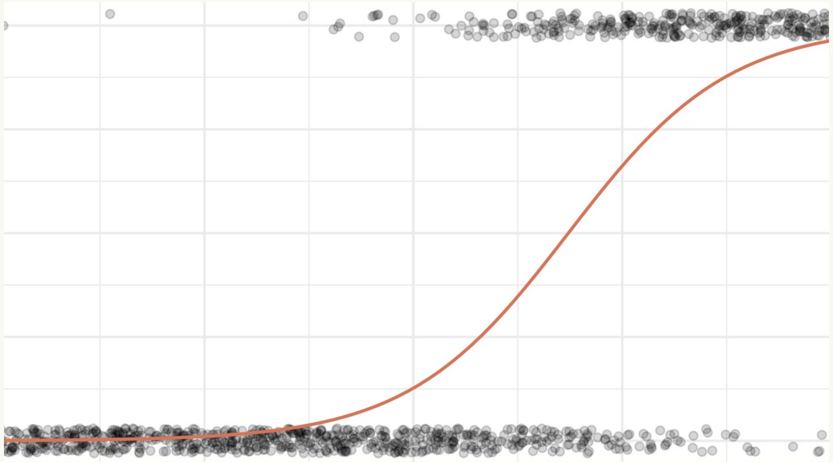
Lecture 6: Logistic Regression

Paul Röttger

Applied Analytical Statistics

24th of February 2026

Plan for today | Logistic regression and GLMs



Today we **keep expanding regression** to more types of outcome variables.

We focus on **logistic regression** for binary outcome variables.

We introduce **generalised linear models** (GLMs) as a more general framework.

We also cover different approaches for **evaluating and comparing model fit**.

Recap | Bernoulli distribution

The **Bernoulli distribution** is the discrete probability distribution of a random variable Y which takes the value 1 with probability p and the value 0 with probability $1-p$.

$$f_Y(p) = \begin{cases} p \text{ if } Y = 1 \\ 1 - p \text{ if } Y = 0 \end{cases}$$

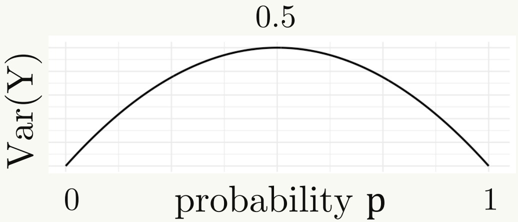
: generalised version of a single coin toss

$$E(Y) = p$$

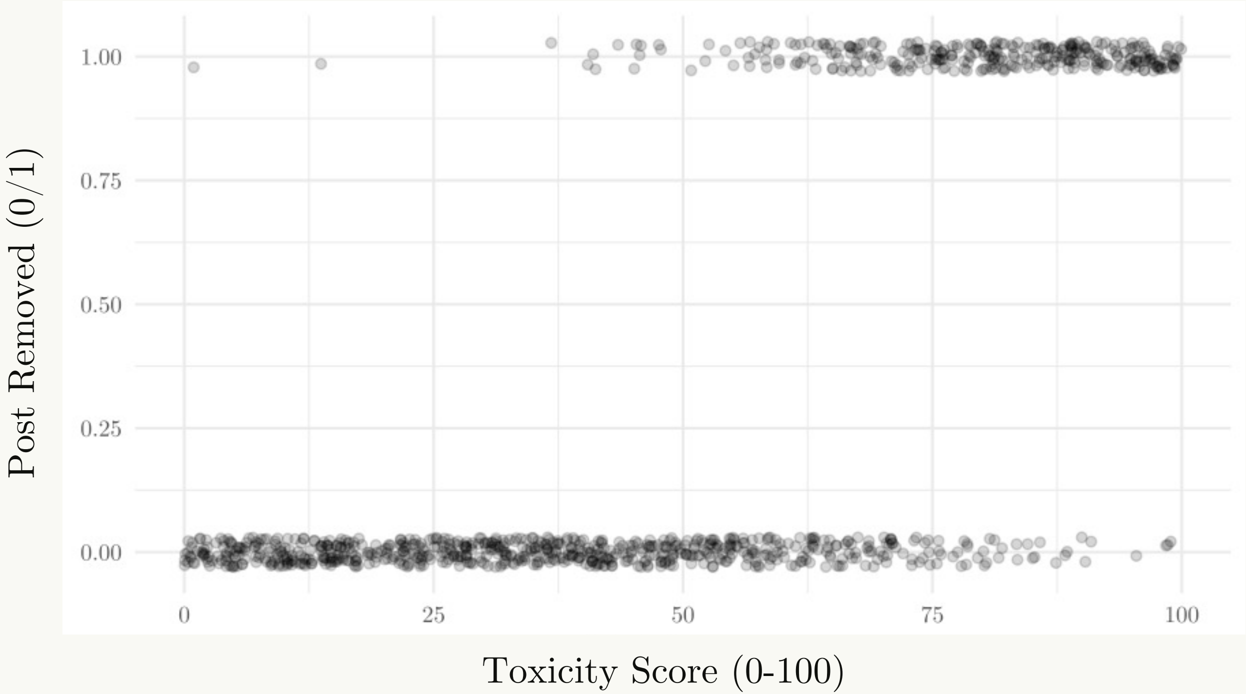
: expected value, i.e. mean = probability p

$$Var(Y) = p(1-p)$$

: variance depends on probability p



Binary outcomes | Working example



Data: 1,000 content moderation decisions

(simulated)

Outcome: post removed?

