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

September 25, 2024

Working Paper AP 2023–01
September 25, 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

September 25, 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 types of regression models with any number of regressors. It is very powerful in large samples and performs very 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.

* 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 of classical econometrics. Among many other scientific areas, it is frequently applied in social and natural sciences, too. We can achieve two goals with regression analysis, namely prediction and description.[1] 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.

Many, if not 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 their regression model $Y = f(X) + \varepsilon$ by demonstrating its capability to *predict* Y. Thus, it seems that description is often confused 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. In that case, it is 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, then 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.

Here, it is shown that even an *optimal* predictor $f(X)$ of Y need not constitute a valid regression model $Y = f(X) + \varepsilon$, whereas the opposite is true. To be more precise, 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 seems to be often misunderstood in practice, and common specification tests do not test for validity at all. Regression analysis plays a significant role in empirical research, and validity is a conditio sine qua non in most applications. Hence, 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, genuine validity tests or measures can rarely be found in the literature. Here, I do not mean the usual goodness-of-fit tests like the ordinary R^2 or Theil's adjusted R^2, hypothesis tests concerning linear- or artificial-regression parameters, or information criteria like the Akaike or the Bayesian information criterion, etc.[2] Actually, such tests and measures do not address the question of validity. They focus on prediction rather than

[1] Some authors distinguish between prediction and control (see, e.g., Fomby et al., 1984, p. 400).

[2] For a discussion of such measures see, e.g., Amemiya (1980) and Desboulets (2018).

description. Further, many tests are based on the classical assumptions of the Gaussian linear 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 validity, but these procedures are insufficient, too. This will be demonstrated later in this work.

The main purpose of this work is (i) to draw attention to the fundamental problem of validity and (ii) to present a genuine validity test. The test is simple and thus easy to implement. More importantly, it can be applied to all types of regression models with any number of regressors. The validity test 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, we should consider another regression model or, at least, change or transform the variables. However, this holds true only if we want to *describe* the impact of X on Y by some regression model $Y = f(X) + \varepsilon$. By contrast, if we wish to *predict* Y by X, the regression model may very well be invalid, and we should focus instead on the prediction power of $f(X)$.

2. Theoretical Background

2.1. Prerequisites

The elements of Euclidean space are considered column vectors. Random variables and random vectors are always denoted by capital Roman letters. The same is true for subsets of Euclidean space and real-valued matrices. A small Roman letter indicates a real number, an Euclidean vector, or a function. For example, $X = (X_1, \ldots, X_m)$ is an m-dimensional random vector, whereas $x = (x_1, \ldots, x_m) \in \mathbb{R}^m$ represents some realization of X. By contrast, $\{X_1, \ldots, X_m\}$ denotes a set of m random variables or random vectors $X_1, \ldots, X_m$. A small Greek letter can either be a random variable, a real number, an Euclidean vector, or a function. An equality or inequality between two (random) vectors means that the assertion holds true componentwise (and almost surely). Let X be an m-dimensional and Y be an n-dimensional random vector. Then, $\mathrm{Cov}(X, Y)$ denotes the $(m \times n)$ matrix of covariances between X and Y, whereas $\mathrm{Var}(X)$ is the $(m \times m)$ covariance matrix of X. Further, $\mathrm{Var}(X) > 0$ means that the covariance matrix of X is positive definite. The transpose of X is denoted by X', 0 symbolizes the zero scalar or a vector of zeros, depending on the given context. The same is 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," $:=$ means "is defined as," $\mathbf{I}_d$ is the $d \times d$ identity matrix, and $\mathbf{1}$ is the indicator function, i.e., $\mathbf{1}_A = 1$ if A is true and $\mathbf{1}_A = 0$ if A is false. The d-dimensional normal distribution with mean vector μ and covariance matrix Σ is symbolized by $\mathcal{N}_d(\mu, \Sigma)$, where the index d is dropped and Σ is substituted with σ^2 in the case of $d = 1$. Further, 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 represents the F-distribution with ν numerator and δ denominator degrees of freedom. Strong convergence, i.e., convergence with probability 1, is simply denoted

by $\to$, whereas weak convergence, i.e., convergence in distribution, is indicated by $\rightsquigarrow$. The additional remark "$n \to \infty$," i.e., that the sample size tends to infinity, is omitted 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 a probability space, where $\mathcal{A}$ is a σ-algebra on Ω, and let L^2 be the Hilbert space of all square-integrable random variables, equipped with the inner product $\mathrm{E}(XY)$ for all $X, Y \in L^2$. 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 tacitly assumed that the covariance matrix of (X, Y) is positive definite unless there is something to the contrary.[3] Further, 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 $f(X) \in L^2$. Two regression functions $\hat{f}$ and $\tilde{f}$ are considered identical if and only if $\hat{f}(X) = \tilde{f}(X)$.

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 response or dependent variable.[4] The given set of regressors, i.e., $\{X_1, \ldots, X_m\}$, is denoted by $\mathcal{S}$. Moreover, the regressors $X_1, \ldots, X_m$ and the dependent variable Y are considered *fixed*. This implies that they do not depend on the choice of the regression function f.[5]

The regression error is defined by $\varepsilon := Y - f(X)$. Hence, ε represents a *residual*, i.e., it is not considered fixed, and thus it is not treated like a regressor. However, for notational convenience, I usually refrain from using some index in order to clarify that ε depends on Y, X, and f. Equation 1 is always satisfied just by the very definition of ε, irrespective of how we choose the dependent variable Y, the regressors $X_1, \ldots, X_m$, and the regression function f. Put another way, if we do not make any assumption about the joint distribution of X and ε, Equation 1 is purely tautological. Consequently, it does not represent a *model*—apart from the very fact that the surrounding framework is just a probabilistic model of reality.

[3] This implies that $\mathrm{Var}(X) > 0$ and $\mathrm{Var}(Y) > 0$.

[4] 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.

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

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 $\mathcal{G}$. 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 $\mathcal{F}$ is a parametric family of regression functions. This means that $\mathcal{F} = \{f(\cdot\,,\theta) \colon \theta \in \Theta \subseteq \mathbb{R}^q\}$, 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. 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 $\theta = (\mu,\eta)$, 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 $\frac{\partial}{\partial\mu} f(X,\theta) = 1$ and thus Equation 2 leads us to the two (necessary and sufficient) conditions $\mathrm{E}(\varepsilon^*) = 0$ and $\mathrm{Cov}\big(\frac{\partial}{\partial\eta} f(X,\theta^*), \varepsilon^*\big) = 0$.[6] 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 exogeneity conditions $\mathrm{E}(\varepsilon^*) = 0$ and $\mathrm{Cov}(X,\varepsilon^*) = 0$ of linear regression.[7] 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.

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

[7] 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.

2.2.2. Description

If our aim is to *predict* some variable Y, we need no structural assumption. Then, the goal is to find any 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 $\mathrm{E}(\varepsilon^2)$. By contrast, if we want to *describe* the impact of X on Y, the situation is completely different. In this case, the regressors are chosen 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.

More precisely, let $x \mapsto g(x) = \mathrm{E}(Y \mid X = x)$ be a real-valued function on D that quantifies the conditional mean of Y given $X = x$. Henceforth, the function g is referred to as the true regression function of Y given X.[8] Correspondingly, $Y = g(X) + \epsilon$ is said to be the true regression equation, while ϵ is called the true regression error or true residual. Now, the regression function $f \in \mathcal{F}$ is said to be *valid* if and only if $f(X) = \mathrm{E}(Y \mid X)$, which is equivalent to

$$\mathrm{E}(\varepsilon \mid X) = 0.$$

Hence, f is valid if and only if $f = g$.

Furthermore, the family $\mathcal{F}$ of regression functions 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

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

Validity is a substantial model assumption.[9] 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.[10]

For example, a linear regression model presumes that

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

[8] We have $\mathrm{E}(Y^2) < \infty$ and thus $\mathrm{Var}(Y) = \mathrm{E}(Y^2) - \mathrm{E}^2(Y) = \mathrm{Var}\big(\mathrm{E}(Y|X)\big) + \mathrm{E}\big(\mathrm{Var}(Y|X)\big) < \infty$, which leads us to $0 \leq \mathrm{E}\big(\mathrm{Var}(Y|X)\big) = \mathrm{E}(Y^2) - \mathrm{E}\big(\mathrm{E}^2(Y \mid X)\big)$. This implies that $\mathrm{E}\big(\mathrm{E}^2(Y \mid X)\big) = \mathrm{E}\big(g^2(X)\big) \leq \mathrm{E}(Y^2)$ and thus $g(X) \in L^2$.

[9] Some authors require even more than validity. For example, Hastie et al. (2009, p. 28) additionally presume that ε is independent of X. In any case, validity is an essential assumption of the celebrated Gauss-Markov theorem.

[10] I will come back to this crucial point in Section 2.7.

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 $l : \mathbb{R} \to \mathbb{R}$ 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. The problem is that the marginal impact of X on Y can be grossly misjudged by $\frac{\partial}{\partial x} f(X)$, i.e., by the partial derivative of the chosen regression function $f \in \mathcal{F}$, if the given regression model $Y = f(X) + \varepsilon$ is invalid.

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),[11] or that we even want to assess the overall distribution of Y given X. This has become increasingly popular in recent 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:

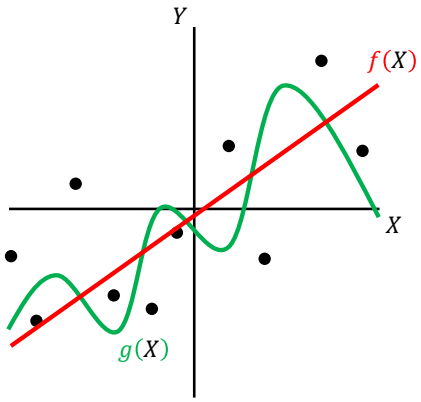
1. Mean regression is still the most popular regression approach in empirical research.
2. It is quite appealing from a mathematical point of view, since we are able to apply well-known rules from probability theory in order to derive the desired results.
3. There is a close relationship between the 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.

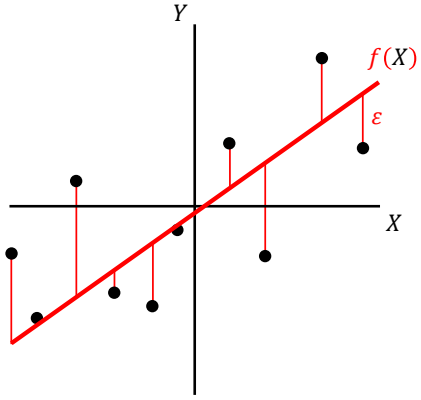
2.3. The Connection between Prediction and Description

The residual $\varepsilon = Y - f(X)$ represents the *prediction error* of the regression model $Y = f(X) + \varepsilon$, whereas its *description error* is given by $g(X) - f(X) = \varepsilon - \epsilon$. Now,

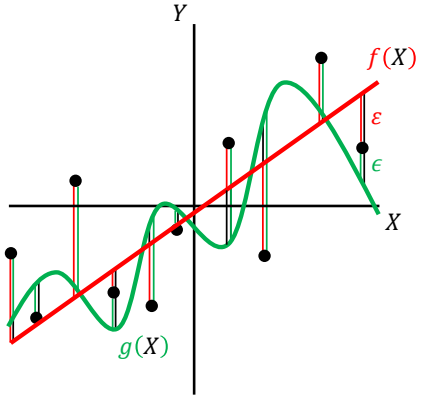
[11] See, e.g., Schulze Waltrup et al. (2015) for a nice overview of these measures.



