Ordinary Least Squares Regression And Regression Diagnostics University of Virginia Charlottesville, VA. April 20, 2001

Jim Patrie.

Department of Health Evaluation Sciences

Division of Biostatistics and Epidemiology

University of Virginia

Charlottesville, VA

jpatrie@virginia.edu

Presentation Outline

I) Overview of Regression Analysis.

II) Different Types of Regression Analysis.

III) The General Linear Model.

IV) Ordinary Least Squares Regression Parameter Estimation.

V) Statistical Inference for the OLS Regression Model.

VI) Overview of the Model Building Process.

VII) An Example Case Study.

Introduction

The term “regression analysis” describes a collection of statistical techniques which serve as the basis for drawing inference as to whether or not a relationship exists between two or more quantities within a system, or within a population.

More specifically, regression analysis is a method to quantitatively characterize the relationship between a response variable Y, which is assumed to be random, and one or more explanatory variables (X), which are generally assumed to have values that are fixed.

I) Regression analysis is typically utilized for one of the following purposes.

- Description

  To assess whether or not a response variable, or perhaps a function of the response variable, is associated with one or more independent variables.

- Control

  To control for secondary factors which may influence the response variable, but are not consider as the primary explanatory variables of interest.

- Prediction

  To predict the value of the response variable at specific values of the explanatory variables.

II) Types of Regression Analysis

- General Linear Regression.*
- Non-Linear Regression.
- Robust Regression.

  Least median squares regression.
  Least absolute deviation regression.
  Weighted least square.

- Non-Parametric Regression.
- Generalized Linear Regression.

  Logistic regression.
  Log-linear regression.

III). In regard to form, the general linear model is expressed as:

$$y_i = \beta_o + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_{p-1} x_{i,p-1} + \varepsilon_i$$

where

y_i is the ith response.

$\beta_o, \beta_1, \cdots \beta_{p-1}$ are the regression parameters.

$x_{i1}, x_{i2}, \cdots x_{i,p-1}$ are known constants.

ε_i is the independent random error associated with the ith response, typically assume to be distributed $N(0, \sigma^2)$.

In matrix notation the general linear model is expressed as:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

where

$\mathbf{y}$ = n x 1 vector of response values.
$\mathbf{X}$ = n x p matrix of known constants (covariates).
β = p x 1 vector of regression parameters.
$\mathbf{e}$ = n x 1 vector of identically distributed random errors, typically assume to be distributed $N(0,\sigma^2)$.

In matrix notation, the general linear model components are:

$$\mathbf{y}_{(n \times 1)} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \mathbf{X}_{(n \times p)} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1,p-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n,p-1} \end{bmatrix}$$

$$\boldsymbol{b}_{(p \times 1)} = \begin{bmatrix} b_o \\ b_1 \\ \vdots \\ b_{p-1} \end{bmatrix} \qquad \boldsymbol{e}_{(n \times 1)} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

The Y and X Model Components

The column vector **y** consists of random variables with a continuous scale measure, while the column vectors of **X** may consist of:

- continuous scale measures, or functions of continuous scale measures (e.g. polynomials, splines).
- binary indicators (e.g. gender).
- nominal or ordinal classification variables (e.g. age class).
- product terms computed between the values of two or more of the columns of **X**; referred to as interaction terms.

Examples of Linear Models

a) $y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots \beta_{p-1} x_{i,p-1} + \varepsilon_i$

b) $y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,1}^2 + \beta_3 x_{i,2} + \beta_4 x_{i,1} x_{i,2} + \varepsilon_i$

c) $y_i = \beta_0 + \beta_1 \sqrt{x_{i,1}} + \beta_2 \log_e(x_{i,2}) + \beta_3 \exp(x_{i,3}) + \varepsilon_i$

* Note that in each case the value y_i is linear in the model parameters.

Examples of Non-Linear Models

a) Exponential Model

$$y_i = \gamma_0 + \gamma_1 \exp(\gamma_2 x_i) + \varepsilon_i$$

b) Logistic Model

$$y_i = \frac{\gamma_0}{1 + \gamma_1 \exp(\gamma_2 x_i)} + \varepsilon_i$$

c) Weibull Model

$$y_i = \alpha - \beta \exp(-\gamma x_i^{\delta}) + \varepsilon_i$$

IV) Parameter Estimation for the Ordinary Least Squares Model.

a) For the estimation of the vector **b,** we minimize *Q*

$$Q = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - , \cdots , - \beta_{p-1} x_{i,p-1})^2$$

by simultaneously solving the p normal equations.

$$\frac{\partial Q}{\partial \beta} = 0 = \begin{bmatrix} \frac{\partial Q}{\partial \beta_o} = 0 \\ \frac{\partial Q}{\partial \beta_1} = 0 \\ \vdots \\ \frac{\partial Q}{\partial \beta_{p-1}} = 0 \end{bmatrix}$$

In Matrix Notation

We minimize

$$Q = (\mathbf{y} - \mathbf{X}\mathbf{b})'(y-\mathbf{X}\mathbf{b})$$

$$\frac{\partial}{\partial \boldsymbol{b}} [(\mathbf{y} - \mathbf{X}\mathbf{b})'(y - \mathbf{X}\mathbf{b})] = 0$$

$$2(\mathbf{X}'\mathbf{X})\beta = 2\mathbf{X}'\mathbf{y}$$

$$(\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{X}'\mathbf{y}$$

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

with the resulting estimator for β expressed as:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

b) Estimation of the variance of y_i; symbolically expressed as $\sigma^2\{y_i\}$.

Let **e** denote the vector of residuals from the model fit

$$\mathbf{e}_{(n \times 1)} = \mathbf{y} - \mathbf{Xb}$$

$$= (\mathbf{y} - \hat{\mathbf{y}}).$$

The sum of squares error (SSE) equals

$$SSE_{(1 \times 1)} = \mathbf{e'e}$$

$$= (\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}}).$$

and the estimator for $\sigma^2\{y_i\}$ is expressed as:

$$\hat{s}^2\{y_i\} = \frac{SSE}{n - p}$$ where p = number of regression parameters.

Typically, $\hat{s}^2\{y_i\}$ is referred to as the residual MSE.

c) Estimation of the variance of **b**; symbolically expressed as $\sigma^2\{\mathbf{b}\}$.

Since

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

$$\text{Var}(\mathbf{b}) = [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'][(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']'\text{Var}(\mathbf{y})$$

$$= \text{Var}(\mathbf{y})[(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}]$$

$$= s^2\{y_i\}(\mathbf{X}'\mathbf{X})^{-1}.$$

the estimator for $\sigma^2\{\mathbf{b}\}$ is expressed as:

$$\hat{s}^2\{\mathbf{b}\} = \hat{s}^2\{y_i\}(\mathbf{X}'\mathbf{X})^{-1}$$

$$= \text{MSE}(\mathbf{X}'\mathbf{X})^{-1}.$$

d) Estimation of the variance of $\hat{y}_i$; symbolically expressed as $s^2\{\hat{y}_i\}$.

Since

$$\hat{\mathbf{y}}_\mathbf{i} = \mathbf{x}_{\mathrm{i},(1..p)}\mathbf{b}$$

$$s^2\{\hat{\mathbf{y}}_\mathbf{i}\} = [\mathbf{x}'_{\mathrm{i},(1..p)}\, s^2\{\mathbf{b}\}\, \mathbf{x}_{\mathrm{i},(1..p)}]$$

the estimator for $s^2\{\hat{y}_i\}$ is expressed as:

$$\hat{s}^2\{\hat{\mathbf{y}}_\mathbf{i}\} = [\mathbf{x}'_{\mathrm{i},(1..p)}\, \hat{s}^2\{\mathbf{y}_\mathrm{i}\}\, (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_{\mathrm{i},(1..p)}]$$

$$= \hat{s}^2\{\mathbf{y}_\mathrm{i}\}[\mathbf{x}'_{\mathrm{i},(1..p)}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_{\mathrm{i},(1..p)}]$$

$$= \mathrm{MSE}\,[\mathbf{x}'_{\mathrm{i},(1..p)}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_{\mathrm{i},(1..p)}]$$

V) Statistical Inference for the Least Squares Regression Model.

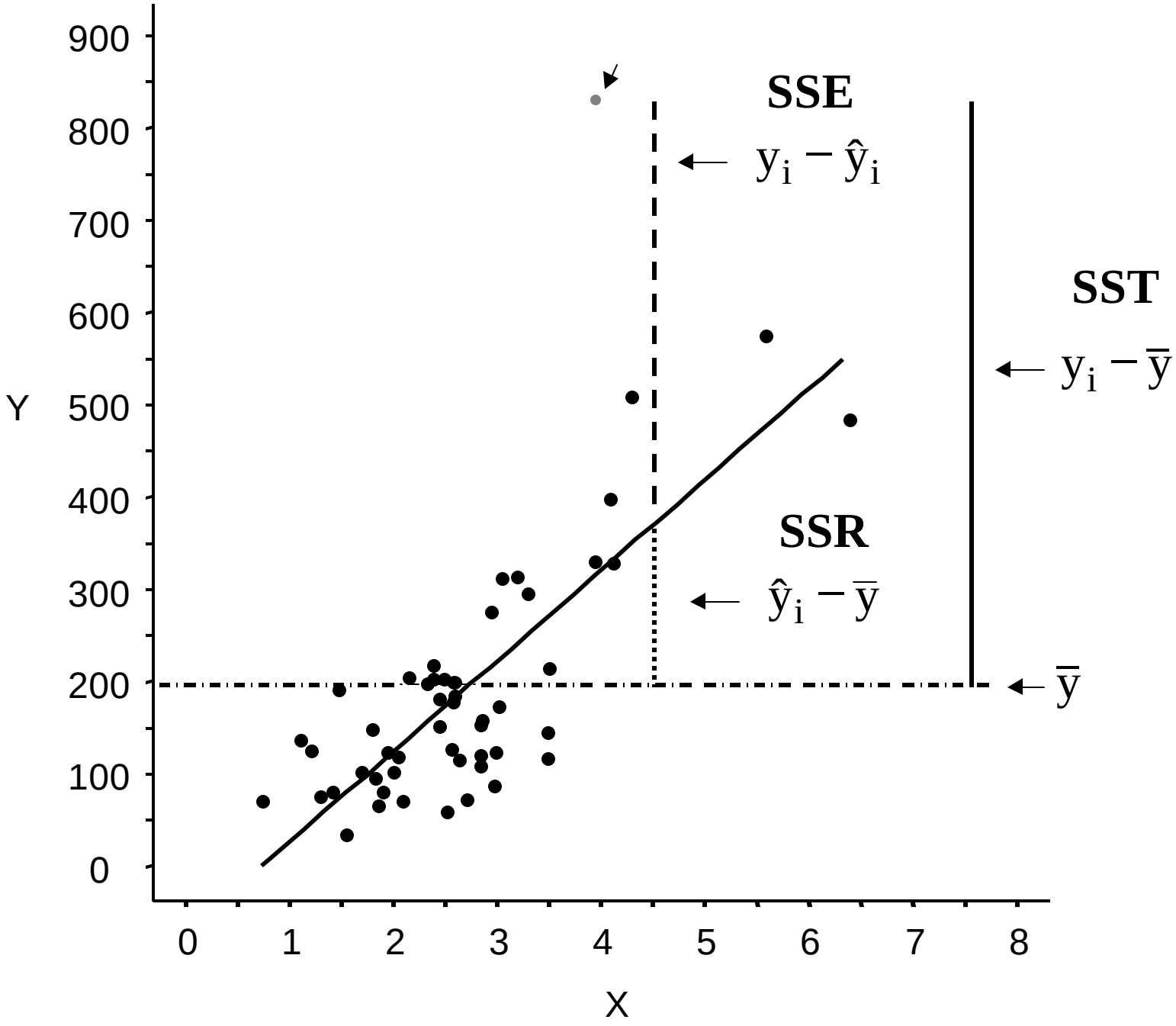
a) ANOVA sum of squares decomposition.

$$\text{SS Total} = \text{SS Regression} + \text{SS Error}$$

$$\sum_{i=1}^{n}(y_i - \bar{y}.)^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y}.)^2 + (y_i - \hat{y}_i)^2$$

Table 1. Regression ANOVA table.

Source	SS	DF	MS	F-test
Regression	SSR	p-1	SSR/(p-1)	MSR/MSE
Error	SSE	n-p	SSE/(n-p)	
Total	SST	n-1		


900
800
700
600
500
400
300
200
100
0
Y
0
1
2
3
4
5
6
7
8
X
SSE
$y_i - \hat{y}_i$
SST
$y_i - \bar{y}$
SSR
$\hat{y}_i - \bar{y}$
$\bar{y}$

b) Hypothesis tests related to **b**

$$H_o: \beta_i = 0$$
$$H_a: \beta_i \neq 0$$

Test Statistic

$$t^* = \frac{b_i - 0}{\sqrt{\hat{s}^2\{b_i\}}}$$

$$= \frac{b_i - 0}{\sqrt{MSE(X'X)^{-1}_{i,i}}} \qquad \text{where } t^* \sim t(n-p)$$

If $|t^*| < t(1-\alpha/2; n-p)$, conclude H_o

If $|t^*| \geq t(1-\alpha/2; n-p)$, conclude H_a

c) Confidence limits for the $\hat{y}_i$ at a specific x_i

$$(1-\alpha)\%\,\text{CL} = \hat{y}_i \pm t(1-\alpha/2, n-p)\,\hat{s}\{\hat{y}_i\}$$

$$= \hat{y}_i \pm t(1-\alpha/2, n-p)\sqrt{\text{MSE}(x'_i (X'X)^{-1} x_i)}$$

d) Prediction limits for a new y_i at a specific x_i.

$$(1-\alpha)\%\text{PL} = \hat{y}_i \pm t(1-\alpha/2, n-p)\,\hat{s}\{y_{new,i}\}$$

$$= \hat{y}_i \pm t(1-\alpha/2, n-p)\sqrt{\text{MSE}\left(1+\frac{1}{n}+x'_i (X'X)^{-1} x_i\right)}$$

e) Simultaneous confidence band for the regression line $Y=X\beta$.

