Leibniz FH

Research Paper Series
Nr. 6 | March 2020

A Geometric View on Linear Regression and Correlation Tests

Autor:
Roland Matthes

Leibniz-Fachhochschule

Autor:

Roland Matthes
Leibniz-Fachhochschule
matthes@leibniz-fh.de

ISSN 2511-7491

Redaktion: Robin Christmann

Leibniz FH
School of Business

Research Paper No. 6

A geometric view on linear regression and correlation tests

ROLAND MATTHES

Zusammenfassung / Abstract:

In this paper we give a geometric approach to linear regression and correlation analysis, especially we give a purely geometric derivation of the density functions for the test statistics of the simple and multiple correlation test.

Further we show that one and two sample t-tests as well as ANOVA can be interpreted as correlation tests and are therefore geometric in nature.

Fachrichtung / Field of Study: Mathematics

Klassifikation / Classification: 62H15,62J10 (Mathematics Subject Classification 2010)

Schlagworte / Keywords: -

1 Introduction

Today the methods of multivariate statistics are widely spread and serve as as a powerful tool for detecting and analyzing dependencies between different sets of data. Of course, the primary goal is to make inferences regarding possible relations between the underlying random variables, such as height, weight and income. Any such hypothetical relation we call a *model*. Typically there is a vector $\mathbf{X} = (X^1, ..., X^k)^T$ of independent variables, called regressors and another vector $\mathbf{Y} = (Y^1, ..., Y^p)^T$ of random variables, called regressands which is assumed to depend on $\mathbf{X}$. We shall write

$$\mathbf{Y} \sim F(\mathbf{X})$$

if the model states a functional dependency F.

The most common approach e.g. in econometrics consists of searching for linear models, i.e. where F is just a (affine)-linear expression, and this is closely related to what is called linear regression and correlation analysis.

The purpose of this paper, which is intended to be continued by subsequent work, is to describe the geometric nature of correlation analysis. More specifically we emphasis the aspect of spherical geometry and work with the standardized variables which can be viewed upon as elements on the hypersphere S^{n-2}. The fact that the sample correlation coefficient can be written as the inner product of the samples and hence measures the angle between them is well known to all statisticians and this geometric point of view was already taken by Fisher [6]. Yet the probability distributions for the test statistics of correlation and multiple correlation coefficients, namely Student's t-distribution and Fisher's F-distribution, are in general derived by standard methods from analysis. Here we shall derive these distributions geometrically using the surface volume measure on the hypersphere. See [4] for a similar approach in the case of Student's t-distribution and [12] 28.29 for a sketch of proof in case of Fisher's F-distribution.

We further show that prominent tests such as the t-test and analysis of variance (ANOVA) can be interpreted as correlation tests and are therefore geometric in nature.

There are several authors, e.g. Saville/Wood in [10], [13], that use a geometric setting also for didactic reasons, since they may help in a very natural way to gain a more intuitive understanding for the occurrence of probability distributions. Nevertheless most standard textbooks of today avoid geometric arguments.

With our paper we intend to give an inspiration for those who are looking for some intuitive approach to correlation analysis, e.g. for teaching purposes.

In a forthcoming paper we want to show, that also nonparametric tests such as the Wilcoxon rank-sum-test can nicely be interpreted as correlation tests and the distribution of the test statistic can naturally be derived in a geometric setting.

For further articles where the geometric picture is taken into account, we refer the reader to [2], [5], [8].

2 Simple linear regression and correlation

Subject to repeating a random experiment several times which means to take a sample of size n (synonymously: n trials, n observations) for some random variable X, we introduce random variables X_i for the value of the i-th trial. We assume that the trials are independent and that each trial follows the same distribution.

We put $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and form the centralized random vector $X_c = (X_1 - \bar{X}, ..., X_n - \bar{X})^T$ and the standardized random vector $X_s = \frac{X_c}{\|X_c\|}$, where $\|.\|$ denotes the Euclidean norm.

2.1 Sample correlation coefficient

Given (real valued) samples $x = (x_1, ..., x_n)^T, y = (y_1, ..., y_n)^T$ of size n for X and Y. We regard x and y as elements of the Euclidean vector space $\mathbb{R}^n$ with its standard inner product. As usual we define the sample mean and the sample variance for x by

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

and

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

respectively.

The sample correlation coefficient due to Pearson is defined by

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}.$$

In terms of the standardized sample data $\xi = (\xi_1, ..., \xi_n)^T, \eta = (\eta_1, ..., \eta_n)^T$ with

$$\xi_i = \frac{x_i - \bar{x}}{\sqrt{n-1}\, s_x}, \quad \eta_i = \frac{y_i - \bar{y}}{\sqrt{n-1}\, s_y} \tag{1}$$

the correlation coefficient can be written as the inner product

$$r_{xy} = \sum_{i=1}^{n} \xi_i \eta_i = \langle \xi, \eta \rangle.$$

In (1) we introduced the factor $1/\sqrt{n-1}$ to give the standardized sample variables the Euclidean length

$$\|\xi\| = \sqrt{\langle \xi, \xi \rangle} = \sqrt{\langle \eta, \eta \rangle} = 1.$$

Hence the Pearson sample correlation coefficient measures the Euclidean angle θ_{xy} between the samples

$$r_{xy} = \cos\theta_{xy}.$$

2.2 Linear regression

It is an easy exercise to show,

$$\theta_{xy} = 0 \vee \theta_{xy} = \pi \Leftrightarrow \forall i \; y_i = a + bx_i$$

with some real constants $a, b \neq 0$.

If this is the case, the data samples $x, y \in \mathbb{R}^n$ are said to be in *perfect* linear correlation.

