STA302/1001 - Methods of Data Analysis I

(Week 01 - lecture B)

Wei (Becky) Lin

May 15-19, 2017

From last lecture (Week 1- Lect A)

- Notes about syllabus.
- A functional relationship vs a statistical relationship.
- Simple linear regression: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$. → $\beta_0, \beta_1, \sigma^2$: 3 parameters
- Least Square Estimation of β_0, β_1.
  - Gauss-Markov conditions
  - LS estimators b_0, b_1

$$b_0 = \bar{y} - b_1 \bar{x}, \quad b_1 = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_1^n (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

  - The unbiased estimator for σ^2.

$$s^2 = \hat{\sigma}^2 = MSE = \frac{SSE}{n-2}, \quad SSE = \sum_1^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n e_i^2$$

  - Variance of LS estimators: $V(b_0)$ and $V(b_1)$

$$V(b_0) = V(\sum w_i Y_i) = \sigma^2 (\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}})$$

$$V(b_1) = V(\sum k_i Y_i) = \frac{\sigma^2}{S_{xx}}$$

  - Estimator of $V(b_0)$ and $V(b_1)$: replace unknown σ^2 by MSE.

Last lecture (Week 1- Lect A)

$$Q = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$$

- Show LS estimators b_0, b_1 which minimize Q are BLUE
  - LS estimators b_0, b_1 are linear estimators

$$b_0 = \sum_{1}^{n} w_i Y_i = \sum_{1}^{n} (\frac{1}{n} - k_i \bar{X}) Y_i = \sum_{1}^{n} \{\frac{1}{n} - \frac{(X_i - \bar{X})\bar{X}}{\sum_1^n (X_i - \bar{X})^2}\} Y_i$$

$$b_1 = \sum_{1}^{n} k_i Y_i = \sum_{1}^{n} \frac{X_i - \bar{X}}{\sum_1^n (X_i - \bar{X})^2} Y_i$$

and

$$\sum k_i = 0, \quad \sum k_i X_i = 1, \quad \sum k_i^2 = 1/S_{xx}$$

$$\sum w_i = 1, \quad \sum w_i X_i = 0$$

  - Unbiasedness: $E(b_0) = \beta_0$, $E(b_1) = \beta_1$
  - Show that b_0, b_1 have the smallest variance among the class of all linear unbiased estimators by linear algebra.

Week 01- Lecture B: Learning objectives & Outcomes

- Properties of Least Squares Fitted Line. same as "estimated reg. line"
- Why squared residuals instead of absolute residuals in OLS estimation?
- Introduction to other types of regressions
- Normal regression model.
- Maximum Likelihood Estimation (MLE) of β_0, β_1.
- Review of statistical inference.

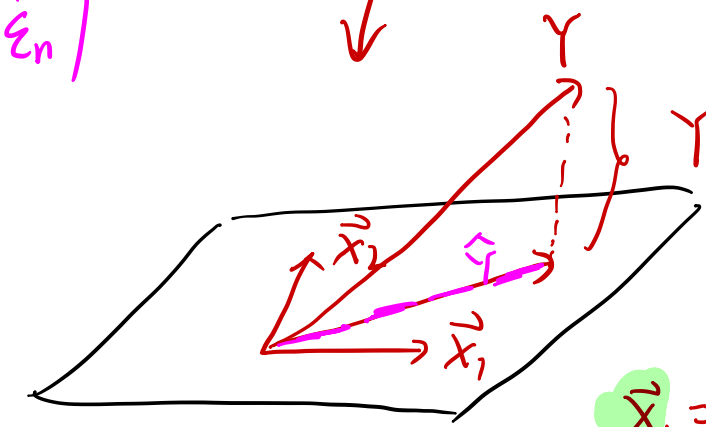
Properties of LS fitted line (Summary)

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$\vec{X}_1$ $\vec{X}_2$

- $\hat{Y}$ is the projection of Y on $\text{span}(\vec{X}_1, \vec{X}_2)$
- $e \perp \text{span}(\vec{X}_1, \vec{X}_2)$



$\vec{X}_1 = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$

$\vec{X}_2 = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$

0. $E(\hat{Y}_i) = E(Y_i)$
1. $\sum_1^n e_i = 0$.
2. $\sum_1^n e_i^2$ is a minimum.
3. $\sum_1^n Y_i = \sum_1^n \hat{Y}_i$
4. $\sum_1^n X_i e_i = 0$
5. $\sum_1^n \hat{Y}_i e_i = 0$
6. The fitted regression line always goes through the point $(\bar{X}, \bar{Y})$.

$e_i = Y_i - \hat{Y}_i$

We will prove them one by one.

- two vectors $\vec{u}, \vec{v}$ are orthogonal iff $\langle u, v \rangle = \sum_i^n u_i v_i = 0$
- e is perpendicular to the subspace spanned by $\vec{X}_1, \vec{X}_2$, so

$\langle e, \vec{X}_1 \rangle = \sum_{i=1}^n e_i \cdot 1 = 0$, $\langle e, \vec{X}_2 \rangle = \sum x_i e_i = 0$, $\langle e, \hat{Y} \rangle = 0$

Properties of LS fitted line: (0) $E(\hat{Y}_i) = E(Y_i)$

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, $E(\varepsilon_i) = 0$, $V(\varepsilon_i) = \sigma^2$, $cov(\varepsilon_i, \varepsilon_j) = 0$

$\Rightarrow E(Y_i) = \beta_0 + \beta_1 X_i + E(\varepsilon_i) = \beta_0 + \beta_1 X_i$

Proof:

Since $\hat{Y}_i = b_0 + b_1 X_i$ so we have

$$E(\hat{Y}_i) = E(b_0) + E(b_1) X_i = \beta_0 + \beta_1 X_i$$

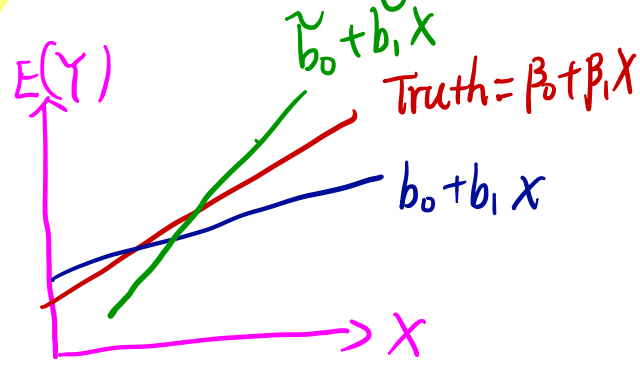
And b_0 and b_1 are unbiased estimators of β_0, β_1 respectively, it follows that

$$E(\hat{Y}_i) = \beta_0 + \beta_1 X_i = E(Y_i)$$

Q: How to understand $E(\hat{Y}_i) = E(Y_i)$

A: Given different samples of $(\vec{X}, \vec{Y})$, we have different estimated lines: $b_0 + b_1 X$

But on avg, it leads to the truth reg. line



Properties of LS fitted line: (1) $\sum_1^n e_i = 0$ → Proof 2: $e_i = Y_i - \hat{Y}_i$

$= Y_i - (\bar{Y} + b_1(X_i - \bar{X}))$

Proof:

$\sum_i^n e_i = \sum_{i=1}^n \{Y_i - \bar{Y} - b_1(X_i - \bar{X})\}$

$= \sum_i^n Y_i - n\bar{Y} - b_1 \sum (X_i - \bar{X})$

$= n\bar{Y} - n\bar{Y} - b_1 \cdot 0$

$= 0$

- Recall b_0, b_1 satisfy the Normal equations:

$$\sum_{i=1}^n Y_i = nb_o + b_1 \sum_{i=1}^n X_i$$

$$\sum_{i=1}^n X_i Y_i = b_0 \sum_{i=1}^n X_i + b_1 \sum_{i=1}^n X_i^2$$

$\hat{Y}_i = b_0 + b_1 X_i$

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

$= \bar{Y} + b_1(X_i - \bar{X})$

- $\sum_1^n e_i = \sum_1^n (Y_i - \hat{Y}_i) = \sum_1^n (Y_i - b_0 - b_1 X_i)$ This gives

$$\sum_1^n e_i = \sum_{i=1}^n Y_i - nb_o - b_1 \sum_{i=1}^n X_i$$

- Thus we have the sum of residuals is zero, $\sum_1^n e_i = 0$, by first Normal equation.

