Narges Goudarzi
Assignment #4

Question 1

a. $\hat{p}(Y=1|x)=\dfrac{e^{-2.7399+3.0287*1-1.2081*.5}}{1+e^{-2.7399+3.0287*1-1.2081*.5}}=0.4218$

b. Test statistic= $\dfrac{-1.2081}{0.4620}=-2.61$ it is larger than 1.96 it means we reject H_0 and β_2 is significant.

c. $Dev_{null}-Dev_{residual}=110.216-56.436=53.78$ It has 99-97=2 degree of freedom. According to chi square table $\chi^2_{0.05,2}=5.99$. It means we reject $H_0:\beta_1=\beta_2=0$.

Question 2

a)

P-value is larger than .05 => reject H_0, we have a big reduction in deviance.
We can reject H_0.

1) A)

		Y	
		0	1
$\hat{Y}$	0	3	2
	1	3	2

The accuracy is proportion correct prediction $\frac{5}{10} = .5$,

sensitivity is proportion of Y=1 is correct prediction $\frac{2}{4} = .5$

and precision of the prediction is proportion of the correctly predicted $\hat{Y} = 1$ among all positive

predicting $\frac{2}{5} = .4$

b)

		Y	
		0	1
$\hat{Y}$	0	5	3
	1	1	1

The accuracy is proportion correct prediction $\frac{6}{10} = .6$,

Sensitivity is proportion of Y=1 is correct prediction $\frac{1}{4} = .25$

and precision of the prediction is proportion of the correctly predicted $\hat{Y} = 1$ among all positive predicting $\frac{1}{2} = .5$

C)

Sensitivity is $p(\hat{Y}=1 \mid Y=1)$ .It means proportion of Y=1 which is correct prediction. If cut off decrease we allowing many positive prediction, many $\hat{Y}=1$ and a few $\hat{Y}=0$ as a result many observation with $Y=1$ will correctly predicted. It means sensitivity will increase.

Question 3

A)

fit_glm1 = glm(chd ~., data = SAheart, family = binomial(link = "logit"))
summary(fit_glm1)
n = nrow(SAheart)
cutoff = 0.5
y_hat = rep(0,n)
idx = which(fitted(fit_glm1)>cutoff)
y_hat[idx] = 1
y_hat
conf_mat=table(predicted=y_hat,actual=SAheart$chd)
conf_mat

```
         actual
predicted   0   1
        0 256  77
        1  46  83
```

```
mean(y_hat == SAheart$chd) # Accuracy
0.7337662
> conf_mat[2, 2] / sum(conf_mat[, 2]) # Sensitivity
[1] 0.51875
> conf_mat[2, 2] / sum(conf_mat[2, ]) # Precision
[1] 0.6434109
```

The accuracy is proportion correct prediction $\frac{256+83}{462}=0.73$,

Sensitivity is proportion of Y=1 is correct prediction $\frac{83}{160}=0.51875$

Specificity $\frac{256}{256+46}=0.848$

precision is $\frac{83}{129}=0.64$

b)

```
fit_back_bic = step(fit_glm1, direction = "backward", k=log(n),trace=0)
> fit_back_bic

Call:  glm(formula = chd ~ tobacco + ldl + famhist + typea + age, family = bi
nomial(link = "logit"),
    data = SAheart)

Coefficients:
   (Intercept)     tobacco      ldl          famhistPresent        typea         age
      -6.44644     0.08038   0.16199            0.90818          0.03712       0.05046

Degrees of Freedom: 461 Total (i.e. Null);  456 Residual
Null Deviance:      596.1
Residual Deviance: 475.7        AIC: 487.7
```

The best subset of predictors to predict chd is **`tobacco,ldl,famhistPresent,typea,age.`**

c)

We are going to test $H_0 : \beta_{alcohol} = \beta_{sbp} = \beta_{adiposity} = \beta_{obesity} = 0$

full model is:

$$\ln(\frac{p(x)}{1-p(x)}) = \beta_0 + \beta_{alcohol}x_1 + \beta_{sbp}x_2 + \beta_{adiposity}x_3 + \beta_{obesity}x_4 + \beta_{tobacco}x_5 + \beta_{ldl}x_6 + \beta_{famhistPresent}x_7 + \beta_{typea}x_8 + \beta_{age}x_9$$

reduced model is baced on backward method is::

$$\ln(\frac{p(x)}{1-p(x)}) = \beta_0 + \beta_{tobacco}x_5 + \beta_{ldl}x_6 + \beta_{famhistPresent}x_7 + \beta_{typea}x_8 + \beta_{age}x_9$$

```
fit_full = glm(chd ~ ., data = SAheart, family = binomial)
summary(fit_full)
fit_reduced = glm(chd ~.-alcohol-sbp-adiposity-obesity, data = SAheart, family = binomial)
summary(fit_reduced)
```

```
Call:
glm(formula = chd ~ ., family = binomial, data = SAheart)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.7781  -0.8213  -0.4387   0.8889   2.5435  

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)    
(Intercept)    -6.1507209  1.3082600  -4.701 2.58e-06 ***
sbp             0.0065040  0.0057304   1.135 0.256374    
tobacco         0.0793764  0.0266028   2.984 0.002847 ** 
ldl             0.1739239  0.0596617   2.915 0.003555 ** 
adiposity       0.0185866  0.0292894   0.635 0.525700    
famhistPresent  0.9253704  0.2278940   4.061 4.90e-05 ***
typea           0.0395950  0.0123202   3.214 0.001310 ** 
obesity        -0.0629099  0.0442477  -1.422 0.155095    
alcohol         0.0001217  0.0044832   0.027 0.978350    
age             0.0452253  0.0121298   3.728 0.000193 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 596.11  on 461  degrees of freedom
Residual deviance: 472.14  on 452  degrees of freedom
AIC: 492.14

Number of Fisher Scoring iterations: 5

> fit_reduced = glm(chd ~.-alcohol-sbp-adiposity-obesity, data = SAheart, fam
ily = binomial)
> summary(fit_reduced)

Call:
glm(formula = chd ~ . - alcohol - sbp - adiposity - obesity, 
    family = binomial, data = SAheart)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.9165  -0.8054  -0.4430   0.9329   2.6139  

Coefficients:
```

```
                 Estimate Std. Error z value Pr(>|z|)    
(Intercept)      -6.44644    0.92087  -7.000 2.55e-12 ***
tobacco           0.08038    0.02588   3.106  0.00190 ** 
ldl               0.16199    0.05497   2.947  0.00321 ** 
famhistPresent    0.90818    0.22576   4.023 5.75e-05 ***
typea             0.03712    0.01217   3.051  0.00228 ** 
age               0.05046    0.01021   4.944 7.65e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 596.11  on 461  degrees of freedom
Residual deviance: 475.69  on 456  degrees of freedom
AIC: 487.69

Number of Fisher Scoring iterations: 5
```

d)

We are going to test $H_0 : \beta_{alcohol} = \beta_{sbp} = \beta_{adiposity} = \beta_{obesity} = 0$

```
L=2* as.numeric(logLik(fit_full) - logLik(fit_reduced))
[1] 3.545546
```

Degree of freedom is 10-6=4

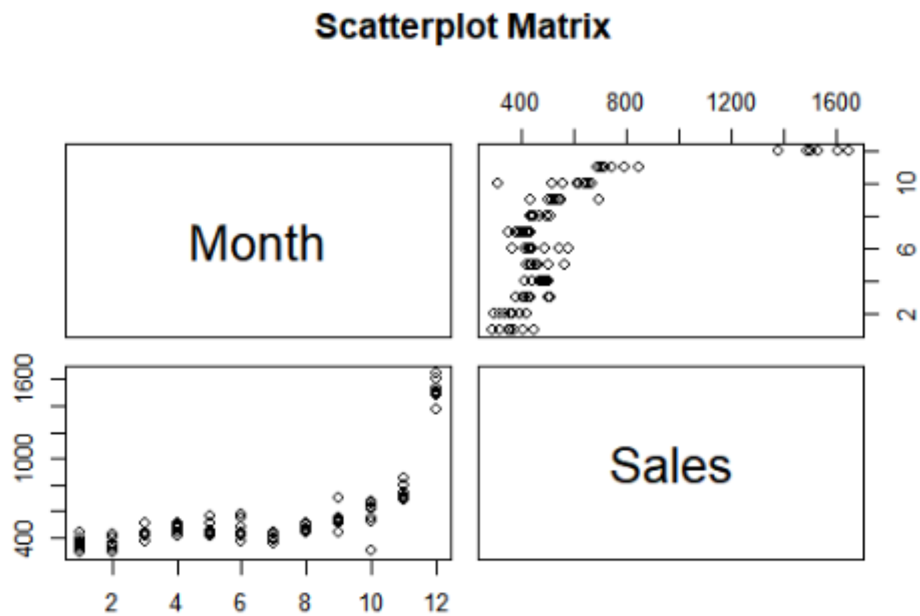
```
1-pchisq(L,4)
[1] 0.4709869
```

It is more than 0.05 it means we fail to reject H_0 and $\beta_{alcohol}, \beta_{sbp}, \beta_{adiposity}, \beta_{obesity}$ are equal to zero and we can ignore them in the full model.

Question 4

a)

```
pairs(~Month+Sales,data=hw4_data1,main="Scatterplot Matrix")
```



At the end of the years (Sep, Oct, Nov and Dec) sales has been increased, but in the Jan and Feb sales decreases.

Month Numerical (model A)

```
fit1 <- lm(Sales~ Month+Year,data=hw4_data1)
> summary(fit1)$adj.r
[1] 0.4321569
Call:
lm(formula = Sales ~ Month + Year, data = hw4_data1)

Residuals:
    Min      1Q  Median      3Q     Max 
-452.05 -157.91  -23.04   75.71  766.25 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2259.840  20525.603  -0.110    0.913    
Month          58.121      6.817   8.526  3.1e-13 ***
Year            1.225     10.296   0.119    0.906    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 225 on 91 degrees of freedom
Multiple R-squared:  0.4444,	Adjusted R-squared:  0.4322 
F-statistic: 36.39 on 2 and 91 DF,  p-value: 2.442e-12
```

Month Categorical (model B)

```
hw4_data1$Month<- as.factor(hw4_data1$Month)
> fit2 <- lm(Sales~ Month+Year,data=hw4_data1)
> summary(fit2)$adj.r
[1] 0.9581081
Call:
lm(formula = Sales ~ Month + Year, data = hw4_data1)

Residuals:
     Min       1Q   Median       3Q      Max 
-254.298  -31.686   -8.024   30.981  167.952 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -10368.909   5585.256  -1.856 0.067021 .  
Month2         -14.125     30.563  -0.462 0.645206    
Month3          82.250     30.563   2.691 0.008647 ** 
Month4         107.000     30.563   3.501 0.000757 ***
Month5          99.000     30.563   3.239 0.001739 ** 
Month6          95.750     30.563   3.133 0.002410 ** 
Month7          31.250     30.563   1.022 0.309600    
Month8          95.875     30.563   3.137 0.002380 ** 
Month9         174.125     30.563   5.697 1.90e-07 ***
Month10        207.375     30.563   6.785 1.75e-09 ***
Month11        382.549     31.667  12.080  < 2e-16 ***
Month12       1159.407     31.667  36.613  < 2e-16 ***
Year             5.384      2.802   1.922 0.058142 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 61.13 on 81 degrees of freedom
Multiple R-squared:  0.9635,	Adjusted R-squared:  0.9581 
F-statistic: 178.3 on 12 and 81 DF,  p-value: < 2.2e-16
```

Adjusted R-squared for model B is higher, it means model B is better.

b)

Coefficient of Year is Positive it shows with increasing year, sales will be increased. According to the seasons at the end of season three and season 4(Months September, October, November and December) sales has been increased significantly. In the first season (January and February) sales has decreased.

For January:(month 1) Sale= -10368.909+5.384*Year

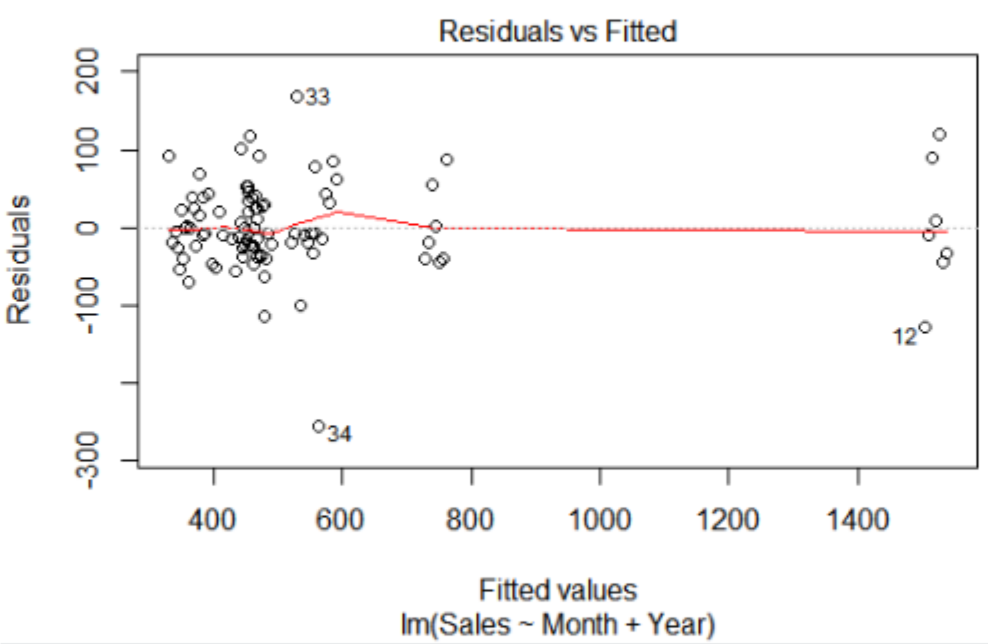
February: (month 2) **Sale=** -10383.03+5.384*Year
March: (month 3) **Sale=** -10286.66+5.384*Year
April: (month 4) **Sale=** -10261.91+5.384*Year
May: (month 5) **Sale=** -10269.91+5.384*Year
June: (month 6) **Sale=** -10273.16+5.384*Year
July: (month 7) **Sale=** -10337.66+5.384*Year
August: (month 8) **Sale=** -10273.03+5.384*Year
September: (month 9) **Sale=** -10194.78+5.384*Year
October: (month 10) **Sale=** -10161.53+5.384*Year

November: (month 11) **Sale=** `-9986.36+5.384*Year`
December: (month 12) **Sale=** `-9209.502+5.384*Year`

