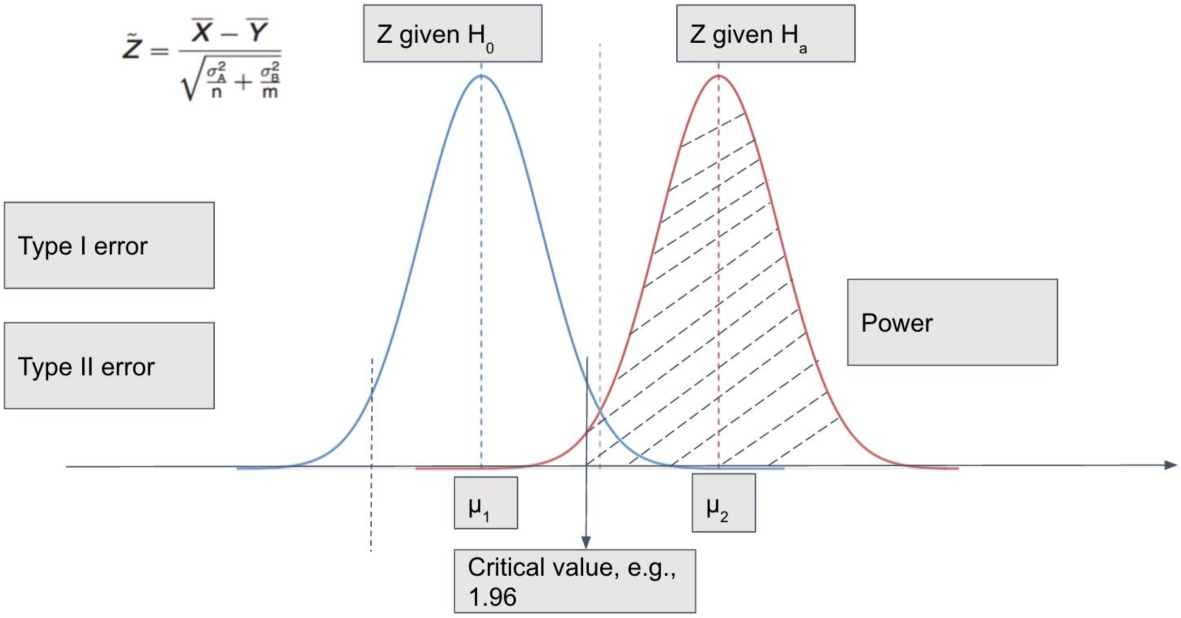
# Hypothesis testing and linear regression

## Question 1

Suppose we have two samples from two population with sample size n and m, one with mean $\mu_1$, variance as $\sigma_1^2$, the other one with mean $\mu_2$, variance $\sigma_2^2$. When detecting the difference in sample mean: $\delta = \mu_1 - \mu_2$, we want power is at least 0.8.
If we know $\delta$= 1, $\sigma_1^2$ = $\sigma_2^2$ = 1, n = m,

- Can you calculate minimal n?
- How n change along with $\delta$ ?
- How n change along with $\sigma_1^2$?

Answer


$\tilde{Z} = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_A^2}{n} + \frac{\sigma_B^2}{m}}}$
Z given $H_0$
Z given $H_a$
Type I error
Type II error
Power
$\mu_1$
$\mu_2$
Critical value, e.g., 1.96

From above plot, blue and red are the distribution for $\underline{X}$ and $\underline{Y}$. E($\underline{X}$) = $\mu_1$, E($\underline{Y}$) = $\mu_2$, using pooled variance, $\sigma_{pooled}^2 \quad = \sigma_1^2/n \ + \sigma_2^2/m$ = 2 $\sigma^2/n$.
From blue line, critical value = $\mu_1$+ $z_{1-\alpha/2} \cdot \sigma_{pooled}$
From red line, critical value = $\mu_2$+ $z_{1-power} \cdot \sigma_{pooled}$
So $n \ = [(z_{1-\alpha/2} \ - \ z_{1-power}) \ / \ (\mu_2 \ - \ \mu_1)\ ]^{\ 2}$ $\cdot$ 2 = 16

## Question 2

A new casino game involves rolling 3 dice. The winnings are directly proportional to the total number of sixes rolled. Suppose a gambler plays the game 101 times, with the following observed counts:

| Number of Sixes | Number of Rolls |
|---|---|
| 0 | 48 |
| 1 | 35 |
| 2 | 15 |
| 3 | 3 |

Test if this is fair dice. What test to use? Calculate stats and p value

Answer:

- Use Chi-square test

$$H_0: this\ is\ fair\ dice$$
$$H_a: this\ is\ not\ fair\ dice$$

- Stats and p value

$$\mathrm{E(0\ of\ sixes)} = 101\times(5/6)^3$$
$$\mathrm{E(1\ of\ sixes)} = 101\times C_3^1(5/6)^2(1/6)$$
$$\mathrm{E(2\ of\ sixes)} = 101\times C_3^2(1/6)^2(5/6)$$
$$\mathrm{E(3\ of\ sixes)} = 101\times(1/6)^3$$

| Number of sixes | Observed number of rolls, $O_i$ | Expected number of rolls, $E_i$ |
|---|---|---|
| 0 | 48 | 58.45 |
| 1 | 35 | 35.07 |
| 2 | 15 | 7.01 |
| 3 | 3 | 0.47 |

$$\chi^2_{df=3} = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} = 24.68$$

$$\mathrm{p - value} = 1.804323\mathrm{e} - 05 < 0.01$$

(in R: pchisq(q = chi_stat, df = df, lower.tail=FALSE))

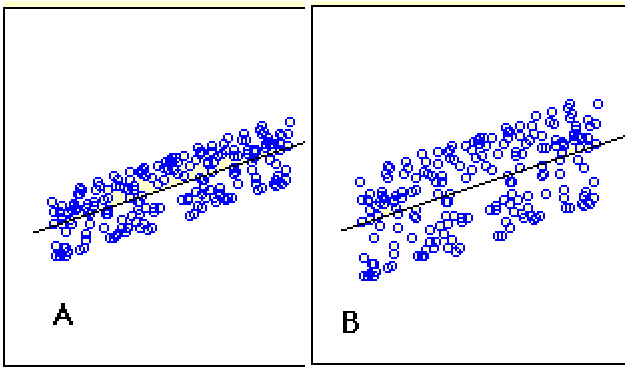
# Question 3

Below graphs show two fitted regression lines (A & B) on randomly generated data. Now, I want to find the sum of residuals in both cases A and B.

Note:

Scale is same in both graphs for both axis.

X axis is independent variable and Y-axis is dependent variable.



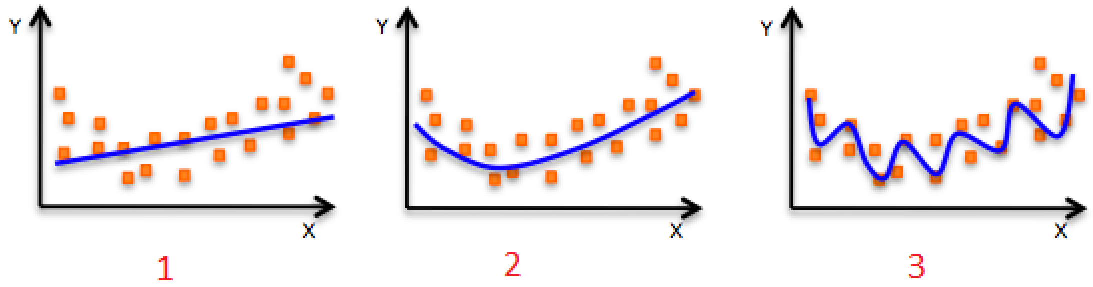
Which of the following statement is true about sum of residuals of A and B?
A) A has higher than B
B) A has lower than B
C) Both have same
D) None of these

Solution: C

Sum of residuals always zero.

# Question 4

The following visualization shows the fit of three different models (in blue line) on same training data. What can you conclude from these visualizations?



1. The training error in first model is higher when compared to second and third model.
2. The best model for this regression problem is the last (third) model, because it has minimum training error.
3. The second model is more robust than first and third because it will perform better on unseen data.
4. The third model is overfitting data as compared to first and second model.
5. All models will perform same because we have not seen the test data.

A. 1 and 3
B. 2

C. 1, 3 and 4
D. Only 5

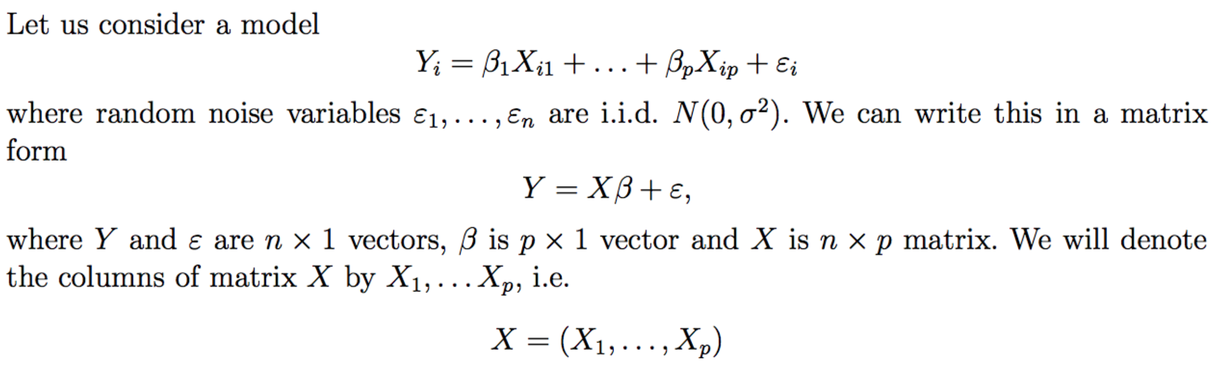
Solution: C

The trend of the data looks like a quadratic trend over independent variable X. A higher degree (Right graph) polynomial might have a very high accuracy on the train population but is expected to fail badly on test dataset. But if you see in left graph we will have training error maximum because it under-fits the training data.

