KATHOLIEKE UNIVERSITEIT
LEUVEN

Faculty of Business and Economics

Order selection tests with multiply-imputed data

Fabrizio Consentino and Gerda Claeskens

DEPARTMENT OF DECISION SCIENCES AND INFORMATION MANAGEMENT (KBI)

KBI 0905

Order Selection Tests with Multiply-Imputed Data

Fabrizio Consentino and Gerda Claeskens

K.U. Leuven

ORSTAT and Leuven Statistics Research Center

Naamsestraat 69, 3000 Leuven, Belgium

Fabrizio.Consentino@econ.kuleuven.be; Gerda.Claeskens@econ.kuleuven.be

March 30, 2009

Abstract

We develop nonparametric tests for the null hypothesis that a function has a prescribed form, to apply to data sets with missing observations. Omnibus nonparametric tests do not need to specify a particular alternative parametric form, and have power against a large range of alternatives, the order selection tests that we study are one example. We extend such order selection tests to be applicable in the context of missing data. In particular, we consider likelihood-based order selection tests for multiply-imputed data. A simulation study and data analysis illustrate the performance of the tests. A model selection method in the style of Akaike's information criterion for multiply imputed datasets results along the same lines.

Keywords: Akaike information criterion, hypothesis test, multiple imputation, lack-of-fit test, missing data, omnibus test, order selection.

1 Introduction

Testing the lack-of-fit of a parametric function is well-studied. Several types of tests exist, ranging from fully parametric tests, to semiparametric and nonparametric omnibus tests. For an overview of nonparametric tests, see Hart (1997). In the setting of missing data, this

is more complicated and not much results are known yet. González-Manteiga and Pérez-González (2006) developed a test based on local linear estimators for a linear regression model with missing response values but a completely observed covariate. We address in particular lack-of-fit tests for missing data situations where multiple imputation is applied. We will focus on a class of smoothing-based tests, that use the idea of order selection. Our tests are applicable in parametric likelihood models and are not restricted to linear models.

Eubank and Hart (1992) introduced the order selection test in linear regression models. The idea is to test the shape of a parametric function, most often the mean of the response, by considering a sequence of alternative models. These alternative models are constructed by means of a series expansion of the function of interest around the hypothesized null model. A data-driven method is then applied to select the "order" of the alternative model. That is, in the sequence of alternative models, a method such as the AIC (Akaike, 1973) will select the most appropriate one. If the selected model coincides with the null model, the test does not reject the null hypothesis. However, if a model different from the null model is selected, the test will reject the null hypothesis. In those instances, the order of the chosen model, that is, the number of parameters in the model, exceeds that of the null model.

By using such a series expansion the class of alternative models is large, and not restricted to a single specified alternative. For example, just testing a linear versus a quadratic fit would miss out on high frequency alternatives for which the quadratic term happens to be zero. We are interested in the development of tests that are sensitive to essentially any departure from the null hypothesis.

The original order selection tests are extended towards testing in general likelihood models by Aerts et al. (1999) and to multiple regression models by Aerts et al. (2000). Recently, these tests have been studied for inverse regression problems by Bissantz et al. (2009). Test statistics can be based on likelihood ratio, Wald or score statistics. All this assumes completely observed data.

In practice, many data contain one or more missing observations. We refer to Little and Rubin (2002) for an overview of methods to deal with such data. Most research focusses on the estimation under missingness. Multiple imputation methods are particularly attractive since once values are imputed, traditional, complete-case methods can be applied. Single imputation replaces an unknown observation by a single value. While this is simple, inference is often improved when imputing values several times, say m times (usually about 5 times), creating m complete data sets. The main problem then arises in the combination of the results over the multiple imputed data. Li et al. (1991a) considered hypothesis testing in this setting. In particular, for a parametric null hypothesis of the form $\theta = \theta_0$, with an alternative of the form $\theta \neq \theta_0$, they construct a Wald test, by combining the results of m Wald tests, one for each of the m imputed data sets. They show that the distribution of such test can be approximated by that of an F-distribution with certain degrees of freedom. Meng and Rubin (1992) extend this idea to combining m likelihood ratio tests. Recently, Reiter (2007) obtained an alternative approximation to the degrees of freedom for such combined Wald test statistics, that should work better for small samples.

The main idea of this paper is to use the combined likelihood ratio tests for the m imputed data sets, in a construction for order selection. In this way we enlarge the testing power by not considering a single parametric test, since order selection tests are constructed to be powerful against a wide range of alternative models. This creates an easy to use lack-of-fit test in the setting of missing data.

Section 2 defines the order selection test first for complete data, and then proposes the new test for the case of multiply imputed data sets. Sections 3 and 4 apply the test to a data example and in a simulation study. A version of Akaike's information criterion that works with multiply imputed datasets is obtained in Section 5. Section 6 presents some extensions of the proposed method.

2 The order selection test

2.1 A model sequence for order selection

We consider a set of data $\boldsymbol{Z}_i = (Y_i, x_i), i = 1, \ldots, n$ with joint density depending on a function $\gamma(\cdot)$ of interest (most often this is the mean response, conditional on covariates) and on some other nuisance parameters $\boldsymbol{\eta}$ (such as an unknown variance). We wish to test the hypothesis

$$H_0 : \gamma(\cdot) \in \mathcal{G} = \{\gamma(\cdot, \boldsymbol{\beta}_p) : \boldsymbol{\beta}_p = (\beta_1, \ldots, \beta_p) \in \Theta\}, \tag{1}$$

where the parameter space $\Theta \subset \mathbb{R}^p$. A simple example is to test for linearity of the mean response, that is, $E(Y|x) = \gamma(x) = \beta_1 + \beta_2 x$. In a parametric hypothesis testing procedure, a specific parametric model would be stated for the alternative hypothesis. In nonparametric or omnibus testing, this is avoided by constructing a sequence of alternative models. These approximations could be quite general. For regression models, we mainly consider additive series expansions of the true underlying function $\gamma(\cdot)$ around the null model. We here mainly follow the approach of Aerts et al. (1999). In particular, we define for $r = 0, 1, 2, \ldots$,

$$\gamma(x; \beta_1, \ldots, \beta_{p+r}) = \gamma(x; \beta_1, \ldots, \beta_p) + \sum_{j=1}^{r} \beta_{p+j} \psi_j(x), \tag{2}$$

where the basis functions $\psi_j(\cdot)$ are known functions and $r = 0$ corresponds to the null model in (1). Most often these functions are taken to be (orthogonalized) polynomials, Legendre polynomials, cosine functions, wavelet functions,... For all further analysis, we consider functions ψ_j that are not of the form of the null model. For example, a polynomial expansion to test for linearity of the mean starts from (orthogonalized) quadratic functions, since the constant and linear function are already included in the null model.

The order selection test actively uses a model selection criterion to perform the test. For each $r = 0, 1, 2, \ldots, R_n$ a model with function $\gamma(\cdot; \beta_1, \ldots, \beta_{p+r})$ is fit to the data. This results in a sequence of $R_n + 1$ fitted models. A model selection criterion such as the AIC

(Akaike, 1973) is applied to select one of these models. If a model different from the null model is selected, in other words, when the selected order $\hat{r} > 0$, then the null hypothesis (1) is rejected. When the selected order $\hat{r} = 0$, the null model cannot be rejected.

