# SO5032 Quantitative Research Methods

Brendan Halpin, Sociology, University of Limerick

Spring 2018

# Logistic regression

- OLS regression requires interval dependent variable
- Binary or "yes/no" dependent variables are not suitable
- Nor are rates, e.g., n successes out of m trials
- 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.

# Linear Probability Model

- OLS gives the "linear probability model" in this case:

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

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

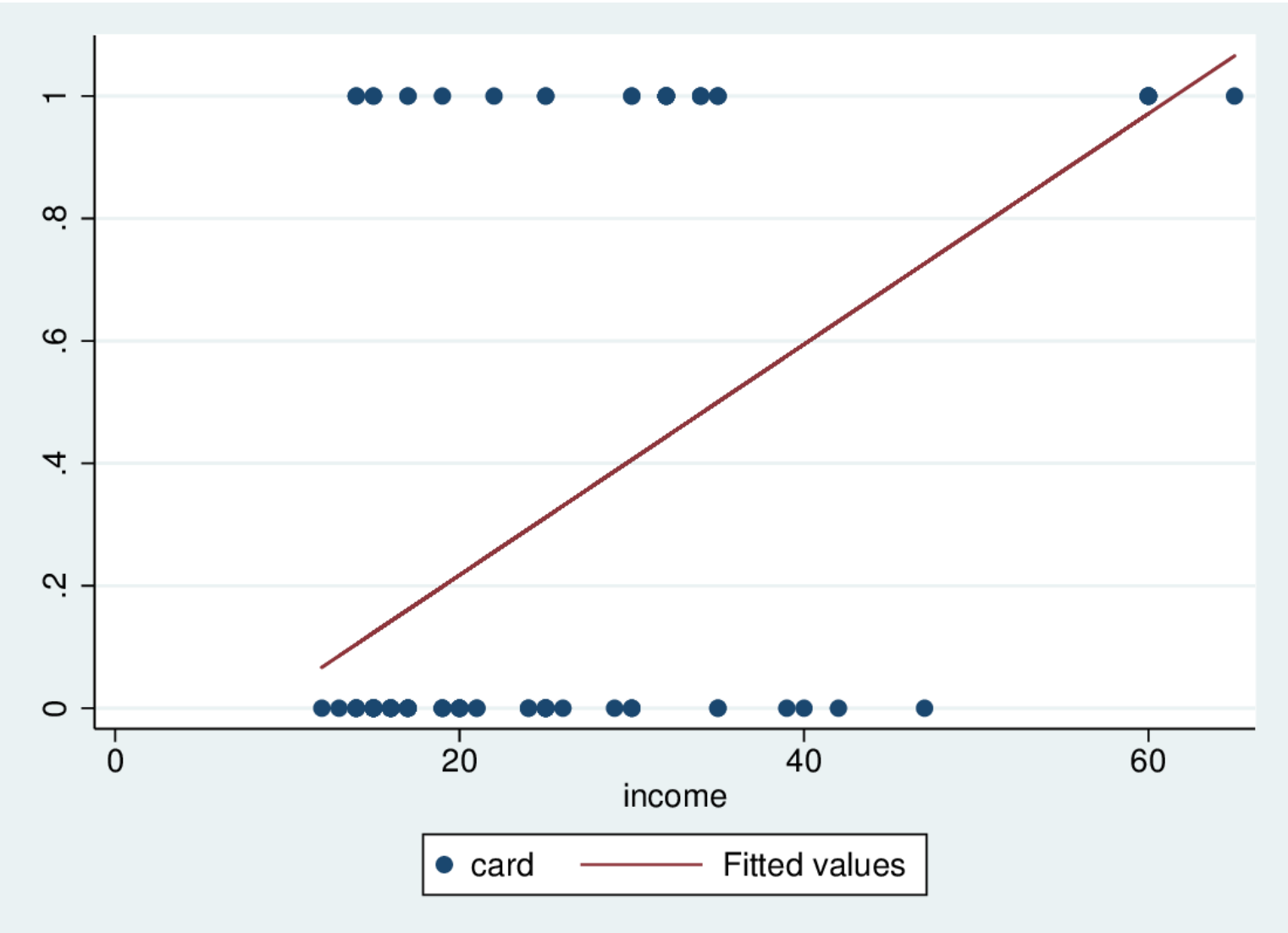
# Credit card example

```
. reg card income
```

