Simple Linear Regression Models

Hongwei Zhang

http://www.cs.wayne.edu/~hzhang

Statistics is the art of lying by means of figures.

--- Dr. Wilhelm Stekhel

Acknowledgement: this lecture is partially based on the slides of Dr. Raj Jain.

Simple linear regression models

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

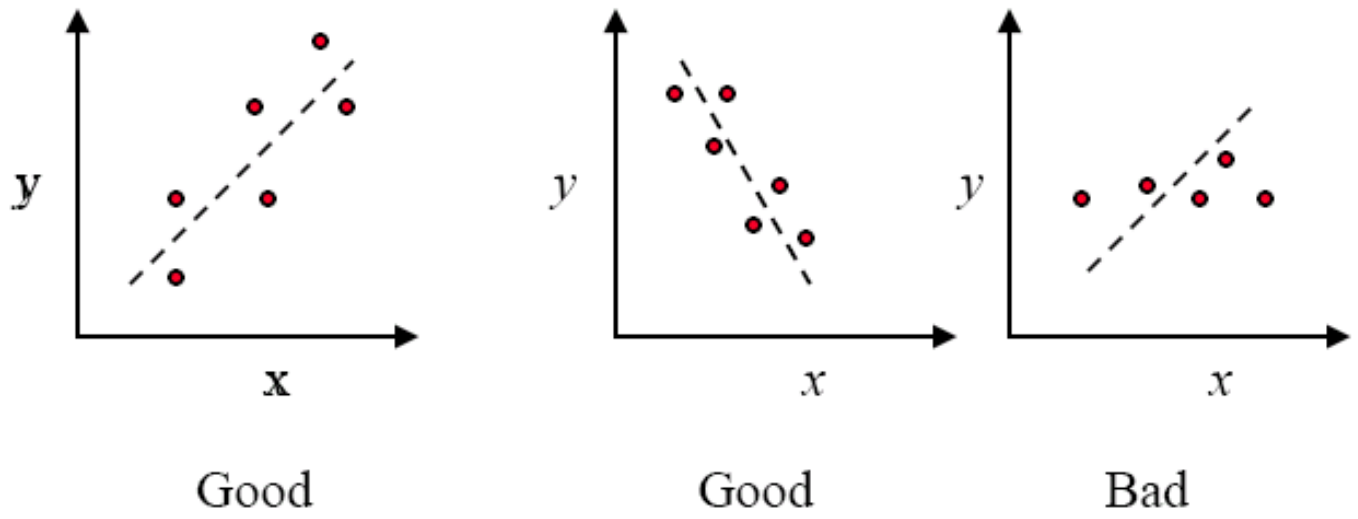
Outline

- Definition of a Good Model
- Estimation of Model parameters
- Allocation of Variation
- Standard deviation of Errors
- Confidence Intervals for Regression Parameters
- Confidence Intervals for Predictions
- Visual Tests for verifying Regression Assumption

Outline

- Definition of a Good Model
- Estimation of Model parameters
- Allocation of Variation
- Standard deviation of Errors
- Confidence Intervals for Regression Parameters
- Confidence Intervals for Predictions
- Visual Tests for verifying Regression Assumption

Definition of a good model?



Good models (contd.)

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

Good models (contd.)

- Formally,
  - $\hat{y} = b_0 + b_1 x$

    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 {(*x1, y1*), ..., (*xn, yn*)}, the estimated response for the i-th observation is:

    $$\hat{y}_i = b_0 + b_1 x_i$$
  - The error is:

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

Good models (contd.)

- 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 overall 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 the unconstrained minimization of the variance of errors (Exercise 14.1)

Outline

- Definition of a Good Model
- Estimation of Model parameters
- Allocation of Variation
- Standard deviation of Errors
- Confidence Intervals for Regression Parameters
- Confidence Intervals for Predictions
- Visual Tests for verifying Regression Assumption

