Stat 306: Finding Relationships in Data.

Lecture 19

6.2 Logistic regression (Part 1)

Age vs. Money

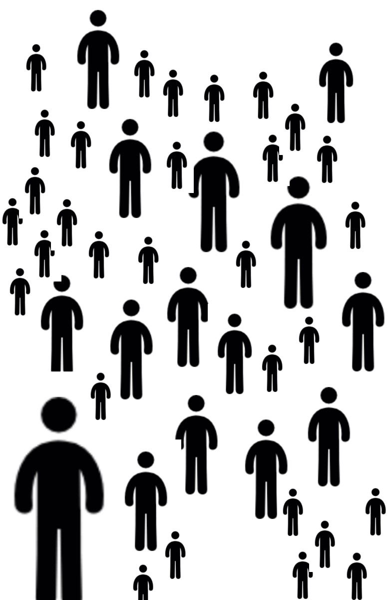
Dependent variable

X → old (0), young (1)

Independent variable

Y → dollars ($) in cash

Population



Population parameters

μ_0, μ_1, σ^2

Hypothesis Test

$H_0 : \mu_0 = \mu_1$

$H_1 : \mu_0 \neq \mu_1$

Sample statistics

$\bar{y}_0 = 56$

$\bar{y}_1 = 27$

$\bar{y}_0 - \bar{y}_1 = 29$

$s_p = 10.81$

$t = 2.68$, $df = 7$

p-value = 0.03

95% C.I. = [3.4, 54.6]

Sample, n=9

x	y
old	71
old	54
old	43
young	45
young	21
young	11
young	30
young	45
young	10

Age vs. Money

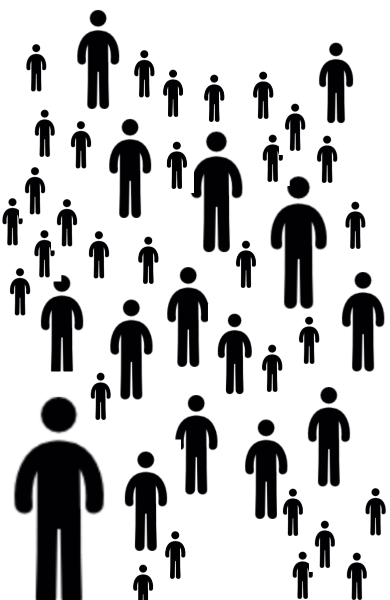
PREDICTOR *variable*

$X \longrightarrow$ Age in Years

RESPONSE *variable*

$Y \longrightarrow$ dollars ($) in cash

Population



Population parameters

$$\beta_0, \beta_1, \sigma^2$$

Hypothesis Test

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

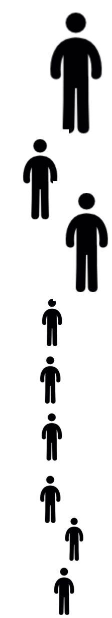
Sample statistics

$b_0 = 17.7$
$b_1 = 0.55$
$s = 15.5$
$R^2 = 0.49$

For parameter β_1 :

95% C.I. = [0.05, 1.05]
p-value = 0.036

Sample, n=9



x	y
82	71
45	54
71	43
22	45
29	21
9	11
12	30
18	45
24	10

Age vs. Money

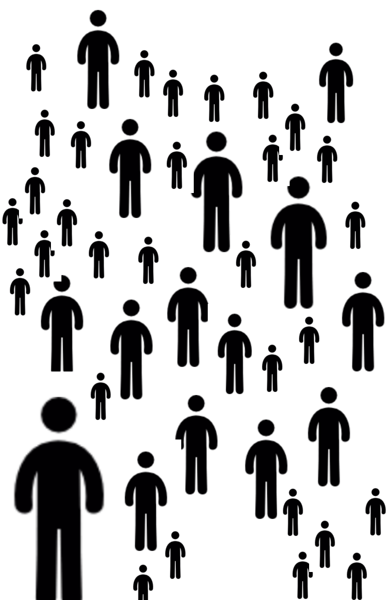
PREDICTOR *variable*

$X \longrightarrow$ Age in Years

RESPONSE *variable*

$Y \longrightarrow$ Use cash regularly? (yes/no)

Population



Population parameters

$$\beta_0 , \beta_1 , \sigma^2$$

Hypothesis Test

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

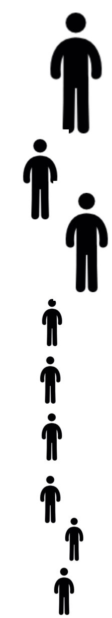
Sample statistics

$b_0 = 17.7$
$b_1 = 0.55$
$s = 15.5$
$R^2 = 0.49$

For parameter β_1 :

95% C.I. $= [0.05, 1.05]$
p-value $= 0.036$

Sample, n=9



X	y
82	1
45	1
71	1
22	0
29	0
9	0
12	0
18	1
24	0

Age vs. Money

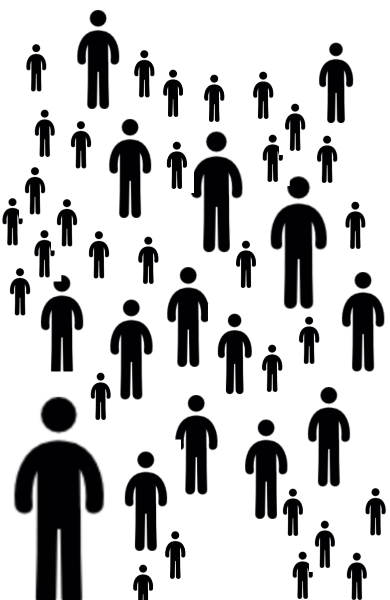
PREDICTOR *variable*

$X \longrightarrow$ Age in Years

RESPONSE *variable*

$Y \longrightarrow$ Use cash regularly? (yes/no)

Population



Population parameters

$\beta_0 , \beta_1 , \sigma^2$

Hypothesis Test

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

Sample statistics

$b_0 = 17.7$

$b_1 = 0.55$

$s = 15$

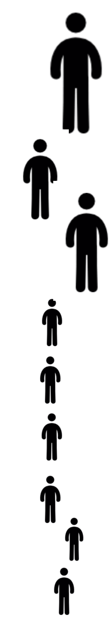
$R^2 = 0.49$

For parameter β_1 :

95% C.I. = [0.05, 1.05]

p-value = 0.0

Sample, n=9



x	y
82	1
45	1
71	1
22	0
29	0
9	0
12	0
18	1
24	0

Age vs. Money

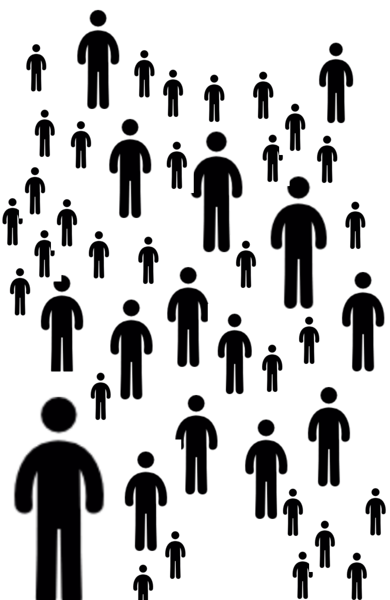
PREDICTOR *variable*

X → Age in Years

RESPONSE *variable*

Y → Use cash regularly? (yes/no)

