A Nonparametric Least-Squares Test for Checking a Polynomial Relationship

Irène Gijbels[1,2] and Valentin Rousson[2]

[1] Institute of Statistics, Catholic University of Louvain, Voie du Roman Pays 20, B-1348 Louvain-la-Neuve, Belgium

[2] Centre for Mathematics and its Applications, Australian National University, Canberra, ACT 0200, Australia

Abstract

In this paper the interest is in testing whether a regression function is a polynomial of a certain degree. One possible approach to this testing problem is to do a parametric polynomial fit and a nonparametric fit and to reject the null hypothesis of a polynomial function if the distance between the two fits if too large. Another approach consists of looking at the residuals from the parametric fit. In this paper we propose an entirely new approach to deal with the testing problem. When testing whether a regression function is a polynomial of degree smaller than or equal to p, the key idea is to first obtain a nonparametric local polynomial estimate of the p-th derivative of the unknown regression function, and then to proceed with a classical least-squares test for a general linear model for testing whether this derivative is constant. This is a quite appealing approach since it just relies on ordinary least-squares tests, and hence is simple to use. The performance of the method is illustrated via a simulation study.

Key words: Derivative estimation; Goodness-of-fit; Least-squares tests; Local polynomial fitting; Model checking; Parametric polynomial regression.

1 Introduction

It is often of interest to find out the relationship between two or more variables. A simple parametric relationship, such as a linear function or a polynomial function, is appealing since it is easy to interpret the influence of one variable on the other. In many situations though it is not clear that the assumption of such a simple parametric function is justified, and such an hypothesis should first be tested using observed information on the variables. Testing whether a regression relationship is polynomial of a certain degree is the objective of this paper.

Consider for simplicity two random variables X and Y, for which we would like to investigate the association via the regression function of Y given X denoted by

$$m(x) = E(Y|X = x).$$

In parametric regression it is assumed that the function $m(\cdot)$ belongs to a parametric family

$$\mathcal{M} = \{m(\cdot;\theta); \theta \in \Theta\},$$

parametrized by some $(p+1)$-dimensional regression parameter θ belonging to a subset Θ of $I\!R^{p+1}$. In this paper we focus on the class $\mathcal{M}_p$ of polynomials of degree less than or equal to p,

$$\mathcal{M}_p = \{m(x;\theta) = a_0 + a_1 x + \cdots + a_p x^p; a_0, a_1, \cdots, a_p \in I\!R\}.$$

Several approaches have been proposed to test the hypothesis

$$H_0 \ : \ m(\cdot) \in \mathcal{M} \qquad \text{versus} \qquad H_1 \ : \ m(\cdot) \notin \mathcal{M}. \tag{1.1}$$

Let $(X_1, Y_1), \cdots, (X_n, Y_n)$ be an i.i.d. sample from (X, Y). One possible approach to the goodness-of-fit testing problem (1.1) consists of tests based on order selection criteria as proposed by for example Eubank and Hart (1992). Stute (1997) suggested to test H_0 by looking at the integrated regression function, which results into studying the empirical process of the regressors marked by the residuals defined as

$$R_n(x) = n^{-1/2} \sum_{i=1}^{n} 1_{\{X_i \leq x\}} \left[Y_i - m(X_i)\right]$$

where $1_{\{X_i \leq x\}}$ takes the value if 1 if $X_i \leq x$ and 0 otherwise. Tests for problem (1.1) can then be based on

$$R_n^1(x) = n^{-1/2} \sum_{i=1}^{n} 1_{\{X_i \leq x\}} \left[Y_i - m(X_i; \widehat{\theta})\right],$$

using the residuals from a parametric fit $m(\cdot;\widehat{\theta})$ which should all be close to zero if the null hypothesis is true. Possible test statistics in this context are based on the distance $d(R_n^1, 0)$ of R_n^1 from the null function, with $d(\cdot,\cdot)$ some appropriate distance function. Examples of distances are the L_2-distance, the Kolmogorov-Smirnov distance (i.e. the supremum norm distance) and the Cramér-Von Mises distance, among others. Stute, González-Manteiga and Presedo-Quindimil (1998) use bootstrap methods to approximate the distribution of the various test statistics.

Another approach to test the null hypothesis in (1.1) is to compare a parametric fit $m(\cdot,\widehat{\theta})$ with a nonparametric estimator $\widehat{m}(\cdot)$ of $m(\cdot)$ and to reject the null hypothesis if the distance $d(m(\cdot;\widehat{\theta}), \widehat{m}(\cdot))$ is too large. See for example Härdle and Mammen (1993) and Stute and González-Manteiga (1996), among others.

Alcalá, Cristóbal and González-Manteiga (1999) followed this approach, using the L_2-distance, to test whether the regression function is polynomial:

$$H_0 \ : \ m(\cdot) \in \mathcal{M}_p \qquad \text{versus} \qquad H_1 \ : \ m(\cdot) \notin \mathcal{M}_p, \tag{1.2}$$

with p a fixed nonnegative integer.

In this paper we focus on this particular testing problem and propose a new appealing approach to tackle it. The key idea is in fact very simple. When $m(\cdot)$ belongs to $\mathcal{M}_p$, its p-th derivative function equals the constant function $p!a_p$. So testing whether $m(\cdot)$ is a polynomial of degree less than or equal to p is equivalent to testing whether it's p-th derivative function $m^{(p)}(\cdot)$ is a constant function or not. Therefore we first estimate $m^{(p)}(\cdot)$ nonparametrically using local polynomial fitting and then use classical least-squares tests for a general linear model to test whether this derivative is a constant function or not.

There are two important ingredients in this new approach: the nonparametric estimation of a derivative function, and the use of available classical least-squares tests for general linear models. The innovation of this approach lies in making use of the classical tests. We opt for estimation of the derivative function via local polynomial fitting since this provides us with a direct estimate of the derivative function, and many properties are available for it (such as for example bias and variance expressions, and an asymptotic normality result). See for example Fan and Gijbels (1996). Indirect estimation of the derivative function would consists of first estimating nonparametrically the regression function $m(\cdot)$ and then calculating the p-th derivative of the obtained estimate (assuming that it exists).

Since local polynomial fitting involves the choice of a smoothing parameter, this is also the case of the resulting testing procedure. The approach followed by Alcalá, Cristóbal and González-Manteiga (1999) also requires the choice of a smoothing parameter. We compare our approach with the latter one via a simulation study.

In the next section we explain in more detail the proposed testing procedure, and highlight its general validity. In Section 3 we provide a simulation study which illustrates the performance of the proposed testing procedure and compare it with the performance of the testing procedure considered by Alcalá, Cristóbal and González-Manteiga (1999).

