AP 2023–01

Chair of Applied Stochastics and Risk Management

HELMUT SCHMIDT UNIVERSITÄT
Universität der Bundeswehr Hamburg

Faculty of Economic and Social Sciences
Department of Mathematics and Statistics

Working Paper

A Test for the Validity of Regression Models

Gabriel Frahm

March 7, 2024

Working Paper AP 2023–01
March 7, 2024

A Test for the Validity of Regression Models

Gabriel Frahm

Helmut Schmidt University
Faculty of Economic and Social Sciences
Department of Mathematics and Statistics
Chair of Applied Stochastics and Risk Management
Holstenhofweg 85, D-22043 Hamburg, Germany

URL: www.hsu-hh.de/stochastik
Phone: +49 (0)40 6541-2791
E-mail: frahm@hsu-hh.de

Working Paper
Please use only the latest version of the manuscript. Distribution is unlimited.

Supervised by: Prof. Dr. Gabriel Frahm
Chair of Applied Stochastics and
Risk Management

URL: www.hsu-hh.de/stochastik

A Test for the Validity of Regression Models*

Gabriel Frahm†
Helmut Schmidt University
Department of Mathematics and Statistics
Chair of Applied Stochastics and
Risk Management

March 7, 2024

Abstract

This work elaborates the connection between prediction and description in regression analysis. Many empirical studies aim at description, which requires a valid regression model. I show that regression models with a strong prediction power can be highly invalid and thus be inappropriate for the purpose of description. Conversely, valid regression models may have a weak prediction power and they even need not fit at all. For this reason, measures of prediction power, or of goodness of fit, are not suitable for assessing the validity of regression models. I develop a simple validity test, which can be applied to all kinds of regression models with an arbitrary number of regressors. It is very powerful in large samples and performs well also in small samples, given that the validity of the regression model is sufficiently low and that there is not too much noise in the true regression equation.

Keywords: Accuracy, Description, Explanation, Goodness of fit, Prediction, Regression, Specification, Validity.

JEL Classification: C01, C52.

*I thank Alexander Jonen very much for his valuable suggestions and helpful comments, which essentially improved the manuscript. I thank also Christian Glöer and André Küster Simic for our inspiring discussions about the topic.
†Phone: +49 40 6541-2791, e-mail: frahm@hsu-hh.de.

Contents

1. Motivation

LINEAR regression is probably the most widely used method of data analysis in economics and it is the main subject matter of classical econometrics. Among many other scientific areas, it is frequently applied in social and natural sciences, too. We can accomplish two goals with regression analysis, viz., prediction and description. Prediction requires no structural assumption about the regression equation $Y = f(X) + \varepsilon$, where Y represents the dependent

variable, f is the regression function, X is some vector of regressors, and ε is the regression error. Thus, prediction does not force us to specify any probabilistic model. Quite the contrary, description is based on the fundamental assumption that the regression model $Y = f(X) + \varepsilon$ is *valid*, i.e., that $f(X)$ corresponds the conditional mean of Y given X.

Apparently, most empirical studies try to *describe* the impact of X on Y rather than to predict Y by a (linear or nonlinear) regression on X. Nonetheless, those studies often try to legitimate some regression model $Y = f(X) + \varepsilon$ by demonstrating its capability to *predict* Y. Thus, it seems that description is often confounded with prediction. Indeed, a regression model may very well have a strong prediction power and it can also possess a good fit, i.e., be accurate in explaining the distribution of Y, but the same model can still be invalid—even to a very high degree—and thus it can be completely unsuitable for the purpose of description. Conversely, a valid regression model $Y = f(X) + \varepsilon$ can have a weak prediction power and $f(X)$ even need not fit to Y at all. However, in that case f is the *only* regression function that describes the impact of X on Y, appropriately. This problem becomes even more serious when model selection is based on measures of prediction power or of goodness of fit rather than validity.

It is shown that even an optimal predictor $f(X)$ of Y need not constitute a valid regression model $Y = f(X) + \varepsilon$, but a valid regression model $Y = f(X) + \varepsilon$ must always be based on an optimal predictor $f(X)$ of Y. Simply put, optimality is a *necessary*, but not a sufficient condition for validity. This link seems to be often misunderstood in practice. Because regression analysis plays such an important role in empirical research, and validity is a conditio sine qua non in most applications, the given problem is relevant from a practical point of view. This work contains some new and surprising insights about the question of validity, which hopefully are interesting for the audience. Thus, I think that it is relevant from a theoretical perspective, too.

Whether or not a regression model is valid can very well depend on the choice of regressors, which means that also the choice of their transformation can have an essential impact on the validity of the model. However, validity tests can rarely be found in the literature. Here, I do not mean goodness-of-fit tests, tests on null hypotheses concerning the regression parameters, or information criteria like, e.g., the Akaike information criterion or the Bayesian information criterion. These kind of specification tests do not address the question of validity. Further, many specification tests are based on the classical assumptions of the Gaussian linear regression model, which are very restrictive and hardly applicable in most real-life situations. Sometimes, also a visual inspection of the regression errors is recommended to assess the validity of a given regression model, but usually this procedure is insufficient, too, which is shown later on.

A main purpose of this work is to present a genuine test for the null hypothesis that a given regression model is valid. It is simple and thus easy to implement. More importantly, the validity test can be applied to all kinds of regression models with an arbitrary number of regressors. It is very powerful in large samples and performs well also in small samples, provided that the validity of the regression model is sufficiently low and that there is not too much noise in the true regression equation. If the test rejects the null hypothesis of validity, the given regression model is (significantly) invalid and so it should be abandoned. Put another way, one should

consider another regression model or, at least, change or transform the variables. However, this holds true only if one wants to *describe* the impact of X on Y by some regression model $Y = f(X) + \varepsilon$. By contrast, if one just wishes to *predict* Y by X, the regression model may very well be invalid and he or she should focus on the prediction power of $f(X)$.

2. Theoretical Background

2.1. Prerequisites

The elements of an Euclidean space are considered column vectors. Random variables and random vectors are always denoted by capital Roman letters. The same holds true for subsets of the Euclidean space and real-valued matrices. A small Roman letter indicates a real number, an Euclidean vector, or a function. For example, $Z = (Z_1, \ldots, Z_d)$ is a d-dimensional random vector, whereas $z = (z_1, \ldots, z_d) \in \mathbb{R}^d$ represents some realization of Z. By contrast, $\{Z_1, \ldots, Z_n\}$ denotes a set of n random variables or random vectors $Z_1, \ldots, Z_n$. A small Greek letter can either be a random variable, a real number, or an Euclidean vector. An equality or inequality between two (random) vectors means that the assertion holds true componentwise (and almost surely). Further, $\mathrm{Var}(Z)$ denotes the covariance matrix of Z and $\mathrm{Var}(Z) > 0$ means that it is positive definite. The transpose of Z is denoted by Z', 0 symbolizes the zero scalar or a vector of zeros, depending on the given context. The same holds true for the symbol 1, respectively. The rank of a (random or real-valued) matrix A is written as $\mathrm{rk}\, A$.

Moreover, $\wedge$ stands for the logical "and," $\vee$ denotes the logical "or," $:=$ means "is defined as," $\mathbb{N} := \{1, 2, \ldots\}$, $\lceil \cdot \rceil$ is the ceiling function, i.e., $\lceil x \rceil$ is the lowest integer that is greater than or equal to $x \in \mathbb{R}$, $\mathbf{I}_d$ is the $d \times d$ identity matrix, and $\mathcal{N}_d(\mu, \Sigma)$ is the d-dimensional normal distribution with mean vector μ and covariance matrix Σ. The univariate normal distribution with mean μ and variance σ^2 is symbolized by $\mathcal{N}(\mu, \sigma^2)$, t_ν denotes Student's t-distribution with ν degrees of freedom, χ^2_δ is the χ^2-distribution with δ degrees of freedom, and F^ν_δ denotes the F-distribution with ν numerator and δ denominator degrees of freedom. Weak convergence, i.e., convergence in distribution, is indicated by $\rightsquigarrow$ and $\xrightarrow{\mathrm{P}}$ denotes convergence in probability. The additional remark "$n \to \infty$," i.e., that the sample size tends to infinity, is dropped for notational convenience. The symbol Φ denotes the cumulative distribution function of the standard normal distribution, whereas ϕ is its probability density function. Finally, if $(x, y) \mapsto f(x, y)$ is a real-valued differentiable function of $x \in A \subseteq \mathbb{R}^k$ and $y \in B \subseteq \mathbb{R}^l$, then $\frac{\partial}{\partial x} f(\cdot, y)$ represents the partial derivative of f with respect to x given y, whereas $\frac{\partial}{\partial y} f(x, \cdot)$ is defined mutatis mutandis.

Now, let $(\Omega, \mathcal{A}, \mathrm{P})$ be some probability space, where $\mathcal{A}$ is a σ-algebra on Ω, and L^2 be the Hilbert space of all square-integrable random variables, which is equipped with the inner product $\mathrm{E}(XY)$ for all $X, Y \in L^2$. Further, let $X_1, \ldots, X_m, Y \in L^2$ be some random variables and (X, Y) with $X = (X_1, \ldots, X_m)$ be the corresponding $(m+1)$-dimensional random vector. It is assumed that $\mathrm{Var}(Y) > 0$ unless otherwise stated. Moreover, let $D \subseteq \mathbb{R}^m$ be some (Borel) set such that $\mathrm{P}(X \in D) = 1$ and f be a *regression function*, i.e., any real-valued function on D such that $f(X) \in L^2$. Henceforth, we may consider $\{X_1, \ldots, X_m\}$ our basic set of regressors.

The corresponding regression equation is given by

$$Y = f(X) + \varepsilon, \tag{1}$$

where the m components of X are called explanatory variables or regressors and Y is referred to as the dependent variable.[1] The regressors, $X_1, \ldots, X_m$, and the dependent variable, Y, are *fixed*, which implies that they do not depend on the choice of the regression function f.[2] By contrast, the regression error is defined by $\varepsilon := Y - f(X)$. Hence, ε represents a residual, i.e., it is not considered fixed and so it is not treated like a regressor. Nevertheless, for notational convenience, I refrain from using any index in order to clarify that ε depends on Y, X, and f.

We conclude that Equation 1 is satisfied just by definition, irrespective of how we choose the dependent variable Y, the regressors $X_1, \ldots, X_m$, and the regression function f. Put another way, as long as we do not make any assumption about the joint distribution of X and ε, Equation 1 is purely tautological. Consequently, it does not represent any *model*—apart from the very fact that the surrounding framework is just a probabilistic model of reality.

2.2. Main Goals of Regression Analysis

2.2.1. Prediction

Let $\mathcal{G} \neq \emptyset$ be the set of all regression functions and $\mathcal{F}$ be a nonempty subset of $\mathcal{G}$. Suppose that the realization of the random variable Y is unknown, i.e., Y is *unobservable*, whereas the random variables $X_1, \ldots, X_m$ are observable. In order to predict Y by X, one typically tries to find an element of $\mathcal{F}$ that minimizes the mean square prediction error

$$\mathrm{E}(\varepsilon^2) = \mathrm{E}\left(\big(Y - f(X)\big)^2\right).$$

Thus, $f(X)$ is called an optimal predictor of Y if and only if f minimizes $\mathrm{E}(\varepsilon^2)$ among all regression functions in $\mathcal{F}$. Consequently, $f \in \mathcal{F}$ is said to be optimal if and only if $f(X)$ is an optimal predictor of Y. In general, an optimal regression function need not be unique.

For example, suppose that $\mathcal{F}$ is a parametric family of regression functions. This means that $\mathcal{F} = \big\{f(\cdot\,, \theta) \colon \theta \in \Theta \subseteq \mathbb{R}^q\big\}$, where $f(\cdot\,, \theta)$ is a function of $x \in D$ and θ is a parameter vector that belongs to some parameter space Θ. Further, assume that $f(X, \cdot)$ is differentiable, almost surely, at each $\theta \in \Theta$ and that $\frac{\partial}{\partial \theta}\mathrm{E}(\varepsilon^2) = \mathrm{E}\big(\frac{\partial}{\partial \theta}\varepsilon^2\big)$. Let $f(X, \theta^*)$ with $\theta^* \in \Theta$ be an optimal predictor of Y. Under the usual regularity conditions of optimization theory (see, e.g., Boyd and Vandenberghe, 2009, Section 5.5), it turns out that θ^* must be a Karush-Kuhn-Tucker (KKT) point. In particular, if $\theta \mapsto \mathrm{E}(\varepsilon^2)$ is convex, we can use Slater's regularity condition. Then, each KKT point θ^* leads us to an optimal predictor $f(X, \theta^*)$ of Y and it is guaranteed that the given minimum is global.

[1] The dependent variable, Y, can be called also "regressand" (Greene, 2012, p. 52), but this term is not commonly used. Further, I do not call the components of X "independent," since they usually depend on each other and also on Y.

[2] A counterexample is $Y = \alpha + \beta X + \varepsilon$ with $X = \beta Z$, where $Z \in L^2$ is a fixed random variable.

Moreover, if there are no (equality or inequality) constraints regarding θ at all, then

$$\mathrm{E}\left(\frac{\partial}{\partial\theta}f(X,\theta^*)\,\varepsilon^*\right)=0 \tag{2}$$

is a necessary and sufficient condition for an optimal prediction of Y, where $\frac{\partial}{\partial\theta}f(X,\theta^*)$ is the derivative of $f(X,\cdot)$ at θ^* and $\varepsilon^* := Y - f(X,\theta^*)$ denotes the associated regression error.

Further, suppose that $\theta=(\mu,\eta)$, where $\mu\in\mathbb{R}$ is a location parameter. Then, the family $\mathcal{F}$ is closed under translations, i.e., $f\in\mathcal{F}\Rightarrow\lambda+f\in\mathcal{F}$ for all $\lambda\in\mathbb{R}$. In this case, θ^* also minimizes the variance of the regression error ε. Moreover, we have that $\frac{\partial}{\partial\mu}f(X,\theta)=1$ and thus Equation 2 leads us to the two (necessary and sufficient) conditions $\mathrm{E}(\varepsilon^*)=0$ and $\mathrm{Cov}\big(\frac{\partial}{\partial\eta}f(X,\theta^*),\varepsilon^*\big)=0$.[3] For example, consider a linear predictor, i.e., $f(X,\alpha,\beta)=\alpha+\beta'X$ with $\alpha\in\mathbb{R}$ and $\beta\in\mathbb{R}^m$. It holds that $\frac{\partial}{\partial\beta}f(X,\alpha,\beta)=X$ and so we obtain the typical orthogonality conditions $\mathrm{E}(\varepsilon^*)=0$ and $\mathrm{Cov}(X,\varepsilon^*)=0$ of linear regression, i.e., the regressors $1,X_1,\ldots,X_m$ are *exogenous*.[4] However, it is worth emphasizing that exogeneity does not represent any model assumption. It is just a simple result of minimizing the mean square error $\mathrm{E}(\varepsilon^2)$ by using a linear regression function.

2.2.2. Description

If our aim is to *predict* some unobservable variable Y, we need no structural assumption.[5] Then, the goal is to find any observable variables $X_1,\ldots,X_m$ in order to minimize the mean square error $\mathrm{E}(\varepsilon^2)$ of the regression equation $Y=f(X)+\varepsilon$. Since the regressors $X_1,\ldots,X_m$ are intended only to predict the variable Y, their choice is rather arbitrary and so they need not have any particular meaning. In any case, the stronger the prediction power of $f(X)$, the lower the mean square error. Hence, prediction means to search for some appropriate regressors and to combine these variables by taking some regression function from $\mathcal{F}$ in order to minimize $\mathrm{E}\big(\varepsilon^2\big)$.

By contrast, if we want to *describe* the impact of X on Y, the situation is completely different. Then, the regressors are chosen in order to explain the relationship between X and Y. Hence, they *have* a particular meaning and so their choice is *not* arbitrary. Moreover, the given regression function should be valid. To be more precise, let $x\mapsto g(x)=\mathrm{E}(Y\,|\,X=x)$ be a real-valued function on D that quantifies the conditional mean of Y given $X=x$. The function g is referred to as the *true* regression function of Y given X.[6] Correspondingly, $Y=g(X)+\epsilon$ is said to be the true regression equation. Further, the regression function $f\in\mathcal{F}$ is called *valid* if and only if

$$f(X)=\mathrm{E}(Y\,|\,X)\quad\Longleftrightarrow\quad\mathrm{E}(\varepsilon\,|\,X)=0\,.$$

Simply put, the regression function f is valid if and only if $f(X)=g(X)$. Two regression

[3] In the special case of $\theta=\mu$, the second condition evaporates.

[4] There exist several nonequivalent definitions of exogeneity in the literature. Throughout this work, a regressor is said to be exogenous if and only if it is not correlated with the residual of the given regression model.

[5] Some authors refer to "control" when speaking about description (see, e.g., Fomby et al., 1984, p. 400).

[6] From $\mathrm{E}(Y^2)<\infty$ it follows that $\infty>\mathrm{Var}(Y)=\mathrm{Var}\big(\mathrm{E}(Y|X)\big)+\mathrm{E}\big(\mathrm{E}(Y^2\,|\,X)\big)+\mathrm{E}\big(\mathrm{E}^2(Y\,|\,X)\big)$, i.e., $\mathrm{E}\big(\mathrm{E}^2(Y\,|\,X)\big)=\mathrm{E}\big(g^2(X)\big)<\infty$, which means that $g(X)\in L^2$. Thus, g is indeed a regression function.

functions $\hat{f}$ and $\tilde{f}$ are considered identical if and only if $\hat{f}(X) = \tilde{f}(X)$. Thus, f is valid if and only if $f = g$. The family $\mathcal{F}$ is called *adequate* if and only if it contains a valid regression function, i.e., $g \in \mathcal{F}$. By contrast, if $\mathcal{F}$ is inadequate, we cannot describe the impact of X on Y by a regression model $Y = f(X) + \varepsilon$ with $f \in \mathcal{F}$, appropriately. Then, we can at best minimize the mean square description error

$$\mathrm{E}\big((\varepsilon - \epsilon)^2\big) = \mathrm{E}\big((g(X) - f(X))^2\big)\,.$$

Validity is a substantial model assumption.[7] Therefore, whenever we use a regression equation $Y = f(X) + \varepsilon$ for the sake of description, it is considered a *regression model*. A regression model is said to be valid if and only if the corresponding regression function, f, is valid. Another expression for validity, which can often be found in the literature, is to say that the regression model is "well-specified." However, this terminology seems to be ambiguous.[8]

For example, a linear regression model presumes that

$$\mathrm{E}(Y \mid X) = \alpha + \beta' X\,.$$

Another well-known example is the probit model, where Y is a binary variable, so that

$$\mathrm{E}(Y \mid X) = \mathrm{P}(Y = 1 \mid X) = \Phi(\alpha + \beta' X).$$

More generally, a generalized linear model implies that

$$\mathrm{E}(Y \mid X) = l^{-1}(\alpha + \beta' X),$$

where $l \colon \mathbb{R} \to \mathbb{R}$ is referred to as a link function and it is presumed that l is invertible.

One typically tries to quantify the *marginal* impact of X on Y, i.e., $\frac{\partial}{\partial x} g(X)$, in which case it is implicitly assumed that the true regression function g is differentiable, almost everywhere. For example, let the linear regression model $Y = \alpha + \beta' X + \varepsilon$ be valid. Then, the marginal impact of X on Y is just β. Further, if the probit model is valid, the marginal impact is $\phi(\alpha + \beta' X)\beta$. More generally, for a (valid) generalized linear model with link function l, we obtain the marginal impact $\beta / l'\big(l^{-1}(\alpha + \beta' X)\big)$. Here, l' symbolizes the derivative of l, which is presumed to exist and to be nonzero at $l^{-1}(\alpha + \beta' X)$, almost surely. However, if the given regression model $Y = f(X) + \varepsilon$ is invalid, the marginal impact of X on Y can be grossly misjudged by $\frac{\partial}{\partial x} f(X)$, i.e., the partial derivative of the chosen regression function $f \in \mathcal{F}$.

I focus on the mean of Y conditional on X, although other characteristics of the (conditional) distribution of Y can be interesting as well. Indeed, one might argue that mean regression is inappropriate if the distribution of Y is skewed or discontinuous. This argument presumes that we aim at quantifying another functional like, e.g., a quantile (Koenker and Basset, 1978, Koenker, 2005) or an expectile (Aigner et al., 1976, Newey and Powell, 1987),[9] or that we even

[7] Some authors require even more than validity. E.g., Hastie et al. (2009, p. 28) presume also that ε is independent of X.

