

# نام و نام خانوادگی : سرمهد زندی گوهر ریزی

شماره دانشجویی : ۸۱۰۱۹۹۱۸۱

درس : استنباط آماری

مدرس : دکتر بهنام بهرگ



## ProjectPhase-2 (Report)

مقدمه

علم آمار تشکیل شده از ۴ جز است که جمع آوری داده های مرتبط با مسئله ای مورد مطالعه، یکی از اجزای علم آمار تلقی می شود. در علم آمار، داده های جمع آوری شده (observations) را درون مجموعه داده ها (dataset) قرار می دهیم. در یک dataset، هر سطر نشان دهنده یک case یا observation است و هر ستون را متغیرهایی می دانیم که برای هر اطلاعاتی را در اختیار ما قرار می دهند.

در این پژوهه مجموعه داده ای که مورد بررسی قرار می گیرد، HealthCare نام دارد که شامل اطلاعات وضعیت سلامتی حدود ۵۰۰۰ نفر و همچنین میزان هزینه ای سالیانه سلامتی آنهاست.

در فاز اول این پژوهه ما به بررسی کامل مجموعه داده HealthCare ویژگی های آن پرداختیم. اما در این فاز از پژوهه در ابتدا قصد داریم تا مروری کوتاه بر روی این مجموعه داده و ویژگی های آن داشته باشیم.

برای بدست آوردن اطلاعات بیشتر در مورد مجموعه داده Healthcare در ابتدا به کمک R آن را import می کنیم. آن چیزی که مشخص است این dataset اطلاعات فردی 5110 نفر شامل جنسیت (male - female - other)، سن و شماره شناسایی آن فرد، و همچنین اطلاعات سلامتی این 5110 نفر که شامل فشار خون (= فشار خون پایین و = فشار خون بالا)، سابقه بیماری قلبی (= داشتن و = نداشتن)، وضعیت تاہل (yes = متأهل بوده است و no = مجرد است)، نوع شغل، نوع محل سکونت (Urban/Rural)، سطح متوسط گلوکز (قد خون)، شاخص bmi فرد، وضعیت استعمال دخانیات و سابقه سکته مغزی (= داشتن و = نداشتن) می باشند را به همراه میزان هزینه سالیانه سلامتی آنها در اختیار ما قرار می دهد.

### کردن dataset و مشاهده آن در (R)

```
# Importing the Healthcare dataset
HealthCare <- read.csv("H:/Second Term/statistical Inference/Projects/ProjectPhase2/HealthCare.csv")
View(HealthCare)
```

	<b>id</b>	<b>gender</b>	<b>age</b>	<b>hypertension</b>	<b>heart_disease</b>	<b>ever_married</b>	<b>work_type</b>	<b>Residence_type</b>	<b>avg_glucose_level</b>	<b>bmi</b>	<b>smoking_status</b>	<b>stroke</b>	<b>health_bills</b>
1	9046	Male	67.00	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1	6011.860
2	51676	Female	61.00	0	0	Yes	Self-employed	Rural	202.21	N/A	never smoked	1	N/A
3	31112	Male	80.00	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1	6384.530
4	60182	Female	49.00	0	0	Yes	Private	Urban	171.23	34.4	smokes	1	5862.754
5	1665	Female	79.00	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1	5461.262
6	56669	Male	81.00	0	0	Yes	Private	Urban	186.21	29.0	formerly smoked	1	5054.021
7	53882	Male	74.00	1	1	Yes	Private	Rural	70.09	27.4	never smoked	1	6795.934
8	10434	Female	69.00	0	0	No	Private	Urban	94.39	22.8	never smoked	1	5158.242
9	27419	Female	59.00	0	0	Yes	Private	Rural	76.15	N/A	Unknown	1	N/A
10	60491	Female	78.00	0	0	Yes	Private	Urban	58.57	24.2	Unknown	1	5205.963
11	12109	Female	81.00	1	0	Yes	Private	Rural	80.43	29.7	never smoked	1	6055.760
12	12095	Female	61.00	0	1	Yes	Govt_job	Rural	120.46	36.8	smokes	1	8146.564
13	12175	Female	54.00	0	0	Yes	Private	Urban	104.51	27.3	smokes	1	5433.346
14	8213	Male	78.00	0	1	Yes	Private	Urban	219.84	N/A	Unknown	1	N/A
15	5317	Female	79.00	0	1	Yes	Private	Urban	214.09	28.2	never smoked	1	5051.455
16	58202	Female	50.00	1	0	Yes	Self-employed	Rural	167.41	30.9	never smoked	1	5975.905
17	56112	Male	64.00	0	1	Yes	Private	Urban	191.61	37.5	smokes	1	5195.593
18	34120	Male	75.00	1	0	Yes	Private	Urban	221.29	25.8	smokes	1	4173.739
19	27458	Female	60.00	0	0	No	Private	Urban	89.22	37.8	never smoked	1	6577.844

متغیرهای Dataset در یک Categorical تعداد محدودی دسته (category) را می‌توانند اختیار کنند. در این سوال، متغیرهای انتخابی، "smoking\_status" است که برای هر فرد (Case) در این Dataset وضعیت استعمال دخانیات آن را با یکی از category های "Unknown" یا "formerly smoked", "never smoked", "smokes" مشخص می‌کند، (لازم به ذکر است که وجود داشتن مقدار Unknown برای یک فرد (Case) به معنی در دست نبودن اطلاعات برای آن است). و متغیر دوم "ever\_married" است که وضعیت متاهل بودن یا مجرد بودن observation های ما را در این مجموعه داده نشان می‌دهد. برای هر case، مقدار yes نشان دهنده متاهل بودن آن فرد، و مقدار no نشان دهنده مجرد بودن آن فرد است.

### ۱. Part A

در این قسمت می‌خواهیم یک بازه اطمینان ۹۵٪ برای اختلاف بین این دو متغیر انتخابی ایجاد کنیم. می‌دانیم احتمال اینکه تخمین نقطه‌ای که برای پارامتر واقعی جامعه آماری مان در یک مطالعه با استفاده از یک sample میزئیم خیلی احتمال کمی دارد که دقیقاً برابر با پارامتری شود که به دنبال آن هستیم. در مفهوم بازه اطمینان به دنبال ارائه یک بازه هستیم که بتوانیم با اطمینان بالای بگوییم که این بازه شامل پارامتر واقعی جامعه هدف می‌شود. در این سوال پارامتری که به دنبال آن هستیم "اختلاف بین دو proportion" است. در این سوال، دو متغیر انتخابی بیش از دو سطح دارند، و من برای همه ی زوج سطح‌های ممکن اقدام به ساخت بازه اطمینان می‌کنم. بدین ترتیب بازه اطمینان برای اختلاف بین نسبت :

- ۱- "افراد مجردی که در دسته smokes هستند، با افراد متاهلی که در دسته smokes هستند"
- ۲- "افراد مجردی که در دسته never smoked هستند، با افراد متاهلی که در دسته never smoked هستند"
- ۳- "افراد مجردی که در دسته formerly smoked هستند، با افراد متاهلی که در دسته formerly smoked هستند"
- ۴- "افراد مجردی که در دسته Unknown هستند، با افراد متاهلی که در دسته Unknown هستند"

می‌دانیم که برای ساخت بازه اطمینان در اولین قدم باید شرایط قضیه حد مرکزی (CLT) را بررسی کنیم.

اما در ابتدا باید بدانیم که در این سوال جامعه هدف ما، مجموعه داده HealthCare است و برای انجام این سوال اولین قدم گرفتن یک نمونه است، که من نمونه‌ای با سایز ۱۰۰ به طور تصادفی انتخاب می‌کنم.

(دستور ایجاد و نمایش نتیجه آن در R)

```
# randomly select a sample (n = 100) without replacement from Healthcare
sample_1_A <- Healthcare[sample(nrow(Healthcare), size = 100, replace = FALSE),]
```

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke	health_bills
4731	64879	Female	8.00	0	0	NO	children	Rural	120.43	23.5	Unknown	0	2436.4100
4115	64196	Male	26.00	0	0	NO	Private	Urban	64.68	23.3	smokes	0	3353.6400
4745	163	Female	20.00	0	0	NO	Private	Rural	94.67	28.8	Unknown	0	2564.9576
2704	16971	Female	26.00	0	0	NO	Private	Urban	100.31	38.6	never smoked	0	3646.4552
2464	45864	Female	36.00	0	0	NO	Private	Rural	55.58	30.0	never smoked	0	3716.5611
622	42553	Female	80.00	0	0	Yes	Private	Rural	148.91	28.3	never smoked	0	2921.9053
3276	70973	Female	50.00	0	0	Yes	Govt_job	Urban	151.25	31.5	never smoked	0	2969.5544
2504	4949	Male	49.00	0	0	Yes	Private	Rural	96.35	35.9	never smoked	0	3660.3242
3615	48459	Male	61.00	0	0	Yes	self-employed	Urban	111.94	26.5	smokes	0	2015.8107
4197	52447	Female	3.00	0	0	NO	children	Rural	131.81	14.1	Unknown	0	2532.4076
3669	11651	Female	25.00	0	0	Yes	Private	Rural	81.21	37.9	never smoked	0	3254.0777
4601	50804	Male	2.00	0	0	NO	children	Rural	65.84	16.1	Unknown	0	3275.3690
2120	25930	Male	42.00	0	0	Yes	Private	Urban	68.24	33.1	formerly smoked	0	2806.1091
3358	32717	Male	16.00	0	0	NO	children	Rural	106.11	22.4	Unknown	0	3065.9866
1534	31415	Female	54.00	0	0	Yes	Private	Urban	207.79	38.6	never smoked	0	2870.1756
1976	6596	Male	0.56	0	0	NO	children	Rural	111.77	21.1	Unknown	0	3038.9521
3564	15136	Male	64.00	0	1	Yes	Private	Rural	109.88	33.9	Unknown	0	893.9138
5013	14688	Female	44.00	0	0	Yes	Private	Urban	73.87	28.8	smokes	0	2410.5348

### بررسی شرایط CLT

شرط independence : از آنجایی که نمونه گیری به صورت تصادفی صورت گرفته است و random sample/assignment داریم، و همچنین سایز نمونه گرفته شده از ۱۰٪ جامعه کوچک تر است ( $100 < 10\% \text{ of population}$ ) بنابراین شرط استقلال درون گروهی برقرار است. همچنین از آنجایی که در هر ۴ حالت ذکر شده، هر دو گروه از هم مستقل هستند و اصطلاحاً non-paired هستند شرط استقلال بین گروهی هم برقرار است. بنابراین در کل شرط استقلال برقرار است.

شرط success-failure باشد :

برای بررسی این شرط برای هر ۴ حالت ، در ابتدا فرکانس تکرار هر دو گروه در هر ۴ حالت ذکر شده را توسط تابع `table()` در بدست می آوریم.

(فرکانس تکرار هر دو گروه در هر ۴ حالت ذکر شده و نمایش نتیجه آن در R)

```
# to understand the frequency
table_1_A <- table(unlist(sample_1_A$ever_married),unlist(sample_1_A$smoking_status))
table_1_A_1 <- table(sample_1_A$ever_married)
```

```
> table_1_A
      formerly smoked never smoked smokes unknown
  No           15          52       26     95
  Yes          70         123       56     63
> table_1_A_1
  No Yes
188 312
```

در این نمونه گرفته شده ، ۳۱۲ نفر متاهل هستند و ۱۸۸ نفر مجرد.

به عنوان مثال ، ۹۵ نفر از بین ۱۸۸ نفری که مجرد هستند در دسته `unknown` قرار گرفته اند و همچنان ۶۳ نفر از بین ۳۱۲ نفری که متاهل هستند در این دسته قرار گرفته اند.

حال باید در هر ۴ حالت ذکر شده ، برای هر دو گروه متاهل و مجرد ، شرط success-failure برقار باشد :

يعني :

$$n_1 \hat{p}_1 \geq 10 \text{ and } n_1(1 - \hat{p}_1) \geq 10$$

$$n_2 \hat{p}_2 \geq 10 \text{ and } n_2(1 - \hat{p}_2) \geq 10$$

حالت اول : "نسبت افراد مجردی که در دسته `smokes` هستند ، با نسبت افراد متاهلی که در دسته `smokes` هستند"

\* بنده برای چک کردن این شرایط برای هر دو گروه ، در R کدی را نوشته ام که اگر برای آن ، شرط برقار باشد عبارت TRUE و اگر برقار نباشد عبارت FALSE را برمیگرداند.

```
#=====
#no ~ smokes
(table_1_A[5] * (table_1_A[5]/table_1_A_1[1])) >= 10
(table_1_A[5] * (1 - (table_1_A[5]/table_1_A_1[1]))) >= 10
#=====
#yes ~ smokes
(table_1_A[6] * (table_1_A[6]/table_1_A_1[2])) >= 10
(table_1_A[6] * (1 - (table_1_A[6]/table_1_A_1[2]))) >= 10
#=====

> #=====
> #no ~ smokes
> (table_1_A[5] * (table_1_A[5]/table_1_A_1[1])) >= 10
  No
  FALSE
  n_1 \hat{p}_1 \geq 10
> (table_1_A[5] * (1 - (table_1_A[5]/table_1_A_1[1]))) >= 10
  No
  TRUE
  n_1(1 - \hat{p}_1) \geq 10
>
> #yes ~ smokes
> (table_1_A[6] * (table_1_A[6]/table_1_A_1[2])) >= 10
  Yes
  TRUE
  n_2 \hat{p}_2 \geq 10
> (table_1_A[6] * (1 - (table_1_A[6]/table_1_A_1[2]))) >= 10
  Yes
  TRUE
  n_2(1 - \hat{p}_2) \geq 10
> #=====
```

حالت دوم : "نسبت افراد مجردی که در دسته never smoked هستند ، با نسبت افراد متاھلی که در دسته smoked هستند"

```
#=====
#no ~ never smoked
(table_1_A[3] * (table_1_A[3]/table_1_A_1[1])) >= 10
(table_1_A[3] * (1 - (table_1_A[3]/table_1_A_1[1]))) >= 10
#=====

> #=====
> #no ~ never smoked
> (table_1_A[3] * (table_1_A[3]/table_1_A_1[1])) >= 10
  No
  TRUE ←  $n_1\hat{p}_1 \geq 10$ 
> (table_1_A[3] * (1 - (table_1_A[3]/table_1_A_1[1]))) >= 10
  No
  TRUE ←  $n_1(1 - \hat{p}_1) \geq 10$ 
>
> #yes ~ never smoked
> (table_1_A[4] * (table_1_A[4]/table_1_A_1[2])) >= 10
  Yes
  TRUE ←  $n_2\hat{p}_2 \geq 10$ 
> (table_1_A[4] * (1 - (table_1_A[4]/table_1_A_1[2]))) >= 10
  Yes
  TRUE ←  $n_2(1 - \hat{p}_2) \geq 10$ 
#_
```

حالت سوم : "نسبت افراد مجردی که در دسته formerly smoked هستند ، با نسبت افراد متاھلی که در دسته formerly smoked هستند"

```
#=====
#no ~ formerly smoked
(table_1_A[1] * (table_1_A[1]/table_1_A_1[1])) >= 10
(table_1_A[1] * (1 - (table_1_A[1]/table_1_A_1[1]))) >= 10
#=====

> #=====
> #no ~ formerly smoked
> (table_1_A[1] * (table_1_A[1]/table_1_A_1[1])) >= 10
  No
  FALSE ←  $n_1\hat{p}_1 \geq 10$ 
> (table_1_A[1] * (1 - (table_1_A[1]/table_1_A_1[1]))) >= 10
  No
  TRUE ←  $n_1(1 - \hat{p}_1) \geq 10$ 
>
> #yes ~ formerly smoked
> (table_1_A[2] * (table_1_A[2]/table_1_A_1[2])) >= 10
  Yes
  TRUE ←  $n_2\hat{p}_2 \geq 10$ 
> (table_1_A[2] * (1 - (table_1_A[2]/table_1_A_1[2]))) >= 10
  Yes
  TRUE ←  $n_2(1 - \hat{p}_2) \geq 10$ 
#_
```

حالت چهارم : "نسبت افراد مجردی که در دسته Unknown هستند ، با نسبت افراد متأهلی که در دسته هستند" هستند

```
#=====
#no ~ unknown
(table_1_A[7] * (table_1_A[7]/table_1_A_1[1])) >= 10
(table_1_A[7] * (1 - (table_1_A[7]/table_1_A_1[1]))) >= 10

#yes ~ unknown
(table_1_A[8] * (table_1_A[8]/table_1_A_1[2])) >= 10
(table_1_A[8] * (1 - (table_1_A[8]/table_1_A_1[2]))) >= 10
#=====

> #
> #no ~ unknown
> (table_1_A[7] * (table_1_A[7]/table_1_A_1[1])) >= 10
  NO
     $n_1 \hat{p}_1 \geq 10$ 
  TRUE ←
> (table_1_A[7] * (1 - (table_1_A[7]/table_1_A_1[1]))) >= 10
  No
  TRUE ←
     $n_1(1 - \hat{p}_1) \geq 10$ 
  >
> #yes ~ unknown
> (table_1_A[8] * (table_1_A[8]/table_1_A_1[2])) >= 10
  Yes
     $n_2 \hat{p}_2 \geq 10$ 
  TRUE ←
> (table_1_A[8] * (1 - (table_1_A[8]/table_1_A_1[2]))) >= 10
  Yes
  TRUE ←
     $n_2(1 - \hat{p}_2) \geq 10$ 
> #=====
```

بنابراین طبق مشاهدات صورت گرفته ، در حالت اول و سوم ، یکی از شرایط (که مشخص شده است) نقض شده است اما همه شرایط برای هر دو گروه در حالت دوم و چهارم برقرار هستند. اما با فرض برقرار بودن همه شرایط برای هر دو گروه در هر ۴ حالت ذکر شده به این سوال پاسخ می دهیم.

ساخت بازه اطمینان :

می دانیم که برای ساخت بازه اطمینان برای اختلاف بین دو proportion ، فرم کلی آن برابر است با :

$$(\hat{p}_1 - \hat{p}_2) \pm z^* SE_{(\hat{p}_1 - \hat{p}_2)}$$

$$SE_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

برای بازه اطمینان 95% ، آماره Z برابر است با :

```
> Z <- qnorm(0.975)
> Z
[1] 1.959964
```

حال به کمک R و فرم کلی بازه اطمینان ، اقدام به ساخت بازه اطمینان برای هر ۴ حالت ذکر شده می کنیم و برای هر حالت ، بازه اطمینان ساخته شده برای آن را تفسیر می کنیم :

**حالت اول** : اختلاف بین نسبت افراد مجردی که در دسته `smokes` هستند، با نسبت افراد متأهلی که در دسته هستند

```
#=====
# Z* for 95% CL
Z <- qnorm(0.975)

# CI for smokes
SE_smokes <- sqrt( ( ((table_1_A[5]/table_1_A_1[1]) * (1 - (table_1_A[5]/table_1_A_1[1]))) / table_1_A[5] ) +
                      ( ((table_1_A[6]/table_1_A_1[2]) * (1 - (table_1_A[6]/table_1_A_1[2]))) / table_1_A[6] ) )

# point estimate
point_estimate_smokes <- ((table_1_A[5]/table_1_A_1[1]) - (table_1_A[6]/table_1_A_1[2]))

# point estimate ± Z*SE
CI_smokes <- c(point_estimate_smokes - (Z * SE_smokes) , point_estimate_smokes + (Z * SE_smokes))
as.numeric(CI_smokes)
#=====
```

> # point estimate ± Z\*SE  
> CI\_smokes <- c(point\_estimate\_smokes - (Z \* SE\_smokes) , point\_estimate\_smokes + (Z \* SE\_smokes))  
> as.numeric(CI\_smokes)  
[1] -0.2076521 0.1252735 ← بازه اطمینان :  
> #=====

کد :

پاسخ :

تفسیر: ما ۹۵ درصد اطمینان داریم که اختلاف بین نسبت افراد مجردی که از نظر عادت سیگار کشیدن در دسته `smokes` قرار می گیرند، با نسبت افراد متأهل که در این دسته قرار دارند، بین ۰.۲۰- تا ۰.۱۲ است.

**حالت دوم** : اختلاف بین نسبت افراد مجردی که در دسته `never smoked` هستند، با نسبت افراد متأهلی که در دسته هستند `never smoked`

```
#=====
# Z* for 95% CL
Z <- qnorm(0.975)

# CI for never smoked
SE_never <- sqrt( ( ((table_1_A[3]/table_1_A_1[1]) * (1 - (table_1_A[3]/table_1_A_1[1]))) / table_1_A[3] ) +
                      ( ((table_1_A[4]/table_1_A_1[2]) * (1 - (table_1_A[4]/table_1_A_1[2]))) / table_1_A[4] ) )

# point estimate
point_estimate_never <- ((table_1_A[3]/table_1_A_1[1]) - (table_1_A[4]/table_1_A_1[2]))

# point estimate ± Z*SE
CI_never <- c(point_estimate_never - (Z * SE_never) , point_estimate_never + (Z * SE_never))
as.numeric(CI_never)
#=====
```

> # point estimate ± Z\*SE  
> CI\_never <- c(point\_estimate\_never - (Z \* SE\_never) , point\_estimate\_never + (Z \* SE\_never))  
> as.numeric(CI\_never)  
[1] -0.26676579 0.03149574 ← بازه اطمینان :  
> #=====

کد :

پاسخ :

تفسیر: ما ۹۵ درصد اطمینان داریم که اختلاف بین نسبت افراد مجردی که تا به حال سیگار نکشیده اند، با نسبت افراد متأهل که تا به حال سیگار نکشیده اند، بین ۰.۲۶۶- تا ۰.۰۳۱ است.