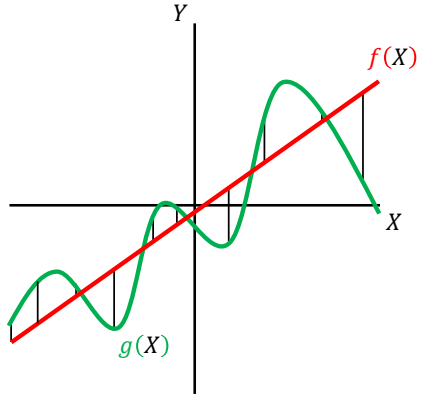
(i) Regression functions f and g



(ii) Prediction: $\min_{f\in\mathcal{F}} \mathrm{E}(\varepsilon^2)$!



(iii) $g(X) - f(X) = \varepsilon - \epsilon$



(iv) Description: $\min_{f\in\mathcal{F}} \mathrm{E}\big((\varepsilon-\epsilon)^2\big)$!

Figure 1: Prediction vs. description.

- the goal of prediction is to minimize the mean square prediction error $\mathrm{E}(\varepsilon^2)$, whereas
- description aims at minimizing the mean square description error $\mathrm{E}\big((\varepsilon-\epsilon)^2\big)$.

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

Figure 1 illustrates the connection between prediction and description. Figure 1 (i) shows 10 observations of Y, based on the true (but unknown) regression function g, while f represents our proposed regression function. Prediction tries to minimize the (mean square) distance between f and the observations of Y (see Figure 1 (ii)). Figure 1 (iii) just demonstrates that the description error $g(X) - f(X)$ equals the difference between the model error, ε, and the true error ϵ. As we can see, description aims at minimizing the distance between the proposed regression function f and the true regression function g (see Figure 1 (iv)), not the observations of Y.

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.[12]

[12] 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.

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 regression model 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* 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.[13]

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

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

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

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

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:

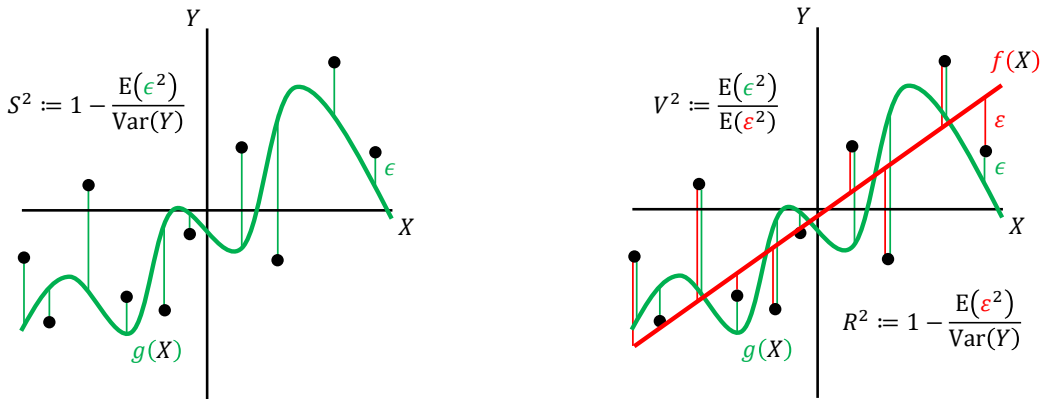
1. The first part, $\mathrm{E}\big((Y - g(X))^2\big)$, measures the fluctuation of Y around $\mathrm{E}(Y \mid X)$, which *shall not* be explained by a regression of Y on X, whereas

2. the second part, $\mathrm{E}\big((g(X) - f(X))^2\big)$, quantifies the deviation of $\mathrm{E}(Y \mid X)$ from $f(X)$, 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

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

[13] Overfitting is not an issue at all when trying to predict Y by X, provided that we know the probability measure P.



(i) Explanation power

(ii) Validity (V^2) and prediction power (R^2)

Figure 2: Regression measures.

provided that $Y \neq f(X)$, i.e., $\varepsilon \neq 0$. Otherwise, we have $f(X) = g(X)$ and so we may set $V^2 = 1$.[14] 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 $Y = g(X) \neq f(X)$, 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

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

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 $R^2 = 1$ 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.[15]

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

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

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 $g(X) = \mathrm{E}(Y \,|\, X)$, we have $\mathrm{E}(\epsilon \,|\, X) = 0$ and thus $\mathrm{E}(\epsilon) = \mathrm{Cov}(g(X), \epsilon) = 0$. Hence, let

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

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

[15] 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., $R^2 = -1$.

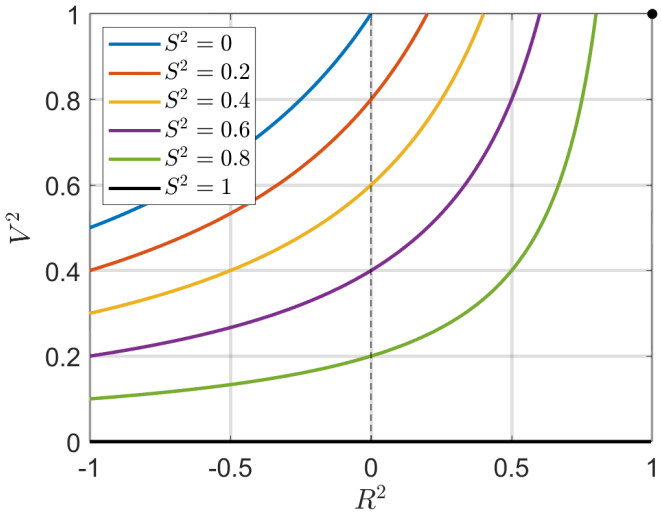


Figure 3: Dependence between V^2 and R^2 given S^2.

be the *explanation power* of X, i.e., the proportion of the total variance of Y that can be explained by the variance of $g(X)$.[16] The explanation power does not depend on the chosen regression function—it only depends on the choice of regressors.

The regression measures are illustrated in Figure 2. The explanation power of X is quantified by S^2, which can be interpreted as the R^2 of the true regression equation $Y = g(X) + \epsilon$ (see Figure 2 (i)). Further, the validity measure V^2 is illustrated in Figure 2 (ii), which contains also the prediction power R^2 of the proposed regression model $Y = f(X) + \varepsilon$.

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*

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

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 $Y = f(X) + \varepsilon$ depends (i) on the explanation power of X and (ii) on the prediction power of $f(X)$. This creates a simple but counterintuitive trade-off, which can be best understood by observing Figure 3. 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 $V^2 = 1$. Now, let us fix any value of R^2. The higher S^2, the lower V^2! Put another way, the higher the explanation power of X, the lower the validity of f, given that the prediction power of the regression model is fixed.

This strange relationship between V^2, R^2, and S^2 can be explained as follows: Suppose you ask your friend to prepare an elaborate dish. You agree that you will bring him all the necessary

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

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

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

ingredients in return. The better the ingredients you get from the supermarket, the higher S^2, and the closer your friend comes to the dish, the higher R^2. Now, suppose the shopping was great, i.e., your goods are fine, but nevertheless your friend screwed up. Put another way, S^2 is close to 1 and R^2 is close to 0. Then, we may conclude that his cooking skills are bad. In other words, V^2 is close to 0, too. 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 is unknown in real life and thus we cannot draw any conclusion about the validity of the regression model just 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, the 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 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 $A^2 > -1$ because, according to Theorem 1, $S^2 = 0$ implies $V^2 > 0$. The following result is an immediate consequence of Theorem 1 and so its proof can be skipped.

Corollary 1 (Accuracy). *Let $Y = f(X) + \varepsilon$ be any regression model. We have $A^2 = V^2 R^2$.*

Hence, the accuracy, A^2, of any regression model equals the product of its validity, V^2, and its prediction power, R^2, where V^2 and R^2 are related to one another according to Figure 3 or, equivalently, by Theorem 1. Table 2 summarizes the regression measures.

Ideally, we should achieve $A^2 = 1$, which implies $S^2 = V^2 = 1$, but this is virtually impossible in real-life applications. The regression model with empty set of regressors is given by

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

This model is always valid, but we have $S^2 = 0$ 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 2 (Necessary Conditions). *Suppose 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) $\mathrm{Var}(\hat{\varepsilon}) = \mathrm{E}\big(\mathrm{Var}(\hat{\varepsilon} \mid X)\big)$

(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 $\tilde{\varepsilon} = \hat{\varepsilon}$.

Theorem 2 immediately leads us to the following corollary. Thus, its proof can be skipped.

Corollary 2. *If the conditions of Theorem 2 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 $Y = f(X) + \varepsilon$ is valid, but they need not be independent. For example, if (X, Y) has an elliptical distribution (Cambanis et al., 1981, Kelker, 1970), 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 satisfied if the regression model $Y = f(X) + \varepsilon$ is valid. This holds true even if f is *nonlinear*. Nevertheless, the exogeneity of the regressors $X_1, \ldots, X_m$ is only a necessary, but not a sufficient condition for validity. By using a parametric family $\mathcal{F} = \{f(\cdot, \theta) : \theta \in \Theta \subseteq \mathbb{R}^q\}$ of regression functions, exogeneity can often be accomplished by specifying θ 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 $Y = \alpha + \beta' X + \varepsilon$ with $\beta = \mathrm{Var}(X)^{-1}\mathrm{Cov}(X, Y)$ are always exogenous just by construction, although the linear regression model can be highly invalid. 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 $n \times m$ matrix

$$\mathbf{X} = \begin{bmatrix} X_{11} & \cdots & X_{1m} \\ \vdots & & \vdots \\ X_{n1} & \cdots & X_{nm} \end{bmatrix}$$

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

$$\mathbf{x} = \begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix}$$

denotes some realization of $\mathbf{X}$. Thus, $\mathbf{x}$ is a real-valued $n \times m$ 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 $\mathrm{E}(\boldsymbol{\varepsilon} \mid \mathbf{X}) = 0$.

Further, the *Gauss-Markov assumption* presumes that $\mathbf{Y} = \alpha 1 + \mathbf{X}\beta + \boldsymbol{\varepsilon}$, $n > m$, $\mathrm{rk}\big[1\ \mathbf{X}\big] = m + 1$,

$\mathrm{E}(\boldsymbol{\varepsilon} \mid \mathbf{X}) = 0$, and $\mathrm{Var}(\boldsymbol{\varepsilon} \mid \mathbf{X}) = \sigma^2 \mathbf{I}_n$ with $\sigma^2 > 0$. It represents another classical assumption of linear regression analysis, which is even stronger than strict exogeneity.

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

Theorem 3 (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 Gauss-Markov assumption 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 4 (Elliptical Distributions). *Suppose $Z = (Z_1, \ldots, Z_d)$ 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 5 (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 are 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 $\mathcal{G}(X) := \{f(X) : f \in \mathcal{G}\}$ 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

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

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 $\mathcal{F}(X) := \{f(X) : f \in \mathcal{F}\}$ 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 6 (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 $\mathrm{E}(\varepsilon) = 0$ and thus $\mathrm{Cov}(f(X), \varepsilon) = 0$.[17] For example, let $\mathcal{L} := \{x \mapsto a + b'x : a \in \mathbb{R}, b \in \mathbb{R}^m\}$ 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.

Proposition 3. *Suppose the regression function $\hat{f} \in \mathcal{F}$ is optimal among $\mathcal{F}$ and consider any other 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 6 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.

[17] As already mentioned in Section 2.3, this implies that $R^2 \in [0, 1]$, in which case the coefficient of determination quantifies the proportion of the total variance of Y that can be explained by the variance of $f(X)$.