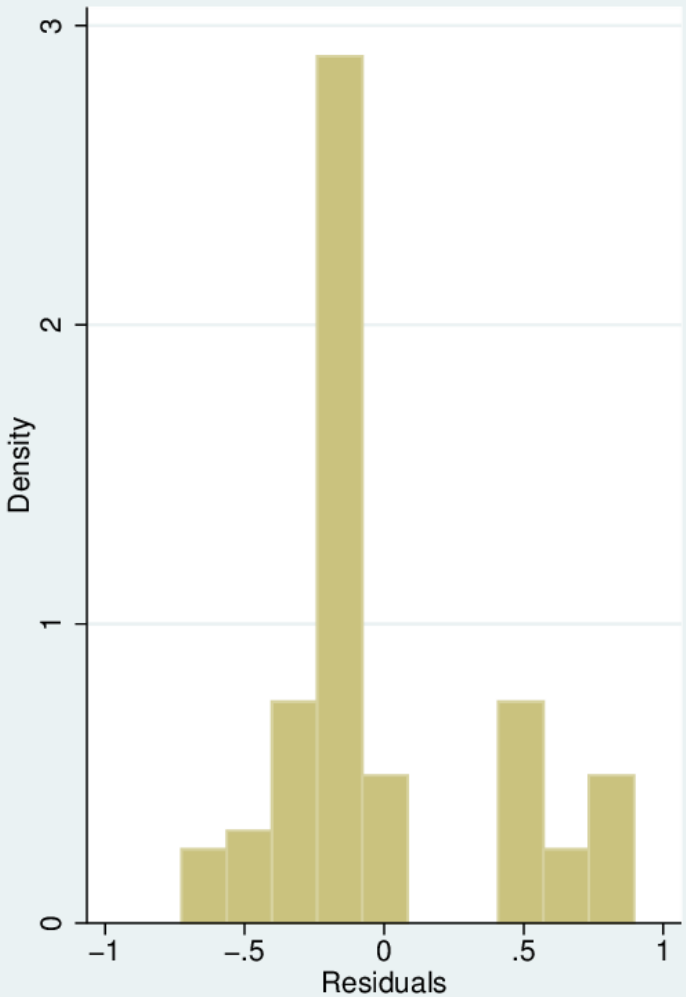
| Source | SS | df | MS | Number of obs | = | 100 |
|---|---|---|---|---|---|---|
| | | | | F(1, 98) | = | 34.38 |
| Model | 5.55556122 | 1 | 5.55556122 | Prob > F | = | 0.0000 |
| Residual | 15.8344388 | 98 | .161575906 | R-squared | = | 0.2597 |
| | | | | Adj R-squared | = | 0.2522 |
| Total | 21.39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

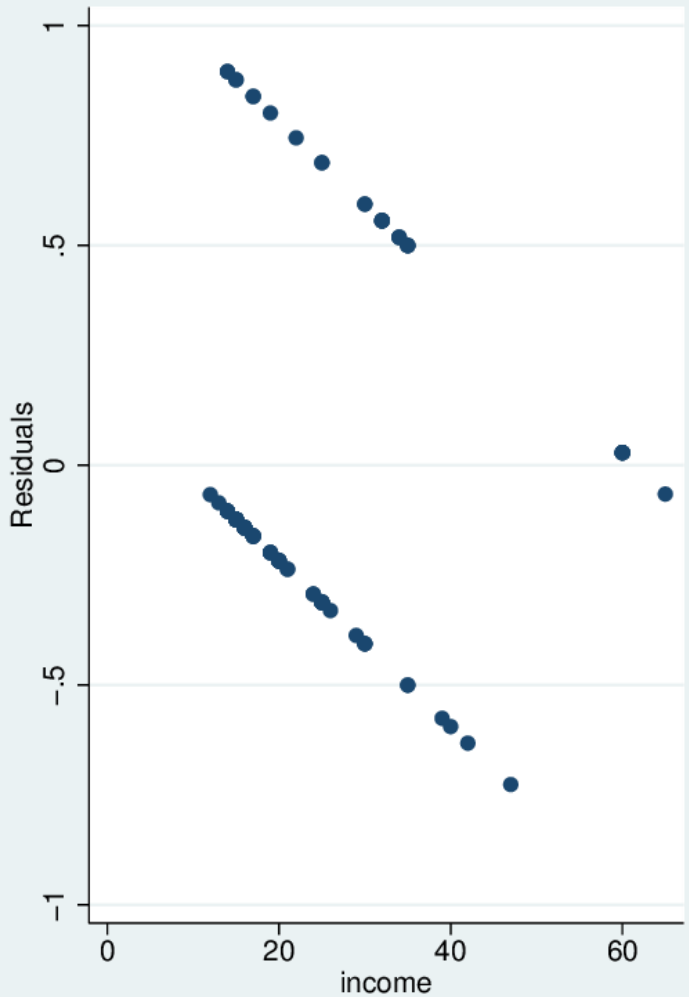
| card | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| income | .0188458 | .003214 | 5.86 | 0.000 | .0124678 | .0252238 |
| _cons | -.1594495 | .089584 | -1.78 | 0.078 | -.3372261 | .018327 |

