Week 12: Logistic Regression

POP88162 Introduction to Quantitative Research Methods

Tom Paskhalis

Department of Political Science, Trinity College Dublin

Research Paper

- Final research paper should focus on a research question in political science, apply and interpret at least one statistical test to answer it.
- Approximately 10 pages and no more than 5,000 words (references excluded)
- Due **23:59 Tuesday, 22 April**
- Key components:
  - Research question;
  - Justification of its importance and relationship to political science literature;
  - Data;
  - Methods: at least one statistical test and its interpretation.
- More details in Research Paper Guidelines.

Topics for Today

- Linear probability model
- Odds
- Odds ratios
- Log odds
- Logistic regression model

Review: Statistical Tests

		Dependent Variable	
		Nominal/ Ordinal	Interval
Independent Variable	Nominal/ Ordinal	χ^2 (chi-squared) test	Mean comparison test
	Interval	Logistic Regression	Linear Regression

Categorical Dependent Variables

Categorical and Discrete Dependent Variable

- Some dependent variables have a limited number of values they can take:
  - two possible values (binary or dichotomous)
  - three or more possible values:
    - without logical ordering (multinomial)
    - with logical ordering (ordinal)
    - with logical ordering and interval-scale (counts)
- For such dependent variables linear regression model can produce undesirable and nonsensical results.

Binary Dependent Variable

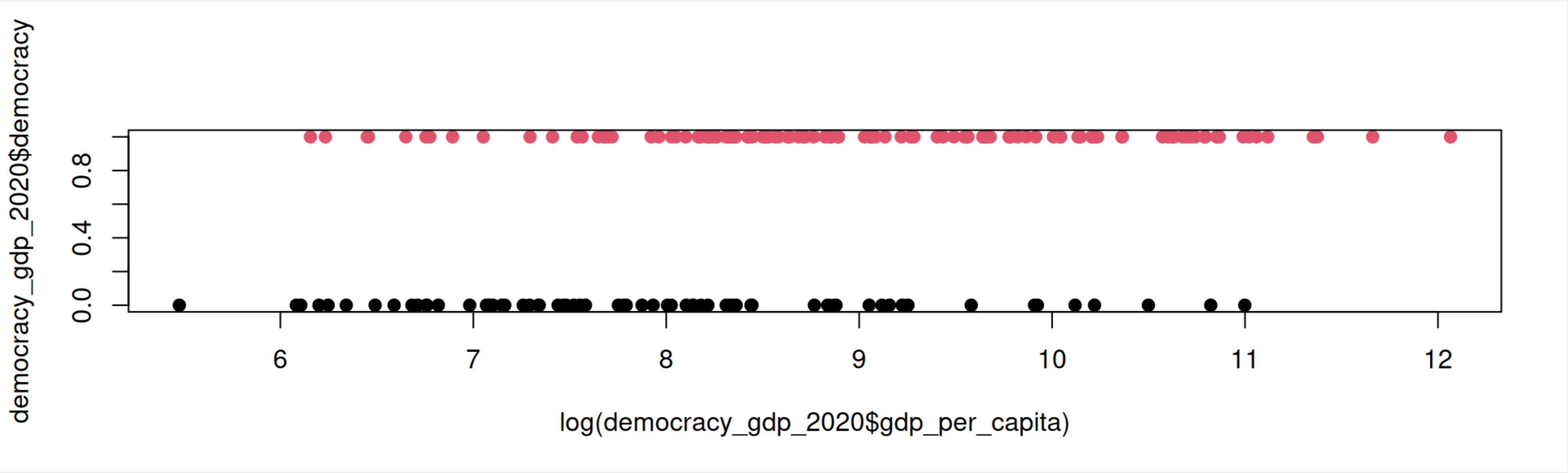
- Binary variables are those with two categories
  - $Y = 1$ if something is "true", or occurred
  - $Y = 0$ if something is "not true", or did not occur

- Examples of binary response variables
  - Survey questions: yes/no; agree/disagree
  - In politics: vote/do not vote
  - In medicine: have/do not have a certain condition
  - In education: correct/incorrect; graduate/do not graduate; pass/fail

Example: GDP and Regime Type

- RQ: Are more economically successful political regimes more likely to be democratic?
- Y: Regime is democratic (1) or authoritarian (0)
- X: Log GDP per capita

Plot | Code



Why Do We Need a New Regression Model?

- Why can't we just run a linear regression?
- Remember that all variables (quantitative and categorical) are represented as numbers.
- Conceivably, we can fit an OLS model with binary outcome.

Linear Probability Model

- The linear regression model that is used for predicting binary outcomes is called **linear probability model**.

