Extension of Linear regression

R Ladies Bogor

17 November 2019

Overview alur permodelan

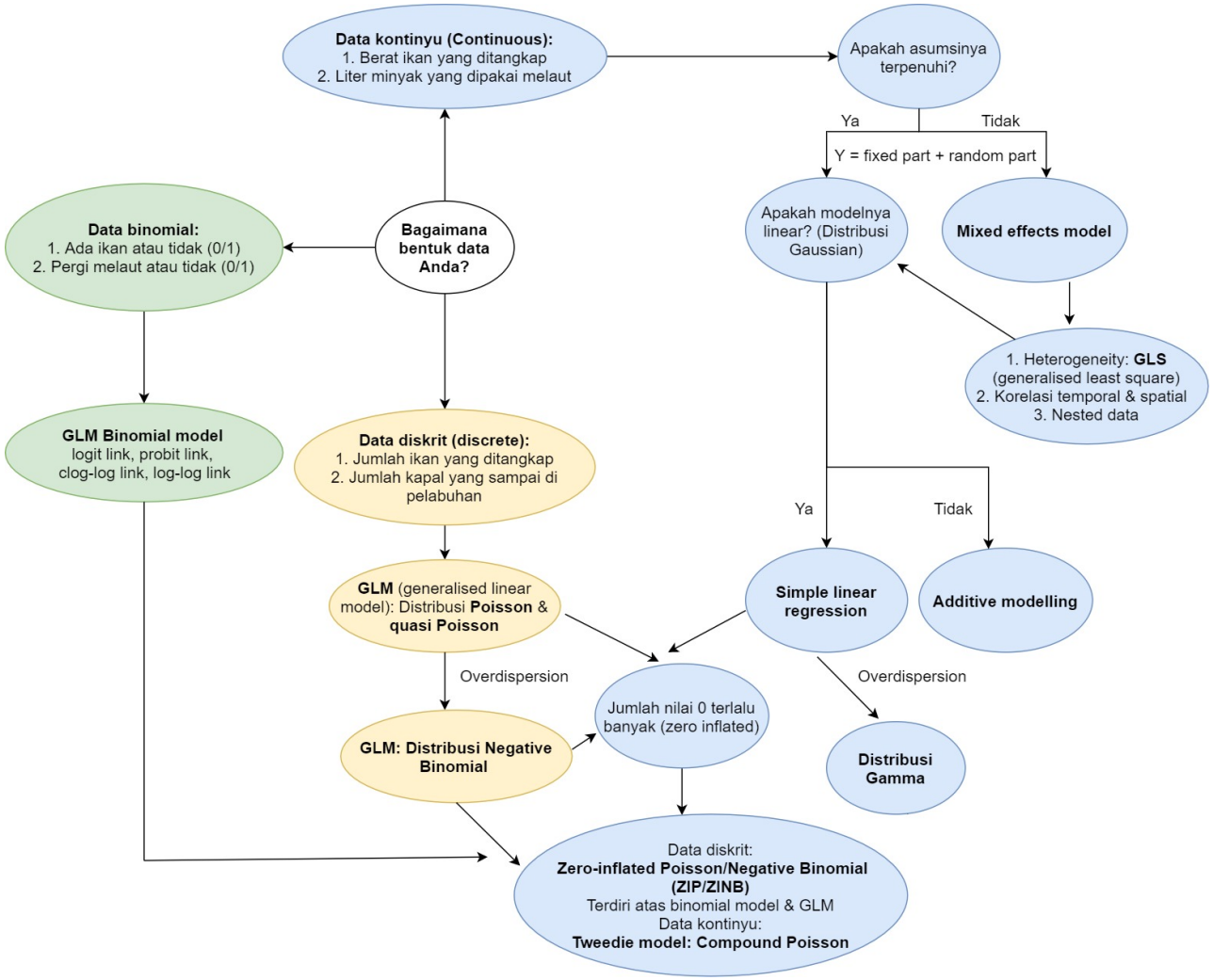


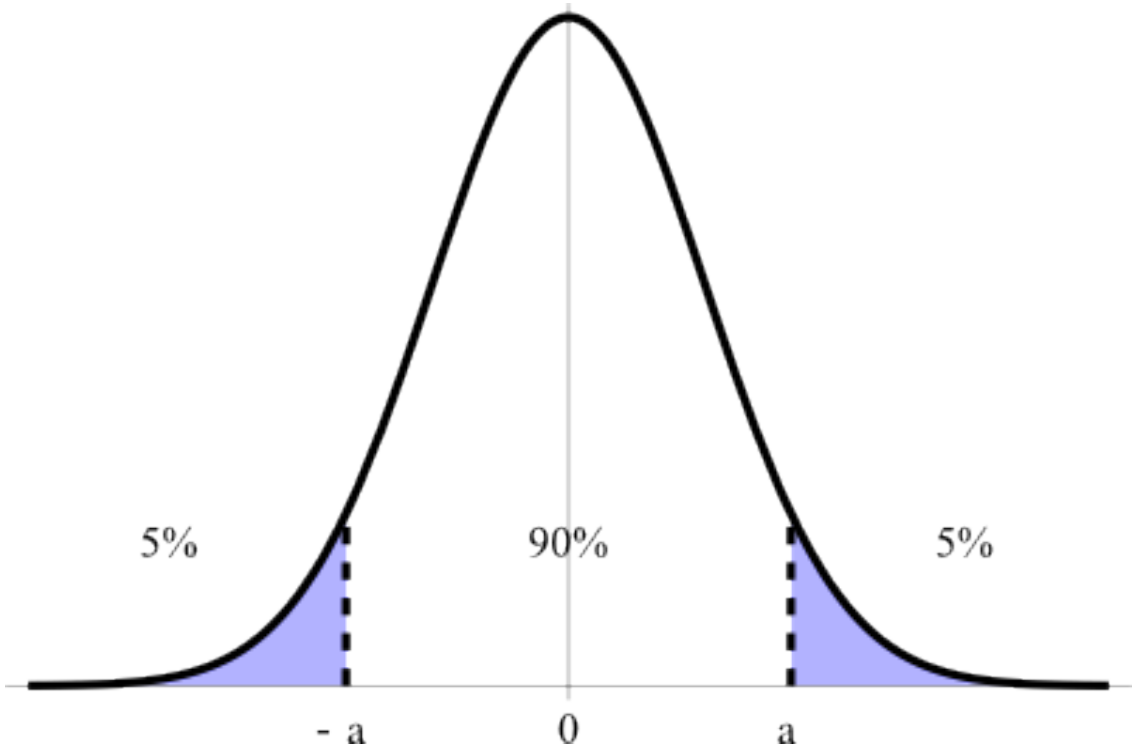
Figure 1: Overview alur permodelan

Extention of linear regression

1. additive model + generalized least square
2. cont. gls + gamma + binomial
3. poisson + neg.binomial
4. zero inflated

Data kontinyu dan metode analisisnya

- Distribusi Gaussian
- Asumsi distribusi Gaussian harus dipenuhi
- Simple linear regression atau additive model.
- . . . atau mixed effects model (koreksi random part) (generalized least square)



Asumsi yang harus dipenuhi dalam menggunakan distribusi Gaussian (normal)

1. Normality = tes formal/histogram
2. Homogeneity = residual data tidak berpola
3. Fixed X = tidak memiliki pengetahuan apriori
4. Independence = tidak tergantung pada variabel lain/kondisi spasial atau temporal

1. Uji normalitas

- Central limit theorem: bila sampel kita cukup besar (n>30), uji normal bisa diabaikan.
- Contoh menggunakan data *C. boenak*.
- Apakah yang mempengaruhi jumlah ikan *C. boenak* (kg) yang ditangkap nelayan?

Central limit theorem = variabel acak yang jumlahnya banyak akan menghasilkan distribusi normal

Cephalopholis boenak

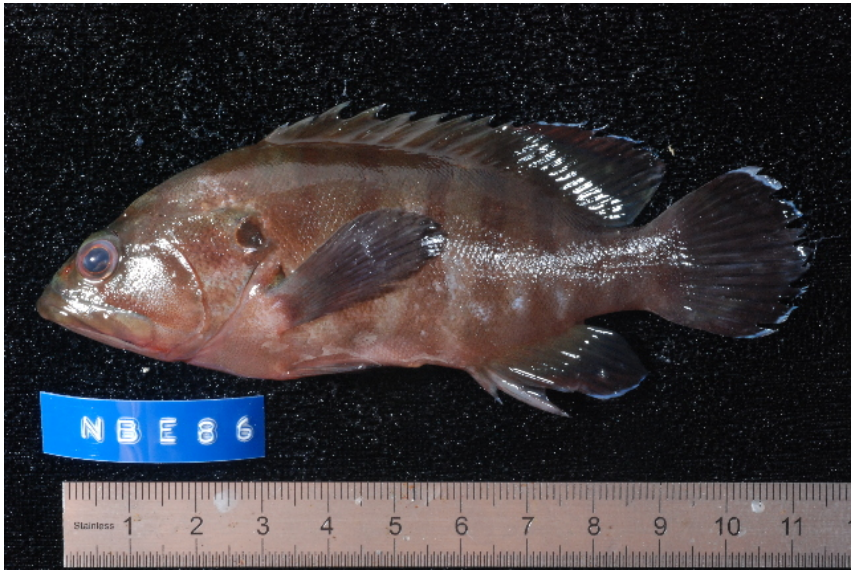
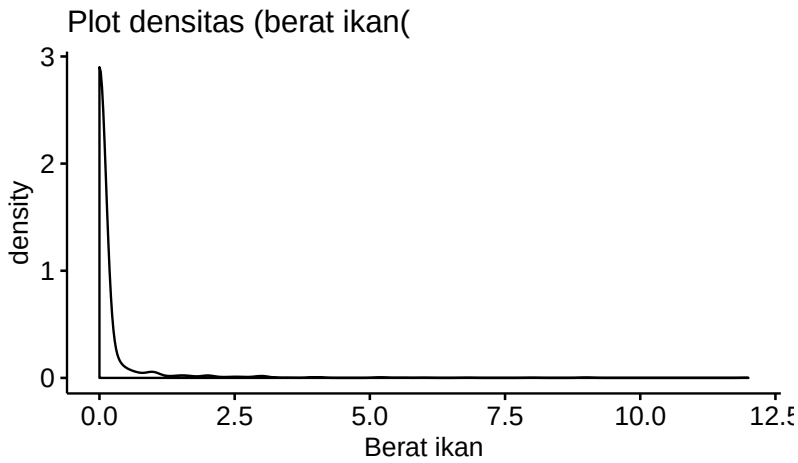
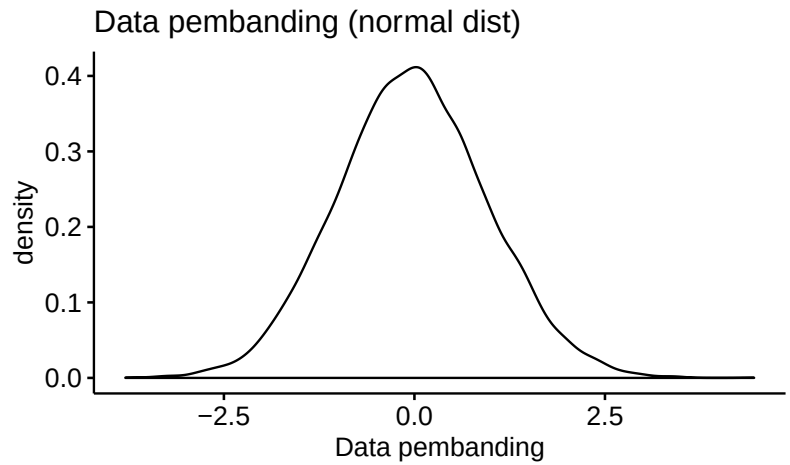
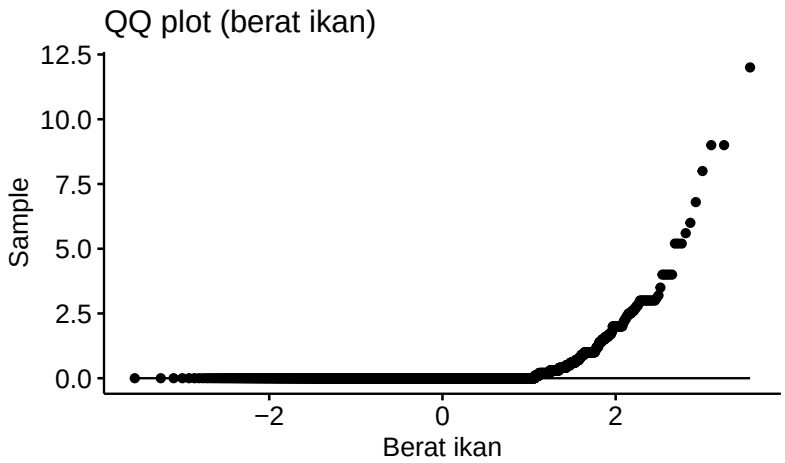


