Introduction to Modelling I Statistical modelling with the Generalised Linear Model

ISAIR

Yang Liu

January 23, 2025

What is modelling?

Statistical models allow

- Description of trends
- Testing of hypotheses (causal inference)
- Prediction from data

All models are collections of assumptions

> *"Essentially, all models are wrong but some models are useful" - George E. P. Box*

What is modelling?

The application of general principles of statistical modelling and inference to geostatistical problems

- formulate a model for the data;
- use likelihood-based methods of inference;
- answer the scientific question.

(P. J. Diggle, Tawn, and Moyeed 1998)

Linear models

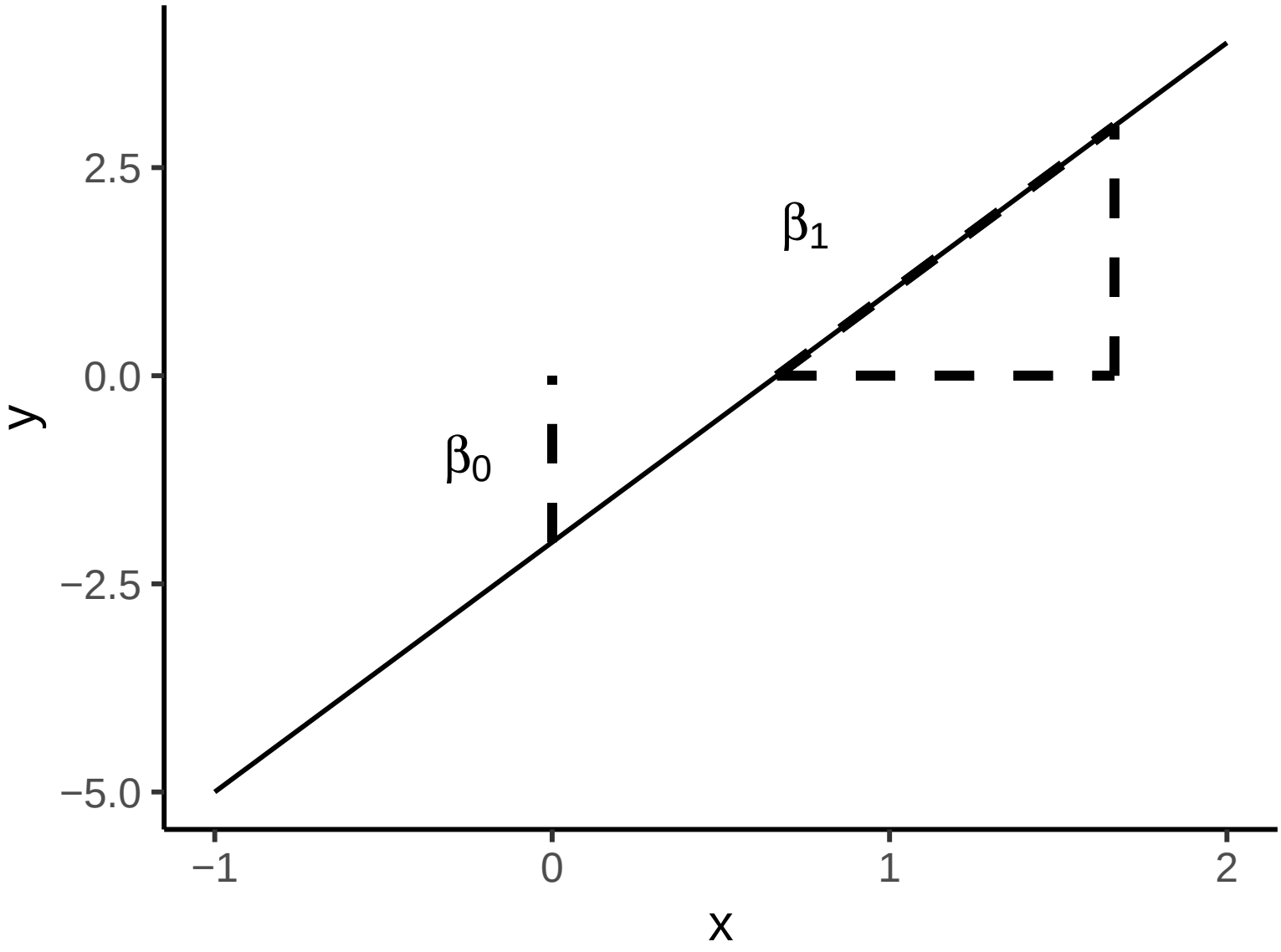
- One of the most basic models is the linear model with one predictor, i.e.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where

- y are the observations of the outcome/response/dependent variable
- x are the observations of the predictor/covariate/independent variable
- β_0, β_1 are the coefficients of the model
- ε are the assumed *errors* which follow a normal distribution, $\varepsilon_i \overset{\text{i.i.d.}}{\sim} N\left(0, \sigma^2\right)$

Linear models



Linear models

- We can extend to a linear model with more than one predictor, e.g. m covariates $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m$

