Project 4: Linear Regression Models

Implement the following material in R and ipython notebook using markdown language to specify the comments.

1 Simple Linear Regression Models

"There are three kinds of lies: lies, damned lies and statistics."
— Mark Twain

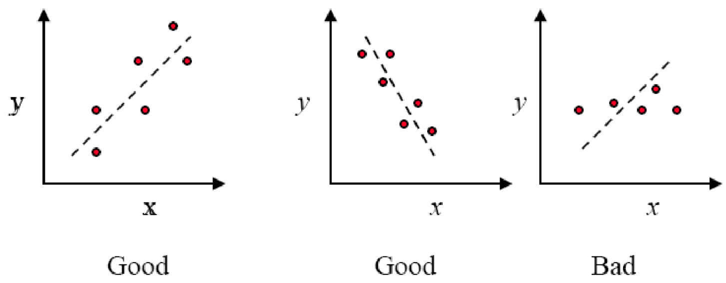
Simple linear regression models

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

2 Definition of a good regression models?



Regression models attempt to fit lines (or curves) to the observation points (data) that minimize the vertical distance between the observation point

and the model line (or curve). The length of this distance is called residual, modeling error, or simply error. The negative and positive errors should cancel out => Zero overall error

It is obvious that many lines will satisfy this criterion.

2.1 Linear Regression Model:

Given n observation pairs $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, the estimated response for the i-th observation is

$\widehat{y}_i = b_0 + b_1 x_i$ where the regression parameters b_0 and b_1 are chosen that minimizes the sum of squares of the errors at the given data (observations).

Formally, the model has the form

$\widehat{y} = b_0 + b_1 x$ where, $\widehat{y}$ is the predicted response when the predictor variable is x.

The error is:

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

The best linear model minimizes the sum of squared errors (SSE), subject to the constraint that the overall mean error is zero:

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

2.2 Linear Regressional Model - the statistical view

Regression analysis is the art and science of fitting straight lines to patterns of data. In a linear regression model, the variable of interest (the so-called “dependent” variable) is predicted from

other variable(s) (the so-called “independent” variable(s)) using a linear equation. If Y denotes the dependent variable, and $X_1, X_2,...,X_\kappa$, are the independent variables, then the assumption is that the value of Y_i in the population is determined by the linear equation $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_\kappa X_{i\kappa} + \varepsilon_i$ where the betas are constants and the epsilons are independent and

identically distributed (i.i.d.) normal random variables with mean zero (the "noise" in the system). β_0 is the so-called intercept of the model—the expected value of Y when all the X's are zero

and β_i is the coefficient (multiplier) of the variable X_i. **The betas together with the mean and standard deviation of the epsilons are the parameters of the model.**

The corresponding equation for predicting Y_i from the corresponding values of the X's is therefore where the b's are estimates of the betas obtained by least-squares, i.e., minimizing the square

prediction error within the sample. <u>Multiple regression allows more than one x variables.</u>

Assumptions

The error terms ε_ι are mutually independent and identically distributed, with mean = 0 and constant variances $E[\varepsilon_\iota] = 0$ $V[\varepsilon_\iota] = \sigma^2$

This is so, because the observations $\mathrm{Y}_1, \Upsilon_2, ..., \Upsilon_\kappa$ are a random sample, they are mutually independent and hence the error terms are also mutually independent.

The distribution of the error term is independent of the joint distribution of X_1, X_2,...,X_κ . The unknown parameters $\beta_0, \beta_1, \beta_2, ..., \beta_\kappa$are constants.

