

Mixed Models

Zack Treisman

Spring 2021

Philosophy

The tools we will investigate this week address several related issues.

- ▶ Heterogeneity of variance
 - ▶ Sometimes this is definitely present, even after building an otherwise good model.
- ▶ Nuisance variables
 - ▶ Often, we have variables that we know we need to measure and account for, but they are not of primary interest.
- ▶ Dependent observations
 - ▶ The independence assumption implicit in all of our models thus far is often false.

We do have tools that deal with these issues. If your data are especially nice and your analysis is fairly straightforward, so are the tools for building appropriate models. If you are not so lucky, then things get complicated.

Heterogeneity of variance

Why is the variance heterogenous?

- ▶ Because of a variable. Hopefully one in your data.

It's possible that the heterogeneity is from not looking at the variables in your model at appropriate scales.

- ▶ Were the data collected on the same scale that you are using, or have you transformed any variables? Recall that inverting a log-transform changes additive error into multiplicative error. Any nonlinear variable transform does this.

Nuisance variables

There is no set definition of nuisance variable. These are any variables that you know are having an effect on your response but are not what you are trying to investigate.

- ▶ Be sure you think about why you aren't interested.

Accounting for these variables is essential for establishing appropriate baselines. Sometimes the signal that you are trying to measure is small compared to environmental variation.

- ▶ Effect of compost treatment on soil moisture retention is going to be smaller than rainfall, soil characteristics, etc. (Cooper)

Dependent observations

Measurements on the same individual are not independent.

- ▶ Individual what? (Organism, forest, drainage, ...)

Even without discrete individuals, nearby observations can still be correlated.

- ▶ Variograms show relationship between distance between observations and covariance.

Having dependent observations means that your data set is effectively smaller.

Tools

- ▶ **Mixed effects models** (LMM, GLMM, GAMM)
 - ▶ Fixed effect variables modify the signal, random effect variables modify the noise.
 - ▶ Many packages for this, no clear winner. (lme4 has lmer, glmer and nlmer; nlme has lme; MASS has glmmPQL; glmmTMB has glmmTMB. Bayesian options include brms::brm and MCMCglmm::MCMCglmm.)
 - ▶ Hopefully this will be easier in another decade or so.
- ▶ Also see **Generalized least squares** (GLS) and **Generalized estimating equations** (GEE).

Many of the tools for mixed modeling also allow the specification of structure to correlation (corStruct) and variance (varStruct).

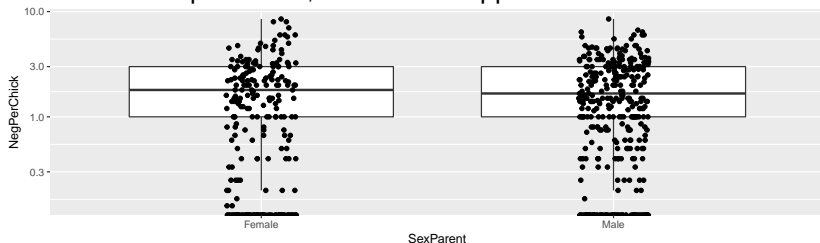
Owls

Roulin and Bersier (2007) (a running example in Zuur et al. (2009)) investigates begging behavior in nestling barn owls. They ask if vocal intensity of owl chick begging differs by parent.

```
head(Owls,3) # Data are in glmmTMB
```

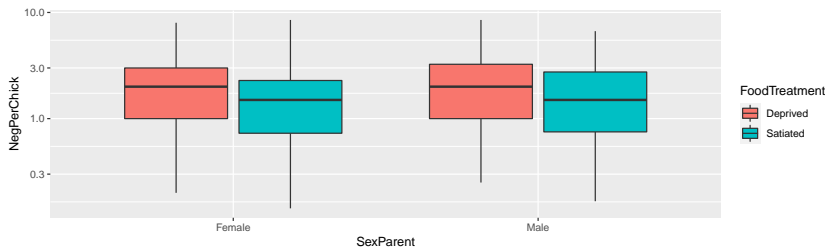
```
##      Nest FoodTreatment SexParent ArrivalTime SiblingNegotiation BroodSize
## 1 AutavauxTV      Deprived      Male      22.25                4          5
## 2 AutavauxTV      Satiated      Male      22.38                0          5
## 3 AutavauxTV      Deprived      Male      22.53                2          5
##      NegPerChick logBroodSize
## 1          0.8      1.609438
## 2          0.0      1.609438
## 3          0.4      1.609438
```

Without other predictors, it does not appear so.



