STAT 704 --- Chapter 1: Regression Models

Model: A mathematical approximation of the relationship between two or more real quantities.

- **We have seen several models for a single variable.**
- **We now consider models relating two or more variables.**

Simple Linear Regression Model

- **Involves a statistical relationship between a response variable (denoted *Y*) and a predictor variable (denoted *X*).**

(Also known as

- **Statistical relationship: Not a perfect line or curve, but a general tendency.**
- **Shown graphically with a scatter plot:**

Example:

- **Must decide what is the proper functional form for this relationship. Linear? Curved? Piecewise?**

Statement of SLR Model: For a sample of data (X_1, Y_1), …, (X_n, Y_n):

• **This model assumes *Y* and *X* are**

• **It is also**

Assumptions about the random errors:

• **We assume**

Note: $\beta_0 + \beta_1 X_i$ **is the <u>deterministic component</u> of the model. It is assumed <u>constant</u> (not random).**
ε_i **is the <u>random component</u> of the model.**

Therefore:

Also,

Example (p.11):

(see picture) When $X = 45$, our expected Y-value is 104, but we might observe a Y-value "somewhere around" 104 when $X = 45$.

Note that our model may also be written using matrix notation:

• This will be valuable later.

Estimation of the Regression Function

• In reality, β_0, β_1 are unknown parameters; we can estimate them through our sample data $(X_1, Y_1), \ldots, (X_n, Y_n)$.
• Typically we cannot find values of β_0, β_1 such that for every (X_i, Y_i).
(No line goes through all the points)

Picture:

<u>Least squares method</u>: Estimate β_0, β_1 using the values that minimize the sum of the *n* squared deviations

Goal: Minimize

• Calculus shows that the estimators (call them b_0 and b_1) that minimize this criterion are:

Then $\hat{Y} = b_0 + b_1 X$ is called the <u>least-squares estimated regression line</u>.

• Why are the "least-squares estimators" b_0 and b_1 "good"?

(1)

(2)

Example in book (p. 15)

***X* = age of subject (in years)**

***Y* = number of attempts to accomplish task**

Data:	**X:**	**20**	**55**	**30**
	Y:	**5**	**12**	**10**

Can verify: For these data, the least squares line is

<u>Note</u>: For the first observation, with *X* = 20, the fitted value $\hat{Y}$ = attempts. The fitted value $\hat{Y}$ is an estimator of the

Interpretation:

Interpretation of b_1:

• The <u>residual</u> (for each observation) is the difference between the observed *Y* value and the fitted value:

• The residual e_i is a type of "estimate" of the unobservable error term ε_i.

Note: For the least-squares line,

Proof:

<u>Other Properties of the Least-Squares Line</u>:

• The least-squares line always

Estimating the Error Variance σ^2

• Since var(Y_i) = σ^2 (an unknown parameter), we need to estimate σ^2 to perform inferences about the regression line.

Recall: With a single sample $Y_1,\ldots, Y_n$, our estimate of var(*Y*) was

• In regression, we estimate the mean of *Y* not by
but rather by

• So an estimate of var(Y_i) = σ^2 is

Why n – 2?

E(MSE) =

$s = \sqrt{MSE}$ **is an estimator of**

Pg. 15 example:

(can calculate automatically in R or SAS)

Normal Error Regression Model

• We have found the least-squares estimates using our previously stated assumptions about ε_i.

• To perform <u>inference</u> about the regression relationship, we make another assumption:

Assume ε_i are

• This implies the response values Y_i are

<u>Fact</u>: Under the assumption of normality, our least-squares estimators b_0 and b_1 are also

Why? Likelihood function = product of the density functions for the *n* observations (considered as a function of the parameters)

• When is this likelihood function <u>maximized</u>?

• Assuming the normal-error regression model, we may obtain CIs and hypothesis tests.