Diagnostics and Remedial Measures

Yang Feng

Remedial Measures

- How do we know that the regression function is a good explainer of the observed data?
  - Plotting
  - Tests
- What if it is not? What can we do about it?
  - Transformation of variables

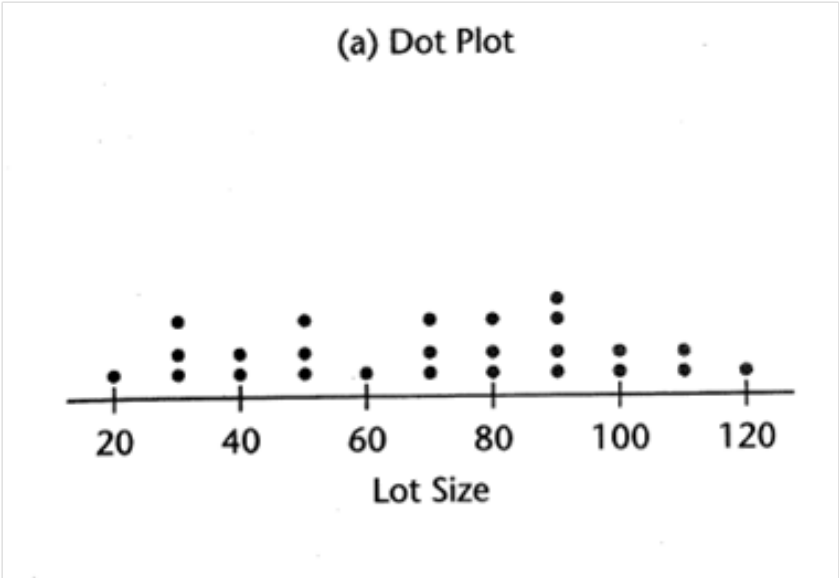
Graphical Diagnostics for the Predictor Variable

- Dot Plot
  - Useful for visualizing distribution of inputs
- Sequence Plot
  - Useful for visualizing dependencies between error terms
- Box Plot - Useful for visualizing distribution of inputs

Toluca manufacturing example again: production time vs. lot size

Dot Plot

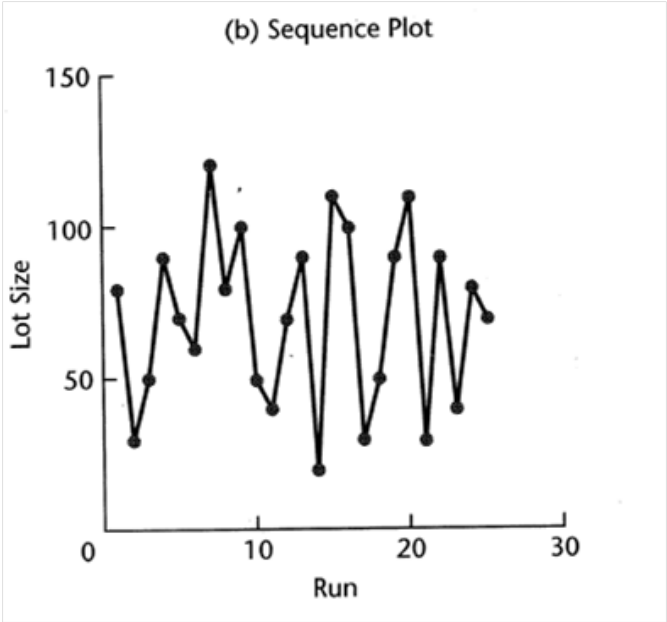
Figure :



- How many observations per input value?
- Range of inputs?

Sequence Plot

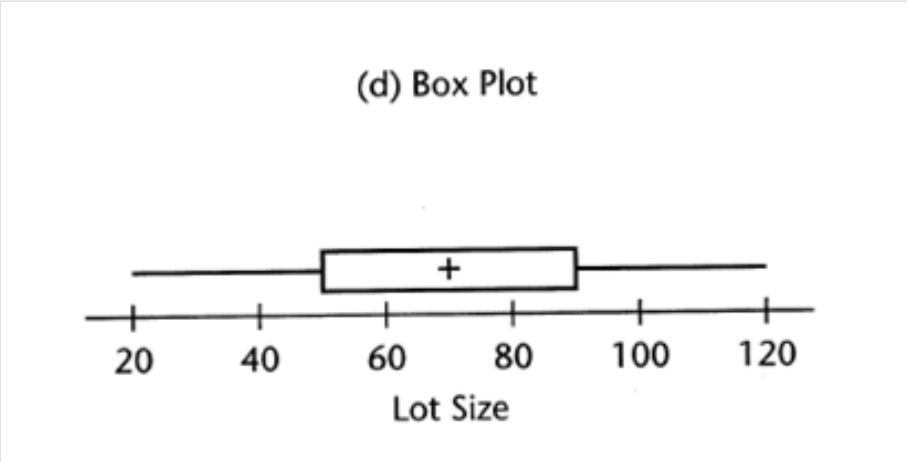
Figure :



- If observations are made over time, is there a correlation between input and position in observation sequence?

Box Plot

Figure :



- Shows
  - Median
  - 1st and 3rd quartiles
  - Maximum and minimum

Residuals

- Recall the definition of residuals:

$$e_i = Y_i - \hat{Y}_i$$

- And the difference between that and the unknown true error

$$\epsilon_i = Y_i - E(Y_i)$$

- In a normal regression model the ϵ_i's are assumed to be iid $N(0, \sigma^2)$ random variables. The observed residuals e_i should reflect these properties.

Remember: residual properties

- Mean

$$\bar{e} = \frac{\sum e_i}{n} = 0$$

- Variance

$$s^2 = \frac{\sum (e_i - \bar{e})^2}{n-2} = \frac{SSE}{n-2} = MSE$$

Nonindependence of Residuals

- The residuals e_i are not independent random variables - The fitted values $\hat{Y}_i$ are based on the same fitted regression line.
  - The residuals are subject to two constraints
    1. - Sum of the e_i's equals 0
    2. - Sum of the products $X_i e_i$'s equals 0
- When the sample size is large in comparison to the number of parameters in the regression model, the dependency effect among the residuals e_i can be safely ignored.

Definition: semistudentized residuals

- It may be useful sometimes to look at a standardized set of residuals, for instance in outlier detection.
- Like usual, since the standard deviation of ϵ_i is σ(itself estimated by square root of MSE) a natural form of standardization to consider is

$$e_i^* = \frac{e_i}{\sqrt{MSE}}$$

- This is called a semistudentized residual.

Departures from Model...

To be studied by residuals

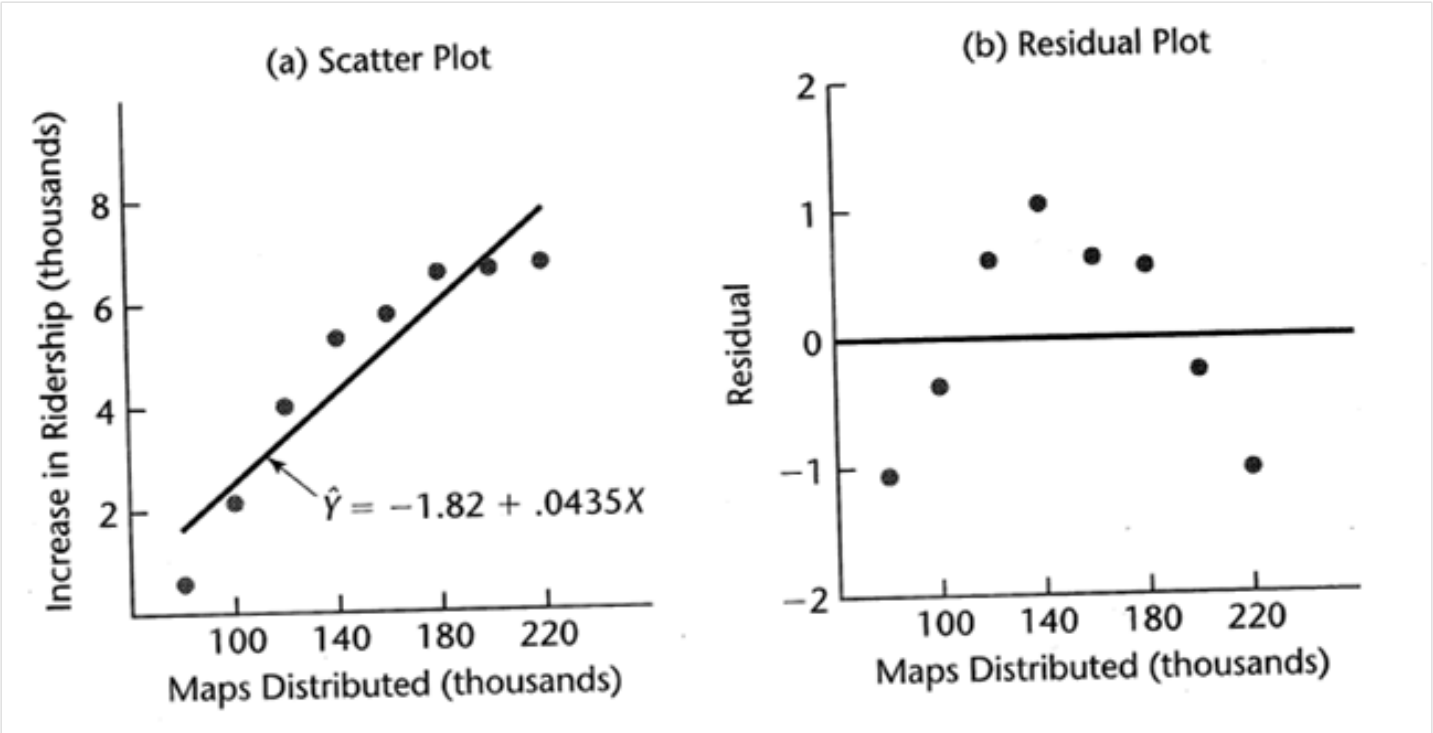
- Regression function not linear
- Error terms do not have constant variance
- Error terms are not independent
- Model fits all but one or a few outlier observations
- Error terms are not normally distributed
- One or more predictor variables have been omitted from the model

