بسمه‌تعالی

دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر

درس استنباط آماری

پروژه_فازدوم

تیر۱۴۰۰

زهرا محقق راد

۸۱۰۱۹۹۲۶۰

فهرست:

فهرست جداول:

فهرست تصاویر:

Question 1

برای این سوال، دو متغیر categoricalای که در نظر گرفته شده‌است، متغیرهای 'work type' و 'Residence type' می‌باشد. که متغیر 'work type' دارای 5 سطح و متغیر 'Residence type' دارای 2 سطح می‌باشد.

```
> levels(factor(data$work_type))
[1] "children"      "Govt_job"      "Never_worked"  "Private"       "Self-employed"
> levels(factor(data$Residence_type))
[1] "Rural" "Urban"
>
```

(A

ابتدا از متغیر 'work_type' یک سطح را انتخاب می‌کنیم، که در این قسمت سطح 'self_employed' انتخاب شده‌است. سپس بررسی می‌کنیم چه درصدی از افرادی که در دسته 'Self_employed' هستند، شهری(urban) بوده و چه درصدی روستایی(rural) می‌باشند. همچنین چه درصد از افراد شهری و روستایی در دسته‌ی 'Self-employed' قرار نمی‌گیرند. به همین منظور یک دیتاست جدا با استفاده از این دو متغیر درست کردیم. در قدم بعد یک ستون به این دیتاست اضافه کردیم و به افرادی که 'Self-employed' بودند، برچسب 'y' و به بقیه نوع کاری برچسب 'n' اختصاص دادیم. نسبت‌های به دست‌آمده برای این گروه در جدول(۱) آورده شده‌است. لازم به ذکر است که برای این سوال از کل دیتاست استفاده شده‌است. (همین عمل را به ازای هر 5 سطح موجود در متغیر 'work_type' کردیم که در قسمت‌های بعد نتیجه‌ی آن‌ها آورده شده‌است)

```
df_a <- data[, c(7,8)]
df_a$new <- ifelse(df_a$work_type == "Self-employed", 'y','n')
table_prop <- addmargins(table( df_a$new, df_a$Residence_type))
```

جدول(۱): نسبت افراد در گروه 'Self-employed'

	Rural	Urban
No	2121	2170
Yes	393	426
Sum	2514	2596
proportion	0.156	0.164

بررسی شرایط قضیه حد مرکزی:

1- independence:

- استقلال درون گروهی: random sampling/assignment داشته باشیم. سایز نمونه از 10% جامعه آماری کمتر است.
- استقلال بین گروهی: این استقلال نیز برقرار است، زیرا افراد نمی‌توانند همزمان در دو گروه قرار گیرند.

2- sample size:

$2514 \times 0.156 \geq 392.184 \geq 10$ and $2514 \times 0.844 \geq 2{,}121.816 \geq 10$

$2596 \times 0.164 \geq 425.744 \geq 10$ and $2596 \times 0.836 \geq 2{,}170.256 \geq 10$

حال به محاسبه‌ی بازه‌ی اطمینان می‌پردازیم:

$$CI = (\hat{P}_1 - \hat{P}_2) \pm Z^* \, SE_{(\hat{P}_1 - \hat{P}_2)} \qquad SE = \sqrt{\frac{\hat{P}_1 (1 - \hat{P}_1)}{n_1} + \frac{\hat{P}_2 (1 - \hat{P}_2)}{n_2}}$$

برای محاسبه‌ی بازه‌ی اطمینان یک تابع به نام compute_CI مطابق با فرمول بالا پیاده‌سازی کردیم که به شرح زیر می‌باشد:

```
compute_CI <- function(P_hat1, P_hat2){

  n1 <- 2514
  n2<- 2596

  SE <- sqrt(((P_hat1 * (1 - P_hat1)) / n9) + ((P_hat2 * (1 - P_hat2)) / n10))
  left <- (P_hat1 - P_hat2) - (-qnorm(0.025) * SE)
  right <- (P_hat1 - P_hat2) + (-qnorm(0.025) * SE)

  return(c(left, right))

}
```

→ CI = (-0.0123, 0.02789)

با توجه به بازه‌ی اطمینان به‌دست آمده، 95% اطمینان داریم که اختلاف بین proportion افرادی که شهری هستند و نوع شغلشان 'Self-employed' است با افرادی که روستایی هستند و نوع شغل آن‌ها نیز 'Self_employed' است در بازه‌ی (-0.0123, 0.02789) قرار می‌گیرد.

حال همین عمل را با سطح دیگری از متغیر 'work_type' انجام می‌دهیم. برای این قسمت از سطح 'Private' استفاده شده‌است. نسبت‌های به دست‌آمده برای این گروه در جدول(۲) آورده شده‌است. از ذکر مجدد کدهای این قسمت صرف‌نظر شده است و تنها نتیجه نوشته داده شده‌است.

جدول(۲): نسبت افراد در گروه 'Private'

	Rural	Urban
No	1052	1133
Yes	1462	1463
Sum	2514	2596
proportion	0.582	0.56

بررسی شرایط قضیه حد مرکزی:

1- independence:

- استقلال درون گروهی: random sampling/assignment داشته باشیم. سایز نمونه از 10% جامعه آماری کمتر است.
- استقلال بین گروهی: این استقلال نیز برقرار است، زیرا افراد نمی‌توانند همزمان در دو گروه قرار گیرند.

2- sample size:

$2514 \times 0.582 \geq 1{,}463.148 \geq 10$ and $2514 \times 0.418 \geq 1{,}050.852 \geq 10$

$2596 \times 0.56 \geq 1{,}453.76 \geq 10$ and $2596 \times 0.44 \geq 1{,}142.24 \geq 10$

→ CI = (-0.00914, 0.045)

با توجه به بازه‌ی اطمینان به‌دست آمده، 95% اطمینان داریم که اختلاف بین proportion افرادی که شهری هستند و نوع شغلشان 'Private' است با افرادی که روستایی هستند و نوع شغل آن‌ها نیز 'Private' است در بازه‌ی (-0.00914,0.045) قرار می‌گیرد.

حال همین عمل را با سطح دیگری از متغیر 'work_type' انجام می‌دهیم. برای این قسمت از سطح 'Never_worked' استفاده شده‌است. نسبت‌های به دست‌آمده برای این گروه در جدول(۳) آورده شده‌است. از ذکر مجدد کدهای این قسمت صرف‌نظر شده است و تنها نتیجه نوشته داده شده‌است.

جدول(۳): نسبت افراد در گروه 'Never_worked'

	Rural	Urban
No	2507	2581
Yes	7	15
Sum	2514	2596
proportion	0.0028	0.00578

بررسی شرایط قضیه حد مرکزی:

1- independence:

- استقلال درون گروهی: random sampling/assignment داشته باشیم. سایز نمونه از 10% جامعه آماری کمتر است.
- استقلال بین گروهی: این استقلال نیز برقرار است، زیرا افراد نمی‌توانند همزمان در دو گروه قرار گیرند.

2- sample size:

$2514 \times 0.0028 > 7$ and $2514 \times 0.9972 > 2507 \geq 10$

$2596 \times 0.00578 > 15 \geq 10$ and $2596 \times 0.99422 > 2581 \geq 10$

باوجود آنکه مقدار اول برابر 7 شد، اما فرض می‌کنیم که شرایط برقرار است و بازه‌ی اطمینان را محاسبه می‌کنیم.

→ CI = (-0.000596, 0.00658)

با توجه به بازه‌ی اطمینان به‌دست آمده، 95% اطمینان داریم که اختلاف بین proportion افرادی که شهری هستند و نوع شغلشان 'Never_worked' است با افرادی که روستایی هستند و نوع شغل آن‌ها نیز 'Never_worked' است در بازه‌ی (-0.000596, 0.00658) قرار می‌گیرد.

حال همین عمل را با سطح دیگری از متغیر 'work_type' انجام می‌دهیم. برای این قسمت از سطح 'Children' استفاده شده‌است. نسبت‌های به دست‌آمده برای این گروه در جدول(۴) آورده شده‌است. از ذکر مجدد کدهای این قسمت صرف‌نظر شده است و تنها نتیجه نوشته داده شده‌است.

جدول(۴): نسبت افراد در گروه 'Children'

	Rural	Urban
No	2174	2249
Yes	340	347
Sum	2514	2596
proportion	0.135	0.1337

بررسی شرایط قضیه حد مرکزی:

1- independence:

- استقلال درون گروهی: random sampling/assignment داشته باشیم. سایز نمونه از 10% جامعه آماری کمتر است.
- استقلال بین گروهی: این استقلال نیز برقرار است، زیرا افراد نمی‌توانند همزمان در دو گروه قرار گیرند.

2- sample size:

$2514 \times 0.135 > 340 \geq 10$ and $2514 \times 0.865 > 2174 \geq 10$

$2596 \times 0.1337 > 347 \geq 10$ and $2596 \times 0.866 > 2249 \geq 10$

→ CI = (-0.017, 0.0203)

با توجه به بازه‌ی اطمینان به‌دست آمده، 95% اطمینان داریم که اختلاف بین proportion افرادی که شهری هستند و نوع شغلشان 'Children' است با افرادی که روستایی هستند و نوع شغل آن‌ها نیز 'Children' است در بازه‌ی (-0.017, 0.0203) قرار می‌گیرد

در نهایت برای سطح آخر یعنی 'Govt_job' بازه‌ی اطمینان را محاسبه می‌کنیم. نسبت‌های به دست‌آمده برای این گروه در جدول(۵) آورده شده‌است. از ذکر مجدد کدهای این قسمت صرف‌نظر شده است و تنها نتیجه نوشته داده شده‌است.

جدول(۵): نسبت افراد در گروه 'Govt_job'

	Rural	Urban
No	2202	2251
Yes	312	345
Sum	2514	2596
proportion	0.124	0.133

بررسی شرایط قضیه حد مرکزی:

1- independence:

- استقلال درون گروهی: random sampling/assignment داشته باشیم. سایز نمونه از 10% جامعه آماری کمتر است.
- استقلال بین گروهی: این استقلال نیز برقرار است، زیرا افراد نمی‌توانند همزمان در دو گروه قرار گیرند.

2- sample size:

$2514 \times 0.124 > 312 \geq 10$ and $2514 \times 0.876 > 2202 \geq 10$

$2596 \times 0.133 > 345 \geq 10$ and $2596 \times 0.866 > 2251 \geq 10$

→ CI = (-0.0096, 0.027)

با توجه به بازه‌ی اطمینان به‌دست آمده، 95% اطمینان داریم که اختلاف بین proportion افرادی که شهری هستند و نوع شغلشان 'Govt_job' است با افرادی که روستایی هستند و نوع شغل آن‌ها نیز 'Govt_job' است در بازه‌ی (-0.0096, 0.027) قرار می‌گیرد.

(B

برای انجام این تست از کل دیتاست استفاده شده است.

H_0 : (nothing going on) → Residence types and work types are independent, work types do not vary by residence types.

H_A : (nothing going on) → Residence types and work types are dependent, work types vary by residence types.

جدول(۷) نتیجه‌ی حاصل از این دو متغیر به شرح زیر می‌باشد:

```
two_catg_table <- table(data$Residence_type, data$work_type)
addmargins(two_catg_table)
```

جدول(۷): توزیع افراد براساس دو متغیر 'work_type' و 'residence_type'

	Children	Govt_job	Never_worked	Private	Self_employed	Total
Rural	340	312	7	1462	393	2514
Urban	347	345	15	1463	426	2596
Total	687	657	22	2925	819	5110

در قدم اول شرایط chi-square را بررسی می‌کنیم:

۱- **شرط independence** : random sampling/assignment داشته‌باشیم، سایز نمونه از 10% جامعه‌آماری کمتر باشد و هر case فقط به یک خانه از جدول contribute کند.

۲_ **شرط sample size** : هرکدام از خانه‌های جدول باید تعدادی بیشتر از 5 داشته‌باشند.

چون شرایط بالا برقرار می‌باشد، با استفاده از تابع chisq.test() بررسی می‌کنیم که آیا این دو متغیر مستقل از هم می‌باشند یاخیر. خروجی تست chi square در تصویر (۱) قابل مشاهده می‌باشد.

```
chisq_indp <- chisq.test(two_catg_table)
```

```
        Pearson's Chi-squared test

data:  two_catg_table
X-squared = 4.6533, df = 4, p-value = 0.3248
```

تصویر(۱): خروجی تابع chi.test() برای بررسی استقلال

همانطور که در تصویر بالا قابل مشاهده می‌باشد، مقدار p-value به‌دست آمده برابر 0.3248 بوده که در مقایسه با $\alpha = 0.05$ بیشتر می‌باشد. بنابراین نمی‌توان فرض H_0 را رد کرد و به عبارتی تلویحا می‌پذیریم که این دو متغیر 'residence type' و 'work type' مستقل از هم می‌باشند.

مقدار Expected به دست‌آمده برای بررسی این دو متغیر در تصویر(۲) قابل مشاهده می‌باشد:

```
chisq_indp$expected
```

	children	Govt_job	Never_worked	Private	Self-employed
Rural	337.9879	323.2286	10.82348	1439.031	402.9288
Urban	349.0121	333.7714	11.17652	1485.969	416.0712

تصویر(۲): مقادیر expected به‌دست آمده با توجه به جدول(۷)

Question 2

برای این قسمت متغیر "hypertesion" در نظر گرفته شده‌است. ابتدا یک سمپل 15تایی به صورت رندم از این دیتاست برمی‌داریم:

```
set.seed(123)
small_sample <- data.frame(data[sample(nrow(data), 15),c(4)])
colnames(small_sample) <- c("hypertension")
```

آزمون فرض به این صورت تعریف می‌شود:

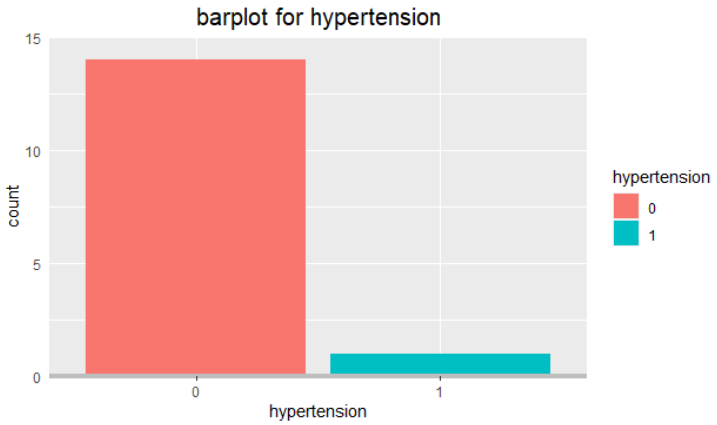
$H_0 : P = 0.5$
$H_A : P < 0.5$

فرض H_0 بررسی می‌کند که آیا احتمال ابتلای افراد به فشارخون به صورت تصادفی می‌باشد. در سمت مقابل فرض H_A بررسی می‌کند که آیا احتمال ابتلای افراد به فشار خون نادر و کم‌تر از حالت تصادفی ست و به عوامل دیگر بستگی دارد.

