Multiple regression: Categorical dependent variables

Johan A. Elkink
School of Politics & International Relations
University College Dublin

28 November 2016

Outline

1. Binary dependent variables
2. Logistic regression
3. Interpretation
4. Model fit

Binary models

Binary models have a dependent variable consisting of two categories.

For example,

- Vote on a particular law
- Turning out in an election
- Approval in a referendum
- Bankrupt or not

Limited dependent variables

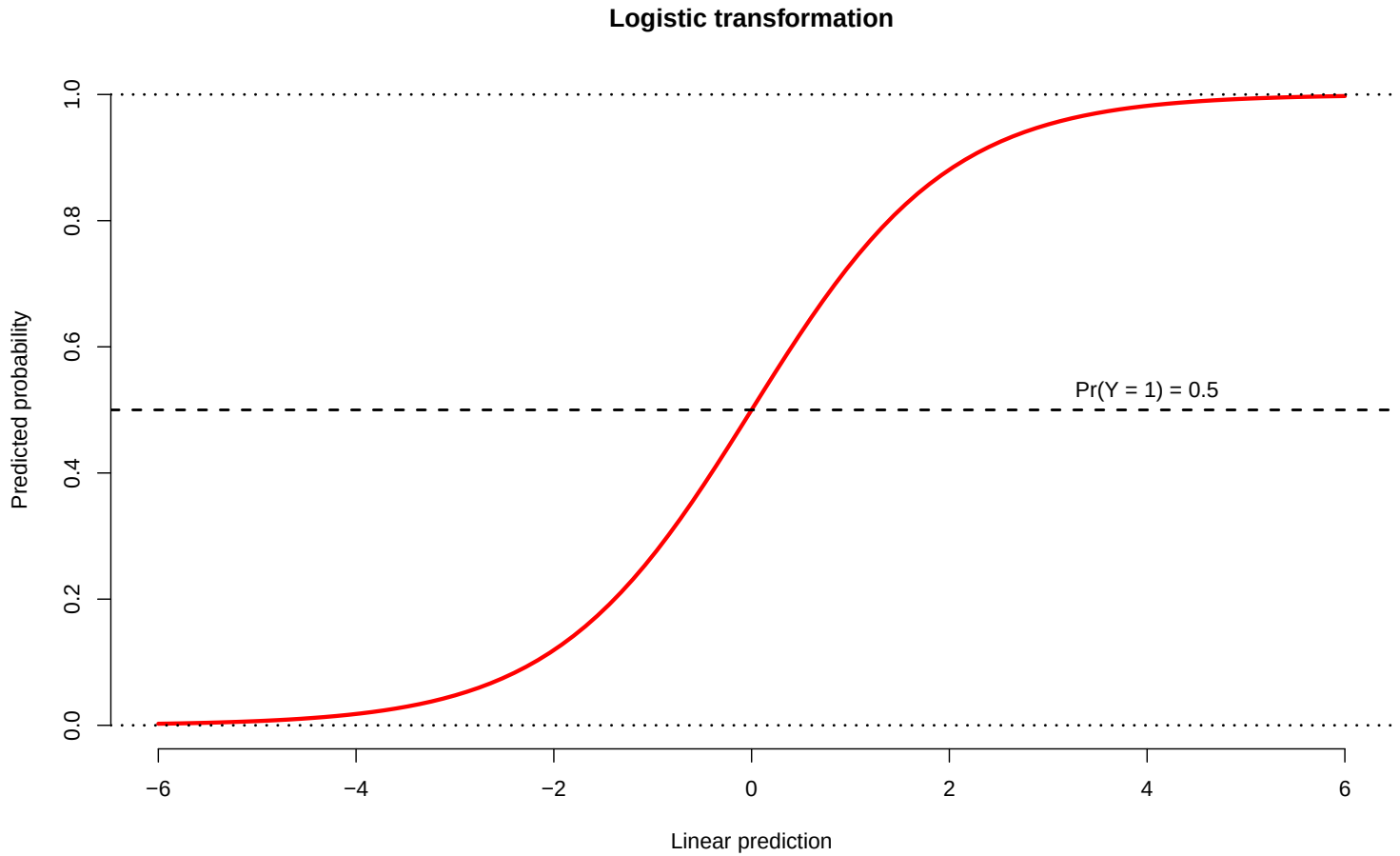
When a dependent variable is not continuous, or is truncated for some reason, a linear model would lead to implausible predictions.