**حالت سوم** : "اختلاف بین نسبت افراد مجردی که در دسته formerly smoked هستند، با نسبت افراد متاهلی که در دسته formerly smoked هستند"

```
#=====#
# Z* for 95% CL
Z <- qnorm(0.975)

# CI for formerly smoked
SE_formerly <- sqrt( ((table_1_A[1]/table_1_A_1[1]) * (1 - (table_1_A[1]/table_1_A_1[1]))) / table_1_A[1] ) +
  ((table_1_A[2]/table_1_A_1[2]) * (1 - (table_1_A[2]/table_1_A_1[2]))) / table_1_A[2] )

# point estimate
point_estimate_formerly <- ((table_1_A[1]/table_1_A_1[1]) - (table_1_A[2]/table_1_A_1[2]))

# point estimate ± Z*SE
CI_formerly <- c(point_estimate_formerly - (Z * SE_formerly) , point_estimate_formerly + (Z * SE_formerly))
as.numeric(CI_formerly)
#=====#
# point estimate ± Z*SE
> CI_formerly <- c(point_estimate_formerly - (Z * SE_formerly) , point_estimate_formerly + (Z * SE_formerly))
> as.numeric(CI_formerly)
[1] -0.31295509  0.02381161 ← بازه اطمینان :
```

کد :

پاسخ :

تفسیر: ما ۹۵ درصد اطمینان داریم که اختلاف بین نسبت افراد مجردی که از نظر عادت سیگار کشیدن در دسته formerly smoked قرار می‌گیرند، با نسبت افراد متأهل که در این دسته قرار دارند، بین ۰.۳۱ تا ۰.۰۲ است.

**حالت چهارم** : "اختلاف بین نسبت افراد مجردی که در دسته Unknown هستند، با نسبت افراد متأهلی که در دسته Unknown هستند"

```
#=====#
# Z* for 95% CL
Z <- qnorm(0.975)

# CI for unknown
SE_unknown <- sqrt( ((table_1_A[7]/table_1_A_1[1]) * (1 - (table_1_A[7]/table_1_A_1[1]))) / table_1_A[7] ) +
  ((table_1_A[8]/table_1_A_1[2]) * (1 - (table_1_A[8]/table_1_A_1[2]))) / table_1_A[8] )

# point estimate
point_estimate_unknown <- ((table_1_A[7]/table_1_A_1[1]) - (table_1_A[8]/table_1_A_1[2]))

# point estimate ± Z*SE
CI_unknown <- c(point_estimate_unknown - (Z * SE_unknown) , point_estimate_unknown + (Z * SE_unknown))
as.numeric(CI_unknown)
#=====#
# point estimate ± Z*SE
> CI_unknown <- c(point_estimate_unknown - (Z * SE_unknown) , point_estimate_unknown + (Z * SE_unknown))
> as.numeric(CI_unknown)
[1] 0.1622076 0.4445845 ← بازه اطمینان :
```

کد :

پاسخ :

تفسیر: ما ۹۵ درصد اطمینان داریم که اختلاف بین نسبت افراد مجردی که اطلاعاتی از وضعیت سیگار کشیدن آن ها در دسترس نیست، با نسبت افراد متأهل که اطلاعاتی از وضعیت سیگار کشیدن آن ها در دسترس نمی باشد، بین ۰.۱۶۲ تا ۰.۴۴۴ است.

در این سوال، متغیرهای انتخابی، "smoking\_status" است که برای هر فرد (Case) در این Dataset وضعیت استعمال دخانیات آن را با یکی از category های : "Unknown" یا "formerly smoked", "never smoked", "smokes" مشخص می کند، (لازم به ذکر است که وجود داشتن مقدار Unknown برای یک فرد (Case) به معنی در دست نبودن اطلاعات برای آن است.) و متغیر دوم "ever\_married" است که وضعیت متاهل بودن یا مجرد بودن observation های ما را در این مجموعه داده نشان می دهد. برای هر case، مقدار yes نشان دهنده متاهل بودن آن فرد، و مقدار no نشان دهنده مجرد بودن آن فرد است.

در این سوال می خواهیم بدانیم که این دو متغیر مستقل از هم اند یا خیر و این کار را با انجام آزمون فرض انجام می دهیم.

بنابراین در ابتدا برای انجام تست استقلال ، فرض صفر و فرض جایگزین را مشخص می کنیم، سپس پس از بررسی شرایط این تست، به محاسبه توزیع expected count هر cell می پردازیم و آن ها را درون جدول نمایش می دهیم، آماره آزمون  $\chi^2$  (chi-square) را محاسبه می کنیم و بر اساس آن به همراه درجه آزادی این تست، اقدام به محاسبه p-value و در نهایت تصمیم گیری برای این تست می پردازیم.

$H_0$  : دو متغیر " ever\_married " و " smoking\_status " مستقل هستند.

$H_A$  : دو متغیر " ever\_married " و " smoking\_status " به هم وابسته هستند.

بهتر است که برای انجام این تست ، در ابتدا مجموعه داده را جامعه هدف بدانیم و یک نمونه با سایز 400 از آن انتخاب کنیم.

(دستور ایجاد sample در R)

```
# randomly select a sample (n=400) without replacement from HealthCare
my_sample_1_B <- HealthCare[sample(nrow(HealthCare), size = 400, replace = FALSE),]
```

#### حال باید به بررسی شرایط آزمون پردازیم :

شرط independence : از آنجایی که نمونه های داخل sample ما به صورت تصادفی انتخاب شده اند، یعنی random sample/assignment رعایت شده است. همچنین به دلیل اینکه شرط  $n < 10\% \text{ population}$  (400 < 10% population) برقرار است. همینطور از آنجایی که هر case فقط در یک cell از جدول مقدار دارد و observation ها مستقل از هم اند. بنابراین شرط استقلال برقرار است.

شرط sample size/skew : برای برقرار بودن این شرط در هر cell باید، حداقل 5 expected case وجود داشته باشد.

بنابراین برای بررسی شرط دوم ، لازم داریم تا expected count هر cell را محاسبه کنیم :

برای محاسبه expected count برای هر سطر ، می گوییم اگر دو متغیر مستقل از هم باشند، انتظار داریم که در هر خانه جدول چه مقداری قرار بگیرد و برای محاسبه این مقدار از رابطه زیر استفاده می کنیم :

$$\text{Expected count} = \frac{(\text{row total}) \times (\text{column total})}{\text{table total}}$$

اما در ابتدا با استفاده از R و تابع table() فرکانس تکرار را بدست می آوریم.

```
> # to understand the frequency
> table_1_B <- table(my_sample_1_B$ever_married, my_sample_1_B$smoking_status)
> table_1_B

  formerly smoked never smoked smokes Unknown
  No              10            44       16      73
  Yes             57           101      53      46
```

حال طبق رابطه ذکر شده برای محاسبه expected count ، به محاسبه این مقدار برای هر cell از جدول می کنیم :

```
> # to understand the frequency
> table_1_B <- table(my_sample_1_B$ever_married,my_sample_1_B$smoking_status)
> table_1_B_B <- addmargins(table_1_B)
> # to calculate the expected count of each cell
> chii <- chisq.test(table_1_B_B)
> expcount_1 <- table_1_B_B
> expcount_1[] <- paste(table_1_B_B,paste0("(",round(chii$expected),")"))
> expcount_1
```

	formerly smoked	never smoked	smokes	Unknown	Sum
No	10 (24)	44 (52)	16 (25)	73 (43)	143
Yes	57 (43)	101 (93)	53 (44)	46 (76)	257
Sum	67	145	69	119	400

همانطور که گفته شد برای برقرار بودن این شرط در هر cell باید، حداقل ۵ expected case وجود داشته باشد. که با توجه به جدول بالا این شرط برقرار است.

حال برای انجام تست استقلال از آنجایی که در این سوال هدف استفاده از R برای انجام این تست است ازتابع chisq.test() برای این کار استفاده می کنیم و سپس با  $\alpha = 0.05$  نتایج را تحلیل می کنیم.

بنابراین در R این تابع را اجرا می کنیم :

```
> # independence test using the chisq.test() function
> chisq.test(table(my_sample_1_B$ever_married,my_sample_1_B$smoking_status))

Pearson's Chi-squared test

data: table(my_sample_1_B$ever_married, my_sample_1_B$smoking_status)
X-squared = 53.173, df = 3, p-value = 1.684e-11 ←
```

از آنجایی که مقدار p-value بدست آمده خیلی کوچکتر از 0.05 است، ما فرض صفر را رد می کنیم. بنابراین نتیجه می گیریم دو متغیر "ever\_married" و "smoking\_status" به هم وابسته هستند.

سوال شماره ۲

در این سوال متغیر categorical "ever\_married" بازیزی انتخابی بندۀ، متغیر "ever\_married" هستش که وضعیت متاهل بودن یا مجرد بودن observation های ما را در این مجموعه داده نشان می دهد. برای هر yes case، مقدار yes نشان دهنده متاهل بودن آن فرد، و مقدار no نشان دهنده ی مجرد بودن آن فرد است.

در ابتدا به کمک دستور زیر در R یک نمونه کوچک با اندازه 15 از مجموعه داده HealthCare بدون جایگزاری و به صورت کاملاً تصادفی می گیریم.

(دستور ایجاد sample و نمایش نتیجه آن در R)

```
# randomly select a small sample (n = 15) without replacement from HealthCare
my_sample <- HealthCare[sample(nrow(HealthCare), size = 15, replace = FALSE),]
```

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke	health_bills
727	50277	Female	51	0	0	Yes	Self-employed	Rural	67.97	29.4	smokes	0	3304.037
2254	1329	Female	43	0	0	No	Govt_job	Rural	101.35	32.2	never smoked	0	3846.142
2430	27007	Male	14	0	0	No	Self-employed	Urban	187.22	29.7	Unknown	0	3136.109
2585	3494	Female	80	0	0	Yes	Private	Rural	102.90	26.7	Unknown	0	2177.496
2963	58061	Female	70	1	0	Yes	Self-employed	Rural	154.60	28.5	formerly smoked	0	2945.962
1988	72311	Male	18	0	0	No	Private	Urban	113.24	24.9	Unknown	0	2705.302
4249	15422	Male	31	0	0	No	Govt_job	Rural	80.57	28.2	formerly smoked	0	3196.138
3999	62396	Female	27	0	0	Yes	Private	Urban	139.20	36.2	never smoked	0	3322.789
2524	30953	Male	75	1	1	Yes	Private	Rural	221.43	32.5	Unknown	0	4225.926
4933	5951	Male	28	1	0	No	Private	Urban	86.61	38.6	smokes	0	4124.425
2968	2579	Female	34	0	0	Yes	Self-employed	Rural	78.12	32.0	Unknown	0	3713.106
2861	59335	Male	38	0	0	Yes	Govt_job	Rural	69.88	27.9	smokes	0	3081.463
4670	29078	Male	39	0	0	Yes	Govt_job	Rural	73.07	26.8	smokes	0	2981.106
3711	12674	Male	44	0	0	Yes	Private	Rural	74.15	34.5	formerly smoked	0	2843.351
526	29933	Female	5	0	0	No	children	Rural	86.11	19.0	Unknown	0	3586.260

همانطور که از متغیر انتخابی مشخص است، یک فرد (observation) یا ازدواج کرده است یا خیر، بنابراین ما موفقیت را "ازدواج کردن" و شکست را "ازدواج نکردن" می‌دانیم.

از آنجایی که در این سوال هدف انجام آزمون فرض برای میزان موفقیت (ازدواج کردن) است، در ابتدا فرض اولیه مان را اینگونه در نظر می‌گیریم که درصد افراد متاهل و درصد افراد مجرد در این مطالعه با هم برابر است. و می‌خواهیم این فرض را در مقابل فرض جایگزین که می‌گوید: "اکثربی افراد مورد مطالعه ما متاهل هستند" برسی کنیم.

بنابراین ما یک متغیر categorical باینری داریم که نشان می‌دهد یک فرد یا ازدواج کرده است (موفقیت =  $p$ ) یا ازدواج نکرده است (شکست =  $1-p$ ) و مطابق فرضیات در نظر گرفته شده، آزمون فرض را برای میزان موفقیت انجام می‌دهیم.

#### -۱- فرض صفر و فرض جایگزین :

$$H_0 : p = 0.5$$

$$H_A : p > 0.5$$

-۲- محاسبه point estimate :

$$n = 15 \quad \hat{p} = 0.6$$

نکته: برای محاسبه point estimate، در R با استفاده از تابع table() نسبت افرادی که متاهل هستند و همچنین نسبت افرادی که مجرد هستند را بدست می‌آوریم: (یاد آوری: متغیر "ever\_married": متغیر "ever\_married" وضعیت متاهل بودن یا مجرد بودن observation های ما را در این sample نشان می‌دهد. برای هر case، مقدار yes نشان دهنده متاهل بودن آن فرد (موفقیت)، و مقدار no نشان دهنده ی مجرد بودن آن فرد (شکست) است).

استفاده از تابع table() و محاسبه نسبت مورد نظر و همچنین نمایش نتیجه آن در R

```
# To calculate the point estimate,
# we first get the ratio of people who are married
# and the ratio of people who are single in our sample.
rat <- table(my_sample$ever_married)
vec <- c((rat[1] / 15) , (rat[2] / 15))
vec
```

```
> vec
No Yes
0.4 0.6
```

#### -۳- بررسی شرایط CLT :

شرط استقلال: به دلیل اینکه observation های درون نمونه گرفته شده، کاملاً تصادفی انتخاب شده اند و random sample/assignment گرفته شده از ۰٪ درصد جامعه (افراد درون مجموعه داده) کمتر است (یعنی  $10\% < 15\%$  population) بنابراین شرط استقلال برقرار است.

#### شرط (success-failure) : sample size/skew

$$np \geq 10 \text{ and } n(1 - p) \geq 10$$

$$15 \times 0.5 = 7.5 \not\geq 10 \text{ and } 15 \times (0.5) = 7.5 \not\geq 10$$

اما از آنجایی که این شرط برقرار نیست، در نتیجه نمی‌توانیم بر مبنای قضیه حد مرکزی آزمون فرض را انجام دهیم و سایز نمونه گرفته شده کوچک است. بنابراین به سراغ روش نادقيق "شبیه سازی" می‌رویم.

#### -۴- ما در این سوال به دنبال محاسبه p-value هستیم :

$$p\_value : P(\hat{p} = 0.6 \text{ or more extreme} | H_0 \text{ true}) \rightarrow P(\hat{p} = 0.6 \text{ or more extreme} | p = 0.5)$$

و می‌دانیم که باید این کار را با شبیه سازی انجام دهیم.

-۵- برای انجام شبیه سازی در این سوال، از یک سکه استفاده می‌کنم که اگر شیر بیاورد نشان دهنده موفقیت، و اگر خط بیاورد نشان دهنده شکست است و این را با استفاده از تابع funcc در R پیاده سازی کرم. در واقع هر بار که این تابع صدای زده می‌شود یک مقدار ۱ یا صفر را به صورت کاملاً تصادفی بر می‌گرداند. سپس با استفاده از حلقه for، ۱۰۰ بار شبیه سازی انجام می‌دهم که در هر بار (در هر simulation) ۱۵ دفعه این تابع صدای زده می‌شود و نتایج آن

در وکتور my\_vector ریخته می شود. سپس با استفاده از تابع table() اقدام به بدست آوردن نسبت موفقیت در این simulation می کنیم. در نهایت نتایج را در یک وکتور به نام res ریخته و آن ها را در یک نمودار Dotplot که با استفاده از کتابخانه ggplot رسم شده است، نمایش می دهم.

قطعه کد این فریم ورک به همراه توضیح کامنت گذاری شده در R در شکل زیر قابل مشاهده است:

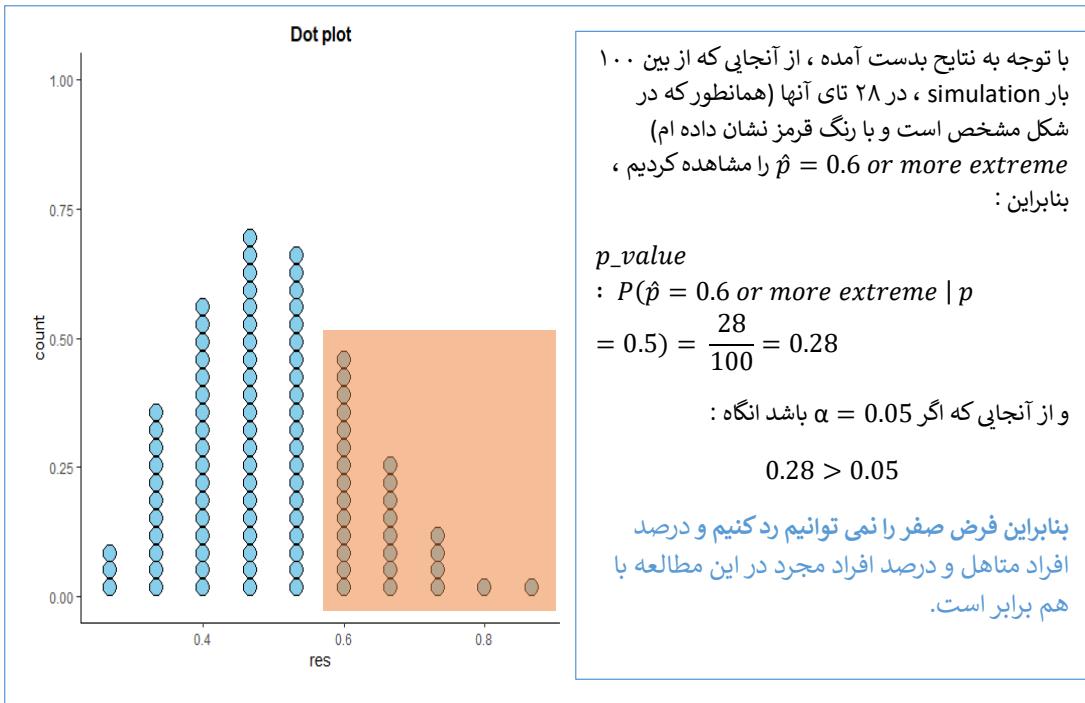
```
# It is a fair coin that returns a value of 1 or 0 whenever it is called
# value of 1 = success
func<- function(){
  rand <- as.integer(runif(1, min=0, max=2))
  return(rand)
}
# define a null vector for final result
res <- c()

# Perform 100 simulations
# in each simulation : i flip the coin 15 times
# and record the proportion of success
# and recording the proportion of success at each iteration
# in res vector
for (y in c(1:100)) {
  my_vector <- c()
  for (z in c(1:15)) {
    my_vector <- c(my_vector,func())
  }
  p_hat_sim <- (as.numeric(table(my_vector)[2]))/15
  res <- c(res,p_hat_sim)
}

# create a data frame for all results
dff <- data.frame(res)

library(ggplot2)
#I use dotplot to display the results of all simulations
ggplot(dff, aes(x = res)) + geom_dotplot(fill="skyblue") +
  labs(title="Dot plot") +           # the title of the plot
  theme_bw() +
  # the grid and background removed from plot,
  theme(panel.border = element_blank(), panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(), axis.line = element_line(colour = "black"),
        # put the title location in the center of the plot.
        plot.title = element_text(size=12,face="bold",hjust = 0.5))
```

### نمودار Dotplot



## : 3. Part A

متغیرهای Dataset در یک Categorical تعداد محدودی دسته (category) را می‌توانند اختیار کنند. در این سوال، متغیر "انتخابی" است که برای هر فرد (Case) در این Dataset وضعیت استعمال دخانیات آن را با یکی از "Unknown", "formerly smoked", "never smoked", "smokes" : category های مشخص می‌کند، (لازم به ذکر است که وجود داشتن مقدار Unknown برای یک فرد (Case) به معنی در دست نبودن اطلاعات برای آن است).  
می‌دانیم که توزیع یک متغیر categorical ، تعداد مقادیری که هر دسته از این متغیر می‌گیرد را لیست می‌کند. ابتدا در R اقدام به محاسبه توزیع متغیر categorical این سوال می‌کنیم.

```
# find the distribution of my categorical variable.
real_dis <- table(HealthCare$smoking_status)/nrow(HealthCare) * 100
```

	formerly smoked	never smoked	smokes	unknown
	17.31898	37.02544	15.44031	30.21526

حال برای گرفتن sample با سایز ۱۰۰ از مجموعه داده ، یکباره صورت کاملاً تصادفی و بدون جایگذاری این کار را انجام می‌دهیم.

```
# randomly select a sample (n=100) without replacement from HealthCare
my_sample_3 <- HealthCare[sample(nrow(HealthCare), size = 100, replace = FALSE), ]
```

همانطور که گفته شد این نمونه کاملاً تصادفی و بدون جایگذاری است ، و خوبی از جامعه هدف (مجموعه داده) است.