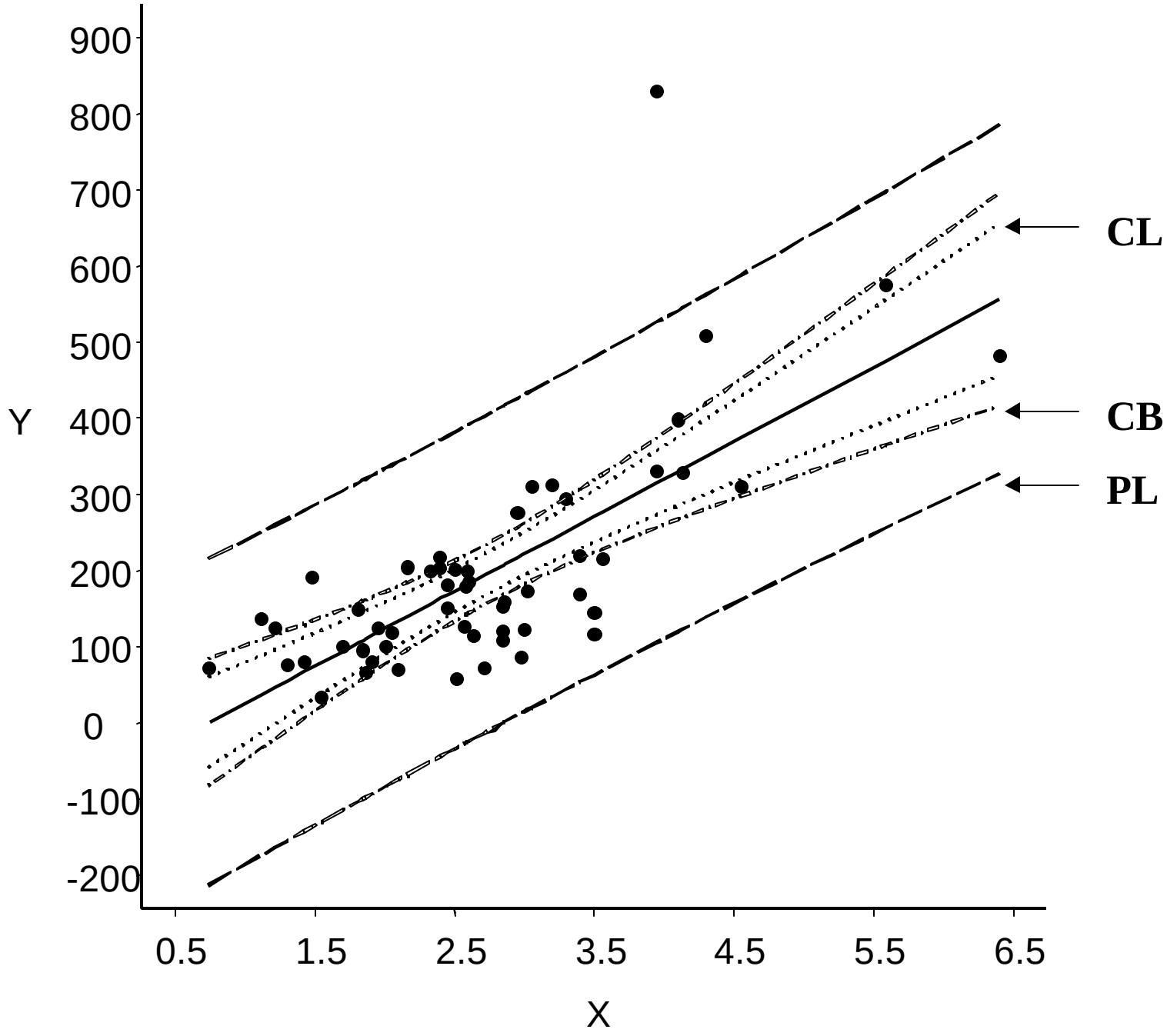
$$(1-\alpha)\%CB = \hat{y}_i \; +/- \; \sqrt{pF_{(1-\alpha;\, p,\, n-p)}\, \hat{s}^2\{\hat{y}_i\}}$$

$$(1-\alpha)\%CB = \hat{y}_i \; +/- \; \sqrt{pF_{(1-\alpha;\, p,\, n-p)} MSE[x'_i (X'X)^{-1} x_i]}$$

where

p = the number of regression parameters.

$F_{(1-\alpha;\, p,\, n-p)}$ = the critical value of a F-distribution with p and n - p df evaluated at the $(1-\alpha)100$ percentile.


900
800
700
600
500
400
300
200
100
0
-100
-200
Y
0.5
1.5
2.5
3.5
4.5
5.5
6.5
X
CL
CB
PL

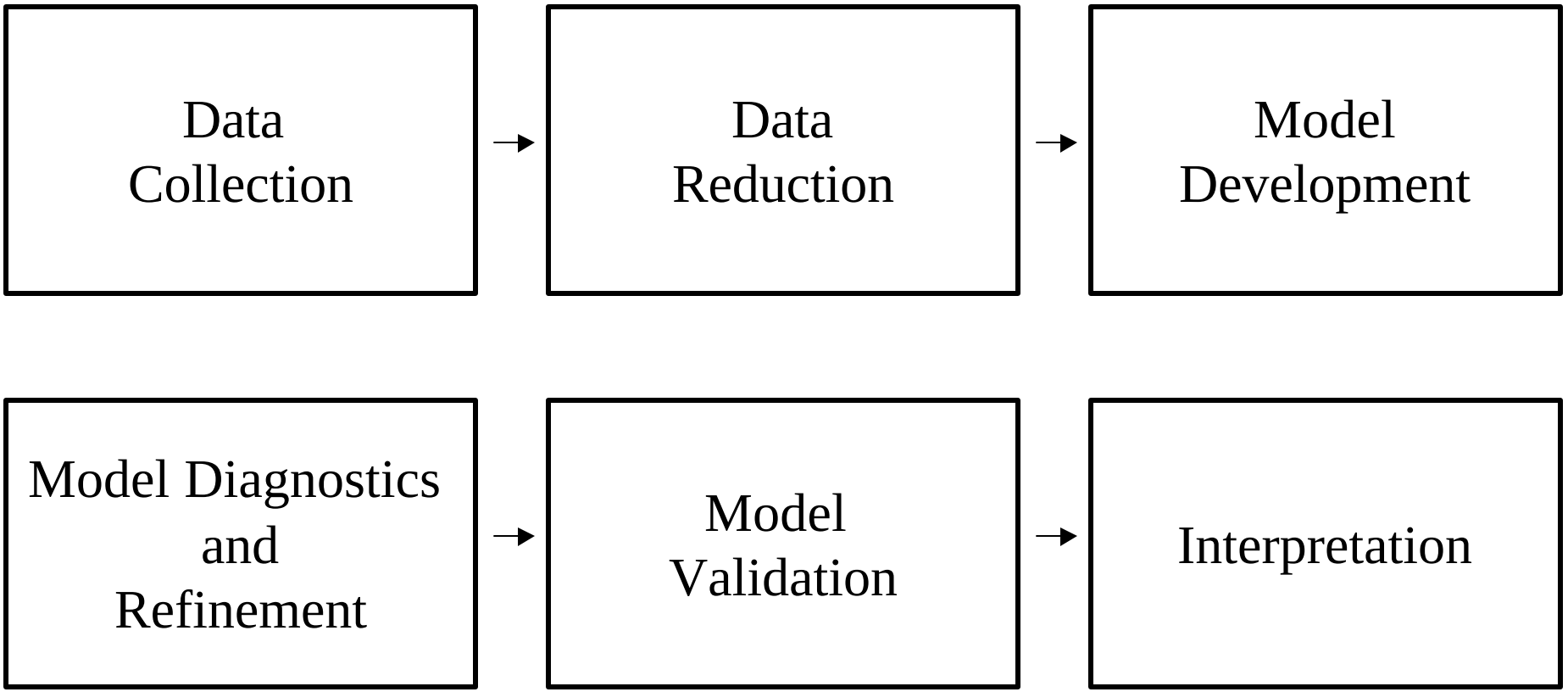
When to Use CL, CB, and PL.

- Apply CL when your goal is to predict the expect value of y_i for one specific vector of predictors x_i within the range of the data.

- Apply CB when your goal is to predict the expect value of y_i for all vectors of predictors x_i within the range of the data.

- Apply PL when your goal is to predict the value of y_i for one specific vector of predictors x_i within the range of the data.

VI) Overview of the Model-Building Process



a) Data Collection

- Controlled experiments.

  Covariate information consists of explanatory variables that are under the experimenter's control.

- Controlled experiments with supplemental variables.

  Covariate information includes supplemental variables related to the characteristics of the experimental units, in addition to the variables that are under the experimenter's control.

- Exploratory observational studies.

  Covariate information may include a large number of variables related to the characteristics of the observational unit, none of which are under the investigator's control.

b) Data Reduction

- Controlled experiments.

  Variable reduction is typically not required because the explanatory variables of interest are predetermined by the experimenter.

- Controlled experiments with supplemental variables.

  Variable reduction is typically not required because the primary explanatory variables of interest, as well as the supplemental variables of interest are predetermined by the experimenter.

- Exploratory observational studies.

  Variable reduction is typically required because numerous sets of explanatory variables will be examined. As a rule of thumb, there should be at least 6-10 observations for every explanatory variable in the model. (e.g. 5 predictors, 50 observations).

Data Reduction Methods.

- Rank Predictors

  Rank your predictors based on their general importance with respect to the subject matter. Select the most important predictors using the rule of thumb that for each predictor you need 6-10 independent observations.

- Cluster Predictors

  Cluster your predictors variables base on a similarity measure, choosing one or perhaps two predictors from within each unique cluster. (“Cluster Analysis” in Johnson et. al 1999).

- Create a Summary Composite Measure

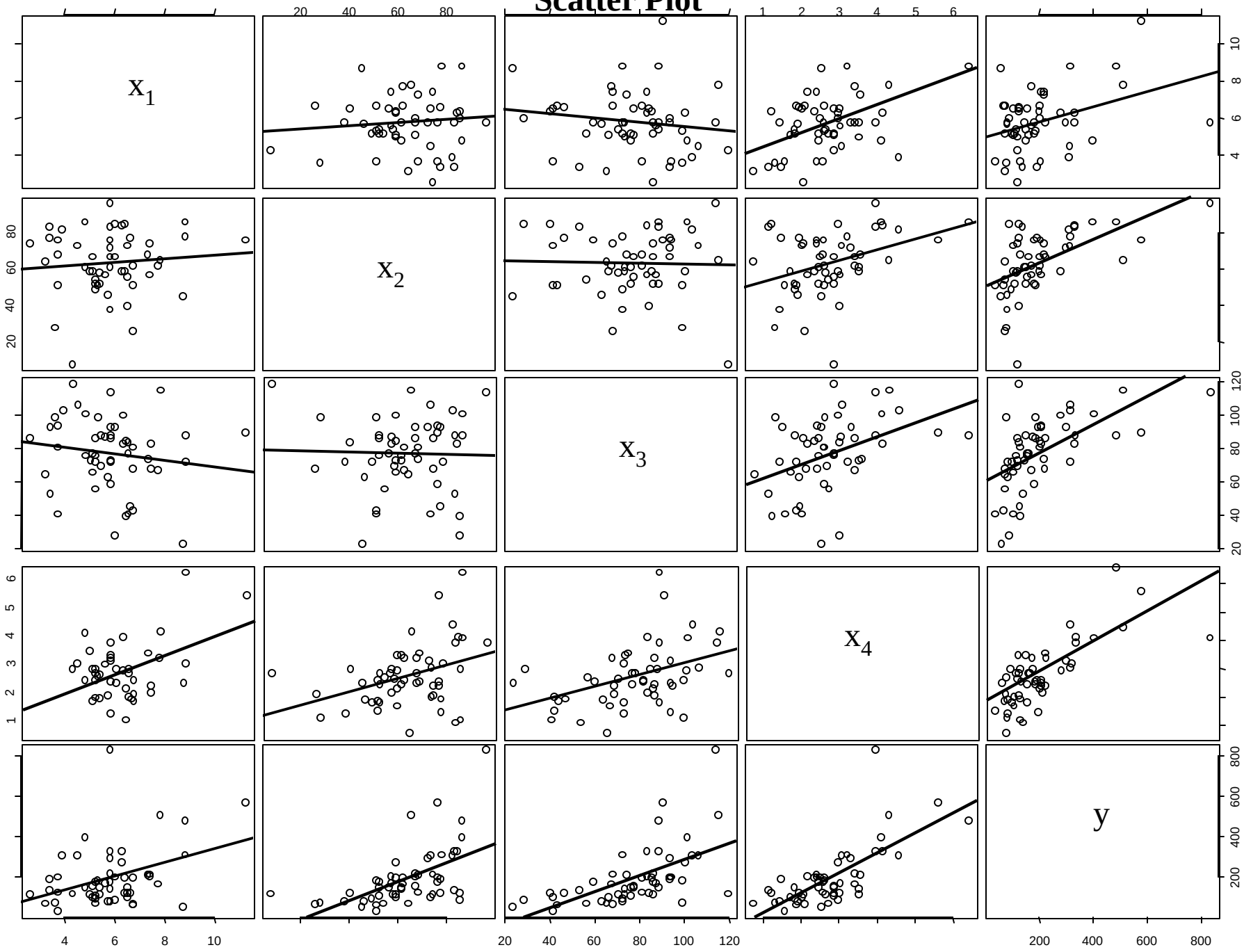
  Produce a composite summary measure (score) that is base on the original set of predictors, which still retains the majority of the information that is contained within the original set of predictors (“Principle Components Analysis”, in Johnson et. al 1999).

c) Model Development.

- Model development should first and foremost be driven by your knowledge of the subject matter and by your hypotheses.
- Graphics, such as a scatter-plot matrix can be utilized to initially examine the univariate relationship between each explanatory variable and the response, as well as the relationship between each pair of the explanatory variables.
- Constructing a correlation matrix may also be informative with regard to quantitatively assessing the degree of the linear association between each explanatory variable and the response variable, as well as between each pair of the explanatory variables.

+Note, that two explanatory variables that are highly correlated essentially provide the same information.

