LINEAR REGRESSION MODEL

Regression analysis is the part of statistics that deals with investigation of the relationship between two or more variables related in a nondeterministic manner.

A familiar example to many students is given by variables

\[ X = \text{high school grade point average (GPA)} \] and
\[ Y = \text{college GPA}. \] The value of \( Y \) cannot be determined
just from knowledge of $X$, and two different students could

have the same $X$ value but have very different $Y$ values.

Yet there is a tendency for those students who have high (low)

high school GPAs also to have high (low) college GPAs.

Knowledge of a student’s high school GPA should be quite

helpful in enabling us to predict how that person will do in college.
* The variable whose value is fixed by the experimenter will be denoted by $x$ and will be called the independent variable.

For fixed $x$, the second variable will be random. We denote this random variable and its observed value by $Y$ and $y$, respectively and refer to it as the dependent variable.

* Assume that the response $Y$ is related to the independent
(input) variable $x$ by

$$Y = \beta_0 + \beta_1 x + e.$$ 

The quantity $e$ is a random variable, assumed to be normally distributed with $\mu = 0$ and $\text{var} = \sigma^2$. 

* Usually $e$ is referred to as random error term in the model.

Note that, without $e$, any observed pair $(x \ y)$ would correspond to a point falling exactly on the line $y = \beta_0 + \beta_1 x$ - called the true regression line.
ESTIMATING THE MODEL

PARAMETERS $\beta_0$ AND $\beta_1$

Assume that we have $n$ pair of observations

$$(x_1, Y_1), \ldots, (x_n, Y_n)$$

such that

* $Y_i = \beta_0 + \beta_1 x_i + e_i,$

for $i = 1, \ldots, n.$

* $x_i$ - are the values of independent variable $x$

(predictor variable)
* $Y_i$ - are the responses corresponding to the i-th experimental run.

* $e_i$ - are unknown error components, $e_i = N(0, \sigma)$.

* the parameters $\beta_0$, $\beta_1$ and $\sigma$ are unknown.

How to estimate these parameters? Let us choose $\beta_0$ and $\beta_1$

such that the sum of squared vertical deviations from the points

$$(x_1, Y_1), \ldots, (x_n, Y_n) \text{ to the line}$$
\[ y = \beta_0 + \beta_1 x \text{ is minimal, i.e} \]

\[ \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 x_i)]^2 = \text{min} \]

Let us denote the corresponding estimators of \( \beta_0 \) and \( \beta_1 \)

by \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \), respectively.

* \( \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \) - is the estimated regression line. Here

\[ \hat{\beta}_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{S_{xy}}{S_{xx}}. \]

* \( \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \)
* The deviations of the observations $Y_i$ from the fitted (estimated) values $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

are called the residuals:

$$\hat{e}_i = Y_i - \hat{Y}_i.$$  

The point estimator of the error variance $\sigma^2$ is

$$S^2 = \hat{\sigma}^2 = \frac{1}{n - 2} \sum_{i=1}^{n} \hat{e}_i^2 := MSE$$

Note that

$$MSE \cdot (n - 2) = SSE = \sum_{i=1}^{n} \hat{e}_i^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

- Sum of Squares due to Error.
INFERENCE CONCERNING THE SLOPE $\beta_1$

In a regression analysis problem, it is of special interest to determine whether the expected response does or not does vary with the magnitude of the input variable $x$. According to the linear regression model,

$$\text{Expected response} = \beta_0 + \beta_1 x.$$ 

This does not change with a change in $x$ if and only if $\beta_1 = 0$. 
We can therefore test the null hypothesis \( H_0 : \beta_1 = 0 \) against

a one- or two- sided alternative, depending of the nature of the

relation that is anticipated.

* Testing hypothesis \( H_0 : \beta_1 = 0 \) against the alternative

\( H_1 : \beta_1 \neq 0 \).

* The test Statistic is

\[
T = \frac{\hat{\beta}_1}{\sqrt{\frac{MSE}{S_{xx}}}}.
\]
The rejection region
\[ R : | T | \geq t_{\alpha/2} \]
is \( \alpha/2 \) upper point of t-distribution
with d.f. = n - 2.