Asymptotic distribution theory was developed by Eubank and Hart (1992) for linear regression models and with a Mallows C_p type of criterion to select the order. Aerts et al. (1999, 2000) extended this to likelihood-based regression models, and related the order selection test statistic to a test statistic that is the supremum of a set of weighted likelihood ratio statistics. Particularly, the null hypothesis (1) is rejected when an AIC-type criterion of the form

$$\text{aic}(r, C_n) = 2\{\log L(\hat{\boldsymbol{\eta}}, \hat{\beta}_1, \ldots, \hat{\beta}_{p+r}) - \log L(\hat{\boldsymbol{\eta}}, \hat{\beta}_1, \ldots, \hat{\beta}_p)\} - C_n r, \quad r = 0, 1, 2, \ldots, R_n,$$

selects $\hat{r} = \arg\max_{r=0,1,2,\ldots,R_n} \text{aic}(r, C_n) > 0$. Note that $\text{aic}(r, C_n)$ is twice the difference of the maximised log-likelihood value at the model with r additional terms in the series expansion, and the corresponding value at the null model, with as penalty C_n times the number of additional terms r. The difference with a traditional AIC difference is that the penalty constant 2 for the AIC is here replaced by a value C_n, which will determine the level of the test. This approach is equivalent to rejecting the hypothesis (1) when the order selection statistic

$$T_{OS} = \max_{r=1,\ldots,R_n} \frac{2(\mathcal{L}_r - \mathcal{L}_0)}{r} > C_n, \tag{3}$$

where $\mathcal{L}_r = \log L(\hat{\boldsymbol{\eta}}_{p+r}, \hat{\boldsymbol{\beta}}_{p+r})$. Note that the dimension of the nuisance parameter $\boldsymbol{\eta}$ stays the same in all models, though the value of the estimator might be different when different approximations of γ are used. This is indicated in the notation by adding a subscript to the estimator. The value C_n is the critical value of the test statistic, which can be chosen to obtain a certain level for the test. In the setting of completely observed data, the asymptotic distribution theory (see Aerts et al., 1999) provides a method to compute P-values of the test. The idea that we work with in this paper is to use similar likelihood-ratio based test

statistics for the data sets after multiple imputation.

2.2 Likelihood ratio tests after multiple imputation

Multiple imputation is a technique to handle with missing data that inserts values for the missing observations in order to create complete sets of data to which standard methods can be applied. The insertion of values is typically repeated a small number of times m (say 3–10) in order to create m sets of completed data. The insertion of multiple values should help to correct the standard errors of estimators for the additional uncertainty introduced by replacing the unknown values by numbers. Indeed, pretending the inserted values to be the true values of the variables would lead to too optimistic inference. In the context of hypothesis testing, with the availability of m completed sets of data, one could perform m likelihood ratio tests to test hypothesis (1). Meng and Rubin (1992) proposed a method to combine the m separate likelihood ratio values into one single test statistic with an approximate F-distribution. This idea builds on a similar combined testing procedure using Wald statistics instead of likelihood ratio statistics, see Li et al. (1991b).

To introduce the notation, fix a value $r > 0$ and consider first the *parametric* testing problem of the null hypothesis (1) against the parametric alternative hypothesis

$$H_{a,r} : \gamma(\cdot) \in \mathcal{G}_r = \{\gamma(\cdot, \boldsymbol{\beta}_{p+r}) : \boldsymbol{\beta}_{p+r} = (\beta_1, \ldots, \beta_{p+r}) \in \Theta_r\}, \tag{4}$$

where the parameter space $\Theta_r \subset \mathbb{R}^{p+r}$. As a concrete example we could be interested in testing whether $H_0 : E(Y|x) = \beta_1 + \beta_2 x$ versus $H_{a,1} : E(Y|x) = \beta_1 + \beta_2 x + \beta_3 x^2$, which in this case is equivalent to testing whether $\beta_3 = 0$ in the quadratic model for the mean. In the next section we will relax this particular form of the alternative hypothesis to allow for omnibus testing.

Denote $\mathcal{L}_{r,\ell}$ the log-likelihood ratio statistic for testing hypothesis (1) against the specific alternative (4) with r additional parameters for the ℓth imputed set of data, with $\ell =$

$1, \ldots, m$. We denote the average of these test statistics as $\bar{\mathcal{L}}_{r,\bullet} = m^{-1} \sum_{\ell=1}^{m} \mathcal{L}_{r,\ell}$.

We denote the parameter estimators for the ℓth imputed data set by $(\hat{\boldsymbol{\eta}}_{p+r,\ell}, \hat{\boldsymbol{\beta}}_{p+r,\ell})$. The average of these m parameter estimators is denoted by $(\bar{\boldsymbol{\eta}}_{p+r,\bullet}, \bar{\boldsymbol{\beta}}_{p+r,\bullet})$ under the alternative model and by $(\bar{\boldsymbol{\eta}}_{p,\bullet}, \bar{\boldsymbol{\beta}}_{p,\bullet})$ under the null model. We now define a 'log likelihood ratio' value for each of the m imputed data sets that is based on the average parameter value over the m sets of completed data, but using the completed data $\boldsymbol{Z}_{i,\ell}, i = 1, \ldots, n$ for the ℓth round of imputation. This leads to

$$\tilde{\mathcal{L}}_{r,\ell} = \log L(\bar{\boldsymbol{\eta}}_{p+r,\bullet}, \bar{\boldsymbol{\beta}}_{p+r,\bullet}; \boldsymbol{Z}_{1,\ell}, \ldots, \boldsymbol{Z}_{n,\ell}),$$

and their average

$$\widetilde{\mathcal{L}}_{r,\bullet} = \frac{1}{m} \sum_{\ell=1}^{m} \tilde{\mathcal{L}}_{r,\ell}.$$

Meng and Rubin (1992) define the combined test statistic for a parametric testing problem

$$D_r = \frac{\widetilde{\mathcal{L}}_{r,\bullet}}{r\{1 + \frac{m+1}{r(m-1)}(\bar{\mathcal{L}}_{r,\bullet} - \widetilde{\mathcal{L}}_{r,\bullet})\}} \tag{5}$$

and argue that this statistic has an approximate F distribution with degrees of freedom r and ν where

$$\nu = \begin{cases} 4 + (t-4)\{1 + (1 - 2t^{-1})D^{-1}\}^2 & \text{if } t = r(m-1) > 4 \\ t(1 + r^{-1})(1 + D^{-1})^2/2 & \text{otherwise,} \end{cases} \tag{6}$$

with $D = \frac{m+1}{r(m-1)}(\bar{\mathcal{L}}_{r,\bullet} - \widetilde{\mathcal{L}}_{r,\bullet})$. We refer to Reiter (2007) for an alternative approximation to the denominator degrees of freedom that should work better for small sample sizes and is defined to not exceed the denominator degrees of freedom for the complete data.

