Extending the linear model

DAAG Chapters 7 and 8

Learning objectives

The linear model framework can be extended in many ways. We will learn about

- Indicator variables for coding factors
- Fitting multiple lines
- Polynomial regression
- Splines

We will also learn about generalized linear models (glm)

- How the glm differs
- Logistic regression
- Ordinal regression
- Poisson regression

The linear model framework

The multiple linear regression model can be written

$$y = X\beta + \epsilon$$

where the distribution for the ϵ's is iid Normal.
Critically important is the design matrix X

- Including an intercept
- Coding factors (multiple intercepts)
- Coding interactions (multiple slopes)
- Polynomial regression
- Splines

Coding factors (separate intercepts)

- Factors are categorical variables that may or may not be ordered.
- In the design matrix, we code factors using 1's and 0's
- For example, if we have a factor for eye colour (blue, brown, other), and the data are:

blue, blue, brown, other, brown, other, blue, brown, blue

$$X = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \qquad X = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

Treatment contrasts   Sum (to zero) contrasts

Coding interactions (separate slopes)

- For a data set with:
  - Continuous response y
  - One three-level factor explanatory variable z
  - One continuous explanatory variable x

What models are available?

1. $y = \beta_0$ (constant)
2. $y = \beta_0 + \beta_1 x$ (single line)
3. $y = \beta_{01} + \beta_{02} z_2 + \beta_{03} z_3$ (three constants)
4. $y = \beta_{01} + \beta_{02} z_2 + \beta_{03} z_3 + \beta_1 x$ (three parallel lines)
5. $y = \beta_{01} + \beta_{02} z_2 + \beta_{03} z_3 + \beta_{11} x + \beta_{12} z_2 x + \beta_{13} z_3 x$ (three separate lines)
6. $y = \beta_0 + \beta_{11} x + \beta_{12} z_2 x + \beta_{13} z_3 x$ (three lines, one intercept)

Polynomial regression

- Polynomials provide a simple way to model curved relationships
- Sometimes there is a good theoretical reason to use a polynomial relationship
- Including higher order terms directly in the design matrix is one option
- Orthogonal polynomials are a good alternative because the correlation between model coefficients will be zero
  - this means greater numerical stability
  - lower-order coefficients won't change if higher-order coefficients are removed from the model
- In R, use `poly()` to specify orthogonal polynomials in a `formula` argument
- In SAS, use `ORPOL` function in `PROC IML` to generate design matrix columns

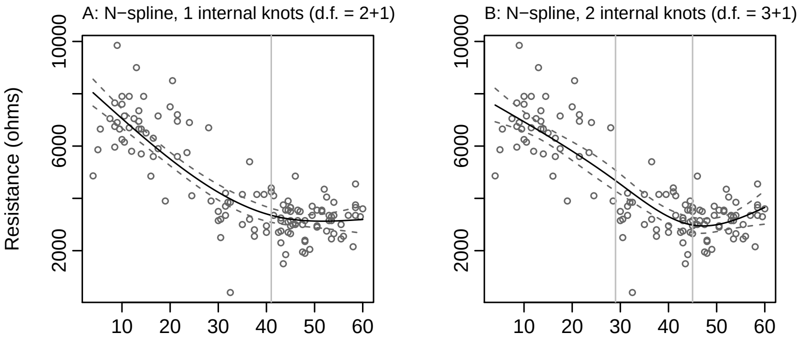
Splines

- Splines extend the idea of polynomial regression
- We do polynomial regression, but piecewise, joining the pieces at *knots*

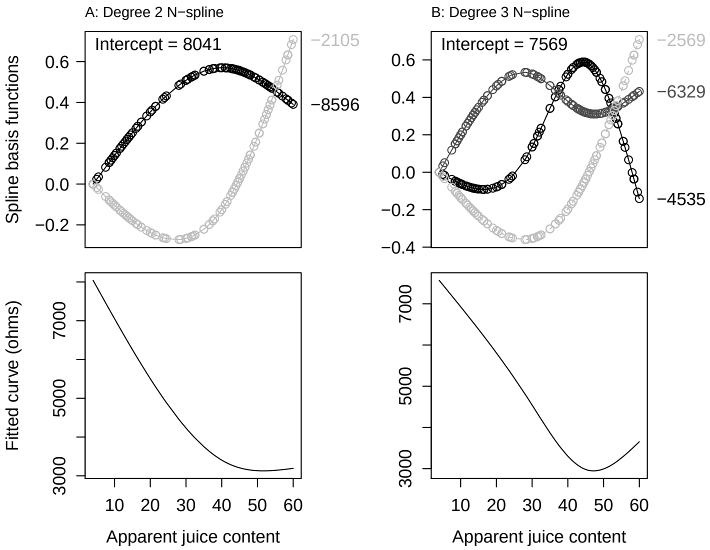
$$y = \beta_0 P_0(x) + \beta_1 P_1(x) + \ldots + \beta_k P_k(x)$$

- The $P_i(x)$ are *basis* functions. They are polynomial functions that are sometimes constrained to be non-zero for only certain values of x.
- Two possible choices for $P_i(x)$ are B-splines and natural splines (linear beyond the data).
- By adding an error term, these spline functions can be fit using the linear model framework
  - $P_i(x)$ is computed for all x in the data and all i
  - These $P_i(x)$ make up the design matrix in the linear model fit

