Assumptions of linear regression & Linear regression

Science unit WCS Indonesia

5-9th of August 2019

Quiz 1

- Pilihlah dari opsi berikut, distribusi mana yang bisa mengantisipasi kondisi *over dispersion* variance dari hasil permodelan?
  - Gaussian
  - Poisson
  - Negative binomial

Jawaban quiz 1

- Pilihlah dari opsi berikut, distribusi mana yang bisa mengantisipasi kondisi *over dispersion* variance dari hasil permodelan?
  - Gaussian
  - Poisson
  - **Negative binomial**

Quiz 2

- Pendekatan apa yang bisa mendeskripsikan berapa banyak variansi yang terjelaskan dari hasil permodelan?
  - AIC
  - multicollinearity
  - R-squared

Jawaban quiz 2

- Pendekatan apa yang bisa mendeskripsikan berapa banyak variansi yang terjelaskan dari hasil permodelan?
  - AIC
  - multicollinearity
  - ***R-squared***

Quiz 3

- Manakah dari distribusi berikut yang bisa menggunakan data bernilai negatif?
  - Gaussian
  - Negative binomial
  - Gamma

Jawaban quiz 3

- Manakah dari distribusi berikut yang bisa menggunakan data bernilai negatif?
  - ***Gaussian***
  - Negative binomial
  - Gamma

Overview alur permodelan

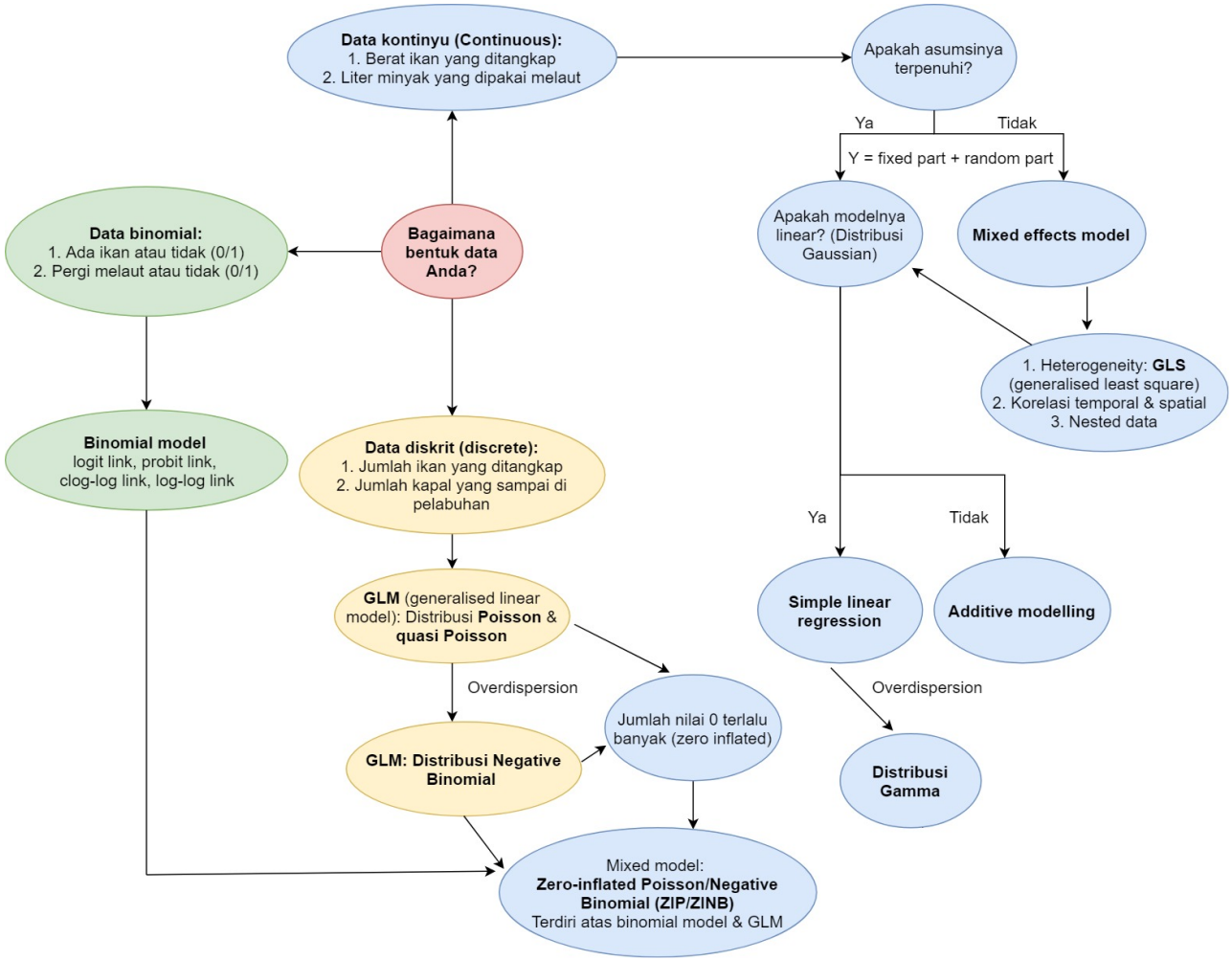


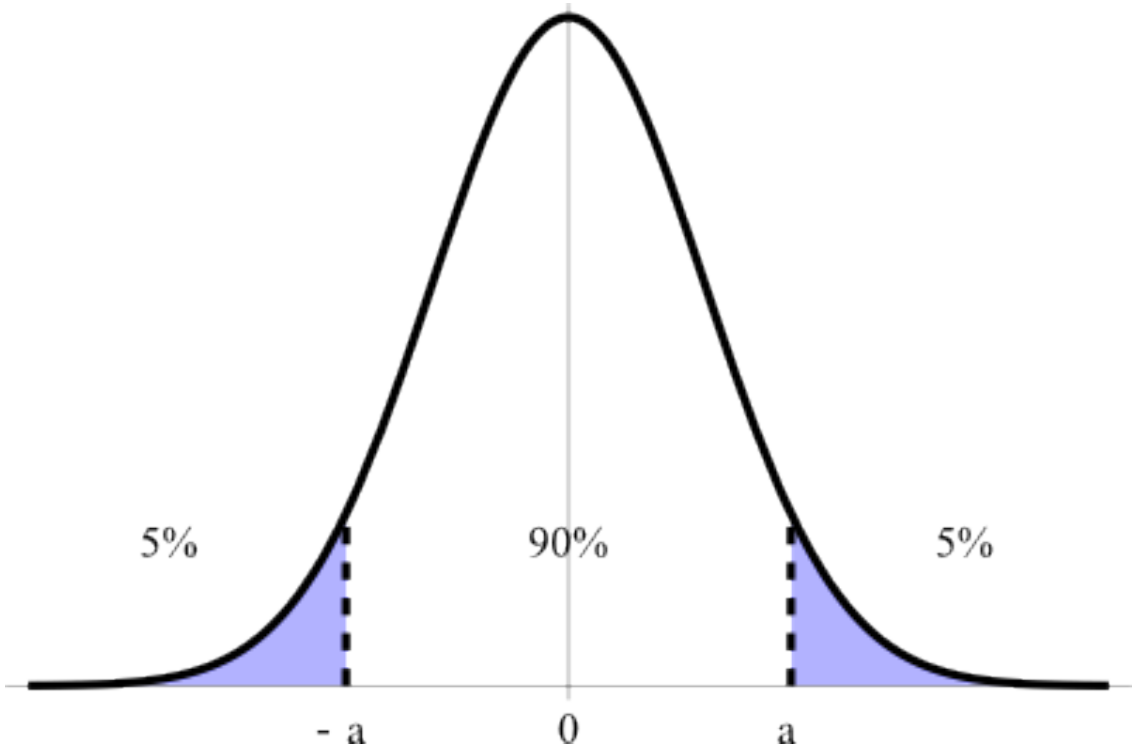
Figure 1: Overview alur permodelan

Dataset yang digunakan

- Data hasil tangkapan *Cephalopholis boenak* di Maluku Utara pada tahun 2017-2018.
- Terdiri atas variabel:
  - Tahun (2017, 2018)
  - Bulan
  - Nama kapal (Boat)
  - Jumlah kru kapal (Crew)
  - Panjang kapal (L)
  - Jumlah hook (Hook)
  - Jumlah jam (Hour)
  - Jumlah hari melaut (Days at Sea)
  - Lokasi
  - Perempat tahun (Quarter)
  - Jumlah individu hasil tangkap (IND)
  - Berat hasil tangkap (KG)
  - Epinephelidae

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)

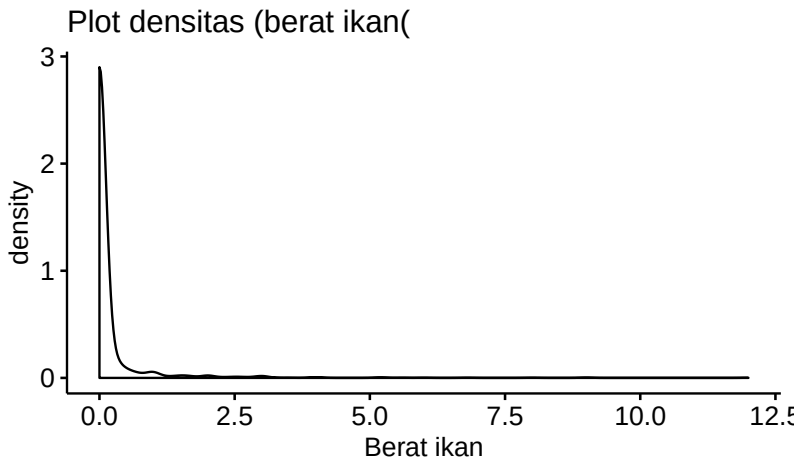
1. Normality = tes formal/histogram
2. Homogeneity = residual data tidak berpola
3. Fixed X = tidak memiliki pengetahuan apriori
4. Independence = tidak tergantung pada variabel lain/kondisi spasial atau temporal