2.3 Combining the test statistics

Instead of assuming a particular alternative model such as in the alternative hypothesis (4), we construct an order selection test to test H_0 against a broad class of alternative models, similar to the order selection idea in complete data cases as described in section 2.1. Again

we consider a sequence of approximations to the function of interest $\gamma(\cdot)$ as in (2). Each such approximation leads to a maximized log likelihood value, and to a statistic D_r as in (10). Similar to combining the log likelihood ratio test statistics $2(\mathcal{L}_r - \mathcal{L}_0)$ in the order selection statistic T_{OS} in (3), our new test statistic combines the log likelihood ratio statistics D_r that are obtained after multiple imputation. We define

$$D_{OS} = \max_{r=1,\ldots,R_n} D_r.$$

Note that the statistic D_r already contains the number of additional parameters r in its denominator. In the complete data case the likelihood ratio statistic $2(\mathcal{L}_r - \mathcal{L}_0)$ has asymptotically a χ^2_r distribution, and under some assumptions on the likelihood, Aerts et al. (1999) obtained that the asymptotic distribution of T_{OS} is given by

$$P(T_{OS} \leq x) = \exp\left[-\sum_{r=1}^{\infty} \frac{P(\chi^2_r > rx)}{r}\right].$$

Since for the case of missing data D_r follows only an *approximate* asymptotic distribution that is $F_{r,\nu}$, with ν depending on the data (Meng and Rubin, 1992), we do not obtain the limiting distribution for D_{OS}. However, by similarity we investigate by simulation whether the approximation

$$P(D_{OS} \leq x) \approx \exp\left[-\sum_{r=1}^{\infty} \frac{P(F_{r,\nu} > x)}{r}\right] \tag{7}$$

holds in practice. This limiting distribution can be used to obtain approximate P-values, as well as to define the appropriate critical value C_n for a given level for the test. Following the same idea of the order selection test when complete data are exploited, the test rejects the hypothesis (1) when the order selection statistic

$$D_{OS} = \max_{r=1,\ldots,R_n} D_r > C_n. \tag{8}$$

To obtain simulated critical values or P-values, we approximate the infinite series in (7) by a finite sum

$$P(D_{OS} \leq x) \approx \exp\left[-\sum_{r=1}^{R_n} \frac{P(F_{r,\nu} > x)}{r}\right], \tag{9}$$

for a large value of R_n, see Table 1 for values of C_n for several choices of ν.

As an alternative to using this approximation one could use a bootstrap approach. For hypothesis testing, data should be generated under the null hypothesis (Hall and Wilson, 1991).

Note that for ν tending to infinity, it holds that for $F_{r,\nu}$ following a F distribution with r and ν degrees of freedom,

$$\lim_{\nu\to\infty} rF_{r,\nu} \sim \chi^2_r.$$

Therefore, for ν large, $P(F_{r,\nu} > x) \approx P(\chi^2_r > rx)$. Hence, for ν large, the distribution in (7) can be further well approximated by the standard distribution for order selection tests in complete data. That is, for ν large,

$$P(D_{OS} \leq x) \approx \exp\left[-\sum_{r=1}^{\infty} \frac{P(\chi^2_r > rx)}{r}\right].$$

This form of the distribution is free of data-dependent values. For most of the simulated datasets (see Section 3) it turned out that the degree of freedom ν computed as in (6) is sufficiently large for the approximation to hold. Under conditions on the imputation scheme which would guarantee that D, defined below (6), converges to zero, the asymptotic distribution of D_{OS} would be the same as that of T_{OS}. However, we do not impose such conditions and allow also for small values of ν by using the form of the distribution in terms of $F_{r,\nu}$ statistics.

3 Simulations

3.1 Simulation settings and methods

We investigate the quality of the approximation to the asymptotic distribution by means of a simulation study. All calculations have been performed using the statistical software package `R`. We consider different simulation settings, related with different sample sizes

and different percentages of missingness. We simulated independent normally distributed response variables Y_i, $i = 1, \ldots, n$ with mean μ=1, standard deviation $\sigma = 1$, and a covariate X_i that is equally spaced in $[0, 1]$. The covariate is fully observed, while the response vector $Y = (Y_1, \ldots, Y_n)$ contains missing observations. We want to test the null hypothesis H_0 : $E(Y|X) = \beta_0$. The missingness in the response variable is introduced by generating a missing data mechanism that depends on the fully observed variable X_i, leading to the MAR condition, in which the missingness depends only on the observed part of the data. Independent standard normal errors ϵ_{ij} are generated, and a data value y_i is set to be missing when $a(y_i - \overline{y}_\bullet) + \epsilon_{ij} \leq z_\tau$ where $\overline{y}_\bullet$ is the sample mean of $(y_1, \ldots, y_n)$ and z_τ is the $(1 - \tau)$-quantile of $N(0, a^2/12 + 1)$ with τ the chosen percentage of missingness and $a = -1$. This scenario is the same for all the different simulation settings used. Two different sample sizes, $n = 30$ and $n = 50$, are considered and for each of them three different percentages of missingness are taken into consideration, 5%, 15%, and 30%. For each setting we run $N = 2000$ simulations. For the order selection test we take an expansion via orthogonal polynomials (using the function `poly` in R) and a cosine basis with $\psi_j(x) = \cos(j\pi x)$ with $j = 1, \ldots, R_n = 15$. additional terms. We have tried with different orders, but this did not change the results significantly. The number of imputations equals 5 for each situation. For the imputation we have used a method that is available in the R library `mice`, short for "multiple imputation by chained equations". We have considered a regression method (norm); in this method the missing data variables are regressed on the complete data variables in order to estimate the unknown parameters, we then draw values from the posterior distribution of the parameters. These estimated parameters are used with the complete data variable in a linear regression and the fitted values serve as imputations for the missing observations. An overview of imputation methods and available software for multiple imputation is given by Horton and Lipsitz (2001) and Horton and Kleinman (2007).

3.2 Critical values and the null distribution of the test

To start the analysis we calculate the theoretical values of the critical point C_n for various values of the type I error α of the test, under the approximate asymptotic distribution in (7). Table 1 shows these values when the second degree of freedom ν increases from 6 to infinity, for different selected choices of α, when using $R_n = 200$. We want to stress that changes in the upper bound for the first degree of freedom do not affect the values of C_n, when the second degree of freedom is bigger than 20. For instance for $\alpha = 0.01$, the second degree of freedom equal to 20 and the first degree of freedom going from 1 to $R_n = 20$ the cut-off point is again 8.502. If the second degree of freedom is between 6 and 20, then small differences can be found; for $\alpha = 0.01$, the second degree of freedom equal to 6 and the first degree of freedom going from 1 to $R_n = 20$, the critical value C_n is 17.7591, which is slightly smaller than the value in Table 1. Furthermore, values of the critical value C_n are not calculated when the second degree of freedom ν is below 6 because such values have never occurred in any of our simulation studies. The critical values decrease when the second degree of freedom increases; for $\alpha = 0.01$ the drop of C_n is more sensitive than for the other selected α's, which show more stable values. For all the chosen nominal levels α the higher is the degree of freedom the more stable is the value of C_n. Furthermore when the second degree of freedom is close to infinity the critical values are similar to the ones computed using the χ^2 distribution. This table is important to have an idea about which value of C_n should be considered to perform the order selection test for hypothesis (1), as shown by formula (3). For the theoretical values of C_n for fully observed data we refer to Hart (1997).

The theoretical values of C_n are used to obtain the simulation results shown in Table 2, using test (8), rejecting the null hypothesis when the observed value of $D_{\mathrm{OS}} > C_n$. We here test the null hypothesis $H_0 : E(Y|X) = \beta_0$. The table shows the simulated significance level of the test, under different choices of α and different bases, when the two different methods for imputing the missing values are used. We observe that the test performs well, with the

Table 1: Simulated critical values C_n using the distribution in (9), for various values of the second degree of freedom of the F distributions, for a given nominal level α and for $R_n = 200$.

