Simple Linear Regression Models

Raj Jain
Washington University in Saint Louis
Saint Louis, MO 63130
Jain@cse.wustl.edu

These slides are available on-line at:

http://www.cse.wustl.edu/~jain/cse567-15/

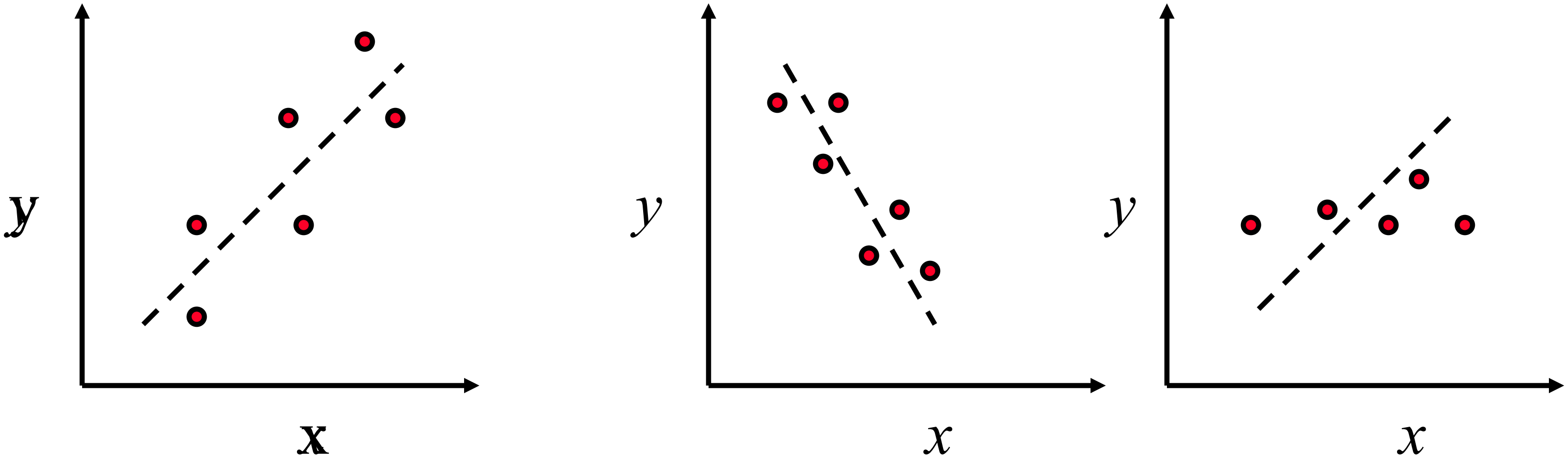
Overview

1. Definition of a Good Model
2. Estimation of Model parameters
3. Allocation of Variation
4. Standard deviation of Errors
5. Confidence Intervals for Regression Parameters
6. Confidence Intervals for Predictions
7. Visual Tests for verifying Regression Assumption

Simple Linear Regression Models

- **Regression Model**: Predict a response for a given set of predictor variables.
- **Response Variable**: Estimated variable
- **Predictor Variables**: Variables used to predict the response. predictors or factors
- **Linear Regression Models**: Response is a linear function of predictors.
- **Simple Linear Regression Models**: Only one predictor

Definition of a Good Model



Good Good Bad

Good Model (Cont)

- Regression models attempt to minimize the distance measured vertically between the observation point and the model line (or curve).
- The length of the line segment is called residual, modeling error, or simply error.
- The negative and positive errors should cancel out
  ⇒ Zero overall error
  Many lines will satisfy this criterion.

Good Model (Cont)

- Choose the line that minimizes the sum of squares of the errors.

$$\hat{y} = b_0 + b_1 x$$

where, $\hat{y}$ is the predicted response when the predictor variable is x. The parameter b_0 and b_1 are fixed regression parameters to be determined from the data.

- Given n observation pairs $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, the estimated response $\hat{y}_i$ for the ith observation is:

$$\hat{y}_i = b_0 + b_1 x_i$$

- The error is:

$$e_i = y_i - \hat{y}_i$$

Good Model (Cont)

- The best linear model minimizes the sum of squared errors (SSE):

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2$$

subject to the constraint that the mean error is zero:

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n}(y_i - b_0 - b_1 x_i) = 0$$

- This is equivalent to minimizing the variance of errors (see Exercise).

Estimation of Model Parameters

- Regression parameters that give minimum error variance are:

$$b_1 = \frac{\Sigma xy - n\bar{x}\bar{y}}{\Sigma x^2 - n\bar{x}^2} \quad \text{and} \quad b_0 = \bar{y} - b_1\bar{x}$$

- where,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$\Sigma xy = \sum_{i=1}^{n} x_i y_i \qquad \Sigma x^2 = \sum_{i=1}^{n} x_i^2$$

Example 14.1

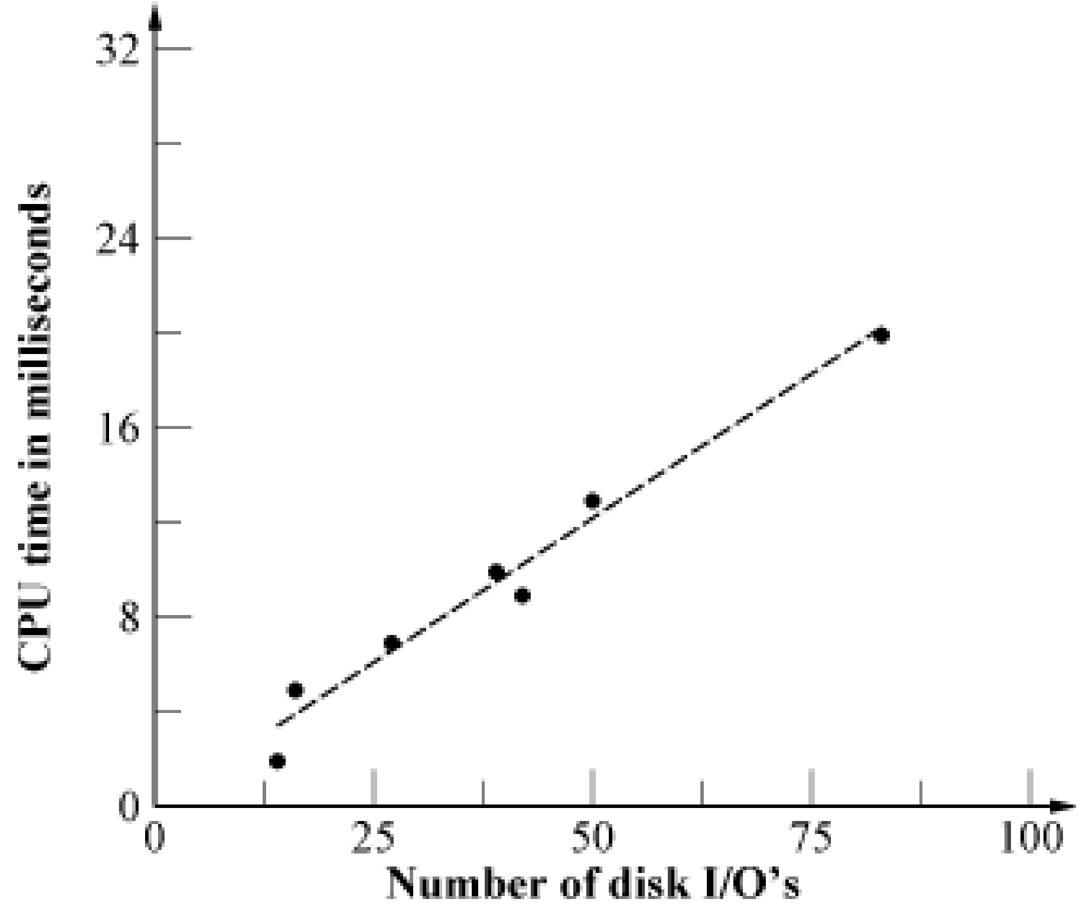
- The number of disk I/O's and processor times of seven programs were measured as: (14, 2), (16, 5), (27, 7), (42, 9), (39, 10), (50, 13), (83, 20)
- For this data: n=7, Σxy=3375, Σx=271, Σx^2=13,855, Σy=66, Σy^2=828, $\bar{x}$= 38.71, $\bar{y}$= 9.43. Therefore,