Properties of LS fitted line: (2) $\sum_1^n e_i^2$ is a minimum

$\hat{Y}_i = b_0 + b_1 X_i$

where b_0, b_1 minimize $Q = \sum_{i=1}^n e_i^2$ $\Rightarrow$ Q reaches its minimum at b_0, b_1.

Proof:

This was the requirement to be satisfied in deriving the LS estimators of the regression parameters since the criterion

$$Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i) = \sum_{i=1}^n (Y_i - b_0 - b_1 X_i)^2$$

to be minimized when the LS estimators b_0 and b_1 are used for estimating β_0 and β_1.

LS approach: Q is minimized at b_0, b_1

step1: $\frac{\partial Q}{\partial \theta^T} = 0 \rightarrow \hat{\theta}$, $\theta = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix}$

step2: check $\det(H) > 0$ and $H(1,1) > 0$

Properties of LS fitted line: (3) $\sum_1^n Y_i = \sum_1^n \hat{Y}_i$

← (1) $\sum_{i=1}^n e_i = \sum (Y_i - \hat{Y}_i) = 0$

↑ observed ↑ Fitted value

Proof:

Want to show that the sum of the observed values Y_i equals the sum of the fitted values $\hat{Y}_i$. From (1) we have $\sum_1^n e_i = \sum_1^n (Y_i - \hat{Y}_i) = 0$, so (3) follows.

$$\sum_1^n \hat{Y}_i = \sum_1^n (b_0 + b_1 X_i)$$

$$= \sum_1^n (\bar{Y} - b_1 \bar{X} + b_1 X_i), \quad b_0 = \bar{Y} - b_1 \bar{X}$$

$$= n\bar{Y} - b_1 n\bar{X} + b_1 \sum_1^n X_i$$

$$= n\bar{Y} - b_1 n\bar{X} + b_1 n\bar{X}$$

$$= n\bar{Y}$$

$$= \sum_{i=1}^n Y_i$$

$\hat{Y}_i = b_0 + b_1 X_i = \bar{Y} + b_1 (X_i - \bar{X})$

⇓

$\sum \hat{Y}_i$

$= \sum (\bar{Y} + b_1 (X_i - \bar{X}))$

$= n\bar{Y} + b_1 \underbrace{\sum (X_i - \bar{X})}_{=0}$

$= n\bar{Y}$

$= \sum Y_i$ QED.

Properties of LS fitted line: (4) $\sum_1^n X_i e_i = 0$

Proof:

Want to prove that the sum of the weighted residuals is zero when the i^{th} residual is weighted by the i^{th} predictor variable value. Again, the Normal equations for b_0 and b_1 are

$$\sum_{i=1}^{n} Y_i = nb_o + b_1 \sum_{i=1}^{n} X_i, \quad \sum_{i=1}^{n} X_i Y_i = b_0 \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2$$

And we have

$$\sum_{1}^{n} X_i e_i = \sum_{1}^{n} X_i (Y_i - \hat{Y}_i) = \sum_{1}^{n} X_i (Y_i - b_0 - b_1 X_i)$$

$$= \sum_{1}^{n} X_i Y_i - b_0 \sum_{1}^{n} X_i - b_1 \sum_{1}^{n} X_i^2 = 0$$

Thus, $\sum_1^n X_i e_i = 0$ by the second Normal Equation.

Properties of LS fitted line: (5) $\sum_1^n \hat{Y}_i e_i = 0$

Proof:

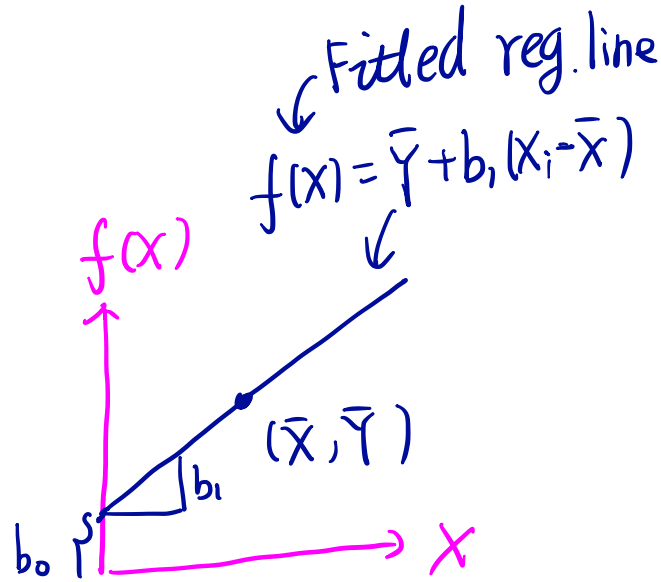
Want to prove that the sum of weighted residuals is zero when the weight of i^{th} residual is $\hat{Y}_i$.

$$\sum_1^n \hat{Y}_i e_i = \sum_1^n (b_0 + b_1 X_i) e_i, \quad \hat{Y}_i = b_0 + b_1 X_i$$

$$= b_0 \sum_1^n e_i + b_1 \sum_1^n X_i e_i$$

$$= 0, \text{ by properties (1) and (4)}$$

Properties of LS fitted line: (6) always goes thru $(\bar{X}, \bar{Y})$



Proof:

Recall the fitted regression line is

$$\hat{Y}_i = b_0 + b_1 X_i.$$

Replacing b_0 by $\bar{Y} - b_1\bar{X}$ gives

$$\hat{Y}_i = \bar{Y} + b_1(X_i - \bar{X}),$$

When $X = \bar{X}$, we have $\hat{Y}_i = \bar{Y}$. That is, the fitted regression line always goes through the point $(\bar{X}, \bar{Y})$.

Why minimizes $Q = \sum e_i^2$ instead of $Q_1 = \sum |e_i|$

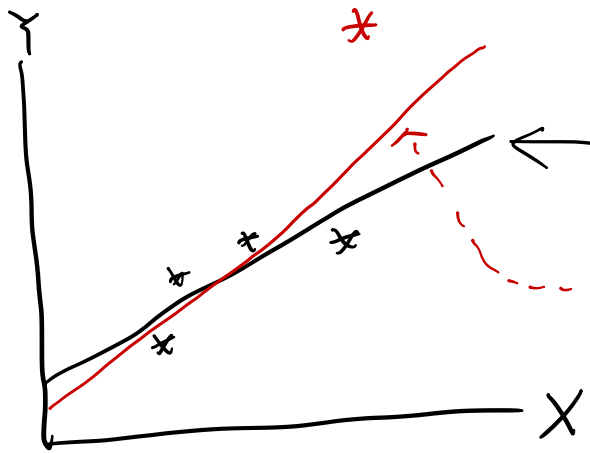
LSE

MAD/LAD

The reasons why we prefer Least Squares estimates instead of the minimum absolute deviation (MAD) estimates (or the Least Absolute Deviation (LAD) Estimator):

1. Squaring always gives a positive value and emphasizes large differences.
2. There is a strong link between OLS estimation and linear algebra. $\hat{Y}$ is a linear function of Y. In fact, it is a projection onto a subspace defined by the independent variables.
3. When ϵ is normally distributed, the maximum likelihood estimate is the OLS estimates (coming soon). OLS VS MLE
4. OLS is BLUE.
5. Least squares method gives a single unique answer, with least absolute value you can get infinite which makes it harder to interpret the results.
6. A lot of nice properties happen with OLS. Such as in the normal error model, it is more efficient. (Reference: Huber, Robust Statistics, p.10)
7. For a sum of independent variables, variances can be added but not for the standard deviation.