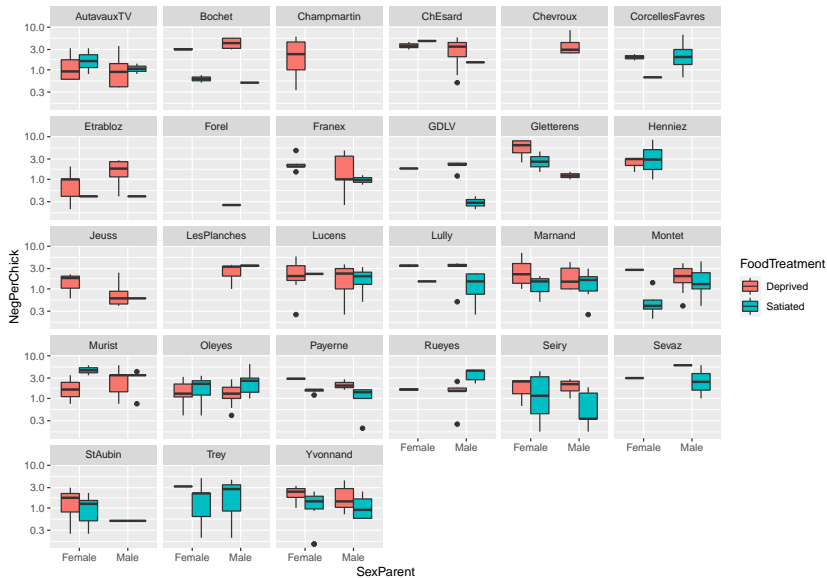
Food Treatment

Begging behavior is likely affected by the hunger levels in the nest. Impose this as an experimental treatment to control for unknown variation. Feed some nests extra, deprive others, then swap treatments the following night.



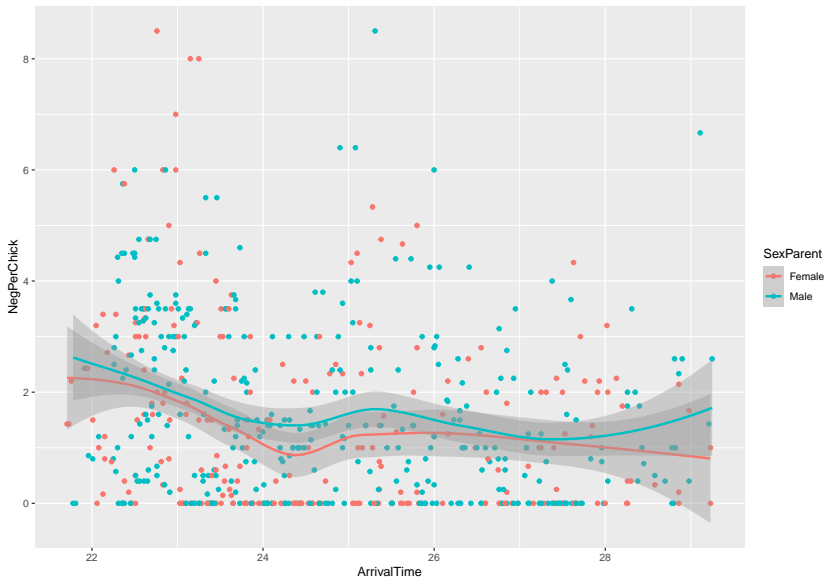
Still no visible difference between parents.

Nest is a nuisance variable.



Arrival time affects the variance.

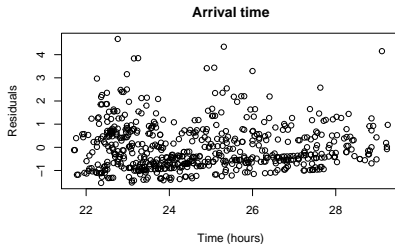
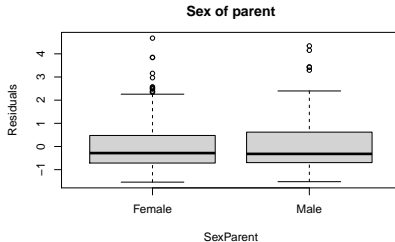
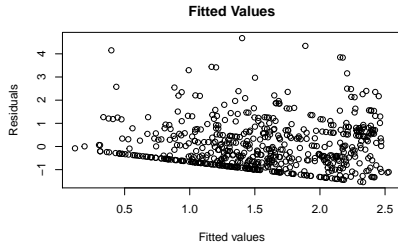
It could also be correlated with the sex of the parent.