2.2.1 Summary of multiple linear regression model

Independent variables: X_1, X_2,...,X_n

Data: $\{(y_1, x_{11}, x_{21}, ..., x_{k1}), .., (y_n, , x_{n1}, x_{2n}, ..., x_{kn})\}$

Population Model: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_\kappa X_{i\kappa} + \varepsilon_i$ where ε_ι are i.i.d. random variables following the normal disribution $N(0, \sigma)$

Regression coefficients: b_0,b_1,...,b_k are estimates of β_0,β_1,...,β_k

Regression Estimates of Y_i: $\widehat{y_i} = b_0 + b_1 x_{i1} + b_2 x_{i2} + ... + b_\kappa x_{i\kappa}$

Goal: Choose b_0,b_1,...,b_k to minimize the residual sum of squares $\sum_{i=1}^n e^2$= $\sum_{i=1}^n (y_i - \widehat{y_i})^2$

2.2.2 Summary of single variable linear regression model

Assuming that the data is a subset of a population then the linear regression model can be described as follows:

<u>**Data:**</u> $\{(x_1, y_1), \ldots, (x_n, y_n)\}$

<u>**Model of the population**</u>:$y_i = \beta_0$+$\beta_1 x_i + \varepsilon_\iota$

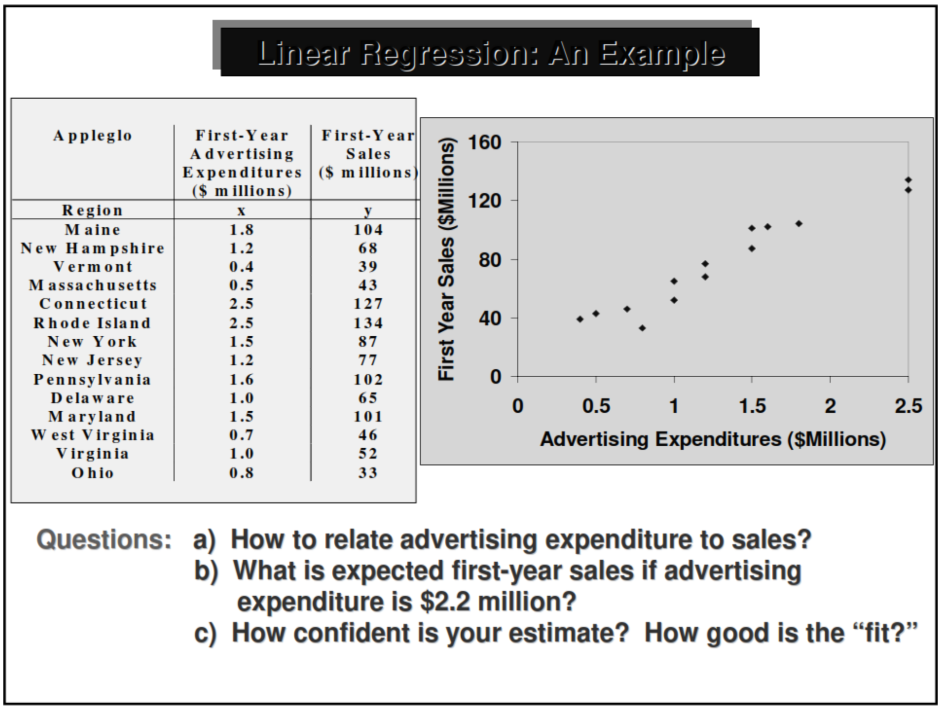
where ε_1,ε_2, ..., ε_n are independent and identically distributed (i.i.d.) random variables, with normal distribution $N(0, \sigma)$

This is the true relation between y and x that depends on the estimation of the unknows β_0 and $\beta_1$based on a sample (data) of the population.

<u>**Comments**</u>:

<u>$E(y_i|x_i)$ =</u>β_0+$\beta_1 x_i$

$SD(y_i|x_i) = \sigma$
Relationship is linear – described by a "line"
β_0= "baseline" value of (i.e., value of y if x is 0)
β_1 = "slope" of line (average change in y per unit change in x)
Prediction regression model:
$\widehat{y}_i = b_0 + b_1 x_i$
where the b's are estimates of the betas obtained by least-squares, i.e., minimizing the square prediction error within the sample.



