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

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

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

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

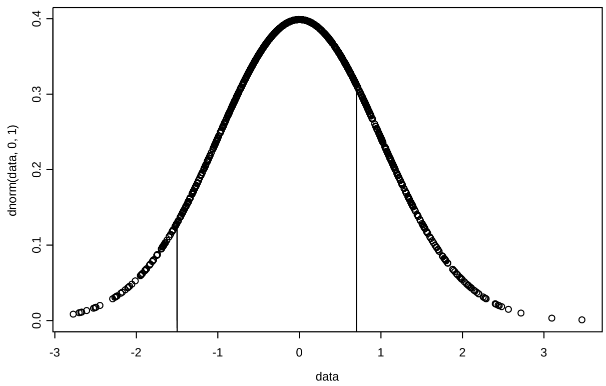
$$[y_i - \alpha - \beta z_{R_i}, y_i - \alpha - \beta z_{L_i}] = [A_i, B_i] \quad \text{if} \quad \beta > 0. \tag{2.3}$$

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

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

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:

$$\varepsilon_i^* = \delta_i \varepsilon_i + (1 - \delta_i)\eta_i, \tag{2.5}$$

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

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

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

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

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

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

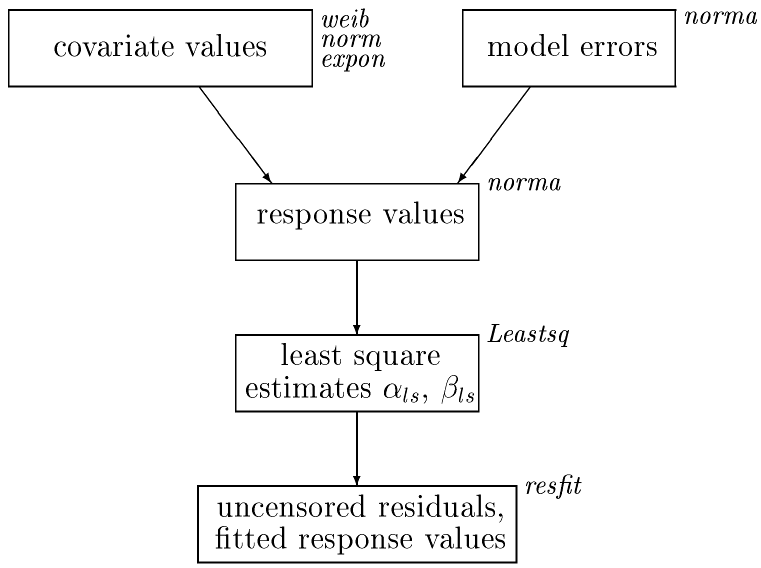
The application of test procedures for normality mentioned in Chapter 1 requires independent observations. But neither the OLS residuals nor the three other residuals to be examined are independent. Thus, for checking the normality of the residuals, the measures skewness and kurtosis are used. It is 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, $i = 1, \ldots, n$, are (Hartung et al., 1993, p.48f):

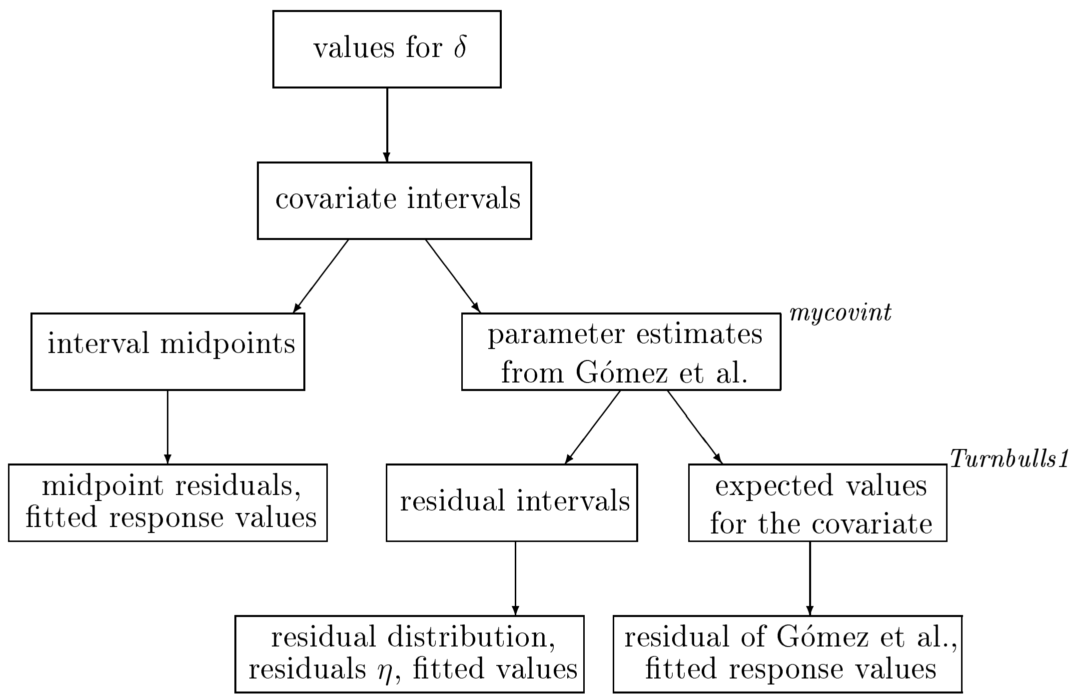
$$S = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\sqrt{(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2)^3}}, \qquad K = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2)^2} - 3.$$

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 $\hat{e}_{lup}$

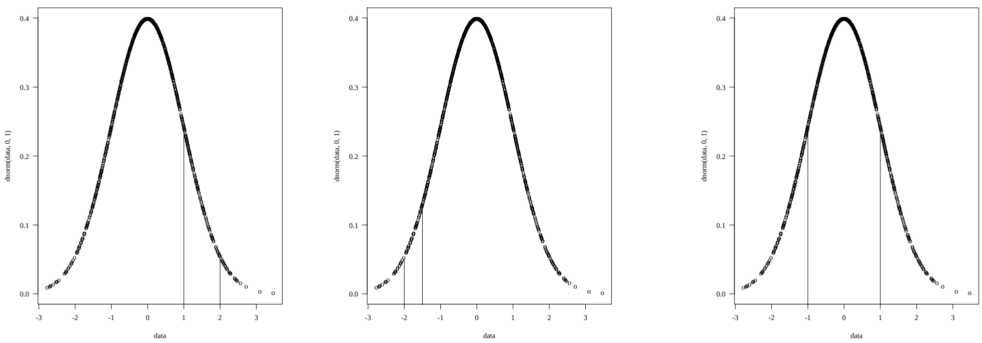
	Skewness		**Kurtosis**	
	Median	**Mean [Std]**	**Median**	**Mean [Std]**
Exponential($\frac{1}{8}$)				
n=100,p=0.3,$\beta = 2$	1.046	0.732 [1.20]	3.013	3.511 [2.37]
n=500,p=0.3,$\beta = 2$	1.372	0.983 [1.17]	4.361	4.609 [1.69]
n=100,p=0.7,$\beta = 2$	-0.141	-0.169 [0.89]	1.649	2.818 [3.92]
n=500,p=0.7,$\beta = 2$	-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.

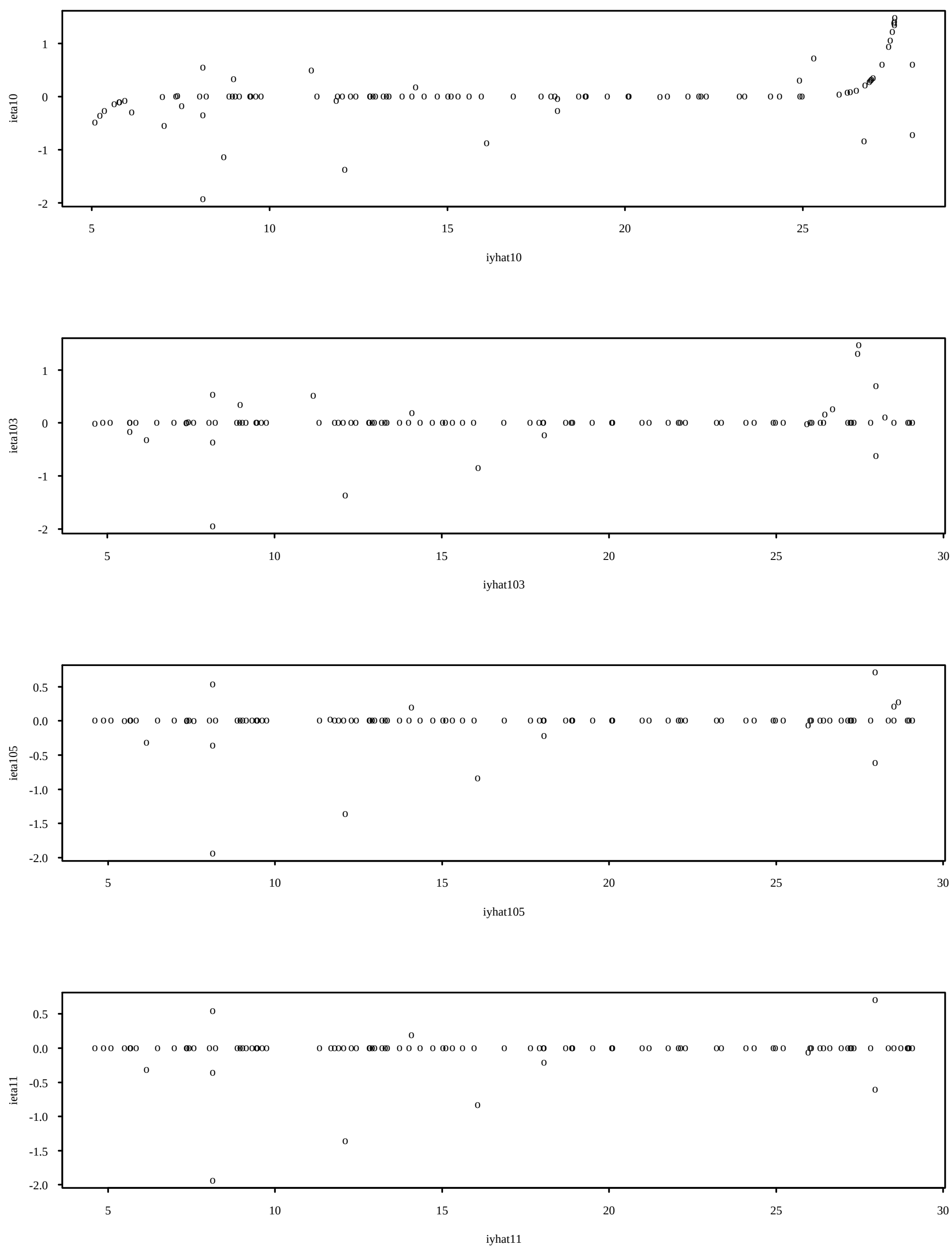


Figure 2: Correctly specified model, censoring level 0.3. First graph: original residual interval width, second graph: 0.3 times increased, third graph: 0.5 times increased, fourth graph: double of the original interval width.

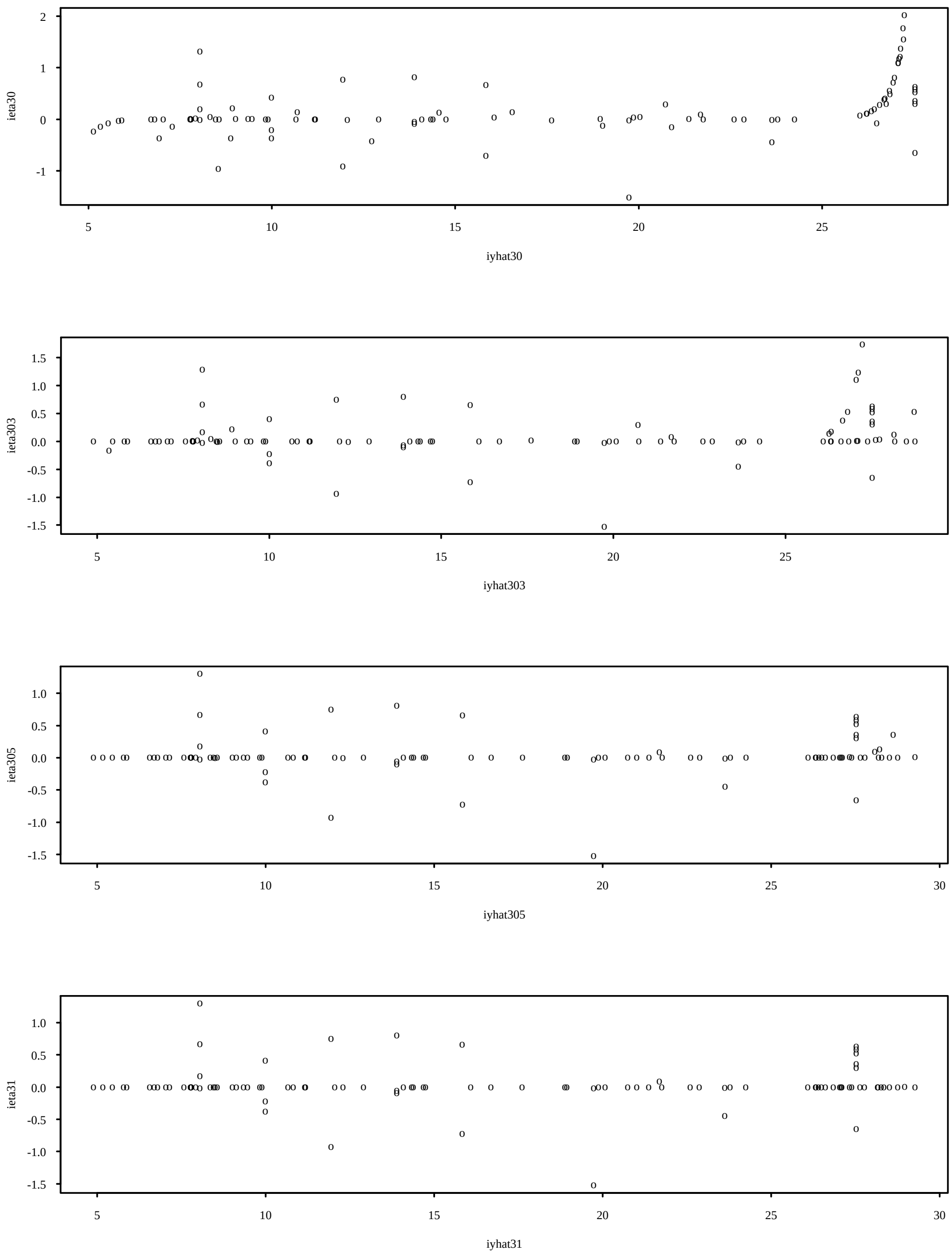


Figure 3: Correctly specified model, censoring level 0.5. First graph: original residual interval width, second graph: 0.3 times increased, third graph: 0.5 times increased, fourth graph: double of the original interval width.

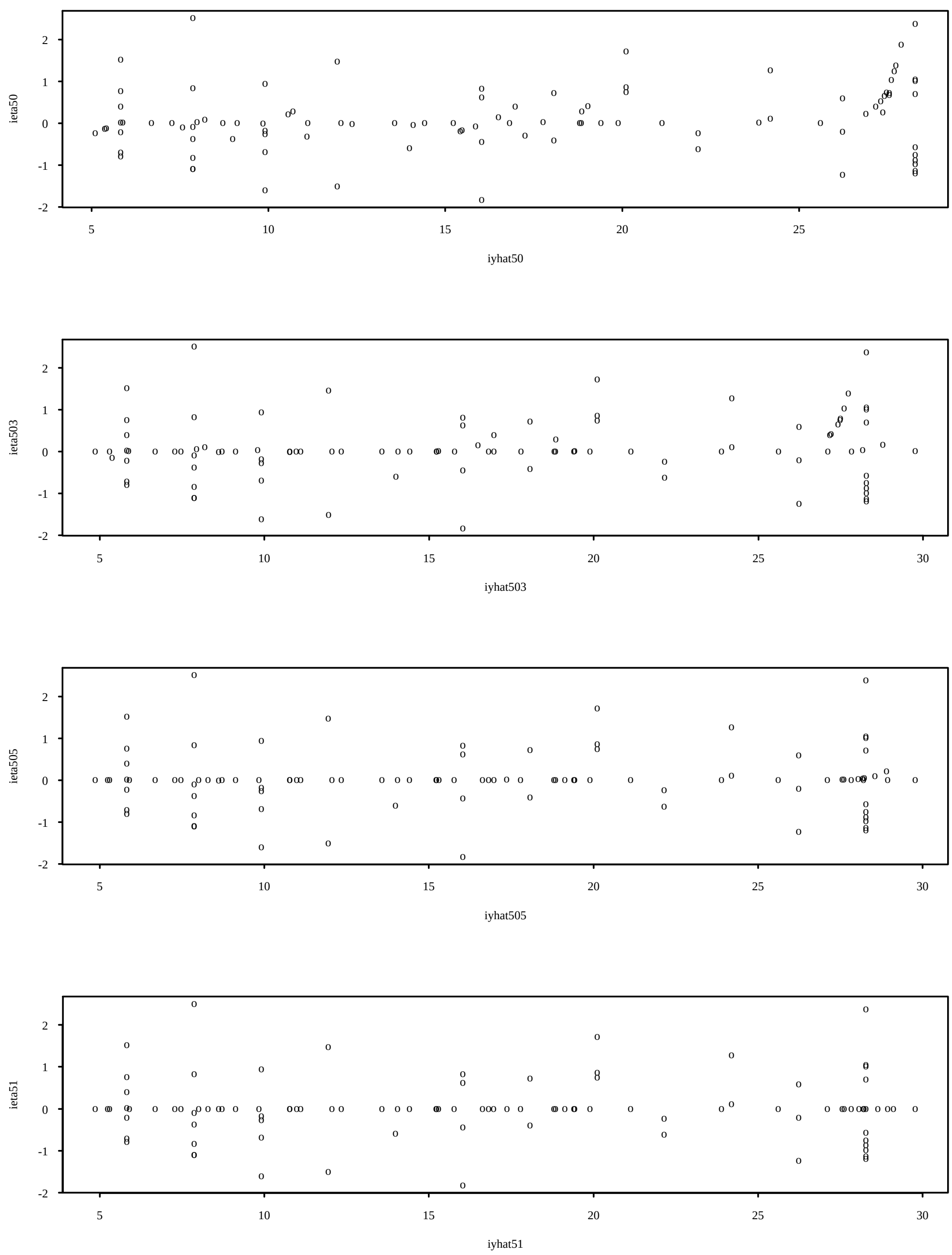


Figure 4: Correctly specified model, censoring level 0.7. First graph: original residual interval width, second graph: 0.3 times increased, third graph: 0.5 times increased, fourth graph: double of the original interval width.

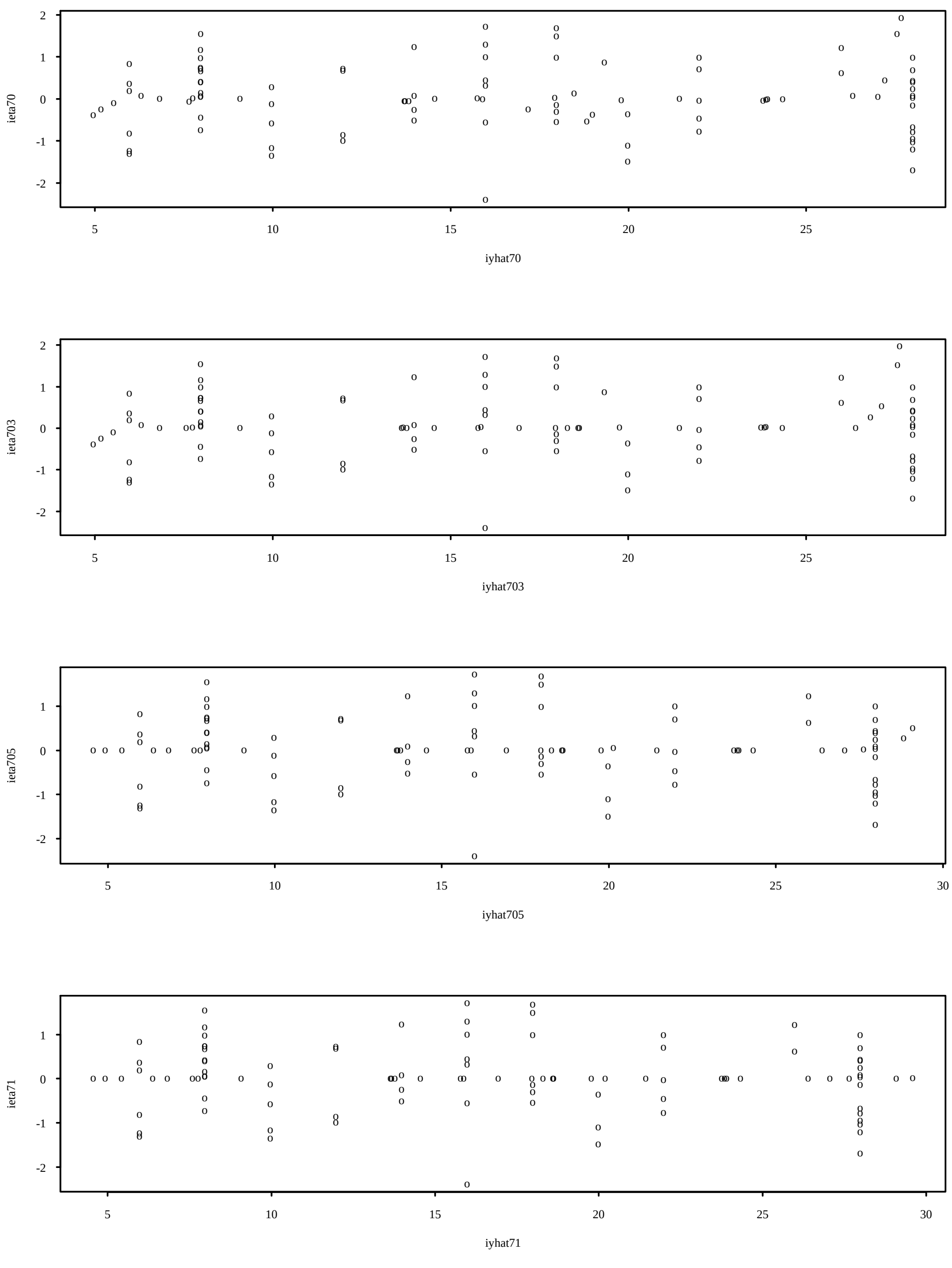


Figure 5: Correctly specified model, censoring level 0.9. First graph: original residual interval width, second graph: 0.3 times increased, third graph: 0.5 times increased, fourth graph: double of the original interval width.

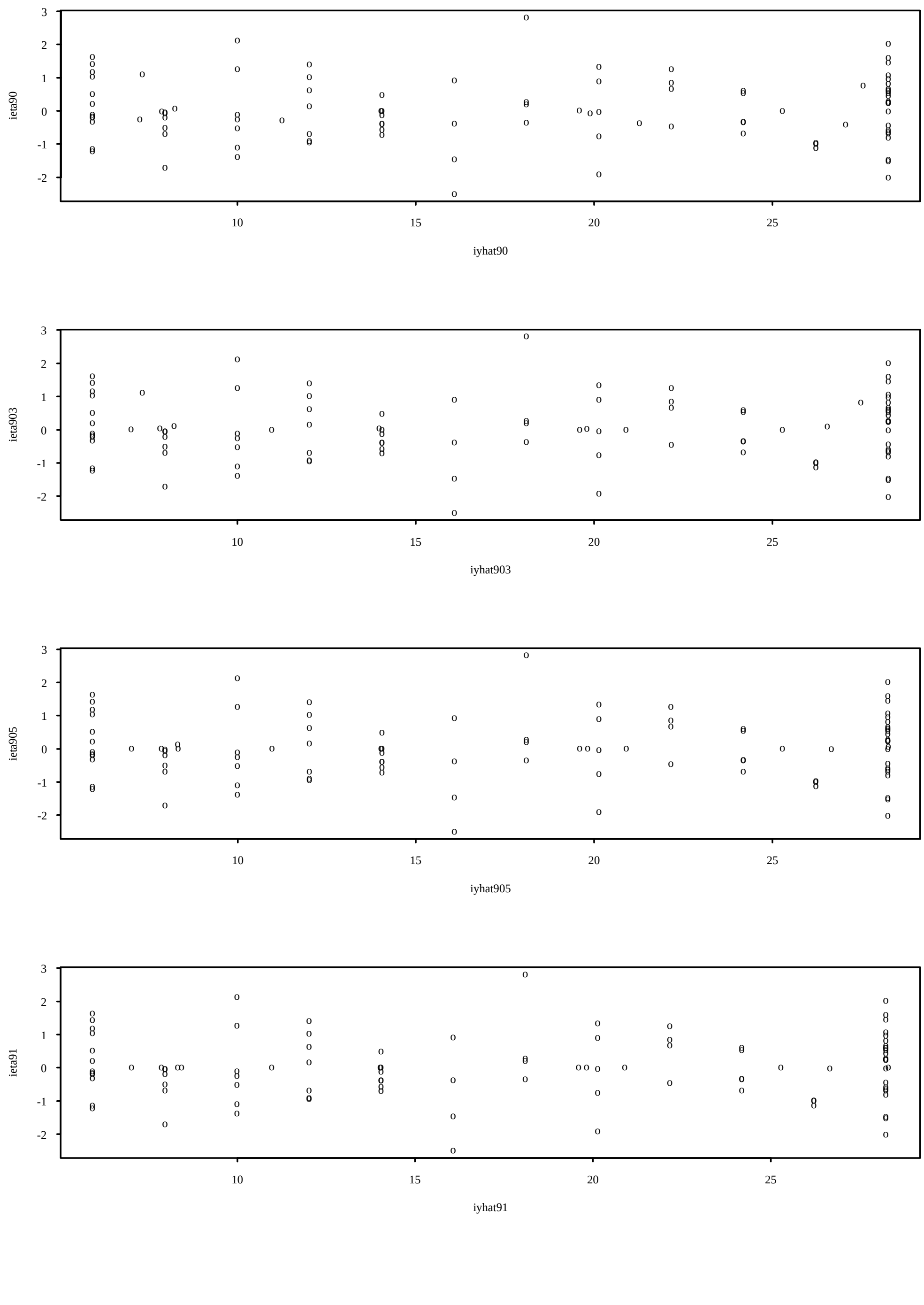


Figure 6: Quadratic model, censoring level 0.1. First graph: original residual interval width, second graph: 0.3 times increased, third graph: 0.5 times increased, fourth graph: double of the original interval width.

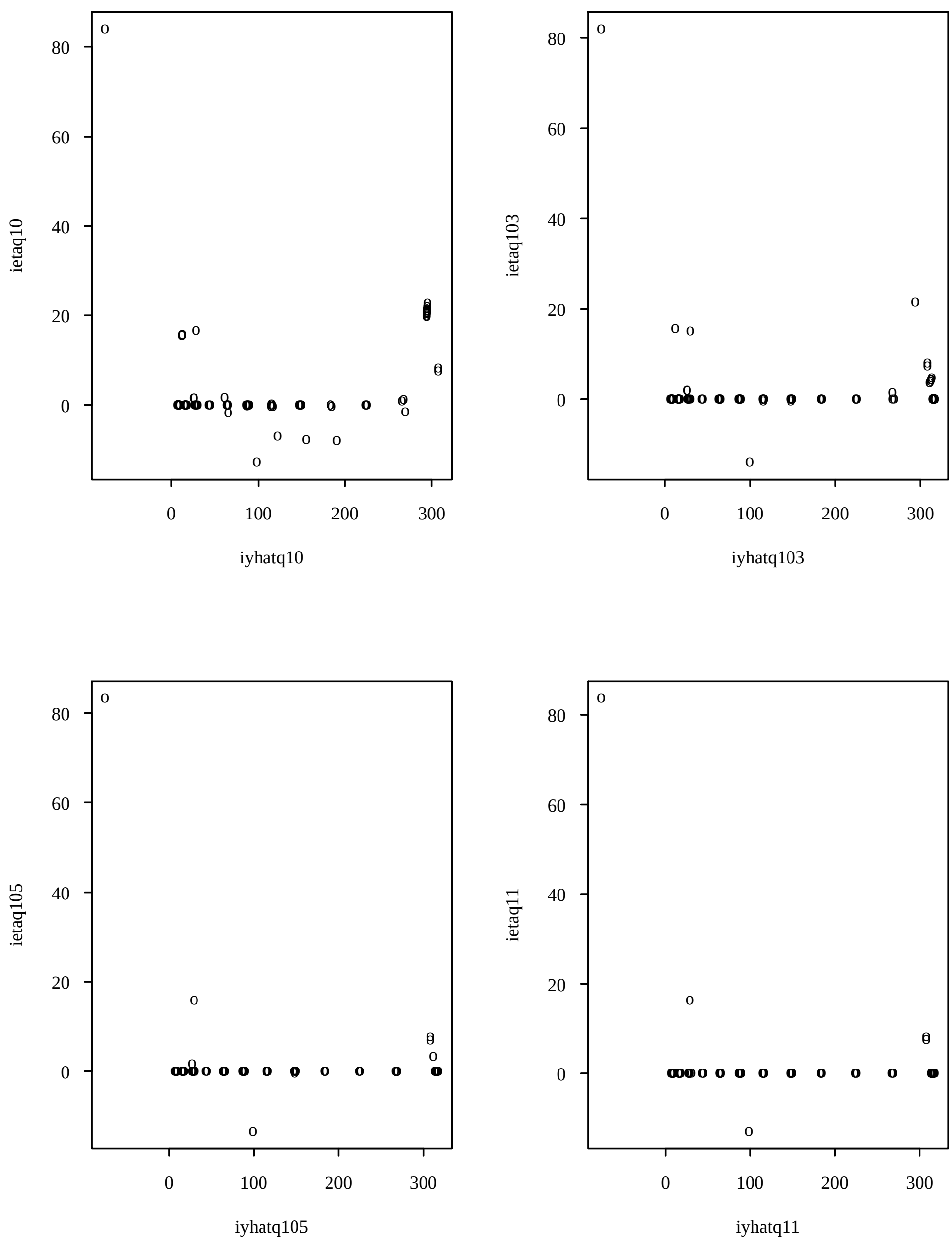


Figure 7: Quadratic model, censoring level 0.3. First graph: original residual interval width, second graph: 0.3 times increased, third graph: 0.5 times increased, fourth graph: double of the original interval width.

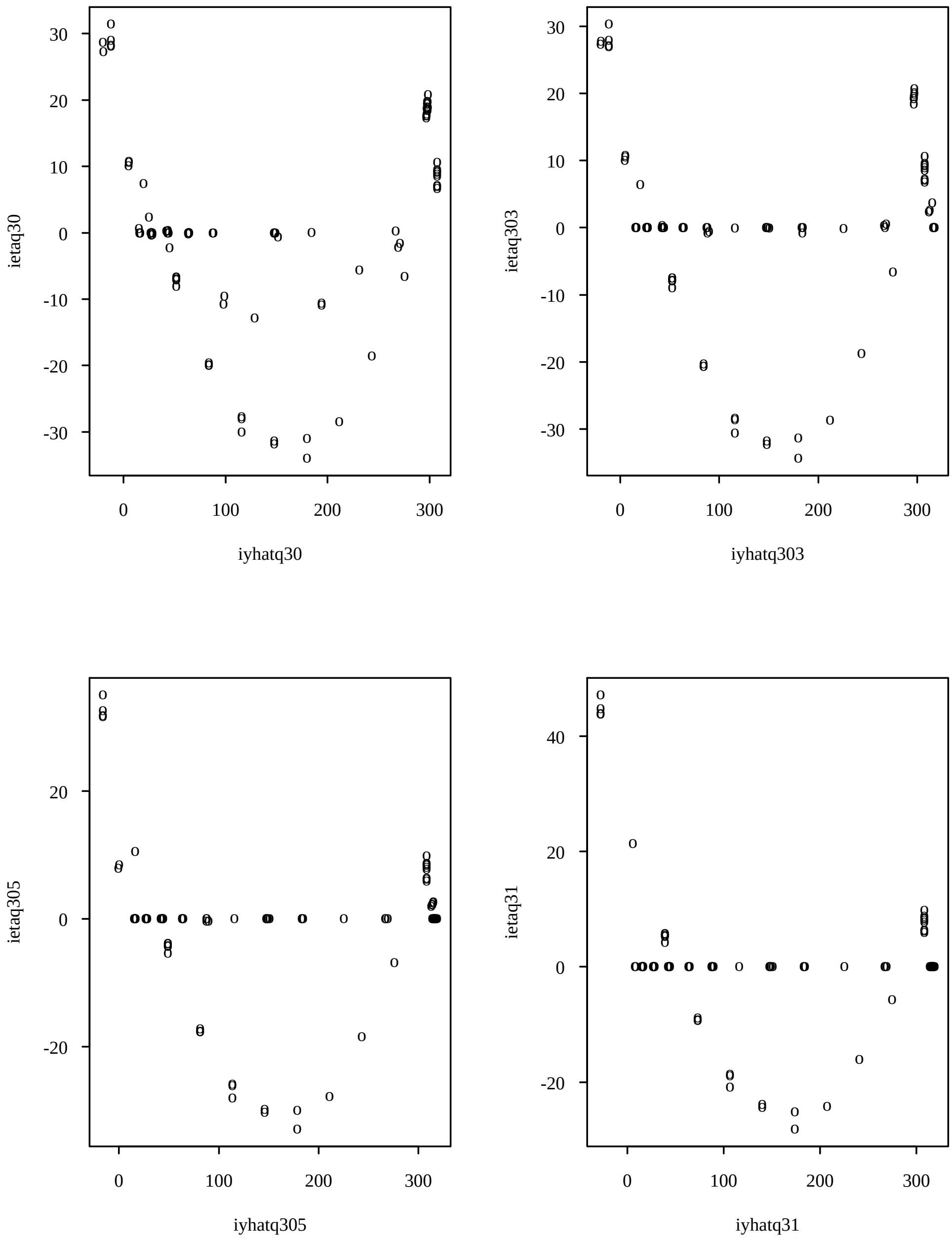


Figure 8: Quadratic model, censoring level 0.5. First graph: original residual interval width, second graph: 0.3 times increased, third graph: 0.5 times increased, fourth graph: double of the original interval width.

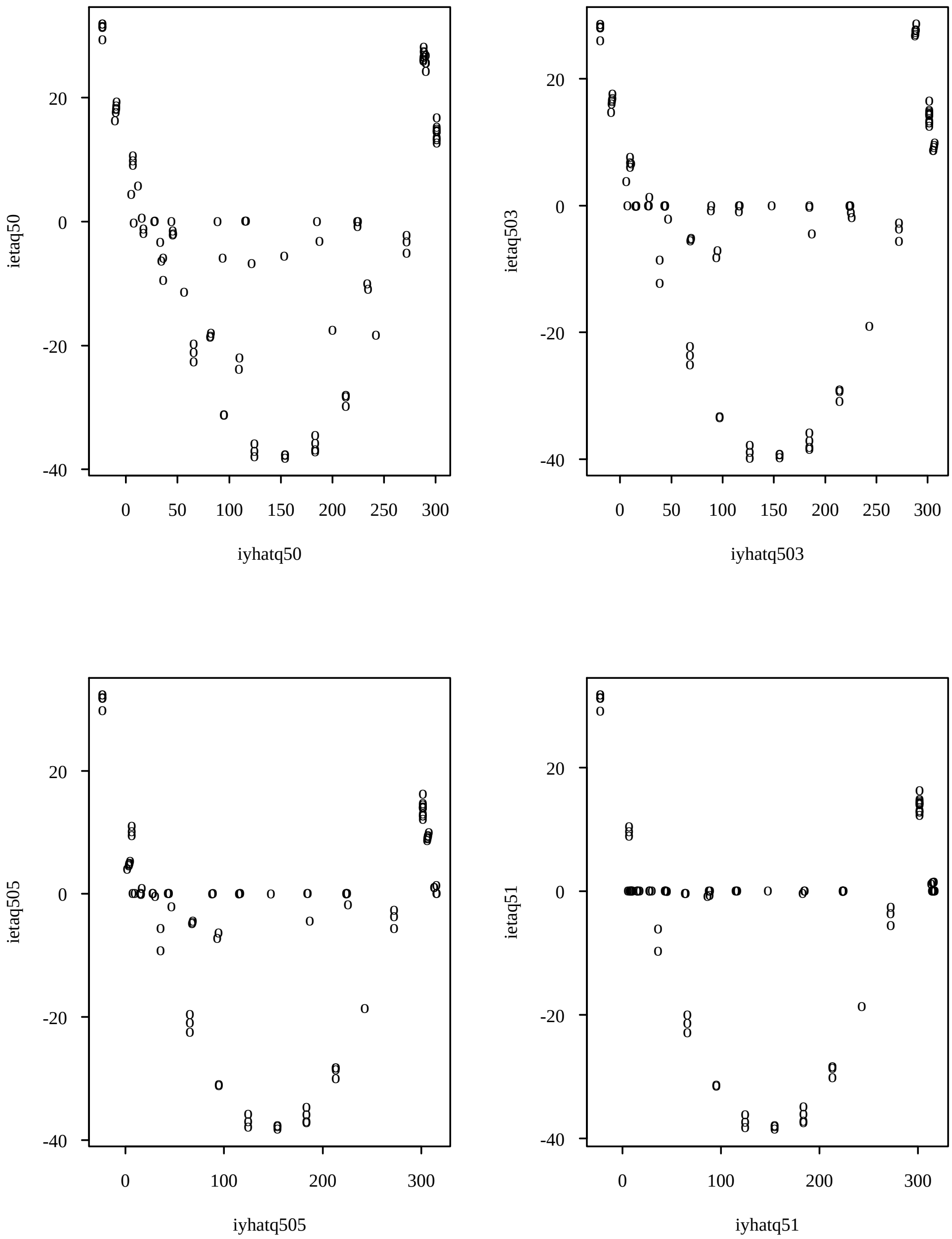


Figure 9: Quadratic model, censoring level 0.7. First graph: original residual interval width, second graph: 0.3 times increased, third graph: 0.5 times increased, fourth graph: double of the original interval width.

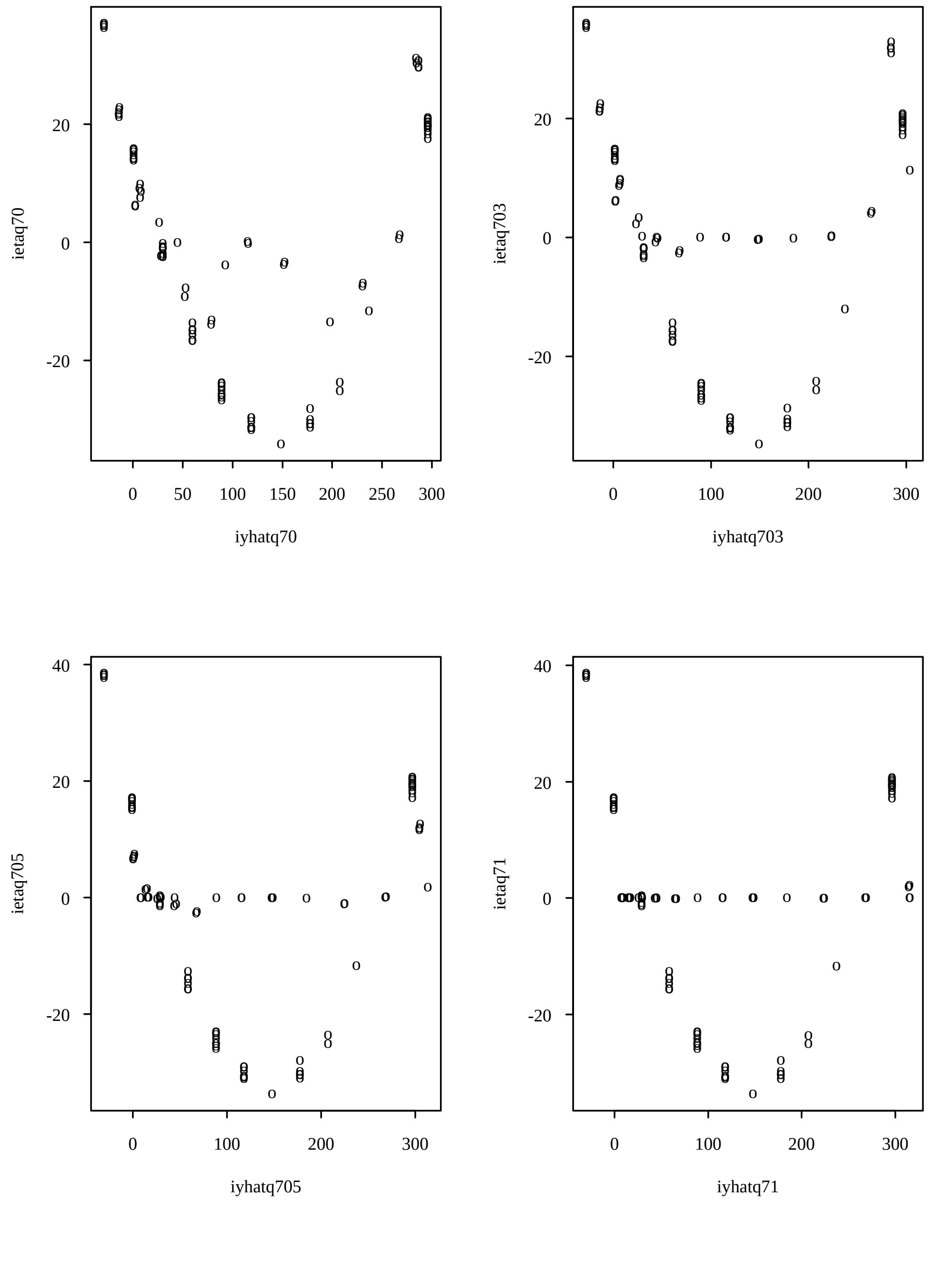


Figure 10: Quadratic model, censoring level 0.9. First graph: original residual interval width, second graph: 0.3 times increased, third graph: 0.5 times increased, fourth graph: double of the original interval width.

