Diagnostics and Remedial Measures

Adapted from Chapter 3 of the textbook
Applied Linear Regression Models, Edition: 4th
Authors: Michael H. Kutner, Christopher J. Nachtsheim and John Neter

Course: Math50 Dartmouth College, Fall 2015
Instructor: Nishant Malik

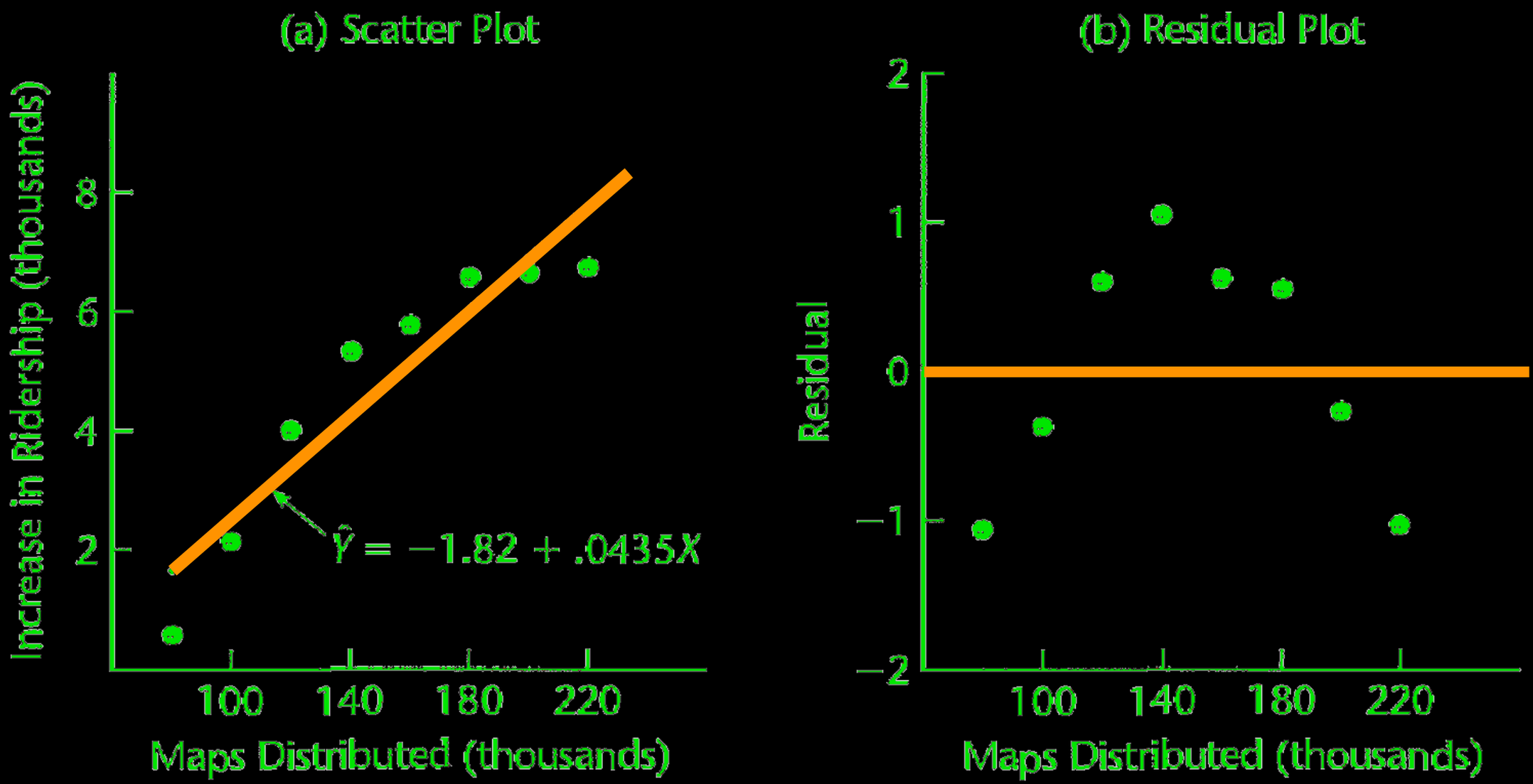
Nonlinearity of Regression Function

Example: Number of Maps Distributed and Increase in the Ridership - Public Transit

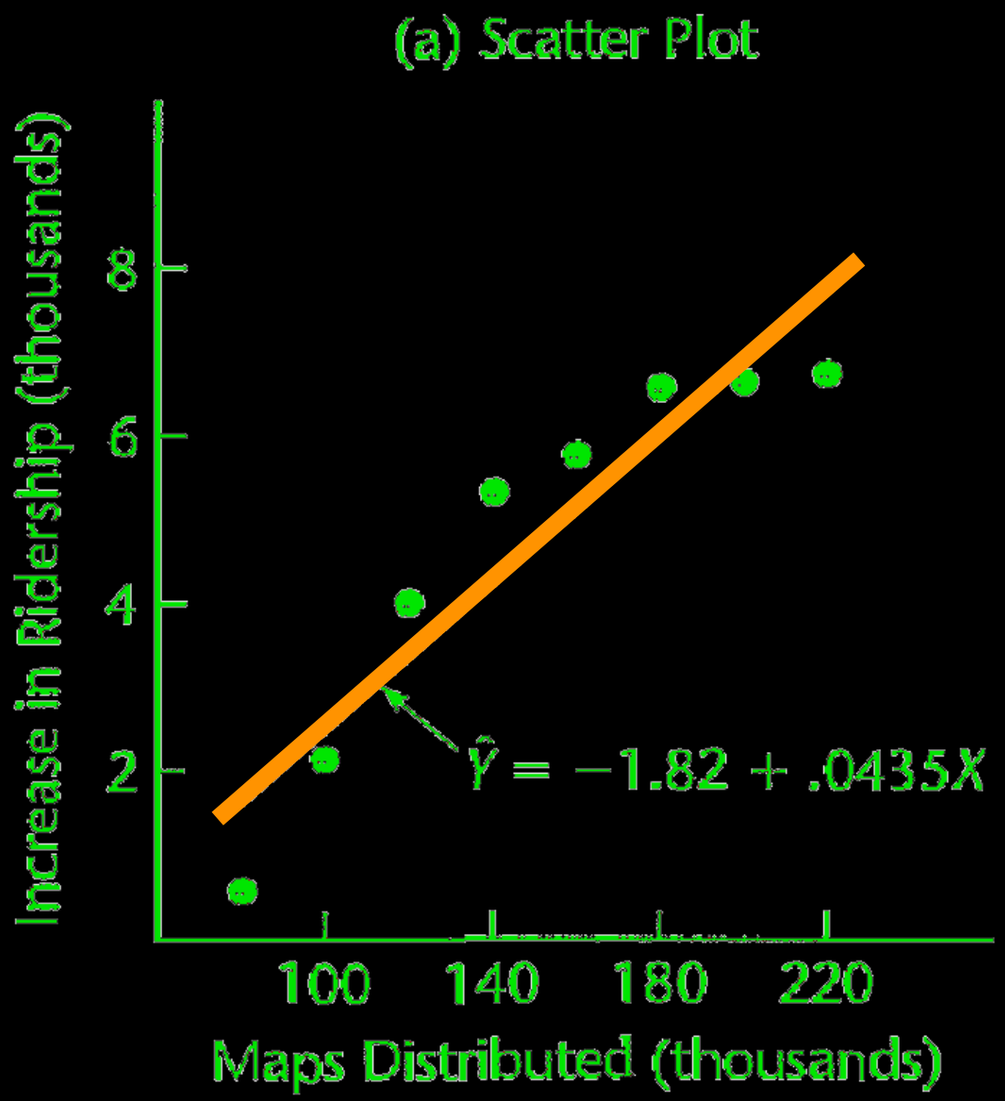
City i	(1) Increase in Ridership (thousands) Y_i	(2) Maps Distributed (thousands) X_i	(3) Fitted Value $\hat{Y}_i$	(4) Residual $Y_i - \hat{Y}_i = e_i$
1	.60	80	1.66	−1.06
2	6.70	220	7.75	−1.05
3	5.30	140	4.27	1.03
4	4.00	120	3.40	.60
5	6.55	180	6.01	.54
6	2.15	100	2.53	−.38
7	6.60	200	6.88	−.28
8	5.75	160	5.14	.61