$$\begin{aligned} b_1 &= \frac{\Sigma xy - n\bar{x}\bar{y}}{\Sigma x^2 - n(\bar{x})^2} = \frac{3375 - 7 \times 38.71 \times 9.43}{13,855 - 7 \times (38.71)^2} = 0.2438 \\ b_0 &= \bar{y} - b_1\bar{x} = 9.43 - 0.2438 \times 38.71 = -0.0083 \end{aligned}$$

- The desired linear model is:

$$\text{CPU time} = -0.0083 + 0.2438(\text{Number of Disk I/O's})$$

Example 14.1 (Cont)



Example 14. (Cont)

- Error Computation

	Disk I/O's	CPU Time	Estimate	Error	Error^2
	x_i	y_i	$\hat{y}_i = b_0 + b_1\ x_i$	$e_i = y_i - \hat{y}_i$	e_i^2
	14	2	3.4043	-1.4043	1.9721
	16	5	3.8918	1.1082	1.2281
	27	7	6.5731	0.4269	0.1822
	42	9	10.2295	-1.2295	1.5116
	39	10	9.4982	0.5018	0.2518
	50	13	12.1795	0.8205	0.6732
	83	20	20.2235	-0.2235	0.0500
Σ	271	66	66.0000	0.00	5.8690

Derivation of Regression Parameters

- The error in the ith observation is:

$$e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_i)$$

- For a sample of n observations, the mean error is:

$$\begin{aligned} \bar{e} &= \frac{1}{n}\sum_i e_i = \frac{1}{n}\sum_i \{y_i - (b_0 + b_1 x_i)\} \\ &= \bar{y} - b_0 - b_1 \bar{x} \end{aligned}$$

- Setting mean error to zero, we obtain:

$$b_0 = \bar{y} - b_1 \bar{x}$$

- Substituting b0 in the error expression, we get:

$$e_i = y_i - \bar{y} + b_1 \bar{x} - b_1 x_i = (y_i - \bar{y}) - b_1 (x_i - \bar{x})$$

Derivation of Regression Parameters (Cont)

- The sum of squared errors SSE is:

$$
\begin{aligned}
\text{SSE} &= \sum_{i=1}^{n} e_i^2 \\
&= \sum_{i=1}^{n} \left\{ \left(y_i - \bar{y}\right)^2 + 2b_1 \left(y_i - \bar{y}\right)\left(x_i - \bar{x}\right) + b_1^2 \left(x_i - \bar{x}\right)^2 \right\} \\
\frac{\text{SSE}}{n-1} &= \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 - 2b_1 \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)\left(x_i - \bar{x}\right) \\
&\quad + b_1^2 \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \\
&= s_y^2 - 2b_1 s_{xy}^2 + b_1^2 s_x^2
\end{aligned}
$$

Derivation (Cont)

- Differentiating this equation with respect to b_1 and equating the result to zero:

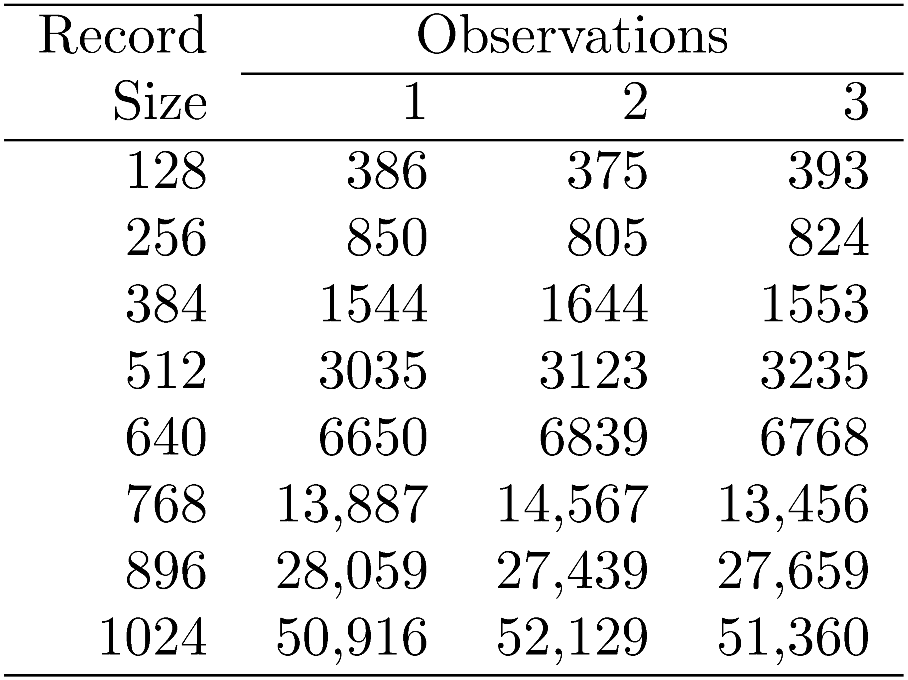
$$\frac{1}{n-1}\frac{d(\mathrm{SSE})}{db_1} = -2s_{xy}^2 + 2b_1 s_x^2 = 0$$

- That is,

$$b_1 = \frac{s_{xy}^2}{s_x^2} = \frac{\Sigma xy - n\bar{x}\bar{y}}{\Sigma x^2 - n(\bar{x})^2}$$

Homework 14A: Exercise 14.7

- The time to encrypt a *k* byte record using an encryption technique is shown in the following table. Fit a linear regression model to this data.



Record	Observations		
Size	1	2	3
128	386	375	393
256	850	805	824
384	1544	1644	1553
512	3035	3123	3235
640	6650	6839	6768
768	13,887	14,567	13,456
896	28,059	27,439	27,659
1024	50,916	52,129	51,360

Allocation of Variation

- Error variance without Regression = Variance of the response

$$\begin{aligned} \text{Error} &= \epsilon_i = \text{Observed Response} - \text{Predicted Response} \\ &= y_i - \bar{y} \end{aligned}$$

and

$$\begin{aligned} \text{Variance of Errors without regression} &= \frac{1}{n-1}\sum_{i=1}^{n}\epsilon_i^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2 \\ &= \text{Variance of y} \end{aligned}$$

Allocation of Variation (Cont)

- The sum of squared errors without regression would be:

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

- This is called **total sum of squares** or (SST). It is a measure of *y*'s variability and is called **variation** of *y*. SST can be computed as follows:

$$\mathrm{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \left(\sum_{i=1}^{n} y_i^2\right) - n\bar{y}^2 = SSY - SS0$$

- Where, SSY is the sum of squares of *y* (or $\Sigma\, y^2$). SS0 is the sum of squares of $\bar{y}$ and is equal to $n\bar{y}^2$