خروجی سمپلی که برای این آزمایش برداشتیم در جدول(۸) و نیز bar plot آن در تصویر(۳) قابل مشاهده می‌باشد:

جدول(۸): دسته‌بندی افراد براساس داشتن فشارخون

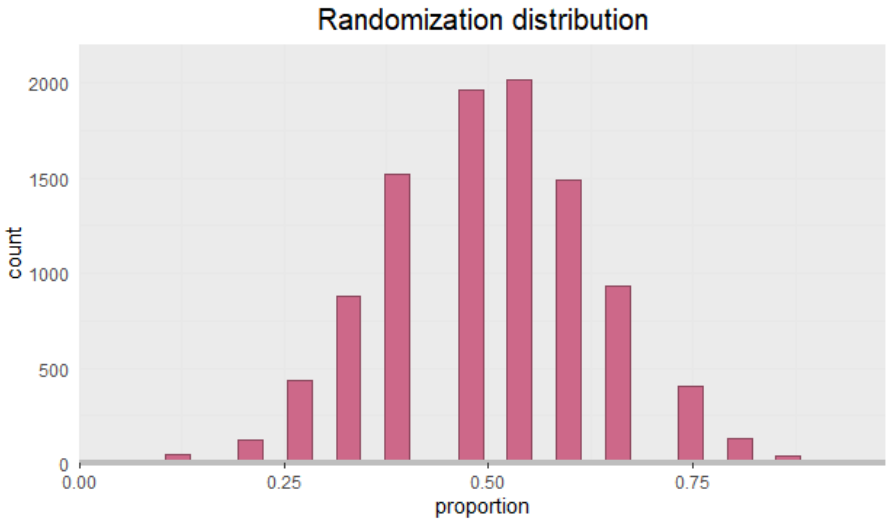
hypertension	
Yes	1
No	14
$\hat{P}$	0.06666667



تصویر(۳): barplot براساس داشتن فشارخون

اگر فرض H_0 صحیح باشد، مثل این است که یک سکه‌ی سالم داریم و این سکه را به 15 مرتبه پرتاب کرده‌ایم و تعداد 'head' آمدن آن را می‌شماریم. این مقدار برابر $\hat{p}_{sim}$ برای یک simulation است. ما این simulation را به تعداد 10000 مرتبه تکرار کرده‌ایم. کدهای زده شده برای این قسمت و خروجی در تصویر(۴) قابل مشاهده می‌باشد:

```
prop_list <- list()

for (i in 1:10000){
  my.sample <- sample( x=c("1", "0"), size=15, replace=TRUE)
  prop_list[[i]] <- mean(my.sample == "1")
}

prop_df <- data.frame(matrix(unlist(prop_list), nrow=1000, byrow=TRUE),stringsAsFactors=FALSE)
colnames(prop_df) <- c('proportion')

ggplot(prop_df,aes(x = proportion)) +
  geom_histogram(bins = 30, fill = "palevioletred3", color = "palevioletred4")+
  scale_y_continuous(expand = c(0, 0))+
  coord_cartesian(ylim = c(0, 230)) +
  ggtitle("Randomization distribution")+
  theme(
    panel.grid = element_line(color = "gray91"),
    axis.line.x = element_line(size = 1.5, linetype = "solid", colour = "gray"),
    axis.ticks.y = element_blank(),
    plot.title = element_text(size = 15, hjust = 0.5)
  )
```



تصویر(۴): هیستوگرام حاصل از شبیه‌سازی

محاسبه‌ی P-value:

P-value = $P(\hat{p}_{sim} \leq 0.06666667 \mid p = 0.5)$

باتوجه به مقدار به‌دست آمده برای $\hat{p}_{sim}$ در هربار اجرای simulation، بررسی می‌کنیم که چه تعداد از آن‌ها مساوی و کمتر از مقدار $\widehat{P}$ می‌باشد:

```
p_val_sim <- mean(prop_df <= prop_hat)
```

→ P-value = 6e-04

باتوجه به مقدار به‌دست آمده برای P-value؛ چون مقدار آن از $\alpha < 0.05$ می‌باشد، پس فرض H_0 را رد می‌کنیم. به عبارتی افراد به صورت تصادفی به فشارخون مبتلا نمی‌شوند و عوامل دیگر از جمله ژنتیک و ... نیز در ابتلا به این بیماری موثر است.

Question 3

برای این قسمت متغیر "work type" در نظر گرفته شده‌است.

(A

ابتدا با استفاد از دستور ()prop.table ، درصد هر level را محاسبه کردیم. خروجی در جدول(۹) قابل مشاهده می‌باشد.

```
work_table <- round(prop.table(table(data$work_type))*100,2)
```

جدول(۹): توزیع احتمالاتی متغیر "work_type"

Children	Govt_job	Never_worked	Private	Self_employed
13.14%	12.86%	0.43%	57.24%	16.03%

توزیع 100فردی که برای نمونه‌گیری اول (که صورت random انجام شده‌است)انتخاب شده‌اند، در جدول(۱۰) قابل مشاهده می‌باشد.

```
sample_1 <-data.frame(data[sample(nrow(data), 100),c(7)])
colnames(sample_1) <- c("work_type")
table(sample_1)
```

جدول(۱۰): توزیع افراد بر حسب متغیر "work_type" بعد از نمونه‌گیری رندم

Children	Govt_job	Never_worked	Private	Self_employed
10	13	1	62	14

برای نمونه‌گیری دوم، برای اینکه بایاس داشته‌باشیم به این صورت عمل شده‌است که levelهای موجود در این متغیر را جدا کردیم و سپس از هر level بدون توجه به درصد توزیع آن در کل دیتاست اصلی، تعدادی برداشتیم. به عنوان مثال از level "Never_worked" با توجه به این‌که تنها 0.43% جامعه را تشکیل می‌دهد، از 100 عدد نمونه‌ای که می‌خواهیم،کل تعداد اعضای این سطح را که برابر 22 نفر می‌باشد.برداشتیم، 35 نمونه‌ی بعدی را از سطح "children" انتخاب می‌کنیم. از43 نفر باقی‌مانده، برای سطح "Gov_job" 5 نفر، سطح "private" 30 نفر و در نهایت 8 نفر از سطح "Self-employed" انتخاب کرده‌ایم. خروجی در جدول(۱۱) قابل مشاهده می‌باشد.

```
children <- data[data$work_type == "children", c(1,7)]
Private <- data[data$work_type == "Private", c(1,7)]
```

```
Govt_job <- data[data$work_type == "Govt_job", c(1,7)]
Self_employed <- data[data$work_type == "Self-employed", c(1,7)]

Never_worked <- data.frame(data[data$work_type == "Never_worked", c(7)])
colnames(Never_worked) <- c("work_type")

b <- data.frame(children[sample(nrow(children), 35),c(2)])
colnames(b) <- c("work_type")

c <- data.frame(Private[sample(nrow(Private), 30),c(2)])
colnames(c) <- c("work_type")

d <- data.frame(Govt_job[sample(nrow(Govt_job), 5),c(2)])
colnames(d) <- c("work_type")

e <- data.frame(Self_employed[sample(nrow(Self_employed), 8),c(2)])
colnames(e) <- c("work_type")

total_sample <- rbind(Never_worked,b,c,d,e)
table(total_sample)
```

جدول(۱۰): توزیع افراد بر حسب متغیر "work_type" بعد از نمونه‌گیری بایاس‌دار

Children	Govt_job	Never_worked	Private	Self_employed
35	5	22	30	8

حال تست goodness of fit را برای هر کدام از این دو نمونه‌ای که گرفتیم اجرا می‌کنیم :

H_0 (nothing going on) : people selected for work type are a simple random sample from the population. The observed counts of people from various work type follow the same distribution in the population.

H_A (something going on) : people selected for work type are not a simple random sample from the population. The observed counts of people from various type do not follow the same distribution in the population.

در قدم اول شرایط تست chi square را بررسی می‌کنیم:

- شرط independence : random sampling/assignment داشته‌باشیم. سایز نمونه از 10% جامعه‌ی آماری کمتر باشد. هر case که داریم فقط در یکی از خانه‌های جدول توزیع وجود داشته‌باشد.
- Sample size : هرکدام از cell ها باید حداقل 5 تا expected case داشته‌باشد.

۱_ سمپل اول که به صورت رندم انتخاب کردیم:

همه‌ی شروط در این سمپل برقرار می‌باشد، به جز شرط sample size و به همین منظور ما دو سطح "children" و "Never_worked" را یک خانه در نظر می‌گیریم. بنابراین تعداد کل اعضای این خانه برابر 11 می‌شود. البته یک مرتبه هم بدون در نظر گرفتن این موضوع حل کردیم. نتایج در تصویر(۵) و (۷) قابل مشاهده می‌باشد.

```
#chi_square test with considering  sample size condition for unbiased sample
count_1 <- c(11, 13, 62, 14)
chi_test_sample_1 <- chisq.test(count_1 , p = c(0.1387, 0.1286, 0.5724, 0.1603))
```

```
        Chi-squared test for given probabilities

data:  count_1
X-squared = 1.2483, df = 3, p-value = 0.7414
```

تصویر(۵): خروجی حاصل از تست chi-square بر روی نمونه‌ی رندم با در نظر گرفتن برقرای شرط

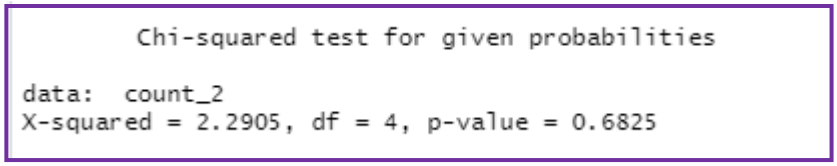
مقادیر expected نیز در تصویر (۶) قابل مشاهده می‌باشد:

```
> chi_test_sample_1$expected
[1] 13.87 12.86 57.24 16.03
```

تصویر(۵): مقادیر expected برای متغیر 'wor_type' در سمپل رندم با در نظر گرفتن برقرای شرط

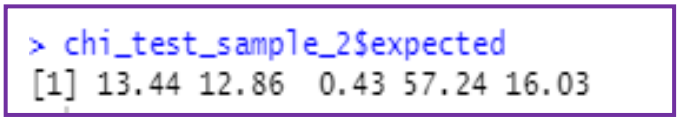
حال برای حالتی که شرط sample size رعایت نشود و یک خانه با کمتر از 5 عدد case داشته‌باشیم:

```
#chi_square test without considering sample size condition for unbiased sample
count_2 <- c(10, 13, 1, 62, 14)
chi_test_sample_2 <- chisq.test(count_1 , p = c(0.1344, 0.1286,0.0043, 0.5724, 0.1603))
```



تصویر(۷): خروجی حاصل از تست chi-square بر روی نمونه‌ی رندم بدون در نظر گرفتن برقرای شرط

مقادیر expected در تصویر(۸) قابل مشاهده می‌باشد:



تصویر(۸): مقادیر expected برای متغیر 'wor_type' در سمپل رندم بدون در نظر گرفتن برقرای شرط

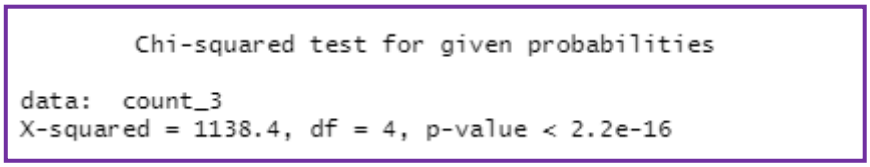
البته مقدار p-value به‌دست آمده در دو حالت تفاوت زیادی باهم ندارند و می‌توان گف تقریبا برابر می‌باشند. حال با توجه به این مقدار به‌دست آمده برای p-value که در تصاویر بالا شاهد آن هستیم، چون این مقدار از $\alpha = 0.05$ بیشتر می‌باشد، آنگاه نمی‌توان فرض H_0 را رد کرد و تلویحا می‌پذیریم که توزیع work type در این نمونه با توزیع آن در جامعه (population) یکسان می‌باشد.

۱_ سمپل دوم که بایاس دارد:

در این قسمت نیز همان مشکل حالت قبل را داریم و شرط سایز سمپل برای expected برقرار نمی‌باشد، به همین دلیل مانند قبل در دو حالت آن را بررسی می‌کنیم:

ابتدا برای حالتی که شرط را در نظر نگیریم که خروجی در تصویر(۹) قابل مشاهده می‌باشد:

```
#chi_square test without considering sample size condition for biased sample
count_3 <- c(35, 5, 22, 30, 8)
chi_test_sample_3  <- chisq.test(count_3 , p = c(0.1344, 0.1286,0.0043, 0.5724, 0.1603))
```



تصویر(۹): خروجی حاصل از تست chi-square بر روی نمونه‌ی بایاس دار بدون در نظر گرفتن برقرای شرط

مقادیر expected نیز مانند بالا می‌باشد، برای حالتی که شرط sample size را در نظر نگرفتیم.

در حالت بعد که شرط sample size را در نظر گرفتیم، مانند بالا دو ستون 'children' و 'No' به این صورت عمل کردیم که ستون 'children' و 'Never_worked' را باهم یکی کردیم و نتیجه در تصویر(۱۰) قابل مشاهده می‌باشد.

```
        Chi-squared test for given probabilities

data:  count_4
X-squared = 72.709, df = 3, p-value = 1.122e-15
```

تصویر(۱۰): خروجی حاصل از تست chi-square بر روی نمونه‌ی بایاس دار با در نظر گرفتن برقرای شرط

البته مقدار p-value به‌دست آمده در دو حالت تفاوت زیادی باهم ندارند و می‌توان گف تقریبا برابر می‌باشند. حال با توجه به این مقدار به‌دست آمده برای p-value که در تصاویر بالا شاهد آن هستیم، چون این مقدار از $\alpha = 0.05$ خیلی کم‌تر می‌باشد، آنگاه فرض H_0 را رد می‌کنیم و تلویحا فرض H_A را می‌پذیریم که بیان می‌کند توزیع work type در این نمونه با توزیع آن در جامعه (population) یکسان نمی‌باشد.

(B

متغیر Categorical دومی که برای این قسمت انتخاب شده‌است، متغیر " gender" می‌باشد.این متغیر دارای سه گروه 'Female' ، 'male' و 'other' می‌باشد.آزمون فرض بر روی کل افراد موجود در دیتاست تست chi-square را انجام داده‌ایم.

H_0 : (nothing going on) → gender work types are independent, work types do not vary by gender.

H_A : (nothing going on) → gender and work types are dependent, work types vary by gender.

در قدم اول شرایط chi-square را بررسی می‌کنیم:

۱- **شرط independence** : random sampling/assignment داشته‌باشیم، سایز نمونه از 10% جامعه‌آماری کمتر باشد و هر case فقط به یک خانه از جدول contribute کند.

۲_ **شرط sample size** : هرکدام از خانه‌های جدول باید تعدادی بیشتر از 5 داشته‌باشند.

جدول(۱۱): توزیع افراد بر حسب متغیر "work_type" و 'gender'

	Children	Govt_job	Never_worked	Private	Self_employed
Female	326	399	11	1757	5044
Male	361	258	11	1170	315
Other	0	0	0	1	0

همانطور که در جدول(۱۱) مشاهده می‌کنید سطرهای مربوط به "Other" شرط sample size را رعایت نکرده‌اند(باتوجه به اینکه نمونه‌ی بالا برای داده‌های مشاهده شده می‌باشد، اما وقتی مقادیر expectedرا به دست‌می‌آوریم نیز مشاهده می‌کنیم که شرط رعایت نمی‌شود. با توجه به توزیعی که داده‌های مشاهده شده دارد)، به همین دلیل ما این سطر را یک بار با سطر Female ادغام کرده و تست را انجام دادیم و یک بار با سطر Male ادغام کرده و تست را انجام دادیم. نتایج و کدهای این قسمت به شرح زیر می‌باشد:

- ابتدا با ستون Female ادغام می‌کنیم که در جدول(۱۲) توزیع افراد قابل مشاهده می‌باشد و سپس تست chi_square را انجام می‌دهیم که نتیجه در تصویر(۱۱) قابل مشاهده می‌باشد.

```
df <- data
df$gender[df$gender == "Other"]  <- "Female"

gender_work_table <- table(df$gender, df$work_type)
indep_test <- chisq.test(df$gender, df$work_type)
```

جدول(۱۲): توزیع افراد بر حسب متغیر 'work_type' و 'gender' بعد از ادغام سطح 'other' با سطح 'Feamle'

	Children	Govt_job	Never_worked	Private	Self_employed
Female	326	399	11	1755	504
Male	361	258	11	1170	315

```
        Pearson's Chi-squared test

data:  df$gender and df$work_type
X-squared = 42.369, df = 4, p-value = 1.399e-08
```

تصویر(۱۱): خروجی حاصل از تست chi-square به منظور بررسی استقلال در دو متغیر 'gender' و 'work_type'

- حال این دفعه با سطر 'Other'را سطر 'Male' ادغام می‌کنیم که در جدول(۱۳) توزیع افراد قابل مشاهده می‌باشد و سپس تست chi_square را انجام می‌دهیم که نتیجه در تصویر(۱۲) قابل مشاهده می‌باشد.

```
df2 <- data
df2$gender[df$gender == "Other"]  <- "Male"

gender_work_table_2 <- table(df2$gender, df2$work_type)

indep_test2 <- chisq.test(df2$gender, df2$work_type)
```

جدول(۱۲): توزیع افراد بر حسب متغیر 'work_type' و 'gender' بعد از ادغام سطح 'other' با سطح 'Male'

	Children	Govt_job	Never_worked	Private	Self_employed
Female	326	399	11	1754	504
Male	361	258	11	1171	315

```
        Pearson's Chi-squared test

data:  df2$gender and df2$work_type
X-squared = 42.25, df = 4, p-value = 1.481e-08
```

تصویر(۱۲): خروجی حاصل از تست chi-square به منظور بررسی استقلال در دو متغیر 'gender' و 'work_type'

مقدار p-value به‌دست آمده در دو حالت تقریبا برابر هم می‌باشد. با توجه به مقداری که برای p-value در دو حالت بالا شاهد هستیم، چون این مقدار از $\alpha = 0.05$ خیلی کمتر می‌باشد، آنگاه فرض H_0 را رد می‌کنیم و به عبارتی تلویحا می‌پذیریم دو متغیر "work_type" و "gender" به‌هم وابسته می‌باشند.

