Model checking and comparison

Bert van der Veen

Department of Mathematical Sciences, NTNU

Typical steps in a real analysis

Before analysis

1. Determine research question
2. Collect data
3. *Plan analysis*

Analysis

1) Explore the data
2) Fit a model
3) Fit another model
4) Determine which model is better
5) Check model assumptions, if violated return to 3)
6) Report all results

What makes a good statistical model?

For prediction or for inference?

- A model is usually only good at one thing
  - Prediction
  - Inference
- But it can do both

What makes a good statistical model?

For prediction or for inference?

- A model is usually only good at one thing
  - Prediction
  - Inference
- But it can do both

This is a matter for **calibration**

What makes a good statistical model?

One that helps to answer your research question.

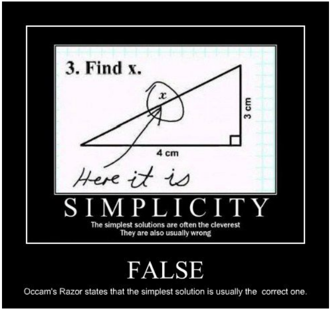
- Accurately represents the data generating process
- Not too difficult to interpret
- Is robust

Principle of Parsimony

The simplest explanation is often the correct one.

A simpler model might

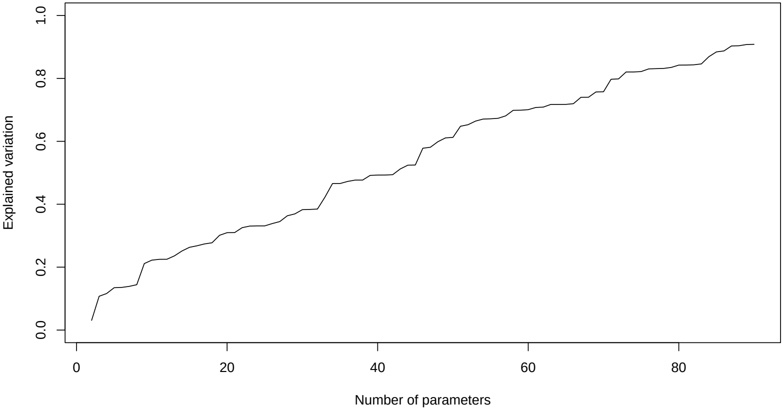
- Make fewer assumptions
- Be easier to interpret
- Be less prone to overfitting
- **Inadequatley accommodate properties of data**


3. Find x.
x
3 cm
4 cm
Here it is
SIMPLICITY
The simplest solutions are often the cleverest
They are also usually wrong
FALSE
Occam's Razor states that the simplest solution is usually the correct one.

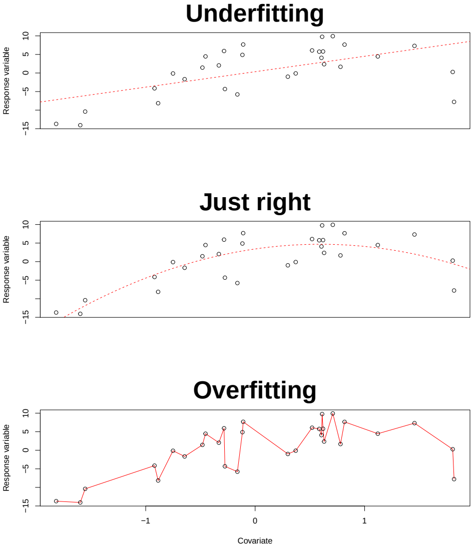
If two competing models fit the data equally well, we continue with the simpler model.

