Questão 1.

$$\begin{aligned}
p(x) &= \frac{exp\{\beta_0 + \beta_1 x\}}{1 + exp\{\beta_0 + \beta_1 x\}} \\
p(x)(1 + exp\{\beta_0 + \beta_1 x\}) &= exp\{\beta_0 + \beta_1 x\} \\
p(x) &= exp\{\beta_0 + \beta_1 x\} - p(x)exp\{\beta_0 + \beta_1 x\}) \\
p(x) &= (1 - p(x))exp\{\beta_0 + \beta_1 x\}) \\
\frac{p(x)}{1 - p(x)} &= exp\{\beta_0 + \beta_1 x\} \\
\log\left(\frac{p(x)}{1 - p(x)}\right) &= \beta_0 + \beta_1 x
\end{aligned}$$

Questão 2.

Temos um conjunto de dados de estudantes de uma disciplina de pós-graduação com as variáveis $X1$: horas de estudo, $X2$: pontuação no exame de admissão e $Y = 1$, se obteve A na disciplina e $Y = 0$, caso contrário. Foi ajustado um modelo de regressão logística e as estimativas dos coeficientes são $b0 = -6$, $b1 = 0,05$ e $b2 = 1$.

a) A probabilidade estimada de um estudante que estudou 40 horas e possui X2=3.5 obter uma nota A na disciplina é,

$$p(y = 1|x_1 = 40, x_2 = 3.5) = \frac{exp\{-6 + 0,05 * 40 + 3,5\}}{1 + exp\{-6 + 0,05 * 40 + 3,5\}} = 37,75\%$$

b)

$$\begin{aligned}
p(y = 1|x_1, x_2 = 3.5) &= 50\% \\
\frac{exp\{-2,5 + 0,05 * x_1\}}{1 + exp\{-2,5 + 0,05 * x_1\}} &= 50\% \\
exp\{-2,5 + 0,05 * x_1\} &= \frac{1}{2}(1 + exp\{-2,5 + 0,05 * x_1\}) \\
exp\{-2,5 + 0,05 * x_1\} &= 1 \\
-2,5 + 0,05 * x_1 &= 0 \\
0,05 * x_1 &= 2,5 \\
x_1 &= 2,5/0,05 \\
x_1 &= 50
\end{aligned}$$

Assim, um estudante com a mesma pontuação considerada em (a) deve estudar 50 horas para ter uma probabilidade de 0,5 de obter uma nota A na disciplina.

Questão 3.

a) Em média, a fração de pessoas com uma chance de 0,37 de não pagar o seu cartão de crédito (inadimplente) não irão de fato pagar é de

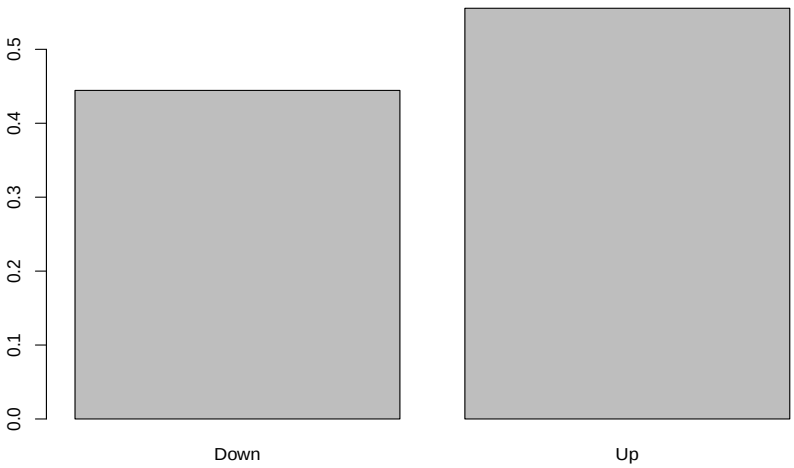
$$\begin{aligned} \frac{p}{1-p} &= 0,37 \\ p &= (1-p)0,37 \\ 1,37p &= 0,37 \\ p &= 0,37/1,37 \\ p &= 27\% \end{aligned}$$

b) Supondo que um indivíduo tenha uma probabilidade de 0,16 de ser inadimplente, a chance dele não pagar o cartão é de,