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_m x_{mi} + \varepsilon_i$$

- We can also write this in matrix notation

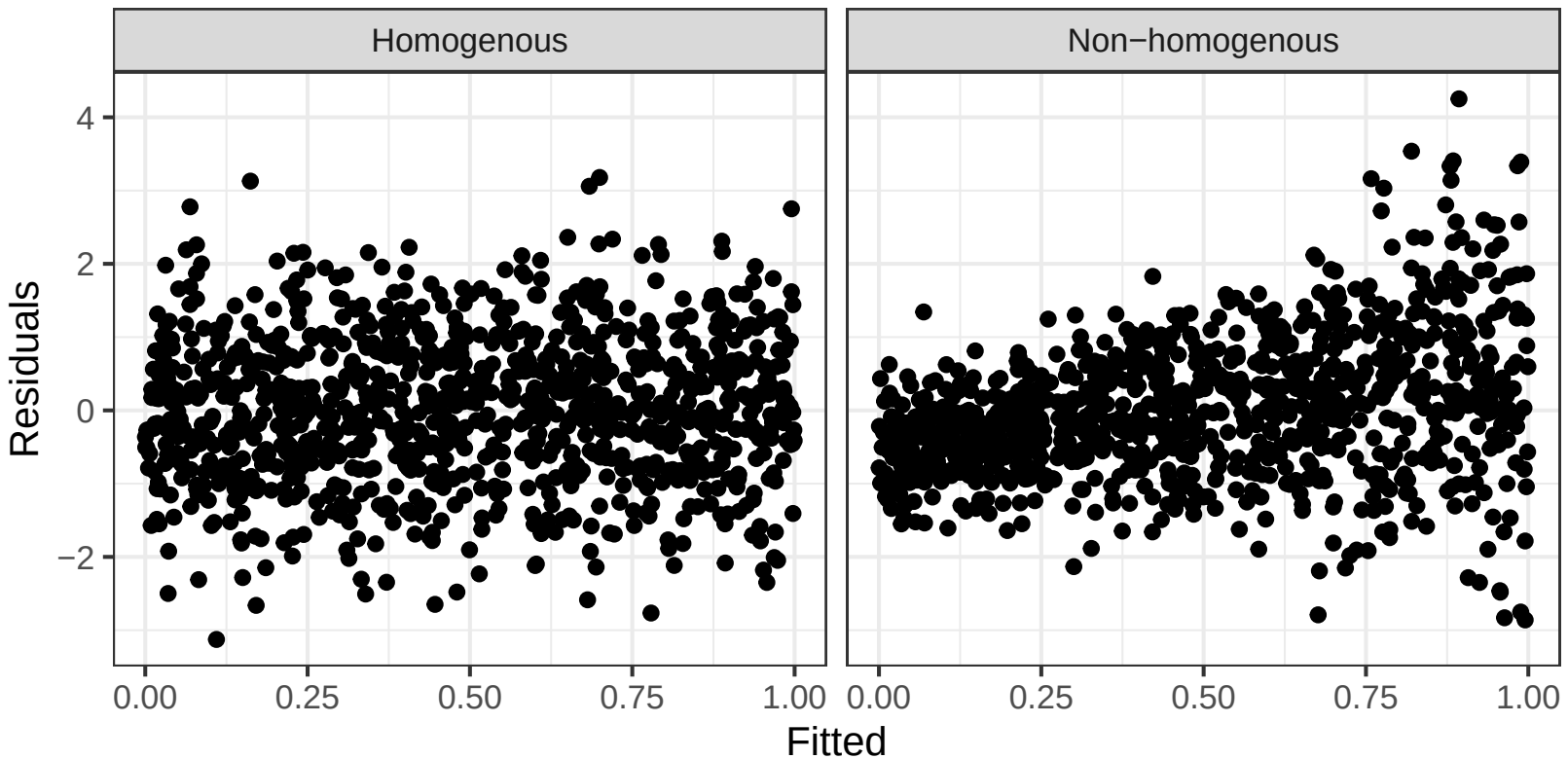
$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\varepsilon_i \overset{\text{i.i.d.}}{\sim} N\left(0, \sigma^2\right)$$

where $X = [\boldsymbol{x}_1 | \boldsymbol{x}_2 | \ldots | \boldsymbol{x}_m]$

Linear models - assumptions

- Linearity: every x_k has a linear effect on $\mathbb{E}(y)$
- Normality of errors - any unexplained variability follows a normal distribution
- Homogeneity of errors - all errors described by the same normal distribution

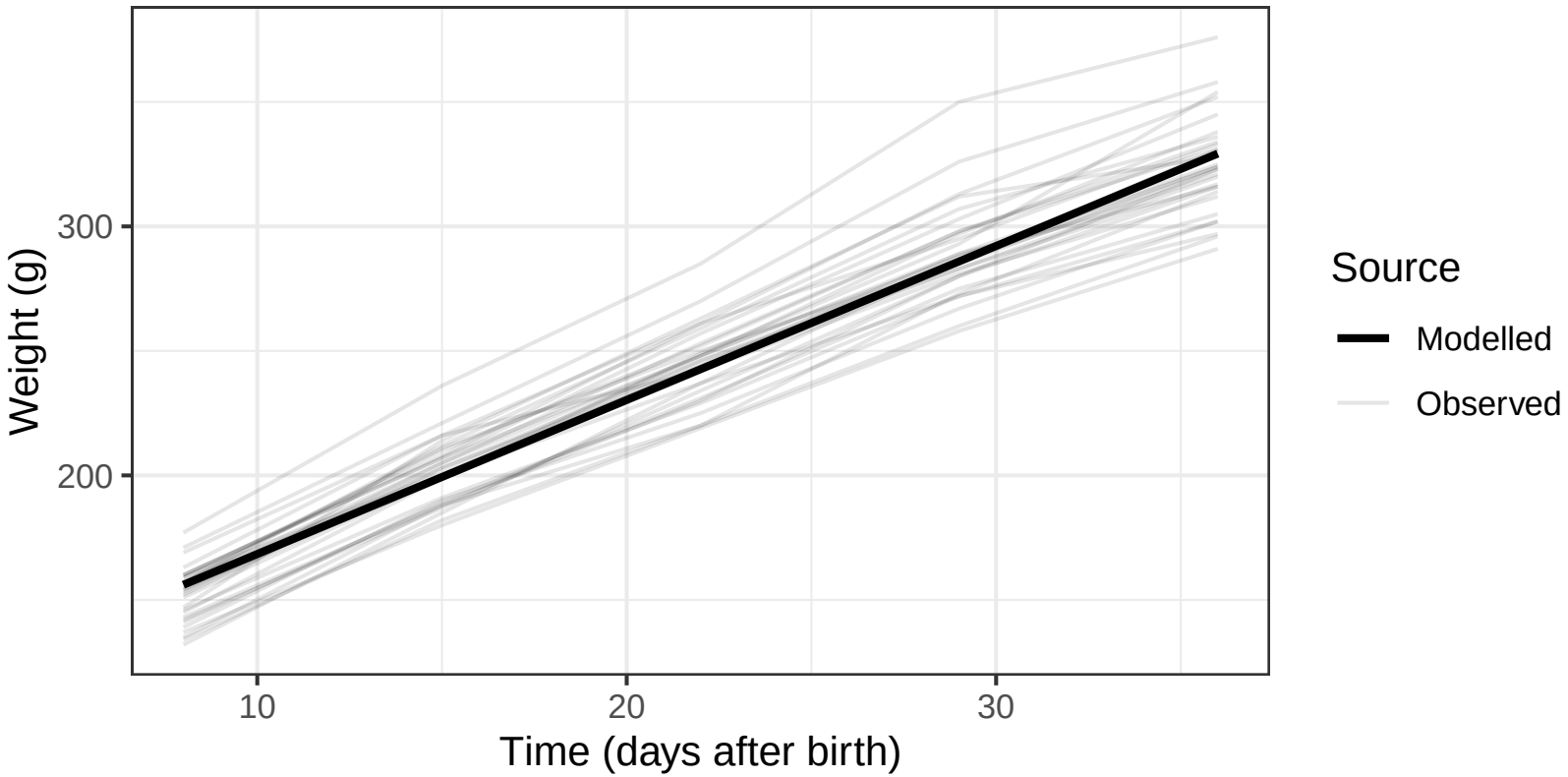


Linear models - reality

- Real epi/trial/health data is collected with some structure
- Measured variables might not explain *all* variation
- We might need to account for some group-specific effect
  - group might be an individual in a repeated measures design
  - group might be spatial location in a surveillance design