Regressors:
- <u>Toxicity score (0-100)</u>
- Verified author (0/1)
- Number of followers

Binary outcomes | Conditional probability

Linear regression models the **conditional mean** of an outcome:

$$E(Y \mid X_1, X_2, \dots, X_k) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

The conditional mean of a binary outcome is the conditional probability of $Y = 1$:

$$E(Y \mid \mathbf{X}) = P(Y = 1 \mid \mathbf{X})$$

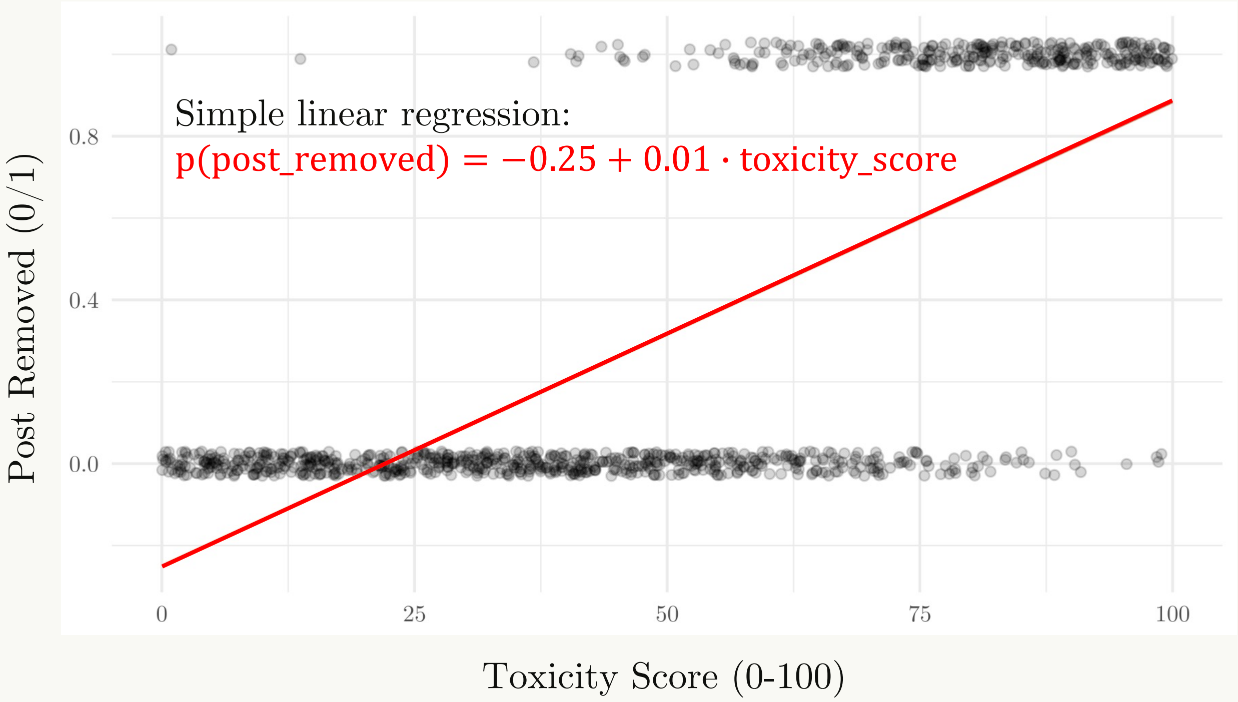
where Y is a **Bernoulli random variable** with $Y \in \{0,1\}$.

In principle, we could still fit **OLS** to binary outcome data:

$$P(Y = 1 \mid \mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$$

where β_1 = change in probability of Y for a one-unit increase in X_1

Binary outcomes | Linear regression



$\hat{\beta}_1 = 0.01$: Each 1-point increase in toxicity score is associated with a 1pp increase in the probability of a post being removed.

$\hat{\beta}_0 = -0.25$: On average, a post with a toxicity score of 0 has a **-25% chance of being removed**.

Binary outcomes | Why linear regression fails

Problem #1: Probabilities are bounded:

$0 \leq P(Y = 1 \mid \mathbf{X}) \leq 1 \quad \text{for all } \mathbf{X}$ whereas linear functions are **unbounded**.

Linear regression for binary outcomes can predict probabilities < 0 and > 1.