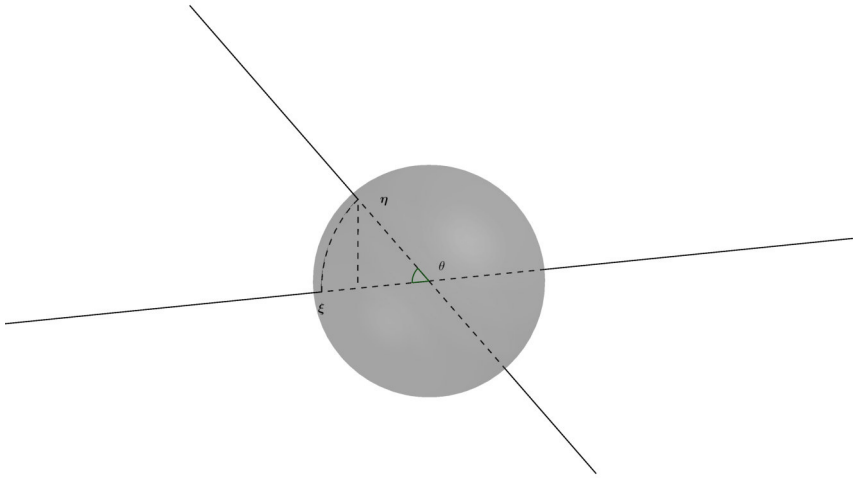

Figure 1: The correlation coefficient measures the angle between samples

The objective of inductive statistics is to infer from given data samples some model for the underlying random variables X, Y.

The process of finding an optimal linear model $Y \sim a + bX$ with $b \neq 0$ on the basis of two real valued data samples $x, y \in \mathbb{R}^n$ of size n is called linear regression. In this context *optimal* is commonly understood in the sense of least squares approximation, which goes back to Gauß and means that a, b should be chosen such that the sum of the squares of the differences $e_i = y_i - (a + bx_i)$ becomes minimal, i.e. we want to minimize

$$\Delta(a, b) = \sum_{i=1}^{n} e_i^2.$$

From the geometrical standpoint this is most natural, since if we think as before of the data samples x, y as elements of $\mathbb{R}^n$ with its natural Euclidean structure, then

$$\Delta(a, b) = \|y - (bx + a)\|^2 = \langle y - (bx + a), y - (bx + a)\rangle$$

and hence is a measure for the difference of the vectors y and $bx + a$. Here we write a instead of $(a, ..., a)^T$.

It is a simple exercise in differential calculus to obtain the optimal solution

$$b = r_{xy}\frac{s_y}{s_x}, \quad a = \bar{y} - b\bar{x}$$

and the linear regression model

$$Y \sim r_{xy}\frac{s_y}{s_x}X + \bar{y} - r_{xy}\frac{s_y}{s_x}\bar{x}. \tag{2}$$

2.3 Linear regression in the context of the standardized variables

Performing linear regression for the standardized samples turns out to make computations simpler. At first remark, that the centralized sample vector $x_c = x - \bar{x}$ is orthogonal to

any constant vector α. (As above we write $\bar{x}, \alpha$ for the corresponding constant column vector.) Indeed

$$\langle x_c, 1\rangle = \sum_{i=1}^{n} x_i - n\bar{x} = 0.$$

Hence $\langle \xi, 1\rangle = \langle \eta, 1\rangle = 0$ which yields

$$\|\eta - \beta\xi - \alpha\|^2 = \|\eta\|^2 + \beta^2\|\xi\|^2 + n\alpha^2 - 2\beta r_{xy} = 1 + \beta^2 + n\alpha^2 - 2\beta r_{xy}.$$

Therefore the optimal values are at once seen to be $\alpha = 0$ and $\beta = r_{xy}$. So β is obtained by orthogonal projection of η onto the line given by $\lambda\xi$.

For the standardized variables then regression model attains the pretty simple form

$$Y_s \sim r_{xy} X_s. \tag{3}$$

We regain (2) simply by substituting the defining expressions for the standardized samples into this formula. So there is no loss of information when working with the standardized samples.

In terms of spherical geometry we obtain the following interpretation:

1. **The sample correlation coefficients r_{xy} measures the distance of the standardized samples as elements of the $n-1$-dimensional hypersphere S^{n-1}.** Since standardized samples have mean zero, they are even elements of an $n-2$-dimensional hypersphere.

2. **Linear regression turns out to be the orthogonal projection of one standardized sample onto the other.**

In other words: Standardizing the samples and interpreting these sample data as geometric objects on a hypersphere gives a very natural infrastructure for linear regression and correlation analysis.

Having in a mind a generalization to multivariate statistics and canonical correlation analysis there is a more sophisticated way to formulate this: Linear subspaces generated by (standardized) samples (here: $\lambda\xi$) correspond to points on a Riemannian manifold (here: S^{n-1}) and correlation is a measure for their (geodesic) distance (here: the round metric $d_s(\xi, \eta)) = \arccos(r_{xy})$).

At least for $n = 3$, the spherical geometry is easy to understand and it thus may serve as an appropriate language to talk about correlation. Great circles are the geodesics and the geodesic distance between two points is given by the angle between them. As a caveat one has to conceed, that for $n > 3$ these concepts can no longer be visualized, but if we are willing to accept that the geometric terms of length and angle in higher dimensions are consistent with our geometric intuition in 3 dimensions, this will help us to gain better insight into the true nature of correlation. In what follows we shall apply these geometric concepts to multiple linear regression models.

Remark. For the standardized sample perfect positive linear correlation means $\eta = \xi$ while negative linear correlation is equivalent to $\eta = -\xi$. So they correspond to the same point or to antipodal points on the sphere. If we do not distinguish between positive or negative linear correlation, then perfectly linear correlated samples correspond to the same points on $S^{n-1}/\sim$ where $\sim$ stands for identifying antipodal points and this is the projective space $\mathbb{P}(\mathbb{R}^n)$.

Now we are going to show, how these geometric arguments carry over to testing the significance of the sample correlation coefficient.

3 Correlation test

3.1 Classical testing

Classically, when we perform a correlation test we test the null hypothesis for the true Pearson correlation coefficient $\rho = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$

$$H_0 : \ \rho = 0 \quad \text{vs.} \quad H_1 : \ \rho \neq 0$$

based on the value of the sample correlation coefficient r_{xy}.

It is assumed, that X, Y have a bivariate normal distribution. Then it is known see e.g. [9], that $\rho = 0$ implies independence, so in this situation uncorrelated and independent means the same.

As a statistic for the test one takes

$$\frac{r_{xy}}{\sqrt{1 - r_{xy}^2}} \sqrt{n-2},$$

see e.g. [3].

This statistic is known to have a Student's distribution with $n-2$ degrees of freedom, the corresponding density is given by

$$h_{n-2}(u) = \frac{\Gamma(\frac{n-1}{2})}{\sqrt{n-2}\,\Gamma(\frac{n-2}{2})\Gamma(\frac{1}{2})}(1 + \frac{u^2}{n-2})^{\frac{1-n}{2}}.$$

In what follows we want to derive this result by a geometric approach.

