Simple Linear Regression Models

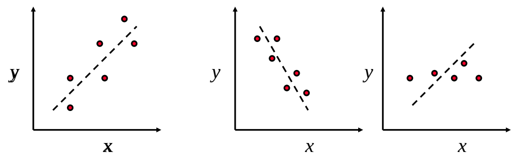
Overview

1. Definition of a Good Model
2. Estimation of Model parameters
3. Allocation of Variation
4. Standard deviation of Errors
5. 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 $\Rightarrow$ 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,

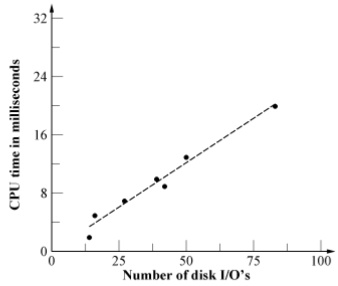
$$b_1 = \frac{\Sigma xy - n\bar{x}\bar{y}}{\Sigma x^2 - n(\bar{x})^2} = \frac{3375 - 7 \times 38.71 \times 9.43}{13,855 - 7 \times (38.71)^2} = 0.2438$$

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

- The desired linear model is:

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

Example 14.1 (Cont)



Example 14. (Cont)

- 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:

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

- 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:

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

Derivation (Cont)

- 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,

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

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:

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

- 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 (Cont)

- 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

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

- 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 (Cont)

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

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

- 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.

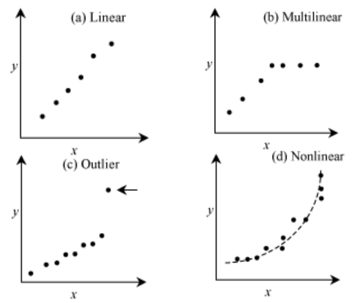
Visual Tests for Regression Assumptions

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.

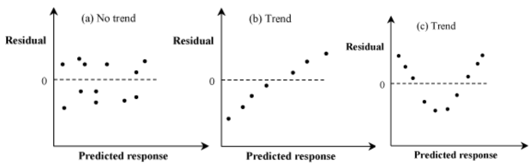
1. Linear Relationship: Visual Test

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



2. Independent Errors: Visual Test

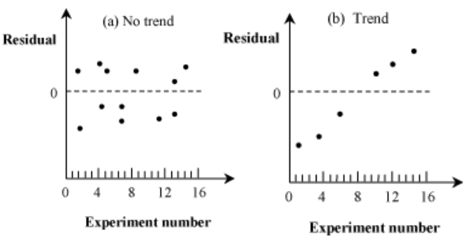
1. Scatter plot of ε_i versus the predicted response $\hat{y}_i$



- All tests for independence simply try to find dependence.

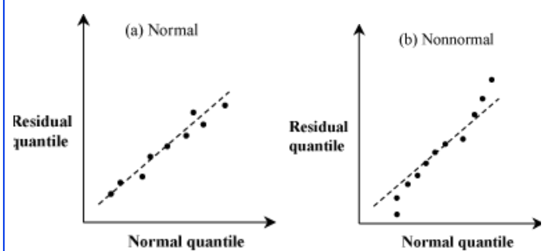
Independent Errors (Cont)

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.



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 $\Rightarrow$ Visual tests