2 The proposed method

We will introduce our testing approach under a fixed design regression model. So assume that data (x_i, Y_i) are generated from

$$Y_i = m(x_i) + \varepsilon_i, \tag{2.1}$$

where $m(\cdot)$ is the unknown regression function and the ε_i are independent and identically distributed errors with mean 0 and variance σ^2.

We consider the null hypothesis

$$H_0 \ : \ m(x) = a_0 + a_1 x \cdots + a_p x^p \quad \forall x,$$

where some of the parameters a_k, $k = 0, \cdots, p$, may possibly be equal to 0 (including a_p). The alternative hypothesis is that $m(\cdot)$ is any other function, either a polynomial of degree larger than p or not a polynomial. Under the null hypothesis, we have that $m^{(p)}(x) = p! a_p$ for each value of x, i.e. is a constant function.

We first estimate $m^{(p)}(x)$ for several values of x by using local polynomial fitting. For this purpose select a set of x values, denoted by $z_1, ..., z_m$, where the derivative function is to be estimated. The integer m depends on n such that $m(n) \to \infty$ and $n \to \infty$. At each point z_j we fit locally a polynomial of degree p, leading to the estimated coefficients

$$\widehat{\beta}_j = (\widehat{\beta}_{j0}, ..., \widehat{\beta}_{jp})^T = \left(X_j^T W_j X_j\right)^{-1} X_j^T W_j Y$$

where $Y = (Y_1, ..., Y_n)^T$, with the superscript T indicating that we take the transposed of the vector (or matrix), and

$$X_j = \begin{pmatrix} 1 & (x_1 - z_j) & \cdots & (x_1 - z_j)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & (x_n - z_j) & \cdots & (x_n - z_j)^p \end{pmatrix},$$

$W_j = \mathrm{Diag}_i\left(K\left((x_i - z_j)/h_j\right)/h_j\right)$, K is a kernel and h_j is a bandwidth (which might depend on j, but also can be taken the same for each j). It turns out that $p!\widehat{\beta}_{jp}$ is an unbiased estimate of $m^{(p)}(z_j)$ under H_0 (see e.g. Section 3.2 in Fan and Gijbels (1996)). Hence we have that, under the null hypothesis, $E[\widehat{\beta}_{jp}] = a_p$ for each j. The variance of $\widehat{\beta}_{jp}$ is given by the $(p+1) \times (p+1)$ element of the matrix

$$\mathrm{Var}\left(\widehat{\beta}_j\right) = \sigma^2 \left(X_j^T W_j X_j\right)^{-1} X_j^T W_j W_j X_j \left(X_j^T W_j X_j\right)^{-1},$$

while the covariance of $\widehat{\beta}_{j_1 p}$ and $\widehat{\beta}_{j_2 p}$ is given by the $(p+1) \times (p+1)$ element of the matrix

$$\mathrm{Cov}\left(\widehat{\beta}_{j_1}, \widehat{\beta}_{j_2}\right) = \sigma^2 \left(X_{j_1}^T W_{j_1} X_{j_1}\right)^{-1} X_{j_1}^T W_{j_1} W_{j_2} X_{j_2} \left(X_{j_2}^T W_{j_2} X_{j_2}\right)^{-1}.$$

Note that if $h_{j_1} = h_{j_2} = h$ and if the support of the Kernel K is $[-1, +1]$, then the covariance between $\widehat{\beta}_{j_1 p}$ and $\widehat{\beta}_{j_2 p}$ vanishes when the difference between z_{j_1} and z_{j_2} is larger than $2h$.

We now form the 'synthetic' data set $(z_j, \widehat{\beta}_{jp})$, $j = 1, ..., m$. Under the null hypothesis H_0, all the estimates $\widehat{\beta}_{jp}$ are unbiased estimates of the same constant a_p. This new 'synthetic' data set can hence be described as

$$\widehat{\beta}_{jp} = a_p + \delta_j,$$

where the $\delta_j = \widehat{\beta}_{jp} - E[\widehat{\beta}_{jp}]$ are random errors with mean 0 and with the same variance-covariance matrix as that of the $\widehat{\beta}_{jp}$. We shall denote this matrix by $\sigma^2 V$. Important to note here it that we are dealing with dependent errors for which we know the exact covariance structure (the only unknown factor is σ^2). Now, if we consider a polynomial model of degree q (where $q \geq 1$ might be chosen quite differently from p)

$$\widehat{\beta}_{jp} = \gamma_0 + \gamma_1 z_j + \gamma_2 z_j^2 + ... + \gamma_q z_j^q + \delta_j,$$

we would have under H_0 that $\gamma_0 = a_p$ and $\gamma_k = 0$ for $k = 1, ..., q$. Our method consists now in rejecting H_0 if we reject the hypothesis

$$K_0 \ : \ \gamma_1 = \gamma_2 = ... = \gamma_q = 0 \tag{2.2}$$

associated to the general linear model

$$B = Z\gamma + \delta$$

where $B = (\widehat{\beta}_{1p}, ..., \widehat{\beta}_{mp})^T$, $\gamma = (\gamma_0, \gamma_1, ..., \gamma_q)^T$, $\delta = (\delta_1, ..., \delta_m)^T$, and

$$Z = \begin{pmatrix} 1 & z_1 & \cdots & z_1^q \\ \vdots & \vdots & \ddots & \vdots \\ 1 & z_m & \cdots & z_m^q \end{pmatrix}.$$

The generalized least-squares theory provides us with an estimate for γ (see for example Section 3.6 in Seber (1977)), given by

$$\widehat{\gamma} = \left(Z^T V^{-1} Z\right)^{-1} Z^T V^{-1} B$$

while the sum of squared residuals is given by

$$\mathrm{SSR} = (B - Z\widehat{\gamma})^T V^{-1} (B - Z\widehat{\gamma}) .$$

Note that the estimates $\widehat{\beta}_{jp}$, $j = 1, \cdots, m$, are asymptotically normal distributed (see for example Tsybakov (1986), Fan, Heckman and Wand (1995) and Masry and Fan (1997), among others), and hence so are the δ_j. These quantities are even exactly normally distributed if the errors ε_i in model (2.1) are normally distributed. It is thus justified to use the classical test of Fisher to test the hypothesis K_0. Denote

$$\mathrm{SST} = \sum_{i=1}^{m} \sum_{j=1}^{m} v_{ij} (\widehat{\beta}_{ip} - \overline{\beta}_p)(\widehat{\beta}_{jp} - \overline{\beta}_p)$$

where

$$\overline{\beta}_p = \frac{\sum_{i=1}^{m} \sum_{j=1}^{m} v_{ij} \widehat{\beta}_{ip}}{\sum_{i=1}^{m} \sum_{j=1}^{m} v_{ij}}$$

and where v_{ij} designs the (i, j) element of the matrix V^{-1}, then the statistic

$$F_q = \frac{\mathrm{SST} - \mathrm{SSR}}{\mathrm{SSR}} \cdot \frac{m - q - 1}{q}$$

follows under the null hypothesis K_0 an F-distribution with q and $m - q - 1$ degrees of freedom. A large value of F_q indicates therefore that K_0, and hence H_0, should be rejected. Note that the testing problem (2.2) differs for different values of q and another statistic is used. We shall see in the next section that for our testing problem (1.2) a value of $q = 2$ or $q = 3$ may be recommended.