3.2 Geometric approach

3.2.1 Spherical distribution

A random vector $\mathcal{Z} = (Z_1, ..., Z_n)^T$ is said to have a *spherical normal distribution* if it possesses a multivariate normal density function

$$f(t) = (2\pi \det \Gamma)^{-n/2} \exp(-\frac{1}{2}(t - \mu_Z)^T \Gamma^{-1}(t - \mu_Z))$$

with t a column vector, $\mu_Z \in \mathbb{R}^n$ the mean vector and where the $n \times n$ covariance matrix has the form $\Gamma = cE$. Here E is the identity matrix and $c > 0$.

If Λ is an orthogonal matrix and the random vector $\mathcal{Z}$ is spherical normal, then also $\Lambda\mathcal{Z} + b$ is spherical normal for any column vector b, since $\Lambda^T c^{-1} E \Lambda = c^{-1} E$.

Now for Λ we choose the Helmert transformation

$$\Lambda = \begin{pmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdot & \cdot & \cdot & \cdot & \frac{1}{\sqrt{n}} \\ \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \cdot & \cdot & \cdot & \frac{1}{\sqrt{n(n-1)}} & -\frac{n-1}{\sqrt{n(n-1)}} \\ \frac{1}{\sqrt{(n-1)(n-2)}} & \frac{1}{\sqrt{(n-1)(n-2)}} & \cdot & \cdot & \frac{1}{\sqrt{(n-1)(n-2)}} & -\frac{n-2}{\sqrt{(n-1)(n-2)}} & 0 \\ & & & \cdot & \cdot & & \\ & & & \cdot & \cdot & & \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & \cdot & \cdot & \cdot & 0 \end{pmatrix}$$

which is easily seen to be an orthogonal map that sends the diagonal $(1, ..., 1)$ to $(1, 0, ...0)$.

Therefore it has the property, that

$$\Lambda(\mathcal{Z} - (\bar{\mathcal{Z}}, ..., \bar{\mathcal{Z}})^T) = (0, Z_L)$$

with $Z_L \in \mathbb{R}^{n-1}$ and Z_L has zero mean.

From the above it follows that Z_L is spherical normal if Z is.

Now we apply this to our random column vectors of trials $(X_1, ..., X_n)^T, (Y_1, ..., Y_n)^T$ introduced above, where we further assume X, Y to be normal. The trials shall be independent and identically distributed which clearly implies that the random vectors are spherical normal.

Therefore also X_L, Y_L are spherical normal and have zero mean.

Since Λ is a Euclidean isometry we obtain with the notation as above

$$\|X_L\| = \|X_c\|, \quad \|Y_L\| = \|Y_c\|.$$

The standardized variables $X_s = X_c/\|X_c\|, Y_s = Y_c/\|Y_c\|$ therefore correspond to points $X_{L,s} := X_L/\|X_L\|, Y_{L,s} := Y_L/\|Y_L\|$ on the $(n-2)$-dimensional hypersphere S^{n-2}. These points can be interpreted as the directions of the vectors X_L, Y_L. The correspondence is isometric with respect to the Euclidean inner product on the ambient $\mathbb{R}^{n-2}$, meaning that

$$\langle X_{L,s}, Y_{L,s} \rangle = \langle X_s, Y_s \rangle.$$

The projected spherical normal distribution for X_L and Y_L on S^{n-2} is the uniform distribution. To see this, introduce polar coordinates and observe that the density measure for the distribution is proportional to

$$\exp(-\frac{1}{2}(t_2^2 + ... + t_n^2))dt_2...dt_n = r^{n-2}\exp(-\frac{r^2}{2})dr\, d\theta_1...d\theta_{n-2}.$$

Integration over r gives the projected distribution. The corresponding density is constant, meaning that $X_{L,s}, Y_{L,s}$ are uniformly distributed on S^{n-2}.

3.2.2 Geometrical testing

Assume that X, Y are bivariate. The Nullhypothesis H_0 states, that X and Y are uncorrelated which in this case means the same as being independent.

Since $X_{L,s}, Y_{L,s}$ are uniformly distributed on S^{n-2} the probability distribution for the angle θ they enclose is clearly given by a ratio of volumes

$$P(|\theta| \leq t | H_0) = 2\frac{\mathrm{Vol}(S_t)}{\mathrm{Vol}(S^{n-2})}$$

where

$$S_t = \{(t_1, ..., t_{n-1}) \in S^{n-2} | \cos t \leq t_{n-1} \leq 1\}$$

is the upper spherical cap corresponding to the angle t as shown in Fig. 2.

Because of rotational symmetry the center of the cap can be chosen arbitrarily.

To compute its volume we use the standard parametrization

$$\pi_{n-2} : (\phi_1, ...\phi_{n-2}) \to \mathbb{R}^{n-1}$$

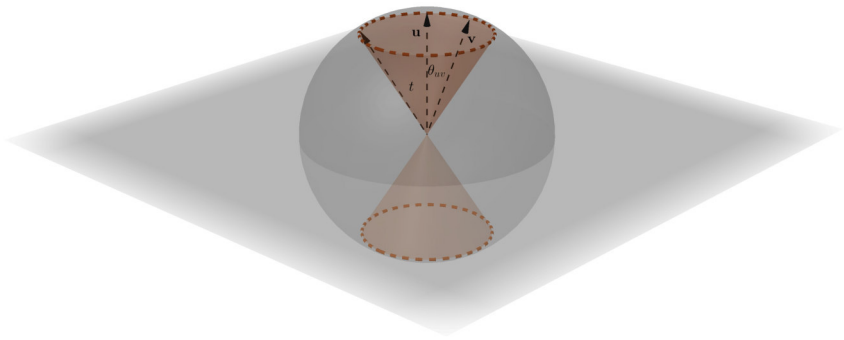

Figure 2: Spherical cap around u

given inductively by $\pi_1 := \begin{pmatrix} \sin(\phi_1) \\ \cos(\phi_1) \end{pmatrix}$ and

$$\pi_l(\phi_1, ..., \phi_l) = \begin{pmatrix} \pi_{l-1}(\phi_1, ..., \phi_{l-1}) \sin(\phi_l) \\ \cos(\phi_l) \end{pmatrix}$$

where $\phi_1 \in [0, 2\pi[$ and $\phi_j \in [0, \pi]$ for $1 < j \leq l$. The Gram determinant gives

$$g = (\sin^{n-3}(\phi_{n-3}) \sin^{n-4}(\phi_{l-4}) ... \sin(\phi_1))^2$$

and one obtains the well known formula

$$\mathrm{Vol}(S^{n-2}) = \int_0^{\pi} ... \int_0^{2\pi} \sin^{n-3}(\phi_{n-3}) \sin^{n-4}(\phi_{n-4}) ... \sin(\phi_1) d\phi_1 ... d\phi_{n-3} = \frac{2\pi^{(n-1)/2}}{\Gamma(\frac{n-1}{2})}.$$

