Introduction to methods in environmental epidemiology 3

Robbie M Parks, PhD

21st July 2025

Email: robbie.parks@columbia.edu

BlueSky: @robbiemparks

Website: sparklabnyc.github.io

Art by Amy Wolfe

COLUMBIA | MAILMAN SCHOOL OF PUBLIC HEALTH

Outline from previous lecture

- Quantile regression
- An example

Outline

- Finding associations from data
- Model likelihood structures
  - Normal
  - Bernoulli
  - Binomial
  - Poisson
- Running models
- Evaluating model fit

Finding associations from data

- Regression: Used to assess presence of a relationship between a dependent (outcome) variable and one or more independent (predictor) variables.

Finding associations from data

- Basic steps for regression models:
  - Establish suitable model for observations.
  - Identify type of relationship between predictor(s) and outcome (linear, non-linear etc).
  - Run the model somehow (e.g., R with packages).
  - Evaluate model fit.

R for normal distribution model

- R code:

5: Fit Regression Model

```
mod <- lm(AveBMI ~ AvePM + PerBlack + PerLatinx + PerAsianAm + MedHInc +
            MedHVal + LTHS + FemaleUnemp + MaleUnemp + ClimateRegion,
          data = dta, na.action = na.omit)
```

R for normal distribution model

- R code:

6: Review Model Results

```
summary(mod)
```

```
Call:
lm(formula = AveBMI ~ AvePM + PerBlack + PerLatinx + PerAsianAm +
    MedHInc + MedHVal + LTHS + FemaleUnemp + MaleUnemp + ClimateRegion,
    data = dta, na.action = na.omit)

Residuals:
    Min      1Q  Median      3Q     Max
-4.9089 -0.6638 -0.1131  0.5264 12.2713

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                2.595e+01  7.663e-01  33.864  < 2e-16 ***
AvePM                      1.303e-01  4.992e-02   2.611  0.00936 **
```

Finding associations from data

- Basic steps for regression models:
  - **Establish suitable model for observations.**
  - **Identify type of relationship between predictor(s) and outcome (linear, non-linear etc).**
  - **Run the model somehow (e.g., R with packages).**
  - Evaluate model fit.

- What are suitable models for observations?
- Assume linear for now.
- How do we run these models?

Generalized linear models

- Extension of linear regression to when distribution not necessarily normally distributed.
- Distributions which are part of the exponential family can describe all sort of non-normally distributed variables.
- Many examples commonly encountered in environmental health.
- We'll go through a few now…

Exponential family of models

- Each distribution must be expressible as:

$$p(y|\theta) = b(y)\exp(\theta^T T(y) - a(\theta)) .$$

- E.g., Normal (Gaussian):

$$\begin{aligned} p(y|\mu) &= \frac{1}{\sqrt{2\pi}}\exp\left(\frac{(y-\mu)^2}{2}\right) \\ &= \frac{1}{\sqrt{2\pi}}\exp\left(\frac{y^2}{2}\right)\exp\left(\mu y - \frac{\mu^2}{2}\right) . \end{aligned}$$

- Natural parameter:

$$\theta = \mu.$$

Exponential family of models

- Each distribution must be expressible as:

$$p(y|\theta) = b(y)\exp(\theta^T T(y) - a(\theta)) .$$

- E.g., Bernoulli:

$$\begin{aligned} p(y|\mu) &= \mu^y(1-\mu)^{1-y} \\ &= \exp(y\log\mu + (1-y)\log(1-\mu)) \\ &= \exp\left(\log\left(\frac{\mu}{1-\mu}\right)y + \log(1-\mu)\right) . \end{aligned}$$

- Natural parameter: $\theta = \log\left(\frac{\mu}{1-\mu}\right).$

Some types of models to know about

- Sometimes dependent variable is not normally distributed.
- Exponential family is family of models described in a certain way (won't get into it in this workshop).
- All of the models we'll look at are exponential family models.
- Others too (exponential, log-normal, gamma, chi-squared, beta, Dirichlet, geometric).

Some types of models to know about

- Normal (we've seen previously).
- Bernoulli (yes/no).
- Binomial (set number of trials).
- Poisson (counts).
  - Conditional Poisson (special case of Poisson useful for some examples in environmental epidemiology).

Normal (Gaussian)

- For:
  - Continuous outcomes (e.g., lung function, birth weight)
  - Assumes symmetric, normally distributed errors

1. Linear Regression (Normal Likelihood)

```
model_normal <- lm(mortality ~ Temp + Ozone, data = data)
summary(model_normal)
```

Bernoulli

- For:
  - Logistic regression (yes/no).
  - When we want to classify observations into binary groups (yes/no).

3. Logistic Regression (Bernoulli Likelihood)

```
model_logistic <- glm(high_mort ~ Temp + Ozone, family = binomial(), data = data)
summary(model_logistic)
```

Binomial

- For:
  - When we want to know how many successes from a set number of trials
  - Proportion outcomes
  - Models grouped binary data

4. Binomial Regression (Grouped Binary Outcomes)

```
model_binomial <- glm(cbind(deaths_over5, total_deaths - deaths_over5) ~ Temp + Ozone,
                      family = binomial(),
                      data = data)
summary(model_binomial)
```

Poisson

- For:
  - With count data (number of events happening within a discrete space and time).

