12) Logistic Regression

Vitor Kamada

February 2018

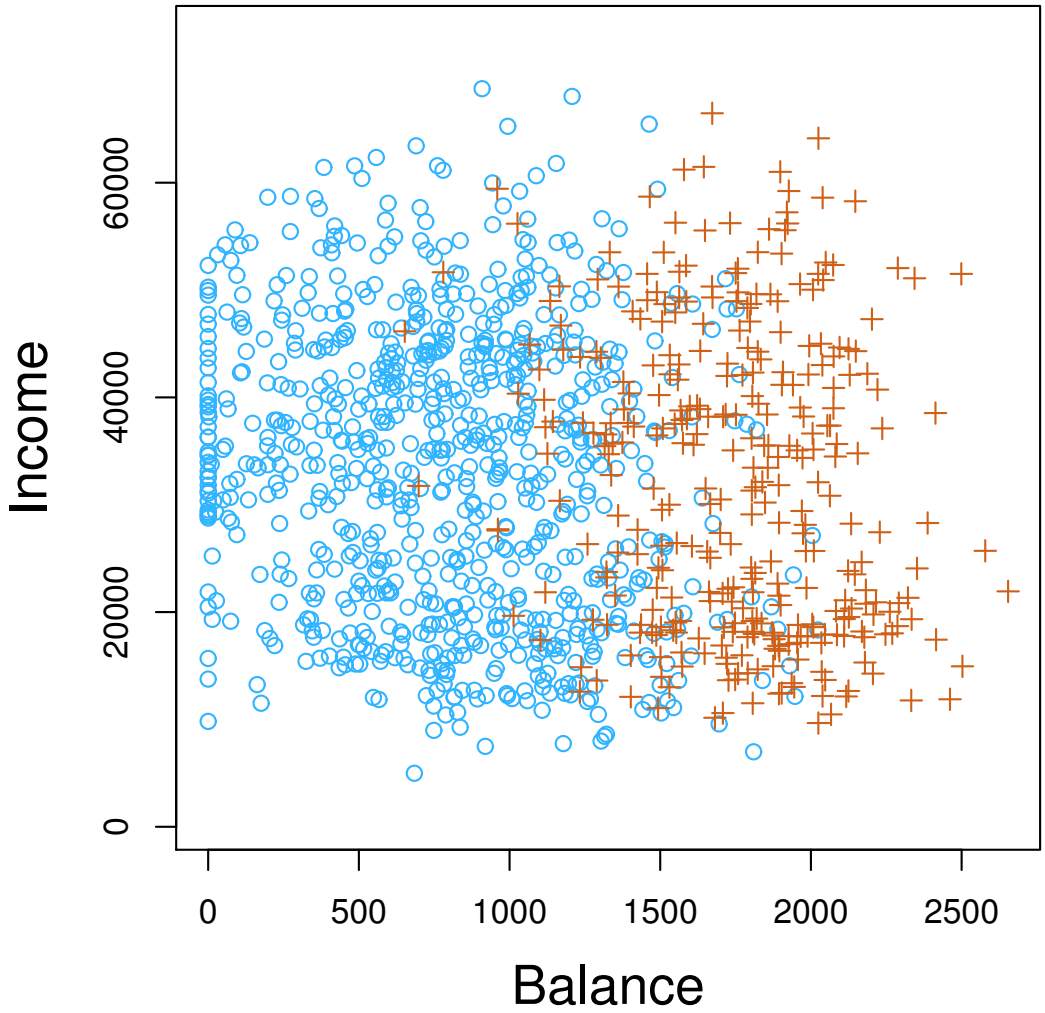
Reference

Tables, Graphics, and Figures from

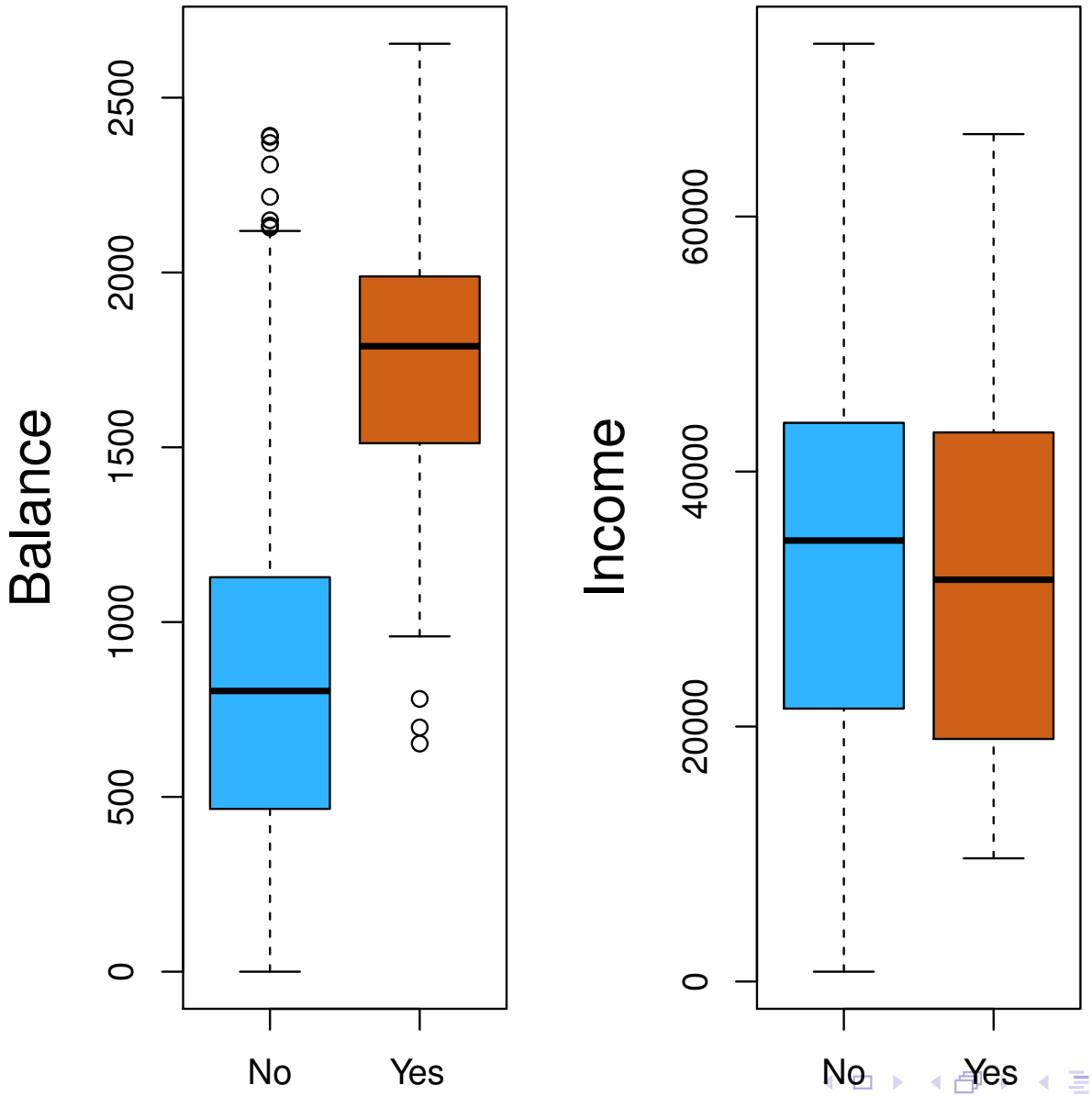
An Introduction to Statistical Learning