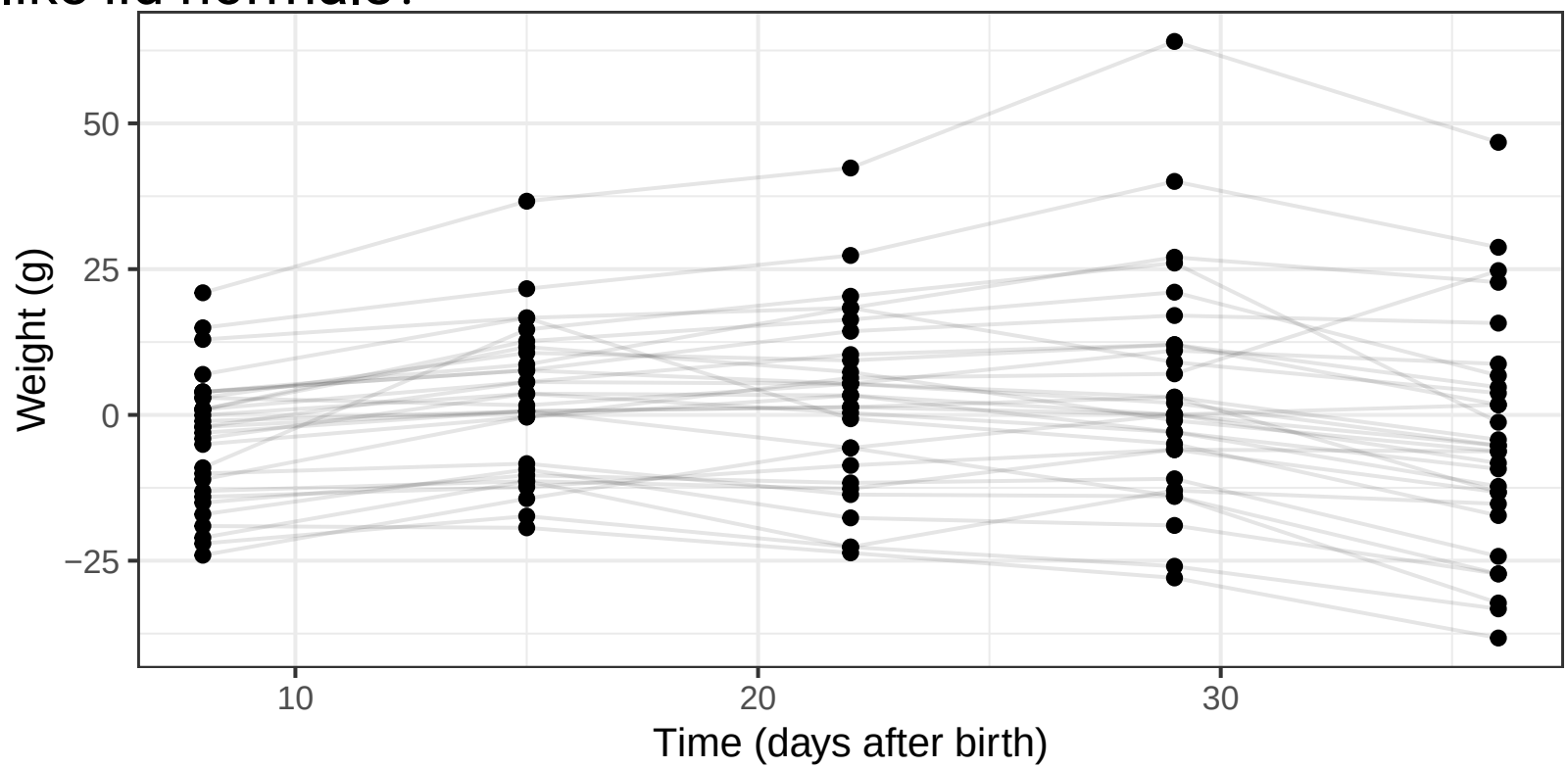
Linear models - example

Data from Gelfand and Smith (1990) on the weight of 30 rats
Each rat has its own birth weight and its own growth rate



Linear models - example

If we look at the residuals from the model above, do they look like iid normals?



Independence

- One of the main assumptions of generalised linear models is that the observations y_i are mutually independent.

- The value of one observation does not give me any information about the value of another observation.

$$\text{Covariance}(y_i, y_j) \equiv 0$$

Do you think this assumption is still valid for the rat data?
Do you think this assumption is still valid for spatial data?

Independence

First law of geography: close things are more related than distant things (Tobler 1970)
There are two main consequences for violating this assumption:

1. Narrower confidence intervals for the regression parameters that leads to an increase in Type I error.

2. We don't exploit the spatial correlation when doing predictions.

The **effective sample size**, a measure of how much information is contained in our data, will be smaller than the actual sample size when looking at correlated data

Statistical workflow

1. Exploratory analysis.
2. Model formulation.
3. Model fitting.
4. Model validation. Evidence against the assumptions?
   - Yes: go back to point 2.
   - No: you can generate predictions and visualise uncertainty.

See also chapter 7 of Peter J. Diggle and Chetwynd (2011)

Prediction

- We may wish to predict $\boldsymbol{y}_{\text{new}}$ given some new X_{new}

- Or we may want to visualise model output

$$\mathbb{E}\left(\boldsymbol{y}_{\text{new}}\right) = X_{\text{new}}\widehat{\boldsymbol{\beta}}$$

where $\mathbb{E}(\cdot)$ is the Expectation operator, and $\widehat{\boldsymbol{\beta}}$ are the estimated coefficients

Prediction

- First, specify X_{new}

```
rats_predict <- data.frame(time = seq(0, 30))
```

- Then add the predictions from the model to the data frame

```
rats_predict <- mutate(
  rats_predict,
  weight = predict(object  = rats_lm,
                   newdata = rats_predict))
```

Complete the rats example in the practical exercise.