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

3 Estimation of model parameters

Regression parameters that give minimum error variance are:
$b_1 = \frac{\sum xy - n\overline{x}\overline{y}}{\sum x^2 - n\overline{x}^2}$, $b_0 = \overline{y} - b_1\overline{x}$
where, $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \sum xy = \sum_{i=1}^{n} x_i y_i, \sum x^2 = \sum_{i=1}^{n} x_i^2$

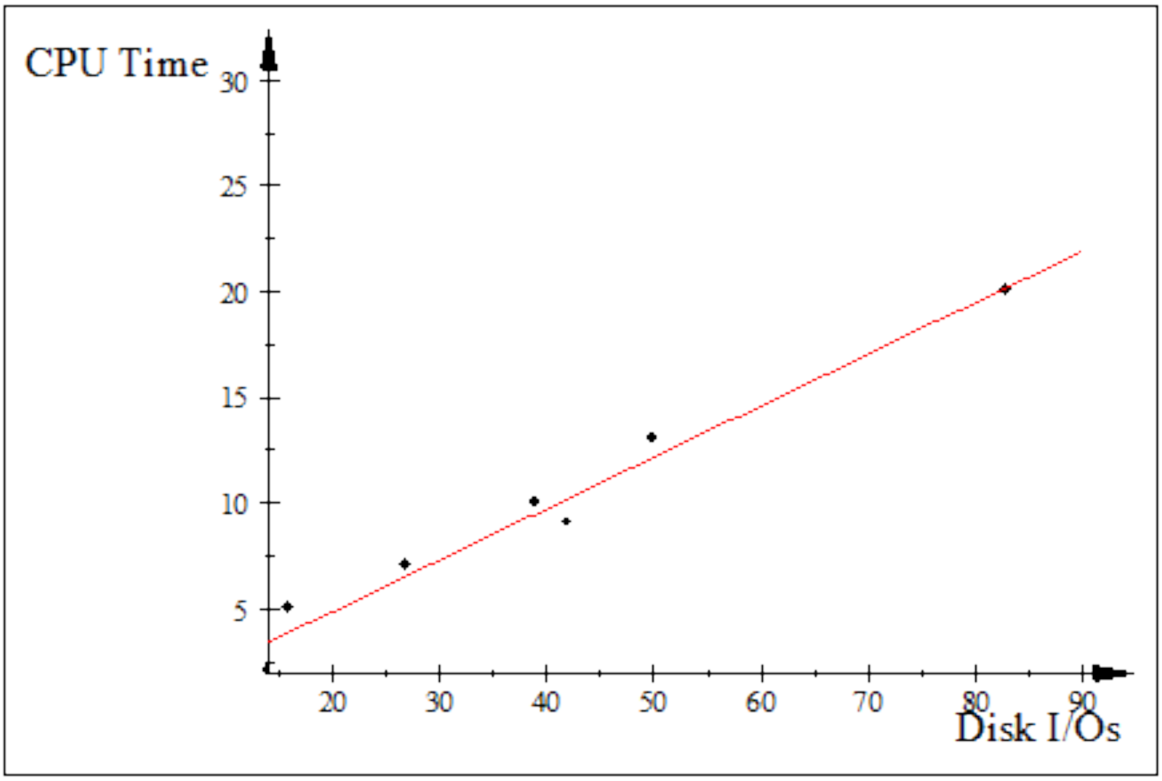
Example 1

The number of disk I/O's and processor time of seven programs were measured as

$$x = \begin{bmatrix} 14 \\ 16 \\ 27 \\ 42 \\ 39 \\ 50 \\ 83 \end{bmatrix} \qquad y = \begin{bmatrix} 2 \\ 5 \\ 7 \\ 9 \\ 10 \\ 13 \\ 20 \end{bmatrix} \qquad x \cdot y : \ 3375 \quad x \cdot x = 13\,855 \ x, \text{ Mean(s):}$$

$\frac{271}{7} = 38.714, \sum_{i=1}^{7} y_i : 66, \sum_{i=1}^{7} y_i^2 : 828, y,$ Mean(s): $\frac{66}{7} = 9.428\,6$

$$\begin{bmatrix} 14 & 2 \\ 16 & 5 \\ 27 & 7 \\ 42 & 9 \\ 39 & 10 \\ 50 & 13 \\ 83 & 20 \end{bmatrix}$$