[8] I will come back to this point in Section 4.

[9] See, e.g., Schulze Waltrup et al. (2015) for a nice overview of these measures.

want to assess the overall distribution of Y given X. This has become increasingly popular during the last years (Kneib et al., 2023). Then, mean regression is certainly not the best choice. However, this is not the intention behind this work. Here, I deliberately refer to the conditional *mean* of Y. I have chosen the mean-regression approach mainly for three reasons:

1. Mean regression is still the most popular regression approach in empirical research.
2. It is quite appealing from a mathematical point of view, since we are able to apply well-known rules from probability theory in order to derive the desired results.
3. There is a close relationship between the mean square error and the mean conditional error, i.e., between prediction and description, which will be elaborated below.

The latter point is probably the main reason for confusion and misunderstanding in applied econometrics. Thus, I will clarify this point before presenting and discussing the validity test.

2.3. The Connection between Prediction and Description

Hence, the goal of prediction is to minimize the mean square *prediction* error

$$\mathrm{E}(\varepsilon^2) = \mathrm{E}\left(\big(Y - f(X)\big)^2\right),$$

whereas description aims at minimizing the mean square *description* error

$$\mathrm{E}\big((\varepsilon - \epsilon)^2\big) = \mathrm{E}\big((g(X) - f(X))^2\big).$$

These goals are distinct and should thus be treated differently in practical applications.

Before going further, I would like to make two basic but quite important observations:

1. There always *exists* a valid regression function, viz., $x \mapsto g(x) = \mathrm{E}(Y \mid X = x)$, and
2. there cannot exist any other valid regression function, i.e., g is *unique*.

This leads us to our first proposition.[10]

Proposition 1. *The true regression function of Y given X is the only valid regression function in $\mathcal{G}$.*

For example, let the regression model $Y = \hat{f}(X) + \hat{\varepsilon}$ with $\hat{f} \in \mathcal{F}$ be valid. If the regression model $Y = \tilde{f}(X) + \tilde{\varepsilon}$ with $\tilde{f} \in \mathcal{F}$ is valid, too, then it holds that $\tilde{f}(X) = g(X) = \hat{f}(X)$ and thus $\tilde{\varepsilon} = \hat{\varepsilon}$. Hence, $Y = \tilde{f}(X) + \tilde{\varepsilon}$ is *essentially* the same as $Y = \hat{f}(X) + \hat{\varepsilon}$. Thus, it makes no difference at all whether we use $\hat{f}$ or $\tilde{f}$ to describe the impact of X on Y.

The following proposition observes that every optimal regression function represents the best possible fit to the true regression function, irrespective of whether or not the given family $\mathcal{F}$ of regression functions is adequate. Hence, it is natural to use an optimal predictor of Y if we want to describe the impact of X on Y as best as possible. Nonetheless, our main goal still is to *describe*

[10] All nontrivial proofs of the following statements can be found in the appendix.

Validity	Case
$V^2 = 0$	$f(X) \neq g(X) = Y$
$0 < V^2 < 1$	$f(X) \neq g(X) \neq Y$
$V^2 = 1$	$f(X) = g(X)$, i.e., f is valid

Table 1: Different cases of validity.

the impact of X on Y and not to *predict* Y. Otherwise, we should take some additional regressors into account in order to increase our prediction power, i.e., to decrease the mean square error.[11]

Proposition 2. *A regression function $\hat{f} \in \mathcal{G}$ is optimal among $\mathcal{F}$ if and only if*

$$\hat{f} \in \underset{f \in \mathcal{F}}{\arg\min}\, \mathrm{E}\left(\big(g(X) - f(X)\big)^2\right).$$

In any case, for each regression function $f \in \mathcal{G}$, we have that

$$\mathrm{E}\big((Y - f(X))^2\big) = \mathrm{E}\big((Y - g(X))^2\big) + \mathrm{E}\big((g(X) - f(X))^2\big). \tag{3}$$

Thus, $\mathrm{E}(\varepsilon^2) = \mathrm{E}(\epsilon^2) + \mathrm{E}\big((\varepsilon - \epsilon)^2\big)$ with $\varepsilon = Y - f(X)$ and $\epsilon = Y - g(X)$, i.e., $\mathrm{E}(\epsilon^2) \leq \mathrm{E}(\varepsilon^2)$.

Thus, we are able to decompose the mean square error $\mathrm{E}\big((Y - f(X))^2\big)$ into two parts:

1. The first part, $\mathrm{E}\big((Y - g(X))^2\big)$, measures the fluctuation of Y around $\mathrm{E}(Y \mid X)$, which *shall not* be explained by a regression of Y on X, whereas
2. the second part, $\mathrm{E}\big((g(X) - f(X))^2\big)$, quantifies the deviation of $f(X)$ from $\mathrm{E}(Y \mid X)$, which is zero if and only if the regression function f is valid.

This suggests to quantify the *validity* of the regression model $Y = f(X) + \varepsilon$ or, equivalently, of the corresponding regression function f, by

$$V^2 := \frac{\mathrm{E}(\epsilon^2)}{\mathrm{E}(\varepsilon^2)} = \frac{\mathrm{E}\big((Y - g(X))^2\big)}{\mathrm{E}\big((Y - f(X))^2\big)} \in \big[0, 1\big],$$

provided that $Y \neq f(X)$, i.e., $\varepsilon \neq 0$. Otherwise, we have that $f(X) = g(X)$ and so we may set $V^2 = 1$.[12] Anyway, f is valid if $V^2 = 1$ and it is invalid if $V^2 < 1$. In the particular case of $V^2 = 0$ we have that $Y = g(X) \neq f(X)$, i.e., f fails to capture the perfect functional relationship between X and Y. Table 1 summarizes the different cases of V^2.

In order to judge whether or not a given regression model $Y = f(X) + \varepsilon$ is appropriate, one typically uses the coefficient of determination

$$R^2 := 1 - \frac{\mathrm{E}(\varepsilon^2)}{\mathrm{Var}(Y)}.$$

[11] In fact, overfitting is not an issue at all when trying to predict Y by X, provided that we know P.

[12] From $Y = f(X)$ it follows that $g(X) = \mathrm{E}(Y \mid X) = \mathrm{E}(f(X) \mid X) = f(X)$.

To be more precise, R^2 measures the *prediction power* of the regression equation $Y = f(X) + \varepsilon$ or, equivalently, of the corresponding predictor $f(X)$. In fact, we have that $R^2 = 1$ if and only if $\mathrm{E}(\varepsilon^2) = 0$. Nonetheless, we can only guarantee that $R^2 \leq 1$ because $\mathrm{E}(\varepsilon^2)$ may very well be greater than $\mathrm{Var}(Y)$, in which case the coefficient of determination becomes negative.[13]

By applying R^2, one usually presumes that the mean of ε is zero and also that $f(X)$ and ε are uncorrelated. For example, consider a linear regression model $Y = \alpha + \beta' X + \varepsilon$ in which the exogeneity conditions $\mathrm{E}(\varepsilon) = 0$ and $\mathrm{Cov}(X, \varepsilon) = 0$ are satisfied. Then, we have that

$$R^2 = \frac{\mathrm{Var}(f(X))}{\mathrm{Var}(Y)} \in \big[0, 1\big],$$

in which case the coefficient of determination quantifies the proportion of the total variance of Y that can be explained by the variance of $f(X)$.

Furthermore, since $\epsilon = Y - g(X)$ with $g(X) = \mathrm{E}(Y \mid X)$, we have that $\mathrm{E}(\epsilon \mid X) = 0$ and thus $\mathrm{E}(\epsilon) = \mathrm{Cov}(g(X), \epsilon) = 0$. Hence, let

$$S^2 := 1 - \frac{\mathrm{E}(\epsilon^2)}{\mathrm{Var}(Y)} = \frac{\mathrm{Var}(g(X))}{\mathrm{Var}(Y)} \in \big[0, 1\big]$$

be the *explanation power* of X, i.e., the proportion of the total variance of Y that can be explained by the variance of $g(X)$.[14] The explanation power does not depend on the chosen regression function—it only depends on the choice of regressors.

The following theorem marks the basic result regarding the validity measure.

Theorem 1 (Validity)**.** *Let $Y = f(X) + \varepsilon$ be any regression model. In the case of $R^2 < 1$, we have that*

$$V^2 = \frac{1 - S^2}{1 - R^2}$$

and otherwise $V^2 = S^2 = 1$. In any case, it holds that $R^2 \leq S^2$, where $R^2 = S^2$ if and only if $V^2 = 1$.

Hence, the validity of a regression model $Y = f(X) + \varepsilon$ essentially depends both on the explanation power of X and on the prediction power of $f(X)$. This creates a trade-off, which can be best understood by observing Figure 1. We can see that a regression model is valid if and only if R^2 attains its maximum S^2. For example, if the explanation power, S^2, is zero, the coefficient of determination, R^2, must be zero, too, in order to obtain a valid regression model. Analogously, if $S^2 = 1$ also $R^2 = 1$ is required for $V^2 = 1$. In any case, the higher S^2 the lower V^2 for any given R^2. Put another way, the higher the explanation power of X, the higher the prediction power of $f(X)$ must be in order to guarantee that the given regression model is valid. This can lead to astonishing phenomena and adverse effects, which might often be overlooked in practical applications. I will come back to this point in Section 3.2.

The problem is that the explanation power of the chosen regressors, i.e., S^2, is unknown in real life and thus we cannot draw any conclusion about the validity of the regression model just

[13] A simple example is $Y = -1 + \varepsilon$ with $Y \sim \mathcal{N}(0, 1)$, so that $\varepsilon = Y + 1$ and thus $\mathrm{E}(\varepsilon^2) = 2$, i.e., $R^2 = -1$.

[14] Thus, S^2 can be considered the coefficient of determination of the true regression equation $Y = g(X) + \epsilon$.

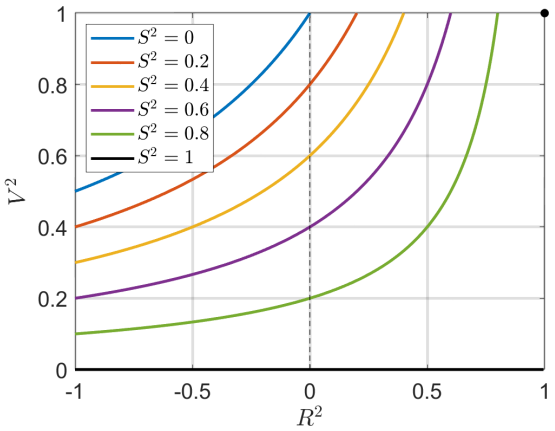


Figure 1: Dependence between R^2 and V^2 given S^2.

Ratio	Definition	Range	Meaning	Object
A^2	$S^2 + V^2 - 1$	$(-1, 1]$	Accuracy	$Y = f(X) + \varepsilon$
R^2	$1 - \frac{\mathrm{E}(\varepsilon^2)}{\mathrm{Var}(Y)}$	$(-\infty, 1]$	Prediction power	$f(X)$
S^2	$1 - \frac{\mathrm{E}(\epsilon^2)}{\mathrm{Var}(Y)}$	$[0, 1]$	Explanation power	X
V^2	$\frac{\mathrm{E}(\epsilon^2)}{\mathrm{E}(\varepsilon^2)}$	$[0, 1]$	Validity	f

Table 2: Regression ratios of $Y = f(X) + \varepsilon = g(X) + \epsilon$ with $g(X) = \mathrm{E}(Y \mid X)$.

by considering its coefficient of determination. In fact, a valid regression model $Y = f(X) + \varepsilon$ must be based on an optimal predictor $f(X)$. Nonetheless, if the explanation power of X is low, a regression model *must* have a weak prediction power. Theorem 1 reveals that the coefficient of determination of a valid regression model must even be zero if $S^2 = 0$. This demonstrates that R^2 shall not be used as a validity measure. The same holds true for any other measure that does not take the explanation power of the regressors into account, even if it controls for the number of parameters, like the Akaike information criterion or the Bayesian information criterion.

In practical applications, we typically want to find some regressors $X_1, \ldots, X_m$ with a strong explanation power. Then, the latter is considered an *additional* goal of description, i.e., it does not supersede the validity of the regression function f. Thus, we try to maximize the sum of

1. the explanation power of the regressors and
2. the validity of the regression model.

The explanation power, S^2, is a measure for our ability to *select* appropriate regressors, whereas the validity, V^2, quantifies our ability to *combine* those regressors. Simply put, for a delicious dinner we need both good ingredients and a careful preparation.

Thus, let us define the *accuracy* of the regression model $Y = f(X) + \varepsilon$ by

$$A^2 := S^2 + V^2 - 1 \in \big(-1, 1\big].$$

In fact, we always have that $A^2 > -1$ because, according to Theorem 1, $S^2 = 0$ implies $V^2 > 0$.

Theorem 2 (Accuracy)**.** *Let $Y = f(X) + \varepsilon$ be any regression model. We have that $A^2 = V^2 R^2$.*

Hence, the accuracy, A^2, of any regression model equals the product of its validity, V^2, and prediction power, R^2, where V^2 and R^2 are related to one another according to Figure 1 or, equivalently, by Theorem 1. Table 2 summarizes the regression ratios.

Ideally, we should achieve $A^2 = 1$, which implies $S^2 = V^2 = 1$, but this is virtually impossible in real-life applications. The trivial regression model is

$$Y = \mathrm{E}(Y) + \varepsilon\,,$$

in which the basic set of regressors is empty. This model is always valid, but we have that $S^2 = 0$ and thus $A^2 = 0$. Hence, we should at least try to accomplish $A^2 > 0$, i.e., $S^2 > 0$ *and* $V^2 > 0$.

2.4. General Conditions for Validity

This section contains some general conditions for the validity of a regression model.

2.4.1. Necessary Conditions

Theorem 3 (Necessary Conditions)**.** *Suppose that the family $\mathcal{F}$ is adequate and let $Y = \hat{f}(X) + \hat{\varepsilon}$ with $\hat{f} \in \mathcal{F}$ be some valid regression model. The following assertions hold true:*

(i) $\mathrm{E}(\hat{\varepsilon}) = 0$

(ii) $\mathrm{Var}(\hat{\varepsilon}) = \mathrm{E}\big(\mathrm{Var}(\hat{\varepsilon} \,|\, X)\big)$

(iii) $\mathrm{Cov}(h(X), \hat{\varepsilon}) = 0$ *for every real-valued function h on D with $h(X) \in L^2$.*

(iv) The regression function $\hat{f}$ is optimal among $\mathcal{F}$.

(v) If the regression function $\tilde{f} \in \mathcal{F}$ is optimal among $\mathcal{F}$, too, then $\tilde{f} = \hat{f}$. This means that also the regression model $Y = \tilde{f}(X) + \tilde{\varepsilon}$ is valid and we have that $\tilde{\varepsilon} = \hat{\varepsilon}$.

Theorem 3 immediately leads us to the following corollary. Thus, its proof can be skipped.

Corollary 1. *If the conditions of Theorem 3 are satisfied, the following assertions hold true:*

(i) $\mathrm{E}(\hat{\varepsilon}^2) = \mathrm{Var}(\hat{\varepsilon})$

(ii) $\mathrm{Cov}(X, \hat{\varepsilon}) = 0$

(iii) $\mathrm{Cov}\big(\hat{f}(X), \hat{\varepsilon}\big) = 0$

(iv) $\mathrm{Cov}\big(f(X),\hat{\varepsilon}\big) = 0$ *for all* $f \in \mathcal{F}$.

(v) $\mathrm{E}\big(f(X)\hat{\varepsilon}\big) = 0$ *for all* $f \in \mathcal{F}$.

Thus, X and ε are uncorrelated if the regression model $Y = f(X) + \varepsilon$ is valid. Nonetheless, they need not be independent. For example, suppose that (X, Y) has an elliptical distribution (Cambanis et al., 1981, Kelker, 1970) with $\mathrm{Var}(X) > 0$. Then, the linear regression model $Y = \alpha + \beta' X + \varepsilon$ with $\beta = \mathrm{Var}(X)^{-1}\mathrm{Cov}(X, Y)$ and $\alpha = \mathrm{E}(Y) - \beta'\mathrm{E}(X)$ is valid. However, ε is independent of X only if the (multivariate) distribution of (X, Y) is normal.

We conclude that the typical exogeneity conditions of linear regression, i.e.,

$$\begin{aligned} \mathrm{E}(\varepsilon) &= 0 \\ \mathrm{Cov}(X, \varepsilon) &= 0, \end{aligned} \tag{4}$$

are always satisfied if the regression model $Y = f(X) + \varepsilon$ is valid, which holds true even if f is *nonlinear*. However, the exogeneity of the chosen regressors $X_1, \ldots, X_m$ is only a necessary, but not a sufficient condition for validity. By using a parametric family of regression functions, exogeneity can often be accomplished by specifying $\theta \in \Theta$ such that $\mathrm{E}\big(f(X,\theta)\big) = \mathrm{E}(Y)$ and $\mathrm{Cov}(X, Y) = \mathrm{Cov}\big(X, f(X,\theta)\big)$, but this does not guarantee that the regression model is valid. For example, the regressors in a linear regression model $Y = \alpha + \beta' X + \varepsilon$ with $\mathrm{Var}(X) > 0$ and $\beta = \mathrm{Var}(X)^{-1}\mathrm{Cov}(X, Y)$ are always exogenous just by construction. Nonetheless, this does not mean that the linear regression model is valid. I will come back to this crucial point later on.

2.4.2. Sufficient Conditions

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a sample of (not necessarily independent) observations of (X, Y). For notational convenience, from now on I will use the random $m \times n$ matrix

$$\mathbf{X} = \begin{bmatrix} X_{11} & \cdots & X_{1n} \\ \vdots & & \vdots \\ X_{m1} & \cdots & X_{mn} \end{bmatrix}$$

to symbolize the sample observations $X_1, \ldots, X_n$ of X. Correspondingly,

$$\mathbf{x} = \begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & & \vdots \\ x_{m1} & \cdots & x_{mn} \end{bmatrix}$$

denotes some realization of $\mathbf{X}$. Thus, $\mathbf{x}$ is a real-valued $m \times n$ matrix of *fixed* regressor values. Moreover, $\mathbf{Y} = (Y_1, \ldots, Y_n)$ contains the sample observations of Y and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)$ contains the corresponding sample errors with $\varepsilon_i = Y_i - f(X_i)$ for $i = 1, \ldots, n$.

Now, let f be a linear regression function. Hayashi (2000, p. 7) says that $\mathbf{X}$ is *strictly exogenous* if and only if $\mathrm{E}(\boldsymbol{\varepsilon} \mid \mathbf{X}) = 0$. Strict exogeneity is a classical, but quite restrictive assumption of linear

regression analysis. However, we can readily extend Hayashi's definition of strict exogeneity to nonlinear regression models. To be more precise, the (linear or nonlinear) regression model $Y = f(X) + \varepsilon$ with $f \in \mathcal{F}$ satisfies the strict exogeneity assumption if and only if $\mathrm{E}(\boldsymbol{\varepsilon} \mid \mathbf{X}) = 0$. The *Gaussian model* (GM) presumes that $\mathbf{Y} = \alpha 1 + \mathbf{X}'\beta + \boldsymbol{\varepsilon}$, $n > m$, $\mathrm{rk}\big[1 \ \mathbf{X}'\big] = m + 1$, and $\boldsymbol{\varepsilon} \mid \mathbf{X} \sim \mathcal{N}_n(0, \sigma^2 \mathbf{I}_n)$ with $\sigma^2 > 0$. It represents another classical assumption of linear regression analysis. Obviously, the GM implies strict exogeneity.

This leads us to the next theorem, which provides sufficient conditions for the validity of a regression model.

Theorem 4 (Sufficient Conditions)**.** *Let $f \in \mathcal{F}$ be some regression function. If anyone of the following assertions holds true, the regression model $Y = f(X) + \varepsilon$ is valid and so the family $\mathcal{F}$ is adequate.*

(i) $\mathrm{E}(\varepsilon) = 0$ *and* ε *is independent of* X*.*

(ii) $\mathrm{E}(\varepsilon) = 0$ *and* $\mathrm{Var}(\varepsilon \mid X) = \mathrm{Var}(\varepsilon)$*.*

(iii) $\mathbf{X}$ *is strictly exogenous.*

(iv) The GM is satisfied.

Thus, if we have found a regression function f such that $\mathrm{E}(\varepsilon) = 0$ and ε is independent of X, we can be sure that f is valid. However, this goes far beyond validity and, in general, there is no regression function at all that satisfies this quite ambitious condition. For example, suppose once again that (X, Y) possesses an elliptical distribution, where the covariance matrix of X is positive definite. It has already been observed that there exists a (unique) valid regression model $Y = \alpha + \beta' X + \varepsilon$, but ε cannot be independent of X unless (X, Y) is normally distributed.