In case of normal errors in the regression model (2.1) this leads to exact tests, whereas in the case of general errors the resulting tests have approximate level.

This test is very simple since it involves the well known least-squares theory. The polynomial model that we use to fit the 'synthetic' data set represents one of the rare cases of a linear model with dependent errors where the covariance structure of the errors is known. So here we have a nice occasion to use the results of the generalized least-squares theory which has hitherto received few possibilities of concrete application in the statistical literature.

We finally mention that our method depends on several parameters which have to be chosen:

- the grid points $z_1, \cdots, z_m$ (the number of grid points as well as their locations have to be chosen);

- the bandwidth parameter h (assuming we take the same bandwidth parameter for each point z_j).

In particular, these parameters should be chosen such that the matrix V is invertible. For example, V is always singular when $m > n - p$. Smaller values of m may also lead to a non-invertible V, depending on h and on the position of the x_i's and z_j's.

While the proposed tests remain valid for any choice of these parameters which do not lead to such computational problems, the power of the tests will be influenced by the specific choices made. We will come back to this issue in the simulation section, where we also recommend some possible appropriate choices of the parameters.

Although we strictly speaking only derived the testing procedures under the fixed design model (2.1), they remain valid under a random design regression model, keeping in mind that all statements should then be read as conditionally upon the observed values $x_1, \cdots, x_n$ of the variable X. In the simulation study we will also illustrate the random design case.

3 Simulation study

We conducted a simulation study which consists of two parts. In a first part we investigate the power of our method. The objective of the second part is to compare our testing approach with the procedure discussed in Alcalá, Cristóbal and González-Manteiga (1999).

Throughout this section we work with a set of grid points $z_1, \cdots, z_m$ which are chosen to be equidistant in a subset $[a, b]$ of the domain of the regression function. One can of course also consider non-equidistant grid points, but we do not elaborate on this. As we shall see, the choice of this subset $[a, b]$ will have an influence on the power of the method. In particular, it is not recommended to take $[a, b]$ as the whole regression domain because of boundary effects (although the validity of the test is not affected by boundary effects).

In all our simulations we considered normal errors with mean zero and variance σ^2, and used the Epanechnikov kernel. Recall that with use of this kernel the 'observations' in the 'synthetic' data set are independent as soon as the z_j's are at least a distance $2h$ apart from each other.

We restricted our simulation study to the test statistics F_1, F_2 and F_3, mainly because this leaves enough choice. In addition taking $q > 3$ may lead to computational problems either caused by nearly-collinearity or by data-sparseness. For a test based on F_1, the null hypothesis was rejected at significance level $\alpha = 0.05$ when the observed value of the statistic was larger than the $(1-\alpha)^{\text{th}}$ quantile of a Fisher distribution with 1 and $m-2$ degrees of freedom. Similarly, for a test based on F_2 the null hypothesis was rejected at significance level $\alpha = 0.05$ when the observed value of the statistic was larger than the $(1-\alpha)^{\text{th}}$ quantile of a Fisher distribution with 2 and $m-3$ degrees of freedom. For the test statistic F_3 we refer to a Fisher distribution with degrees of freedom 3 and $m-4$.

3.1 Simulation study 1: Investigation of the power of the proposed tests

In order to investigate the performance of our tests, we simulated data from model (2.1) with polynomial regression function

$$m(x) = (x - 0.5)^r, \tag{3.1}$$

with $r = 0, 1, 2, 3$, where the true regression function was respectively a constant, linear, quadratic or cubic function defined on the interval $[0, 1]$. We tested the null hypothesis

$$H_0 \ : \ m(x) = a_0 + a_1 x + \cdots + a_p x^p \ , \qquad\qquad a_0, a_1, \cdots, a_p \in \mathbb{R},$$

i.e. that the regression function is a polynomial of degree less than or equal to p. We considered this testing problem for $p = 0, 1, 2$ and 3. A schematic overview of the situations investigated in this setup is given in Table 1 below.

Table 1. *The situations investigated in the setup of simulation study 1.*

		true regression function			
		constant $r=0$	linear $r=1$	quadratic $r=2$	cubic $r=3$
hypothesis testing	$p=0$	H_0 true	H_0 false	H_0 false	H_0 false
	$p=1$	H_0 true	H_0 true	H_0 false	H_0 false
	$p=2$	H_0 true	H_0 true	H_0 true	H_0 false
	$p=3$	H_0 true	H_0 true	H_0 true	H_0 true

We chose the x_i to be the data set $\{0.02, 0.05, 0.08, ..., 0.95, 0.98\}$, which size is $n = 33$, while the Y_i were generated according to model (3.1). We considered two different noise levels of the error random variable: $\sigma = 0.01$ and $\sigma = 0.05$.

The test statistics F_q $(q \geq 1)$ depend on the bandwidth parameter h and on the grid points z_j. In this simulation study we considered three different values for h, namely $h = 0.1, 0.2, 0.3$, and five different values for m, namely $m = 5, 10, 15, 20, 25$. Finally, we tried three kinds of intervals $[a, b]$, namely $[a, b] = [0, 1], [h/2, 1 - h/2], [h, 1 - h]$. In order to measure the boundary effects on the power of our method, we also considered the case $[a, b] = [0, 1]$ with extra data generated outside the boundaries (that is we added some x_i outside $[0, 1]$, and generated the corresponding Y_i from the true model investigated).

For each of these settings, we simulated 100 samples and reported each time the percentage of times that the null hypothesis was rejected among these 100 times. In all our simulations we did not encounter any computational problem for inverting the matrix V (m was always small enough).