Polynomial fit: $s = 0.24376t - 8.2824 \times 10^{-3}$

$$x = \begin{matrix} 14 \\ 16 \\ 27 \\ 42 \\ 39 \\ 50 \\ 83 \end{matrix} \quad f(t) = 0.24376t - 8.2824 \times 10^{-3}$$

Error Computation

Disk I/Os	CPU Time	Estimate	Error	Error$^{\mathbf{2}}$
x_i	y_i	$\widehat{y_i} = 0.243\,76x_i - 8.282\,4 \times 10^{-3}$	$e_i = y_i - \widehat{y_i}$	e_i^2
14	2	3.404 4	−1.404 4	1.972 2
16	5	3.891 9	1.108 1	1.227 9
27	7	6.573 2	0.426 76	0.182 13
42	9	10.23	−1.229 6	1.512
39	10	9.498 4	0.501 64	0.251 65
50	13	12.18	0.820 28	0.672 86
83	20	20.224	−0.223 8	$5.008\,5 \times 10^{-2}$
271	66	$\sum_{i=1}^{7}(0.243\,76x_i - 8.282\,4 \times 10^{-3}) = 66.001$	0	$\sum_{i=1}^{7}(y_i - (0.243\,76x_i - 8.282\,4 \times 10^{-3}))^2 = 5.8689$

3.1 Derivation of regression parameters

The error in the ith observation is:

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

For a sample of n observations, the mean error is: $\overline{e} = \overline{y} - b_0 - b_1\overline{x}$

Setting the mean error to zero, we obtain: $b_0 = \overline{y} - b_1\overline{x}$ and $e_i = y_i - \widehat{y} = (y_i - \overline{y}) - b_1(x_i - \overline{x})$

For a sample of n observations the mean error is : $\overline{e} = \frac{1}{n}\sum e_i = \overline{y} - b_0 - b_1\overline{x}$

The sum of squared errors SSE is:

$SSE = \sum_{i=1}^{n} e_i^2 = \sum((y_i - \overline{y})^2 - 2(y_i - \overline{y})b_1(x_i - \overline{x}) + b_1^2(x_i - \overline{x})^2)$

$\frac{SSE}{n-1} = \frac{1}{n-1}\sum(y_i - \overline{y})^2 - \frac{2}{n-1}\sum(y_i - \overline{y})b_1(x_i - \overline{x}) + b_1^2\frac{1}{n-1}\sum(x_i - \overline{x})^2 = s_y^2 - 2b_1 s_{xy}^2 + b_1 s_x^2$

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

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

3.2 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

4 Allocation of variation

Error variance from the sample mean = Variance of the response from the mean value of the observation

Error = $\in_i$= Observed Response - Predicted Response from the mean value $= y_i - \overline{y}$

Variance of Errors from the sample mean $= \frac{1}{n-1}\sum_{i=1}^{n} \in_i^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2 =$ variance of y

Note that the standard error of the model is not the square root of the average value of the squared

errors within the historical sample of data. Rather, the sum of squared errors is divided by $n-1$

rather than n under the square root sign because this adjusts for the fact that a "degree of freedom for error"

has been used up by estimating one model parameter (namely the mean) from the sample of n data points.

The sum of squared errors from the sample mean $SST = \sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2$ is called total sum of squares.

It is a measure of y's variability and is called variation of y. SST can be computed as follows:

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

Where, SSY is the sum of squares of y and SS0 is the sum of squares of $\overline{y}$ and is equal to $n\overline{y}^2$

The difference between SST ans SSE is the sum of squares explained by the regression.

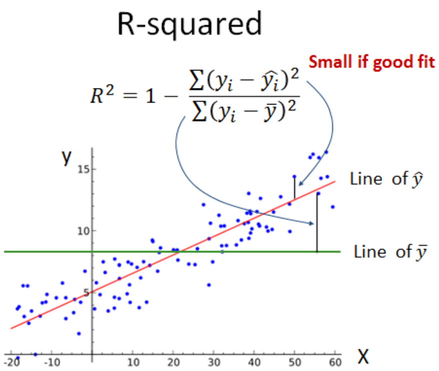
It is called SSR: SSR = SST - SSE or SST = SSR + SSE

The fraction of the variation that is explained determines the goodness of the regression and

it is called the coefficient of determination, $R^2 = \frac{SSR}{SST} = \frac{SST-SSE}{SST} = 1-\frac{SSE}{SST}$.

