به نام خدا

دانشگاه تهران

دانشکده مهندسی برق و کامپیوتر

استنباط آماری

فاز دوم پروژه

دیتاست insurance

نام و نام خانوادگی	مهسا تاجیک
شماره دانشجویی	810198126
تاریخ ارسال گزارش	99/4/3

فهرست گزارش سوالات

QUESTION1

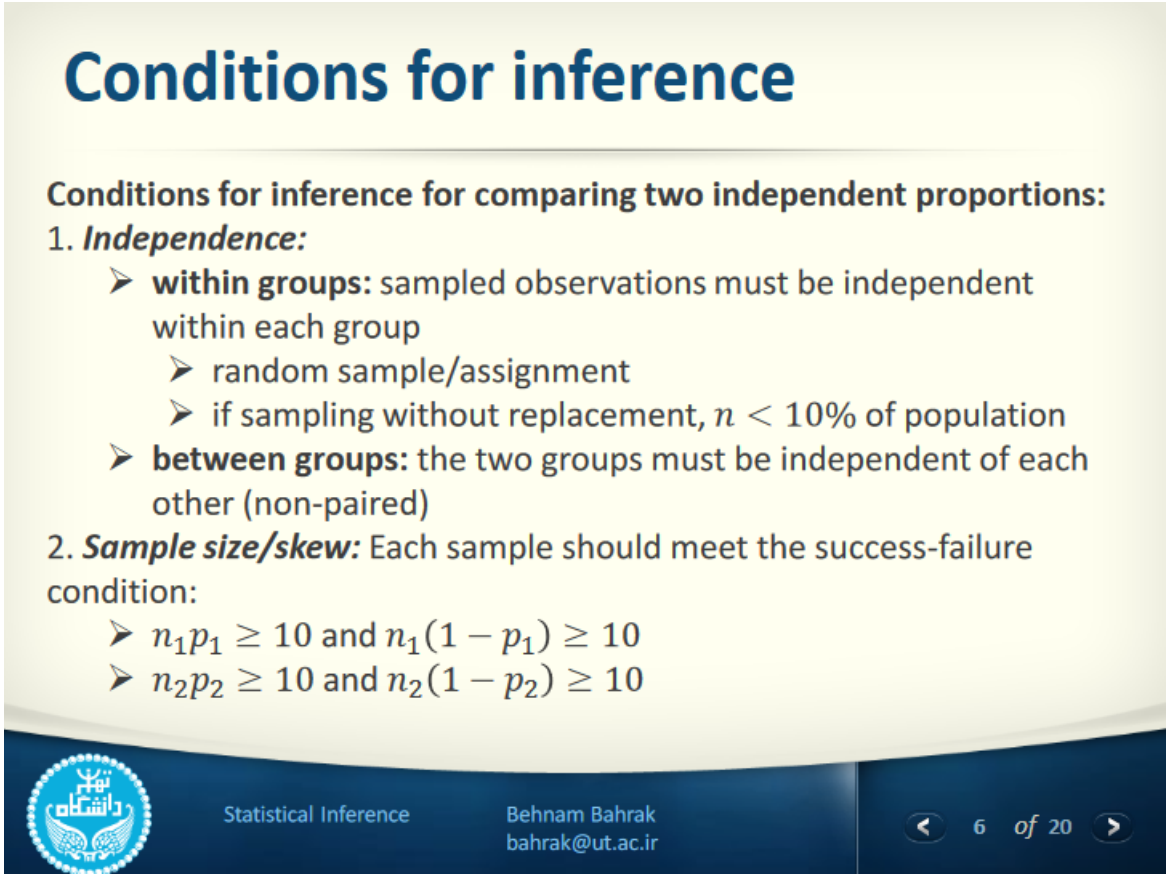
در این بخش خواسته شده که دو متغیر کتگوریکال را در نظر بگیریم که حداقل یکی از آنها بیشتر از دو سطح داشته باشد. متغیر های smoker دارای 2 سطح و region دارای 4 سطح را انتخاب می کنیم.

.a

روش 1:

می خواهیم بازه ی اطمینان 95 درصد را برای تفاضل نسبت های این دو متغیر محاسبه کنیم. سطح northeast از متغیر region را انتخاب می کنیم و می خواهیم نسبت سیگاری های این ناحیه را با سیگاری های بقیه ی نواحی(مجموع 3 ناحیه ی دیگر) مقایسه کنیم.

برای اینکه بازه ی اطمینان را محاسبه کنیم ابتدا باید شرایط قضیه ی حدمرکزی را بررسی کنیم که برقرار باشند. شرایط در اسلاید زیر آمده اند.



1) استقلال : از آنجاییکه تعداد کل داده های دیتاست برابر است با 1338 و جمعیت کوچکی است و خودش یک نمونه است که از جامعه گرفته شده بنابراین همان را بعنوان یک نمونه در نظر میگیریم و

فرض می کنیم که بصورت رندوم نمونه برداری انجام شده و برای هر فرد تنها یکبار اطلاعات وارد شده یعنی نمونه برداری بدون جاگذاری است پس استقلال درون گروهی داریم. استقلال برون گروهی نداریم زیرا برای بررسی هر دو متغیر از یک دسته داده استفاده می کنیم.

2) اندازه نمونه : میخواهیم نسبت سیگاری های ناحیه ی northeast را با نسبت سیگاری های مجموع سه ناحیه ی دیگر مقایسه کنیم و ببینیم که با هم برابرند یا اختلاف دارند. برای چک کردن شرط اندازه برای بازه اطمینان از تخمین گرهای نقطه ای $\hat{p}_1$ و $\hat{p}_2$ استفاده میکنیم که مقادیر آنها و تعداد نمونه در ناحیه ی northeast و بقیه نواحی را به شکل زیر محاسبه میکنیم :

```
> n_northeast
northeast
      324
> n_others
northwest
     1014
> p_hat1
[1] 0.2067901
> p_hat2
[1] 0.1685595
>
```

$\hat{p}_1 = 0.2$

$\hat{p}_2 = 0.16$

n1 = 324

n2 = 1014

p1*n1 = 0.2*324=64.8> 10 , (1-p1)*n1=0.8*324=259.2>10

p2*n2=0.16*1014=162.24>10 , (1-p2)*n2=0.84*1014=851.76>10

همانطور که میبینیم شرط اندازه نمونه هم برقرار است با فرض برقراری شرط استقلال بنابراین می توانیم از قضیه حد مرکزی برای محاسبه بازه اطمینان استفاده کنیم:

CI = point estimate ± margin of error

$CI = (\hat{p}_1 - \hat{p}_2) \pm z^* \, SE_{(\hat{p}1 - \hat{p}2)}$

بازه اطمینان 95 درصد خواسته شده بنابراین $z^* = 1.96$

$\hat{p}_1$ نسبت سیگاری های ناحیه northeast و $\hat{p}_2$ نسبت سیگاری های بقیه نواحی است که در R آن ها را محاسبه کردیم سپس مقدار SE و درنهایت با استفاده از رابطه بالا مقدار بازه اطمینان را محاسبه می کنیم :

```
ct1 <- table(insurance$smoker, insurance$region, dnn=c("insurance","region"))
ptable <- prop.table(ct1)

n_region <- table(insurance$region)
n_northeast <- n_region[1]
n_others <- n_region[2]+n_region[3]+n_region[4]
p_hat1 <- ptable[2]/(ptable[1]+ptable[2])
p_hat2 <- ptable[4]+ptable[6]+ptable[8]/(ptable[3]+ptable[4]+ptable[5]+ptable[6]+ptable[7]+ptable[8])


SE <- sqrt((p_hat1*(1-p_hat1)/n_northeast)+(p_hat2*(1-p_hat2)/n_others))

lower <- (p_hat1-p_hat2)-1.96 * SE
upper <- (p_hat1-p_hat2)+1.96 * SE
CI <- c(lower,upper)
CI
```

مقادیر SE و CI در تصویر زیر آورده شده اند:

```
> SE
 northeast
0.02538644
> CI
  northeast   northeast
-0.01152675  0.08798810
>
```

تفسیر بازه ی اطمینان : 95 درصد اطمینان داریم که اختلاف نسبت سیگاری ها به غیر سیگاری های ناحیه ی northeast با بقیه نواحی در بازه ی بدست آمده قرار دارد.

روش2 : از هر متغیر یک سطح را در نظر میگیریم و اختلاف نسبت های آن ها را پیدا کرده و بازه ی اطمینان را تشکیل می دهیم . سطح yes از متغیر smoker و سطح northeast از متغیر region را انتخاب می کنیم و داریم:

```
ct1 <- table(insurance$smoker, insurance$region, dnn=c("insurance","region"))
ptable <- prop.table(ct1)
ptable
smokeyes <- ptable[2]+ptable[4]+ptable[6]+ptable[8]
northeast <- ptable[1]+ptable[2]
pe <- northeast - smokeyes

se <- sqrt((northeast*(1-northeast)/324)+(smokeyes*(smokeyes)/68))

low <- (pe)-1.96 * se
up <- (pe)+1.96 * se
ci <- c(low,up)
ci
```

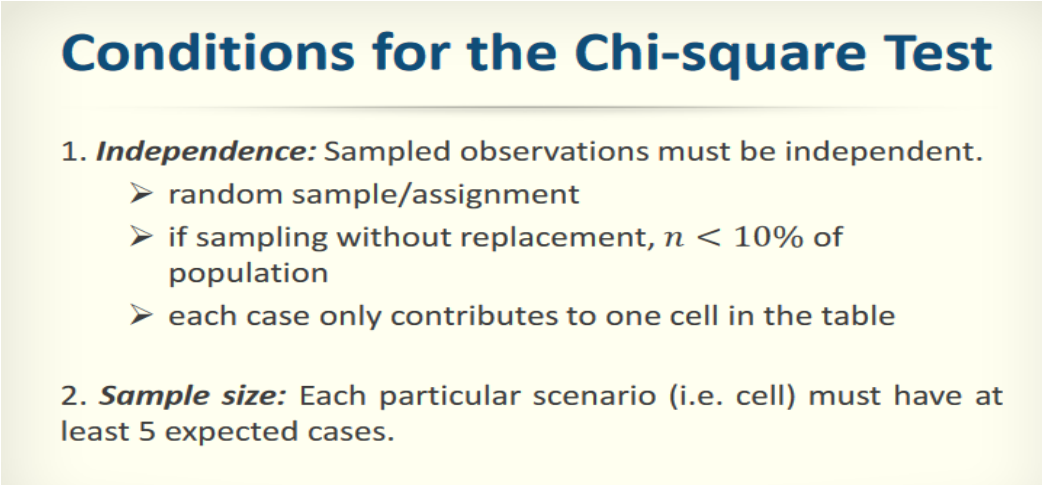
مقدار ci برابر است با :

```
> ci
[1] -0.03004772  0.10478614
>
```

تفسیر بازه اطمینان : 95 درصد اطمینان داریم اختلاف نسبت سیگاری ها به غیر سیگاری ها و نسبت northeast ها به بقیه در بازه ی بالا قرار دارد.