Theorem 7 (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 regression function may even be unique. Nonetheless, $\hat{f}$ cannot be valid.[18] In this case, the error $\hat{\varepsilon} = Y - \hat{f}(X)$ even need not be orthogonal to $\mathcal{F}(X)$. For example, suppose $Y = 0X + \epsilon$, 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 $\mathrm{E}(\hat{\varepsilon}) = 0$. 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 typical 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 assume 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 $\mathrm{E}(XY) = 1$ 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 $\mathrm{E}\big(X^3\varepsilon\big) = -2$.[19] 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.

To sum up, the Gauss-Markov assumption is stronger than strict exogeneity, which implies validity, which is equivalent to global optimality, which is sufficient for optimality, orthogonality, and exogeneity. Furthermore, 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

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

[19] 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.

analysis, i.e., $\mathcal{F} = \mathcal{L}$, then optimality, orthogonality, and exogeneity are even equivalent. Moreover, if $\mathcal{F}$ is adequate, then optimality, orthogonality, validity, and global optimality are equivalent. Finally, if $\mathcal{F} = \mathcal{L}$ is adequate, it turns out that optimality, orthogonality, validity, global optimality, and exogeneity are equivalent.[20] Nonetheless, in general, exogeneity is only a *necessary* but not a sufficient condition for validity. That is, it is not enough to guarantee that the typical exogeneity conditions are satisfied if we want to construct a valid regression model.

The following theorem summarizes our previous findings, where the abbreviation

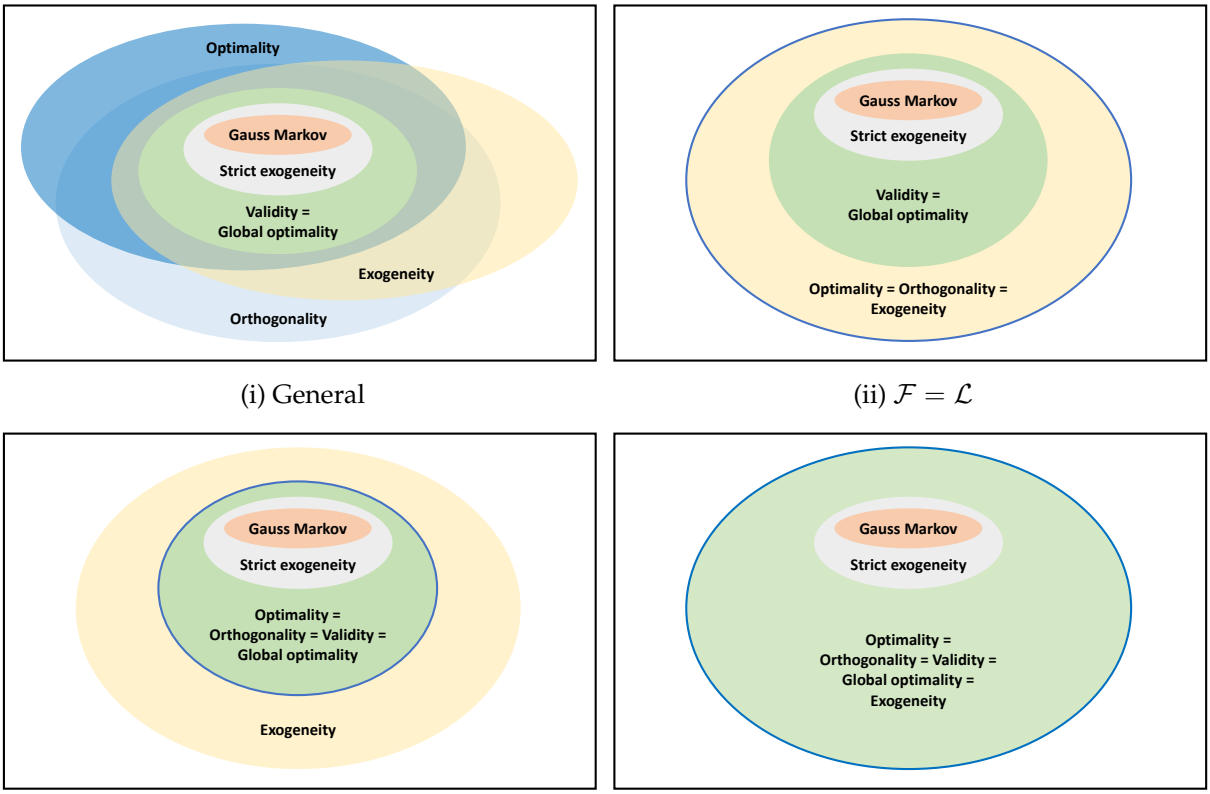
- V means that $f \in \mathcal{F}$ is *valid*, i.e., $\mathrm{E}(\varepsilon \mid X) = 0$,
- GM means that the *Gauss-Markov assumption* is satisfied,
- SE means that $\mathbf{X}$ is *strictly exogenous*, i.e., $\mathrm{E}(\boldsymbol{\varepsilon} \mid \mathbf{X}) = 0$,
- OP means that f is *optimal* 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 8 (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 4 illustrates some important aspects of Theorem 8. 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.

[20] A typical example is a linear regression of Y on X where (X, Y) is elliptically distributed.



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

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

Figure 4: Hierarchy of regression properties.

2.7. Variable Selection and Model Specification

A regression equation $Y = f(X) + \varepsilon$ is satisfied just by the very definition of $\varepsilon := Y - f(X)$. It becomes a regression *model* only if we make some specific assumption A about the joint distribution of X and ε. Thus, we may consider the given model *well specified* if and only if A is satisfied. For example, $A\colon \mathrm{E}(\varepsilon \mid X) = 0$ claims that the regression model $Y = f(X) + \varepsilon$ is valid.[21] However, there is no common understanding of what constitutes a well-specified regression model, and there are many myths and legends circulating. These are highly misleading in the light of what we know so far. Hence, they shall be clarified in this section before going further.

There is a common misconception in regression analysis, namely that there exists only one set of relevant regressors. Thus, we should not use any other set of regressors to describe Y. The objective is to find the relevant regressors and estimate the parameters of the corresponding model. This seems to be one of the most persistent mistakes in regression analysis.

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

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

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

in which case

$$Y = \frac{2}{3}X_1 + \frac{2}{3}X_2 + \varepsilon_2 \tag{5}$$

is the only linear regression model that involves X_1 and X_2 such that all exogeneity conditions are satisfied. The common misconception mentioned above seems to be due to the following arguments (see, e.g., Fomby et al., 1984, Section 18.2.1).[22] Equation 5 turns into

$$Y = \frac{2}{3}X_1 + \varepsilon_1 \tag{6}$$

with $\varepsilon_1 = \frac{2}{3}X_2 + \varepsilon_2$ if we omit X_2, in which case X_1 is correlated with ε_1. More precisely, it holds that

$$\mathrm{Cov}(X_1, \varepsilon_1) = \mathrm{Cov}\left(X_1, \frac{2}{3}X_2 + \varepsilon_2\right) = \frac{2}{3}\frac{1}{2} + 0 = \frac{1}{3} \neq 0,$$

i.e., X_1 is endogenous. Thus, (i) taking only X_1 as a single regressor into account leads us astray, while (ii) adding X_3 to Equation 5 does not make any sense at all because X_3 is irrelevant. This means that its regression coefficient must be zero and all other parameters remain unaffected.

Unfortunately, both arguments are flawed. We can very well ignore any regressor without producing endogeneity. Furthermore, the regression coefficient of each additional regressor can (and, in general, it *will*) be nonzero, too. Ignoring some variable just means to reduce the set of regressors, whereas taking some additional variable into account means to extend that set. However, a linear regression model obeys all exogeneity conditions if its parameters are specified in the usual way, i.e., by $\beta = \mathrm{Var}(X)^{-1}\mathrm{Cov}(X, Y)$ and $\alpha = \mathrm{E}(Y) - \beta'\mathrm{E}(X)$.[23] Thus, we can choose any *arbitrary set* $\mathcal{S} = \{X_1, \ldots, X_m\}$ of regressors to create a linear regression model in which all exogeneity conditions are satisfied just by construction.[24]

For example, what happens if we exclude X_2? Then, our set of regressors shrinks to $\mathcal{S} = \{X_1\}$, and so the linear regression model turns into

$$Y = X_1 + \varepsilon_1$$

with $\varepsilon_1 = Y - X_1$, in which X_1 and ε_1 are still uncorrelated. Further, if we include X_3, our set of regressors expands to $\mathcal{S} = \{X_1, X_2, X_3\}$, and the linear regression model turns into

$$Y = \frac{1}{2}X_1 + \frac{1}{2}X_2 + \frac{1}{2}X_3 + \varepsilon_3$$

with $\varepsilon_3 = Y - \frac{1}{2}X_1 - \frac{1}{2}X_2 - \frac{1}{2}X_3$. Thus, we have $\beta_3 = \frac{1}{2} \neq 0$, which means that X_3 is in fact relevant, and we can see that all exogeneity conditions are again satisfied.[25] These arguments do *not* require us to make any distributional assumption about X and ε.

[22] Although Fomby et al. (1984) refer to the estimation of α and β, their arguments are based on model specification.

[23] Here, any component of β may be zero, in which case the corresponding regressor disappears.

[24] Theorem 4 even guarantees that the resulting regression model is *valid* if the random variables possess a joint (nonsingular) elliptical distribution and the regression parameters are specified in the usual way.

[25] However, none of these models need to be *valid*, since (according to Figure 4 or, equivalently, Theorem 8) exogeneity is only a necessary but not a sufficient condition for validity.

To sum up, the specification of a linear regression model essentially depends on the given set of regressors, provided that we choose the regression parameters in the usual way, i.e., such that all exogeneity conditions are satisfied by construction. This holds true for each set of regressors. Hence, endogeneity is no problem at all in linear regression analysis, provided that the regressors are observable and that there are no measurement errors.

We conclude that there is not only one appropriate way to describe Y.[26] If we have found a well-specified regression model based on some set of regressors, a model based on any other set of regressors can be well specified, too. Hence, it makes no sense at all to say that a regression model is well specified only if we limit ourselves to specific regressors. In fact, we can choose any other set of regressors to construct a well-specified model, and the number of well-specified regression models for Y can even be infinite, since the number of potential regressors is infinite, too.[27] Whether or not some regression model is well specified only depends on A, i.e., on our assumption about the distribution of (X, ε), given the set $\mathcal{S} = \{X_1, \ldots, X_m\}$ of regressors.

Now, I would like to discuss another issue related to variable selection and model specification. Suppose the true regression equation is $Y = X_1 + X_2 + \epsilon$, where the variables $X_1, X_2, \epsilon \sim \mathcal{N}(0,1)$ are (mutually) independent and $\mathcal{F} = \mathcal{L}$. Since we have $g(X) = X_1 + X_2$ with $X = (X_1, X_2)$, the family of linear regression functions, $\mathcal{L}$, is adequate and Theorem 8 tells us that validity and exogeneity are equivalent in this particular case. Now, consider the simple 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 have also $\mathrm{E}(\varepsilon \mid X_1) = 0$, which suggests that the given regression model is valid. Nevertheless, it is in fact *invalid* because $f(X) = X_1 \neq X_1 + X_2 = g(X)$! What is the reason for this paradox?