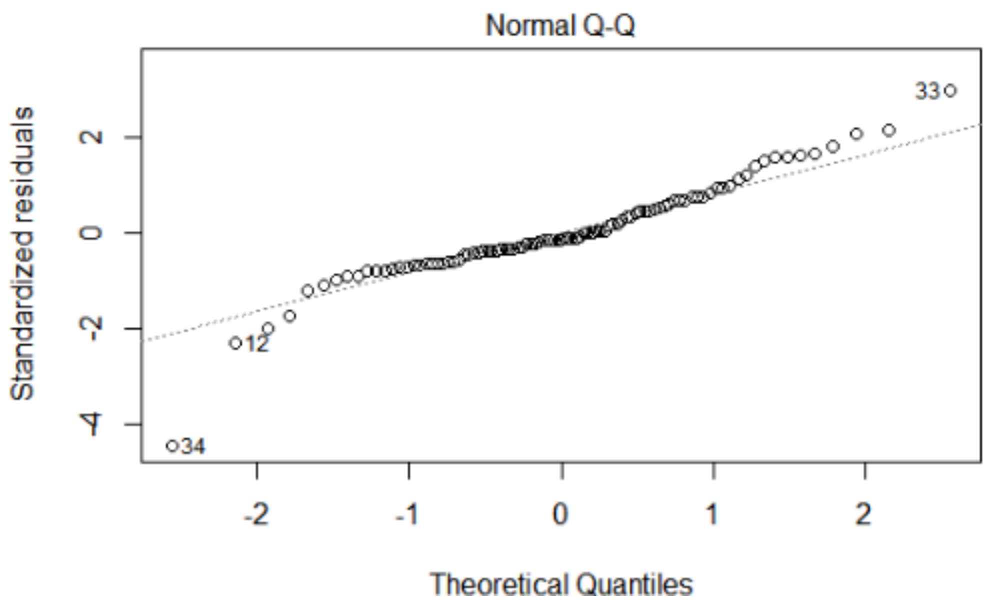
assumptions that made by our predictions are Linear relationship, normality, there is not multicollinearity, there is **not auto-correlation** and equality of variances. The errors are independent.

c)
```
dwtest(fit2,alternative="two.sided")

        Durbin-Watson test

data:  fit2
DW = 2.4509, p-value = 0.03902
alternative hypothesis: true autocorrelation is not 0
```

We can reject null hypothesis => there is auto correlation.

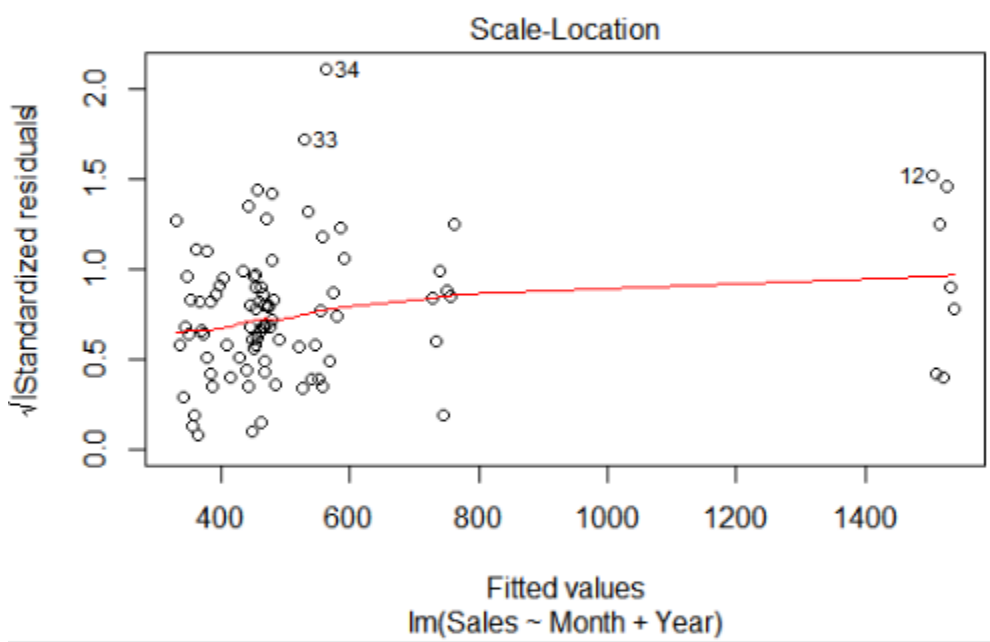


Residuals vs Fitted: **we can't accept linearity assumption.**




```
Shapiro-Wilk normality test

data:  residuals(fit2)
W = 0.93187, p-value = 0.0001059
```

The QQ plot of residuals can be used to check the normality assumption. The normal probability plot of residuals should approximately follow a straight line. In this problem, the pattern is non-linear, so plot give evidence against assume normality. It means we reject normality assumption. However, we can double check by using Shapiro test that it shows we reject normality assumption. Therefore, **Normality assumption is rejected.**



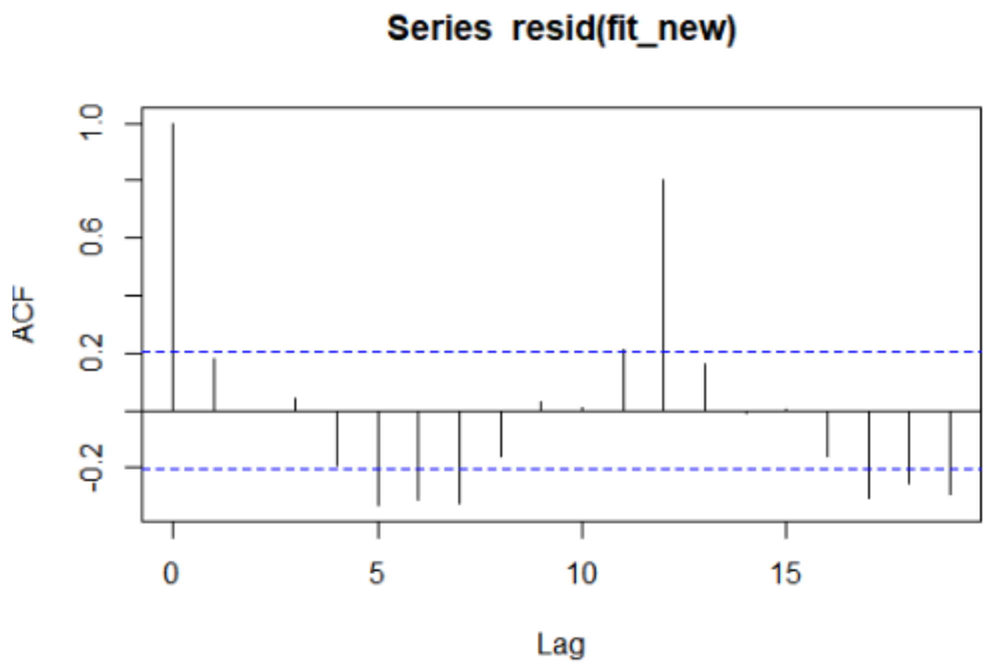
```
studentized Breusch-Pagan test

data:  fit2
BP = 13.9, df = 12, p-value = 0.3071
```

Equality of variance: Regarding to above plots for each value of X, the distribution of residuals has the same variance. This means that the level of error in the model is approximately the same regardless of the value of the predictor variable. Also, we can use Breusch-Pagan Test that **it proves again the assumption of equality of variances is accepted.**

D)

```
fit2 <- lm(Sales~ Month+Year,data=hw4_data1)
rho_hat_dw = (1-dwtest(fit2)$statistic/2)
rho_hat_dw
     DW
-0.225457


num_obs=94
> e_t = resid(fit2)
> e_t_1 = e_t[-num_obs]
> cor(e_t_1,e_t[-1])
[1] -0.2348575

y_t = hw4_data1$Sales[-1]
y_t_1 = hw4_data1$Sales[-num_obs]
y_new = y_t - rho_hat_dw*y_t_1

x_t = hw4_data1$Year[-1]
x_t_1 = hw4_data1$Year[-num_obs]
x_new = x_t - rho_hat_dw*x_t_1
#hw4_data1$Month<- as.numeric(hw4_data1$Month)
x1_t = hw4_data1$Month[-1]
x1_t_1 = hw4_data1$Month[-num_obs]
x1_new = x1_t - rho_hat_dw*x1_t_1
fit_new <- lm(y_new~x_new+x1_new,data=hw4_data1)
acf(resid(fit_new))
dwtest(fit_new,alternative="two.sided")
AIC(fit2)
AIC(fit_new)


Durbin-Watson test

data:  fit_new
DW = 1.6353, p-value = 0.05427
alternative hypothesis: true autocorrelation is not 0
 AIC(fit2)
  [1] 1054.002
 AIC(fit_new)
 [1] 1264.77
```

Based on AIC model B is better.



The error independence assumption is rejected.