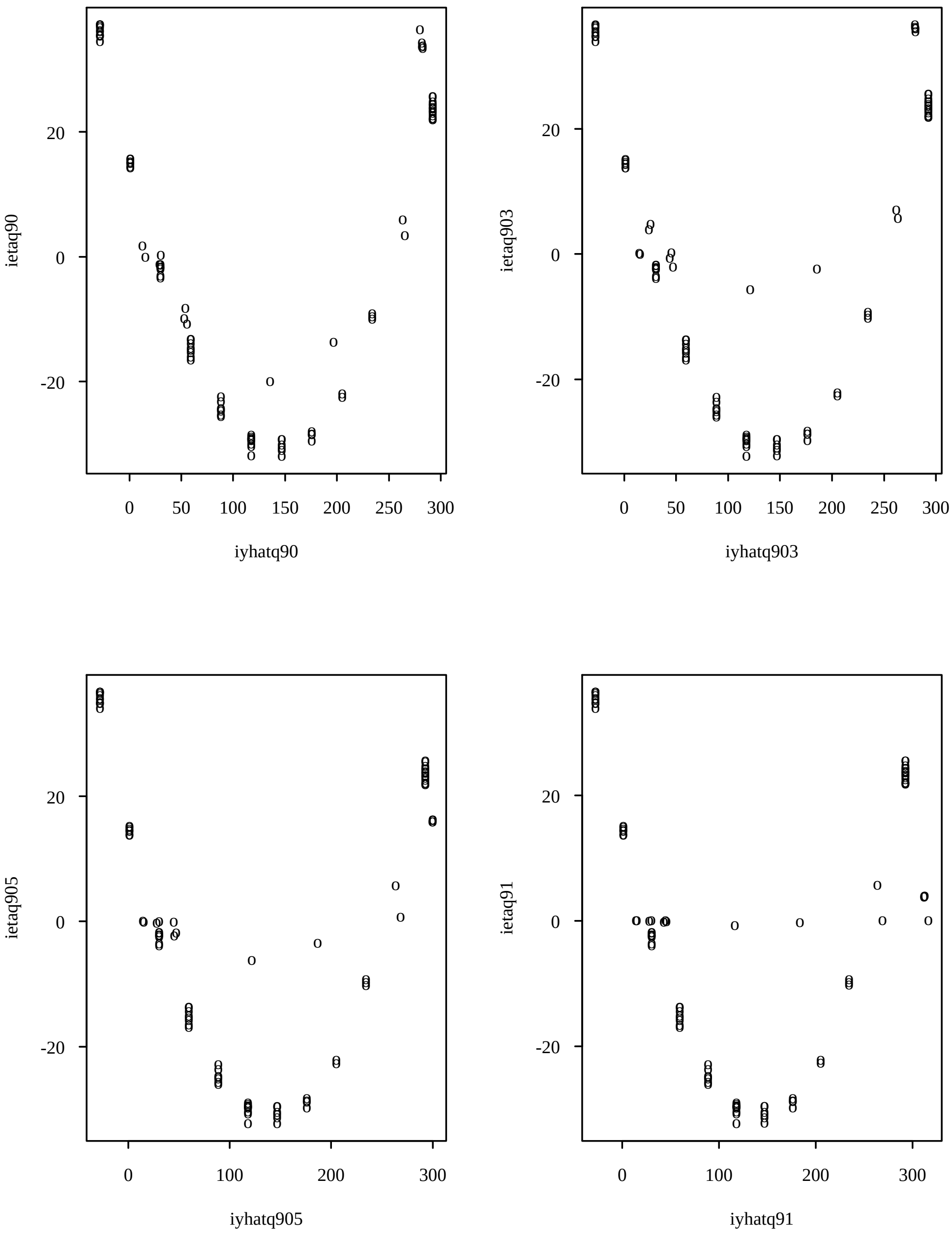


Figure 11: Covariate depending model, censoring level 0.1. First graph: original residual interval width, second graph: 0.3 times increased, third graph: 0.5 times increased, fourth graph: double of the original interval width.

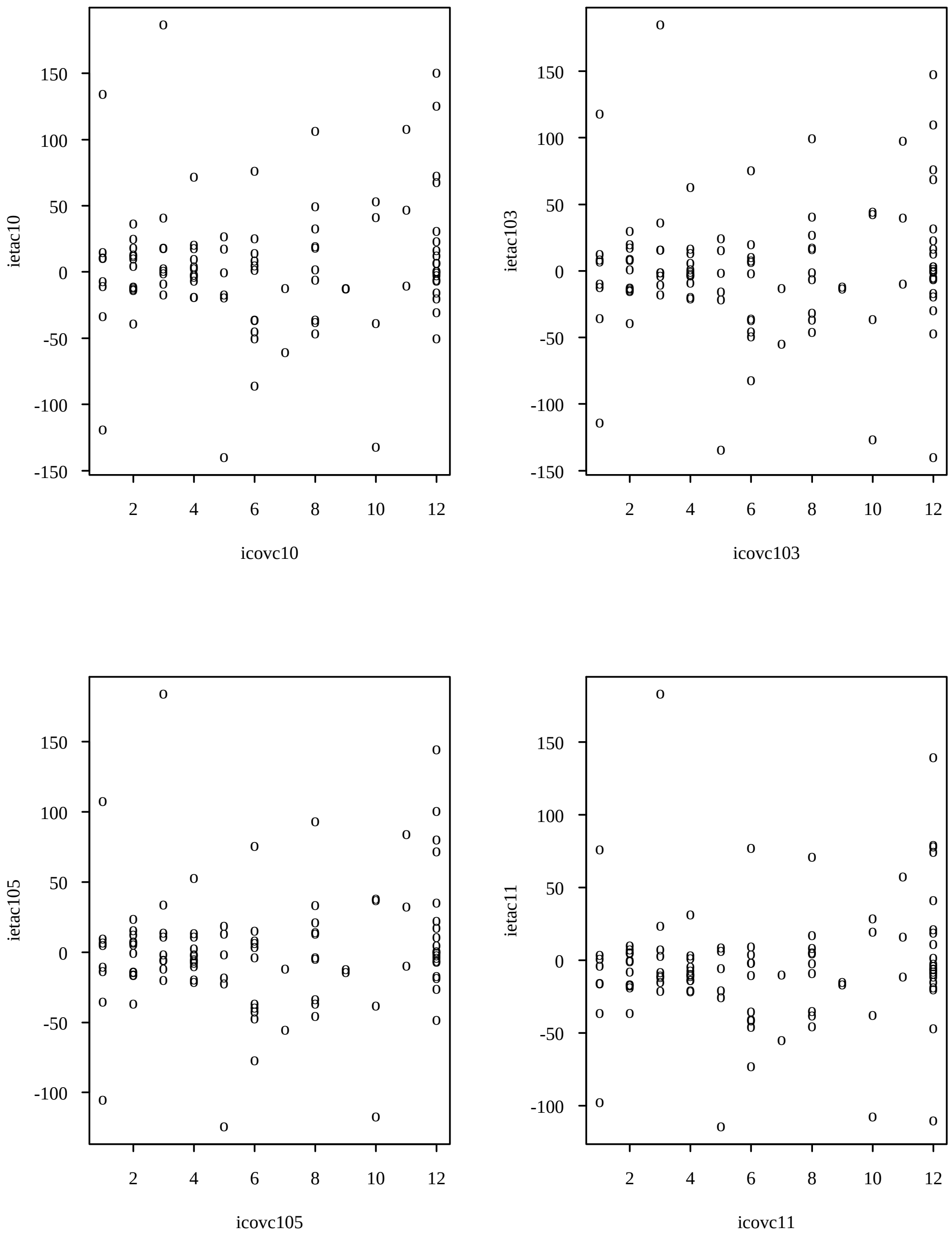


Figure 12: Covariate depending model, censoring level 0.3. First graph: original residual interval width, second graph: 0.3 times increased, third graph: 0.5 times increased, fourth graph: double of the original interval width.

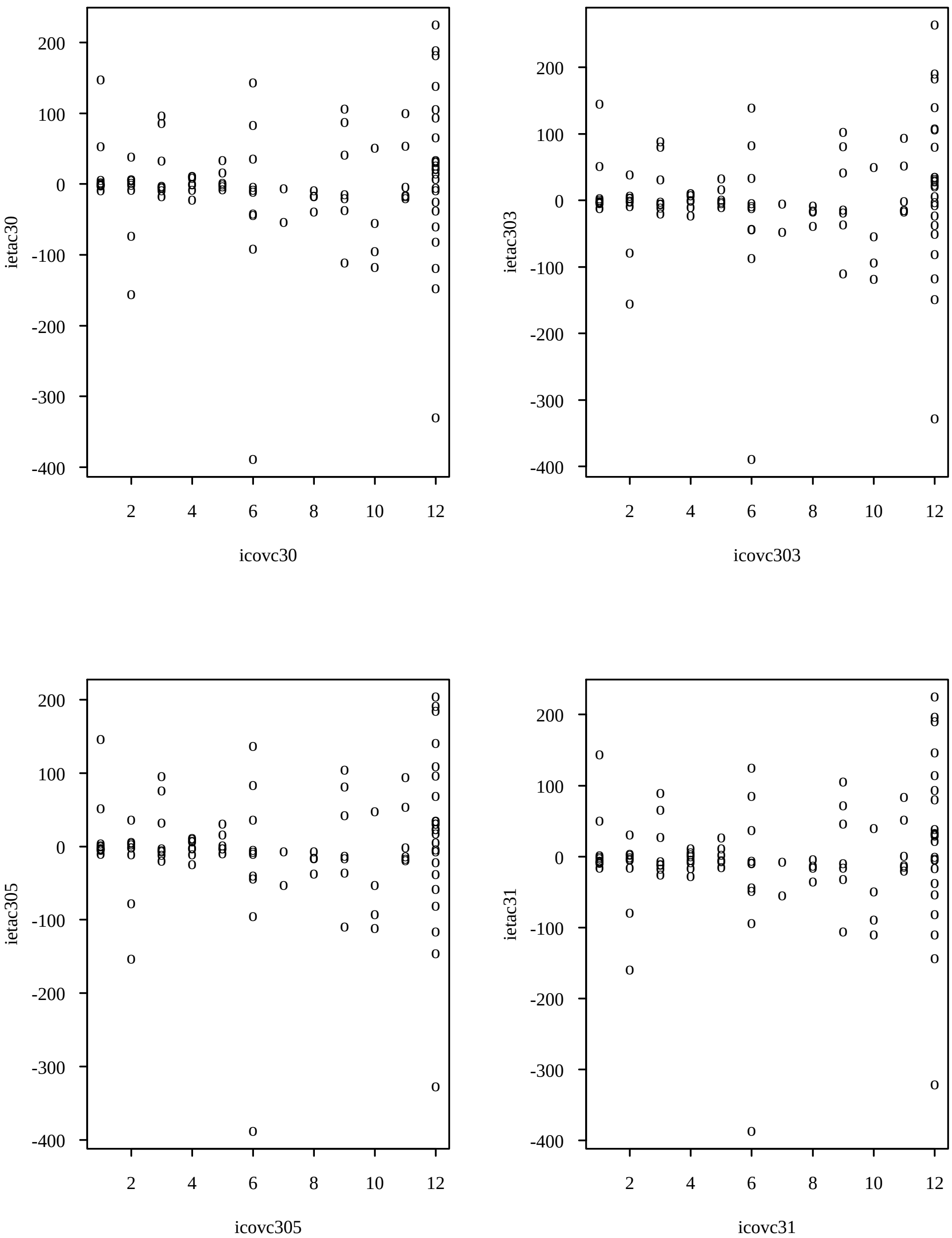


Figure 13: Covariate depending model, censoring level 0.5. First graph: original residual interval width, second graph: 0.3 times increased, third graph: 0.5 times increased, fourth graph: double of the original interval width.

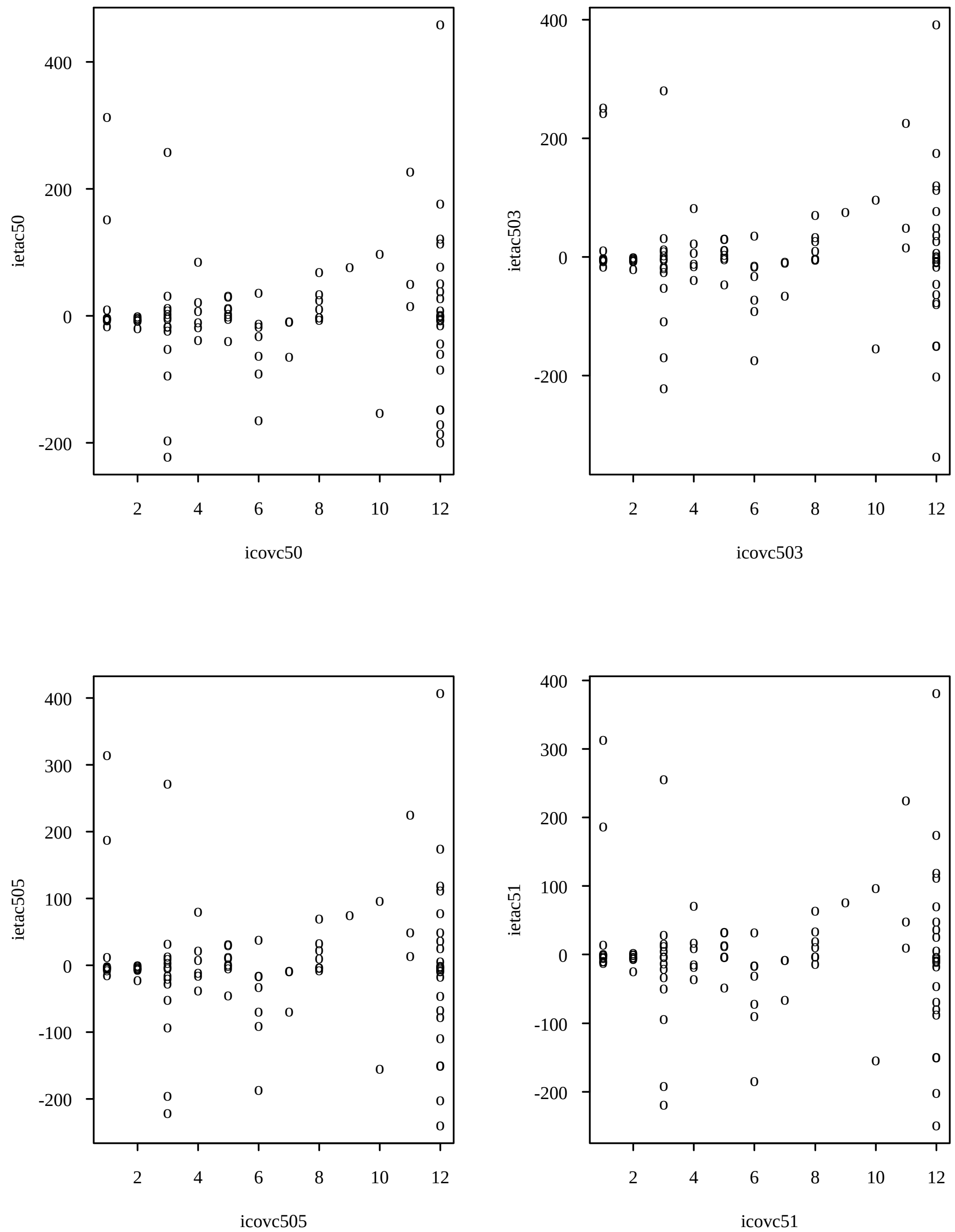


Figure 14: Covariate depending model, censoring level 0.7. First graph: original residual interval width, second graph: 0.3 times increased, third graph: 0.5 times increased, fourth graph: double of the original interval width.

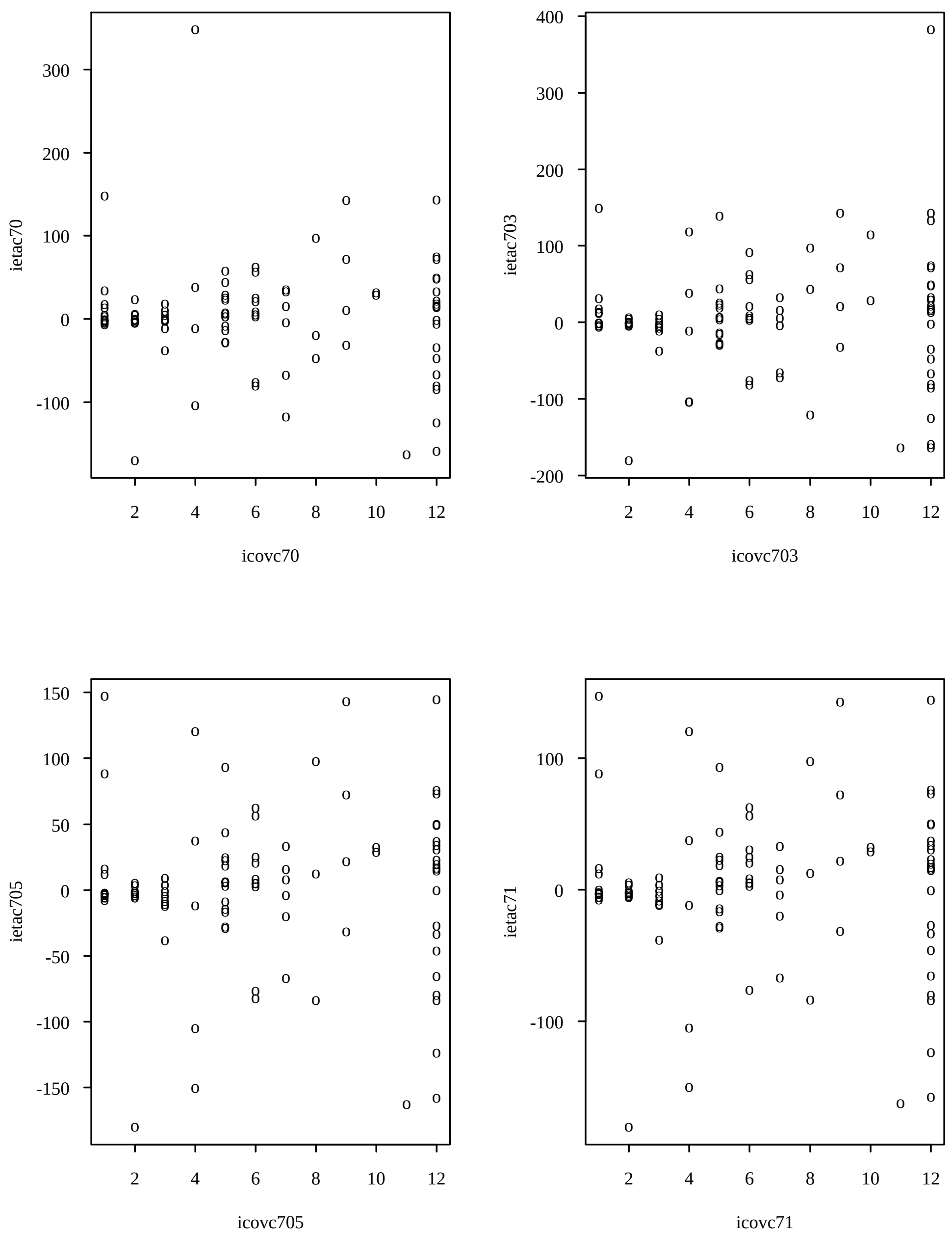
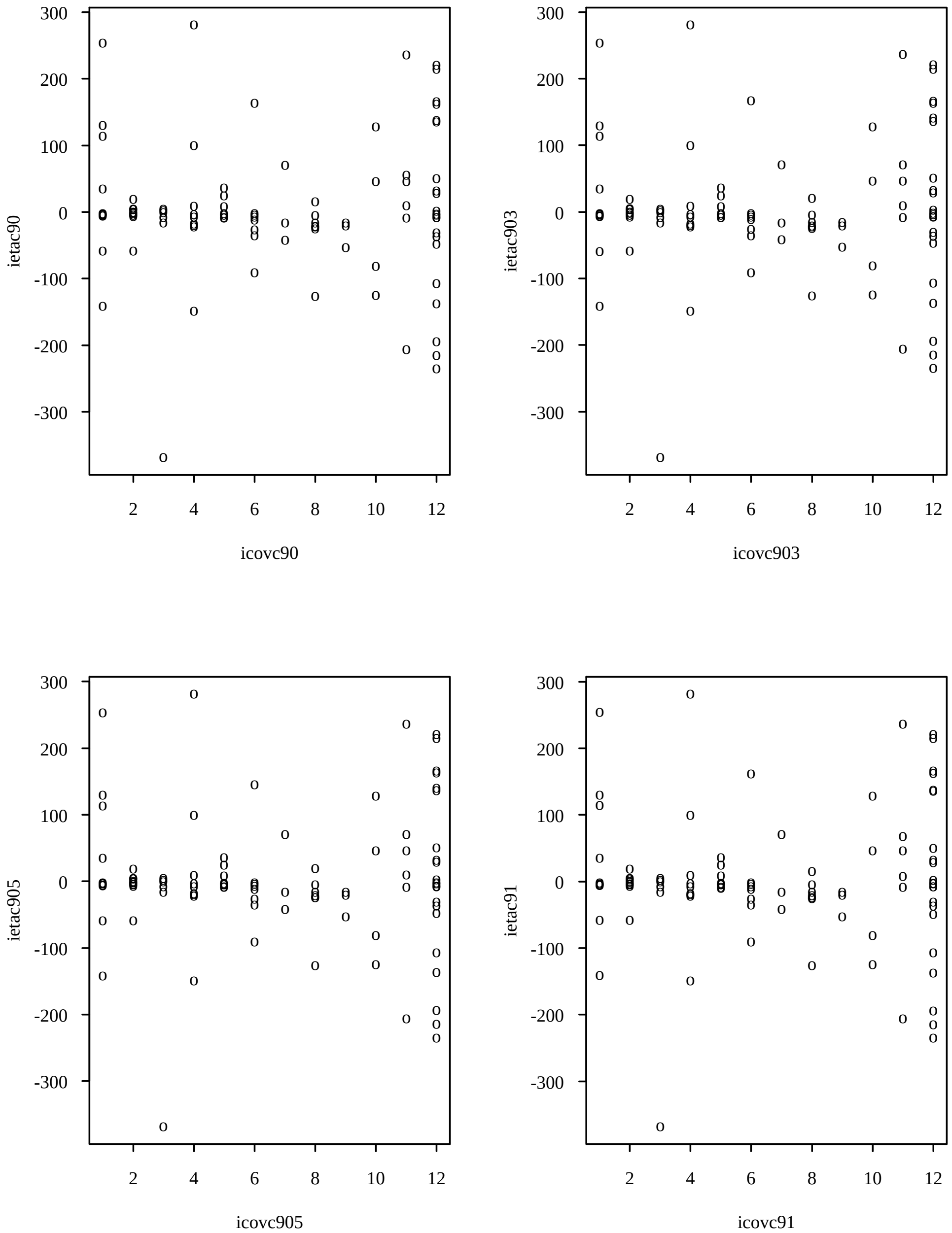


Figure 15: Covariate depending model, censoring level 0.9. First graph: original residual interval width, second graph: 0.3 times increased, third graph: 0.5 times increased, fourth graph: double of the original interval width.



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).

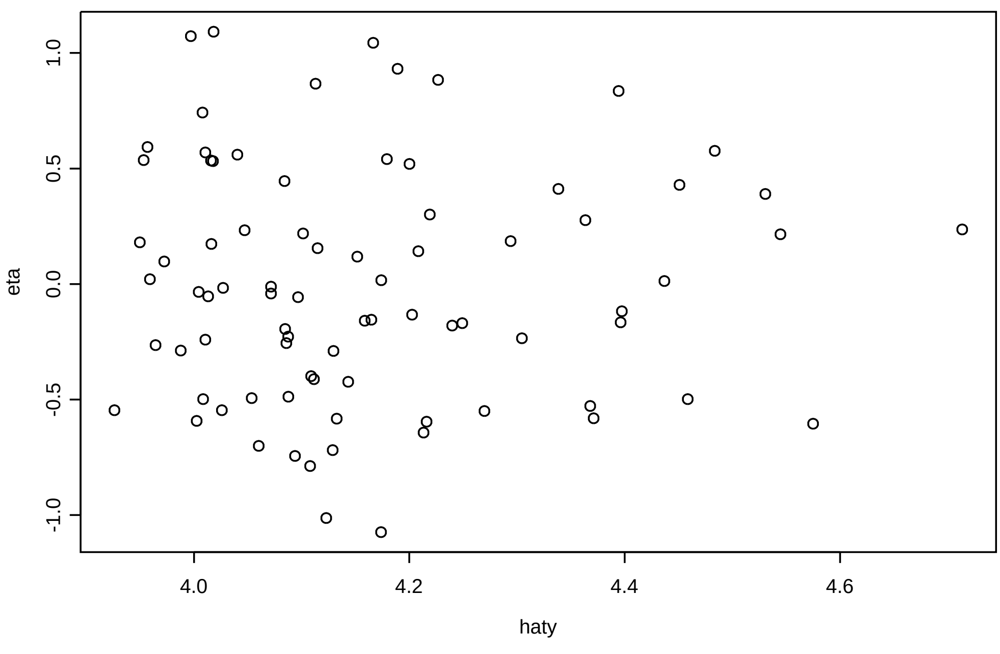
The covariate Z is taken to be the patient's time between Indinavir failure and enrollment. It is of interest examining whether there is an association between Z and age X with the log10 viral load level Y at the time of enrollment. The covariate Z was of interest because delays in initiating ACTG359 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.

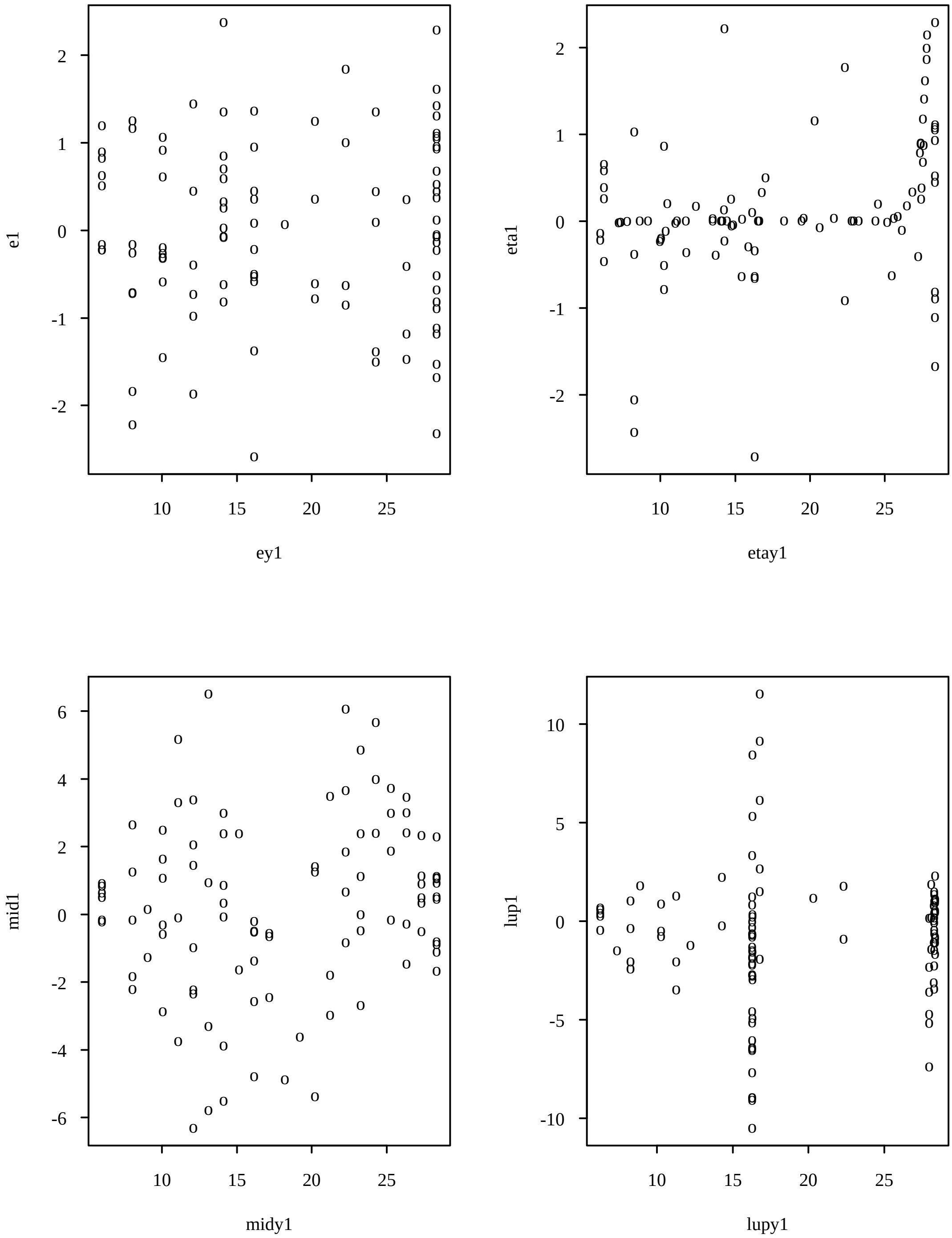
Yu, Q.; Li, L.; Wong, G.Y.C. (2000): On consistency of the self-consistent estimator of survival functions with interval-censored data. *Scandinavian Journal of Statistics, 27*, 35-44.

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.

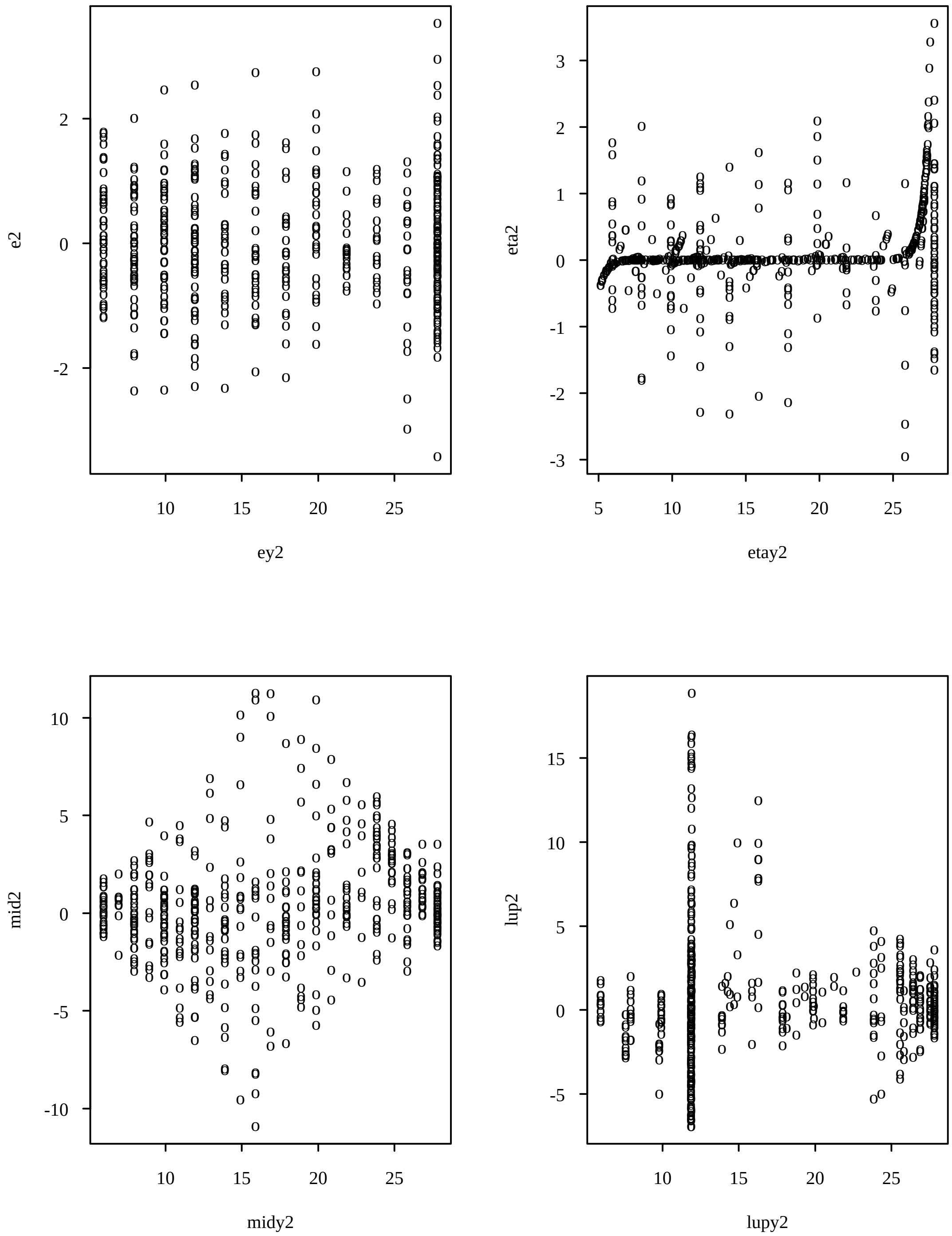
Appendix A

Residual plots when the model is correctly specified

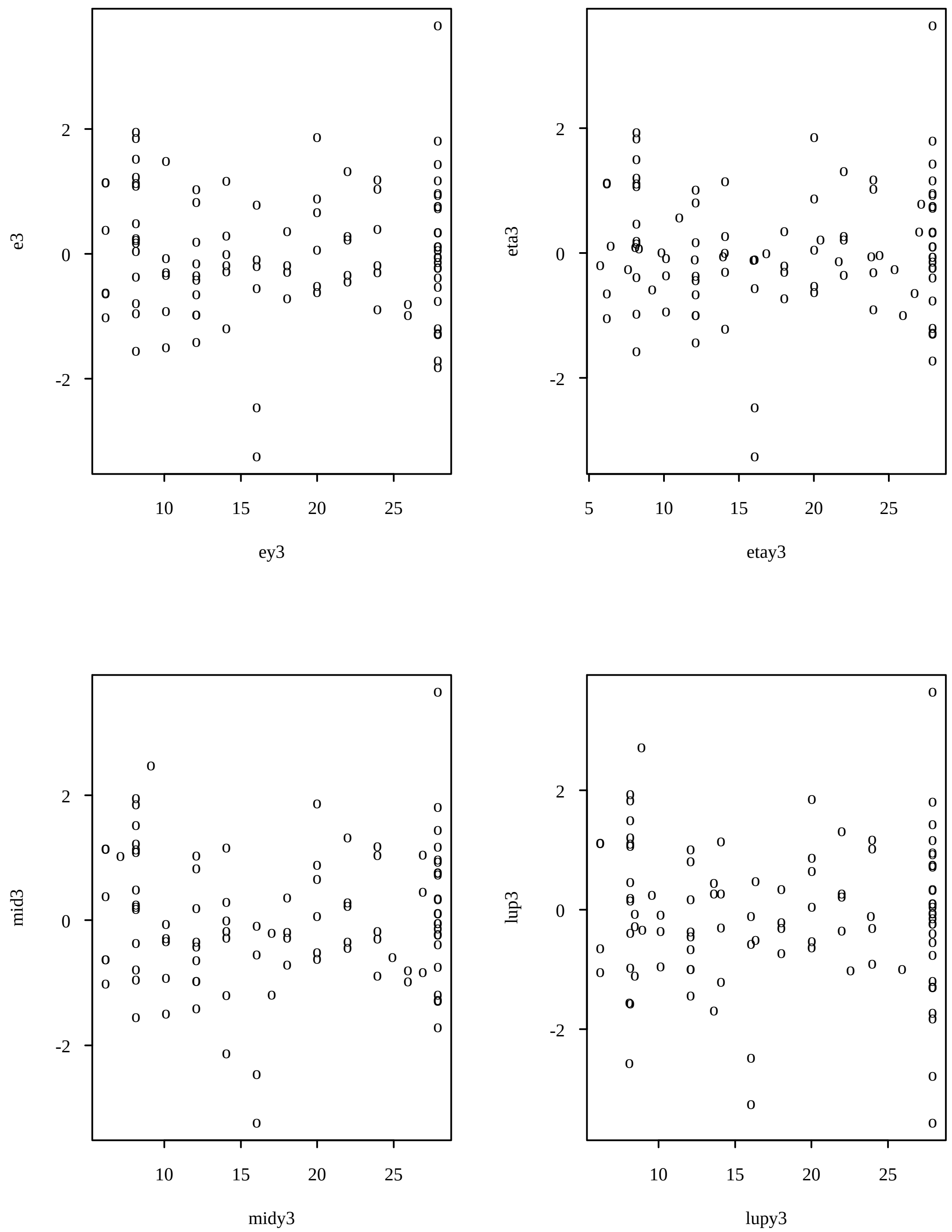
Scenario 1: covariate distribution $\text{Exp}(\frac{1}{8})$, β=2, p=0.3, n=100:



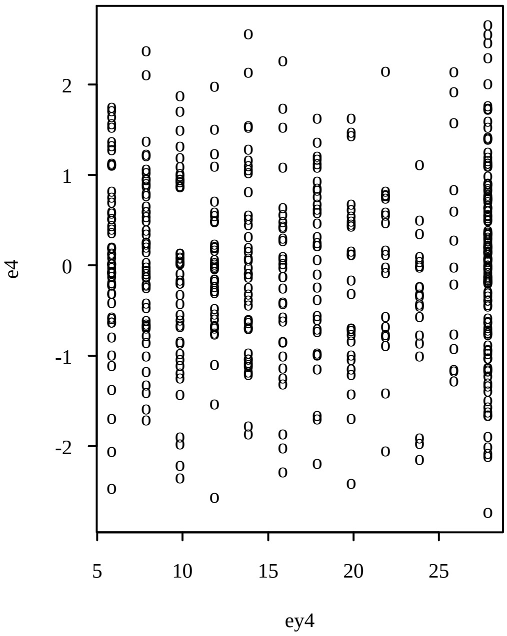
Scenario 2: covariate distribution $\text{Exp}(\frac{1}{8})$, β=2, p=0.3, n=500:

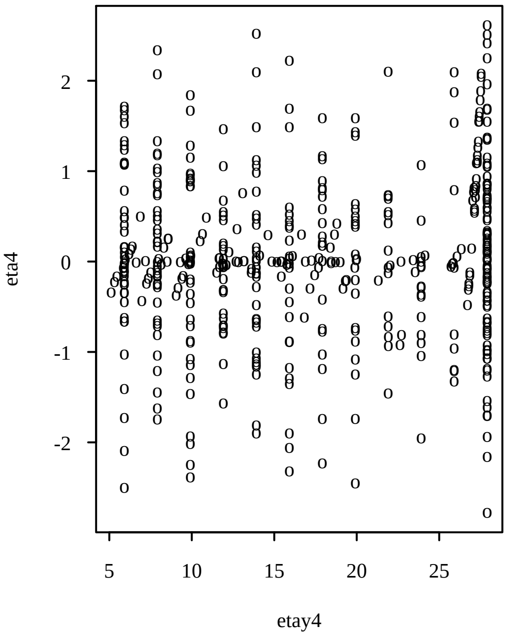


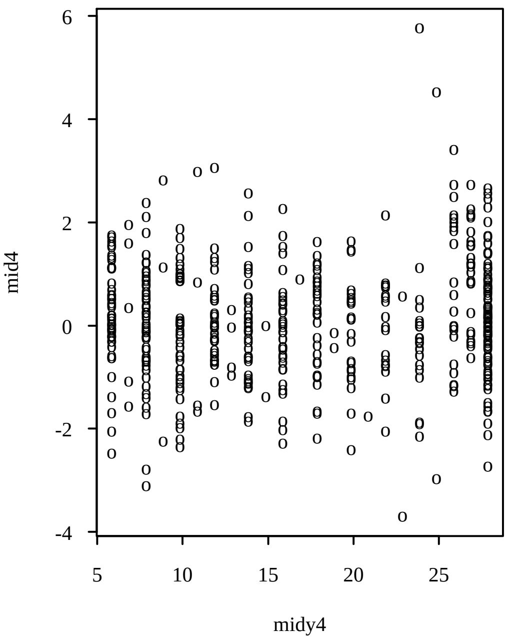
Scenario 3: covariate distribution $\text{Exp}(\frac{1}{8})$, β=2, p=0.7, n=100:

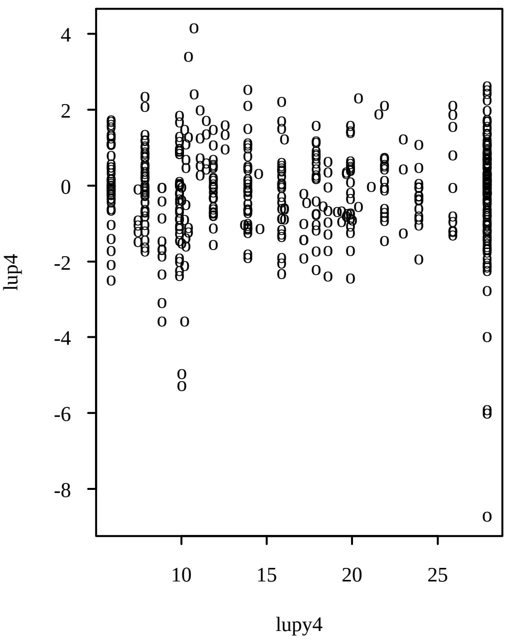


Scenario 4: covariate distribution Exp($\frac{1}{8}$), β=2, p=0.7, n=500:

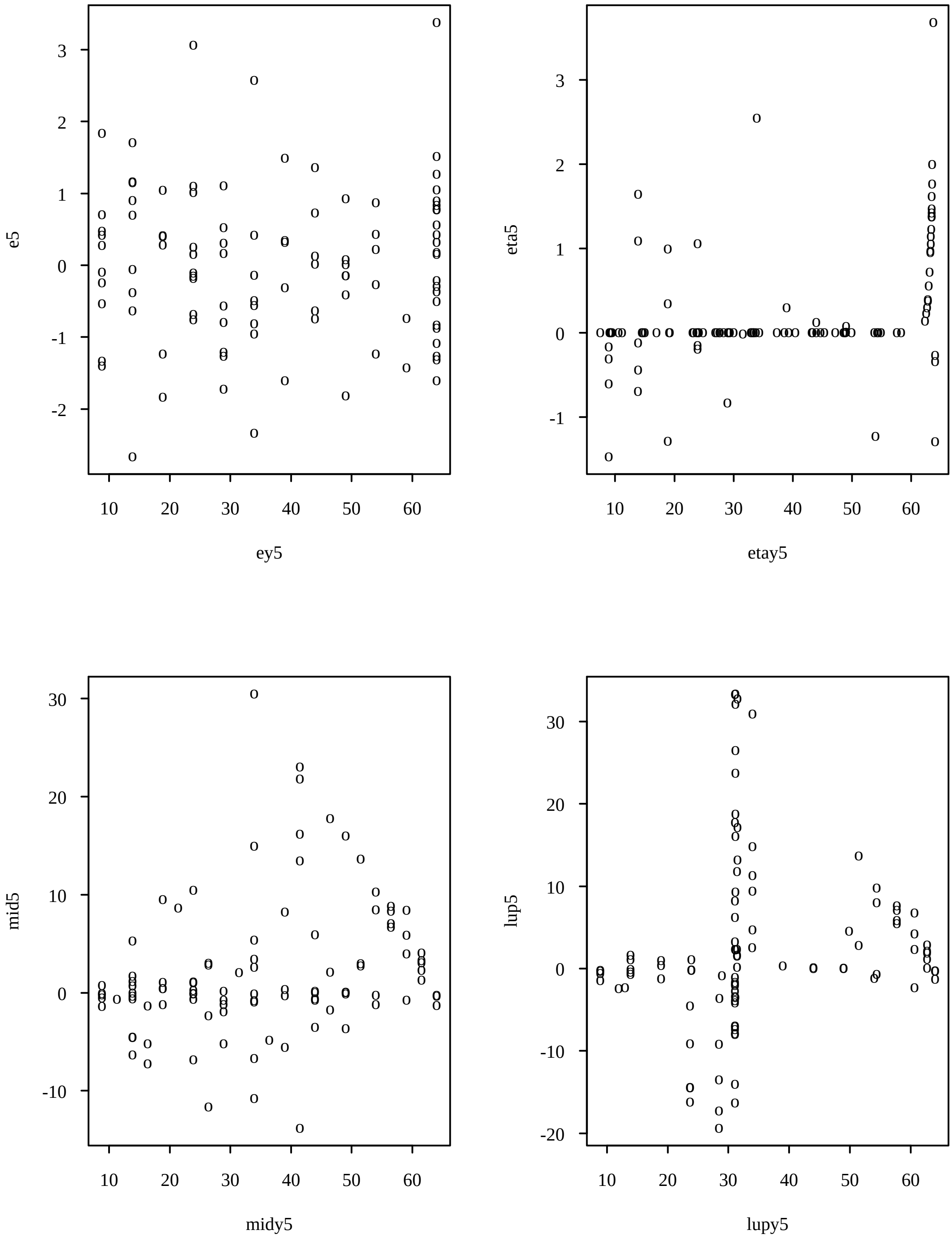




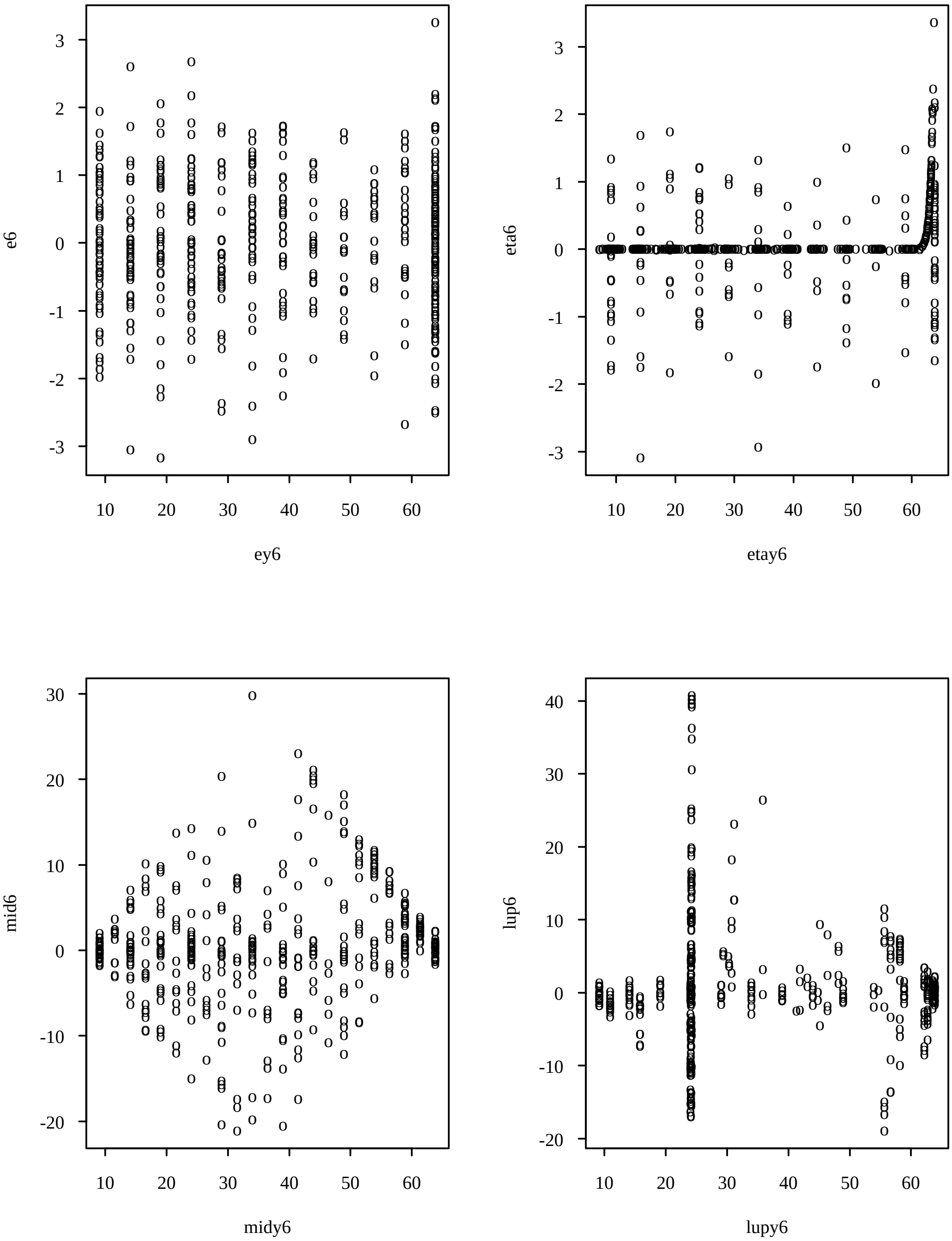




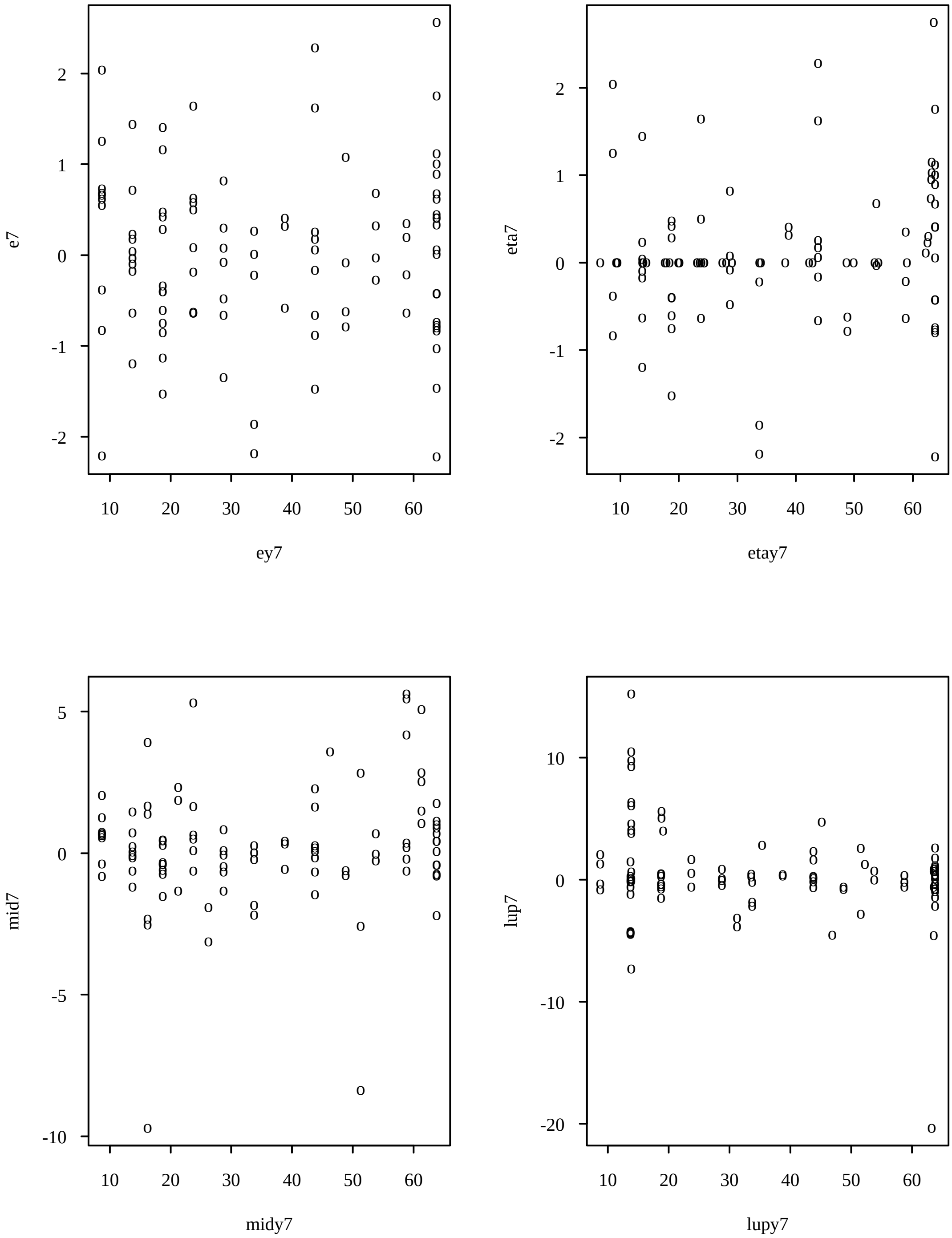
Scenario 5: covariate distribution $\text{Exp}(\frac{1}{8})$, β=5, p=0.3, n=100:



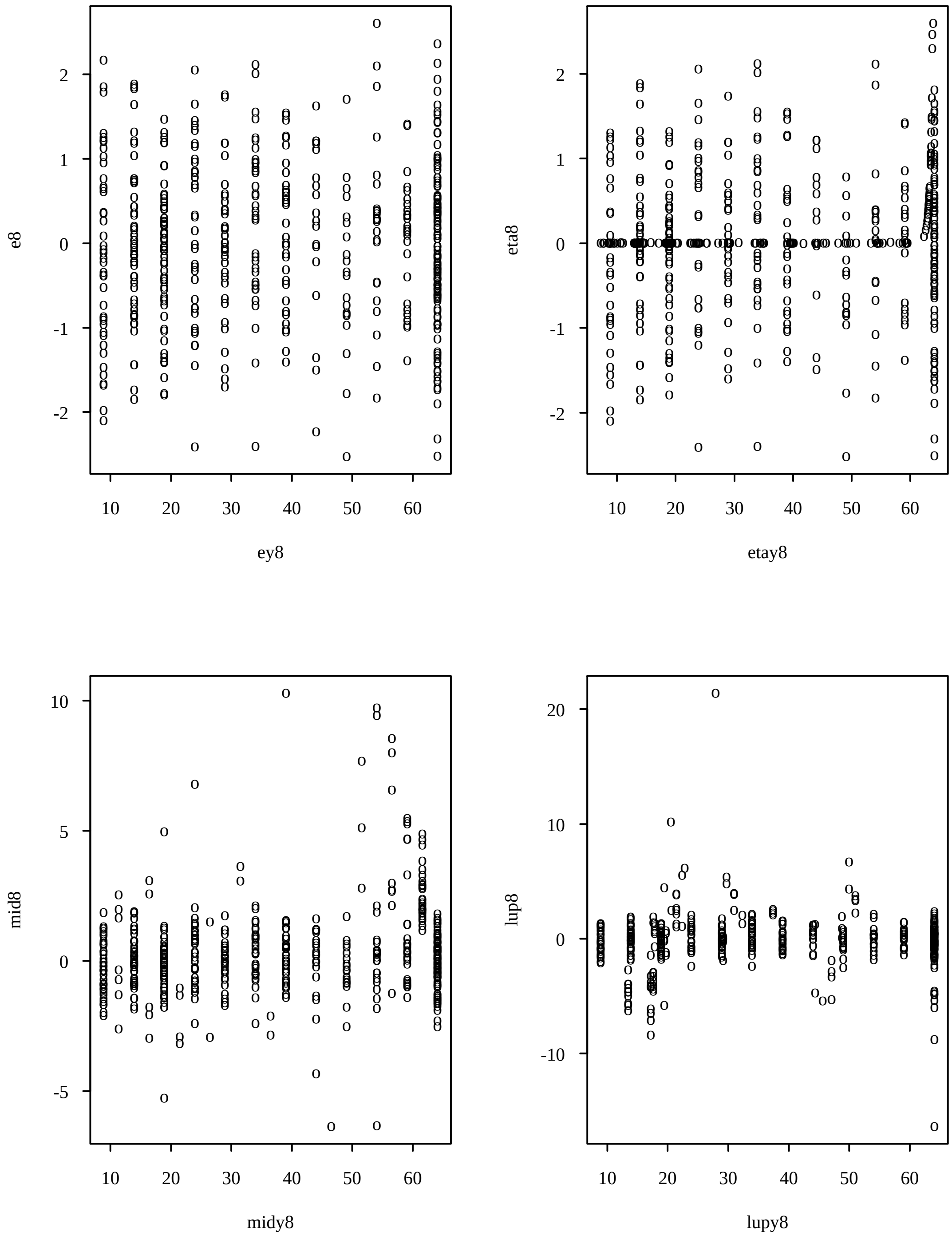
Scenario 6: covariate distribution $\mathrm{Exp}(\frac{1}{8})$, β=5, p=0.3, n=500:



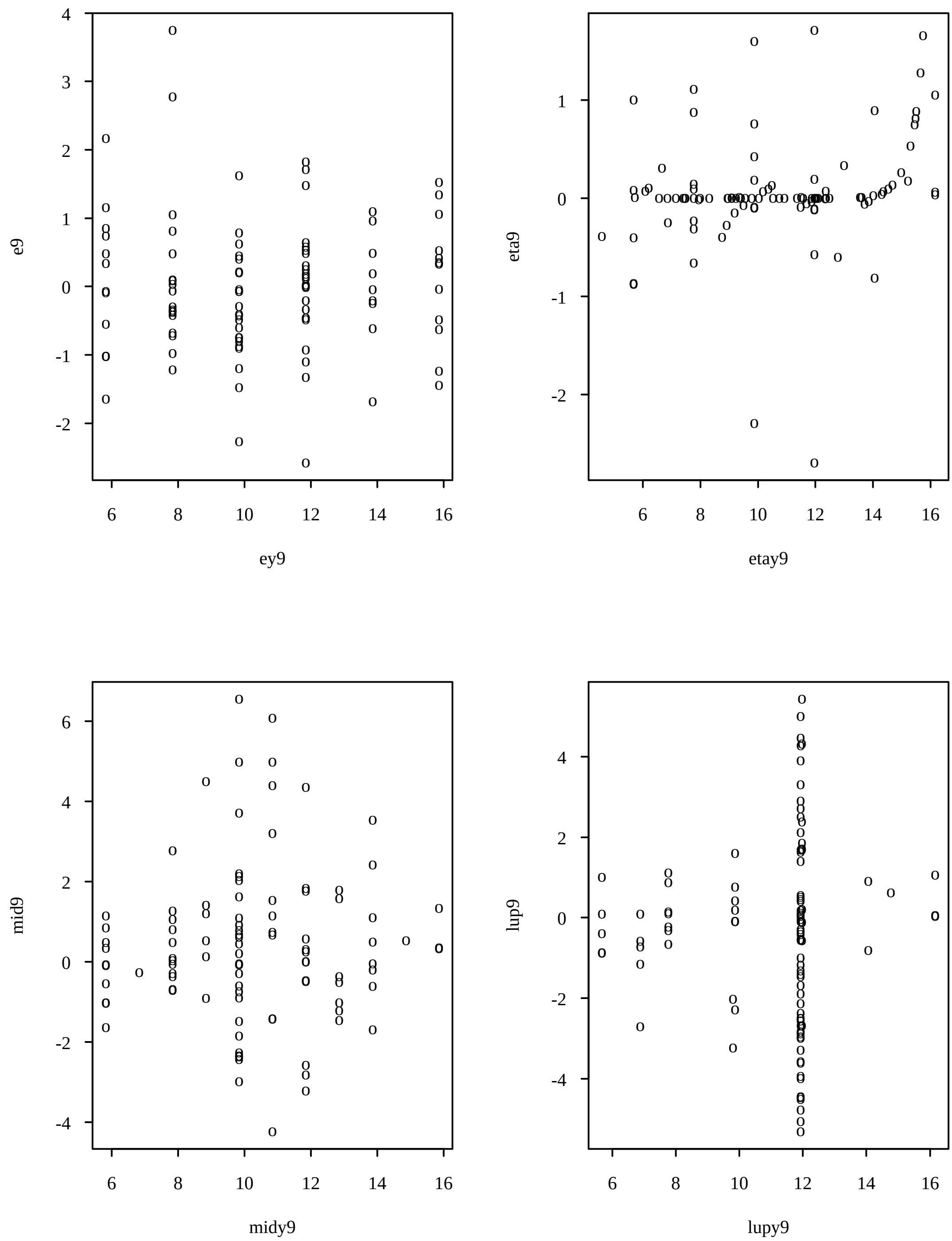
Scenario 7: covariate distribution Exp($\frac{1}{8}$), β=5, p=0.7, n=100:



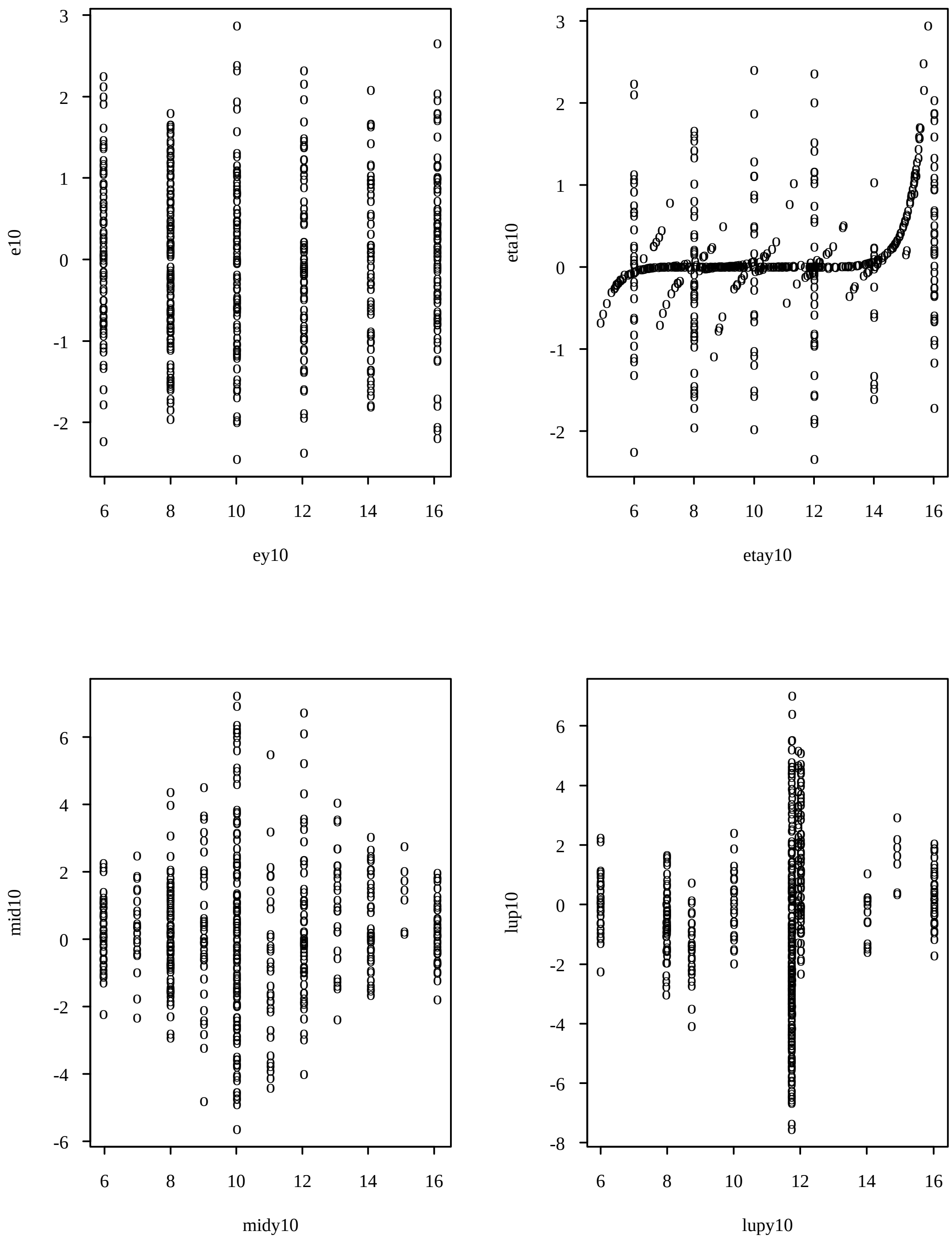
Scenario 8: covariate distribution Exp($\frac{1}{8}$), β=5, p=0.7, n=500:



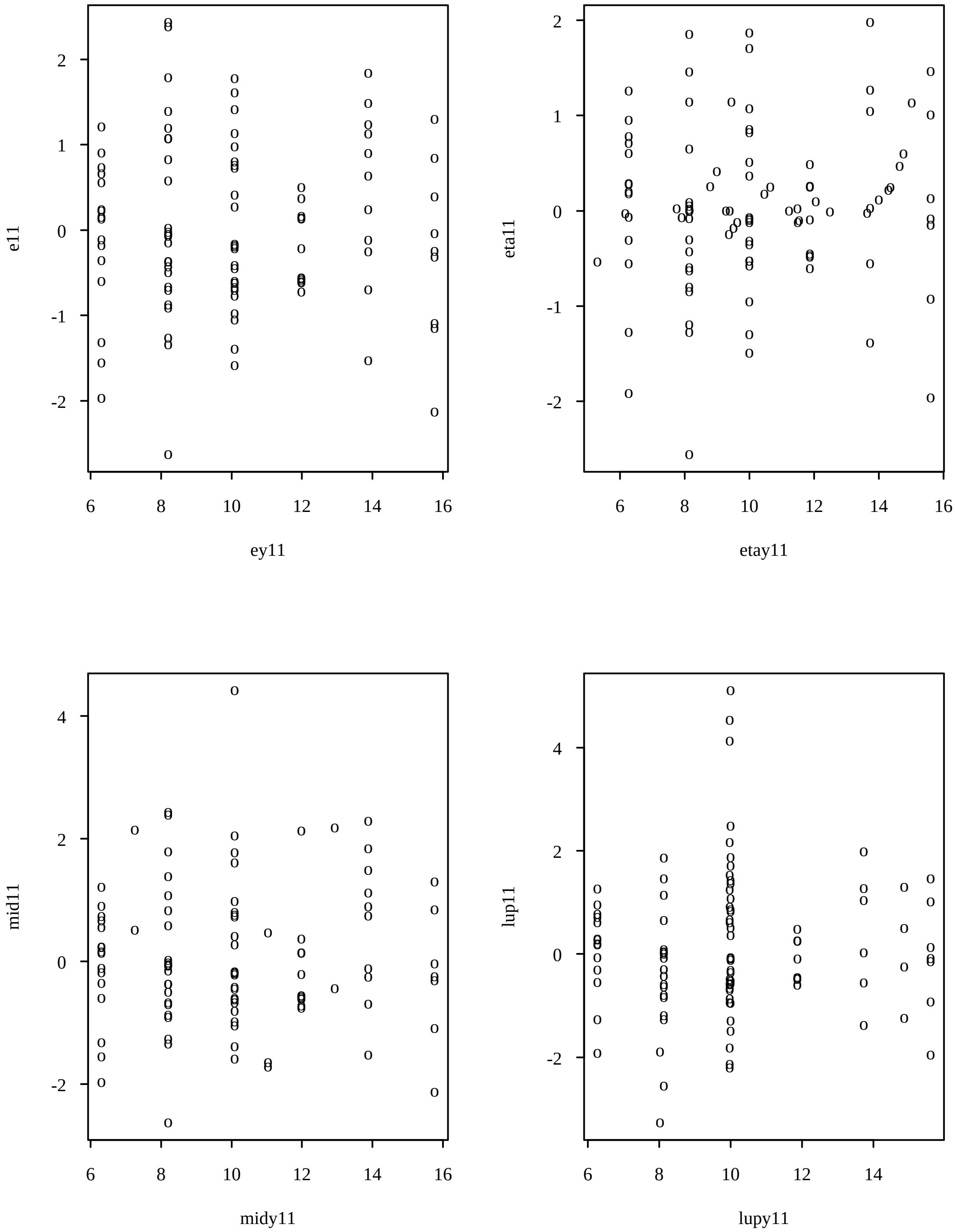
Scenario 9: covariate distribution $W(\frac{1}{6}, \frac{3}{2})$, β=2, p=0.3, n=100:



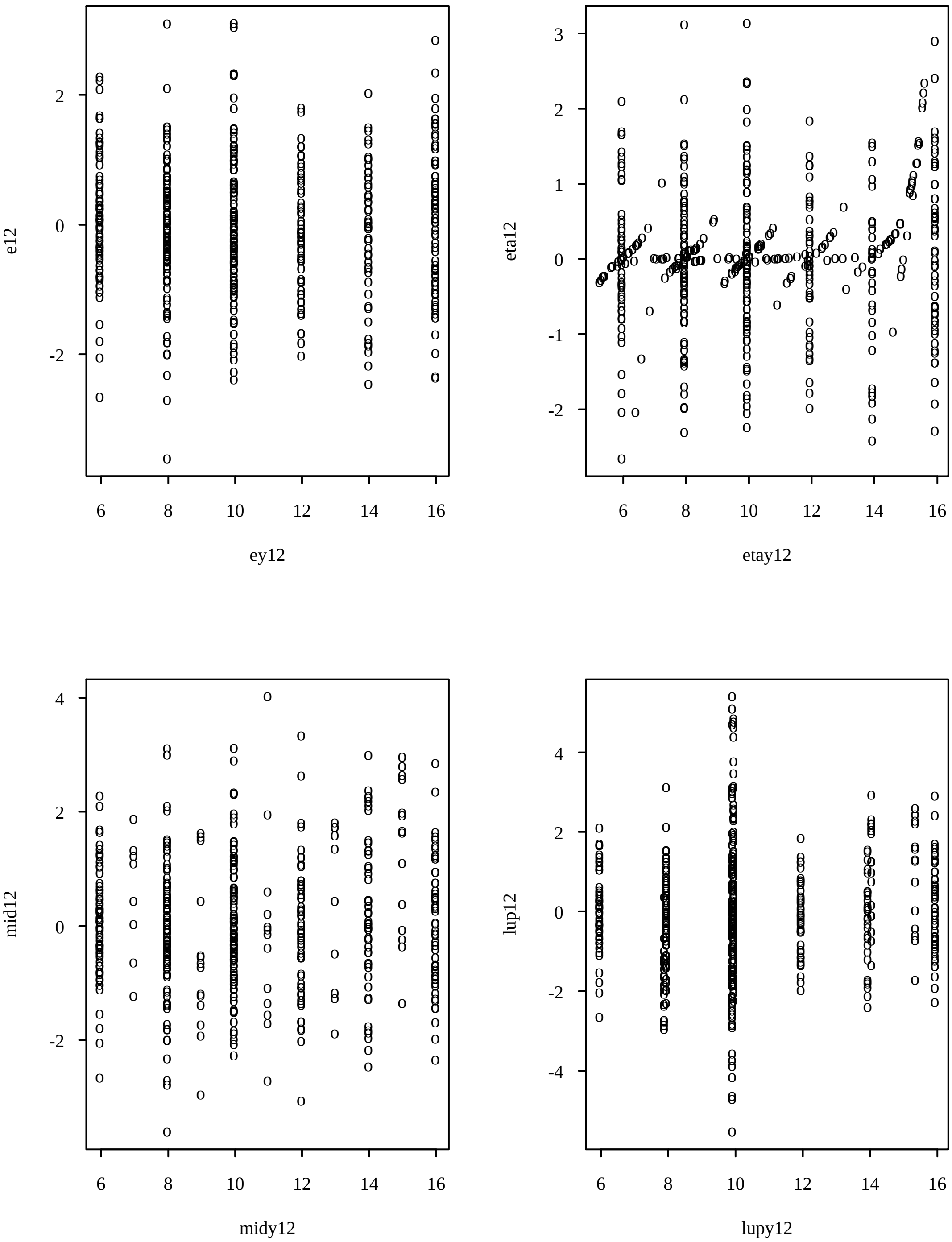
Scenario 10: covariate distribution $W(\frac{1}{6}, \frac{3}{2})$, β=2, p=0.3, n=500:



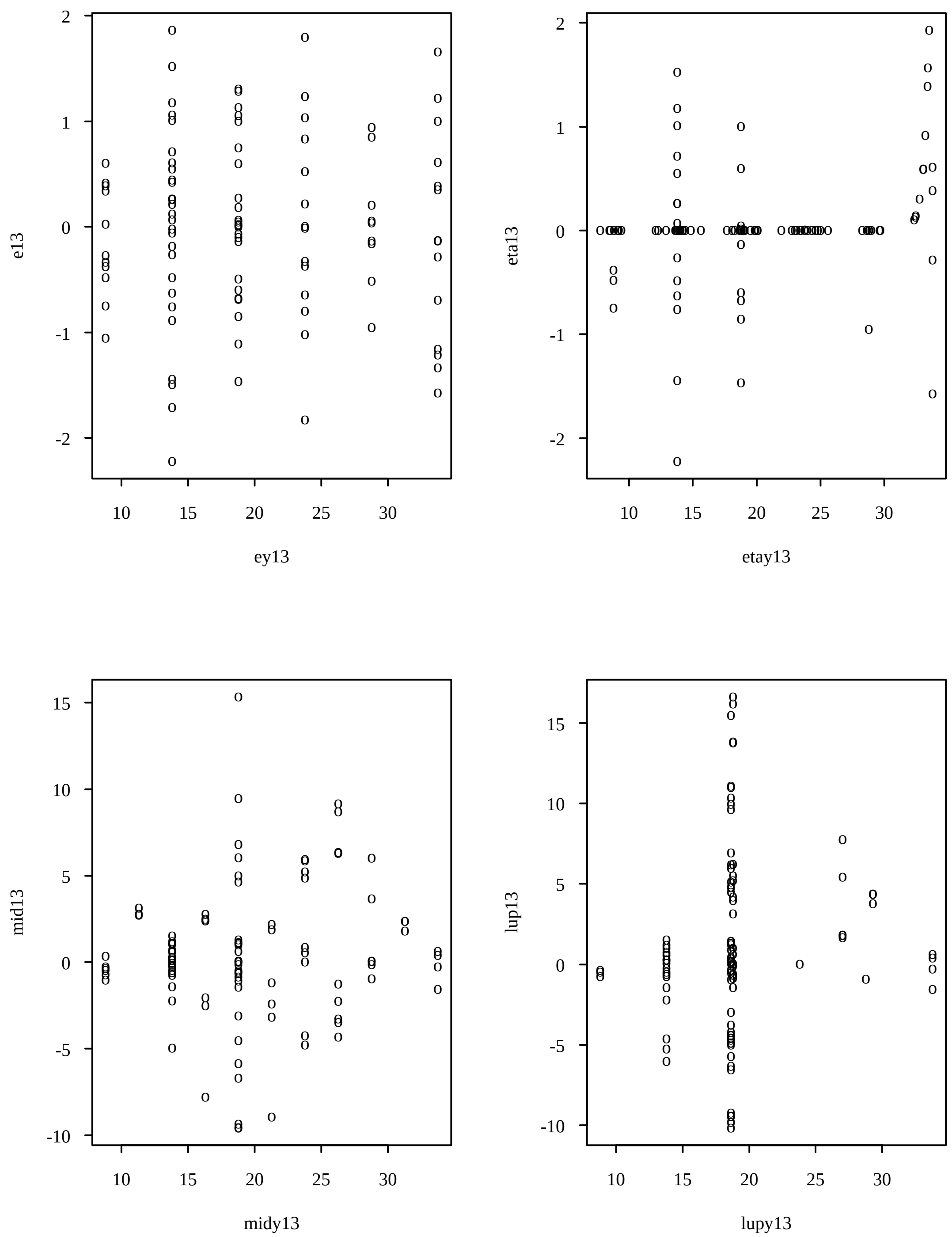
Scenario 11: covariate distribution W($\frac{1}{6}$, $\frac{3}{2}$), β=2, p=0.7, n=100:



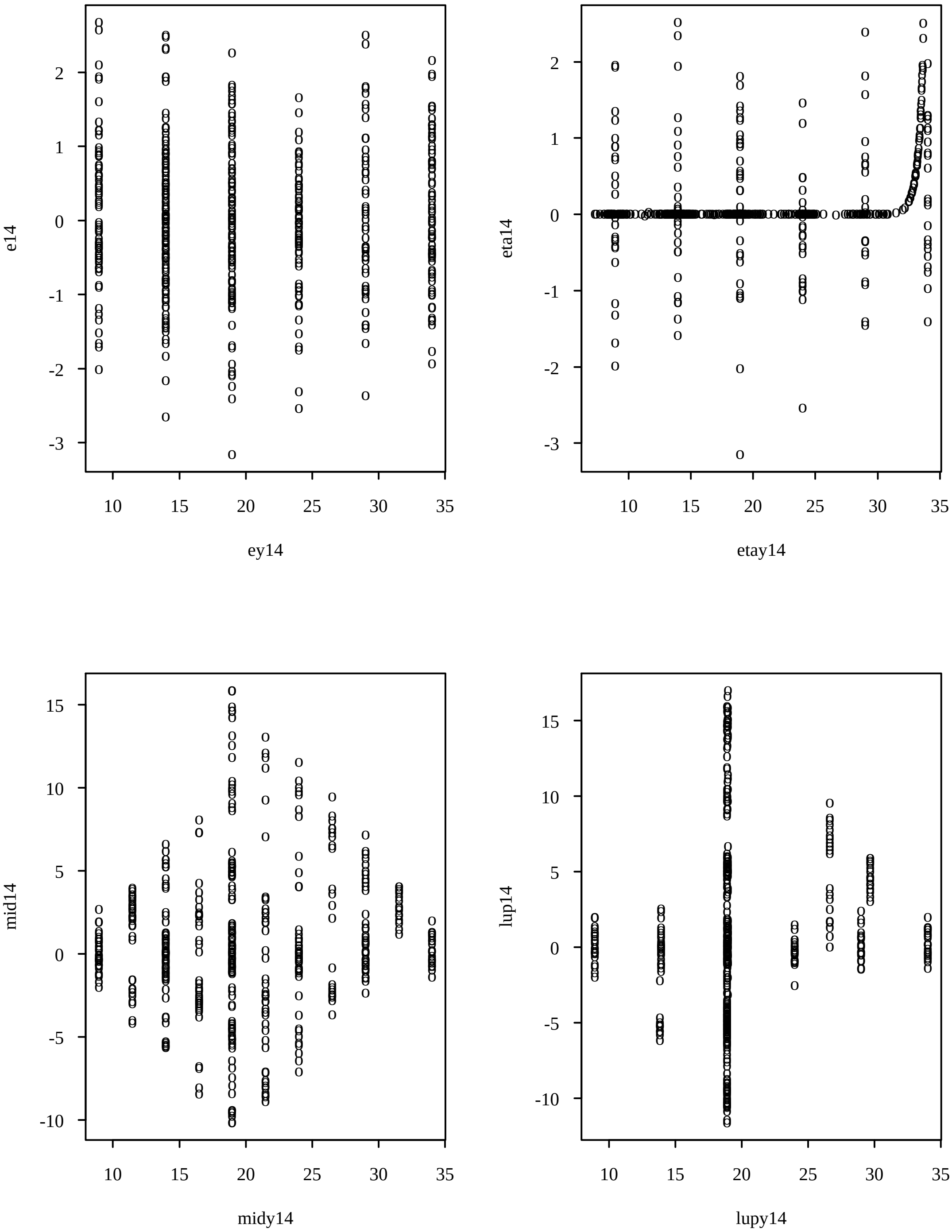
Scenario 12: covariate distribution $W(\frac{1}{6}, \frac{3}{2})$, β=2, p=0.7, n=500:



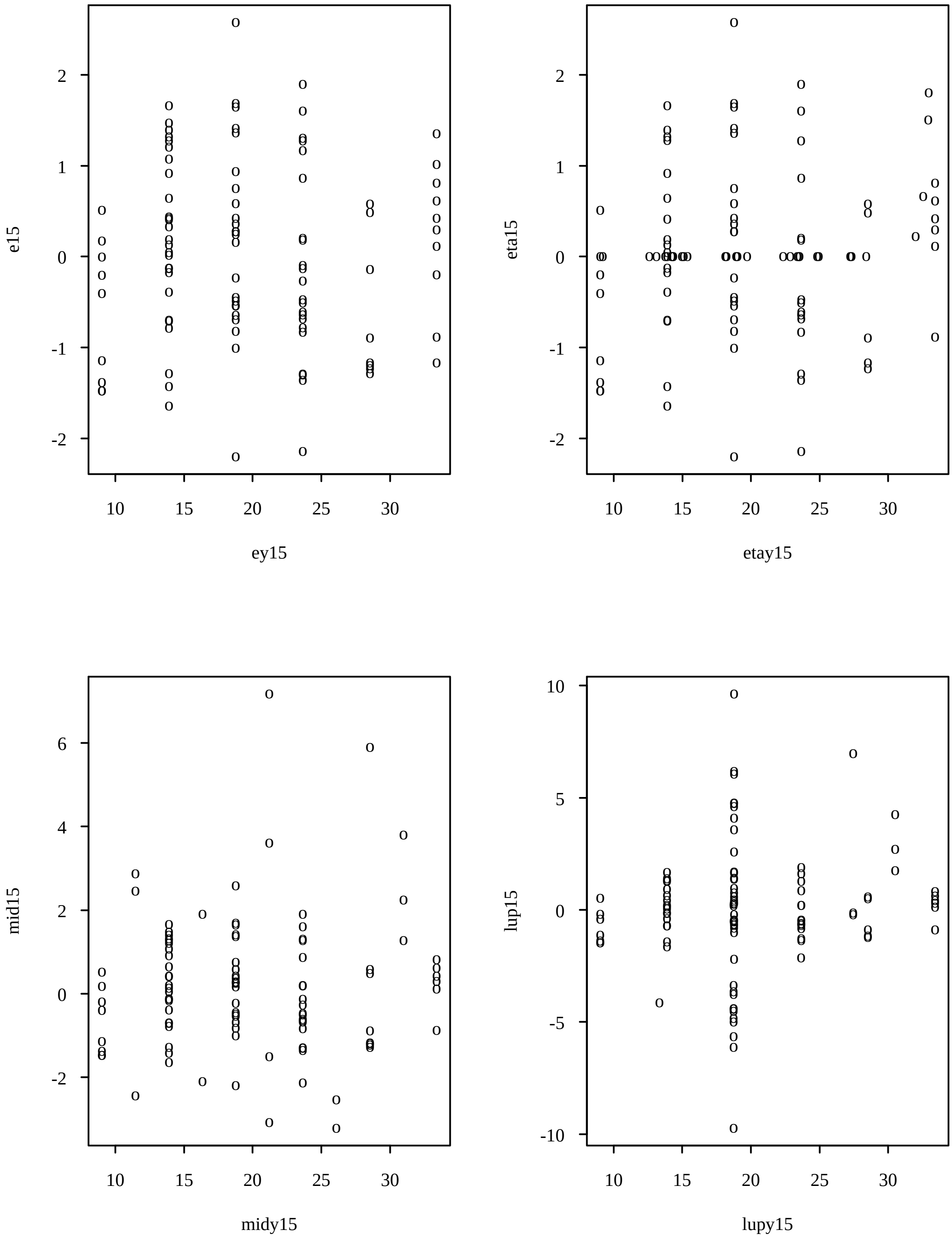
Scenario 13: covariate distribution W($\frac{1}{6}$, $\frac{3}{2}$), β=5, p=0.3, n=100:



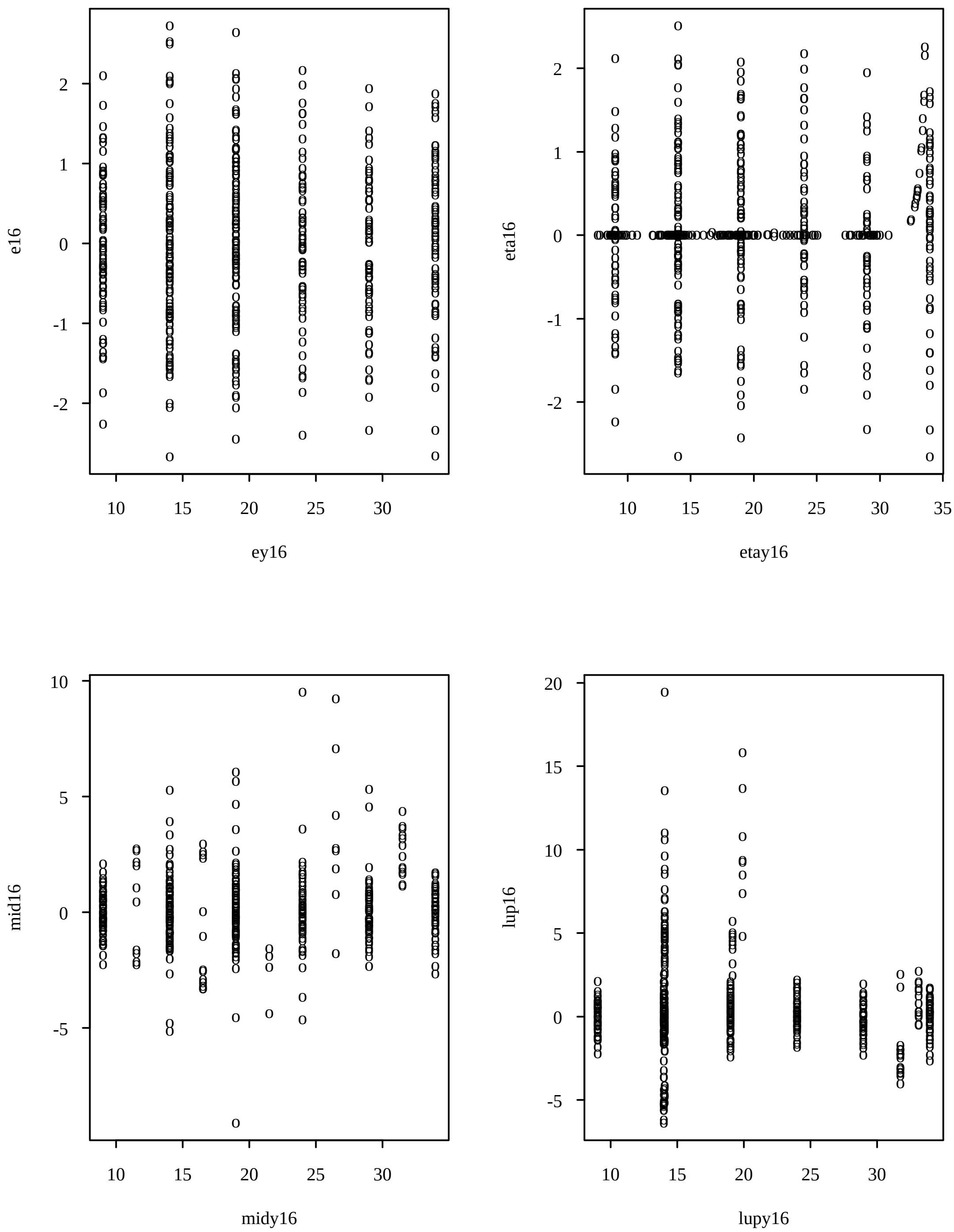
Scenario 14: covariate distribution $W(\frac{1}{6}, \frac{3}{2})$, β=5, p=0.3, n=500:



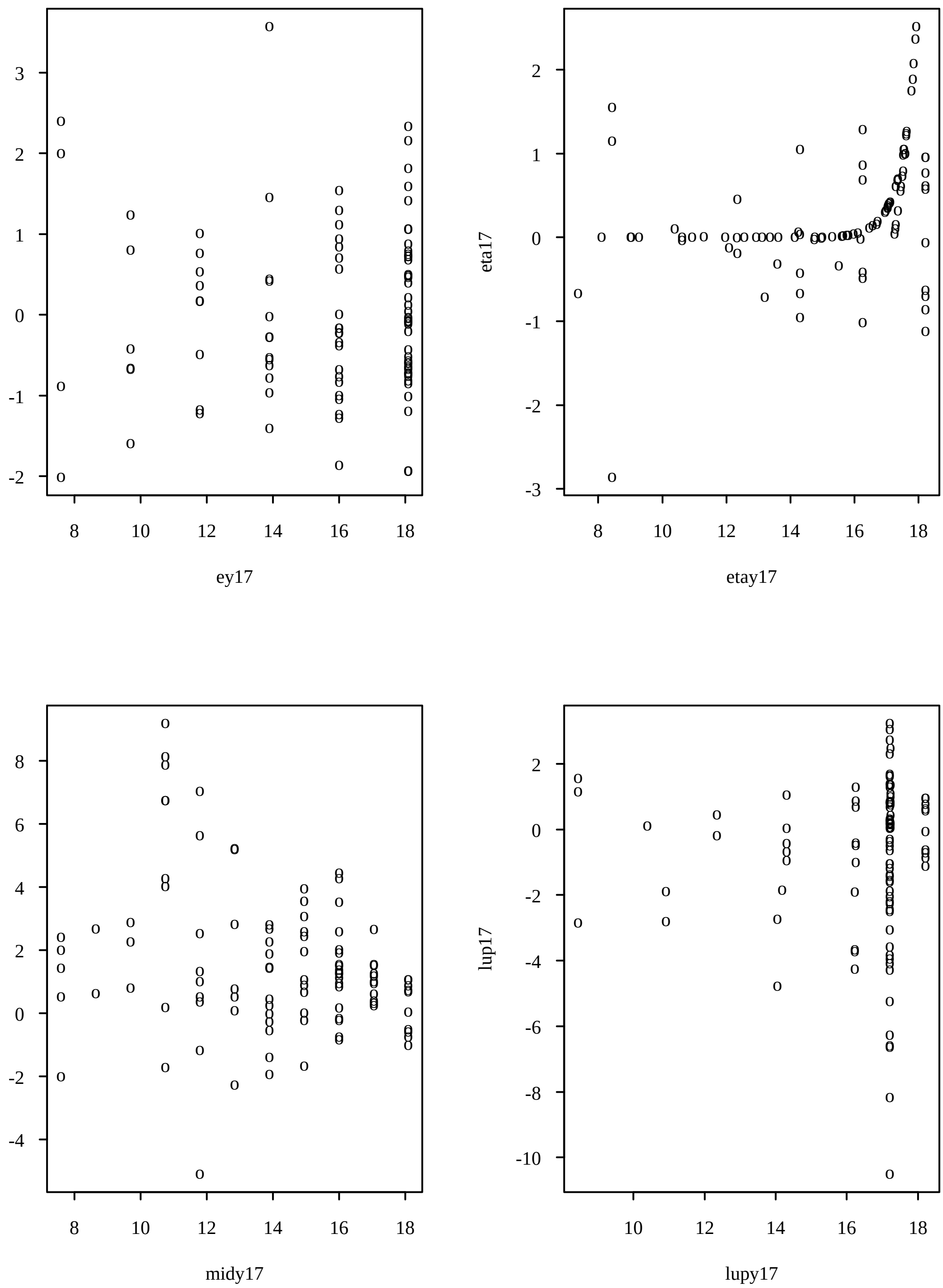
Scenario 15: covariate distribution W($\frac{1}{6}$, $\frac{3}{2}$), β=5, p=0.7, n=100:



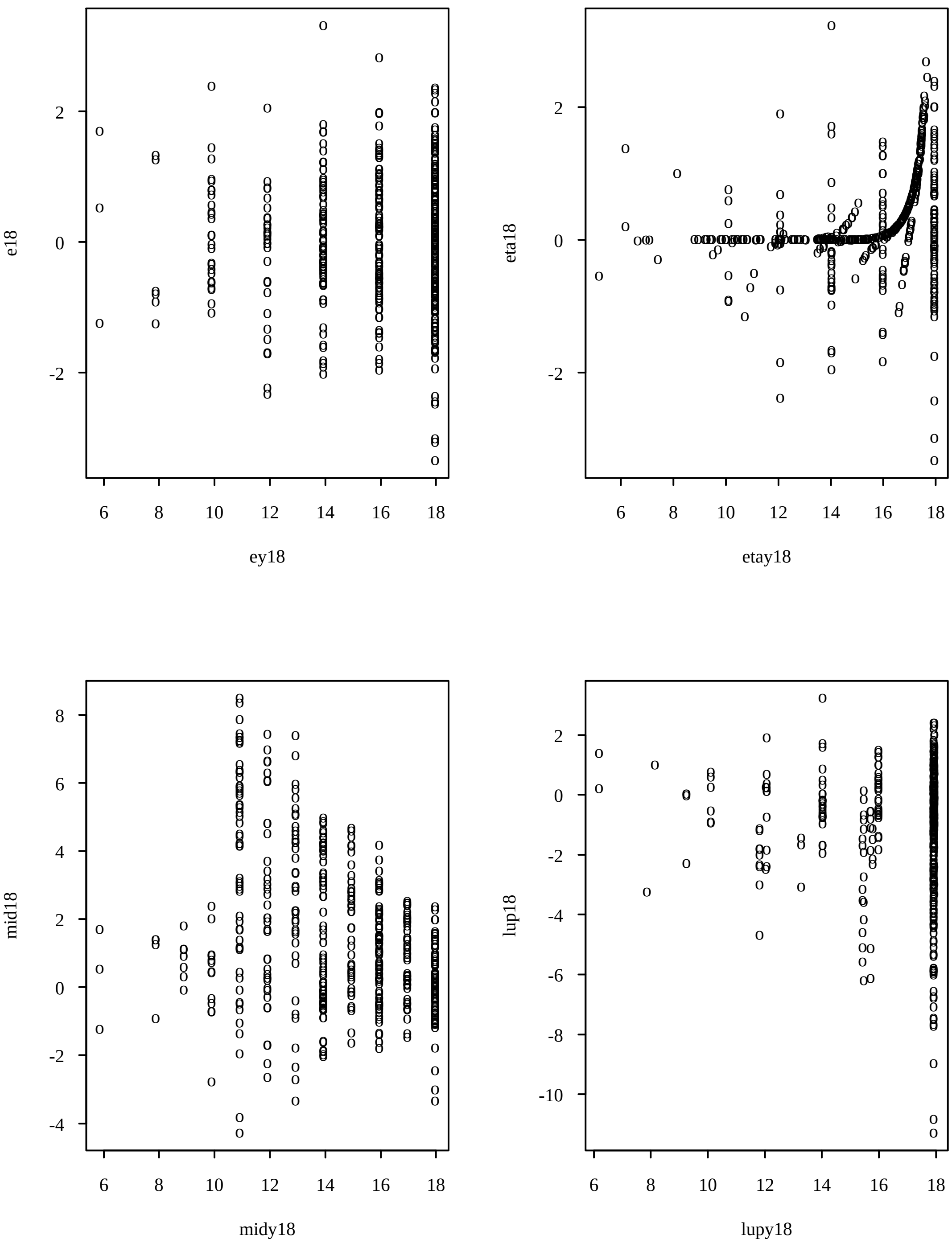
Scenario 16: covariate distribution $W(\frac{1}{6},\frac{3}{2})$, β=5, p=0.7, n=500:



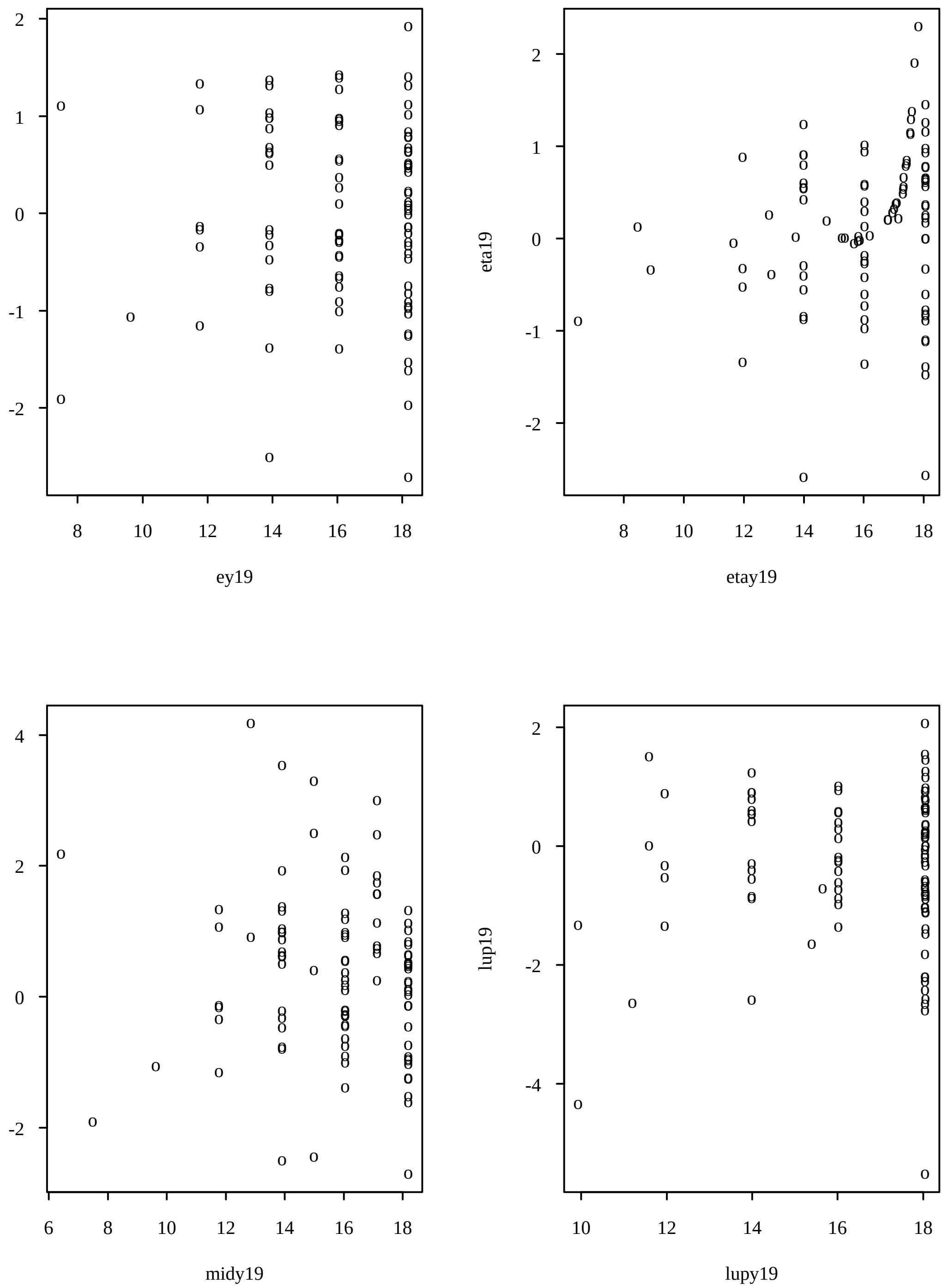
Scenario 17: covariate distribution N(4,4), β=2, p=0.3, n=100:



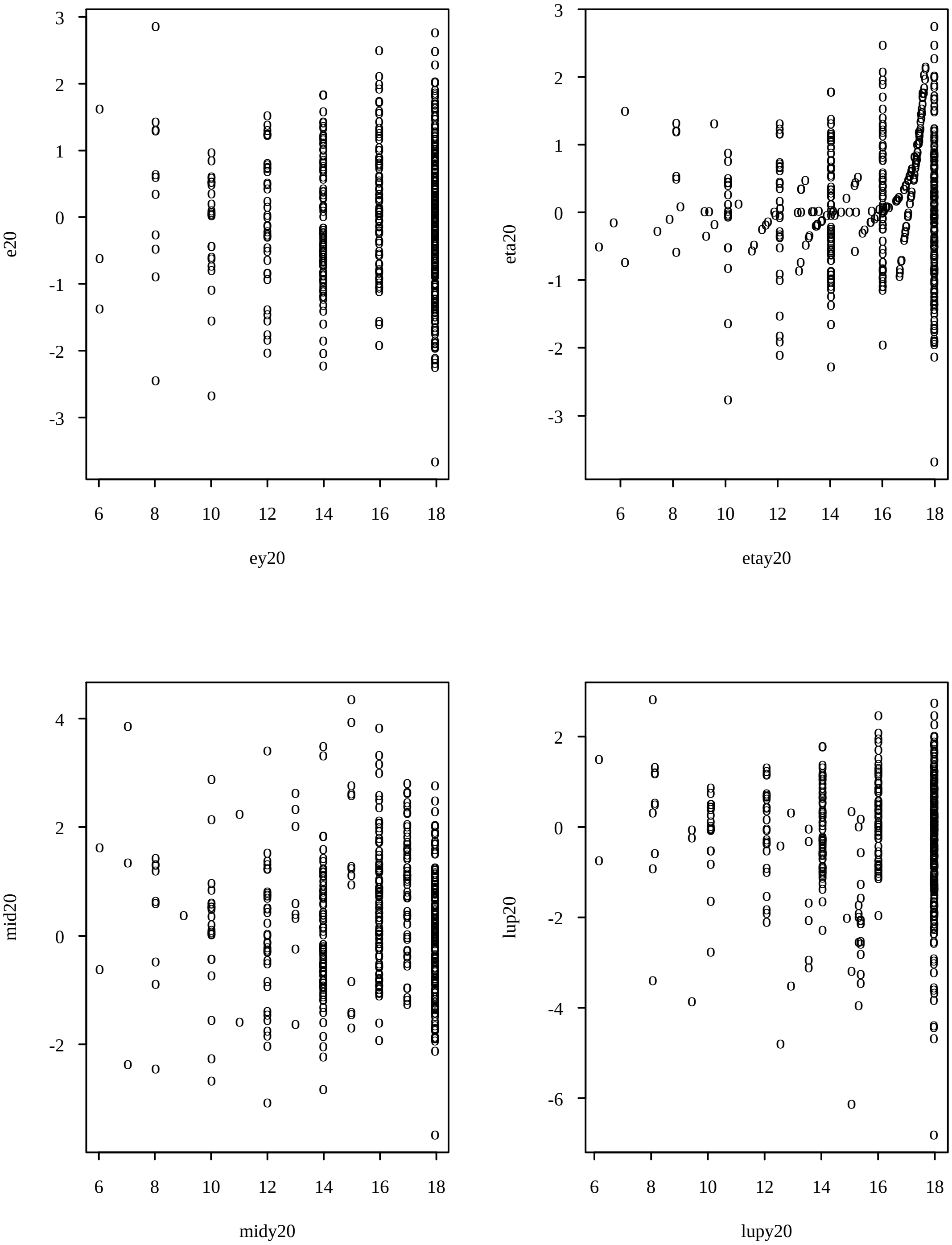
Scenario 18: covariate distribution N(4,4), β=2, p=0.3, n=500:



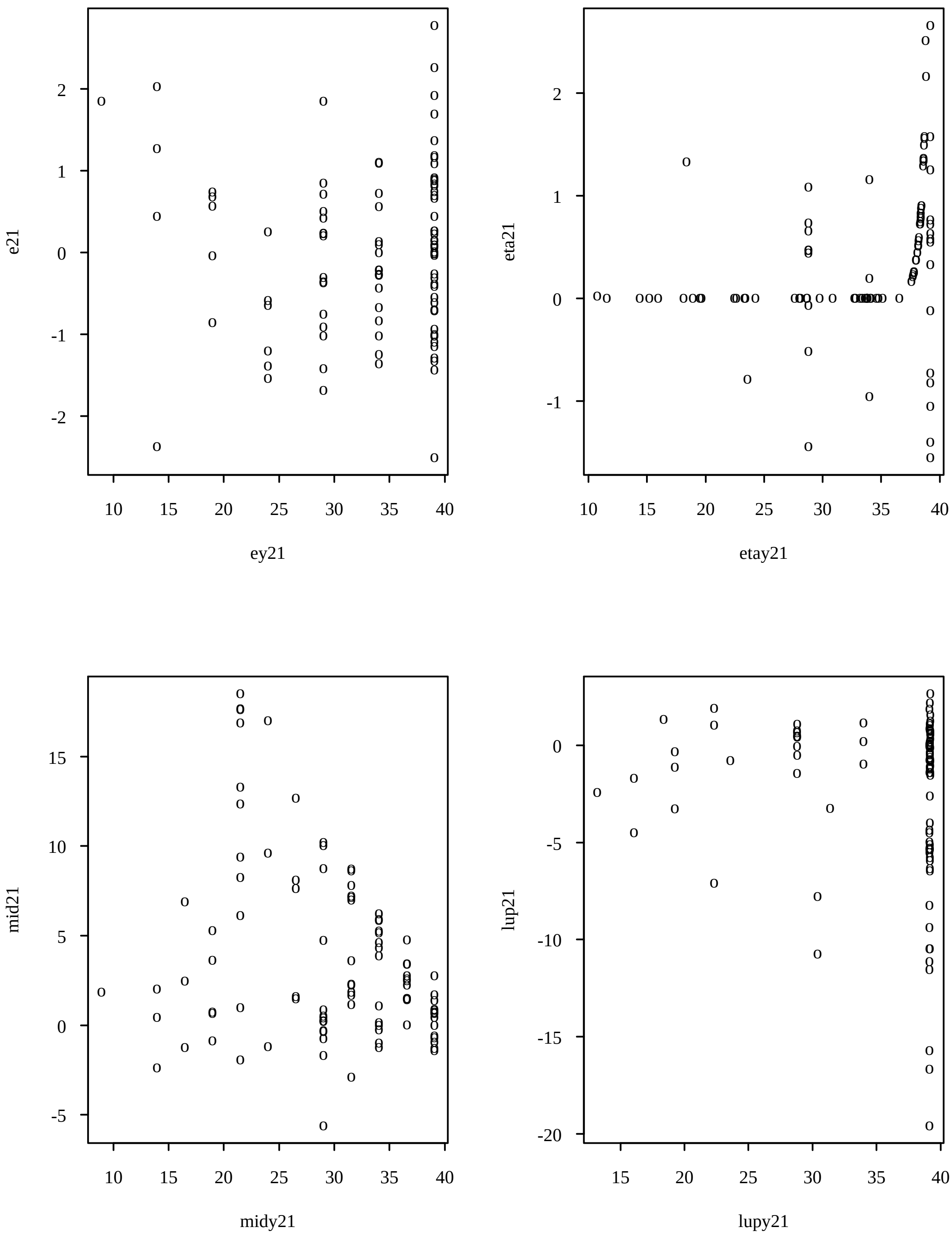
Scenario 19: covariate distribution N(4,4), β=2, p=0.7, n=100:



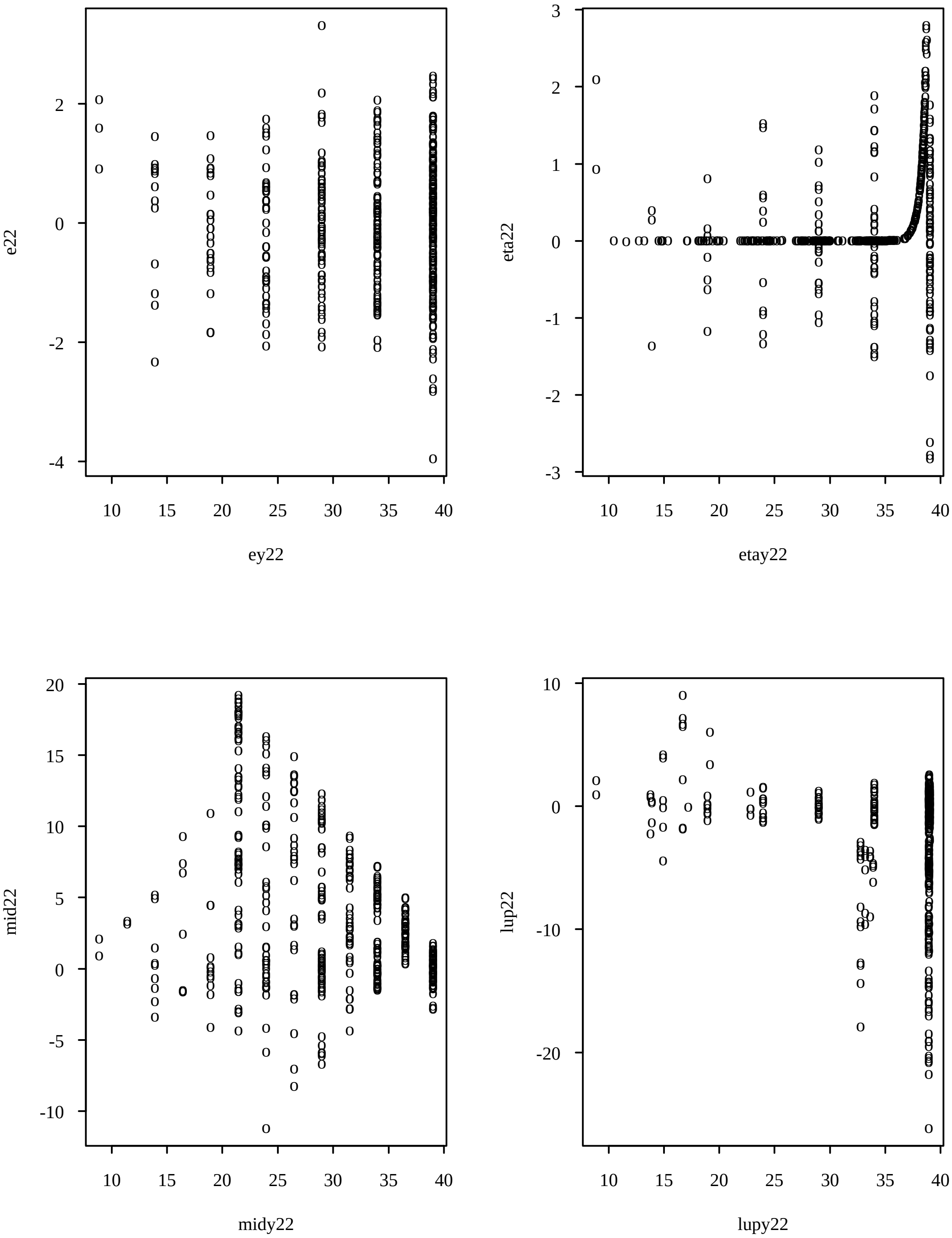
Scenario 20: covariate distribution N(4,4), β=2, p=0.7, n=500:



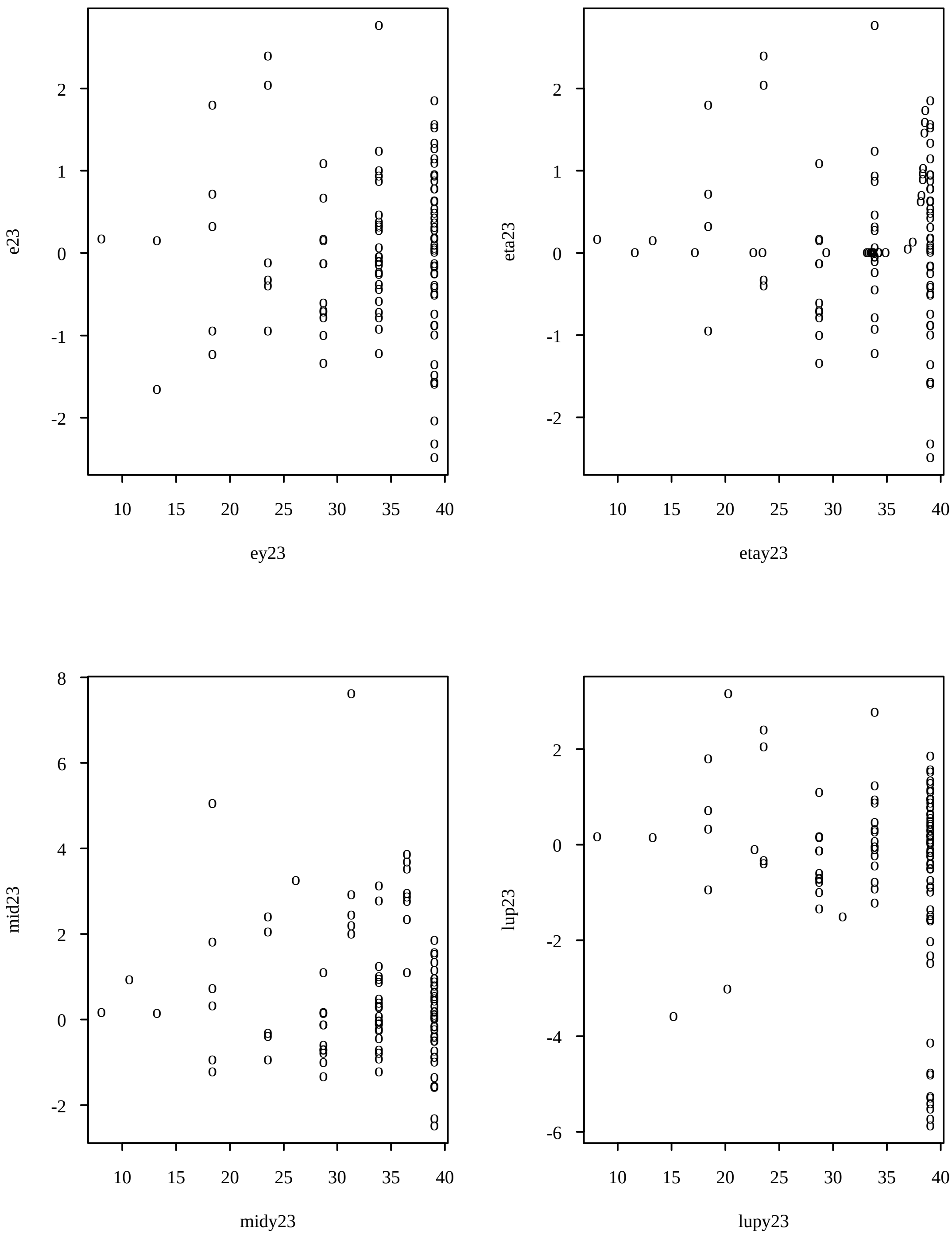
Scenario 21: covariate distribution N(4,4), β=5, p=0.3, n=100:



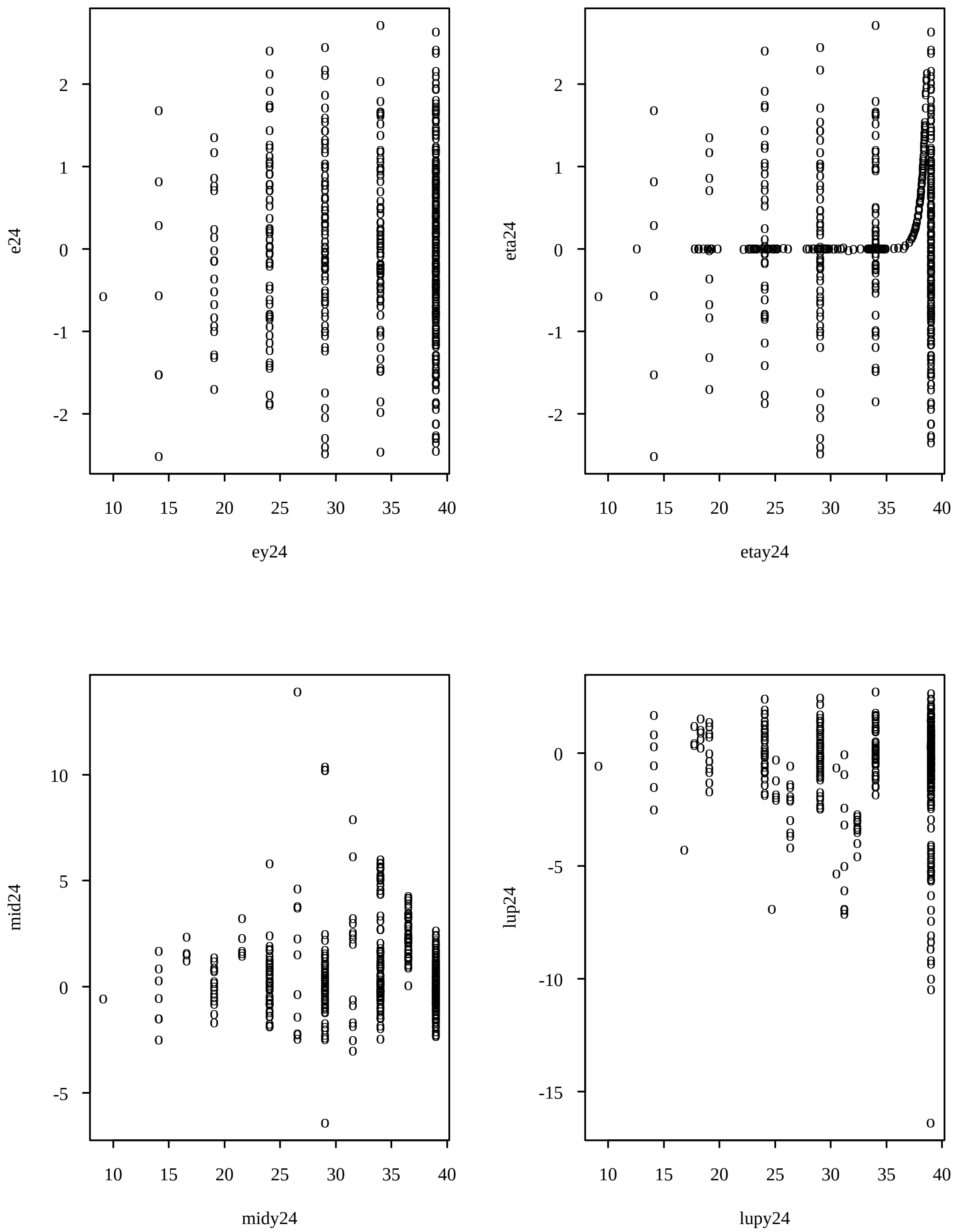
Scenario 22: covariate distribution N(4,4), β=5, p=0.3, n=500:



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



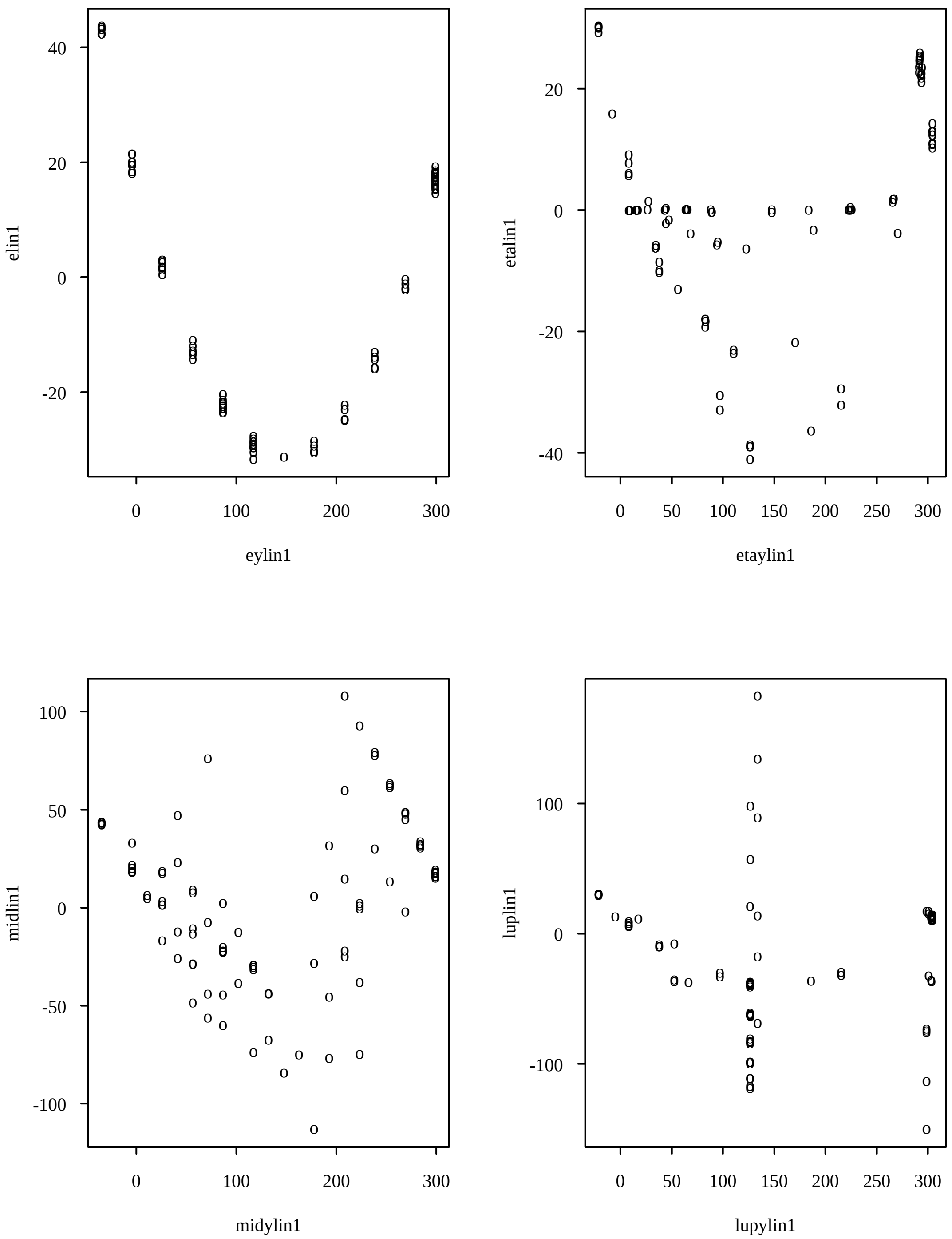
Scenario 24: covariate distribution N(4,4), β=5, p=0.7, n=500:



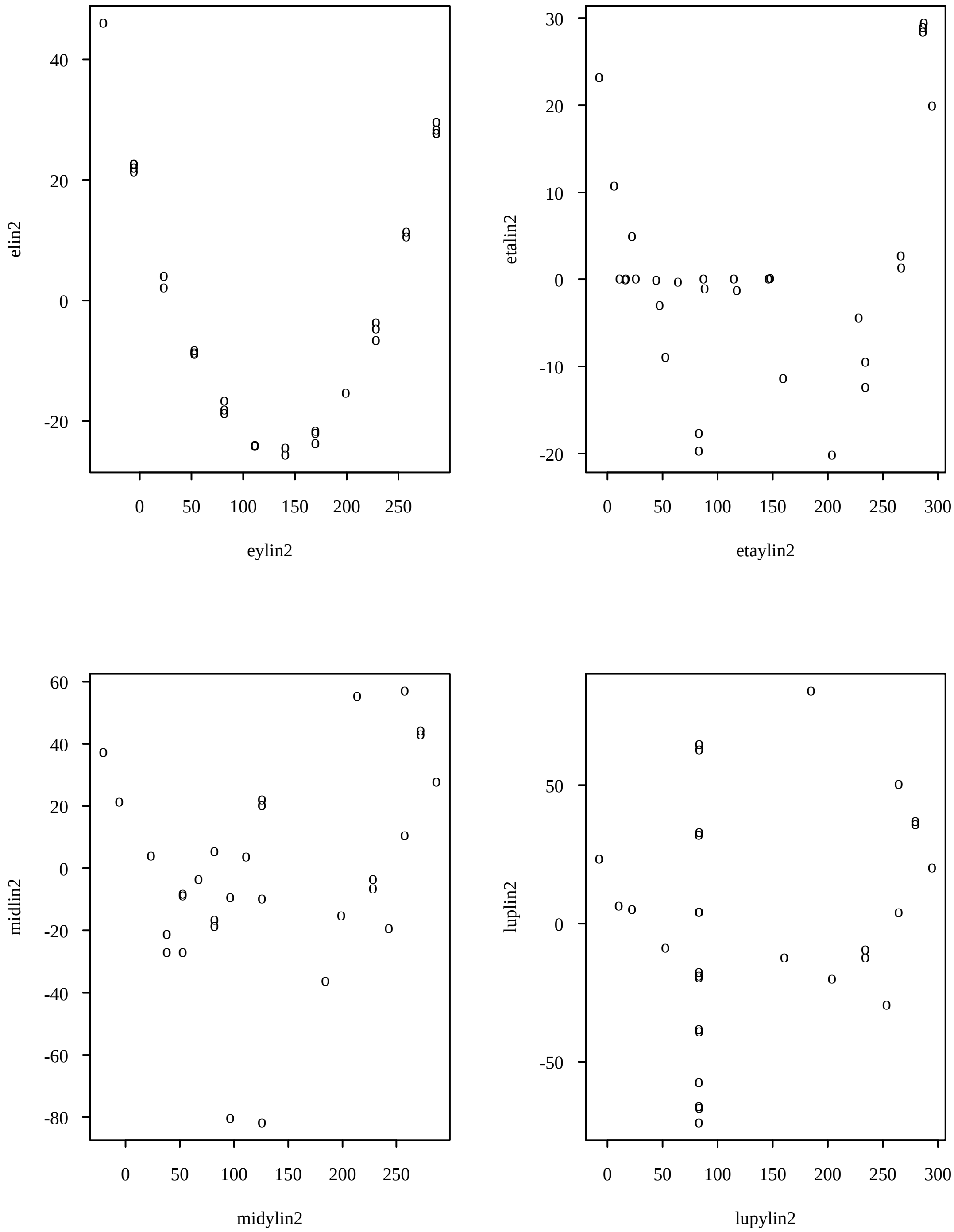
Appendix B

Residual plots when a quadratic term is missing

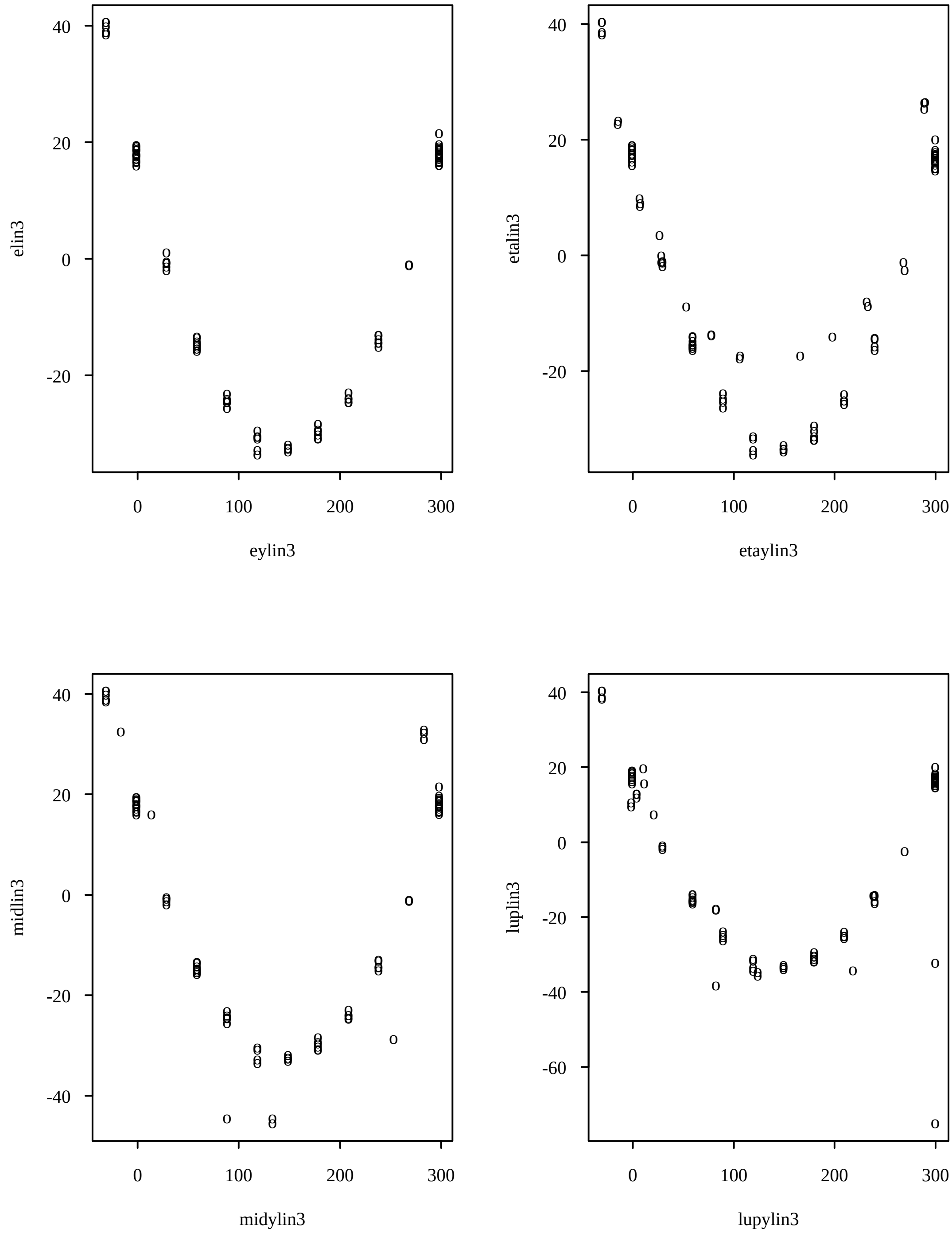
Scenario 1: covariate distribution $\mathrm{Exp}(\frac{1}{8})$, β=2, p=0.3, n=100:



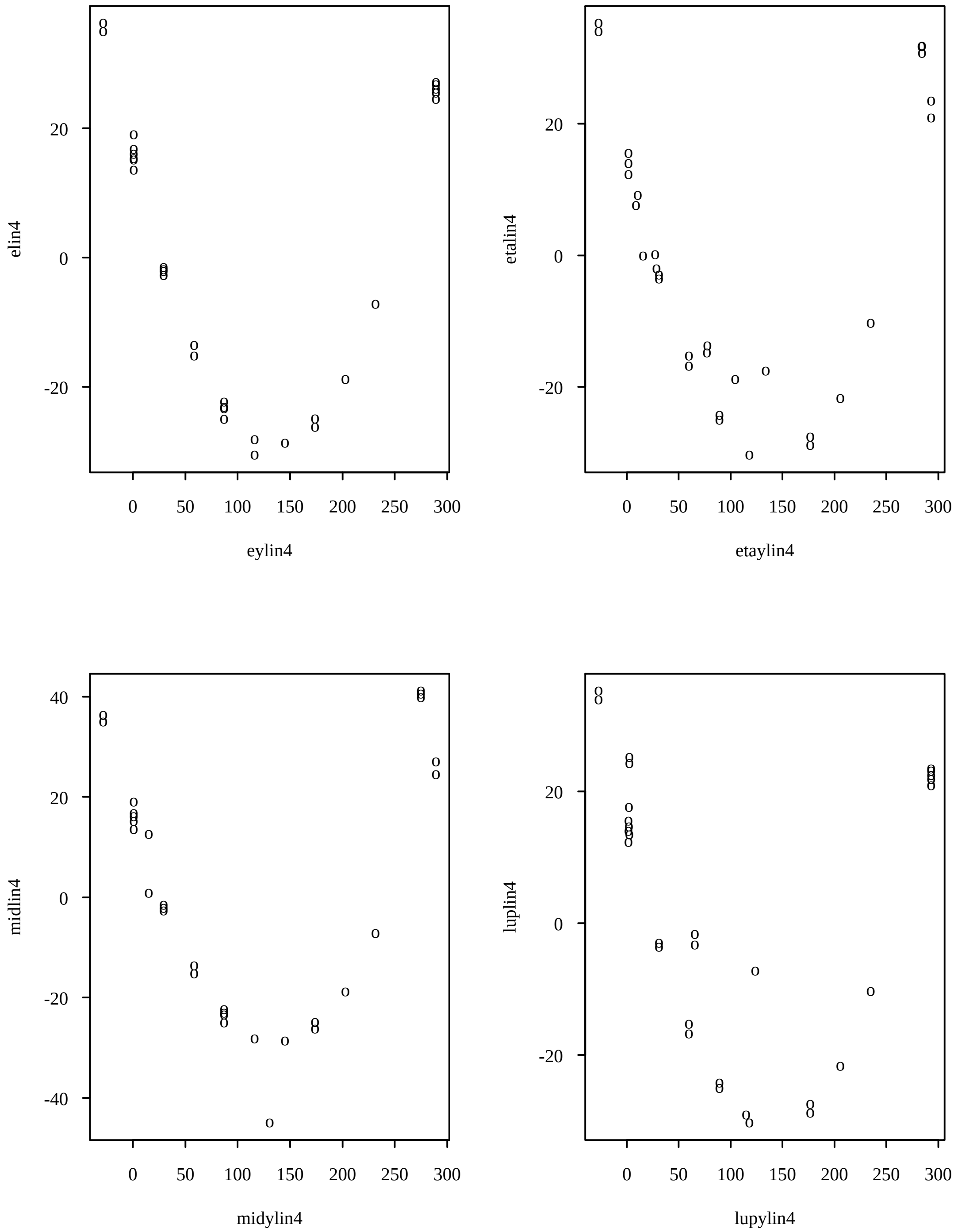
Scenario 2: covariate distribution Exp($\frac{1}{8}$), β=2, p=0.3, n=30:



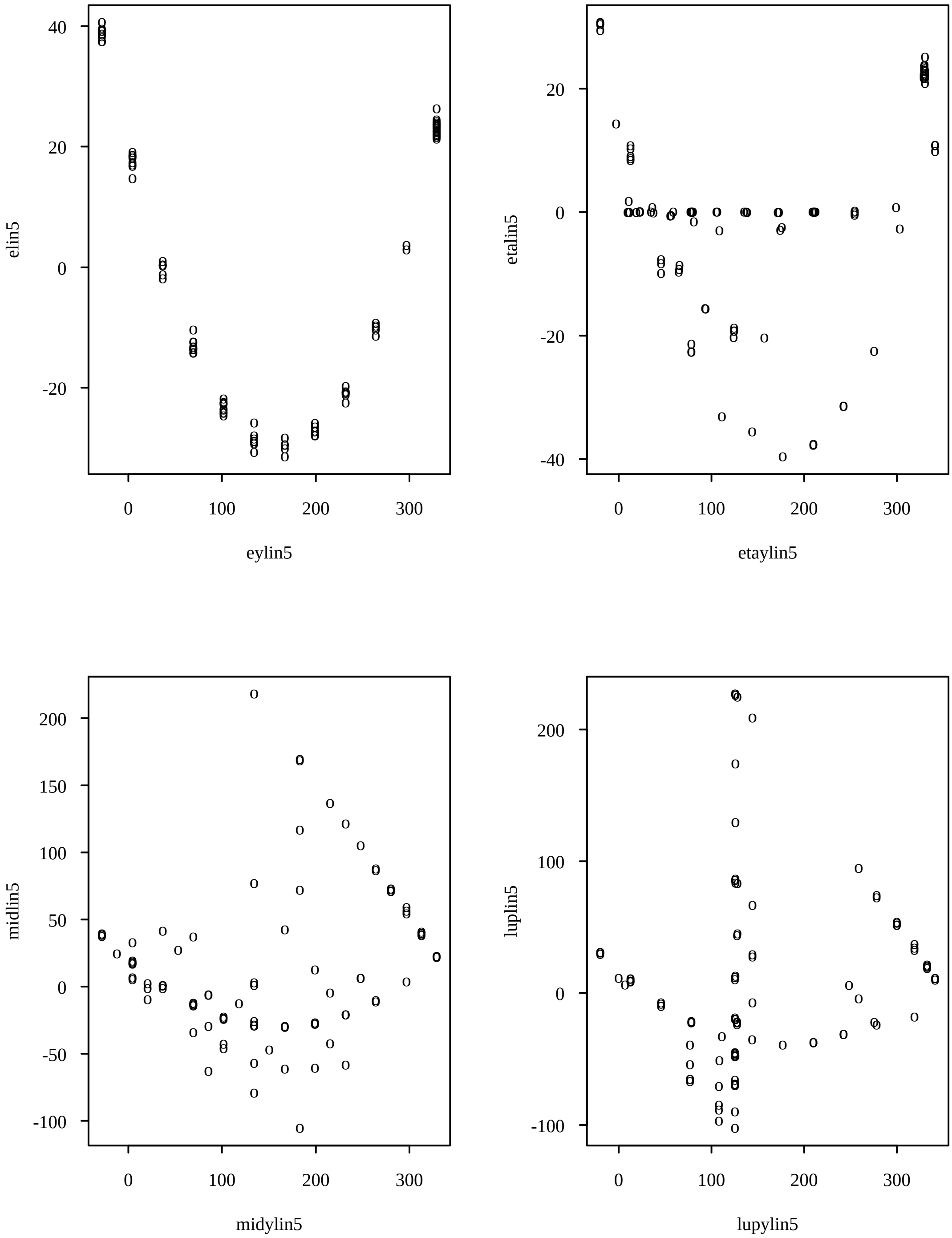
Scenario 3: covariate distribution $\mathrm{Exp}(\frac{1}{8})$, β=2, p=0.7, n=100:



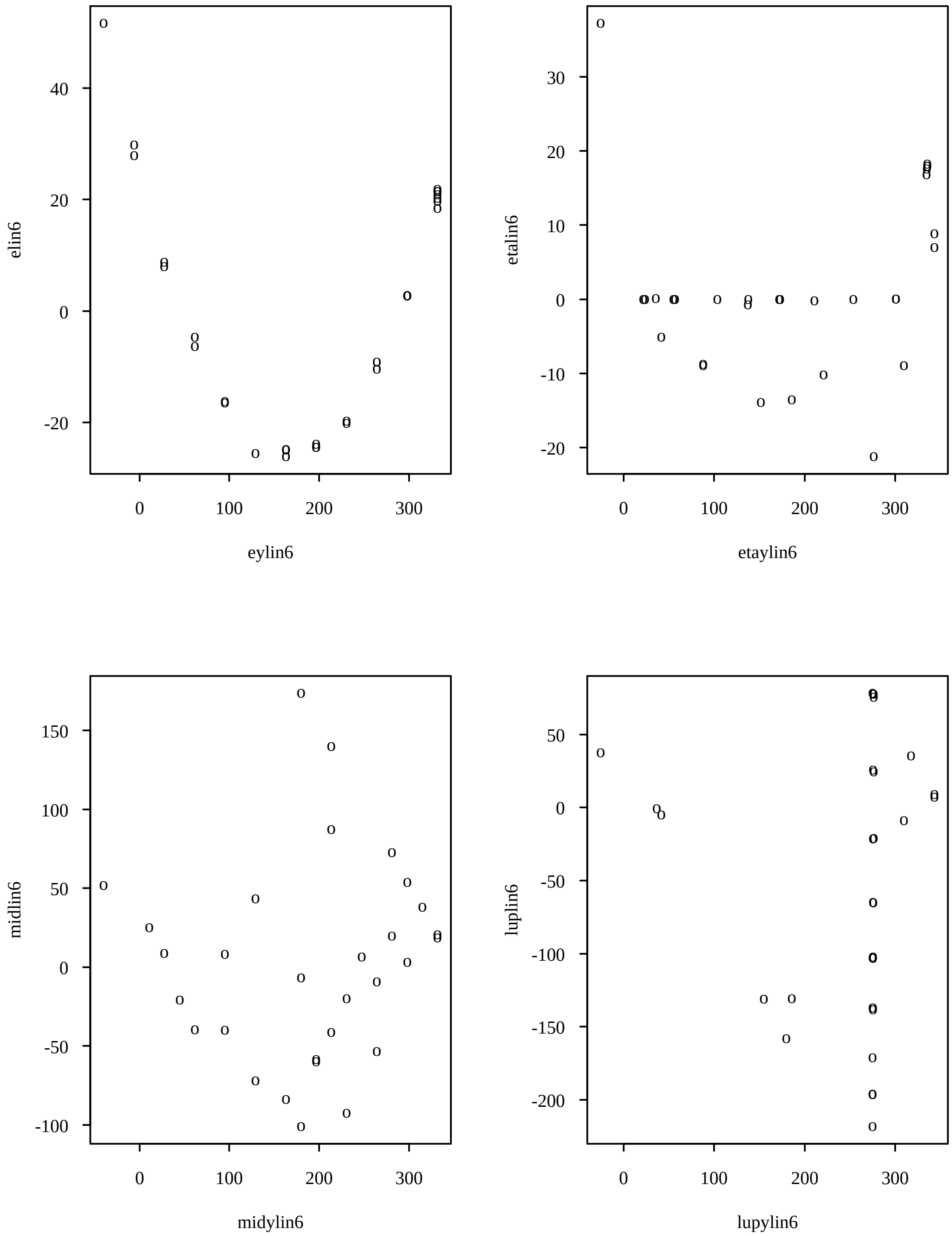
Scenario 4: covariate distribution Exp($\frac{1}{8}$), β=2, p=0.7, n=30:



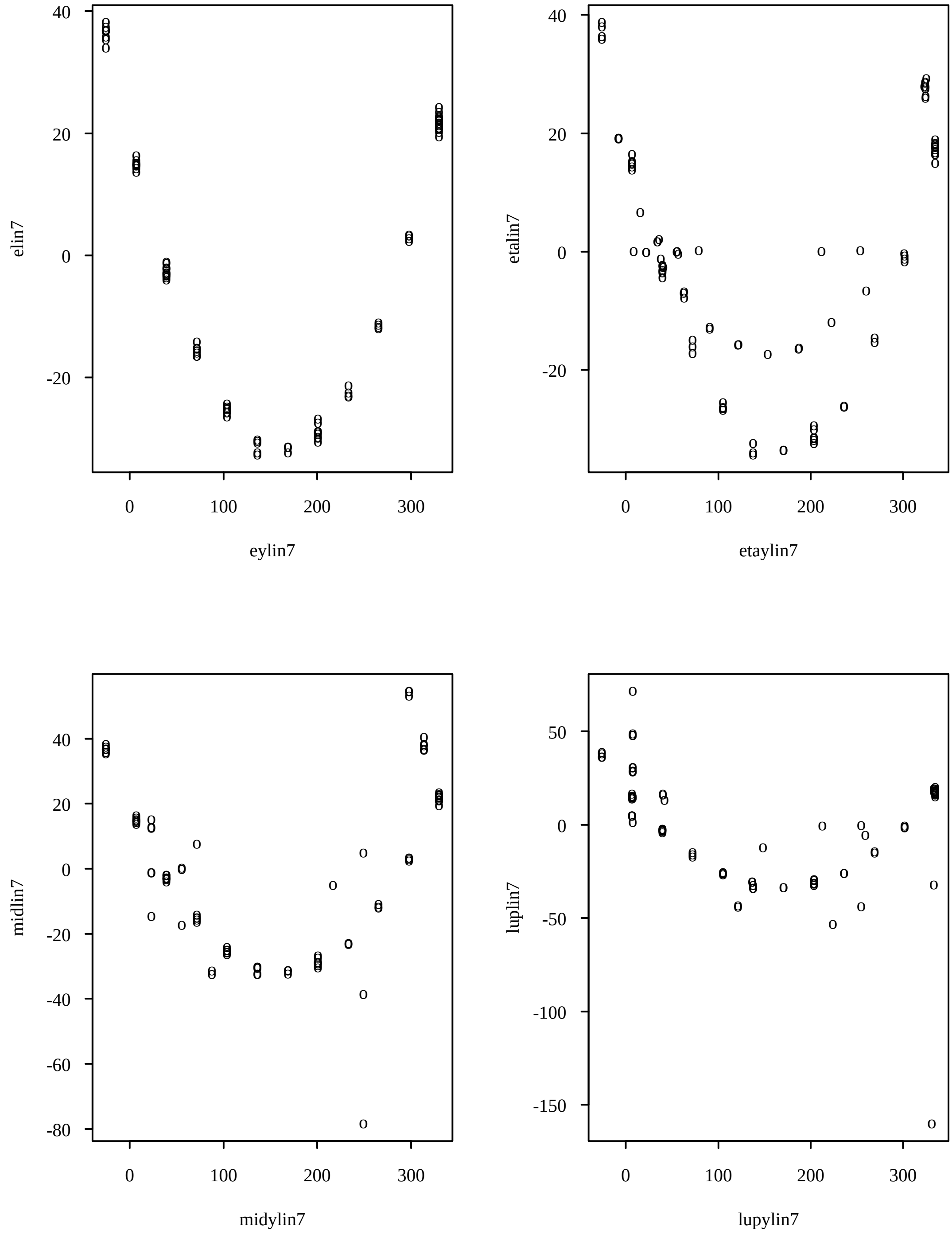
Scenario 5: covariate distribution $\mathrm{Exp}(\frac{1}{8})$, β=5, p=0.3, n=100:



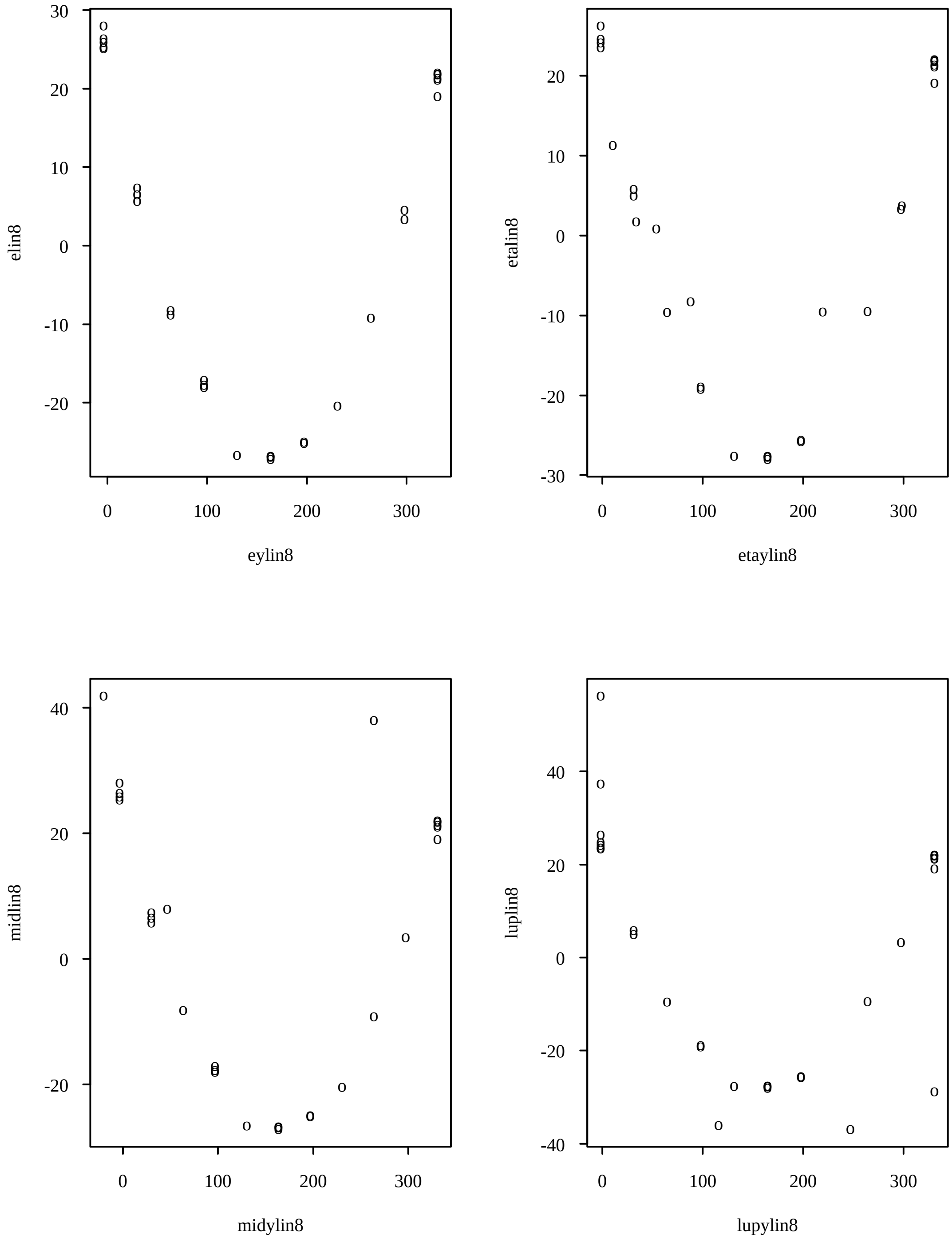
Scenario 6: covariate distribution Exp($\frac{1}{8}$), β=5, p=0.3, n=30:



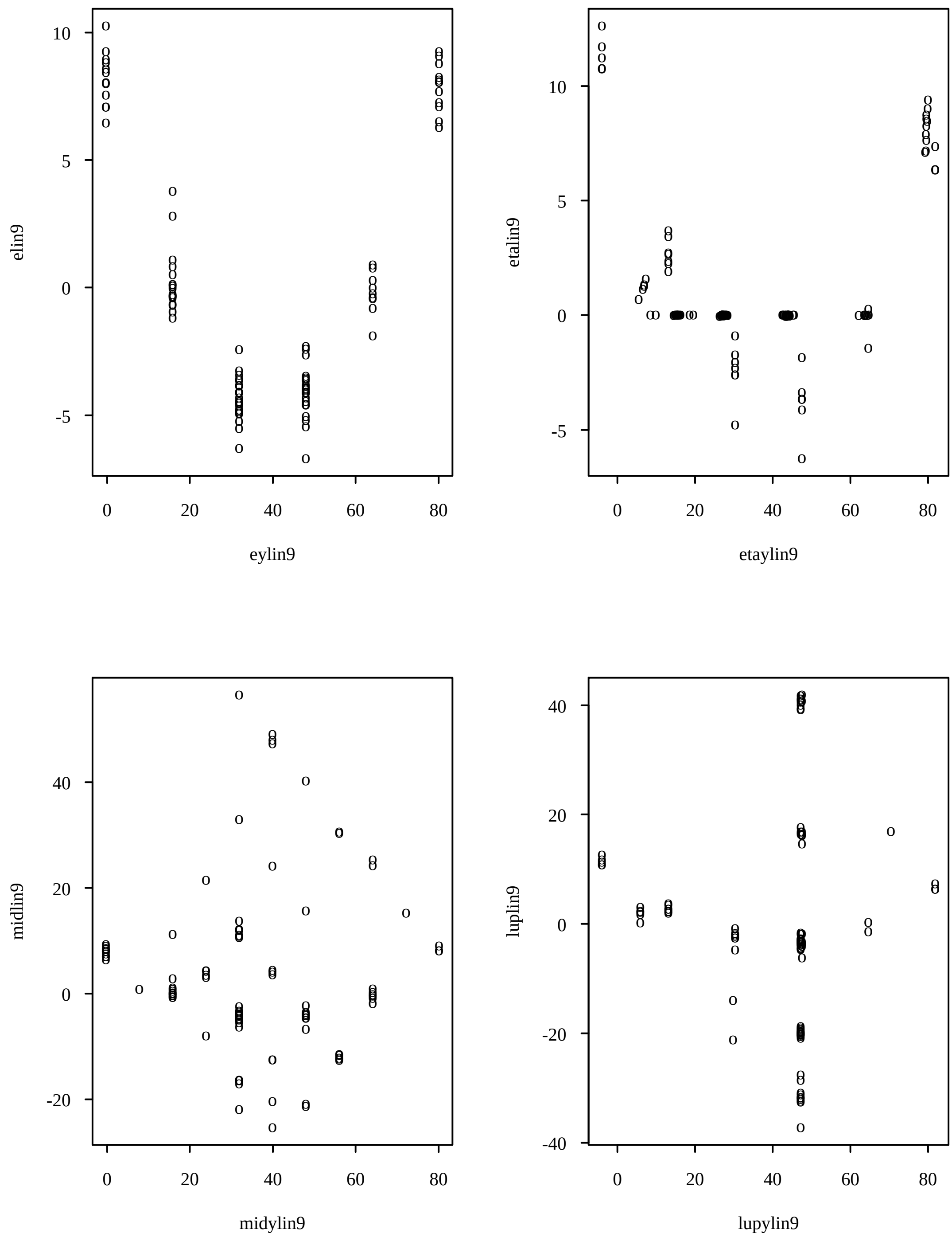
Scenario 7: covariate distribution $\mathrm{Exp}(\frac{1}{8})$, β=5, p=0.7, n=100:



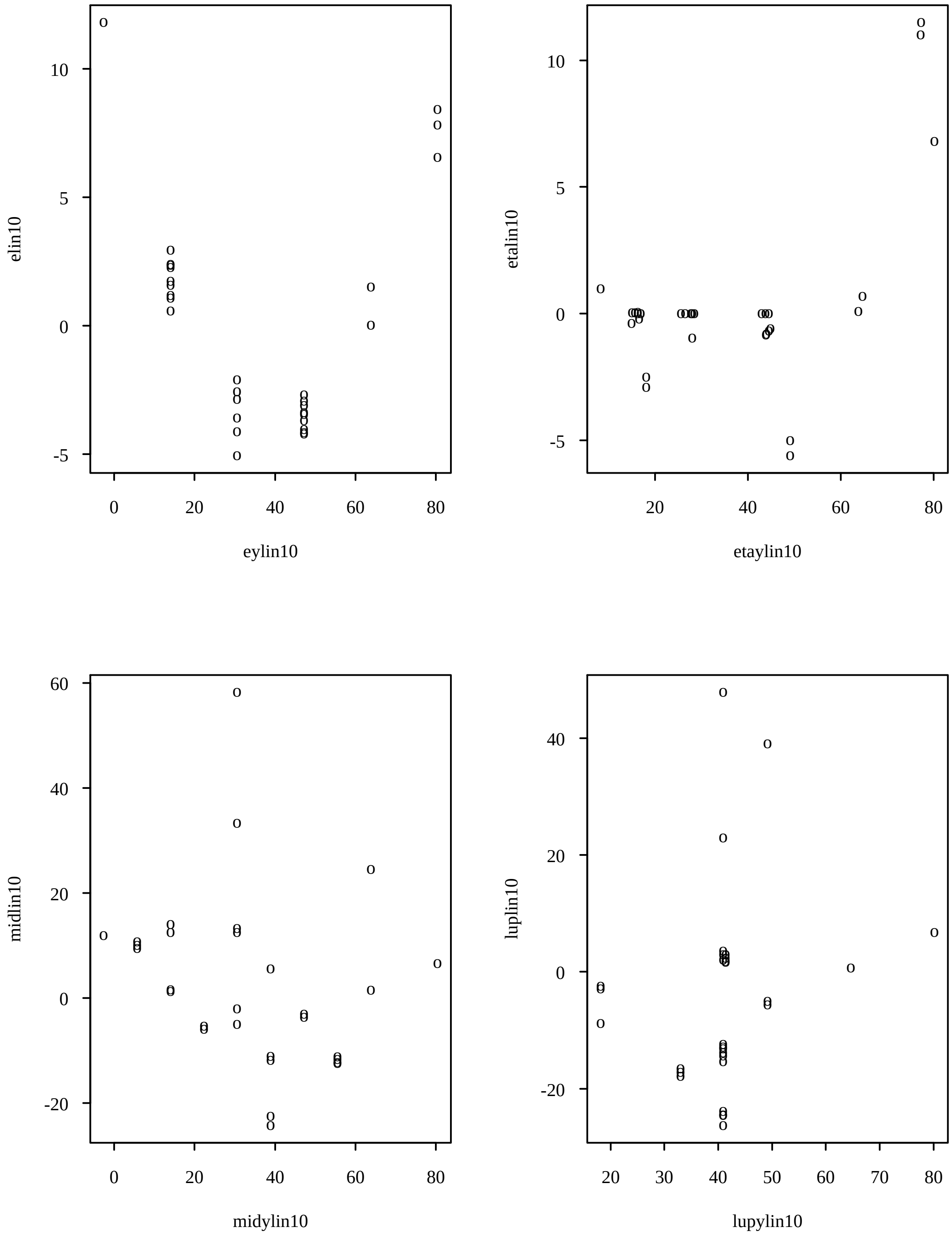
Scenario 8: covariate distribution $\mathrm{Exp}(\frac{1}{8})$, β=5, p=0.7, n=30:



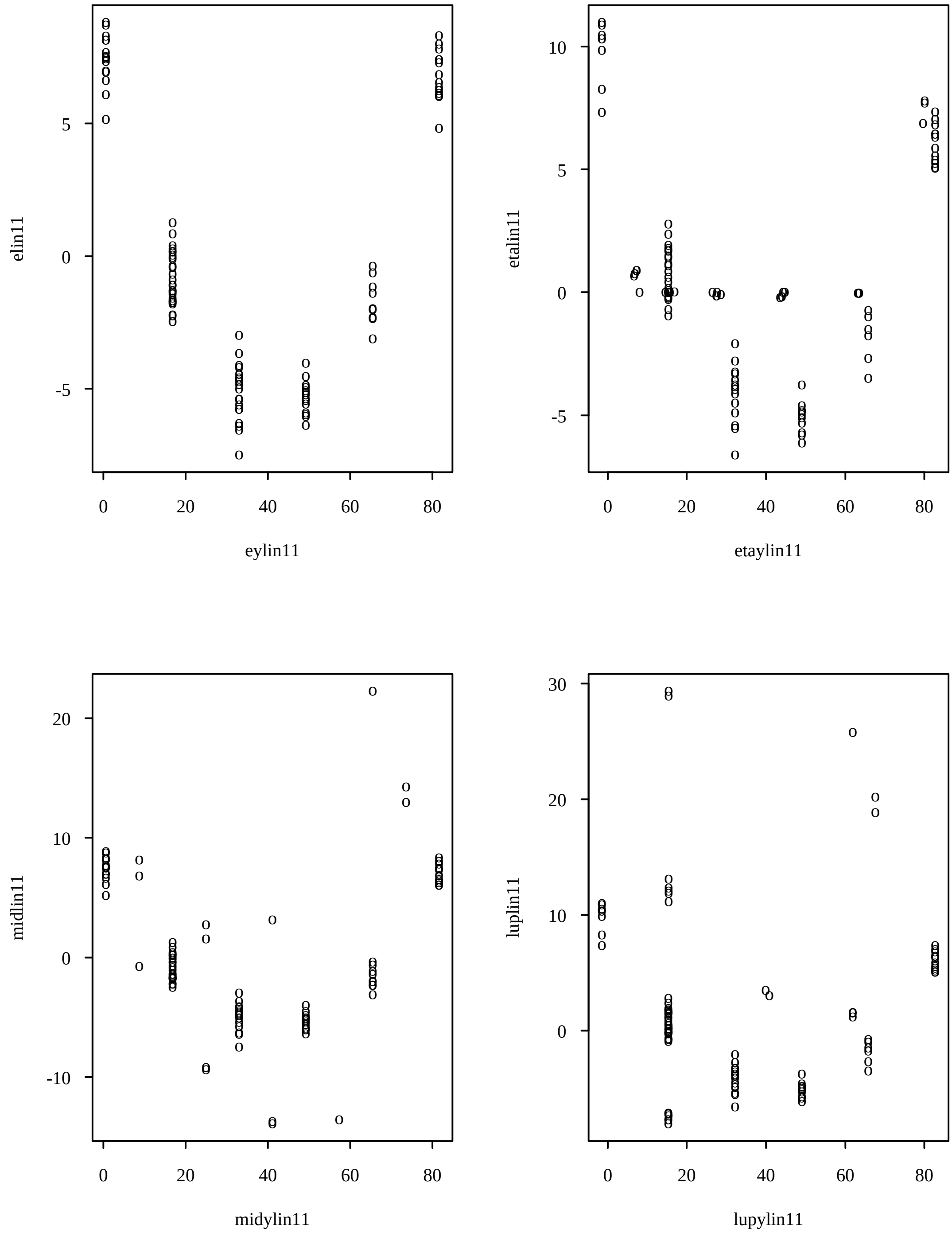
Scenario 9: covariate distribution $W(\frac{1}{6}, \frac{3}{2})$, β=2, p=0.3, n=100:



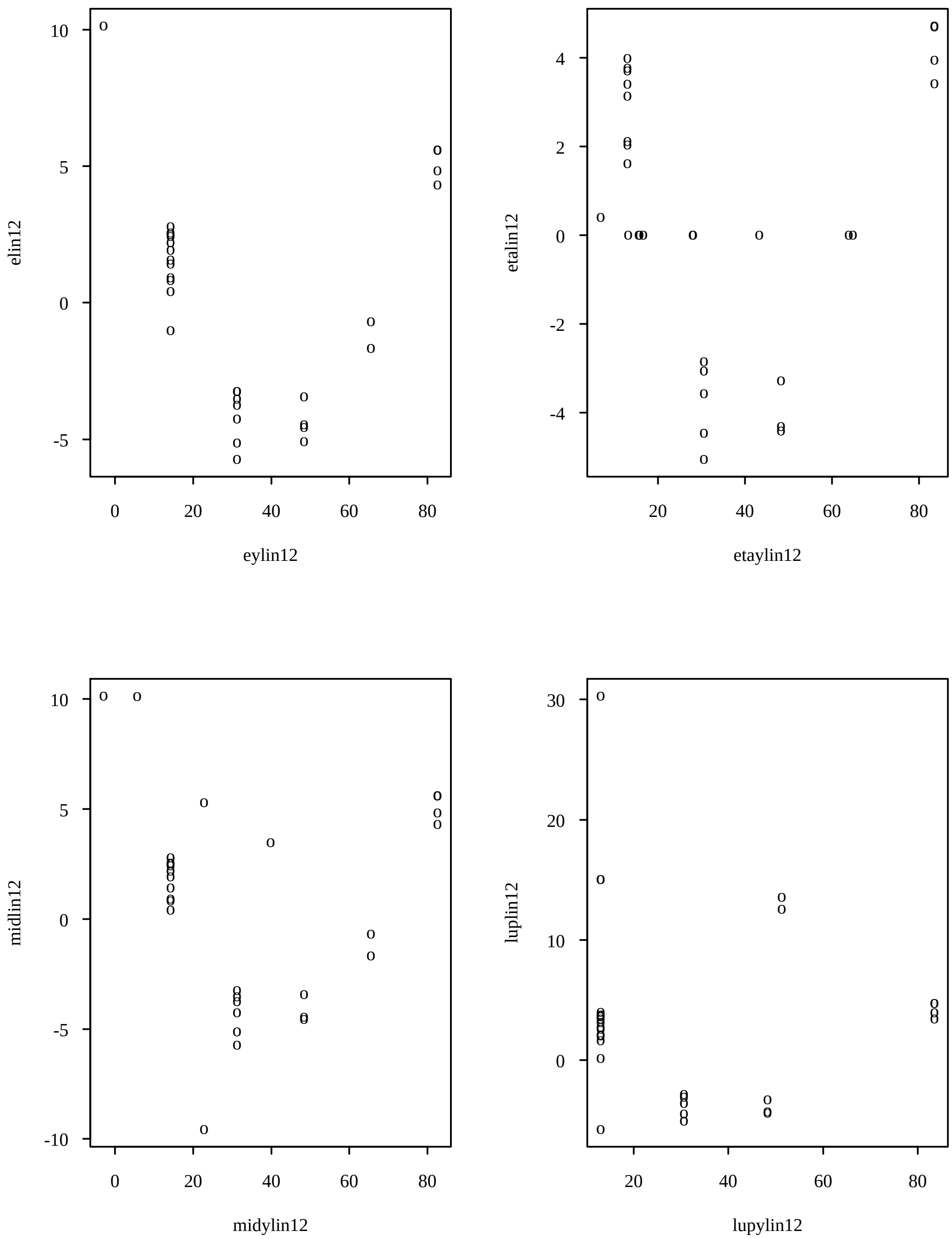
Scenario 10: covariate distribution W($\frac{1}{6}$, $\frac{3}{2}$), β=2, p=0.3, n=30:



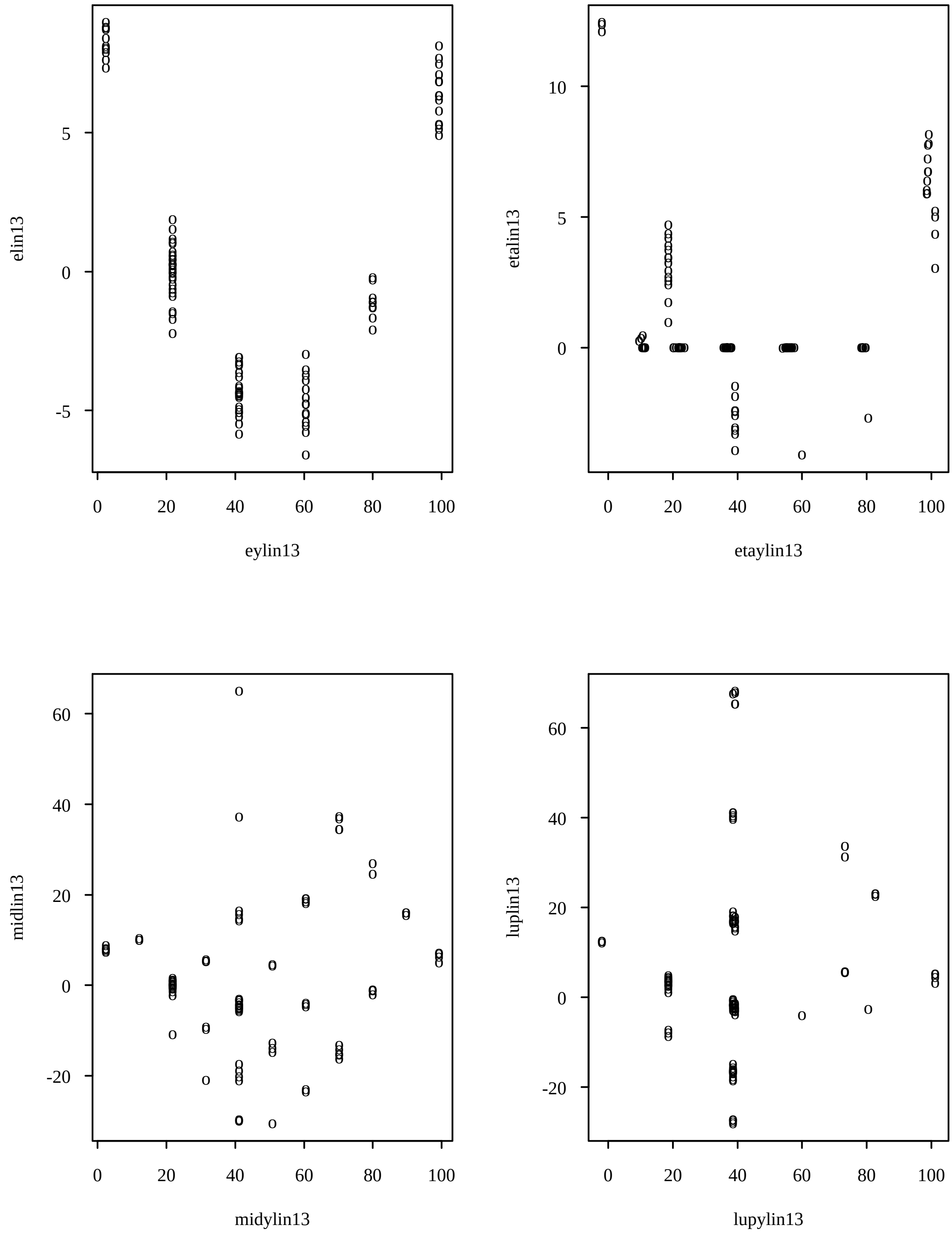
Scenario 11: covariate distribution W($\frac{1}{6}$, $\frac{3}{2}$), β=2, p=0.7, n=100:



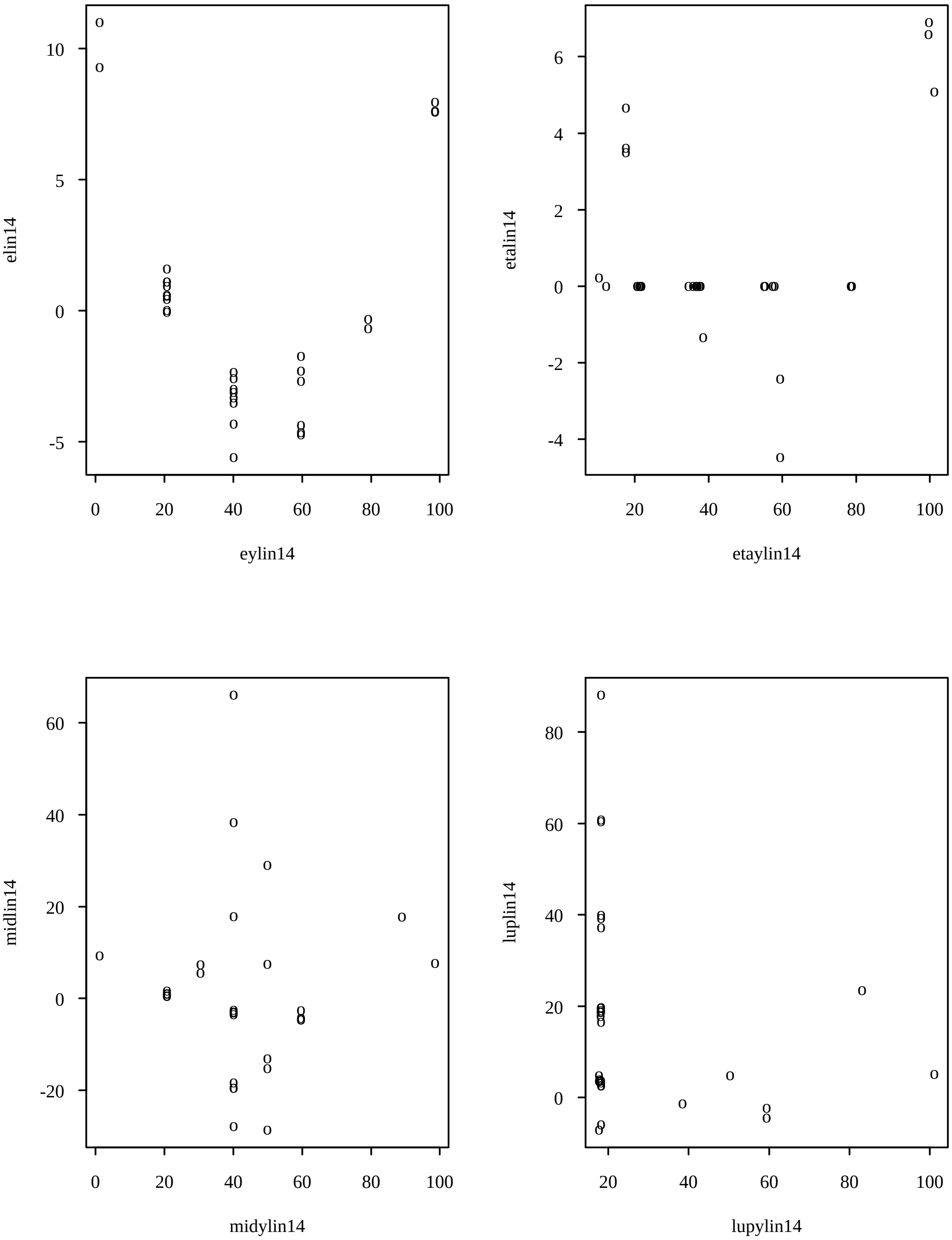
Scenario 12: covariate distribution W($\frac{1}{6}$, $\frac{3}{2}$), β=2, p=0.7, n=30:



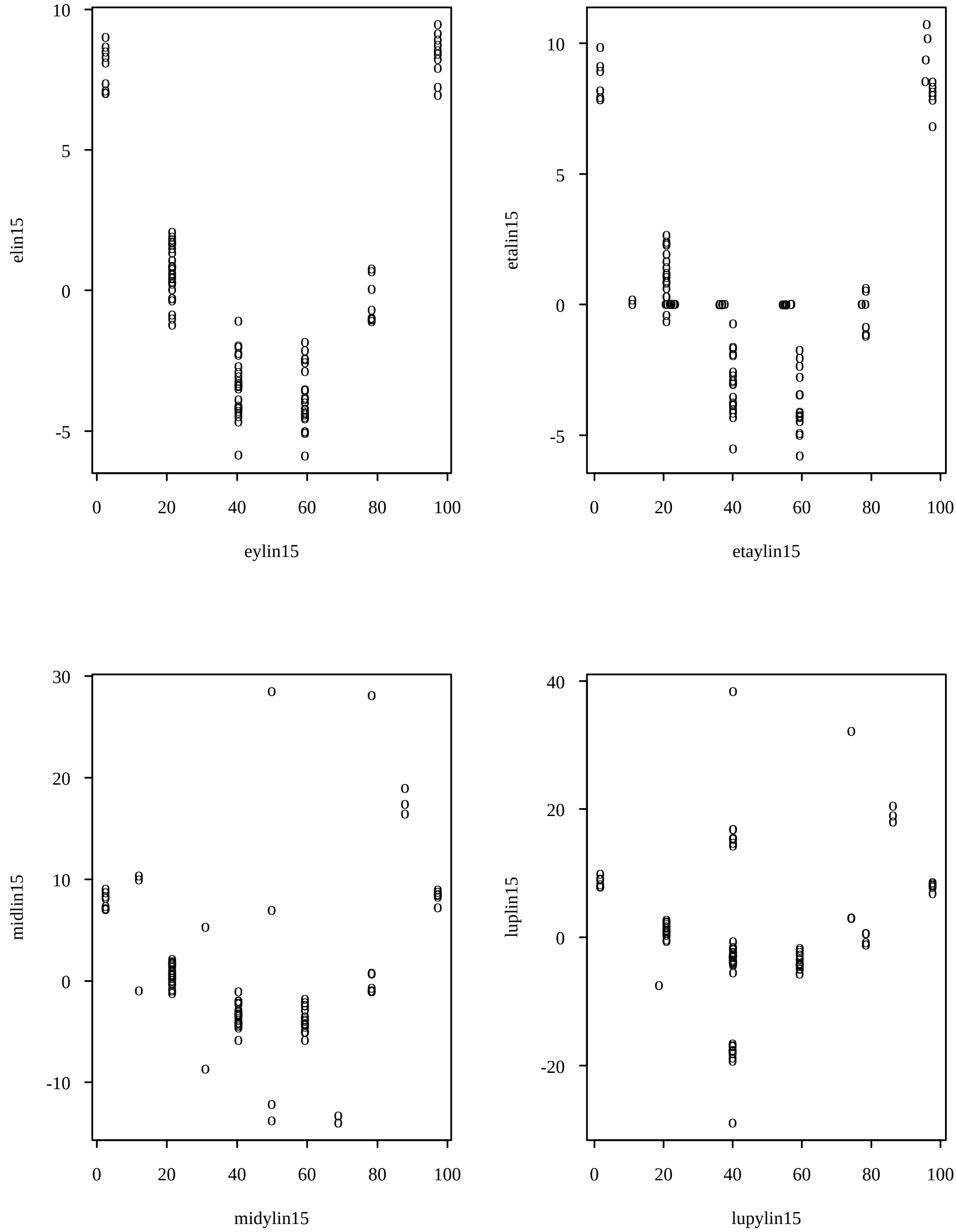
Scenario 13: covariate distribution W($\frac{1}{6}$, $\frac{3}{2}$), β=5, p=0.3, n=100:



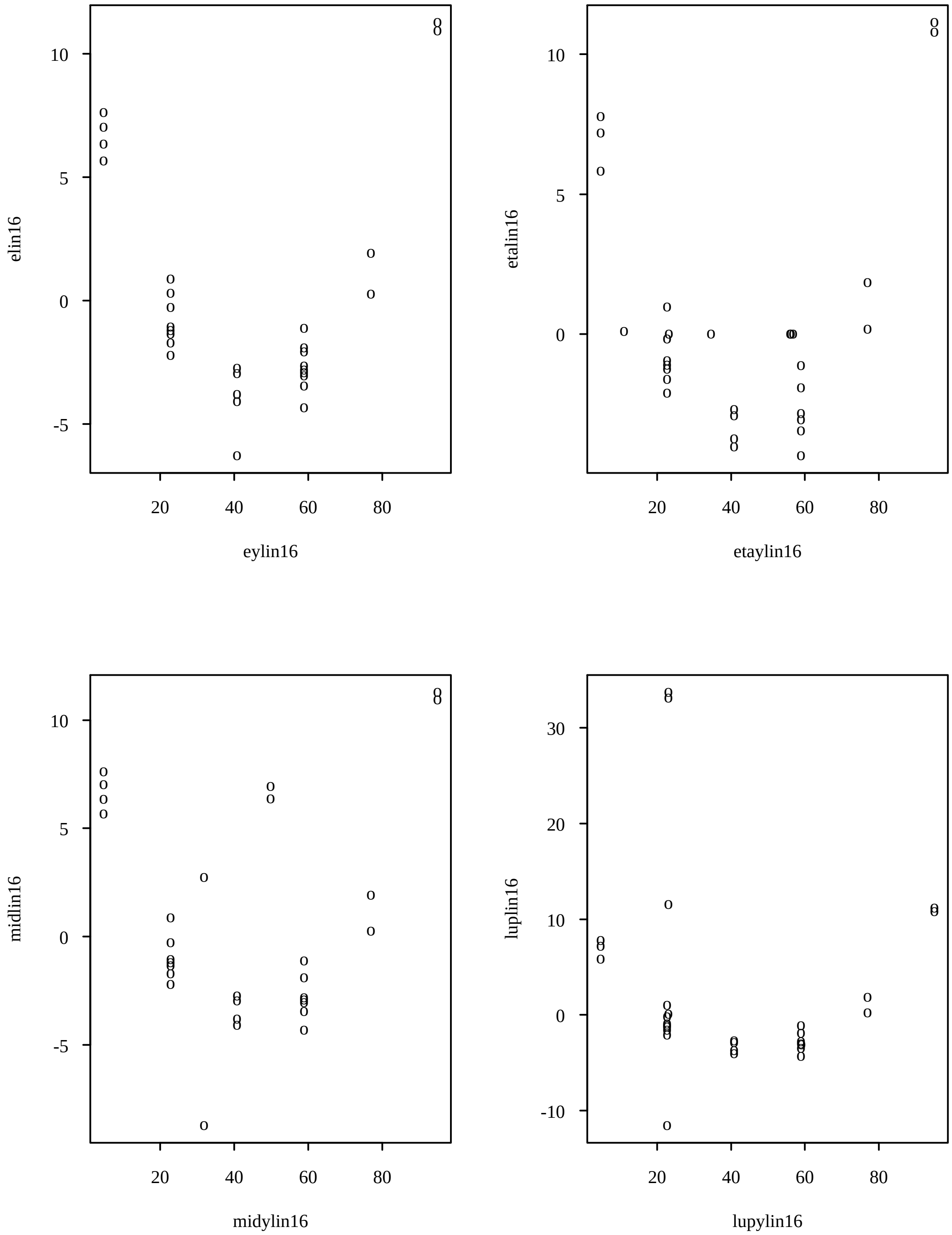
Scenario 14: covariate distribution $W(\frac{1}{6}, \frac{3}{2})$, β=5, p=0.3, n=30:



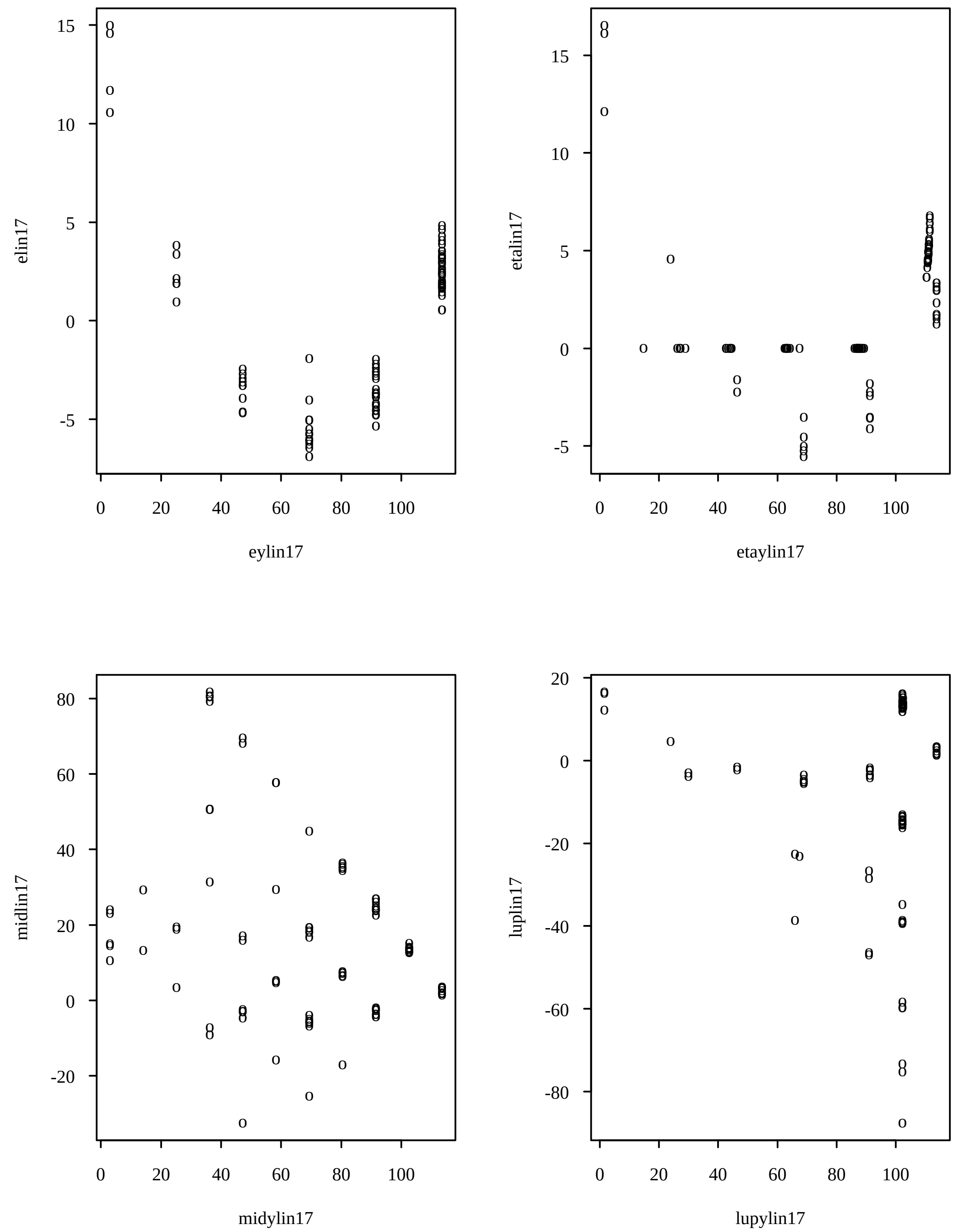
Scenario 15: covariate distribution $W(\frac{1}{6}, \frac{3}{2})$, β=5, p=0.7, n=100:



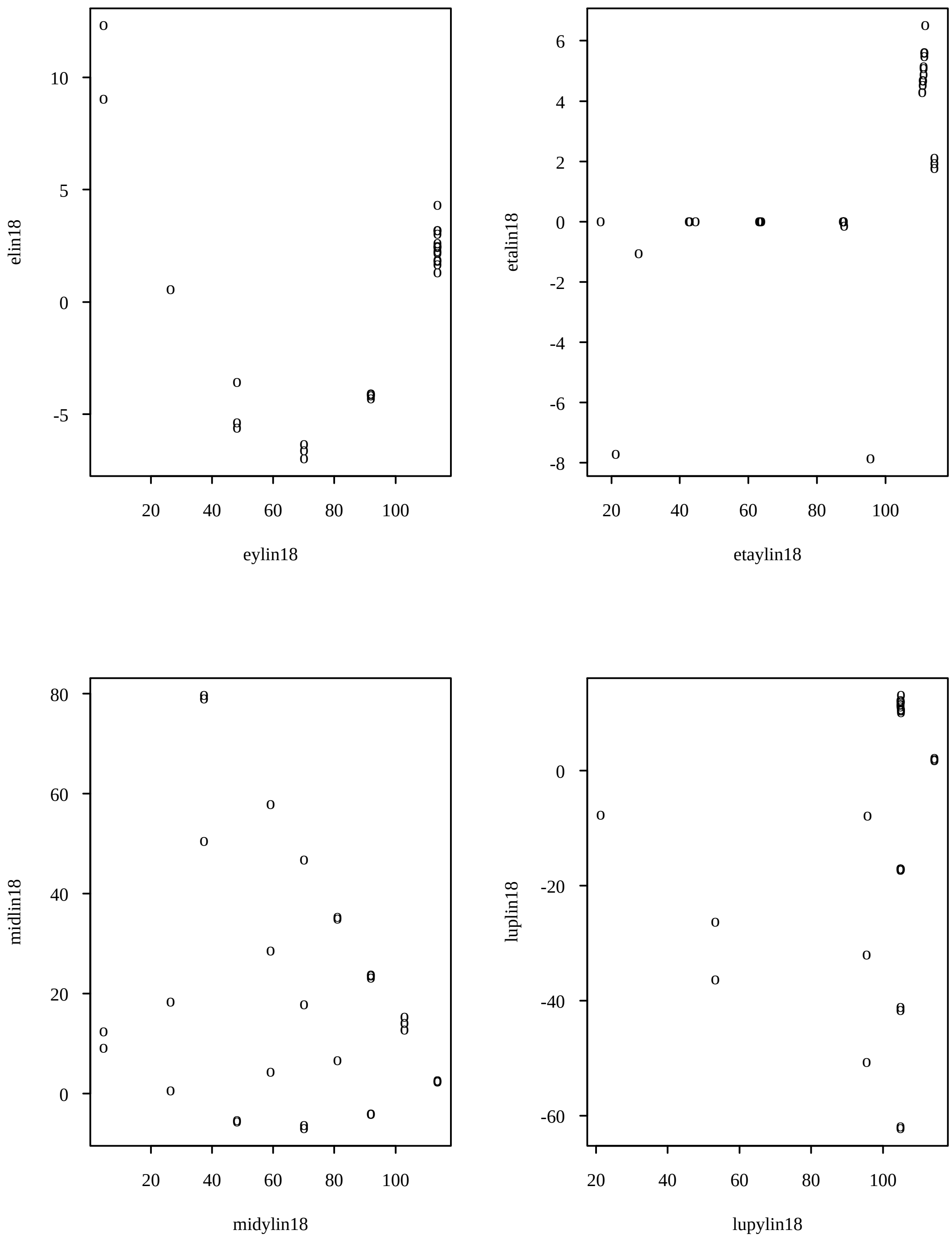
Scenario 16: covariate distribution $W(\frac{1}{6}, \frac{3}{2})$, β=5, p=0.7, n=30:



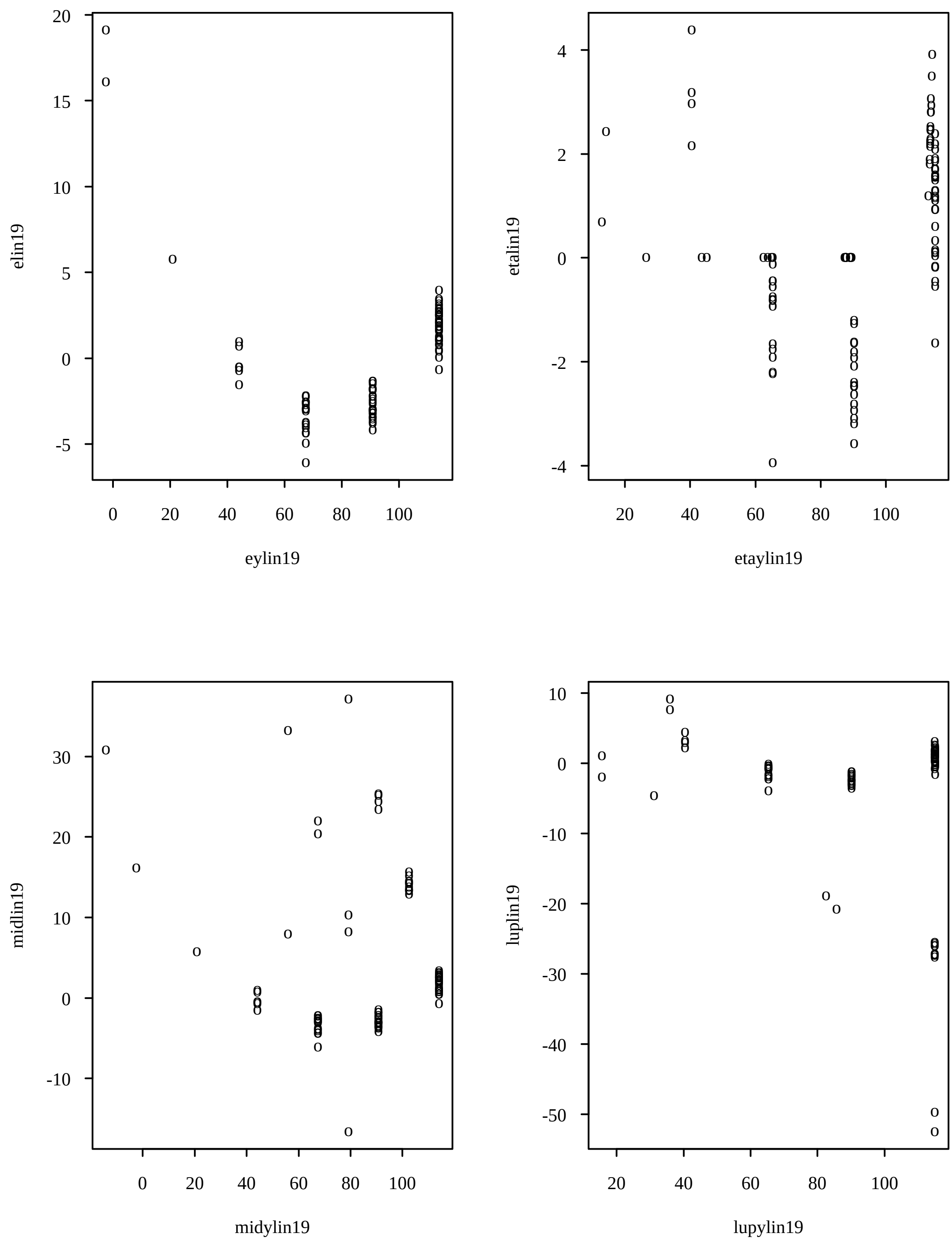
Scenario 17: covariate distribution N(6,4), β=2, p=0.3, n=100:



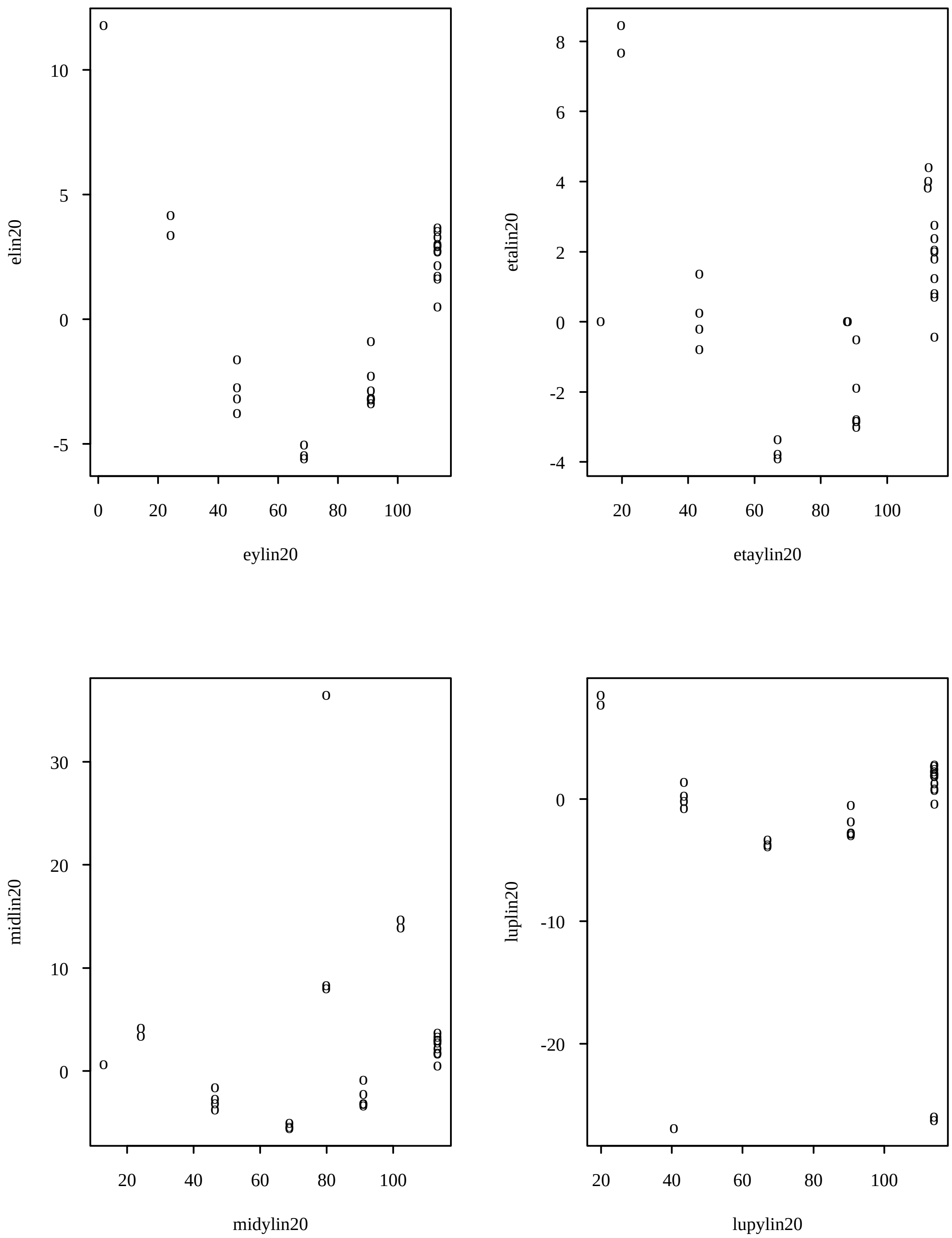
Scenario 18: covariate distribution N(6,4), β=2, p=0.3, n=30:



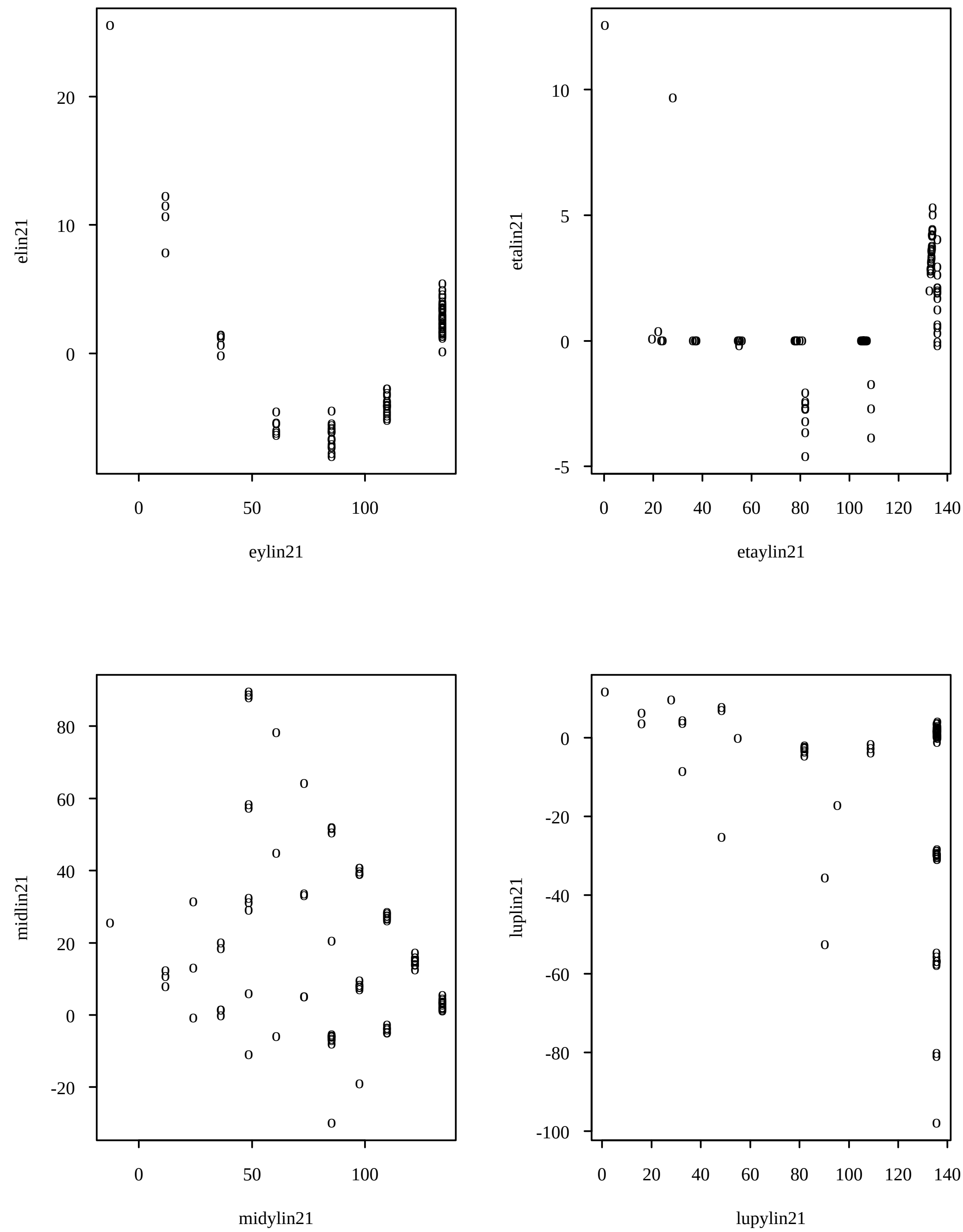
Scenario 19: covariate distribution N(6,4), β=2, p=0.7, n=100:



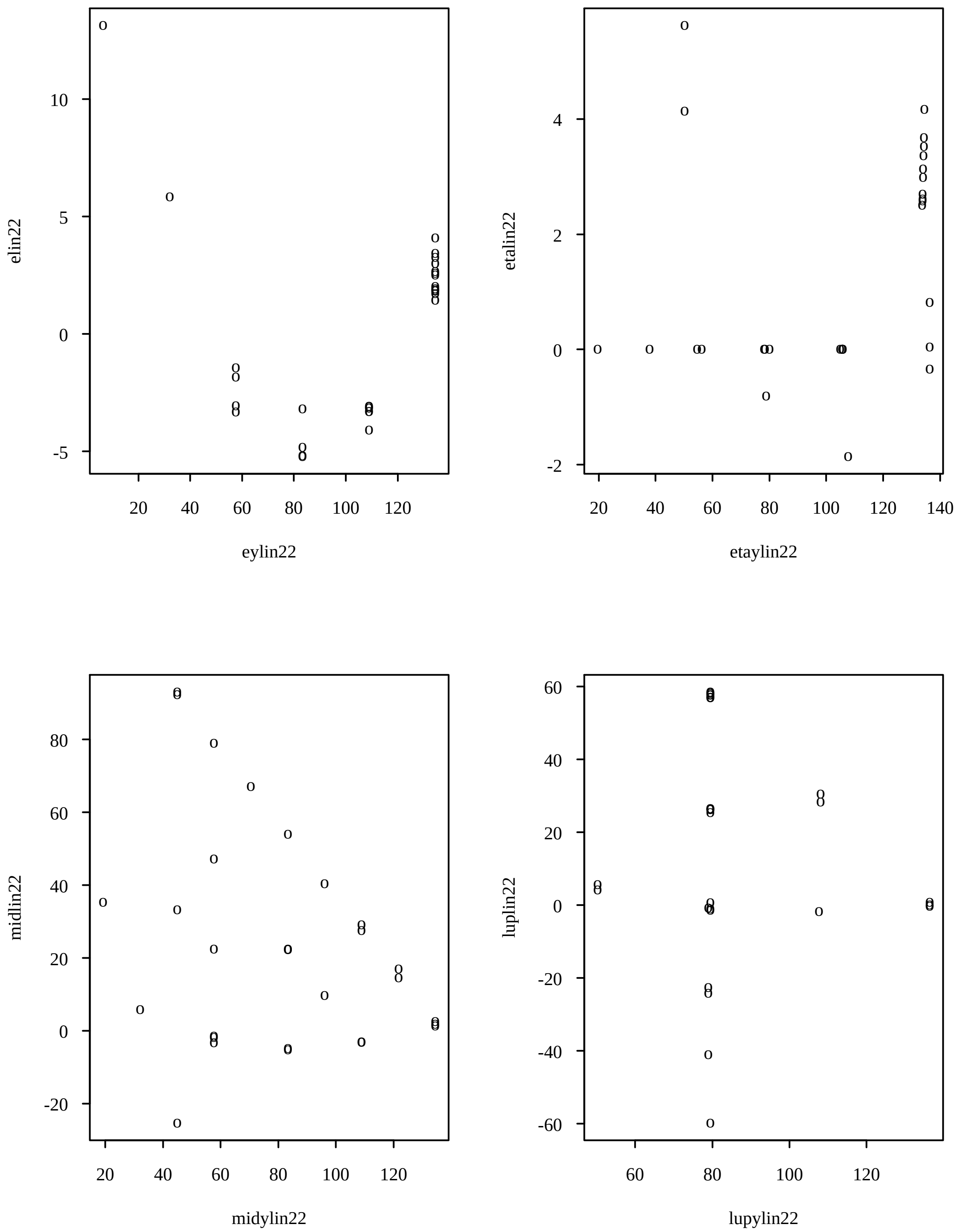
Scenario 20: covariate distribution N(6,4), β=2, p=0.7, n=30:



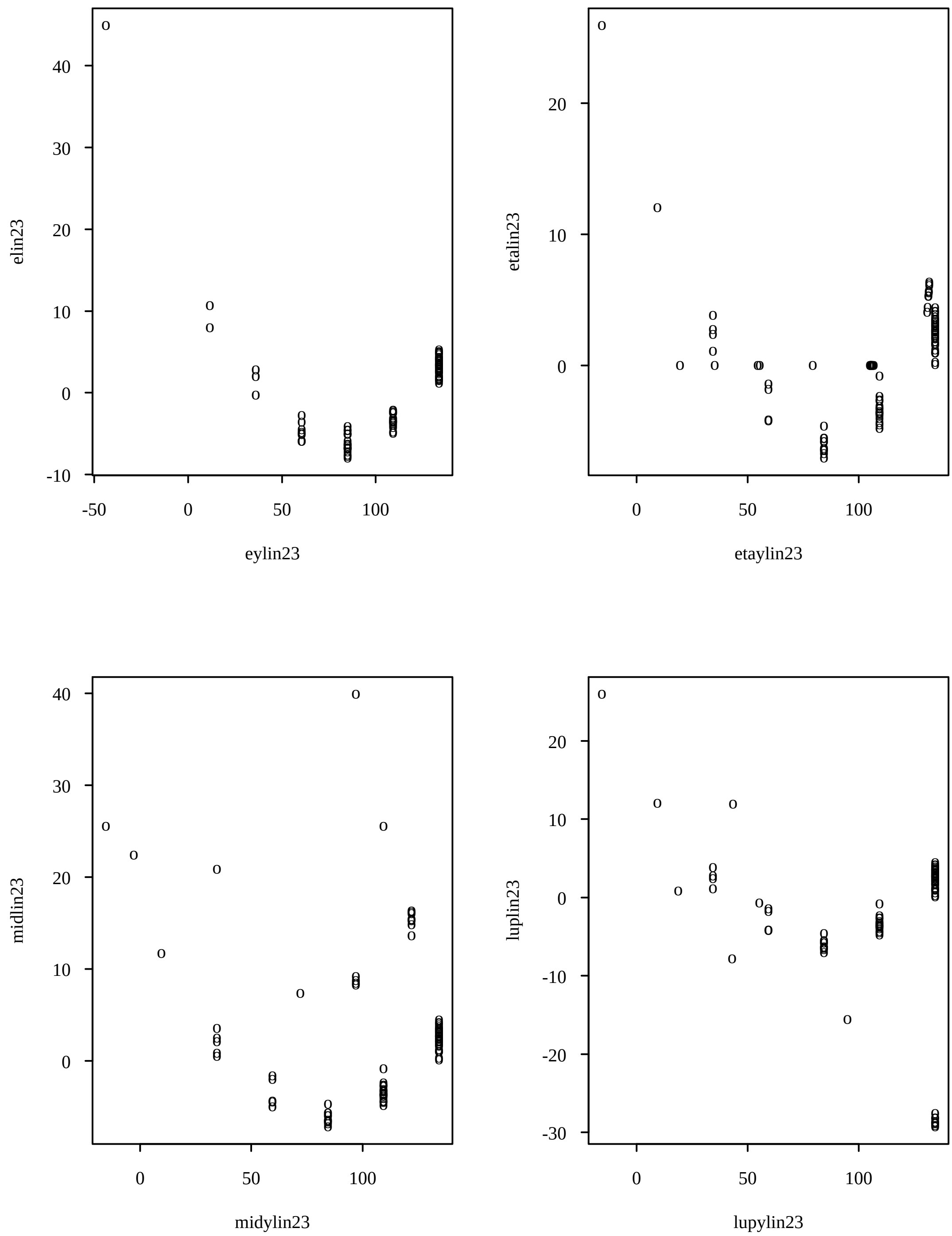
Scenario 21: covariate distribution N(6,4), β=5, p=0.3, n=100:



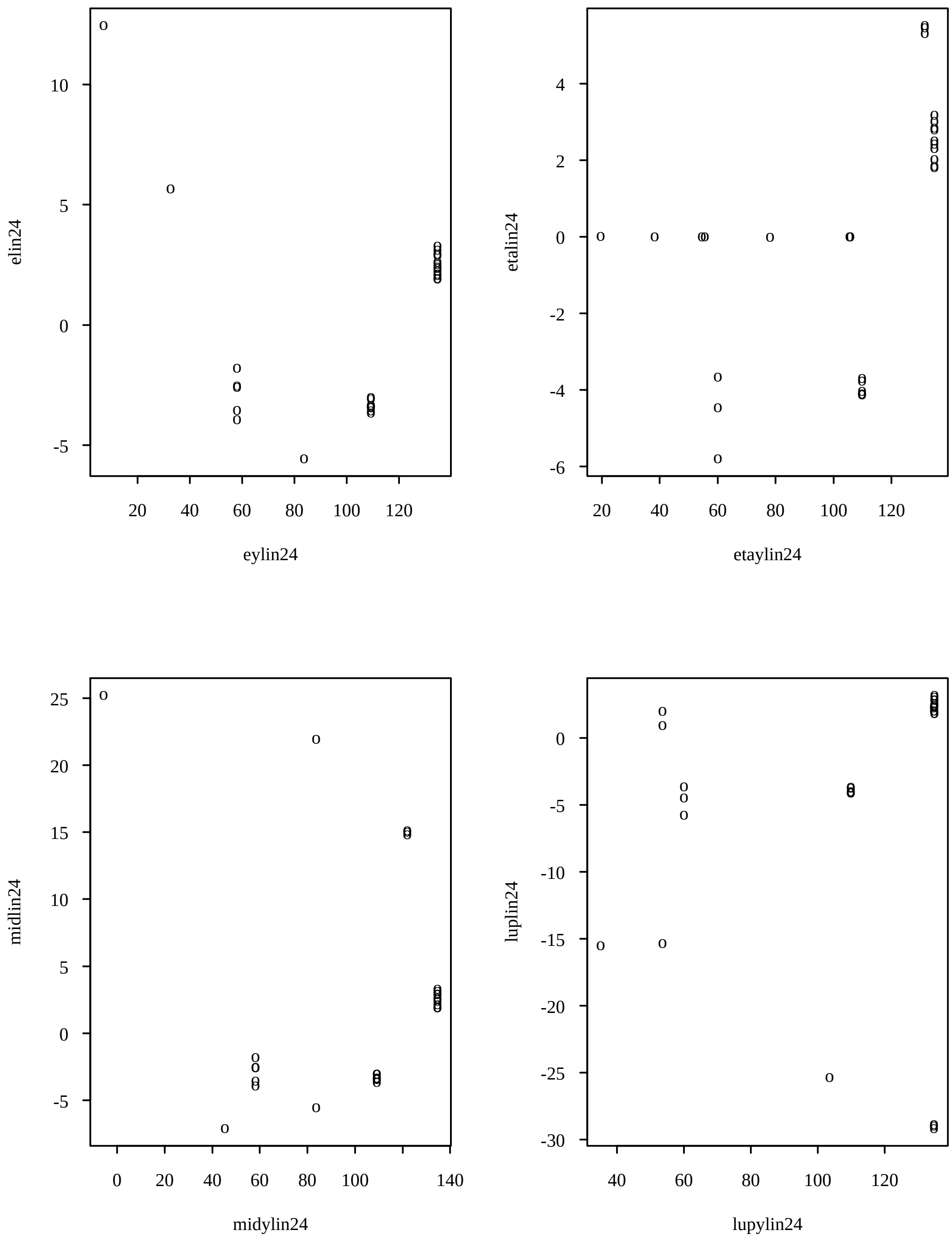
Scenario 22: covariate distribution N(6,4), β=5, p=0.3, n=30:



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



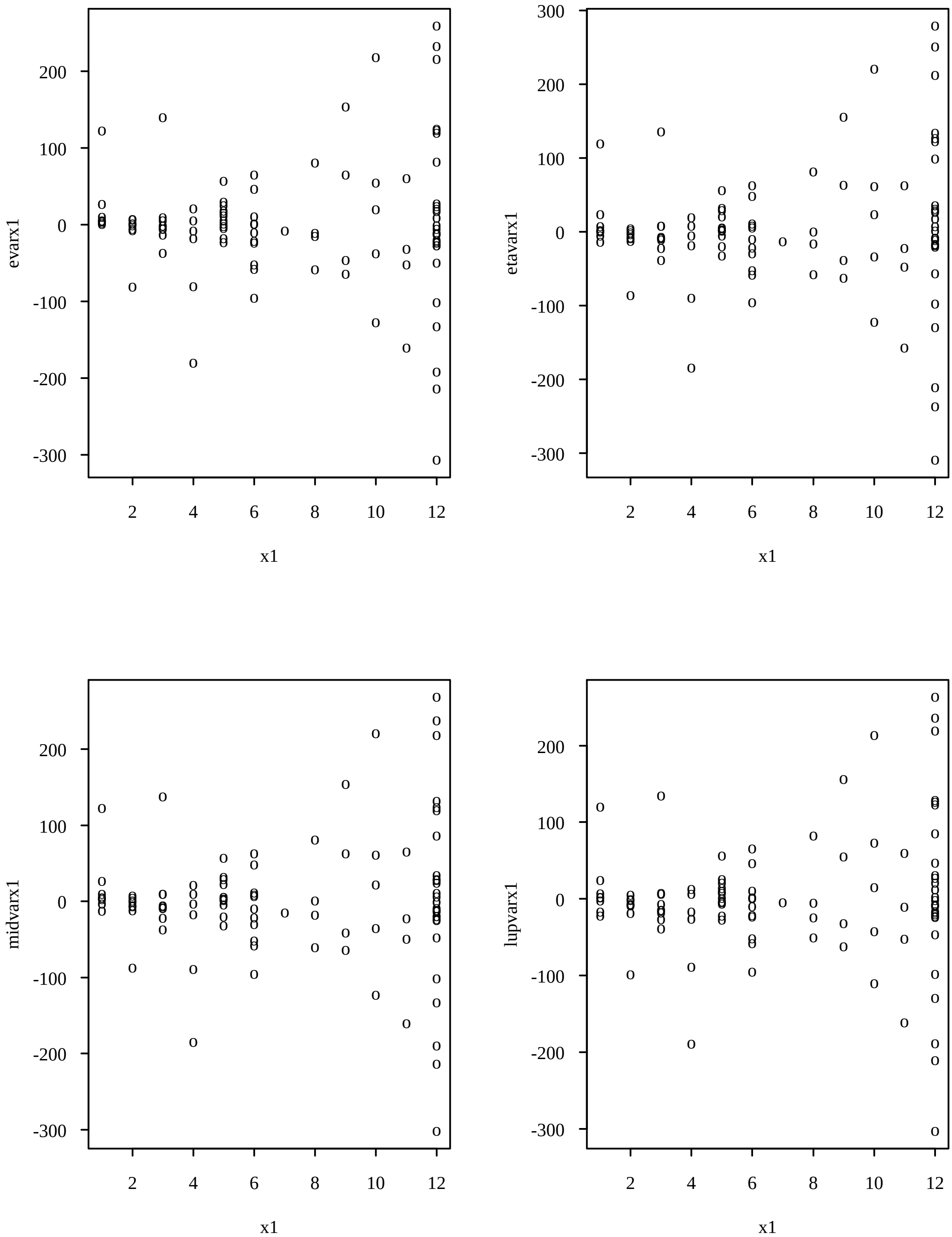
Scenario 24: covariate distribution N(6,4), β=5, p=0.7, n=30:



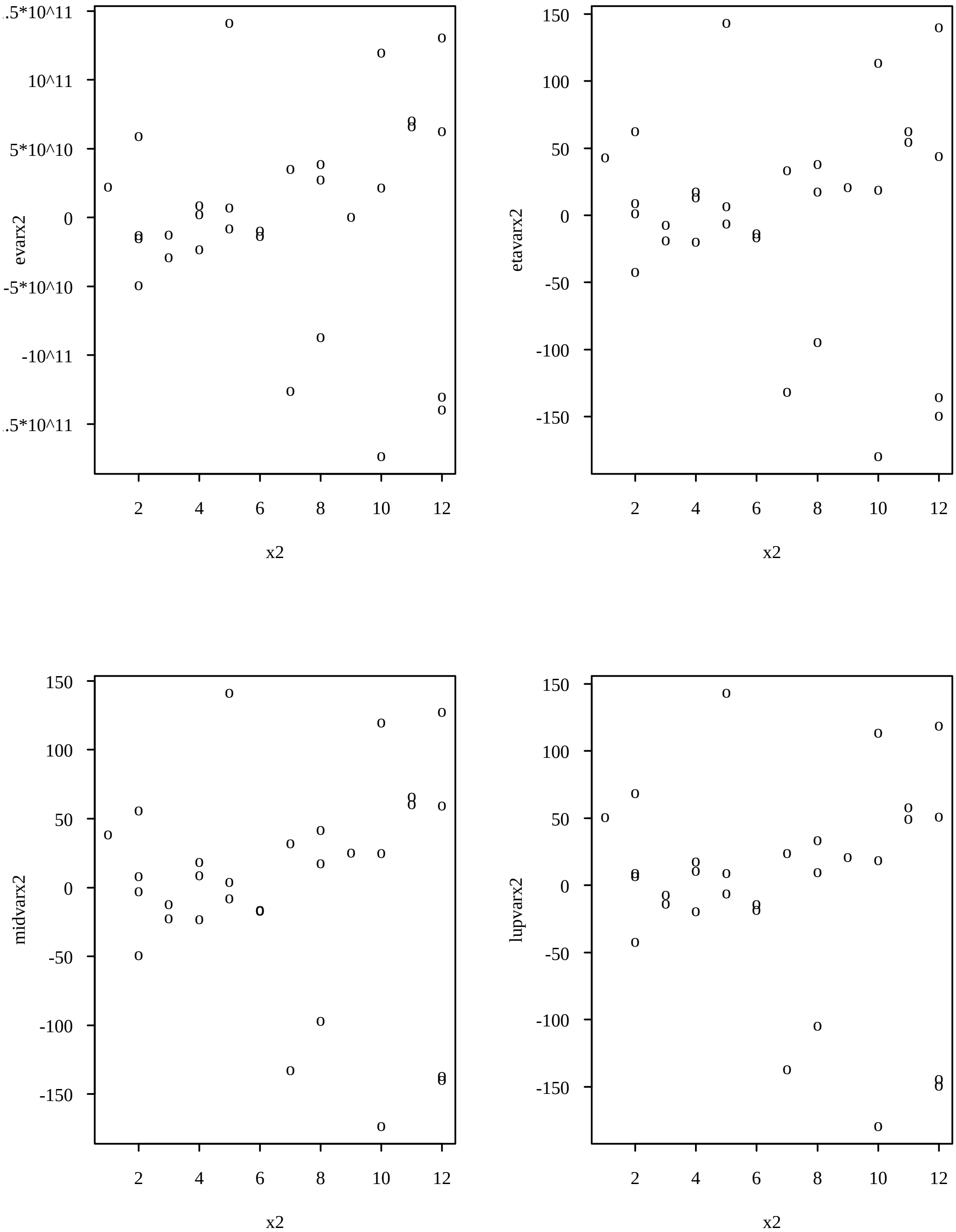
Appendix C

Residual plots when the error variance depends on the covariate

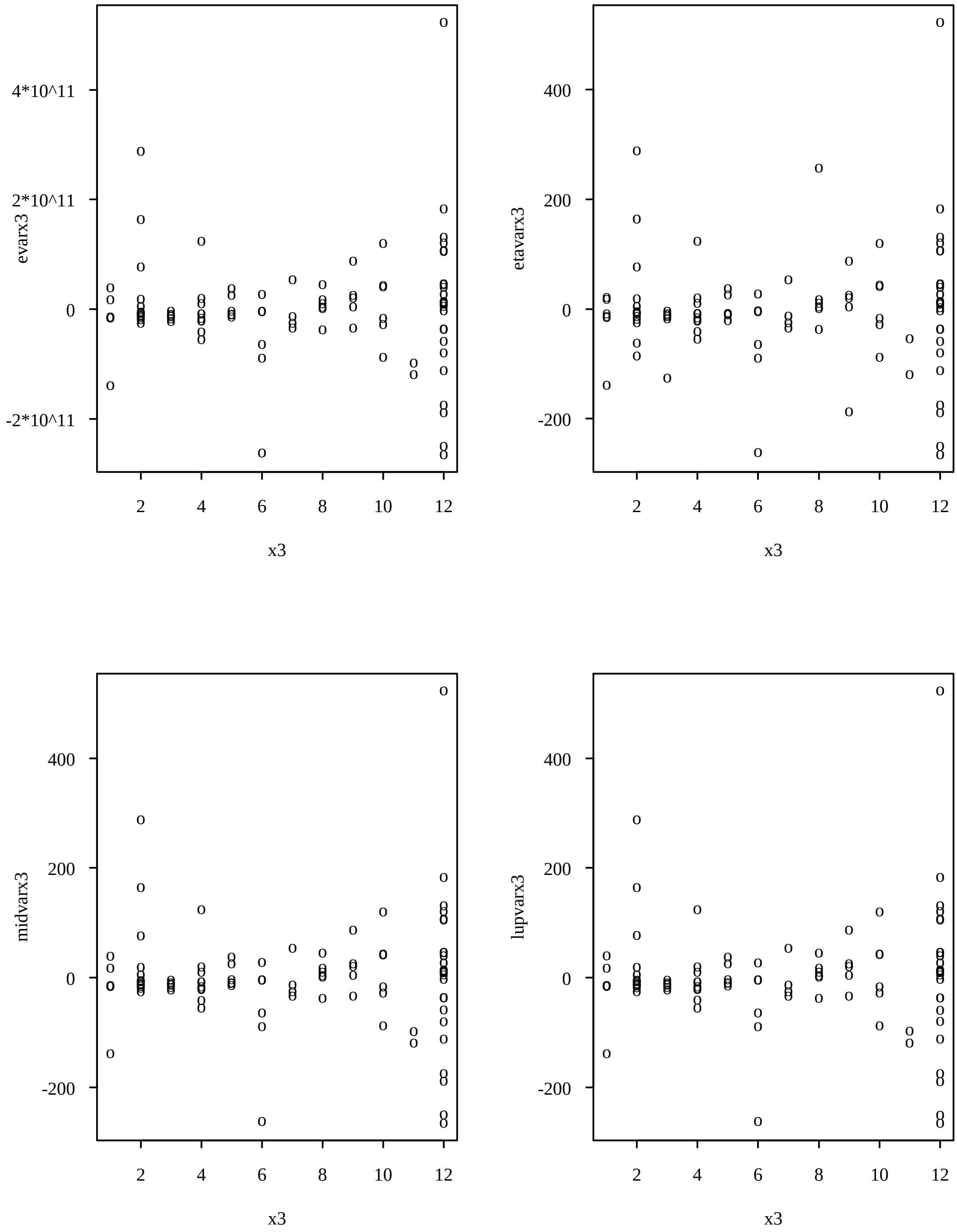
Scenario 1: covariate distribution Exp($\frac{1}{8}$), β=2, p=0.3, n=100:



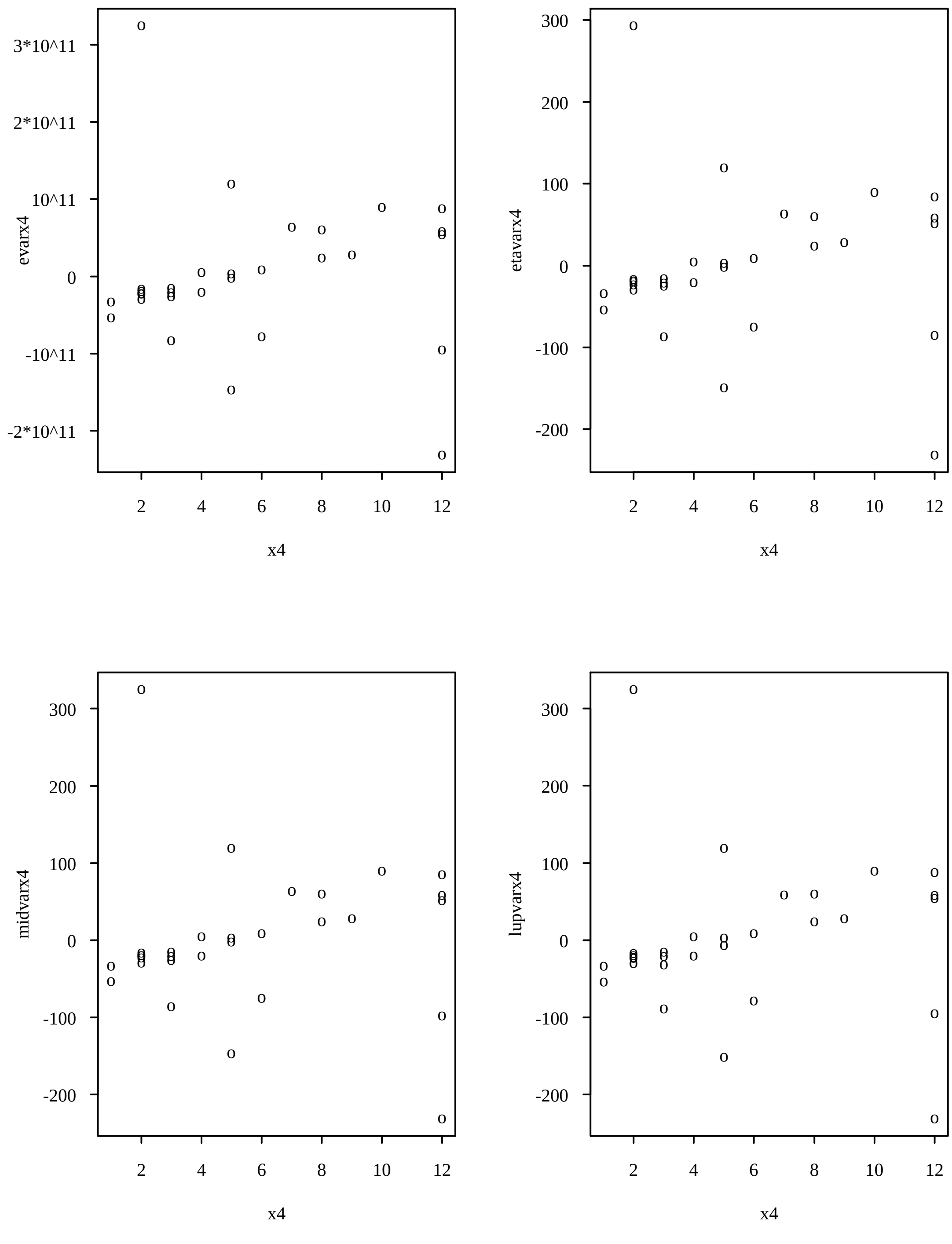
Scenario 2: covariate distribution Exp($\frac{1}{8}$), β=2, p=0.3, n=30:



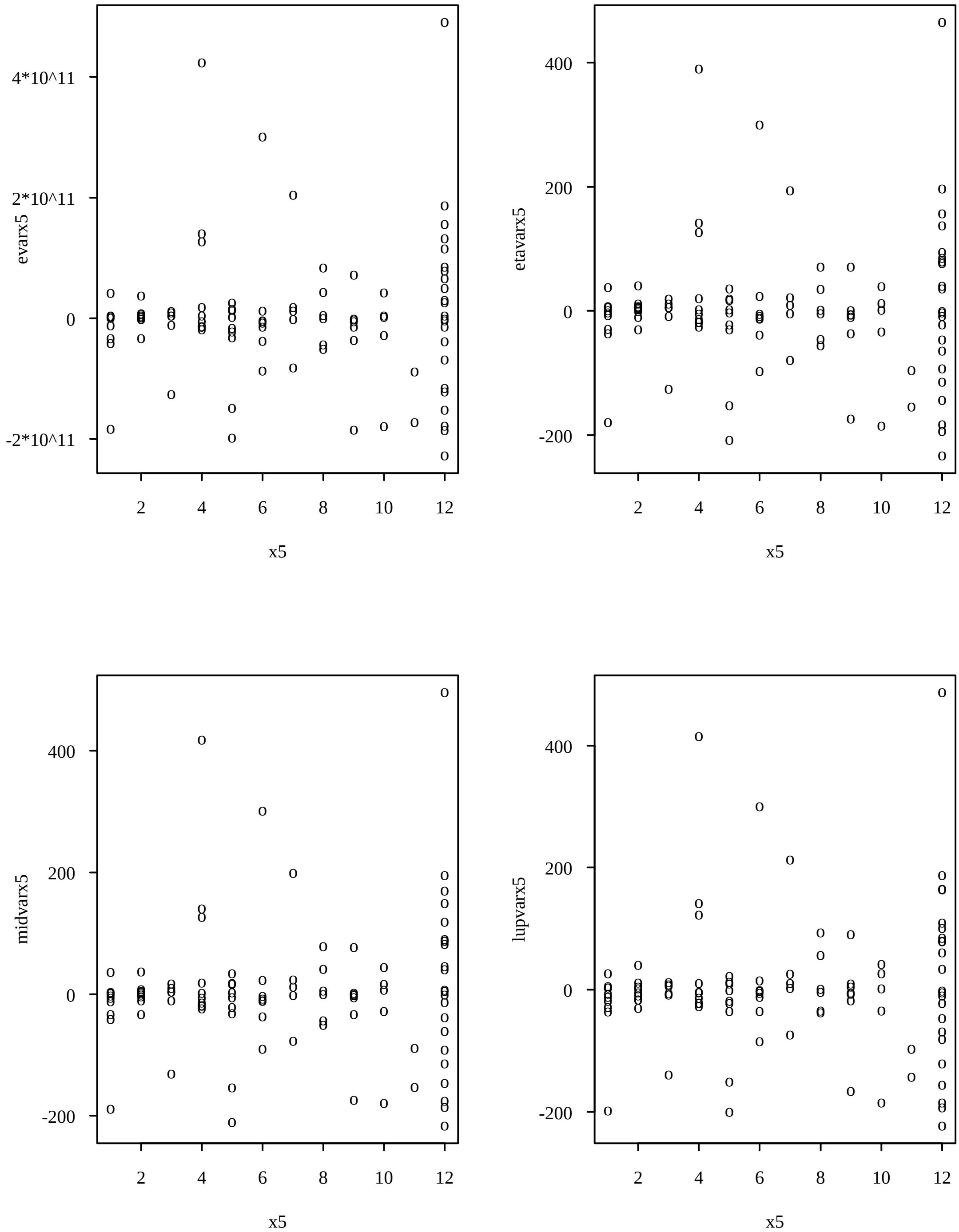
Scenario 3: covariate distribution $\mathrm{Exp}(\frac{1}{8})$, β=2, p=0.7, n=100:



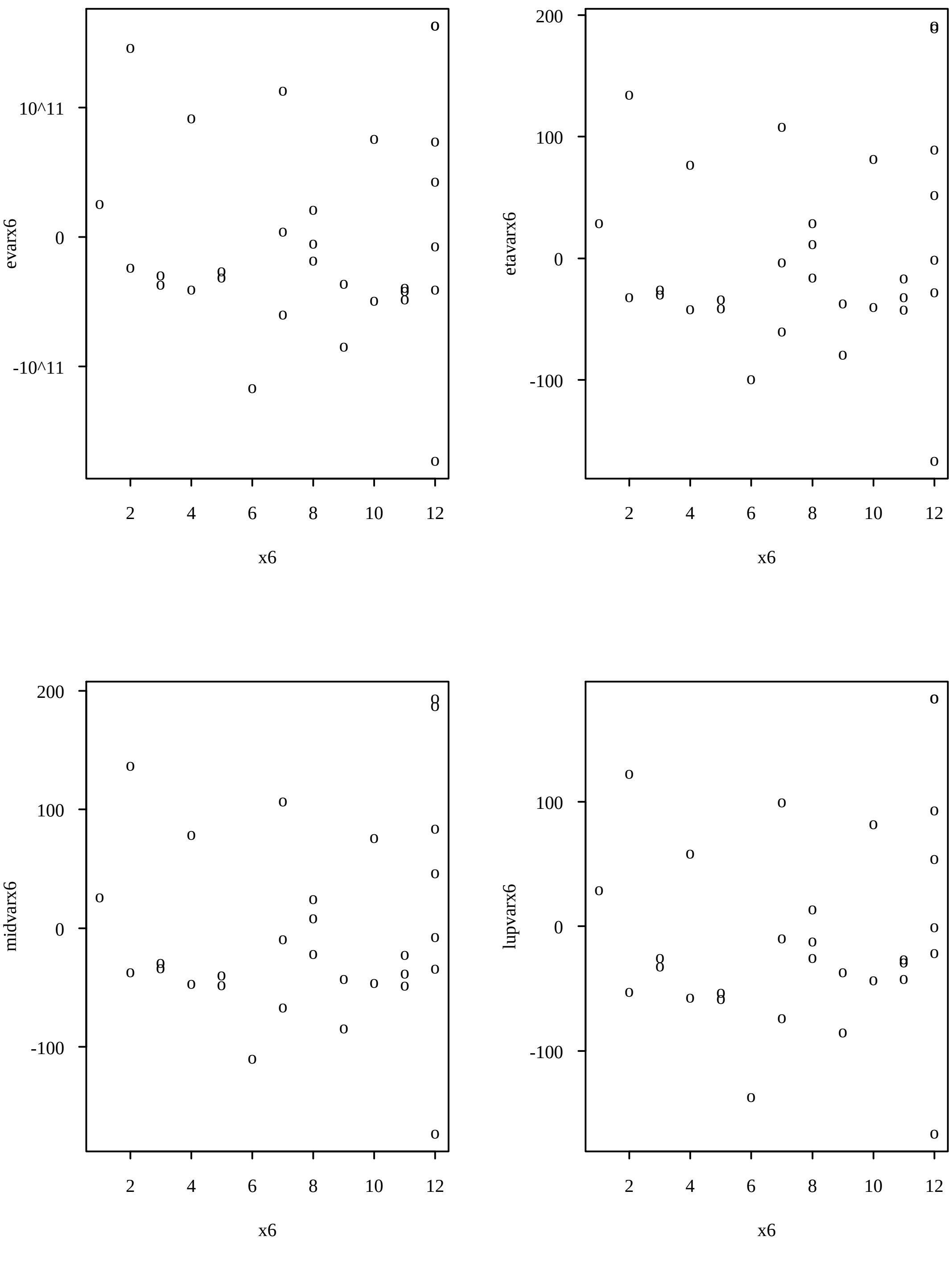
Scenario 4: covariate distribution Exp($\frac{1}{8}$), β=2, p=0.7, n=30:



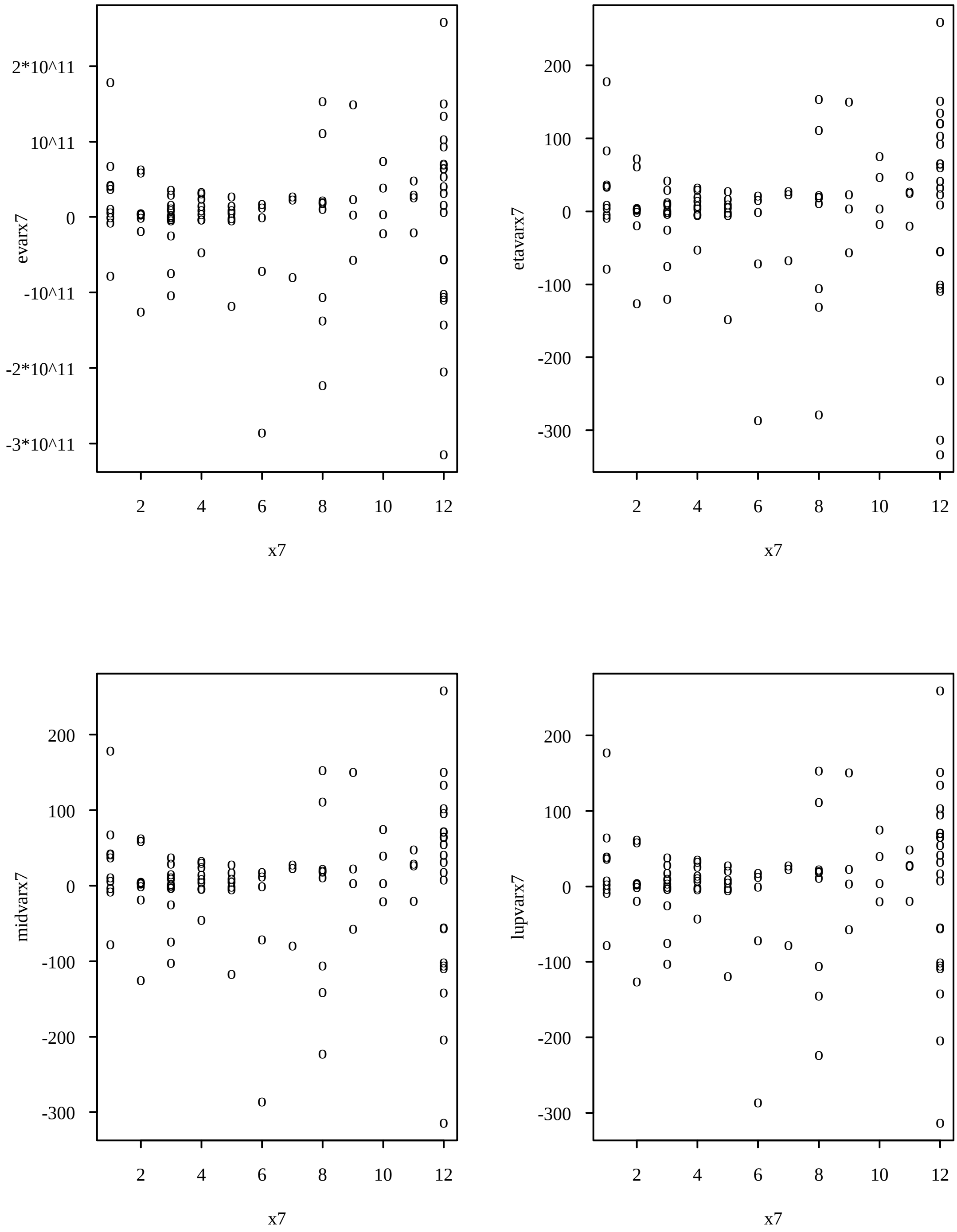
Scenario 5: covariate distribution $\mathrm{Exp}(\frac{1}{8})$, β=5, p=0.3, n=100:



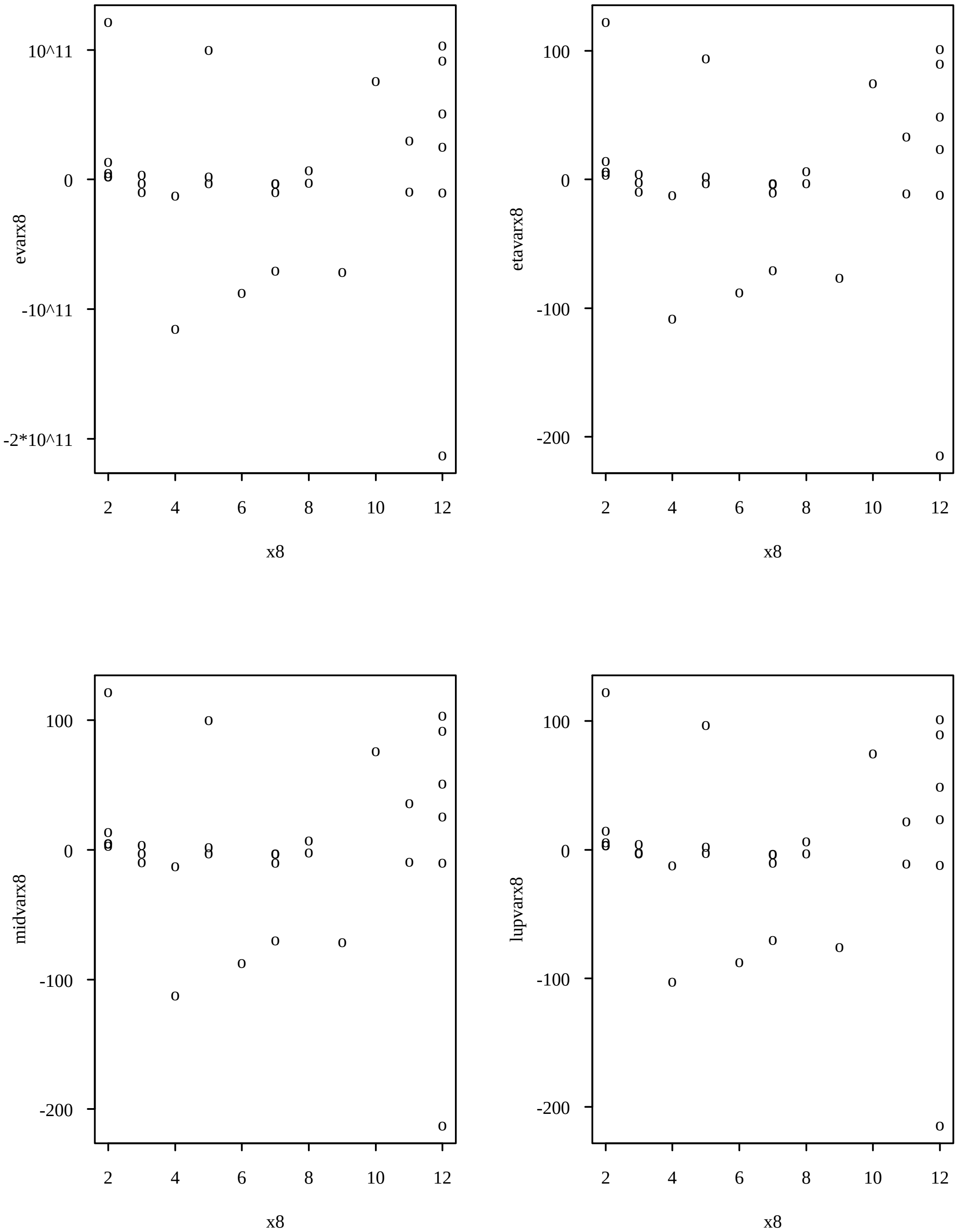
Scenario 6: covariate distribution Exp($\frac{1}{8}$), β=5, p=0.3, n=30:



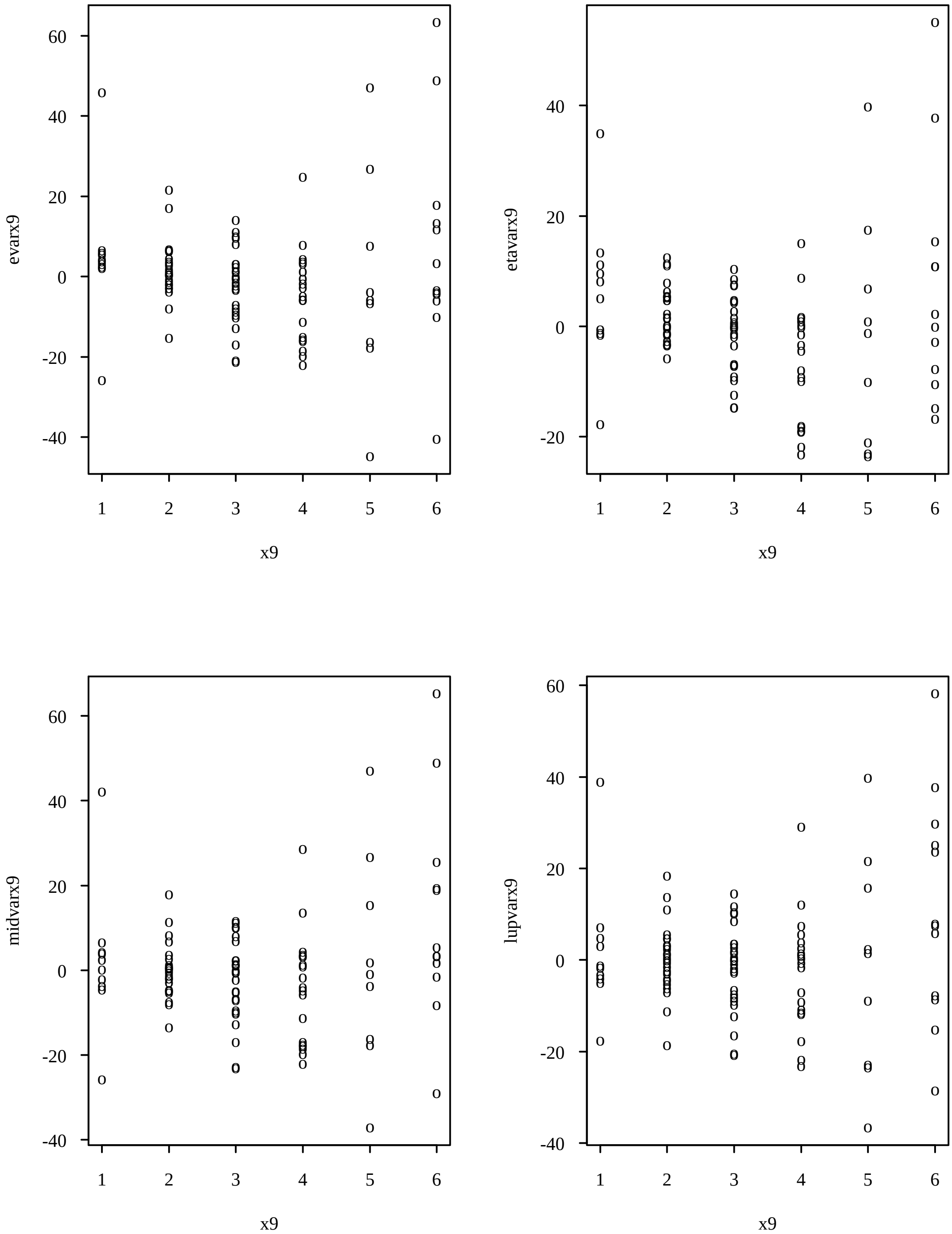
Scenario 7: covariate distribution Exp($\frac{1}{8}$), β=5, p=0.7, n=100:



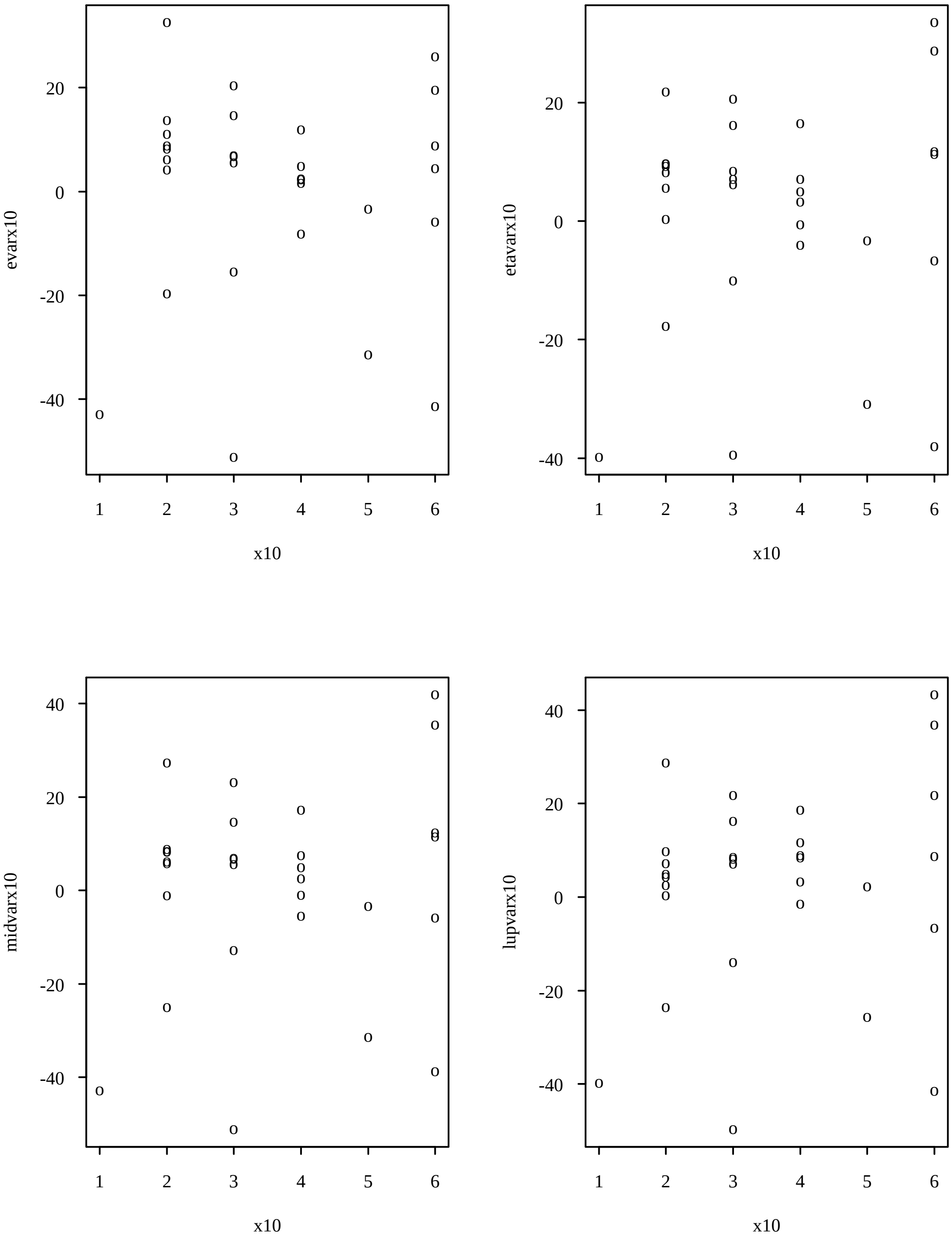
Scenario 8: covariate distribution Exp($\frac{1}{8}$), β=5, p=0.7, n=30:



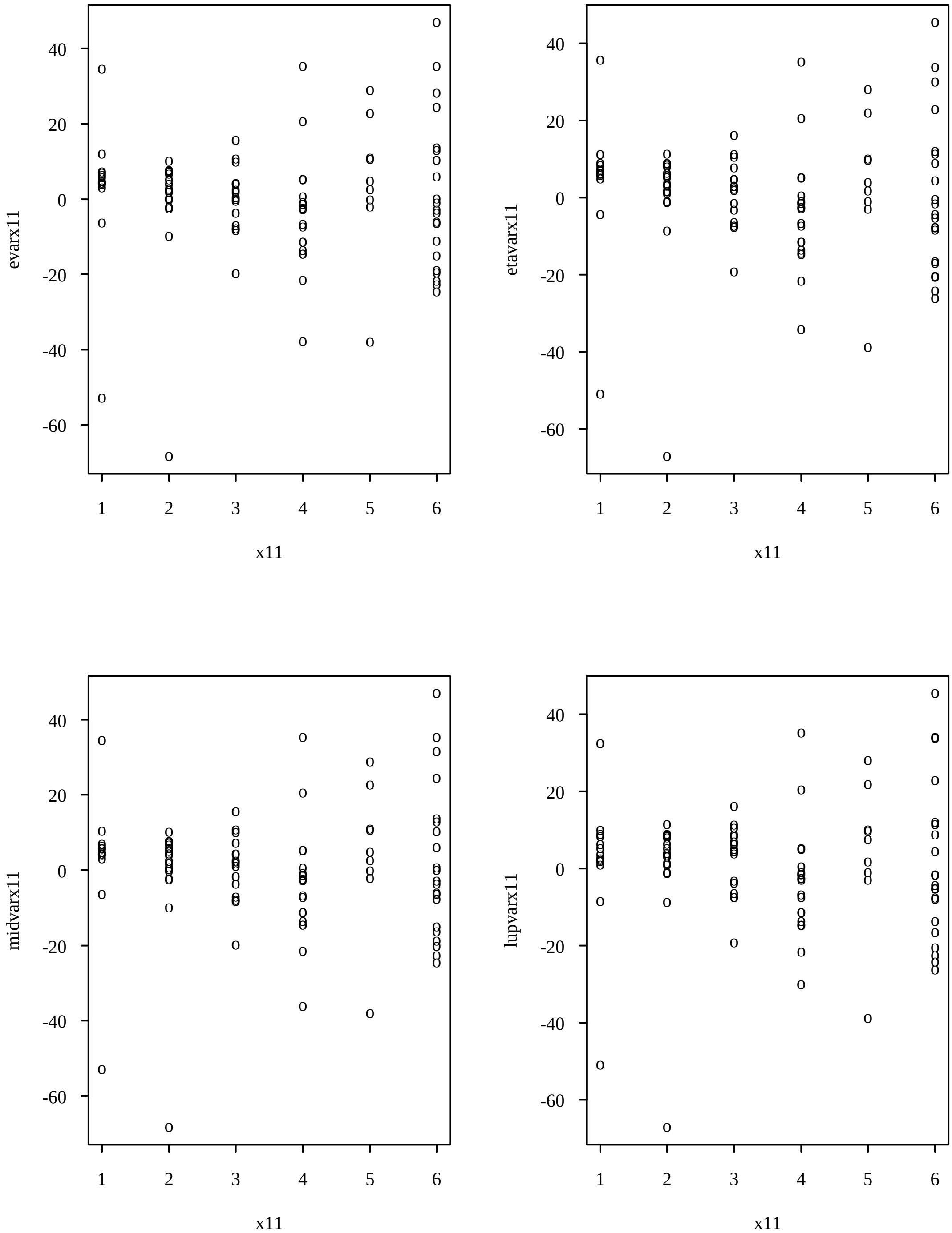
Scenario 9: covariate distribution $W(\frac{1}{6}, \frac{3}{2})$, β=2, p=0.3, n=100:



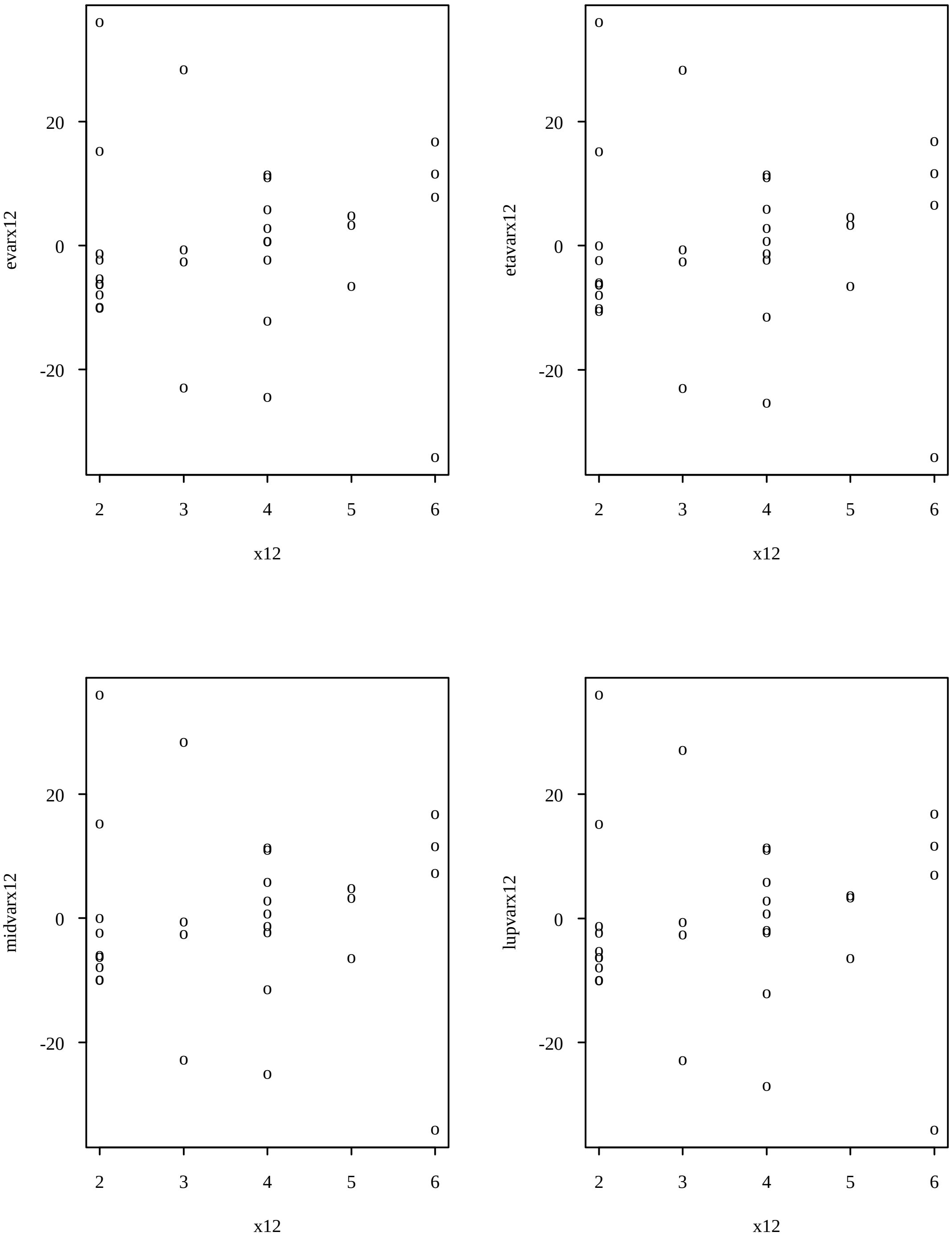
Scenario 10: covariate distribution $W(\frac{1}{6}, \frac{3}{2})$, β=2, p=0.3, n=30:



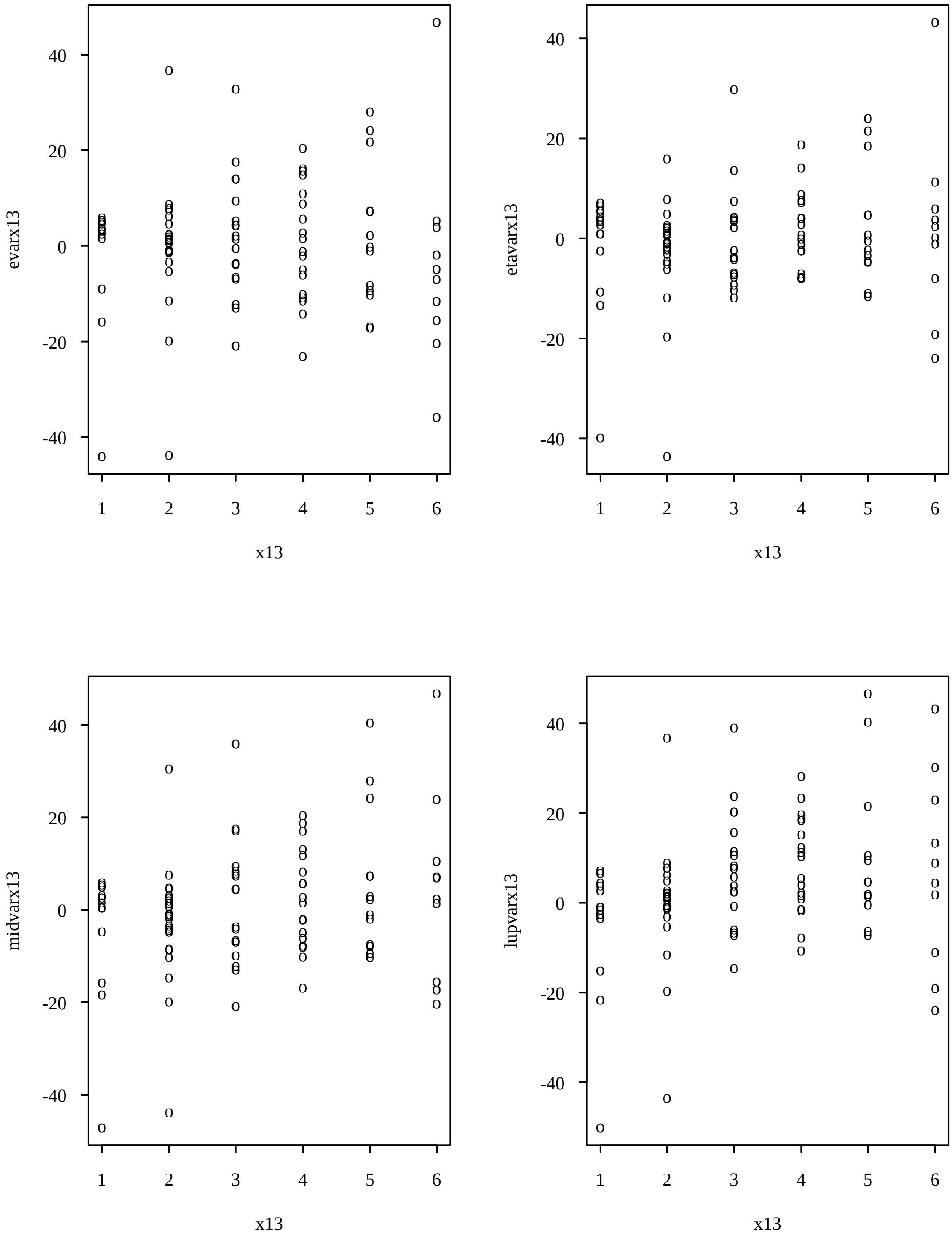
Scenario 11: covariate distribution $W(\frac{1}{6}, \frac{3}{2})$, β=2, p=0.7, n=100:



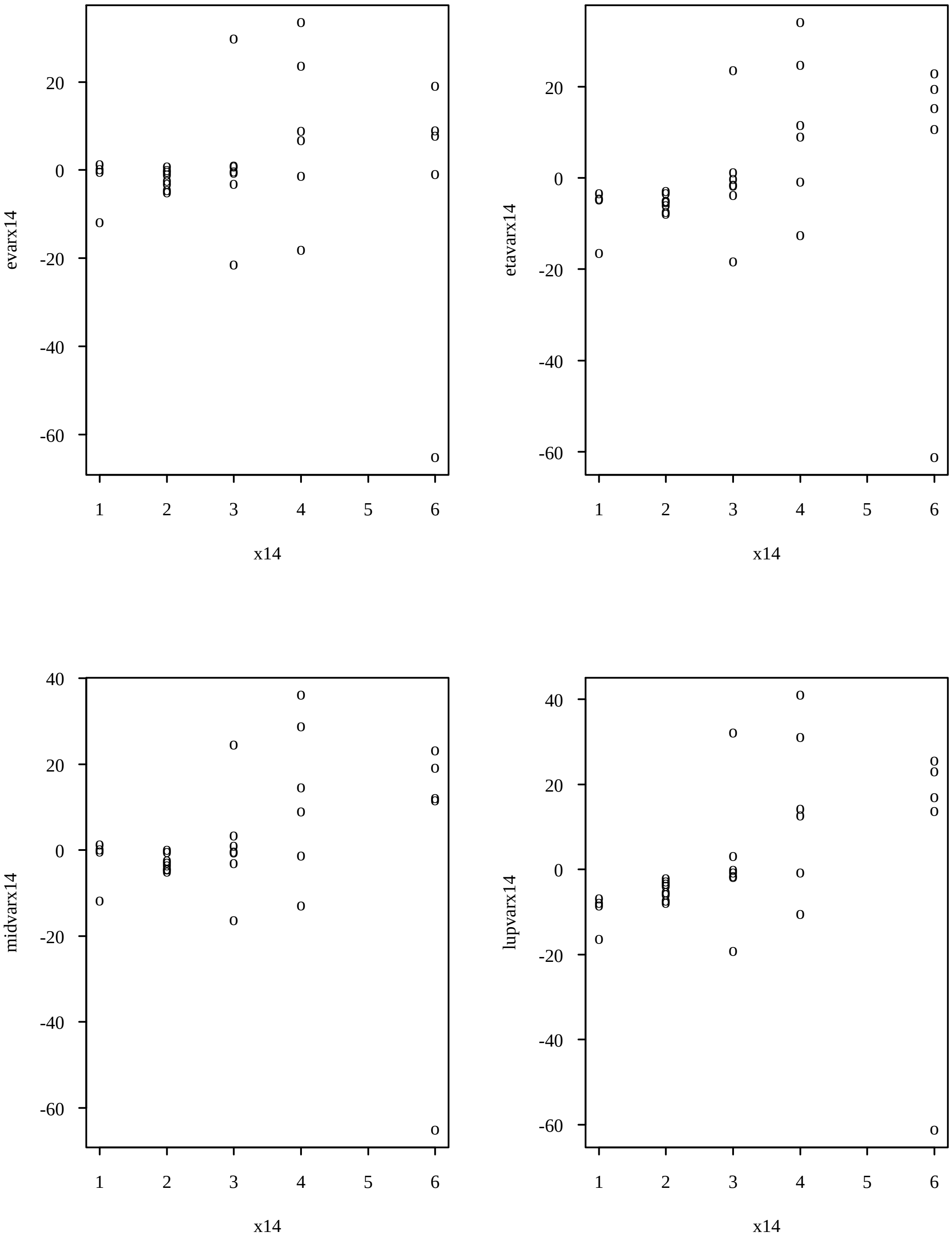
Scenario 12: covariate distribution $W(\frac{1}{6}, \frac{3}{2})$, β=2, p=0.7, n=30:



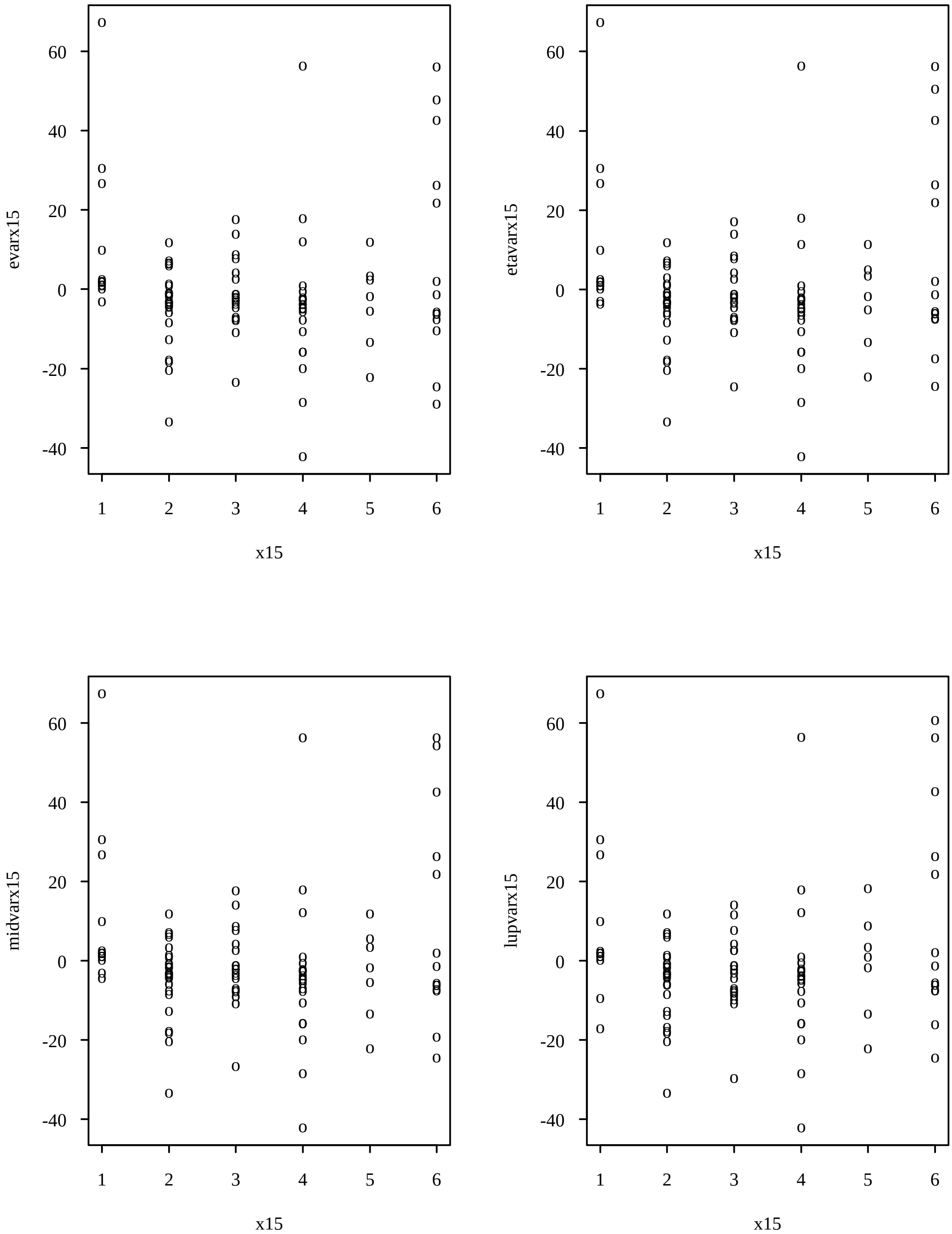
Scenario 13: covariate distribution $W(\frac{1}{6},\frac{3}{2})$, β=5, p=0.3, n=100:



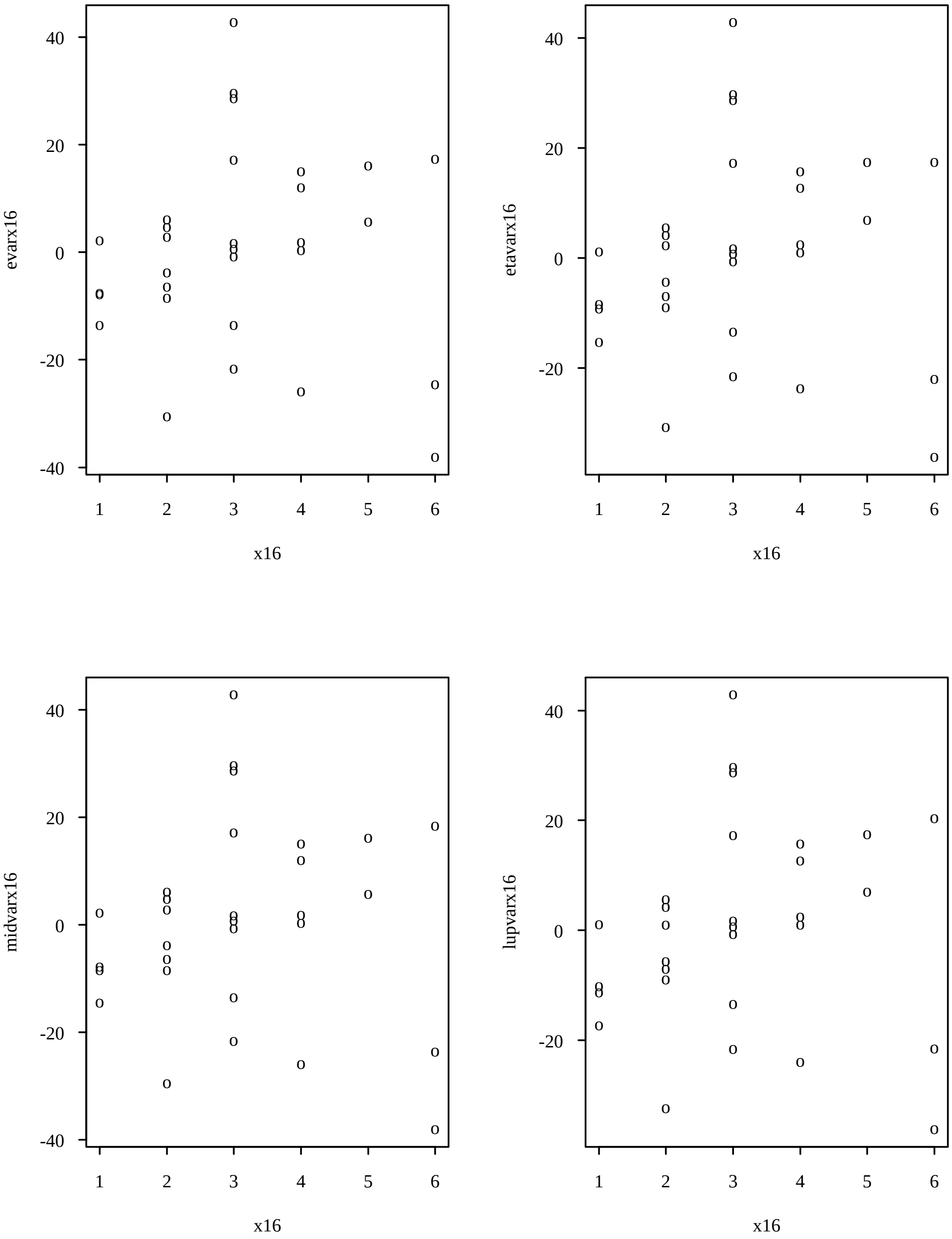
Scenario 14: covariate distribution W($\frac{1}{6}$, $\frac{3}{2}$), β=5, p=0.3, n=30:



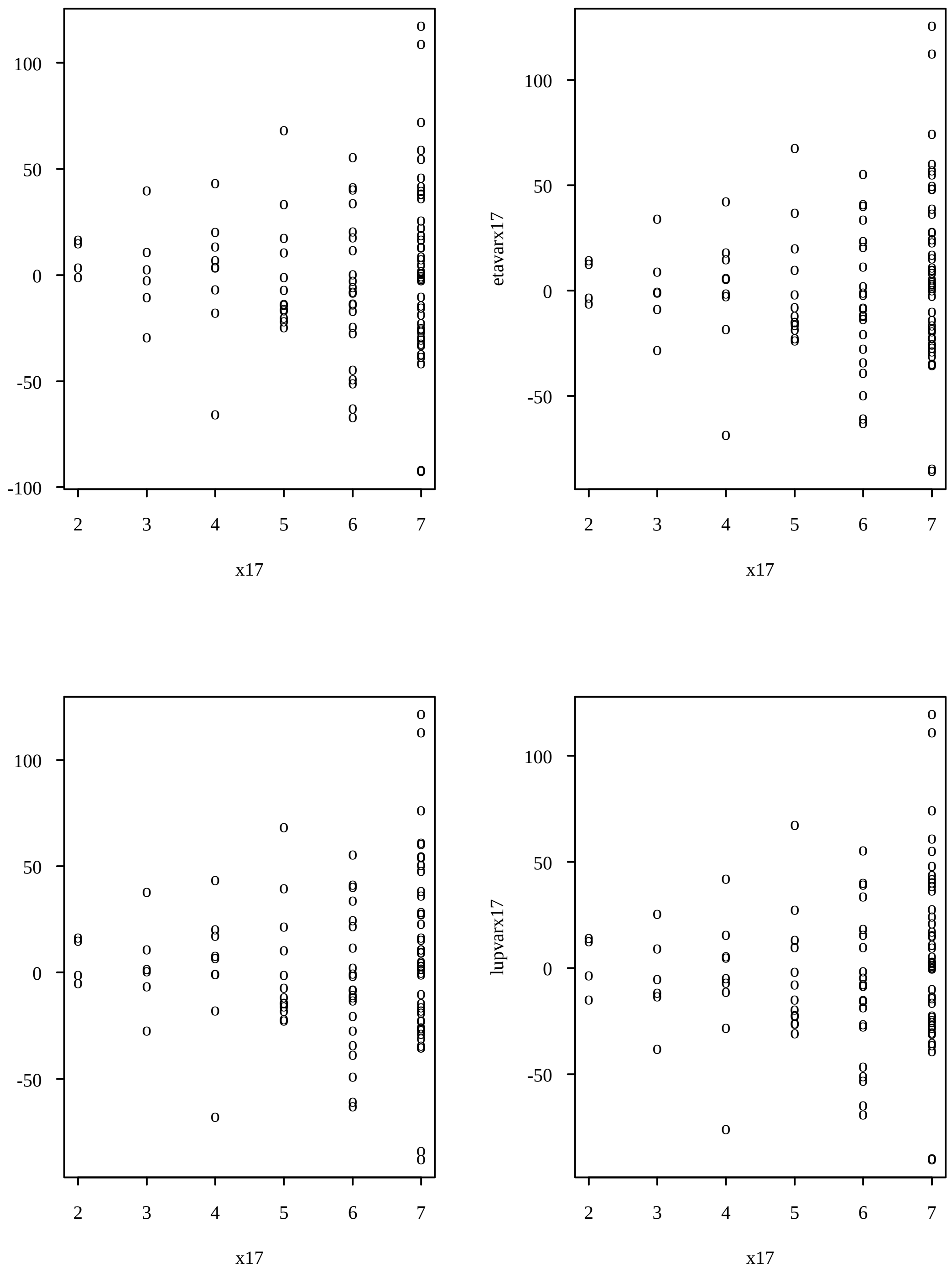
Scenario 15: covariate distribution $W(\frac{1}{6}, \frac{3}{2})$, β=5, p=0.7, n=100:



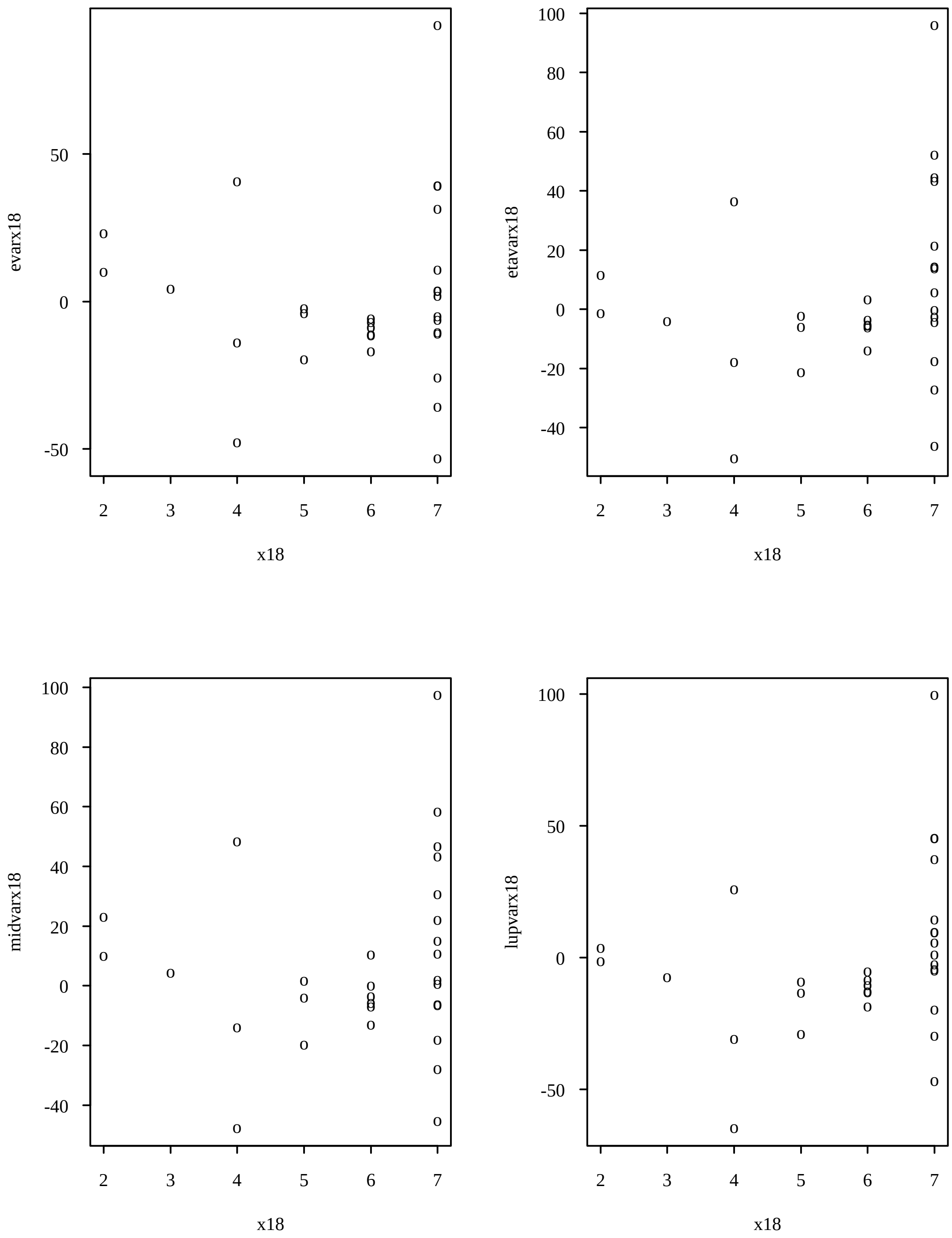
Scenario 16: covariate distribution $W(\frac{1}{6}, \frac{3}{2})$, β=5, p=0.7, n=30:



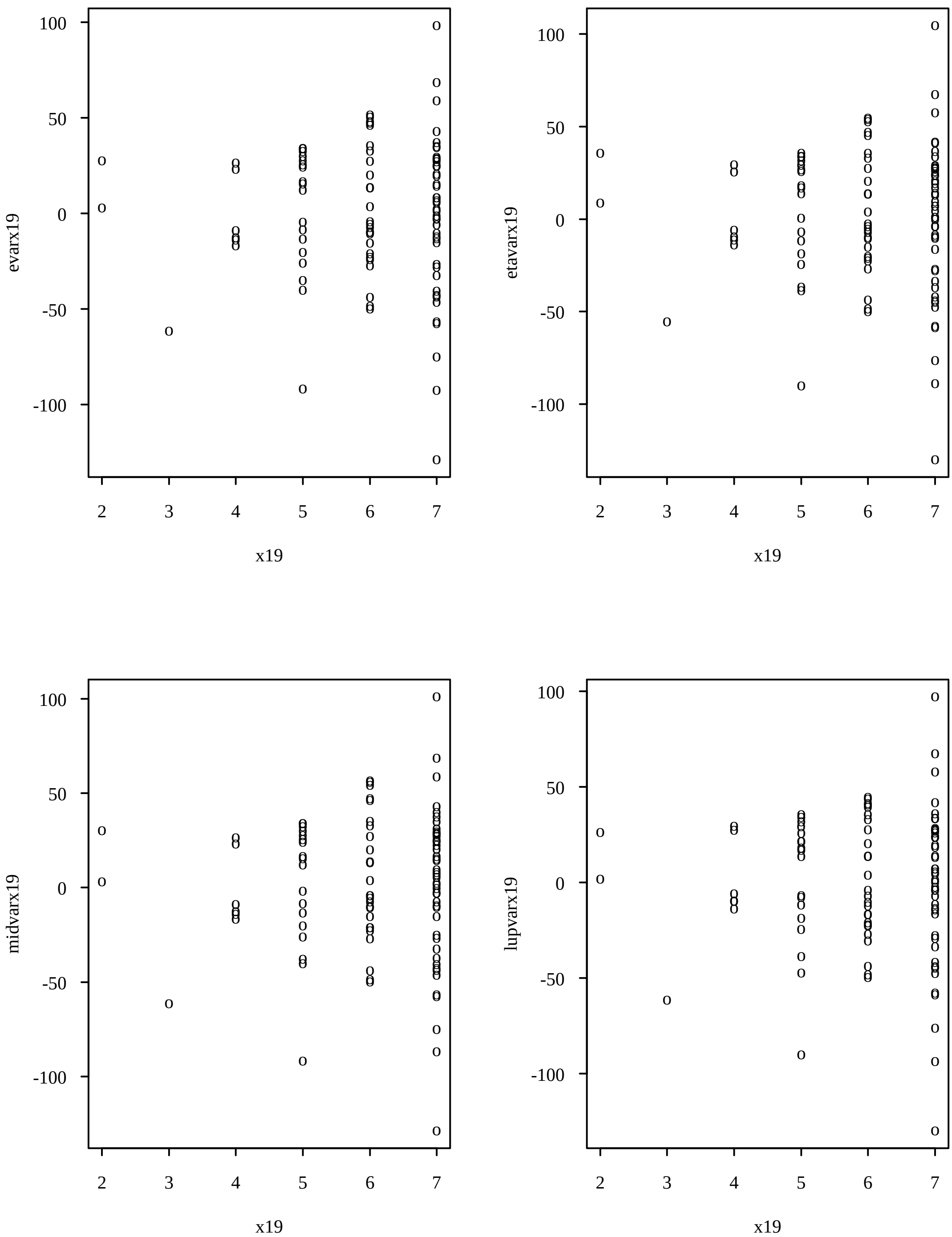
Scenario 17: covariate distribution N(6,4), β=2, p=0.3, n=100:



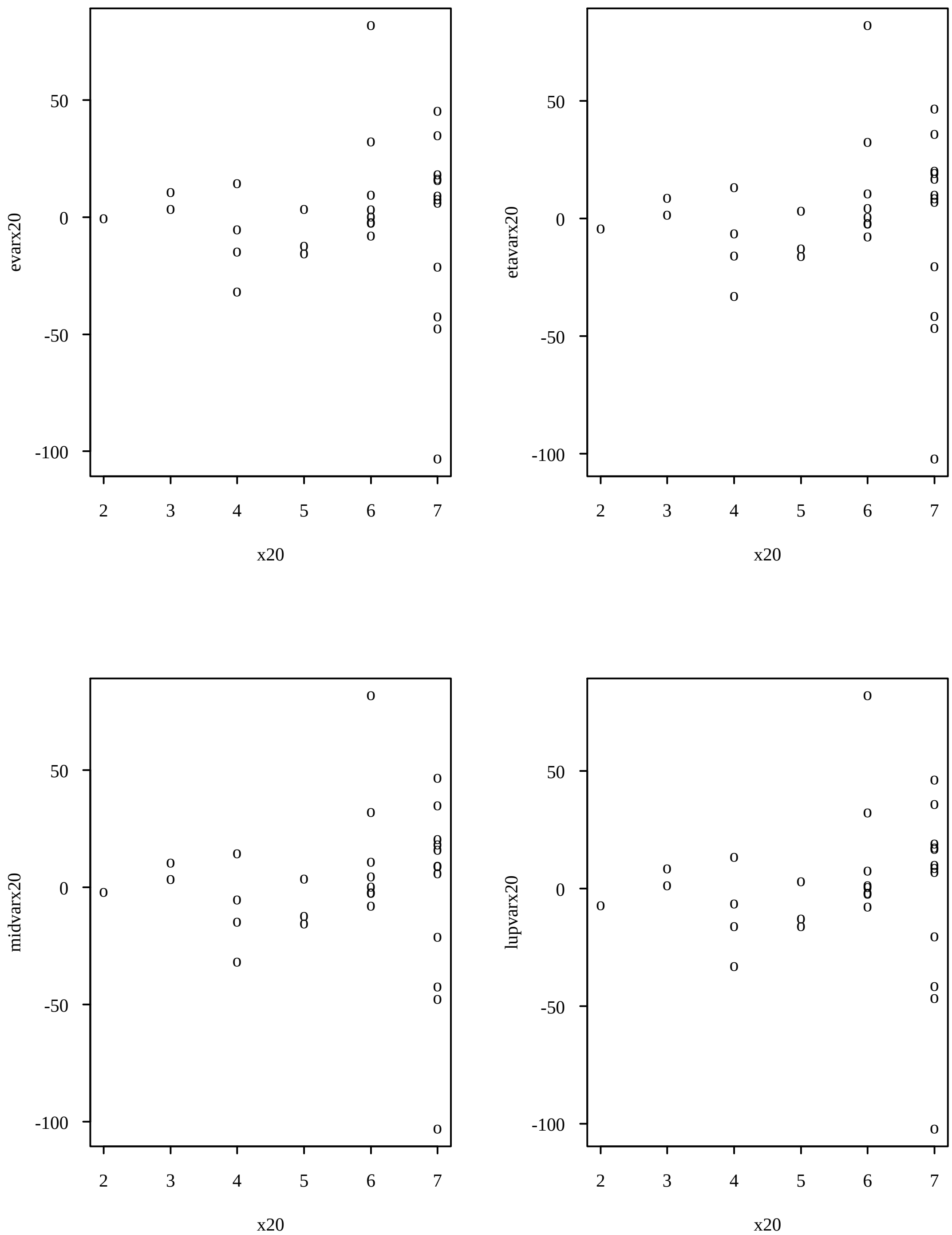
Scenario 18: covariate distribution N(6,4), β=2, p=0.3, n=30:



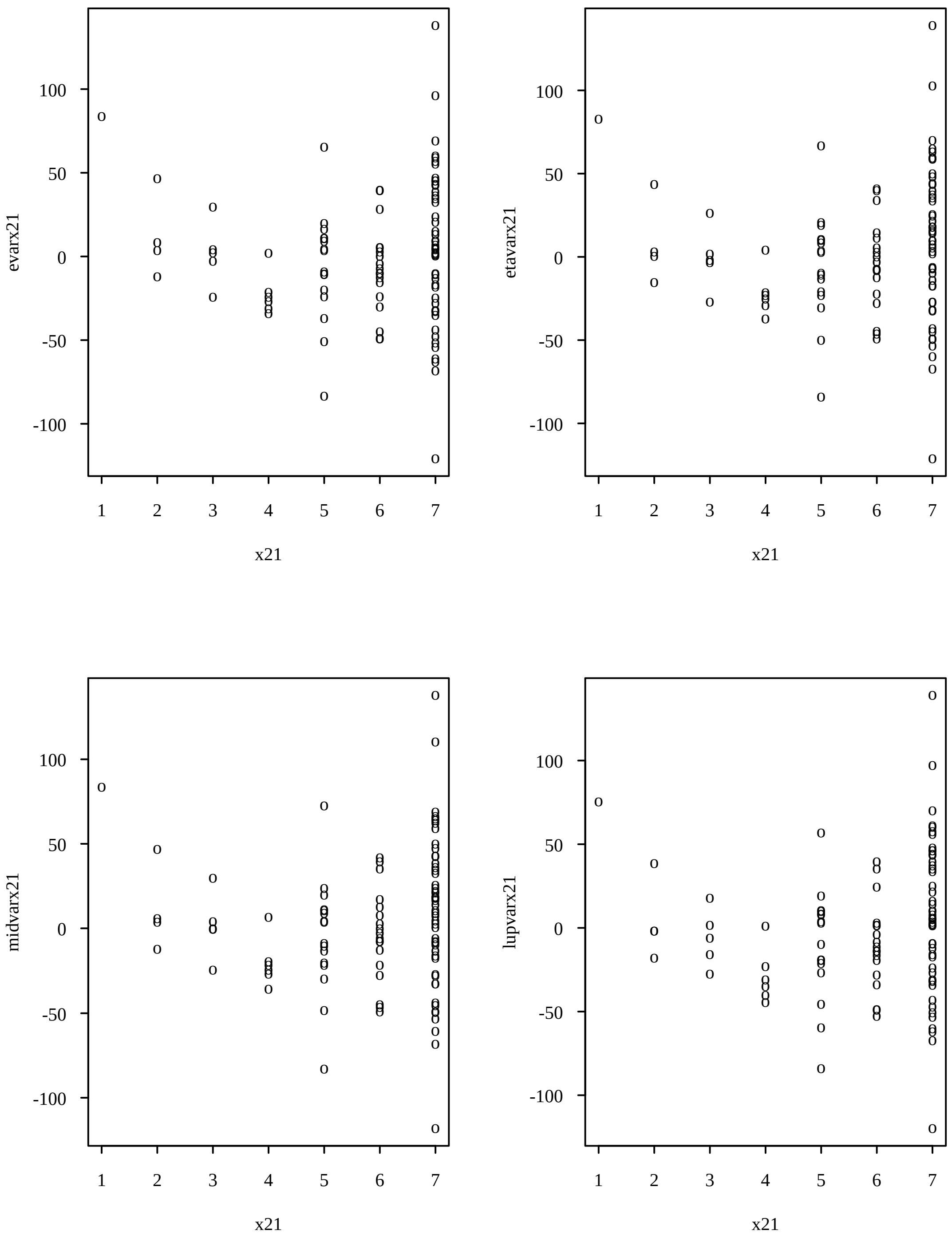
Scenario 19: covariate distribution N(6,4), β=2, p=0.7, n=100:



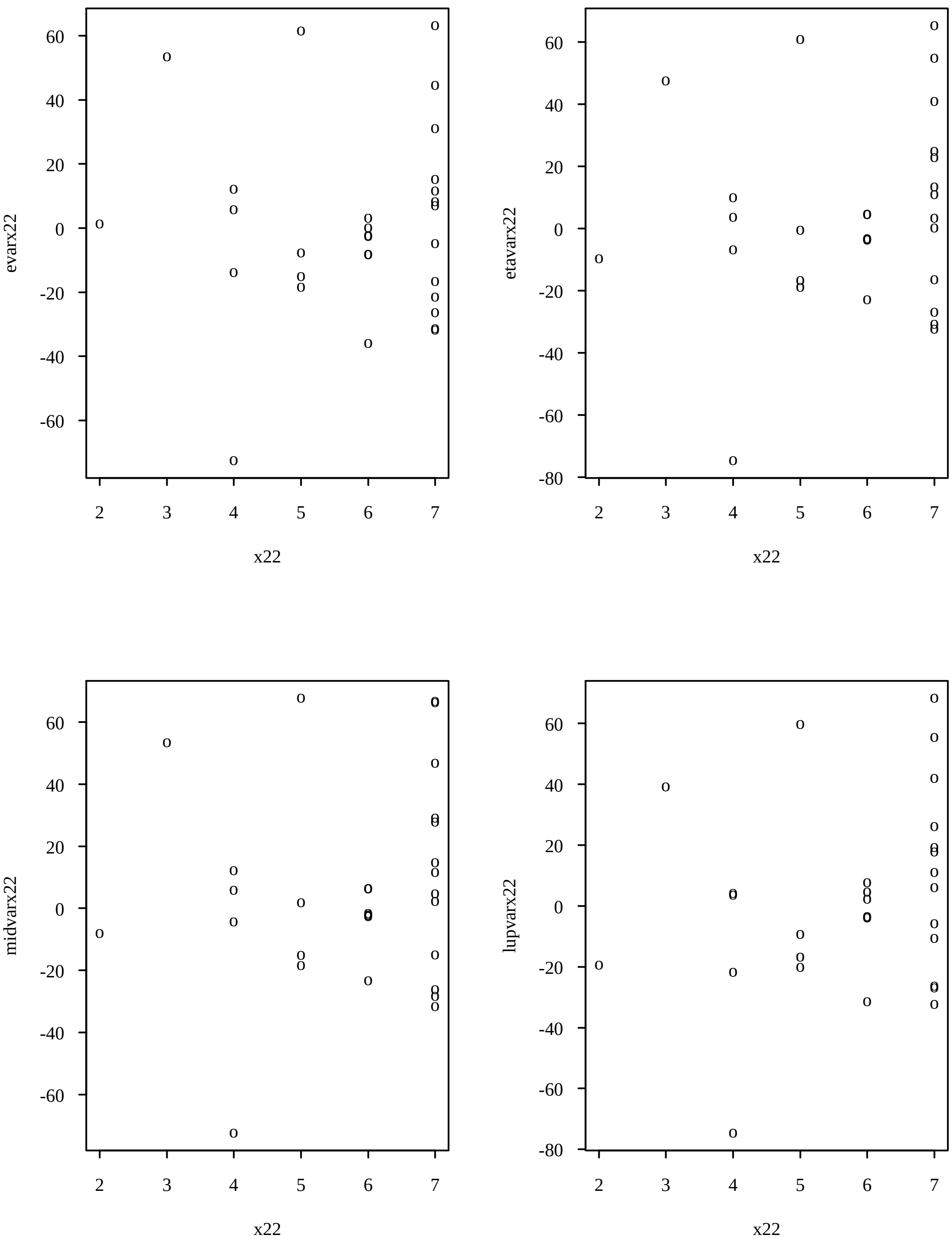
Scenario 20: covariate distribution N(6,4), β=2, p=0.7, n=30:



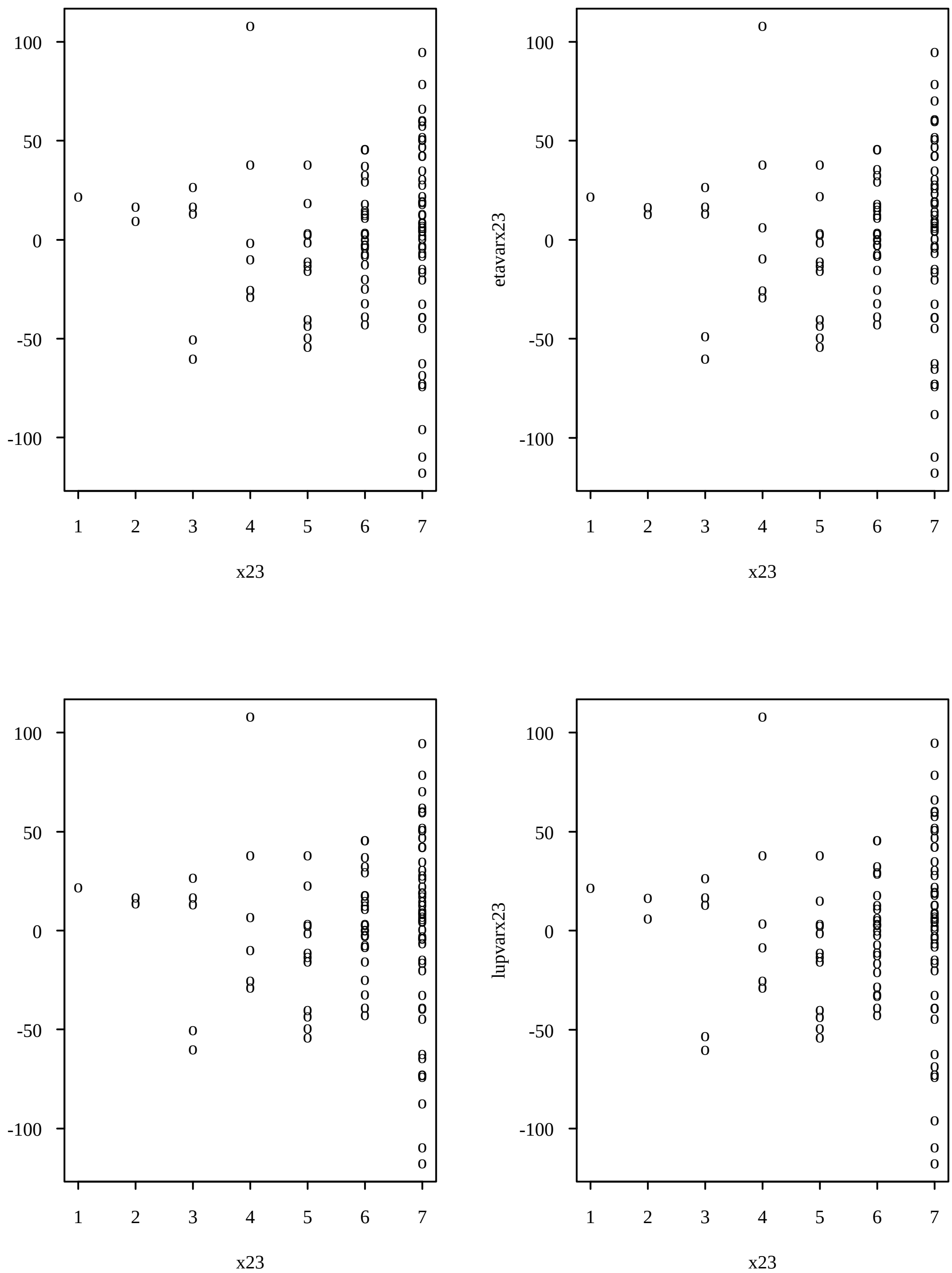
Scenario 21: covariate distribution N(6,4), β=5, p=0.3, n=100:



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: