LINEAR MODELS IN TEN MINUTES

DANIEL V. TAUSK

Introductory statistics textbooks are annoying to people with a reasonable math background because everything is dumbed down for people with almost no math background. For example, while studying such texts, sometimes its hard to figure out what is really going on amidst so many ugly matrix formulas. So if you have some math background (which here means mostly some linear algebra and measure theory), let's learn the abstract theory of linear models in ten minutes. I consider only the linear models admitting a simple exact inference theory, which are the linear models with a Gaussian response variable whose variance is known up to a multiplicative constant. Those include multiple linear regression, t tests and ANOVAs (but not mixed models). Proofs will be somewhat sketchy so main ideas are not lost and I can keep my "ten minutes" promise. Some extra details are given in footnotes which can be mostly skipped in a first reading.

We assume the reader might have very little familiarity with probability theory and statistics, so we start with a lightning course on what is needed from probability theory (Section 1) and on Gaussian random variables and random vectors (Section 2). In Section 3 we then present the abstract theory of linear models. These first three sections is what should take about ten minutes of learning. The connections with concrete applications (t tests, linear regression, etc) are given in Section 4 and this could take a few more minutes. Finally, in Appendix A we give a short rigorous presentation of conditional probability and we use that formalism to properly present the theory of linear models with a random explanatory variable.

1. Very basic probability theory in five minutes

We consider a fixed *probability space*, i.e., a measure space in which the total measure is equal to 1. The measurable subsets of the probability space are called *events*. We don't need to give a name to the fixed probability space because we won't need to talk about it much — it's just there. What matters in probability theory are the random objects and their distributions. By a *random object* X we mean a measurable function defined on the fixed probability space and taking values in some measurable space (i.e., a set endowed with a σ-algebra); we call X a *random variable* if it takes values in the real numbers and a (V-valued) *random vector* if it takes values in a real finite-dimensional vector space V. Both $\mathbb{R}$ and V are endowed with their respective Borel σ-algebras.

Date: July 1st, 2024.

The probability $\mathbb{P}(X \in A)$ that a random object X belongs to a measurable subset A of its counterdomain is understood as the probability (i.e., the measure) of the event $[X \in A]$ which is defined as the inverse image $X^{-1}[A]$. The *distribution* of a random object X is defined as the probability measure on its counterdomain given by the image under X of the probability measure on its domain, i.e., it is the map that associates $\mathbb{P}(X \in A)$ to every measurable subset A of the counterdomain of X.

We can form new random objects as functions of other random objects in the following way: if f is a measurable function defined on the counterdomain of a random object X, we define a new random object $f(X)$ by setting $f(X) = f \circ X$. Obviously the distribution of a measurable function of a random object X depends only on the distribution of X.

A *probability density* for a random object X with respect to some positive σ-finite measure ν on its counterdomain is defined as the Radon–Nikodym derivative of the distribution of X with respect to ν, assuming that such distribution is absolutely continuous with respect to ν. A family of random objects is said to be *independent* if the distribution of the family (seen as a single random object taking values in the product measurable space) is the product measure of the distributions of the individuals of the family. Clearly, measurable functions of independent random objects are again independent random objects.

The *expected value* (or simply *mean*) of an integrable random variable X, denoted $E(X)$, is defined as the integral of X with respect to the probability measure on its domain. Note that the expected value of X or of any real-valued function of X depends only on the distribution of X as it is equal to the integral of such a function with respect to the distribution of X. The *covariance* $\mathrm{Cov}(X, Y)$ of two square-integrable random variables X and Y is defined as the L^2-inner product of their expectation free components[1] or, more explicitly:

$$\mathrm{Cov}(X, Y) = E\big[\big(X - E(X)\big)\big(Y - E(Y)\big)\big] = E(XY) - E(X)E(Y).$$

If X and Y are independent random variables then Fubini's Theorem yields $E(XY) = E(X)E(Y)$ and thus $\mathrm{Cov}(X, Y) = 0$. The *variance* of a square-integrable random variable X is defined by $\mathrm{Var}(X) = \mathrm{Cov}(X, X)$ and the *standard deviation* of X is defined as the square root of $\mathrm{Var}(X)$. Note that the standard deviation is just the semi-norm associated to the positive semi-definite inner product Cov on the space of square-integrable random variables. We have $\mathrm{Var}(X) = 0$ if and only if X is constant almost everywhere.

If X is an integrable V-valued random vector for some real finite-dimensional vector space V, we again define the expected value (or mean) $E(X)$

[1]Note that $E(X)$ is just the L^2-orthogonal projection of X onto the one-dimensional subspace of constant maps and that the expectation free component $X - E(X)$ of X is the L^2-orthogonal projection of X onto the hyperplane consisting of random variables with zero expectation.

of X as the integral of X with respect to the probability measure on its domain, so that now $E(X)$ is an element of V. If X is square-integrable, the variance $\mathrm{Var}(X)$ of X is defined as the positive semi-definite symmetric bilinear form on the dual space V^* given by:

$$\mathrm{Var}(X) : V^* \times V^* \ni (\alpha, \beta) \longmapsto \mathrm{Cov}\big(\alpha(X), \beta(X)\big).$$

Typical statisticians (which only work in coordinates and don't like abstract vector spaces) would call some matrix representation of $\mathrm{Var}(X)$ the *covariance matrix* of X. It is easy to see that the variance of X is degenerate if and only if X takes values with probability 1 in a proper affine subspace of V. If $\mathrm{Var}(X)$ is nondegenerate then it is a legitimate (positive definite) inner product on V^* and hence it induces a linear isomorphism

$$V^* \ni \alpha \longmapsto \mathrm{Var}(X)(\alpha, \cdot) \in V^{**} \cong V \tag{1}$$

from V^* to V which then induces an inner product on V. The matrix that represents the inner product induced on V is the inverse of the matrix that represents the inner product on V^* (assuming one uses on V^* the dual of the basis used on V).

Clearly, if $T : V \to W$ is a linear map between real finite-dimensional vector spaces and X is an integrable V-valued random vector then:

$$E\big(T(X)\big) = T\big(E(X)\big).$$

Moreover, if X is square-integrable then the bilinear map $\mathrm{Var}\big(T(X)\big)$ is the pull-back of the bilinear map $\mathrm{Var}(X)$ by the adjoint $T^* : W^* \to V^*$ of the linear map T, i.e.:

$$\mathrm{Var}\big(T(X)\big) = \mathrm{Var}(X)(T^*\cdot, T^*\cdot). \tag{2}$$

It is often convenient to identify the bilinear map $\mathrm{Var}(X)$ on V^* with the linear map (1) because for linear maps we can write formulas involving compositions that readily translate into matrix multiplications[2] when bases are chosen. Using such identification, equality (2) becomes:

$$\mathrm{Var}\big(T(X)\big) = T \circ \mathrm{Var}(X) \circ T^*. \tag{3}$$

2. Gaussian random vectors in two minutes

A random variable X is said to be *Gaussian* (or *normally distributed*) if it is either constant almost everywhere or if its probability density with respect to the Lebesgue measure is the exponential of a second degree polynomial with a negative leading coefficient. The distribution of a Gaussian random variable X is completely determined by its mean $\mu \in \mathbb{R}$ and standard deviation $\sigma \geq 0$ and its probability density with respect to the Lebesgue measure

[2]However, one should keep in mind that if $B : V \times W \to \mathbb{R}$ is a bilinear form then the standard matrix representation of B is the transpose of the standard matrix representation of the corresponding linear map $V \ni v \mapsto B(v, \cdot) \in W^*$. This problem does not arise, of course, if $V = W$ and B is symmetric.

is given by $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ if $\sigma > 0$. A Gaussian random variable with mean zero and unit variance is called *standard normal*. Since an affine transformation of a Gaussian random variable is Gaussian, it follows that if X is Gaussian with mean μ and standard deviation $\sigma > 0$ then $\frac{X-\mu}{\sigma}$ is standard normal.