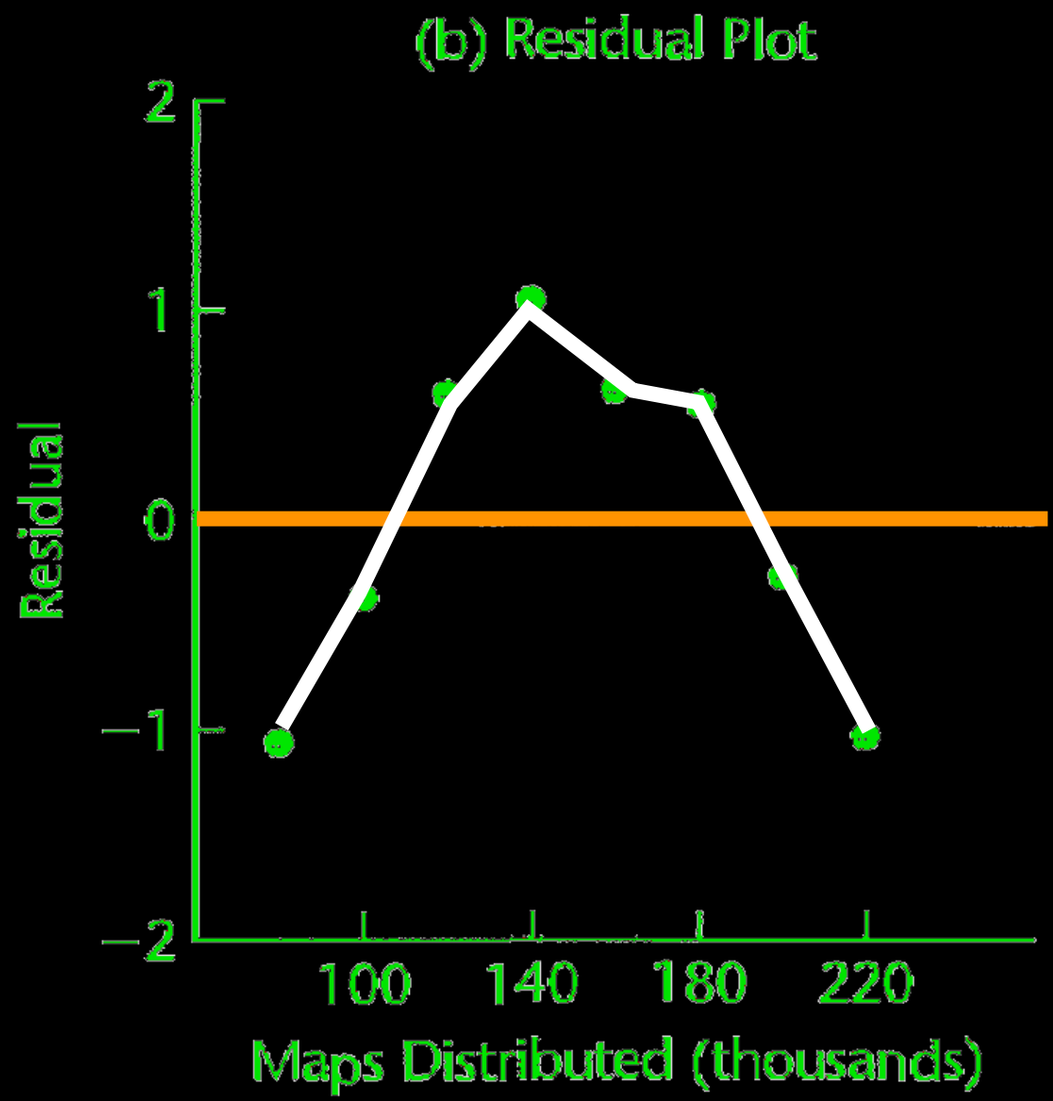
$$\hat{Y} = -1.82 + .0435X$$

Nonlinearity of Regression Function

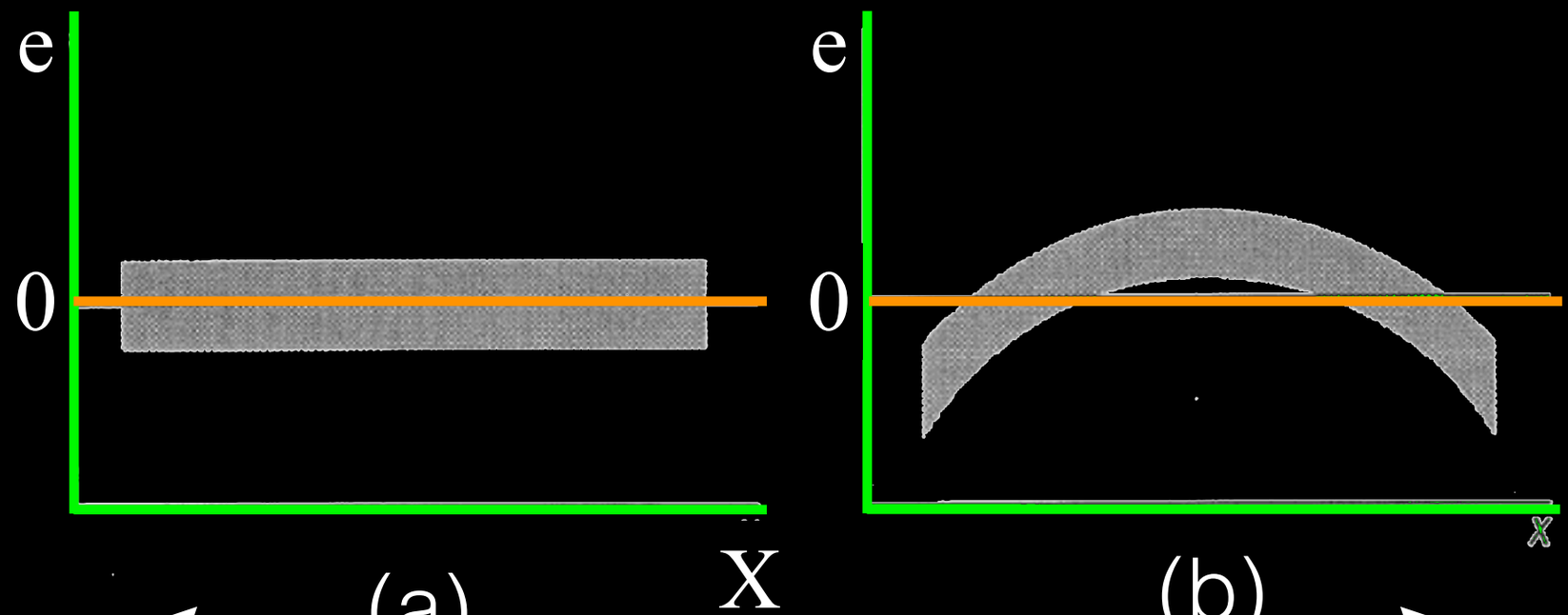


Nonlinearity of Regression Function





Prototype Residual Plots



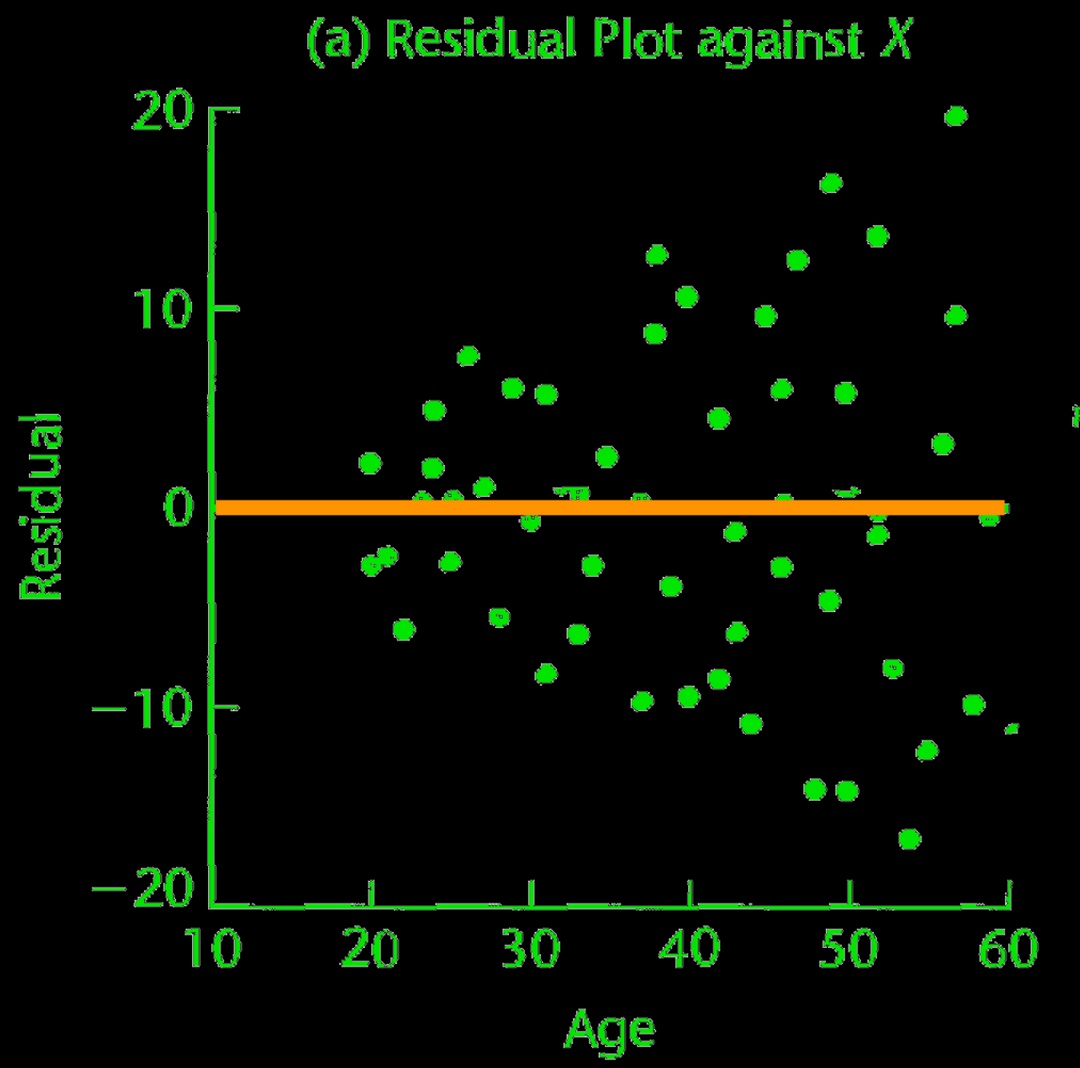
This is how it is supposed to be

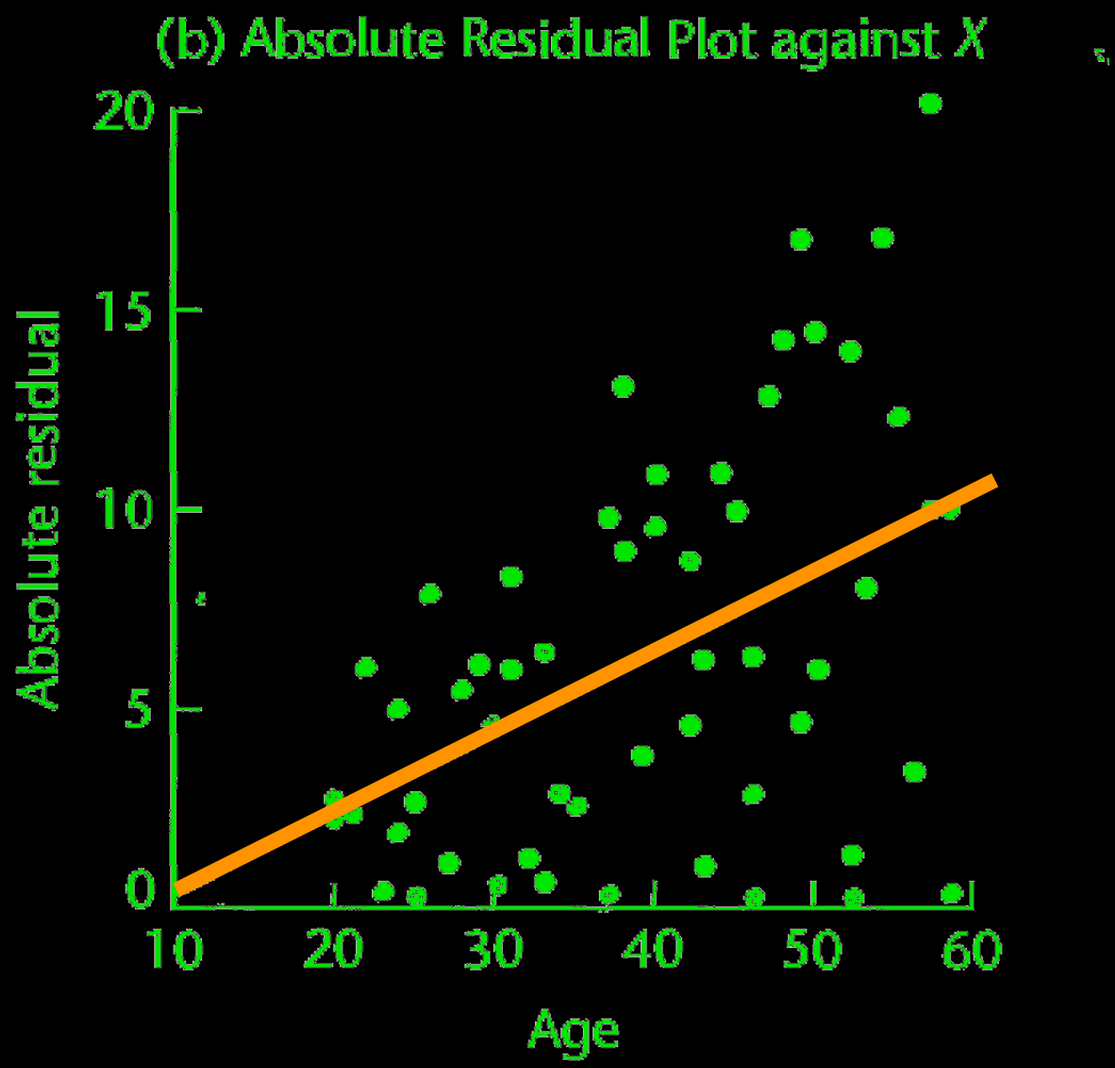
This is how it is looks like because of nonlinearity