Our set $\mathcal{S}$ of regressors consists of X_1 and X_2. Therefore, we actually have *three* exogeneity conditions, but the third one is violated, since

$$\mathrm{Cov}(X_2, \varepsilon) = \mathrm{Cov}(X_2, X_2 + \epsilon) = \mathrm{Var}(X_2) + \underbrace{\mathrm{Cov}(X_2, \epsilon)}_{=0} = 1.$$

Furthermore,

$$\mathrm{E}(\varepsilon \mid X_1, X_2) = \mathrm{E}(X_2 + \epsilon \mid X_1, X_2) = \mathrm{E}(X_2 \mid X_1, X_2) + \underbrace{\mathrm{E}(\epsilon \mid X_1, X_2)}_{=0} = X_2 \neq 0$$

clearly demonstrates that the given regression model is invalid. By contrast, if we had initially chosen X_1 as a single regressor, i.e., $\mathcal{S} = \{X_1\}$, the simple linear regression model $Y = X_1 + \varepsilon$ would have satisfied all (two) exogeneity conditions, and its mean conditional error $\mathrm{E}(\varepsilon \mid X_1)$ would have been zero. Hence, in that case, this linear regression model would have been valid.

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. Hence, if the family $\mathcal{F}$ of regression

[26] This is supported also by the simple fact that the trivial regression model $Y = Y$ is always valid.

[27] For any set of regressors, there exists a unique valid regression function f, namely the true regression function g.

functions is parametric, we have to estimate the parameters of the (optimal) regression function. This creates estimation risk and can lead to overfitting, which 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 choice of $\mathcal{S}$, i.e., of the set of regressors, should depend on our principal goal:

- If we want to *predict* Y by X, the choice of $\mathcal{S}$ is rather arbitrary, since we only try to achieve a strong prediction power. Hence, $X_1, \ldots, X_m$ are just chosen for pure statistical reasons.
- By contrast, if we want to *describe* the impact of X on Y, the choice of $\mathcal{S}$ 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. This goal can be accomplished without validity, and in general the regressors even need not be exogenous. By contrast, description aims at maximizing V^2. In that case, we try to come as close as possible to the true regression function g, given our set of regressors, by searching for some regression function $f \in \mathcal{F}$ that minimizes the mean square description error. Then, once again we should try to find an optimal regression function in $\mathcal{F}$. In general, however, this is far from sufficient—even not if we focus on linear regression analysis, i.e., $\mathcal{F} = \mathcal{L}$, and guarantee that all exogeneity conditions are satisfied. The problem is that $\mathcal{L}$ can be inadequate, in which case each regression function $f \in \mathcal{F}$ is invalid. Thus, I strongly recommend to apply a validity test, i.e., a test for the null hypothesis $H_0 \colon f = g$, which is equivalent to $H_0 \colon \mathrm{E}(\varepsilon \,|\, X) = 0$. Such a test is developed in the next section.

3. The Validity Test

3.1. Test Statistic

3.1.1. Basic Motivation

Theorem 9 (Residuals). *Let $f \in \mathcal{F}$ be any regression function, $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a sample of $n \geq 1$ 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. Then, we have*

$$\mathrm{P}(\boldsymbol{\varepsilon} \leq \boldsymbol{e} \,|\, \mathbf{X}) = \prod_{i=1}^{n} \mathrm{P}(\varepsilon_i \leq e_i \,|\, \mathbf{X})$$

for all $\boldsymbol{e} = (e_1, \ldots, e_n) \in \mathbb{R}^n$*. Further, it holds that*

$$\mathrm{P}(\varepsilon_i \leq e_i \,|\, \mathbf{X}) = \mathrm{P}(\varepsilon_i \leq e_i \,|\, X_i)$$

for all $e_i \in \mathbb{R}$ *and* $i = 1, \ldots, n$*.*

Hence, if the given observations of (X, Y) are independent, 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. In particular, if the regression function f is valid, we have

$$\mathrm{E}(\varepsilon_i \,|\, \mathbf{X}) = \mathrm{E}(\varepsilon_i \,|\, X_i) = 0$$

for $i = 1, \ldots, n$. Further, the residuals $\varepsilon_1, \ldots, \varepsilon_n$ are independent conditionally on $\mathbf{X}$.

Now, let $\mathcal{P}$ be any $n \times n$ permutation matrix, which may depend on the random matrices $\mathbf{X}$, $\mathbf{Y}$, or on any other (random) quantity. Hence, $\boldsymbol{\varepsilon}^* = (\varepsilon_1^*, \ldots, \varepsilon_n^*) = \mathcal{P}\boldsymbol{\varepsilon}$ is a permutation of $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)$. For example, assume that the errors $\varepsilon_1, \ldots, \varepsilon_n$ are sorted in ascending order $\varepsilon_1^* \leq \ldots \leq \varepsilon_n^*$. Then, we can expect that the mean of ε_1^* is less than the mean of ε_2^*, etc. Thus, in general, the distribution of $\boldsymbol{\varepsilon}$ conditional on $\mathbf{X}$ is not invariant against some permutation.

Suppose the permutation is based on $\mathbf{X}$, i.e., that $\mathcal{P}$ is a function of $X_1, \ldots, X_n$. Simply put, assume that $\mathcal{P}$ is fixed with $\mathbf{X}$. Then, if the regression function f is valid, we have

$$\mathrm{E}(\varepsilon_i^* \,|\, \mathbf{X}) = \mathrm{E}(\varepsilon_i \,|\, \mathbf{X}) = \mathrm{E}(\varepsilon_i \,|\, X_i) = 0$$

for $i = 1, \ldots, n$, i.e., $\mathrm{E}(\boldsymbol{\varepsilon}^* \,|\, \mathbf{X}) = 0$. Moreover, according to Theorem 9, even the joint distribution of $\boldsymbol{\varepsilon}$ given $\mathbf{X}$ does not change after the permutation, provided that each ε_i is independent of X_i.

Let $\left(\varepsilon_1^*, \ldots, \varepsilon_k^*\right)$ be any selection of $k \in \left\{1, \ldots, n\right\}$ elements of $\boldsymbol{\varepsilon}$ that is determined by $\mathbf{X}$. The selection can be viewed as a result of the following two-step procedure: (i) Consider some permutation $\boldsymbol{\varepsilon}^* = \mathcal{P}\boldsymbol{\varepsilon}$ that is based on $\mathbf{X}$ and (ii) choose the first k elements of $\boldsymbol{\varepsilon}^*$. Theorem 9 implies that $\mathrm{E}(\varepsilon_i^* \,|\, \mathbf{X}) = 0$ for $i = 1, \ldots, k$, provided that the given regression function, f, is valid. Hence, the conditional mean of $\frac{1}{\sqrt{k}} \sum_{i=1}^{k} \varepsilon_i^*$ is zero, too. Therefore, a natural quantity for testing the null hypothesis that $\mathrm{E}(\varepsilon_i \,|\, X_i) = 0$ for $i = 1, \ldots, n$, i.e., that f is valid, appears to be

$$T_k = \frac{1}{k} \left(\sum_{i=1}^{k} \varepsilon_i^* \right)^2 .$$

There exist $s_n = 2^n - 1$ possible selections of errors and s_n grows exponentially to infinity as $n \to \infty$. However, it makes not much sense to consider all possible selections in order to test for the validity of f. Instead, I will concentrate on n^2 specific selections. More precisely, let $\varepsilon_{i,j}$ be the error that belongs to the jth nearest neighbor of X_i for $i, j = 1, \ldots, n$.[28] Let us take the Euclidean distance as the default metric, although nothing speaks against using any other metric.

[28] Here, it is implicitly assumed that each realization of X has precisely one jth nearest neighbor for $j = 1, \ldots, n$. Further, the nearest neighbor of each observation is the observation itself.

For example, suppose $n \geq 3$ and set $k = 3$. Let us assume that $X_1 = 1$, $X_2 = -3$, and $X_3 = 4$ with $\varepsilon_1 = 0$, $\varepsilon_2 = -1$, and $\varepsilon_3 = 2$. The 1st nearest neighbor of X_1 is $X_1 = 1$ itself and so we have $\varepsilon_{1,1} = \varepsilon_1 = 0$. The 2nd nearest neighbor of X_1 is $X_3 = 4$, which leads us to $\varepsilon_{1,2} = \varepsilon_3 = 2$. Finally, the 3rd nearest neighbor of X_1 is $X_2 = -3$ and thus we obtain $\varepsilon_{1,3} = \varepsilon_2 = -1$. Now, the corresponding quantity is $T_{3,1} = \frac{1}{3}\big(\sum_{j=1}^{3} \varepsilon_{1,j}\big)^2$. In the same way, we obtain $T_{3,2} = \frac{1}{3}\big(\sum_{j=1}^{3} \varepsilon_{2,j}\big)^2$ and $T_{3,3} = \frac{1}{3}\big(\sum_{j=1}^{3} \varepsilon_{3,j}\big)^2$, which need not equal $T_{3,1}$ if $n > 3$. This leads us to the test statistic

$$T = \frac{1}{n^2} \sum_{k=1}^{n} \sum_{i=1}^{n} T_{k,i} = \frac{1}{n^2} \sum_{k=1}^{n} \sum_{i=1}^{n} \frac{1}{k} \left(\sum_{j=1}^{k} \varepsilon_{i,j} \right)^2 ,$$

which can be re-written as

$$T = n \left\{ \frac{1}{n} \sum_{k=1}^{n} \frac{k}{n} \left[\frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{k} \sum_{j=1}^{k} \varepsilon_{i,j} \right)^2 \right] \right\} .$$

The latter representation of T clarifies that

$$n^{-1} T \to \mathrm{E}\Big(W \mathrm{E}^2\big(\varepsilon \,|\, X \in B_W(X)\big) \Big)$$

under very mild regularity conditions concerning the errors $\varepsilon_1, \ldots, \varepsilon_n$, where $B_w(x) \subseteq \mathbb{R}^m$ with $w \in [0, 1]$ and $x \in D$ is the smallest ball around x such that $\mathrm{P}\big(X \in B_w(x)\big) = w$.[29] Further, W is uniformly distributed on $[0, 1]$ and independent of (X, ε).

Hence, if the sample size n is large, a realization of T can be understood as n times the mean of the squared and weighted expectation of ε under the condition that X belongs to some ball. The center of that ball is given by x, i.e., the realization of X, whereas its size is quantified by w, i.e., the realization of W. By this means, the given test tries to detect both local deviations and global deviations of the proposed regression function, f, from the true regression function g. The smaller w, the more the test focuses on errors that are associated with regressor values around x. Further, since $\mathrm{E}^2\big(\varepsilon \,|\, X \in B_w(x)\big)$ is multiplied by w, the smaller the contribution of the squared expectation of those errors. Hence, deviations of f from g that occur on a broader spectrum have more weight than deviations that occur in specific regions of the support of X.

We can assume, more generally, that W is an absolutely continuous random variable between 0 and 1 with density function ψ, which leads us to the test statistic

$$T = \frac{1}{n^2} \sum_{k=1}^{n} \sum_{i=1}^{n} \psi\left(\frac{k}{n}\right) \frac{1}{k} \left(\sum_{j=1}^{k} \varepsilon_{i,j} \right)^2 .$$

This enables us to control the potential size of the ball around X. The more we focus on X, i.e., the steeper the distribution of W on the left, the better we can detect complex deviations of f from g. However, this can be at the expense of the test power. Conversely, the more we reduce

[29] Once again, it is implicitly assumed that $B_W(X)$ exists, almost surely.

