Chapter 3

Diagnostics and Remedial Measures

Instructor: Li, Caixia

Outline

Diagnostics (using plots and tests)

- Diagnostics for prediction variable
- Diagnostics for residuals
  - L.I.N.E(Linearity; Independence; Normality; Equality of variance)
  - Outliers
  - Lack of fit

Remedial Measures

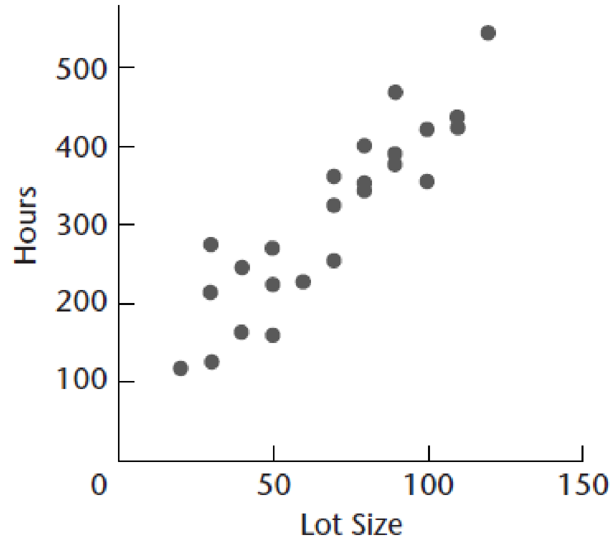
- Remedial action when violation of L.I.N.E.
- Presence of outliers
- Lack of fit

Diagnostics

- Procedures to determine appropriateness of the model and check assumptions used in the standard inference
- If there are violations, inference and model may not be reasonable thereby resulting in faulty conclusions
- Always check before any inference!
- Procedures involve both graphical methods and formal statistical tests

3.1 Diagnostics for Predictor Variables

- Scatterplot of Y vs X common diagnostic
  - Is linear trend reasonable?
  - Any unusual/influential (X, Y) observations?

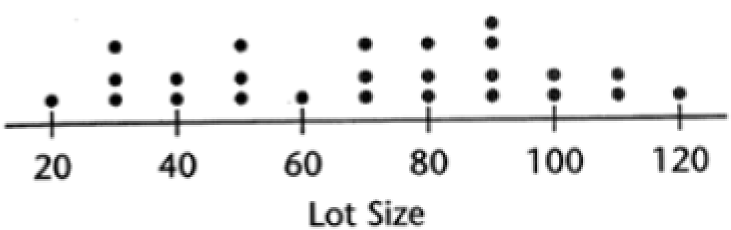


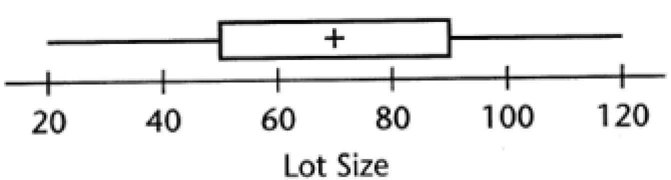
- Can also look at distribution of X alone
  - Unusual or outlying values?
  - Does X have pattern over time (order collected)?
- If Y depends on X, looking at Y alone may be deceiving (i.e.,mixture of normal dists)

Graphical diagnostics for X

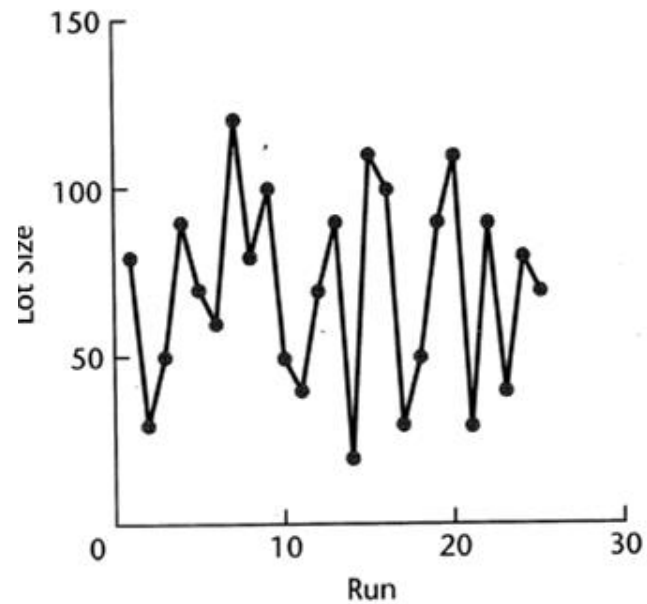
Useful plots of X levels

- Dot plot or bar plot for discrete variable
- Histogram or stem-and-leaf plot
- Box Plot
- Sequence Plot (X versus Run #)





2	0
3	000
4	00
5	000
6	0
7	000
8	000
9	0000
10	00
11	00
12	0



3.2 Residuals

- In a normal regression model the ε_i's are assumed to be i.i.d $N(0, \sigma^2)$ distributed.

$$\varepsilon_i = Y_i - E(Y_i) = Y_i - \left(\beta_0 + \beta_1 X_i\right) \quad \sim i.i.d.\ N(0, \sigma^2)$$

- Recall the definition of residuals:

$$e_i = Y_i - \hat{Y}_i = Y_i - \left(b_0 + b_1 X_i\right) = Y_i - \bar{Y} - b_1\left(X_i - \bar{X}\right) \quad i = 1, \ldots, n$$

- The properties of the residuals

(1) $\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} X_i e_i = \sum_{i=1}^{n} \hat{Y}_i e_i = 0$

(2) In normal error model, e_i's are normal distributed, but not independent. When n large, the dependency can be ignored.

$$e_i = Y_i - \hat{Y}_i \sim N\left(0, (1 - h_{ii})\sigma^2\right), \quad \operatorname{cov}(e_i, e_j) = -h_{ij}\sigma^2 \neq 0, \quad i \neq j$$

where $h_{ij} = \frac{1}{n} + \frac{\left(X_i - \bar{X}\right)\left(X_j - \bar{X}\right)}{SS_{XX}}$

- Proof

$$\operatorname{var}(e_i)=\operatorname{var}(Y_i)-\operatorname{var}(\bar{Y})-\operatorname{var}(b_1)\left(X_i-\bar{X}\right)^2-2\operatorname{cov}(Y_i,\bar{Y})-2\left(X_i-\bar{X}\right)\operatorname{cov}(Y_i,b_1)$$

$$=\sigma^2+\frac{\sigma^2}{n}+\frac{\left(X_i-\bar{X}\right)^2\sigma^2}{SS_{XX}}-\frac{2\sigma^2}{n}-\frac{2\left(X_i-\bar{X}\right)^2\sigma^2}{SS_{XX}}$$

$$=\sigma^2\left(1-\frac{1}{n}-\frac{\left(X_i-\bar{X}\right)^2}{SS_{XX}}\right)$$

$$\operatorname{cov}(e_i,e_j)=\operatorname{cov}\left(Y_i-\bar{Y}-b_1\left(X_i-\bar{X}\right),Y_j-\bar{Y}-b_1\left(X_j-\bar{X}\right)\right)$$

$$=\operatorname{cov}\left(Y_i-\bar{Y},Y_j-\bar{Y}\right)-\left(X_j-\bar{X}\right)\operatorname{cov}\left(Y_i,b_1\right)-\left(X_i-\bar{X}\right)\operatorname{cov}\left(Y_j,b_1\right)+\left(X_i-\bar{X}\right)\left(X_j-\bar{X}\right)\operatorname{var}(b_1)$$

$$=-\frac{\sigma^2}{n}-2\frac{\left(X_i-\bar{X}\right)\left(X_j-\bar{X}\right)\sigma^2}{SS_{XX}}+\frac{\left(X_i-\bar{X}\right)\left(X_j-\bar{X}\right)\sigma^2}{SS_{XX}}$$

$$=-\sigma^2\left(\frac{1}{n}+\frac{\left(X_i-\bar{X}\right)\left(X_j-\bar{X}\right)}{SS_{XX}}\right)$$

Semi-studentized residuals

Actually, $$e_i \sim N\left(0, \sigma^2\left[1-\left(\frac{1}{n}+\frac{(X_i-\bar{X})^2}{\sum_{j=1}^{n}(X_j-\bar{X})^2}\right)\right]\right)$$

- It may be useful sometimes to look at a standardized set of residuals, for instance in outlier detection.
- Studentized residual

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

- Semi-studentized residual.

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

Departures from Model…

To be studied by residuals

- Regression function not linear **(L)**
- Error terms are not independent **(I)**
- Error terms are not normally distributed **(N)**
- Error terms do not have equal variance **(E)**

- Model fits all but one or a few **outlier** observations
- One or more **predictor** variables have been **omitted** from the model

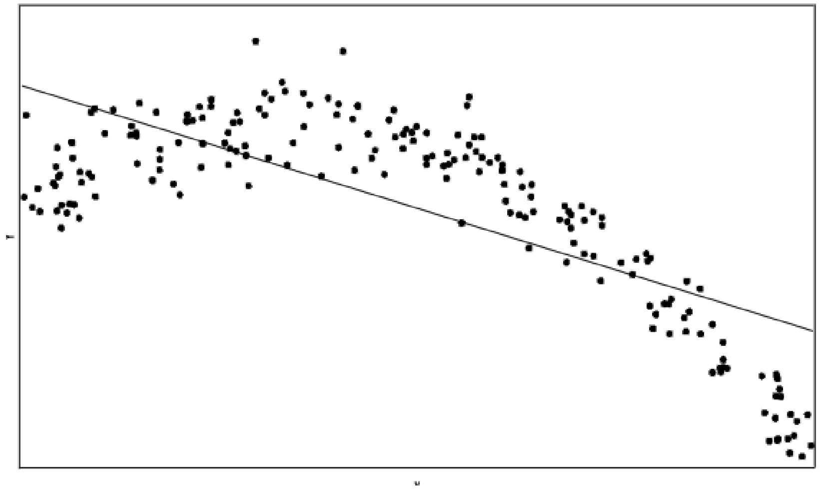
3.3 Diagnostics for Residuals

Common Plots

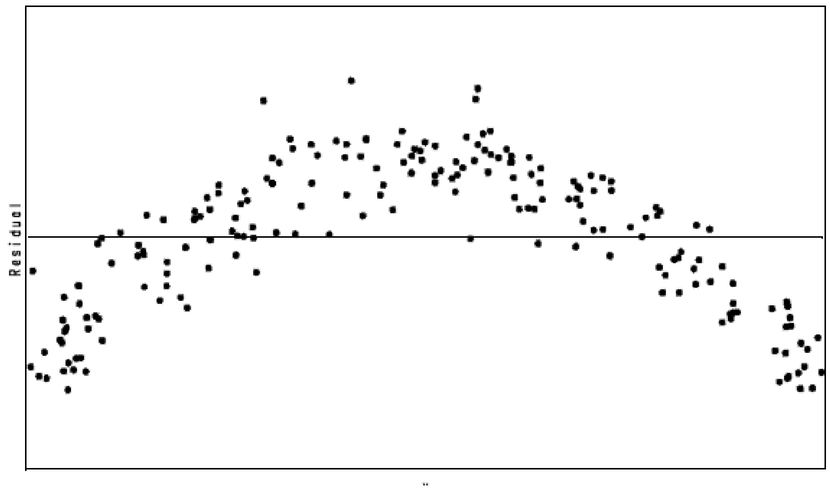
- Residuals/Absolute Residuals versus Predictor Variable
- Residuals/Absolute Residuals versus Predicted Values
- Residuals versus Omitted variables
- Residuals versus Time
- Box Plots, Histograms, Normal Probability Plots

Nonlinearity of Regression Function

- Plot Y versus X
- Plot Residuals versus X
  - Random Cloud around regression line/0 $\Rightarrow$ Linear Relation
  - U-Shape or Inverted U-Shape $\Rightarrow$ Nonlinear Relation



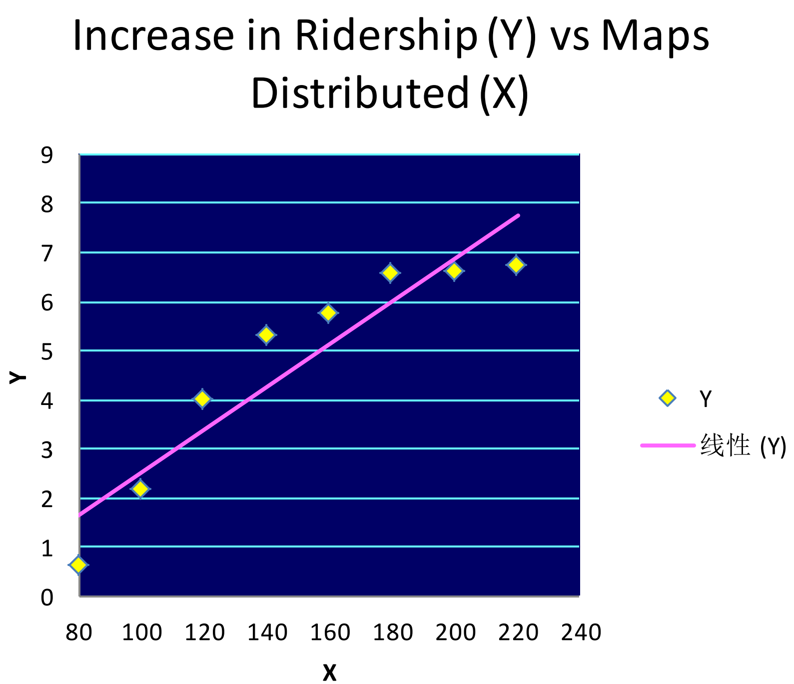
Y vs **X**

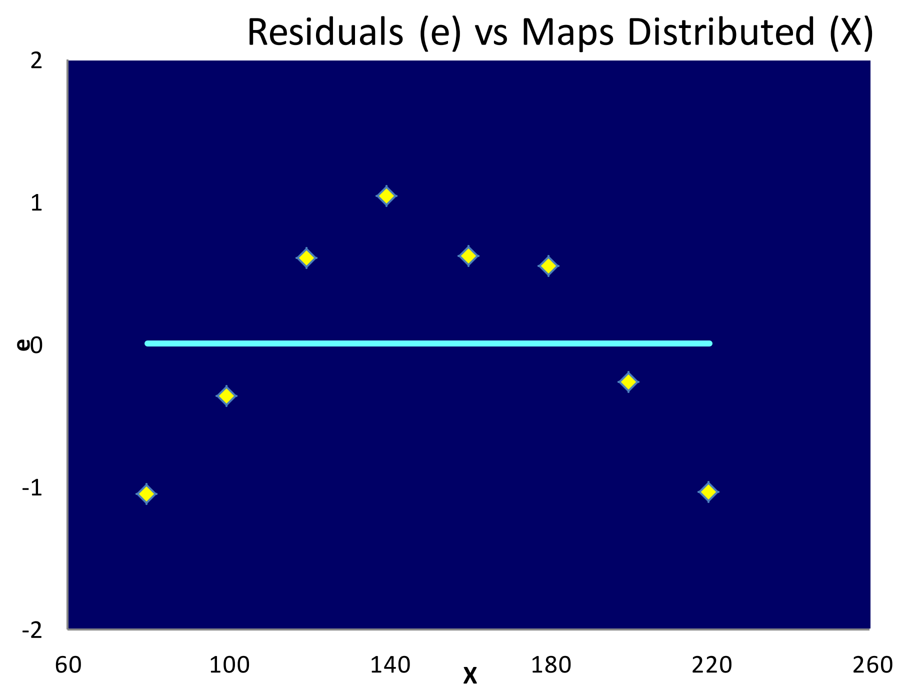


Residual vs **X**

Nonlinearity of Regression Function

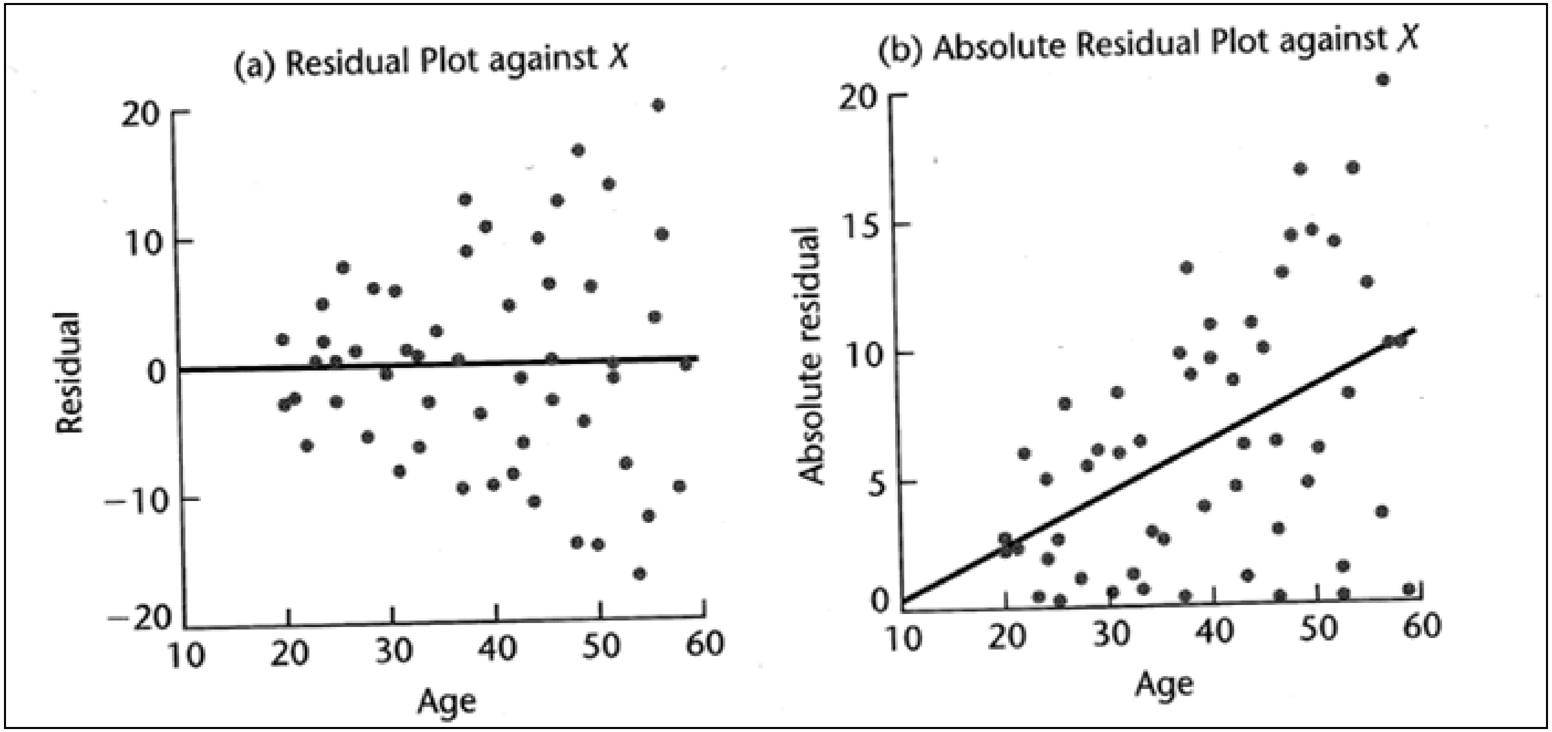
- Transit example : ridership increase vs. num. maps distributed (Table 3.1,Figure 3.5, p.10)





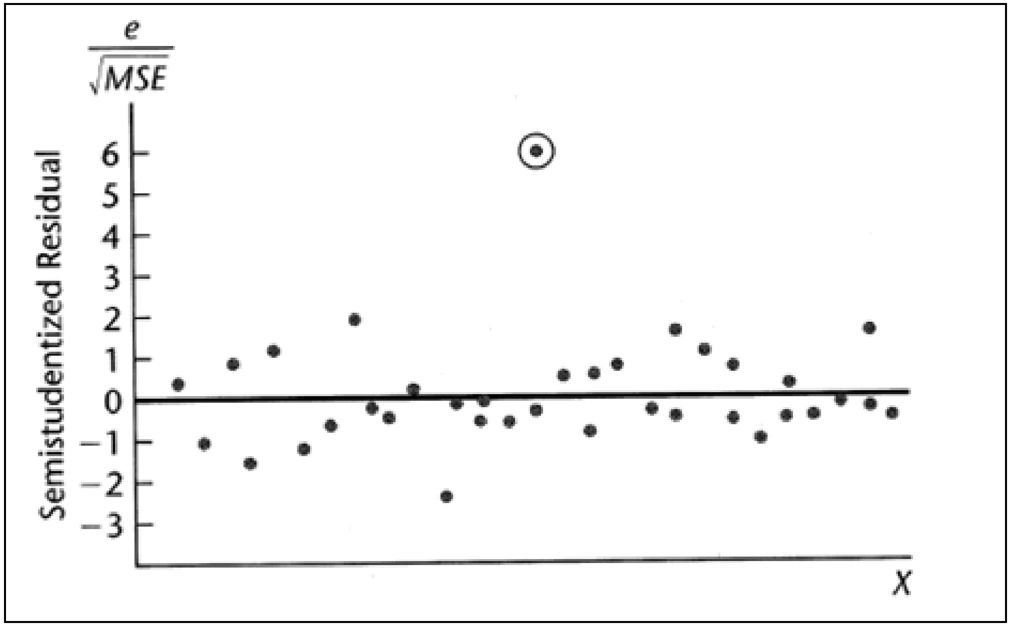
Nonconstancy of Error Variance

- Plot Residuals versus X or Predicted Values
  - Funnel Shape ⇒ Non-constant Variance
- Plot absolute Residuals/squared residuals
  - Positive Association ⇒ Non-constant Variance



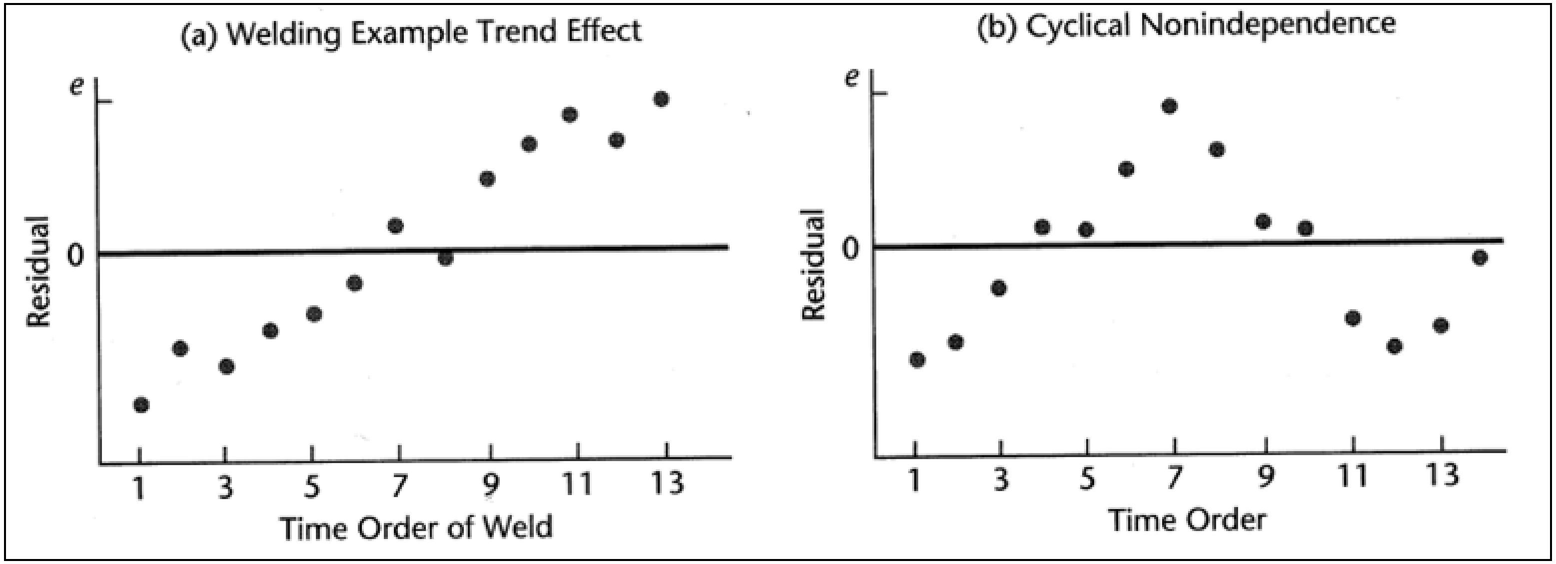
Presence of Outliers

- Outliers can strongly effect the fitted values of the regression line.
- Rule of thumb: say it is an outlier, if $e_i^* = \dfrac{e_i}{\sqrt{MSE}} > 4$



Non-independence of Error Terms

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



Linear Trend

Cyclical Trend

Non-Normal Errors

- Distribution plots of residuals
  - e.g., boxplot – Can confirm symmetry and lack of outliers
- Check Proportion that lie within 1 standard deviation from 0, 2 SD, etc, where SD=sqrt(MSE)
- Normal probability plot of residual
  - Q-Q plot– should fall approximately on a straight line (Only works well with moderate to large samples)
  - **qqnorm(e); qqline(e)** in R

Normal Quantile-Quantile(Q-Q) Plot

Step 1. First sort the sample data by arranging the values in order from lowest to highest. $e_{(1)}, e_{(2)}, \ldots, e_{(n)}$

Step 2. With a sample of size n, each $e_{(i)}$ represents a sample quantile corresponding to a proportion p_i. Roughly speaking, we expect the first ordered value to be in the interval (0, 1/n), the second to be in the interval (1/n, 2/n), and the last to be in of the interval ((n−1)/n, 1).

$p_i = (i - a)/(n + 1 - 2a)$, a in the range from 0 to ½

e.g. $a=0$, $p_i = i/(n + 1)$; $a=1/2$, $p_i = (i - 1/2)/n$

$a=3/8$, $p_i = (i - 3/8)/(n + 1/4)$.

In *R*, $a=3/8$ if $n \le 10$ and $a=1/2$ if $n > 10$.

Step 3. Use the standard normal distribution to find the theoretical quantile q_i corresponding to p_i in step 2.

$$q_i = \Phi^{-1}(p_i)$$

Step 4. plot the points $(q_i, e_{(i)})$, where each $e_{(i)}$ is a sample quantile and q_i is the theoretical quantile.

Step 5. Examine the normal quantile plot and determine whether or not the distribution is normal.

Example 随机选取10个零件，测得其直径与标准尺寸的偏差如下：（单位：丝）

9.4 8.8 9.6 10.2 10.1 7.2 11.1 8.2 8.6 9.6

Q-Q图步骤如下：

(1) 首先将数据排序：

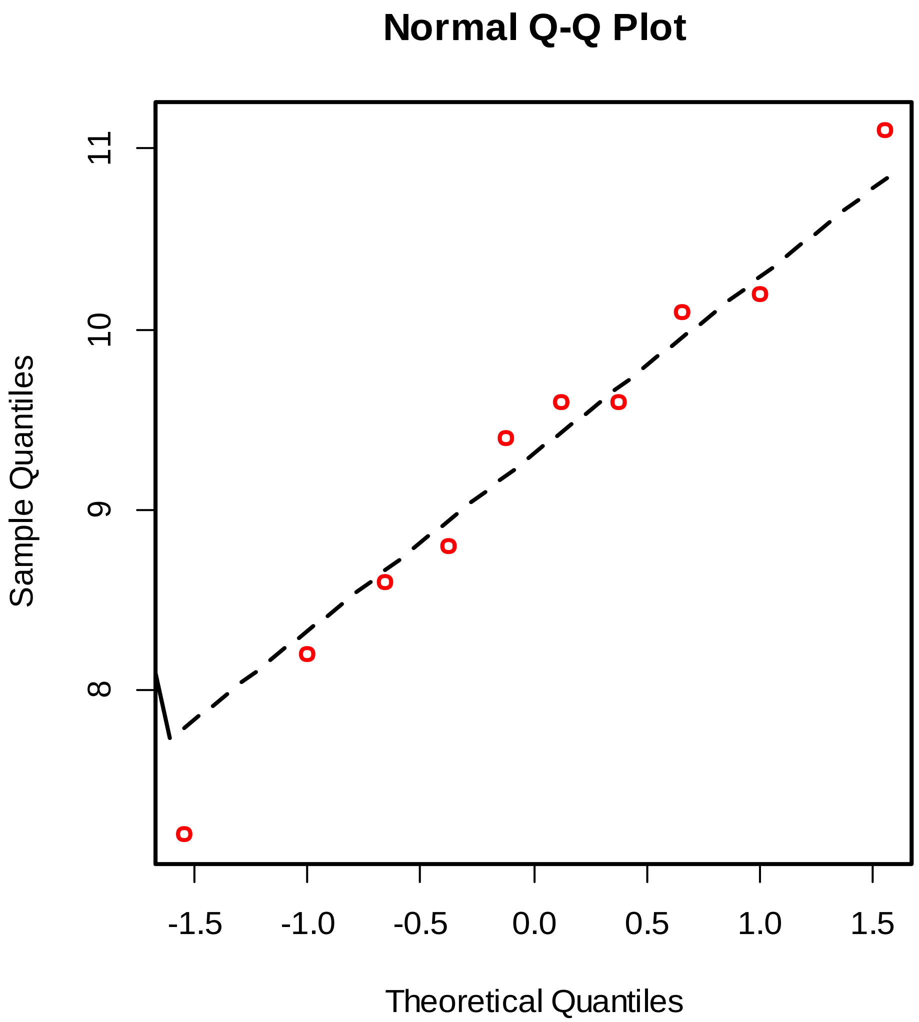
7.2 8.2 8.6 8.8 9.4 9.6 9.8 10.1 10.2 11.1;

(2) 对每一个i，计算a=3/8=0.375对应的修正频率 $p_i=(i-0.375)/(n+0.25)$, $i=1,2,\ldots,n$；

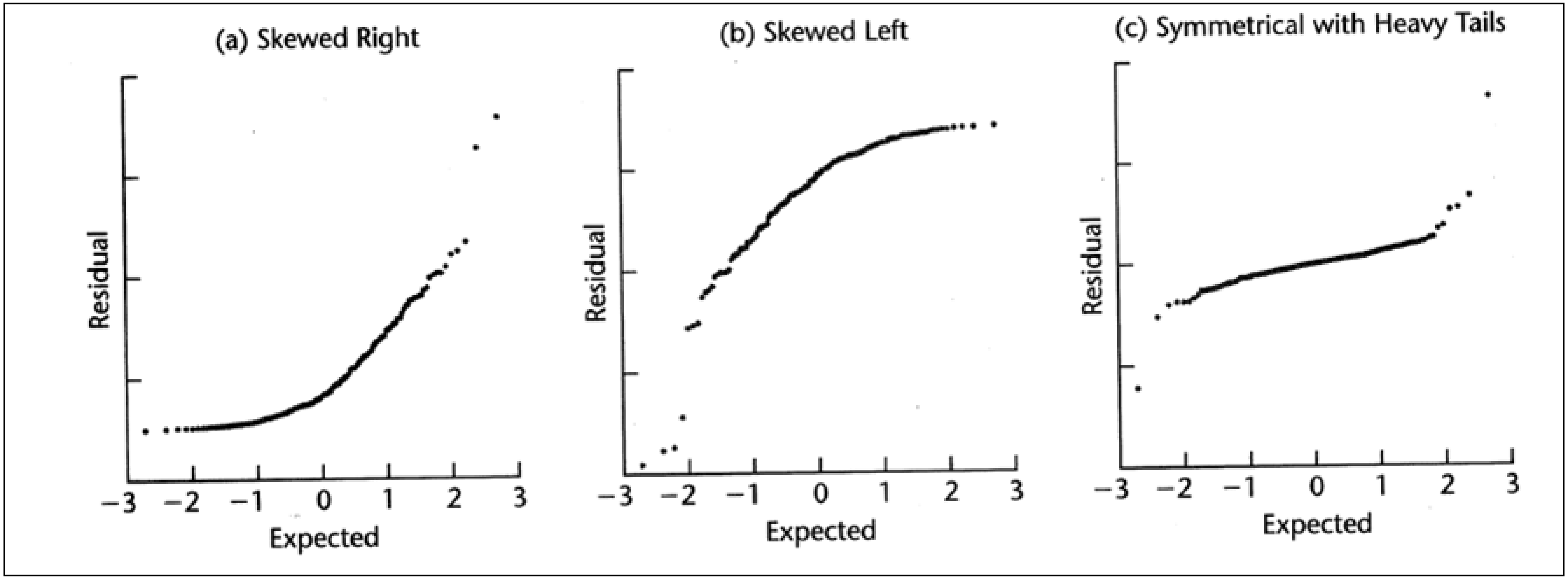
(3) 对每一个i，计算p_i对应的理论分位数 $q_i=\Phi^{-1}(p_i)$, $i=1,2,\ldots,n$；

(4) 将二维坐标系中绘出n个点$(q_i, x_{(i)})$, $i=1,2,\ldots,n$

rank	$x_{(i)}$	p_i	q_i
1	7.2	0.061	-1.547
2	8.2	0.159	-1.000
3	8.6	0.256	-0.655
4	8.8	0.354	-0.375
5	9.4	0.451	-0.123
6	9.6	0.549	0.123
7	9.6	0.646	0.375
8	10.1	0.744	0.655
9	10.2	0.841	1.000
10	11.1	0.939	1.547

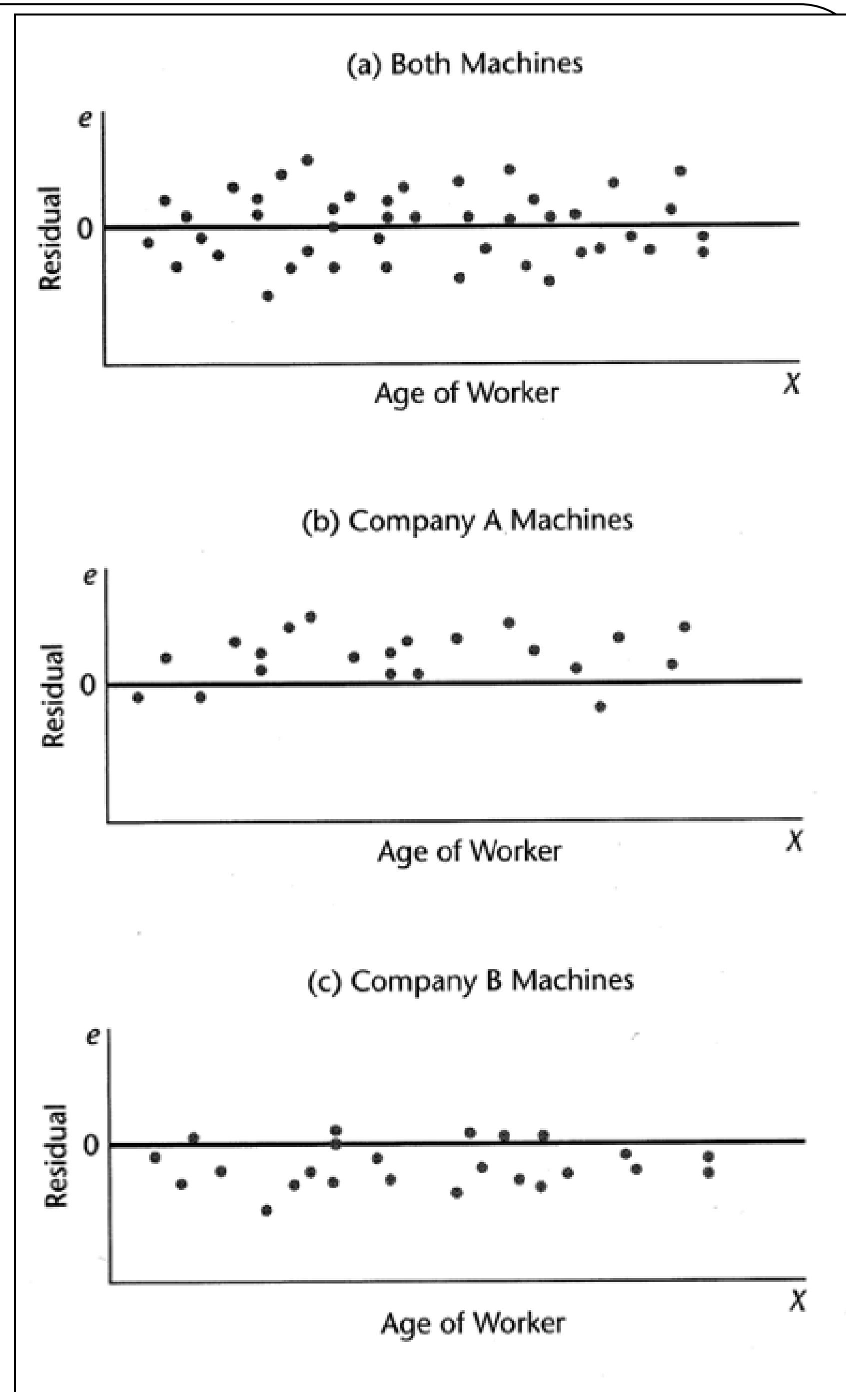


- Figure : Examples of non-normality in distribution of error terms



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



3.4 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)

3.5 Tests for Normality of Residuals

- Correlation Test
  1) Obtain correlation between observed residuals and expected values under normality
  2) Compare correlation with critical value based on α-level from Table B.6, page 1329. A good approximation for the α=0.05 critical value is: $1.02-1/\mathrm{sqrt}(10n)$
  3) Reject the null hypothesis of normal errors if the correlation falls below the table value

- Shapiro-Wilk Test – Performed by most software packages. Related to correlation test, but more complex calculations

3.6 Tests for Constancy of Error Variance

1. Brown-Forsythe test for constant variance. (It is applicable when the variance increasing or decreasing in X)

- Divide dataset into 2 groups based on levels of with sample size n_1, n_2. Compute $d_{ij} = \left| e_{ij} - \tilde{e}_j \right| \quad i = 1, \dots, n_j \quad j = 1, 2$
- The test statistic for comparing the means of the absolute deviations of the residuals around the group medians

$$t_{BF} = \frac{\bar{d}_1 - \bar{d}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \overset{H_0}{\sim} t\left(n_1 + n_2 - 2\right), \text{approximately.}$$

$$s^2 = \frac{\left(n_1 - 1\right) s_1^2 + \left(n_2 - 1\right) s_2^2}{n_1 + n_2 - 2}$$

where $\bar{d}_1, s_1^2, \bar{d}_2, s_2^2$ are the mean and variance for each group of d_{ij}.

Tests for Equal Variance - II

2. Breusch-Pagan (aka Cook-Weisberg) Test:

H_0: Equal Variance Among Errors $\sigma^2\{\varepsilon_i\} = \sigma^2 \ \forall\, i$

H_A: Unequal Variance Among Errors $\sigma^2\{\varepsilon_i\} = \sigma^2 h\left(\gamma_1 X_{i1} + \ldots + \gamma_p X_{ip}\right)$

1) Let $SSE = \sum_{i=1}^{n} e_i^2$ from original regression

2) Fit Regression of e_i^2 on $X_{i1}, \ldots X_{ip}$ and obtain $SS(\text{Reg}^*)$

Test Statistic: $X_{BP}^2 = \dfrac{SS(\text{Reg}^*)/2}{(SSE/n)^2} \overset{H_0}{\sim} \chi_p^2$, approximately

Reject H_0 if $X_{BP}^2 \geq \chi^2(1-\alpha; p)$ $\qquad p$ = # of predictors

3.7 F test for lack of fit

- Formal test for determining whether a specific type of regression function adequately fits the data.

- Test for linearity of regression

  Assumes $Y|X \overset{ind}{\sim} N(\mu(X), \sigma^2)$

  $H_0 : \mu(X) = \beta_0 + \beta_1 X$ (Reduced model)

  $H_a : \mu(X) \neq \beta_0 + \beta_1 X$ (Full model)

- Will use full/reduced model framework

Test for Linearity

$$H_0 : E(Y_i) = \beta_0 + \beta_1 X_i$$

$$H_A : E(Y_i) = \mu_i \neq \beta_0 + \beta_1 X_i$$

Requires: repeat observations at one or more X levels (called replicates)

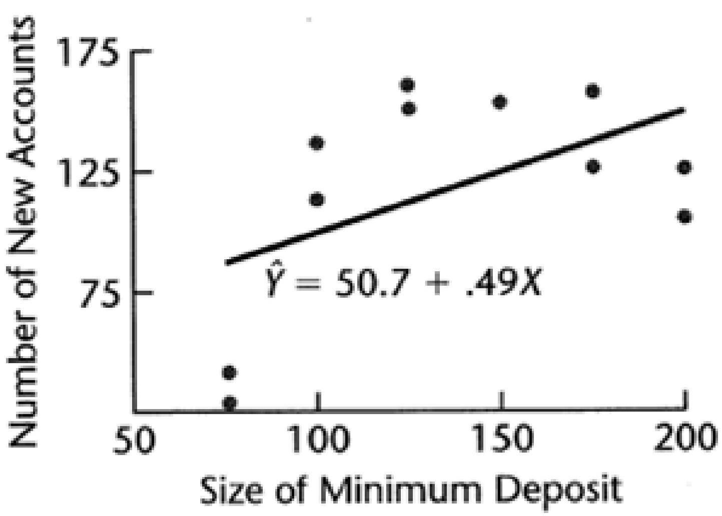
Branch i	Size of Minimum Deposit (dollars) X_i	Number of New Accounts Y_i
1	125	160
2	100	112
3	200	124
4	75	28
5	150	152
6	175	156
7	75	42
8	175	124
9	125	150
10	200	104
11	100	136

- Notation
  - Define X levels as $X_1, X_2, \ldots, X_c$
  - There are n_j replicates at level X_j $(\sum n_j = n)$
  - Y_{ij} is the i^{th} replicate at X_j

	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

Bank example

	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



(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	

Test for Linearity

- The SSE for the reduced model is as before. $\hat{Y}_{ij} = b_0 + b_1 X_{ij}$

$$SSE(R) = \sum_{j=1}^{c} \sum_{i=1}^{n_j} \left(Y_{ij} - \hat{Y}_{ij} \right)^2 \quad df_R = n - 2$$

- In full model, there are c parameters $\hat{\mu}_j = \overline{Y}_j$,

$$SSE(F) = \sum_{j=1}^{c} \sum_{i=1}^{n_j} \left(Y_{ij} - \overline{Y}_j \right)^2 \quad df_F = n - c$$

- Test statistic

$$F^* = \frac{\left[\left(SSE(R) - SSE(F) \right) / \left(df_R - df_F \right) \right]}{\left[SSE(F) / df_F \right]}$$

- Decision rule:
  - Reject H_0 if $F^* \geq F\left(1 - \alpha; df_R - df_F, df_F\right)$.

Bank example

- $SSE(R)=14716.6, \quad df_R=11-2=9$
- $SSE(F)=1148.0, \quad df_F=11-6=5$
- $SSLF=SSE(R)-SSE(F)=13593.6, \quad df_{LF}=4$

$$F^* = \frac{13{,}593.6}{4} \div \frac{1{,}148.0}{5}$$
$$= \frac{3{,}398.4}{229.6} = 14.80 \quad > F\left(0.95;4,5\right)$$

- Decision rule: Reject H_0

3.8 Overview of Remedial Measures

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

- Abandon simple regression model and develop and use a more appropriate model, e.g., generalized linear regression, nonparametric regression, etc…
  - may yield better insights, but a more complex model lead to more complex procedures for estimating the parameters.

- Employ some transformation of the data so that the simple regression model is appropriate for the transformed data. (This chapter)

Remedial Measures

- Nonlinearity of regression function - Transformation(s) or nonlinear regression(Chapter13)
- Nonconstancy of error variance - Weighted least squares (Chapter 11) and transformations
- Non-independence of error terms - Directly model correlation or use first differences (Chapter 12)
- Non-normality of error terms - Transformation(s) or fit Generalized Linear Model(Chapter 14)
- Omission of Important Predictor Variables - Include important predictors in a multiple regression model (Chapter 6 and later on)
- Outlying observations - Robust regression (Chapter 11)

Nonlinear Relationships

- Can model many nonlinear relationships with linear models, some with several explanatory variables

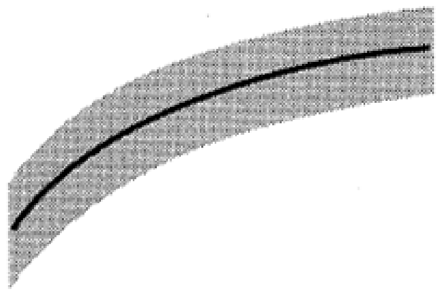
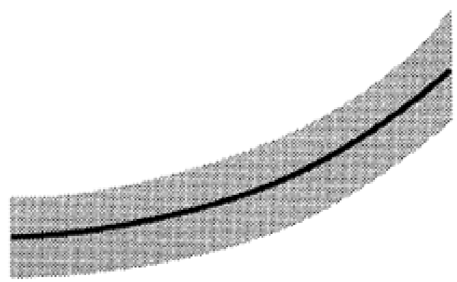
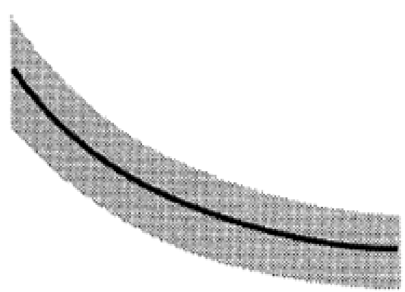
$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i \\ Y_i &= \beta_0 + \beta_1 \log(X_i) + \varepsilon_i \end{aligned}$$

- Can sometimes transform nonlinear model into a linear model

$$\begin{aligned} Y_i &= \beta_0 \exp(\beta_1 X_i)\varepsilon_i \\ &\downarrow \\ \log(Y_i) &= \log(\beta_0) + \beta_1 X_i + \log(\varepsilon_i) \end{aligned}$$

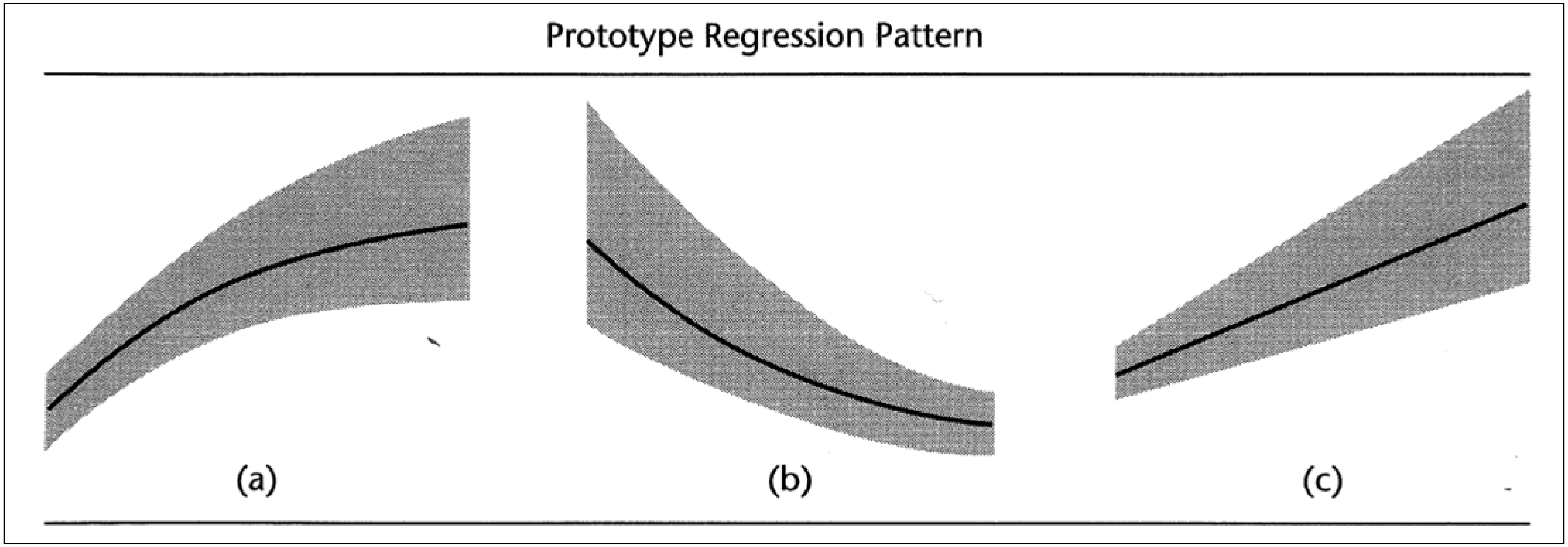
  - Have altered our assumptions about error

Transformation on X

	Prototype Regression Pattern	Transformations of X
(a)		$X' = \log_{10} X$ $\quad X' = \sqrt{X}$
(b)		$X' = X^2$ $\quad X' = \exp(X)$
(c)		$X' = 1/X$ $\quad X' = \exp(-X)$

Transformations on Y

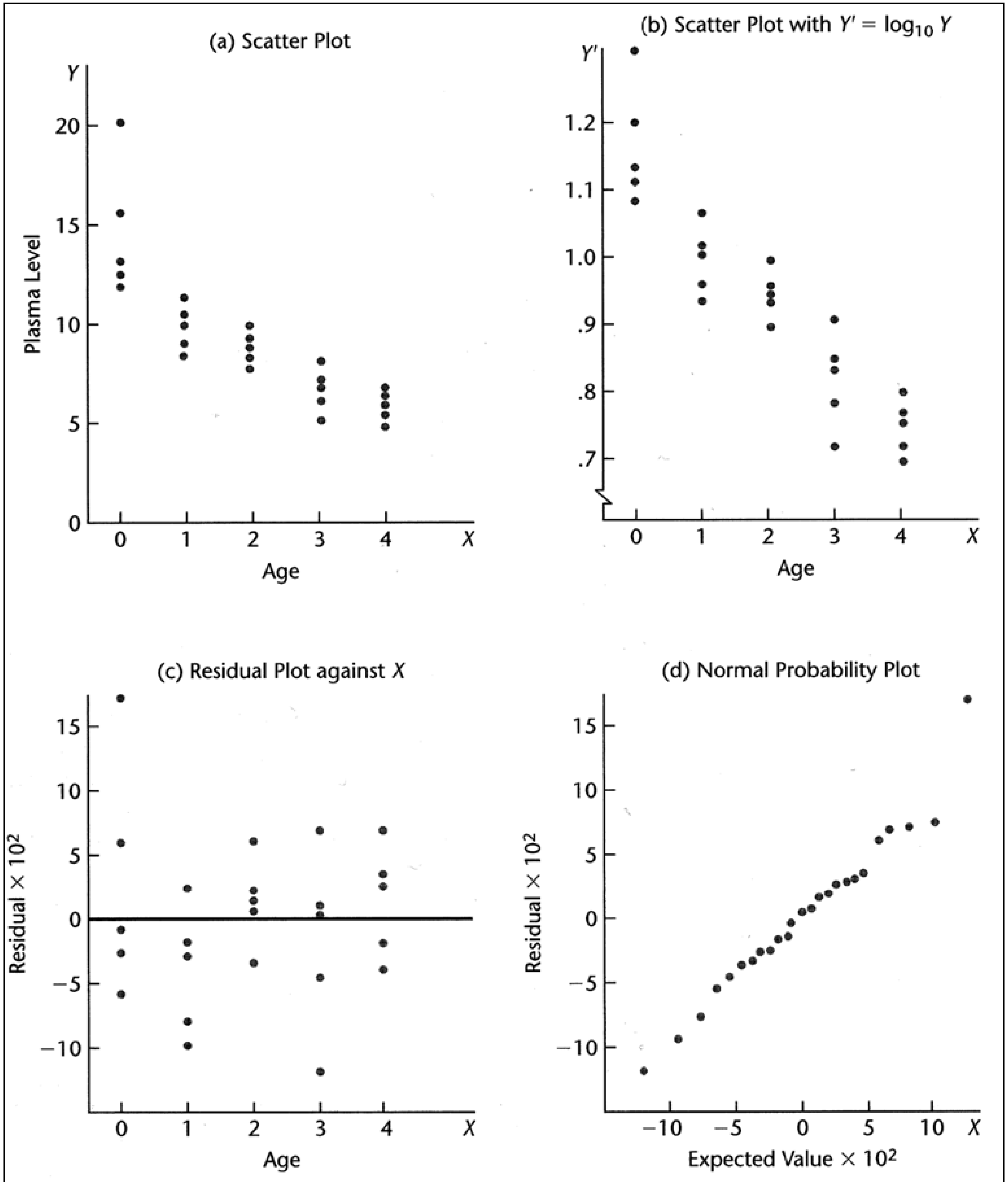
- Nonconstancy of error variance--Transformations on Y
- Can be combined with transformation on X



$$y' = \sqrt{Y}$$
$$y' = \log_{10} Y$$
$$y' = 1/Y$$

Plasma example

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



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

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

$$y_i^{(\lambda)} = \begin{cases} \dfrac{y_i^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \ln(y_i) & \text{if } \lambda = 0, \end{cases}$$

- Maximum likelihood is a way to estimate λ.

Box Cox Transforms

- If you take least square criterion, care must be taken because the variance of the residuals is not comparable as λ varies.
- The observations are first standardized

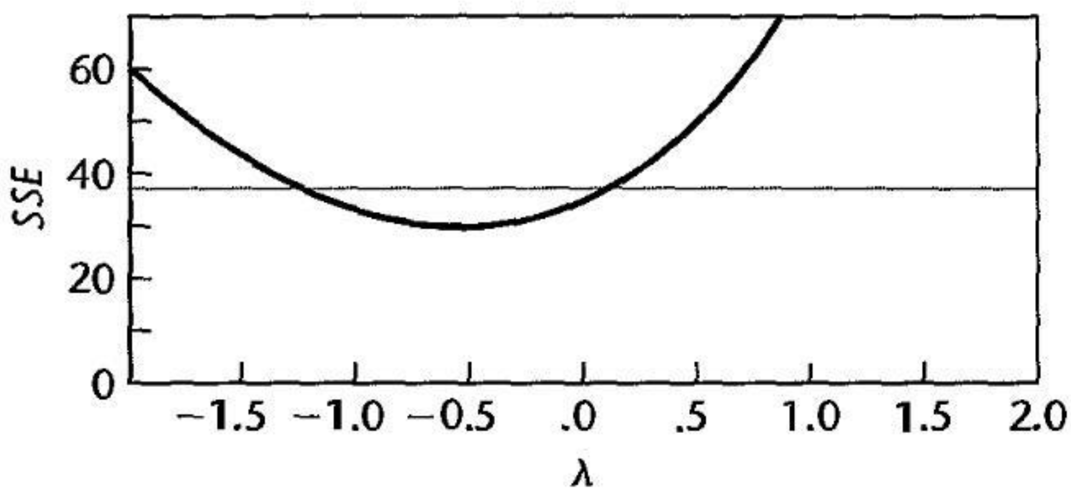
$$y_i^{(\lambda)} = \begin{cases} \dfrac{y_i^\lambda - 1}{\lambda(\mathrm{GM}(y))^{\lambda-1}}, & \text{if } \lambda \neq 0 \\ \mathrm{GM}(y) \ln y_i, & \text{if } \lambda = 0 \end{cases}$$

where GM(y) represent the *geometric mean* of $y_1, y_2, \ldots y_n$

- The linear regression of $y_i^{(\lambda)}$ on X_i is fitted.
  - work on a grid of λ values, say $\lambda = -2, -1{:}75, \ldots, 1{:}75, 2$, and construct regression models and calculate a list of SSE according to different 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



Homework

- Page 148: 3.5 (e),(f),(g)
- Page 150: 3.15; 3.16 (b) (c) (e) (f)

R code

```
#### Plots for predictor x
toluca = read.table("D:\\Data_4e\\CH01TA01.txt",header=F)
x = toluca[,1]
y = toluca[,2]
library(graphics)
boxplot(x,horizontal =T)
stem(x,scale=3)
hist(x)
```

R code

```
####Diagnostic plots and tests for residual
fit = lm(y~x)
resi = fit$residuals
yfit = predict(fit)
plot(x, fit$resi,xlab="x",ylab="Residual")
plot(yfit, fit$resi,xlab="Fitted y",ylab="Residual")
qqnorm(resi)
qqline(resi)
shapiro.test(resi)  ##Shapiro-Wilk Normality Test
```

R code

```
####Brown-Forsythe Test using Toluca example
ind1 = which(x<80);   ind2 = which(x>=80)
resi1 = resi[ind1];   resi2 = resi[ind2]
d1 = abs(resi1-median(resi1))
d2 = abs(resi2-median(resi2))
t.test(d1,d2)
####Lack of Fit Test using Bank example
data = read.table(' CH03TA04.txt',header=F)
y = data[,2];  x = data[,1]
Reduced=lm(y~x)
Full=lm(y~as.factor(x)-1)
anova(Reduced, Full)
```

R code

```
####Box-Cox transformation using Plasma example
library(MASS)
data = read.table(' CH03TA08.txt',header=F)
y = data[,2];  x = data[,1]
fit = lm(y~x)
a = boxcox(fit)
a$x[which.max(a$y)] ##best transformation parameter
```