Allocation of Variation (Cont)

- The difference between SST and SSE is the sum of squares explained by the regression. It is called SSR:

$$\mathrm{SSR} = \mathrm{SST} - \mathrm{SSE}$$

or

$$\mathrm{SST} = \mathrm{SSR} + \mathrm{SSE}$$

- The fraction of the variation that is explained determines the goodness of the regression and is called the coefficient of determination, R^2:

$$R^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} = \frac{\mathrm{SST} - \mathrm{SSE}}{\mathrm{SST}}$$

Allocation of Variation (Cont)

- The higher the value of R^2, the better the regression. $R^2=1 \Rightarrow$ Perfect fit $R^2=0 \Rightarrow$ No fit

$$\text{Sample Correlation}(x, y) = R_{xy} = \frac{s^2_{xy}}{s_x s_y}$$

- Coefficient of Determination = {Correlation Coefficient (x,y)}^{2}
- Shortcut formula for SSE:

$$\text{SSE} = \Sigma y^2 - b_0 \Sigma y - b_1 \Sigma xy$$

Example 14.2

- For the disk I/O-CPU time data of Example 14.1:

$$\begin{aligned} \text{SSE} &= \Sigma y^2 - b_0 \Sigma y - b_1 \Sigma xy \\ &= 828 + 0.0083 \times 66 - 0.2438 \times 3375 = 5.87 \end{aligned}$$

$$\begin{aligned} \text{SST} &= \text{SSY} - \text{SS0} = \Sigma y^2 - n(\bar{y})^2 \\ &= 828 - 7 \times (9.43)^2 = 205.71 \end{aligned}$$

$$\text{SSR} = \text{SST} - \text{SSE} = 205.71 - 5.87 = 199.84$$

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{199.84}{205.71} = 0.9715$$

- The regression explains 97% of CPU time's variation.

Standard Deviation of Errors

- Since errors are obtained after calculating two regression parameters from the data, errors have *n-2* degrees of freedom

$$s_e = \sqrt{\frac{\text{SSE}}{n-2}}$$

- SSE/*(n-2)* is called **mean squared errors** or (MSE).
- Standard deviation of errors = square root of MSE.
- SSY has *n* degrees of freedom since it is obtained from *n* independent observations without estimating any parameters.
- SS0 has just one degree of freedom since it can be computed simply from $\bar{y}$
- SST has *n-1* degrees of freedom, since one parameter $\bar{y}$ must be calculated from the data before SST can be computed.

Standard Deviation of Errors (Cont)

- SSR, which is the difference between SST and SSE, has the remaining one degree of freedom.
- Overall,

$$\begin{array}{ccccccccc} \mathrm{SST} & = & \mathrm{SSY} & - & \mathrm{SS0} & = & \mathrm{SSR} & + & \mathrm{SSE} \\ n-1 & = & n & - & 1 & = & 1 & + & (n-2) \end{array}$$

- Notice that the degrees of freedom add just the way the sums of squares do.

Example 14.3

- For the disk I/O-CPU data of Example 14.1, the degrees of freedom of the sums are:

$$
\begin{array}{lccccccccc}
\text{SS}: & \text{SST} & = & \text{SSY} & - & \text{SS0} & = & \text{SSR} & + & \text{SSE} \\
 & 205.71 & = & 828 & - & 622.29 & = & 199.84 & + & 5.87 \\
\text{DF}: & 6 & = & 7 & - & 1 & = & 1 & + & 5
\end{array}
$$

- The mean squared error is:

$$\text{MSE} = \frac{\text{SSE}}{\text{DF for Errors}} = \frac{5.87}{5} = 1.17$$

- The standard deviation of errors is:

$$s_e = \sqrt{\text{MSE}} = \sqrt{1.17} = 1.08$$

Confidence Intervals for Regression Params

- Regression coefficients b_0 and b_1 are estimates from a single sample of size *n* ⇒ Random
  ⇒ Using another sample, the estimates may be different. If β_0 and β_1 are true parameters of the population. That is,

$$y = \beta_0 + \beta_1 x$$

- Computed coefficients b_0 and b_1 are estimates of β_0 and β_1, respectively.

$$s_{b_0} = s_e \left[\frac{1}{n} + \frac{\bar{x}^2}{\Sigma x^2 - n\bar{x}^2} \right]^{1/2}$$

$$s_{b_1} = \frac{s_e}{[\Sigma x^2 - n\bar{x}^2]^{1/2}}$$

Confidence Intervals (Cont)

- The 100(1-α)% confidence intervals for b_0 and b_1 can be be computed using $t_{[1-\alpha/2;\, n-2]}$ --- the 1-α/2 quantile of a t variate with n-2 degrees of freedom. The confidence intervals are:

$$b_0 \mp t s_{b_0}$$

And

$$b_1 \mp t s_{b_1}$$

- If a confidence interval includes zero, then the regression parameter cannot be considered different from zero at the at 100(1-α)% confidence level.

Example 14.4

- For the disk I/O and CPU data of Example 14.1, we have n=7, $\bar{x}$=38.71, Σx^2=13,855, and s_e=1.0834.
- Standard deviations of b_0 and b_1 are:

$$
\begin{aligned}
s_{b_0} &= s_e\left[\frac{1}{n}+\frac{\bar{x}^2}{\Sigma x^2 - n\bar{x}^2}\right]^{1/2} \\
&= 1.0834\left[\frac{1}{7}+\frac{(38.71)^2}{13,855 - 7\times 38.71\times 38.71}\right]^{1/2} = 0.8311 \\
s_{b_1} &= \frac{s_e}{[\Sigma x^2 - n\bar{x}^2]^{1/2}} \\
&= \frac{1.0834}{[13,855 - 7\times 38.71\times 38.71]^{1/2}} = 0.0187
\end{aligned}
$$

Example 14.4 (Cont)