Linear model

First thing to try is basic linear regression (ANCOVA).

```
M.lm <- lm(NegPerChick~SexParent*(FoodTreatment+ArrivalTime),data=Owls)
```



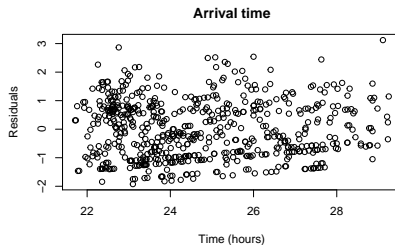
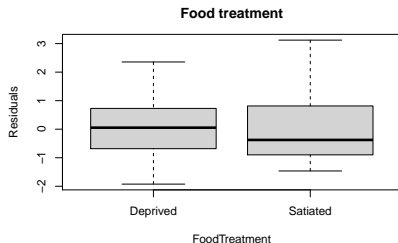
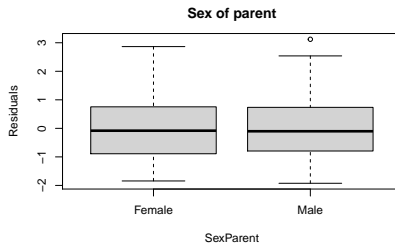
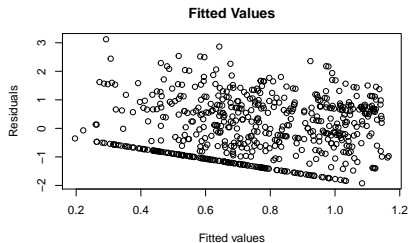
Summary of M.lm

```
##
## Call:
## lm(formula = NegPerChick ~ SexParent * (FoodTreatment + ArrivalTime),
##     data = Owls)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.3317 -1.0833 -0.4573  0.8630  7.0978
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   6.766988   1.259815   5.371 1.12e-07 ***
## SexParentMale                 -0.044492   1.652324  -0.027  0.979
## FoodTreatmentSatiated         -0.850112   0.195174  -4.356 1.56e-05 ***
## ArrivalTime                   -0.198358   0.050578  -3.922 9.81e-05 ***
## SexParentMale:FoodTreatmentSatiated  0.086027   0.254579   0.338  0.736
## SexParentMale:ArrivalTime       0.007349   0.066230   0.111  0.912
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.527 on 593 degrees of freedom
## Multiple R-squared:  0.1134, Adjusted R-squared:  0.1059
## F-statistic: 15.16 on 5 and 593 DF,  p-value: 4.941e-14
```

Log-transform the negotiations per chick

This helps some with the heteroscedasticity in variance.

```
Owls$LogNeg<-log(Owls$NegPerChick+1)
M2.lm=lm(LogNeg~SexParent*(FoodTreatment+ArrivalTime),data=Owls)
```



Summary of M2.1m

```
##
## Call:
## lm(formula = LogNeg ~ SexParent * (FoodTreatment + ArrivalTime),
##     data = Owls)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.08353 -0.46500 -0.05645  0.41842  1.74442
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   2.574961   0.466240   5.523 5.00e-08 ***
## SexParentMale                 0.176201   0.611502   0.288 0.773336
## FoodTreatmentSatiated        -0.364079   0.072231  -5.040 6.18e-07 ***
## ArrivalTime                 -0.068917   0.018718  -3.682 0.000253 ***
## SexParentMale:FoodTreatmentSatiated  0.006885   0.094216   0.073 0.941767
## SexParentMale:ArrivalTime     -0.003275   0.024511  -0.134 0.893745
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5649 on 593 degrees of freedom
## Multiple R-squared:  0.1402, Adjusted R-squared:  0.133
## F-statistic: 19.35 on 5 and 593 DF,  p-value: < 2.2e-16
```

A mixed model with nlme::lme

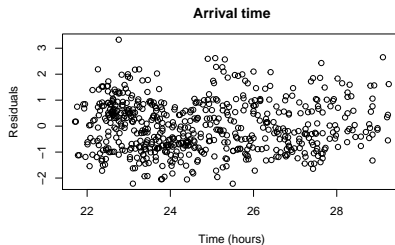
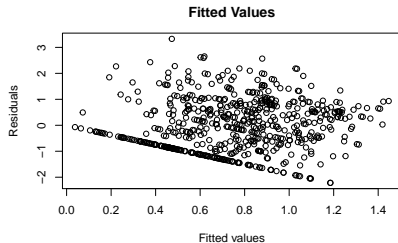
Refit the linear model with `nlme::gls` so that it can be more easily compared to the mixed model. The fitted model is identical to the one above, but R puts it in a different namespace.