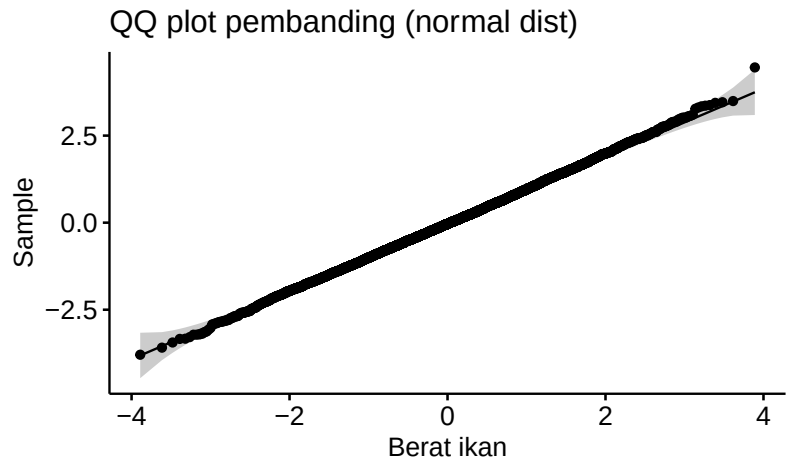
Figure 2: By Serge Planes - https://commons.wikimedia.org/w/index.php?curid=20092429

Plot densitas & QQ-plot untuk berat ikan

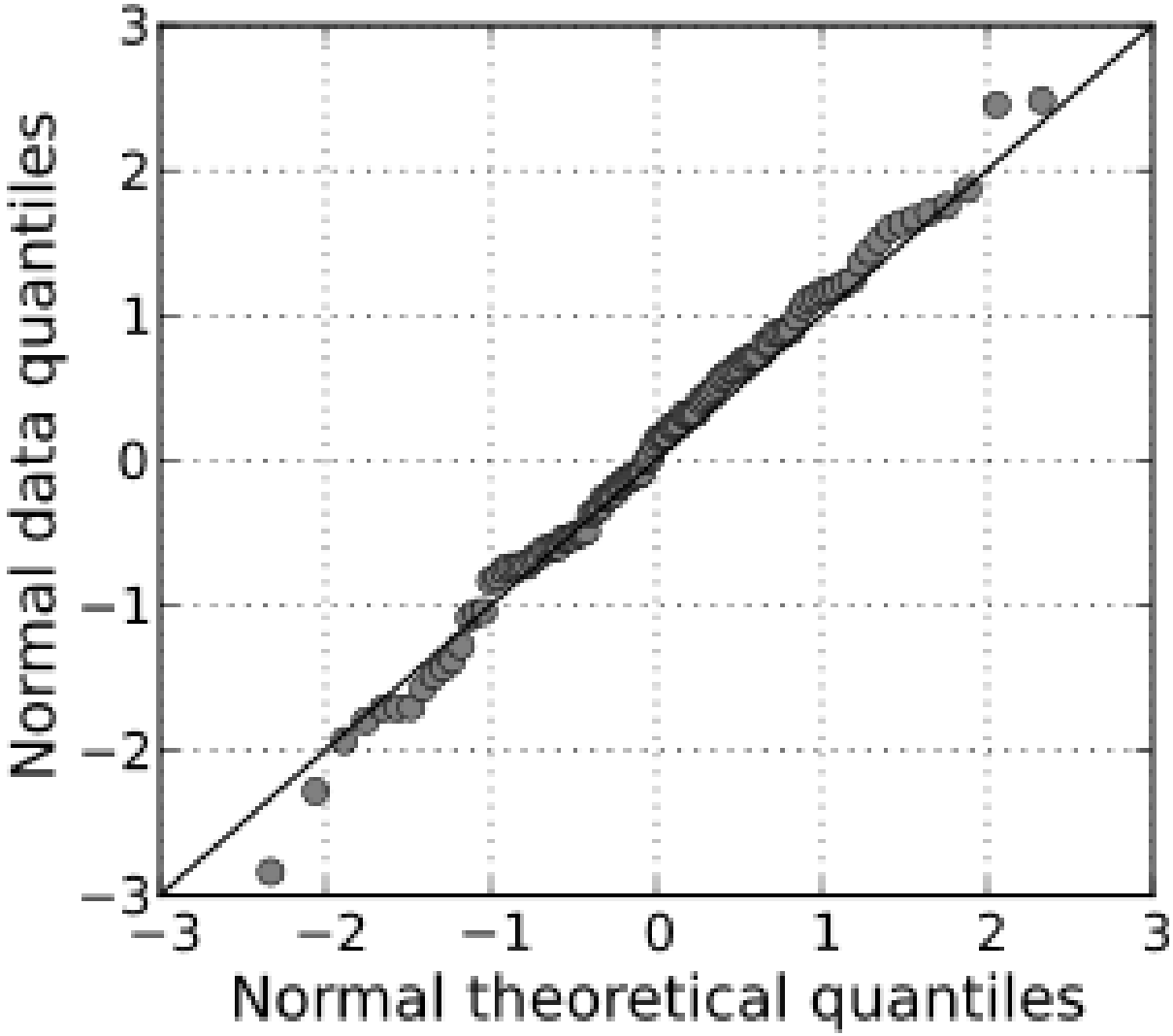








Contoh Q-Q plot dengan distribusi normal:



Menggunakan Shapiro-Wilk normality test:

```
shapiro.test(db.data$KG)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  db.data$KG
## W = 0.24532, p-value < 2.2e-16
```

p-value >0.05 = distribusi data tidak berbeda secara signifikan dari distribusi normal (data berdistribusi normal).

2. Heterogeneity of variances

- Pada linear regression = explanatory variables berasal dari populasi yang sama.
- Ketika kita mem-fit-kan model, residual data tidak menunjukkan pola apapun.
- Bila ada = teori linear regression model kita tidak valid

Cara menguji pelanggaran homogeneity

- Diagnosa plot residual vs fitted values model yang kita uji.
- Variance dari variabel yang dimodelkan = konstan dan tidak mengikuti pola tertentu

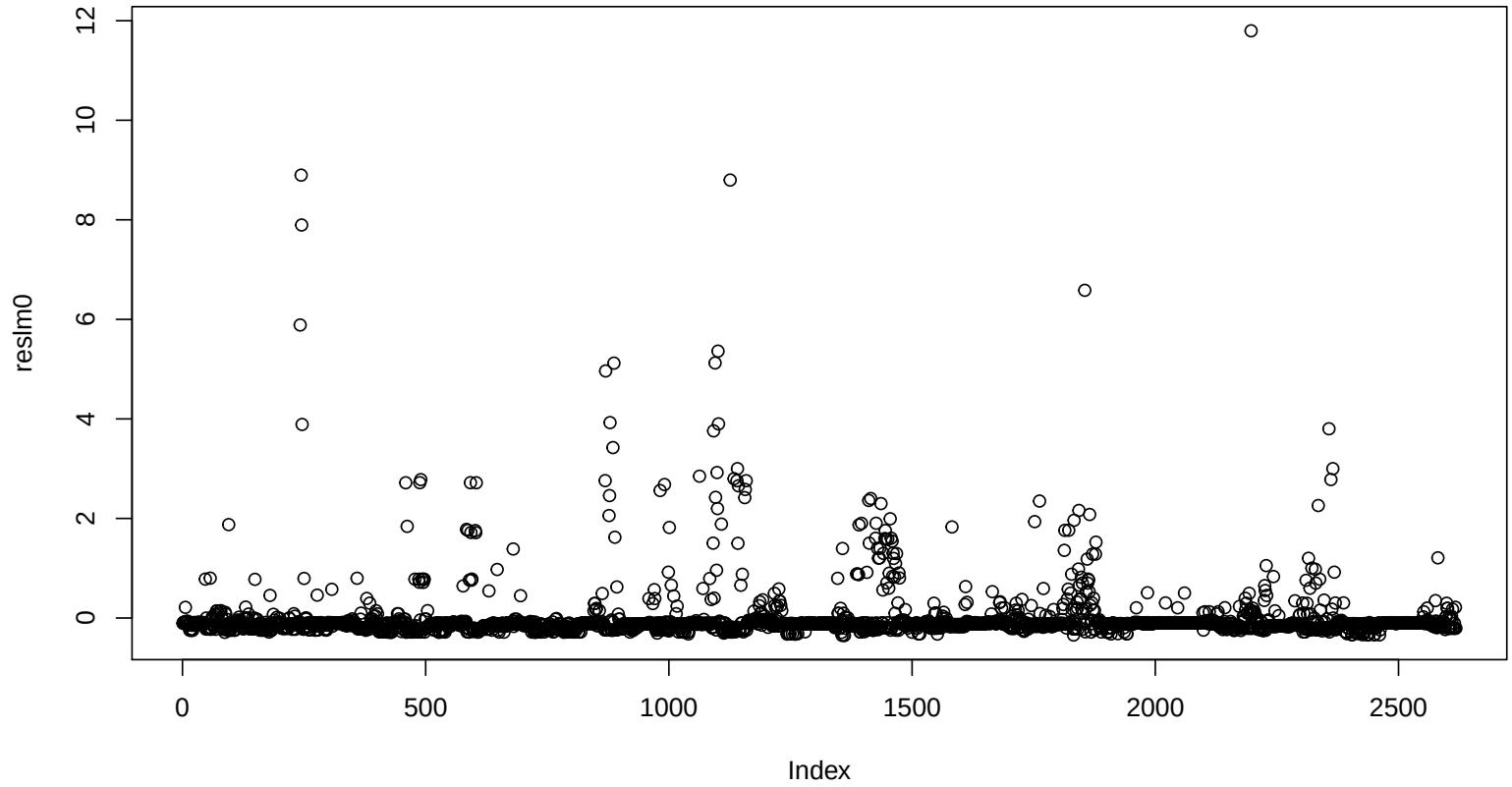
Contoh observasi visual homogeneity

- Plot kg ikan dengan variabel-variabel penjelas.
- Random parts (sumber error) diabaikan.

$$Y = fixed.parts$$

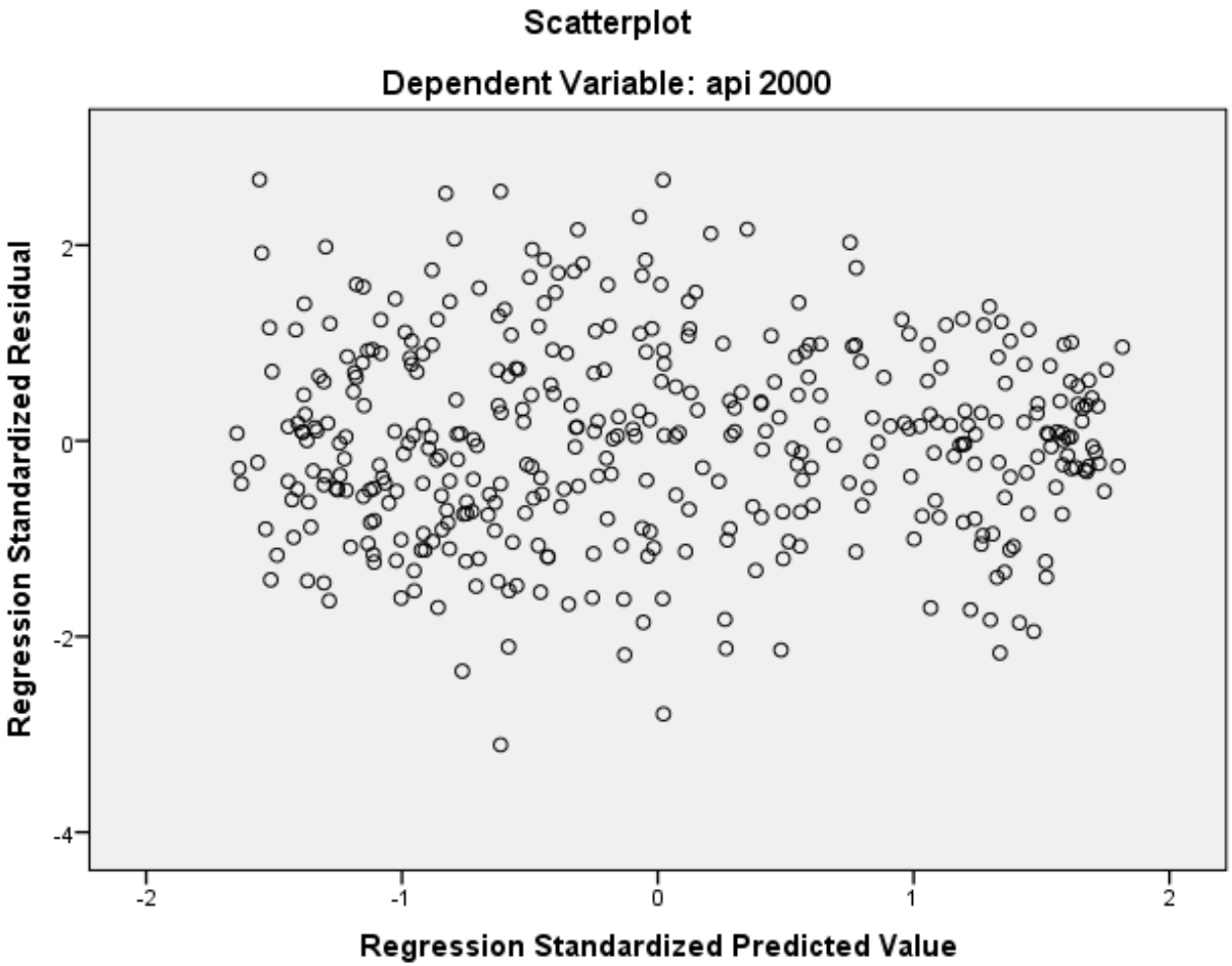
$$kg.ikan.ditangkap(Y) = jumlah.hook + jumlah.jam$$