Nonconstancy of Error Variance

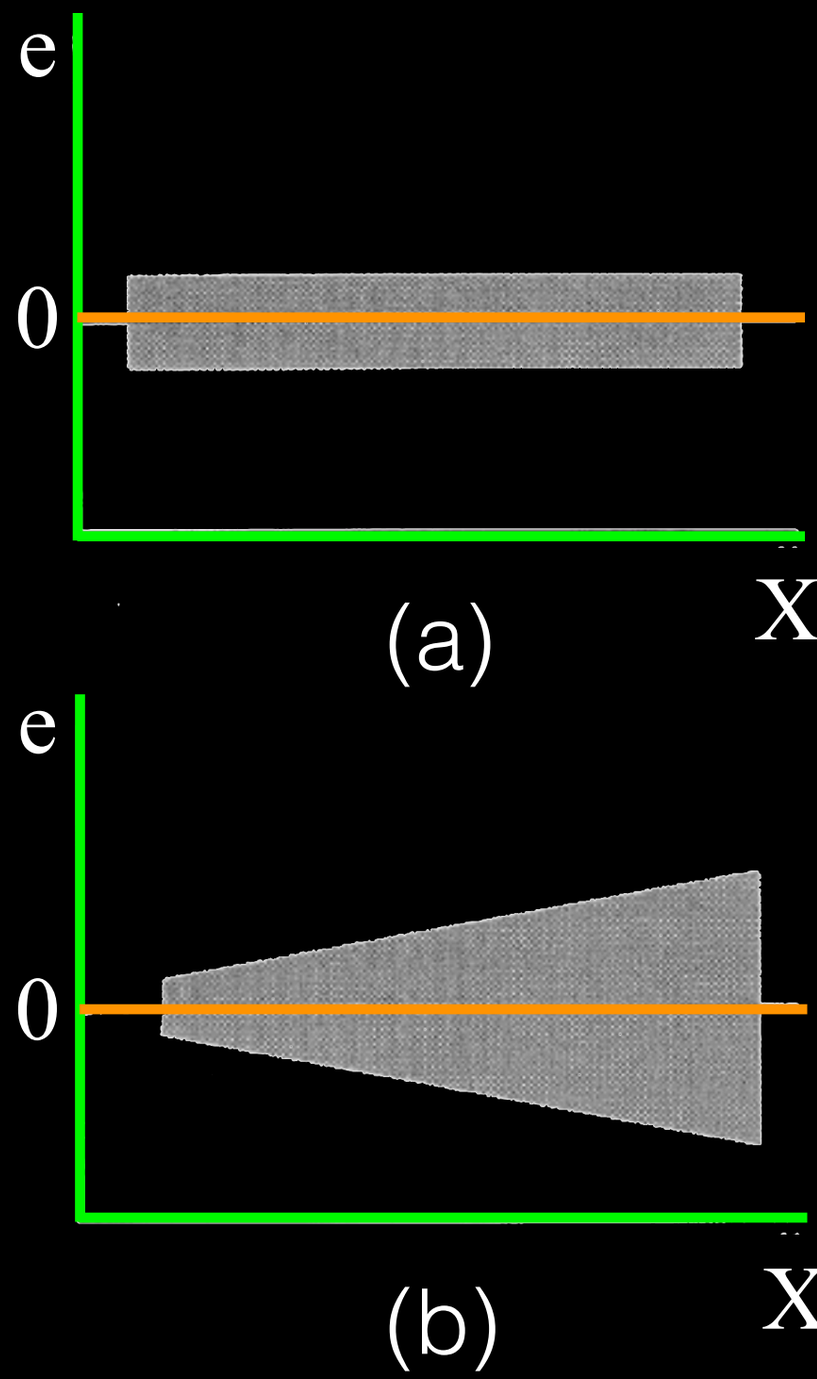
Example: Diastolic Blood pressure of healthy adult woman and age