- ▶ The argument `random=~1|Nest` in `lme` indicates that the intercept varies depending on the nest.
- ▶ Since observations from the same nest are correlated, the effective number of observations is reduced.
- ▶ The `anova` command carries out a likelihood ratio test since we give it `lme` objects.

```
Form <- formula(LogNeg~SexParent*(FoodTreatment+ArrivalTime))
M.gls <- gls(Form,data=Owls)
M1.lme <- lme(Form,random=~1|Nest,data=Owls)
anova(M.gls,M1.lme)
```

##	Model	df	AIC	BIC	logLik	Test	L.Ratio	p-value
##	M.gls	1 7	1053.537	1084.233	-519.7683			
##	M1.lme	2 8	1026.878	1061.959	-505.4390	1 vs 2	28.65875	<.0001

Diagnostic plots of the mixed model



Summary of the mixed model

```
## Linear mixed-effects model fit by REML
##   Data: Owls
##       AIC      BIC    logLik
##   1026.878 1061.96 -505.439
##
## Random effects:
##   Formula: ~1 | Nest
##           (Intercept) Residual
##   StdDev:    0.2143996 0.5345205
##
## Fixed effects: list(Form)
##                                     Value Std.Error DF   t-value p-value
## (Intercept)                      2.5872800 0.4495127 567   5.755744 0.0000
## SexParentMale                     0.2491715 0.5861657 567   0.425087 0.6709
## FoodTreatmentSatiated             -0.4188292 0.0705245 567  -5.938776 0.0000
## ArrivalTime                       -0.0667932 0.0180023 567  -3.710251 0.0002
## SexParentMale:FoodTreatmentSatiated 0.0322772 0.0914373 567   0.352998 0.7242
## SexParentMale:ArrivalTime          -0.0088322 0.0234809 567  -0.376144 0.7070
## Correlation:
##                                     (Intr) SxPrnM FdTrtS ArrvlT SPM:FT
## SexParentMale                      -0.749
## FoodTreatmentSatiated              -0.102   0.080
## ArrivalTime                        -0.989   0.746   0.022
## SexParentMale:FoodTreatmentSatiated 0.078 -0.113 -0.756 -0.019
## SexParentMale:ArrivalTime          0.747 -0.994 -0.019 -0.755   0.037
##
## Standardized Within-Group Residuals:
##           Min           Q1           Med           Q3           Max
##   -2.2182205 -0.7902406 -0.0778161  0.6991138  3.3273193
##
## Number of Observations: 599
## Number of Groups: 27
```

Model selection

```
M1.lmeML=lme(Form,random=~1|Nest,method="ML",data=Owls)
dredge(M1.lmeML)
```

```
## Global model call: lme.formula(fixed = Form, data = Owls, random = ~1 | Nest, method = "ML")
## ---
## Model selection table
##      (Int)      ArT FdT SxP ArT:SxP FdT:SxP df   logLik   AICc delta weight
## 4  2.7200 -0.07137  +                               5 -491.813  993.7  0.00  0.476
## 8  2.7030 -0.07191  +  +                               6 -491.347  994.8  1.11  0.273
## 16 2.5750 -0.06670  +  +             +                7 -491.273  996.7  3.01  0.106
## 24 2.7110 -0.07185  +  +             +                7 -491.280  996.8  3.02  0.105
## 32 2.5870 -0.06682  +  +             +                8 -491.211  998.7  4.94  0.040
## 3  0.9475                +                4 -509.674 1027.4 33.69  0.000
## 7  0.9253                +  +                5 -509.444 1029.0 35.26  0.000
## 23 0.9371                +  +                6 -509.354 1030.9 37.12  0.000
## 6  2.3950 -0.06804                +                5 -526.370 1062.8 69.11  0.000
## 2  2.4220 -0.06712                4 -527.605 1063.3 69.55  0.000
## 14 2.3170 -0.06488                +  +                6 -526.346 1064.8 71.11  0.000
## 1  0.7604                3 -541.668 1089.4 95.65  0.000
## 5  0.7189                +                4 -540.826 1089.7 95.99  0.000
## Models ranked by AICc(x)
## Random terms (all models):
## '1 | Nest'
```

It does not appear that sex of the parent is a significant predictor. It is close though. Perhaps there is more work to do.

Fit the optimal linear mixed model

Sex of the parent is not a predictor.

```
M1.opt <- lme(LogNeg~FoodTreatment+ArrivalTime,random=~1|Nest,data=Owls)
summary(M1.opt)
```

```
## Linear mixed-effects model fit by REML
##   Data: Owls
##       AIC       BIC    logLik
##   1009.241 1031.192 -499.6203
##
## Random effects:
## Formula: ~1 | Nest
##      (Intercept) Residual
## StdDev:    0.2180265 0.5333704
##
## Fixed effects: LogNeg ~ FoodTreatment + ArrivalTime
##              Value Std.Error DF   t-value p-value
## (Intercept)    2.7219747 0.29697570 570   9.165648     0
## FoodTreatmentSatiated -0.4031260 0.04597355 570  -8.768650     0
## ArrivalTime        -0.0714293 0.01177156 570  -6.067954     0
## Correlation:
##              (Intr) FdTrtS
## FoodTreatmentSatiated -0.112
## ArrivalTime          -0.984 0.039
##
## Standardized Within-Group Residuals:
##      Min           Q1           Med           Q3           Max
## -2.22283609 -0.78307304 -0.07461892  0.68690000  3.29183331
##
## Number of Observations: 599
## Number of Groups: 27
```