(1) Also implies that Q is sensitive to outliers while Q_1 is less sensitive to outliers.

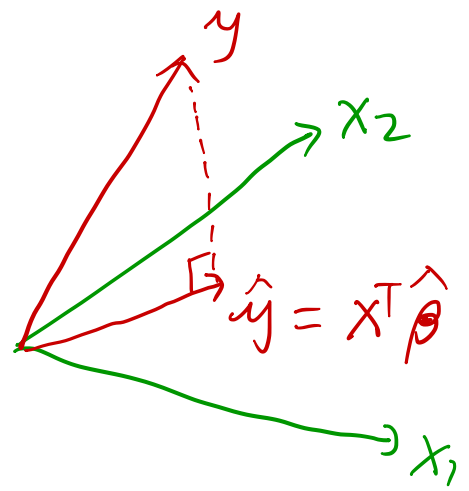


← fitted line without outlier *

fitted line with outlier

(2)

Predicted outcome $\hat{Y}$ are the orthogonal projection of y on subspace spanned by $\vec{X}$ ⟹



Different criterion of regression

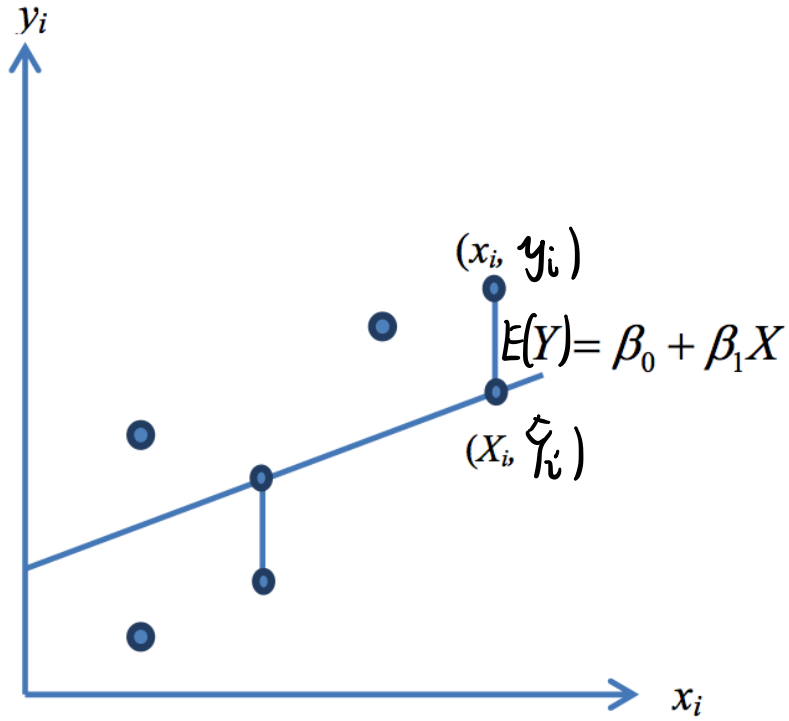
Direct regression

LSE: minimizes

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

LAD: minimizes

$Q = \sum_{i=1}^{n} |e_i|$



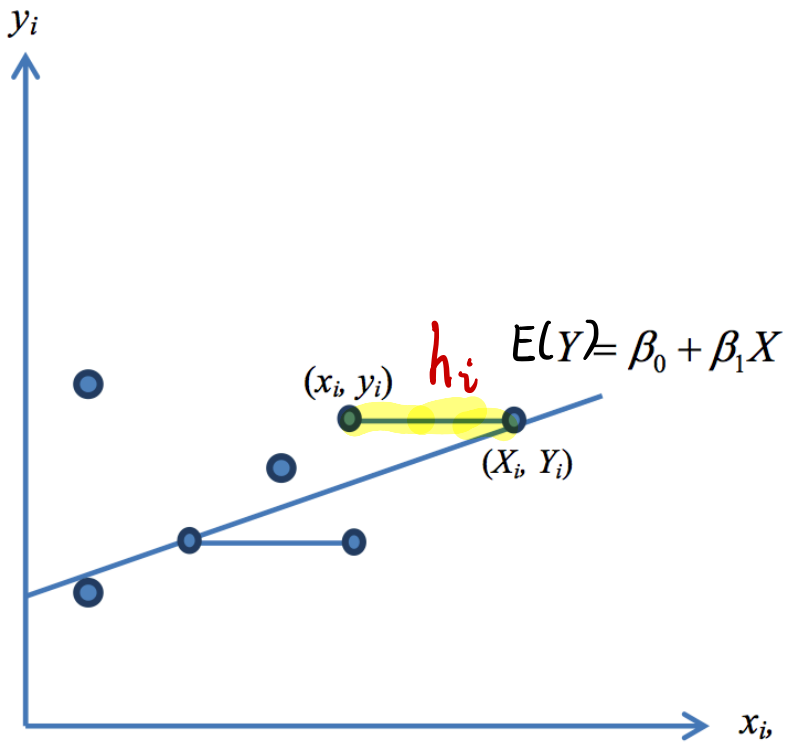
Direct regression

OLS or LS: refers to the same Least square procedure

- This method is often referred to the Ordinary Least Squares (OLS) estimation which minimizes $Q = \sum (Y_i - b_0 - b_1 X_i)^2$
- Least Absolute Deviation (LAD) Regression Method: the estimates of β_0, β_1 are chosen such that the sum of absolute deviations is minimum.

$Q = \sum |Y_i - b_0 - b_1 X_i|$

Reverse regression method

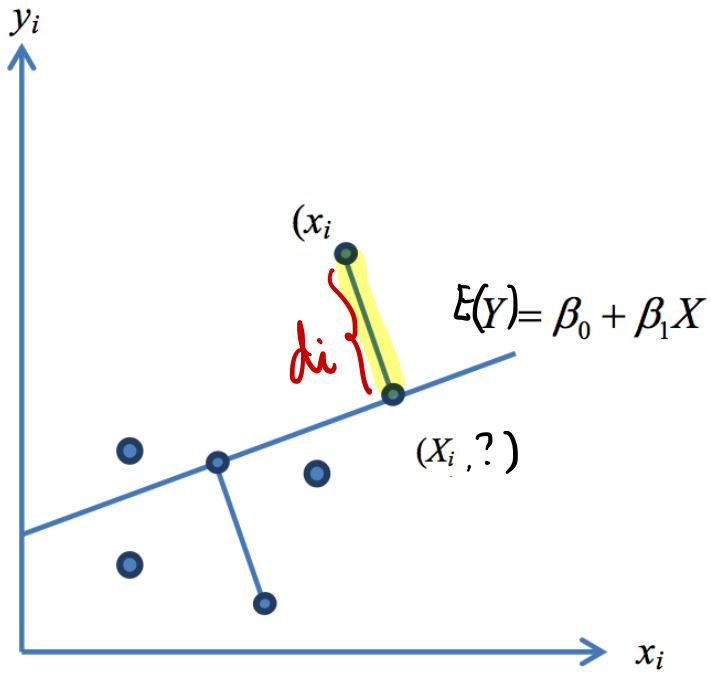


Reverse regression method

$Q = \sum_{i=1}^{n} h_i^2$

This reverse (or inverse) regression approach minimizes the sum of squares of horizontal distance between the observed data points and the fitted line to estimate β_0, β_1

