743- Regression and Time Series

Mamikon S. Ginovyan

Simple Linear Regression Model

Two-Variable Regression Model

Statistical Inferences

The general **Statistical Inference** procedure of the **least-squares regression model** with

one dependent (random) variable (Y), and
one independent (non-random) variable (x)

involves the following steps:

1. **Specification of the underlying model.**

2. Point estimation of the unknown regression parameters

3. Properties and distributions of the point estimators.

Two-Variable Regression Model

4. **Construction of confidence intervals for the regression unknown parameters, and prediction intervals for the future values.**

5. **Hypotheses testing about the model.**

6. **A measure of the fit of the regression model (i.e., how good the model describes the data?).**

Two-Variable Regression Model

1. Specification of the two-variable regression model.

The general two-variable regression model is given by the equation

$$Y = f(x) + e, \qquad (1)$$

where e $(E[e]=0)$ is the **random** component,

x is the independent **(non-random)** variable,

Y is the dependent **(random)** variable,

f(x) is the **deterministic** component.

Two-Variable Regression Model

If we denote by $f_{Y|x}(y)$ the **conditional** ***pdf*** of the random variable ***Y*** for given value of ***x***,

and by $E[Y|x]$ the expected value associated with ***pdf*** $f_{Y|x}(y)$ then

$$y = f(x) = E[Y|x]$$

is the **regression** of ***Y*** on ***x***.

Two-Variable Regression Model

- **Example 1.**

Suppose that corresponding to each value of $\boldsymbol{x}$ in the interval $0 \le x \le 1$ the conditional ***pdf*** of a RV $\boldsymbol{Y}$ is given by

$$f(y) = f_{Y|x}(y) = \frac{x+y}{x+1/2}, \quad 0 \le y \le 1; \quad 0 \le x \le 1. \qquad (2)$$

Find and graph the regression curve of $\boldsymbol{Y}$ on $\boldsymbol{x}$.

Example 1

Solution.

Observe, first, that for any $x \in [0,1]$, $f(y)$ given by (2) is indeed a ***pdf***.

1. $f(y) \geq 0$ for all $0 \leq y \leq 1$ and $0 \leq x \leq 1$.

2. $$\int_0^1 f(y)dy = \int_0^1 \frac{x+y}{x+1/2}dy$$

$$= \frac{x}{x+1/2} \cdot y\Big|_0^1 + \frac{1}{x+1/2} \cdot \frac{y^2}{2}\Bigg|_0^1$$

$$= \frac{x}{x+1/2} + \frac{1}{x+1/2} \cdot \frac{1}{2} = \frac{x+1/2}{x+1/2} = 1.$$

Example 1

Next,

$$E[Y|x] = \int_0^1 y \cdot f(y)dy = \int_0^1 y \cdot \frac{x+y}{x+1/2}dy$$

$$= \left[\frac{xy^2}{2(x+1/2)} + \frac{y^3}{3(x+1/2)} \right]\Bigg|_0^1 = \frac{3x+2}{6x+3}, 0 \le x \le 1.$$

Thus, the **regression curve** of ***Y*** on ***x*** is given by

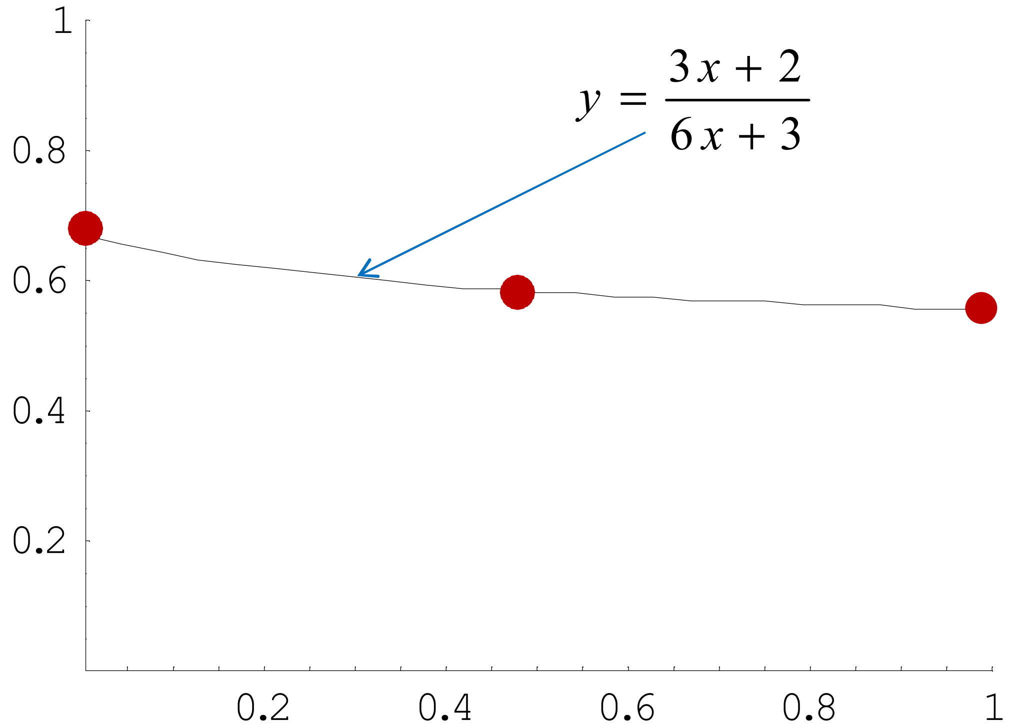
$$y = f(x) = E[Y|x] = \frac{3x+2}{6x+3}, \quad 0 \le x \le 1.$$

Example 1

$$E[Y|x=0] = \frac{2}{3}$$

$$E[Y|x=\frac{1}{2}] = \frac{7}{12}$$

$$E[Y|x=1] = \frac{5}{9}$$



Simple Linear Regression Model

We will consider the special case of the two-variable regression model, called **simple linear model**, where the relationship between $\boldsymbol{Y}$ and $\boldsymbol{x}$ is **linear.**