Hasil plot residual dan AIC




```
## [1] 5055.287
```

Contoh plot residual yang tidak berpola



Contoh plot residual berpola

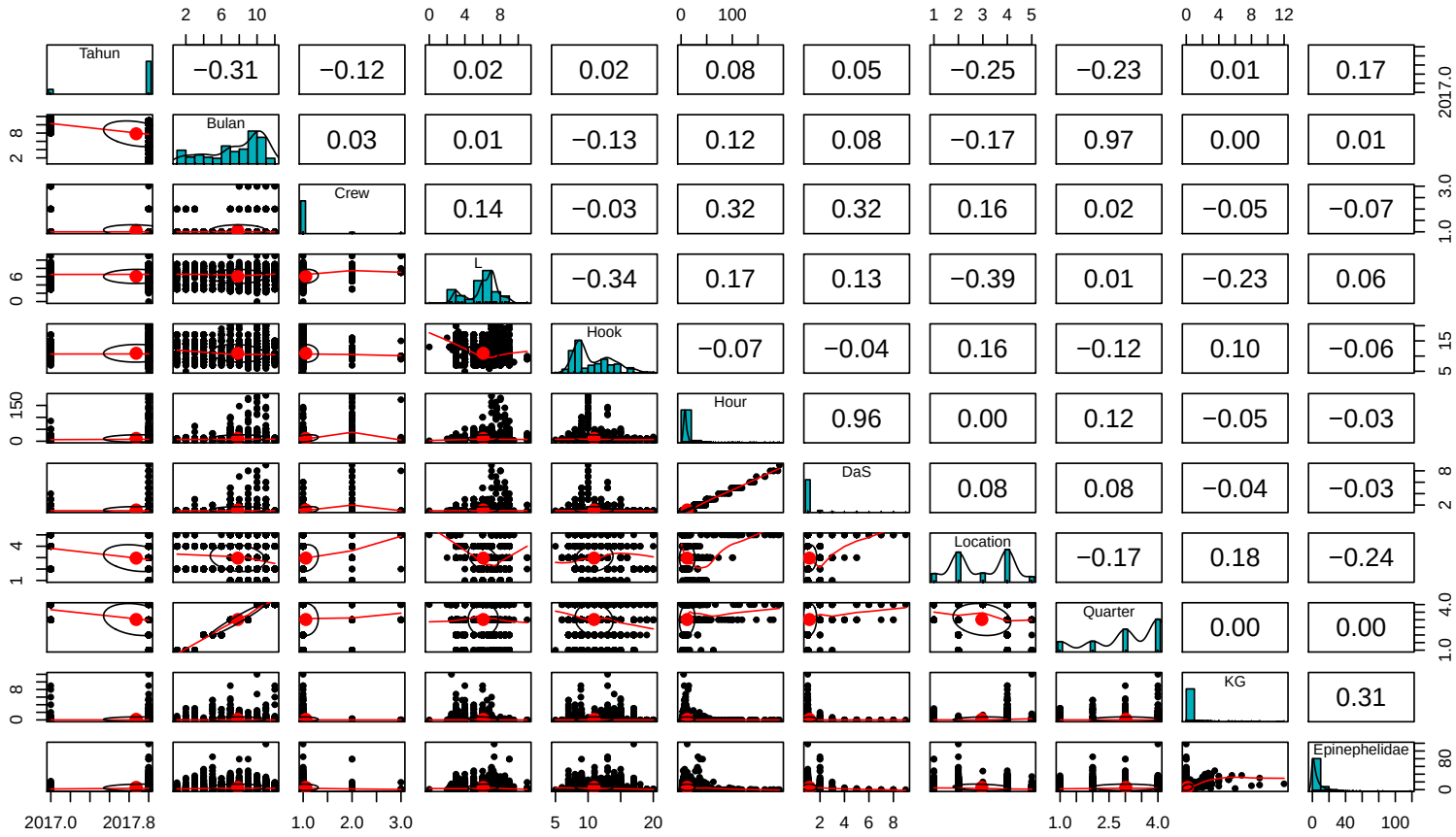
3. Fixed X

- Asumsi yang mengimplikasikan = explanatory variables sudah diketahui nilainya (apriori)
- Contoh: memilih situs yang konsentrasi racunnya sudah kita ketahui
- Eror sampling bukan masalah besar apabila kecil dibanding variasi sampel/rentang data yang kita miliki. Eg., 20 sampel dengan variasi nilai 15-20 derajat Celcius, eror termometer kita adalah 0.1.

4. Independence

- Masalah paling = bisa mem-disvalidasi tes-tes penting seperti F-test atau t-test.
- Muncul ketika = nilai Y pada Xi dipengaruhi oleh Xi lainnya.
  - Disebabkan oleh: pemilihan model yang tidak tepat dan kondisi alami data tersebut.
  - Plot garis lurus = ada pola non-linear antara Y dan X.
  - Plot residual terhadap X = ada pola yang teratur
- Gunakan model yang lebih baik/transformasi data = hubungan menjadi linear.
- Gunakan uji collinearity = identifikasi variabel-variabel yang berkorelasi

Uji multicollinearity (visual)



Uji multicollinearity (VIF)

- Gunakan Variance Inflation Factors (VIF).
  - VIF $= 1$: tidak ada korelasi antara variable
  - VIF $=$ 1-5: korelasi 'medium'
  - VIF > 5: korelasi tinggi
- Akibat korelasi tinggi $=$ estimasi koefisien buruk; nilai p-value dipertanyakan.

Uji multicollinearity (VIF)

```
library(caret)
library(car)
model1 <- lm(IND ~ Bulan + Quarter + Boat +
                Crew + L + Hook + offset(DaS),
              data = db.data)
vif(model1)
```

```
##                GVIF  Df GVIF^(1/(2*Df))
## Bulan     20.926889   1        4.574592
## Quarter   21.017404   1        4.584474
## Boat     417.293630 338        1.008966
## Crew       4.833682   1        2.198564
## L         21.061244   1        4.589253
## Hook       2.390691   1        1.546186
```

Pelanggaran Independence lainnya

- Karena kondisi data itu sendiri.
- Apa yang kita makan menit ini bergantung pada apa yang kita minum 5 menit yang lalu.
- Apabila hujan turun pada jarak 200 m dari tanah, maka sebetulnya hujan juga tengah terjadi pada 100 m dari tanah.
- Diatasi dengan memasukkan struktur dependesi temporal ataupun spasial pada model.

Analisis lanjutan data continuous ikan *C. boenak*

1. Menambahkan interaction terms

$$kg(Y) = hook + jam + hook \times jam$$

2. Menambahkan fixed effect berupa variabel non-linear

$$kg(Y) = hook + hook^2$$

3. Menambahkan variabel penjelas lainnya

$$kg(Y) = hook + jam + bulan$$

4. Mentransformasi data
5. Menguji apakah terdapat spatial autocorrelation (e.g., koordinat X-Y) atau temporal autocorrelation (e.g., bulan, tahun pemancingan) dan mengkoreksinya.

Additive model

Smoothing models memungkinkan adanya hubungan non-linear antara response variable dan explanatory variables (additive models).

Additive model

Regresi linear dengan satu variabel penjelas biasanya akan terlihat sebagai berikut:

$$Yi = \alpha + \beta \times Xi + \varepsilon i$$

Dimana hubungan antara

$$Yi$$

dan

$$Xi$$

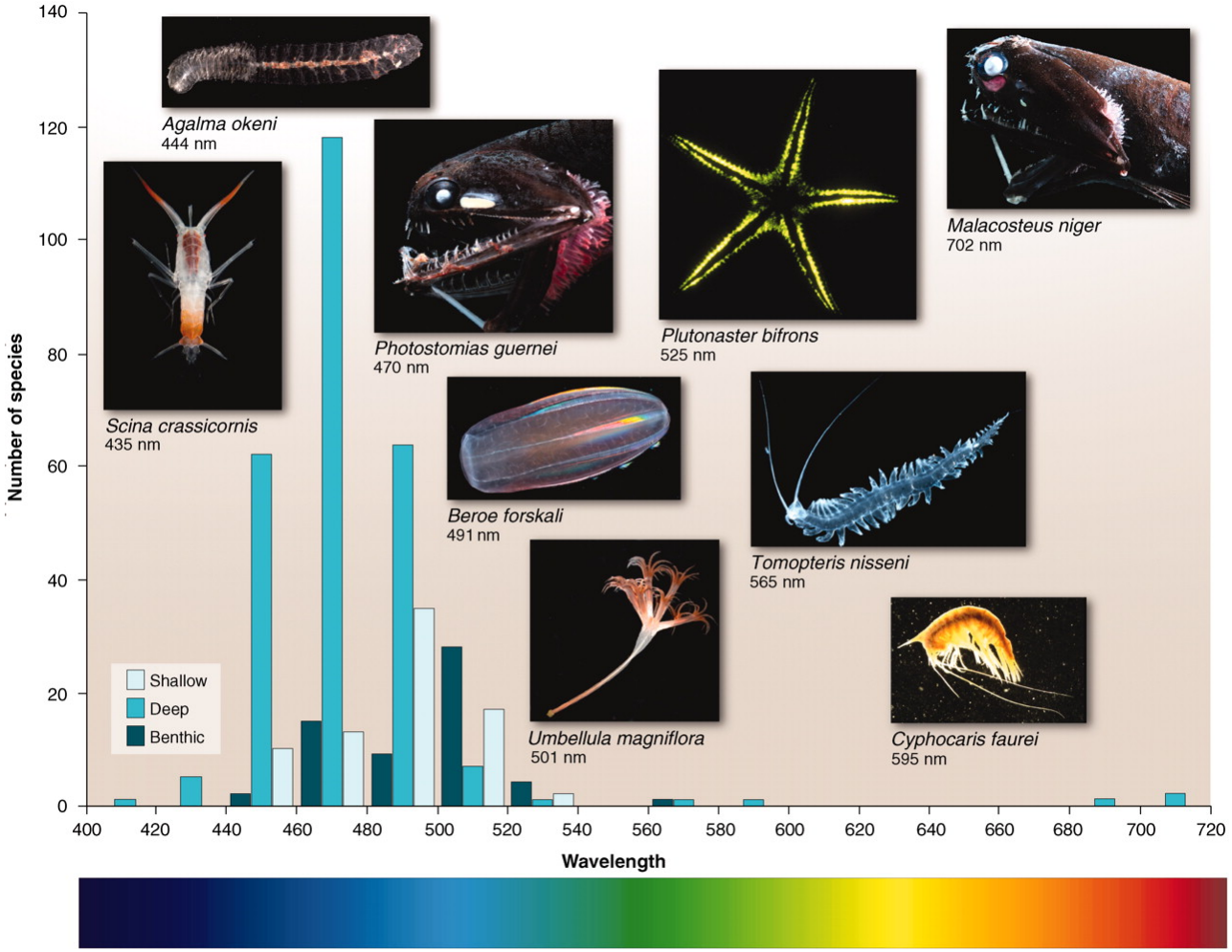
dijelaskan oleh parameter

$$\beta$$

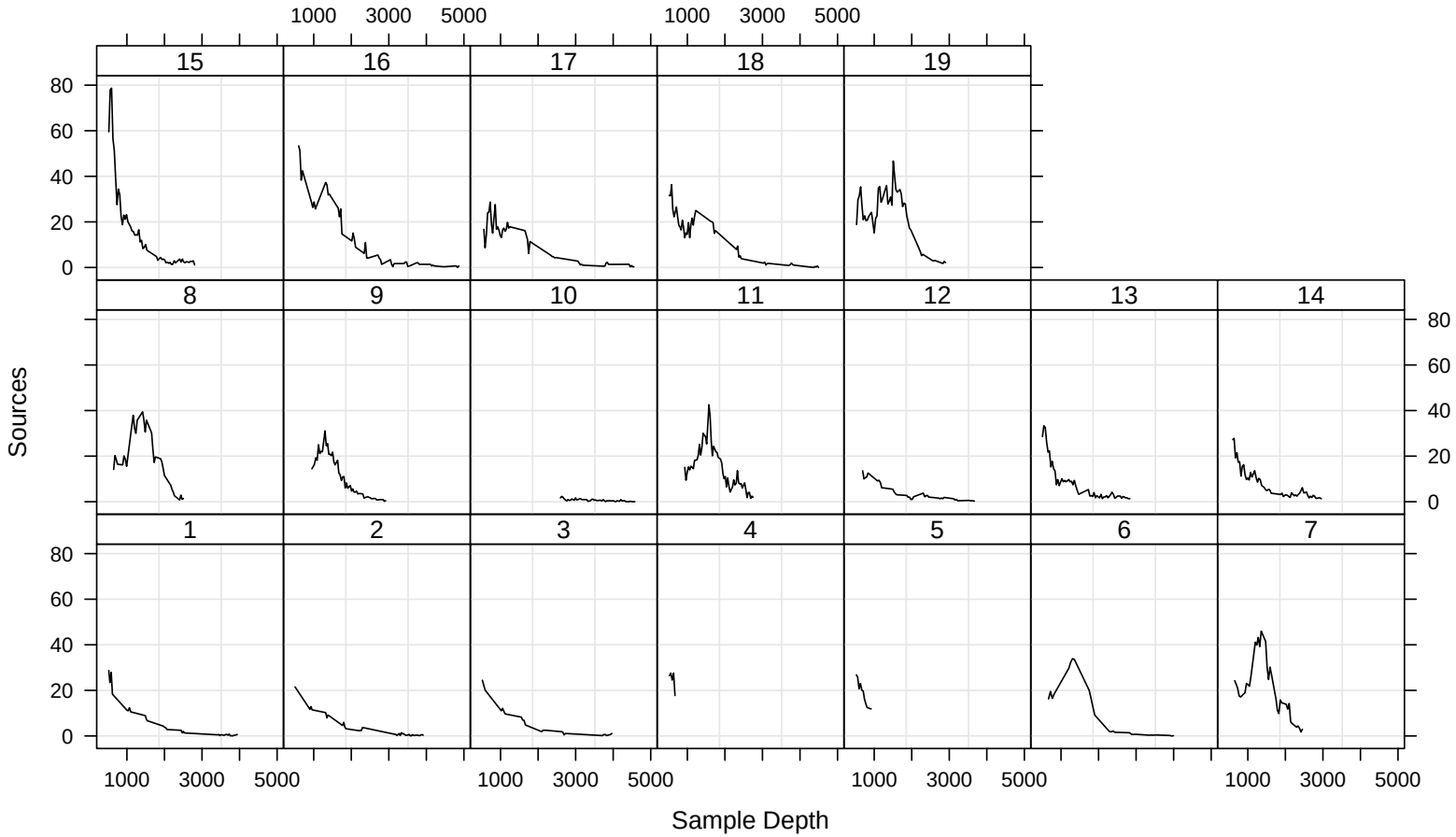
.