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: $SSE = \sum y^2 - b_0 \sum y - b_1 \sum xy$.



Example 3

For the disk I/O-CPU time data: $SSE = 5.87$ and $SST = 205.71$ and $SSR = 199.84$ and $R^2 = 0.9715$

The linear 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

5 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)

$S_e^2 = \frac{SSE}{n-2}$

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 $\overline{y}$

SST has $n-1$ degrees of freedom, since one parameter must be calculated from the data before SST can be computed

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

Overall,
$SST = SSY - SS0 = SSR + SSE$
$n-1 = n-1 = 1 + (n-2)$
Notice that the degrees of freedom add just the way the sums of squares do.

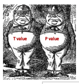
Example

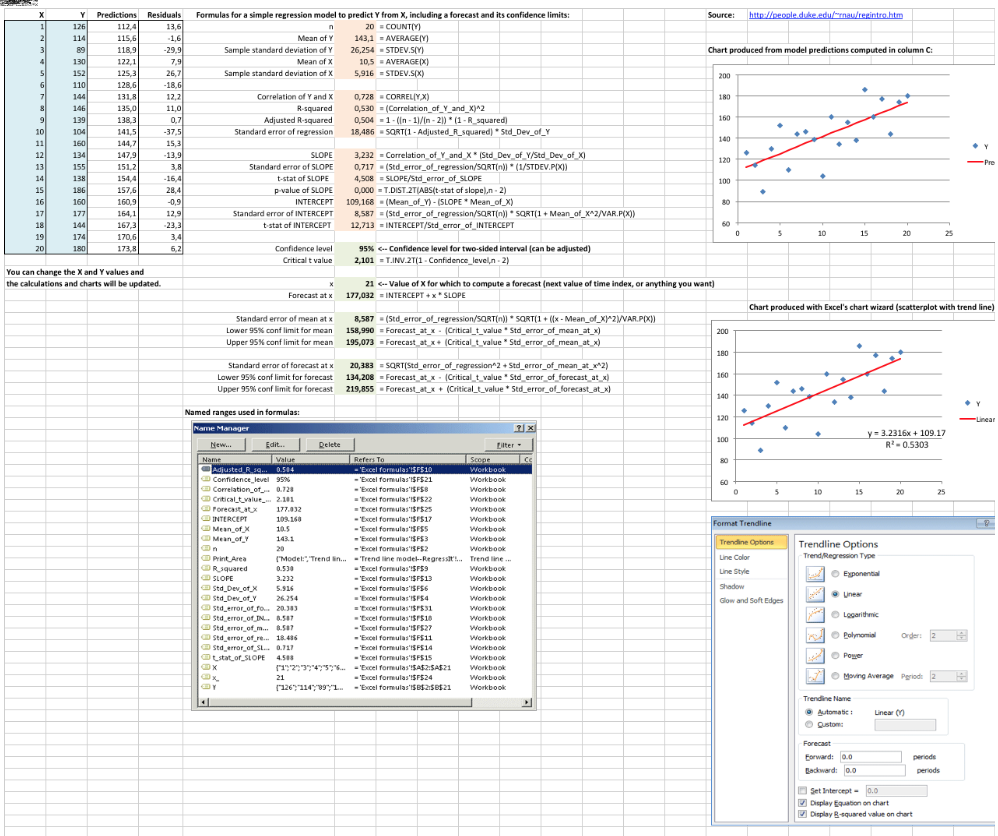
For the disk I/O-CPU data we have
SS: SST(205.71) = SSy(828) - SS0 (622.29) = SSR (199.84) + SSE(5.87)
DF: SST(6) = SSy(7) - SS0 (1) = SSR (1) + SSE(5)
The mean squared error is:
$MSE = \frac{SSE}{DF\ for\ Errors} = \frac{5.87}{5} = 1.174$
The standard deviation of errors is:
$s_e = \sqrt{MSE} = \sqrt{1.174} = 1.083\,5$

Outline

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

6 Regression Statistics





7 CIs for regression parameters

1. Regression coefficients b_0 and b_1 are estimates from a single random sample of size $n \geq 1$.

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 b_0 and b_1 are estimates of β_0 and β_1, respectively.

