INSTITUTE OF LAW, POLITICS AND DEVELOPMENT

Sant'Anna School of Advanced Studies – Pisa

Intro to R

Logistic Regression in R

Prof. Kevin Koehler
kevin.koehler@santannapisa.it

- Did a respondent vote?

- Did a respondent vote?
- Did a (civil) war start?

- Did a respondent vote?
- Did a (civil) war start?
- Did a coup happen?

- Did a respondent vote?
- Did a (civil) war start?
- Did a coup happen?

- Did a respondent vote?
- Did a (civil) war start?
- Did a coup happen?

Binary outcomes are ubiquitous in social/political science

- Did a respondent vote?
- Did a (civil) war start?
- Did a coup happen?

Binary outcomes are ubiquitous in social/political science

What happens if we run an OLS model on a binary dependent variable?

The linear probability model

As we discussed, OLS predicts a **continuous** outcome using a linear combination of predictors:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

We estimate coefficients by minimizing the sum of squared residuals:

$$\text{SSR} = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

The linear probability model

As we discussed, OLS predicts a **continuous** outcome using a linear combination of predictors:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

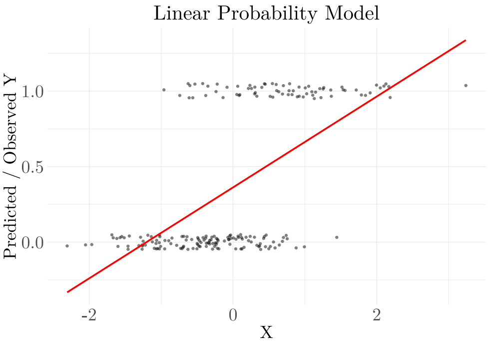
We estimate coefficients by minimizing the sum of squared residuals:

$$\text{SSR} = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

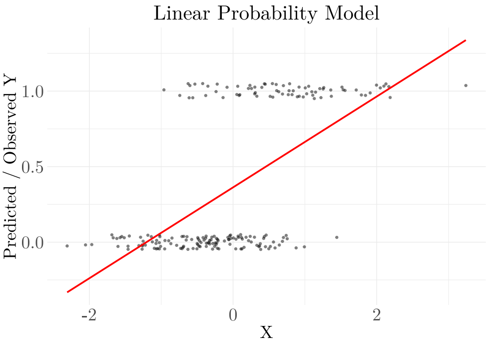
If we estimate an OLS model with a binary dependent variable, we call this a **linear probability model**.

What happens if we use OLS to estimate the probability of an event?

Limitations of linear probability models (1)

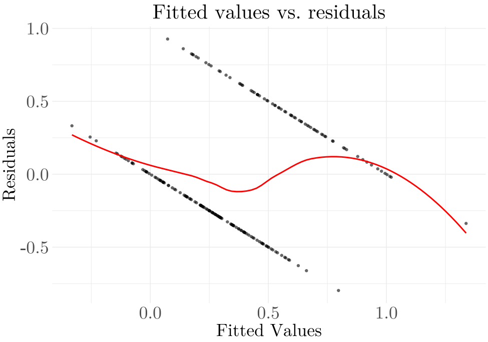


Limitations of linear probability models (1)

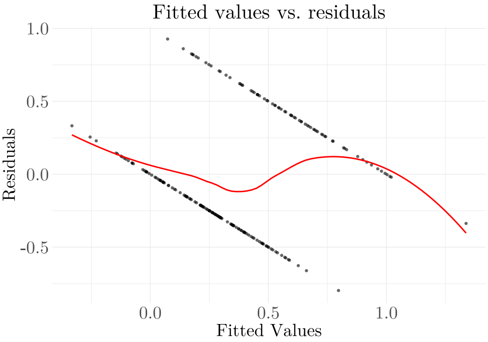


Predictions can fall **outside [0,1]**

Limitations of linear probability models (2)