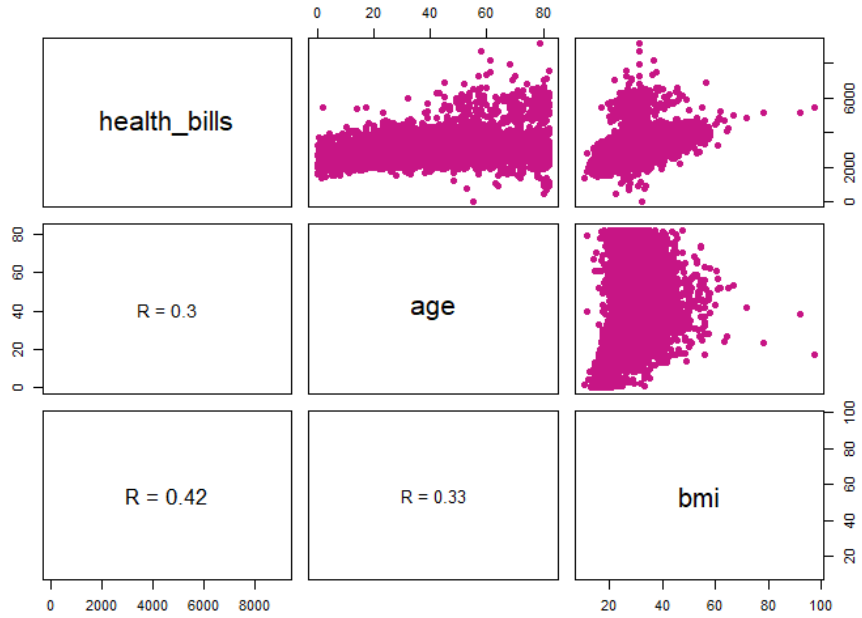
Question 4

متغیرعددی که برای این قسمت به عنوان متغیر response انتخاب شده‌است، متغیر "Health_Bills' می‌باشد. دو متغیر دیگری که به عنوان explanatory انتخاب شده‌است؛ متغیرهای 'bmi' و 'age' می‌باشد. لازم به ذکر است که آز آنجایی که برای دو ستون 'health_bills' و 'bmi' مقادیر'NA' وجود داشت، ابتدا این مقادیر را حذف کردیم و تعداد سطرهای باقی مانده برابر 4909 می‌باشد.

```
num_no_miss_df <- data[rowSums(is.na(data)) == 0, c(13, 3, 10)]
```

(A

نمودار correlogram که برای این متغیرها رسم کردیم در تصویر(۱۳) قابل مشاهده می‌باشد:



تصویر(۱۳): نمودار correlogram برای سه متغیر 'age' و 'bmi' و 'health_bills'

باتوجه به تصویر(۱۳)، به دلیل آنکه correaltion بین دو متغیر 'bmi' و 'health_bills' بیشتر می‌باشد، به نظر می‌رسد که متغیر 'bmi' نسبت به متغیر 'age'، significant تر باشد.

(B

a) ابتدا شرایط linear regression را بررسی می‌کنیم:

شرایط linear regression:

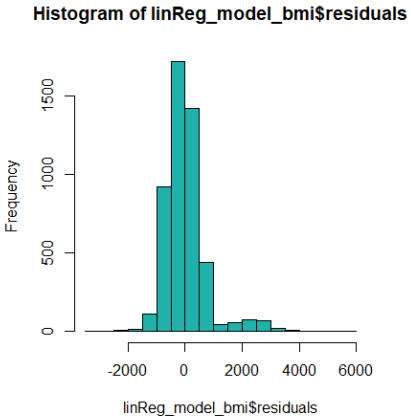
1- **linearity**: همانطور که در نمودار بالا مشاهده می‌شود رابطه‌ی بین متغیرهای 'bmi' و 'age'با متغیر 'health_bills' خطی می‌باشد.

2-**nearly normal residuals**: طبق تصاویر(۱۴) و (۱۵) برای متغیر 'bmi' و طبق تصاویر (۱۷) و (۱۸) برای متغیر 'age'، با اغماض برقرار می‌باشد.

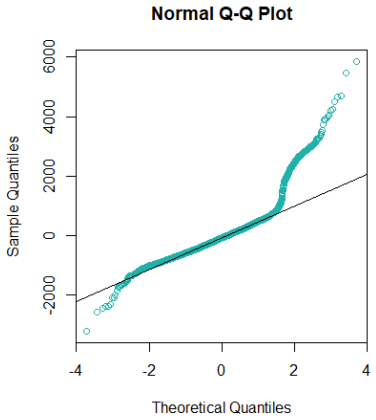
3- **constant variability**: طبق تصاویر(۱۶) برای متغیر 'bmi' و طبق تصاویر (۱۹) برای متغیر 'age'، این شرط نیز برقرار می‌باشد.

bmi:

Nearly normal residuals

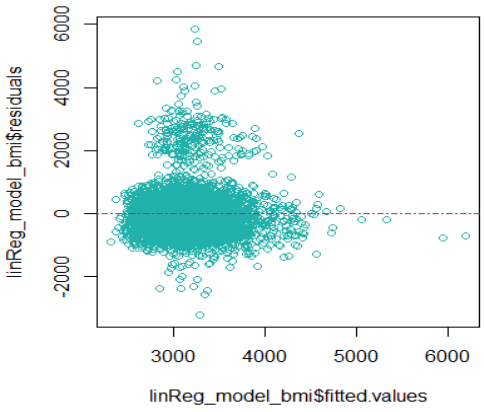


تصویر(۱۵): نمودار هیستوگرام برای بررسی شرط Nearly normal residuals برای مدلی که توسط متغیر'bmi' ساخته شده‌است.



تصویر(۱۴): نمودارQ-Q plot برای بررسی شرط Nearly normal residuals برای مدلی که توسط متغیر'bmi' ساخته شده‌است.

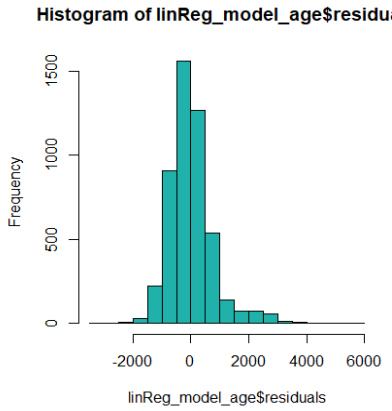
Constant variability



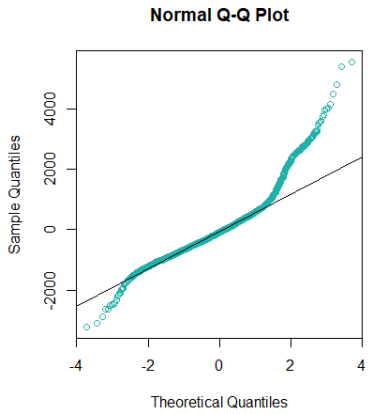
تصویر(۱۶): residual plot برای بررسی شرط constant variabilityبرای مدلی که توسط متغیر'bmi' ساخته شده‌است.

Residual plot Age:

Nearly normal residuals

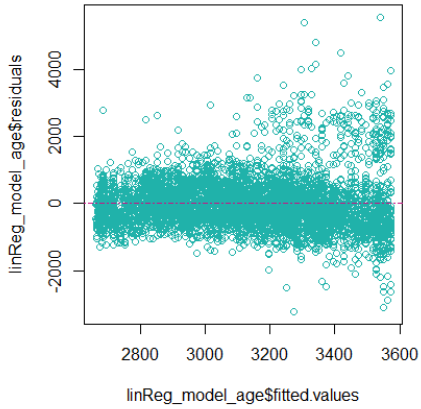


تصویر(۱۸): نمودار هیستوگرام برای بررسی شرط Nearly normal residuals برای مدلی که توسط متغیر'age' ساخته شده‌است.



تصویر(۱۷): نمودار Q-Q plot برای بررسی شرط Nearly normal residuals برای مدلی که توسط متغیر'age' ساخته شده‌است.

Constant variability



تصویر(۱۹): residual plot برای بررسی شرط constant variabilityبرای مدلی که توسط متغیر'age' ساخته شده‌است.

بعد از بررسی شروط، حال باتوجه به صورت سوال برای هر متغیر explanatory که در بالا معرفی کردیم، یک linear regression فیت نموده‌ایم. کدهای زده شده برای این قسمت و خروجی آن در تصویر(۲۰) و (۲۱) قابل مشاهده می‌باشد:

```
#least squares regression for bmi
linReg_model_bmi <- lm(health_bills ~ bmi, data = num_no_miss_df)
summary(linReg_model_bmi)

#least squares regression for age
```

```
linReg_model_age <- lm(health_bills ~ age, data = num_no_miss_df)
summary(linReg_model_age)
```

خروجی برای متغیر **'bmi'** :

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

Residuals:
    Min      1Q  Median      3Q     Max
-3236.1  -449.6   -85.5   271.7  5868.4

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1856.124     40.679   45.63   <2e-16 ***
bmi           44.386      1.359   32.67   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 747.5 on 4907 degrees of freedom
Multiple R-squared:  0.1787,    Adjusted R-squared:  0.1785
F-statistic:  1067 on 1 and 4907 DF,  p-value: < 2.2e-16
```

تصویر(۲۰): خروجی مدل linear regression ساخته شده توسط متغیر 'bmi'

خروجی برای متغیر **'age'** :

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

Residuals:
    Min      1Q  Median      3Q     Max
-3228.4  -482.8   -84.7   351.8  5561.2

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2663.2295    24.0935  110.54   <2e-16 ***
age           11.0895     0.4974   22.29   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 786 on 4907 degrees of freedom
Multiple R-squared:  0.09197,   Adjusted R-squared:  0.09179
F-statistic:   497 on 1 and 4907 DF,  p-value: < 2.2e-16
```

تصویر(۲۱): خروجی مدل linear regression ساخته شده توسط متغیر 'age'

(b

برحسب متغیر **'bmi'**:

$\widehat{health\_bills}$ = 1856.124 + 44.386 bmi

تفسیر عرض از مبدا (intercept): زمانی که مقدار 'bmi' برابر صفر باشد، تخمین مدل ما برای مقدار 'health_bills' برابر 1856.124 خواهد بود. که البته بی‌معنی می‌باشد، زیرا امکان ندارد مقدار 'bmi' یک فرد برابر صفر باشد.

تفسیر شیب خط (slope) : شیب خط بیان می‌کند که به ازای هر واحد افزایش در مقدار 'bmi'، به اندازه‌ی 44.386 مقدار 'health_bills' افزایش می‌یابد.

برحسب متغیر **'age'**:

$\widehat{health\_bills}$ = 2663.2295 + 11.0895 age

تفسیر عرض از مبدا (intercept): زمانی که مقدار 'age' برابر صفر باشد، تخمین مدل ما برای مقدار 'health_bills' برابر 2663.2295 خواهد بود. که البته بی‌معنی می‌باشد، زیرا امکان ندارد مقدار 'bmi' یک فرد برابر صفر باشد.

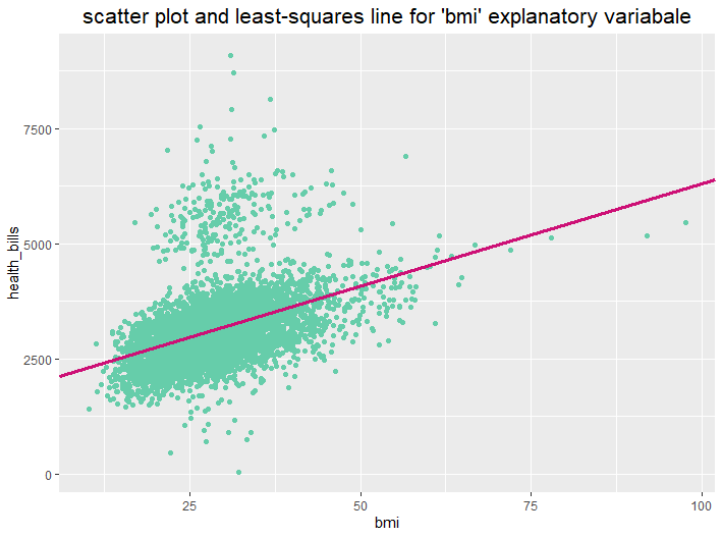
تفسیر شیب خط (slope) : شیب خط بیان می‌کند که به ازای هر واحد افزایش در سن افراد ('Age')، به اندازه‌ی 11.0895 مقدار 'health_bills' افزایش می‌یابد.

(c

به منظور رسم نمودار scatter و فیت کردن خط linear regression از دستورات زیر استفاده شده‌است و خروجی‌های به دست آمده در تصاویر(۲۲) و (۲۳) قابل مشاهده می‌باشد.

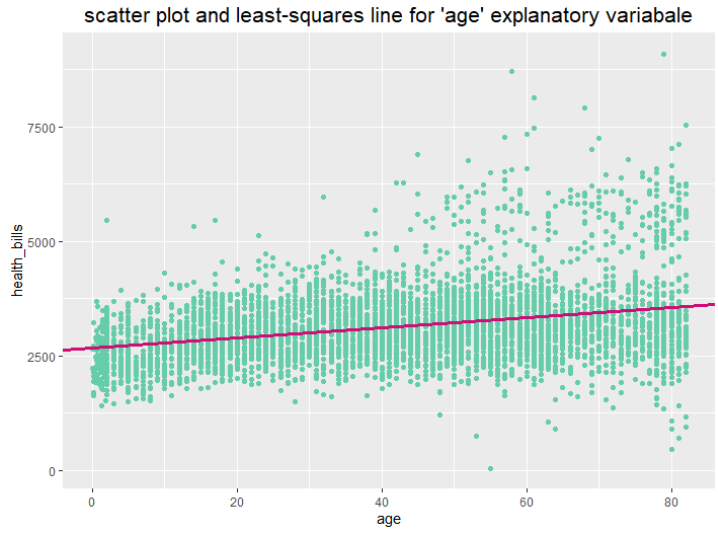
```
ggplot(num_no_miss_df , aes(x = bmi, y = health_bills) ) +
 geom_point(color = 'darkmagenta') +
 geom_abline(slope = bmi_reg$coefficients[[2]],
             intercept = bmi_reg$coefficients[[1]],
             color = 'seagreen4', lwd = 1.3)+
 ggtitle("scatter plot and least-squares line for 'bmi' explanatory variabale")+
 theme(
  plot.title = element_text(hjust = 0.5, size = 16),
 )
```

نمودار برای برحسب **'bmi'**:



تصویر(۲۲): scatter plot برای نمایش رابطه‌ی بین دو متغیر 'bmi' و 'health_bills' و همچنین فیت کردن خط regression بر روی آن

نمودار برای برحسب **'Age'**:



تصویر(۲۳): scatter plot برای نمایش رابطه‌ی بین دو متغیر 'age' و 'health_bills' و همچنین فیت کردن خط regression بر روی آن

(C

باتوجه به دو مدل، که با استفاده از دو متغیر explanatory، 'bmi' و 'age' در مورد 'B' ساختیم، مقدار p-value به‌دست آمده برای هر دو متغیر 2e-16> می‌باشد، بنابراین هردو متغیر significant هستند، مقدار دقیق p-valueها به شرح زیر می‌باشد:

p-value for bmi :

age_reg$coefficients[8]

> 5.198011e-212

p-value for age :

age_reg$coefficients[8]

> 5.912627e-105

باتوجه به مقدار p-value دقیق به دست آمده، متغیر 'bmi' بیشتر از متغیر 'age'، significant می‌باشد. زیرا مقدار p-value آن کمتر می‌باشد.

(D

مقایسه براساس $\mathbf{R^2_{adj}}$:

باتوجه به خروجی مدل در مورد 'B'، مقدار به دست‌آمده برای R^2_{adj} در مدلی که با استفاده از متغیر 'bmi' و در مدلی که با استفاده از متغیر 'age' ساختیم، به شرح زیر می‌باشد:

R^2_{adj} for bmi

bmi_reg$adj.r.squared

> 0.1784898

R^2_{adj} for age

age_reg$adj.r.squared

> 0.09178569

می‌دانیم که هرچه مقدار R^2_{adj} بیشتر باشد، توان مدل بیشتر خواهد بود و مدل بهتری خواهیم داشت، بنابراین مدلی که توسط متغیر 'bmi'ساخته‌ایم از مدل دیگر بهتر می‌باشد و توانایی بیشتری در پیش‌بینی کردن مقدار متغیر response (health_bills) دارد.

مقایسه براساس تست **ANOVA**:

باتوجه به خروجی مدل در مورد 'B'، سطر آخر از این خروجی بیانگر اعمال تست ANOVA بر روی این مدل‌ها می‌باشد، البته می‌توان به طور جداگانه تست ANOVA را با استفاده از تابع anova() نیز بر روی مدل‌های ساخته شده نیز اعمال کرد که نتیجه‌ی هر دو حالت یکی می‌شود. خروجی این قسمت در تصاویر(۲۴) و (۲۵) قابل مشاهده می‌باشد.