James et al. (2017): Chapters: 4.3, 4.6.1, 4.6.2

The Default Data Set: Default Rate $\cong 3\%$



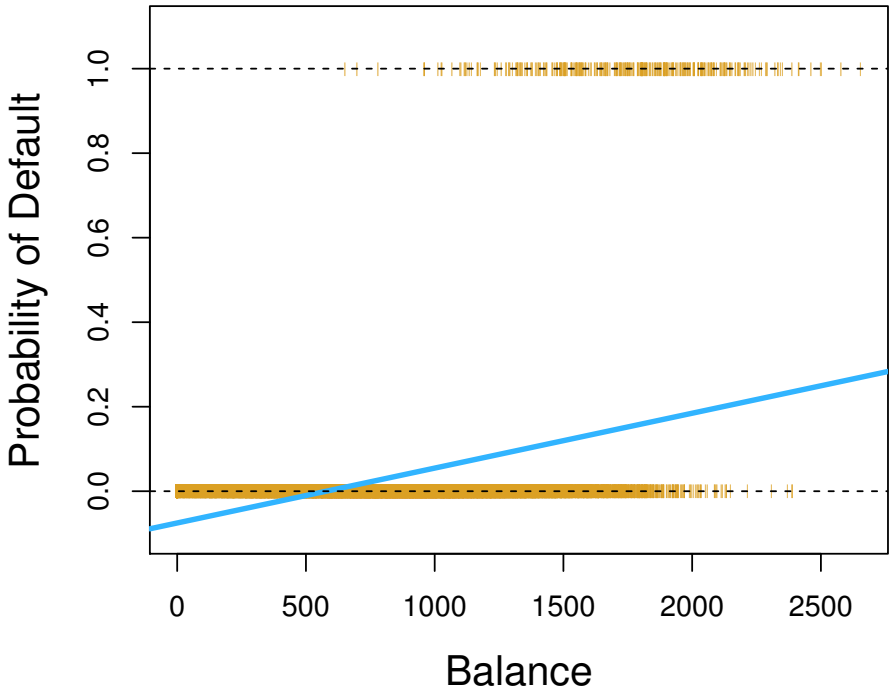
A Subset of 10,000 Individuals

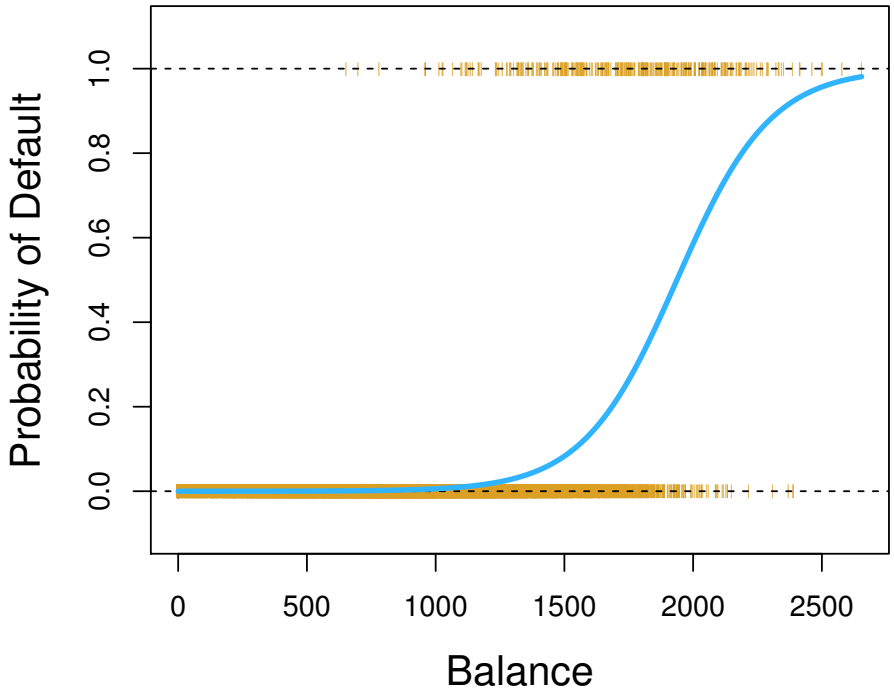


Why Not Linear Regression?

$$y = \begin{cases} 1 \ \textit{if} \ \textit{stroke} \\ 2 \ \textit{if} \ \textit{drug} \ \textit{overdose} \\ 3 \ \textit{if} \ \textit{seizure} \end{cases}$$