اما برای اینکه یک نمونه دیگر با سایز ۱۰۰ از مجموعه داده بگیریم که به عمد با Bias همراه باشد ، به جای اینکه نمونه را از کل مجموعه داده به صورت تصادفی بگیریم ، به عمد از چند سطر مخصوص و با جایگذاری گرفتیم.

```
# select a Biased sample (n=100) with replacement from HealthCare
my_sample_3_Bias <- HealthCare[sample(HealthCare[3,10,], size = 100, replace = TRUE), ]
```

حال توزیع هر دو نمونه گرفته شده را محاسبه می‌کیم. توزیع نمونه ای که Bias ندارد را first\_dis و توزیع نمونه ای که با Bias همراه است را sec\_dis نام‌گذاری می‌کنیم.

```
> # find the distribution of my Non-Biased sample.
> first_dis <- table(my_sample_3$smoking_status)
> first_dis
formerly smoked    never smoked      smokes      unknown
           17            34            16            33
> # find the distribution of my Biased sample.
> sec_dis <- table(my_sample_3_Bias$smoking_status)
> sec_dis
formerly smoked    never smoked      smokes      unknown
           22            33            23            22
```

خب حال می خواهیم بررسی کنیم که آیا هر کدام از این دو توزیع ، با توزیع اصلی برابر هستند یا که خیر از نظر آماری اختلاف قابل توجهی بینشان وجود دارد. برای این کار از تست goodness of fit استفاده می کنیم.

### ۱. مقایسه توزیع first\_dis با توزیع اصلی :

طبق توزیع اصلی و با توجه به توزیع مشاهده شده از نمونه (observed) در ابتدا برای اینکه هر دو توزیع هم جنس باشند ، اقدام به محاسبه توزیع expected می کنیم :

```
# to find the expected #
Expected <- real_dis
as.integer(Expected)
```

```
> Expected <- real_dis
> as.integer(Expected)
[1] 17 37 15 30
```

\*\* نکته : می دانیم که باید حتما مجموع این مقادیر برابر با ۱۰۰ (که سایز sample) است بشود. اما از آنجایی که مجموع این مقادیر برابر با ۹۹ شده است ، باید به یکی از دسته ها یک واحد اضافه کنیم.

بنابراین جدول ما بدین صورت خواهد شد:

Smoking status	Formerly smoked	Never smoked	smokes	unknown	total
Expected #	17	37	15	31	100
Observed #	17	34	16	33	100

در واقع در این آزمون فرض می خواهیم ببینیم این دو توزیع چه قدر با هم اختلاف (فاصله) دارند. در ابتدا فرض صفر و فرض جایگزین را مشخص کرده ، سپس آماره آزمون ( $\chi^2$ ) را محاسبه می کنیم و بر اساس آن به همراه درجه آزادی این تست ، اقدام به محاسبه p-value و در نهایت تصمیم گیری برای این تست انجام می دهیم.

$H_0$  : دو توزیع یکسان هستند و اگر اختلافی وجود دارد به خاطر randomness در انتخاب آن ۱۰۰ نفر است. در واقع توزیع وضعیت سیگار کشیدن افراد مجموعه داده HealtCare با توزیع وضعیت سیگار کشیدن افرادی که به صورت تصادفی و بدون بایاس از این مجموعه داده نمونه گیری شده اند یکسان است.

$H_A$  : دو توزیع یکسان نیستند و اختلاف معنی داری دارند. در واقع توزیع وضعیت سیگار کشیدن افراد مجموعه داده HealtCare با توزیع وضعیت سیگار کشیدن افرادی که به صورت تصادفی و بدون بایاس از این مجموعه داده نمونه گیری شده اند یکسان نیست.

در ابتدا لازم است که شرایط تست chi-square بررسی شوند :

شرط independence : از انجایی که افراد در هر دو sample با سایز ۱۰۰ ، به صورت تصادفی انتخاب شده اند و وضعیت سیگار کشیدن آن ها مستقل از هم هستند شرط independence را داریم. لازم به ذکر است که هر فرد فقط یک مقدار در این متغیر دارد که وضعیت سیگار کشیدن آن فقط در یک cell از جدول دریک آمده است و همچنین از آنجایی که سایز هر دو نمونه (n=100) از 10 درصد جامعه کمتر است، می توان گفت که شرط استقلال برقرار است.

شرط sample size/skew : در هر cell جدول حداقل ۵ expected case وجود دارد. بنابراین این شرط هم برقرار است.

آماره آزمون ( $\chi^2$ ) بدین صورت محاسبه می شود :

$$\chi^2 \text{ statistic} = \sum_{i=1}^k \frac{(observed - expected)^2}{expected} = 0.43$$

همچنین درجه آزادی این تست بدین صورت است :

$$df = k - 1 = 4 - 1 = 3$$

*k : number of cells*

از آنجایی کهتابع ( chisq.test() ) برای انجام آزمون goodness of fit هم استفاده می شود، در اینجا از این تابع برای انجام این تست و محاسبه p-value استفاده می کنیم.

```
> # Goodness of fit test
> chisq.test(x=as.integer(first_dis), p=as.integer(Expected)/100)

chi-squared test for given probabilities

data: as.integer(first_dis)
X-squared = 0.43894, df = 3, p-value = 0.9321 ←
```

اگر به مقدار دقیق به دست آمده برای p-value نگاه کنیم میبینیم که بسیار بزرگ است پس می توانیم نتیجه بگیریم که فرض صفر رد نمی شود.

تفسیر: توزیع وضعیت سیگار کشیدن افراد مجموعه داده HealtCare با توزیع وضعیت سیگار کشیدن افرادی که به صورت تصادفی و بدون بایاس از این مجموعه داده نمونه گیری شده اند یکسان است.

## ۲. مقایسه توزیع sec\_dis با توزیع اصلی:

طبق توزیع اصلی و با توجه به توزیع مشاهده شده از نمونه (observed) در ابتدا برای اینکه هر دو توزیع هم جنس باشند، اقدام به محاسبه توزیع expected می کنیم :

<pre># to find the expected # Expected &lt;- real_dis as.integer(Expected)</pre>	<pre>&gt; Expected &lt;- real_dis &gt; as.integer(Expected) [1] 17 37 15 30</pre>
--	---

\*\* نکته: می دانیم که باید حتما مجموع این مقادیر برابر با ۱۰۰ است بشود. اما از آنجایی که مجموع این مقادیر برابر با ۹۹ شده است، باید به یکی از دسته ها یک واحد اضافه کنیم.

بنابراین جدول ما بدین صورت خواهد شد:

Smoking status	Formerly smoked	Never smoked	smokes	unknown	total
Expected #	17	37	15	31	100
Observed #	22	33	23	22	100

$H_0$  : دو توزیع یکسان هستند و اگر اختلاف وجود دارد به خاطر randomness در انتخاب آن ۱۰۰ نفر است. در واقع توزیع وضعیت سیگار کشیدن افراد مجموعه داده HealtCare با توزیع وضعیت سیگار کشیدن افرادی که در یک نمونه همراه با بایاس از مجموعه داده قرار دارند، یکسان است.

$H_A$  : دو توزیع یکسان نیستند و اختلاف معنی داری دارند. در واقع توزیع وضعیت سیگار کشیدن افراد مجموعه داده HealtCare با توزیع وضعیت سیگار کشیدن افرادی که در یک نمونه همراه با بایاس از مجموعه داده قرار دارند، یکسان نیست.

در ابتدا لازم است که شرایط تست chi-square بررسی شوند که در قسمت قبل برقرار بودند.

آماره آزمون  $\chi^2$  بدين صورت محاسبه می شود :

$$\chi^2 statistic = \sum_{i=1}^k \frac{(observed - expected)^2}{expected} = 8.7$$

همچنین درجه آزادی اين تست بدين صورت است :

$$df = k - 1 = 4 - 1 = 3$$

*k : number of cells*

از آنجايي که تابع `chisq.test()` ، برای انجام آزمون goodness of fit هم استفاده می شود، در اينجا از اين تابع برای انجام اين تست و محاسبه p-value استفاده می کنيم.

```
> # Goodness of fit test #2  
> chisq.test(x=as.integer(sec_dis), p=as.integer(Expected)/100)  
  
chi-squared test for given probabilities  
  
data: as.integer(sec_dis)  
X-squared = 8.7826, df = 3, p-value = 0.03233 ←
```

اگر به مقدار دقیق به دست آمده برای p-value نگاه کنیم میبینیم که اگر  $\alpha = 0.05$  در نظر بگیریم ، مقدار از آلفا کوچکتر است پس می توانیم نتیجه بگیریم که فرض صفر را رد می کنیم.

تفسیر: دو توزيع يكسان نیستند و اختلاف معنی داري دارند. در واقع توزيع وضعیت سیگار کشیدن افراد مجموعه داده HealtCare با توزيع وضعیت سیگار کشیدن افرادی که در يك نمونه همراه با بایاس از مجموعه داده قرار دارند، يكسان نیست.

### 3. Part B

در قسمت قبل از متغير "smoking\_status" استفاده کردیم که برای هر فرد (Case) در این Dataset وضعیت استعمال دخانیات آن را مشخص می کند. حال در این قسمت متغير دوم را ، "ever\_married" انتخاب می کنیم که وضعیت متاهل بودن یا مجرد بودن observation های ما را در این مجموعه داده نشان می دهد. برای هر case، مقدار yes نشان دهنده متاهل بودن آن فرد، و مقدار no نشان دهنده مجرد بودن آن فرد است. در این سوال می خواهیم بدانیم که این دو متغير مستقل از هم اند یا خیر.

بنابراین در ابتدا برای انجام تست استقلال ، فرض صفر وفرض جایگزین را مشخص می کنیم، سپس پس از بررسی شرایط این تست، به محاسبه توزيع expected count (cell) می پردازیم و آن ها را درون جدول نمایش می دهیم، آماره آزمون  $\chi^2$  (chi-square) را محاسبه می کنیم و بر اساس آن به همراه درجه آزادی این تست، اقدام به محاسبه p-value و در نهایت تصمیم گیری برای این تست می پردازیم.

$H_0$  : دو متغير "ever\_married" و "smoking\_status" مستقل هستند.

$H_A$  : دو متغير "ever\_married" و "smoking\_status" به هم وابسته هستند.

بهتر است که برای انجام اين تست ، در ابتدا مجموعه داده را جامعه هدف بدانیم و يك نمونه با سایز 300 از آن انتخاب کنیم.

(دستور ایجاد sample در R)

```
# randomly select a sample (n=300) without replacement from Healthcare  
my_sample_3_B <- Healthcare[sample(nrow(Healthcare), size = 300, replace = FALSE), ]
```

شرط independence : از آنجایی که نمونه های داخل sample ما به صورت تصادفی انتخاب شده اند، یعنی random رعایت شده است. همچنین به دلیل اینکه شرط  $n < 10\% \text{ population}$  (300) برقرار است. همینطور از آنجایی که هر case فقط در یک cell از جدول مقدار دارد و observation ها مستقل از هم اند. بنابراین شرط استقلال برقرار است.

شرط sample size/skew : برای برقار بودن این شرط در هر cell باید، حداقل 5 expected case وجود داشته باشد.

بنابراین برای بررسی شرط دوم ، لازم داریم تا expected count هر cell را محاسبه کنیم :

برای محاسبه expected count برای هر سطر ، می گوییم اگر دو متغیر مستقل از هم باشند، انتظار داریم که در هر خانه جدول چه مقداری قرار بگیرد و برای محاسبه این مقدار از رابطه زیر استفاده می کنیم :

$$\text{Expected count} = \frac{(\text{row total}) \times (\text{column total})}{\text{table total}}$$

اما در ابتدا با استفاده از R و تابع () فرکانس تکرار را بدست می آوریم.

```
> # to understand the frequency
> table_3_B <- table(my_sample_3_B$ever_married, my_sample_3_B$smoking_status)
> table_3_B

      formerly smoked never smoked smokes unknown
No           9          34       10      54
Yes          40         72       38      43
```

حال طبق رابطه ذکر شده برای محاسبه expected count ، به محاسبه این مقدار برای هر cell از جدول می کنیم :

```
> # to understand the frequency
> table_3_B <- table(my_sample_3_B$ever_married, my_sample_3_B$smoking_status)
> table_3_B_B <- addmargins(table_3_B)
> # to calculate the expected count of each cell
> chi <- chisq.test(table_3_B_B)
> expcount[] <- paste(table_3_B_B,paste0("(",round(chi$expected),")"))
> expcount

      formerly smoked never smoked smokes unknown sum
No   9 (17)          34 (38)        10 (17)  54 (35) 107
Yes 40 (32)          72 (68)        38 (31)  43 (62) 193
Sum 49              106            48          97      300
```

همانطور که گفته شد برای برقار بودن این شرط در هر cell باید، حداقل 5 expected case وجود داشته باشد. که با توجه به جدول بالا این شرط برقار است.

حال برای انجام تست استقلال از آنجایی که در این سوال هدف استفاده از R برای انجام این تست است از تابع () chisq.test() برای این کار استفاده می کنیم و سپس با  $\alpha = 0.05$  نتایج را تحلیل می کنیم.

بنابراین در R این تابع را اجرا می کنیم :

```
> # independence test using the chisq.test() function
> chisq.test(table(my_sample_3_B$ever_married, my_sample_3_B$smoking_status))

Pearson's Chi-squared test

data: table(my_sample_3_B$ever_married, my_sample_3_B$smoking_status)
X-squared = 28.505, df = 3, p-value = 2.846e-06
```

از آنجایی که مقدار p-value بدست آمده خیلی کوچکتر از 0.05 است، ما فرض صفر را رد می کنیم. بنابراین نتیجه می گیریم دو متغیر "ever\_married" و "smoking\_status" به هم وابسته هستند.

برای این سوال ، متغیر عددی "health\_bills" را به عنوان متغیر Response انتخاب کرده ام که میزان هزینه سالیانه ای را که هر فرد درون این مجموعه داده برای سلامتی خود می پردازد را نشان می دهد. همچنین دو متغیر "age" و "avg\_glucose\_level" که به ترتیب نشان دهنده سن هر فرد ، و متوسط قند خون (سطح گلوكز) برای هر فرد هستند را به عنوان متغیرهای Explanatory انتخاب کرده ام.

نکته ۱ : طبق آن چیزی که در فاز اول این پروژه عنوان شد ، متغیر health\_bills دارای مقادیر ( N/A values ) میباشد و رویکرد من برای مقادیر گمشده در این متغیر ، استفاده از روش جایگذاری میانه به جای مقادیر گم شده است. بنابراین در ابتدا این عمل را روی مجموعه داده HealthCare در R با استفاده از دستور زیر انجام می دهیم. در واقع هر سطر از مجموعه داده HealthCare که برای متغیر ذکر شده مقدار گم شده داشته باشد ، برای آن سطر میانه مقادیری که آن متغیر دارد را جایگذاری میکنم.

```
# at first lets replace missing values with the mean
HealthCare2 <- HealthCare
HealthCare2$health_bills[is.na(HealthCare2$health_bills)]<-median(HealthCare2$health_bills,na.rm=TRUE)
```

نکته ۲ : از آنجاکه درون این مجموعه داده حدود 5000 تا داده وجود دارد ، در این سوال فرض را بر این گذاشته ام که جامعه هدف همان مجموعه داده است و بنابراین در ابتدا یک نمونه با سایز 100 به صورت کاملاً تصادفی و بدون جایگذاری می گیرم و از همین نمونه برای پاسخ دادن به این سوال استفاده می کنم.

(دستور ایجاد sample در R)

```
# randomly select a sample (n=100) without replacement from HealthCare
my_sample_4 <- HealthCare2[sample(nrow(HealthCare2), size = 100, replace = FALSE),]
```

## : 4.Part A

در ابتدا لازم است که عنوان کنم در علم پژوهشی ، فشار خون بالا غالباً در کنار دیابت رخ می دهد و مطالعات نشان می دهد که ممکن است رابطه ای بین آنها وجود داشته باشد. دیابت در واقع عبارت است از بالا بودن مقادیری از قند در خون یک فرد (سطح گلوكز). افرادی که قند خون بالای دارند ، خطر بیشتری در ابتلا بیماری های قلبی را دارند و در نتیجه به نظر میرسد که در سال هزینه های بیشتری را هم باید برای سلامتی خود پردازند. همینطور غالباً با افزایش سن یک فرد احتمال مبتلا شدن به بیماری هایی مثل قند خون بالا و در نتیجه بیماری های قلبی بیشتر می شود (البته صرفاً یک احتمال است). بنابراین به نظر میرسد که رابطه ای بین سن و هزینه های سالیانه ای که یک فرد برای سلامتی خود می پردازد ، و همچنین رابطه ای بین متوسط قند خون یک فرد با این هزینه ها وجود دارد.

بنابراین با توجه به توضیحات داده شده ، بنده حدس میزنم که متغیر "avg\_glucose\_level" که متوسط قند خون (سطح گلوكز) برای هر فرد درون این مجموعه داده را مشخص می کند Predictor مهم و بهتری برای پیش بینی متغیر Response ما در این سوال است.

## : 4.Part B

: a

در این قسمت می خواهیم برای هر دو متغیر Explanatory که انتخاب کردیم ، یک مدل خطی بسازیم. ما در مفهوم رگرسیون ، بر اساس متغیر Explanatory می خواهیم رابطه بین دو متغیر را با یک مدل خطی ، مدل کنیم. در واقع با استفاده از خط رگرسیونی یک تخمین برای متغیر Response می نزیم.

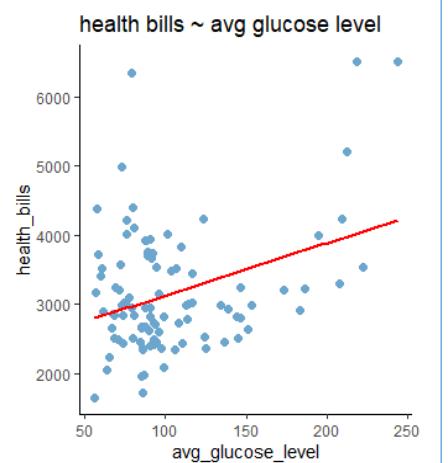
- ۱- ساخت مدل Regression که بر اساس متوسط قند خون (سطح گلوکز) هر فرد (بر اساس متغیر "avg\_glucose\_level") ، هزینه های سالیانه ای که یک فرد برای سلامتی خود می پردازد را پیش بینی می کند :

در ابتدا برای داشتن یک مدل Regression باید شرایط لازم برقرار باشند :

**شرط Linearity :** طبق این شرط رابطه بین دو متغیر "health\_bills" و "avg\_glucose\_level" باید خطی باشد. برای بررسی این شرط اقدام به رسم یک نمودار scatterplot در R به کمک کتابخانه ggplot می کنیم.

```
# checking the linearity condition :
library(ggplot2)
# plot a scatter plot
ggplot(data = my_sample_4,aes(x=avg_glucose_level,y=health_bills)) +
  geom_point(aes(x = avg_glucose_level, y = health_bills),
             color = "skyblue3", size = 2) +
  labs(title = "health bills ~ avg glucose level") +
  # put the title location in the center of the plot,
  theme(plot.title = element_text(size=12,face="bold",hjust = 0.5)) +
  theme_classic() +
  # for create a linear model
  geom_smooth(method=lm,se=FALSE,color="red")
```

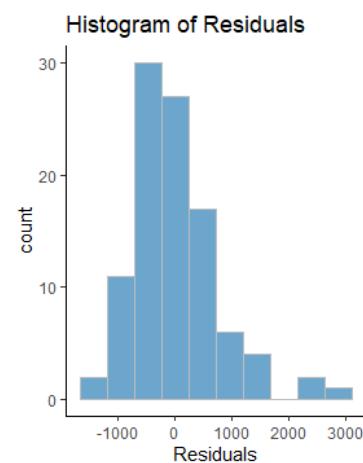
با توجه به این نمودار ، رابطه بین دو متغیر "health\_bills" و "avg\_glucose\_level" تقریبا خطی است. بنابراین شرط Linearity را برقرار می دانیم.



**شرط Nearly normal Residuals :** طبق این شرط ، توزیع نقاط Residual ها (فاصله نقاط اصلی با نقاط تخمین) باید نرمال باشد. هرچه این توزیع نرمال تر باشد ، مدل Regression ما بهتر است.

```
# checking the Nearly normal residuals condition
library(ggplot2)
#create histogram of residuals
ggplot(data = my_sample_4, aes(x = modelnum1$residuals)) +
  geom_histogram(bins = 10,
                 fill = 'skyblue3',
                 color = 'gray') +
  labs(title = 'Histogram of Residuals', x = 'Residuals') +
  theme_classic()
```

با توجه به این نمودار ، میتوان مشاهده کرد که توزیع چوله به راست است اما با فرض نرمال بودن این توزیع ، این شرط را برقرار میدانیم.



**شرط Constant Variability :** طبق این شرط که به homoscedasticity معروف است ، باید نقاط دو متغیر یک پراکندگی ثابت ، اطراف خط Regression داشته باشند. که فرض می کنیم این شرط هم برقرار است.