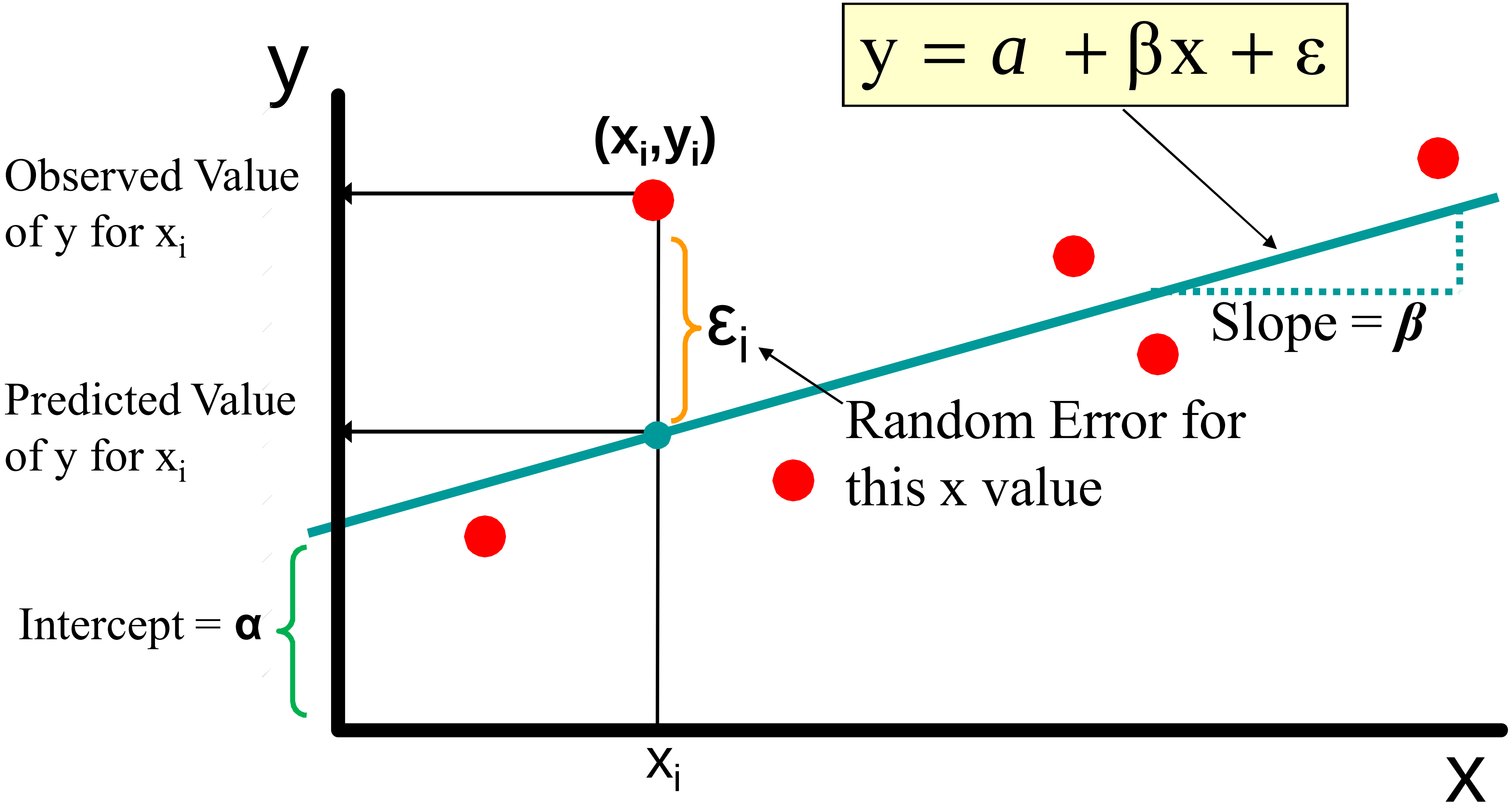
This model is described by equation

$$Y_i = a + b\,x_i + e_i, \quad i = 1, 2, \ldots, n, \qquad (3)$$

where a and b **are unknown (regression)** parameters.

The linear model (3) is completely specified by the following.

Simple Linear Regression Model



Assumptions - SLRM

Assumptions (imposed on the RV's $e_i, i = \overline{1, n}$).

1. $E[e_i] = 0$ for all $i = \overline{1, n}$.

2. $E[e_i^2] = s^2$ for all $i = \overline{1, n}$.

3. The RV's $e_i, i = \overline{1, n}$, are **independent**, implying that $E[e_i e_j] = 0, i \neq j$.

4. The RV's e_i are **normally distributed**:

$$e_i \sim N(0, s^2), i = \overline{1, n}.$$

Assumptions - SLRM

In terms of RV's $Y_i, i = \overline{1, n},$

the model can be specified as follows:

1'. $E[Y \mid x_i] = a + b\, x_i, \ \ i = \overline{1, n}.$

2'. $Var[Y \mid x_i] = s^2, \ \ i = \overline{1, n}.$

3'. The RV's $Y_i, \ \ i = \overline{1, n},$ are **independent**.

4'. $Y_i \sim N(a + b\, x_i, s^2),$

Assumptions - SLRM

that is, the **conditional** ***pdf*** $f_{Y|x_i}(y)$ is given by

$$f_{Y|x_i}(y) = \frac{1}{\sqrt{2p} \cdot s} e^{-\frac{(y-a-b\,x_i)^2}{2s^2}}, \quad i = \overline{1,n}. \qquad (4)$$

These assumptions are illustrated in Fig.1.