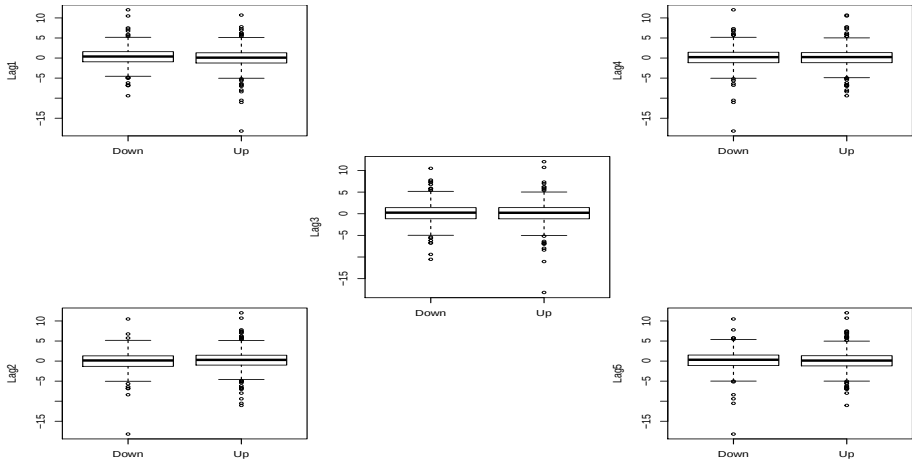
$$\begin{aligned} Chance &= \frac{p}{1-p} \\ &= \frac{0,16}{1-0,16} \\ &= \frac{0,16}{0,84} \\ &= 19\% \end{aligned}$$

Questão 4.

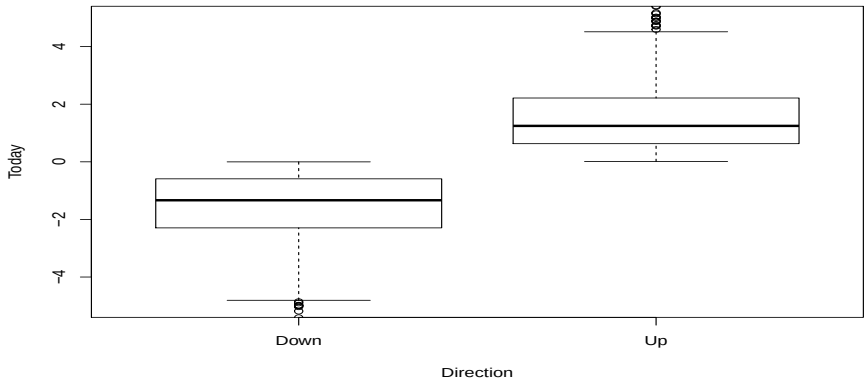
a) No gráfico abaixo notamos que existe uma proporção maior de *Up*.



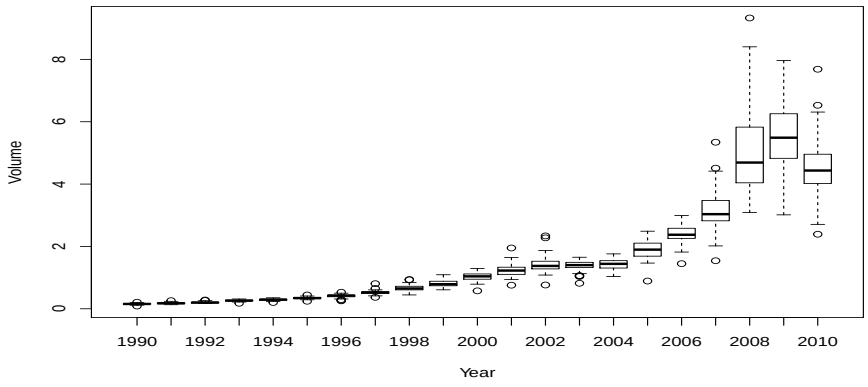
Notamos que as distribuições das variáveis *lags* versus a variável *Direction* são similares.



No gráfico da variável *Today* versus *Direction*, notamos que a categoria *Down* apresentou uma assimetria negativa e a categoria *Up* apresentou uma assimetria positiva.



No gráfico da variável *Volume* versus *Year*, notamos que houve um aumento na dispersão e nos valores assumidos pela variável *Volume* ao longo do intervalo temporal observado.



b) Apresentamos agora o ajuste do modelo

```
glm( Direction ~ Lag1 + Lag2 +Lag3 +Lag4 + Lag5 + Volume, data = Weekly, family = binomial )
```

	Estimativa	Erro padrão	Z	Pr(>|z|)
Intercepto	0,2669	0,0859	3,11	0,0019
Lag1	-0,0413	0,0264	-1,56	0,1181
Lag2	0,0584	0,0269	2,18	0,0296
Lag3	-0,0161	0,0267	-0,60	0,5469
Lag4	-0,0278	0,0265	-1,05	0,2937
Lag5	-0,0145	0,0264	-0,55	0,5833
Volume	-0,0227	0,0369	-0,62	0,5377

Apenas o intercepto e o preditor *Lag2* são significativos para o modelo ao nível de significância de 10%, pois o valor p dos testes realizados estão abaixo de 10%.

