

Binary Outcomes

Francisco Villamil

Applied Quantitative Methods II
MA in Social Sciences, Spring 2026

1/1

Today's goals

- Understand why OLS is problematic for binary outcomes
- Learn the linear probability model and its trade-offs
- Understand logistic regression and how to estimate it in R
- Interpret logit results using marginal effects and predicted probabilities
- Compare LPM and logit in practice

2/1

This session introduces the first major extension beyond standard OLS: what to do when the outcome is binary (0/1). We'll work through two approaches — the linear probability model (which is just OLS applied to a binary outcome) and logistic regression (which is purpose-built for binary outcomes). The key practical skill is interpretation: logit coefficients are not directly meaningful, so we'll spend significant time on marginal effects and predicted probabilities.

Roadmap

3/1

Binary outcomes are everywhere

- Many outcomes in social science are binary (yes/no):
 - Did someone vote?
 - Did a war break out?
 - Did a bill pass?
 - Did a country democratize?
- Our outcome $Y \in \{0,1\}$
- We want to model: $\Pr(Y = 1 | X)$

4/1

Start by asking students for examples of binary outcomes in their own research or fields. The key insight is that when Y is binary, $E[Y | X] = \Pr(Y = 1 | X)$. So modeling the conditional expectation is the same as modeling the probability. This is what makes binary outcomes special: we're modeling a probability, which must be between 0 and 1.

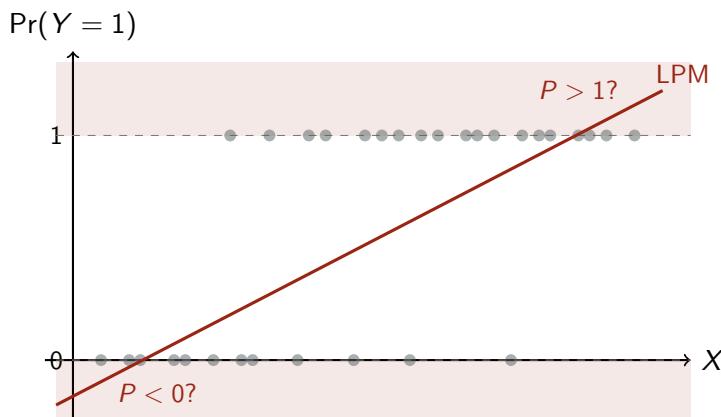
What happens if we just use OLS?

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- OLS gives us: $E[Y | X] = \beta_0 + \beta_1 X$
- Since $Y \in \{0, 1\}$: $E[Y | X] = \Pr(Y = 1 | X)$
- So OLS is modeling a probability as a **linear function** of X
- This is the **linear probability model** (LPM)

5/1

The LPM in pictures



6/1

Walk through this step by step. The LPM is not a new model — it's literally just OLS applied to a binary outcome. The math works because for a 0/1 variable, the mean equals the proportion of 1s. So the regression line estimates a probability. The name “linear probability model” comes from the fact that it models probability as a linear function of the predictors.

This is the fundamental problem with the LPM visualized. The data points can only be at 0 or 1, but the linear fit extends beyond both bounds. For low values of X , the model predicts negative probabilities; for high values, it predicts probabilities greater than 1. Both are nonsensical. Draw attention to the shaded regions. This motivates the need for a model that naturally constrains predictions to $[0, 1]$.

The LPM: Simple and intuitive

- β_1 has a direct interpretation:
 - A one-unit increase in X changes the probability of $Y = 1$ by β_1
- Easy to estimate: just `lm(y ~ x, data = df)`
- Easy to interpret: same as OLS
- Many applied researchers use the LPM in practice
- Especially common in economics and political science

7/1

LPM limitations

- **Problem 1:** Predictions outside $[0, 1]$
 - A linear function can produce $\hat{P} < 0$ or $\hat{P} > 1$
 - Probabilities must be between 0 and 1
- **Problem 2:** Heteroskedasticity by construction
 - $\text{Var}(\varepsilon | X) = P(1 - P)$, which varies with X
 - Always use robust SEs with the LPM
- **Problem 3:** Non-linearity at the extremes
 - True relationship between X and $\Pr(Y = 1)$ is S-shaped
 - LPM forces it to be linear

8/1

Despite the problems, the LPM is widely used because of its simplicity. The coefficient is directly interpretable as a change in probability, which is exactly what we want. In economics, the LPM is often the default for binary outcomes, with logit used as a robustness check. The key selling point: you don't need to learn any new interpretation tricks. It's just OLS.