Diagnostics for Residuals

- Plot of residuals against predictor variable
- Plot of absolute or squared residuals against predictor variable
- Plot of residuals against fitted values
- Plot of residuals against time or other sequence
- Plot of residuals against omitted predictor variables
- Box plot of residuals
- Normal probability plot of residuals

1. Test for nonlinearity of Regression Function: Residual Plot against the predictor variable

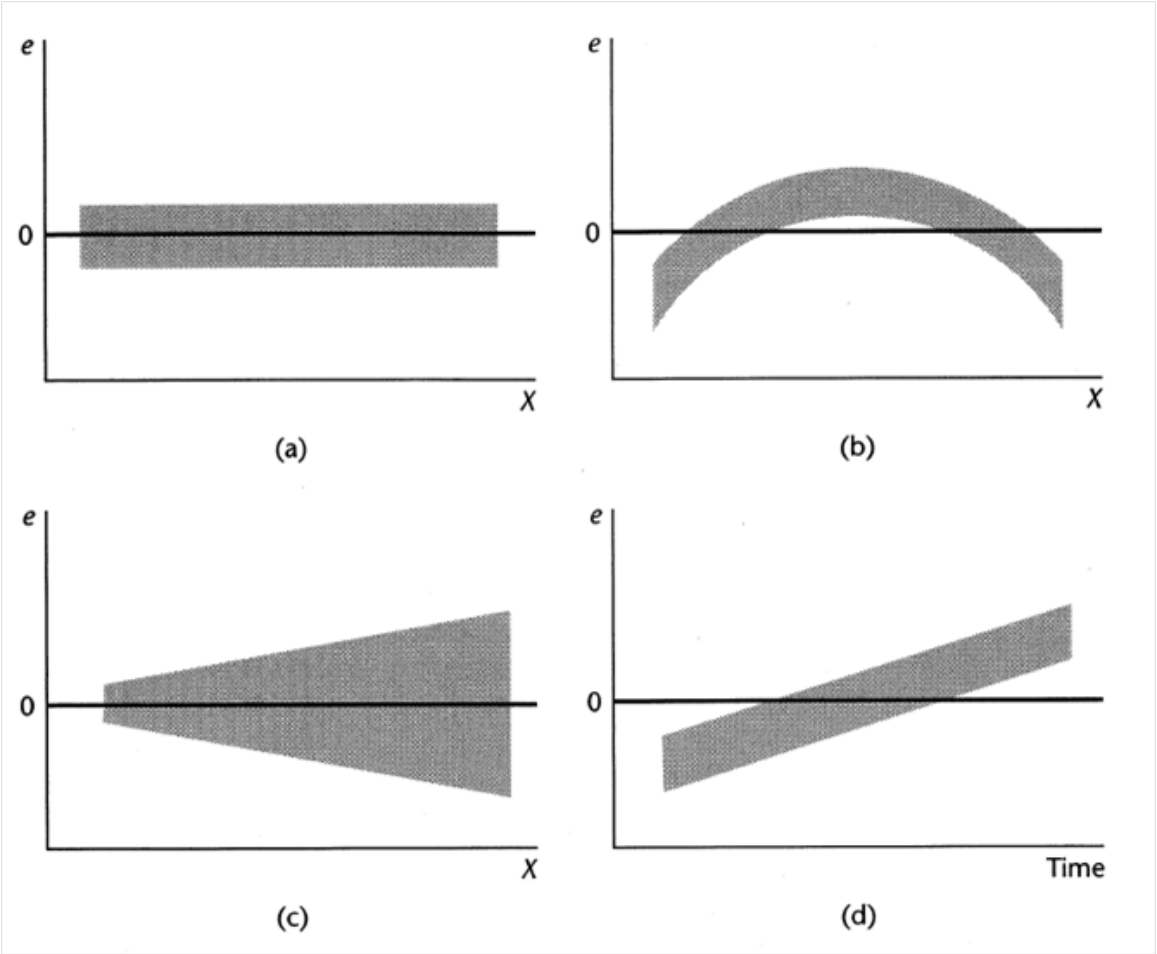
Figure : Transit example : ridership increase vs. num. maps distributed



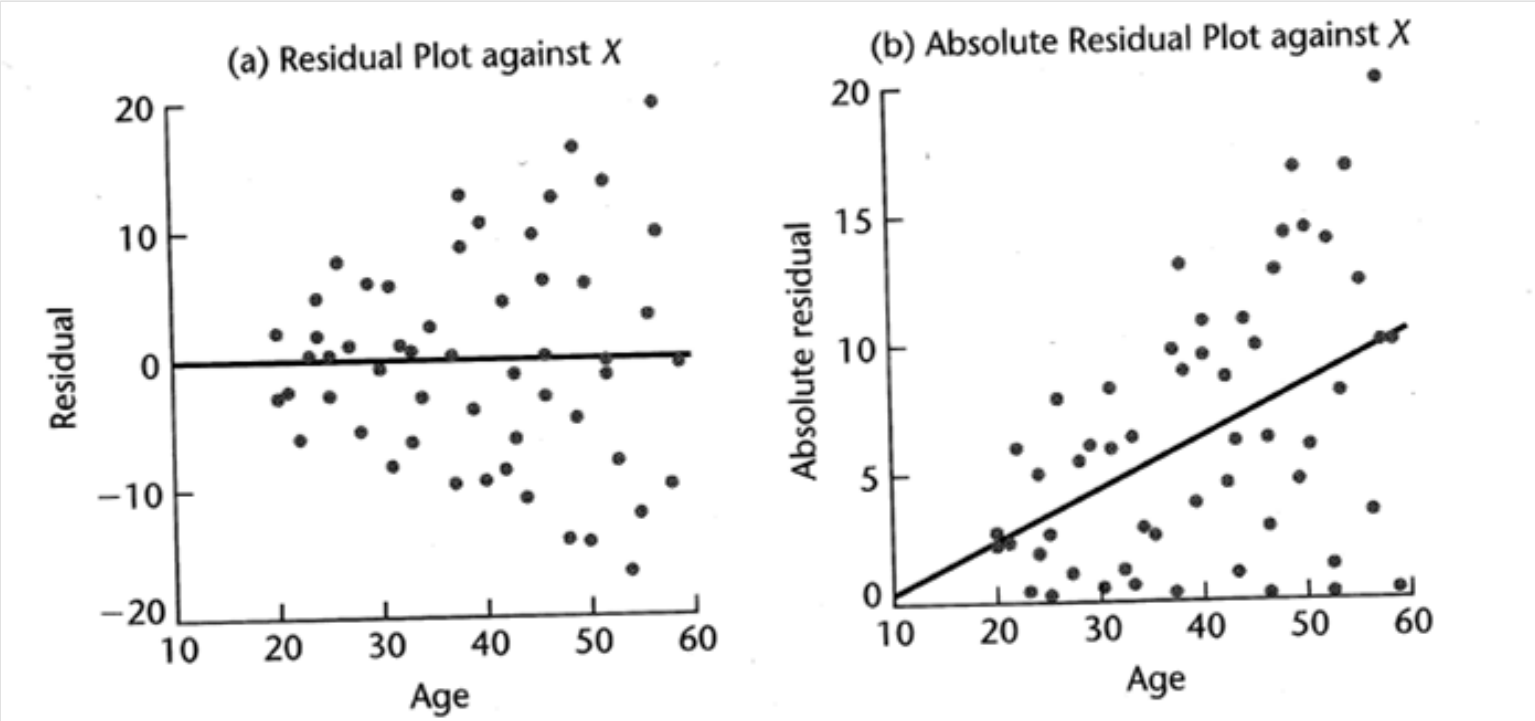
Should be no systematic relationship between residual and predictor variable if it is linear related.