When $r \leq p$ the null hypothesis is true and the percentage of times the null hypothesis was rejected indicates the actual size of the tests, and in fact reports on the validity of the tests. We found that this percentage was about 5% for each setting considered, and for each of the three tests F_1, F_2 and F_3, illustrating clearly the validity of the tests for all choices of the parameters.

When $r > p$, the null hypotheses is false and the percentage of times the null hypothesis was rejected refers to the power of the tests. As may be seen in Table 1, there are six situations to be considered here: testing the hypothesis of a constant relationship when the true relationship is linear, quadratic or cubic; testing the hypothesis of a linear relationship when the true relationship is quadratic or cubic; and testing the hypothesis that the regression relationship is quadratic when the true relationship is cubic. We report the obtained results graphically for the case $\sigma = 0.05$ and for all of the six situations. Figures 1-6 present our findings. In each figure, the solid line refers to the case where $[a, b] = [0, 1]$ with extra data generated outside $[0, 1]$. The dotted, short-dashed and long-dashed lines refer to the cases where $[a, b] = [0, 1]$, $[h/2, 1-h/2]$ and $[h, 1-h]$ respectively, without extra data outside $[0, 1]$.

As can be seen from these figures, our tests were found more powerful for large values of h and m if we do not take into account the boundary effects (see the solid lines). But because of boundary effects, our tests with $[a, b] = [0, 1]$ were found sometimes worse for large values of h and m than for small values (see the dotted lines). The choice $[a, b] = [h, 1-h]$ avoids boundary effects but reduces the information available. Moreover, our plots show sometimes a loss of power when m is larger than the number of data in $[a, b]$ (see the long-dashed lines). The compromise obtained by taking $[a, b] = [h/2, 1-h/2]$ did not really improve the performance of our method in none of the situations considered, and boundary effects were still present.

Note that the test which uses the statistic F_1 is not able to make a distinction between a quadratic relationship and a constant relationship, neither between a cubic relationship and a linear relationship (see Figures 4 and 5 respectively). This is not surprising for anyone who is used to the least-squares theory and who knows that that theory will indeed be unable to distinguish, for example, the quadratic function in (3.1) from a constant function. If we would have considered instead a quadratic function with minimum away from the center 0.5 of the interval $[0, 1]$ then the test based on F_1 would have shown higher power. In other words, the difficulties encountered with the power here are caused by the power difficulties experienced by the classical least-squares tests in similar situations.

We finally mention that when the variance of the errors was taken to be $\sigma = 0.01$, our tests were found logically more powerful.

As a conclusion to this section, we recommend the use of the tests defined by F_2 or F_3, rather than the test defined by F_1. Further, it seems natural to choose a bandwidth h which fits reasonably well the regression curve. This will have the advantage to provide the user with a graphical display of what the true curve looks like when the null hypothesis is rejected. Then, if h is small enough, one may choose the domain $[a, b]$ of the grid points z_j to be h apart from the boundaries of the regression domain, along with m as large as (but not larger than) the number of data in this interval. Of course, if h is too large such that this interval $[a, b]$ does not contain enough information about the regression curve, one may reduce the bandwidth, or define $[a, b]$ to be a fraction of h apart from the boundaries (such as $h/2$). In this latter case, one should avoid taking a too large value for m, as can be seen from our simulations. A detailed theoretical study of the power properties of our tests might possibly indicate a (theoretical) optimal choice for the parameter m (as a function of h and for some kind of $[a, b]$). This will be subject of future research.

3.2 Simulation study 2: Comparison

We compared our method with that of Alcalá, Cristóbal and González-Manteiga (1999). In their paper they considered two regression models. The data were generated according to a quadratic model

$$Y_i = 5X_i + cX_i^2 + \varepsilon_i \tag{3.2}$$

and to a cubic model

$$Y_i = X_i + cX_i^3 + \varepsilon_i. \tag{3.3}$$

We will refer to these models as respectively Model 1 and Model 2.

The X_i were generated uniformly in $[0, 1]$ for the quadratic model (Model 1) and uniformly in $[-1, 1]$ for the cubic model (Model 2). Choices of $c = 0, 1, 5$ and $\sigma^2 = 1, 2, 3$ were investigated along with samples of size $n = 50, 100, 200$. The nonparametric fits in Alcalá, Cristóbal and González-Manteiga (1999) were calculated using a local-linear approach or a local quadratic approach with different bandwidths $h = 0.1, 0.25, 0.6$ (which all lead essentially to the same results). For each case they have computed the percentage of times that the null hypothesis was rejected out of 500 simulated samples. The level $\alpha = 0.05$ was used. Cases with $c = 0$ correspond to the null hypothesis and therefore the percentage of rejection should be approximately α. Cases with $c = 1$ and $c = 5$ correspond respectively to moderate and important departures from the null hypothesis. We report their results for the case $h = 0.25$ in the last two columns of Tables 2 and 3 (where LL and LQ stand for a local linear and a local quadratic fit respectively). It should be mentioned that Alcalá, Cristóbal and González-Manteiga (1999) incorporated a weight function $w(\cdot) = I_{[\varepsilon,1-\varepsilon]}(\cdot)$ for Model 1, and $w(\cdot) = I_{[-1+\varepsilon,1-\varepsilon]}(\cdot)$ for Model 2, with ε small (typically something like 0.1 or larger), and $I_J(\cdot)$ the indicator function on the interval J, into their distance function. Hence, they avoid boundary effects in this way.

We tried our tests defined by the statistics F_1, F_2 and F_3 for each of the cases considered in the last paragraph. For Model 1, we tried two different bandwidths, namely $h = 0.125$ and $h = 0.25$, and the z_j were taken equidistant in $[a, b] = [h/2, 1 - h/2]$ since the regression domain was here rather small. In order to avoid computational problems and to save computational time we took $m = 15$, although a much larger value could be chosen when $n = 100$ or $n = 200$. A pilot simulation study seemed to indicate though that we do not gain much power for larger values of m. For Model 2, we restricted our attention to the case $h = 0.25$ since a smaller bandwidth often led to computational problems. As the regression domain was here twice larger, we tried both the cases $[a, b] = [-1+h/2; 1-h/2]$ and $[a, b] = [-1+h; 1-h]$. The number of grid points was kept to $m = 15$. The percentage of times that our tests rejected the null hypothesis are given in Tables 2 and 3 for Models 1 and 2 respectively. One may observe that the results associated to the case $c = 0$ suggest that our tests are still approximately valid in the uniformly random design case.