Definition 1. For a finite-dimensional real vector space V, we call a V-valued random vector X *Gaussian* (or we say that the distribution of X is *multivariate normal*) if the random variable $\alpha(X)$ is Gaussian for every linear functional $\alpha \in V^*$.

Obviously a linear (or affine) function of a Gaussian random vector is again a Gaussian random vector.

Since two V-valued random vectors have the same distribution if and only if they have the same characteristic function[3], it follows that the distribution of a V-valued random vector X is determined by the distribution of $\alpha(X)$ for all $\alpha \in V^*$ and hence that the distribution of a Gaussian random vector is determined by its mean and variance. Using this observation one checks easily[4] that a V-valued random vector X is Gaussian with nondegenerate variance if and only if its density with respect to a Lebesgue measure[5] is proportional to $f(x) = \exp\left(-\frac{1}{2}\|x-\mu\|^2\right)$, where $\mu \in V$ is the mean of X and $\|\cdot\|$ is the norm associated to the inner product in V induced by the variance of X.

In the following propositions we assume that X is a Gaussian V-valued random vector with a nondegenerate variance, where V is a real finite-dimensional vector space. We also assume V to be endowed with the inner product induced by the variance of X.

Whenever V is a real finite-dimensional vector space endowed with an inner product and W is a subspace of V, we will denote by $P_W : V \to W$ the orthogonal projection onto W.

Proposition 2. *If $V = \bigoplus_{i=1}^k V_i$ is an orthogonal direct sum decomposition then the random vectors $P_{V_i}(X)$, $i = 1, \ldots, k$, are independent.*

---

[3]The *characteristic function* $\varphi_X : V^* \to \mathbb{C}$ of a V-valued random vector X is defined by $\varphi_X(\alpha) = E\big[\exp\big(i\alpha(X)\big)\big]$, for all $\alpha \in V^*$. Up to a multiplicative constant and replacement of α with $-\alpha$, this is just the Fourier transform of the distribution of X seen as a tempered distribution (in the sense of Schwartz).

[4]To see that a random vector with such a probability density is Gaussian with the corresponding mean and variance, write a linear functional $\alpha \in V^*$ as a constant times a coordinate functional corresponding to an orthonormal basis of V with respect to the inner product to which the norm is associated.

[5]By a *Lebesgue measure* on V we mean any measure that corresponds to the Lebesgue measure of $\mathbb{R}^n$ through some linear isomorphism between V and $\mathbb{R}^n$. Equivalently, a Lebesgue measure on V is any locally finite translation invariant measure. Two Lebesgue measures on V are always constant multiples of each other.

Proof. Note that when we identify V with the cartesian product $\prod_{i=1}^k V_i$, the probability density of X with respect to a Lebesgue measure becomes a product of functions of the coordinates of $x = (x_i)_{i=1}^k \in \prod_{i=1}^k V_i$. □

Definition 3. Given a positive integer m, we say that a random variable has a *chi-squared distribution with m degrees of freedom* if it has the same distribution as the sum of m squares of independent standard normal random variables.

Clearly, the expected value of a random variable having a chi-squared distribution with m degrees of freedom is m.

Proposition 4. *If W is a nonzero subspace of V and if the orthogonal projection P_W annihilates the mean of X then the random variable $\|P_W(X)\|^2$ has a chi-squared distribution whose number of degrees of freedom is the dimension of W.*

Proof. Pick an orthonormal basis of V that extends a basis of W and note that the coordinates of X corresponding to the elements of the basis of W are independent and standard normal random variables. Moreover, the squared norm $\|P_W(X)\|^2$ is the sum of the squares of such coordinates of X. □

3. Abstract theory of linear models in five minutes

Let V be a real finite-dimensional vector space and Y be a V-valued Gaussian random vector with nondegenerate variance. We use here the letter Y instead of X because in the context of the theory of linear models the letter X is more often used for the explanatory variables, while here we think of Y as the response variable. We assume that the mean of Y is known to be a linear function of some unknown parameter. More specifically, we consider some real finite-dimensional vector space Θ as a parameter space and some known linear map $L : \Theta \to V$ such that the mean $\mu = E(Y)$ of Y is equal to $L(\theta)$ for some unknown $\theta \in \Theta$. The linear map L is assumed to be injective so that the unknown parameter is identifiable.

In this section we will only present the abstract theory and it will be convenient to use the map L to identify the parameter space Θ with the image of L, so we assume from now on the parameter space to be a subspace W of V and $L : W \to V$ to be the inclusion map. The unknown parameter is then just the mean μ of Y which is assumed to belong to W. We also need to assume that W is a proper subspace of V. When discussing practical applications (notably, in Subsection 4.3) the identification of Θ with W will be revoked as it tends to cause confusion.

The variance of Y is assumed to be known up to a multiplicative constant, so we assume V^* to be endowed with a known inner product $\langle \cdot, \cdot \rangle$ such that the variance of Y is σ^2 times $\langle \cdot, \cdot \rangle$ for some unknown constant $\sigma > 0$. We let V be endowed with the inner product induced by $\langle \cdot, \cdot \rangle$, which will also be denoted by $\langle \cdot, \cdot \rangle$. The inner product on V induced by the variance of

Y is then $\frac{1}{\sigma^2}$ times $\langle \cdot, \cdot \rangle$. We denote by $\| \cdot \|$ the norms corresponding to both inner products $\langle \cdot, \cdot \rangle$. Note that for matters regarding orthogonality and orthogonal projections, a multiplicative constant on the inner product is irrelevant.

We need to recall the definitions of some basic probability distributions.

Definition 5. Given a positive integer m, a random variable is said to have a *Student's t distribution with m degrees of freedom* if it has the same distribution as the quotient

$$\frac{Z}{\sqrt{\frac{U}{m}}},$$

where Z and U are independent random variables such that Z has a standard normal distribution and U has a chi-squared distribution with m degrees of freedom.

Definition 6. Given positive integers m_1 and m_2, a random variable is said to have a *Snedecor's F distribution with degrees of freedom m_1 and m_2* if it has the same distribution as the quotient

$$\frac{\frac{1}{m_1}U_1}{\frac{1}{m_2}U_2},$$

where U_1 and U_2 are independent random variables such that U_1 has a chi-squared distribution with m_1 degrees of freedom and U_2 has a chi-squared distribution with m_2 degrees of freedom.