The next theorem emphasizes the special role of elliptical distributions in linear regression analysis. It states that we can select, i.e., exclude or include, any component of an elliptically distributed random vector at discretion in order to create a valid linear regression model.

Theorem 5 (Elliptical Distributions)**.** *Suppose that $Z = (Z_1, \ldots, Z_d)$ with $Z_1, \ldots, Z_d \in L^2$ is some random vector possessing an elliptical distribution with $\mathrm{Var}(Z) > 0$. Further, let $\{X_1, \ldots, X_m, Y\}$ be any subset of $\{Z_1, \ldots, Z_d\}$. Then, the linear regression model*

$$Y = \alpha + \beta' X + \varepsilon$$

with $X = (X_1, \ldots, X_m)$ is valid if and only if the parameters α and β are such that System 4 is satisfied, which is equivalent to $\beta = \mathrm{Var}(X)^{-1}\mathrm{Cov}(X, Y)$ and $\alpha = \mathrm{E}(Y) - \beta'\mathrm{E}(X)$.

2.4.3. Equivalent Conditions

The following theorem provides equivalent conditions for validity.

Theorem 6 (Equivalent Conditions)**.** *The regression model $Y = f(X) + \varepsilon$ with $f \in \mathcal{F}$ is valid, and so the family $\mathcal{F}$ is adequate, if and only if the following equivalent assertions hold true:*

(i) The regression function f is optimal among $\mathcal{G}$.

(ii) $\mathrm{E}(\varepsilon) = 0$ *and* $\mathrm{Var}(\varepsilon) = \mathrm{E}(\mathrm{Var}(\varepsilon \mid X))$.

(iii) $\mathrm{E}(\varepsilon^2) = \mathrm{E}\big((Y - g(X))^2\big)$

The first part of that theorem asserts that a regression function f is valid if and only if it is optimal among the set $\mathcal{G}$ of *all* regression functions. Put another way, validity is equivalent to *global* optimality. Thus, if we want to describe the impact of X on Y, appropriately, we must find the best predictor $f(X)$ of Y among the set $\mathcal{G}(X) := \big\{f(X) : f \in \mathcal{G}\big\}$ of all possible predictors of Y based on X. This goal can be highly ambitious if the true regression function, g, is not simple.

2.5. Projection Theorems

Let $\mathcal{F}$ be any family of regression functions. We have that

$$\mathrm{E}\big((Y - f(X))^2\big) = \mathrm{Var}\big(Y - f(X)\big) + \mathrm{E}^2\big(Y - f(X)\big)$$

for all $f \in \mathcal{F}$. A translation of f affects only the mean, but not the variance of $Y - f(X)$. Thus, if $\mathcal{F}$ is closed under translations and the regression function $\hat{f}$ is optimal among $\mathcal{F}$, it must hold that $\mathrm{E}(\hat{\varepsilon}) = 0$ with $\hat{\varepsilon} = Y - \hat{f}(X)$. Further, let $\mathcal{F}(X) := \big\{f(X) : f \in \mathcal{F}\big\}$ be the set of all predictors of Y that are obtained by choosing some regression function f from $\mathcal{F}$ and applying f to the vector X of regressors. The error of the regression model $Y = \tilde{f}(X) + \tilde{\varepsilon}$ with $\tilde{f} \in \mathcal{F}$ is said to be orthogonal to $\mathcal{F}(X)$ if and only if $\mathrm{E}\big(f(X)\tilde{\varepsilon}\big) = 0$ for all $f \in \mathcal{F}$. The following theorem is an immediate consequence of Hilbert's projection theorem and thus its proof can be skipped.

Theorem 7 (Projection Theorem I)**.** *Let $\mathcal{F}(X)$ be a closed and convex subset of L^2.*

(i) The family $\mathcal{F}$ contains a unique regression function f that is optimal among $\mathcal{F}$.

(ii) If $\mathcal{F}(X)$ is a vector subspace of L^2, then f is the unique element of $\mathcal{F}$ such that $\varepsilon = Y - f(X)$ is orthogonal to $\mathcal{F}(X)$.

Let $\mathcal{V}$ be any family of regression functions such that $\mathcal{V}(X)$ is a vector subspace of L^2. Thus, if f is the optimal regression function among $\mathcal{V}$, it holds that $\mathrm{E}(f(X)\varepsilon) = 0$ with $\varepsilon = Y - f(X)$. Moreover, if $\mathcal{V}$ is closed under translations, we have that $\mathrm{E}(\varepsilon) = 0$ and thus $\mathrm{Cov}(f(X), \varepsilon) = 0$.[15] For example, let $\mathcal{L} := \big\{x \mapsto a + b'x : a \in \mathbb{R}, b \in \mathbb{R}^m\big\}$ be the family of linear regression functions. In this case, the typical exogeneity conditions $\mathrm{E}(\varepsilon) = 0$ and $\mathrm{Cov}(X, \varepsilon) = 0$ are satisfied if and only if ε is orthogonal to $\mathcal{L}(X)$, which means that $f(X)$ is the best linear predictor based on X.

The next proposition guarantees that $f \in \mathcal{F}$ is optimal if the corresponding error $\varepsilon = Y - f(X)$ is orthogonal to $\mathcal{F}(X)$, provided that the family $\mathcal{F}$ contains an optimal regression function at all. This holds true irrespective of whether or not the family $\mathcal{F}$ is adequate.

[15] This implies that $R^2 \geq 0$, i.e., in this case the coefficient of determination quantifies the proportion of the total variance of Y that can be explained by the variance of $f(X)$.

Proposition 3. *Suppose that the regression function $\hat{f} \in \mathcal{F}$ is optimal among $\mathcal{F}$ and consider another regression function $\tilde{f} \in \mathcal{F}$. Then, the error $\tilde{\varepsilon} = Y - \tilde{f}(X)$ cannot be orthogonal to $\mathcal{F}(X)$.*

The next theorem is similar to Theorem 7 and thus it can be considered a variant of Hilbert's projection theorem. However, it does not require that $\mathcal{F}(X)$ is some vector subspace of L^2. It even need not be closed and convex. The essential requirement is that the family $\mathcal{F}$ is adequate.

Theorem 8 (Projection Theorem II)**.** *If the family $\mathcal{F}$ is adequate, it contains a unique regression function f that is optimal among $\mathcal{F}$. The regression function f coincides with the unique valid regression function in $\mathcal{F}$, and f is the unique element of $\mathcal{F}$ such that $\varepsilon = Y - f(X)$ is orthogonal to $\mathcal{F}(X)$.*

Hence, if the family $\mathcal{F}$ is adequate, a regression function that is optimal among $\mathcal{F}$ (and thus valid) is always characterized by an orthogonal projection of Y onto $\mathcal{F}(X)$. Otherwise, i.e., if $\mathcal{F}$ is inadequate, there can very well exist some regression function $\hat{f}$ that is optimal among $\mathcal{F}$ and this may even be unique. Nonetheless, $\hat{f}$ cannot be valid.[16] In this case, the error $\hat{\varepsilon} = Y - \hat{f}(X)$ even need not be orthogonal to $\mathcal{F}(X)$. For example, suppose that $Y = 0X + \epsilon$, where $X, \epsilon \sim \mathcal{N}(0,1)$ are independent, and let $\mathcal{F}(X) = \{\lambda + X \colon \lambda \in \mathbb{R}\}$ be the set of predictors of Y. Obviously, the family $\mathcal{F}$ is inadequate. It turns out that $\hat{f}(X) = X$ is the optimal predictor of Y among $\mathcal{F}(X)$. However, the error $\hat{\varepsilon} = Y - \hat{f}(X)$ is *not* orthogonal to $\mathcal{F}(X)$, since $\hat{\varepsilon} = \epsilon - X$ and thus $\mathrm{E}\big(\hat{f}(X)\hat{\varepsilon}\big) = \mathrm{E}\big(X(\epsilon - X)\big) = -1$. In fact, $\mathcal{F}(X)$ is closed and convex, but it is not a vector subspace of L^2. However, $\mathcal{F}$ is closed under translations and so we have that $\mathrm{E}(\hat{\varepsilon}) = 0$. Nonetheless, X is endogenous, since $\mathrm{Cov}(X, \hat{\varepsilon}) = -1$, too.

2.6. Hierarchy of Regression Properties

Although validity implies both optimality and exogeneity, in general, the latter properties are neither necessary nor sufficient for one another. It has been shown at the end of Section 2.5 that an optimal regression function need not produce exogenous regressors and thus it can very well violate the typical exogeneity conditions of linear regression. Conversely, if $\mathcal{F}$ is a family of *nonlinear* regression functions, it can happen that $f \in \mathcal{F}$ is suboptimal although it satisfies the exogeneity conditions. Then, exogeneity can even *prevent* f from being optimal.

For example, consider the family of cubic regression functions of the form $x \mapsto f(x, \alpha, \beta) = \alpha + \frac{\beta}{3}x^3$ with $\alpha, \beta \in \mathbb{R}$ and suppose that $Y = X + \epsilon$, where $X, \epsilon \sim \mathcal{N}(0,1)$ are independent. Thus, we obtain $\frac{\partial}{\partial x} f(x, \alpha, \beta) = \beta x^2$, $\frac{\partial}{\partial \alpha} f(x, \alpha, \beta) = 1$, and $\frac{\partial}{\partial \beta} f(x, \alpha, \beta) = \frac{1}{3}x^3$. The regressor X is exogenous if and only if $\mathrm{Cov}(X, Y) = \mathrm{Cov}\big(X, f(X, \alpha, \beta)\big)$, i.e., $\mathrm{E}(XY) = \mathrm{E}\big(Xf(X, \alpha, \beta)\big)$. We have that $\mathrm{E}(XY) = 1$ and Stein's lemma reveals that $\mathrm{E}\big(Xf(X, \alpha, \beta)\big) = \mathrm{E}\big(\frac{\partial}{\partial x} f(X, \alpha, \beta)\big) = \beta$. Hence, the exogeneity conditions $\mathrm{E}(\varepsilon) = \mathrm{Cov}(X, \varepsilon) = 0$ are satisfied if and only if $\alpha = 0$ and $\beta = 1$. Due to Equation 2, optimality requires $\mathrm{E}\big(\frac{\partial}{\partial \beta} f(X, \alpha, \beta)\, \varepsilon\big) = \frac{1}{3}\mathrm{E}\big(X^3\varepsilon\big) = 0$, i.e., $\mathrm{E}\big(X^3\varepsilon\big) = 0$, but with $\alpha = 0$ and $\beta = 1$ we obtain $\mathrm{E}\big(X^3\varepsilon\big) = -2$.[17] By contrast, for $\alpha = 0$ and $\beta = \frac{3}{5}$, the cubic regression function becomes optimal. Hence, if the given regression function satisfies the typical exogeneity conditions, it cannot be optimal.

[16] A typical example is a linear regression of Y on X where the true regression function of Y given X is nonlinear.

[17] This can be shown by using the formula $\mathrm{E}(X^k) = k!/\big(2^{\frac{k}{2}}\tfrac{k}{2}!\big)$ for each even integer k.

We conclude that the GM is stronger than strict exogeneity, which implies validity, which is equivalent to global optimality, which is sufficient for optimality, orthogonality, and exogeneity. Further, if the predictor $f(X)$ stems from some vector subspace of L^2, i.e., $\mathcal{F} = \mathcal{V}$, optimality and orthogonality are equivalent, and if $\mathcal{F}$ contains all linear regression functions, orthogonality implies exogeneity. In particular, if we focus on linear regression analysis, i.e., $\mathcal{F} = \mathcal{L}$, then optimality, orthogonality, and exogeneity are even equivalent. Moreover, if $\mathcal{F}$ is adequate, optimality, orthogonality, validity, and global optimality are equivalent. Finally, if $\mathcal{F} = \mathcal{L}$ is adequate, optimality, orthogonality, validity, global optimality, and exogeneity are equivalent. A typical example is a linear regression of Y on X where (X, Y) is elliptically distributed. However, in general, exogeneity is only a necessary but not a sufficient condition for validity.

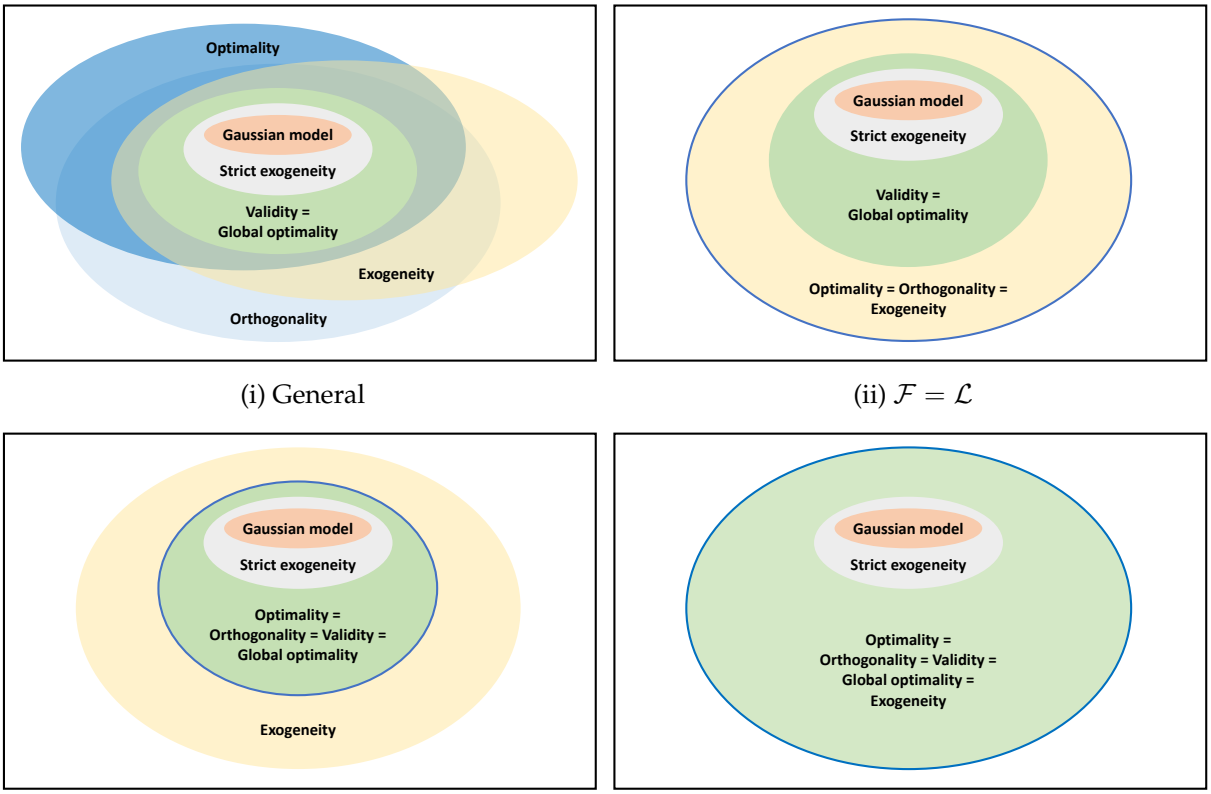
The following theorem summarizes our previous findings, where the abbreviation

- V means that $f \in \mathcal{F}$ is *valid*, i.e., $\mathrm{E}(\varepsilon \mid X) = 0$,
- GM means that the *Gaussian model* is satisfied,
- SE means that $\mathbf{X}$ is *strictly exogenous*, i.e., $\mathrm{E}(\boldsymbol{\varepsilon} \mid \mathbf{X}) = 0$,
- OP means that f is *optimal* among $\mathcal{F}$, i.e., $f \in \arg\min_{\mathcal{F}} \mathrm{E}(\varepsilon^2)$,
- GOP means that f is *globally optimal*, i.e., it is optimal among $\mathcal{G}$,
- EX means *exogeneity*, i.e., $\mathrm{E}(\varepsilon) = 0$ and $\mathrm{Cov}(X, \varepsilon) = 0$, whereas
- OR means *orthogonality*, i.e., $\mathrm{E}\big(f(X)\varepsilon\big) = 0$ for all $f \in \mathcal{F}$.

Theorem 9 (Hierarchy)**.** *Let $\mathcal{F} \subseteq \mathcal{G}$ be any family of regression functions, $\mathcal{V}$ be a family of regression functions such that $\mathcal{V}(X)$ is a vector subspace of L^2, and $\mathcal{L}$ be the family of linear regression functions.*

- GM $\Rightarrow$ SE $\Rightarrow$ V $\Leftrightarrow$ GOP $\Rightarrow$ OP $\wedge$ OR $\wedge$ EX
- $\mathcal{F} = \mathcal{V}$*:* OP $\Leftrightarrow$ OR
- $\mathcal{F} \supseteq \mathcal{L}$*:* OR $\Rightarrow$ EX
- $\mathcal{F} = \mathcal{L}$*:* OP $\Leftrightarrow$ OR $\Leftrightarrow$ EX
- $g \in \mathcal{F}$*:* OP $\Leftrightarrow$ OR $\Leftrightarrow$ V $\Leftrightarrow$ GOP $\Rightarrow$ EX
- $g \in \mathcal{F} = \mathcal{L}$*:* OP $\Leftrightarrow$ OR $\Leftrightarrow$ V $\Leftrightarrow$ GOP $\Leftrightarrow$ EX

Figure 2 illustrates some important aspects of Theorem 9. In most cases, exogeneity is only a necessary but not a sufficient condition for validity. This holds true even if we concentrate on the family of linear regression functions, i.e., $\mathcal{F} = \mathcal{L}$, or if $\mathcal{F}$ is adequate, i.e., $g \in \mathcal{F}$. Hence, a (linear) regression model can very well satisfy the exogeneity conditions without being valid. This underpins the importance of a *genuine* validity check in practical applications, provided that we want to describe the impact of X on Y and not to predict Y by X.



(i) General (ii) $\mathcal{F} = \mathcal{L}$

(iii) $\mathcal{F}$ is adequate (iv) $\mathcal{F} = \mathcal{L}$ is adequate

Figure 2: Hierarchy of regression properties.

2.7. The Basic Set of Regressors

Suppose for the sake of simplicity but without loss of generality that the true regression equation is $Y = X_1 + X_2 + \epsilon$, where $X_1, X_2, \epsilon \sim \mathcal{N}(0,1)$ are independent, and assume that $\mathcal{F} = \mathcal{L}$. Since we have that $g(X) = X_1 + X_2$ with $X = (X_1, X_2)$, the family of linear regression functions is adequate and Theorem 9 tells us that validity and exogeneity are equivalent in this particular case. Now, consider the linear regression model $Y = X_1 + \varepsilon$, which means that $\varepsilon = X_2 + \epsilon$. Since it holds that $\mathrm{E}(\varepsilon) = \mathrm{Cov}(X_1, \varepsilon) = 0$, the typical exogeneity conditions of linear regression seem to be satisfied. Further, we also have that $\mathrm{E}(\varepsilon \,|\, X_1) = 0$, which suggests that the mean conditional error of the given regression model is zero, too. Nonetheless, it turns out that the simple regression model is invalid, since $f(X) = X_1 \neq X_1 + X_2 = g(X)$! What is the reason? Actually, we represent the dependent variable Y by two regressors, i.e., X_1 and X_2. Hence, there are in fact *three* exogeneity conditions, but the third one is violated because $\mathrm{Cov}(X_2, \varepsilon) = 1 \neq 0$. Further, $\mathrm{E}(\varepsilon \,|\, X_1, X_2) = X_2 \neq 0$ clearly indicates that the given regression model is invalid.

Therefore, we should always keep in mind that the entire concept of description requires us to fix some (random) vector $X = (X_1, \ldots, X_m)$ of regressors *in advance*, which must not be changed, intermediately. Even if we choose some regression function $f \in \mathcal{F}$ that is not influenced by some component of X at all, we still try to describe the impact of *all* regressors, i.e., $X_1, \ldots, X_m$, on the dependent variable Y. Put another way, we try to find $\mathrm{E}(Y \,|\, X_1, \ldots, X_m)$. By contrast, suppose

that we would have chosen only X_1 as regressor right from the start. Then, the linear regression model $Y = X_1 + \varepsilon$, in fact, would have satisfied all exogeneity conditions and so it would have been valid. Consequently, the mean conditional error of this (simple) linear regression model, i.e., $\mathrm{E}(\varepsilon \mid X_1)$, would have been zero, too. To sum up, whether or not a linear regression model suffers from an *omitted-variable bias* depends on our basic set of regressors, i.e., $\{X_1, \ldots, X_m\}$.