Additive model

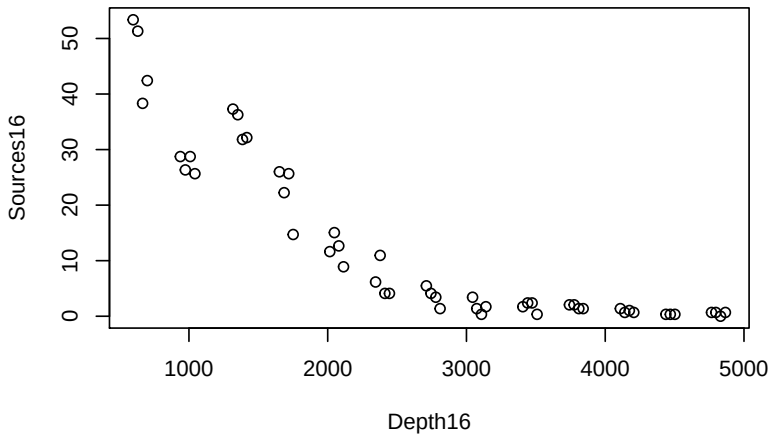
Kita coba lihat hubungan antara kelimpahan spesies bioluminescence pelagic sepanjang gradien kedalaman di timur laut Samudra Atlantik.



Additive model



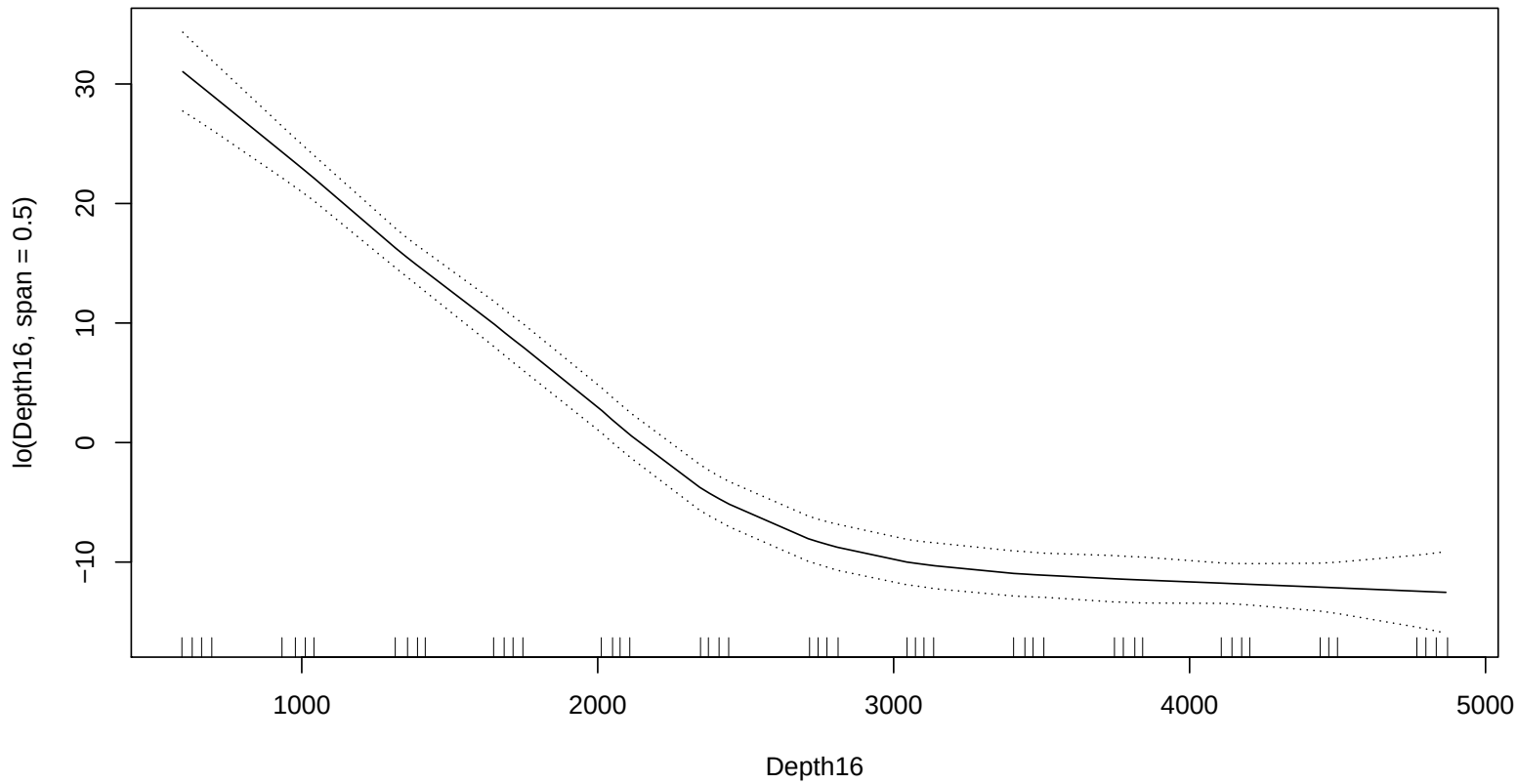
Additive model



Additive model

Additive model diaplikasikan dimana kita melihat perubahan kelimpahan Pelagic Bioluminescence (variabel *Yi*) terhadap kedalaman (variabel *Xi*)

```
library(gam)
M2<-gam(Sources16~lo(Depth16,span=0.5))
plot(M2,se=T)                    #Figure 3.1B
```

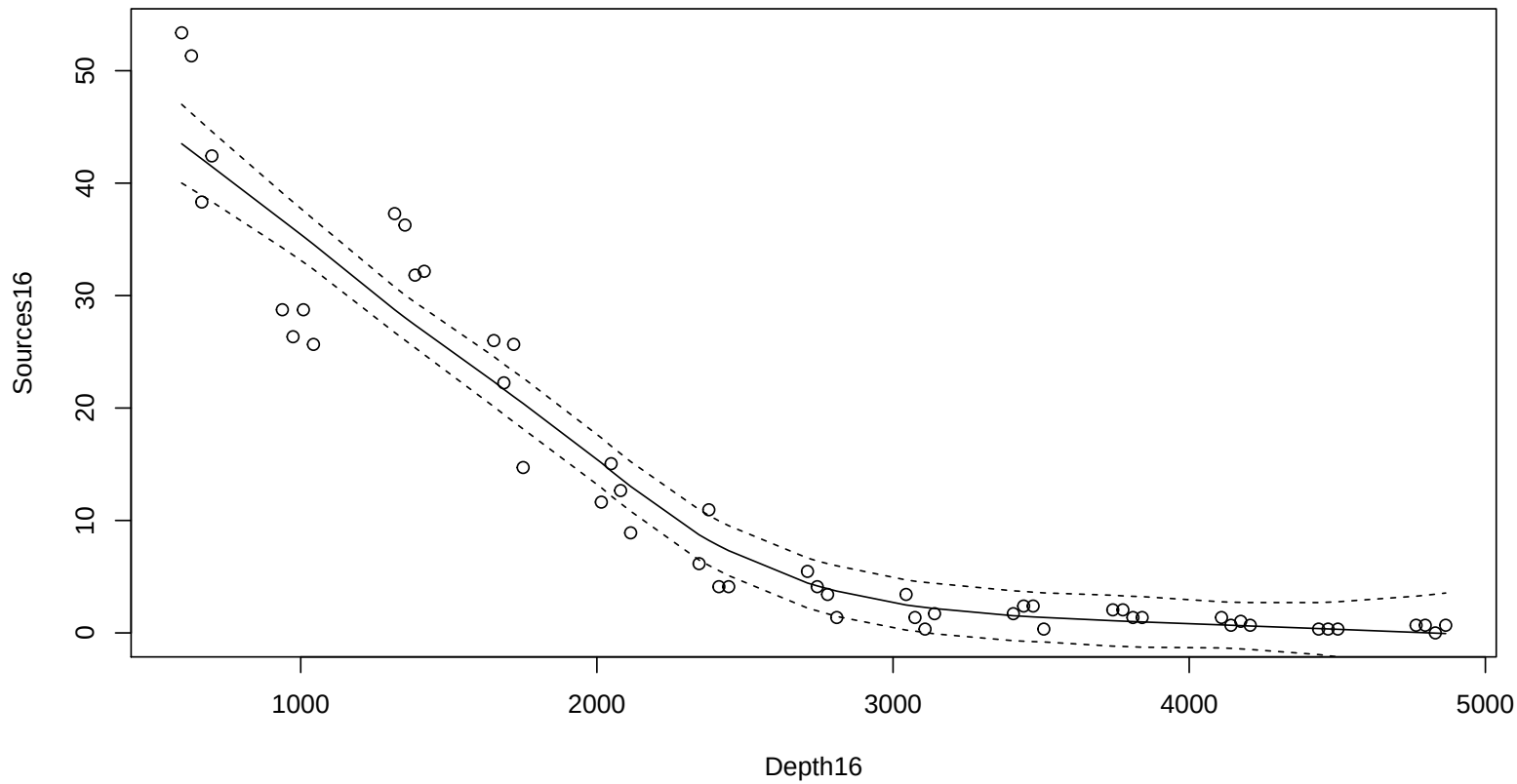


Additive model

Sekarang kita menggunakan perintah `predict` untuk menghitung nilai X dan Y (fitted values) berikut nilai standar errornya.

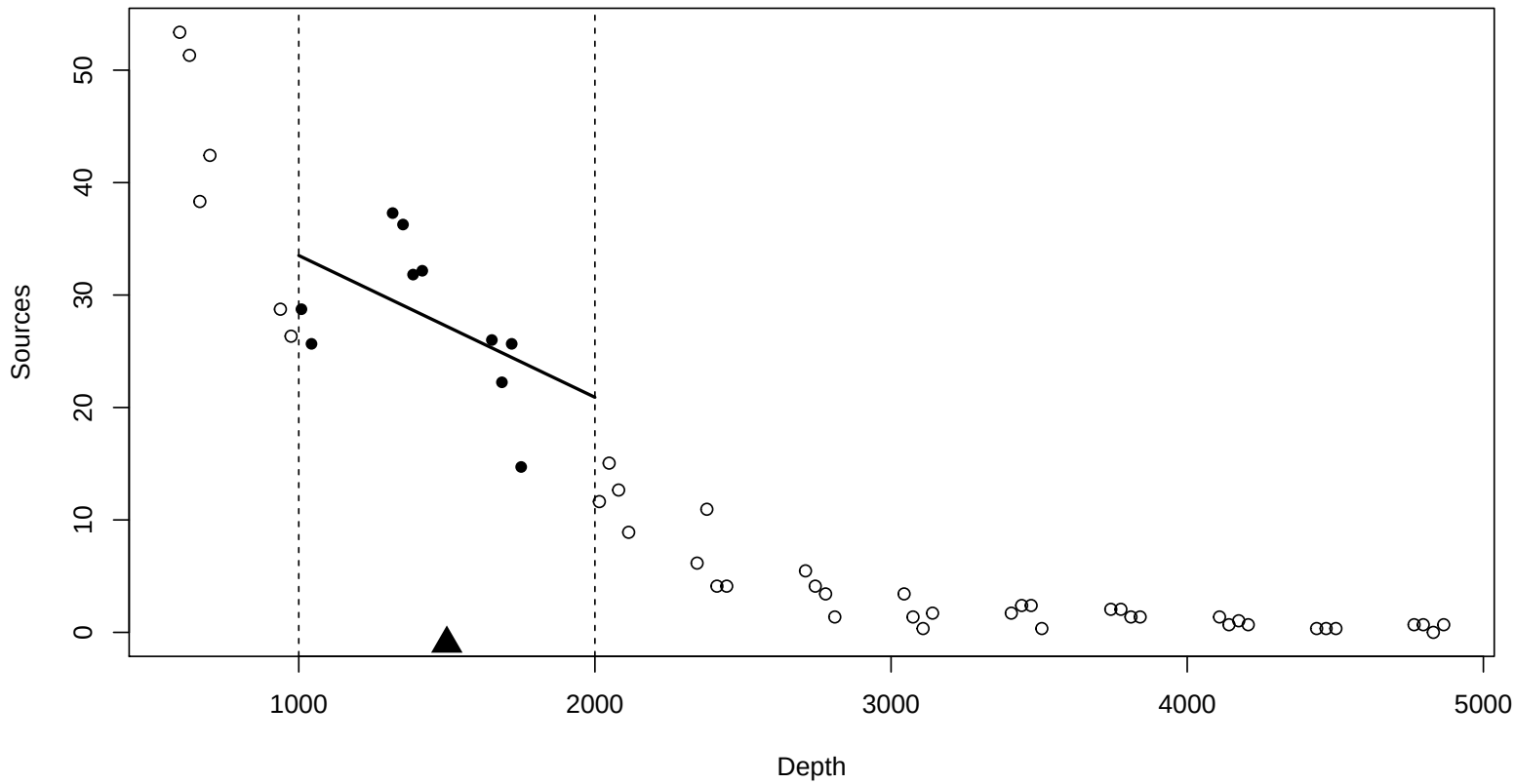
Additive model

```
#It may be better to predict along an equidistant gradient
P2 <- predict(M2, se = TRUE)
plot(Depth16, Sources16, type = "p")
I1 <- order(Depth16)
lines(Depth16[I1], P2$fit[I1], lty = 1)
lines(Depth16[I1], P2$fit[I1] + 2 * P2$se[I1], lty = 2)
lines(Depth16[I1], P2$fit[I1] - 2 * P2$se[I1], lty = 2)
```



Additive model

Untuk suatu permodelan dimana kita tertarik melihat rentang tertentu saja (e.g., nilai sekitar 1500 m), maka kita bisa menggunakan weighted linear regression, menghitung median dalam rentang, atau menggunakan linear regression hanya pada rentang tersebut.