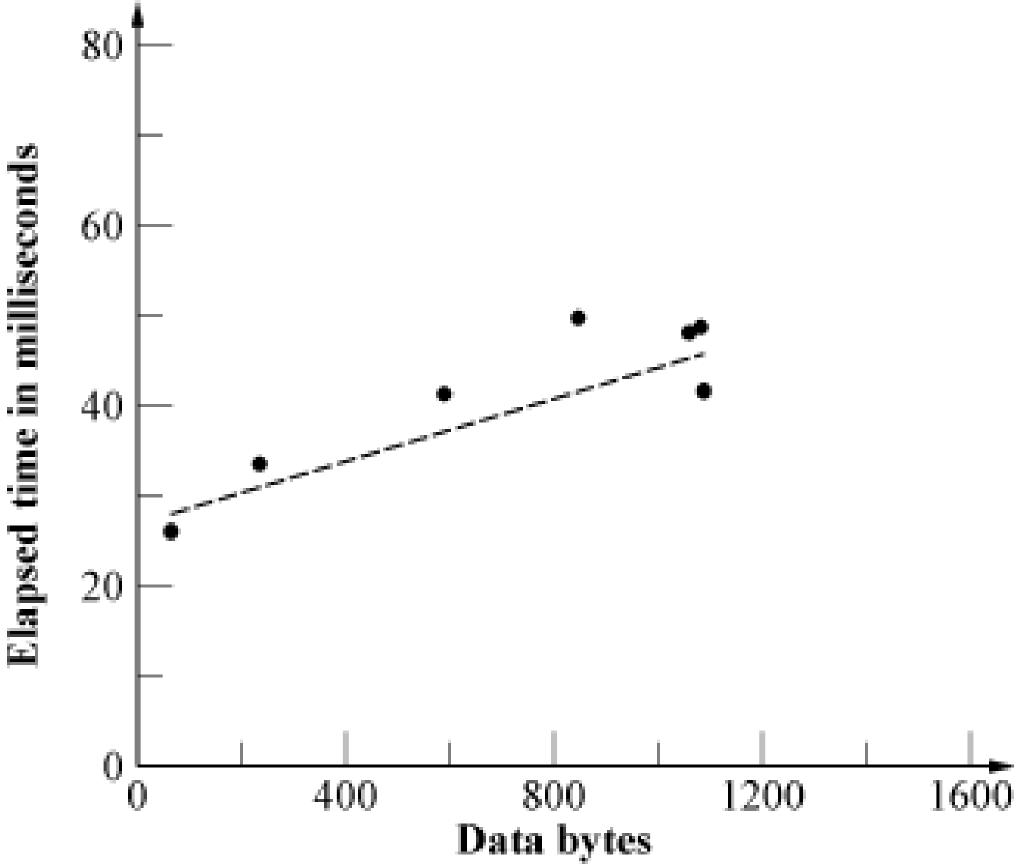
- From Appendix Table A.4, the 0.95-quantile of a *t*-variate with 5 degrees of freedom is 2.015.
  ⇒ 90% confidence interval for b_0 is:
  $$\begin{aligned} -0.0083 \mp (2.015)(0.8311) &= -0.0083 \mp 1.6747 \\ &= (-1.6830,\ 1.6663) \end{aligned}$$
- Since, the confidence interval includes zero, the hypothesis that this parameter is zero cannot be rejected at 0.10 significance level. ⇒ b_0 is essentially zero.
- 90% Confidence Interval for b_1 is:
  $$\begin{aligned} 0.2438 \mp (2.015)(0.0187) &= 0.2438 \mp 0.0376 \\ &= (0.2061,\ 0.2814) \end{aligned}$$
- Since the confidence interval does not include zero, the slope b_1 is significantly different from zero at this confidence level.

Case Study 14.1: Remote Procedure Call

UNIX		ARGUS	
Data Bytes	Time	Data Bytes	Time
64	26.4	92	32.8
64	26.4	92	34.2
64	26.4	92	32.4
64	26.2	92	34.4
234	33.8	348	41.4
590	41.6	604	51.2
846	50.0	860	76.0
1060	48.4	1074	80.8
1082	49.0	1074	79.8
1088	42.0	1088	58.6
1088	41.8	1088	57.6
1088	41.8	1088	59.8
1088	42.0	1088	57.4

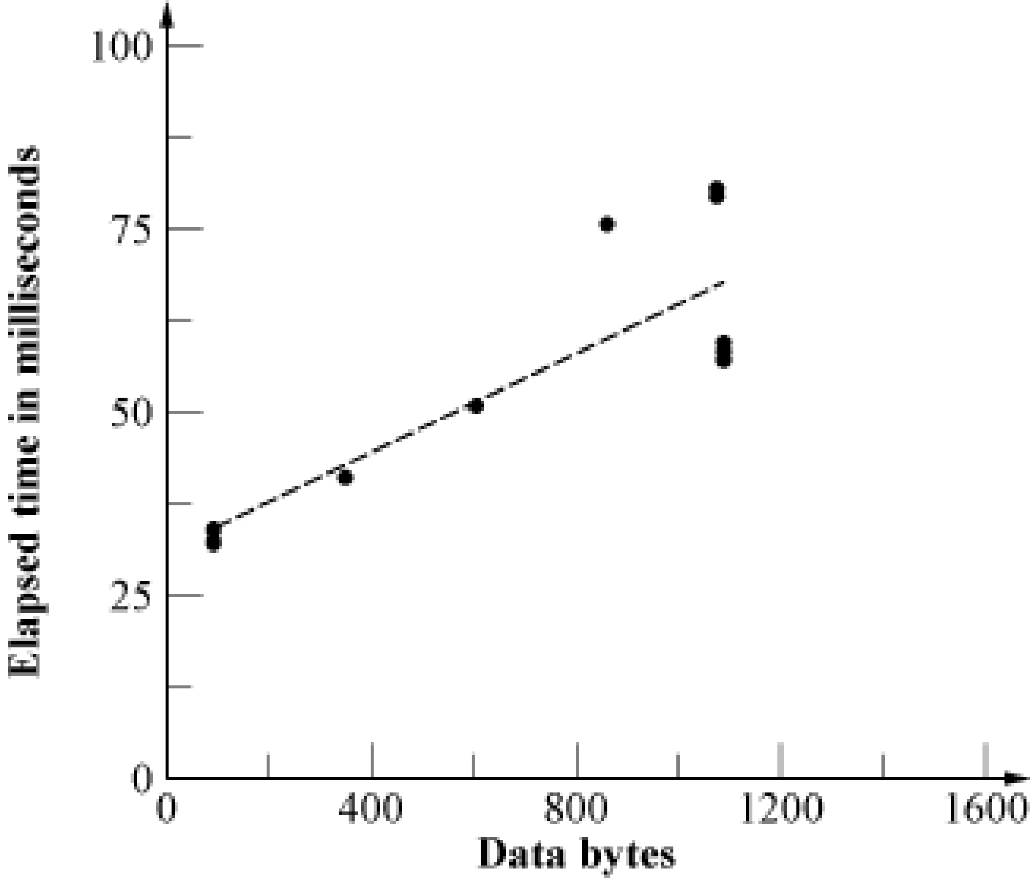
Case Study 14.1 (Cont)

- UNIX:



Case Study 14.1 (Cont)

- ARGUS:



Case Study 14.1 (Cont)

- Best linear models are:

$$\begin{aligned}\text{Time on UNIX} &= 0.017\ (\text{Data size in bytes}) + 26.898\\ \text{Time on ARGUS} &= 0.034\ (\text{Data size in bytes}) + 31.068\end{aligned}$$

- The regressions explain 81% and 75% of the variation, respectively.

  Does ARGUS takes larger time per byte as well as a larger set up time per call than UNIX?

Case Study 14.1 (Cont)

UNIX:

Parameter	Mean	Std. Dev.	Confidence Interval
b_0	26.898	2.005	(23.2968, 30.4988)
b_1	0.017	0.003	(0.0128, 0.0219)

ARGUS:

Parameter	Mean	Std. Dev.	Confidence Interval
b_0	31.068	4.711	(22.6076, 39.5278)
b_1	0.034	0.006	(0.0231, 0.0443)

- Intervals for intercepts overlap while those of the slopes do not. ⇒ Set up times are not significantly different in the two systems while the per byte times (slopes) are different.

Homework 14B: Exercise 14.7

- For the data of Exercise 14.7 (Homework 14A), compute R2 and 90% confidence intervals for regression parameters.

