MATHEMATICS &
STATISTICS

AUBURN UNIVERSITY

COLLEGE OF
SCIENCES & MATHEMATICS

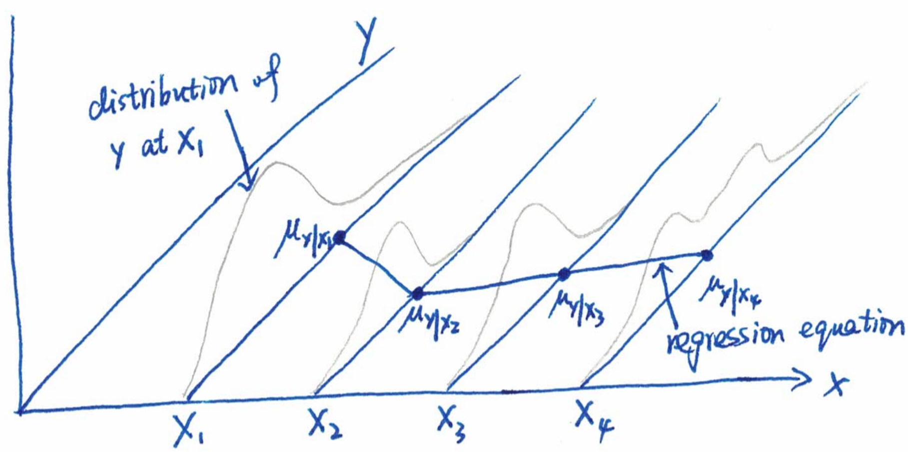
Simple Linear Regression

1. Statement of Assumptions

1.1. Existence : applies to any regression model

For any fixed value of the variable X, Y is a random variable with a certain probability distribution having finite mean $\mu_{y|x}$ and variance $\sigma^2_{y|x}$

y|x : mean and variance of the r.v. Y depend on X.



1.2. Independence

The Y-values are statistically independent of one another.

violated examples : different observations are made on the same individual at different times. (longitudinal data).

mixed model → involving repeated measurements

multivariate linear model

generalized estimating equations (GEE) → analyzing correlated response data.