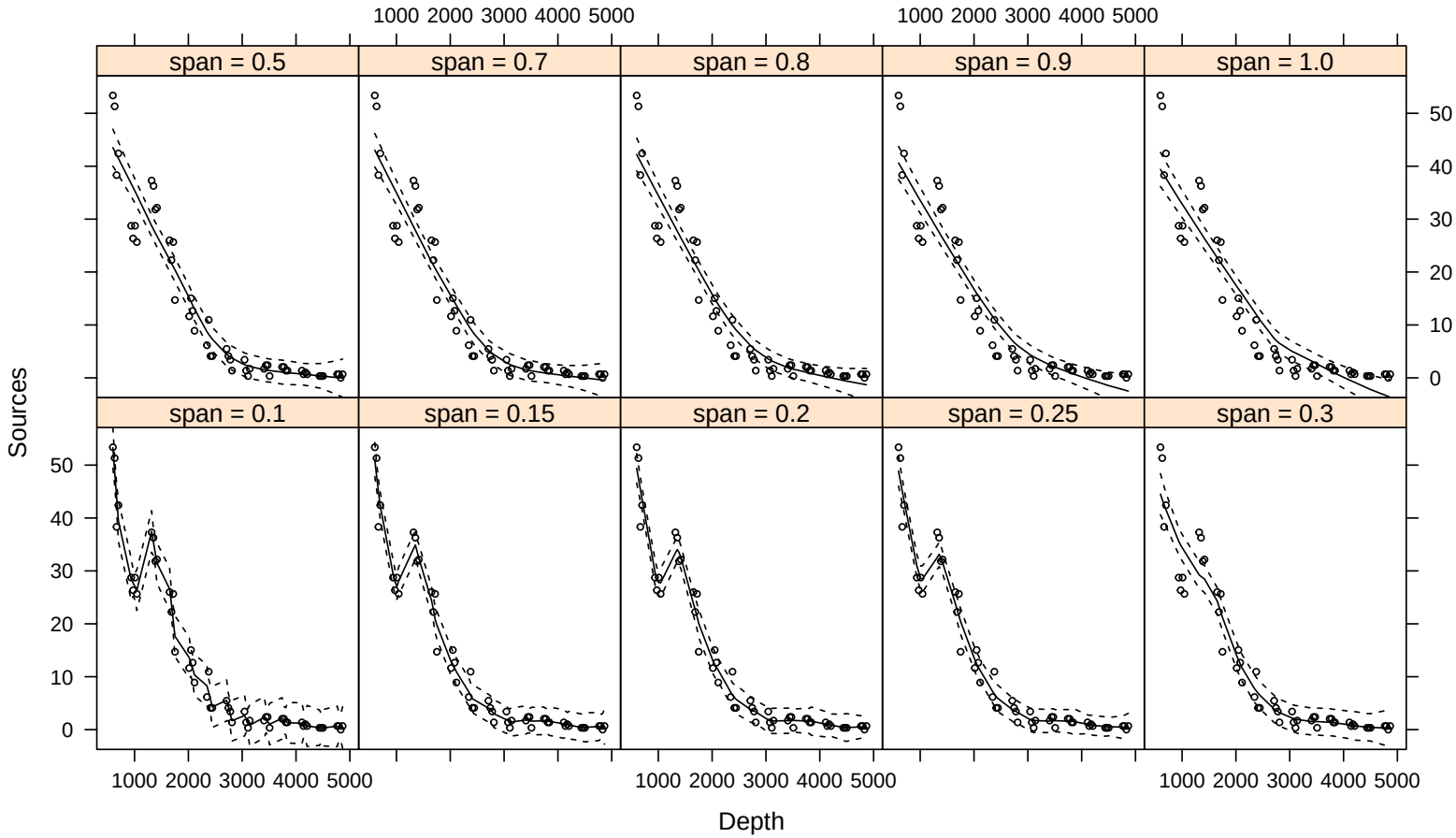
Additive model

Kita kemudian bisa meneruskan melakukan prediksi per bagian rentang tersebut. Ini dilakukan dengan pendekatan LOESS (local regression smoother) yang sudah kita coba di atas. Fungsi R yang bisa dipakai adalah `loess` atau `lo`, dimana kita mencoba mem-fitkan model polinomial pangkat 2.

Yang perlu kita perhatikan adalah bagaimana menentukan rentang yang tepat, apakah 1500 cukup, kurang, atau lebih.

`lo(Depth, span=0.5)` membantu kita menetapkan bahwa ukuran dari rentang adalah 50%dari data. Apabila nilainya 1, maka kita akan membuat sebuah regresi linear.

Additive model

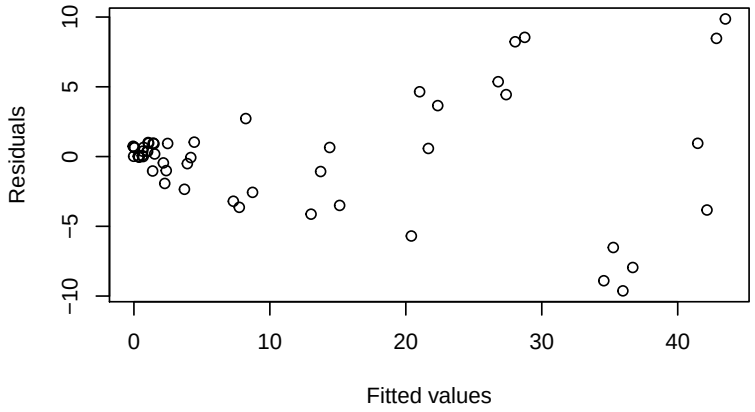


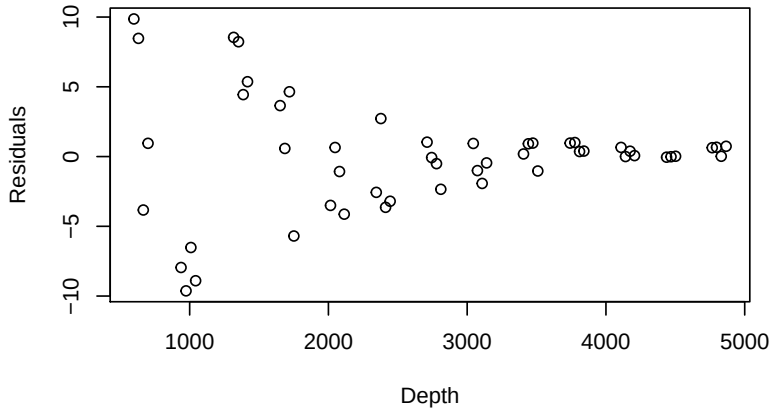
Additive model

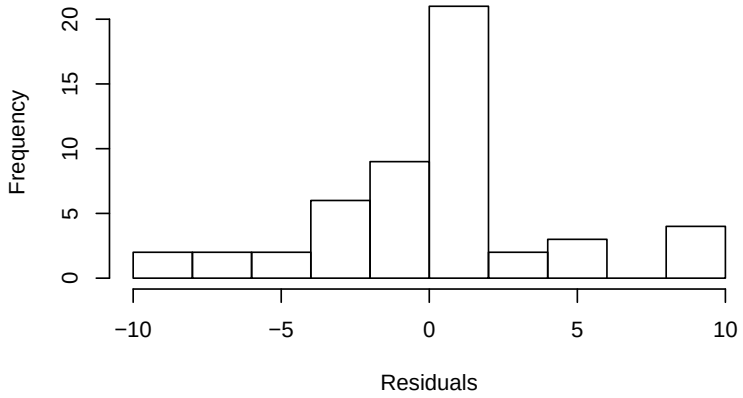
```
##      df      AIC
## M1    3 247.6038
## M2    3 244.8716
## M3    3 247.4270
## M4    3 249.0272
## M5    3 288.6282
## M6    3 296.3008
## M7    3 296.7522
## M8    3 299.0385
## M9    3 306.8628
## M10   3 314.3067
```

Additive model

Tapi, apakah residualnya sudah aman?

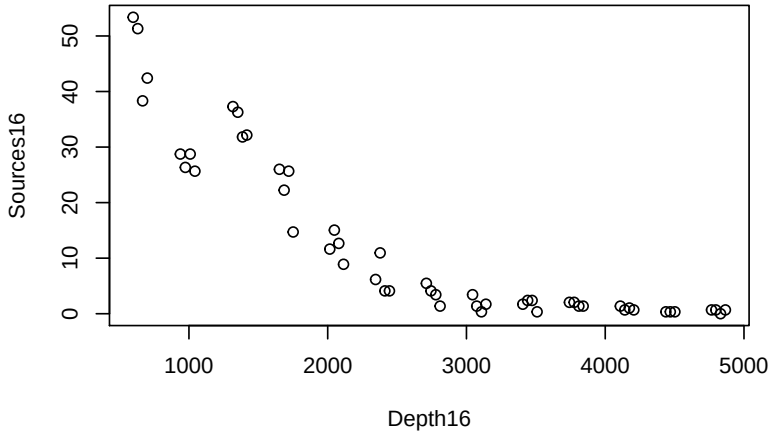


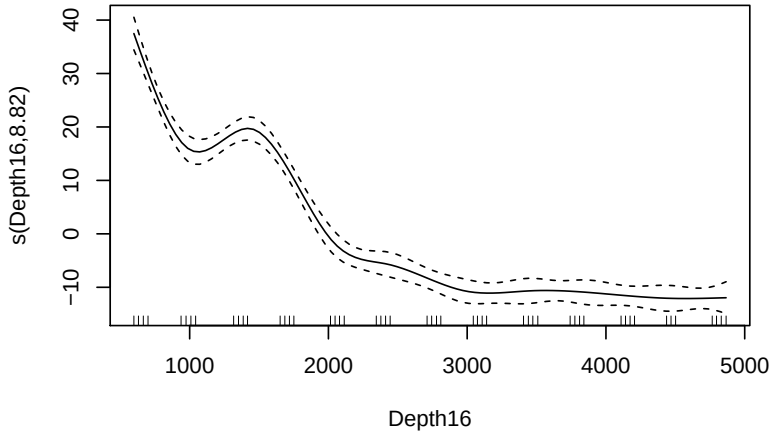


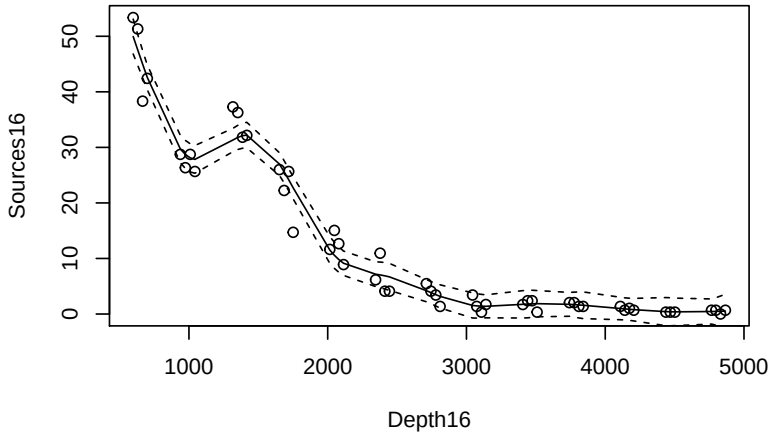


Additive model

Kalau belum kita bisa melakukan hal yang sama menggunakan package `mgcv`.