Problem #2: The relationship between $\mathbf{X}$ and $P(Y = 1 \mid \mathbf{X})$ must be non-linear.

In linear regression $\partial P(Y = 1 \mid \mathbf{X}) / \partial X_1 = \beta_1$ is constant for all $\mathbf{X}$.

BUT a constant rate of change is not compatible with hard limits at $Y = 0$ and $Y = 1$.

Therefore, the linear model is misspecified for binary outcomes.

Logistic regression | Key ingredients

Our goal is to **ensure boundedness of linear predictions to range (0,1)**.

Let $Y \in \{0,1\}$ be a Bernoulli random variable with $p = P(Y = 1) = E(Y)$. Then:

$$\text{odds} = \frac{p}{1-p}$$

map p from $(0,1) \rightarrow (0,\infty)$

$p = 0.8 \leftrightarrow$ odds of 4:1

$p = 0.5 \leftrightarrow$ odds of 1:1

$p = 0.01 \leftrightarrow$ odds of 1:99

🤔 Which transformation can get rid of the other boundary at 0?

By taking the natural logarithm, we arrive at the **logit of p** (aka log-odds):

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

maps p from $(0,1) \rightarrow (-\infty,\infty)$

Logistic regression | The logistic model

For **univariate logistic regression** we model:

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

which we denote as

$$\mathrm{logit}(p_i) = \beta_0 + \beta_1 X_i$$

$$p_i = E(Y|X_i)$$

→ Both sides of the equation take values in (-∞,∞).

By exponentiating we can show this is equivalent to:

linear model still in here!

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

which we denote as

$$p_i = \mathrm{logit}^{-1}(\beta_0 + \beta_1 X)$$

"linked" to conditional mean of outcome by logit

→ Both sides of the equation take values in (0,1).

Interpretation | Predicted probabilities

We now have a **well-specified logistic model** that we can fit for binary outcomes, where every observation X_i corresponds to a predicted probability p_i.

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

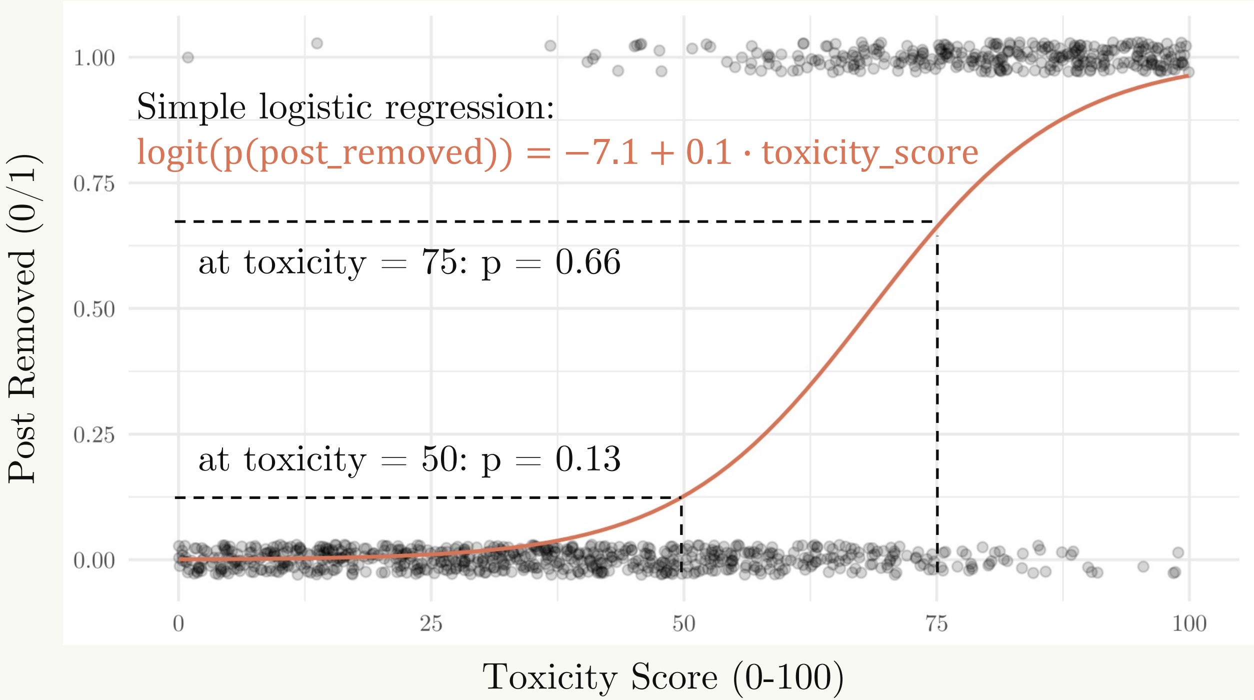
which is equivalent to

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

Direct interpretation of β_1 is **unintuitive**: change in **log-odds** for a one-unit increase in X.

