Coupon Dataset Analysis

Abolfazl Joukar

2024-11-17

1. Dataset Description

A study of the effectiveness of coupons offering a price reduction on a given product. Suppose 1,000 homes were selected at random and (at random) 200 of them were mailed a coupon good for 5% reduction in price of the product, 200 were mailed 10% coupons, 200 were mailed 15% coupons, 200 were mailed 20% coupons and 200 were mailed 30% coupons.The following displays details of the dataset.

```
suppressWarnings({library(kableExtra)})
Variable<-c("ID","coupon","redemption")
Description<-c("Sample index number",
               "Percentage of price reduction",
               "Status of redemption: 1 = redeemed, 0 = not redeemed")
details_dataset<-data.frame(Variable,Description)
colnames(details_dataset)<-c("Variable","Description of variable")
details_data<-knitr::kable(details_dataset,caption="Details of the dataset",
                           booktabs=T,label="kabletable",digits=4)%>%
kable_styling(bootstrap_options=c("striped","hover","condensed","responsive"))
kable_classic_2(details_data,full_width=F,html_font="Cambria")
```

Table 1: Details of the dataset

Variable	Description of variable
ID	Sample index number
coupon	Percentage of price reduction
redemption	Status of redemption: 1 = redeemed, 0 = not redeemed

The variable **ID** is the household identification number, variable **coupon** is the price reduction in percentage when the coupon is redeemed and variable **redemption** a binary variable taking the value **one** if the coupon was used during the six-month period of the study and taking the value **zero** if the coupon was not redeemed during the six month period. The following show a partial data provided in the dataset:

```
Data<-read.table(file="D:/coupon.txt",header=F)
colnames(Data)<-c("ID","coupon","redemption")
Data_table<-knitr::kable(head(Data),caption="Coupon Data",
                           booktabs=T,label="kabletable",digits=4)%>%
kable_styling(bootstrap_options=c("striped","hover","condensed","responsive"))
kable_classic_2(Data_table,full_width=F,html_font="Cambria")
```

