**STA 471 – Regression Analysis**

Readding assignment: Chapter 0.

## Chapter 1 Fitting a Straight Line by Least Squares

### 1.1 Model and Assumptions

*Regression*: Statistical methods for identifying and fitting relationships between two or more variables, for example, son's height $Y$ and father's height $X$.

— Consider the simplest 2-variable case first.

- Model

Regression model: $Y = f(X) + \varepsilon$.

$Y$ — response (or dependent) variable.
$X$ — predictor (explanatory or independent) variable.
$f(\cdot)$ — functional relationship between $X$ and $Y$ (model function).
$\varepsilon$ — random error.

Generally, $f(\cdot)$ can be any function with known form but unknown parameters.

Examples: $f(X) = \beta_0 + \beta_1 X$ — linear in $X$.
$f(X) = \beta_0 + \beta_1 X + \beta_2 X^2$ — quadratic in $X$.
$f(X) = \alpha(1 - e^{\gamma X})$ — exponential.

The form of $f(\cdot)$ is chosen by investigators based on:
(1) Theoretical relationship between variables.
(2) Approximation of a complex relationship. Ex: Approximate $f(X)$ by a polynomial.

Nature of variables $X, Y$

In measuring/recording bivariate data, one or both variables may be considered measured with error (or subject to random variation).

A. When both variables are measured with error, there are two possible regression relationships that can be fit to the data:
   (1) Model the mean value of $Y$ given a particular value of $X$, $E(Y|X) = f(X)$.
   (2) Model the mean value of $X$ given a particular value of $Y$, $E(X|Y) = g(Y)$.

B. When one variable is measured with error, designate that variable $Y$ and the other $X$. (The practical case is that the random variation of $X$ is negligible.)

Standard regression techniques deal with case B.

The Error Term

— For $X$ considered fixed and $Y$ random, $\varepsilon$ explains why the assumed functional relationship does not hold exactly.
— $\varepsilon$ consists of one or more of
(1) Measurement error,
(2) Error due to natural variation in the population of measurements,
(3) Model misspecification error.

Simple Linear Regression (SLR)

If $f(X) = \beta_0 + \beta_1 X$, the model is called a *simple linear regression* model: $Y = \beta_0 + \beta_1 X + \varepsilon$.

$\beta_0$ — the *intercept*, i.e., the value of $Y$ when $X = 0$.
$\beta_1$ — the *slope* of the line, i.e., the amount by which $Y$ increases when $X$ increases by 1 unit.

If the SLR model is appropriate, then for observations on $(X, Y)$: $(X_1, Y_1)$, $(X_2, Y_2)$, $\cdots$, $(X_n, Y_n)$, i.e.,

$$X\colon\ X_1\ , X_2\ , \cdots\ , X_n$$
$$Y\colon\ Y_1\ , Y_2\ , \cdots\ , Y_n$$

the model becomes: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i\ ,\ \ i = 1, 2, \cdots, n.$

— If $\varepsilon_i = 0$ for all $i = 1, 2, \cdots, n$, then points would fall exactly on a line.

- Assumptions

(1) $E(\varepsilon_i) = 0$, (2) $Var(\varepsilon_i) = \sigma^2$, (3) $\varepsilon_i's$ are independent, (4) $\varepsilon_i$ is normally distributed. (i.e., $\varepsilon_i's$ are iid $N(0, \sigma^2)$.) They are equivalent to

(1') $E(Y_i) = \beta_0 + \beta_1 X_i$, (2') $Var(Y_i) = \sigma^2$, (3') $Y_i's$ are independent, (4') $Y_i$ is normally distributed. (i.e., $Y_i's$ are independent and $Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$.)