More intuitive: how does **predicted probability change as X changes**?

Interpretation | Predicted probabilities (cont'd)



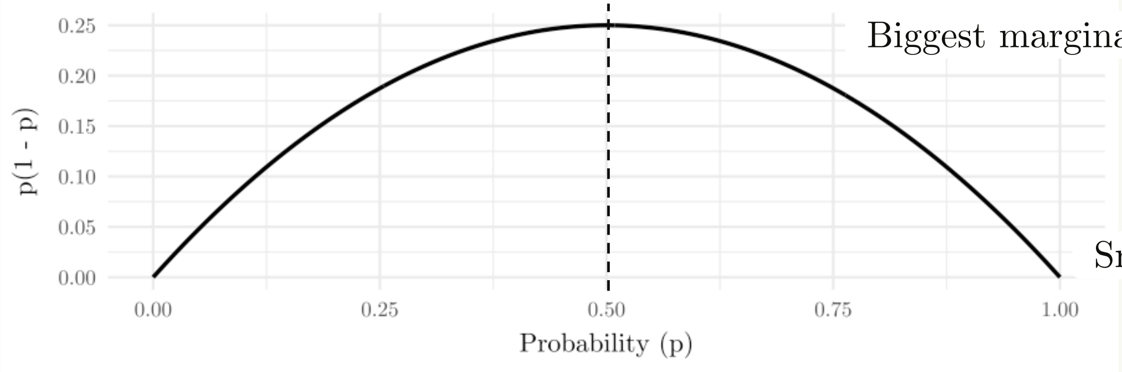
An increase in toxicity score from 50 to 75 is associated with an increase in probability of a post being removed by 53pp.

Interpretation | Marginal effects

More generally, we can quantify the **slope of the fitted logistic regression curve**:

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$ taking derivative: $$\frac{\partial p}{\partial X_i} = \beta_1 p_i (1 - p_i)$$ → "marginal effect" of X (not causal)

Slope is not constant, depends on predicted probability level p_i multiplied by constant β_1.

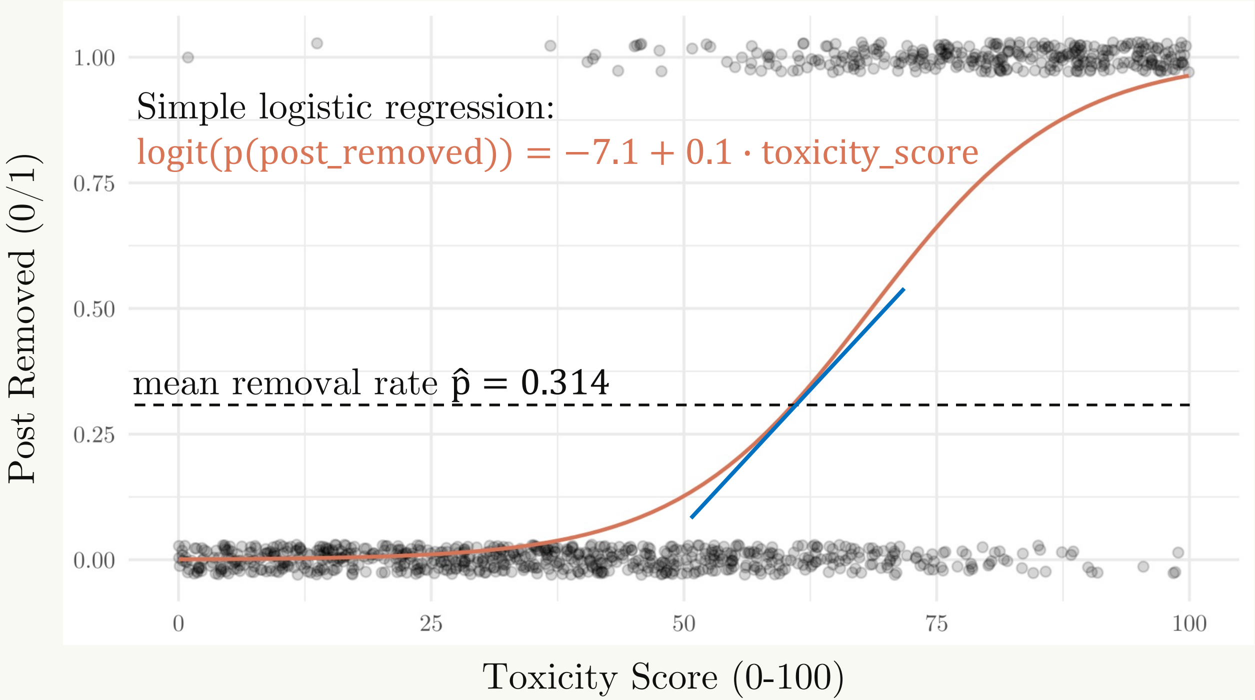


Biggest marginal effect at $p_i = 0.5$

→ S-shape of logistic curve

Small marginal effect at extremes

Interpretation | Marginal effects (cont'd)



How do we interpret $\hat{\beta}_1$?