.b

در این قسمت خواسته شده با تعریف یک آزمون فرض بررسی کنیم آیا دو متغیر مستقل از هم اند یا خیر. از آزمون chi-square برای بررسی استقلال دو متغیرکتگوریکال smoker و region استفاده می کنیم. فرض صفر را مستقل بودن دو متغیر و فرض مقابل را وابسته بودن آن ها در نظر میگیریم و باید ابتدا شرایط تست را بررسی کنیم:



1)شرط استقلال مانند حالت قبل است و استقلال درون گروهی داریم ولی استقلال بین گروهی نداریم.

2) شرط اندازه نمونه برقرار است چون تعداد سیگاری ها و غیر سیگاری ها در هر ناحیه از 5 بزرگتر است. در قطعه کد زیر ابتدا نسبت سیگاری ها و غیرسیگاری ها را در هر ناحیه بدست اوردیم سپس جمعیت هر ناحیه را پیدا کردیم و این دو مقدار را برای هر ناحیه در هم ضرب کردیم و تعداد سیگاری های هر ناحیه پیدا شد که از 5 بزرگتر است و نتیجه را در ادامه اورده ایم:

```
ct1 <- table(insurance$smoker, insurance$region, dnn=c("insurance","region"))
ptable <- prop.table(ct1)
ptable
n_region <- table(insurance$region)
n_region

n_region[1]*ptable[2]
n_region[2]*ptable[4]
n_region[3]*ptable[6]
n_region[4]*ptable[8]
```

تعداد سیگاری های هر ناحیه :

```
> n_region[1]*ptable[2]
northeast
 16.22422
> n_region[2]*ptable[4]
northwest
 14.08819
> n_region[3]*ptable[6]
southeast
 24.75635
> n_region[4]*ptable[8]
southwest
 14.08819
>
```

با فرض برقراری شرایط ، تست را اعمال می کنیم :

```
> q1_b_table <- table(insurance$region, insurance$smoker)
> chisq.test(q1_b_table)

        Pearson's Chi-squared test

data:  q1_b_table
X-squared = 7.3435, df = 3, p-value = 0.06172

>
```

مقدار p-value برابر است با 0.06 که اگر سطح معناداری را 0.05 در نظر بگیریم نمی توانیم فرض صفر را رد کنیم و دو متغیر مستقل از هم اند.

QUESTION2

در این سوال خواسته شده که یک متغیر کتگوریکال باینری را انتخاب کنیم. در این دیتاست دو متغیر با این ویژگی ها داریم : sex , smoker که متغیر smoker را انتخاب می کنیم. یک نمونه به سایز 14 بدون جایگذاری از این متغیر برمیداریم. تعداد "yes" ها را در آن می شماریم و از بین این 14 داده 4 تای آن ها "yes" هستند.

```
> sample_smokers <- sample(insurance$smoker, size=14, replace = FALSE)
> count_smokers <- sum(c(sample_smokers == 'yes'))
> count_smokers
[1] 4
>
```

فرض صفر آزمون را به این شکل تعریف می کنیم که میانگین تعداد "yes" ها برابر است با np و فرض مقابل اینست که این میانگین برابر np نباشد.

$H0: \mu = np = 14 * 0.5 = 7$

$Ha: \mu \neq 7$

در روش شبیه سازی می خواهیم در نهایت به p-value برسیم که این مقدار برابر است با احتمال مشاهده ای که در نمونه داشتیم و یا extremتر از آن به شرط صحیح بودن فرض صفر.

شرایط تست را بررسی می کنیم :

1)استقلال درون گروهی برقرار است زیرا نمونه برداری برای هر 14 سمپل بدون جایگذاری انجام شده است .

استقلال بین گروهی برقرار نیست زیرا 2000 نمونه را با جایگذاری برداشتیم.(تعداد کل داده ها 1338تاست)

2) اندازه نمونه: این شرط هم برقرار نیست چون : $np<10$, $n(1-p)<10$

با روش شبیه سازی و تکرار نمونه برداری بررسی می کنیم ببینیم احتمال آنکه در هربار نمونه برداری تعداد کمتر از 4 تا یا بیشتر از 10تا "yes" ببینیم چقدر است زیرا تست دو طرفه است.

2000 بار نمونه برداری انجام میدهیم:

و در این 2000 نمونه برداری می بینیم که مقدار p-value برابر است با 0.0105 بنابراین فرض صفر رد می شود.

```r
sample_smokers <- sample(insurance$smoker, size=14, replace = FALSE)
count_smokers <- sum(c(sample_smokers == 'yes'))
count_smokers
success_per_iter = c()
for(i in cbind(1:2000)){
  yes <- sum(sample(0:1,14,replace = TRUE))
  success_per_iter[i] <- yes
}
total_success <- sum(c(success_per_iter > 14-count_smokers)+c(success_per_iter < count_smokers))
success_rate <- total_success/2000
success_rate
```

```
> success_rate <- total_success/2000
> success_rate
[1] 0.0105
>
```

QUESTION3

a.

تنها متغیر کتگوریکال این دیتاست که بیشتر از دو سطح دارد، region است که توضیح احتمال آن را در زیر محاسبه کردیم :

```
> n_region <- table(insurance$region)/1338
> n_region

northeast northwest southeast southwest
0.2421525 0.2428999 0.2720478 0.2428999
>
```

دو نمونه سمپل 100 تایی از دیتاست برمیداریم یکی بدون بایاس و به دیگری بایاس اضافه می کنیم و هردو را با استفاده از تست goodness of fit با توزیع اصلی که در بالا حساب کردیم ، مقایسه می کنیم:

ابتدا آزمون فرض را تعریف می کنیم : فرض صفر اینست که دو توزیع برابرند و اگر اختلافی ببینیم تصادفی است، فرض مقابل اینست که بین دو توزیع تفاوت معناداری وجود دارد و دو توزیع یکسان نیستند و بایاس وجود دارد.

شرایط تست را بررسی می کنیم :

1) شرط استقلال : نمونه برداری تصادفی و بدون جایگذاری انجام شده است بنابراین استقلال برقرار است.

2) شرط اندازه نمونه برقرار است و در هر خانه حداقل 5 داده را داریم.

حالت اول : سمپل بدون بایاس

```
> sample_unbiased <- insurance[sample(nrow(insurance), 100), ]
> prop_dist1 <- table(sample_unbiased$region)
> prop_dist1

northeast northwest southeast southwest
       26        28        23        23
> chisq.test(prop_dist1,n_region)

        Pearson's Chi-squared test

data:  prop_dist1 and n_region
X-squared = 5, df = 4, p-value = 0.2873
```

در این حالت می بینیم که مقدار p-value از 0.05 بیشتر شده بنابراین دو توزیع یکسان هستند و فرض صفر را نمیتوان رد کرد.

حالت دوم : سمپل بایاس دار

```
> filter_by_sex <- filter(select(insurance,sex,region), insurance$sex == 'female')
> sample_biased <- filter_by_sex[sample(nrow(filter_by_sex), 100), ]
> prop_dist2 <- table(sample_biased$region)
> prop_dist2

northeast northwest southeast southwest
       23        24        32        21
> chisq.test(prop_dist2,n_region)

        Pearson's Chi-squared test

data:  prop_dist2 and n_region
X-squared = 8, df = 6, p-value = 0.2381
```

در این حالت سطرهایی از جدول را انتخاب کردیم که جنسیت آن ها مونث بوده و به این شکل بایاس ایجاد کردیم و میبینیم که مقدار X-squared از حالت قبل بیشتر شده و خطای بیشتری داریم وp-value هم از 0.05 بزرگتر است پس فرض صفر رد نمی شود و دو توزیع یکسان هستند.

.b

برای این قسمت متغیر کتگوریکال sex را انتخاب می کنیم و با region مقایسه می کنیم.

با تعریف یک آزمون فرض بررسی می کنیم آیا دو متغیر مستقل از هم اند یا خیر. از آزمون chi-square برای بررسی استقلال دو متغیرکتگوریکال sex و region استفاده می کنیم. فرض صفر را مستقل بودن دو متغیر و فرض مقابل را وابسته بودن آن ها در نظر میگیریم و باید ابتدا شرایط تست را بررسی کنیم:

1) شرط استقلال : نمونه ها همان داده های اصلی هستند و فرض می کنیم که تصادفی جمع آوری شده اند و از 10 درصد جمعیت کل کمتر اند و هر مورد فقط در یکی از خانه های جدول قرار دارد (برای هر متغیر داده های جدول هم پوشانی ندارند ولی داده های دو متغیر یکسان هستند و از این نظر مستقل از هم نیستند.)

2) شرط اندازه نمونه هم برقرار است و تعداد هر مورد در هر خانه ی جدول از 5 بزرگتر است.

```
  Chi-squared approximation may be incorrect
> q3_b_table <- table(insurance$region, insurance$sex)
> chisq.test(q3_b_table)

        Pearson's Chi-squared test

data:  q3_b_table
X-squared = 0.43514, df = 3, p-value = 0.9329
```

مقدار p-vale از 0.05 بزرگتر است بنابراین نمی توان فرض صفر را رد کرد و دو متغیر از هم مستقل اند.

QUESTION 4

متغیر charges را بعنوان متغیر پاسخ در نظر میگیریم و در فاز اول پروژه دیدیم که بیشترین همبستگی را با متغیر age دارد بنابراین age را بعنوان explanatory در نظر میگیریم.

.a

```
> model_4_a <- lm(insurance$charges ~ insurance$age)
> summary(model_4_a)

Call:
lm(formula = insurance$charges ~ insurance$age)

Residuals:
   Min     1Q Median     3Q    Max
 -8059  -6671  -5939   5440  47829

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    3165.9      937.1   3.378 0.000751 ***
insurance$age   257.7       22.5  11.453  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11560 on 1336 degrees of freedom
Multiple R-squared:  0.08941,   Adjusted R-squared:  0.08872
F-statistic: 131.2 on 1 and 1336 DF,  p-value: < 2.2e-16
```

.b ضرایب شیب و عرض از مبدا را می توانیم با صدا کردن summary روی مدل بدست آوریم که در قسمت قبل اینکار را انجام دادیم و از آن ها برای نوشتن رابطه رگرسیون خطی استفاده می کنیم :

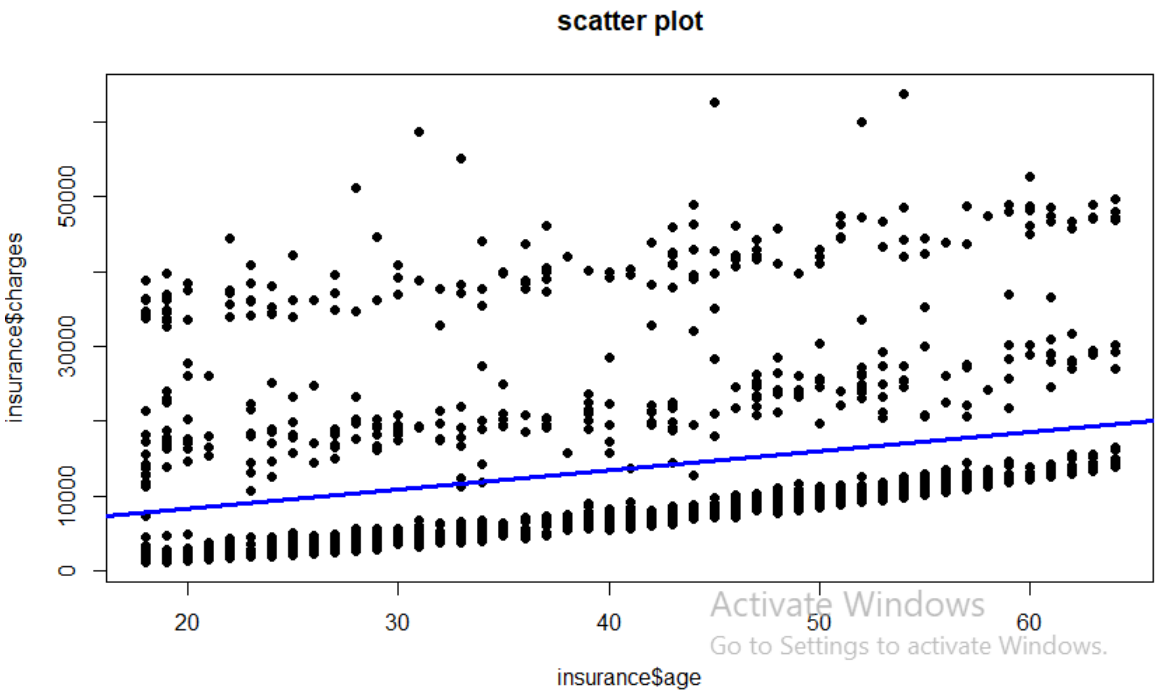
Charges = 3165.9 + 257.7 * age

تفسیر شیب : به ازای هر یک سالی که سن افراد زیاد میشود انتظار داریم که 257.7 واحد به charges آن ها اضافه شود.

تفسیر عرض از مبدا : این تفسیر در این مثال بی معناست چون سن نمی تواند 0 باشد.

.c کد این قسمت را در زیر می بینیم :

```
plot(x=insurance$age, y=insurance$charges, main = "scatter plot",
     xlab = "insurance$age", ylab = "insurance$charges",
     pch = 19, frame = TRUE)
abline(model_4_a, col = "blue", lwd=3)
```



.d

.1

در این قسمت خواسته شده تا یک نمونه 27 تایی از دیتاست اصلی برداریم سپس برای این نمونه با در نظر گرفتن همان متغیرهای response و explanatory قسمت های قبل ، یک مدل رگرسیون خطی تعریف کنیم که کد این قسمت را در زیر میبینیم :

```
sample_4_d <- insurance[sample(nrow(insurance), 27), ]
model_4_d1 <- lm(charges ~ age , data = sample_4_d)
summary(model_4_d1)
```

در ادامه سوال خواسته تا ببینیم متغیر explanatory ما که همان سن است ، predictor خوبی برای متغیر پاسخ هست یا خیر. یک آزمون فرض روی شیب خط رگرسیون تعریف می کنیم :

H0: $\beta = 0$

Ha: $\beta \neq 0$

Summary را روی مدل صدا میزنیم تا مقدار تخمین گر نقطه ای برای شیب خط(b1) و SE_{b1} را بدست آوریم:

```
Call:
lm(formula = charges ~ age, data = sample_4_d)

Residuals:
   Min     1Q Median     3Q    Max
 -5576  -4428  -4089  -2978  33281

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    302.0     6162.3   0.049   0.9613
age            272.2      135.3   2.011   0.0552 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9851 on 25 degrees of freedom
Multiple R-squared:  0.1393,    Adjusted R-squared:  0.1049
F-statistic: 4.046 on 1 and 25 DF,  p-value: 0.05519
```

مقدار p-value در جدول بالا آمده که برابر است با 0.05519 و از 0.05 بزرگتر است بنابراین فرض صفر رد نمی شود و در نتیجه در این سطح معناداری متغیر age پیش بینی کننده ی خوبی برای متغیر پاسخ charges نیست.

همچنین مقدار p-value را می توان با محاسبه t-statistic بدست آورد:

$T = (\text{point estimate} - \text{null value}) / SE_{b1}$

$T = (272.2 - 0) / 135.3 = 2.011$

```
> 2*pt(-abs(2.011), df=25)
[1] 0.05522643
>
```

میبینیم که مقدار p-value محاسبه شده تقریبا همان مقدار درج شده در جدول است .

.2

بازه اطمینان 95 درصد برای شیب خط رگرسیون مدل حالت قبل را باید محاسبه کنیم:

$Ci = \text{point estimate} \pm \text{margin of error}$

$Ci = b1 \pm t^* SE_{b1}$

```
> tstar <- abs(qt(0.025,df=25))
> pointEstimate = 272.2
> SE_b1 = 135.3
> lower <- pointEstimate - tstar*SE_b1
> upper <- pointEstimate + tstar*SE_b1
> ci <- c(lower,upper)
> ci
[1]  -6.455566 550.855566
>
```

تفسیر بازه اطمینان : 95 درصد اطمینان داریم به ازای هر یکسال افزایش سن افراد انتظار داریم مقدار charges آن فرد در بازه (-6.45,55.85) تغییر کند.

.3

برای این قسمت متغیر children را انتخاب می کنیم که بعد از age و bmi بیشترین همبستگی را با charges دارد منتها متغیر bmi با age همبستگی نسبتا زیادی دارد و بدرد نمی خورد. با همان سمپل 27تایی که قسمت اول گرفتیم برای متغیر charges , children یک مدل رگرسیون خطی تعریف می کنیم:

```
model_4_d3 <- lm(charges ~ children+age ,data = sample_4_d)
summary(model_4_d3)
```

نتیجه بصورت زیر است :

```
Call:
lm(formula = charges ~ children + age, data = sample_4_d)

Residuals:
     Min       1Q   Median       3Q      Max
-12724.1  -3548.9  -1206.9    -34.5  28790.0

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6226.6     6196.3  -1.005   0.3250
children      3849.2     1550.8   2.482   0.0205 *
age            337.3      126.0   2.678   0.0132 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8968 on 24 degrees of freedom
Multiple R-squared:  0.3151,    Adjusted R-squared:  0.258
F-statistic: 5.521 on 2 and 24 DF,  p-value: 0.01065
```

مقایسه دو مدل بر مبنای adjusted R squared : در مدل اول این مقدار برابر است با 0.1049 و در مدل دوم برابر است با 0.258 که نشان می دهد متغیر دومی که به مدل اضافه کردیم predictor خوبی برای متغیر پاسخ ماست.

مقایسه دو مدل بر مبنای ANOVA table : مقدار R^2 را از جدول برای دو مدل محاسبه میکنیم می خواهیم ببینیم مدل ما چقدر خوب است و چقدر از تغییرات متغیر پاسخ توسط متغیرهای explanatory توضیح داده می شود:

R^2 = explained var / total var

خروجی مدل اول:

```
> anova(model_4_d1)
Analysis of Variance Table

Response: charges
          Df     Sum Sq   Mean Sq F value  Pr(>F)
age        1  392561819 392561819  4.0455 0.05519 .
Residuals 25 2425909126  97036365
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

R^2 = 392561819 /(392561819+ 2425909126) = 0.13

خروجی مدل دوم:

```
> anova(model_4_d3)
Analysis of Variance Table

Response: charges
           Df     Sum Sq   Mean Sq F value  Pr(>F)
children    1  311359417 311359417  3.8711 0.06079 .
age         1  576743386 576743386  7.1706 0.01316 *
Residuals 24 1930368141  80432006
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

R^2 = (311359417+576743386) / (311359417+576743386 +1930368141) = 0.31

همانطور که میبینیم مقادیر R^2 در هر مدل از مقادیر R^2_{adj} آن مدل بیشتر است . می دانیم که هرچقدر R^2 بزرگتر باشد رابطه ی خطی بین متغیر پاسخ و explanatory قوی تر خواهد بود ولی با اضافه کردن هر متغیری این مقدار افزایش می یابد حتی اگه خیلی متغیر خوبی را اضافه نکرده باشیم و ممکن است باعث overfit شدن مدل شود. برای همین از معیار R^2_{adj} برای بررسی مدل استفاده می کنیم که به ازای هر predictor که اضافه می کنیم جریمه ای در نظر میگیرد و همیشه مقدار آن از R^2 کمتر است و یک متغیر زمانی خوب است که اضافه شدن آن به مدل مقدار R^2_{adj} را افزایش دهد.