Nonconstancy of Error Variance





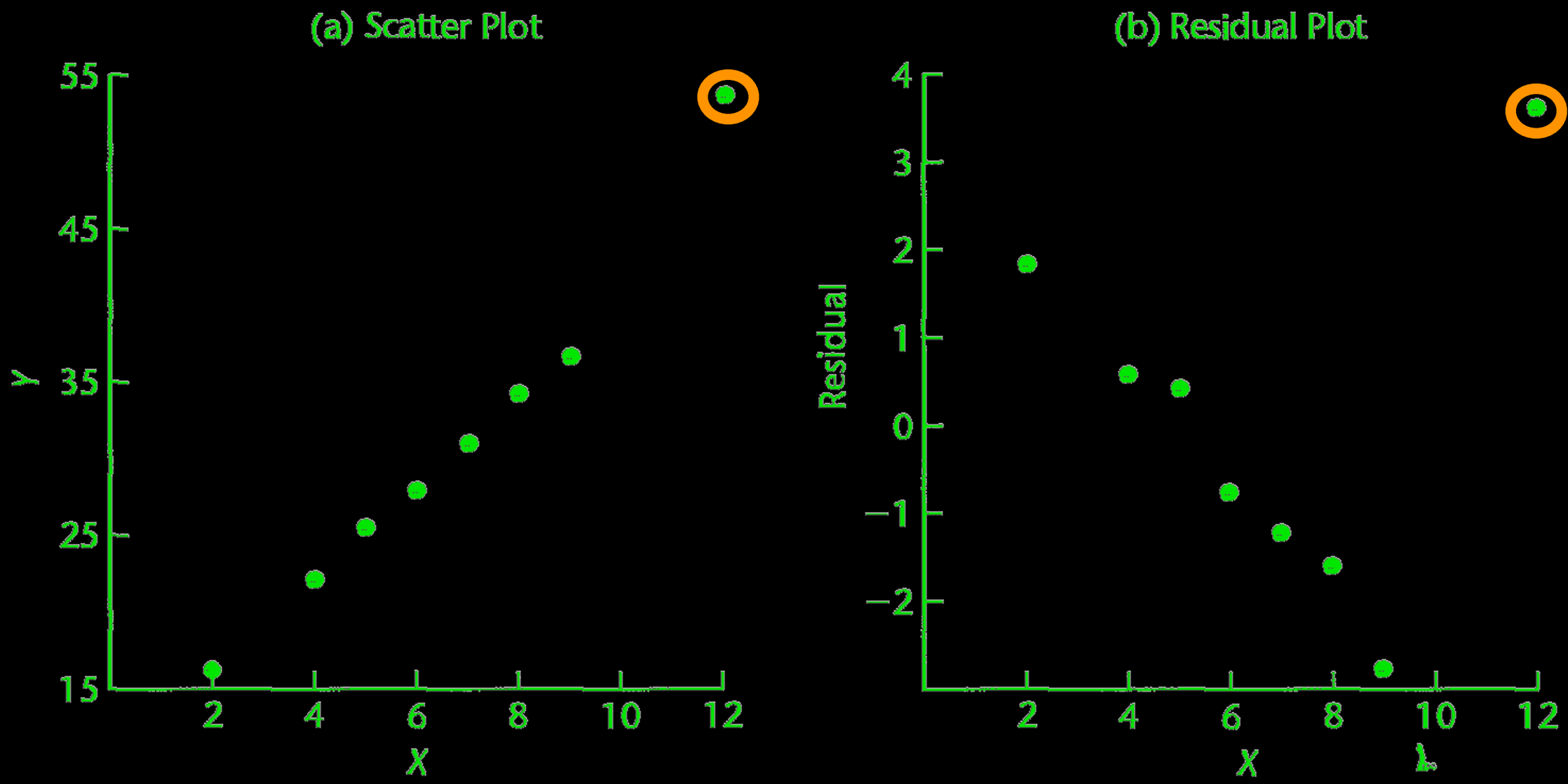
Prototype Residual Plots



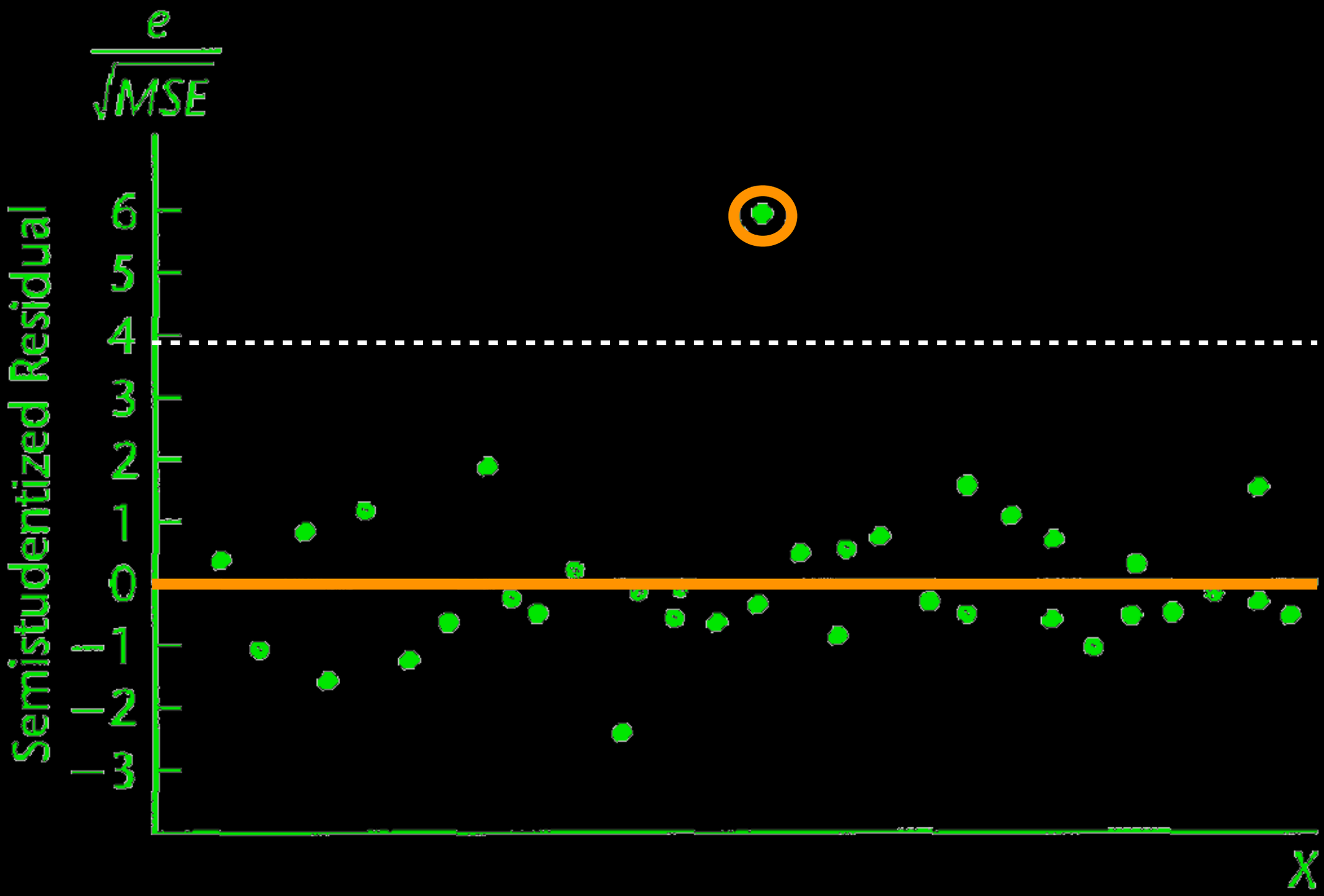
This is how it is supposed to be

This is how it is looks likes because of non-constant error variance

Presence of Outliers



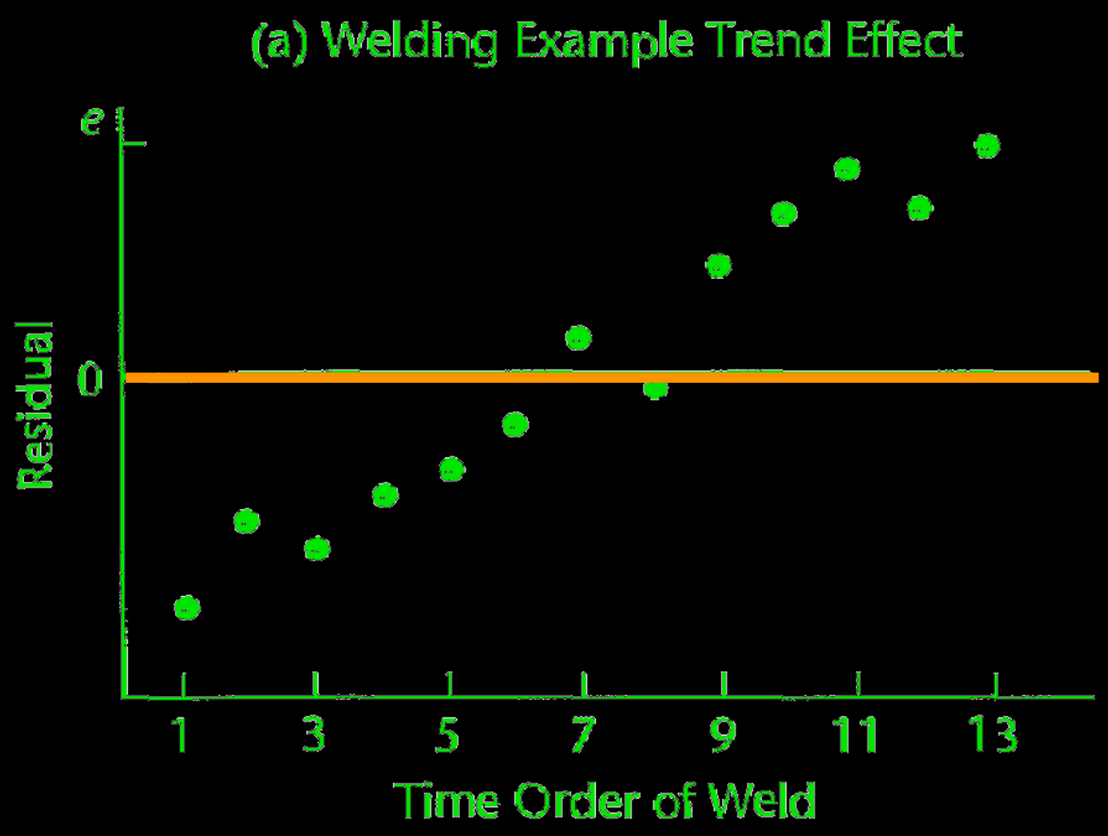
Presence of Outliers



Nonindependence of Error Terms

Example: Diameter of weld and the shear strength of the weld