For binary dependent variables we estimate the **probability of observing a one**:

- Prediction below 0 and above 1 would not make sense.
- For any case where the predicted probability is already high, it cannot increase much with a change in $\boldsymbol{X}$ (and vice versa for low probabilities).
- A linear model would imply high levels of **heteroskedasticity**.

Estimators

A typical approach is to have an estimator that is "linear in the parameters" – i.e. it generates a linear prediction based on $\boldsymbol{X}$ and $\boldsymbol{\beta}$ – but then transforms this linear prediction into one bounded between 0 and 1.



Outline

Logistic regression

The most common transformation is the logistic transformation, which relates to the log-odds:

$$\log\left(\frac{Pr(y_i = 1)}{Pr(y_i = 0)}\right) = \beta_1 + \beta_2 x_{i1} + \beta_3 x_{i2},$$

which can also be formulated as:

$$Pr(y_i = 1) = \frac{1}{1 + e^{-(\beta_1 + \beta_2 x_{i1} + \beta_3 x_{i2})}}.$$

Estimating a logistic regression

Estimating a logistic regression is straightforward and output will look similar to that of linear regression.

E.g. explaining "Yes" in the Marriage Equality Referendum.

Note the use of continuous and discreet independent variables.

Age 25-34	-0.152
	(0.410)
35-44	-0.707^{*}
	(0.386)
45-54	-0.865^{**}
	(0.390)
55-64	-1.084^{***}
	(0.399)
65+	-1.857^{***}
	(0.374)
Urban	0.305^{*}
	(0.168)
Pro-abortion attitude	0.221^{***}
	(0.028)
intercept	0.358
	(0.372)
N	851

Outline

Derivatives

For linear regression, we interpret using the first derivative – i.e. the effect of X on Y is:

$$\frac{\partial \hat{y}}{\partial \mathbf{x}_j} = \beta_j.$$

In a logistic regression, however, the derivative is more complicated:

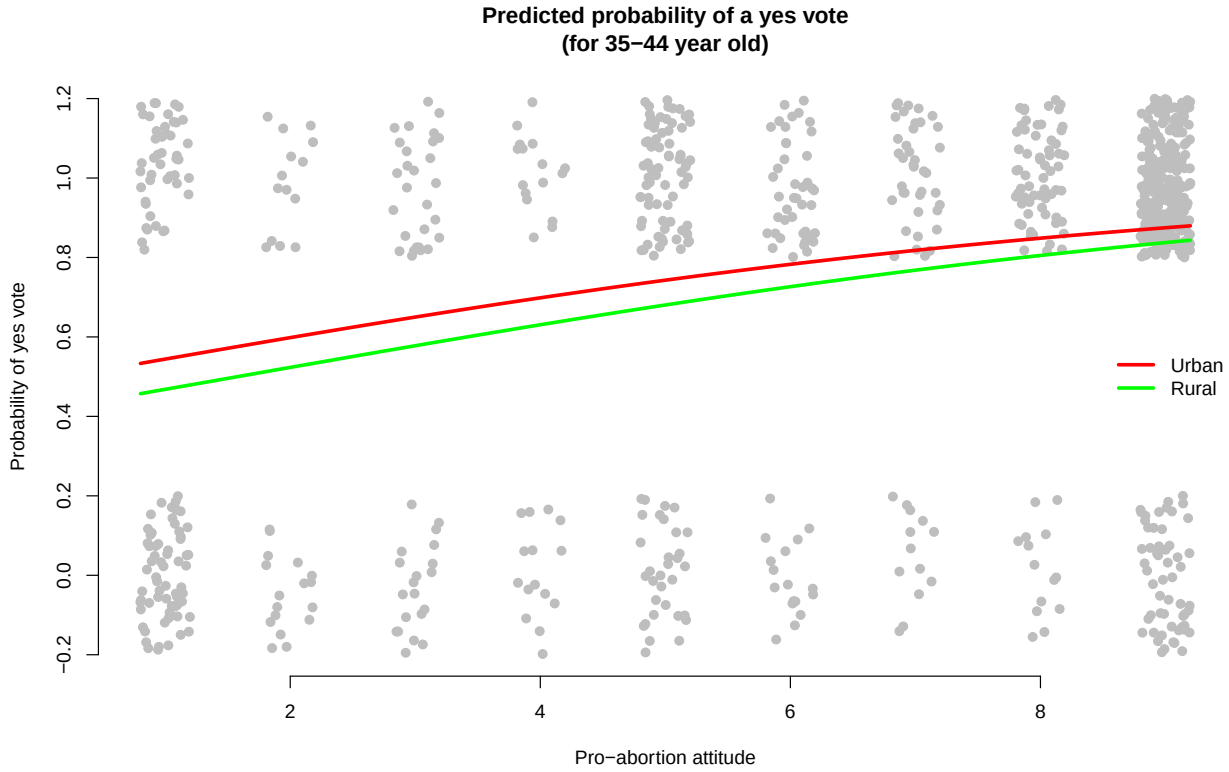
$$\frac{\partial \hat{\pi}}{\partial \mathbf{x}_j} = \beta_j \hat{\pi}(1 - \hat{\pi}).$$

