۱-

a) مدل linear regression علی به صورت زیر خواهد بود:

$$\widehat{studying\ hours} = 2.544 + 0.164 \times caffeine$$

$df = n - 2 = 20 - 2 = 18$, $SE_{b_1} = 0.057$, $b_1 = 0.164$

```
> qt(0.975,df=18)
[1] 2.100922
```

$t_{18}^{*} = 2.1$

$confidence\ Interval: b_1 \pm t_{df}^{*} SE_{b_1} \rightarrow 0.164 \pm 2.1 \times 0.057 =$ (0.0443,0.2837)

---

b)

i. شروط استفاده از linear regression به شرح زیر است:

- Linearity: رابطه‌ی بین متغیر explanatory (متغیر first exam score) و متغیر Response (متغیر second exam score) باید خطی باشد.
- Nearly normal residuals: توزیع residualها باید تقریباً نرمال باشد.
- Constant variability: variability نقاط در اطراف خط regression باید تقریباً ثابت باشد.

ii. مدل regression به صورت زیر خواهد بود:

$$\widehat{second\ exam\ score} = 6.985 + 0.881 \times first\ exam\ score$$

$H_0: \beta_1 = 0$ متغیر $first\ exam\ score$ پیش‌بینی‌کننده خوبی برای متغیر $second\ exam\ score$ نمی‌باشد

$H_A: \beta_1 \neq 0$ متغیر $first\ exam\ score$ پیش‌بینی‌کننده خوبی برای متغیر $second\ exam\ score$ می‌باشد

$$t - statistic\ for\ the\ slope(coefficient\ of\ \text{first exam score}\ variable):\ T = \frac{b_1 - 0}{SE_{b_1}} \ ,\ df = n - 2$$

$$T = \frac{b_1 - 0}{SE_{b_1}} = \frac{0.881}{0.11} \approx 8 \ , df = 41 - 2 = 39$$

```
> 2*pt(8, df=18, lower.tail =FALSE)
[1] 2.450721e-07
```

$p - value = P(|T| > 8) = 0 < \alpha = 0.05 \rightarrow$ $\boldsymbol{Reject\ H_0}$

پس می‌توانیم نتیجه بگیریم که شیب خط regression مخالف صفر بوده و درنتیجه متغیر first exam score پیش‌بینی‌کننده‌ی خوبی برای متغیر second exam score می‌باشد.

---

c) نقطه‌ی C اگر حذف شود، بیشترین میزان تغییر در شیب‌خط regression ایجاد می‌شود و در واقع با حذف C، شیب‌خط بیشترین کاهش را پیدا کرده و به شیب افقی نزدیک می‌شود. در نتیجه R^2 نیز کاهش می‌یابد.

۲-

$Least\ Square\ Line: \hat{y} = b_0 + b_1 x$

$b_1 = \frac{s_y}{s_x} R \quad , \qquad R = cor(x, y)$

```
> sd(c(95,85,80,70,60))
[1] 13.50926
```

$s_x = 13.50926$

```
> sd(c(85,95,70,65,70))
[1] 12.5499
```

$s_y = 12.5499$

```
> x =c(95,85,80,70,60)
> y = c(85,95,70,65,70)
> cor(x,y)
[1] 0.6930525
```

$R = 0.6930525 \approx \qquad b_1 = \frac{s_y}{s_x} R = \frac{12.5499}{13.50926} \times 0.6930525 = 0.6438354 \approx 0.64$

می‌دانیم که خطر رگرسیون از نقطه $(\bar{x}, \bar{y})$ می‌گذرد. حال با داشتن مقدار b_1 و قرار دادن $(\bar{x}, \bar{y})$ در معادله خط رگرسیون می‌توانیم b_0 را بدست آوریم:

```
> mean(x)
[1] 78
> mean(y)
[1] 77
```

$77 = b_0 + 0.64 \times 78 = b_0 + 49.92 \ \rightarrow \ b_0 = 27.08$

معادله‌ی خط regression به صورت زیر خواهد شد:

$$\hat{y} = 27.08 + 0.64\, x$$

۳-

a) فرض $H_0: \beta_1 = 0$ به این معناست که متغیر دما predictor خوبی برای متغیر مسافت نیست.

$$0 \in (-0.02, 0.12) \rightarrow \boldsymbol{fail\ to\ Reject\ H_0}$$

پس H0 را نمی‌توانیم reject کنیم. یعنی که متغیر دما predictor خوبی برای متغیر مسافت نیست.

پس گزینه (ii) درست می‌باشد.