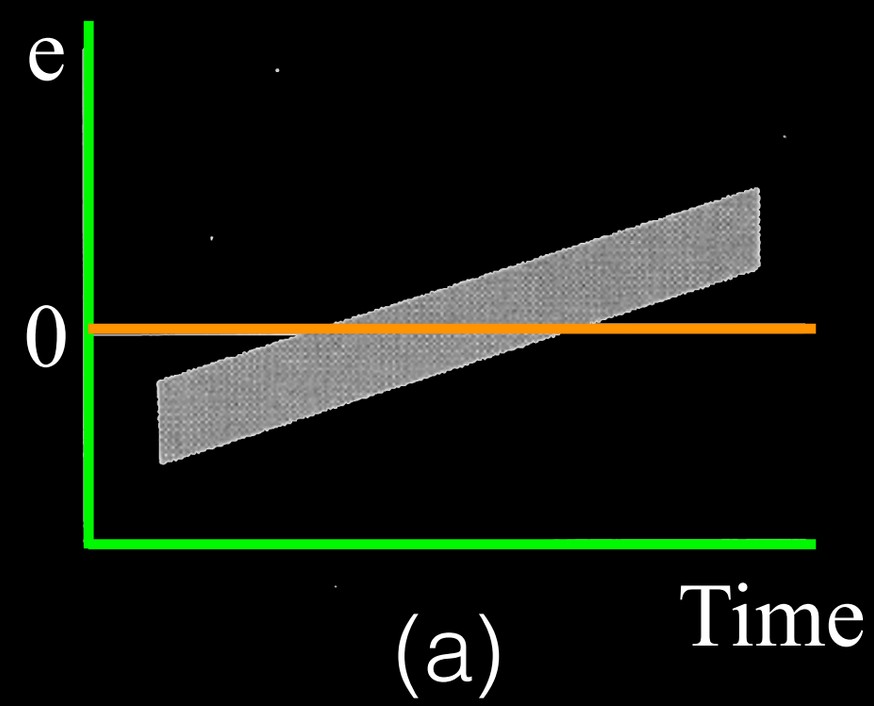
Nonindependence of Error Terms



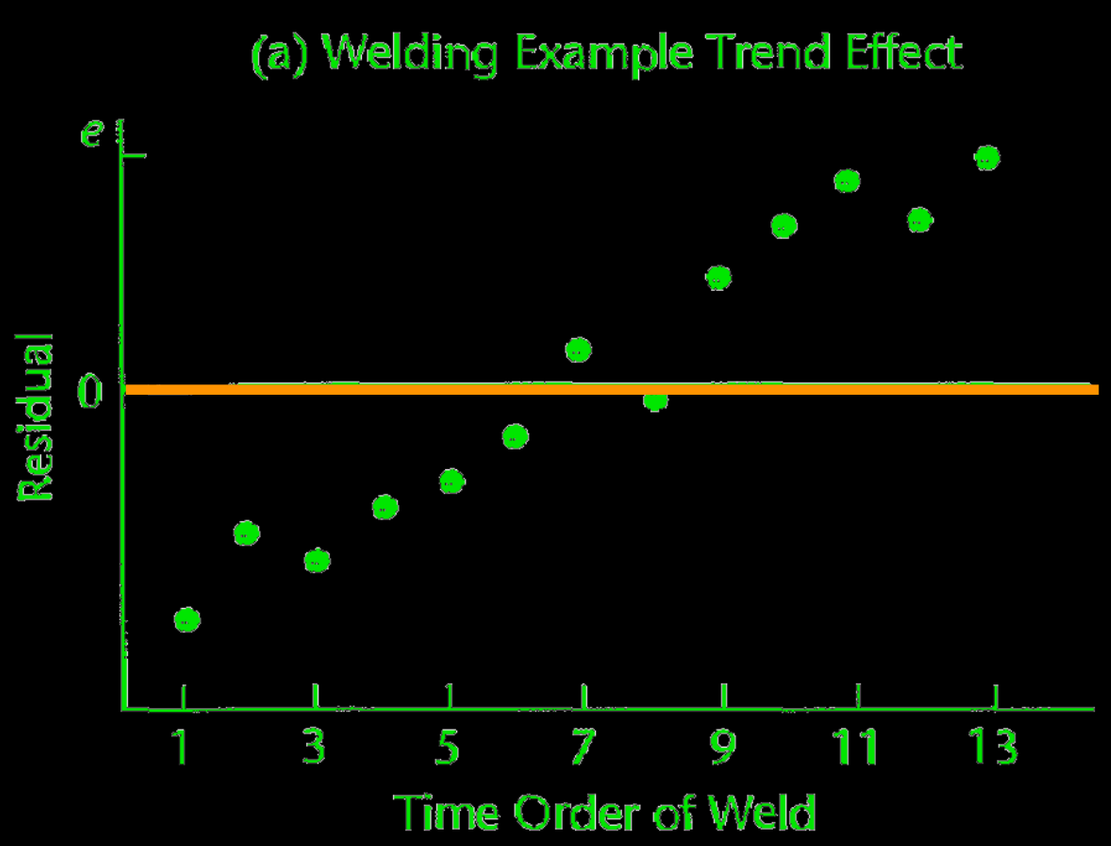
Prototype Residual Plots

Trend in residuals
i.e., they are not independent

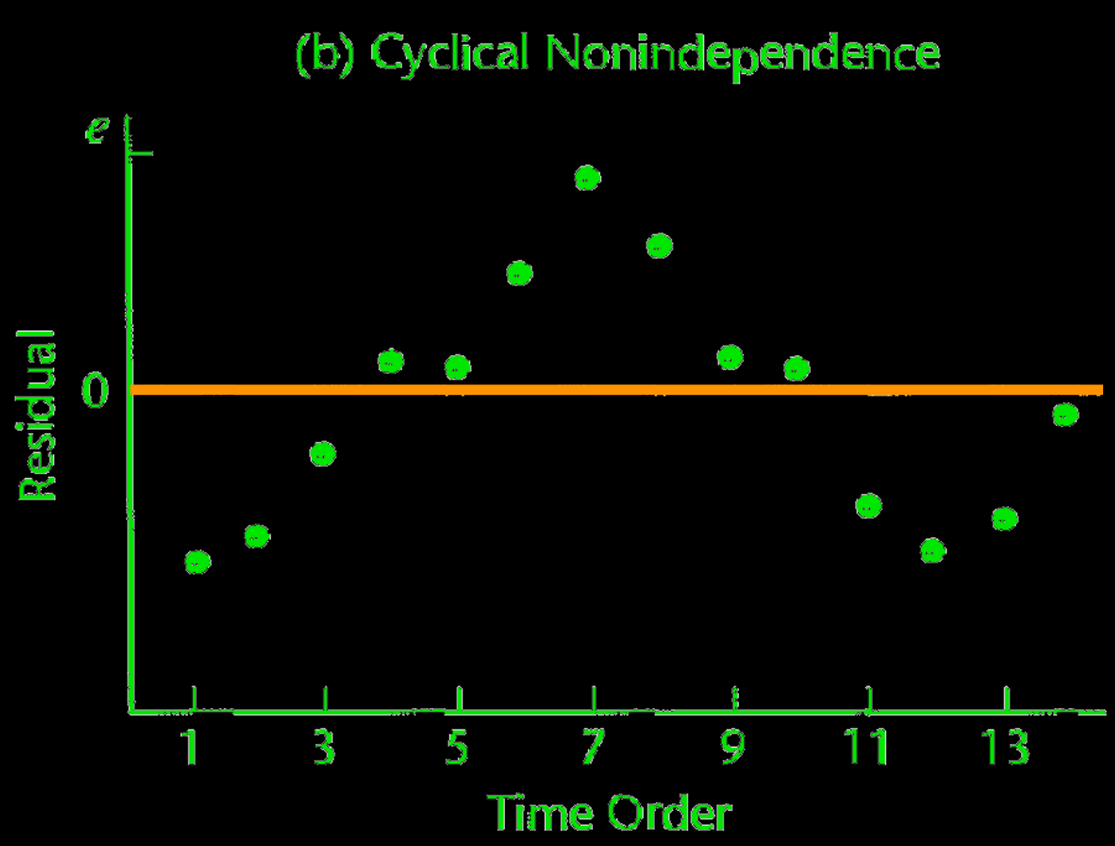
Not desired property



Nonindependence of Error Terms

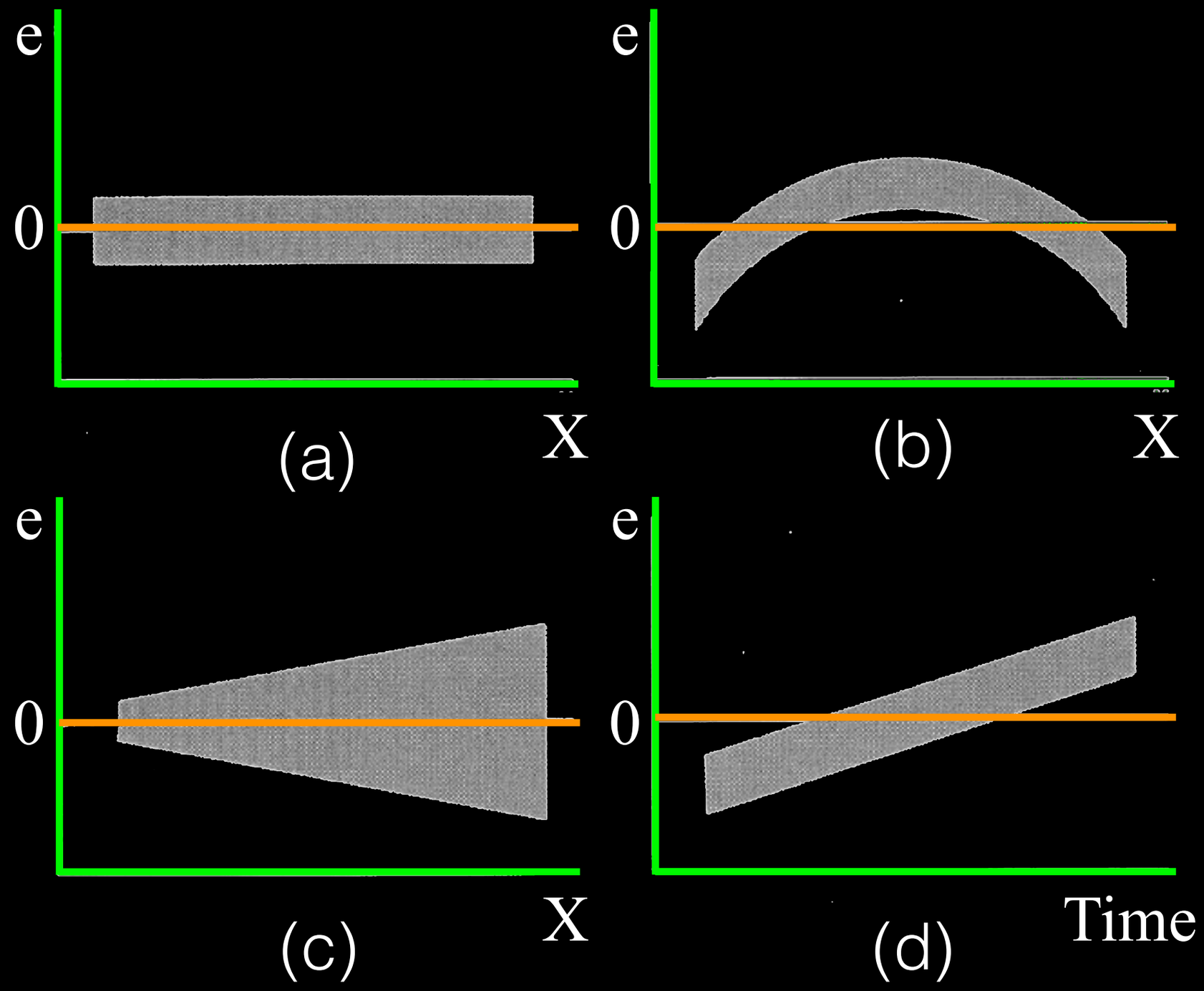


Trend

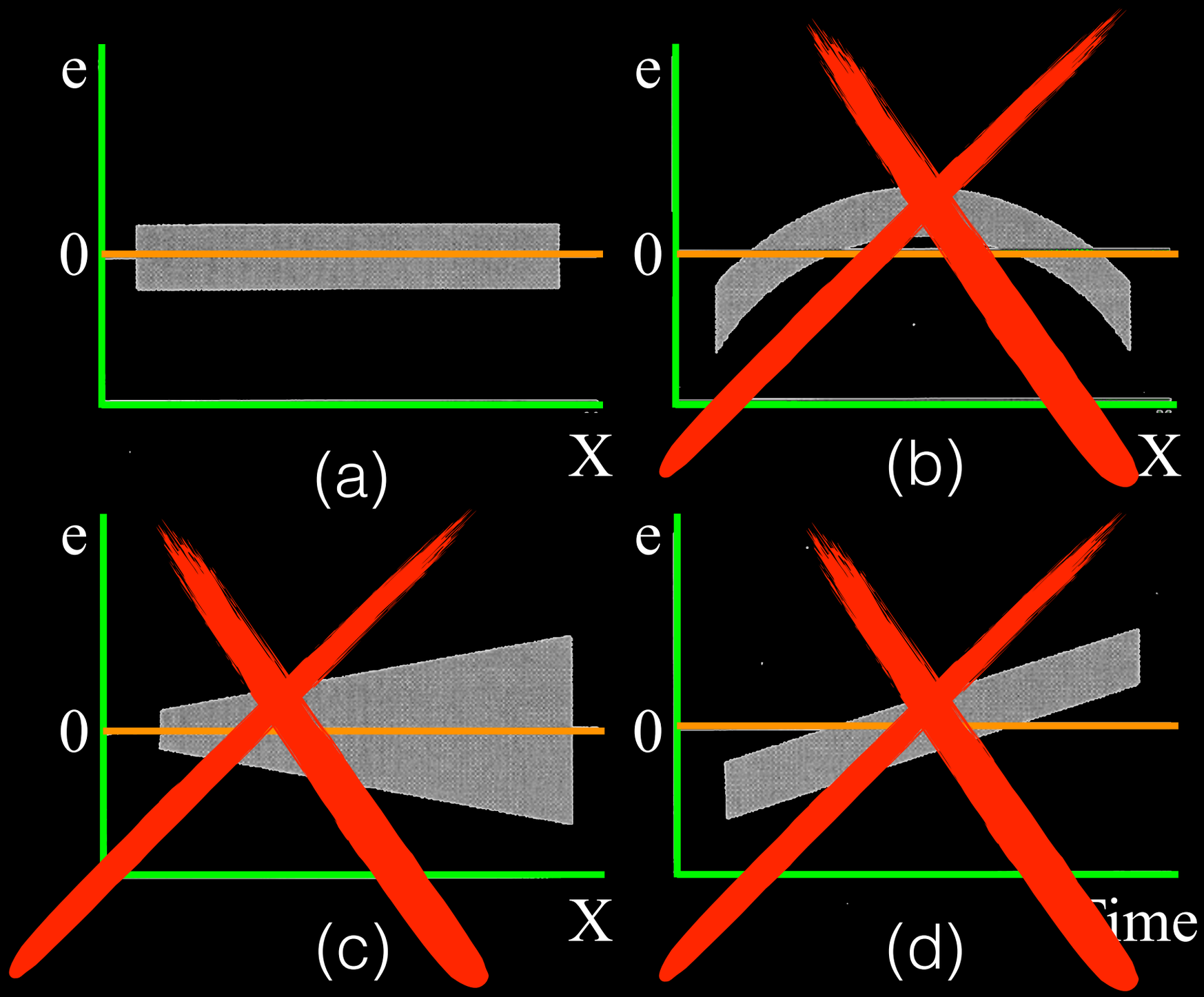


Cyclic

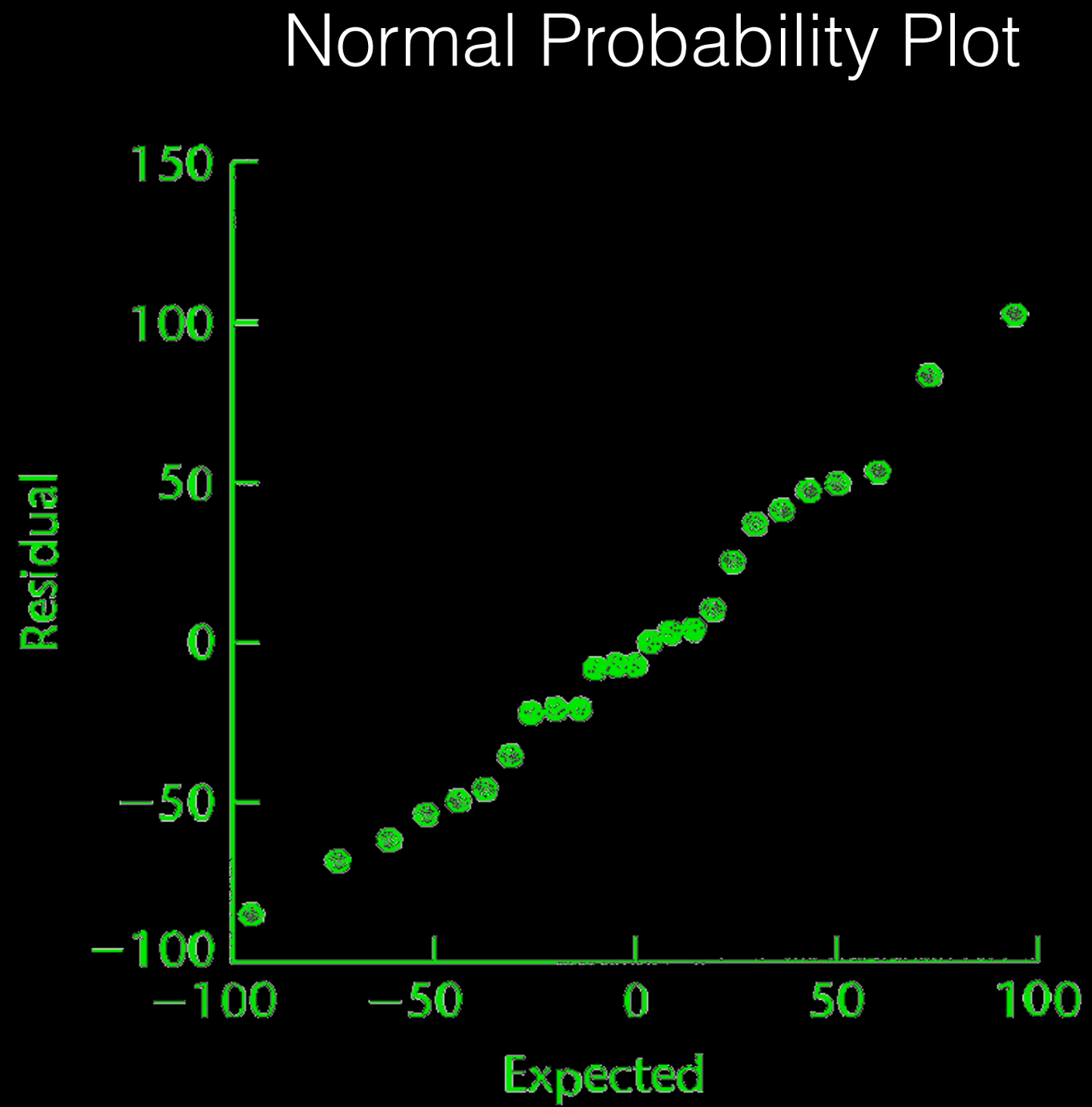