* البته لازم به ذکر است که چون هر دو مدل به یک اندازه predictor داشتند، زمانی که تابع ANOVA را به صورت anova(model1, model2) استفاده می‌کنیم تا دو مدل را مقایسه کند، مقدار df = 0 می‌شود و بنابراین p-value نخواهیم داشت. باید حتما تعداد predictorهای مدل‌ها یکسان نباشد تا بتوان از این حالت استفاده کرد، به همین منظور ما به طور جداگانه تابع ANOVA را بر روی هر مدل استفاده کردیم و مقدار p-value به دست آمده در دو حالت را باهم مقایسه کردیم.

bmi:

```
bmi_aov <- anova(linReg_model_bmi)
```

```
Analysis of Variance Table

Response: health_bills
            Df     Sum Sq   Mean Sq F value    Pr(>F)
bmi          1  596471821 596471821  1067.4 < 2.2e-16 ***
Residuals 4907 2742166684    558828
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

تصویر(۲۴): خروجی تابع anova() بر روی مدل ساخته شده توسط متغیر 'bmi'

exact p-value:

```
bmi_aov$`Pr(>F)`
```

```
> 5.198011e-212
```

age:

```
age_aov <- anova(linReg_model_age)
```

```
Analysis of Variance Table

Response: health_bills
            Df     Sum Sq   Mean Sq F value    Pr(>F)
age          1  307057057 307057057  497.01 < 2.2e-16 ***
Residuals 4907 3031581449    617808
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

تصویر(۲۵): خروجی تابع anova() بر روی مدل ساخته شده توسط متغیر 'age'

exact p-value:

age_aov$`Pr(>F)`

> 5.912627e-105

باتوجه به مقادیر به دست آمده در تست ANOVA نیز قابل مشاهده می‌باشد؛ مدلی که با استفاده از متغیر 'bmi' ساخته شده‌است، مدل بهتری می‌باشد، زیرا مقدار p-value آن کمتر است.

(E

باتوجه به مورد 'D' لیست بهترین predictorها به ترتیب به شرح زیر می‌باشد:

1) bmi

2) age

اگر هردوی این متغیر را در پیش‌بینی متغیر 'health_bills' استفاده کنیم، R^2_{adj} نسبت به حالتی که با هر کدام به تنهایی مدل بسازیم، افزایش پیدا خواهد کرد. خروجی در تصویر(۲۶) قابل مشاهده می‌باشد.

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

Residuals:
    Min      1Q  Median      3Q     Max
-3296.7  -442.2   -73.9   293.8  5640.6

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1754.5934    40.6378   43.18   <2e-16 ***
bmi           37.9916     1.4149   26.85   <2e-16 ***
age            6.6789     0.4927   13.56   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 734 on 4906 degrees of freedom
Multiple R-squared:  0.2083,    Adjusted R-squared:  0.208
F-statistic: 645.4 on 2 and 4906 DF,  p-value: < 2.2e-16
```

تصویر(۲۶): خروجی مدل linear regression ساخته شده توسط دو متغیر 'age' و 'bmi'

همانطور که در تصویر بالا قابل مشاهده می‌باشد، مقدار R^2_{adj} برابر 0.208 شده است و نسبت به دو مدل که در مورد 'B' ایجاد کردیم، بیشتر است. بنابراین این دو متغیر predictorهای خوبی برای پیش‌بینی مقدار 'health_bills' می‌باشند.

(F

ابتدا طبق خواسته‌ی سوال یک نمونه‌ی 100تایی از مجموعه‌داده برداشتیم:

```
samp_1 <- num_no_miss_df[sample(nrow(num_no_miss_df), 100), ]
```

(a

طبق صورت سوال 90% از دیتا را برداشته و دو linear regression به ازای هر کدام از متغیرهای explanatory که داشتیم، ایجاد کردیم. اما در قدم اول به بررسی شرایط استفاده از linear regression می‌پردازیم:

```
df_90percent <- samp_1[sample.int(n = nrow(samp_1), size = floor(0.9*nrow(samp_1)), replace = F),]
```

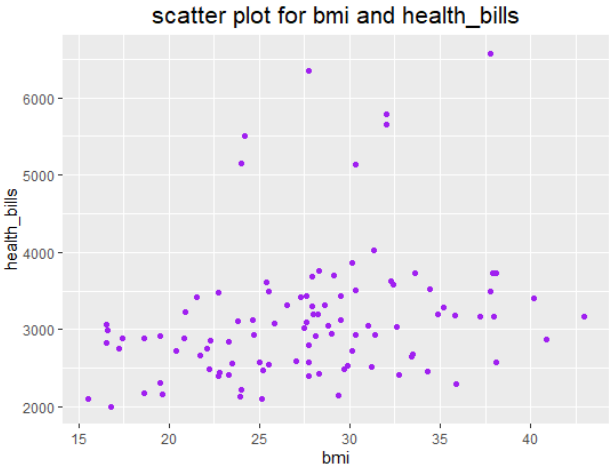
شرایط linear regression:

1- **linearity**: رابطه‌ی بین متغیرهای 'bmi' و 'age'با متغیر 'health_bills' باتوجه به scatter plot که در تصاویر (۲۷) و (۲۸) قابل مشاهده می‌باشد، خطی است.

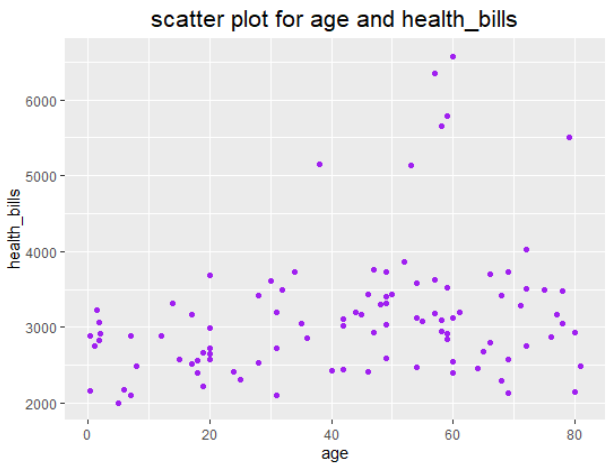
2- **nearly normal residuals**: طبق تصاویر(۲۹) و (۳۰) برای متغیر 'bmi' و طبق تصاویر (۳۲) و (۳۳) برای متغیر 'age'، برقرار می‌باشد.

3- **constant variability**: طبق تصاویر(۳۱) برای متغیر 'bmi' و طبق تصاویر (۳۴) برای متغیر 'age'، این شرط نیز برقرار می‌باشد.

Scatter plot:



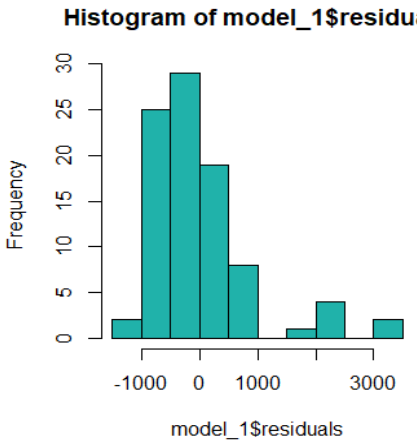
تصویر(۲۷): scatter plot برای نمایش رابطه‌ی بین دو متغیر 'bmi' و 'health_bills' در سمپل ۱۰۰تایی



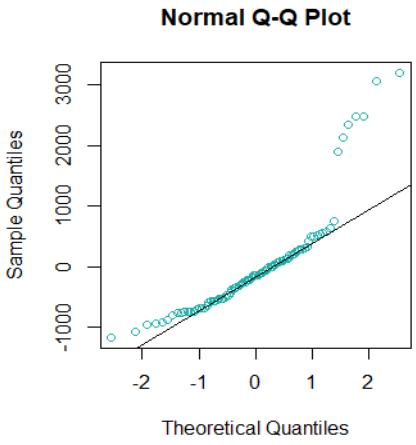
تصویر(۲۸): scatter plot برای نمایش رابطه‌ی بین دو متغیر 'age' و 'health_bills' در سمپل ۱۰۰تایی

bmi:

nearly normal residuals

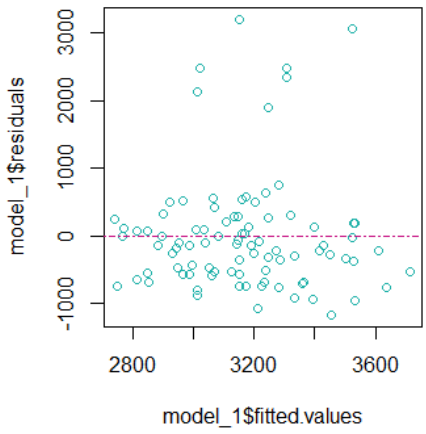


تصویر(۳۰): نمودار هیستوگرام برای بررسی شرط Nearly normal residuals برای مدلی که توسط متغیر'bmi' در سمپل با سایز ۹۰ ساخته شده‌است.



تصویر(۲۹): نمودار Q-Q plot برای بررسی شرط Nearly normal residuals برای مدلی که توسط متغیر'bmi' در سمپل با سایز ۹۰ ساخته شده‌است.

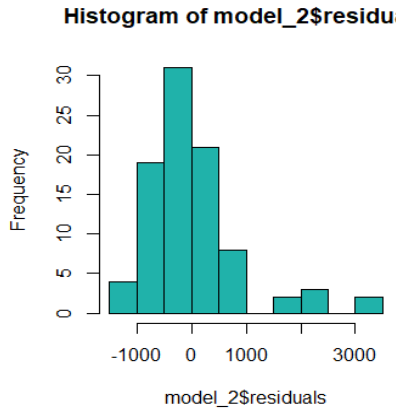
constant variability



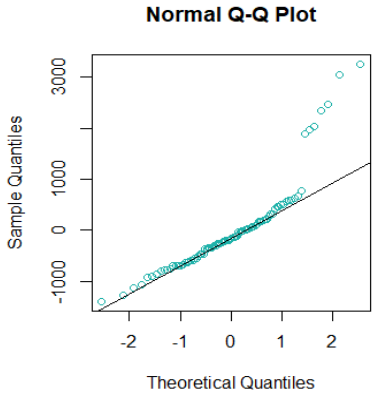
تصویر(۳۱): residual plot برای بررسی شرط constant variabilityبرای مدلی که توسط متغیر 'bmi' در سمپل با سایز ۹۰ساخته شده‌است.

Age:

nearly normal residuals

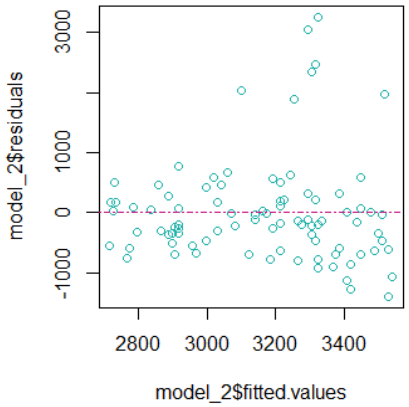


تصویر(۳۳): نمودار هیستوگرام برای بررسی شرط Nearly normal residuals برای مدلی که توسط متغیر'Age' در سمپل با سایز ۹۰ ساخته شده‌است.



تصویر(۳۲): نمودار Q-Q plot برای بررسی شرط Nearly normal residuals برای مدلی که توسط متغیر'age' در سمپل با سایز ۹۰ ساخته شده‌است.

constant variability



تصویر(۳۴): residual plot برای بررسی شرط constant variabilityبرای مدلی که توسط متغیر'age' در سمپل با سایز ۹۰ساخته شده‌است.

حال در قدم بعدی linear regression model را برای هر دو متغیر می‌سازیم که نتایج آن در تصاویر(۳۵) و (۳۶) قابل مشاهده می‌باشد:

```
model_1 <- lm(health_bills ~ bmi, data = df_90percent)
summary(model_1)
```

bmi:

```
Call:
lm(formula = health_bills ~ bmi, data = df_90percent)

Residuals:
    Min      1Q  Median      3Q     Max
-1162.9  -540.8  -141.8   205.7  3192.6

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2127.49     435.72   4.883 4.64e-06 ***
bmi            36.90      15.18   2.431   0.0171 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 864.8 on 88 degrees of freedom
Multiple R-squared:  0.06292,   Adjusted R-squared:  0.05227
F-statistic: 5.909 on 1 and 88 DF,  p-value: 0.01709
```

تصویر(۳۵): خروجی مدل linear regression ساخته شده توسط متغیر 'bmi' برای مجموعه داده با سایز ۹۰

Age:

```
Call:
lm(formula = health_bills ~ age, data = df_90percent)

Residuals:
    Min      1Q  Median      3Q     Max
-1387.5  -529.1  -161.0   202.4  3251.6

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2713.009    194.699  13.934   <2e-16 ***
age           10.220      3.911   2.613   0.0105 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 860.6 on 88 degrees of freedom
Multiple R-squared:  0.07201,   Adjusted R-squared:  0.06147
F-statistic: 6.829 on 1 and 88 DF,  p-value: 0.01055
```

تصویر(۳۶): خروجی مدل linear regression ساخته شده توسط متغیر 'age' برای مجموعه داده با سایز ۹۰

ابتدا یک آزمون فرض برای بررسی متغیر 'Age' طرح می‌کنیم:

$H_0 : \beta_{age} = 0$

$H_A : \beta_{age} \neq 0$

فرض H_0 به این معنی می‌باشد که متغیر 'Age' یک significant predictor نمی‌باشد و متغیری که انتخاب کرده‌ایم یک متغیر بد می‌باشد.

در مقابل فرض H_A بیان می‌کند که یک رابطه‌ای بین متغیر 'health_bills' و 'Age' وجود دارد و این متغیر یک significant predictor می‌باشد. مقدار p-value در خروجی مدل بالا قابل مشاهده می‌باشد:

smry_model_2$coefficients[8]

> 0.0105

طبق مقدار p-value به‌دست آمده، این مقدار از $\alpha = 0.05$ کمتر می‌باشد، بنابراین فرض H_0 را رد می‌کنیم. به عبارتی دیگر متغیر 'Age' یک significant predictor می‌باشد.

حال همین آزمون را برای متغیر 'bmi' نیز بررسی می‌کنیم:

$H_0 : \beta_{bmi} = 0$

$H_A : \beta_{bmi} \neq 0$

فرض H_0 به این معنی می‌باشد که متغیر 'bmi' یک significant predictor نمی‌باشد و متغیری که انتخاب کرده‌ایم یک متغیر بد می‌باشد.

در مقابل فرض H_A بیان می‌کند که یک رابطه‌ای بین متغیر 'health_bills' و 'bmi' وجود دارد و این متغیر یک significant predictor می‌باشد. مقدار p-value در خروجی مدل بالا قابل مشاهده می‌باشد:

```
smry_model_1$coefficients[8]
```

> 0.0171

طبق مقدار p-value به‌دست آمده، این مقدار از $\alpha = 0.05$ کمتر می‌باشد، بنابراین فرض H_0 را رد می‌کنیم. به عبارتی دیگر متغیر 'bmi' یک significant predictor می‌باشد.

طبق آزمون‌های فرض انجام شده، هردو متغیر significant predictor می‌باشد، اما متغیر 'bmi' به علت آنکه p-value به دست‌آمده برای آن کمتر بود، نسبت به متغیر 'Age' بیشتر significant می‌باشد.

(b

بازه اطمینان برای شیب خط به صورت زیر می‌باشد:

$$CI = b_1 \pm t^*_{df} \, SE_{b_1}$$

ابتدا برای متغیر 'bmi' بازه‌اطمینان را محاسبه می‌کنیم. برای این قسمت از تابع confint() استفاده شده‌است.