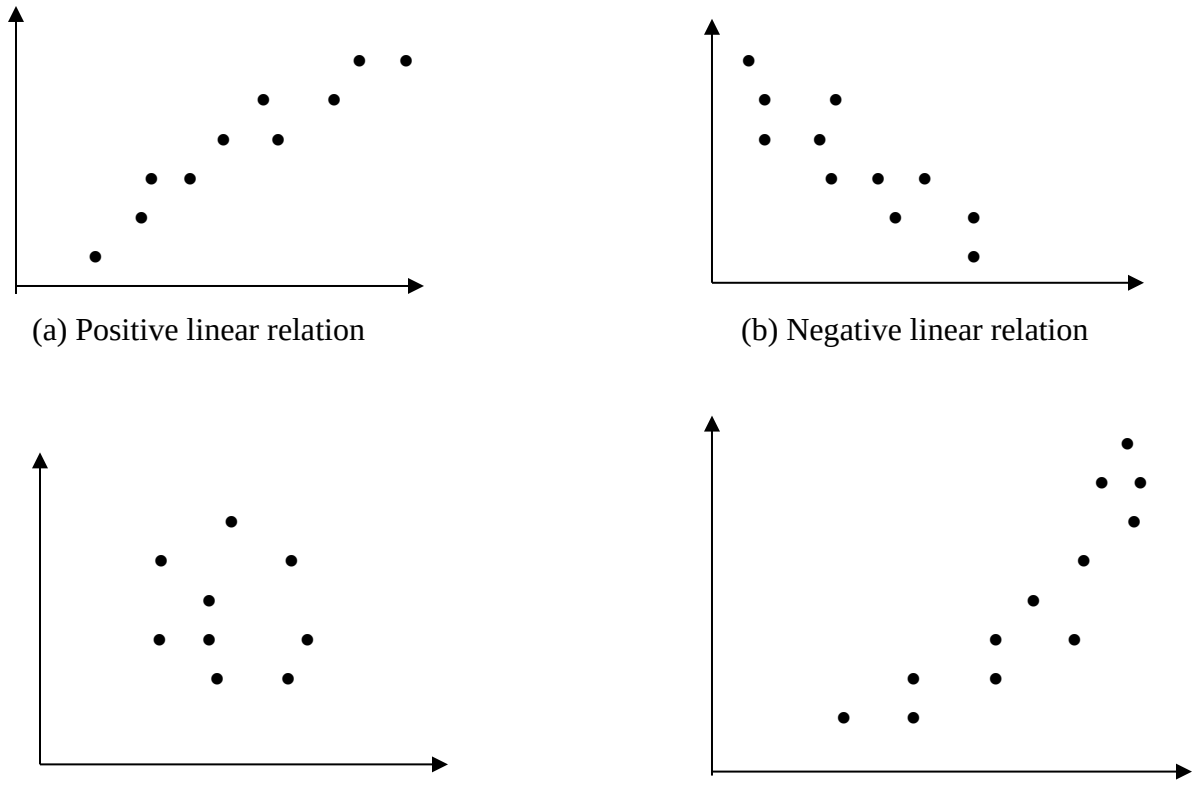
Is it appropriate to fit a bivariate data set by a SLR model?

## 1.2 Scatter Plot and Correlation

- Identify relationship between $X$ and $Y$

A *scatter plot*: A plot of bivariate numerical data in which each observation on $(X, Y)$ is represented as a point on a rectangular coordinate system. It can reveal the relationship between $X$ and $Y$.

A *positive relationship* between $X$ and $Y$: $Y$ increases as $X$ increases.
A *negative relationship* between $X$ and $Y$: $Y$ decreases as $X$ increases.



(a) Positive linear relation (b) Negative linear relation

(c) No relation (d) Positive curved relation

**Figure 1.1** Scatterplots illustrating various types of relationships.

- Correlation

Observations $X$: $X_1, X_2, \cdots, X_n \rightarrow \bar{X}\ s_X$

$Y$: $Y_1, Y_2, \cdots, Y_n \rightarrow \bar{Y}\ s_Y$

$z_X$: $\frac{X_1-\bar{X}}{s_X}, \frac{X_2-\bar{X}}{s_X}, \cdots, \frac{X_n-\bar{X}}{s_X},$

$z_Y$: $\frac{Y_1-\bar{Y}}{s_Y}, \frac{Y_2-\bar{Y}}{s_Y}, \cdots, \frac{Y_n-\bar{Y}}{s_Y},$

where $\bar{X}$ and $\bar{Y}$ are the sample means of $X's$ and $Y's$, respectively, and $s_X$and $s_Y$are the sample standard deviations of $X's$ and $Y's$, respectively.