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,

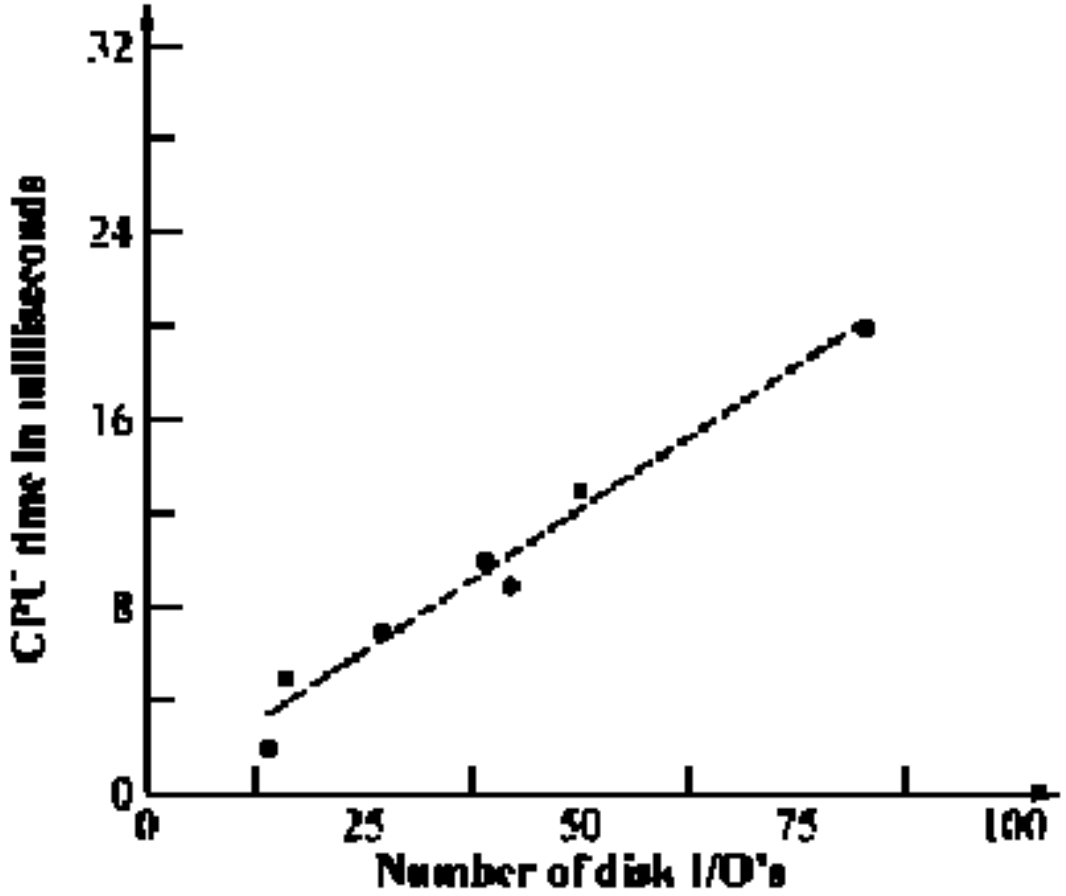
$$b_1 = \frac{\Sigma xy - n\bar{x}\bar{y}}{\Sigma x^2 - n(\bar{x})^2} = \frac{3375 - 7 \times 38.71 \times 9.43}{13,855 - 7 \times (38.71)^2} = 0.2438$$

$$b_0 = \bar{y} - b_1\bar{x} = 9.43 - 0.2438 \times 38.71 = -0.0083$$

- The desired linear model is:

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

Example (contd.)



Example (contd.)

- Error Computation

	Disk I/O's x_i	CPU Time y_i	Estimate $\hat{y}_i=b_0+b_1\ x_i$	Error $e_i=y_i-\hat{y}_i$	Error^2 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 (contd.)

- The sum of squared errors SSE is:

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

Next step?

Derivation (contd.)

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

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

- That is,

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

Least Squares Regression vs. Least Absolute Deviations Regression?

Least Squares Regression	Least Absolute Deviations Regression
Not very robust to outliers	Robust to outliers
Simple analytical solution	No analytical solving method (have to use iterative computation-intensive method)
Stable solution	Unstable solution
Always one unique solution	Possibly multiple solutions

The *unstable* property of the method of least absolute deviations means that, for any small horizontal adjustment of a data point, the regression line may jump a large amount. In contrast, the least squares solutions is *stable* in that, for any small horizontal adjustment of a data point, the regression line will always move only slightly, or continuously.

Outline

- Definition of a Good Model
- Estimation of Model parameters
- Allocation of Variation
- Standard deviation of Errors
- Confidence Intervals for Regression Parameters
- Confidence Intervals for Predictions
- Visual Tests for verifying Regression Assumption

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 (contd.)

- 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:

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

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

Allocation of variation (contd.)

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

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

or

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

Variation not explained by the regression

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

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

Allocation of variation (contd.)

- The higher the value of R^2, the better the regression.
  $R^2=1 \Rightarrow$ Perfect fit; $R^2=0 \Rightarrow$ No fit
- Shortcut formula for SSE:

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

Example

- 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.

Outline

- Definition of a Good Model
- Estimation of Model parameters
- Allocation of Variation
- Standard deviation of Errors
- Confidence Intervals for Regression Parameters
- Confidence Intervals for Predictions
- Visual Tests for verifying Regression Assumption

Standard deviation of errors

- Since errors are obtained after calculating two regression parameters from the data, errors have *n-2* degrees of freedom
- SSE/*(n-2)* is called **mean squared errors** or (MSE)

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

- Standard deviation of errors = square root of MSE

- Note:
  - 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 (contd.)

- 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

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

SS :	SST	=	SSY	−	SS0	=	SSR	+	SSE
	205.71	=	828	−	622.29	=	199.84	+	5.87
DF :	6	=	7	−	1	=	1	+	5

- 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:

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

Outline

- Definition of a Good Model
- Estimation of Model parameters
- Allocation of Variation
- Standard deviation of Errors
- Confidence Intervals for Regression Parameters
- Confidence Intervals for Predictions
- Visual Tests for verifying Regression Assumption

CIs for regression parameters

- Regression coefficients *b0* and *b1* are estimates from a single sample of size *n* => 1) Random; 2) Using another sample, the estimates may be different.

- If β_0 and β_1 are true parameters of the population (i.e., $y = \beta_0 + \beta_1 x$), then the computed coefficients *b0* and *b1* are estimates of β0 and β1, respectively.