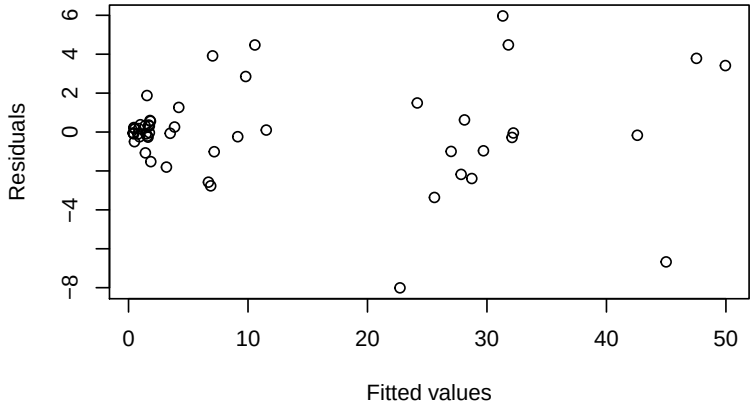


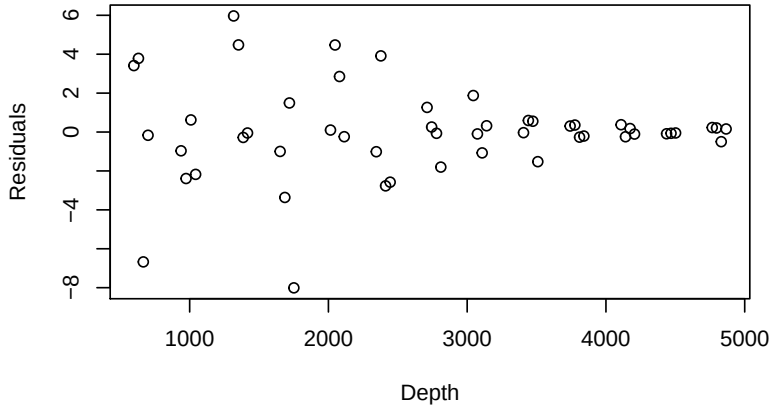


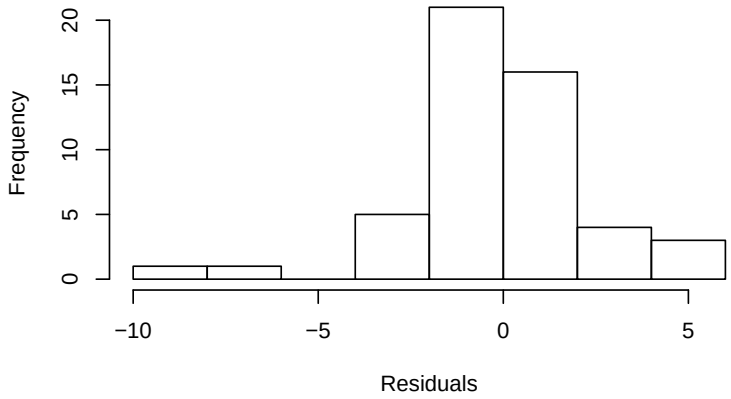


Additive model

Dimana sekarang residualnya:







Additive model

Apa yang berbeda?

Additive model

Kali ini, setelah membagi kedalaman ke beberapa rentang tertentu, kita memfitkan model dengan bentuk `cubic polynomial`, yakni $Yi = \alpha + \beta i \times Xi + \beta_2 \times Xi^2 + \beta_3 \times Xi^3$ dimana kita menentukan jumlah smoothing yang tidak fixed.

Generalized least square models

- Linear regression model biasanya terdiri atas bagian:

$$Y = \textit{fixed.part} + \textit{random.part}$$

$$\alpha + \beta_1 + .. + \beta_q X_q$$

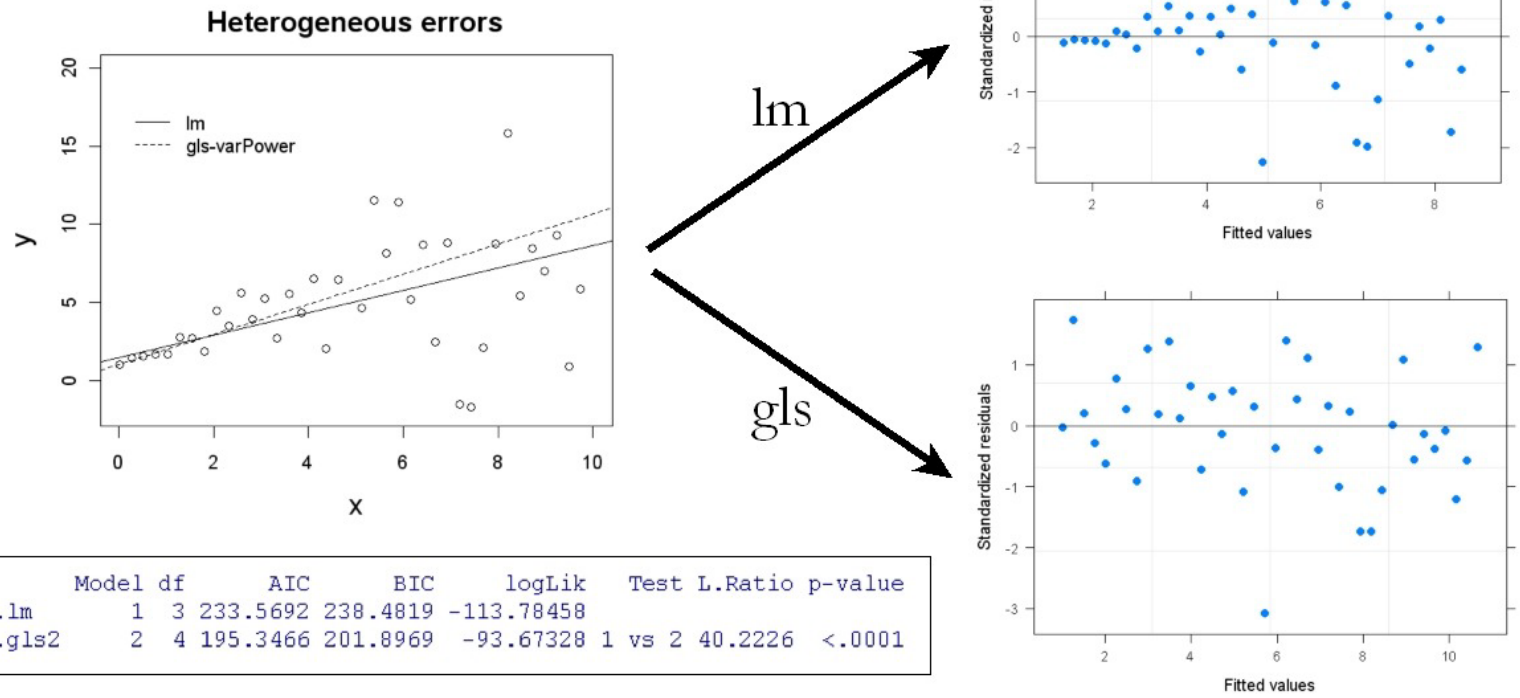
```
Heterogeneity
Nested data
Temporal correlation
Spatial correlation
Random noise
```

Generalized least square models

- Random part= real random term; linear regression/additive modelling.
- Random part= nested data; mixed effects model.
- Random part= heterogeneity; generalised least squares/weighted linear regression.
- Random part= violation of independence; spatial and or temporal autocorrelation.

Generalized least square models

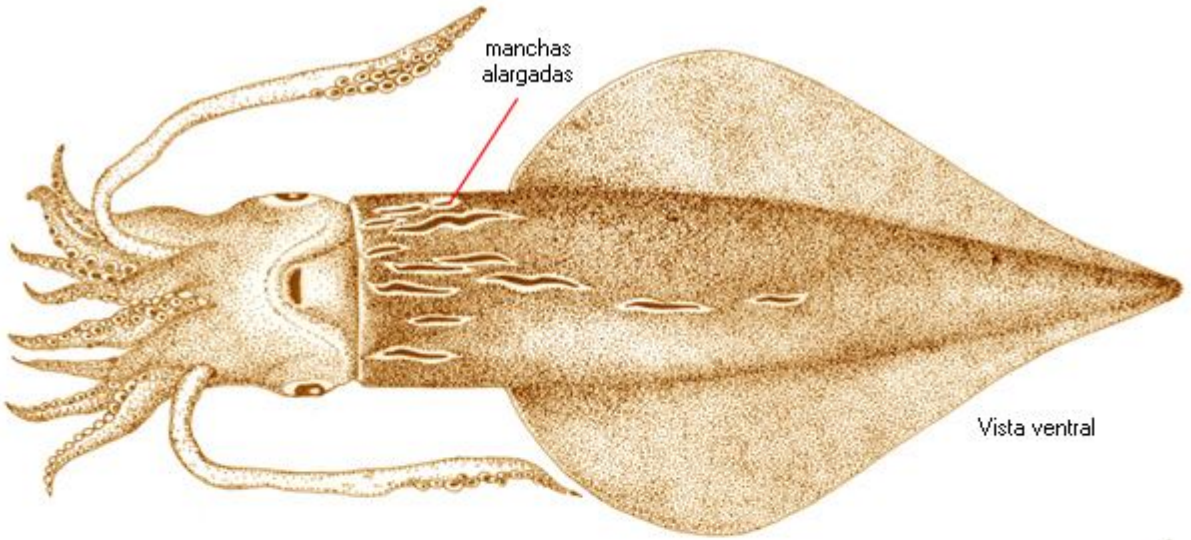
Linear model versus generalized least squares model:

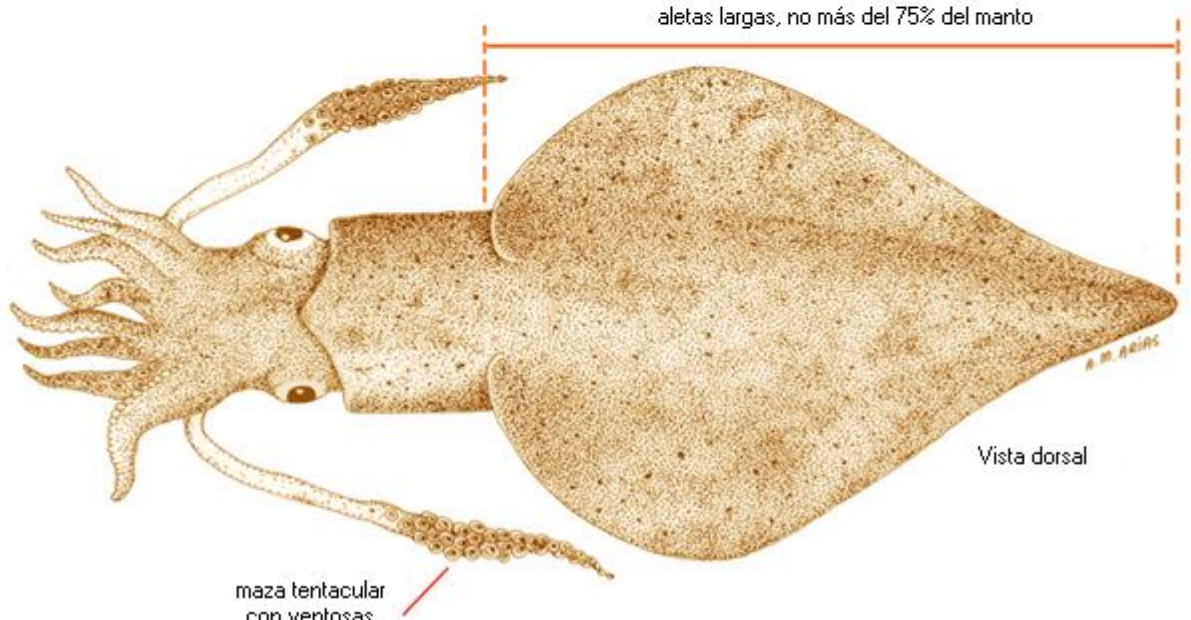


	Model	df	AIC	BIC	logLik	Test	L.Ratio	p-value
fit.lm	1	3	233.5692	238.4819	-113.78458			
fit.gls2	2	4	195.3466	201.8969	-93.67328	1 vs 2	40.2226	<.0001

Figure 3: Generalized least square vs simple linear regression

Generalized least square models: heterogeneity





Generalized least square models

```
Squid <- read.delim("Squid.txt", header = TRUE)
names(Squid)
```

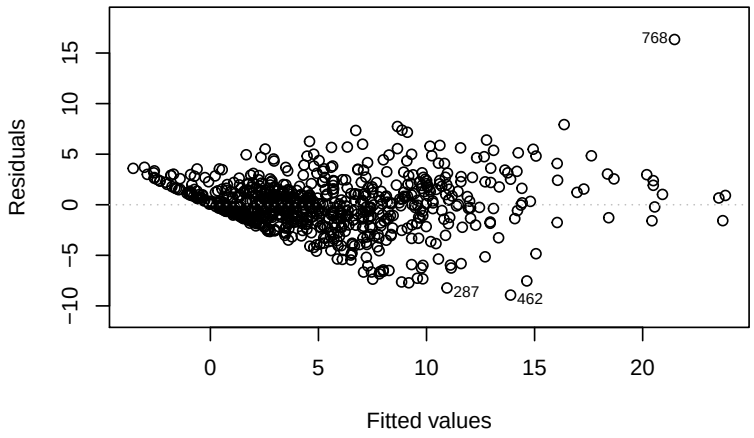
```
## [1] "Specimen"     "YEAR"         "MONTH"        "DML"
## [5] "Testisweight"
```