Degree of freedom ν	$\alpha = 0.01$	0.05	0.10
6	17.7592	8.8685	6.3072
7	14.9495	7.8246	5.6748
8	13.2251	7.1435	5.2499
9	12.0745	6.6676	4.9460
10	11.2590	6.3180	4.7186
20	8.5020	5.0560	3.8790
30	7.8327	4.7200	3.6217
40	7.5327	4.5710	3.5096
50	7.3624	4.4865	3.4467
60	7.2526	4.4320	3.4064
70	7.1759	4.3940	3.3783
80	7.1193	4.3659	3.3577
90	7.0759	4.3443	3.3418
100	7.0414	4.3272	3.3292
200	6.8900	4.2519	3.2740
300	6.8408	4.2274	3.2560
400	6.8164	4.2153	3.2471
500	6.8019	4.2081	3.2418
600	6.7922	4.2032	3.2383
700	6.7853	4.1998	3.2358
800	6.7801	4.1972	3.2339
900	6.7761	4.1952	3.2325
1000	6.7729	4.1936	3.2313
∞	6.7442	4.1793	3.2208

significance levels close to the nominal levels, when the data are imputed using the regression method (norm), this for both sample sizes 30 and 50 and for all the different percentages of missingness. The results in the table also show that the complete cases method, where cases with missing observations are completely left out for performing the test, performs badly, especially for a small sample size, independently of the percentage of missingness. Hence, discarding the missing observations leads to biased results for estimation, and to a too high simulated value for the significance level of the test. For instance when sample size equals 30 and the percentage of missingness is 30, at the nominal level of $\alpha = 0.05$, the simulated significance level for the test we propose is 0.0575, while under the complete cases the level is 0.1685. There is, however, an improvement when the sample size increases to 50.

In addition to the previous analyses we investigate the approximation of the asymptotic distribution of D_{OS} by using the bootstrap, as in (7). We consider $B = 1000$ bootstrap replicates by resampling with replacement the pairs (Y_i, X_i), for data sets with sample size $n = 30$ and percentage of missingness equal to 30%. Note that the bootstrap data might contain a different percentage of missingness due to the resampling. For each of the B bootstrap data sets, we compute the test statistic D_{OS} in precisely the same way as for the original data. Those 1000 bootstrap values of D_{OS} are used to construct a bootstrap density plot. We also generate data from the approximate asymptotic distribution (9) using, as second degree of freedom in the F-distributions, the degree obtained by performing the order selection test to the original dataset. Figure 1 displays the bootstrap density function of D_{OS} and the density of the distribution of (9) for the different settings. The shape of the distributions is quite similar, even with large percentages of missing values.

3.3 Simulated power of the tests

To evaluate the performance of the test, we now investigate the power of the order selection test using multiple imputation. We considered four different settings: the sample size is

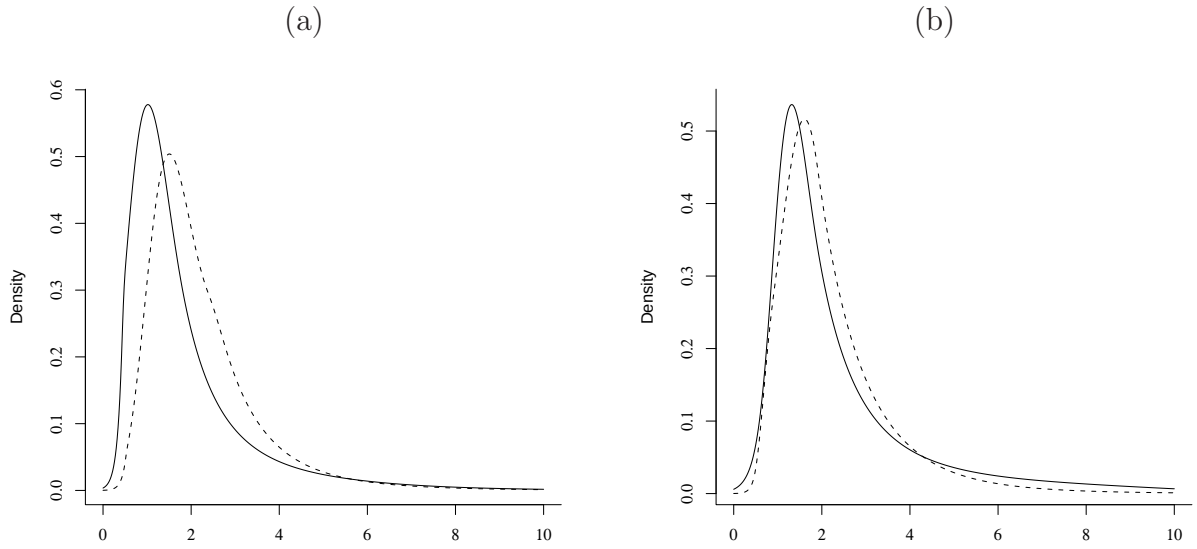
Table 2: Results of a simulation study. The table shows, based on cosine and polynomial basis, simulated significance levels of the test D_{OS} when the theoretical critical values C_n are used, for different values of the nominal level α. The imputation methods is 'norm', also a complete case analysis (CC) using T_{OS} is performed. The original data analysis (before introducing missingness) is shown in the last line.

		$n = 30$			$n = 50$		
% missing	Method	$\alpha = 0.01$	0.05	0.10	$\alpha = 0.01$	0.05	0.10
				Cosine basis			
5%	Mis_{norm}	0.0115	0.0545	0.0985	0.0090	0.0430	0.0920
	CC	0.0175	0.0810	0.1630	0.0135	0.0565	0.1195
15%	Mis_{norm}	0.0065	0.0305	0.0660	0.0055	0.0400	0.0820
	CC	0.0215	0.0940	0.1900	0.0105	0.0585	0.1210
30%	Mis_{norm}	0.0055	0.0335	0.0585	0.0090	0.0535	0.0875
	CC	0.0790	0.1905	0.3470	0.0165	0.0645	0.1365
	Orig	0.0135	0.0835	0.1655	0.0150	0.0595	0.1285
				Polynomial basis			
5%	Mis_{norm}	0.0130	0.0540	0.1040	0.0090	0.0460	0.0995
	CC	0.0160	0.0795	0.1585	0.0115	0.0595	0.1260
15%	Mis_{norm}	0.0140	0.0475	0.0900	0.0085	0.0450	0.0915
	CC	0.0195	0.0880	0.1885	0.0105	0.0590	0.1150
30%	Mis_{norm}	0.0165	0.0575	0.1000	0.0105	0.0470	0.0765
	CC	0.0600	0.1685	0.3260	0.0160	0.0650	0.1385
	Orig	0.0140	0.0790	0.1610	0.0130	0.0620	0.1295

equal to 30 or 50, and two different percentages of missingness 5% and 30%. We generate normal data Y with $E(Y|X) = 1 + \beta_1 X$, where X is a equally spaced variable in $[0, 1]$ and β_1 takes values in a grid $(0, 0.2, 0.5, 0.7, 1, 1.2, 1.5, 1.7, 2)$. We performed the order selection test D_{OS}, using cosine and polynomial bases, for testing the no effect null hypothesis $H_0 : E(Y|X) = \beta_0$.