Sample standard deviation of b_0 and b_1

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

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

The $100(1-a)\%$ confidence intervals for b_0 and b_1 can be computed using $t[1-a/2; n-2]$ — the $1-a/2$ quantile of a t variate with $n-2$ degrees of freedom.

The confidence intervals are:

$b_0 \mp t s_{b_0}$

$b_1 \mp t s_{b_1}$

If a confidence interval includes zero, then the regression parameter cannot be considered different from zero at the

$100(1-a)\%$ confidence level

Example

For the disk I/O and CPU example , we have $n = 7$, $\overline{x} = 38.71$, $\sum x^2 = 13,855$, and $s_e = 1.0834$

1) standard deviations of b_0 and b_1 are

$s_{b_0} = s_e \left[\frac{1}{n} + \frac{\overline{x}^2}{\sum x^2 - n\overline{x}^2}\right]^{1/2} = 0.8311$

$s_{b_1} = \frac{s_e}{[\sum x^2 - n\overline{x}^2]^{1/2}} = 0.0187$

2) For the 0.95-quantile of a t-variate with 5 degrees of freedom is 2.015 $\Rightarrow$ 90% compute the confidence interval for b_0 and b_1

Since, the confidence interval includes zero, the hypothesis that this parameter is zero cannot be rejected at 0.10 significance level => b is essentially zero.

90% Confidence Interval for b_1 is: $0.2438 \mp 0.0376 = (0.2061, 0.2814)$

Since the confidence interval does not include zero, the slope $b_1$is significantly different from zero at this confidence level.

Case study: 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

8_Users_eliashoustis_OneDrive_PSE_2016_simple_regression_ODFVYJ2F.png

9_Users_eliashoustis_OneDrive_PSE_2016_simple_regression_ODFVYJ2G.png

1. Compute the Best linear models for UNIX and ARGUS

Best linear models are:
Time on Unix = 0.030(data size in bytes) + 24
Time on ARGUS = 0.034 (Data size in bytes) + 30

1. Verify that the regressions explain 81% and 75% of the variation, respectively

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

3. 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.

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)

Outline

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

8 CI for predications

$$\widehat{y_p} = b_0 + b_1 x_p$$

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

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

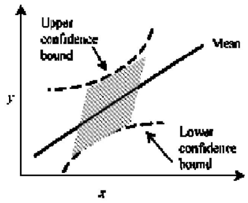
$m = 1 \rightarrow$Standard deviation of a single future observation:

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

$s_{\widehat{y}_{mp}} = s_e \left[\frac{1}{n} + \frac{(x_p - \overline{x})^2}{\sum x^2 - n\overline{x}^2} \right]^{1/2}$

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

Standard deviation of the prediction is minimal at the center of the measured range (i.e., when x = 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, let us estimate the CPU time for a program with 100 disk I/O's

CPU time = 0.0083 + 0.2438(Number of disk I/Os)=24.3674

Standard deviation of errors $s_e = 1.0834$

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

$s_{\widehat{y}_p} = 1.0834 \left[\frac{1}{7} + \frac{(100-38.71)^2}{13855-7(38.71)^2} \right]^{1/2} = 1.215\,6$

From table above, the 0.95-quantile of the t-variate with 5 degrees of freedom is 2.015

$\rightarrow$ 90% CI for the predicted mean $= 24.3674 \mp (2.015)(1.2159) = (21.9174, 26.8174)$

CPU time of a single future program with 100 disk I/O's: $s = 1.6286$ 90% CI for a single prediction:

$24.3674 \mp (2.015)\,(1.6286) = (21.086, 27.6489)$

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

9 Visual test for regress assumptions

Regression assumptions:

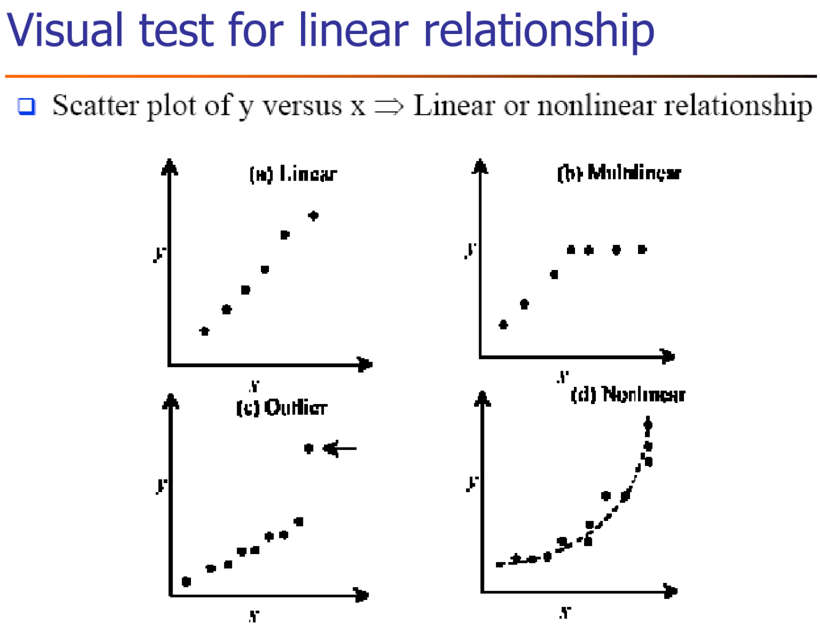
The true relationship between the response variable y and the predictor variable x is linear.

The predictor variable x is non-stochastic and it is measured without any error.

The model errors are statistically independent.

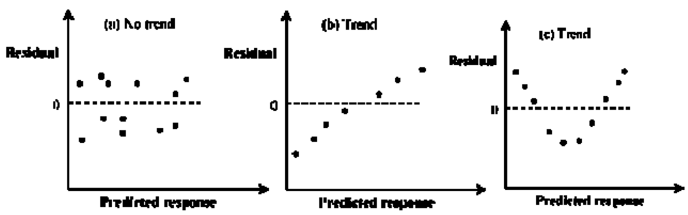
The errors are normally distributed with zero mean and a constant standard deviation.

Visual test for linear relationship



10 Visual test for independent errors

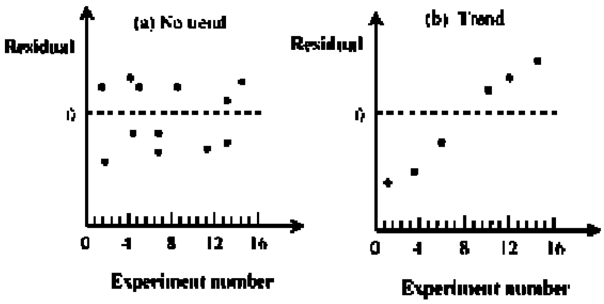
Scatter plot of ε_i versus the predicted response $\widehat{y}_i$



Any trend would imply the dependence of errors on predictor variable $\Rightarrow$ curvilinear model or transformation.

In practice, dependence can be proven yet independence cannot.

Plot the residuals as a function of the experiment number

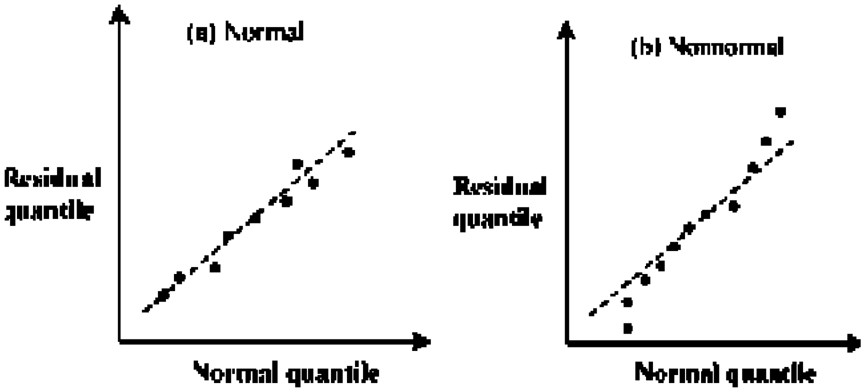


Any trend would imply that other factors (such as environmental conditions or side effects) should be considered in the modeling.