Scatter Plot



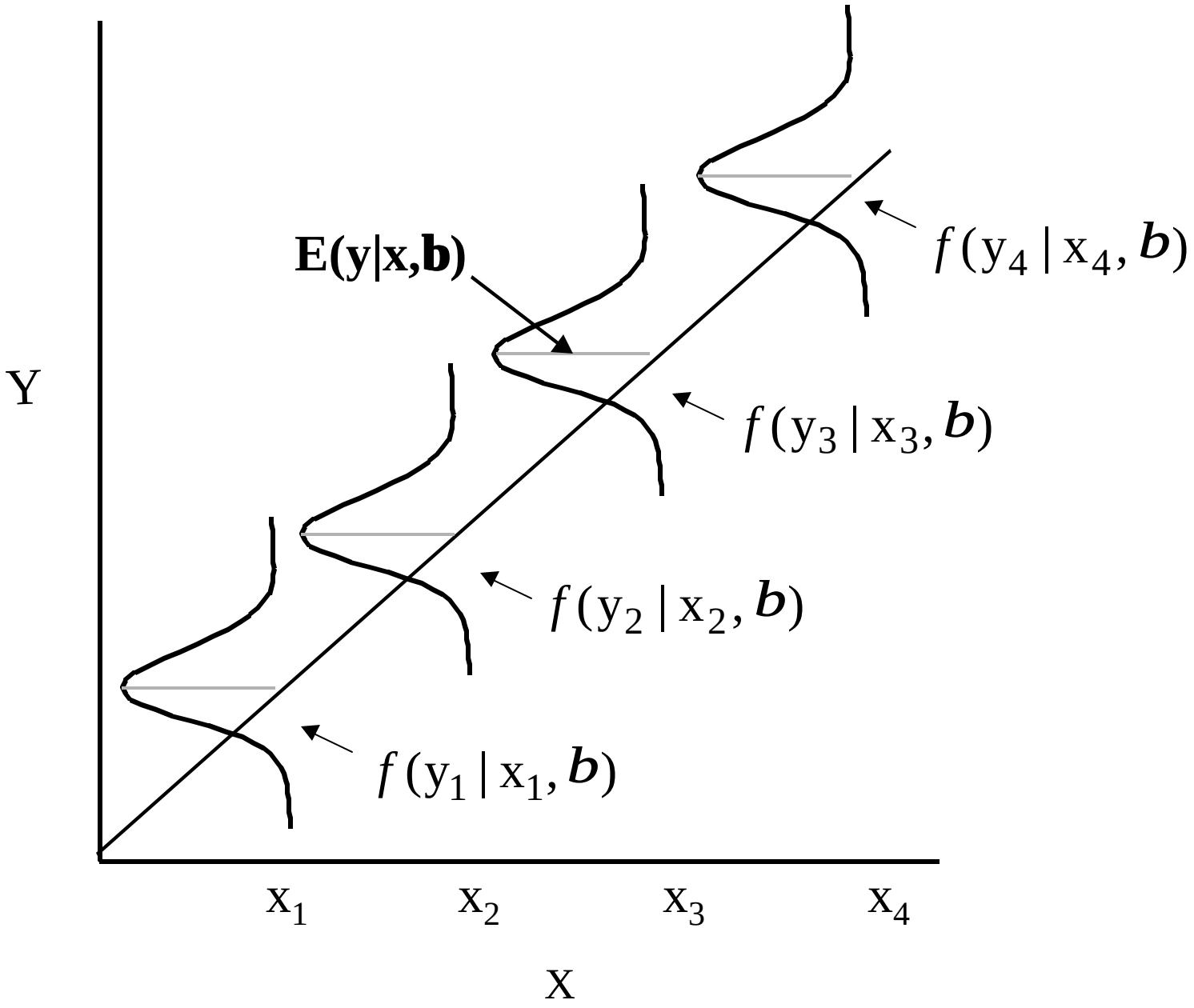
d) Model Diagnostics and Refinement.

Once you have fit your initial regression model the things to assess include the following:

- The Assumption of Constant Variance.
- The Assumption of Normality.
- The Correctness of Functional Form.
- The Assumption of Additivity.
- The Influence of Individual Observations on Model Stability.

+All these model assessments can be carried out by using standard residual diagnostics provided in statistical packages such as SAS .

Model Assumptions



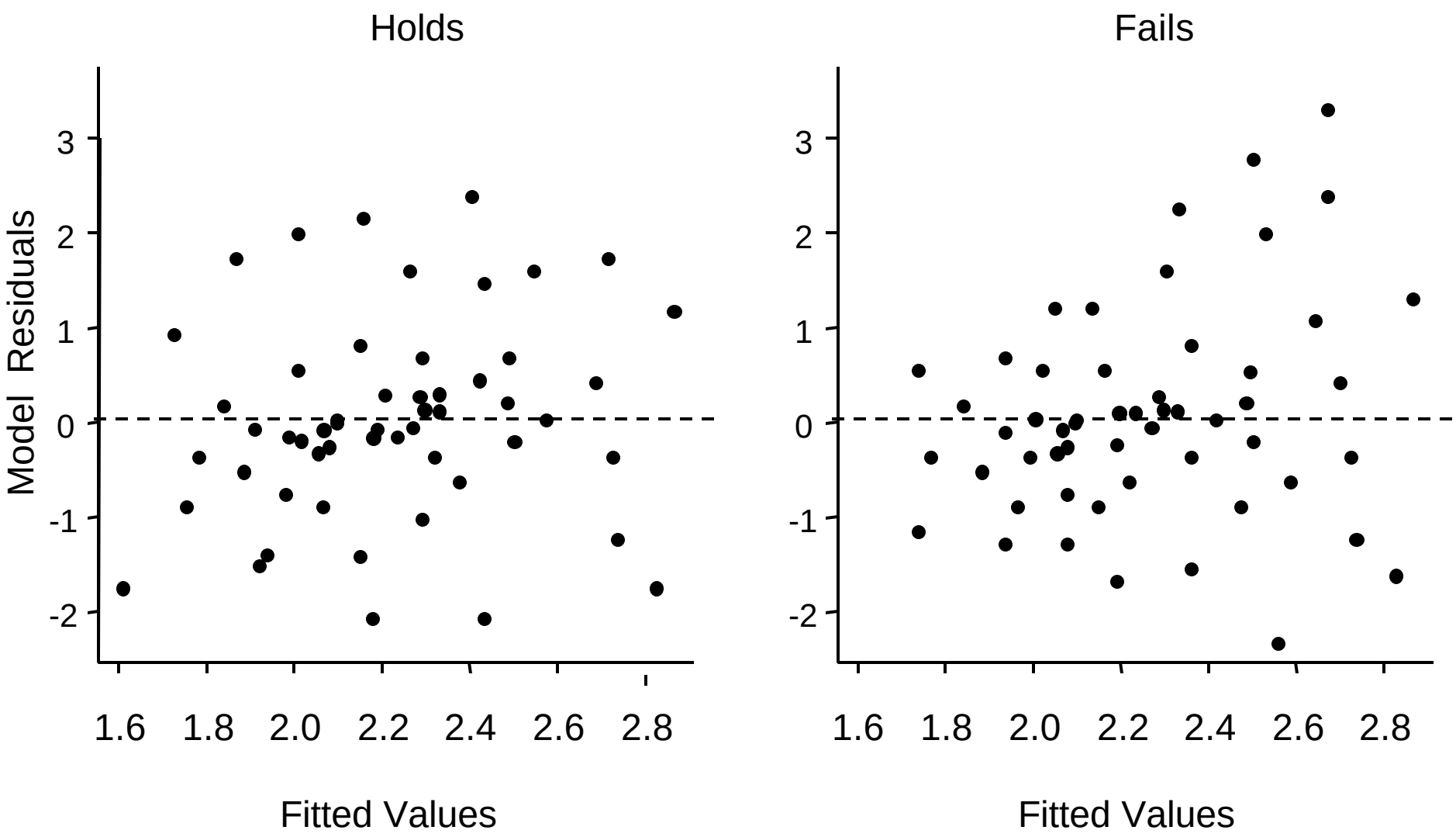
- Assessment of the Assumption of Constant Variance.

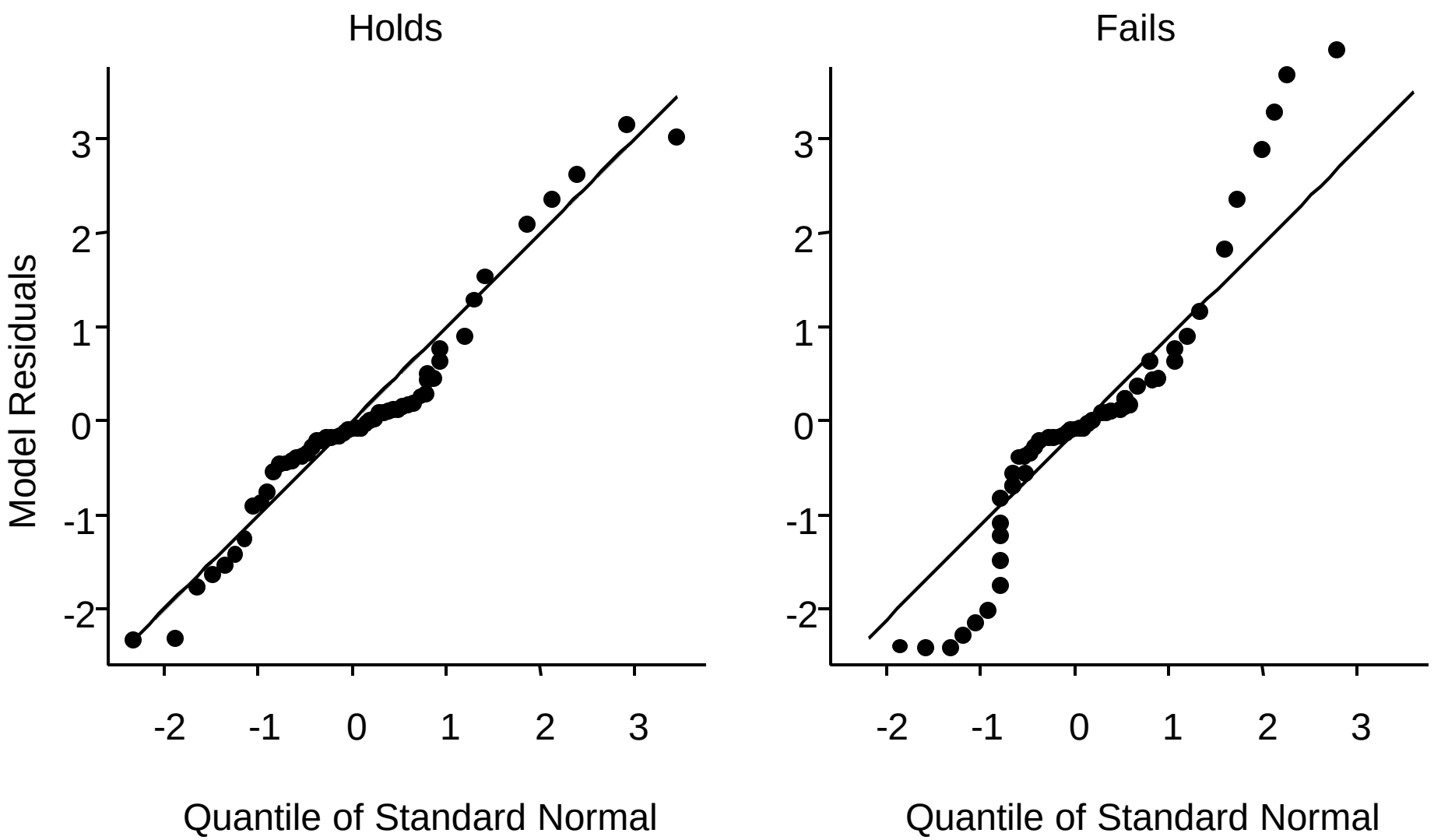
  Plot the residuals from your the model versus the fitted values. Examine if the variability between the residuals remains relatively constant across the range of the fitted values.

- Assessment of the Assumption of Normality.

  Plot the residuals from your model versus their expected value under normality (Normal Probability Plot). The expected value of the kth order residual (e_k) is determined by the formula:

  $E(e_k) = \sqrt{MSE}\left[z\left(\frac{k - 0.375}{n + 0.25}\right) \right]$. Where MSE is the estimated residual variance from the regression model, $z(\alpha)$ denotes the (α) quantile value of the standard normal distribution, and n is the total sample size.


Assumption of Constant Variance
Holds
Fails
Model Residuals
3
2
1
0
-1
-2
1.6
1.8
2.0
2.2
2.4
2.6
2.8
Fitted Values
Fitted Values


Assumption of Normality
Holds
Fails
Model Residuals
3
2
1
0
-1
-2
-2
-1
0
1
2
3
Quantile of Standard Normal
Quantile of Standard Normal

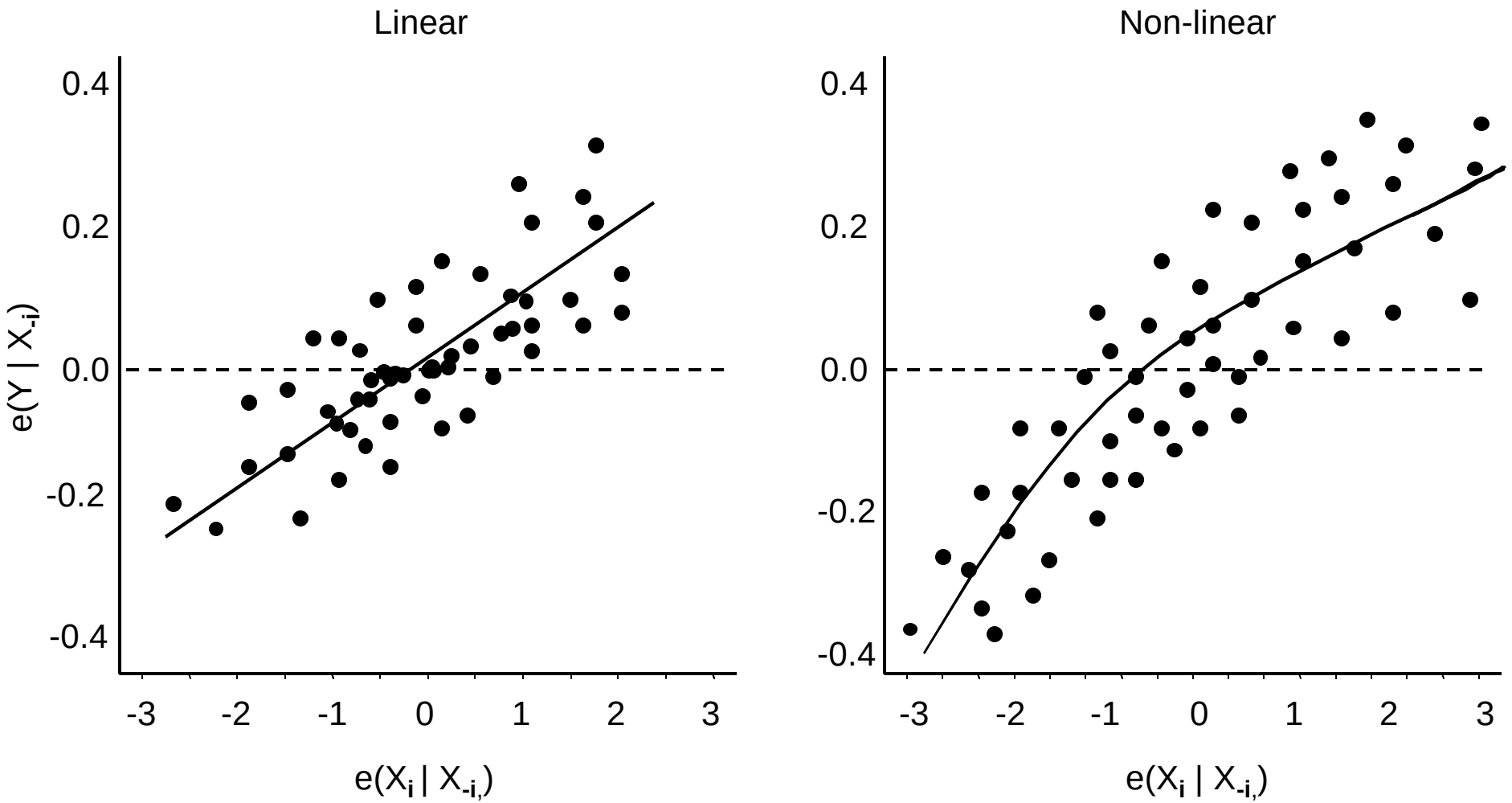
- Assessment of Correct Functional Form.