To calculate the power we consider the theoretical values of C_n as shown in Table 1. Figure 2 shows good results also when the sample size is small (equal to 30) and the percentage of missingness is large (30%). The order selection test D_{OS} that uses imputation, is able to

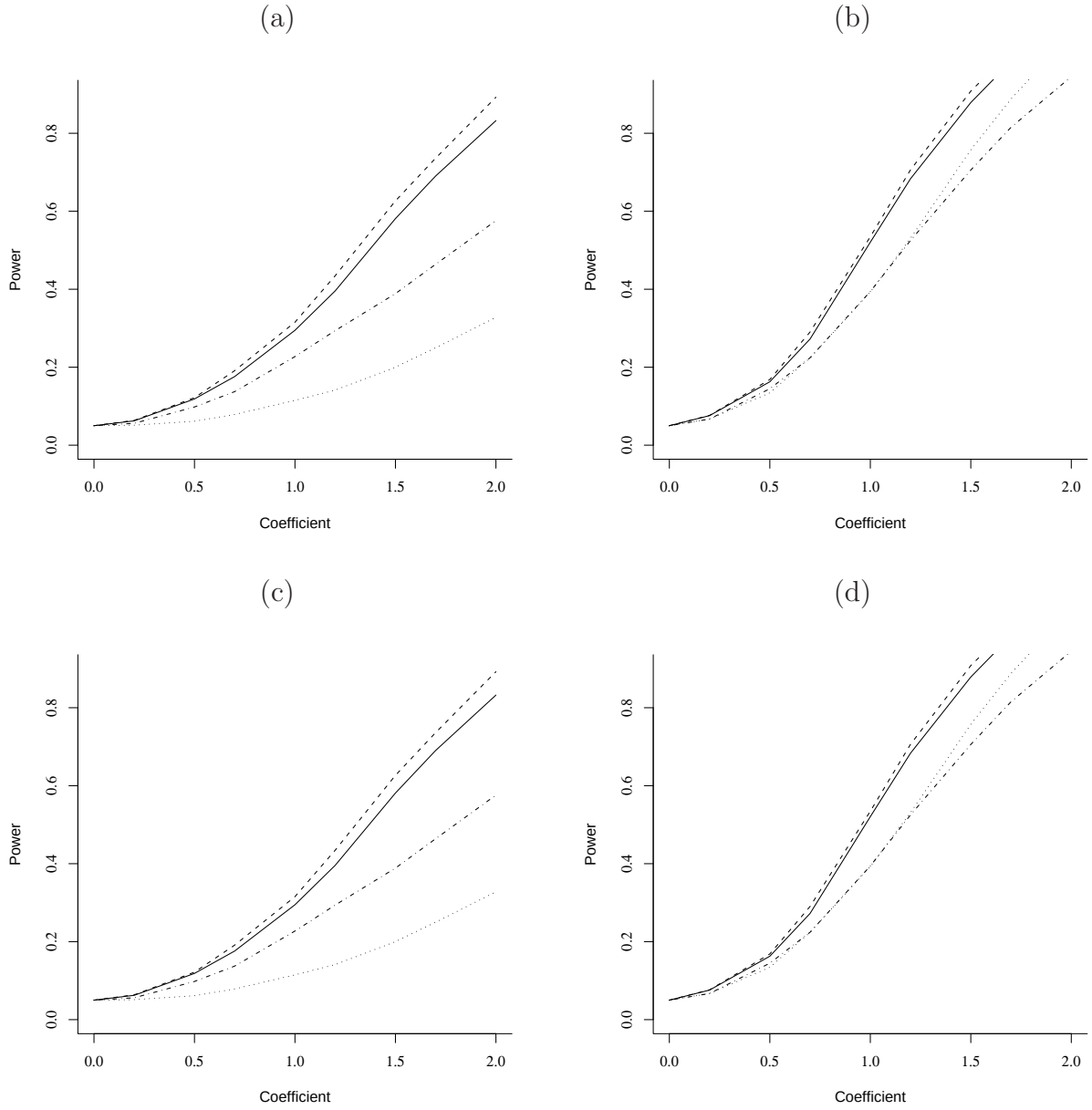
Figure 1: Density plots under H_0 of D_{OS} for testing the no-effect null hypothesis for a data set with missing observations for the responses and sample size equal to 30. The density obtained by bootstrap resampling is shown with the solid line, while the dashed line displays the density when data are simulated from the approximate asymptotic distribution (9). Plots (a) displays the distribution under cosine basis, while (b) uses a polynomial basis, with percentage of missingness equal to 30%.



detect departures from the null hypothesis, also for small values of β_1, while the complete cases order selection test does not work well. Performing a size adjustment for the power, we can see how low the real power is when the complete cases analysis is performed, compared with the order selection test after imputation. The proposed test D_{OS} is performing nicely, also for large percentages of missingness. When the percentage of missing observations is small, then the two tests show the same performance. When the sample size increases to 50, there is an increase in power, and a better behaviour of the test T_{OS} for the complete cases when the percentage of missingness is not severe, due to the larger sample size.

Next, we consider a more complicated true alternative model for investigating the power of the test. We used again two different sample sizes ($n = 30$ and $n = 50$) and two different percentages of missingness (5% and 30%). The response variable Y is now generated from a

Figure 2: Power curves for testing no effect with a true linear alternative at the 5% level. In each plot the power curve of D_{OS} for the missing data and of T_{OS} for the complete cases are displayed. We used a cosine basis with (a) $n = 30$ and (b) $n = 50$ and a polynomial basis with (c) $n = 30$ and (d) $n = 50$. Simulated power curves are shown of T_{OS} for the complete cases with 5% missingness (solid line) and with 30% missingness (dotted line). The proposed test D_{OS} with 5% missingness (dashed line), and with 30% missingness (dot-dashed line).



normal distribution with $E(Y|X) = \exp(-2+\beta_1 X)$, where X is a equally spaced variable in $[0,1]$; β_1 belongs to the grid $(0,1,2,3,4,5)$. We performed the order selection test D_{OS}, using polynomial and cosine bases, for testing the no effect null hypothesis $H_0 : E(Y|X) = \beta_0$.

Again, when the sample size equals 50 the order selection test D_{OS} using multiple imputation and the complete case order selection test T_{OS} have similar power curves. When the sample size is smaller, then the T_{OS} test based on the complete cases only is not working well due to discarding the missing observations, and a too large rejection probability under the null hypothesis, while the D_{OS} test is making a correct decision more often. In fact using a corrected size the power for the complete cases is low compared to the one of the test after imputation.

3.4 Sensitivity analysis

The above sections showed that the order selection test when missing data are present works quite properly; which means that the imputation is done properly as well. In the above simulation the model for performing the imputation is correct; in this section we want to perform a sensitivity analysis in order to verify if a misspecification problem could arise in the multiple imputation method. We start considering the same simulation setting described in section 3.1, where Y_i, $i = 1, \ldots, n$ are independent normally distributed response variables with mean μ=1, standard deviation $\sigma = 1$, and a covariate X_i that is equally spaced in $[0,1]$. The covariate is fully observed, while the response vector $Y = (Y_1, \ldots, Y_n)$ contains missing observations. We want to test the null hypothesis $H_0 : E(Y|X) = \beta_0$, considering an orthogonal polynomial expansion. The model used for the multiple imputation is different from that in the above sections; there we used X_i as a variable in a linear regression model to perform the imputation, here we consider a variable transformation and use for imputation a mean model of the form $\beta_0 + \beta_1 g(X)$, where $g(X)$ is one of the functions given in the first column of table 3. Note that `poly(X, degree=5)` stands for orthogonalized polynomials of

Figure 3: Power curves for testing no effect with a true exponential alternative at 5% level. In each plot the power curve of D_{OS} for the missing data and of T_{OS} for the complete cases are displayed. We used a cosine basis with (a) $n = 30$ and (b) $n = 50$ and a polynomial basis with (c) $n = 30$ and (d) $n = 50$. Simulated power curves are shown of T_{OS} for the complete cases with 5% missingness (solid line) and with 30% missingness (dotted line). The proposed test D_{OS} with 5% missingness (dashed line), and with 30% missingness (dot-dashed line).