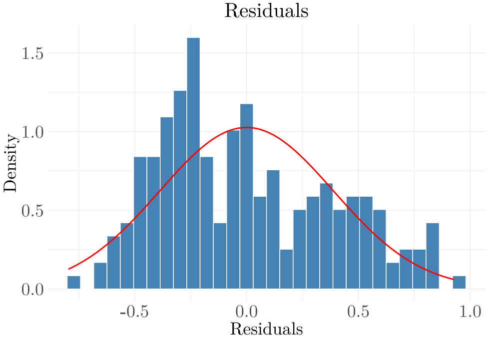


Limitations of linear probability models (2)

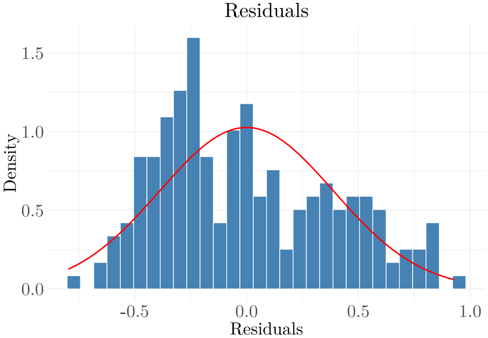


Violates **homoscedasticity** assumption

Limitations of linear probability models (3)

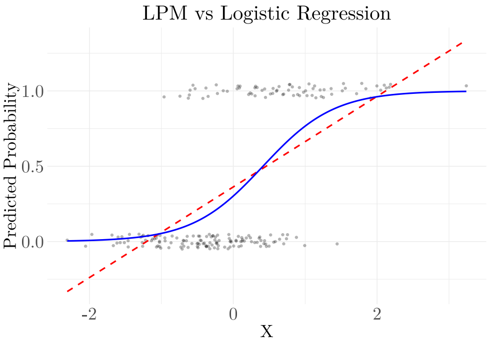


Limitations of linear probability models (3)



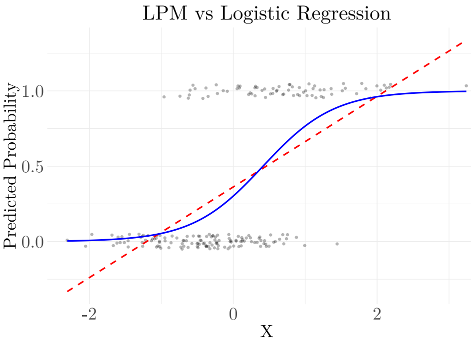
Residuals are **not** normally distributed

Limitations of linear probability models (4)



Constant marginal effect (OLS), Variable effect (Logistic).

Limitations of linear probability models (4)



Constant marginal effect (OLS), Variable effect (Logistic).

Interpreting coefficients as **probabilities** is misleading.

Limitations of linear probability models

In brief, LPMs:

1. ... produce predictions outside the range of the dependent variable (i.e., above 1 and below 0);
2. ... have heteroskedastic errors by definition;
3. ... have residuals which are not normally distributed;
4. ... produce coefficients which cannot properly be interpreted as probabilities.

Limitations of linear probability models

In brief, LPMs:

1. ... produce predictions outside the range of the dependent variable (i.e., above 1 and below 0);
2. ... have heteroskedastic errors by definition;
3. ... have residuals which are not normally distributed;
4. ... produce coefficients which cannot properly be interpreted as probabilities.

! LPMs are not BLUE

As a result, the OLS estimator is no longer BLUE, the Best Linear Unbiased Estimator. In particular, it is not the best (i.e., most efficient) estimator.

We need a better model

We want:

- Predictions **in [0,1]**
- A **nonlinear relationship** between predictors and the probability
- Interpretability in terms of **probabilities**

We need a better model

We want:

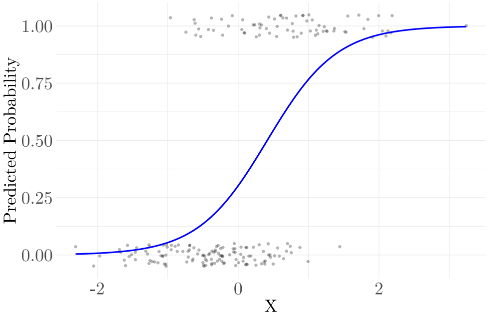
- Predictions **in [0,1]**
- A **nonlinear relationship** between predictors and the probability
- Interpretability in terms of **probabilities**

Logistic regression models the **log-odds** of the outcome:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_i$$

This is called the **logit** function

We need a better model



The logistic function

What are log-odds?

