

Business Statistics 41000

Multiple Linear Regression

Mladen Kolar

Binary response data

Often we want to predict a response variable that is binary: $Y = 0$ or 1 .

The goal is generally to predict the probability that $Y = 1$. You can then do **classification** based on this estimate.

- ▶ Buy or not buy.
- ▶ Win or lose.
- ▶ Sick or healthy.
- ▶ Pay or default.
- ▶ Thumbs up or down.

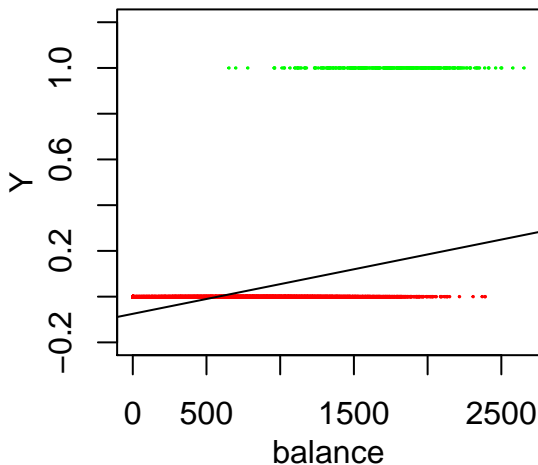
Can we use linear regression for this task?

```
library(ISLR)
df = Default
df$Y = as.numeric(df$default)-1
head(df)
```

##	default	student	balance	income	Y
## 1	No	No	730	44362	0
## 2	No	Yes	817	12106	0
## 3	No	No	1074	31767	0
## 4	No	No	529	35704	0
## 5	No	No	786	38463	0
## 6	No	Yes	920	7492	0

Can we use linear regression for this task?

```
par(mar=c(3,3,3,1), mgp=c(2,1,0))  
plot(df$balance, df$Y, col=c("red", "green")[df$Y+1],  
      xlab="balance", ylab="Y", ylim = c(-0.2, 1.2), cex = 0.1)  
abline(lm(Y~balance, data=df))
```



Can we use linear regression for this task?

We can, but the fit is very unappealing, as is the interpretation.

```
summary(lm(Y~balance, data=df))
```

```
##  
## Call:  
## lm(formula = Y ~ balance, data = df)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max   
## -0.2353 -0.0694 -0.0263  0.0200  0.9905   
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)      
## (Intercept) -7.52e-02  3.35e-03  -22.4   <2e-16 ***  
## balance      1.30e-04  3.47e-06   37.4   <2e-16 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 0.168 on 9998 degrees of freedom  
## Multiple R-squared:  0.123, Adjusted R-squared:  0.122   
## F-statistic: 1.4e+03 on 1 and 9998 DF,  p-value: <2e-16
```

This is just a bad fit.

Modelling probability $P(Y = 1 | X)$

Y is an indicator: $Y = 0$ or 1 .

The conditional mean is thus

$$\begin{aligned}\mathbb{E}[Y|X] &= p(Y = 1|X) \times 1 + p(Y = 0|X) \times 0 \\ &= p(Y = 1|X).\end{aligned}$$

So the mean function is a probability.

We need a model that gives mean/probability values between 0 and 1.

Logistic regression

Model for conditional distribution of Y given $X = x$

$$p_1(x; \beta) = P(Y = 1 \mid x; \beta) = S(\beta_0 + \sum_{j=1}^p \beta_j \cdot x_j)$$

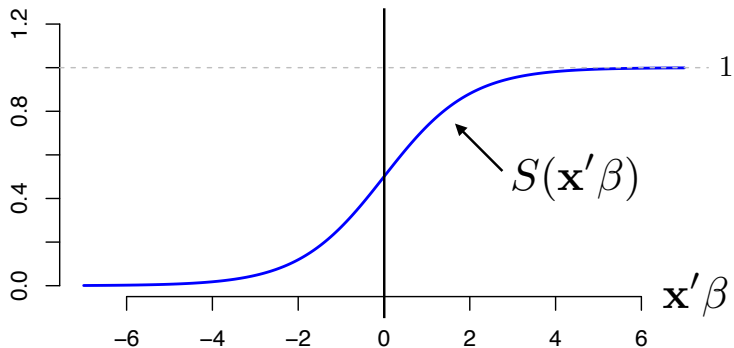
$$p_0(x; \beta) = P(Y = 0 \mid x; \beta) = 1 - p_1(x; \beta)$$

where $S(z) = \frac{e^z}{1 + e^z}$ is the logistic sigmoid function.