Population



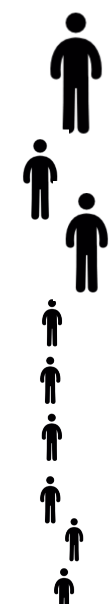
Instead of :

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad , \quad i = 1, ..., n$$

We have:

$$p_i = \frac{1}{1+e^{-(\beta_0+\beta_1 x_i)}} \quad , \quad i = 1, ..., n$$

Sample, n=9



x	y
82	1
45	1
71	1
22	0
29	0
9	0
12	0
18	1
24	0

Age vs. Money

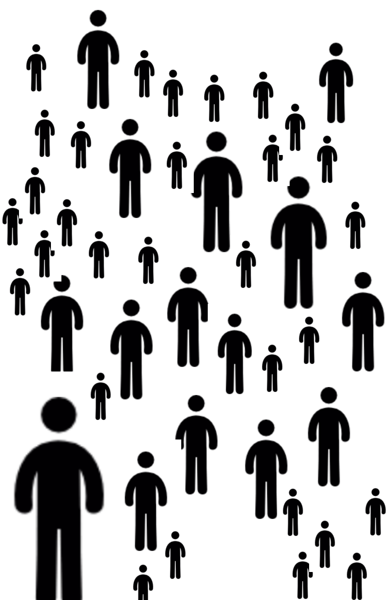
PREDICTOR *variable*

X ⟶ Age in Years

RESPONSE *variable*

Y ⟶ Use cash regularly? (yes/no)

Population



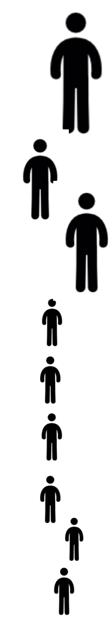
Instead of :

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad , \quad i = 1, ..., n$$

We have:

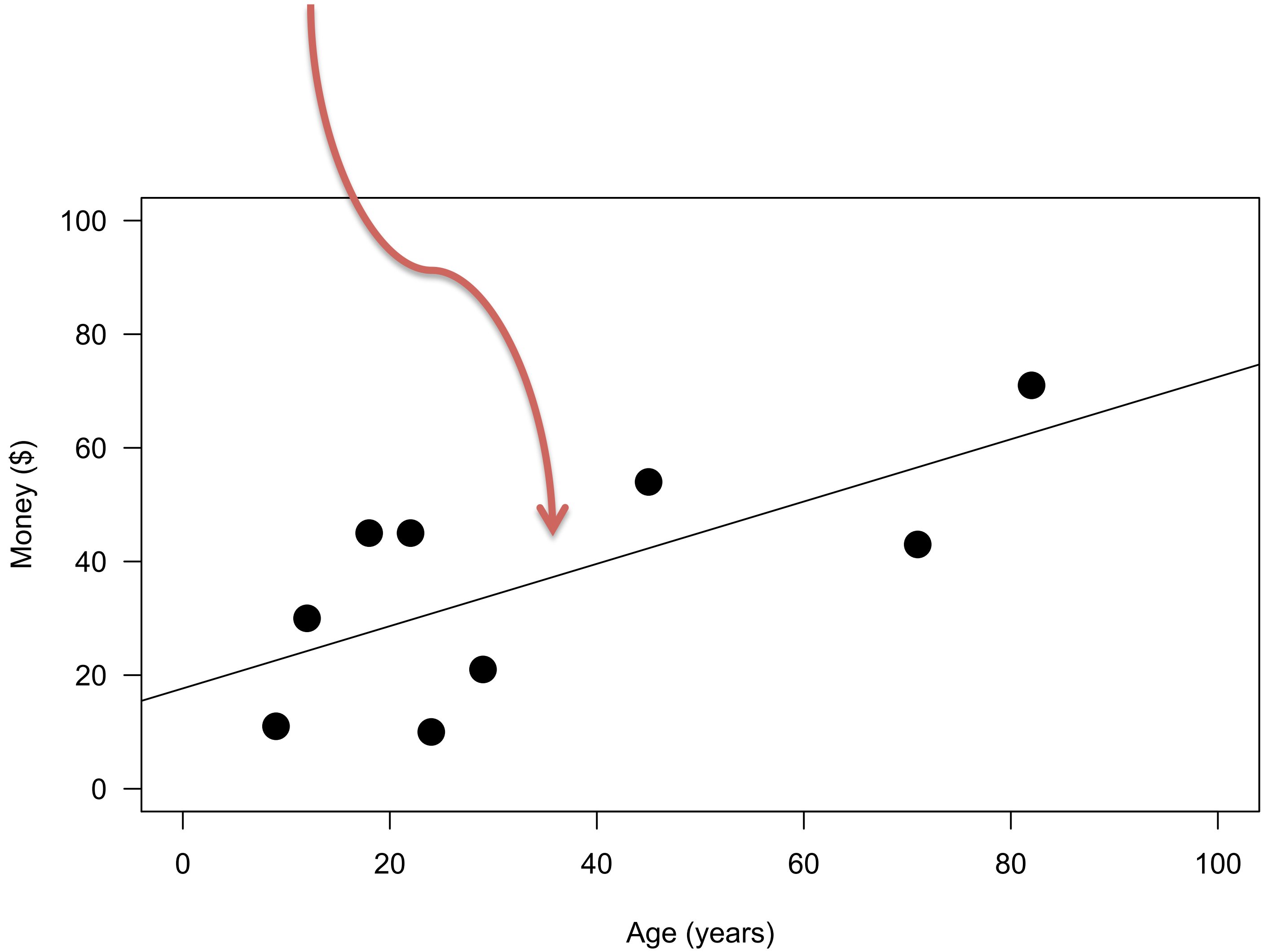
$$p_i = \frac{1}{1+e^{-(\beta_0+\beta_1 x_i)}} \quad , \quad i = 1, ..., n$$

Sample, n=9

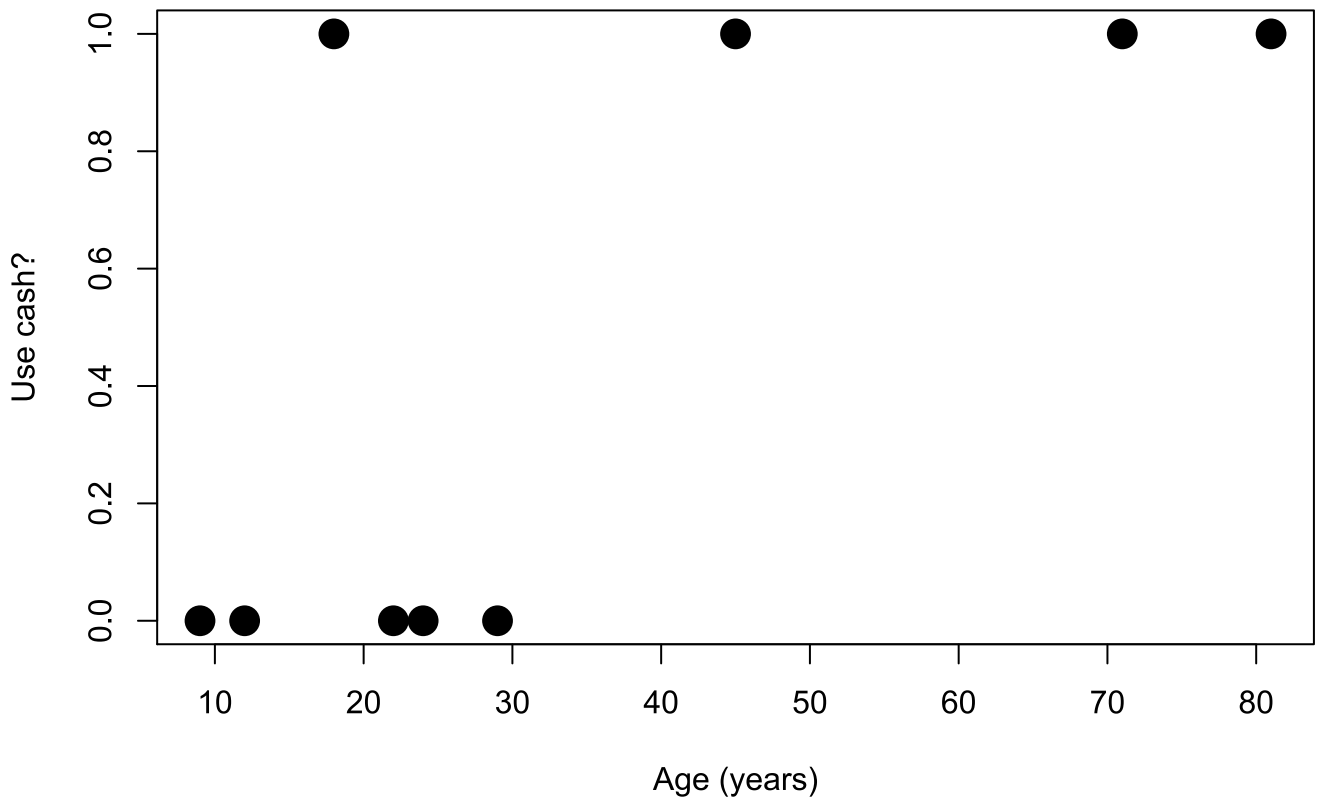


x	y
82	1
45	1
71	1
22	0
29	0
9	0
12	0
18	1
24	0

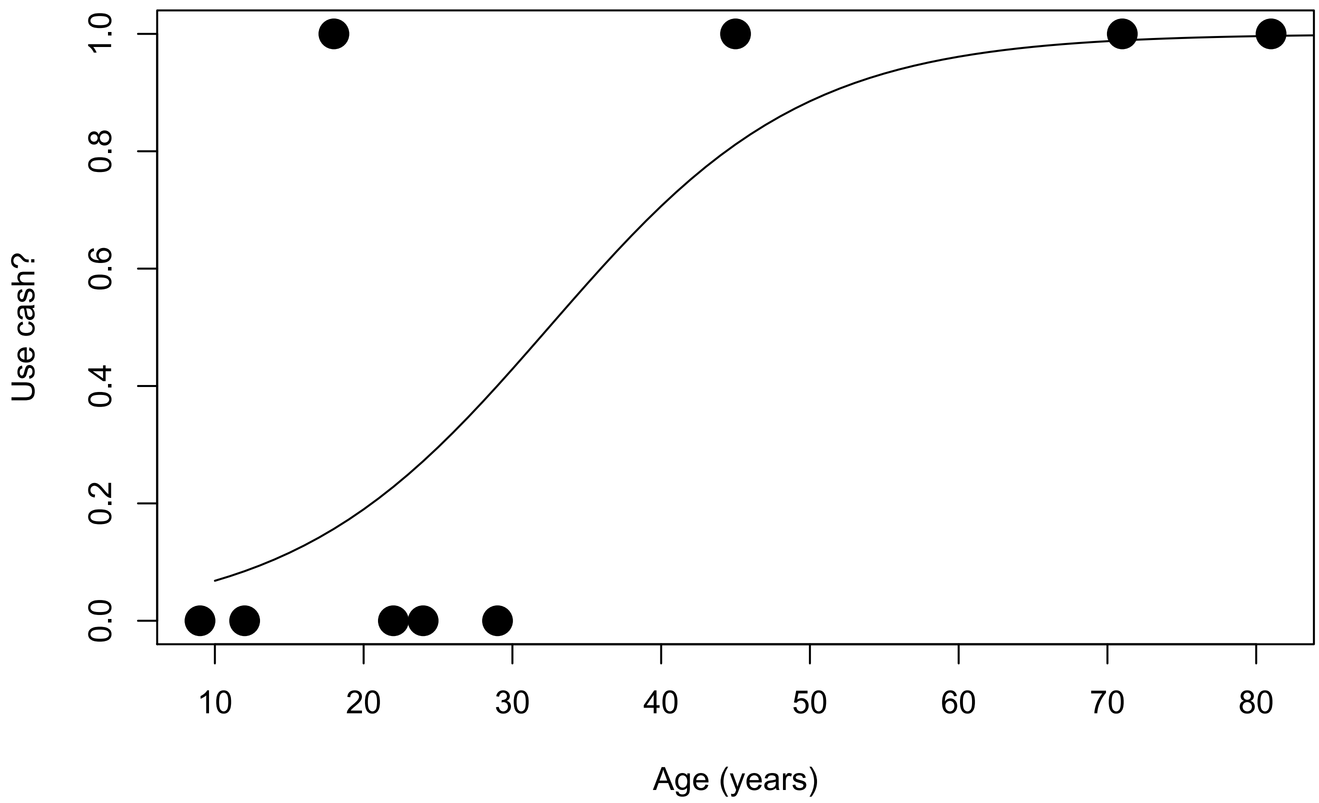
prediction equation : $y = b_0 + b_1x$



prediction equation ?



prediction equation ?



Standard **linear** regression:

1. We assume outcomes are normally distributed:

$$Y \sim N(\mu, \sigma^2)$$

2. Depending on the values of covariates, each outcome (*i* = 1,…,*n*) is normally distributed with a different mean:

$$Y_i \sim N(\mathbf{x}_i^T \boldsymbol{\beta}, \sigma^2)$$

or we write this out:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i$$

3. In other words, we model how the mean of Y changes given different values for X:

$$\mathrm{E}\left(Y_i; \mathbf{x}_i\right) = \mu(\mathbf{x}_i) = \mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$$

Standard **logistic** regression:

1. We assume outcomes are distributed according to a Bernoulli distribution:

$$Y \sim \text{Bernoulli}(\pi) \quad \text{, where } \pi = \Pr(Y = 1) \in (0,1)$$

2. Depending on the values of covariates, each outcome (*i* = 1,…,*n*) is a Bernoulli distributed with a different mean:

$$Y_i \sim \text{Bernoulli}(\pi(\mathbf{x}_i))$$

or we write this out:

$$\Pr(Y_i = 1) = \pi(\mathbf{x}_i)$$
$$= \exp\{\mathbf{x}_i^T \boldsymbol{\beta}\}/[1 + \exp\{\mathbf{x}_i^T \boldsymbol{\beta}\}]$$

3. In other words, we model how the log-odds of Y=1 vs. Y=0 changes given different values for X:

$$\log\left\{\frac{P(Y_i = 1; \mathbf{x}_i)}{P(Y = 0; \mathbf{x}_i)}\right\} = \log\left\{\frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)}\right\} = \mathbf{x}_i^T \boldsymbol{\beta} \in (-\infty, \infty)$$

Standard **logistic** regression:

$$\Pr(Y_i = 1) = \pi(\mathbf{x}_i)$$
$$= \exp\{\mathbf{x}_i^T \boldsymbol{\beta}\}/[1 + \exp\{\mathbf{x}_i^T \boldsymbol{\beta}\}]$$
$$= \frac{1}{1+exp(-x_i^T \beta)}$$

In other words, we model how the log-odds of Y=1 vs. Y=0 changes given different values for X:

$$\log\left\{\frac{P(Y_i = 1; \mathbf{x}_i)}{P(Y = 0; \mathbf{x}_i)}\right\} = \log\left\{\frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)}\right\} = \mathbf{x}_i^T \boldsymbol{\beta} \in (-\infty, \infty)$$

Age vs. Money

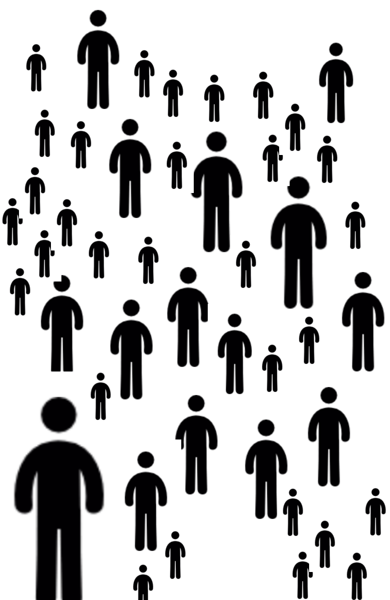
PREDICTOR variable

X → Age in Years

RESPONSE variable

Y → Use cash regularly? (yes/no)

Population



Population parameters

β_0, β_1

Hypothesis Test

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

Sample statistics

b_0 = -3.78

b_1 = 0.116

Null deviance = 12.37

Residual deviance = 6.63

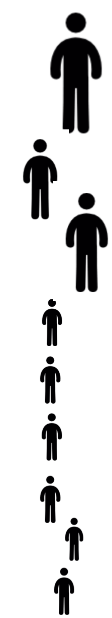
AIC = 10.6

For parameter β_1:

95% C.I. = [0.01, 0.41]

p-value = 0.179

Sample, n=9



X	y
82	1
45	1
71	1
22	0
29	0
9	0
12	0
18	1
24	0

Age vs. Money

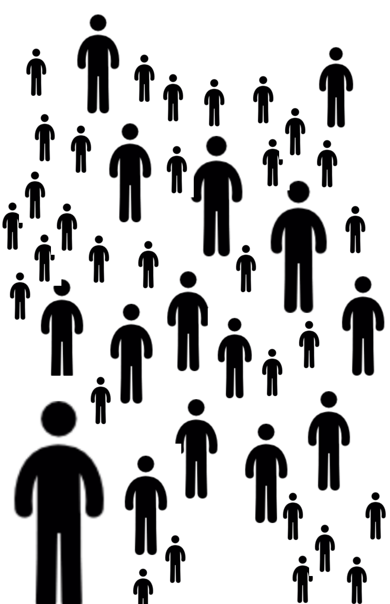
Interpretation: There is a 12.3% increase in the odds of carrying cash for every additional year in age. Note that exp(0.116) = 1.123 and consider the following:

$$Pr(Y = 1|X = x^*) = \frac{1}{1+\exp(-(-3.78+0.12x^*))}$$

$$Pr(Y = 0|X = x^*) = 1 - Pr(Y = 1|X = x^*)$$

$$\frac{Pr(Y=1|X=x^*)}{Pr(Y=0|X=x^*)} = \exp(-3.78 + 0.12x^*)$$

Population



Population parameters

β_0, β_1

Hypothesis Test

$H_0 : \beta_1 = 0$

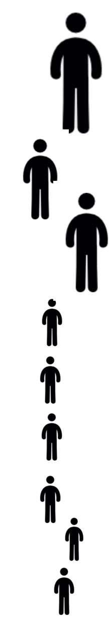
$H_1 : \beta_1 \neq 0$

Sample statistics

b_0 = -3.78
b_1 = 0.116
Null deviance = 12.37
Residual deviance = 6.63
AIC = 10.6

For parameter β_1:
95% C.I. = [0.01, 0.41]
p-value = 0.179

Sample, n=9



x	y
82	1
45	1
71	1
22	0
29	0
9	0
12	0
18	1
24	0

Age vs. Money

Interpretation: There is a 12.3% increase in the odds of carrying cash for every additional year in age. Note that exp(0.1163) = 1.123 and consider the following:

$$Pr(Y = 1|X = x^*) = \frac{1}{1+\exp(-(-3.78+0.12x^*))}$$

$$Pr(Y = 0|X = x^*) = 1 - Pr(Y = 1|X = x^*)$$

$$\frac{Pr(Y=1|X=x^*)}{Pr(Y=0|X=x^*)} = \exp(-3.78 + 0.12x^*)$$

Therefore:

$$odds_{x^*} = \frac{Pr(Y=1|X=x^*)}{Pr(Y=0|X=x^*)} = \exp(-3.78 + 0.12x^*)$$

$$odds_{x^*+1} = \frac{Pr(Y=1|X=x^*+1)}{Pr(Y=0|X=x^*+1)} = \exp(-3.78 + 0.12(x^* + 1))$$

Age vs. Money

Interpretation: There is a 12.3% increase in the odds of carrying cash for every additional year in age. Note that exp(0.1163) = 1.123 and consider the following:

$$Pr(Y=1|X=x^*) = \frac{1}{1+\exp(-(-3.78+0.12x^*))}$$

$$Pr(Y=0|X=x^*) = 1 - Pr(Y=1|X=x^*)$$

$$\frac{Pr(Y=1|X=x^*)}{Pr(Y=0|X=x^*)} = \exp(-3.78+0.12x^*)$$

Therefore:

$$odds_{x^*} = \frac{Pr(Y=1|X=x^*)}{Pr(Y=0|X=x^*)} = \exp(-3.78+0.12x^*)$$

$$odds_{x^*+1} = \frac{Pr(Y=1|X=x^*+1)}{Pr(Y=0|X=x^*+1)} = \exp(-3.78+0.12x^*+0.12)$$

Age vs. Money

Interpretation: There is a 12.3% increase in the odds of carrying cash for every additional year in age. Note that exp(0.1163) = 1.123 and consider the following:

$$Pr(Y = 1|X = x^*) = \frac{1}{1+\exp(-(-3.78+0.12x^*))}$$

$$Pr(Y = 0|X = x^*) = 1 - Pr(Y = 1|X = x^*)$$

$$\frac{Pr(Y=1|X=x^*)}{Pr(Y=0|X=x^*)} = \exp(-3.78 + 0.12x^*)$$

Therefore:

$$odds_{x^*} = \frac{Pr(Y=1|X=x^*)}{Pr(Y=0|X=x^*)} = \exp(-3.78 + 0.12x^*)$$

$$odds_{x^*+1} = \frac{Pr(Y=1|X=x^*+1)}{Pr(Y=0|X=x^*+1)} = \exp(-3.78 + 0.12x^*)\exp(0.12)$$

Age vs. Money

Interpretation: There is a 12.3% increase in the odds of carrying cash for every additional year in age. Note that exp(0.1163) = 1.123 and consider the following:

$$Pr(Y = 1|X = x^*) = \frac{1}{1+\exp(-(-3.78+0.12x^*))}$$

$$Pr(Y = 0|X = x^*) = 1 - Pr(Y = 1|X = x^*)$$

$$\frac{Pr(Y=1|X=x^*)}{Pr(Y=0|X=x^*)} = \exp(-3.78 + 0.12x^*)$$

Therefore:

$$odds_{x^*} = \frac{Pr(Y=1|X=x^*)}{Pr(Y=0|X=x^*)} = \exp(-3.78 + 0.12x^*)$$

$$odds_{x^*+1} = \frac{Pr(Y=1|X=x^*+1)}{Pr(Y=0|X=x^*+1)} = (odds_{x^*})\exp(0.1163)$$

$$= (odds_{x^*})1.123$$

Age vs. Money

Interpretation: There is a 12.3% increase in the odds of carrying cash for every additional year in age. Note that exp(0.1163) = 1.123 and consider the following:

$$Pr(Y = 1|X = x^*) = \frac{1}{1+\exp(-(-3.78+0.12x^*))}$$

$$Pr(Y = 0|X = x^*) = 1 - Pr(Y = 1|X = x^*)$$

$$\frac{Pr(Y=1|X=x^*)}{Pr(Y=0|X=x^*)} = \exp(-3.78 + 0.12x^*)$$

Therefore:

$$odds_{x^*} = \frac{Pr(Y=1|X=x^*)}{Pr(Y=0|X=x^*)} = \exp(-3.78 + 0.12x^*)$$

$$odds_{x^*+1} = \frac{Pr(Y=1|X=x^*+1)}{Pr(Y=0|X=x^*+1)} = (odds_{x^*})\exp(0.1163)$$

$$= (odds_{x^*})1.123$$

Therefore: the odds is increased by 12.3% for every additional 1 unit of x.

Age vs. Money

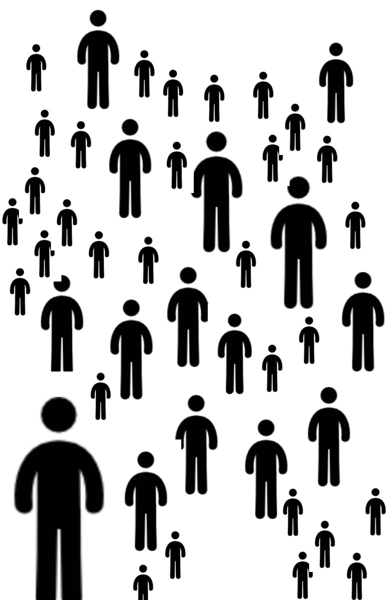
PREDICTOR variable

$X \longrightarrow$ Age in Years

RESPONSE variable

$Y \longrightarrow$ Use cash regularly? (yes/no)

Population



Population parameters

β_0 , β_1

Hypothesis Test

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

Sample statistics

b_0 = -3.78

b_1 = 0.116, exp(0.116) =1.123

Null deviance = 12.37

Residual deviance = 6.63

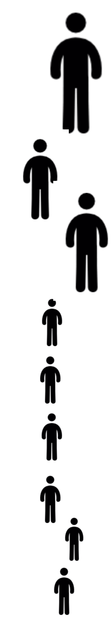
AIC = 10.6

For parameter β_1:

95% C.I. = [0.01, 0.41]

p-value = 0.179

Sample, n=9

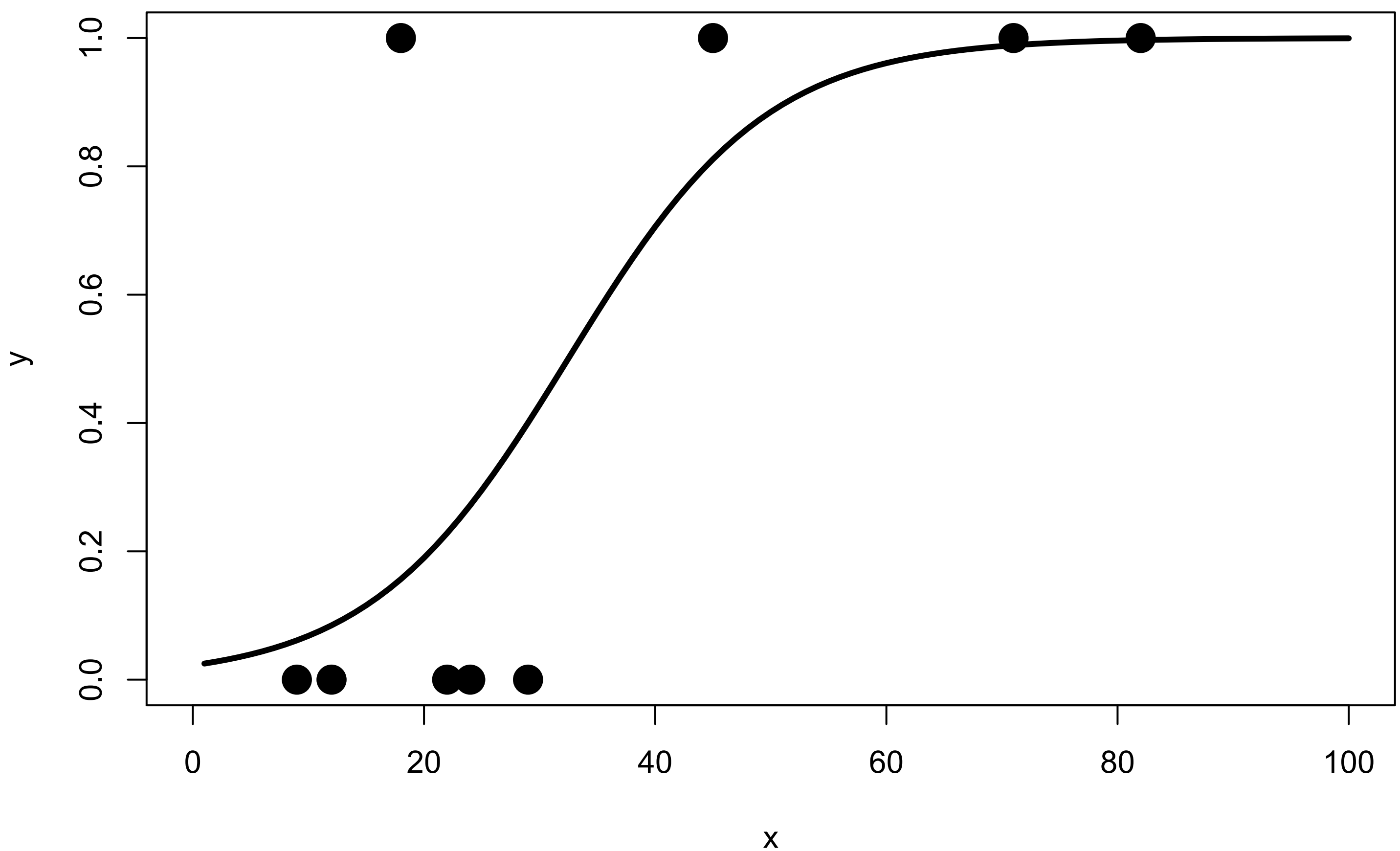


X	y
82	1
45	1
71	1
22	0
29	0
9	0
12	0
18	1
24	0

The fitted logistic model:

$$Z = -3.777 + 0.11X$$

$$Pr(Y=1) = 1/(1+\exp(-Z))$$



The fitted logistic model:

$$Z = -3.777 + 0.11X$$

$$Pr(Y=1) = 1/(1+exp(-Z))$$

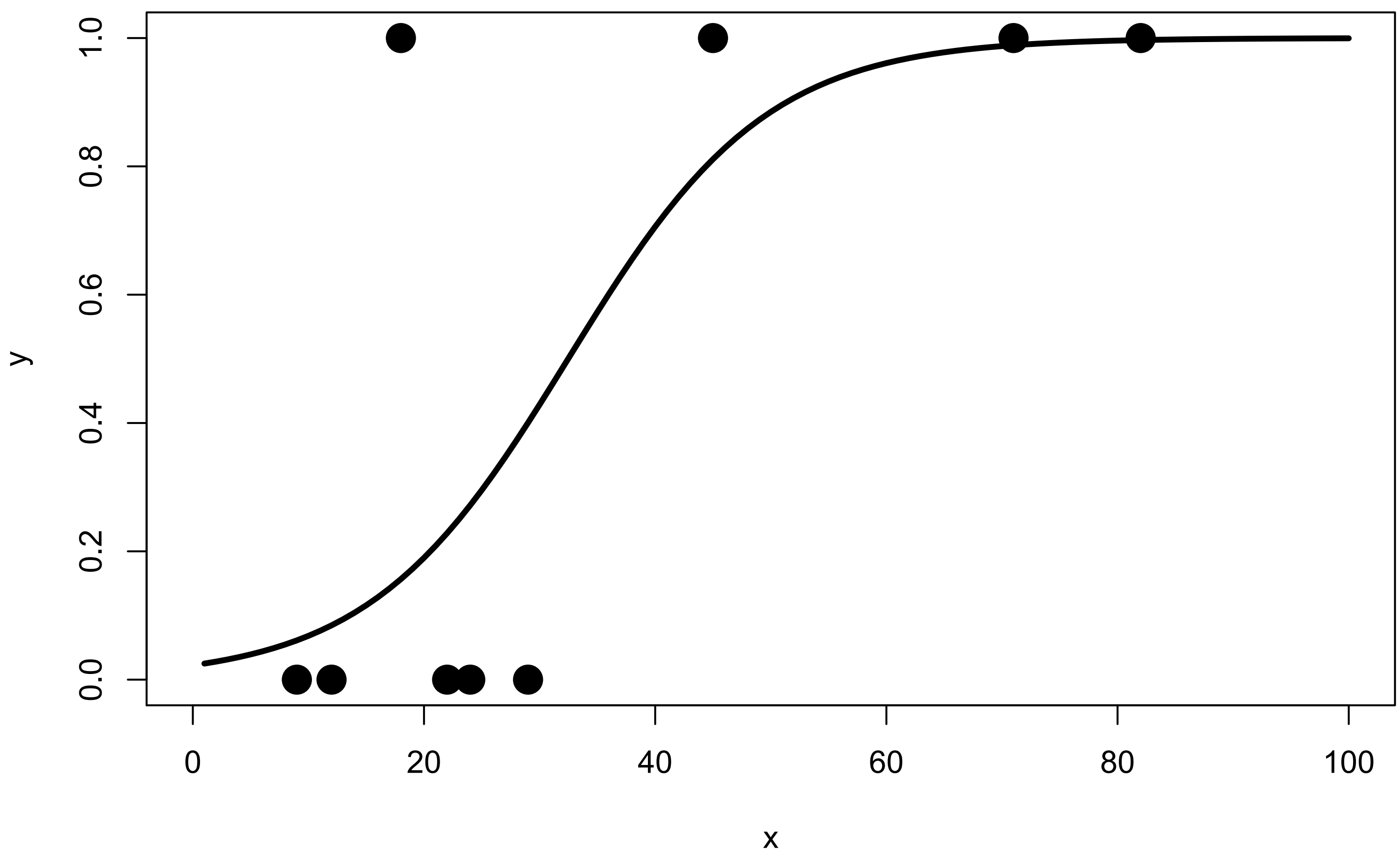
```
> x <- c(82, 45, 71, 22, 29, 9, 12, 18, 24)
> y <- c(1, 1, 1, 0, 0, 0, 0, 1, 0)
> mod1<-glm(y~x, family="binomial")
> summary(mod1)

Call:
glm(formula = y ~ x, family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0115  -0.7200  -0.3554   0.1499   1.9254

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.77725    2.45949  -1.536    0.125
x            0.11633    0.08655   1.344    0.179

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 12.3653  on 8  degrees of freedom
Residual deviance:  6.6337  on 7  degrees of freedom
AIC: 10.634

Number of Fisher Scoring iterations: 6
```

The fitted logistic model:

```
> Z <- cbind(1,1:100)%*%coef(mod1)
> fitted_line <- 1/(1+exp(-Z))
>
> plot(y~x, pch=20, cex=3, xlim=c(0,100), ylim=c(0,1))
> lines(1:100, fitted_line, lwd=3)
```

$$Z = -3.777 + 0.11X$$

$$Pr(Y=1) = 1/(1+exp(-Z))$$



multiple logistic regression

Age vs. Money

PREDICTOR *variables*

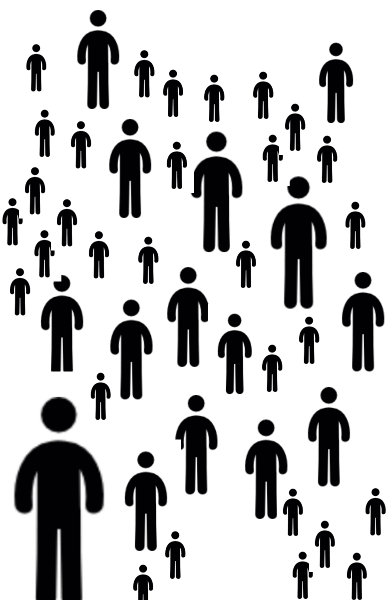
x_1 → Age in Years

x_2 → Student? (Yes/No)

RESPONSE *variable*

Y → Use cash regularly? (yes/no)

Population



Population parameters

$\beta_0, \beta_1, \beta_2$

Hypothesis Test

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

Sample statistics

$b_0 = -10.43$

$b_1 = 0.278$

$b_2 = 4.685$

Null deviance = 12.37

Residual deviance = 4.69

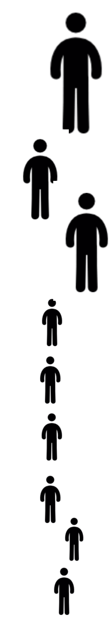
AIC = 10.69

For parameter β_1 :

95% C.I. = [0.03, 1.00]

p-value = 0.198

Sample, n=9



x_1	x_2	y
82	1	1
45	0	1
71	0	1
22	0	0
29	1	0
9	1	0
12	0	0
18	1	1
24	1	0

multiple logistic regression

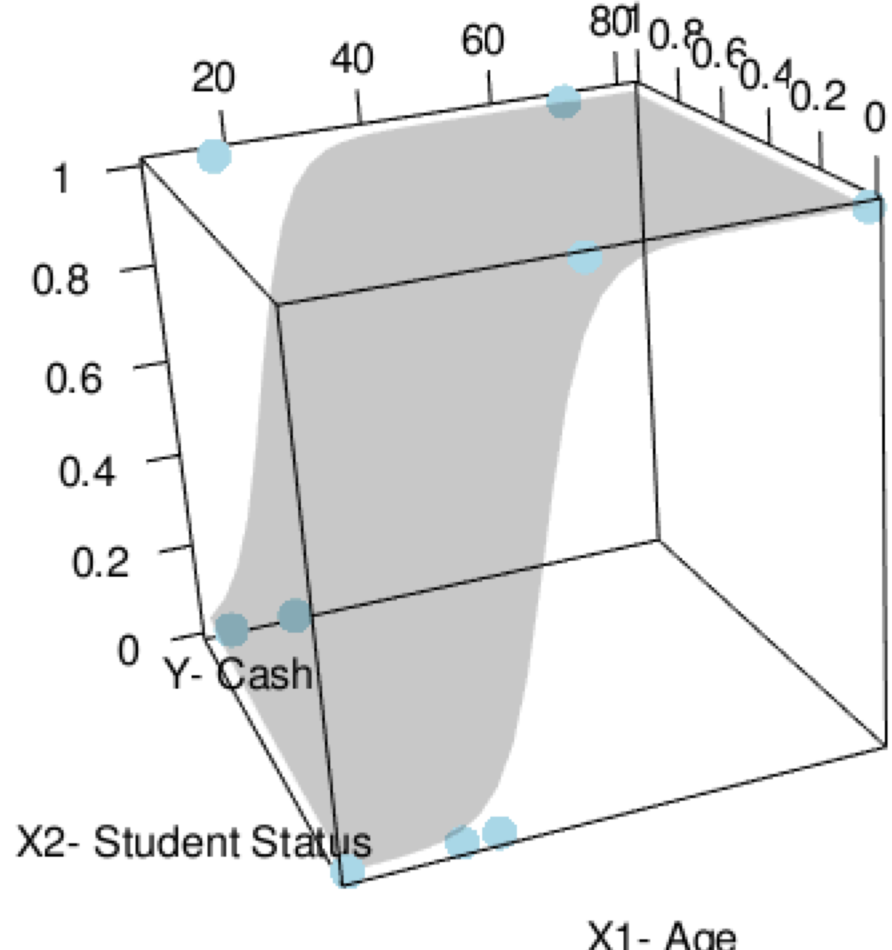
```
> x1 <- c(82, 45, 71, 22, 29, 9, 12, 18, 24)
> x2 <- c(0, 0, 1, 1, 0, 0, 1, 1, 0)
> y <- c(1, 1, 1, 0, 0, 0, 0, 1, 0)
> mod2<-glm(y~x1+x2, family="binomial")
> summary(mod2)

Call:
glm(formula = y ~ x1 + x2, family = "binomial")

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.34142  -0.41561  -0.02688   0.00288   1.50144

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.4309     8.1514  -1.280    0.201
x1            0.2783     0.2163   1.287    0.198
x2            4.6852     4.5674   1.026    0.305

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 12.3653  on 8  degrees of freedom
Residual deviance:  4.6866  on 6  degrees of freedom
AIC: 10.687

Number of Fisher Scoring iterations: 7
```

multiple logistic regression

How do we interpret the β coefficients?

Each estimated coefficient
is the expected change in
the log odds of using cash for a unit
increase in the corresponding
predictor variable holding the other
predictor variables
constant at certain value.

Each exponentiated coefficient is the
ratio of two odds, or the change in odds
in the multiplicative scale for a unit increase
in the corresponding predictor variable holding
other variables at certain value.



https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-how-do-i-interpret-odds-ratios-in-logistic-regression/

multiple logistic regression

Age vs. Money

Interpretation: There is a 32% increase in the odds of carrying cash for every additional year in age, when "student status" is fixed.

Note that exp(0.28) = 1.32 and consider the following:

$$Pr(Y=1|X_1=x_1^*, X_2=x_2) = \frac{1}{1+\exp(-(-10.43+0.28x_1^*+4.69x_2))}$$

$$Pr(Y=0|X=x^*) = 1 - Pr(Y=1|X_1=x_1^*, X_2=x_2)$$

$$odds_{x_1^*,x_2} = \frac{Pr(Y=1|X_1=x_1^*,X_2=x_2)}{Pr(Y=0|X_1=x_1^*,X_2=x_2)} = \exp(-10.43+0.28x_1^*+4.69x_2)$$

$$odds_{x_1^*+1,x_2} = \frac{Pr(Y=1|X_1=x_1^*+1,X_2=x_2)}{Pr(Y=0|X_1=x_1^*+1,X_2=x_2)} = \exp(-10.43+0.28(x_1^*+1)+4.69x_2)$$

$$odds_{x_1^*+1,x_2} = \frac{Pr(Y=1|X_1=x_1^*+1,X_2=x_2)}{Pr(Y=0|X_1=x_1^*+1,X_2=x_2)} = \exp(-10.43+0.28x_1^*+0.28+4.69x_2)$$

$$odds_{x_1^*+1,x_2} = \frac{Pr(Y=1|X_1=x_1^*+1,X_2=x_2)}{Pr(Y=0|X_1=x_1^*+1,X_2=x_2)} = (odds_{x_1^*,x_2})\exp(0.28)$$

$$= (odds_{x_1^*,x_2})1.32$$

multiple logistic regression with interaction effects

For non-students (x2=0), a one-year increase in age yields a change in log odds of 2.5.
On the other hand, for the students (x2=1), a one-year increase in age
yields a change in log odds of (2.5 – 2.4) = 0.1.

```
> x1 <- c(82, 45, 71, 22, 29, 9, 12, 18, 24)
> x2 <- c(0, 0, 1, 1, 0, 0, 1, 1, 0)
> y <- c(1, 1, 1, 0, 0, 0, 0, 1, 0)
> mod3 <- glm(y~x1*x2, family="binomial")
```

multiple logistic regression with interaction effects

For non-students (x2=0), a one-year increase in age yields a change in log odds of 2.5. On the other hand, for the students (x2=1), a one-year increase in age yields a change in log odds of (2.5 – 2.4) = 0.1.

$$Pr(Y=1|X_1=x_1^*,X_2=x_2)=\frac{1}{1+\exp(-(-92.289+2.497x_1^*+89.803x_2-2.394x_1^*x_2))}$$

$$Pr(Y=0|X=x^*)=1-Pr(Y=1|X_1=x_1^*,X_2=x_2)$$

$$odds_{x_1^*,x_2}=\frac{Pr(Y=1|X_1=x_1^*,X_2=x_2)}{Pr(Y=0|X_1=x_1^*,X_2=x_2)}=\exp(-92.3+2.5x_1^*+89.8x_2-2.4x_1^*x_2)$$

$$odds_{x_1^*+1,x_2}=\frac{Pr(Y=1|X_1=x_1^*+1,X_2=x_2)}{Pr(Y=0|X_1=x_1^*+1,X_2=x_2)}=\exp(-92.3+2.5(x_1^*+1)+89.8x_2-2.4(x_1^*+1)x_2)$$

$$odds_{x_1^*+1,x_2}=\exp(-92.3+2.5x_1^*+2.5+89.8x_2-2.4x_1^*x_2-2.4x_2)$$

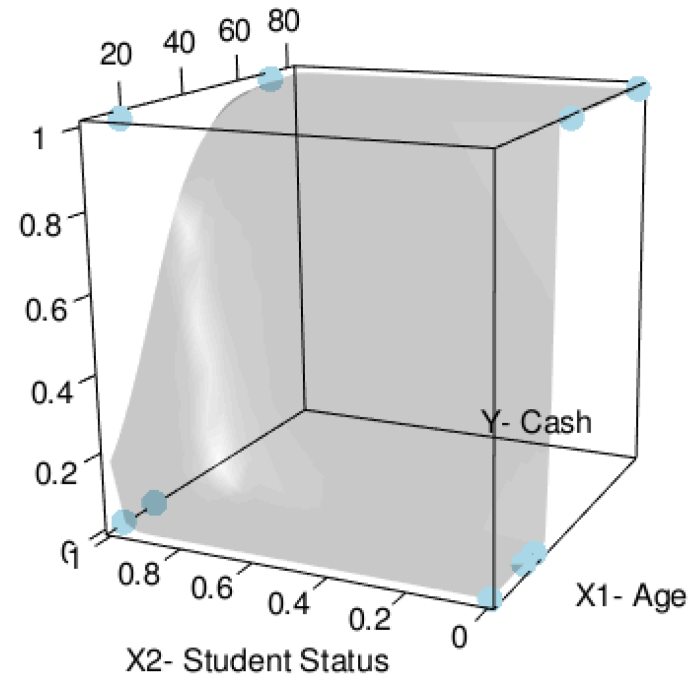
multiple logistic regression with interaction effects

For non-students (x2=0), a one-year increase in age yields a change in log odds of 2.5. On the other hand, for the students (x2=1), a one-year increase in age yields a change in log odds of (2.5 – 2.4) = 0.1.

Notice a few things:

exp(2.5)=12.2
exp(0.1)=1.11

So for a non-student moving from 20 years old to 30 years old the odds of having cash increase from
exp(-92.3)*exp(2.5*20) = 0.000
to
exp(-92.3)*exp(2.5*30) = 0.000



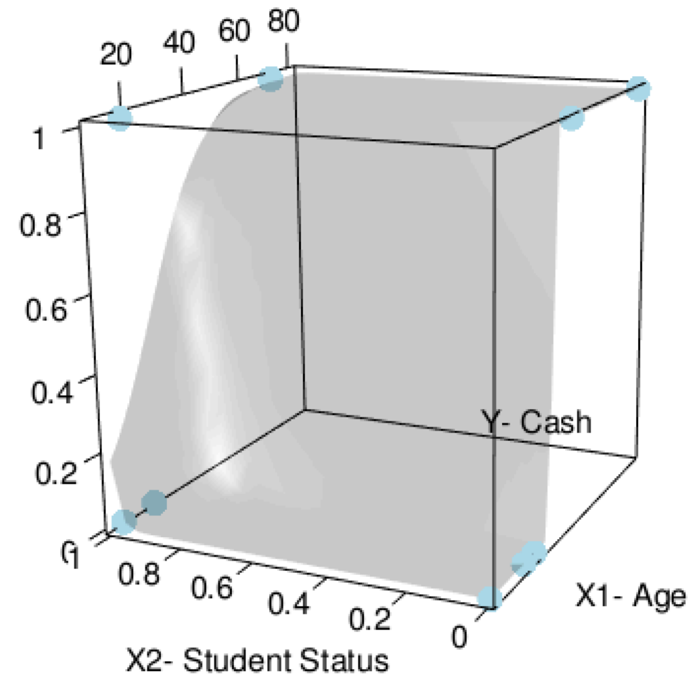
multiple logistic regression with interaction effects

For non-students (x2=0), a one-year increase in age yields a change in log odds of 2.5. On the other hand, for the students (x2=1), a one-year increase in age yields a change in log odds of (2.5 – 2.4) = 0.1.

Notice a few things:

exp(2.5)=12.2
exp(0.1)=1.11

So for a non-student moving from 30 years old to 40 years old the odds of having cash increase from
exp(-92.3)*exp(2.5*30) = 0.000
to
exp(-92.3)*exp(2.5*40) = 1980.293



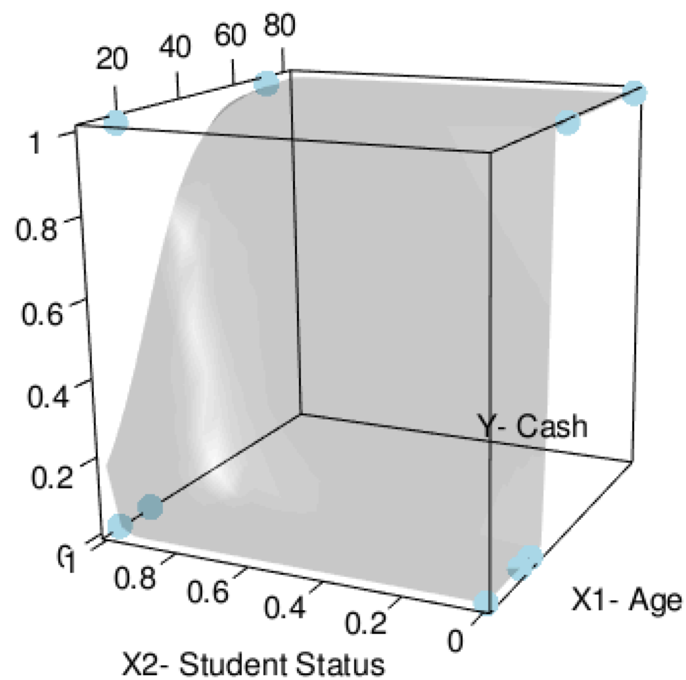
multiple logistic regression with interaction effects

For non-students (x2=0), a one-year increase in age yields a change in log odds of 2.5. On the other hand, for the students (x2=1), a one-year increase in age yields a change in log odds of (2.5 – 2.4) = 0.1.

Notice a few things:

exp(2.5)=12.2
exp(0.1)=1.11

So for a student moving from 30 years old to 40 years old the odds of having cash increase from
exp(-92.3 +89.8)*exp((2.5-2.4)*30) = 1.65
to
exp(-92.3 +89.8)*exp((2.5-2.4)*40) = 4.48



multiple logistic regression with interaction effects

For non-students (x2=0), a one-year increase in age yields a change in log odds of 2.5. On the other hand, for the students (x2=1), a one-year increase in age yields a change in log odds of (2.5 – 2.4) = 0.1.

So for a student moving from 30 years old to 40 years old the odds of having cash increase from
exp(-92.3 +89.8)*exp((2.5-2.4)*30) = 1.65
to
exp(-92.3 +89.8)*exp((2.5-2.4)*40) = 4.48

Pr(30 y.o. student has cash)=
1.65/(1+1.65) = 64%

Pr(40 y.o. student has cash)=
4.48/(1+4.48) = 83%

(small differences are due to rounding.)