For example, consider the true regression equation

$$Y = a + bX + cW + \epsilon,$$

where the 3-dimensional random vector (W, X, Y) possesses an elliptical distribution with positive definite covariance matrix and the parameters $a, b, c \in \mathbb{R}$ are as described in Theorem 5, i.e., System 4 is satisfied. Further, let us assume that $\mathrm{Cov}(X, W) \neq 0$ and $c \neq 0$. Hence, the linear regression model

$$Y = a + bX + \varepsilon \tag{5}$$

is invalid, since X becomes endogenous after omitting the regressor W.[18] By contrast, the linear regression model

$$Y = \alpha + \beta X + \varepsilon, \tag{6}$$

in which α and β are chosen according to Theorem 5, too, is still valid and so the single regressor X is exogenous. The reason is that, in fact, we do *not* omit the variable W in Equation 6, since in this case our basic set of regressors is just $\{X\}$. Thus, we try to explain only the impact of X on Y. By contrast, in Equation 5 we try to explain the impact of X *and* W on Y. More precisely, the linear regression model reads

$$Y = a + bX + 0W + \varepsilon\,.$$

Thus, our basic set of regressors is $\{X, W\}$, but we omit the regressor W and so the resulting regression model becomes invalid.

2.8. Variable Selection

Correspondingly, there seems to be a widespread opinion, namely that we should select the "right" (number of) regressors when applying regression analysis. More precisely, consider the linear regression model

$$Y = \alpha + \beta' X + \varepsilon$$

based on $X = (X_1, \ldots, X_m)$ and suppose that all exogeneity conditions are satisfied. Further, assume that $\beta_1, \ldots, \beta_m \neq 0$ and let $X_1, \ldots, X_m$ be correlated with each other. Now, it is typically argued that one should not exclude or include any (other) regressor (see, e.g., Fomby et al., 1984, Section 18.2.1). The arguments go like this: Excluding some regressor leads to endogeneity and including another regressor makes no sense at all because its regression coefficient is zero.

Unfortunately, both arguments are flawed. In fact, we can very well ignore each regressor

[18] From $\varepsilon = cW + \epsilon$ we conclude that $\mathrm{Cov}(X, \varepsilon) = c\,\mathrm{Cov}(X, W) \neq 0$.

without producing endogeneity and, in general, the regression coefficient of any additional regressor is nonzero. Ignoring some variable just means to reduce our basic set of regressors, whereas taking some additional variable into account means to extend that set. However, for each arbitrary set of regressors, we can calculate the regression parameters $\alpha, \beta_1, \ldots, \beta_m$ by

$$\beta = \mathrm{Var}(X)^{-1}\mathrm{Cov}(X,Y) \qquad \text{and} \qquad \alpha = \mathrm{E}(Y) - \beta'\mathrm{E}(X),$$

so that the linear regression model $Y = \alpha + \beta' X + \varepsilon$ obeys all exogeneity conditions.

For example, suppose that $\mathrm{E}(X_1) = \mathrm{E}(X_2) = \mathrm{E}(X_3) = \mathrm{E}(Y) = 0$,

$$\mathrm{Var}(X) = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 1 \end{bmatrix}, \qquad \text{and} \qquad \mathrm{Cov}(X,Y) = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

Then, the regressors of the linear regression model

$$Y = \frac{2}{3}X_1 + \frac{2}{3}X_2 + \varepsilon$$

are exogenous. What happens if we exclude X_2? In that case, our basic set of regressors just reduces to $\{X_1\}$ and so our regression model is now

$$Y = X_1 + \varepsilon\,,$$

in which X_1 and ε are still uncorrelated. By contrast, if we include X_3, our basic set of regressors becomes $\{X_1, X_2, X_3\}$ and the regression model turns into

$$Y = \frac{1}{2}X_1 + \frac{1}{2}X_2 + \frac{1}{2}X_3 + \varepsilon\,.$$

Once again, all exogeneity conditions are satisfied and the regression coefficient of X_3 is, in fact, nonzero. However, none of these models need to be *valid*, since exogeneity is only a necessary but not a sufficient condition for validity—according to Figure 2 or, equivalently, Theorem 9.

Moreover, the choice of regressors is typically motivated by arguments that focus on optimality rather than validity (see, e.g., Shibata, 1981). This is usually associated with the common problem that the probability measure P is unknown in real life. Thus, if the family $\mathcal{F}$ of regression functions is parametric, we have to estimate the parameters of the (optimal) regression function. This creates estimation risk, which can lead to overfitting. This must be taken into account, too. However, optimality is not the same as validity. More precisely, prediction aims at minimizing the mean square prediction error, $\mathrm{E}(\varepsilon^2)$, whereas description means to minimize the mean square description error $\mathrm{E}\big((\varepsilon - \epsilon)^2\big)$. Hence, the selection criteria of prediction do not apply to description. Therefore, mixing up the main goals of regression analysis, i.e., prediction and description, and applying flawed arguments of variable selection can be highly misleading.

As already mentioned at the beginning of Section 2.2.2, the selection of variables should

depend on our principal goal:

- If we want to *predict* Y by X, the choice of X is rather arbitrary, since we only try to achieve a strong prediction power. Hence, $X_1, \ldots, X_m$ are selected for pure statistical reasons.
- By contrast, if we want to *describe* the impact of X on Y, the choice of X is *not* arbitrary. It is driven by theoretical considerations that go beyond statistics.

More precisely, prediction aims at maximizing R^2. This requires us to find some variables $X_1, \ldots, X_m$ with a strong explanation power, S^2, and also an optimal regression function $f \in \mathcal{F}$ in order to minimize the mean square prediction error. In general, this goal can be accomplished without exogeneity and it is not necessary at all that f is valid. By contrast, description aims at maximizing V^2. In that case, we search within $\mathcal{F}$ for the true regression function g, given the regressors $X_1, \ldots, X_m$, in order to minimize the mean square description error. Then, in fact we should guarantee that f satisfies the typical exogeneity conditions of linear regression. However, in general, exogeneity alone does not suffice—even not if we focus on linear regression analysis, i.e., search for g within the set $\mathcal{L}$ of linear regression functions. The problem is that $\mathcal{L}$ need not be adequate, in which case each $f \in \mathcal{F}$ is invalid. Thus, I recommend to apply a genuine validity test, i.e., a test for the null hypothesis that $f = g$. Such a test is developed in the next section.

3. The Validity Test

3.1. Test Statistic

Let $f \in \mathcal{F}$ be a regression function, $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a sample of $n \geq 1$ (not necessarily independent) observations of (X, Y), and $\varepsilon_1, \ldots, \varepsilon_n$ with $\varepsilon_i = Y_i - f(X_i)$ for $i = 1, \ldots, n$ be the associated sample errors. Consider any partition

$$\mathcal{P} = \Big\{ \{\varepsilon_{1,1}, \ldots, \varepsilon_{n_1,1}\}, \ldots, \{\varepsilon_{1,r}, \ldots, \varepsilon_{n_r,r}\} \Big\}$$

of $\{\varepsilon_1, \ldots, \varepsilon_n\}$ into $r \in \{1, \ldots, n\}$ subsamples with $n_1, \ldots, n_r \geq 1$ such that $\sum_{j=1}^{r} n_j = n$.

Henceforth, each statement about $\varepsilon_{i,j}$, i.e., the ith error within the jth subsample, is tacitly understood to hold true for $i = 1, \ldots, n_j$ and $j = 1, \ldots, r$. Similarly, each assertion about ε_i implicitly refers to $i = 1, \ldots, n$. In general, the distribution of $\varepsilon_{i,j}$ depends on the information that has been used to create the partition of errors. For example, assume that the errors $\varepsilon_1, \ldots, \varepsilon_n$ are sorted in ascending order, $\varepsilon_1^* \leq \ldots \leq \varepsilon_n^*$, and grouped into $r = n \geq 2$ singular subsamples $\{\varepsilon_1^*\}, \ldots, \{\varepsilon_n^*\}$. Then, we can expect that the mean of ε_1^* is less than the mean of ε_2^*, etc.

Suppose that $\mathcal{P}$ depends *only* on $\mathbf{X}$, i.e., on the sample observations $X_1, \ldots, X_n$ of X.[19] Then, conditional on $\mathbf{X}$, it makes no difference at all whether we consider ε_i before or after it has been assigned to any subsample. Splitting $\{\varepsilon_1, \ldots, \varepsilon_n\}$ into r subsamples by using only information that is contained in $\mathbf{X}$ does not even change the *joint* distribution of errors conditional on $\mathbf{X}$. Simply put, $\mathcal{P}$ has no influence on the (conditional) distribution of $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)$.

[19] This does not mean that $\mathcal{P}$ must depend on $\mathbf{X}$, but that it must not depend on any information that goes *beyond* $\mathbf{X}$.

If $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independent, then $(\varepsilon_1, X_1, Y_1), \ldots, (\varepsilon_n, X_n, Y_n)$ are independent, too, since ε_i is a function of (X_i, Y_i). In this case, $\varepsilon_1, \ldots, \varepsilon_n$ are even independent *conditionally* on $\mathbf{X}$.[20] Further, the (conditional) distribution of ε_i depends on $\mathbf{X}$ only through X_i, i.e., the sample observation of X that is associated with ε_i.

Let ε_k be the element of the entire sample $\{\varepsilon_1, \ldots, \varepsilon_n\}$ of errors that coincides with the element $\varepsilon_{i,j}$ of the jth subsample of errors. We conclude that

$$\mathrm{E}(\varepsilon_{i,j} \,|\, \mathbf{X}) = \mathrm{E}(\varepsilon_k \,|\, X_k).$$

Moreover, if the regression function f is valid, we have that $\mathrm{E}(\varepsilon_k \,|\, X_k) = 0$ and thus

$$\mathrm{E}(\varepsilon_{i,j} \,|\, \mathbf{X}) = 0.$$

This means that the (conditional) means of the errors within each subsample are zero, irrespective of how we create the subsamples according to the given realizations $x_1, \ldots, x_n$ of $X_1, \ldots, X_n$. This is the key observation for developing the validity test.

The following theorem, whose proof can be skipped, strengthens the aforementioned results.

Theorem 10 (Partition of errors)**.** *Let $f \in \mathcal{F}$ be some regression function, $(X_1, Y_1), (X_2, Y_2), \ldots$ be an infinite sample of independent observations of (X, Y), $1 \leq r \leq n$, and $\mathcal{X} := \{X_1, X_2, \ldots\}$. Assume that the partition $\mathcal{P}$ of $\{\varepsilon_1, \ldots, \varepsilon_n\}$ into r subsamples depends only on $X_1, \ldots, X_n$.*

(i) $\mathcal{P}$ has no influence on the joint distribution of $\varepsilon_1, \ldots, \varepsilon_n$ conditional on $\mathcal{X}$.

(ii) The errors $\varepsilon_1, \ldots, \varepsilon_n$ are independent conditionally on $\mathcal{X}$.

(iii) The distribution of ε_i given $\mathcal{X}$ depends only on X_i.

In particular, if the regression function f is valid, the mean of $\varepsilon_{i,j}$ given $\mathcal{X}$ is zero.

Consider any regression model $Y = f(X) + \varepsilon$ with $f \in \mathcal{F}$ and fix some number $r \in \mathbb{N}$. Let $(X_1, Y_1), (X_2, Y_2), \ldots$ be an infinite sample of independent observations of (X, Y) and $x_1, x_2, \ldots$ be the corresponding sample realizations of X. The following arguments implicitly refer to the *conditional* probability measure $\mathrm{P}(\cdot \,|\, X_1 = x_1, X_2 = x_2, \ldots)$. Put another way, the sample realizations $x_1, x_2, \ldots$ of X are considered *fixed*. Further, let

$$\mathcal{P}_n := \Big\{ \{\varepsilon_{1,1,n}, \ldots, \varepsilon_{n_{1,n},1,n}\}, \ldots, \{\varepsilon_{1,r,n}, \ldots, \varepsilon_{n_{r,n},r,n}\} \Big\}$$

with $n \geq r$ be some partition of errors based on $x_1, \ldots x_n$. This means that $\mathcal{P}_n$ depends on the entire sample size n and, especially, it is determined only by the first n sample realizations of X, where $\varepsilon_{i,j,n}$ denotes the ith error within the jth subsample, which possesses the size $n_{j,n}$. Hence, we are concerned with r triangular arrays $\{\varepsilon_{i,1,n}\}_{n=r,r+1,\ldots}^{i=1,\ldots,n_{1,n}}, \ldots, \{\varepsilon_{i,r,n}\}_{n=r,r+1,\ldots}^{i=1,\ldots,n_{r,n}}$ of errors.

[20] $\mathrm{P}(\boldsymbol{\varepsilon} \leq e \,|\, \mathbf{X}) = \prod_{i=1}^{n} \mathrm{P}(\varepsilon_i \leq e_i \,|\, X_i) = \prod_{i=1}^{n} \mathrm{P}(\varepsilon_i \leq e_i \,|\, \mathbf{X})$ for all $e = (e_1, \ldots, e_n) \in \mathbb{R}^n$.

For each $j \in \{1, \ldots, r\}$ and $n \geq r$, define

$$T_{j,n} := \frac{\frac{1}{\sqrt{n_{j,n}}} \sum_{i=1}^{n_{j,n}} \varepsilon_{i,j,n}}{\sqrt{\frac{1}{n_{j,n}} \sum_{i=1}^{n_{j,n}} \left(\varepsilon_{i,j,n} - \frac{1}{n_{j,n}} \sum_{i=1}^{n_{j,n}} \varepsilon_{i,j,n} \right)^2}},$$

provided that $\frac{1}{n_{j,n}} \sum_{i=1}^{n_{j,n}} \left(\varepsilon_{i,j,n} - \frac{1}{n_{j,n}} \sum_{i=1}^{n_{j,n}} \varepsilon_{i,j,n} \right)^2 > 0$, whereas

$$T_{j,n} := \begin{cases} 0, & n_{j,n} < 2 \quad \vee \quad \varepsilon_{1,j,n} = \ldots = \varepsilon_{n_{j,n},j,n} = 0 \\ -\infty, & n_{j,n} \geq 2 \quad \wedge \quad \varepsilon_{1,j,n} = \ldots = \varepsilon_{n_{j,n},j,n} < 0 \\ +\infty, & n_{j,n} \geq 2 \quad \wedge \quad \varepsilon_{1,j,n} = \ldots = \varepsilon_{n_{j,n},j,n} > 0 \end{cases},$$

given that $\frac{1}{n_{j,n}} \sum_{i=1}^{n_{j,n}} \left(\varepsilon_{i,j,n} - \frac{1}{n_{j,n}} \sum_{i=1}^{n_{j,n}} \varepsilon_{i,j,n} \right)^2 = 0$.

The principal assumption is that

$$\left(T_{1,n}, \ldots, T_{r,n} \right) \rightsquigarrow \mathcal{N}_r(0, \mathbf{I}_r)$$

if f is valid, which is motivated by Theorem 10. More precisely, if (i) the observations of (X, Y) are independent, (ii) the partition of errors depends only on the first n observations of X, and (iii) f is valid, then the (conditional) mean of $\varepsilon_{i,j,n}$ is zero for $i = 1, \ldots, n_{j,n}$, $j = 1, \ldots, r$, and $n = r, r+1, \ldots$. Moreover, since the errors are (conditionally) independent, the same holds true for $T_{1,n}, \ldots, T_{r,n}$ for each $n \geq r$. Finally, from the continuous mapping theorem we conclude that

$$T_n := \sum_{j=1}^{r} T_{j,n}^2 = \sum_{j=1}^{r} \frac{\left(\sum_{i=1}^{n_{j,n}} \varepsilon_{i,j,n} \right)^2}{\sum_{i=1}^{n_{j,n}} \left(\varepsilon_{i,j,n} - \frac{1}{n_{j,n}} \sum_{i=1}^{n_{j,n}} \varepsilon_{i,j,n} \right)^2} \rightsquigarrow \chi_r^2 \,.$$

However, it can happen that $n_{j,n} < 2$ or $\varepsilon_{1,j,n} = \ldots = \varepsilon_{n_{j,n},j,n} = 0$, in which case we set $T_{j,n}$ to zero and reduce the number of degrees of freedom of χ^2 by one. In fact, we can ignore the given subsample if its size is too small in order to provide any evidence against the validity of f. Similarly, if we observe that $\varepsilon_{1,j,n}, \ldots, \varepsilon_{n_{j,n},j,n} = 0$, there is absolutely no evidence at all against the validity of f, too. By contrast, if we observe that $\varepsilon_{1,j,n} = \ldots = \varepsilon_{n_{j,n},j,n} \neq 0$ (and $n_{j,n}$ is large), we have a clear evidence against the validity of f. Thus, we do *not* ignore this subsample.

Hence, let $0 \leq \delta \leq r$ be the number of subsamples with $n_{j,n} \geq 2$ and $\frac{1}{n_{j,n}} \sum_{i=1}^{n_{j,n}} \varepsilon_{i,j,n}^2 > 0$. If the entire sample size n is large, the test statistic T_n is approximately χ_δ^2-distributed, provided that the regression function f is valid. Thus, we should reject the null hypothesis that f is valid if and only if

$$p = \begin{cases} 1 - F_{\chi_\delta^2}(t_n), & \delta > 0 \\ 1, & \delta = 0 \end{cases}$$

falls below some low (nominal) level of significance, where $F_{\chi_\delta^2}$ is the cumulative distribution function of χ_δ^2 and t_n is the given realization of the test statistic T_n. Synonymously, we should

reject the *regression model* $Y = f(X) + \varepsilon$ if the p-value is sufficiently low.

Before going further, I would like to point out some practical aspects of the validity test. In most cases, we can guarantee that $n_{j,n} \geq 2$ and we also have that $\frac{1}{n_{j,n}} \sum_{i=1}^{n_{j,n}} \varepsilon_{i,j,n}^2 > 0$ for $j = 1, \ldots, r$. This means that δ coincides with r. Further, even if all errors $\varepsilon_{1,j,n}, \ldots, \varepsilon_{n_{j,n},j,n}$ are in fact zero, it can actually happen that $\frac{1}{n_{j,n}} \sum_{i=1}^{n_{j,n}} \varepsilon_{i,j,n}^2 > 0$, *numerically*. This can lead to erroneous results, e.g., when running a Monte Carlo simulation with $V^2 = 1$ and $S^2 = 1$, i.e., $Y = f(X)$. Thus, if the validity test is implemented by a number cruncher, I recommend to take only those subsamples into account for which $\frac{1}{n_{j,n}} \sum_{i=1}^{n_{j,n}} \varepsilon_{i,j,n}^2 \geq \upsilon > 0$, where υ is the machine precision.

3.2. Linear Regression Models

In the following, I will frequently refer to the true regression equation $Y = g(X) + \epsilon$, where g is the true regression function of Y given X and ϵ is the corresponding residual. Thus, it holds that $\mathrm{E}(\epsilon) = 0$ and I assume that ϵ is independent of X. The true regression equation can be considered the data-generating process of Y. By contrast, $Y = f(X) + \varepsilon$ always represents some regression model. More precisely, I focus on *linear* regression models that are specified such that the exogeneity conditions of linear regression are satisfied. Hence, $f(X) = \alpha + \beta' X$ is the (unique) optimal linear predictor of Y. Proposition 2 tells us that $f(X)$ is the best choice, among the set $\mathcal{L}(X)$ of linear predictors of Y based on X, also if we want to describe the impact of X on Y—irrespective of whether or not the family $\mathcal{L}$ of linear regression functions is adequate.

3.2.1. Simple Regression

Consider a simple linear regression model $Y = \alpha + \beta X + \varepsilon$, i.e., X is a random variable. Hence, the given regression function is $x \mapsto f(x) = \alpha + \beta x$. Further, let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a sample of n independent observations of (X, Y) and suppose that the subsample $\{\varepsilon_{1,1,n}, \ldots, \varepsilon_{n_{1,n},1,n}\}$ contains all errors that are associated with $f(X_i) \leq 0$, i.e., $\alpha + \beta X_i \leq 0$, whereas the errors in the subsample $\{\varepsilon_{1,2,n}, \ldots, \varepsilon_{n_{2,n},2,n}\}$ stem from $f(X_i) > 0$, i.e., $\alpha + \beta X_i > 0$, for $i = 1, \ldots, n$.