Further for the spherical cap

$$\mathrm{Vol}(S_t) = \int_0^{t} \int_0^{\pi} .... \int_0^{2\pi} \sin^{n-3}(\phi_{n-3}) \sin^{n-4}(\phi_{n-4}) ... \sin(\phi_1) d\phi_1 ... d\phi_{n-3}$$

which can be written in the form

$$\mathrm{Vol}(S_t) = \mathrm{Vol}(S^{n-3}) \int_0^{t} \sin^{n-3}(\phi) d\phi.$$

Summarizing these considerations leads to

$$P(|\theta| \leq t | H_0) = \frac{\Gamma(\frac{n-1}{2})}{\sqrt{\pi}\Gamma(\frac{n-2}{2})} (\int_0^{t} + \int_{\pi-t}^{\pi}) \sin^{n-3}(\phi) d\phi. \tag{4}$$

We are going to link this with the Student's t-distribution and recall the relation $r_{xy} = \cos(\theta_{xy})$ which yields

$$\frac{r_{xy}}{\sqrt{1 - r_{xy}^2}} = \cot(\theta_{xy}).$$

With this in mind, substitute $u = \sqrt{n-2}\cot(\phi)$ in the integral of Eq.(4), so

$$du = -\sqrt{n-2}(1 + \frac{u^2}{n-2})d\phi$$

and we obtain

$$-\sqrt{n-2}\sin^{n-3}(\phi)d\phi = (1 + \frac{u^2}{n-2})^{\frac{1-n}{2}}du. \tag{5}$$

Recall that $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ and so the distribution in (4) is revealed to be the t- distribution.

Alternative assumptions on the distribution on X and Y For the classical t-test on an existing correlation between X and Y one generally assumes that they are bivariate normally distributed. Yet in this situation uncorrelated and independent, means the same. So in the end, we are testing for independence.

From the above it further becomes apparent, that alternative we may impose the condition that their standardizations (directions) are uniformly distributed and perform a test on independence of the directions.

4 Multivariate case

In the multivariate case we meet the situation, that the set of variables is separated into two subsets of independent and dependent variables. We shall confine ourselves with the case of one dependent random variable Y and k independent random variables $X^1, ..., X^k$. We write $\mathbf{X} = (X^1, ..., X^k)^T$ for the column vector of independent variables.

When taking samples $x^j = (x_1^j, ..., x_n^j)^T$, $(y_1, ..., y_n)^T$ of size n the random variables Y_i, X_i^j again read out the corresponding values of the i-th trial. In the sequel we shall confine ourselves to the case of one dependent variable.

4.1 Multiple correlation coefficient

The multiple correlation coefficient $\mathbf{R} \geq 0$ is by definition built from the bivariate ones ρ_{X^iY} by the formula

$$\mathbf{R}^2 = \rho_{XY}^T \mathbf{R}_X^{-1} \rho_{XY} \tag{6}$$

where $\mathbf{R}_X$ is the correlation matrix

$$\mathbf{R}_X = (\rho_{X^iX^j}) \tag{7}$$

and ρ_{XY} is the vector

$$\rho_{XY} = (\rho_{X^1Y}, ..., \rho_{X^kY}). \tag{8}$$

The corresponding multiple sample correlation coefficient is given by

$$R^2 = r_{xy}^T R_x^{-1} \rho_{xy} \tag{9}$$

where R_x is the sample correlation matrix

$$R_x = \begin{pmatrix} 1 & r_{x^1x^2} & . & . & r_{x^1x^k} \\ r_{x^2x^1} & 1 & . & . & r_{x^2x^k} \\ . & . & . & . & . \\ . & . & . & . & . \\ r_{x^kx^1} & r_{x^kx^2} & . & . & 1 \end{pmatrix}. \tag{10}$$

and r_{xy} is the vector

$$r_{xy} = (r_{x^iy}). \tag{11}$$

4.2 Multiple linear regression

A linear model in this case has the form

$$Y \sim a + \sum_{i=1}^{k} b_j X^j = a + \mathbf{X}^T b.$$

with column vector $b = (b_1, ..., b_k)^T$. Again linear regression is concerned with finding an optimal linear model on the basis of some sample of size n with values

$$y = (y_1, ..., y_n)^T, x^j = (x_1^j, ..., x_n^j)^T, \quad j = 1, ..., k.$$

Write $\mathbf{x}$ for the data matrix

$$\mathbf{x} = (x^1, ..., x^k).$$

Linear regression chooses the least square method for the optimal linear model, which means that one has to find the minimum of

$$D(a, b) = \sum_{j=1}^{n} (y_j - (a + b_1 x_j^1 + ... + b_k x_j^k))^2.$$

Using vector notation this takes the form

$$D(a, b) = \|y - (a + \mathbf{x}b)\|^2$$

with $a = (a, .., a)^T$, $b = (b_1, ..., b_k)^T$ and $\text{argmin} D(a, b)$ are the parameters for the optimal linear model

$$Y \sim a + \mathbf{X}^T b.$$

It is not hard to solve this optimization problem by use of some standard calculus. We shall not give the formulas here, but derive the equivalent ones for the standardized variables as in the previous section.

4.3 Standardization

Just as in the bivariate case it will not only simplify the formulas but also provide more insight into the underlying geometric principles when we work with the standardized

sample variables η_j, ξ_j^i with corresponding vectors η, ξ^i and matrix $\boldsymbol{\xi} = (\xi^1, ..., \xi^k)$. We are looking for

$$\text{argmin}(\Delta(\alpha, \beta)) = \text{argmin}(\|\eta - (\alpha + \boldsymbol{\xi}\beta)\|^2)$$

with $\alpha = (\alpha, .., \alpha)^T$, $\beta = (\beta_1, ..., \beta_k)^T$.