* $H_0 : B_0 = 0 \quad vs \quad H_1 : B_0 \neq 0$, rejeitamos H_0 ao nível significância de 10%. Portanto, o intercepto é significativo para o modelo.
* $H_0 : B_1 = 0 \quad vs \quad H_1 : B_1 \neq 0$, não rejeitamos H_0 ao nível significância de 10%. Portanto, *Lag1* não é significativo para o modelo.
* $H_0 : B_2 = 0 \quad vs \quad H_1 : B_2 \neq 0$, rejeitamos H_0 ao nível significância de 10%. Portanto, *Lag2* é significativo para o modelo.
* $H_0 : B_3 = 0 \quad vs \quad H_1 : B_3 \neq 0$, não rejeitamos H_0 ao nível significância de 10%. Portanto, *Lag3* não é significativo para o modelo.
* $H_0 : B_4 = 0 \quad vs \quad H_1 : B_4 \neq 0$, não rejeitamos H_0 ao nível significância de 10%. Portanto, *Lag4* não é significativo para o modelo.
* $H_0 : B_5 = 0 \quad vs \quad H_1 : B_5 \neq 0$, não rejeitamos H_0 ao nível significância de 10%. Portanto, *Lag5* não é significativo para o modelo.
* $H_0 : B_6 = 0 \quad vs \quad H_1 : B_6 \neq 0$, não rejeitamos H_0 ao nível significância de 10%. Portanto, *Volume* não é significativo para o modelo.

c) Apresentamos a matriz de confusão.

Classe verdadeira	Classe predita		
	Down	Up	Total
Down	54	430	484
Up	48	557	605
Total	102	987	1089

Considerando "Up"como a classe positiva, temos que a sensibilidade é dada por $\frac{557}{605} = 92,1\%$, assim o modelo tem uma taxa alta de acerto de classificar como "Up"quando realmente é "Up". Já a especifi-

cidade é dada por $\frac{54}{484} = 11,2\%$, ou seja, o modelo tem uma taxa baixa de acerto de classificar como "Down"quando realmente é "Down". O total de predições corretas é $\frac{54+557}{1089} = 56,1\%$.

d) Usando como dados de treinamento o período de 1990 à 2008, ajustamos o modelo de regressão logístico com apenas a co-variável *Lag2* e o intercepto.

	Estimativa	Erro padrão	Z	Pr(>|z|)
Intercepto	0,2033	0,0643	3,16	0,0016
Lag2	0,0581	0,0287	2,02	0,0430

Portanto, as duas componentes do modelo são significativas usando um nível de significância de 10%.

Tendo como amostra de teste os anos seguintes (2009 e 2010), obtemos a seguinte *confusion matrix*.

Classe verdadeira	Classe predita		
	Down	Up	Total
Down	9	34	43
Up	5	56	61
Total	14	90	104

Considerando "Up"como a classe positiva, temos que a sensibilidade é dada por $\frac{56}{61} = 92\%$, assim o modelo tem uma taxa alta de acerto de classificar como "Up"quando realmente é "Up". Já a especificidade é dada por $\frac{9}{43} = 21\%$, ou seja, o modelo tem uma taxa baixa de acerto de classificar como "Down"quando realmente é "Down". O total de predições corretas nos dados de teste é $\frac{9+56}{104} = 62,5\%$.

Questão 5.

a) Da tabela podemos obter intervalos de 95% de confiança para os coeficientes estimados dos preditores

	Estimativa	Erro padrão	Z	Pr(>|z|)
Intercepto	-10,7495	0,3692	-29,12	< 0,001
studentYes	-0,7149	0,1475	-4,85	<0,001
balance	0,0057	0,0002	24,75	<0,001

e estes intervalos são:

β_1 associado ao studentYes

$$IC(\beta_1, 95\%) = [-0,7149 - 1,96 \times 0,1475 \; ; \; -0,7149 + 1,96 \times 0,1475] = [-1,004 \; ; \; -0,4258].$$