Assumptions - SLRM

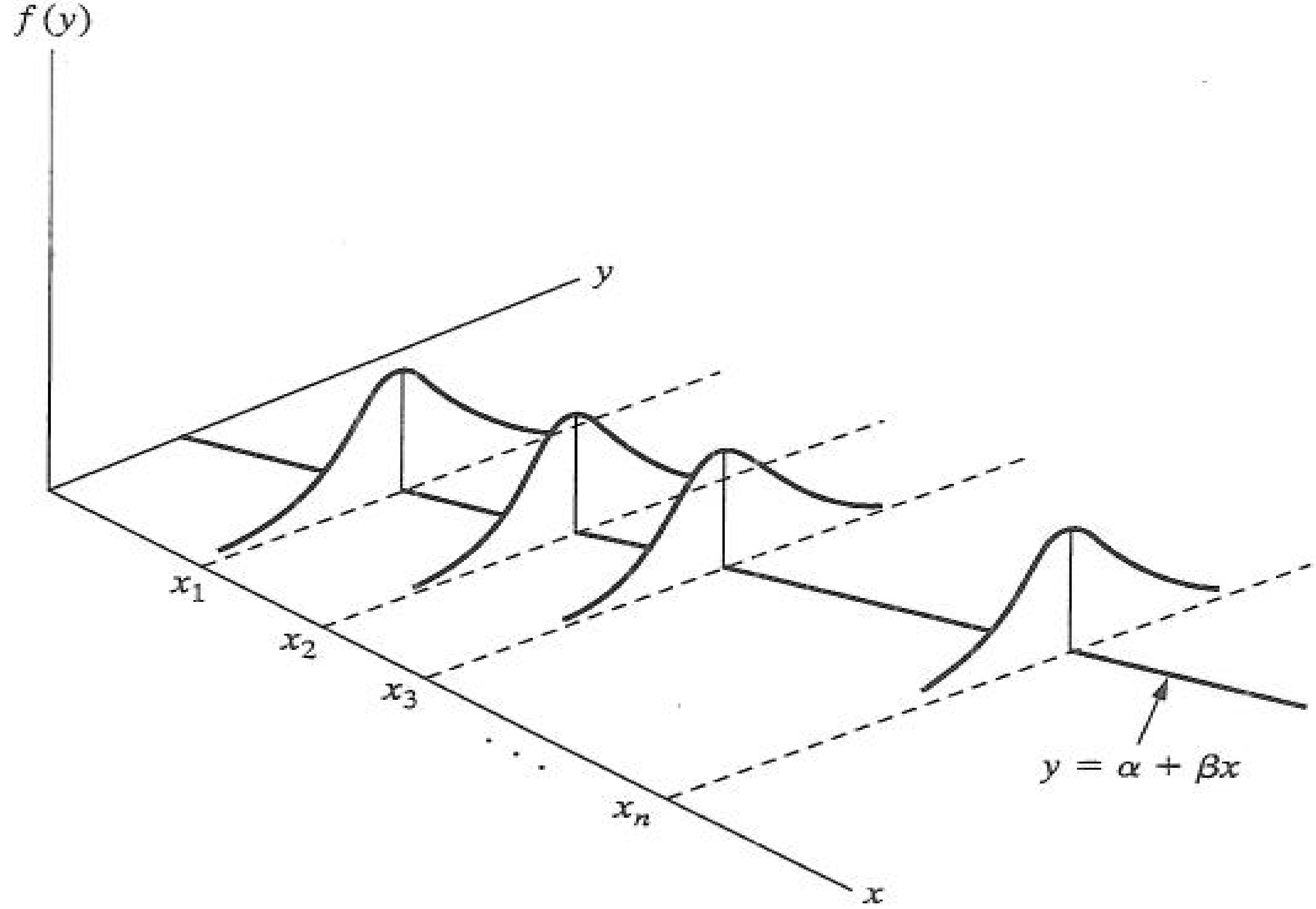


Fig.1

Assumptions - SLRM

Comments to the Model Assumptions:

Comment 1. If the error terms e_i have **constant variances** (as assumed above), that is,

$$Var(e_i) = s^2 \text{ for all } i = \overline{1, n},$$

then e_i (and also the model) are called **homoscedastic**.

But if the variance is changing, that is,

$$Var(e_i) \neq Var(e_j), i \neq j,$$

then e_i (and also the model) are called **heteroscedastic**.

Serially Correlated (SC)

Comment 2. The assumption that the errors corresponding to different observations are **independent** and therefore **uncorrelated** is **important** in both **time-series** and **cross-section studies**.

- When the error terms from different observations are **correlated**, we say that the error process e_i is **serially correlated (SC)**.
- We distinguish **negative** and **positive** serial correlations in a time-series study.

Assumptions - SLRM

Comment 3. Since the **independent variable** $\boldsymbol{x}$ is **non-random**, x_i and e_i are **uncorrelated**:

$$E[x_i e_i] = x_i E[e_i] = 0.$$

In the cases where the **independent variable** $\boldsymbol{X}$ is **also random**, we will assume that

$$E[x_i e_i] = 0.$$

Assumptions - SLRM

Comment 4. The variance $s^2 = Var(e_i)$ is **unknown model parameter**, and must be **estimated** as part of the regression model.

Thus, the **simple regression model** has **three unknown parameters** (a, b, s^2),
while the **curve-fitting model** has only **two** (a and b).

2. Estimation of unknown parameters

The Model:

$$Y_i = \alpha + \beta x_i + \varepsilon_i, \; E[\varepsilon_i] = 0, \; E[\varepsilon_i \varepsilon_j] = \sigma^2 \delta_{ij}. \qquad (1)$$

The unknown parameters: ***α*, *β*** and ***σ*²*.***

The Data:

n independent observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
from the distribution with ***pdf*** $f_{Y|x_i}(y)$ given by

$$f_{Y|x_i}(y) = \frac{1}{\sqrt{2\pi} \cdot \sigma} e^{-\frac{(y-\alpha-\beta x_i)^2}{2\sigma^2}}, \quad i = \overline{1, n}.$$

Estimation of unknown parameters

The Problem:

Find **point estimators and point estimates** for unknown parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$.

First recall the **least square point estimates** for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ that we have obtained in the statistical solution of the prediction problem.

Estimation of unknown parameters

They are

$$\hat{b} = \frac{s_{xy}}{s_{xx}} = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}, \qquad (2)$$

$$\hat{a} = \overline{y} - \frac{s_{xy}}{s_{xx}}\overline{x} = \overline{y} - \hat{b}\,\overline{x}. \qquad (3)$$

The corresponding **point estimators (RV's)** which we again denote by $\hat{a}$ and $\hat{b}$ are given by

$$\hat{b} = \frac{s_{xY}}{s_{xx}} \qquad (Y \ \ is \ \ a \ \ RV) \qquad (2')$$

$$\hat{a} = \overline{Y} - \hat{b}\,\overline{x} \qquad (3')$$

Estimation of unknown parameters

Point estimation of $\boldsymbol{\sigma}^2$

To estimate $\boldsymbol{\sigma}^2$ we use the regression **residuals**.