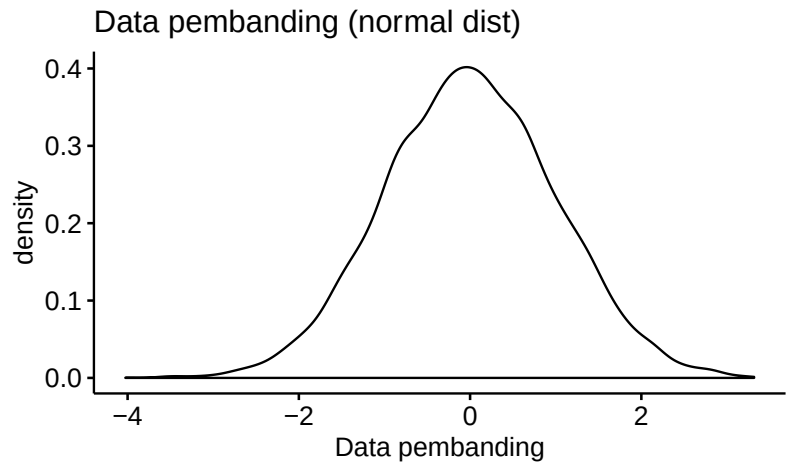
1. Uji normalitas

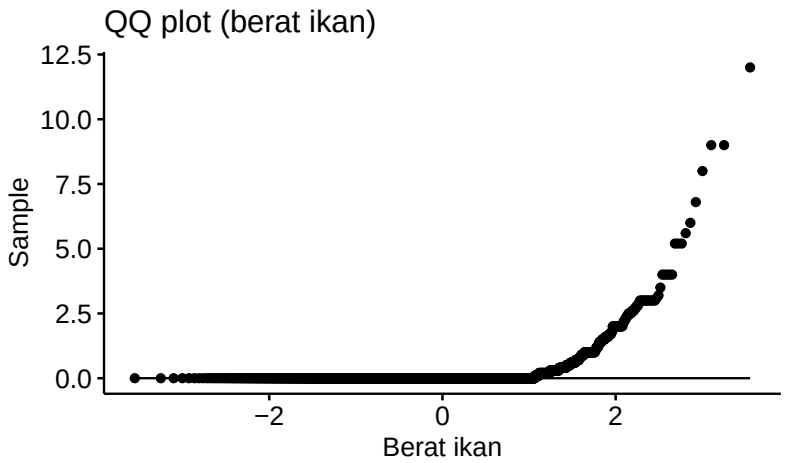
- Central limit theorem: bila sampel kita cukup besar (n>30), uji normal bisa diabaikan.
- Contoh menggunakan data *C. boenak*.
- Apakah yang mempengaruhi jumlah ikan *C. boenak* (kg) yang ditangkap nelayan?

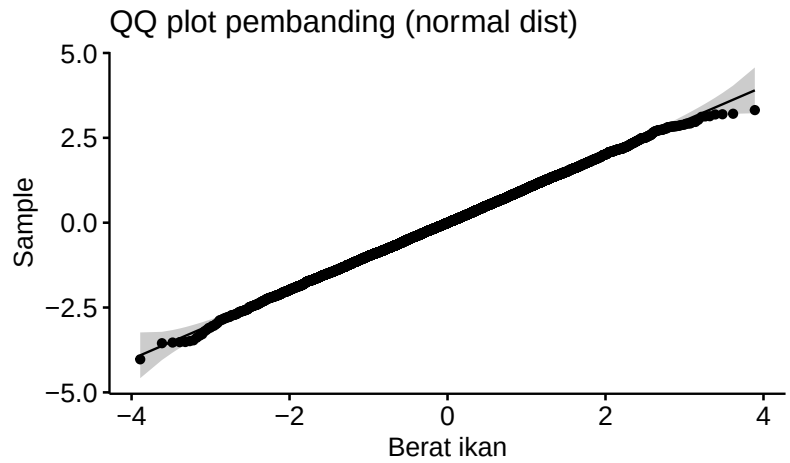
Central limit theorem = variabel acak yang jumlahnya banyak akan menghasilkan distribusi normal

Plot densitas & QQ-plot untuk berat ikan

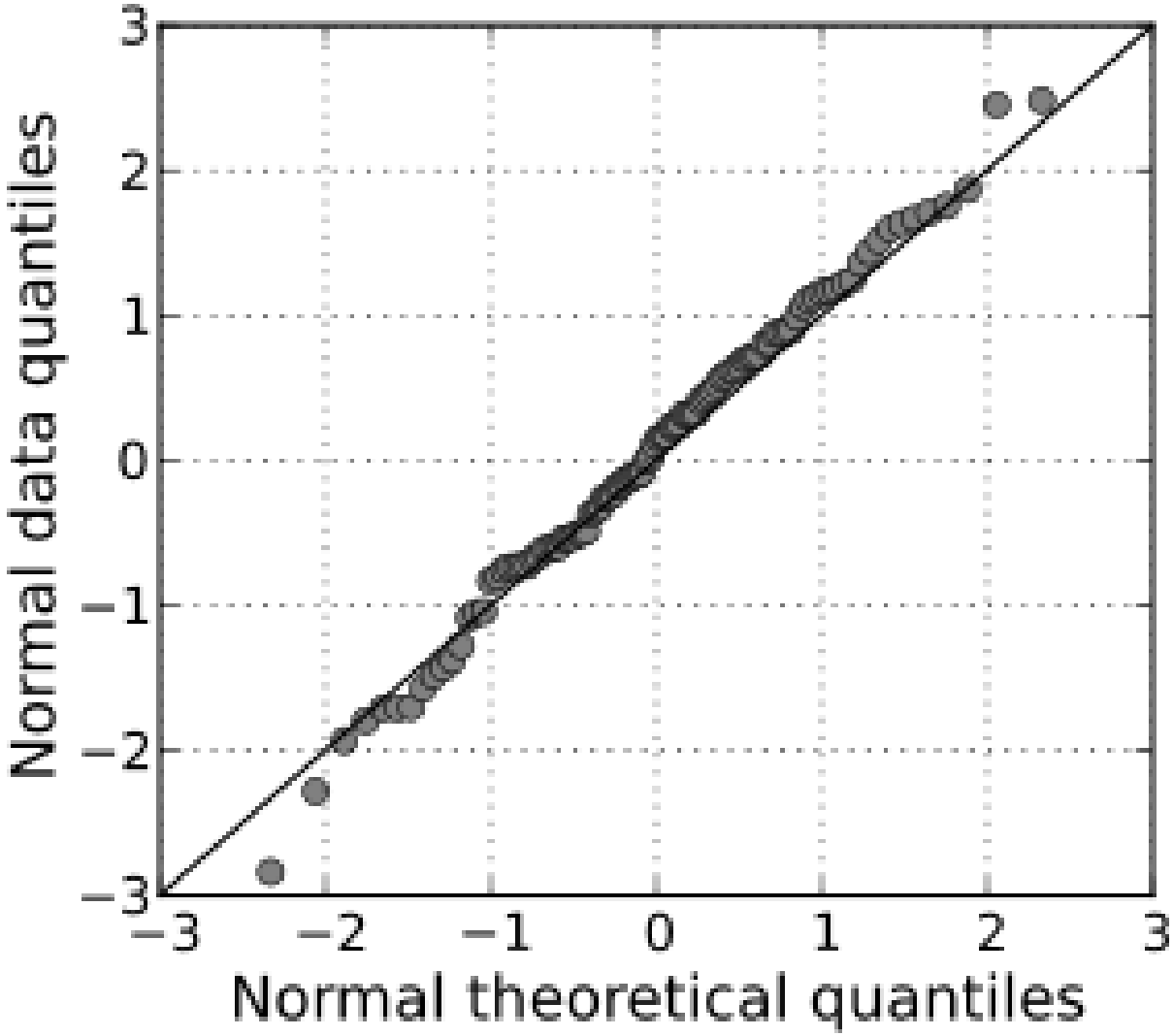








Contoh Q-Q plot dengan distribusi normal:



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

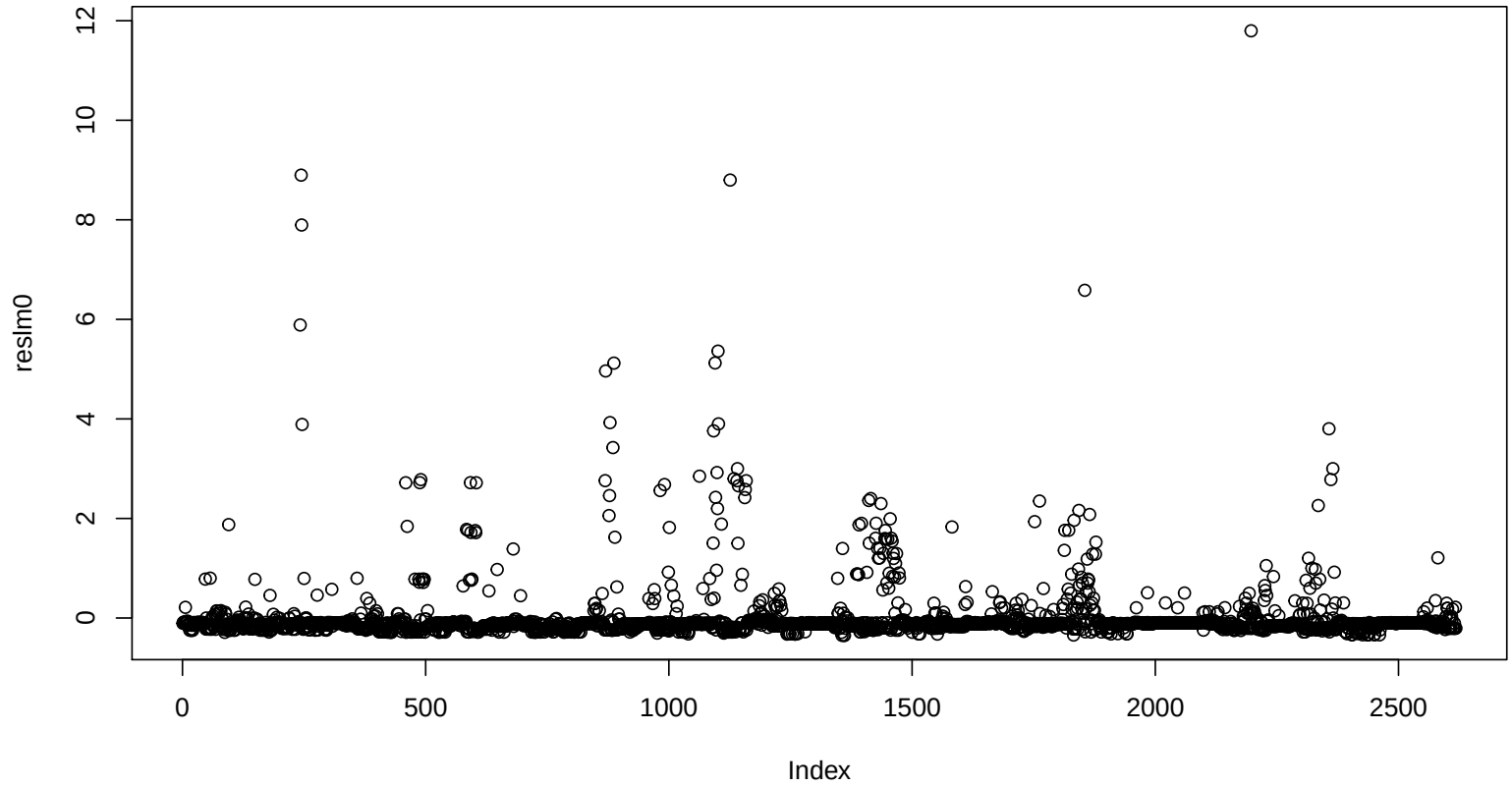
Contoh observasi visual homogeneity

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

$$Y = fixed.parts$$

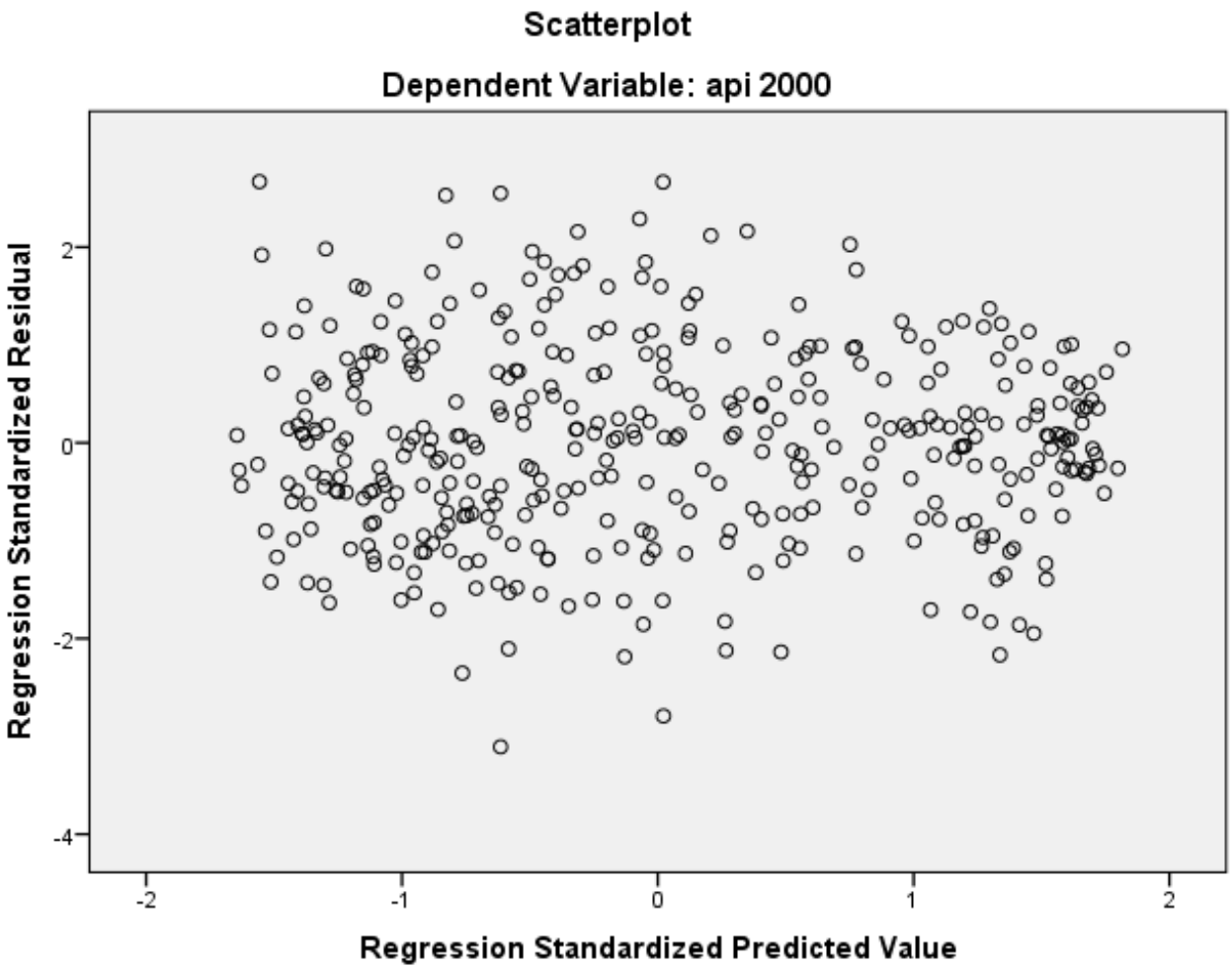
$$kg.ikan.ditangkap(Y) = jumlah.hook + jumlah.jam$$

Hasil plot residual dan AIC




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

Contoh plot residual yang tidak berpola



Contoh plot residual berpola

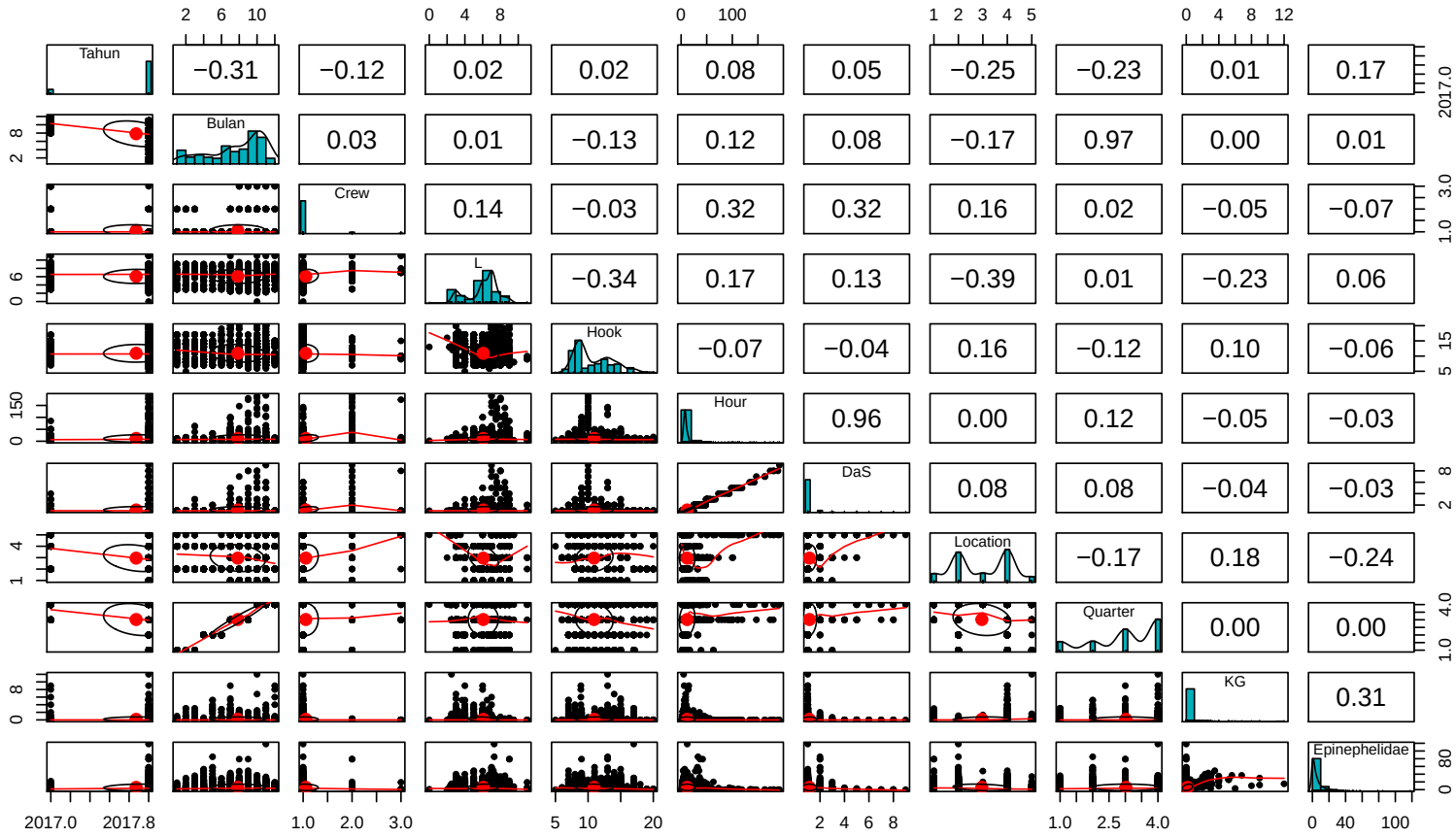
3. Fixed X

- 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.