در این دو مدل با اضافه کردن متغیر دوم مقدار R^2_{adj} افزایش یافته و مدل دوم از مدل اول بهتر است اما باز هم مقادیر R^2 و R^2_{adj} کوچک هستند.

QUESTION5

متغیر پاسخ سوال قبل یعنی charges را در نظر میگیریم:

a. بهترین مدل رگرسیون خطی را از دو روش backward elimination و forward selection بدست می آوریم :

- **روش backward elimination** : در این روش مدلی با تمام متغیرهای دیتاست تعریف می کنیم و متغیری که p-value آن از 0.05 بزرگتر بوده و از p-value بقیه متغیرها هم بیشتر است از مدل حذف می کنیم سپس دوباره مدل را بدون آن متغیر می سازیم و همینکار را با مدل جدید تکرار می کنیم و این روش را ادامه می دهیم تا جاییکه تمام مقادیر p-value از 0.05 کمتر شوند. کد این قسمت را در زیر میبینیم:

```
model_5_a_1 <- lm(charges ~ age+sex+bmi+children+smoker+region, data = insurance)
summary(model_5_a_1)
model_5_a_2 <- lm(charges ~ age+bmi+children+smoker+region, data = insurance)
summary(model_5_a_2)
model_5_a_3 <- lm(charges ~ age+bmi+children+smoker, data = insurance)
summary(model_5_a_3)
```

نتیجه حاصل از تعریف مدل با تمام متغیرهای explanatory :

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -11938.5      987.8 -12.086  < 2e-16 ***
age                256.9       11.9  21.587  < 2e-16 ***
sexmale           -131.3      332.9  -0.394 0.693348
bmi                339.2       28.6  11.860  < 2e-16 ***
children           475.5      137.8   3.451 0.000577 ***
smokeryes        23848.5      413.1  57.723  < 2e-16 ***
regionnorthwest   -353.0      476.3  -0.741 0.458769
regionsoutheast  -1035.0      478.7  -2.162 0.030782 *
regionsouthwest   -960.0      477.9  -2.009 0.044765 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6062 on 1329 degrees of freedom
Multiple R-squared:  0.7509,    Adjusted R-squared:  0.7494
F-statistic: 500.8 on 8 and 1329 DF,  p-value: < 2.2e-16

>
```

در این مدل متغیرهای sex,region مقدار p-value بزرگتر از 0.05 دارند و مقدار آن برای متغیر sex بیشتر است پس آن را حذف می کنیم و مدلی بدون این متغیر می سازیم:

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -11990.27     978.76 -12.250  < 2e-16 ***
age                256.97      11.89  21.610  < 2e-16 ***
bmi                338.66      28.56  11.858  < 2e-16 ***
children           474.57     137.74   3.445 0.000588 ***
smokeryes        23836.30     411.86  57.875  < 2e-16 ***
regionnorthwest   -352.18     476.12  -0.740 0.459618
regionsoutheast  -1034.36     478.54  -2.162 0.030834 *
regionsouthwest   -959.37     477.78  -2.008 0.044846 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6060 on 1330 degrees of freedom
Multiple R-squared:  0.7509,    Adjusted R-squared:  0.7496
F-statistic: 572.7 on 7 and 1330 DF,  p-value: < 2.2e-16

> |
```

در مدل دوم متغیر region مقدار p-value بزرگتر از 0.05 دارد و حذف می شود و مدل جدید ساخته می شود :

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -12102.77     941.98 -12.848  < 2e-16 ***
age            257.85      11.90  21.675  < 2e-16 ***
bmi            321.85      27.38  11.756  < 2e-16 ***
children       473.50     137.79   3.436 0.000608 ***
smokeryes    23811.40     411.22  57.904  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6068 on 1333 degrees of freedom
Multiple R-squared:  0.7497,    Adjusted R-squared:  0.7489
F-statistic: 998.1 on 4 and 1333 DF,  p-value: < 2.2e-16

> |
```

با حذف متغیرهای region, sex تمام مقادیر p-value از 0.05 کمتر می شوند و این بهترین مدلی است که میتوانیم با این روش تعریف کنیم.

- **روش forward selection** : در این روش با مدلی تنها با یک متغیر explanatory شروع میکنیم و هر بار متغیری را به مدل اضافه می کنیم که مقدار adjusted R squared مدل بیشتر باشد و تاجایی ادامه می دهیم که با اضافه کردن متغیر جدید شاهد افزایش مقدار adjusted R squared نباشیم.

```
model1 <- lm(charges ~ age, data = insurance)
model2 <- lm(charges ~ sex, data = insurance)
model3 <- lm(charges ~ bmi, data = insurance)
model4 <- lm(charges ~ children, data = insurance)
model5 <- lm(charges ~ smoker, data = insurance)
model6 <- lm(charges ~ region, data = insurance)
summary(model1)$adj.r.squared
summary(model2)$adj.r.squared
summary(model3)$adj.r.squared
summary(model4)$adj.r.squared
summary(model5)$adj.r.squared
summary(model6)$adj.r.squared
```

مقادیر adjusted R squared عبارتند از :

```
> summary(model1)$adj.r.squared
[1] 0.08872432
> summary(model2)$adj.r.squared
[1] 0.002536334
> summary(model3)$adj.r.squared
[1] 0.03862008
> summary(model4)$adj.r.squared
[1] 0.003878717
> summary(model5)$adj.r.squared
[1] 0.6194802
> summary(model6)$adj.r.squared
[1] 0.00440006
```

بیشترین مقدار را مدلی با متغیر smoker دارد پس آن را در مدل نگه می داریم و متغیرهای دیگر را اضافه می کنیم :

```
model21 <- lm(charges ~ smoker + age, data = insurance)
model22<- lm(charges ~ smoker + sex, data = insurance)
model23 <- lm(charges ~ smoker + bmi, data = insurance)
model24 <- lm(charges ~ smoker + children, data = insurance)
model25 <- lm(charges ~ smoker + region, data = insurance)
summary(model21)$adj.r.squared
summary(model22)$adj.r.squared
summary(model23)$adj.r.squared
summary(model24)$adj.r.squared
summary(model25)$adj.r.squared
```

مقادیر بصورت زیر است :

```
> model25 <- lm(charges ~ smoker +
> summary(model21)$adj.r.squared
[1] 0.7209834
> summary(model22)$adj.r.squared
[1] 0.6192024
> summary(model23)$adj.r.squared
[1] 0.6574295
> summary(model24)$adj.r.squared
[1] 0.6230399
> summary(model25)$adj.r.squared
[1] 0.6191738
>
```

متغیر age هم باعث افزایش R^2_{adj} می شود پس در مدل نگه می داریم :

```
model31 <- lm(charges ~ smoker + age + sex, data = insurance)
model32 <- lm(charges ~ smoker + age + bmi, data = insurance)
model33 <- lm(charges ~ smoker + age + children, data = insurance)
model34 <- lm(charges ~ smoker + age + region, data = insurance)
summary(model31)$adj.r.squared
summary(model32)$adj.r.squared
summary(model33)$adj.r.squared
summary(model34)$adj.r.squared
```

نتایج این مرحله :

```
> summary(model31)$adj.r.squared
[1] 0.7207857
> summary(model32)$adj.r.squared
[1] 0.7469093
> summary(model33)$adj.r.squared
[1] 0.723122
> summary(model34)$adj.r.squared
[1] 0.7210637
```

متغیر bmi هم به مدل اضافه میشود و مرحله ی بعد را با ثابت نگه داشتن متغیرهای smoker,age,bmi انجام می دهیم :

```
model41 <- lm(charges ~ smoker + age + bmi + sex, data = insurance)
model42 <- lm(charges ~ smoker + age + bmi + children, data = insurance)
model43 <- lm(charges ~ smoker + age + bmi + region, data = insurance)
summary(model41)$adj.r.squared
summary(model42)$adj.r.squared
summary(model43)$adj.r.squared
```

نتایج این مرحله :

```
> summary(model41)$adj.r.squared
[1] 0.7467396
> summary(model42)$adj.r.squared
[1] 0.7489434
> summary(model43)$adj.r.squared
[1] 0.7475274
>
```

متغیر children را هم به مدل اضافه می کنیم :

```
model51 <- lm(charges ~ smoker + age + bmi + children + sex, data = insurance)
model52 <- lm(charges ~ smoker + age + bmi + children + region, data = insurance)
summary(model51)$adj.r.squared
summary(model52)$adj.r.squared
```

نتایج این مرحله :

```
> summary(model51)$adj.r.squared
[1] 0.748783
> summary(model52)$adj.r.squared
[1] 0.7495727
>
```

متغیر region هم باعث افزایش مقدار R^2_{adj} می شود بنابراین یک مرحله دیگر ادامه میدهیم و متغیر region را هم در مدل ثابت نگه می داریم و متغیر sex را اضافه میکنیم :

```
> model61 <- lm(charges ~ smoker + age + bmi + children + region + sex, data = insurance)
> summary(model61)$adj.r.squared
[1] 0.7494136
>
```

می بینیم که با اضافه کردن متغیر sex مقدار R^2_{adj} نسبت به حالت قبل کمتر می شود بنابراین بهترین حالت زمانیست که این متغیر را در مدل نداریم .

.b