Linear Probability vs Logistic Model





The Logistic Model

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

$$1 - p(X) = \frac{1}{1+e^{\beta_0+\beta_1 X}}$$

$$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}$$

$$log\left[\frac{p(X)}{1-p(X)}\right] = \beta_0 + \beta_1 X$$

Logistic Regression

	Coefficient	Std. error	Z-statistic	P-value
Intercept	−10.6513	0.3612	−29.5	<0.0001
balance	0.0055	0.0002	24.9	<0.0001

	Coefficient	Std. error	Z-statistic	P-value
Intercept	−3.5041	0.0707	−49.55	<0.0001
student[Yes]	0.4049	0.1150	3.52	0.0004

	Coefficient	Std. error	Z-statistic	P-value
Intercept	−10.8690	0.4923	−22.08	<0.0001
balance	0.0057	0.0002	24.74	<0.0001
income	0.0030	0.0082	0.37	0.7115
student[Yes]	−0.6468	0.2362	−2.74	0.0062

Predictions given Balance

	Coefficient	Std. error	Z-statistic	P-value
`Intercept`	−10.6513	0.3612	−29.5	<0.0001
`balance`	0.0055	0.0002	24.9	<0.0001

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}}$$

$$\frac{e^{-10.65+0.0055\times 1{,}000}}{1+e^{-10.65+0.0055\times 1{,}000}} = 0.57\%$$

$$X = 2000 \rightarrow \hat{p}(X) = 58.6\%$$

Predictions given Student

	Coefficient	Std. error	Z-statistic	P-value
`Intercept`	−3.5041	0.0707	−49.55	<0.0001
`student[Yes]`	0.4049	0.1150	3.52	0.0004

$$\hat{Pr}(default = Yes|student = Yes)$$

$$\frac{e^{-3.5+0.405\times1}}{1+e^{-3.5+0.405\times1}} = 4.3\%$$

$$\hat{Pr}(default = Yes|student = No)$$

$$\frac{e^{-3.5}}{1+e^{-3.5}} = 2.9\%$$

Multiple Logistic Regression

	Coefficient	Std. error	Z-statistic	P-value
Intercept	−10.8690	0.4923	−22.08	<0.0001
balance	0.0057	0.0002	24.74	<0.0001
income	0.0030	0.0082	0.37	0.7115
student[Yes]	−0.6468	0.2362	−2.74	0.0062

$$\hat{p}(X) = \frac{e^{-10.87+0.0057\times 1{,}500+0.003\times 40-0.65\times 1}}{1+e^{-10.87+0.0057\times 1{,}500+0.003\times 40-0.65\times 1}} = 5.8\%$$

$$\hat{p}(X) = \frac{e^{-10.87+0.0057\times 1{,}500+0.003\times 40-0.65\times 0}}{1+e^{-10.87+0.0057\times 1{,}500+0.003\times 40-0.65\times 0}} = 10.5\%$$

S&P 500 Stock Index Over 1,250 Days

library(ISLR); summary(Smarket)