1. Odds are the ratio between the probability that an event occurs and the probability that it does not. Formally, $\text{Odds} = \frac{p}{1-p}$

What are log-odds?

1. Odds are the ratio between the probability that an event occurs and the probability that it does not. Formally, $\text{Odds} = \frac{p}{1-p}$
2. Odds vary in [0,1] (like all ratios). Taking the natural logarithm, we transform [0,1] into $[-\infty, \infty]$, obtaining log-odds. Formally, $\log\left(\frac{p_i}{1-p_i}\right)$

What are log-odds?

1. Odds are the ratio between the probability that an event occurs and the probability that it does not. Formally, $\text{Odds} = \frac{p}{1-p}$
2. Odds vary in [0,1] (like all ratios). Taking the natural logarithm, we transform [0,1] into $[-\infty, \infty]$, obtaining log-odds. Formally, $\log\left(\frac{p_i}{1-p_i}\right)$

What are log-odds?

1. Odds are the ratio between the probability that an event occurs and the probability that it does not. Formally, Odds $= \frac{p}{1-p}$
2. Odds vary in [0,1] (like all ratios). Taking the natural logarithm, we transform [0,1] into $[-\infty, \infty]$, obtaining log-odds. Formally, $\log\left(\frac{p_i}{1-p_i}\right)$

We now have an interval-scaled variable ranging from negative to positive infinity. A linear model fits well on a log-odds scale. Formally:

$$\boxed{\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_i}$$

How can we estimate β_0 and β_1 in logistic regression?

Why is there no version of minimizing SSR with binary outcomes?

1. LPM (OLS) produces estimates above 1 and below 0

Why is there no version of minimizing SSR with binary outcomes?

1. LPM (OLS) produces estimates above 1 and below 0
2. Residuals are non-normal and heteroskedastic by definition

Why is there no version of minimizing SSR with binary outcomes?

1. LPM (OLS) produces estimates above 1 and below 0
2. Residuals are non-normal and heteroskedastic by definition
3. LPM (OLS) coefficients cannot be interpreted as probabilities

Why is there no version of minimizing SSR with binary outcomes?

1. LPM (OLS) produces estimates above 1 and below 0
2. Residuals are non-normal and heteroskedastic by definition
3. LPM (OLS) coefficients cannot be interpreted as probabilities

Why is there no version of minimizing SSR with binary outcomes?

1. LPM (OLS) produces estimates above 1 and below 0
2. Residuals are non-normal and heteroskedastic by definition
3. LPM (OLS) coefficients cannot be interpreted as probabilities

Instead, logistic regression maximizes the likelihood that we would observe the given data under the model parameters.

Maximum likelihood estimation (MLE)

While OLS minimizes the sum of the squared residuals (and thus the **difference between the observed and predicted values**), MLE maximizes the likelihood that we would observe the given data under the model parameters.

Maximum likelihood estimation (MLE)

While OLS minimizes the sum of the squared residuals (and thus the **difference between the observed and predicted values**), MLE maximizes the likelihood that we would observe the given data under the model parameters.

This approach is more general. It applies to any model which can be expressed as a likelihood (e.g., logistic, Poisson, Negative Binomial, etc.)

Maximum likelihood estimation (MLE)

While OLS minimizes the sum of the squared residuals (and thus the **difference between the observed and predicted values**), MLE maximizes the likelihood that we would observe the given data under the model parameters.

This approach is more general. It applies to any model which can be expressed as a likelihood (e.g., logistic, Poisson, Negative Binomial, etc.)

Example: Imagine you flip a coin 1,000 times and obtain 570 heads (and 430 tails). MLE would ask: **What is the probability of heads, p_h, which would make this data most likely?**

(The answer is obvious: $p_h = \frac{570}{1000} = 0.57$).