  Plot the residuals from a model which excludes the explanatory variable of interest against the residuals from a model in which the explanatory variable of interest is regressed on the remaining explanatory variables. This type of plot is referred to as a partial regression plot (Neter et al. 1996). Examine whether there is a non-linear relationship between the two sets of residuals. If there is a non-linear relationship, fit a higher order term in addition to the linear term.

- Assessment of the Assumption of Additivity.

  For each plausible interaction use a partial residual plot to examine whether or not there is a systematic pattern in the residuals. If so it may indicate that by adding the interaction term to the model it will enhance your model's predictive performance.

Partial Regression Plots



- Assess the Influence of Individual Observations on Model Stability.

1) Identify outlying Y observations by examining the studentized residuals

$$r_i = \frac{e_i}{\sqrt{MSE(1-h_{ii})}}$$

where e_i is the value of the ith residual, MSE is the residual variance, and $h_{ii} = x'_i (X'X)^{-1} x_i$ (hat matrix).

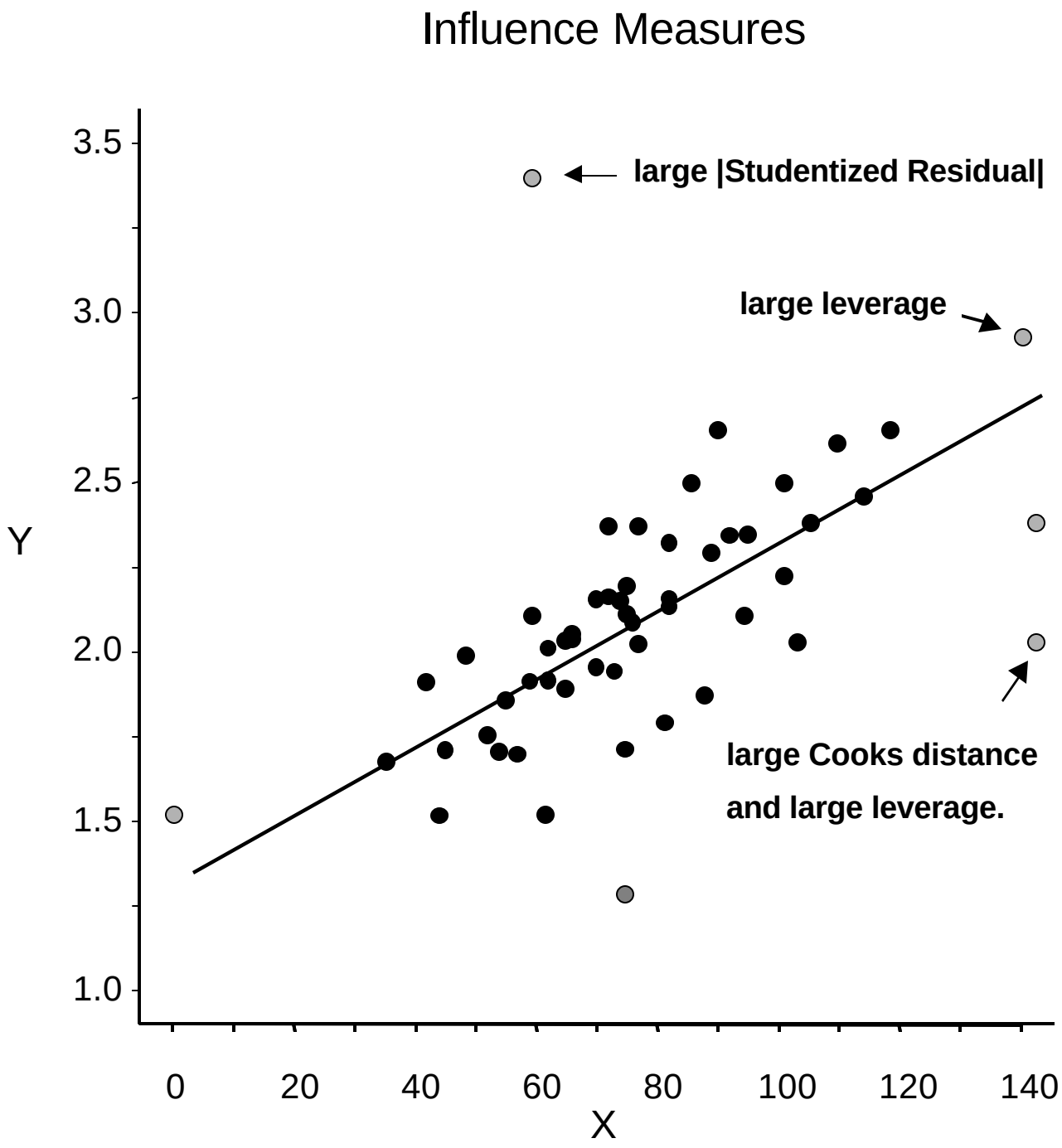
2) Identify outlying X observations by examining what is referred to as the "leverage".

$$h_{ii} = x'_i (X'X)^{-1} x_i$$

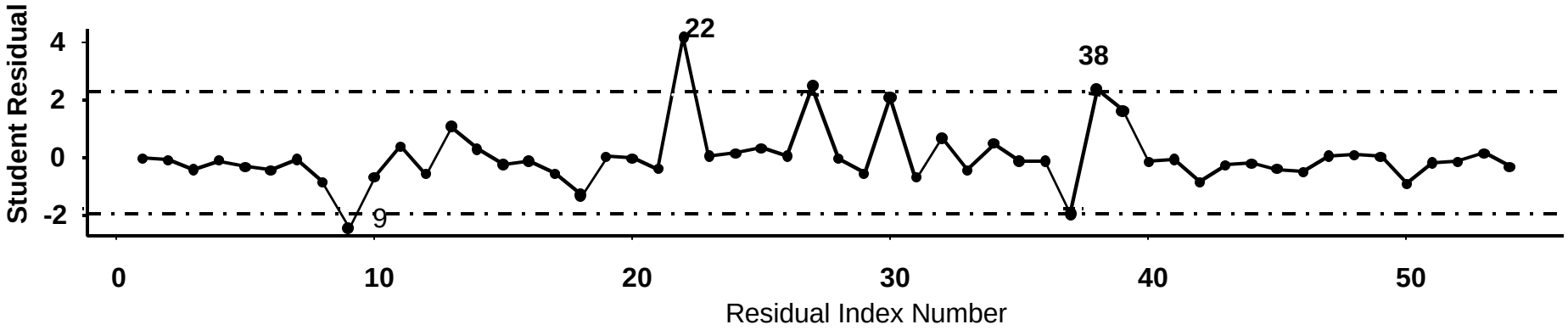
3) Evaluate the influence of the *i*th case on all n fitted values by examining the Cook distance (D_i).

$$D_i = \frac{e_i^2}{pMSE}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right]$$

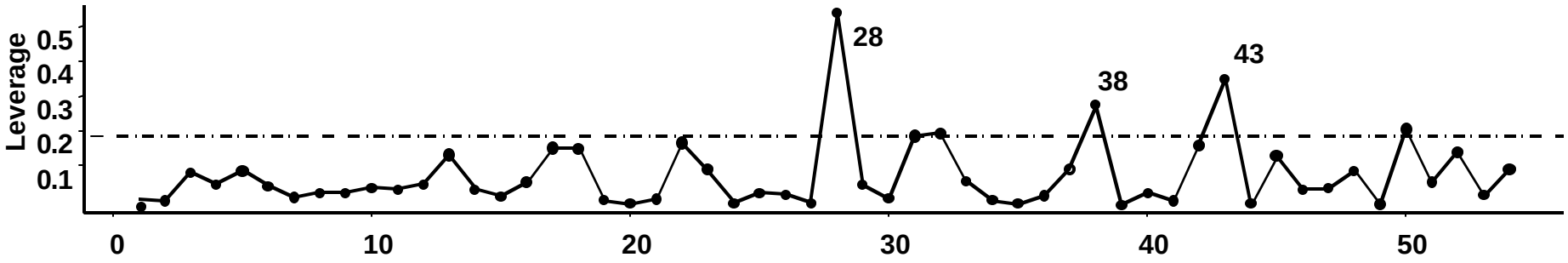
where p is the number of model parameter, and e_i, MSE, and h_{ii} are defined as previously stated.


Influence Measures
large |Studentized Residual|
large leverage
large Cooks distance
and large leverage.
Y
X
3.5
3.0
2.5
2.0
1.5
1.0
0
20
40
60
80
100
120
140

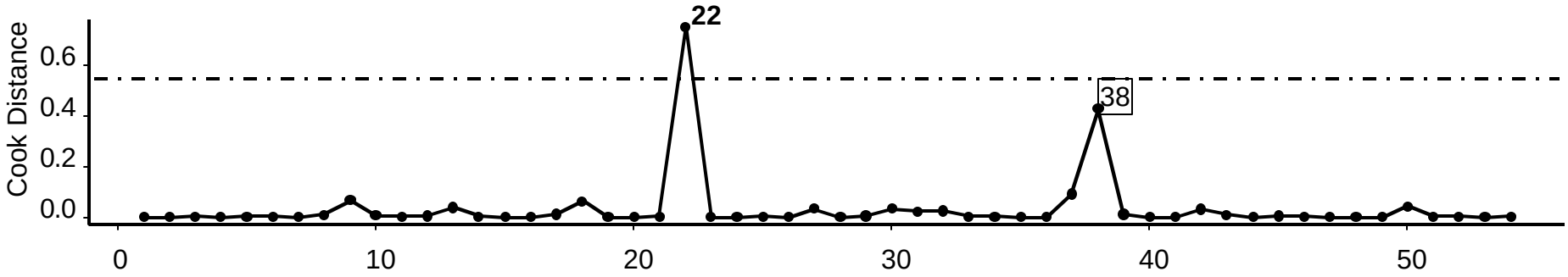
Student Residual



Residual Leverage



Cooks Distance



Some Remedial Measures

- Non-Constant Variance

  A transformation of the response variable to a new scale (e.g. log) is often helpful in attaining equal residual variation across the range of the predicted values. The Box-Cox transformation method (Neter et al., 1996) can be utilized to determine the proper form of the transformation. If there is no variance stabilizing transformation which rectifies the situation an alternative approach is to use a more robust estimator, such as iterative weighted least squares (Myers, 1990).

- Non-Normality

  Generally, non-normality and non-constant variance go hand in hand. An appropriate variance stabilizing transformation will more than likely also be remedial in attaining normally distributed residual error.

The Box-Cox Transformation

The Box-Cox procedure functions to identify a transformation from the family of power transformations on Y which corrects for skewness in the response distribution and for unequal error variance. The family of power transformations is of the form $Y' = Y^{\lambda}$, where λ is determined from the data. This family encompasses the following simple transformations.

Value of λ	Transformation
2.0	$Y' = Y^2$
0.5	$Y' = Y^{1/2}$
0	$Y' = \log_e(Y)$
-0.5	$Y' = 1/Y^{1/2}$
1.0	$Y' = 1/Y$

For each λ value, the Y_i^{λ} observations are first standardized so that the magnitude of the error sum of squares does not depend on the value of λ.

$$W_i = \begin{cases} K_1(Y^{l} - 1) & l \neq 0 \\ K_2(\log_e Y_i) & l = 0 \end{cases}$$

where

$$K_1 = \frac{1}{l K_2^{l-1}}$$

$$K_2 = \left(\prod_{i=1}^{n} Y_i\right)^{1/n}$$

Once the standardized observations W_i have been obtained for a given λ value, they are regressed on the predictors X and the error sum of squares SSE is obtained. It can be shown that the maximum likelihood estimate for λ is the value of λ for which SSE is minimum. We therefore choose the value λ which produces the smallest SSE.

- Outliers

Outliers, either with respect to the response variable, or with respect to the explanatory variables can have a major influence on the values of the regression parameter estimates. Gross outliers should always be checked first for data authenticity. In terms of the response variable, if there is no legitimate reason to remove the offending observations it may be informative to fit the regression model with and without the outliers. If statistical inference changes depending on the inclusion or the exclusion of the outliers, it is probably best to use a robust form of regression, such as least median squares or least absolute deviation regression (Myers, 1990). For the explanatory variables, if your data set is reasonably large, its is generally recommended to use some from of truncation that reduces the range of the offending explanatory variable.

e) Model Validation

Model validation applies mainly to those models that will be utilized as a predictive tool. Types of validation procedure include:

- External Validation.

  Predictive accuracy is determined by applying your model to a new sample of data. External validation, when feasible should always be your first choice for the method of model validation.

- Internal Validation.

  Predictive accuracy can be determined by first fitting your model to a subset of the data and then applying your model to the data that you withheld from the model building process (Cross-validation). Alternatively, measures of predictive accuracy can be evaluated by a bootstrap re-sampling procedure. The bootstrap procedure provides a measure of the optimism that is induced by optimizing your model's fit to your sample of data.

VII) An Example Case Study.

A hospital surgical unit was interested in predicting survival in patients undergoing a particular type of liver operation. A random sample of 54 patients was available for analysis. From each patient, the following information was extracted from the patient's pre-operative records.