Prototype Residual Plots



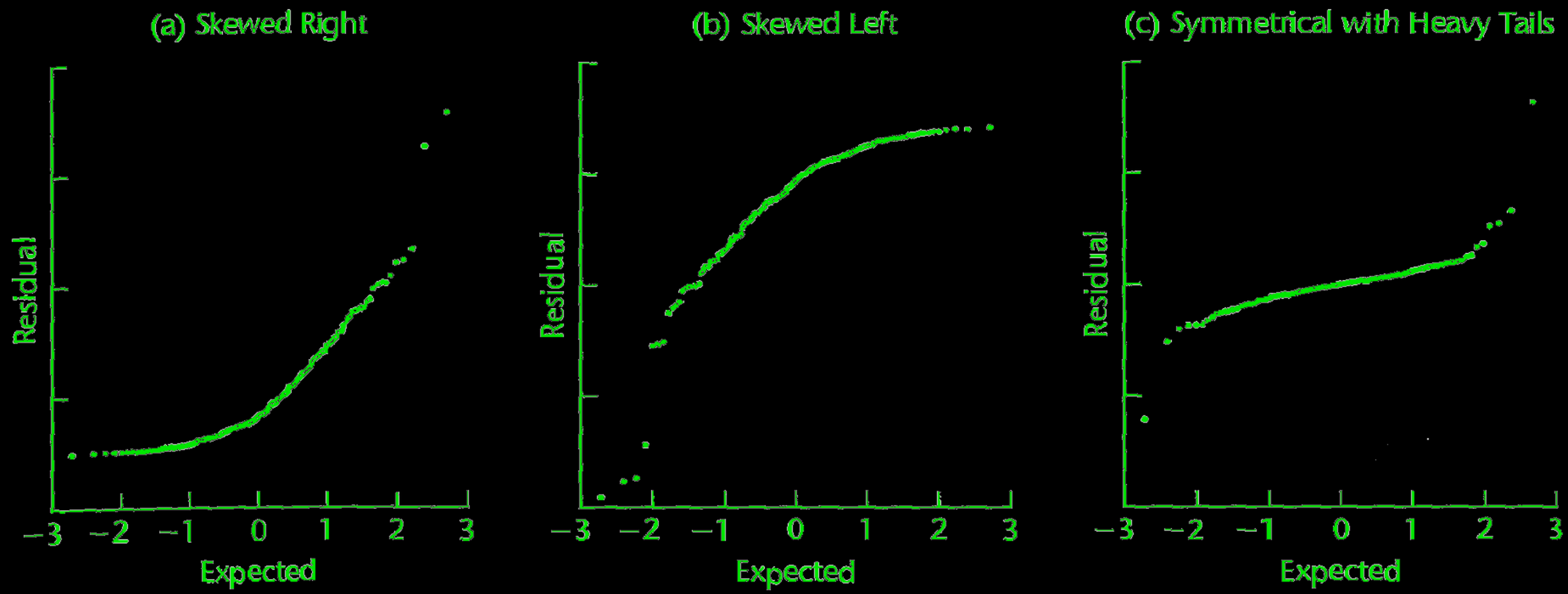
Prototype Residual Plots



Nonnormality of Error Terms



Nonnormality of Error Terms



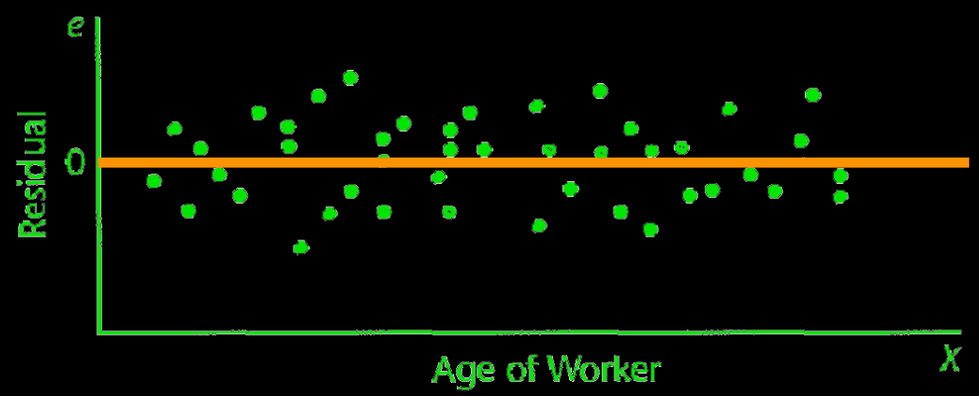
Undesirable Normal Probability Plot

Omission of Important Predictor variable

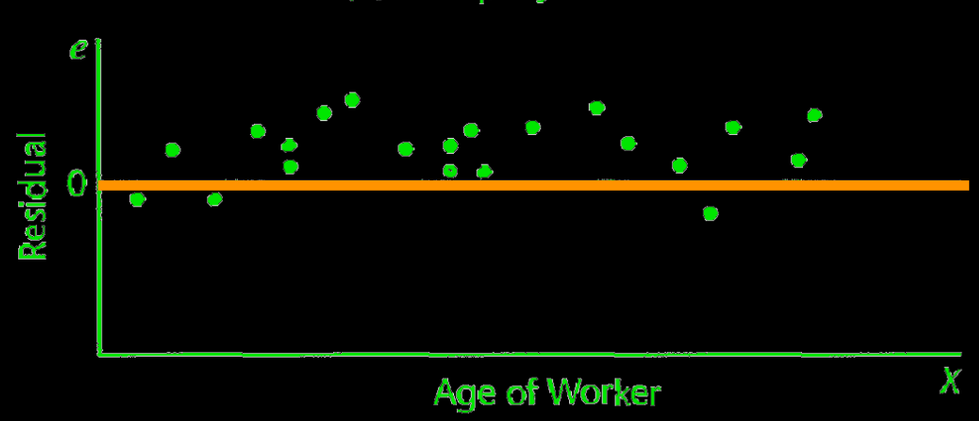
Example: Piece rate worker in an assembling operation, the relation between output and age of the worker