- Sample standard deviation of b0 and b1

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

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

CIs for regression parameters (contd.)

- The 100(1-α)% confidence intervals for b0 and b1 can be computed using t[1-α/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 100(1-α)% confidence level

Example

- 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:

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

$$= 1.0834 \left[\frac{1}{7} + \frac{(38.71)^2}{13,855 - 7 \times 38.71 \times 38.71}\right]^{1/2} = 0.8311$$

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

$$= \frac{1.0834}{[13,855 - 7 \times 38.71 \times 38.71]^{1/2}} = 0.0187$$

Example (contd.)

- The 0.95-quantile of a *t*-variate with 5 degrees of freedom is 2.015

=> 90% confidence interval for b0 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 => b0 is essentially zero.

=> 90% Confidence Interval for b1 is:

$$\begin{aligned} 0.2438 \mp (2.015)(0.0187) &= 0.2438 \mp 0.0376 \\ &= (0.2061,\ 0.2814) \end{aligned}$$

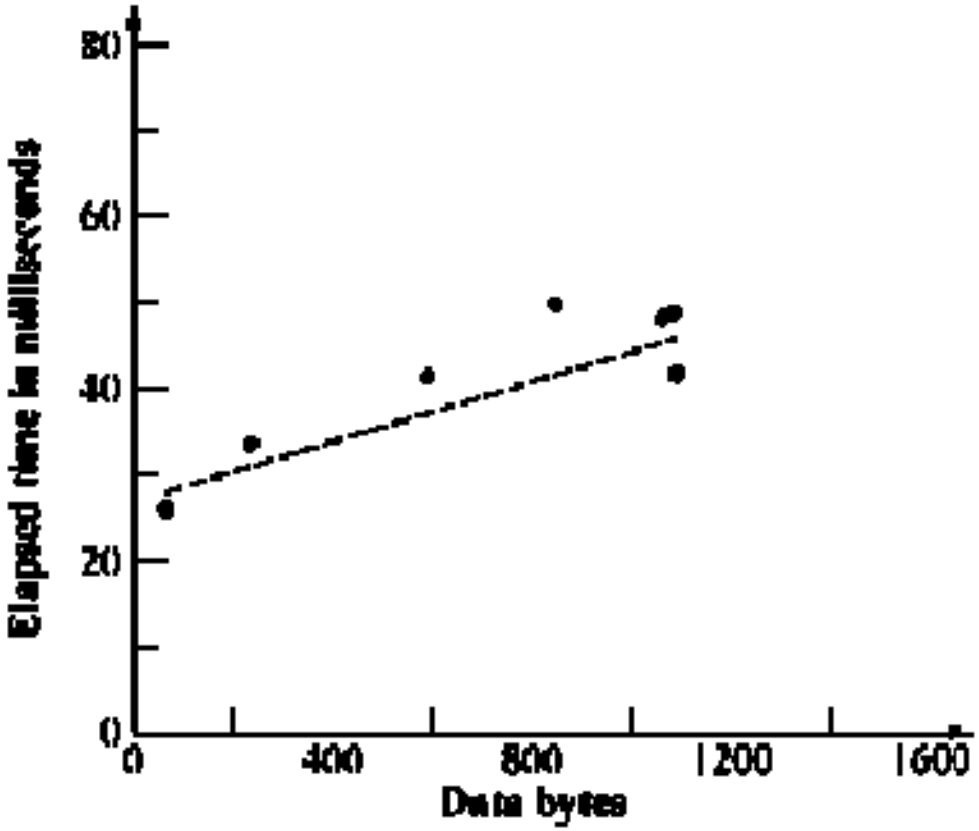
- Since the confidence interval does not include zero, the slope b1 is significantly different from zero at this confidence level.