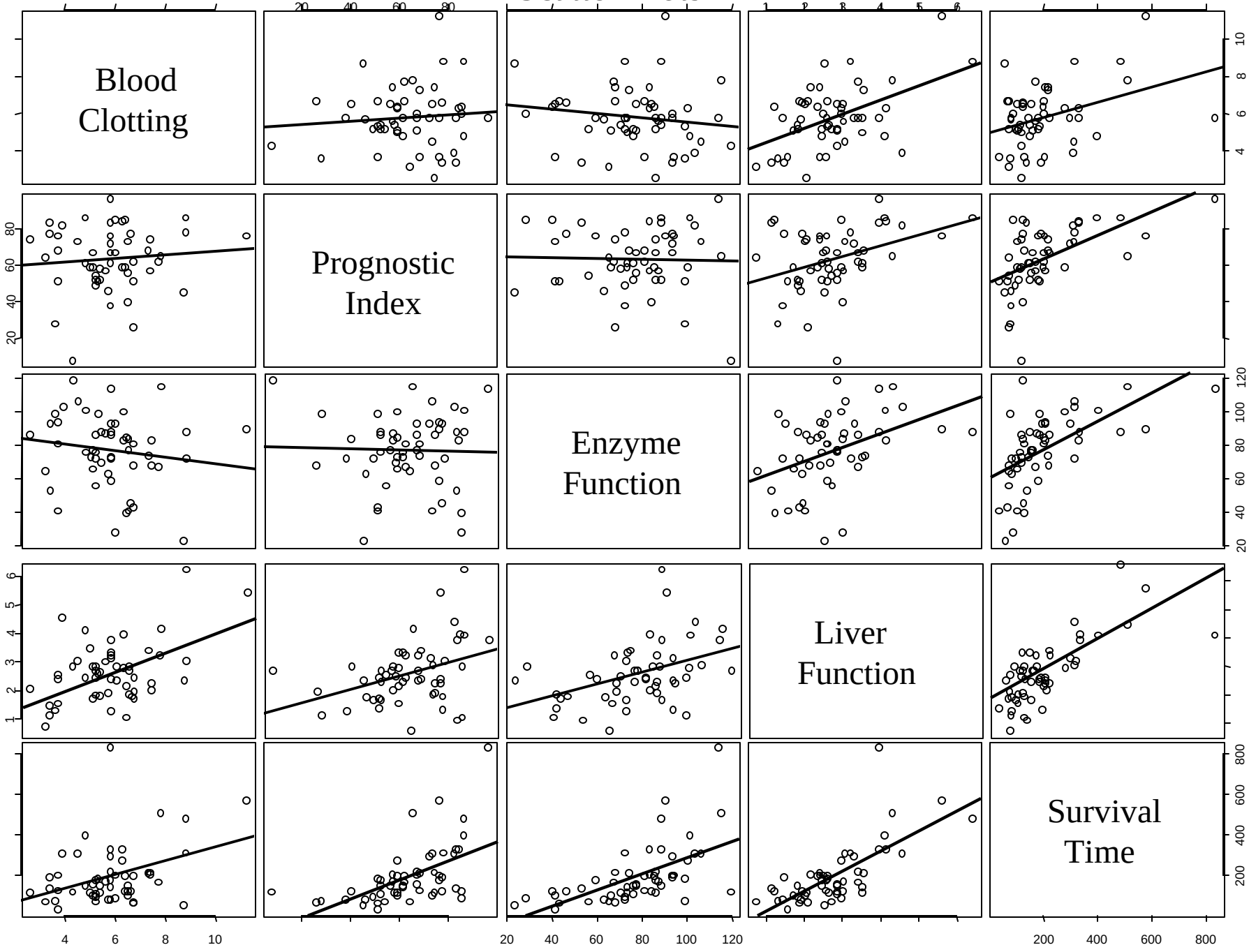
- $\mathbf{x}_1$ blood clotting score.
- $\mathbf{x}_2$ prognostic index, which includes the age of the patient.
- $\mathbf{x}_3$ enzyme function test score.
- $\mathbf{x}_4$ liver function test score.

Taken from Neter et al. (1996).

Summary Statistics

Variable	n	Mean	SD	SE	Median	Min	Max
Bld Clotting	54	5.78	1.60	0.22	5.80	2.60	11.20
Prog Index	54	63.24	16.90	2.30	63.00	8.00	96.00
Enzyme Fun	54	77.11	21.25	2.89	79.00	23.00	119.00
Liver Fun	54	2.74	1.07	0.15	2.60	0.74	6.40
Survival	54	197.17	145.30	19.77	155.50	34.00	830.00

Scatter Plots



Pearson Correlation Matrix

Correlate	Blood Clotting	Prognostic Index	Enzyme Function	Liver Function	Survival Time
Bld Clotting	1.000	0.090	-0.150	0.502	0.372
p-value		0.517	0.280	0.000	0.000
Prog. Index		1.000	-0.024	0.369	0.554
p-value			0.865	0.006	0.000
Enzyme Function			1.000	0.416	0.580
p-value				0.002	0.000
Liver Function				1.000	0.722
p-value					0.000

Things to Consider.

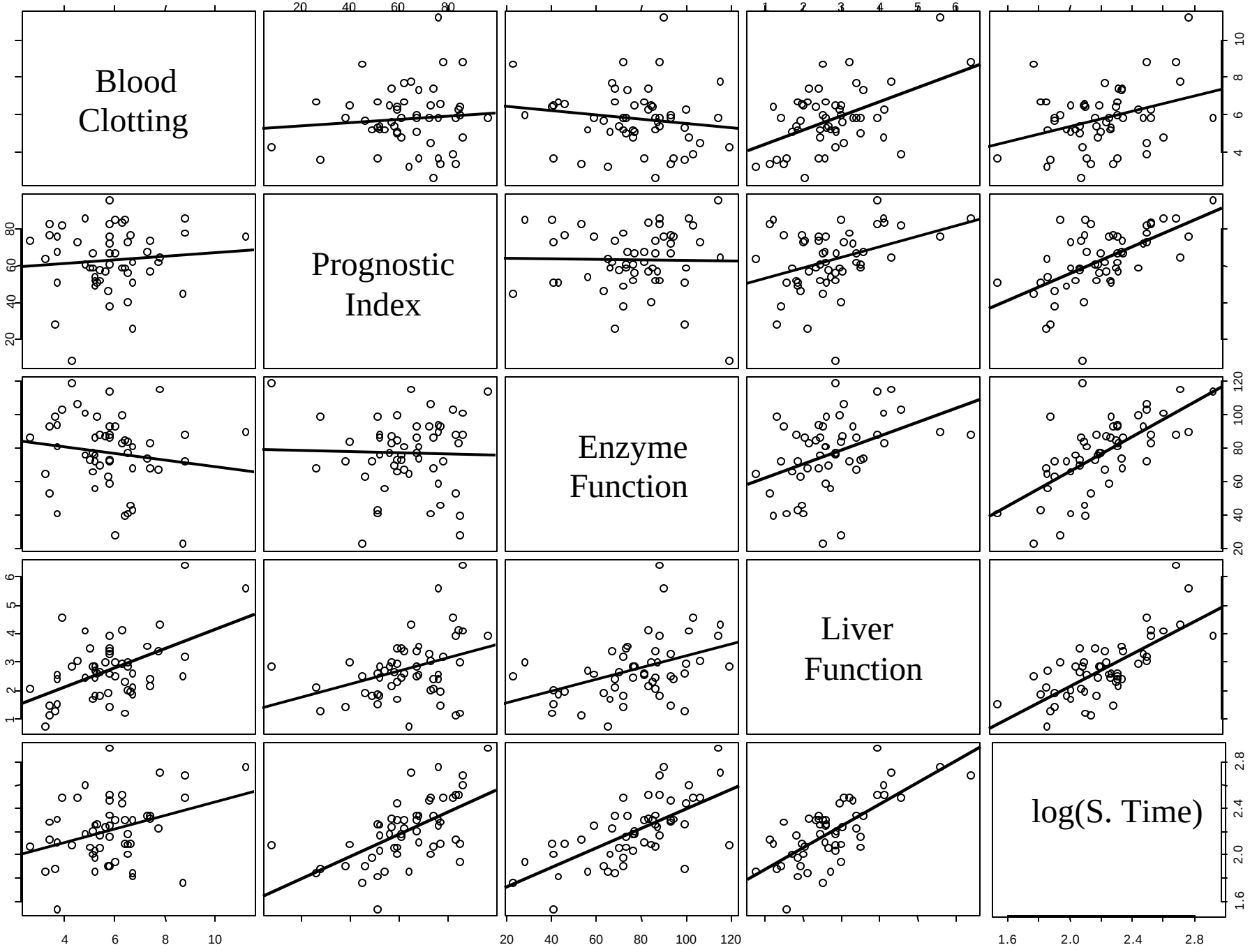
- Functional Form

  Are the explanatory variables linearly related to the response? If not, is there a transformation of the explanatory or the response variable that leads to a linear relationship. If only a few of the relationships between the explanatory and response variable are non-linear, it is best to begin by transforming the Xs, or by modeling the non-linearity.

- Multicollinearity

  Are there pairs of explanatory variables that appear to be highly correlated?. A high degree of collinearity may cause the parameter standard errors to be substantially inflated, as well as induce the regression coefficients to flip sign.

Scatter Plot



Pearson Correlation Matrix

Correlate	Blood Clotting	Prognostic Index	Enzyme Function	Liver Function	$\log_{10}$(S Time)
Bld Clotting	1.000	0.090	-0.150	0.502	0.346
p-value		0.517	0.280	0.000	0.010
Prog. Index		1.000	-0.024	0.369	0.593
p-value			0.865	0.006	0.000
Enzyme Function			1.000	0.416	0.665
p-value				0.002	0.000
Liver Function				1.000	0.726
p-value					0.000

Model Statement

$E(y|X) = \beta_o + \beta_1(\text{Bld Clotting}) + \beta_2(\text{Prog. Index}) + \beta_3(\text{Enzyme Fun.}) + \beta_4(\text{Liver Function})$.

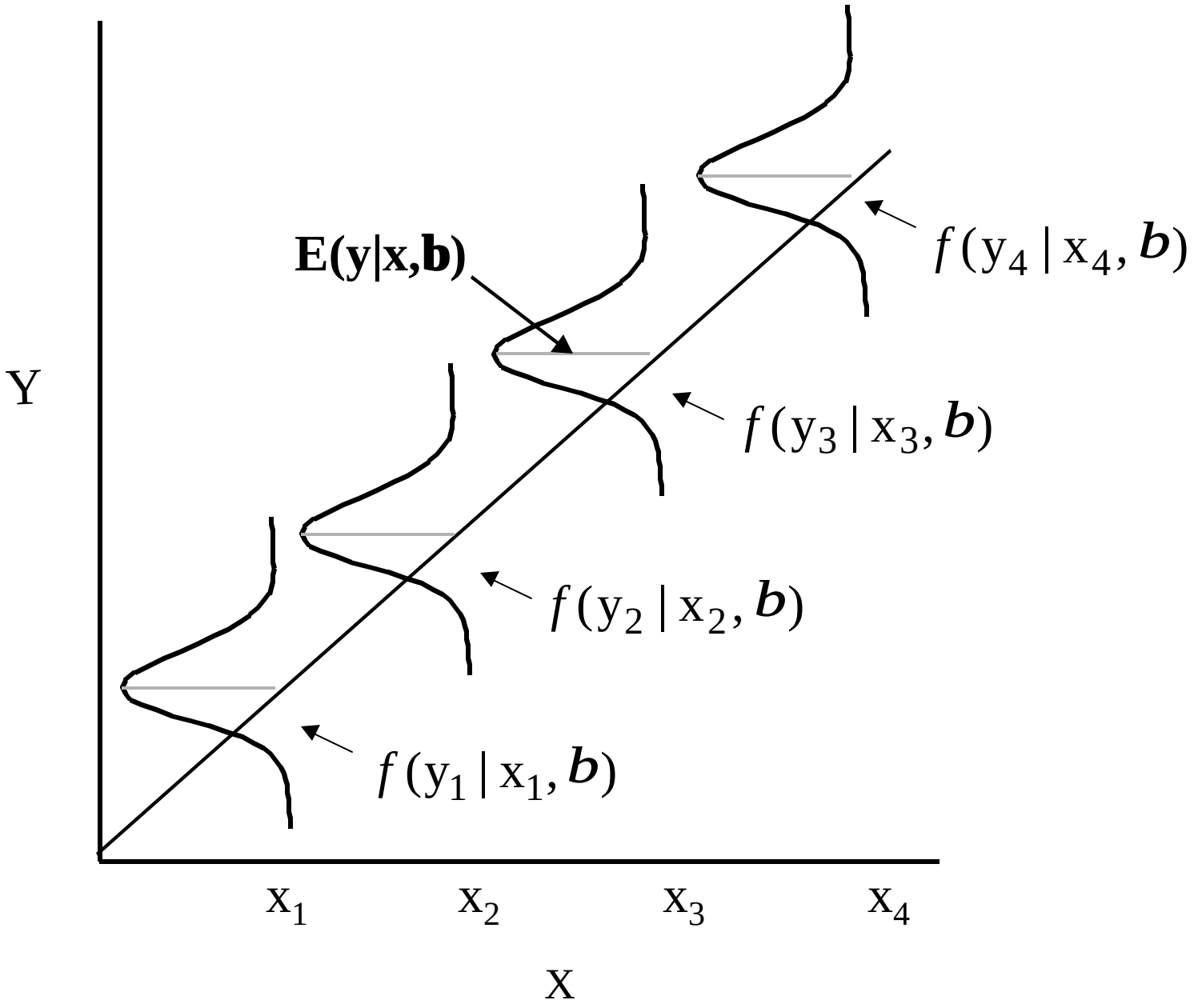
Term	b	SE(b)	t	P(T≥ t)
Intercept	0.4888	0.0502	9.729	<0.0001
Bld. Clotting	0.0685	0.0054	12.596	<0.0001
Prog. Index	0.0098	0.0004	21.189	<0.0001
Enzyme Fun.	0.0095	0.0004	23.910	<0.0001
Liver Fun.	0.0019	0.0097	0.198	0.8436