In this case the calculations are less involved. The reason for this comes from the fact, that by construction the standardized variables have zero mean

$$\sum_{i=1}^{n} \xi_i^j = \sum_{i=1}^{n} \eta_i = 0$$

and norm $\|\xi^i\| = \|\eta\| = 1$. As a consequence

$$\langle \eta, \alpha \rangle = 0, \quad \boldsymbol{\xi}^T \alpha = 0.$$

We use this in the following computation:

$$\|\eta - (\alpha + \boldsymbol{\xi}\beta)\|^2 \;\; = \;\; 1 + \|\boldsymbol{\alpha}\|^2 - 2\langle \eta, \boldsymbol{\xi}\beta \rangle + \|\boldsymbol{\xi}\beta\|^2.$$

The minimal value is obtained for $\alpha = 0$. In order to determine $\text{argmin}(-2\langle \eta, \boldsymbol{\xi}\beta \rangle + \|\boldsymbol{\xi}\beta\|^2)$ the necessary condition

$$\nabla(-2\langle \eta, \boldsymbol{\xi}\beta \rangle + \|\boldsymbol{\xi}\beta\|^2) = 2(-\boldsymbol{\xi}^T\eta + \boldsymbol{\xi}^T\boldsymbol{\xi}\beta) = \mathbf{0}. \tag{12}$$

leads to

$$\beta = R_x^{-1} r_{xy} \tag{13}$$

where $r_{xy} = \boldsymbol{\xi}^T\eta$ is the vector in (11) and we use the fact, that the sample covariance matrix R_x in (10) can be written as

$$R_x = \boldsymbol{\xi}^T\boldsymbol{\xi}.$$

The linear regression model based on the the standardized samples reads

$$Y_s \sim \mathbf{X}_s R_x^{-1} r_{xy}. \tag{14}$$

In order to obtain the original linear model and the hyperplane of regression for the non-standardized variables one may again substitute the defining expressions for the standardized variables into this formula.

4.4 Geometric interpretation

We shall visualize the case where we have two samples x^1, x^2. In what follows we already assume that the Helmert transform Λ has been applied to the standardized variables and by abuse of notation we shall omit the index L.

The perfect linear correlation for the standardized variables

$$\eta = \beta_1\xi^1 + \beta_2\xi^2$$

would mean, that η lies in the plane E spanned by ξ^1 and ξ^2. If this is not the case the distance between η and E, which means orthogonal projection of η onto E, may serve as a natural measure for the deviance from perfect correlation.

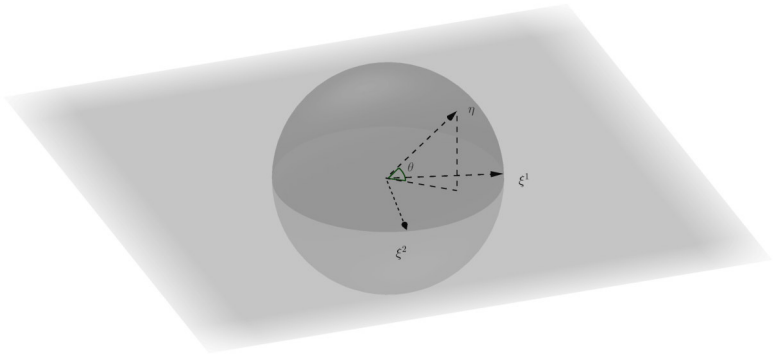

Figure 3: Orthogonal projection of η onto E

For more than two independent variables everything works in complete analogy. (Again we omit the index L.) We consider the orthogonal projection of η onto the flat spanned by $\xi^1, ..., \xi^k$. Of course this amounts to finding β_i which minimize $\|\eta - \beta_1\xi^1 + ... + \beta_k\xi^k\|^2$. This of course is the same thing as least square optimization and we are looking for the same β as in (12) to obtain the orthogonal decomposition

$$\eta = \boldsymbol{\xi}\beta + (\eta - \boldsymbol{\xi}\beta).$$

Let θ denote the angle between η and the flat. Then

$$\|\eta - \boldsymbol{\xi}\beta\| = \sin(\theta)$$

which gives us the Euclidean distance between η and E.

A computation gives:

$$\sin^2(\theta) = \langle \eta - \boldsymbol{\xi}\beta, \eta - \boldsymbol{\xi}\beta\rangle = \langle \eta, \eta - \boldsymbol{\xi}\beta\rangle = 1 - \langle \eta, \boldsymbol{\xi}\beta\rangle$$

and this implies

$$\cos^2(\theta) = 1 - \sin^2(\theta) = \langle \eta, \boldsymbol{\xi}\beta\rangle = r_{xy}^T R_{\mathbf{x}}^{-1} r_{xy}$$

by means of (13) and (11).

The rhs of the above equation corresponds to the square R^2 of the usual multiple correlation coefficient as defined in (9).

So we again see, that there is a very natural geometric interpretation of the multiple correlation coefficient in terms of the angle between η and the flat spanned by the ξ^i.

We summarize the above and obtain in complete analogy to the bivariate case:

1. **The multiple sample correlation coefficients R measures the angle between one standardized sample and the flat spanned by the others.**

2. **Linear regression turns out to be the orthogonal projection of one standardized samples onto the flat spanned by the others.**

In most textbooks on statistics the square R^2 of the multiple correlation coefficient, also called the coefficient of determination, is explained due to a decomposition of variance into an explained and an unexplained part. On the basis of this analysis of variance (ANOVA) one can perform an F-test on significance.

There is a geometrical derivation for the distribution of R^2 which we want to present in the next chapter. In [12], 28.29 a similar approach is sketched.

5 Multiple correlation test

As just mentioned, the classical significance test for the coefficient of determination uses the F-distribution, which is the quotient of two χ^2-distributions for the explained and unexplained variance, respectively.

