Regression and residual analysis in linear models with interval censored data

Dissertation

zur Erlangung des Grades

eines Doktors der Naturwissenschaften

der Universität Dortmund

Dem Fachbereich Statistik der Universität Dortmund

vorgelegt von

Rebekka Topp

Dortmund 2002

Gutachter:

Prof. Dr. G. Gómez
Prof. Dr. S. Schach
Prof. Dr. W. Urfer

Tag der mündlichen Prüfung: 19. Juli 2002

Introduction to interval censored data and overview of the two parts of the thesis

Interval censored data arises naturally in medical longitudinal follow-up studies in which the event of interest can not be easily observed, for instance cancer recurrence or the elevation of levels of a biomarker without noticeable symptoms. In these situations, the patients are usually examined at clinical visits that take place only in certain time intervals, and the event of interest may then occur between two consecutive clinical visits. Then, one observes only a certain time interval $[X_L, X_R]$ which is known to include the true time X of onset of the event of interest. This type of interval censoring is called interval censoring case II. As special cases it includes left censoring and right censoring for X_L equal to zero and X_R infinity, respectively. Another type of interval censoring occurs when the event is only known to be smaller or larger than an observed monitoring time. This kind of data is referred to interval censoring case I, or current status data. Finally, one speaks of doubly censored data if one observes $\min\{\max\{X, X_L\}, X_R\}$. For a more extensive review of the different types of interval censored data see Gómez et al. (2001b). In this thesis, interval censoring case II will be considered and the censoring intervals will be taken to be closed on both sides in order to account for exact observations.

An example for interval censored data is given in Betensky and Finkelstein (1999) who introduce the AIDS clinical trial group protocol 181, a natural history substudy of a comparative trial of three anti-pneumocystis drugs. The patients were monitored periodically for evidence of bacterial and viral infections, with the objective of understanding the relationship between these two events, and eventually the natural history of AIDS. Many patients missed several of the prescheduled clinic visits, and when they returned to the hospital for examination, new laboratory indications for the two events were found. Thus, their times until occurrence of the bacterial or viral infection were censored into the time intervals between their last and their new clinic visits.

Another example is the AIDS clinical trial group protocol 359, a randomized clinical trial designed to compare six different anti-retroviral treatment regimens for HIV-infected persons who had previously failed on the protease inhibitor Indinavir (see Gulick et al., 2000). The patients were monitored periodically for their viral load levels with the aim to determine the time period these levels remained below the threshold of 500 viral copies/ml. It happened that the viral load levels climbed above the threshold between two consecutive clinic visits so that the exact time below 500 copies/ml was interval censored into the time interval $[X_1, X_2]$, where X_1 is the elapsed time between the first viral load observation below 500 copies/ml and the last observation before the viral load is subsequently observed to be above 500 copies/ml. Similarly, X_2 is the elapsed time between the visit prior to the first viral load observed below 500 copies/ml and the first visit that the viral load is subsequently observed to be above this threshold.

Methods for interval censored data have been strongly developed in the past decades. An approach for the estimation of the distribution function when the data is interval censored is found in the article by Peto (1973). Turnbull in 1976 presented a theory for nonparametrically estimating the distribution function of interval censored variables, incorporating in the estimation process the idea of self-consistency developed by Efron (1967). Turnbull's work had a strong impact on the further development of all kind of statistical methods for interval censored data, including the field of linear regression. The statistical properties of Turnbull's nonparametric maximum likelihood estimator (NPMLE) have been studied very extensively. Concerning uniqueness, consistency and asymptotic properties see for example Gentleman and Geyer (1994), Yu, Schick, Li and Wong (1998), Pan and Chappell (1999) or Yu, Li and Wong (2000). Resulting from problems in developing a distribution theory of Turnbull's NPMLE, Groeneboom and Wellner (1992) characterized the NPMLE using isotonic regression theory and thereof derived a distribution theory for it.

Some research has also been done on variance estimation of the estimated survival function for interval censored data. Two methods for this problem are studied in Sun (2001). Since the underlying survival function can be assumed to be smooth in many applications, and the NPMLE as a step function does not efficiently use this information, some proposals for smooth estimation of the survival function for interval censored data have been made. See for example Li, Watkins and Yu (1997) or Pan (2000). Recently, an extension of Turnbull's NPMLE to the case of bivariate interval censored data was proposed by Betensky and Finkelstein (1999).

Concerning parameter estimation in linear models with interval censored data, Finkelstein and Wolfe in 1985 developed estimation theory for linear models when the response is interval censored. They proposed a semiparametric approach using an EM algorithm for the maximization of the likelihood function under different parametric models for the covariate distribution, but without assuming a parametric form for the distribution of the response variable. Li and Pu (1999) applied a least squares approach to the log-linear model with interval censored response. For regression analysis with an interval censored covariates, Gómez, Espinal and Lagakos (2002) proposed a semiparametric approach by maximizing the data likelihood under the assumption of a normal distribution for the response. The covariate distribution is estimated nonparametrically via Turnbull's (1976) method. Recently, Gil, López-García, Lubiano and Montenegro (2001) considered linear relations between two interval censored variables by defining a metric for the distance between the observed values of the response and those predicted from the model.

The estimation of the regression parameters of a linear model is also considered in the first part of this thesis where a new estimation theory is presented for models with both interval censored response and covariate. Unlike Gil et al. (2001), it does not use certain distances between the observed and predicted data but is an extension of the method of Gómez et al. (2002) and considers a semiparametric maximum likelihood approach.

Closely related to linear model estimation is the field of residual analysis. In regression theory, the analysis of residuals is an integrated tool necessary to complete the process of fitting linear models. However, in connection with interval censored data, only very few research has been done. For proportional hazard models, Farrington (2000) derived interval censored counterparts to the right censored Cox-Snell, martingale, deviance, and Schoenfeld residuals.

For linear models, Gómez et al. (2002) proposed an intuitive definition of residuals coming from linear models that incorporate interval censored covariates. The second part of this thesis presents a new residual theory for regression analysis with interval censored covariates, which is shown to be superior to that proposed by Gómez et al. (2002).

Introduction

The first part of this thesis deals with linear regression analysis when both response and covariate are interval censored. Linear regression analysis is a statistical technique for investigating and modelling relationships between different variables. A statistical relation between two random variables (Y and Z, say) is defined such that one variable can be expressed in terms of a mathematical function of the other variable, for example $Y = f(Z) + \varepsilon$. In this case, Y is called the dependent variable or response, Z is the independent variable or covariate, and ε is an error term. To examine the linear relationship between Y and Z (or some more Z), an appropriate model should be chosen on the nature of the statistical relation and the variable types under consideration.

When saying a relationship between some variables is 'linear', this usually refers to linearity in the parameters. In contrast, the value of the highest power of the independent variable in the model is called the 'order' of the model. For example, $Y = \beta_0 + \beta_1 Z + \beta_2 Z^2 + \varepsilon$ is a second-order (in the covariate Z) linear (in the parameters β_i, $i = 0, 1, 2$) regression model. The ε are called 'model errors' and are a random component reflecting the inaccuracy of the relationship between the variables which can never be exact due to e.g. measurement errors in the observations.

The history of linear models can be traced back to the early 19th century where Legendre was the first to introduce a linear model. The principle for the determination of the unknown parameters β_i, $i = 0, 1, 2$, was to minimize the sum of squares of the residuals $e = Y - \beta_0 - \beta_1 Z - \beta_2 Z^2$. Among the various approaches of performing regression, the least squares method is probably the most widely used.

Applications of linear regression analysis are numerous and occur in almost every field, including engineering, physical sciences, economics, management, life and biological science, and the social sciences. In this thesis, the main focus is on variables coming from the field of medicine, and more specifically, the interest will be on variables that are interval censored, that is, the response Y and the covariate Z are not observed directly but only known to lie in some interval $[Y_L, Y_R]$ and $[Z_L, Z_R]$, respectively.

Chapter 1 of this part of the thesis presents the statistical methods necessary for the development of the new regression theory. It contains an introduction of the theory for nonparametrically estimating the distribution function of interval censored variables, both in the one-dimensional case and the two-dimensional case. Furthermore, it introduces the regression method of Gómez et al. (2002) who proposed an approach for parameter estimation in linear models with exactly observed response and interval censored covariates. Their method will be extended in Chapter 2 when developing a new regression theory for the case that the response variable is interval censored as well. It uses a maximum likelihood approach for the estimation of the regression parameters while estimating at the same time the unknown distribution function of the interval censored covariate. The performance of the proposed method is assessed via a simulation study as described in Chapter 3. Finally, Chapter 4 contains a discussion of possible alternative approaches for the estimation of the regression parameters in the given context.

Contents

Chapter 1

Methods for interval censored variables

This chapter gives an overview of the methods used in the development of the new regression theory for interval censored data. It describes density estimation in the context of interval censored random variables as introduced by Turnbull (1976) for the one-dimensional case, and generalized by Betensky and Finkelstein (1999) for the two-dimensional case. Furthermore, the regression theory for linear models with observed response and interval censored covariate as proposed by Gómez et al. (2002) is presented. Their method will be extended later to the case that both covariate and response are interval censored.

1.1 Nonparametric estimation of the distribution of an interval censored variable

Suppose X to be a continuous, interval censored random variable with distribution function F and realizations x_i, $i = 1, \ldots, n$. Due to interval censoring, the x_i are not observed directly but only their respective censoring intervals $[x_{L_i}, x_{R_i}]$. These are known to include the true value x_i with probability one.

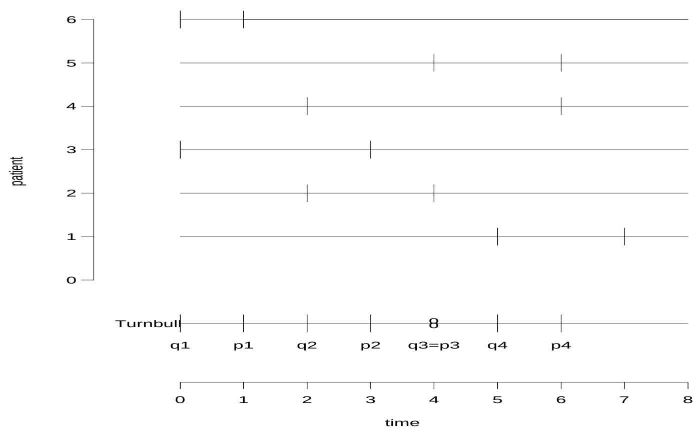
Turnbull (1976) proposed a maximum likelihood approach for determining an estimate for the distribution function F. It is a maximum likelihood approach which makes use of the equivalence between maximum likelihood estimates and self-consistent estimates as described in the following.

The construction of the likelihood for the data in the given context follows from the fact that the contribution of each individual i is $F(x_{R_i}) - F(x_{L_i})$, which results from X being interval censored. The complete likelihood accounting for all individuals is therefore given by

$$L(F) = \prod_{i=1}^{n} \left(F(x_{R_i}) - F(x_{L_i}) \right).$$

Maximizing this likelihood with respect to F would yield the maximum likelihood estimate for the distribution function of X. Turnbull shows that this maximization problem can be reduced to a simpler one: After sorting all observed interval endpoints x_{L_i} and x_{R_i} in ascending order, one constructs a set of disjoint intervals $[q_1, p_1]$,. . .,$[q_m, p_m]$ in the following way: Firstly, each $[q_j, p_j]$ must not contain any other member x_{L_i} or x_{R_i} except at their endpoints, and secondly, it must hold that $q_1 \leq p_1 < q_2 \leq \ldots < q_m \leq p_m$. An example for the construction of the Turnbull intervals $[q_j, p_j]$, $j = 1, \ldots, m$, is given in Figure 1.1. It shows six observed patient time intervals [0,1], [4,6], [2,6], [0,3], [2,4], [5,7] and the resulting Turnbull intervals [0,1], [2,3], [4,4], [5,6] obtained with the two construction rules given above.

Figure 1.1: Illustration of the construction of Turnbull's intervals



Turnbull proved that:

1. Any cumulative distribution function which increases outside the set $\bigcup_{j=1}^{m}[q_j, p_j]$ can not be a maximum likelihood estimate of F, and

2. for fixed values of $F(p_j+)$ and $F(q_j-)$, the likelihood is independent of the behavior of F within each interval $[q_j, p_j]$.

This means that it suffices to consider only those distribution functions which increase in some or all of the intervals $[q_j, p_j]$ and are constant outside these intervals. Furthermore, the behavior of the distribution function inside these intervals is not defined but can be imagined to be arbitrary. Thus, the problem of maximizing $L(F)$ reduces to that of maximizing

$$L(s_1, \ldots, s_m) = \prod_{i=1}^{n} \sum_{j=1}^{m} \alpha_{ij} s_j,$$

where $s_j = F(p_j+) - F(q_j-)$ with $\sum_{j=1}^{m} s_j = 1$, and $\alpha_{ij} = 1$ if $[q_j, p_j] \subseteq [x_{L_i}, x_{R_i}]$ and 0 otherwise. The meaning of the indicator α_{ij} is that only those individuals contribute to the likelihood, whose observed censoring intervals contain one or more Turnbull intervals. The estimate of the density of X is given through the weight vector $\mathbf{s} = (s_1, \ldots, s_m)$.

In order to determine the maximum likelihood estimate of $\mathbf{s}$, Turnbull proposed to apply an algorithm which is based on the equivalence between the maximum likelihood estimates and the self-consistent estimates and is described in the following. For details on the self-consistency equations see Efron, 1967.

Define I_{ij}=1 if $x_i \in [q_j, p_j]$ and 0 otherwise. Because of censoring the value of I_{ij} is not known, but its expectation is given by

$$E_s(I_{ij}) = \alpha_{ij} s_j = \mu_{ij}(\mathbf{s}).$$

That is, $\mu_{ij}(\mathbf{s})$ represents the probability that the i-th observation lies in $[q_j, p_j]$. Furthermore, the proportion of observations in the interval $[q_j, p_j]$ is

$$\sum_{i=1}^{n} \mu_{ij}(\mathbf{s})/M(\mathbf{s}) = \pi_j(\mathbf{s}),$$

where

$$M(\mathbf{s}) = \sum_{i=1}^{n} \sum_{j=1}^{m} \mu_{ij}(\mathbf{s}).$$

The self-consistent estimate of the s_j is then defined to be any solution of the simultaneous equation

$$s_j = \pi_j(s_1, \ldots, s_m).$$

Turnbull incorporates these formulas in an iterative procedure in order to derive the nonparametric estimate for the s_j:

Step 1: Chose initial estimates s_j^0, $j = 1, ..., m$. This can be any set of positive numbers summing to unity, e.g. $s_j = \frac{1}{m}$ for all j.

Step 2: Evaluate $\mu_{ij}(\mathbf{s^0})$, $M(\mathbf{s^0})$ and $\pi_j(\mathbf{s^0})$ using the formulas given above.

Step 3: Obtain the improved estimates s_j^1 by setting $s_j^1 = \pi_j(\mathbf{s^0})$.

Step 4: Return to Step B replacing $\mathbf{s^0}$ by $\mathbf{s^1}$.

Step 5: Stop when the values of $\mathbf{s^1}$ and $\mathbf{s^0}$ do not differ anymore.

Turnbull shows that the algorithm converges monotonely for those initial vectors $\mathbf{s^0}$ that are close to the true density vector $\mathbf{s}$. Gentleman and Geyer (1994) provide easily verifiable conditions for the self-consistent estimator to be a maximum likelihood estimator and for checking whether the maximum likelihood estimate is unique.

1.2 Nonparametric estimation of the distribution of two interval censored variables

Betensky and Finkelstein (1999) generalized Turnbull's estimation procedure to bivariate discrete interval censored data. Unlike Turnbull, the likelihood function is not maximized using a self-consistent algorithm, but an extension of the method of Gentleman and Geier (1994) is applied.

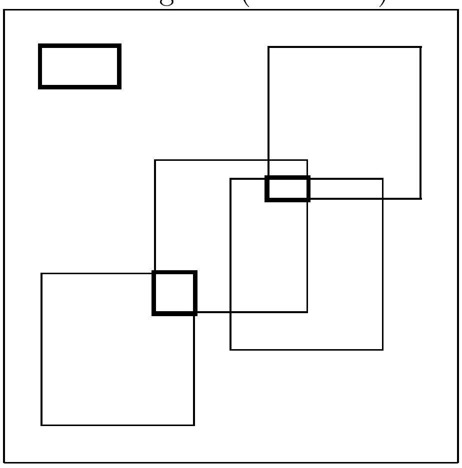
In the bivariate case, one observes for each individual i, $i = 1, \ldots, n$, the data rectangle $[x_{L_{1i}}, x_{R_{1i}}] \times [x_{L_{2i}}, x_{R_{2i}}]$ which are known to contain the realizations of X_{1i} and X_{2i}. Denoting $F(x_1, x_2)$ the joint cumulative distribution function of X_1 and X_2, the likelihood for the data in this setting is

$$\prod_{i=1}^{n} \left(F(x_{R_{i1}}+, x_{Ri2}+) - F(x_{R_{i1}}+, x_{L_{i2}}-)\right.$$

$$\left. -F(x_{L_{i1}}-, x_{R_{i2}}+) + F(x_{L_{i1}}-, x_{L_{i2}}-)\right).$$

Similar to the one-dimensional case, the support of the maximum likelihood estimate of F is contained in that set of rectangles which is formed by intersecting the observed data rectangles such that no other rectangle is contained within them. This mechanism is equivalent to the one used in the construction of the Turnbull intervals explained in the previous section. Figure 1.2 gives an illustration.

Figure 1.2: Final rectangles (thick lines), resulting from intersecting the observed regions (thin lines)



Denote the final rectangles as $[r_j, s_j] \times [t_j, u_j]$, $j = 1, ..., J$. Define furthermore the probability associated with rectangle j to be $p_j = F(s_j+, u_j+) - F(s_j+, t_j-) - F(r_j-, u_j+) + F(r_j-, t_j-)$. Then, adopting the argumentation of Turnbull (1976), the search for the maximum likelihood estimate for F can be restricted to those vectors $\mathbf{p} = (p_1, ..., p_J)$ having strictly non-negative components and summing to one. The maximum likelihood estimate eventually results from maximizing

$$L(\mathbf{p}) = \prod_{i=1}^{n} \sum_{j=1}^{J} \alpha_{ij} p_j,$$

where α_{ij} equals 1 if $[r_j, s_j] \subseteq [x_{L1_i}, x_{R1_i}]$ and $[t_j, u_j] \subseteq [x_{L2_i}, x_{R2_i}]$, and 0 otherwise.

Under the constraints for the p_j given above, the authors propose to maximize the likelihood $L(\mathbf{p})$ directly by solving a concave programming problem with linear constraints as described in Gentleman and Geier (1994).

1.3 Linear regression models with exactly observed response and interval censored covariate

Gómez et al. (2002) proposed a theory for linear regression analysis with interval censored covariates. The idea of their approach is to simultaneously maximize the data likelihood and estimate the unknown distribution function of the covariate.

The authors consider a continuous response variable Y with exactly observed realizations y_i, and a discrete and interval censored covariate Z whose realizations z_i are not observed but only the corresponding covariate intervals $[z_{L_i}, z_{R_i}]$, $i = 1, \ldots, n$. These intervals are known to include z_i with probability one. The model to be established is

$$Y = \alpha + \beta Z + \varepsilon, \qquad \textit{model } 1$$

where the error term ε is said to be independent of Z and normally distributed with expectation zero and variance σ^2. The aim is to estimate the parameter vector $\theta = (\alpha, \beta, \sigma^2)$ from the observed data $(y_i, [z_{L_i}, z_{R_i}])$.

Since Z is taken to be a discrete random variable, the authors suppose that it assigns positive mass w_j to the points s_j, $j = 1, \ldots, m$. From the normality of the model errors follows that the conditional density f of the response Y given s_j as a realization of Z is also normally distributed, with expectation $\alpha + \beta s_j$ and variance σ^2:

$$f(y|s_j;\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\alpha-\beta s_j)^2}{2\sigma^2}\right).$$

This density is used in the construction of the data likelihood

$$L(\theta, w_j) = \prod_{i=1}^{n} \sum_{j=1}^{m} \alpha_{ij} w_j f(y_i|s_j;\theta),$$

where

$$\alpha_{ij} = \begin{cases} 1 & : \quad s_j \in [z_{L_i}, z_{R_i}] \\ 0 & : \quad s_j \notin [z_{L_i}, z_{R_i}] \end{cases}, \quad \text{and} \quad w_j = P(Z = s_j).$$

Due to the unknown covariate distribution $\mathbf{w} = (w_1, ..., w_m)$, this likelihood can not be maximized directly to obtain the maximum likelihood estimates for the model parameters. Therefore, the authors maximize L simultaneously for θ and $\mathbf{w}$ using a two-step algorithm which first maximizes L with respect to $\mathbf{w}$ for fixed θ, and then resolves the maximization problem for θ with $\mathbf{w}$ known. These two steps are described in detail below.

1.3.1 Nonparametric estimation of w when θ is known

Assuming that the value for θ is known, the maximization of the likelihood L reduces to the problem of finding a vector $\mathbf{w}$ that maximizes

$$L^*(\mathbf{w}) = \prod_{i=1}^{n} \sum_{j=1}^{m} \alpha_{ij} w_j f(y_i|s_j),$$

subject to the constraints $\sum_{j=1}^{m} w_j = 1$ and $w_j \geq 0$ for all j.

The authors propose an algorithm for this maximization problem which is similar to Turnbull's density estimation procedure described in Chapter 1.1. It consists of the following steps: First, the authors fix a value for θ and chose start values for $\mathbf{w}$. With these, they calculate the probability ν_{ij} that the covariate of the i-th individual is equal to s_j. This quantity is then used to determine the expected number τ_j of individuals with $Z_i = s_j$. Finally,

τ_j is taken to be an improved estimate of the covariate density $\mathbf{w}$, and can later be used to recalculate ν_{ij} and τ_j. This procedure is repeated until the improved estimate and the old estimate are sufficiently close. The following scheme illustrates this estimation procedure:

Step 1a: Fix the value for θ using $\theta^0 = (\alpha^0, \beta^0, \sigma_0^2)$, where

$$\alpha^0 = \bar{y} - \frac{\beta^0}{n}\sum_{i=1}^{n}\hat{e}_i,$$

$$\beta^0 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})\hat{e}_i}{\sum_{i=1}^{n}(\hat{v_i^2} - \hat{e_i^2}) - (1/n)(\sum_{i=1}^{n}\hat{e}_i)^2},$$

$$n\sigma_0^2 = \sum_{i=1}^{n}(y_i - \alpha^0)^2 - (\beta^0)^2\sum_{i=1}^{n}(\hat{v_i^2} + \hat{e_i^2}),$$

and

$$\hat{e}_i = (x_{L_i} + x_{R_i})/2, \quad \text{and} \quad \hat{v_i^2} = ((x_{L_i} - \hat{e}_i)^2 + (x_{R_i} - \hat{e}_i)^2)/2.$$

Step 1b: Chose initial estimates for the w_j^0, for instance take $w_j^0 = \frac{1}{m}$.

Step 1c: Evaluate $\nu_{ij}(\theta, \mathbf{w^0})$ defined as

$$\nu_{ij} := P(X = s_j | y_i, [x_{L_i}, x_{R_i}]) = \frac{\alpha_{ij} f(y_i|s_j;\theta) w_j}{\sum_{k=1}^{m}\alpha_{ik} f(y_i|s_k;\theta) w_k},$$

replacing w_j by w_j^0, and calculate

$$\tau_j(\theta, \mathbf{w^0}) = \frac{1}{n}\sum_{i=1}^{n}\nu_{ij}(\theta, \mathbf{w^0}).$$

Step 1d: Obtain the improved estimate w_j^1 by setting $w_j^1 = \tau_j(\theta, \mathbf{w^0})$.

Step 1e: Return to step 1c replacing $\mathbf{w^0}$ by $\mathbf{w^1}$.

Step 1f: Repeat steps 1c to 1e until the value of $\mathbf{w^1}$ does not change anymore. Denote it by $\hat{\mathbf{w}}^{\mathbf{1}}$.

1.3.2 Maximum likelihood estimation of θ when $\mathbf{w}$ is known

When the covariate density $\mathbf{w}$ is known, the maximization of the likelihood

$$L^{**}(\theta) = \prod_{i=1}^{n}\sum_{j=1}^{m} \alpha_{ij} w_j f(y_i|\theta)$$

with respect to θ can be achieved via the usual maximum likelihood approach: The logarithm of L^{**} is derived with respect to α, β and σ^2, and these derivations are set to zero and solved for the parameters. The authors show that the solution of the maximum likelihood equations $\frac{\partial}{\partial\theta} log L^{**}(\theta) = 0$ is

$$\hat{\alpha} = \bar{y} - \frac{\beta}{n}\sum_{i=1}^{n} e_i(\theta, \mathbf{w}), \tag{1}$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(y_i - \alpha)e_i(\theta, \mathbf{w})}{\sum_{i=1}^{n}(v_i(\theta, \mathbf{w}) + e_i^2(\theta, \mathbf{w}))}, \tag{2}$$

$$n\hat{\sigma}^2 = \sum_{i=1}^{n}(y_i - \alpha)^2 - \beta^2 \sum_{i=1}^{n}(v_i(\theta, \mathbf{w}) + e_i^2(\theta, \mathbf{w})), \tag{3}$$

where

$$e_i(\theta, \mathbf{w}) = \frac{\sum_{k=1}^{m} \alpha_{ik} s_k w_k \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{1}{2\sigma^2}(y_i - \alpha - \beta s_k)^2\}}{\sum_{k=1}^{m} \alpha_{ik} w_k \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{1}{2\sigma^2}(y_i - \alpha - \beta s_k)^2\}}, \tag{4}$$

and

$$v_i(\theta, \mathbf{w}) = \frac{\sum_{k=1}^{m} \alpha_{ik}(s_k - exp_i(\theta, w))^2 w_k \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{1}{2\sigma^2}(y_i - \alpha - \beta s_k)^2\}}{\sum_{k=1}^{m} \alpha_{ik} w_k \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{1}{2\sigma^2}(y_i - \alpha - \beta s_k)^2\}}. \tag{5}$$

The algorithm proposed by the authors maximizes L^{**} by first choosing initial values for equations (4) and (5), which are then used to calculate the estimates given in (1) to (3). Afterwards, (4) and (5) are determined again using the newly calculated estimates and the covariate density vector that resulted from the algorithm of the previous section. This procedure is repeated until the values for $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\sigma}^2$ stabilize. The following scheme illustrates the estimation process.

Step 2a: Calculate θ^0 from formulas (1) to (3) by choosing the initial values for $e_i(\theta, \mathbf{w})$ and $v_i(\theta, \mathbf{w})$ to be

$$e_i^0(\theta, \mathbf{w}) = \frac{z_{L_i} + z_{R_i}}{2} \quad \text{and}$$

$$v_i^0(\theta, \mathbf{w}) = \frac{(z_{L_i} - e_i^0)^2 + (z_{R_i} - e_i^0)^2}{2}.$$

Step 2b: Evaluate $e_i(\theta^0, \hat{\mathbf{w}}^{\mathbf{1}})$ and $v_i(\theta^0, \hat{\mathbf{w}}^{\mathbf{1}})$ using equations (4) and (5) employing θ^0 and $\hat{\mathbf{w}}^{\mathbf{1}}$ from step 1f above.

Step 2c: Obtain the improved estimate θ^1 from formulas (1) to (3), replacing $e_i(\theta, \mathbf{w})$ and $v_i(\theta, \mathbf{w})$ by $e_i(\theta^0, \hat{\mathbf{w}}^{\mathbf{1}})$ and $v_i(\theta^0, \hat{\mathbf{w}}^{\mathbf{1}})$.

Step 2d: Return to step 2a replacing θ^0 by θ^1.

Step 2e: Repeat steps 2a to 2c until the difference between θ^0 and θ^1 is sufficiently small. Denote the final estimate by $\hat{\theta}^1$.

In total, the two-step algorithm for calculating simultaneously the density $\mathbf{w}$ of the interval censored covariate and the estimator for the parameter vector θ, results in the combination of the two algorithms given above and is summarized in the following scheme:

Step I: Execute Step 1a up to Step 1f.
Step II: Execute Step 2a up to Step 2e.
Step III: Return to Step 1c replacing $\theta^{\mathbf{0}}$ by $\hat{\theta}^{\mathbf{1}}$ and $\mathbf{w}^{\mathbf{0}}$ by $\hat{\mathbf{w}}^{\mathbf{1}}$.
Step IV: Repeat steps I to III until convergence of θ and $\mathbf{w}$.

Chapter 2

Linear regression with interval censored response and covariate

This chapter presents a new estimation theory for linear regression models when both covariate and response are interval censored. It is an extension of the method of Gómez et al. (2002) introduced previously. The model to be considered here is

$$Y_i = \alpha + \beta Z_i + \varepsilon_i, \qquad i = 1, \ldots, n \qquad \textit{model } 2$$

where the response Y_i is continuous and censored into the interval $[Y_{L_i}, Y_{R_i}]$, and the covariate Z_i is discrete and censored into the interval $[Z_{L_i}, Z_{R_i}]$. The model errors ε are assumed to have a normal distribution with mean zero and variance σ^2.

Let s_j be the possible values for Z with corresponding weights w_j, $j = 1, \ldots, m$, and denote the covariate density and distribution function as $\mathbf{w}$ and $\mathbf{W}$, respectively. From the errors' normal distribution follows that the distribution of Y given s_j as a value of Z is also normal with mean $\alpha + \beta s_j$ and variance σ^2:

$$f(y|s_j, \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(y - \alpha - \beta s_j)^2}{2\sigma^2}\right).$$

Here, $\theta = (\alpha, \beta, \sigma^2)$ is the vector of the model parameters which we want to estimate.

It will be assumed that the interval censoring for the covariate and the response occurs noninformatively. If a variable X is subject to noninformative censoring, this means that for any given values x_0, x_1, x_2, the conditional density of this variable is the same as the density of the uncensored variable truncated into the observed censoring interval:

$$P(X = x_0 | X_L = x_1, X_R = x_2) = \begin{cases} \frac{P(X=x_0)}{P(X \in [x_1,x_2])} & : \quad x_0 \in [x_1, x_2] \\ 0 & : \quad \text{otherwise} \end{cases} .$$

Gómez et al. (2001b) show that the contribution to the likelihood of an unique individual with observed censoring interval $[x_L, x_R]$ which includs the true value of interest x, is proportional to $\int_{x_L}^{x_R} dW(x)$ where $W = P(X \leq x)$. With this fact, the likelihood for the observed data of *model* 2 can be constructed as given in the next section.

2.1 Estimation procedure

The observed data for *model* 2 consists of n independent and identically distributed realizations of Y and Z. Since these two variables are interval censored, one observes the intervals $([y_{L_i}, y_{R_i}], [z_{L_i}, z_{R_i}])$, $i = 1, ..., n$. In order to obtain the estimates for the model parameters α, β and σ^2, a maximum likelihood approach will be proposed as described in the following.

The likelihood for the observed data can be constructed by noting the following facts: The contribution of an arbitrary individual i to the likelihood consists of the contribution of this individual with respect to both the covariate and the response. Since the covariate Z is interval censored, its density must be estimated with a method similar to the one given in Turnbull (1976), yielding as a result the weights w_j (for more details on the method of Turnbull see Chapter 1.1). Thus, the contribution of individual i with respect to Z is $\sum_{j=1}^{m} \alpha_{ij} w_j$, where the indicator variable α_{ij} specifies whether or not the covariate value s_j is contained in the observed covariate interval $[z_{L_i}, z_{R_i}]$. On the other hand, the contribution of this individual with respect to the response Y given a fixed value of Z, is determined by the conditional density $f(y|s_j, \theta)$. Since the value of Y is not exactly observed but only its censoring interval $[y_{L_i}, y_{R_i}]$, the conditional density must be integrated over the range of this censoring interval in order to obtain the respective contribution to the likelihood. The total contribution of individual i to the likelihood is then the combination of these two single likelihood contributions, and the complete likelihood accounting for all individuals is therefore given by

$$L(\theta, w_j) = \prod_{i=1}^{n} P(Y_i \in [Y_{L_i}, Y_{R_i}], Z_i \in [Z_{L_i}, Z_{R_i}])$$

$$= \prod_{i=1}^{n} \sum_{j=1}^{m} \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} f(y|s_j; \theta) dy, \tag{2.1}$$

where $\alpha_{ij} = \begin{cases} 1 & : \quad s_j \in [z_{L_i}, z_{R_i}] \\ 0 & : \quad \text{otherwise} \end{cases}$,

and $w_j = P(Z = s_j)$ is the weight the covariate assigns to the point s_j.

The estimation of the parameter vector θ will be achieved through maximizing L. Similar as in the context of the regression theory of Gómez et al. (2002), this maximization can not be carried out directly because of the unknown covariate density function $\mathbf{w} = (w_1, \ldots, w_m)$. Thus, L is maximized through an algorithm that iterates between maximizing L with respect to $\mathbf{w}$ while holding θ fixed, and maximizing L with respect to θ while holding $\mathbf{w}$ fixed. These two steps are described in detail below.

Nonparametric estimation of w when θ is known

For a fixed value of θ, the maximum likelihood estimate of the vector $\mathbf{w}$, given the constraints $\sum_{j=1}^{m} w_j = 1$ and $w_j \geq 0$ for all j, is determined by using a procedure based on the equivalence between the maximum likelihood and the self-consistent estimators as explained in Turnbull (1976): First, initial values for the covariate density weights w_j, are chosen. With these, the conditional probabilities ν_{ij} that the covariate Z_i equals a given value s_j are calculated. Summing these probabilities over all individuals i leads to the expected number τ_j of individuals with a covariate value equal to s_j. This expected number is then taken to be an improved estimate of the covariate density $\mathbf{w}$, and can be used to recalculate ν_{ij} and τ_j. The whole procedure is repeated until the difference of the values of the improved and the old estimate is sufficiently small. The following scheme gives a summary:

Step A1 Take initial estimates for the w_j^0, for example $w_j^0 = \frac{1}{m}$ for $j = 1, \ldots, m$. Denote $\mathbf{w^0} = (w_1^0, ..., w_m^0)$.

Step A2 Evaluate $\nu_{ij}(\mathbf{w^0}, \theta)$ and $\tau_j(\mathbf{w^0}, \theta)$ defined as

$$\nu_{ij}(\mathbf{w^0}, \theta) = P(Z_i = s_j | [z_{L_i}, z_{R_i}], [y_{L_i}, y_{R_i}])$$

$$= \frac{\alpha_{ij} w_j^0 \int_{y_{L_i}}^{y_{R_i}} f(y|s_j; \theta)}{\sum_{j=1}^m \alpha_{ij} w_j^0 \int_{y_{L_i}}^{y_{R_i}} f(y|s_j; \theta)},$$

$$\tau_j(\mathbf{w^0}, \theta) = \frac{1}{n} \sum_{i=1}^n \nu_{ij}(\mathbf{w^0}, \theta).$$

Step A3 Obtain the improved estimate $\mathbf{w^1}$ setting

$$\mathbf{w^1} = \tau_j(\mathbf{w^0}, \theta).$$

Step A4 Go to step A2 replacing $\mathbf{w^0}$ by $\mathbf{w^1}$ and repeat the whole procedure until their values are sufficiently close.

Maximum likelihood estimation of θ when w is known

When the covariate density is known, the maximization of the likelihood L with respect to θ can be achieved by solving the score equation $\frac{\partial}{\partial \theta} log L = 0$. The resulting estimates for α, β and σ^2 are derived in Appendix A. They are calculated to

$$\hat{\beta} = \frac{\bar{d} - \bar{a}\bar{b}}{\bar{c} - \bar{b}^2}, \tag{2.2}$$

$$\hat{\alpha} = \bar{a} - \hat{\beta}\bar{b}, \tag{2.3}$$

$$\hat{\sigma}^2 = \bar{e} - 2\hat{\alpha}\bar{a} + \hat{\alpha}^2 - \hat{\beta}^2\bar{c}, \tag{2.4}$$

where $\bar{a}$, $\bar{b}$, $\bar{c}$, $\bar{d}$ and $\bar{e}$ is the average of a_i, b_i, c_i, d_i and e_i, $i = 1, ..., n$, respectively, defined as

$$a_i = E(Y_i | [Z_{Li}, Z_{Ri}], [Y_{Li}, Y_{Ri}]),$$

$$b_i = E(Z_i | [Z_{Li}, Z_{Ri}], [Y_{Li}, Y_{Ri}]),$$

$$c_i = E(Z_i^2|[Z_{Li}, Z_{Ri}], [Y_{Li}, Y_{Ri}]),$$
$$d_i = E(Z_iY_i|[Z_{Li}, Z_{Ri}], [Y_{Li}, Y_{Ri}]),$$
$$e_i = E(Y_i^2|[Z_{Li}, Z_{Ri}], [Y_{Li}, Y_{Ri}]).$$

The following propositions show that the estimates $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\sigma}^2$ are similar to the maximum likelihood estimators in a simple linear model with exactly observed response and covariate.

Proposition 1
It holds that $\hat{\beta}$ as defined above converges in probability to the value $\frac{Cov(Z,Y)}{Var(Z)}$.

Proof
Applying the law of large numbers, it holds that

$$\bar{a} = \frac{1}{n}\sum_{i=1}^{n} a_i \stackrel{n\to\infty}{\longrightarrow} E(a_i) = E(E(Y_i|[Z_{L_i}, Z_{R_i}], [Y_{L_i}, Y_{R_i}])) = E(Y_i),$$

$$\bar{b} = \frac{1}{n}\sum_{i=1}^{n} b_i \stackrel{n\to\infty}{\longrightarrow} E(b_i) = E(E(Z_i|[Z_{L_i}, Z_{R_i}], [Y_{L_i}, Y_{R_i}])) = E(Z_i),$$

$$\bar{c} = \frac{1}{n}\sum_{i=1}^{n} c_i \stackrel{n\to\infty}{\longrightarrow} E(c_i) = E(E(Z_i^2|[Z_{L_i}, Z_{R_i}], [Y_{L_i}, Y_{R_i}])) = E(Z_i^2),$$

and

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i \stackrel{n\to\infty}{\longrightarrow} E(d_i) = E(E(Z_iY_i|[Z_{L_i}, Z_{R_i}], [Y_{L_i}, Y_{R_i}])) = E(Z_iY_i).$$

Thus, it holds for the numerator of $\hat{\beta}$ that

$$\bar{d} - \bar{a}\bar{b} =\stackrel{n\to\infty}{\longrightarrow} E(ZY) - E(Y)E(Z) = Cov(Z,Y),$$

and for the denominator that

$$\bar{c} - \bar{b}^2 \stackrel{n\to\infty}{\longrightarrow} E(Z^2) - E(Z)^2 = Var(Z,Y).$$

In total, this means that

$$\hat{\beta} = \frac{\bar{d} - \bar{a}\bar{b}}{\bar{c} - \bar{b}^2} \stackrel{n\to\infty}{\longrightarrow} \frac{Cov(Z,Y)}{Var(Y)}. \qquad \square$$

Proposition 2
It holds that $\hat{\alpha}$ as defined above converges in probability to the value $E(Y) - \beta E(Z)$.

Proof
Applying the law of large numbers it holds that

$$\bar{a} = \frac{1}{n}\sum_{i=1}^{n} a_i \overset{n\to\infty}{\longrightarrow} E(a_i) = E(E(Y_i|[Z_{L_i}, Z_{R_i}], [Y_{L_i}, Y_{R_i}])) = E(Y_i)$$

and

$$\bar{b} = \frac{1}{n}\sum_{i=1}^{n} b_i \overset{n\to\infty}{\longrightarrow} E(b_i) = E(E(Z_i|[Z_{L_i}, Z_{R_i}], [Y_{L_i}, Y_{R_i}])) = E(Z_i).$$

Thus, together with Proposition 1, this means that

$$\hat{\alpha} = \bar{a} - \hat{\beta}\bar{b} \overset{n\to\infty}{\longrightarrow} E(Y) - \beta E(Z). \qquad \square$$

Proposition 3
It holds that $\hat{\sigma}^2$ as defined above converges in probability to the value $Var(Y) - \beta Var(Z)$.

Proof
Again, with the law of large numbers and Proposition 1, one obtains

$$\begin{aligned}
\hat{\sigma}^2 &\overset{n\to\infty}{\longrightarrow} E(Y^2) - 2\hat{\alpha}E(Y) + \hat{\alpha}^2 - \beta^2 E(Z^2) \\
&= E(Y^2) - 2\left(E(Y) - \beta E(Z)\right) E(Y) \\
&\quad + \left(E(Y) - \beta E(Z)\right)^2 - \beta^2 E(Z^2) \\
&= E(Y^2) - E(Y)^2 - \beta^2 \left(E(Z^2) - E(Z)^2\right) \\
&= Var(Y) - \beta^2 Var(Z). \qquad \square
\end{aligned}$$

For the determination of the parameter estimates of *model* 2, a procedure is proposed that uses start values for $\bar{a}$ to $\bar{e}$. It iterates between calculating $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\sigma}^2$ and re-determining the values for $\bar{a}$ to $\bar{e}$ as explained in the scheme given below.

Step B1 Take initial estimates for a_i, b_i, c_i, d_i and e_i, for example

$$a_i^0 = \frac{y_{L_i} + y_{R_i}}{2},$$

$$b_i^0 = \frac{z_{L_i} + z_{R_i}}{2},$$

$$c_i^0 = \frac{(z_{L_i} - b_i^0)^2 + (z_{R_i} - b_i^0)^2}{2},$$

$$d_i^0 = \frac{(z_{L_i} - b_i)(y_{L_i} - a_i) + (z_{R_i} - b_i)(y_{R_i} - a_i)}{2},$$

$$e_i^0 = \frac{(y_{L_i} - a_i^0)^2 + (y_{R_i} - a_i^0)^2}{2}.$$

Step B2 Use these values in (2.2) to (2.4) to compute the initial estimate $\theta^0 = (\hat{\alpha}^0, \hat{\beta}^0, \hat{\sigma}_0^2)$.

Step B3 Re-evaluate a_i up to e_i with their theoretical formulas given in Appendix A by employing θ_0.

Step B4 Obtain the improved estimate θ^1 by solving equations (2.2) to (2.4).

Step B5 Go to step B3 substituting θ^0 by θ^1.

Step B6 Cycle steps B3 to B5 until the difference between the values of θ^0 and θ^1 is sufficiently small.

The complete algorithm to obtain the joint maximum likelihood estimate for $\mathbf{w}$ and θ follows from the combination of the two conditional algorithms given above. It has been implemented in the program *semipara.cpp* and can be found on the floppy disc. The criteria for convergence of the estimates was chosen to be the relative norm differences of the estimates at iteration stage l:

$$\frac{||\hat{w}^{l-1} - \hat{w}^l||}{||\hat{w}^{l-1}||} \quad \text{and} \quad \frac{||\hat{\theta}^{l-1} - \hat{\theta}^l||}{||\hat{\theta}^{l-1}||}.$$

The estimates were defined to converge if the respective relative norm difference was less than 0.001. A flow-chart of the structure of this program is given in Chapter 3.

2.2 Confidence intervals for the model parameters

The MAPLE program given in Appendix B can be used to construct approximate confidence intervals for the parameter estimates resulting from the newly proposed estimation procedure. It uses the observed information matrix and quantiles of the normal distribution, and the different steps in the calculation process of the program are explained in the following:

Consider a given data set which consists of values y_{L_i} and y_{R_i} for the observed response intervals, values s_j for the discrete covariate with respective density weights w_j, and the estimated regression parameters $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\sigma}^2$. The first part of the program reads this data into variables. With these, the log-likelihood as defined in equation (2.1) is constructed and its first and second derivatives with respect to the regression parameters are calculated. Then, the Hessian matrix is formed from all second derivatives and the observed information matrix is calculated by multiplying the Hessian with minus one. Eventually, the inversion of the observed information matrix provides an estimate for the variances of $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\sigma}^2$. These estimated variances are then employed in the construction of the approximate confidence intervals.

2.3 Multiple regression

This section extends the proposed regression theory to the case that *model* 2 additionally incorporates an exactly observed covariate vector. This means, the model now under consideration is

$$Y = \alpha + \vec{\beta}_1'\vec{X} + \beta_2 Z + \varepsilon,$$

where $\vec{X} = (X_1, \ldots, X_p)$ is a vector of exactly observed covariates, $\vec{\beta}_1'$ is the corresponding p-dimensional parameter vector, Y is the interval censored response, Z is an interval censored covariate, and ε is a continuous $N(0, \sigma^2)$ random variable independent of $\vec{X}$ and Z.

The observed data for individual i is then $\vec{x}_i = (x_{1_i}, \ldots, x_{p_i})'$, $[z_{L_i}, z_{R_i}]$ and $[y_{L_i}, y_{R_i}]$. By defining $\theta = (\alpha, \vec{\beta}_1', \beta_2, \sigma^2)$ and using the notation and assumptions of *model* 2, the likelihood function in the new context is given as

$$L_n^*(\mathbf{w}, \theta) = \prod_{i=1}^{n} \sum_{j=1}^{m} \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} f(y|(\vec{x}_i, s_j); \theta)), \tag{2.5}$$

where $\mathbf{w} = (w_1, \ldots, w_m)$, $w_j = P(Z = s_j)$, $\alpha_{ij} = I\{s_j \in [z_{L_i}, z_{R_i}]\}$ and $f(y|(\vec{x}_i, s_j); \theta))$ is the conditional density of Y given $(\vec{X} = \vec{x}_i, Z = s_j)$:

$$f(y|(\vec{x}_i, s_j); \theta)) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \alpha - \vec{\beta}_1' \vec{x}_i - \beta_2 s_j)^2}{2\sigma^2}\right)^2.$$

The idea of the estimation procedure for the model parameters α, $\vec{\beta}_1'$, β_2 and σ^2 is the same as for *model* 2, only that the likelihood function is now given by (2.5). This means, L^* is maximized simultaneously for $\mathbf{w}$ and θ by cycling between steps A und B of the earlier proposed algorithm. In the present context, Step A now consists of the same self-consistent equations as given earlier but using the new expression for $\nu_{ij}(\mathbf{w}, \theta)$, which is

$$\nu_{ij}(\mathbf{w}, \theta) = P(Z_i = s_j|[y_{L_i}, y_{R_i}], [z_{L_i}, z_{R_i}], \vec{x}_i) = \frac{\alpha_{ij} w_j \int_{y_{L_i}}^{y_{R_i}} f(y|(\vec{x}_i, s_j); \theta)}{\sum_{j=1}^{m} \alpha_{ij} w_j \int_{y_{L_i}}^{y_{R_i}} f(y|(\vec{x}_i, s_j); \theta)}.$$

Step B is modified in so far that it now incorporates the maximum likelihood estimators resulting from the new context of the multiple regression. These are obtained from maximizing the logarithm of likelihood (2.5) for fixed $\mathbf{w}$ and are derived in Appendix C.

2.4 Model errors coming from the exponential family or Weibull distribution

In the previous sections, the regression parameters were estimated assuming the model errors to be normally distributed with mean zero and variance σ^2. The normal distribution is known to be a member of the so-called exponential family of distributions, which is defined in the following way:

Definition
Let X be a random variable with density function f determined by the parameter vector η. One says that f belongs to the exponential family of distributions if it can be expressed as

$$f(x;\eta) = h(x)c(\eta)exp[Q(\eta)t(x)],$$

where $Q(\eta)$ and $t(x)$ are vectors of common dimension k such that $Q(\eta)t(x) = \sum_{i=1}^{k} Q_i(\eta)t_i(x)$.

For example, the $N(0,\sigma^2)$-distribution is obtained when taking $h(x) = 1$, $c(\eta) = (2\pi\sigma^2)^{-1/2}$, $Q(\eta) = (0, \frac{1}{2\sigma^2})$ and $t(x) = (x, -x^2)$. Other members of the exponential family are the gamma, binomial and Poisson distribution.

In what follows it will be shown that the proposed regression theory still holds when the model errors come from any distribution which is a member of the exponential family. This means that the likelihood to be considered now is

$$L^{**}(\theta, w_j) = \prod_{i=1}^{n} \sum_{j=1}^{m} \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} f(y_i|s_j;\theta)dy,$$

where

$$f(y_i|s_j;\theta) = h(y_i - \alpha - \beta s_j)c(\eta)exp[Q(\eta)t(y_i - \alpha - \beta s_j)]$$

and $\theta = (\alpha, \beta, \eta)$.

The proceeding for obtaining the maximum likelihood estimate for θ in the new context is the same as in the original setting, namely maximizing the logarithm of L^{**} with respect to the parameters α, β and η. The resulting partial derivatives are given in Appendix D. It can be shown that the solutions $(F1) - (F3)$ of Appendix D include equations $(E1) - (E3)$ for normally distributed ε:

Corollary 1
Equation (F1) reduces to equation (E1) when the model errors are normal.

Proof
When the ε are normally distributed, it holds that $h(\varepsilon_i) = 1$, $c(\eta) = \frac{1}{\sqrt{2\pi\sigma^2}}$,

$Q(\eta) = -\frac{1}{2\sigma^2}$ and $t(\varepsilon_i) = (y_i - \alpha - \beta s_j)^2$. With that, equation (F1) results to

$$(F1) = \sum_{i=1}^{n} \left(\frac{\sum_{j=1}^{m} \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} -\frac{h'(\varepsilon_i)}{h(\varepsilon_i)} f(y_i|s_j;\theta)dy}{C_i(\theta)} \right.$$

$$\left. - \frac{\sum_{j=1}^{m} \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} f(y_i|s_j;\theta)[Q(\eta)t'(\varepsilon_i)]dy}{C_i(\theta)} \right)$$

$$= \sum_{i=1}^{n} \left(\frac{\sum_{j=1}^{m} \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} -\frac{0}{1} f(y_i|s_j;\theta)dy}{C_i(\theta)} \right.$$

$$\left. - \frac{\sum_{j=1}^{m} \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} f(y_i|s_j;\theta)[-\frac{1}{2\sigma^2}(-2)(y_i - \alpha - \beta s_j)]dy}{C_i(\theta)} \right)$$

$$= \frac{\sum_{j=1}^{m} \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} f(y_i|s_j;\theta)[\frac{1}{\sigma^2}(y_i - \alpha - \beta s_j)]dy}{C_i(\theta)} = (E1) \qquad \square$$

Corollary 2
Equation (F2) reduces to equation (E2) when the model errors are normal.

Proof
When the ε are normally distributed, it holds that $h(\varepsilon_i) = 1$, $c(\eta) = \frac{1}{\sqrt{2\pi\sigma^2}}$, $Q(\eta) = -\frac{1}{2\sigma^2}$ and $t(\varepsilon_i) = (y_i - \alpha - \beta s_j)^2$. With that, equation (F2) results to

$$(F2) = \sum_{i=1}^{n} \left(\frac{\sum_{j=1}^{m} \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} -\frac{s_j h'(\varepsilon_i)}{h(\varepsilon_i)} f(y_i|s_j;\theta)dy}{C_i(\theta)} \right.$$

$$\left. - \frac{\sum_{j=1}^{m} \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} f(y_i|s_j;\theta)[Q(\eta)s_j t'(\varepsilon_i)]dy}{C_i(\theta)} \right)$$

$$= \sum_{i=1}^{n} \left(\frac{\sum_{j=1}^{m} \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} -\frac{s_j 0}{1} f(y_i|s_j;\theta)dy}{C_i(\theta)} \right.$$

$$- \frac{\sum_{j=1}^{m} \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} f(y_i|s_j;\theta)[-\frac{1}{2\sigma^2}(-2)(y_i-\alpha-\beta s_j)s_j]dy}{C_i(\theta)} \Bigg)$$

$$= \frac{\sum_{j=1}^{m} \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} f(y_i|s_j;\theta)[\frac{1}{\sigma^2}(y_i-\alpha-\beta s_j)s_j]dy}{C_i(\theta)} = (E2) \qquad \square$$

Corollary 3
Equation (F3) reduces to equation (E3) when the model errors are normal.

Proof
When the ε are normally distributed, it holds that $h(\varepsilon_i) = 1$, $c(\eta) = \frac{1}{\sqrt{2\pi\sigma^2}}$, $Q(\eta) = -\frac{1}{2\sigma^2}$ and $t(\varepsilon_i) = (y_i - \alpha - \beta s_j)^2$. With that, equation (F3) results to

$$(F3) = \sum_{i=1}^{n} \Bigg(\frac{\sum_{j=1}^{m} \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} -\frac{c'(\eta)}{c(\eta)} f(y_i|s_j;\theta)dy}{C_i(\theta)}$$

$$+ \frac{\sum_{j=1}^{m} \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} f(y_i|s_j;\theta)[Q'(\eta)t(\varepsilon_i)]dy}{C_i(\theta)} \Bigg)$$

$$= \sum_{i=1}^{n} \frac{\sum_{j=1}^{m} \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} \left(-\frac{c'(\eta)}{c(\eta)} + [Q'(\eta)t(\varepsilon_i)]\right) f(y_i|s_j;\theta)dy}{C_i(\theta)}$$

$$= \sum_{i=1}^{n} \frac{\sum_{j=1}^{m} \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} \left(-\frac{1}{\sigma^2} + [\frac{1}{\sigma^4}(y_i-\alpha-\beta s_j)^2]\right) f(y_i|s_j;\theta)dy}{C_i(\theta)} = (E3)\square$$

2.4.1 Weibull distribution

The proposed regression theory can also be applied when the model errors come from the Weibull distribution, as will be shown in the following. The likelihood of the data in this context is

$$L^{***}(\theta, w_j) = \prod_{i=1}^{n} \sum_{j=1}^{m} \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} f(y_i|s_j;\theta)dy,$$

where

$$f(y_i|s_j;\theta) = \alpha\beta s_j^{\beta-1} \exp\left(-\alpha s_j^{\beta}\right)$$

and $\theta = (\alpha, \beta)$.

Setting the partial derivatives of $l^{***} = log\ L^{***}$ to zero and solving for α and β yields the maximum likelihood estimates given in Appendix E.

Chapter 3

Simulations

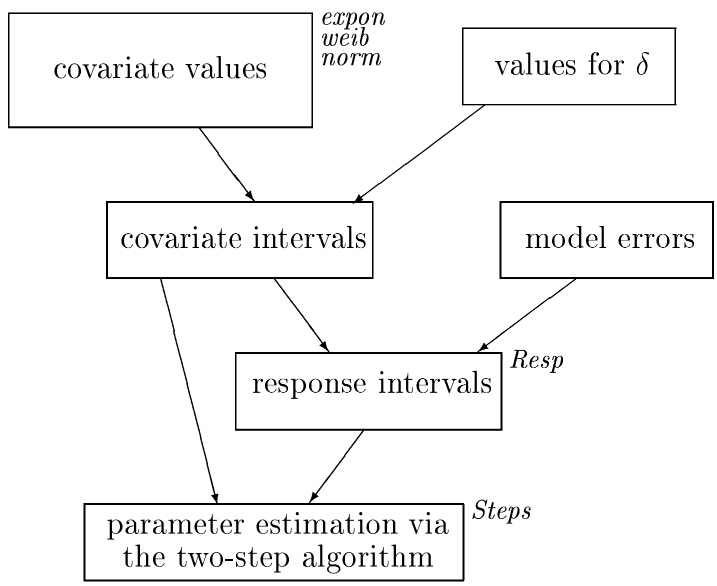
Since theoretical results for the goodness of the proposed estimates are difficult to obtain, their performance is checked through a simulation study. It involves different data scenarios for *model* 2 with the aim to assess to what extend the proposed parameter estimates are able to reflect these data situations. Table 3.1 shows the simulation scenarios used in the study.

Table 3.1: Scenarios for the simulation study

number of observations	200 and 500
covariate distributions	$Exp(\frac{1}{8})$, $Weib(\frac{1}{6}, \frac{3}{2})$, $N(6, 4)$
percentage of censoring	0.3 and 0.7
value for α	4
values for β	2 and 5
value for σ^2	1

The simulations are carried out by the program *semipara.cpp* on the floppy disc, and a short summary of how this program works is given now: The model errors ε are generated from a $N(0, 1)$-distribution, and the values for the covariate Z are simulated from the exponential, Weibull or normal distribution. These values are used to construct the covariate intervals $[Z_L, Z_R]$ after the following scheme: Depending on the covariate distribution, there is a certain number of values j, $j = 1, \ldots, k$, which the covariate can take on. An indicator variable δ_{ij} determines with a given probability p, whether or not the covariate for individual i is observed at value j. Then, one looks at each value z_i and goes back to the nearest observed value j and takes it as the value for z_{L_i}. Similarly, z_{R_i} is that observed value j coming first after z_i. The corresponding response intervals $[y_{L_i}, y_{R_i}]$ result from the formulas $y_{L_i} = \alpha + \beta z_{L_i} + \varepsilon_i$ and $y_{R_i} = \alpha + \beta z_{R_i} + \varepsilon_i$. Eventually, the two-step algorithm described in Chapter 2 is applied to the generated response and covariate intervals for the estimation of the model parameters α, β and σ^2.

The following flow-chart illustrates the simulation process of the program *semipara.cpp*. The steps of the program are written inside the boxes and the arrows indicate which step enters in the calculation of another step. As most calculations are executed by procedures within the program, their names are written outside the corresponding box which will make it easier to find one's way when looking at the code of the program.



Other procedures used in this program are listed below together with a short description of their usage:

FileOpen: opens all files needed for reading and writing.
Spalloc: allocates memory for the vectors and matrices.
ran2: generates random uniform variates.
Simpson: integrates an user-defined function applying Simpson's method.

The last mentioned procedure *Simpson* is used for the calculation of different integrals over the conditional density $f(y|s_j;\theta)$ which is needed among others in the calculation of the conditional means a_i to e_i given in Chapter 2. As these integrals cannot be calculated analytically in C, a numerical approximation applying Simpson's method is used. The idea of Simpson's method is to approximate the area under a given graph by a sequence of quadratics. That is, the range of the upper and lower interval limit is divided into an even number of subintervals and their width is calculated. Then, the function value at the left endpoints of the first three subintervals in calculated as well as the area of the parabola through these three points. This process is repeated moving two subintervals to the right. Simpson's method is said to be the most exact among those existing for numerical integration. Though, it is obviously not as exact as the analytical form. This must be taken into consideration when assessing the simulation results of the estimates.

The performance of the program *semipara.cpp* with respect to speed and convergence is highly satisfying. Running it on a 400 megahertz Pentium II processor with 128 MB RAM main memory using the SUSE LINUX 7.1 operating system yielded convergence of the parameter estimates after 5 to 30 iterations depending on the number of observations and the level of censoring. The time needed for the calculations varied between 5 and 60 seconds.

3.1 Simulation theory

The simulation study involves the generation of data coming from different statistical distributions. The theory applied for the generation of these distributions is given now (for references see Box and Müller, 1958, or Morgan, 1984).

Uniform distribution
For the generation of a Uniform(0,1) random variable, a *Congruential Pseudo-Random Number Generator* is used. By applying the recursion formula $x_{n-1} = ax_n + b \; mod \; m$ with seed x_0 and a, b, m given numbers, a sequence of integers will be obtained, each of which lies between 0 and $m-1$. An approximation to Uniform(0,1) random variables u_i can then be achieved by setting $u_i = x_i/m$.

Exponential and Weibull distribution
As the Exponential and Weibull distributions are continuous, one can make use of the *Inversion Method* to generate their distribution functions. Suppose one wishes to simulate a continuous random variable X with distribution function $F(x) = P(X \leq x)$, and suppose further that the inverse function $F^{-1}(u)$ is well-defined for $u \in [0,1]$. Then, it is well known that if U is a $(0,1)$-Uniform random variable, $X = F^{-1}(U)$ has the required distribution.

Normal distribution
For the simulation of the Normal distribution, the *Polar Marsagliar Method* is applied: If U is a Uniform(0,1) random variable, then $V = 2U - 1$ is a Uniform(-1,1) random variable. By selecting two independent Uniform(-1,1) random variables V_1 and V_2, a random point in the square $[-1,1] \times [-1,1]$ can be specified which has polar coordinates $(\tilde{R}, \Theta)$ given by $\tilde{R}^2 = V_1^2 + V_2^2$ and $tan(\Theta) = V_2/V_1$. The repeated selection of such points provides a random scatter of points inside this square, and rejection of points outside the unit-circle produces a uniform random scatter of points within this circle. For any of these points, the polar coordinates $\tilde{R}$ and Θ are independent random variables, Θ is a Uniform(0,2π) random variable and $\tilde{R}^2$ is a Uniform(0,1) random variable. One can write

$$sin(\Theta) = \frac{V_2}{\tilde{R}} = \frac{V_2}{\sqrt{V_1^2 + V_2^2}}, \quad cos(\Theta) = \frac{V_1}{\sqrt{V_1^2 + V_2^2}}.$$

Eventually, a pair of independent $N(0,1)$-variables is obtained by defining M_1 and M_2 as

$$M_1 = \sqrt{-2log(\tilde{R}^2)}\, \frac{V_2}{\sqrt{V_1^2 + V_2^2}}, \quad M_2 = \sqrt{-2log(\tilde{R}^2)}\, \frac{V_1}{\sqrt{V_1^2 + V_2^2}}.$$

3.2 Results of the simulations

Table 3.2 and 3.3 show the results of the simulation study for *model* 2 under the different scenarios given in Table 3.1 above. Each column gives the median and mean value [standard deviation] calculated using 500 replicates for the estimated model parameters.

Table 3.2: Estimated regression parameters when $\alpha = 4$, $\beta = 2$ and $\sigma^2 = 1$

	Median	**Mean [Std]**	**Median**	**Mean [Std]**	**Median**	**Mean [Std]**
	for α		**for β**		**for σ^2**	
Exponential($\frac{1}{8}$)						
n=200,p=0.3	3.801	3.799 [0.228]	2.011	2.011 [0.032]	1.193	1.192 [0.130]
n=500,p=0.3	3.827	3.823 [0.141]	2.007	2.007 [0.032]	1.199	1.201 [0.084]
n=200,p=0.7	3.971	3.972 [0.159]	1.997	1.997 [0.021]	0.994	1.100 [0.111]
n=500,p=0.7	3.974	3.977 [0.099]	1.997	1.997 [0.013]	1.005	1.010 [0.068]
Weibull($\frac{1}{6}$,$\frac{3}{2}$)						
n=200,p=0.3	4.030	4.028 [0.246]	1.973	1.972 [0.069]	1.309	1.317 [0.134]
n=500,p=0.3	4.043	4.033 [0.163]	1.971	1.973 [0.044]	1.330	1.327 [0.091]
n=200,p=0.7	3.977	3.979 [0.183]	1.999	1.999 [0.049]	0.958	0.961 [0.101]
n=500,p=0.7	3.981	3.981 [0.117]	2.001	2.000 [0.032]	0.978	0.980 [0.071]
Normal(6,4)						
n=200,p=0.3	4.219	4.215 [0.497]	1.937	1.940 [0.085]	0.950	0.945 [0.118]
n=500,p=0.3	4.213	4.223 [0.303]	1.939	1.938 [0.052]	0.948	0.952 [0.069]
n=200,p=0.7	4.055	4.033 [0.358]	1.983	1.984 [0.059]	0.930	0.933 [0.105]
n=500,p=0.7	4.059	4.058 [0.222]	1.980	1.981 [0.037]	0.931	0.938 [0.069]

Table 3.3: Estimated regression parameters when $\alpha = 4$, $\beta = 5$ and $\sigma^2 = 1$

	Median	**Mean [Std]**	**Median**	**Mean [Std]**	**Median**	**Mean [Std]**
	for α		**for β**		**for σ^2**	
Exponential($\frac{1}{8}$)						
n=200,p=0.3	3.531	3.510 [0.288]	5.062	5.064 [0.045]	1.882	1.879 [0.235]
n=500,p=0.3	3.559	3.549 [0.180]	5.056	5.058 [0.028]	1.866	1.885 [0.138]
n=200,p=0.7	3.944	3.939 [0.168]	5.004	5.003 [0.022]	1.093	1.093 [0.123]
n=500,p=0.7	3.952	3.951 [0.100]	5.003	5.002 [0.012]	1.106	1.105 [0.072]
Weibull($\frac{1}{6}$,$\frac{3}{2}$)						
n=200,p=0.3	3.817	3.817 [0.306]	5.040	5.043 [0.090]	2.233	2.253 [0.303]
n=500,p=0.3	3.836	3.833 [0.193]	5.038	5.039 [0.056]	2.267	2.258 [0.203]
n=200,p=0.7	3.970	3.974 [0.198]	5.005	5.004 [0.053]	1.042	1.042 [0.110]
n=500,p=0.7	3.972	3.974 [0.118]	5.004	5.005 [0.032]	1.078	1.076 [0.071]

Normal(6,4)						
n=200,p=0.3	4.423	4.409 [0.568]	4.921	4.920 [0.100]	1.283	1.294 [0.166]
n=500,p=0.3	4.436	4.433 [0.347]	4.918	4.917 [0.060]	1.277	1.282 [0.100]
n=200,p=0.7	4.107	4.125 [0.378]	4.971	4.969 [0.063]	0.960	0.970 [0.110]
n=500,p=0.7	4.124	4.124 [0.232]	4.970	4.970 [0.039]	0.984	0.984 [0.075]

Both tables show that the values of the median and the mean do not differ much within the simulation scenarios. For $\beta = 2$, the estimation results for the parameter α are best when the covariate distribution is Weibull. For an exponential covariate distribution, this parameter is slightly underestimated, and for a normal distribution it is slightly overestimated. It can be also noticed that the standard deviation is twofold when the covariate distribution is normal. The estimation of the parameter β is very accurate for all covariate distributions and the standard deviations are also smaller than those for the parameter α. The results for the estimation of the error variance σ^2 is most satisfying for an exponential and Weibull covariate distribution with a low level of censoring ($p = 0.7$). At a high censoring level, the value of the error variance is overestimated. The results for a normally distributed covariate are similar for both low and high censoring levels but generally underestimate the error variance.