If the linear regression model is valid and the subsample sizes are large, we can expect that

$$T_n = \sum_{j=1}^{2} \frac{\left(\sum_{i=1}^{n_{j,n}} \varepsilon_{i,j,n}\right)^2}{\sum_{i=1}^{n_{j,n}} \left(\varepsilon_{i,j,n} - \frac{1}{n_{j,n}} \sum_{i=1}^{n_{j,n}} \varepsilon_{i,j,n}\right)^2}$$

is approximately χ_2^2-distributed, provided that $\delta = r = 2$. Hence, we should reject the linear regression model if $p = 1 - F_{\chi_2^2}(t_n)$ falls below some low level, where t_n is the realization of T_n.

To illustrate the practical importance of the validity test, let us consider the following example: Suppose that $Y = g(X) + \epsilon$ with

$$x \mapsto g(x) = \begin{cases} -c, & x \leq 0 \\ c, & x > 0 \end{cases} \tag{7}$$

and $c \geq 0$, where $X, \epsilon \in \mathcal{N}(0, 1)$ are independent.

True:	$Y = \begin{cases} -c + \epsilon, & X \leq 0 \\ c + \epsilon, & X > 0 \end{cases}$ with $c \geq 0$ and $X, \epsilon \in \mathcal{N}(0,1)$ being independent.
Model:	$Y = \alpha + \beta X + \varepsilon$ with $\alpha = 0$ and $\beta = 0.7979c$.
Impact:	$\frac{\partial}{\partial x} f(X) = 0.7979c$, $\frac{\partial}{\partial x} g(X) = 0$
Moments:	$\mathrm{E}(\varepsilon) = \mathrm{Cov}(X, \varepsilon) = 0$, $\mathrm{E}(\varepsilon \mid X) = \begin{cases} -c\big(0.7979X + 1\big), & X \leq 0 \\ -c\big(0.7979X - 1\big), & X > 0 \end{cases}$
Ratios:	$A^2 = \dfrac{0.6366c^2}{\big(1 + 0.3634c^2\big)\big(1 + c^2\big)}$, $R^2 = \dfrac{0.6366c^2}{1 + c^2}$, $S^2 = \dfrac{c^2}{1 + c^2}$, $V^2 = \dfrac{1}{1 + 0.3634c^2}$

Table 3: Fact sheet of the piecewise constant regression equation.

We can express Y, *equivalently*, by the linear regression model

$$Y = \alpha + \beta X + \varepsilon$$

with $\alpha = 0$ and $\beta = 2c\phi(0)$, where $\phi(0) = 0.3989$ represents the density of the standard normal distribution at 0. Hence, the error of the linear regression model is $\varepsilon = Y - 2c\phi(0)X$ and it is evident that $\mathrm{E}(Y) = 0$, i.e., $\mathrm{E}(\varepsilon) = 0$, too. Further, we have that

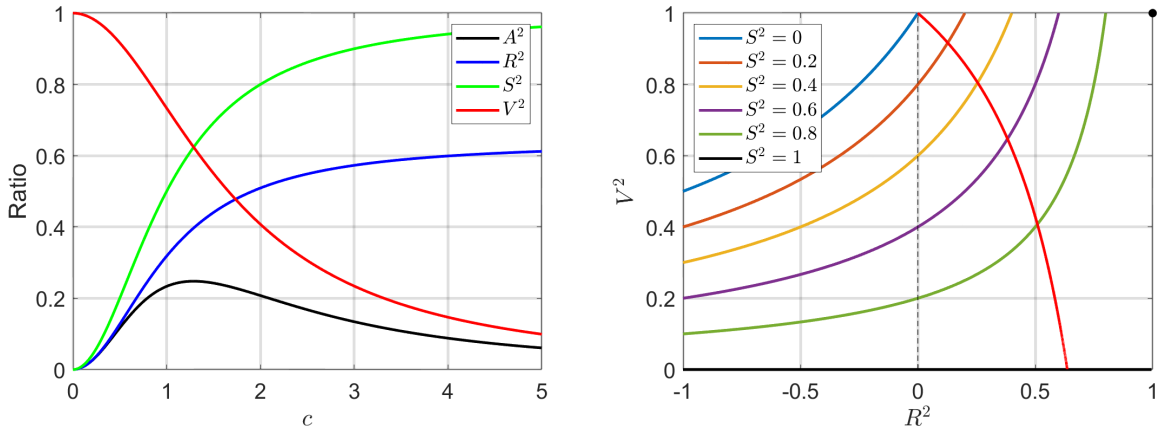
$$\mathrm{Cov}(X, \varepsilon) = \mathrm{E}(X\varepsilon) = \mathrm{E}(XY) - 2c\phi(0) = 0$$

because

$$\mathrm{E}(XY) = \int_{-\infty}^{\infty} x\mathrm{E}(Y \mid X = x)\phi(x)\,\mathrm{d}x = -c\int_{-\infty}^{0} x\phi(x)\,\mathrm{d}x + c\int_{0}^{\infty} x\phi(x)\,\mathrm{d}x = 2c\phi(0).$$

Hence, $Y = 0.7979cX + \varepsilon$ satisfies the typical exogeneity conditions of linear regression. That is, X is exogenous and $f(X) = 0.7979cX$ is the best linear predictor of Y based on X. However, we have that $g(X) = \mathrm{E}(Y \mid X) = \pm c$, depending on whether $X \leq 0$ or $X > 0$. Thus, for all $c > 0$, the conditional mean of Y essentially differs from $f(X)$ and so the linear regression model is clearly invalid. Actually, the true marginal impact of X on Y is $\frac{\partial}{\partial x} g(X) = 0$, almost surely, but the linear regression model suggests a marginal impact of $\frac{\partial}{\partial x} f(X) = 0.7979c$, which is positive if $c > 0$. Simply put, we have a *spurious regression*.

Table 3 summarizes the given example. It contains the true regression equation ("True"), the linear regression model ("Model"), the suggested and the true marginal impact of X on Y ("Impact"), the unconditional moments and the mean conditional error ("Moments"), and



(i) Regression ratios

(ii) Prediction power vs. validity

Figure 3: Piecewise constant regression equation.

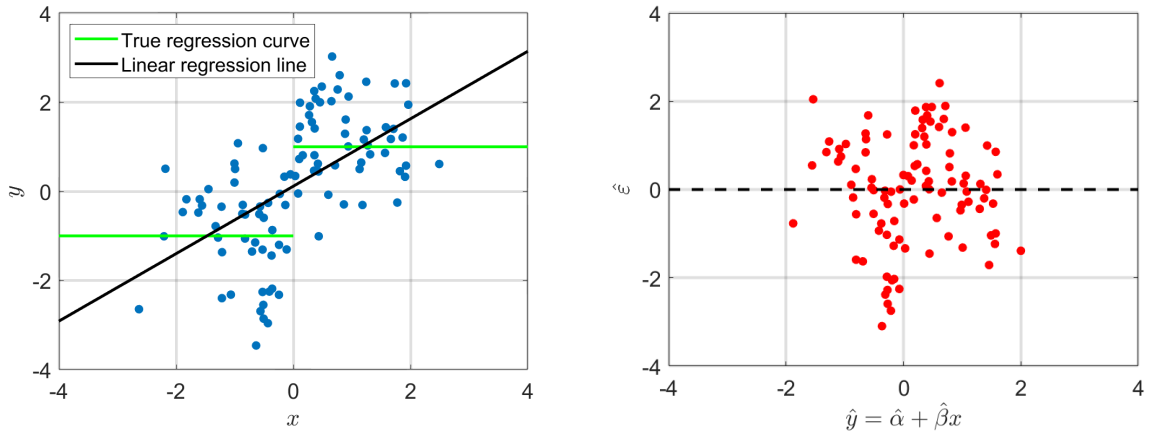
the corresponding regression ratios ("Ratios"). Figure 3 (i) clarifies how the regression ratios depend on the parameter c. For example, let us assume that $c = 1$, which leads us to the validity $V^2 = 0.7335$. Hence, although the linear regression model is clearly invalid, the coefficient of determination amounts to $R^2 = 0.3183$, which is quite high. In fact, as is shown by Figure 3 (i), the lower V^2, the higher R^2. More precisely, the linear regression model is valid if and only if it does not fit at all ($c = 0$), and the better it fits ($c \to \infty$), the more it is invalid. This adverse effect can be seen also in Figure 3 (ii), which clarifies how V^2 and R^2 are connected through S^2. Hence, the stronger the explanation power of X, i.e., S^2, the more invalid the linear regression model.

This is a prime example of why we should not rely on R^2 in order to verify the validity of any regression model. The regressor X is exogenous for all $c \geq 0$. Hence, endogeneity is no problem at all in this context. Furthermore, the regression coefficient $\beta = 0.7979c$ is positive for all $c > 0$. Finally, the coefficient of determination approaches 0.6366 as c tends to infinity. Thus, even the prediction power can be very strong. For these reasons, the linear regression model appears to be well-specified and the usual validity checks would never reveal that it is actually invalid.

How to apply the validity test in this situation? In the first step, we can estimate the regression parameters α and β by ordinary least squares (OLS). Thus, let $\hat{\alpha}$ and $\hat{\beta}$ be the corresponding OLS estimators, which lead us to the sample predictions $\hat{Y}_1, \ldots, \hat{Y}_n$ with $\hat{Y}_i := \hat{\alpha} + \hat{\beta} X_i$ for $i = 1, \ldots, n$. Further, let $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n$ with $\hat{\varepsilon}_i := Y_i - \hat{Y}_i$ for $i = 1, \ldots, n$ be the corresponding prediction errors. Now, we split the entire sample $\{\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n\}$ of errors into two subsamples, where the error $\hat{\varepsilon}_i$ is assigned to the first subsample if $\hat{Y}_i \leq 0$, whereas it is put into the second subsample if $\hat{Y}_i > 0$ for $i = 1, \ldots, n$. Thus, our test statistic is

$$T_n = \sum_{j=1}^{2} \frac{\left(\sum_{i=1}^{n_{j,n}} \hat{\varepsilon}_{i,j,n}\right)^2}{\sum_{i=1}^{n_{j,n}} \left(\hat{\varepsilon}_{i,j,n} - \frac{1}{n_{j,n}} \sum_{i=1}^{n_{j,n}} \hat{\varepsilon}_{i,j,n}\right)^2},$$

where we can assume that $n_{1,n}, n_{2,n} \geq 2$, $\frac{1}{n_{1,n}} \sum_{i=1}^{n_{1,n}} \hat{\varepsilon}_{i,1,n}^2 > 0$, and $\frac{1}{n_{2,n}} \sum_{i=1}^{n_{2,n}} \hat{\varepsilon}_{i,2,n}^2 > 0$. Hence, the



(i) Scatter plot (ii) OLS predictions vs. prediction errors

Figure 4: 100 realizations of X vs. $Y = \pm 1 + \epsilon$.

number of degrees of freedom of T_n is $\delta = r = 2$.[21]

Figure 4 (i) contains $n = 100$ simulated observations of X and $Y = \pm 1 + \epsilon$, i.e., the parameter c equals 1. The OLS estimates of $\alpha = 0$ and $\beta = 0.7979$ are $\hat{\alpha} = 0.1148$ and $\hat{\beta} = 0.7571$, respectively. The corresponding regression line can be found in Figure 4 (i), too. Further, the ordinary R^2 based on the OLS estimates amounts to 0.3202, and the corresponding p-value of the F-test for $H_0 : \beta = 0$ is virtually zero. Figure 4 (ii) is a residual plot. It contains the OLS predictions of Y together with the associated prediction errors, which look fine, too.

All in all, Figure 4 suggests that the linear regression model is appropriate, but we know that the opposite is true. In fact, the true regression curve, which is depicted on the left-hand side of Figure 4, is piecewise constant and it jumps up at $x = 0$. Thus, it is far away from being linear. With the best will in the world, this cannot be seen just by a visual inspection of the original or the fitted data. What does the validity test tell us in this situation? The value of the test statistic is 10.2570, i.e., its p-value amounts to 0.0059. Thus, after applying the validity test, the linear regression model can be rejected on every meaningful significance level.

Here, we have assumed that the regression model is linear just for the sake of simplicity but without loss of generality. Indeed, we could have used any other parametric regression function $f(\cdot\,, \theta)$ with $\theta \in \Theta \subseteq \mathbb{R}^q$, instead, and estimate the parameter vector θ in some appropriate way. In the light of our previous findings, a natural estimator is given by

$$\hat{\theta} \in \underset{\theta \in \Theta}{\arg\min}\, \widehat{\mathrm{E}}\big((Y - f(X, \theta))^2\big),$$

where $\widehat{\mathrm{E}}$ denotes the empirical mean. Then, given that the conditions mentioned in Section 2.2.1

[21] We can always guarantee that $n_{1,n}, n_{2,n} \geq 2$ by choosing some appropriate threshold for the sample predictions, without destroying the fundamental result expressed by Theorem 10, given that our choice is based only on $\mathbf{X}$ and $n \geq 4$. This holds true even if $\hat{\beta} = 0$, but then we cannot provide any meaningful threshold for the sample predictions.

are satisfied, we obtain the estimating equation

$$\widehat{\mathrm{E}}\left(\frac{\partial}{\partial\theta}f(X,\hat{\theta})\,\hat{\varepsilon}\right)=0,$$

which leads us to a generalized method-of-moments estimator. Of course, any other estimation procedure can be applied as well, but at least we should guarantee that the resulting estimator is consistent for the parameter vector θ^* from Equation 2, which minimizes the mean square error $\mathrm{E}(\varepsilon^2)$. Nonetheless, the validity test applies to *any* choice of $\theta\in\Theta$. That is, we need not limit ourselves to an *optimal* regression function in order to test for validity.

Hence, suppose that $f(\cdot\,,\theta)\in\mathcal{F}$ is some parametric regression function, where the (unknown) parameter vector θ is specified in some arbitrary way, and let the sample size be $n\geq q$. Further, let $\hat{\theta}\in\Theta$ be some consistent estimator for θ and $\hat{Y}_1=f(X_1,\hat{\theta}),\ldots,\hat{Y}_n=f(X_n,\hat{\theta})$ be the sample predictions of Y with their associated prediction errors $\hat{\varepsilon}_1,\ldots,\hat{\varepsilon}_n$. In the previous example, the threshold for each sample prediction, which has been used in order to create the partition of $\{\hat{\varepsilon}_1,\ldots,\hat{\varepsilon}_n\}$, was deliberately set to 0. However, in most applications it is not clear how to set the threshold, and usually we prefer to split the entire sample into more than two subsamples.

In order to solve this problem, we can simply rearrange the sample predictions $\hat{Y}_1,\ldots,\hat{Y}_n$ in ascending order, which leads us to the sorted predictions $\hat{Y}_1^*\leq\ldots\leq\hat{Y}_n^*$. Next, we split the set $\{\hat{Y}_1^*,\ldots,\hat{Y}_n^*\}$ into $r\in\{1,\ldots,n\}$ subsets $P_{1,n},\ldots,P_{r,n}$ of the form

$$P_{j,n}:=\left\{\hat{Y}^*_{\left\lceil\frac{(j-1)n}{r}\right\rceil+1},\ldots,\hat{Y}^*_{\left\lceil\frac{jn}{r}\right\rceil}\right\}$$

for $j=1,\ldots,r$. Since r does not exceed n, the subsets $P_{1,n},\ldots,P_{r,n}$ are pairwise disjoint and nonempty. Hence, each subset $P_{j,n}$ contains $n_{j,n}\geq 1$ predictions and each prediction leads us to the associated prediction error. In this way, we construct a partition $\mathcal{P}_n=\{P_{1,n},\ldots,P_{r,n}\}$ of $\{\hat{\varepsilon}_1,\ldots,\hat{\varepsilon}_n\}$, i.e., we create r subsamples of prediction errors. If the sample size n is large, each subsample contains $n_{j,n}\approx\frac{n}{r}$ errors, i.e., it shares approximately $\frac{1}{r}$ of the entire sample. Since $\hat{\theta}$ is a consistent estimator for θ, $\mathcal{P}_n$ depends only on $X_1,\ldots,X_n$ as n tends to infinity. More precisely, the sample predictions $\hat{Y}_1,\ldots,\hat{Y}_n$ depend on $\hat{\theta}$, which in turn depends on the whole sample $(X_1,X_1),\ldots,(X_n,Y_n)$. Nonetheless, if we assume that $f(X,\cdot)$ is continuous at θ, almost surely, we can expect that $\mathcal{P}_n$ is asymptotically equivalent to the partition of errors that we would obtain, in the same way, if the parameter vector θ were known.

Now, I would like to explain why the partition of errors should be created in that particular way, irrespective of whether or not the regression function f is parametric. If f is invalid, we want to find some partition of errors that is based only on $\mathbf{X}$ (and on $\hat{\theta}$, given that f is parametric), such that the errors are either positive or negative in most subsamples. In this case, we can expect that T_n is high enough in order to reject the (invalid) regression model $Y=f(X)+\varepsilon$. By contrast, if the errors take both positive values and negative values in most subsamples, they cancel out each other. Then, T_n might be too low to reject the given regression model. The principal idea is that the sign of $\mathrm{E}(\varepsilon\,|\,X=x)$ depends on x mainly through the prediction $f(x)$.

True:	$Y = a + bX + c(X^2 - 1) + \epsilon$ with $a, b, c \in \mathbb{R}$ and $X, \epsilon \sim \mathcal{N}(0,1)$ being independent.
Model:	$Y = \alpha + \beta X + \varepsilon$ with $\alpha = a$ and $\beta = b$.
Impact:	$\frac{\partial}{\partial x} f(X) = b$, $\frac{\partial}{\partial x} g(X) = b + 2cX$
Moments:	$\mathrm{E}(\varepsilon) = \mathrm{Cov}(X, \varepsilon) = 0$, $\mathrm{E}(\varepsilon \mid X) = c(X^2 - 1)$
Ratios:	$A^2 = \dfrac{b^2}{(2c^2+1)(b^2+2c^2+1)}$, $R^2 = \dfrac{b^2}{b^2+2c^2+1}$, $S^2 = \dfrac{b^2+2c^2}{b^2+2c^2+1}$, $V^2 = \dfrac{1}{2c^2+1}$

Table 4: Fact sheet of the quadratic regression equation.

To be more precise, consider some prediction $f(x_0)$ with $x_0 \in D$. If $\mathrm{E}(\varepsilon \mid X = x_0) \gtrless 0$, it should hold that $\mathrm{E}(\varepsilon \mid X = x) \gtrless 0$ for most other $x \in D$ such that $f(x) \approx f(x_0)$. Thus, all errors that are associated with a similar prediction should be grouped together when creating the subsamples.

Our basic assumption about the mean conditional error can be violated in some applications. Put another way, it can happen that the sign of $\mathrm{E}(\varepsilon \mid X = x)$ depends on x *not* mainly through $f(x)$.[22] In this case, we could apply any other function $h \colon D \to \mathbb{R}$ instead of f to the sample realizations of X. However, this could require more effort both from a conceptual and from a numerical point of view. Further, the main problem is that, in real life, we do not know the true regression function and thus we are not able to calculate the mean conditional error. Anyway, the method presented here is not wrong in any sense. At most, it might be inefficient with regard to the power of the validity test. The particular charm of creating the subsamples of errors by using the sample predictions of Y is its simple applicability to multiple regression analysis.

The true regression function, g, of the dependent variable $Y = \pm c + \epsilon$ is piecewise constant. Now, consider another example in which g is quadratic. More precisely, suppose that

$$Y = a + bX + c(X^2 - 1) + \epsilon \tag{8}$$

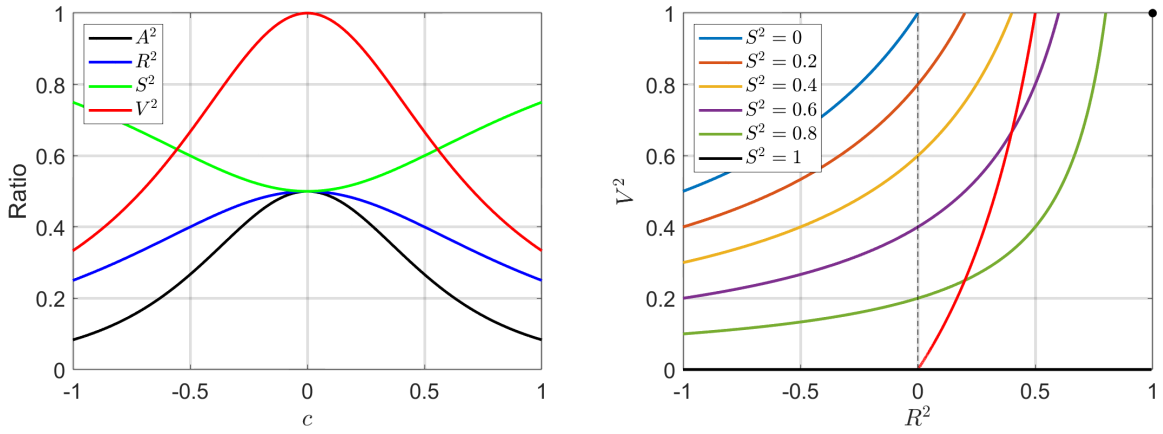
with $a, b, c \in \mathbb{R}$ and $X, \epsilon \sim \mathcal{N}(0,1)$ being independent. Further, let the (simple) linear regression model be

$$Y = \alpha + \beta X + \varepsilon \tag{9}$$

with $\alpha = a$ and $\beta = b$. Hence, we obtain the regression error $\varepsilon = c(X^2 - 1) + \epsilon$ with

$$\mathrm{E}(\varepsilon) = c\mathrm{E}\big(X^2 - 1\big) + \mathrm{E}(\epsilon) = 0$$

[22] In particular, this holds true if some component of x has no impact on $f(x)$ but on the sign of $\mathrm{E}(\varepsilon \mid X = x)$.