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \epsilon_i$$

$$P(Y_i = 1) = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki}$$

- Advantages:
  - Simple and well-known model for a new class of dependent variable.
  - Easy to interpret coefficients.
- Disadvantages:
  - Model produces fitted values that are outside of the $[0, 1]$ range.
  - Relationship certainly non-linear.

Example: Coefficient in LPM

```r
lpm_fit <- lm(democracy ~ log(gdp_per_capita), data = democracy_gdp_2020)
summary(lpm_fit)
```

```
Call:
lm(formula = democracy ~ log(gdp_per_capita), data = democracy_gdp_2020)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.9363 -0.4372  0.1502  0.3861  0.7243 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)         -0.56415    0.21644  -2.606  0.00995 ** 
log(gdp_per_capita)  0.13642    0.02468   5.527 1.18e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4507 on 173 degrees of freedom
  (20 observations deleted due to missingness)
Multiple R-squared:  0.1501,	Adjusted R-squared:  0.1452 
F-statistic: 30.54 on 1 and 173 DF,  p-value: 1.183e-07
```

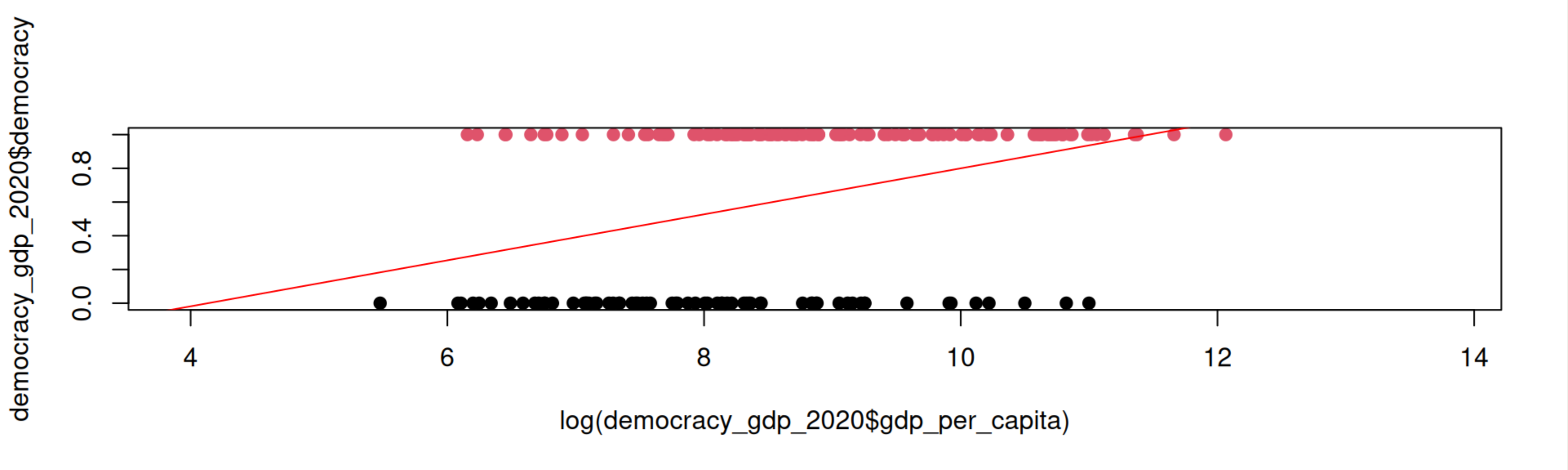
- The estimate $\hat{\beta}$ indicates that an increase of 1 log GDP per capita is associated with a 0.136 increase in the probability of a state being democratic.

Example: Fitted Values in LPM

- Fitted values $\hat{Y}_i$ are interpreted as probability of $Y_i = 1$
- E.g. for a regime with 50K USD per capita GDP:

$$P(Y_i = 1) = -0.564 + 0.136 \times log(50000) = -0.564 + 0.136 \times 10.81 = 0.9$$

