**Linear Least-Squares Regression**

**Simple Regression : Least-Square fit**

When two variables are highly correlated, the points on the scatter plot more or less follow a diagonal

line. The least –square line passes through the mean of the data points.

* *The question is how do we measure* ***the line that best fits*** *the model?*

****The method of least squares assumes that the best-fit curve of a given type is the curve that has the minimal sum of the deviations squared (least square error) from the given data set.

From the diagram, the data points that are not close to the line of best fit are called ***outliers***.

A line that fits the data well therefore makes the residuals ***minimal***.

The question here is how ***small*** should the residual be?

From the simple linear regression model, $y=β\_{0}+β\_{1}x+ε$. This equation is a population regression model whiles this is a $y\_{i}=β\_{0}+β\_{1}x\_{i}+ε\_{i}$ is a sample regression model.

The parameters $β\_{0}$ and $β\_{1}$ are unknown and must be estimated using the sample data and the $ε\_{i}$ is the error term which assumed to have mean zero and the unknown variance$ σ^{2}$. That is $ε\_{i}\~N(0, σ^{2})$.

The slope $β\_{1}$ is the change in the mean of the distribution of $y$ produced by a limit change in $x$. If the range of data on $x$ include $x=0$, then the intercept $β\_{0}$ is the mean of the distribution of the response $y$ when $x=0$. If the range of $x$ does not include zero, the $β\_{0}$ has no practical interpretation.

The errors are uncorrelated to with the fitted values and with each of the $X$s. This means that the value of the error does not depend on the value of any other error.

 Let us consider the following $n $pairs of

data points $\left(x\_{1},y\_{1}\right), \left(x\_{2},y\_{2}\right), …, (x\_{n},y\_{n})$, where $x$ and $y$ are predictors and response variables. The sample regression model is $y\_{i}=β\_{0}+β\_{1}x\_{i}+ε\_{i}$ .

The fitted model from the data points is given by $\hat{y}\_{i}=\hat{β}\_{0}+\hat{β}\_{1}x\_{i}$.

According the least-squares criterion, the best fitting line should have a minimum squared error.

We want to treat the residuals as a continuous differentiable quantity, so we will take the sum of residuals instead of the absolute value.

The error term is given by $ε\_{i}=y\_{i}-\hat{y}\_{i}=y\_{i}-(\hat{β}\_{0}+\hat{β}\_{1}x\_{i}$)

 $deviation= e\_{1}^{2}+e\_{2}^{2}+…+e\_{n}^{2}=\sum\_{i=1}^{n}e\_{i}^{2}= \sum\_{i=1}^{n}(y\_{i}-\hat{y}\_{i})^{2} $

 = ***minimum***.

*Sum of squares cannot be negative. No line can pass perfectly through all of the data points.*

*The residual reflects this fact!*

 **A geometrical interpretation of least Squares**

The diagram below illustrates the $y^{'}=[y\_{1},y\_{1},$…,$ y\_{n}]$ as defining the a vector from the origin to the point $A$. We should note that $y\_{1},y\_{1},$…,$ y\_{n}$ , form a coordinates of an $n-$ dimensional sample space in 3 dimensions. The $X$ matrix consist of $ p (n×1)$ column vectors, for examples, **1** (a column vector of 1’s), $x\_{1},x\_{2},$**…,**$ x\_{k}$ . Each of these columns defines a vector from the origin in the sample space. These $p$ vectors form a $p-$ dimensional subspace called the ***estimation space.***  We may represent any point on the subspace by a linear combination of the vectors $1,x\_{1},x\_{2},$**…,**$ x\_{k}$ . That is to say, any point in the estimation space is of the form$ Xβ.$ Let the vector $Xβ$ determine the point of $B.$

1

3

2

A

C

B

$$\hat{Y}=X\hat{β}$$

$$Xβ$$

$$0$$

$$ Y$$

$$Y-\hat{Y}$$

The least square estimate of the data is the orthogonal projection of the data vector onto the independent variable subspace.