(i) Regression ratios (ii) Prediction power vs. validity

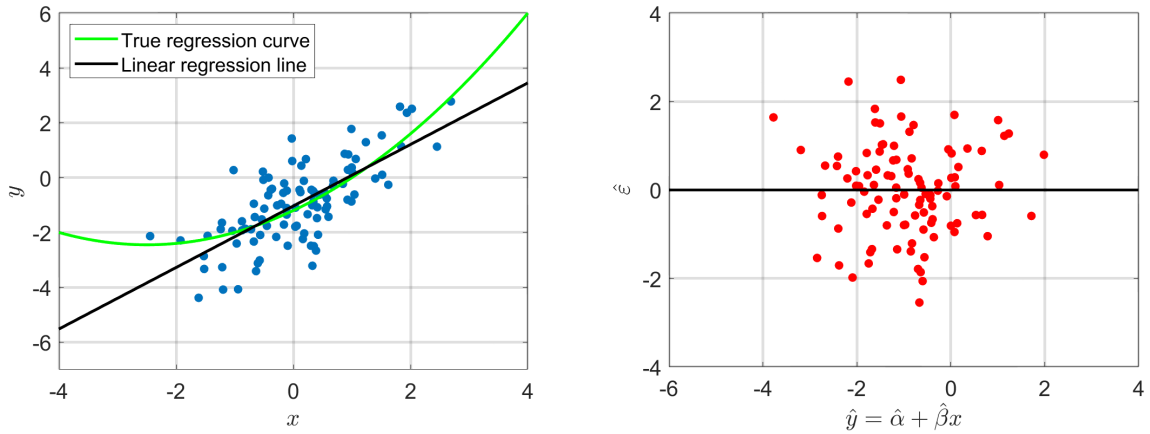
Figure 5: Ratios for the quadratic regression equation with $b = 1$.

and

$$\mathrm{Cov}(X, \varepsilon) = c\mathrm{Cov}\big(X, X^2\big) + \mathrm{Cov}(X, \epsilon) = 0\,.$$

This means that the typical exogeneity conditions of linear regression are satisfied. Actually, it does not matter how we choose the parameters a and b of the true regression equation (8). It always turns out that $\widehat{Y} = \alpha + \beta X$ is the best linear predictor of Y based on X, provided that $\alpha = a$ and $\beta = b$. Moreover, $\hat{Y}$ has some prediction power for all $\beta \neq 0$ and X is always exogenous (even if $\beta = 0$). Further, the mean conditional error is $\mathrm{E}(\varepsilon \,|\, X) = c(X^2 - 1)$. Hence, the linear regression model given by Equation 9 is invalid if and only if $c \neq 0$. In this case, the conditional mean of Y is a quadratic function of X and the (true) marginal impact of X on Y is $\frac{\partial}{\partial x} g(X) = b + 2cX$. Thus, it depends on X itself, which is completely overlooked if we use a linear regression model. Table 4 summarizes the given example. Once again, this fact sheet also contains the ratios A^2, R^2, S^2, and V^2, i.e., the accuracy, the coefficient of determination, the explanation power, and the validity. Figure 5 (i) shows how the regression ratios depend on the parameter c if $b = 1$ and the red line in Figure 5 (ii) clarifies how the validity and the prediction power of the linear regression model are connected with one another for different values of S^2.

Figure 6 (i) contains a scatter plot based on 100 independent copies of X and $Y = -1 + X + 0.2(X^2 - 1) + \epsilon$, which have been obtained by Monte Carlo simulation. Hence, the regression coefficients are given by $\alpha = a = -1$ and $\beta = b = 1$, whereas the hidden parameter c of the true regression function amounts to 0.2. The coefficient of determination is $R^2 = 0.4808$. Thus, we conclude that the prediction power is fairly strong. Further, the validity is $V^2 = 0.9259$, which indicates that the regression model is slightly invalid. The OLS estimates of α and β are $\hat{\alpha} = -1.0345$ and $\hat{\beta} = 1.1221$, respectively, which have been used to create the regression line in Figure 6 (i). The ordinary R^2 based on the OLS estimates amounts to 0.5195. Hence, the fit is very good, compared with values that can usually be observed in real life. The F-test for the null hypothesis that $\beta = 0$ leads us to a p-value of virtually zero. Finally, the residual plot can be found in Figure 5 (ii), where the residuals are based on the given OLS estimates of α and β.



(i) Scatter plot (ii) OLS predictions vs. prediction errors

Figure 6: 100 realizations of X vs. $Y = -1 + X + 0.2(X^2 - 1) + \epsilon$.

Both the numerical and the graphical results appear good. I think that nobody of us would recognize that the given regression model is invalid just by applying the usual validity checks. The graph of the true regression function $x \mapsto g(x) = -1 + x + 0.2(x^2 - 1)$ can be found in Figure 6 (i). Thus, we have that $\frac{\partial}{\partial x} g(x) = 1 + 0.4x$. Hence, especially for higher absolute values of x, the impact of X on Y is severely misunderstood when using the linear regression model. We conclude that the linear regression model serves well in order to *predict* Y, but it cannot *describe* the impact of X on Y, appropriately. In fact, the crux of the matter is that we ignore the regressor X^2 in Equation 9, but there is *no* omitted-variable bias at all, since X is exogenous.

How does the validity test perform in this situation? For applying this test, we may simply use the OLS estimates of α and β. For example, by creating $r = 5$ subsamples, the validity test leads us to $t_n = 12.8996$, i.e., to a p-value of 0.0243. Hence, the linear regression model can be rejected, at least, on a significance level of 5%. This holds true although the validity, V^2, is quite high in this case, i.e., the linear regression model is not so far away from being valid.

3.2.2. Multiple Regression

Now, consider another example, namely the Cobb-Douglas production function

$$(x_1, x_2) \mapsto \pi(x_1, x_2) = b_0 x_1^{b_1} x_2^{b_2}$$

with $b_0, x_1, x_2 > 0$ and $b_1, b_2 \in \mathbb{R}$. Here, $\pi(x_1, x_2)$ quantifies the total production, i.e., the output, of some economy, given the capital input x_1 and the labor input x_2. Further, b_0 is some scale parameter. Thus, we can express the Cobb-Douglas production function, equivalently, by

$$(\log x_1, \log x_2) \mapsto \log \pi(x_1, x_2) = \log b_0 + b_1 \log x_1 + b_2 \log x_2 \,.$$

True:	$Y = a + b_1 K + b_2 L + cKL + \epsilon$ with $\begin{bmatrix} K \\ L \\ \epsilon \end{bmatrix} \sim \mathcal{N}_3 \left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho & 0 \\ \rho & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right)$, where $a, b_1, b_2, c \in \mathbb{R}$ and $-1 < \rho < 1$.
Model:	$Y = \alpha + \beta_1 K + \beta_2 L + \varepsilon$ with $\alpha = a + c\rho$, $\beta_1 = b_1$, and $\beta_2 = b_2$.
Impact:	$\frac{\partial}{\partial(k,l)} f(K,L) = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$, $\frac{\partial}{\partial(k,l)} g(K,L) = \begin{bmatrix} b_1 + cL \\ b_2 + cK \end{bmatrix}$
Moments:	$\mathrm{E}(\varepsilon) = \mathrm{Cov}(K,\varepsilon) = \mathrm{Cov}(L,\varepsilon) = 0$, $\mathrm{E}(\varepsilon \mid K, L) = c(KL - \rho)$
Ratios:	$A^2 = \frac{b_1^2 + b_2^2 + 2b_1 b_2 \rho}{\left[c^2(1+\rho^2)+1\right]\left[b_1^2 + b_2^2 + 2b_1 b_2 \rho + c^2(1+\rho^2) + 1\right]}$
	$R^2 = \frac{b_1^2 + b_2^2 + 2b_1 b_2 \rho}{b_1^2 + b_2^2 + 2b_1 b_2 \rho + c^2(1+\rho^2) + 1}$,
	$S^2 = \frac{b_1^2 + b_2^2 + 2b_1 b_2 \rho + c^2(1+\rho^2)}{b_1^2 + b_2^2 + 2b_1 b_2 \rho + c^2(1+\rho^2) + 1}$, $V^2 = \frac{1}{c^2(1+\rho^2)+1}$

Table 5: Fact sheet of the Cobb-Douglas-like regression equation.

However, in real life, we cannot expect that the given quantities are related to one another in that *precise* manner.[23] Moreover, both the capital input and the labor input can be considered stochastic, which means that the output of the economy is stochastic, too.

Thus, let Y, K, and L be the (natural) logarithms of the output, the capital input, and the labor input, respectively, of the given economy. Suppose that

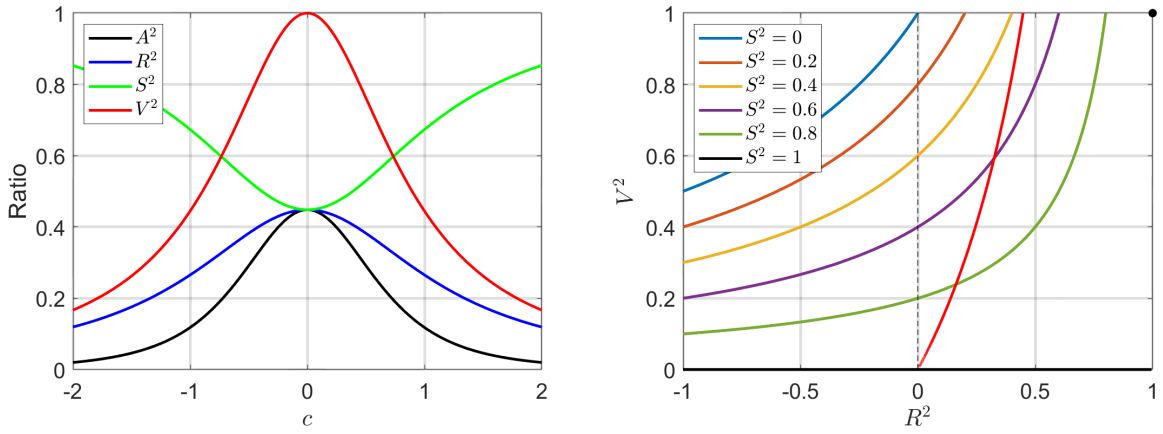
$$Y = a + b_1 K + b_2 L + cKL + \epsilon \tag{10}$$

with $a = \log b_0$ and

$$\begin{bmatrix} K \\ L \\ \epsilon \end{bmatrix} \sim \mathcal{N}_3 \left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho & 0 \\ \rho & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right),$$

where $c \in \mathbb{R}$ and $-1 < \rho < 1$. Hence, log-capital input and log-labor input are correlated if $\rho \neq 0$. Furthermore, their impact on the log-output of the economy is not linear if $c \neq 0$. In that case, there is a synergy of capital and labor, which is quantified by the parameter c.

[23] Here, I decidedly refrain from discussing whether or not that function is appropriate at all to describe the total production of an economy. Here, it just serves as a standard example of a multiple regression function in econometrics.



(i) Scatter plot (ii) OLS predictions vs. prediction errors

Figure 7: Ratios for the Cobb-Douglas regression with $b_1 = 0.25$, $b_2 = 0.75$, and $\rho = 0.5$.

Consider the (multiple) linear regression model

$$Y = \alpha + \beta_1 K + \beta_2 L + \varepsilon\,. \tag{11}$$

The parameters $\alpha = a + c\rho$, $\beta_1 = b_1$, and $\beta_2 = b_2$ lead us to $\varepsilon = -c\rho + cKL + \epsilon$, where the regression error satisfies the typical exogeneity conditions of linear regression, i.e.,

1. $\mathrm{E}(\varepsilon) = -c\rho + c\mathrm{E}(KL) + \mathrm{E}(\epsilon) = -c\rho + c\rho = 0$,
2. $\mathrm{Cov}(K, \varepsilon) = \mathrm{E}(K\varepsilon) = c\mathrm{E}(K^2 L) + \mathrm{E}(K\epsilon) = 0$, and
3. $\mathrm{Cov}(L, \varepsilon) = \mathrm{E}(L\varepsilon) = c\mathrm{E}(KL^2) + \mathrm{E}(L\epsilon) = 0$.

This holds true irrespective of how we choose a, b_1, b_2, c, and ρ, i.e., the parameters of the true regression equation (10). Hence, the regressors K and L are always exogenous. Put another way, there is no omitted-variable bias—although we ignore KL in our linear regression model. Further, the regression parameters β_1 and β_2 of the linear regression model (11), in fact, coincide with the regression parameters b_1 and b_2, respectively, of the true (but nonlinear) regression equation. All this said, the linear regression model is still invalid if $c \neq 0$. Table 5 contains the fact sheet of the Cobb-Douglas-like regression equation.[24]

The mean conditional error amounts to $\mathrm{E}(\varepsilon \,|\, K, L) = c(KL - \rho)$. Thus, suppose that there is a synergy of capital and labor, i.e., $c > 0$. Then, the log-output of the economy is systematically overestimated by the linear regression model if $KL < \rho$ and it is systematically underestimated if $KL > \rho$. Further, Table 5 reveals that the marginal impact of log-capital and of log-labor on the log-output of the economy is always underestimated by an amount of cL and cK, respectively.

For example, suppose that $b_1 = 0.25$, $b_2 = 0.75$, and $\rho = 0.5$. Hence, log-capital and log-labor are positively correlated, where labor has a stronger impact on output than capital. Figure 7 contains the corresponding regression ratios. As it can be seen, even for higher absolute values of

[24]The regression ratios can be calculated by applying Isserlis' theorem.

c, the prediction power of the linear regression model, i.e., R^2, can still be satisfactory, although the model becomes highly invalid. Once again, this underpins our insights of Section 2.3, where it has been shown that R^2 shall not be used as a validity measure.

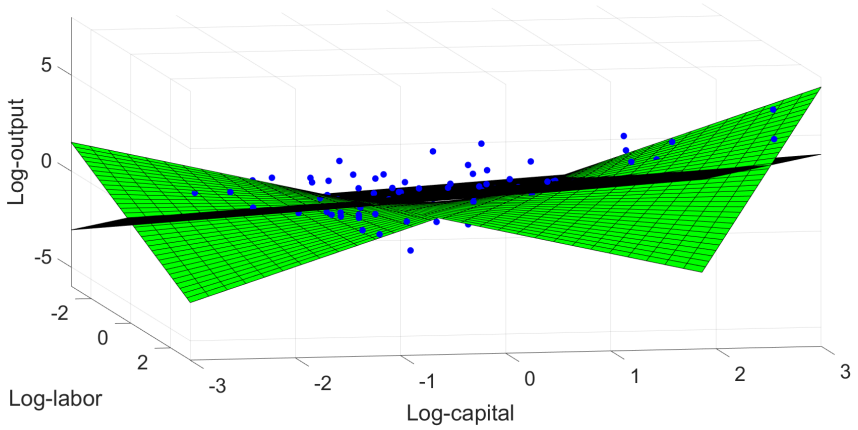


Figure 8: Scatter plot for the Cobb-Douglas regression, where the plane represents the fitted linear regression function and the bent surface illustrates the true regression function.

Consider a Monte Carlo simulation of $n = 100$ independent observations of (K, L, Y), given that $a = 0$, $b_1 = 0.25$, $b_2 = 0.75$, $\rho = 0.5$, and $c = 0.5$, in which case the linear regression model is invalid. The resulting OLS estimates of $\alpha = 0.25$, $\beta_1 = 0.25$, and $\beta_2 = 0.75$ are $\hat{\alpha} = 0.4661$, $\hat{\beta}_1 = 0.3155$, and $\hat{\beta}_2 = 0.8615$, respectively. Further, the coefficient of determination is 0.3824, whereas the ordinary R^2, obtained by the OLS estimates, even amounts to 0.4939. Finally, the F-test leads us to a p-value of virtually zero. The scatter plot in Figure 8 contains the realized data points in $\mathbb{R}^3$. One can see that the graph of the linear regression function, i.e., the plane, which is based on the given OLS estimates, fits good to the data.

Figure 9 contains the corresponding residual plot. The linear predictions of the log-output (based on the OLS estimates) can be found on the x-axis and the associated prediction errors are

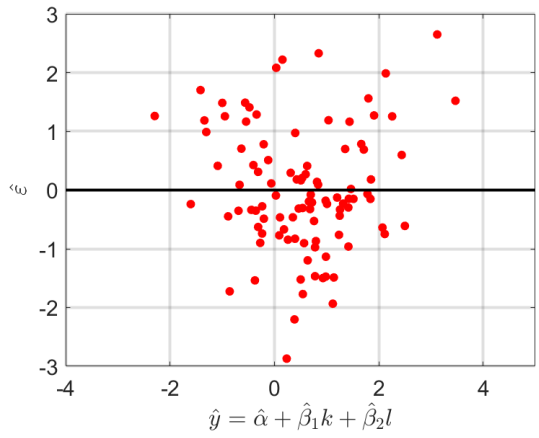


Figure 9: Residual plot for the Cobb-Douglas regression.

given on the y-axis.[25] Obviously, the quantitative results look fine and also a visual inspection of the data does not reveal any anomaly. Nonetheless, we already know that the linear regression model is invalid, since the relationship between log-output, log-capital, and log-labor is nonlinear. This is illustrated by the bent surface in Figure 8, which represents the true Cobb-Douglas-like regression function $(k, l) \mapsto 0.25k + 0.75l + 0.5kl$. Now, how does the validity test perform in that situation, where we apply a *multiple* regression? With $r = 5$ subsamples, it comes to $t_n = 19.6724$, which corresponds to a p-value of 0.0014. Thus, again we can reject the linear regression model on every meaningful significance level.

3.3. Size and Power

Here, I present the size and power of the validity test for the three examples discussed in the previous sections, i.e.,

1. the piecewise constant regression equation,
2. the quadratic regression equation, and
3. the Cobb-Douglas-like regression equation.

The given results are obtained by Monte Carlo simulation, where each setting consists of the following attributes:

- The example $e \in \{1, 2, 3\}$, containing the true regression equation and the corresponding linear regression model,
- the parameter $c \in C_e$ of the true regression equation, where the parameter set C_e depends on the given example e, viz.,
  - $C_1 = \{0, 0.25, 0.5, 0.75, 1\}$,
  - $C_2 = \{-0.4, -0.2, 0, 0.2, 0.4\}$, and
  - $C_3 = \{-0.5, -0.25, 0, 0.25, 0.5\}$,
- the sample size $n \in \{100, 500, 1000, 5000, 10000\}$, and
- the number $r \in \{5, 10\}$ of subsamples of the validity test.

The number of Monte Carlo repetitions in every setting amounts to 10000 and each repetition $w \in \{1, \ldots, 10000\}$ creates n independent observations of (X, Y). The given regression model is rejected if the p-value in Repetition w falls below the (nominal) level of 5%.