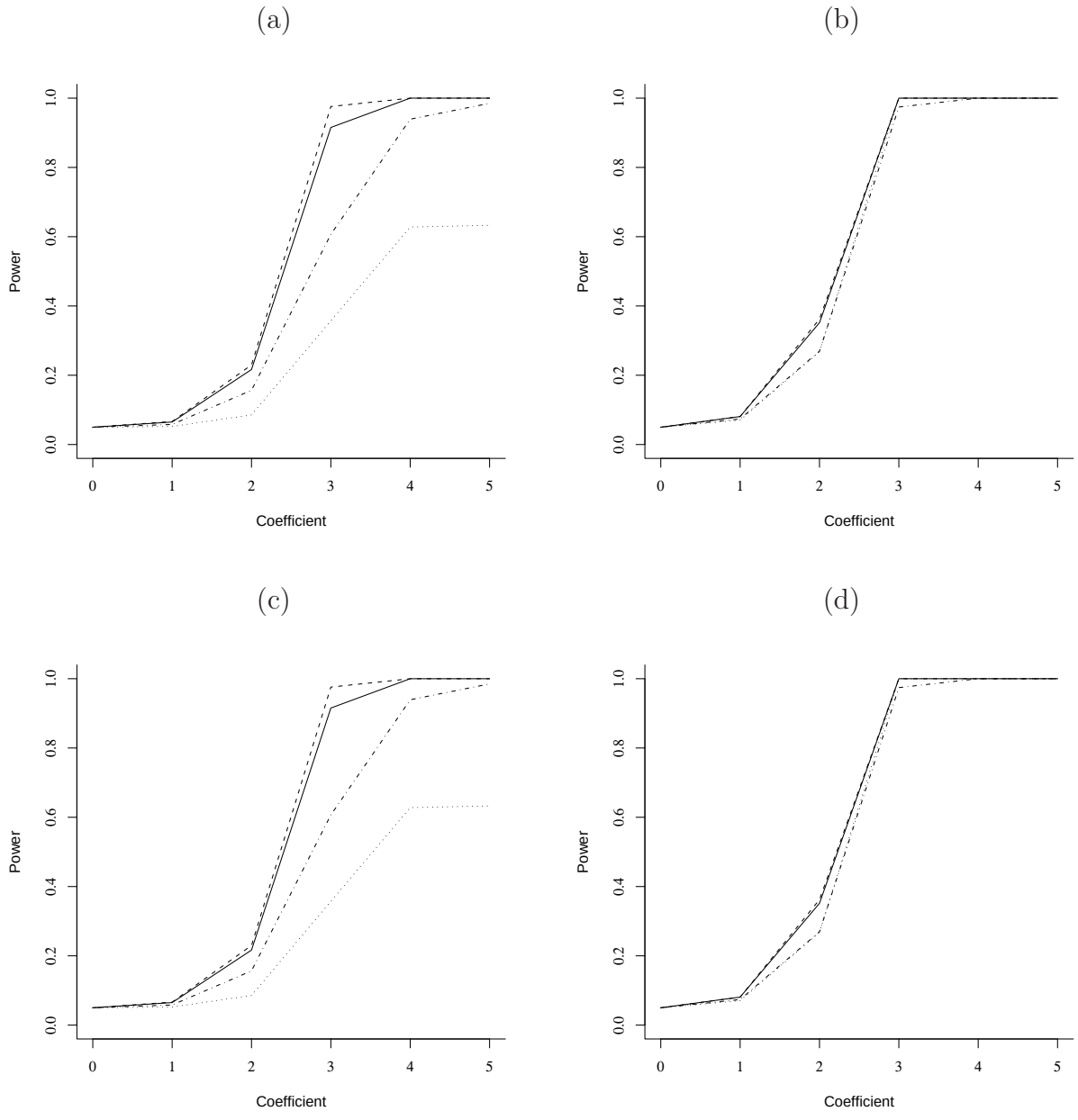


Table 3: Results of a simulation study. The table shows, based on polynomial basis, simulated significance levels of the test D_{OS} when the theoretical critical values C_n are used, for different values of the nominal level α, performing a sensitivity analysis. The imputation methods is 'norm'. The null hypothesis is $H_0 : E(Y|X) = \beta_0$. Different variable transformations have been used for imputation.

Var in imputed Model	% missing	$n = 30$		
		$\alpha = 0.01$	0.05	0.10
		Polynomial basis		
$\log(X + 10)$	5%	0.0135	0.0555	0.1070
	15%	0.0110	0.0460	0.0880
	30%	0.0110	0.0435	0.0755
X^2	5%	0.0135	0.0535	0.1060
	15%	0.0125	0.0430	0.0850
	30%	0.0155	0.0610	0.1045
$\cos(6\pi X)$	5%	0.0060	0.0370	0.0775
	15%	0.0015	0.0090	0.0285
	30%	0.0000	0.0025	0.0110
`poly(X, degree=5)`	5%	0.0110	0.0530	0.1095
	15%	0.0100	0.0445	0.0910
	30%	0.0065	0.0520	0.0960

degree 5. Hence this last model is the only one that contains structures as used in fitting the alternative models for the construction of the order selection test.

Table 3 displays some nice results about the sensitivity of the imputation model. The order selection test is working fine for all the settings except for the usage of the high frequency cosine function which results in too small simulated type I errors when the percentage of missingness is large.

We deepen the question by analyzing a different setting; Y_i, $i = 1, \ldots, n$ are independent normally distributed response variables with mean $\mu = E(Y|X)$ as specified below, standard deviation $\sigma = 1$, and a covariate X_i that is equally spaced in $[0, 1]$. The covariate is fully observed, while the response vector $Y = (Y_1, \ldots, Y_n)$ contains missing observations. We want

to test the null hypothesis $H_0 : E(Y|X) = \beta_0 + \beta_1 X$, considering an orthogonal polynomial expansion. The model used for the multiple imputation is summarized in Table 4, by showing the variable $g(X)$ that is used in a model for the mean of the form $\beta_0 + \beta_1 g(X)$. Unlike the previous setting where all models contained the model under the null hypothesis (which was there a constant function), in this setting this is only true for the first two models. The first setting gives the correct imputation model (a linear model for the mean), in the second setting, the model used for imputation is richer than necessary (it contains a fifth degree orthogonal polynomial in X). We see that this only slightly reduces the observed significance levels in our simulation study. For the other three models, the model that is used for imputation does not contain the null model, and is hence misspecified. For a small percentage of missingness, all imputation methods are still doing reasonably well, even the high frequency cosine model. When the missingness increases, this particular model has problems in keeping the level, but the other misspecified imputation models are still giving reasonable results.

While these simulation results show that it is not really crucial to know the correct model for imputations, it is still advised to pay attention to this part of the modeling process. It might be interesting to further search for methods that are robust against misspecification of the imputation model.