As can be seen in Table 2, our tests were found less powerful than those of Alcalà, Cristóbal and González-Manteiga and (1999) in the quadratic case. This poor performance may be explained by the fact that the regression domain was not large enough. As a consequence, there were too many dependencies among the z_j's when $h = 0.25$, such that we had not enough information for detecting a departure from the null hypothesis. For $h = 0.125$ these dependencies were reduced and the power of our tests improved but it is to be noted that such a bandwidth was clearly too small to fit the data.

Table 2. *Percentage of times that the null hypothesis is rejected when the data are generated from Model 1.*

σ^2	c	n	F_1	F_2	F_3	F_1	F_2	F_3	LL	LQ
			$h = 0.125$, $[a,b] = [h/2, 1-h/2]$			$h = 0.25$, $[a,b] = [h/2, 1-h/2]$			$h = 0.25$	
1	0	50	5.4	5.6	4.4	5.4	5.6	5.2	4.8	4.4
		100	6.0	4.8	4.8	4.8	4.2	4.2	5.0	5.1
		200	5.4	4.8	6.0	5.6	5.8	5.4	4.9	5.2
	1	50	7.6	5.4	4.4	7.2	6.4	6.0	8.7	11.0
		100	10.8	9.0	9.2	10.8	9.0	9.0	18.1	20.3
		200	18.8	12.2	9.0	16.8	11.4	10.2	26.0	31.2
	5	50	62.4	49.6	41.6	47.8	37.8	32.6	76.1	82.5
		100	91.6	82.0	72.4	85.0	74.8	68.6	96.8	100.0
		200	99.6	98.8	96.2	99.6	97.0	96.0	100.0	100.0
2	0	50	5.4	4.4	4.6	5.2	5.4	5.4	4.5	4.1
		100	4.4	4.6	5.4	5.4	4.2	3.6	5.5	4.9
		200	6.0	4.6	5.0	7.0	3.8	4.2	5.8	5.2
	1	50	6.2	6.2	5.8	7.2	5.4	6.4	7.2	7.6
		100	9.8	7.2	7.6	8.4	7.2	6.4	12.5	14.6
		200	9.4	7.8	5.8	9.0	8.8	6.8	17.0	24.1
	5	50	41.4	30.8	23.0	33.4	25.6	18.8	53.6	75.2
		100	66.4	48.8	39.6	58.2	43.8	38.8	84.8	91.6
		200	90.6	82.8	74.6	89.8	76.6	67.6	96.7	100.0
3	0	50	5.0	5.8	5.4	4.8	6.2	5.6	4.5	4.4
		100	5.0	5.4	5.4	4.8	4.6	4.4	4.8	5.3
		200	6.2	4.4	4.8	6.2	4.6	5.4	5.1	4.8
	1	50	6.6	8.6	7.4	7.0	7.8	6.2	6.9	8.0
		100	8.4	5.6	8.2	8.0	6.0	5.6	8.4	9.6
		200	11.0	7.8	7.8	9.6	8.4	6.2	13.5	14.6
	5	50	24.2	18.8	16.2	23.2	16.8	15.6	35.2	41.2
		100	47.0	34.2	26.8	44.6	34.6	29.8	68.1	83.6
		200	78.8	64.8	54.8	74.6	60.2	51.2	85.1	94.2

Table 3. *Percentage of times that the null hypothesis is rejected when the data are generated from Model 2.*

σ^2	c	n	F_1	F_2	F_3	F_1	F_2	F_3	LL	LQ
			$h = 0.25$, $[a, b] = [-1 + h/2, 1 - h/2]$			$h = 0.25$, $[a, b] = [-1 + h, 1 - h]$			$h = 0.25$	
1	0	50	5.4	5.6	4.4	6.0	6.6	6.0	3.6	3.5
		100	6.0	4.8	4.8	5.4	5.4	4.8	4.6	3.9
		200	5.4	4.8	6.0	5.0	5.6	4.2	4.3	4.5
	1	50	5.2	9.8	7.8	5.0	9.4	7.6	5.2	4.8
		100	3.6	19.4	17.6	2.4	14.8	14.6	9.3	8.7
		200	3.4	39.2	28.8	2.8	32.8	24.2	19.0	20.1
	5	50	2.4	95.0	91.4	1.8	92.6	84.6	45.8	56.0
		100	0.2	100.0	100.0	0.0	99.8	99.6	65.8	74.1
		200	0.0	100.0	100.0	0.0	100.0	100.0	87.2	95.3
2	0	50	5.4	4.4	4.6	5.8	5.6	4.6	4.5	4.2
		100	4.4	4.6	5.4	6.2	5.2	4.8	4.8	4.2
		200	6.0	4.6	5.0	5.6	4.0	4.8	5.3	4.2
	1	50	6.0	8.2	6.8	5.4	9.6	8.0	5.7	5.0
		100	4.8	11.2	10.2	5.6	11.6	7.8	7.6	5.9
		200	2.8	19.6	14.8	3.2	17.2	15.2	10.0	9.5
	5	50	2.0	77.4	65.8	3.0	71.2	60.8	24.6	24.5
		100	0.8	96.8	93.8	0.2	92.4	86.6	47.6	49.6
		200	0.0	100.0	99.8	0.0	99.8	99.4	79.5	90.1
3	0	50	5.0	5.8	5.4	3.8	5.6	3.6	3.9	3.0
		100	5.0	5.4	5.4	5.8	3.8	4.6	4.0	3.9
		200	6.2	4.4	4.8	6.6	4.6	5.6	5.1	4.7
	1	50	6.0	9.0	8.2	7.0	10.2	7.4	4.8	4.5
		100	5.0	9.6	8.4	6.6	7.4	6.4	6.3	6.5
		200	5.8	12.4	11.2	5.4	11.6	9.8	9.6	10.2
	5	50	4.4	57.4	48.2	4.0	50.4	43.2	18.6	22.7
		100	1.6	89.2	80.4	1.4	85.0	74.6	41.1	48.1
		200	0.0	100.0	98.4	0.0	99.4	98.2	76.3	85.4

For the cubic model however, as can be seen in Table 3, our tests were found more powerful than those of Alcalá, Cristóbal and González-Manteiga (1999) (except the method defined by F_1 which is not able to distinguish a cubic from a linear relationship as we discussed before). Here the dependencies between the z_j's were reduced as a consequence of a larger regression domain. We conclude that our method is especially efficient when there is sufficiently 'independent' information among the z_j's, along with a bandwidth which fits reasonably well the regression curve.

Acknowledgments: The first author was supported by 'Projet d'Actions de Recherche Concertées', No. 98/03–217. The second author was supported by the Swiss National Science Foundation (Grant No 81NE-54413). Most part of this research was carried out when both authors were visiting the Centre for Mathematics and its Applications at the Australian National University, Canberra. The authors gratefully acknowledge the Centre, and in particular Professor Peter Hall, for its hospitality and the financial support.