Table 6 contains the results of the simulation study, where the true regression equations can be found on the upper left of each panel. The linear regression models are valid if and only if $c = 0$. Hence, this table contains the size, i.e., the rejection rate if $c = 0$, and the power, i.e., the rejection rate in the case of $c \neq 0$, for each combination of c, n, and r. On the left-hand side

[25] An obvious advantage of the residual plot is that it can be applied for an arbitrary number of regressors.

				n									
c	R^2	S^2	V^2	100	500	1000	5000	10000	100	500	1000	5000	10000
$Y = \pm c + \epsilon$						$r = 5$					$r = 10$		
0	0	0	1.0000	0.0309	0.0138	0.0141	0.0105	0.0120	0.1389	0.0343	0.0265	0.0195	0.0188
0.25	0.0374	0.0588	0.9778	0.0775	0.2169	0.5009	0.9996	1.0000	0.2231	0.4330	0.8190	1.0000	1.0000
0.50	0.1273	0.2000	0.9167	0.2562	0.8818	0.9981	1.0000	1.0000	0.5103	0.9918	1.0000	1.0000	1.0000
0.75	0.2292	0.3600	0.8303	0.5638	0.9990	1.0000	1.0000	1.0000	0.8329	1.0000	1.0000	1.0000	1.0000
1	0.3183	0.5000	0.7335	0.8289	1.0000	1.0000	1.0000	1.0000	0.9714	1.0000	1.0000	1.0000	1.0000
$Y = -1 + X + c(X^2 - 1) + \epsilon$						$r = 5$					$r = 10$		
−0.4	0.4310	0.5690	0.7576	0.7400	1.0000	1.0000	1.0000	1.0000	0.8825	1.0000	1.0000	1.0000	1.0000
−0.2	0.4808	0.5192	0.9259	0.2256	0.8831	0.9985	1.0000	1.0000	0.4229	0.9465	0.9996	1.0000	1.0000
0	0.5000	0.5000	1.0000	0.0300	0.0133	0.0131	0.0134	0.0095	0.1344	0.0299	0.0239	0.0216	0.0209
0.2	0.4808	0.5192	0.9259	0.2225	0.8806	0.9983	1.0000	1.0000	0.4101	0.9455	1.0000	1.0000	1.0000
0.4	0.4310	0.5690	0.7576	0.7458	1.0000	1.0000	1.0000	1.0000	0.8869	1.0000	1.0000	1.0000	1.0000
$Y = 0.25K + 0.75L + cKL + \epsilon$						$r = 5$					$r = 10$		
−0.50	0.3824	0.5294	0.7619	0.5212	0.9994	1.0000	1.0000	1.0000	0.7053	0.9996	1.0000	1.0000	1.0000
−0.25	0.4298	0.4711	0.9275	0.1522	0.6978	0.9689	1.0000	1.0000	0.3298	0.8018	0.9911	1.0000	1.0000
0	0.4483	0.4483	1.0000	0.0309	0.0152	0.0117	0.0123	0.0122	0.1423	0.0318	0.0258	0.0204	0.0184
0.25	0.4298	0.4711	0.9275	0.1497	0.6951	0.9673	1.0000	1.0000	0.3259	0.8064	0.9919	1.0000	1.0000
0.50	0.3824	0.5294	0.7619	0.5263	0.9995	1.0000	1.0000	1.0000	0.6927	0.9995	1.0000	1.0000	1.0000

Table 6: Size ($c = 0$) and power ($c \neq 0$) of the validity test.

one can find also the prediction power, R^2, the explanation power, S^2, and the validity, V^2, for each parameter c that is taken into consideration. The accuracy, A^2, follows immediately by $S^2 + V^2 - 1$ (or V^2R^2) and so this measure is dispensed with in order not to overload the table.

The power of the validity test is very strong for sample sizes $n \geq 1000$ or if the parameter c exceeds some critical threshold. Even for values of V^2 above 90%, its power turns out to be satisfactory whenever $n \geq 500$. In the case of $Y = \pm c + \epsilon$, which is depicted in the first panel of Table 6, the coefficient of determination decreases with the validity of the linear regression model. This counterintuitive relationship between V^2 and R^2, which is also illustrated by Figure 3 (ii), has already been discussed in Section 3.2.1. However, the validity test is not affected by R^2. It reliably indicates the invalidity of the regression model even if R^2 is high.

If we choose $r = 10$ subsamples for the validity test, and the sample size is small, i.e., $n = 100$, the size of the test always exceeds its nominal level of 5%. Hence, one should split the sample of regression errors into 5 rather than 10 subsamples when applying the validity test to small samples. By contrast, for larger sample sizes, the size of the test is much smaller than 5%. This is because the unknown regression parameters are estimated by OLS. If we would use instead the true regression parameters, the rejection rates would, in fact, increase to 5% if $c = 0$. However, the overall power of the validity test is significantly stronger with 10 subsamples.

In many econometric applications, the sample size is quite small. More precisely, one often uses quarterly data for regression analysis, in which case the validity test might be questionable. The results of the Monte Carlo study provided in Table 6 are based on the assumption that the explanation power of X, i.e., S^2, is quite low. To be more precise, the variance of the error ϵ of the true regression equation $Y = g(X) + \epsilon$ equals 1, which is relatively high, compared with the variance of $g(X)$. Hence, it seems reasonable to ask whether or not the validity test works well

	τ										
c	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
$Y = \pm c + \epsilon$											
0	0	0.0854	0.0900	0.0889	0.0882	0.0946	0.0903	0.0904	0.0880	0.0916	0.0894
0.25	1.0000	0.9884	0.7001	0.4023	0.2667	0.2037	0.1705	0.1497	0.1341	0.1209	0.1177
0.50	1.0000	1.0000	0.9901	0.8817	0.6968	0.5342	0.4028	0.3316	0.2736	0.2336	0.2163
0.75	1.0000	1.0000	0.9999	0.9916	0.9287	0.8219	0.6964	0.5786	0.4846	0.4089	0.3464
1	1.0000	1.0000	1.0000	0.9997	0.9878	0.9540	0.8764	0.7882	0.6958	0.6033	0.5281
$Y = -1 + X + c(X^2 - 1) + \epsilon$											
−0.4	1.0000	1.0000	0.9995	0.9915	0.9489	0.8689	0.7770	0.6712	0.5948	0.5120	0.4385
−0.2	1.0000	0.9998	0.9533	0.7766	0.5795	0.4455	0.3442	0.2867	0.2391	0.2103	0.1843
0	0	0.0833	0.0914	0.0828	0.0896	0.0872	0.0876	0.0852	0.0949	0.0873	0.0865
0.2	1.0000	0.9996	0.9549	0.7784	0.5817	0.4446	0.3411	0.2874	0.2390	0.2105	0.1887
0.4	1.0000	1.0000	0.9998	0.9907	0.9516	0.8789	0.7732	0.6767	0.5913	0.5005	0.4415
$Y = 0.25K + 0.75L + cKL + \epsilon$											
−0.50	0.9926	0.9867	0.9505	0.8746	0.7649	0.6620	0.5563	0.4704	0.4030	0.3489	0.3096
−0.25	0.9973	0.9708	0.7961	0.5855	0.4303	0.3122	0.2586	0.2111	0.1858	0.1553	0.1457
0	0	0.0914	0.0854	0.0923	0.0864	0.0889	0.0886	0.0900	0.0857	0.0929	0.0906
0.25	0.9983	0.9652	0.8034	0.5850	0.4184	0.3122	0.2592	0.2201	0.1899	0.1664	0.1445
0.50	0.9917	0.9858	0.9517	0.8747	0.7725	0.6578	0.5650	0.4739	0.4075	0.3473	0.3046

Table 7: Rejection rates with $n = 40$ and $r = 5$.

in small samples, given that the variance of ϵ is less than 1.

For example, suppose that we have only 10 years of quarterly data, i.e., $n = 40$, and that the true regression equation is $Y = g(X) + \epsilon$ with $\text{Var}(\epsilon) = \tau^2$ for $0 \leq \tau \leq 1$. Further, let $r = 5$ be the number of subsamples. If $\tau = 0$, the explanation power amounts to 1—except for $Y = \pm c + \epsilon$ with $c = \tau = 0$, in which case S^2 is not defined at all because $\text{Var}(Y) = 0$. By contrast, in the case of $\tau = 1$ we are, basically, in the same situation as in Table 6, but now the sample size is much smaller. Table 7 contains the given results for each combination of c and τ.

The rejection rates are quite satisfactory even for relatively high levels of τ. Of course, if τ is close to 1 and c is close to 0, i.e., if the explanation power of X is low and the validity of the linear regression model is high, the rejection rates become low. However, they always exceed the nominal level of 5%. More importantly, the rejection rates exceed that level also if $c = 0$. We can see that the (real) level of significance is rather 10%. This is owed by the fact that the sample size, $n = 40$, is quite small and the number of subsamples, $r = 5$, is relatively high. Thus, each subsample of errors consists only of 8 observations. It is surprising enough that the validity test works at all under these unpleasant circumstances.

4. Other Specification Tests

The validity test presented here should be distinguished from other specification tests that can be found in the literature. There seems to be no common understanding about what should make up a well-specified regression model. In Section 2.1, I already mentioned that every regression *equation* $Y = f(X) + \varepsilon$ is satisfied just by the very definition of ε. By contrast, if we

make any assumption about the joint distribution of X and ε, the regression equation becomes a regression *model*, which can very well be violated. Thus, a regression model could be considered well-specified if and only if the given assumption about the distribution of (X, ε) is satisfied. For example, we could assume that $\mathrm{E}(\varepsilon \mid X) = 0$, i.e., that the regression model $Y = f(X) + \varepsilon$ is valid. In that case, the regression model would be well-specified if and only if it is valid.[26]

However, we can find many alternative concepts of model specification in the literature. It has already been mentioned in Section 2.8 that there is no single way to predict or explain Y by a regression equation or model. That is, if we have found a well-specified regression model $Y = f(X) + \varepsilon$, a model based on any other set of regressors can be well-specified, too.[27] Hence, it makes no sense to say that a regression model is well-specified only if we choose $X_1, \ldots, X_m$ as explanatory variables. Put another way, it is very well possible to choose any other basic set of regressors in order to construct a well-specified regression model and we can imagine that the number of well-specified models for some dependent variable Y can even be infinite.

Thus, let $\{X_1, \ldots, X_m\}$ be a given set of regressors. Let us call $Y = f(X) + \varepsilon$ well-specified if and only if the joint distribution of X and ε satisfies some specific condition A.[28] For example, A can be some regression property that is discussed in Section 2.6, e.g., that the regression model is valid or optimal, or that the regressors are exogenous, given the chosen regression function f.

Suppose that we want to test for the null hypothesis A, i.e., that the given regression model is well-specified. Then, we can apply any other test for the null hypothesis $B \supset A$. Here, $B \supset A$ means that B is implied by A. Put another way, B is a necessary condition for A. For example, according to Figure 2, we can test for validity by testing for optimality or for exogeneity, etc. However, we can expect that a genuine test for B does not perform as well as a genuine test for A. More precisely, if the regression model $Y = f(X) + \varepsilon$ satisfies B but not A, then the test that is constructed in order to test for B will not reject the null hypothesis A although it is false.

The reader might ask why we should apply a genuine test for B and not a genuine test for A if we are interested in testing A? There are many responses to this question. For example, it could be that we already have a test for B and thus decide to apply the same test for A just for practical reasons. Another possibility could be that constructing a genuine test for A is difficult, whereas testing for B is easy. Anyway, I think that everybody of us agree that—for pure statistical reasons—it is better to apply a genuine test for A and not for $B \supset A$ if we actually want to test for A. However, most specification tests that can be found in the literature actually suffer from that particular problem, i.e., they do not represent genuine tests for validity.

4.1. Linear Regression Tests

Consider any regression function $f \in \mathcal{F}$ and let $\hat{f} \in \mathcal{F}$ be optimal among $\mathcal{F}$. From Theorem 8 we conclude that $f(X) = \hat{f}(X)$ if f is valid. Thus, f cannot be valid if $f(X) \neq \hat{f}(X)$ and so it should be rejected. To sum up, indeed we can test for the optimality of f in order to test for its

[26] This precise notion of model specification is shared, e.g., by MacKinnon (1992).

[27] In particular, the trivial regression model $Y = Y$ is always well-specified but meaningless.

[28] More generally, we could also consider some regression model involving $\mathbf{X}$, $\mathbf{Y}$, and $\boldsymbol{\varepsilon}$.

validity, since optimality is a necessary condition for validity. However, a test for optimality is not a genuine test for validity. For example, suppose that $\mathrm{Var}(X) > 0$ and let $Y = a + b'X + \epsilon$ be a linear regression model where a and b are such that the typical exogeneity conditions are satisfied. Since the covariance matrix of X is positive definite, the regression parameters are uniquely determined by $b = \mathrm{Var}(X)^{-1}\mathrm{Cov}(X, Y)$ and $a = \mathrm{E}(Y) - b'\mathrm{E}(X)$. Further, due to Theorem 7, $\hat{f}(X) = a + b'X$ is the unique optimal predictor of Y among the set $\mathcal{L}(X)$ of linear predictors based on X. Write $b = (b_1, b_2)$ and $\beta = (\beta_1, \beta_2)$, where the parameter vectors b_1 and β_1 refer to the same subvector X_1 of $X = (X_1, X_2)$.

Now, many specification tests are based on the hypotheses

H_0**:** $b_2 = \beta_2$ vs.

H_1**:** $b_2 \neq \beta_2$.

Thus, if H_0 is false, $f(X) = \alpha + \beta'X$ cannot coincide with $\hat{f}(X) = a + b'X$.[29] This means that the linear predictor $\alpha + \beta'X$ cannot be optimal. Hence, the linear regression model $Y = \alpha + \beta'X + \varepsilon$ cannot be valid either and so it should be rejected.

A particular version of these kind of specification tests (see, e.g., Greene, 2012, p. 177) is given by

H_0**:** $b_2 = 0$ vs.

H_1**:** $b_2 \neq 0$.

Here, H_0 states that X_2 can be dispensed with in order to predict Y by X. If H_0 is false, we should reject the linear regression model $Y = \alpha + \beta_1'X_1 + \varepsilon$, since the predictor $f(X) = \alpha + \beta_1'X_1$ is suboptimal given the basic set $\{X_1, \ldots, X_m\}$ of regressors.[30]

All these specification tests do not represent genuine validity tests, since the regression model $Y = \alpha + \beta'X + \varepsilon$ can be invalid even if the null hypothesis H_0 is true. Actually, if we cannot reject H_0, we may only accept the hypothesis that $\alpha + \beta'X$ is the (unique) optimal predictor of Y among all linear predictors based on X. Otherwise, we may reject also the null hypothesis that $Y = \alpha + \beta'X + \varepsilon$ is valid, but the power of such a validity test might be very poor. Another problem is that these specification tests are typically restricted to linear regression models.

4.2. Artificial Regression Tests

Now, assume that $X \neq 0$ is a random variable and consider a (simple) linear regression model $Y = \alpha + \beta X + \varepsilon$ with $\alpha, \beta \in \mathbb{R}$. Further, suppose that the parameter $\hat{\gamma}$ of the quadratic regression model $Y = \alpha + \beta X + \hat{\gamma}X^2 + \epsilon$ minimizes the mean square error $\mathrm{E}(\epsilon^2)$, where the parameters α and β stem from the linear regression model. The expanded model is said to be an artificial

[29] Since the covariance matrix of X is positive definite, $\alpha + \beta'X$ must differ from $a + b'X$ if H_1 is true.

[30] Nonetheless, the same predictor can still be optimal if our basic set of regressors contains only the regressors in X_1.

regression (MacKinnon, 1992). Hence, the family of regression functions that is taken into consideration is

$$\mathcal{F} = \left\{ x \mapsto f(x) = \alpha + \beta x + \gamma x^2 \colon \gamma \in \mathbb{R} \right\},$$

where $\hat{f} \in \mathcal{F}$ with $x \mapsto \hat{f}(x) = \alpha + \beta x + \hat{\gamma} x^2$ is optimal among $\mathcal{F}$. Thus, we should reject the linear regression model $Y = \alpha + \beta X + \varepsilon$ if $\hat{\gamma} \neq 0$ because then $\hat{\gamma} X^2 \neq 0$, i.e., $f(X) = \alpha + \beta X \neq \alpha + \beta X + \hat{\gamma} X^2 = \hat{f}(X)$. Similarly, Ramsey's (1969) RESET tests whether the prediction power of $\alpha + \beta X$ (with $\beta \neq 0$) can be increased by adding regressors of the form $(\alpha + \beta X)^k$ with $k > 1$. However, these kind of specification tests are designed to test for the *optimality* of $Y = \alpha + \beta X + \varepsilon$ but not for its validity. Another problem is that the power of tests based on artificial regressions essentially depend on the expansion of $Y = \alpha + \beta X + \varepsilon$, i.e., on the artificial regression equation.

The validity test developed here goes into another direction. In order to test for validity, we need not find any alternative model whose predictor has a stronger prediction power than the predictor of the original model. We have seen that the specification tests discussed above just aim at rejecting the hypothesis that $f(X)$ is *optimal* among $\mathcal{F}(X)$, but the problem is that optimality is weaker than validity if $\mathcal{F} \subset \mathcal{G}$. Hence, if $\mathcal{F}$ is not rich enough, specification tests based on the prediction power might be less powerful than genuine validity tests. Nonetheless, a fair comparison is nearly impossible because the power of specification tests that are based on the prediction power essentially depends on the alternative model that is taken into consideration. More precisely, the more we can reduce the mean square error of an invalid (linear) regression model by applying some alternative (nonlinear) regression model, the more powerful is the test for optimality. See, e.g., Greene (2012, Section 5.9) for a further discussion of that topic.

4.3. The Durbin-Wu-Hausman Test

A well-known further specification test is developed by Hausman (1978). This test presumes that we propose a parametric regression function $f(\cdot\,, \theta)$ where the parameter vector θ is not explicitly specified. Instead, it is assumed that we have some estimator $\hat{\theta}_0$ for θ that is asymptotically efficient and thus consistent under the null hypothesis H_0 that $f(\cdot\,, \theta)$ is valid, whereas $\hat{\theta}_0$ is inconsistent if $f(\cdot\,, \theta)$ is invalid. Further, there shall be another estimator $\hat{\theta}_1$ for θ that is consistent both under H_0 and under any specific alternative hypothesis H_1, for which reason it is asymptotically inefficient under H_0. Thus, if H_0 is true, we have that

$$n(\hat{\theta}_1 - \hat{\theta}_0)' V^{-1} (\hat{\theta}_1 - \hat{\theta}_0) \rightsquigarrow \chi^2_q$$

under the usual conditions of asymptotic theory, where V is the asymptotic covariance matrix of $\sqrt{n}(\hat{\theta}_1 - \hat{\theta}_0)$ under H_0 and q is the number of parameters.[31] By contrast, if H_1 is true, the test statistic should exceed a critical value, given that the sample size is large enough.

There is a close relationship between the test proposed by Hausman and other specification tests already developed by Durbin (1954) and Wu (1973), for which reason the presented test

[31] Here, it is implicitly assumed that V has full rank. Otherwise, we have to choose the Moore-Penrose inverse of V, in which case the number of degrees of freedom of χ^2 reduces to $\operatorname{rk} V$.

is called Durbin–Wu–Hausman (DWH) test. Obviously, the DWH test requires us to specify some parametric family $\mathcal{F}$ of regression functions. Further, there is a general shortcoming due to the very construction of the test statistic: The DWH test is just designed to detect significant deviations of $\hat{\theta}_1$ from $\hat{\theta}_0$, but such deviations need not occur at all if $f(\cdot\,,\theta)$ is invalid, i.e., if H_0 is violated. Hence, the DWH test is not a genuine test for validity. This shall be demonstrated by its most well-known implementation, namely a test for exogeneity.

Thus, consider a linear regression model $Y = \alpha + \beta' X + \varepsilon$ with $X = (X_1, \ldots, X_m)$ and

$$\beta = \mathrm{Cov}(Z, X)^{-1}\mathrm{Cov}(Z, Y),$$

where $Z = (Z_1, \ldots, Z_m)$ is any vector of instrumental variables.[32] It is implicitly assumed that the covariance matrices $\mathrm{Cov}(Z, X)$ and $\mathrm{Var}(X)$ are regular. This implies that $\mathrm{Cov}(Z, \varepsilon) = 0$, i.e., the instrumental variables are always exogenous. Further, consider the vector

$$\beta_0 = \mathrm{Var}(X)^{-1}\mathrm{Cov}(X, Y).$$

Now, the hypotheses are given by

$H_0\colon\ \beta = \beta_0$ vs.

$H_1\colon\ \beta \neq \beta_0$.

The regressors $X_1, \ldots, X_m$ are exogenous under H_0, whereas the exogeneity condition is violated under H_1. Hence, if the null hypothesis is true, the OLS estimator $\hat{\beta}_0$ is consistent, since it estimates $\beta_0 = \beta$. Further, if we assume that the random vector (X, Y, Z) is normally distributed, $\hat{\beta}_0$ is even asymptotically efficient. In any case, the OLS estimator becomes inconsistent under H_1. By contrast, the instrumental-variables (IV) estimator $\hat{\beta}_1$ is consistent both under H_0 and under H_1 because it always estimates β. However, $\hat{\beta}_1$ is asymptotically inefficient under H_0.