Orthogonal regression method



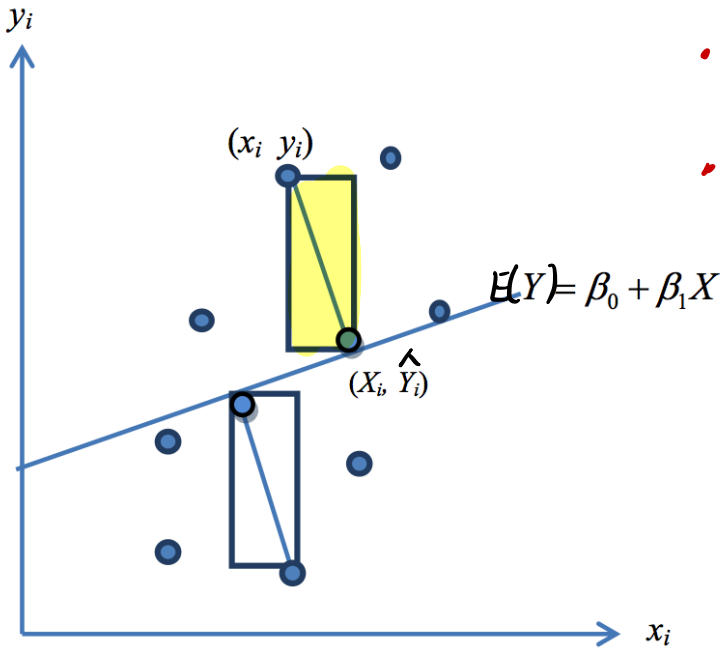
Major axis regression method

- both x_i and Y_i are random
- $Q = \sum_{i=1}^{n} d_i^2$

d_i = Perpendicular distance between Y_i and the fitted line.

Generally when uncertainties are involved in dependent and independent variables both, then orthogonal regression is more appropriate. The least squares principle in orthogonal regression minimizes the squared perpendicular distance between the observed data points and the fitted line to obtain the estimates of regression coefficients. This is also known as major axis regression method.

Reduced Major Axis regression



Reduced major axis method

- X_i, Y_i : both are random
- $Q = \sum_{i=1}^{n}$ (area of rect)$_i$

The area of rectangles defined between corresponding observed data points and nearest point on the fitted line can also be minimized. Such an approach is more appropriate when the uncertainties are present in study and explanatory variables both and is called as reduced major axis regression.

More discussion of different criterion of regression: Chapter 2.2,2.3,2.11-2.13, Linear Models and Generalizations, 3rd Edition, by C.R. Rao, H. Toutenburg, Shalabh, C.Heumann.

Normal Error Regression Model

Normal Error regression model

needs only GM assumptions (doesn't specify the distrn of ε) ↙

① $Y = \beta_0 + \beta_1 X + \varepsilon$
② $E(\varepsilon_i) = 0, \ \forall i$
③ $V(\varepsilon_i) = \sigma^2, \ \forall i$
④ $Cov(\varepsilon_i, \varepsilon_j) = 0, \ j \neq i$

- **Motivation:** Least Squared estimators b_0, b_1 are BLUE no matter what distribution of the error terms ϵ_i. However, to set up interval estimates and make tests, we need to make an assumption about the form of the distribution of the ϵ_i.
- **Normal error model description**

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, \ldots, n.$$

where $\epsilon \sim_{IID} N(0, \sigma^2)$. → stronger assumption than G-M.

① $E(\varepsilon_i) = 0$
② $V(\varepsilon_i) = 0$
③ $\varepsilon_i \perp\!\!\!\perp \varepsilon_j$
④ Normal

- **It follows that Y_i are independent normals (since mean varies with X_i).**

$$Y_i \sim_{\text{indep}} N(\beta_0 + \beta_1 X_i, \sigma^2) \qquad E(Y_i) \quad V(Y_i)$$

- $\varepsilon_i \sim N(0, \sigma^2)$, $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$
- The first 2 moments determine the Normal distrn.

Review: Likelihood function and MLE

- Likelihood of the parameters given the data is proportional to the probability of the data given the parameters. i.e. $L(\theta|y) = P(Y|\theta)$
- If $Y_i \sim_{IID} f(\theta), i = 1, \ldots n$, then the likelihood function is

not a pdf

$$L(\theta) = cf(y_1;\theta)f(y_2;\theta)\ldots f(y_n;\theta) = c\prod_{i=1}^{n} f(y_i;\theta).$$

It's information in relative sense.

where c is an arbitrary non-negative constant.

- The Maximum Likelihood Estimate for a parameter is the value that maximize the likelihood function.

$$\hat{\theta} = \max_{\theta\in\Omega} L(\theta)$$

- To maximize $L(\theta)$ is equivalent to maximize the log likelihood function $\ell(\theta)$

$$\ell(\theta) = \log L(\theta), \quad \hat{\theta} = \max_{\theta\in\Omega} \ell(\theta)$$

Symbol log in this course stands for the natural logarithm.

Maximum Likelihood Estimation: example

Consider a random sample $y_i, i = 1, \ldots, n$ from $N(\mu, \sigma^2 = 1)$

step 1

- Likelihood function of μ

$$L(\mu) = \prod_{i=1}^{n} f(y_i; \mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\{-\frac{(y_i - \mu)^2}{2}\} = c \exp\{-\frac{\sum_1^n (y_i - \mu)^2}{2}\}$$

step 2

- Log likelihood function of μ

$$\ell(\mu) = \log L(\mu) = \log c - \frac{\sum_1^n (y_i - \mu)^2}{2}$$

step 3

- Find the MLE of μ, set the derivative to μ to zero and solve it for MLE

$$\frac{\partial \ell(\theta)}{\partial \theta} = \sum_1^n (y_i - \mu) = 0$$

- $\hat{\mu}_{MLE} = \bar{y}$.

$\frac{\partial^2 \ell(\theta)}{\partial \theta^2} = -n < 0 \Rightarrow L(\mu)$ at $\bar{Y}$ reaches it maximum.

Assume p.d.f $f(y;\theta)$ is smooth enough

standard procedure to find MLE:

① Likelihood $L(\theta; y_1, \cdots y_n) = C \prod_{i=1}^{n} f(y_1;\theta) \cdots f(y_n;\theta)$

$\Downarrow$ monotonic transf.