**Definition 1.1** Pearson's sample correlation coefficient of $X$ and $Y$ is given by

$$r_{XY} = \frac{\sum_{i=1}^{n} z_{X_i} z_{Y_i}}{n-1} = \frac{\sum_{i=1}^{n} \frac{X_i - \bar{X}}{s_X} \frac{Y_i - \bar{Y}}{s_Y}}{n-1}$$

— $r_{XY}$ measures the strength of the linear relationship between $X$ and $Y$.

**Note:** If there is not any relationship between $X$ and $Y$, $r_{XY}$ is close to zero.
Q: Is the converse true?

- Some convenient notations in regression analysis

$S_{XY} = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{n} X_i Y_i - (\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)/n$
$= (n-1)[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})] = (n-1)$(The sample covariance of $X's$ and $Y's$.
Proof:

$S_{XX} = \sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2/n$
$= (n-1)(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2) = (n-1)s_X^2 = (n-1)$(the sample variance of $X's$.
$S_{YY} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n} Y_i^2 - (\sum_{i=1}^{n} Y_i)^2/n$
$= (n-1)(\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2) = (n-1)s_Y^2 = (n-1)$(the sample variance of $Y's$.

Then $r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)\sqrt{s_X^2}\sqrt{s_Y^2}} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{(n-1)s_X^2 \cdot (n-1)s_Y^2}} = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}$

- Properties of $r_{XY}$

(1) $-1 \le r_{XY} \le 1$. The strength of linear relationship based on $r_{XY}$ can be summarized as follows.

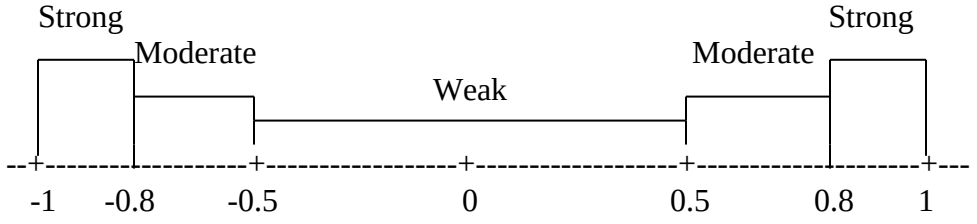



**Figure 1.2** The strength of linear relationship based on *r*

(2) $r_{XY} = 1$ *iff* all points in a scatter plot lie exactly on a straight line that slopes upward. $r_{XY} = -1$ *iff* all points lie exactly on a straight line that slopes downward.
(3) The value of $r_{XY}$ is a measure of the strength of linear relationship between $X$ and $Y$. $r_{XY} = 0$ does not rule out any other strong relationship between $X$ and $Y$.

**Ex. 1.1** The National Institute of Health is studying the relationship between number of cigarettes smoked per day and birthweight of babies born to mothers who smoke cigarettes. The following data are observed.

| No. of cigarettes per day ($X$) | 21 | 12 | 28 | 10 | 24 | 5 |
|---|---|---|---|---|---|---|
| Birthweight ($Y$) | 6.0 | 8.0 | 5.6 | 7.5 | 6.2 | 8.5 |

(1) Construct a scatter plot for this data set. Can a straight line adequately summarize the relationship between $X$ and $Y$? Explain.
(2) Compute the Pearson's sample correlation coefficient $r_{XY}$ and use it to judge the strength of linear relationship between $X$ and $Y$.

### 1.3 Fit a Line to Bivariate Data

In practice, observe $(X_i, Y_i)$ and do not know $\beta_0$ and $\beta_1$.

— look for the best-fitting line.