$\hat{s}(y_i)$=MSE	df	R^2	R_a^2
0.0473	49	0.9724	0.9701

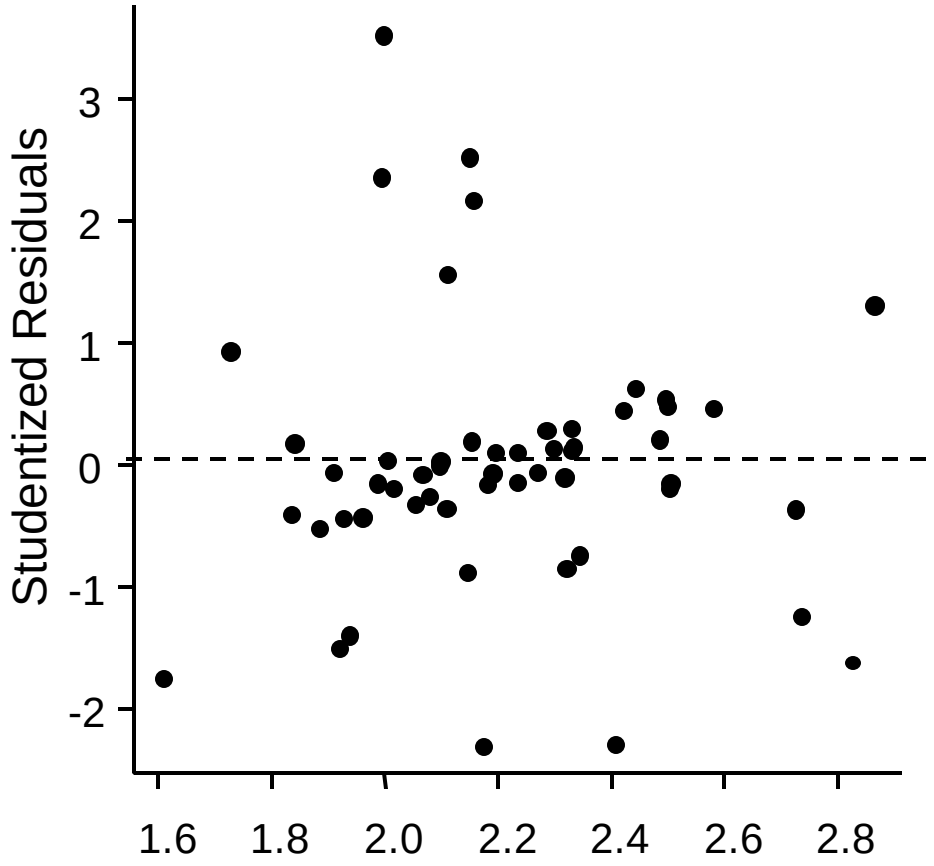
Regression ANOVA Table

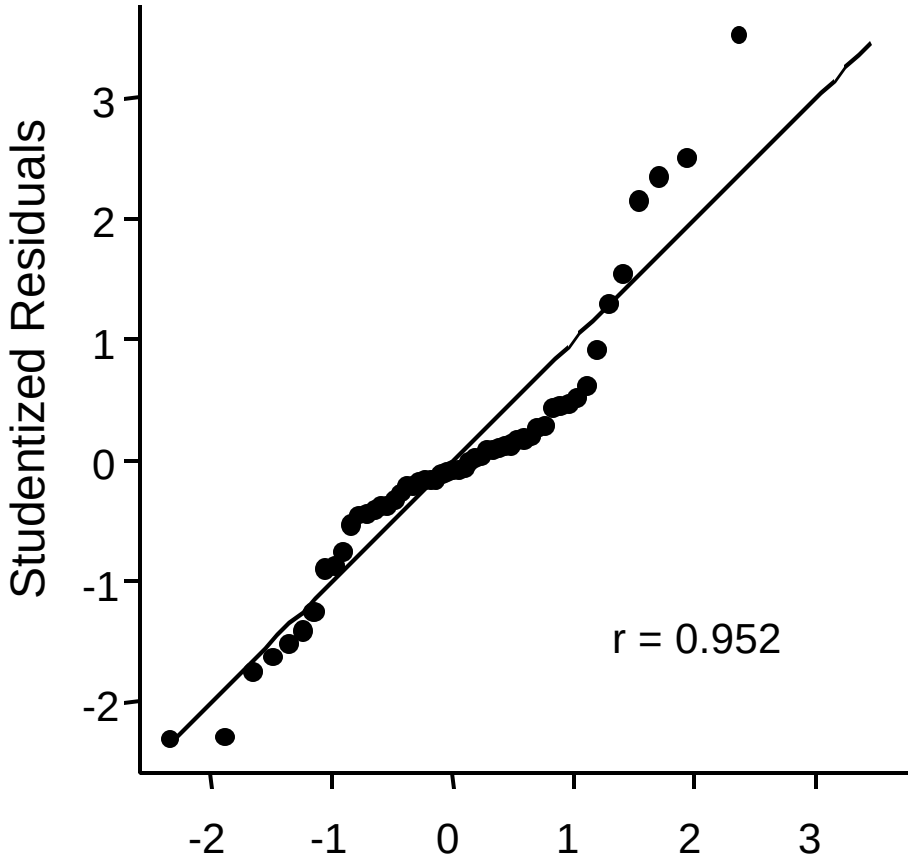
Term	df	SS	MS	f	P(F≥ f)
Regression	4	3.9727	0.9932	431.10	<0.0001
Bld. Clotting	1	0.4767	0.4767	212.79	<0.0001
Prog. Index	1	1.2635	1.2635	564.04	<0.0001
Enzyme Fun.	1	2.1122	2.1126	947.51	<0.0001
Liver Fun.	1	0.0000	0.0000	0.0393	0.8436
Residual	49	0.1097	0.0022		
Total	53	4.0477			

Model Assumptions

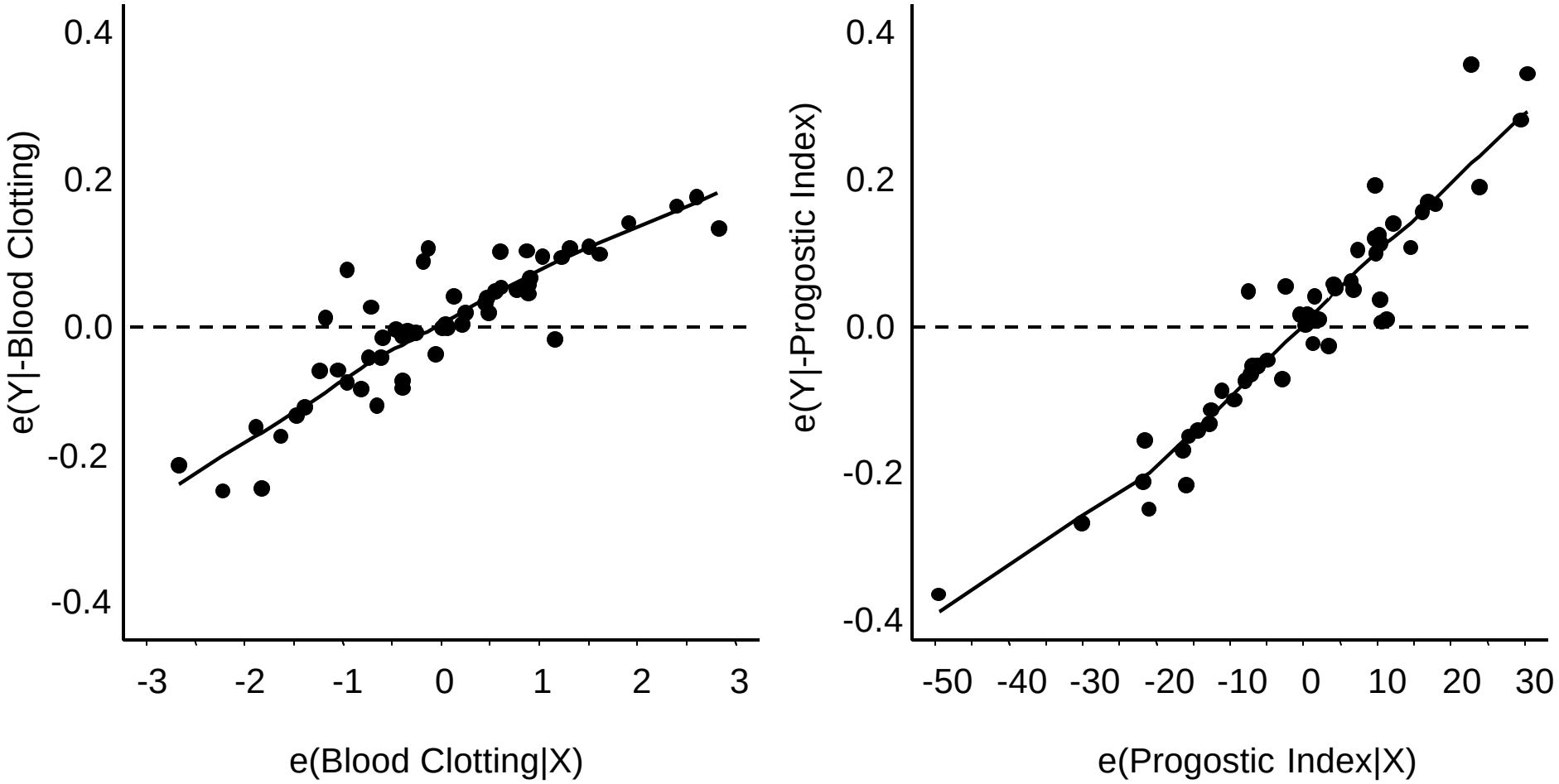


Residual Diagnostics

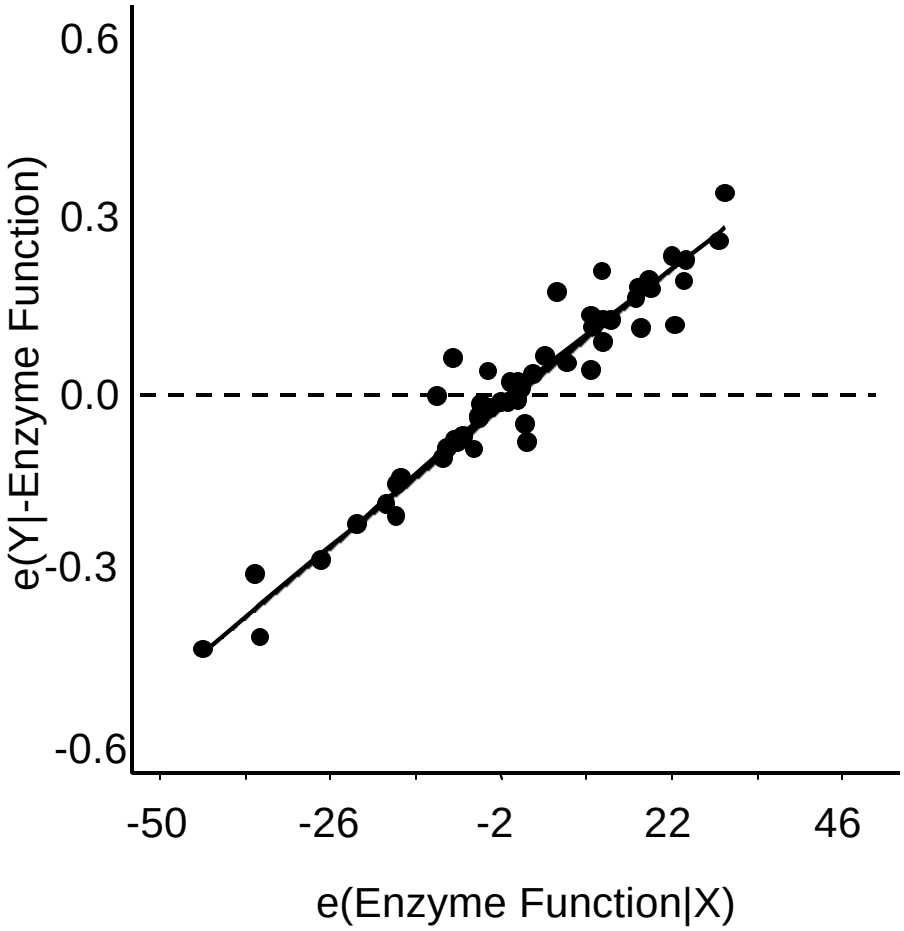


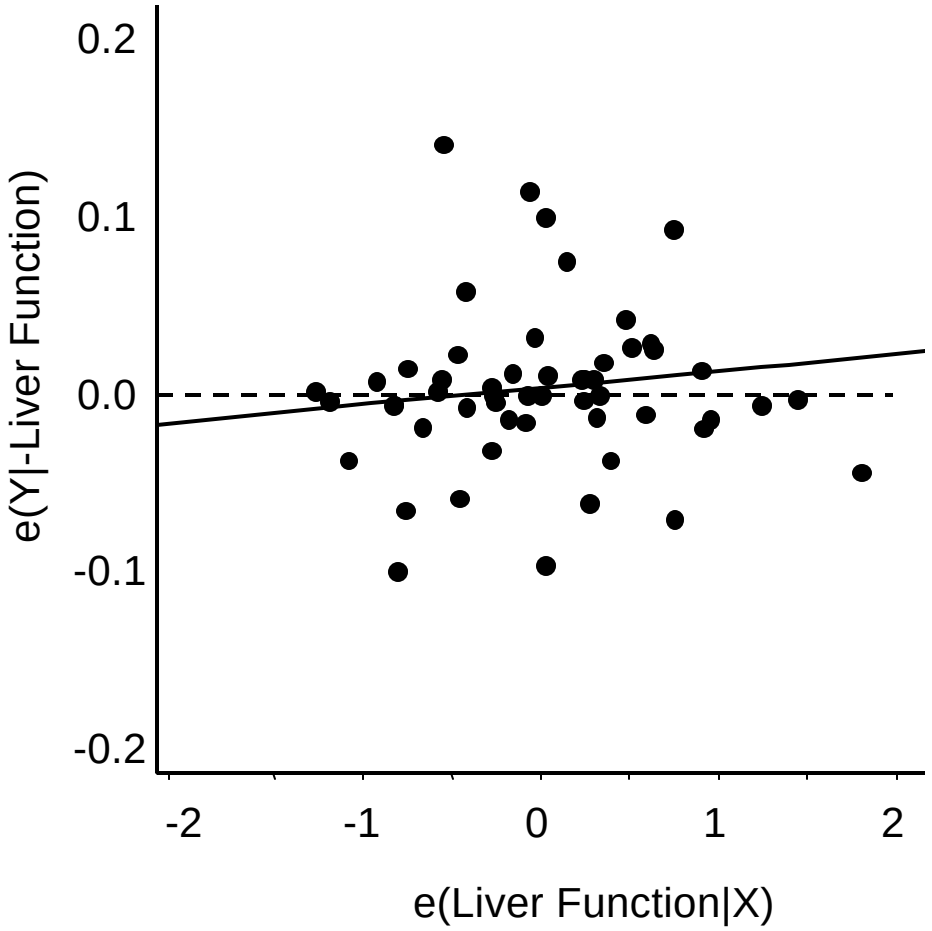


Partial Regression Plots
(Functional Form)