Prototype Residual Plots

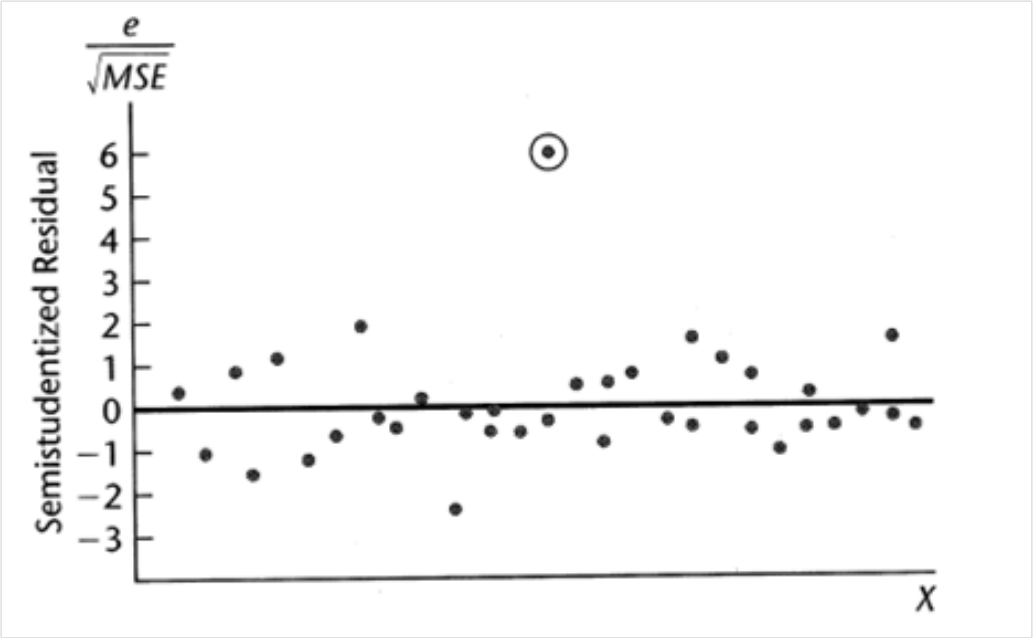
Figure : Indicate residual plots



2. Nonconstancy of Error Variance

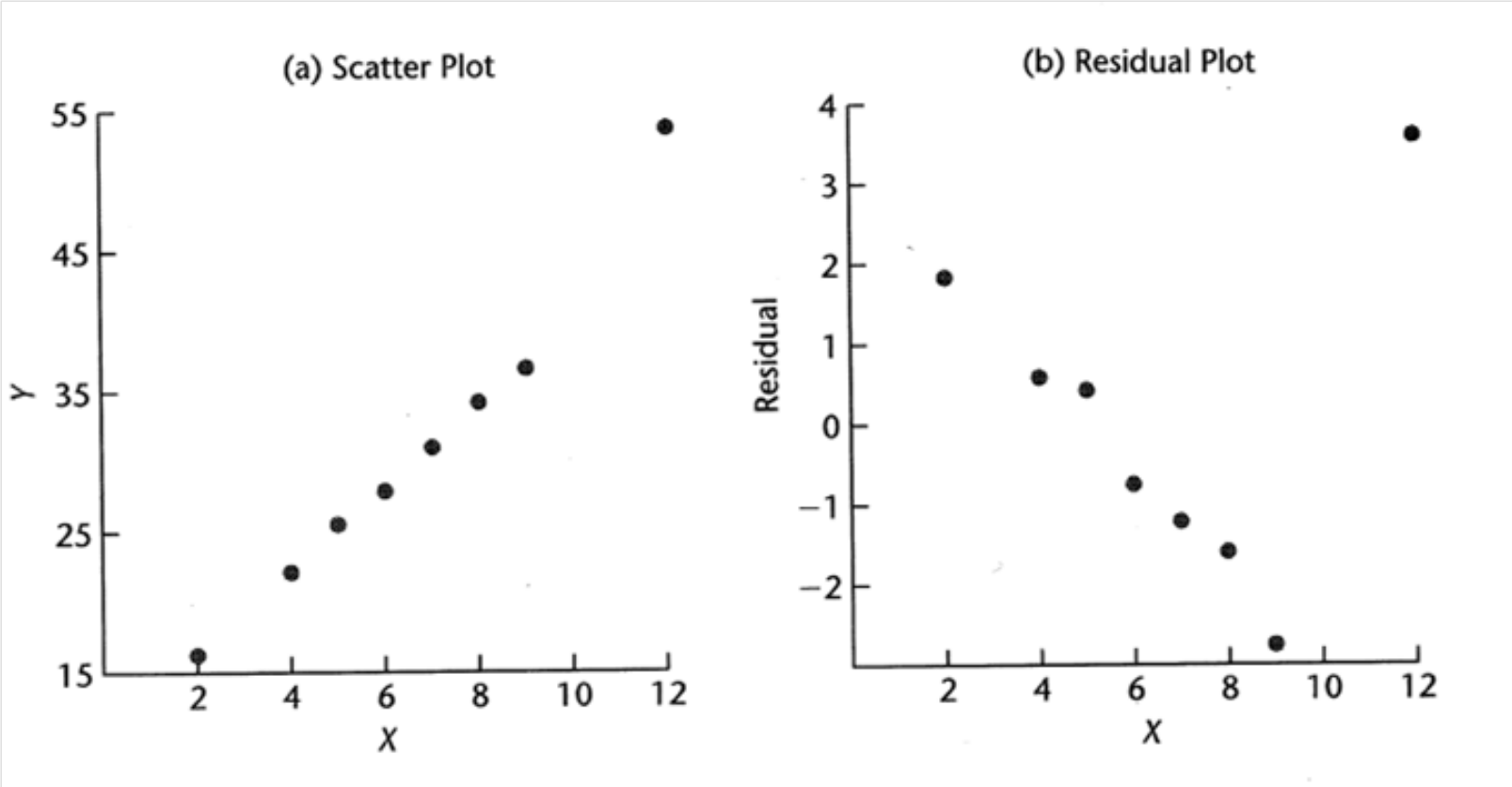


3. Presence of Outliers

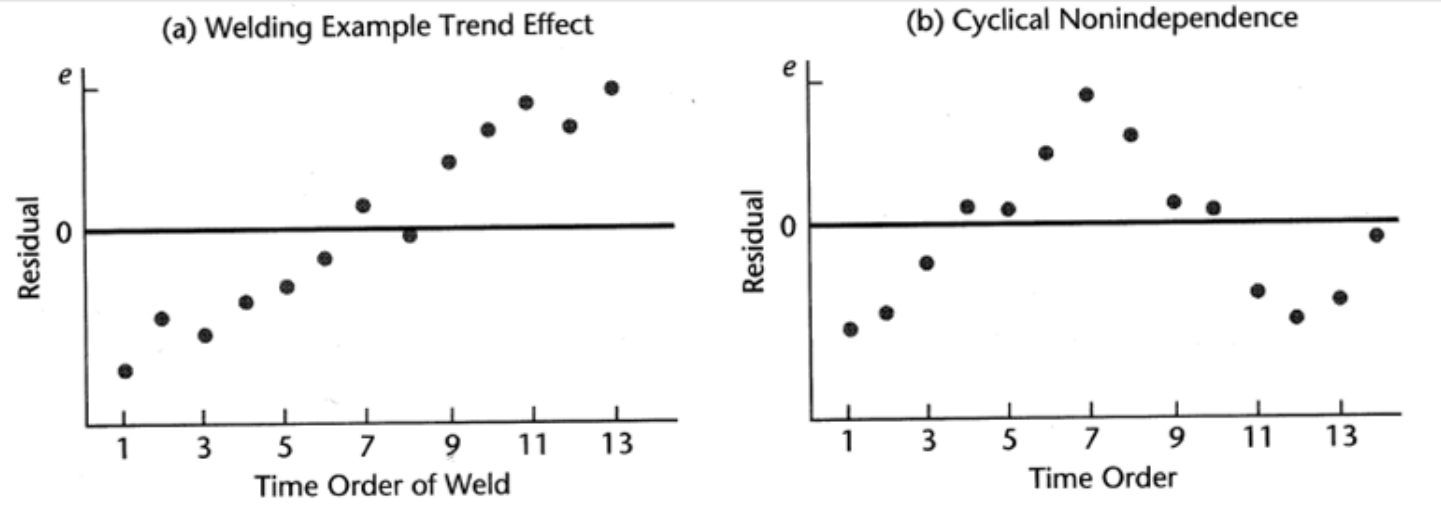


Outliers can strongly effect the fitted values of the regression line.

Outlier effect on residuals



4. Nonindependence of Error Terms

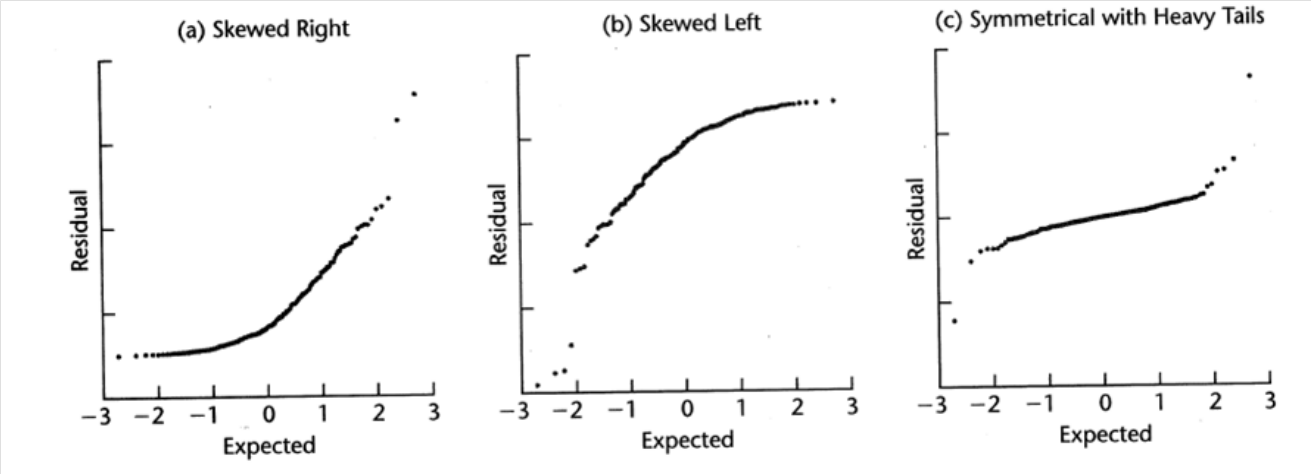


Sequential observations can exhibit observable trends in error distribution.
Application: Time series.