The log-odds of class 1 is a linear function of X

$$\log \left(\frac{p_1(x; \beta)}{p_0(x; \beta)} \right) = \beta_0 + \sum_{j=1}^p \beta_j \cdot x_j$$

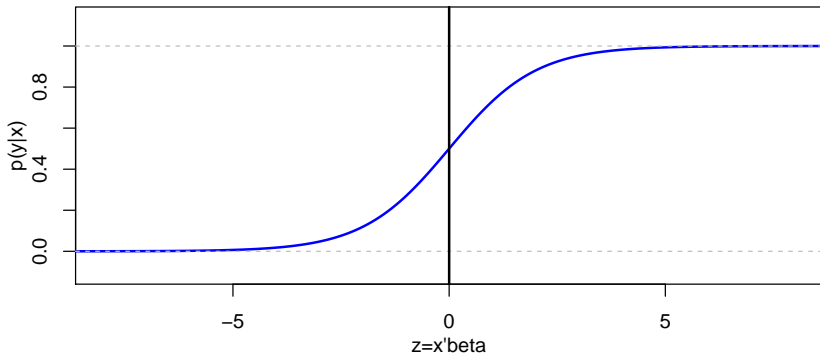
Representation with logistic regression



A linear function $\mathbf{x}'\beta$ can take values from $-\infty$ to ∞ .
By passing it through sigmoid, we get values in $[0, 1]$.

These values also sum to 1 to make a probability.

Logistic sigmoid function



Step 1: Linear combination

$$z = x'\beta$$

Step 2: Nonlinear transformation

$$P(Y = 1 \mid x; \beta) = F(z) = \frac{\exp(z)}{1 + \exp(z)}$$

z	$F(z)$
-3	0.05
-2	0.12
-1	0.27
0	0.5
1	0.73
2	0.88
3	0.95

Logistic regression

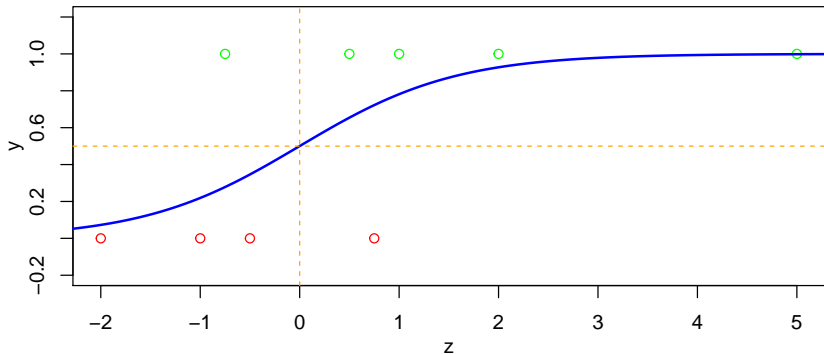
Our model for $Y \mid X$ is the following

$$p(Y = 1 | X_1 \dots X_d) = S(\mathbf{X}'\beta) = \frac{\exp[\beta_0 + \beta_1 X_1 \dots + \beta_d X_d]}{1 + \exp[\beta_0 + \beta_1 X_1 \dots + \beta_d X_d]}.$$

These models are easy to fit in R:

```
glm(Y ~ X1 + X2, family=binomial)
```

- ▶ “g” is for **generalized**; **binomial** indicates $Y = 0$ or 1 .
- ▶ Otherwise, **glm** uses the same syntax as **lm**.

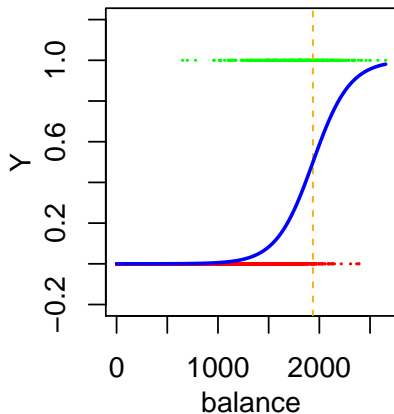
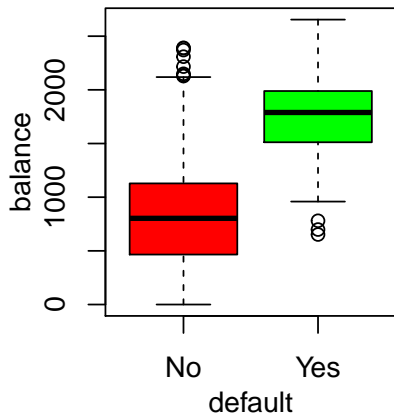


Suppose that $\hat{Y}_i = 1$ if $P(Y = 1 | z_i) > 0.5$ and $\hat{Y}_i = 0$ otherwise.

When do we assign a new observation to a positive class?

For what values of z do we have $\hat{Y} = 1$?