Three problems, in order of how much they matter in practice. Problem 1 (out-of-bounds predictions) is the most commonly cited but often the least practically important — if most predicted probabilities are in a reasonable range, a few going below 0 or above 1 may not change your conclusions. Problem 2 (heteroskedasticity) has a simple fix: always use robust SEs with the LPM, same as we discussed last week. Problem 3 (non-linearity) is arguably the most important: if the true relationship is S-shaped, a linear approximation will be poor at the extremes, biasing your estimates of marginal effects for observations far from the mean.

When is the LPM “good enough”?

- When probabilities are in the middle range (0.2–0.8)
 - The linear approximation is reasonable here
- When you care about **average marginal effects**
 - LPM and logit often give similar AMEs
- When simplicity of interpretation matters
- When is it **not** good enough?
 - Rare events (many observations near 0)
 - You need predicted probabilities to be bounded
 - The relationship is clearly non-linear

9/1

This is a practical question that students will face in their own research. The honest answer is: in many applied settings, the LPM and logit give very similar answers, especially for average marginal effects. The difference matters most when (a) the outcome is rare (e.g., civil war onset, which happens in less than 2% of country-years), (b) you need to make predictions that are valid probabilities, or (c) you care about how effects vary across the probability scale. Angrist and Pischke (2009) make a strong case for the LPM; other methodologists disagree. Students should know both approaches.

You estimate an LPM predicting civil war onset. You find that 8% of predicted probabilities are negative.

Is this a problem? What would you do?

10/1

Discussion prompt. Let students debate. Key points to draw out:
(1) It depends on what you’re using the model for. If you just want average effects, the out-of-bounds predictions may not matter much.
(2) If you need actual predicted probabilities (e.g., for a risk score), then yes, it’s a serious problem. (3) The percentage of out-of-bounds predictions gives you a sense of how bad the linear approximation is. 8% is moderate. (4) The practical solution: estimate both LPM and logit and see if the key conclusions change. If they agree, use whichever is easier to communicate.

Roadmap

11/1

The logistic function

$$\Pr(Y = 1 | X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

- This is an S-shaped (sigmoid) curve
- Output is always between 0 and 1
- Steep in the middle, flat at the extremes
- A natural model for probabilities

12/1

This is the key formula of the lecture. Emphasize that the logistic function takes any real number and maps it to the interval (0, 1). When $\beta_0 + \beta_1 X$ is very negative, the probability approaches 0; when it's very positive, it approaches 1. The S-shape captures the intuition that going from 0.01 to 0.02 is a bigger deal than going from 0.50 to 0.51, and that probabilities can't keep increasing linearly forever.

The sigmoid curve



13/1

Walk through the shape of the curve. At the center ($X\beta = 0$), the probability is exactly 0.5 and the curve is steepest. As $X\beta$ becomes very positive, the probability approaches 1 but never reaches it. As $X\beta$ becomes very negative, the probability approaches 0 but never reaches it. This S-shape solves the LPM's main problem: predictions are always valid probabilities. Contrast with the straight line from the LPM diagram.

The logit transformation

We can rearrange the logistic function:

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

- The left side is the **log-odds** (or “logit”)
- $\frac{P}{1 - P}$ is the **odds** of the event
- The model is linear in the log-odds, not in the probability
- This is why we call it “logistic regression” or “logit”

14/1

The logit transformation is the inverse of the logistic function. It maps probabilities in $(0, 1)$ to the entire real line $(-\infty, +\infty)$. This is the key mathematical trick: we take a bounded quantity (probability) and transform it to an unbounded quantity (log-odds), then model the unbounded quantity as a linear function of X . The word “logit” comes from “log” + “unit” (as in probability unit). The odds $P/(1 - P)$ are familiar from betting: if $P = 0.75$, the odds are 3 to 1.