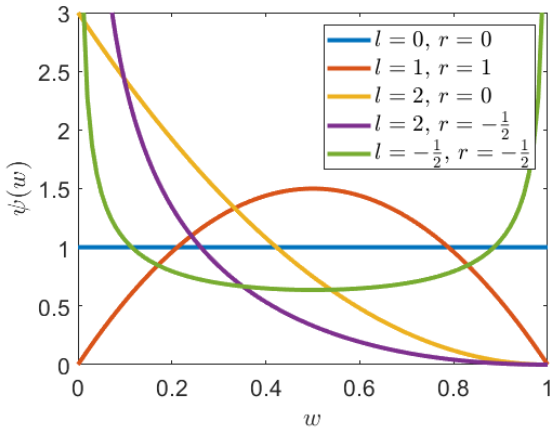


Figure 5: Different choices of the weight function.

our focus, i.e., the steeper the distribution of W on the right, the stronger the power of the test can be, but then it might happen that complex deviations are not well recognized. I recommend to use the probability density function of the beta distribution, i.e.,

$$w \mapsto \psi(w) \propto (1-w)^l w^r,$$

where $l, r > -1$ control how much weight we give to the left and right side of $[0,1]$. The default setting is $l = r = 0$, in which case we obtain the density function of the uniform distribution on $[0,1]$, i.e., $\psi \equiv 1$. Other choices of l and r are visualized in Figure 5.

In most practical applications, we estimate the regression function f by some estimator $\hat{f}$, in which case we can approximate ε_i by $\hat{\varepsilon}_i = Y_i - \hat{f}(X_i)$ for $i = 1, \ldots, n$. Thus, we finally obtain the test statistic

$$T = \frac{1}{n^2} \sum_{k=1}^{n} \sum_{i=1}^{n} \psi\left(\frac{k}{n}\right) \frac{1}{k} \left(\sum_{j=1}^{k} \hat{\varepsilon}_{i,j} \right)^2 .$$

The distribution of T under the null hypothesis that f is valid can be very well approximated by a residual bootstrap, which is explained in the next section.

3.1.2. Bootstrap

Let $\hat{f}$ by any consistent estimator for the regression function f. A typical example is a parametric regression function $f = f(\cdot\,, \theta)$ that is continuous in θ. Then, $\hat{f} = f(\cdot\,, \hat{\theta})$ is consistent for f if $\hat{\theta}$ is a consistent estimator for θ. We could assume, e.g., that $\hat{\theta}$ is an ordinary least-squares (OLS) estimator, which minimizes $\theta \mapsto \frac{1}{n} \sum_{i=1}^{n} \big(Y_i - f(X_i\,, \theta)\big)^2$ and corresponds to the solution of

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} f(X_i, \hat{\theta})\, \hat{\varepsilon}_i = 0$$

with $\hat{\varepsilon}_i = Y_i - f(X_i\,, \hat{\theta})$ for $i = 1, \ldots, n$, provided that the conditions mentioned in Section 2.2.1 are satisfied. However, the given validity test applies to *any* choice of the parameter (vector)

$\theta \in \Theta$, provided that its estimator $\hat{\theta}$ is consistent for θ and $\theta \mapsto f(\cdot, \theta)$ is continuous. Hence, we need not limit ourselves to an optimal regression function f among $\mathcal{F}$ in order to test for validity, although this approach is natural in the light of Theorem 8. In particular, we can also choose some *explicit* regression function $f \in \mathcal{F}$, in which case $\hat{f} \equiv f$. Anyway, $\hat{f}$ should be given before starting the bootstrap. The approximate residuals are represented by $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n$ with $\hat{\varepsilon}_i = y_i - \hat{f}(x_i)$, where (x_i, y_i) denotes the given observation of (X_i, Y_i) for $i = 1, \ldots, n$.

Now, one should proceed as follows:

1. Centralize the (approximate) residuals $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n$.[30]
2. Draw $N \gg 0$ bootstrap samples of size n with replacement from the vector of centralized residuals. The given result is an $n \times N$ matrix $\left[\epsilon_{i,b}\right]$ of simulated *true* residuals.[31]
3. Simulate N realizations of $\mathbf{Y}$ by $Y_{i,b} = \hat{f}(x_i) + \epsilon_{i,b}$ for $i = 1, \ldots, n$ and $b = 1, \ldots, N$.[32]
4. Apply the given estimator $\hat{f}$ for the regression function f to each simulated realization of $\mathbf{Y}$ by using the fixed regressor matrix $\mathbf{x}$. The simulated estimates of f are $\hat{f}_1, \ldots, \hat{f}_N$.
5. Compute the corresponding error $\hat{\varepsilon}_{i,b} = Y_{i,b} - \hat{f}_b(x_i)$ for $i = 1, \ldots, n$ and $b = 1, \ldots, N$.
6. Compute T on the basis of $\hat{\varepsilon}_{1,b}, \ldots, \hat{\varepsilon}_{n,b}$ for each bootstrap sample $b \in \{1, \ldots, N\}$.

This leads us to N simulated realizations $t_1, \ldots, t_N$ of T. The cumulative distribution function of T at $s \in \mathbb{R}$ under the null hypothesis that f is valid can be approximated by

$$\hat{F}_T(s) = \frac{1}{N} \sum_{b=1}^{N} \mathbf{1}_{t_b \leq s} \,.$$

In simulation studies, it can happen that the realization t of T is zero, in which case we have a perfect fit, i.e., $Y_i = \hat{f}(X_i)$ and thus $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n = 0$ for $i = 1, \ldots, n$. Nonetheless, due to numerical imprecisions, a number cruncher might come to the wrong conclusion that t is just *almost* zero. In order to prevent such inaccuracies, one should set t to zero whenever $t \leq \eta$, where η is the machine precision. In any case, if t is de facto zero, it is clear that we should not consider f invalid. Hence, let us define the (realized) p-value of T as

$$p := \begin{cases} 1 - \hat{F}_T(t), & t > 0 \\ 1, & t = 0 \end{cases} .$$

Now, one should reject the *null hypothesis* that the regression function f is valid if and only if p falls below some low level of significance. Synonymously, one should reject the *regression model* $Y = f(X) + \varepsilon$ if p is less than the given significance level.[33]

[30] If we already have $\frac{1}{n} \sum_{i=1}^{n} \hat{\varepsilon}_i = 0$ by construction, as in OLS regression, this step can be ignored.

[31] I use $N = 1000$ bootstrap replications in my own implementation of the validity test.

[32] Hence, $\hat{f}(x_i)$ is the bootstrap counterpart of $g(x_i)$ and it is presumed that ϵ_i is independent of X_i for $i = 1, \ldots, n$.

[33] An R implementation of this test can be found at `https://github.com/floschuetze/Validity`.

3.2. Linear Regression Models

Henceforth, 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 $\mathrm{E}(\epsilon \mid X) = 0$ 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 typical 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

To illustrate the practical importance of the validity test, let us consider the following example: Suppose $Y = g(X) + \epsilon$ 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. Further, consider the simple linear regression model $Y = \alpha + \beta X + \varepsilon$ with $\alpha = 0$ and $\beta = 2\phi(0)c$, where $\phi(0) = 0.3989$ represents the density of X at 0. Hence, the error of the given regression model is $\varepsilon = Y - 2\phi(0)cX$. It is evident that $\mathrm{E}(Y) = 0$ and thus $\mathrm{E}(\varepsilon) = 0$, too. Further, we have

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

because

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

Further, let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a sample of n independent observations of (X, Y).

To sum up, the linear regression model $Y = 0.7979cX + \varepsilon$ satisfies all exogeneity conditions. This means that X is exogenous and $f(X) = 0.7979cX$ is the best linear predictor of Y based on X. However, we have $g(X) = \mathrm{E}(Y \mid X) = \pm c$, depending on whether $X \leq 0$ or $X > 0$. Hence, 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 $\frac{\partial}{\partial x} f(X) = 0.7979c > 0$ for all $c > 0$. Thus, we have a spurious regression, which becomes all the more serious the higher c.

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 the corresponding regression measures ("Measures"). Figure 6 (i) clarifies how the regression measures depend on the parameter c. For example, let us assume that $c = 1$, which leads us

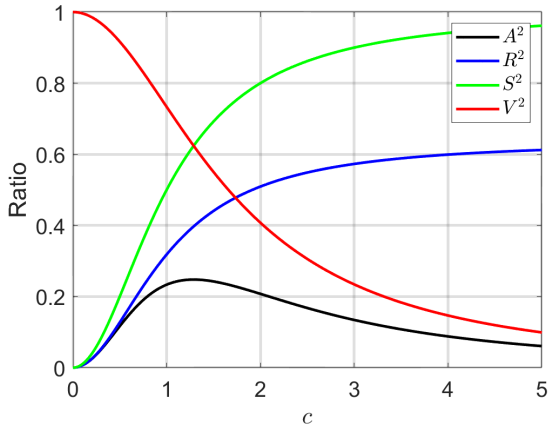
True:	$Y = \begin{cases} -c + \epsilon, & X \leq 0 \\ c + \epsilon, & X > 0 \end{cases}$ with $c \geq 0$ and $X, \epsilon \in \mathcal{N}(0,1)$ being independent.
Model:	$Y = \alpha + \beta X + \varepsilon$ with $\alpha = 0$ and $\beta = 0.7979c$.
Impact:	$\frac{\partial}{\partial x} f(X) = 0.7979c$,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

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

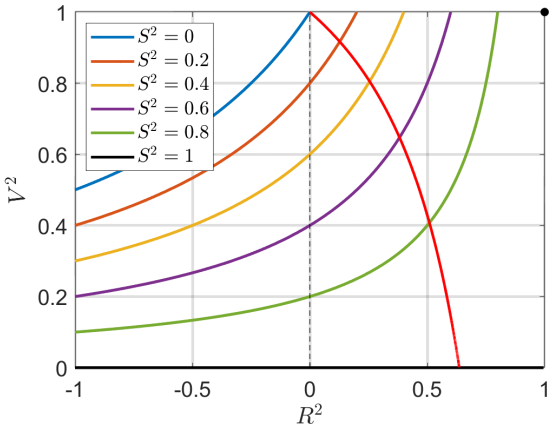
to the validity $V^2 = 0.7335$. Hence, although the linear regression model is clearly invalid, the coefficient of determination amounts to $R^2 = 0.3183$, which is quite high. In fact, as is shown by Figure 6 (i), the higher R^2, i.e., the stronger the prediction power of $f(X)$, the lower V^2, i.e., the more invalid the linear regression model. 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 becomes invalid! This adverse effect can be seen also in Figure 6 (ii), where the red line clarifies how R^2 and V^2 are connected through S^2. This is a striking example of why we should not rely on R^2 in order to verify the validity of any regression model.

Figure 6 (iii) 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 $\hat{\alpha} = -0.0002$ with a standard error of 0.1199 and $\hat{\beta} = 0.8047$ with a standard error of 0.1120. The corresponding regression line can be found in Figure 6 (iii), too. The ordinary R^2, based on the OLS estimates, is 0.3450. This is quite good compared with values that are usually obtained in econometric applications. Further, the corresponding p-value of the F-test for $H_0 \colon \beta = 0$ is virtually zero. Hence, X seems to have a significant (linear) impact on Y. Figure 6 (iv) is a residual plot. It contains the OLS predictions of Y together with the associated prediction errors, which look fine, too. Thus, at first glance, the linear regression model appears to be well specified, and the usual validity checks would never reveal that it is in fact invalid.