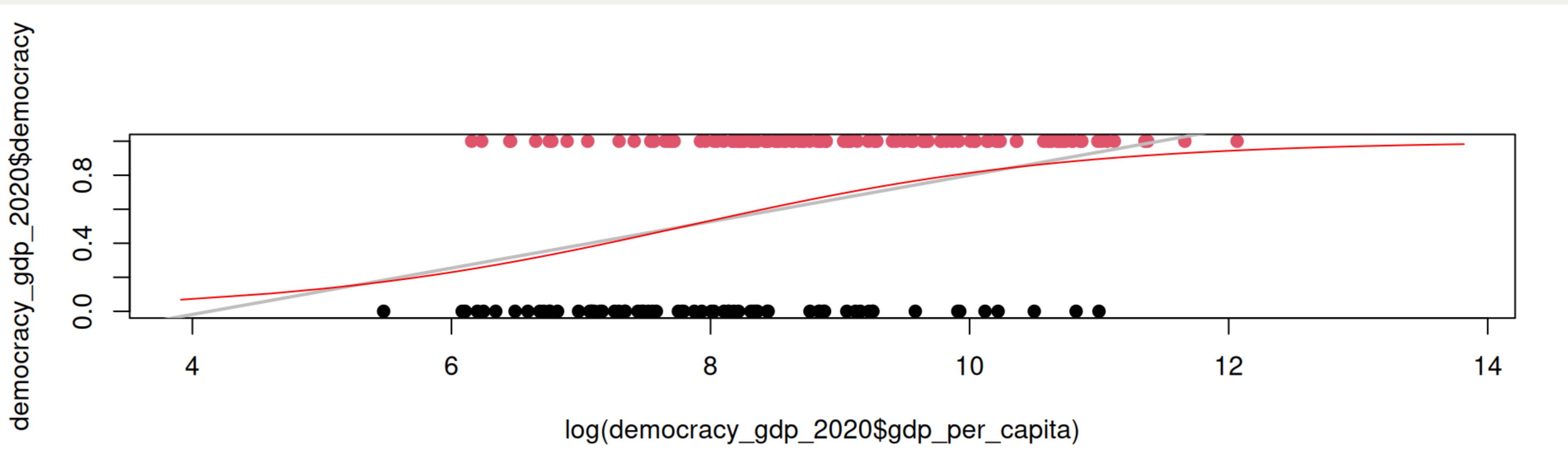
Plot | Code



Example: Alternatives to LPM

- Since probabilities of > 1 and < 0 do not make sense, we might want a different approach for modelling binary dependent variables.
- The s-shaped line below could offer one such alternative.

Plot | Code



Proportions and Probabilities

- For binary dependent variables, we are interested in the **proportion** of the subjects in the population for whom $Y = 1$.
- We can also think of this as the probability π that a randomly selected member of the population will have the value $Y = 1$ rather than $Y = 0$.

$$\pi = P(Y = 1)$$

$$1 - \pi = P(Y = 0)$$

  - If $\pi = 0$, no unit in the population has $Y = 1$;
  - If $\pi = 1$ every unit in the population has $Y = 1$.
- We want to model π, given one or more independent (explanatory) variables X.

Binary Predictor of Regime Type

- Does political regime depend on former colonial status?

```r
democracy_gdp_2020$democracy <- factor(democracy_gdp_2020$democracy, labels = c("autocracy", "democracy"))
democracy_gdp_2020$noncol <- factor(democracy_gdp_2020$noncol, labels = c("colony", "non-colony"))
```

```r
table(democracy_gdp_2020$noncol, democracy_gdp_2020$democracy)
```

```
           autocracy democracy
colony            66        90
non-colony         6        19
```

```r
prop.table(table(democracy_gdp_2020$noncol, democracy_gdp_2020$democracy), margin = 1)
```

```
           autocracy democracy
colony     0.4230769 0.5769231
non-colony 0.2400000 0.7600000
```

Odds

Conditional Probabilities

- Consider the dummy variable $X = 0$ if a state had been a colony at some point, and $X = 1$ if not.
- We can then estimate conditional probabilities of having democratic regime separately for these two groups:

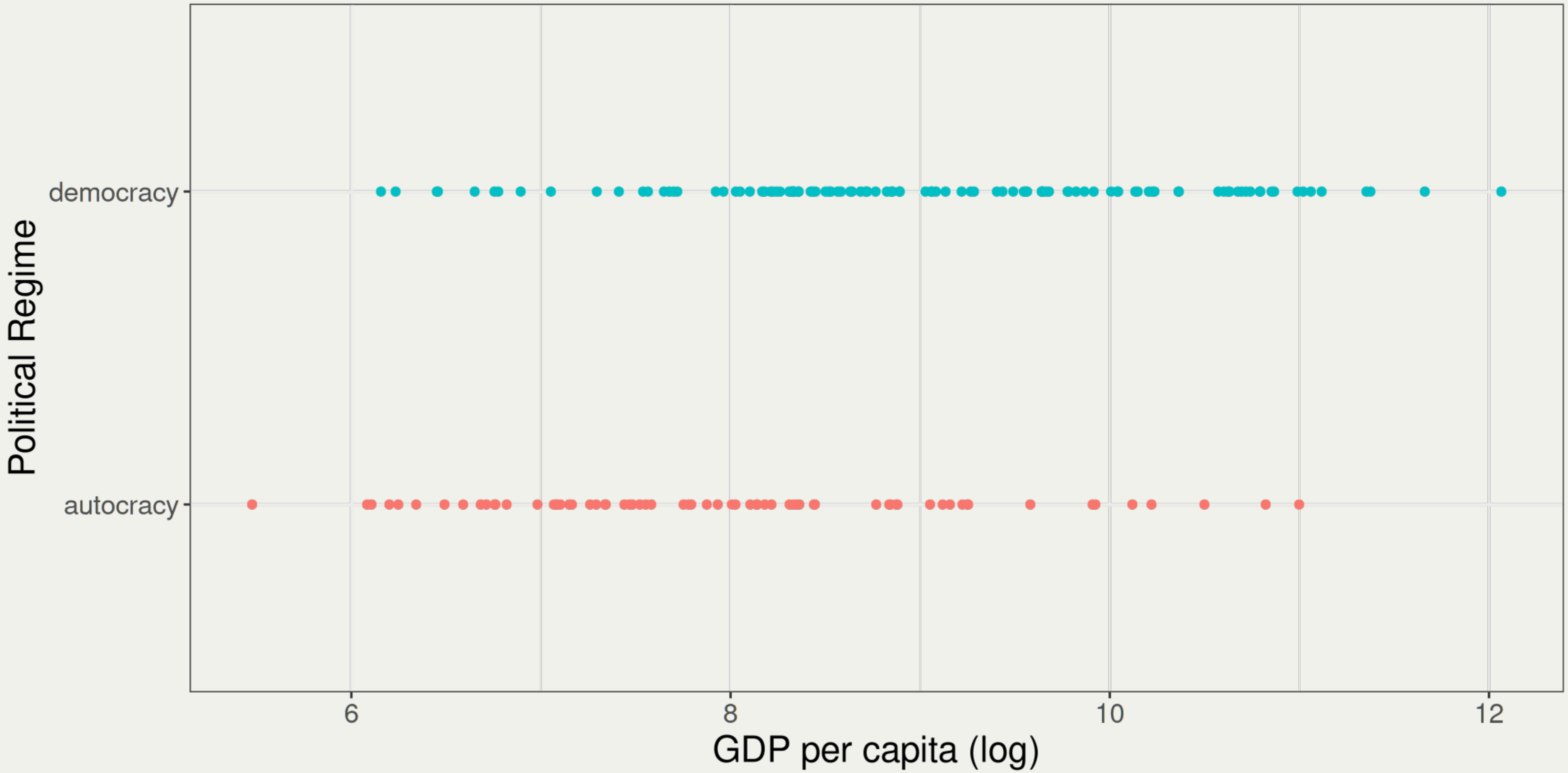
$$\hat{P}(Y = 1 | X = 0) = 0.58$$

$$\hat{P}(Y = 1 | X = 1) = 0.76$$

- The estimated probability of having democratic regime is higher for non-colonies than for former colonies.
- More generally, we would like to model how the probability $\pi = P(Y = 1)$ depends on one or more explanatory variables, which might be continuous.

Continuous Predictor

- The linear regression model implicitly assumed a normal distribution for the response variable.
- But binary outcomes cannot have a normal distribution!



Political Regime
democracy
autocracy
6
8
10
12
GDP per capita (log)

How to Model π?

- Linear regression model: *conditional mean* is equal to a linear combination of explanatory variables:

$$E(Y_i | X_{1i}, \ldots) = \mu_i = \alpha + \beta_1 X_{1i} + \ldots$$

- Linear probability model: *conditional probability* is equal to a linear combination of explanatory variables:

$$E(Y_i | X_{1i}, \ldots) = P(Y_1 = 1 | X_{1i}, \ldots) = \pi_i = \alpha + \beta_1 X_{1i} + \ldots$$

- But we need some way to make sure $0 \leq \pi_i \leq 1$
  - We cannot model a linear model for π directly.
  - Instead, we build a linear model for a transformation of π!