# Credit card example



# Credit card example

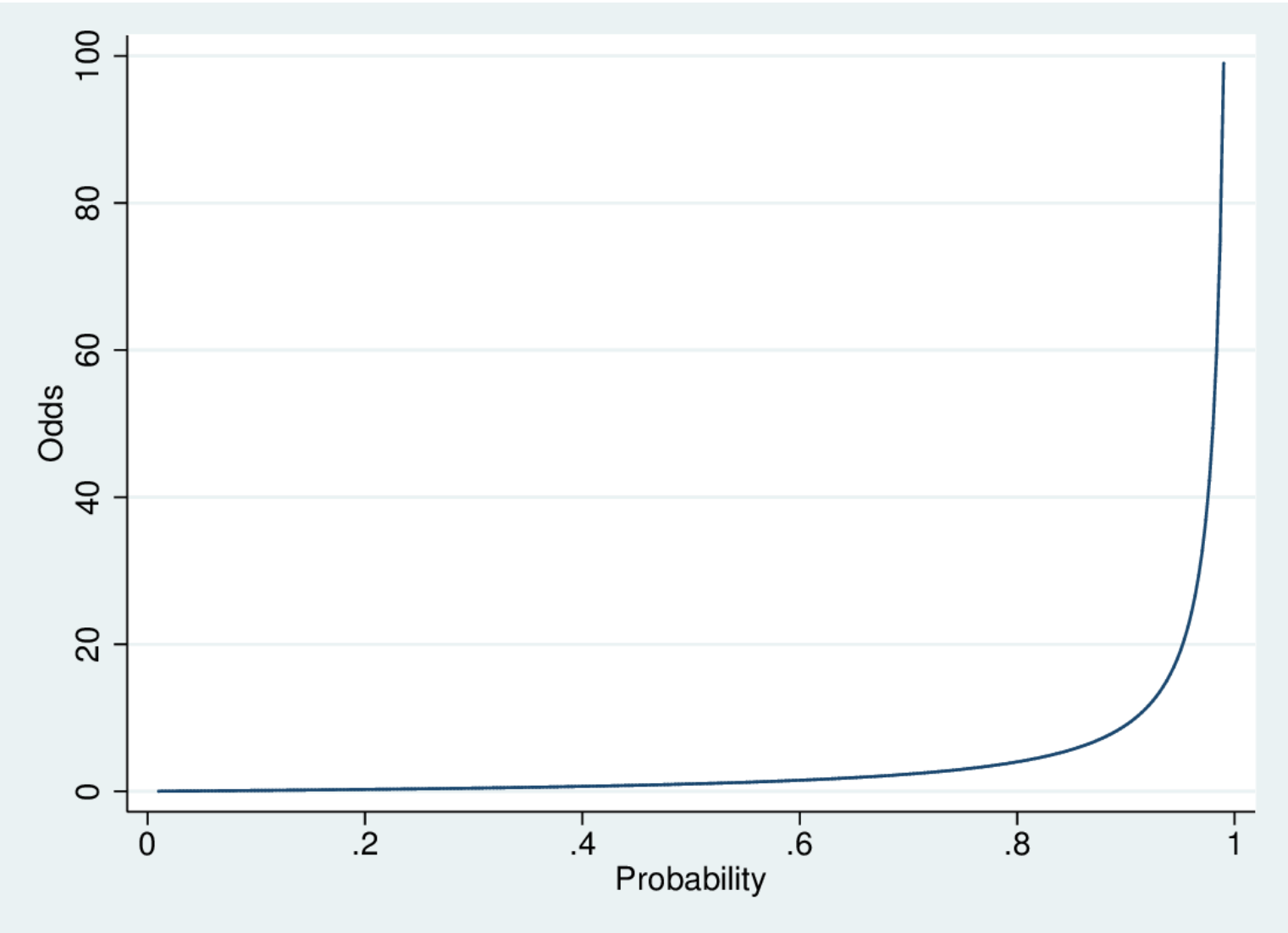


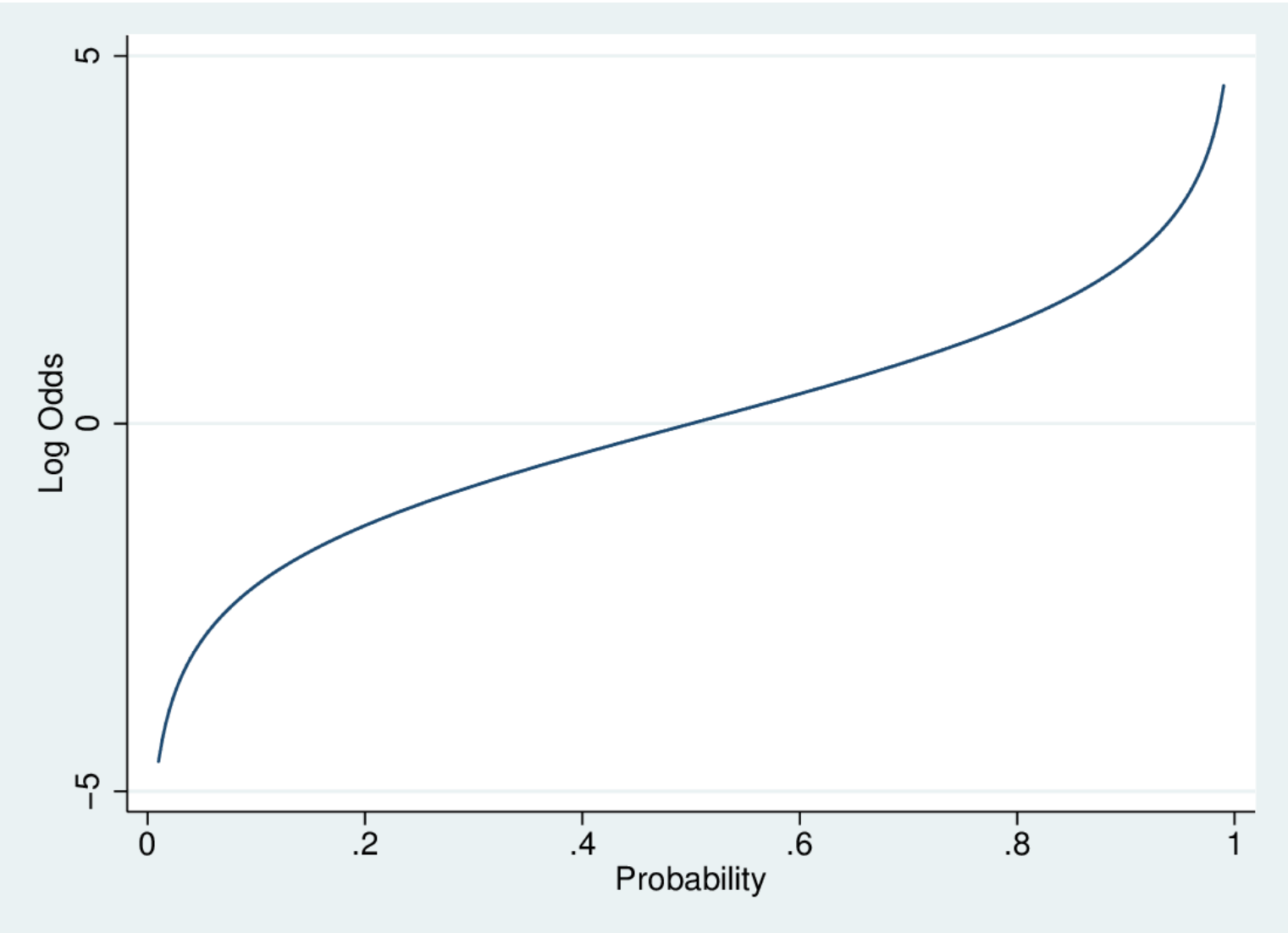
# Logistic transformation

- 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



# Logistic regression

- Logistic regression uses this as the dependent variable:

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

- Alternatively:

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

- Or:

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

# Parameters

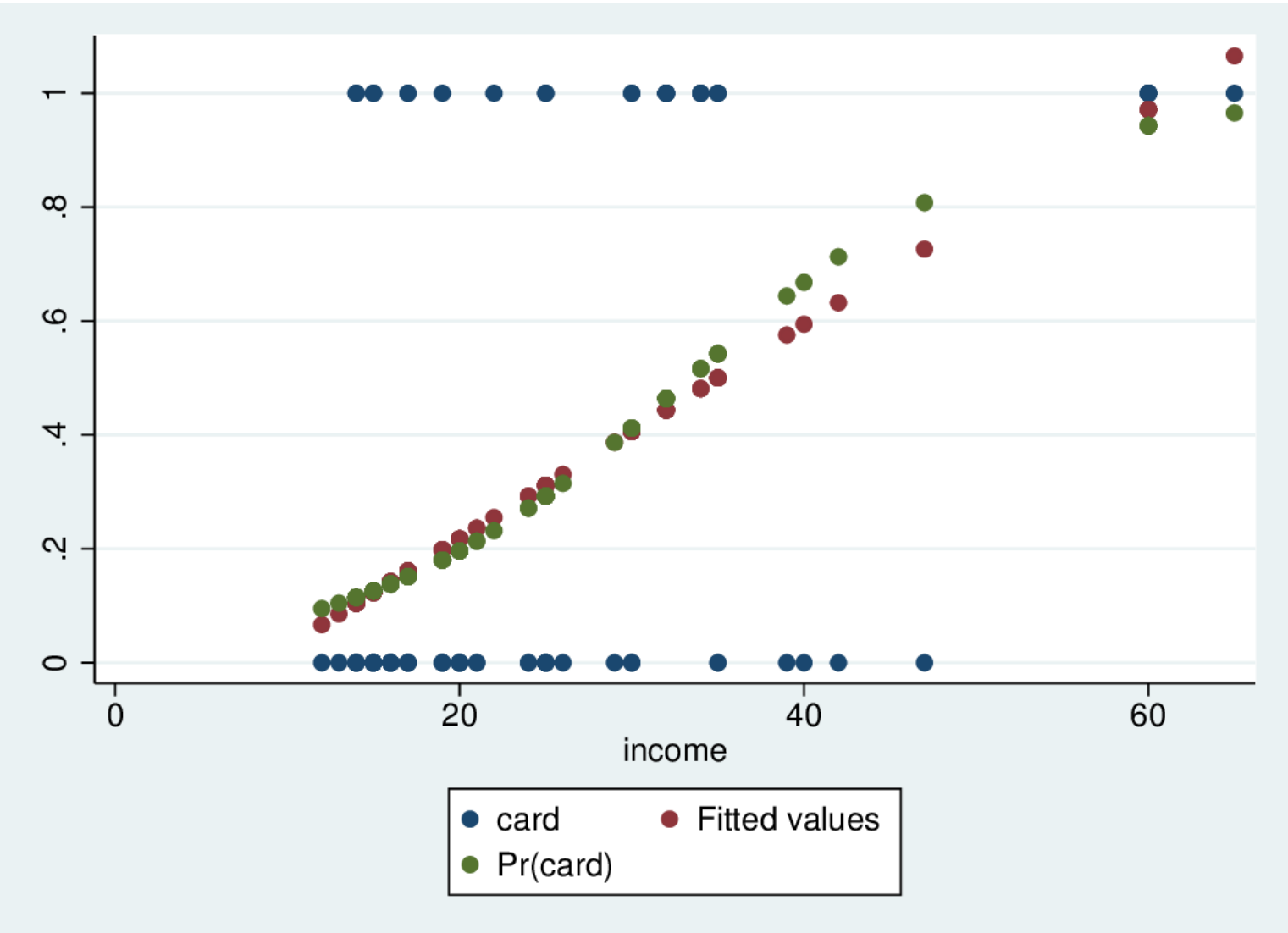
- The b parameter is the effect of a unit change in X on $\log\left(\frac{Pr(Y=1)}{1-Pr(Y=1)}\right)$
- This implies a multiplicative change of $e^b$ in $\frac{Pr(Y=1)}{1-Pr(Y=1)}$, in the Odds
- Thus an odds ratio
- But the effect of b on P depends on the level of b