Confidence Intervals for Predictions

$$\hat{y}_p = b_0 + b_1 x_p$$

- This is only the mean value of the predicted response. Standard deviation of the mean of a future sample of m observations is:

$$s_{\hat{y}_{mp}} = s_e \left[\frac{1}{m} + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma x^2 - n\bar{x}^2} \right]^{1/2}$$

- m =1 ⇒ Standard deviation of a single future observation:

$$s_{\hat{y}_{1p}} = s_e \left[1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma x^2 - n\bar{x}^2} \right]^{1/2}$$

CI for Predictions (Cont)

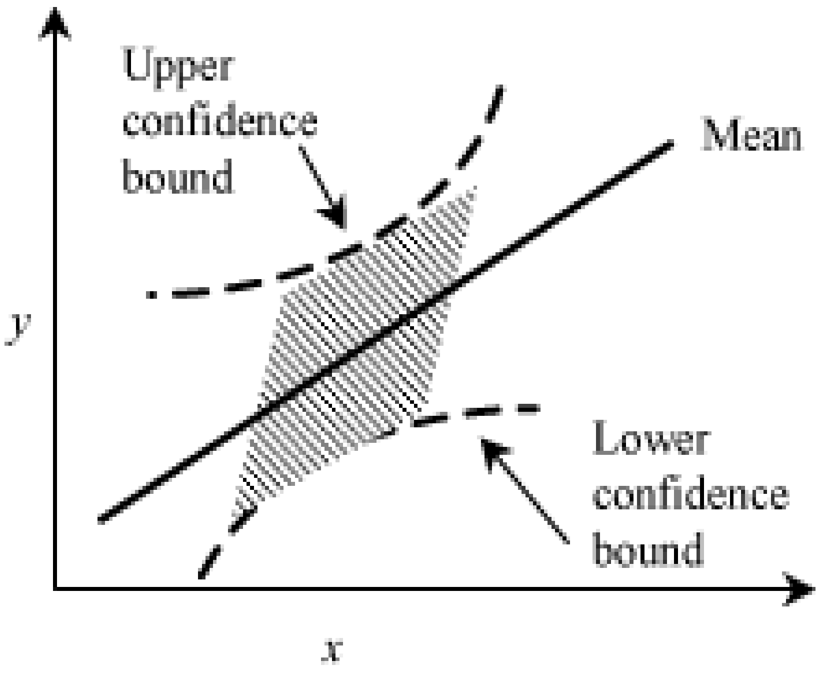
- m = ∞ ⇒ Standard deviation of the mean of a large number of future observations at x_p:

$$s_{\hat{y}_p} = s_e \left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma x^2 - n\bar{x}^2} \right]^{1/2}$$

- 100(1-α)% confidence interval for the mean can be constructed using a t quantile read at *n-2* degrees of freedom.

CI for Predictions (Cont)

- Goodness of the prediction decreases as we move away from the center.



Example 14.5

- Using the disk I/O and CPU time data of Example 14.1, let us estimate the CPU time for a program with 100 disk I/O's.

$$\text{CPU time} = -0.0083 + 0.2438(\text{Number of disk I/O's})$$

- For a program with 100 disk I/O's, the mean CPU time is:

$$\text{CPU time} = -0.0083 + 0.2438(100) = 24.3674$$

$$\text{Standard deviation of errors } s_e = 1.0834$$

Example 14.5 (Cont)

- The standard deviation of the predicted mean of a large number of observations is:

$$s_{\hat{y}_p} = 1.0834\left[\frac{1}{7} + \frac{(100-38.71)^2}{13,855 - 7(38.71)^2}\right]^{1/2} = 1.2159$$

- From Table A.4, the 0.95-quantile of the t-variate with 5 degrees of freedom is 2.015.

  $\Rightarrow$ 90% CI for the predicted mean

$$\begin{aligned} &= \quad 24.3674 \mp (2.015)(1.2159) \\ &= \quad (21.9174,\ 26.8174) \end{aligned}$$

Example 14.5 (Cont)

- CPU time of a single future program with 100 disk I/O's:

$$s_{\hat{y}_{1p}} = 1.0834\left[1+\frac{1}{7}+\frac{(100-38.71)^2}{13,855-7(38.71)^2}\right]^{1/2} = 1.6286$$

- 90% CI for a single prediction:

$$\begin{aligned} &= \quad 24.3674 \mp (2.015)(1.6286) \\ &= \quad (21.0858,\ 27.6489) \end{aligned}$$

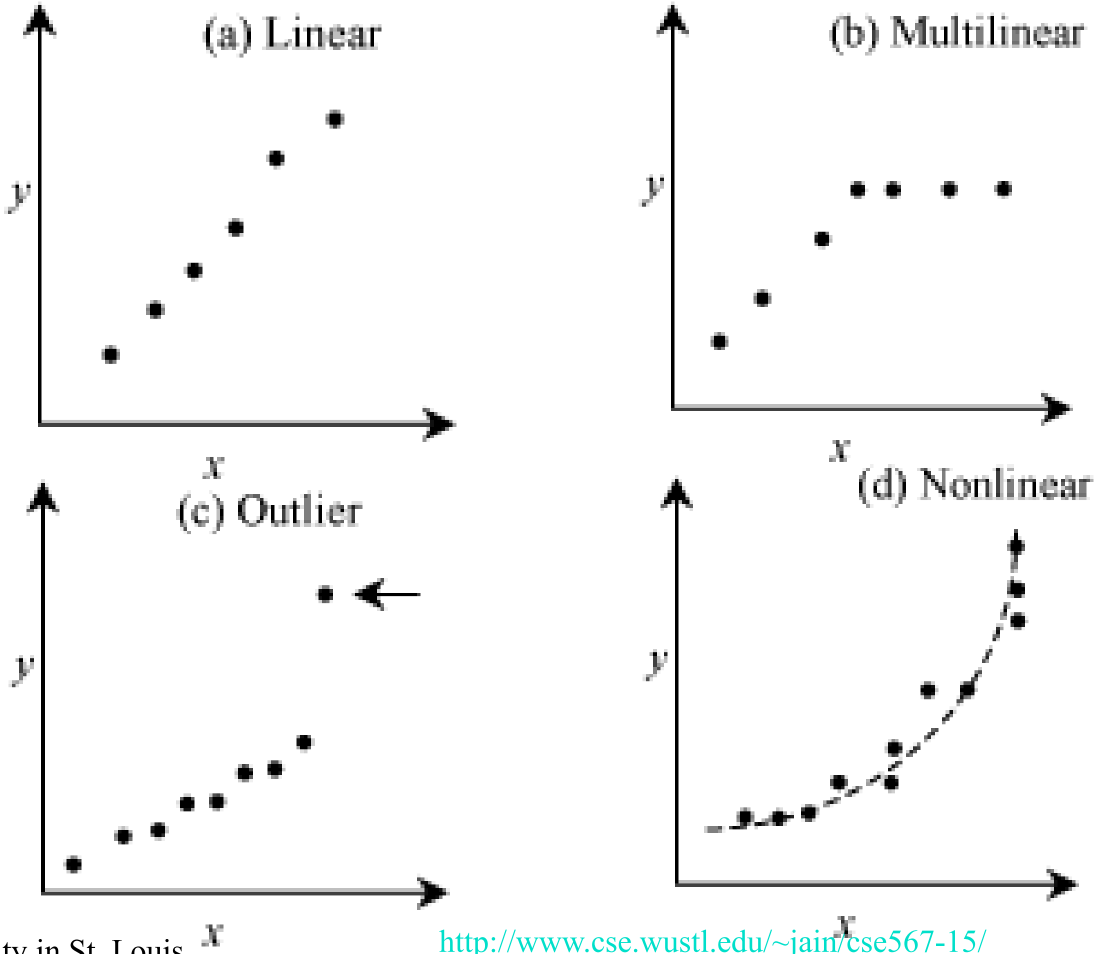
Visual Tests for Regression Assumptions

Regression assumptions:

1. The true relationship between the response variable *y* and the predictor variable *x* is linear.
2. The predictor variable *x* is non-stochastic and it is measured without any error.
3. The model errors are statistically independent.
4. The errors are normally distributed with zero mean and a constant standard deviation.