5.1 Geometrical testing

Geometrically the situation is depicted in figure 3:

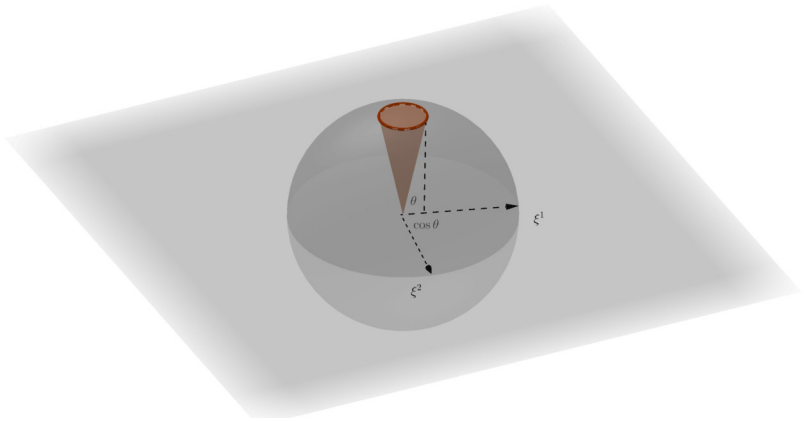

Figure 4: Spherical cap

Let for any vector η denote $\theta_\eta \in [0, \pi/2]$ the angle between η and the flat spanned by $\xi^1, .., \xi^k$. If $\theta_\eta \geq \theta$ then for the corresponding coefficient of determination

$$R^2 \leq \cos^2 \theta \tag{15}$$

We assume that X^j, Y are multinormal. Then as in the bivariate case, the corresponding directions $X^j_{L,s}, Y_{L,s}$ are uniformly distributed. We want to derive the distribution for R^2 under $H_0 : \mathbf{R^2} = 0$. Then as in the bivariate case, all variates are pairwise directional independent.

We have to compute the conditional probability $P(R^2 \geq t|H_0)$. In order to make this explicit, we assume w.k.o.g. that the flat spanned by $\xi^1, .., \xi^k$ is spanned by the unit vectors $e_{n-k}, ..., e_{n-1}$. For a fixed angle θ we denote by C^{n-2}_θ the region of all $s = (s_1, ..., s_{n-1})^T \in S^{n-2}$ such that

$$\sum_{i=n-k}^{n-1} s_i^2 \geq \cos^2(\theta).$$

Now $P(R^2 \geq \cos^2(\theta)|H_0)$ corresponds to the ratio of volume of C^{n-2}_θ and the volume of the hypersphere. In order to compute the volume, we choose an appropriate parametrization of S^{n-2}.

For the hyperspheres S^l we already introduced the standard parametrizations π_l. Furthermore for $m, l \geq 1$ we consider a map $\rho_{m,l} : S^m \times S^l \times [0, \pi/2] \mapsto S^{m+l+1}$ which assigns to (P, Q, τ) the point $(P\sin(\tau), Q\cos(\tau))$ and gives rise to another parametrization for S^{m+l+1}:

$$\pi_{m,l}(\phi_1, ..., \phi_m, \psi_1, ..., \psi_l, \tau) = \begin{pmatrix} \pi_m(\phi_1, ..., \phi_m)\sin(\tau) \\ \pi_l(\psi_1, ..., \psi_l)\cos(\tau) \end{pmatrix}$$

with Gram determinant

$$g = (\sin^{m-1}(\phi_{m-1})...\sin(\phi_1))^2(\sin^{l-1}(\psi_{l-1})...\sin(\psi_1))^2\sin(\tau)^{2l}\cos(\tau)^{2m}$$

We use this parametrization with $l = k - 1$ where k corresponds to the k independent random variables and $m = n - 3 - l = n - k - 2$.

$$\begin{aligned} P(R^2 \geq \cos^2(\theta)|H_0) = \frac{\mathrm{Vol}(C_\theta^{n-2})}{\mathrm{Vol}(S^{n-2})} &= \frac{\mathrm{Vol}(S^{k-1})\mathrm{Vol}(S^{n-k-2})}{\mathrm{Vol}(S^{n-2})}\int_0^\theta \sin(\tau)^{k-1}\cos(\tau)^{n-k-2}d\tau \\ &= \frac{2\Gamma(\frac{n-1}{2})}{\Gamma(\frac{k}{2})\Gamma(\frac{n-k-1}{2})}\int_0^\theta \sin(\tau)^{k-1}\cos(\tau)^{n-k-2}d\tau. \end{aligned}$$

5.2 Comparison with the classical F-test

Assuming H_0 the test-statistic $F = \frac{R^2}{1-R^2}\frac{k}{n-k-1}$ is known to be $F(n-k-1, k)$-distributed, which has density

$$g_{n-k-1,k}(x) = \frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{k}{2})\Gamma(\frac{n-k-1}{2})}k^{\frac{k}{2}}(n-k-1)^{\frac{n-k-1}{2}}\frac{x^{\frac{n-k-1}{2}-1}}{(k+(n-k-1)x)^{\frac{n-1}{2}}}.$$

It is an easy exercise in calculus, that the substitution $\tau \to x = \cot^2(\tau)\frac{k}{n-k-1}$ in the upper integral yields

$$P(R^2 \geq \cos^2(\theta)|H_0) = \int_{F_\theta}^\infty g_{n-k-1,k}(x)dx = P(F \geq F_\theta)$$

with $F_\theta = \cot^2(\theta)\frac{k}{n-k-1}$.

6 T-test, ANOVA and linear regression

In applied statistics the probably most frequent test situations consists in testing whether the mean μ_X of some random variable X equals a fixed μ_0 or if the means of k random variable are equal.

Here we take a sample $x^j = (x_1^j, ..., x_{n_j}^j)^T$ of size n_j corresponding to each random variable X^j and call it a group. In applications the random variables measure the same quantity, say weight, but within different groups, say male, female.

If $k = 2$ one usually distinguishes between paired and unpaired samples and if $k > 2$ there is a balanced and an unbalanced design. For $k \leq 2$ a t-test is applied, while for $k > 2$ an analysis of variance (ANOVA) is performed which uses an F-test for the quotient of between-group and within-group variance. In a slightly more general setting we distinguish between one and more-factorial ANOVA. Here a factor causes the sample

to be split into groups, corresponding to the levels of the factor. If there are more factors, the combination of levels just causes the existence of more different groups.

In this section we want to show that t-test and also ANOVA are special cases of a significance test for correlation or coefficient of determination after introducing categorial variables.

6.1 Two sample t-test

We assume that we have two samples $\{x_i\}$ and $\{y_j\}$ of size n_x and n_y, resp.. It is assumed that the corresponding random variables are both normal with the same unknown standard deviation σ and means μ_X, μ_Y. If one is willing to test whether their means are different, we want to reject

$$H_0 : \quad \mu_X = \mu_Y.$$

For this one introduces the test statistic (see e.g. [1], 14.6.1)

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}} \sqrt{\frac{(n-2)n_y n_x}{n}} \tag{16}$$

which is known to have a Student's t-distribution with $n-2$ degrees of freedom. Here as before $\bar{x}, \bar{y}$ are the sample means and s_x^2, s_y^2 the sample variances.