Maximum likelihood estimation

- We can't use OLS for logistic regression
- Instead, we use **maximum likelihood estimation** (MLE)
- The intuition:
 - For each observation, the model predicts $\Pr(Y_i = 1)$
 - MLE finds the coefficients that make the observed data most likely
- No need to derive this — R does it for us
- The math is different, but the workflow is the same

15/1

Keep this brief — students don't need the derivation. The key idea is that MLE is a different estimation method from OLS, but conceptually similar: OLS minimizes the sum of squared residuals, MLE maximizes the probability of observing the data we actually observed. The likelihood function for a binary model is the product of predicted probabilities for the 1s and predicted (1-probability) for the 0s. R's `glm()` does all the computation. What matters for students is understanding the output, not the optimization algorithm.

Estimating logit in R

- The function: `glm(y ~ x, family = binomial, data = df)`
- `glm`: generalized linear model
- `family = binomial`: tells R to use logistic regression
- The syntax is identical to `lm()`, just swap to `glm()`
- Works with `broom::tidy()`, `modelsummary()`, etc.

16/1

Emphasize the simplicity. The only change from OLS is replacing `lm()` with `glm(..., family = binomial)`. Everything else — formula syntax, `tidy()`, `modelsummary()` — works the same. The `family = binomial` argument tells R that the outcome is binary and it should use the logistic link function. Technically, `family = binomial(link = "logit")` is the full specification, but `logit` is the default for `binomial` so we can omit it. Mention that `family = binomial(link = "probit")` gives probit regression, which we'll briefly mention later.

Interpreting logit output: Log-odds

- The direct output gives coefficients in **log-odds**
- $\beta_1 = 0.5$ means:
 - A one-unit increase in X increases the log-odds by 0.5
- This is hard to interpret!
- Nobody thinks in log-odds

17/1

Be honest with students: log-odds coefficients are the “native” output of logistic regression, but they are not intuitive. A log-odds of 0.5 doesn’t map onto any natural quantity that people understand. You can’t say “the probability increases by 0.5” because that’s not what it means. This is the central interpretation challenge of logistic regression and why we need additional tools (odds ratios, marginal effects, predicted probabilities).

Interpreting logit output: Odds ratios

- Exponentiate the coefficient: $e^{\beta_1} = \text{odds ratio}$
- In R: `exp(coef(model))`
- $e^{0.5} \approx 1.65$ means:
 - A one-unit increase in X **multiplies** the odds by 1.65
 - Or: the odds increase by 65%
- Slightly more intuitive, but still not probabilities
- The change in probability depends on where you start

18/1

Odds ratios are commonly reported in epidemiology and public health, but they can be misleading. The key subtlety: “the odds increase by 65%” is NOT the same as “the probability increases by 65 percentage points.” For rare events, odds ratios approximate relative risk ratios, so they’re somewhat interpretable. For common events, the connection between odds and probability is non-linear and confusing. Many methodologists recommend against reporting odds ratios as the primary result. We’ll learn better tools in the next section.

Worked example: From log-odds to probability

$$\text{Logit model: } \log\left(\frac{P}{1-P}\right) = -2 + 0.5 \cdot \text{Education}$$

- For someone with Education = 4:

- For someone with Education = 4:
 - Log-odds = $-2 + 0.5 \times 4 = 0$
 - Odds = $e^0 = 1$ (50-50 chance)
 - Probability = $\frac{1}{1+e^0} = 0.50$

- For someone with Education = 8:

- For someone with Education = 8:
 - Log-odds = $-2 + 0.5 \times 8 = 2$
 - Odds = $e^2 \approx 7.4$
 - Probability = $\frac{1}{1+e^{-2}} \approx 0.88$

- Going from 4 to 8 years: P goes from 0.50 to 0.88
- The same 4-unit change gives $\Delta P = 0.38$

19/1

Walk through each step carefully. This is the slide where the abstraction becomes concrete. The key lesson: even though $\beta_1 = 0.5$ is constant, the change in probability depends on where you start. Going from Education = 0 to 4 gives a different ΔP than going from 4 to 8. You can compute P at Education = 0: log-odds = -2 , $P = 1/(1+e^2) \approx 0.12$. So going from 0 to 4, $\Delta P = 0.50 - 0.12 = 0.38$. Going from 4 to 8, $\Delta P = 0.88 - 0.50 = 0.38$. Actually similar here, but try Education = 12: log-odds = 4 , $P \approx 0.98$, so from 8 to 12: $\Delta P = 0.10$. The effect diminishes at the extremes.