β_2 associado ao balance

$$IC(\beta_2, 95\%) = [0,0057 - 1,96 \times 0,0002 \; ; \; 0,0057 + 1,96 \times 0,0002] = [0,005308 \; ; \; 0,006092].$$

b) A probabilidade de um estudante ser inadimplente, considerando que ele tenha um balance igual a 2.000 reais é de,

$$p(default = Yes|student = Yes(1), balance = 2000) = \frac{\exp\{\widehat{\beta}_0 + \widehat{\beta}_1 + \widehat{\beta}_2 \times 2000\}}{1 + \exp\{\widehat{\beta}_0 + \widehat{\beta}_1 + \widehat{\beta}_2 \times 2000\}} = 50,3\%.$$

Para um não estudante, a probabilidade é

$$p(default = Yes|student = No(0), balance = 2000) = \frac{\exp\{\widehat{\beta}_0 + \widehat{\beta}_2 \times 2000\}}{1 + \exp\{\widehat{\beta}_0 + \widehat{\beta}_2 \times 2000\}} = 67,4\%.$$

Portanto, apesar de ambas as probabilidades serem altas, o mais arriscado é oferecer crédito para o não estudante, pois temos uma probabilidade maior de não pagar a dívida.

Questão 6. Após fazer a transformação na variável *suburb*, dividimos os dados em treinamento e teste com as respectivas proporções 80% e 20%. O ajuste do modelo com os dados de treinamento é dado por,

	Estimativa	Erro padrão	Z	Pr(>|z|)
Intercepto	-38,6301	7,7258	-5,00	<0,001
zn	-0,0956	0,0412	-2,32	0,0202
indus	-0,0750	0,0489	-1,53	0,1249
chas1	0,6565	0,7742	0,85	0,3964
nox	49,8372	8,5454	5,83	<0,001
rm	-0,3643	0,7783	-0,47	0,6397
age	0,0288	0,0141	2,05	0,0407
dis	0,7762	0,2540	3,06	0,0022
rad	0,6267	0,1726	3,63	<0,001
tax	-0,0045	0,0029	-1,53	0,1261
ptratio	0,3693	0,1352	2,73	0,0063
black	-0,0120	0,0063	-1,89	0,0594
lstat	0,0926	0,0559	1,66	0,0975
medv	0,2149	0,0819	2,63	0,0087

Notamos que o intercepto e as variáveis zn, nox, age,dis, rad, ptraio, black, lstat e medv foram significativas ao nível de 10% de significância.

Retirando individualmente as variáveis mais não significativas e ajustando o modelo novamente, obtemos o seguinte modelo final:

	Estimativa	Erro padrão	Z	Pr(>|z|)
Intercepto	-36,7865	7,3431	-5,01	$<$ 0,001
zn	-0,0996	0,0382	-2,61	0,0090
nox	43,7290	7,3265	5,97	$<$ 0,001
age	0,0262	0,0120	2,18	0,0291
dis	0,7324	0,2394	3,06	0,0022
rad	0,7251	0,1628	4,45	$<$ 0,001
tax	-0,0062	0,0026	-2,38	0,0172
ptratio	0,3237	0,1229	2,63	0,0084
black	-0,0108	0,0060	-1,81	0,0698
lstat	0,0932	0,0504	1,85	0,0641
medv	0,1843	0,0540	3,41	$<$ 0,001

Para verificar a qualidade das predições do nosso modelo, usamos agora o conjunto de dados de teste para construir a matriz de confusão.

Classe verdadeira	Classe predita		
	Abaixo	Acima	Total
Abaixo	44	3	47
Acima	6	48	54
Total	50	51	101