```
confint(model_1,"bmi", level=0.95)
```

CI = (6.733325, 67.07627)

تفسیر بازه‌ی اطمینان برای شیب خط به این صورت می‌باشد که ما 95% اطمینان داریم که هر یک واحد افزایش در 'bmi'، انتظار داریم که به طور متوسط مقدار 'health_bills' بین (6.733325, 67.07627) افزایش پیدا کند.

حال برای متغیر 'Age' بازه‌ی اطمینان را محاسبه می‌کنیم:

```
confint(model_2,"age", level=0.95)
```

CI = (2.448041, 17.99295)

تفسیر بازه‌ی اطمینان برای شیب خط به این صورت می‌باشد که ما 95% اطمینان داریم که هر یک سال افزایش در سن افراد('Age')، انتظار داریم که به طور متوسط مقدار 'health_bills' بین (2.448041, 17.99295) افزایش پیدا کند.

(c

به منظور پیش‌بینی مقدار 'health_bills' برای 10% باقی‌مانده از دیتا، ازتابع predict استفاده کردیم و مدل‌هایی که با استفاده از دو متغیر explanatory برای 90% دیتا ساخته‌بودیم، به عنوان ورودی این تابع قرار دادیم. کدهای زده شده برای این قسمت به شرح زیر می‌باشد:

```
test_data <- samp_1[-sample.int(n = nrow(samp_1), size = floor(0.9*nrow(samp_1)), replace = F),]

y_pred_1 <- predict(model_1, test_data)
y_pred_2 <- predict(model_2, test_data)
```

مقادیر پیش‌بینی شده توسط مدل اول و مقادیر واقعی برای 'health_bills' در جدول(۱۳) قابل مشاهده می‌باشد.

جدول(۱۳): مقادیر واقعی و پیش‌بینی شده برای متغیر 'health_bills' براساس مدل ساخته شده توسط متغیر 'bmi'

Actual	2729.8	2569.7	3166.8	2559.1	2793.2	2918.8	3293.9	2931.6	3126.8	2648.9
predicted	2880.4	3533.6	3500.4	2994.8	3149.8	3164.5	3157.1	3286.3	3216.2	3360.1

مقادیر پیش‌بینی شده توسط مدل دوم و مقادیر واقعی برای 'health_bills' در جدول(۱۴) قابل مشاهده می‌باشد.

جدول(۱۴): مقادیر واقعی و پیش‌بینی شده برای متغیر 'health_bills' براساس مدل ساخته شده توسط متغیر 'age'

Actual	2729.8	2569.7	3166.8	2559.1	2793.2	2918.8	3293.9	2931.6	3126.8	2648.9
predicted	2917.4	3418.2	3172.9	2896.9	3387.5	3316.0	3203.6	3193.4	3264.9	2917.4

d)

جدول‌های (۱۳) و (۱۴) نماینگر مقادیر واقعی و پیش‌بینی شده برای متغیر response می‌باشد. همانطور که قابل مشاهده می‌باشد در برخی نقاط overestimate و در برخی نقاط underestimate داریم. برای مثال در جدول(۱۳) در ستون اول overestimate داریم زیرا مقدار واقعی برابر 2729.8 و مقدار تخمین زده برابر 2880.4 می‌باشد. اما در ستون 7 دارای underestimate می‌باشیم، زیرا مقدار واقعی برابر 3293.9 و مقدار تخمین‌زده برابر 3157.1 می‌باشد.

به منظور محاسبه‌ی نرخ موفقیت دو روش در نظر گرفته شده‌است:

روش اول:

همانطور که می‌دانیم چون regression داشتیم، مقادیر واقعی و پیش‌بینی شده دقیقا یکسان نمی‌باشند؛ به همین منظور ما این دو مقادیر(مقادیر پیش‌بینی شده و مقادیر واقعی) را برهم تقسیم کرده و هر کدام از این مقدارها را که به عدد '1' نزدیک‌تر بود، فرض کردیم که مقادیر واقعی و پیش‌بینی شده برای آن‌ها دقیقا باهم برابر می‌باشد(مقدار خطای چشم‌پوشی را تا 0.15 بالاتر و پایین‌تر از یک در نظر گرفتیم.). سپس درصد مقادیری که مدل درست پیش‌بینی کرده بود را محاسبه نمودیم.

```
actuals_preds_1 <- data.frame(cbind(actuals=test_data$health_bills, predicteds=y_pred_1))

ratio <- list()
for(i in 1:10){
  ratio[i] <- actuals_preds_1$actuals[i]/actuals_preds_1$predicteds[i]
}

ratio <- data.frame(matrix(unlist(ratio), nrow=10, byrow=TRUE),stringsAsFactors=FALSE)
colnames(ratio) <- c("ratio")

actuals_preds_1$label <- ratio$ratio
actuals_preds_1$new <- ifelse((1 - actuals_preds_1$label) <0.15, '1','0')
success_rate <- length(actuals_preds_1$new[actuals_preds_1$new == 1]) / 10
```

Success rate in model(with 'bmi' explanatory) : 0.8
Success rate in model(with 'age' explanatory) : 0.8

روش دوم:

برای محاسبه‌ی نرخ موفقیت در این روش، از پکیج 'Metrics' مقادیر MSE، MAE، RMSE و MAPE را محاسبه نموده ایم. همچنین Min_Max accuracy را نیز به دست‌آوردیم. توضیحاتی راجب این معیارها در جدول (۱۵) آورده شده‌است.

جدول(۱۵): توضیحاتی راجب MSE، MAE، RMSE، MAPE و Min_Max accuracy

MAPE	Mean absolute percentage error	Lower the better
MSE	Mean squared error	Lower the better
MAE	Mean absolute error	Lower the better
RMSE	Root Mean Square Error	Lower the better
Min_Max accuracy	Mean(min(actuals,predicteds)/max(actuals, predicteds))	Higher the better

```
#success rate for first model
actuals_preds_1 <- data.frame(cbind(actuals=test_data$health_bills, predicteds=y_pred_1))
mse(actuals_preds_1$actuals, actuals_preds_1$predicteds)
mae(actuals_preds_1$actuals, actuals_preds_1$predicteds)
mape(actuals_preds_1$actuals, actuals_preds_1$predicteds)
rmse(actuals_preds_1$actuals, actuals_preds_1$predicteds)
min_max_accuracy <- mean(apply(actuals_preds_1, 1, min) / apply(actuals_preds_1, 1, max))
```

نتایج به‌دست آمده برای مدل اول و مدل دوم در جدول(۱۶) قابل مشاهده می‌باشد.

جدول(۱۶): خروجی success rate برای دوم مدل ساخته‌شده

	First model	Second model
Min_Max accuracy	0.89	0.9
Adjusted R squared	0.052	0.061
MAPE	0.138	0.114
MSE	209870.8	154847.3
MAE	377.8	313.051
RMSE	458.1	393.5

این معیارهای استفاده شده در واقع میزان نزدیک بودن مقدار پیش‌بینی شده توسط مدل با مقدار واقعی متغیر response را بررسی می‌کنند. برای مثال معیار MSE، میانگین مجموع مربعات خطا(منظور از خطای فاصله‌ی بین مقدار واقعی با مقدار پیش‌بینی شده‌است) می‌باشد که هرچه این مقدار کمتر باشد به این معنی است که مقادیر به هم نزدیک‌تر بوده و در نتیجه مدلی که ایجاد کرده‌ایم مدل خوبی می‌باشد.

Question 5

ابتدا سطرهایی missing value را حذف می‌کنیم. متغیرهایی که به عنوان explanatory برای این قسمت انتخاب شده‌است به شرح زیر می‌باشد:

age, work_type, bmi, stroke

* این متغیرها به این دلیل انتخاب شده‌اند که correlation نسبتا بالایی نسبت به متغیرهای باقی‌مانده، با متغیر 'health_bills' به عنوان متغیر response دارند.(باتوجه به نمودار پایین)

(A

در قدم اول ابتدا ستون‌های categorical را با استفاده از دستور زیر به numerical تبدیل می‌کنیم. تایپ ستون‌های این مجموعه داده در تصویر(۳۷) قابل مشاهده می‌باشد.

```
MakeNum <- function(x) as.numeric(as.factor(x))
non_miss_df <- mutate(non_miss_df, across(c(2,6,7,8,11), MakeNum))
```

```
> str(non_miss_df)
'data.frame':   4909 obs. of  13 variables:
 $ id                : int  9046 31112 60182 1665 56669 53882 10434 60491 12109 12095 ...
 $ gender            : chr  "Male" "Male" "Female" "Female" ...
 $ age               : num  67 80 49 79 81 74 69 78 81 61 ...
 $ hypertension      : int  0 0 0 1 0 1 0 0 1 0 ...
 $ heart_disease     : int  1 1 0 0 0 1 0 0 0 1 ...
 $ ever_married      : chr  "Yes" "Yes" "Yes" "Yes" ...
 $ work_type         : chr  "Private" "Private" "Private" "Self-employed" ...
 $ Residence_type    : chr  "Urban" "Rural" "Urban" "Rural" ...
 $ avg_glucose_level: num  229 106 171 174 186 ...
 $ bmi               : num  36.6 32.5 34.4 24 29 27.4 22.8 24.2 29.7 36.8 ...
 $ smoking_status    : chr  "formerly smoked" "never smoked" "smokes" "never smoked" ...
 $ stroke            : int  1 1 1 1 1 1 1 1 1 1 ...
 $ health_bills      : num  6012 6385 5863 5461 5054 ...
```

تصویر(۳۷): تایپ ستون‌های موجود در مجموعه داده

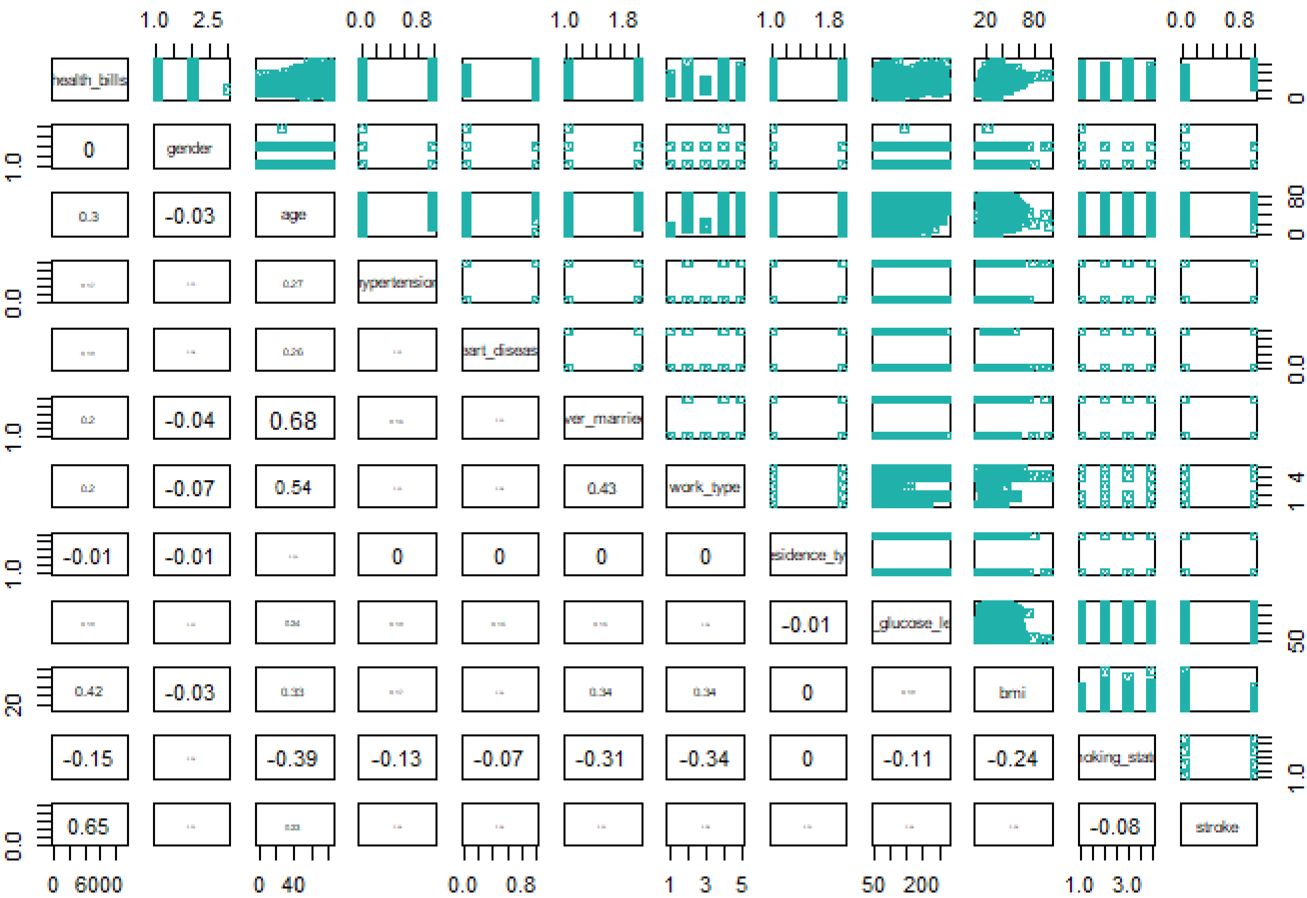
ستون‌هایی که تایپ آن‌ها "chr" می‌باشد، از نوع Categorical هستند. بعد از تبدیل، تایپ ستون‌های مجموعه‌داده در تصویر(۳۸) قابل مشاهده می‌باشد.

```
> str(non_miss_df)
'data.frame':   4909 obs. of  13 variables:
 $ id                : int  9046 31112 60182 1665 56669 53882 10434 60491 12109 12095 ...
 $ gender            : num  2 2 1 1 2 2 1 1 1 1 ...
 $ age               : num  67 80 49 79 81 74 69 78 81 61 ...
 $ hypertension      : int  0 0 0 1 0 1 0 0 1 0 ...
 $ heart_disease     : int  1 1 0 0 0 1 0 0 0 1 ...
 $ ever_married      : num  2 2 2 2 2 2 1 2 2 2 ...
 $ work_type         : num  4 4 4 5 4 4 4 4 4 2 ...
 $ Residence_type    : num  2 1 2 1 2 1 2 2 1 1 ...
 $ avg_glucose_level: num  229 106 171 174 186 ...
 $ bmi               : num  36.6 32.5 34.4 24 29 27.4 22.8 24.2 29.7 36.8 ...
 $ smoking_status    : num  1 2 3 2 1 2 2 4 2 3 ...
 $ stroke            : int  1 1 1 1 1 1 1 1 1 1 ...
 $ health_bills      : num  6012 6385 5863 5461 5054 ...
```

تصویر(۳۸): تایپ ستون‌های موجود در مجموعه داده بعد از تبدیل ستون‌های categorical به numerical

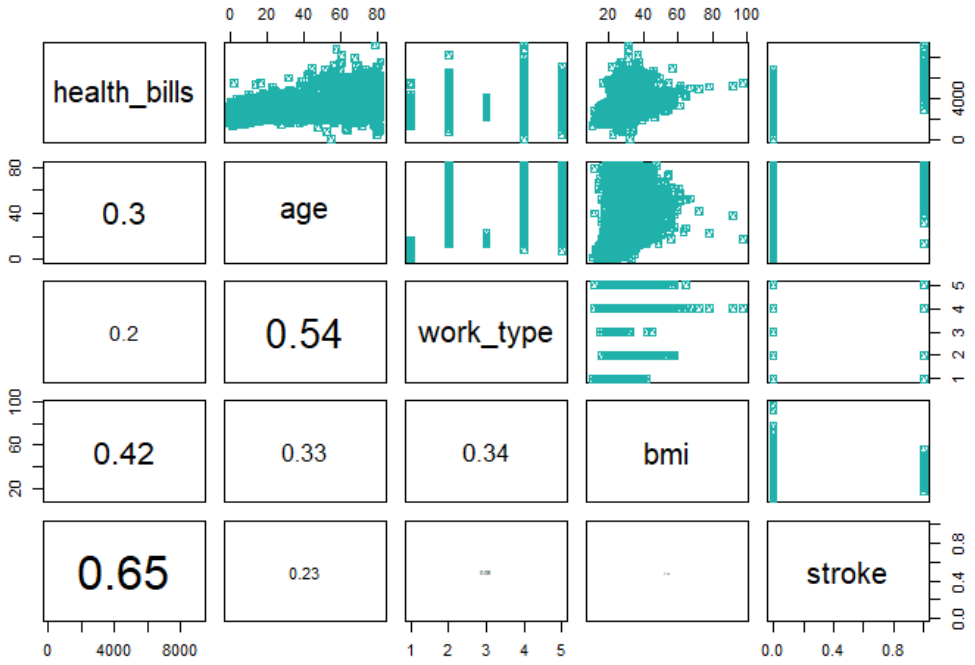
حال correlogram را برای این متغیرها رسم می‌کنیم که در تصویر(۳۹) قابل مشاهده می‌باشد.