$\partial p / \partial X = \hat{\beta}_1 \hat{p}(1-\hat{p})$
$= 0.1 \cdot 0.31 \cdot (1-0.31)$
$= 0.02$

At mean removal rate $\hat{p}$, a 1-point increase in toxicity score is associated with a 2pp increase in probability of a post being removed.

Interpretation | Odds ratios

Finally, we can interpret coefficients in terms of **odds ratios** (OR).

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$$ is equivalent to $$\text{odds} = \frac{p}{1-p} = e^{\beta_0 + \beta_1 X}$$

As we increase regressor X by one unit:

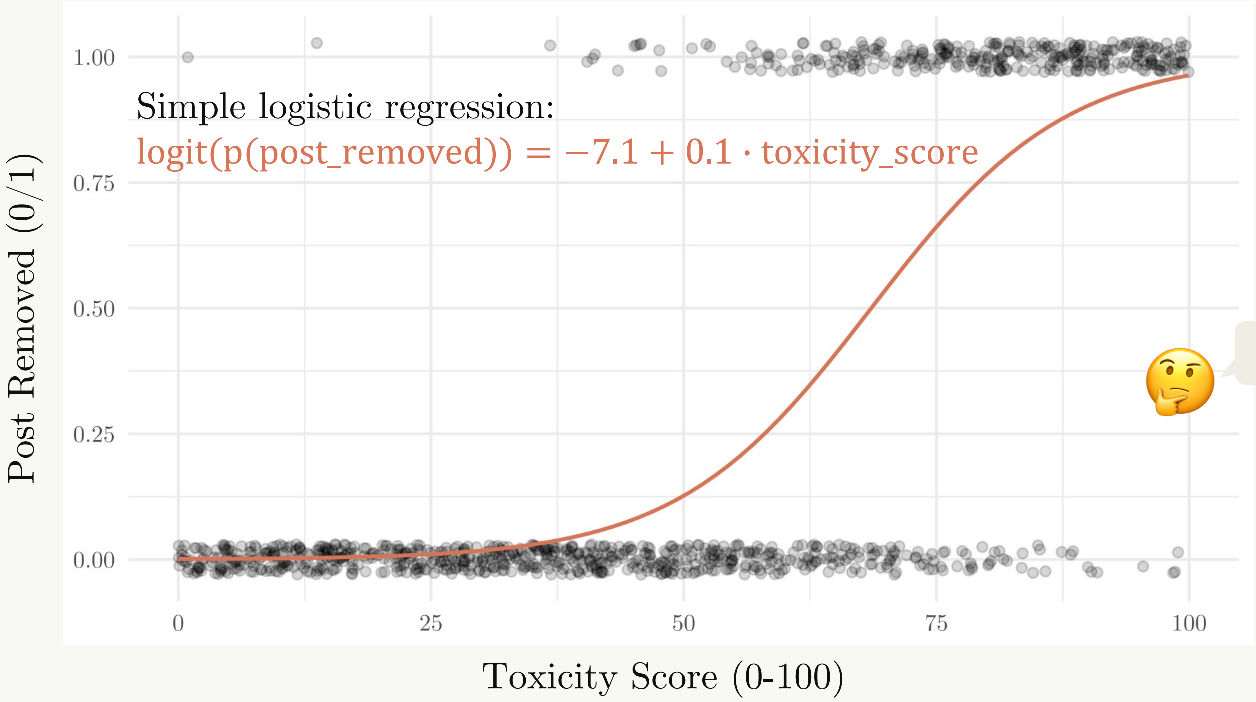
$$\text{OR} = \frac{\text{odds}(X+1)}{\text{odds}(X)} = \frac{e^{\beta_0+\beta_1(X+1)}}{e^{\beta_0+\beta_1 X}} = e^{\beta_1}$$ which is constant across X!

For every one-unit increase in X, the odds in favour of Y=1 multiply by e^{β_1}.

→ **OR>1**: on average, Y=1 becomes more likely as X grows

→ **OR<1**: on average, Y=1 becomes less likely as X grows

Interpretation | Odds ratios (cont'd)



$\hat{\beta}_1 = 0.1 \rightarrow \mathrm{OR} = e^{0.1} = 1.1$:

Each one-point increase in toxicity score is associated with a 10% increase in the odds of a post being removed.

🤔 What about a 10-point increase?

$\mathrm{OR} = e^{0.1*10} = 2.7$:

Each 10-point increase in toxicity score is associated with a 170% increase in the odds of a post being removed.