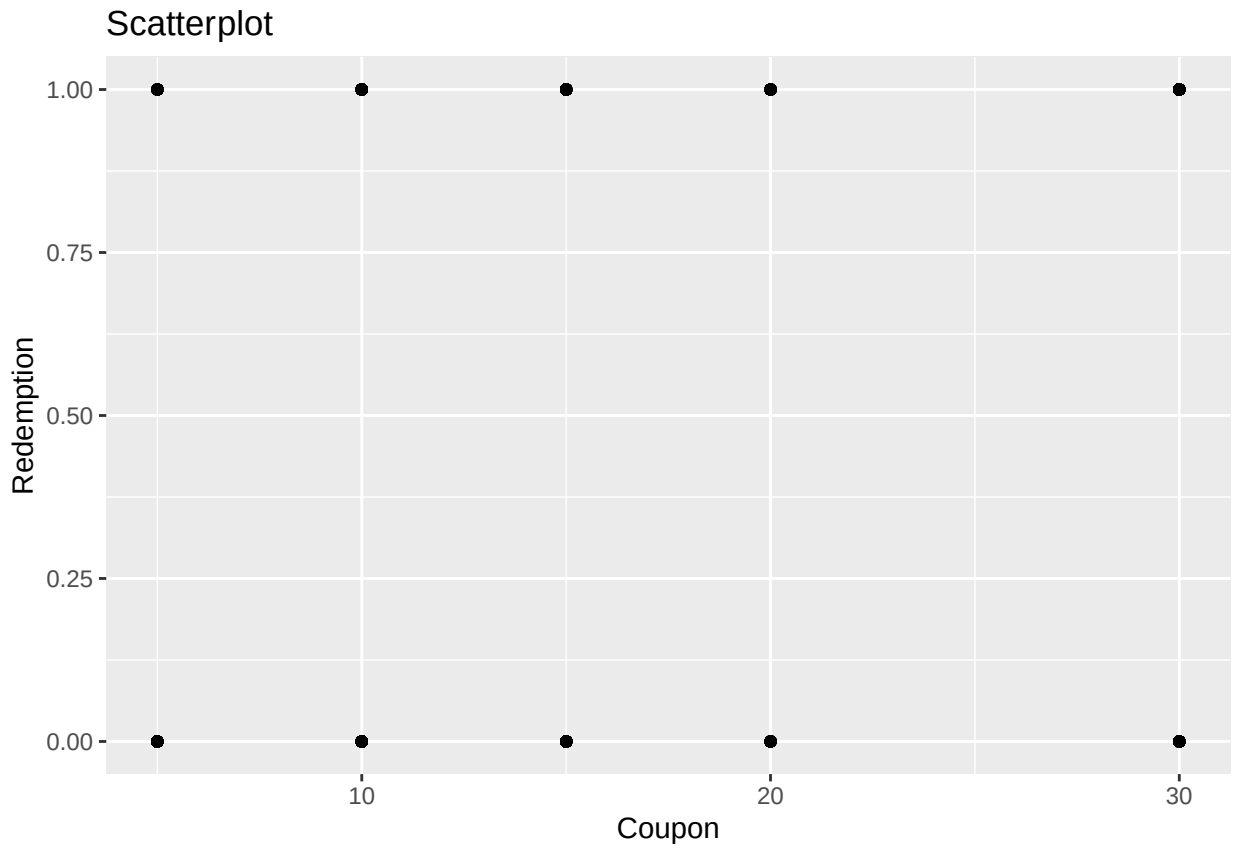
Table 2: Coupon Data

ID	coupon	redemption

1	5	0
2	5	0
3	5	0
4	5	0
5	5	0
6	5	0

In data analysis and statistics, scatter plots are used to display the relationship between two numerical variables. To investigate the relationship between variables coupon (X, as the explanatory variable) and redemption (Y, as the response variable), we draw the scatter plots.

```
suppressWarnings({library(ggplot2)})
ggplot(Data,aes(x=coupon,y=redemption))+geom_point()+
        xlab("Coupon")+ylab("Redemption")+ggtitle("Scatterplot")
```



The scatter plots do not display specific relationship between two variables.So, we transform the data to give it a relationship.So, we transform the data to investigate the relationship between variables.The new data is shown in Table 2.Then, we fit four different models to investigate the relationship between variables coupon (X, as the explanatory variable) and redemption (Y, as the response variable) based on new data and choose the best model using Chi-Squared goodness-of-fit tests.

```
Table<-table(Data[,-1])
coupon_new<-c(5,10,15,20,30)
redemption_new<-rep(c(0,1),5)
count<-c(-1,-1)
for(i in 1:nrow(Table)){
count<-c(count,c(Table[i,]))
```

```
}
count<-count[-c(1,2)]
rate<-count/200
Data2<-data.frame(coupon_new,count[seq(1,10,2)],rate[seq(1,10,2)]
                  ,count[seq(2,10,2)],rate[seq(2,10,2)])
colnames(Data2)<-c("coupon","count0","Rate0","count1","Rate1")
Data_table2<-knitr::kable(Data2,caption="The transformed coupon data",
                          booktabs=T,label="kabletable",digits=4)%>%
kable_styling(bootstrap_options=c("striped","hover","condensed","responsive"))
kable_classic_2(Data_table2,full_width=F,html_font="Cambria")
```

Table 3: The transformed coupon data

coupon	count0	Rate0	count1	Rate1
5	168	0.840	32	0.160
10	149	0.745	51	0.255
15	130	0.650	70	0.350
20	97	0.485	103	0.515
30	52	0.260	148	0.740

2. Linear Model

In this section, we build a linear regression model with **rate** being the dependent variable and **coupon** being the independent (explanatory) variables. Then, we predict the **rate** or **redemption** using this model. Results are presented in Table 4.