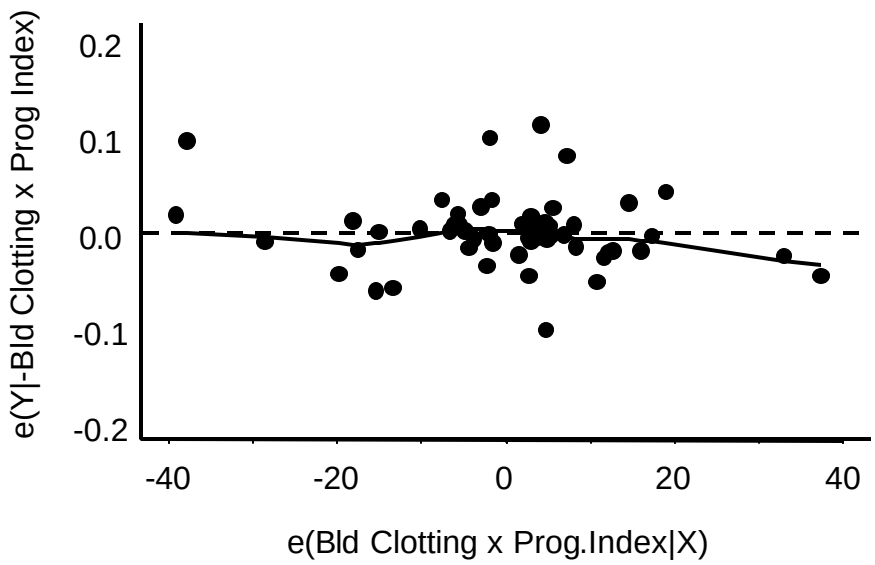


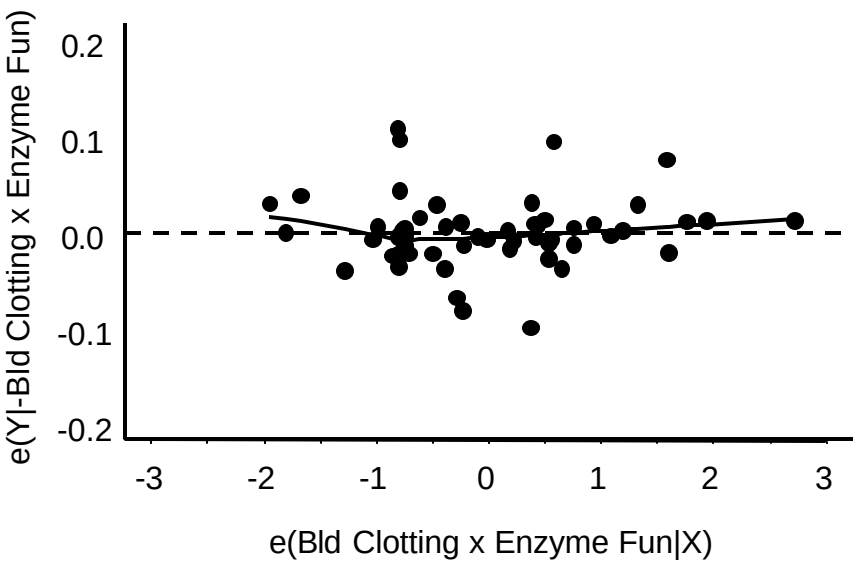
Partial Regression Plots
(Functional Form)

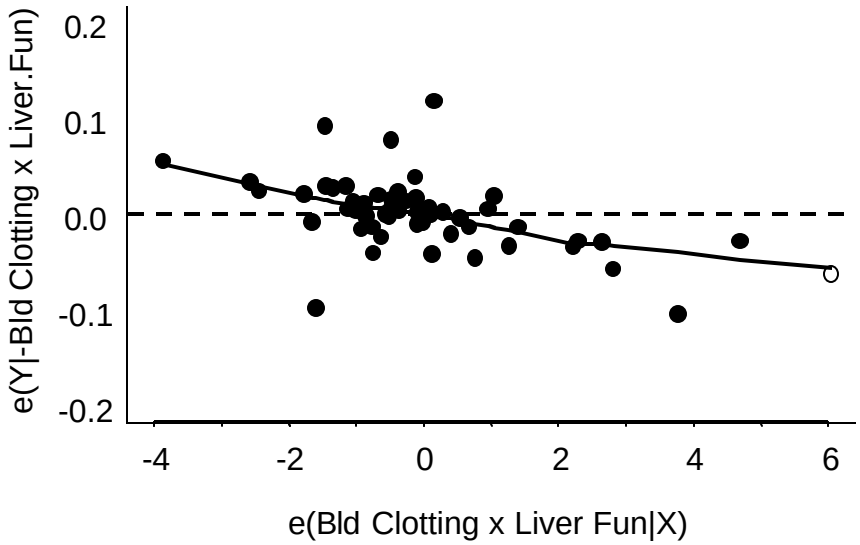


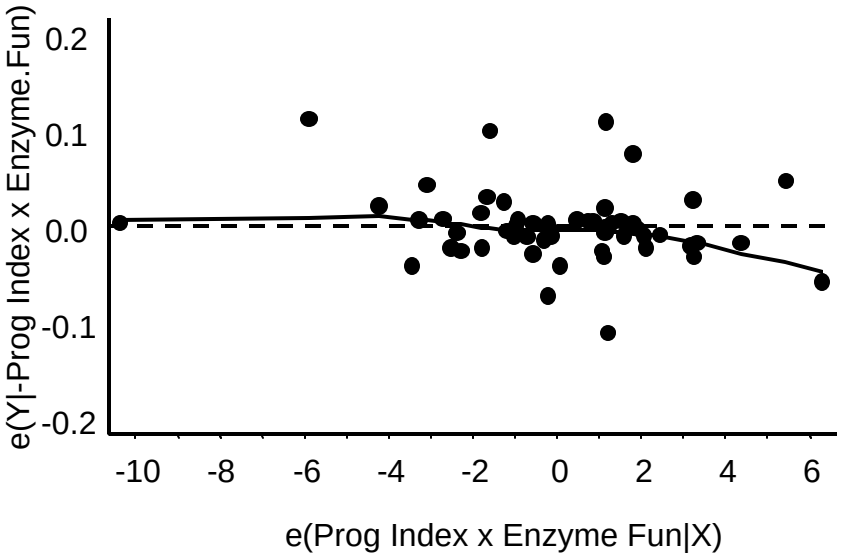


Partial Regression Plots (Additivity)

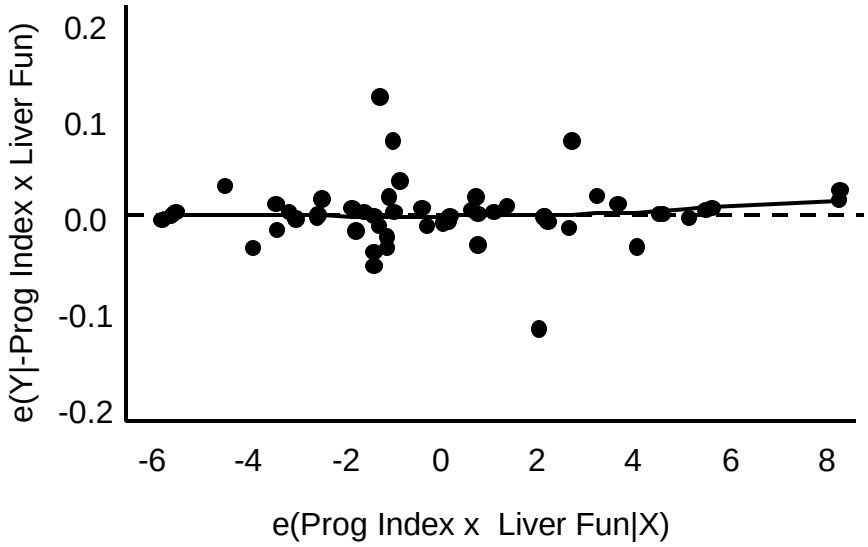


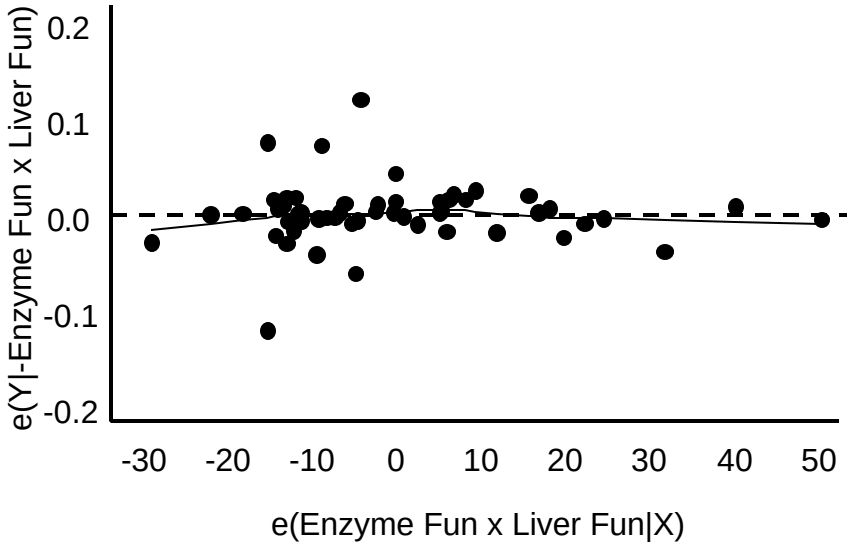






Partial Regression Plots (Additivity)





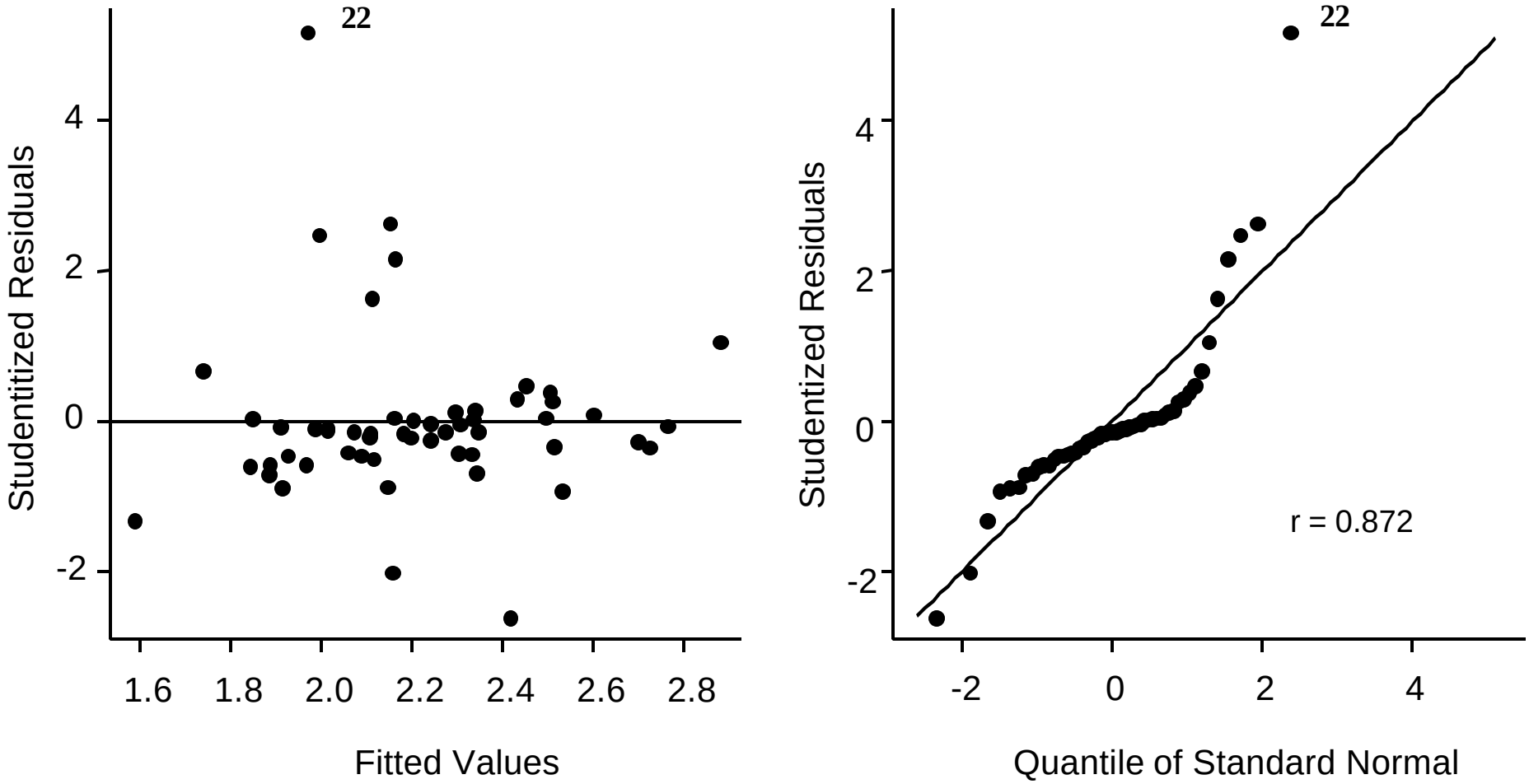
Model Statement

$E(y|X) = \beta_o + \beta_1(\text{Bld Clotting}) + \beta_2(\text{Prog. Index}) + \beta_3(\text{Enzyme Fun.}) + \beta_4(\text{Liver Fun}) + \beta_5(\text{Bld. Clotting x Liver Fun.})$.