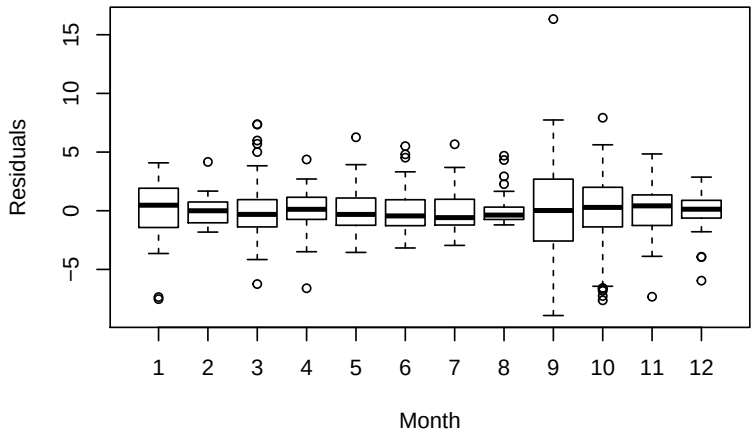
DML = dorsal mantel length

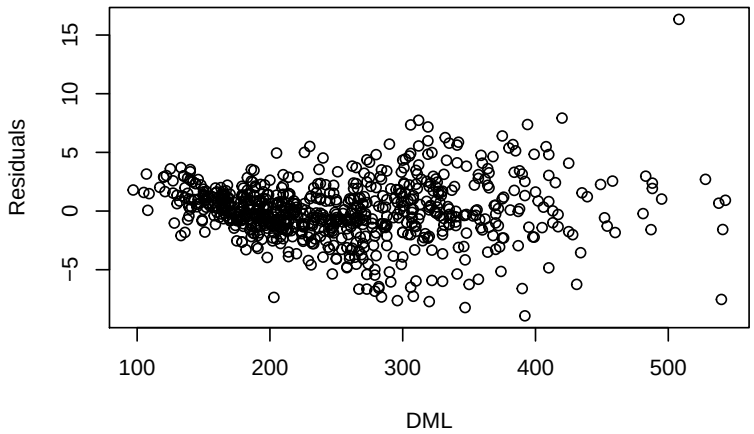
Generalized least square models: heterogeneity

$$Testisweighti = intercept + DML_i + Month_i : Month_i + Residuals_i$$

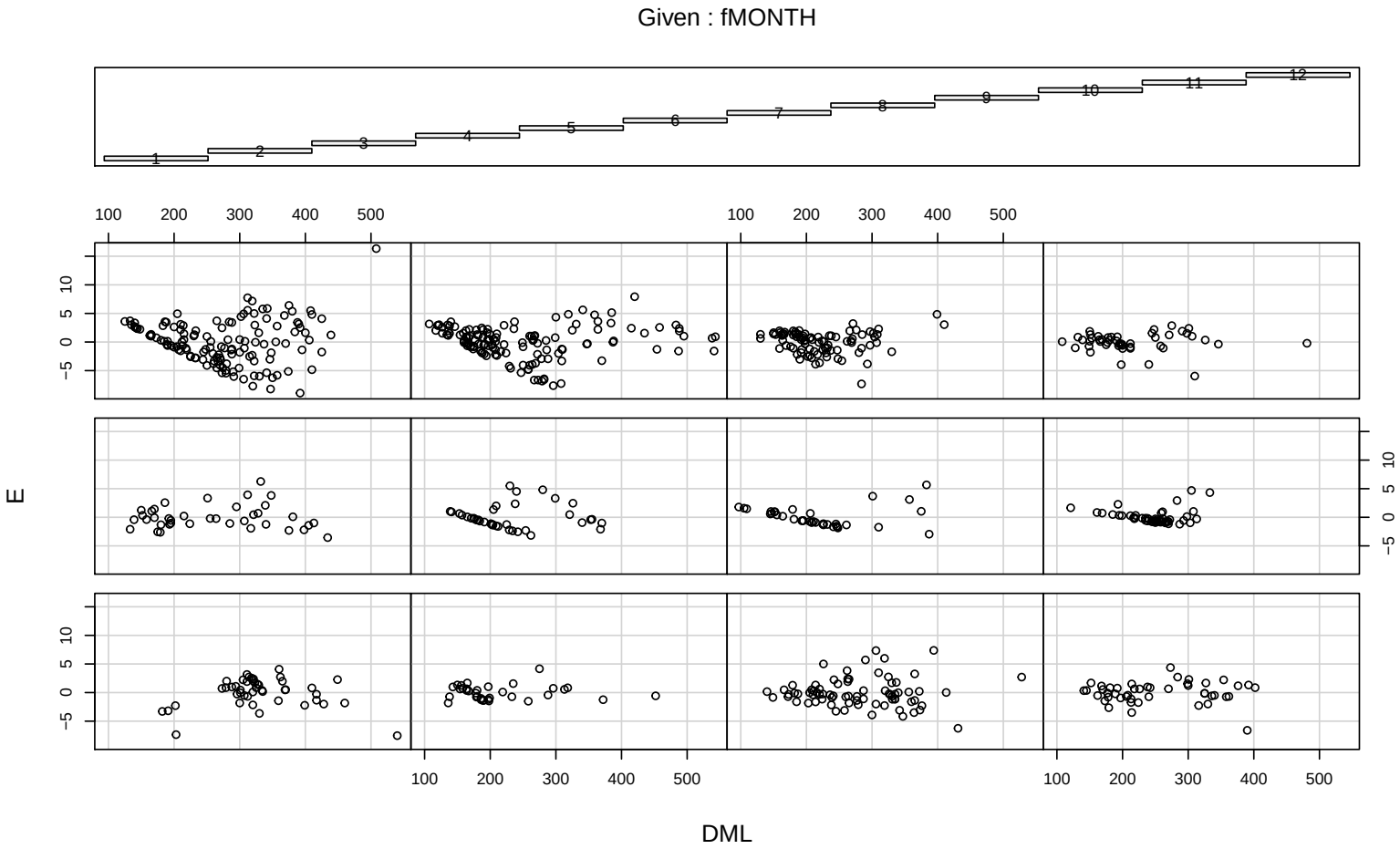
Generalized least square models: heterogeneity







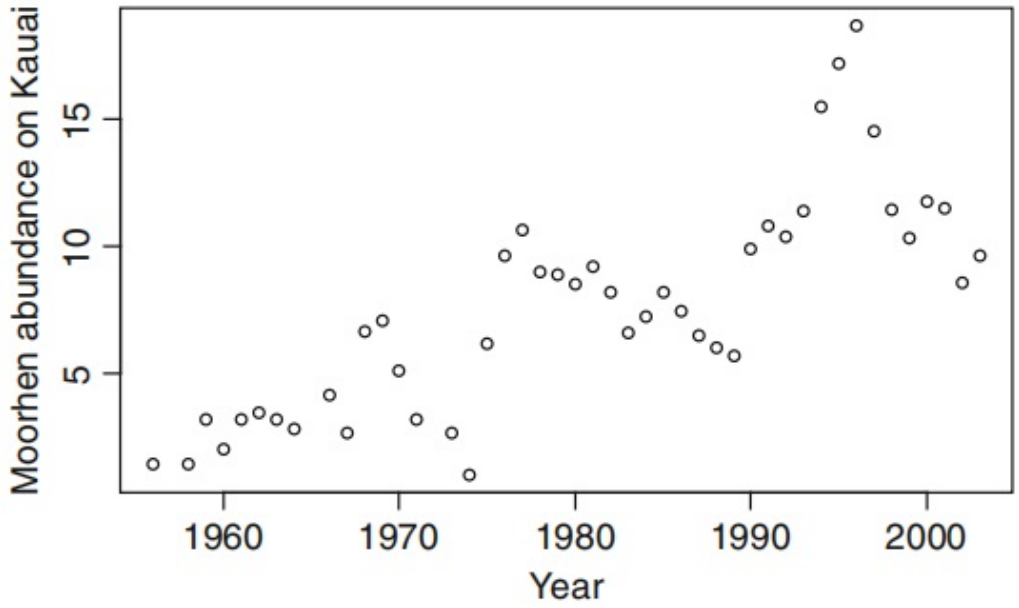
Generalized least square models: heterogeneity



Generalized least square models: Temporal autocorrelation

Generalized least square models: Temporal autocorrelation

Fig. 6.1 Time series plot of square-root-transformed moorhen abundance measured on the island of Kauai



Generalized least square models: Temporal autocorrelation

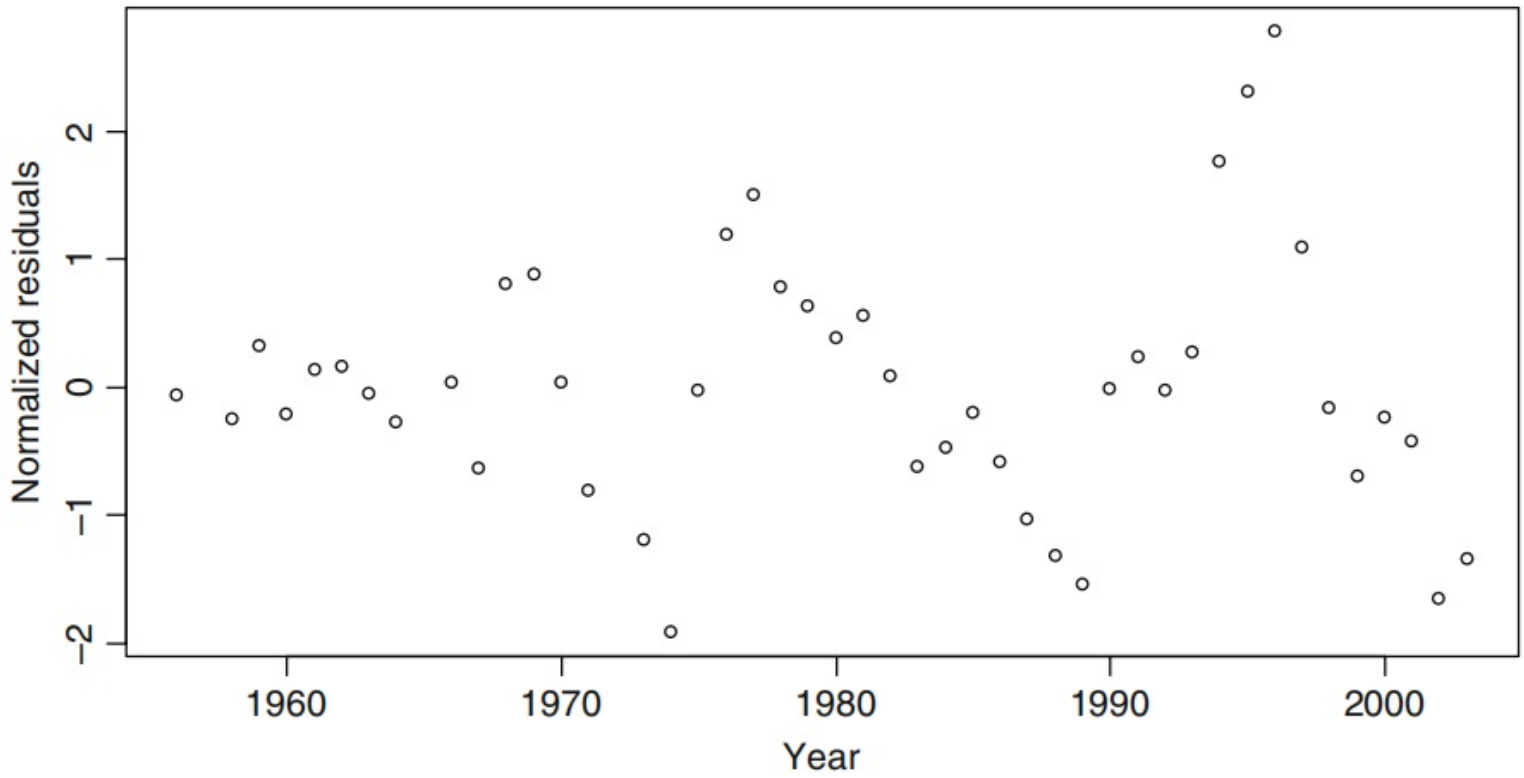


Fig. 6.2 Normalised residuals plotted versus time. Note the pattern in the residuals