From the diagram,

The squared distance from $B$ to $A$ is given by $S\left(β\right)=\left(y-Xβ\right)^{'}(y-Xβ)$**.** In order to minimize the squared distance of point $A$ defined by the observation vector $y$to the estimation space required finding the point in the estimation space that is closest to $A$. This is achieved when the point $A$ is orthogonal to a point in the model space. From the diagram, the point $C$, is the required point.. This point is defined by a vector $\hat{y}=X\hat{β}$. Therefore we have $y-\hat{y}=y-X\hat{β}$.

This result is orthogonal to the model space, so we have

$X^{'}\left(y-X\hat{β}\right)=0$ , expanding gives

$X^{'}X\hat{β}=X^{'}y$ , $ $

$X'X$ is non-singular if no explanatory variable is a perfect linear function of the others, meaning its inverse can be evaluated,

So we have $\hat{β}=(X^{'}X)^{-1}X'y$ .

That is, minimizing the residual sum of squares leads to the least-squares normal equation in a matrix form .

We illustrate it by a small data set taken from a class notes.

|  |  |
| --- | --- |
| $x$ | $$y$$ |
|  **4** |  **2** |
|  **2** |  **3** |

*The solution will be demonstrated in class*.

**Properties of the least-squares estimators.**

Assumptions on the random error $ε\_{i}$

* $E\left(ε\_{i}\right)=0$, the error is normally distributed with mean zero.
* $V\left(ε\_{i}\right)=σ^{2}$, the error is normally distributed constant variance.
* Mean of predicted values equals mean of response $\frac{1}{n}\sum\_{i=1}^{n}y\_{i}= \frac{1}{n}\sum\_{i=1}^{n}\hat{y}\_{i}$.
* Least squares line always passes through the centroid of data . ie $(\overbar{x},\overbar{y})$
* The residual and predictors are uncorrelated. $\sum\_{i=1}^{n}x\_{i}e\_{i}=0$
* The predicted values and residuals are uncorrelated. $\sum\_{i=1}^{n}\hat{y}\_{i}e\_{i}=0$

**Geometric Approach to obtain the parameters.**

In least squares regressions, we seek that the values of $β\_{o}$ and $β\_{1}$ that minimize

 $S\left(β\_{0},β\_{1}\right)=\sum\_{i=1}^{n}ε\_{i}^{2}=\sum\_{i=1}^{n}(y\_{i}- β\_{o}-β\_{1}X\_{i})^{2}$.

*I will demonstrate how to obtain the parameters by calculus approach in class*.

**Optimality of least squares estimates**

OLS estimates have some strong statistical properties.

Specifically when ;

* The data obtained constitutes a random sample from a well-defined population,
* The population model is linear,
* The error has a zero expected value,
* The independent variables are linearly are independent and
* The error is normally distributed and uncorrelated with the independent variables ( the so called

homoscedasticity assumptions); then the OLS estimate is the best linear unbiased estimate often denoted with the acronym ***“ BLUE”*** . That is has the smallest sample variance .

These five conditions and the proof are called the Gauss-Markov conditions and theorem). In addition, when the Gauss-Markov conditions hold, OLS estimates are also maximum likelihood estimates.

Refereces :

 John Fox (2008), *Applied Regression Analysis and Generalized Linear Models*, second edition.

Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining ,(2006), *Introduction to Linear* *Regression Analysis*, fourth edition.

[http://ancastermath.wikispaces.com/Statistics+2](http://ancastermath.wikispaces.com/Statistics%2B2), (accessed on 04/15/2011)

<http://www.efunda.com/math/leastsquares/leastsquares.cfm>, (accessed on 04/15/2011)

<http://en.wikipedia.org/wiki/Linear_least_squares_%28mathematics%29>, (accessed on 04/15/2011)

[http://en.wikipedia.org/wiki/File:Linear\_least\_squares\_geometric\_interpretation.png](http://en.wikipedia.org/wiki/File%3ALinear_least_squares_geometric_interpretation.png)

 (accessed on 04/15/2011)

<http://serc.carleton.edu/mathyouneed/bestfit.html> (accessed on 04/15/2011)

<http://www.lx.it.pt/~mtf/Figueiredo_Linear_Regression.pdf> (accessed on 04/15/2011)