5. Non-normality of Error Terms

- Distribution plots. (e.g., boxplot)
- Comparison of Frequencies. (e.g., 68 percent of the residuals fall between $\pm\sqrt{\text{MSE}}$ or about 90 percent between $\pm 1.645\sqrt{\text{MSE}}$.
- Normal probability plot
  Q-Q plot with numerical quantiles on the horizontal axis

Figure : Examples of non-normality in distribution of error terms



Normal probability plot

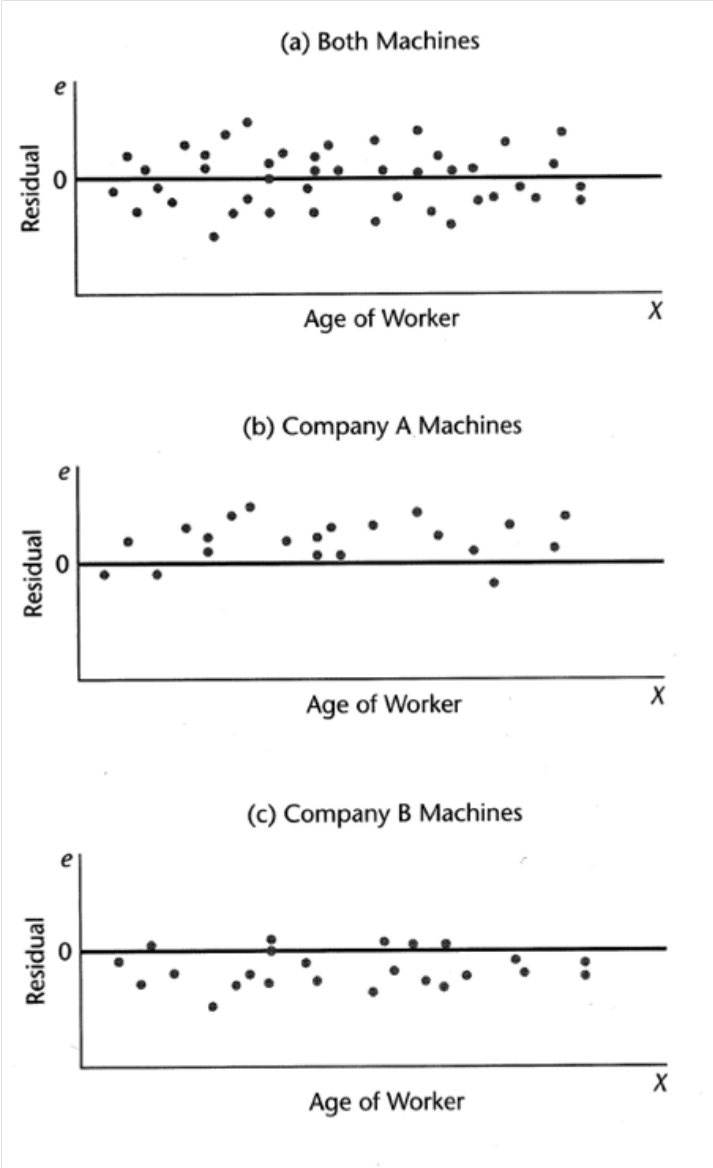
- For a $N(0, MSE)$ random variable, a good approximation of the expected value of the k-th smallest observation in a random sample of size n is

$$\sqrt{MSE}[z(\frac{k - .375}{n + .25})]$$

- A normal probability plot consists of plotting the expected value of the k-th smallest observation against the observed k-th smallest observation

6. Omission of Important Predictor Variables

- Example
  - Qualitative variable
  - Type of machine
- Partitioning data can reveal dependence on omitted variable(s)
- Works for quantitative variables as well
- Can suggest that inclusion of other inputs is important



Tests Involving Residuals

- Tests for randomness (run test, Durbin-Watson test, Chapter 12)
- Tests for constancy of variance (Brown-Forsythe test, Breusch-Pagan test, Section 3.6)
- Tests for outliers (fit a new regression line to the other $n-1$ observations. detail in Chapter 10)
- Tests for normality of error distribution (will discuss now.)

Correlation Test for Normality of Error Distribution

For one way to run this correlation test let

$$e = \{e_{\sigma(1)}, \ldots, e_{\sigma(n)}\}$$

be the ordered sequence (from smallest to the largest) of the observed errors. Let

$$r = \{r_1, \ldots, r_k, \ldots, r_n\},$$

where r_k is the expected value of the k^{th} residual under the normality assumption, i.e.

$$r_k \approx \sqrt{MSE}[z(\frac{k - .375}{n + .25})]$$

Then compute the sample correlation between e and r.

Correlation Test for Normality of Error Distribution

- A formal test for the normality of the error terms can be developed in terms of the coefficient of correlation between the residuals e_i and their expected values under normality. High value indicates normality!
- Tables (B.6 in the book) gives critical values for the null hypothesis (normally distributed errors).
- Less than the critical value, reject the null hypothesis!
- Toluca Company example: coefficient of correlation: $0.991 > 0.959$, which is the critical value for α risk at 0.05. Then, conclude, it does not depart from normal substantially!

Tests for Constancy of Error Variance

- Brown-Forsythe test does not depend on normality of error terms. The Brown-Forsythe test is applicable to simple linear regression when
  - The variance of the error terms either increases or decreases with X ("megaphone" residual plot)
  - Sample size is large enough to ignore dependencies between the residuals
- The Brown-Forsythe test is essentially a t-test for testing whether the means of two normally distributed populations are the same where the populations are the absolute deviations between the prediction and the observed output space in two non-overlapping partitions of the input space.

Brown-Forsythe Test

- Divide X into X_1 (the low values of X) and X_2 (the high values of X)
- Let e_{i1} be the i-th residual for X_1 and vice versa
- $\tilde{e_1}$ and $\tilde{e_2}$ denote the median of the residuals.
- let $n = n_1 + n_2$
- The Brown-Forsythe test uses the absolute deviations of the residuals around their group median

$$d_{i1} = |e_{i1} - \tilde{e_1}|, d_{i2} = |e_{i2} - \tilde{e_2}|$$

Brown-Forsythe Test

- The test statistic for comparing the means of the absolute deviations of the residuals around the group medians is

$$t^*_{BF} = \frac{\bar{d}_1 - \bar{d}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where the pooled variance

$$s^2 = \frac{\sum(d_{i1} - \bar{d}_1)^2 + \sum(d_{i2} - \bar{d}_2)^2}{n - 2}$$

Brown-Forsythe Test

- If n_1 and n_2 are not extremely small

$$t^*_{BF} \sim t(n-2)$$

  approximately
- From this confidence intervals and tests can be constructed.

Breusch-Pagan Test

- Another test for the constancy of error variance.
- Assume error terms are independently and normally distributed and

$$\sigma_i^2 = \gamma_0 + \gamma_1 X_i$$

- Then the problem reduces to

$$H_0 : \gamma_1 = 0 \text{ v.s. } H_1 : \gamma_1 \neq 0$$

- Test statistics is derived in the following two steps
  1. Regress Y on X, get SSE and residual vector e.
  2. Regress e^2 on X, get SSR^*.
-

$$X_{BP}^2 = \frac{SSR^*}{2(\frac{SSE}{n})^2}$$

BP-test

- Under H_0, when n is reasonably large,

$$X^2_{BP} \sim \chi^2(1)$$

- If $X^2_{BP} > \text{qchisq}(1-\alpha, 1)$, reject H_0.
- Direct function in R:
  ncvTest in the package car.

F test for lack of fit

- Formal test for determining whether a specific type of regression function adequately fits the data.
- Assume the observations $Y|X$ are
  1. independent
  2. normally distributed
  3. same variance σ^2
- Requires: repeat observations at one or more X levels (called replicates)

Example

- 11 similar branches of a bank offered gifts for setting up money market accounts
- Minimum initial deposits were specific to qualify for the gift
- Value of gift was proportional to the specified minimum deposit
- Interested in: relationship between specified minimum deposit and number of new accounts opened

F Test Example Data and ANOVA Table

Figure :

(a) Data

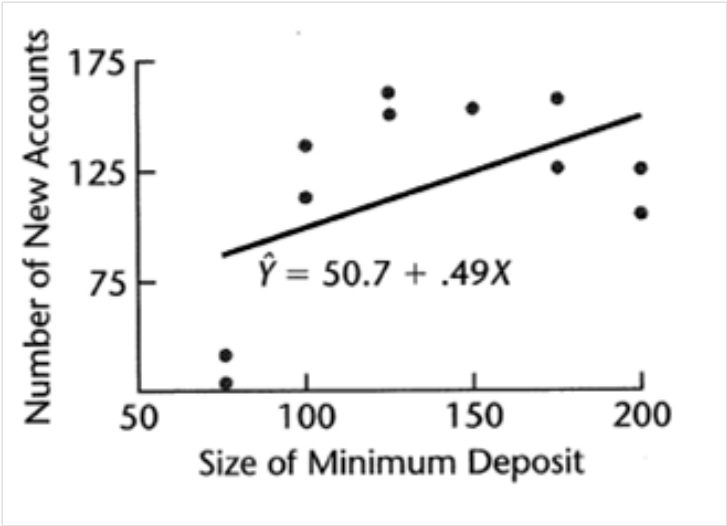
Branch i	Size of Minimum Deposit (dollars) X_i	Number of New Accounts Y_i	Branch i	Size of Minimum Deposit (dollars) X_i	Number of New Accounts Y_i
1	125	160	7	75	42
2	100	112	8	175	124
3	200	124	9	125	150
4	75	28	10	200	104
5	150	152	11	100	136
6	175	156			

(b) ANOVA Table

Source of Variation	SS	df	MS
Regression	5,141.3	1	5,141.3
Error	14,741.6	9	1,638.0
Total	19,882.9	10	

Fit

Figure :