Generalized least square models:: Temporal autocorrelation

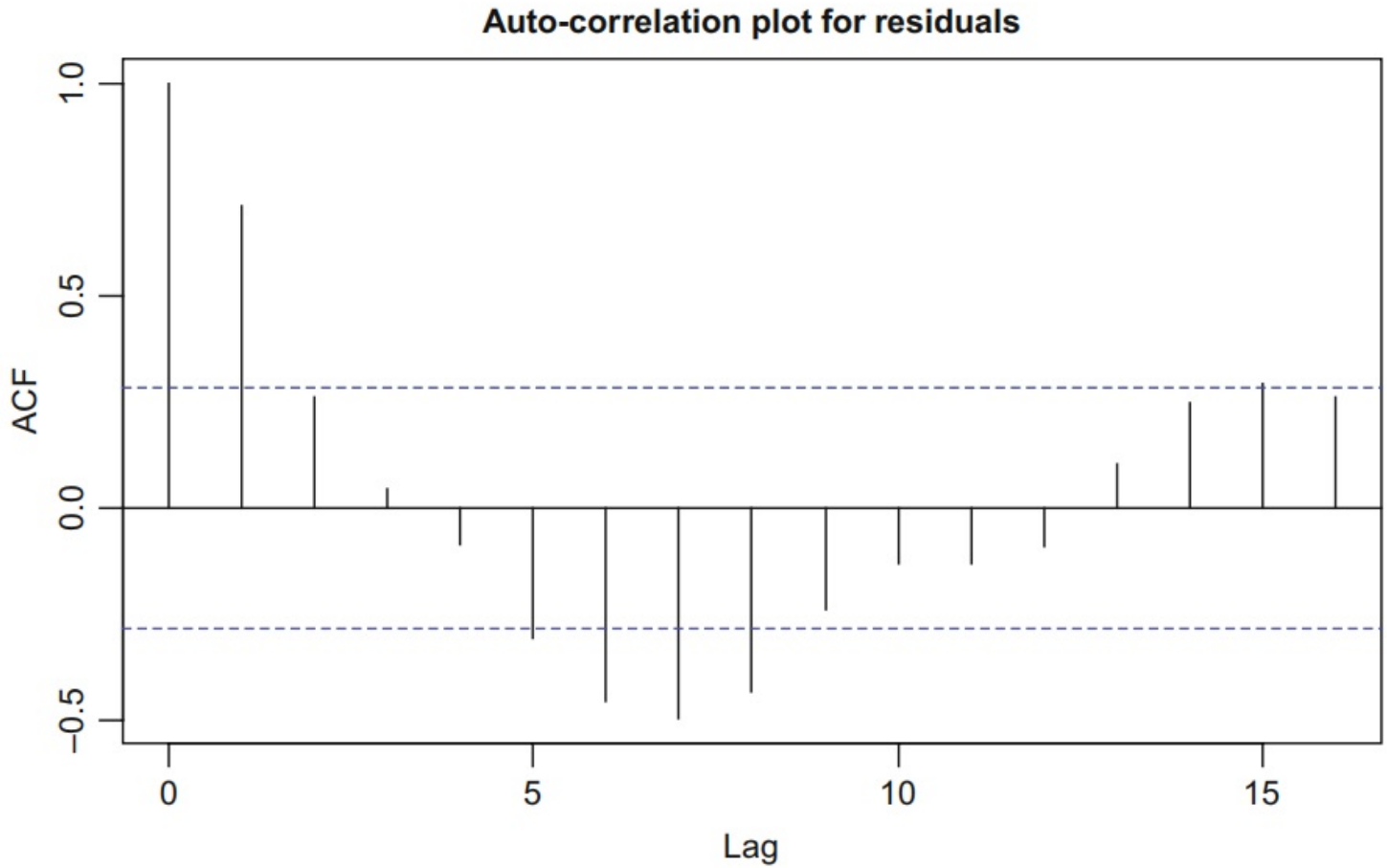


Fig. 6.3 Auto-correlation plot for the residuals obtained by applying linear regression on the Bird time series. Note that there is a clear indication of violation of independence

Generalized least square models: Spatial autocorrelation

What influences a boreal forster index (Tatarstan, Russia)? : relief, soil, climatic factors?

```
Boreality <- read.delim("Boreality.txt", header = TRUE)
names(Boreality)
```

```
##  [1] "point"  "x"      "y"      "Oxalis" "boreal" "nBor"   "nTot"
##  [8] "Grn"    "NDVI"   "T61"    "Wet"
```

Generalized least square models: Spatial autocorrelation

```
Boreality$Bor<-sqrt(1000*(Boreality$nBor+1)/(Boreality$nTot))
B.lm<-lm(Bor~Wet,data=Boreality)
summary(B.lm)
```

```
##
## Call:
## lm(formula = Bor ~ Wet, data = Boreality)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.0918 -2.4214 -0.1382  1.9783 17.6054
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  18.4880     0.3787   48.82   <2e-16 ***
## Wet         165.8036    10.5991   15.64   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.491 on 531 degrees of freedom
## Multiple R-squared:  0.3155, Adjusted R-squared:  0.3142
## F-statistic: 244.7 on 1 and 531 DF,  p-value: < 2.2e-16
```

Generalized least square models: Spatial autocorrelation

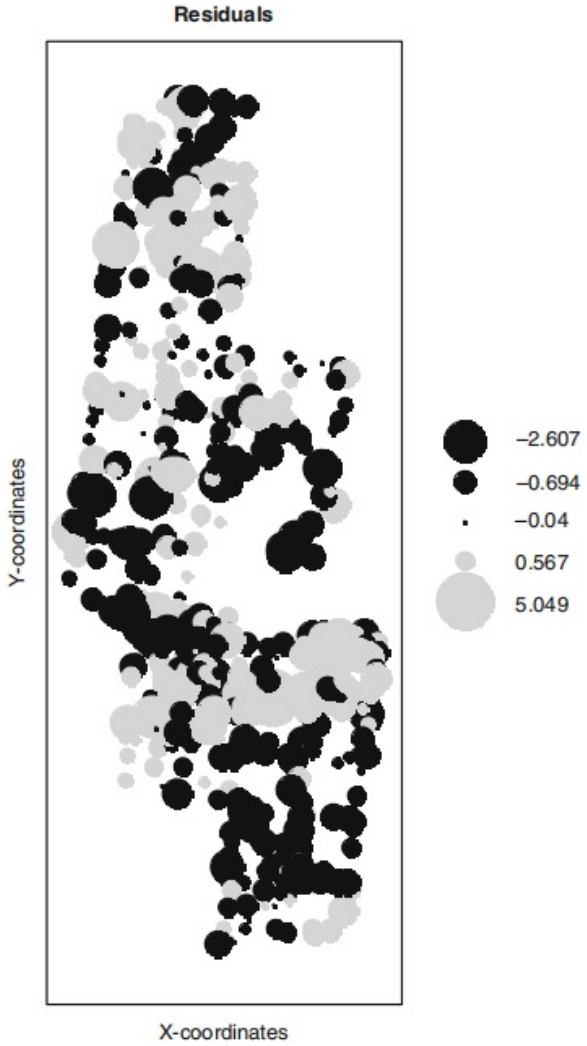


Fig. 7.1 Standardised residuals obtained by the linear regression model plotted versus their spatial coordinates. Black dotes are negative residuals, and grey dots are positive residuals

Generalized least square models: Spatial autocorrelation

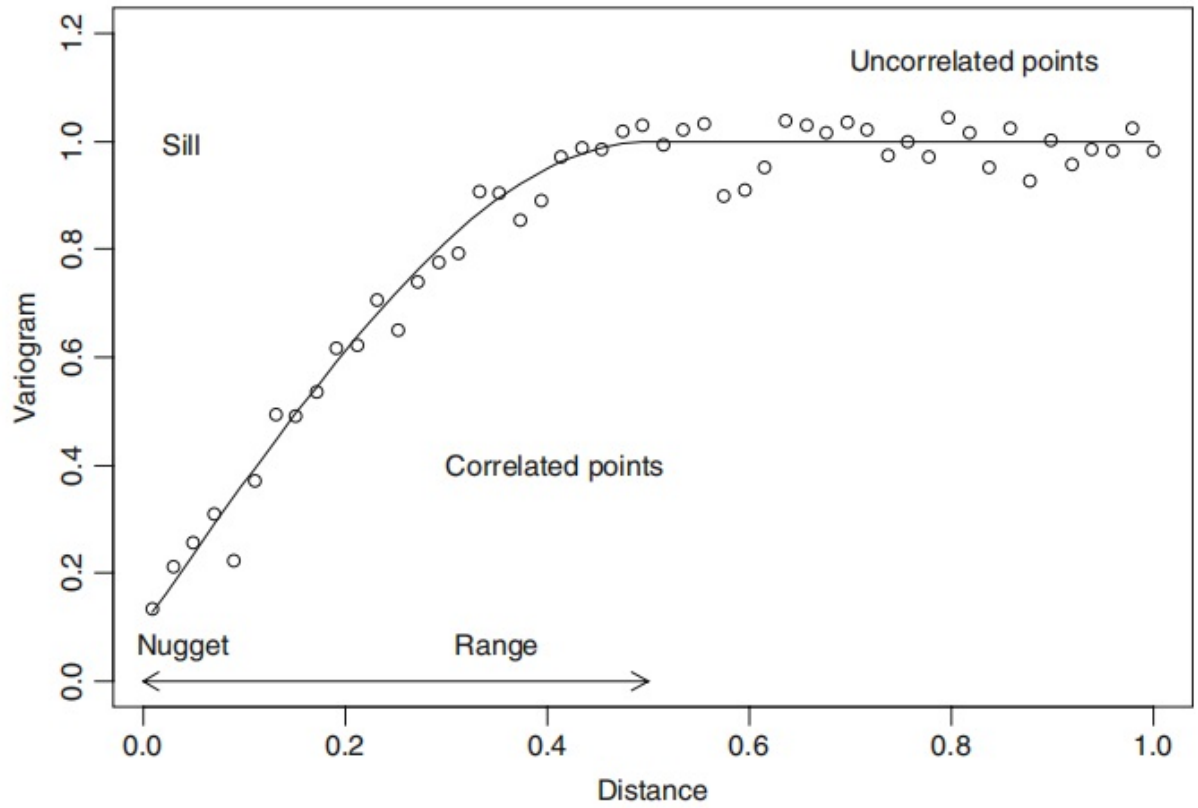
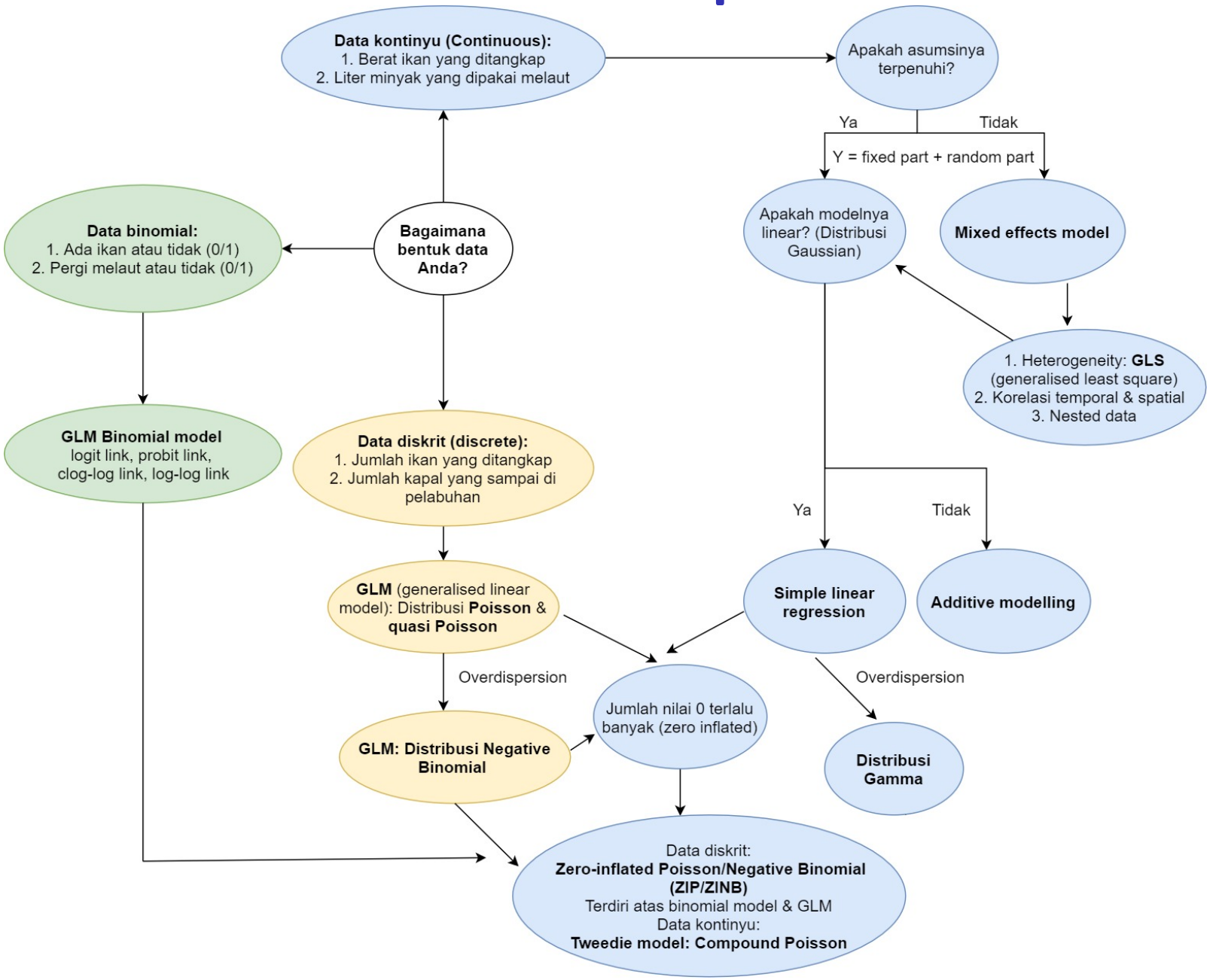


Fig. 7.2 Variogram with fitted line. The sill is the asymptotic value and the range is the distance where this value occurs. Pairs of points that have a distance larger than the range are uncorrelated. The nugget effect occurs if $\hat{\gamma}(\mathbf{h})$ is far from 0 for small h

Alur permodelan



References:

Zuur, et al. 2009. Sumber utama konten modul ini.