بنابراین حال می توانیم اقدام به ساخت مدل Regression بکنیم.

برای ساخت مدل ، از تابع lm() در R استفاده می کنیم :

```
> modelnum1 <- summary(lm(health_bills ~ avg_glucose_level,data=my_sample_4))
> modelnum1
call:
lm(formula = health_bills ~ avg_glucose_level, data = my_sample_4)

Residuals:
    Min      1Q  Median      3Q     Max 
-1319.2 -570.5 -111.8  365.1 2982.6 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2661.256   210.820 12.623 <2e-16 ***
avg_glucose_level 4.657     1.729  2.694  0.0083 **  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

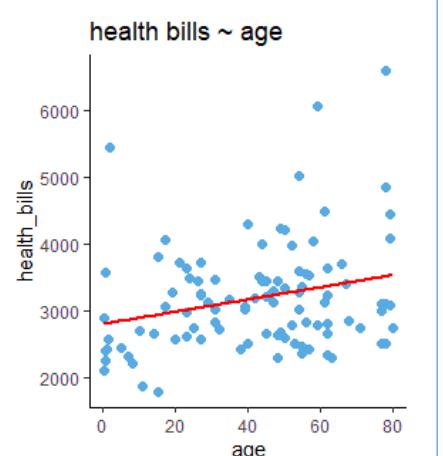
Residual standard error: 789.3 on 98 degrees of freedom
Multiple R-squared:  0.06896, Adjusted R-squared:  0.05946 
F-statistic: 7.259 on 1 and 98 DF,  p-value: 0.0083
```

-۲ ساخت مدل Regression که بر اساس سن هر فرد (بر اساس متغیر "age") ، هزینه های سالیانه ای که یک فرد برای سلامت خود می پردازد را پیش بینی می کند :

در ابتدا برای داشتن یک مدل Regression باید شرایط لازم برقرار باشند :  
**شرط Linearity** : طبق این شرط رابطه بین دو متغیر "health\_bills" و "age" خطی باشد. برای بررسی این شرط اقدام به رسم یک نمودار scatterplot در R به کمک کتابخانه ggplot می کنیم.

```
# checking the linearity condition :
library(ggplot2)
# plot a scatter plot
ggplot(data = my_sample_4,aes(x=age,y=health_bills)) +
  geom_point(aes(x = age, y = health_bills),
             color = "skyblue3", size = 2) +
  labs(title = "health bills ~ age") +
  # put the title location in the center of the plot.
  theme(plot.title = element_text(size=12,face="bold",hjust = 0.5)) +
  theme_classic()
# for create a linear model
geom_smooth(method=lm,se=FALSE,color="red")
```

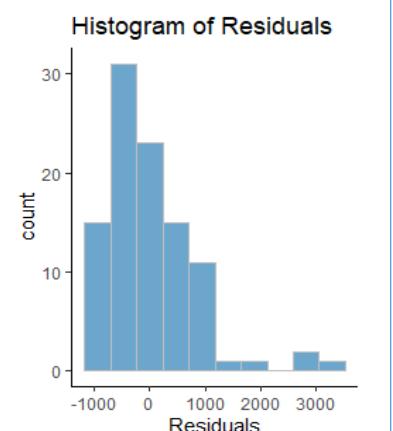
با توجه به این نمودار ، رابطه بین دو متغیر "age" و "health\_bills" خطی است. بنابراین شرط Linearity را برقرار می دانیم.



**شرط Nearly normal Residuals** : طبق این شرط ، توزیع Residual ها (فاصله نقاط اصلی با نقاط تخمین) باید نرمال باشد. هرچه این توزیع نرمال تر باشد ، مدل Regression ما بهتر است.

```
# checking the Nearly normal residuals condition
library(ggplot2)
#create histogram of residuals
ggplot(data = my_sample_4, aes(x = modelnum2$residuals)) +
  geom_histogram(bins = 10,
                 fill = 'skyblue3',
                 color = 'gray') +
  labs(title = 'Histogram of Residuals', x = 'Residuals') +
  theme_classic()
```

با توجه به این نمودار ، میتوان مشاهده کرد که توزیع چوله به راست است اما با فرض نرمال بودن این توزیع ، این شرط را برقرار میدانیم.



شرط **Constant Variability** : طبق این شرط که به homoscedasticity معروف است ، باید نقاط دو متغیر یک پراکندگی ثابت ، اطراف خط Regression داشته باشند. که فرض می کنیم این شرط هم برقرار است.

بنابراین حال می توانیم اقدام به ساخت مدل Regression بکنیم.  
برای ساخت مدل ، ازتابع (`lm`) در R استفاده می کنیم :

```
> modelnum2 <- summary(lm(health_bills ~ age,data=my_sample_4)) ← کد
> modelnum2

Call:
lm(formula = health_bills ~ age, data = my_sample_4)

Residuals:
    Min      1Q  Median      3Q     Max 
-1148.8 -550.8 -102.9  296.8 3077.6 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2799.532   167.037   16.76 < 2e-16 ***
age          9.182     3.479    2.64  0.00966 **  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 790.4 on 98 degrees of freedom
Multiple R-squared:  0.06638, Adjusted R-squared:  0.05685 
F-statistic: 6.967 on 1 and 98 DF,  p-value: 0.009659
```

نتیجه :

: b

۱- بنابراین **Linear model** ما برای پیش بینی هزینه های سالیانه سلامتی (بر اساس متغیر متوسط سطح قند خون) برابر است با :

$$\widehat{\text{health\_bills}} = 2661.256 + 4.657 \times \text{avg\_glucose\_level}$$

بنابراین مدل Regression ای ساختیم که بر اساس متوسط قند خون هر فرد درون این مجموعه داده و با استفاده از این مدل که شیب آن برابر با 4.657 و عرض از مبدأ آن برابر با 2661.256 است ، بتوانیم هزینه های سالیانه سلامتی را برای آن فرد پیش بینی کنیم و تخمین بزنیم.

تفسیر شیب خط رگرسیون و عرض از مبدأ خط رگرسیون که بدست آوردهیم :

تفسیر **Slope** : از آنجایی که شیب خط بدست آمده مقداری مثبت دارد، بنابراین به ازای هر یک واحد افزایش در متوسط قند خون (سطح گلوکز) یک فرد ، انتظار داریم که هزینه های سالیانه ای که آن فرد برای سلامتی خود باید پرداخت کند به میزان ۴/۶۵۷ واحد افزایش پیدا کند.

تفسیر **intercept** : می دانیم که عرض از مبدأ ، در واقع جایی است که خط رگرسیون ، محور y را قطع می کند. (جایی که  $x=0$ ). و این بدین معنی است که انتظار داریم زمانی که متوسط قند خون یک فرد برابر با صفر باشد ! ، میزان هزینه ای که در سال برای سلامتی خود باید پرداخت کند برابر با 2661.256 میباشد !

\*\* عرض از مبدأ یک وضعیت فرضی است که خیلی معنی ندارد. همینطور که مشاهده می شود در این سوال ، تفسیر عرض از مبدأ بی معنی است. \*\*

۲- همچنین **Linear model** ما برای پیش بینی هزینه های سالیانه سلامتی (بر اساس متغیر سن) برابر است با :

$$\widehat{\text{health\_bills}} = 2799.532 + 9.182 \times \text{age}$$

بنابراین مدل Regression ای ساختیم که بر اساس سن هر فرد درون این مجموعه داده و با استفاده از این مدل که شیب آن برابر با 9.182 و عرض از مبدأ آن برابر با 2799.532 است ، بتوانیم هزینه های سالیانه سلامتی را برای آن فرد پیش بینی کنیم و تخمین بزنیم.

## تفسیر شیب خط رگرسیون و عرض از مبدأ خط رگرسیون که بدست آوردهیم :

**تفسیر Slope :** از آنجایی که شیب خط بدست آمده مقداری مثبت دارد، بنابراین به ازای هر یک واحد افزایش در سن یک فرد، انتظار داریم که هزینه های سالیانه ای که آن فرد برای سلامتی خود باید پرداخت کند به میزان ۹/۱۸۲ واحد افزایش پیدا کند.

**تفسیر intercept :** می دانیم که عرض از مبدأ، در واقع جایی است که خط رگرسیون، محور  $y$  را قطع می کند. (جایی که  $x=0$ ). و این بدین معنی است که انتظار داریم زمانی که سن یک فرد برابر با صفر باشد! (بعنی فرد هنوز متولد نشده است)، میزان هزینه ای که در سال برای سلامتی خود باید پرداخت کند برابر با 2799.532 میباشد!

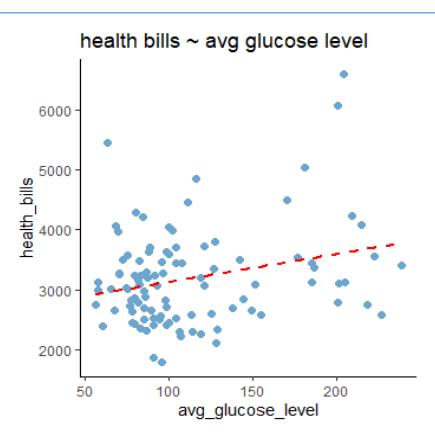
\*\* عرض از مبدأ یک وضعیت فرضی است که خیلی معنی ندارد. همینطور که مشاهده می شود در این سوال، تفسیر عرض از مبدأ بی معنی است. \*\*

: C

در این قسمت می خواهیم یک نمودار scatterplot برای نمایش مدلی که با استفاده از هر دو متغیر avg\_glucose\_level و age برای تخمین متغیر Response فیت کردیم ایجاد کنیم و مدل خطی را درون نمودار با استفاده یک از خط dashed نمایش بدهیم. برای این کار در R از کتابخانه ggplot استفاده می کنیم.

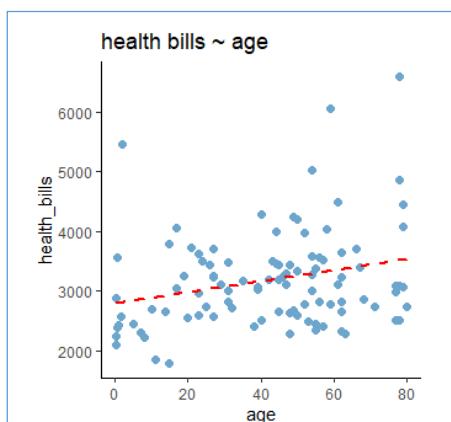
نمودار health bills ~ avg glucose level برای Scatterplot

```
library(ggplot2)
# plot a scatter plot
ggplot(data = my_sample_4,aes(x=avg_glucose_level,y=health_bills)) +
  geom_point(aes(x = avg_glucose_level, y = health_bills),
             color = "skyblue3", size = 2) +
  labs(title = "health bills ~ avg glucose level") +
  # put the title location in the center of the plot.
  theme(plot.title = element_text(size=12,face="bold",hjust = 0.5)) +
  theme_classic() +
  # for create a linear model with dashed Line
  geom_smooth(method=lm,se=FALSE,color="red",lty=2,size=1.5)
```



نمودار health bills ~ age برای Scatterplot

```
library(ggplot2)
# plot a scatter plot
ggplot(data = my_sample_4,aes(x=age,y=health_bills)) +
  geom_point(aes(x = age, y = health_bills),
             color = "skyblue3", size = 2) +
  labs(title = "health bills ~ age") +
  # put the title location in the center of the plot.
  theme(plot.title = element_text(size=12,face="bold",hjust = 0.5)) +
  theme_classic() +
  # for create a linear model with dashed Line
  geom_smooth(method=lm,se=FALSE,color="red",lty=2,size=1)
```



## 4. Part C

می دانیم که برای بررسی کردن اینکه آیا predictor های ما significant می شوند یا خیر می توانیم از آزمون فرض (برای کل مدل یا برای slop) استفاده کنیم. می دانیم که اگر شیب خط رگرسیون ما صفر بشود یعنی آن متغیر Explanatory که انتخاب کردیم، predictor خوبی نیست اما اگر از صفر فاصله بگیرد predictor خوبی است. بنابراین در انجام آزمون فرض برای slop ما به دنبال این هستیم که آیا شیب خط رگرسیون صفر می شود یا مخالف با صفر است. در واقع فرض صفر و فرض جایگزین در این آزمون فرض را این گونه تعریف می کنیم:

$H_0 : \beta_1 = 0$  متغير Explanatory در پیش بینی متغیر Response، متغیر خوبی نیست.

$H_A : \beta_1 \neq 0$  متغير Explanatory در پیش بینی متغیر Response، متغیر خوبی است.

سپس آماره آزمون Z و p-value را حساب می کنیم. اما از آنجا که در این قسمت باید از اطلاعاتی که در قسمت قبل بدست آورده استفاده کنیم بنابراین از مقدار p-value که توسط خود مدل محاسبه شده است استفاده می کنیم و در نهایت به بررسی significant بودن یا نبودن آن متغیر با  $a = 0.05$  می پردازیم :

از آنجا که در قسمت قبل برای هر متغیر Explanatory یک مدل ایجاد کردیم بنابراین به ترتیب به بررسی significant بودن یا نبودن آن متغیر می پردازیم.

۱- مدل Regression که بر اساس متوسط قند خون (سطح گلوکز) هر فرد ، هزینه های سالیانه ای که یک فرد برای سلامتی خود می پردازد را پیش بینی می کند :

```
> modelnum1 <- summary(lm(health_bills ~ avg_glucose_level, data=my_sample_4))
> modelnum1

call:
lm(formula = health_bills ~ avg_glucose_level, data = my_sample_4)

Residuals:
    Min      1Q  Median      3Q     Max 
-1319.2 -570.5 -111.8  365.1 2982.6 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2661.256   210.820 12.623 <2e-16 ***
avg_glucose_level 4.657     1.729  2.694  0.0083 **  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 789.3 on 98 degrees of freedom
Multiple R-squared:  0.06896,  Adjusted R-squared:  0.05946 
F-statistic: 7.259 on 1 and 98 DF,  p-value: 0.0083
```

برای سهولت بررسی ، مقدار p-value را با رنگ قرمز مشخص کرده ام. از آنجا که از مقدار آلفا (0.05) کمتر شده است ، پس متغیر avg\_glucose\_level پیش بینی متغیر Response معنی دار یک Predictor معنی دار (significant) است.

۲- مدل Regression که بر اساس سن هر فرد (بر اساس متغیر "age") ، هزینه های سالیانه ای که یک فرد برای سلامتی خود می پردازد را پیش بینی می کند :

```
> modelnum2 <- summary(lm(health_bills ~ age, data=my_sample_4))
> modelnum2

call:
lm(formula = health_bills ~ age, data = my_sample_4)

Residuals:
    Min      1Q  Median      3Q     Max 
-1148.8 -550.8 -102.9  296.8 3077.6 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2799.532   167.037 16.76 <2e-16 ***
age          9.182     3.479  2.64  0.00966 **  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 790.4 on 98 degrees of freedom
Multiple R-squared:  0.06638,  Adjusted R-squared:  0.05685 
F-statistic: 6.967 on 1 and 98 DF,  p-value: 0.009659
```

از آنجا که p-value از مقدار آلفا (0.05) کمتر شده است ، پس متغیر age برای پیش بینی متغیر Response یک Predictor معنی دار (significant) است.

\*\* نکته : پس هر دو متغیر کرده ام significant شدند و متغیر های Explanatory خوبی هستند ، اما از آنجا که مقدار p-value از مقدار avg\_glucose\_level متغیر p-value کمتر شده است بنابراین میتوان گفت از بین این دو متغیر Explanatory که انتخاب کرده ام ، متغیر avg\_glucose\_level معنی دار تری است (significat). همچنین میدانیم اگر مدلی Adjusted R-square بالاتری داشته باشد مدل بهتری تلقی می شود. بنابراین از آنجا که مدل ساخته شده با متغیر avg\_glucose\_level ، مقدار Adjusted R-square بالاتری دارد بنابراین می توان گفت این متغیر نسبت به متغیر Predictor ، age بهتری است.

می دانیم که برای اینکه بدانیم مدل Regression ای که ایجاد کرده ایم ، مدل خوبی است یا خیر (اصطلاحا متغیر Explanatory خوبی انتخاب کردیم یا خیر) می توانیم از دو روش استفاده کنیم.

۱- معیار  $R^2$

۲- تست ANOVA

ما در این سوال یکبار با استفاده از متغیر avg\_glucose\_level ، و بار دیگر با استفاده از متغیر age اقدام به ساخت مدل Regression برای تخمین health\_bills کردیم.

برای مقایسه این دو مدل با هم دیگر در ابتدا از معیار Adjusted R<sup>2</sup> (Adjusted R-square) استفاده می کنیم.

### ۱- بررسی دو مدل با استفاده از Adjusted R<sup>2</sup>

می دانیم که  $R^2$  معیار برای اندازه گیری قدرت یک Linear model است و می گوید که چه میزان از variability که در متغیر Response دیده می شود توسط مدل توضیح داده می شود. و می دانیم که هر چه  $R^2$  (که مقداری بین ۰ و ۱ دارد) بزرگتر باشد یعنی ما مدل بهتری داریم. اما می دانیم که با اضافه کردن متغیر های الگی این معیار حتماً زیاد می شود و لزوماً مدل پنهانه تری نخواهیم داشت. بنابراین برای حل این مشکل از معیار Adjusted R<sup>2</sup> برای مقایسه دو مدل استفاده می کنیم و میدانیم که مدلی که دارای مقدار Adjusted R<sup>2</sup> بالاتری باشد مدل بهتری است.

```
> modelnum1 <- summary(lm(health_bills ~ avg_glucose_level, data=my_sample_4))
> modelnum1

call:
lm(formula = health_bills ~ avg_glucose_level, data = my_sample_4)

Residuals:
    Min      1Q  Median      3Q     Max 
-1319.2  -570.5 -111.8   365.1  2982.6 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2661.256   210.820 12.623 <2e-16 ***
avg_glucose_level 4.657    1.729   2.694  0.0083 **  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 789.3 on 98 degrees of freedom
Multiple R-squared:  0.06896,  Adjusted R-squared:  0.05946 
F-statistic: 7.259 on 1 and 98 DF,  p-value: 0.0083
```

مدل Regression که بر اساس متغیر قند خون (سطح گلوکز) هر فرد ، هزینه های سالیانه ای که یک فرد برای سلامتی خود می پردازد را پیش بینی می کند :

```
> modelnum2 <- summary(lm(health_bills ~ age, data=my_sample_4))
> modelnum2

call:
lm(formula = health_bills ~ age, data = my_sample_4)

Residuals:
    Min      1Q  Median      3Q     Max 
-1148.8  -550.8 -102.9   296.8  3077.6 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2799.532   167.037 16.76 < 2e-16 ***
age          9.182     3.479   2.64  0.00966 **  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 790.4 on 98 degrees of freedom
Multiple R-squared:  0.06638,  Adjusted R-squared:  0.05685 
F-statistic: 6.967 on 1 and 98 DF,  p-value: 0.009659
```

مدل Regression که بر اساس سن هر فرد (بر اساس متغیر "age" ، هزینه های سالیانه ای که یک فرد برای سلامتی خود می پردازد را پیش بینی می کند :

با توجه به مقدار Adjusted R-square که با رنگ سبز برای هر دو مدل مشخص کردم ، می توان مشاهده کرد از آنجا که مدل ساخته شده با متغیر avg\_glucose\_level Adjusted R-square بالاتری دارد بنابراین می توان گفت این متغیر نسبت به متغیر age ، Predictor بهتری است و بنابراین مدل Regression که بر اساس متغیر قند خون (سطح گلوکز) هر فرد ، هزینه های سالیانه ای که یک فرد برای سلامتی خود می پردازد را پیش بینی می کند ، مدل بهتری است.