From Probabilities to Odds

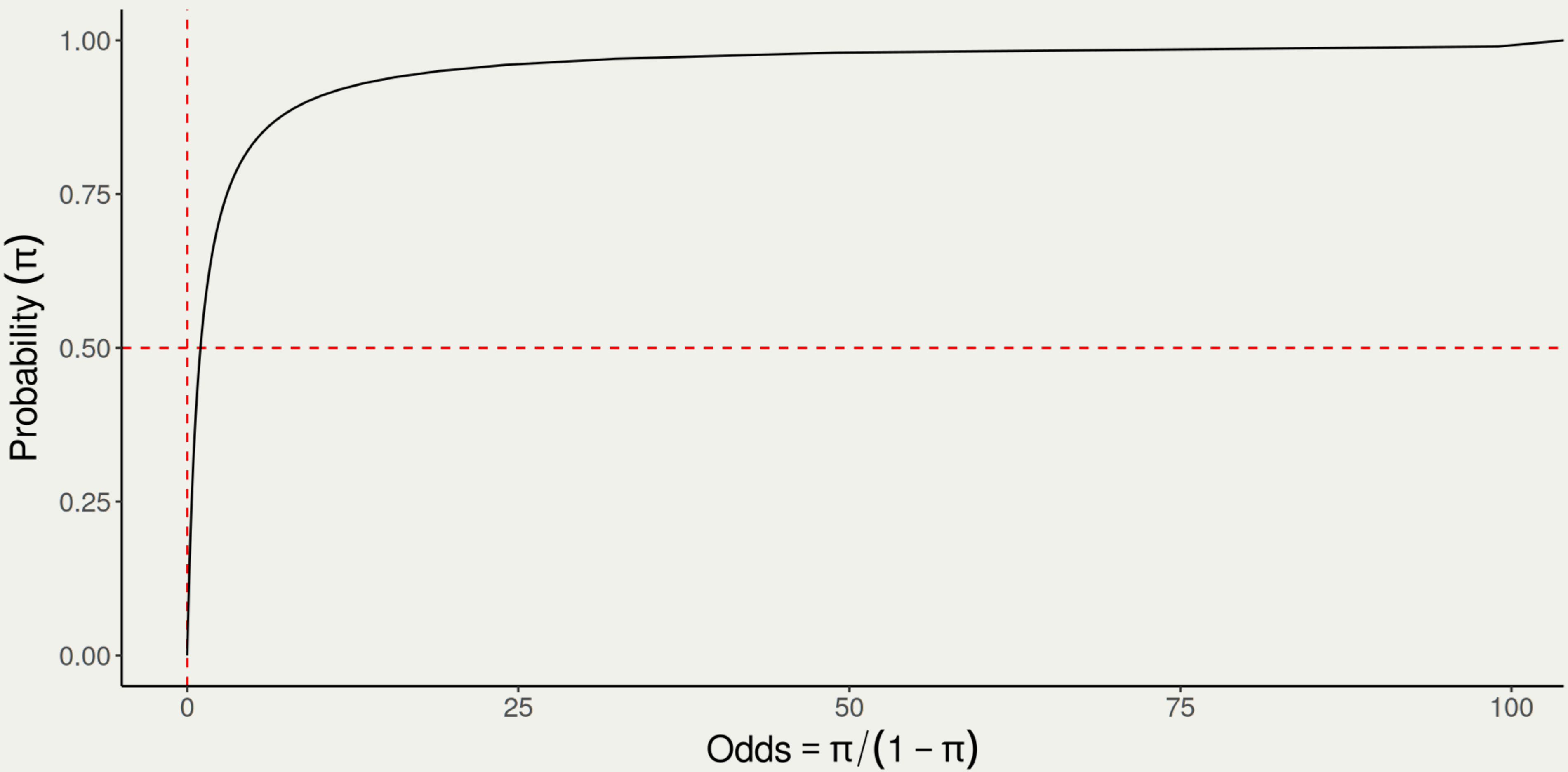
- The **odds** are the ratio of the probabilities of the event and the non-event:

$$Odds = \frac{P(Y=1)}{1-P(Y=1)} = \frac{\pi}{1-\pi}$$

- If the probability of having a democratic regime is $\pi = 0.9$
  - the odds of having a democratic regime are $= 0.9/0.1 = 9$
  - the odds of having an autocratic regime are $= 0.1/0.9 = 0.11$
- Odds vs. probabilities π:
  - If $odds = 1$, $P(Y=1) = P(Y=0)$, i.e., $\pi = 0.5$
  - If $odds > 1$, $P(Y=1) > P(Y=0)$, i.e., $\pi > 0.5$
  - If $odds < 1$, $P(Y=1) < P(Y=0)$, i.e., $\pi < 0.5$

From Probabilities to Odds

- Range of π is $(0, 1)$
- Range of odds is $(0, +\infty)$



Conditional Odds

```r
prop.table(table(democracy_gdp_2020$noncol, democracy_gdp_2020$democracy), margin = 1)
```

```
           autocracy democracy
colony     0.4230769 0.5769231
non-colony 0.2400000 0.7600000
```