Because of the non-linear relationship, the effect of X on Y depends on all other independent variables.

Nevertheless, a quick method to interpret logit coefficients is to divide them by 4 to get the slope at $\hat{\pi} = 0.5$.

Graphical interpretation

An alternative method is to plot the relationship between one **x** and π, holding the other values of **X** constant (e.g. at the mean, median, etc.).



Because the **link function** $g(\mathbf{X}\beta)$ is not linear (but instead $g(\mathbf{X}\beta) = \frac{1}{1+e^{-\mathbf{X}\beta}}$), the effect of **X** on **y** depends on all **X**.

Fitted values

A third useful way of interpreting **logit** regression coefficients is by describing typical cases or interesting examples.

Age	Region	$P(Yesvote)$
18–24	Urban	0.89
35–44	Urban	0.79
65+	Urban	0.59
18–24	Rural	0.85
35–44	Rural	0.74
65+	Rural	0.51

(This assumes attitude towards abortion at the median value.)

Presentation

Bottom line: it is much better to present *interpretable and understandable* inferences, with an indication of the level of *uncertainty*, than to present simply estimated coefficients.

E.g. "An increase in automobile support for a Republican senator from \$10000 to \$20000 in total increases his or her probability to vote for the Corporate Average Fuel Economy standard bill by 11%, give or take 7%, all else equal."

Outline

R^2 for logistic regression

Although various authors have proposed pseudo-R^2 estimators that roughly do the same thing as an R^2 for linear regression, there is no good alternative.

They cannot be interpreted as "the proportion of variance in Y explained."

Instead, it is typically better to look at the quality of the predictions – do I get high $Pr(Y = 1)$ for the observed ones in the data and low $Pr(Y = 1)$ for the observed zeros?

Confusion matrix

Evaluating the performance of the binary model can be done by using the **confusion matrix**:

		True value 1	True value 0
Prediction	1	True positive	False positive (Type I error)
Prediction	0	False negative (Type II error)	True negative

Confusion matrix

Evaluating the performance of the binary model can be done by using the **confusion matrix**:

		True value 1	True value 0	
Prediction	1	True positive	False positive (Type I error)	Precision: $\frac{TP}{TP+FP}$
	0	False negative (Type II error)	True negative	
		Sensitivity: $\frac{TP}{TP+FN}$	Specificity: $\frac{TN}{FP+TN}$	Accuracy: $\frac{TP+TN}{N}$
		TPR: $\frac{TP}{TP+FN}$	FPR: $\frac{FP}{FP+TN}$	