Logistic regression | Multiple regressors

Multivariate logistic regression is the direct analogue of multivariate linear regression:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}$$

where

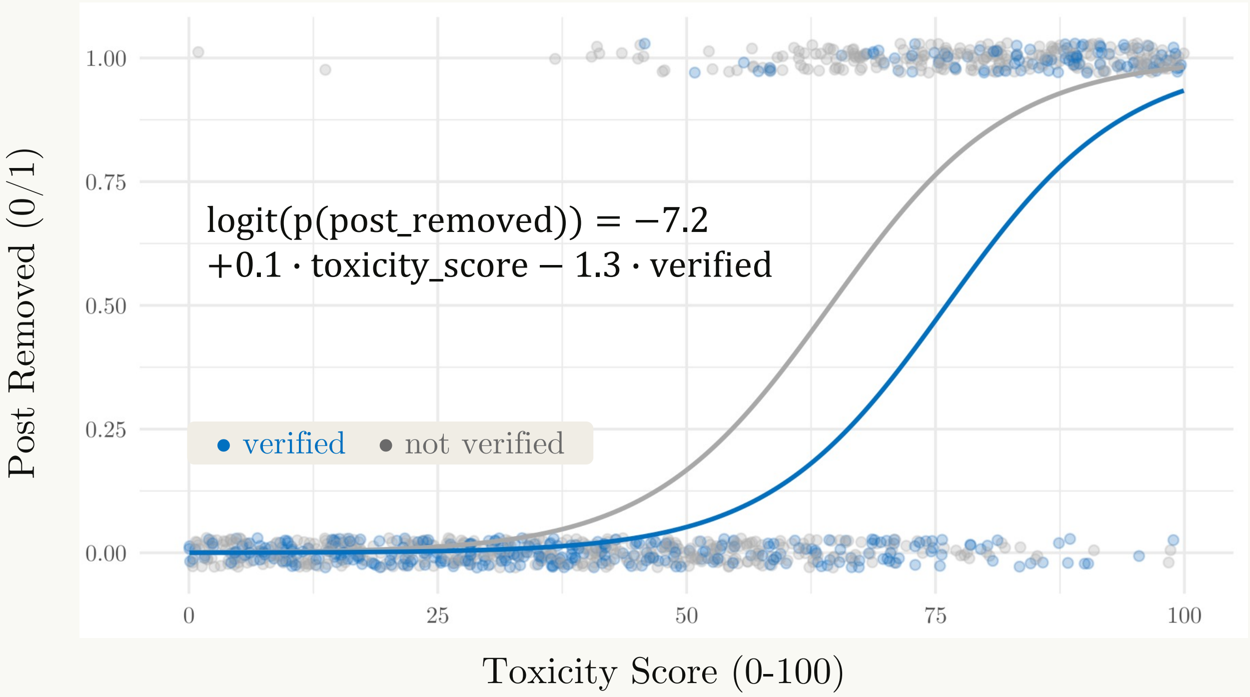
$$p_i = P(Y_i = 1 \mid X_i)$$
$$Y_i \sim \text{Bernoulli}(p_i)$$

We now interpret coefficients **ceteris paribus**, i.e. holding other regressors constant.

β_j = change in log-odds of $Y = 1$ for a one-unit increase in X_j, ceteris paribus

e^{β_j} = multiplicative change in odds of $Y = 1$ for a one-unit increase in X_j, ceteris paribus

Logistic regression | Multiple regressors (cont'd)



$\hat{\beta}_1 = 0.1 \rightarrow \mathrm{OR} = 1.1$:
Each 1-point increase in toxicity score is associated with an increase in the odds of a post being removed by 10%, **ceteris paribus**.

$\hat{\beta}_2 = -1.3 \rightarrow \mathrm{OR} = 0.27$:
Posts from verified accounts have 73% lower odds of being removed than from non-verified accounts, **ceteris paribus.**

Logistic regression | Estimating coefficients

In linear regression, we minimised squared residuals:

$\hat{\beta} = \arg\min \sum (Y_i - \hat{Y}_i)^2$ by OLS, producing closed-form solution $\hat{\beta} = (X'X)^{-1}X'Y$

This breaks down for logistic regression (and other non-linear models).

Instead, we estimate parameters using **Maximum Likelihood Estimation** (**MLE**):

$L(\beta) = \prod_i p_i^{Y_i} (1 - p_i)^{1-Y_i}$ under a Bernoulli model, where $p_i = \frac{1}{1 + e^{-X_i\beta}}$

MLE chooses β that maximises $L(\beta)$ via p_i, by numerical optimisation.
We **find the coefficients that make the observed outcomes most likely**.

Logistic regression | Assumptions

Assumptions for consistent estimates, i.e. unbiased in large samples:

Correct functional form: The log—odds are a linear function of parameters $\boldsymbol{\beta}$.

Exogeneity: Regressors X are uncorrelated with unobserved determinants of Y.

Assumptions for MLE to function:

No perfect multicollinearity: Regressors X are not perfectly correlated with each other.

No complete separation: Outcome Y is not perfectly predicted by X.

Specific to logistic regression

Assumptions for correct standard errors and inference:

Independence: Observations are not correlated with each other.