## ۲- مقایسه دو مدل با استفاده از تست ANOVA

در ابتدا در R با استفاده از دستور() anova اقدام به انجام این تست برای هر دو مدل ساخته شده در مرحله قبل می کنیم. (نتایج این تست در شکل زیر قابل مشاهده است).

```
> # 1- ANOVA for model number one : health_bills ~ avg_glucose_level
> m1 <- lm(health_bills ~ avg_glucose_level,data=my_sample_4)
> anova(m1)
Analysis of Variance Table

Response: health_bills
          Df  Sum Sq Mean Sq F value Pr(>F)
avg_glucose_level  1 4522375 4522375 7.2592 0.0083 **
Residuals        98 61052559 622985
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

>
> # 2- ANOVA for model number two : health_bills ~ age
> m2 <- lm(health_bills ~ age,data=my_sample_4)
> anova(m2)
Analysis of Variance Table

Response: health_bills
          Df  Sum Sq Mean Sq F value Pr(>F)
age            1 4352618 4352618 6.9673 0.009659 **
Residuals      98 61222316 624718
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

کد مدل اول : ANOVA جدول

کد مدل دوم: ANOVA جدول

همانطور که مشاهده می شود مقدار p-value هر دو تست ANOVA نشان از significant بودن هر دو متغیر انتخابی من در این سوال دارد، پس هر دو متغیر significant که انتخاب کرده ام Explanatory شدند و متغیر های Explanatory خوبی هستند، اما از آنجا که مقدار p-value متغیر avg\_glucose\_level از مقدار p-value متغیر age کمتر شده است بنابراین میتوان گفت از بین این دو متغیر avg\_glucose\_level که انتخاب کرده ام، متغیر معنی دار تری است (significant). همچنین میدانیم اگر مدلی ANOVA بالاتری داشته باشد مدل بهتری تلقی می شود. و ما میتوانیم از روی جدول Adjusted R-square مقدار  $R^2$  را محاسبه کنیم.

$$R^2 = \frac{\text{explained variability}}{\text{total variability}} = \frac{SS_{Reg}}{SS_{total}} = \frac{4522375}{65574934} = 0.068 \quad \text{مدل اول :}$$

$$R^2 = \frac{\text{explained variability}}{\text{total variability}} = \frac{SS_{Reg}}{SS_{total}} = \frac{4352618}{65574934} = 0.066 \quad \text{مدل دوم :}$$

بنابراین از آنجا که مدل ساخته شده با متغیر avg\_glucose\_level (مدل اول)، مقدار Adjusted R-square بالاتری دارد بنابراین این مدل ، مدل بهتری است.

## 4.Part E

در قسمت های قبل مشاهده کردیم که هر چه یک متغیر Explanatory خوب باشد می تواند به ساخت مدل بهتری کمک کند. در واقع یکی از ویژگی های یک Predictor خوب ساخت مدلی برای تخمين متغیر Response که دارای حداقل خطای ممکن باشد است. از طرف دیدیم که هرچه یک متغیر Predictor بهتری باشد می تواند مقدار Adjusted  $R^2$  مدل را افزایش دهد. بنابراین ویژگی های یک predictor خوب عبارتند از:

- دارای مقدار p-value کمتر از آلفا ( $\alpha$ ) است.
- با اضافه شدن به مدل باعث افزایش مقدار Adjusted  $R^2$  میشود.
- دارای یک رابطه خطی قوی با متغیر Response است.

## : 4.Part F

نکته : همانطور که در ابتدای حل سوال ۴ گفته شد ، برای حل این سوال یک نمونه با سایز ۱۰۰ گرفته ام و با همان نمونه این سوال را حل می کنم. بنابراین برای حل این قسمت هم از همان نمونه ی گرفته شده استفاده میکنم. (جهت یادآوری ما این نمونه را با نام my\_sample\_4 میشناسیم).

: a

در ابتدا باید ۹۰ درصد داده های این sample را انتخاب کرد. برای این کار در R دستور زیر را اجرا میکنم.

```
# I considered 90% of the data as new_data  
ned <- sample(nrow(my_sample_4), nrow(my_sample_4)*.90)  
new_data <- my_sample_4[ned,]
```

جهت یاد آوری : برای این سوال ، متغیر عددی "health\_bills" را به عنوان متغیر Response انتخاب کرده ام که میزان هزینه سالیانه ای را که هر فرد درون این مجموعه داده برای سلامت خود می پردازد را نشان می دهد. همچنین دو متغیر "age" و "avg\_glucose\_level" که به ترتیب نشان دهنده سن هر فرد ، و متوسط قند خون (سطح گلوکز) برای هر فرد هستند را به عنوان متغیر های Explanatory انتخاب کرده ام.

بنابراین باید دو مدل با استفاده از این دو متغیر Explanatory بسازیم :

- ۱- مدل Regression که با ۹۰ درصد داده های sample و بر اساس متوسط قند خون (سطح گلوکز) هر فرد ، هزینه های سالیانه ای که یک فرد برای سلامت خود می پردازد را پیش بینی می کند :  
در ابتدا برای داشتن یک مدل Regression باید شرایط لازم برقرار باشند که من قبلابرا این دو متغیر شرایط را بررسی کرده ام و شرایط برقرارند.

برای ساخت مدل ، از تابع (`lm`) در R استفاده می کنیم :

```
> # create my Regression models  
> modelnum11 <- summary(lm(health_bills ~ avg_glucose_level, data=new_data))  
> modelnum11  
  
Call:  
lm(formula = health_bills ~ avg_glucose_level, data = new_data)  
  
Residuals:  
    Min      1Q   Median      3Q     Max  
-1296.40  -540.54   -33.95   332.13  2907.20  
  
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) 2551.271   213.873 11.929 < 2e-16 ***  
avg_glucose_level 5.567     1.806  3.083 0.00274 **  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 763.3 on 88 degrees of freedom  
Multiple R-squared:  0.09749, Adjusted R-squared:  0.08723  
F-statistic: 9.506 on 1 and 88 DF,  p-value: 0.002736
```

کد :

نتیجه :

- ۲- مدل Regression که با ۹۰ درصد داده های sample و بر اساس سن هر فرد (بر اساس متغیر "age" ، هزینه های سالیانه ای که یک فرد برای سلامت خود می پردازد را پیش بینی می کند :  
در ابتدا برای داشتن یک مدل Regression باید شرایط لازم برقرار باشند که من قبلابرا این دو متغیر شرایط را بررسی کرده ام و شرایط برقرارند.

برای ساخت مدل ، از تابع  $\text{Im}(z)$  در R استفاده می کنیم :

```
> # create my Regression models
> modelnum22 <- summary(lm(health_bills ~ age, data=new_data))
> modelnum22

call:
lm(formula = health_bills ~ age, data = new_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-1099.8 -552.9   -77.1   296.8  3034.7 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2689.170    170.825   15.742 < 2e-16 ***
age          11.148      3.553    3.138  0.00232 **  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 761.9 on 88 degrees of freedom
Multiple R-squared:  0.1006,    Adjusted R-squared:  0.0904 
F-statistic: 9.845 on 1 and 88 DF,  p-value: 0.002317
```

حال در اینجا با استفاده از آزمون فرض می خواهیم بررسی کنیم که آیا predictor ها significant می شوند یا خیر (یعنی متغیر خوبی برای تخمین متغیر Response هستند یا خیر). در واقع باید هر دو متغیر Explanatory را بررسی کرد تا دید که متغیر خوبی است یا خیر. اگر بخواهیم فریم ورک این آزمون را معرفی کنیم وفرض صفر و فرض جایگزین را تعریف کنیم خواهیم داشت:

۱- باید در ابتدا برای هر متغیر فرض صفر و فرض جایگزین را چنین تعریف کرد:  
**(برای متغیر age)**

**متغیر age** در بیش بینی متغیر Response ، متغیر معنا داری نیست.

متغیر age در بیش بینه متغیر Response ، متغیر معناداری است.

## (برای متغیر avg\_glucose\_level)

**متغیر avg glucose level** در بیش بینه متغیر Response ، متغیر معنا داری نیست.

**متغیر avg\_glucose\_level** در پیش بینی متغیر Response ، متغیر معنا داری است.

-۲ سپس باید آماره  $t$  را به صورت زیر محاسبه کنیم :  
**(برای متغیر age)**

$$T = \frac{\hat{\beta}_1 - \text{null value}}{SE_{b1}} = \frac{11.148 - 0}{3.553} = 3.138$$

(برای متغیر avg\_glucose\_level)

$$T = \frac{\hat{\beta}_1 - null\ value}{SE_{b1}} = \frac{5.567 - 0}{1.806} = 3.083$$

-۳- حال باید  $p$ -value را محاسبه کنیم. (تست دوطرفه است).

(برای متغیر age)

$$pvalue : P(|T| > 3.138) = P(T > 3.138) + P(T < -3.138) = 0.00232$$

## (برای متغیر avg\_glucose\_level)

$$pvalue : P(|T| > 3.083) = P(T > 3.083) + P(T < -3.083) = 0.00274$$

۴- در نهایت به بررسی significant بودن یا نبودن آن متغیر با  $\alpha = 0.05$  پردازیم:  
**(برای متغیر age)**

چون  $0.00232 < 0.05$  است بنابراین فرض صفر را رد می‌کنیم. پس این متغیر Significant می‌شود.

(برای متغیر avg\_glucose\_level)

چون  $0.05 < 0.00274$  است بنابراین فرض صفر را رد می‌کنیم. پس این متغیر Significant می‌شود.

\*\* نکته: تمام این موارد به صورت خودکار توسط  $R$  و دستور ( $\text{Im}$ ) تولید شده اند در در بالا قابل مشاهده هستند.

برای ساخت بازه اطمینان برای شیب خط رگرسیون فرم کلی آن بدین شرح است :

$$\text{point estimate} \pm ME \rightarrow b_1 \pm t_{df}^* SE_{b1}$$

بنابراین اگر بخواهیم برای ضریب هر در متغیر Explanatory در این سوال (همان شیب خط) بازه اطمینان ۹۵ درصد بسازیم خواهیم داشت :

۱- بازه اطمینان برای شیب مدلی که با متغیر avg\_glucose\_level ساخته شده :

$$df = n - 2 = 90 - 2 = 89$$

$$t_{89}^* = 1.98$$

```
> # to calculate the t-statistic
> qt(0.025, df=89)
[1] -1.986979
```

$$b_1 \pm t_{df}^* SE_{b1} = 5.567 \pm 1.98 \times 1.806 = (1.97, 9.15)$$

کد محاسبات بالا در R :

```
> # df
> dff <- 90 - 2
> # to calculate the t-statistic
> t_sta <- qt(0.025, df=dff)
> # CI : point estimate ± ME
> CI_1 <- c(modelnum11$coefficients[2] - (abs(t_sta) * modelnum11$coefficients[4]),
+            modelnum11$coefficients[2] + (abs(t_sta) * modelnum11$coefficients[4]))
> CI_1
[1] 1.978583 9.154792
```

تفسیر: ما ۹۵ درصد اطمینان داریم که به ازای هر یک واحد افزایش در سطح قند خون هر فرد ، انتظار داریم که میزان هزینه ای که این فرد سالیانه برای سلامتی خود باید پردازد ، به طور متوسط بین ۱/۹۷ و ۹/۱۵ واحد افزایش یابد.

۲- بازه اطمینان برای شیب مدلی که با متغیر age ساخته شده :

$$df = n - 2 = 90 - 2 = 89$$

$$t_{89}^* = 1.98$$

```
> # to calculate the t-statistic
> qt(0.025, df=89)
[1] -1.986979
```

$$b_1 \pm t_{df}^* SE_{b1} = 11.148 \pm 1.98 \times 3.553 = (4.08, 18.20)$$

کد محاسبات بالا در R :

```
> # create 95% CI for health_bills ~ age
> # df
> dff <- 90 - 2
> # to calculate the t-statistic
> t_sta <- qt(0.025, df=dff)
> # CI : point estimate ± ME
> CI_2 <- c(modelnum22$coefficients[2] - (abs(t_sta) * modelnum22$coefficients[4]),
+            modelnum22$coefficients[2] + (abs(t_sta) * modelnum22$coefficients[4]))
> CI_2
[1] 4.087226 18.208048
```

تفسیر: ما ۹۵ درصد اطمینان داریم که به ازای هر یک واحد افزایش در سن هر فرد ، انتظار داریم که میزان هزینه ای که این فرد سالیانه برای سلامتی خود باید پردازد ، به طور متوسط بین ۴/۰۸ و ۱۸/۲۰ واحد افزایش یابد.

در این قسمت می خواهیم با استفاده ۱ درصد داده های باقی مانده از sample ، دو مدل ایجاد کنیم و مقادیری از متغیر Response همان متغیر (Response) را تخمین بزنم.

در ابتدا دو مدل را با ۱ درصد داده های باقی مانده ایجاد می کنیم ، سپس معادله هر مدل Regression را می نویسیم و در نهایت بر اساس یک مقدار از متغیر Explanatory و با استفاده از مدل ساخته شده اقدام به تخمین متغیر Response می کنیم :

```
# the remaining percent of samples.
new_data_2 <- my_sample_4[-ned,]

# create my Regression models --> health_bills ~ avg_glucose_level
modelnum111 <- summary(lm(health_bills ~ avg_glucose_level,data=new_data_2))

# create my Regression models --> health_bills ~ age
modelnum222 <- summary(lm(health_bills ~ age,data=new_data_2))
```

```
> modelnum111
call:
lm(formula = health_bills ~ avg_glucose_level, data = new_data_2)

Residuals:
    Min      1Q  Median      3Q     Max 
-1394.39 -650.07   52.91  219.29 1628.70 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 4149.289   1042.587   3.980  0.00406 ***
avg_glucose_level -5.101     6.958  -0.733  0.48442  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 978.7 on 8 degrees of freedom
Multiple R-squared:  0.06295, Adjusted R-squared:  -0.05418 
F-statistic: 0.5374 on 1 and 8 DF,  p-value: 0.4844
```

```
> modelnum222
call:
lm(formula = health_bills ~ age, data = new_data_2)

Residuals:
    Min      1Q  Median      3Q     Max 
-1322.16 -625.33  -30.18  169.53 1790.22 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 3679.619    631.067   5.831 0.000391 ***
age          -6.331    13.301  -0.476 0.646810  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 997 on 8 degrees of freedom
Multiple R-squared:  0.02754, Adjusted R-squared:  -0.09402 
F-statistic: 0.2266 on 1 and 8 DF,  p-value: 0.6468
```

معادله این مدل :

$$\text{health\_bills} = 4149.289 - 5.101 \text{ avg\_glucose\_level}$$

حال می خواهیم برای یک فرد که متوسط سطح گلوکز آن برابر با ۱۲۰ است ، میزان هزینه ای که در سال برای سلامتی خود باید پردازد را با استفاده از مدل رگرسیونی که ساختیم تخمین بزنیم :

$$\text{health\_bills} = 4149.289 - 5.101 \times 120 = 3537.169$$

معادله این مدل :

$$\text{health\_bills} = 3679.619 - 6.331 \text{ age}$$

حال می خواهیم برای یک فرد که سن آن ۴۰ سال است ، میزان هزینه ای که در سال برای سلامتی خود باید پردازد را با استفاده از مدل رگرسیونی که ساختیم تخمین بزنیم :

$$\begin{aligned} \text{health\_bills} &= 3679.619 - 6.331 \times 40 \\ &= 3426.379 \end{aligned}$$

برای مقایسه مقادیر واقعی متغیر Response با مقادیر تخمین زده برای این متغیر توسط هر دو مدل (که با استفاده از دو متغیر Explanatory انتخابی در این سوال ایجاد شده اند) ، ابتدا در R اقدام به بدست آوردن مقادیر actual می کنیم و آن را درون یک وکتور به نام vec\_1 میریزیم. سپس به کمکتابع predict() مقادیر تخمین زده توسط مدل ای که با استفاده از متغیر توضیحی avg\_glucose\_level ایجاد کردم را بدست آورده و درون وکتور دیگری به نام vec\_p میریزیم.

کد این قسمت در زیر قابل مشاهده است :

```
# obtain actual values for response variable from the data set
vec_1 <- c(unlist(new_data$health_bills))
# obtain predicted values for response variable from my Linear model
vec_p <- c(predict(lm(health_bills ~ avg_glucose_level,data=new_data)))
```

```
> # compare :
> vec_p == vec_1
 2045 3145 164 5096 2705 664 551 2794 3104 3430 899 1782 673 4331
FALSE FALSE
 2876 3955 4400 1958 2064 5020 1847 2830 2802 1917 2725 1085 4975 355
FALSE FALSE
 3687 1570 2224 2502 4525 1145 2007 3572 4520 217 2580 2702 1680 4054
FALSE FALSE
 1941 2692 923 4248 3267 2950 196 3481 1227 595 3892 2608 3576 5010
FALSE FALSE
 4745 78 4722 425 3555 1793 3149 336 3505 4499 3052 340 2564 3239
FALSE FALSE
 426 2598 3983 661 4582 1345 1656 3056 4769 2168 4824 2905 60 3672
FALSE FALSE
 3766 4997 1061 1922 1386 2467
FALSE FALSE
```

مشاهده می شود که هیچ کدام از تخمین های زده شده توسط مدلمان کاملا دقیق نتوانسته مقادیر را پیش بینی کند اما تا حد زیادی به مقادیر واقعی نزدیک هستند. به دلیل اینکه ذات تخمین زدن لزوما با مقدار واقعی برابر نخواهد شد و همیشه یک خطای تمامی مدل های ایجاد شده خواهد داشت (در واقع همان residual ها).

همچنین اگر مقادیر تخمین زده شده توسط مدلی که با متغیر توضیحی age ایجاد شده است را درون وکتور دیگری ریخته و با مقادیر actual مقایسه کنیم خواهیم دید :

```
> vec_a <- c(predict(lm(health_bills ~ age,data=new_data)))
>
> # second compare :
> vec_a == vec_1
 2045 3145 164 5096 2705 664 551 2794 3104 3430 899 1782 673 4331
FALSE FALSE
 2876 3955 4400 1958 2064 5020 1847 2830 2802 1917 2725 1085 4975 355
FALSE FALSE
 3687 1570 2224 2502 4525 1145 2007 3572 4520 217 2580 2702 1680 4054
FALSE FALSE
 1941 2692 923 4248 3267 2950 196 3481 1227 595 3892 2608 3576 5010
FALSE FALSE
 4745 78 4722 425 3555 1793 3149 336 3505 4499 3052 340 2564 3239
FALSE FALSE
 426 2598 3983 661 4582 1345 1656 3056 4769 2168 4824 2905 60 3672
FALSE FALSE
 3766 4997 1061 1922 1386 2467
FALSE FALSE
```

باز هم وضعیت مشابه با مقایسه قبل را خواهیم داشت. هیچ کدام از تخمین های زده شده توسط مدلمان کاملا دقیق نتوانسته مقادیر را پیش بینی کند اما تا حد زیادی به مقادیر واقعی نزدیک هستند.

سوال شماره ۵

از آنجاکه در سوال چهارم ، متغیر "health\_bills" بود برای این سوال هم همان را به عنوان متغیر Response انتخاب کرده ام. این متغیر میزان هزینه سالیانه ای را که هر فرد درون این مجموعه داده برای سلامتی خود می پردازد را نشان می دهد. همچنین همانند سوال چهارم ، از دو متغیر "age" و "avg\_glucose\_level" که به ترتیب نشان دهنده سن هر فرد ، و متوسط قند خون (سطح گلوکز) برای هر فرد هستند ، و متغیر "bmi" که شاخص bmi را برای هر فرد نشان میدهد ، به عنوان متغیر های Explanatory استفاده می کنم.

نکته ۱ : طبق آن چیزی که در فاز اول این پروژه عنوان شد ، متغیر health\_bills دارای مقادیر missing (N/A values) میباشد و رویکرد من برای مقادیر گمشده در این متغیر ، استفاده از روش جایگذاری میانه به جای مقادیر گم شده است. بنابراین در ابتدا این عمل را روی مجموعه داده HealthCare در R با استفاده از دستور زیر انجام می دهیم. در واقع هر سطر از مجموعه داده HealthCare که برای متغیر ذکر شده مقدار گم شده داشته باشد ، برای آن سطر میانه مقادیری که آن متغیر دارد را جایگذاری میکنم.