What if we don't use Nest as a random effect?

```
M1.wrong=lm(LogNeg~SexParent*(FoodTreatment+ArrivalTime)+Nest,data=Owls)
Anova(M1.wrong)
```

```
## Anova Table (Type II tests)
##
## Response: LogNeg
##
```

	Sum Sq	Df	F value	Pr(>F)
## SexParent	0.106	1	0.3745	0.5408
## FoodTreatment	22.383	1	79.0186	< 2.2e-16 ***
## ArrivalTime	10.865	1	38.3571	1.132e-09 ***
## Nest	28.654	26	3.8907	8.526e-10 ***
## SexParent:FoodTreatment	0.049	1	0.1716	0.6789
## SexParent:ArrivalTime	0.087	1	0.3074	0.5795
## Residuals	160.607	567		

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The result is similar, but it applies *only* to these 27 nests.

Comparison of parameter estimates

They are almost identical. The mixed model predicts a 3-6% smaller effect of food treatment.

```
intervals(M1.lme)$fixed[2:4, c(1,3)]
```

##		lower	upper
##	SexParentMale	-0.9021498	1.40049288
##	FoodTreatmentSatiated	-0.5573504	-0.28030806
##	ArrivalTime	-0.1021526	-0.03143379

```
apply(intervals(M1.lme)$fixed[2:4, c(1,3)], 1, diff)
```

##	SexParentMale	FoodTreatmentSatiated	ArrivalTime
##	2.30264266	0.27704233	0.07071885

```
confint(M1.wrong)[2:4,]
```

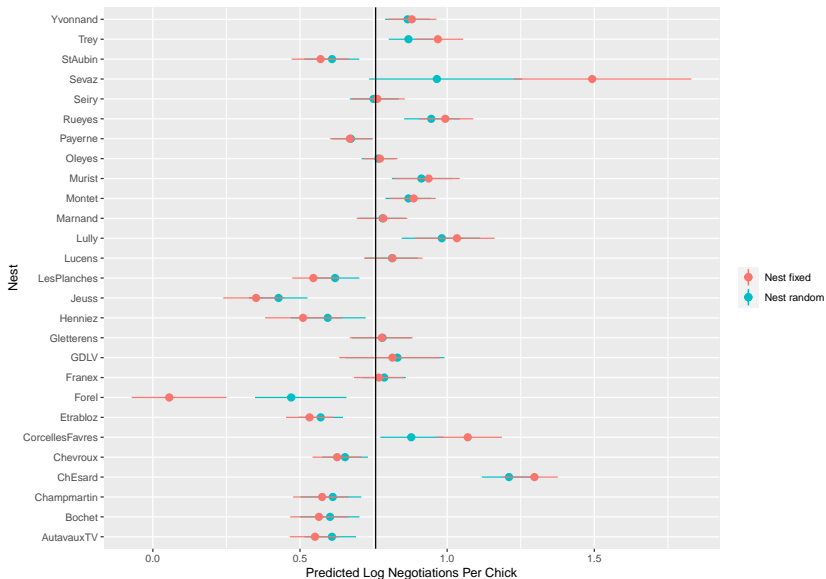
##		2.5 %	97.5 %
##	SexParentMale	-0.8178743	1.4865913
##	FoodTreatmentSatiated	-0.5772142	-0.2988176
##	ArrivalTime	-0.1015671	-0.0308154

```
apply(confint(M1.wrong)[2:4,], 1, diff)
```

##	SexParentMale	FoodTreatmentSatiated	ArrivalTime
##	2.30446563	0.27839658	0.07075168

Comparison of predictions

The model where Nest is a random effect makes predictions that are closer to the overall mean.



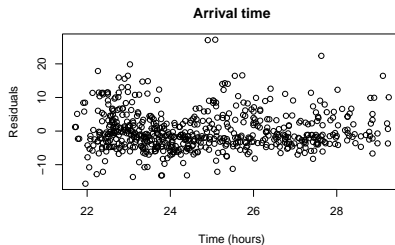
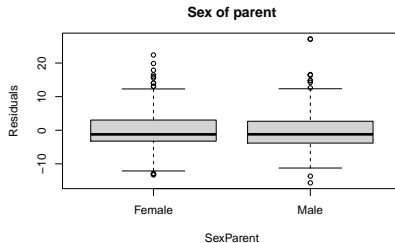
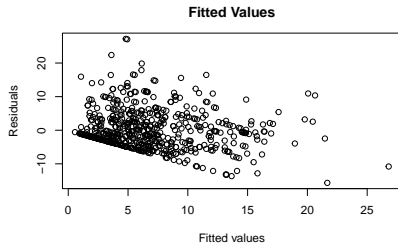
A model built with glmmTMB

Instead of log-transforming the negotiations, we could use a log link and a negative binomial error distribution.

- ▶ Include `offset(BroodSize)` instead of using `NegPerChick`.
- ▶ Random effect of `1|Nest` is included in the main formula.
- ▶ Include zero inflation with `zi = ~1`. The formula `~1` means that we expect there to be excess zeros in the data but we do not propose a reason for them.

```
M.tmb <- glmmTMB(SiblingNegotiation ~ SexParent*(FoodTreatment + ArrivalTime) +  
  (1|Nest) + offset(BroodSize),  
  family = nbinom1(), zi = ~1, data=Owls)
```

Diagnostic plots of the TMB model



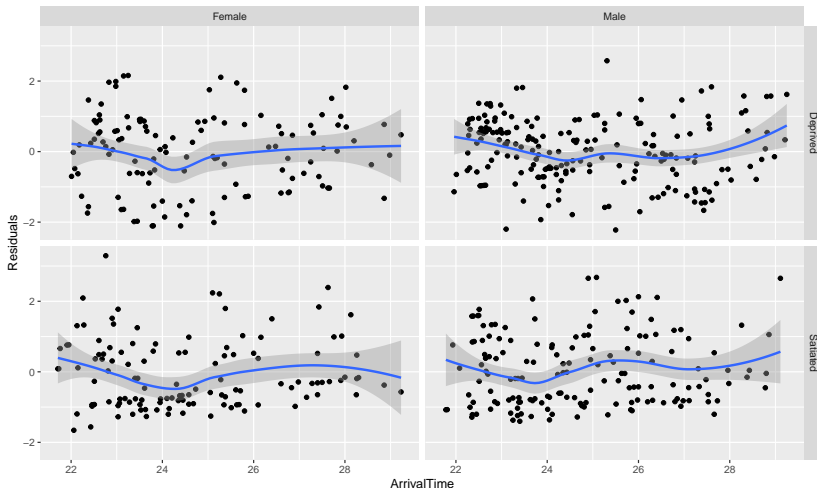
Summary of M1.tmb

```
## Family: nbinom1 ( log )
## Formula:
## SiblingNegotiation ~ SexParent * (FoodTreatment + ArrivalTime) +
## (1 | Nest) + offset(BroodSize)
## Zero inflation: ~1
## Data: Owls
##
##      AIC      BIC    logLik deviance df.resid
## 3394.1    3433.6   -1688.0   3376.1      590
##
## Random effects:
##
## Conditional model:
##   Groups Name      Variance Std.Dev.
##   Nest  (Intercept) 1.157    1.076
## Number of obs: 599, groups: Nest, 27
##
## Overdispersion parameter for nbinom1 family (): 5.01
##
## Conditional model:
##                                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       0.86649    0.85366   1.015 0.310092
## SexParentMale                     0.57261    1.04563   0.548 0.583950
## FoodTreatmentSatiated             -0.94619    0.13788  -6.862 6.77e-12 ***
## ArrivalTime                      -0.11557    0.03400  -3.399 0.000676 ***
## SexParentMale:FoodTreatmentSatiated 0.28153    0.17318   1.626 0.104017
## SexParentMale:ArrivalTime          -0.02681    0.04294  -0.624 0.532409
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Zero-inflation model:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.5605    0.3625  -7.063 1.63e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The effect of arrival time is not linear

SexParent-FoodTreatment combinations show residual patterns.

```
Owls$Eopt <- resid(M1.opt,type="normalized")  
ggplot(Owls, aes(ArrivalTime, Eopt))+ geom_point()+geom_smooth()+  
  facet_grid(FoodTreatment~SexParent)+ylab("Residuals")
```



GAMM

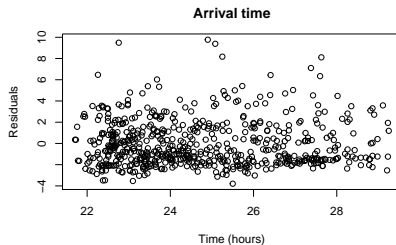
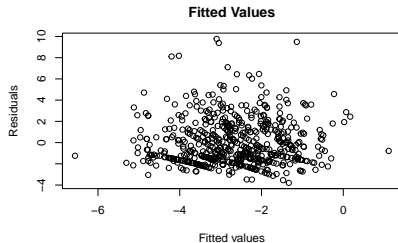
A generalized additive mixed model gives flexibility to include nonlinear effects.

- ▶ A black-box smooth function relating ArrivalTime to SiblingNegotiation is created with `s()`.
- ▶ Including `by=SexParent` substitutes for the interaction.
- ▶ Random effects go in a `list`.
- ▶ Negative binomial errors aren't an option so we use Poisson.

```
M.gamm<-gamm(SiblingNegotiation~offset(BroodSize)+  
              SexParent*FoodTreatment+s(ArrivalTime, by=SexParent),  
              random=list(Nest=~1),data=Owls,  
              family=poisson)
```

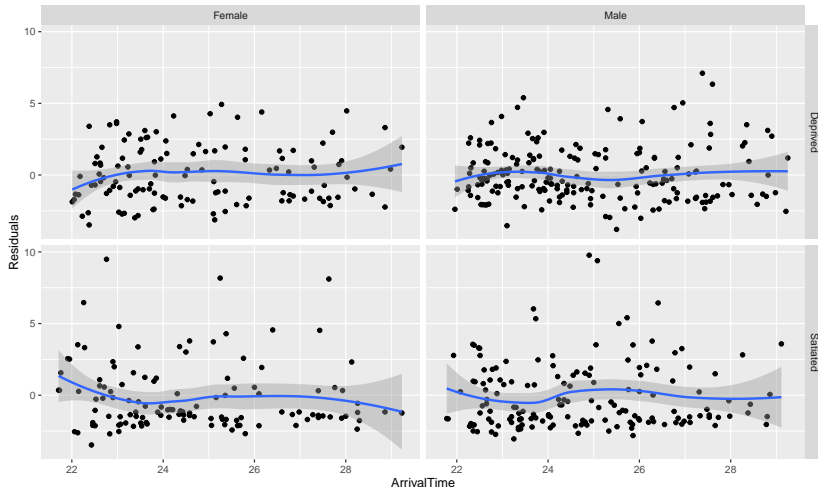
```
##  
## Maximum number of PQL iterations: 20
```

Diagnostic plots of the generalized additive mixed model



Using a GAMM straightens out the residuals

```
Owls$Egamm <- resid(M.gamm,type="normalized")  
ggplot(Owls, aes(ArrivalTime, Egamm))+ geom_point()+geom_smooth()+  
  facet_grid(FoodTreatment~SexParent)+ylab("Residuals")
```



Summary of M.gamm\$gam

```
##
## Family: poisson
## Link function: log
##
## Formula:
## SiblingNegotiation ~ offset(BroodSize) + SexParent * FoodTreatment +
##   s(ArrivalTime, by = SexParent)
##
## Parametric coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -2.17039    0.23417  -9.268   <2e-16 ***
## SexParentMale   -0.00743    0.04830  -0.154   0.8778
## FoodTreatmentSatiated -0.68125    0.05721 -11.908   <2e-16 ***
## SexParentMale:FoodTreatmentSatiated  0.14333    0.07172   1.998   0.0461 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##               edf Ref.df      F p-value
## s(ArrivalTime):SexParentFemale 8.324  8.324 24.82  <2e-16 ***
## s(ArrivalTime):SexParentMale   8.096  8.096 35.85  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  -16.2
##   Scale est. = 1          n = 599
```

Baby Owl says, "Hey Papa, bring us another mouse! The one the scientists left for us smells funny."

F tests on the GAMM

```
anova(M.gamm$gam)
```

```
##
## Family: poisson
## Link function: log
##
## Formula:
## SiblingNegotiation ~ offset(BroodSize) + SexParent * FoodTreatment +
##      s(ArrivalTime, by = SexParent)
##
## Parametric Terms:
```

	df	F	p-value
## SexParent	1	0.024	0.8778
## FoodTreatment	1	141.794	<2e-16
## SexParent:FoodTreatment	1	3.994	0.0461

```
##
## Approximate significance of smooth terms:
```

	edf	Ref.df	F	p-value
## s(ArrivalTime):SexParentFemale	8.324	8.324	24.82	<2e-16
## s(ArrivalTime):SexParentMale	8.096	8.096	35.85	<2e-16

Summary of M.gamm\$lme

`summary(M.gamm$lme)` includes the correlation matrix.

- ▶ Highly correlated parameters may warrant an explanation.
- ▶ Intercept is often highly correlated with slope for variables in linear models that are not centered.
- ▶ Interaction terms tend to be correlated with the interacting variables.

```
sml <- summary(M.gamm$lme)
colnames(sml$corFixed) <- abbreviate(colnames(sml$corFixed), minlength=8)
rownames(sml$corFixed) <- abbreviate(rownames(sml$corFixed), minlength=8)
round(sml$corFixed,3)
```

##	X(Intrc)	XSxPrntM	XFdTrtmS	XSPM:FTS	X(AT):SPF	X(AT):SPM
## X(Intrc)	1.000	-0.130	-0.092	0.069	0.007	0.001
## XSxPrntM	-0.130	1.000	0.457	-0.569	-0.043	-0.090
## XFdTrtmS	-0.092	0.457	1.000	-0.769	0.093	-0.001
## XSPM:FTS	0.069	-0.569	-0.769	1.000	-0.067	0.034
## X(AT):SPF	0.007	-0.043	0.093	-0.067	1.000	0.003
## X(AT):SPM	0.001	-0.090	-0.001	0.034	0.003	1.000

AIC and significance of SexParent

Fitting a model without SexParent gives a much higher AIC.