Denote by $W^\perp$ the orthogonal complement of W in V. Since the mean μ of Y is in W, we have $P_{W^\perp}(\mu) = 0$ and thus Proposition 4 yields that the random variable $\frac{1}{\sigma^2}\|P_{W^\perp}(Y)\|^2$ has a chi-squared distribution with $\dim(W^\perp)$ degrees of freedom. Keeping in mind that the expected value of a chi-squared distributed random variable is equal to the number of degrees of freedom, such observation yields the following result.

Proposition 7. *Under the assumptions above, the random variable S defined by*

$$S = \frac{\|P_{W^\perp}(Y)\|}{\sqrt{\dim(W^\perp)}}$$

is such that S^2 is an unbiased estimator *of σ^2, i.e., σ^2 is the expected value of S^2. Moreover, the random variable*

$$\dim(W^\perp)\frac{S^2}{\sigma^2}$$

has a chi-squared distribution having $\dim(W^\perp)$ degrees of freedom. □

Note that since $\mu \in W$ we have $P_W(\mu) = \mu$ and therefore

$$\hat{\mu} = P_W(Y)$$

is a linear unbiased estimator of the paramater μ, i.e., $\hat{\mu}$ is a linear function of the data Y whose expected value is μ. In a certain sense $\hat{\mu}$ is the best linear unbiased estimator of μ, as we now explain. Let $\alpha \in W^*$ be a linear functional on W. We can think of $\alpha(\mu)$ as a "coordinate" of the unknown mean μ of Y. Given a linear functional $\hat{\alpha} \in V^*$, we have that the expected value of $\hat{\alpha}(Y)$ is $\hat{\alpha}(\mu)$ and thus $\hat{\alpha}(Y)$ is an unbiased estimator of $\alpha(\mu)$ — in the sense that the expected value of $\hat{\alpha}(Y)$ is $\alpha(\mu)$ for every possible $\mu \in W$ — if and only if $\hat{\alpha}$ is an extension of α. Since the variance of $\hat{\alpha}(Y)$ is $\sigma^2\|\hat{\alpha}\|^2$, we have that the linear extension $\hat{\alpha}$ of α such that $\hat{\alpha}(Y)$ has the least variance is the one having the least possible norm. This minimum is clearly attained at the linear extension of α that vanishes on $W^\perp$, i.e., at the linear functional $\hat{\alpha}$ given by $\hat{\alpha} = \alpha \circ P_W$. For such $\hat{\alpha}$, the random variable $\hat{\alpha}(Y) = \alpha(\hat{\mu})$ is known as the *best linear unbiased estimator* (BLUE) of the parameter $\alpha(\mu)$. From now on, $\hat{\alpha}$ always denotes $\alpha \circ P_W$. Clearly

$$\frac{\hat{\alpha}(Y) - \alpha(\mu)}{\sigma\|\hat{\alpha}\|}$$

is a standard normal random variable. Moreover, since $\hat{\alpha}(Y)$ is a function of the orthogonal projection $P_W(Y)$ and S^2 is a function of the orthogonal projection $P_{W^\perp}(Y)$, it follows from Proposition 2 that $\hat{\alpha}(Y)$ and S^2 are independent. The next result then immediately follows from the definition of Student's t distribution using Proposition 7.

Proposition 8. *Under the assumptions above, if $\alpha \in W^*$ and $\hat{\alpha} \in V^*$ is the linear extension of α that vanishes on $W^\perp$ then the random variable*

$$\frac{\hat{\alpha}(Y) - \alpha(\mu)}{S\|\hat{\alpha}\|}$$

has a Student's t distribution whose number of degrees of freedom is the dimension of $W^\perp$. □

Proposition 8 can be used to construct confidence intervals for the parameter $\alpha(\mu)$. Recall that, for $\gamma \in [0, 1]$, a *γ-confidence set* for a certain parameter is a random subset[6] of the parameter space that contains that parameter with probability γ.

Corollary 9. *Under the assumptions of Proposition 8, given $\gamma \in]0, 1[$, we have that*

$$[\hat{\alpha}(Y) - cS\|\hat{\alpha}\|, \hat{\alpha}(Y) + cS\|\hat{\alpha}\|]$$

is a γ-confidence interval for the parameter $\alpha(\mu)$, where $c > 0$ is chosen in such a way that a random variable having a Student's t distribution with $\dim(W^\perp)$ degrees of freedom has a probability γ of belonging to the interval $[-c, c]$. □

[6]If M is a set, we can identify the set $\wp(M)$ of all subsets of M with the product $\{0, 1\}^M$ and this turns $\wp(M)$ naturally into a measurable space. A *random subset* of M is then a $\wp(M)$-valued random object.

Recall that in order to test a certain null hypothesis about an unknown parameter of interest we can define a p-value[7] by considering the probability under the null hypothesis that a certain real-valued *statistic* T (i.e., a measurable function of the observed data Y) has a value greater than or equal to the observed value of T. The statistic T should be chosen in such a way that larger values of T are "more incompatible" with the null hypothesis than smaller values of T and in such a way that the distribution of T under the null hypothesis is known. For example, in order to test the null hypothesis $\alpha(\mu) = 0$, we can let T be the absolute value of

$$\frac{\hat{\alpha}(Y)}{S\|\hat{\alpha}\|}$$

and Proposition 8 tells us that under the null hypothesis the statistic T has the distribution of the absolute value of a Student's t distribution with $\dim(W^\perp)$ degrees of freedom.

In order to test more general null hypotheses about μ we need another statistic.

Proposition 10. *Under the assumptions above, if W_0 is a proper subspace of W containing the mean μ and W_1 denotes the orthogonal complement of W_0 in W then*

$$\frac{\frac{1}{\dim(W_1)}\|P_{W_1}(Y)\|^2}{S^2} \tag{4}$$

has a Snedecore's F distribution whose degrees of freedom are $\dim(W_1)$ *and* $\dim(W^\perp)$.

Proof. Under $\mu \in W_0$ we have $P_{W_1}(\mu) = 0$ and thus $\frac{1}{\sigma^2}\|P_{W_1}(Y)\|^2$ has a chi-squared distribution with $\dim(W_1)$ degrees of freedom by Proposition 4. Moreover, since W_1 and $W^\perp$ are orthogonal, Proposition 2 implies that $P_{W_1}(Y)$ and S^2 are independent. The conclusion then follows from Proposition 7. □

Proposition 10 can be used to test the null hypothesis that the mean μ of Y belongs to a certain proper subspace W_0 of W. Namely, one simply defines a p-value by computing the probability that a random variable with a Snedecore's F distribution with degrees of freedom $\dim(W_1)$ and $\dim(W^\perp)$ is greater than or equal to the observed value of the statistic (4).