To sum up, all prerequisites required by Hausman (1978) are satisfied and so we can test for the null hypothesis that the linear regression model is well-specified—in the sense that $\beta = \beta_0$, i.e., that the components of X are *exogenous*. According to Theorem 9, exogeneity is equivalent to optimality if we focus on linear regression. Nonetheless, as we have already seen above, a linear regression model can even be highly invalid although the chosen regressors are exogenous, i.e., the given regression model is optimal. Hence, the DWH test is not a genuine test for validity.

4.4. The Harvey-Collier Test

Another well-known specification test is the ψ-test proposed by Harvey and Collier (1977). It is based on forecast errors. The authors presume that the GM is satisfied. Before applying the test, the (fixed) regressor matrix $\mathbf{x}$, together with the associated sample realizations $y_1, \ldots, y_n$ of Y, is arranged in ascending order, along some pre-specified row of $\mathbf{x}$. Hence, if $m > 1$, i.e., in the case of a multiple regression, one has to choose a leading regressor. Then, linear (ex-post) forecasts

[32] Some components of Z can be identical with the corresponding components of X, but it must hold that $Z \neq X$.

for $y_{m+2}, \ldots, y_n$ are calculated, recursively, by taking the first $m+1, \ldots, n-1$ observations. The corresponding $n-m-1$ forecast errors $u_{m+2}, \ldots, u_n$ are used to calculate the Harvey-Collier test statistic

$$\psi = \sqrt{\frac{n-m-2}{n-m-1}} \frac{\sum_{i=m+2}^{n} u_i}{\sqrt{\sum_{i=m+2}^{n} \left(u_i - \frac{1}{n-m-1} \sum_{i=m+2}^{n} u_i\right)^2}}.$$

According to Harvey and Collier (1977), it holds that $\psi \sim t_{n-m-2}$, given that the GM is satisfied.

The Harvey-Collier test differs in several aspects from the validity test presented here:

1. In contrast to the validity test, we have to specify some leading regressor in order to apply the Harvey-Collier test in the case of $m > 1$.

2. The Harvey-Collier test is designed to falsify the GM, which is even stronger than strict exogeneity and hardly satisfied in most applications of econometrics.

3. Its power is weak if the forecast errors are symmetrically distributed around 0, which can very well happen if the true regression function is neither convex nor concave.

Anyway, we may conclude that the Harvey-Collier test is not a validity test.

4.5. Utts' Rainbow Test

The rainbow test developed by Utts (1982) appears to be similar to the validity test developed here. Like for the Harvey-Collier test, its basic assumption is that the sample observations $Y_1, \ldots, Y_n$ of Y with $n > m+1$ obey the GM. It consists of two OLS regressions: The first one is made by using the entire sample and the second one is based on a subsample with size $n_S > m+1$, where the observations of X in that subsample are selected from some central region of X. The test statistic is

$$F = \frac{n_S - m - 1}{n - n_S} \left(\frac{\text{SSE}}{\text{SSE}_S} - 1 \right),$$

where SSE stands for the sum of squared errors of the entire sample and SSE_S denotes the sum of squared errors of the subsample. According to Utts (1982), it holds that $F \sim F^{n-n_S}_{n_S-m-1}$, provided that the GM is satisfied and the null hypothesis H_0 is true. Actually, despite of the similarities, the rainbow test has not much to do with the validity test presented here—except for the very fact that subsamples are created, too, in order to apply the test. On the one hand, this test is based on the very strong assumption that the GM is satisfied. On the other hand, it represents an analysis-of-variance test, which refers to (conditional) homoscedasticity rather than validity. However, a valid regression model need not possess a homoscedastic error, i.e., it is not required that $\text{Var}(\varepsilon \mid X) = \sigma^2 > 0$. We conclude that the rainbow test is a test for the null hypothesis that the GM is satisfied, but this is much stronger than validity. Hence, it is not a validity test, either.

5. Conclusion

Evaluating regression models by applying the usual validity checks of regression analysis can lead us to highly erroneous conclusions. Measures of prediction power, or of goodness of fit, are misleading when trying to describe the impact of some explanatory variable(s) on a dependent variable. Regression models with a strong prediction power can be highly inappropriate for the given purpose, even if they fit well to the data. Conversely, valid regression models may have a weak prediction power and they even need not fit at all. Also the typical exogeneity conditions of linear regression are far from sufficient to guarantee that the given regression model is valid.

Genuine tests for the validity of regression models can rarely be found in the literature and a visual inspection of the data often leads nowhere. The validity test developed here is simple and it can be applied to all kinds of regression models with an arbitrary number of regressors. It is very powerful in large samples and performs well also in small samples, given that the validity of the regression model is sufficiently low and that there is not too much noise in the true regression equation. Hence, the presented test pursues its mission and thus it should be applied whenever the main goal of regression is description rather than prediction.

Proofs

Proof of Proposition 2

We have that

$$\mathrm{E}\big((Y-f(X))^2\big) = \mathrm{E}\big((Y-g(X))^2\big) + 2\mathrm{E}\big((Y-g(X))(g(X)-f(X))\big) + \mathrm{E}\big((g(X)-f(X))^2\big)$$

with

$$\begin{aligned}\mathrm{E}\big((Y-g(X))(g(X)-f(X))\big) &= \mathrm{E}\left(\mathrm{E}\big((Y-g(X))(g(X)-f(X)) \mid X\big)\right)\\ &= \mathrm{E}\big((g(X)-g(X))(g(X)-f(X))\big)\\ &= \mathrm{E}\big(0(g(X)-f(X))\big) = \mathrm{E}(0) = 0,\end{aligned}$$

i.e.,

$$\mathrm{E}\big((Y-f(X))^2\big) = \mathrm{E}\big((Y-g(X))^2\big) + \mathrm{E}\big((g(X)-f(X))^2\big).$$

This is equivalent to

$$\mathrm{E}(\varepsilon^2) = \mathrm{E}(\epsilon^2) + \mathrm{E}\big((\varepsilon-\epsilon)^2\big)$$

for all $f \in \mathcal{G}$, which implies $\mathrm{E}(\epsilon^2) \leq \mathrm{E}(\varepsilon^2)$. Hence, $\mathrm{E}\big((Y-f(X))^2\big)$ is minimal if and only if $\mathrm{E}\big((g(X)-f(X))^2\big)$ is minimal.

Proof of Theorem 1

In the case of $R^2 < 1$, we obtain

$$\frac{1 - S^2}{1 - R^2} = \frac{\mathrm{E}(\epsilon)/\mathrm{Var}(Y)}{\mathrm{E}(\varepsilon)/\mathrm{Var}(Y)} = \frac{\mathrm{E}(\epsilon)}{\mathrm{E}(\varepsilon)} = V^2.$$

By contrast, $R^2 = 1$ implies that $\varepsilon = 0$ and thus $V^2 = 1$. Further, Proposition 2 leads us to $S^2 = 1$, too. The same proposition implies also that $R^2 \leq S^2$, where $R^2 = S^2$ if and only if $V^2 = 1$.

Proof of Theorem 2

In the case of $V^2 > 0$, it follows that

$$\frac{A^2}{V^2} = \frac{S^2 + V^2 - 1}{V^2} = 1 - \frac{1 - S^2}{V^2} = R^2$$

and thus $A^2 = V^2 R^2$. Otherwise, i.e., if $V^2 = 0$, we have that $S^2 = 1$ and so $A^2 = 1 + 0 - 1 = 0$, in which case the given formula is valid, too.

Proof of Theorem 3

(i) We have that $\mathrm{E}(\hat{\varepsilon}) = \mathrm{E}\big(\mathrm{E}(\hat{\varepsilon} \,|\, X)\big) = \mathrm{E}(0) = 0$.

(ii) By applying the variance decomposition theorem, we conclude that

$$\begin{aligned} \mathrm{Var}(\hat{\varepsilon}) &= \mathrm{E}\big(\mathrm{Var}(\hat{\varepsilon} \,|\, X)\big) + \mathrm{Var}\big(\mathrm{E}(\hat{\varepsilon} \,|\, X)\big) \\ &= \mathrm{E}\big(\mathrm{Var}(\hat{\varepsilon} \,|\, X)\big) + \mathrm{Var}(0) = \mathrm{E}\big(\mathrm{Var}(\hat{\varepsilon} \,|\, X)\big). \end{aligned}$$

(iii) From the law of total expectation and $\mathrm{E}(\hat{\varepsilon}) = 0$, we conclude that

$$\begin{aligned} \mathrm{Cov}(h(X), \hat{\varepsilon}) &= \mathrm{E}\big(h(X)\hat{\varepsilon}\big) = \mathrm{E}\big(\mathrm{E}(h(X)\hat{\varepsilon} \,|\, X)\big) \\ &= \mathrm{E}\big(h(X)\mathrm{E}(\hat{\varepsilon} \,|\, X)\big) = \mathrm{E}\big(h(X)0\big) = \mathrm{E}(0) = 0\,. \end{aligned}$$

(iv) Since $\hat{f}$ is valid, we have that $\hat{f} = g$. Thus, due to Proposition 2, $\hat{f}$ is optimal among $\mathcal{F}$.

(v) Let $\tilde{f} \in \mathcal{F}$ be optimal among $\mathcal{F}$, too, and $\tilde{\varepsilon} = Y - \tilde{f}(X)$ be the associated regression error. Then,

$$\begin{aligned} \mathrm{E}(\tilde{\varepsilon}^2) = \mathrm{E}\left((Y - \tilde{f}(X))^2\right) &= \mathrm{E}\left(\left[(Y - \hat{f}(X)) + (\hat{f}(X) - \tilde{f}(X))\right]^2\right) \\ &= \mathrm{E}\left(\hat{\varepsilon}^2\right) + 2\mathrm{E}\left((\hat{f}(X) - \tilde{f}(X))\hat{\varepsilon}\right) + \mathrm{E}\left((\hat{f}(X) - \tilde{f}(X))^2\right) \end{aligned}$$

with

$$\mathrm{E}\big((\hat{f}(X) - \tilde{f}(X))\hat{\varepsilon}\big) = \mathrm{E}\big(\hat{f}(X)\hat{\varepsilon}\big) - \mathrm{E}\big(\tilde{f}(X)\hat{\varepsilon}\big) = 0,$$

since both $\hat{f}(X)$ and $\tilde{f}(X)$ are square-integrable. Thus, we obtain

$$\mathrm{E}(\tilde{\varepsilon}^2) = \mathrm{E}\left(\hat{\varepsilon}^2\right) + \mathrm{E}\left(\left(\hat{f}(X) - \tilde{f}(X)\right)^2\right)$$

and because $\tilde{f}$ is optimal, too, it must hold that $\mathrm{E}(\tilde{\varepsilon}^2) = \mathrm{E}(\hat{\varepsilon}^2)$. We conclude that

$$\mathrm{E}\left(\left(\hat{f}(X) - \tilde{f}(X)\right)^2\right) = 0,$$

which means that $\tilde{f}(X) = \hat{f}(X) = g(X)$ and thus $\tilde{\varepsilon} = \hat{\varepsilon}$. Hence, also the regression model $Y = \tilde{f}(X) + \tilde{\varepsilon}$ is valid.

Proof of Theorem 4

(i) Since ε is independent of X, we have that $\mathrm{E}(\varepsilon \mid X) = \mathrm{E}(\varepsilon) = 0$.

(ii) The variance decomposition theorem tells us that

$$\mathrm{Var}(\varepsilon) = \mathrm{E}\big(\mathrm{Var}(\varepsilon \mid X)\big) + \mathrm{Var}\big(\mathrm{E}(\varepsilon \mid X)\big).$$

From $\mathrm{Var}(\varepsilon \mid X) = \mathrm{Var}(\varepsilon)$ it follows that $\mathrm{Var}\big(\mathrm{E}(\varepsilon \mid X)\big) = 0$ and thus $\mathrm{E}(\varepsilon \mid X) = \mathrm{E}(\varepsilon) = 0$.

(iii) Fix any sample observation (X_i, Y_i) and let $\varepsilon_i = Y_i - f(X_i)$ be the associated sample error. If $\mathbf{X}$ is strictly exogenous, we have that

$$\mathrm{E}(\varepsilon_i \mid X_i) = \mathrm{E}\big(\mathrm{E}(\varepsilon_i \mid \mathbf{X}) \mid X_i\big) = \mathrm{E}(0 \mid X_i) = 0$$

and from $(X_i, Y_i) \sim (X, Y)$ we conclude that $(\varepsilon_i, X_i) \sim (\varepsilon, X)$, i.e., $\mathrm{E}(\varepsilon \mid X) = 0$.

(iv) The GM implies that $\mathbf{X}$ is strictly exogenous, which means that f is valid.

Proof of Theorem 5

It is well-known that the random vector (X, Y) is elliptically distributed, too. Further, $\mathrm{Var}(Z) > 0$ implies that $\mathrm{Var}(X) > 0$, i.e., the exogeneity conditions given by System 4 are equivalent to $\beta = \mathrm{Var}(X)^{-1}\mathrm{Cov}(X, Y)$ and $\alpha = \mathrm{E}(Y) - \beta'\mathrm{E}(X)$. Now, from Corollary 5 in Cambanis et al. (1981), we conclude that $\mathrm{E}(Y \mid X) = \alpha + \beta' X$, which means that the linear regression model is valid. Conversely, if the linear regression model is valid, Theorem 3 (i) and Corollary 1 (ii) guarantee that the exogeneity conditions given by System 4 are satisfied.

Proof of Theorem 6

(i) This is an immediate consequence of Proposition 2.

(ii) From the variance decomposition theorem, we conclude that

$$\mathrm{Var}(\varepsilon) = \mathrm{E}\big(\mathrm{Var}(\varepsilon \mid X)\big) + \mathrm{Var}\big(\mathrm{E}(\varepsilon \mid X)\big).$$

Further, $\mathrm{Var}(\varepsilon) = \mathrm{E}(\mathrm{Var}(\varepsilon \mid X))$ implies that $\mathrm{Var}\big(\mathrm{E}(\varepsilon \mid X)\big) = 0$, i.e., $\mathrm{E}(\varepsilon \mid X) = \mathrm{E}(\varepsilon) = 0$, which means that the regression model is valid. Conversely, if the regression model is valid, we obtain $\mathrm{E}(\varepsilon) = 0$ by Theorem 3 (i) and $\mathrm{Var}(\varepsilon) = \mathrm{E}(\mathrm{Var}(\varepsilon \mid X))$ by Theorem 3 (ii).

(iii) We already know from Proposition 2 that

$$\mathrm{E}\big((Y - f(X))^2\big) = \mathrm{E}\big((Y - g(X))^2\big) + \mathrm{E}\big((g(X) - f(X))^2\big)$$

for all $f \in \mathcal{G}$. Thus, if $E(\varepsilon^2) = \mathrm{E}\big((Y - f(X))^2\big) = \mathrm{E}\big((Y - g(X))^2\big)$, we have that

$$\mathrm{E}\big((g(X) - f(X))^2\big) = 0$$

and so $f(X) = g(X)$. Hence, the regression model $Y = f(X) + \varepsilon$ is valid. Conversely, if it is valid, i.e., $f(X) = g(X)$, it follows that $\mathrm{E}(\varepsilon^2) = \mathrm{E}\big((Y - g(X))^2\big)$.

Proof of Proposition 3

We have that

$$\mathrm{E}(\hat{\varepsilon}^2) = \mathrm{E}(\tilde{\varepsilon}^2) + 2\mathrm{E}\big((\tilde{f}(X) - \hat{f}(X))\tilde{\varepsilon}\big) + \mathrm{E}\big((\tilde{f}(X) - \hat{f}(X))^2\big),$$

where $\hat{\varepsilon} = Y - \hat{f}(X)$, and if $\tilde{\varepsilon} = Y - \tilde{f}(X)$ is orthogonal to $\mathcal{F}(X)$, we obtain

$$\mathrm{E}\big((\tilde{f}(X) - \hat{f}(X))\tilde{\varepsilon}\big) = \mathrm{E}\big(\tilde{f}(X)\tilde{\varepsilon}\big) - \mathrm{E}\big(\hat{f}(X)\tilde{\varepsilon}\big) = 0 - 0 = 0\,.$$

Then, it holds that

$$\mathrm{E}(\hat{\varepsilon}^2) = \mathrm{E}(\tilde{\varepsilon}^2) + \mathrm{E}\big((\tilde{f}(X) - \hat{f}(X))^2\big).$$

Since the regression function $\hat{f}$ is optimal, we must have that $\mathrm{E}(\hat{\varepsilon}^2) \leq \mathrm{E}(\tilde{\varepsilon}^2)$, i.e.,

$$\mathrm{E}\big((\tilde{f}(X) - \hat{f}(X))^2\big) = 0\,.$$

However, this cannot be true because $\tilde{f} \neq \hat{f}$. Thus, $\tilde{\varepsilon}$ cannot be orthogonal to $\mathcal{F}(X)$.

Proof of Theorem 8

Suppose that $\mathcal{F}$ is adequate and let $\hat{f}$ be the valid element of $\mathcal{F}$. Due to Theorem 3 (iv,v), $\hat{f}$ is the unique optimal regression function among $\mathcal{F}$ and Corollary 1 (v) asserts that $\hat{\varepsilon} = Y - \hat{f}(X)$ is orthogonal to $\mathcal{F}(X)$. Now, consider any regression function $\tilde{f} \in \mathcal{F}$. From Proposition 3 we conclude that $\tilde{\varepsilon} = Y - \tilde{f}(X)$ is orthogonal to $\mathcal{F}(X)$ only if $\tilde{f} = \hat{f}$. This means that $\hat{f}$ is the unique element of $\mathcal{F}$ that produces an error being orthogonal to $\mathcal{F}(X)$.

References

Aigner, D., Amemiya, T., Poirier, D. (1976): "On the estimation of production frontiers: Maximum likelihood estimation of the parameters of a discontinuous density function," *International Economic Review* **17**, pp. 377–396.

Boyd, S., Vandenberghe, L. (2009): *Convex Optimization*, Cambridge University Press, 7th edition.

Cambanis, S., Huang, S., Simons, G. (1981): "On the theory of elliptically contoured distributions," *Journal of Multivariate Analysis* **11**, pp. 368–385.

Durbin, J. (1954): "Errors in variables," *Review of the International Statistical Institute* **22**, pp. 23–32.

Fomby, T., Hill, R., Johnson, S. (1984): *Advanced Econometric Methods*, Springer.

Greene, W. (2012): *Econometric Analysis*, Pearson, 7th edition.

Harvey, A., Collier, P. (1977): "Testing for functional misspecification in regression analysis," *Journal of Econometrics* **6**, pp. 103–119.

Hastie, T., Tibshirani, R., Friedman, J. (2009): *The Elements of Statistical Learning*, Springer, 2nd edition.

Hausman, J. (1978): "Specification tests in econometrics," *Econometrica* **46**, pp. 1251–1271.

Hayashi, F. (2000): *Econometrics*, Princeton University Press.

Kelker, D. (1970): "Distribution theory of spherical distributions and a location-scale parameter generalization," *Sankhya A* **32**, pp. 419–430.

Kneib, T., Silbersdorff, A., Säfken, B. (2023): "Rage against the mean – A review of distributional regression approaches," *Econometrics and Statistics* **26**, pp. 99–123.

Koenker, R. (2005): *Quantile Regression*, Cambridge University Press.

Koenker, R., Basset, G. (1978): "Regression quantiles," *Econometrica* **46**, pp. 33–50.

MacKinnon, J. (1992): "Model specification tests and artificial regressions," *Journal of Economic Literature* **30**, pp. 102–146.

Newey, W., Powell, J. (1987): "Asymmetric least squares estimation and testing," *Econometrica* **55**, pp. 819–847.

Ramsey, J. (1969): "Tests for specification errors in classical linear least squares regression analysis," *Journal of the Royal Statistical Society, Series B* **31**, pp. 350–371.

Schulze Waltrup, L., Sobotka, F., Kneib, T., Kauermann, G. (2015): "Expectile and quantile regression—David and Goliath?" *Statistical Modelling* **15**, pp. 433–456.

Shibata, R. (1981): "An optimal selection of regression variables," *Biometrika* **68**, pp. 45–54.

Utts, J. (1982): "The rainbow test for lack of fit in regression," *Communications in Statistics: Theory and Methods* **11**, pp. 2801–2815.

Wu, D. (1973): "Alternative tests of independence between stochastic regressors and disturbances," *Econometrica* **41**, pp. 733–750.