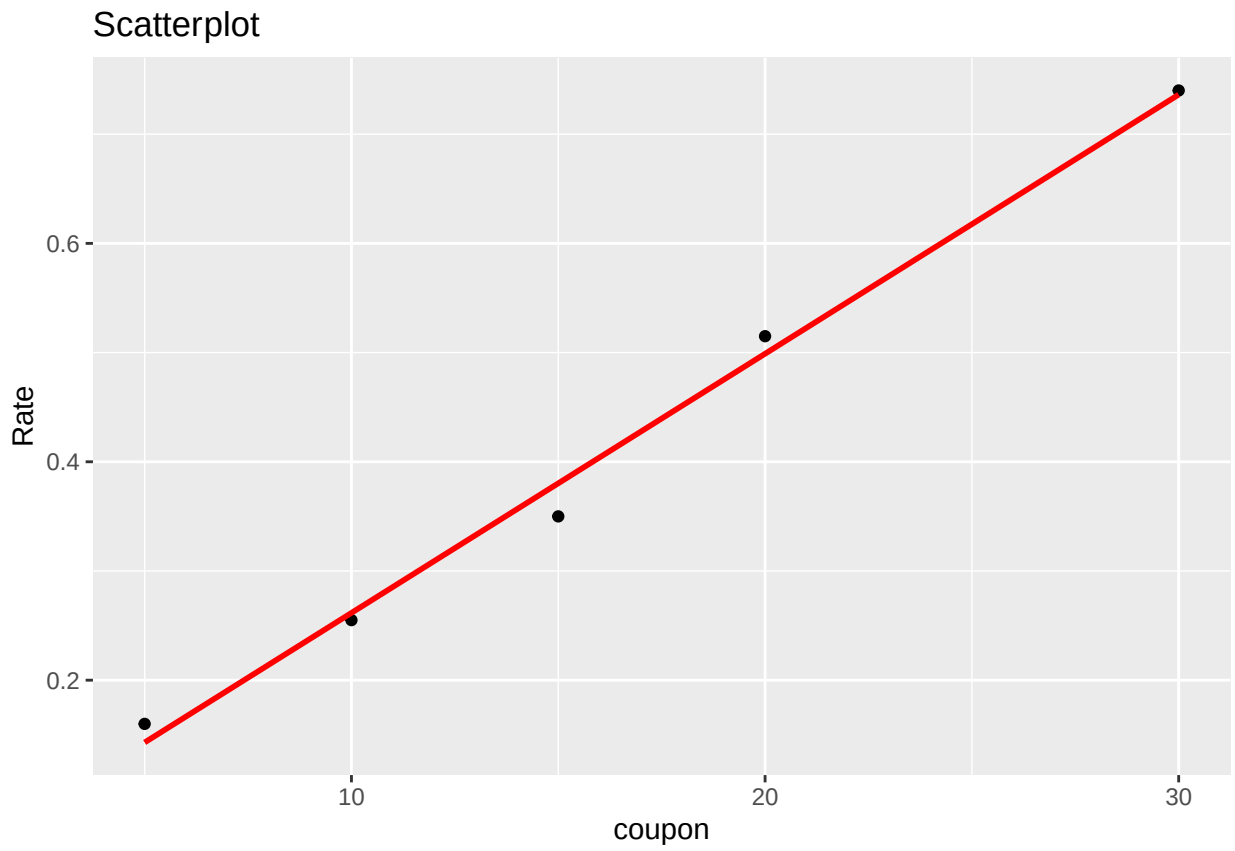
```
suppressWarnings({library(MASS)})

test_func<-function(observed,expected,df){
X2<-sum(((observed-expected)^2)/expected)
p_value1<-pchisq(X2,df,lower.tail=F)
result<-matrix(c(X2,p_value1),1,2)
colnames(result)<-c("Statistic","P-value")
rownames(result)<-c("X2")
return(list(result=result))
}
#-------------------------------------------
round_func<-function(x){
x_floor<-x-floor(x)
x_ceiling<-ceiling(x)-x
x_round<-c()
for(i in 1:length(x)){
if(x_floor[i]<=x_ceiling[i]){
x_round[i]<-floor(x[i])
}else{
x_round[i]<-ceiling(x[i])
}
}
return(list(x_round=x_round))
}
#-------------------------------------------
lin_model<-lm(Rate1~coupon,data=Data2)
Rate1_linear<-lin_model$fitted
count1_linear<-round_func(Rate1_linear*200)
Data2<-data.frame(Data2,Rate1_linear,count1_linear)
colnames(Data2)<-c("coupon","count0","Rate0","count1","Rate1",
                   "Rate1_linear","count1_linear")
result_linear<-test_func(Data2$count1,Data2$count1_linear,4)$result
#-------------------------------------------------
Data_table2<-knitr::kable(Data2,caption="Prediction the rate and redemption using the linear model",
                          booktabs=T,label="kabletable",digits=4)%>%
kable_styling(bootstrap_options=c("striped","hover","condensed","responsive"))
kable_classic_2(Data_table2,full_width=F,html_font="Cambria")
```