---

b)

$$\ln\widehat{(LE)} = 6.33 - 0.78\ln(\mathrm{HR})$$

$$hr = 60 \rightarrow \ln\widehat{(LE)} = 6.33 - 0.78\ln(60) = 6.33 - 0.78 * 4.094 = 3.13668$$

$$\widehat{LE} = e^{3.13668} = 23.027$$

پس طول عمر مورد انتظار فردی با میانگین ضربان قلب ۶۰ حدوداً برابر است با ۲۳ سال(به سمت پایین گرد شد)

۴-

a)

```
library(MASS)
absenteeism <- quine
#a)Converting the Eth, Sex, and Lrn variables to binary variables
absenteeism$Eth <- ifelse(absenteeism$Eth == "N",1,0)
absenteeism$Sex <- ifelse(absenteeism$Sex == "M",1,0)
absenteeism$Lrn <- ifelse(absenteeism$Lrn == "SL",1,0)
```

b)

```
#b)fitting the model
## Explanatory variables: Eth,Sex,Lrn
##Response variable: Days
absenteeism_mlr <- lm(Days ~ Eth + Sex + Lrn,data = absenteeism)
summary(absenteeism_mlr)
```

The Result:

```
Call:
lm(formula = Days ~ Eth + Sex + Lrn, data = absenteeism)

Residuals:
    Min      1Q  Median      3Q     Max 
-22.190 -10.078  -4.928   5.768  59.914 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   18.932      2.570   7.365 1.32e-11 ***
Eth           -9.112      2.599  -3.506 0.000609 ***
Sex            3.104      2.637   1.177 0.241108    
Lrn            2.154      2.651   0.813 0.417732    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.67 on 142 degrees of freedom
Multiple R-squared:  0.08933,	Adjusted R-squared:  0.07009 
F-statistic: 4.643 on 3 and 142 DF,  p-value: 0.003967
```

c)

$$\widehat{Days} = 18.932 - 9.112 Eth{:}NotAboriginal + 3.104 Sex{:}male + 2.154 Lrn{:}SL$$

تفسیر ضریب متغیر Eth: با فرض ثابت ماندن سایر چیزها(دو دانش‌آموز وجود داشته باشند که جنسیت و سطح یادگیری یکسان داشته باشند) ، مدل ما پیش‌بینی می‌کند تعداد روزهایی که یک دانش‌آموز غیربومی(NotAboriginal) در طول سال از مدرسه غیبت می‌کند به طور متوسط 9.112 روز از دانش‌آموزان بومی کمتر است.

تفسیر ضریب متغیر Sex: با فرض ثابت ماندن سایر چیزها(دو دانش‌آموز وجود داشته باشند که جنسیت و سطح یادگیری یکسان داشته باشند) ، مدل ما پیش‌بینی می‌کند تعداد روزهایی که یک دانش‌آموز مذکر(male) در طول سال از مدرسه غیبت می‌کند به طور متوسط 3.104 روز از دانش‌آموزان مونث بیشتر است.

تفسیر ضریب متغیر Lrn: با فرض ثابت ماندن سایر چیزها(دو دانش‌آموز وجود داشته باشند که جنسیت و قومیت یکسان داشته باشند) ، مدل ما پیش‌بینی می‌کند تعداد روزهایی که یک دانش‌آموز با سطح یادگیری کند(SL) در طول سال از مدرسه غیبت می‌کند به طور متوسط 2.154 روز از دانش‌آموزان با سطح یادگیری متوسط بیشتر است.

تفسیر عرض از مبدأ(intercept): انتظار داریم یک دانش‌آموز مونث بومی با سطح یادگیری متوسط، در طول سال 18.932 روز از مدرسه غیبت کند.

d) می‌توان از خروجی دستور summary مدلمون adjusted_R_squared را استخراج کرد:

```r
#outputting the adjusted_R_Square
sprintf("Adjusted_R_Squared: %s",summary(absenteeism_mlr)$adj.r.squared)
```

The Result:

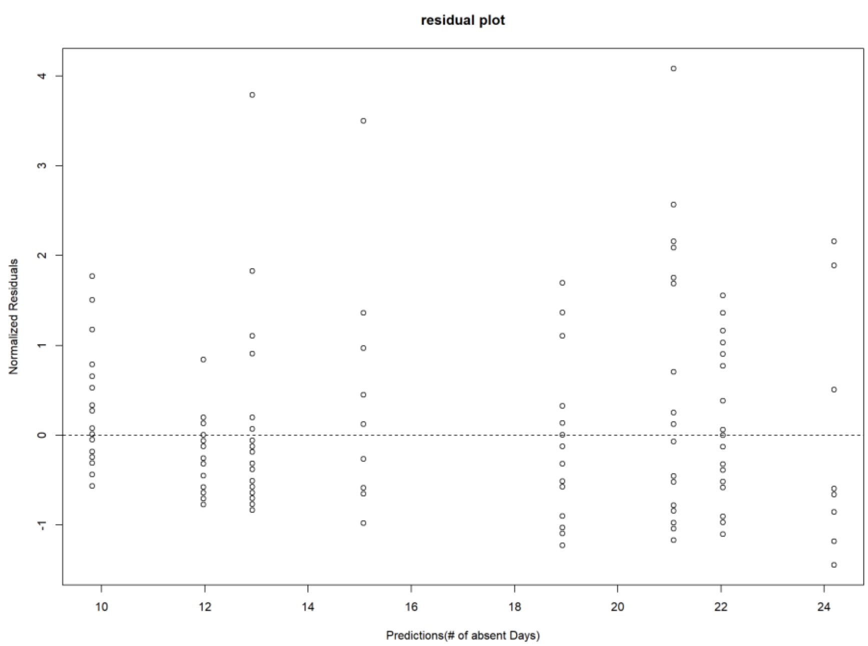
```
> #outputting the adjusted_R_Square
> sprintf("Adjusted_R_Squared: %s",summary(absenteeism_mlr)$adj.r.squared)
[1] "Adjusted_R_Squared: 0.0700920155850691"
```

معیار $adjusted\_R^2$ همواره از R^2 مقدار کمتری دارد چرا که برای هر predictor اضافه شده به مدل مقداری جریمه در نظر می‌گیرد و بیان می‌کند که چه درصد از variability موجود در متغیر Days توسط مدل‌ما (سه متغیر Eth و Sex و Lrn)، با در نظر گرفتن جریمه برای آنها، توضیح داده می‌شود.

e) برای رسم residual plot از مقادیر استاندارد شده(نرمال شده) residual ها استفاده کردیم. علت این کار این است که در تحلیل رگرسیون یک توزیع multivariate واریانس residualها برای مقادیر مختلف متغیر ورودی متفاوت است.

```
#Residual Plot
plot(absenteeism_mlr$fitted.values, rstudent(absenteeism_mlr) ,main="residual plot",
     xlab="Predictions(# of absent Days)",ylab="Normalized Residuals")
abline(h=0, lty="dashed")
```

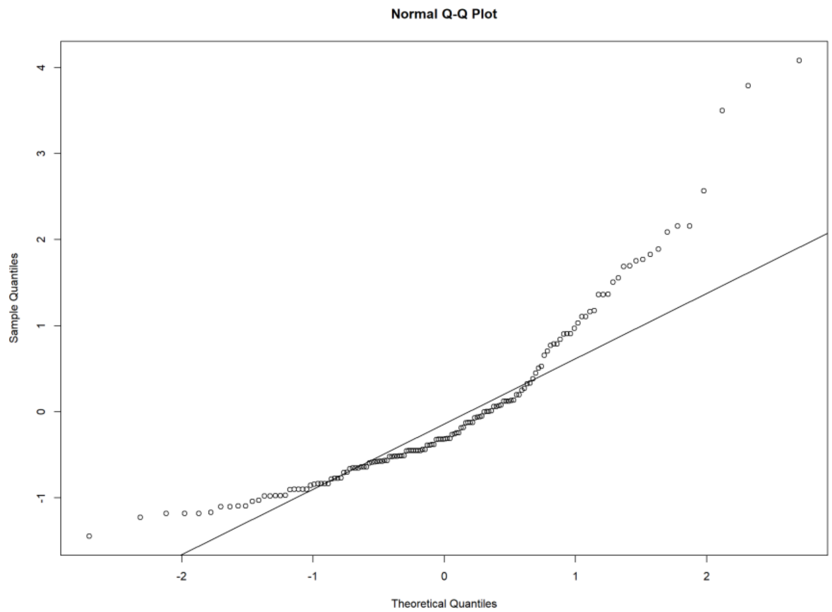
The Result:



با توجه به residual plot می‌توانیم ببینیم که این مدل، مدل خوبی نمی‌باشد. چرا که واریانس residualها حول خط residual=0 دارای مقادیر نسبتاً ثابتی نمی‌باشند.(یعنی شرط Constant variability برقرار نمی‌باشد) همچنین می‌توان مشاهده کرد که توزیع residualها اختلاف زیادی با توزیع نرمال دارند. برای بررسی دقیق‌تر این شرط می‌توانیم Normal Q-Q plot را به صورت زیر رسم کنیم.

```
#checking the Normality of Residuals
qqnorm(rstudent(absenteeism_mlr))
points(qqline(rstudent(absenteeism_mlr)))
```

The Result:



طبق normal QQ plot بالا، توضیحات بالای ما درست بوده و شرط توزیع تقریباً نرمال residualها(nearly normal residuals) هم نقض می‌شود.

۵-

a) نادرست- فقط کافیست که correlation متغیر explanatory اضافه شده با متغیر response صفر نباشد در این صورت اضافه کردن این متغیر به مدل ما باعث افزایش R^2 می‌شود ولی اگر correlation متغیر explanatory اضافه شده با متغیر response دقیقاً برابر با صفر باشد، R^2 تغییری نمی‌کند. همچنین از آنجایی که گفته شده این متغیر predictor خوبی برای متغیر response نمی‌باشد، و در معیار $adjusted\_R^2$ ما جریمه برای اضافه کردن تعداد متغیرها در مدل در نظر می‌گیریم، پس میزان $adjusted\_R^2$ کاهش می‌یابد.

b) درست

c) نادرست- correlation معیاری برای نشان دادن میزان همبستگی خطی بین دو متغیر می‌باشد نه هر نوع رابطه‌ای. مثلاً ممکن است بین دو متغیر رابطه‌ی درجه ۲ قوی‌ای وجود داشته باشد اما correlation بین آنها بسیار ناچیز است چون رابطه خطی بین آنها وجود ندارد.

d) نادرست- وقتی متغیرهای explanatoryمون collinear هستند یعنی correlation بالایی بین آنها وجود دارد. از این رو قرار دادن/ندادن یکی از این متغیرها در مدل، در مقدار ضریب متغیر دیگر تأثیرگذار است.

۶-

a)

$$\widehat{annual\ murder\ rate} = -29.901 + 2.559\%poverty$$

b) تفسیر intercept: در ناحیه‌ی شهری‌ای که درصد فقر برابر با صفر است به طور میانگین سالانه 29.901- میلیون قتل رخ می‌دهد. (می‌بینیم که intercept در اینجا بی‌معنی است.)

تفسیر slope: با افزایش هر واحد (در اینجا یک درصد) در درصد فقر یک ناحیه‌ی شهری، انتظار داریم تعداد میلیون قتل‌های سالیانه به طور متوسط ۲.۵۵۹ واحد(۲.۵۵۹ میلیون قتل در سال) اضافه شود.

تفسیر R^2: 70.52 درصد از variability موجود در تعداد قتل‌های سالیانه توسط مدل ما (توسط متغیر درصد فقر) توضیح داده می‌شود.

---

c) با توجه به scatter plot رسم شده می‌بینیم که بین دو متغیر مذکور یک correlation مثبت وجود دارد. پس کافیست از R^2 جذر گرفته و همان مقدار بدست آمده رو به عنوان correlation coefficient درنظر بگیریم:

$R^2 = 70.52\% \ \rightarrow |R| = R = \sqrt{R} = \sqrt{0.7052} = 0.84$

---

d)

$H_0: \beta_1 = 0$ متغیر $\%poverty$ پیش‌بینی‌کننده خوبی برای متغیر $annual\ murder\ rate$ نمی‌باشد

$H_A: \beta_1 \neq 0$ متغیر $\%poverty$ پیش‌بینی‌کننده خوبی برای متغیر $annual\ murder\ rate$ می‌باشد

$$t - statistic\ for\ the\ slope(coefficient\ of\ \%\text{poverty}\ variable):\ T = \frac{b_1 - 0}{SE_{b_1}}\ ,\ \ df = n - 2$$

$$T = \frac{b_1 - 0}{SE_{b_1}} = \frac{2.599}{0.390} = 6.664103\ \ ,df = 20 - 2 = 18$$

```
> 2*pt(6.664103,df=18,lower.tail = FALSE)
[1] 2.977667e-06
```

$p - value = P(|T| > 6.664103) \approx\ 0 < 0.05 \quad \rightarrow\ \boldsymbol{Reject\ H_0}$

پس می‌توانیم نتیجه بگیریم که شیب خط regression مخالف صفر بوده و درنتیجه متغیر poverty% پیش‌بینی‌کننده خوبی برای متغیر annual murder rate می‌باشد.