```
M2.gamm <- gamm(SiblingNegotiation~offset(BroodSize)+  
                FoodTreatment+s(ArrivalTime),  
                random=list(Nest=~1),data=Owls,  
                family=poisson)
```

```
##
```

```
## Maximum number of PQL iterations: 20
```

```
AIC(M.gamm, M2.gamm)
```

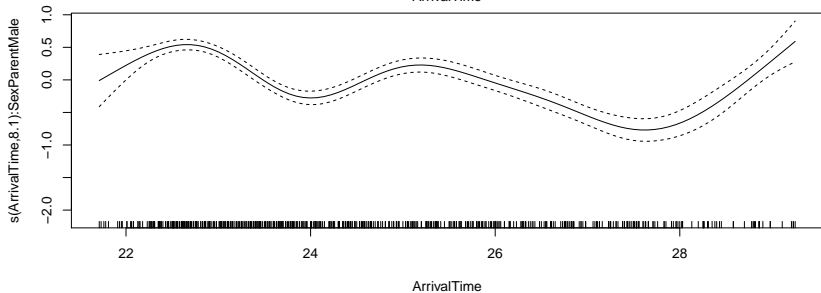
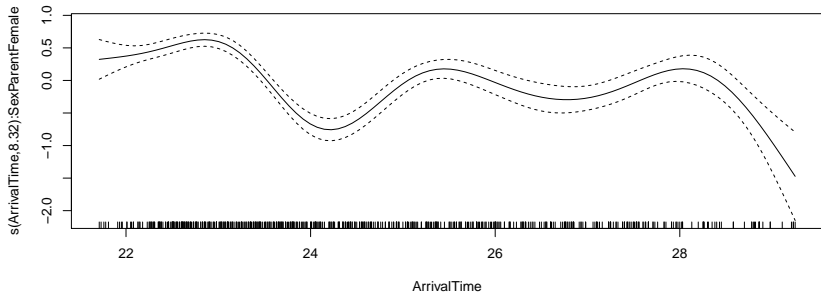
```
##          df      AIC
```

```
## M.gamm    9 3107.012
```

```
## M2.gamm   5 3177.889
```

The smooth approximators

```
plot(M.gamm$gam)
```



Enough owl models already

Roulin and Bersier (2007) finds that the sex of the parent *is* significant without going further than a linear mixed model.

The analysis there seems to hinge on the variable *amount of time in nestbox* which is missing from the data provided in Zuur et al. (2009) (which is the source of the data in `glmmTMB::Owls`).

Mixed model formula cheatsheet

Y is response, X is a fixed effect variable, S is a random effect variable.

Modeling goal	Formula specification	Mathematical specification $Y \sim$
fixed effect only model	NA	$\beta_0 + \beta_1 X + \epsilon_s$
random group intercept	(1group)	$(\beta_0 + b_{0Ss}) + \beta_1 X + \epsilon_s$
random slope of x within group with correlated intercept	$(x \text{group}) = (1+x \text{group})$	$(\beta_0 + b_{0Ss}) + (\beta_1 + b_{1Ss})X + \epsilon_s$
random slope of x within group: no variation in intercept	$(0+x \text{group}) = (-1+x \text{group})$	$\beta_0 + (\beta_1 + b_{1Ss})X + \epsilon_s$
intercept varying among sites and among blocks within sites (nested random effects)	$(1 \text{site/block}) = (1 \text{site}) + (1 \text{site:block})$	$(\beta_0 + b_{0Ss} + b_{0STst}) + \beta_1 X + \epsilon_{st}$
intercept varying among crossed random effects (e.g. site, year)	$(1 \text{group1}) + (1 \text{group2})$	$(\beta_0 + b_{0Ss} + b_{0Tt}) + \beta_1 X + \epsilon_{st}$

References

Roulin, Alexandre, and Louis-Felix Bersier. 2007. "Nestling Barn Owls Beg More Intensely in the Presence of Their Mother Than in the Presence of Their Father." *Animal Behaviour* 74 (4): 1099–1106.
<https://doi.org/https://doi.org/10.1016/j.anbehav.2007.01.027>.

Zuur, A., E. N. Ieno, N. Walker, A. A. Saveliev, and G. M. Smith. 2009. *Mixed Effects Models and Extensions in Ecology with R*. Statistics for Biology and Health. Springer New York.
<https://books.google.com/books?id=vQUNprFZKHsC>.