Table 4: Prediction the rate and redemption using the linear model

coupon	count0	Rate0	count1	Rate1	Rate1_linear	count1_linear
5	168	0.840	32	0.160	0.1430	29
10	149	0.745	51	0.255	0.2616	52
15	130	0.650	70	0.350	0.3803	76
20	97	0.485	103	0.515	0.4989	100
30	52	0.260	148	0.740	0.7362	147

```r
suppressWarnings({library(ggplot2)})
ggplot(Data2,aes(x=coupon,y=Rate1))+geom_point()+
        geom_smooth(method=lm,color="red",se=FALSE)+
        xlab("coupon")+ylab("Rate")+ggtitle("Scatterplot")
```




```r
suppressWarnings({library(gridExtra)})
suppressWarnings({library(lmtest)})
data2<-data.frame(Residuals=lin_model$resid,Fitted=fitted(lin_model))
Plot1<-ggplot(data2,aes(x=Fitted,y=Residuals))+geom_point(color="blue")+
            geom_hline(yintercept=0,color="red")+
            ggtitle("Residuals ~ Fitted values")
Plot2<-ggplot(data2,aes(sample=lin_model$resid))+
            geom_qq(distribution=qnorm,col="blue")+geom_qq_line(distribution=qnorm,col="red")+
            ggtitle("QQ-plot")+xlab("Fitted Normal Quantiles")+ylab("Sample quantiles")
#-----------------------------------------
bptest(lin_model,~coupon,data=Data2,studentize=F)
```

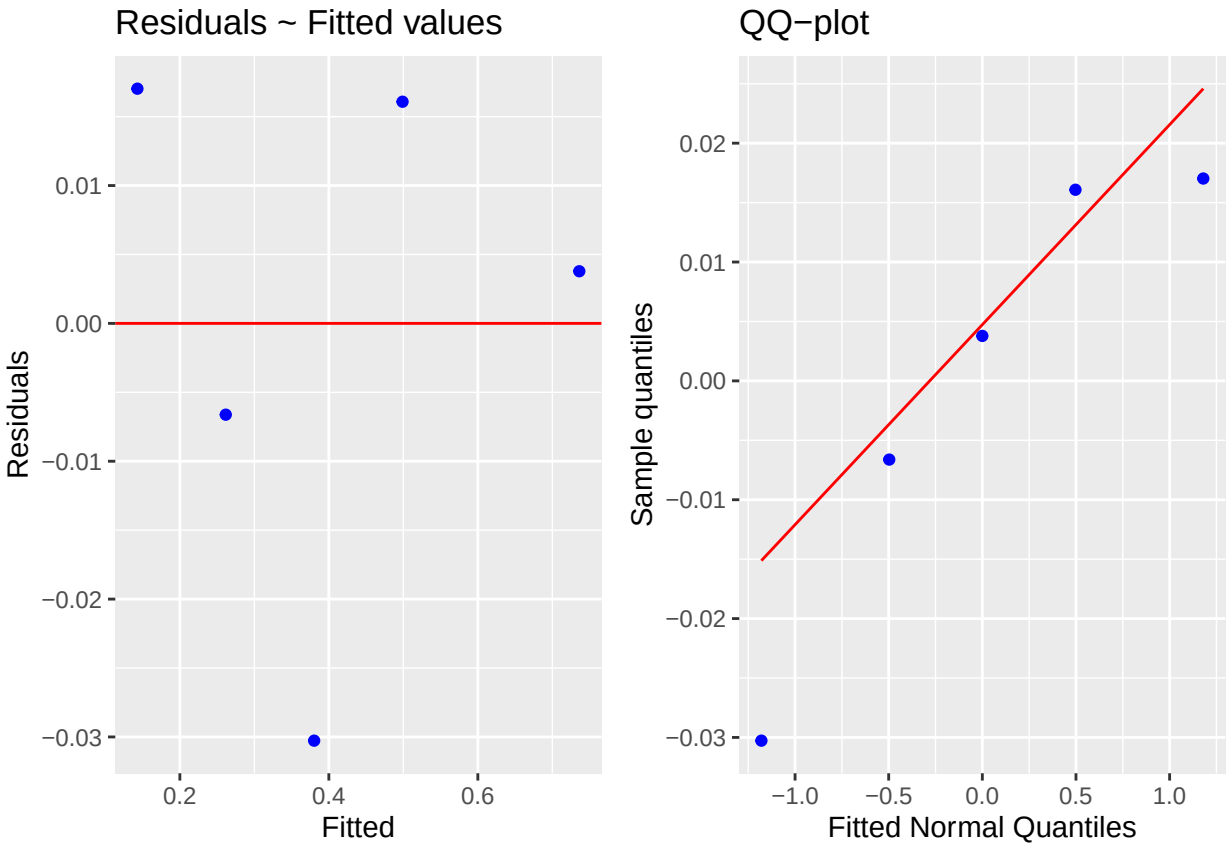
```
##
##  Breusch-Pagan test
##
## data:  lin_model
## BP = 0.14303, df = 1, p-value = 0.7053
```

```r
shapiro.test(residuals(lin_model))
```

```
##
##  Shapiro-Wilk normality test
```

```
##
## data:  residuals(lin_model)
## W = 0.89478, p-value = 0.3817
```

```
#-----------------------------------------
grid.arrange(Plot1,Plot2,ncol=2)
```