The problem of model complexity

A model with always fit better if you add a parameter

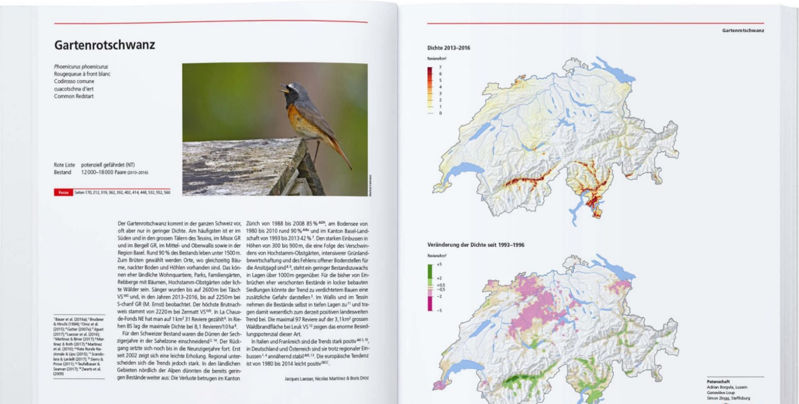


Model complexity



Example: Swiss birds

- Data by Schmid et al. (1998): the Swiss breeding bird atlas
- Occurrence of 56 species at 2524 locations recorded over a 4-year period



Gartenrotschwanz

Phoenicurus phoenicurus
Rougequeue à front blanc
Codirosso comune
cuacotschna d'iert
Common Redstart

Rote Liste	potenziell gefährdet (NT)
Bestand	12000–18000 Paare (2013–2016)

Focus Seiten 170, 212, 319, 362, 392, 402, 414, 448, 532, 552, 560

Der Gartenrotschwanz kommt in der ganzen Schweiz vor, oft aber nur in geringer Dichte. Am häufigsten ist er im Süden und in den grossen Tälern des Tessins, im Misox GR und im Bergell GR, im Mittel- und Oberwallis sowie in der Region Basel. Rund 90 % des Bestands leben unter 1500 m. Zum Brüten gewählt werden Orte, wo gleichzeitig Bäume, nackter Boden und Höhlen vorhanden sind. Das können eher ländliche Wohnquartiere, Parks, Familiengärten, Rebberge mit Bäumen, Hochstamm-Obstgärten oder lichte Wälder sein. Sänger wurden bis auf 2600 m bei Täsch VS[VdS] und, in den Jahren 2013–2016, bis auf 2250 m bei S-chanf GR (M. Ernst) beobachtet. Der höchste Brutnachweis stammt von 2220 m bei Zermatt VS[VdS]. In La Chaux-de-Fonds NE hat man auf 1 km² 31 Reviere gezählt[6]. In Riehen BS lag die maximale Dichte bei 8,1 Revieren/10 ha[8].

Für den Schweizer Bestand waren die Dürren der Sechzigerjahre in der Sahelzone einschneidend[2, 14]. Der Rückgang setzte sich noch bis in die Neunzigerjahre fort. Erst seit 2002 zeigt sich eine leichte Erholung. Regional unterscheiden sich die Trends jedoch stark. In den ländlichen Gebieten nördlich der Alpen dünnten die bereits geringen Bestände weiter aus: Die Verluste betrugen im Kanton Zürich von 1988 bis 2008 85 %[AtZH], am Bodensee von 1980 bis 2010 rund 90 %[AtBo] und im Kanton Basel-Landschaft von 1993 bis 2013 42 %[7]. Den starken Einbussen in Höhen von 300 bis 900 m, die eine Folge des Verschwindens von Hochstamm-Obstgärten, intensiverer Grünlandbewirtschaftung und des Fehlens offener Bodenstellen für die Ansitzjagd sind[8, 9], steht ein geringer Bestandszuwachs in Lagen über 1000 m gegenüber. Für die bisher von Einbrüchen eher verschonten Bestände in locker bebauten Siedlungen könnte der Trend zu verdichtetem Bauen eine zusätzliche Gefahr darstellen[3]. Im Wallis und im Tessin nehmen die Bestände selbst in tiefen Lagen zu[11] und tragen damit wesentlich zum derzeit positiven landesweiten Trend bei. Die maximal 97 Reviere auf der 3,1 km² grossen Waldbrandfläche bei Leuk VS[12] zeigen das enorme Besiedlungspotenzial dieser Art.

In Italien und Frankreich sind die Trends stark positiv[AtF, 5, 10], in Deutschland und Österreich sind sie trotz regionaler Einbussen[1, 4] annähernd stabil[AtD, 13]. Die europäische Tendenz ist von 1980 bis 2014 leicht positiv[EBCC].