[7]More formally, if $\mathbb{P}_0$ denotes probabilities under the null hypothesis and if we set $F(t) = \mathbb{P}_0(T \geq t)$ for all $t \in \mathbb{R}$ then the p-value is defined by $\mathfrak{p} = F(T)$. It is a simple exercise to check that $\mathbb{P}_0(\mathfrak{p} \leq \gamma) \leq \gamma$ for all $\gamma \in [0, 1]$, so if we use $\mathfrak{p} \leq \gamma$ as the rejection criterion for the null hypothesis then we have a probability of at most γ of committing a *type I error*, i.e., incorrectly rejecting the null hypothesis. It is also easy to check that if the distribution of T under the null hypothesis is continuous (i.e., if $\mathbb{P}_0(T = t) = 0$ for all $t \in \mathbb{R}$) then actually $\mathbb{P}_0(\mathfrak{p} \leq \gamma) = \gamma$ for all $\gamma \in [0, 1]$.

4. Concrete applications

Let us now look into some concrete applications of the theory of Section 3. We start with the simplest case.

4.1. **Single sample t test.** Let $Y = (Y_1, \ldots, Y_n)$ be an independent and identically distributed (i.i.d.) sample of size $n \geq 2$ from a normal distribution with mean μ and standard deviation $\sigma > 0$, i.e., $(Y_i)_{i=1}^n$ is an independent family of random variables and all Y_i have a normal distribution with mean μ and standard deviation σ. The parameters μ and σ are regarded as unknown. We have that Y is an $\mathbb{R}^n$-valued Gaussian random vector with mean $\mu\mathbf{1}_n$ and variance equal to σ^2 times the canonical inner product $\langle \cdot, \cdot \rangle$ of $\mathbb{R}^{n*}$, where $\mathbf{1}_n \in \mathbb{R}^n$ denotes the vector whose coordinates are all equal to 1 and the canonical inner product of $\mathbb{R}^{n*}$ is the one for which the dual of the canonical basis is orthonormal. The inner product induced by $\langle \cdot, \cdot \rangle$ on $\mathbb{R}^n$ is just the canonical inner product of $\mathbb{R}^n$.

The subspace W of $\mathbb{R}^n$ in which we know that the mean of Y lies is the one-dimensional subspace spanned by $\mathbf{1}_n$ and the orthogonal complement $W^\perp$ is the subspace consisting of vectors with zero sum. If $\alpha \in W^*$ is the linear functional defined by $\alpha(\mu\mathbf{1}_n) = \mu$, for all $\mu \in \mathbb{R}$, then the linear extension $\hat{\alpha}$ of α to $\mathbb{R}^n$ that vanishes on $W^\perp$ is the linear functional that gives the arithmetic mean of a vector. The BLUE for the parameter $\alpha(\mu\mathbf{1}_n) = \mu$ is therefore

$$\hat{\alpha}(Y) = \overline{Y},$$

where $\overline{Y}$ is the *sample mean* defined by:

$$\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i.$$

The orthogonal projections $P_W : \mathbb{R}^n \to W$ and $P_{W^\perp} : \mathbb{R}^n \to W^\perp$ satisfy

$$P_W(Y) = \overline{Y}\mathbf{1}_n, \quad P_{W^\perp}(Y) = Y - \overline{Y}\mathbf{1}_n = (Y_1 - \overline{Y}, \ldots, Y_n - \overline{Y})$$

and therefore the unbiased estimator S^2 of σ^2 is the *sample variance* defined by:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \overline{Y})^2.$$

Since $\|\hat{\alpha}\| = \frac{1}{\sqrt{n}}$, Proposition 8 says that the random variable

$$\frac{\overline{Y} - \mu}{\frac{1}{\sqrt{n}} S}$$

has a Student's t distribution with $n-1$ degrees of freedom. Using this fact one can construct confidence intervals for μ and test hypotheses about μ.

4.2. **Two samples t test (with equal variances).** Let Y^1 be an i.i.d. sample of size $n_1 \geq 1$ from a normal distribution with mean μ_1 and standard deviation $\sigma > 0$ and Y^2 be an i.i.d. sample of size $n_2 \geq 1$ from a normal distribution with mean μ_2 and the same standard deviation σ. Assume also that Y^1 and Y^2 are independent. The parameters μ_1, μ_2 and σ are regarded as unknown. We have that $Y = (Y^1, Y^2)$ is a Gaussian V-valued random vector with $V = \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$. The mean of Y is $(\mu_1 \mathbf{1}_{n_1}, \mu_2 \mathbf{1}_{n_2})$ and the variance of Y is σ^2 times the canonical inner product of $V^* \cong \mathbb{R}^{n_1+n_2*}$. The subspace W of V in which we know that the mean of Y lies is the two-dimensional subspace spanned by $(\mathbf{1}_{n_1}, 0)$ and $(0, \mathbf{1}_{n_2})$. The orthogonal complement $W^\perp$ is the space of pairs of zero-sum vectors. The orthogonal projections P_W and $P_{W^\perp}$ satisfy

$$P_W(Y) = (\overline{Y^1}\mathbf{1}_{n_1}, \overline{Y^2}\mathbf{1}_{n_2}), \quad P_{W^\perp}(Y) = (Y^1 - \overline{Y^1}\mathbf{1}_{n_1}, Y^2 - \overline{Y^2}\mathbf{1}_{n_2})$$

and therefore the unbiased estimator of σ^2 is given by

$$S^2 = \frac{1}{n_1 + n_2 - 2}\Big(\sum_{i=1}^{n_1}(Y_i^1 - \overline{Y^1})^2 + \sum_{i=1}^{n_2}(Y_i^2 - \overline{Y^2})^2\Big),$$

assuming $n_1 + n_2 \geq 3$. Note that S^2 is the mean of the sample variances of Y^1 and Y^2 with weights given by the degrees of freedom $n_1 - 1$ and $n_2 - 1$. The statistic S^2 is often called the *pooled sample variance*.

In the context of a two samples problem the parameter that one normally wishes to estimate is the difference $\mu_1 - \mu_2$ of the means of the two normal distributions from which one is sampling. Thus let $\alpha \in W^*$ be the linear functional defined by $\alpha(\mu_1 \mathbf{1}_{n_1}, \mu_2 \mathbf{1}_{n_2}) = \mu_1 - \mu_2$ and note that the BLUE $\hat{\alpha}(Y)$ is given by:

$$\hat{\alpha}(Y) = \overline{Y^1} - \overline{Y^2}.$$

Since $\|\hat{\alpha}\| = \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$, Proposition 8 yields that the random variable

$$\frac{(\overline{Y^1} - \overline{Y^2}) - (\mu_1 - \mu_2)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

has a Student's t distribution with $n_1 + n_2 - 2$ degrees of freedom. This fact can be used to construct a confidence interval for $\mu_1 - \mu_2$ and to test the null hypothesis that $\mu_1 = \mu_2$.

Remark 11. If the variances of the normal distributions from which one is sampling are different then to test the null hypothesis $\mu_1 = \mu_2$ one uses the so called *Welch's t test*. This test is only approximate and it is not a particular case of the theory of Section 3.