```
library(caret)
data_ctrl <- trainControl(method = "cv", number = 5)
model_caret <- train(charges ~ age + bmi + children + smoker + sex + region, data = insurance,
                     trControl = data_ctrl, method = "lm",na.action = na.pass)

model_caret
model_caret$finalModel
```

خروجی بصورت زیر است :

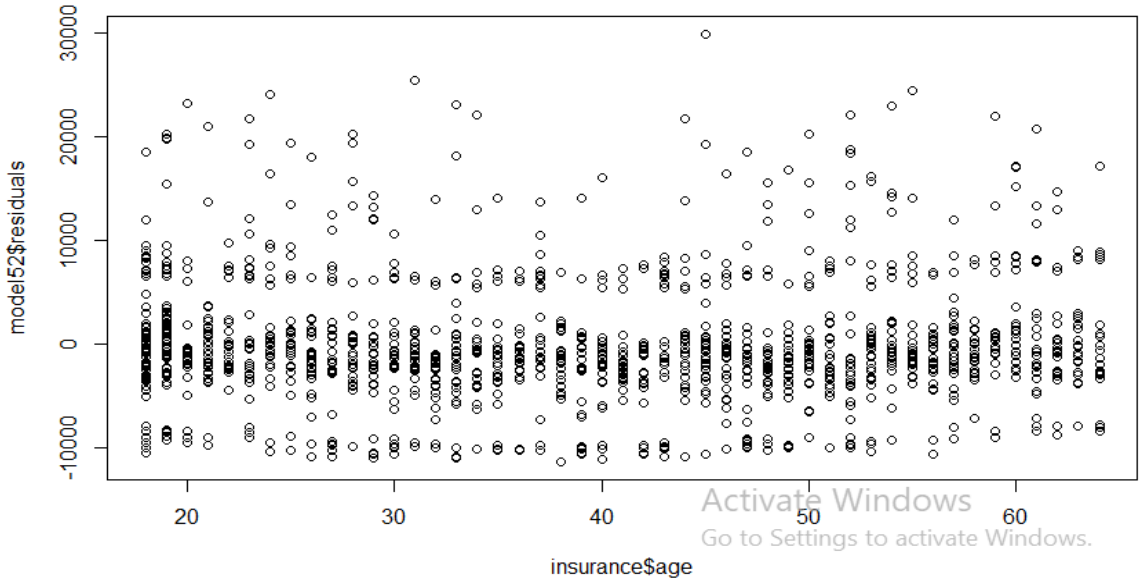
```
> model_caret
Linear Regression

1338 samples
   6 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 1070, 1071, 1070, 1071, 1070
Resampling results:

  RMSE      Rsquared   MAE
  6066.375  0.7500223  4191.234
```

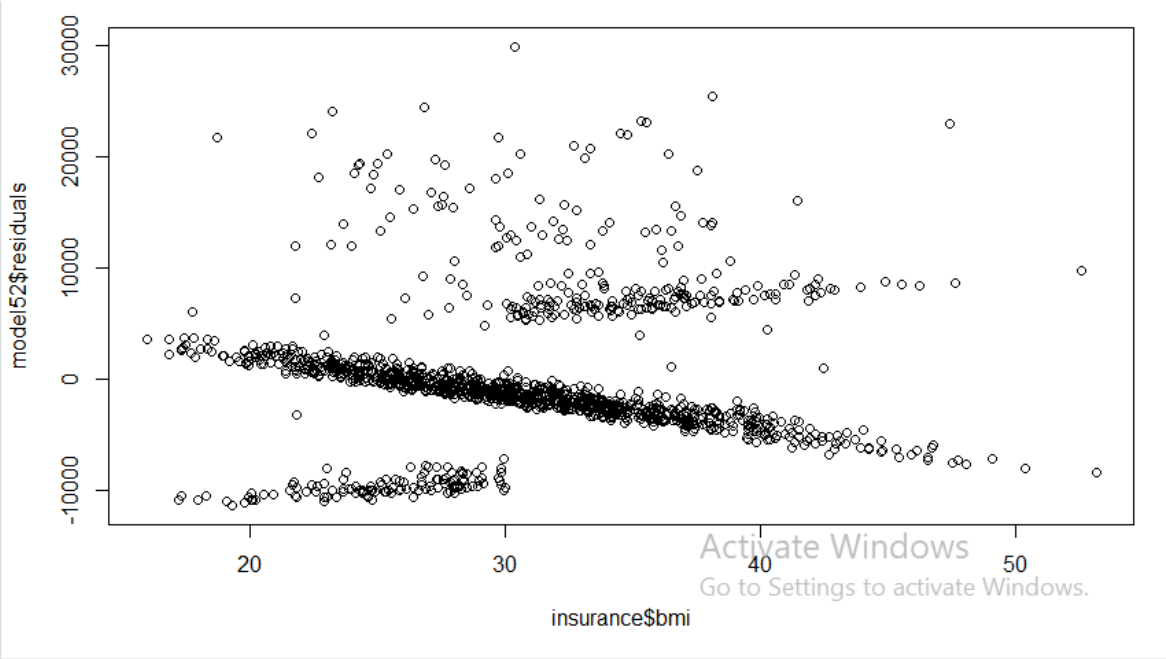
مقدار RMSE برابر است با 6066.375

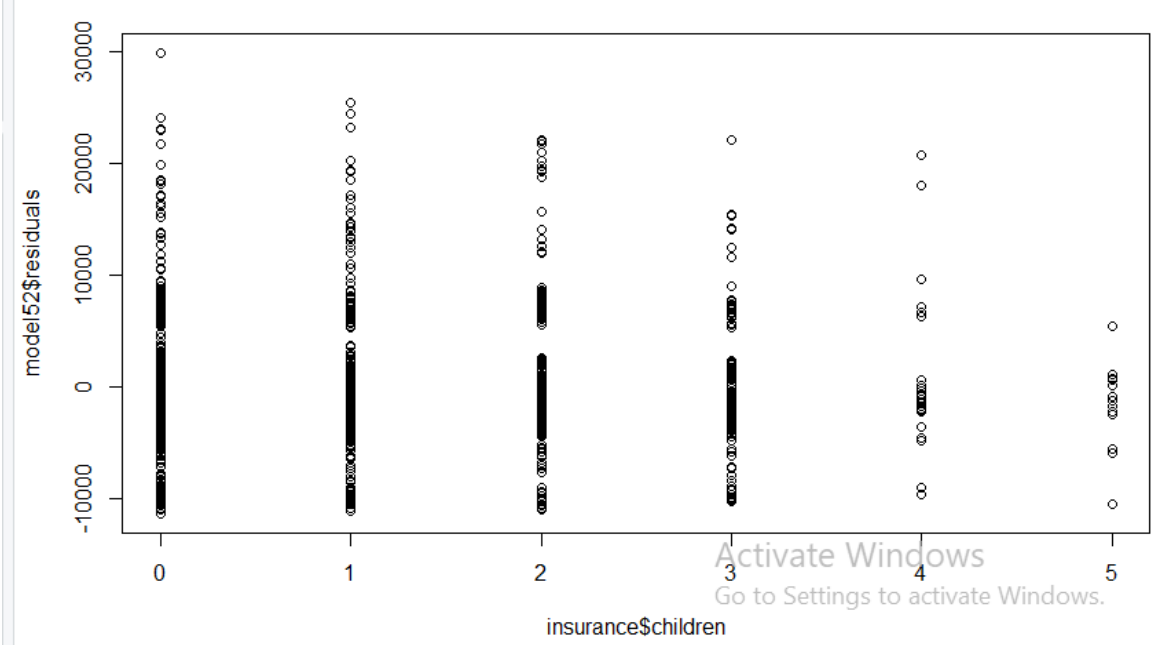
تفسیر RMSE : میانگین تفاضل بین مقدار خروجی مشاهده شده و خروجی پیش بینی شده توسط مدل است که هرچقدر مقدار آن کمتر باشد مدل بهتری داریم. و در این مدل با 6 پیش بینی کننده این مقدار زیاد است و مدل خوبی نداریم.

c. سه شرط رگرسیون خطی را برای مدل بررسی می کنیم :

- Linearity : باید ببینیم رابطه ی بین متغیرهای explanatory با متغیر response خطی است یا خیر. این را می توانیم با scatter plot یا residuals plot نشان دهیم :

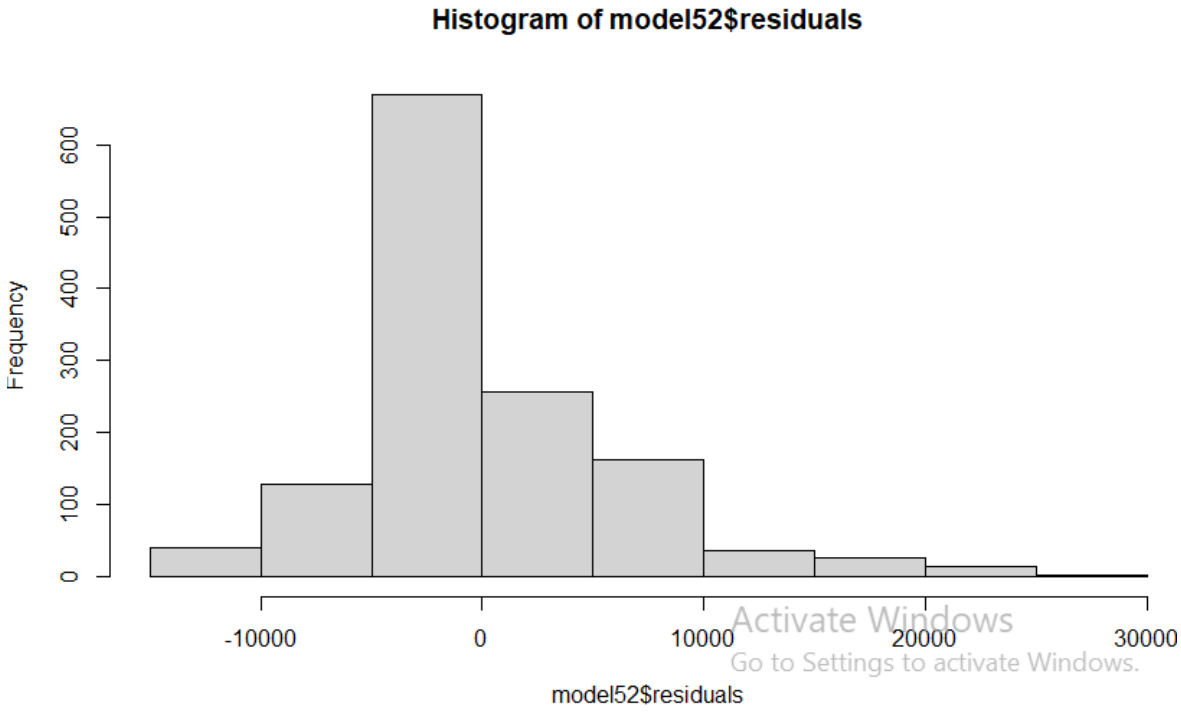


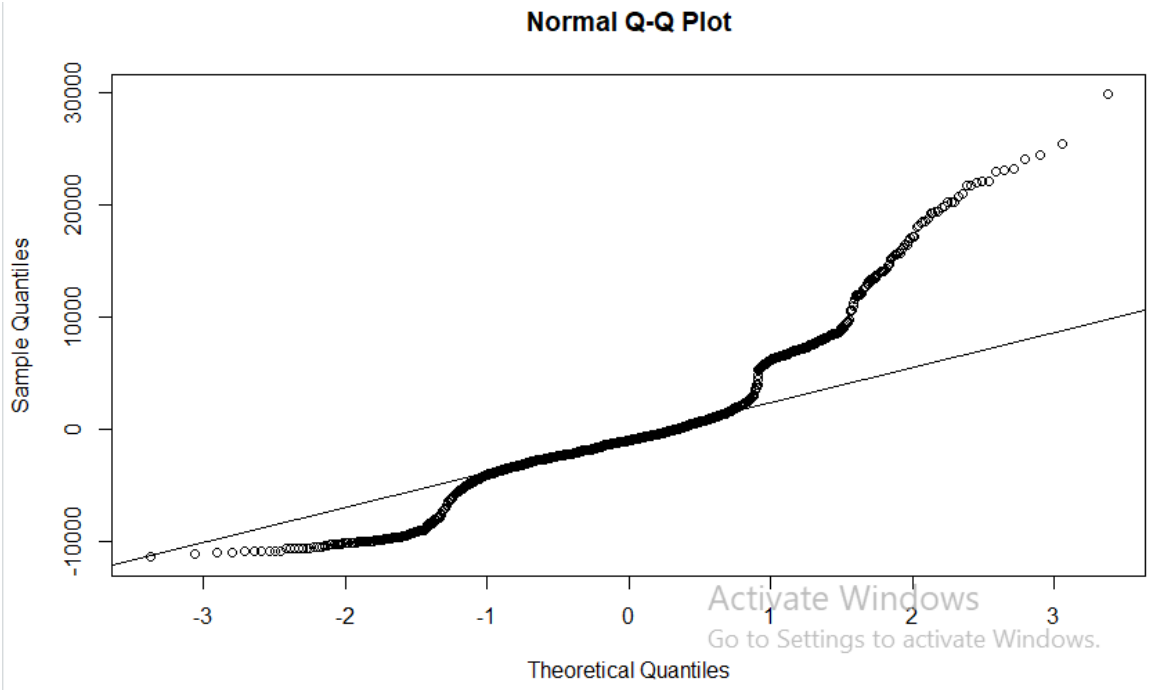




برای اینکه این شرط برقرار باشد باید داده ها اطراف y=0 پراکنده باشند ولی میبینیم که اینگونه نیست پس این شرط برقرار نیست.

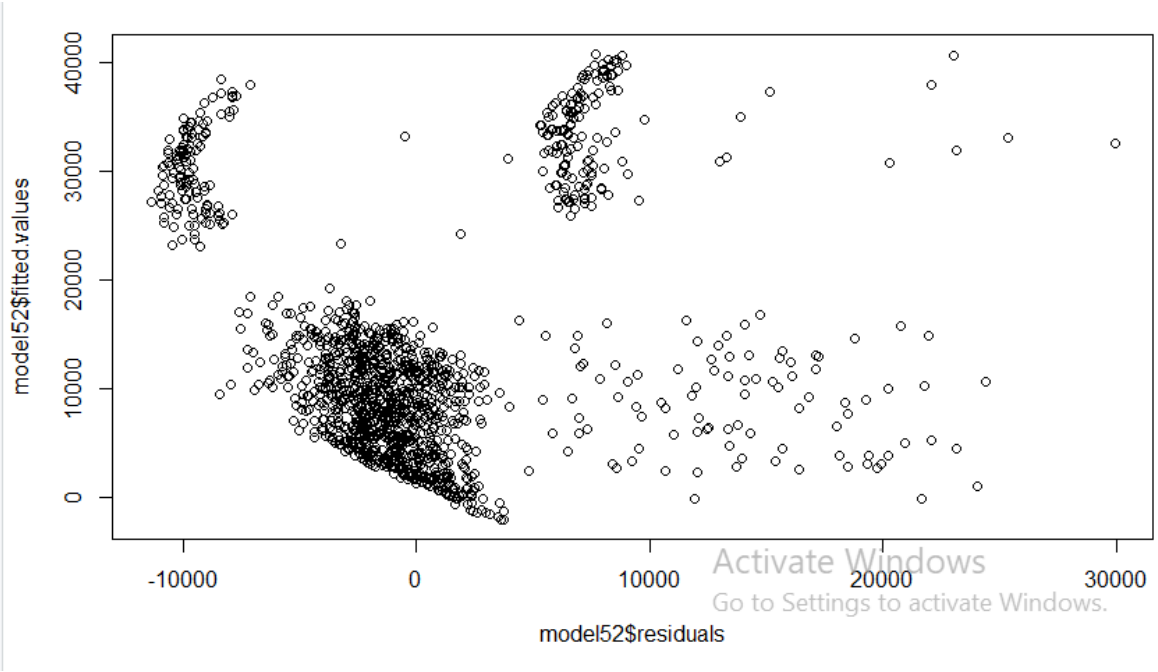
- Nearly normal residuals : باید توزیع residuals نزدیک نرمال باشد که میتوانیم هیستوگرام یا qqplot را برای آن رسم کنیم و این موضوع را بررسی کنیم :





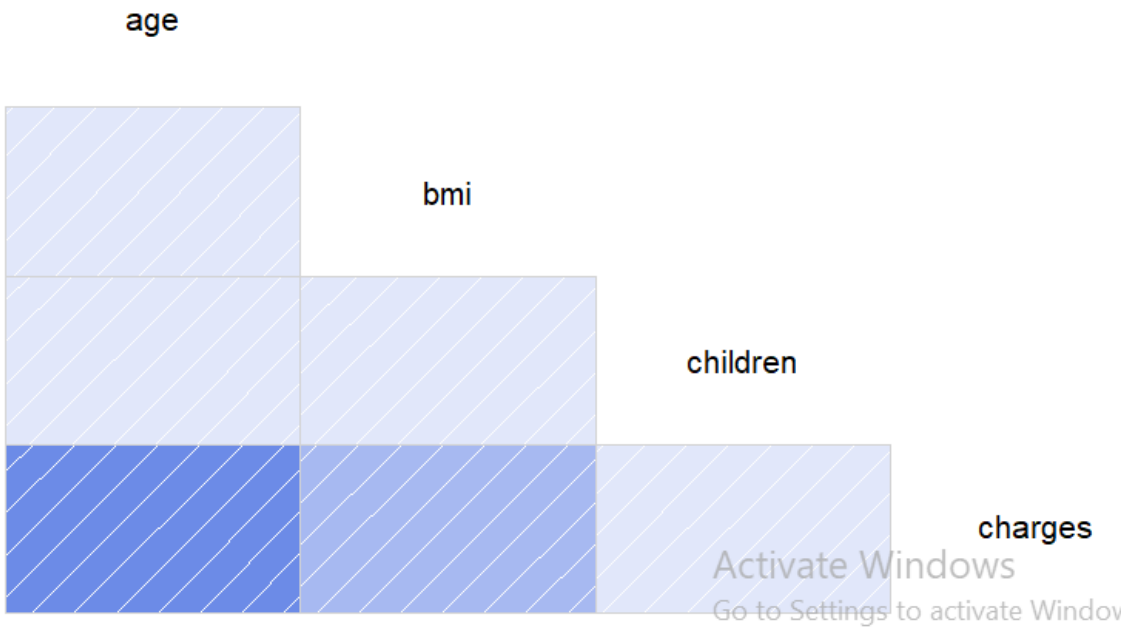
هیستوگرام residuals نرمال نیست و qqplot آن هم خطی نیست و نشان می دهد توزیعی نزدیک نرمال ندارد بنابراین این شرط برقرار نیست.

- Constant variability : تغییرات ثابت برای residuals باید داشته باشیم برای چک کردن این شرط باید نمودار residuals و predicted را رسم کنیم و در شکل زیر میبینیم که این تغییرات ثابت نیست و شرط برقرار نیست.



هیچکدام از شرایط برقرار نیست بنابراین بهترین مدلی که در قسمت اول این سوال پیدا کردیم ، مدل قابل اعتمادی نیست و باید موارد دیگری را برای انتخاب مدل در نظر بگیریم.

d. نمودار correlogram را در زیر آورده ایم همانطور که پیداست متغیر age و charges بیشترین همبستگی را دارند و بعد از آن age و bmi بیشترین میزان همبستگی را با یکدیگر دارند بنابراین بهتر است این دو متغیر با هم بعنوان explanatory انتخاب نشوند چون چیزی به مدل اضافه نمی کنند.



e. درصد تغییرات متغیر پاسخ که توسط مدل توضیح داده می شود همان مقدار R squared است که مقدارش برابر است با :

```
> model52 <- lm(charges ~ smoker + age + bmi + children + region, data = insurance)
> summary(model52)$r.squared
[1] 0.7508839
>
```

f. از آن جاییکه شرایط رگرسیون برای مدل ما برقرار نبود و همچنین مقدار RMSE زیاد شد، بنابراین این بررسی را با فرض برقرار بودن شرایط انجام می دهیم . مقدار adjusted R squared می توانند معیاری برای ما باشد که مدلی که تعریف کردیم خوب بوده یا نه ولی زمانیکه شرایط رگرسیون برقرار نیست این مقدار مفهومی نخواهد داشت. این مقدار برای مدل ما برابر است با 0.7495 که برای مدلی که در آن شرایط برقرار است مقدار خوبی است.