```
upper.panel_2<-function(x, y){

  points(x,y, pch = 14, col = "lightseagreen")
}

panel.cor2 <- function(x, y){
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  r <- round(cor(x, y), digits=2)
  txt <- paste0( r)
  cex.cor <- 0.9/strwidth(txt)
  text(0.5, 0.5, txt, cex = cex.cor * r)
}

pairs(non_miss_df[-c(1,6,9,11)],
     lower.panel = panel.cor2,
     upper.panel = upper.panel_2
)
```



تصویر(۳۹): نمودار correlogram برای تمام ستون‌های موجود در مجموعه داده

باتوجه به نمودار بالا و مقدار correlation هر متغیر با متغیر 'health_bills'، متغیرهای 'age'، 'bmi' و 'stroke' به عنوان متغیر explanatory انتخاب شده‌اند، زیرا نسبت به متغیرهای دیگر Correlation بالاتری با متغیر response دارند. از بین این متغیرهای انتخاب شده؛ همانطور که در تصویر بالا قابل مشاهده می‌باشد، correlation بین متغیر 'age' و 'health_bills' برابر 0.3، بین متغیر 'work_type' و 'health_bills' برابر 0.2، بین متغیر 'bmi' و 'health_bills' برابر 0.42 و در نهایت بین متغیر 'stroke' و 'health_bills' برابر 0.65 می‌باشد. متغیر 'stroke' نسبت به متغیرهای دیگر significantتر می‌باشد. سپس متغیر 'bmi' و بعد از آن متغیر 'age' قرار می‌گیرند. تصویر(۴۰) نمودار correlogram بین متغیرهای explanatory انتخاب شده و متغیر response می‌باشد.



تصویر(۴۰): نمودار correlogram برای متغیر response و متغیرهای explanatory انتخاب شده

(B

باتوجه به متغیرهای explanatory انتخاب شده، یک مدل ساختیم که خروجی آن در تصویر(۴۱) قابل مشاهده می‌باشد.

```
multi_model <- lm(health_bills ~ stroke + bmi + age, data = non_miss_df)
summary(multi_model)
```

```
Call:
lm(formula = health_bills ~ stroke + bmi + age, data = new_df)

Residuals:
    Min      1Q  Median      3Q     Max
-3127.5  -359.2    -5.4   322.3  3632.9

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1811.7935    29.5966  61.216  < 2e-16 ***
stroke1     2563.9980    38.8656  65.971  < 2e-16 ***
bmi           40.5935     1.0308  39.380  < 2e-16 ***
age            1.0441     0.3687   2.832  0.00465 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 534.3 on 4905 degrees of freedom
Multiple R-squared:  0.5805,    Adjusted R-squared:  0.5803
F-statistic:  2263 on 3 and 4905 DF,  p-value: < 2.2e-16
```

تصویر(۴۱): مدل multiple linear regression ساخته شده برای پیش‌بینی مقدار 'health_bills'

(C

همانطور که می‌دانیم، متغیر R^2، نمایانگر درصدی از variability از متغیر response است که توسط مدل توضیح داده می‌شود. همچنین می‌دانیم هر متغیر explanatory که به مدل اضافه کنیم، هرچقدر هم که بی‌ربط باشد، باعث می‌شود R^2 افزایش یابد و این باعث این تفکر اشتباه می‌شود که هرچقدراز متغیرهای بیشتری استفاده کنیم، مدل بهتر خواهد شد. برای حل این مشکل از R^2_{adj} استفاده می‌کنیم تا واقعا predictorهایی که خوب هستند و به‌اندازه‌ی زیادی R^2 را افزایش می‌دهند در مدل ظاهر شوند. باتوجه به این توضیحات و مدلی که در بالا برای این متغیرها ساختیم، مقدار R^2 و R^2_{adj} به شرح زیر می‌باشد:

R^2 :

multi_model_sumry$r.squared

> 0.5805146

R^2_{adj}:

multi_model_sumry$adj.r.squared

> 0.5802581

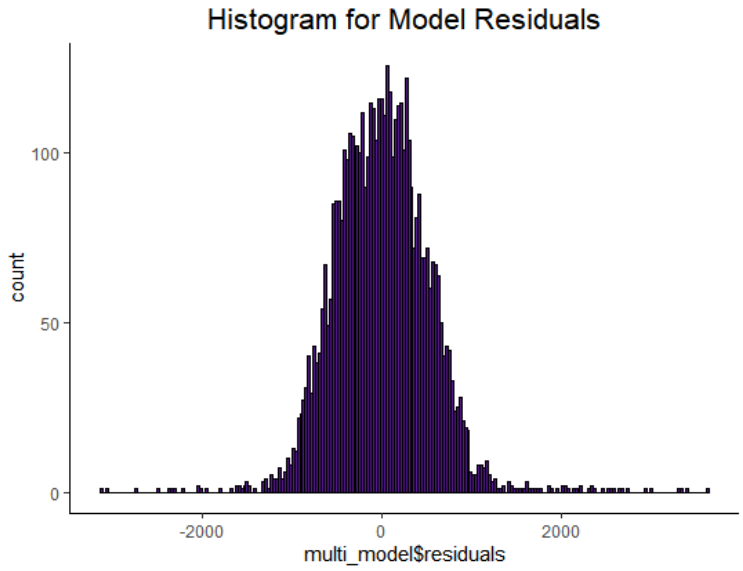
طبق مقادیر به‌دست آمده، دو مقدار R^2 و R^2_{adj} باهم برابرمی‌باشند و می‌گوییم 58% از variability متغیر response توسط این مدل توضیح داده می‌شود.

(D

باتوجه به مقدار R^2_{adj} که در مورد قبل به دست آوردیم، مدلی که ساختیم، 58% از variability موجود در متغیر response را توضیح می‌دهد، بنابراین 58% مدل به دیتا فیت شده‌است. هرچقدر مقدار R^2_{adj} به یک(به طور درصدی به 100) نزدیکتر باشد، مدلی که ساختیم مدل به دیتا بهتر فیت می‌شود.

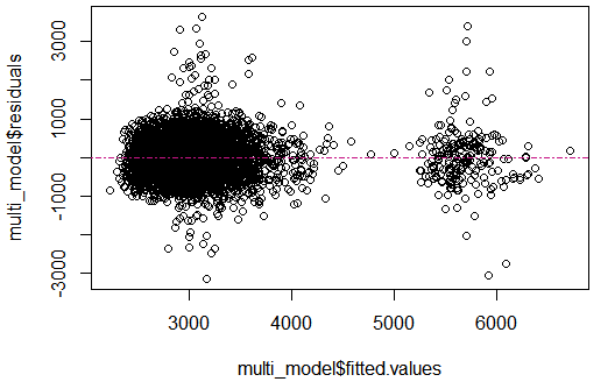
راه دیگر بررسی اینکه چقد مدلی که داریم به دیتا فیت شده‌است؛ استفاده از p-value به دست آمده از تست ANOVA بر روی این مدل می‌باشد. هرچه مقدار p-value کمتر باشد، مدلی که داریم بهتر است. طبق تصویر(۴۱) مقدار p-value حاصل از F test قابل مشاهده می‌باشد که 2.2e-16> می‌باشد.

یکی دیگر از راه‌های بررسی خوب بودن مدل هیستوگرام residualها می‌باشد که هرچقدر توزیع آن‌ها به نرمال با میانگین نزدیک تر باشد، مدل بهتر است. به همین منظور نمودار هیستوگرام برای residualها در ایم مدل را رسم نموده‌ایم که به در تصویر(۴۲) قابل مشاهده می‌باشد.

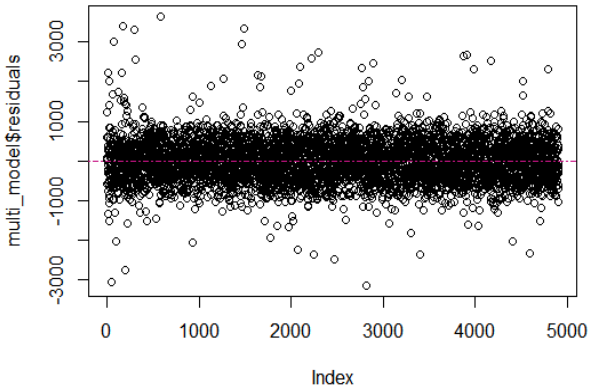


تصویر(۴۲): هیستوگرام برای نمایش توزیع residualها در مدل multiple linear regression

همانطور که در تصویر(۴۲) قابل مشاهده می‌باشد، توزیع residualها یک توزیع نرمال است، بنابراین مدلی که ساخته‌ایم یک مدل خوب است و می‌تواند به خوبی به دیتا فیت شود. رابطه‌ی بین متغیرهای explanatory و response هم همانطور که در تصویر(۴۰) مشاهده کردیم، رابطه‌ی خطی بود. با توجه به دو تصویر(۴۳) و (۴۴) نیز مشاهده می‌کنیم که constant variability در residualها برقرار است و همچنین از هم مستقل هستند. بنابراین مدلی که داریم مدل reliable و خوبی می‌باشد.



تصویر(۴۳): residual plot برای بررسی constant variability



تصویر(۴۴): residual plot برحسب index برای بررسی شرط استقلال

(E

به منظور پیدا کردن بهترین مدل، باید بهترین متغیرهای explanatory را از بین متغیرهای این مجموعه داده انتخاب کنیم. به همین منظور ما از دو روش 'Backward elimination' و 'forward selection' برای انتخاب متغیرها براساس معیارهای 'R^2_{adj}' و 'P-value' استفاده شده است. برای پیاده‌سازی این روش برمبنای 'R^2_{adj}'از دو تابع به نام‌های forwardSelection() و backwardSelection() استفاده شده‌است. برای 'p-value' نیز از تابع SignifReg() موجود در پکیج 'SignifReg' استفاده شده است.

- R^2_{adj}: هرچه بیشتر باشد مدلی که داریم بهتر است و توان بیشتری دارد.
- P-value: هرچه از مقدار significant level کمتر باشد، متغیرexplanatory داده شده significant تر است و در نتیجه مدل بهتری خواهیم داشت

Backward Elimination based on R^2_{adj} :

```
#adjusted_r_squared
backwardSelection <- function(df, response, max_steps = NaN, verbose = TRUE){

  selected_vars <- names(df)
  selected_vars <- selected_vars[-which(selected_vars == response)] # removing response variable
  if (is.na(max_steps)){
    max_steps <- length(selected_vars) - 1
  }
  current_adjR <- 0
  best_formula <- paste(response, paste(selected_vars, collapse = ' + '), sep = ' ~ ')
```

```
 step = 1
 while(step <= max_steps){
  results = c()
  for (var in selected_vars){
   vars <- paste(selected_vars[-which(selected_vars == var)], collapse = ' + ')

   formula = paste(response, vars, sep = ' ~ ')
   summary <- summary(lm(as.formula(formula), data = df))
   results <- c(results, summary$adj.r.squared)
  }

  if (max(results) > current_adjR){
   current_adjR <- max(results)
   selected_vars <- selected_vars[-which.max(results)]
   best_formula <- paste(response, paste(selected_vars, collapse = ' + '), sep = ' ~ ')
   if (verbose){
    cat(paste("\n\nStep ", step, ":\n"))
    print(best_formula)
    cat(paste("Adjusted R Squared: ", current_adjR))
   }
  } else{
   if (verbose){
    cat(paste("\n\nNo improvment in Adjusted R Squared, finished in step ", step-1))
   }
   break
  }
  step <- step + 1
 }

 return (lm(formula = as.formula(best_formula), data = df))
}
reg_backw <- backwardSelection(non_miss_df[-c(1)], response = "health_bills", verbose = TRUE)
```

خروجی تابع Backward Elimination براساس R^2_{adj} در تصویر(۴۵) قابل مشاهده می‌باشد.

```
Step  1 :
[1] "health_bills ~ gender + age + hypertension + heart_disease + ever_married + work_type + Residence_type + avg_glucose_level + bmi + stroke"
Adjusted R Squared:  0.585272884866462

Step  2 :
[1] "health_bills ~ age + hypertension + heart_disease + ever_married + work_type + Residence_type + avg_glucose_level + bmi + stroke"
Adjusted R Squared:  0.585339622477796

Step  3 :
[1] "health_bills ~ age + heart_disease + ever_married + work_type + Residence_type + avg_glucose_level + bmi + stroke"
Adjusted R Squared:  0.585398312503123

No improvment in Adjusted R Squared, finished in step  3
```

تصویر(۴۵): خروجی الگوریتم Backward Elimination براساس R^2_{adj}

Backward Elimination based on P-value:

```
nullmodel = lm(health_bills~1, new_df)
fullmodel = lm(health_bills~., new_df)
scope = list(lower=formula(nullmodel),upper=formula(fullmodel))

fitt_2 <- lm(health_bills ~ ., data =new_df)

reg_backw_pval <- SignifReg(fitt_2, scope = scope,alpha = 0.05,
                    direction = "backward",criterion = "p-value",trace = TRUE)
```

به دلیل آنکه خروجی تابع هرمرحله را نشان می‌دهد و خیلی طولانی بود، از آوردن آن خودداری کرده و فقط نتیجه‌ی آخر در جدول(۱۷) آورده شده است.

جدول(۱۷): متغیرهای انتخاب شده در روش Backward Elimination براساس p-value

explanatory	P-value
age	9.06e-11
hypertension	0.000711
ever_married	6.27e-05
avg_glucose_level	3.17e-07
stroke	< 2e-16

Forward Selection based on R^2_{adj} :

```
#adjusted_r_squared
forwardSelection <- function(df, response, max_steps = NaN, verbose = TRUE){

 selected_vars = c()
 remain_vars <- names(df)
 remain_vars <- remain_vars[-which(remain_vars == response)] # removing response variable
 if (is.na(max_steps)){
  max_steps <- length(remain_vars)
 }
 current_adjR <- 0
 best_formula <- ""
 step = 1
 while(step <= max_steps){
```

```
   results = c()
   for (var in remain_vars){
    if (is.null(selected_vars)){
     vars <- var
    } else{
     vars <- paste(paste(selected_vars, collapse = ' + '), var, sep = ' + ')
    }
    formula = paste(response, vars, sep = ' ~ ')
    summary <- summary(lm(as.formula(formula), data = df))
    results <- c(results, summary$adj.r.squared)
   }

   if (max(results) > current_adjR){
    current_adjR <- max(results)
    selected_vars <- c(selected_vars, remain_vars[which.max(results)])
    remain_vars <- remain_vars[-which.max(results)]
    best_formula <- paste(response, paste(selected_vars, collapse = ' + '), sep = ' ~ ')
    if (verbose){
     cat(paste("\n\nStep ", step, ":\n"))
     print(best_formula)
     cat(paste("Adjusted R Squared: ", current_adjR))
    }
   } else{
    if (verbose){
     cat(paste("\n\nNo improvment in Adjusted R Squared, finished in step ", step-1))
    }
    break
   }
   step <- step + 1
  }
  return (lm(formula = as.formula(best_formula), data = df))
}

reg_forw <- forwardSelection(non_miss_df[-c(1)], response = "health_bills", verbose = TRUE)
```

Forward Selection based on P-value:

```
fitt_1 <- lm(health_bills ~ 1, data =new_df)
nullmodel = lm(health_bills~1, new_df)
fullmodel = lm(health_bills~., new_df)
scope = list(lower=formula(nullmodel),upper=formula(fullmodel))
reg_forw_pval <- SignifReg(fitt_1, scope = scope,alpha = 0.05,
                    direction = "forward",criterion = "p-value",trace = TRUE)
```

خروجی تابع Forward Selection براساس R^2_{adj} در تصویر(۴۵) قابل مشاهده می‌باشد.