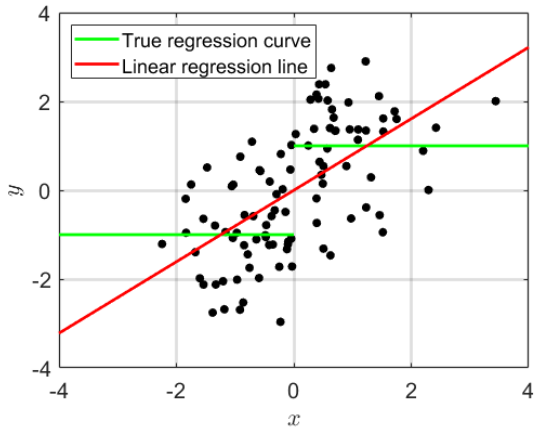
All in all, Figure 6 suggests that the linear regression model is appropriate, but we know that the exact opposite is true. To be more precise, the true regression curve, which is depicted in Figure 6 (i), is piecewise constant and it jumps up at $x = 0$. Thus, it is far away from being linear. With the best will in the world, this cannot be seen just by a visual inspection both of the scatter plot and of the residual plot. How does the validity test perform in this situation? The p-value amounts to 0.0030. Thus, after applying the validity test, the linear regression model can be rejected on every common significance level.



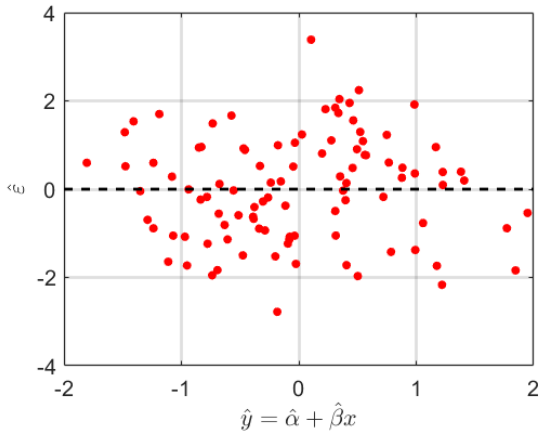
(i) Regression measures



(ii) Prediction power vs. validity



(iii) Scatter plot



(iv) Residual plot

Figure 6: Piecewise constant regression equation.

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:	$\mathrm{E}(\varepsilon) = \mathrm{Cov}(X, \varepsilon) = 0$, $\mathrm{E}(\varepsilon \mid X) = c(X^2 - 1)$
Measures:	$A^2 = \dfrac{b^2}{(2c^2+1)(b^2+2c^2+1)}$, $R^2 = \dfrac{b^2}{b^2+2c^2+1}$, $S^2 = \dfrac{b^2+2c^2}{b^2+2c^2+1}$, $V^2 = \dfrac{1}{2c^2+1}$

Table 4: Fact sheet of the quadratic regression equation.

Now, consider another example, in which the true regression function, g, is quadratic. More precisely, suppose

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

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

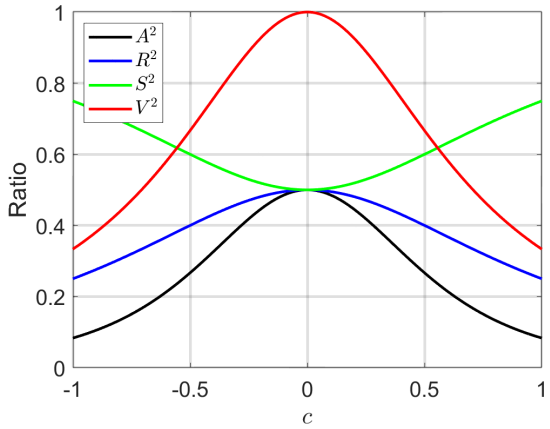
$$\mathrm{E}(\varepsilon) = c\mathrm{E}\big(X^2 - 1\big) + \mathrm{E}(\epsilon) = 0$$

and

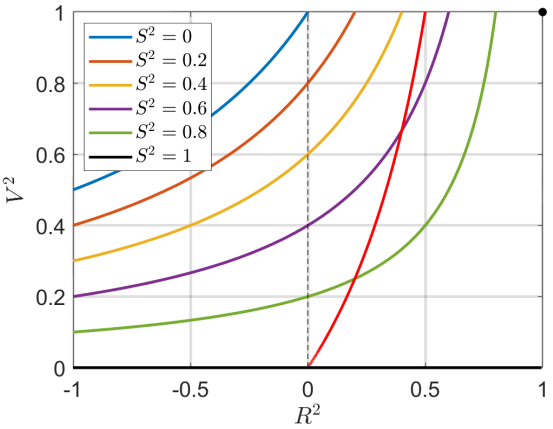
$$\mathrm{Cov}(X, \varepsilon) = c\mathrm{Cov}\big(X, X^2\big) + \mathrm{Cov}(X, \epsilon) = 0\,.$$

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 $\mathrm{E}(\varepsilon \mid X) = c(X^2 - 1)$. 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.

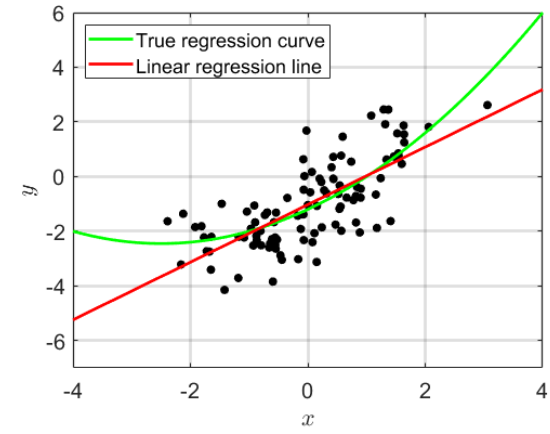
Let us assume that $b = 1$. Figure 7 (i) shows how the regression measures depend on the parameter c in that case, and the red line in Figure 7 (ii) clarifies how the validity and the prediction power of the linear regression model are connected with one another for different values of S^2. Further, Figure 7 (iii) contains a scatter plot based on 100 independent copies of



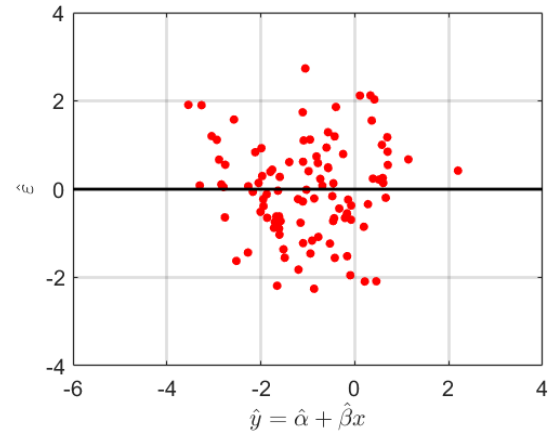
(i) Regression measures



(ii) Prediction power vs. validity



(iii) Scatter plot



(iv) Residual plot

Figure 7: Quadratic regression equation with $a = -1$ and $b = 1$.

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 true coefficient of determination is $R^2 = 0.4808$. This means 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, which have been used to create the regression line in Figure 7 (iii), are $\hat{\alpha} = -1.0273$ and $\hat{\beta} = 1.0533$ with standard errors 0.1086 and 0.1017, respectively. The ordinary R^2 based on the OLS estimates amounts to 0.5227. 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 7 (iv), where the residuals are based on the given OLS estimates of α and β.

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, i.e., $x \mapsto g(x) = -1 + x + 0.2(x^2 - 1)$, can be found in Figure 7 (iii). According to Table 4, we have a true marginal impact of $\frac{\partial}{\partial x} g(x) = 1 + 0.4x$. 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.[34]

Despite of the inconspicuous results of the usual validity checks, the p-value of the validity test proposed here amounts to 0.0210. Hence, the linear regression model can be rejected 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

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

However, in real life, we cannot expect that the given quantities are related to one another in that precise manner.[35] Moreover, both the capital input and the labor input can be considered stochastic, which means that the output of the economy is stochastic, too.

[34] However, there is *no* endogeneity at all, i.e., X is in fact exogenous.

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

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:	$\mathrm{E}(\varepsilon) = \mathrm{Cov}(K,\varepsilon) = \mathrm{Cov}(L,\varepsilon) = 0$, $\mathrm{E}(\varepsilon \mid K, L) = c(KL - \rho)$
Measures:	$A^2 = \frac{b_1^2 + b_2^2 + 2b_1 b_2 \rho}{\left[c^2(1+\rho^2)+1\right]\left[b_1^2 + b_2^2 + 2b_1 b_2 \rho + c^2(1+\rho^2) + 1\right]}$
	$R^2 = \frac{b_1^2 + b_2^2 + 2b_1 b_2 \rho}{b_1^2 + b_2^2 + 2b_1 b_2 \rho + c^2(1+\rho^2) + 1}$,
	$S^2 = \frac{b_1^2 + b_2^2 + 2b_1 b_2 \rho + c^2(1+\rho^2)}{b_1^2 + b_2^2 + 2b_1 b_2 \rho + c^2(1+\rho^2) + 1}$, $V^2 = \frac{1}{c^2(1+\rho^2)+1}$

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

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

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

with $a = \log b_0$ and

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

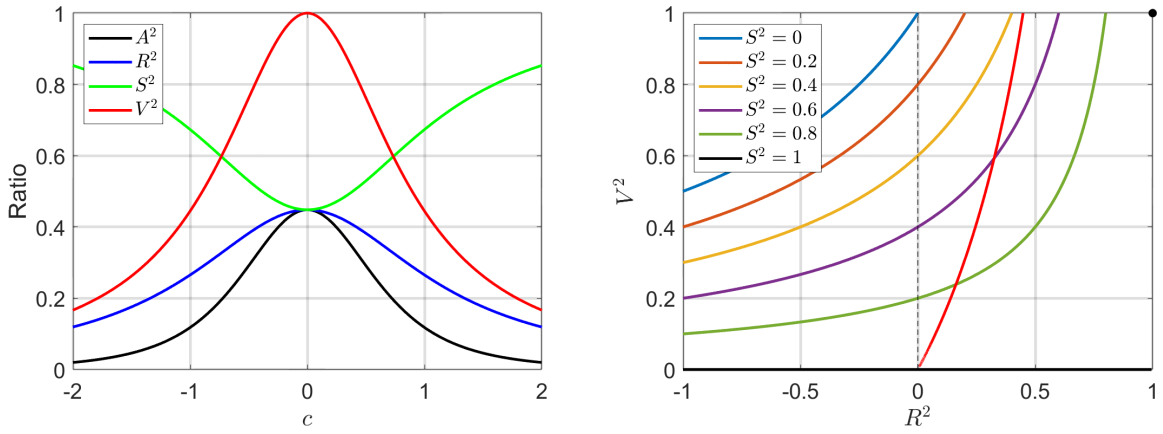
where $c \in \mathbb{R}$ and $-1 < \rho < 1$. Thus, log-capital input and log-labor input are correlated for $\rho \neq 0$. Furthermore, their impact on the log-output of the economy is not linear if $c \neq 0$. In that case, there is a synergy of capital and labor, which is quantified by the parameter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

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. $\mathrm{E}(\varepsilon) = -c\rho + c\mathrm{E}(KL) + \mathrm{E}(\epsilon) = -c\rho + c\rho = 0$,



(i) Regression measures

(ii) Prediction power vs. validity

Figure 8: Cobb-Douglas regression with $b_1 = 0.25$, $b_2 = 0.75$, and $\rho = 0.5$.

2. $\mathrm{Cov}(K, \varepsilon) = \mathrm{E}(K\varepsilon) = c\mathrm{E}(K^2L) + \mathrm{E}(K\epsilon) = 0$, 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 endogeneity—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. Despite all this, the linear regression model is still invalid if $c \neq 0$. Table 5 contains the fact sheet of the Cobb-Douglas regression equation.[36]

The mean conditional error amounts to $\mathrm{E}(\varepsilon \mid K, L) = c(KL - \rho)$. Suppose 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 $KL < \rho$, whereas 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, let us assume 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 8 (i) contains the corresponding regression measures. As it can be seen, even for higher absolute values of 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.