```
> summary(model52)$r.squared
[1] 0.7508839
> summary(model52)$adj.r.squared
[1] 0.7495727
>
```

QUESTION6

متغیر smoker را که کتگوریکال باینری است ، بعنوان متغیر پاسخ و متغیرهایage و children و charges را بعنوان متغیرهای explanatory در نظر می گیریم .

a. مدل را بصورت زیر تعریف می کنیم :

```
model_6_a <- glm(as.factor(smoker) ~ charges+age+children,data = insurance,family = binomial(link = "logit"))
summary(model_6_a)
```

نتایج در زیر گزارش شده است :

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.789e+00  4.081e-01  -6.834 8.26e-12 ***
charges      2.966e-04  1.938e-05  15.304  < 2e-16 ***
age         -8.626e-02  1.063e-02  -8.118 4.72e-16 ***
children    -1.601e-01  1.081e-01  -1.481    0.139
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1356.63  on 1337  degrees of freedom
Residual deviance:  418.08  on 1334  degrees of freedom
AIC: 426.08

Number of Fisher Scoring iterations: 7
```

b. تفسیر عرض از مبدا : برای فردی که هر سه متغیر children و age و charges صفر است ، لگاریتم شانس سیگاری بودن چقدر است. این تفسیر بی معناست چون سن نمیتواند صفر باشد.

تفسیر شیب برای متغیر age : به ازای هر یکسال افزایش سن فرد ، با فرض ثابت بودن مقدارpredictor های دیگر، انتظار داریم لگاریتم شانس سیگاری بودن فرد 8.626e-02 واحد کاهش یابد.

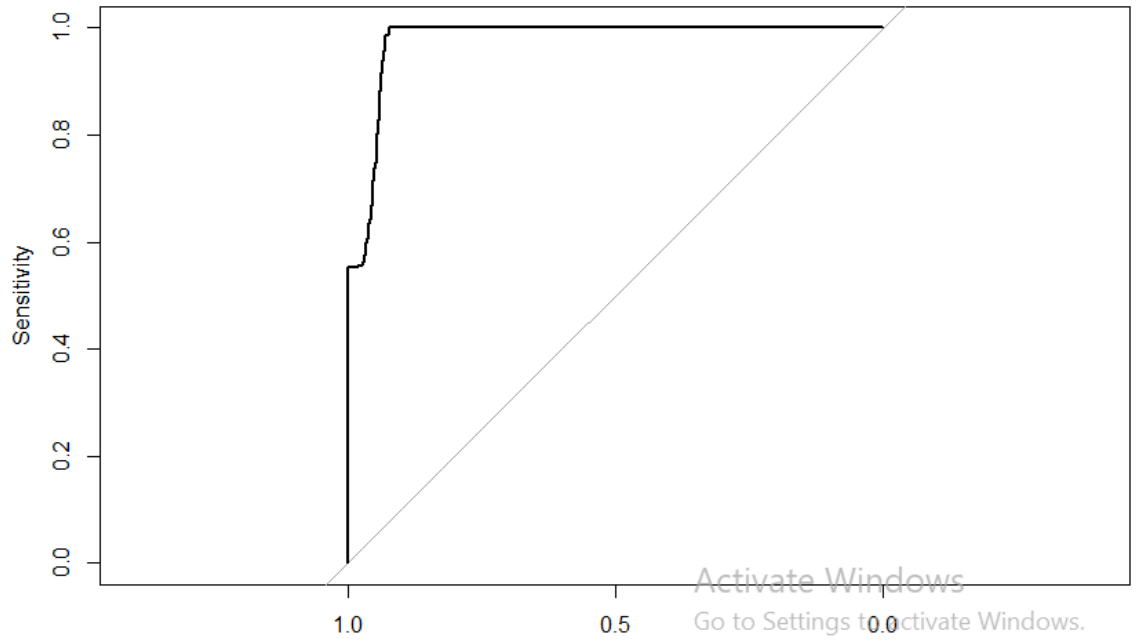
تفسیر شیب برای متغیر children : به ازای هر یک فرزند بیشتر ، با فرض ثابت بودن مقدارpredictor های دیگر، انتظار داریم لگاریتم شانس سیگاری بودن فرد 1.601e-01 واحد کاهش یابد.

تفسیر شیب برای متغیر charges : به ازای هر یک واحد افزایش charges ، با فرض ثابت بودن مقدارpredictor های دیگر، انتظار داریم لگاریتم شانس سیگاری بودن 2.966e-04 واحد افزایش یابد.

c. منحنی roc را با قطعه کد زیر می توان رسم کرد و با دستور آخر هم auc محاسبه می شود:

```
filter_dataa <- select(insurance,age,children,smoker,charges)

library(pROC)

prob <- predict(model_6_a,type=c("response"))
ROC <- roc(filter_dataa$smoker ~ prob, data = filter_dataa)
plot(ROC)
auc(ROC)
```

نتایج بصورت زیر است :



مقدار auc که همان سطح زیر منحنی roc است ، برابر است با :

```
> plot(ROC)
> auc(ROC)
Area under the curve: 0.9761
>
```

d. بازه اطمینان 98 درصد را برای سه متغیر explanatory بصورت زیر محاسبه می کنیم.

98% CI for log odds ratio = PE ± CV*SE

مقدار CV برای بازه 98 براب راست با 2.33 و مقادیر PE, SE را از جدول برمیداریم . کد این قسمت و نتیجه را در شکل زیر می توان دید.

```
> zstar <- abs(qnorm(0.01))
> PE <- c(2.966e-04,-8.626e-02,-1.601e-01)
> SE <- c( 1.938e-05, 1.063e-02,1.081e-01)
> lower <- PE - zstar*SE
> upper <- PE + zstar*SE
> log_ci_charges <- c(lower[1],upper[1])
> log_ci_age <-  c(lower[2],upper[2])
> log_ci_children <-  c(lower[3],upper[3])
> exp(log_ci_charges)
[1] 1.000252 1.000342
> exp(log_ci_age)
[1] 0.8949485 0.9403239
> exp(log_ci_children)
[1] 0.6626037 1.0956833
>
```

همچنین برای محاسبه سریع تر بازه اطمینان می توان از دستور زیر استفاده کرد که میبینیم نتایج یکی است:

```
> exp(cbind("Odds ratio" = coef(model_6_a), confint.default(model_6_a, level = 0.98)))
            Odds ratio       1 %      99 %
(Intercept) 0.06148426 0.0237929 0.1588841
charges     1.00029668 1.0002516 1.0003418
age         0.91735481 0.8949572 0.9403129
children    0.85202878 0.6625468 1.0957008
```

QUESTION7

a. تمام explanatory هایی که مقدار p-value آن ها از 0.05 کمتر باشد می توانند پیش بینی کننده ی خوبی باشند و هرچه این مقدار کمتر باشد آن متغیر بهتر است. در مدل سوال 6 تمام متغیر ها را در مدل نداشتیم یکبار دیگر مدل را با همان متغیر پاسخ و با تمام explanatory ها می سازیم و نتیجه بصورت زیر است :

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      5.503e+00  1.073e+00   5.129 2.92e-07 ***
charges          3.934e-04  3.114e-05  12.634  < 2e-16 ***
age             -1.006e-01  1.331e-02  -7.555 4.18e-14 ***
children        -2.439e-01  1.278e-01  -1.909   0.0563 .
sexmale          5.478e-01  3.018e-01   1.815   0.0695 .
bmi             -3.708e-01  4.623e-02  -8.020 1.06e-15 ***
regionnorthwest  1.459e-01  3.985e-01   0.366   0.7143
regionsoutheast  6.419e-01  4.206e-01   1.526   0.1269
regionsouthwest  3.200e-01  4.386e-01   0.730   0.4657
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1356.63  on 1337  degrees of freedom
Residual deviance:  302.57  on 1329  degrees of freedom
AIC: 320.57
```

کمترین میزان p-value را متغیر charges دارد بنابراین بهترین پیش بینی کننده است و بعد از آن به ترتیب bmi و age پیش بینی کننده های خوبی هستند.

b. متغیر کتگوریکال sex را انتخاب می کنیم و متغیر پاسخ هم که در قسمت قبل smoker انتخاب کردیم .با این دو متغیر یک مدل تعریف می کنیم و ضریب متغیر explanatory ما در این مدل همان لگاریتم odds ratio است بنابراین مقدار odds ratio را محاسبه می کنیم که نسبت دو احتمال را به ما میدهد سپس برای رسم منحنی OR به یکی از احتمال ها در بازه ی (0,1) مقدار می دهیم و مقدار احتمال دیگر را محاسبه می کنیم و نمودار را برای آن ها رسم می کنیم:

```
model_7_b <- glm(as.factor(smoker) ~ sex,data = insurance,family = binomial(link = "logit"))
summary(model_7_b)
```

شیب خط که همان لگاریتم شانس است می توانیم در نتیجه ی زیر ببینیم:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.5595     0.1026 -15.202  < 2e-16 ***
sexmale       0.3804     0.1369   2.778  0.00547 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

بنابراین مقدار OR برابر است با :