References

Alcalá, J.T., Cristóbal, J.A. and González-Manteiga, W. (1999). Goodness-of-fit test for linear models based on local polynomials. *Statistics and Probability Letters*, **42**, 39–46.

Eubank, R. and Hart, J, (1992). Testing goodness-of-fit in regression via order selection criteria. *Annals of Statistics*, **20**, 1412–1425.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications.* Chapman and Hall, London.

Fan, J., Heckman, N.E. and Wand, M.P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of the American Statistical Association*, **90**, 141–150.

Härdle, W. and Mammen, E. (1993). Comparing noonparametric versus parametric regression fits. *Annals of Statistics*, **21**, 1926–1947.

Masry, E. and Fan, J. (1997). Local polynomial estimation of regression functions for mixing conditions. *Scandinavian Journal of Statistics*, **24**, 165–179.

Seber, G.A.F. (1977). *Linear Regression Analysis.* John Wiley and Sons, New York.

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

Stute, W. and González-Manteiga, W. (1996). NN goodness-of-fit test for linear models. *Journal of Statistical Planning and Inference*, **53**, 75–92.

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

Tsybakov, A.B. (1986). Robust reconstruction of functions by the local-approximation method. *Problems in Information Transmission*, **22**, 133–146.

Figure 1: Power of the tests when $r = 1$ and $p = 0$ for different values of h and m.

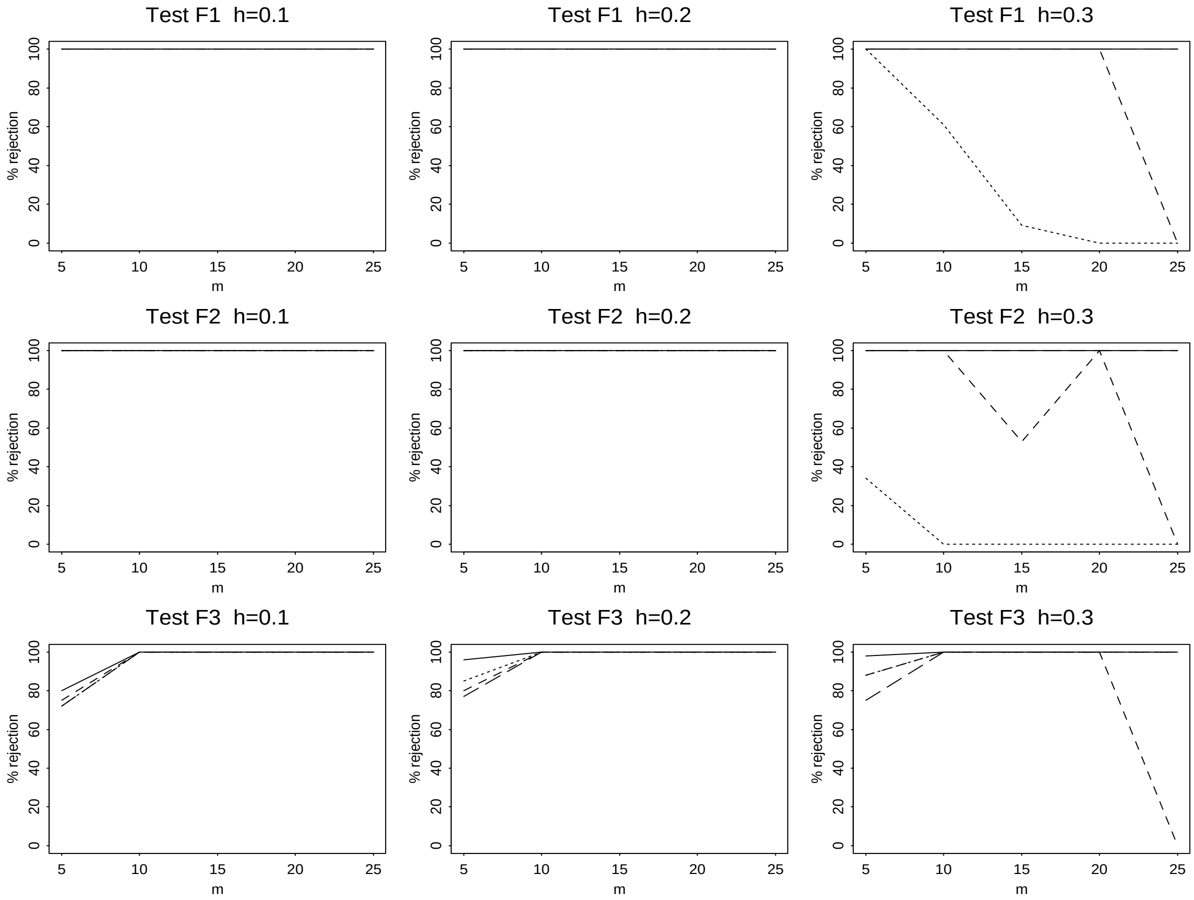

Figure 2: Power of the tests when $r = 2$ and $p = 1$ for different values of h and m.