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

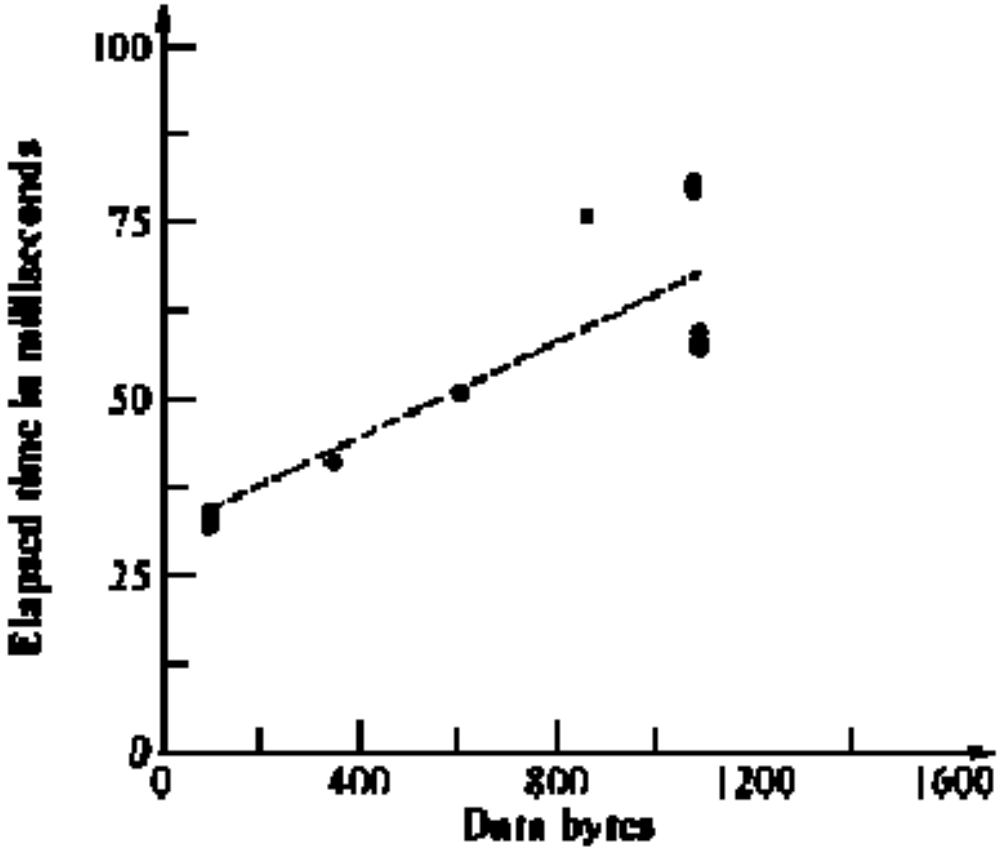
Case study (contd.)

- UNIX:



Case study (contd.)

- ARGUS:



Case study (contd.)

- Best linear models are:

$$\begin{aligned} \text{Time on UNIX} &= 0.030\ (\text{Data size in bytes}) + 24 \\ \text{Time on ARGUS} &= 0.034\ (\text{Data size in bytes}) + 30 \end{aligned}$$

- 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 (contd.)

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.

Outline

- Definition of a Good Model
- Estimation of Model parameters
- Allocation of Variation
- Standard deviation of Errors
- Confidence Intervals for Regression Parameters
- Confidence Intervals for Predictions
- Visual Tests for verifying Regression Assumption

CI for predications

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

- 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 $\Rightarrow$ 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 predications (contd.)

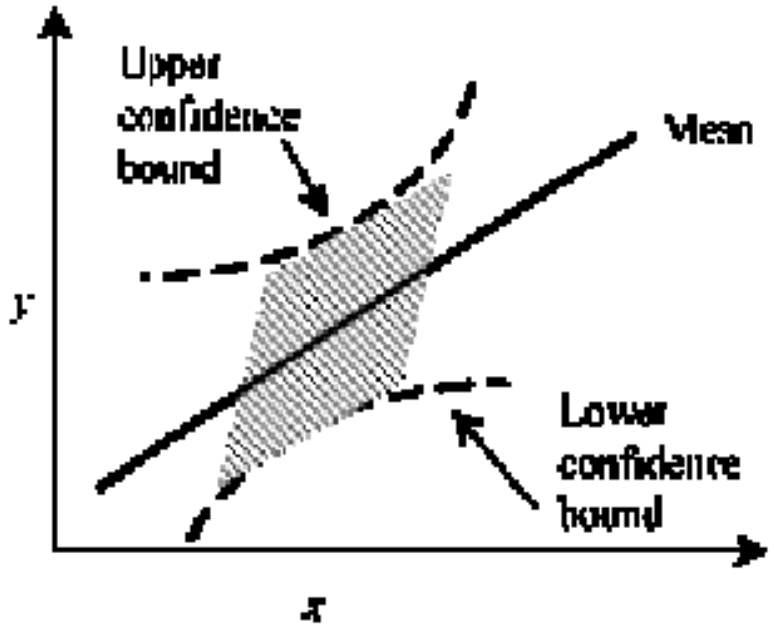
- m = $\infty \Rightarrow$ Standard deviation of the mean of a large number of future observations at x_p:

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

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

CI for predications (contd.)

- Standard deviation of the prediction is minimal at the center of the measured range (i.e., when $x = \bar{x}$); Goodness of the prediction decreases as we move away from the center.



Example

- 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 (contd.)

- 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

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

Example (contd.)

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

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

- 90% CI for a single prediction:

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

Outline

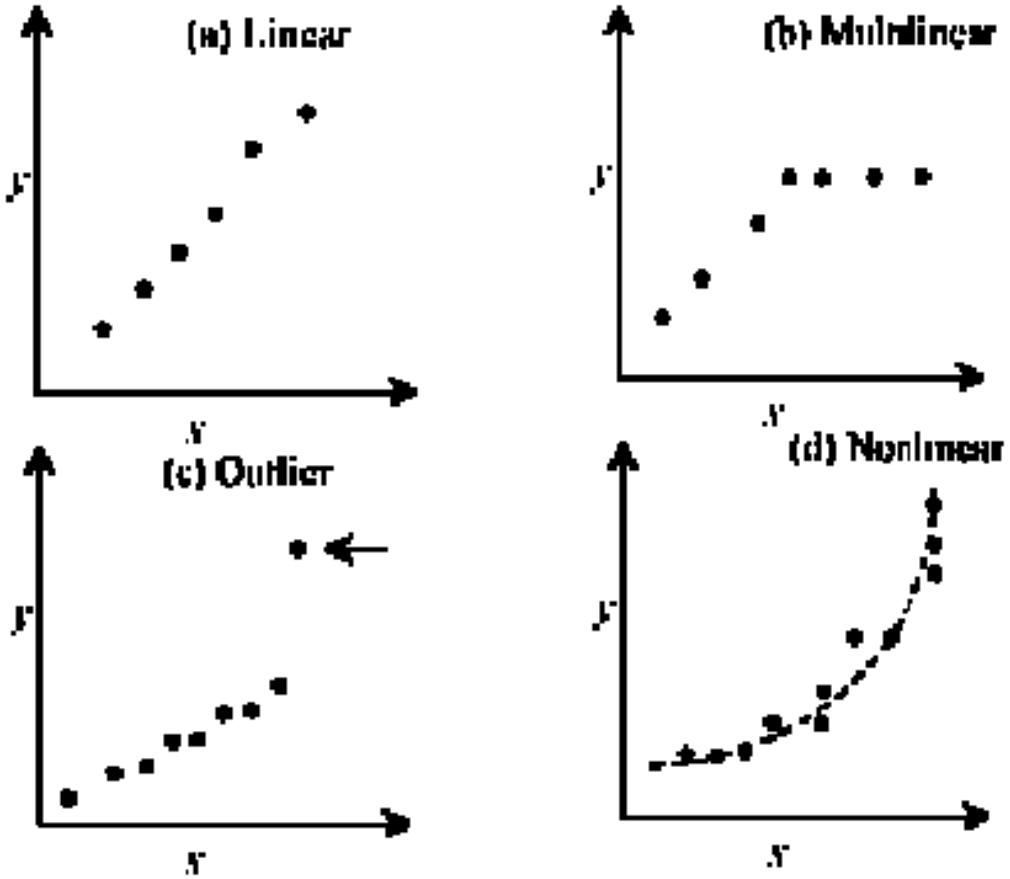
- Definition of a Good Model
- Estimation of Model parameters
- Allocation of Variation
- Standard deviation of Errors
- Confidence Intervals for Regression Parameters
- Confidence Intervals for Predictions
- Visual Tests for verifying Regression Assumption