Generalized least square models

- Linear regression model biasanya terdiri atas bagian:

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

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

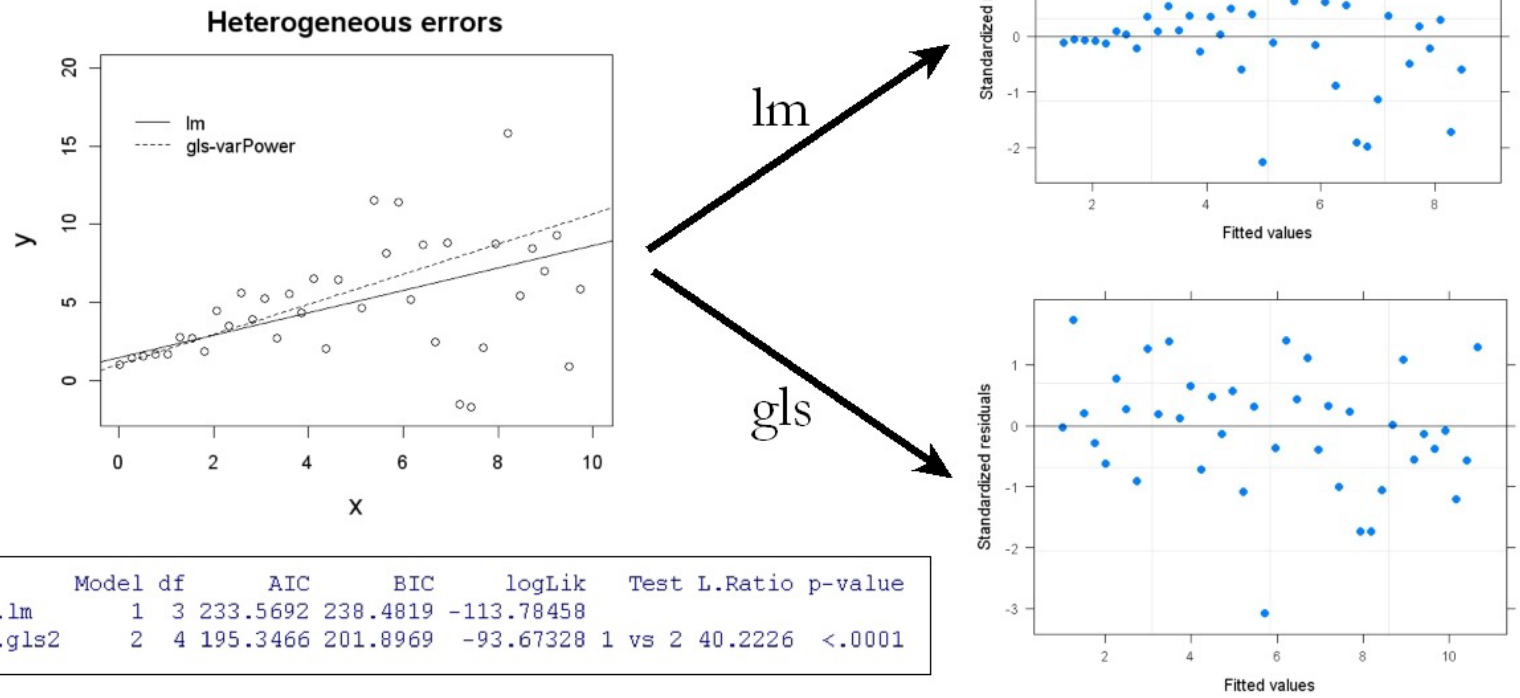
```
Heterogeneity
Nested data
Temporal correlation
Spatial correlation
Random noise
```

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 2: Generalized least square vs simple linear regression

Bagaimana bila masih ada pola residual?

Bila kita tidak ingin melakukan transformasi = gunakan distribusi lain (distribusi gamma) atau gunakan smoothing models.

Distribusi Gamma

- Distribusi normal (Gaussian) = total luas permukaan di bawah Normal density curve : 1.
- Probabilitas mengukur burung gagak 20 gr = 0.21
- Probabilitas mengukur berat gagak 5 gr = sangat kecil, walau masih terukur dalam distribusi normal.

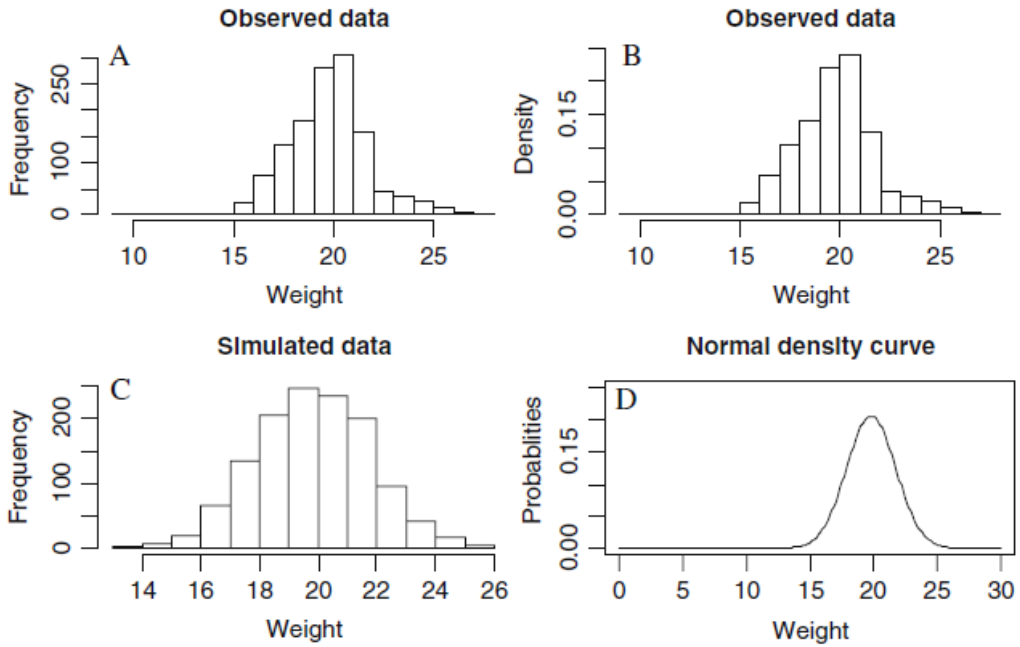


Fig. 8.1 **A**: Histogram of weight of 1281 sparrows. **B**: As panel A, but now scaled so that the total area in the histogram is equal to 1. **C**: Histogram of simulated data from a Normal distribution with mean and variance taken from the 1281 sparrows. **D**: Normal probability curve with values for the mean and the variance taken from the sample of 1281 sparrows. The surface under the Normal density curve adds up to 1

PDF distribusi normal

$$f(y_i; \mu, \alpha) = \frac{1}{\alpha\sqrt{2\pi}}$$

Mean $= E(Y) = \mu$ dan Variance $= var(Y) = \sigma^2$

Maka, variabel y bisa bernilai $-\infty$ dan ∞.

Penggunaan distribusi Gamma

- Bila data continuous kita masih memiliki residual berpola = gunakan distribusi gamma
- Persyaratan Y bernilai positif.

Mean $= E(Y) = \mu$ dan Variance $= var(Y) = \frac{\sigma^2}{v}$

- Distribusi Gamma memungkinkan overdispersion
- Dispersion ditentukan oleh nilai v^{-1}. Nilai v yang kecil menunjukkan persebaran data cukup besar. Sebaliknya, dengan nilai v yang besar, maka persebaran data menjadi kecil dan bentuk distribusi Gamma akan mendekati distribusi Normal (Gaussian).