GLMs

- *Generalised* linear models extend linear model
- Used when we have data where normal errors don't make sense

Data	Values	Distribution
Counts	$\{0, 1, 2, \ldots\}$	Poisson
Trials	$\{0, 1, 2, \ldots, n\}$	Binomial
Proportions	$[0, 1]$	Beta
Strictly positive values	$(0, \infty)$	log-Normal
Strictly non-negative values	$[0, \infty)$	Gamma

GLMs

The general form of a GLM is

$$g\left(\mathbb{E}(\boldsymbol{y})\right) = X\boldsymbol{\beta}$$

Here, $g(\cdot)$ is the **link** between

- the expected outcome, $\mathbb{E}(\boldsymbol{y})$, and
- the **linear predictor**, $X\boldsymbol{\beta}$, a linear combination of the variables used to predict the outcome
- A GLM requires us to specify the *link function* that links the expected value of the outcome to the linear predictor

GLMs - Poisson

- The Poisson likelihood describes how some count varies with the x variables

$$\log(\boldsymbol{\lambda}) = X\boldsymbol{\beta}$$

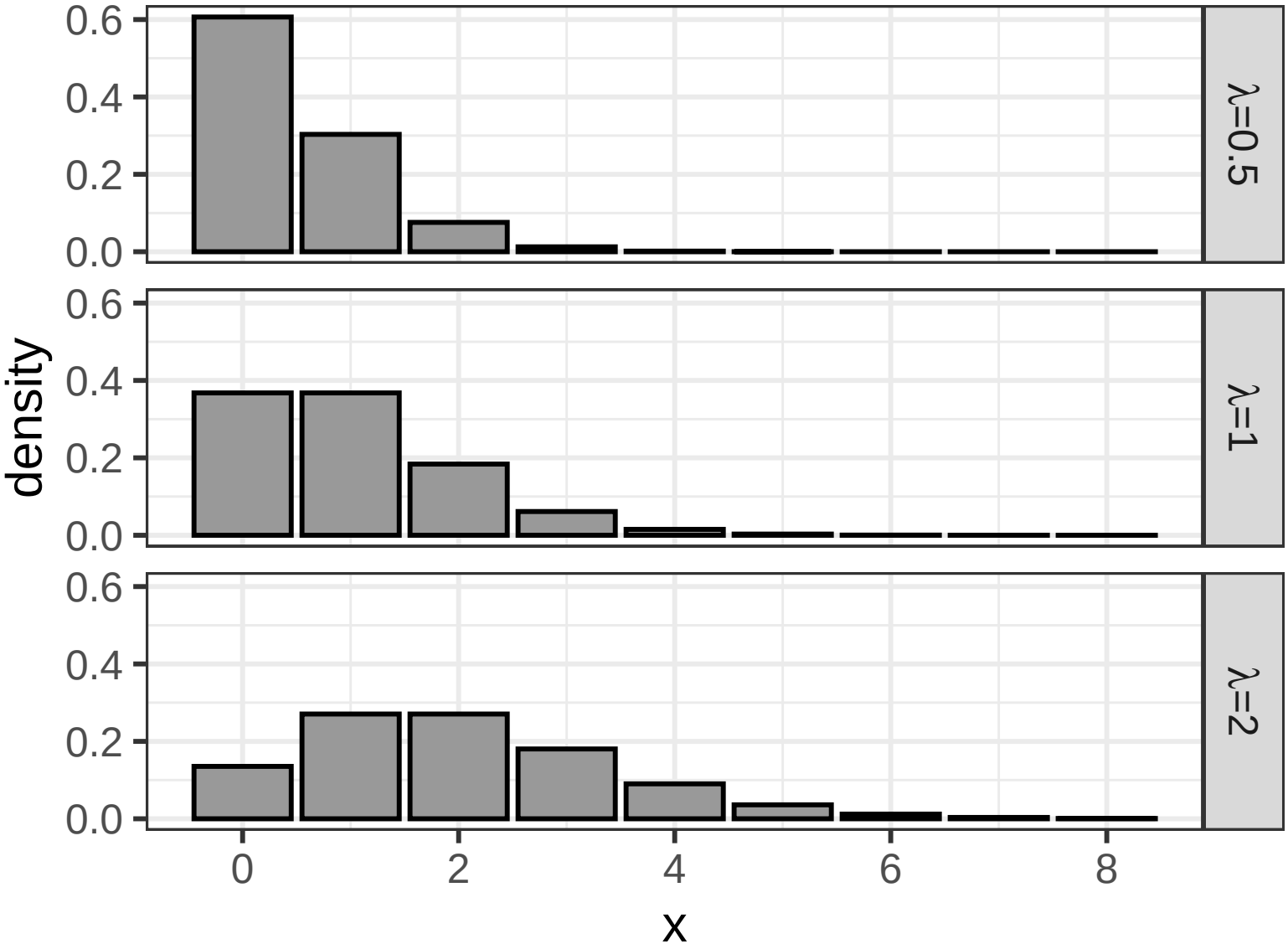
where λ_i is the **rate parameter**, the expected number of times an event occurs

- For those needing a refresher of the Poisson distribution,

$$P(y = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

describes the number of times that the outcome occurs

GLMs - Poisson



GLMs - Poisson

- If we want to parameterise the model as a rate *per unit*, we include an offset regarding of the known number of units
- e.g. counts, y_i, in regions with different expected numbers of cases (based on population), E_i (the offset)
- With a common per-capita rate, μ, we have $\mu = \lambda_i / E_i$, and

$$y_i \sim \text{Poisson}\left(\lambda_i\right) \leftrightarrow y_i \sim \text{Poisson}\left(E_i \mu\right)$$

- Alternatively, we can allow each region to have its own rate, μ_i, and have the rate vary according to predictor variables