Jacques Laesser, Nicolas Martinez & Boris Droz

[1] Bauer et al. (2016a); [2] Bruderer & Hirschi (1984); [3] Droz et al. (2015); [4] Gatter (2007a); [5] Jiguet (2017); [6] Laesser et al. (2016); [7] Martinez & Birrer (2017); [8] Martinez & Roth (2017); [9] Martinez et al. (2010); [10] Rete Rurale Nazionale & Lipu (2015); [11] Scandolara & Lardelli (2017); [12] Sierro & Posse (2011); [13] Teufelbauer & Seaman (2017); [14] Zwarts et al. (2009)

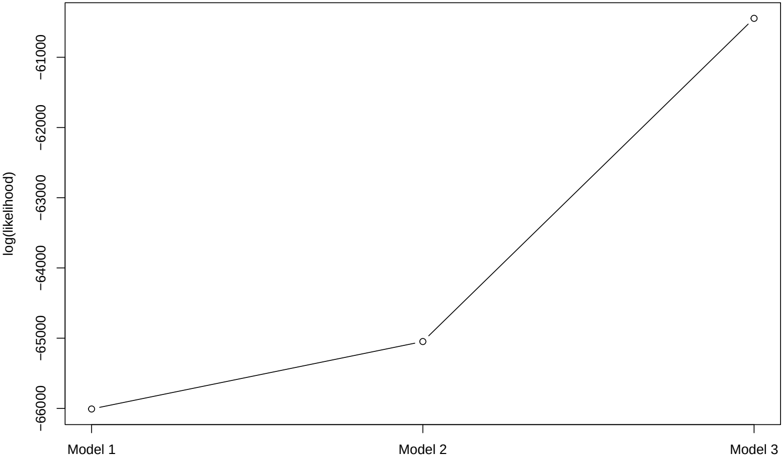
Gartenrotschwanz

Patenschaft
Adrian Borgula, Luzern
Geneviève Loup
Simon Zingg, Steffisburg

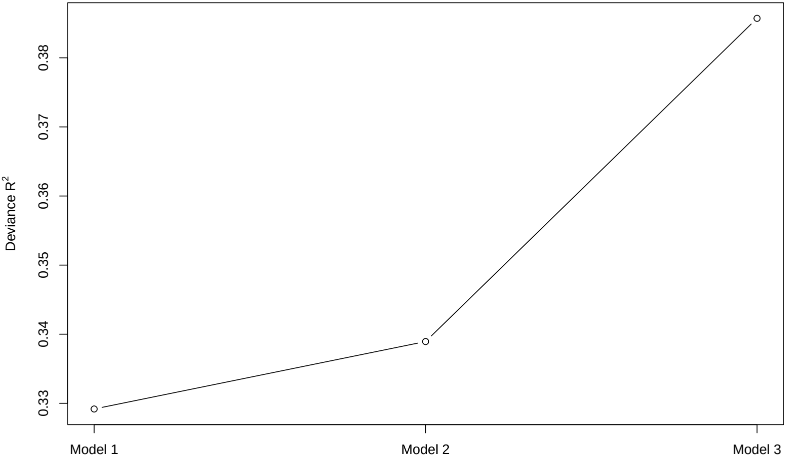
Swiss birds: fit some models

```
model1 <- glm(occ~0+Species + slp:Species, data = datalong, family = "binomial")
model2 <- glm(occ~0+Species + (mind + slp):Species, data = datalong, family = "binomial")
model3 <- glm(occ~0+Species + (mind + slp + ddeg0):Species, data = datalong, family = "binomial")
```

Swiss birds: log-likelihoods



Baseball attendance: explained variance



Confirmatory vs. Exploratory analysis

Confirmatory: test hypothesis
Exploratory: find a good model

Usually indicated by the research question.

Hypothesis testing

1) Define null hypothesis
2) Define alternative hypothesis
3) Choose a test statistic (e.g., deviance in GLMs)
4) Calculate its distribution under the null hypothesis
5) Check if the observed test statistic falls within the null distribution
6) Accept/Reject

Likelihood ratio test

Is improved fit due to noise or is the alternative model actually better?