PDF distribusi Gamma

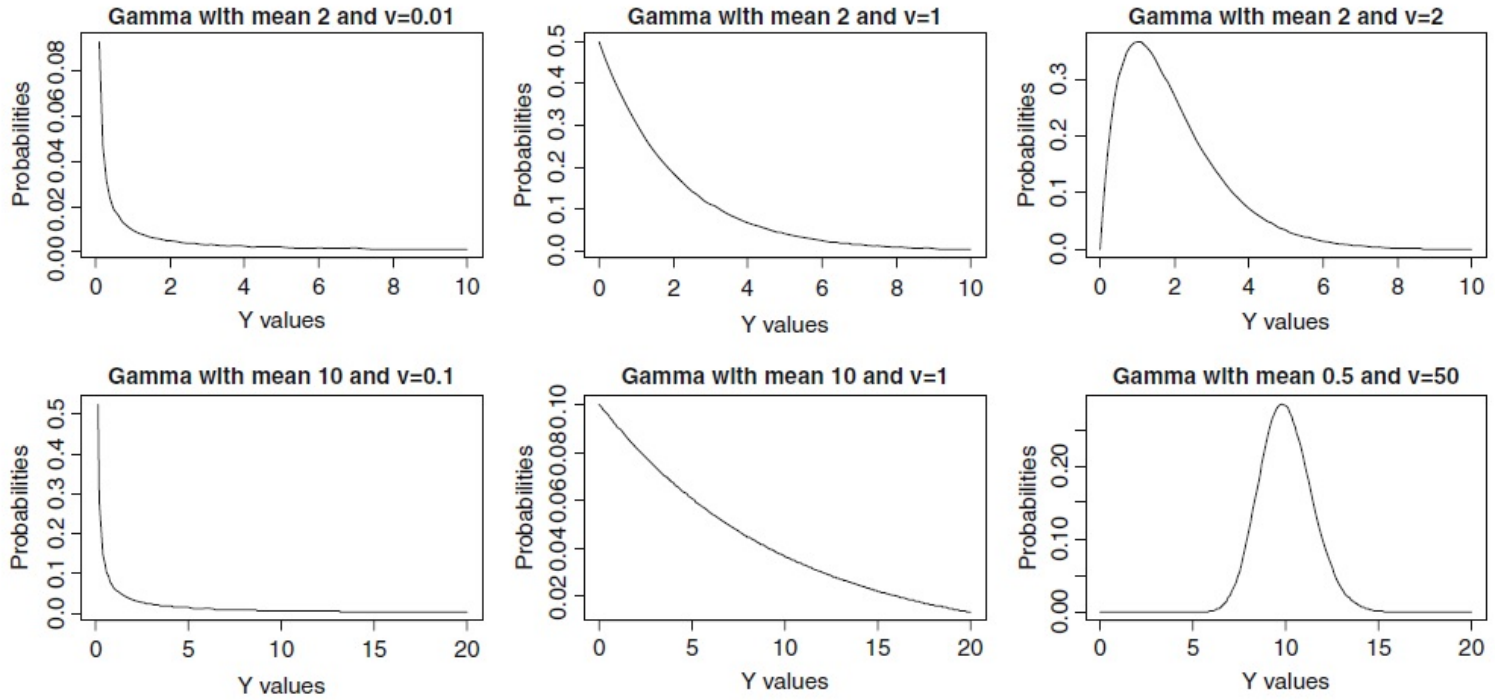
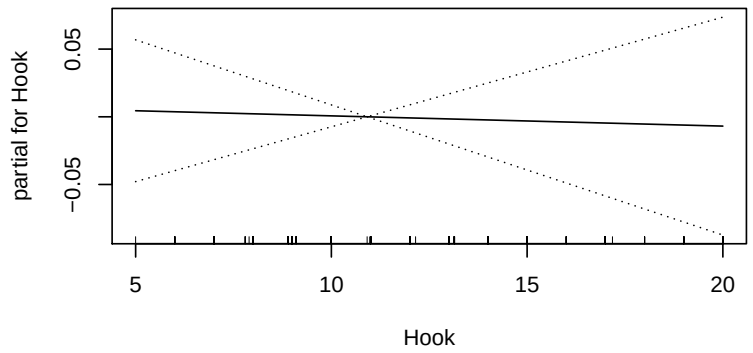
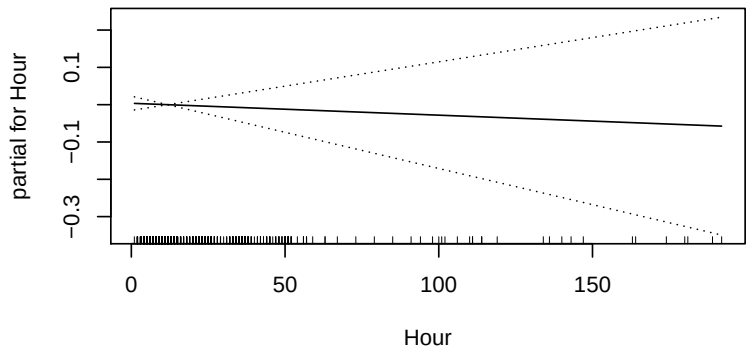


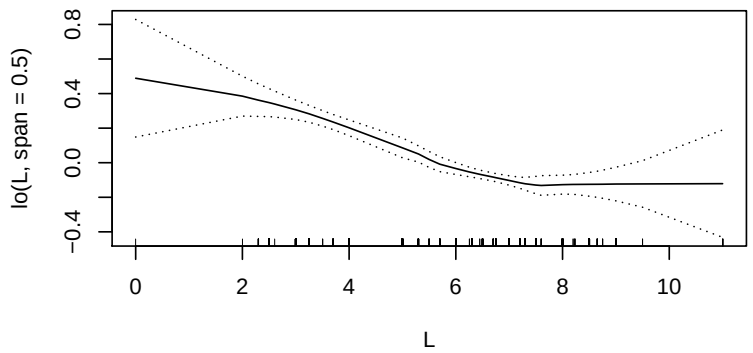
Fig. 8.4 Gamma distributions for different values of μ and v. The R function `dgamma` was applied, which uses a slightly different parameterisation: $\mathrm{E}(Y) = a \times s$ and $\mathrm{var}(Y) = a \times s^2$, where a is called the shape and s the scale. In our parameterisation, $v = a$ and $\mu = a \times s$