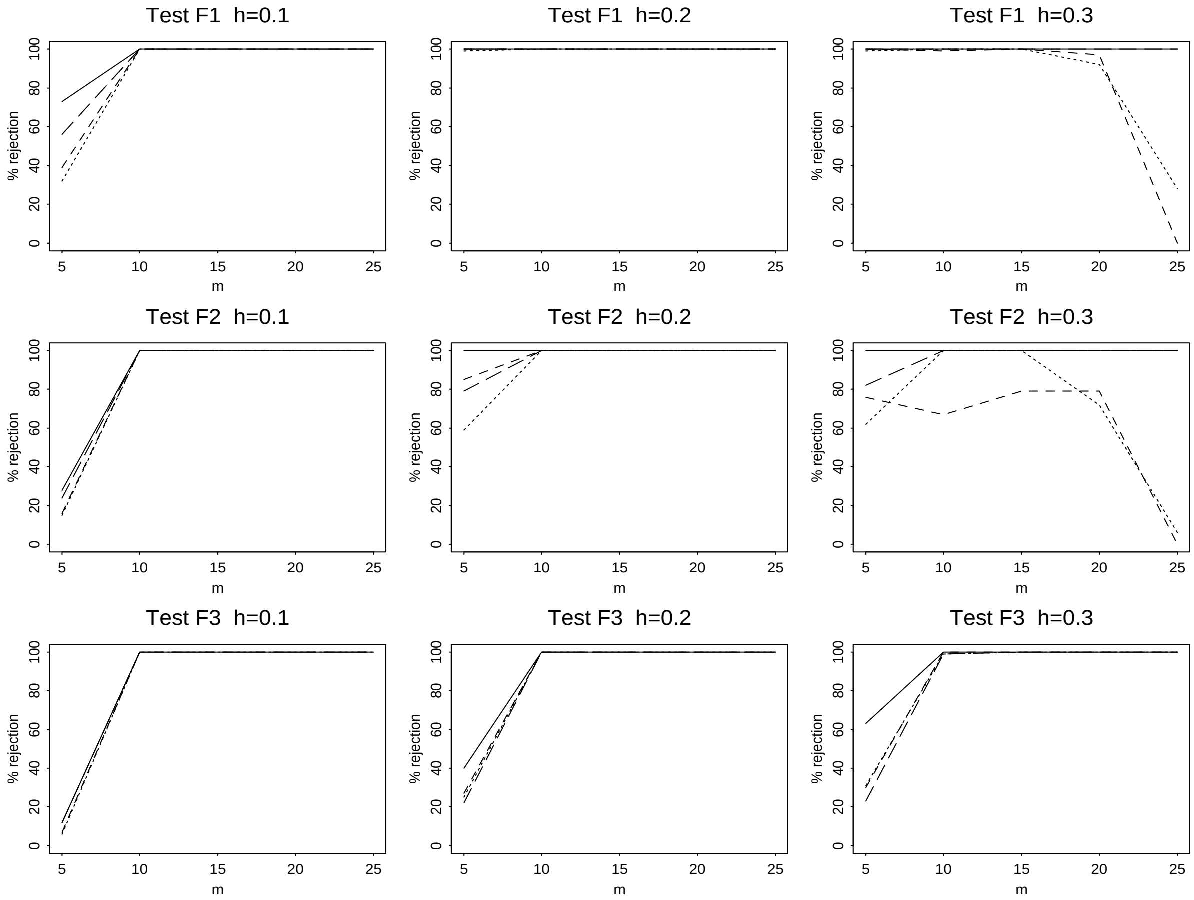

Figure 3: Power of the tests when $r = 3$ and $p = 2$ for different values of h and m.

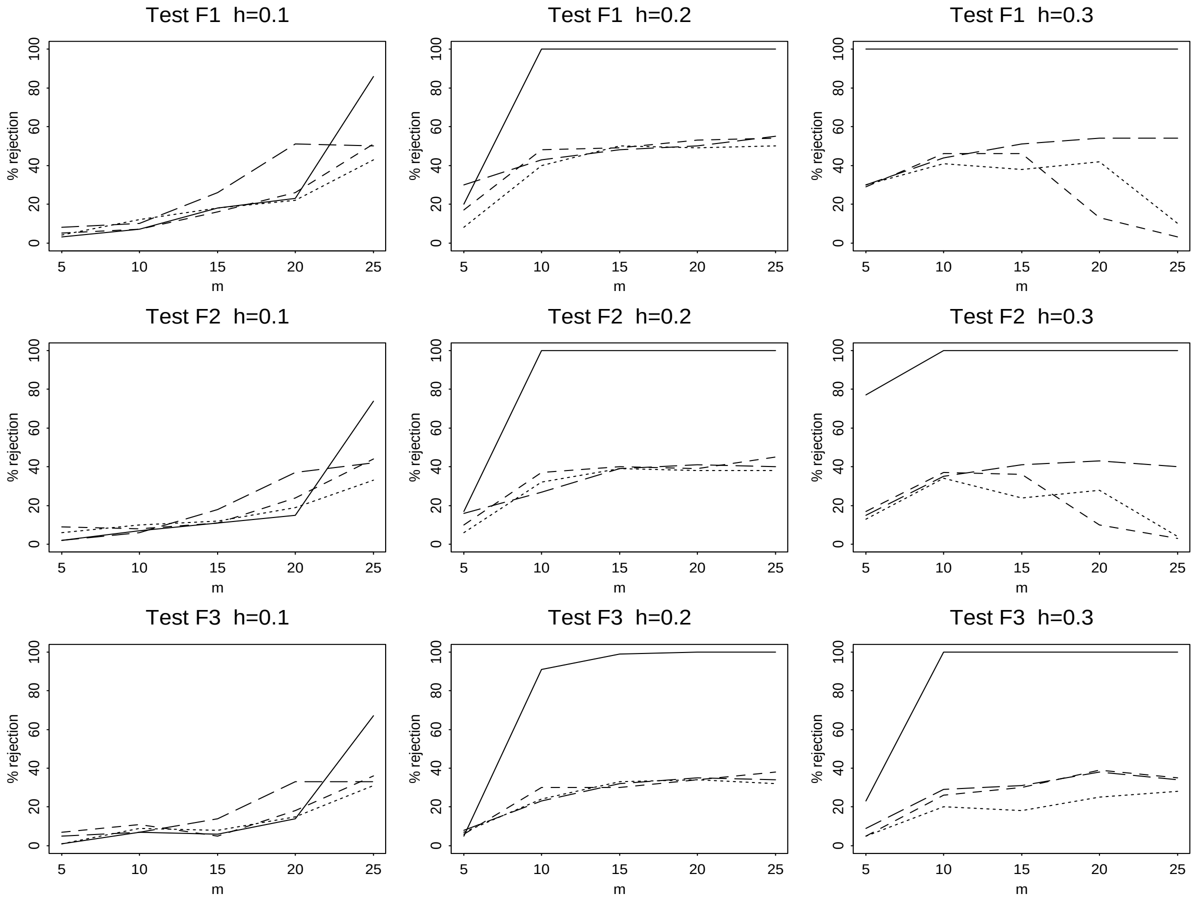

Figure 4: Power of the tests when $r = 2$ and $p = 0$ for different values of h and m.