Some maths

Maximum Likelihood Estimation (MLE)

Step 1: We start with the log-odds (or logit):

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_i$$

where $p_i = P(y_i = 1|X_i)$, the probability of success

Step 2: Exponentiate both sides:

$$\frac{p_i}{1-p_i} = e^{\beta_0+\beta_1 X_i}$$

Step 3: Solve for p_i:

Multiply by $1 - p_i$

$$p_i = (1-p_i)e^{\beta_0+\beta_1 X_i} \tag{a}$$

Multiply out and rearrange:

$$p_i = e^{\beta_0+\beta_1 X_i} - p_i e^{\beta_0+\beta_1 X_i} \tag{b}$$

$$p_i + p_i e^{\beta_0+\beta_1 X_i} = e^{\beta_0+\beta_1 X_i} \tag{c}$$

$$p_i(1 + e^{\beta_0+\beta_1 X_i}) = e^{\beta_0+\beta_1 X_i} \tag{d}$$

Divide both sides by $1 + e^{\beta_0+\beta_1 X_i}$:

$$p_i = \frac{e^{\beta_0+\beta_1 X_i}}{1 + e^{\beta_0+\beta_1 X_i}} \tag{e}$$

Step 4: Final form:

Divide by $e^{\beta_0+\beta_1 X_i}$

$$p_i = \frac{e^{\beta_0+\beta_1 X_i}/e^{\beta_0+\beta_1 X_i}}{\frac{1}{e^{\beta_0+\beta_1 X_i}} + \frac{e^{\beta_0+\beta_1 X_i}}{e^{\beta_0+\beta_1 X_i}}} = \frac{1}{e^{-(\beta_0+\beta_1 X_i)} + 1}$$

Which can be rearranged as:

$$\boxed{p_i = \frac{1}{1 + e^{\beta_0+\beta_1 X_i}}}$$

This is the **standard logistic function**.

Maximum Likelihood Estimation (MLE)

Step 5: The Bernoulli Likelihood:

Each observation $y_i \in \{0, 1\}$ can be modeled as a **Bernoulli trial**:

$$P(y_i|X_i) = p_i^{y_i}(1 - p_i)^{1-y_i}$$

This works because under this formula:

$$P(y_i|X_i) = \begin{cases} p_i & \text{if} \quad y_i = 1 \\ 1 - p_i & \text{if} \quad y_i = 0 \end{cases}$$

Step 6: Full likelihood for all observations:

We can now write the likelihood of the data as:

$$\boxed{\mathcal{L}(\beta_0, \beta_1) = \prod_{i=1}^{n} \left[p_i^{y_i}(1 - p_i)^{1-y_i}\right]}$$

This is the **joint probability** of observing the entire dataset, given the model parameters β_0 and β_1 (assuming independent observations).

Step 7: The Log-Likelihood:

Taking the **log of the full likelihood** makes the math more tractable, so this quantity, the **log-likelihood**, is maximized in MLE:

$$\ell(\beta_0, \beta_1) = \log\left(\prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1-y_i}\right)$$

We use the log of product rule–$\log\left(\prod a_i\right) = \sum \log a_i$:

$$\ell(\beta_0, \beta_1) = \sum_{i=1}^{n} \left[\log(p_i^{y_i}(1 - p_i)^{1-y_i})\right]$$

We use the log-of-power rule–$\log(a^b) = b\log(a)$:

$$\ell(\beta_0, \beta_1) = \sum_{i=1}^{n} [y_i \log(p_i) + (1 - y_i)\log(1 - p_i)]$$

This is the **log-likelihood function** which is maximized in MLE.

The last slide on MLE, I promise!

We can thus state the MLE problem as:

$$\boxed{(\hat{\beta}_0, \hat{\beta}_1) = \arg\max_{\beta_0, \beta_1} \; \ell(\beta_0, \beta_1)}$$

In words: We want to find the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that *maximize* the likelihood function $\ell(\beta_0, \beta_1)$.

The last slide on MLE, I promise!

We can thus state the MLE problem as:

$$\boxed{(\hat{\beta}_0, \hat{\beta}_1) = \arg\max_{\beta_0, \beta_1} \; \ell(\beta_0, \beta_1)}$$

In words: We want to find the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that *maximize* the likelihood function $\ell(\beta_0, \beta_1)$.

No closed-form solution

There is **no closed-form solution**, meaning this maximization problem cannot be solved directly. Instead, MLE uses **numerical optimization**.