221 Parker Hall, Auburn, AL 36849-5310; Telephone: 334-844-4290; Fax: 334-844-6555

w w w . a u b u r n . e d u

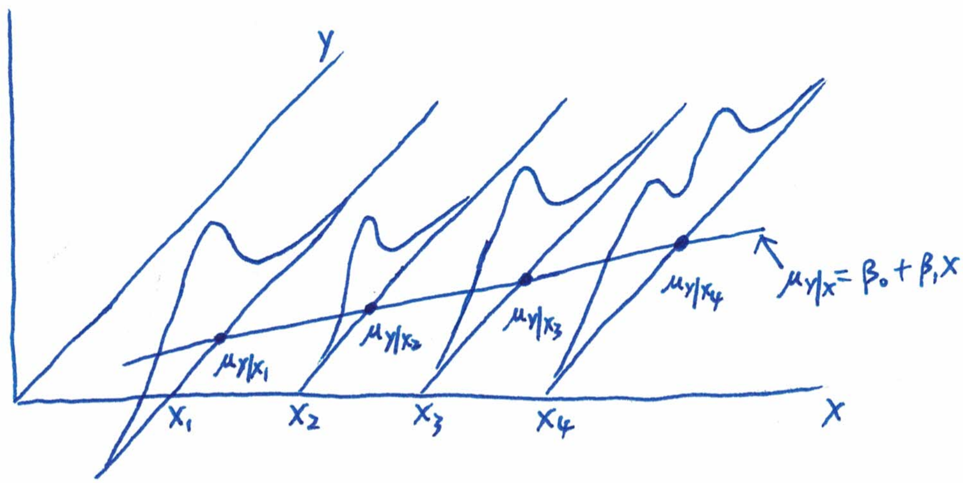
MATHEMATICS &
STATISTICS

AUBURN UNIVERSITY

COLLEGE OF
SCIENCES & MATHEMATICS

1.3 Linearity

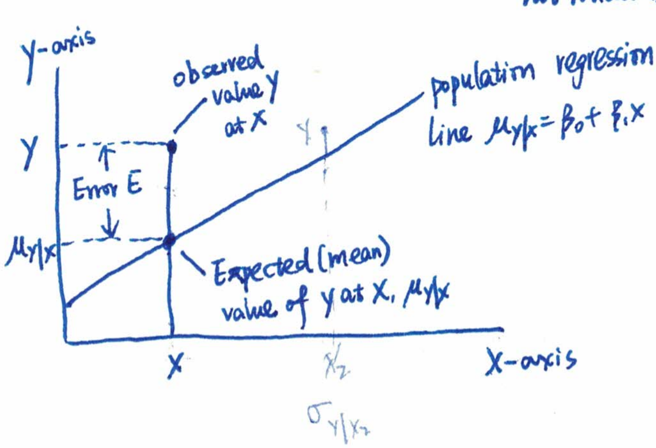
The mean value of Y, $\mu_{Y|X}$, is a straight-line function of X.



$$\mu_{Y|X} = \underset{\text{intercept}}{\beta_0} + \underset{\text{slope}}{\beta_1} X$$

$Y = \beta_0 + \beta_1 X + E$ (ε), E is a r.v. with mean 0 at fixed X.
i.e. $\mu_{E|X} = 0$ for any X.

Y: r.v.; $\beta_1 X$: fixed, not random; E: r.v., the error component.



The variable E describes how distant an individual's response can be from the population regression line.

What we observe at a given X (namely Y) is in error from that expected on the average (namely $\mu_{Y|X}$) by an amount E, which is random and varies from individual to individual.

221 Parker Hall, Auburn, AL 36849-5310; Telephone: 334-844-4290; Fax: 334-844-6555

www.auburn.edu

MATHEMATICS &
STATISTICS

AUBURN UNIVERSITY

COLLEGE OF
SCIENCES & MATHEMATICS

1.4 homoscedasticity

The variance of Y is the same for any X.

example of violation: Page 2. $Y|X_1$ vs $Y|X_2$, $Y|X_1$ more spread $\sigma^2_{Y|X_1} > \sigma^2_{Y|X_2}$

$$\sigma^2_{Y|X} \equiv \sigma^2$$

1.5. Normal distribution

For any fixed value of X, Y has a normal distribution.

This assumption makes it possible to evaluate the statistical significance (eg. C.I., testin

Violation: Y transformation. eg: $\log(Y)$, $\sqrt{Y}$, ... box-cox procedure, ...

Summary

Maintain distinctions among random variables, parameters, and point estimates.

Y: r.v.

X: assumed to be measured without error. fixed value.

β_0, β_1: parameters (unknown) for a population.

E/ε random, unobservable ~~unobserved~~ variable.

$\hat{\beta}_0$: point estimate for β_0

$\hat{E} = Y - \hat{Y} = Y - (\hat{\beta}_0 + \hat{\beta}_1 X)$: point estimate of E. at value X. $\rightarrow$ residual

Normal distribution = Gaussian distribution.

221 Parker Hall, Auburn, AL 36849-5310; Telephone: 334-844-4290; Fax: 334-844-6555

www.auburn.edu

2. Estimate the coefficients

$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ n observation pairs

Goal: find $\hat{\beta}_0, \hat{\beta}_1$ such that the resulting line is as close as possible to the data points

different measures of closeness

2.1 The Least-squares Method

$$(\hat{\beta}_0, \hat{\beta}_1)_{LS} = \underset{\beta_0, \beta_1}{\text{argmin}} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \text{argmin} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

$$= \text{argmin} \sum_{i=1}^{n} e_i^2$$

$e_i = y_i - \hat{y}_i$: i^{th} residual

$\sum_{i=1}^{n} e_i^2$: residual sum of squares (RSS), or sum of squares about the regression line

sum of squares due to error (SSE), Error sum of squares (SSE)

2.2. The minimum-variance method.

$\hat{\beta}_0, \hat{\beta}_1$: unbiased for β_0, β_1, and have minimum variance among all unbiased estimators.

2.3. Maximum Likelihood method.

$\hat{\beta}_0, \hat{\beta}_1$: maximize the likelihood function.

$$L(Y; \beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} f(y_i; \beta_0, \beta_1, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 X)]\right.$$

Under the assumption of Gaussian and mutual independence of Y_i's, maximum-likelihood estimates, LS estimates, and minimum-variance estimates are all the same.