- Odds of having democratic regime in former colonies:

$$\widehat{Odds}_C = \frac{\hat{\pi}}{1-\hat{\pi}} = \frac{0.58}{1-0.58} = 1.38$$

- Odds of having democratic regime in non-colonies:

$$\widehat{Odds}_{NC} = \frac{\hat{\pi}}{1-\hat{\pi}} = \frac{0.76}{1-0.76} = 3.17$$

From Odds to Odds Ratios

- An **odds ratio** is the ratio of two conditional odds that describes the association between two variables.

$$\widehat{OR}_{NC/C} = \frac{\widehat{Odds}_{NC}}{\widehat{Odds}_{C}} = \frac{3.17}{1.38} = 2.3$$

- The odds of having a democratic regime for non-colonies are 2.3 higher (130% higher) than for former colonies.
- The probability of having a democratic regime is higher for non-colonies than for former colonies.
- Having past colonial history is associated with lower odds of having democratic regime.

Odds Ratios

- In our example,
  - Y = political regime (1 = democracy, 0 = autocracy)
  - X = colonial past (1 = non-colony, 0 = colony)
- The association is described by comparing odds of $Y = 1$ for levels of variable X
  - If odds ratio $= 1$, odds are equal for groups 0 and 1 (no association between X and Y)
  - If odds ratio > 1, odds for group $1 >$ odds for group 0 (positive association between X and Y)
  - If odds ratio < 1, odds for group $1 <$ odds for group 0 (negative association between X and Y)

From Odds to Log Odds

- Recall that we need to solve the problem that:
  - The linear predictor $\alpha + \beta_1 X_{1i} + \ldots$ can take values from $-\infty$ to $+\infty$.
  - The probability π_i must be between 0 and 1.
- We now have the necessary pieces to solve the problem.
  - Turning π_i into the odds expanded the range to:

$$0 < \frac{\pi}{1-\pi} < +\infty$$

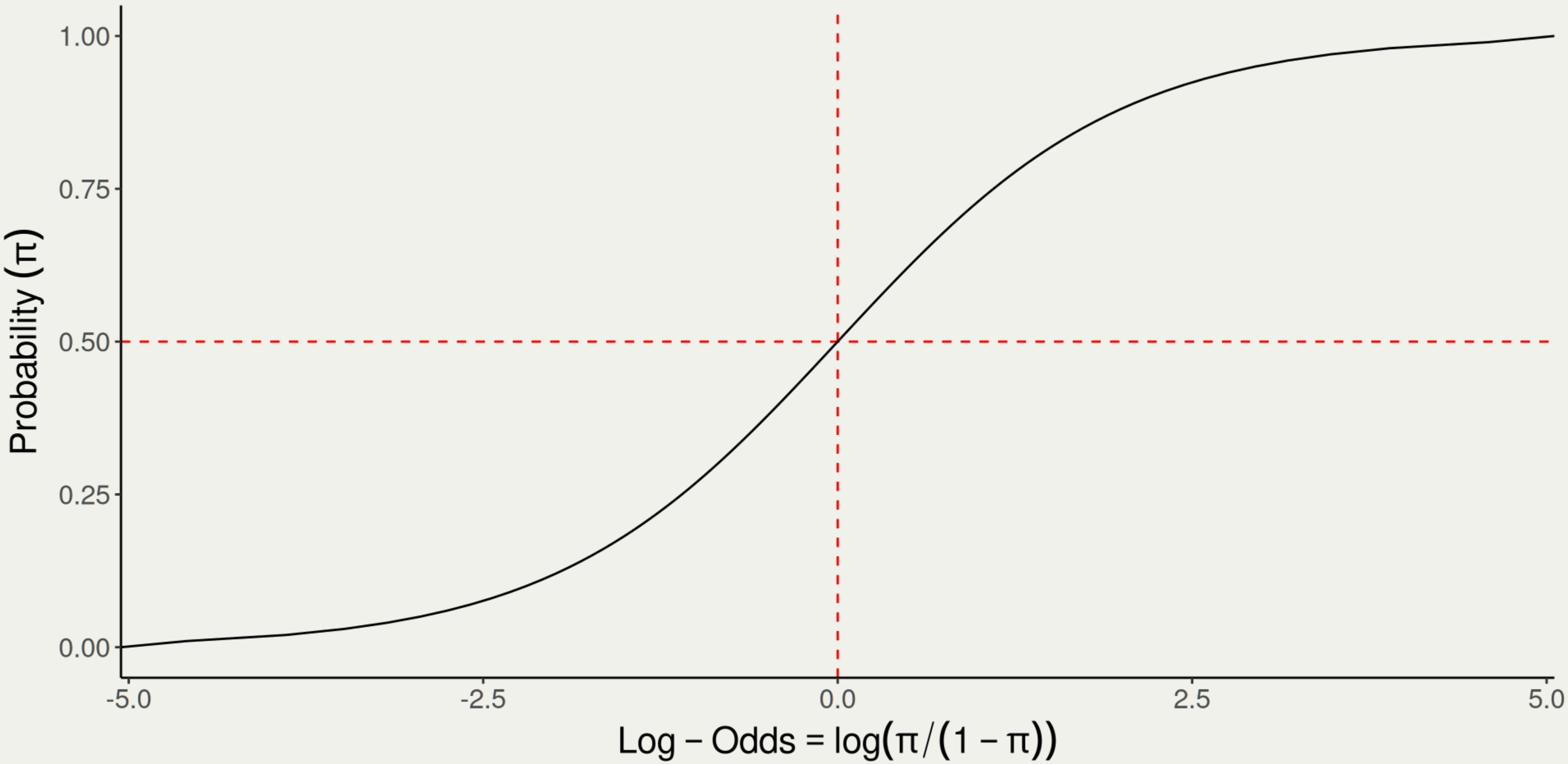
  - By taking the logarithm of the odds:

$$-\infty < log\left(\frac{\pi}{1-\pi}\right) < +\infty$$

- This transformation is known as the **logit**.

From Probabilities to Log Odds

- Range of π is $(0, 1)$
- Range of log-odds is $(-\infty, +\infty)$



Logistic Regression Model

Logistic Regression Model

- We can now express a logit transformed probability of $Y_i = 1$ as *binary logistic regression model*:

$$log(Odds_i) = log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki}$$

where:

- Observations $i = 1, \ldots, n$
- π_i is the probability of dependent variable $Y_i = 1$
- $X_{1i}, \ldots, X_{ki}$ are k independent variables
- α is the **intercept** or **constant**
- $\beta_1, \ldots, \beta_k$ are **coefficients**

Model for the Probabilities

- Although the model is written first for the log-odds, it also implies a model for the probabilities, π_i:

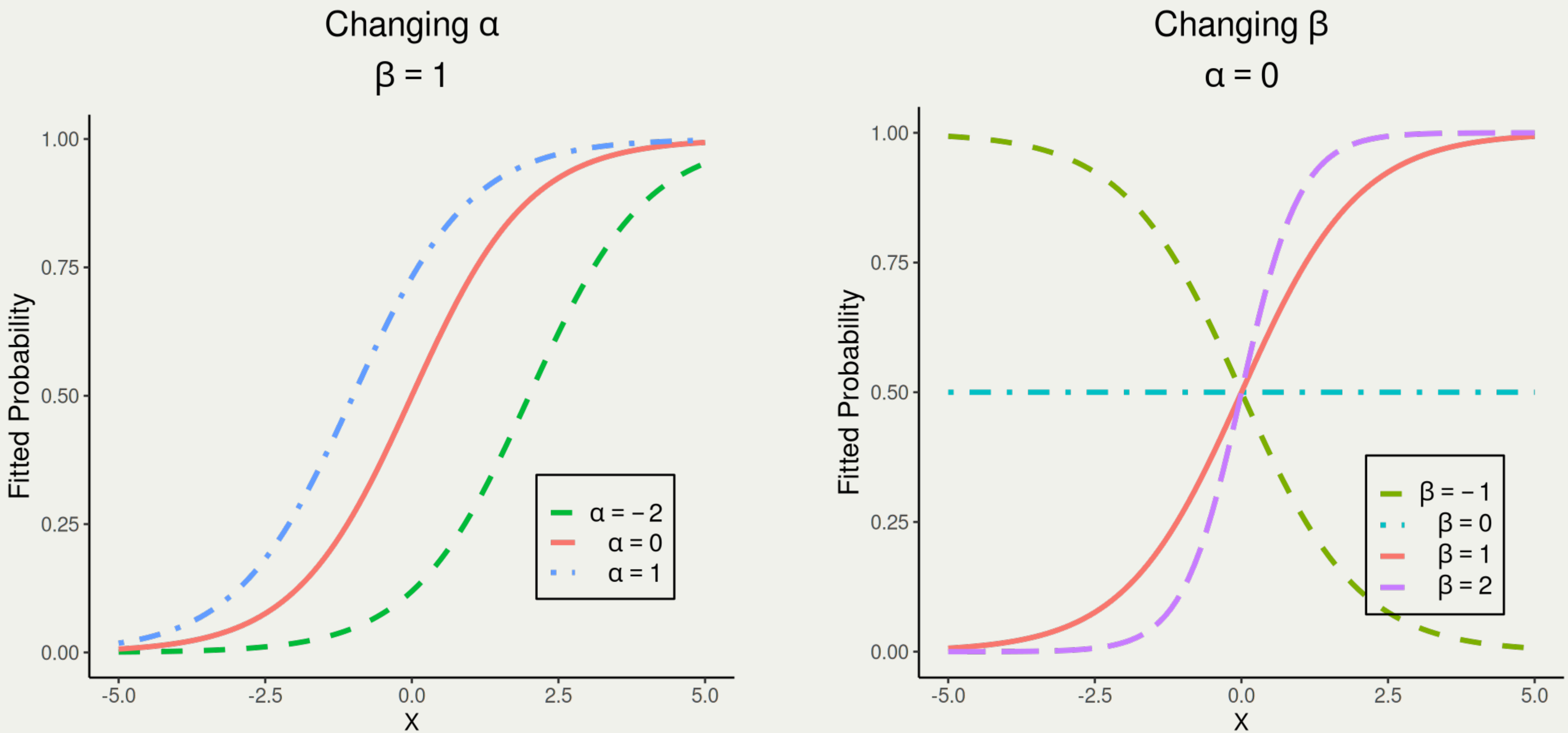
$$\pi_i = \frac{exp(\alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki})}{1 + exp(\alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki})}$$