4 Data analysis

Climate change is having a large impact in political decisions and and it is nowadays one of the most serious challenges to face. Climate change may result from both natural factors and human activities. Environmental agencies play an important role in measuring the effects of climate change in our daily life and in different economic sectors. An important effect of climate change is the global warming, which represents the increase in the temperature of the atmosphere near the earth's surface. Temperature change may occur because of the increase

Table 4: Results of a simulation study. The table shows, based on polynomial basis, simulated significance levels of the test D_{OS} when the theoretical critical values C_n are used, for different values of the nominal level α, performing a sensitivity analysis. The imputation methods is 'norm'. The null hypothesis is $H_0 : E(Y|X) = \beta_0 + \beta_1 X$, with $\beta_0 = 1$ and $\beta_1 = 2$.

Var in imputed model	% missing	$n = 30$		
		$\alpha = 0.01$	0.05	0.10
		Polynomial basis		
X	5%	0.0080	0.0470	0.0945
	15%	0.0080	0.0370	0.0680
	30%	0.0090	0.0400	0.0700
`poly(X, degree=5)`	5%	0.0065	0.0440	0.0835
	15%	0.0050	0.0285	0.0545
	30%	0.0030	0.0320	0.0605
$\log(X + 10)$	5%	0.0070	0.0435	0.0940
	15%	0.0060	0.0285	0.0685
	30%	0.0025	0.0300	0.0600
X^2	5%	0.0055	0.0385	0.0805
	15%	0.0025	0.0160	0.0445
	30%	0.0020	0.0170	0.0390
$\cos(6\pi X)$	5%	0.0050	0.0335	0.0780
	15%	0.0015	0.0130	0.0345
	30%	0.0000	0.0080	0.0270

of the emission of greenhouse gasses, due to human activities. Greenhouse gasses are found in the atmosphere and are emitted through natural or artificial processes; for this reason they represent a strategic aspect to measure and control. Among the economic sectors that contribute to global warming, agriculture is an important one, since it is highly sensitive to climate change, because its activities depend directly on climate conditions, and because of its greenhouse gasses release. Crop and meat production, milk products, livestock, are some of the agricultural activities that contribute to the global warning. The European Union has developed climate change policies to reduce the emission of greenhouse gases

by agricultural activities, following the guidelines of the Kyoto Protocol. For instance the Common Agricultural Policy (CAP) is used to regulate the production, trade, and processing of agricultural products in the EU. Several factors directly connect climate change and agricultural productivity, such as average temperature increase, change in rainfall amount, atmospheric concentrations of CO_2, etc.

We want to investigate the relationship between the emission of greenhouse gasses and the production of wheat. The data come from Eurostat, the Statistical Office of the European Communities, which gathers and analyses figures from the national statistical offices across Europe and provides statistical information. Data to be analyzed are (Y_i, X_i), $i = 1, \ldots, 33$, where the response variable Y_i is the total greenhouse gas (GHG) emissions in thousands of tons, for the agricultural sector, and X_i is the yield $(100kg/ha)$ of wheat in 33 European countries for the year 2006; the response variable contains 7 missing observations. For the analysis we rescale the explicative variable to the interval [0,1]. We want to analyze the relationship between the yield and the emission of greenhouse gas; we consider a linear regression model. We test two different null hypotheses

$$H_0 : E(Y|X) = \beta_0$$

and

$$H_0 : E(Y|X) = \beta_0 + \beta_1 X$$

We consider again the polynomial and the cosine bases used in the simulation study, with $R_n = 15$.

Table (5) displays the results when the order selection test is performed. At the 10% level, the test D_{OS} rejects the null hypothesis that the expected value of the conditional distribution of Y given X is constant, which is expected since the wheat production has an impact in the total emission of greenhouse gases; on the contrary, the null hypothesis that the expected value of $(Y|X)$ is linear, at the 10% level, is not rejected.

Table 5: Results for the climate data. The table shows critical values of the test D_{OS}, the P-value and, for the missing data approach, the second degree of freedom used to calculate the corresponding P-value.

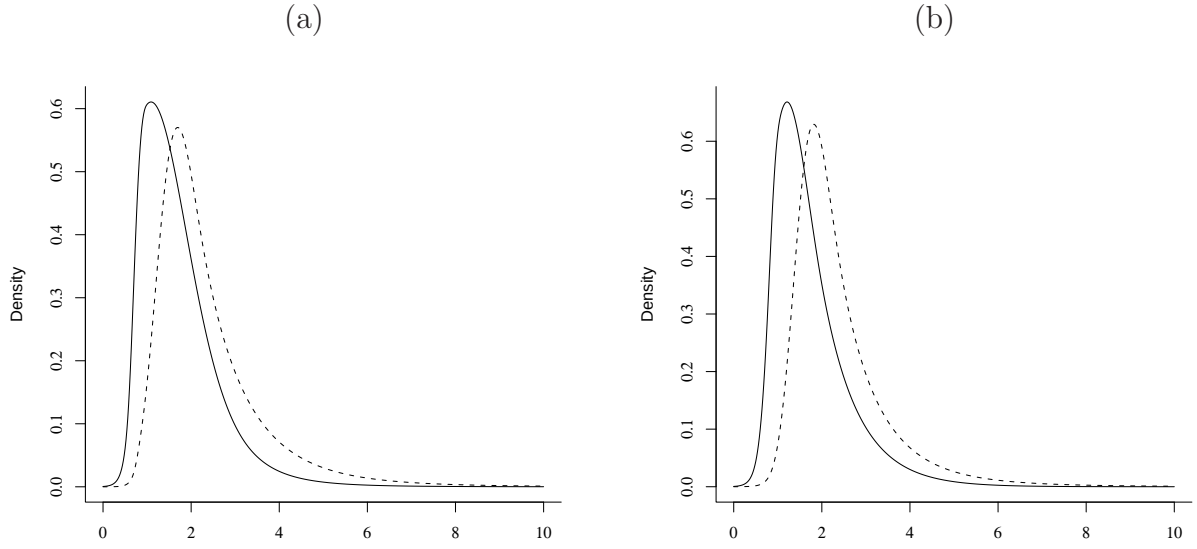
	Cosine basis			Polynomial basis		
Method	C_n	P-value	df ν	C_n	P-value	df ν
			μ constant			
Missing	3.943	0.061	300.92	3.353	0.083	138.24
CC	4.159	0.051	–	4.329	0.045	-
			μ linear			
Missing	0.826	0.797	197.20	0.898	0.767	243.24
CC	1.874	0.323	–	1.916	0.310	-

Furthermore we want to check whether the distribution of the test D_{OS}, applied to the dataset resembles the asymptotic distribution of (9). We drew 2000 bootstrap for the dataset, using the cosine and the polynomial bases, to estimated the distribution. We carried out the analysis testing the null hypothesis $H_0 : E(Y|X) = \beta_0 + \beta_1 X$. To approximate the asymptotic distribution (9) we use as second degree of freedom ν=197.20 for the test with cosine basis and ν=243.24 when using an orthogonal polynomial basis. Figure 4 displays the result, where we can see that the shape of the bootstrap distribution is quite similar to the approximated asymptotic distribution.

