



UNIVERSITY OF LIMERICK

## **SO5032 Lecture 9**

---

Brendan Halpin

March 20, 2023

## SO5032 Lecture 9

# SO5032 Lecture 9

---

## Logistic regression

Today we introduce logistic regression: for binary outcomes

See Agresti Ch 15 Sec 1.

# Binary outcomes and regression

- OLS (linear regression) requires an interval dependent variable
- Binary or “yes/no” dependent variables are not suitable
- Nor are rates, e.g.,  $n$  successes out of  $m$  trials

# Problems with OLS

- Errors are distinctly not normal
- While predicted value can be read as a probability, can depart from 0:1 range
- Particular difficulties with multiple explanatory variables
- Nonetheless still often used

# Linear Probability Model

- If we use OLS with binary outcomes, it is called “linear probability model”:

$$Pr(Y = 1) = a + bX$$

- data is 0/1, prediction is probability
- Assumptions violated, but if predicted probabilities in range 0.2–0.8, not too bad

# Credit card example

```
. reg card income
```

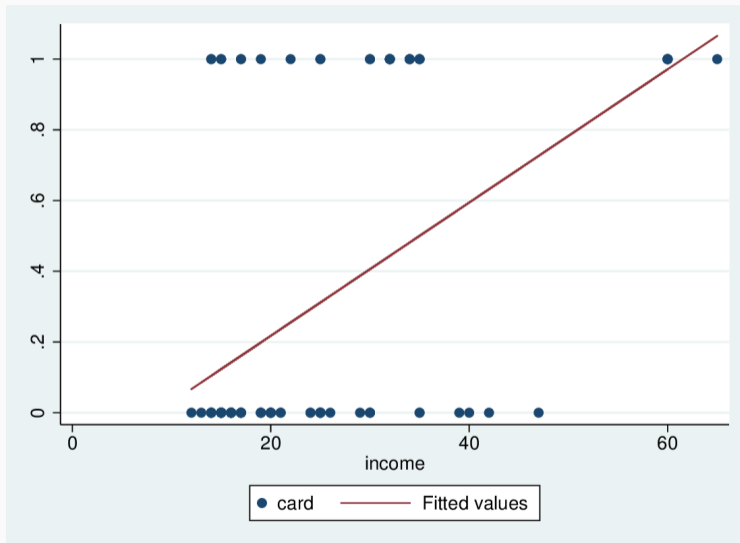
Source	SS	df	MS	Number of obs	=	100
Model	5.55556122	1	5.55556122	F(1, 98)	=	34.38
Residual	15.8344388	98	.161575906	Prob > F	=	0.0000
Total	21.39	99	.216060606	R-squared	=	0.2597
				Adj R-squared	=	0.2522
				Root MSE	=	.40197

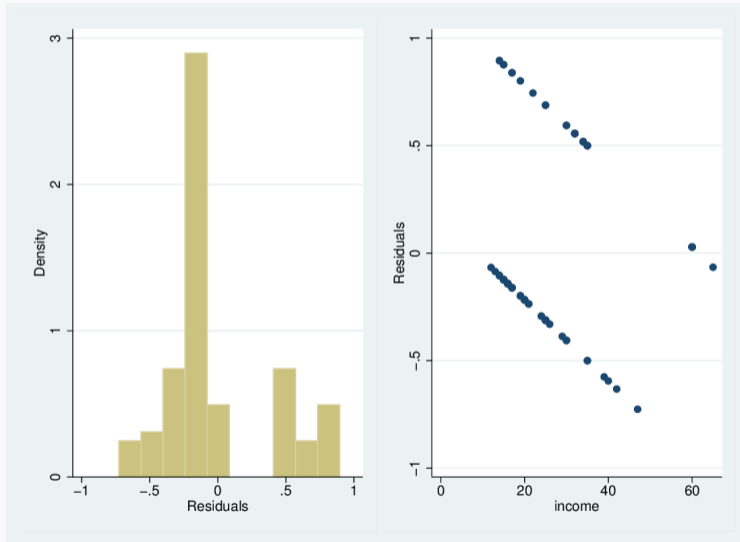
card	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
income	.0188458	.003214	5.86	0.000	.0124678	.0252238
_cons	-.1594495	.089584	-1.78	0.078	-.3372261	.018327