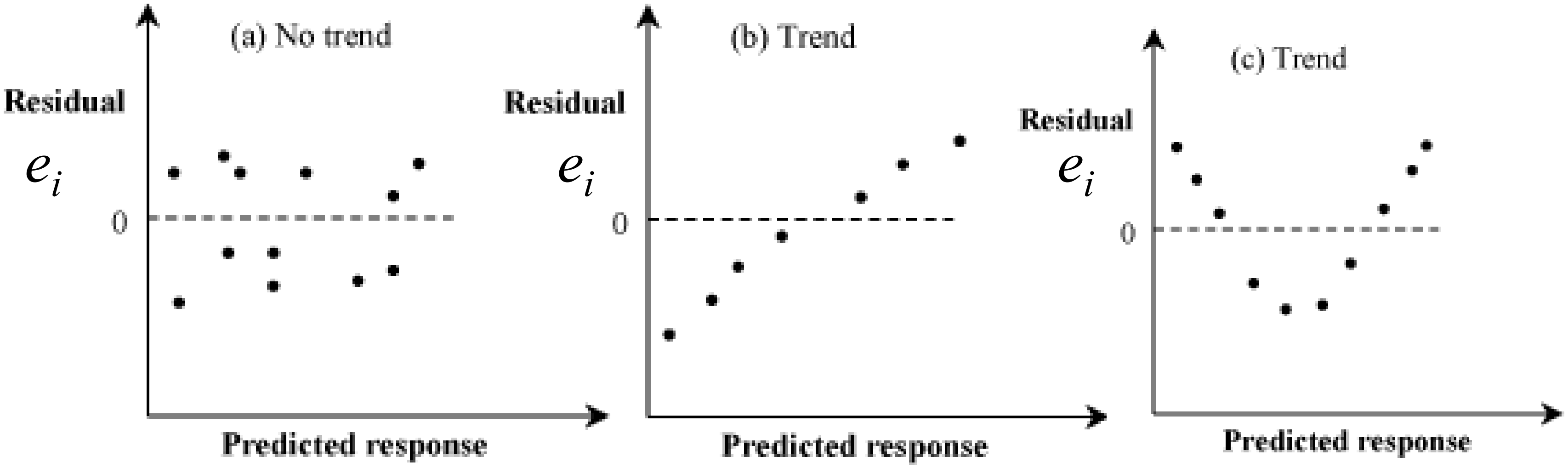
1. Linear Relationship: Visual Test

- Scatter plot of y versus x ⇒ Linear or nonlinear relationship



2. Independent Errors: Visual Test

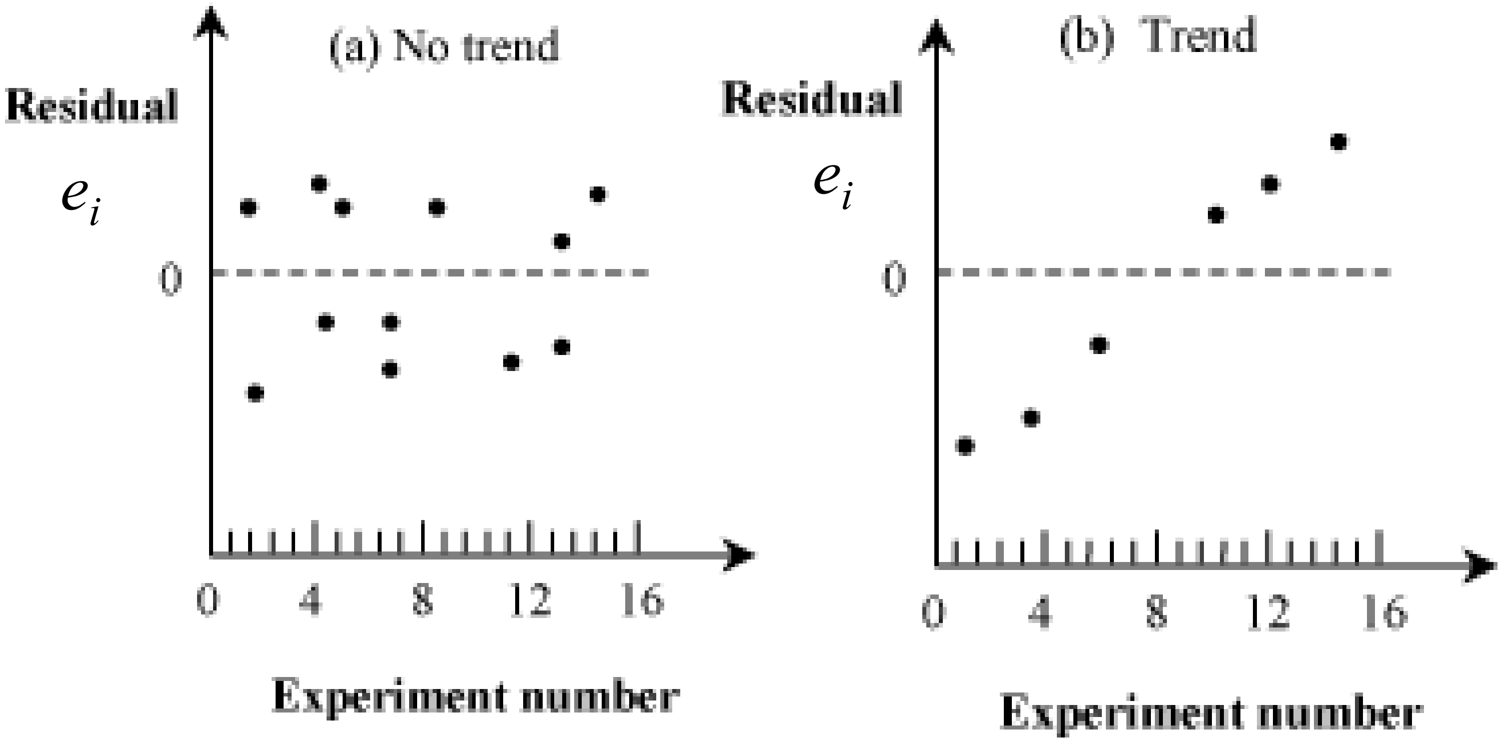
1. Scatter plot of e_i versus the predicted response $\hat{y}_i$



- All tests for independence simply try to find dependence.

