بسمه تعالی

موضوع

محاسبات آماری در R

گردآورنده

سارا معصومی

استاد راهنما

سرکار خانم فرزانه احمدپور

تیر ۱۴۰۱

فهرست

عنوان **صفحه**

"سری زمانی در R"

داده های زیر مربوط به تعداد تماس های گرفته شده با یک مرکز مشاوره طی ۳۱ روز میباشد.که مربوط به ماه فروردین سال ۱۴۰۱ است.

مشاهدات	ردیف	مشاهدات	ردیف
۳۱	۱۷	۱۱	۱
۳۵	۱۸	۱۲	۲
۴۱	۱۹	۱۳	۳
۴۱	۲۰	۱۴	۴
۴۰	۲۱	۱۴	۵
۳۸	۲۲	۱۵	۶
۴۰	۲۳	۱۶	۷
۴۲	۲۴	۱۵	۸
۴۵	۲۵	۱۷	۹
۴۴	۲۶	۱۸	۱۰
۴۷	۲۷	۲۰	۱۱
۴۹	۲۸	۲۱	۱۲
۵۱	۲۹	۲۳	۱۳
۵۵	۳۰	۲۵	۱۴
۲۹	۳۱	۲۶	۱۵
		۲۸	۱۶

۱. **ابتدا لازم است مشاهدات را با استفاده از دستور زیر به نرم افزار معرفی کنیم.**

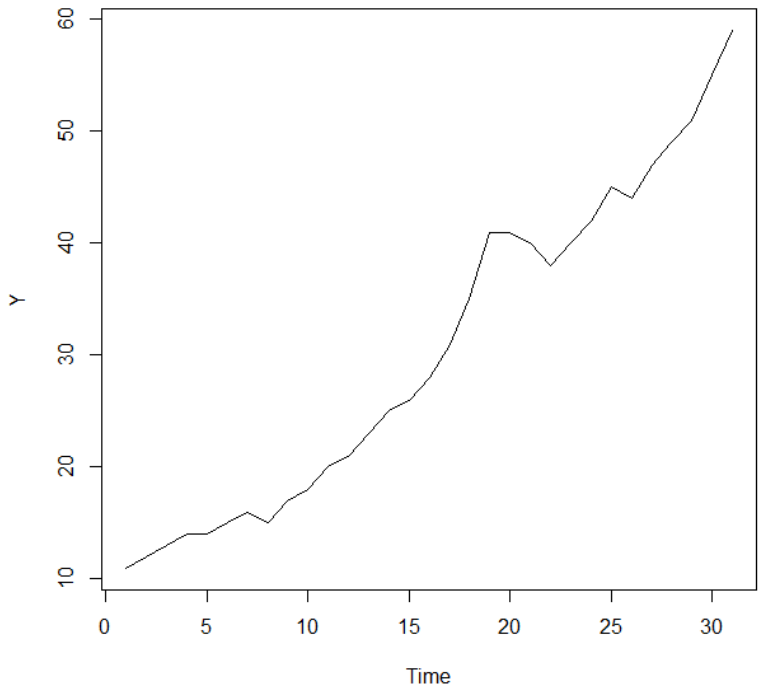
```
x <- c (11,12,13,14,14,15,16,15,17,18,20,21,23,25,26,28,31,35,41,41,40,38,40,42,45,44,47,49,51,55,59)
```

۲. **حال با استفاده از دستورات زیر نمودار سری زمانی مربوط به مشاهدات را رسم میکنیم.**

```
Y <- ts (x, start=1, end=31)
plot(Y)
```

۳. **بعد از ران کردن دستورات بالا میتوانیم در خروجی نمودار سری زمانی را ببینیم.**

خروجی:

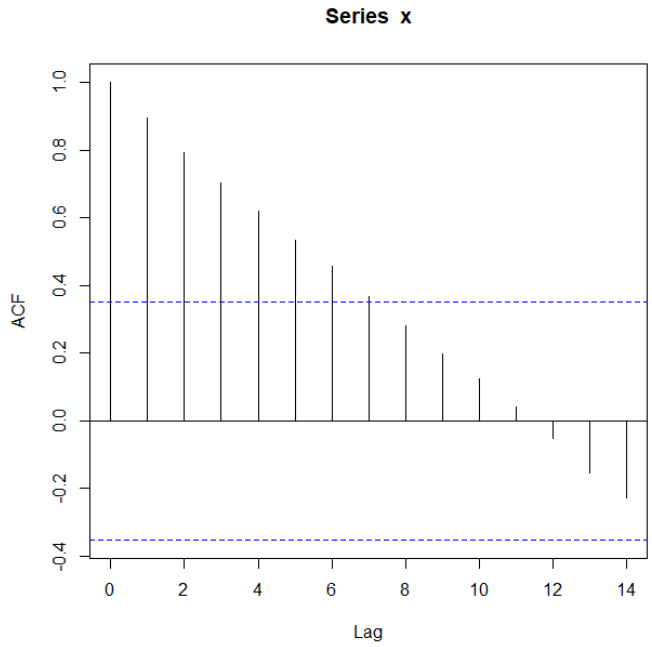


تفسیر: همانطور که در نمودار فوق میبینیم یک روند صعودی بین مشاهدات وجود دارد یعنی با گذر زمان تعداد تماس های دریافتی افزایش یافته است.

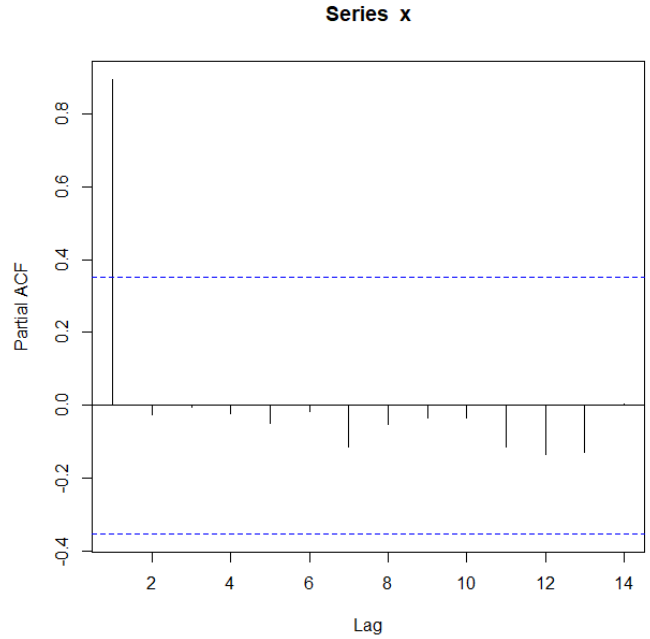
۴. با استفاده از دستورات زیر میتوانیم نمودار تابع خودهمبستگی و تابع خودهمبستگی جزئی را به ترتیب رسم کنیم.

acf(x)

pacf(x)

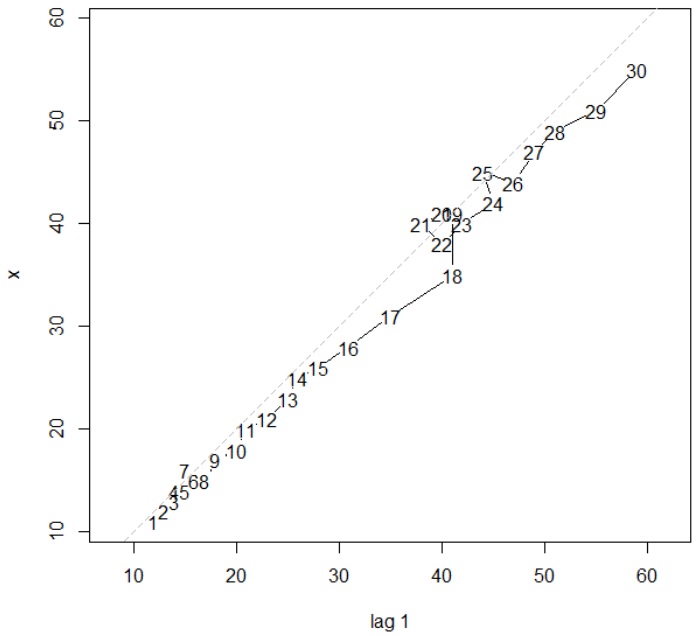


همانطور که در نمودار فوق میبینیم بیشتر از ۵٪ مقادیر خارج بازه ی اطمینان قرار دارند بنابراین سری ما ایستا نیست.

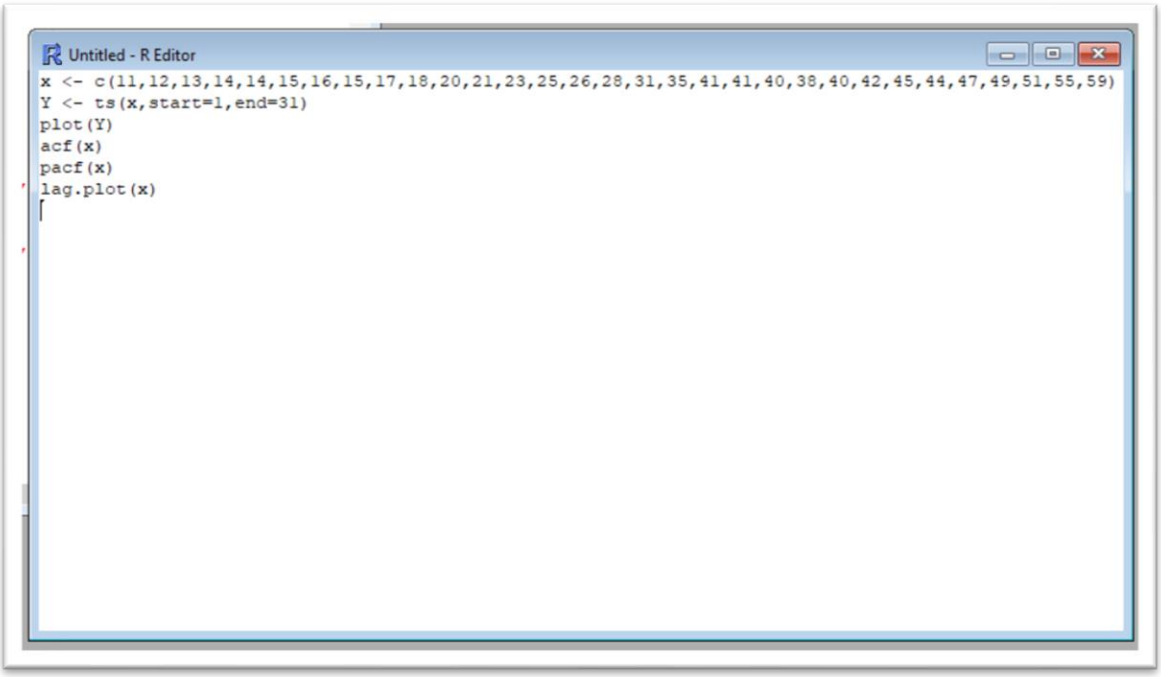


۵. **میتوانیم نمودار پراکندگی تاخیرها را نیز برای تشخیص خودهمبستگی از طریق کدهای زیر رسم کنیم.**

Lag.plot(x)



کدهای اجرا شده در R :



"طرح آزمایش در R"

آنالیز واریانس یک راهه

داده های زیر مربوط به تغذیه ی دامهای ضعیف یک گوداری میباشد برای افزایش وزن آنها تغذیه آنها را کنترل کردیم و به ۵ مدل تغذیه دست یافتیم.میخواهیم ببینیم آیا این 5 روش بپنتایج یکسانی داشته اند یا خیر.اگر نتیجه حاصل از این چهار روش باهم یکسان باشد تنها یکی از این روش هارا پیش میبریم اگر نتایج حاصل از ۴ روش باهم یکسان نباشند میخواهیم ببینیم کدام روش نتیجه متفاوتی در افزایش وزن دام ها داشته است.بنابراین ما 5 روش متفاوت داریم و در هر روش 4 مشاهده داریم.حال داده هارا وارد نرم افزار میکنیم و سپس جدول انالیز واریانس را رسم میکنیم.این سطوح تنها سطوح ممکن برای عامل مورد نظر بودند بنابراین طرح را با اثر ثابت در نظر میگیریم.

$$\begin{cases} H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 \\ H_1: \mu_i \neq \mu_j \qquad\qquad i \neq j \end{cases}$$

```
resp <- c(21,22,21,23,26,27,29,28,24,23,21,22,20,21,23,24,29,31,32,33)
fact <- rep(1:5,each=4)
fact <- factor(fact)
summary(aov(resp~fact))
```

خروجی:

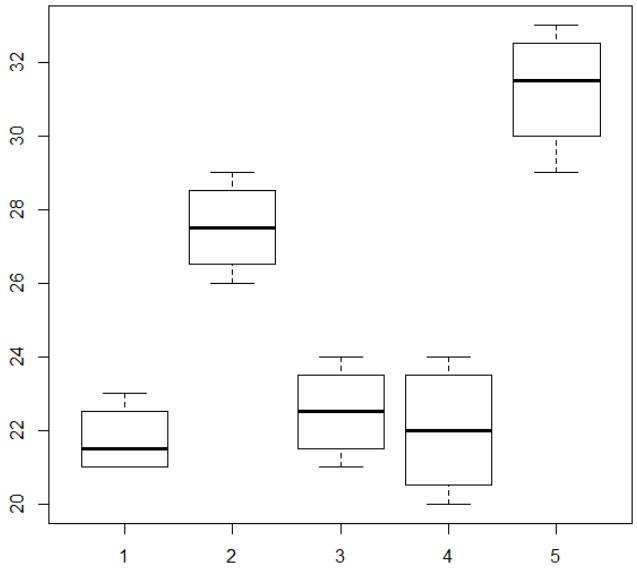
```
            Df Sum Sq Mean Sq F value   Pr(>F)
fact         4  284.5   71.12   33.87 2.39e-07 ***
Residuals   15   31.5    2.10
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

تفسیر: همانطور که در خروجی فوق جدول آنالیز واریانس را میبینیم در ستون df درجات ازادی محاسبه شده است در ستون بعدی مقدار مجموع مربعات برای خطای بین گروهی (factor)و برای خطای درون گروهی (residuals)محاسبه شده است که مقدار مجموع مربعات بین گروهی سهم خیلی بیشتری نسبت به خطای درون گروهی دارد به این معناست که میان میانگین سطوح مختف عامل تفاوت وجود دارد برای اطمینان از این فرض آزمون F را انجام داده و نتیجه در ستون F نوشته شده است که برابر ۳۳/۸۷ است که عدد بزرگی است و میتوان فرض برابری میانگین ۵ گروه را رد کرد و حداقل میانگین یکی از گروه ها با سایر گروه ها تفاوت معنا داری دارد.

رسم نمودار جعبه ای برای هر ۵ سطح با استفاده از دستور زیر

```
boxplot(resp~fact)
```

خروجی:



تفسیر:همانطور که در نمودار فوق میبینیم برای هر کدام از سطوح (گروه ها) یک نمودار جعبه ای رسم شده است و وزن دام های موجود در گروه ۲ و ۵ با سایر گروه ها تفاوت چشمگیری دارد بنابراین میتوان گفت که روش ۲ به خصوص روش ۵ در افزایش وزن دام ها تاثیر زیاده داشته اند اما وزن دام های حاصل از روش ۱ و ۳ و ۴ تقریبا باهم برابرند.

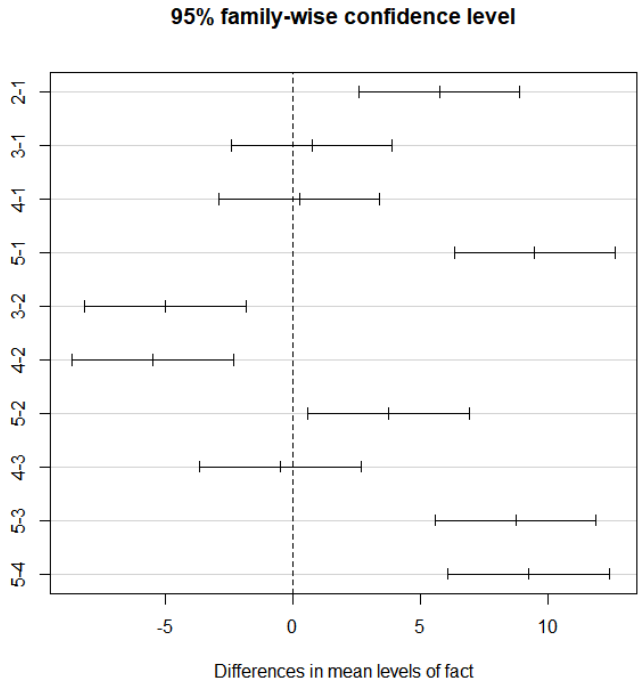
حال لازم است با استفاده از دستورات زیر مقادیر برازش داده شده (fitted) و باقیمانده ها (resid) را محاسبه کنیم.

```
mod <- aov(resp~fact)
mod[['resid']]
```

انجام آزمون توکی در R و رسم نمودار برای مقایسه ی میانگین های جفتی.

```
TukeyHSD(aov(resp~fact))

plot(TukeyHSD(aov(resp~fact)))
```

```
Fit: aov(formula = resp ~ fact)

$fact
     diff       lwr       upr     p adj
2-1  5.75  2.585819  8.914181 0.0004088
3-1  0.75 -2.414181  3.914181 0.9455907
4-1  0.25 -2.914181  3.414181 0.9991215
5-1  9.50  6.335819 12.664181 0.0000012
3-2 -5.00 -8.164181 -1.835819 0.0015991
4-2 -5.50 -8.664181 -2.335819 0.0006401
5-2  3.75  0.585819  6.914181 0.0168374
4-3 -0.50 -3.664181  2.664181 0.9873207
5-3  8.75  5.585819 11.914181 0.0000033
5-4  9.25  6.085819 12.414181 0.0000017
```



"رگرسیون در R"

مدل خطی ساده :

میخواهیم ببینیم سن نوجوانان (۱۳تا۱۸ سال) در میزان استفاده ی آنها از برنامه ی اینستاگرام در طول روز آیا ارتباطی دارد یا خیر.بنابراین متغیر وابسته دقایق استفاده از اینستاگرام در طول شبانه روز است و متغیر مستقل سن نوجوانان است.تعداد ۱۵ مشاهده داریم.

به دست آوردن مقادیر بتا صفر و بتا یک:

x<-c(۱۵’۱۴’۱۳’۱۶’۱۷’۱۸’۱۴’۱۳’۱۵’۱۷’۱۸’۱۶’۱۷’۱۴’۱۳)

y<-c(۲۱۰’۷۰’۷۵’۱۹۰’۲۰۰’۶۵’۸۰’۲۵’۹۵’۱۶۵’۸۰’۱۵۰’۱۰۰’۴۵’۲۰)

lm(y~x)

خروجی:

```
> x<-c(15,14,13,16,17,18,14,13,15,17,18,16,17,14,13)
> y<-c(210,70,75,190,200,65,80,25,95,165,80,150,100,45,20)
> lm(y~x)

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
    -147.88         16.47
```

مقدار بتا صفر برابر ۱۴۷٫۸۸- است و مقدار بتا یک برابر ۱۶٫۴۷ است.

محاسبه ی مقادیر مینیمم ، ماکسیمم و چارک ها برای باقی مانده ها :

همچنین محاسبه ی ضرایب استاندارد شده و آزمون صفر بودن آنها :

```
result <- summary(lm(y~x))
result
```

خروجی:

```
> result <- summary(lm(y~x))
> result

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-83.588 -39.471  -4.176  33.618 110.824

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -147.882    131.700  -1.123   0.2818
x             16.471      8.534   1.930   0.0757 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 57.46 on 13 degrees of freedom
Multiple R-squared:  0.2227,    Adjusted R-squared:  0.1629
F-statistic: 3.724 on 1 and 13 DF,  p-value: 0.07573
```

با توجه به مقدار pr که برای بتا صفر برابر ۰/۲۸ است میتوان فرض صفر بودن بتا صفر را پذیرفت.همچنین در سطح ۰/۱ میتوان فرض صفر بودن سن نوجوانان را رد کرد.

جدول آنالیز واریانس

ابتدا مدل را تشکیل داده و سپس جدول آنالیز واراینس را رسم میکنیم.

```
y <- -147.882+131.7*x+rnorm(15)
m1 <- lm(y~x);
anova(m1)
```

خروجی:

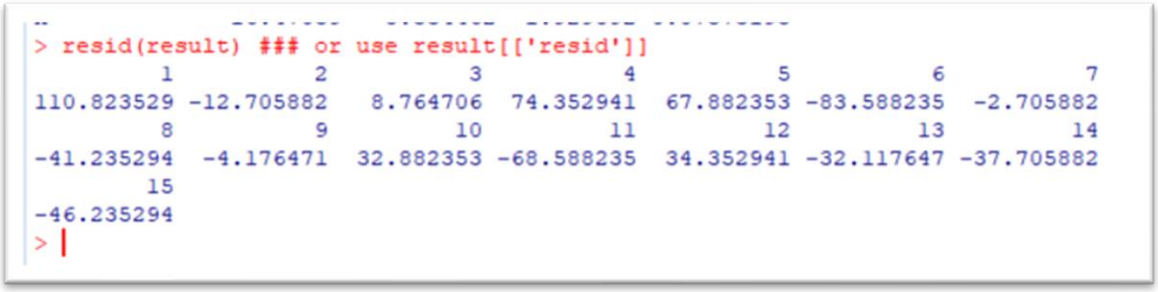
```
> m1 <- lm(y~x);
> anova(m1)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x          1 785872  785872  964502 < 2.2e-16 ***
Residuals 13     11       1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

همانطور که در جدول فوق میبینیم با توجه به مقدار F و Pr میبینیم که مدل معنا دار است و سهم بسیار زیاد از تغییرات کل را مدل تبیین کرده است و سهم کمی شامل خطا شده است بنابراین مدل ما مدل مناسبی است و قدرت تبیین خوبی دارد.

از طریق کد های زیر باقی مانده هارا محاسبه میکنیم.

```
resid(result) ### or use result[['resid']]
```

خروجی:

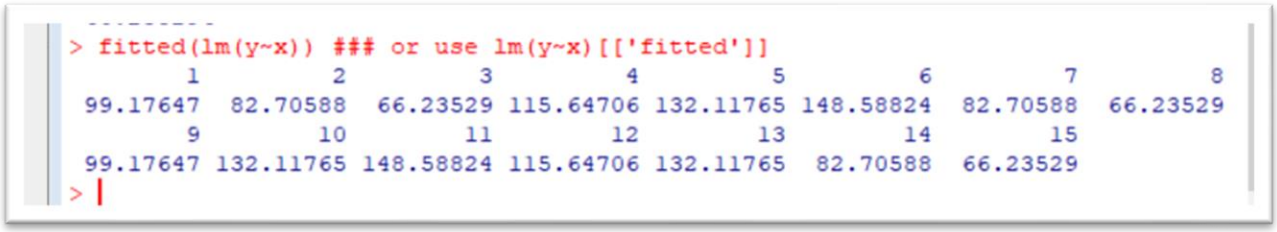


```
> resid(result) ### or use result[['resid']]
        1         2         3         4         5         6         7
110.823529 -12.705882   8.764706  74.352941  67.882353 -83.588235  -2.705882
        8         9        10        11        12        13        14
-41.235294  -4.176471  32.882353 -68.588235  34.352941 -32.117647 -37.705882
       15
-46.235294
>
```

محاسبه ی مقادیر برازش داده شده از طریق کد های زیر

```
fitted(lm(y~x)) ### or use lm(y~x)[['fitted']]
```

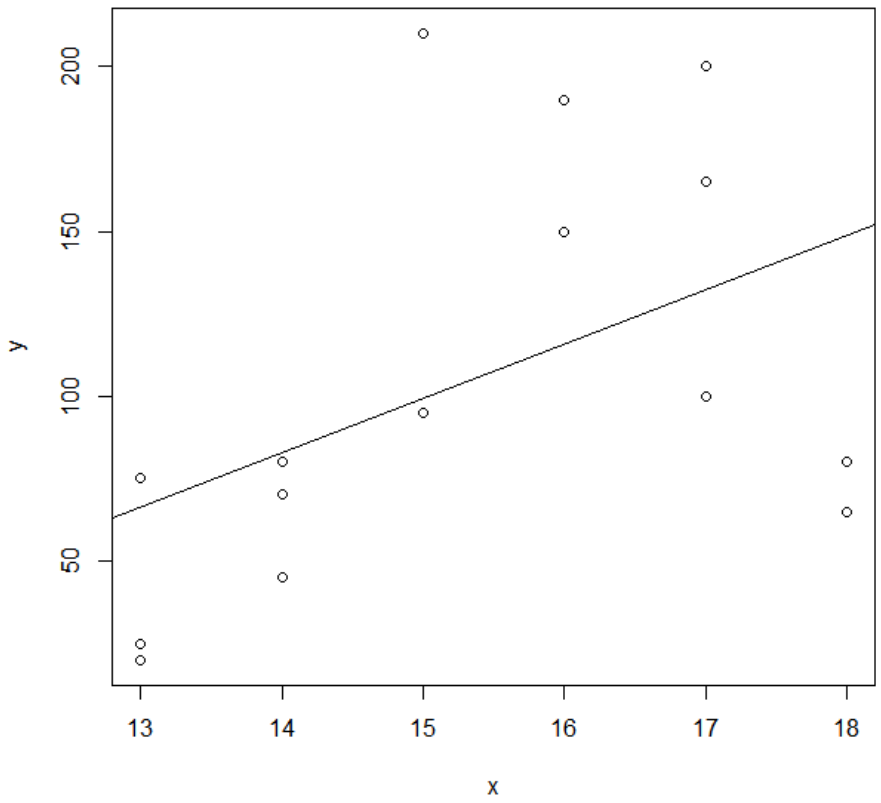
خروجی:



```
> fitted(lm(y~x)) ### or use lm(y~x)[['fitted']]
       1        2        3        4        5        6        7        8
 99.17647  82.70588  66.23529 115.64706 132.11765 148.58824  82.70588  66.23529
       9       10       11       12       13       14       15
 99.17647 132.11765 148.58824 115.64706 132.11765  82.70588  66.23529
>
```

رسم نمودار پراکنش

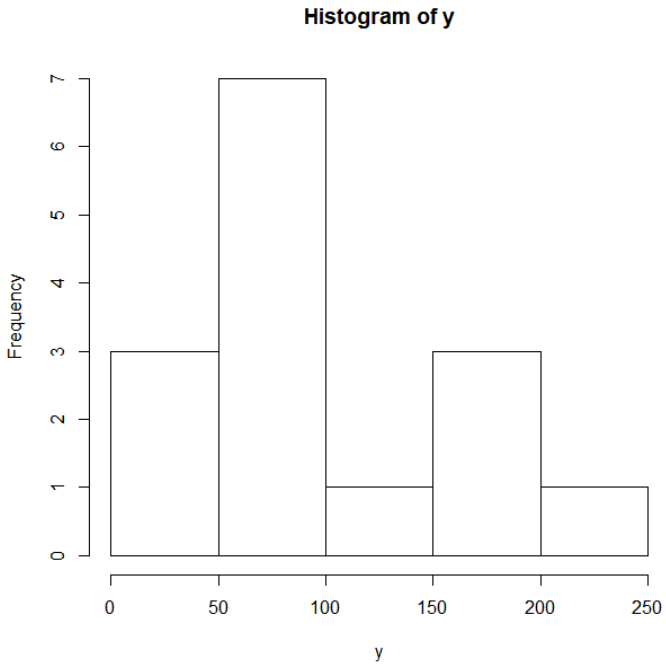
```
plot(x,y)
abline(lm(y~x))
```



همانطور که میبینیم یک رابطه خطی صعودی بین دو متغیر برقرار است.

رسم نمودار هیستوگرام

Hist(y)



همانطور که در نمودار فوق میبینیم بیشترین فراوانی مربوط به ۵۰تا ۱۰۰ دقیقه میباشد.

آزمون شاپیرویلک برای بررسی نرمالیتی

shapiro.test(y)

```
        Shapiro-Wilk normality test

data:  y
W = 0.9113, p-value = 0.142
```

با توجه به مقدار p-value که بزرگتر از ۰/۰۵ است فرض نرمال بودن را برای متغیر پاسخ میپذیریم.

"آزمون فرض در R"

آزمون T-test :

۱. **آزمون تی تک نمونه‌ای** :(One sample t test) **آزمون در مورد برابری میانگین جامعه یا مقدار ثابت و معلوم مثل** a.

یک کارخانه میخواهد بداند به طور میانگین وزن محصولات تولیدی اش چقد است یک نمونه ی ۱۶ تایی از محصولات را انتخاب نمودیم.مدیر کارخانه میگوید وزن نرمال ۲۷ کیلو گرم میباشد بنابراین میخواهیم آزمون کنیم که آیا میانگین وزن محصولات تولیدی ۲۷ هست یا خیر.

$$\begin{cases} H_0: \mu = 27 \\ H_1: \mu \neq 27 \end{cases}$$

با استفاده از کد های زیر مشاهدات را وارد نرم افزار کرده و آزمون را انجام میدهیم.

```
x<-(25,33,41,19,22,35,18,25,36,21,27,31,39,24,28,36)

x

t.test(x, mu = 27)
```

خروجی:

```
        One Sample t-test

data:  x
t = 1.3439, df = 30, p-value = 0.189
alternative hypothesis: true mean is not equal to 27
95 percent confidence interval:
 25.17299 35.85927
sample estimates:
mean of x
 30.51613
```

تفسیر: با توجه به مقدار p-value **که برابر ۰.۱۸۹ است و بزرگتر از ۰.۰۵ است میتوانیم فرض صفر را بپذیریم و میانگین ۲۷ برای محصولات کارخانه مناسب است یک فاصله اطمینان ۹۵٪ نیز برای محصولات محاسبه کرده است که بازه ی ۰.۲۵۱۷و ۳۵.۸۵ میباشد.**

۲. **آزمون تی دو جامعه مستقل** (Two independent sample t test): **آزمون در مورد برابری میانگین دو جامعه مستقل.**

میخواهیم تعداد تماس های گرفته شده با دو مطب مختلف را باهم مقایسه کنیم و ببینیم آیا تعداد تماس های گرفته شده با این دو مطب به طور میانگین باهم برابر است یا خیر.تعداد تماس ها برای مطب اول و مطب دوم را وارد نرم فزار میکنیم. ازمون تی دو جامعه مستقل را انجام میدهیم.

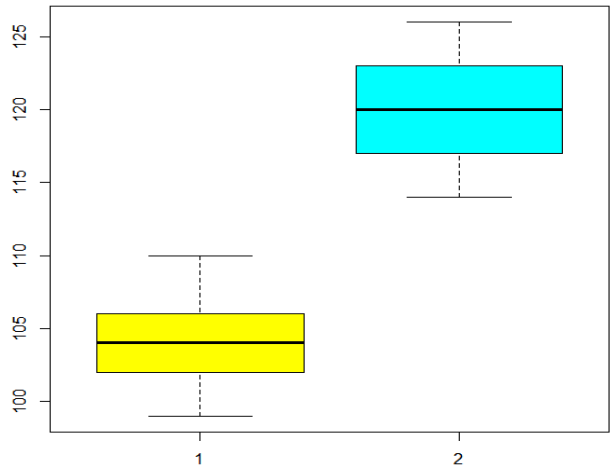
$$\begin{cases} H_0: \mu_A = \mu_B \\ H_1: \mu_A \neq \mu_B \end{cases}$$

matab1 <- c (۱۰۰,۱۰۵,۱۰۴,۹۹,۱۰۶,۱۰۴,۱۰۲,۱۰۸,۱۱۰,۱۰۳)

matab2 <- c (۱۱۴,۱۲۶,۱۱۷,۱۲۴,۱۱۹,۱۲۱,۱۱۵,۱۲۱,۱۱۹,۱۲۳)

boxplot (matab1, matab2, col = c(31,5), names=c("1","2"))

خروجی:



تفسیر: همانطور که میبینیم ارتفاع دو نمودار تقریبا باهم برابر است بنابراین میتوانیم فرض برابری واریانس را برای دو جامعه در نظر بگیریم.

انجام آزمون برای اطمینان از برابری واریانس دو گروه:

var.test(matab1,matab2)

$$\begin{cases} H_0: \sigma_A/\sigma_B = 1 \\ H_1: \sigma_A/\sigma_B \neq 1 \end{cases}$$

خروجی:

```
        F test to compare two variances

data:  matab1 and matab2
F = 0.7628, num df = 9, denom df = 9, p-value = 0.6932
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1894656 3.0709770
sample estimates:
ratio of variances
         0.7627872
```

تفسیر: با توجه به مقدار p-value **که برابر ۰.۶۹۳۲ و از ۰.۰۵ بزرگتر است میتوان گفت که فرض صفر یعنی برابری واریانس دو گروه را میپذیریم.**

انجام آزمون برابری میانگین دو جامعه:

t.test(matab2,matab2,paired = FALSE,var.equal = TRUE ,alternative = "less")

خروجی:

```
        Two Sample t-test

data:  matab2 and matab2
t = 0, df = 18, p-value = 0.5
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf 3.002374
sample estimates:
mean of x mean of y
    119.9     119.9
```

تفسیر: با توجه به مقدار p-value **که بزرگتر از ۰.۰۵ است فرض صفر یعنی برابری میانگین دو جامعه را میپذیریم.**

۳. **آزمون تی برای زوج متغیرها** :(Paired sample t test) **آزمون برابر میانگین دو متغیر از یک جامعه.**

میخواهیم بررسی کنیم که میانگین سطح اکسیژن خون افراد قبل از بیماری کرونا و پس از بهبودی تغییر کرده است یا خیر.بنابراین لازم است فرض برابری میانگین در دو جامعه ی وابسته را بررسی کنیم.

$$\begin{cases} H_0: \mu_A = \mu_B \\ H_1: \mu_A \neq \mu_B \end{cases}$$

ghabl=c(۹۷,۹۸,۹۷,۹۴,۹۶,۹۲,۹۷,۹۴,۹۶,۹۸,۹۹,۹۹,۱۰۰,۱۰۱,۹۸)

behbudi=c(۹۹,۹۹,۱۰۰,۱۰۱,۹۹,۹۸,۹۶,۹۷,۹۷,۹۶,۹۸,۹۹,۱۰۱,۹۷,۹۸)

بررسی فرض برابر واریانس دو گروه

var.test(bimar,behbudi)

خروجی:

```
        F test to compare two variances

data:  bimar and behbudi
F = 0.5358, num df = 14, denom df = 14, p-value = 0.2553
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1799004 1.5960733
sample estimates:
ratio of variances
         0.5358491
```

تفسیر: با توجه به مقدار p-vlue که بیشتر ۰/۰۵ است میتوان فرض برابری واریانس را برای دو جامعه در نظر گرفت.

بررسی فرض برابر میانگین های دو جامعه :

```
t.test(bimar,behbudi,paired = TRUE,var.equal = TRUE,alternative="greater")
```

خروجی:

```
        Paired t-test

data:  bimar and behbudi
t = -15.405, df = 14, p-value = 1
alternative hypothesis: true difference in means is greater than 0.01
95 percent confidence interval:
 -9.584415       Inf
sample estimates:
mean of the differences
                  -8.6

>
```

تفسیر: با توجه به مقدار p-value میبینیم که بیشتر از ۰/۰۵ است بنابراین میانگین سطح قند خون قبل از بیماری کرونا و پس بهبودی باهم برابر است.

"آزمون های ناپرامتری در R"

۱. آزمون شاپیرو ویلک برای بررسی فرض نرمال بودن.

با استفاده از دستورات زیر ابتدا ۲۰ داده ی نرمال تولید میکنیم و سپس آزمون شاپیوویلک را برای داده ها اجرا میکنیم و انتظار داریم نتیجه نرمال بدن داده ها باشد.

```
x <- rnorm(۲۰)

x

shapiro.test(x)
```

خروجی:

```
> x <- rnorm(20)
> x
 [1] -0.45601491 -1.40794886  1.29780550  1.31153098  2.25127964  0.43089830
 [7]  0.42787441  0.66874840 -0.84206940 -1.72623210  0.07486034  0.92130788
[13]  0.46167746 -0.90950043  1.43103307  0.55953919 -1.79328208  0.53448239
[19]  1.81280637  1.13701402
> shapiro.test(x)

        Shapiro-Wilk normality test

data:  x
W = 0.9478, p-value = 0.3346

>
```

تفسیر: همانطور که در تصویر بالا میبینیم مقدار p-value بیشتر از ۰/۰۵ است بنابراین فرض نرمال بودن را برای داده های تولید شده میپذیریم.

۲. آزمون من ویتنی

با استفاده از دستورات زیر متغیر هارا معرفی کرده و آزمون من ویتنی را اجرا میکنیم

x <- c(۱۰’۵’۸’۹’۱۱’۱۲’۱۴’۹’۱۰’۸)

y <- c(۷’۸’۵’۳’۴’۶’۴’۵’۶’۷)

wilcox.test(x,y,exact=F)

خروجی:

```
> x <-c(10,5,8,9,11,12,14,9,10,8)
> y <-c(7,8,5,3,4,6,4,5,6,7)
> wilcox.test(x,y,exact=F)

        Wilcoxon rank sum test with continuity correction

data:  x and y
W = 93, p-value = 0.001244
alternative hypothesis: true location shift is not equal to 0
```