Visual test for regress assumptions

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

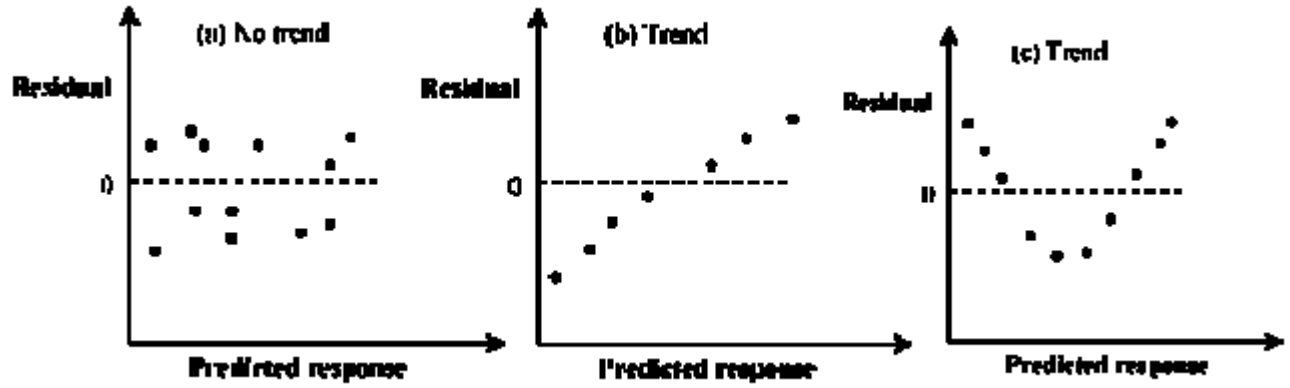
Visual test for linear relationship

- Scatter plot of y versus x $\Rightarrow$ Linear or nonlinear relationship



Visual test for independent errors

- Scatter plot of ε_i versus the predicted response $\hat{y}_i$

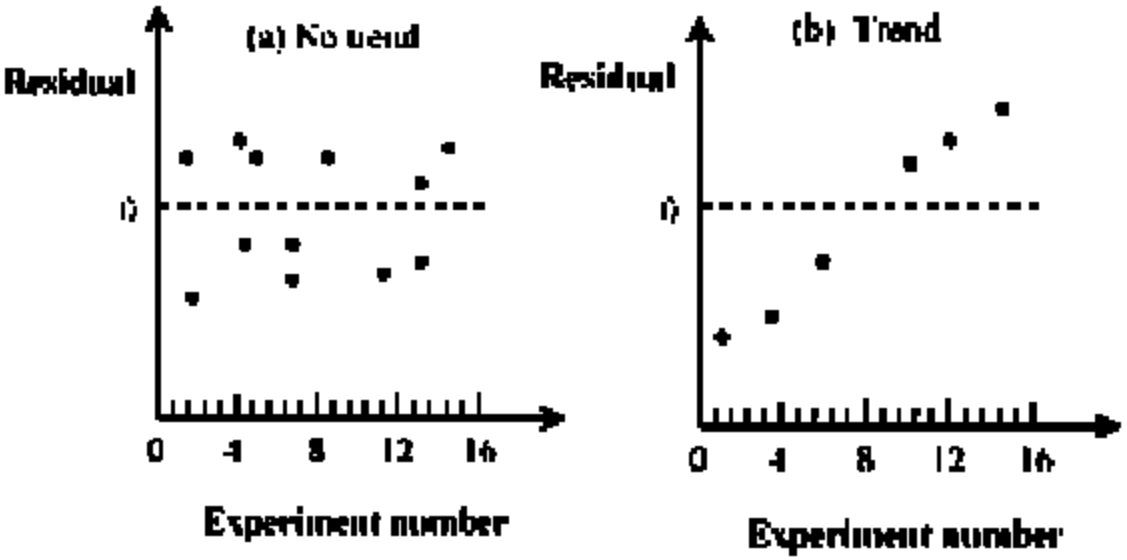


  - Any trend would imply the dependence of errors on predictor variable => curvilinear model or transformation

- In practice, dependence can be proven yet independent cannot

Visual test for independent errors (contd.)

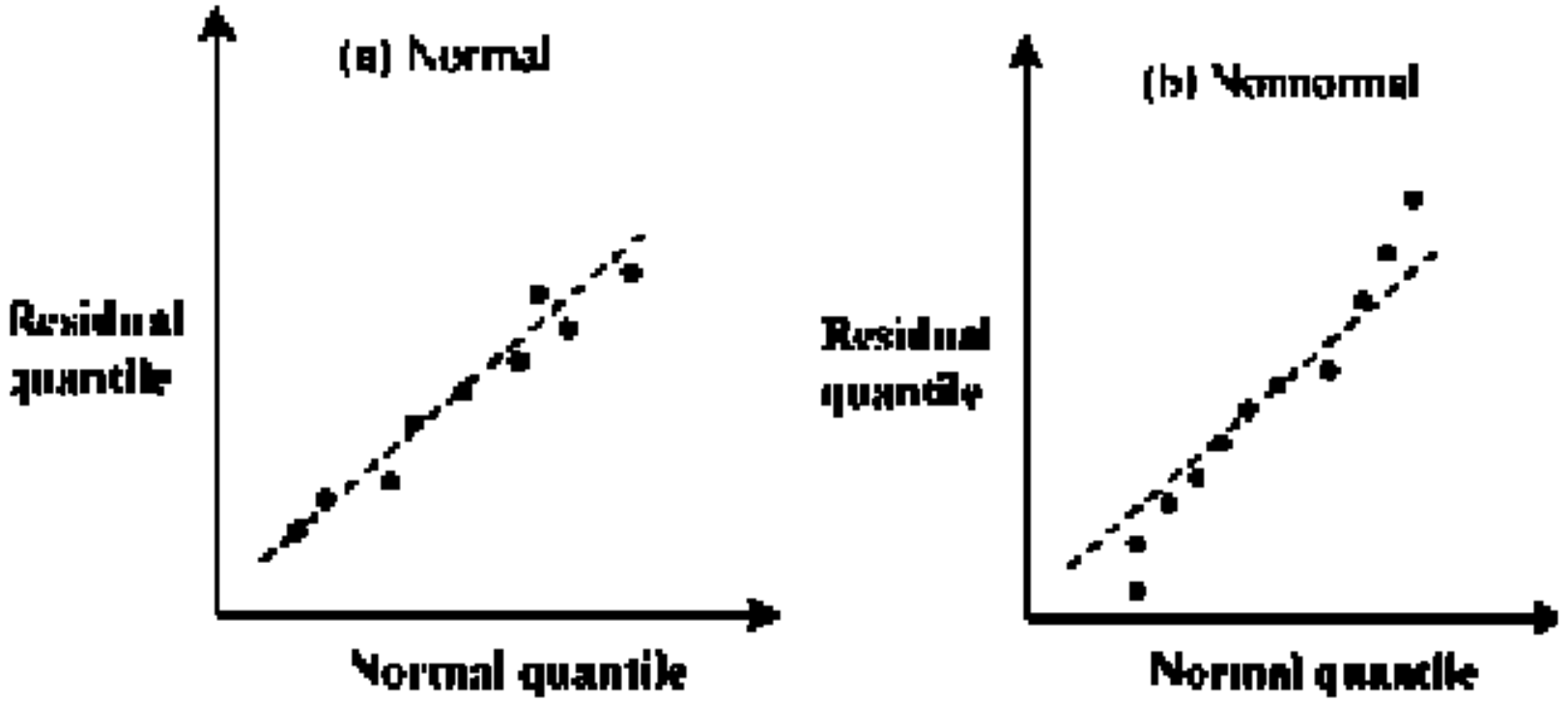
- Plot the residuals as a function of the experiment number



  - Any trend would imply that other factors (such as environmental conditions or side effects) should be considered in the modeling