Recall that the *i*-th residual we defined by

$$e_i = \hat{e}_i = y_i - \hat{y}_i = y_i - \hat{a} - \hat{b}\, x_i . \qquad (4)$$

Denote by ***SSE*** the **residual sum of squares**
(or the **sum of squares of errors**) defined by

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{a} - \hat{b}\, x_i)^2 . \qquad (5)$$

Estimation of unknown parameters

The **least squares point estimate** $\hat{s}^2 = s^2$ of unknown model variance s^2 is defined to be

$$\hat{s}^2 = s^2 = \frac{SSE}{n-2}$$

$$= \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i)^2. \qquad (6)$$

The corresponding point estimator is

$$S^2 = \hat{s}^2$$

$$= \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{a} - \hat{b}x_i)^2. \qquad (6')$$

MLE of unknown parameters

Now we show that the **maximum likelihood estimators (MLE)** for a and b coincide with the **least squares estimators** $\hat{a}$ and $\hat{b}$ given by (2`) and (3`), respectively,

while the **MLE** of s^2 has divisor $\boldsymbol{n}$ rather then $\mathbf{(\boldsymbol{n} - 2)}$, and so it is a **biased estimator** for s^2.

MLE of unknown parameters

Theorem 1. (MLE's).

Under the Model Assumptions 1-4, the **MLEstimators** of a, b and s^2 are given by formulas

$$\hat{b} = \frac{S_{xY}}{S_{xx}} = (2') \qquad (8)$$

$$\hat{a} = \overline{Y} - \hat{b}\,\overline{x} = (3') \qquad (9)$$

$$\hat{s}_1^2 = S_1^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \qquad (10)$$

where $\hat{Y}_i = \hat{a} + \hat{b}\,x_i, i = \overline{1, n}.$

Theorem 1.

Proof.

Since the random variables $Y_i, i = \overline{1, n}$ are **independent** $N(a + b x_i, s^2)$ - **normally** distributed, for **likelihood function** $\boldsymbol{L}$, we have

$$L = L(a, b, s^2) = \prod_{i=1}^{n} f_{Y|x_i}(y_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2p} \cdot s} e^{-\frac{(y - a - b x_i)^2}{2 s^2}}.$$

The **log-likelihood** function is

$$l = \ln L = -\frac{n}{2}\ln(2p) - \frac{n}{2}\ln s^2 - \frac{1}{2 s^2}\sum_{i=1}^{n}(y_i - a - b x_i)^2.$$

Theorem 1.

The maximum of $\boldsymbol{L}$ (or equivalently, $\boldsymbol{l}$) occurs when the partial derivatives w.r.t. a , b and s^2 all vanish.

Setting these partial derivatives equal to 0 gives

$$\frac{\partial l}{\partial a} = -\frac{1}{s^2}\sum_{i=1}^{n}(y_i - a - b\,x_i)(-1) = 0 \qquad (11)$$

$$\frac{\partial l}{\partial b} = -\frac{1}{s^2}\sum_{i=1}^{n}(y_i - a - b\,x_i)(-x_i) = 0 \qquad (12)$$

$$\frac{\partial l}{\partial s^2} = -\frac{n}{2s^2} - \frac{1}{(s^2)^2}\sum_{i=1}^{n}(y_i - a - b\,x_i)^2 = 0 \qquad (13)$$

Theorem 1.

We see that (after simplification) the equations (11) and (12) coincide with the **normal equations** in the **Method of Least Squares.**

Hence solving them for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ we obtain (8) and (9).

Now plugging (8) and (9) into (13) and solving for s^2 we obtain (10).

Properties of the Estimators $\hat{a}$ and $\hat{b}$

Unbiasedness: (Expectations of Estimators $\hat{a}$ and $\hat{b}$).

Theorem 1.

Let $(x_1, Y_n), \mathbf{K}, (x_n, Y_n)$ be $\boldsymbol{n}$ observations satisfying the **Model Assumptions 1-4**, and let $\hat{a}$ and $\hat{b}$ be the **LS Estimators** (or, equivalently, the **MLE estimators**) of the regression parameters $\boldsymbol{\alpha}$ (intercept) and $\boldsymbol{\beta}$ (slope), respectively, defined by (8) and (9):

$$\hat{b} = \frac{S_{xY}}{S_{xx}}; \quad \hat{a} = \overline{Y} - \hat{b}\,\overline{x}. \qquad (1)$$

Then $\hat{a}$ and $\hat{b}$ are unbiased estimators for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$:

$$E[\hat{b}] = b \quad \text{and} \quad E[\hat{a}] = a\,.$$

Unbiasedness of $\hat{a}$ and $\hat{b}$

<u>**Proof.**</u>

We use the following equalities

$$E[aX + b] = aE[X] + b; \quad \sum_{i=1}^{n} c = n \cdot c.$$

We have (since S_{xx} is a constant)

$$E[\hat{b}] = E\left[\frac{S_{xY}}{S_{xx}}\right] = \frac{1}{S_{xx}} E[S_{xY}]. \quad (2)$$

Now we show that

$$E[S_{xY}] = b\, S_{xx}.$$

Unbiasedness of $\hat{a}$ and $\hat{b}$

Indeed, we have

$$E[nS_{xY}] = E[n \cdot \sum_{i=1}^{n} x_i Y_i - (\sum_{i=}^{n} x_i)(\sum_{i=}^{n} Y_i)]$$

$$= n \cdot \sum_{i=1}^{n} x_i E[Y_i] - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} E[Y_i]) \quad (\text{since } E[Y_i] = a + b\, x_i)$$

$$= a\, n \sum_{i=1}^{n} x_i + b\, n \sum_{i=1}^{n} x_i^{\,2} - a\, n \sum_{i=1}^{n} x_i - b\, (\sum_{i=1}^{n} x_i)^2$$

$$= b\, [n \sum_{1}^{n} x_i^{\,2} - (\sum_{i=1}^{n} x_i)^2] = b\, (nS_{xx}).$$

Thus

$$E[S_{xY}] = b\, S_{xx}. \qquad (3)$$

Unbiasedness of $\hat{a}$ and $\hat{b}$

From (2) and (3) we get

$$E[\hat{b}] = \frac{1}{S_{xx}} E[S_{xY}] = \frac{1}{S_{xx}} \cdot b\, S_{xx} = b\,.$$

Now we show that $E[\hat{a}] = a$.