# Question 5

Using MLE to achieve coefficient estimator for multiple linear regression (i.e., more than one feature in model)

Answer:



Let us consider a model

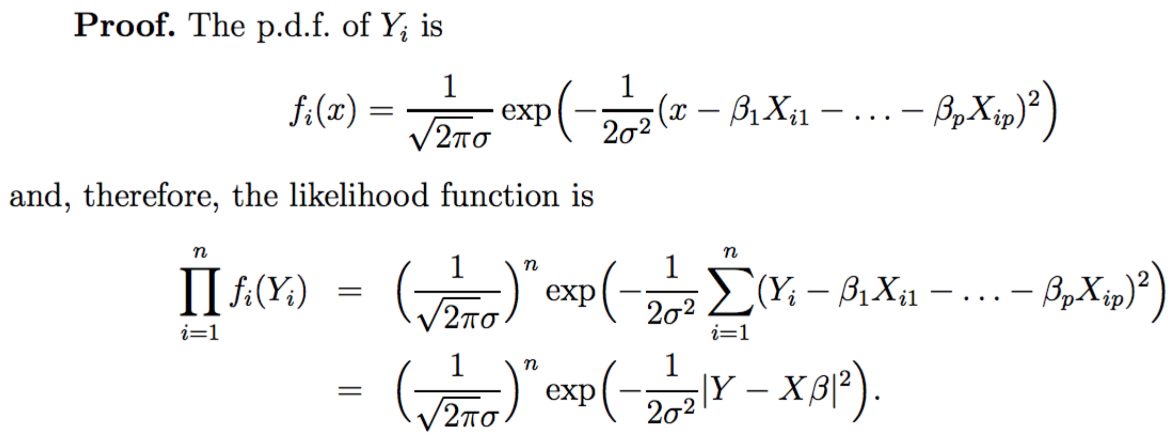
$$Y_i = \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + \varepsilon_i$$

where random noise variables $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. $N(0, \sigma^2)$. We can write this in a matrix form

$$Y = X\beta + \varepsilon,$$

where $Y$ and $\varepsilon$ are $n \times 1$ vectors, $\beta$ is $p \times 1$ vector and $X$ is $n \times p$ matrix. We will denote the columns of matrix $X$ by $X_1, \ldots X_p$, i.e.

$$X = (X_1, \ldots, X_p)$$



**Proof.** The p.d.f. of $Y_i$ is

$$f_i(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(-\frac{1}{2\sigma^2}(x - \beta_1 X_{i1} - \ldots - \beta_p X_{ip})^2\Big)$$

and, therefore, the likelihood function is

$$\begin{aligned} \prod_{i=1}^{n} f_i(Y_i) &= \Big(\frac{1}{\sqrt{2\pi}\sigma}\Big)^n \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_1 X_{i1} - \ldots - \beta_p X_{ip})^2\Big) \\ &= \Big(\frac{1}{\sqrt{2\pi}\sigma}\Big)^n \exp\Big(-\frac{1}{2\sigma^2}|Y - X\beta|^2\Big). \end{aligned}$$

To maximize the likelihood function, first, we need to minimize $|Y - X\beta|^2$. If we rewrite the norm squared using scalar product:

$$\begin{aligned} |Y - X\beta|^2 &= (Y - \sum_{i=1}^{p} \beta_i X_i, Y - \sum_{i=1}^{p} \beta_i X_i) \\ &= (Y, Y) - 2\sum_{i=1}^{p} \beta_i (Y, X_i) + \sum_{i,j=1}^{p} \beta_i \beta_j (X_i, X_j). \end{aligned}$$

Then setting the derivatives in each $\beta_i$ equal to zero

$$-2(Y, X_i) + 2\sum_{j=1}^{p} \beta_j (X_i, X_j) = 0$$

we get

$$(Y, X_i) = \sum_{i=1}^{p} \beta_j (X_i, X_j) \quad \text{for all} \quad i \leq p.$$

In matrix notations this can be written as $X^T Y = X^T X\beta$. Matrix $X^T X$ is a $p \times p$ matrix. Is is invertible since by assumption $X$ has rank $p$. So we can solve for $\beta$ to get the MLE

$$\hat{\beta} = (X^T X)^{-1} X^T Y.$$

# Question 6

Use new features you generated, either delete or impute existing features with missing values. Build the best linear regression model to explain interest rate.