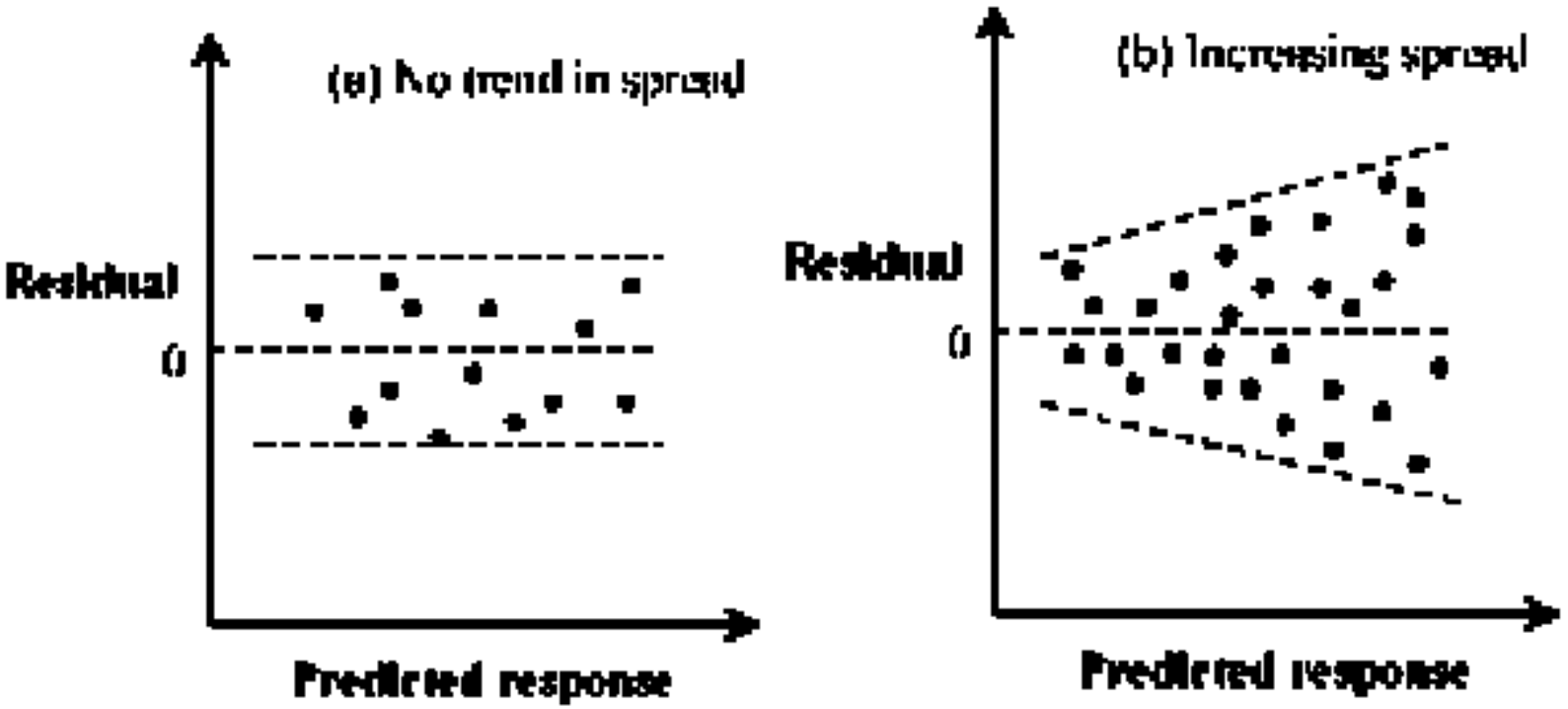
Visual test for “normal distribution of errors”?