Number of New Accounts
175
125
75
$\hat{Y} = 50.7 + .49X$
50
100
150
200
Size of Minimum Deposit

Data Arranged To Highlight Replicates

Figure :

	Size of Minimum Deposit (dollars)					
Replicate	$j=1$ $X_1=75$	$j=2$ $X_2=100$	$j=3$ $X_3=125$	$j=4$ $X_4=150$	$j=5$ $X_5=175$	$j=6$ $X_6=200$
$i=1$	28	112	160	152	156	124
$i=2$	42	136	150		124	104
Mean $\bar{Y}_j$	35	124	155	152	140	114

- The observed value of the response variable for the i-th replicate for the j-th level of X is Y_{ij}
- The mean of the Y observations at the level $X = X_j$ is $\bar{Y}_j$

Full Model vs. Regression Model

- The full model is

$$Y_{ij} = \mu_j + \epsilon_{ij}$$

where

1. μ_j are parameters, $j = 1, ..., c$
2. ϵ_{ij} are iid $N(0, \sigma^2)$

- Since the error terms have expectation zero

$$E(Y_{ij}) = \mu_j$$

Full Model

- In the full model there is a different mean (a free parameter) for each X_i
- In the regression model the mean responses are constrained to lie on a line

$$E(Y) = \beta_0 + \beta_1 X$$

Fitting the Full Model

- The estimators of μ_j are simply

$$\hat{\mu}_j = \bar{Y}_j$$

- The error sum of squares of the full model therefore is

$$SSE(F) = \sum\sum(Y_{ij} - \bar{Y}_j)^2 = SSPE$$

SSPE: Pure Error Sum of Squares

Degrees of Freedom

- Ordinary total sum of squares had n-1 degrees of freedom.
- Each of the j terms is a ordinary total sum of squares
  - Each then has $n_j - 1$ degrees of freedom
- The number of degrees of freedom of SSPE is the sum of the component degrees of freedom

$$df_F = \sum_j (n_j - 1) = \sum_j n_j - c = n - c$$

General Linear Test

- Remember: the general linear test proposes a reduced model

$$H_0 : E(Y) = \beta_0 + \beta_1 X \text{ (Normal regression model)}$$

$$H_a : E(Y) \neq \beta_0 + \beta_1 X \text{ (The full model, one independent mean for each level of X)}$$

SSE For Reduced Model

The SSE for the reduced model is as before

- remember

$$SSE(R) = \sum_i \sum_j [Y_{ij} - (b_0 + b_1 X_j)]^2$$
$$= \sum_i \sum_j (Y_{ij} - \hat{Y}_{ij})^2$$

- and has n-2 degrees of freedom $df_R = n - 2$

SSE(R)

Figure :

(a) Data

Branch i	Size of Minimum Deposit (dollars) X_i	Number of New Accounts Y_i	Branch i	Size of Minimum Deposit (dollars) X_i	Number of New Accounts Y_i
1	125	160	7	75	42
2	100	112	8	175	124
3	200	124	9	125	150
4	75	28	10	200	104
5	150	152	11	100	136
6	175	156			

(b) ANOVA Table

Source of Variation	*SS*	*df*	*MS*
Regression	5,141.3	1	5,141.3
Error	14,741.6	9	1,638.0
Total	19,882.9	10	

F Test Statistic

From the general linear test approach

$$F^* = \frac{SSE(R) - SSE(F)}{df_R - df_F} \div \frac{SSE(F)}{df_F}$$
$$= \frac{SSE - SSPE}{(n-2) - (n-c)} \div \frac{SSPE}{n-c}$$

Lack of fit sum of squares:

$$SSLF = SSE - SSPE$$

Then

$$F^* = \frac{SSLF}{(n-2) - (n-c)} \div \frac{SSPE}{n-c} = \frac{MSLF}{MSPE}$$

F Test Rule

- From the F test we know that large values of F^* lead us to reject the null hypothesis:
  If $F^* \leq F(1-\alpha; c-2, n-c)$, conclude H_0
  If $F^* > F(1-\alpha; c-2, n-c)$, conclude H_a
- For this example we have

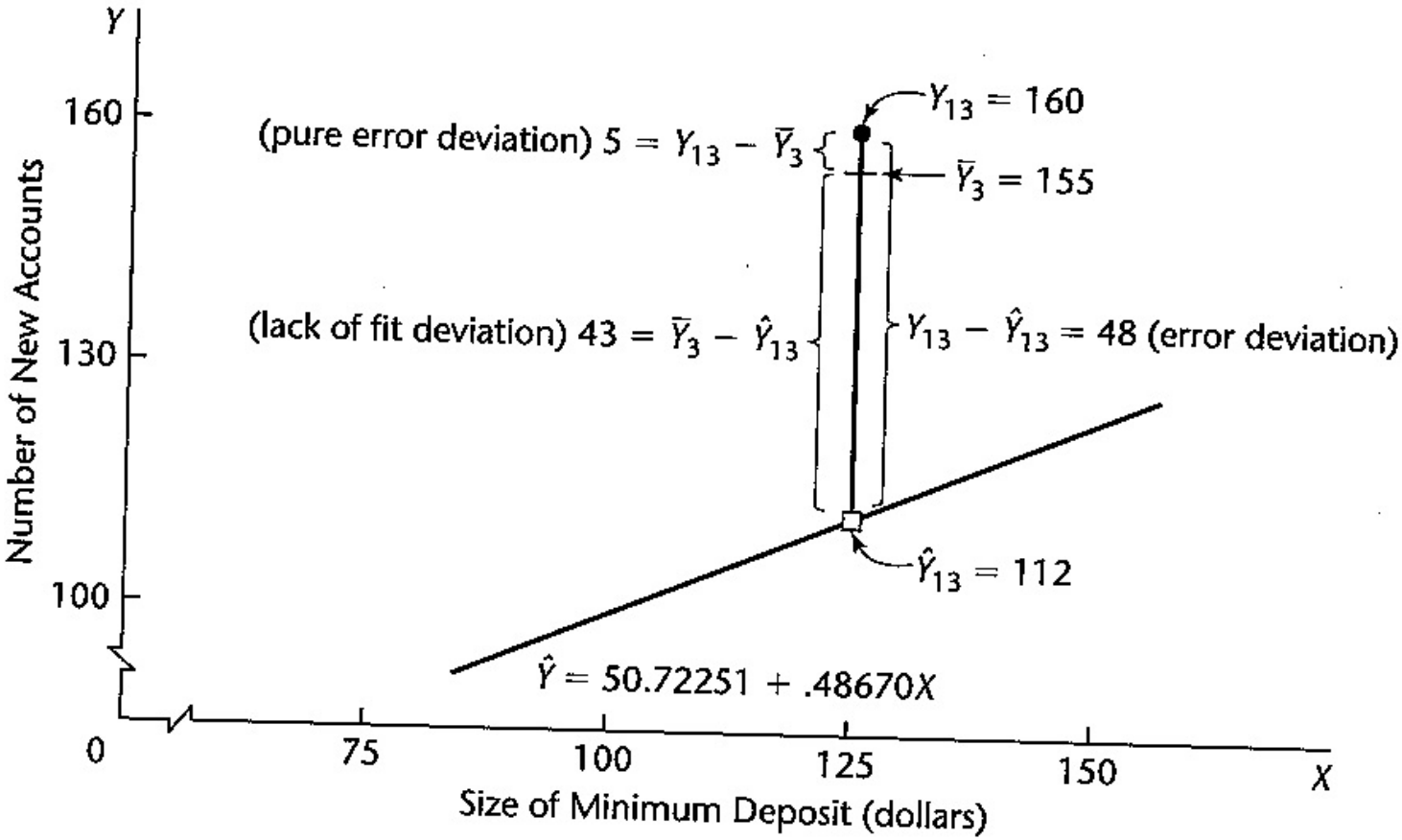
$$SSPE = 1{,}148.0 \qquad n - c = 11 - 6 = 5$$
$$SSE = 14{,}741.6$$
$$SSLF = 14{,}741.6 - 1{,}148.0 = 13{,}593.6 \qquad c - 2 = 6 - 2 = 4$$
$$F^* = \frac{13{,}593.6}{4} \div \frac{1{,}148.0}{5}$$
$$= \frac{3{,}398.4}{229.6} = 14.80$$

Variance decomposition

SSE = SSPE + SSLF.

$$\sum\sum(Y_{ij} - \hat{Y}_{ij})^2 = \sum\sum(Y_{ij} - \bar{Y}_j)^2 + \sum\sum(\bar{Y}_j - \hat{Y}_{ij})^2$$

Example decomposition



Example decomposition

(a) General

Source of Variation	*SS*	*df*	*MS*
Regression	SSR	1	MSR
Error	SSE	$n-2$	MSE
Lack of fit	$SSLF$	$c-2$	$MSLF$
Pure error	$SSPE$	$n-c$	$MSPE$
Total	$SSTO$	$n-1$	

(b) Bank Example

Source of Variation	*SS*	*df*	*MS*
Regression	$SSR = 5{,}141.3$	1	$MSR = 5{,}141.3$
Error	$SSE = 14{,}741.6$	9	$MSE = 1{,}638.0$
Lack of fit	$SSLF = 13{,}593.6$	4	$MSLF = 3{,}398.4$
Pure error	$SSPE = 1{,}148.0$	5	$MSPE = 229.6$
Total	$SSTO = 19{,}882.9$	10	

Example Conclusion

- If we set the significance level to $\alpha = .01$
- And look up the value of the F inv-cdf $F(.99, 4, 5) = 11.4$
- We can conclude that the null hypothesis should be rejected.