Omission of Important Predictor variable

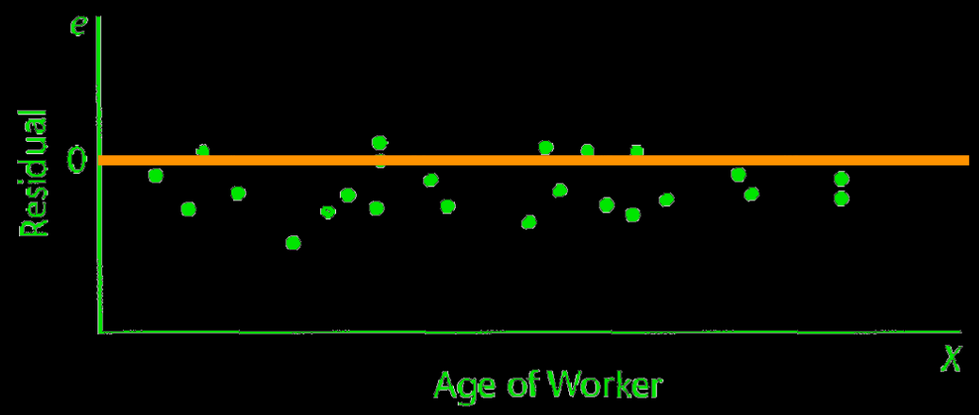
(a) Both Machines



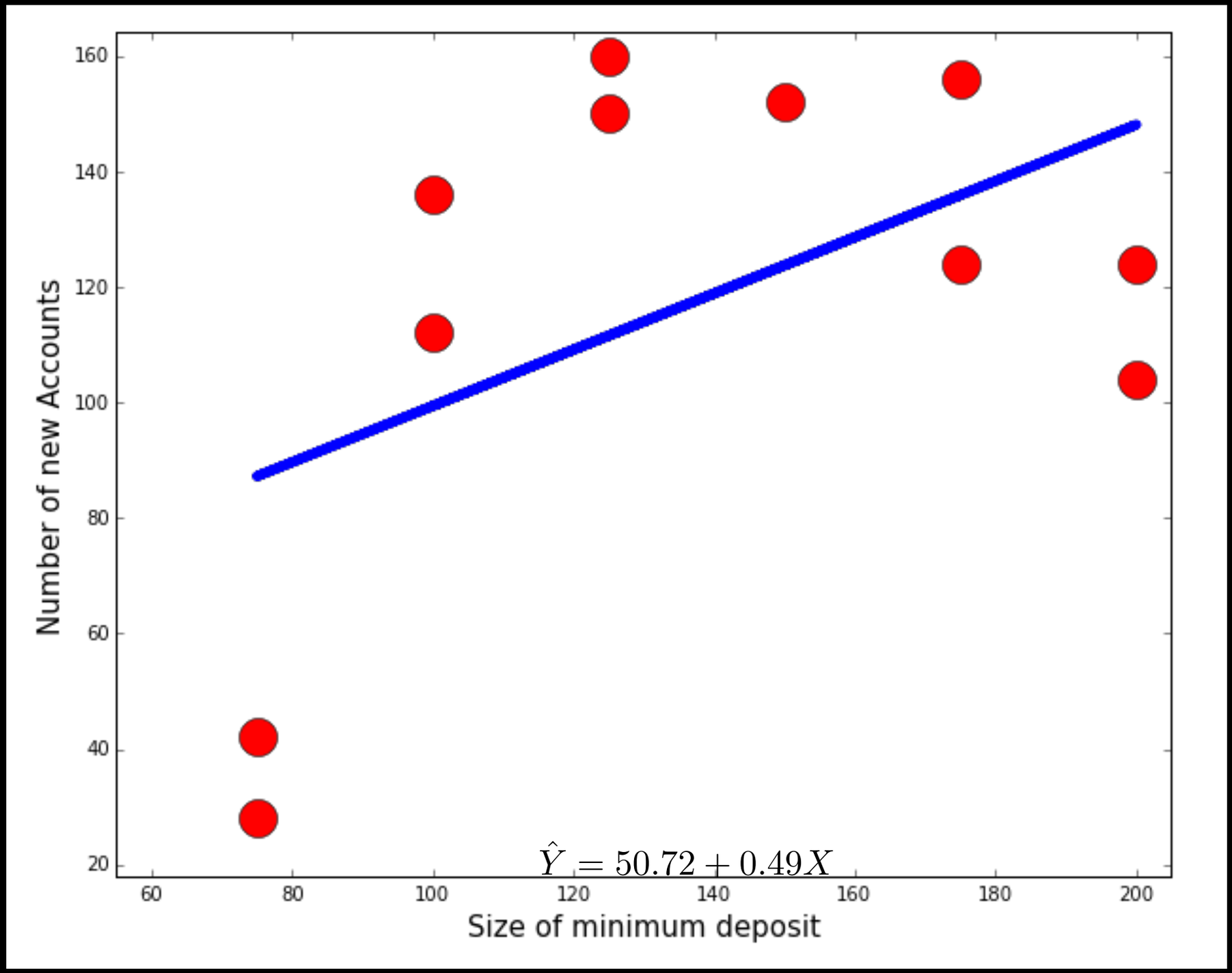
(b) Company A Machines



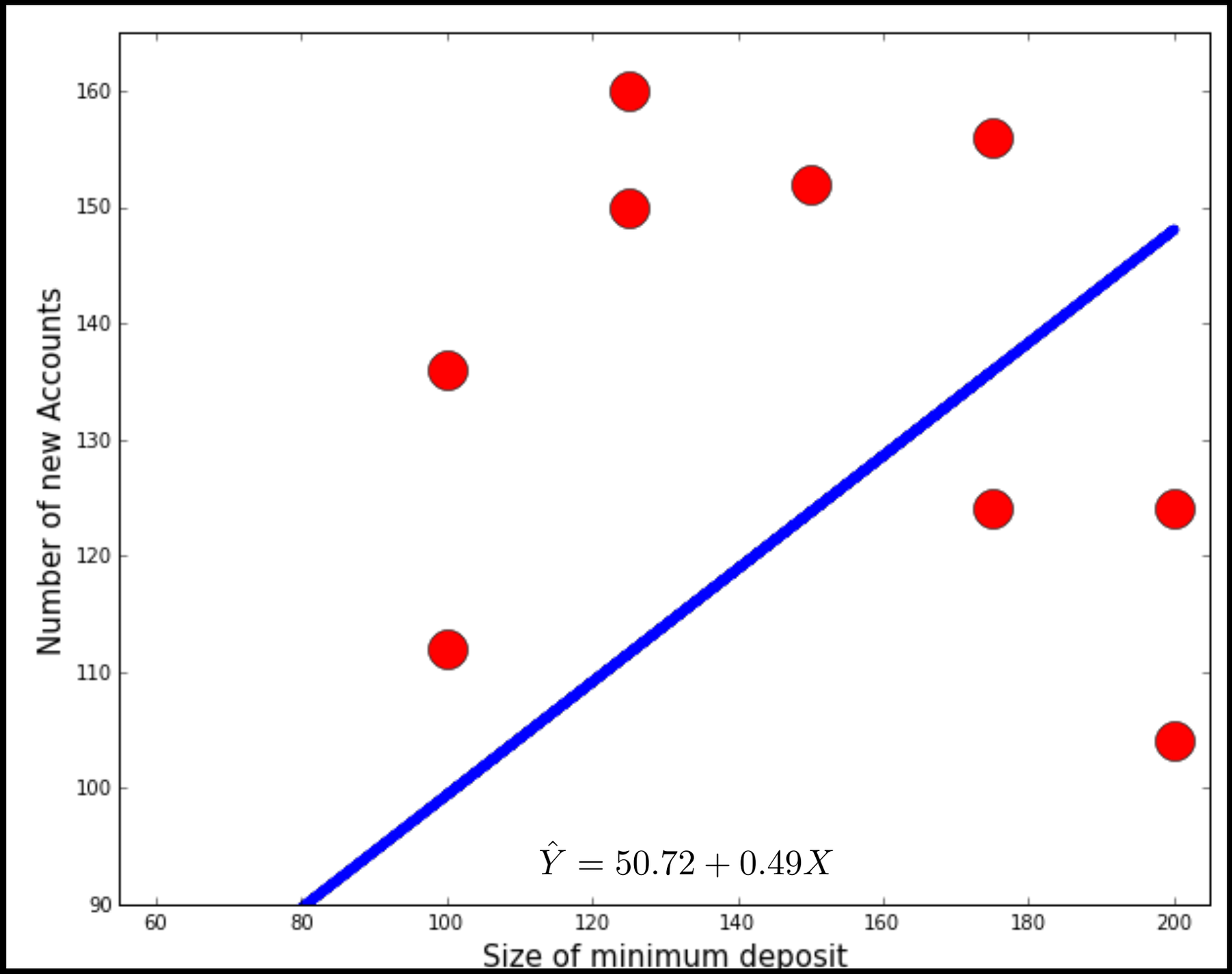
(c) Company B Machines



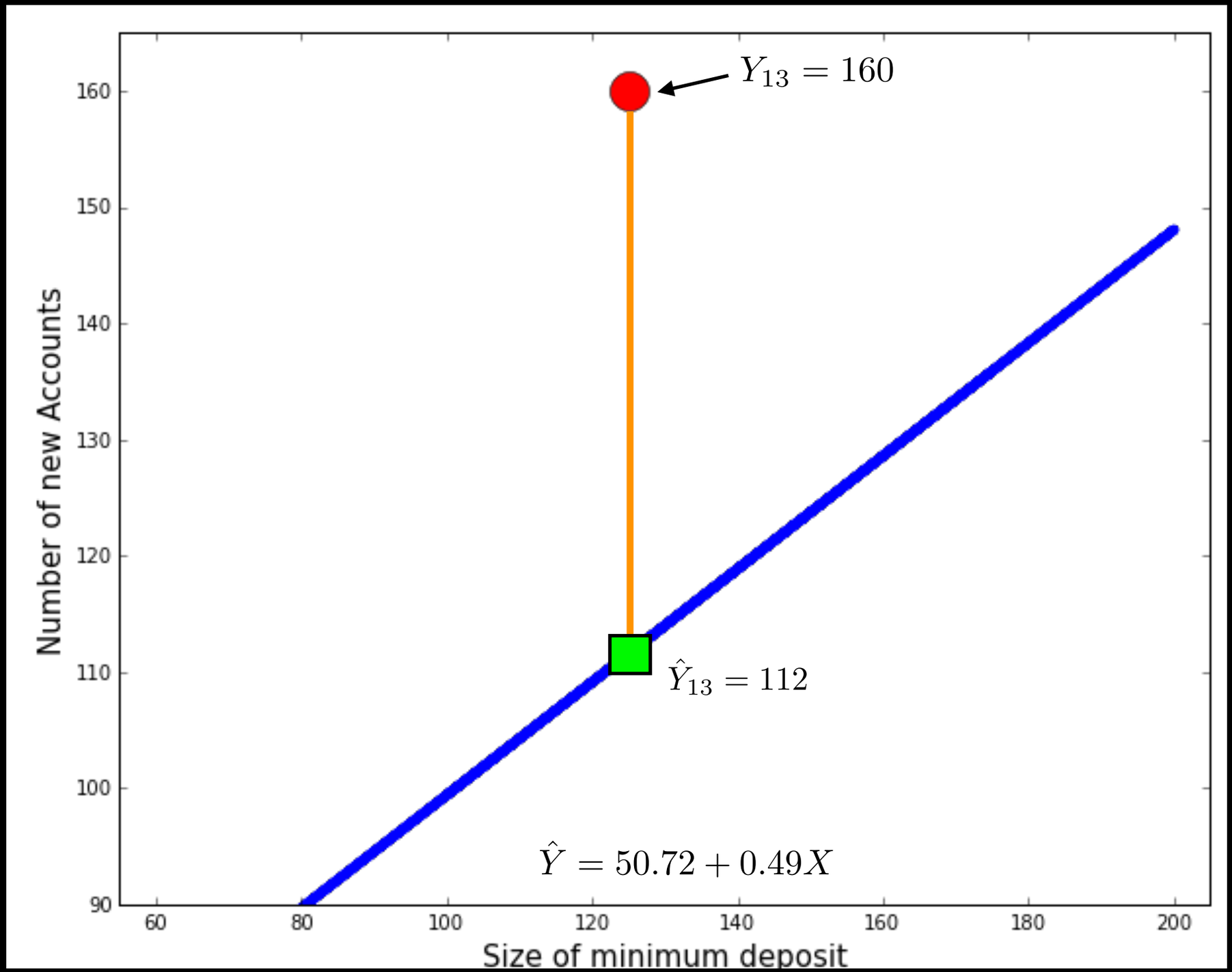
Decomposition of error deviation



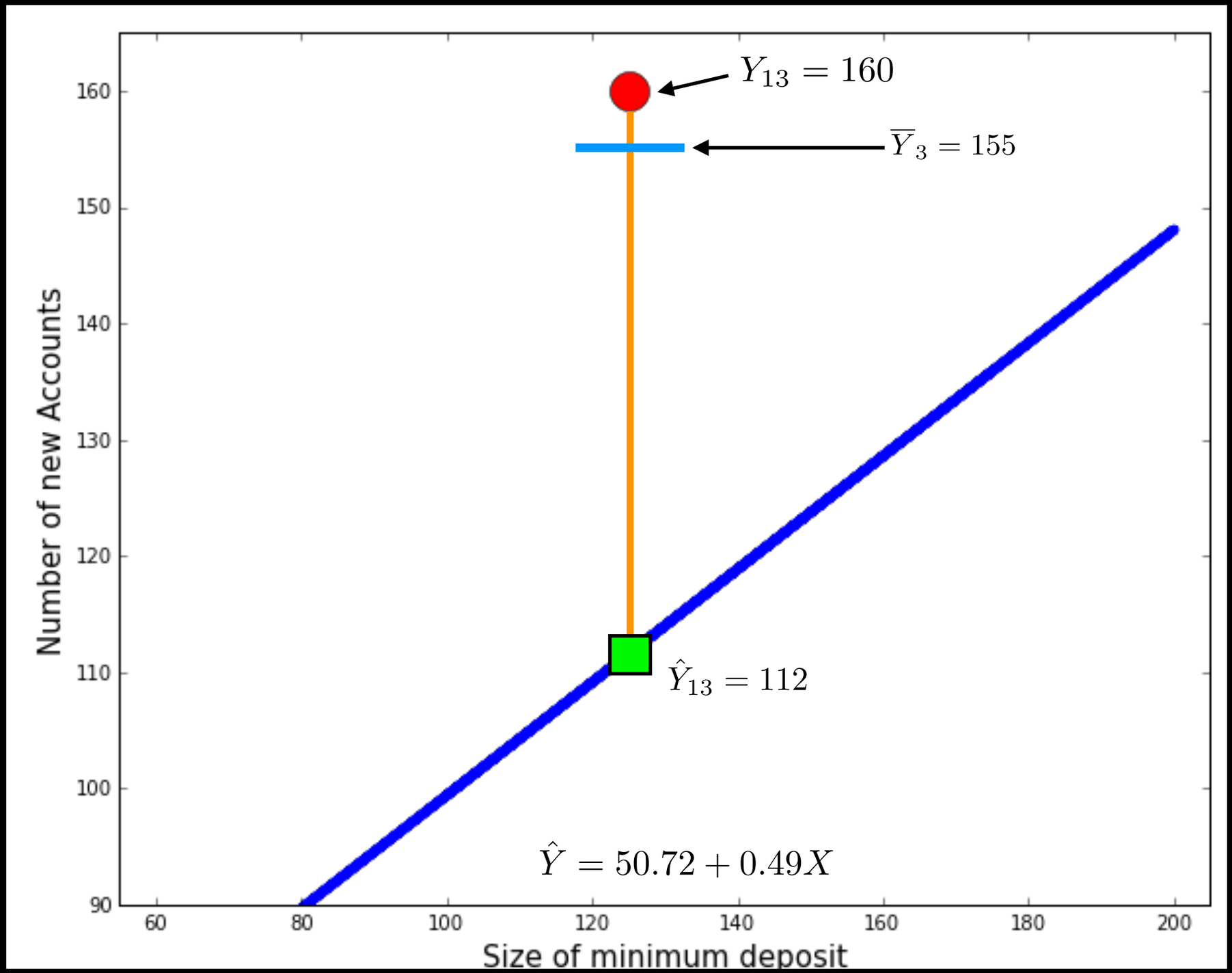
Decomposition of error deviation



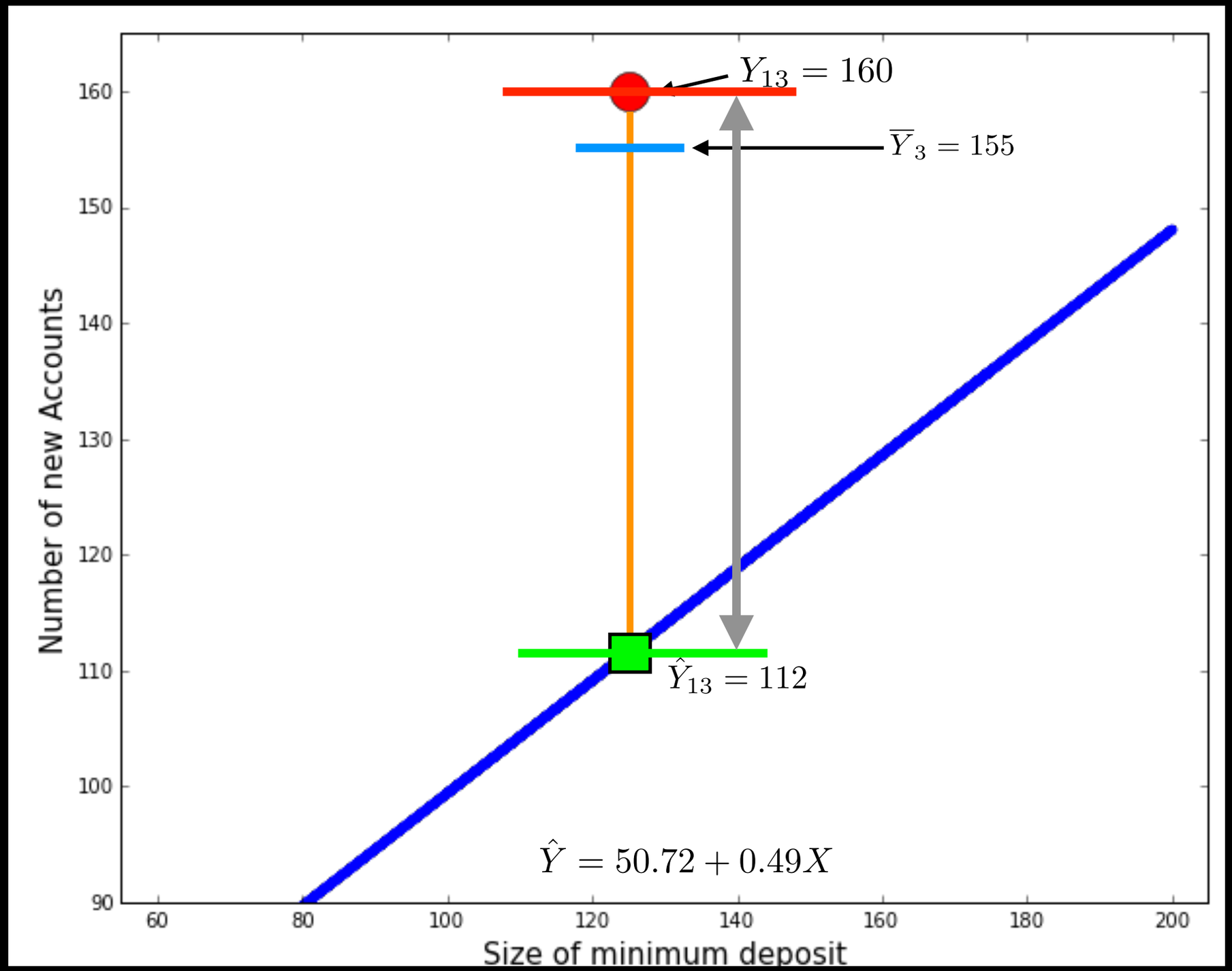
Decomposition of error deviation



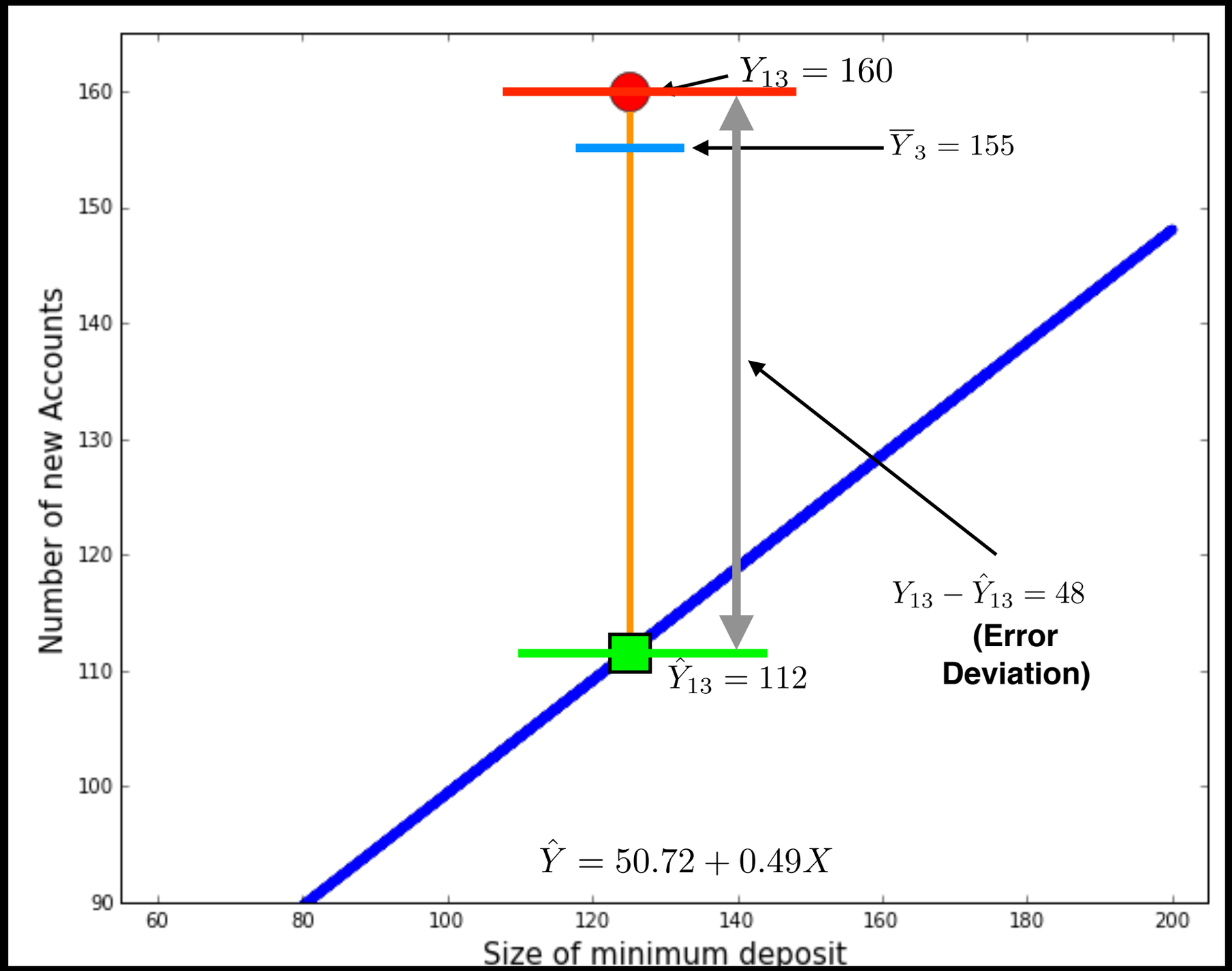
Decomposition of error deviation



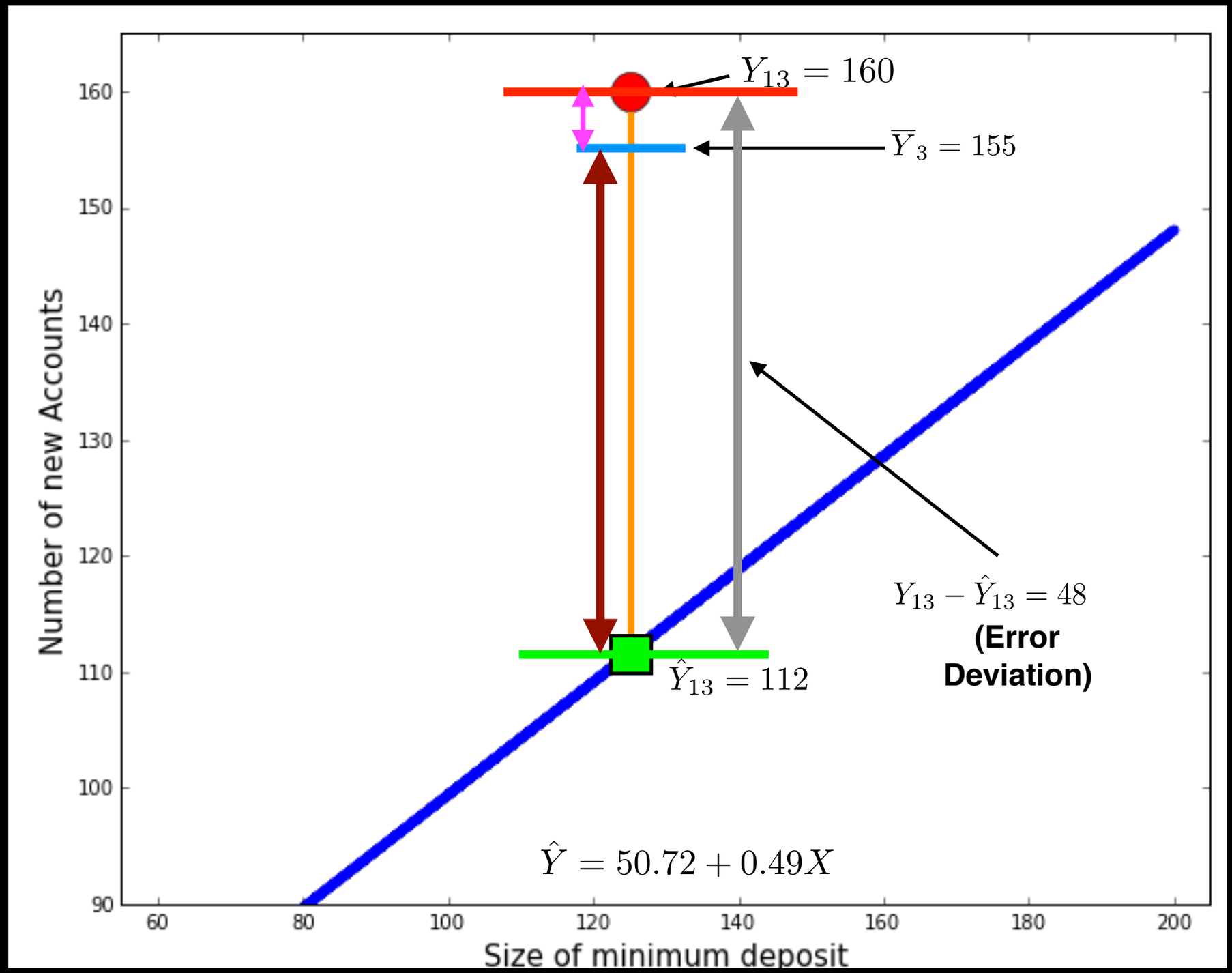
Decomposition of error deviation



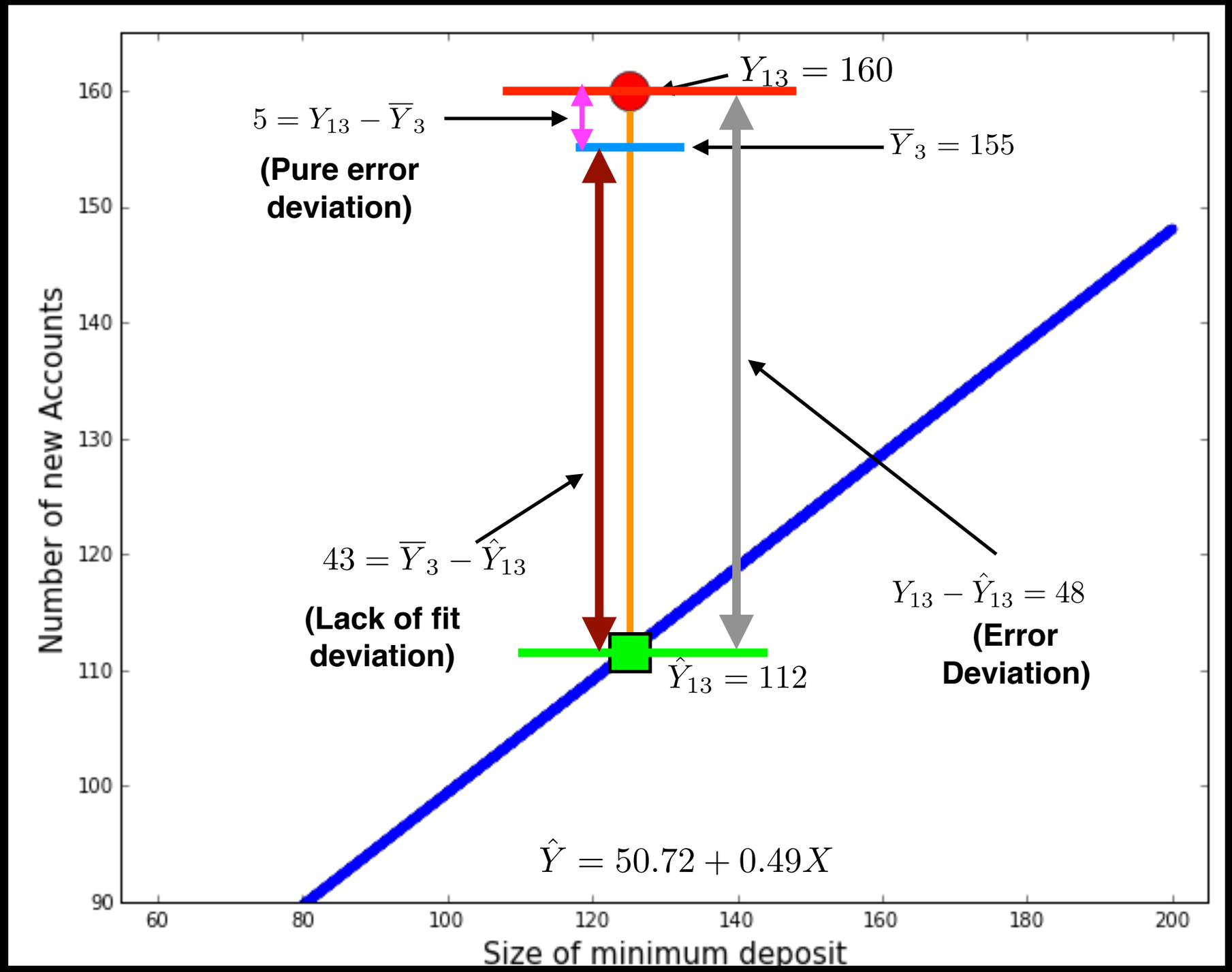
Decomposition of error deviation



Decomposition of error deviation



Decomposition of error deviation



General ANOVA table

Source of Variation	SS	df	MS
Regression	$SSR = \sum\sum(\hat{Y}_{ij} - \bar{Y})^2$	1	$MSR = \frac{SSR}{1}$
Error	$SSE = \sum\sum(Y_{ij} - \hat{Y}_{ij})^2$	$n-2$	$MSE = \frac{SSE}{n-2}$
Lack of fit	$SSLF = \sum\sum(\bar{Y}_j - \hat{Y}_{ij})^2$	$c-2$	$MSLF = \frac{SSLF}{c-2}$
Pure error	$SSPE = \sum\sum(Y_{ij} - \bar{Y}_j)^2$	$n-c$	$MSPE = \frac{SSPE}{n-c}$
Total	$SSTO = \sum\sum(Y_{ij} - \bar{Y})^2$	$n-1$	

General ANOVA table

Source of Variation	SS	df	MS
Regression	$SSR = \sum\sum(\hat{Y}_{ij} - \bar{Y})^2$	1	$MSR = \frac{SSR}{1}$