```
> oddsratio <- exp(0.3804)
> oddsratio
[1] 1.46287
> |
```

این مقدار را در رابطه زیر جایگذاری می کنیم و رابطه ی بین دو احتمال را خواهیم داشت و با مقدار دهی به p(smoker|female) مقدار احتمال دیگر را محاسبه می کنیم:

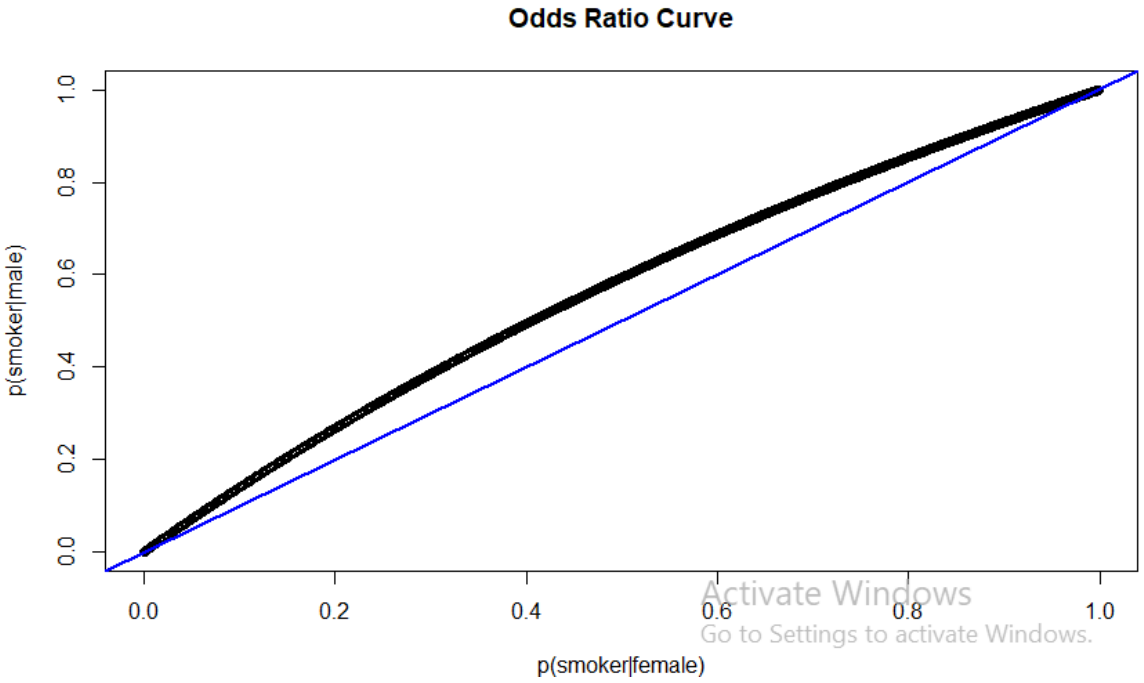
$$OR = \frac{p(smoker|male)/(1-p(smoker|male))}{p(smoker|female)/(1-p(smoker|female))}$$

$$p(smoker|male) = \frac{OR * p(smoker|female)/(1 - p(smoker|female)}{(1 + OR * p(smoker|female)/(1 - p(smoker|female))}$$

کد این قسمت بصورت زیر است:

```
model_7_b <- glm(as.factor(smoker) ~ sex,data = insurance,family = binomial(link = "logit"))
summary(model_7_b)
oddsratio <- exp(0.3804)
oddsratio
p1 <- c()
p2 <- c()
j <- 1
for(i in seq(from=0, to=1, by=0.001)){
  p_smoke_female <-i
  p_smoke_male <- (oddsratio*p_smoke_female/(1-p_smoke_female))/(1+(oddsratio*p_smoke_female/(1-p_smoke_female)))
  p1[j] <- p_smoke_female
  p2[j] <- p_smoke_male

  j = j +1

}

plot(p1,p2,xlim=c(0,1),ylim=c(0,1),xlab="p(smoker|female)",ylab="p(smoker|male)",main = "Odds Ratio Curve")
abline(c(0,0),c(1,1),col= 'blue', lwd = 2)
```

منحنی OR نیز بصورت زیر است:



تفسیر منحنی OR :

c. بهترین پیش بینی کننده برای smoker دیدیم که متغیر chargesاست که pvalue کمتری نسبت به بقیه داشت و مدل جدید را با این متغیر می سازیم:

```
model_7_c <- glm(as.factor(smoker) ~ charges,data = insurance,family = binomial(link = "logit"))
summary(model_7_c)
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.698e+00  3.064e-01  -18.60   <2e-16 ***
charges      2.535e-04  1.593e-05   15.91   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1356.63  on 1337  degrees of freedom
Residual deviance:  503.65  on 1336  degrees of freedom
AIC: 507.65

Number of Fisher Scoring iterations: 6
```

d. برای پیدا کردن بهترین threshold ماتریس confusion را تشکیل می دهیم و مقادیر sensitivity و specificity را به ازای مقادیر آستانه مختلف محاسبه میکنیم جاییکه هر دو مقدار نسبتا خوبی داشته باشند بهترین threshold را داریم :

```
predictTrain = predict(model_7_c, type="response")
summary(predictTrain)
tapply(predictTrain, insurance$smoker, mean)
sens <- c()
spec <- c()
j <- 1
for(i in seq(0,1,0.1)){
  table_7_d <- table(insurance$smoker, predictTrain > i)
  sens[j] <- table_7_d[4]/(table_7_d[4]+table_7_d[3])
  spec[j] <- table_7_d[1]/(table_7_d[2]+table_7_d[1])
  j = j+1
}
sens
spec
```

مقادیر به صورت زیر است :

مقادیر sensitivity :

```
0.6783042 0.7613293 0.7697595 0.7945736 0.8008658 0.8194444 0.8505155 0.9080460 0.9615385
```

مقادیر specificity :

```
0.9978655 0.9781529 0.9522445 0.9361111 0.9196025 0.9135472 0.9047203 0.9003436 0.8950931
```

به ترتیب از سمت چپ به راست به ازای مقادیر آستانه 0.1 تا 0.9 با گام 0.1 بدست آمده اند و میبینیم که مقدار 0.9 sensitivity و specificity خوبی دارد.

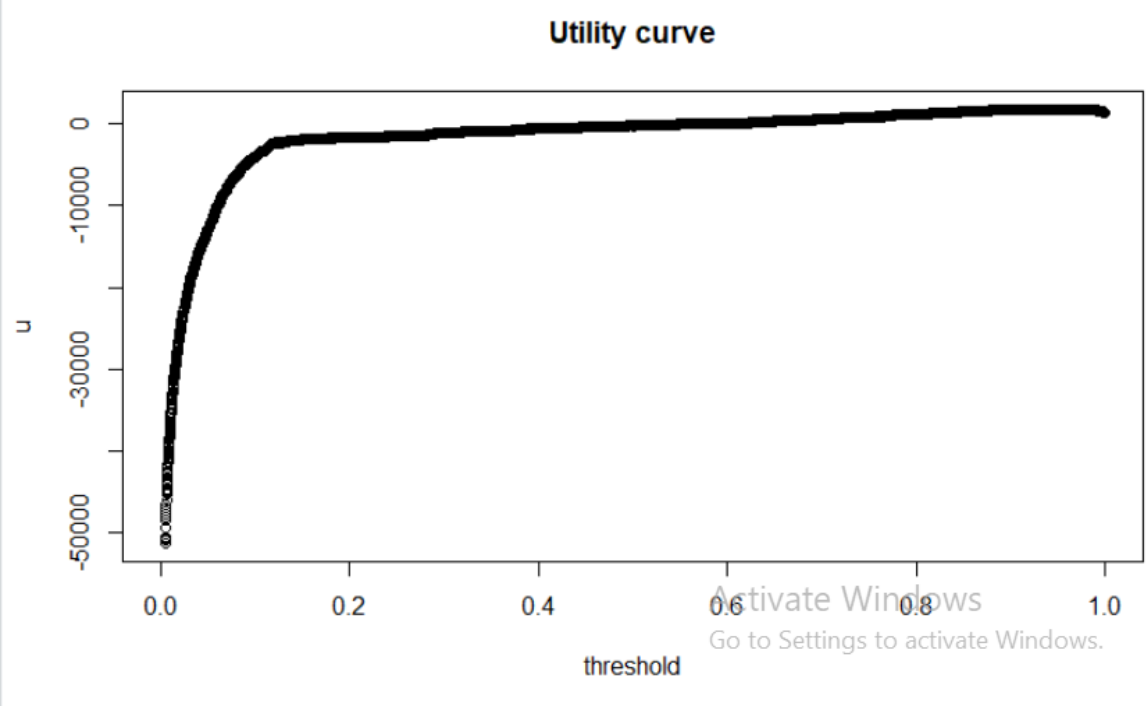
.e

مشابه حالت قبل به ازای threshold های مختلف و اینبار با گام 0.0001 تابع utility را محاسبه کردیم و منحنی آن هم در ادامه آمده است و این تابع در مقدار استانه 0.975 بیشینه میشود و از نمودار هم همین را می توان برداشت کرد.

```r
u <- c()
threshold <- c()
sensitivity <- c()
specificity <- c()
j <- 1
for(i in seq(0,1,0.0001)){
  table_7_e <- table(insurance$smoker, predictTrain > i)

  sensitivity[j] <- table_7_e[4]/(table_7_e[4]+table_7_e[3])
  specificity[j] <- table_7_e[1]/(table_7_e[2]+table_7_e[1])
  u[j] <- table_7_e[1]+table_7_e[2]+(-50)*table_7_e[3]+5*table_7_e[4]
  threshold[j] <- i
  print(u[j])
  j = j+1
}

max_index <- which.max(u)
threshold[max_index]
plot(threshold,u,main='Utility curve')
sensitivity[max_index]
specificity[max_index]
```

نتیجه به صورت زیر است :

```
> threshold[max_index]
[1] 0.9749
> plot(threshold,u,main='Utility curve')
> sensitivity[max_index]
[1] 1
> specificity[max_index]
[1] 0.869281
>
```



QUESTION8

متغیر high medical cost را با بصورت کتگوریکال باینری تعریف می کنیم که برای داده هایی که متغیر charges آنها از میانه ی این متغیر بیشتر است مقدار متغیر جدید را TRUE و برای مقادیر کمتر مقدار متغیر جدید را FALSE قرار می دهیم.

سپس مدل جدید را با استفاده از این متغیر جدید به عنوان متغیر پاسخ و متغیرهای دیگر دیتاست بعنوان explanatory می سازیم.

```
high_medical_cost <- numeric(nrow(insurance))

high_medical_cost[insurance$charges < median(insurance$charges)] <- FALSE
high_medical_cost[insurance$charges > median(insurance$charges)] <- TRUE

# mean(insurance$charges)
# median(insurance$charges)
# high_medical_cost
model_8 <- glm(factor(high_medical_cost) ~ age+sex+bmi+children+region+smoker,
               data = insurance,family = binomial)
summary(model_8)
```

خروجی را در زیر میبینیم که متغیرهای age ، bmi و region معنادار می شوند.

```
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -8.17993    0.67948 -12.038  < 2e-16 ***
age               0.16683    0.01004  16.624  < 2e-16 ***
sexmale          -0.35313    0.18188  -1.942  0.05219 .
bmi               0.03268    0.01582   2.065  0.03891 *
children          0.14483    0.07495   1.932  0.05333 .
regionnorthwest  -0.41109    0.25915  -1.586  0.11267
regionsoutheast  -0.86119    0.26801  -3.213  0.00131 **
regionsouthwest  -0.77646    0.25872  -3.001  0.00269 **
smokeryes        22.32977  509.88463   0.044  0.96507
---
```