Indeed, we have

$$E[\hat{a}] = E[\overline{Y} - \hat{b}\,\overline{x}] = E[\overline{Y}] - \overline{x}E[\hat{b}]$$

$$= \frac{1}{n}\sum_{i=1}^{n} E[Y_i] - \overline{x}\cdot b = \frac{1}{n}\sum_{i=1}^{n}(a + b\,x_i) - \overline{x}\,b$$

$$= \frac{1}{n}\cdot a\,n + b\cdot\frac{1}{n}\sum_{i=1}^{n} x_i - \overline{x}\,b$$

$$= a + b\,\overline{x} - \overline{x}\,b = a\,.$$

The Variances of Estimators $\hat{a}$ and $\hat{b}$

Theorem 2.

For the variances $Var(\hat{a})$ and $Var(\hat{b})$ we have

$$a)\quad Var(\hat{b}) = \frac{s^2}{S_{xx}} = \frac{s^2}{\sum_{i=1}^{n}(x-\overline{x})^2}. \qquad (4)$$

$$b)\quad Var(\hat{a}) = \frac{\frac{s^2}{n}\sum_{i=1}^{n} x_i^2}{S_{xx}} = s^2\left[\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right]. \qquad (5)$$

The Variance of Estimators $\hat{a}$ and $\hat{b}$

Proof.

We use the following facts

1) $Var(aX + b) = a^2 Var(X)$

2) $Var(\sum_{k=1}^{n} X_k) = \sum_{k=1}^{n} Var(X_k)$ for **independent** X_k

3) $\sum_{i=1}^{n} (x_i - \overline{x}) = n\overline{x} - n\overline{x} = 0.$

The Variance of Estimators $\hat{a}$ and $\hat{b}$

Proof of a). We have

$$Var(\hat{b}) = Var\left[\frac{S_{xY}}{S_{xx}}\right] = \frac{1}{S_{xx}^2} Var(S_{xY}) \qquad (6)$$

To compute $Var(S_{xY})$, first observe that by Fact 3)

$$S_{xY} = \sum_{i=1}^{n} (x_i - \overline{x})(Y_i - \overline{Y}) = \sum_{i=1}^{n} (x_i - \overline{x}) Y_i. \qquad (7)$$

Therefore, by Facts 1) and 2) (**independence** of Y_i !)

$$Var(S_{xY}) = Var\left[\sum_{i=1}^{n} (x_i - \overline{x}) Y_i\right] = \sum_{i=1}^{n} Var[(x_i - \overline{x}) Y_i]$$

$$= \sum_{i=1}^{n} (x_i - \overline{x})^2 Var(Y_i) = \sum_{i=1}^{n} (x_i - \overline{x})^2 \cdot s^2 = S_{xx} \cdot s^2. \quad (8)$$

The Variance of Estimators $\hat{a}$ and $\hat{b}$

From (6) and (8) we get (4):

$$Var(\hat{b}) = \frac{1}{S^2_{xx}} \cdot (S_{xx} \cdot s^2) = \frac{s^2}{S_{xx}}.$$

To prove b) we use the following fact.

Lemma 1.

The RV's $\hat{b}$ and $\overline{Y}$ are independent.

The Variance of Estimators $\hat{a}$ and $\hat{b}$

Using **Lemma 1** and **Facts** 1) and 2), we can write

$$Var(\hat{a}) = Var(\overline{Y} - \hat{b}\,\overline{x})$$

$$= Var(\overline{Y}) + \overline{x}^2 Var(\hat{b})$$

$$= \frac{s^2}{n} + \overline{x}^2 \cdot \frac{s^2}{S_{xx}} = s^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}} \right],$$

yielding the second equality in (5). To prove the first equality:

Remark – Problem. Show that

$$\frac{1}{nS_{xx}} \sum_{i=1}^{n} x_i^2 = \frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}$$

The Distribution of Estimators $\hat{a}$ and $\hat{b}$

Theorem 3.

Under the **Model Assumptions** both estimators $\hat{a}$ and $\hat{b}$ have **normal** distributions:

$$a)\quad \hat{b} \sim N\left(b\,,\ \frac{s^2}{S_{xx}} \right)$$

$$b)\quad \hat{a} \sim N\left(a\,,\ \frac{s^2 \sum_{i=1}^{n} x_i^2}{n S_{xx}} \right).$$

The Distribution of Estimators $\hat{\alpha}$ and $\hat{\beta}$

Proof.

We use the following result.

Lemma 2.

Let $X_1, \ldots, X_n$ be **independent** and **normally** distributed RV's: $X_k \sim N(\mu_k, \sigma_k^2)$.

Then for any constants $a_k, k = \overline{1, n}$,

$$X = \sum_{i=1}^{n} a_k X_k \sim N\left(\sum_{i=1}^{n} a_k \mu_k , \sum_{i=1}^{n} a_k^2 \sigma_k^2 \right).$$

The Distribution of Estimators $\hat{a}$ and $\hat{b}$

Proof of (a).

Using equation (7) we can write

$$\hat{b} = \frac{S_{xY}}{S_{xx}} = \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \overline{x}) Y_i = \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{S_{xx}}\right) Y_i = \sum_{i=1}^{n} a_i Y_i,$$

where $a_i = \dfrac{x_i - \overline{x}}{S_{xx}}, i = \overline{1, n}.$

Taking into account that $Y_i \sim N(a + b x_i, s^2)$,
and applying **Lemma 2** we obtain (a).

The Distribution of Estimators $\hat{a}$ and $\hat{b}$

Indeed, taking into account the equality

$$\sum (x_i - \overline{x}) = n\overline{x} - n\overline{x} = 0,$$

we can write

$$m = \sum_{i=1}^{n} a_i m_i = \sum_{i=1}^{n} \frac{x_i - \overline{x}}{S_{xx}} (a + b x_i)$$

$$= \frac{a}{S_{xx}} \sum_{i=1}^{n} (x_i - \overline{x}) + b \sum_{i=1}^{n} \frac{(x_i - \overline{x}) x_i}{S_{xx}}$$

$$= \frac{b}{S_{xx}} \sum_{i=1}^{n} (x_i - \overline{x})(x_i - \overline{x})$$

$$= \frac{b}{S_{xx}} \cdot S_{xx} = b .$$

The Distribution of Estimators $\hat{a}$ and $\hat{b}$

Next, for variance we have

$$\sum_{i=1}^{n} a_i \sigma_i^2 = \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{S_{xx}^2} \sigma^2 = \frac{\sigma^2}{S_{xx}^2} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

$$= \frac{\sigma^2}{S_{xx}^2} S_{xx} = \frac{\sigma^2}{S_{xx}}.$$

Proof of (b).