② log-likelihood $\ell(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(y_i;\theta)$

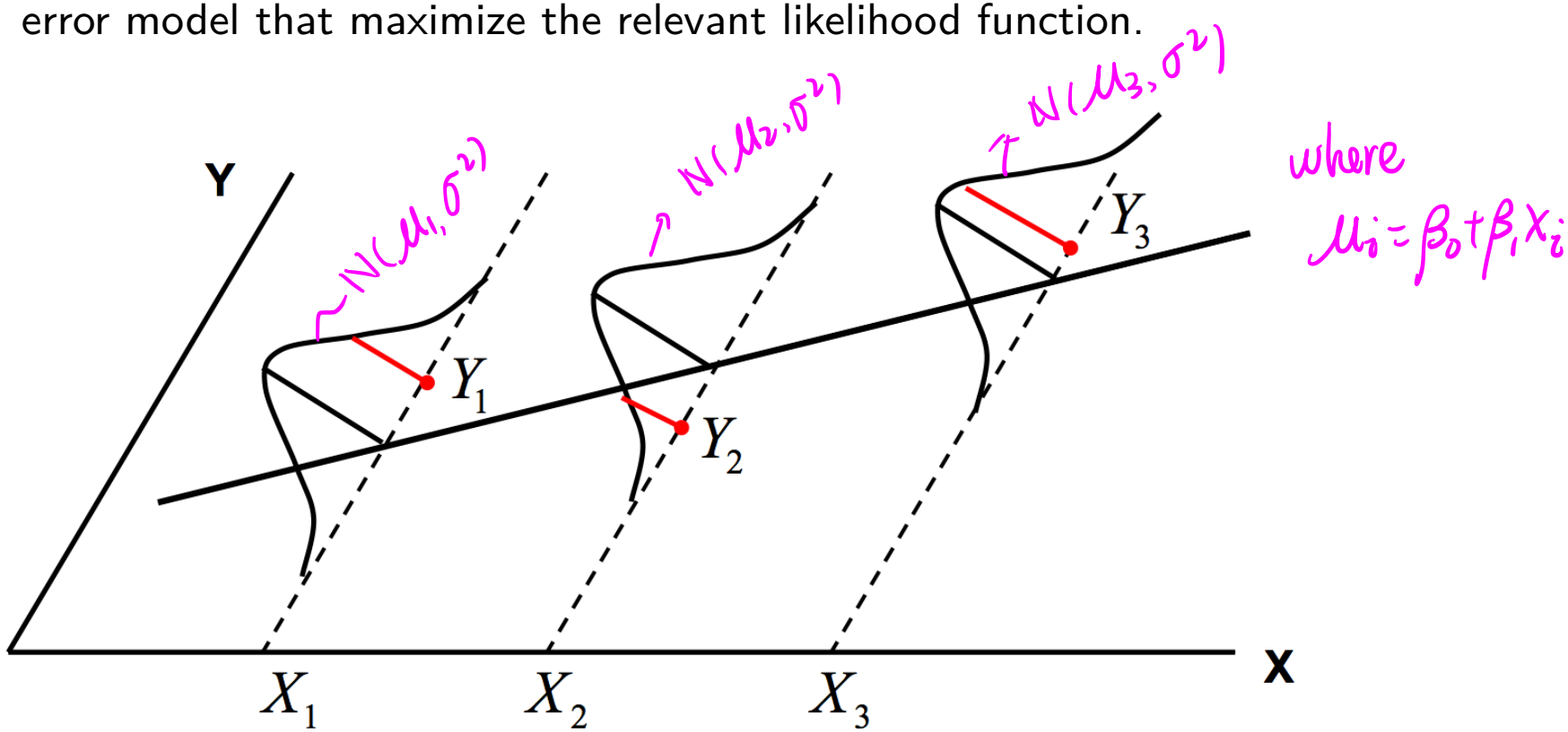
③ $\frac{\partial \ell(\theta)}{\partial \theta} = 0$, if $\theta = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}$, $\ell_\theta(\theta) = \begin{pmatrix} \partial \ell(\theta)/\partial \theta_1 \\ \partial \ell(\theta)/\partial \theta_2 \end{pmatrix} = 0$

$\rightarrow \hat{\theta} \leftarrow$

④ check $\left.\frac{\partial^2 \ell(\theta)}{\partial \theta^2}\right|_{\theta=\hat{\theta}} < 0$

Maximum Likelihood Estimation for Regression

- Similarly, we find the values of parameters $\theta = (\beta_0, \beta_1, \sigma^2)$ in the normal error model that maximize the relevant likelihood function.



- $Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$, the density function at y_i is

$$f(y_i; \beta_0, \beta_1, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\}$$

MLE for Regression

(1) • Likelihood function

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} f(y_i; \theta) = \{\frac{1}{2\pi\sigma^2}\}^{n/2} \exp\{-\frac{\sum_1^n (y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\}$$

(2) • Log likelihood function

$$\ell(\beta_0, \beta_1, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{\sum_1^n (y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}$$

(3) Partial differentiation of $\ell(\beta_0, \beta_1, \sigma^2)$ yields

$\hat{\beta}_0 = \bar{Y} - b_1\bar{X}$

$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}}$ $\Leftarrow$

$\hat{\sigma}^2 = \frac{\Sigma e_i^2}{n}$

$$\begin{cases} \frac{\partial \ell(\beta_0, \beta_1, \sigma^2)}{\partial \beta_0} = \frac{1}{\sigma^2}\sum(y_i - \beta_0 - \beta_1 x_i) & (1) \\ \frac{\partial \ell(\beta_0, \beta_1, \sigma^2)}{\partial \beta_1} = \frac{1}{\sigma^2}\sum x_i(y_i - \beta_0 - \beta_1 x_i) & (2) \\ \frac{\partial \ell(\beta_0, \beta_1, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum(y_i - \beta_0 - \beta_1 x_i)^2 & (3) \end{cases}$$

MLE vs LSE for Regression

Set partial derivatives equal to zero and solve them to obtain the MLE of $\beta_0, \beta_1, \sigma^2$.

Parameters	MLE	Same as LSE?
β_0	$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$	Same. $\hat{\beta}_0 = b_0$
β_1	$\hat{\beta}_1 = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_1^n (x_i - \bar{x})^2}$	Same. $\hat{\beta}_1 = b_1$
σ^2	$\hat{\sigma}^2 = \frac{\sum_1^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n}$	Different. $\hat{\sigma}^2 = \frac{n-2}{n} MSE$

Note that the MLE of σ^2 is a biased estimator of the error term variance.

Upcoming: Inference for SLR

- We knew how to point estimate of β_0, β_1 in SLR from data, Using Ordinary Least Squares (OLS) or MLE.
- Now we take up the inferences concerning the regression parameters β_0, β_1
  - How accurate are their estimates?
  - How to obtain an interval estimate for each of them?
  - How to test a specific parameter value of interest?
- Use Statistical Inference
  - Confidence intervals
  - Hypothesis Tests

Before we give answers, we will have a short review of statistical inference.

Review on Statistical Inference

Review: Central Limit Theorem (sample mean distribution)

CLT: $X_1, \cdots X_n \sim$ one distrn with $\mu, \sigma^2 < 0$, then $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = \frac{\sum X_i - n\mu}{\sigma\sqrt{n}} \overset{D}{\sim} N(0,1)$

1. Consider a random sample $X_1, \ldots, X_n$ from a distribution with mean μ and variance $\sigma^2 < \infty$.
2. Point estimate for the unknown mean is the sample mean, $\bar{X} = \sum X_i/n$.
3. Central Limit Theorem (CLT): as n approaches infinity, the random variables $\sqrt{n}(\bar{X} - \mu)$ converge in distribution to a normal $N(0, \sigma^2)$.

$$\sqrt{n}(\bar{X} - \mu) = \frac{1}{\sqrt{n}}(\sum X_i - n\mu) \to_D N(0, \sigma^2)$$

or put it in a brief way, $\bar{X} \to_D N(\mu, \sigma^2/n)$. ← Normal is the limiting distrn

4. Note that if $X_i \sim_{iid} N(\mu, \sigma^2)$, then $\bar{X}$ is exactly Normally distributed.

$$\sqrt{n}(\bar{X} - \mu) \sim N(0, \sigma^2)$$

Make sure you understand the difference between 3 and 4.