```
# at first lets replace missing values with the mean
HealthCare2 <- HealthCare
HealthCare2$health_bills[is.na(HealthCare2$health_bills)]<-median(HealthCare2$health_bills,na.rm=TRUE)
```

نکته ۲ : از آنجاکه درون این مجموعه داده حدود 5000 تا داده وجود دارد ، در این سوال فرض را بر این گذاشته ام که جامعه هدف همان مجموعه داده است و بنابراین در ابتدا یک نمونه با سایز 100 به صورت کاملاً تصادفی و بدون جایگذاری می گیرم و از همین نمونه برای پاسخ دادن به این سوال استفاده می کنم.

(دستور ایجاد sample در R)

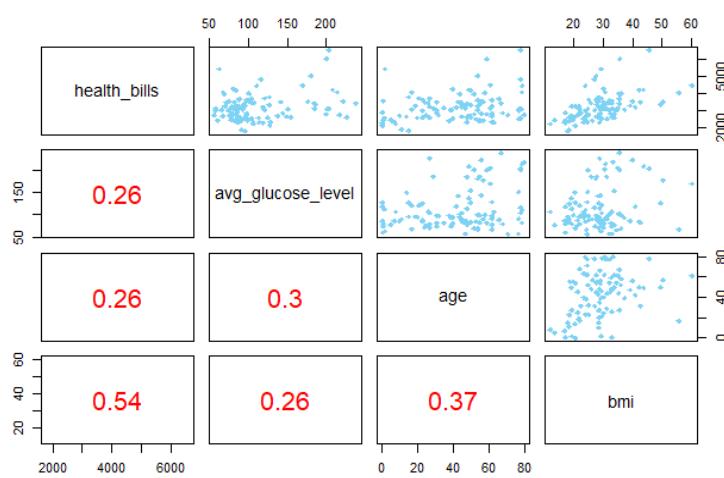
```
# randomly select a sample (n=100) without replacement from HealthCare
my_sample_5 <- HealthCare[sample(nrow(HealthCare), size = 100, replace = FALSE),]
```

## : 5.Part A

در ابتدا در R نمودار pairwise scatterplot را رسم می کنیم تا بتوانیم در مورد رابطه بین متغیرهایی که انتخاب کردیم صحبت کنیم.

(دستور ایجاد نمودار pairwise scatterplot در R و نتیجه اجرای آن)

```
# to create the correlation panel and texts
func2 <- function(x, y){
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  value <- round(cor(x, y), digits=2)
  text(0.5, 0.5, value, cex = 2,col="red")
}
# to create scatterplots
func2 <- function(x, y){
  points(x,y, pch = 18, col = "skyblue")
}
# Create the pairwise scatter plot
pairs(health_bills ~ avg_glucose_level + age + bmi,
      data=my_sample_4,
      upper.panel = func2,
      lower.panel = func2)
```



در این نمودار رابطه بین متغیرهای Explanatory با Response قابل مشاهده است. همانطور که از نمودار پیدا شده است متغیر bmi برای تخمين متغیر health\_bills متریک بهتری نسبت به سایرین است به دلیل اینکه ضریب همبستگی بالاتری دارد. از طرف دو متغیر age و bmi دارای correlation قوی تری نسبت به بقیه متغیرها هستند ، و اصطلاحاً این دو متغیر collinear محسوب میشوند. بنابراین وجود این دو متغیر با هم در مدل مان چیز زیادی به مقدار Adjusted R<sup>2</sup> اضافه نمی کند. بنابراین طبق توضیحات bmi داده شده و براساس اطلاعات نمودار ، متغیر برای متغیر Response بهتری است. (significant)

## : 5.Part B

در این قسمت می خواهیم با استفاده از هر سه متغیر Explanatory که انتخاب کردیم ، یک مدل خطی بسازیم. ما در مفهوم رگرسیون ، براساس متغیر Explanatory می خواهیم یک تخمین گر برای تخمین متغیر Response بسازیم. در واقع با استفاده از خط رگرسیونی یک تخمین برای مقادیر متغیر Response می نزیم.

در اینجا به دلیل اینکه از چندین متغیر استفاده می کنیم ، در واقع داریم یک Multiple linear Regression تحت شرایط خاص به ما جواب خوبی می دهد. این شرایط عبارتند از:

- ۱- خطی بودن رابطه بین هر متغیر Explanatory با متغیر Response
- ۲- نرمال بودن توزیع Residual ها
- ۳- تغییرات ثابت در Residual ها (برابر بودن مقدار underestimate ها و overestimate ها)
- ۴- مستقل بودن تک تک Residual ها

که در اینجا فرض را بر این میگذاریم که هر ۴ شرط برقرار اند (البته اگه برقرار نباشند هم مدل را ایجاد می کنیم چون گفتیم که اگر این ۴ شرط برقرار باشند ما یک مدل خوب خواهیم داشت.)

برای ساخت مدل ، ازتابع `lm()` در R استفاده می کنیم :

```
> # fit a multiple linear regression model :  
> bills_model <- summary(lm(health_bills ~ avg_glucose_level + age + bmi , data=my_sample_4))  
> bills_model
```

Call:  
`lm(formula = health_bills ~ avg_glucose_level + age + bmi, data = my_sample_4)`

Residuals:

Min	1Q	Median	3Q	Max
-931.1	-415.4	-105.4	272.3	2424.3

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1514.333	271.205	5.584	2.19e-07 ***
avg_glucose_level	2.160	1.600	1.350	0.180
age	1.504	3.335	0.451	0.653
bmi	46.929	8.847	5.304	7.24e-07 ***

---  
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 687.8 on 96 degrees of freedom  
Multiple R-squared: 0.3075, Adjusted R-squared: 0.2859  
F-statistic: 14.21 on 3 and 96 DF, p-value: 9.782e-08

کد :

نتیجه :

معادله این مدل :

$$\text{health\_bills} = 1514.333 + 2.160 \text{ avg\_glucose\_level} + 1.504 \text{ age} + 46.929 \text{ bmi}$$

## : 5.Part C

می دانیم که R-square معیاری است که قدرت مدل Regression را نشان می دهد و می گوید که چه میزان از تغییرات متغیر Response توسط مدلی که ساختیم توضیح داده می شود.  $R^2$  مقداری بین ۰ و ۱ دارد و هرچه این مقدار بیشتر باشد به معنی بهتر بودن مدل است. از آنجا که این مقدار توسط مدلی که در قسمت قبل ساختیم محاسبه شده است بنابراین ، ۳۰.۷۵ درصد از variability که در میزانه هزینه ای که در سال هر فرد برای سلامتی خود باید پردازد (همان متغیر Response) ، توسط مدل توضیح داده می شود.

## : 5.Part D

اگر بخواهیم کل مدل را در نظر بگیریم و استنباط برای کل مدل انجام بدھیم می توانیم از آزمون ANOVA برای این کار استفاده کنیم تا ببینیم که آبا کل مدل یک prediction خوب است یا خیر.

اگر فرض صفر و فرض جایگزین را اینگونه در نظر بگیریم :

$H_0: \beta_1 = \beta_2 = \beta_3 = 0$  هر سه متغیر Explanatory که انتخاب کردیم خوب نیستند.

$H_A : \text{at least one } \beta_i \text{ is different than 0}$  حداقل یکی از متغیرهایی که انتخاب کردیم significant است و در کل مدل prediction خوبی است.

حال با توجه به جدولی که داشتیم مقدار p-value که از تست ANOVA برای کل مدل بدست آمده است برابر است با :

```
> # fit a multiple linear regression model :
> bills_model <- summary(lm(health_bills ~ avg_glucose_level + age + bmi , data=my_sample_4))
> bills_model

call:
lm(formula = health_bills ~ avg_glucose_level + age + bmi, data = my_sample_4)

Residuals:
    Min      1Q Median      3Q     Max 
-931.1 -415.4 -105.4  272.3 2424.3 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1514.333   271.205   5.584 2.19e-07 ***
avg_glucose_level 2.160    1.600   1.350   0.180    
age          1.504    3.335   0.451   0.653    
bmi          46.929   8.847   5.304 7.24e-07 ***  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 687.8 on 96 degrees of freedom
Multiple R-squared:  0.3075,   Adjusted R-squared:  0.2859 
F-statistic: 14.21 on 3 and 96 DF,  p-value: 9.782e-08
```

همانطور که مشاهده می شود این مقدار خیلی کوچک است. بنابراین فرض صفر رد می شود و این یعنی کل مدلی که من ایجاد کرده ام با هر سه متغیر با هم ، در واقع مدل significant است. من یک prediction خوب است.

همچنین گفتیم که R-square معیاری است که قدرت مدل Regression را نشان می دهد و می گوید که چه میزان از تغییرات متغیر Response توسط مدلی که ساختیم توضیح داده می شود. ما در قسمت قبل دیدیم که 30.75 درصد از variability که در میزانه هزینه ای که در سال هر فرد برای سلامتی خود باید پردازد (همان متغیر Response) ، توسط مدل توضیح داده می شود.

## : 5.Part E

برای انتخاب یک مدل بهینه در این سوال ابتدا از روش **Adjusted R<sup>2</sup>** و **forward selection** استفاده می کنیم.

در این روش از یک مدل با یک متغیر شروع می کنیم و  $R^2$  را Adjusted  $R^2$  را حساب می کنیم. آن مدلی که بالاترین  $R^2$  را داشت انتخاب می کنیم و متغیر بعدی را اضافه می کنیم. آنقدر این کار را تکرار می کنیم تا هیچ متغیری باعث نشود که  $R^2$  افزایش پیدا کند و متوقف شویم.

**\*\* نکته :** کد انجام این کار در R هم زده شده است و من به جای اسکرین شات ، همان کد را دقیقا در هر سطر جدول می گذارم و  $R^2$  حساب شده آن در R را در جدول جلوی کد می نویسم.

### : STEP 1

Variables included	Adjusted R <sup>2</sup>
lm(health_bills ~ avg_glucose_level , data=my_sample_4)	0.05946
lm(health_bills ~ age , data=my_sample_4)	0.05685
lm(health_bills ~ bmi , data=my_sample_4)	0.2828

مشاهده می شود که با متغیر bmi به بالاترین  $R^2$  می رسیم. بنابراین این را انتخاب می کنیم و به مرحله بعد می رویم.

## : STEP 2

Variables included	Adjusted R <sup>2</sup>
lm(health_bills ~ bmi + avg_glucose_level , data=my_sample_4)	0.2917
lm(health_bills ~ bmi + age , data=my_sample_4)	0.2798

مشاهده می شود که با متغیرهای bmi و avg\_glucose\_level به بالاترین Adjusted R<sup>2</sup> می رسیم. بنابراین این را انتخاب می کنیم و به مرحله بعد می رویم.

## : STEP 3

Variables included	Adjusted R <sup>2</sup>
lm(health_bills ~ bmi + avg_glucose_level + age , data=my_sample_4)	0.2859

از آنجایی که در 2 STEP به مقدار 0.2917 رسیدیم و در 3 STEP هم همانطور که مشاهده می شود با اضافه شدن متغیر age مقدار Adjusted R-square کاهش یافته است. بنابراین بهترین مدل و بهینه ترین مدل ، همان است که در 2 STEP بدست آورده (به دلیل یک متغیر کمتر و Adjusted R<sup>2</sup> بیشتر).

سپس برای انتخاب یک مدل بهینه در این سوال از روش forward selection و معیار P-value استفاده می کنیم.

در این روش با یک Predictor شروع می کنیم ، و متغیری را انتخاب می کنیم که کمترین p-value را دارد. سپس متغیر بعدی را اضافه می کنیم و این قدر این کار را تکرار می کنیم تا جایی که هیچ متغیر دیگری نتوانیم اضافه کنیم.

## : STEP 1

```
> # STEP 1 :
> # variable included : health_bills ~ avg_glucose_level
> summary(lm(health_bills ~ avg_glucose_level , data=my_sample_4))

Call:
lm(formula = health_bills ~ avg_glucose_level, data = my_sample_4)

Residuals:
    Min      1Q  Median      3Q     Max 
-1319.2  -570.5 -111.8  365.1 2982.6 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2661.256   210.820 12.623 <2e-16 ***
avg_glucose_level 4.657     1.729  2.694  0.0083 **  
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 789.3 on 98 degrees of freedom
Multiple R-squared:  0.06896,  Adjusted R-squared:  0.05946 
F-statistic: 7.259 on 1 and 98 DF,  p-value: 0.0083

>
> # variable included : health_bills ~ age
> summary(lm(health_bills ~ age , data=my_sample_4))

Call:
lm(formula = health_bills ~ age, data = my_sample_4)

Residuals:
    Min      1Q  Median      3Q     Max 
-1148.8 -550.8 -102.9  296.8 3077.6 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2799.532   167.037 16.76 < 2e-16 ***
age          9.182     3.479  2.64  0.00966 ** 
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 790.4 on 98 degrees of freedom
Multiple R-squared:  0.06638,  Adjusted R-squared:  0.05685 
F-statistic: 6.967 on 1 and 98 DF,  p-value: 0.009659

> # variable included : health_bills ~ bmi
> summary(lm(health_bills ~ bmi , data=my_sample_4))

Call:
lm(formula = health_bills ~ bmi, data = my_sample_4)

Residuals:
    Min      1Q  Median      3Q     Max 
-895.9 -461.4 -120.7  261.4 2551.6 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1690.920   246.410  6.862 6.18e-10 ***
bmi         51.441    8.129  6.328 7.47e-09 *** 
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 689.2 on 98 degrees of freedom
Multiple R-squared:  0.2901,  Adjusted R-squared:  0.2828 
F-statistic: 40.04 on 1 and 98 DF,  p-value: 7.469e-09
```

همان طور که مشاهده می شود مدلی که با متغیر bmi ساخته شده است دارای کمترین p-value ممکن است. بنابراین این متغیر را انتخاب کرده و به مرحله بعد می رویم.

## : STEP 2

```

> # STEP 2 :
> # variable included : health_bills ~ bmi + avg_glucose_level
> summary(lm(health_bills ~ bmi + avg_glucose_level , data=my_sample_4))

Call:
lm(formula = health_bills ~ bmi + avg_glucose_level, data = my_sample_4)

Residuals:
    Min      1Q Median      3Q     Max 
-922.6 -416.8 -106.6  277.2 2395.1 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1523.540   269.322   5.657 1.56e-07 ***
bmi          48.176    8.369   5.756 1.01e-07 ***
avg_glucose_level 2.320    1.554   1.493   0.139    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 684.9 on 97 degrees of freedom
Multiple R-squared:  0.306, Adjusted R-squared:  0.2917 
F-statistic: 21.39 on 2 and 97 DF,  p-value: 2.019e-08

>
> # variable included : health_bills ~ bmi + age
> summary(lm(health_bills ~ bmi + age , data=my_sample_4))

Call:
lm(formula = health_bills ~ bmi + age, data = my_sample_4)

Residuals:
    Min      1Q Median      3Q     Max 
-899.25 -417.30 -75.09  302.60 2505.58 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1656.343   251.014   6.599 2.20e-09 ***
bmi          48.990    8.751   5.598 2.01e-07 ***
age          2.504    3.265   0.767   0.445    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 690.7 on 97 degrees of freedom
Multiple R-squared:  0.2944, Adjusted R-squared:  0.2798 
F-statistic: 20.23 on 2 and 97 DF,  p-value: 4.534e-08

```

همان طور که مشاهده می شود با اضافه شدن متغیر avg\_glucose\_level به مدل ساخته شده است ، این متغیر نسبت به متغیر age دارای کمترین p-value ممکن است. بنابراین این متغیر را انتخاب کرده و به مرحله بعد می رویم.

بنابراین دیگر نمیتوانیم هیچ متغیری اضافه کنیم و درنتیجه بهترین مدل همان مدلی است که در STEP 2 دوم به آن رسیدیم.

همانطور که مشاهده می شود دو مدلی که با استفاده از روش Forward selection و با معیار های R2 و Adjusted R2 بدست آورده ایم مثل هم شدند.

**سپس برای انتخاب یک مدل بهینه در این سوال از روش Backward Elimination و معیار Adjusted R<sup>2</sup> استفاده می کنیم.**

در این روش از یک مدل کامل با همه ی متغیر های انتخابی شروع می کنیم و سپس یکی یکی متغیر ها را حذف می کنیم و هر بار را حساب می کنیم. سپس مدلی را انتخاب می کنیم که Adjusted R2 بالاتری دارد و این کار را تا جایی که حذف هیچ متغیری باعث افزایش Adjusted R2 نشود ادامه می دهیم.

## : STEP 1

Variables included	removed	Adjusted R <sup>2</sup>
lm(health_bills ~ bmi + avg_glucose_level + age , data=my_sample_4)		<b>0.2859</b>
lm(health_bills ~ bmi + avg_glucose_level , data=my_sample_4)	-age	<b>0.2917</b>
lm(health_bills ~ bmi + age , data=my_sample_4)	-avg	<b>0.2798</b>
lm(health_bills ~ avg_glucose_level + age , data=my_sample_4)	-bmi	<b>0.08605</b>

مشاهده می شود که با حذف متغیر age به بالاترین R<sup>2</sup> می رسیم. بنابراین این را حذف می کنیم و به مرحله بعد می رویم.

## : STEP 2

Variables included	removed	Adjusted R <sup>2</sup>
lm(health_bills ~ bmi , data=my_sample_4)	-avg	<b>0.2828</b>
lm(health_bills ~ avg_glucose_level , data=my_sample_4)	-bmi	<b>0.05946</b>

مشاهده می شود که در این مرحله Adjusted R<sup>2</sup> هر دو مدل از مدلی که در مرحله اول بدست آوریم کمتر شده است. بنابراین همین جا متوقف می شویم و مدل بهینه مدلی است که در STEP1 بدست آوردیم.

\*\* مشاهده می شود که در روش Forward selection با هر دو معیار ذکر شده هم به همین مدل رسیدیم.

حال می خواهیم برای انتخاب یک مدل بهینه در این سوال از روش Backward Elimination و معیار P-value استفاده کنیم.

در این روش با کل مدل و همه متغيرها شروع می کنیم. سپس متغیری را حذف می کنیم که بالاترین P-value را دارد (در واقع کمتری دارد). و این کار را آنقدر تکرار می کنیم تا جایی که همه متغیرهای ما significant شوند.

## : STEP 1

```
> # STEP 1 :
> # variable included : health_bills ~ bmi + avg_glucose_level + age
> summary(lm(health_bills ~ bmi + avg_glucose_level + age , data=my_sample_4))

Call:
lm(formula = health_bills ~ bmi + avg_glucose_level + age, data = my_sample_4)

Residuals:
    Min      1Q Median      3Q     Max 
-931.1 -415.4 -105.4  272.3 2424.3 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1514.333   271.205   5.584 2.19e-07 ***
bmi          46.929    8.847   5.304 7.24e-07 ***
avg_glucose_level 2.160    1.600   1.350   0.180    
age           1.504    3.335   0.451   0.653  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 687.8 on 96 degrees of freedom
Multiple R-squared:  0.3075,   Adjusted R-squared:  0.2859 
F-statistic: 14.21 on 3 and 96 DF,  p-value: 9.782e-08
```

همانطور که مشاهده می شود متغیر p-value دارای بالاترین مقدار نسبت بقیه است و بنابراین آن را حذف می کنیم و به مرحله بعد میرویم.

## : STEP 2

```
> # STEP 2 :
> # variable included : health_bills ~ bmi + avg_glucose_level
> summary(lm(health_bills ~ bmi + avg_glucose_level , data=my_sample_4))

Call:
lm(formula = health_bills ~ bmi + avg_glucose_level, data = my_sample_4)

Residuals:
    Min      1Q Median      3Q     Max 
-922.6 -416.8 -106.6  277.2 2395.1 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1523.540   269.322   5.657 1.56e-07 ***
bmi          48.176    8.369   5.756 1.01e-07 ***
avg_glucose_level 2.320    1.554   1.493   0.139  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 684.9 on 97 degrees of freedom
Multiple R-squared:  0.306,   Adjusted R-squared:  0.2917 
F-statistic: 21.39 on 2 and 97 DF,  p-value: 2.019e-08
```

همانطور که مشاهده می شود متغیر avg\_glucose\_level از آنجا که مقدار p-value آن کمتر از 0.05 است و بنابراین آن را significant نمیدانیم و حذف می کنیم.

بنابراین مدل بهینه بدست آمده در این روش مدلی است که فقط با متغیر bmi ساخته می شود. مشاهده می شود که در این روش مدل بهینه بدست آمده با بقیه روش ها متفاوت است. می دانیم که لزوماً روش Forward و Backward با معیارهای یکسان

هم منجر به مدل بھینه یکسان نمیشوند. اما با توجه به اینکه مدلی که با متغیر های bmi و avg\_glucose\_level ایجاد کردیم در بین همه بھینه ترین بود. بنابراین همین مدل را انتخاب می کنیم.

مدل انتخابی :

```
> summary(lm(health_bills ~ bmi + avg_glucose_level , data=my_sample_4))
call:
lm(formula = health_bills ~ bmi + avg_glucose_level, data = my_sample_4)
Residuals:
    Min      1Q  Median      3Q     Max 
-922.6 -416.8 -106.6  277.2 2395.1 
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1523.540   269.322   5.657 1.56e-07 ***
bmi          48.176     8.369   5.756 1.01e-07 ***
avg_glucose_level 2.320     1.554   1.493   0.139    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
Residual standard error: 684.9 on 97 degrees of freedom
Multiple R-squared:  0.306, Adjusted R-squared:  0.2917 
F-statistic: 21.39 on 2 and 97 DF, p-value: 2.019e-08
```

معادله این مدل :

$$\text{health\_bills} = 1523.540 + 48.176 \text{ bmi} + 2.320 \text{ avg\_glucose\_level}$$

## : 5.Part F

از آنجاکه در مدل انتخابی به دلیل اینکه از چندین متغیر Explanatory استفاده می کنیم ، در واقع داریم یک Multiple linear Regression میسازیم و می دانیم که تحت شرایطی خاص به ما جواب خوبی می دهد. این شرایط عبارتند از :

- خطی بودن رابطه بین هر متغیر Explanatory با متغیر Response
- نرمال بودن توزیع Residual ها (Nearly normal residuals)
- تغییرات ثابت در Residual ها (برابر بودن مقدار underestimate ها و overestimate ها) یا (variability Constant)

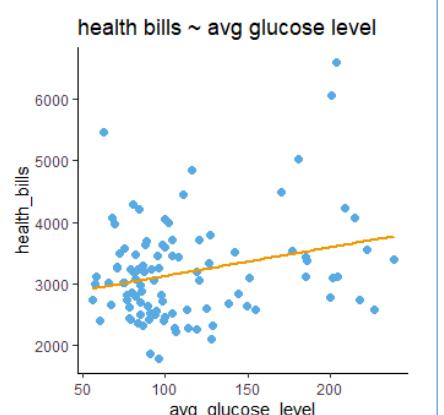
در این قسمت می خواهیم بررسی کنیم که آیا برای مدل بھینه انتخابی در قسمت قبل این شرایط برقرار هستند یا خیر.

**شرط Linearity :** طبق این شرط رابطه بین دو متغیر "health\_bills" و "avg\_glucose\_level" و همچنین دو متغیر "bmi" و "health\_bills" باید خطی باشد. برای بررسی این شرط اقدام به رسم یک نمودار scatterplot در R به کمک کتابخانه ggplot می کنیم.

health bills ~ avg\_glucose\_level

```
# 1
# checking the linearity condition for health bills ~ avg_glucose_level:
library(ggplot2)
# plot a scatter plot
ggplot(data = my_sample_4,aes(x=avg_glucose_level,y=health_bills)) +
  geom_point(aes(x = avg_glucose_level, y = health_bills),
             color = "skyblue3", size = 2) +
  labs(title = "health bills ~ avg glucose level") +
  # put the title location in the center of the plot.
  theme(plot.title = element_text(size=12,face="bold",hjust = 0.5)) +
  theme_classic() +
  # for create a linear model
  geom_smooth(method=lm,se=FALSE,color="orange2")
```

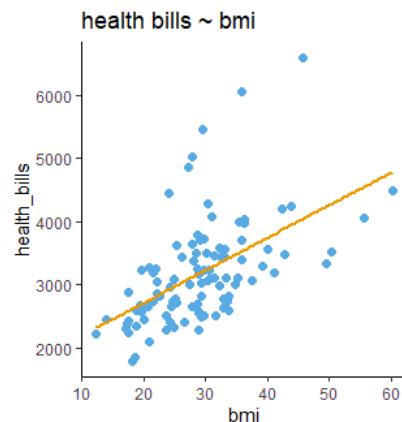
با توجه به این نمودار ، رابطه بین دو متغیر "health\_bills" و "avg\_glucose\_level" خطی است. بنابراین شرط Linearity را برقرار می دانیم.



## health bills ~ bmi

```
# checking the linearity condition for health bills ~ bmi:  
library(ggplot2)  
# plot a scatter plot  
ggplot(data = my_sample_4,aes(x=bmi,y=health_bills)) +  
  geom_point(aes(x = bmi, y = health_bills),  
             color = "skyblue3", size = 2) +  
  labs(title = "health bills ~ bmi") +  
  # put the title location in the center of the plot.  
  theme(plot.title = element_text(size=12,face="bold",hjust = 0.5)) +  
  theme_classic() +  
  # for create a linear model  
  geom_smooth(method=lm,se=FALSE,color="orange2")
```

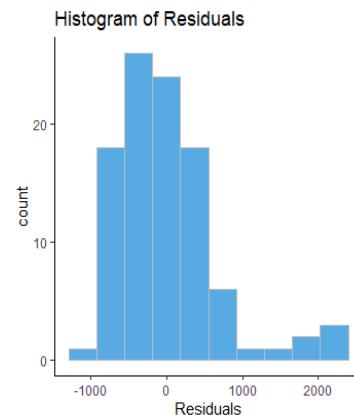
با توجه به این نمودار ، رابطه بین دو متغیر "bmi" و "health\_bills" خطی است. بنابراین شرط **Linearity** را برقرار می دانیم.



شرط **Nearly normal Residuals** : طبق این شرط ، توزیع Residual ها (فاصله نقاط اصلی با نقاط تخمین) باید نرمال باشد. هرچه این توزیع نرمال تر باشد ، مدل Regression ما بهتر است. برای بررسی این شرط در R و به کمک کتابخانه ggplot اقدام به رسم نمودار هیستوگرام می کنیم تا شکل توزیع را ببینیم.