4.3. **Multiple linear regression.** We consider a random variable Y, which we think of as a response variable, and we wish to account for the value of Y in terms of some explanatory variable X. We assume that X takes values in some arbitrary set $\mathcal{X}$ and that Y is normal with mean $\mathcal{L}(\theta, X)$ and unknown

standard deviation $\sigma > 0$, where θ is some unknown parameter belonging to some real-finite dimensional vector space Θ and $\mathcal{L} : \Theta \times \mathcal{X} \to \mathbb{R}$ is a known map that is linear in the variable θ. Note that since $\mathcal{X}$ is an arbitrary set, the explanatory variable X could actually correspond to something that in practical applications we would regard as "multiple variables", as "multiple variables" is really the same thing as a single variable taking values in a cartesian product. For simplicity, we choose to focus on the case in which the response variable Y is real-valued, though the case in which Y is a Gaussian random vector could also be easily handled using the theory of Section 3 as long as we assume the variance of Y to be known up to a multiplicative constant.

The model considered above is the "population model", i.e., the model for some generic individual sampled from a (possibly idealized) population of data. In concrete applications, the empirical data used to make estimates about the unknown parameters is a sample of that population. So consider an independent sample $Y = (Y_1, \ldots, Y_n)$ of size n for the response variable and a corresponding sample $X = (X_1, \ldots, X_n)$ of size n for the explanatory variable. We will regard here X as a fixed point of $\mathcal{X}^n$ instead of as a random object. The case of a random explanatory variable X can be easily handled as a corollary of the theory we develop in this subsection by using conditional probabilities and it will be discussed in Subsection A.1. For each $i = 1, \ldots, n$, we assume that Y_i is normal with mean $\mathcal{L}(\theta, X_i)$ and standard deviation $\sigma > 0$ for some unknown parameter $\theta \in \Theta$. Hence Y is a Gaussian $\mathbb{R}^n$-valued random vector whose variance is σ^2 times the canonical inner product of $\mathbb{R}^{n*}$ and whose mean μ is $L(\theta)$, where $L : \Theta \to \mathbb{R}^n$ is the linear map given by:

$$L(\theta) = \big(\mathcal{L}(\theta, X_1), \ldots, \mathcal{L}(\theta, X_n)\big). \tag{5}$$

We assume L to be injective so that the parameter θ is identifiable and we denote by W the image of L. As before, W must be a proper subspace of $\mathbb{R}^n$. For the development of the abstract theory in Section 3 it was convenient to identify Θ and W using L, but here it is not.

Let $\hat{\theta}$ be the Θ-valued random vector such that:

$$L(\hat{\theta}) = P_W(Y) = \hat{\mu}.$$

Clearly $\hat{\theta}$ is an unbiased estimator for θ. Moreover, for any $\alpha \in \Theta^*$, the BLUE for the parameter $\alpha(\theta) = \alpha\big(L^{-1}(\mu)\big)$ is $\alpha\big(L^{-1}(\hat{\mu})\big) = \alpha(\hat{\theta}) = \hat{\alpha}(Y)$, where $\hat{\alpha} = \alpha \circ L^{-1} \circ P_W$.

Let us find a convenient formula for $\hat{\theta}$ and for its variance. Denote by $\mathcal{R} : \mathbb{R}^n \to \mathbb{R}^{n*}$ the linear isomorphism $\mathcal{R}(y) = \langle y, \cdot \rangle$ induced by the canonical inner product of $\mathbb{R}^n$ and by $L^* : \mathbb{R}^{n*} \to \Theta^*$ the adjoint of the linear map L. Since the kernel of $L^* \circ \mathcal{R}$ is $W^\perp$, we have:

$$(L^* \circ \mathcal{R})\big(Y - L(\hat{\theta})\big) = 0. \tag{6}$$

Now $L^* \circ \mathcal{R} \circ L$ has the same kernel as L and thus it is an isomorphism from Θ onto Θ^*. Therefore (6) yields the following formula for $\hat{\theta}$:

$$\hat{\theta} = (L^* \circ \mathcal{R} \circ L)^{-1}\big((L^* \circ \mathcal{R})(Y)\big). \tag{7}$$

Using formula (3), the identification of $\mathrm{Var}(Y) : \mathbb{R}^{n*} \times \mathbb{R}^{n*} \to \mathbb{R}$ with the linear map $\sigma^2\mathcal{R}^{-1}$ from $\mathbb{R}^{n*}$ to $\mathbb{R}^n$ and keeping in mind that $\mathcal{R}^* = \mathcal{R}$ we then obtain:

$$\mathrm{Var}(\hat{\theta}) = \sigma^2(L^* \circ \mathcal{R} \circ L)^{-1}. \tag{8}$$

Since $\mathrm{Var}\big(\hat{\alpha}(Y)\big) = \sigma^2\|\hat{\alpha}\|^2$ and $\hat{\alpha}(Y) = \alpha(\hat{\theta})$, equality (8) yields:

$$\|\hat{\alpha}\|^2 = \alpha\big((L^* \circ \mathcal{R} \circ L)^{-1}(\alpha)\big).$$

Proposition 8 says that

$$\frac{\hat{\alpha}(Y) - \alpha(\theta)}{S\|\hat{\alpha}\|} \tag{9}$$

has a Student's t distribution with $\dim(W^\perp)$ degrees of freedom, where the unbiased estimator S^2 of σ^2 is given by:

$$S^2 = \frac{\|Y - L(\hat{\theta})\|^2}{\dim(W^\perp)}.$$

This fact can be used to make tests and to construct confidence intervals for the parameter $\alpha(\theta)$. In order to test a more general null hypothesis that θ belongs to some proper subspace Θ_0 of Θ, we simply use the F statistic given by Proposition 10 with $W_0 = L[\Theta_0]$. The projection $P_{W_1}(Y)$ appearing in the numerator of (4) can often be conveniently computed as the difference $P_W(Y) - P_{W_0}(Y)$, noting that $P_{W_0}(Y) = L(\hat{\theta}_0)$ with $\hat{\theta}_0$ given by the formula obtained from (7) by replacing L with its restriction to Θ_0.

Remark 12. One might naively think that a "linear model" for the response variable Y should be a model that is linear in the explanatory variable X, but the linearity that matters is actually the linearity with respect to the parameter θ. The linearity of $\mathcal{L}(\theta, X)$ in X plays no role in the development of the theory and in fact one can easily transform the explanatory variable X into a new variable X' in order to force the model to be linear in the new explanatory variable X'. Namely, let $\mathcal{X}'$ denote the dual space Θ^* of Θ and set $X' = \mathcal{L}(\cdot, X)$, so that $\mathcal{L}(\theta, X) = \mathcal{L}'(\theta, X')$ where $\mathcal{L}' : \Theta \times \mathcal{X}' \to \mathbb{R}$ is the bilinear map given by simple evaluation $\mathcal{L}'(\theta, \alpha) = \alpha(\theta)$ of a linear functional $\alpha \in \Theta^*$ on a parameter $\theta \in \Theta$.