Furthermore, we check the normality assumption. Let, that at any given value of x, the population of error term values has a normal distribution.To do this, we have used the Shapiro- Wilk test and the Normal Q-Q plot.The result of Shapiro- Wilk test and Normal Q-Q plot show that the distribution of errors is normal.

We also use the Breusch-Pagan test to check the assumption that the error variance is constant.Since the p-value of this test is greater than the $\alpha = 0.05$, then the null hypothesis is accepted and therefore the error variances are constant.

3. Logit Linear Model

In this section, we fit a logit linear regression model with **rate** being the dependent variable and **coupon** being the independent (explanatory) variables.Then, we also predict the **rate** or **redemption** using this model. Results are given in Table 5.

```
logit_linear_model<-lm(log(Rate1/(1-Rate1))~coupon,data=Data2)
prop<-exp(logit_linear_model$fitted)/(1+exp(logit_linear_model$fitted))
count_logit_linear<-round_func(prop*200)
Data2<-Data2[,-c(6,7)]
Data2<-data.frame(Data2,prop,count_logit_linear)
colnames(Data2)<-c("coupon","count0","Rate0","count1","Rate1",
                   "Rate1_logit_linear","count1_logit_linear")
result_logit_linear<-test_func(Data2$count1,Data2$count1_logit_linear,4)$result
#-------------------------------------------------
Data_table2<-knitr::kable(Data2,caption="Prediction the rate and redemption using the logit linear model
                          booktabs=T,label="kabletable",digits=4)%>%
kable_styling(bootstrap_options=c("striped","hover","condensed","responsive"))
kable_classic_2(Data_table2,full_width=F,html_font="Cambria")
```

Table 5: Prediction the rate and redemption using the logit linear model

coupon	count0	Rate0	count1	Rate1	Rate1_logit_linear	count1_logit_linear
5	168	0.840	32	0.160	0.1620	32
10	149	0.745	51	0.255	0.2497	50
15	130	0.650	70	0.350	0.3642	73
20	97	0.485	103	0.515	0.4964	99
30	52	0.260	148	0.740	0.7449	149

4. Logistic Model

Logistic regression model is used to model dichotomous outcome variables. In this section, we use logistic model to predict the **rate** or **redemption**. Results are reported in Table 6.

```
logit_model<-glm(redemption~coupon,family=binomial(logit),data=Data)
Rate1_logit<-unique(logit_model$fitted)
count1_logit<-round_func(Rate1_logit*200)
Data2<-Data2[,-c(6,7)]
Data2<-data.frame(Data2,Rate1_logit,count1_logit)
colnames(Data2)<-c("coupon","count0","Rate0","count1","Rate1",
                   "Rate1_logit","count1_logit")
result_logit<-test_func(Data2$count1,Data2$count1_logit,4)$result
#-------------------------------------------------
Data_table2<-knitr::kable(Data2,caption="Prediction the rate and redemption using the logistic model",
                          booktabs=T,label="kabletable",digits=4)%>%
kable_styling(bootstrap_options=c("striped","hover","condensed","responsive"))
kable_classic_2(Data_table2,full_width=F,html_font="Cambria")
```

Table 6: Prediction the rate and redemption using the logistic model

coupon	count0	Rate0	count1	Rate1	Rate1_logit	count1_logit
5	168	0.840	32	0.160	0.1622	32
10	149	0.745	51	0.255	0.2501	50
15	130	0.650	70	0.350	0.3648	73
20	97	0.485	103	0.515	0.4972	99
30	52	0.260	148	0.740	0.7457	149