For $\beta = 5$, the estimation results for the parameter α are most satisfying when the percentage of censored data is low, regardless of the covariate distribution. When the percentage of censoring is high, the value of α is underestimated in case of the exponential and Weibull distribution, and overestimated in case of the normal distribution. Among these three covariate distributions, the Weibull performs best. With respect to the model parameter β, the simulation results show that the estimation procedure performs well for all three covariate distributions and estimates close to the true parameter value are obtained. The error variance σ^2 is estimated most satisfactorily for a low censoring level, otherwise it is overestimated. The value of the slope β has obviously an effect in the estimation of the error variance because the overestimation was not that high for $\beta = 2$.

It can be also noticed that the number of observations affects the value of the standard deviation of the estimates in so far that it gets smaller if the number of observations gets larger.

Table 3.4 gives a summary of those simulation scenarios for which the parameter estimates perform best.

Table 3.4: Summary of the simulation results

	best performance for $\beta = 2$	**best performance for $\beta = 5$**
$\hat{\alpha}$	Weib, exp/norm and p=0.7	exp/norm/Weib and p=0.7
$\hat{\beta}$	all scenarios	all scenarios
$\hat{\sigma}^2$	exp/Weib and p=0.7	exp/norm/Weib and p=0.7

Chapter 4

Discussion of other approaches

Two other approaches for the estimation problem of *model* 2 were investigated in addition to the semiparametric approach described in Chapter 2. The first approach is an empirical one with the idea of adapting the well-known uncensored regression estimators to the context of interval censored data. The second approach imitates the least squares method of uncensored regression analysis and transfers it to the interval censored setting. The following sections summarize the problems encountered in the process of examining these approaches.

4.1 Empirical approach

Consider the linear model $Y = \alpha + \beta Z + \varepsilon$ where Y is the response variable and Z the covariate, both uncensored. It is known from regression theory that for this model the least squares estimates

$$\hat{\beta} = \frac{c\hat{o}v(Y,Z)}{v\hat{a}r(Z)} \quad \text{and} \quad \hat{\alpha} = \hat{E}(Y) - \hat{\beta}\hat{E}(Z) \qquad (*)$$

are unbiased and have minimum variance when the conditions of the Gauss-Markov theorem are met.

When Y and Z are interval censored, one could think in trying to estimate the involved covariance, variance and expected values through the common density function of Z and Y, which can be calculated with the method developed by Betensky and Finkelstein (1999) described in Chapter 1.2. From the estimated common density $\hat{h}$, say, one could then calculate the marginal densities $\hat{f}$ and $\hat{g}$, say, of Y and Z, respectively. From these three distribution functions one could finally estimate the covariance, variance and expected values from

$$\hat{E}(Z) = \int_{Z_{L_i}}^{Z_{R_i}} zg(z)dz, \qquad \hat{E}(Y) = \int_{Y_{L_i}}^{Y_{R_i}} yf(y)dz,$$

$$v\hat{a}r(Z) = \int_{Z_{L_i}}^{Z_{R_i}} (z - \hat{E}(Z))^2 g(z)dz,$$

$$c\hat{o}v(Y,Z) = \int_{Z_{L_i}}^{Z_{R_i}} \int_{Y_{L_i}}^{Y_{R_i}} (z - \hat{E}(Z))(y - \hat{E}(Y))h(y,z)dydz,$$

and calculate the estimators $\hat{\alpha}$ and $\hat{\beta}$ with the formulas given in $(*)$.

Simulations using the same simulation scenarios as in the semiparametric approach showed that the estimates for α resulting from the empirical approach are not very accurate. Table 4.1 below gives the means [mean squared errors] of $\hat{\alpha}$ and $\hat{\beta}$, calculated from 1000 replications of each setting.

Table 4.1: Simulation results for the empirical approach where $\alpha = 4$ and $\beta = 2$

distribution	**parameters**	$\hat{\alpha}$ **[MSE]**	$\hat{\beta}$ **[MSE]**
Exponential($\frac{1}{8}$)	n=100, p=0.3	4.390 [0.41]	2.086 [0.02]
	n=500,p=0.3	4.329 [0.41]	2.086 [0.02]
	n=100,p=0.7	4.547 [0.40]	2.022 [<0.01]
	n=500,p=0.7	4.478 [0.25]	2.024 [<0.01]
Weibull($\frac{1}{6}$,$\frac{3}{2}$)	n=100,p=0.3	4.611 [0.52]	1.984 [0.03]
	n=500,p=0.3	4.662 [0.47]	1.939 [<0.01]
	n=100,p=0.7	4.593 [0.42]	1.995 [<0.01]
	n=500,p=0.7	4.559 [0.33]	1.976 [<0.01]
Normal(6,4)	n=100,p=0.3	4.145 [0.98]	2.112 [0.05]
	n=500,p=0.3	4.099 [0.20]	2.103 [0.02]
	n=100,p=0.7	4.168 [0.48]	2.111 [0.03]
	n=500,p=0.7	4.091 [0.10]	2.103 [0.01]

It can be seen that the value of α is strongly overestimated when the covariate distribution is exponential or Weibull. Only in case of a normally distributed covariate, this estimate is near the true value. The mean squared error is quite high for all three covariate distributions, so it must be concluded that the values of the estimator differ considerately within the 1000 replications. With respect to the parameter β, the simulation results show that the estimates are quite accurate and the mean squared errors are small.

One could conclude from Table 4.1 that the estimation results for a normally distributed covariate are not too bad, but this conclusion is not very appropriate due to the high mean squared errors for $\hat{\alpha}$. Furthermore, the estimation results are only stable when the number of observation is very high ($n = 500$), which does indicate a poor performance on small data sets. Also, the percentage of censoring effects the value of the mean squared error, but the influence seems not to be as high as that of the number of observations, especially in the case of a normally distributed covariate.

The main disadvantage, though, of the empirical approach is that it does not provide an estimate for the model error variance σ^2. In the uncensored data setting, $\hat{\sigma}^2$ is calculated from the formula

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} z_i)^2,$$

which has no proper equivalent in the interval censored data setting. The method of replacing the unobserved values y_i and z_i by the midpoints of their observed censoring intervals is generally known to lead to considerable biases in the estimators and is also not a methodologically correct approach.

4.2 Least squares approach

The least squares method in uncensored regression analysis achieves parameter estimation by minimizing the sum of squares

$$\sum_{i=1}^{n} (y_i - \alpha - \beta z_i)^2,$$

that is, the vertical distances between the observed data points and the fitted line. One could think in applying this method to the interval censored data setting by minimizing the distances between the observed data rectangles and the fitted line. To avoid the definition of such a distance, one could directly try to minimize

$$\sum_{i=1}^{n} E\left((y_i - \alpha - \beta z_i)^2 | z_i \in [z_{L_i}, z_{R_i}], y_i \in [y_{L_i}, y_{R_i}]\right),$$

which is the expected sum of squares conditioned on the observed data rectangles $[z_{L_i}, z_{R_i}] \times [y_{L_i}, y_{R_i}]$. This would be equivalent to minimizing

$$\sum_{i=1}^{n} \int_{z_{L_i}}^{z_{R_i}} \int_{y_{L_i}}^{y_{R_i}} (y - \alpha - \beta z)^2 h_i(z, y) dy dz, \qquad (**)$$

where $h_i(z, y)$ is the joint density of Z and Y truncated into the rectangle $[z_{L_i}, z_{R_i}] \times [y_{L_i}, y_{R_i}]$.

The solution of this equations would require the calculation of the density h_i, which can be achieved with the method of Betensky and Finkelstein (1999), as well as the mathematical minimization of the given sum with respect to the parameters α and β, which could be carried out by a mathematical software like MAPLE. For the purpose of running simulations in order to assess the performance of the estimators, the problem occurrs how to connect these two steps so that they can be executed consecutively by the computer without interference from the outside. This problem could not be solved until now because of two facts: The MAPLE software is too inefficient to calculate the common density h_i, and the C language can not be used to solve minimization problems. Trying to calculate first h_i in C and then solving the minimization problem in MAPLE fails because it does not seem to exist a command that automatically starts a MAPLE program from the C interface. Theoretical calculations of the properties of the parameter estimates resulting from minimizing $(**)$ are quite complex and difficult to interpret.

Chapter 5

Outlook

For the purpose of assessing the goodness of the estimated *model* 2, a residual theory should be developed in the future to complete the proposed regression theory. It is not sufficient to consider an ad-hoc approach like Gómez et al. (2002) did, because it could be seen from the results of the simulation study in Chapter 3 that these residuals perform quite unsatisfactorily in most of the considered data situations. It is rather desirable to extend the concept of the residual theory given in Part II of this thesis to the case that the response variable is interval censored as well.

Appendix A

Derivation of the ML equations when the errors are normally distributed

Consider the likelihood function

$$L = \prod_{i=1}^{n} \sum_{j=1}^{m} \alpha_{ij} \int_{Y_{L_i}}^{Y_{R_i}} f(y|s_j, \theta) w_j dy = \prod_{i=1}^{n} \sum_{j=1}^{m} \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} f(y|s_j, \theta) dy,$$

where α_{ij} equals one if $s_j \in [z_{L_i}, z_{R_i}]$ and zero elsewhere. $\theta = (\alpha, \beta, \sigma^2)$ is the parameter vector to be estimated, and $f(y|s_j, \theta)$ is given by

$$f(y|s_j, \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{1}{2\sigma^2}(y - \alpha - \beta s_j)^2).$$

Define

$$L := \prod_{i=1}^{n} C_i(\theta),$$

where $C_i(\theta)$ is the contribution of the i-th individual to the likelihood L. Then,

$$logL = \sum_{i=1}^{n} logC_i(\theta).$$

In order to get the ML-estimators of θ, the ML equations are solved:

$$(E1) \qquad \frac{\partial logL}{\partial \alpha} = 0,$$

$$(E2) \qquad \frac{\partial logL}{\partial \beta} = 0,$$

$$(E3) \qquad \frac{\partial logL}{\partial \sigma^2} = 0.$$

Consider the quantities a_i, b_i, c_i, d_i and e_i defined as

$$a_i := E(Y|[Z_{L_i}, Z_{R_i}], [Y_{L_i}, Y_{R_i}]) = \frac{\sum_{j=1}^m \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} y f(y|s_j, \theta) dy}{C_i(\theta)},$$

$$b_i := E(Z|[Z_{L_i}, Z_{R_i}], [Y_{L_i}, Y_{R_i}]) = \frac{\sum_{j=1}^m s_j \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} f(y|s_j, \theta) dy}{C_i(\theta)},$$

$$c_i := E(Z^2|[Z_{L_i}, Z_{R_i}], [Y_{L_i}, Y_{R_i}]) = \frac{\sum_{j=1}^m s_j^2 \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} f(y|s_j, \theta) dy}{C_i(\theta)},$$

$$d_i := E(ZY|[Z_{L_i}, Z_{R_i}], [Y_{L_i}, Y_{R_i}]) = \frac{\sum_{j=1}^m s_j \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} y f(y|s_j, \theta) dy}{C_i(\theta)},$$

$$e_i := E(Y^2|[Z_{L_i}, Z_{R_i}], [Y_{L_i}, Y_{R_i}]) = \frac{\sum_{j=1}^m \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} y^2 f(y|s_j, \theta) dy}{C_i(\theta)}.$$

Then, solving equation (E1) leads to

$$(E1) \Leftrightarrow \sum_{i=1}^n \frac{1}{C_i(\theta)} \sum_{j=1}^m \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} \frac{y - \alpha - \beta s_j}{\sigma^2} f(y|s_j, \theta) dy = 0$$

$$\Leftrightarrow \sum_{i=1}^n \frac{1}{C_i(\theta)} \sum_{j=1}^m \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} y f(y|s_j, \theta) dy$$

$$= \sum_{i=1}^n \frac{1}{C_i(\theta)} \sum_{j=1}^m \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} (\alpha + \beta s_j) f(y|s_j, \theta) dy$$

$$\Leftrightarrow \sum_{i=1}^{n} a_i = \sum_{i=1}^{n} \frac{1}{C_i(\theta)} \sum_{j=1}^{m} \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} (\alpha + \beta s_j) f(y|s_j, \theta) dy$$

$$\Leftrightarrow \sum_{i=1}^{n} a_i = \alpha \sum_{i=1}^{n} \frac{\sum_{j=1}^{m} \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} f(y|s_j, \theta) dy}{C_i(\theta)}$$

$$+\beta \sum_{i=1}^{n} \frac{\sum_{j=1}^{m} s_j \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} f(y|s_j, \theta) dy}{C_i(\theta)}$$

$$\Leftrightarrow \sum_{i=1}^{n} a_i = n\alpha + \beta \sum_{i=1}^{n} b_i \Leftrightarrow n\alpha = \sum_{i=1}^{n} a_i - \beta \sum_{i=1}^{n} b_i$$

$$\Rightarrow \hat{\alpha} = \frac{1}{n} \sum_{i=1}^{n} a_i - \hat{\beta} \frac{1}{n} \sum_{i=1}^{n} b_i = \bar{a} - \hat{\beta} \bar{b}.$$

Equally, solving equation (E2) results in

$$(E2) \Leftrightarrow \sum_{i=1}^{n} \frac{1}{C_i(\theta)} \sum_{j=1}^{m} \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} \frac{y - \alpha - \beta s_j}{\sigma^2} s_j f(y|s_j, \theta) dy = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} \frac{1}{C_i(\theta)} \sum_{j=1}^{m} \alpha_{ij} w_j s_j \int_{Y_{L_i}}^{Y_{R_i}} (y - \alpha) f(y|s_j, \theta) dy$$

$$= \sum_{i=1}^{n} \frac{1}{C_i(\theta)} \sum_{j=1}^{m} \alpha_{ij} w_j \beta s_j^2 \int_{Y_{L_i}}^{Y_{R_i}} f(y|s_j, \theta) dy$$

$$\Leftrightarrow \sum_{i=1}^{n} \frac{1}{C_i(\theta)} \sum_{j=1}^{m} \alpha_{ij} w_j s_j \int_{Y_{L_i}}^{Y_{R_i}} (y - \alpha) f(y|s_j, \theta) dy = \beta \sum_{i=1}^{n} c_i$$

$$\Leftrightarrow \sum_{i=1}^{n} \frac{1}{C_i(\theta)} \left(\sum_{j=1}^{m} \alpha_{ij} w_j s_j \int_{Y_{L_i}}^{Y_{R_i}} y f(y|s_j, \theta) dy - \alpha \sum_{j=1}^{m} \alpha_{ij} w_j s_j \int_{Y_{L_i}}^{Y_{R_i}} f(y|s_j, \theta) dy \right)$$

$$= \beta \sum_{i=1}^{n} c_i$$

$$\Leftrightarrow \sum_{i=1}^{n} \frac{1}{C_i(\theta)} \sum_{j=1}^{m} \alpha_{ij} w_j s_j \int_{Y_{L_i}}^{Y_{R_i}} y f(y|s_j, \theta) dy - \alpha \sum_{i=1}^{n} b_i = \beta \sum_{i=1}^{n} c_i$$

$$\Leftrightarrow \sum_{i=1}^{n} d_i - \alpha \sum_{i=1}^{n} b_i = \beta \sum_{i=1}^{n} c_i$$

$$\Leftrightarrow \bar{d} - \alpha \bar{b} = \beta \bar{c} \Leftrightarrow \beta = \frac{\bar{d} - \alpha \bar{b}}{\bar{c}},$$

and replacing α by its estimate $\hat{\alpha}$ from (E1) results that

$$\hat{\beta} = \frac{\bar{d} - \bar{a}\bar{b}}{\bar{c} - \bar{b}^2}.$$

Finally, from equation (E3) one obtains

$$(E3) \Leftrightarrow \sum_{i=1}^{n} \frac{1}{C_i(\theta)} \sum_{j=1}^{m} \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} f(y|s_j, \theta) \left(\frac{-1}{2\sigma^2} + \frac{(y - \alpha - \beta s_j)^2}{2\sigma^4} \right) dy = 0$$

$$\Leftrightarrow \frac{1}{\sigma^4} \sum_{i=1}^{n} \frac{1}{C_i(\theta)} \sum_{j=1}^{m} \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{Ri}} f(y|s_j, \theta)(y - \alpha - \beta s_j)^2 dy$$

$$= \frac{1}{\sigma^2} \sum_{i=1}^{n} \frac{1}{C_i(\theta)} \sum_{j=1}^{m} \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} f(y|s_j, \theta) dy$$

$$\Leftrightarrow \sum_{i=1}^{n} \frac{1}{C_i(\theta)} \sum_{j=1}^{m} \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} f(y|s_j, \theta)(y - \alpha - \beta s_j)^2 dy = n\sigma^2.$$

Noting that $(y - \alpha - \beta s_j)^2 = (y - \alpha)^2 + \beta^2 s_j^2 - 2\beta s_j (y - \alpha)$, this is equal to

$$\sum_{i=1}^{n} \frac{1}{C_i(\theta)} \sum_{j=1}^{m} \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} (y - \alpha)^2 f(y|s_j, \theta) dy = n\sigma^2 - \beta^2 \sum_{i=1}^{n} c_i + 2\beta^2 \sum_{i=1}^{n} c_i$$

$$\Leftrightarrow \sum_{i=1}^{n} \frac{1}{C_i(\theta)} \sum_{j=1}^{m} \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} (y - \alpha)^2 f(y|s_j, \theta) dy = n\sigma^2 + \beta^2 \sum_{i=1}^{n} c_i$$

$$\Leftrightarrow \sum_{i=1}^{n} \frac{1}{C_i(\theta)} \sum_{j=1}^{m} \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} (y^2 - 2\alpha y + \alpha^2) f(y|s_j, \theta) dy = n\sigma^2 + \beta^2 \sum_{i=1}^{n} c_i$$

$$\Leftrightarrow \sum_{i=1}^{n} \frac{1}{C_i(\theta)} (\sum_{j=1}^{m} \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} y^2 f(y|s_j, \theta) dy - 2\alpha \sum_{j=1}^{m} \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} y f(y|s_j, \theta) dy$$

$$
\begin{aligned}
&+\alpha^2 \sum_{j=1}^{m} \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} f(y|s_j,\theta) dy = n\sigma^2 + \beta^2 \sum_{i=1}^{n} c_i \\
\Leftrightarrow & \sum_{i=1}^{n} e_i - 2\alpha \sum_{i=1}^{n} a_i + n\alpha^2 = n\sigma^2 + \beta^2 \sum_{i=1}^{n} c_i \\
\Leftrightarrow & \sum_{i=1}^{n} e_i - 2\alpha \sum_{i=1}^{n} a_i + n\alpha^2 - \beta^2 \sum_{i=1}^{n} c_i = n\sigma^2 \\
\Rightarrow & \hat{\sigma^2} = \bar{e} - 2\hat{\alpha}\bar{a} + \hat{\alpha}^2 - \hat{\beta}^2 \bar{c}.
\end{aligned}
$$

Appendix B

Maple program for the calculation of approximate confidence intervals

```
> with(LinearAlgebra):
```

Specifying the number of observations n and the number of examinations m

```
> n:=2; m:=6;
```

Reading the data

```
> data:=matrix(9,6,readdata(‘A:\\data.txt‘,9));
```

Assigning the variables needed in the loglikelihood

```
> ID:=matrix(n,m);
> for i from 1 to n do for j from 1 to m do
  ID[i,j]:=data[i,j] end do end do;
> for j from 1 to m do w[j]:=data[3,j] end do;
> for j from 1 to m do s[j]:=data[4,j] end do;
> for i from 1 to n do yl[i]:=data[5,i] end do;
> for i from 1 to n do yr[i]:=data[6,i] end do;
> alphahat:=data[7,1];
```

```
> betahat:=data[8,1];
> sigma2hat:=data[9,1];
```

Definition of the log-likelihood

```
> i:='i'; j:='j';
> for j from 1 to m do
  f[j]:=(1/(sqrt(2*Pi*sigma^2)))*exp(-((y-alpha-beta*s[j])^2)/
           (2*sigma^2))  end do;
> loglike:=
  sum('log(sum('ID[i,j]*w[j]*int(f[j],y=yl[i]..yr[i])',
               'j'=1..6))','i'=1..n);
```

Calculation of the score function of loglike

```
> i:='i'; j:='j';
> der11:=diff(loglike,alpha);
> der12:=diff(loglike,beta);
> der13a:=algsubs(sigma^2=V,loglike);
> der13b:=subs(sigma=sqrt(V),der13a);
> der13c:=diff(der13b,V);
> der13:=subs(V=sigma^2,der13c);
```

Calculation of the second derivatives of loglike

```
> der111:=diff(der11,alpha);
> der112:=diff(der11,beta);
> der113a:=algsubs(sigma^2=V,der11);
> der113b:=subs(sigma=sqrt(V),der113a);
> der113c:=diff(der113b,V);
> der113:=subs(V=sigma^2,der113c);
>
> der122:=diff(der12,beta);
> der123a:=algsubs(sigma^2=V,der12);
> der123b:=subs(sigma=sqrt(V),der123a);
> der123c:=diff(der123b,V);
> der123:=subs(V=sigma^2,der123c);
>
> der133a:=algsubs(sigma^2=V,der13);
```

```
> der133b:=subs(sigma=sqrt(V),der133a);
> der133c:=diff(der133b,V);
> der133:=subs(V=sigma^2,der133c);
```

Construction of the Hessian matrix

```
> matt:=Matrix(1..3,1..3,[[der111,der112,der113],
        [der112,der122,der123],[der113,der123,der133]]);
>
```

Calculating the observed information matrix

```
> alpha:=alphahat;beta:=betahat;sigma:=sqrt(sigma2hat);
> evalf(matt);
> fish:=evalf(-1*matt);
```

Inverting the observed information matrix which is an estimate for the variance of \hat{alpha}, \hat{beta} and \hat{sigma}^2

```
> variance:=MatrixInverse(fish);
```

Constructing the confidence intervals for the regression parameters

```
> alpha:='alpha';beta:='beta';sigma:='sigma';
> CI(alpha):=[alphahat-1.96*sqrt(variance[1,1])/sqrt(n),
              alphahat+1.96*sqrt(variance[1,1])/sqrt(n)];
> CI(beta):=[betahat-1.96*sqrt(variance[2,2])/sqrt(n),
             betahat+1.96*sqrt(variance[2,2])/sqrt(n)];
> CI(sigma):=[sigma2hat-1.96*sqrt(variance[3,3])/sqrt(n),
              sigma2hat+1.96*sqrt(variance[3,3])/sqrt(n)];
```

Appendix C

Derivation of the MLE for the multiple regression setting

With the notations given in Appendix A, setting the partial derivations of the likelihood to zero and solving for the parameters, one yields the following solutions:

For the parameter α it holds that

$$\frac{\partial logL^*}{\partial \alpha} = \sum_{i=1}^{n} \frac{1}{C_i(\theta)} \sum_{j=1}^{m} \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} \frac{y - \alpha - \vec{\beta}_1' \vec{x}_i - \beta_2 s_j}{\sigma^2} f(y|(\vec{x}_i, s_j); \theta) dy \overset{!}{=} 0$$

$$\Leftrightarrow \sum_{i=1}^{n} \frac{1}{C_i(\theta)} \sum_{j=1}^{m} \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} y f(y|(\vec{x}_i, s_j); \theta) dy$$

$$= \sum_{i=1}^{n} \frac{1}{C_i(\theta)} \sum_{j=1}^{m} \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} (\alpha + \vec{\beta}_1' \vec{x}_i + \beta_2 s_j) f(y|(\vec{x}_i, s_j); \theta) dy$$

$$\Leftrightarrow \sum_{i=1}^{n} a_i = n\alpha + \sum_{i=1}^{n} \frac{1}{C_i(\theta)} \sum_{j=1}^{m} \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} \vec{\beta}_1' \vec{x}_i f(y|(\vec{x}_i, s_j); \theta) dy + \beta_2 \sum_{i=1}^{n} b_i$$

$$\Leftrightarrow \sum_{i=1}^{n} a_i - \sum_{l=1}^{p} \beta_{1l} \sum_{i=1}^{n} x_{li} - \beta_2 \sum_{i=1}^{n} b_i = n\alpha.$$

For the parameter $\vec{\beta}_1$ it holds that

$$\frac{\partial logL^*}{\partial \vec{\beta}_1'} = \sum_{i=1}^{n} \frac{1}{C_i(\theta)} \sum_{j=1}^{m} \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} \frac{\vec{x}_i(y - \alpha - \vec{\beta}_1'\vec{x}_i - \beta_2 s_j)}{\sigma^2} f(y|(\vec{x}_i, s_j); \theta)dy$$

$$\stackrel{!}{=} 0$$

$$\Leftrightarrow \sum_{i=1}^{n} x_{ki} a_i = \alpha \sum_{i=1}^{n} x_{ki} + \sum_{l=1}^{p} \beta_{1l} \sum_{i=1}^{n} x_{li} x_{ki} + \beta_2 \sum_{i=1}^{n} x_{ki} b_i,$$

$$\Leftrightarrow \sum_{i=1}^{n} x_{ki} a_i - \alpha \sum_{i=1}^{n} x_{ki} - \beta_2 \sum_{i=1}^{n} x_{ki} b_i = \sum_{l=1}^{p} \beta_{1l} \sum_{i=1}^{n} x_{li} x_{ki},$$

$$\text{for} \quad k = 1, \ldots, p.$$

For the parameter β_2 it holds that

$$\frac{\partial logL^*}{\partial \beta_2} = \sum_{i=1}^{n} \frac{1}{C_i(\theta)} \sum_{j=1}^{m} \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} \frac{s_j(y - \alpha - \vec{\beta}_1'\vec{x}_i - \beta_2 s_j)}{\sigma^2} f(y|(\vec{x}_i, s_j); \theta)dy$$

$$\stackrel{!}{=} 0$$

$$\Leftrightarrow \sum_{i=1}^{n} \frac{1}{C_i(\theta)} \sum_{j=1}^{m} \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} \beta_2 s_j^2 f(y|(\vec{x}_i, s_j); \theta)dy$$

$$= \sum_{i=1}^{n} \frac{1}{C_i(\theta)} \sum_{j=1}^{m} \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} s_j(y - \alpha - \vec{\beta}_1'\vec{x}_i) f(y|(\vec{x}_i, s_j); \theta)dy$$

$$\Leftrightarrow \beta_2 \sum_{i=1}^{n} c_i = \sum_{i=1}^{n} d_i - \alpha \sum_{i=1}^{n} b_i - \sum_{l=1}^{p} \beta_{1l} \sum_{i=1}^{n} x_{li} b_i.$$

For the parameter σ^2 it holds that

$$\frac{\partial logL^*}{\partial \sigma^2} = \sum_{i=1}^{n} \frac{1}{C_i(\theta)} \sum_{j=1}^{m} \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} \left(-\frac{1}{2\sigma^2} + \frac{2(y - \alpha - \vec{\beta}_1'\vec{x}_i - \beta_2 s_j)^2}{4\sigma^4} \right)$$

$$f(y|(\vec{x}_i, s_j); \theta)dy \stackrel{!}{=} 0$$

$$\Leftrightarrow \frac{1}{2\sigma^2} \sum_{i=1}^{n} \frac{1}{C_i(\theta)} \sum_{j=1}^{m} \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} f(y|(\vec{x}_i, s_j); \theta)dy$$

$$= \frac{1}{2\sigma^4} \sum_{i=1}^{n} \frac{1}{C_i(\theta)} \sum_{j=1}^{m} \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} (y - \alpha - \vec{\beta}_1' \vec{x}_i - \beta_2 s_j)^2 f(y|(\vec{x}_i, s_j); \theta) dy. \qquad (*)$$

Noting that $(y - \alpha - \vec{\beta}_1' \vec{x}_i - \beta_2 s_j)^2$ is equivalent to $(y - \alpha - \vec{\beta}_1' \vec{x}_i)^2 + \beta_2^2 s_j^2 - 2\beta_2 s_j (y - \alpha - \vec{\beta}_1' \vec{x}_i)$, it holds that

$$(*) \Leftrightarrow n\sigma^2 - \beta^2 \sum_{i=1}^{n} c_i + 2\beta^2 \sum_{i=1}^{n} c_i$$

$$= \sum_{i=1}^{n} \frac{1}{C_i(\theta)} \sum_{j=1}^{m} \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} (y - \alpha - \vec{\beta}_1' \vec{x}_i)^2 f(y|(\vec{x}_i, s_j); \theta) dy$$

$$\Leftrightarrow n\sigma^2 + \beta^2 \sum_{i=1}^{n} c_i = \sum_{i=1}^{n} e_i - 2\alpha \sum_{i=1}^{n} a_i - 2 \sum_{l=1}^{p} \beta_{1l} \sum_{i=1}^{n} x_{li} a_i$$

$$-2\alpha \sum_{l=1}^{p} \beta_{1l} \sum_{i=1}^{n} x_{li} + n\alpha^2 + \sum_{l=1}^{p} \beta_{1l}^2 \sum_{i=1}^{n} x_{li}^2.$$

Appendix D

Derivation of the MLE when the errors come from the exponential family

Setting $logL^{**}$ to zero and solving for the parameters yields the following maximum likelihood estimates:

With respect to α one gets

$$\frac{\partial logL^{**}}{\partial\alpha} = \sum_{i=1}^{n}\left(\frac{\sum_{j=1}^{m}\alpha_{ij}w_j\int_{Y_{L_i}}^{Y_{R_i}}\frac{\partial h(y_i-\alpha-\beta s_j)}{\partial\alpha}c(\eta)exp[Q(\eta)t(y_i-\alpha-\beta s_j)]}{C_i(\theta)}\right.$$

$$\left.+\frac{\sum_{j=1}^{m}\alpha_{ij}w_j\int_{Y_{L_i}}^{Y_{R_i}}f(y_i|s_j;\theta)\left[Q(\eta)\frac{\partial t(y_i-\alpha-\beta s_j)}{\partial\alpha}\right]dy}{C_i(\theta)}\right)$$

$$=\sum_{i=1}^{n}\left(\frac{\sum_{j=1}^{m}\alpha_{ij}w_j\int_{Y_{L_i}}^{Y_{R_i}}-\frac{h'(\varepsilon_i)}{h(\varepsilon_i)}f(y_i|s_j;\theta)dy}{C_i(\theta)}\right.$$

$$\left.-\frac{\sum_{j=1}^{m}\alpha_{ij}w_j\int_{Y_{L_i}}^{Y_{R_i}}f(y_i|s_j;\theta)[Q(\eta)t'(\varepsilon_i)]dy}{C_i(\theta)}\right)=(F1).$$

With respect to β one gets

$$\frac{\partial logL^{**}}{\partial\beta}=\sum_{i=1}^{n}\left(\frac{\sum_{j=1}^{m}\alpha_{ij}w_j\int_{Y_{L_i}}^{Y_{R_i}}\frac{\partial h(y_i-\alpha-\beta s_j)}{\partial\beta}c(\eta)exp[Q(\eta)t(y_i-\alpha-\beta s_j)]}{C_i(\theta)}\right.$$

$$\left.+\frac{\sum_{j=1}^{m}\alpha_{ij}w_j\int_{Y_{L_i}}^{Y_{R_i}}f(y_i|s_j;\theta)\left[Q(\eta)\frac{\partial t(y_i-\alpha-\beta s_j)}{\partial\beta}\right]dy}{C_i(\theta)}\right)$$

$$=\sum_{i=1}^{n}\left(\frac{\sum_{j=1}^{m}\alpha_{ij}w_j\int_{Y_{L_i}}^{Y_{R_i}}-\frac{s_jh'(\varepsilon_i)}{h(\varepsilon_i)}f(y_i|s_j;\theta)dy}{C_i(\theta)}\right.$$

$$\left.-\frac{\sum_{j=1}^{m}\alpha_{ij}w_j\int_{Y_{L_i}}^{Y_{R_i}}f(y_i|s_j;\theta)[Q(\eta)s_jt'(\varepsilon_i)]dy}{C_i(\theta)}\right)=(F2).$$

With respect to η one gets

$$\frac{\partial logL^{**}}{\partial\eta}=\sum_{i=1}^{n}\left(\frac{\sum_{j=1}^{m}\alpha_{ij}w_j\int_{Y_{L_i}}^{Y_{R_i}}h(y_i-\alpha-\beta s_j)c'(\eta)exp[Q(\eta)t(y_i-\alpha-\beta s_j)]}{C_i(\theta)}\right.$$

$$\left.+\frac{\sum_{j=1}^{m}\alpha_{ij}w_j\int_{Y_{L_i}}^{Y_{R_i}}f(y_i|s_j;\theta)\left[Q'(\eta)t(y_i-\alpha-\beta s_j)\right]dy}{C_i(\theta)}\right)$$

$$=\sum_{i=1}^{n}\left(\frac{\sum_{j=1}^{m}\alpha_{ij}w_j\int_{Y_{L_i}}^{Y_{R_i}}-\frac{c'(\eta)}{c(\eta)}f(y_i|s_j;\theta)dy}{C_i(\theta)}\right.$$

$$\left.+\frac{\sum_{j=1}^{m}\alpha_{ij}w_j\int_{Y_{L_i}}^{Y_{R_i}}f(y_i|s_j;\theta)[Q'(\eta)t(\varepsilon_i)]dy}{C_i(\theta)}\right)=(F3).$$

Setting these equations to zero and solving for the parameters one obtains the maximum likelihood equations

$$\sum_{i=1}^{n} h_i = -Q(\eta) \sum_{i=1}^{n} t'_i,$$

$$\sum_{i=1}^{n} zh_i = -Q(\eta) \sum_{i=1}^{n} zt'_i,$$

$$\sum_{i=1}^{n} n\frac{c'(\eta)}{c(\eta)} = -Q'(\eta) \sum_{i=1}^{n} t_i,$$

where

$$t_i = E\left(t(\varepsilon)|[y_{L_i}, y_{R_i}]; [x_{L_i}, x_{R_i}]\right),$$

$$t'_i = E\left(t'(\varepsilon)|[y_{L_i}, y_{R_i}]; [x_{L_i}, x_{R_i}]\right),$$

$$zt'_i = E\left(Zt'(\varepsilon)|[y_{L_i}, y_{R_i}]; [x_{L_i}, x_{R_i}]\right),$$

$$h_i = E\left(\frac{h'(\varepsilon)}{h(\varepsilon)}||[y_{L_i}, y_{R_i}]; [x_{L_i}, x_{R_i}]\right),$$

$$zh_i = E\left(Z\frac{h'(\varepsilon)}{h(\varepsilon)}|[y_{L_i}, y_{R_i}]; [x_{L_i}, x_{R_i}]\right).$$

Appendix E

Derivation of the MLE when the errors come from the Weibull distribution

Setting l^{***} to zero and solving for the parameters yields the following solutions:
For the parameters $\hat{\alpha}$ it holds that

$$\frac{\partial l^{***}}{\partial \alpha} = \sum_{i=1}^{n} \frac{1}{C_i(\theta)} \sum_{j=1}^{m} \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} \beta s_j^{\beta-1} \exp\left(-\alpha s_j^{\beta}\right) - \alpha\beta s_j^{\beta-1} \exp\left(-\alpha s_j^{\beta}\right) s_j^{\beta} dy$$

$$= \sum_{i=1}^{n} \frac{1}{C_i(\theta)} \sum_{j=1}^{m} \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} f(y_i|s_j;\theta) \left(\frac{1}{\alpha} - s_j^{\beta}\right) dy \stackrel{!}{=} 0$$

$$\Leftrightarrow \frac{n}{\alpha} = \sum_{i=1}^{n} f_i \qquad \Rightarrow \qquad \hat{\alpha} = n / \sum_{i=1}^{n} f_i.$$

For the parameters $\hat{\beta}$ it holds that

$$\frac{\partial l^{***}}{\partial \beta} = \sum_{i=1}^{n} \frac{1}{C_i(\theta)} \sum_{j=1}^{m} \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} \left(\alpha s_j^{\beta-1} + \alpha\beta(\beta-1) s_j^{\beta-2}\right) \exp\left(-\alpha s_j^{\beta}\right)$$

$$+ \alpha\beta s_j^{\beta-1} \exp\left(-\alpha s_j^{\beta}\right) \left(-\alpha\beta s_j^{\beta-1}\right) dy$$

$$= \sum_{i=1}^{n} \frac{1}{C_i(\theta)} \sum_{j=1}^{m} \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} f(y_i|s_j;\theta) \left(\frac{1}{\beta} + (\beta-1) s_j - \alpha\beta s_j^{\beta-1}\right) dy \stackrel{!}{=} 0$$

$$\Leftrightarrow \frac{n}{\beta} + (\beta - 1)\sum_{i=1}^{n} b_i = \alpha\beta \sum_{i=1}^{n} g_i$$

$$\Leftrightarrow \frac{n}{\beta} + \beta \sum_{i=1}^{n} b_i - \alpha\beta \sum_{i=1}^{n} g_i = \sum_{i=1}^{n} b_i.$$

In these expressions, the following conditional expected values are used:

$$\frac{1}{C_i(\theta)} \sum_{j=1}^{m} \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} s_j^{\beta} f(y_i|s_j;\theta)dy = E(Z^{\beta}|[y_{L_i}, y_{R_i}], [z_{L_i}, z_{R_i}]) = f_i$$

and

$$\frac{1}{C_i(\theta)} \sum_{j=1}^{m} \alpha_{ij} w_j \int_{Y_{L_i}}^{Y_{R_i}} s_j^{\beta-1} f(y_i|s_j;\theta)dy = E(Z^{\beta-1}|[y_{L_i}, y_{R_i}], [z_{L_i}, z_{R_i}]) = g_i.$$

Introduction

The second part of this thesis deals with residual analysis in the context of linear regression models with interval censored data. Residual analysis is the general class of techniques for detecting problems in regression models, based on the fact that residuals carry important information concerning the appropriateness of the assumptions made in linear regression analysis.

Many of today's common methods of residual analysis were developed in the early 1960s in works by F. Anscombe, J.W. Tukey, G.E.P. Box and D.R Cox. During the late 1970s interest in residual analysis was renewed by the development of methods for assessing the influence of individual observations in model estimation. Residual based methods for detecting model deficiencies or influential observations include informal graphics to display general features of the residuals as well as formal tests to detect specific departures from underlying model assumptions.

In uncensored regression analysis, the residuals are defined as the difference between the observed and the fitted response, and a plot of these quantities against the covariate or the fitted response values is a standard tool for the evaluation of the fitted model. An overview of the properties of uncensored residuals and the different ways of using them for model evaluation is given in Chapter 1.1.

When a linear model incorporates censored data, the difficulty of defining appropriate residuals occurs. Since the realizations of the censored variables are not directly observable, one can not calculate the difference between the observed and the fitted response. One approach to solve this problem is given in Hillis (1995), who developed a residual theory for linear models with right censored data, which is presented in Chapter 1.2. A definition of residuals in the context of regression analysis with interval censored covariates was proposed by Gómez et al. (2002) and is summarized in Chapter 1.3.

The method of Gómez et al. (2002) is the only one existing for linear models with interval censored data, and their performance has not been investigated yet. For this reason, Chapter 2 of this thesis develops a new residual theory for regression models incorporating interval censored covariates. It is shown that the residuals resulting from this context are interval censored as well, which leads to the proposal of determining these unobserved residuals through their distribution function inside the respective censoring intervals. The performance of the so-defined residuals is investigated in Chapter 3 by means of a simulation study. The study includes the residuals defined by Gómez et al. (2002) as well as residuals resulting from taking the midpoints of the covariate intervals as the observed covariate values. The results show that the residuals proposed in this thesis are superior to the other two. In Chapter 4, an application of the new method to a real data set is given.

Contents

Chapter 1

Residual analysis in regression models

This chapter gives an overview of some of the existing theories for residual analysis in linear regression models. The first section introduces residual analysis for uncensored data: important properties of uncensored residuals are presented along with the most common devices for using them in model evaluation. The second section presents right censored residuals as introduced by Hillis (1995). This concept will be the basis for the construction of new residuals in context with interval censored data as proposed in Chapter 2. Finally, a residual theory for linear models with interval censored covariates as proposed by Gómez et al. (2002) is presented.

1.1 Uncensored residuals

In the uncensored data situation, one considers the linear model

$$y_i = \alpha + \beta z_i + \varepsilon_i, \quad i = 1, \ldots, n, \qquad \textit{model } 1$$

where the pair (y_i, z_i) for the response variable and the covariate is observed directly. The model errors ε_i are usually assumed to be independent and identically distributed and to have a normal distribution with mean zero and constant variance. In this context, the so called least squares estimates $\hat{\alpha}$

and $\hat{\beta}$ for the unknown regression parameters α and β are defined as

$$\hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{\beta}\frac{1}{n}\sum_{i=1}^{n} z_i \quad \text{and} \quad \hat{\beta} = \frac{\sum_{i=1}^{n} y_i x_i - \frac{1}{n}\sum_{i=1}^{n} y_i \sum_{i=1}^{n} z_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} z_i\right)^2}.$$

The fitted values $\hat{y}_i$ are $\hat{y}_i = \hat{\alpha} + \hat{\beta} z_i$, and the least squares residuals $\hat{e}_i$ result from the difference between the i-th observed response value and the corresponding fitted value: $\hat{e}_i = y_i - \hat{y}_i$. That is, one can think of the residuals as the observed errors if the model is correct, or the quantity the regression equation has not been able to explain.

The least squares residuals have some important properties which are summarized in the following proposition (Montgomery and Peck, 1982, p.16f):

Proposition 1
Under the assumptions of the *model* 1, the least squares residuals have the following properties:

- They have zero mean.
- Their approximate average variance is $\frac{1}{n-2}\sum_{i=1}^{n}(\hat{e}_i - \bar{\hat{e}})^2$. This quantity is also known as the mean squared error (MSE).
- The residuals are not independent and the correlation between $\hat{e}_i$ and $\hat{e}_j$ is $\rho_{ij} = \frac{cov(\hat{e}_i, \hat{e}_j)}{\sqrt{var(\hat{e}_i) var(\hat{e}_j)}}$, $i, j = 1, \ldots, n$.
- The sum of the residuals in any regression model that contains an intercept is zero: $\sum_{i=1}^{n} \hat{e}_i = 0$.
- The sum of the residuals weighted by the corresponding value of the regressor variable always equals zero: $\sum_{i=1}^{n} z_i \hat{e}_i = 0$.
- The sum of the residuals weighted by the corresponding fitted value always equals zero: $\sum_{i=1}^{n} \hat{y}_i \hat{e}_i = 0$.

Sometimes it is useful to work with the standardized residuals $\hat{d}_i = \frac{\hat{e}_i}{\sqrt{MSE}}$ which have zero mean and approximately unit variance.

Another type of residuals is the so called studentized residual. They result from standardizing each residual with an estimate $s_{(i)}$ of the residuals' standard deviation independent of that residual: $s_{(i)}$ is calculated by applying the formula of the MSE but leaving out the i-th observation. The studentized residuals are then defined as $\hat{r}_i = \frac{\hat{e}_i}{s_{(i)}}$. In a regression model where p parameters are estimated, each studentized residual is distributed as Student's t with $n-p-1$ degrees of freedom when normality of the error term ε holds. Like the standardized and ordinary residuals, the studentized residuals are not independent of each other (Rawlings, 1988, p.249f).

1.2 Regression diagnostics with uncensored residuals

Residuals can generally be used to assess both the validity of the data and how well the assumptions of the model are satisfied. The main focus here will be on the latter issue, an extensive review of methods for the detection of influential observations is given in Cook and Weisberg (1982).

Usually, the following assumptions are to be checked after the model is fitted to the data:

Distribution
Most analytical methods for fitting regression models assume some parametric distribution of the dependent variable, in most cases the normal distribution. This distribution is usually determined via the model errors, and the residuals are considered to be able to reflect it. Since it is known that an incorrect specification of the error distribution leads to not efficient parameter estimates and invalid inferential statements, it is important to check the assumed distribution.

Fit of the relationship
Residuals can also be used to assess whether the assumed relationship between the dependent and the independent variable adequately fits the data. For example, one may check whether or not the mean of the dependent variable is a linear function of a given independent variable.

Error variance
Furthermore, residual diagnostics can be used to assess whether or not the variance of the model errors is constant (homoscedastic).

Available methods for studying these assumptions via residuals include both graphical and nongraphical procedures. The most common ways for examining residuals for the validation of the estimated model is using graphical devices. The principal ways of plotting residuals are overall plots, plots against the fitted values and plots against the covariate:

The overall plot

If the n residuals are plotted overall and the fitted model is correct, then one should obtain n observations from a normal distribution with mean zero. To prove if the residuals contradict this idea one can construct a normal plot where the observations should fall approximately on a straight line. When the number of residuals is very large, a histogram can be used which should then have the form of a Gauss curve with mean zero. (Draper and Smith 1981, p.142f).

Plot of the residuals versus the fitted values

A plot of the residuals $\hat{e}_i$ versus the corresponding fitted values $\hat{y}_i$ is useful for detecting several common types of model inadequacies. It is important, though, not to use the observed values y_i in the plot, because the $\hat{e}_i$ and the y_i are usually correlated while the $\hat{e}_i$ and the $\hat{y}_i$ are uncorrelated (for a proof see Draper and Smith, 1981, p.147f).

If the residual plot resembles data points which are distributed in the same way above and below the zero-axis, that is, one can include the points in a horizontal band, then there are no obvious model defects.

In contrast, if the plotted positive residuals get larger as the $\hat{y}_i$ get larger, and the plotted negative residuals get more negative as the $\hat{y}_i$ get larger (that is, it appears an outward opening funnel pattern), then this indicates that the variance of the errors is not constant but an increasing function of Y. An inward-opening funnel would mean that the variance increases as Y decreases. The usual approach to deal with inequality of the variance is to apply a suitable transformation to either the regressor or the response variable.

On the other hand, a curved pattern of the plotted residuals would indicate nonlinearity. This means that other regressor variables, for example a squared term, are needed to be included in the model.

A plot of the residuals versus the fitted values $\hat{y}_i$ may also reveal one or more unusually large residuals. These points are potential outliers. But large residuals occurring at the extreme $\hat{y}_i$-values can also indicate that either the variance is not constant or the true relationship between y and x is not linear. These possibilities should be investigated before the points are considered as outliers.

Plot of the residuals versus an independent variable

Alternatively, one can construct a plot of the residuals against the corresponding values of a regressor variable. This will reveal wrong model specifications in the same way as described for the plot of the residuals against the fitted values.

Formal tests based on residuals

Formal test procedures for regression diagnostics are also available. Tests for normality based on uncensored residuals usually make use of the score test or the Lagrange multiplier test. See for example Jarque and Bera (1987) who also provide a comparison study between various of these tests. A Monte Carlo study comparing the performance of procedures like the Kolmogorov-Smirnov or Chi-square test when applied to regression residuals is given in Huang and Bolch (1974). Diagnostic tests of the distributional shape other than the normal are given for example in Spiegelhalter (1983). Though, a common feature of all the existing tests is that they require independent observations to be tested on. Therefore, one has to assure that the residuals one wishes to apply to these tests are independent.

Diagnostic tests for homoscedasticity in uncensored regression analysis are manifold. Goldfeld and Quandt (1965) distinguished between constructive and nonconstructive tests: constructive testing procedures are designed to test for and at the same time estimate the specific form of heteroscedasticity. This means that in case of the rejection of homoscedasticity, an estimate of the covariance matrix is directly available. See for example Rutemiller and Bowers (1968), Glejser (1969) or White (1980). Nonconstructive procedures as those of Goldfeld and Quandt (1965), Theil (1971) and Harrison and McCabe (1979) are designed to establish the absence or presence of heteroscedasticity without regard to subsequent estimation. Also, different types of heteroscedasticity can be specified, for example that the error variance is a function of the independent variable or that it depends on the values of the dependent variable. Most of the test procedures for heteroscedasticity are parametric and assume a normal distribution of the residuals. Those not assuming an underlying parametric distribution are rather complicated to compute or rely on weight functions and other parameters that have to be specified according to somewhat difficult patterns.

Test diagnostics for the linear relationship between uncensored variables in regression models do not exist in the current literature.

1.3 Residuals for right censored data

Hillis (1995) proposed residuals for linear models when the response variable is not exactly observed but censored to the right. He considers the model

$$t_i = \beta z_i + \varepsilon_i, \quad i = 1, \ldots, n \qquad \textit{model } 2$$

where the ε_i are independently and identically distributed with distribution function F, t_i is the survival time of the i-th individual, and z_i is the value of the corresponding covariate. The censoring time for t_i is denoted as c_i with the assumption that the distribution of the ε_i does not depend on the value of z_i or c_i. The observed data for *model* 2 is the triple (y_i, z_i, δ_i), where $y_i = \min(t_i, c_i)$ and $\delta_i = I(t_i \leq c_i)$ with I the indicator function.

For the development of the residuals in this context, the author defines a sequence of random variables

$$\varepsilon_i^* = \delta_i \varepsilon_i + (1 - \delta_i) U_i,$$

where each U_i comes from the distribution function F_i defined as

$$F_i(x) = P(\varepsilon_i \leq x | \varepsilon_i > c_i - z_i\beta) = \begin{cases} 0 & : \quad x \leq c_i - z_i\beta \\ \frac{F(x) - F(c_i - z_i\beta)}{1 - F(c_i - z_i\beta)} & : \quad x > c_i - z_i\beta \end{cases}.$$

This means that ε_i^* equals ε_i for an uncensored observation, and for a censored observation ε_i^* is equal to a randomly generated observation from the conditional distribution of ε_i given that $\varepsilon_i > c_i - z_i\beta$.

The author shows that the ε_i^* have the same joint distribution as the ε_i and suggests to replace β and F with their estimates for defining the residuals of *model* 2:

$$\hat{e}_i^* = \delta_i \hat{e}_i + (1 - \delta_i)\hat{u}_i,$$

where the $\hat{u}_i$ are randomly generated observations from the distribution $\hat{F}_i$ given by

$$\hat{F}_i(x) = \begin{cases} 0 & : \quad x \leq c_i - z_i\hat{\beta} \\ \frac{\hat{F}(k) - \hat{F}(c_i - z_i\hat{\beta})}{1 - \hat{F}(c_i - z_i\hat{\beta})} & : \quad x > c_i - z_i\hat{\beta} \end{cases}.$$

In this expression $\hat{F}$ is the product-limit estimate based on the censored and uncensored residuals $\hat{e}_i = y_i - z_i\hat{\beta}$, and $\hat{\beta}$ is the Buckley-James estimate for the parameter β (see Buckley and James, 1979).

If the model assumptions are correct, plots of the $\hat{e}_i^*$ versus the independent variable or the fitted values exhibit a random scatter.

1.4 Residuals in models with an interval censored covariate

For models where the response variable Y is continuous and exactly observed and the covariate Z is discrete and interval censored, Gómez et al. (2002) proposed residuals to graphically assess the fit of the model. The authors consider the model

$$y_i = \alpha + \beta z_i + \varepsilon_i, \qquad i = 1, \ldots, n$$

where the y_i are the exactly observed response values and z_i is the realization of the interval censored covariate Z_i for which only the corresponding censoring intervals $[z_{L_i}, z_{R_i}]$ can be observed. The model errors ε_i are assumed to be normally distributed and independent of Z_i.

For the estimation of the regression parameters, the authors propose an algorithm that simultaneously maximizes the data likelihood and estimates the distribution function of the covariate. For details see Chapter 1.3 of the first part of this thesis.

The authors define the residuals of this context to be $r_i = y_i - \hat{\alpha} - \hat{\beta} z_i$. Because of the fact that the value of z_i is not directly observed but only the corresponding censoring interval $[z_{L_i}, z_{R_i}]$, they propose to replace it by the conditional expected value $\hat{z}_i = E_{\hat{W}_T}(Z|z_{L_i}, z_{R_i})$. Here, $\hat{W}_T$ is the estimated distribution function of the covariate that results from applying Turnbull's (1976) method on the observed covariate intervals (for details on the method of Turnbull see Chapter 1.2 of the first part of this thesis). So, the model residuals proposed by the authors are $\hat{r}_i = y_i - \hat{\alpha} - \hat{\beta} \hat{z}_i$.

The authors show that $E(\hat{r}_i) = E(r_i) = 0$, so that a plot of $\hat{r}_i$ versus $\hat{z}_i$ should show a random scatter around zero if the regression model is correctly specified.

Chapter 2

New residuals for models with interval censored covariates

This chapter presents a new methodology for residual analysis in linear models that incorporate an interval censored covariate. It is based on the assumption of normality for the model errors and the fact that they can not be observed directly but only their respective censoring intervals, as explained in the following.

The linear regression model considered here is given by

$$Y_i = \alpha + \beta Z_i + \varepsilon_i, \quad i = 1, \ldots, n \qquad \textit{model } 3$$

where Y_i is the continuous response variable with realizations y_i, Z_i is the discrete, interval censored covariate with realizations $[z_{L_i}, z_{R_i}]$, and the model errors ε_i have distribution function $F = N(0, \sigma^2)$ and are independent of the Z_i. An extension of this situation to a more general setting is when allowing both interval censored and exactly observed data for the covariate. This case will be considered here, and the observed data then consists of the triple $(y_i, [z_{L_i}, z_{R_i}], \delta_i)$, where δ_i equals zero if the covariate for the i-th individual is interval censored, and δ_i equals one if the covariate is exactly observed. In the latter case the interval $[z_{L_i}, z_{R_i}]$ becomes the point $\{z_i\}$.

The aim is to assess the goodness of the fitted *model* 3 using residuals. The regression parameters α, β and σ^2 will be estimated by applying the method of Gómez et al. (2002) described in Chapter 1.3 of the first part of this thesis.

Consider *model* 3 for an individual i with exactly observed covariate value z_i. This case resembles the situation in the simple linear model with uncensored data where the respective model errors ε_i are given by

$$\varepsilon_i = y_i - \alpha - \beta z_i. \tag{2.1}$$

The situation changes when the covariate for individual i is interval censored, that is, only the interval $[z_{L_i}, z_{R_i}]$ is observed. Then, it follows from *model* 3 that the resulting model errors are interval censored as well, and are included by the error intervals

$$[y_i - \alpha - \beta z_{L_i}, y_i - \alpha - \beta z_{R_i}] \quad \text{if} \quad \beta < 0 \quad \text{and} \tag{2.2}$$

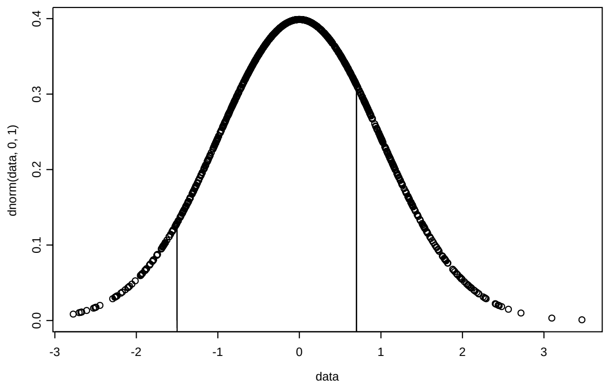
$$[y_i - \alpha - \beta z_{R_i}, y_i - \alpha - \beta z_{L_i}] = [A_i, B_i] \quad \text{if} \quad \beta > 0. \tag{2.3}$$

In the following only the case $\beta > 0$ will be considered.

For illustrative purposes, and in order to be able to distinguish clearly between those residuals resulting from exact and those coming from interval censored covariate observations, we introduce the following notation: Those model errors coming from exactly observed data as given in equation (2.1) will be denoted as ε_i. The model errors coming from interval censored data will be called η_i.

When we deal with an interval censored Z_i, the resulting model error η_i is not known directly but we observe only the corresponding error intervals in equations (2.2) and (2.3) which are known to contain η_i with probability one. In order to obtain some more information about where η_i may be located inside this error interval, one can look at its distribution function. The distribution of the η_i is determined by the assumption that the model errors have a $N(0, \sigma^2)$-distribution and the above stated fact that they are interval censored. This leads to the conclusion that the η_i have a $N(0, \sigma^2)$-distribution truncated in the error interval limits A_i and B_i as illustrated in Figure 2.1.

Figure 2.1: distribution of the interval censored model errors



Its formula is given by

$$G_i(x) = P(\eta_i \leq x | \eta_i \in [A_i, B_i]) = \begin{cases} 0 & : \quad x < A_i \\ \frac{\Phi(x/\sigma) - \Phi(A_i/\sigma)}{\Phi(B_i/\sigma) - \Phi(A_i/\sigma)} & : \quad x \in [A_i, B_i] \\ 1 & : \quad x > B_i \end{cases}, \quad (2.4)$$

where Φ is the distribution function of the standard normal distribution.

From this follows that the model errors accommodating simultaneously for exact and interval censored covariate observations in *model* 3 are given by:

$$\varepsilon_i^* = \delta_i \varepsilon_i + (1 - \delta_i) \eta_i, \quad (2.5)$$

with ε_i^* equal to $\varepsilon_i = y_i - \alpha - \beta z_i$ if the i-th covariate is not censored, and ε_i^* equal to η_i coming from the conditional distribution G_i defined above when the covariate is interval censored.

Then, the residuals corresponding to the model errors defined in (2.5) result to

$$\hat{e}_i^* = \delta_i \hat{e}_i + (1 - \delta_i) \hat{\eta}_i, \quad (2.6)$$

where $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\sigma}^2$ are the estimates for the model parameters resulting from the procedure of Gómez et al. (2002). This means that $\hat{e}_i^*$ equals $\hat{e}_i = y_i - \hat{\alpha} - \hat{\beta} z_i$ if the i-th covariate is not censored, and for an interval censored covariate, $\hat{e}_i^*$ equals $\hat{\eta}_i$ defined as the expected value of the distribution $\hat{G}_i$ given by

$$\hat{G}_i(x) = \begin{cases} 0 & : \quad x < \hat{A}_i \\ \frac{\Phi(x/\hat{\sigma}) - \Phi(\hat{A}_i/\hat{\sigma})}{\Phi(\hat{B}_i/\hat{\sigma}) - \Phi(\hat{A}_i/\hat{\sigma})} & : \quad x \in [\hat{A}_i, \hat{B}_i] \\ 1 & : \quad x > \hat{B}_i \end{cases} , \tag{2.7}$$

where $[\hat{A}_i, \hat{B}_i] = [y_i - \hat{\alpha} - \hat{\beta} z_{R_i}, y_i - \hat{\alpha} - \hat{\beta} z_{L_i}]$ are the estimated residual intervals.

The value of the residual $\hat{\eta}_i$, where $\hat{\eta}_i$ is the mean of a $N(0, \hat{\sigma}^2)$-distribution truncated in $\hat{A}_i$ and $\hat{B}_i$, can be calculated by using standard results of probability theory (see for example Hartung et al., 1993):

$$\hat{\eta}_i = \frac{\varphi(\hat{A}_i/\hat{\sigma}) - \varphi(\hat{B}_i/\hat{\sigma})}{\Phi(\hat{B}_i/\hat{\sigma}) - \Phi(\hat{A}_i/\hat{\sigma})} \, \hat{\sigma}, \tag{2.8}$$

where φ is the density function of the standard normal distribution.

2.1 Theoretical properties of the residuals

It is usually of interest to calculate the expected value and the variance of the proposed estimates. As can be seen in the formulas of the previous section, the proposed estimate for the residuals of *model* 3 result quite complicate and straightforward computations of the mean and the variance of these estimates are not possible. Nevertheless, some approximate results will be given below.

Consider the estimated residual vector $\hat{\mathbf{e}}^* = (\hat{e}_1^*, \ldots, \hat{e}_n^*)$ where $\hat{e}_i^*$ equals $\hat{\eta}_i$ when individual i has an interval censored covariate, and $\hat{e}_i^*$ equals $\hat{e}_i$ when the covariate of individual i is exactly observed. The expected value of the residual vector $\hat{\mathbf{e}}^*$ is therefore composed of the expected value of the vector $\hat{\mathbf{e}}$ of all uncensored residuals and the expected value of the vector $\hat{\eta}$ of all residuals coming from an interval censored covariate.

Proposition 1
The expected value of the estimated residual vector $\hat{\mathbf{e}}$, *which has entry* $\hat{e}_i$ *at position* i *for all individuals* i *with exactly observed covariate, and zero otherwise, is given by*

$$E(\hat{\mathbf{e}}) = \alpha - E(\hat{\alpha}) + \beta\mathbf{z} - E(\hat{\beta})\mathbf{z},$$

where $\mathbf{z}$ *is the vector of all exactly observed covariate values.*

Proof

$$E(\hat{\mathbf{e}}) = E(\mathbf{y} - \hat{\alpha} - \hat{\beta}\mathbf{z}) = E(\mathbf{y}) - E(\hat{\alpha}) - E(\hat{\beta})\mathbf{z} =$$

$$= E(\alpha + \beta\mathbf{z} + \varepsilon) - E(\hat{\alpha}) - E(\hat{\beta})\mathbf{z} = \alpha + \beta\mathbf{z} - E(\hat{\alpha}) - E(\hat{\beta})\mathbf{z},$$

where $\mathbf{y}$ and ε is the vector of the response values and model errors, respectively, of those individuals who have an exactly observed covariate. □

Proposition 2
The expected value of the estimated residual vector $\hat{\eta}$, *which has entry* $\hat{\eta}_i$ *at position* i *for all individuals* i *with an interval censored covariate, and zero otherwise, can be approximated by*

$$E(\hat{\eta}) \approx \frac{\varphi(\mathbf{A}/\sigma) - \varphi(\mathbf{B}/\sigma)}{\Phi(\mathbf{B}/\sigma) - \Phi(\mathbf{A}/\sigma)}\,\sigma,$$

under the assumption that the parameter estimates for α, β *and* σ^2 *are unbiased.*

Proof

$$E(\hat{\eta}) = E\left(\frac{\varphi(\hat{\mathbf{A}}/\hat{\sigma}) - \varphi(\hat{\mathbf{B}}/\hat{\sigma})}{\Phi(\hat{\mathbf{B}}/\hat{\sigma}) - \Phi(\hat{\mathbf{A}}/\hat{\sigma})}\,\hat{\sigma}\right) \approx \frac{E(\varphi(\hat{\mathbf{A}}/\hat{\sigma})) - E(\varphi(\hat{\mathbf{B}}/\hat{\sigma}))}{E(\Phi(\hat{\mathbf{B}}/\hat{\sigma})) - E(\Phi(\hat{\mathbf{A}}/\hat{\sigma}))}\,E(\hat{\sigma}).$$

If $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\sigma}^2$ are unbiased estimates, this expression is equivalent to

$$\frac{\varphi(\mathbf{A}/\sigma) - \varphi(\mathbf{B}/\sigma)}{\Phi(\mathbf{B}/\sigma) - \Phi(\mathbf{A}/\sigma)}\,\sigma = \eta.$$ □

Summarizing above two terms, the approximated expected value of the residual vector $\hat{\mathbf{e}}^*$ is given by

$$E(\hat{\mathbf{e}}^*) = \delta \left(\alpha - E(\hat{\alpha}) + \mathbf{z}(\beta - E(\hat{\beta}))\right) + (1-\delta)\frac{\varphi(\mathbf{A}/\sigma) - \varphi(\mathbf{B}/\sigma)}{\Phi(\mathbf{B}/\sigma) - \Phi(\mathbf{A}/\sigma)}\ \sigma,$$

where the components δ_i of δ equal one when the covariate is exactly observed, and zero when it is interval censored.