5 Model selection via AIC for multiply imputed data

In the previous sections we discussed a nonparametric testing method that works with missing data. We here use the obtained results to develop a version of Akaike's information criterion to handle multiple imputations. While it is straightforward to apply any traditional variable selection criterion such as the AIC to the separate imputed sets of data, the biggest question is how the results of those separate AIC selections should be combined? In a Bayesian setting Yang et al. (2005) compute for each imputed dataset separately the

Figure 4: Density plots of D_{OS} for testing the null hypothesis of linearity for the data set. The density obtained by bootstrap resampling is shown with the solid line, while the dashed line displays the density when data are simulated from the approximate asymptotic distribution (9). Plot (a) displays the distribution using a cosine basis, while for (b) a polynomial basis is used.



posterior for each of the candidate models and then take for each model separately the average of the posterior probabilities over the different imputed data sets. Schomaker et al. (2007) work with the AIC and imputation. They mention two approaches. A first one is to compute the classical AIC for each imputed dataset separately and then compute the average of the AIC values to make a ranking of the candidate models. Their second method is the one that they actually apply in their paper. This consists of computing an averaged dataset that consists of the average of each data value after imputation. Now they have a single dataset to which the classical AIC can be applied. We here propose a theoretically solid version of the AIC that is related though different from the two mentioned proposals. We will see that actually a combination of the two proposals is required.

Multiple imputation for a model S leads to m different datasets, each with its own maximized log likelihood function. Often the candidate models are nested, in which case

we denote by S_0 the smallest model under consideration. If we were in a testing setting to compare a model S (under the alternative hypothesis) with a simpler model S_0 (under the null hypothesis) we could apply the results of Meng and Rubin (1992) who proposed to combine the m separate likelihood ratio values into one single test statistic with an approximate F-distribution, as in Section 2.2. Denote the number of parameters in model S by $|S|$, and the difference in numbers of parameters of the two models by $p_S = |S| - |S_0|$. By the results in Meng and Rubin (1992), we obtain that the combined test statistic for testing model S_0 versus model S is

$$\frac{\tilde{D}_S}{p_S} = \frac{\widetilde{\mathcal{L}}_{S,\bullet}}{p_S\{1 + \frac{m+1}{p_S(m-1)}(\bar{\mathcal{L}}_{S,\bullet} - \widetilde{\mathcal{L}}_{S,\bullet})\}}. \tag{10}$$

This statistic has an approximate F distribution with degrees of freedom p_S and ν where ν is as in (6). The second degree of freedom ν is expected to be large under a good imputation scheme where D will be small. Therefore we can work with p_S only as a penalty term in the AIC difference for model S compared to model S_0, see also Section 2.3. Thus we arrive at the definition of the AIC difference for model S compared to model S_0 as

$$\text{aic}(S, S_0) = -\tilde{D}_S + 2\,p_S. \tag{11}$$

Note that the constant 2 is already absorbed in the notation for the log likelihood ratio statistics. These differences can be computed for all models S under consideration, with $\text{aic}(S_0, S_0) = 0$. The model with the smallest AIC difference is considered the best one. Criterion (11) is generally applicable for use with multiple imputation for likelihood models.

6 Discussion and extensions

We introduced an order selection test to apply to data with missing observations. In the simulations we have considered the situation of a missing response variable with a completely observed covariate. Since the likelihood ratio test on which the order selection test is based,

can also be applied to data sets with missing covariates, the tests are equally well applicable to data with missing covariates. One requirement is that a proper imputation method should be used to lead to a valid asymptotic distribution of the, for imputation combined, likelihood ratio test. In cases where the approximate asymptotic distribution is not expected to work well, a bootstrap procedure can be used to provide P-values.

While the illustrations in this paper are restricted to the case of a simple regression model, the order selection testing idea is readily extended to be applicable to multiple regression. We refer to Aerts et al. (2000) and Bissantz et al. (2009) for examples and the construction of a series expansion in more than one variable.

Since the Wald test is asymptotically equivalent to the likelihood ratio test, one could modify the proposed test statistic D_{OS} to use the Wald statistics instead of the likelihood ratio statistics. This test is expected to have similar power behaviour. One other related construction that could be of particular interest would be to combine score statistics instead. However, we are not aware of results on the construction and asymptotic distribution of score tests, combined after multiple imputation. This seems an interesting topic for further research since such score tests could be applied to models that are not likelihood based (for example based on generalized estimating equations, or quasi-likelihood), and can be made robust for model misspecification.

Other related test statistics following the order selection idea could be constructed for the situation of missing data, following their equivalent ideas for complete sets of data. In particular, one could consider the Bayesian information criterion BIC for order selection, hereby leaving out the order zero as a possibility, due to the consistency of the BIC as a model selection method. Such test was first considered for goodness-of-fit testing by Ledwina (1994). Claeskens and Hjort (2004) discuss some alternative schemes based on both BIC and AIC that have better power properties. Such tests could be of interest to investigate in the missing data setting as well.

References

Aerts, M., Claeskens, G., and Hart, J. D. (1999). Testing the fit of a parametric function. *Journal of the American Statistical Association*, 94:869–879.

Aerts, M., Claeskens, G., and Hart, J. D. (2000). Testing lack of fit in multiple regression. *Biometrika*, 87:405–424.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. and Csáki, F., editors, *Second International Symposium on Information Theory*, pages 267–281. Akadémiai Kiadó, Budapest.

Bissantz, N., Claeskens, G., Holzmann, H., and Munk, A. (2009). Testing for lack of fit in inverse regression – with applications to biophotonic imaging. *Journal of the Royal Statistical Society, Series B*, 71(1):25–48.

Claeskens, G. and Hjort, N. L. (2004). Goodness of fit via nonparametric likelihood ratios. *Scandinavian Journal of Statistics*, 31:487–513.

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

González-Manteiga, W. and Pérez-González, A. (2006). Goodness-of-fit tests for linear regression models with missing response data. *Canad. J. Statist.*, 34(1):149–170.

Hall, P. and Wilson, S. R. (1991). Two guidelines for bootstrap hypothesis testing. *Biometrics*, 47(2):757–762.

Hart, J. D. (1997). *Nonparametric Smoothing and Lack-of-fit Tests*. Springer-Verlag, New York.

Horton, N. J. and Kleinman, K. P. (2007). Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models. *Amer. Statist.*, 61(1):79–90.

Horton, N. J. and Lipsitz, S. R. (2001). Multipe imputation in practice: Comparison of software packages for regression models with missing variables. *Amer. Statist.*, 55(3):244–254.

Ledwina, T. (1994). Data-driven version of Neyman's smooth test of fit. *Journal of the American Statistical Association*, 89:1000–1005.

Li, K. H., Meng, X.-L., Raghunathan, T. E., and Rubin, D. B. (1991a). Significance levels from repeated p-values with multiply-imputed data. *Statist. Sinica*, 1(1):65–92.

Li, K. H., Raghunathan, T. E., and Rubin, D. B. (1991b). Large-sample significance levels from multiply imputed data using moment-based statistics and an F reference distribution. *J. Amer. Statist. Assoc.*, 86(416):1065–1073.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical analysis with missing data*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition.

Meng, X.-L. and Rubin, D. B. (1992). Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika*, 79(1):103–111.

Reiter, J. P. (2007). Small-sample degrees of freedom for multi-component significance tests for multiple imputation for missing data. *Biometrika*, 94(2):502–508.

Schomaker, M., Heumann, C., and Toutenburg, H. (2007). New approaches for model selection under missing data. Technical report, Department of Statistics, Ludwig-Maximilians-Universität München.

Yang, X., Belin, T. R., and Boscardin, W. J. (2005). Imputation and variable selection in linear regression models with missing covariates. *Biometrics*, 61(2):498–506.