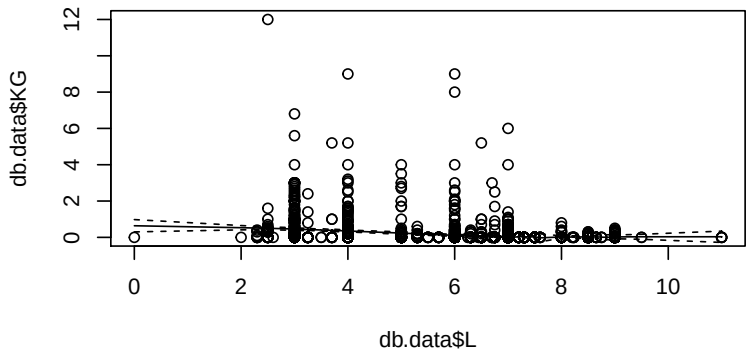
Additive model

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



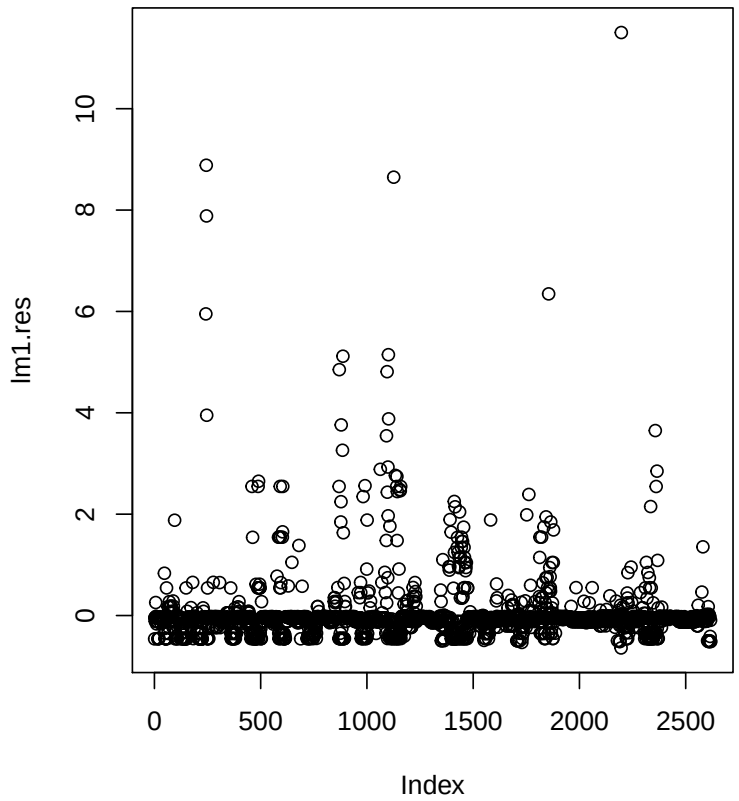


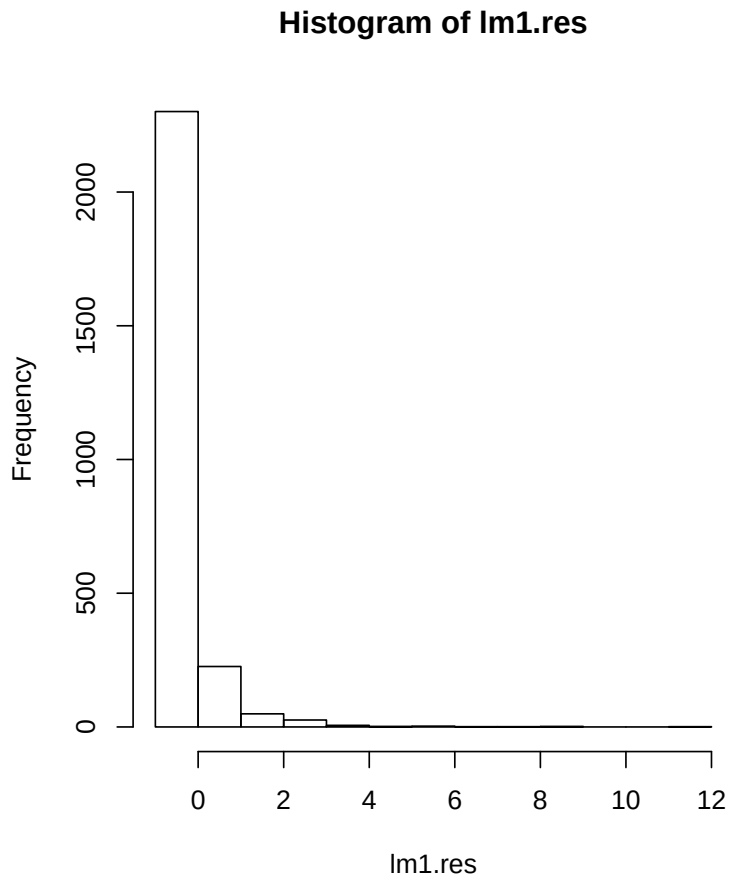




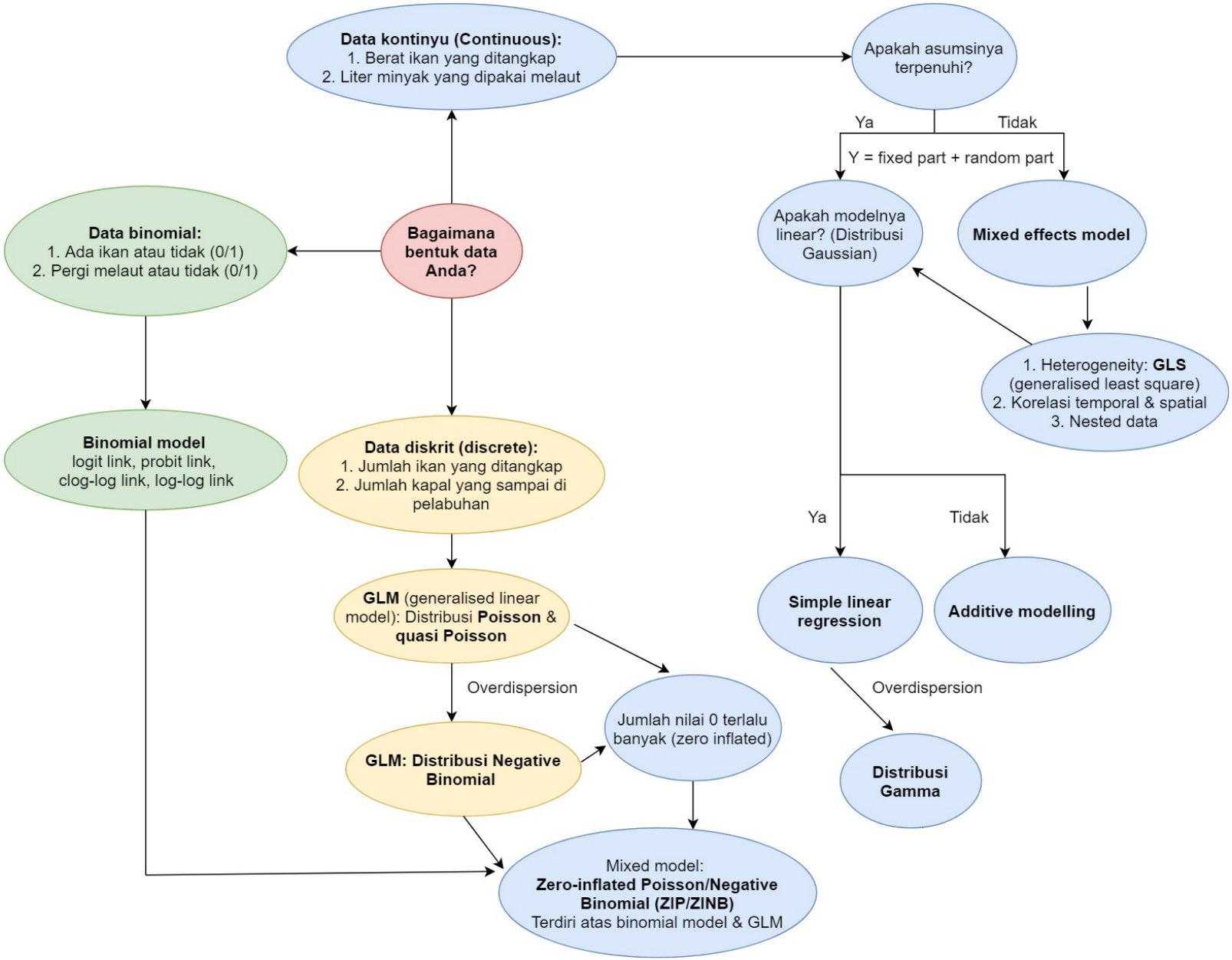
Plot residual dan AIC additive model

```
## [1] 4931.96
```





Alur permodelan



Data count

- Generalised linear modelling (GLM) dan generalised additive modelling (GAM) merupakan model perpanjangan dari regresi linear dan additive model.
- GLM dan GAM memiliki distribusi non-Gaussian = hubungan (relationship atau link) antara response variable dan explanatory variable berbeda.
- Distribusi Bernoulli atau binomial: data presence-absence, proportional data (0-100%).
- Distribusi Poisson atau negative binomial: count data.

Tahapan dalam menggunakan GLM/GAM

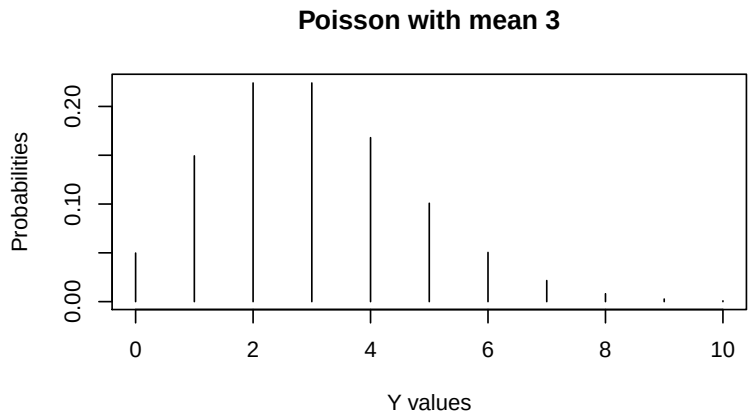
1. memilih distribusi response variable
2. mendefinisikan/menentukan kovariat (systematic part)
3. menspesifikasikan relationship/link antara expected value (response variable) dengan systematic part (kovariat penjelas)

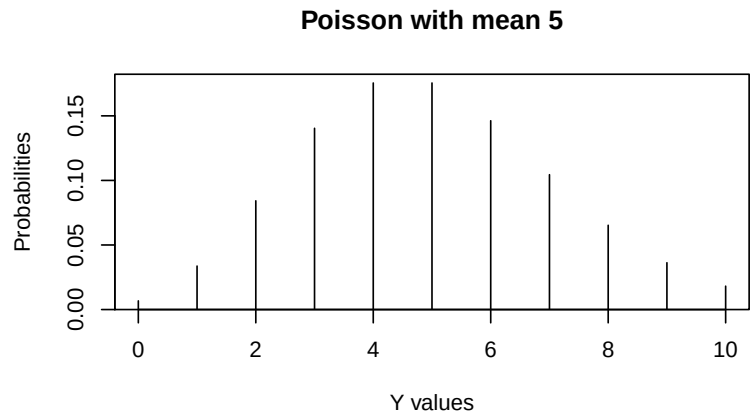
Distribusi Poisson

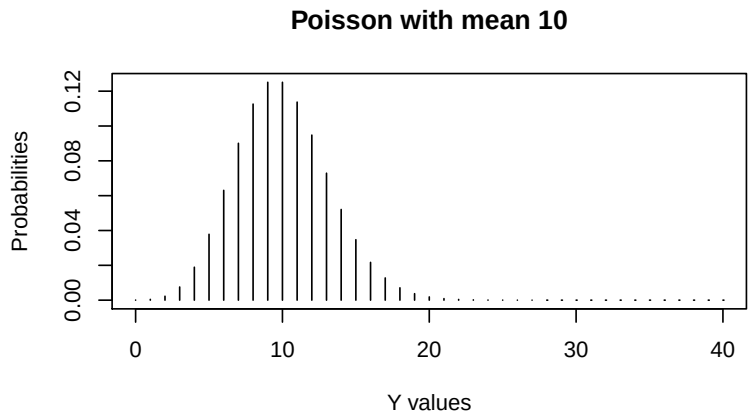
$$f(y;\mu) = \frac{\mu^y}{\times} e^{-\mu} y!$$

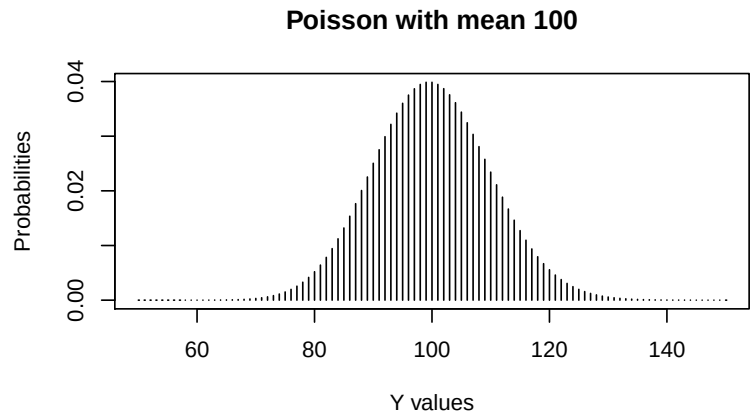
Dimana y nilainya harus ≥ 0 dan berupa bilangan bulat/integer. Persamaan ini mendeskripsikan mengenai probability Y dengan nilai mean μ.

Distribusi Poisson









μ kecil, kurva densitas = skewed

μ besar, kurvanya menjadi simetris, menyerupai distribusi normal.

Distribusi Poisson

- $P(Y < 0) = 0$ (tidak bernilai < 0)
- Mean $= E(Y) = \mu$
- Variance $= var(Y) = \mu$

Sepintas mengenai variabel offset

- Ketika kita mengumpulkan data count, observasi memiliki tingkatan upaya (effort) yang mmungkin berbeda.
- e.g., pengamatan jumlah semut yang sampai pada sumber makanan tertentu.
- Idealnya, setiap sumber makanan akan diobservasi dengan jumlah 'effort' yang sama (misalnya: lama pengamatan sama).
- Bisa jadi effort pengamatannya berbeda, sehingga analisisnya harus mencakup nilai effort ini.

Variabel offset

- Distribusi Poisson dan negative binomial bisa mencakup nilai offset.
- Persamaan Poisson misalnya bisa ditulis sebagai:

$$log(\mu) = \beta_0 + \beta_1 x_1 + \beta_p x_p$$

Variabel offset

- Variabel offset = beta coefficientnya 1.
- Ketika dimasukkan dalam model Poisson atau negative binomial, yang dimodelkan adalah nilai offset sebagai 'rate' (faktor pengali) dan bukan nilai 'count'nya sendiri.
- Contoh semut = variabel offset: nilai log waktu yang dipakai untuk mengamati sumber makanan.
- Persamaan dengan nilai offset atau logged effort variable (A):

$$log(\mu) = 1.log(A) + \beta_0 + \beta_1 x_1 + \beta_p x_p$$

Beta coefficient atau standardized regression coefficient, merupakan estimasi dari sebuah analisis regresi yang telah distandardisasi sehingga variance dari variabel dependent dan explanatorynya totalnya bernilai 1.