By part (a) we have

$$\hat{b} \sim N(b, \frac{\sigma^2}{S_{xx}}).$$

Next, observe that by **Lemma 2,**

$$\overline{Y} \sim N(a + b\overline{x}, \frac{\sigma^2}{n}).$$

The Distribution of Estimators $\hat{a}$ and $\hat{b}$

Since by **Lemma 1** the RV's $\hat{b}$ and $\overline{Y}$ are **independent**, by **Lemma 2** for $\boldsymbol{n} = \mathbf{2}$, $(\hat{a} = \overline{Y} - \hat{b}\,\overline{x})$

$$X_1 = \overline{Y}, X_2 = \hat{b}, a_1 = 1, a_2 = -\overline{x},$$

we obtain

$$\sum_{i=1}^{2} a_i m_i = a + b\,\overline{x} - b\,\overline{x} = a,$$

$$\sum_{i=1}^{2} a_i^2 s_i^2 = (1)\left(\frac{s^2}{n}\right) + (-\overline{x})^2 \cdot \frac{s^2}{S_{xx}}$$

$$= \frac{s^2}{n} + \overline{x}^2 \cdot \frac{s^2}{S_{xx}} = s^2\left[\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right].$$

Theorem 3 is proved.

Properties of the Estimator S^2

Recall that the least squares point estimator (S_{LS}^2) and the maximum likelihood estimator S_{ML}^2 for model variance s^2 are given by

$$S_{LS}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{Y})^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{a} - \hat{b}x_i)^2 \qquad (1)$$

and

$$S_{ML}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y})^2. \qquad (2)$$

Observe that

$$S_{LS}^2 = \frac{n}{n-2}S_{ML}^2. \qquad (3)$$

Properties of the Estimator s^2

We will use the following result.

Lemma 3.
Under the **Model Assumptions 1-4**,

a) the RV's $\hat{b}$, $\bar{Y}$ and S^2_{ML} are **mutually independent**.

b) the RV $\dfrac{nS^2_{ML}}{s^2}$ has **chi square distribution** with $(\boldsymbol{n}$ **- 2**$)$ degrees of freedom, that is,

$$\frac{nS^2_{ML}}{s^2} \sim c^2(n-2). \qquad (4)$$

Properties of the Estimator s^2

- **<u>Theorem 4</u>.**

The least squares estimator S_{LS}^2 is an **<u>unbiased</u>** estimator for s^2, while
the maximum likelihood estimator S_{ML}^2 is **<u>biased</u>**.

Properties of the Estimator s^2

Proof.

Since if $X \sim c^2(k)$, then $E[X] = k$,

using **Lemma 3(b)** and formula (3) we obtain

$$E[S^2_{LS}] = E[\frac{n}{n-2} S^2_{ML}] = \frac{s^2}{n-2} E[\frac{nS^2_{ML}}{s^2}]$$

$$= \frac{s^2}{n-2} \cdot (n-2) = s^2.$$

Remark 1 - Problem.

It can be shown that $Cov(\hat{a}, \hat{b}) = -\frac{\overline{x} s^2}{S_{xx}}.$

Properties of the Estimator s^2

Remark 2.
Using the estimate s_{LS}^2 for s^2, we obtain the following **estimated** variances and covariance for estimators $\hat{a}$ and $\hat{b}$:

$$\hat{V}ar(\hat{b}) = s_{\hat{b}}^2 = \frac{s^2}{S_{xx}}.$$

$$\hat{V}ar(\hat{a}) = s_{\hat{a}}^2 = s^2\left(\sum_{i=1}^{n} x_i^2 / (nS_{xx})\right).$$

$$\hat{C}ov(\hat{a}, \hat{b}) = -\frac{\overline{x}s^2}{S_{xx}},$$

where $s^2 = s_{LS}^2$.

Best Linear Unbiased Estimators (BLUE)

1. BLUE for unknown μ.

Let $a_1, \ldots, a_n$ be any set of real numbers such that

$$\sum_{k=1}^{n} a_k = 1, \qquad (1)$$

and let $X_1, \ldots, X_n$ be independent RV's with common mean μ and variances $\sigma_k^2 = Var(X_k), k = \overline{1, n}$.

Then the statistic (a linear combination of X_k's)

$$T = \sum_{k=1}^{n} a_k X_k \qquad (2)$$

is an **unbiased** $(E[T] = \mu)$ estimator for μ with variance

$$Var(T) = \sum_{i=1}^{n} a_i^2 \sigma_i^2. \qquad (3)$$

BLUE for unknown m

Definition 1.
The estimator $\boldsymbol{T}$ given by (2) with $\{a_k\}$ satisfying (1) is called **linear unbiased estimator** of m.

Definition 2.
A linear unbiased estimator T_0 of m that has **minimum variance** (among all linear unbiased estimators) is called **best linear unbiased estimator (BLUE)** of m.

BLUE for unknown m

Theorem 1.

If $X_i, i = \overline{1, n}$, are ***iid*** with common **(unknown)** m and common **(known)** s^2, then the **BLUE** T_0 of m is the **sample mean**, that is,

$$T_0 = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

BLUE for unknown m

Proof.

Let $T = T(\underline{X}) = \sum_{k=1}^{n} b_k X_k, \qquad \sum_{k=1}^{n} b_k = 1.$

We have

$$Var(T) = Var\left(\sum_{k=1}^{n} b_k X_k\right) = s^2 \sum_{k=1}^{n} b_k^2$$

which is **least** ***iff*** we choose the coefficients $b_k, k = 1, \dots, n,$ so that $\sum_{k=1}^{n} b_k^2$ is **smallest**, subject to the condition

$$\sum_{k=1}^{n} b_k = 1.$$

BLUE for unknown m

We have
$$\sum_{k=1}^{n} b_k^2 = \sum_{k=1}^{n}\left(b_k - \frac{1}{n} + \frac{1}{n}\right)^2$$
$$= \sum_{k=1}^{n}\left(b_k - \frac{1}{n}\right)^2 + \frac{2}{n}\sum_{k=1}^{n}\left(b_k - \frac{1}{n}\right) + \frac{1}{n}$$
$$= \sum_{k=1}^{n}\left(b_k - \frac{1}{n}\right)^2 + \frac{1}{n},$$

which is **minimized** for the choice $b_k = 1/n,\ k = 1,\dots,n,$
Thus, the **BLUE** $T_0 = T_0(\underline{X})$ of q is

$$T_0 = T_0(\underline{X}) = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Gauss-Markov Theorem

Theorem 2 (Gauss-Markov).