---

e) در قسمت d کامل نوشته شد.

f)

```
> qt(0.975,df=18)
[1] 2.100922
```

$t_{18}^* = 2.1$

$confidence\ Interval: b_1 \pm t^*_{df} SE_{b_1} \quad \rightarrow 2.599 \pm 2.1 \times 0.390 = (1.78,3.418)$

$0 \notin (1.78\,,3.418) \rightarrow$ ***Reject H_0***

---

g) بله- باتوجه به اینکه مقدار null value (یعنی 0) در بازه‌ی اطمینان وجود ندارد، معادل این است که null hypothesis را رد کنیم که طبق آزمون فرض بالا این اتفاق افتاده.

۷-

a) تست ANOVA. چرا که بیش از یک متغیر explanatory داریم که میخواهیم برای coefficient آنها آزمون فرض طراحی کنیم.(میخواهیم بررسی کنیم آیا ضریب این متغیرها صفر هست یا نه. یعنی آیا predictor خوبی برای مدل رگرسیون ما هستند یا نه)

b) مدل رگرسیون ما به صورت زیر خواهد بود:

$$\widehat{commute}\ time = b_0 + b_1 Day{:}Mon + b_2 Day{:}Tue + b_3 Day{:}Wed + b_4 Day{:}Thu$$

با توجه به مدل رگرسیون بالا، فرض‌های آزمون فرض به صورت زیر خواهد بود:

$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ میان هیچکدام از روزها و زمان رفت‌وآمد رابطه‌خطی وجود ندارد

$H_A: \beta_i \neq 0$ بین حداقل یکی از روزها و زمان رفت‌وآمد رابطه‌خطی وجود دارد یعنی $predictor$ خوبی برای زمان رفت‌وآمد می‌باشد

c)

Source	DF	Sum of Squares	Mean Square	F_value	Prob.
Day (Groups)	**4**	14.28	**3.57**	**3.36**	**0.03746967**
Error (Residuals)	**15**	**15.92**	**1.061**		
Total	19	30.2			

$SS_{reg} = SS_{Tot} - SS_{res} \ \rightarrow \ SS_{res} = SS_{Tot} - SS_{reg} = \ 30.2 - 14.28 = 15.92$

$5\ Days\ in\ week(5\ categories) \rightarrow (4\ binary\ variables) 4\ predictors \ \rightarrow \ df_{Reg} = 4$

$df_{Res} = \ df_{Tot} - \ df_{Reg} = 19 - 4 = 15$

$$MS_{Reg} = \frac{SS_{Reg}}{df_{Reg}} = \frac{14.28}{4} = 3.57$$

$$MS_{Res} = \frac{SS_{Res}}{df_{Res}} = \frac{15.92}{15} = 1.061$$

$$F\_Value(4,15) = \frac{MS_{Reg}}{MS_{Res}} = \frac{3.57}{1.061} = 3.36$$

```
> pf(3.36 ,4 ,15 ,lower.tail = FALSE)
[1] 0.03746967
```

d) با توجه به میزان احتمال که در بخش (c) بدست آوردیم(prob. = 0.037) فرض null را رد می‌کنیم.

$$probability(> F) = P(F > 3.36) = 0.037 < \alpha = 0.05 \quad \rightarrow \boldsymbol{Reject\ H_0}$$

e) یعنی بین حداقل یکی از روزهای هفته و زمان رفت‌وآمد یک رابطه خطی وجود دارد. به عبارت دیگر حداقل یکی از روزهای هفته پیش‌بینی کننده‌ی خوبی برای مدت زمان رفت‌وآمد می‌باشد.

۸-

a) ابتدا مدل را با متغیرهای explanatory مشخص شده و متغیر life expectancy به عنوان متغیر Response می‌سازیم (fit می‌کنیم):

```
##Explanatory variables: Population,Income,Illiteracy,Murder,HS.Grad,Frost,Area
df_states <- data.frame(state.x77)
#a)
model1 <- lm(Life.Exp ~ Population + Income + Illiteracy + Murder + HS.Grad + Frost + Area,data = df_states)
summary(model1)
```