In the uncensored case, the residuals are known for the property of having mean zero. The expression above for the approximated expected value for the interval censored residuals $\hat{e}^*$ is zero only if the model parameters $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\sigma}^2$ are unbiased and the A_i and B_i are symmetric around zero.

The computations of the variance of $\hat{\mathbf{e}}^*$ is also divided into one part regarding the residual vector $\hat{\mathbf{e}}$ and another part concerned with the residual vector $\hat{\eta}$.

Proposition 3
The variance of the estimated residual vector $\hat{\mathbf{e}}$, which has entry $\hat{e}_i$ at position i for all individuals i with exactly observed covariate, and zero otherwise, is given by

$$Var(\mathbf{e}) = Var(\varepsilon) + Var(\hat{\alpha} - \hat{\beta}\mathbf{z}).$$

Proof

$$Var(\mathbf{e}) = Var(\mathbf{y} - \hat{\alpha} - \hat{\beta}\mathbf{z}) = Var(\alpha + \beta\mathbf{z} + \varepsilon - \hat{\alpha} - \hat{\beta}\mathbf{z})$$

$$= Var(\varepsilon) + Var(\hat{\alpha} - \hat{\beta}\mathbf{z}). \qquad \Box$$

Proposition 4
The variance of the estimated residual vector $\hat{\eta}$, which has entry $\hat{\eta}_i$ at position i for all individuals i with an interval censored covariate, and zero otherwise, can be approximated by

$$Var(\hat{\eta}) \approx \hat{\sigma}^2\ \frac{\varphi(Var(\hat{\mathbf{A}})/\hat{\sigma}^2) - \varphi(Var(\hat{\mathbf{B}})/\hat{\sigma}^2)}{\Phi(Var(\hat{\mathbf{B}})/\hat{\sigma}^2) - \Phi(Var(\hat{\mathbf{A}})/\hat{\sigma}^2)}.$$

Here, $\hat{\mathbf{A}}$ *and* $\hat{\mathbf{B}}$ *is the vector with entry* $\hat{A}_i$ *and* $\hat{B}_i$*, respectively, at position* i *for individuals* i *with an interval censored covariate, and zero otherwise.*

Proof

$$Var(\hat{\eta}) = Var\left(\frac{\varphi(\hat{\mathbf{A}}/\hat{\sigma}) - \varphi(\hat{\mathbf{B}}/\hat{\sigma})}{\Phi(\hat{\mathbf{B}}/\hat{\sigma}) - \Phi(\hat{\mathbf{A}}/\hat{\sigma})}\;\hat{\sigma}\right) = \hat{\sigma}^2 Var\left(\frac{\varphi(\hat{\mathbf{A}}/\hat{\sigma}) - \varphi(\hat{\mathbf{B}}/\hat{\sigma})}{\Phi(\hat{\mathbf{B}}/\hat{\sigma}) - \Phi(\hat{\mathbf{A}}/\hat{\sigma})}\right)$$

$$\approx \hat{\sigma}^2\;\frac{\varphi(Var(\hat{\mathbf{A}}/\hat{\sigma})) - \varphi(Var(\hat{\mathbf{B}}/\hat{\sigma}))}{\Phi(Var(\hat{\mathbf{B}}/\hat{\sigma})) - \Phi(Var(\hat{\mathbf{A}}/\hat{\sigma}))}$$

$$= \hat{\sigma}^2\;\frac{\varphi(Var(\hat{\mathbf{A}})/\hat{\sigma}^2) - \varphi(Var(\hat{\mathbf{B}})/\hat{\sigma}^2)}{\Phi(Var(\hat{\mathbf{B}})/\hat{\sigma}^2) - \Phi(Var(\hat{\mathbf{A}})/\hat{\sigma}^2)}. \qquad \square$$

The value of this expression depends on the variance of the interval censored covariate and the estimated model parameters as well as on the value of the observed covariate interval as can be seen in the following two formulas:

$$Var(\hat{\mathbf{A}}) = Var(\mathbf{y} - \hat{\alpha} - \hat{\beta}\mathbf{z_R}) = Var(\alpha + \beta\mathbf{Z} + \varepsilon - \hat{\alpha} - \hat{\beta}\mathbf{z_R}) =$$

$$= \beta^2 Var(\mathbf{Z}) + \sigma^2 + Var(\hat{\alpha} - \hat{\beta}\mathbf{z_R^2})$$

and

$$Var(\hat{\mathbf{B}}) = \beta^2 Var(\mathbf{Z}) + \sigma^2 + Var(\hat{\alpha} - \hat{\beta}\mathbf{z_L^2}).$$

Consistency of the residual distribution function

The residuals $\hat{\eta}$ were defined as the mean of the truncated residual distribution function $\hat{G}$. In the following, it will be shown that $\hat{G}$ is a consistent estimate of the truncated error distribution G.

Proposition 5
It holds that $\hat{\sigma}^2$ *resulting from the estimation procedure of Gómez et al. (2002) is an consistent estimate of the true error variance* σ^2*.*

Proof
As explained in Chapter 1.3 of the first part of this thesis, the estimation procedure for σ^2 consists of two steps: the estimation of the unknown covariate distribution via a self-consistent algorithm and the maximization of the resulting likelihood function. Yu et al. (1989) proved the strong consistency of the generalized maximum likelihood estimate resulting from a self-consistent procedure. Thus, the estimated covariate distribution resulting from the first step is a consistent estimate for the true covariate distribution.

In the second step, this consistent estimate is used when deriving the formulas for the maximum likelihood estimate of σ^2. So, it can be said that, for n large, this estimate is equivalent to the 'true' maximum likelihood estimate which would result from using the true covariate distribution in the likelihood instead of the estimated one. And as commonly known, the maximum likelihood estimator is a consistent estimate for the true parameter under consideration (a proof is given for example in Wald, 1949).

Proposition 6
The estimated error distribution function $\hat{F} = N(0, \hat{\sigma}^2)$ *is a consistent estimate of the true error distribution function* $F = N(0, \sigma^2)$.

Proof
$\hat{F}$ is a simple plug-in estimate obtained by replacing the unknown variance σ^2 of F by an estimate. As shown in Proposition 5, this estimate is consistent. Bickel and Fan (1996) showed that in density estimation, plug-in estimates are consistent when the estimator used to substitute the unknown parameter is consistent itself.

With this result and the fact that $\hat{G}$ is a continuous function of $\hat{F}$, the consistency of $\hat{G}$ follows straightforwardly.

Chapter 3

Simulations

In order to find out whether the newly proposed residuals can be used to check the underlying assumptions of the model, it will be examined if they reflect the normal distribution of the model errors and if they are sensitive to deviations from the model assumptions. For this purpose, several simulation studies are conducted which include the newly proposed residuals $\hat{e}^*$ as well as three other types of residuals: The ordinary least squares (OLS) residuals $\hat{e}$, the residuals $\hat{e}_{mid}$ resulting from taking the midpoints of the intervals $[z_L, z_R]$ as the observed values for the covariate, and the residuals $\hat{e}_{lup}$ proposed by Gómez et al. (2002) defined in Chapter 1.3:

The four types of residuals involved in the simulation study are

- the least squares residuals: $\hat{e}_i = y_i - \alpha_{ls} - \beta_{ls} z_i$,

- the midpoint residuals: $\hat{e}_{mid_i} = y_i - \alpha_{ls} - \beta_{ls} z_{mid_i}$, where $z_{mid_i} = \frac{z_{L_i} + z_{R_i}}{2}$,

- the residuals following Gómez et al. (2002): $\hat{e}_{lup_i} = y_i - \hat{\alpha} - \hat{\beta} z_{lup_i}$, where $z_{lup_i} = E_{\hat{W}_T}(Z | z_{L_i}, z_{R_i})$,

- the newly proposed residuals: $\hat{e}^*_i = \delta_i \hat{e}_i + (1 - \delta_i)\hat{\eta}_i$ with $\hat{e}_i$ as defined above and $\hat{\eta}_i = \frac{\varphi(\hat{A}_i/\hat{\sigma}) - \varphi(\hat{B}_i/\hat{\sigma})}{\Phi(\hat{B}_i/\hat{\sigma}) - \Phi(\hat{A}_i/\hat{\sigma})}\ \hat{\sigma}$.

The behavior of these residuals is studied under different data scenarios including various covariate distributions, high and low percentages of censoring in the data, and different number of observations. A summary is given in Table 3.1.

Table 3.1: Scenarios for the simulation study

number of observations	200 and 500
covariate distributions	$Exp(\frac{1}{8})$, $Weib(\frac{1}{6},\frac{3}{2})$, $N(6,4)$
percentage of censoring	0.3 and 0.7
value for α	4
values for β	2 and 5
value for σ^2	1

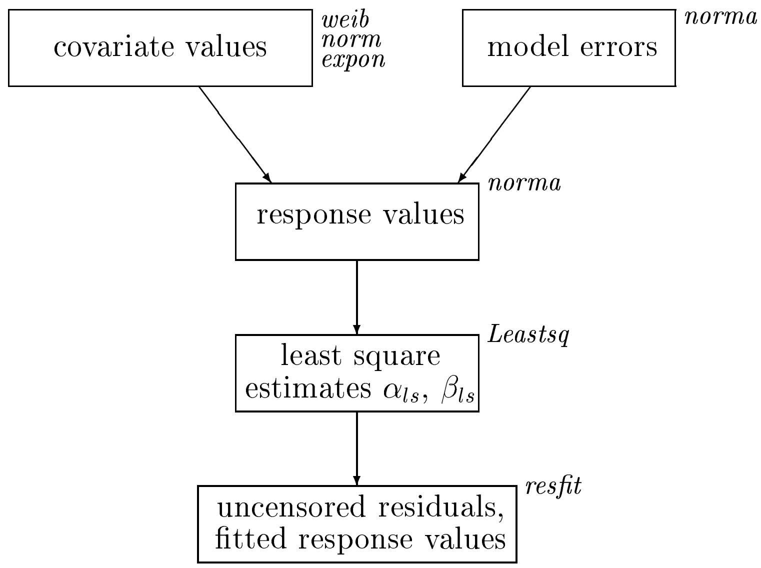
The application of test procedures for normality mentioned in Chapter 1 requires independent observations. But neither the OLS residuals nor the three other residuals to be examined are independent. Thus, for checking the normality of the residuals, the measures skewness and kurtosis are used. It is known that the value for the skewness is zero for symmetric data distributions. The more negative (positive) this value is, the more skewed to the left (right) is the data distribution. The kurtosis is a measure for unimodal distributions and compares the data distribution's absolute maximum with that of the density of a normal distribution. A value bigger (smaller) than zero indicates that the data's absolute maximum is bigger (smaller) than that of the normal distribution. This means that the theoretic distribution of the underlying population is not normal if the values for the skewness and the kurtosis differ substantially from zero. The formulas for the calculation of the skewness S and the kurtosis K of n observations x_i, $i = 1, \dots, n$, are (Hartung et al., 1993, p.48f):

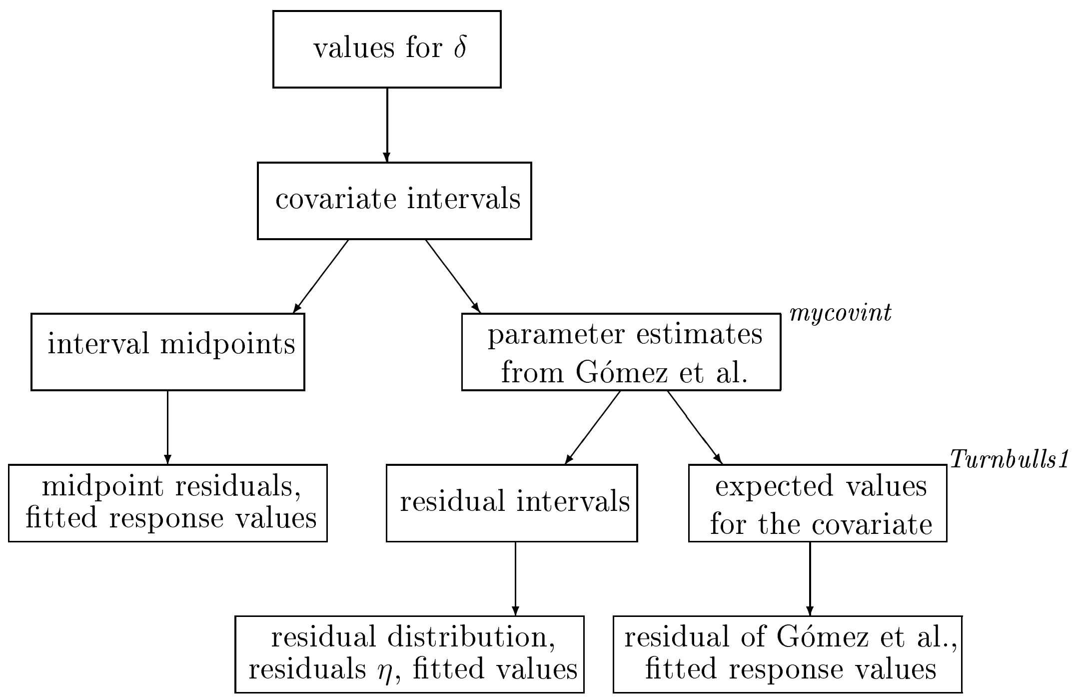
$$S = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\sqrt{(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2)^3}}, \qquad K = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2)^2} - 3.$$

When checking the model assumptions of linearity and constant error variance, residual plots will be used instead of formal test procedures for the same reason as mentioned in the context with normality.

The simulations are carried out by the computer program *residuals.cpp* which can be found on the floppy disc. It includes the following steps: First, the values z_i for the covariate are generated from an exponential, Weibull or normal distribution. The model errors ε_i are simulated from a $N(0, \sigma^2)$ distribution. The values for the true model parameters α, β and σ^2 are fixed, and from these and the two previously generated variables one can calculate the values y_i of the response variable via. The estimators $\hat{\alpha}_{ls}$ and $\hat{\beta}_{ls}$ for the model parameters of the uncensored data are determined via the least squares method, and the resulting uncensored residuals $\hat{e}_i$ can then be calculated straightforwardly. The covariate intervals $[z_L, z_R]$ are generated using the following scheme: Depending on the covariate distribution, there is a certain number of values j, $j = 1, \ldots, k$, which can be assigned to the covariate. An indicator variable δ_{ij} determines with a given probability p if the covariate for individual i is observed at value j or not. Then, one looks at each value z_i and goes back to the nearest observed value j and takes it as the value for z_{L_i}. Similarly, z_{R_i} is that observed value j which is the first after z_i. For the midpoint residuals $\hat{e}_{mid_i}$, one takes the center z_{mid_i} of each covariate interval $[z_{L_i}, z_{R_i}]$ and calculates $\hat{e}_{mid_i} = y_i - \hat{\alpha}_{ls} - \hat{\beta}_{ls} z_{mid_i}$. For the residuals $\hat{e}_{lup}$ defined by Gómez et al. (2002), one needs to apply their algorithm in order to estimate the parameters $\hat{\alpha}$ and $\hat{\beta}$. The $\hat{e}_{lup}$ are then calculated using the formula $\hat{e}_{lup_i} = y_i - \hat{\alpha} - \hat{\beta} z_{lup_i}$, where z_{lup} is the expected value of Z under the covariate distribution resulting from applying Turnbull's algorithm to the intervals $[z_L, z_R]$. Finally, our newly proposed residuals $\hat{e}_i^*$ are calculated via formula 2.6 in Chapter 2.

The following flow-charts illustrate the simulation process of the program *residuals.cpp*. The steps of the program are written inside the boxes and the arrows indicate which step enters in the calculation of another step. The first flow-chart represents that part of the program where the uncensored residuals are generated, and the second flow-chart describes the simulation process of the three types of residuals resulting from the data of the interval censored covariate. As most calculations are executed by procedures within the program, their names are written outside the corresponding box, which will make it easier to find one's way when looking at the program code.


covariate values
weib
norm
expon
model errors
norma
response values
norma
least square
estimates α_{ls}, β_{ls}
Leastsq
uncensored residuals,
fitted response values
resfit


values for δ
covariate intervals
interval midpoints
parameter estimates
from Gómez et al.
mycovint
midpoint residuals,
fitted response values
residual intervals
expected values
for the covariate
Turnbulls1
residual distribution,
residuals η, fitted values
residual of Gómez et al.,
fitted response values

Other procedures used in this program are listet below, together with a short description of their usage:

FileOpen: opens all files needed for reading and writing.
Spalloc: allocates memory for the vectors and matrices.
ran2: generates random uniform variates.
probnormal: calculates for any given value the value of the $N(0, \sigma^2)$ distribution function.
schiefe: calculates the skewness of a given data set.
kurto: calculates the kurtosis of a given data set.
sign: determines the sign of a given expression.
Turnbulls1: calculates the distribution function of an interval censored variable using the method proposed by Turnbull (1976).

The performance of the program *residuals.cpp* with respect to speed and convergence is highly satisfying. Running it on a 400 megahertz Pentium II processor with 128 MB RAM main memory using the SUSE LINUX 7.1 operating system yielded the values of the four types of residuals within seconds regardless of the number of observations and percentage of censoring.

Simulation theory

The simulation study involves the generation of data coming from different statistical distributions. The theory applied for the generation of the used distributions is given in the following (for references see Box and Müller, 1958, or Morgan, 1984).

1. *Uniform distribution*
   For the generation of a Uniform(0,1) random variable, a *Congruential Pseudo-Random Number Generator* is used. By applying the recursion formula $x_{n-1} = ax_n + b \; mod \; m$ with seed x_0 and a, b, m given numbers, a sequence of integers will be obtained, each of which lies between 0 and $m-1$. An approximation to Uniform(0,1) random variables u_i can then be achieved by setting $u_i = x_i/m$.

2. *Exponential and Weibull distribution*
   As the Exponential and Weibull distributions are continuous, one can make use of the *Inversion Method* to generate their distribution functions. Suppose one wishes to simulate a continuous random variable X

with distribution function $F(x) = P(X \leq x)$, and suppose further that the inverse function $F^{-1}(u)$ is well-defined for $u \in [0,1]$. Then, it is well known that if U is a $(0,1)$-Uniform random variable, $X = F^{-1}(U)$ has the required distribution.

3. *Normal distribution*
   For the simulation of the Normal distribution, the *Polar Marsagliar Method* is applied: If U is a Uniform(0,1) random variable, then $V = 2U - 1$ is a Uniform(-1,1) random variable. By selecting two independent Uniform(-1,1) random variables V_1 and V_2, a random point in the square $[-1,1] \times [-1,1]$ can be specified which has polar coordinates $(\tilde{R}, \Theta)$ given by $\tilde{R}^2 = V_1^2 + V_2^2$ and $tan(\Theta) = V_2/V_1$. The repeated selection of such points provides a random scatter of points inside this square, and rejection of points outside the unit-circle produces a uniform random scatter of points within this circle. For any of these points, the polar coordinates $\tilde{R}$ and Θ are independent random variables, Θ is a Uniform(0,2π) random variable and $\tilde{R}^2$ is a Uniform(0,1) random variable. One can write

$$sin(\Theta) = \frac{V_2}{\tilde{R}} = \frac{V_2}{\sqrt{V_1^2 + V_2^2}}, \quad cos(\Theta) = \frac{V_1}{\sqrt{V_1^2 + V_2^2}}.$$

   Eventually, a pair of independent $N(0,1)$-variables is obtained by defining M_1 and M_2 as

$$M_1 = \sqrt{-2log(\tilde{R}^2)}\,\frac{V_2}{\sqrt{V_1^2 + V_2^2}}, \quad M_2 = \sqrt{-2log(\tilde{R}^2)}\,\frac{V_1}{\sqrt{V_1^2 + V_2^2}}.$$

3.1 Checking for normality

Tables 3.2 to 3.5 show the simulation results for the skewness and kurtosis of each of the four types of residuals. For each scenario, median and mean values [standard deviation] are calculated using 1000 replicates.

The uncensored residuals $\hat{e}$ (Table 3.2) resemble the normal distribution of the model errors satisfactorily in each of the studied scenarios. Their median and mean values are about the same and always around zero. It can be noticed that the standard deviation of the skewness and kurtosis for $n = 100$

is two fold that for $n = 500$. This means that the residuals fit better to the normal distribution for large n than for small n. This phenomenon occurs with all four types of residuals.

The newly proposed residuals $\hat{e}^*$ (Table 3.3) have a symmetric distribution for those scenarios involving a low percentage of censoring ($p = 0.7$). Otherwise their distribution seems to be skewed to the right. The values for the kurtosis are quite large at a high percentage of censoring, but also those for a low censoring level are too big for possibly coming from a normal distribution.

The distribution of the residuals $\hat{e}_{mid}$ coming from the covariate midpoints (Table 3.4) seems to be symmetric only in the case of a Weibull-distributed covariate, but then the kurtosis is substantially above zero and therefore not similar to that of the corresponding normal distribution. In all other scenarios the values for the skewness and kurtosis differ substantially from zero.

The distribution of the residuals $\hat{e}_{lup}$ (Table 3.5) is even less normal than the one of the $\hat{e}_{mid}$. They perform best within some scenarios for the Weibull distribution but as in the case of the $\hat{e}_{mid}$ the kurtosis differs substantially from zero.

In summary, the results of the simulation study show that of the three types of residuals coming from interval censored data, the newly proposed residuals $\hat{e}^*$ perform best.

3.2 The residuals when the model is correctly specified

Residual plots will be used to examine whether the four types of residuals can be applied to validate the assumption of linearity. For that purpose it must be investigated first how the residuals behave when there are no model misspecifications. Therefore, residual plots are simulated for each residual type under the assumptions of *model* 3 in Chapter 2. For each of the 24 data scenarios, the simulated residuals are plotted versus the corresponding fitted values as shown in Appendix A. The first plot is always the one coming ht

Table 3.2: Skewness and kurtosis for the least squares residuals $\hat{e}$

	Skewness		**Kurtosis**	
	Median	**Mean [Std]**	**Median**	**Mean [Std]**
Exponential($\frac{1}{8}$)				
n=100,p=0.3,$\beta=2$	-0.044	-0.042 [0.23]	-0.009	0.055 [0.45]
n=500,p=0.3,$\beta=2$	0.004	0.004 [0.11]	-0.010	0.013 [0.22]
n=100,p=0.7,$\beta=2$	-0.031	-0.029 [0.23]	0.006	0.078 [0.46]
n=500,p=0.7,$\beta=2$	0.003	0.008 [0.11]	-0.001	0.010 [0.21]
n=100,p=0.3,$\beta=5$	-0.035	-0.037 [0.24]	-0.031	0.052 [0.46]
n=500,p=0.3,$\beta=5$	0.006	0.007 [0.11]	-0.009	0.003 [0.22]
n=100,p=0.7,$\beta=5$	-0.043	-0.037 [0.23]	0.008	0.061 [0.44]
n=500,p=0.7,$\beta=5$	0.008	0.007 [0.11]	-0.012	0.012 [0.22]
Weibull($\frac{1}{6}$,$\frac{3}{2}$)				
n=100,p=0.3,$\beta=2$	-0.038	-0.034 [0.22]	-0.017	0.057 [0.44]
n=500,p=0.3,$\beta=2$	0.008	0.006 [0.11]	-0.015	0.005 [0.21]
n=100,p=0.7,$\beta=2$	-0.021	-0.027 [0.23]	-0.032	0.056 [0.45]
n=500,p=0.7,$\beta=2$	0.004	0.006 [0.11]	0.012	0.021 [0.21]
n=100,p=0.3,$\beta=5$	-0.041	-0.042 [0.23]	-0.008	0.072 [0.45]
n=500,p=0.3,$\beta=5$	0.005	0.006 [0.11]	-0.002	0.022 [0.22]
n=100,p=0.7,$\beta=5$	-0.016	-0.024 [0.22]	-0.003	0.065 [0.43]
n=500,p=0.7,$\beta=5$	0.009	0.010 [0.11]	-0.007	0.016 [0.23]
Normal(6,4)				
n=100,p=0.3,$\beta=2$	0.009	0.012 [0.24]	-0.039	0.038 [0.46]
n=500,p=0.3,$\beta=2$	0.007	0.011 [0.11]	-0.006	0.010 [0.22]
n=100,p=0.7,$\beta=2$	-0.001	-0.004 [0.23]	-0.033	0.039 [0.45]
n=500,p=0.7,$\beta=2$	0.006	0.005 [0.11]	-0.014	0.009 [0.21]
n=100,p=0.3,$\beta=5$	0.018	0.013 [0.24]	-0.044	0.034 [0.48]
n=500,p=0.3,$\beta=5$	0.001	-0.0003 [0.11]	-0.009	0.011 [0.21]
n=100,p=0.7,$\beta=5$	0.016	0.021 [0.24]	-0.035	0.037 [0.47]
n=500,p=0.7,$\beta=5$	0.010	0.008 [0.11]	-0.014	0.008 [0.22]

Table 3.3: Skewness and kurtosis for the residuals $\hat{\eta}$

	Skewness		**Kurtosis**	
	Median	**Mean [Std]**	**Median**	**Mean [Std]**
Exponential($\frac{1}{8}$)				
n=100,p=0.3,$\beta = 2$	0.215	0.204 [0.50]	2.539	2.793 [1.39]
n=500,p=0.3,$\beta = 2$	0.275	0.274 [0.21]	2.503	2.586 [0.65]
n=100,p=0.7,$\beta = 2$	-0.030	-0.030 [0.29]	0.664	0.775 [0.66]
n=500,p=0.7,$\beta = 2$	-0.002	0.002 [0.13]	0.650	0.677 [0.29]
n=100,p=0.3,$\beta = 5$	0.270	0.295 [0.59]	2.992	3.231 [1.54]
n=500,p=0.3,$\beta = 5$	0.300	0.315 [0.25]	2.995	3.057 [0.65]
n=100,p=0.7,$\beta = 5$	-0.064	-0.065 [0.30]	0.864	0.947 [0.66]
n=500,p=0.7,$\beta = 5$	-0.013	-0.013 [0.14]	0.819	0.847 [0.33]
Weibull($\frac{1}{6}$,$\frac{3}{2}$)				
n=100,p=0.3,$\beta = 2$	0.153	0.182 [0.57]	3.145	3.502 [1.64]
n=500,p=0.3,$\beta = 2$	0.248	0.253 [0.24]	3.248	3.302 [0.71]
n=100,p=0.7,$\beta = 2$	-0.027	-0.035 [0.30]	0.766	0.857 [0.66]
n=500,p=0.7,$\beta = 2$	0.005	0.006 [0.14]	0.785	0.812 [0.31]
n=100,p=0.3,$\beta = 5$	0.200	0.223 [0.67]	3.816	4.107 [1.72]
n=500,p=0.3,$\beta = 5$	0.316	0.315 [0.28]	4.004	4.094 [0.87]
n=100,p=0.7,$\beta = 5$	-0.029	-0.038 [0.31]	0.976	1.070 [0.68]
n=500,p=0.7,$\beta = 5$	0.001	-0.002 [0.15]	0.958	1.004 [0.36]
Normal(6,4)				
n=100,p=0.3,$\beta = 2$	0.087	0.106 [0.39]	1.291	1.463 [0.90]
n=500,p=0.3,$\beta = 2$	0.087	0.090 [0.17]	1.422	1.470 [0.43]
n=100,p=0.7,$\beta = 2$	-0.081	-0.074 [0.27]	0.375	0.452 [0.58]
n=500,p=0.7,$\beta = 2$	-0.044	-0.048 [0.13]	0.394	0.413 [0.27]
n=100,p=0.3,$\beta = 5$	0.145	0.161 [0.42]	1.375	1.536 [1.02]
n=500,p=0.3,$\beta = 5$	0.127	0.129 [0.19]	1.526	1.583 [0.45]
n=100,p=0.7,$\beta = 5$	-0.062	-0.061 [0.29]	0.438	0.514 [0.62]
n=500,p=0.7,$\beta = 5$	-0.070	-0.070 [0.13]	0.491	0.516 [0.28]

Table 3.4: Skewness and kurtosis for the midpoint residuals $\hat{e}_{mid}$

	Skewness		**Kurtosis**	
	Median	**Mean [Std]**	**Median**	**Mean [Std]**
Exponential($\frac{1}{8}$)				
n=100,p=0.3,$\beta = 2$	0.244	0.220 [0.52]	2.088	2.213 [1.14]
n=500,p=0.3,$\beta = 2$	0.293	0.288 [0.23]	2.442	2.484 [0.54]
n=100,p=0.7,$\beta = 2$	0.142	0.118 [0.46]	0.653	1.026 [1.51]
n=500,p=0.7,$\beta = 2$	0.158	0.170 [0.23]	0.986	1.215 [0.96]
n=100,p=0.3,$\beta = 5$	0.216	0.244 [0.59]	2.575	2.818 [1.33]
n=500,p=0.3,$\beta = 5$	0.328	0.322 [0.26]	3.059	3.083 [0.56]
n=100,p=0.7,$\beta = 5$	0.544	0.545 [1.10]	4.591	5.556 [4.24]
n=500,p=0.7,$\beta = 5$	0.642	0.637 [0.64]	6.327	7.055 [3.22]
Weibull($\frac{1}{6}$,$\frac{3}{2}$)				
n=100,p=0.3,$\beta = 2$	-0.038	-0.035 [0.22]	0.841	0.925 [0.68]
n=500,p=0.3,$\beta = 2$	0.008	0.006 [0.11]	0.992	0.991 [0.30]
n=100,p=0.7,$\beta = 2$	-0.021	-0.027 [0.23]	0.364	0.645 [1.01]
n=500,p=0.7,$\beta = 2$	0.004	0.006 [0.11]	0.650	0.782 [0.62]
n=100,p=0.3,$\beta = 5$	-0.041	-0.042 [0.23]	1.490	1.606 [0.80]
n=500,p=0.3,$\beta = 5$	0.005	0.006 [0.11]	1.590	1.596 [0.32]
n=100,p=0.7,$\beta = 5$	-0.016	-0.024 [0.22]	3.660	4.522 [3.62]
n=500,p=0.7,$\beta = 5$	0.009	0.010 [0.11]	5.517	6.025 [2.54]
Normal(6,4)				
n=100,p=0.3,$\beta = 2$	0.723	0.720 [0.23]	0.362	0.435 [0.58]
n=500,p=0.3,$\beta = 2$	0.725	0.723 [0.10]	0.397	0.410 [0.24]
n=100,p=0.7,$\beta = 2$	0.396	0.439 [0.42]	0.599	1.020 [1.46]
n=500,p=0.7,$\beta = 2$	0.446	0.458 [0.19]	0.906	1.075 [0.80]
n=100,p=0.3,$\beta = 5$	0.924	0.932 [0.23]	0.572	0.619 [0.58]
n=500,p=0.3,$\beta = 5$	0.946	0.942 [0.10]	0.577	0.586 [0.26]
n=100,p=0.7,$\beta = 5$	1.432	1.459 [0.67]	3.484	4.520 [3.62]
n=500,p=0.7,$\beta = 5$	1.554	1.576 [0.38]	4.964	5.590 [2.60]

Table 3.5: Skewness and kurtosis for the midpoint residuals $\hat{e}_{lup}$

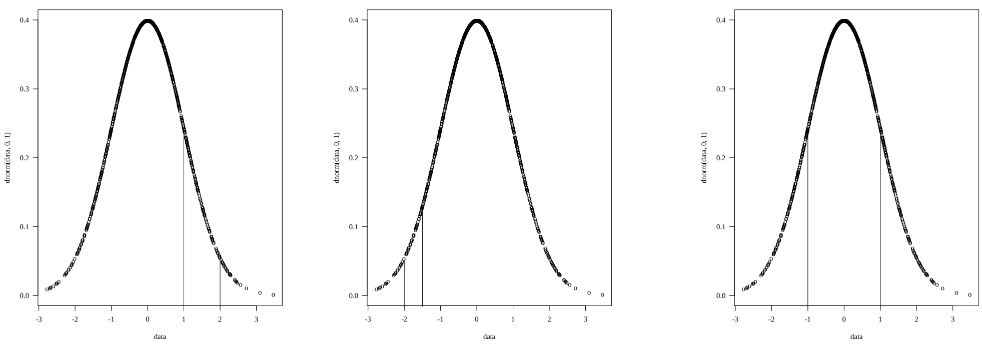
	Skewness		**Kurtosis**	
	Median	**Mean [Std]**	**Median**	**Mean [Std]**
Exponential($\frac{1}{8}$)				
n=100,p=0.3,$\beta = 2$	1.046	0.732 [1.20]	3.013	3.511 [2.37]
n=500,p=0.3,$\beta = 2$	1.372	0.983 [1.17]	4.361	4.609 [1.69]
n=100,p=0.7,$\beta = 2$	-0.141	-0.169 [0.89]	1.649	2.818 [3.92]
n=500,p=0.7,$\beta = 2$	-0.271	-0.348 [0.58]	2.328	3.517 [4.14]
n=100,p=0.3,$\beta = 5$	1.087	0.738 [1.32]	3.252	3.961 [2.65]
n=500,p=0.3,$\beta = 5$	1.499	1.054 [1.31]	5.063	5.290 [1.95]
n=100,p=0.7,$\beta = 5$	-0.297	-0.337 [1.70]	6.768	8.538 [6.51]
n=500,p=0.7,$\beta = 5$	-0.882	-0.942 [1.32]	10.509	13.273 [10.11]
Weibull($\frac{1}{6}$,$\frac{3}{2}$)				
n=100,p=0.3,$\beta = 2$	0.497	0.488 [0.37]	0.302	0.394 [0.57]
n=500,p=0.3,$\beta = 2$	0.463	0.476 [0.176]	0.351	0.384 [0.27]
n=100,p=0.7,$\beta = 2$	0.395	0.408 [0.596]	1.319	1.741 [1.64]
n=500,p=0.7,$\beta = 2$	0.529	0.517 [0.384]	2.039	2.177 [0.98]
n=100,p=0.3,$\beta = 5$	0.596	0.609 [0.402]	0.564	0.654 [0.66]
n=500,p=0.3,$\beta = 5$	0.559	0.580 [0.201]	0.486	0.522 [0.30]
n=100,p=0.7,$\beta = 5$	0.767	0.774 [0.986]	3.894	4.702 [3.13]
n=500,p=0.7,$\beta = 5$	1.181	1.087 [0.735]	5.629	6.008 [2.11]
Normal(6,4)				
n=100,p=0.3,$\beta = 2$	-1.239	-1.263 [0.41]	1.835	2.282 [1.91]
n=500,p=0.3,$\beta = 2$	-1.321	-1.320 [0.19]	2.292	2.403 [0.91]
n=100,p=0.7,$\beta = 2$	-0.828	-0.894 [0.54]	1.448	2.163 [2.59]
n=500,p=0.7,$\beta = 2$	-0.875	-0.906 [0.28]	2.171	2.527 [1.64]
n=100,p=0.3,$\beta = 5$	-1.638	-1.641 [0.47]	2.682	3.095 [2.30]
n=500,p=0.3,$\beta = 5$	-1.811	-1.807 [0.21]	3.495	3.621 [1.12]
n=100,p=0.7,$\beta = 5$	-2.201	-2.269 [0.72]	5.880	7.321 [5.19]
n=500,p=0.7,$\beta = 5$	-2.509	-2.541 [0.41]	8.921	9.666 [3.94]

from the uncensored residuals $\hat{e}$, the second one using the newly proposed residuals $\hat{e}^*$, the third one coming from the midpoint residuals $\hat{e}_{mid}$, and the last one using the residuals $\hat{e}_{lup}$.

Considering the different residual plots in Appendix A, one can observe that the least squares residuals $\hat{e}_i$ scatter randomly in the plane and show no special patterns throughout the different simulation scenarios, thus confirming the correctly specified model.

In the plots using the $\hat{e}^*$, a curve can be noticed that is mostly zero but at large values goes up and at small values goes down (see for example Scenario 2). In the following this special shape will be referred to as "S-shape". To understand where this pattern comes from, one has to look at the generation mechanism for the residuals $\hat{\eta}_i$, those residuals which come from the interval censored data: Their values depend on the residual intervals $[\hat{A}_i, \hat{B}_i]$, which on their part determine where the corresponding error normal distribution is to be truncated. So, if both $\hat{A}_i$ and $\hat{B}_i$ are large (small), then the resulting $\hat{\eta}_i$ gets large (small) as well. In case that $\hat{A}_i$ and $\hat{B}_i$ are of opposite sign, the resulting value for $\hat{\eta}_i$ is more probable to be around zero. Figure 3.1 illustrates this idea.

Figure 3.1: Truncation schemes for the residual distribution depending on the values of the residual intervals $[\hat{A}_i, \hat{B}_i]$



In a correctly specified model, the values for $\hat{A}_i$ and $\hat{B}_i$ will be mostly of opposite sign, but in some occasions both $\hat{A}_i$ and $\hat{B}_i$ will be small or large, leading to the appearance of the S-shape. So, when interpreting a residual plot using the $\hat{e}^*$, it is nothing unusual to encounter the S-form pattern but it is not imperative either. The S-shaped curve does not point at possible model violations but is an inherent structure of these residuals when the model is correctly specified. The generation mechanism of the S-shape will be studied more extensively in Chapter 3.5.

The performance of the $\hat{e}_{mid}$ differs from scenario to scenario. It can be noticed that especially for a high percentage of censoring (p=0.3), the plot resembles a growing and then falling variance of the residuals (see for example Scenario 6) which would lead to the wrong conclusion of a not constant error variance. This makes it difficult to use them for regression diagnostics.

In the plots coming from the $\hat{e}_{lup}$, one finds very often a certain $\hat{y}$-value for which these residuals have a far bigger variance than otherwise (for example in Scenario 6). The distribution of these residuals within the plot does not seem to follow a special pattern but they are not evenly spread in the plane, either. Using this plot in regression diagnostic could therefore cause irritations about possible model deviations.

3.3 Checking for deviations from linearity

Appendix B shows the simulated residual plots when the true model includes a quadratic term but the fitted model is only linear. That is, the true response values y_i are generated from the model $y_i = \alpha + \beta_1 z_i + \beta_2 z_i^2 + \varepsilon_i$ but the residuals are calculated using only the linear relationship $y_i = \alpha + \beta_1 z_i + \varepsilon_i$. Following the residual theory for uncensored data, the residual plots should reveal the misspecified model by showing a quadratic structure in the plotted points.

As the previous simulation showed that the performance of the residuals does not vary between $n = 100$ and 500, and because it is of general interest to examine the small sample size behavior of the residuals, $n = 500$ is dropped and replaced by $n = 30$, but $n = 100$ is still kept.

From the 24 simulation scenarios shown in Appendix B, it can be seen that the least square residuals $\hat{e}$ perfectly reproduce the hidden quadratic structure in the data.

The $\hat{e}^*$ reflect the quadratic structure well in all scenarios where the covariate distribution is exponential and in most scenarios for a Weibull distributed covariate. For the normal distribution, the quadratic structure can be seen in those scenarios where $n = 100$ whereas for $n = 30$ the pattern is not that clear. What strikes in especially those plots where the percentage of censoring is high (p=0.3), is the line of residuals at zero (for example in Scenario 1). These points are the values of those $\hat{\eta}$ which are calculated from the estimated truncated error distribution and result mostly zero because of the following facts: the y_i are generated from the model with the additional quadratic term $\beta_2 z_i^2$. As a consequence, the values for the y_i result very large. These large y_i-values are used in the estimation of the $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\sigma}^2$ (the formulas are given in Chapter 1.3 of the first part of this thesis) with the consequence that these estimates result very large as well. As the values of the estimated model parameters enter in the calculation of the residual intervals $[\hat{A}_i, \hat{B}_i]$, and above all $\hat{\beta}$ has a huge influence as a multiplier of the z-values, the values of $\hat{A}_i$ and $\hat{B}_i$ result being of opposite sign more often as in the correctly specified model, and large or small values for both $\hat{A}_i$ and $\hat{B}_i$ do almost not occur. Though, the line at zero does not disturb the quadratic pattern of the uncensored residuals in the plot and it still can be clearly recognized. For a low percentage of censoring, the quadratic structure of the uncensored residuals is very dominant anyway.

The residuals $\hat{e}_{mid}$ and $\hat{e}_{lup}$ do not seem to be able to reflect the missing quadratic term in the model and neither do they seem to support the quadratic structure of the uncensored residuals. In contrary, they often make it rather impossible to recognize the pattern of the uncensored residuals (see for example Scenario 5). In all scenarios for the normal distribution as well as in most scenarios for the Weibull distribution, the residual plots do not show a quadratic curve. In case of a exponentially distributed covariate the quadratic structure is only reflected when the percentage of censoring is low.

3.4 Checking for constant variance

All computationally reasonable test procedures for heteroscedasticity are based on the assumption of normality of the residuals. As seen previously, normality is not given for the $\hat{e}_{mid}$, $\hat{e}_{lup}$ and $\hat{e}^*$. So, the ordinary residual plot is used again to check this assumption for the four different types of residuals.

Simulations are carried out for a linear model where the error variance is not constant but depends on the covariate. Appendix C shows the 24 simulated scenarios where ε_i is generated from a normal distribution with mean zero and variance x_i^2. All four types of residuals perform similarly and it can be seen that for $n = 100$ most scenarios show the growing variance of the residuals as the values of the covariate get larger (see for example Scenario 1). For a small number of observations, though, the variance structure is not resembled at all, which leads to the conclusion that there should be a reasonable large number of observations when using the residual plots for regression diagnostics.

3.5 Examining the S-shape

As seen in the previous simulations, the performance of the residuals $\hat{e}^*$ vary considerately depending on the percentage of censoring in the data. Another important factor affecting these residuals is the width of the residual interval, as mentioned in connection with the appearance of the S-shape. The influence of these two factors will now be examined more extensively by studying one data scenario under various censoring levels and interval widths. The censoring level p will range from 0.1 to 0.9 (in steps of 0.2), and the residual interval width will be increased by 0, 0.3, 0.5 and 1 times the original interval width. The model under consideration will be specified by $\alpha = 4$ and $\beta = 2$, the distribution of the covariate is chosen to be exponential with mean $\frac{1}{8}$, and the number of observations n will be 100.

First, simulations for a correctly specified linear model and constant error variance are carried out and the results are shown in Figures 1-5. As expected, at high censoring levels (p=0.1 to 0.5) and a small residual interval width, the typical "S-form" as described before can be observed. With growing interval width, though, this structure disappears, and at the end there is only one straight line at zero. This behavior is reasonable because with growing interval width the truncated normal error distribution approximates better to the not truncated $N(0, \sigma^2)$ distribution, and the resulting means are therefore mostly zero.

When the true model is quadratic but the residuals are calculated using only a linear term (see Figures 6-10), it can be observed that the quadratic structure is not visible at a high percentage of censoring (p=0.1), especially when the residual intervals are very wide and produce residuals with values near zero. In contrast, the quadratic data structure is resembled quite well for higher percentages of censoring, again with the observation that wide residual intervals produce a line of residuals at zero.

Figures 11-15 show the residual plots for the case that the model is linear but the error variance depends on the covariate values. Here, it can be seen that both the percentage of censoring and the residual interval width do not affect the shape of the residuals in the plots, and the growing error variance is resembled well in all cases.

3.6 Summary of the simulation results

The simulation results can be summarized in the following way: For checking an underlying normal distribution of the model errors, none of the three types of residuals coming from interval censored data can be used, though the newly proposed residuals $\hat{e}^*$ perform best. With respect to checking whether the included variables specify the fitted model correctly, the simulations showed that the residuals $\hat{\eta}$ are able to detect missing terms in the model in all scenarios when the number of observations is sufficiently large. For a small number of observations, they still perform satisfactorily in case of an exponential or Weibull distributed covariate. In contrast, the residuals $\hat{e}_{lup}$ and $\hat{e}_{mid}$ perform well only in case of a low percentage of censoring and an exponentially distributed covariate. All three types of residuals can be used to detect a covariate depending error variance as long as there is a sufficiently large number of observations.

Figure 1: Correctly specified model, censoring level 0.1. First graph: original residual interval width, second graph: 0.3 times increased, third graph: 0.5 times increased, fourth graph: double of the original interval width.

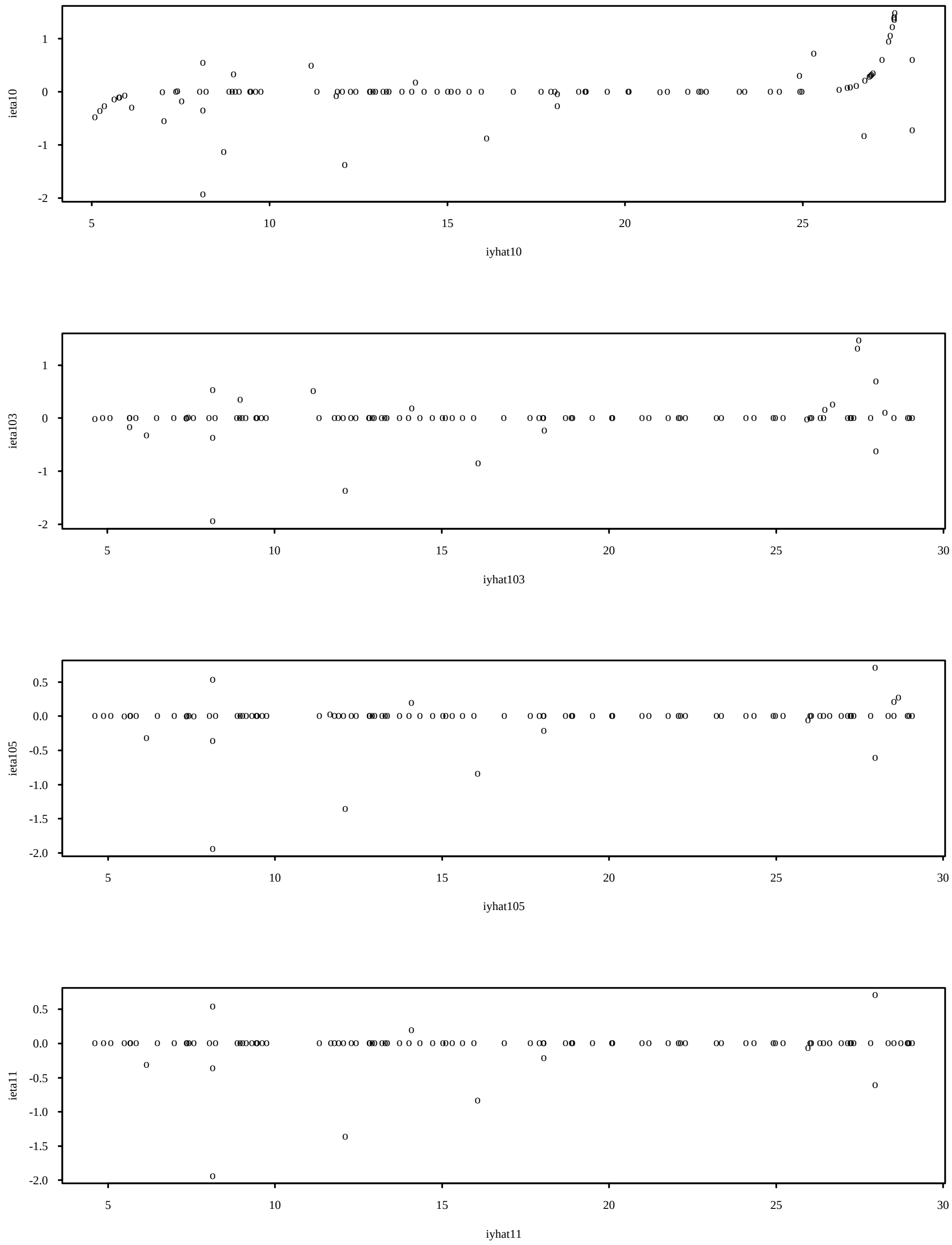


Figure 2: Correctly specified model, censoring level 0.3. First graph: original residual interval width, second graph: 0.3 times increased, third graph: 0.5 times increased, fourth graph: double of the original interval width.

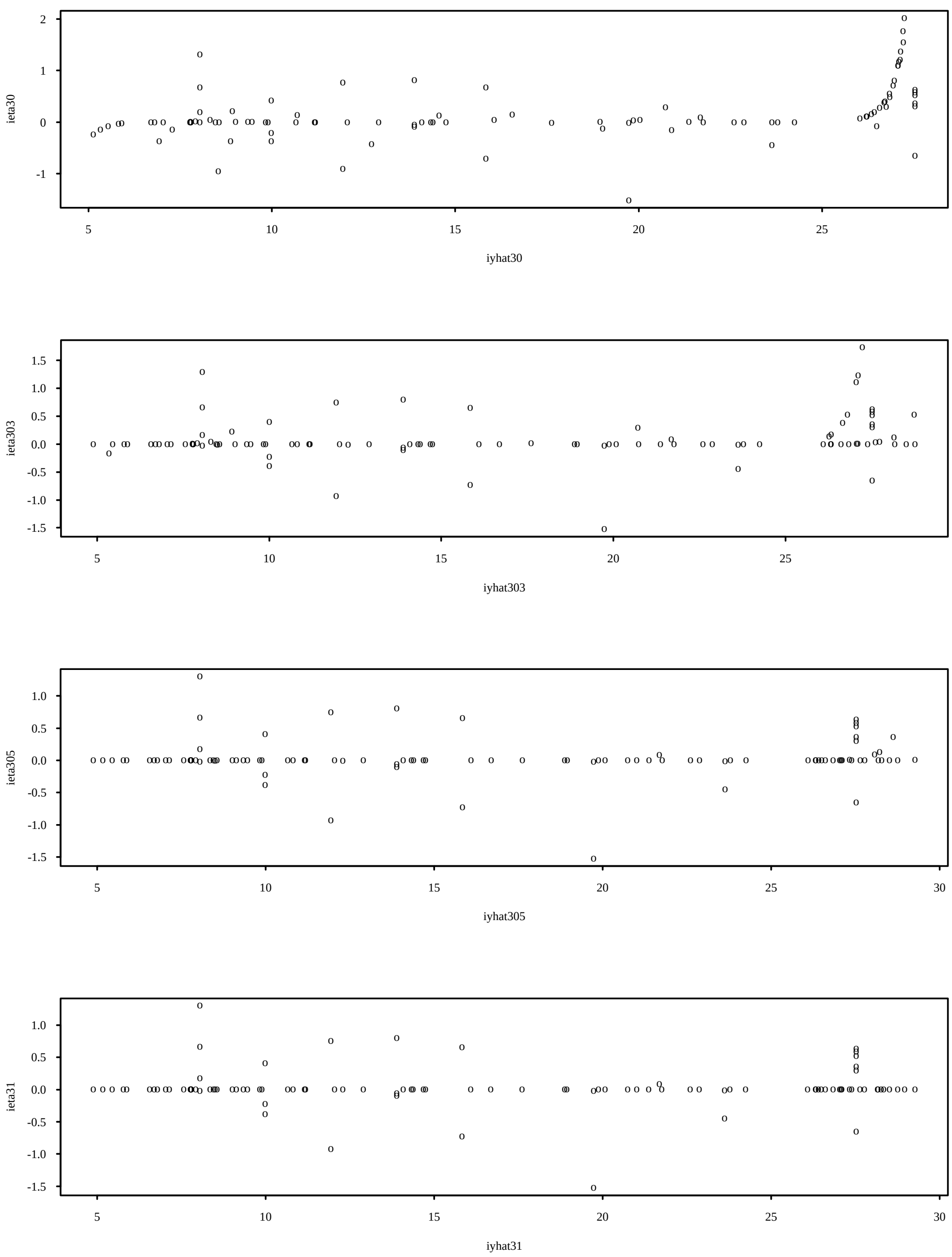


Figure 3: Correctly specified model, censoring level 0.5. First graph: original residual interval width, second graph: 0.3 times increased, third graph: 0.5 times increased, fourth graph: double of the original interval width.

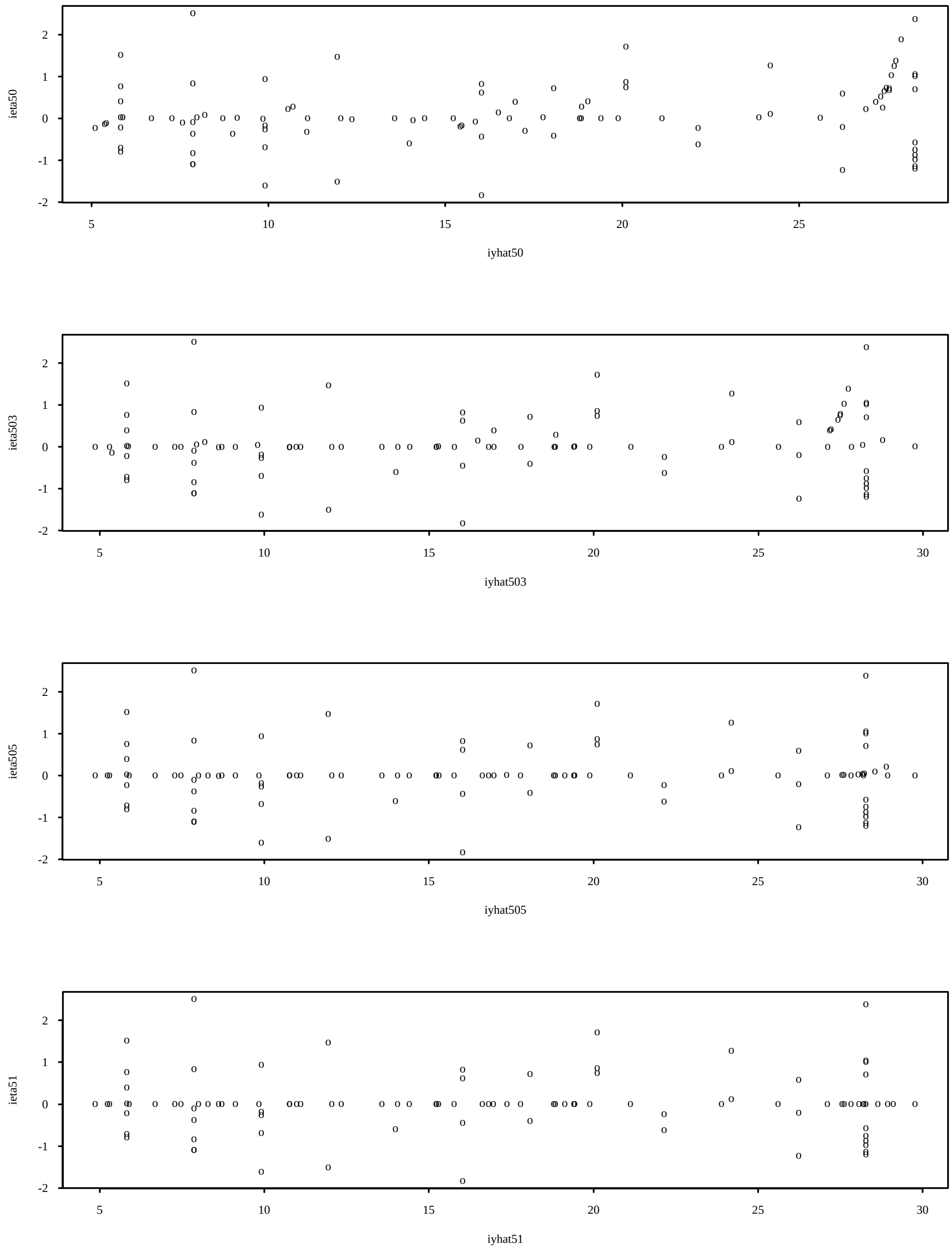


Figure 4: Correctly specified model, censoring level 0.7. First graph: original residual interval width, second graph: 0.3 times increased, third graph: 0.5 times increased, fourth graph: double of the original interval width.

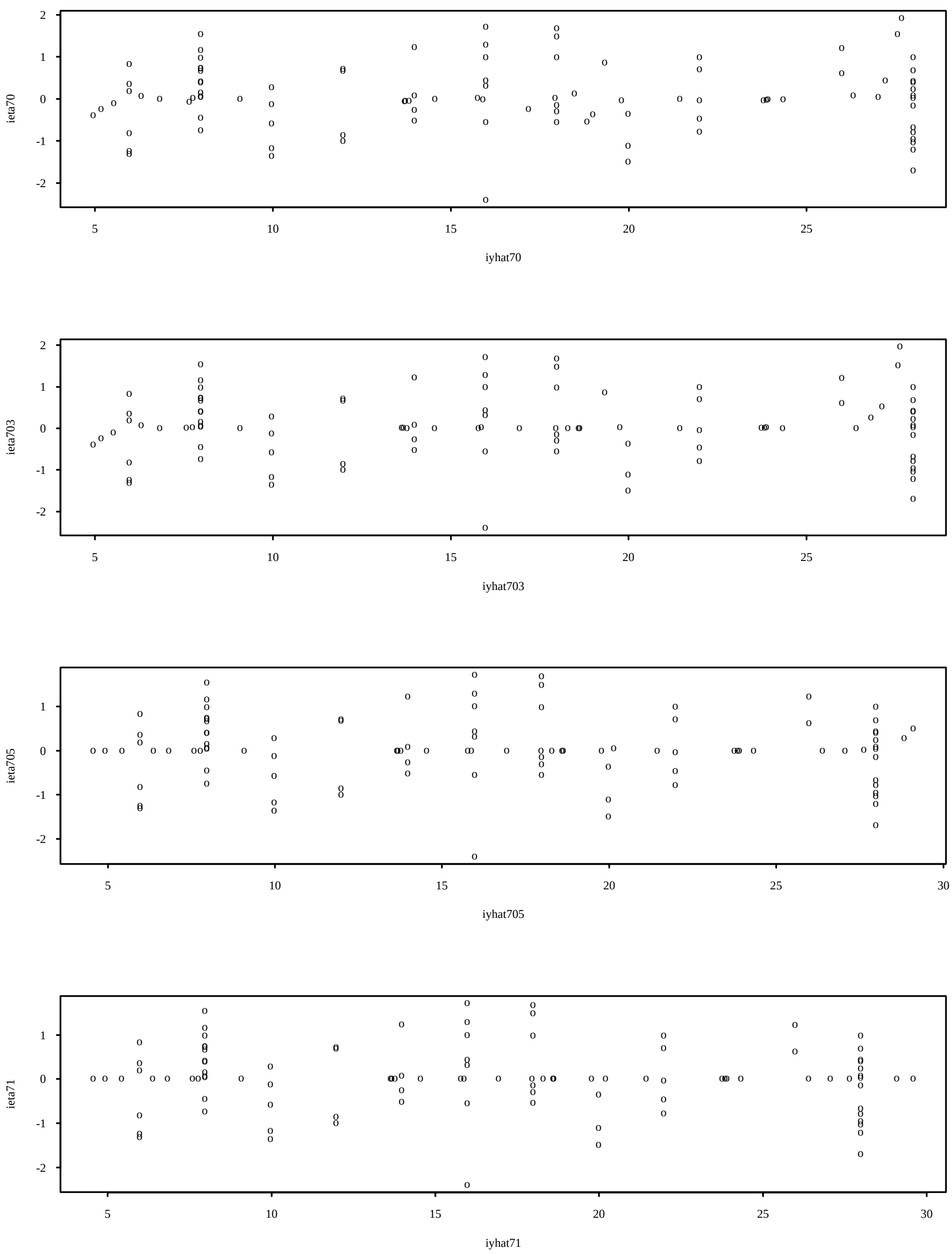


Figure 5: Correctly specified model, censoring level 0.9. First graph: original residual interval width, second graph: 0.3 times increased, third graph: 0.5 times increased, fourth graph: double of the original interval width.

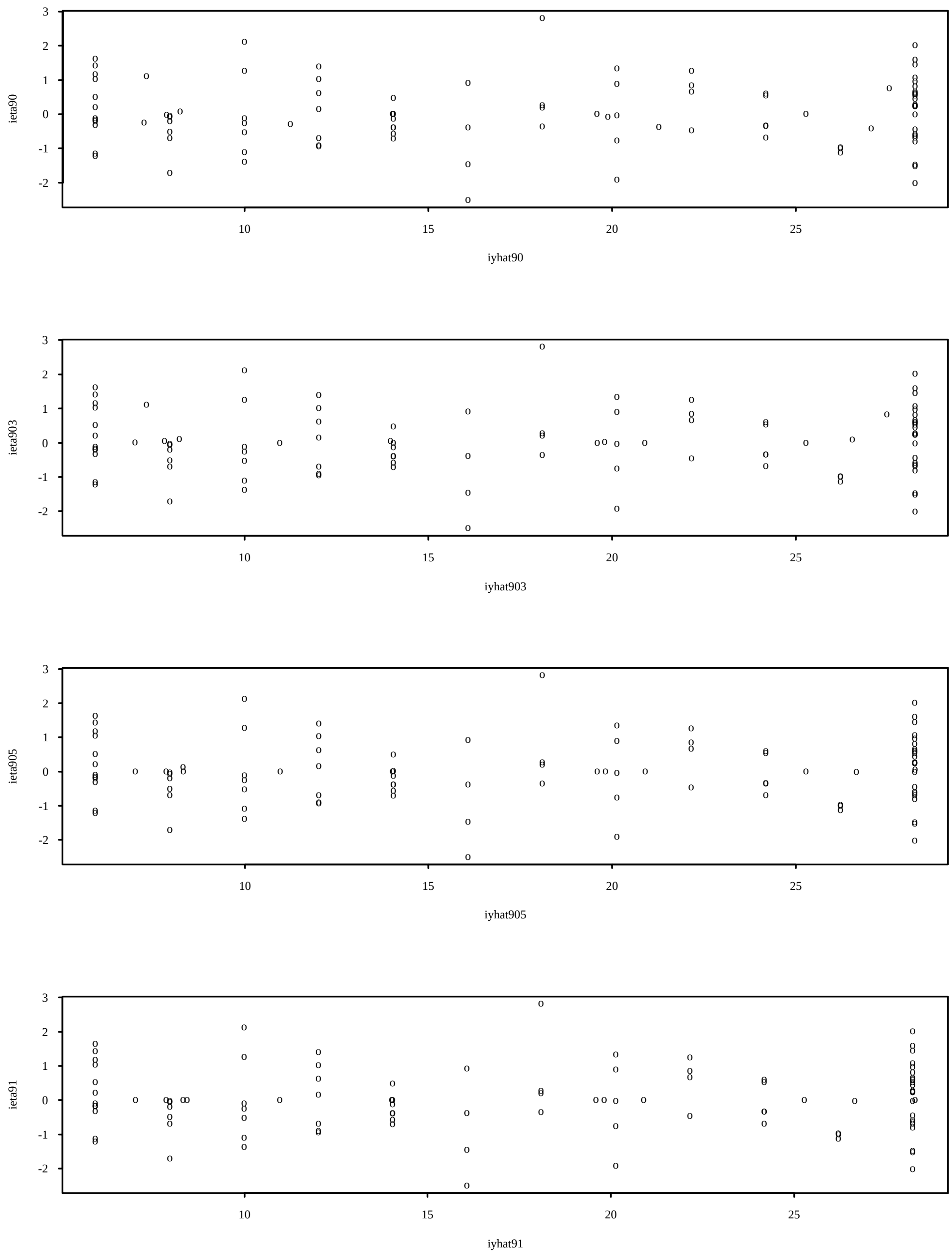


Figure 6: Quadratic model, censoring level 0.1. First graph: original residual interval width, second graph: 0.3 times increased, third graph: 0.5 times increased, fourth graph: double of the original interval width.