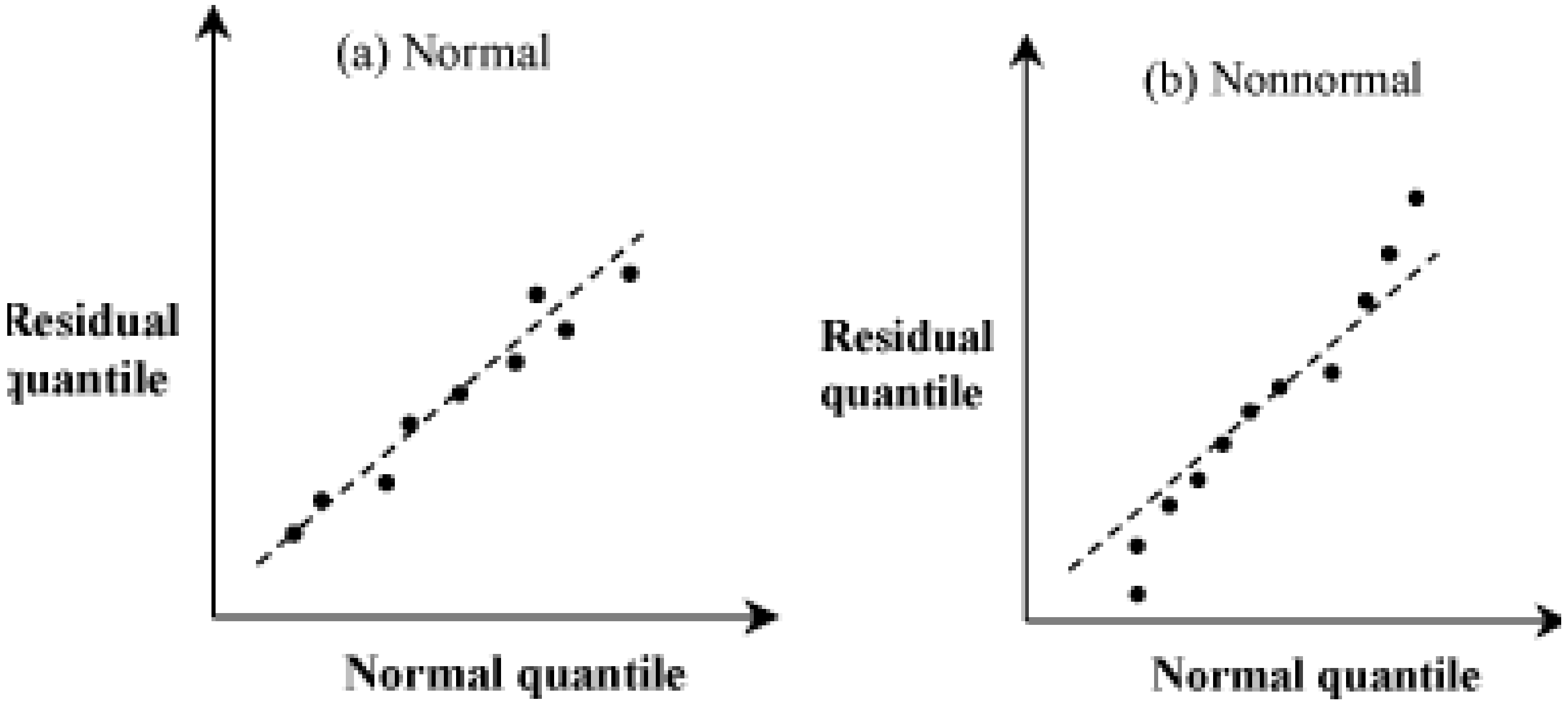
Independent Errors (Cont)

2. Plot the residuals as a function of the experiment number



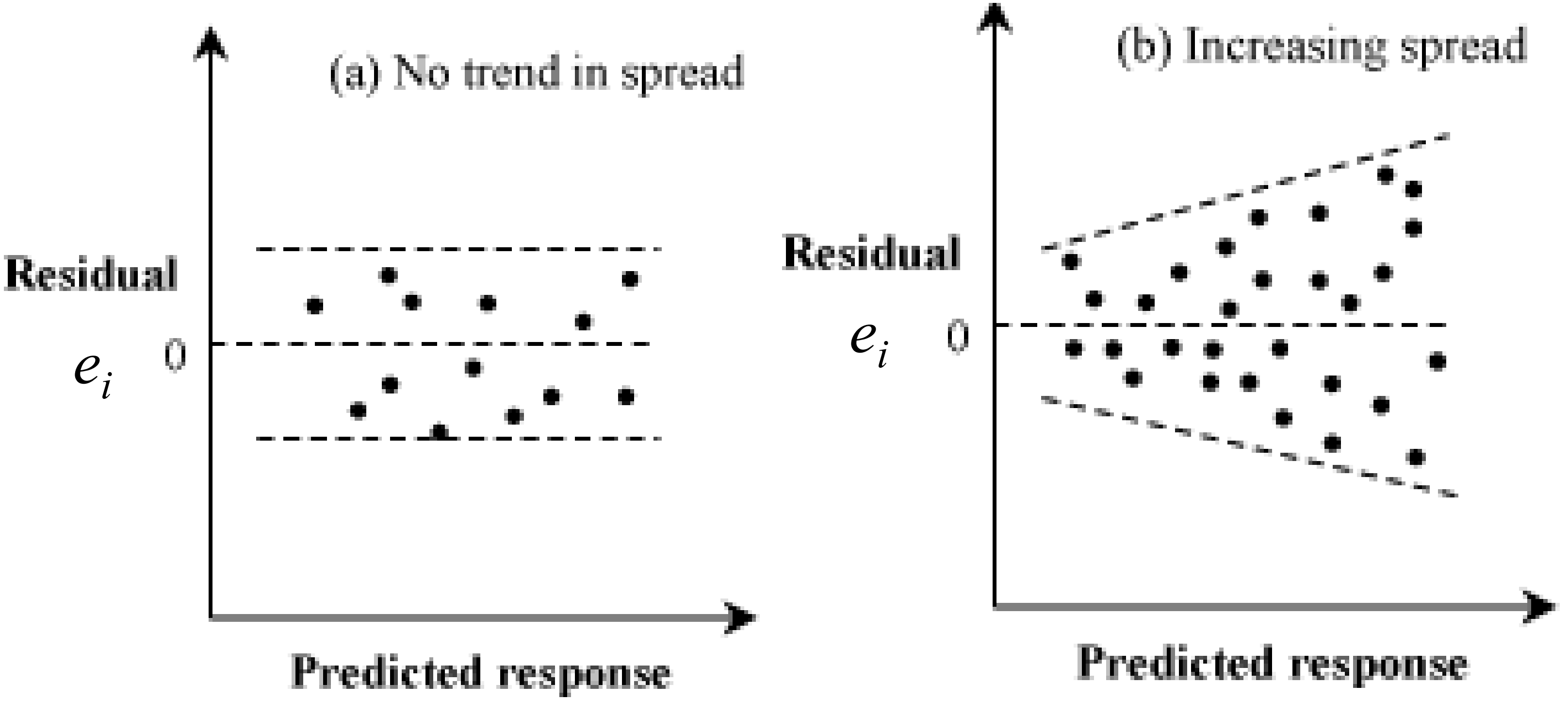
3. Normally Distributed Errors: Test

- Prepare a normal quantile-quantile plot of errors. Linear ⇒ the assumption is satisfied.