We compare this with a sample correlation test, where we form the union of the two samples $\{z_k\} = \{x_i\} \cup \{y_j\}$ and introduce the categorial (characteristic) variable C which is 0 for those k where z_k belongs to the first sample and 1 else. As before c is the corresponding sample variable.

The null hypothesis $H_0 : \mu_X = \mu_Y$ implies that X and Y are identically distributed and therefore the sample $z = (x_1, ..., x_{n_x}, y_1, ..., y_{n_y})^T$ can be viewed as $n_x + n_y$ observations for the same normally distributed random variable Z. So clearly Z and C are independent under H_0

Let $\zeta = \frac{z - \bar{z}}{\sqrt{n-1}s_z}$ denote the standardized variable for z and $\gamma = \frac{c - \bar{c}}{\sqrt{n-1}s_c}$.

An elementary calculation shows, that $\sqrt{n-1}\, s_c = \sqrt{n_x n_y / n}$ and the sample correlation coefficient equals

$$r_{zc} = \langle \zeta, \gamma \rangle = \frac{(\sum_{k=1}^{n} z_k c_k) - n_y \bar{z}}{\sqrt{n-1}\, s_z \sqrt{\frac{n_x n_y}{n}}}.$$

Of course

$$\sum_{k=1}^{n} z_k c_k = n_y \bar{y}$$

and

$$n_y \bar{z} = \frac{n_y}{n} (\sum_{i=1}^{n_x} x_i + \sum_{j=1}^{n_y} y_j) = \frac{n_y}{n} (n_x \bar{x} + n_y \bar{y})$$

hence

$$r_{zc} = \frac{n_y(\bar{y} - \bar{z})}{\sqrt{n-1}\, s_z \sqrt{\frac{n_x n_y}{n}}} = \frac{\frac{n_y n_x}{n}(\bar{y} - \bar{x})}{\sqrt{n-1}\, s_z \sqrt{\frac{n_x n_y}{n}}}.$$

Further

$$\begin{aligned}\sum_{i=1}^{n_x}(x_i-\bar{z})^2 &= \sum_{i=1}^{n_x}(x_i-\bar{x}+\frac{n_y}{n}(\bar{x}-\bar{y}))^2\\ &= \sum_{i=1}^{n_x}(x_i-\bar{x})^2+2\frac{n_y}{n}(\bar{x}-\bar{y})\sum_{i=1}^{n_x}(x_i-\bar{x})\\ &\quad +n_x\left(\frac{n_y}{n}(\bar{x}-\bar{y})\right)^2\\ &= \sum_{i=1}^{n_x}(x_i-\bar{x})^2+\frac{n_x n_y^2}{n^2}(\bar{y}-\bar{x})^2,\end{aligned}$$

and similarly

$$\sum_{j=1}^{n_y}(y_j-\bar{z})^2=\sum_{j=1}^{n_y}(y_j-\bar{y})^2+\frac{n_x^2 n_y}{n^2}(\bar{y}-\bar{x})^2.$$

Therefore

$$(n-1)s_z^2=\sum_{i=1}^{n_x}(x_i-\bar{z})^2+\sum_{j=1}^{n_y}(y_j-\bar{z})^2=(n_x-1)s_x^2+(n_y-1)s_y^2+\frac{n_x n_y}{n}(\bar{y}-\bar{x})^2.$$

Eventually we arrive at

$$r_{zc}=\sqrt{\frac{n_x n_y}{n}}\frac{\bar{x}-\bar{y}}{\sqrt{(n_x-1)s_x^2+(n_y-1)s_y^2+\frac{n_x n_y}{n}(\bar{y}-\bar{x})^2}}. \tag{17}$$

For a significance test with $H_0:\rho_{ZC}=0$ we choose as above the statistic

$$\frac{r_{zc}}{\sqrt{1-r_{zc}^2}}\sqrt{n-2},$$

and by an elemtary computation from (17) we obtain that this statistic equals the statistic t from the t-test (16):

$$\frac{r_{zc}}{\sqrt{1-r_{zc}^2}}\sqrt{n-2}=\sqrt{\frac{(n-2)n_y n_x}{n}}\frac{\bar{x}-\bar{y}}{\sqrt{(n_x-1)s_x^2+(n_y-1)s_y^2)}} \tag{18}$$

So in the end, we see, that the above t-test and the (dichotomic) correlation test turn out to be equivalent.

Remark. Also the one sample t-test can be treated as a correlation test. Here we have for a random variable X one sample $\{x_i\}$ of size n and the null hypothesis $H_0:\mu=\mu_0$. In this case we proceed as in the two sample case with an additional "constant" sample $y_j\equiv\mu_0, j=1,...,n$. The same calculations as above give

$$\frac{r_{zc}}{\sqrt{1-r_{zc}^2}}\sqrt{n-2}=\sqrt{n}\,\frac{\bar{x}-\mu_0}{s_x}$$

which coincides with the test statistic for the one sample test, see e.g.[1], 14.4.

6.2 ANOVA

When more than two random variables groups shall be compared then as described above ANOVA generalizes the t-test. For k random normal variables X^j, all assumed to have the same stadard deviation, we have a sample x^j of size n_j, $j = 1, ..., k$. We want to test whether all means μ_{X^j} are equal,

$$H_0 : \quad \mu_1 = ... = \mu_k.$$

The test statistic is given by

$$F = \frac{SQA}{SQR}\frac{(n-k)}{(k-1)},$$

see e.g. [7], p.182, where

$$SQA = \sum_{i,j} n_j(\bar{y}^j - \bar{z})^2, \quad SQR = \sum_j (n_j - 1)s_{y^j}^2.$$

It follows an $F_{k-1,n-k}$-distribution. Let $SQT := (n-1)s_z^2$ then it is easy to verify that

$$SQT = SQR + SQA. \tag{19}$$

We shall now see, that **ANOVA is the same as the significance test for multiple correlation** as done above.

For this we form the union of all samples $\{z_l\} = \bigcup_{i,j}\{y_i^j\}$ and introduce k categorial variables C^j which are just the characteristic variables for the k groups and c_l^j be the corresponding sample variables, i.e. $c_l^j = 1$ if $z_l \in \{y_1^j ..., y_{n_j}^j\}$ and zero else.

Remark. For our linear regression we actually only need the first $k-1$ characteristic variables $C^1, ..., C^{k-1}$ since the elements of the k-th group are automatically identified as those elements for which all other characteristic variables vanish.

Under H_0 all X^j are identically distributed. We may now perform multiple correlation analysis with $z = (z_1, ..., z_n)^T$ the sample for the dependent variable Z and C^j, $j = 1, .., k-1$ are the independent variables. Our claim will follow from

Lemma 1. *Let R^2 be the coefficient of determination for the multiple linear regression model*

$$Z \sim a + \sum_{j=1}^{k-1} b_j C^j.$$

Then

$$1 - R^2 = \frac{SQR}{SQT}.$$

Proof. First observe that $\sum_{j=1}^k C^j = 1$, where as before 1 shall denote the constant vector. Therefore we can rewrite the model as

$$Z \sim \sum_{j=1}^{k} \alpha_j C^j.$$

Let $\zeta = \frac{z-\bar{z}}{\sqrt{n-1}s_z}$ denote the standardized variable for z and $\gamma^j = \frac{c^j-\bar{c}^j}{\sqrt{n-1}s_{c^j}}$, then we find from our previous calculations that

$$1 - R^2 = \min_{\beta_j} \|\zeta - \sum_{j=1}^{k-1} \beta_j \gamma^j\|^2.$$

Re-substitute the standardized variables, then with appropriate α_j

$$\|\zeta - \sum_{j=1}^{k-1} \beta_j \gamma^j\| = \frac{1}{\sqrt{n-1}s_z} \|z - \sum_{j=1}^{k} \alpha_j c^j\|$$

and therefore

$$SQT(1 - R^2) = \min_{\alpha_j} \|z - \sum_{j=1}^{k} \alpha_j c^j\|^2.$$

Remark, that in the latter sum the k-th sample variable is reintroduced by means of the relation $1 = \sum_{j=1}^{k} c^j$.

The minimum is attained, if α_j are chosen such that $\sum_{j=1}^{k} \alpha_j c^j$ is the orthogonal projection of z onto the flat spanned by the c^j, so we have to choose α_j subject to the conditions

$$\langle z - \sum_{j=1}^{k} \alpha_j c^j, c^i \rangle = 0$$

for all i. Since $\langle c^i, c^j \rangle = \delta_{ij} n_i$ this gives $\alpha_j = \bar{y}^j$ and hence

$$SQT(1 - R^2) = \|z - \sum_{j=1}^{k} \bar{y}^j c^j\|^2 = SQR.$$

This proves the lemma. □

Now by (19) we deduce

$$R^2 = \frac{SQA}{SQT}$$

and therefore

$$F = \frac{R^2}{1 - R^2} \frac{n-k}{k-1}.$$

So we see that the test statistic F from ANOVA is identical to the test statistic of the significance test for the corresponding multiple correlation coefficient.

Conclusion. As we have seen that t-test and variance analysis can be explained as correlation tests they eventually can be given a geometric meaning. This may perhaps help to explain the underlying concepts for the tests more intuitively.

References

[1] BAMBERG, Günter; BAUR, Franz; KRAPP, Michael. Statistik, 18. Auflage. Oldenburg, 2017.

[2] BILIN ZENG, Kang Chen; WANG, Cong. Geometric views of partial correlation coefficient in regression analysis. International Journal of Statistics and Probability, 2017, 6. Jg., Nr. 3.

[3] BOSCH, Karl. Statistik-Taschenbuch, 3.Auflage. Oldenburg, 1998

[4] CHANCE, William A. A geometric derivation of the distribution of the correlation coefficient $|r|$ when $\rho = 0$. The American Mathematical Monthly, 1986, 93. Jg., Nr. 2, S. 94-98.

[5] DEMESHEV, Boris; GNILOVA, Olya. How Gauss and Markov Met Pythagoras: Geometry in Econometrics. 2018.

[6] FISHER, Ronald A. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika 10.4 ,1915, 507-521.

[7] KRENGEL, Ulrich. Einführung in die Wahrscheinlichkeitstheorie und Statistik. Braunschweig: Vieweg, 1988.

[8] LANGSRUD, Øyvind. The geometrical interpretation of statistical tests in multivariate linear regression. Statistical Papers, 2004, 45. Jg., Nr. 1, S. 111-122.

[9] RAHMAN, Najeeb Abdur. A Course in Theoretical Statistics, Charles Griffin and Company, 1968

[10] SAVILLE, David J.; WOOD, Graham R. Statistical methods: The geometric approach. Springer Science and Business Media, 2012.

[11] SMALL, Christopher G. The statistical theory of shape. Springer Science and Business Media, 2012.

[12] STUART, Alan; ORD, John Keith; ARNOLD, Steven F. Kendall's Advanced Theory of Statistics: Classical Inference and the Linear Model. Volume 2A. John Wiley, 2004.

[13] WOOD, Graham R.; SAVILLE, David J. A new angle on the t-test. Journal of the Royal Statistical Society: Series D (The Statistician), 2002, 51. Jg., Nr. 1, S. 99-104.

Bisher erschienene Research Papers:

Nr. 1 | Februar 2017
Berlemann, Michael; Christmann, Robin.
The Role of Precedents on Court Delay.
Evidence from a Civil Law Country.

Nr. 2 | September 2017
Matthes, Roland.
A note on the Saito-Kurokawa lift for Hermitian forms.

Nr. 3 | Februar 2018
Christmann, Robin.
Prosecution and Conviction under Hindsight Bias in Adversary Legal Systems.

Nr. 4 | März 2019
Broere, Mark; Christmann, Robin.
Takeovers, Shareholder Litigation, and the Free-riding Problem

Nr. 5 | März 2019
Matthes, Roland.
Vector-valued Cusp Forms and Orthogonal Modular Forms

Nr. 6 | März 2020
Matthes, Roland.
A Geometric View on Linear Regression and Correlation Tests

Nr. 7 | März 2020
Matthes, Roland.
A Note on the Geometry of Partial Correlation and the Grassmannian

Leibniz-Fachhochschule
Expo Plaza 11
30539 Hannover

leibniz-fh.de