2.4 LS estimates

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_i) \right)^2 = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)$$

The value of β_0, β_1 that minimize $Q(\beta_0, \beta_1)$ satisfy:

$$\frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_0} = 0, \qquad \frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_1} = 0$$

This leads to the normal equations:

$$\begin{cases} n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases}$$

useful identities:

$$\sum (x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = \sum (x_i - \bar{x}) y_i = \sum x_i y_i - n\bar{x}\bar{y}$$

$$\Rightarrow \begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 = \dfrac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = r_{xy} \cdot \dfrac{S_y}{S_x} \end{cases}$$

The fitted regression line (LS line):

$$y = \hat{\beta}_0 + \hat{\beta}_1 x = \bar{y} + \hat{\beta}_1 (x - \bar{x})$$, pass through the point $(\bar{x}, \bar{y})$ — center of data

The fitted value for the i^{th} case:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = \bar{y} + \hat{\beta}_1 (x_i - \bar{x}), \quad i = 1, 2, \cdots, n$$

Residual $e_i = y_i - \hat{y}_i = y_i - \bar{y} - \hat{\beta}_1 (x_i - \bar{x})$, an "estimator" of the error term ε_i

Properties of residuals:

1) $\sum_{i=1}^{n} e_i = 0$ $\quad \sum (y_i - \bar{y}) - \hat{\beta}_1 \sum (x_i - \bar{x}) = 0$

2) $\sum_{i=1}^{n} x_i e_i = 0$ $\quad \sum x_i (y_i - \bar{y}) - \hat{\beta}_1 \sum x_i (x_i - \bar{x}) = \sum x_i y_i - n\bar{x}\bar{y} - \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\cancel{\sum x_i^2 - n\bar{x}^2}} \cdot (\cancel{\sum x_i^2 - n\bar{x}^2})$

3) $\sum_{i=1}^{n} \hat{y}_i e_i = 0$ $\quad \sum (\hat{\beta}_1 (x_i - \bar{x}) + \bar{y}) \cdot e_i = \sum \bar{y} e_i + \hat{\beta}_1 (\sum x_i e_i - \bar{x} \sum e_i)$

221 Parker Hall, Auburn, AL 36849-5310; Telephone: 334-844-4290; Fax: 334-844-6555

MATHEMATICS &
STATISTICS

AUBURN UNIVERSITY

COLLEGE OF
SCIENCES & MATHEMATICS

2.5. Estimate of Error Variance.

$$Var(\varepsilon_i) = \sigma^2$$

$$\sum (y_i - \bar{y} - \hat{\beta}_1(x_i - \bar{x})) = \sum (y_i - \bar{y})^2 + \hat{\beta}_1^2 \sum (x_i - \bar{x})^2 - 2\hat{\beta}_1 \underbrace{\sum (x_i - \bar{x})(y_i - \bar{y})}_{= \hat{\beta}_1 \sum (x_i - \bar{x})^2}$$

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 - \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$E(SSE) = (n-2)\sigma^2$$

Mean squared error (MSE), unbiased estimator of σ^2.

$$MSE = \frac{SSE}{n-2}, \quad E(MSE) = \sigma^2$$

Analogy

usually $var = \frac{1}{n-1} \sum (y_i - \bar{y})^2$ (μ)

y_i's are indep. with the same mean μ and variance σ^2

$n-1$: to estimate σ^2, we need to estimate μ first.

our case: the population mean $\mu_{y|x}$ changes with X.

$\hat{y}$ is the estimate of $\mu_{y|x}$, so we have $(y_i - \hat{y}_i)^2$

$n-2$: we need to estimate β_0, β_1 first $\qquad \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

3. Properties of LS Estimators

3.1. LS estimators are linear functions of the responses y_is.

$$\hat{\beta}_1 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{\sum_{j=1}^{n} (x_j - \bar{x})^2} y_i = \sum_{i=1}^{n} k_i y_i$$

$$\hat{\beta}_0 = \sum_{i=1}^{n} \left(\frac{1}{n} - \bar{x} k_i\right) y_i$$

$\hat{Y}_i$ and e_i are also linear functions of Y_i

221 Parker Hall, Auburn, AL 36849-5310; Telephone: 334-844-4290; Fax: 334-844-6555

www.auburn.edu