↑ approximation. ← exact.

$$\frac{\hat{\theta}-\theta}{\sqrt{Var(\hat{\theta})}} \xrightarrow[n\to\infty]{D} N(0,1), \text{ where } \theta = E(\hat{\theta})$$

CLT: $\theta = \mu$, $\hat{\theta} = \bar{X}$, $E(\hat{\theta}) = E(\bar{X}) = \mu$

$$\Rightarrow Var(\hat{\theta}) = Var(\bar{X}) = Var\left(\frac{1}{n}\sum_1^n X_i\right) = \frac{1}{n^2}(\sigma^2 + \cdots + \sigma^2) = \frac{\sigma^2}{n}$$

$$\Rightarrow \frac{\hat{\theta}-\mu}{\sqrt{\sigma^2/n}} = \frac{\sqrt{n}(\hat{\theta}-\mu)}{\sigma} * \frac{n}{n}$$

$$= \frac{\sqrt{n}(\sum X_i - n\mu)}{n\sigma}$$

$$= \frac{\sum X_i - n\mu}{\sqrt{n}\,\sigma} \xrightarrow{D} N(0,1)$$

all follow the same formula.

$E(\sum X_i) = n E(X_1) = n\mu$

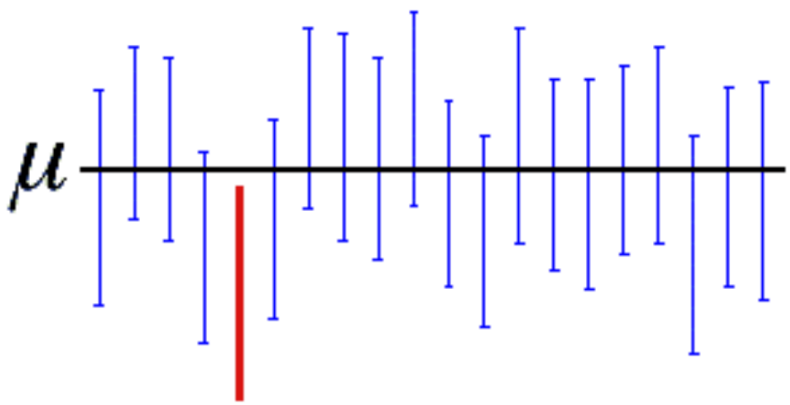
$Var(\sum X_i) = n\sigma^2 \to \sqrt{n\sigma^2} = \sqrt{n}\,\sigma$

Review: Confidence Interval for μ given σ^2

Two-side 95% CI for μ (assume σ^2 is known)

$$\bar{X} \pm Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} = (\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}})$$

- A confidence interval (CI) is a type of interval estimate of a population parameter. It is an observed interval which is calculated from the observations, in principle different from sample to sample.
- How frequently the observed interval contains the parameter is determined by the confidence level or confidence coefficient.
- More specifically, confidence level suggests that if CI are constructed across many separate data analyses of replicated experiments, the proportion of such intervals that contain the true value of the parameter will match the given confidence level.



Replicate the same experiments 100 times, (1-2)*100 times, the CI contains the true μ.

Example 2.1: Confidence Interval for μ given σ^2

- Observed a random sample (5,6,2,5,3) of size n=5, further assume $\sigma^2 = 1$.
- 95% CI (two-sided) for μ

$$
\begin{aligned}
0.95 &= P\left(\frac{|\bar{X} - \mu|}{1/\sqrt{n}} < 1.96\right) \\
&= P\left(|\bar{X} - \mu| < 1.96\frac{1}{\sqrt{n}}\right) \\
&= P\left(\bar{X} - 1.96\frac{1}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{1}{\sqrt{n}}\right)
\end{aligned}
$$

$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \to N(0,1)$

$P\left(\left|\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right| \le 1.96\right) = 0.95$

$P\left(\bar{X} - 1.96\frac{1}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{1}{\sqrt{n}}\right) = 0.95$

- For the given data, we find $\bar{X} = 4.2$ and the 95% CI for μ

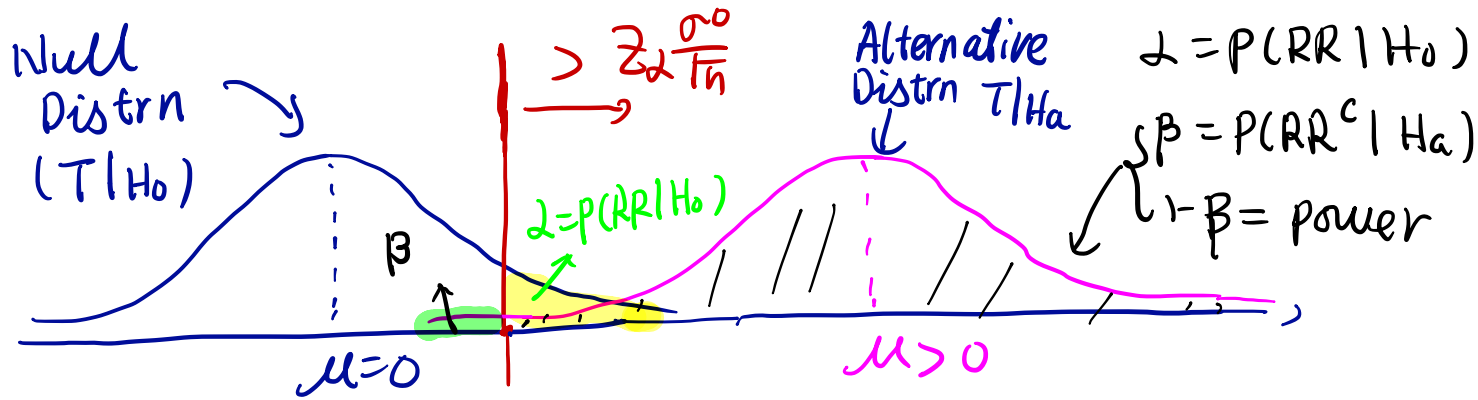
$$(4.2 - 1.96/\sqrt{5}, 4.2 + 1.96/\sqrt{5}) = (3.323461, 5.076539)$$

Review: Hypothesis Testing

$X_1, \cdots X_n \sim N(\mu, 1)$

① $H_0: \mu = 0$, $H_a: \mu > 0$ ② test statistic $T = \frac{\bar{x} - \mu}{\sigma_0/\sqrt{n}}$ ③ $T|_{H_0} = \frac{\bar{x}-0}{\sigma_0/\sqrt{n}} \sim N(0,1)$ (distrn under H_0)

- A hypothesis test is a statistical test that is used to determine whether there is enough evidence in a sample of data to infer the null hypothesis is true.
- P-value for a testing indicates where the observed data sits on the null distribution.
- In some textbook, they view P-value as a way to determine "likely" or "unlikely" of observing a more extreme statsitic in the direction of alternative hypothesis than the one observed assuming null hypothesisis were true.