```
# checking the Nearly normal residuals condition  
library(ggplot2)  
#create histogram of residuals  
ggplot(data = my_sample_4, aes(x = final_model$residuals)) +  
  geom_histogram(bins = 10,  
                 fill = 'skyblue3',  
                 color = 'gray') +  
  labs(title = 'Histogram of Residuals', x = 'Residuals') +  
  theme_classic()
```

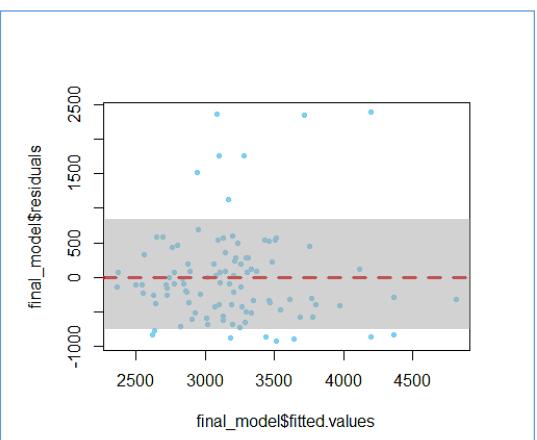
با توجه به این نمودار ، میتوان مشاهده کرد که توزیع چوله به راست است اما با فرض نرمال بودن این توزیع ، این شرط را برقرار میدانیم.



شرط **Constant Variability** : طبق این شرط که به homoscedasticity معروف است ، باید نقاط دو متغیر یک پراکندگی ثابت ، اطراف خط Regression داشته باشند. این شرط را می توان با رسم نمودار زیر در R بررسی کرد.

```
# checking the Constant variability of residuals  
plot(final_model$residuals ~ final_model$fitted.values,  
     pch=20,col="skyblue")  
abline(h=0,col="red2",lwd=3,lty=2)
```

با توجه به این نمودار ، نقاط دو متغیر یک پراکندگی ثابت ، اطراف خط Regression دارند بنابراین می توان گفت این شرط هم برقرار است.



همانطور که گفته شد **Multiple Linear Regression** تحت شرایطی خاص به ما جواب خوبی می دهد. بنابراین با توجه به برقرار بودن شروط لازم برای داشتن یک **Multiple Linear Regression** خوب می توان گفت که مدلی که انتخاب کردیم یک مدل قابل اعتماد و **prediction** خوبی برای متغیر **Response** است.

## 5. Part G

روش اعتبارسنجی **k-fold cross** یک روش برای ارزیابی کردن مدل است که می توان با استفاده از آن تعیین کرد که نتایج یک تحلیل آماری بر روی یک مجموعه داده تا چه اندازه قابل تعمیم و مستقل از داده های آموزشی است. در این روش داده های sample ای که با آن ها مدل های قسمت E و B را ساخته ایم را به 5 قسمت مساوی تقسیم می کنیم، سپس هر بار یکی از این قسمت ها را به عنوان مجموعه Test و بقیه را به عنوان Train در نظر می گیریم. در هر بار مدل را با داده های Train ایجاد می کنیم و سپس، معیار **Root mean square error** (یا بهتر است بگم standard deviation) خطاهای پیش بینی مدل (همان Residual ها) را برای مدل مشخص می کند، را محاسبه می کنم. از این معیار همانند Adjusted R-square میتوان برای مقایسه مدل ها استفاده کرد.

بنابراین در ابتدا داده های sample ای که با آن ها مدل های قسمت E و B را ساخته ایم را به 5 قسمت مساوی تقسیم می کنیم به طوری که یک پنجم داده ها را به مجموعه test اختصاص می دهیم.

```
# to divide my sample into "training" and "test" parts :
my_sample_44 <- sample(nrow(my_sample_4), nrow(my_sample_4)*0.2)
test_part <- my_sample_4[my_sample_44,]
train_part <- my_sample_4[-my_sample_44,]
```

حال با مابقی داده ها، که مجموعه train ما می شوند، اقدام به ساخت مدلی که در قسمت B این سوال ساختیم، میکنیم.

```
# build the model that I created in part B with train data :
model_B <- lm(health_bills ~ avg_glucose_level + age + bmi , data=train_part)
```

و همین کار را دقیقا برای بهترین مدلی که در قسمت E به آن رسیدیم هم انجام می دهیم :

```
# build the model that I created in part E with train data :
model_E <- lm(health_bills ~ avg_glucose_level + bmi , data=train_part)
```

در نهایت باید هر بار یکی از این 5 قسمت را به عنوان مجموعه Test و بقیه را به عنوان Train در نظر بگیریم. سپس در هر بار مدل را با داده های Train ایجاد می کنیم و در نهایت معیار Root mean square error را برای هر دو مدل محاسبه کنیم.

```
# doing predictions with test part with my predictors
# and computing the R2 and RMSE metrics
model_BB <- train(health_bills ~ avg_glucose_level + age + bmi,
                   data = train_part, method = "lm",
                   trControl = trainControl(number = 5,method = "cv"))
# Root mean square error (RMSE) for first model
model_BB$results[2]
```

```
> model_BB$results[2]
      RMSE
1 673.5681
```

برای مدل قسمت B

```
# doing predictions with test part with my predictors
# and computing the R2 and RMSE metrics
model_EE <- train(health_bills ~ avg_glucose_level + bmi,
                   data = train_part, method = "lm",
                   trControl = trainControl(number = 5,method = "cv"))
# Root mean square error (RMSE) for best model
model_EE$results[2]
```

```
> model_EE$results[2]
      RMSE
1 711.1878
```

برای مدل قسمت E

بنابراین مدلی که دارای RMSE کمتر است، دارای خطای کمتری است و بهتر است.

صحبی با دستیار آموزشی محترم : بنده برای فهمیدن مفهوم این سوال به جستجوی اینترنتی پرداختم و از آنجا که هدف این قسمت از سوال، بیشتر مقایسه دو مدل با معیار **Root mean square error** بوده است، ایده حل این سوال را از لینک زیر بدست آوردم و اقدام به محاسبه RMSE کرده ام.

<http://www.sthda.com/english/articles/38-regression-model-validation/157-cross-validation-essentials-in-r/>

برای این سوال متغیر "stroke" binary categorical Response را که نشان دهنده سابقه سکته مغزی (۰ = نداشتن و ۱ = داشتن) برای یک فرد است را به عنوان متغیر Response انتخاب کرده ام. همچنین از سه متغیر "avg\_glucose\_level" و "gender" و همچنین "hypertension" که به ترتیب نشان دهنده متوسط قند خون هر فرد ، داشتن یا نداشتن فشار خون بالا برای آن فرد و همچنین جنسیت آن فرد هستند ، به عنوان متغیر های Explanatory برای تخمین داشتن یا نداشتن سابقه سکته مغزی برای یک فرد استفاده می کنم.

نکته : همانند سوال چهارم و پنجم از آنجا که درون این مجموعه داده حدود 5000 تا داده وجود دارد ، در این سوال فرض را براي گذاشته ام که جامعه هدف همان مجموعه داده است و بنابراین در ابتدا یک نمونه با سایز 100 به صورت کاملاً تصادفی و بدون جایگذاری می گیرم و از همین نمونه برای پاسخ دادن به این سوال استفاده می کنم.

(دستور ایجاد sample در R)

```
# randomly select a sample (n=100) without replacement from HealthCare
my_sample_6 <- HealthCare[sample(nrow(HealthCare), size = 100, replace = FALSE),]
```

## : 6.Part A

از آنجا که متغیر Response در این سوال از نوع binary categorical است. بنابراین اگر بخواهیم با استفاده بقیه متغیرها (که به عنوان predictor در نظر گرفته شوند) یک مدل برای پیش بینی یک بودن یا صفر بودن Response ایجاد کنیم باید از Logistic Regression استفاده کنیم. البته می دانیم که با یک ترکیب خطی نمی توانیم به ۰ و ۱ برسیم. بنابراین برای متغیر Response یک توزیع احتمالاتی گسسته مثل Binomial در نظر می گیریم. سپس یک مدل خطی به صورت زیر در نظر می گیریم که  $X_i$  ها Predictor های ما هستند.

$$\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$$

با استفاده از یک تابع logit ، مدل خطی  $\eta$  را به پارامتر های توزیع احتمال مان تبدیل می کنیم.

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right). \text{ for } 0 \leq p \leq 1$$

حال  $\eta$  را می توان به صورت ترکیب خطی ۰ و ۱ داشت . بنابراین یک توزیع برنولی خواهیم داشت که یک پارامتر  $p$  دارد و این پارامتر احتمال موفقیت (یک بودن) هست و به جای اینکه  $p$  را بر حسب  $X_i$  ها تخمین بزنیم ،  $\log\left(\frac{p}{1-p}\right)$  را بر حسب  $X_i$  ها تخمین می زنیم.

به این کار می گوییم استفاده از logistic regression یکی از این هاست. برای کردن یک مدل logistic fit در R از دستور glm() استفاده می کنیم.

```
> # To fit a logistic Regression model
> log_model<-summary(glm(stroke ~ avg_glucose_level + hypertension + gender
+                               data = my_sample_4 , family = binomial))
> log_model
```

Call:  
`glm(formula = stroke ~ avg_glucose_level + hypertension + gender,  
 family = binomial, data = my_sample_4)`

Deviance Residuals:

Min	1Q	Median	3Q	Max
-0.9312	-0.2765	-0.1981	-0.1693	2.7516

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-5.647358	1.594135	-3.543	0.000396 ***
avg_glucose_level	0.017001	0.009247	1.838	0.065992 .
hypertension	1.330107	1.055701	1.260	0.207696
genderMale	0.268145	1.009072	0.266	0.790444

کد :

مدل :

این جدول به ما می‌گوید که مدلی که با آن پیش‌بینی انجام می‌دهیم بدین صورت است :

$$\log\left(\frac{P}{1-P}\right) = -5.647358 + 0.017001 \times \text{avg\_glucose\_level} + 1.330107 \times \text{hypertension} \\ + 0.268145 \times \text{gender: male}$$

: **تفسیر intercept**

لوگاریتم شانس (**log odds**) داشتن سابقه سکته مغزی برای یک زن که دارای فشار خون بالا نیست ( $\text{hypertension} = 0$ ) و متوسط قند خون آن ۰ است برابر با  $-5.647358$  است.

: **تفسیر slope**

اگر به شبیه متغیر **ave\_glucose\_level** دقت شود ، در میابیم که به ازای هر یک واحد افزایش در قند خون هر فرد ، لوگاریتم شانس (**log odds ratio**) آن فرد  $0.01$  واحد افزایش پیدا میکند.

همچنین اگر بخواهیم نسبت شانس را برای کسانی که یک واحد قند خون بیشتری نسبت به کسی که یک واحد قند خون کمتری دارد (**odds ratio**) بیابیم خواهیم داشت :

$$\log\left(\frac{p_1}{1-p_1}\right) = -5.647358 + 0.017001 \times (\text{avg\_glucose\_level} + 1)$$

$$\log\left(\frac{p_2}{1-p_2}\right) = -5.647358 + 0.017001 \times \text{avg\_glucose\_level}$$

$$\log\left(\frac{p_1}{1-p_1}\right) - \log\left(\frac{p_2}{1-p_2}\right) = 0.017001$$

$$\frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}} = e^{0.017001} = 0.98$$

این نسبت شانس است (**odds ratio**) و گفتیم که با هر واحد افزایش در قند خون یک فرد ، آن فرد به میزان  $0.01$  واحد افزایش میابد.

همچنین اگر به شبیه متغیر **hypertension** دقت شود ، در میابیم که اگر بقیه predictor ها را ثابت در نظر بگیریم ، لوگاریتم شانس داشتن سابقه سکته مغزی برای فردی که دارای فشار خون بالا است ، به میزان  $1.330107$  واحد بیشتر از شانس داشتن سابقه سکته مغزی برای فردی که دارای فشار خون بالا است ، در مقابل فردی که فشار خون بالا ندارد برابر است با :  $e^{1.330107} = 3.78$

همچنین اگر به شبیه متغیر **gender** دقت شود ، در میابیم که اگر بقیه predictor ها را ثابت در نظر بگیریم ، لوگاریتم شانس داشتن سابقه سکته مغزی برای مرد ها ، به میزان  $0.268145$  واحد بیشتر از شانس داشتن سابقه سکته مغزی برای زنان است. و این یعنی نسبت شانس (**odds ratio**) داشتن سابقه سکته مغزی برای مرد ها ، در مقابل زن ها برابر است با :

$$e^{0.268145} = 1.30$$

## : 6.Part B

متغیر انتخابی بندۀ برای این قسمت ، متغیر "hypertension" است که نشان دهنده داشتن یا نداشتن فشار خون بالا است. با توجه به مدل Logistic ساخته شده در قسمت قبل ، اگر تمامی Predictor های دیگر را ثابت نگه داریم ، نسبت شانس (**odds ratio**) داشتن سابقه سکته مغزی برای فردی که دارای فشار خون بالا است ، در مقابل فردی که فشار خون بالا ندارد برابر است با :  $e^{1.330107} = 3.78$

```
> # calculate the odds ratio for selected variable :
> e <- 2.718
> odds_ratio <- e^1.330107
> odds_ratio
[1] 3.780927
```

مقداری که بدست آوردهیم در واقع OR است که از طریق رابطه زیر بدست می‌آید:

$$OR = \frac{P(stroke|hypertension)/[1 - P(stroke|hypertension)]}{P(stroke|no hypertension)/[P(stroke|no hypertension)]} = 3.78$$

ما برای رسم نمودار OR باید بتوانیم مقادیر مختلفی را بین ۰ و ۱ برای یکی از احتمال هایمان در رابطه بالا در نظر بگیریم تا با استفاده از OR که به ما داده است، مقادیر مختلف احتمال دیگر را حساب کنیم. به عنوان مثال در این قسمت بندۀ برای  $P(stroke|no hypertension)$  به تعداد ۵۰ دفعه مقادیر مختلف بین ۰ و ۱ را در نظر می‌گیریم و با استفاده از مقدار OR اقدام به محاسبه احتمال  $P(stroke|hypertension)$  می‌کنم.

حال تمامی این مقادیر را برای  $P(stroke|hypertension)$  و  $P(stroke|no hypertension)$  در دو vector جدا می‌بریم تا بتوانیم Odds ratio curve را برای این مقادیر رسم کنم.

\*\* تمام کارهای توضیح داده شده را در R کد زنی کرده ام و با کامنت گذاری مشخص کرم. کد این قسمت در شکل زیر قابل مشاهده است. لازم به ذکر است که برای رسم نمودار از کتابخانه ggplot استفاده کرده ام.

```
# define 2 empty vectors
vec_1 <- c()
vec_2 <- c()

# to repeat 50 times...
for (i in c(1:50)) {
  # calculate the odds ratio for selected variable :
  e <- 2.718
  odds_ratio <- e^1.330107

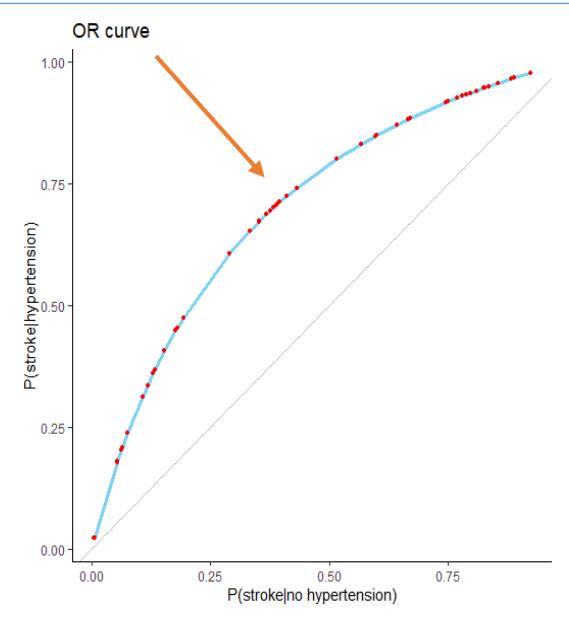
  # generate a random number between 0 and 1 for P(stroke|no hypertension) :
  randd <- runif(1,0,1)

  # calculate the probability of P(stroke|hypertension) :
  P_stroke_hypertension <- ((odds_ratio * (randd/(1-randd))) / (1 + (odds_ratio * (randd/(1-randd)))))

  # save all values to two vectors :
  vec_1 <- c(vec_1,randd)
  vec_2 <- c(vec_2,P_stroke_hypertension)
}

daf <- data.frame(vec_1,vec_2)