# Credit card example



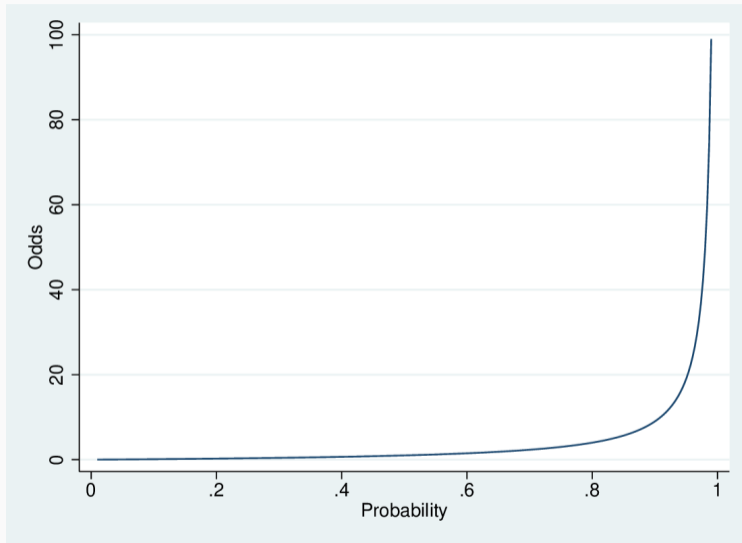
# Credit card example



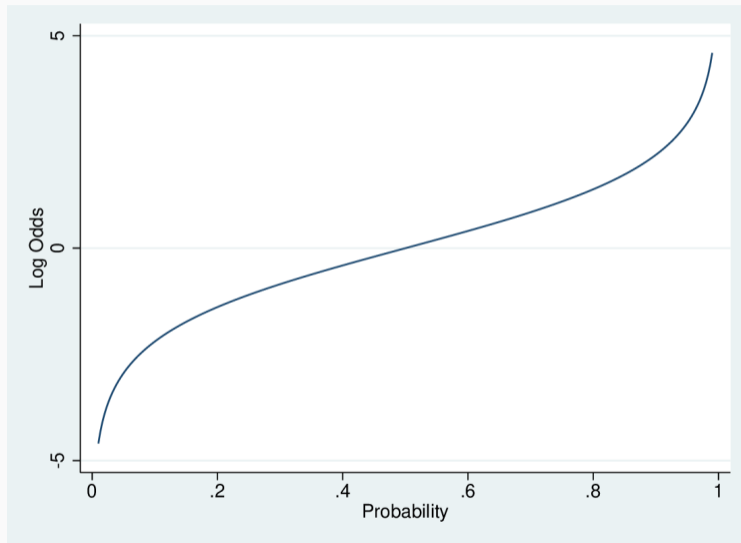
# Logistic transformation

- Probability is bounded [0 : 1]
- OLS predicted value is unbounded
- How to transform probability to  $-\infty : \infty$  range?
- Odds:  $\frac{p}{1-p}$  – range is 0 :  $\infty$
- Log of odds:  $\log \frac{p}{1-p}$  has range  $-\infty : \infty$

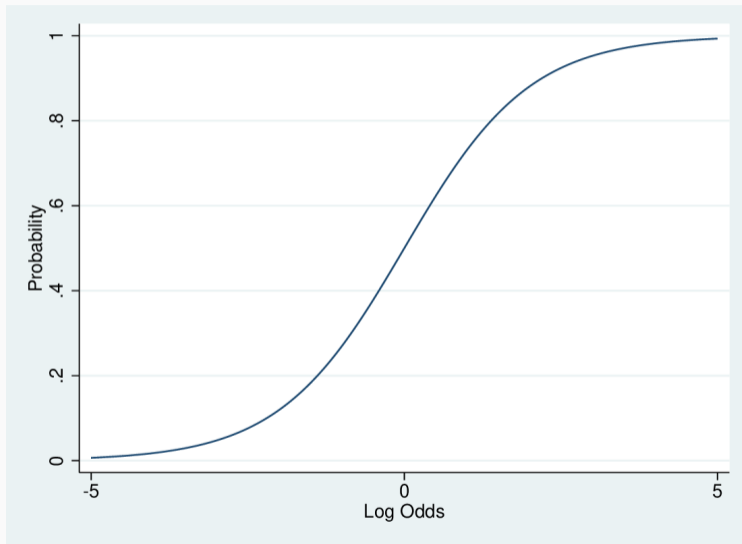
# Probability to odds



# Probability to log-odds



## Rotated: the "S-shaped" curve



- Logistic regression uses this as the dependent variable:

$$\log \left( \frac{p}{1-p} \right) = a + bX$$

# Alternatives

We can look at this in three ways

- In terms of log-odds:

$$\log \left( \frac{\Pr(Y = 1)}{1 - \Pr(Y = 1)} \right) = a + bX$$

- In terms of odds:

$$\frac{\Pr(Y = 1)}{1 - \Pr(Y = 1)} = e^{a+bX}$$

- In terms of probability:

$$\Pr(Y = 1) = \frac{e^{a+bX}}{1 + e^{a+bX}} = \frac{1}{1 + e^{-a-bX}}$$



- The b parameter is the effect of a unit change in X on  $\log \left( \frac{Pr(Y=1)}{1-Pr(Y=1)} \right)$
- This implies a multiplicative change of  $e^b$  in  $\frac{Pr(Y=1)}{1-Pr(Y=1)}$ , in the Odds
- Thus an odds ratio
- But the effect of b on P depends on the level of b

# Credit card logistic regression

```
. logit card income
```

```
Iteration 0:  log likelihood = -61.910066  
Iteration 1:  log likelihood = -48.707265  
Iteration 2:  log likelihood = -48.613215  
Iteration 3:  log likelihood = -48.61304  
Iteration 4:  log likelihood = -48.61304
```

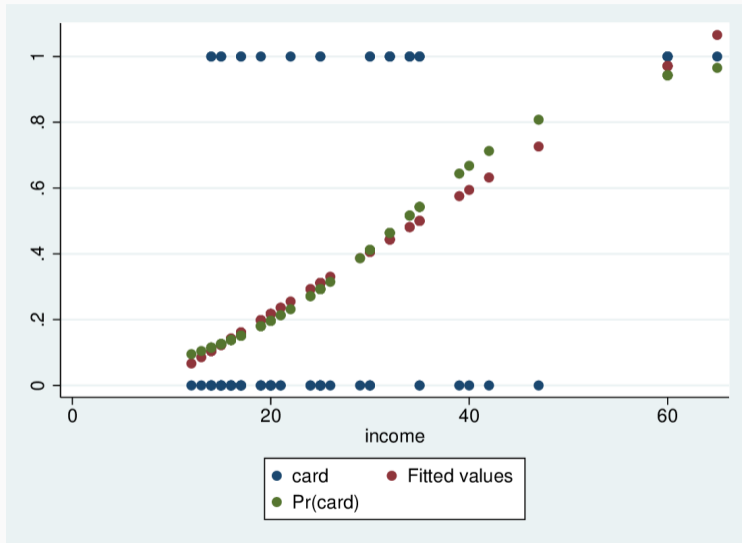
```
Logistic regression
```

```
Number of obs      =      100  
LR chi2(1)         =      26.59  
Prob > chi2        =      0.0000  
Pseudo R2         =      0.2148
```

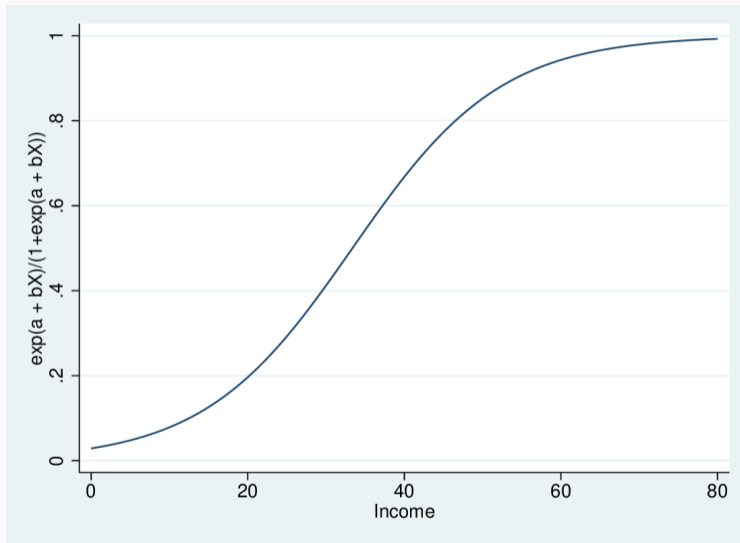
```
Log likelihood = -48.61304
```

card	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
income	.1054089	.0261574	4.03	0.000	.0541413	.1566765
_cons	-3.517947	.7103358	-4.95	0.000	-4.910179	-2.125714

# Credit card logistic regression



# Sigmoid curve from $a+bX$



## Calculating predicted probabilities by hand

- We can calculate the predicted probability for any combination of values of the independent variables
- First, plug them into the  $a + bX$  part to get the predicted log-odds
- Then take the anti-log of the log-odds to get the odds
- Then  $\text{odds}/(1+\text{odds})$  gives us the probability

# Calculating predicted probabilities

- Example:  $\log(\text{odds}) = 0.25 + 0.12X$
- Predict for  $X = 10$ 
  - Predicted log-odds =  $0.25 + 0.12 \cdot 10 = 1.45$
  - Predicted odds =  $e^{1.45} = 4.263$
  - Predicted probability =  $4.263 / (1 + 4.263) = 0.810$

# Web applet for practicing

`http://teaching.sociology.ul.ie:3838/logabx/`