Let $(x_i, Y_n), \ldots, (x_n, Y_n)$ be $\boldsymbol{n}$ **independent** observations satisfying **Model Assumptions**:

$$E[Y_i] = E\left[Y_i \middle| x_i\right] = a + b\, x_i, \quad i = \overline{1, n},$$

$$Var(Y_i) = s^2, \quad i = \overline{1, n}.$$

Then, of all estimators of a and b that are **linear functions** of $Y_1, \ldots, Y_n$ and that are **unbiased**, the **least squares estimators**

$$\hat{b} = \frac{S_{xy}}{S_{xx}} \quad \text{and} \quad \hat{a} = \overline{Y} - \hat{b}\,\overline{x}$$

have the **smallest respective variances**, that is, are **BLUE's** for a and b, respectively.

Gauss-Markov Theorem

Proof.

Consider a general linear estimator of a, say

$$T = \sum_{i=1}^{n} a_i Y_i. \qquad (1)$$

Now $\boldsymbol{T}$ will be an unbiased estimator of a if

$$E[T] = a. \qquad (2)$$

Since, by (1)

$$E[T] = E\left[\sum_{i=1}^{n} a_i Y_i\right] = \sum_{i=1}^{n} a_i E[Y_i] = \sum_{i=1}^{n} a_i (a + b x_i)$$

$$= a \sum_{i=1}^{n} a_i + b \sum_{i=1}^{n} a_i x_i, \qquad (3)$$

Gauss-Markov Theorem

then (2) will be satisfied if and only if

$$\sum_{i=1}^{n} a_i = 1 \quad \text{and} \quad \sum_{i=1}^{n} a_i x_i = 0. \qquad (4)$$

The variance of $\boldsymbol{T}$, which is to be minimized subject to (4), is

$$Var(T) = Var(\sum_{i=1}^{n} a_i Y_i) = s^{2}(\sum_{i=1}^{n} a_i^2). \qquad (5)$$

Hence, it is enough to minimize $\sum_{i=1}^{n} a_i^2$ subject to

$$\sum_{i=1}^{n} a = 1 \quad \text{and} \quad \sum_{i=1}^{n} a_i x_i = 0.$$

Gauss-Markov Theorem

For $i = \overline{1, n}$, we denote

$$b_i = \frac{\sum_{k=1}^{n} x_k^2 - x_i(\sum_{k=1}^{n} x_k)}{n\sum_{k=1}^{n} x_k^2 - (\sum_{k=1}^{n} x_k)^2}. \tag{6}$$

Then we can express a_i as

$$a_i = b_i + \Delta_i, i = \overline{1, n}, \tag{7}$$

for some Δ_i.

From (4), it follows that we must have.

$$\sum_{i=1}^{n} b_i + \sum_{i=1}^{n} \Delta_i = 1 \quad \text{and} \quad \sum_{i=1}^{n} b_i x_i + \sum_{i=1}^{n} \Delta_i x_i = 0. \tag{8}$$

Gauss-Markov Theorem

It is easy to check that (using (6))

$$\sum_{i=1}^{n} b_i = 1 \quad \text{and} \quad \sum_{i=1}^{n} b_i x_i = 0. \qquad (9)$$

Hence, from (8) and (9) we obtain

$$\sum_{i=1}^{n} \Delta_i = 0 \quad \text{and} \quad \sum_{i=1}^{n} \Delta_i x_i = 0.$$

Now using (7) we can write

$$\sum_{i=1}^{n} a_i^2 = \sum_{i=1}^{n} (b_i + \Delta_i)^2 = \sum_{i=1}^{n} b_i^2 + \sum_{i=1}^{n} \Delta_i^2 + 2\sum_{i=1}^{n} b_i \Delta_i$$

$$= \sum_{i=1}^{n} b_i^2 + \sum_{i=1}^{n} \Delta_i^2 + 0 \geq \sum_{i=1}^{n} b_i^2,$$

and the result follows.

Inferences about regression parameters: HT and CI's

Inferences about the slope b.

Recall that

(a) $\hat{b} = \frac{S_{XY}}{S_{XX}}$ is an unbiased point estimator for b.

(Theorem 1)

(b) $Var(\hat{b}) = \frac{s^2}{S_{XX}}$ **(Theorem 2)**

and $\hat{Var}(\hat{b}) = s_{\hat{b}}^2 = \frac{s^2}{S_{xx}}$. **(Remark 2)**

(c) $\hat{b} \sim N(b, \frac{s^2}{S_{xx}})$ **(Theorem 3)**

Statistical inferences about b are based on the following theorem.

Inferences about the Slope

Theorem 1.

Let $(x_1, Y_n), \ldots, (x_n, Y_n)$ be $\boldsymbol{n}$ observations satisfying the **Model Assumptions 1-4.** Then the statistic

$$T = \frac{\hat{b} - b}{S / \sqrt{S_{xx}}} = \frac{\hat{b} - b}{S_{\hat{b}}} \sim t_{(n-2)} \qquad (1)$$

has **Student *t*-distribution** with (***n*** **-2**) degrees of freedom,

where
$$S = S_{LS} = \sqrt{\frac{SSE}{n-2}}.$$

Inferences about the Slope

Proof.

The statistic ***T*** we can write

$$T = \frac{\hat{b} - b}{S \,/ \sqrt{S_{xx}}} = \frac{\hat{b} - b}{s \,/ \sqrt{S_{xx}}} \left[\sqrt{\frac{(n-2)S^2}{s^2} \cdot \frac{1}{(n-2)}} \right]^{-1}.$$

Now, by (c)

$$Z = \frac{\hat{b} - b}{s \,/ \sqrt{S_{xx}}} \sim N(0,1).$$

Next,

$$\frac{(n-2)S^2}{s^2} \sim c^2(n-2),$$

and $\hat{b}$ and S^2 are **independent.**

Inferences about the Slope

Therefore the result follows from the definition of Student distribution, because

$$T = \frac{Z}{\sqrt{c^2(n-2)/(n-2)}} \sim t_{(n-2)}.$$

Thus, we can use the standard ***T*** -**procedure** to make inferences about the slope parameters b .