# 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
Log likelihood =  -48.61304                       Pseudo R2         =     0.2148
```

| 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



# Inference

- In practice, inference is similar to OLS though based on a different logic
- For each explanatory variable, $H_0 : \beta = 0$ is the interesting null
- $z = \frac{\hat{\beta}}{SE}$ is approximately normally distributed (large sample property)
- More usually, the Wald test is used: $\left(\frac{\hat{\beta}}{SE}\right)^2$ has a $\chi^2$ distribution with one degree of freedom

# Likelihood ratio tests

- The "likelihood ratio" test is thought more robust than the Wald test for smaller samples
- Where $l_0$ is the likelihood of the model without $X_j$, and $l_1$ that with it, the quantity

$$-2\left(\log\frac{l_0}{l_1}\right) = -2\left(\log l_0 - \log l_1\right)$$

is $\chi^2$ distributed with one degree of freedom

# Nested models

- More generally, $-2\left(\log \frac{l_o}{l_1}\right)$ tests nested models: where model 1 contains all the variables in model 0, plus $m$ extra ones, it tests the null that all the extra $\beta$ coefficients are zero ($\chi^2$ with $m$ df)
- If we compare a model against the null model (no explanatory variables, it tests

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$$

- Strong analogy with $F$ test in OLS

# Maximum likelihood estimation

- What is this "likelihood"?
- Unlike OLS, logistic regression (and many, many other models) are extimated by *maximum likelihood estimation*
- In general this works by choosing values for the parameter estimates which maximise the probability (likelihood) of observing the actual data
- OLS can be ML estimated, and yields exactly the same results

# Iterative search

- Sometimes the values can be chosen analytically
  - A likelihood function is written, defining the probability of observing the actual data given parameter estimates
  - Differential calculus derives the values of the parameters that maximise the likelihood, for a given data set
- Often, such "closed form solutions" are not possible, and the values for the parameters are chosen by a systematic computerised search (multiple iterations)
- Extremely flexible, allows estimation of a vast range of complex models within a single framework

# Likelihood as a quantity

- Either way, a given model yields a specific maximum likelihood for a give data set
- This is a probability, henced bounded $[0 : 1]$
- Reported as log-likelihood, hence bounded $[-\infty : 0]$
- Thus is usually a large negative number
- Where an iterative solution is used, likelihood at each stage is usually reported – *normally* getting nearer 0 at each step

# Tabular data

- If all the explanatory variables are categorical (or have few fixed values) your data set can be represented as a table
- If we think of it as a table where each cell contains $n$ yeses and $m - n$ noes ($n$ successes out of $m$ trials) we can fit grouped logistic regression
- $n$ successes out of $m$ trials implies a binomial distribution of degree $m$

$$\log \frac{n}{m-n} = \alpha + \beta X$$

- The parameter estimates will be exactly the same as if the data were treated individually

# Tabular data and goodness of fit

- But unlike with individual data, we can calculate goodness of fit, by relating observed successes to predicted in each cell
- If these are close we cannot reject the null hypothesis that the model is incorrect (i.e., you want a high p-value)
- Where $l_i$ is the likelihood of the current model, and $l_s$ is the likelihood of the "saturated model" the test statistic is

$$-2\left(\log\frac{l_i}{l_s}\right)$$

- The saturated model predicts perfectly and has as many parameters as there are "settings" (cells in the table)
- The test has *df* of number of settings less number of parameters estimated, and is $\chi^2$ distributed

# Fit with individual data

- Where the number of "settings" (combinations of values of explanatory variables) is large, this approach to fit is not feasible
- Cannot be used with continuous covariates
- Hosmer-Lemeshow statistic attempts to create an analogy
  - Divide sample into deciles of predicted probability
  - Calculate a fit measure based on observed and predicted numbers in the ten groups
  - Simulation shows this is $\chi^2$ distributed with 2 df
  - Not a perfect solution, sensitive to how the cuts are made
- Pseudo-$R^2$ measures exist, but none approaches the clean interpretation as in OLS
- See `http://www.ats.ucla.edu/stat/mult_pkg/faq/general/Psuedo_RSquareds.htm`

# Predicting outcomes

- Another way of assessing the adequacy of a logit model is its accuracy of classification:

| | True yes | True no |
|---|---|---|
| Predicted yes | a | c |
| Predicted no | b | d |

- Proportion correctly classified: $\frac{a+d}{a+b+c+d}$
- Sensitivity: $\frac{a}{a+b}$; Specificity: $\frac{d}{c+d}$
- False positive: $\frac{c}{a+c}$; False negative: $\frac{b}{b+d}$
- Stata: `estat class`

# Some problems

- Zero cells in tables can cause problems: no yeses or no noes for particular settings
- Not automatically a problem but can give rise to attempts to estimate a parameter as $-\infty$ or $+\infty$
- If this happens, you will see a large parameter estimate and a huge standard error
- In individual data, sometimes certain combinations of variables have only successes or only failures
- In Stata, these cases are dropped from estimation – you need to be aware of this as it changes the interpretation (you may wish to drop one of the offending variables instead)