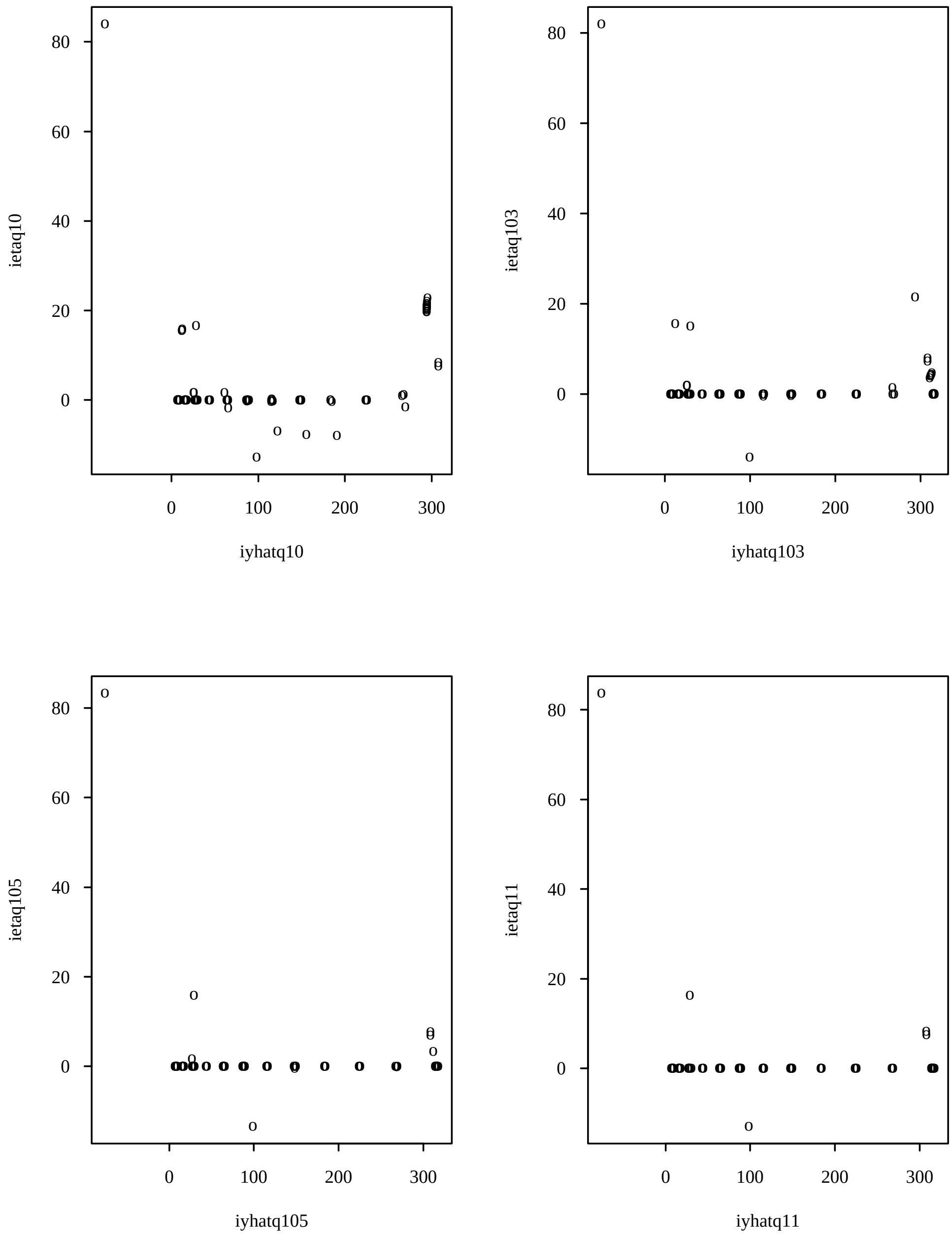


Figure 7: Quadratic model, censoring level 0.3. First graph: original residual interval width, second graph: 0.3 times increased, third graph: 0.5 times increased, fourth graph: double of the original interval width.

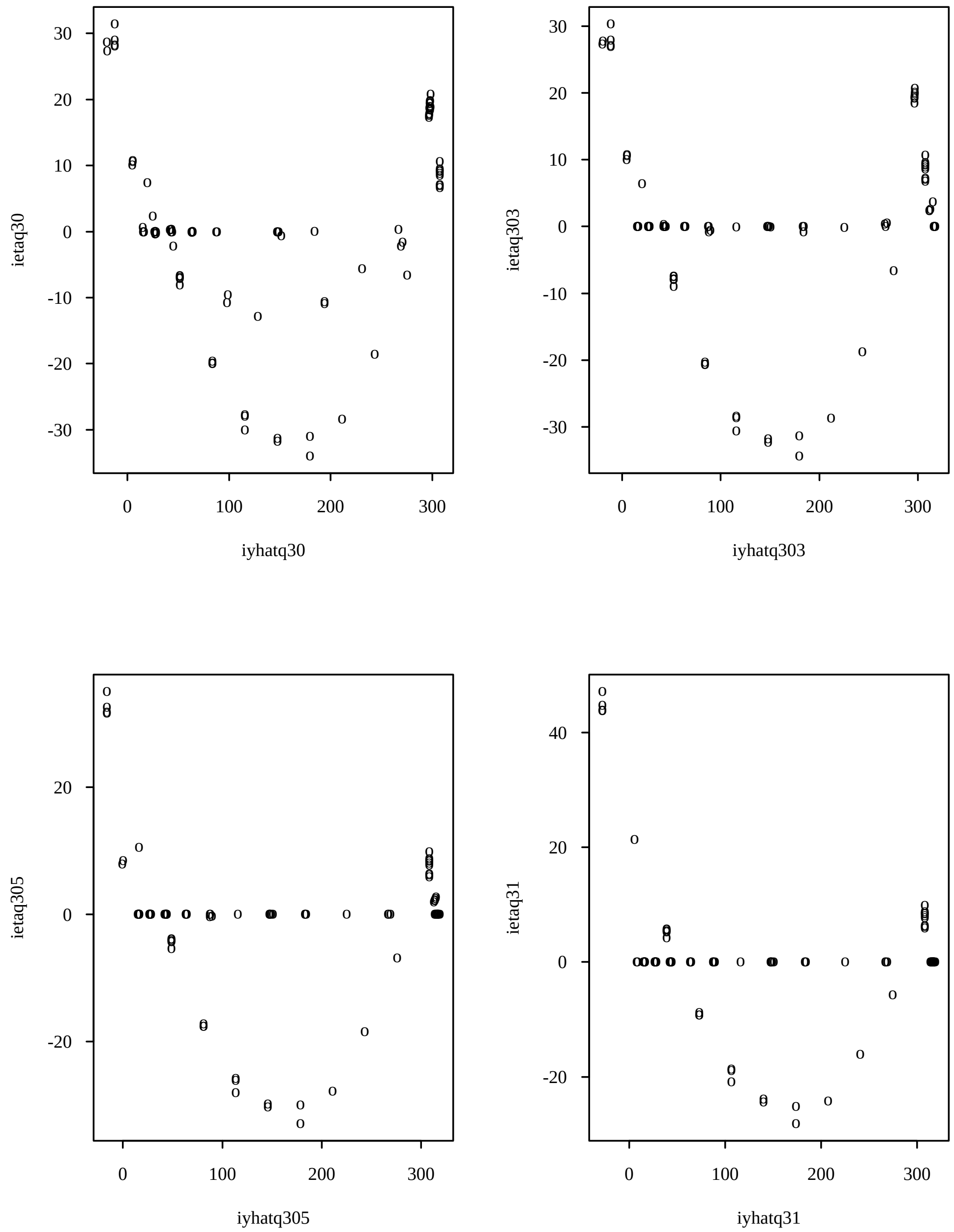


Figure 8: Quadratic model, censoring level 0.5. First graph: original residual interval width, second graph: 0.3 times increased, third graph: 0.5 times increased, fourth graph: double of the original interval width.

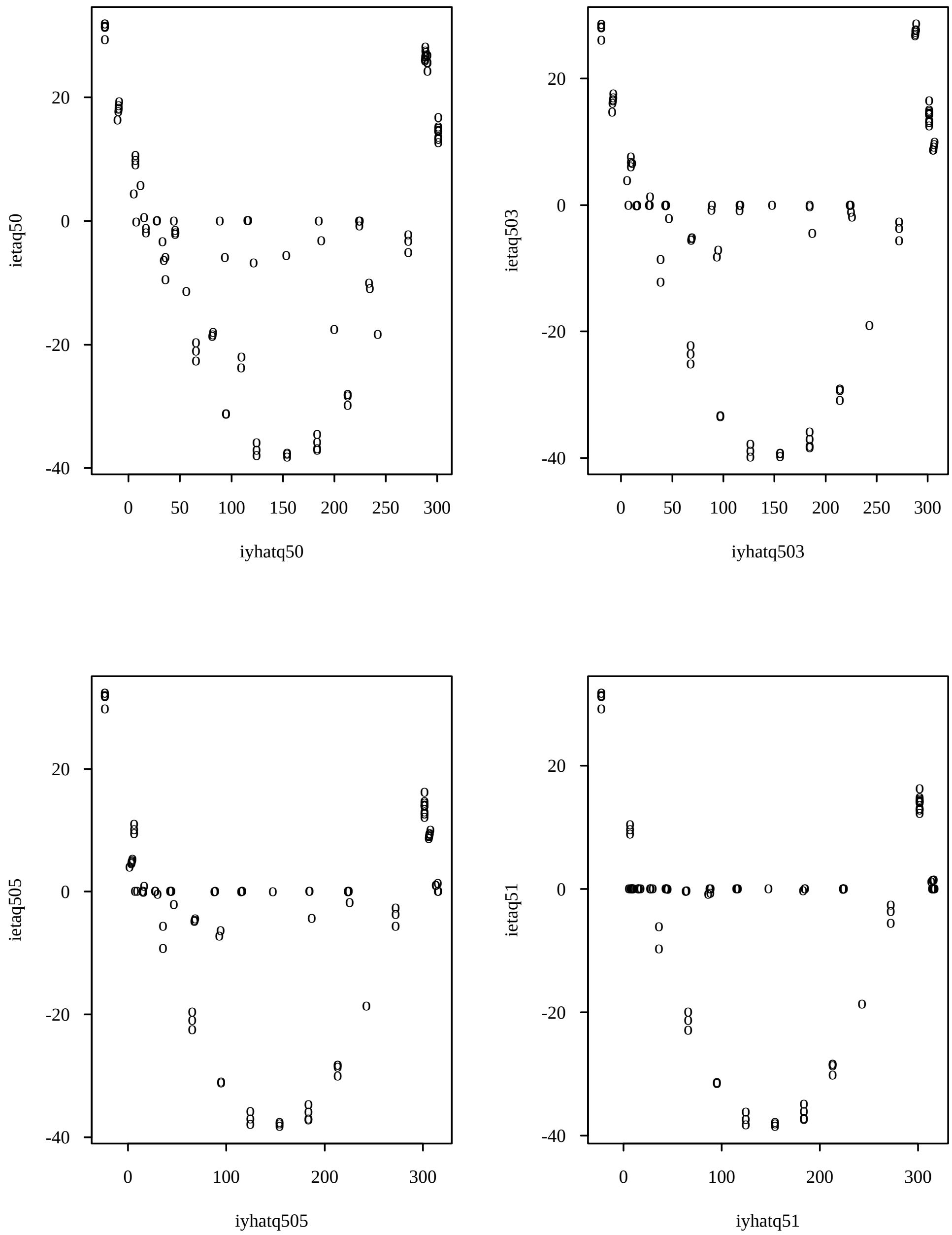


Figure 9: Quadratic model, censoring level 0.7. First graph: original residual interval width, second graph: 0.3 times increased, third graph: 0.5 times increased, fourth graph: double of the original interval width.

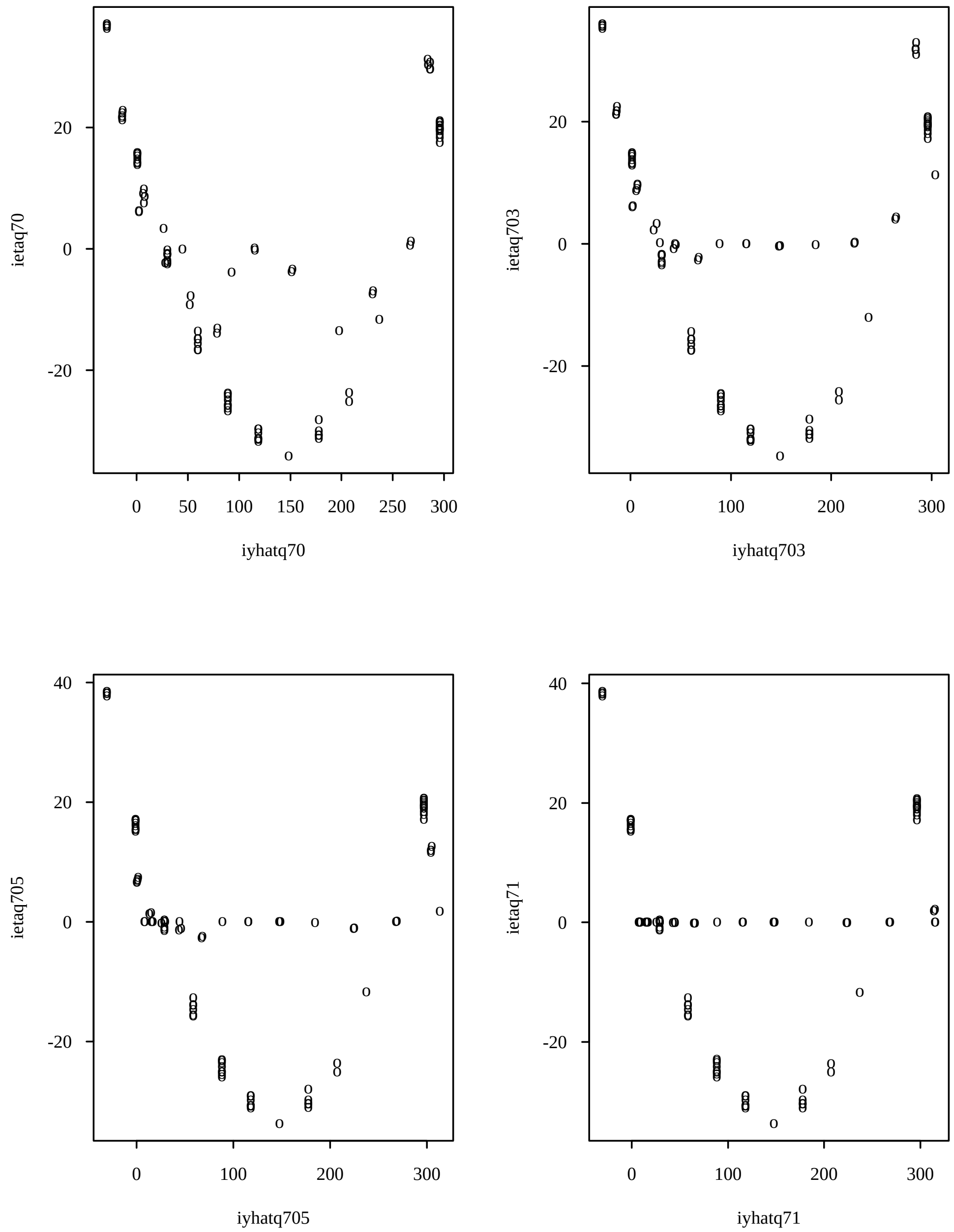


Figure 10: Quadratic model, censoring level 0.9. First graph: original residual interval width, second graph: 0.3 times increased, third graph: 0.5 times increased, fourth graph: double of the original interval width.

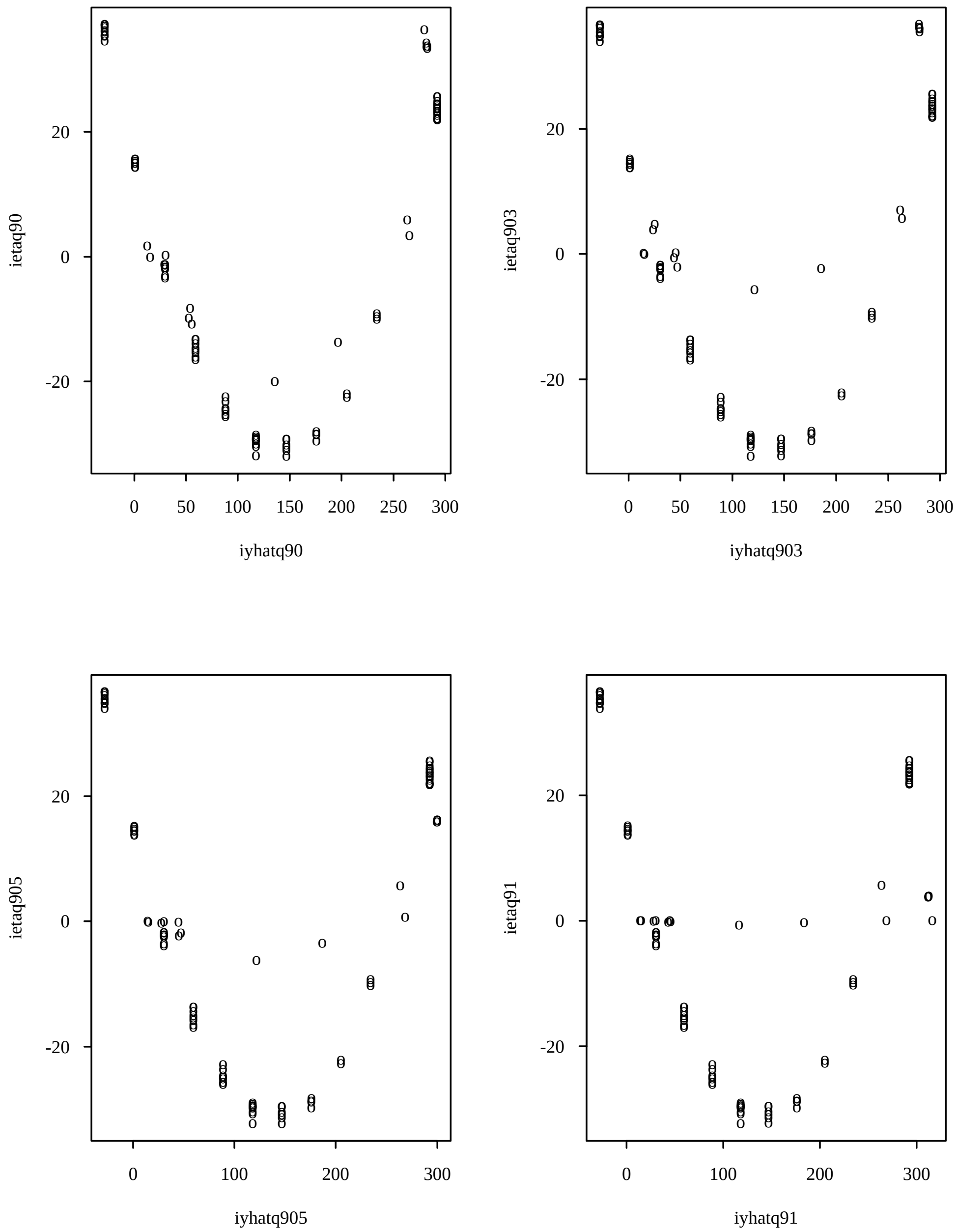


Figure 11: Covariate depending model, censoring level 0.1. First graph: original residual interval width, second graph: 0.3 times increased, third graph: 0.5 times increased, fourth graph: double of the original interval width.

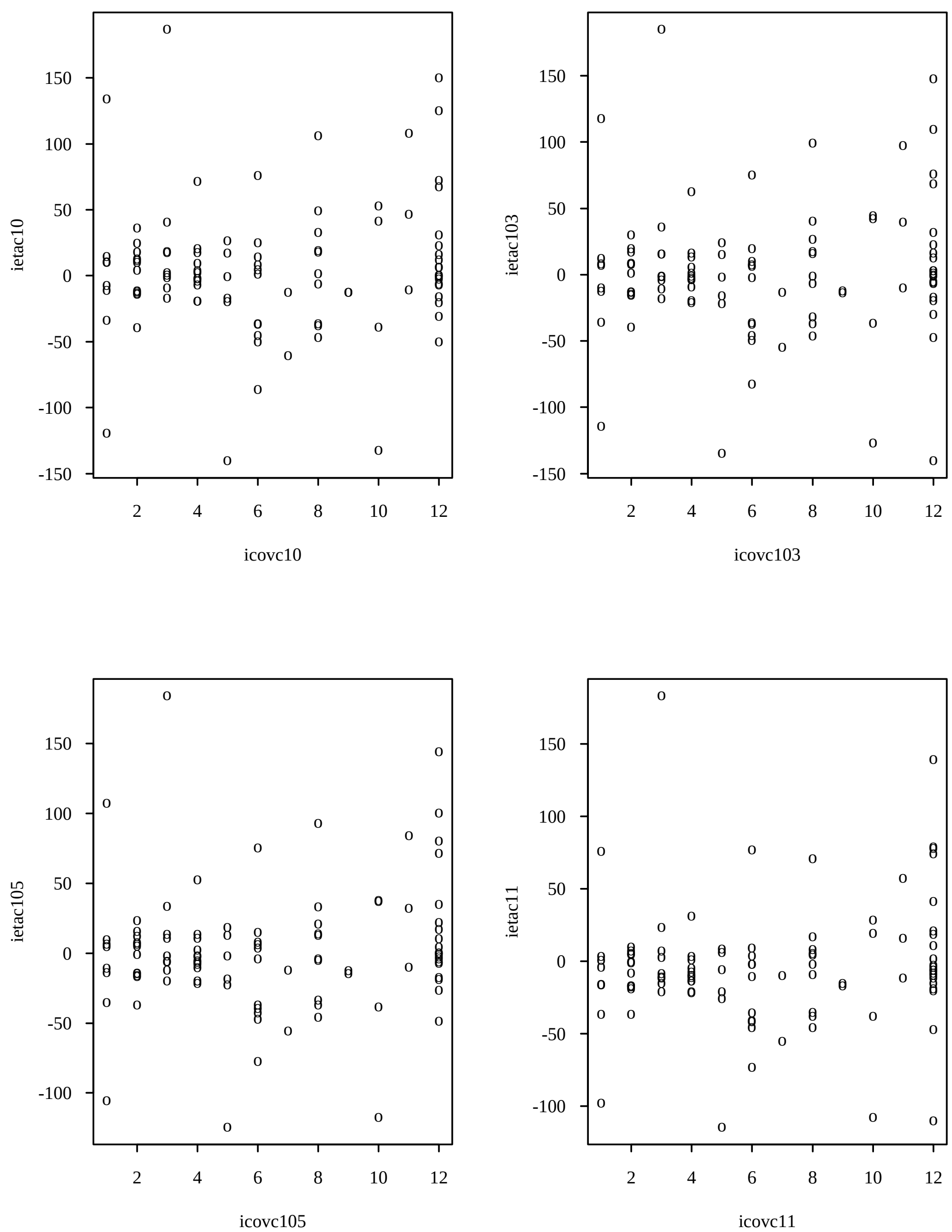


Figure 12: Covariate depending model, censoring level 0.3. First graph: original residual interval width, second graph: 0.3 times increased, third graph: 0.5 times increased, fourth graph: double of the original interval width.

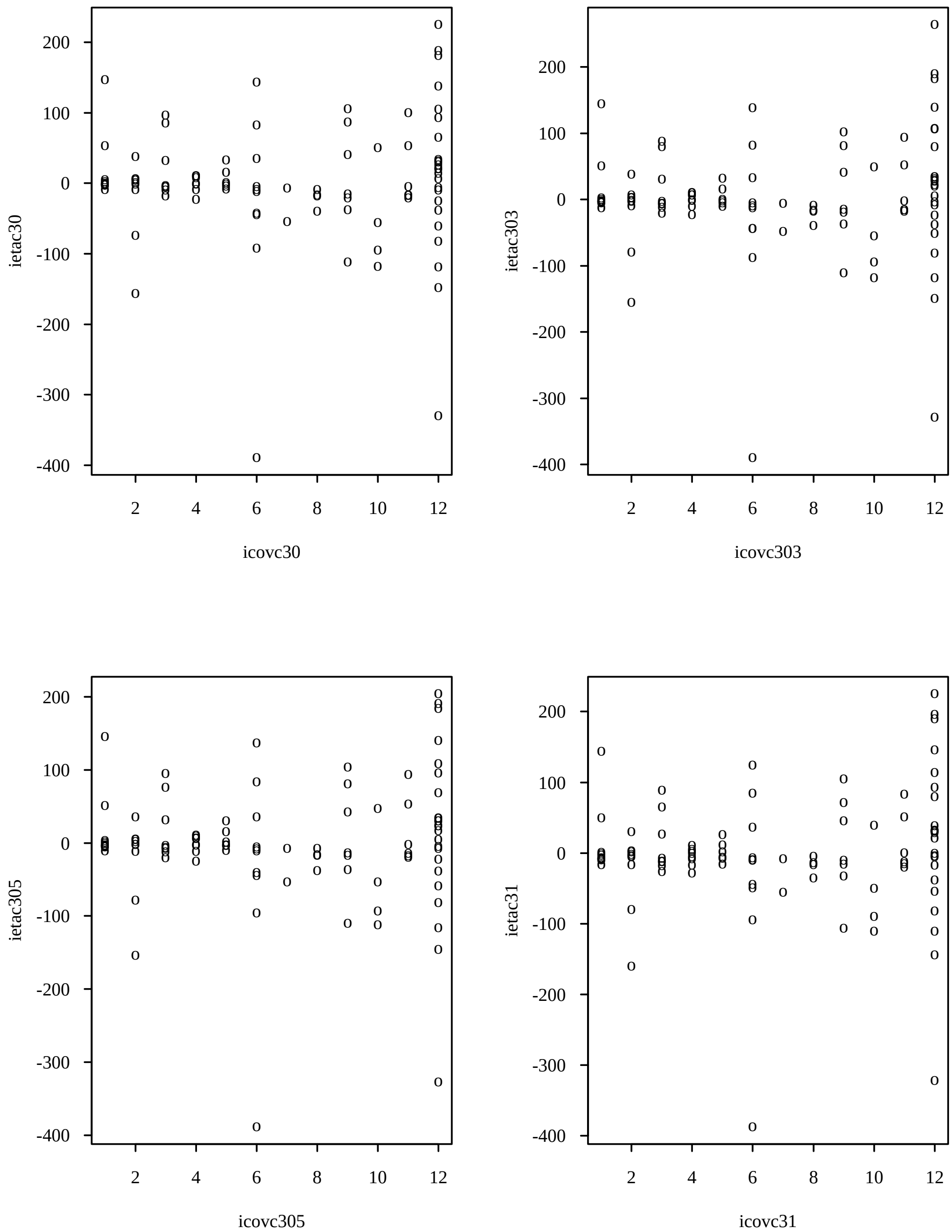


Figure 13: Covariate depending model, censoring level 0.5. First graph: original residual interval width, second graph: 0.3 times increased, third graph: 0.5 times increased, fourth graph: double of the original interval width.

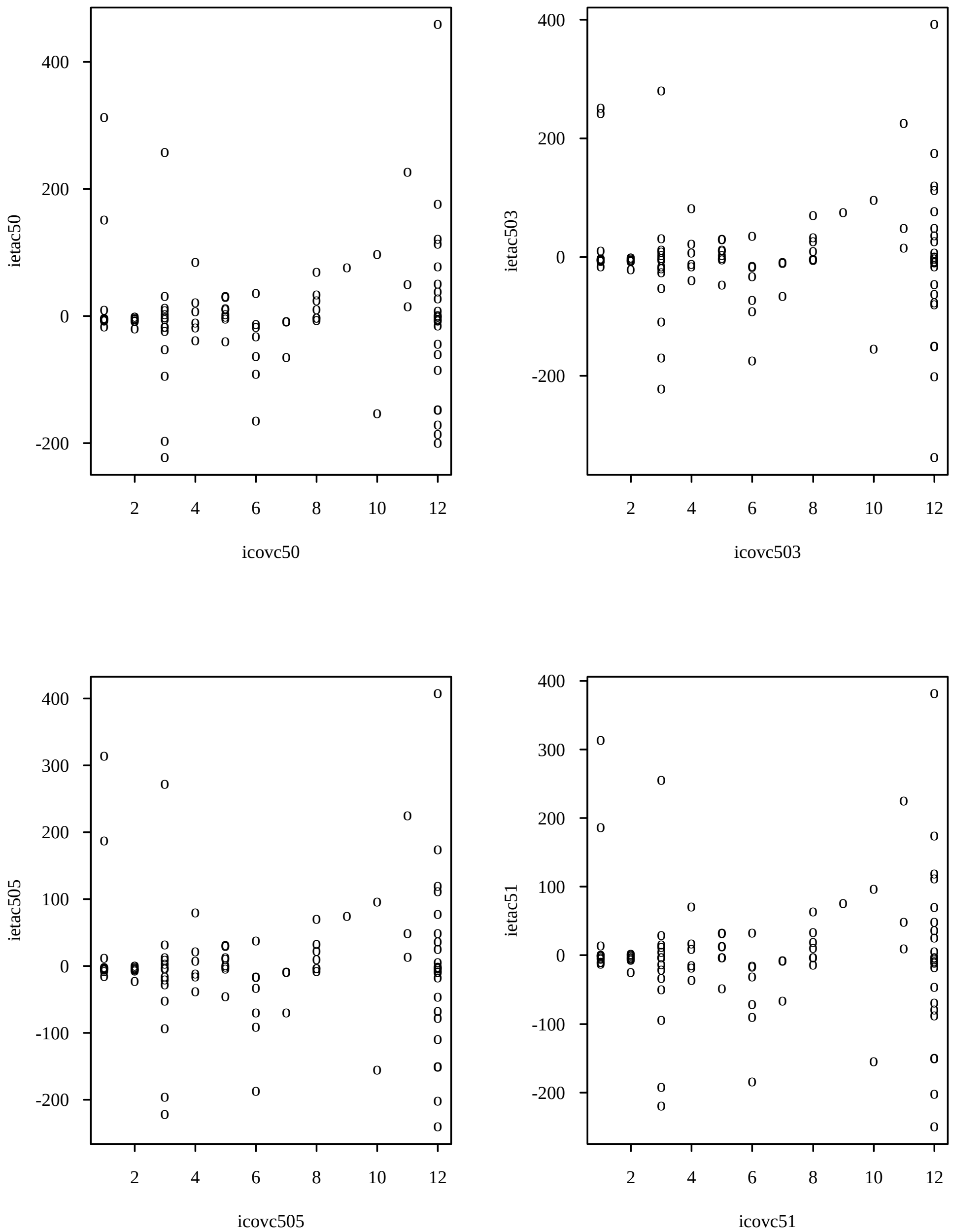


Figure 14: Covariate depending model, censoring level 0.7. First graph: original residual interval width, second graph: 0.3 times increased, third graph: 0.5 times increased, fourth graph: double of the original interval width.

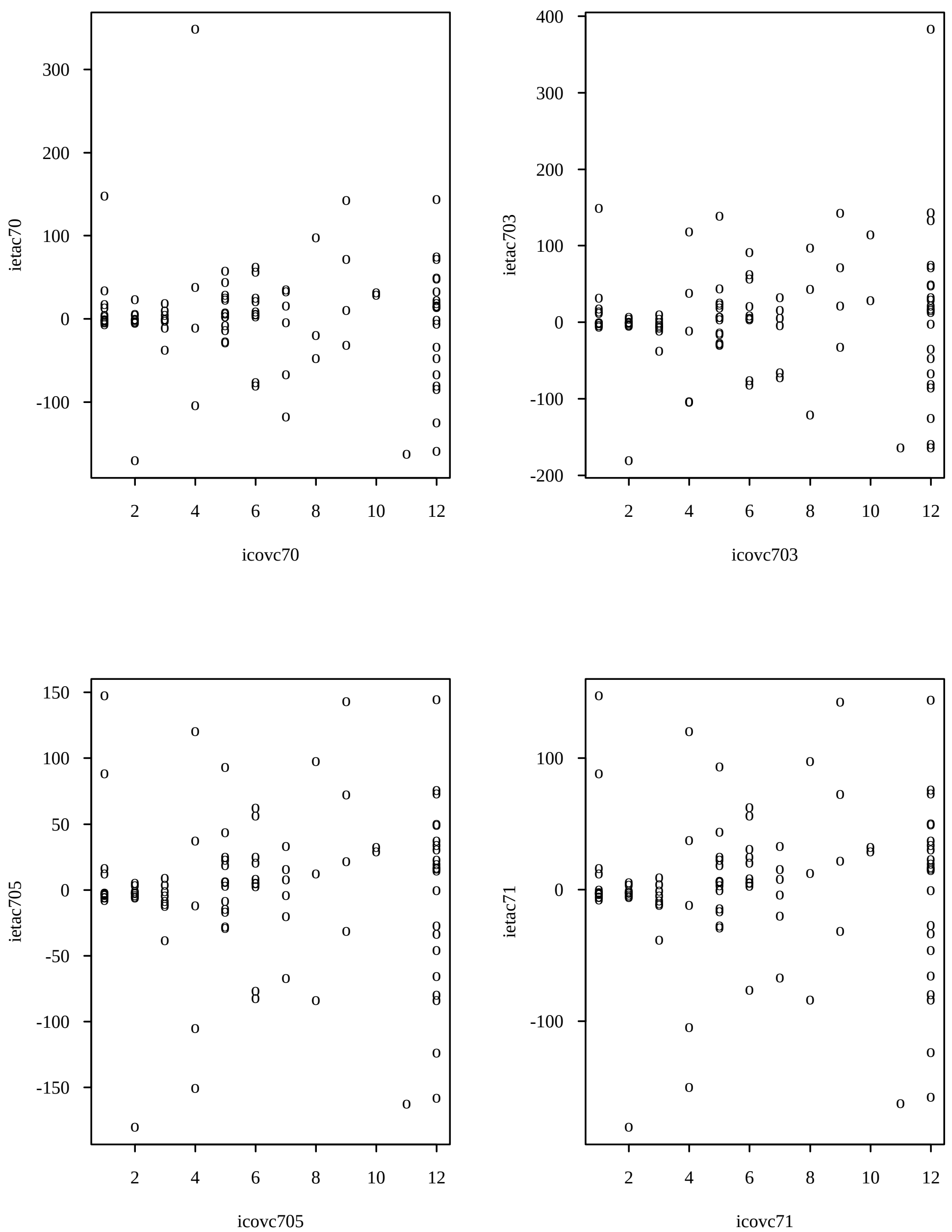
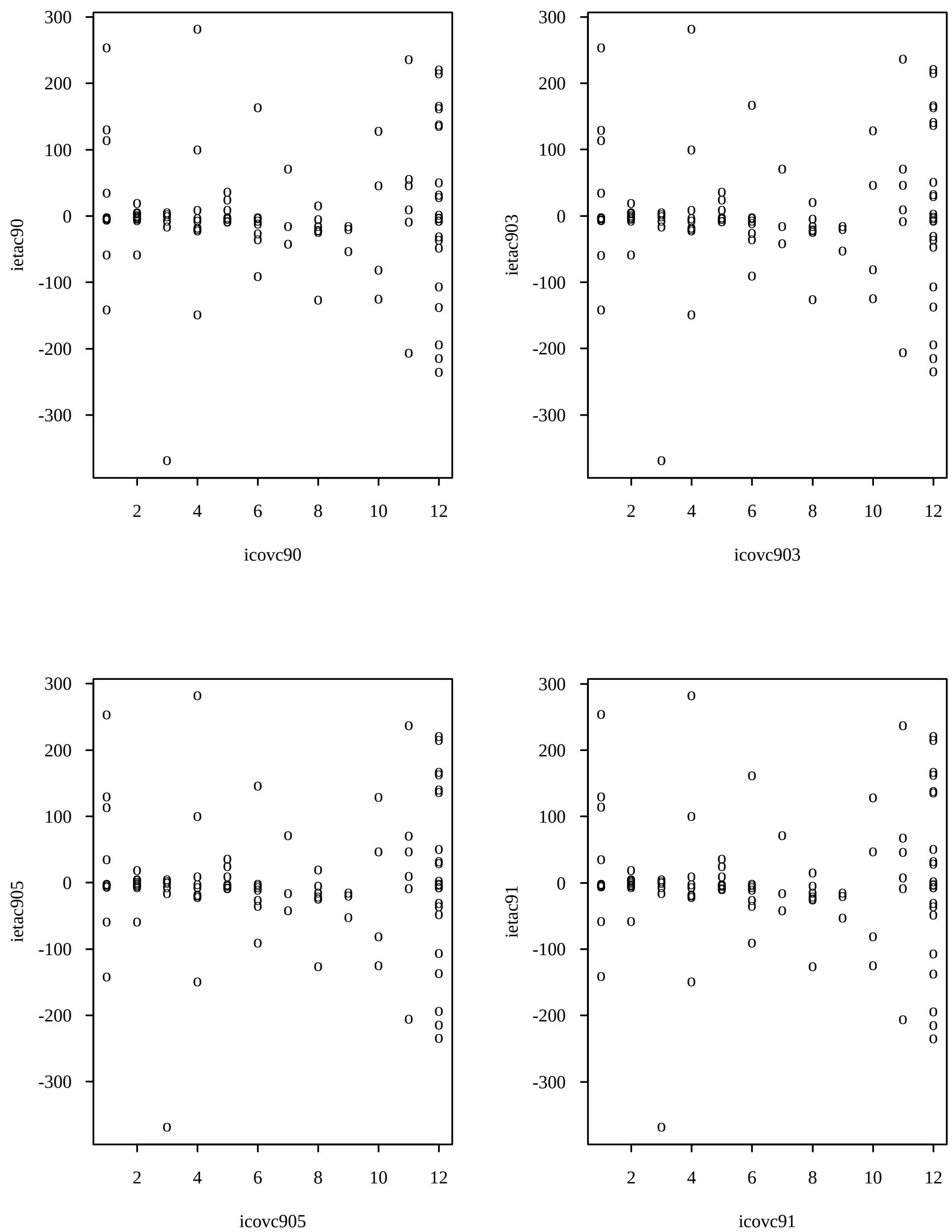


Figure 15: Covariate depending model, censoring level 0.9. First graph: original residual interval width, second graph: 0.3 times increased, third graph: 0.5 times increased, fourth graph: double of the original interval width.



Chapter 4

Data application

To illustrate the proposed new residual theory, it is applied to data of the randomized clinical trial ACTG358. This trial was designed to compare six different antiretroviral treatment regimens for HIV-infected persons who had previously failed combination therapy involving the protease inhibitor Indinavir. For details of the study see Gulick et al. (2000).

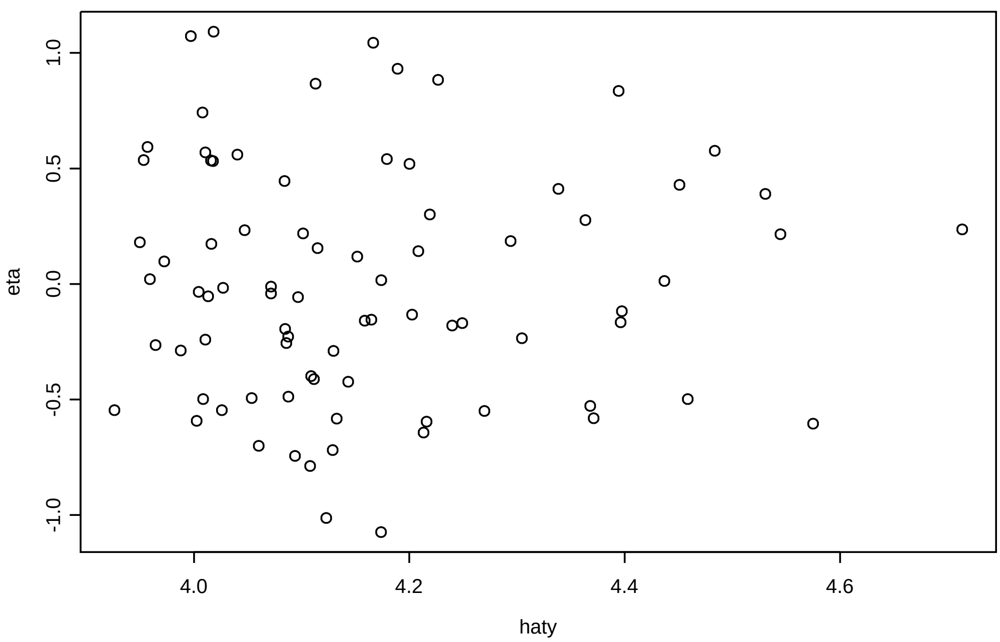
The covariate Z is taken to be the patient's time between Indinavir failure and enrollment. It is of interest examining whether there is an association between Z and age X with the log10 viral load level Y at the time of enrollment. The covariate Z was of interest because delays in initiating ACTG359 led to concerns that patients who had failed Indinavir several months before might behave differently from those who had just recently failed.

The analysis includes 81 patients whose viral load dropped below 500 copies during their prior treatment with Indinavir. Because the viral load was monitored only periodically, the exact time at which a patient's viral load fell below or climbed above 500 could not be observed directly. Thus, the covariate Z, the time between Indinavir failure and enrollment, is censored into the interval of the elapsed time between the first viral load record above 500 copies and randomization, and the elapsed time between the last viral load below 500 copies and randomization.

Fitting the model $Y = \alpha + \beta Z + \gamma X + \varepsilon$ to the data yields estimates for the regression parameters (estimated standard errors) of $\hat{\alpha} = 4.0877$ (0.1596), $\hat{\beta} = -0.0028$ (0.0031), $\hat{\gamma} = 0.0071$ (0.0031), and $\hat{\varepsilon}^2 = 0.2732$ (0.0455). Thus, the positive coefficient for Z suggests that patients with longer delays between Indinavir failure and study entry tend to have higher baseline viral load levels (p=0.02). Similar results were obtained when X was not included in the model. Age was not significantly associated with the baseline viral load (p=0.37), see Gómez et al. (2002).

Gómez et al. (2002) evaluate the goodness of the fitted model with their proposed residuals $\hat{e}_{lup}$. As seen in the simulation results of Chapter 3, these residuals are not generally able to detect violations of the underlaying model assumptions. Hence, we repeat the residual analysis of the data from ACTG359 and re-check the fitted model with a residual plot applying the newly proposed $\hat{e}^*$. The resulting plot of the $\hat{e}^*$ against the fitted response values $\hat{y}_i$ is given in Figure 4.1.

Figure 4.1: Residual plot for the fitted model of the ACTG358 data



It can be seen that the residuals scatter randomly in the plane. They show no special patterns indicating possible model violations like missing regressor variables or non-constant error variance. So, it can be concluded that the fitted model represents the data adequately and support the hypothesis that patients with longer delays between Indinavir failure and study entry tend to have higher baseline viral load levels.

Bibliography

Betensky, R.A.; Finkelstein, D.M. (1999): A nonparametric maximum likelihood estimator for bivariate interval-censored data. *Statistics in Medicine, 18*, 3089-100.

Bickel, P.J.; Fan, J. (1996): Some problems on the estimation of unimodal densities. *Statistica Sinica, 6*, 23-45.

Box, G.E.P.; Müller, M.E. (1958): A note on the generation of random normal deviates. *Annals of Mathematical Statistics, 29*, 610-11.

Buckley, J.; James, L. (1979): Linear regression with censored data. *Biometrika, 66*, 429-36.

Chesher, A.; Irish, M. (1987): Residual analysis in the grouped and censored normal linear model. *Journal of Econometrics, 34*, 33-61.

Cook, R.D.; Weisberg, S. (1982): Residuals and influence in regression. *Chapman and Hall, New York, London.*

Draper, N.R.; Smith, H. (1981): Applied regression analysis. *Wiley and Sons Inc., New York, London, Sydney*

Efron, B. (1967): The two-sample problem with censored data. *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability, 4*, 831-53.

Feller, W. (1966): An introduction to probability theory and its application, Volume I. *Wiley and Sons Inc., New York, London, Sydney*

Finkelstein, D.; Wolfe, R. (1985): A semiparametric model for regression analysis of interval-censored failure time data. *Biometrics, 41*, 933-45.

Gentleman, R.; Geier, C.J. (1994): Maximum likelihood for interval censored data: Consistency and computation. *Biometrika, 81*, 618-23.

Gil, M.A.; López-García, M.T.; Lubiano M.A.; Montenegro, M. (2001): Regression and correlation analysis of a linear relation between random intervals.

Glejser, H. (1969): A new test for heteroscedasticity. *Journal of the American Statistical Association, 64*, 316-23.

Goldstein, L.; Messer, K. (1992): Optimal plug-in estimators for nonparametric functional estimation. *Journal of the American Statistical Association, 60*, 539-47.

Goldfeld, S.M.; Quandt, R.E. (1965): Some tests for homoscedasticity. *Journal of the American Statistical Association, 60*, 539-47.

Gómez, G.; Espinal, A.; Lagakos, S.W. (2001a): Inference for a linear regression model with an interval-censored covariate. *Technical Report of the Politecnical University of Catalonia, Barcelona, Spain.*

Gómez, G.; Espinal, A.; Lagakos, S.W. (2002): Inference for a linear regression model with an interval-censored covariate. *Statistics in Medicine, accepted for publication.*

Gómez, G.; Calle, M.L.; Oller, R. (2001b): A walk through interval-censored survival data. *Technical Report of the Politecnical University of Catalonia, Barcelona, Spain.*

Gray, R.J.; Pierce, D.A. (1985): Goodness-of-fit tests for censored survival data. *Annals of Statistics, 13*, 552-63.

Groeneboom, P.; Wellner, J.A. (1992): Information bounds and nonparametric maximum likelihood estimation. *Birkhäuser Verlag, Basel, Boston, Berlin*

Gulick, R.M. et al. (2000): Randomized study of Saquinavir in combination with Ritonavir or Nelfinavir together with Delavirdine, Adefovir Dipivoxil or both in HIV-infected subjects with virologic failure on Indinavir. *Journal of Infectious Diseases 182*, 1375-1384.

Harrison, M.J.; McCabe, B.P.M. (1979): A test for heteroscedasticity based on ordinary least squares residuals. *Journal of the American Statistical Association, 74*, 494-99.

Hartung, J.; Elpelt, B.; Klösener, K.-H. (1993): Statistik. *R. Oldenbourg Verlag, München, Berlin.*

Hillis, S.L. (1995): Residual Plots for the censored data linear regression model. *Statistics in Medicine, 14*, 2023-36.

Huang, C.J.; Bolch, B.W. (1974): On the testing of regression distrurbances for normality. *Journal of the American Statistical Association, 69*, 330-35.

Jarque, C.M.; Bera, A.K. (1987): A test for normality of observations and regression residuals. *International Statistical Review, 55*, 163-72.

Li, L.; Pu, Z. (1999): Regression models with arbitrarily interval-censored observations. *Communications in Statistics, 28(7)*, 1547-63.

Li, L.; Watkins, T.; Yu, Q. (1997): An EM algorithm for smoothing the self-consistent estimator of survival functions with interval-censored data. *Scandinavian Journal of Statistics, 24*, 531-42.

Montgomery, D.C.; Peck, E.A. (1982): Introduction to linear regression *John Wiley and Sons, New York, Chichester, Brisbane, Toronto, Singapore.*

Morgan, B.J.T. (1984): Elements of simulation. *Chapman and Hall, London, New York.*

Pan W. (200): Smooth estimation of the survival function for interval censored data. *Statistics in Medicine, 19*, 2611-24.

Pan, W.; Chappell, R. (1999): A note on inconcistency of NPMLE of the distribution function from left truncated and case I interval censored data. *Lifetime Data Analysis, 5(3)*, 281-91.

Peto, R. (1973): Experimental survival curves for interval-censored data. *Journal of the Royal Statistical Society, Series C 22*, 86-91.

Rawlings, J.O. (1988): Applied regression analysis. *Wadsworth&Brooks, Pacific Grove, California.*

Rutemiller, H.C.; Bowers, D.A. (1968): Estimation in a heteroscedastic regression model. *Journal of the American Statistical Association, 63*, 552-57.

Spiegelhalter, D.J. (1983): Diagnostic tests of distributional shape. *Biometrika, 70*, 401-09.

Sun, J. (2001): Variance estimation of a survival function for interval-censored survival data. *Statistics in Medicine, 20*, 1249-57.

Theil, H. (1971): Principles of Econometrics. *North-Holland Publishing Co., Amsterdam.*

Turnbull, B. (1976): The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society, Series B 38*, 290-95.

Wald, A. (1949): Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics, 20*, 595-601.

White, H. (1980): A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity. *Econometrica, 48*, 817-38.

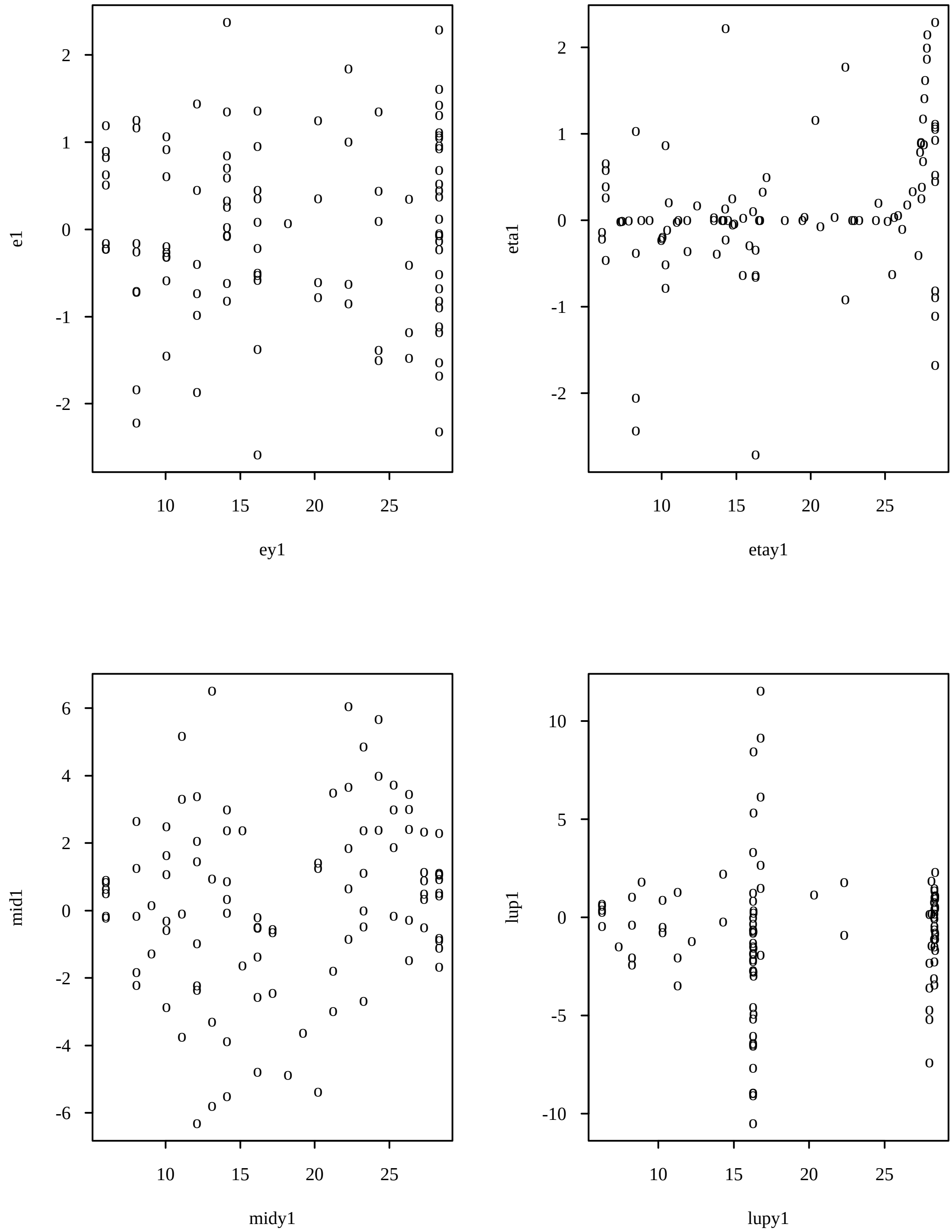
Yu, Q.; Li, L.; Wong, G.Y.C. (2000): On consistency of the self-consistent estimator of survival functions with interval-censored data. *Scandinavian Journal of Statistics, 27*, 35-44.

Yu, Q.; Schick, A.; Li, L.; Wong, G.Y.C. (1998): Asymptotic properties of the GLME with case 2 interval-censored data. *Statistics and Probability Letters, 37*, 223-28.

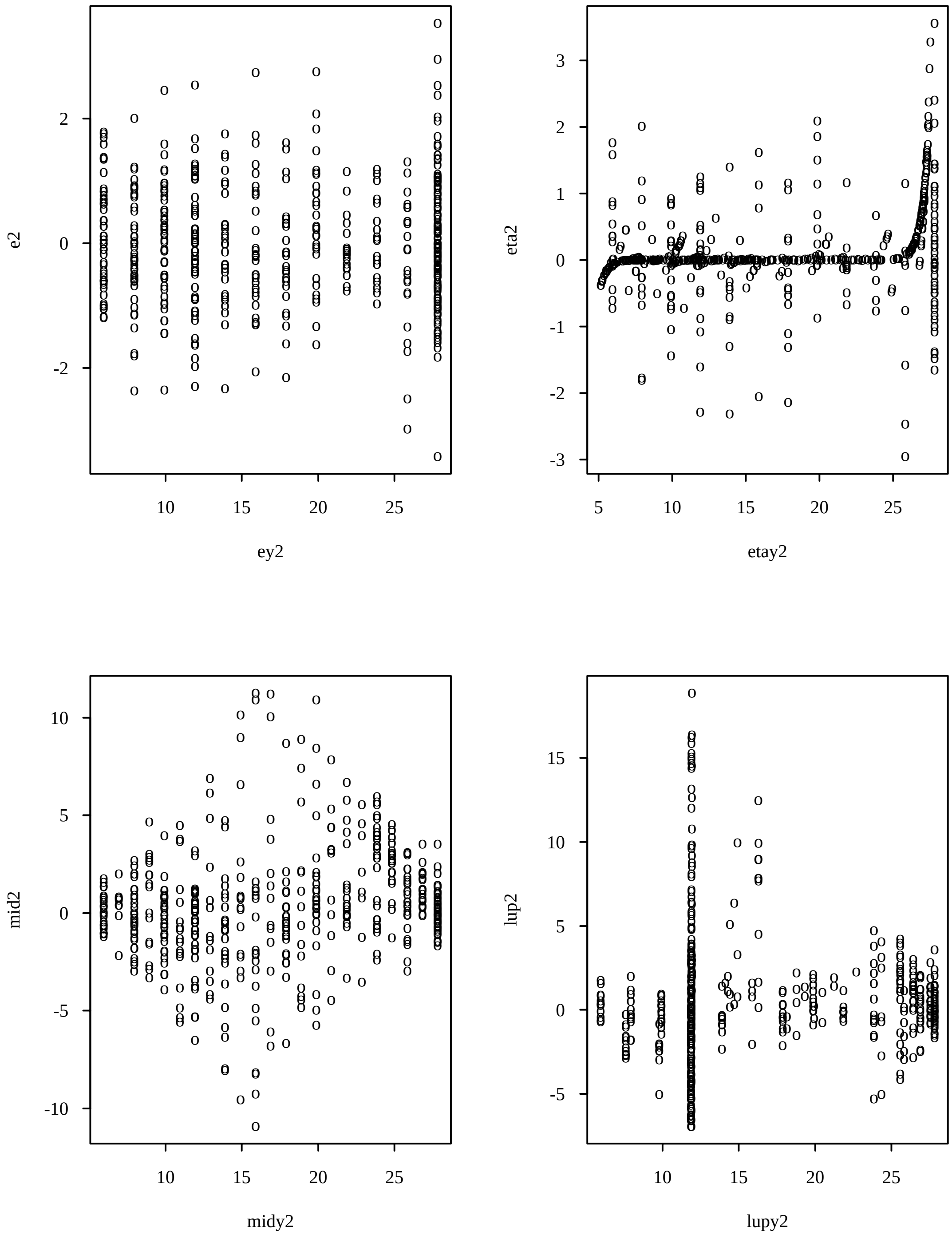
Appendix A

Residual plots when the model is correctly specified

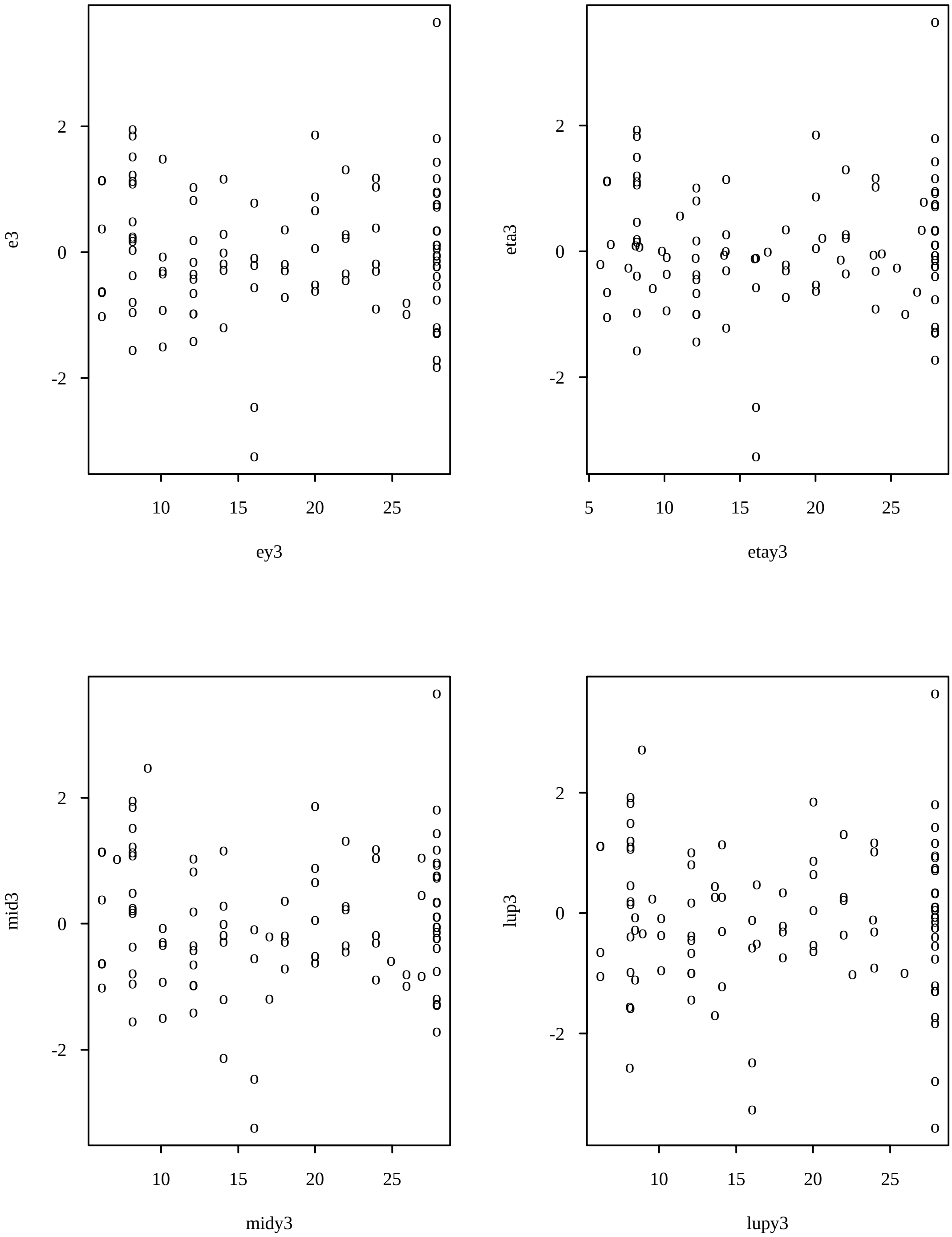
Scenario 1: covariate distribution $\mathrm{Exp}(\frac{1}{8})$, β=2, p=0.3, n=100:



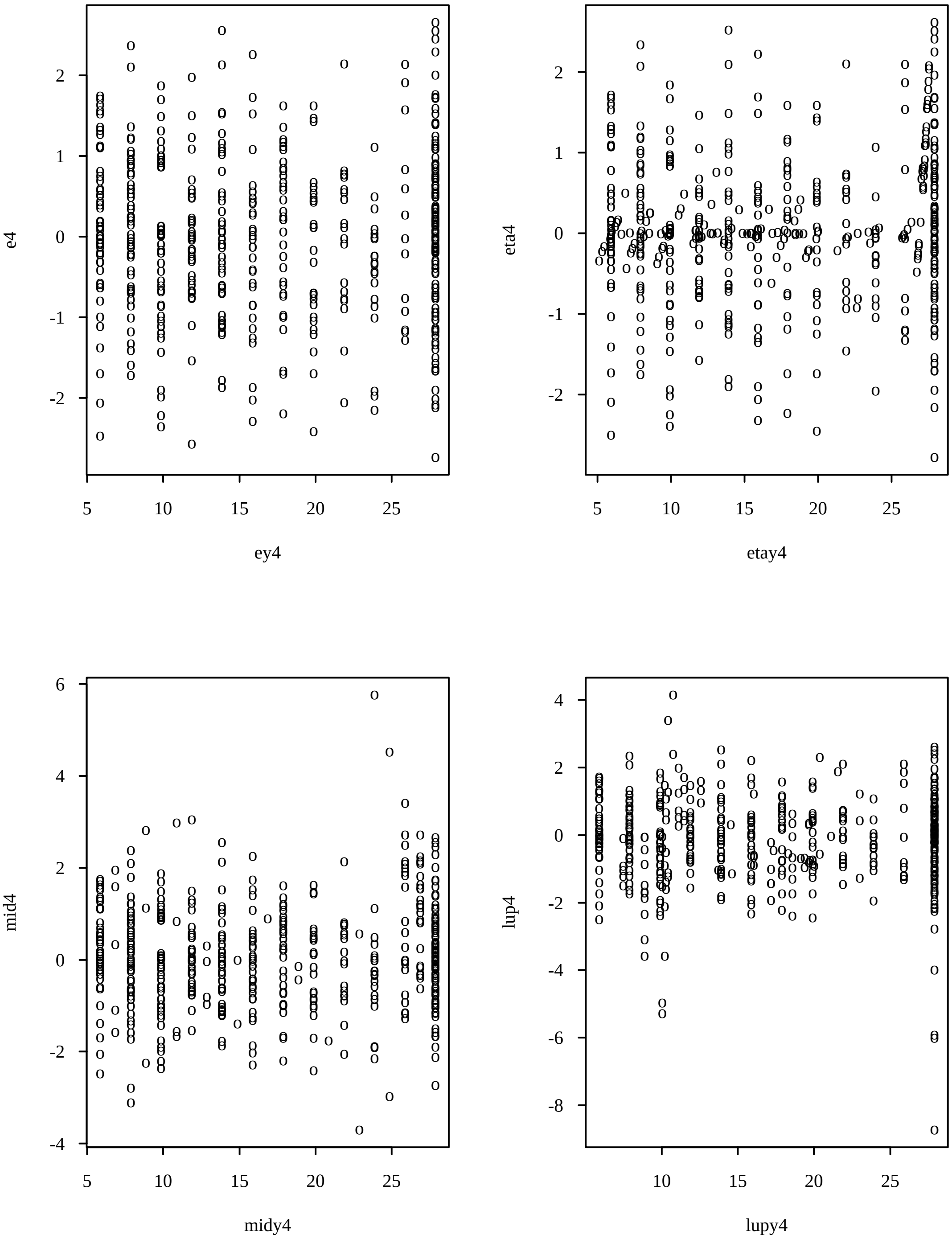
Scenario 2: covariate distribution Exp($\frac{1}{8}$), β=2, p=0.3, n=500:



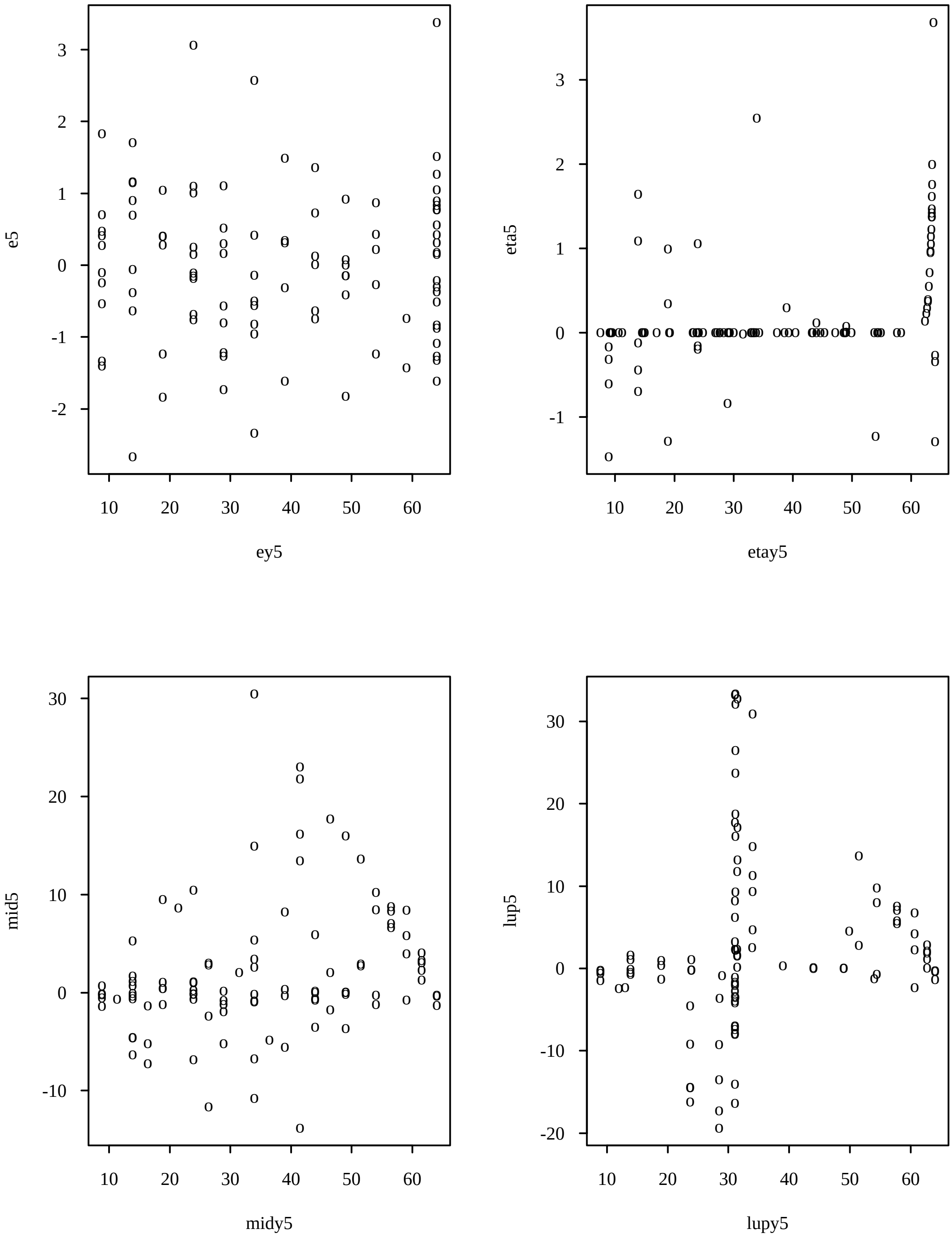
Scenario 3: covariate distribution $\text{Exp}(\frac{1}{8})$, β=2, p=0.7, n=100:



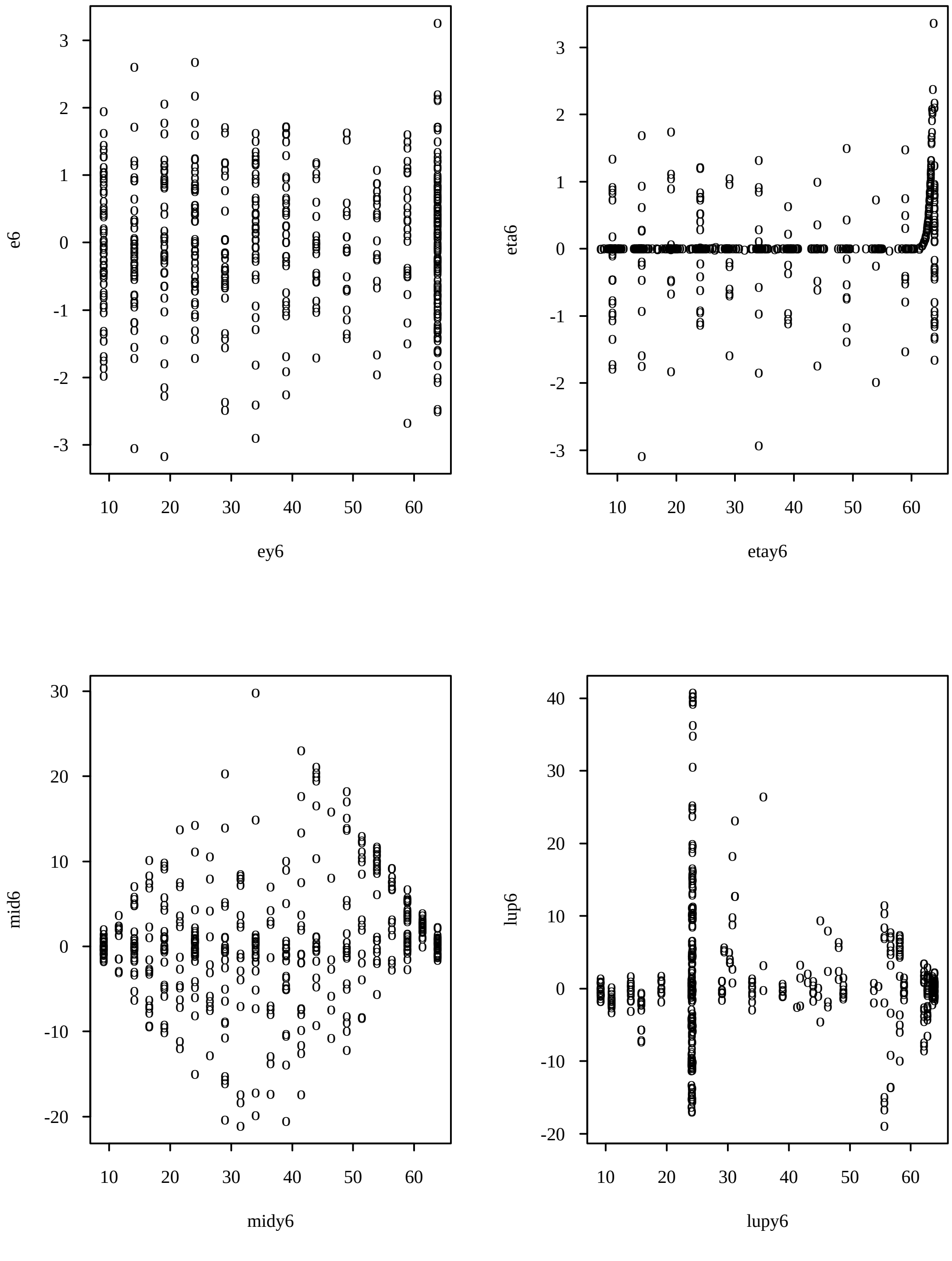
Scenario 4: covariate distribution Exp($\frac{1}{8}$), β=2, p=0.7, n=500:



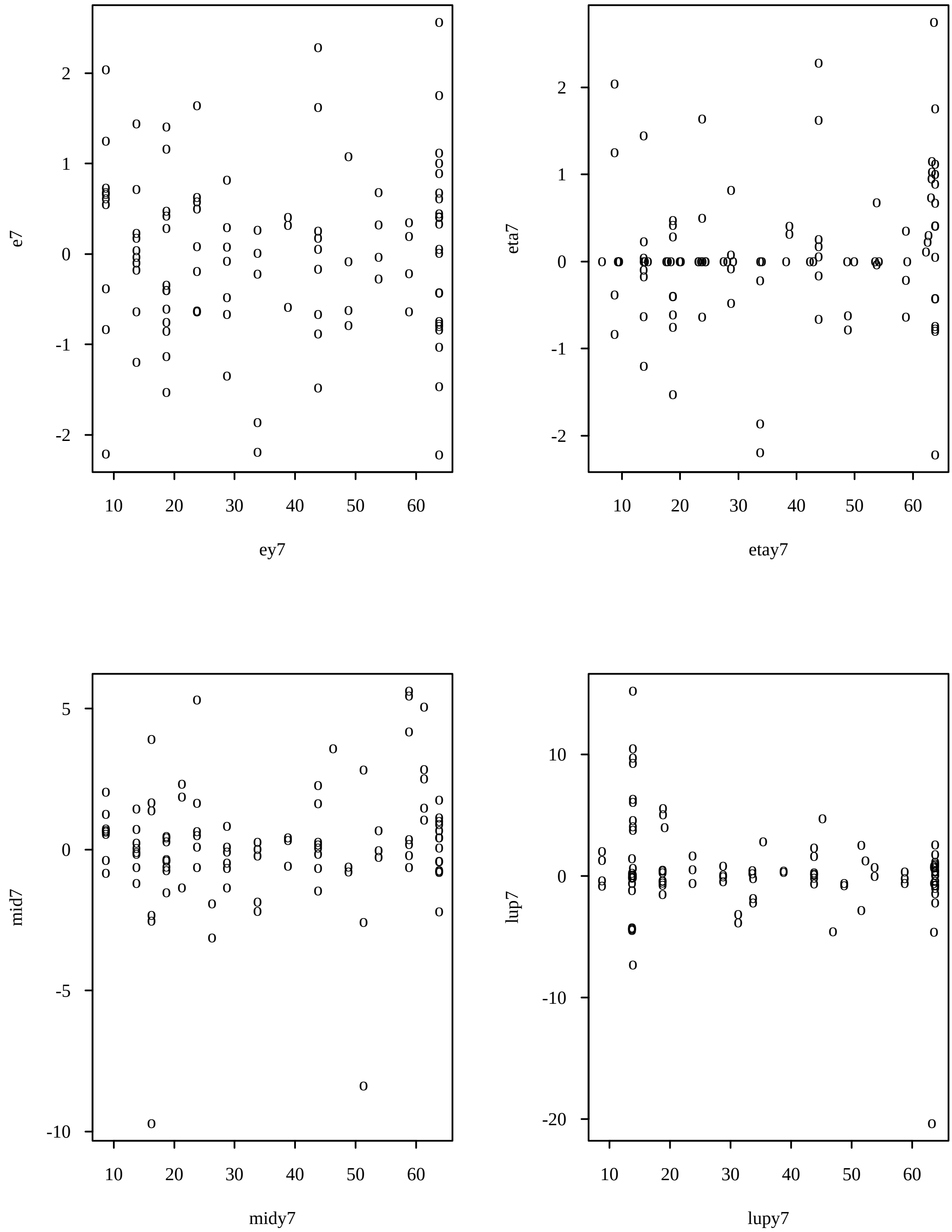
Scenario 5: covariate distribution Exp($\frac{1}{8}$), β=5, p=0.3, n=100:



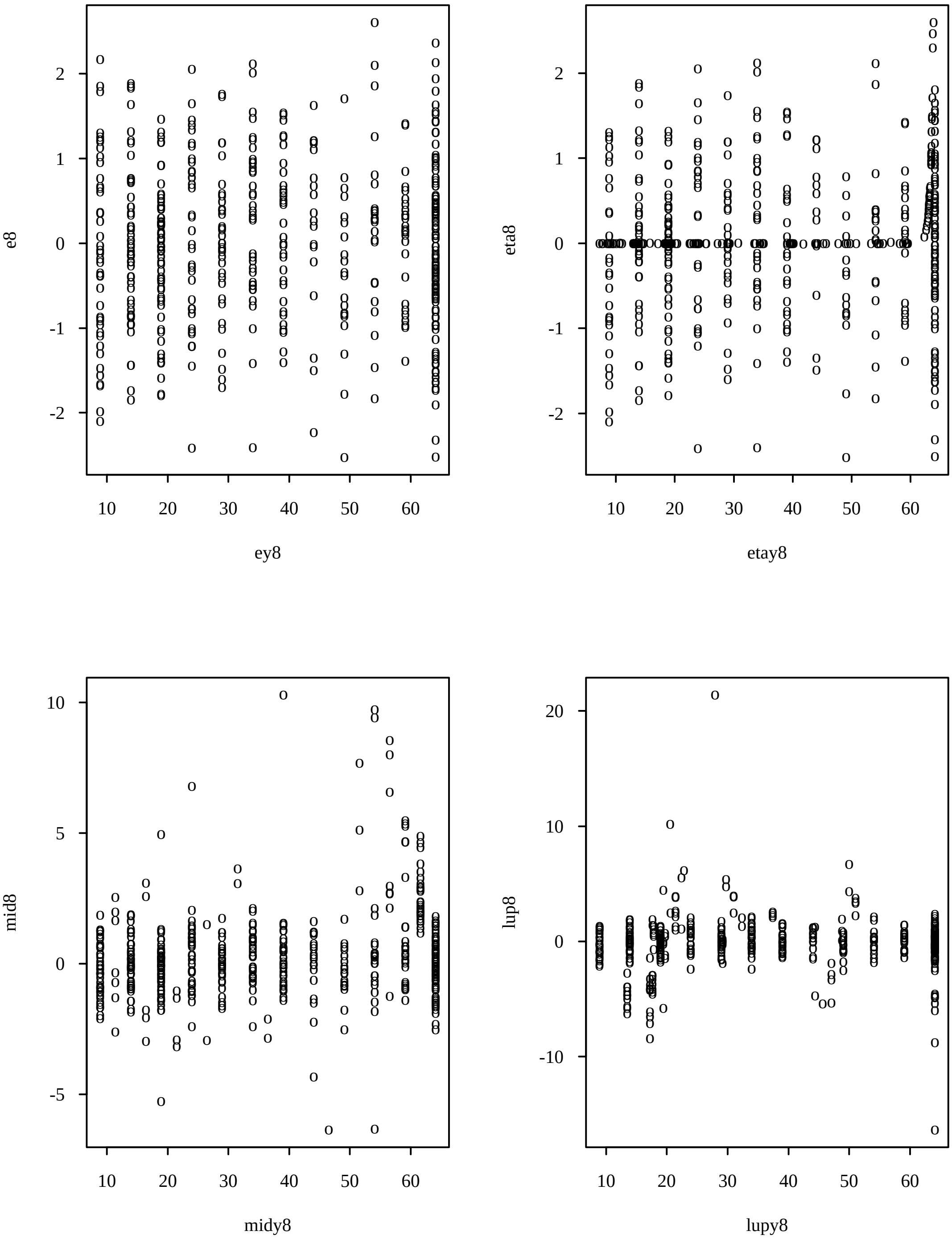
Scenario 6: covariate distribution Exp$(\frac{1}{8})$, β=5, p=0.3, n=500:



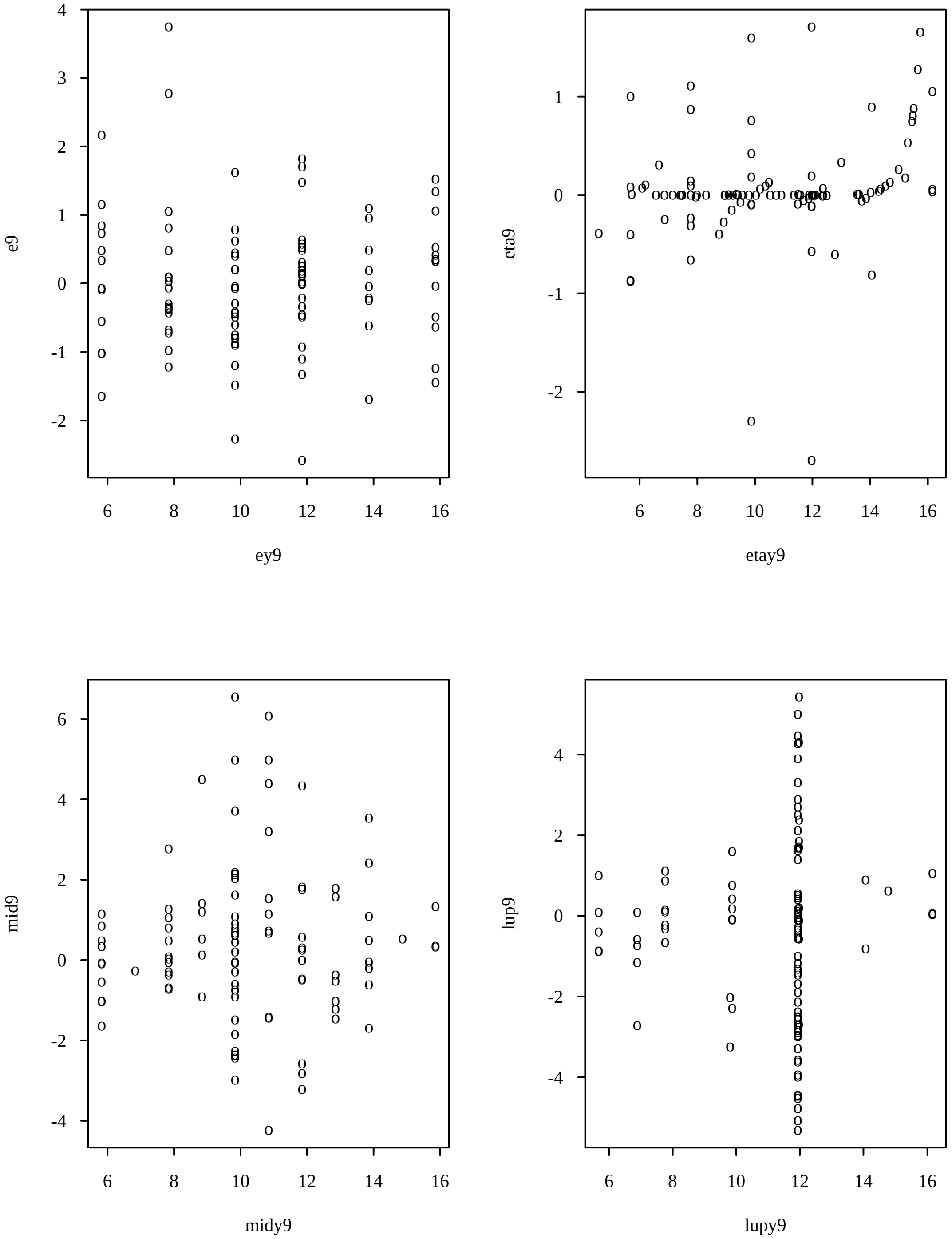
Scenario 7: covariate distribution Exp($\frac{1}{8}$), β=5, p=0.7, n=100:



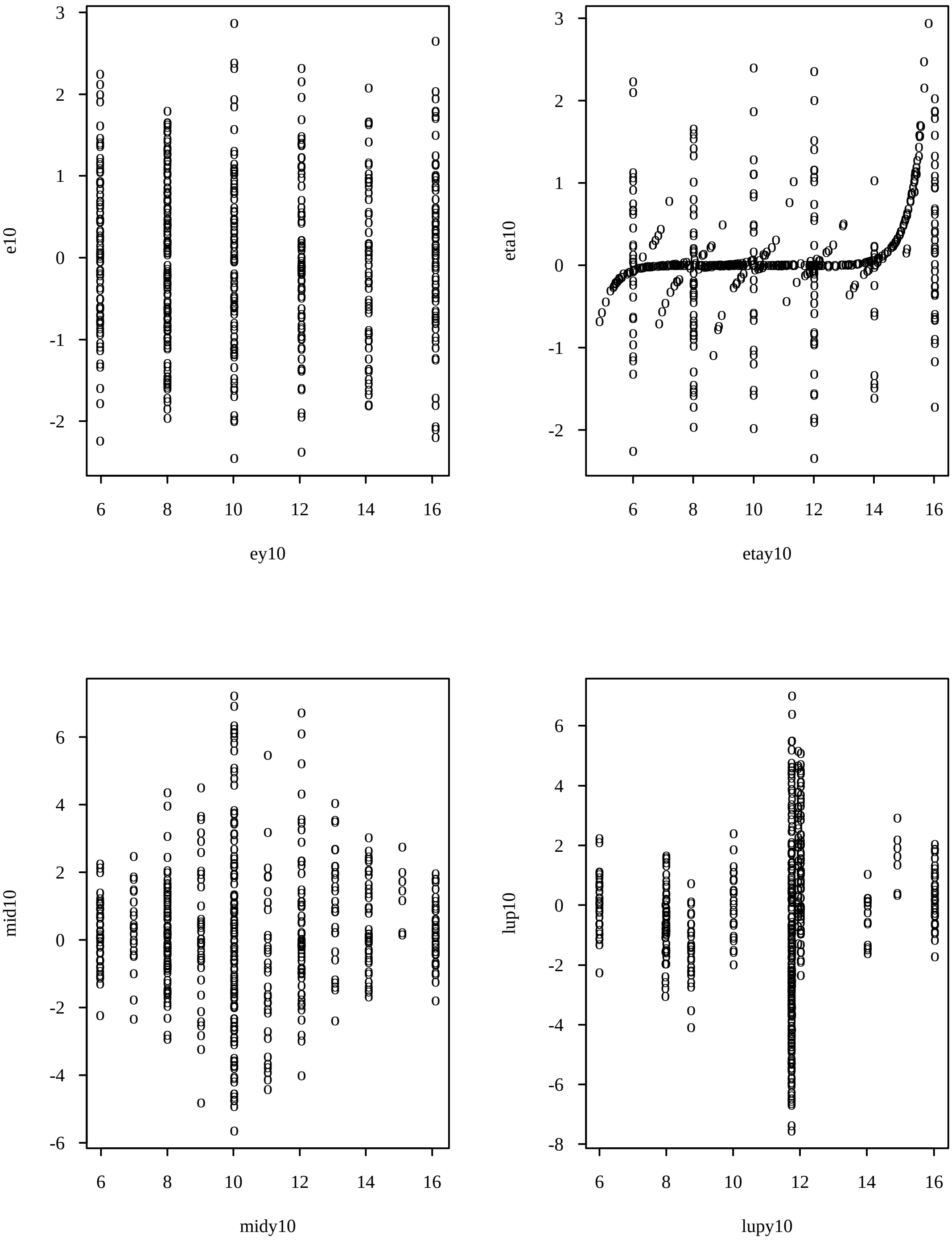
Scenario 8: covariate distribution Exp($\frac{1}{8}$), β=5, p=0.7, n=500:



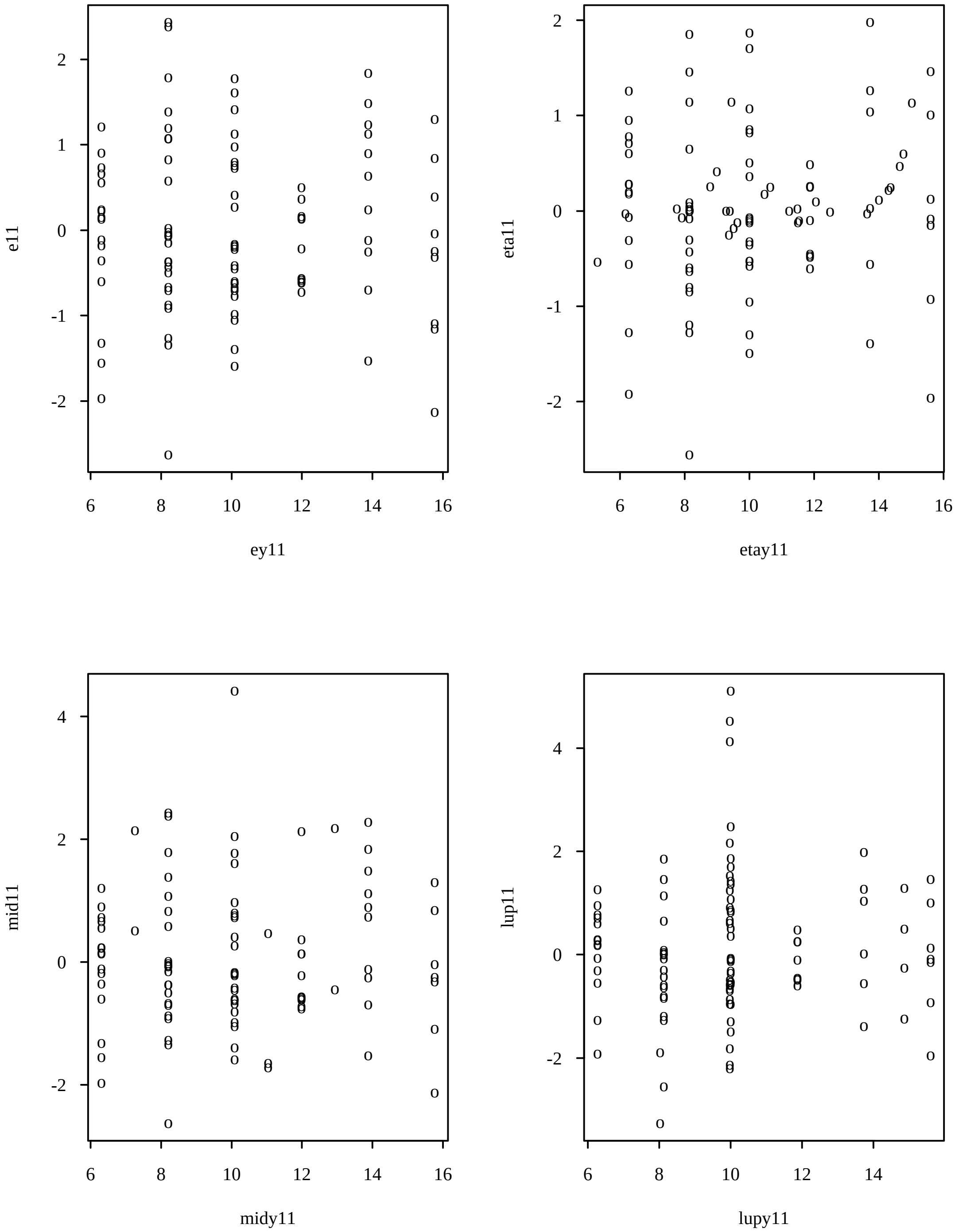
Scenario 9: covariate distribution $W(\frac{1}{6}, \frac{3}{2})$, β=2, p=0.3, n=100:



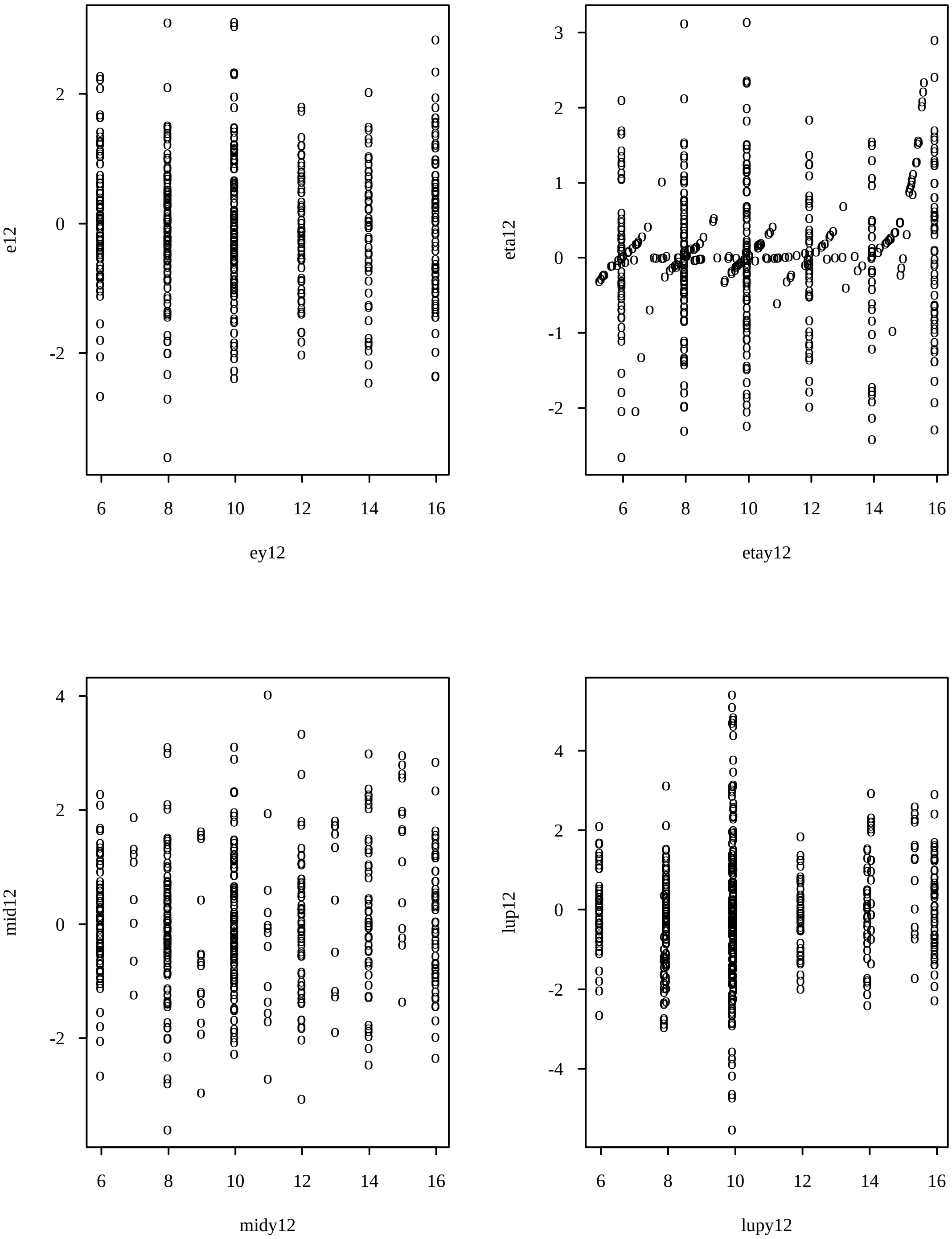
Scenario 10: covariate distribution $W(\frac{1}{6}, \frac{3}{2})$, β=2, p=0.3, n=500:



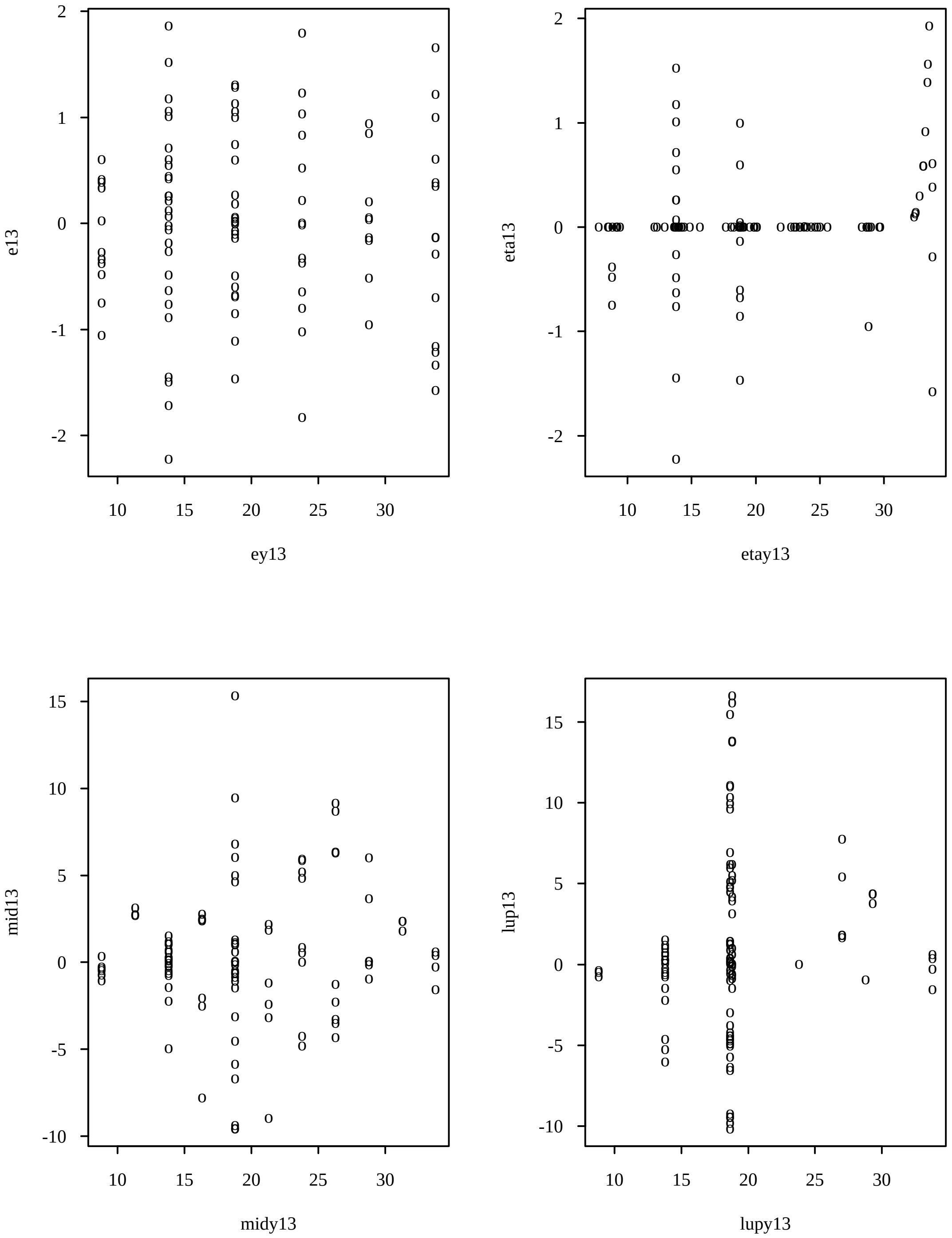
Scenario 11: covariate distribution $W(\frac{1}{6}, \frac{3}{2})$, β=2, p=0.7, n=100:



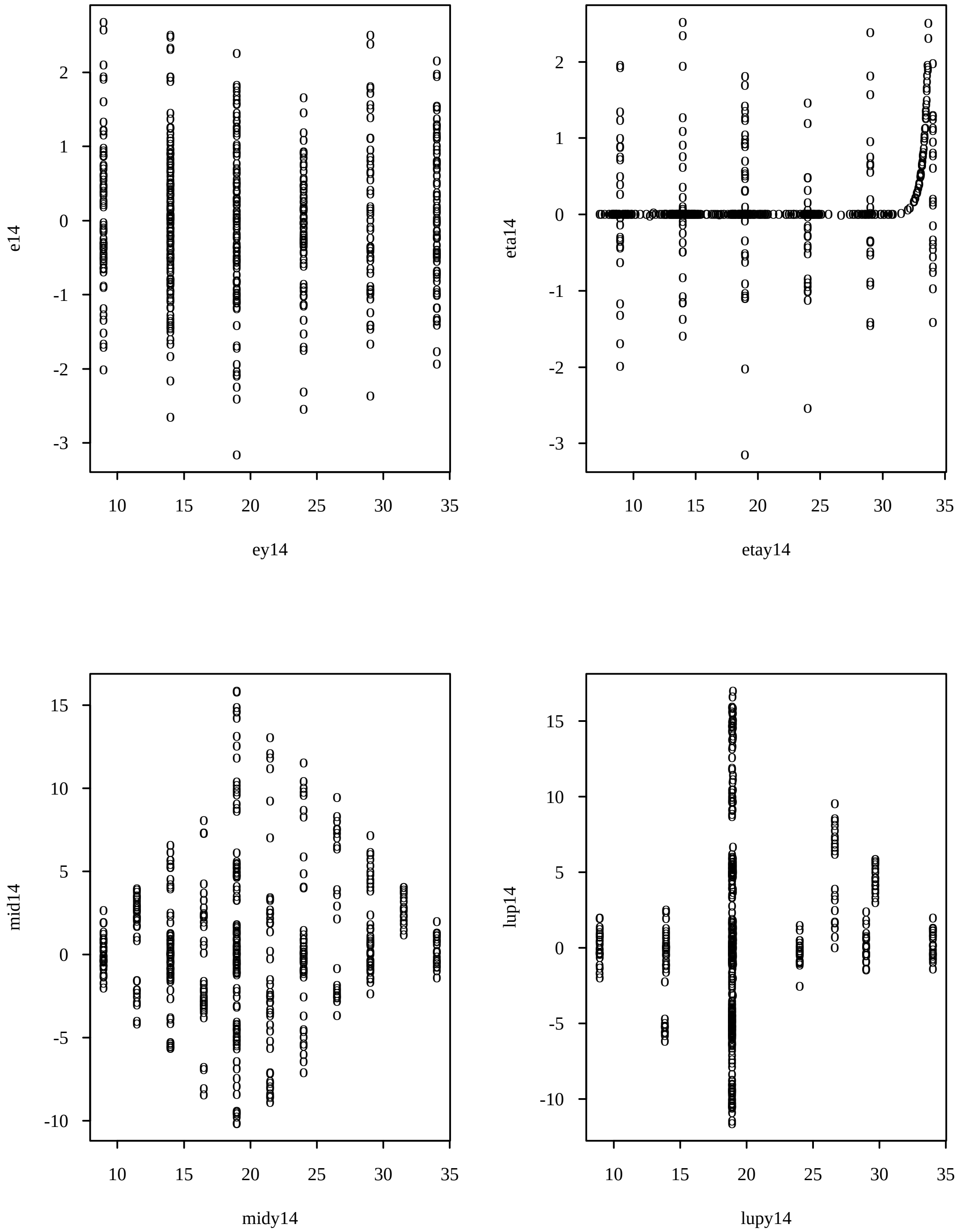
Scenario 12: covariate distribution $W(\frac{1}{6}, \frac{3}{2})$, β=2, p=0.7, n=500:



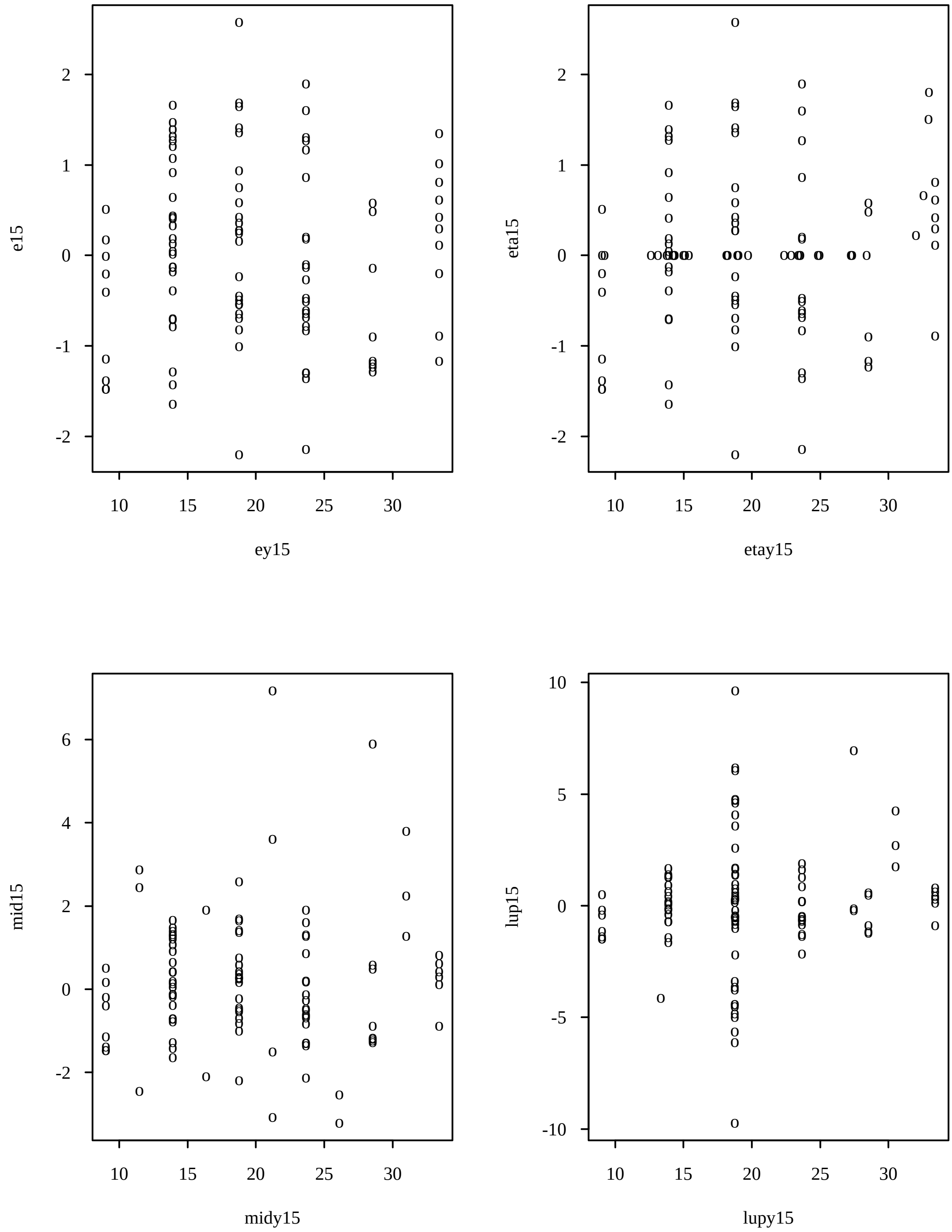
Scenario 13: covariate distribution $W(\frac{1}{6}, \frac{3}{2})$, β=5, p=0.3, n=100:



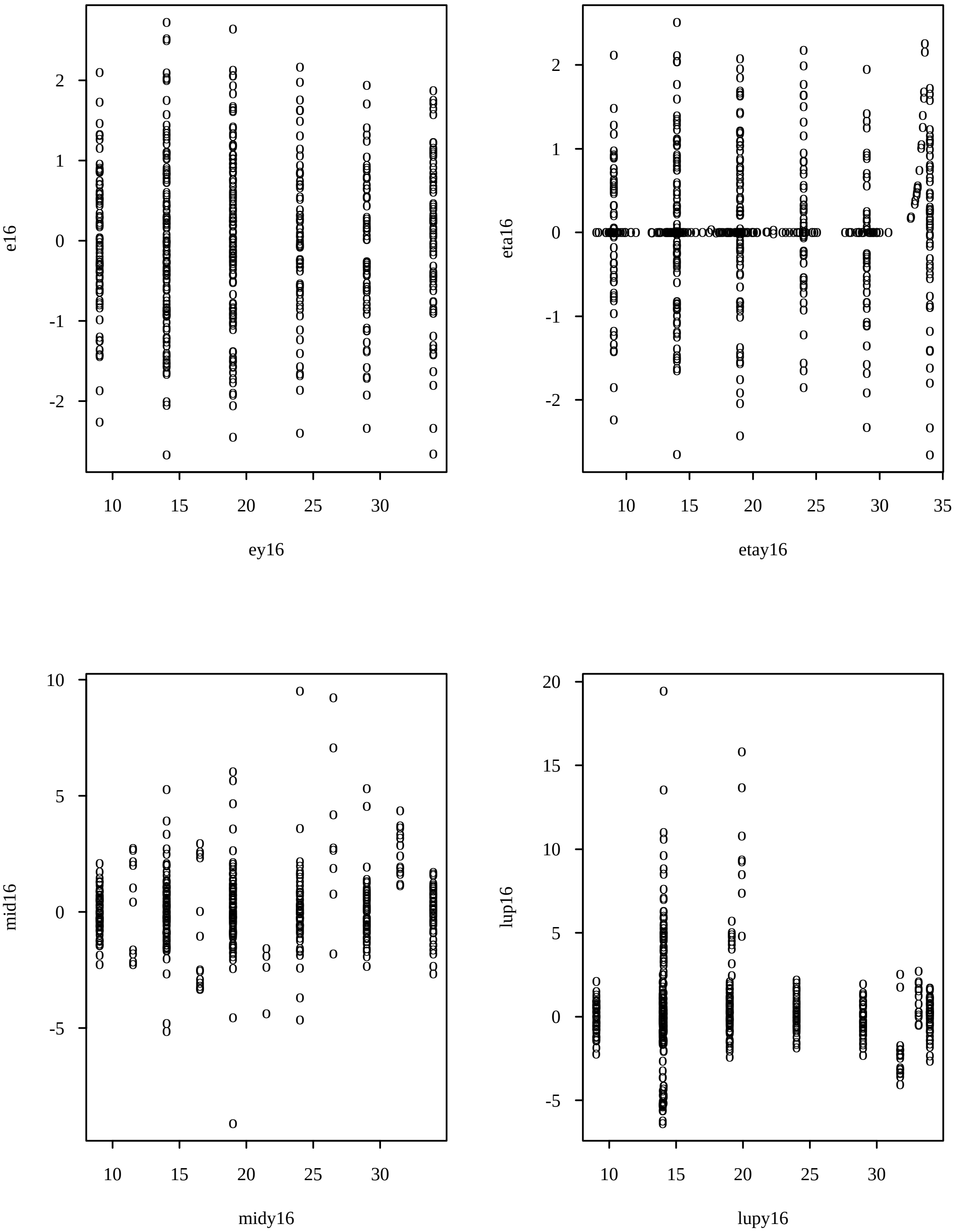
Scenario 14: covariate distribution $W(\frac{1}{6}, \frac{3}{2})$, β=5, p=0.3, n=500:



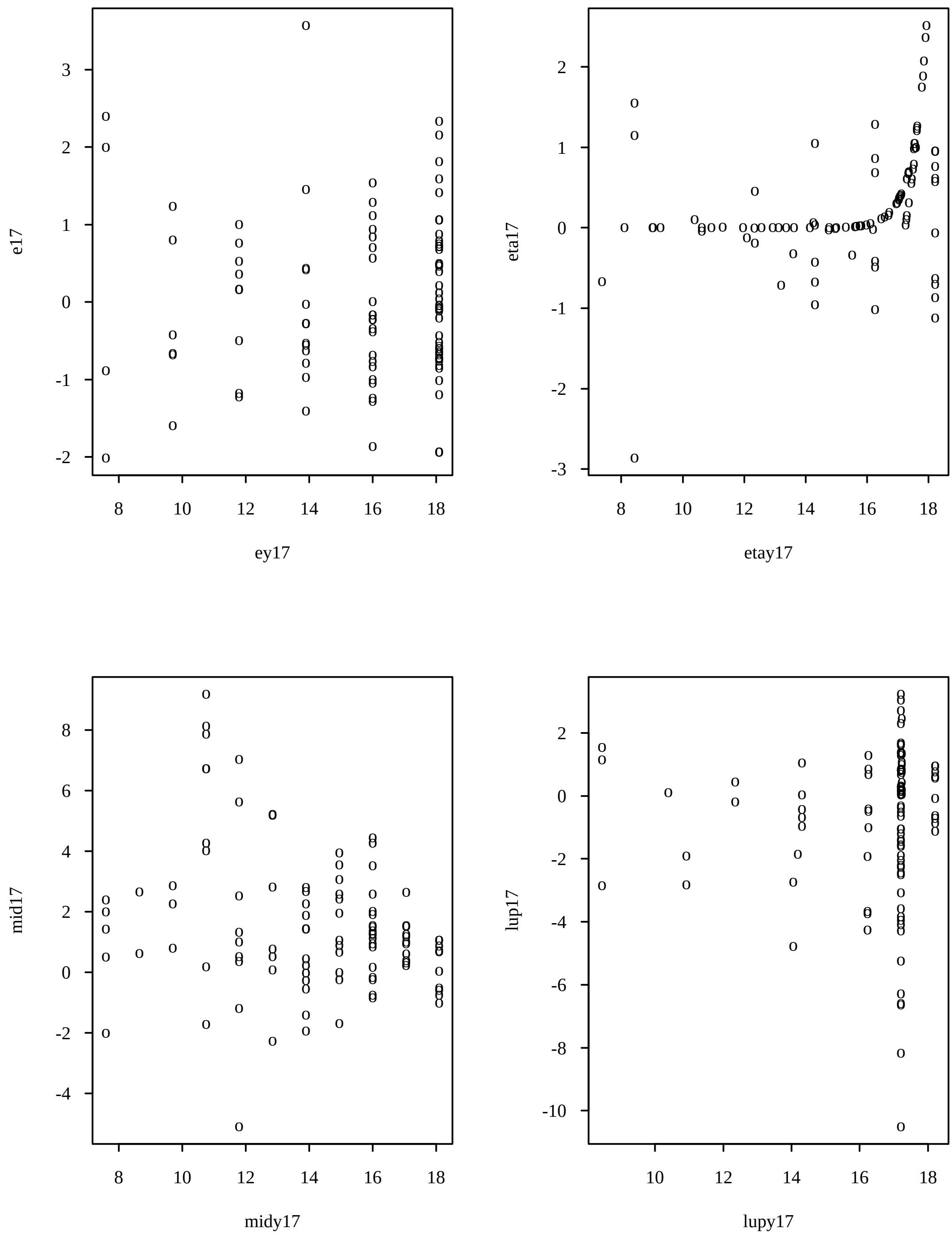
Scenario 15: covariate distribution W($\frac{1}{6}$, $\frac{3}{2}$), β=5, p=0.7, n=100:



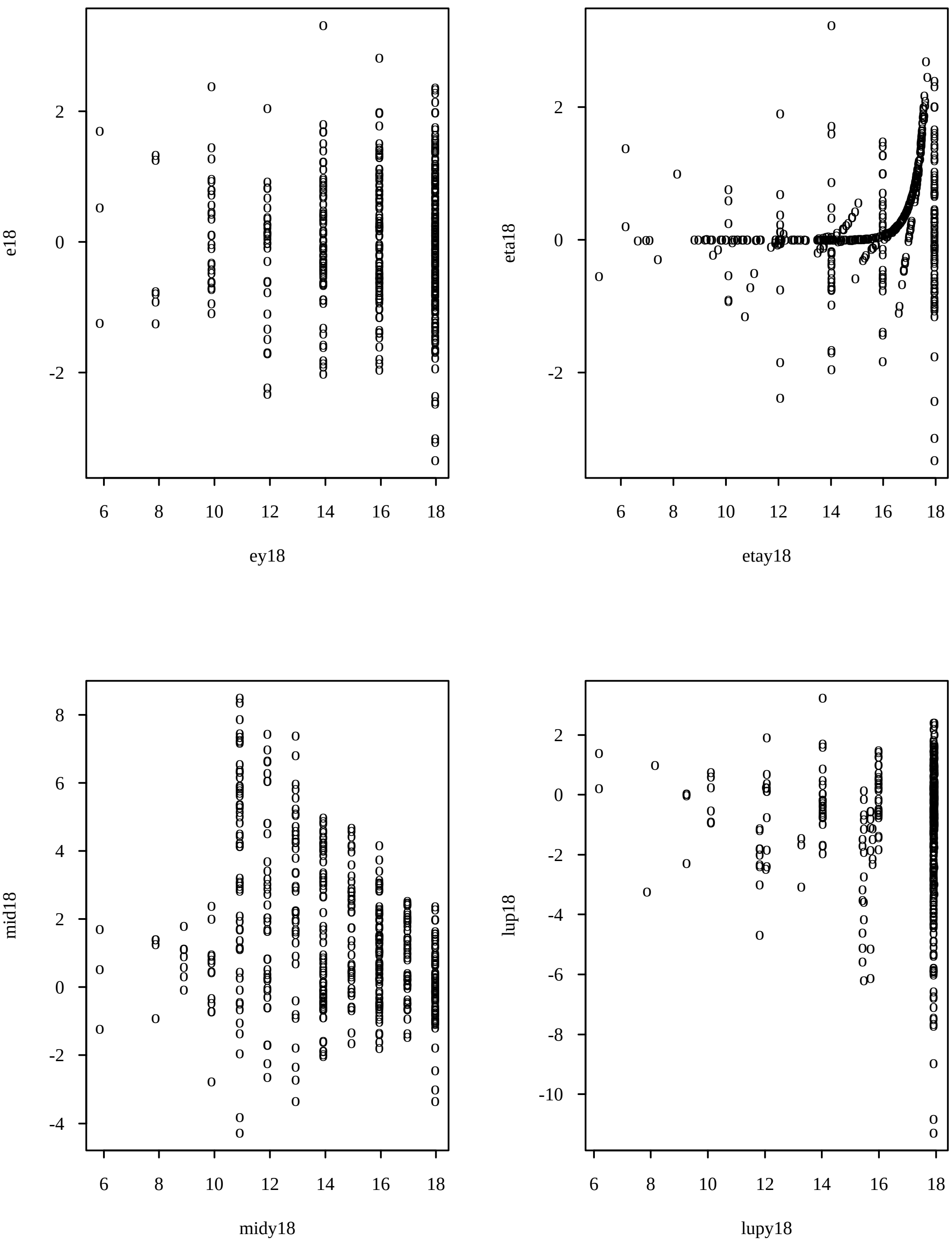
Scenario 16: covariate distribution $W(\frac{1}{6}, \frac{3}{2})$, β=5, p=0.7, n=500:



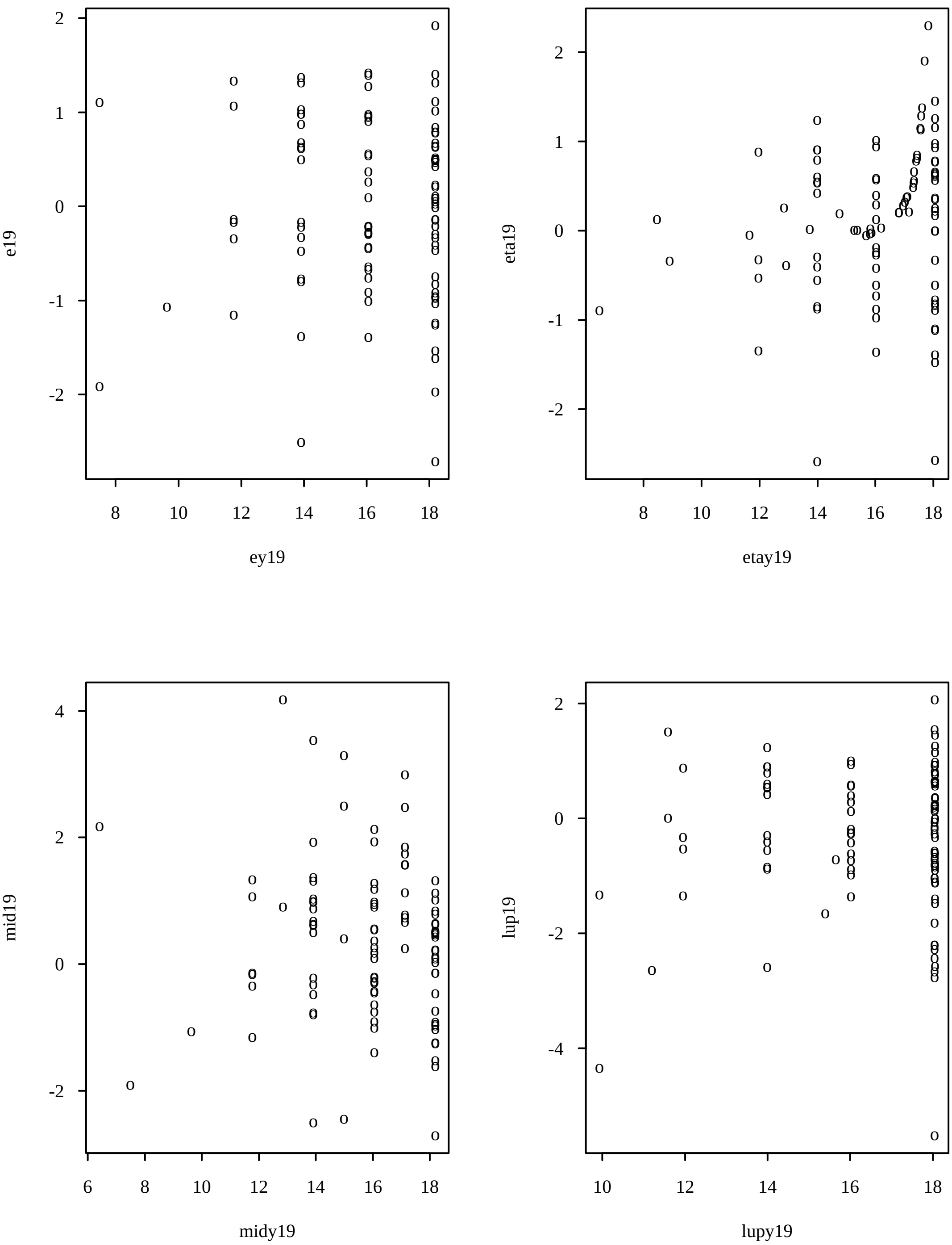
Scenario 17: covariate distribution N(4,4), β=2, p=0.3, n=100:



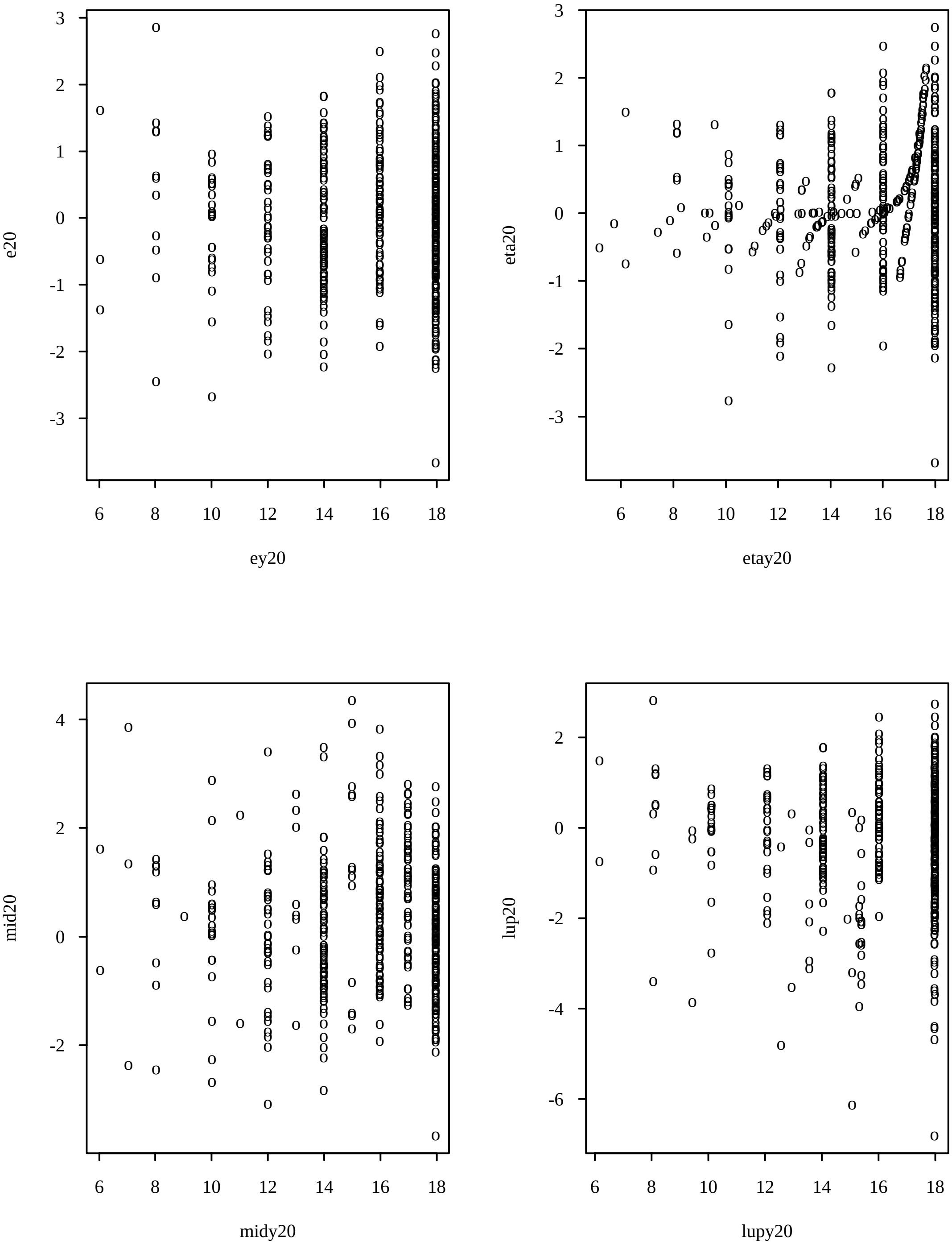
Scenario 18: covariate distribution N(4,4), β=2, p=0.3, n=500:



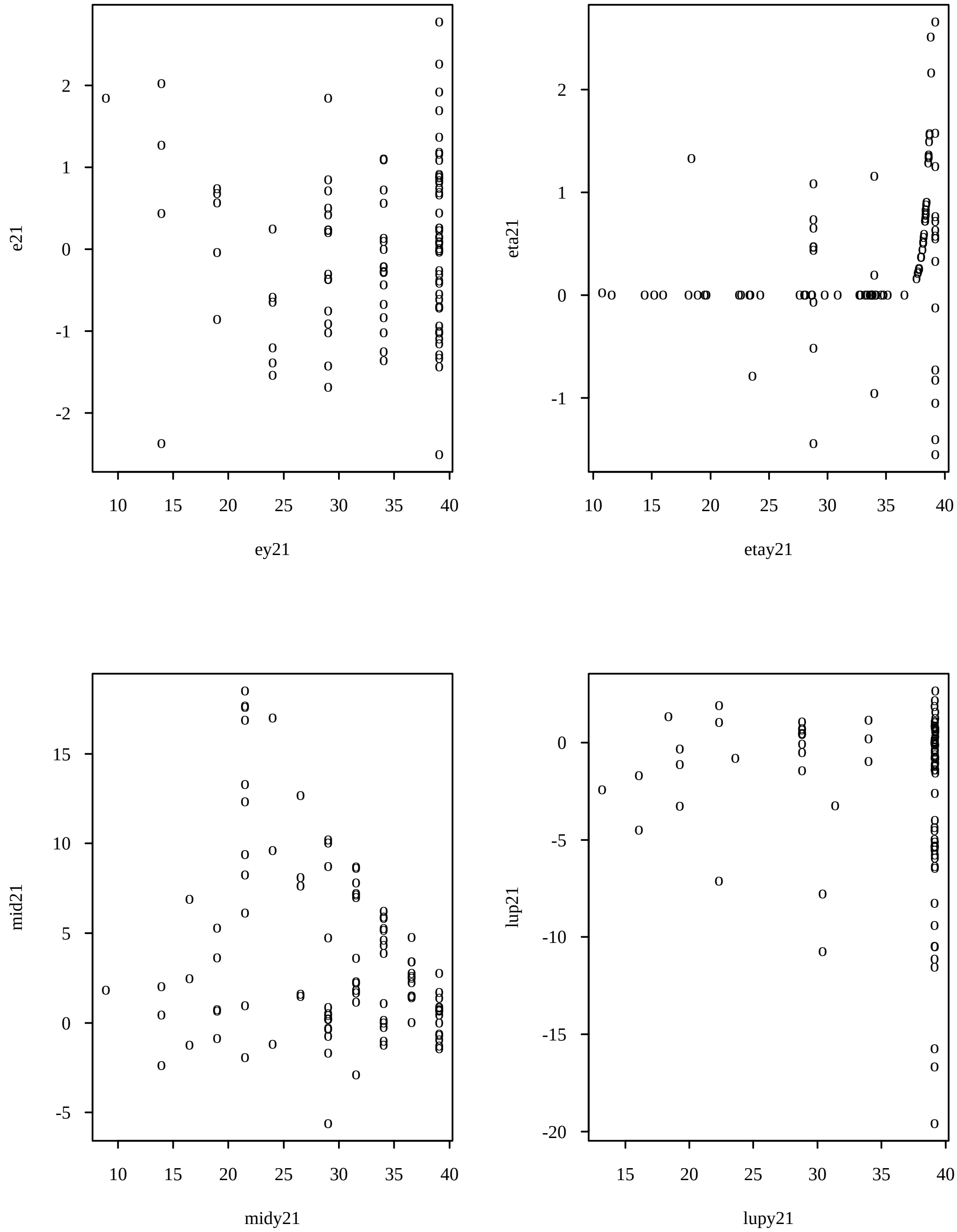
Scenario 19: covariate distribution N(4,4), β=2, p=0.7, n=100:



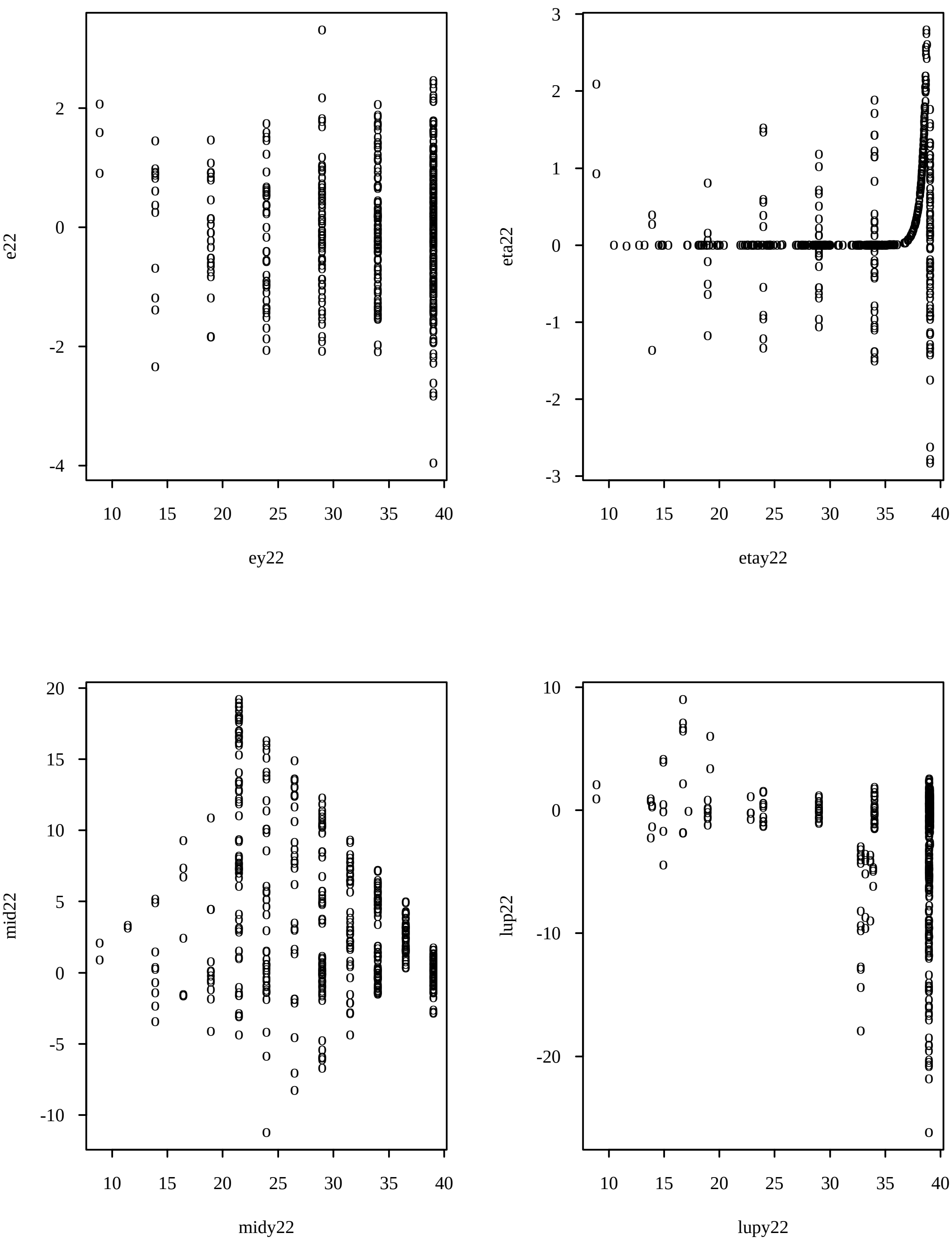
Scenario 20: covariate distribution N(4,4), β=2, p=0.7, n=500:



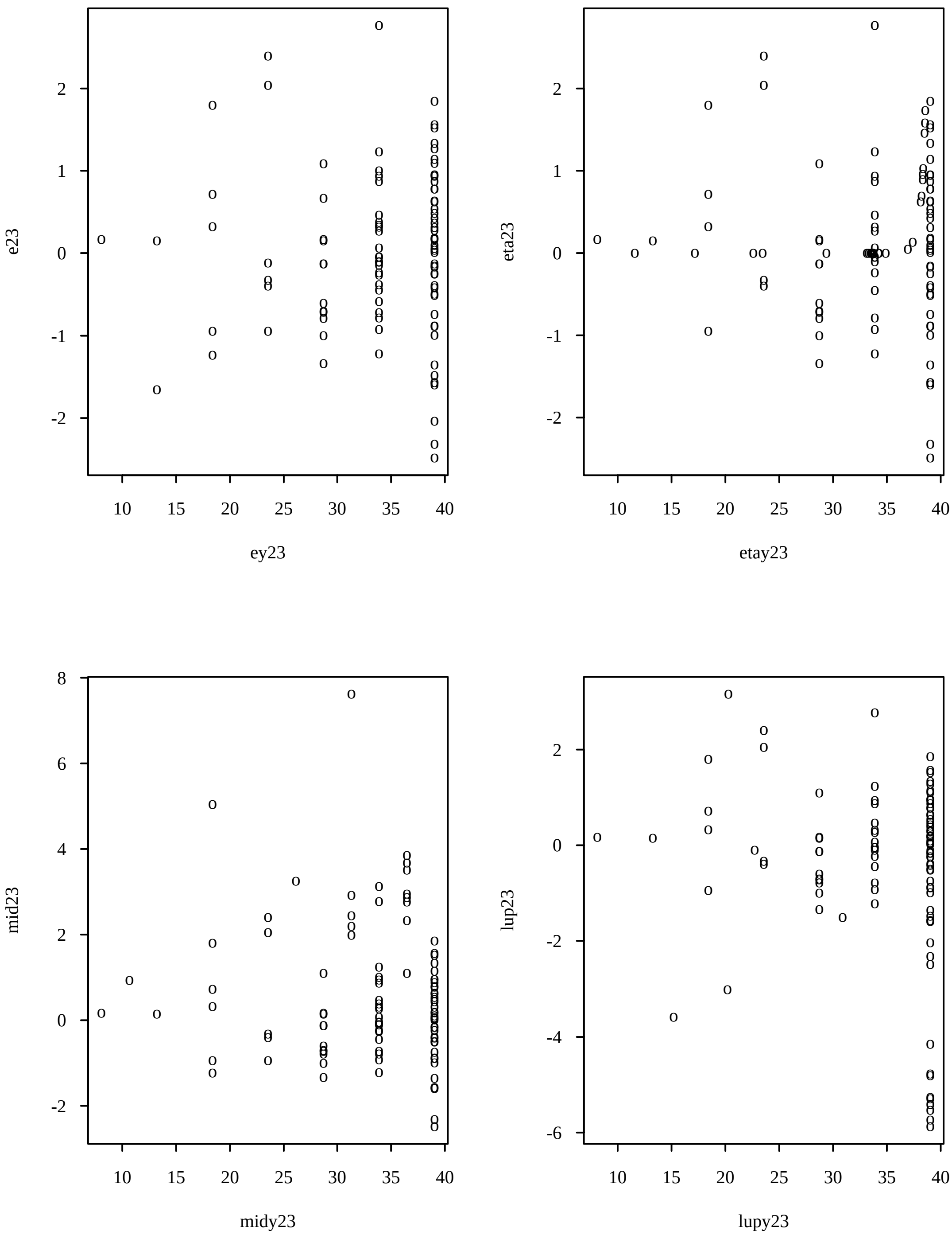
Scenario 21: covariate distribution N(4,4), β=5, p=0.3, n=100:



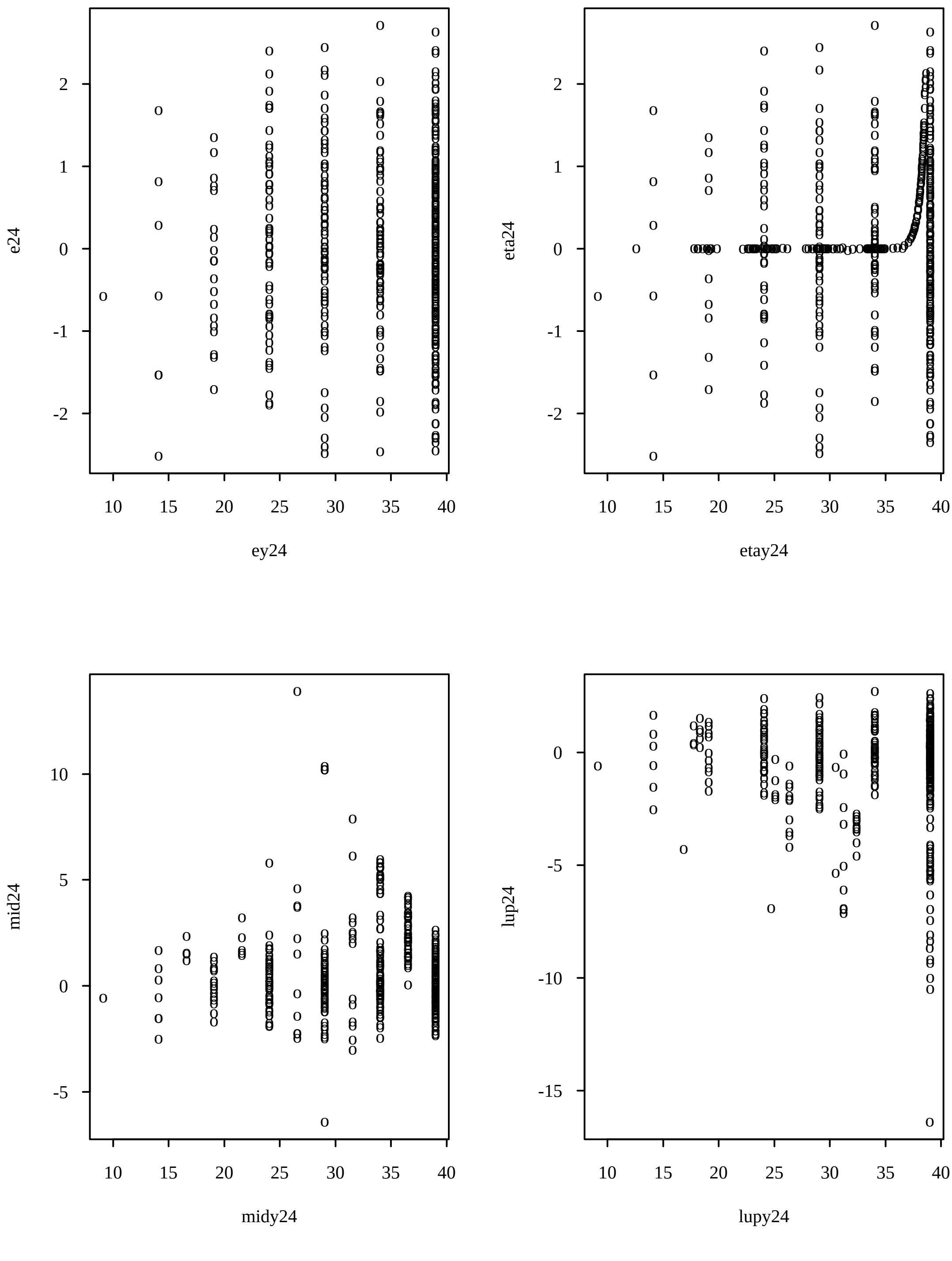
Scenario 22: covariate distribution N(4,4), β=5, p=0.3, n=500:



Scenario 23: covariate distribution N(4,4), β=5, p=0.7, n=100:



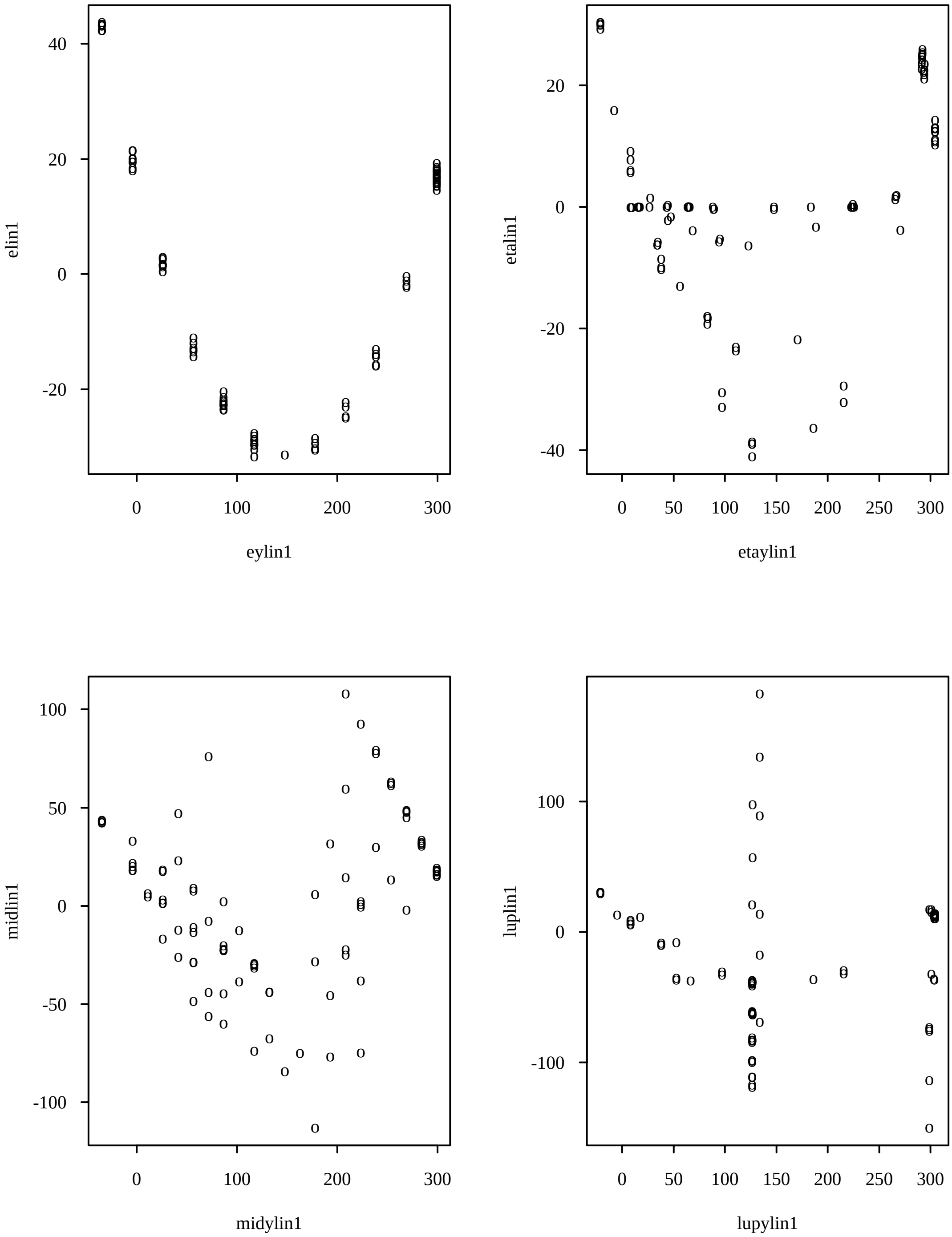
Scenario 24: covariate distribution N(4,4), β=5, p=0.7, n=500:



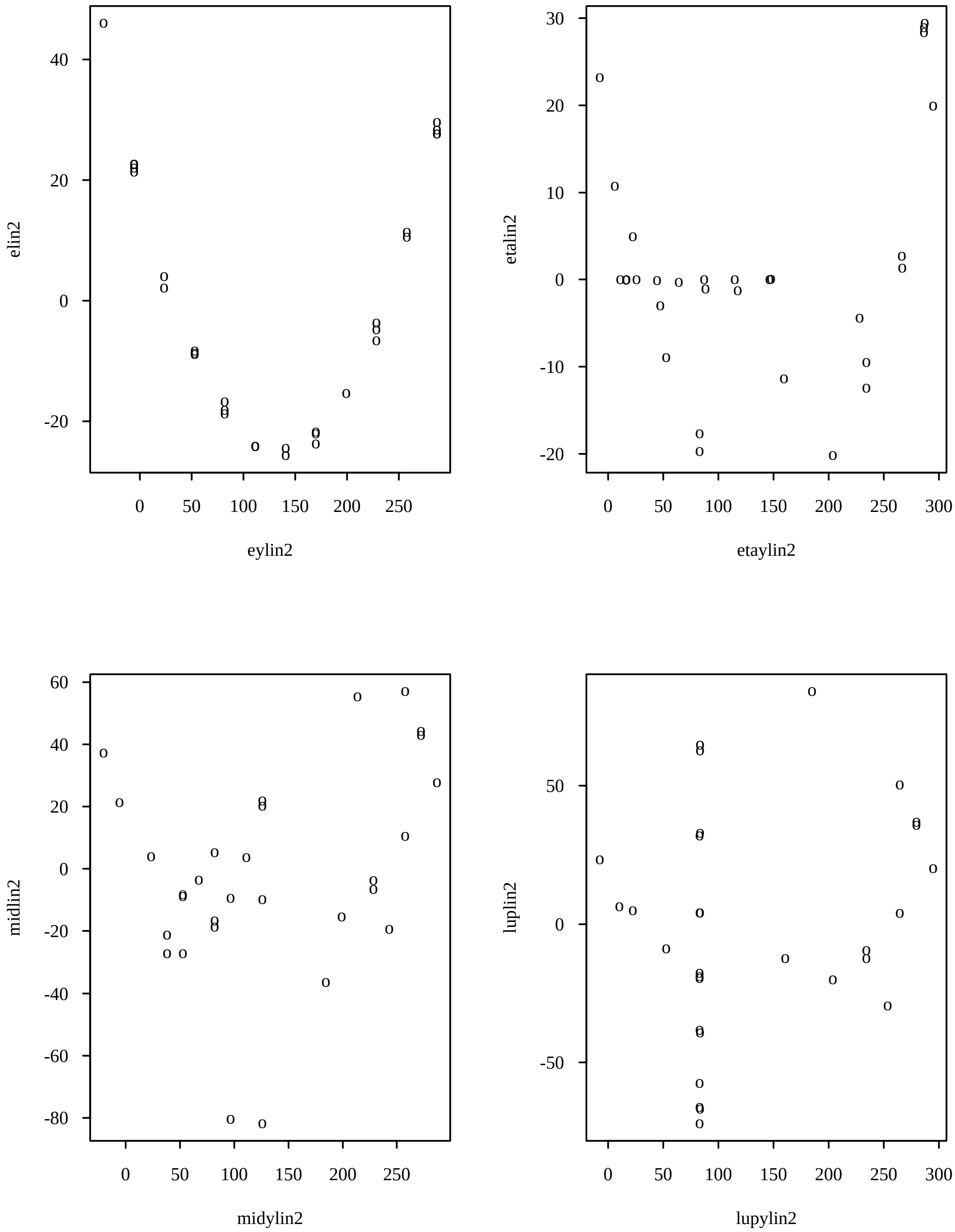
Appendix B

Residual plots when a quadratic term is missing

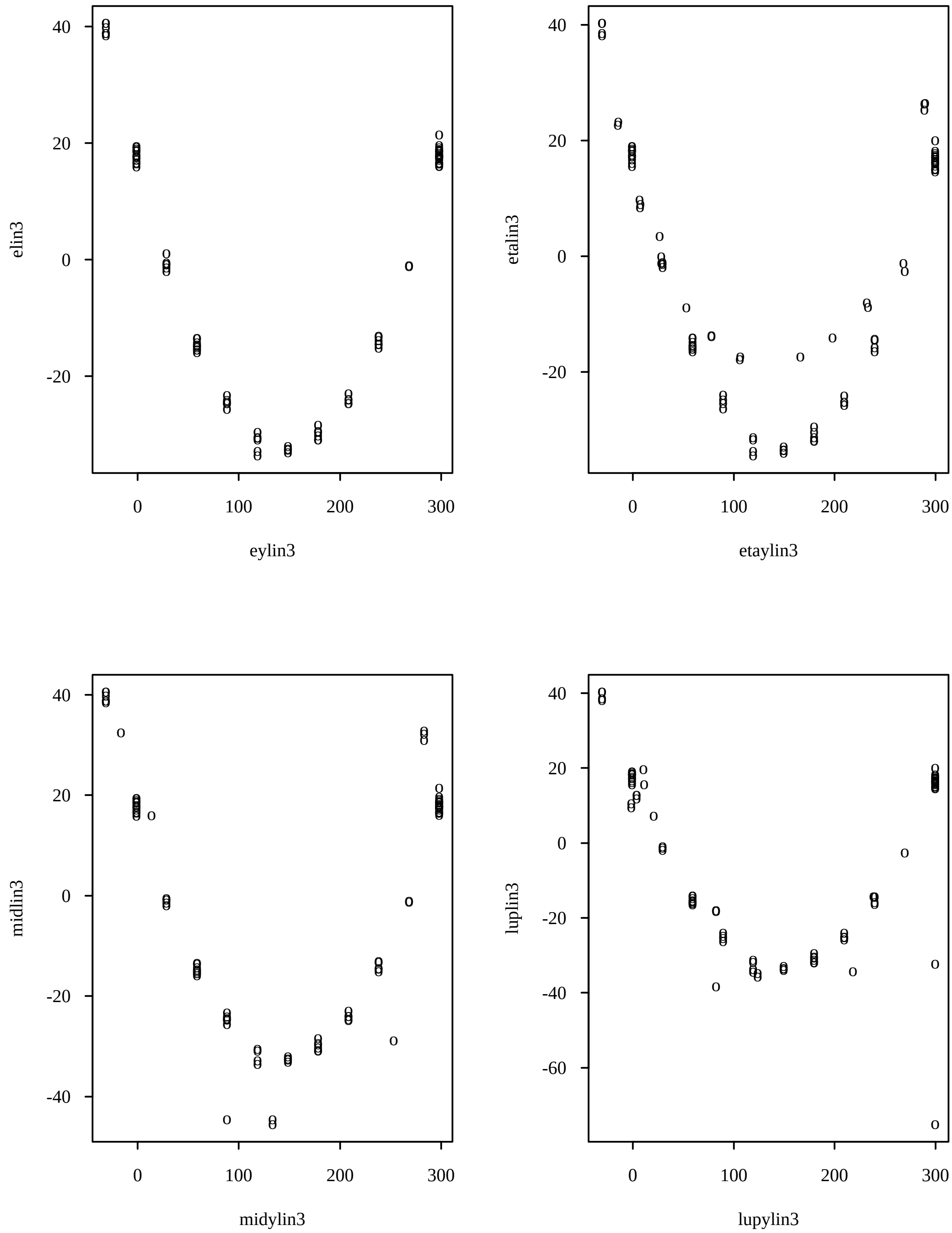
Scenario 1: covariate distribution $\mathrm{Exp}(\frac{1}{8})$, β=2, p=0.3, n=100:



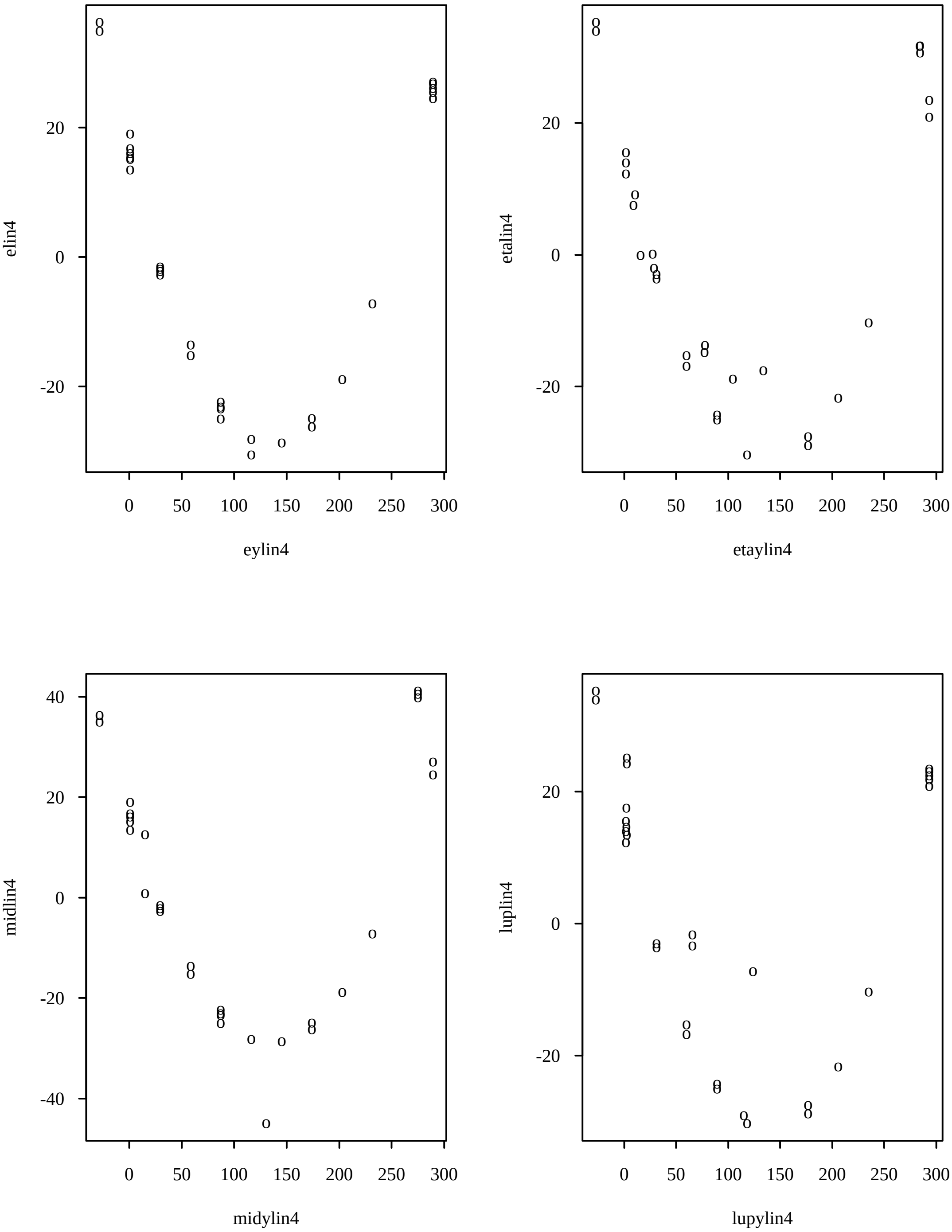
Scenario 2: covariate distribution Exp($\frac{1}{8}$), β=2, p=0.3, n=30:



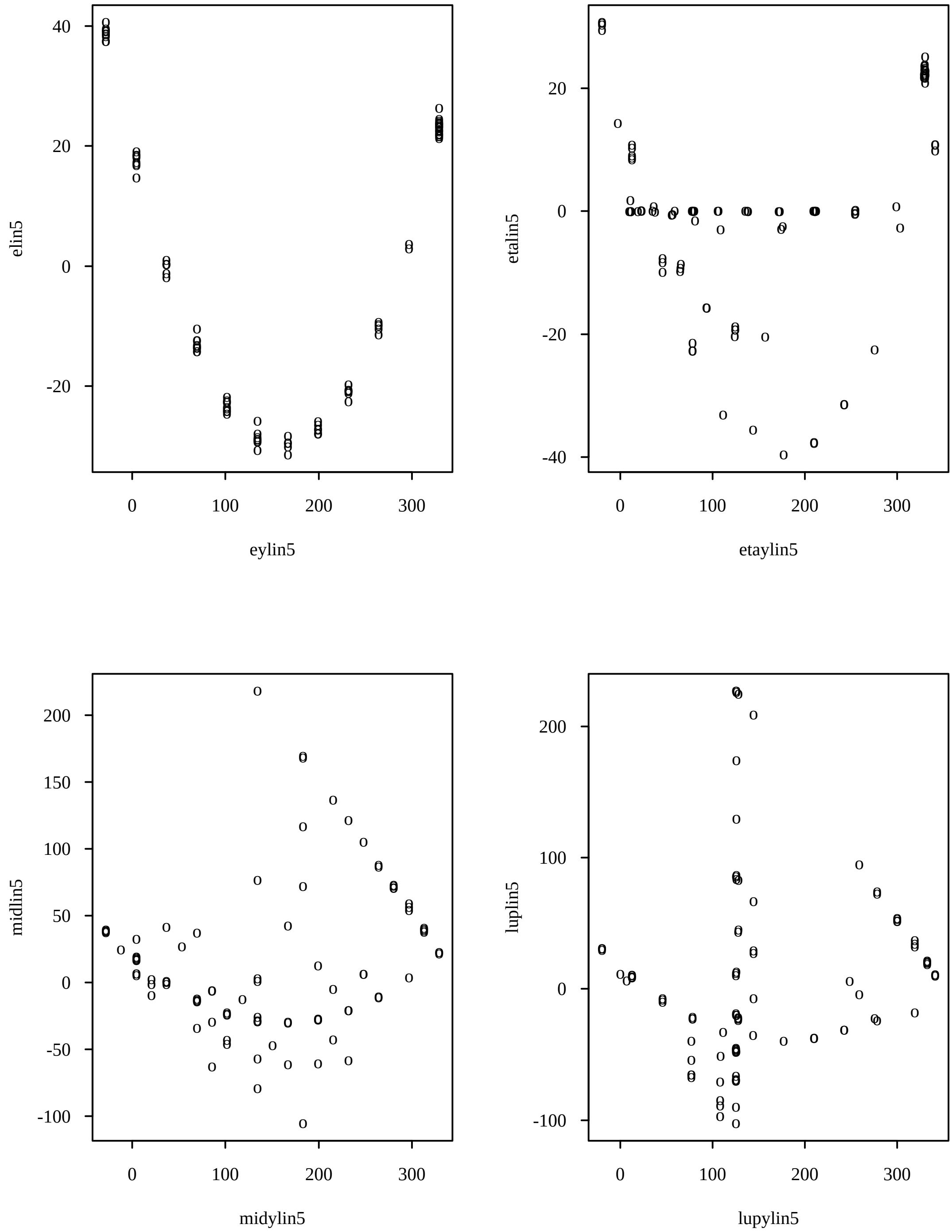
Scenario 3: covariate distribution Exp($\frac{1}{8}$), β=2, p=0.7, n=100:



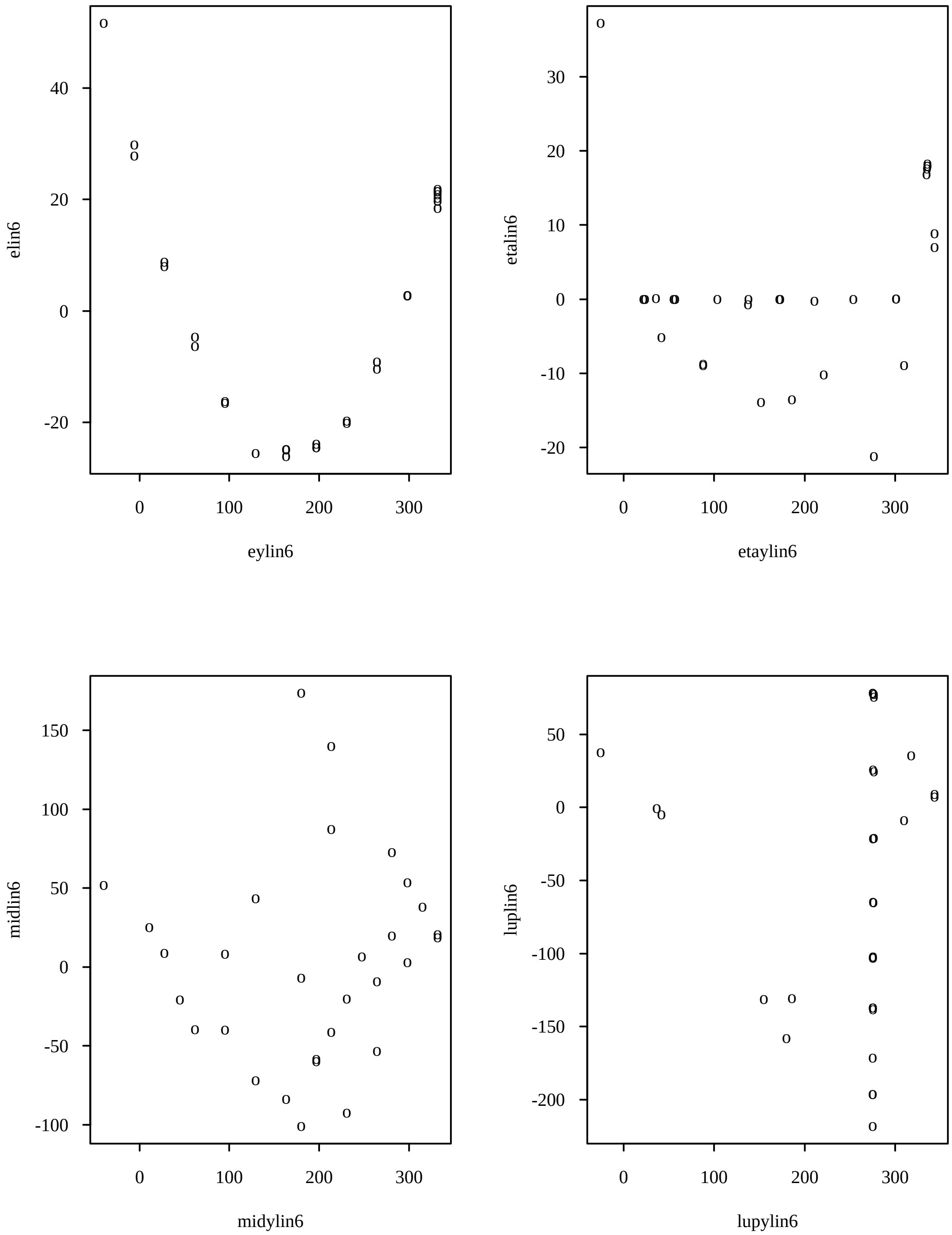
Scenario 4: covariate distribution $\text{Exp}(\frac{1}{8})$, β=2, p=0.7, n=30:



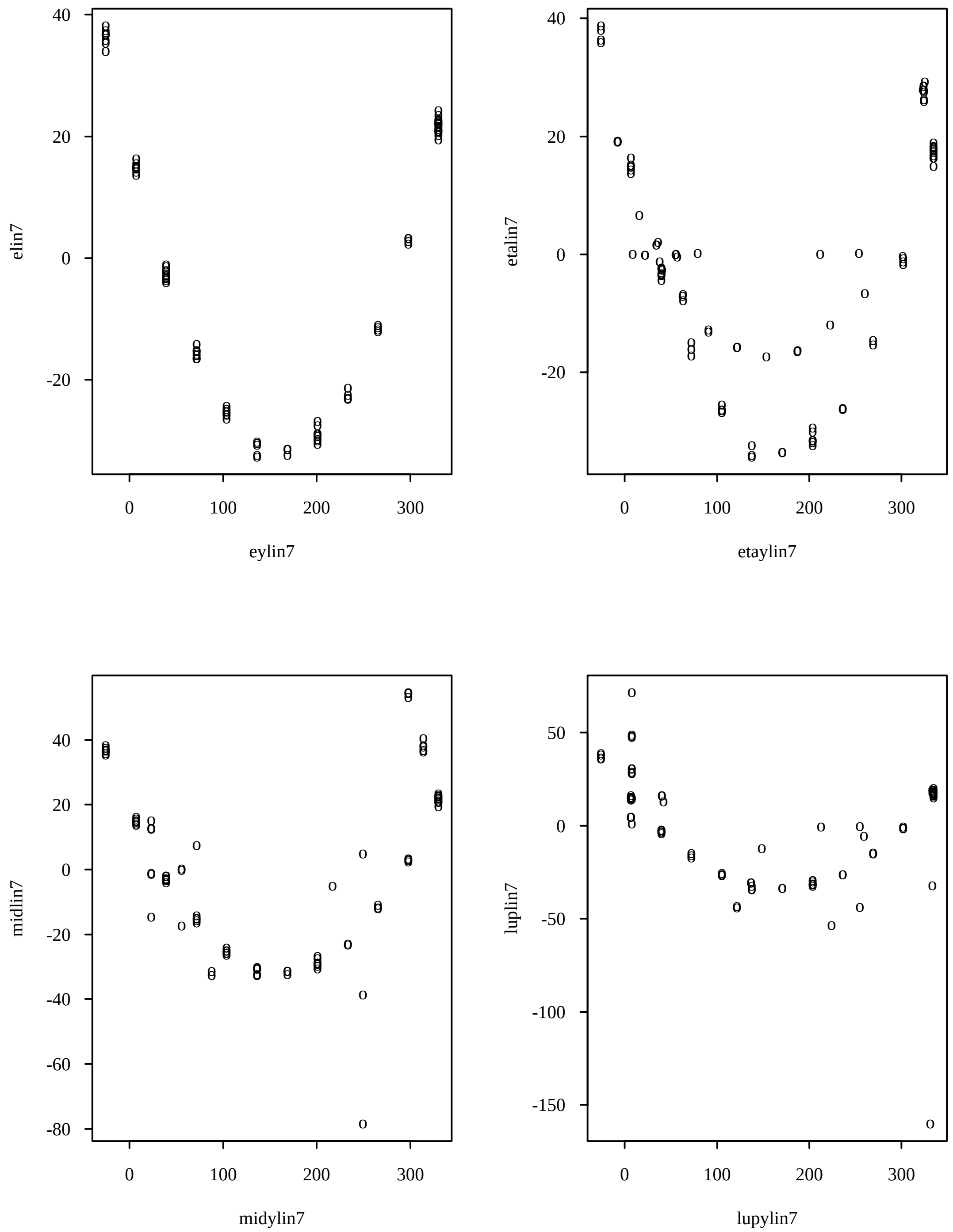
Scenario 5: covariate distribution Exp($\frac{1}{8}$), β=5, p=0.3, n=100:



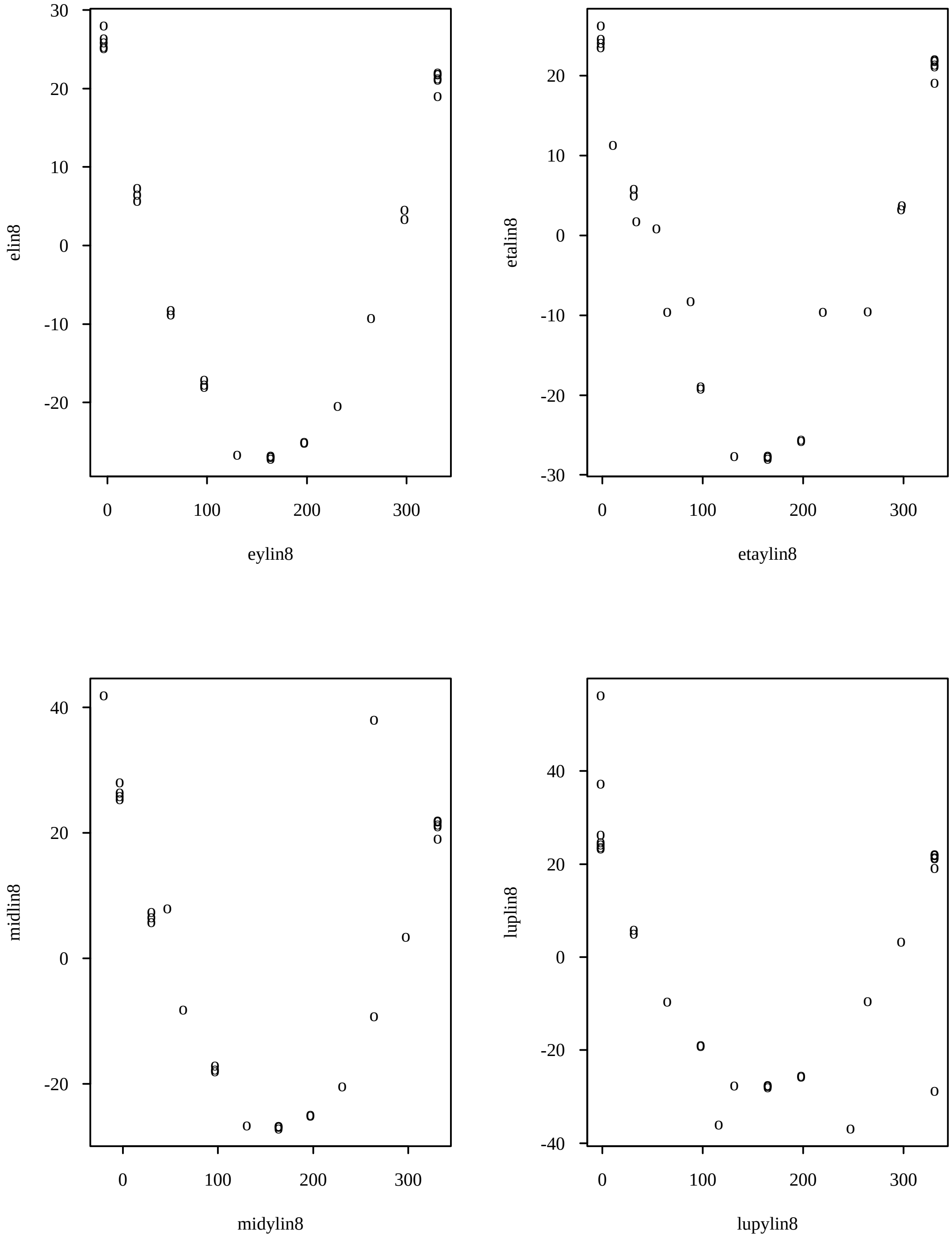
Scenario 6: covariate distribution Exp($\frac{1}{8}$), β=5, p=0.3, n=30:



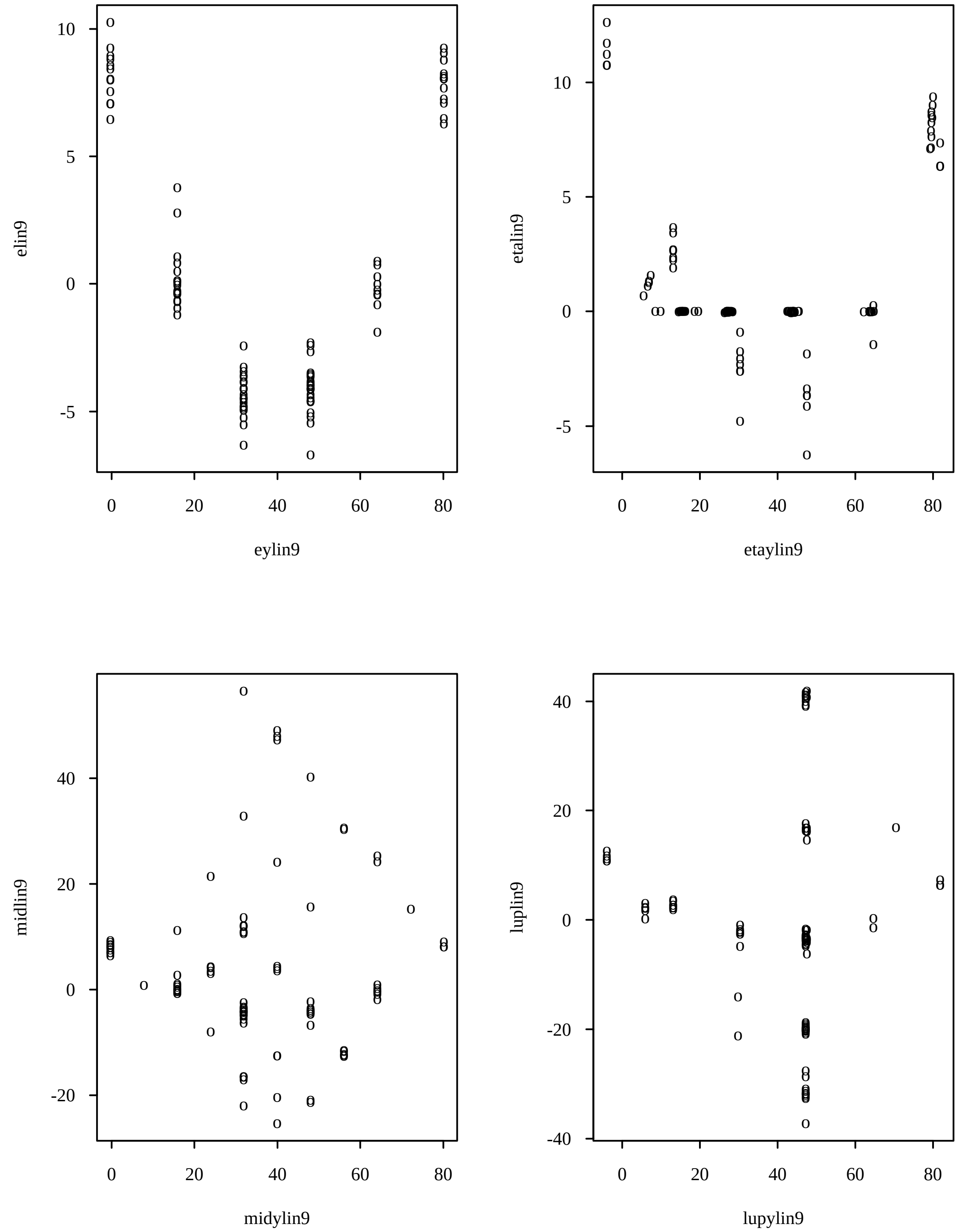
Scenario 7: covariate distribution $\mathrm{Exp}(\frac{1}{8})$, β=5, p=0.7, n=100:



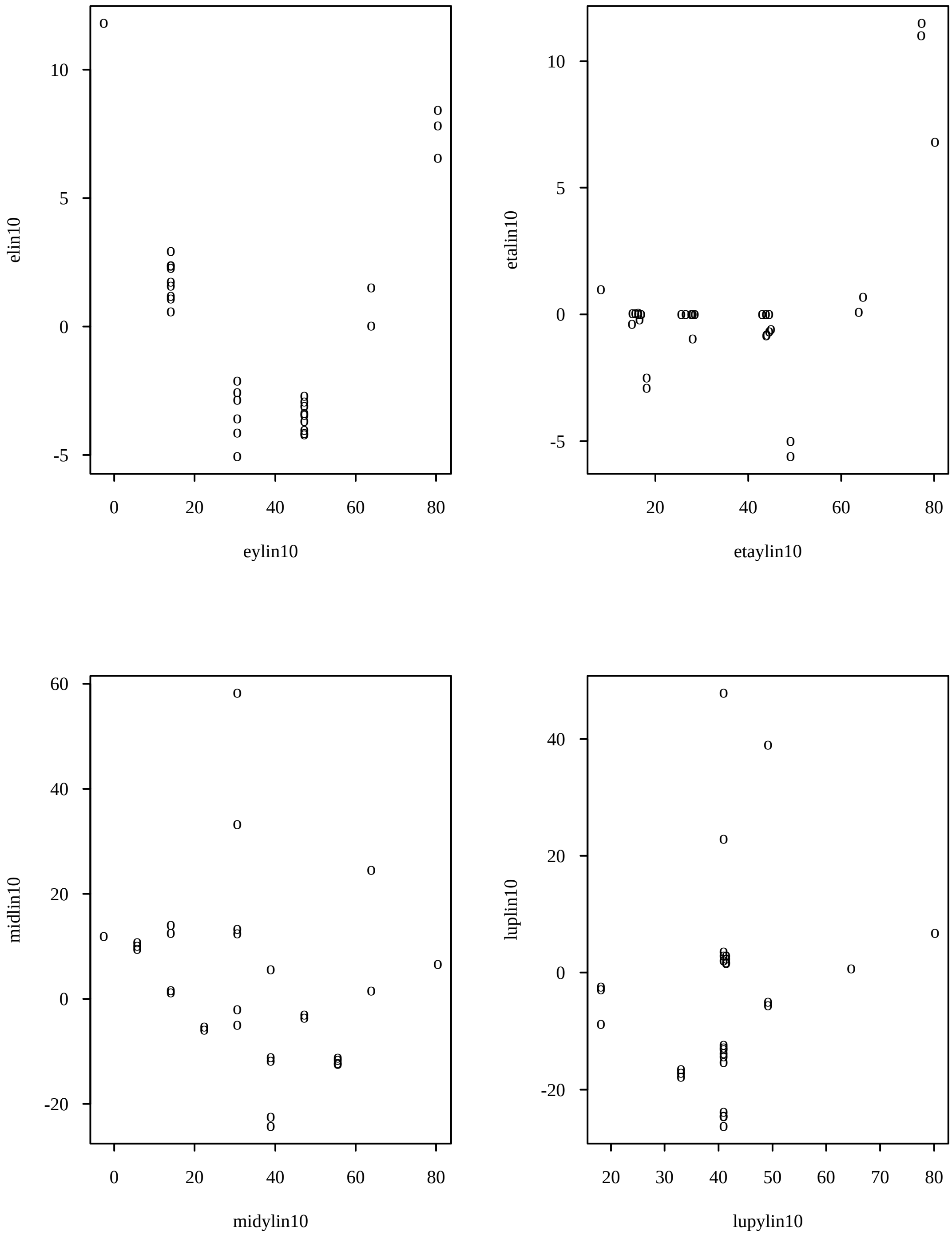
Scenario 8: covariate distribution Exp($\frac{1}{8}$), β=5, p=0.7, n=30:



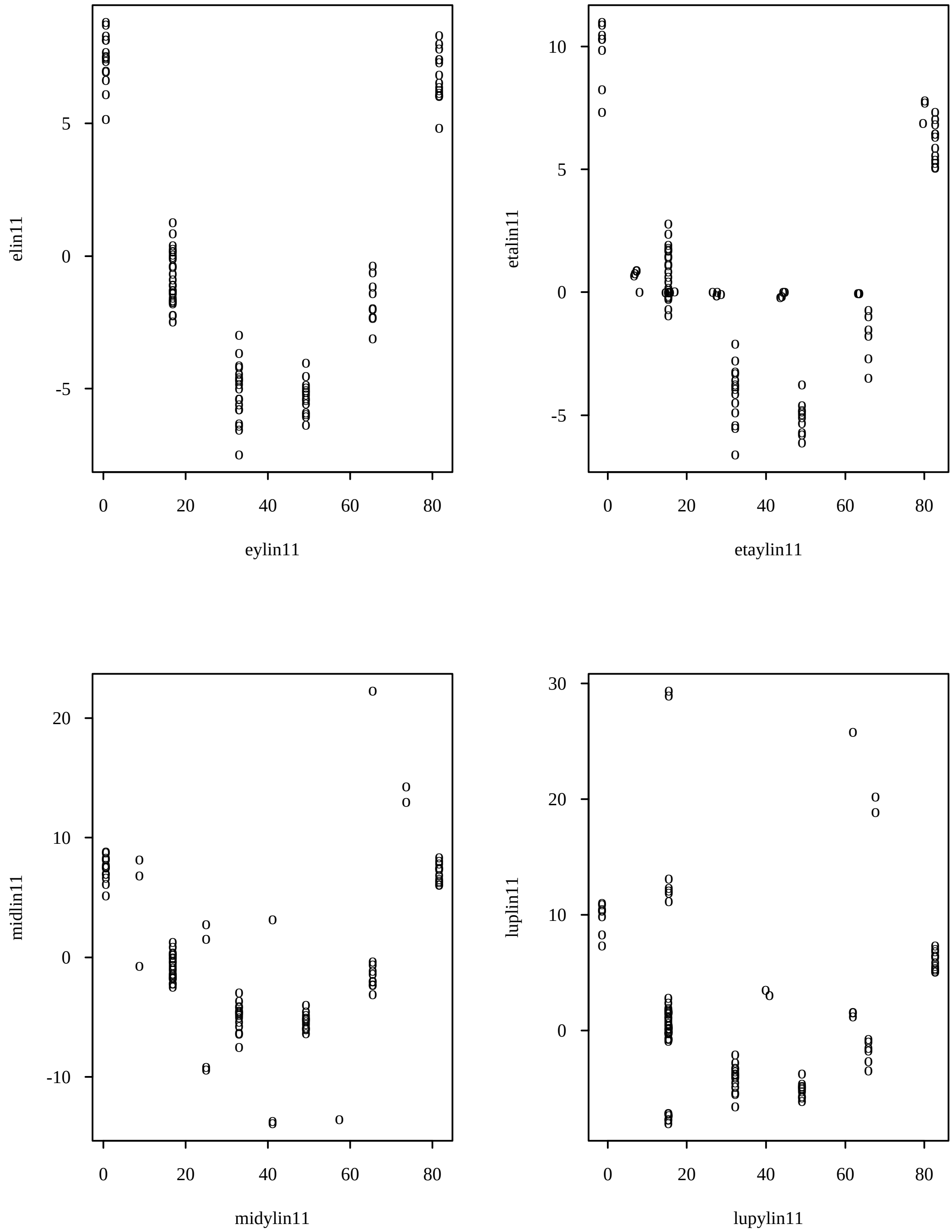
Scenario 9: covariate distribution $W(\frac{1}{6}, \frac{3}{2})$, β=2, p=0.3, n=100:



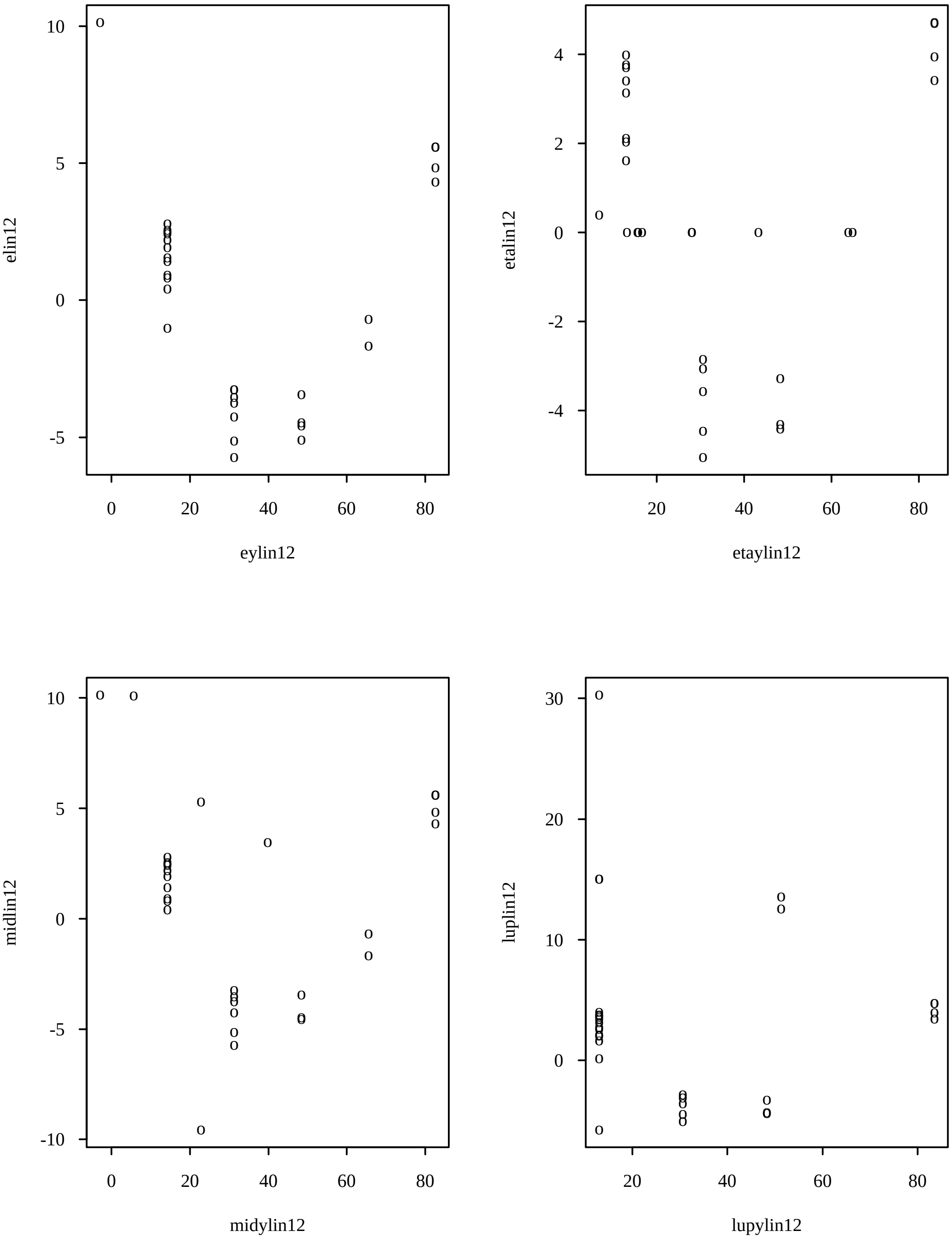
Scenario 10: covariate distribution $W(\frac{1}{6}, \frac{3}{2})$, β=2, p=0.3, n=30:



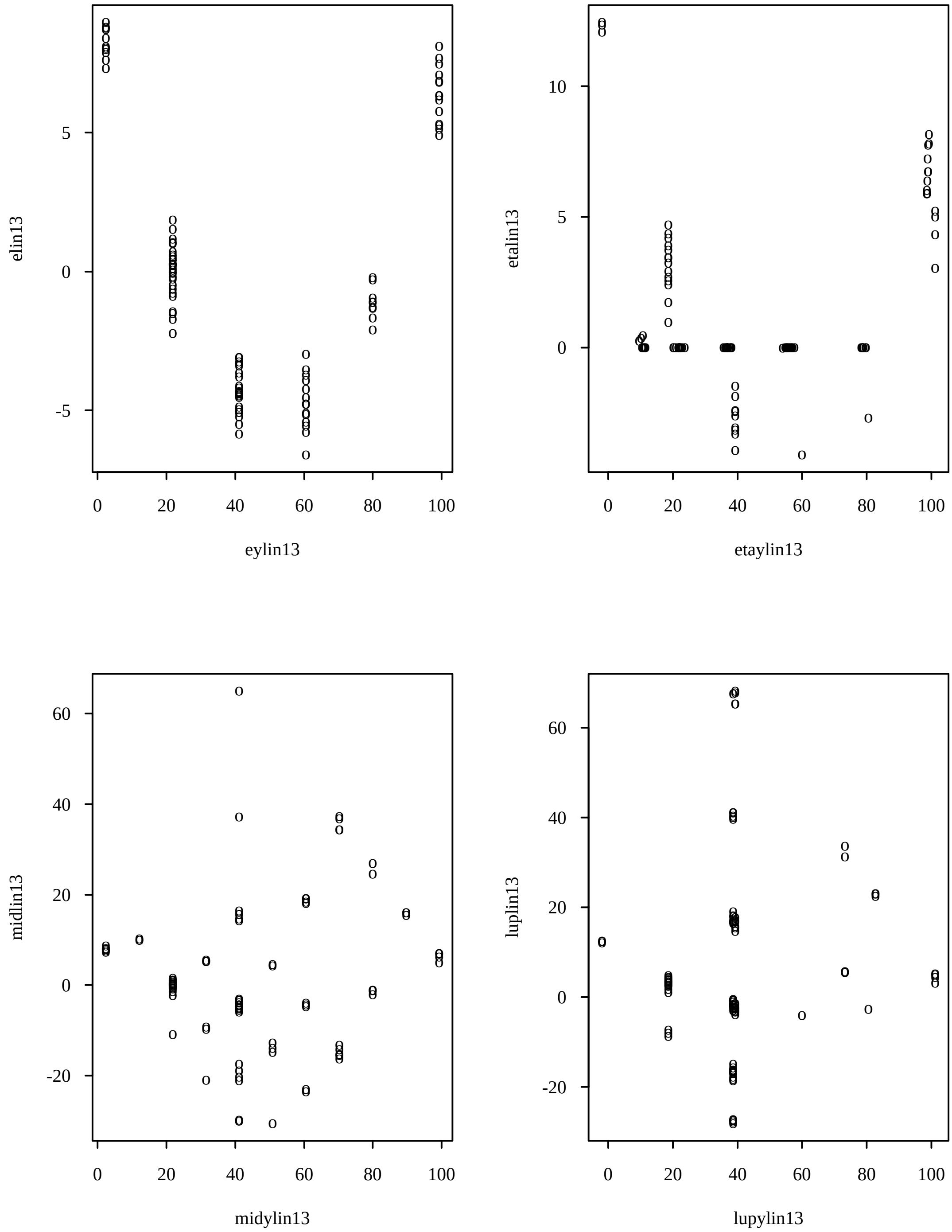
Scenario 11: covariate distribution $W(\frac{1}{6}, \frac{3}{2})$, β=2, p=0.7, n=100:



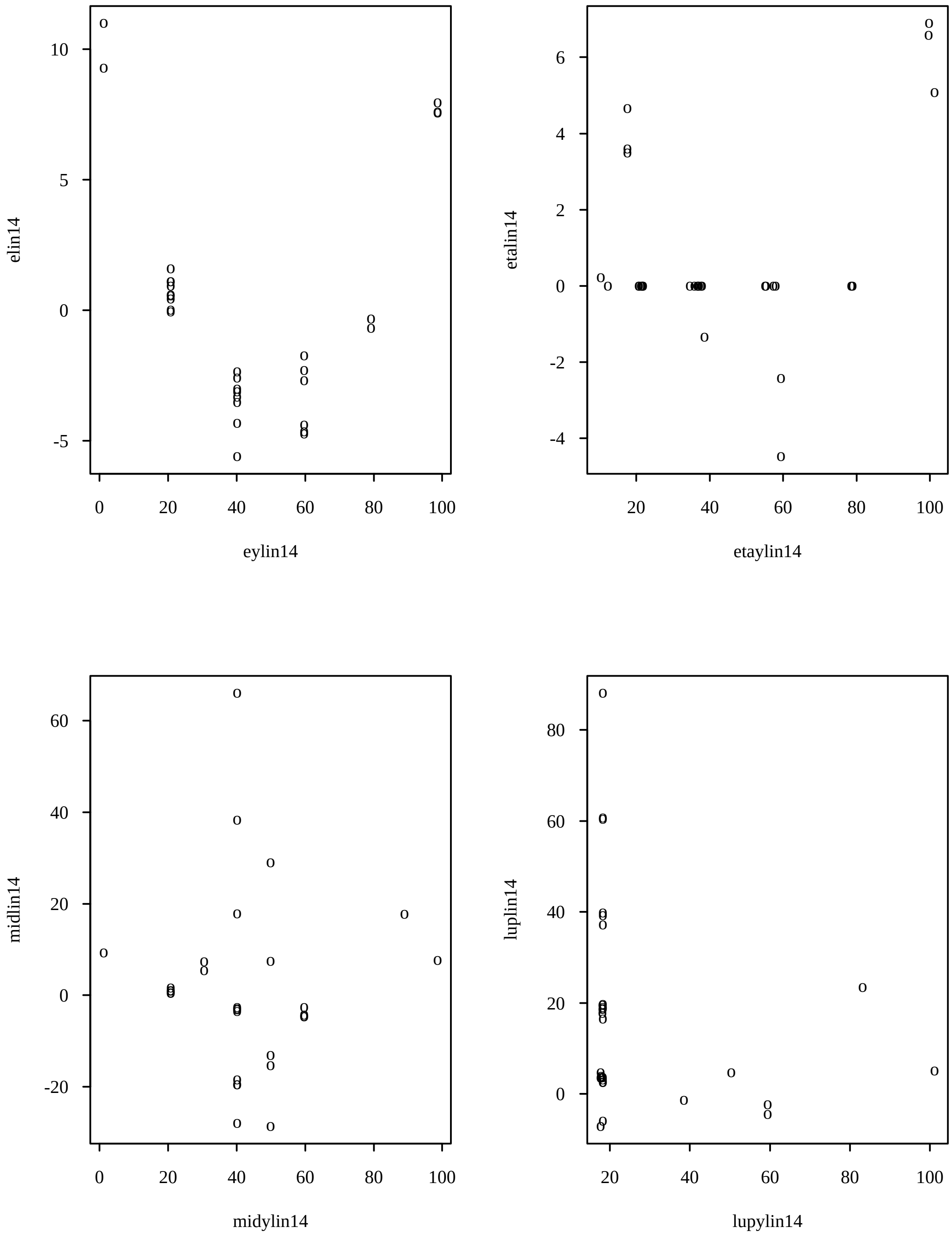
Scenario 12: covariate distribution $W(\frac{1}{6}, \frac{3}{2})$, β=2, p=0.7, n=30:



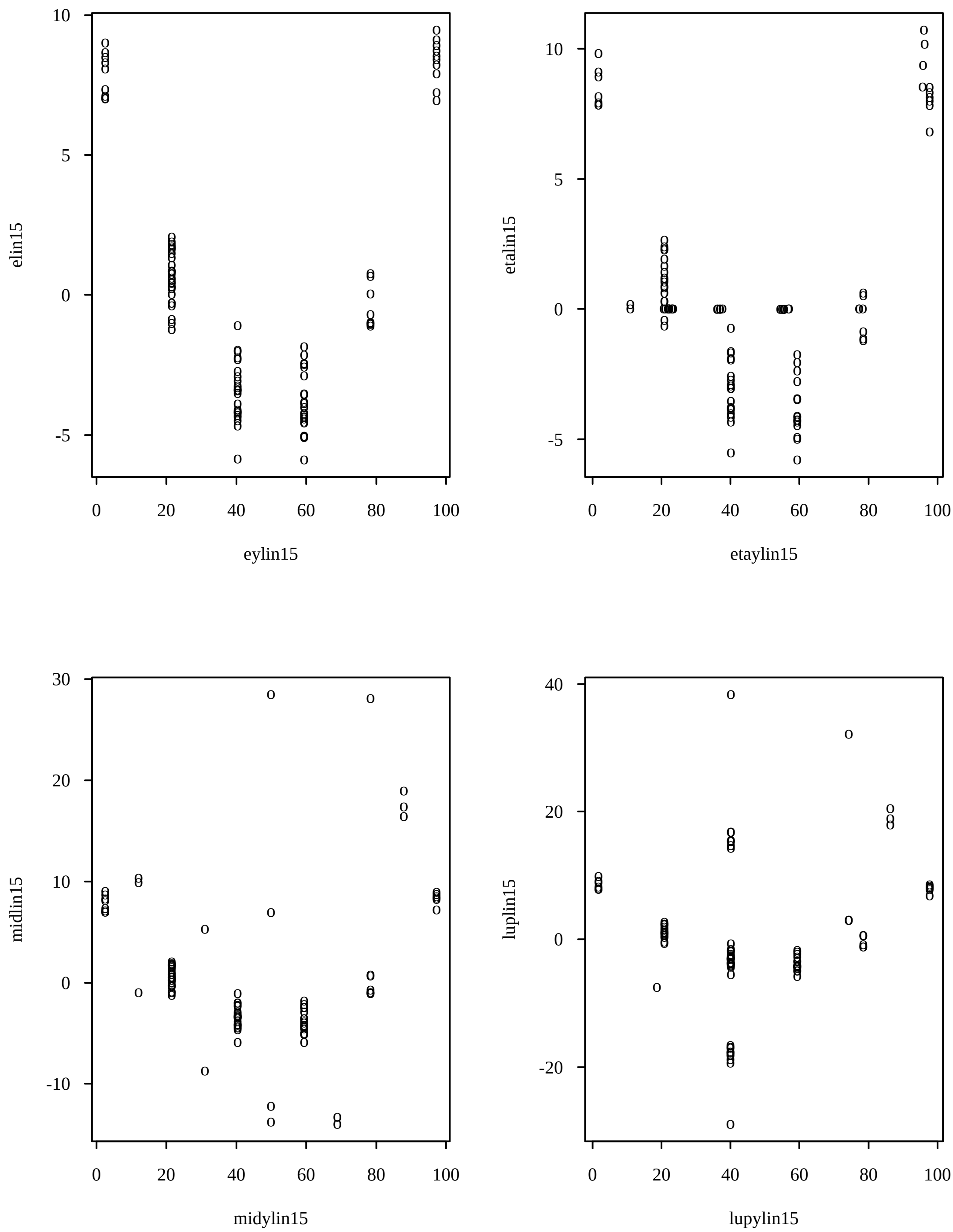
Scenario 13: covariate distribution $W(\frac{1}{6}, \frac{3}{2})$, β=5, p=0.3, n=100:



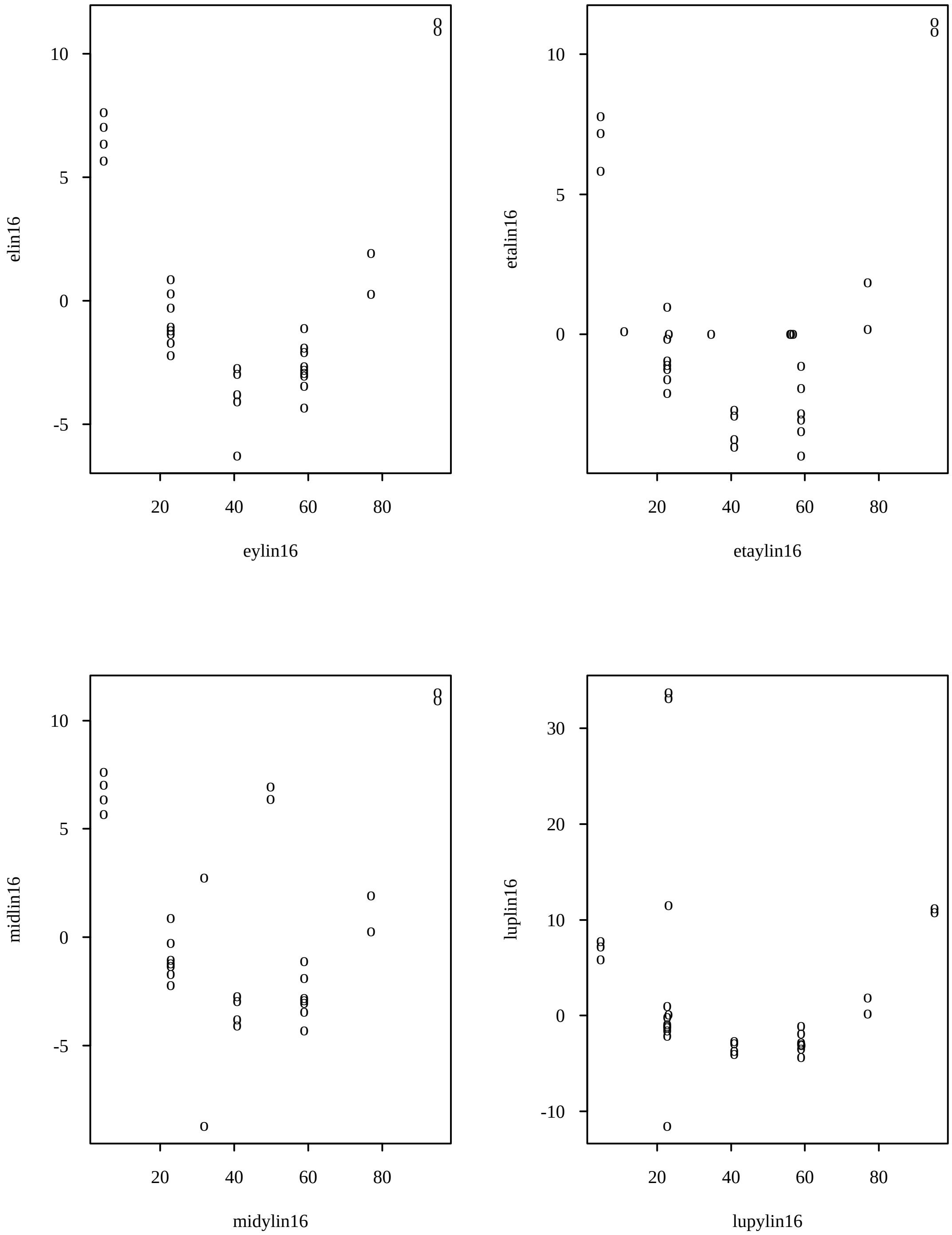
Scenario 14: covariate distribution $W(\frac{1}{6}, \frac{3}{2})$, β=5, p=0.3, n=30:



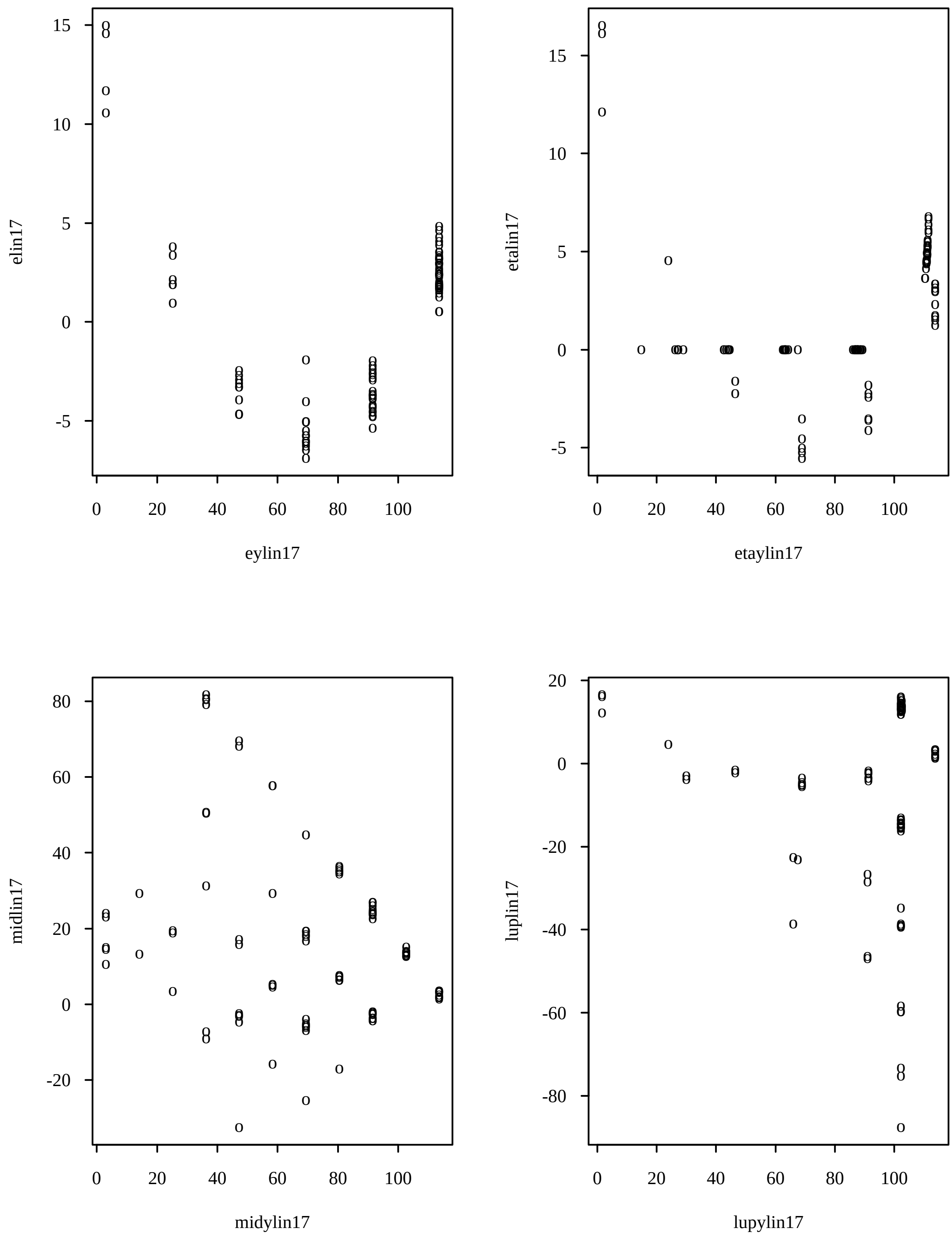
Scenario 15: covariate distribution $W(\frac{1}{6}, \frac{3}{2})$, β=5, p=0.7, n=100:



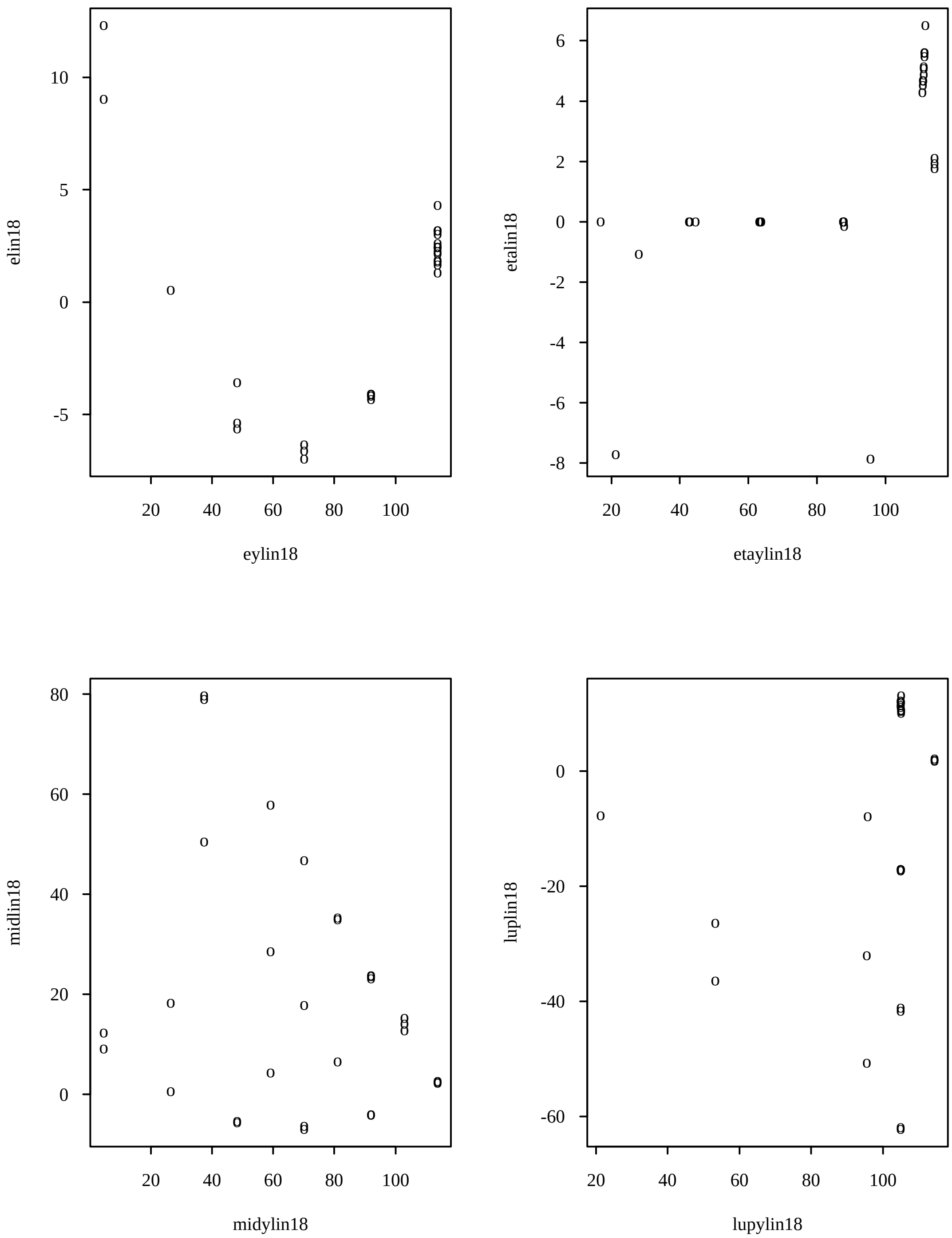
Scenario 16: covariate distribution $W(\frac{1}{6}, \frac{3}{2})$, β=5, p=0.7, n=30:



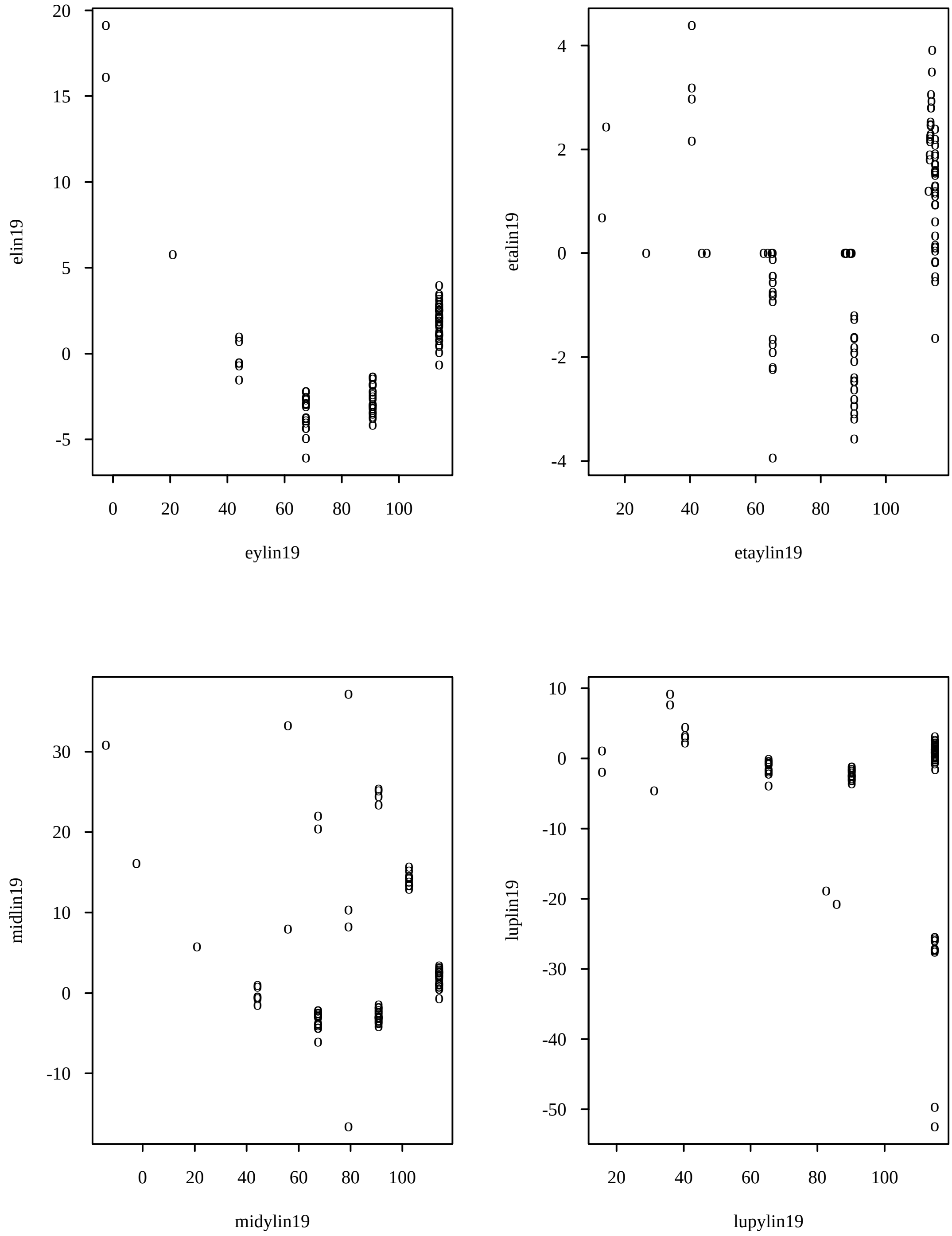
Scenario 17: covariate distribution N(6,4), β=2, p=0.3, n=100:



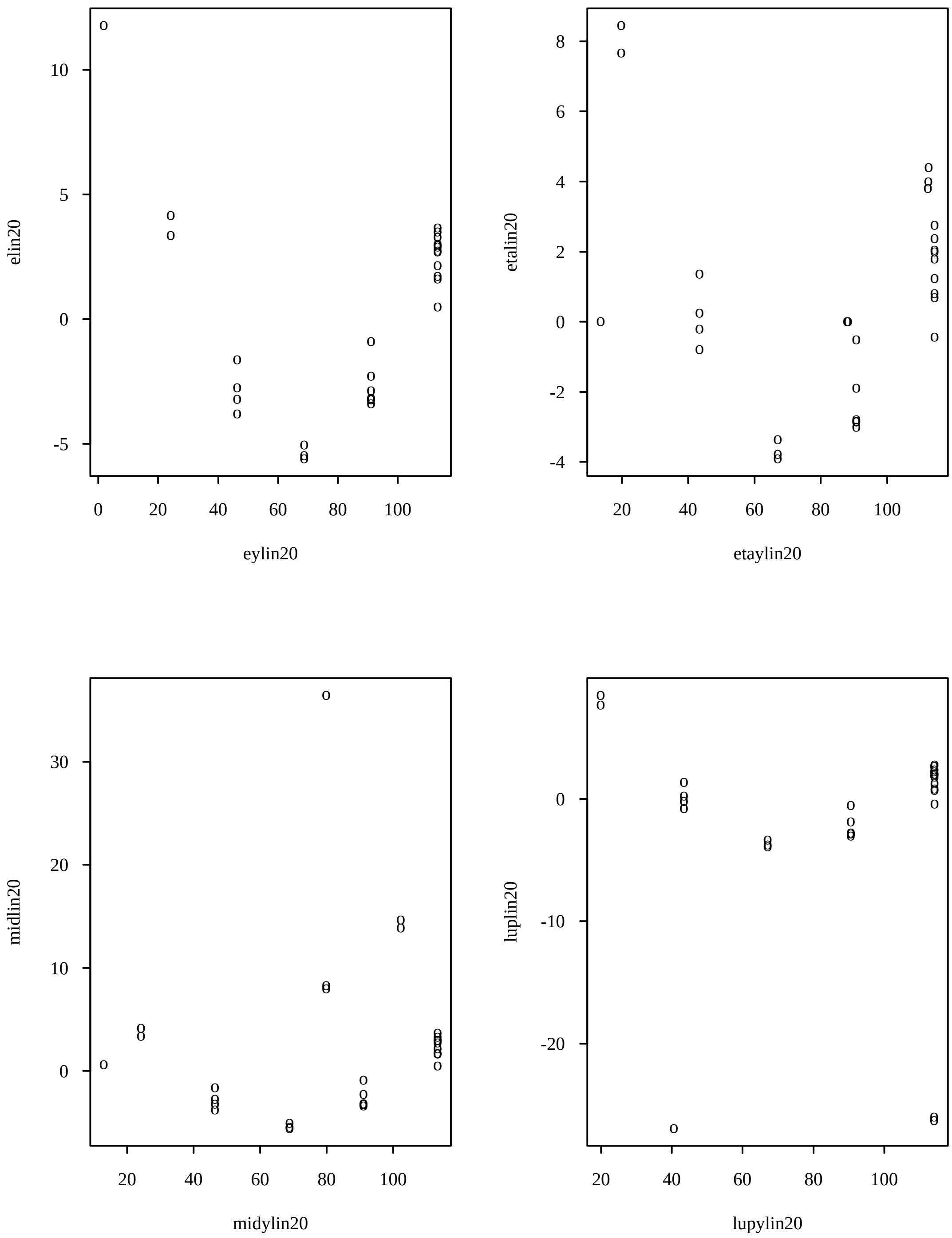
Scenario 18: covariate distribution N(6,4), β=2, p=0.3, n=30:



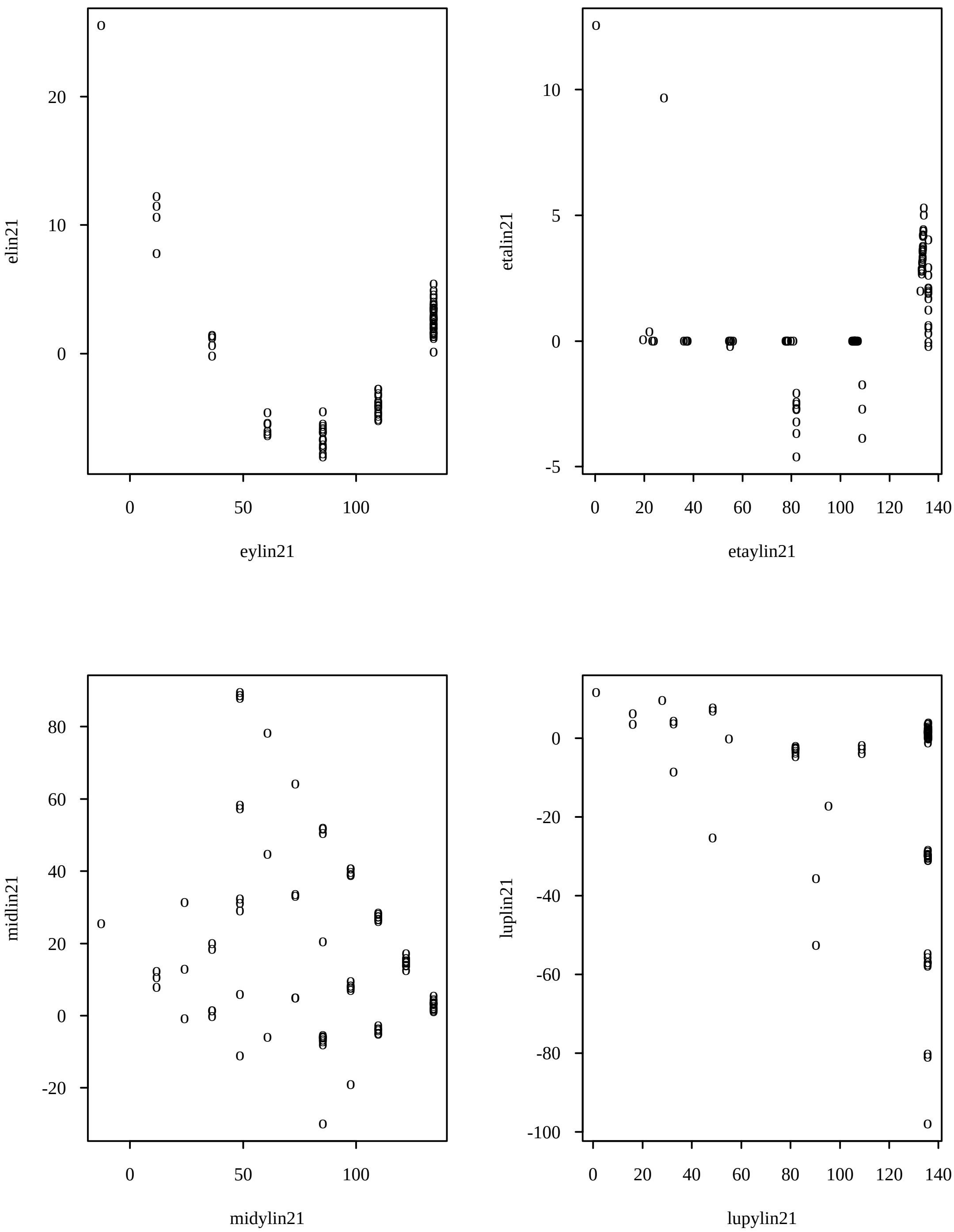
Scenario 19: covariate distribution N(6,4), β=2, p=0.7, n=100:



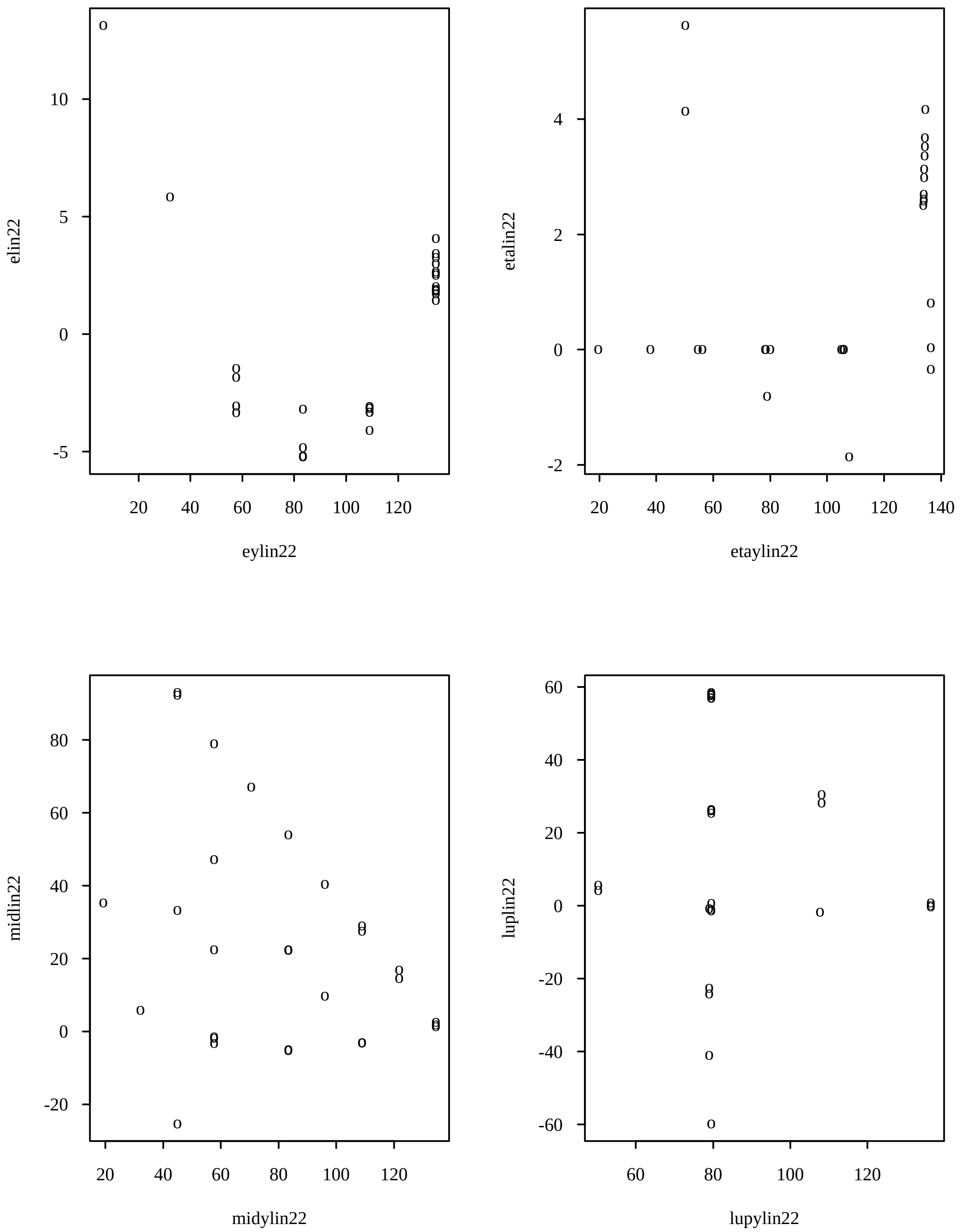
Scenario 20: covariate distribution N(6,4), β=2, p=0.7, n=30:



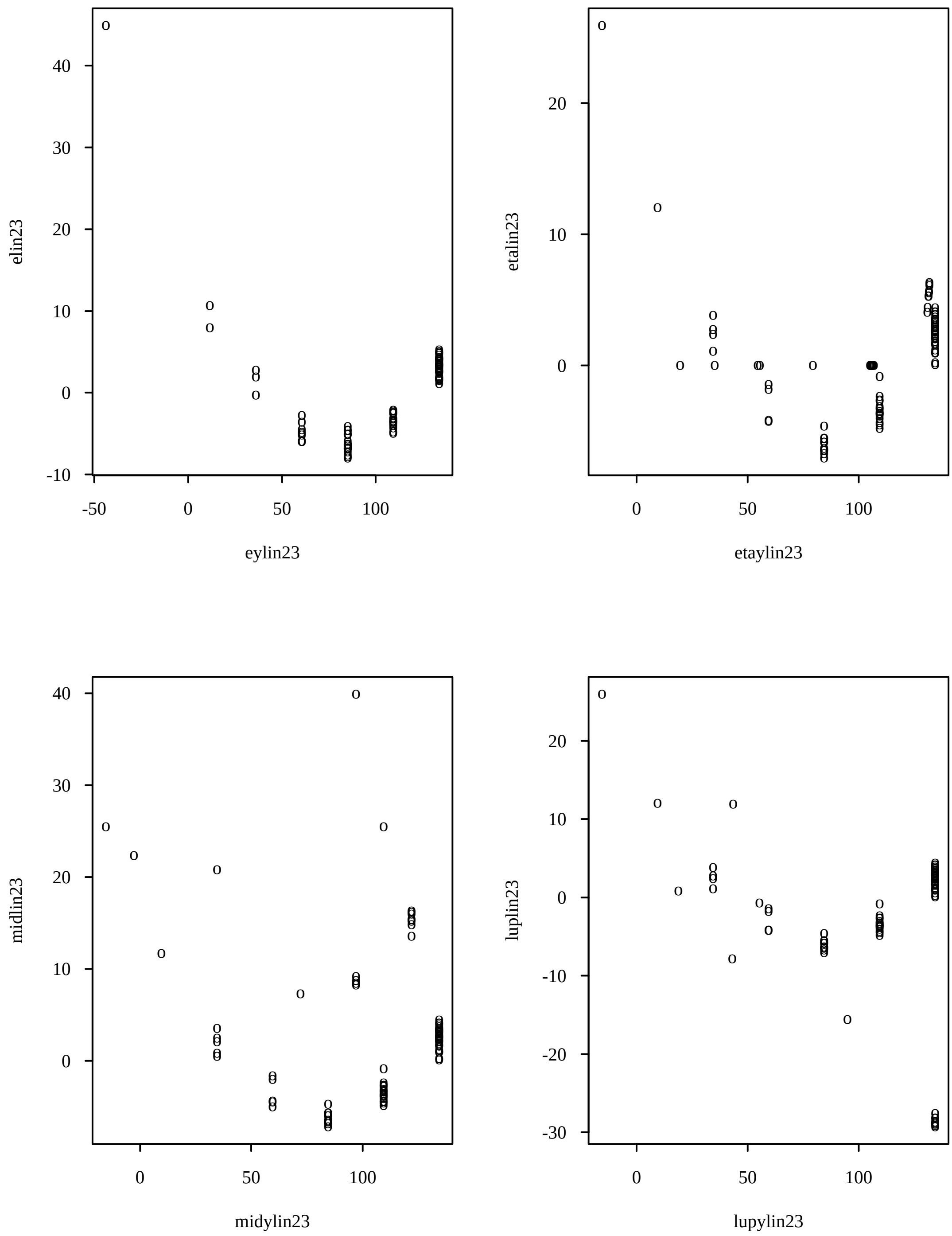
Scenario 21: covariate distribution N(6,4), β=5, p=0.3, n=100:



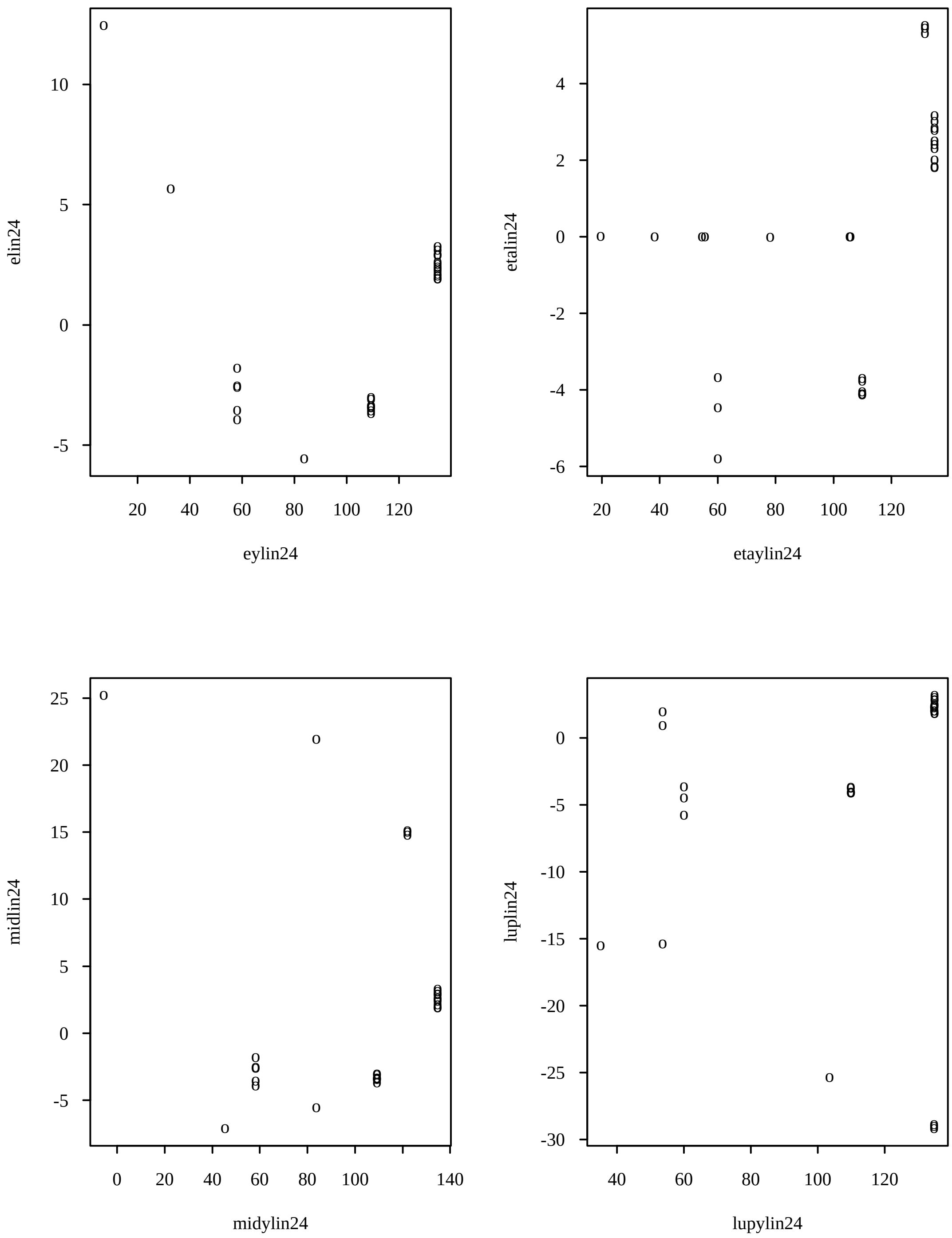
Scenario 22: covariate distribution N(6,4), β=5, p=0.3, n=30:



Scenario 23: covariate distribution N(6,4), β=5, p=0.7, n=100:



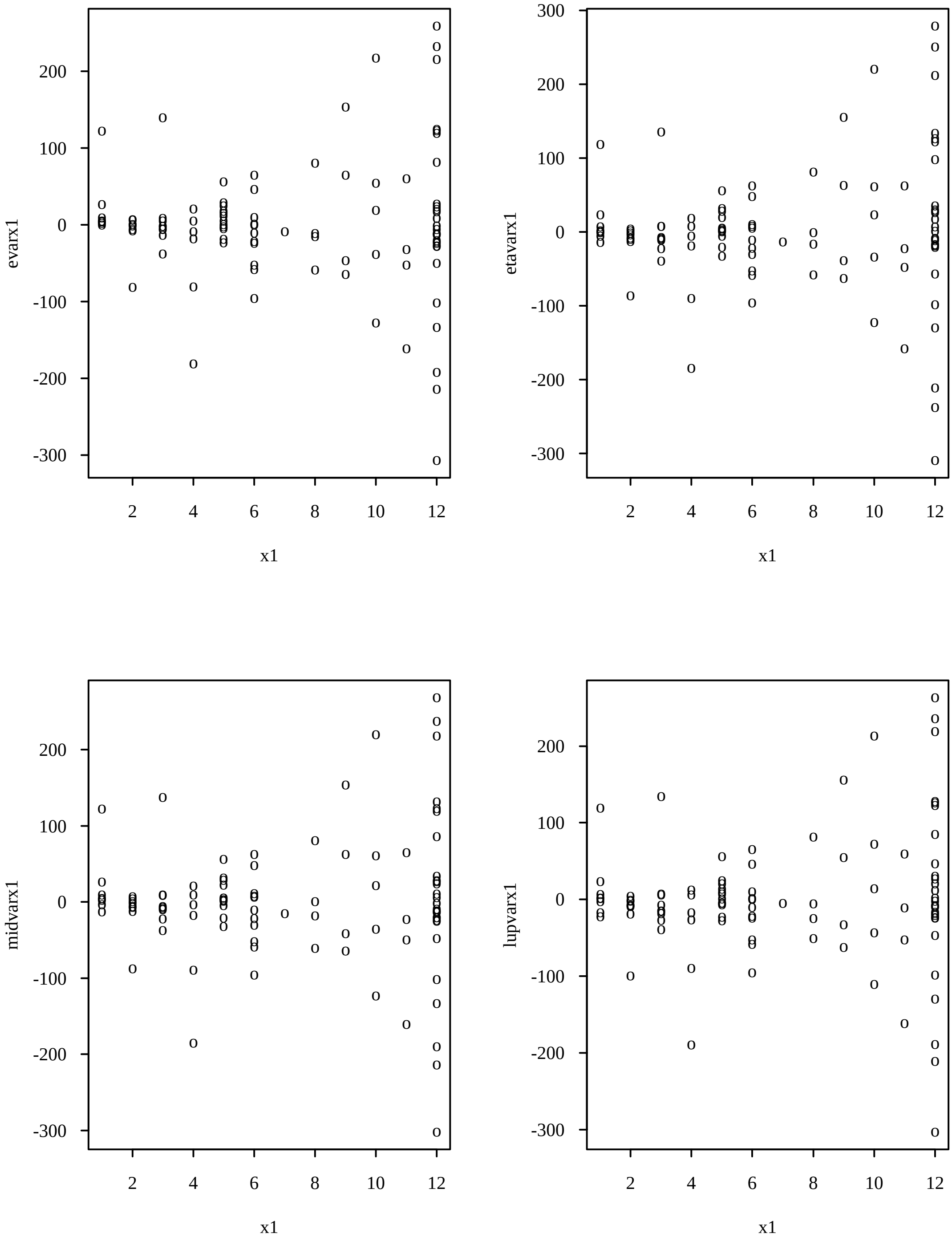
Scenario 24: covariate distribution N(6,4), β=5, p=0.7, n=30:



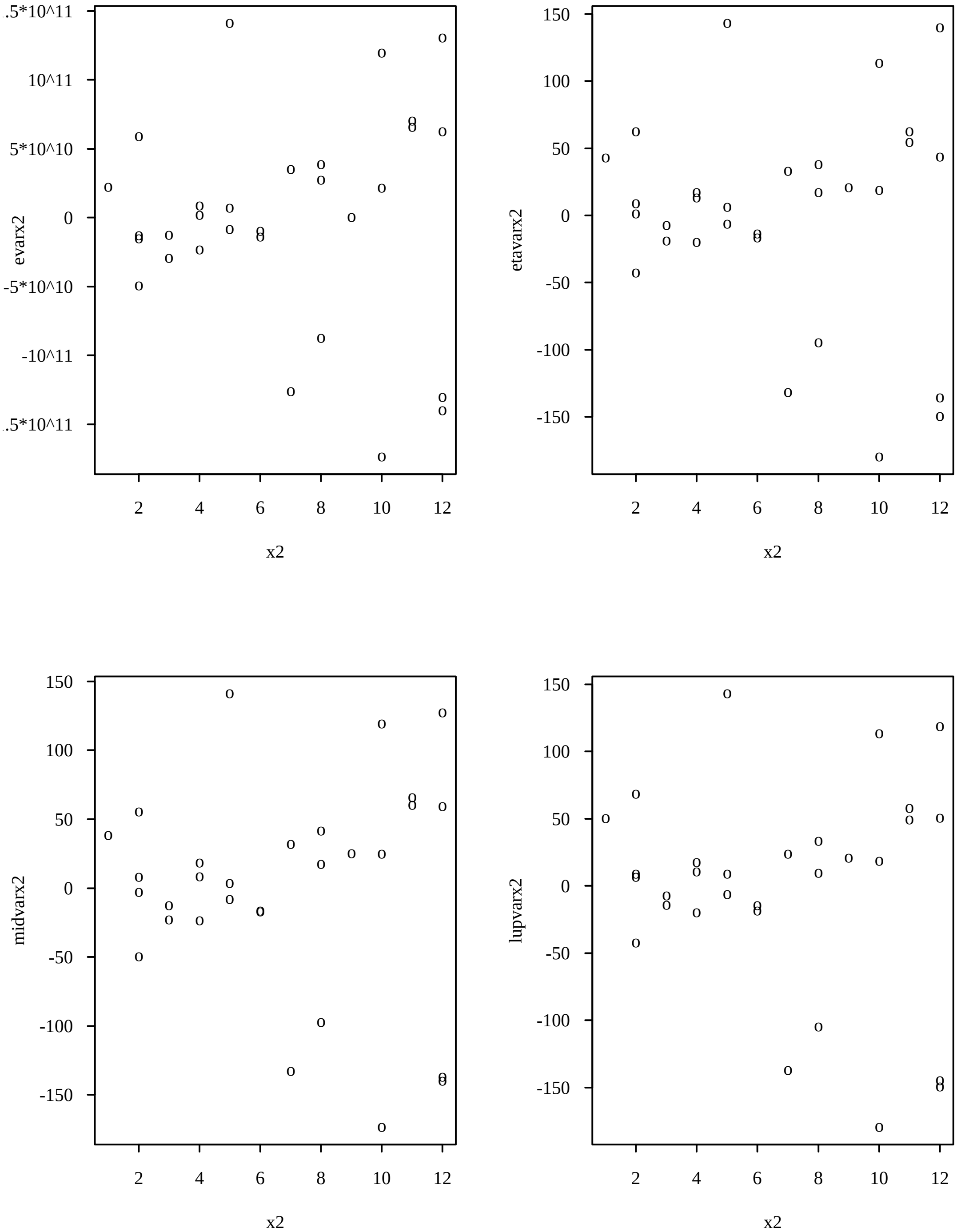
Appendix C

Residual plots when the error variance depends on the covariate

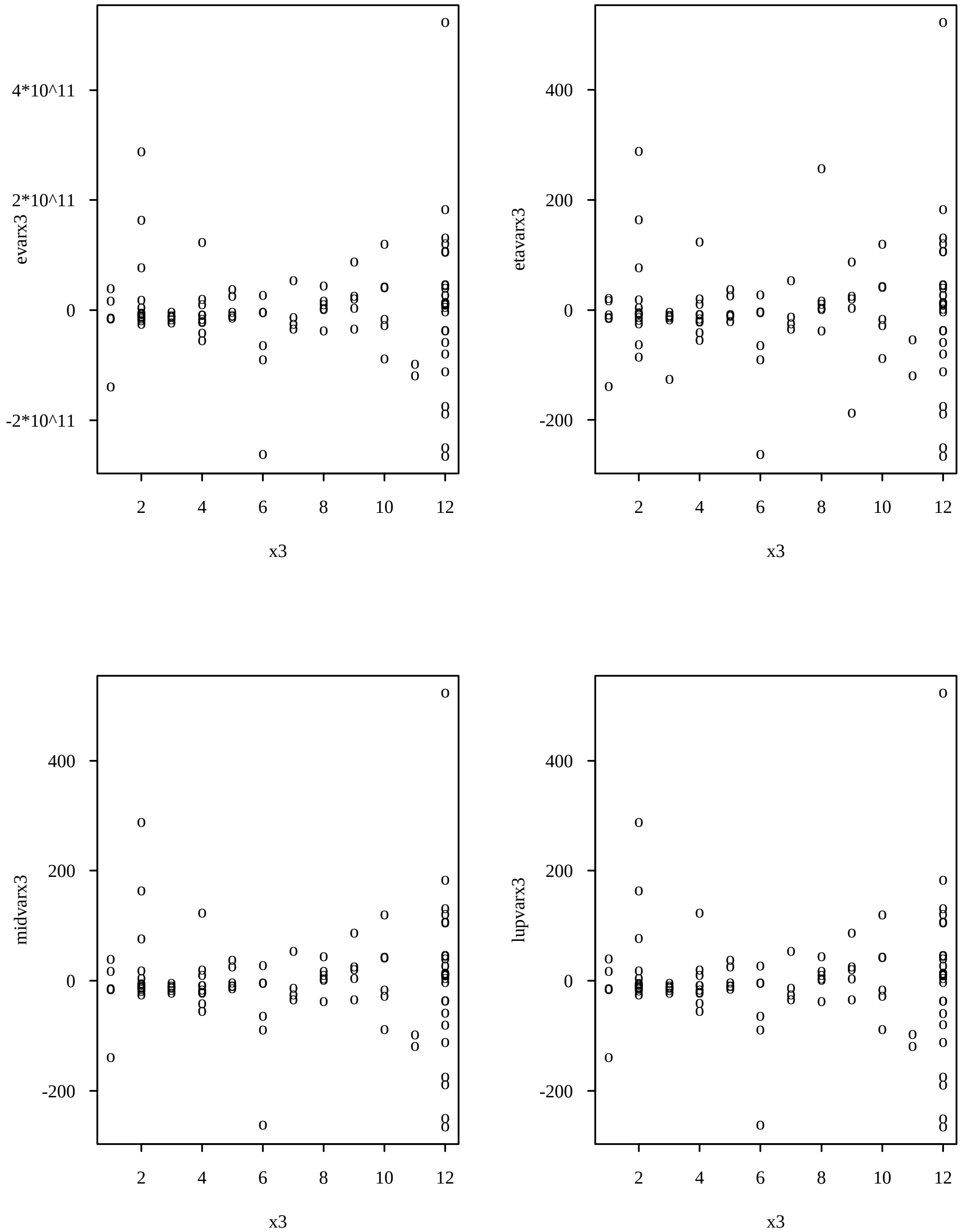
Scenario 1: covariate distribution $\mathrm{Exp}(\frac{1}{8})$, β=2, p=0.3, n=100:



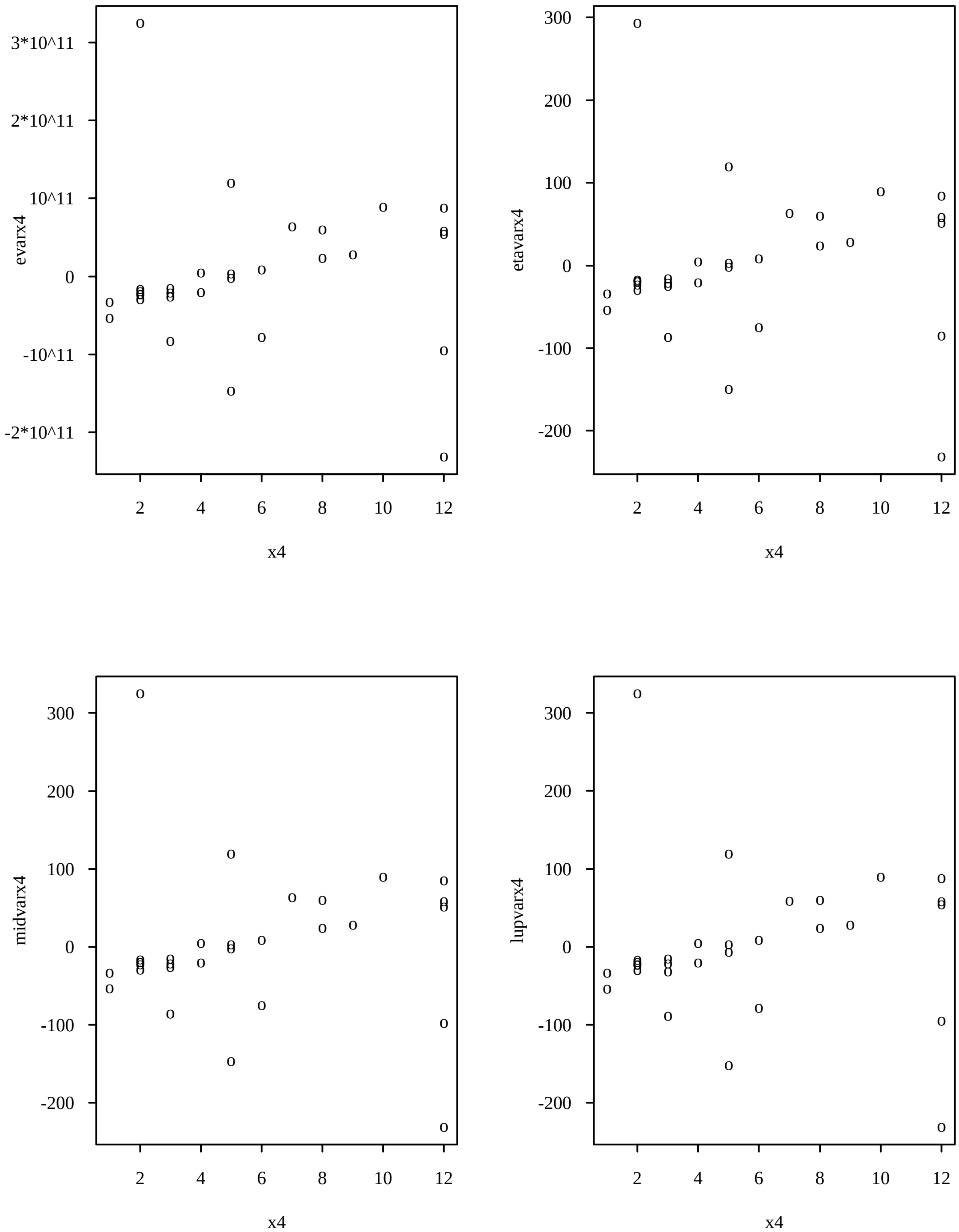
Scenario 2: covariate distribution Exp($\frac{1}{8}$), β=2, p=0.3, n=30:



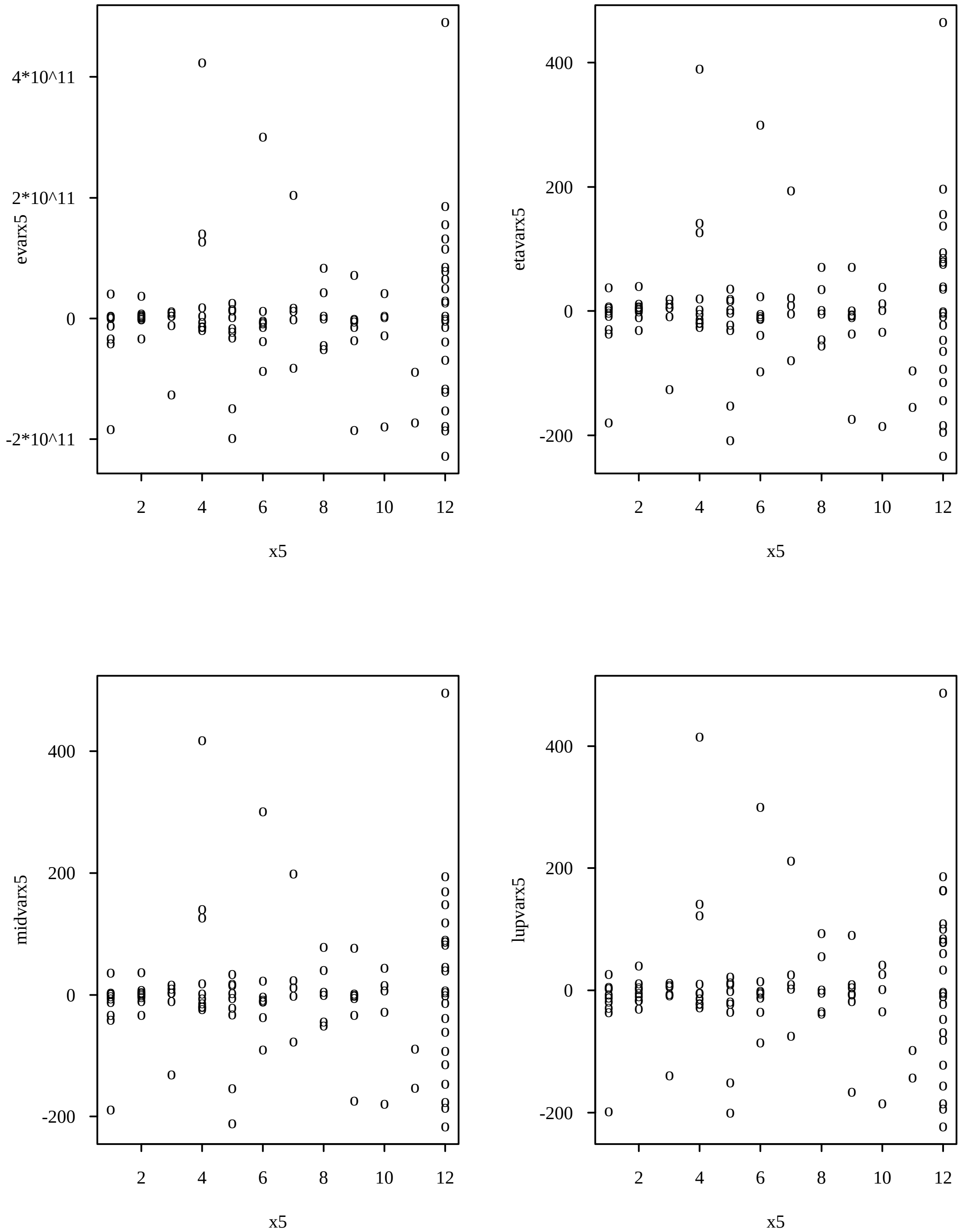
Scenario 3: covariate distribution Exp($\frac{1}{8}$), β=2, p=0.7, n=100:



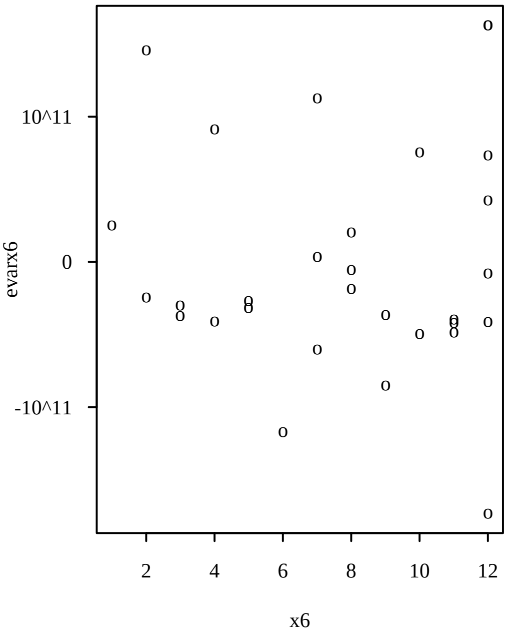
Scenario 4: covariate distribution Exp($\frac{1}{8}$), β=2, p=0.7, n=30:

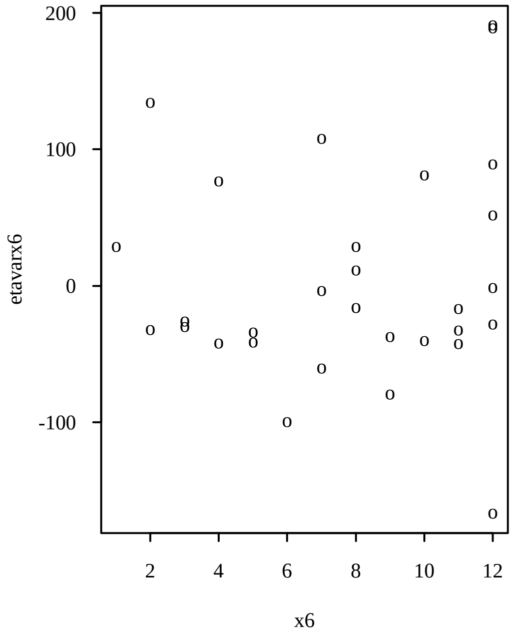


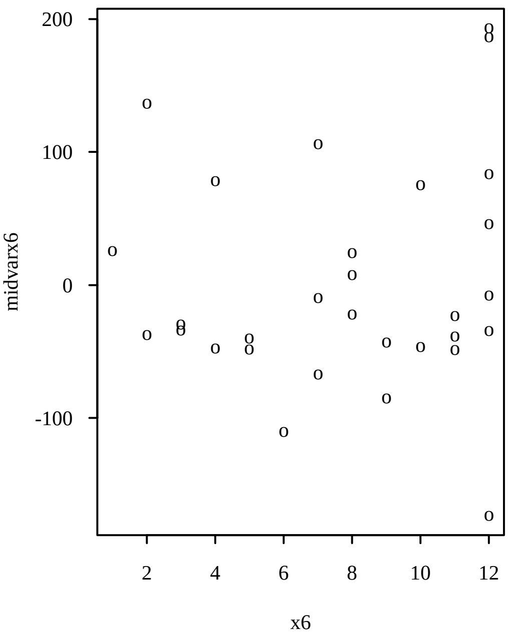
Scenario 5: covariate distribution Exp($\frac{1}{8}$), β=5, p=0.3, n=100:

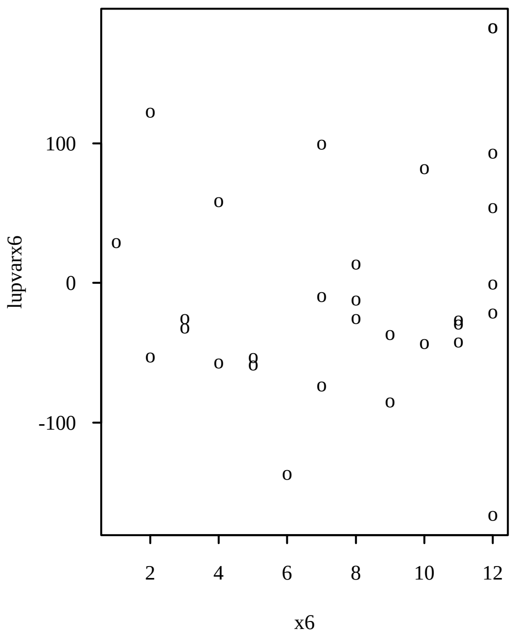


Scenario 6: covariate distribution $\text{Exp}(\frac{1}{8})$, β=5, p=0.3, n=30:

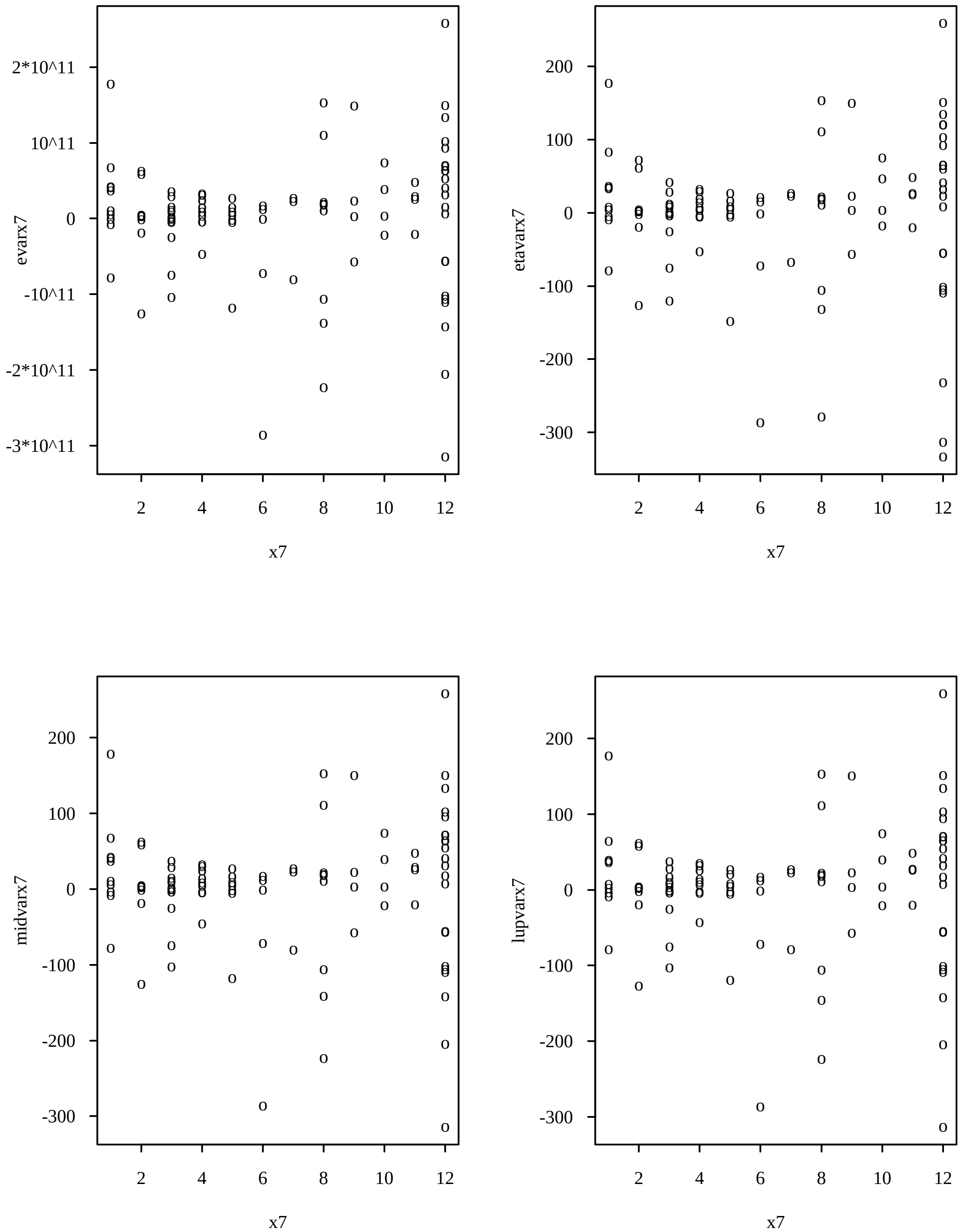




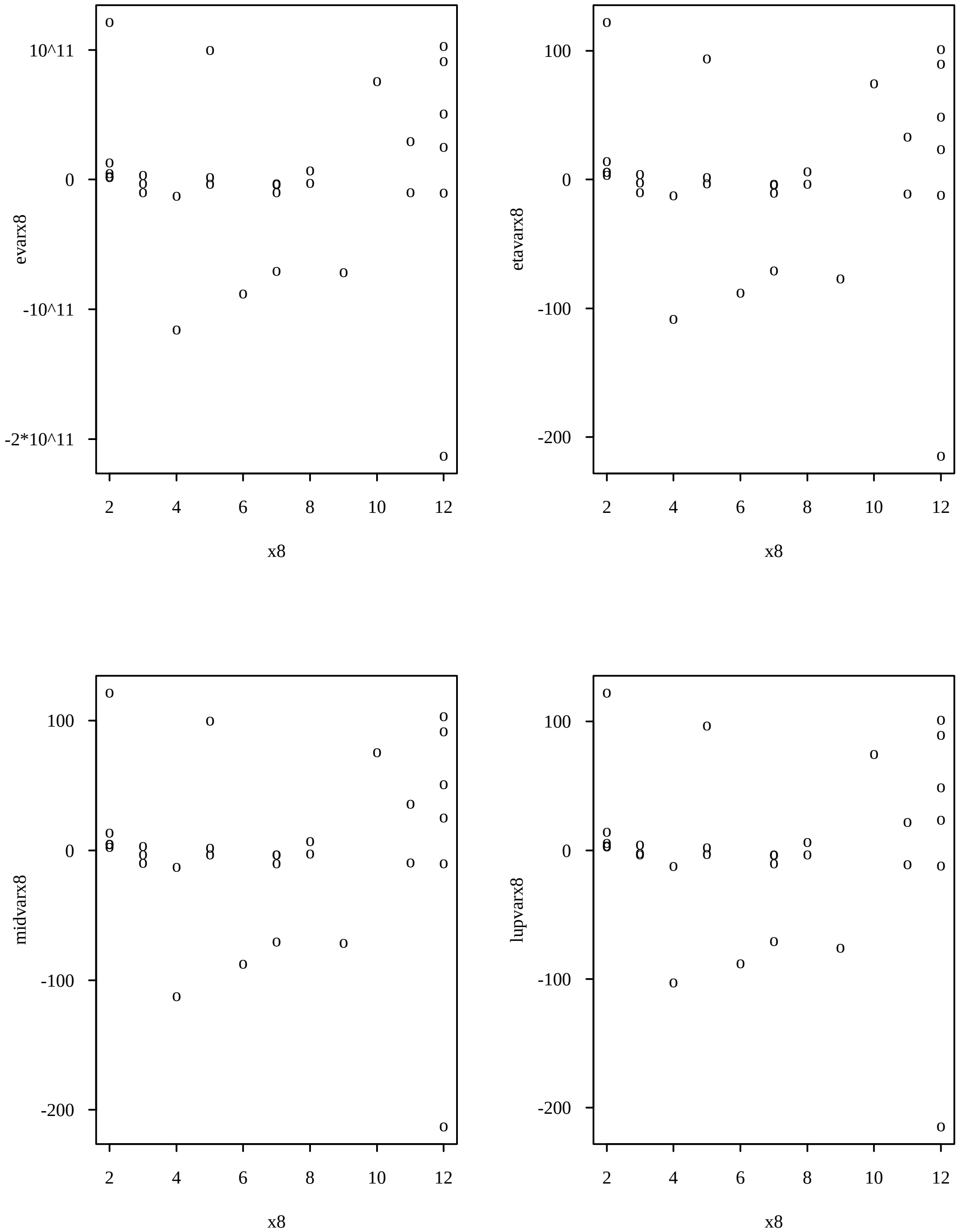




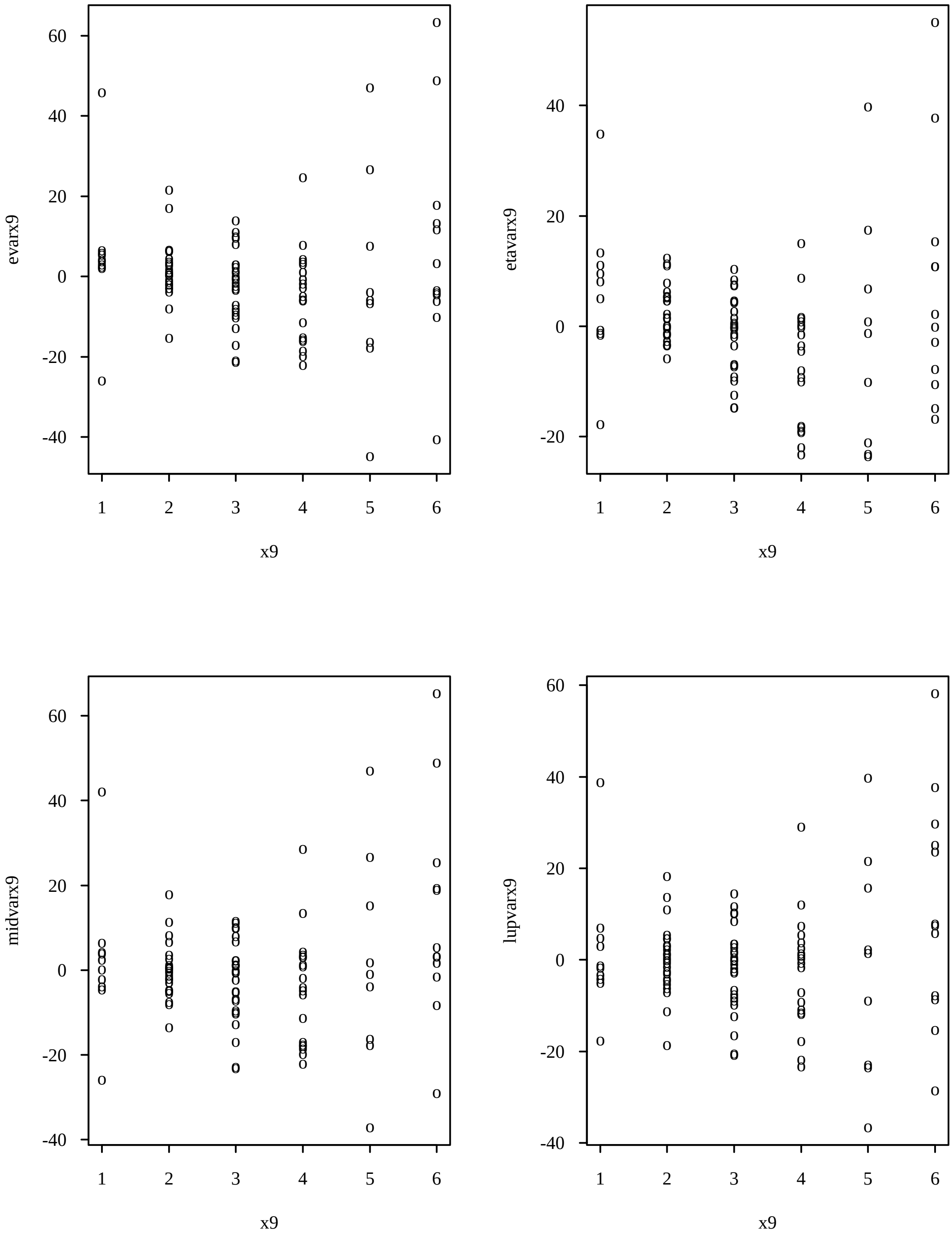
Scenario 7: covariate distribution Exp($\frac{1}{8}$), β=5, p=0.7, n=100:



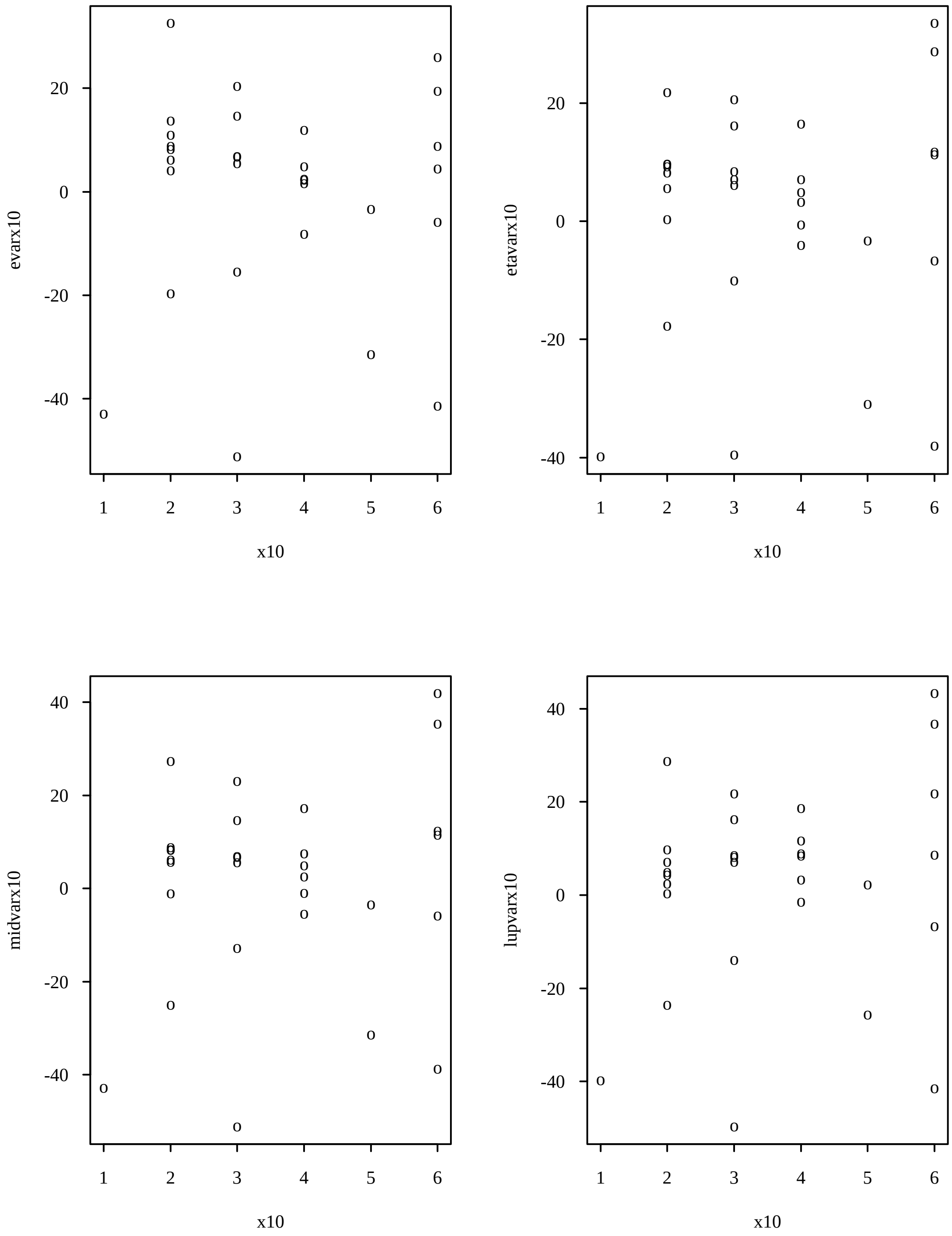
Scenario 8: covariate distribution $\mathrm{Exp}(\frac{1}{8})$, β=5, p=0.7, n=30:



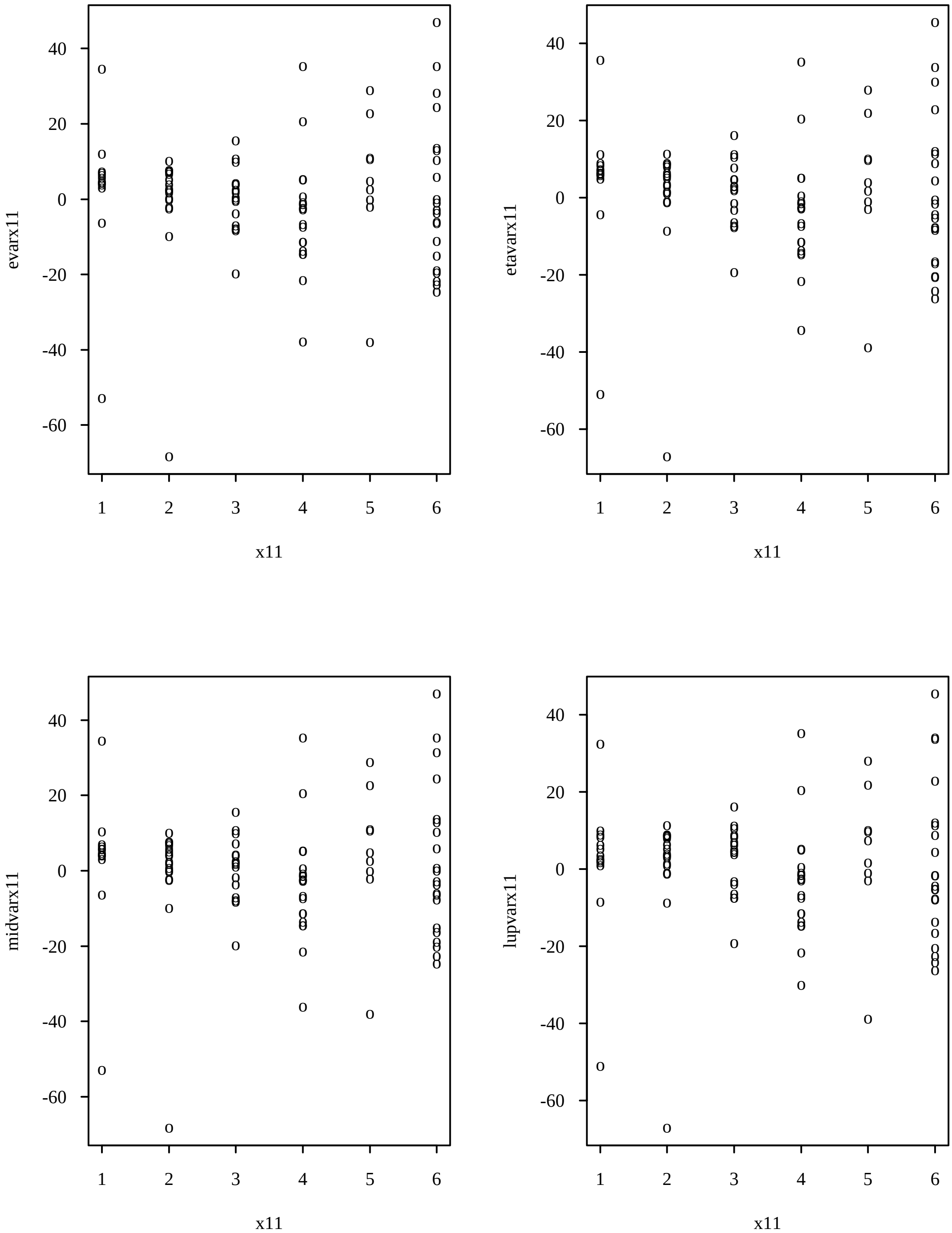
Scenario 9: covariate distribution $W(\frac{1}{6}, \frac{3}{2})$, β=2, p=0.3, n=100:



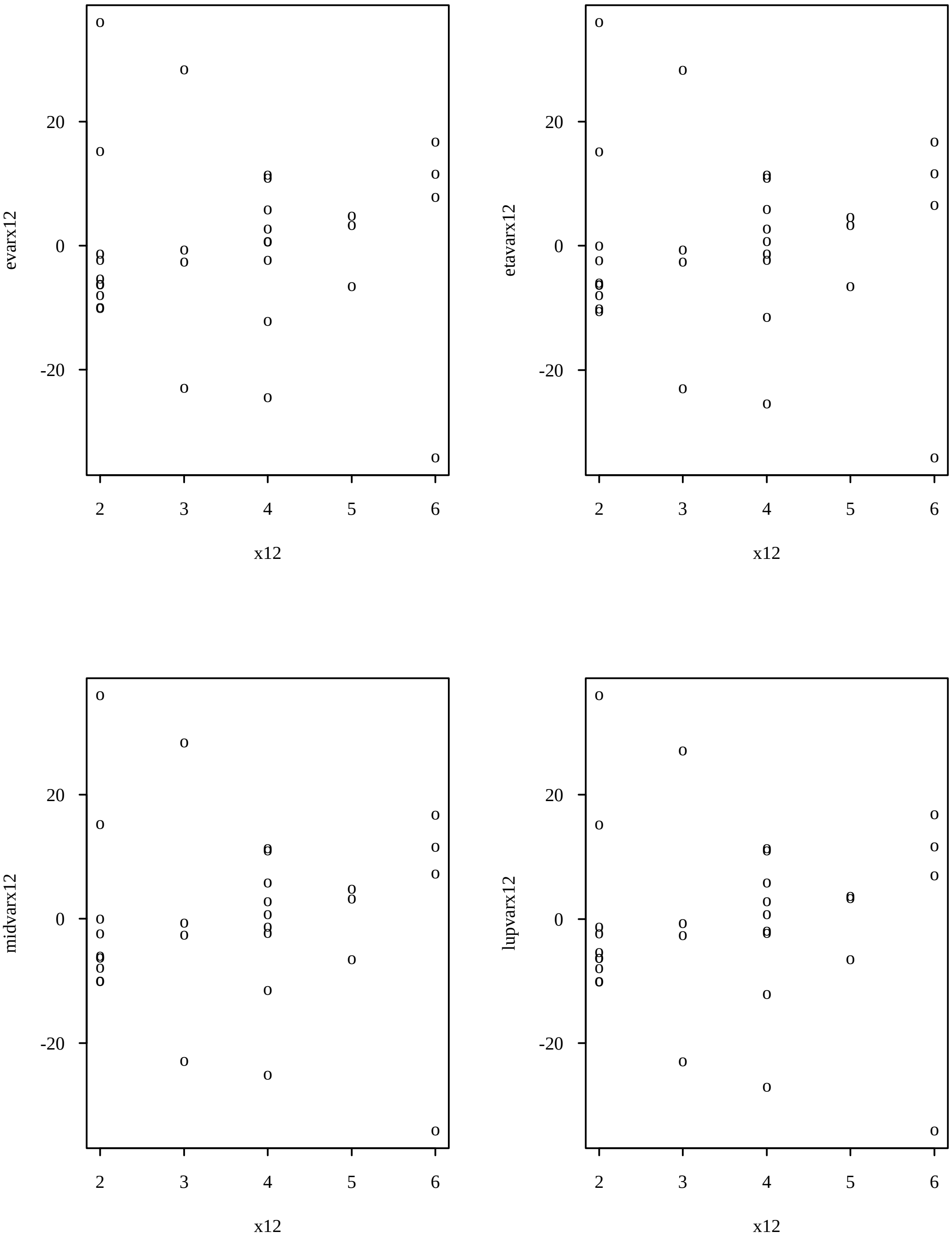
Scenario 10: covariate distribution $W(\frac{1}{6}, \frac{3}{2})$, β=2, p=0.3, n=30:



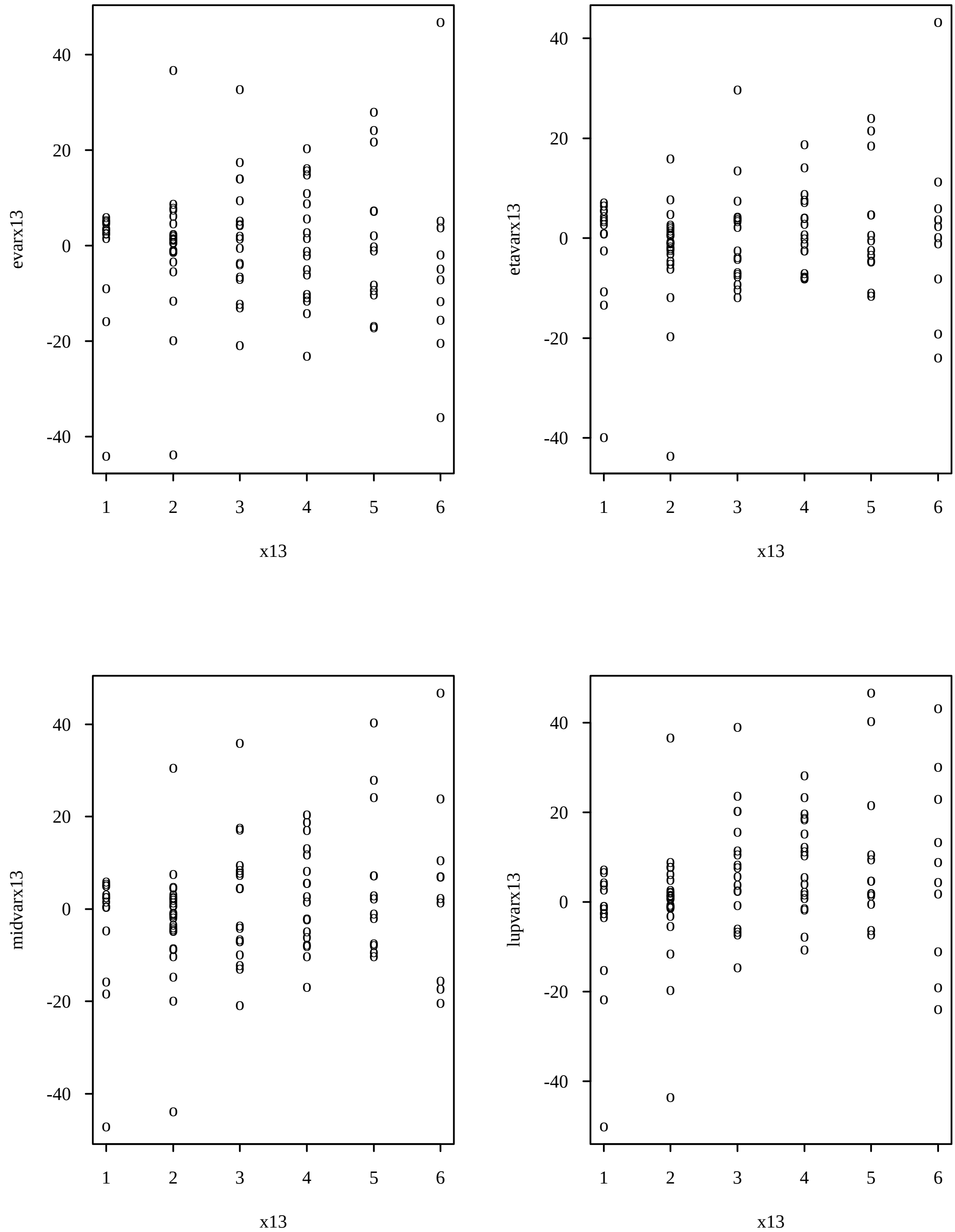
Scenario 11: covariate distribution $W(\frac{1}{6}, \frac{3}{2})$, β=2, p=0.7, n=100:



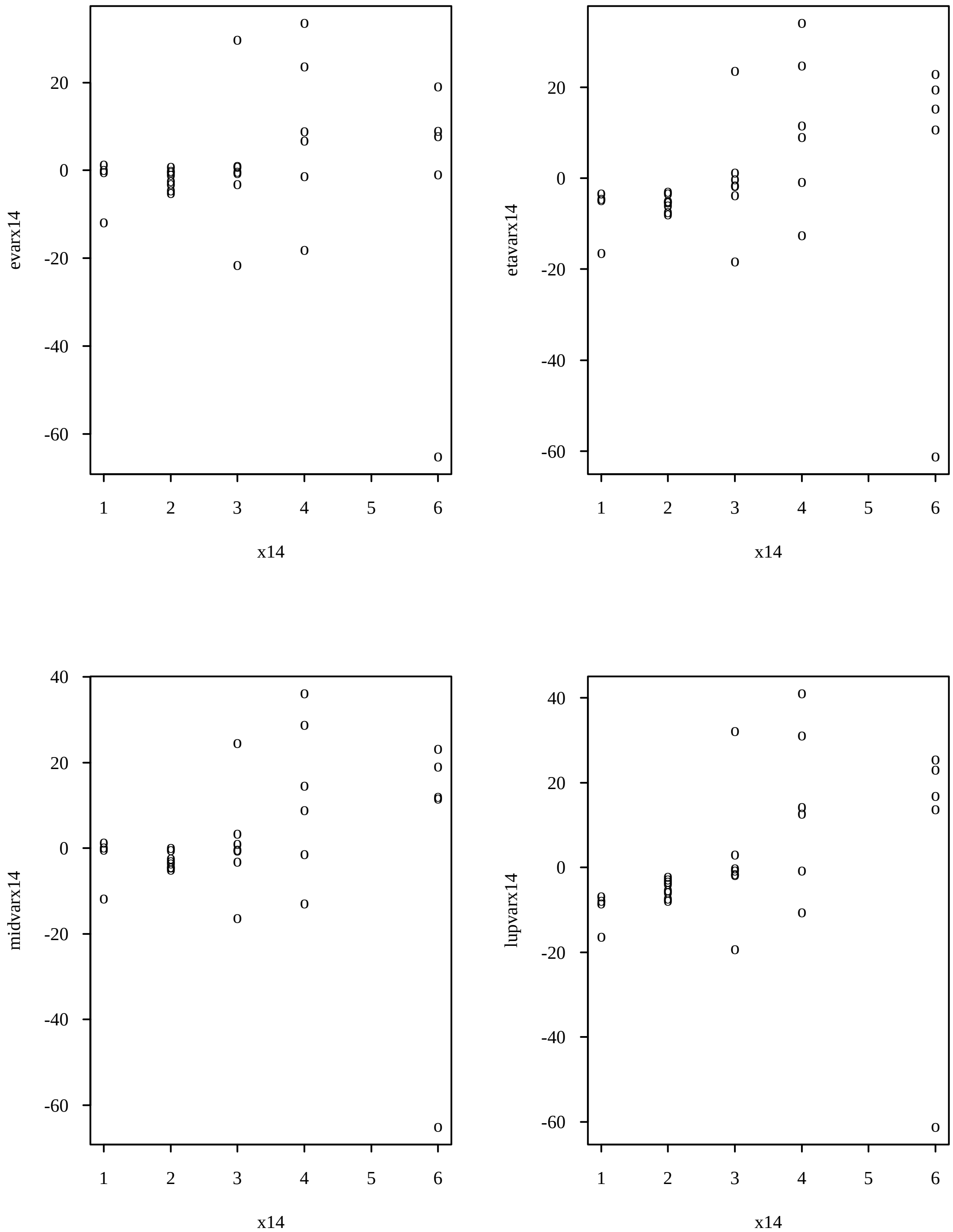
Scenario 12: covariate distribution $W(\frac{1}{6}, \frac{3}{2})$, β=2, p=0.7, n=30:



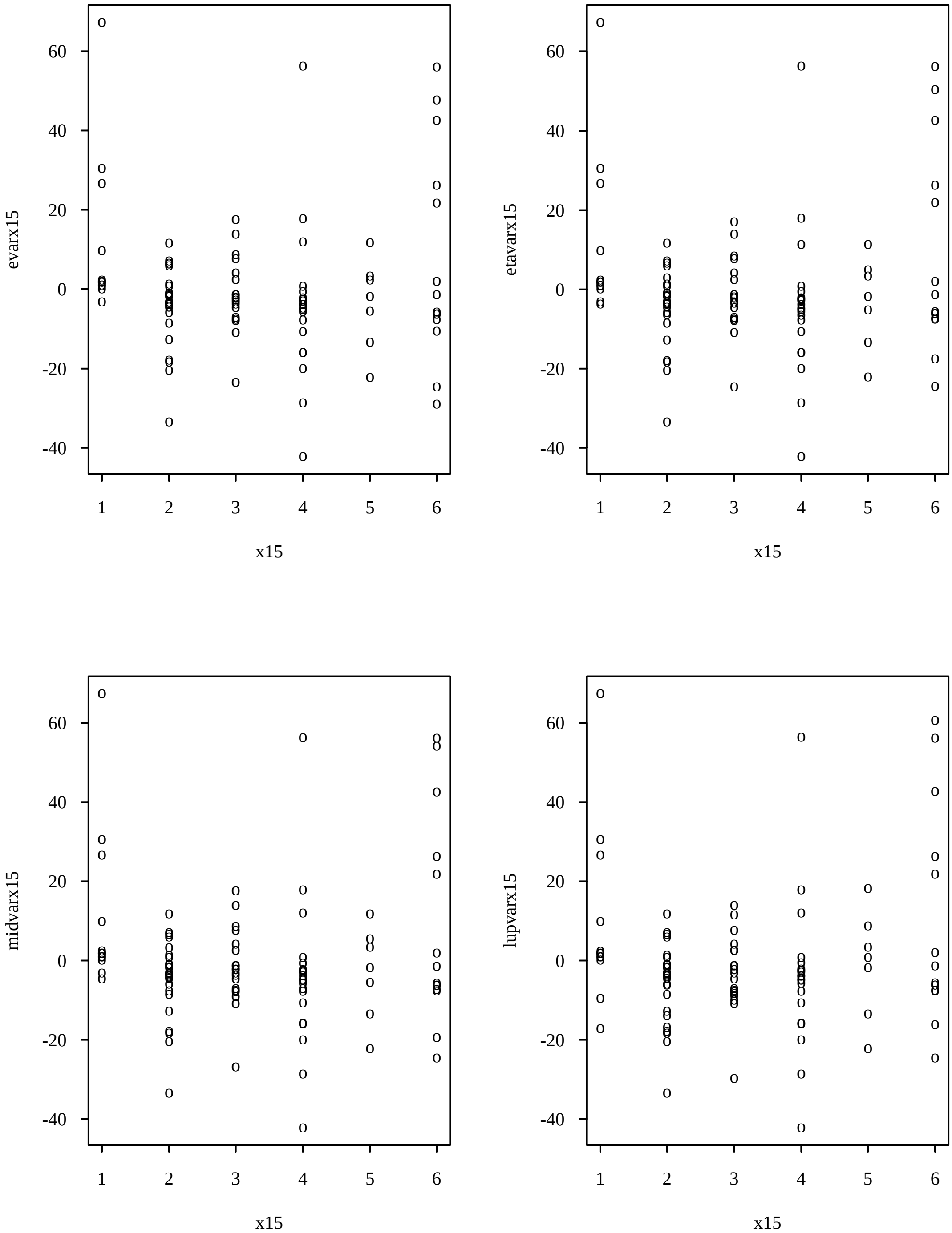
Scenario 13: covariate distribution W($\frac{1}{6}$, $\frac{3}{2}$), β=5, p=0.3, n=100:



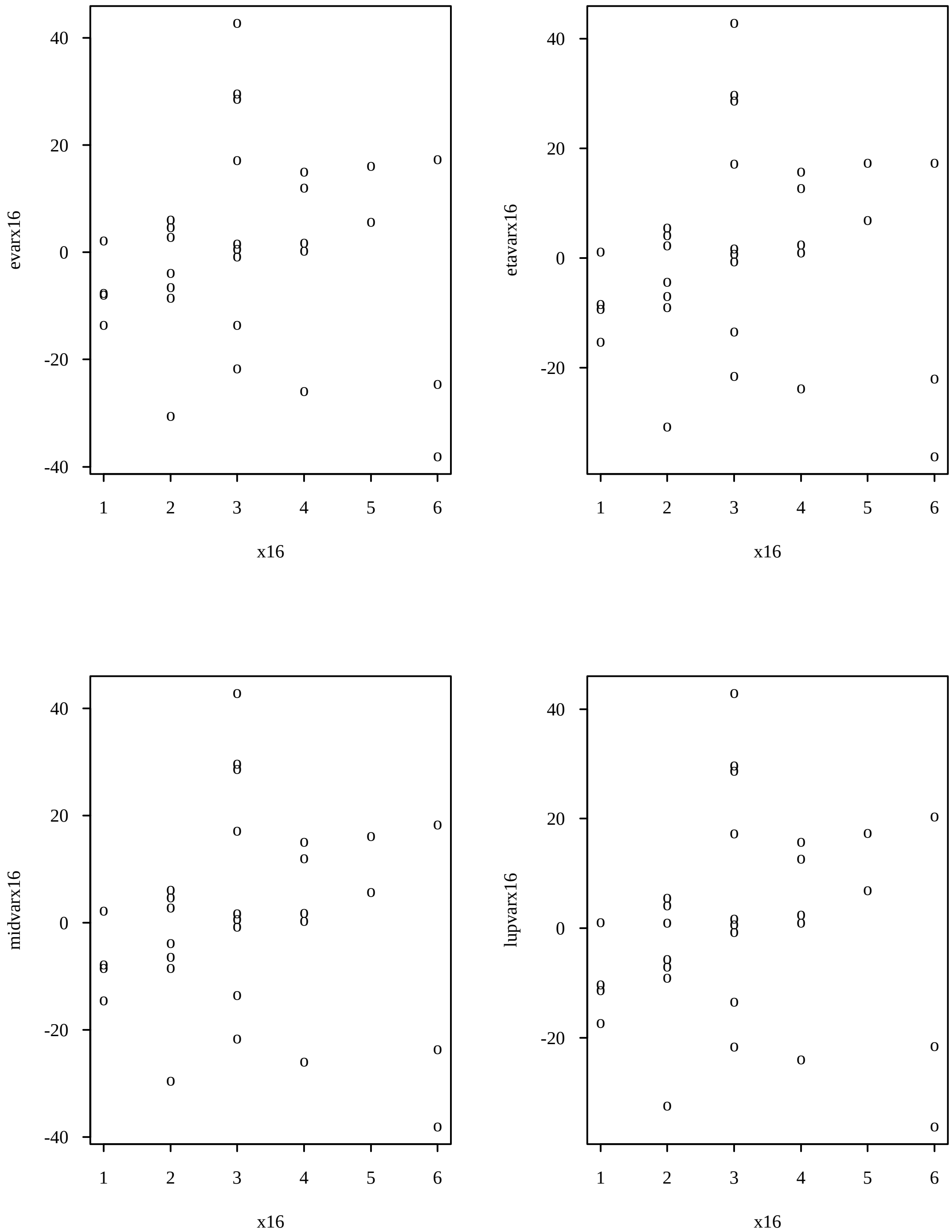
Scenario 14: covariate distribution $W(\frac{1}{6}, \frac{3}{2})$, β=5, p=0.3, n=30:



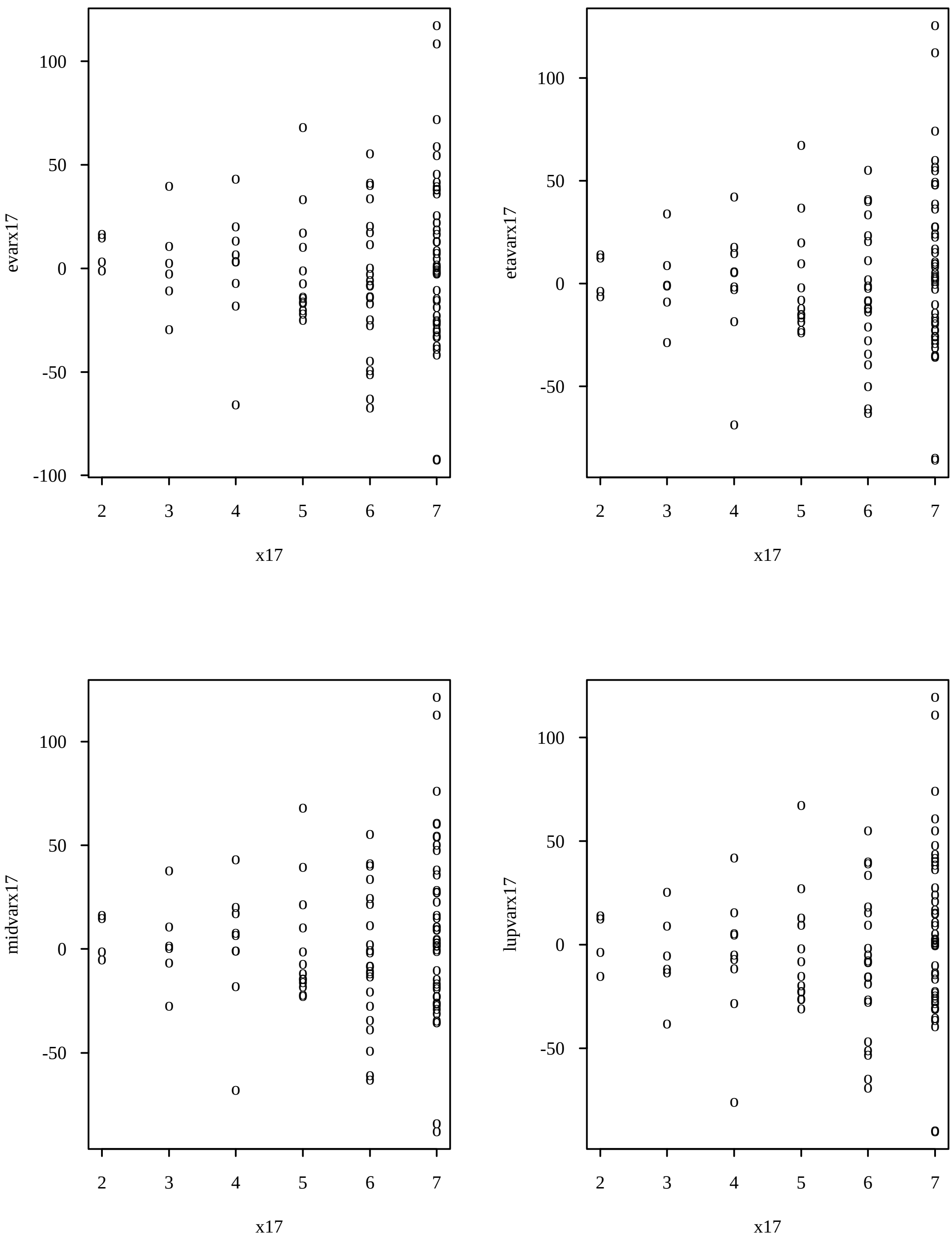
Scenario 15: covariate distribution $W(\frac{1}{6}, \frac{3}{2})$, β=5, p=0.7, n=100:



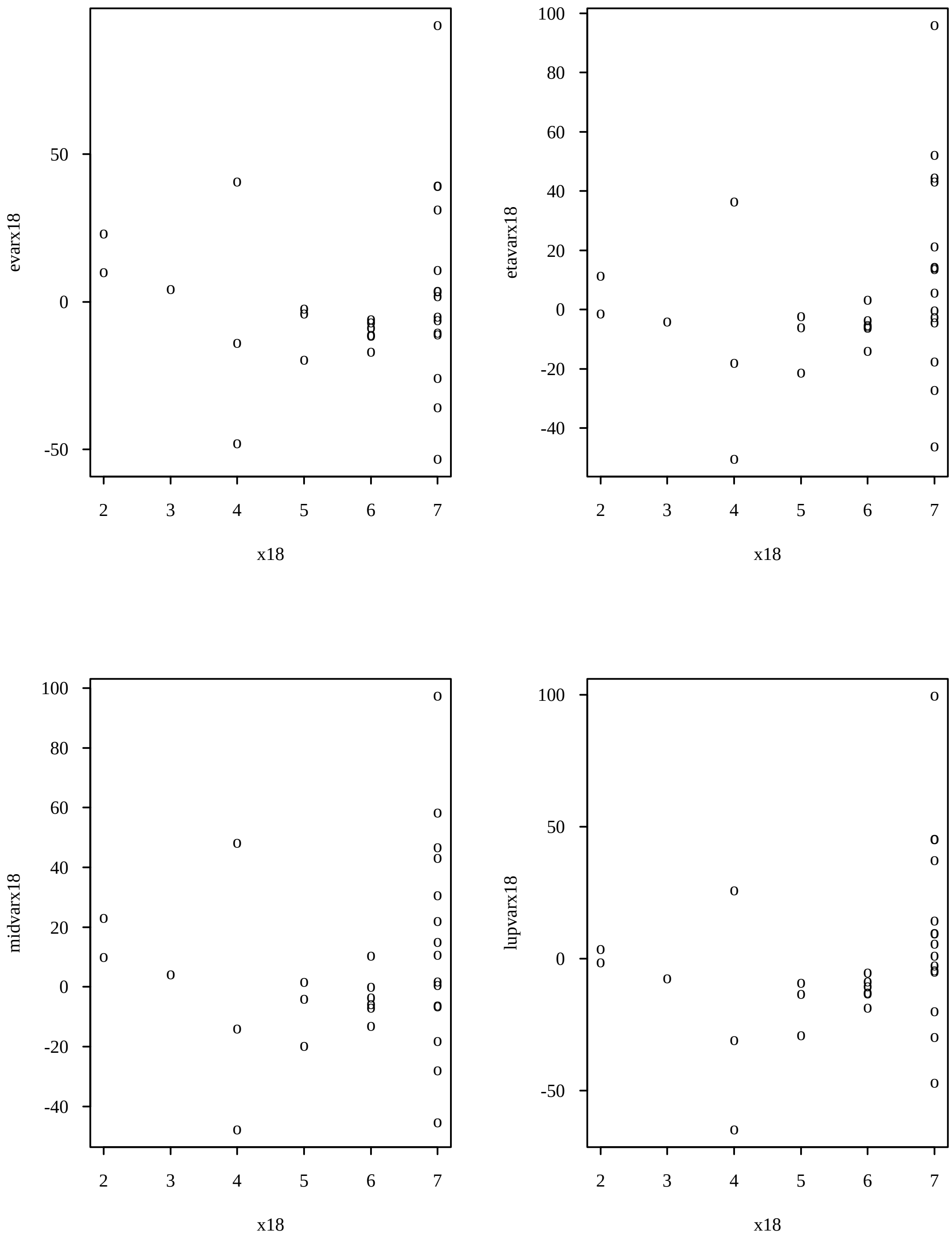
Scenario 16: covariate distribution $W(\frac{1}{6},\frac{3}{2})$, β=5, p=0.7, n=30:



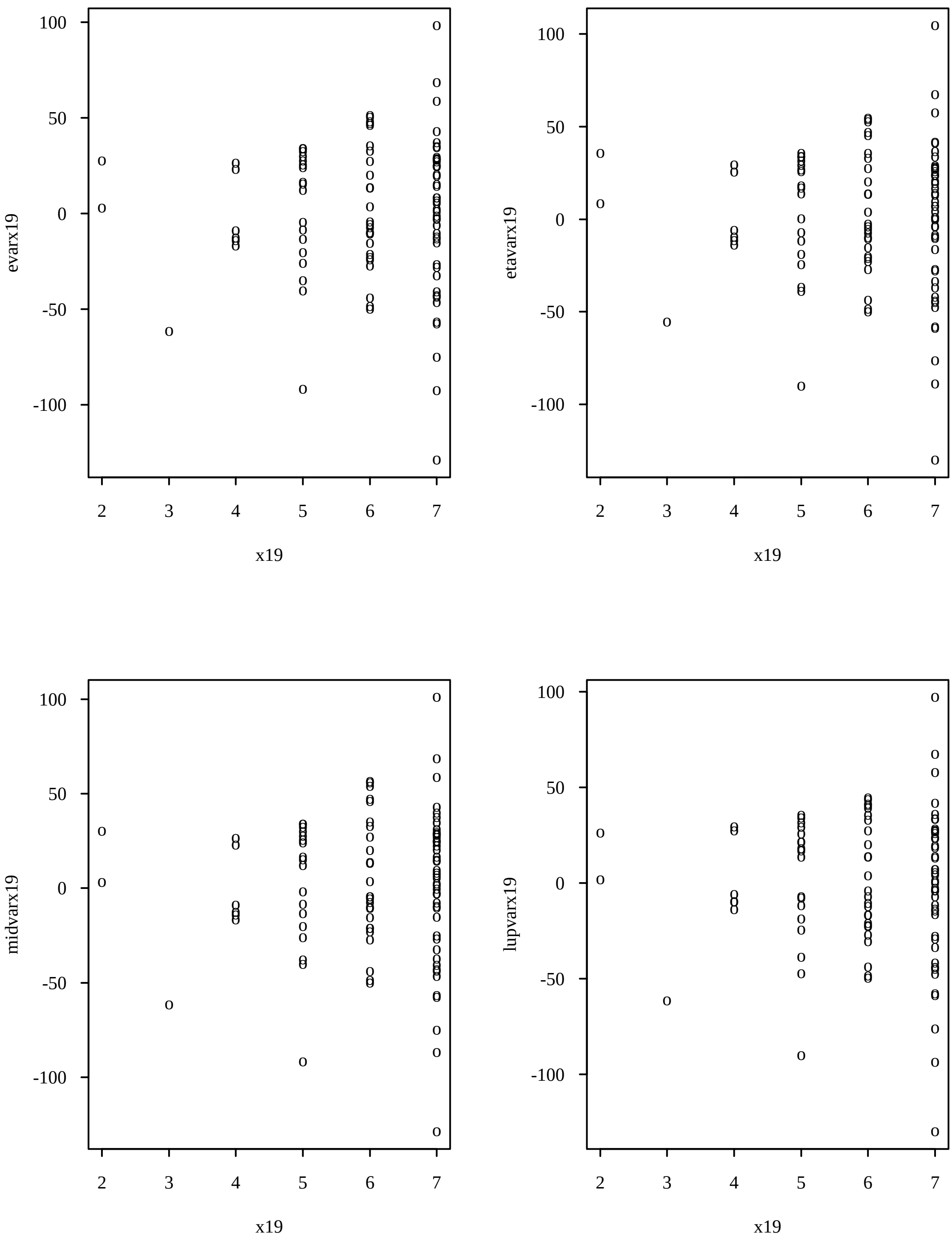
Scenario 17: covariate distribution N(6,4), β=2, p=0.3, n=100:



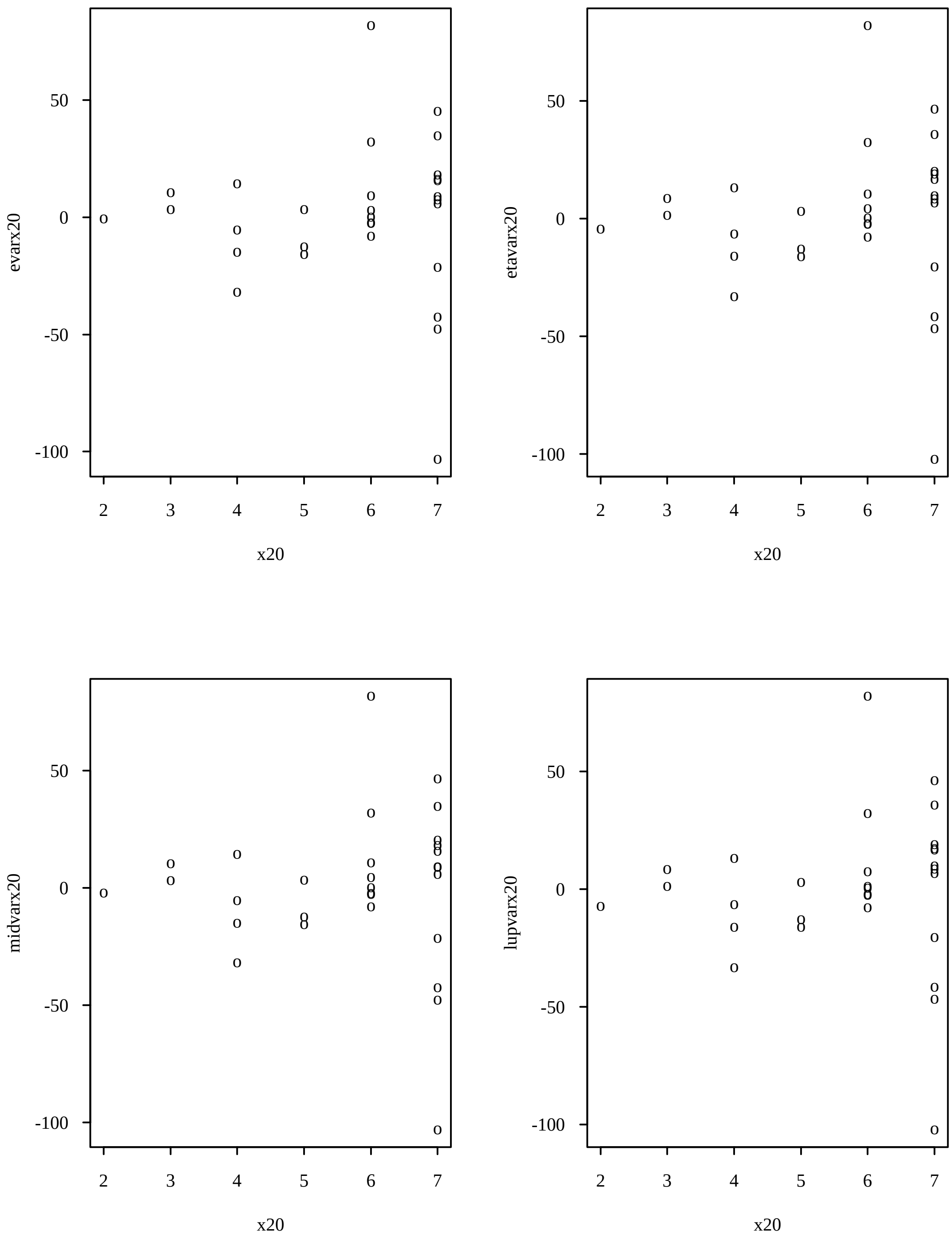
Scenario 18: covariate distribution N(6,4), β=2, p=0.3, n=30:



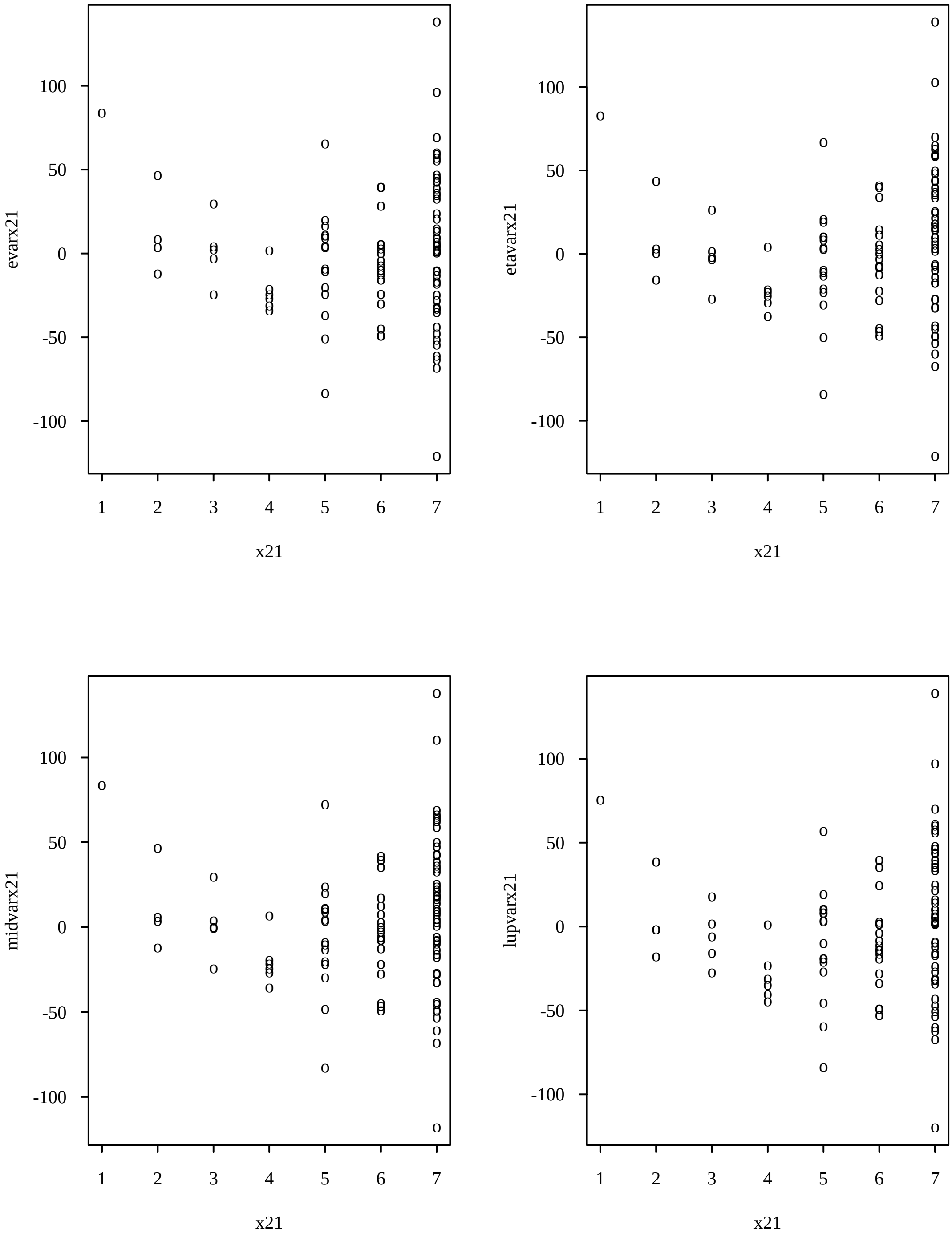
Scenario 19: covariate distribution N(6,4), β=2, p=0.7, n=100:



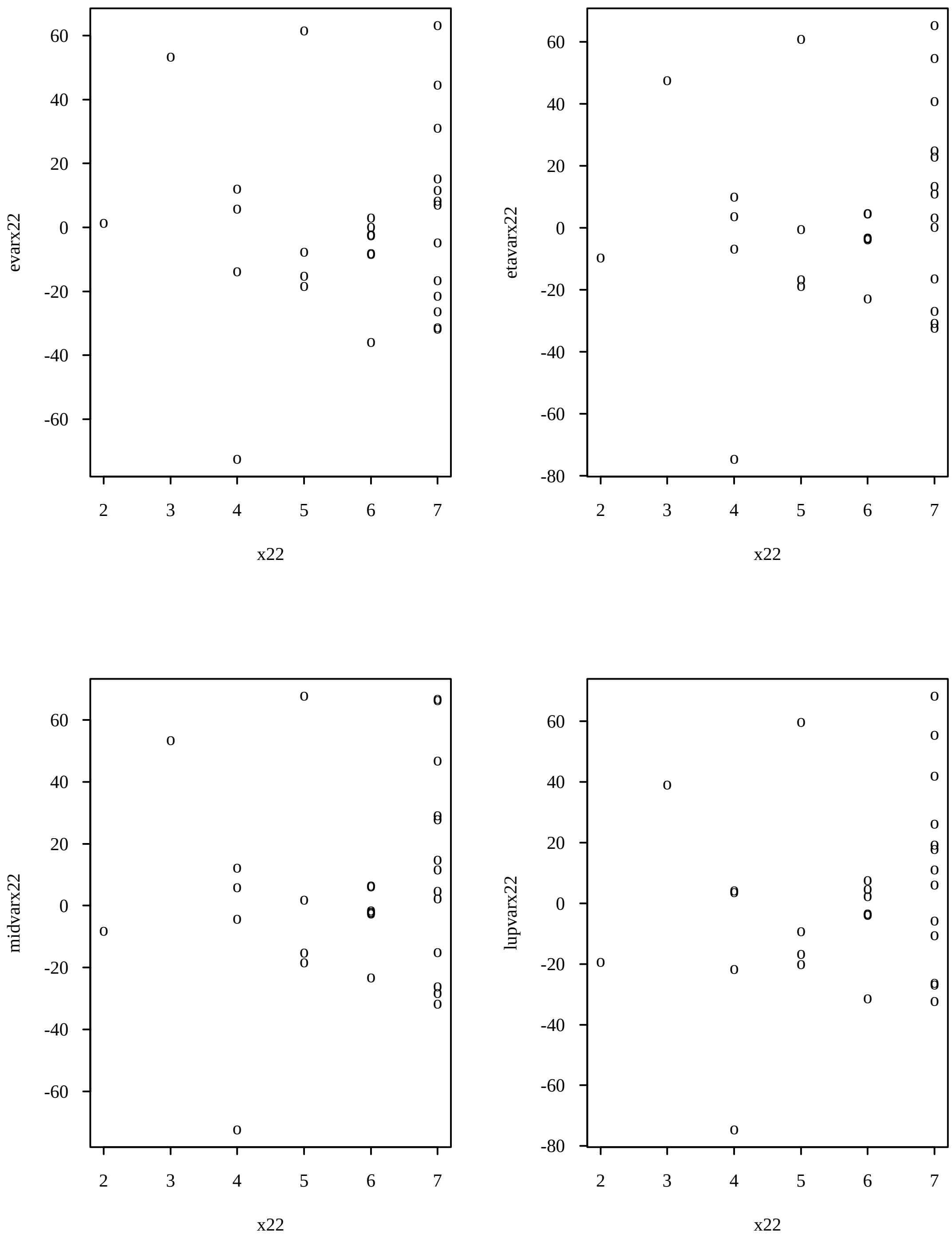
Scenario 20: covariate distribution N(6,4), β=2, p=0.7, n=30:



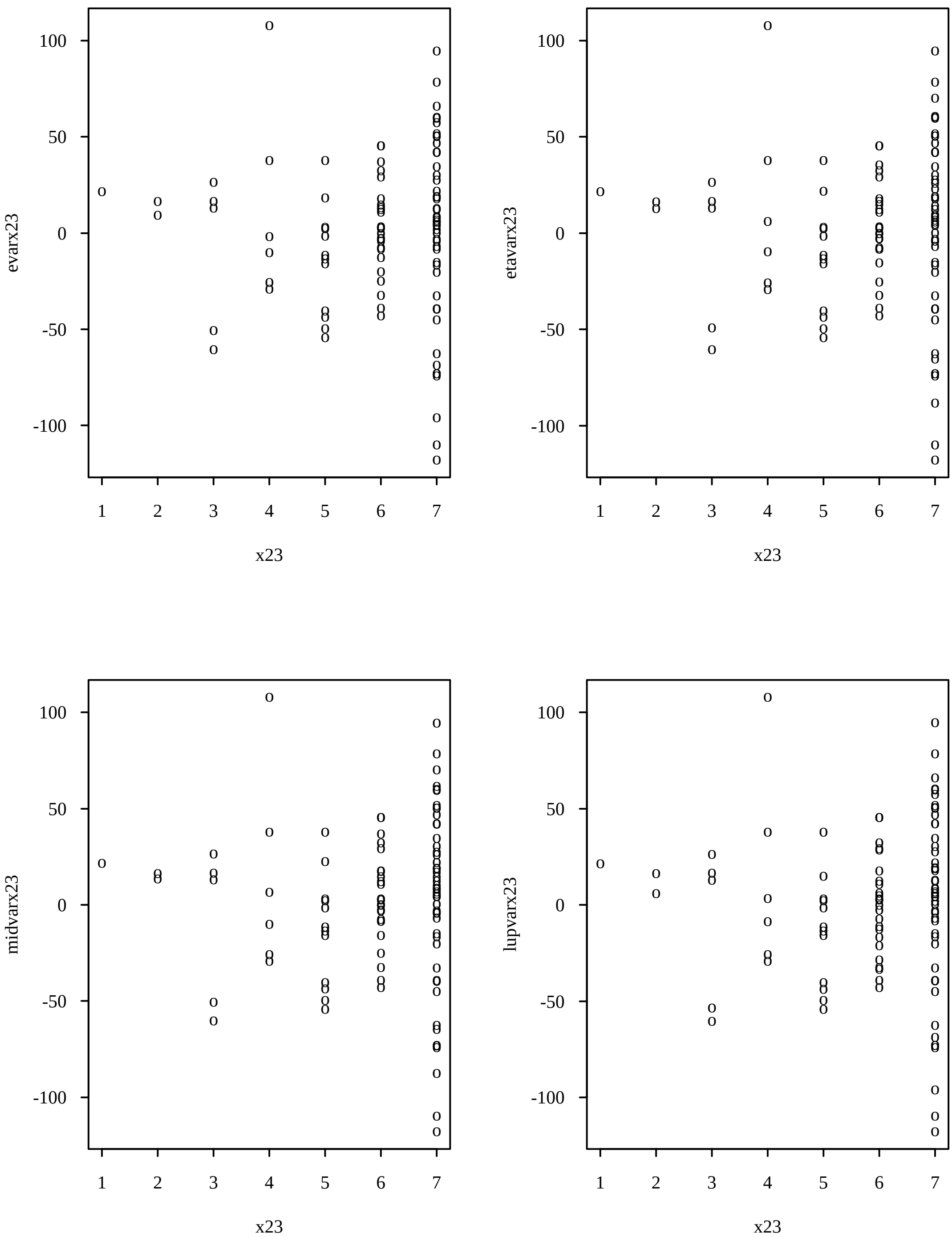
Scenario 21: covariate distribution N(6,4), β=5, p=0.3, n=100:



Scenario 22: covariate distribution N(6,4), β=5, p=0.3, n=30:



Scenario 23: covariate distribution N(6,4), β=5, p=0.7, n=100:



Scenario 24: covariate distribution N(6,4), β=5, p=0.7, n=30: