

پروژه نهایی درس استنباط آماری

پوریا آزادی مقدم

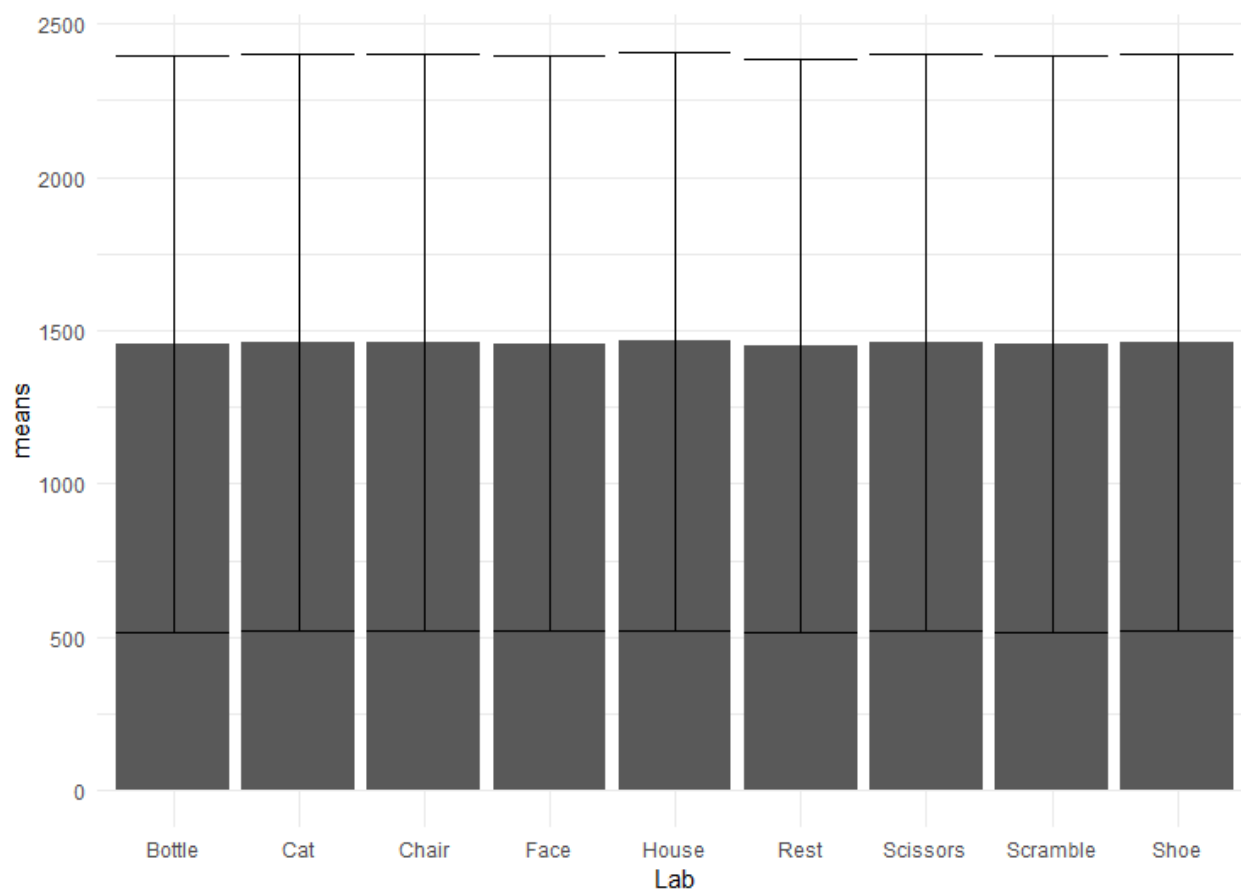
810193331

Q1:

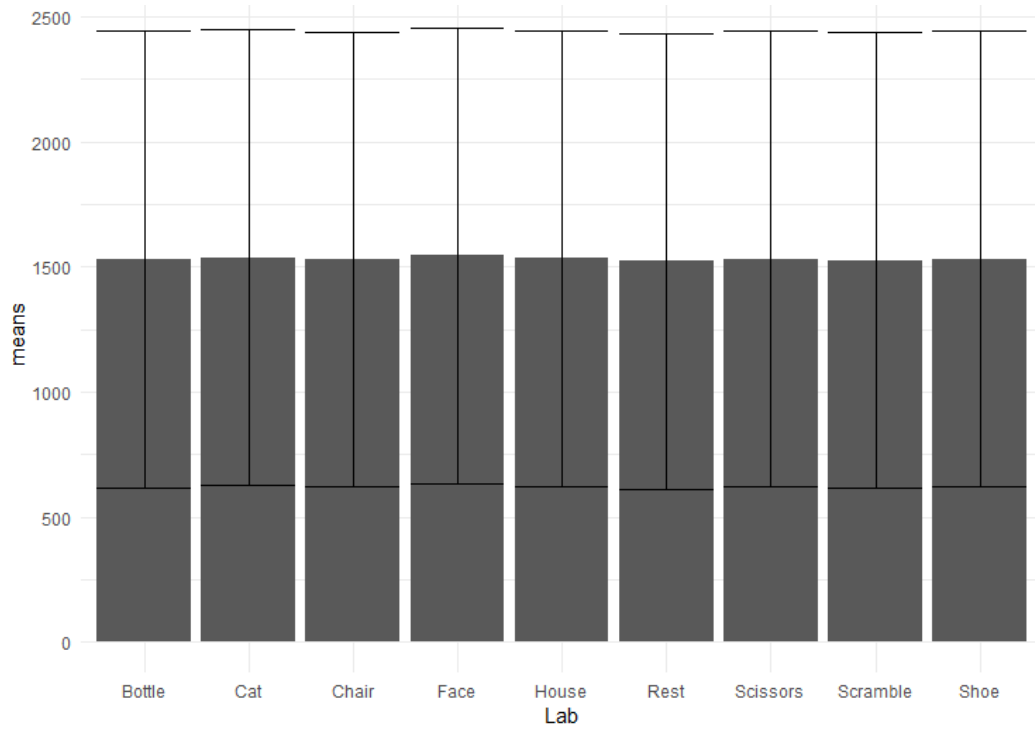
بخش الف:

در ادامه شکل های BarPlot مربوط به میانگین پاسخ ها در ناحیه های متفاوت و Standard Deviation مربوط به آنها در حالتی که الف یا همان سطح معناداری را برابر با 0.05 در نظر گرفته ایم, آمده است.

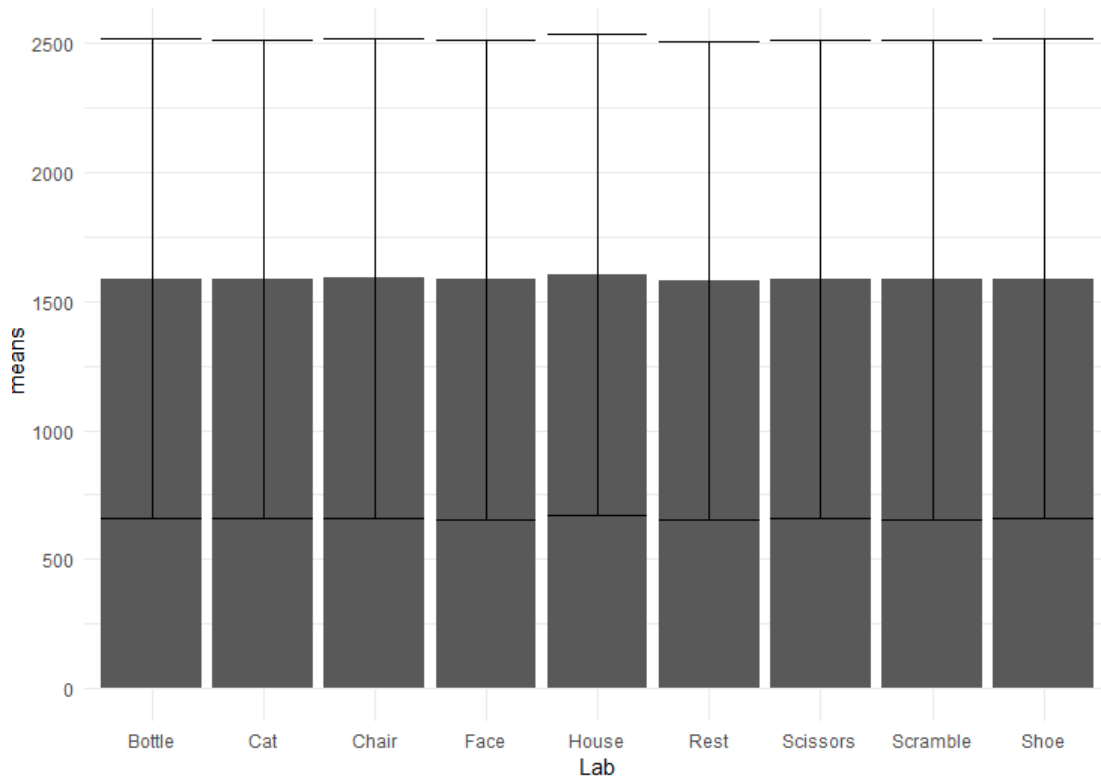
Area#1 or Temporal Region:



Area#2 or Face Region:



Area#3 or House Region:



بخش ب:

برای انجام این تست آماری از روش های Pairwise_t_Test استفاده خواهیم نمود. توجه شود که این تست ها را با سطح معناداری 0.05 درصد , در هر 3 ناحیه Temporal, Face, House و بین تمامی 8 نوع تصاویر متفاوت و Rest انجام خواهیم داد.

در اینجا از روش anova نمی توان استفاده نمود چرا که بین نوع های مختلف تصاویر استقلال وجود ندارد اما در شروط ANOVA داریم که باید این شرط برقرار باشد, پس نمی توان از این تست برای این مرحله استفاده نمود. از طرف دیگر این داده ها شرایط تست t_tset را اغنا می نمایند داد ها تعداد زیادی دارند که می توان با حد مرکزی به شرایط مناسب جهت t_test رسید.

```
[1] "Bottle"
[1] "in Area1 for Rest and "
[1] "Bottle"

Paired t-test

data: as.numeric(area1s[[i, 1]]) and as.numeric(area1s[[6, 1]])
t = 32.693, df = 2792, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 6.763462 7.626513
sample estimates:
mean of the differences
      7.194987

[1] "in Area2 for Rest and "
[1] "Bottle"

Paired t-test

data: as.numeric(area2s[[i, 1]]) and as.numeric(area2s[[6, 1]])
t = 15.161, df = 300, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 6.938577 9.008532
sample estimates:
mean of the differences
      7.973555

[1] "in Area3 for Rest and "
[1] "Bottle"

Paired t-test

data: as.numeric(area3s[[i, 1]]) and as.numeric(area3s[[9, 1]])
t = -11.043, df = 1033, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.500001 -1.745608
sample estimates:
mean of the differences
      -2.122805
```

Area#1 =Temporal Region
Area#2 =Face Region
Area#3 =House Region

```
[1] "Cat"
[1] "in Area1 for Rest and "
[1] "Cat"
```

Paired t-test

```
data: as.numeric(areals[[i, 1]]) and as.numeric(areals[[6, 1]])
t = 37.181, df = 2792, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 9.126915 10.143153
sample estimates:
mean of the differences
 9.635034
```

Area#1 =Tempral Region
Area#2 =Face Region
Area#3 =House Region

```
[1] "in Area2 for Rest and "
[1] "Cat"
```

Paired t-test

```
data: as.numeric(area2s[[i, 1]]) and as.numeric(area2s[[6, 1]])
t = 23.49, df = 300, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
14.88627 17.60861
sample estimates:
mean of the differences
16.24744
```

```
[1] "in Area3 for Rest and "
[1] "Cat"
```

Paired t-test

```
data: as.numeric(area3s[[i, 1]]) and as.numeric(area3s[[9, 1]])
t = -4.7594, df = 1033, p-value = 2.218e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-1.3396295 -0.5574691
sample estimates:
mean of the differences
-0.9485493
```

```
[1] "Chair"  
[1] "in Area1 for Rest and "  
[1] "Chair"
```

Paired t-test

```
data: as.numeric(area1s[[i, 1]]) and as.numeric(area1s[[6, 1]])  
t = 39.205, df = 2792, p-value < 2.2e-16  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
 8.33085 9.20804  
sample estimates:  
mean of the differences  
      8.769445
```

Area#1 =Tempral Region
Area#2 =Face Region
Area#3 =House Region

```
[1] "in Area2 for Rest and "  
[1] "Chair"
```

Paired t-test

```
data: as.numeric(area2s[[i, 1]]) and as.numeric(area2s[[6, 1]])  
t = 13.564, df = 300, p-value < 2.2e-16  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
 6.137858 8.221079  
sample estimates:  
mean of the differences  
      7.179468
```

```
[1] "in Area3 for Rest and "  
[1] "Chair"
```

Paired t-test

```
data: as.numeric(area3s[[i, 1]]) and as.numeric(area3s[[9, 1]])  
t = 4.0832, df = 1033, p-value = 4.785e-05  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
 0.5095458 1.4523885  
sample estimates:  
mean of the differences  
      0.9809671
```

```
[1] "Face"  
[1] "in Area1 for Rest and "  
[1] "Face"
```

Paired t-test

```
data: as.numeric(area1s[[i, 1]]) and as.numeric(area1s[[6, 1]])  
t = 23.753, df = 2792, p-value < 2.2e-16  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
 6.553451 7.732764  
sample estimates:  
mean of the differences  
      7.143108
```

Area#1 =Tempral Region
Area#2 =Face Region
Area#3 =House Region

```
[1] "in Area2 for Rest and "  
[1] "Face"
```

Paired t-test

```
data: as.numeric(area2s[[i, 1]]) and as.numeric(area2s[[6, 1]])  
t = 28.401, df = 300, p-value < 2.2e-16  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
 22.39392 25.72834  
sample estimates:  
mean of the differences  
      24.06113
```

```
[1] "in Area3 for Rest and "  
[1] "Face"
```

Paired t-test

```
data: as.numeric(area3s[[i, 1]]) and as.numeric(area3s[[9, 1]])  
t = -15.579, df = 1033, p-value < 2.2e-16  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
 -5.597186 -4.344902  
sample estimates:  
mean of the differences  
      -4.971044
```

```
[1] "House"
```

```
[1] "in Area1 for Rest and "
```

```
[1] "House"
```

Paired t-test

```
data: as.numeric(area1s[[i, 1]]) and as.numeric(area1s[[6, 1]])
t = 50.742, df = 2792, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 13.92873 15.04850
sample estimates:
mean of the differences
      14.48862
```

Area#1 =Tempral Region

Area#2 =Face Region

Area#3 =House Region

```
[1] "in Area2 for Rest and "
```

```
[1] "House"
```

Paired t-test

```
data: as.numeric(area2s[[i, 1]]) and as.numeric(area2s[[6, 1]])
t = 16.623, df = 300, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
   9.827195 12.466493
sample estimates:
mean of the differences
      11.14684
```

```
[1] "in Area3 for Rest and "
```

```
[1] "House"
```

Paired t-test

```
data: as.numeric(area3s[[i, 1]]) and as.numeric(area3s[[9, 1]])
t = 41.33, df = 1033, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 14.38197 15.81569
sample estimates:
mean of the differences
      15.09883
```



```
[1] "Scissors"
[1] "in Area1 for Rest and "
[1] "Scissors"
```

Paired t-test

```
data: as.numeric(area1s[[i, 1]]) and as.numeric(area1s[[6, 1]])
t = 37.634, df = 2792, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 8.530571 9.468348
sample estimates:
mean of the differences
      8.999459
```

Area#1 =Tempral Region
Area#2 =Face Region
Area#3 =House Region

```
[1] "in Area2 for Rest and "
[1] "Scissors"
```

Paired t-test

```
data: as.numeric(area2s[[i, 1]]) and as.numeric(area2s[[6, 1]])
t = 18.716, df = 300, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 9.358365 11.557582
sample estimates:
mean of the differences
      10.45797
```

```
[1] "in Area3 for Rest and "
[1] "Scissors"
```

Paired t-test

```
data: as.numeric(area3s[[i, 1]]) and as.numeric(area3s[[9, 1]])
t = -6.8516, df = 1033, p-value = 1.253e-11
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-1.7046115 -0.9456012
sample estimates:
mean of the differences
      -1.325106
```

```

[1] "Scramble"
[1] "in Area1 for Rest and "
[1] "Scramble"

Paired t-test

data: as.numeric(areals[[i, 1]]) and as.numeric(areals[[6, 1]])
t = 20.436, df = 2792, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 3.345152 4.055221
sample estimates:
mean of the differences
      3.700186

[1] "in Area2 for Rest and "
[1] "Scramble"

Paired t-test

data: as.numeric(area2s[[i, 1]]) and as.numeric(area2s[[6, 1]])
t = 7.3849, df = 300, p-value = 1.517e-12
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 2.455966 4.240379
sample estimates:
mean of the differences
      3.348173

[1] "in Area3 for Rest and "
[1] "Scramble"

Paired t-test

data: as.numeric(area3s[[i, 1]]) and as.numeric(area3s[[9, 1]])
t = -12.764, df = 1033, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.982165 -2.920891
sample estimates:
mean of the differences
     -3.451528

```

Area#1 =Tempral Region
Area#2 =Face Region
Area#3 =House Region

```
[1] "shoe"  
[1] "in Area1 for Rest and "  
[1] "Shoe"
```

Paired t-test

```
data: as.numeric(areals[[i, 1]]) and as.numeric(areals[[6, 1]])  
t = 36.471, df = 2792, p-value < 2.2e-16  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
 8.496672 9.462211  
sample estimates:  
mean of the differences  
      8.979441
```

Area#1 =Tempral Region
Area#2 =Face Region
Area#3 =House Region

```
[1] "in Area2 for Rest and "  
[1] "Shoe"
```

Paired t-test

```
data: as.numeric(area2s[[i, 1]]) and as.numeric(area2s[[6, 1]])  
t = 17.037, df = 300, p-value < 2.2e-16  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
 8.726294 11.005401  
sample estimates:  
mean of the differences  
      9.865847
```

```
[1] "in Area3 for Rest and "  
[1] "Shoe"
```

Paired t-test

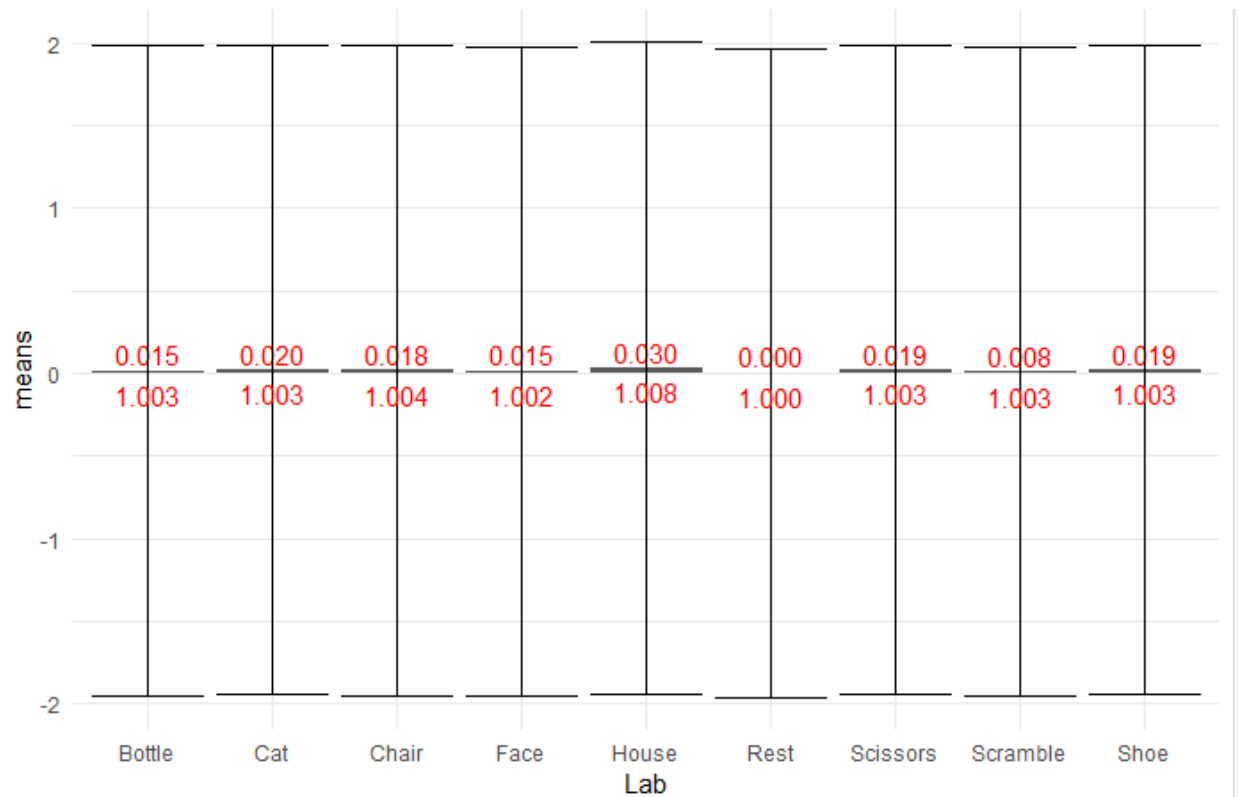
```
data: as.numeric(area3s[[i, 1]]) and as.numeric(area3s[[9, 1]])  
t = NaN, df = 1033, p-value = NA  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
 NaN NaN  
sample estimates:  
mean of the differences  
      -0.401000
```

بخش ج:

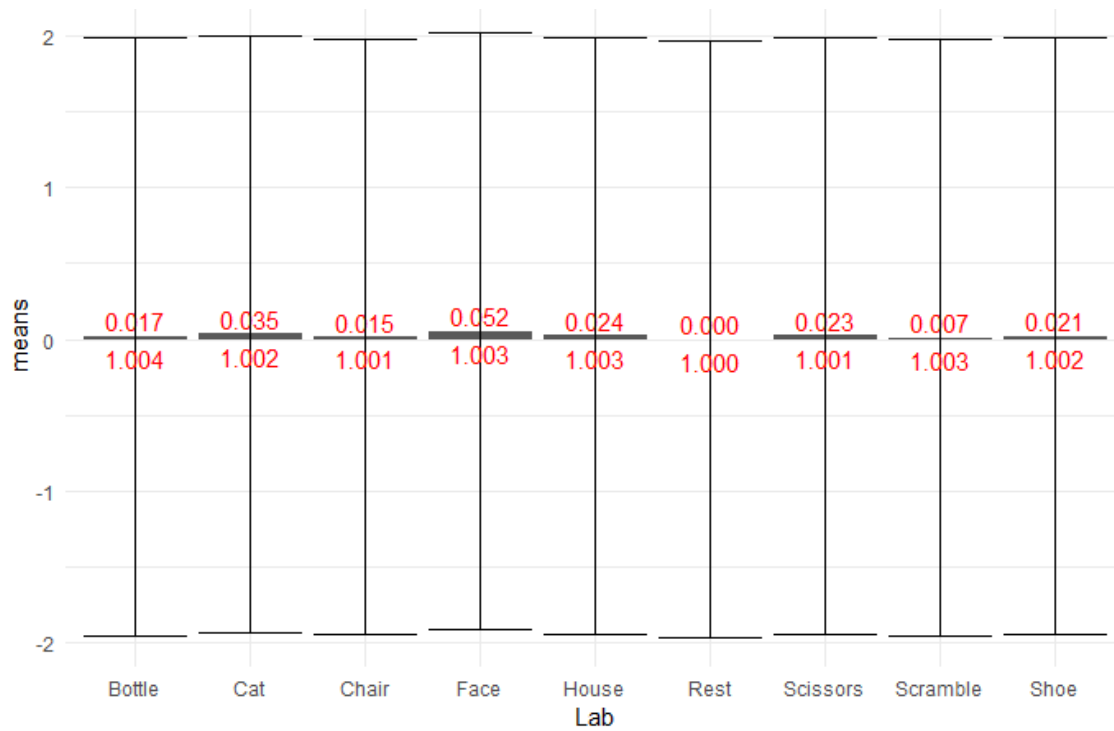
در این بخش به وسیله داده های Rest نرمالایز نمودیم, این عملیات را برای بخش الف یعنی بر روی میانگین داده ها و Standard Deviation انجام دادیم.

توجه شود که این نرمال سازی مکم خواهد کرد تا بایاس احتمالی بر روی داده ها از بین برود, در نتیجه این عمل نتیجه مشاهده شده و مقایسه انجام شده معتبر تر خواهد بود.

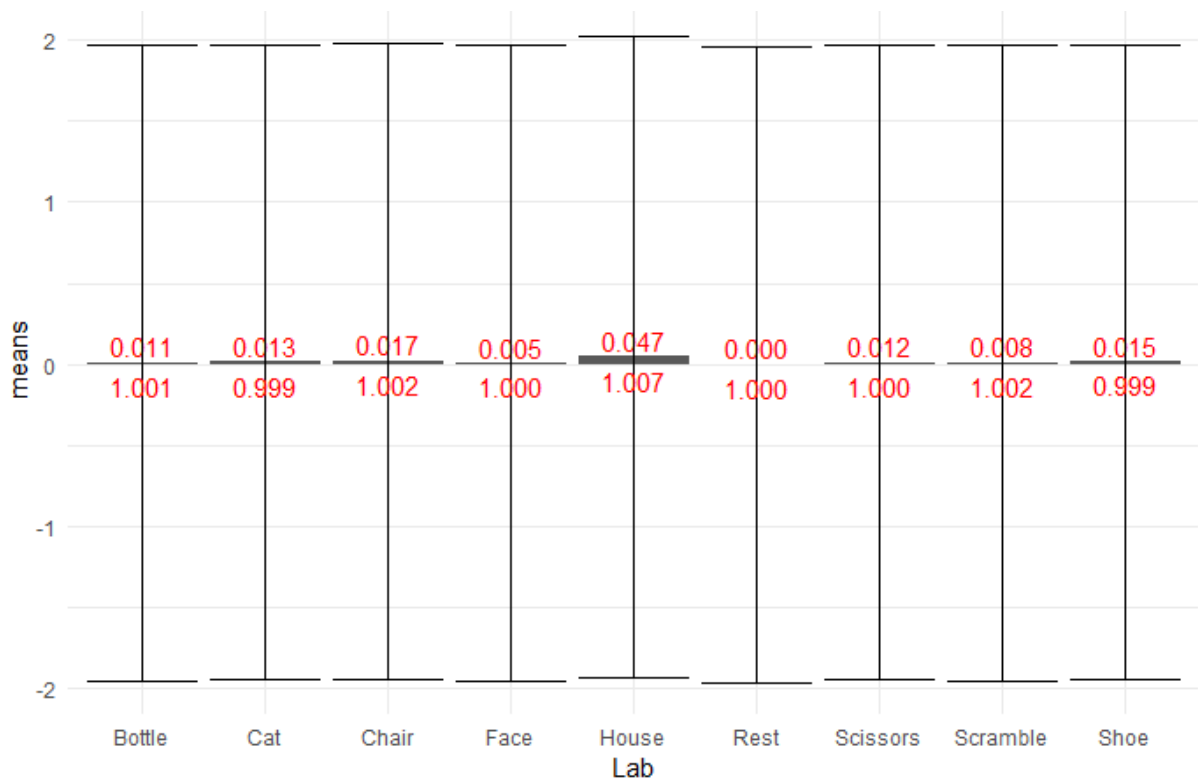
Area#1 or Temporal Region:



Area#2 or Face Region:



Area#3 or Face Region:



بخش د:

در این بخش، تست های T_test و Signed Ranked Test را به صورت paired انجام خواهیم داد.

تست T_test در زمانی عملکرد بهتری از خود نشان می دهد که اگر تعداد نمونه ها کم است، نمونه ها از توزیعی نرمال بدست آمده باشند یا به قدر کافی جهت استفاده از قضیه حد مرکزی، نمونه ها بزرگ باشند و تقارن نیز در آن ها وجود داشته باشد. در این دو حالت است که میتوان از T_test استفاده نمود و به نتایج حاصل از آن اعتماد کرد.

ویژگی Signed Rank Test آن است که با توجه به خود داده ها تست را انجام می دهد و اثر داده های پرت را نیز کم می کند. در حقیقت این نوع تست به توزیع نمونه ها نگاه نمی کند اما چولگی را مد نظر قرار می دهد.

T_TEST:

Person #5, Temporal Region or Area#1

```
> t.test(as.numeric(data_set[[5,1]][7,]), as.numeric(data_set[[5,1]][1,]), paired = TRUE)
```

Paired t-test

data: as.numeric(data_set[[5, 1]][7,]) and as.numeric(data_set[[5, 1]][1,])

t = 13.265, df = 421, p-value < 2.2e-16

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

4.667868 6.291942

sample estimates:

mean of the differences

5.479905

Person #5, Face Region or Area#2

```
> t.test(as.numeric(data_set[[5,2]][7,]), as.numeric(data_set[[5,2]][1,]), paired = TRUE)
```

Paired t-test

data: as.numeric(data_set[[5, 2]][7,]) and as.numeric(data_set[[5, 2]][1,])

t = 6.4933, df = 66, p-value = 1.276e-08

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

4.704473 8.882095

sample estimates:

mean of the differences

6.793284

Person #5, House Region or Area#3

```
> t.test(as.numeric(data_set[[5,3]][7,]), as.numeric(data_set[[5,3]][1,]), paired = TRUE)
```

Paired t-test

data: as.numeric(data_set[[5, 3]][7,]) and as.numeric(data_set[[5, 3]][1,])

t = 5.3621, df = 95, p-value = 5.765e-07

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

2.762754 6.011205

sample estimates:

mean of the differences

4.386979

Singed Renk Test:

Person #5, Temporal Region or Area#1

```
> wilcox.test(as.numeric(data_set[[5,1]][7,]), as.numeric(data_set[[5,1]][1,]), paired = TRUE, alternative = "two.sided")

wilcoxon signed rank test with continuity correction

data: as.numeric(data_set[[5, 1]][7, ]) and as.numeric(data_set[[5, 1]][1, ])
V = 73960, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0

> wilcox.test(as.numeric(data_set[[5,2]][7,]), as.numeric(data_set[[5,2]][1,]), paired = TRUE, alternative = "two.sided")

wilcoxon signed rank test with continuity correction

data: as.numeric(data_set[[5, 2]][7, ]) and as.numeric(data_set[[5, 2]][1, ])
V = 1998, p-value = 8.192e-08
alternative hypothesis: true location shift is not equal to 0

> wilcox.test(as.numeric(data_set[[5,3]][7,]), as.numeric(data_set[[5,3]][1,]), paired = TRUE, alternative = "two.sided")

wilcoxon signed rank test with continuity correction

data: as.numeric(data_set[[5, 3]][7, ]) and as.numeric(data_set[[5, 3]][1, ])
V = 3556.5, p-value = 2.175e-06
alternative hypothesis: true location shift is not equal to 0
```

Person #5, Face Region or Area#2

Person #5, House Region or Area#3

از آن جایی که داده های مربوط به این بخش دارای چولگی می باشد و همچنین داده های پرت نیز را شامل خواهد شد پس در نتیجه t_test عملکرد به نسبت ضعیف تری نسبت به signed rank دارد و rank مناسب تر است. هرچند که هر دو تست نشان داده اند که اختلاف معناداری را نشان دادند.

بخش ه:

در این قسمت, با استفاده از یکی از تست paired پارامتری و همچنین یکی از تست های paired ناپارامتری تمایز میانگین فعالیت وکسل ها در حالتی که فرد چهره مشاهده می کند با حالتی که فرد در حالت استراحت است را بررسی نمودیم, به همین منظور تست های `t_test` و `Wilcoxn` انتخاب شدند.

هر دو تست فرض در این مرحله رد شدند و اختلاف معناداری مشاهده شد:

T_test:

Person #3, Temporal Region or Area#1

```
> #T-Test in all areas for 3rd person for rest and  
> t.test(as.numeric(data_set[[3,1]][3,]), as.numeric(data_set[[3,1]][6,]), paired = TRUE) -----
```

Paired t-test

```
data: as.numeric(data_set[[3, 1]][3, ]) and as.numeric(data_set[[3, 1]][6, ])  
t = 7.1542, df = 306, p-value = 6.249e-12  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
 1.968985 3.463067  
sample estimates:  
mean of the differences  
      2.716026
```

Person #3, Face Region or Area#2

```
> t.test(as.numeric(data_set[[3,2]][3,]), as.numeric(data_set[[3,2]][6,]), paired = TRUE)
```

Paired t-test

```
data: as.numeric(data_set[[3, 2]][3, ]) and as.numeric(data_set[[3, 2]][6, ])  
t = 1.9418, df = 10, p-value = 0.08084  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
 -0.4127013  6.0108831  
sample estimates:  
mean of the differences  
      2.799091
```

Person #3, House Region or Area#3

```
> t.test(as.numeric(data_set[[3,3]][3,]), as.numeric(data_set[[3,3]][6,]), paired = TRUE)
```

Paired t-test

```
data: as.numeric(data_set[[3, 3]][3, ]) and as.numeric(data_set[[3, 3]][6, ])  
t = 3.4152, df = 125, p-value = 0.0008601  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
 0.7034298 2.6422845  
sample estimates:  
mean of the differences  
      1.672857
```

```
>
```


Wilcoxon:

Person #3, Temporal Region or Area#1

```
> #wilcox_test in all areas for 3rd person for rest and face
> wilcox.test(as.numeric(data_set[[3,1]][3,]), as.numeric(data_set[[3,1]][6,]), paired = TRUE, alternative = "two.sided")

wilcoxon signed rank test with continuity correction
```

```
data: as.numeric(data_set[[3, 1]][3, ]) and as.numeric(data_set[[3, 1]][6, ])
V = 34696, p-value = 4.602e-13
alternative hypothesis: true location shift is not equal to 0
```

```
> wilcox.test(as.numeric(data_set[[3,2]][3,]), as.numeric(data_set[[3,2]][6,]), paired = TRUE, alternative = "two.sided")
```

wilcoxon signed rank test

```
data: as.numeric(data_set[[3, 2]][3, ]) and as.numeric(data_set[[3, 2]][6, ])
V = 50, p-value = 0.1475
alternative hypothesis: true location shift is not equal to 0
```

```
> wilcox.test(as.numeric(data_set[[3,3]][3,]), as.numeric(data_set[[3,3]][6,]), paired = TRUE, alternative = "two.sided")
```

wilcoxon signed rank test with continuity correction

```
data: as.numeric(data_set[[3, 3]][3, ]) and as.numeric(data_set[[3, 3]][6, ])
V = 5351, p-value = 0.0004985
alternative hypothesis: true location shift is not equal to 0
```

بخش و:

در این بخش میزان ضریب همبستگی بین پاسخ به تصاویر صورت و پاسخ به تصاویر خانه را در نواحی حساس به صورت، با روش Pearson و Spearman در شرکت کننده اول بررسی شد.

تست pearson تستی است که به صورت پارامتری انجام میشود. این تست میتواند رابطه خطی بین دو متغیر را نشان بدهد. از طرف دیگر تست Spearman تستی بدون پارامتر می باشد. این تست میزان غیر یکنواختی بین دو متغیر را بررسی خواهد کرد.

توجه شود که در حالت نرمال، این دو تست عملکرد نزدیک به هم و حتی تغییر خطی را نشان می دهند. با توجه به این دو تست، رابطه مورد بررسی در این سوال خطی می باشد:

نکته: توجه شود که منظور از یکنواختی آن است که دو متغیر با هم تغییر نمایند اما رابطه آن ها لزوما خطی نباشد.

Perrson#1:

```
> #paerson and pearmen test for person1 for House and face
> cor.test(as.numeric(data_set[[1,1]][5,]), as.numeric(data_set[[1,1]][4,]), method = "pearson")

Pearson's product-moment correlation

data: as.numeric(data_set[[1, 1]][5, ]) and as.numeric(data_set[[1, 1]][4, ])
t = 431.8, df = 575, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9981884 0.9986936
sample estimates:
      cor
0.9984616

> cor.test(as.numeric(data_set[[1,1]][5,]), as.numeric(data_set[[1,1]][4,]), method = "spearman")

Spearman's rank correlation rho

data: as.numeric(data_set[[1, 1]][5, ]) and as.numeric(data_set[[1, 1]][4, ])
S = 67188, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.9979015
```

Person#2:

```
> #paerson and pearmen test for person2 for House and face
> cor.test(as.numeric(data_set[[1,2]][5,]), as.numeric(data_set[[1,2]][4,]), method = "pearson")

Pearson's product-moment correlation

data: as.numeric(data_set[[1, 2]][5, ]) and as.numeric(data_set[[1, 2]][4, ])
t = 102.86, df = 19, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9977427 0.9996440
sample estimates:
      cor
0.9991033

> cor.test(as.numeric(data_set[[1,2]][5,]), as.numeric(data_set[[1,2]][4,]), method = "spearman")

Spearman's rank correlation rho

data: as.numeric(data_set[[1, 2]][5, ]) and as.numeric(data_set[[1, 2]][4, ])
S = 8, p-value = 4.158e-06
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.9948052
```

Person#3:

```
> #paerson and pearmen test for person3 for House and face
> cor.test(as.numeric(data_set[[1,3]][5,]), as.numeric(data_set[[1,3]][4,]), method = "pearson")

Pearson's product-moment correlation

data: as.numeric(data_set[[1, 3]][5, ]) and as.numeric(data_set[[1, 3]][4, ])
t = 284.33, df = 223, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9982098 0.9989419
sample estimates:
      cor
0.9986236

> cor.test(as.numeric(data_set[[1,3]][5,]), as.numeric(data_set[[1,3]][4,]), method = "spearman")

Spearman's rank correlation rho

data: as.numeric(data_set[[1, 3]][5, ]) and as.numeric(data_set[[1, 3]][4, ])
S = 3792.5, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.9980023
```

بخش ز:

در این بخش مقدرا p_value در مرحله قبل را به کمک permutation test (از مجموعه روش‌های resampling) به دست آوردیم. برای این منظور با شافل کردن، جفت بودن داده‌ها را بر هم زدیم و هر بار همبستگی حساب نمودیم. به این ترتیب توزیع فرض صفر به دست می‌آید و p_value را به کمک محاسبه احتمال مشاهده واقعی (ضریب همبستگی بدون شافل کردن) در توزیع صفر به دست آمد.

توجه شود که این permutation با می شود تست pearson عملکرد متعادل ترو نرمال تری را از خود نشان بدهد، پس مشکل حساسیت نداشتن به چوگس و داده‌های پرت برای آن بر طرف خواهد شد.

```
for (sample_number in 1:1000) {
  #sampling and calculating correlation
  r_sam[sample_number] = cov(sample(as.numeric(data_set[[1,2]][4,]), 21), sample(as.numeric(data_set[[1,2]][5,]), 21)) / (sd(sample(as.numeric(data_set[[1,2]][5,]), 21)) *
                                                                    sd(sample(as.numeric(data_set[[1,2]][4,]), 21)))
}
#finding the distribution
cdf = ecdf(as.numeric(r_sam))
#the p-value is:
p_value = cdf(cov(as.numeric(data_set[[1,2]][4,]), as.numeric(data_set[[1,2]][5,])) / (sd(as.numeric(data_set[[1,2]][4,])) * sd(as.numeric(data_set[[1,2]][5,]))))
p_value
] 1
```

p_value
] 1

بخش ح:

تفاوت میانگین پاسخ در ناحیه تمپورال به دسته‌های مختلف تصاویر را در ماتریسی که مولفه زام آن تفاوت میانگین پاسخ به تصاویر دسته ام از میانگین پاسخ به تصاویر دسته زام است، نمایش داده ایم که در ادامه قابل مشاهده می باشد:

Difference Matrix:

```
> print(diff_matrix)
      [,1]      [,2]      [,3]      [,4]      [,5]      [,6]      [,7]      [,8]      [,9]
[1,] 0      -2.440047 -1.574458 0.0518797 -7.293631 7.194987 -1.804472 3.494801 -1.784454
[2,] 2.440047 0      0.865589 2.491926 -4.853584 9.635034 0.6355747 5.934848 0.6555926
[3,] 1.574458 -0.865589 0      1.626337 -5.719173 8.769445 -0.2300143 5.069259 -0.2099964
[4,] -0.0518797 -2.491926 -1.626337 0      -7.34551 7.143108 -1.856352 3.442922 -1.836334
[5,] 7.293631 4.853584 5.719173 7.34551 0      14.48862 5.489159 10.78843 5.509177
[6,] -7.194987 -9.635034 -8.769445 -7.143108 -14.48862 0      -8.999459 -3.700186 -8.979441
[7,] 1.804472 -0.6355747 0.2300143 1.856352 -5.489159 8.999459 0      5.299273 0.0200179
[8,] -3.494801 -5.934848 -5.069259 -3.442922 -10.78843 3.700186 -5.299273 0      -5.279255
[9,] 1.784454 -0.6555926 0.2099964 1.836334 -5.509177 8.979441 -0.0200179 5.279255 0
```

حال بر روی هر مولفه این ماتریس تست T_TEST می زنیم و سپس آن را به یک لیست 81 درایه ای به روش Bonferroni و روش FDR تبدیل می کنیم:

Bonferroni:

```
> print(p_value_1_adjust_boffernt)
[1] 0.000000e+00 1.575662e-07 7.118262e-06 1.163720e-24 1.324238e-32 7.958319e-71 1.042489e-09 5.694809e-33 1.481301e-12 1.575662e-07 0.000000e+00 1.000000e+00 7.193106e-57 1.055666e-16
[15] 3.053981e-78 1.000000e+00 7.850331e-38 1.000000e+00 7.118262e-06 1.000000e+00 0.000000e+00 1.255590e-24 6.035569e-27 4.932338e-84 1.000000e+00 1.350326e-29 1.000000e+00 1.163720e-24
[29] 7.193106e-57 1.255590e-24 0.000000e+00 1.624427e-53 2.252560e-14 2.620051e-37 1.000000e+00 2.855775e-40 1.324238e-32 1.055666e-16 6.035569e-27 1.624427e-53 0.000000e+00 1.934521e-108
[43] 7.750495e-20 1.483286e-69 3.510929e-20 7.958319e-71 3.053981e-78 4.932338e-84 2.252560e-14 1.934521e-108 0.000000e+00 8.917702e-77 1.093299e-32 1.654210e-69 1.042489e-09 1.000000e+00
[57] 1.000000e+00 2.620051e-37 7.750495e-20 8.917702e-77 0.000000e+00 3.973245e-54 1.000000e+00 5.694809e-33 7.850331e-38 1.350326e-29 1.000000e+00 1.483286e-69 1.093299e-32 3.973245e-54
[71] 0.000000e+00 1.639759e-45 1.481301e-12 1.000000e+00 1.000000e+00 2.855775e-40 3.510929e-20 1.654210e-69 1.000000e+00 1.639759e-45 0.000000e+00
> print(p_value_4_adjust_boffernt)
[1] 0.000000e+00 4.518625e-16 1.000000e+00 1.000000e+00 7.780076e-50 3.207485e-23 1.372724e-01 1.680754e-34 5.736799e-13 4.518625e-16 0.000000e+00 2.848602e-13 5.800755e-10 9.277225e-26 6.625635e-42
[16] 3.114817e-08 1.659950e-58 2.951735e-01 1.000000e+00 2.848602e-13 0.000000e+00 1.000000e+00 1.465299e-54 8.683941e-27 3.556965e-02 9.973172e-26 1.124128e-08 1.000000e+00 5.800755e-10 1.000000e+00
[31] 0.000000e+00 1.278797e-24 2.593260e-16 1.000000e+00 2.167015e-15 5.839366e-02 7.780076e-50 9.277225e-26 1.465299e-54 1.278797e-24 0.000000e+00 3.714039e-74 2.465973e-35 5.947925e-91 3.874005e-41
[46] 3.207485e-23 6.625635e-42 8.683941e-27 2.593260e-16 3.714039e-74 0.000000e+00 1.496507e-34 2.679217e-01 5.185260e-37 1.372724e-01 3.114817e-08 3.556965e-02 1.000000e+00 2.465973e-35 1.496507e-34
[61] 0.000000e+00 4.299158e-39 3.669727e-03 1.680754e-34 1.659950e-58 9.973172e-26 2.167015e-15 5.947925e-91 2.679217e-01 4.299158e-39 0.000000e+00 1.484363e-60 5.736799e-13 2.951735e-01 1.124128e-08
[76] 5.839366e-02 3.874005e-41 5.185260e-37 3.669727e-03 1.484363e-60 0.000000e+00
```

```
> print(p1_final)
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
[1,] 1    1    1    1    1    1    1    1    1
[2,] 1    1    0    1    1    1    0    1    0
[3,] 1    0    1    1    1    1    0    1    0
[4,] 1    1    1    1    1    1    1    0    1
[5,] 1    1    1    1    1    1    1    1    1
[6,] 1    1    1    1    1    1    1    1    1
[7,] 1    0    0    1    1    1    1    1    0
[8,] 1    1    1    0    1    1    1    1    1
[9,] 1    0    0    1    1    1    0    1    1
> print(p4_final)
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
[1,] 1    1    0    0    1    1    0    1    1
[2,] 1    1    1    1    1    1    1    1    0
[3,] 0    1    1    0    1    1    1    1    1
[4,] 0    1    0    1    1    1    0    1    0
[5,] 1    1    1    1    1    1    1    1    1
[6,] 1    1    1    1    1    1    1    0    1
[7,] 0    1    1    0    1    1    1    1    1
[8,] 1    1    1    1    1    0    1    1    1
[9,] 1    0    1    0    1    1    1    1    1
```

BH(FDR):

```
> print(p_value_1_adjust_fdr)
[1] 0.000000e+00 2.424096e-09 1.062427e-07 2.374939e-26 3.079622e-34 4.188589e-72 1.654744e-11 1.460207e-34 2.428362e-14 2.424096e-09 0.000000e+00 3.599188e-01 2.877242e-58 1.852046e-18
[15] 2.035987e-79 2.870321e-01 2.242952e-39 5.881140e-01 1.062427e-07 3.599188e-01 0.000000e+00 2.461942e-26 1.284164e-28 3.794106e-85 7.951141e-01 3.000725e-31 6.222869e-01 2.374939e-26
[29] 2.877242e-58 2.461942e-26 0.000000e+00 5.601472e-55 3.817898e-16 7.081218e-39 1.417426e-01 8.653864e-42 3.079622e-34 1.852046e-18 1.284164e-28 5.601472e-55 1.758655e-109
[43] 1.409181e-21 7.063268e-71 6.624394e-22 4.188589e-72 2.035987e-79 3.794106e-85 3.817898e-16 1.758655e-109 0.000000e+00 5.245707e-78 2.666582e-34 7.192219e-71 1.654744e-11 2.870321e-01
[57] 7.951141e-01 7.081218e-39 1.409181e-21 5.245707e-78 0.000000e+00 1.471572e-55 4.110064e-01 1.460207e-34 2.242952e-39 3.000725e-31 1.417426e-01 7.063268e-71 2.666582e-34 1.471572e-55
[71] 0.000000e+00 5.289544e-47 2.428362e-14 5.881140e-01 6.222869e-01 8.653864e-42 6.624394e-22 7.192219e-71 4.110064e-01 5.289544e-47 0.000000e+00
> print(p_value_4_adjust_fdr)
[1] 0.000000e+00 9.221685e-18 3.397329e-01 6.862648e-01 3.704798e-51 7.127744e-25 1.989455e-03 4.802155e-36 1.043054e-14 9.221685e-18 0.000000e+00 5.374721e-15 1.017676e-11 2.378776e-27 2.880711e-43
[16] 5.106257e-10 9.764414e-60 4.043473e-03 3.397329e-01 5.374721e-25 0.000000e+00 3.953353e-01 7.712099e-56 2.347011e-28 5.472254e-04 2.432481e-27 1.905302e-10 6.862648e-01 1.017676e-11 3.953353e-01
[31] 0.000000e+00 2.973948e-26 5.517575e-18 1.947668e-01 4.249050e-17 8.715472e-04 3.704798e-51 2.378776e-27 7.712099e-56 2.973948e-26 0.000000e+00 2.856953e-75 7.954753e-37 5.407205e-92 1.549602e-42
[46] 7.127744e-25 2.880711e-43 2.347011e-28 5.517575e-18 2.856953e-75 0.000000e+00 4.534870e-36 3.773545e-03 1.788021e-38 1.989455e-03 5.106257e-10 5.472254e-04 1.947668e-01 7.954753e-37 4.534870e-36
[61] 0.000000e+00 1.592281e-40 5.824964e-05 4.802155e-36 9.764414e-60 2.432481e-27 4.249050e-17 5.407205e-92 3.773545e-03 1.592281e-40 0.000000e+00 9.895753e-62 1.043054e-14 4.043473e-03 1.905302e-10
[76] 8.715472e-04 1.549602e-42 1.788021e-38 5.824964e-05 9.895753e-62 0.000000e+00
```

```
> print(p1_final_fdr)
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
[1,] 1    1    1    1    1    1    1    1    1
[2,] 1    1    0    1    1    1    0    1    0
[3,] 1    0    1    1    1    1    0    1    0
[4,] 1    1    1    1    1    1    1    0    1
[5,] 1    1    1    1    1    1    1    1    1
[6,] 1    1    1    1    1    1    1    1    1
[7,] 1    0    0    1    1    1    1    1    0
[8,] 1    1    1    0    1    1    1    1    1
[9,] 1    0    0    1    1    1    0    1    1
> print(p4_final_fdr)
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
[1,] 1    1    0    0    1    1    1    1    1
[2,] 1    1    1    1    1    1    1    1    1
[3,] 0    1    1    0    1    1    1    1    1
[4,] 0    1    0    1    1    1    0    1    1
[5,] 1    1    1    1    1    1    1    1    1
[6,] 1    1    1    1    1    1    1    1    1
[7,] 1    1    1    0    1    1    1    1    1
[8,] 1    1    1    1    1    1    1    1    1
[9,] 1    1    1    1    1    1    1    1    1
>
```

همان طور که در ماتریس های نهایی بالا می بینید روش FDR تعداد درایه های 1 بیشتری دارد که به معنای رد کرد فرض صفر در آن درایه های خاص می باشد که این مورد با بخش نظری نیز همخوانی دارد.

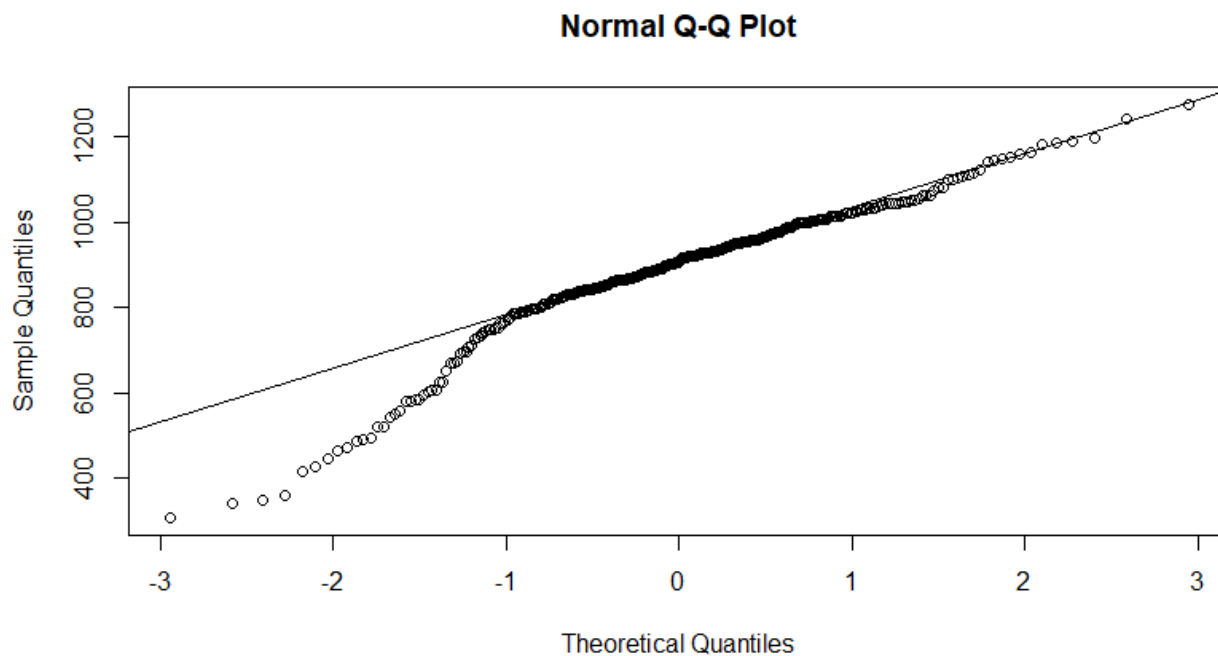
بخش ط:

در این قسمت توزیع داده مربوط به واکسل های ناحیه تمپورال و ناحیه حساس به خانه را با استفاده از تست KS و نمودار QQ-plot بررسی کردیم:

Peron#3, Temporal Region or Area#1 and Face Images:

```
One-sample kolmogorov-Smirnov test  
data: as.numeric(data_set[[3, 1]][4, ])  
D = 0.10567, p-value = 0.002107  
alternative hypothesis: two-sided
```

Person#3, Temporal Region or Area#1



Peron#3, Temporal Region or Area#1 and House Images:

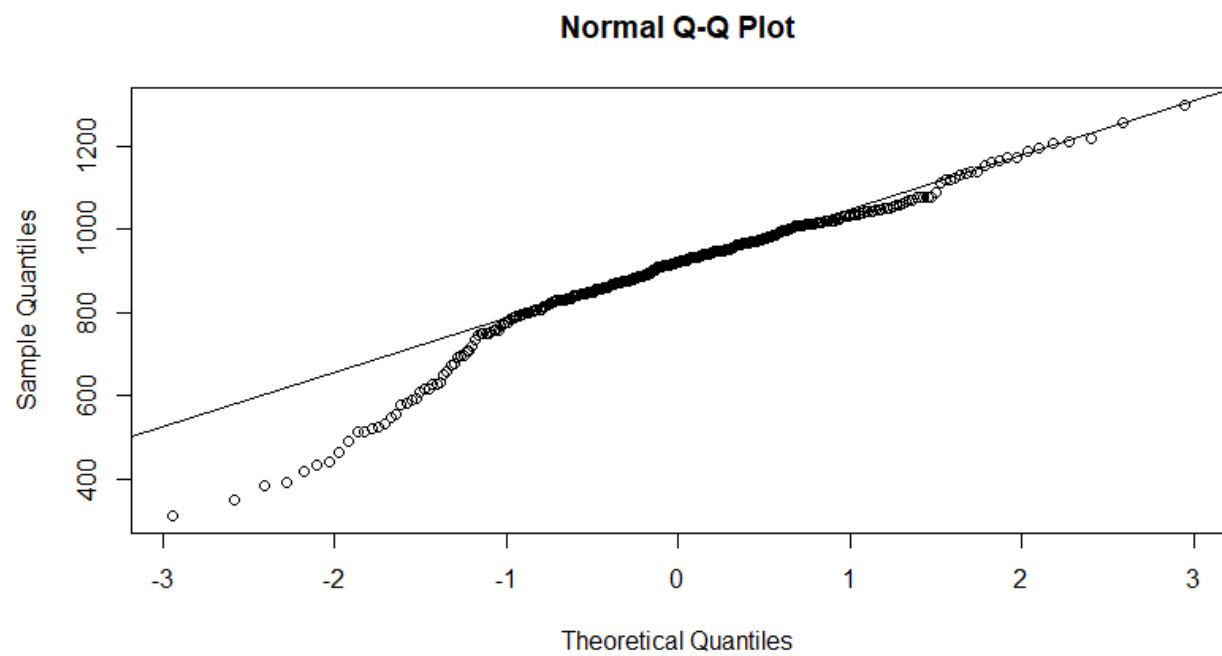
One-sample kolmogorov-smirnov test

data: as.numeric(data_set[[3, 1]][5,])

D = 0.098601, p-value = 0.005111

alternative hypothesis: two-sided

Person#3, Temporal Region or Area#1 and House Images

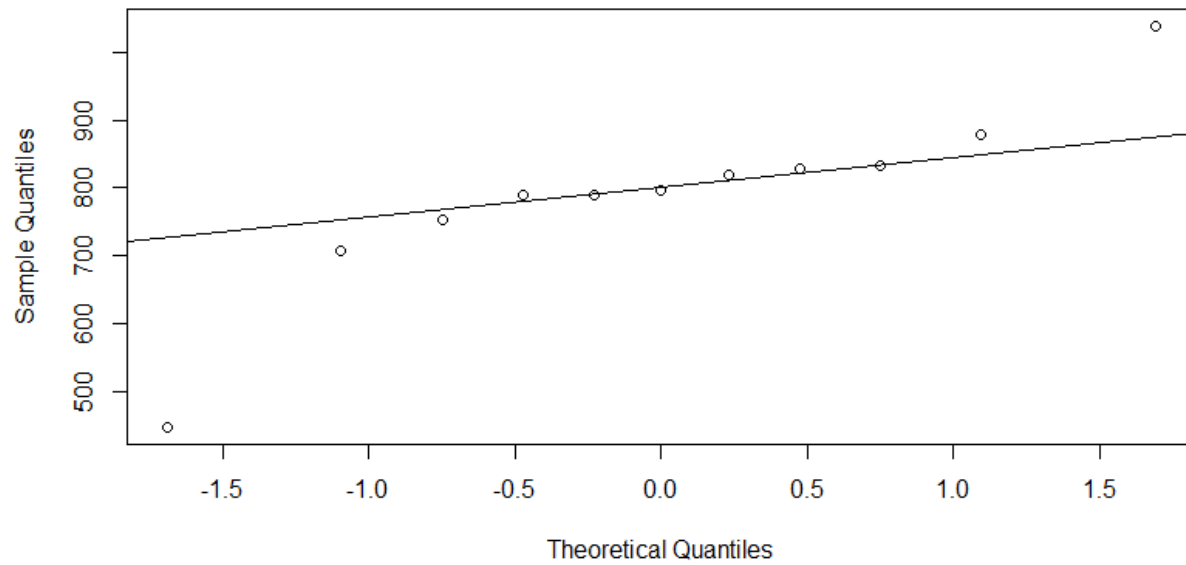


Peron#3, Face Region or Area#2 and Face Images:

Person#3, Face Region or Area#2 and Face Images

```
one-sample Kolmogorov-Smirnov test  
data: as.numeric(data_set[[3, 2]][4, ])  
D = 0.228, p-value = 0.5429  
alternative hypothesis: two-sided
```

Normal Q-Q Plot

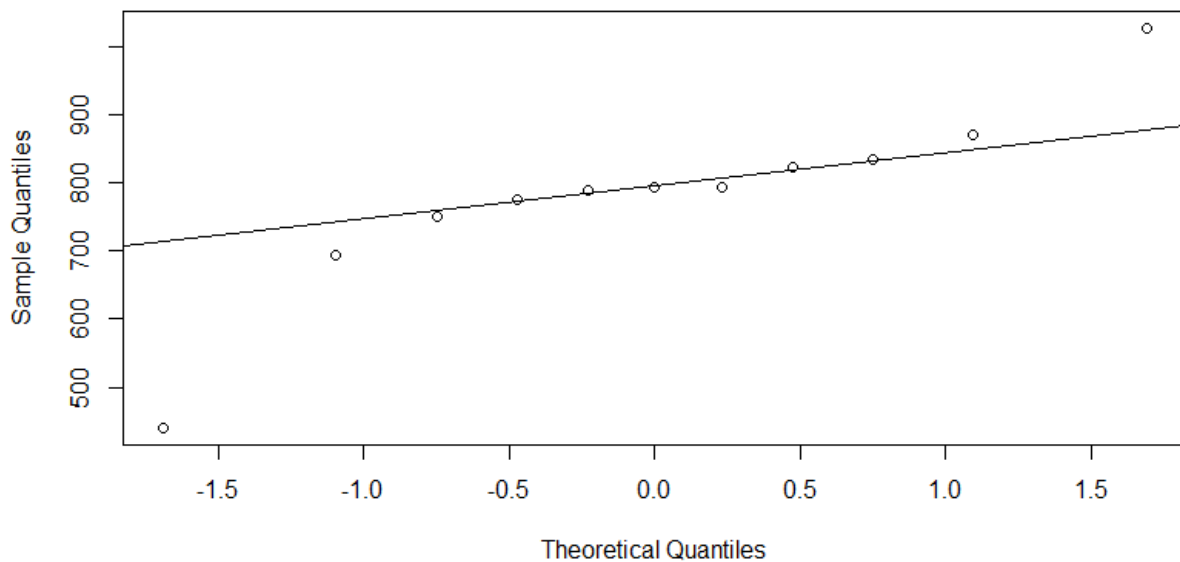


Peron#3, Face Region or Area#2 and House Images:

```
one-sample kolmogorov-smirnov test  
data: as.numeric(data_set[[3, 2]][5, ])  
D = 0.23268, p-value = 0.5172  
alternative hypothesis: two-sided
```

Person#3, Face Region or Area#2 and House Images

Normal Q-Q Plot



همان طور که می بینیم در ناحیه تمپورال مقدار p_value کم تر از 0.05 می باشد پس فرض نرمال بودن برای آن ها رد خواهد شد اما در ناحیه صورت مقدار p_value ها بزرگتر از 0.05 است که نتیجه می دهد که فرض نرمال بودن برای آن ها رد نمی تواند بشود و نرمال تر هستند.

بخش ی:

تست T_Test :

این تست زمانی مناسب است که یا تعداد نمونه ها به اندازه ی کافی بزرگ باشد که بتوان با توجه به قضیه حد مرکزی از آن برای مقایسه ی میانگین استفاده کرد و یا آن که توزیع نرمال باشد، در نتیجه توان تست بالا خواهد بود و استفاده از باعث رسیدن به نتیجه منطقی خواهد شد. اما اگر توزیع نمونه ها نرمال نباشد یا تعداد نمونه ها به اندازه ی کافی بزرگ نباشد در این صورت این تست پاسخگویی نیاز مورد نظر نیست و از توان بالایی برخوردار نیست .

Wilcoxon:

برای زمانی مناسب است که توزیع داده ها نرمال و یا تعریف آماره قابل استفاده از حد مرکزی نباشد. حساسیت به چولگی و داده پرت را کم میکند و توزیع داده ها برای آن مهم نمی باشد. توجه شود که اگر توزیع ها متقارن و تعداد نمونه ها زیاد باشد، هر دو تست wilcoxon و t_test قدرت برابر ای دارند اما اگر تعداد داده ها کم باشد نسبت به t_test توان کمی خواهد داشت.

تست کلموگروف-اسمیرنوف :

اگر چه این تست به توزیع داده ها حساسیت ندارد اما این تست به توزیع نمونه ها به طور کلی نگاه ای میاندازد ولی تاثیر پارامترهای توزیع را تقریباً در نظر نمیگیرد. توجه نمایید که تست زمانی میتواند تفاوت های ریز را تشخیص دهد که تعداد نمونه ها کم باشد یا توزیعها تفاوت واضحی از هم داشته باشند.

توجه شود که اگر دو تا داده ی نمونه از یک توزیع باشند به سختی میتواند مشابه بودن توزیع دو دسته داده- را رد کند. این به آن علت است که دیدن تفاوت کم برای این نوع تست ممکن نمیشود. به همین خاطر توان تست در برابر فرض صفر پایین است. این تست، از نوه غیر پارامتری می باشد و همچنین تست به تعداد نمونه وابسته نیست و عمومیت بالایی دارد.

Bootstrap :

این تست اگر تعداد نمونه ها کم باشد، توان و دقت پایینی خواهد داشت. فارغ از نوع توزیع نمونه ها بازه های اطمینان با روش ناپارامتری محاسبه میکند که تفاوت های ریز را نیز در نظر میگیرد. و به طور کلی روش مناسبی هم در روش های پارامتری و هم غیر پارامتری است.

```
> print(t.test(person_1_areal_area2, person_1_areal_rest, alternative = "two.sided", paired = TRUE))
```

Paired t-test

```
data: person_1_areal_area2 and person_1_areal_rest
t = -7.8651, df = 576, p-value = 1.829e-14
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.802386 -1.682432
sample estimates:
mean of the differences
      -2.242409
```

```
> print(wilcox.test(person_1_areal_area2, person_1_areal_rest, alternative = "two.sided", paired = TRUE))
```

Wilcoxon signed rank test with continuity correction

```
data: person_1_areal_area2 and person_1_areal_rest
V = 53030, p-value = 2.661e-13
alternative hypothesis: true location shift is not equal to 0
```

```
> print(ks.test(person_1_areal_area2, person_1_areal_rest, alternative = "two.sided"))
```

Two-sample Kolmogorov-Smirnov test

```
data: person_1_areal_area2 and person_1_areal_rest
D = 0.013865, p-value = 1
alternative hypothesis: two-sided
```

Question 2:

بخش الف:

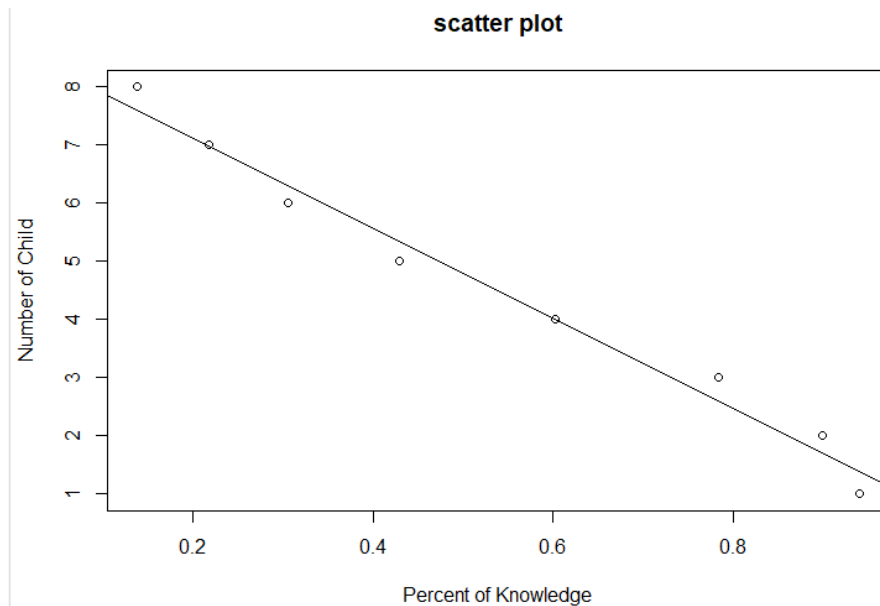
جدول بدست آمده برابر است با:

| نسبت تعداد دختران به پسران | سن مادر | | درصد با سواد | درص از کل مادران | تعداد مادر | تعداد فرزندان مادر |
|-------------------------------|---------|---------|--------------|---------------------|------------|--------------------------|
| | متوسط | میانہ | | | | |
| 0.9189 | 28 | 30.48 | 0.9391 | 0.2305 | 85074 | 1 |
| 0.8944 | 36 | 37.55 | 0.8983 | 0.2594 | 95727 | 2 |
| 0.9085 | 42 | 44.5035 | 0.7828 | 0.15996 | 59020 | 3 |
| 0.9513 | 48 | 50.34 | 0.6021 | 0.1113 | 41089 | 4 |
| 0.9680 | 53 | 55.03 | 0.4297 | 0.0783 | 28892 | 5 |
| 0.9913 | 56 | 57.79 | 0.3066 | 0.06141 | 22657 | 6 |
| 1.0193 | 59 | 60.089 | 0.2180 | 0.04300 | 15866 | 7 |
| 1.0426 | 61 | 61.37 | 0.1382 | 0.05588 | 20620 | +8 |
| 0.9593 | 27 | 29.91 | 0.7098 | %100 | 368945 | کل |

بخش ب:

در این بخش به دو متد مختلف میتوان عمل کرد،

روش نخست آن است که داده های بدست آمده در جدول مرحله قبل استفاده نماییم و رگرسیون و اسکتر پلات را برای آن بدست بیاوریم:



The summary is:

```
call:
lm(formula = y ~ x)

Residuals:
    Min       1Q   Median       3Q      Max
-0.39192 -0.32037 -0.00531  0.31340  0.39284

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.6935     0.2639   32.94 5.21e-08 ***
x             -7.7751     0.4303  -18.07 1.85e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3554 on 6 degrees of freedom
Multiple R-squared:  0.982,    Adjusted R-squared:  0.9789
F-statistic: 326.4 on 1 and 6 DF, p-value: 1.85e-06

> |
```

توجه شود که با توجه به مقدار P_Value مقادیر ضرایب معتبر و معنا دار است.

از طرف دیگر میتوان از تمام داده های مادران موجود در دیتا ست اصلی استفاده کرد و اثر سواد بر روی تعداد فرزندان مادر را بدست آورد:

```
> summary(fit)

Call:
lm(formula = y ~ x)

Residuals:
    Min       1Q   Median       3Q      Max
-0.046502 -0.042539 -0.003005  0.041465  0.054007

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.10768    0.03530   31.38 6.96e-08 ***
x           -0.12629    0.00699  -18.07 1.85e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0453 on 6 degrees of freedom
Multiple R-squared:  0.982,    Adjusted R-squared:  0.9789
F-statistic: 326.4 on 1 and 6 DF, p-value: 1.85e-06

> print(fit$coefficients)
(Intercept)      x
  1.1076786 -0.1262952
> print(standard_error)
[1] 0.04194043
> res
      1      2      3      4      5      6      7      8
-0.042283333  0.043211905  0.054007143 -0.000397619 -0.046502381 -0.043307143 -0.005611905  0.040883333
> |
```

Handwritten notes:

- P-value* (pointing to the p-value in the coefficients table)
- coefficients* (pointing to the coefficients table)
- residuals* (pointing to the residuals table)

همان طور که می بینید در این جا نیز مقادیر ضریب بدست آمده با توجه به P_Value معنا دار می باشد. همچنین مشخص شد که هر چه سواد داشتن مادر ها با تعداد فرزندان مادران اثر منفی میگذارد یعنی اثر بی سوادی بر افزایش تعداد فرزندان بیشتر از اثر سواد بر افزایش تعداد فرزندان است. مقدار باقی مانده ها و R_Value نیز نشان می دهد که بع علت ان که داده ها گسسته می باشند و کتگوریکال هستند، این مدل، مدل مناسبی نمی باشد.

بخش ج:

خطای بدست آمده در مرحله قبل برابر با 0.0419 می باشد.

در این مرحله دو پارامتر دیگر را به پارامتر سواد برای مدل کردن تعداد فرزندان مادرها اضافه می کنیم. این پارامترهای اضافه معیار وضعیت شغلی مادرها و معیار شهری یا روستایی بودن می باشد:

LM

Shahri

Education Employed

```
Call:
lm(formula = y ~ x + x2 + x3)

Residuals:
    Min       1Q   Median       3Q      Max
-3.9154 -1.4329 -0.4329  0.9993 29.0846

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.949042   0.012130  -78.236 < 2e-16 ***
x             2.432242   0.008129 299.209 < 2e-16 ***
x2           -0.050295   0.009097  -5.529 3.23e-08 ***
x3           -0.112343   0.004718 -23.811 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.835 on 184996 degrees of freedom
Multiple R-squared:  0.3308,    Adjusted R-squared:  0.3308
F-statistic: 3.049e+04 on 3 and 184996 DF, p-value: < 2.2e-16

> print(fit$coefficients)
(Intercept)          x          x2          x3
-0.94904159  2.43224245 -0.05029465 -0.11234330
> res = fit$residuals
> standard_error = sd(res)
> print(standard_error)
[1] 1.834617
```

بخش د:

با به کارگیری GLM برای مدل بدست آمده در مرحله قبل میبینیم که :

```

      Education      Shahri      Employed

Call:
glm(formula = y ~ x + x2 + x3, family = poisson())

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0493 -1.4891 -0.7067  0.7597  9.2326

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.387435   0.006451  -215.08 <2e-16 ***
x             1.433482   0.003486   411.19 <2e-16 ***
x2            0.057152   0.003794    15.06 <2e-16 ***
x3           -0.037235   0.002283   -16.31 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 534084  on 184999  degrees of freedom
Residual deviance: 353668  on 184996  degrees of freedom
AIC: 626752

Number of Fisher Scoring iterations: 6

> print(fit$coefficients)
(Intercept)          x          x2          x3
-1.38743494  1.43348239  0.05715219 -0.03723474
> res = fit$residuals
> standard_error = sd(res)
> print(standard_error)
[1] 1.440347
```

همان طور که دیدیم مقدار خطا در این حالت از مقدار خطا در مرحله قبل کم تر می شود و همچنین مقدار P-Value نیز کاهش پیدا میکند. پس در نتیجه هنگامی که مدل مورد استفاده پیچیده می شود، استفاده از glm به جای lm باعث خطای استاندارد کمتری خواهد شد. از عوامل آن که مدل خطی توصیف کننده خوبی نیست می توان به کتگوریکال بودن نوع داده ها اشاره کرد که نیاز به توضیف و مدل پیچیده تری را ایجاد میکند.

بخش ه:

ماتریس کوواریانس برای متغیرهای ذکر شده به شکل زیر می باشد:

```
> cor(df,use="pairwise")
      num_of_child age_of_mother education working_Statues
num_of_child      1.00000000    0.75838159 0.57329847    -0.03698754
age_of_mother      0.75838159    1.00000000 0.68971723     0.05041066
education          0.57329847    0.68971723 1.00000000     0.01422403
working_Statues    -0.03698754    0.05041066 0.01422403     1.00000000
```

حال هم به روش lm اثر سن مادر و وضعیت شغلی بر تعداد فرزند را بررسی خواهیم کرد:

```
Call:
glm(formula = num_of_child ~ age_of_mother, data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-7.5953  -0.7850  -0.0865   0.6830  27.9368

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.0485757  0.0063838  -164.3  <2e-16 ***
age_of_mother  0.0873118  0.0001745   500.4  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 2.136944)

    Null deviance: 930501  on 184999  degrees of freedom
Residual deviance: 395330  on 184998  degrees of freedom
AIC: 665496

Number of Fisher Scoring iterations: 2

>
> fit2 <- glm(num_of_child~working_Statues,data=df)
> summary(fit2)

Call:
glm(formula = num_of_child ~ working_Statues, data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.676  -1.676  -0.676   1.324   31.324

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.676034  0.005365  312.40  <2e-16 ***
working_Statues -0.091740  0.005763  -15.92  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 5.02291)

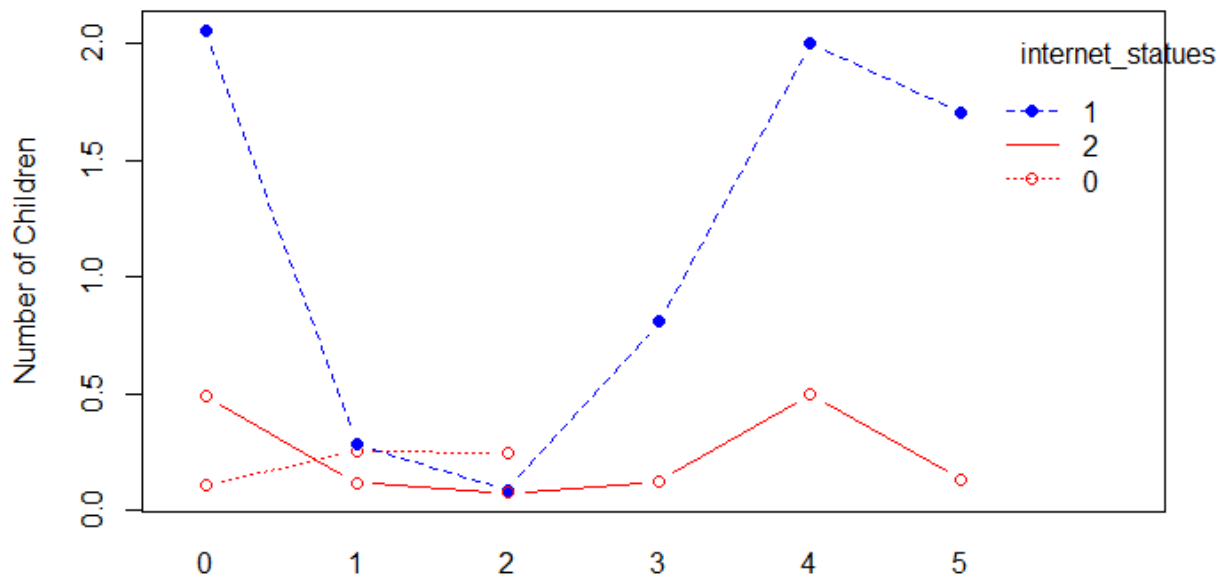
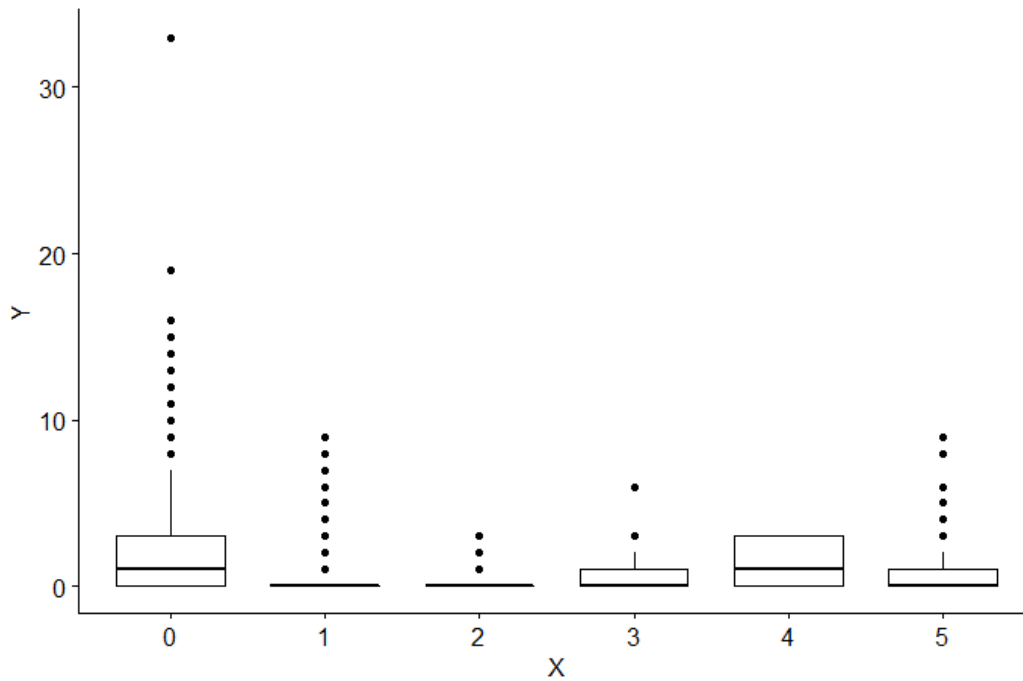
    Null deviance: 930501  on 184999  degrees of freedom
Residual deviance: 929228  on 184998  degrees of freedom
AIC: 823603

Number of Fisher Scoring iterations: 2
```

از آن جایی که مقدار P_Value در هر دو حالت بالا مقدار اندکی را دارد در نتیجه به ماتریس کوواریانس مراجعه خواهیم نمود. در این ماتریس همبستگی بین سن مادر و تعداد فرزندان برابر با 0.75 و بین وضعیت شغلی و تعداد فرزندان مادر برابر با -0.03 می باشد. پس اثر سن بسیار بیشتر است و همچنین از آنجایی که مقدار کوواریانس بین تعداد فرزندان مادر و شغل منفی است یعنی آن که اثری معکوس دارند هرچند این اثر مقدار بسیار کمی است چرا که کوواریانس خیلی کوچک است. توجه شود که با مشاهده ی همبستگی متوجه میشویم که متغیرها به تنهایی نمیتوانند مدل را توصیف کنند و باید متغیرهای جدیدی را وارد کرد تا بتواند واریانس و تغییرات بیشتری از متغیر پاسخ را توضیح دهد. در مجموع با توجه به این بررسی های انجام شده در این سوال و سوال بخش ب میتوان گفت که سن و سواد و معیار شهری یا روستایی بودن برای توصیف تعداد فرزندان متغیرهای بهتری میباشد و همبستگی بیشتری دارند.

بخش ح:

به وسیله ی interaction.plot اثر دو متغیر وضعیت آماده به کار بودن و اینترنت را بر تعداد فرزندان بررسی خواهیم کرد:



حال دو مدل با استفاده از lm و Glm در حالی که همکاری و تعامل بین وضعیت شغلی فرد مورد نظر و شاخصه اینترنت در آن ها در نظر گرفته شده است و در نظر گرفته نشده است را طراحی خواهیم کرد. سپس از تست ANOVA استفاده میکنیم تا ببینیم در حالت داشتن تعامل در مدل، توصیف بهتری نتیجه شده است یا خیر:

در حالت lm :

```
> summary(residuals_statel)

call:
lm(formula = num_of_child ~ internet_statues + IsJobReady)

Residuals:
    Min       1Q   Median       3Q      Max
-1.8886 -1.6882 -0.6882  1.3118 31.3118

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.48784    0.01252   118.83  <2e-16 ***
internet_statues 0.20037    0.01141    17.56  <2e-16 ***
IsJobReady    -1.04901    0.02469   -42.49  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.231 on 184997 degrees of freedom
Multiple R-squared:  0.01049, Adjusted R-squared:  0.01048
F-statistic: 980.2 on 2 and 184997 DF, p-value: < 2.2e-16

> summary(residuals_state2)

call:
lm(formula = num_of_child ~ internet_statues + IsJobReady + internet_statues *
    IsJobReady)

Residuals:
    Min       1Q   Median       3Q      Max
-1.8967 -1.6881 -0.6881  1.3119 31.3119

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.47954    0.01266  116.862  < 2e-16 ***
internet_statues 0.20856    0.01156   18.039  < 2e-16 ***
IsJobReady    -0.75115    0.07179  -10.463  < 2e-16 ***
internet_statues:IsJobReady -0.21464    0.04858   -4.418  9.95e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.231 on 184996 degrees of freedom
Multiple R-squared:  0.01059, Adjusted R-squared:  0.01057
F-statistic: 660.1 on 3 and 184996 DF, p-value: < 2.2e-16
```

```
> anova(residuals_statel,residuals_state2)
```

Analysis of Variance Table

Model 1: num_of_child ~ internet_statues + IsJobReady

Model 2: num_of_child ~ internet_statues + IsJobReady + internet_statues * IsJobReady

| | Res.Df | RSS | Df | Sum of Sq | F | Pr(>F) |
|---|--------|--------|----|-----------|--------|---------------|
| 1 | 184997 | 920744 | | | | |
| 2 | 184996 | 920647 | 1 | 97.149 | 19.521 | 9.954e-06 *** |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

در GLM خواهیم داشت:

```
Call:
glm(formula = num_of_child ~ internet_statues + IsJobReady, family = poisson())
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9601  -1.8395  -0.5763   0.9057  12.7689
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.398893   0.004453   89.57  <2e-16 ***
internet_statues 0.126980   0.003985   31.86  <2e-16 ***
IsJobReady    -1.677319   0.025204  -66.55  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(Dispersion parameter for poisson family taken to be 1)

```
Null deviance: 534084 on 184999 degrees of freedom
Residual deviance: 524133 on 184997 degrees of freedom
AIC: 797215
```

Number of Fisher Scoring iterations: 6

```
> summary(residuals_state4_GLM)
```

```
Call:
glm(formula = num_of_child ~ internet_statues + IsJobReady +
    internet_statues * IsJobReady, family = poisson())
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9649  -1.8394  -0.5762   0.9059  11.8876
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.393803   0.004468   88.146 < 2e-16 ***
internet_statues 0.131949   0.003997   33.016 < 2e-16 ***
IsJobReady    -0.522423   0.070022   -7.461 8.6e-14 ***
internet_statues:IsJobReady -0.913412   0.056538  -16.156 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(Dispersion parameter for poisson family taken to be 1)

```
Null deviance: 534084 on 184999 degrees of freedom
Residual deviance: 523846 on 184996 degrees of freedom
AIC: 796930
```

Number of Fisher Scoring iterations: 6

```
> anova(residuals_state3_GLM, residuals_state4_GLM)
```

Analysis of Deviance Table

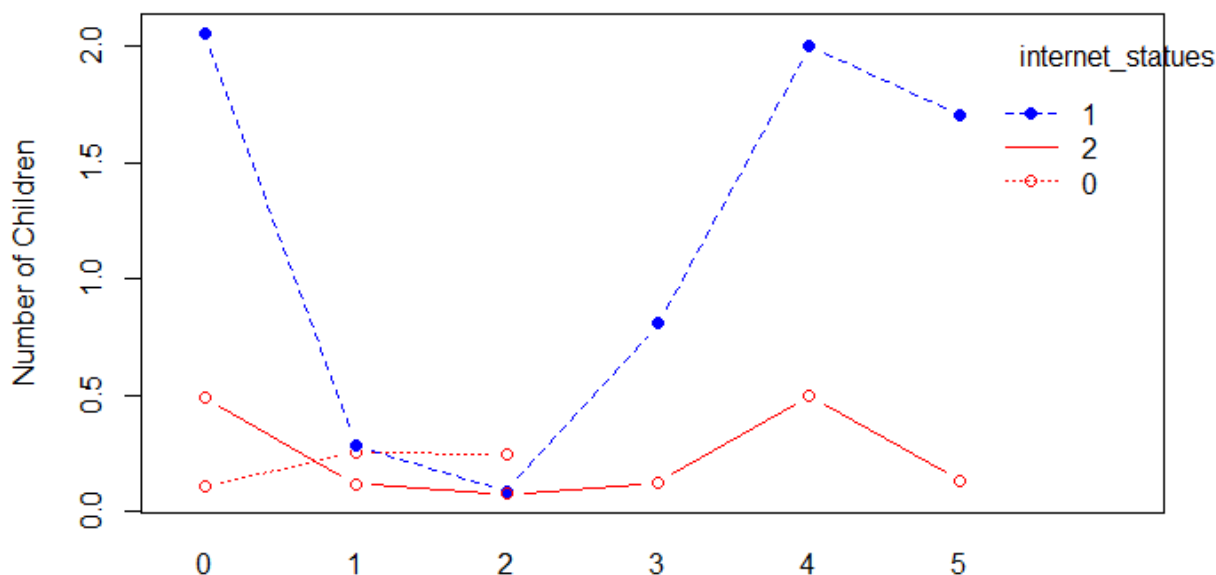
Model 1: num_of_child ~ internet_statues + IsJobReady

Model 2: num_of_child ~ internet_statues + IsJobReady + internet_statues * IsJobReady

| | Resid. Df | Resid. Dev | Df | Deviance |
|---|-----------|------------|----|----------|
| 1 | 184997 | 524133 | | |
| 2 | 184996 | 523846 | 1 | 287.27 |

با بررسی این تست های ANOVA در بالا, متوجه شدیم که بین این دو متغیر تعامل وجود دارد و مدل های با توصیف بهتری را می سازند.

اگر به نموداری که در جلوتر رسم کردیم بازگردیم:



و با توجه به آن که محور افقی آن برابر با وضعیت شغلی می باشد که برابر است با :

- 1: آماده برای کار است
- 2: تحصیل
- 3: خانه داری
- 4: داشتن درآمد بدون کار
- 5: سایر

در نتیجه استفاده از اینترنت با آماده برای کار بود و یا آماده نبودن به جهت تحصیل تعامل بیشتری را دارد و تعامل کمتری با سطوح دیگر دارد.