Example: Predicting default



```
fit = glm(default~balance, df, family="binomial")
summary(fit)
```

```
## Call:
## glm(formula = default ~ balance, family = "binomial", data = df)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -10.65133    0.36116   -29.5    <2e-16 ***
## balance      0.00550    0.00022    24.9    <2e-16 ***
```

$$\hat{p}_1(X) = \hat{p}(\text{default} \mid X) = \frac{\exp(-10.65 + X \cdot 0.0055)}{1 + \exp(-10.65 + X \cdot 0.0055)}$$

New predictions

The predict function works as before, but add `type = "response"` to get $\hat{p} = \exp[\mathbf{x}'\mathbf{b}]/(1 + \exp[\mathbf{x}'\mathbf{b}])$ (otherwise it just returns the linear function $\mathbf{x}'\mathbf{b}$).

```
xf = data.frame(balance=c(1500, 2000, 2500))
phat = predict(fit, newdata=xf, type="response")
cbind(xf, phat)
```

```
##   balance   phat
## 1    1500 0.0829
## 2    2000 0.5858
## 3    2500 0.9567
```

More features

```
summary(glm(default~balance+income, data=df, family="binomial"))
```

```
## Call:
```

```
## glm(formula = default ~ balance + income, family = "binomial",
```

```
##      data = df)
```

```
##
```

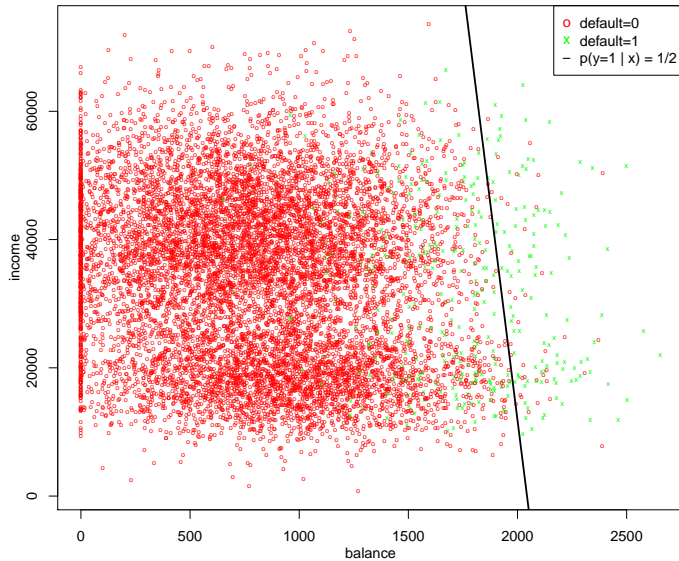
```
## Coefficients:
```

```
##           Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept) -1.15e+01  4.35e-01  -26.54   <2e-16 ***
```

```
## balance      5.65e-03  2.27e-04   24.84   <2e-16 ***
```

```
## income       2.08e-05  4.99e-06    4.17    3e-05 ***
```



Confounding

```
summary(glm(default~student, data=df, family="binomial"))
```

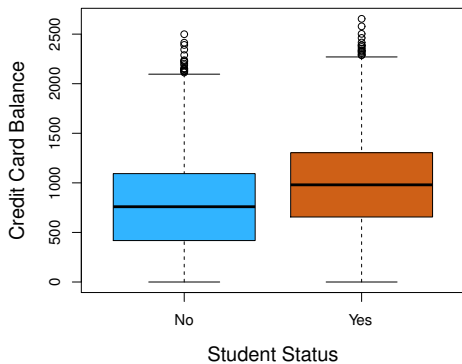
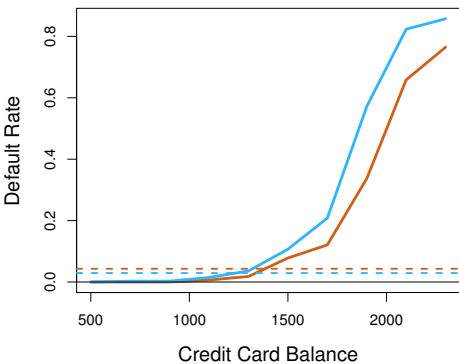
```
## Coefficients:
```

##	Estimate	Std. Error	z value	Pr(> z)
## (Intercept)	-3.5041	0.0707	-49.55	< 2e-16 ***
## studentYes	0.4049	0.1150	3.52	0.00043 ***

```
summary(glm(default~balance+student, data=df, family="binomial"))
```

```
## Coefficients:
```

##	Estimate	Std. Error	z value	Pr(> z)
## (Intercept)	-1.07e+01	3.69e-01	-29.12	< 2e-16 ***
## balance	5.74e-03	2.32e-04	24.75	< 2e-16 ***
## studentYes	-7.15e-01	1.48e-01	-4.85	1.3e-06 ***



```
summary(glm(default~balance+income+student, data=df, family="binomial"))
```

```
## Call:
```

```
## glm(formula = default ~ balance + income + student, family = "binomial",  
##      data = df)
```

```
##
```

```
## Coefficients:
```

##		Estimate	Std. Error	z value	Pr(> z)
##	(Intercept)	-1.09e+01	4.92e-01	-22.08	<2e-16 ***
##	balance	5.74e-03	2.32e-04	24.74	<2e-16 ***
##	income	3.03e-06	8.20e-06	0.37	0.7115
##	studentYes	-6.47e-01	2.36e-01	-2.74	0.0062 **