```
Step  1 :
[1] "health_bills ~ stroke"
Adjusted R Squared:  0.423322588497727

Step  2 :
[1] "health_bills ~ stroke + bmi"
Adjusted R Squared:  0.57965763886088

Step  3 :
[1] "health_bills ~ stroke + bmi + heart_disease"
Adjusted R Squared:  0.584930340118203

Step  4 :
[1] "health_bills ~ stroke + bmi + heart_disease + work_type"
Adjusted R Squared:  0.585118741936992

Step  5 :
[1] "health_bills ~ stroke + bmi + heart_disease + work_type + avg_glucose_level"
Adjusted R Squared:  0.585221003242891

Step  6 :
[1] "health_bills ~ stroke + bmi + heart_disease + work_type + avg_glucose_level + Residence_type"
Adjusted R Squared:  0.58526310684758

Step  7 :
[1] "health_bills ~ stroke + bmi + heart_disease + work_type + avg_glucose_level + Residence_type + age"
Adjusted R Squared:  0.585296735549773

Step  8 :
[1] "health_bills ~ stroke + bmi + heart_disease + work_type + avg_glucose_level + Residence_type + age + ever_married"
Adjusted R Squared:  0.585398312503124
```

تصویر(۴۶): خروجی الگوریتم Forward Selection براساس R^2_{adj}

نتیجه‌ی حاصل از Forward Selection براساس P-value در جدول(۱۸) قابل مشاهده می‌باشد.

جدول(۱۸): متغیرهای انتخاب شده در روش Forward Selection براساس p-value

explanatory	P-value
stroke	< 2e-16
bmi	< 2e-16
heart_disease	< 2e-16

همانطور که قابل مشاهده می‌باشد، دو روش "Forward Selection" و "Backward Elimination" جواب یکسانی را باتوجه به مقدار R^2_{adj} تولید کردند. این متغیرها را به عنوان متغیرهای بهترین مدل انتخاب می‌کنیم و مدلی باتوجه به این متغیرها می‌سازیم. متغیرهای انتخاب شده به شرح زیر می‌باشد:

age, heart_disease, ever_married, work_type, Residence_type, avg_glucose_level, bmi, stroke

گزارش مدل ساخته‌شده توسط متغیرهای انتخاب شده در تصویر(۴۷) قابل مشاهده می‌باشد:

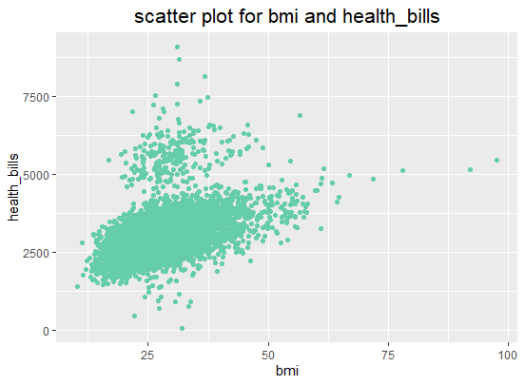
```
best_model <- lm(heatlth_bills ~ age + heart_disease + ever_married +
                 Residence_type + avg_glucose_level + bmi + stroke, data = non_miss_df[-c(1)])

summary(best_model)
```

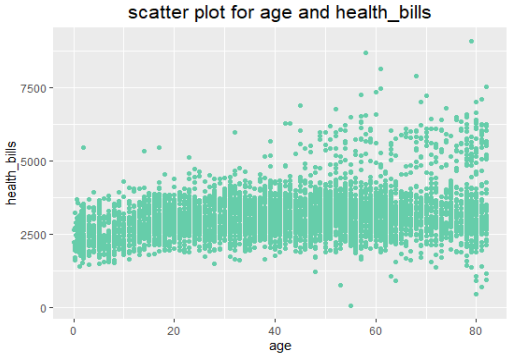
```
Call:
lm(formula = health_bills ~ age + heart_disease + ever_married +
    work_type + Residence_type + avg_glucose_level + bmi + stroke,
    data = new_df[-c(1)])

Residuals:
    Min      1Q  Median      3Q     Max
-3382.0  -350.1     5.1   331.9  3336.9

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           1799.8539    33.8485  53.174  < 2e-16 ***
age                      0.9926     0.5658   1.754   0.0794 .
heart_disease          263.2892    36.6936   7.175 8.29e-13 ***
ever_marriedYes        -33.3500    22.4845  -1.483   0.1381
work_typeGovt_job      -17.1277    37.8391  -0.453   0.6508
work_typeNever_worked  139.1658   115.3738   1.206   0.2278
work_typePrivate        18.0399    31.1662   0.579   0.5627
work_typeSelf-employed -29.4992    38.8527  -0.759   0.4477
Residence_typeUrban    -18.8005    15.1698  -1.239   0.2153
avg_glucose_level        0.2257     0.1790   1.261   0.2075
bmi                     40.8000     1.1004  37.078  < 2e-16 ***
stroke                2530.6621    39.0640  64.782  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 531.1 on 4897 degrees of freedom
Multiple R-squared:  0.5863,    Adjusted R-squared:  0.5854
F-statistic:   631 on 11 and 4897 DF,  p-value: < 2.2e-16
```

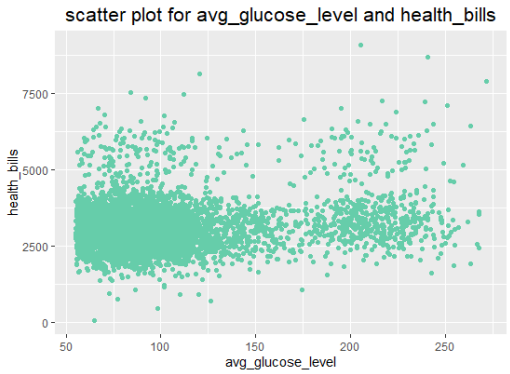
تصویر(۴۷): مدل ساخته شده با استفاده از متغیرهای انتخاب شده توسط روش‌های Backward و Forward براساس R^2_{adj}

(F

- **linearity:** به منظور بررسی این ویژگی scatter plot دوبه‌دوی متغیرهای explanatory با متغیر response را رسم می‌کنیم. لازم به ذکر است که نمودار فقط برای متغیرهای numerical رسم شده‌است، زیرا برای متغیرهای categorical این فرض برقرار است. به دلیل آنکه متغیرهای categorical طبق تعریف، فرض خطی بودن را برآورده می‌کنند، زیرا آن‌ها دو نقطه داده ایجاد کرده و دو نقطه خط مستقیم را تعریف می‌کند. چیزی به عنوان رابطه غیرخطی برای یک متغیر تنها با دو مقدار وجود ندارد. نمودارهای مربوط به این قسمت در تصویر(۴۸) قابل مشاهده می‌باشد.





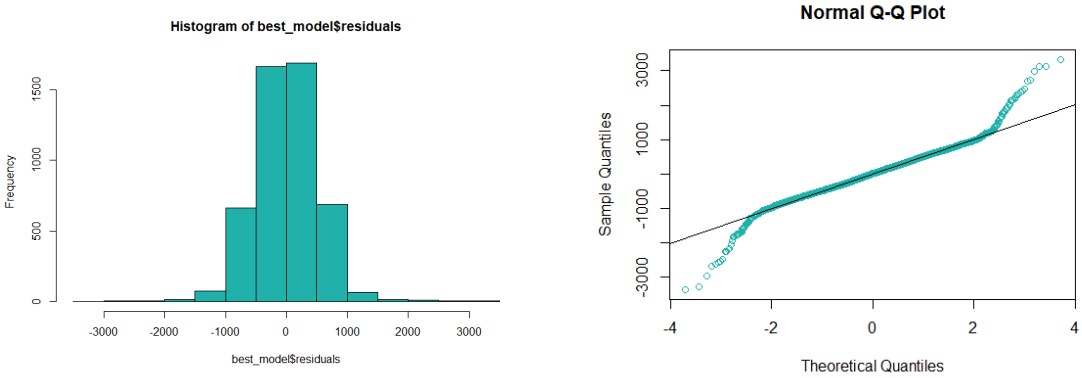


تصویر(۴۸): نمودار scatter plot به منظور بررسی شرط linearity بین متغیر response و متغیرهای explanatory

همانطور که در نمودارهای بالا قابل مشاهده می‌باشد، رابطه‌ی بین متغیرهای explanatory با متغیر responce از از نوع خطی می‌باشد.

- **Nearly normal residuals:** برای بررسی این فرض نمودار Q-Q plot و هیستوگرام را رسم می‌کنیم نمودارهای این قسمت در تصویر(۴۹) قابل مشاهده می‌باشد.

```
hist(best_model$residuals, col="lightseagreen")

qqnorm(best_model$residuals, col="lightseagreen")
```



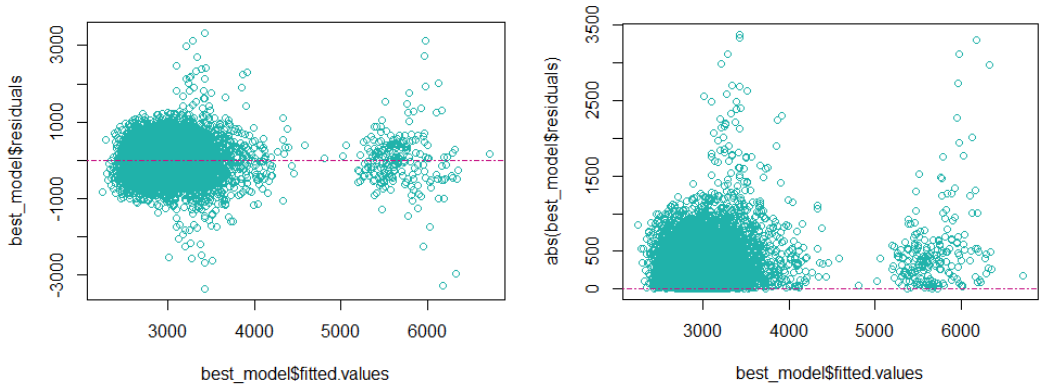
تصویر(۴۹): نمودار Q-Q plot و هیستوگرام به منظور بررسی شرط nearly normal residuals

طبق نمودارهای بالا residualها توزیع تقریبا نرمال دارند.

- **Constant variability:** برای نشان دادن برقراری این شرط نیز، نمودار residualها بر حسب $\hat{y}$ را رسم می‌کنیم. نمودارها در تصویر(۵۰) قابل مشاهده می‌باشد.

```
plot(best_model$residuals~ best_model$fitted.values, col = "lightseagreen")
```

```
plot(abs(best_model$residuals)~ best_model$fitted.values, col = "lightseagreen")
```



تصویر(۵۰): نمودار residual plot برای بررسی شرط Constant variability

باتوجه به نمودارهای بالا این شرط نیز برقرار می‌باشد. زیرا نقطه‌ای وجود ندارد که بتوانیم بگوییم در آن نقطه مقدار residualها خیلی متفاوت باشد.

(G

الگوریتم k-fold-cross-validation عملکرد مدل را در زیرمجموعه مختلف داده‌های آموزش ارزیابی می‌کند و سپس میانگین خطای پیش بینی را محاسبه می کند. الگوریتم به شرح زیر است:

۱. مجموعه داده ها را به طور تصادفی به k زیرمجموعه تقسیم می‌کند.
۲. یک زیر مجموعه را رزرو کرده و مدل را روی زیر مجموعه‌های دیگر آموزش می‌دهد.
۳. مدل را در زیرمجموعه رزرو شده آزمایش می‌کند و خطای پیش‌بینی را محاسبه می‌کند.
۴. این روند را تکرار می‌کند تا زمانی که هر یک از k زیرمجموعه به عنوان مجموعه تست مورد بررسی قرار گیرند.
۵. در نهایت میانگین k خطاهای ثبت شده را محاسبه می‌کند. این خطای اعتبارسنجی متقابل است که به عنوان معیار عملکرد برای مدل عمل می کند.

برای پیاده‌سازی این الگوریتم از توابع trainControl() و train() موجود در پکیج 'caret' استفاده شده‌است. کدهای زده شده برای این قسمت و خروجی به ازای هر دو مدل 'B' و 'E' به شرح زیر می‌باشد:

```
#training control
set.seed(123)
train.control <- trainControl(method = "cv", number = 5)

# Train the model in B
model_B <- train(health_bills ~ stroke + bmi + age, data = new_df, method = "lm",
                 trControl = train.control)

# Summarize the results
print(model_B)
```

```
# Train the model in E
model_E <- train(health_bills ~ age + heart_disease + ever_married + work_type +
                   Residence_type + avg_glucose_level + bmi + stroke,
                 data = new_df, method = "lm",
                 trControl = train.control)

# Summarize the results
print(model_E)
```

برای مدل ایجاد شده در مورد 'B' خروجی در تصویر(۵۱) قابل مشاهده می‌باشد.

```
Linear Regression

4909 samples
   3 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 3928, 3926, 3927, 3928, 3927
Resampling results:

  RMSE      Rsquared   MAE
  534.1477  0.5780687  408.1944

Tuning parameter 'intercept' was held constant at a value of TRUE
```

تصویر(۵۱): خروجی حاصل از الگوریتم 5_fold-cross-validation برای مدل 'B'

برای مدل ایجاد شده در مورد 'E' خروجی در تصویر(۵۲) قابل مشاهده می‌باشد.

```
Linear Regression

4909 samples
   8 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 3927, 3928, 3928, 3925, 3928
Resampling results:

  RMSE      Rsquared   MAE
  532.1124  0.5846193  407.4851

Tuning parameter 'intercept' was held constant at a value of TRUE
```

تصویر(۵۲): خروجی حاصل از الگوریتم 5_fold-cross-validation برای مدل 'E'

همانطور که در تصاویر بالا قابل مشاهده می‌باشد، مقدار MSE به‌دست آمده در مدل 'E' کمتر است و می‌دانیم هرچه مقدار MSE کمتر باشد، مدل بهتر خواهد بود. بنابراین نتیجه می‌گیریم که مدلی که در مورد 'E' با پیدا کردن بهترین متغیرهای explanatory ایجاد کردیم، مدل بهتری نسبت به مدل ساخته شده در مورد 'B' می‌باشد.

Question 6

متغیر باینری انتخاب شده برای این قسمت، متغیر'heart_disease' می‌باشد. متغیرهای explanatory انتخاب شده نیز 'age'، 'gender'، 'smoking_type'، 'Residence_type'، 'avg_glucose_level' می‌باشد.

*لازم به ذکر است که در متغیر 'gender' سه سطح 'Other'، 'Male' و 'Female' وجود دارد. به دلیل آنکه فقط یک سطر با سطح 'Other' داریم، آن را جزو متغیر 'Male' در نظر گرفتیم.

(A

با استفاده از تابع glm() یک logistic regression بر روی داده‌ها فیت کردیم. خروجی در تصوی(۵۳) قابل مشاهده می‌باشد.

```
glm_model <- glm(heart_disease ~ age + gender + avg_glucose_level +
                 smoking_status, data = new_df,family = binomial)

summary(glm_model)
```

```
Call:
glm(formula = heart_disease ~ age + gender + avg_glucose_level +
    smoking_status + Residence_type, family = binomial, data = new_df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4662  -0.3083  -0.1471  -0.0610   3.8927

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)               -8.570329   0.442573 -19.365  < 2e-16 ***
age                        0.080462   0.005586  14.403  < 2e-16 ***
genderMale                 0.777968   0.144349   5.389 7.07e-08 ***
avg_glucose_level          0.005466   0.001180   4.634 3.60e-06 ***
smoking_statusnever smoked -0.201445   0.182026  -1.107  0.26843
smoking_statussmokes        0.593660   0.206696   2.872  0.00408 **
smoking_statusUnknown      -0.179217   0.224082  -0.800  0.42384
Residence_typeUrban        -0.109274   0.141892  -0.770  0.44123
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1934.6  on 4908  degrees of freedom
Residual deviance: 1460.3  on 4901  degrees of freedom
AIC: 1476.3

Number of Fisher Scoring iterations: 7
```

تصویر(۵۳): خروجی مدل ساخته شده توسط logistic regression

تفسیر برای intercept: لگاریتم شانس ابتلا به بیماری قلبی در حالتی که همه‌ی متغیرهای explanatory را برابر صفر قرار داده‌ایم، برابر8.570329- می‌باشد.

تفسیر برای متغیر 'age': به ازای هر یکسال افزایش در سن افراد، لگاریتم شانس ابتلا شدن به بیماری قلبی 0.082989 افزایش پیدا می‌کند.

تفسیر برای متغیر 'gender': اگر همه‌ی predictorهای دیگر را ثابت در نظر بگیریم، لگاریتم شانس ابتلا شدن مردان به بیماری قلبی، 0.7738 واحد بیشتر از شانس ابتلا شدن زنان به بیماری قلبی می‌باشد.

تفسیر برای متغیر 'avg_glucose_level': به ازای هر یک واحد افزایش در متوسط سطح گلوکز افراد، لگاریتم شانس ابتلا شدن به بیماری قلبی 0.005446 افزایش پیدا می‌کند.

تفسیر برای متغیر 'smoking_status(smoke)': اگر همه‌ی predictorهای دیگر را ثابت در نظر بگیریم، لگاریتم شانس ابتلا شدن افرادی که سیگار می‌کشند، به بیماری قلبی، 0.5975 واحد بیشتر از شانس ابتلا شدن افرادی است که برخی اوقات سیگار می‌کشند.

تفسیر برای متغیر 'smoking_status(never smoke)': اگر همه‌ی predictorهای دیگر را ثابت در نظر بگیریم، لگاریتم شانس ابتلا شدن افرادی که سیگار نمی‌کشند، به بیماری قلبی، 0.203 واحد کمتر از شانس ابتلا شدن افرادی است که برخی اوقات سیگار می‌کشند.

تفسیر برای متغیر 'smoking_status(never smoke)': اگر همه‌ی predictorهای دیگر را ثابت در نظر بگیریم، لگاریتم شانس ابتلا شدن افرادی وضعیت سیگار کشیدن آن‌ها نامشخص است ، به بیماری قلبی، 0.1927 واحد کمتر از شانس ابتلا شدن افرادی است که برخی اوقات سیگار می‌کشند.

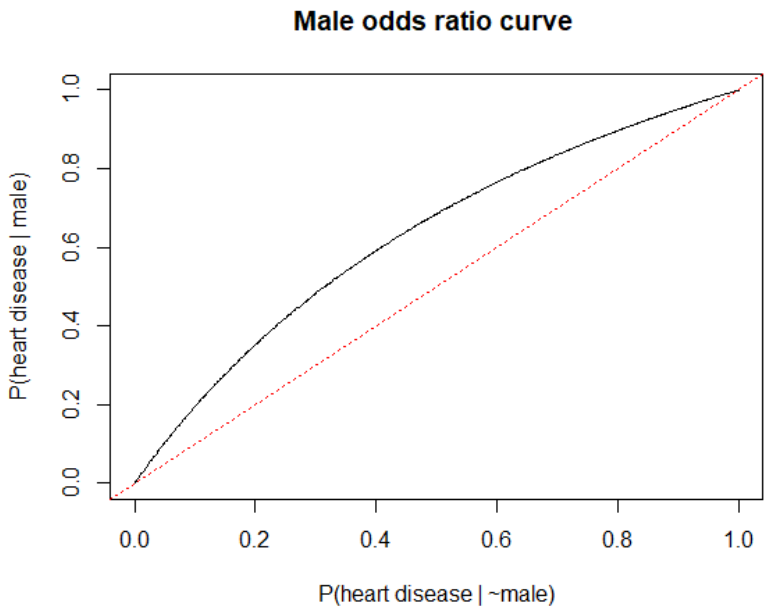
تفسیر برای متغیر 'Residense_type': اگر همه‌ی predictorهای دیگر را ثابت در نظر بگیریم، لگاریتم شانس ابتلا شدن افراد شهری به بیماری قلبی، 0.109 واحد کمتر از شانس ابتلا شدن افراد روستایی به بیماری قلبی می‌باشد.

(B

به منظور رسم نمودار odds ratio curve، متغیر 'gender' انتخاب شده‌است. برای P(heart disease | not male) مقادیر مختلف را در نظر گرفته و به ازای آن، طبق فرمول زیر، مقادیر P(heart disease | male) محاسبه می‌شود.

$$\text{P(heart disease | male)} = \frac{e^{0.777987} \times \frac{\text{P(heart disease | not male)}}{(1 - \text{P(heart disease | not male)})}}{(1 + e^{0.777987} \times \frac{\text{P(heart disease | not male)}}{(1 - \text{P(heart disease | not male)})})}$$