11 Visual test for "normal distribution of errors"?

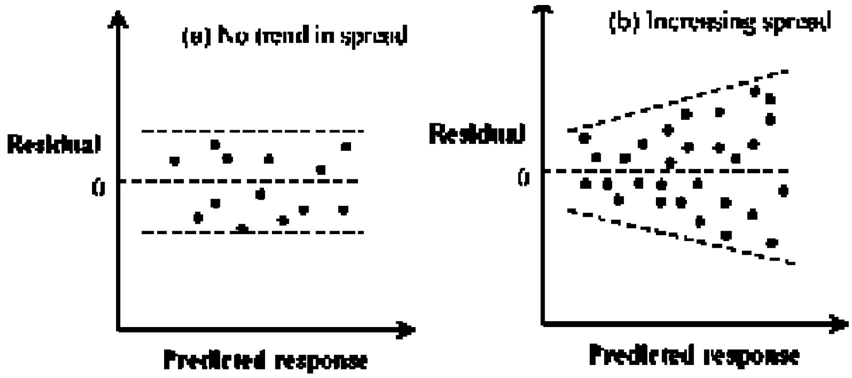
Prepare a normal quantile-quantile plot of errors.

Linear $\Longrightarrow$ the assumption is satisfied



12 Visual test for constant standard deviation of errors

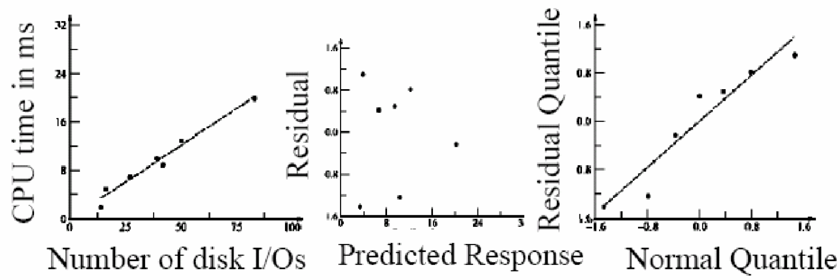
Also known as **homoscedasticity**



Trend$\Longrightarrow$Try curvilinear regression or transformation

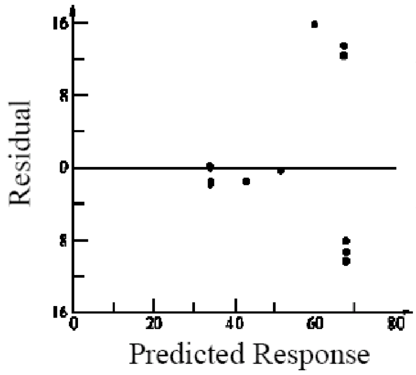
Example

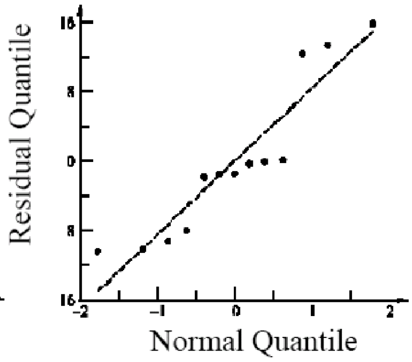
For the disk I/O and CPU time data of Example 14.1



1. Relationship is linear
2. No trend in residuals $\rightarrow$seams 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
- Confidence Intervals for Predictions
- Visual Tests for verifying Regression Assumption

Homework

(100 points) 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