The Result:

```
Call:
lm(formula = Life.Exp ~ Population + Income + Illiteracy + Murder + 
    HS.Grad + Frost + Area, data = df_states)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.48895 -0.51232 -0.02747  0.57002  1.49447 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7.094e+01  1.748e+00  40.586  < 2e-16 ***
Population   5.180e-05  2.919e-05   1.775   0.0832 .  
Income      -2.180e-05  2.444e-04  -0.089   0.9293    
Illiteracy   3.382e-02  3.663e-01   0.092   0.9269    
Murder      -3.011e-01  4.662e-02  -6.459 8.68e-08 ***
HS.Grad      4.893e-02  2.332e-02   2.098   0.0420 *  
Frost       -5.735e-03  3.143e-03  -1.825   0.0752 .  
Area        -7.383e-08  1.668e-06  -0.044   0.9649    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7448 on 42 degrees of freedom
Multiple R-squared:  0.7362,	Adjusted R-squared:  0.6922 
F-statistic: 16.74 on 7 and 42 DF,  p-value: 2.534e-10
```

حال با توجه به مقدار p-value هر متغیر می‌توانیم بدترین متغیر(متغیری که بیشترین مقدار p-value را دارد) حذف کنیم و مدل را بدون آن متغیر بسازیم. که در اینجا متغیر Area بدترین متغیر بوده و بیشترین مقدار p-value(0.9649) را دارد. پس Area را حذف کرده و مدل را با سایر متغیرها fit می‌کنیم:

```
model2 <- lm(Life.Exp ~ Population + Income + Illiteracy + Murder + HS.Grad + Frost ,data = df_states)
summary(model2)
```

The Result:

```
Call:
lm(formula = Life.Exp ~ Population + Income + Illiteracy + Murder + 
    HS.Grad + Frost, data = df_states)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.49047 -0.52533 -0.02546  0.57160  1.50374 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7.099e+01  1.387e+00  51.165  < 2e-16 ***
Population   5.188e-05  2.879e-05   1.802   0.0785 .  
Income      -2.444e-05  2.343e-04  -0.104   0.9174    
Illiteracy   2.846e-02  3.416e-01   0.083   0.9340    
Murder      -3.018e-01  4.334e-02  -6.963 1.45e-08 ***
HS.Grad      4.847e-02  2.067e-02   2.345   0.0237 *  
Frost       -5.776e-03  2.970e-03  -1.945   0.0584 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7361 on 43 degrees of freedom
Multiple R-squared:  0.7361,	Adjusted R-squared:  0.6993 
F-statistic: 19.99 on 6 and 43 DF,  p-value: 5.362e-11
```

می‌بینیم که با حذف متغیر Area ، مقدار معیار $adjusted\_R^2$ مدل جدید نسبت به مدل قبلی افزایش پیدا کرده. پس حذف متغیر Area درست بوده و الگوریتم را ادامه می‌دهیم. در اینجا متغیر Illiteracy بدترین متغیر بوده و بیشترین مقدار p-value (0.9340) را دارد. پس Illiteracy را حذف کرده و مدل را با سایر متغیرها fit می‌کنیم:

```
model3 <- lm(Life.Exp ~ Population + Income + Murder + HS.Grad + Frost ,data = df_states)
summary(model3)
```

The Result:

```
Call:
lm(formula = Life.Exp ~ Population + Income + Murder + HS.Grad + 
    Frost, data = df_states)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.4892 -0.5122 -0.0329  0.5645  1.5166 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7.107e+01  1.029e+00  69.067  < 2e-16 ***
Population   5.115e-05  2.709e-05   1.888   0.0657 .  
Income      -2.477e-05  2.316e-04  -0.107   0.9153    
Murder      -3.000e-01  3.704e-02  -8.099 2.91e-10 ***
HS.Grad      4.776e-02  1.859e-02   2.569   0.0137 *  
Frost       -5.910e-03  2.468e-03  -2.395   0.0210 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7277 on 44 degrees of freedom
Multiple R-squared:  0.7361,	Adjusted R-squared:  0.7061 
F-statistic: 24.55 on 5 and 44 DF,  p-value: 1.019e-11
```

می‌بینیم که با حذف متغیر Illiteracy ، همچنان مقدار معیار $adjusted\_R^2$ مدل جدید نسبت به مدل قبلی افزایش پیدا کرده. پس حذف متغیر Income درست بوده و الگوریتم را ادامه می‌دهیم. در اینجا متغیر Income بدترین متغیر بوده و بیشترین مقدار p-value (0.9153) را دارد. پس Income را حذف کرده و مدل را با سایر متغیرها fit می‌کنیم:

```r
model4 <- lm(Life.Exp ~ Population + Murder + HS.Grad + Frost ,data = df_states)
summary(model4)
```