Splines


A: N−spline, 1 internal knots (d.f. = 2+1)
B: N−spline, 2 internal knots (d.f. = 3+1)
Resistance (ohms)
2000
6000
10000
10
20
30
40
50
60

Splines



A: Degree 2 N−spline
B: Degree 3 N−spline
Intercept = 8041
Intercept = 7569
−2105
−8596
−2569
−6329
−4535
Spline basis functions
Fitted curve (ohms)
Apparent juice content
Apparent juice content

Generalized linear models

GLMs extend the linear modelling framework by allowing

- Non-Gaussian errors
- A *link function* that transforms the linear model response

The linear models we have considered so far had the form

$$y = X\beta + \epsilon,\ \epsilon \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

$$E[y] = X\beta$$

The generalized linear model is

$$f(E[y]) = X\beta$$

where $f()$ is the link function. Also

$$y = E[y] + \epsilon \qquad \text{or} \qquad y \sim (E[y], \theta)$$

but here ϵ can have a non-Gaussian distribution.

Logistic regression

In binomial logistic regression, the errors are binomial and the link function is logistic

$$f(E[y]) = \log\left(\frac{E[y]}{1 - E[y]}\right)$$

In this context, the $E[y] = p$, the binomial probability.
The model for $E[y]$ is

$$f(p) = \log\left(\frac{p}{1-p}\right) = X\beta$$

or

$$p = \frac{\exp(X\beta)}{1 + \exp(X\beta)}$$

and $y \sim \mathrm{Binom}(n, p)$, or $y \sim \mathrm{Bern}(p)$.

- Fit by maximizing likelihood of y as a function of β.
- Model comparison via deviance ($-2\log L(y|\hat{\beta})$).
- Confidence intervals for β using the likelihood.

Ordinal regression

- Ordinal response, link is usually logistic
- Here we look at the cumulative probabilities $\gamma_j = P(y \leq j)$

$$\log\left(\frac{\gamma_j}{1-\gamma_j}\right) = \eta_j - X\beta$$

- The η_j are *cutpoints* between the response categories

$$\eta_i < \eta_j \text{ for } i < j$$

- Assumption: β-effects are proportional to the odds for all j

$$\frac{\gamma_j}{1-\gamma_j} = \frac{\exp(\eta_j)}{\exp(X\beta)} \qquad \text{or} \qquad \frac{1-\gamma_j}{\gamma_j} = \exp(X\beta)\exp(-\eta_j)$$

  - Or, can include separate β_j for each j.

Poisson regression

- Errors are Poisson, link function most commonly log
- Recall that Poisson is for count data that arise from a Poisson process
- $E[y] = \lambda$, the rate parameter. The model is

  $$f(\lambda) = \log(\lambda) = X\beta$$

  or

  $$E[y] = \lambda = \exp(X\beta)$$

  and $y \sim \mathrm{Poisson}(\lambda)$.
- Note that the Poisson distribution has $\mathrm{Var}(y) = \lambda$. If we have over- or under- dispersion, we can relax this requirement and estimate a dispersion parameter ϕ *(quasipoisson)*.

Example: Head injuries

- Data: (simulated) patient data that present with head injuries
  - Q: Can we identify patients that would be classified as *high risk* using available criteria?
- Response: Whether a patient is classified as *high risk* by a clinician
- Explanatory variables:
  - Whether over age 65
  - Amount of amnesia before impact (threshold 30 mins)
  - Basal skull fracture present
  - Open skull fracture present
  - Whether vomiting
  - Whether loss of consciousness occurred
- Use logistic regression

Example: Head injuries

```
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)             -1.34880    0.05612 -24.036  < 2e-16 ***
age.65                   0.27891    0.12511   2.229  0.02579 *
amnesia.before           0.03770    0.10382   0.363  0.71652
basal.skull.fracture     0.31854    0.15474   2.059  0.03953 *
loss.of.consciousness    0.36088    0.12553   2.875  0.00404 **
open.skull.fracture      0.33752    0.20753   1.626  0.10387
vomiting                 0.76134    0.12595   6.045  1.5e-09 ***
---
(Dispersion parameter for binomial family taken to be 1)
Null deviance: 3460.4  on 3120  degrees of freedom
Residual deviance: 3401.3  on 3114  degrees of freedom
AIC: 3415.3
```

Example: Head injuries

The model is $\log\left(\frac{p}{1-p}\right) = X\beta$, so $p = \frac{\exp(X\beta)}{1+\exp(X\beta)}$.

- At the baseline, $X\beta = -1.349$ (the model intercept), or $\hat{p} = 0.206$
- What would get us to $p = 0.5$? We would need $\exp(X\beta) \geq 1$, or $X\beta \geq 0$
- If a patient is vomiting ($\hat{\beta} = 0.761$), then we also need at least two of
  - Whether over age 65 ($\hat{\beta} = 0.279$)
  - Basal skull fracture present ($\hat{\beta} = 0.319$)
  - Open skull fracture present ($\hat{\beta} = 0.338$)
  - Whether loss of consciousness occurred ($\hat{\beta} = 0.361$)
- If a patient is not vomiting, then even with all other conditions present, $\hat{p} \leq 0.5$
- Amount of amnesia before impact (threshold 30 mins) has little to no effect ($\hat{\beta} = 0.038$)