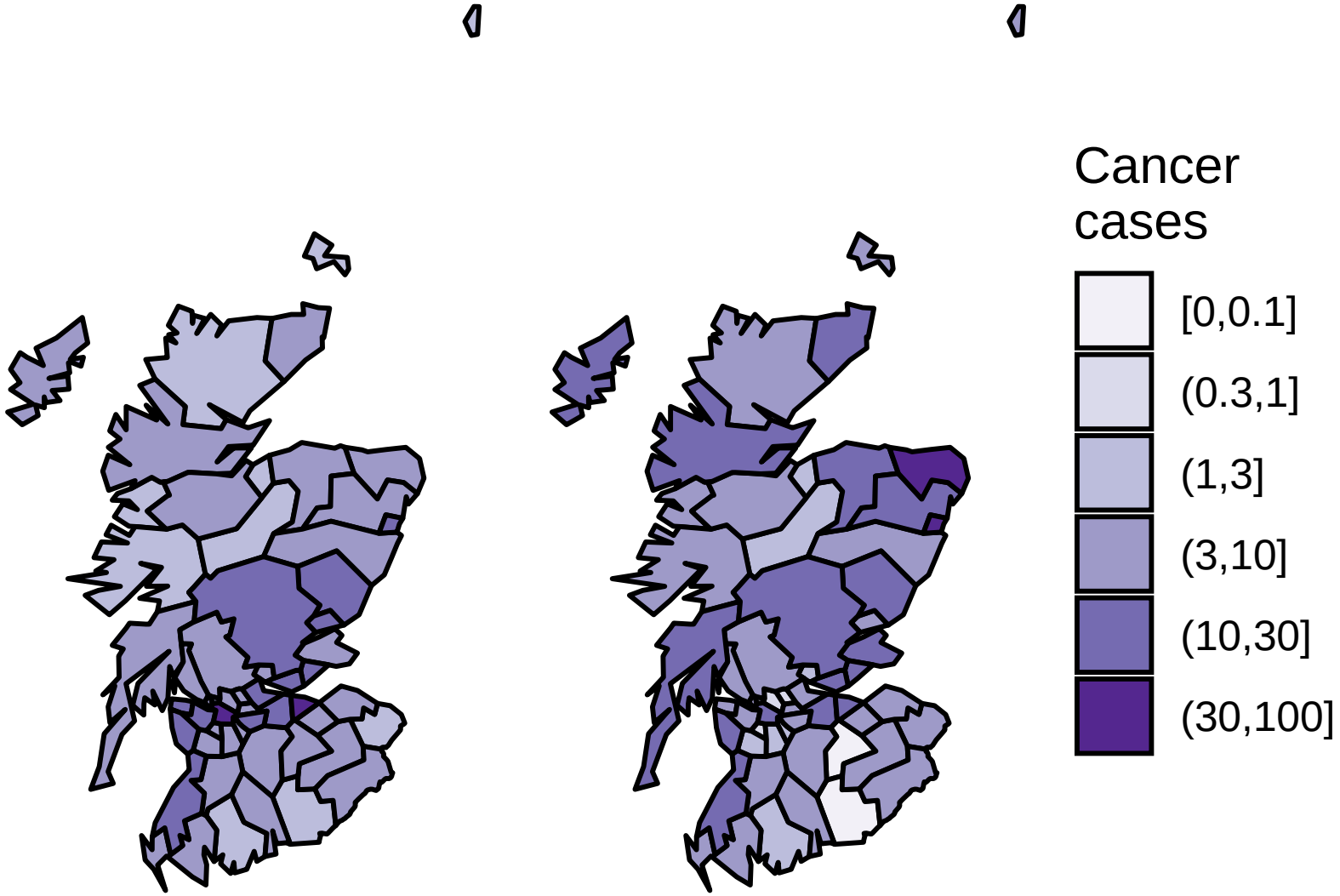
$$\log \lambda_i = \log E_i + X_{i*}\boldsymbol{\beta}$$

and we call $\log E_i$ the offset (and it is known)

GLMs - Poisson

- The offset is based on expected number of times something happens
- If rate across locations was uniform, how many in each location?
- Different applications require different consideration of expected number
    - differential risk by demographics
    - catchment area may be more relevant than population

GLMs - Scottish lip cancer



GLMs - Scottish lip cancer

- In the Scottish lip cancer data
  - y = `CANCER` is observed outcome
  - E = `CEXP` is expected cases, based on age-specific risk and district age structure
  - x = `AFF` is the proportion of the population working in agriculture, fishing and forestry

A non-spatial approach

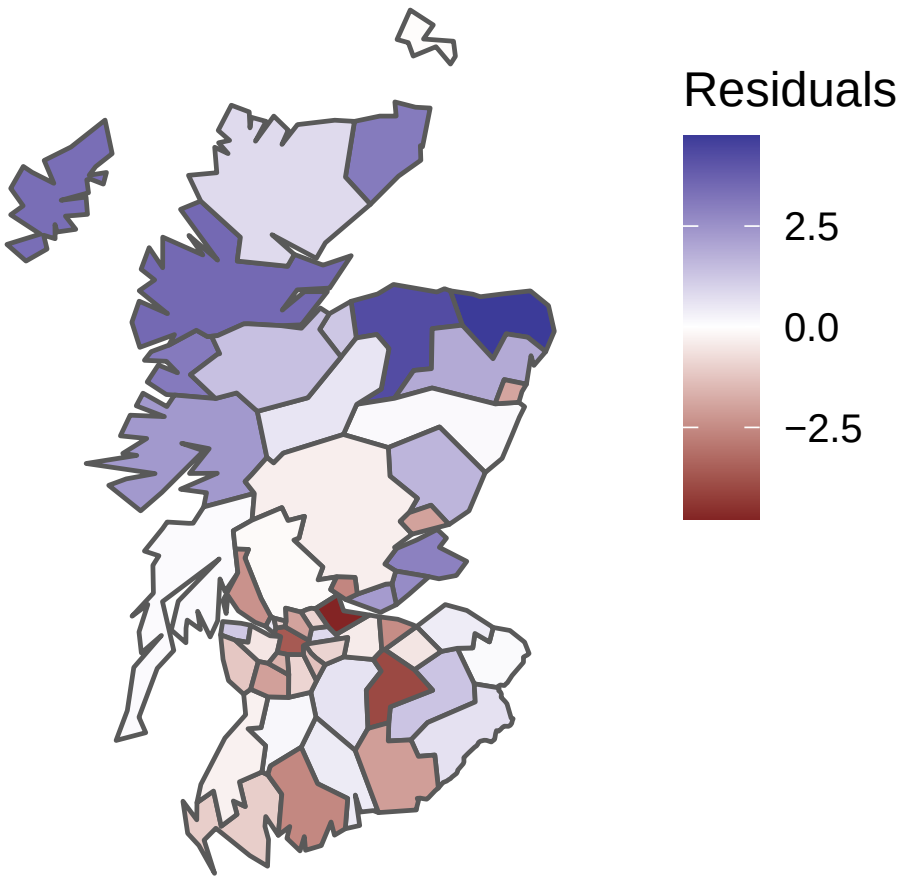
```
scotland_lip <- read_csv("../data/scotlip/scotlip.csv")

scotland_glm <- glm(data = scotland_lip,
                    CANCER ~ offset(log(CEXP)) + AFF,
                    family = poisson())
```

- Specifying with `offset()` informs `glm()` that we aren't using `log(CEXP)` as an explanatory variable

GLMs - Scottish lip cancer

- Residuals on link (log) scale
- Do these residuals look iid?
- Is this a random pattern?
- Are there any trends?