2. Poisson Regression (Count Data)

```
model_poisson <- glm(mortality ~ Temp + Ozone + day_of_week,
                     family = poisson(),
                     data = data)
summary(model_poisson)
```

Link function

- When modelling, we must make a decision about how the linear predictors are related to the key parameters in our chosen relationship.
- You will see several examples.
- This is linked to the natural parameter for each distribution in the exponential family form
- For example, link function on Poisson regression is given a log-link function to prevent counts from going negative.

$$\log(\mu) = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Finding associations from data

- Basic steps for regression models:
  - Establish suitable model for observations.
  - Identify type of relationship between predictor(s) and outcome.
  - Run the model somehow (e.g., R with packages).
  - **Evaluate model fit.**

- How to evaluate model fit?

Evaluating model fit

- Goodness-of-fit: AIC, BIC, deviance, residual plots.
- Predictive accuracy: cross-validation, ROC/AUC for binary outcomes.
- Diagnostics:
  - Check for overdispersion in Poisson models.
  - Assess multicollinearity.
  - Examine influential data points.

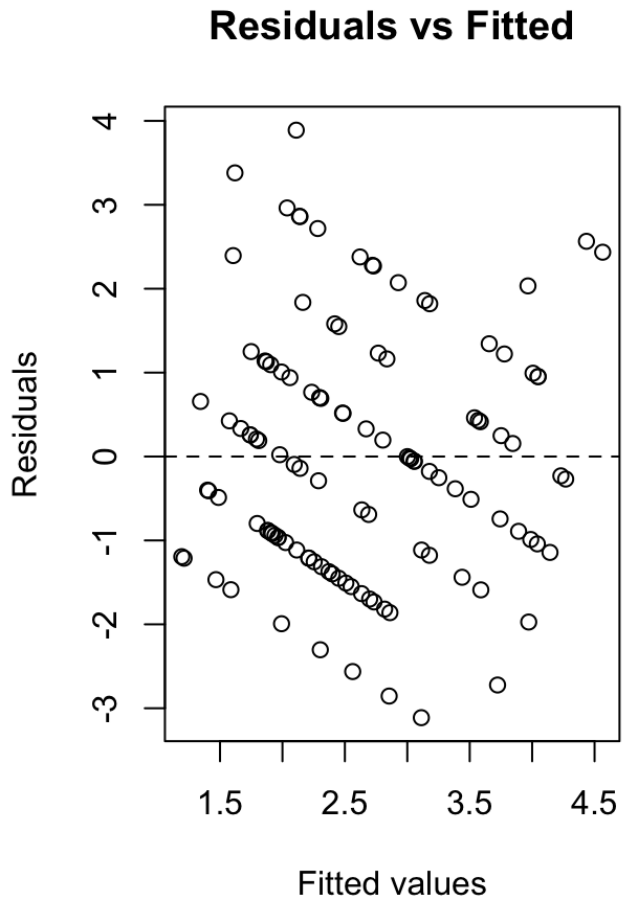
Evaluating model fit

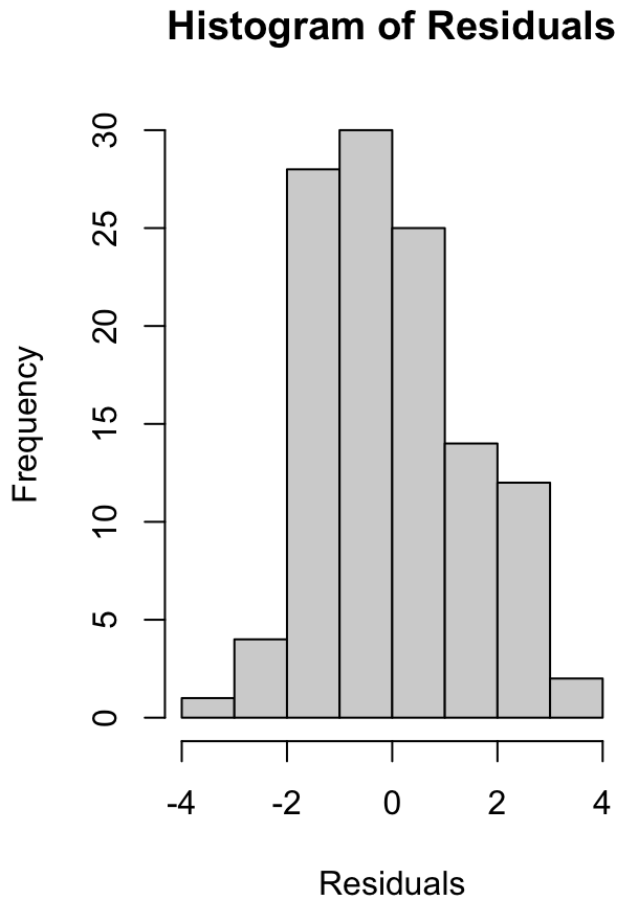
- R code:

```r
par(mfrow = c(1, 2))
plot(model_normal$fitted.values, resid(model_normal),
     main = "Residuals vs Fitted",
     xlab = "Fitted values", ylab = "Residuals")
abline(h = 0, lty = 2)
hist(resid(model_normal), main = "Histogram of Residuals", xlab = "Residuals")
```

Evaluating model fit

- R code:





Outline

- Finding associations from data
- Model likelihood structures
  - Normal
  - Bernoulli
  - Binomial
  - Poisson
- Running models
- Evaluating model fit

Getting ready for the lab

- This lab will involve taking some models and concepts from the **Introduction to methods in environmental epidemiology 3** lecture and introduce you to how to code them in R and assess their suitability.

Application

- How can you imagine applying this learning to your data and your research questions?

Questions

- Questions?

Introduction to methods in environmental epidemiology 3

Robbie M Parks, PhD

21st July 2025

Email: robbie.parks@columbia.edu

BlueSky: @robbiemparks

Website: sparklabnyc.github.io

Art by Amy Wolfe

COLUMBIA | MAILMAN SCHOOL OF PUBLIC HEALTH