A review

- Graphical procedures for determining appropriateness of regression fit
  - Various Residual plots
- Tests to determine
  - Constancy of error variance
  - Lack of fit test
- what do we do if we determine(through testing or otherwise) that the linear regression fit is not good?

How to fix

If simple regression model is not appropriate then there are two choices:

1. Abandon simple regression model and develop and use a more appropriate model
   e.g., logistic regression, nonparametric regression, etc...
   may yield better insights, but a more complex model lead to more complex procedures for estimating the parameters. (Later in the course)
2. Employ some transformation of the data so that the simple regression model is appropriate for the transformed data. (This chapter)

Fixes For...

- Nonlinearity of regression function - Transformation(s) (today)
- Nonconstancy of error variance - Weighted least squares (Chapter 11) and transformations
- Nonindependence of error terms - Directly model correlation or use first differences (Chapter 12)
- Non-normality of error terms - Transformation(s) (today)
- Omission of Important Predictor Variables - Modify the model—-Multiple regression analysis, Chapter 6 and later on.
- Outlying observations - Robust regression (Chapter 11)

Nonlinearity of regression function

Direct approach

- Modify regression model by altering the nature of the regression function. For instance, a quadratic regression function might be used

$$E(Y) = \beta_0 + \beta_1 X + \beta_2 X^2$$

- or an exponential function

$$E(Y) = \beta_0 \beta_1^X$$

- Such approaches employ a tranformation to (approximately) linearize a regression function

Quick Questions

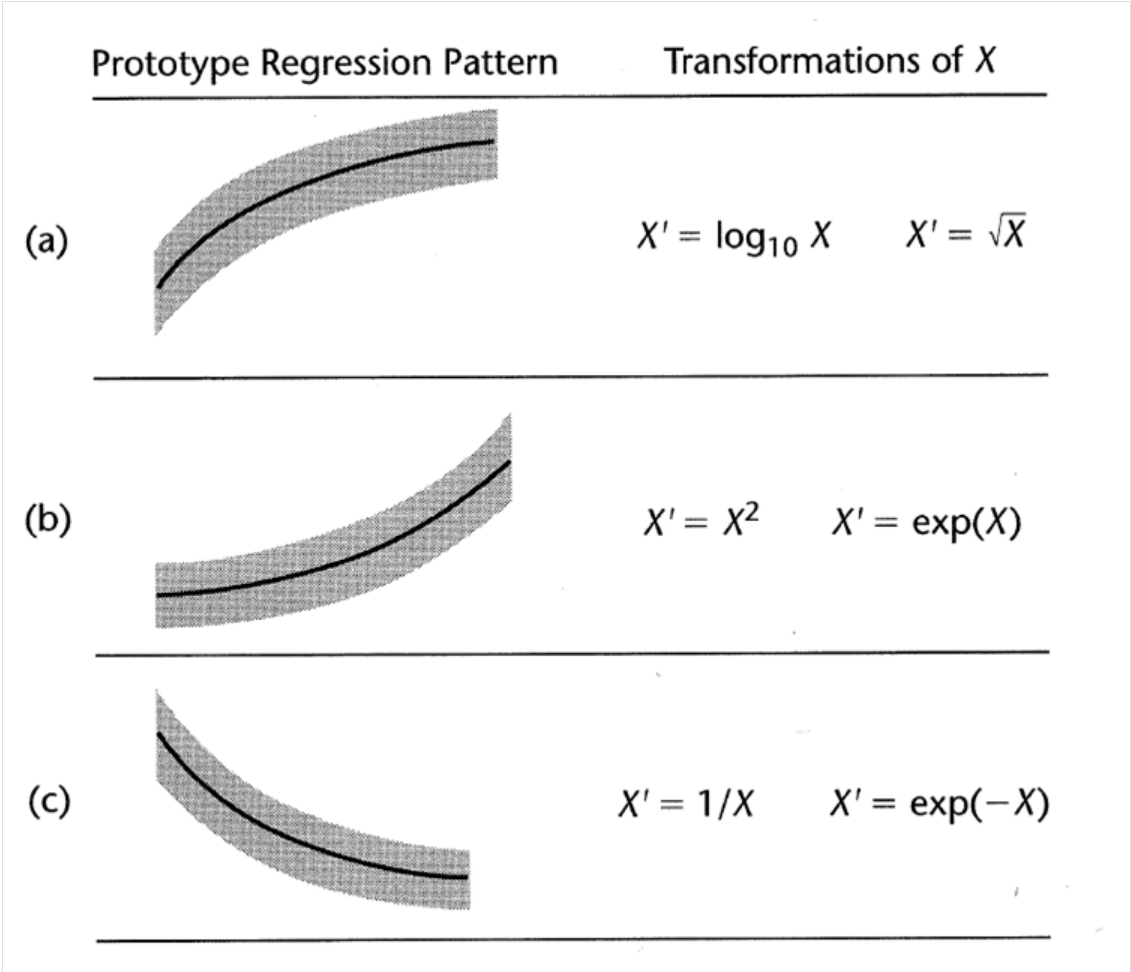
- How would you fit such models?
- How does the exponential regression function relate to regular linear regression?
- Where did the error terms go?

Transformations

Transformations for nonlinearity relation only
- Appropriate when the distribution of the error terms is reasonably close to a normal distribution
- In this situation
1. transformation of X should be attempted;
2. transformation of Y should not be attempted because it will materially effect the distribution of the error terms.

Prototype Regression Patterns

Figure :



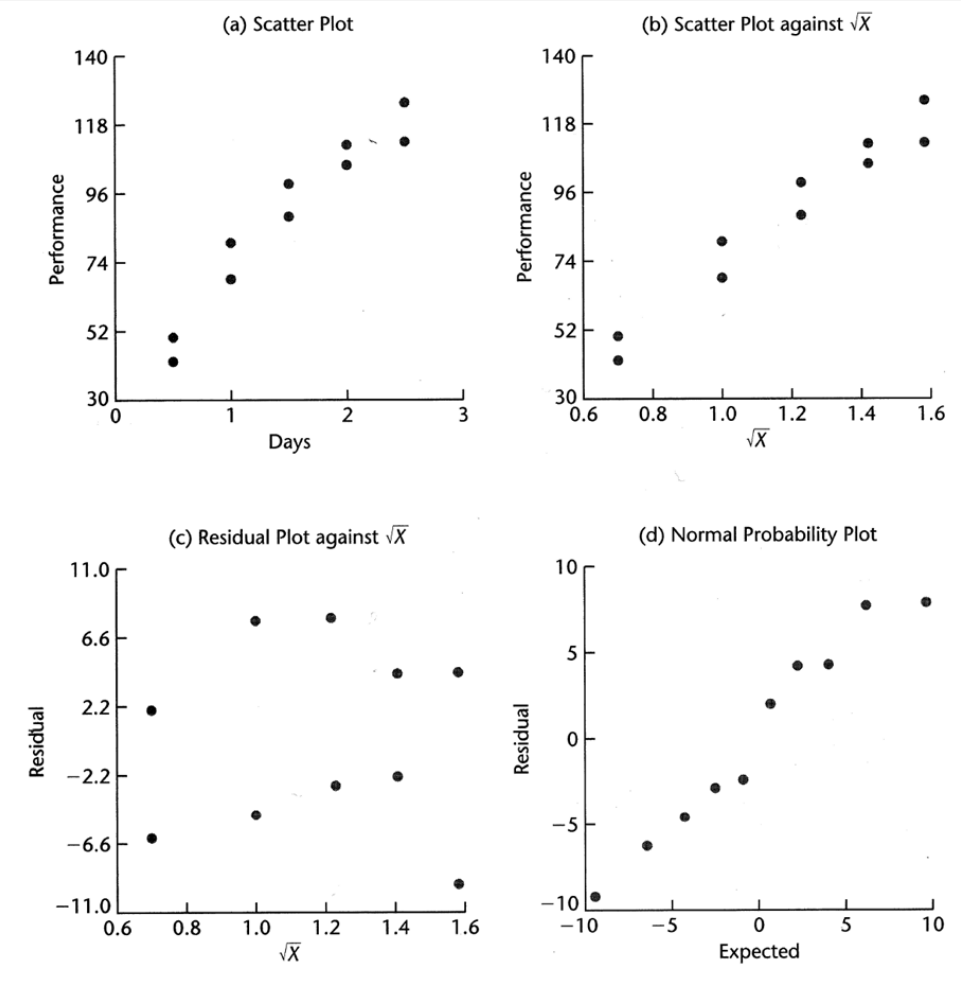
Example

Experiment

- X: days of training received
- Y: sales performance(score)

$X' = \sqrt{X}$

Figure :



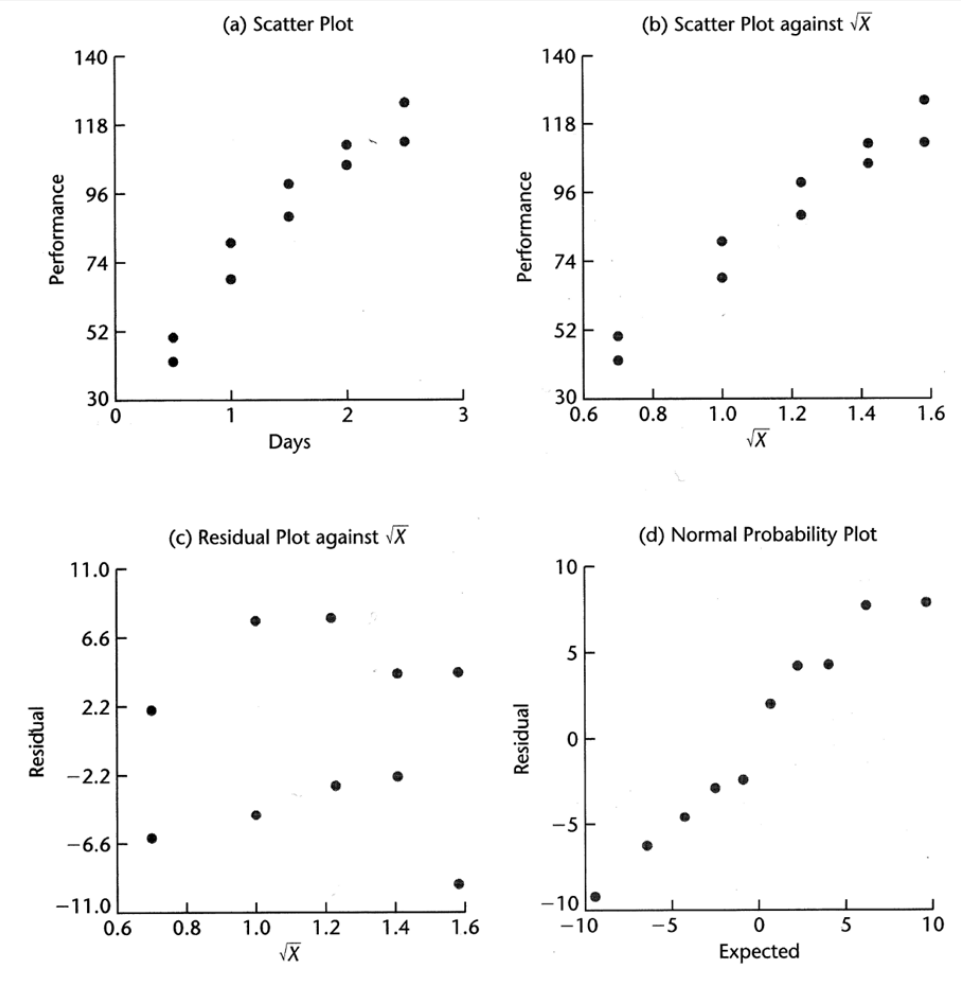
Example Data Transformation

Figure :

Sales Trainee i	(1) Days of Training X_i	(2) Performance Score Y_i	(3) $X_i' = \sqrt{X_i}$
1	.5	42.5	.70711
2	.5	50.6	.70711
3	1.0	68.5	1.00000
4	1.0	80.7	1.00000
5	1.5	89.0	1.22474
6	1.5	99.6	1.22474
7	2.0	105.3	1.41421
8	2.0	111.8	1.41421
9	2.5	112.3	1.58114
10	2.5	125.7	1.58114

Graphical Residual Analysis

Figure :

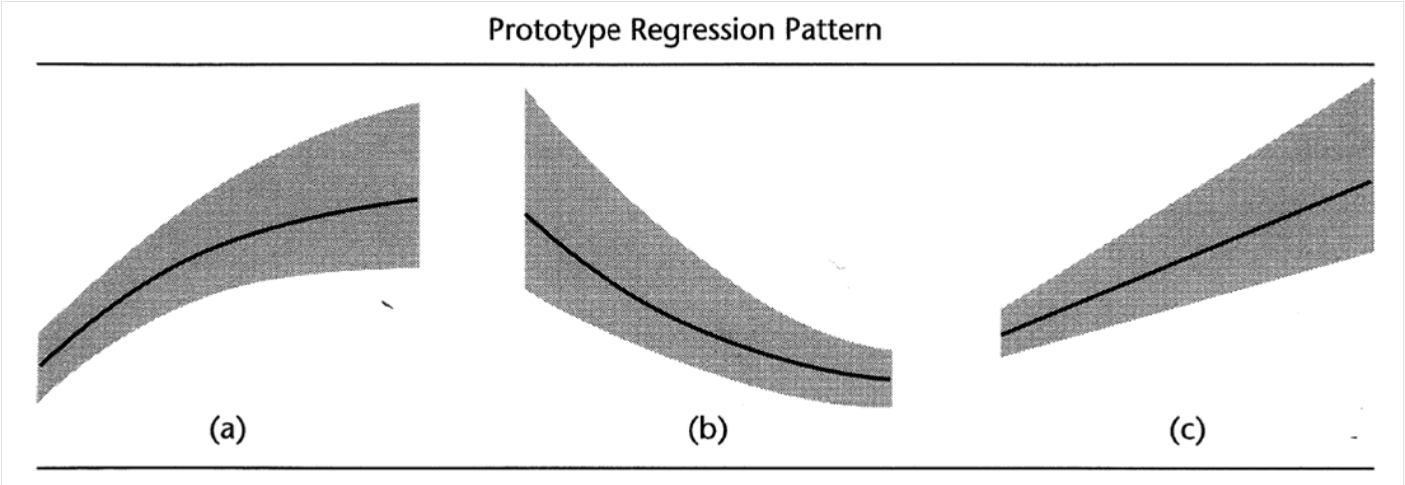


Transformations on Y

- Non-normality and unequal variances of error terms frequently appear together
- To remedy these in the normal regression model we need a transformation on Y
- This is because
  - Shapes and spreads of distributions of Y need to be changed
  - May help linearize a curvilinear regression relation
- Can be combined with transformation on X

Prototype Regression Patterns and Y Transformations

Figure :



Transformations on Y:

$y' = \sqrt{Y}$

$y' = \log_{10} Y$

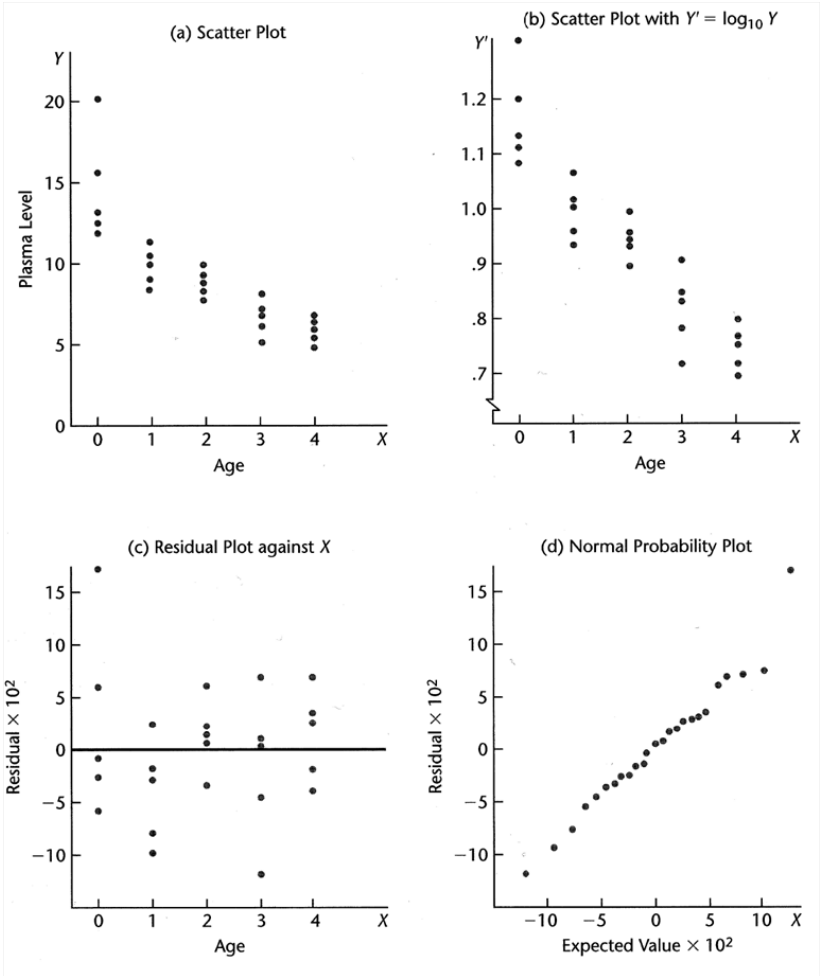
$y' = 1/Y$

Example

- Use of logarithmic transformation of Y to linearize regression relations and stabilize error variance
- Data on age(X) and plasma level of a polyamine (Y) for a portion of the 25 healthy children in a study. Younger children exhibit greater variability than older children.

Plasma Level vs. Age

Figure :



Associated Data

Figure :

Child i	(1) Age X_i	(2) Plasma Level Y_i	(3) $Y_i' = \log_{10} Y_i$
1	0 (newborn)	13.44	1.1284
2	0 (newborn)	12.84	1.1086
3	0 (newborn)	11.91	1.0759
4	0 (newborn)	20.09	1.3030
5	0 (newborn)	15.60	1.1931
6	1.0	10.11	1.0048
7	1.0	11.38	1.0561
...	...	...	...
19	3.0	6.90	.8388
20	3.0	6.77	.8306
21	4.0	4.86	.6866
22	4.0	5.10	.7076
23	4.0	5.67	.7536
24	4.0	5.75	.7597
25	4.0	6.23	.7945

Associated Data (Cont')

- if we fit a simple linear regression line to the log transformed Y data we obtain:

$$\hat{Y}' = 1.135 - .1023X$$

- And the coefficient of correlation between the ordered residuals and their expected values under normality is .981(for $\alpha = .05$ B.6 in the book shows a critical value of .959)
- Normality of error terms supported, regression model for transformed Y data appropriate.

Box Cox Transforms

- It can be difficult to graphically determine which transformation of Y is most appropriate for correcting
  - skewness of the distributions of error terms
  - unequal variances
  - nonlinearity of the regression function
- The Box-Cox procedure automatically identifies a transformation from the family of power transformations on Y

Box Cox Transforms

- This family is of the form

$$Y' = Y^{\lambda}$$

- Examples include

$$\lambda = 2 \quad Y' = Y^2$$
$$\lambda = .5 \quad Y' = \sqrt{Y}$$
$$\lambda = 0 \quad Y' = lnY \text{ (by definition)}$$
$$\lambda = -.5 \quad Y' = \frac{1}{\sqrt{Y}}$$
$$\lambda = -1 \quad Y' = \frac{1}{Y}$$

Box Cox Cont.

- The normal error regression model with the response variable a member of the family of power transformations becomes

$$Y_i^\lambda = \beta_0 + \beta_1 X_i + \epsilon_i$$

- This model has an additional parameter that needs to be estimated
- Maximum likelihood is a way to estimate this parameter

Box Cox Maximum Likelihood Estimation

- Before setting up MLE, the observations are further standardized so that the magnitude of the error sum of squares does not depend on the value of λ
- The transformation is give by

$$W_i = K_1(Y_i^\lambda - 1) \quad \lambda \neq 0$$
$$\text{or } K_2(log_e Y_i) \quad \lambda = 0$$

where

$$K_2 = (\prod Y_i)^{1/n}$$
$$K_1 = \frac{1}{\lambda K_2^{\lambda-1}}$$

Box Cox Maximum Likelihood Estimation

- Maximize

$$log(L(X, Y, \sigma, \lambda, b_1, b_0)) = -\sum_i \frac{(W_i - (b_1 X_i + b_0))^2}{2\sigma^2} - nlog(\sigma)$$

w.r.t λ σ b_1 b_0

- How?
  - Take partial derivatives
  - Solve
  - or... gradient ascent methods

Box Cox Maximum Likelihood Estimation

- Maximize

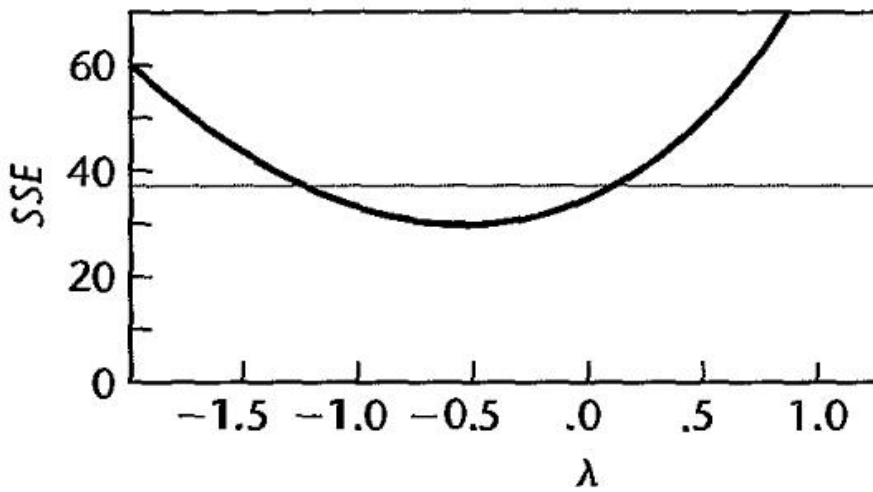
$$log(L(X, Y, \sigma, \lambda, b_1, b_0)) = -\sum_i \frac{(W_i - (b_1 X_i + b_0))^2}{2\sigma^2} - nlog(\sigma)$$

  w.r.t λ σ b_1 b_0

- How?
  - Take partial derivatives
  - Solve
  - or... gradient ascent methods
- Not easy task, so we work on a grid of λ values, say $\lambda = -2, -1.75, \ldots, 1.75, 2$. And calculate a list of MSE according to different λ values.

The plasma levels example

λ	SSE	λ	SSE
1.0	78.0	−.1	33.1
.9	70.4	−.3	31.2
.7	57.8	−.4	30.7
.5	48.4	−.5	30.6
.3	41.4	−.6	30.7
.1	36.4	−.7	31.1
0	34.5	−.9	32.7
		−1.0	33.9



Comments on Box Cox

- The Box-Cox procedure is ordinarily used only to provide a guide for selecting a transformation
- At time, theoretical considerations or prior information can be utilized to help in choosing an appropriate transformation
- It is important to perform residual analysis after the transformation to ensure the transformation is appropriate
- When transformed models are employed, b_0 and b_1 obtained via least squares have the least squares property w.r.t the transformed observations not the original ones.
- Usually take a nearby λ value for which the power transformation is easier to understand. Say $\hat{\lambda} = 0.03$, we may just use $\hat{\lambda} = 0$. Of course, one should examine the flatness of likelihood function in the neighborhood of $\hat{\lambda}$.
- When λ near 1, no transformation of Y needed.

Exploration of Shape of Regression Function: Nonparametric Regression Curves

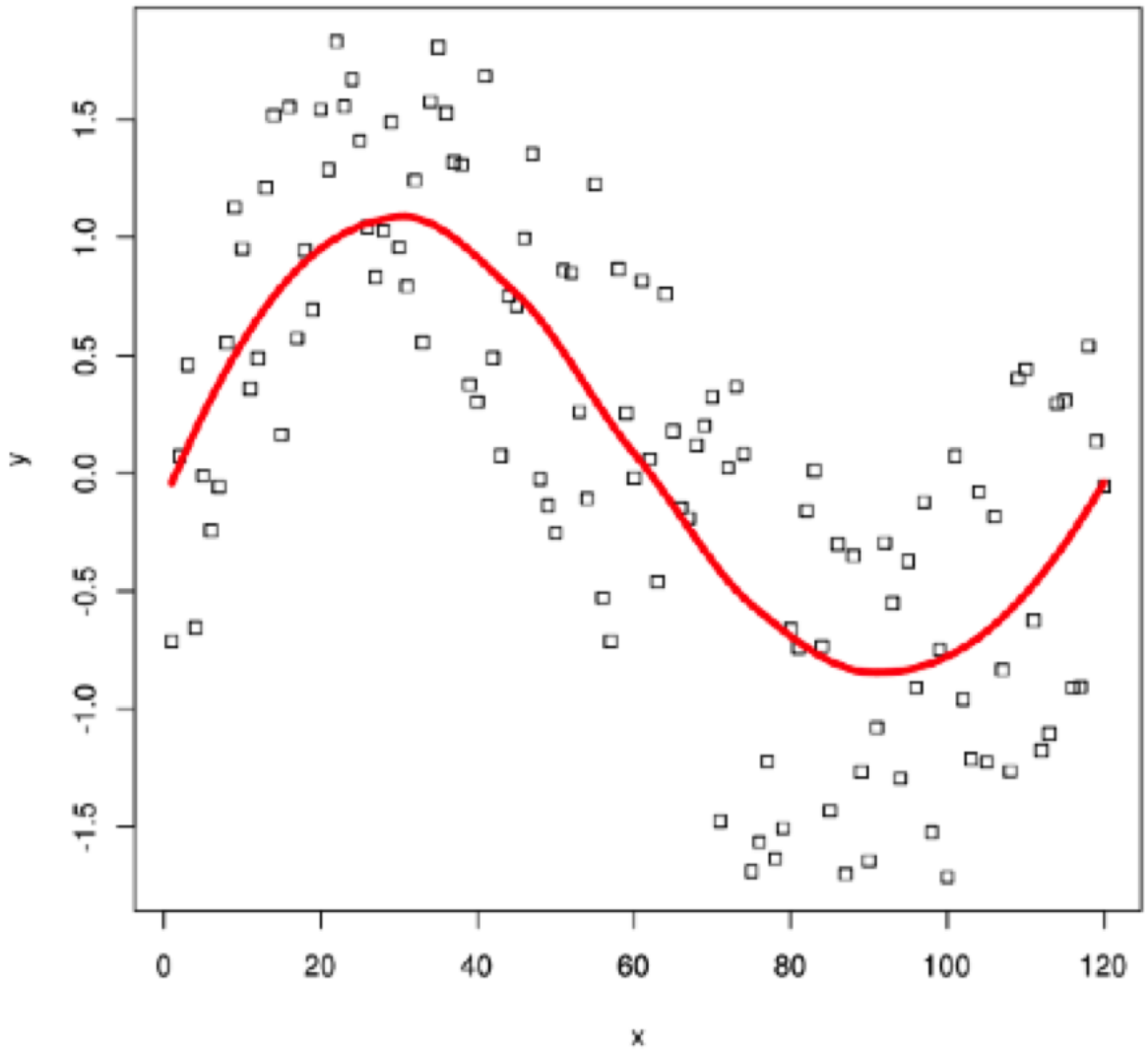
- So far: parametric regression approaches
  - Linear
  - Linear with transformed inputs and outputs
  - etc.
- Other approaches
  - Method of moving averages : interpolate between mean outputs at adjacent inputs
  - Lowess : "LOcally WEighted Scatterplot Smoothing"

Lowess Method

- Intuition
  - Fit low-order polynomial (linear) regression models to points in a neighborhood
    - The neighborhood size is a parameter
    - Determining the neighborhood is done via a nearest neighbors algorithm

    Produce predictions by weighting the regressors by how far the set of points used to produce the regressor is from the input point for which a prediction is wanted
- While somewhat ad-hoc, it is a method of producing a nonlinear regression function for data that might seem otherwise difficult to regress

Lowess Method Example



R example

```
require(graphics)

plot(cars, main = "lowess(cars)")
lines(lowess(cars), col = 2)
lines(lowess(cars, f=.2), col = 3)
legend(5, 120, c(paste("f = ", c("2/3", ".2"))),
lty = 1, col = 2:3)
```