Binomial for proportion data

- Proportion data on the range $p_i \in [0, 1]$ typically from a trial
    - successes, y_i, of
    - attempts, n_i, therefore
    - failures, $n_i - y_i$
- The binomial model likelihood is given by

$$\mathcal{L}\left(\boldsymbol{y}|\boldsymbol{p}, \boldsymbol{n}\right) = \prod_{i=1}^{N} \binom{n_i}{y_i} p_i^{y_i} \left(1 - p_i\right)^{n_i - y_i}$$

- For regression, p_i varies as a function of other variables

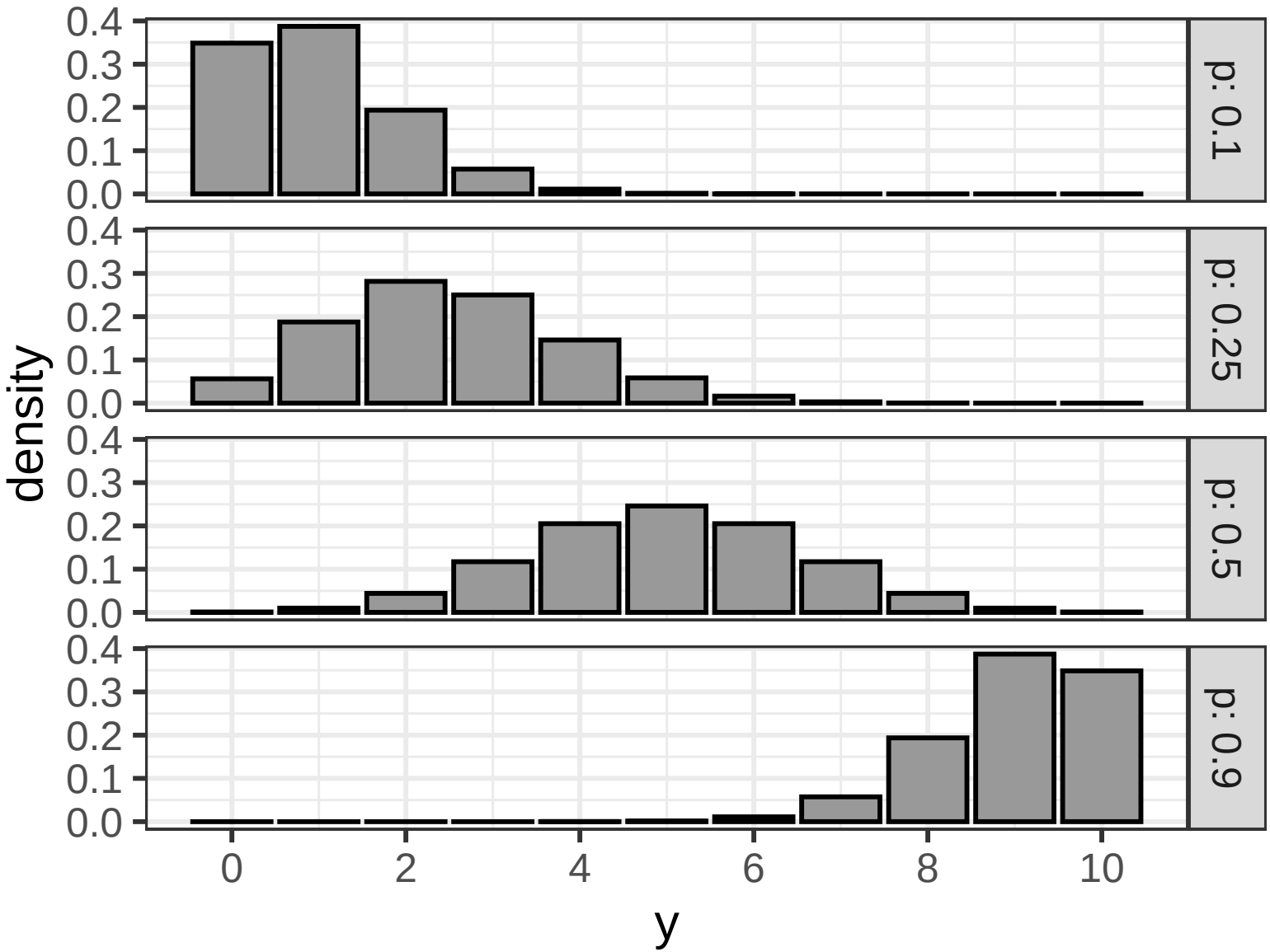
Binomial for proportion data

- Our link function, $g(\cdot)$, is the logit,

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

where $p_i/(1-p_i)$ is the odds of an event with probability p_i occurring