$\approx$ homoskedasticity in OLS

Correct variance specification: Conditional variance depends only on mean: $\mathrm{Var}(Y_i \mid X_i) = p_i(1 - p_i)$

Logistic regression | Uncertainty

Unlike linear regression, the logistic model does not contain a Gaussian error term ε:

$\mathrm{Y_i} \sim \mathrm{Bernoulli}(\mathrm{p_i})$ where $\mathrm{p_i} = \mathrm{logit}^{-1}(\boldsymbol{X_i}\beta)$ and $\mathrm{Var}(\mathrm{Y_i} \mid \mathrm{X_i}) = \mathrm{p_i}(1 - \mathrm{p_i})$

Under logistic regression assumptions, for large n, we can show that:

$\hat{\boldsymbol{\beta}} \sim \mathrm{N}(\boldsymbol{\beta}, [\mathbf{X}'\mathbf{W}\mathbf{X}]^{-1})$ where $\mathbf{W} = \mathrm{diag}\big(\mathrm{p_i}(1 - \mathrm{p_i})\big)$

Note that there is no separate variance parameter σ^2.
Noise is intrinsic to the Bernoulli outcome distribution and depends on $\mathrm{p_i}$.

Logistic regression | Inference

To test for significance of coefficients in logistic regression, we use a **Wald test**:

$$z = \frac{\hat{\beta}_j - \beta_{j,0}}{SE(\hat{\beta}_j)}$$

The test statistic measures the distance between our sample coefficient and the coefficient value under the null in SE units (see Week 3).

For large n, under H_0, z approximately follows a standard normal distribution: $z \sim N(0,1)$.

🤔 Why can we use standard normal rather than t-distribution?

In logistic regression, there is no constant variance σ^2 that requires separate estimation.
→ **no df correction required**, standard errors follow directly from data and MLE

Generalised linear models | Motivation

We now covered two regression models for two types of outcome variables:

Linear regression for unbounded continuous outcome variables

Logistic regression for binary outcome variables

These models share a common structure:

$$E(Y_i|\mathbf{X}_i) = \mathbf{X}_i\boldsymbol{\beta} \quad \text{and} \quad p_i = \text{logit}^{-1}(\mathbf{X}_i\boldsymbol{\beta}) \;\; \text{where} \;\; p_i = E(Y_i|\mathbf{X}_i)$$

Both model **conditional means** and include a **linear component**.

Generalised linear modelling (GLM) extends this structure to other outcomes.

→ <u>all</u> regression is about describing how the expected value of Y changes with X.

Generalised linear models | Three components

Generalised linear modelling (GLM) is a framework for statistical analysis that includes linear regression and logistic regression as special cases. GLMs have three components:

The **systematic component** is the linear predictor $X\beta$.

→ same in all GLMs

The **random component** specifies the distribution of the outcome variable Y.

→ modelling assumption based on type of outcome variable, determines variance structure

$Y_i \sim N(\mu_i, \sigma^2)$ where $\mu_i = E(Y_i|X)$

$Y_i \sim \text{Bernoulli}(p_i)$ where $p_i = E(Y_i|X)$

The **link function** connects the expected value of Y to the linear predictor: $g(E(Y|X)) = X\beta$

→ ensures that transformed outcome is linearly related to predictors

Identity link: $g(\mu_i) = \mu_i$

Logit link: $g(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$

Generalised linear models | Estimating coefficients

We fit all GLMs, like logistic regression, using **maximum likelihood estimation** (MLE).
$\rightarrow$ finding the parameters that make the observed data most likely

The likelihood function depends on the random component and link function.

```
glm(
  post_removed ~ toxicity_score + is_verified,
  data = df,
  family = binomial(link = "logit")
  )
```

logistic regression in R

GLMs using MLE are fitted by numerical optimisation.

Poisson regression | GLM version

RQ: Is higher ad spend associated with higher engagement on social media ads?

Data: Number of likes for 1,000 Instagram ads.

The outcome is a non-negative integer with no upper boundary $Y \in \{0,1,2,\ldots\}$

The corresponding **random component** is a Poisson distribution:

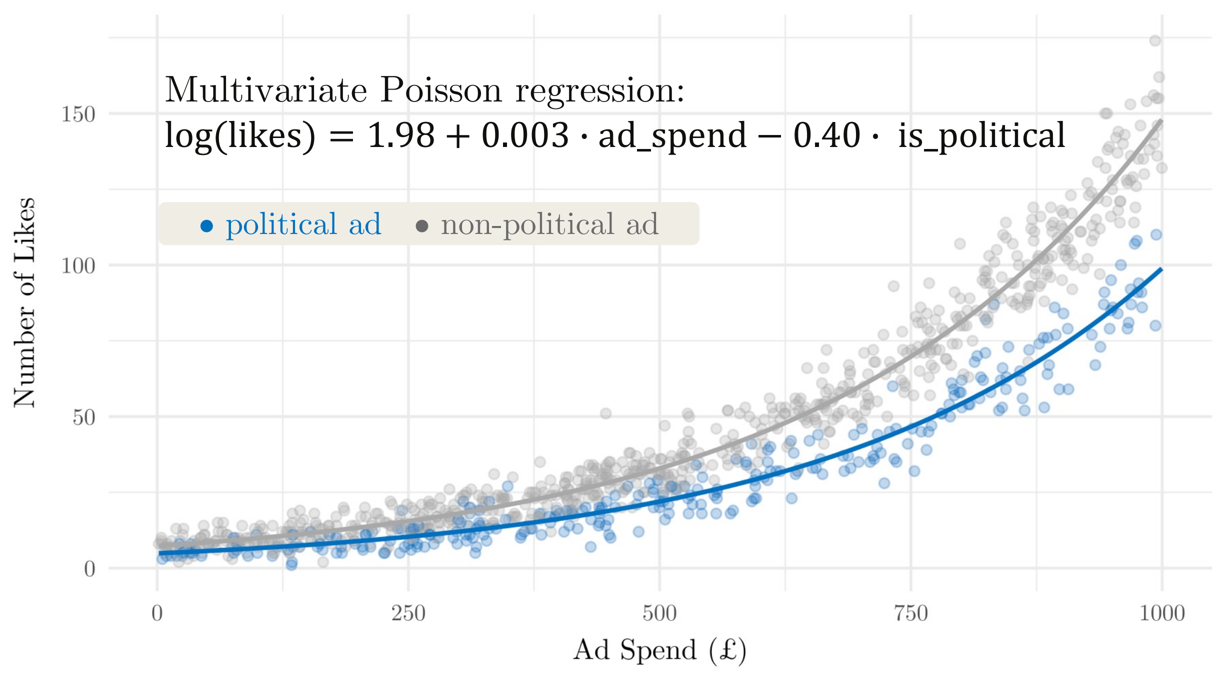
$Y_i \sim \text{Poisson}(\lambda_i)$ where $\lambda_i = E(Y_i|X)$ and $\underline{\text{Var}(Y_i|X) = \lambda_i}$

Strong assumption: variance = mean.
Fit negative binomial if violated.

The **link function** needs to map $(0,\infty) \rightarrow (-\infty,\infty)$:

$g(\lambda_i) = \log(\lambda_i)$ from which follows the **Poisson GLM** $\log(\lambda_i) = \mathbf{X}_i\beta$

Poisson regression | Coefficient interpretation



$\hat{\beta}_1 = 0.003$, $e^{0.003*100} = 1.35$: Each 100£ increase in ad spend is associated with a 35% increase in expected likes, **ceteris paribus**.

$\hat{\beta}_2 = -1.3$, $e^{-0.4} = 0.67$: Political ads, on average, receive 33% fewer likes than non-political ads, **ceteris paribus.**

Model comparison | Nested models

Model A is **nested** in model B if B contains all regressors in A plus additional regressors:

A: `likes ~ ad_spend` is **nested** in B: `likes ~ ad_spend + is_political`

Nested models allow for **stepwise theoretical expansion**:
→ fit baseline THEN add controls THEN add interactions etc.

Nested models allow us to **understand omitted variable bias**:
→ how much of **ad_spend** coefficient in A was indirect association via **is_political**?

Nested models enable **joint hypothesis testing:**
→ do **is_political** and other controls **jointly** improve our model?

Model comparison | The Likelihood Ratio (LR) test

A **likelihood ratio (LR) test** compares how well **two <u>nested</u> models** explain observed data:

$\mathrm{LR} = -2(\log \mathrm{L}_{\mathrm{restricted}} - \log \mathrm{L}_{\mathrm{full}})$ where $\log \mathrm{L}$ is the fitted model log-likelihood.

Under H_0 of no difference, the test statistic follows a chi-squared distribution: $\mathrm{LR} \sim \chi^2_{\mathrm{df}}$
where df = number of additional parameters (coefficients) in the full model.

This is a very flexible test for significance of one or multiple coefficients:
Does adding these regressors significantly improve model fit?
(full vs. restricted)

For adding a single predictor, for large n, LR test ≈ Wald test.

Model comparison | AIC

We may also want to compare **non-nested models with different functional forms**.

For this, we can use the Akaike Information Criterion (AIC):

$\mathrm{AIC} = -2\ \mathrm{LogL} + 2\mathrm{k}$ where k is the number of parameters

When comparing models, a **lower AIC indicates better model fit**.

When comparing two models A and B, where AIC of A is smaller than AIC of B:

$\exp((\mathrm{AIC}_\mathrm{A} - \mathrm{AIC}_\mathrm{B})/2)$ is the relative likelihood of B with respect to A

Example: value of 0.5 $\rightarrow$ B is 50% as likely as A to minimise expected information loss.

Recap | Key takeaways from Week 6

[TO ADD]