Procedure

- Fit two models: M_0 with k parameters and M_1 with r
- Calculate likelihood ratio $\Lambda = \log\left(\frac{\mathcal{L}(\mathbf{y};\Theta_0)_{M_0}}{\mathcal{L}(\mathbf{y};\Theta_1)_{M_1}}\right)$
- $\mathcal{L}(\mathbf{y};\Theta_0)_{M_0} \leq \mathcal{L}(\mathbf{y};\Theta_1)_{M_1}$
- $-2\Lambda \sim \chi^2(k_1 - k_0)$ under the null
- $p \geq 0.05$ difference in likelihood is due to sampling

LRT approximation assumptions

- $n \to \infty$
- Θ_0 contained in Θ_1: nested models
- The true parameter is in the interior of the parameter space
- Model is "identifiable"
- Hessian matrix is sufficiently close to the Fisher information
- y_i are independent

These assumptions may fail, especially in models more complex than GLMs

Alternatively: LRT by simulation.

LRT and deviance

$$\Lambda = \frac{D(\mathbf{y}; \hat{\boldsymbol{\mu}}_{M0}) - D(\mathbf{y}; \boldsymbol{\mu}\hat{}_{M1})}{\phi} \tag{1}$$

Example: Lizard habitat preference

- Data originally by Schoener (1970)
- Counts of two species of lizard in Jamaica

Figure 2: wikipedia.org: grahami

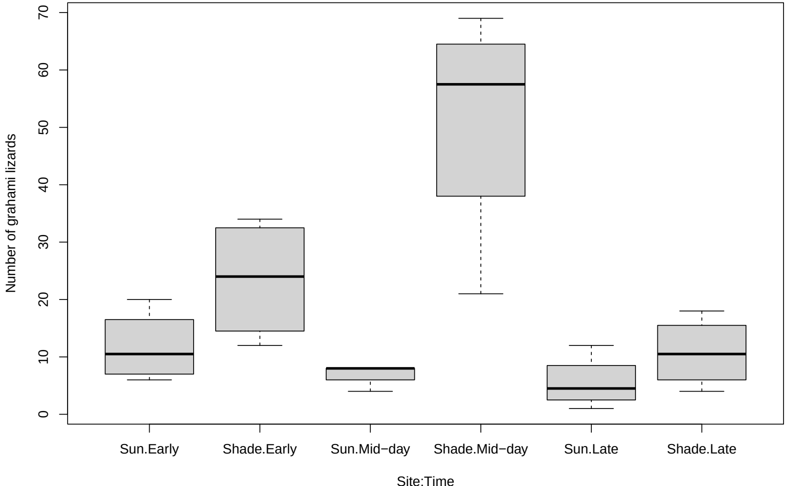
Figure 3: wikipedia.org: opalinus

Lizards: the data

Dataset also covered in McCullagh and Nelder (1989)

	Site	Diameter	Height	Time	grahami	opalinus
1	Sun	D <= 2	H < 5	Early	20	2
2	Sun	D <= 2	H < 5	Mid-day	8	1
3	Sun	D <= 2	H < 5	Late	4	4
4	Sun	D <= 2	H >= 5	Early	13	0
5	Sun	D <= 2	H >= 5	Mid-day	8	0
6	Sun	D <= 2	H >= 5	Late	12	0
7	Sun	D > 2	H < 5	Early	8	3
8	Sun	D > 2	H < 5	Mid-day	4	1
9	Sun	D > 2	H < 5	Late	5	3
10	Sun	D > 2	H >= 5	Early	6	0
12	Sun	D > 2	H >= 5	Late	1	1
13	Shade	D <= 2	H < 5	Early	34	11
14	Shade	D <= 2	H < 5	Mid-day	69	20
15	Shade	D <= 2	H < 5	Late	18	10
16	Shade	D <= 2	H >= 5	Early	31	5
17	Shade	D <= 2	H >= 5	Mid-day	55	4
18	Shade	D <= 2	H >= 5	Late	13	3
19	Shade	D > 2	H < 5	Early	17	15
20	Shade	D > 2	H < 5	Mid-day	60	32
21	Shade	D > 2	H < 5	Late	8	8

Example: Lizards



Lizards: interaction

Null hypothesis: There is no interaction of Time and Site

Fit M_0:

```
nmodel <- glm(cbind(grahami, opalinus)~Time+Site,
              data = lizards, family="binomial")
```

Alternative hypothesis: There is an interaction of Time and Site

Fit M_1:

```
amodel <- update(nmodel, formula = . ~ Time*Site)
```

Lizards: interaction

```
##
## Call:
## glm(formula = cbind(grahami, opalinus) ~ Time + Site + Time:Site,
##     family = "binomial", data = lizards)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)            2.24071    0.47037   4.764  1.9e-06 ***
## TimeMid-day            0.06188    0.87818   0.070   0.9438
## TimeLate              -1.22911    0.62585  -1.964   0.0495 *
## SiteShade             -1.16315    0.51296  -2.268   0.0234 *
## TimeMid-day:SiteShade  0.07270    0.91343   0.080   0.9366
## TimeLate:SiteShade     0.69387    0.70486   0.984   0.3249
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

Lizards: LRT

```
(Lambda <- 2*(logLik(amodel)-logLik(nmodel)))
```

```
## 'log Lik.' 1.098943 (df=6)
```

```
k <- attr(logLik(amodel),"df")-attr(logLik(nmodel),"df")
pchisq(Lambda, k,lower.tail=FALSE)
```

```
## 'log Lik.' 0.5772548 (df=6)
```

We reject the alternative hypothesis.

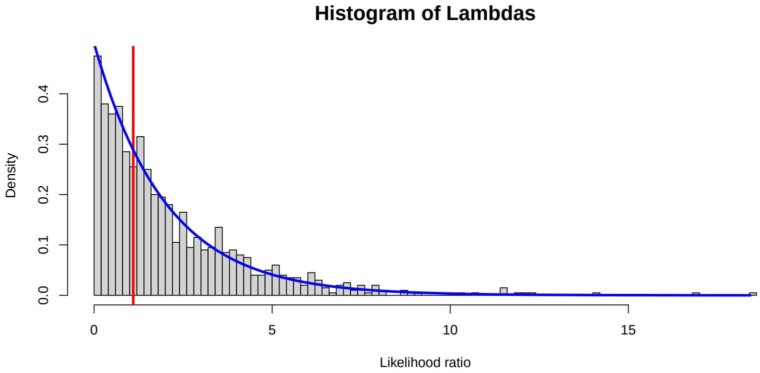
Lizards: LRT simulation

```r
Lambdas <- NULL
for(i in 1:1000){
ynew <- as.matrix(stats::simulate(nmodel))
nmodel2 <- glm(ynew~Time+Site,
             data = lizards, family="binomial")
amodel2 <- update(nmodel2, formula = .~Time*Site)
# Store test statistic
Lambdas <- c(Lambdas,
             2*(logLik(amodel2)-logLik(nmodel2)))
}
# if <0.05 our test statistic in the tail.
sum(Lambdas>Lambda)/1000
```

```
## [1] 0.598
```

Lizards: LRT simulation

Red: Observed statistic
Blue: $\chi^2(2)$



Information criteria

A different paradigm:
Find the best model amongst a set of models.

Best:

- Penalise complexity (number of parameters)
- By fit (likelihood)

Most commonly:

1) AIC: Akaike's Information Criterion (Akaike 1974)
2) BIC: Bayesian Information Criterion (Schwarz 1978)

Lower = better

Akaike's Information Criterion

$$\text{AIC} = -2\mathcal{L}(\mathbf{y};\Theta) + 2k \tag{2}$$

- Penalizes model complexity
- (approximately) Measures information loss to the true data generating process
- Asymptotically

AIC tends to select too complex models with little data. Finite sample correction (Sugiura 1978):

$$\text{AIC}_\text{c} = \text{AIC} + \frac{2k(k+1)}{n-k-1} \tag{3}$$

Find the model that predicts best

Bayesian Information Criterion

$$\text{BIC} = -2\mathcal{L}(\mathbf{y};\Theta) + k\log(n) \tag{4}$$

So the penalty is different.

Find the model closest to the ”true” model

Connection of AIC and LRT

Rule of thumb: difference of 2 points means a model is better

$$\begin{aligned}\Delta\text{AIC} =&\text{AIC}_{M_1} - \text{AIC}_{M_0} \\ &= 2\mathcal{L}(\mathbf{y};\Theta_0) - 2\mathcal{L}(\mathbf{y};\Theta_1) + 2k_1 - 2k_0 \\ &= -2\Lambda + 2(k_1 - k_0)\end{aligned} \tag{5}$$

So AIC with a rule of = 2 can be seen as a more liberal LRT (Sutherland et al. 2023)

The cult of (A)IC

Presentation by Mark Brewer

"Always use (A)IC for model comparison"

My perspectives

- Use common sense
- Do not blindly test all models ("dredging")
- Use model comparison techniques in moderation

Don't take the "best" model paradigm too seriously

Freedman's paradox

Just by chance, predictors with no relationship to the response will be selected.

Lizards: interaction selection

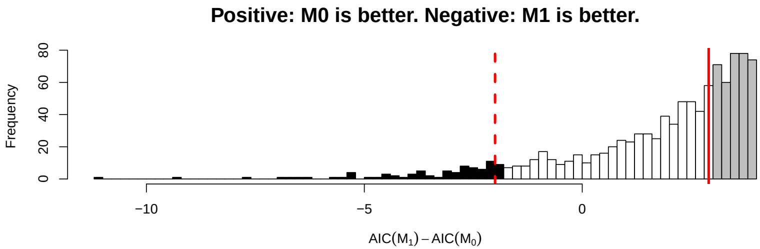
```
##        df      AIC
## nmodel  4 119.7237
## amodel  6 122.6248
```

```
##        df     AICc
## nmodel  4 121.9460
## amodel  6 127.8748
```

```
##        df      BIC
## nmodel  4 124.2657
## amodel  6 129.4378
```

ΔAIC by simulation under the simpler model

Solid: observed ΔAIC
Dashed: -2 ΔAIC



Black: We would incorrectly select the more complex model (7%)
White: We would consider the models equivalent (34%)
Gray: Under a 2 ΔAIC rule we would select the more simpler model (59%)

ΔAIC by simulation under the additive model

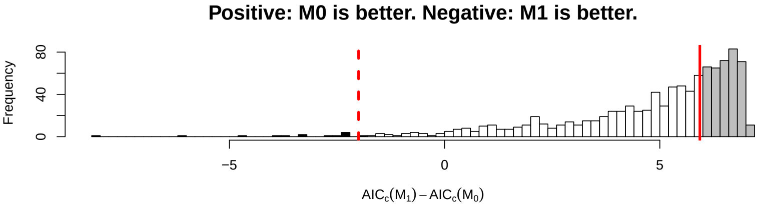
```
## Models with a more extreme test statistic: 0.607
```

```
## Models with more than 2 AIC difference: 0.591
```

93% of the time we would have selected the simpler model

ΔAICc by simulation under the simpler model

Solid: observed ΔAICc
Dashed: -2 ΔAICc



Black: We would incorrectly select the more complex model (1%)
White: We would consider the models equivalent (40%)
Gray: Under a 2 ΔAIC rule we would select the more simpler model (89%)

ΔAICc by simulation under the simpler model

```
## Models with more than 2 AIC difference: 0.591
```

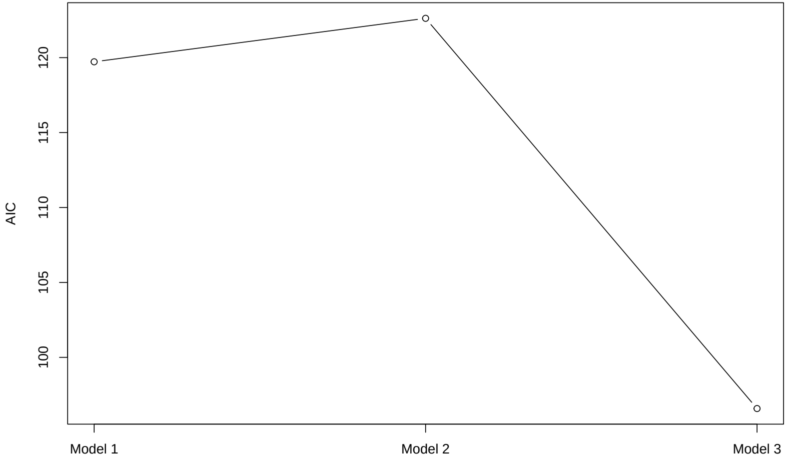
```
## Models with more than 2 AICc difference: 0.885
```

99% of the time we would select the simpler model.

Lizards: interaction and height