Binomial for proportion data



Binomial for proportion data

As $p \in [0, 1]$, we have $\text{logit}\, p \in (-\infty, \infty)$

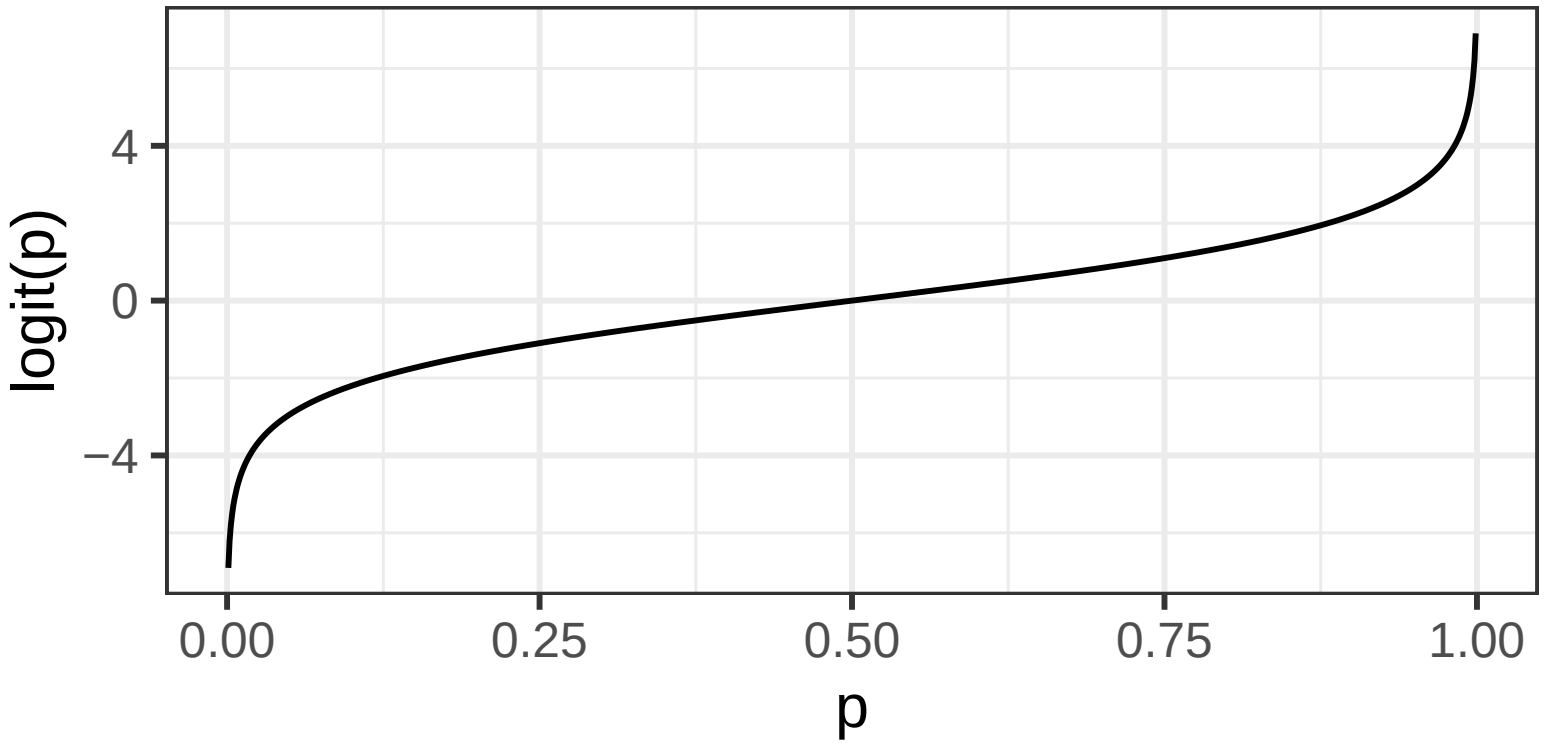
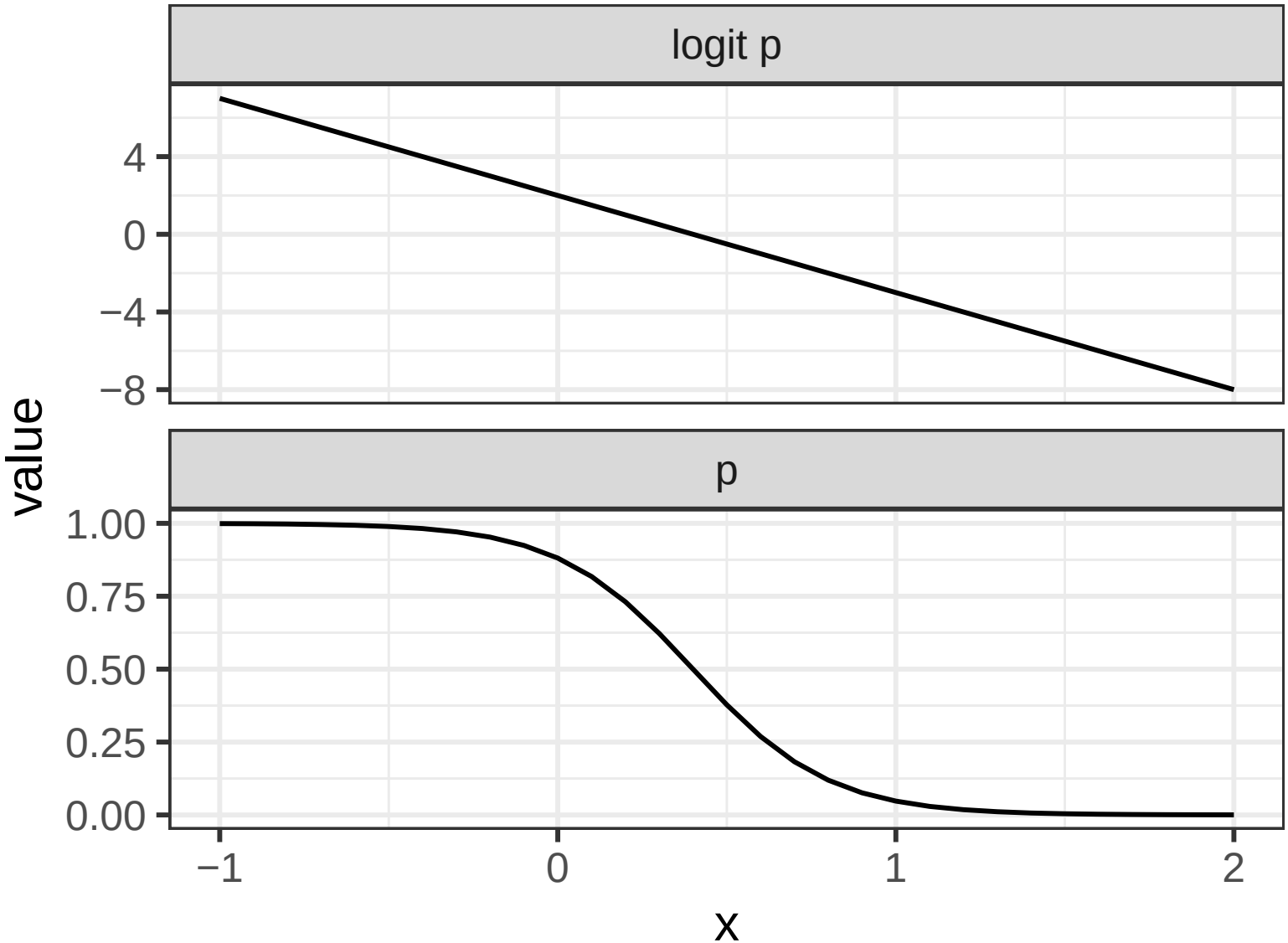


Figure 1: Logit link function showing how p and its logit are related

Binomial for proportion data

Consider $\text{logit}\, p = 2 - 5x$



Binomial for proportion data

- Our binomial GLM is then

$$y_i \sim \text{Binomial}\left(p_i, n_i\right)$$
$$\text{logit}\left(p_i\right) = X_{i*}\boldsymbol{\beta}$$

- Remember that $X_{i*}\boldsymbol{\beta}$ can take on any real value, and inverting the logit maps it to $(0, 1)$