```
set.seed(42)
x_axis <-round(sort(runif(1000, min=0, max=1)),4)

var <- exp(glm_summary$coefficients[3])
y_axis <-(var * (x_axis / (1 - x_axis))) / (1 + (var * (x_axis / (1 - x_axis))))

plot(x_axis, y_axis, type="l",xlab="P(heart disease | ~male)", ylab="P(heart disease | male)")
abline(a = 0 , b = 1, col = "red", lty = 3)
title("Male odds ratio curve ")
```



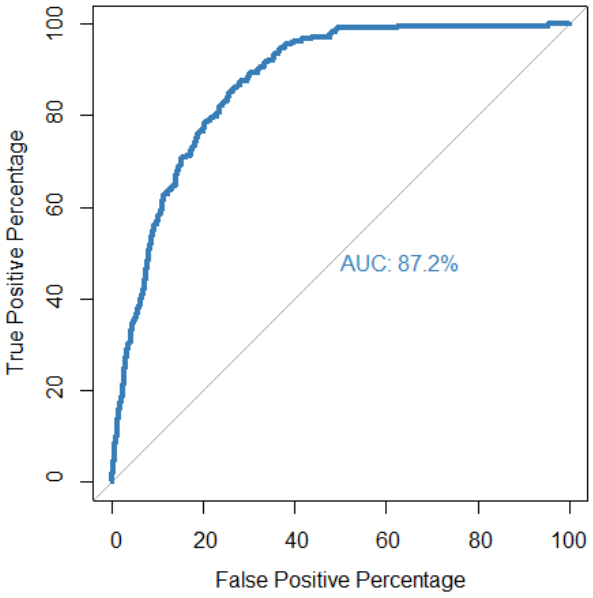
تصویر(۵۴): odds ration curve برای متغیر gender:Male

تصویر(۵۴) نماینگر odds ratio curve می‌باشد. در نمودار بالا مقادیر مختلفی برای احتمال بیماری قلبی داشتن به شرط مرد نبودن را در نظر گرفتیم و احتمال بیماری قلبی داشتن به شرط مرد بودن را از روی OR محاسبه نموده‌ایم. هرچه مقدار ORبزرگتر باشد، سطح زیر نمودار بیشتر می‌شود. برای نمودار ما مقدار OR برابر $e^{0.777987} = 2.177044$ می‌باشد. برای مثال زمانی که مقدار P(heart disease | not male) برابر 0.2 است، احتمال P(heart disease | male) تقریبا نزدیک به 0.4 می‌باشد.

(C

برای رسم ROC Curve از تابع roc() موجود در پکیج 'pROC' استفاده شده است. برای مدل ما این نمودار در تصویر(۵۵) قابل مشاهده می‌باشد.

```
library(pROC)

pred <- predict(glm_model, type = "response")

par(pty = "s")
roc_ <- roc(new_df$heart_disease, pred, plot = TRUE, legacy.axes = TRUE, percent = TRUE,
        xlab = "False Positive Percentage", ylab = "True Positive Percentage",
        col = "#377eb8", lwd = 4, print.auc = TRUE)
```



تصویر(۵۵): نمودار ROC برای مدل ساخته شده

این نمودار یک معیار مناسب است برای آنکه نشان دهد این مدل یا به عبارتی این طبقه‌بندی که ساختیم، طبقه‌بند خوبی است یا خیر. علت اهمیت این نمودار به این دلیل است که trade off بین sensitivity و specificity را برای همه‌ی thresholdهای ممکن نمایش می‌دهد و بسته به نیاز می‌توانیم threshold مورد نظر را انتخاب کنیم. همچنین performance مدلی که ساختیم را در برابر تصمیم-گیری شانسی، می‌توانیم بسنجیم(خط راست y=x موجود در نمودار بالا نماینگر تصمیم‌گیری شانسی می‌باشد که خط مربوط به مدل ما از آن دورتر باشد، به این معناست که مدلی که ساختیم بهتر است.)

باتوجه به مقدار AUC(مساحت زیر سطح نمودار) می‌توان خوب بودن یک مدل را بررسی کرد. هرچه این مقدار بالاتر باشد مدل یا همان طبقه‌بندی که داریم بهتر می‌شود. اگر مساحت بالا 90% باشد یک طبقه‌بند خیلی خوب داریم و بین 80% تا 90% نیز طبقه‌بند خوب محسوب می‌شود.

(D

یک متغیر explanatory را زمانی می‌گوییم significant است که مقدار p-value آن در مدلی که ایجاد کردیم، کمتر از significant level (α) باشد. باتوجه به این موضوع و خروجی مشاهده شده از مدل، متغیرهای 'age'، 'gender'، 'avg_glucose_level' و 'smoking_status' متغیرهای significant می‌باشند زیرا براساس آزمون فرض مقدار p-value به دست آمده برای این متغیرها کمتر از 0.05 می‌باشد. متغیر 'smoking_status' به دلیل اینکه یکی از سطح‌های آن (smoking_status:smoke) مقدار p-value کمتر از 0.05 دارد، بنابراین کل متغیر siginicant محسوب می‌شود، زیرا برای متغیرهای Categroical باید تمام سطوح آن متغیر را در نظر بگیریم و اگر حتی یکی از سطوح significant شده باشد، آنگاه کل متغیر significant می‌شود.

البته لازم به ذکر است این متغیرها در این مدلsignificant شده‌اند و امکان دارد اگر به طور جداگانه فقط با همین متغیر یک مدل دیگر بسازیم در آن حالت significant نشود، یا اگر یک متغیر دیگر به این مدل اضافه کنیم، دیگر این متغیر significant نباشد.

البته می‌توانیم به صورت جداگانه برای آزمون فرض را برای هر متغیر بیان کنیم، عنوان مثال برای متغیر age داریم:

H_0 : age = 0

H_A : age ≠ 0

فرض H_0 بیان می‌کند که متغیر 'age' یک significant predictor نمی‌باشد. در مقابل فرض H_A بیان می‌کند که این متغیر یک significant predictor می‌باشد.

p-value :

glm_summary$coefficient[26]

> 4.941313e-47

باتوجه به مقدار p-value به‌دست آمده، چون مقدار آن از $\alpha = 0.05$ کمتر می‌باشد، بنابراین فرض H_0 رد می‌شود و به عبارتی این متغیر یک متغیر significant می‌باشد.

(E

حال باتوجه به متغیرهایی که بیشترین significant contribution را دارند، که در مورد قبل مشخص شد، یک مدل جدید می‌سازیم، خروجی در تصویر(۵۶) قابل مشاهده می‌باشد:

```
Call:
glm(formula = heart_disease ~ age + gender + avg_glucose_level,
    family = binomial, data = new_df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1721  -0.3107  -0.1502  -0.0682   3.7798

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -8.463689   0.391578 -21.614  < 2e-16 ***
age               0.077807   0.005363  14.508  < 2e-16 ***
genderMale        0.826023   0.142570   5.794 6.88e-09 ***
avg_glucose_level 0.005395   0.001172   4.604 4.14e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1934.6  on 4908  degrees of freedom
Residual deviance: 1477.6  on 4905  degrees of freedom
AIC: 1485.6

Number of Fisher Scoring iterations: 7
```

تصویر(۵۶): مدل ساخته شده با استفاده از significant predictorهای مشخص شده در مورد 'D'

مدل ساخته شده در مقایسه با مدل 'A'، قابل مشاهده می‌باشد که مقدار شیب برای متغیر 'age' کاهش یافته است. ولی برای متغیر 'gender' افزایش یافته‌است. به همین نسبت مقادیر p-value تغییر داشته‌است ولی همچنان هر سه متغیر در این مدل نیز، متغیر significant می‌باشند. همچنین مقدار عرض از مبدا نیز خیلی کم افزایش داشته‌است.

تفسیر عرض از مبدا و شیب برای این مدل به شرح زیر می‌باشد:

تفسیر برای intercept: لگاریتم شانس ابتلا به بیماری قلبی در حالتی که همه‌ی متغیرهای explanatory را برابر صفر قرار داده‌ایم، برابر8.4637- می‌باشد.

تفسیر برای متغیر 'age': به ازای هر یکسال افزایش در سن افراد، لگاریتم شانس ابتلا شدن به بیماری قلبی 0.0778 افزایش پیدا می‌کند.

تفسیر برای متغیر 'gender': اگر همه‌ی predictorهای دیگر را ثابت در نظر بگیریم، لگاریتم شانس ابتلا شدن مردان به بیماری قلبی، 0.826 واحد بیشتر از شانس ابتلا شدن زنان به بیماری قلبی می‌باشد.

تفسیر برای متغیر 'avg_glucose_level': به ازای هر یک واحد افزایش در متوسط سطح گلوکز افراد، لگاریتم شانس ابتلا شدن به بیماری قلبی 0.00539 افزایش پیدا می‌کند.

(F

به منظور رسم نمودار Utility، ابتدا به ازای هر خروجی یک ارزشی در نظر گرفته‌ایم که در جدول(۱۹) قابل مشاهده می‌باشد.

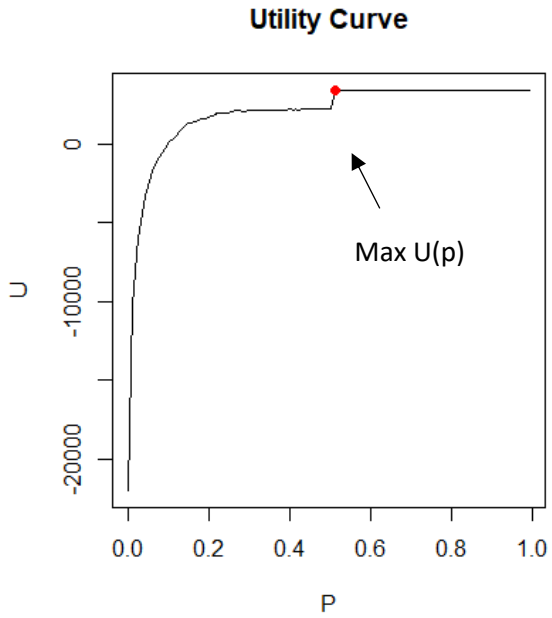
جدول(۱۹): ارزش در نظر گرفته شده به ازای هر خروجی

outcome	Utility
True positive	1
True negative	1
False positive	-10
False negative	-5

سپس 200 مقدار برای threshold (بین 0 و 1) در نظر می‌گیریم. باتوجه به هرکدام از این thresholdها مقادیر TP، TN، FP و FN یا به عبارتی confusion matrix را حساب می‌کنیم. در نهایت با استفاده از مقادیر به‌دست آمده و باتوجه به فرمول تابع Utility، خروجی را محاسبه می‌کنیم و نمودار را رسم می‌کنیم. کدهای زده شده برای این قسمت در پایین اورده شده‌است. نمودار به دست‌آمده نیز در تصویر(۵۷) قابل مشاهده می‌باشد.

$$U(p) = TP(p) + TN(p) - 10\ FP(p) - 5\ FN(p)$$

```
set.seed(123)
threshold <- round(sort(runif(200, min=0, max=1)),4)
utility_list <- list()

c <- 0
for (i in threshold){
  predicted_values <- ifelse(predict(significant_model,type="response")>i,1,0)
  actual_values <- significant_model$y
  conf_matrix <- table(predicted_values,actual_values)
  c <- c + 1
  if(is.na(conf_matrix[3]) && is.na(conf_matrix[4])){
    utility <- conf_matrix[1] - (5 * conf_matrix[2])
    utility_list[[c]] <- utility
  }
  else{
    utility <- conf_matrix[1] + conf_matrix[4] - (5 * conf_matrix[2]) - (10 * conf_matrix[3])
    utility_list[[c]] <- utility
  }
}

utility_df <- data.frame(matrix(unlist(utility_list), nrow=200, byrow=TRUE),stringsAsFactors=FALSE)
colnames(utility_df) <- c('utility')
utility_df$threshold <- threshold

plot(utility_df$threshold, utility_df$utility, type="l",xlab="P",
     ylab="U")
y <- c(max(utility_df$utility))
x <- utility_df$threshold[utility_df$utility == y][1]
points(x, y, pch = 19, col ="red")
title("Utility Curve")
```



تصویر(۵۷): نمودار Utility Curve

بهترین مقدار threshold در نمودار با نقطه‌ی قرمز رنگ مشخص شده‌است. مختصات این نقطه به شرح زیر می‌باشد:

```
y <- c(max(utility_df$utility))
x <- utility_df$threshold[utility_df$utility == y][1]
```

→ threshold : 0.5115
Utility : 3451

Question 7

طبق آنچه در صورت سوال خواسته شده‌است، ابتدا یک ستون با نام 'high_medical_costs' به مجموعه داده اضافه کردیم. مقادیر این ستون به این صورت تعریف می‌شود که برای مقادیر بیشتر از median ستون 'health_bills'، برچسب '1' و برای گروه مقابل برچسب '0' در نظر گرفته‌شده است.

median : 3031.724

```
median_healthBills <- median(new_df$health_bills)
new_df$high_medical_costs <- ifelse(new_df$health_bills > median_healthBills , '1','0')
```

سپس در قدم بعد، این ستون را به عنوان متغیر response در نظر گرفته و یک مدل logistic regression با متغیرهای explanatory 'work_type'، 'age'، 'gender' و 'hypertension' ساخته شده‌است. خروجی در تصویر(۵۸) قابل مشاهده می‌باشد.

```
high_medical_costs_glm <- glm(high_medical_costs ~ age + work_type +
                              hypertension + gender, data = new_df, family = binomial)

summary(high_medical_costs_glm)
```

```
Call:
glm(formula = high_medical_costs ~ age + work_type + hypertension +
    gender, family = binomial, data = new_df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6537  -1.1826  -0.6646   1.1116   1.7938

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -1.398663   0.099550 -14.050  < 2e-16 ***
age                      0.010580   0.001815   5.829 5.57e-09 ***
work_typeGovt_job        0.880423   0.145959   6.032 1.62e-09 ***
work_typeNever_worked    0.993917   0.438707   2.266   0.0235 *
work_typePrivate         1.019376   0.121679   8.378  < 2e-16 ***
work_typeSelf-employed   0.890410   0.151332   5.884 4.01e-09 ***
hypertension             0.493445   0.109558   4.504 6.67e-06 ***
genderMale               0.102184   0.060504   1.689   0.0912 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6805.3  on 4908  degrees of freedom
Residual deviance: 6469.0  on 4901  degrees of freedom
AIC: 6485

Number of Fisher Scoring iterations: 4
```

تصویر(۵۸): مدل logistic regression ساخته شده برای پیش‌بینی مقدار 'high_medical_costs'

همانطور که در تصویر بالا قابل مشاهده می‌باشد، تمام متغیرهای explanatory به جز 'gender'، متغیرهای significant می‌باشند. اما از بین این significant predictorها، بیش‌ترین تاثیر را متغیر 'work_type' دارد، زیرا p-value برای سطح 'Private' این متغیر، از بقیه p-valueها کمتر می‌باشد. متغیر بعدی متغیر 'age' می‌باشد که بیشترین تاثیر را در پیش‌بینی دارد.