Receiver Operating Characteristic curve

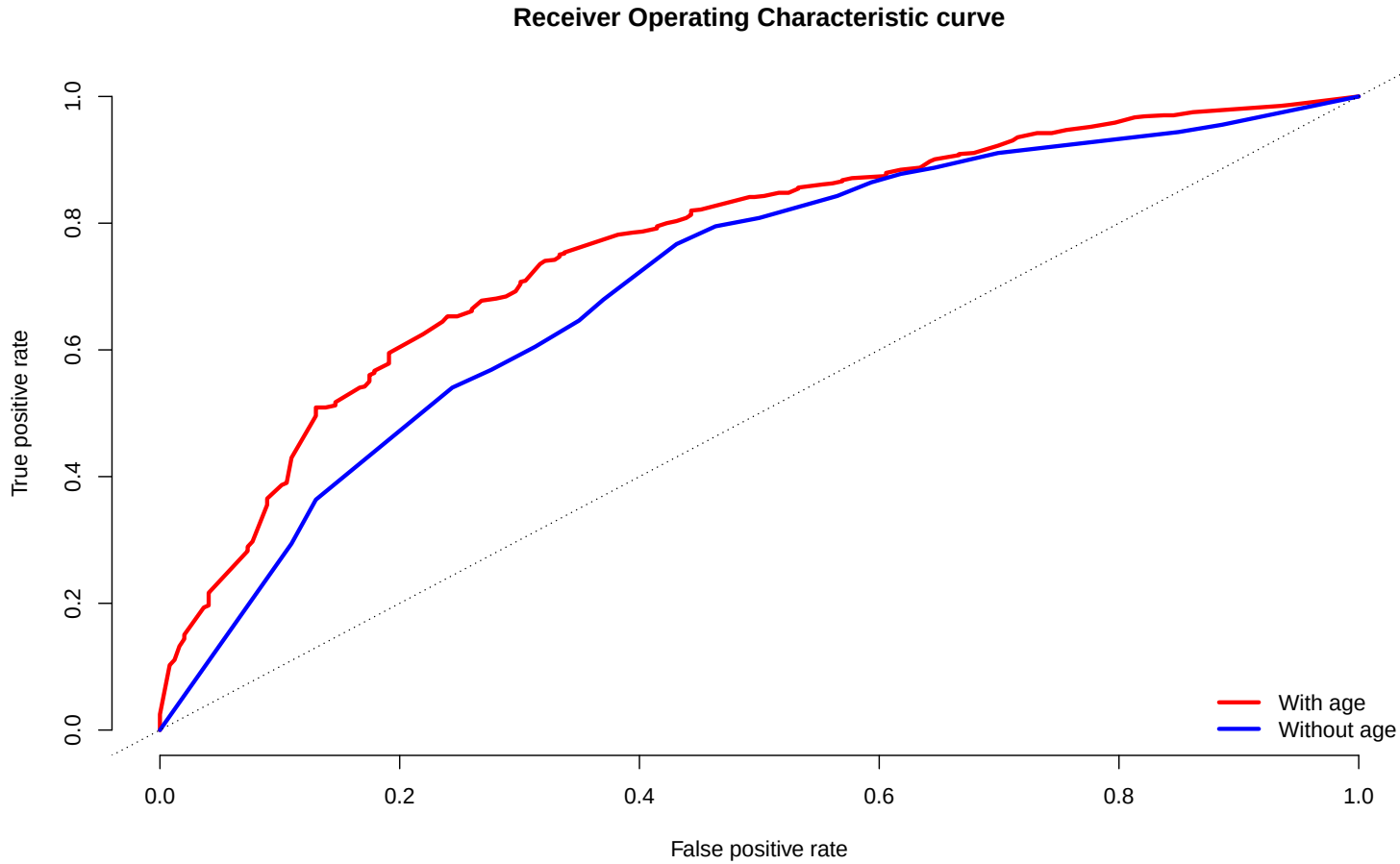
The accuracy of predictions will depend on the threshold probability – variations on default of $\hat{\pi} = 0.5$ are possible.

Depending on the application, it might be better or worse to over- or underestimate ones relative to zeros.

The ROC-curve plots, for all possible thresholds, the true positive rate against the false positive rate.

An ROC-curve further from (above) the 45 degree line indicates a better predictive performance; any predictions under this line indicate worse than random prediction.

Receiver Operating Characteristic curve



Given the above, we can also calculate the area under the ROC-curve as a measure of prediction quality, called AUC. This is somewhat related to the Gini coefficient for income distributions ($G = 2AUC - 1$).

Variations

Logistic regression is the most common and easiest to use. Other models exist for specific uses:

- Probit regression—similar to logistic regression, but with less fat tails.
- Ordered probit—similar to probit regression, but with multiple category ordinal dependent variable.
- Multinomial logistic regression—similar to logistic regression, but for a multiple category nominal dependent variable.
- Poisson / negative binomial—regression models for discrete, positive dependent variables, such as counts.