# plot the OR curve :
library(ggplot2)
ggplot(data=daf, aes(x=vec_1, y=vec_2, group=1)) +
  # OR curve
  geom_line(color="skyblue",size=1.2) +
  # y = x curve
  geom_abline(slope=1, intercept=0,color="gray") +
  geom_point(color="red", size=1) +
  # to put titles
  labs(title = 'OR curve', y = 'P(stroke|hypertension)', x = 'P(stroke|no hypertension)') +
  theme_classic()
```



#### تفسیر نمودار:

همانطور که مشاهده می‌شود این نمودار شامل ۵۰ مقدار برای احتمال داشتن سابقه سکته مغزی به شرط نداشتن فشار خون بالا ( $P(stroke|no hypertension)$ ) است که در محور x نمایش داده شده است و مقادیر حساب برای احتمال داشتن سابقه سکته مغزی به شرط داشتن فشار خون بالا ( $P(stroke|hypertension)$ ) را متناظر می‌کند. همانطور که می‌بینیم یک شکل منحنی مانند شده است به دلیل اینکه OR با مقدار ۱ فاصله دارد. اگر OR برابر با ۱ می‌شد مشخصاً این curve خطی می‌شد.

تا به الان دیدیم که با استفاده از یک مدل Logistic Regression می توانیم برای یک متغیر binary categorical شناس یک بودن آن بر حسب ترکیب خطی یکسری متغیر Explanatory بنویسیم. و همانطور که گفته شد می توانیم از روی لوگاریتم شناس ، خود شناس و احتمال یک بودن متغیر Response را بدست بیاوریم. اما این برای پیش بینی کافی نیست و ما نیاز به یک Threshold داریم تا با استفاده از آن و احتمالی که بدست می آوریم تعیین کنیم که طبق پیش بینی این مدل ، متغیر ما مقدار 1 می گیرد یا مقدار 0. اما برای تعیین Threshold مناسب باید بتوانیم مقادیر Sensitivity و Specificity را در مورد Specificity مقدار Threshold های مختلف محاسبه کنیم و در نهایت با در نظر گرفتن مقدار Specificity و Sensitivity مورد نظرمان ، حد آستانه مناسب تست را انتخاب کنیم. و می دانیم که بین این دو مقدار Specificity و Sensitivity Trade off پک است. نمودار ROC با نمایش رابطه بین Specificity و Sensitivity ، به ما برای انتخاب حد آستانه مناسب کمک می کند. همچنین این نمودار Performance مدل ایجاد شده را با تصمیم گیری کامل شناسی مقایسه می کند. همچنین با استفاده از این نمودار می توانیم مقدار محاسبت زیر منحنی ROC که به آن AUC گفته می شود را به عنوان معیاری برای بررسی اینکه مدلی که ساختیم خوب است یا خیر، در نظر بگیریم و برای مقایسه مدل های از آن استفاده کنیم.

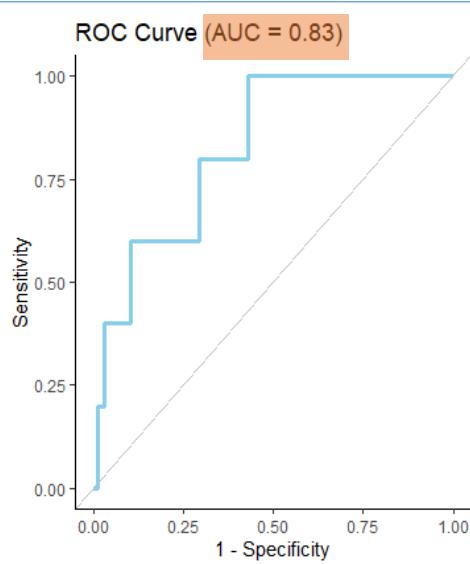
برای این که برای مدل Logistic Regression ایجاد شده در قسمت A این سوال ، این نمودار مهم را رسم کنیم ، در R از پکیج plotROC و کتابخانه ggplot برای رسم این نمودار استفاده می کنیم. کد این نمودار و نتیجه حاصل از اجرای کد و همچنین محاسبه مساحت زیر منحنی ROC در شکل زیر قابل مشاهده است :

```
#load necessary packages
library(ggplot2)
library(plotROC)
library(pROC)

#use model to make predictions
log_model22 <- glm(stroke ~ avg_glucose_level + hypertension + gender,
                     data = my_sample_4 , family = binomial)
my_per <- predict(log_model22,
                    my_sample_4,
                    type="response")

#create ROC curve with predictions and calculate the AUC
ROC_curve <- roc(my_sample_4$stroke, my_per)
AUC <- round(auc(my_sample_4$stroke, my_per),2)

#to create ROC plot
ggroc(ROC_curve, colour = 'skyblue', legacy.axes = TRUE, size = 1.2) +
  # # to put titles
  labs(title = paste0('ROC Curve ', '(AUC = ', AUC, ')'), y = 'sensitivity' ,
       x = '1 - specificity') +
  # y = x curve
  geom_abline(slope=1, intercept=0, color="gray")+
  theme_classic()
```



همانطور که بالاتر هم توضیح دادم ، این نمودار رابطه بین Specificity و Sensitivity که گفته شد این منحنی یک معیار مناسب برای تعیین اینکه این Logistic classifier خوبی است یا خیر. مساحت زیر این منحنی در واقع همان AUC این مدل را نشان می دهد که از این معیار برای مقایسه دو مدل Logistic استفاده می شود. مشخصا هر چه این مساحت بیشتر باشد (هر چه از خط  $y=x$  منحنی دورتر باشد) یعنی مدل بهتری داریم و پرiformنس بیشتری نسبت به تصمیم گیری کاملا شناسی دارد.

مقدار AUC محاسبه شده برای این مدل ، تقریباً مقدار خوبی است و نشان از خوب بود مدل Logistic است. همانطور که شده دارد. می توان با حذف یک متغیر از مدل ، مجدداً AUC مدل را محاسبه کرد و اگر کاهش یافت بنابراین نباید متغیر را حذف کنیم. ما به دنبال AUC مدل های Logistic با AUC بالاتر هستیم. بنابراین این مدل به دلیل داشتن AUC بالا ، تقریباً مدل خوبی است.

## 6. Part D

می دانیم که برای بررسی کردن اینکه آیا predictor های ما در مدلی که ساخته ایم significant می شوند یا خیر می توانیم از آزمون فرض (برای کل مدل یا برای تک تک slop ها) استفاده کنیم. می دانیم که اگر شیب خط رگرسیون ما صفر بشود یعنی آن متغیر آنکه انتخاب کردیم ، predictor خوبی نیست اما اگر از صفر فاصله بگیرد predictor خوبی است. بنابراین در انجام آزمون فرض برای slop ما به دنبال این هستیم که آیا شیب خط رگرسیون صفر می شود یا مخالف با صفر است. در واقع فرض صفر و فرض جایگزین در این آزمون فرض را این گونه تعریف می کنیم :

$H_0 : \beta_1 = 0$  متغیر Explanatory در پیش بینی متغیر Response ، متغیر خوبی نیست.

$H_A : \beta_1 \neq 0$  متغیر Explanatory در پیش بینی متغیر Response ، متغیر خوبی است.

سپس آماره آزمون Z و p-value را حساب می کنیم. اما جدول مدل Logistic ای که در قسمت A بدست آورده ایم این مقادیر را برای ما محاسبه کرده است. بنابراین از مقدار p-value که توسط خود مدل محاسبه شده است استفاده می کنیم و در نهایت به بررسی significant بودن یا نبودن آن متغیر با  $a = 0.05$  می پردازیم :

مدل Logistic Regression که بر اساس متوسط قند خون (سطح گلوکز) هر فرد ، داشتن یا نداشتن فشار خون بالا و جنسیت هر فرد اقدام به پیش بینی داشتن یا نداشتن سابقه سکته مغزی برای آن فرد می کند :

```
> # To fit a logistic Regression model
> log_model<-summary(glm(stroke ~ avg_glucose_level + hypertension + gender
+ 
+                               , data = my_sample_4 , family = binomial))
> log_model

call:
glm(formula = stroke ~ avg_glucose_level + hypertension + gender,
      family = binomial, data = my_sample_4)

Deviance Residuals:
    Min      1Q   Median      3Q      Max
-0.9312 -0.2765 -0.1981 -0.1693  2.7516

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.647358  1.594135 -3.543 0.000396 ***
avg_glucose_level 0.017001  0.009247  1.838 0.065992 .
hypertension   1.330107  1.055701  1.260 0.207696
genderMale     0.268145  1.009072  0.266 0.790444
```

همانطور که گفته شد برای این مدل از دو متغیر Explanatory به نام "hypertension" و "gender" استفاده شده است که اینکه مقدار p-value ای با توجه به مقدار جدول بدست آمده است ، هر دو متغیر significant نیستند به دلیل اینکه مقدار p-value آن ها از مقدار آلفا (0.05) بیشتر شده است. اما می توان مشاهده کر که متغیر "hypertension" به دلیل مقدار p-value کمتر نسبت به متغیر "gender" یک Predictor معنی دار تر (تر) است.

بنابراین متغیر hypertension که برای هر فرد مشخص می کند که فشار خون بالای دارد یا خیر ، در این مدل به عنوان Explanatory متغیر ن نقش بیشتری در پیش بینی داشتن یا نداشتن سابقه سکته مغزی برای یک فرد دارد ، شناخته می شود.

## 6. Part E

در قسمت قبل مشاهده کردیم که متغیر hypertension که برای هر فرد داشتن یا نداشتن فشار خون بالا را مشخص میکند ، به عنوان Explanatory متغیر که نقش مهمتری در پیش بینی داشتن یا نداشتن سابقه سکته مغزی برای یک فرد ایفا میکند انتخاب شد. حال می خواهیم در این قسمت مدل Logistic Regression جدید مان را با استفاده از این متغیر ایجاد کنیم.

برای *fit* کردن این مدل logistic در R از دستور *glm()* استفاده می کنیم.

(نتیجه در صفحه بعد است)

```

> # To fit a logistic Regression model
> log_model11<-summary(glm(stroke ~ hypertension ,
+                               data = my_sample_4 , family = binomial))
> log_model11

call:
glm(formula = stroke ~ hypertension, family = binomial, data = my_sample_4)

Deviance Residuals:
    Min      1Q  Median      3Q     Max 
-0.6335 -0.2619 -0.2619 -0.2619  2.6038 

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -3.3557    0.5873 -5.714 1.11e-08 ***  
hypertension  1.8517    0.9778  1.894   0.0583 .    

```

کد :  
مدل :

این جدول به ما می‌گوید که معادله مدلی که با آن پیش‌بینی انجام می‌دهیم بدین صورت است :

$$\log \left( \frac{P}{1 - P} \right) = -3.3557 + 1.8517 \text{ hypertension}$$

همانطور که مشاهده می‌شود ، متغیر خوبی است و در این مدل جدید برعکس مدل قبل ، significant شده است. بنابراین یکی از اولین نتایجی که میشه گرفت متغیر gender برای پیش‌بینی اصلاً متغیر خوبی نیست و با حذف آن می‌توان به مدل بهتری دست پیدا کرد.

حال اگر بخواهیم این دو مدل را پس از حذف متغیر gender با هم مقایسه کنیم نیاز به محاسبه مقدار AUC برای هر دو مدل داریم.

```

# to compare two logistic models with AUC :
logmodelnum1 <- glm(stroke ~ avg_glucose_level + hypertension + gender ,
                      data = my_sample_4 , family = binomial)

logmodelnum2 <- glm(stroke ~ avg_glucose_level + hypertension ,
                      data = my_sample_4 , family = binomial)

my_per_num1 <- predict(logmodelnum1,
                        my_sample_4,
                        type="response")
my_per_num2 <- predict(logmodelnum2,
                        my_sample_4,
                        type="response")

# calculate the AUC
AUC_num1 <- round(auc(my_sample_4$stroke, my_per_num1),4)
AUC_num2 <- round(auc(my_sample_4$stroke, my_per_num2),4)

```

> AUC\_num1  
[1] 0.8253  
> AUC\_num2  
[1] 0.8253

با توجه به اینکه مقدار AUC مدل جدید (مدلی که با متغیر Explanatory significant تر بود) پس از حذف متغیر gender تغییر نکرده است. پس می‌توان نتیجه گرفت که متغیر gender اضافی است و بهر است از مدل حذف شود. و همان متغیر hypertension ، برای این مدل Predictor مهم تری است. ☺

: 6.Part F

در این سوال می خواهیم با استفاده از یک مدل Logistic Regression برای هر فرد پیش بینی کنیم که آیا هزینه های بالای را برای سلامتی خود متحمل می شود یا خیر.

در ابتدا لازم است تا یک Threshold برای این کار تعیین کنیم. حد آستانه ای که در این سوال در نظر گرفته ام ، میانه ی مقادیر health\_bills است. از آنجا که این متغیر درون مجموعه داده HealthCare نشان دهنده میزان هزینه سالیانه ای است که هر فرد برای سلامتی خود باید پردازد ، بنابراین بهترین گزینه برای تعیین حد آستانه مورد نظر در این سوال است ، همچنین از آنجا که میانگین ، یک آماره Non-robust است و به شدت به outlier ها حساس است ، از آماره میانه برای تعیین Threshold موردنظر استفاده میکنم. بنابراین Threshold انتخابی بندۀ برای این سوال میانه هزینه سالیانه ای است که افراد این مجموعه داده برای سلامتی خود باید پردازند و برای محاسبه این مقدار در R ، ابتدا میانه این متغیر را پیدا کرده و درون متغیر Threshold می ریزم.

\*\* نکته : طبق آن چیزی که در فاز اول این پروژه عنوان شد ، متغیر health\_bills دارای مقادیر missing ( N/A values ) میباشد و رویکرد من برای مقادیر گمشده در این متغیر ، استفاده از روش جایگذاری میانه به جای مقادیر گم شده است. بنابراین در ابتدا این عمل را روی مجموعه داده HealthCare در R با استفاده از دستور زیر انجام می دهیم. در واقع هر سطر از مجموعه داده HealthCare که برای متغیر ذکر شده مقدار گم شده داشته باشد ، برای آن سطر میانه مقادیری که آن متغیر دارد را جایگذاری میکنم. (که این کار قبلا انجام شده است).

```
# at first lets replace missing values with the mean
HealthCare2 <- HealthCare
HealthCare2$health_bills[is.na(HealthCare2$health_bills)]<-median(HealthCare2$health_bills,na.rm=TRUE)
```

(کد تعیین Threshold در R)

```
> # Get the median of health_bills variable in dataset
> # and specifying the Threshold :
> Threshold <- median(HealthCare2$health_bills)
> Threshold
[1] 3031.724
```

حال می خواهیم یک متغیر binary categorical به نام "high\_medical\_costs" به مجموعه داده HealthCare اضافه کنیم که برای هر فرد درون این مجموعه داده اگر هزینه ای که برای سلامتی خود پرداخت می کند از Threshold تعیین شده بیشتر شد مقدار 1 و اگر کمتر شد مقدار 0 را به خود بگیرد. بنابراین در R نیاز به تعریف یک تابع دارم که به کمک تابع lapply هر زمان برای یک سطر از مجموعه داده HealthCare صدا زده شد ، برای هر فرد که مقدار متغیر health\_bills آن از این Threshold تعیین شده بیشتر شد مقدار 1 را برگرداند در غیر این صورت مقدار 0 را برگرداند. (برای افرادی که مقدار متغیر health\_bills آنها دقیقا با Threshold تعیین شده برابر است هم مقدار 0 در نظر گرفته می شود).

```
# To determine whether a person is in category 1 or in category 0.
high_cost <- function(inp){
  if(inp > Threshold){
    return(1)
  }else{
    return(0)
  }
}

# Add a new column to dataset and named it "high_medical_costs"
# call function on each row of the dataset with lapply.
# To determine whether a person is in category 1 or in category 0 in new variable.
HealthCare2$high_medical_costs <- lapply(HealthCare2$health_bills,high_cost)
```

ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke	health_bills	high_medical_costs
Yes	Private	Urban	228.69	36.6	formerly smoked	1	6011.860	1
Yes	self-employed	Rural	202.21	NA	never smoked	1	3031.724	0
Yes	Private	Rural	105.92	32.5	never smoked	1	6384.530	1
Yes	Private	Urban	171.23	34.4	smokes	1	5862.754	1
Yes	self-employed	Rural	174.12	24.0	never smoked	1	5461.262	1
Yes	Private	Urban	186.21	29.0	formerly smoked	1	5054.021	1
Yes	Private	Rural	70.09	27.4	never smoked	1	6795.934	1
No	Private	Urban	94.39	22.8	never smoked	1	5158.242	1

کد در R :

قسمتی از مجموعه داده  
پس از اجرای کد و  
مشاهده نتیجه :

حال می خواهیم با استفاده از یک مدل Logistic Regression برای هر فرد پیش بینی کنیم که آیا هزینه های بالای را برای سلامتی خود متحمل می شود یا خیر.

برای این کار متغیر Response را همان متغیر جدیدی که به مجموعه داده اضافه کردیم در نظر می گیریم (یعنی متغیر "stroke" و متغیر های Explanatory (high\_medical\_costs) را به ترتیب : "stroke" ، "heart\_disease" و "age" در نظر گرفته شده اند. (در واقع می خواهیم با استفاده از سن ، داشتن یا نداشتن سابقه بیماری قلبی ، سابقه سکته مغزی و داشتن یا نداشتن فشار خون بالا ، یک تخمین برای هر فرد بزنیم و پیش بینی کنیم که آیا هزینه های بالای را برای سلامتی خود متحمل می شود یا خیر).

برای fit کردن یک مدل logistic در R از دستور `glm()` استفاده می کنیم.

```
> # To fit a Logistic Regression model
> log_model_q7 <- summary(glm(unlist(high_medical_costs) ~ hypertension + age +
+                                         stroke + heart_disease ,
+                                         data = HealthCare2 , family = binomial))
> log_model_q7

call:
glm(formula = unlist(high_medical_costs) ~ hypertension + age +
stroke + heart_disease, family = binomial, data = HealthCare2)

Deviance Residuals:
Min      1Q   Median      3Q      Max
-2.0682 -1.1011 -0.8753  1.1856  1.5999

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.789636  0.064125 -12.314 < 2e-16 ***
hypertension 0.229553  0.102871  2.231  0.0256 *
age          0.014816  0.001405 10.547 < 2e-16 ***
stroke       1.403086  0.177304  7.913  2.5e-15 ***
heart_disease -0.193949  0.134150 -1.446  0.1482
```

: کد

: مدل

این جدول به ما می گوید که مدلی که با آن پیش بینی انجام می دهیم بدین صورت است :

$$\log \left( \frac{P}{1-P} \right) = -0.789636 + 0.229553 \times \text{hypertension} + 0.014816 \times \text{age} \\ + 1.403086 \times \text{stroke} - 0.193949 \times \text{heart_disease}$$

: تفسیر *intercept*

لوگاریتم شانس (*log odds*) متحمل شدن هزینه بالا برای سلامتی برای یک فرد که دارای فشار خون بالا نیست (*hypertension = 0*) و سن آن برابر با 0 است و تا به حال سابقه سکته مغزی و سابقه بیماری قلبی هم نداشته است برابر با 0.789636 است.

: تفسیر *slop*

اگر به شبیه متغیر *age* دقت شود ، در میابیم که به ازای هر یک سال افزایش در سن هر فرد ، لوگاریتم شانس (*log odds*) آن فرد 0.014814 واحد افزایش پیدا میکند.

اگر به شبیه متغیر *heart\_disease* دقت شود ، در میابیم که اگر بقیه predictor ها را ثابت در نظر بگیریم ، لوگاریتم شانس متحمل شدن هزینه بالای سلامتی برای فردی که دارای سابقه بیماری قلبی است ، به میزان 0.193949 واحد کمتر از شانس متحمل شدن هزینه بالای سلامتی برای فردی که سابقه بیماری قلبی نداشته است. و این یعنی نسبت شانس (odds ratio) داشتن متحمل شدن هزینه بالای سلامتی برای فردی که دارای سابقه بیماری قلبی است ، در مقابل فردی که دارای سابقه بیماری قلبی نیست ، برابر است با :  $e^{-0.193949} = 0.82$

همچنین اگر به شبیه متغیر *hypertension* دقت شود ، در میابیم که اگر بقیه predictor ها را ثابت در نظر بگیریم ، لوگاریتم شانس متحمل شدن هزینه بالای سلامتی برای فردی که دارای فشار خون بالا است ، به میزان 0.229553 واحد بیشتر از شانس متحمل شدن هزینه بالای سلامتی برای فردی که فشار خون بالا ندارد است. و این یعنی نسبت شانس (odds ratio) داشتن متحمل شدن هزینه بالای سلامتی برای فردی که دارای فشار خون بالا است ، در مقابل فردی که فشار خون بالا ندارد برابر است  $e^{0.229553} = 1.25$

در نهایت اگر به شبیه متغیر **stroke** دقیق شود ، در میابیم که اگر بقیه **predictor** ها را ثابت در نظر بگیریم ، لوگاریتم شانس متحمل شدن هزینه بالای سلامتی برای فردی که دارای سابقه سکته مغزی است ، به میزان 1.403086 واحد بیشتر از شانس (odds ratio) داشتن متحمل شدن هزینه بالای سلامتی برای فردی که دارای سابقه سکته مغزی است ، در مقابل فردی که دارای سابقه سکته مغزی نیست ، برابر است با :  $e^{1.403086} = 4.06$

```
> # To fit a logistic Regression model
> log_model_q7 <- summary(glm(unlist(high_medical_costs) ~ hypertension + age +
+                                         stroke + heart_disease ,
+                                         data = Healthcare2 , family = binomial))
> log_model_q7

Call:
glm(formula = unlist(high_medical_costs) ~ hypertension + age +
    stroke + heart_disease, family = binomial, data = Healthcare2)

Deviance Residuals:
    Min      1Q   Median      3Q      Max 
-2.0682 -1.1011 -0.8753  1.1856  1.5999 

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -0.789636  0.064125 -12.314 < 2e-16 ***
hypertension  0.229553  0.102871   2.231  0.0256 *  
age          0.014816  0.001405  10.547 < 2e-16 ***
stroke       1.403086  0.177304   7.913 2.5e-15 ***
heart_disease -0.193949  0.134150  -1.446  0.1482    

```

با توجه به اینکه از بین چهار متغیر اختحابی بنده به عنوان Predictor در این سوال ، فقط سه تا از آنها مقدار کمتر از 0.05 دارند و p-value می شوند ، بنابراین متغیر significant که **heart\_disease** نشده است متغیر اثرگذار و خوبی نیست اما از بین سه متغیر دیگر ، متغیر **age** به دلیل داشتن p-value کمتر از دو متغیر دیگر ، تاثیرگذارترین متغیر برای پیش بینی اینکه هر فرد آیا هزینه های بالای را برای سلامتی خود متحمل می شود یا خیر، میباشد.