Reject H_0 when $\frac{\bar{x}-0}{\sigma^0/\sqrt{n}} > z_\alpha \Leftrightarrow \bar{x} \geq z_\alpha \frac{\sigma^0}{\sqrt{n}}$

Review: Hypothesis Testing

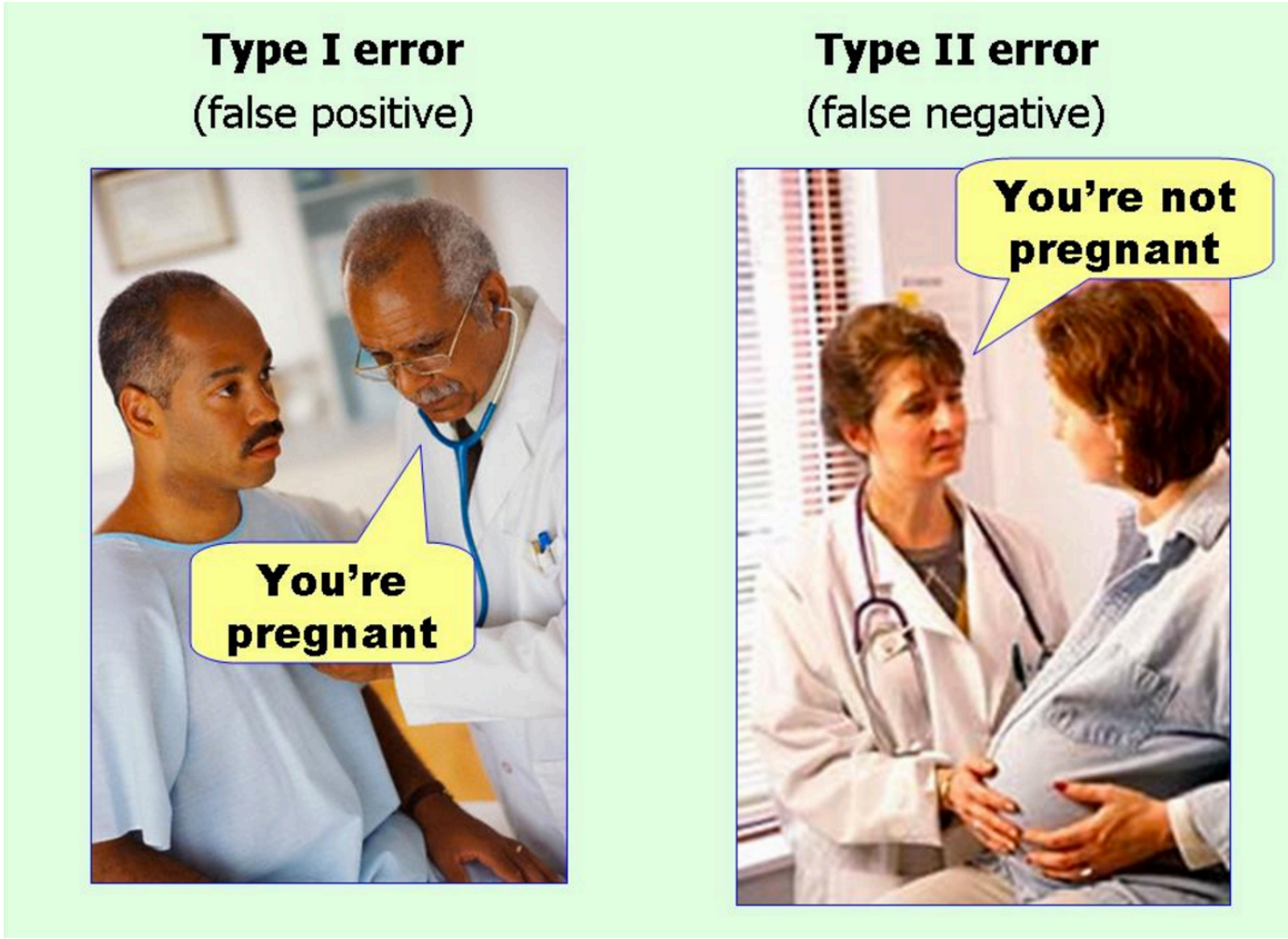
R.R. = Rejection Region

		Reality/ Truth: Null is True H0: T	Reality/ Truth: Null is False H0: F
Decision	Reject Null	Type I error "False Positive"	Correct Decision
	Fail to Reject Null	Correct Decision	Type II error "False Negative"

- Type I error = P(reject $H_0|H_0$ is true). $= P(R.R. | H_0) = \alpha$
- Type II error = P(fail to reject $H_0|H_0$ is false)= P(fail to reject $H_0|H_a$ is true) $= P(RR^c | H_a) = \beta$
- Significance level α: maximum allowable probability of making type I error (usually 5%)
- Test statistics: a statistics (function of data variable) whose distribution under H_0 is known.

Null distribution

Review: Hypothesis Testing



https://effectsizefaq.com/index/faqs/page/3/

Review: Hypotheis Testing

H_0: $\theta \in \Omega_0$

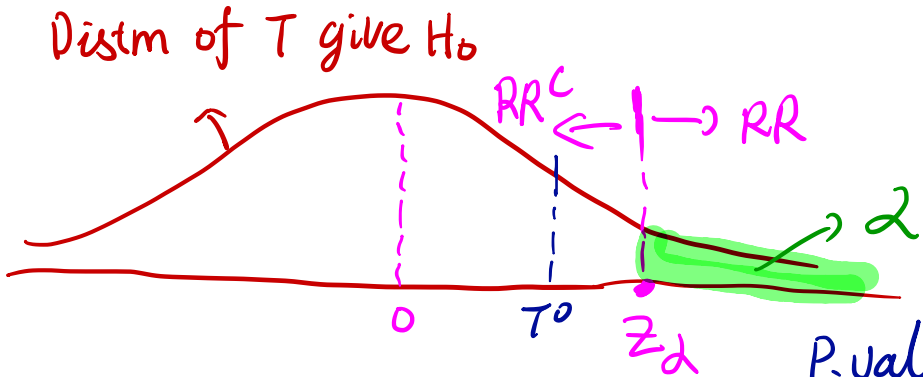
H_a: $\theta \in \Omega / \Omega_0$

- If we reject H_0 (i.e. H_0 is false), then we accept H_a.
- However, if we fail to reject H_0, it doesn't imply that H_0 is true.

- P-value: the probability (under H_0) that the test statistics is equally or more extreme than the one actually observed.
  - If P-value$< \alpha$, reject H_0. $\Rightarrow$ H_a is True.
  - If P-value$> \alpha$, do not reject H_0. This doesn't imply that H_0 is true.
- Smaller P-value $\rightarrow$ less likely H_0 is to be true
  - p-value$< 0.001 \rightarrow$ strong evidence against H_0
  - p-value$< 0.05 \rightarrow$ moderate evidence against H_0
  - p-value$< 0.1 \rightarrow$ weak evidence against H_0
  - p-value$> 0.1 \rightarrow$ no evidence against H_0

$H_0: \mu = 0$

$H_a: \mu > 0$



$T^0 = \frac{\bar{x}^0 - 0}{\sigma^0/\sqrt{n}}$

= observed T (plugging data)