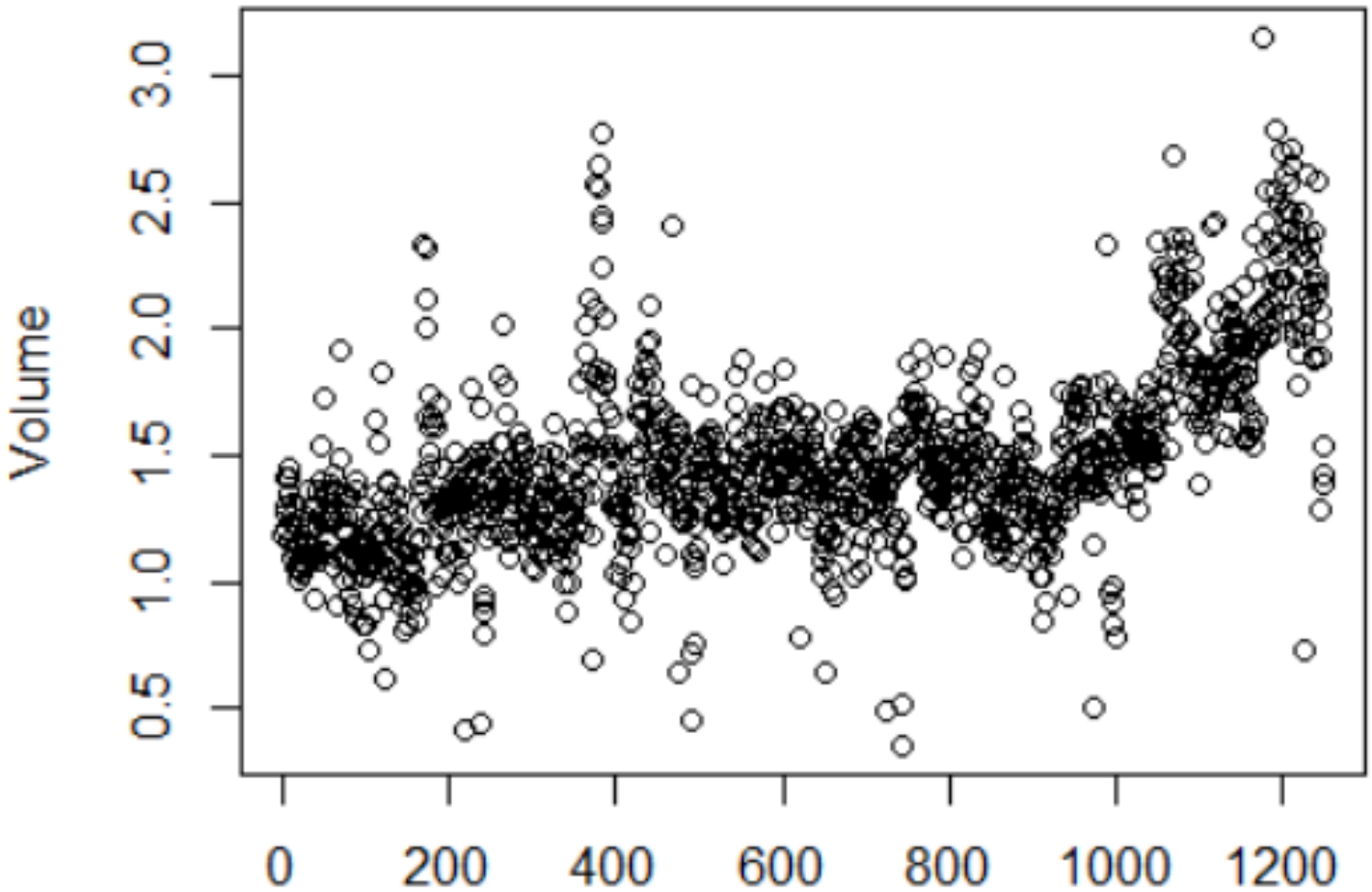
Statistic	N	Mean	St. Dev.	Min	Max
Year	1,250	2,003.016	1.409	2,001	2,005
Lag1	1,250	0.004	1.136	−4.922	5.733
Lag2	1,250	0.004	1.136	−4.922	5.733
Lag3	1,250	0.002	1.139	−4.922	5.733
Lag4	1,250	0.002	1.139	−4.922	5.733
Lag5	1,250	0.006	1.148	−4.922	5.733
Volume	1,250	1.478	0.360	0.356	3.152
Today	1,250	0.003	1.136	−4.922	5.733

round(cor(Smarket[,-9]),2)