- π_i is always between 0 and 1.
- The plots on the next slide give examples of

$$\pi_i = \frac{exp(\alpha + \beta X_i)}{1 + exp(\alpha + \beta X_i)}$$

for a simple logistic model with one continuous X

Probabilities from a Logistic Model



Example: GDP and Regime Type

- Let's return to our example:
  - RQ: Are more economically successful political regimes more likely to be democratic?
  - Y: Regime is democratic (1) or authoritarian (0)
  - X_1: Log GDP per capita
  - X_2: Colonial past (1 non-colony, 0 colony)

```r
glm_fit <- glm( # Note that we use glm() function rather than lm()
  democracy ~ log(gdp_per_capita) + noncol,
  family = binomial(link = "logit"), # tells R to use logit
  data = democracy_gdp_2020
)
```

Summarising Logistic Regression Model

```
summary(glm_fit)
```

```
Call:
glm(formula = democracy ~ log(gdp_per_capita) + noncol, family = binomial(link = "logit"),
    data = democracy_gdp_2020)

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         -5.08270    1.21876  -4.170 3.04e-05 ***
log(gdp_per_capita)  0.65378    0.14540   4.497 6.91e-06 ***
noncolnon-colony    -0.08905    0.56937  -0.156    0.876
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 226.17  on 169  degrees of freedom
Residual deviance: 200.15  on 167  degrees of freedom
  (25 observations deleted due to missingness)
AIC: 206.15
```

Interpretation of the Coefficient Estimates

$$log\widehat{\left(\frac{\pi_i}{1-\pi_i}\right)} = -5.08 + 0.65 \times \text{log(GDP)}_i - 0.08 \times \text{Non-colony}_i$$

- The signs of the coefficient estimates show the directions of associations:
  - $\hat{\beta}_{log(GDP)} > 0 \rightarrow$ higher log GDP per capita is associated with higher probability of regime being democratic, controlling for former colony status.
  - $\hat{\beta}_{non-colony} < 0 \rightarrow$ non-colonial status is associated with lower probability of regime being democratic, holding log GDP constant.

Interpretation of the Coefficient Estimates

- Log-odds is not a very intuitive concept!
- Exponentiating converts them into more intuitive odds ratios

```
round(exp(coef(glm_fit)), 2)
```

```
      (Intercept)  log(gdp_per_capita)     noncolnon-colony
             0.01                 1.92                 0.91
```

- $exp(\hat{\beta}_{log(GDP)}) = 1.92 \rightarrow$ an increase of 1 log GDP per capita multiplies the odds of regime being democratic by 1.92, controlling for former colony status.
  - i.e. it increases the odds by 92%
- $exp(\hat{\beta}_{non-colony}) = 0.91 \rightarrow$ holding log GDP per capita constant, having no colonial past multiplies the odds of regime being democratic by 0.91
  - i.e. it decreases the odds by 9%

Next

- Workshop:
  - RQ Presentations V
- Research paper due:
  - **23:59 Tuesday, 22 April**