P.value $= P(Z > T^0) > \alpha$

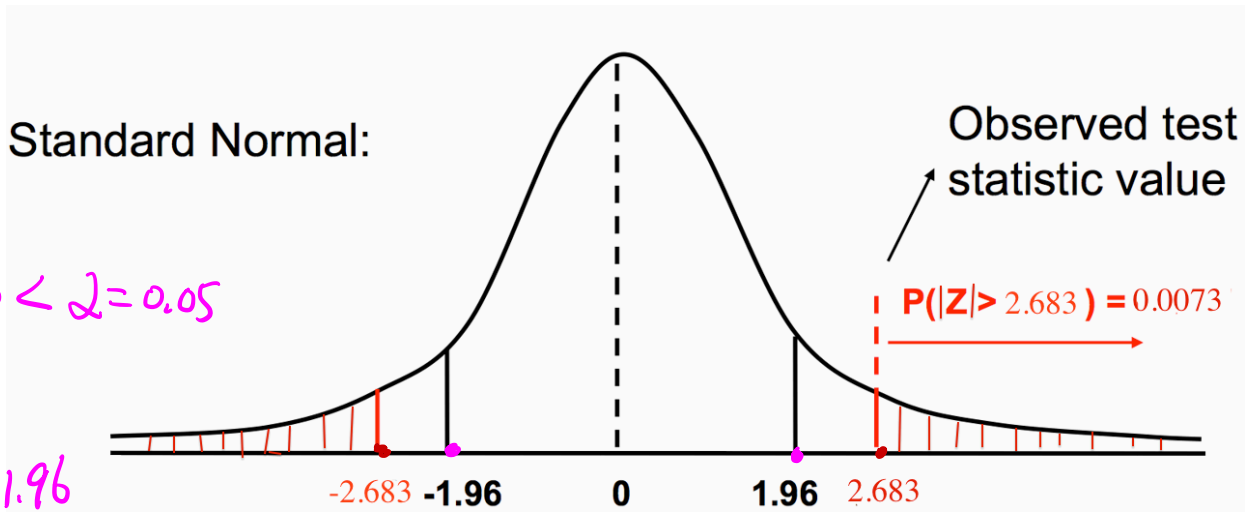
Example 2.1: Hypothesis Testing

- Observed a random sample (5,6,2,5,3) of size n=5, assume $\sigma^2 = 1$.
- 95% CI (two-sided) for μ: (3.323461,5.076539) ← $\bar{X} \pm 1.96 \frac{1}{\sqrt{n}}$
- Now interested in testing $H_0 : \mu = 3, \quad H_a : \mu \neq 3$
  - test statistics:

$$T = \frac{\bar{X} - \mu}{1/\sqrt{5}} \sim_{H_0} Z = N(0,1)$$

  - given data, under H_0, $T_{obs} = (4.2 - 3)/\sqrt{5} = 2.683$ (H_0)
  - p-value= $P(|Z| > T_{obs}) = P(|Z| > 2.683) = 0.0073$
  - make decision: we have strong evidence to reject the H_0 since p-value is less than 0.01, or because the $T_{obs} > 1.96$ for $\alpha = 5\%$.
  - Note that $\mu = 3 \notin$ 95% CI: (3.323461,5.076539). ③

Methods ①, ②, ③: consistent.



① P.value = 0.0073 < α = 0.05

② T_{obs} = 2.683

$|T_{obs}| > z_{\alpha/2} = 1.96$

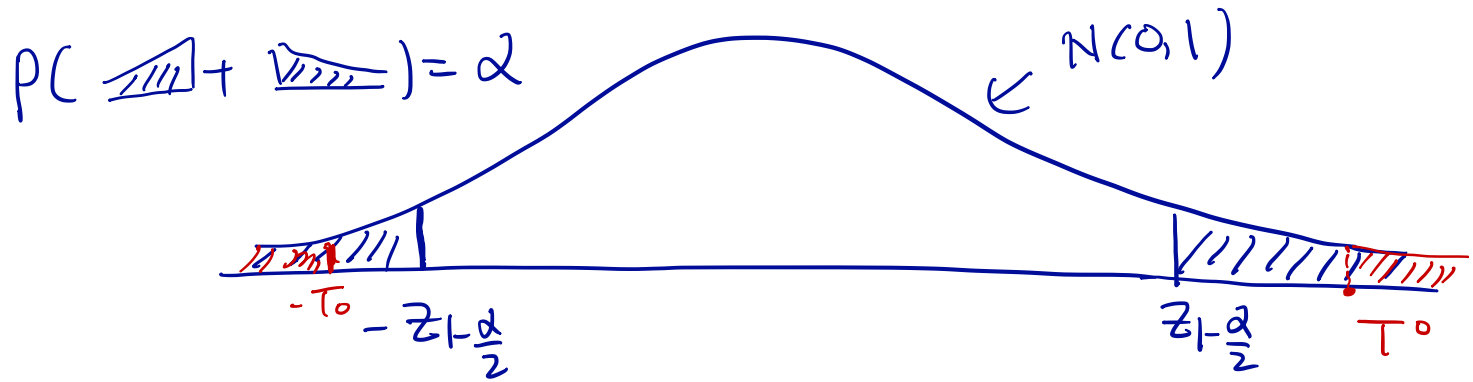
①② give consistent result: reject H_0.

Review: P-value and CI agree about statistical significance

- You can use either P values or confidence intervals to determine whether your results are statistically significant. (Reject H_0)
- The confidence level is equivalent to $1 - \alpha$ level. So, if your significance level is 0.05, the corresponding confidence level is 95%.
  - If the P value is less than your significance (α) level, the hypothesis test is statistically significant.
  - If the CI does not contain the null hypothesis value (H_0 value), the results are statistically significant.
  - If P-value$< \alpha$, the CI will not contain the null hypothesis value.
- To understand why the results always agree, let's recall how both the significance level and confidence level work.
  - The significance level defines the distance the sample mean must be from H_0 to be considered statistically significant.
  - The confidence level defines the distance for how close the confidence limits are to sample mean.
  - Both the significance level and the confidence level define a distance from a limit to a mean. Guess what? The distances in both cases are exactly the same!

Assume two-sided test: $H_0: \theta = \theta_0$ $H_a: \theta \neq \theta_a$

$T|_{H_0} \sim$ A known distn, say $N(0,1)$



3 ways to make decision: Accept or Reject H_0

$T^\circ = T$(plugging in observed data)$|_{H_0}$

① P-value $= P(|Z| > T_0)$, P-value $< \alpha$: Reject

② $|T^\circ| > z_{1-\frac{\alpha}{2}}$: Reject

③ $\theta_0 \notin (1-\alpha)$ confidence interval.

Upcoming topic and reading

- Review of distribution theory.
- Inference concerning β_0, β_1.
- Interval estimation of mean response.
- Prediction interval.
- Difference between prediction interval and confidence interval.

Reading for next week: CH1: 1.8, CH2: 2.1,2.2, 2.4,2.5,2.6, 2.7

Practice problems after Week 1- lecture B

1. Keep trying the practice problem
   - 1.3, 1.5, 1.6, 1.7, 1.8, 1.11, 1.16, 1.18, 1.20(*), 1.21(*), 1.24(*), 1.29, 1.30, 1.33, 1.36, 1.39a, 1.40, 1.41a.
   - For questions marked (*), the SAS code & output is posted with the solutions.
   - You only need to interpret that output.
2. Show that
   - $Cov(\bar{Y}, b_1) = 0$ (Hint: $Cov(\bar{Y}, b_1) = Cov\{\sum \frac{1}{n} Y_i, \sum k_i Y_i\}$)
   - Using above prove that $V(b_0) = \sigma^2(1/n + \bar{X}^2/S_{xx})$
   - $Cov(b_0, b_1) = -\sigma^2 \bar{X}/S_{xx}$
   - $Cov(\bar{Y}, b_0) = \sigma^2/n$
3. A no-intercept model is: $Y_i = \beta_1 X_i + \epsilon_i, \epsilon \sim_{iid} N(0, \sigma^2)$
   - find the LS estimate of β_1, denoted it as b_1
   - find the variance of b_1, $V(b_1)$
   - find an unbiased estimate of $V(b_1)$