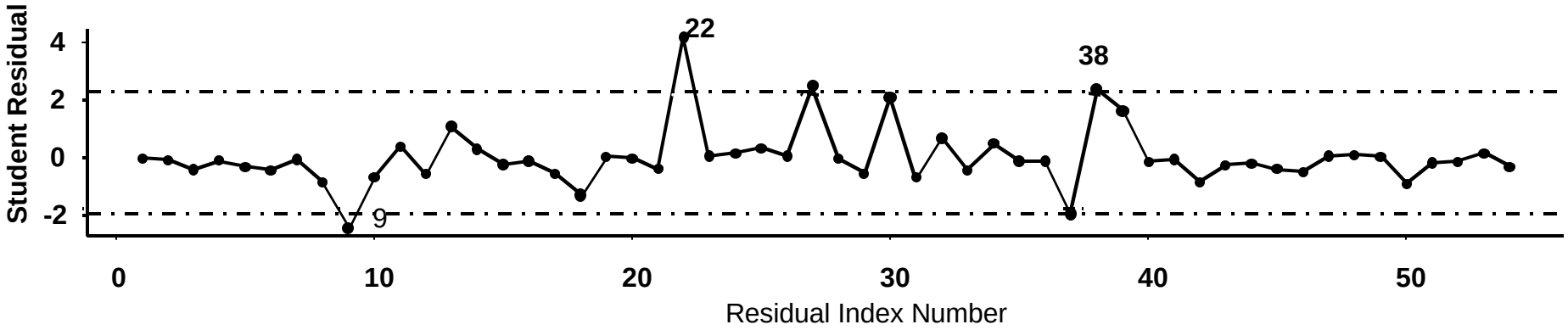
Term	b	SE(b)	t	P(T≥ t)
Intercept	0.3671	0.0682	5.386	0.0028
Bld. Clotting	0.0876	0.0092	9.511	<0.0001
Prog. Index	0.0093	0.0004	22.419	<0.0001
Enzyme Fun.	0.0095	0.0004	25.262	<0.0001
Liver Fun	0.0389	0.0175	2.231	0.0304
Bld C. x Liver Fun	-0.0059	0.0024	-2.500	0.0159

$\hat{s}(y_i)$=MSE	df	R^2	R_a^2
0.0450	48	0.9756	0.9724

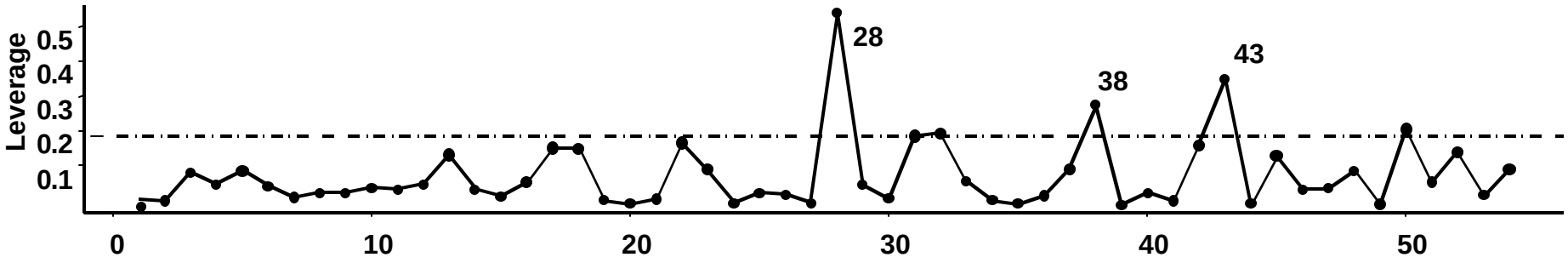
Residual Diagnostics



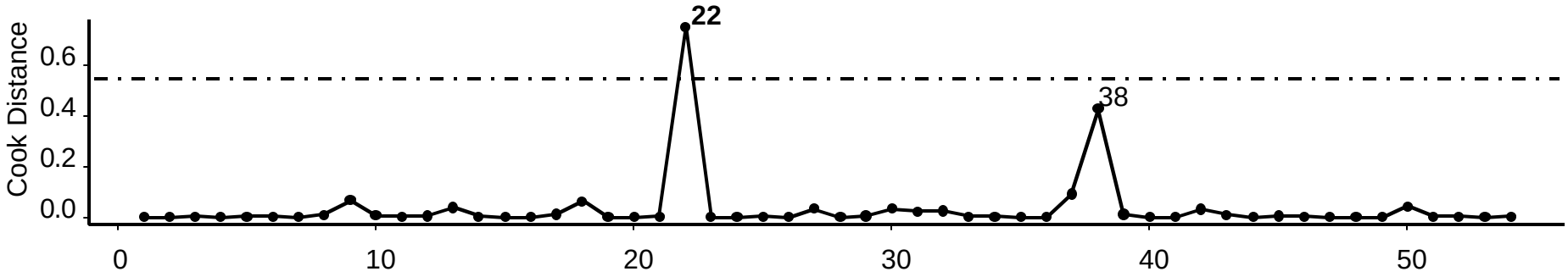
Student Residual



Residual Leverage



Cooks Distance



Summary of Influential Observations

Index	Bld.C	Prog.Ind	Enyz.Fun	Liv.Fun	Surv	log.Surv
9	6.00	67.00	93.00	2.50	202.00	2.31
22	3.40	83.00	53.00	1.12	136.00	2.13
28	11.20	76.00	90.00	5.59	574.00	2.76
38	4.30	8.00	119.00	2.85	120.00	2.08
43	8.80	86.00	88.00	6.40	483.00	2.68
Mean	5.78	63.24	77.11	2.74	197.16	19.77
SD	0.21	2.30	2.90	0.14	19.77	0.04

Model Statement (excluding obs. 22, 38, and 43)

$E(y|X) = \beta_o + \beta_1(\text{Bld Clotting}) + \beta_2(\text{Prog. Index}) + \beta_3(\text{Enzyme Fun.}) + \beta_4(\text{Liver Fun}) + \beta_5(\text{Bld. Clotting x Liver Fun.})$.

Term	b	SE(b)	t	P(T≥ t)
Intercept	0.2768	0.0574	4.822	<0.0001
Bld. Clotting	0.0996	0.0081	12.343	<0.0001
Prog. Index	0.0094	0.0004	24.684	<0.0001
Enzyme Fun.	0.0096	0.0003	31.571	<0.0001
Liver Fun	0.0566	0.0145	3.913	0.0003
Bld C. x Liver Fun	-0.0084	0.0021	-4.004	0.0002

$\hat{s}(y_i)$=MSE	df	R^2	R_a^2
0.0432	44	0.9853	0.9832

Bootstrap Validation

Original Index

Parameter estimate (P_{org}) obtained from the original fit of the OLS model.

$$E(Y_{org}) = X_{org}b$$

Train Index

For bootstrap random samples i=1…b
Parameter estimate ($P_{training,i}$) is obtained from an OLS model fit, in which $X_{training,i} = X_{boot,i.}$ and $Y_{training,i} = Y_{boot,i}$

$$E(Y_{training,i}) = X_{training,i}\ b_{training}$$

Test Index

Parameter estimate ($P_{test,i}$) obtained from a model, in which $X_{test.i} = X_{org}$ and $Y_{test,i} = Y_{org,}$

$$E(Y_{test,i}) = X_{test,i}\ b_{training,i}$$

Optimism

The optimism for the *i*th bootstrap sample is estimated by:

$$O_i = P_{training,i} - P_{test,i}$$

Index (Bias) Corrected

The bootstrap corrected estimate is computed as :

$$P_{corrected} = P_{org} - \frac{1}{B}\sum_{i}^{B} O_i$$

Full Model

$$E(y|X) = \beta_o + \beta_1(\text{Bld Clotting}) + \beta_2(\text{Prog. Index}) + \beta_3(\text{Enzyme Fun.}) + \beta_4(\text{Liver Fun}) + \beta_5(\text{Bld. Clotting x Liver Fun.}).$$

Bootstrap Model Validation

Parameter	Index original	training	test	Optimism	Index corrected	N Bootstraps
R-square	0.9756	0.9772	0.9712	0.0060	0.9695	500
MSE	0.0018	0.0016	0.0021	-0.0005	0.0023	500
Intercept	0.0000	0.0000	0.0093	-0.0093	0.0093	500
Slope	1.0000	1.0000	0.9959	0.0041	0.9959	500

Reduced Model

$E(y|X) = \beta_o + \beta_1(\text{Bld Clotting}) + \beta_2(\text{Prog. Index}) + \beta_3(\text{Enzyme Fun.})$

Bootstrap Model Validation

Parameter	Index original	training	test	Optimism	Index corrected	N Bootstraps
R-square	0.9723	0.9773	0.9691	0.0048	0.9678	500
MSE	0.0020	0.0018	0.0022	-0.0005	0.0025	500
Intercept	0.0000	0.0000	0.0070	-0.0070	0.0070	500
Slope	1.0000	1.0000	0.9966	0.0034	0.9966	500

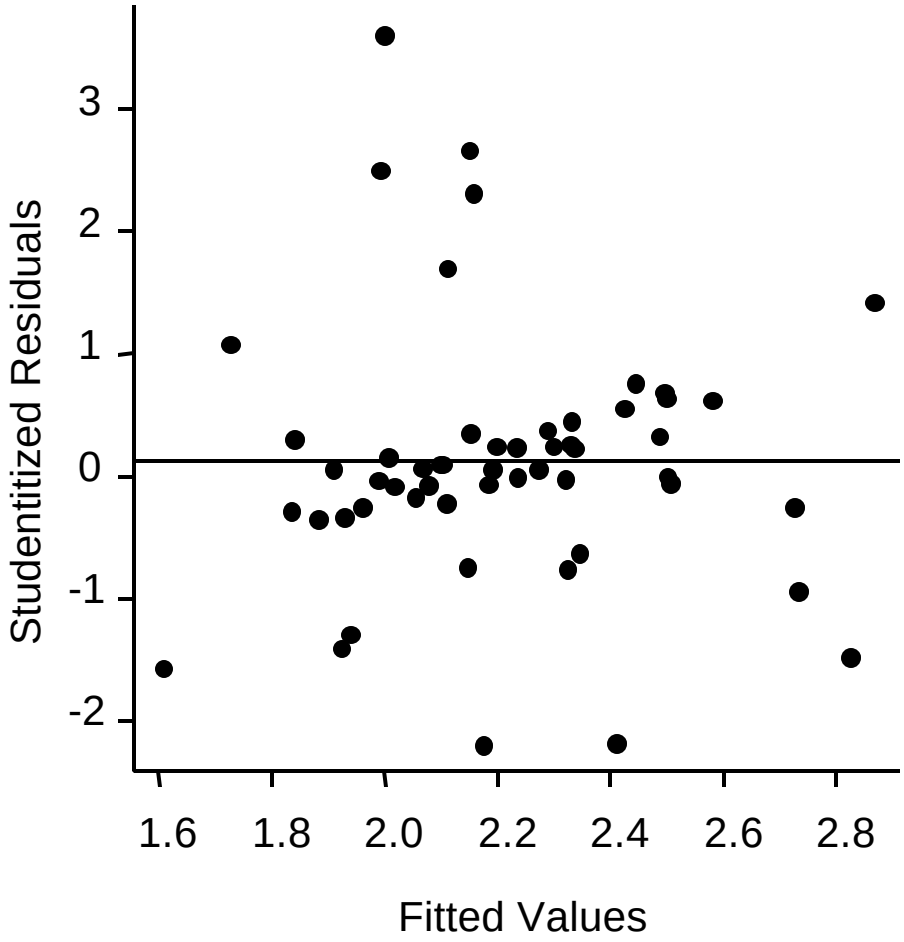
Model Statement

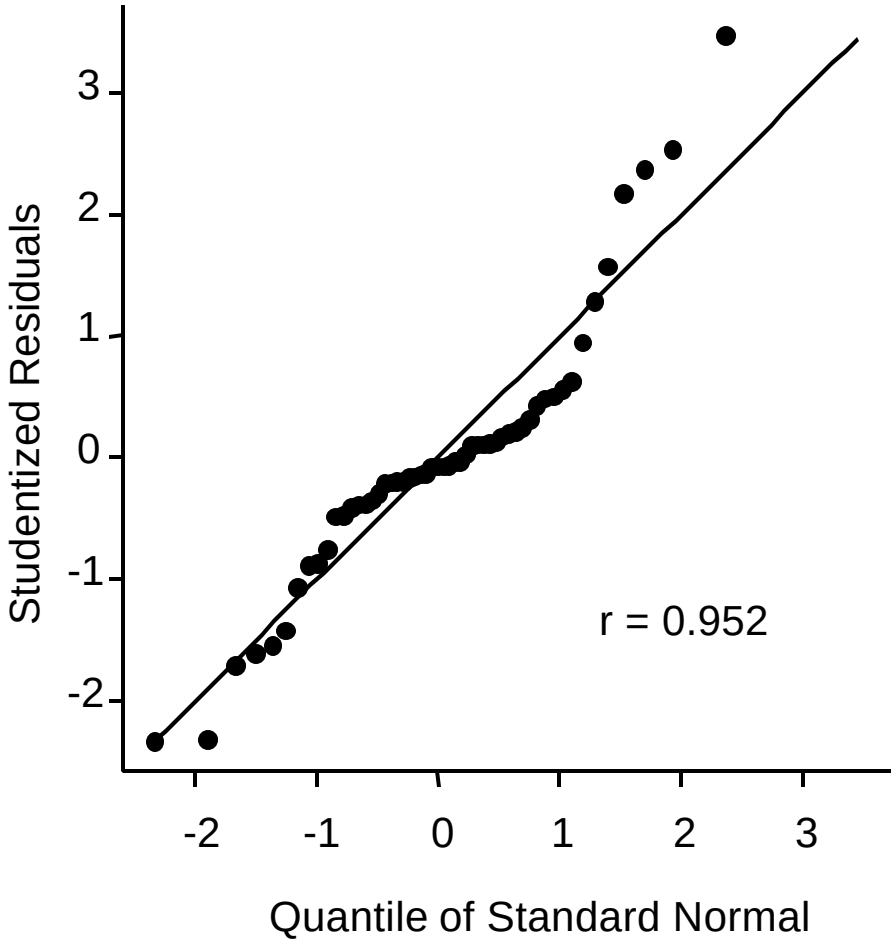
$E(y|X) = \beta_o + \beta_1(\text{Bld Clotting}) + \beta_2(\text{Prog. Index}) + \beta_3(\text{Enzyme Fun.})$

Term	b	SE(b)	t	P(T≥ t)
Intercept	0.4836	0.0426	11.345	<0.0001
Bld. Clotting	0.0699	0.0041	16.975	<0.0001
Prog. Index	0.0093	0.0003	24.300	<0.0001
Enzyme Fun.	0.0095	0.0003	31.082	<0.0001

$\hat{s}(y_i)$=MSE	df	R^2	R_a^2
0.0432	50	0.9723	0.9694

Residual Diagnostics





Interpretation

The model validation suggests that if we were to use the regression coefficients from the model which included terms for the patient's pre-operative blood clotting score, prognostic index, and enzyme function score on a new sample of patients, approximately 96.8% of the variation in postoperative survival time would be explained by this model.

If we were to use the regression coefficients from the model that also included the patient's pre-operative liver function score and a term for blood clotting by liver function interaction on a new sample of patient, we would expect that approximately 96.9% of the variation in postoperative survival time would be explained by this model.

References

Johnson RA, Wichern DW**,** *Applied Multivariate Statistical Analysis.* (1999) Prentice Hall, Upper Saddle River, NJ.

Neter J, Kutner MH, Nachtsheim CJ, Wasserman W. *Applied Linear Statistical Models.* Fourth Edition. (1996) IRWIN, Chicago, IL.

Myers RH. *Classical and Modern Regression with Applications* Second Edition. (1990) Duxbury Press, Belmont, Cal.