Remark 13. Consider the linear map

$$\Theta \ni \theta \longmapsto \mathcal{L}(\theta, \cdot) \in \mathbb{R}^{\mathcal{X}}, \tag{10}$$

where $\mathbb{R}^{\mathcal{X}}$ denotes the vector space of all real-valued functions on the set $\mathcal{X}$. The linear map (10) has to be assumed injective in order for the map (5) to be injective. We note that the substance of the model is in the finite-dimensional subspace of $\mathbb{R}^{\mathcal{X}}$ given by the image of the injective linear map

(10). Namely, such injective linear map gives an isomorphism between the parameter space Θ and its image and changing this isomorphism amounts to making a linear reparametrization of the model. Though often the choice of a particular concrete parametrization for the model is convenient, one might also in some cases prefer to avoid the choice of some parametrization altogether, letting Θ be a finite-dimensional subspace of $\mathbb{R}^{\mathcal{X}}$ and setting $\mathcal{L}(\theta, X) = \theta(X)$ for all $\theta \in \Theta$ and $X \in \mathcal{X}$.

Remark 14. The theory developed above is readily adaptable to the case of a *multiple linear regression with weights*. This is the model in which the i-th element Y_i of the sample of the response variable is assumed to have variance $\frac{\sigma^2}{w_i}$ for some known positive real numbers w_1, ..., w_n. In this case one simply replaces the canonical inner product of $\mathbb{R}^{n*}$ with the inner product whose matrix with respect to the dual of the canonical basis is diagonal with diagonal elements $\frac{1}{w_i}$. The induced inner product on $\mathbb{R}^n$ is then the one whose matrix with respect to the canonical basis is diagonal with diagonal elements w_i. The linear isomorphism $\mathcal{R} : \mathbb{R}^n \to \mathbb{R}^{n*}$ appearing on all the formulas should of course be replaced with the isomorphism induced by the latter inner product.

4.4. **ANOVA.** The acronym ANOVA, which stands for "analysis of variance", is just a bad name for the particular case of the linear model considered in Subsection 4.3 in which the value set $\mathcal{X}$ for the explanatory variable X is finite. It can also be seen as a generalization of the theory of two samples t tests (Subsection 4.2) to the case of a finite number of samples. The parameter space Θ can be regarded as the space $\mathbb{R}^{\mathcal{X}}$ of finite families $(\mu_x)_{x\in\mathcal{X}}$ and the map $\mathcal{L} : \mathbb{R}^{\mathcal{X}} \times \mathcal{X} \to \mathbb{R}$ is just the evaluation map $\mathcal{L}\big((\mu_x)_{x\in\mathcal{X}}, x\big) = \mu_x$.

Given samples $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_n)$ of the explanatory variable and the response variable, respectively, we denote by n_x the number of coordinates of X that are equal to x, for each $x \in \mathcal{X}$, and we assume all n_x to be positive. It is more convenient here to think of Y as taking values in the product space

$$V = \prod_{x\in\mathcal{X}} \mathbb{R}^{n_x},$$

which means that we regard Y as an independent finite family $(Y^x)_{x\in\mathcal{X}}$, with each Y^x being an i.i.d. random sample of size n_x from a normal distribution with unknown mean μ_x and unknown standard deviation $\sigma > 0$. We then have that Y is a Gaussian random vector with mean

$$\mu = L\big((\mu_x)_{x\in\mathcal{X}}\big) = (\mu_x \mathbf{1}_{n_x})_{x\in\mathcal{X}}$$

and variance given by σ^2 times the canonical inner product of $V^* \cong \mathbb{R}^{n*}$. The subspace W of V given by the image of the linear map $L : \mathbb{R}^{\mathcal{X}} \to V$ consists of the families whose x-th coordinate is a scalar multiple of $\mathbf{1}_{n_x}$, for all $x \in \mathcal{X}$. The orthogonal complement $W^\perp$ is then the space of families

whose coordinates are zero-sum vectors and the orthogonal projections P_W and $P_{W^\perp}$ satisfy:

$$P_W(Y) = (\overline{Y^x}\mathbf{1}_{n_x})_{x\in\mathcal{X}} \quad \text{and} \quad P_{W^\perp}(Y) = (Y^x - \overline{Y^x}\mathbf{1}_{n_x})_{x\in\mathcal{X}}.$$

The dimension of $W^\perp$ is $n - |\mathcal{X}|$, where $|\mathcal{X}|$ denotes the cardinality of $\mathcal{X}$. Assuming $n > |\mathcal{X}|$, the unbiased estimator S^2 of σ^2 is given by:

$$S^2 = \frac{1}{n-|\mathcal{X}|}\sum_{x\in\mathcal{X}} \|Y^x - \overline{Y^x}\mathbf{1}_{n_x}\|^2.$$

Note that S^2 is just the mean of the sample variances of the samples Y^x weighted by the corresponding degrees of freedom $n_x - 1$.

The standard goal of the ANOVA procedure is to test the null hypothesis that all means μ_x are equal. One could ask why not simply use several two samples t tests to compare the means μ_x pairwise. The answer is that doing multiple comparisons inflates the probability of type I error, i.e., the probability of committing a type I error in at least one comparison gets larger as the number of comparisons grows. It's best therefore to use a single test. To this aim, we consider the one-dimensional subspace Θ_0 of $\Theta = \mathbb{R}^{\mathcal{X}}$ consisting of families whose coordinates are all equal and we use the F statistic from Proposition 10 with $W_0 = L[\Theta_0]$. We assume the cardinality of $\mathcal{X}$ to be at least 2. The space W_0 is the one-dimensional space spanned by $(\mathbf{1}_{n_x})_{x\in\mathcal{X}}$ and the orthogonal projection P_{W_0} satisfies

$$P_{W_0}(Y) = (\overline{Y}\mathbf{1}_{n_x})_{x\in\mathcal{X}},$$

where $\overline{Y}$ is the *grand mean* given by:

$$\overline{Y} = \frac{1}{n}\sum_{x\in\mathcal{X}}\sum_{i=1}^{n_x} Y_i^x = \frac{1}{n}\sum_{x\in\mathcal{X}} n_x\overline{Y_x}.$$

If W_1 denotes the orthogonal complement of W_0 in W then

$$P_{W_1}(Y) = P_W(Y) - P_{W_0}(Y) = \big((\overline{Y^x} - \overline{Y})\mathbf{1}_{n_x}\big)_{x\in\mathcal{X}}$$

and therefore the numerator of the F statistic (4) is given by:

$$\frac{1}{\dim(W_1)}\|P_{W_1}(Y)\|^2 = \frac{1}{|\mathcal{X}|-1}\sum_{x\in\mathcal{X}} n_x(\overline{Y^x} - \overline{Y})^2.$$

Proposition 10 then says that, under the null hypothesis that all means μ_x are equal, the statistic

$$\frac{1}{S^2(|\mathcal{X}|-1)}\sum_{x\in\mathcal{X}} n_x(\overline{Y^x} - \overline{Y})^2$$

has a Snedecore's F distribution with degrees of freedom $|\mathcal{X}|-1$ and $n-|\mathcal{X}|$.