- `logit()` and `inv.logit()` are in the boot package

Binomial for proportion data

```
## Simple feature collection with 3 features and 18 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: -72.46907 ymin: 18.37332 xmax: -72.25662 ymax: 1
## Geodetic CRS:  WGS 84
##   adm2code adm1code adm1_en        adm2_en adm0code adm0_en under_18
## 1   HT0111     HT01    West Port-au-Prince       HT   Haiti   399381
## 2   HT0112     HT01    West         Delmas       HT   Haiti   166714
## 3   HT0113     HT01    West      Carrefour       HT   Haiti   225495
##    urban  total vaccine_type dtp3coverage2016 vacc_num vacc_denom
## 1 977790 987310         DTP3             76.8    20820      27118
## 2 395260 395260         DTP3             84.3    11918      14138
## 3 501768 511345         DTP3             71.4    10028      14045
##     vacc_denom_type imputed  vacc_exp  urbanpct
## 1 Surviving Infants       0 18182.022 0.9903576 MULTIPOLYGON (((-72.3
## 2 Surviving Infants       1  9479.218 1.0000000 MULTIPOLYGON (((-72.2
## 3 Surviving Infants       0  9416.863 0.9812710 MULTIPOLYGON (((-72.4
```

Binomial for proportion data

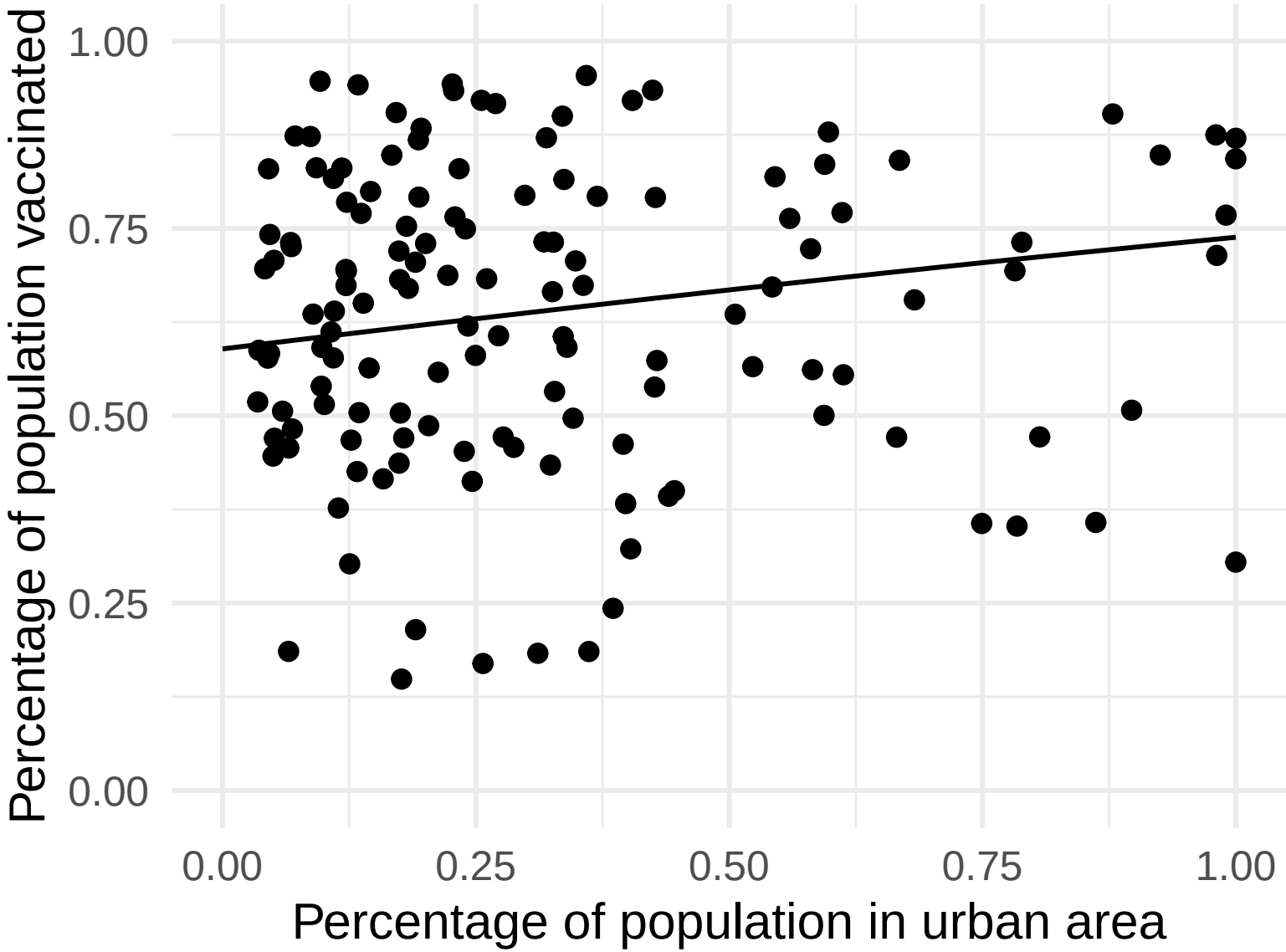
Our number of failures is $n_i - y_i$

```
adm2 <- mutate(adm2,
                vacc_fail = vacc_denom - vacc_num)
```

and the model call is

```
adm2_binom <-
  glm(data = adm2,
      cbind(vacc_num, vacc_fail) ~ urbanpct,
      family = binomial())
```

Binomial for proportion data



Binomial for proportion data

```
summary(adm2_binom)
```

```
##
## Call:
## glm(formula = cbind(vacc_num, vacc_fail) ~ urbanpct, family = binomia
##     data = adm2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -77.242   -8.782    1.858   12.611   35.249
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 0.360855   0.006778   53.24   <2e-16 ***
## urbanpct    0.675942   0.011145   60.65   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 48588  on 137  degrees of freedom
## Residual deviance: 44852  on 136  degrees of freedom
## AIC: 45875
##
## Number of Fisher Scoring iterations: 4
```

Binomial for proportion data

Complete the practical exercises on the Haiti vaccine coverage data

References I

Diggle, P. J., J. A. Tawn, and R. A. Moyeed. 1998. "Model-Based Geostatistics." *Journal of the Royal Statistical Society. Series C: Applied Statistics* 47 (3): 299–325.

Diggle, Peter J., and Amanda G. Chetwynd. 2011. *Statistics and Scientific Method*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199543182.001.0001.

Gelfand, Alan E, and Adrian FM Smith. 1990. "Sampling-Based Approaches to Calculating Marginal Densities." *Journal of the American Statistical Association* 85 (410): 398–409.

Tobler, W. R. 1970. "A Computer Movie Simulating Urban Growth in the Detroit Region." *Economic Geography* 46 (sup1): 234–40. https://doi.org/10.2307/143141.