Least Squares Criterion: Find the line that minimizes the sum of squared deviations of points about the line, i.e.,

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2.$$

— Minimize $S$ as a function of $\beta_0$ and $\beta_1$.

**Notes:** (1) $S(\beta_0, \beta_1)$ is a quadratic function of $\beta_0$ & $\beta_1$.
(2) $S(\beta_0, \beta_1)$ is non-negative for all values of $(\beta_0, \beta_1)$.

As long as $n \geq 2$ and there are at least two distinct values of $X$, then a unique minimum will exist and can be found by differentiating $S$ with respect to $\beta_0$ and $\beta_1$ and setting the partial derivatives to 0. The solution is

$$b_1 = \frac{\sum_{i=1}^n X_i Y_i - (\sum_{i=1}^n X_i)(\sum_{i=1}^n Y_i)/n}{\sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2/n},$$

$$b_0 = \bar{Y} - b_1 \bar{X}$$

— The least-squares (LS) estimators of $\beta_1$ and $\beta_0$.

Details for the solution:

**Notes:** (1) Since $b_1 = \frac{\sum_{i=1}^n X_i Y_i - (\sum_{i=1}^n X_i)(\sum_{i=1}^n Y_i)/n}{\sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2/n} = \frac{S_{XY}}{S_{XX}} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - X)^2} = r_{XY}\frac{s_Y}{s_X}$, $b_1$ and $r_{XY}$ have the same sign.

(2) When the assumptions for the SLR model are satisfied, the LS estimators are most efficient (i.e., the best estimators). However, they are not robust (i.e., they are greatly affected by outlier(s).

(3) Why does the solution give the minimum for $S$? Without (1) and (2), need to check

$$\left| \begin{matrix} \frac{\partial^2 S}{\partial {\beta_0}^2} & \frac{\partial^2 S}{\partial \beta_0 \partial \beta_1} \\ \frac{\partial^2 S}{\partial \beta_0 \partial \beta_1} & \frac{\partial^2 S}{\partial {\beta_1}^2} \end{matrix} \right|_{(\beta_0, \beta_1) = (b_0, b_1)} = \left| \begin{matrix} 2n & 2\sum_{i=1}^n X_i \\ 2\sum_{i=1}^n X_i & 2\sum_{i=1}^n {X_i}^2 \end{matrix} \right| = 4n\sum_{i=1}^n {X_i}^2 - 4(\sum_{i=1}^n X_i)^2$$

$$= 4n[\sum_{i=1}^n {X_i}^2 - (\sum_{i=1}^n X_i)^2/n] = 4n\sum_{i=1}^n (X_i - \bar{X})^2 \geq 0.$$

The *least squares line* (or *sample regression line*): The line fitted by the least-squares method.

$$\hat{Y} = b_0 + b_1 X.$$

Property: The least squares line passes through the point $(\bar{X}, \bar{Y})$.

- Prediction

We can use the least squares line to predict the $Y$ value at an $X$ value. The predicted $Y$ value at $X = X_0$ is $\hat{Y}_0 = b_0 + b_1 X_0$.

**Note:** (Danger of extrapolation) The least squares line should not be used to predict $Y$ values for $X$ values that are much outside the range of $X$ values in the data set since we do not know whether the linear pattern observed continues outside the range.

**Ex. 1.1 (continued)** The National Institute of Health is studying the relationship between number of cigarettes smoked per day and birthweight of babies born to mothers who smoke cigarettes. The following data are observed.

| No. of cigarettes per day($X$) | 21 | 12 | 28 | 10 | 24 | 5 |
|---|---|---|---|---|---|---|
| Birthweight ($Y$) | 6.0 | 8.0 | 5.6 | 7.5 | 6.2 | 8.5 |

(3) Obtain the equation of the least squares line and use it to predict the birthweight of a baby whose mother smokes 25 cigarettes per day. Is it reasonable to use the least squares line to predict the $Y$ value for $X$ = 80? Explain.

(4) Interpret the values of $b_0$ and $b_1$ in the context of the problem.