Why coefficients alone are not enough

- In OLS: β_1 = change in Y for one-unit change in X (always)
- In logit: the change in **probability** depends on:
 - The current value of X
 - The values of all other variables
- A coefficient of $\beta_1 = 0.5$ could mean:
 - Going from $P = 0.01$ to $P = 0.016$ (tiny change)
 - Going from $P = 0.50$ to $P = 0.62$ (large change)
- We need better tools to interpret logit models

20/1

This slide hammers home the core message. In OLS, the coefficient IS the marginal effect, everywhere, always. In logit, the coefficient is the marginal effect on the log-odds scale, but on the probability scale — which is what we actually care about — the effect varies. This is not just a mathematical curiosity; it has real implications. A researcher who reports “education has a log-odds coefficient of 0.5” has told you almost nothing about the substantive size of the effect. You need to know where on the curve you are. This motivates the next section on marginal effects and predicted probabilities.

Roadmap

21/1

A logit model estimates $\hat{\beta}_1 = 0.8$ for education.

Your colleague says: "Education increases the probability of voting by 0.8."

What's wrong with this statement?

22/1

Discussion prompt. Key errors to identify: (1) The coefficient 0.8 is in log-odds, not probability. A log-odds change of 0.8 is NOT a probability change of 0.8. (2) Even if they meant "0.8 percentage points" or "80 percentage points," it would still be wrong because the effect on probability varies depending on where you are on the curve. (3) The correct statement requires either specifying the baseline ("for someone at the mean of all variables, education increases the probability by...") or reporting the average marginal effect. This motivates the tools we're about to learn.

Predicted probabilities

- The most intuitive way to interpret logit models
- “What is the predicted probability of $Y = 1$ for a person with these characteristics?”
- In R:
 - `marginaleffects::predictions(model)`
 - Returns predicted probabilities for each observation
 - Or at specific values: `predictions(model, newdata = datagrid(...))`

23/1

Predicted probabilities are the most natural and intuitive quantity to report from a logit model. Instead of saying “the coefficient is 0.8,” you can say “a college-educated 40-year-old woman has a 78% predicted probability of voting, compared to 52% for a high-school-educated 40-year-old woman.” These are quantities that any audience can understand. The `predictions()` function from the `marginaleffects` package makes this easy. You can compute predictions for every observation in your data, or for specific hypothetical profiles using `datagrid()`.

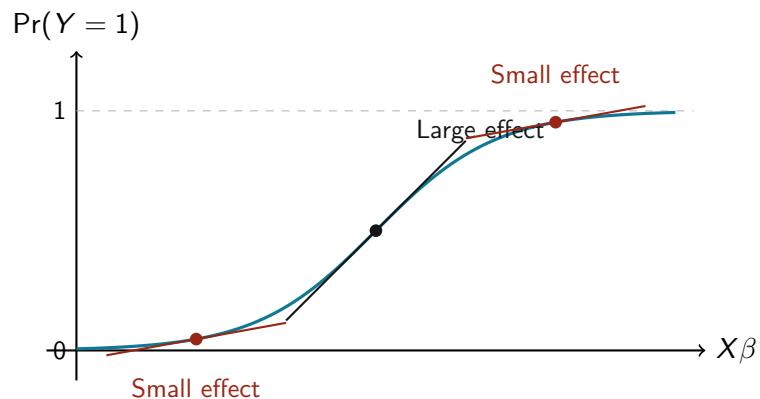
Average marginal effects (AME)

- The marginal effect of X on $\Pr(Y = 1)$ varies across observations
- The **AME** averages across all observations in the sample
- In R: `marginaleffects::avg_slopes(model)`
- Interpretation (just like OLS):
 - “On average, a one-unit increase in X changes $\Pr(Y = 1)$ by ΔP ”
- AMEs from logit are often similar to LPM coefficients

24/1

The AME is the most direct analog to the OLS coefficient: it gives you a single number summarizing the effect of X on the probability of $Y = 1$, averaged across the sample. The computation works as follows: for each observation, compute the marginal effect (the slope of the probability curve at that point), then average across all observations. This is why AMEs from logit and LPM coefficients are often similar: both are averaging the effect across the sample, just using different functional forms. The `avg_slopes()` function does all this automatically. Use it.