Variabel Offset

- Yang equivalen dengan:

$$log(\frac{\mu}{A}) = \beta_0 + \beta_1 x_1 + \beta_p x_p$$

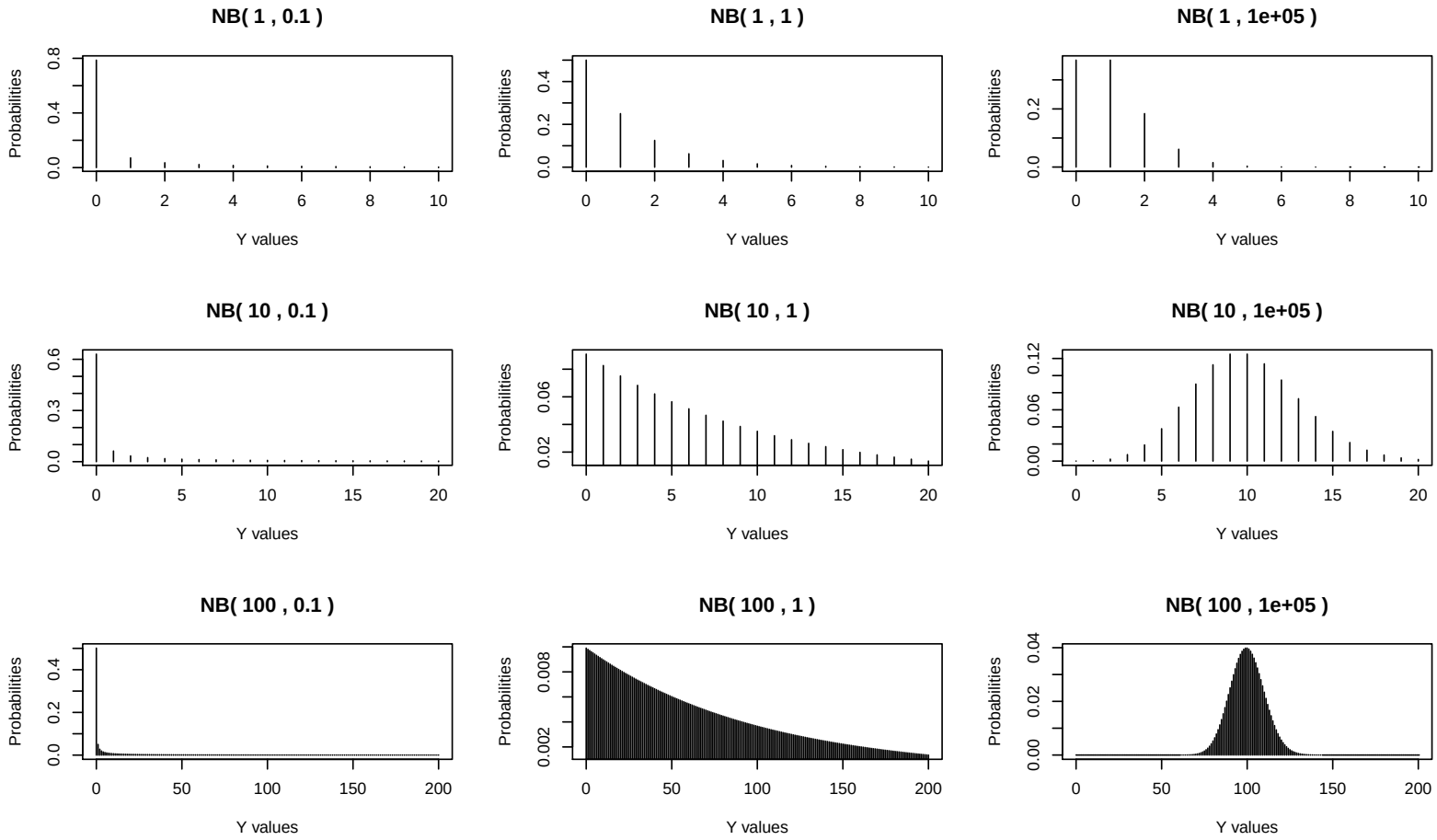
- Fungsi offset menjadi faktor “pengali”.
- Ketika mem-fitkan model dengan offset, nilai exponensial dari beta coefficient sebuah predictor variable menjelaskan seberapa banyak perubahan ‘rate’ akan terjadi apabila dikalikan penambahan satu unit dari predictor variable.
- Dengan offset = dua kali effort menyebabkan penambahan sebanyak dua kali dari nilai count.
- Apabila asumsi ini tidak sesuai = variabel sebagai covariate, daripada sebagai offset.

Distribusi Negative Binomial

- Distribution function dari negative binomial terdiri atas 2 parameter: μ dan k.
- Nilai mean dan variance Y adalah:
- Mean $= E(Y) = \mu$ dan Variance $= var(Y) = \mu + \frac{\mu^2}{k}$
- Bila variance $>$ mean, maka model kita memiliki *overdispersion*.
- Ditentukan oleh nilai k (parameter dispersion).
- Bila nilai k sangat besar (relatif terhadap μ^2) $=$ maka $\frac{\mu^2}{k}$ mendekati 0, atau variance-nya $= \mu$.
- Dalam kondisi ini, negative binomial $=$ distribusi Poisson.
- Semakin kecil nilai k, maka akan semakin besar juga overdispersionnya.

Distribusi Negative Binomial

μ & k = nilai mean dan dispersion parameter



Distribusi Bernoulli dan binomial

- Biasanya dicontohkan dari distribusi yang muncul dari melontarkan koin.
- Didefinisikan sebagai N percobaan melontar yang identik dan independen
- Dengan probabilitas sukses $P(Y_i = 1) = \pi$
- Dan probabilitas gagal $P(Y_i = 1) = 1 - \pi$.
- Sukses dan gagal adalah nilai 1 dan 0.
- Independen = setiap lontaran tidak berkaitan dengan lontaran sebelum atau sesudahnya.
- Identikal = setiap lontaran memiliki probabilitas sukses yang sama besarnya.

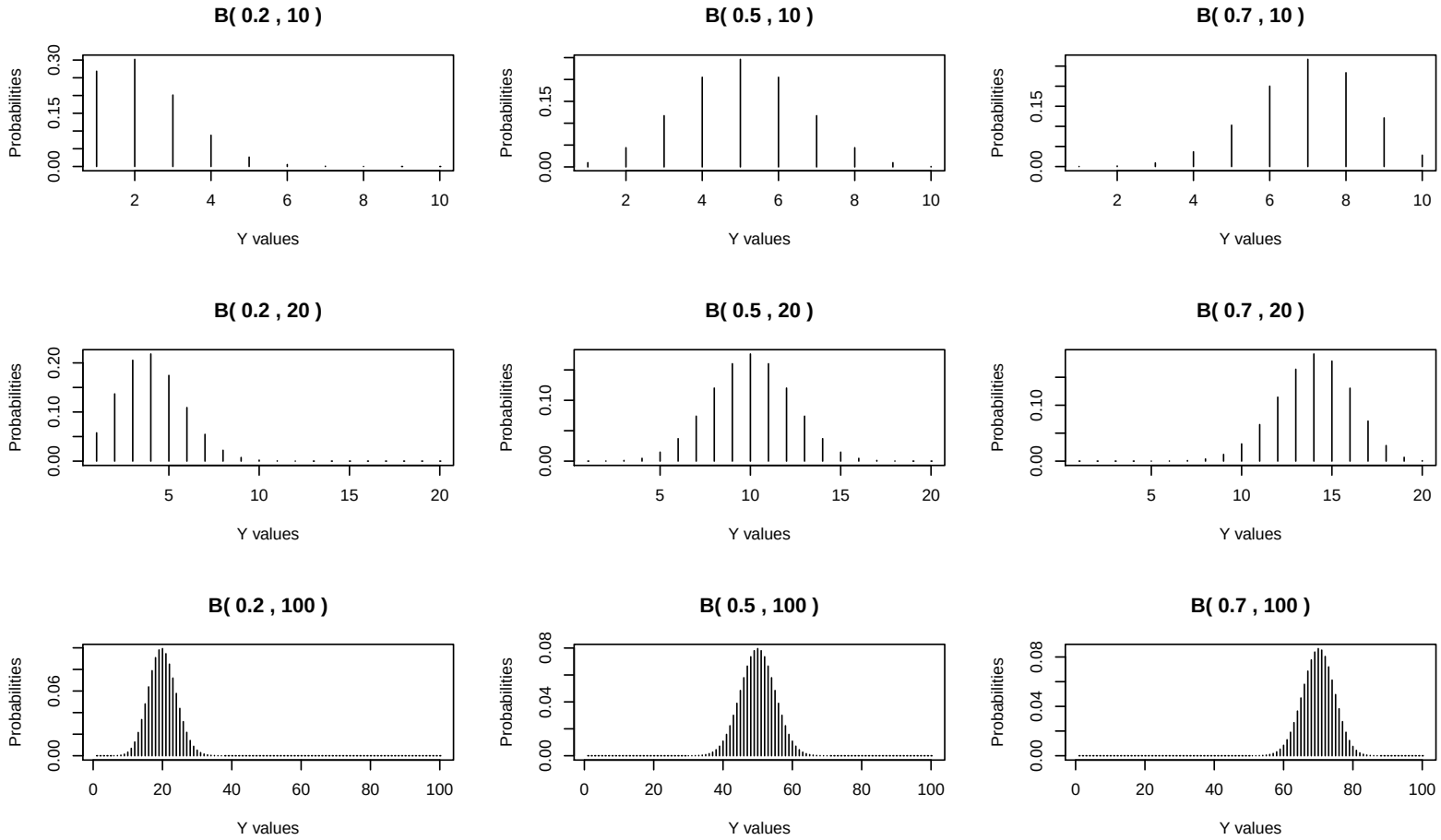
PDF distribusi Binomial

$$f(y;\pi) = \binom{N}{y} \times \pi^{y} \times (1-\pi)^{N-y}$$

- Mean $= E(Y) = N \times \pi$
- Variance $= var(Y) = N \times \pi \times (1-\pi)$

Distribusi Binomial

π & N



Distribusi Binomial dan Bernoulli

- Studi presence absence dari suatu spesies, e.g., presence absence suatu spesies ikan tertentu dalam 62 situs estuari (Zuur et al 2001).
- Distribusi Bernoulli diperoleh bila $N = 1 =$ melontarkan koin sekali dan mencatat probabilitasnya.
- Biasanya distribusi Bernoulli tidak dibedakan dengan distribusi Binomial.