	Year	Lag1	Lag2	Lag3	Lag4	Lag5	Volume	Today
Year	1.00	0.03	0.03	0.03	0.04	0.03	0.54	0.03
Lag1	0.03	1.00	-0.03	-0.01	0.00	-0.01	0.04	-0.03
Lag2	0.03	-0.03	1.00	-0.03	-0.01	0.00	-0.04	-0.01
Lag3	0.03	-0.01	-0.03	1.00	-0.02	-0.02	-0.04	0.00
Lag4	0.04	0.00	-0.01	-0.02	1.00	-0.03	-0.05	-0.01
Lag5	0.03	-0.01	0.00	-0.02	-0.03	1.00	-0.02	-0.03
Volume	0.54	0.04	-0.04	-0.04	-0.05	-0.02	1.00	0.01
Today	0.03	-0.03	-0.01	0.00	-0.01	-0.03	0.01	1.00

Average Number of Shares Traded Daily Increased from 2001 to 2005

plot(Volume)



glm.fit=glm(Direction~Lag1+Lag2+Lag3+Lag4+Lag5+Volume, data = Smarket, family=binomial)

summary(glm.fit)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.126000	0.240736	-0.523	0.601
Lag1	-0.073074	0.050167	-1.457	0.145
Lag2	-0.042301	0.050086	-0.845	0.398
Lag3	0.011085	0.049939	0.222	0.824
Lag4	0.009359	0.049974	0.187	0.851
Lag5	0.010313	0.049511	0.208	0.835
Volume	0.135441	0.158360	0.855	0.392

exp(cbind(OR = coef(glm.fit), confint(glm.fit)))

```
                    OR      2.5 %    97.5 %
(Intercept) 0.8816146 0.5493177 1.412613
Lag1        0.9295323 0.8420468 1.025314
Lag2        0.9585809 0.8686451 1.057370
Lag3        1.0111468 0.9167701 1.115303
Lag4        1.0094029 0.9151142 1.113442
Lag5        1.0103664 0.9168466 1.113533
Volume      1.1450412 0.8398920 1.563564
```

glm.probs=predict(glm.fit,type="response")

glm.probs[1:10]

```
        1         2         3         4         5         6         7
0.5070841 0.4814679 0.4811388 0.5152224 0.5107812 0.5069565 0.4926509
        8         9        10
0.5092292 0.5176135 0.4888378
```

glm.pred=rep("Down",1250)
glm.pred[glm.probs>.5]="Up"

table(glm.pred,Direction)

```
        Direction
glm.pred Down  Up
    Down  145 141
    Up    457 507
```

$\frac{145+507}{1250} \simeq 52.16\%$

Training Error vs Test Error

```
train=(Year<2005)

Smarket.2005=Smarket[!train,]

dim(Smarket.2005)

Direction.2005=Direction[!train]

glm.fit=glm(Direction~Lag1+Lag2+Lag3+Lag4+Lag5+
Volume, data=Smarket, family=binomial,subset=train)

glm.probs=predict(glm.fit,Smarket.2005,type="response")

glm.pred=rep("Down",252)

glm.pred[glm.probs>.5]="Up"
```

table(glm.pred,Direction.2005)

```
            Direction.2005
glm.pred Down Up
    Down   77 97
    Up     34 44
```

mean(glm.pred==Direction.2005)

$$\frac{(77+45)}{252} \cong 48.1\%$$

mean(glm.pred!=Direction.2005)

$$\frac{(34+97)}{252} \cong 51.9\%$$

glm.fit=glm(Direction~Lag1+Lag2, data=Smarket, family=binomial,subset=train)

```
glm.probs = predict(glm.fit,Smarket.2005,
   type ="response")
glm.pred=rep("Down",252)
glm.pred[glm.probs>.5]="Up"
table(glm.pred,Direction.2005)
```

glm.pred	Down	Up
Down	35	35
Up	76	106

$$\frac{(35+106)}{252} \simeq 55.9\%$$

summary(glm.fit)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.03222	0.06338	0.508	0.611
Lag1	-0.05562	0.05171	-1.076	0.282
Lag2	-0.04449	0.05166	-0.861	0.389

Specific Prediction

```
predict(glm.fit,newdata =
data.frame(Lag1=c(5,-3),Lag2=c(4,-6)),
type="response")
```

1	2
0.39	0.61