Logistic regression in R

Logistic regression in R

```r
data <- readRDS("V-Dem-CY-Full+Others-v14.rds")
data <- data %>%
  filter(year>1969 & year<1980) %>%
  mutate(gdp=log(e_gdppc),
         dem=ifelse(v2x_polyarchy>=0.42,1,0))

logit1 <- glm(dem~gdp,
              data = data,
              family = "binomial")

stargazer(logit1,
          type="latex",
          style = "apsr",
          header = F,
          font.size = "tiny",
          title="Basic logistic regression",
          no.space = T,
          omit.stat = c("adj.rsq","f","ser"),
          dep.var.caption = "",
          dep.var.labels = "Democracy",
          covariate.labels = "GDP/capita (log)")
```

Table 1: Basic logistic regression

	Democracy
GDP/capita (log)	1.229^{***}
	(0.074)
Constant	-3.304^{***}
	(0.164)
N	1,544
Log Likelihood	−682.967
AIC	1,369.934

$^{*}p < .1$; $^{**}p < .05$; $^{***}p < .01$

Sant'Anna
School of Advanced Studies – Pisa

Logistic regression in R

Table 2: Basic logistic regression

	Democracy
GDP/capita (log)	1.229***
	(0.074)
Constant	−3.304***
	(0.164)
N	1,544
Log Likelihood	−682.967
AIC	1,369.934

*p < .1; **p < .05; ***p < .01

The log-odds of being a democracy increase by 1.229 for each one-unit increase in `log(gdp)`.

Logistic regression in R

Table 2: Basic logistic regression

	Democracy
GDP/capita (log)	1.229***
	(0.074)
Constant	−3.304***
	(0.164)
N	1,544
Log Likelihood	−682.967
AIC	1,369.934

*p < .1; **p < .05; ***p < .01

The log-odds of being a democracy increase by 1.229 for each one-unit increase in `log(gdp)`.

This does not tell me anything!

Logistic regression in R

Table 2: Basic logistic regression

	Democracy
GDP/capita (log)	1.229***
	(0.074)
Constant	−3.304***
	(0.164)
N	1,544
Log Likelihood	−682.967
AIC	1,369.934

*p < .1; **p < .05; ***p < .01

The log-odds of being a democracy increase by 1.229 for each one-unit increase in `log(gdp)`.

This does not tell me anything!

We can convert the log-odds to odds ratios: $\text{OR} = e^{1.229} \approx 3.42$. This means that the odds of being democratic increase by 3.42 if `log(GDP)` increases by one unit.

Logistic regression in **R**

Table 2: Basic logistic regression

	Democracy
GDP/capita (log)	1.229***
	(0.074)
Constant	−3.304***
	(0.164)
N	1,544
Log Likelihood	−682.967
AIC	1,369.934

*p < .1; **p < .05; ***p < .01

The log-odds of being a democracy increase by 1.229 for each one-unit increase in `log(gdp)`.

This does not tell me anything!

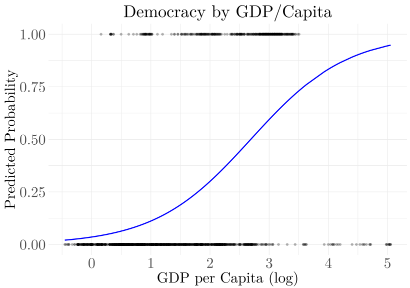
We can convert the log-odds to odds ratios: $\text{OR} = e^{1.229} \approx 3.42$. This means that the odds of being democratic increase by 3.42 if `log(GDP)` increases by one unit.

Doubling GDP leads to an increase in the odds of being democratic of about 2.3
(since $\log(2) \cdot 1.229 \approx 0.85$ and $e^{0.85} \approx 2.34$.

Predicted probabilities

```r
data_model <- model.frame(logit1)
data_model$p_p <- predict(logit1,
                          type = "response")

ggplot(data_model, aes(x = gdp,
                       y = p_p)) +
  geom_line(color = "blue", size = 1) +
  geom_point(aes(y = dem), alpha = 0.3) +
  labs(
    title = "Democracy by GDP/Capita",
    x = "GDP per Capita (log)",
    y = "Predicted Probability"
  ) +
  custom_theme
```



Predicted probabilities