Why marginal effects vary



25/1

This diagram shows WHY the marginal effect varies. The tangent line at each point represents the marginal effect — the instantaneous rate of change in probability. At the extremes (near 0 or 1), the curve is flat, so the marginal effect is small. In the middle (near 0.5), the curve is steepest, so the marginal effect is largest. This is the fundamental difference from OLS, where the marginal effect is the same everywhere (a constant slope). The AME averages across all these different tangent lines, giving a single summary number.

Marginal effects at representative values

- Instead of averaging, evaluate at specific values
- Example: “What is the effect of education on voting for a 40-year-old woman?”
- In R:
→ `avg_slopes(model, newdata = datagrid(age = 40, female = 1))`
- Useful when the effect varies substantially across groups
- E.g., the effect may be larger for young people than for old people

26/1

Marginal effects at representative values (MER) let you ask more specific questions than the AME. The AME gives you one number for the whole sample; MER gives you the effect for a specific type of observation. This is especially useful when you have interactions or when the predictor's effect differs substantially across the sample. For example, the marginal effect of income on voting might be large for young people (who are on the steep part of the curve) and small for old people (who already have high voting probabilities).

Plotting predicted probabilities

- The best way to communicate logit results to any audience
- Show how $\Pr(Y = 1)$ changes across values of X
- Include confidence bands
- In R:
 - `plot_predictions(model, condition = "x")`
 - Plots the S-curve with 95% confidence interval
- Much more informative than a table of log-odds

27/1

If students remember one thing from this section, it should be this: always plot predicted probabilities. A table of log-odds coefficients is nearly uninterpretable for a general audience. A plot showing how the predicted probability changes across a predictor is immediately understandable. The `plot_predictions()` function from the `marginaleffects` package makes this trivial. You can also condition on multiple variables: `plot_predictions(model, condition = c("education", "female"))` will show separate S-curves for men and women.

Comparing LPM and logit

- In many cases, LPM and logit give similar results
 - Especially for average marginal effects
 - Especially when probabilities are in the 0.2–0.8 range
- Where they differ:
 - Predicted probabilities near 0 or 1
 - LPM can go outside [0, 1]; logit cannot
 - Marginal effects at extreme values
- A good practice: estimate both and compare

28/1

Summarize the key practical takeaway: for most applied research, the LPM and logit tell a similar story. When they disagree, the logit is usually more appropriate because it respects the bounded nature of probabilities. A common workflow in applied papers is to present the LPM as the main specification (because the coefficients are easy to interpret) and show logit AMEs as a robustness check, or vice versa. If both give similar results, your findings are robust to the functional form assumption.

Model fit for logit

- No R^2 in the usual sense (MLE, not OLS)
- Alternative measures:
 - **Pseudo- R^2** : compares model to null model (McFadden's)
 - **AIC / BIC**: penalized likelihood (lower = better)
 - **Classification accuracy**: percent correctly predicted
 - **ROC / AUC**: trade-off between true and false positives
- None is perfect; use as rough guides
- Reported automatically by `modelsummary()` and `performance::r2()`

29/1

Don't spend too long here. The main point is that the familiar R^2 doesn't exist for logit models. McFadden's Pseudo- R^2 is the most common substitute, but its values tend to be much lower than OLS R^2 (a Pseudo- R^2 of 0.2 is considered quite good). AIC is useful for comparing models with different predictors. Classification accuracy (what percent of observations does the model correctly classify as 0 or 1) is intuitive but depends on the threshold you choose. ROC curves show how the true positive rate varies with the false positive rate as you change the threshold. `modelsummary()` reports several of these automatically.