Considerando "Acima"como a classe positiva, temos que a sensibilidade é dada por $\frac{48}{54} = 88,9\%$, assim o modelo tem uma taxa alta de acerto de classificar como "Acima"quando realmente é "Acima". Já a especificidade é dada por $\frac{44}{47} = 93,6\%$, ou seja, o modelo tem uma taxa alta de acerto de classificar como "Abaixo"quando realmente é "Abaixo". O total de predições corretas nos dados de teste é $\frac{44+48}{101} = 91,1\%$. Portanto, o modelo está fazendo boas predições.

```
require(ISLR)
require(caret)
#Q4

#a)
str(Weekly)
attach(Weekly)
barplot(prop.table(table(Direction)), beside=TRUE)

layout(mat = matrix(c(1, 0, 2, 0,3,0,4,0,5), nrow = 3,ncol = 3),heights = c(1, 1), widths = c(1, 1))

par(mar = c(2,4,0,0))
boxplot(Lag1 ~Direction, data = Weekly, ylab = "Lag1")
par(mar = c(2, 4, 0, 0))
boxplot(Lag2 ~Direction, data = Weekly, ylab = "Lag2")
par(mar = c(1, 4, 0, 0))
boxplot(Lag3 ~Direction, data = Weekly, ylab = "Lag3")
par(mar = c(2, 4, 0, 0))
boxplot(Lag4 ~Direction, data = Weekly, ylab = "Lag4")
par(mar = c(2, 4, 0, 0))
boxplot(Lag5 ~Direction, data = Weekly, ylab = "Lag5")

boxplot(Today ~Direction, data = Weekly, ylim = c(-5,5))
boxplot(Volume ~ Year, data = Weekly)

#b)

glm.fit_4b <- glm( Direction ~ Lag1+Lag2+Lag3+Lag4+Lag5+Volume, data = Weekly,family = binomial )
summary(glm.fit_4b)

#c)
probabilidades <- predict(glm.fit_4b, newdata = Weekly[,c(2:6,7)], type = "response")

predicted.classes <- ifelse(probabilidades > 0.5, "Up", "Down")

Mconfusao <- table(Dados = as.factor(Weekly$Direction),predicao = as.factor(predicted.classes))

addmargins(Mconfusao)
```

```
#outra forma
MatrizConfu <-confusionMatrix(data = as.factor(predicted.classes),
reference = as.factor(Weekly$Direction),positive = "Up")

#d)
#treinamento
treinamento_4d <- subset(Weekly, Year <= 2008)
glm.fit_4d <- glm( Direction ~ Lag2 , data = treinamento_4d, family = binomial )
summary(glm.fit_4d)

#Teste
Teste_4d <- subset(Weekly, Year > 2008)
probabilidades.teste_4d <- predict(glm.fit_4d, newdata = data.frame(Lag2 = Teste_4d[,3]),
type = "response")
predicted.classes.teste_4d <- ifelse(probabilidades.teste_4d > 0.5, "Up", "Down")

Mconfusao.teste_4d <- table(Dados = as.factor(Teste_4d$Direction),
predicao = as.factor(predicted.classes.teste_4d))

addmargins(Mconfusao.teste_4d)

#Q5
#a)
str(Default)

glm.fit_5a <- glm( default~ student  +  balance , data = Default , family = binomial )
summary(glm.fit_5a)

confint.default(glm.fit_5a)
 c(-7.149e-01  - qnorm(0.975)*1.475e-01,
   -7.149e-01  + qnorm(0.975)*1.475e-01)
 c(5.738e-03  - qnorm(0.975)*2.318e-04,
   5.738e-03  + qnorm(0.975)*2.318e-04)

#b)
 probabilidade_5b <- predict(glm.fit_5a, newdata = data.frame(student = c("Yes","No"),
 balance = c(2000,2000)), type = "response")
 probabilidade_5b
```

```
 #Q6
require(MASS)
str(Boston)

Boston[,1] <- factor(ifelse(Boston$crim > quantile(Boston$crim,0.5),"Acima","Abaixo"),
levels = rev(c("Acima","Abaixo")))
Boston[,4] <- as.factor(Boston[,4])

#80% treino e 20% teste
set.seed(2020)
Ntreino <- sample(1:506, 405,replace = F)
#selecionando aleatoriamente a amostra de treino
Dados_treino <- Boston[Ntreino,]
Dados_teste <-Boston[-Ntreino,]

glm.fit_6a <- glm( crim ~ . , data = Dados_treino , family = binomial )

summary(glm.fit_6a)

#retirando as variáveis não significativas
glm.fit_6aa <- glm( crim ~ . , data = Dados_treino[,-c(3,4,6)] , family = binomial )

summary(glm.fit_6aa)

#predição
probabilidades.teste6 <- predict(glm.fit_6aa, newdata = Dados_teste[,-c(1,3,4,6)], type = "response")

predicted.classes.teste6 <- ifelse(probabilidades.teste6 > 0.5, "Acima", "Abaixo")

Mconfusao.teste6 <- table(Dados = as.factor(Dados_teste$crim),predicao =
as.factor(predicted.classes.teste6))

addmargins(Mconfusao.teste6)

#outra forma
MatrizConfu6 <- confusionMatrix(data = as.factor(predicted.classes.teste6), reference =
as.factor(Dados_teste$crim),positive = "Acima")
MatrizConfu6
```