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



Visual test for constant standard deviation of errors

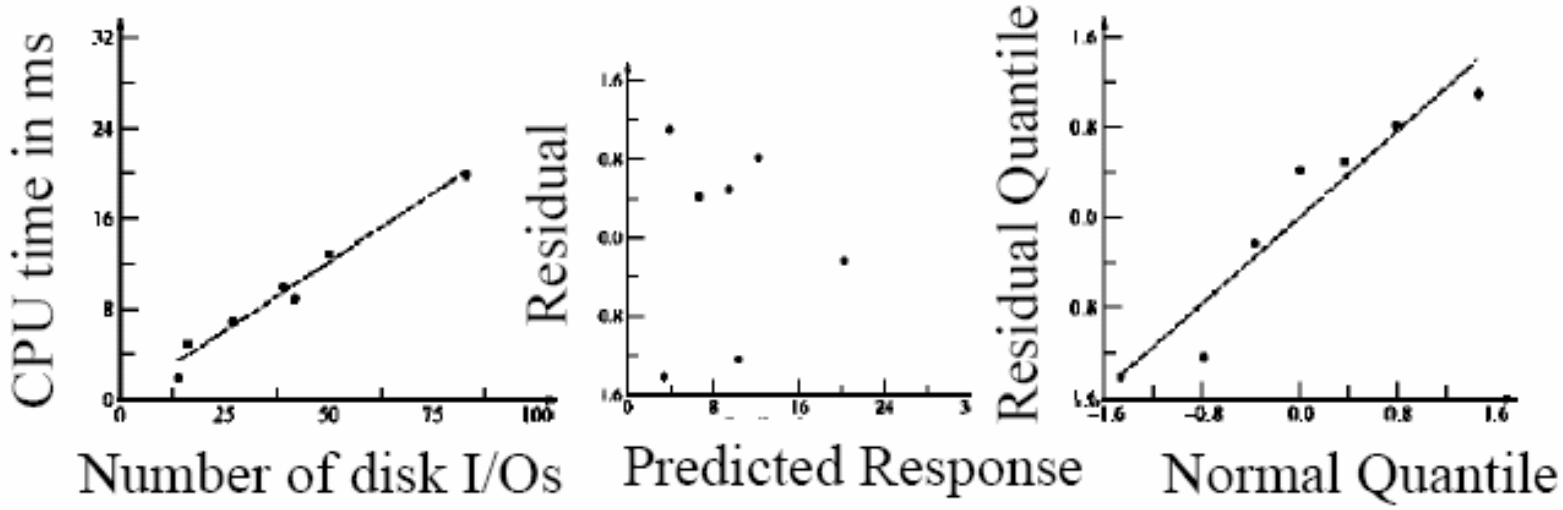
- Also known as **homoscedasticity**



- Trend $\Rightarrow$ Try curvilinear regression or transformation

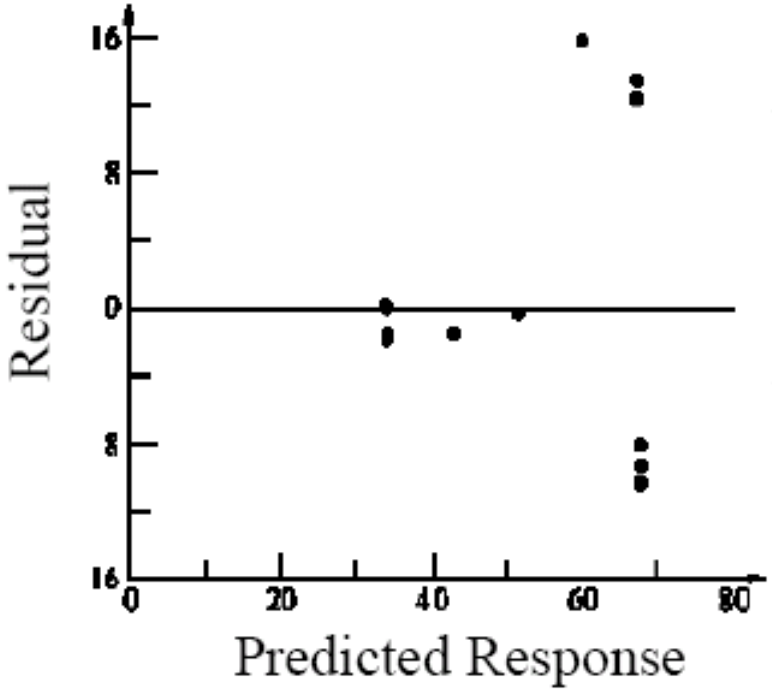
Example

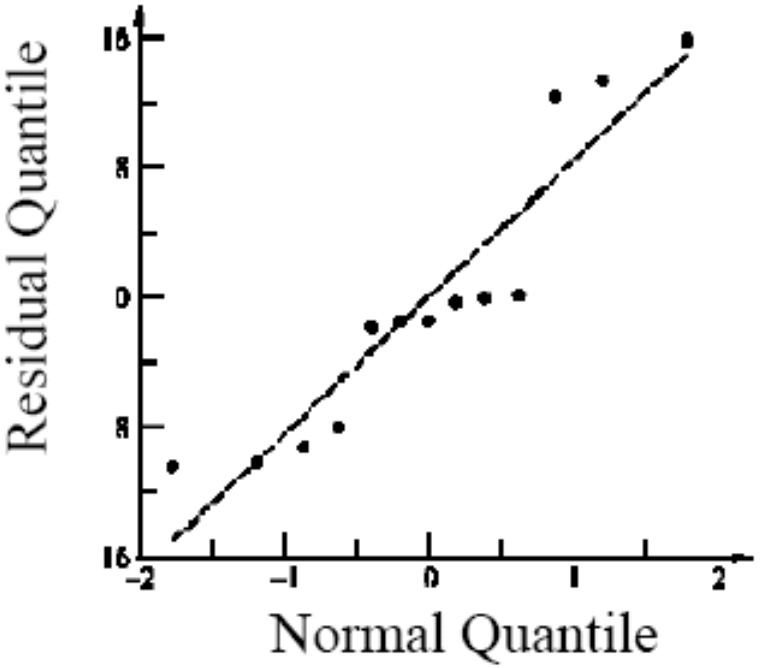
For the disk I/O and CPU time data of Example 14.1



1. Relationship is linear
2. No trend in residuals ⇒ Seem independent
3. Linear normal quantile-quantile plot

Another example: RPC performance





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

Summary

- Definition of a Good Model
- Estimation of Model parameters
- Allocation of Variation
- Standard deviation of Errors
- Confidence Intervals for Regression Parameters & Predictions
- Visual Tests for verifying Regression Assumption

Exercise

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. Use visual tests to verify the regression assumptions.

Record Size	Observations		
	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