A note on probit

- Logit uses the logistic CDF as the link function
- **Probit** uses the normal (Gaussian) CDF instead
- In R: `glm(y ~ x, family = binomial(link = "probit"), data = df)`
- In practice, logit and probit give very similar results
 - Probit coefficients \approx logit coefficients $\times 0.625$
 - Predicted probabilities are nearly identical
- Logit is more common in political science; probit in economics
- Use whichever is conventional in your field

30/1

This is a brief mention — don't dwell on it. The only practical difference between logit and probit is the shape of the link function: logistic has slightly fatter tails than the normal distribution. For applied work, they give virtually indistinguishable results in terms of marginal effects and predicted probabilities. The coefficient scale differs (multiply logit by about 0.625 to get probit, or by 1.6 the other way), but once you convert to marginal effects, the numbers are nearly the same. The choice between them is mostly a matter of disciplinary convention. Some economists prefer probit because it connects to the latent variable interpretation; political scientists tend to use logit because odds ratios are more familiar.

Roadmap

31/1

Complete workflow in R

```
lpm <- lm(vote ~ age + income + educ, data = df)
logit <- glm(vote ~ age + income + educ,
              family = binomial, data = df)
modelsummary(list("LPM" = lpm, "Logit" = logit),
             vcov = list("robust", NULL))
avg_slopes(logit)                      # compare AMEs to LPM
plot_predictions(logit, condition = "income")
```

32/1

Walk through this as a template workflow. Step 1: estimate both the LPM and logit with the same predictors. Step 2: present them side by side with `modelsummary()` — note that we use robust SEs for the LPM (because of heteroskedasticity) and default SEs for logit. Step 3: compute AMEs for the logit and compare them to the LPM coefficients. Step 4: plot predicted probabilities across a key predictor. This four-step workflow gives you a complete picture of the binary outcome analysis. Students should follow this pattern in the lab and assignment.

Decision tree: When to use which?

- **Use LPM when:**

- You want simple, direct interpretation
- Probabilities are in the middle range
- You mainly care about average effects

- **Use logit when:**

- You need bounded predicted probabilities
- Many observations have extreme probabilities (near 0 or 1)
- You want to properly account for the binary nature of Y

- In practice: estimate both, report the one most appropriate
- Always report marginal effects, not just log-odds

33/1

This is practical advice for their research. The honest answer is: in most applied settings, it doesn't matter much which one you use, as long as you interpret the results properly. The exception is rare events (civil war onset, corporate bankruptcy, etc.), where the LPM can perform poorly. The last point is crucial: if you use logit, you MUST convert to marginal effects or predicted probabilities. Reporting log-odds coefficients is not informative. If you use the LPM, the coefficients are already in probability units, which is one of its main advantages.

Summary: Key takeaways

- Binary outcomes require special treatment
- The LPM is simple but has known limitations
- Logistic regression bounds probabilities between 0 and 1
- Log-odds are not intuitive — use marginal effects and predicted probabilities
- AMEs from logit are often similar to LPM coefficients
- Always estimate both and compare; always plot predicted probabilities

34/1

Recap the session. The three most important practical takeaways: (1) Know when and why to use logit instead of LPM. (2) Never interpret logit coefficients on the log-odds scale — always convert to marginal effects or predicted probabilities using the `marginaleffects` package. (3) Always plot predicted probabilities — this is the single most effective way to communicate logit results. These skills will come up again in the sessions on panel data and other outcome types.

For next week

- Read Gelman et al., chapters 11–12
- Read Arel-Bundock, Greifer, and Heiss (2025), chapters 1–4
- Complete Assignment 3
- Next session: Model interpretation and diagnostics
 - Beyond coefficient tables
 - Visualizing model results
 - Residual diagnostics
 - Influence and outliers

35/1

Mention that the Gelman chapters cover model checking and diagnostics, which we'll discuss next week. The Arel-Bundock et al. book (the `marginaleffects` online textbook) gives practical guidance on interpreting model results, which applies to both the binary outcome models from today and the continuous outcome models from last week. Assignment 3 has two parts: the in-class lab they started today and the take-home exercises due before next session. Encourage them to start the take-home early.

Questions?

36/1