Remark 15. The standard literature names the procedure described above as "one-way ANOVA". The procedure usually referred to as "two-way ANOVA" corresponds to the same model, but with "two explanatory variables" instead of "one explanatory variable". Of course this simply means that the

finite set $\mathcal{X}$ is taken as a cartesian product $\mathcal{X}_1 \times \mathcal{X}_2$ of two finite sets and no additional theory needs to be developed. The only additional thing worth mentioning here is that by regarding the explanatory variable X as a pair of variables (X_1, X_2) we can test for the absence of interaction between the two variables. *Absence of interaction* here means that the parameter $(\mu_{(x_1,x_2)})_{(x_1,x_2)\in\mathcal{X}_1\times\mathcal{X}_2}$ belongs to the subspace Θ_0 of the parameter space $\Theta = \mathbb{R}^{\mathcal{X}_1\times\mathcal{X}_2}$ consisting of maps that are the sum of a function of the first projection with a function of the second projection, i.e., $\mu_{(x_1,x_2)} = \mu^1_{x_1} + \mu^2_{x_2}$ for $\mu^1 \in \mathbb{R}^{\mathcal{X}_1}$ and $\mu^2 \in \mathbb{R}^{\mathcal{X}_2}$. To test the null hypothesis of absence of interaction one can then use the F statistic given by Proposition 10 with $W_0 = L[\Theta_0]$.

Appendix A. Conditional probability

The conditional probability $\mathbb{P}(A|B)$ of an event A conditioned on an event B such that $\mathbb{P}(B) > 0$ is defined by:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \tag{11}$$

More generally, one is interested in conditional probabilities of the form $\mathbb{P}(Y \in A|X = x)$, where X and Y are random objects, A is a measurable subset of the counterdomain $\mathcal{Y}$ of Y and x is a point of the counterdomain $\mathcal{X}$ of X. Though $[Y \in A] = Y^{-1}[A]$ and $[X = x] = X^{-1}(x)$ are events, one often has $\mathbb{P}(X = x) = 0$, so the simple definition (11) does not apply.

In order to define this more general notion of conditional probability, the following definition is useful.

Definition 16. Given measurable spaces $\mathcal{X}$ and $\mathcal{Y}$, by a *kernel* with source $\mathcal{X}$ and target $\mathcal{Y}$ we mean a map K that assigns a probability measure K_x on $\mathcal{Y}$ to each point $x \in \mathcal{X}$ in such a way that for every measurable subset A of $\mathcal{Y}$ the map $\mathcal{X} \ni x \mapsto K_x(A) \in \mathbb{R}$ is measurable.

Given a probability measure $\mathbb{P}$ on $\mathcal{X}$ and a kernel K with source $\mathcal{X}$ and target $\mathcal{Y}$ we define a probability measure $\mathbb{P} \star K$ on the product space $\mathcal{X} \times \mathcal{Y}$ by setting[8]

$$(\mathbb{P} \star K)(C) = \int_{\mathcal{X}} K_x(C_x)\, \mathrm{d}\mathbb{P}(x),$$

for every measurable subset C of $\mathcal{X} \times \mathcal{Y}$, where $C_x = \{y \in \mathcal{Y} : (x, y) \in C\}$. When the map K is constant, i.e., if K_x does not depend on x then $\mathbb{P} \star K$ is simply the standard product of the measure $\mathbb{P}$ by the measure K_x.

We note that if X is an $\mathcal{X}$-valued random object and Y is a $\mathcal{Y}$-valued random object and if $\mathbb{P} \star K$ is the distribution of (X, Y) then $\mathbb{P}$ is the distribution of X, i.e., $\mathbb{P}$ is the image of $\mathbb{P} \star K$ under the first projection of

[8]One has to show first that the map $\mathcal{X} \ni x \mapsto K_x(C_x)$ is measurable for every measurable subset C of the product $\mathcal{X} \times \mathcal{Y}$. This is done as in standard proofs of Fubini's Theorem, by noting that the class of sets C for which the thesis holds contains the measurable rectangles and is closed under certain set operations.

the product $\mathcal{X} \times \mathcal{Y}$. A good way of intuitively understanding the meaning of the probability measure $\mathbb{P} \star K$ is in terms of the corresponding sampling strategy for a value (x, y) for the random object (X, Y) having $\mathbb{P} \star K$ as its distribution: first sample the value x of X using $\mathbb{P}$ as the distribution of X and then sample the value y of Y using K_x as the distribution of Y.

Definition 17. Given an $\mathcal{X}$-valued random object X and a $\mathcal{Y}$-valued random object Y, by a *regular conditional probability* of Y given X we mean any kernel K with source $\mathcal{X}$ and target $\mathcal{Y}$ such that the distribution of the random object (X, Y) is equal to $\mathbb{P}_X \star K$, where $\mathbb{P}_X$ denotes the distribution of X.

Since a probability measure on $\mathcal{X} \times \mathcal{Y}$ is characterized by its value on measurable rectangles $B \times A$, we have that K is a regular conditional probability of Y given X if and only if

$$\mathbb{P}\big([Y \in A] \cap [X \in B]\big) = \mathbb{P}\big((X, Y) \in B \times A\big) = \int_B K_x(A) \, \mathrm{d}\mathbb{P}_X(x), \tag{12}$$

for every measurable subset A of $\mathcal{Y}$ and every measurable subset B of $\mathcal{X}$. In other words, a kernel K is a regular conditional probability of Y given X if and only if for every measurable subset A of $\mathcal{Y}$, the map $\mathcal{X} \ni x \mapsto K_x(A)$ is a Radon–Nikodym derivative of the measure $B \mapsto \mathbb{P}\big([Y \in A] \cap [X \in B]\big)$ on $\mathcal{X}$ with respect to the distribution $\mathbb{P}_X$ of X. In particular, if K and K' are two regular conditional probabilities of Y given X then for every measurable subset A of $\mathcal{Y}$ we have that $K_x(A) = K'_x(A)$, for $\mathbb{P}_X$-almost every $x \in \mathcal{X}$. It follows that if the σ-algebra of $\mathcal{Y}$ is countably generated, then $K_x = K'_x$ for $\mathbb{P}_X$-almost every $x \in \mathcal{X}$.

Even though a regular conditional probability K of Y given X is not unique, it is almost unique in the sense explained above and we thus write

$$\mathbb{P}(Y \in A | X = x) = K_x(A), \tag{13}$$

for every $x \in \mathcal{X}$ and every measurable subset A of $\mathcal{Y}$. Unfortunately, there are pathological situations in which a regular conditional probability does not exist because one cannot choose all the relevant Radon–Nikodym derivatives in such a way that the map $A \mapsto K_x(A)$ is countably additive for all $x \in \mathcal{X}$. However, existence holds under fairly general conditions, for example, it holds if the measurable space $\mathcal{Y}$ is a *standard Borel space*, i.e., if it is isomorphic as a measurable space to a Borel subset of a complete separable metric space endowed with its Borel σ-algebra[9].