```
model1 <- glm(cbind(grahami, opalinus)~Time+Site,
              data = lizards, family="binomial")
model2 <- update(model1, formula = . ~ Time*Site)
model3 <- update(model2, formula = ~.+Height)
```

Lizards: AIC plotted



Omitted variable bias

Occurs if we omit a variable, i.e., we have the model:

$$\mathrm{g}\{\mathbb{E}(y_i|x_{i1}, x_{i2})\} = \alpha + x_{i1}\beta_1 + x_{i2}\beta_2 \qquad (6)$$

but we fit without x_{i2}. The consequence is especially clear for linear models:

$$y_i = \alpha + x_{i1}\beta_1 + \epsilon_i, \qquad \text{with } \epsilon_i = x_{i2}\beta_2 + \epsilon_i^\star \qquad (7)$$

- ϵ_i may be correlated with x_{i1}
- Residual variance changes with x_{i1}
- Causes bias in parameter estimates and incorrect standard errors

Multimodel inference

What to do when you have multiple models that are equally good?

- The simpler model
- Do not "model average" for inference
- Report multiple models

GLM assumptions

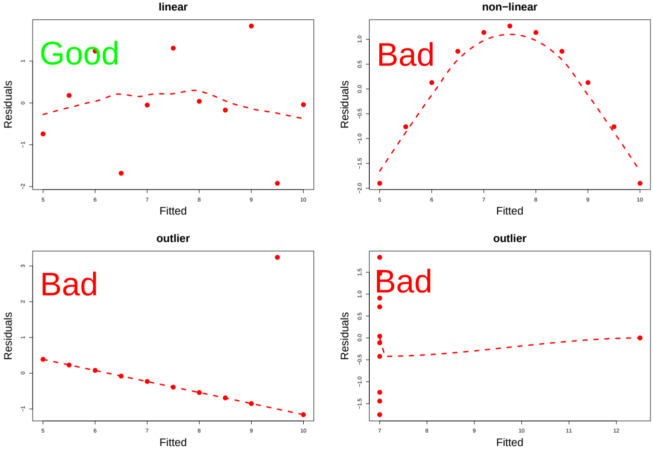
- Lack of outliers
- Independence of observations
- Correct distribution
- Correct link function
- Correct variance function
- No systematic departure from the model

GLM assumptions

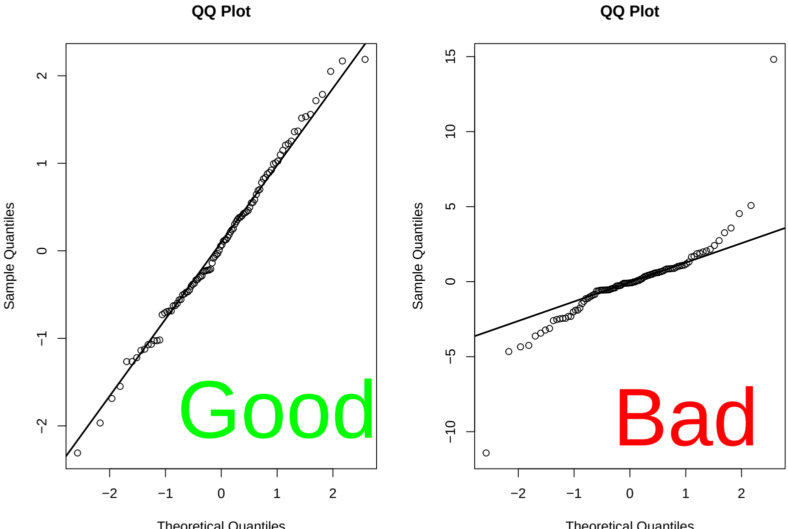
- Lack of outliers
- Independence of observations
- Correct distribution
- Correct link function
- Correct variance function
- No systematic departure from the model

We use plots to check if any assumptions are violated.

Residual plots: residuals vs. fitted



Residual plots: QQ plot



Residuals

You can usually construct residual plots with `plot` or `DHARMa`.

Summary

- Hypothesis testing and information criteria: two different paradigms
- Do not mix them
- Do not fall for “the best model” trap
- The full model is often a good model
- Keep things simple
- See also Murtaugh (2014) and Burnham and Anderson, and a lot of others
- Check your assumptions