Now, consider a Monte Carlo simulation of $n = 100$ independent observations of (K, L, Y) with $a = 0$, $b_1 = 0.25$, $b_2 = 0.75$, $\rho = 0.5$, and $c = 0.25$, in which case the linear regression model is invalid. The resulting OLS estimates of $\alpha = 0.125$, $\beta_1 = 0.25$, and $\beta_2 = 0.75$ are $\hat{\alpha} = 0.1314$, $\hat{\beta}_1 = 0.3151$, and $\hat{\beta}_2 = 0.9081$, respectively, where the corresponding standard errors are 0.1026,

[36] The given regression measures can be calculated by applying Isserlis' theorem.

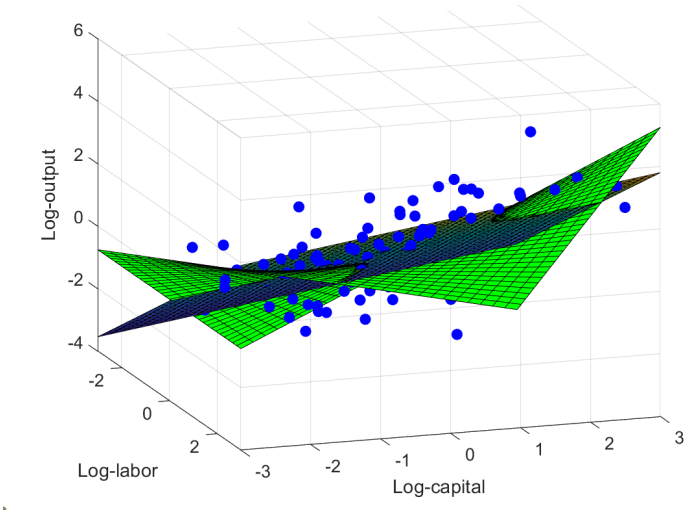


Figure 9: Scatter plot for the Cobb-Douglas regression, where the plane represents the fitted linear regression function and the bent surface illustrates the true regression function.

0.1157, and 0.1301. Further, the true coefficient of determination is 0.4298, whereas the ordinary R^2, obtained by the OLS estimates, even amounts to 0.5522. Once again, the F-test leads us to a p-value of virtually zero. The scatter plot in Figure 9 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.

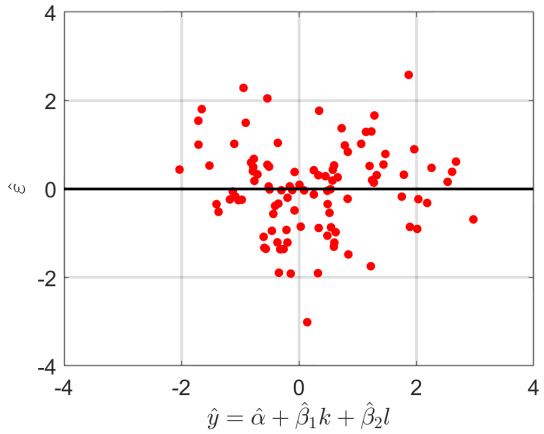


Figure 10: Residual plot for the Cobb-Douglas regression.

Figure 10 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 given on the y-axis.[37] The quantitative results look fine and also a visual inspection of the data does not reveal any anomaly. However, 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 9, which represents the true Cobb-Douglas regression

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

function $(k, l) \mapsto 0.25k + 0.75l + 0.5kl$. Now, how does the validity test perform in that situation, where we now apply a *multiple* regression? We obtain a p-value of 0.0170. Thus, again we can clearly reject the linear regression model on a significance level of 5%.

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 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 $n \in \{50, 100, 250, 500, 1000\}$.

The number of Monte Carlo repetitions in every setting amounts to R and each repetition $r \in \{1, \ldots, R\}$ creates n independent observations of (X, Y). I have chosen the uniform distribution on $[0, 1]$ for W, i.e., the weight function $\psi \equiv 1$, in order to demonstrate the default setting of the test. The regression model is rejected if the p-value in Repetition r falls below the level of 5%.

Table 6 contains the results of the simulation study with $R = 1000$ repetitions, 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 and n. On the left-hand side one can find also the accuracy, A^2, 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 size of the validity test is always about 5% and thus it satisfies the nominal significance level. Moreover, its power is quite strong even for $n = 50$. This holds true also for the Cobb-Douglas regression equation, which involves two regressors. In the case of $Y = \pm c + \epsilon$, which is depicted in the first panel of Table 6, we can see that the validity, V^2, decreases with the prediction power, R^2, of the linear regression model. This counterintuitive relationship between

	Measure				Sample size				
c	A^2	R^2	S^2	V^2	50	100	250	500	1000
Piecewise constant regression equation ($Y = \pm c + \epsilon$)									
0	0.0000	0.0000	0.0000	1.0000	0.0430	0.0540	0.0440	0.0540	0.0480
0.25	0.0366	0.0374	0.0588	0.9778	0.0990	0.1550	0.3480	0.6770	0.9400
0.50	0.1167	0.1273	0.2000	0.9167	0.2680	0.5390	0.9330	0.9990	1.0000
0.75	0.1903	0.2292	0.3600	0.8303	0.5520	0.8930	1.0000	1.0000	1.0000
1	0.2335	0.3183	0.5000	0.7335	0.8060	0.9950	1.0000	1.0000	1.0000
Quadratic regression equation ($Y = a + bX + c(X^2 - 1) + \epsilon$)									
−0.40	0.3265	0.4310	0.5690	0.7576	0.6860	0.9680	1.0000	1.0000	1.0000
−0.20	0.4452	0.4808	0.5192	0.9259	0.2510	0.4650	0.8870	0.9960	1.0000
0	0.5000	0.5000	0.5000	1.0000	0.0400	0.0520	0.0560	0.0630	0.0470
0.20	0.4452	0.4808	0.5192	0.9259	0.2410	0.4540	0.8800	0.9960	1.0000
0.40	0.3265	0.4310	0.5690	0.7576	0.7080	0.9490	1.0000	1.0000	1.0000
Cobb-Douglas regression equation ($Y = a + b_1K + b_2L + cKL + \epsilon$)									
−0.50	0.2913	0.3824	0.5294	0.7619	0.5130	0.8520	0.9980	1.0000	1.0000
−0.25	0.3986	0.4298	0.4711	0.9275	0.1870	0.3520	0.7110	0.9670	1.0000
0	0.4483	0.4483	0.4483	1.0000	0.0520	0.0520	0.0440	0.0550	0.0490
0.25	0.3986	0.4298	0.4711	0.9275	0.1850	0.3260	0.7050	0.9590	1.0000
0.50	0.2913	0.3824	0.5294	0.7619	0.5070	0.8160	0.9980	1.0000	1.0000

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

V^2 and R^2 has already been discussed in Section 3.2.1. However, the validity test is not affected by R^2. That is, it reliably indicates the invalidity of the regression model even if R^2 is high.

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 in small samples, given that the variance of ϵ is less than 1. For example, suppose we have only 5 years of quarterly data, i.e., $n = 20$, and that the true regression equation is $Y = g(X) + \epsilon$ with $\text{Var}(\epsilon) = \tau^2$ for $0 \leq \tau \leq 1$. 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 of a Monte Carlo study based on $R = 10000$ repetitions for each combination of c and τ. Although the sample size, n, is very small, the rejection rates are satisfactory even for relatively high levels of τ. In particular, despite of the small sample size, the validity test fairly satisfies its nominal significance level of 5%.

	τ										
c	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Piecewise constant regression equation ($Y = \pm c + \epsilon$)											
0	0.0000	0.0461	0.0546	0.0474	0.0483	0.0443	0.0509	0.0538	0.0490	0.0484	0.0500
0.25	0.9943	0.8734	0.4365	0.2253	0.1406	0.1075	0.0893	0.0797	0.0727	0.0710	0.0605
0.50	0.9944	0.9808	0.8724	0.6408	0.4361	0.3082	0.2272	0.1782	0.1501	0.1246	0.1125
0.75	0.9951	0.9916	0.9626	0.8777	0.7171	0.5645	0.4362	0.3384	0.2792	0.2196	0.1991
1	0.9949	0.9929	0.9814	0.9495	0.8716	0.7544	0.6500	0.5333	0.4361	0.3600	0.3054
Quadratic regression equation ($Y = a + bX + c(X^2 - 1) + \epsilon$)											
−0.40	0.9997	0.9987	0.9786	0.9151	0.8033	0.6902	0.5675	0.4745	0.4062	0.3406	0.2843
−0.20	0.9999	0.9782	0.8009	0.5721	0.4008	0.2868	0.2221	0.1751	0.1460	0.1205	0.1141
0	0.0000	0.0468	0.0460	0.0501	0.0504	0.0499	0.0498	0.0538	0.0492	0.0523	0.0506
0.20	1.0000	0.9782	0.8054	0.5675	0.3988	0.2862	0.2219	0.1812	0.1457	0.1256	0.1130
0.40	0.9998	0.9990	0.9809	0.9190	0.7976	0.6861	0.5720	0.4783	0.3927	0.3331	0.2857
Cobb-Douglas regression equation ($Y = a + b_1K + b_2L + cKL + \epsilon$)											
−0.50	0.8627	0.8498	0.7810	0.6923	0.5987	0.5036	0.4184	0.3617	0.3089	0.2618	0.2274
−0.25	0.8612	0.7768	0.5912	0.4198	0.2971	0.2247	0.1865	0.1553	0.1319	0.1226	0.1020
0	0.0000	0.0568	0.0606	0.0639	0.0574	0.0637	0.0643	0.0593	0.0604	0.0631	0.0560
0.25	0.8670	0.7799	0.5860	0.4125	0.2960	0.2281	0.1831	0.1503	0.1367	0.1169	0.1045
0.50	0.8632	0.8476	0.7838	0.6912	0.5852	0.4961	0.4129	0.3607	0.3017	0.2639	0.2232

Table 7: Rejection rates for $n = 20$.

4. Other Specification Tests

The validity test presented here shall be distinguished from other specification tests that can be found in the literature. For this purpose, let $X = (X_1, \ldots, X_m)$ be some vector of regressors and Y be the dependent variable. We call the regression model $Y = f(X) + \varepsilon$ well specified if and only if the joint distribution of X and ε satisfies some specific condition A.[38] 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 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 $B \supset A$, where "$B \supset A$" means that B is implied by A.[39] For example, according to Figure 4, we can test for validity by testing for optimality or for exogeneity, etc. Nonetheless, we can expect that a genuine test for B does not perform as well as a genuine test for A if our aim is to reject 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 the null hypothesis B will not reject the null hypothesis A although A 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 actually 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

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

[39] Put another way, B is a necessary condition 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 in fact suffer from that particular problem, i.e., they do not represent genuine tests for validity.

4.1. Linear-Regression Tests

Theorem 8 tells us that optimality is a necessary condition for validity. Thus, we can test for the optimality of f (Hypothesis B) in order to test for its validity (Hypothesis A). However, a genuine test for optimality is not a genuine test for validity. For example, let $Y = a + b'X + \epsilon$ be a linear regression model where a and b are such that the typical exogeneity conditions are satisfied. 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 6, $\hat{f}(X) = a + b'X$ 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 $\hat{f}(X) = a + b'X$.[40] 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 the regressor vector 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 set $\mathcal{S} = \{X_1, \ldots, X_m\}$ of regressors.[41]

None of these specification tests refer to validity, since the regression model $Y = \alpha + \beta'X + \varepsilon$ can very well 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 drawback is that these specification tests are restricted to linear regression models.

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

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

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 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 represents an artificial regression (MacKinnon, 1992). To sum up, 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 $k > 1$. Once again, these kind of specification tests refer to the optimality of $Y = \alpha + \beta X + \varepsilon$, not to 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 $\mathcal{F}$ is inadequate. 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 is another estimator $\hat{\theta}_1$ for θ that is consistent both under H_0 and under some particular alternative hypothesis H_1, for which reason it is asymptotically inefficient under H_0. Thus, if H_0 is true, we have

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

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.[42] By contrast, if H_1 is true, the test statistic should exceed a critical threshold, 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 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 the Hausman test for exogeneity.

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.[43] It is implicitly assumed that the covariance matrices $\mathrm{Cov}(Z, X)$ and $\mathrm{Var}(X)$ are regular. The given specification implies that $\mathrm{Cov}(Z, \varepsilon) = 0$, i.e., the instrumental variables are exogenous just by construction. 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 some regressor is endogenous 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 8, exogeneity is equivalent to optimality if we focus on linear regression. Nonetheless, as we have already seen above, a linear regression model can 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.

[42] 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 $\mathrm{rk}\, V$.

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

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 rely on the Gaussian linear model, i.e., they presume that the Gauss-Markov assumption is satisfied, and that ε has a normal distribution conditional on $\mathbf{X}$. 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 column 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 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 Gaussian linear model is satisfied. Hence, this model can be rejected if ψ exceeds some critical threshold.

That test differs in several aspects from the validity test presented in this work:

1. One must specify some leading regressor in order to apply the test in the case of $m > 1$.
2. The Harvey-Collier test is designed to falsify the Gaussian linear model, 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 conclude that the Harvey-Collier test is not a genuine validity test.

4.5. Utts' Rainbow Test

The rainbow test developed by Utts (1982) appears to be similar to the validity test developed here. However, once again its basic assumption is that the sample observations $Y_1, \ldots, Y_n$ of Y with $n > n_S > m+1$ obey the Gaussian linear model. It consists of two OLS regressions: The first one is made by using the entire sample with size n, while the second one is based on a subsample with size n_S, 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 Gaussian linear model is satisfied.

Despite of the similarities, the rainbow test has little to do with the validity test presented here—except for the very fact that subsamples are also created to apply the test. On the one hand, this test is based on the very strong assumption that the Gaussian linear model is satisfied. On the other hand, it is 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 $\mathrm{Var}(\varepsilon \mid X) = \sigma^2 > 0$. We conclude that the rainbow test is a test for the null hypothesis that the Gaussian linear model is satisfied, but this is much stronger than validity. Hence, it is not a genuine validity test, either.

4.6. Stute's Cusum Test

This test is developed by Stute (1997) and further studied by Stute et al. (1998a,b). He proposes a cusum approach for testing the validity of parametric regression functions. His test is based on the empirical process

$$R(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{1}_{X_i \leq x} \hat{\varepsilon}_i$$

for all $x \in \mathbb{R}^m$, which is marked by the (approximate) errors $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n$ with $\hat{\varepsilon}_i = Y_i - f(X_i, \hat{\theta})$ for $i = 1, \ldots, n$. Two metrics are considered for testing the null hypothesis that $f(\cdot, \theta)$ is valid:

1. The Kolmogorov-Smirnov (KS) statistic $D = \max_{x \in S} |R(x)|$ with $S = \{X_1, \ldots, X_n\}$ and
2. the Cramér-von-Mises (CvM) statistic

$$W^2 = \frac{1}{n} \sum_{i=1}^{n} R^2(X_i).$$

Stute et al. (1998a) report that the KS statistic and the CvM statistic are comparably well in the univariate case, whereas the CvM statistic seems to be favorable in the multivariate case.

A closer look reveals that W^2 is similar to T, i.e., the test statistic developed in this work: Let $\{\hat{\varepsilon}_{i,1}, \ldots, \hat{\varepsilon}_{i,k_i}\}$ be the set of all errors associated with the sample observations of X that are less than or equal to X_i for $i = 1, \ldots, n$.[44] This leads us to $R^2(X_i) = \frac{1}{n}\big(\sum_{j=1}^{k_i} \hat{\varepsilon}_{i,j}\big)^2$ and thus

$$W^2 = \frac{1}{n^2} \sum_{i=1}^{n} \left(\sum_{j=1}^{k_i} \hat{\varepsilon}_{i,j} \right)^2.$$

In particular, in the univariate case, i.e., $m = 1$, we simply have

$$W^2 = \frac{1}{n^2} \sum_{i=1}^{n} \left(\sum_{j=1}^{i} \hat{\varepsilon}_j^* \right)^2,$$

where $\hat{\varepsilon}_1^*, \ldots, \hat{\varepsilon}_n^*$ are the errors associated with the sample observations of X after they have been sorted in ascending order. Stute et al. (1998a) consider different bootstrap techniques in order to

[44] Here, "less than or equal to" is understood in the usual, i.e., componentwise, sense of matrix algebra.

approximate the distribution of their test statistics under the null hypothesis of validity. One of these techniques corresponds to the residual bootstrap explained in Section 3.1.2.

The essential difference between T and W^2 is threefold:

1. T is based on the squared sums of residuals that can be found in some neighborhood of each observation of X, whereas W^2 considers the squared sums of all residuals that are associated with the sample observations of X that are less than or equal to themselves.

2. When computing T, the squared sums of residuals are divided by $k \leq n$, i.e., the number of residuals in each neighborhood, whereas they are divided by n when computing W^2.

3. T makes use of n^2 sums of residuals, whereas W^2 involves only n sums of residuals.

In contrast to most other specification tests in the literature, Stute's test is in fact a genuine test for the validity of regression models. However, it is not the only validity test that is based on cumulative sums of regression errors. See González-Manteiga and Crujeiras (2013) for a broad overview.[45] These tests seem to be largely overlooked in the empirical literature.

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 can be applied to all types of regression models with any 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 used whenever the main goal of regression is description rather than prediction.

Acknowledgements

I thank Alexander Jonen very much for his valuable suggestions and helpful comments, which essentially improved the manuscript. The same goes for Florian Schütze, who has also taken on the programming of a publicly available R code for the validity test. Further, I thank Christian Glöer and André Küster Simic for our very stimulating discussions on this topic.

[45] Further, Lin et al. (2002) investigate several diagnostic tools based on cumulative residuals.

Proofs

Proof of Proposition 2

We have

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

with

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

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

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

(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

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

(iv) Since $\hat{f}$ is valid, we have $\hat{f} = g$. 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 3

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

(ii) Suppose $\mathrm{E}(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon \mid X) = \mathrm{Var}(\varepsilon)$, which implies that $\mathrm{Var}(\varepsilon \mid X)$ is deterministic. Thus, we have

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

and from Theorem 5 (ii) it follows that f is valid.

(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

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

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 Gaussian linear model implies that $\mathbf{X}$ is strictly exogenous and thus also that f is valid.

Proof of Theorem 4

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 2 (i) and Corollary 2 (ii) guarantee that the exogeneity conditions given by System 4 are satisfied.

Proof of Theorem 5

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

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

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

Therefore, $\mathrm{Var}(\varepsilon) = \mathrm{E}(\mathrm{Var}(\varepsilon \mid X))$ 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 2 (i) and $\mathrm{Var}(\varepsilon) = \mathrm{E}(\mathrm{Var}(\varepsilon \mid X))$ by Theorem 2 (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

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

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

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

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 $\mathrm{E}(\hat{\varepsilon}^2) \leq \mathrm{E}(\tilde{\varepsilon}^2)$, 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 7

Suppose $\mathcal{F}$ is adequate and let $\hat{f}$ be the valid element of $\mathcal{F}$. Due to Theorem 2 (iv,v), $\hat{f}$ is the unique optimal regression function among $\mathcal{F}$ and Corollary 2 (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 $\mathcal{F}(X)$.

Proof of Theorem 9

Let us start with the second statement and choose the first residual, ε_1, without loss of generality. It holds that

$$\begin{aligned}
\mathrm{P}(\varepsilon_1 \leq e_1 \,|\, X_1 = x_1) &= \frac{\mathrm{P}(\varepsilon_1 \leq e_1, X_1 = x_1)}{\mathrm{P}(X_1 = x_1)} \\
&= \frac{\mathrm{P}(\varepsilon_1 \leq e_1, X_1 = x_1) \prod_{i=2}^n \mathrm{P}(X_i = x_i)}{\mathrm{P}(X_1 = x_1) \prod_{i=2}^n \mathrm{P}(X_i = x_i)} \\
&= \frac{\mathrm{P}(\varepsilon_1 \leq e_1, X_1 = x_1, \ldots, X_n = x_n)}{\mathrm{P}(X_1 = x_1, \ldots, X_n = x_n)} = \mathrm{P}(\varepsilon_1 \leq e_1 \,|\, \mathbf{X} = \mathbf{x})
\end{aligned}$$

for all $e_1 \in \mathbb{R}$. This holds true for each other residual and so we have

$$\mathrm{P}(\varepsilon_i \leq e_i \,|\, \mathbf{X} = \mathbf{x}) = \mathrm{P}(\varepsilon_i \leq e_i \,|\, X_i = x_i)$$

for all $e_i \in \mathbb{R}$ and $i = 1, \ldots, n$. Thus, it follows that

$$\begin{aligned}
\mathrm{P}(\boldsymbol{\varepsilon} \leq \boldsymbol{e} \,|\, \mathbf{X} = \mathbf{x}) &= \frac{\mathrm{P}(\varepsilon_1 \leq e_1, \ldots, \varepsilon_n \leq e_n, X_1 = x_1, \ldots, X_n = x_n)}{\mathrm{P}(X_1 = x_1, \ldots, X_n = x_n)} \\
&= \frac{\prod_{i=1}^n \mathrm{P}(\varepsilon_i \leq e_i, X_i = x_i)}{\prod_{i=1}^n \mathrm{P}(X_i = x_i)} = \prod_{i=1}^n \mathrm{P}(\varepsilon_i \leq e_i \,|\, X_i = x_i) \\
&= \prod_{i=1}^n \mathrm{P}(\varepsilon_i \leq e_i \,|\, \mathbf{X} = \mathbf{x})
\end{aligned}$$

for all $\boldsymbol{e} = (e_1, \ldots, e_n) \in \mathbb{R}^n$.

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.

Amemiya, T. (1980): "Selection of regressors," *International Economic Review* **21**, pp. 331–354.

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.

Desboulets, L. (2018): "A review on variable selection in regression analysis," *Econometrics* **6**, DOI: 10.3390/econometrics6040045.

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.

González-Manteiga, W., Crujeiras, R. (2013): "An updated review of goodness-of-fit tests for regression models," *Test* **22**, pp. 361–411.

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.

Lin, D., Wei, L., Ying, Z. (2002): "Model-checking techniques based on cumulative residuals," *Biometrics* **58**, pp. 1–12.

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.

Stute, W. (1997): "Nonparametric model checks for regression," *Annals of Statistics* **25**, pp. 613–641.

Stute, W., González-Manteiga, W., Presedo-Quindimil, M. (1998a): "Bootstrap approximations in model checks for regression," *Journal of the American Statistical Association* **93**, pp. 141–149.

Stute, W., Thies, S., Zhu, L.X. (1998b): "Model checks for regression: an innovation process approach," *Annals of Statistics* **26**, pp. 1916–1934.

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.