y] = [X + Y]$ and $\lambda \cdot [X] = [\lambda \cdot X]$.

In the rest of the paper we will use the notation X instead of [X] for brevity remembering that equivalent random variables are considered equal.

Theorem 1. The set H with the operations of addition and multiplication by a real number is a linear space.

For any $X, Y \in H$, $E(X \cdot Y)$ is defined, since $|XY| \leq 1/2\, X^2 + 1/2\, Y^2$, and $E(X^2)$ and E(Y2) are finite. It is easy to check that the function $(X, Y) = E(X \cdot Y)$ has all properties of a scalar product on H.

Theorem 2.
The scalar product given by
$(X, Y) = E(X \cdot Y)$
makes H a Euclidean space.

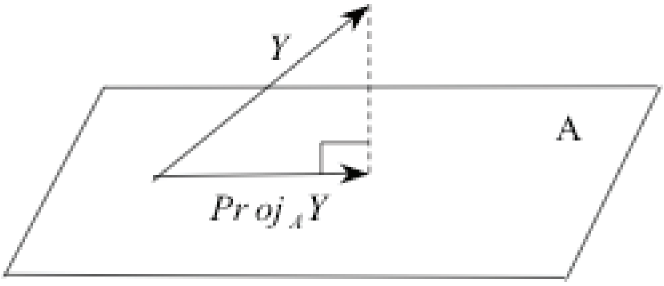
Definition.
1) For $X \in H$, the length is given by
$\|X\| = \sqrt{(X,X)} = \sqrt{E(X^2)}$.

2) For $X, Y \in H$, the distance is defined by
$d(X, Y) = \|X-Y\| = \sqrt{(E(X-Y)^2)}$.

A similar approach is used by Grimmett and Stirzaker (2004, pg. 343-347) but they do not introduce scalar product on random variables though the scalar product is very relevant to orthogonal projections and makes proofs shorter.

REGRESSION ESTIMATE AS AN ORTHOGONAL PROJECTION

For any linear subspace A of H we denote $\mathrm{Proj}_A Y$ the orthogonal projection of vector Y on A.



The following is a well known algebraic fact.

Theorem 3.
Suppose A is a linear subspace of H, $Y \in H$ and $Proj_A Y$ exists. Then $Proj_A Y$ is the closest to Y vector from A.

Choosing different A's we can get different types of regression: simple linear, multiple linear, quadratic, polynomial, etc.

Theorem 4.
The conditional expectation E(Y | X) is the closest to Y function of X.

This is based on the following fact: $E(Y \mid X) = Proj_A Y$ for $A = \{ f(X) \mid f: R \rightarrow R \text{ and } f(X) \in H \}$.

APPLICATION TO SIMPLE LINEAR REGRESSION

Theorem 5 (simple linear regression). If $\sigma_X \neq 0$, then the closest to Y linear function of X is given by

$$\alpha + \beta X, \text{ where } \begin{cases} \beta = \dfrac{Cov(Y,X)}{\sigma_X^{\ 2}} \\ \alpha = \mu_Y - \beta \mu_X \end{cases}. \qquad (1)$$

Here μ X and μY are the mean values of X and Y respectively, and $\sigma_X^{\ 2}$ is the variance of X.

Geometrical proof

Denote $A = \{a+b X \mid a, b \in R\}$. $Proj_A Y \in A$, so $Proj_A Y = \alpha + \beta X$ for some $\alpha, \beta \in R$. We just need to show that α and β sdfasd are given by the formula (1).

For ε = Y - Proj$_A$Y = Y - (α+βX), we have $\varepsilon \perp$ 1 and $\varepsilon \perp$X, 1 X$\in$A. So (ε, 1) = 0 and (ε, X) = 0, (α+β X, 1) = (Y, 1) and (α+β X, X) = (Y, X), which lead to a system of linear equations:

$$\begin{cases} E(\alpha + \beta X) = E(Y) \\ E(\alpha X + \beta X \cdot X) = E(Y \cdot X) \end{cases} \quad \text{and} \quad \begin{cases} \alpha + \beta \mu_X = \mu_Y \\ \alpha \mu_X + \beta E(X^2) = E(Y \cdot X) \end{cases}$$

Subtract the first equation multiplied by μ_X from the second equation:

$$\begin{cases} \alpha + \beta\mu_X = \mu_Y \\ \beta\left[E(X^2) - \mu_X{}^2\right] = E(Y \cdot X) - \mu_Y \mu_X \end{cases}$$

Since E(X2) - $\mu_X{}^2$ = $\alpha_X{}^2$ and E(Y • X) - $\mu_Y\mu_X$ = Cov (X, Y), we get

$$\begin{cases} \alpha + \mu_X \beta = \mu_Y \\ \sigma_X{}^2 \beta = Cov(Y, X) \end{cases} \quad \text{and the solution} \quad \begin{cases} \beta = \dfrac{Cov(Y, X)}{\sigma_X{}^2} \\ \alpha = \mu_Y - \mu_X \beta \end{cases}$$

We believe that the geometrical proof for the coefficients of simple linear regression is shorter and conceptually clearer than the usual proofs minimising mean-square error.

Corollary.
If $\hat{Y} = \alpha + \beta$X is the best linear estimator of Y from theorem 5, then the residual ε)= Y – $\hat{Y}$ω has the following properties:
1) μ_ε= 0,
2) Cov (ε, X) = 0

This immediately follows from the fact that $\varepsilon \perp$1 (hence E(ε) = 0 and $\varepsilon \perp$X (hence E (ε•X=0 and Cov (ε,X) = E (ε•X) - E(ε) • E(X) = 0).

Thus, according to the corollary, the residuals (estimation errors) equal 0 on average and are uncorrelated with the predictor X; this is another evidence that $\hat{Y}$ is the best linear estimator of Y.

Example.
Create a linear regression model for a response variable Y versus a predictor variable X if the expectations of X and Y are −8 and 5 respectively, their variances are 4 and 9 respectively and their correlation coefficient equals 0.15.

Solution

The standard deviations of X and Y are 2 and 3 respectively. Their covariance equals Cov (X, Y) = σ_X σ_Y ρ_{XY} = 2 x 3 x 0.15 = 0.9.

According to theorem 5, $\beta = \frac{Cov\left(Y,X\right)}{\sigma_X{}^2} = \frac{0.9}{4}$, β = 0.225 and

α = μ_Y - μ_X β , α= 5 -(-8)x0.225, α =6.8.

Hence $\hat{Y}$ = 6.8 + 0.225 X is the linear regression model, that is the linear function of X closest to Y.

After the population regression model was introduced we create its sample estimate. We follow the common pattern in estimation theory when a population object is estimated from a sample. For example, the population mean μ is estimated by a sample mean $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$.

Similarly the equation Y= α + βX + ε of the simple linear regression is estimated from a sample by the equation Y = a + b X + e, where a, b and e are sample estimates of α, β and ε respectively. Substituting corresponding sample estimates for the parameters in (1), we get formulas for the coefficients a and b:

$$\begin{cases} b = \dfrac{s_{yx}}{s_x{}^2} \\ a = \bar{y} - b\,\bar{x} \end{cases}$$

where x, y, $s_X{}^2$, s_{XY} are the same estimates of μ_X, μ_Y, $\sigma_X{}^2$, Cov (Y,X) respectively.

APPLICATION TO MULTIPLE LINEAR REGRESSION

Theorem 6 (multiple linear regression). Suppose random variables $X_1,\ldots, X_n$ are linearly independent. Then the closest to Y linear function of $X_1,\ldots, X_n$ is given by

$$\alpha + \beta_1 X_1 + \ldots + \beta_n X_n , \qquad \begin{cases} \beta = \begin{bmatrix} \beta_1 \\ \ldots \\ \beta_n \end{bmatrix} = S^{-1} S_0 \\ \alpha = \mu_Y - M^T \beta \end{cases} \qquad (2)$$

where S is the covariance matrix of $X_1,\ldots,X_n$;

$$S_0 = \begin{bmatrix} \sigma_{01} \\ \cdots \\ \sigma_{0n} \end{bmatrix} \text{ with } \sigma_{0k} = Cov(Y, X_k), \quad k = 1, \ldots, n;$$

$$M = \begin{bmatrix} \mu_1 \\ \ldots \\ \mu_n \end{bmatrix}$$ is the colum of the mean values of $X_1,\ldots,X_n$ and M^T is its transpose.

Similar to theorem 5, theorem 6 has a short proof, since the best linear estimator of Y is its projection $Proj_A Y$ on $A = \{ a + b_1 X_1 + \ldots + b_n X_n \mid a, b_1, \ldots, b_n \in R \}$.

The sample estimates for the coefficients of multiple linear regression can be derived from the formula (2).

CONCLUSION

Using the described approach we justify basic formulas for regression and at the same time avoid lengthy and tedious proofs. Even students without knowledge of linear algebra have an intuitive understanding of orthogonal projections in two-dimensional and three-dimensional spaces. The authors used the described approach to teaching regression in courses on statistics, probability theory and financial mathematics at the Auckland University of Technology (New Zealand) and the Moscow Technological University (Russia). The case studies show that the students gained a better understanding of the concept of regression, regression formulas and their logical connections. This improves the students' critical thinking and conceptual knowledge of regression as a complement to the procedural knowledge provided in traditional statistics courses.

REFERENCES

Chatterjee, S. (2000). Regression analysis by example. (3rd ed.). New York: Wiley.
Chiang, C.L. (2003). Statistical methods of analysis. River Edge, N.J.: World Scientific.
Dowdy, S.M. (2004). Statistics for research. (3rd ed.). Hoboken, N.J.: Wiley-Interscience.
Freund, R.J. (1998). Regression analysis: statistical modelling of a response variable. San Diego: Academic Press.
Grimmett, G. & Stirzaker, D. (2004). Probability and random processes. (3rd ed.). New York: Oxford University Press.
Hsu, H.P. (1997). Probability, random variables, and random processes. New York: McGraw-Hill.
Kachapova, F. & Kachapov, I. (2006). Mathematical approach to portfolio analysis. Auckland: Maths Ken.
Moore, D.S. & McCabe, G.P. (2006). Introduction to the practice of statistics. (5th ed.). New York: W.H.Freeman and Company.
Rao, K.K. (1966). A simplified proof of Gauss-Markov theorem when the regression matrix is of less than full rank. The American Mathematical Monthly, Vol. 73, No. 4, pp. 394-395. Mathematical Association of America.
Saville, D.J. & Wood, G.R. (1996). Statistical methods. A geometric primer. New York: Springer-Verlag.
ScheffeωH. (1959). The analysis of variance. John Wiley & Sons.
Seber, G.A.F. (1980). The linear hypothesis: a general theory. (2th ed.). Charles Griffin & Company Ltd.
Seber, G.A.F. & Lee, A.J. (2003). Linear regression analysis. (2th ed.). John Wiley & Sons.
Wild, C.J. & Seber, G.A.F. (2000). Chance encounters. John Wiley & Sons.