A Confidence Interval for the Slope

For given number a $(0 \le a \le 1)$, a $100(1-a)\%$ **CI** for the **slope** b of the **true regression line** is the interval

$$\hat{b} \pm t_{a/2,(n-2)} \cdot s_{\hat{b}}, \qquad (2)$$

where $t_{a/2,(n-2)}$ is the upper $\frac{a}{2}$ - percentile of ***T***-distribution with **(*n* - 2)** ***df***, that is,

$$P(T > t_{a/2,(n-2)}) = \frac{a}{2}, \quad s_{\hat{b}} = \frac{s}{\sqrt{S_{xx}}},$$

and

$$s^2 = \frac{SSE}{n-2} = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i)^2.$$

A Confidence Interval for the Slope

Indeed,

By Theorem 1

$$T = \frac{\hat{b} - b}{s_{\hat{b}}} \sim t_{(n-2)}.$$

Hence, for given a $(0 \le a \le 1)$,

$$P(-t_{a/2,(n-2)} < \frac{\hat{b} - b}{s_{\hat{b}}} < t_{a/2,(n-2)}) = 1 - a$$

solving the inside inequality for b we obtain (2).

Hypothesis Testing for the Slope

The null (basic) hypothesis H_0 and the alternative H_a are specified to be

$$H_0 : b = b_0 \quad \text{vs.} \quad H_a : b \vee b_0 \qquad (\vee = \{\neq, >, <\}).$$

The **Test Statistic** is

$$T = \frac{\hat{b} - b}{s_{\hat{b}}} \sim t_{(n-2)} \quad \text{under} \quad H_0 : b = b_0.$$

For a specific sample $(x_1, y_1), \ldots, (x_n, y_n)$, the **observed value** of **Test Statistic** is

$$t_0 = T(obs) = \frac{\hat{b} - b_0}{s_{\hat{b}}}.$$

Hypothesis Testing for the Slope

Then use standard ***t*** -**critical value**, ***P*** -**value** and **CI's** Methods to test the hypotheses.

Remark.

The **model utility test** is the test

$$H_0 : b = 0 \quad \text{vs.} \quad H_a : b \vee 0,$$

in which case the **test statistic** and **observed value** are given by

$$T = \frac{\hat{b}}{S_{\hat{b}}}; \qquad t_0 = T(obs) = \frac{\hat{b}}{s_{\hat{b}}}.$$

Inference about the Intercept a

Recall:

(a) $\hat{a} = \overline{Y} - \hat{b}\,\overline{x}$ is an **unbiased point estimator** for a **(Th.1),**

(b) $Var(\hat{a}) = s^2\left[\dfrac{1}{n} + \dfrac{\overline{x}^2}{S_{xx}}\right]$ **(Th.2),**

$\hat{V}ar(\hat{a}) = S^2_{\hat{a}} = S^2\left(\sum_{i=1}^{n} x_i^2 / (nS_{xx})\right)$ **(Remark 2)**, and

(c) $\hat{a} \sim N(a, s^2 \sum_{i=1}^{n} x_i^2 / (nS_{xx}))$.

So, statistical inferences about a can be based on the following:

Inference about the Intercept

Theorem 2.

Under the Model Assumptions, the statistic

$$T = \frac{\hat{a} - a}{s_{\hat{a}}} \sim t_{(n-2)} \qquad (3)$$

has **Student *t* -distribution** with (***n* -2**) ***df***.

Proof is similar to that of **Theorem 1**.

A Confidence Interval for the Intercept

For given number a $(0 \le a \le 1)$, a $100(1-a)\%$ **CI** for the **intercept** a of the **true regression line** is the interval

$$\hat{a} \pm t_{a/2,(n-2)} \cdot S_{\hat{a}},$$

where

$$S_{\hat{a}} = S\sqrt{\sum_{i=1}^{n} x_i^2 / (nS_{xx})}$$

Hypothesis Testing for the Intercept

The **hypothesis** are

$$H_0 : a = a_0 \quad \text{vs.} \quad H_a : a \vee a_0.$$

The **Test Statistic** is

$$T = \frac{\hat{a} - a}{S_{\hat{a}}} \sim t_{(n-2)} \qquad \text{(under } H_0 : a = a_0\text{)}$$

The **observed value** of **TS** is

$$T_0 = T(obs) = \frac{\hat{a} - a_0}{s_{\hat{a}}}.$$

Then, use **standard *T* -procedure**.

Inferences about the Variance

By **Lemma 3**, the statistic

$$c^2 = \frac{(n-2)S^2}{s^2} \sim c^2(n-2),$$

where $S^2 = S^2_{LS}$,

has c^2 - **distribution** with (***n* - 2**) ***df***.

So, statistical inferences about variance is based on

c^2**- procedure**.

A Confidence Interval for the Variance

It follows that

$$P\left(c^{2}_{a/2,(n-2)} \le \frac{(n-2)s^{2}}{s^{2}} \le c^{2}_{1-a/2,(n-2)} \right) = 1-a\,.$$

Solving the inside inequality for s^{2} we obtain

$$P\left(\frac{(n-2)s^{2}}{c^{2}_{1-a/2,(n-2)}} \le s^{2} \le \frac{(n-2)s^{2}}{c^{2}_{a/2,(n-2)}} \right) = 1-a\,.$$

Thus, for given a $(0 \le a \le 1)$, a $100(1-a)\%$ **CI** for s^{2} is the interval

$$\left(\frac{(n-2)s^{2}}{c^{2}_{1-a/2,(n-2)}}, \quad \frac{(n-2)s^{2}}{c^{2}_{a/2,(n-2)}} \right).$$

Hypothesis Testing about the Variance

The **hypotheses** are

$$H_0 : s^2 = s_0^2 \quad \text{vs.} \quad H_a : s^2 \vee s_0^2 .$$

The **TS** is

$$c^2 = \frac{(n-2)s^2}{s^2} \sim c^2(n-2) \qquad (\text{under } H_0 : s^2 = s_0^2)$$

The **observed value of TS** is

$$c^2(obs) = \frac{(n-2)s^2}{s_0^2} .$$

Then use standard c^2 **- procedure** with (***n*-2**) ***df*** , instead of (***n*-1**) ***df***.