The Result:

```
Call:
lm(formula = Life.Exp ~ Population + Murder + HS.Grad + Frost,
    data = df_states)

Residuals:
     Min       1Q   Median       3Q      Max
-1.47095 -0.53464 -0.03701  0.57621  1.50683

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.103e+01  9.529e-01  74.542  < 2e-16 ***
Population   5.014e-05  2.512e-05   1.996  0.05201 .
Murder      -3.001e-01  3.661e-02  -8.199 1.77e-10 ***
HS.Grad      4.658e-02  1.483e-02   3.142  0.00297 **
Frost       -5.943e-03  2.421e-03  -2.455  0.01802 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7197 on 45 degrees of freedom
Multiple R-squared:  0.736,	Adjusted R-squared:  0.7126
F-statistic: 31.37 on 4 and 45 DF,  p-value: 1.696e-12
```

می‌بینیم که با حذف متغیر Income ، همچنان مقدار معیار $adjusted\_R^2$ مدل جدید نسبت به مدل قبلی افزایش پیدا کرده. پس حذف متغیر Income درست بوده و الگوریتم را ادامه می‌دهیم. در اینجا مقادیر p-valueها خیل کوچک شده‌اند و همه به غیر از متغیر Population مقدار کوچک تر از 0.05 دارند. البته مقدار p-value متغیر Population خیلی نزدیک به 0.05می‌باشد. ولی اگر بخواهیم الگوریتم را یک گام دیگر پیش ببریم و متغیر Population را حذف کنیم و مدل را با سایر متغیرها fit می‌کنیم خواهیم داشت:

```r
model5 <- lm(Life.Exp ~ Murder + HS.Grad + Frost ,data = df_states)
summary(model5)
```

The Result:

```
Call:
lm(formula = Life.Exp ~ Murder + HS.Grad + Frost, data = df_states)

Residuals:
    Min      1Q  Median      3Q     Max
-1.5015 -0.5391  0.1014  0.5921  1.2268

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 71.036379   0.983262  72.246  < 2e-16 ***
Murder      -0.283065   0.036731  -7.706 8.04e-10 ***
HS.Grad      0.049949   0.015201   3.286  0.00195 **
Frost       -0.006912   0.002447  -2.824  0.00699 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7427 on 46 degrees of freedom
Multiple R-squared:  0.7127,	Adjusted R-squared:  0.6939
F-statistic: 38.03 on 3 and 46 DF,  p-value: 1.634e-12
```

می‌بینیم با حذف متغیر Population ، مقدار معیار $adjusted\_R^2$ مدل جدید نسبت به مدل قبلی کاهش پیدا کرده(مقدار معیار $adjusted\_R^2$ در مدل جدید برابر 0.6939 می‌باشد در حالیکه در مدل قبلی برابر 0.7126 بود) پس حذف متغیر Population درست نبوده و مدل مرحله قبل(یعنی model4) مدل نهایی ما خواهد بود.

b) ابتدا مدل را با توجه به متغیرهای explanatory و response گفته شده fit می‌کنیم:

```
## Response variable: Life.Exp
##Explanatory variable: Murder
model_b <- lm(Life.Exp ~ Murder ,data = df_states)
summary(model_b)
```

The Result:

```
Call:
lm(formula = Life.Exp ~ Murder, data = df_states)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.81690 -0.48139  0.09591  0.39769  2.38691 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 72.97356    0.26997  270.30  < 2e-16 ***
Murder      -0.28395    0.03279   -8.66 2.26e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8473 on 48 degrees of freedom
Multiple R-squared:  0.6097,	Adjusted R-squared:  0.6016 
F-statistic: 74.99 on 1 and 48 DF,  p-value: 2.26e-11
```

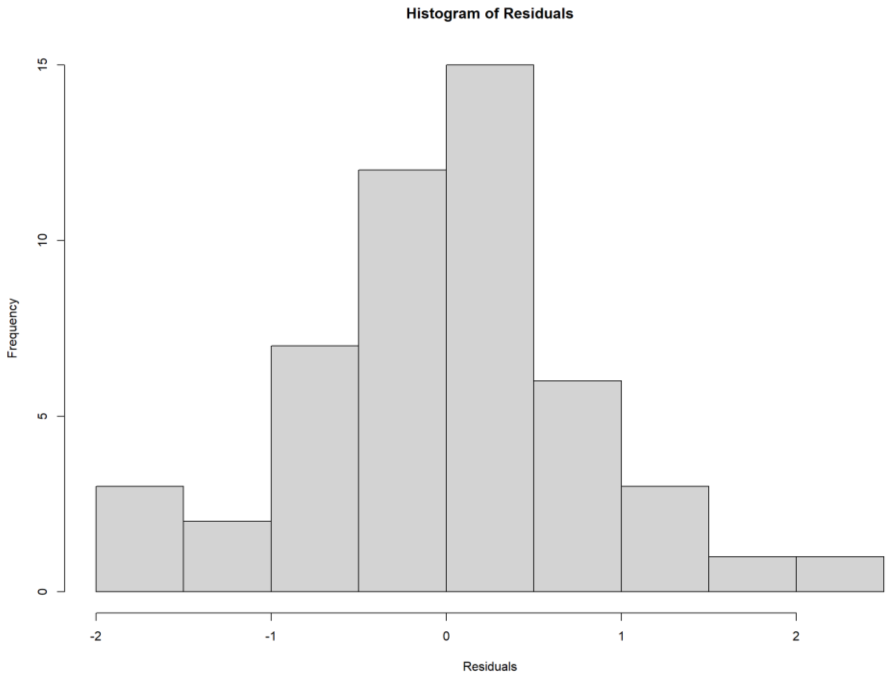
حال با توجه به مدل بدست آمده معادله‌ی رگرسیون را می‌نویسیم:

$$\widehat{Life.Exp} = 72.97356 - 0.28395 \times Murder$$

c)

```
#c)histogram of residuals
hist(model_b$residuals, xlab = "Residuals", main = "Histogram of Residuals")
```

The Result:



حال میانگین و انحراف معیار Residualها را محاسبه می‌کنیم:

```
##Calculating mean and SD of Residuals
sprintf("Mean of Residuals: %s",mean(model_b$residuals))
sprintf("Standard deviation of Residuals: %s",sd(model_b$residuals))
```

The Result:

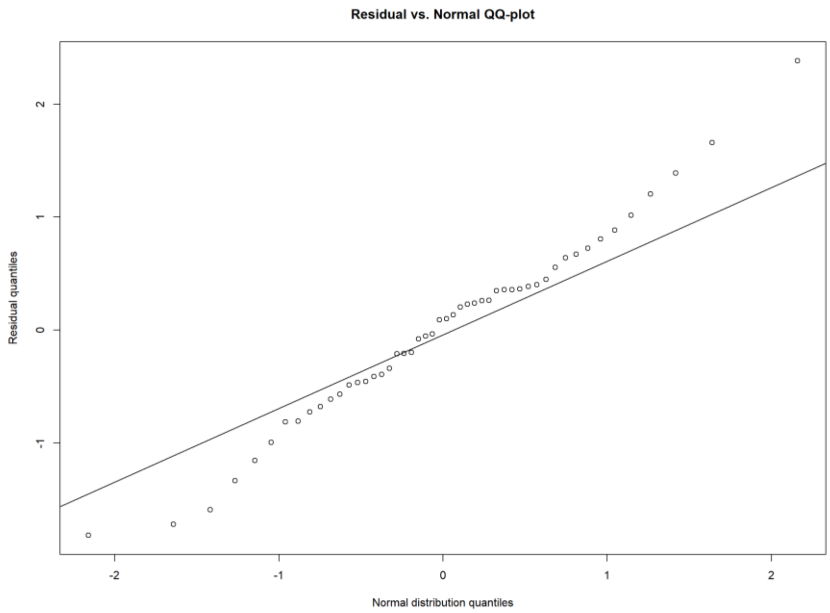
```
> ##Calculating mean and SD of Residuals
> sprintf("Mean of Residuals: %s",mean(model_b$residuals))
[1] "Mean of Residuals: 1.16204788846996e-17"
> sprintf("Standard deviation of Residuals: %s",sd(model_b$residuals))
[1] "Standard deviation of Residuals: 0.838625292285084"
```

$$\mu_{res} \approx 0 \quad ,\sigma_{res} \approx 0.84$$

d) با ۱۰۰ نقطه از یک توزیع نرمال با میانگین 0 و انحراف معیار بدست آمده در قسمت (b) quantile می‌سازیم و آن را با توزیع residualها مقایسه می‌کنیم.

```
#d)QQ-plot of residuals vs. Normal distribution with mean=0 and sd = model sd
qqplot(qnorm(ppoints(100),sd = sd(model_b$residuals)), model_b$residuals,
       main = "Residual vs. Normal QQ-plot",xlab = "Normal distribution quantiles",
       ylab = "Residual quantiles")
qqline(model_b$residuals)
```

The Result:



با توجه به QQ-plot بالا می‌توان دید که توزیع residualها توزیعی نزدیک به یک توزیع نرمال با میانگین 0 و انحراف معیار حدود 0.84 دارد. پس residualها nearly normal هستند.