4. Constant Standard Deviation of Errors

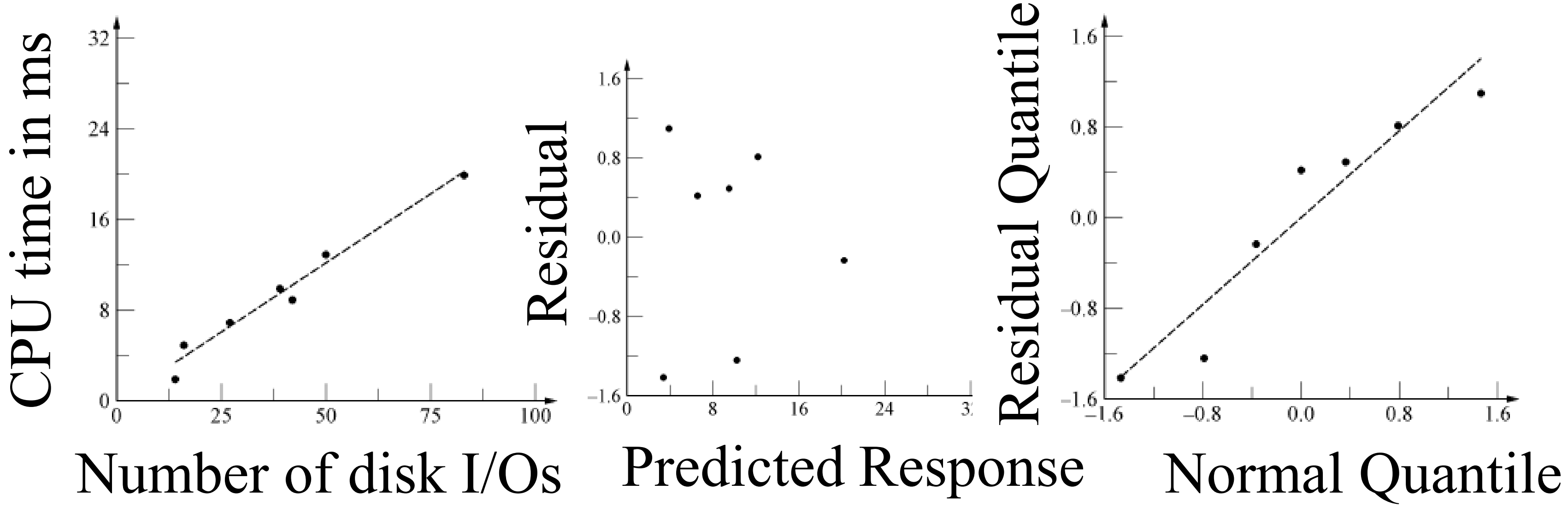
- Also known as **homoscedasticity**



- Trend ⇒ Try curvilinear regression or transformation

Example 14.6

For the disk I/O and CPU time data of Example 14.1

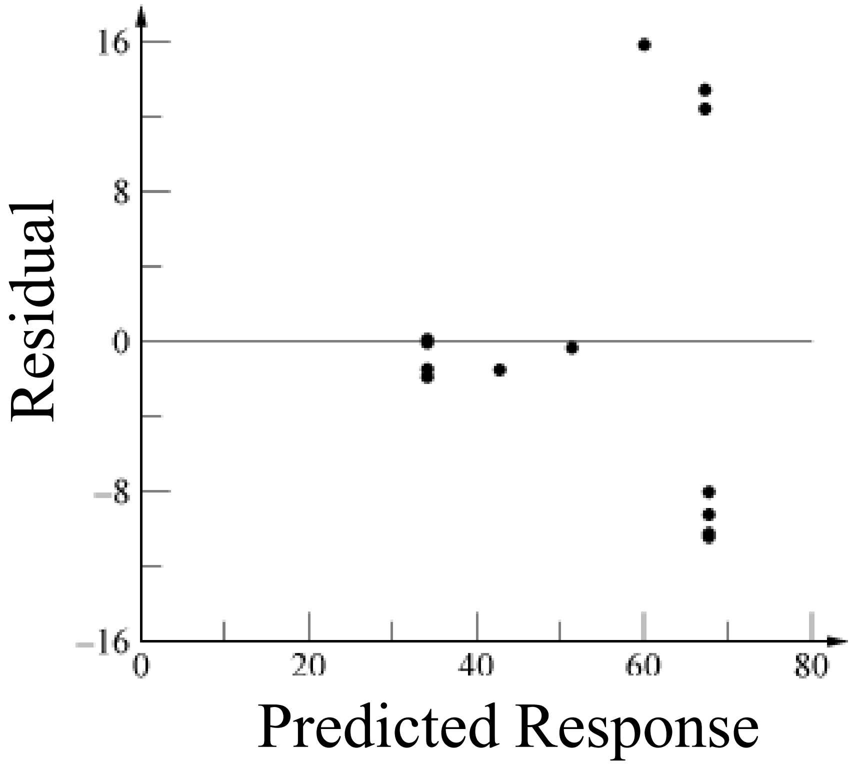


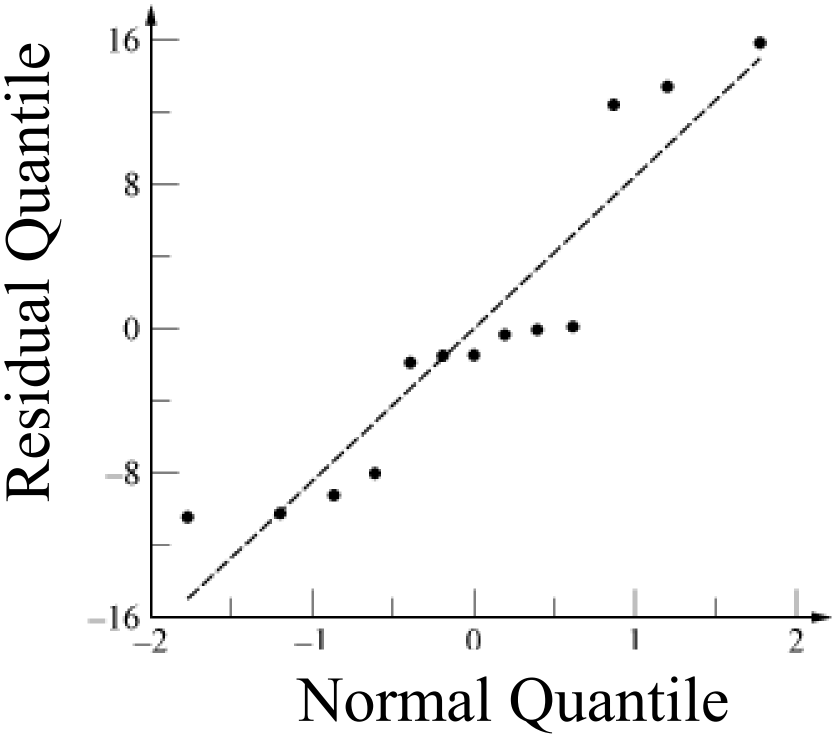
1. Relationship is linear

2. No trend in residuals ⇒ Seem independent

3. Linear normal quantile-quantile plot ⇒ Larger deviations at lower values but all values are small

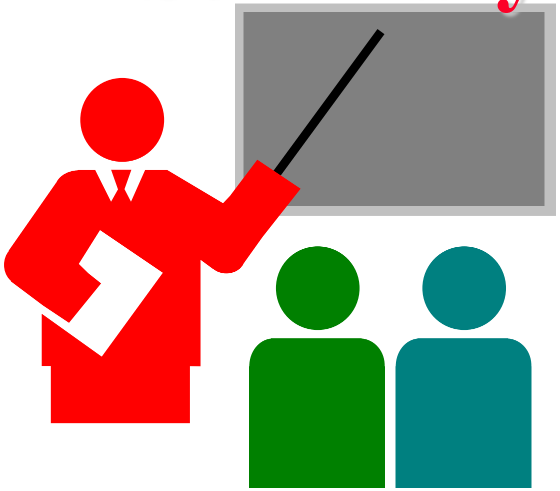
Example 14.7: RPC Performance





1. Larger errors at larger responses
2. Normality of errors is questionable

Summary



- **Terminology**: Simple Linear Regression model, Sums of Squares, Mean Squares, degrees of freedom, percent of variation explained, Coefficient of determination, correlation coefficient
- Regression parameters as well as the predicted responses have confidence intervals
- It is important to verify assumptions of linearity, error independence, error normality ⇒ Visual tests

Homework 14C: Exercise 14.7

- For the data of Exercise 14.7 (Homework 14B), use visual tests to verify the regression assumptions. Write your observations from the graphs.