Error	$SSE = \sum\sum(Y_{ij} - \hat{Y}_{ij})^2$	$n-2$	$MSE = \frac{SSE}{n-2}$
Lack of fit	$SSLF = \sum\sum(\bar{Y}_j - \hat{Y}_{ij})^2$	$c-2$	$MSLF = \frac{SSLF}{c-2}$
Pure error	$SSPE = \sum\sum(Y_{ij} - \bar{Y}_j)^2$	$n-c$	$MSPE = \frac{SSPE}{n-c}$
Total	$SSTO = \sum\sum(Y_{ij} - \bar{Y})^2$	$n-1$	

Overview of the Remedial Measures

If normal/simple error linear regression model is not appropriate then you have two choices

1. Abandon regression model

2. Employ some transformation on the data so that regression model is appropriate for the transformed data.

Overview of the Remedial Measures

FIXES:

Nonlinearity of regression function

Either transform the data or use a different regression function altogether for example

Quadratic: $E\{Y\} = \beta_0 + \beta_1 X + \beta_2 X^2$

Exponential: $E\{Y\} = \beta_0 \beta_1^X$

Overview of the Remedial Measures

FIXES:

Nonconstancy of error variance

Transformations or weighted least squares (when variance varies in systematic fashion)

Overview of the Remedial Measures

FIXES:

Nonindependence of error terms

New model that assumes correlated error terms

Non-normality of error terms

Transformation of the data

[Sometimes the transformation that stabilizes the variance also fixes the normality]

Overview of the Remedial Measures

FIXES:

Outlying observations

Either discard the outliers or use robust estimation/ regression