5. Poisson Regression Model

A poisson regression model is used to model count data and model response variables (Y-values) that are counts.In this section of the project, we predict the **rate** or **redemption** using poisson model which results are detailed in Table 7.

```
poiss_model<-glm(count1~coupon,family=poisson,data=Data2)
count1_poiss<-round_func(poiss_model$fitted)
Rate1_poiss<-poiss_model$fitted/200
Data2<-Data2[,-c(6,7)]
Data2<-data.frame(Data2,Rate1_poiss,count1_poiss)
colnames(Data2)<-c("coupon","count0","Rate0","count1","Rate1",
                   "Rate1_poiss","count1_poiss")
result_poiss<-test_func(Data2$count1,Data2$count1_poiss,4)$result
#-------------------------------------------------
Data_table2<-knitr::kable(Data2,caption="Prediction the rate and redemption using the poisson model",
                          booktabs=T,label="kabletable",digits=4)%>%
kable_styling(bootstrap_options=c("striped","hover","condensed","responsive"))
kable_classic_2(Data_table2,full_width=F,html_font="Cambria")
```

Table 7: Prediction the rate and redemption using the poisson model

coupon	count0	Rate0	count1	Rate1	Rate1_poiss	count1_poiss
5	168	0.840	32	0.160	0.1932	39
10	149	0.745	51	0.255	0.2557	51
15	130	0.650	70	0.350	0.3384	68
20	97	0.485	103	0.515	0.4479	90
30	52	0.260	148	0.740	0.7847	157

6. Selection the best model

In this section, we use Chi-Squared goodness-of-fit tests to select the best model.

```
result_n<-rbind(result_linear,result_logit_linear,
                result_logit,result_poiss)
row.names(result_n)<-c("Linear model","Logit linear model",
                       "Logistic model","Poisson model")
result_table<-knitr::kable(result_n,caption="Goodness-of-Fit Tests for different model",
                           booktabs=T,label="kabletable",digits=4)%>%
kable_styling(bootstrap_options=c("striped","hover","condensed","responsive"))
kable_classic_2(result_table,full_width=F,html_font="Cambria")
```

Table 8: Goodness-of-Fit Tests for different model

	Statistic	P-value
Linear model	0.9001	0.9246
Logit linear model	0.3116	0.9891
Logistic model	0.3116	0.9891
Poisson model	3.7089	0.4468

It is observed from Table 8 that the logistic model and logit linear model have largest values of P-value of Chi-Squared goodness-of-fit tests. On the other hand, the redemption variable is binary variable and so the logistic model is the best model to describe the relationship between variables **coupon** and **redemption**.

7. Confusion matrix and Odds ratio for the best Model

In this section of the project, we compute the confusion matrix and odds ratio for the logistic regression model.

```
set.seed(1369)
suppressWarnings({library(caTools)})
split<-sample.split(Data$ID,SplitRatio=0.7)
Data_new<-data.frame(Data,split)
Data_train<-Data_new%>%dplyr::filter(Data_new$split=="TRUE")
Data_test<-Data_new%>%dplyr::filter(Data_new$split=="FALSE")

Data_train<-Data_train[,-c(1,4)]
Data_test<-Data_test[,-c(1,4)]

fit_logis_train<-glm(redemption~coupon,family="binomial",data=Data_train)
#-----------------------------------------------------------------------
models.probs_test<-predict(fit_logis_train,Data_test,type="response")
models.pred_test<-rep(0,length(models.probs_test))
models.pred_test[models.probs_test>0.5] <- 1
Table_model_performance2<-table(models.pred_test,Data_test$redemption,
                           dnn = c("Predicted Status","Observed Status"))
cat("\n Confusion matrix \n")
```

```
##
##  Confusion matrix
```

```
Table_model_performance2
```

```
##                  Observed Status
## Predicted Status   0   1
##                 0 134  43
##                 1  51  72
```

```
cat("\n Odds ratio \n")
```

```
##
##  Odds ratio
```

```
odds_ratio<-exp(coef(logit_model))
round(odds_ratio,2)
```

```
## (Intercept)      coupon
##        0.11        1.11
```

```
#------------------
```

Observation from the confusion matrix, the information that we can get is that, with total of 185 (redemption=0) the model has predicted 134 correctly and 51 wrongly. Similarly, total of 115 (redemption=1) the model has predicted 72 correctly and 43 wrongly. Percentage of correct predictions is $\frac{(134+72)}{(134+51+43+72)} \times 100 =$ 69%, which tells us 69% of the responses are predicted correctly.

The odds ratio of 1.11 for coupon variable implies that 1 unit increase in coupon variable , the odds of that the redemption increases by a factor of 1.11.