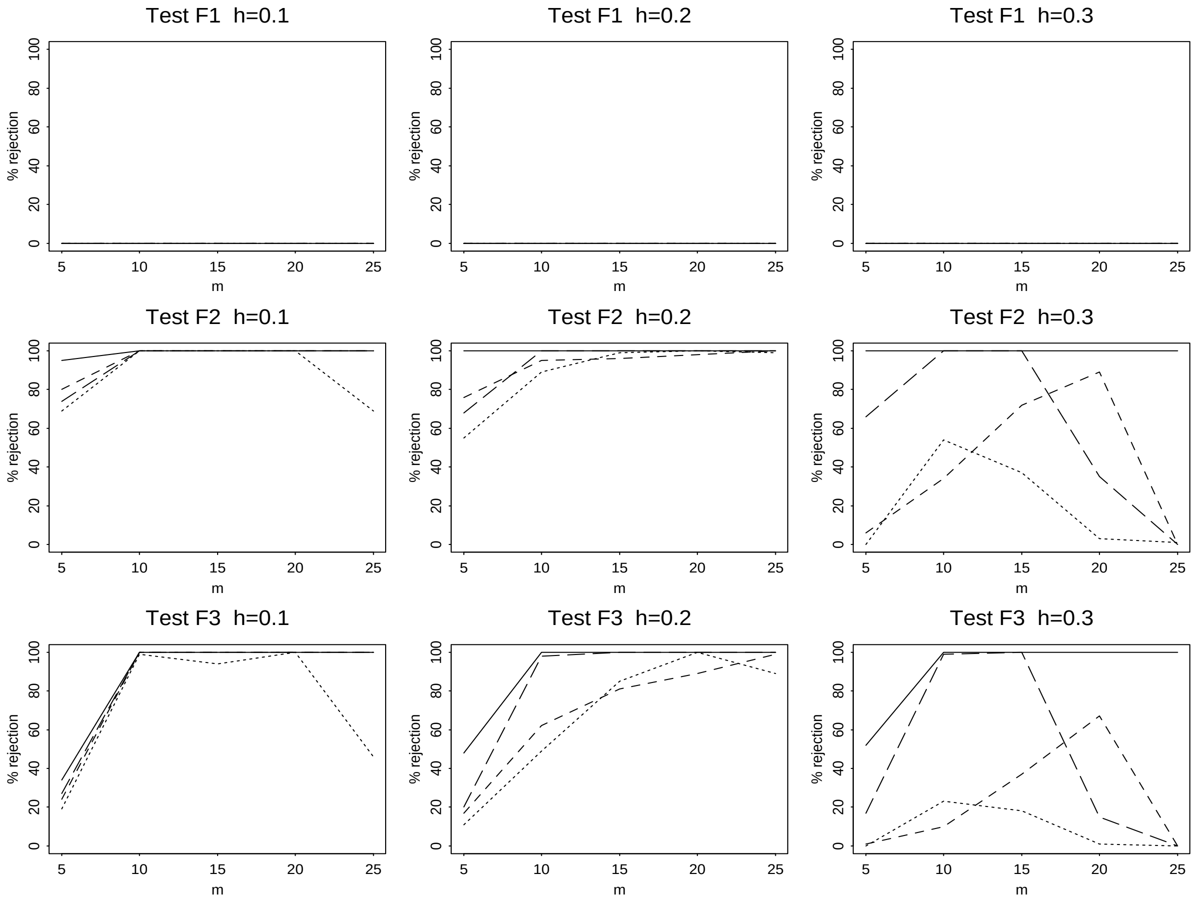

Figure 5: Power of the tests when $r = 3$ and $p = 1$ for different values of h and m.

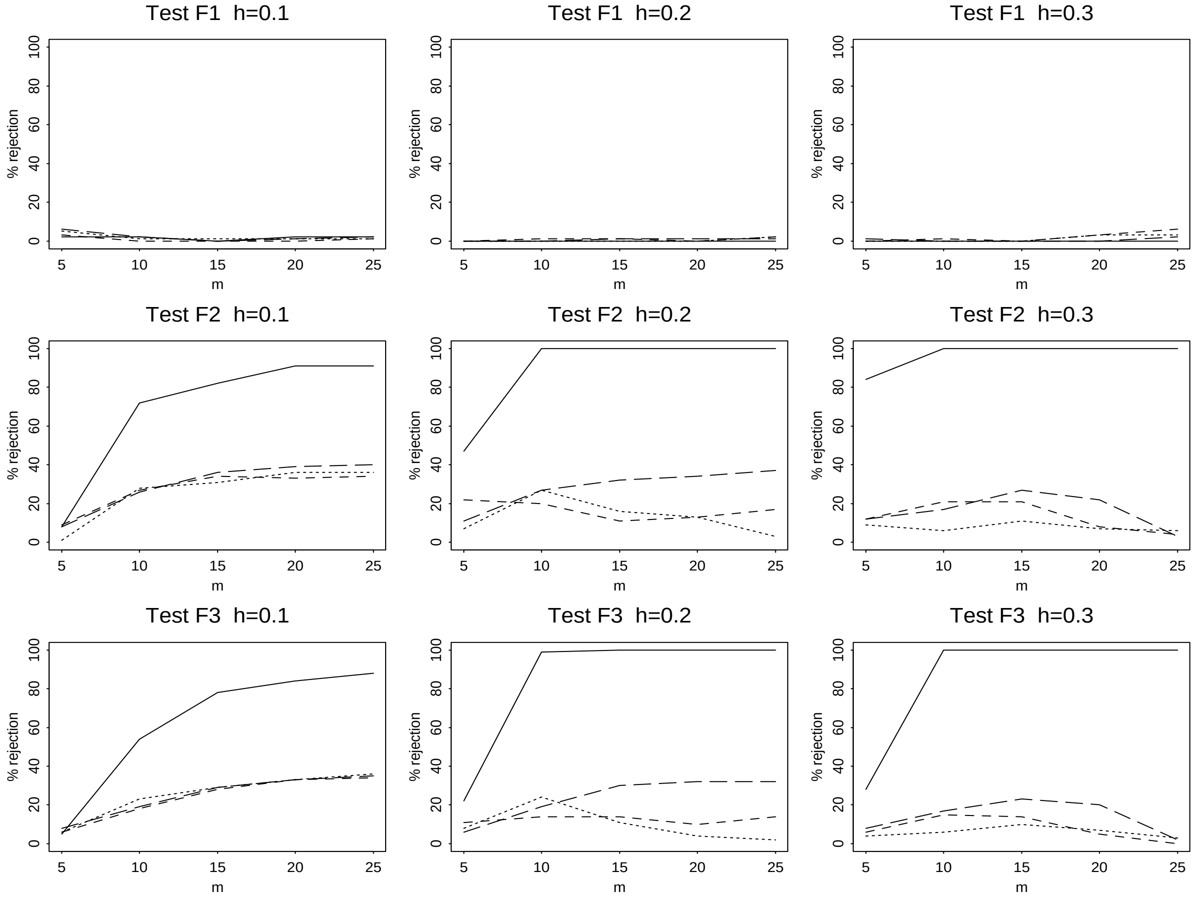

Figure 6: Power of the tests when $r = 3$ and $p = 0$ for different values of h and m.