[9]It is a standard result on basic descriptive set theory that two uncountable standard Borel spaces are isomorphic, so it is sufficient to prove existence if $\mathcal{Y}$ is the real line endowed with its Borel σ-algebra. In this case, one first defines the conditional cumulative distribution function $F_x(y) = \mathbb{P}(Y \leq y | X = x)$ for $x \in \mathcal{X}$ and $y \in \mathbb{Q}$ using Radon–Nikodym derivatives. Since $\mathbb{Q}$ is countable, we have that for $\mathbb{P}_X$-almost every $x \in \mathcal{X}$ the map F_x is increasing, right-continuous and satisfies $\lim_{y \to -\infty} F_x(y) = 0$ and $\lim_{y \to +\infty} F_x(y) = 1$. We can then uniquely extend each F_x to an increasing right-continuous function defined on the entire real line and obtain using the standard theory of extensions of measures a

Using the more familiar notation (13) for conditional probabilities, formula (12) becomes

$$\mathbb{P}\big([Y \in A] \cap [X \in B]\big) = \int_B \mathbb{P}(Y \in A | X = x)\, \mathrm{d}\mathbb{P}_X(x)$$

and it is known as the *law of total probability*. It follows that

$$\mathbb{P}(Y \in A) = \int_{\mathcal{X}} \mathbb{P}(Y \in A | X = x)\, \mathrm{d}\mathbb{P}_X(x)$$

and that

$$\mathbb{P}(Y \in A | X \in B) = \frac{1}{\mathbb{P}_X(B)} \int_B \mathbb{P}(Y \in A | X = x)\, \mathrm{d}\mathbb{P}_X(x)$$

if $\mathbb{P}_X(B) = \mathbb{P}(X \in B)$ is positive.

If $Z = f(Y)$ is a measurable function of Y then a regular conditional probability of Z given X can be obtained from a regular conditional probability of Y given X by taking the images of the probability measures $\mathbb{P}(Y \in \cdot | X = x)$ under f, i.e., by setting

$$\mathbb{P}(Z \in A | X = x) = \mathbb{P}\big(Y \in f^{-1}[A] | X = x\big),$$

for every measurable subset A of the counterdomain of Z and every $x \in \mathcal{X}$. More generally, if $Z = f(X, Y)$ is a measurable function of X and Y, a regular conditional probability of Z given X is obtained as

$$\mathbb{P}(Z \in A | X = x) = \mathbb{P}\big(Y \in f_x^{-1}[A] | X = x\big),$$

where f_x denotes the function $f_x = f(x, \cdot)$.

A.1. **Multiple linear regression with a random explanatory variable.** In Subsection 4.3 we developed the theory of multiple linear regression assuming that the sample $(X_1, \ldots, X_n)$ for the explanatory variable X was a fixed element of the set $\mathcal{X}^n$ instead of a random object. Using conditional probabilities, it is straightforward to generalize such theory to the case of a random explanatory variable. In this context, we need to assume that the set $\mathcal{X}$ in which the explanatory variable takes values is endowed with some σ-algebra and that the map $\mathcal{L} : \Theta \times \mathcal{X} \to \mathbb{R}$ is such that $\mathcal{L}(\theta, \cdot)$ is measurable for every $\theta \in \Theta$.

For the "population model" we assume that Y is a random variable, that X is an $\mathcal{X}$-valued random object and that the distribution of (X, Y) is such that, for every $x \in \mathcal{X}$, the conditional distribution of Y given $X = x$ is normal with mean $\mathcal{L}(\theta, x)$ and standard deviation $\sigma > 0$, where $\theta \in \Theta$ and σ are unknown parameters. No assumptions are needed about the distribution of X. The empirical data consists of a sample $Y = (Y_1, \ldots, Y_n)$ of size n and a corresponding sample $X = (X_1, \ldots, X_n)$, where X is an arbitrary $\mathcal{X}^n$-valued random object and for every $x \in \mathcal{X}^n$ we have that, conditionally on $X = x$, the variables $Y_1, \ldots, Y_n$ are independent and Y_i

unique probability measure $A \mapsto \mathbb{P}(Y \in A | X = x)$ on the Borel σ-algebra of $\mathbb{R}$ such that $\mathbb{P}(Y \leq y | X = x) = F_x(y)$, for all $y \in \mathbb{R}$.

has a normal distribution with mean $\mathcal{L}(\theta, x_i)$ and standard deviation σ, for all $i = 1, \ldots, n$. We do not need to assume that the coordinates of X are independent. Clearly, for every $x \in \mathcal{X}^n$, the distribution of Y conditioned on $X = x$ is that of a Gaussian $\mathbb{R}^n$-valued random vector with mean

$$L_x(\theta) = \big(\mathcal{L}(\theta, x_1), \ldots, \mathcal{L}(\theta, x_n)\big)$$

and variance given by σ^2 times the canonical inner product of $\mathbb{R}^{n*}$. Here $L_x : \Theta \to \mathbb{R}^n$ is a linear map for every $x \in \mathcal{X}^n$ and L_X (i.e., the composition of $x \mapsto L_x$ with X) is a *random linear map* from Θ to $\mathbb{R}^n$, i.e., a random object taking values in the space of linear maps from Θ to $\mathbb{R}^n$.

The development of the theory now goes through exactly as in Subsection 4.3, but with everything conditioned on $X = x$ for some $x \in \mathcal{X}^n$. We will then obtain, for instance, that the random variable (9) has a Student's t distribution with $\dim(W^\perp)$ degrees of freedom, conditionally on $X = x$. Similarly, the statistic obtained from Proposition 10 will have a Snedecore's F distribution under the appropriate null hypothesis, conditionally on $X = x$.

The important point is that confidence intervals and p-values obtained from the theory of Subsection 4.3 will have the correct properties also *unconditionally*. For example, if $\mathrm{CI}(X, Y)$ is a γ-confidence interval for some parameter $\alpha(\theta)$ conditionally on $X = x$, i.e., if

$$\mathbb{P}\big(\mathrm{CI}(X, Y) \ni \alpha(\theta) | X = x\big) = \gamma \tag{14}$$

for all $x \in \mathcal{X}^n$ then also:

$$\mathbb{P}\big(\mathrm{CI}(X, Y) \ni \alpha(\theta)\big) = \gamma. \tag{15}$$

Namely, (15) follows directly from (14) by integrating in x with respect to the distribution of X. There is, however, one caveat: the theory of Subsection 4.3 requires the linear map (5) to be injective, so $\mathrm{CI}(x, Y)$ is only defined and equality (14) holds only if $x \in \mathcal{X}^n$ is such that L_x is injective. Thus we obtain (15) only if the random linear map L_X is injective with probability 1. In the general case, we have the equality

$$\mathbb{P}\big(\mathrm{CI}(X, Y) \ni \alpha(\theta) | L_X \text{ is injective}\big) = \gamma,$$

assuming that L_X is injective with positive probability.

Similar remarks are valid for p-values. For instance, if we use the theory of Subsection 4.3 to obtain a p-value $\mathfrak{p}(X, Y)$ satisfying

$$\mathbb{P}\big(\mathfrak{p}(X, Y) \leq \gamma | X = x\big) = \gamma$$

for all $\gamma \in [0, 1]$ under some null hypothesis then integrating over x with respect to the distribution of X we obtain

$$\mathbb{P}\big(\mathfrak{p}(X, Y) \leq \gamma | L_X \text{ is injective}\big) = \gamma$$

for all $\gamma \in [0, 1]$ under the same null hypothesis.

Departamento de Matemática,
Universidade de São Paulo, Brazil
Email address: tausk@ime.usp.br
URL: http://www.ime.usp.br/~tausk