Bagaimana cara memilih distribusi yang tepat untuk data kita?

- Data count = distribusi Poisson.
  - High overdispersion = gunakan distribusi negative binomial.
  - Distribusi Normal juga bisa dipakai, tapi ingat: tidak meng-exclude nilai negatif.
- Continuous data= distribusi Normal
  - High overdispersion = distribusi gamma
  - Additive modelling.
- Perlu diingat: distribusi-distribusi ini berlaku untuk response variable, bukan explanatory variable.
- Bila ragu, plot terlebih dahulu nilai mean terhadap variance dari response variable untuk melihat hubungan mean-variancenya

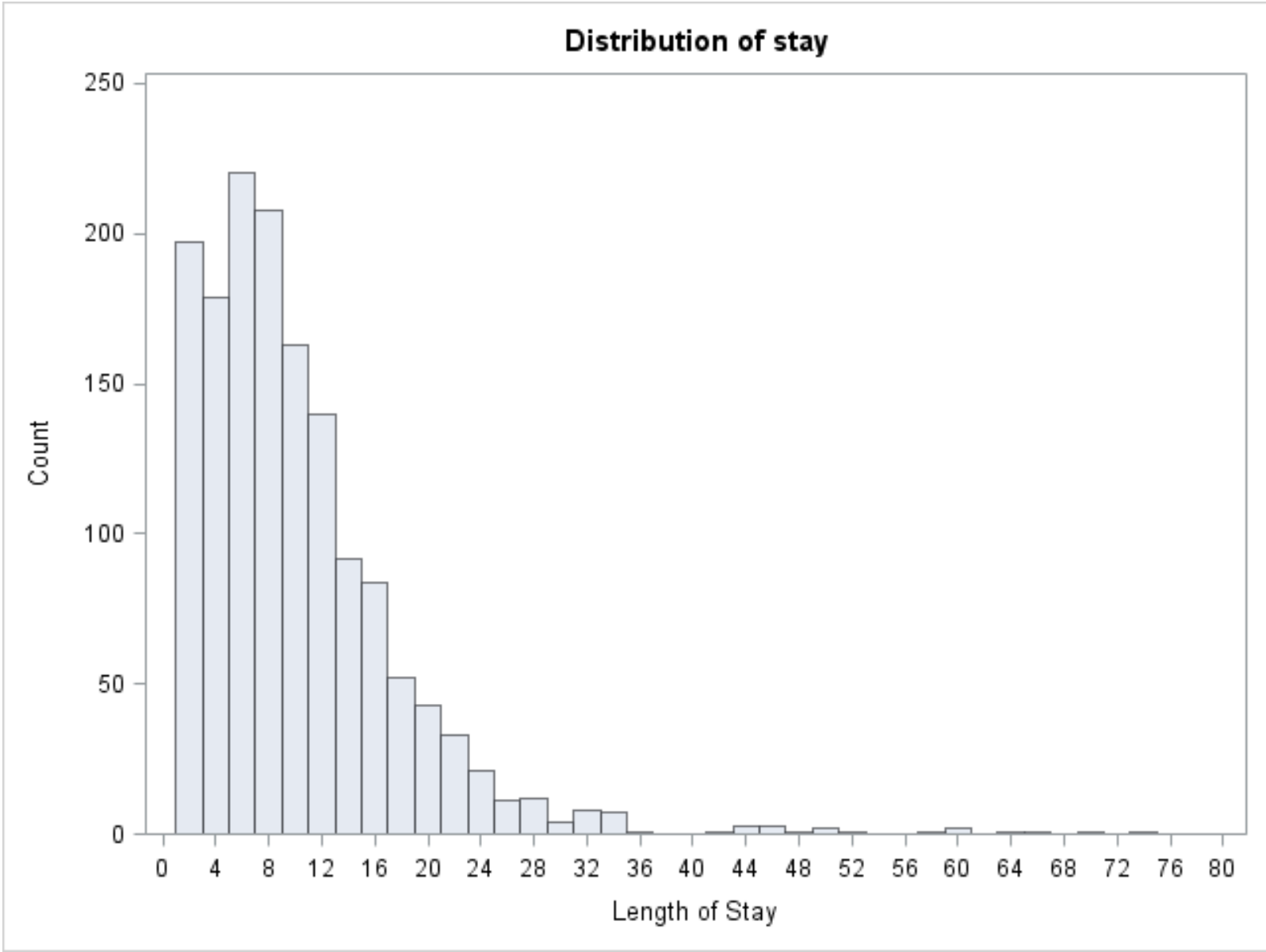
Zero truncated dan zero inflated

- Zero truncated dan zero inflated distribution bisa diaplikasikan untuk data count
- Berlaku untuk distribusi Poisson, negative binomial dan geometric.
- Zero truncated = data yang tidak bisa mengambil nilai 0, e.g., studi kedokteran: seperti berapa lama pasien tinggal di rumah sakit; transek tanaman, tidak mungkin kita menemukan transek yang tidak ada abundancenya (akibat experimental design).
- Variabel-variabel ini tidak mungkin memiliki nilai 0, tapi karena 0 pasti masuk dalam probabilitas distribusi untuk data count, ini bisa membuat nilai μ menjadi bermasalah kalau nilai meannya kecil.

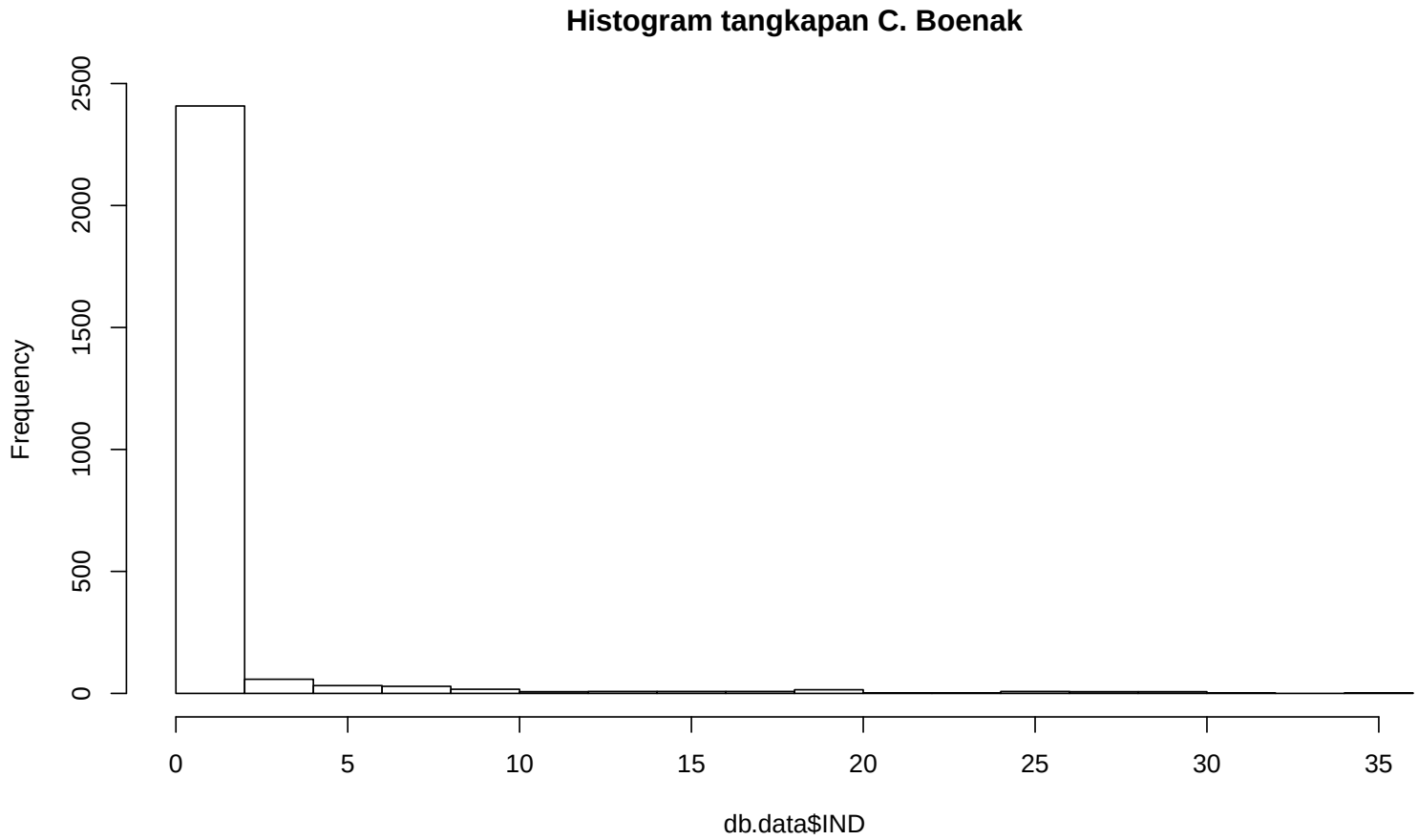
Zero truncated dan zero inflated

- Solusinya = modifikasi distribusi dan meng-exclude kemungkinan observasi nilai 0 = zero truncated distribution.
- Prinsip yang sama diaplikasikan pada kasus-kasus dimana terdapat banyak nilai 0 (zero inflated Poisson).

Distribusi Zero Truncated



Distribusi Zero Inflated



Bagaimana cara memilih distribusi terbaik?

- Membandingkan nilai AIC = semakin kecil AIC, semakin baik
- Inspeksi visual data residual
- Perbandingan nilai R-squared = semakin besar, semakin banyak informasi yang terjelaskan oleh model yang kita pakai (hati-hati dengan overfitting!).

Bagaimana menentukan best fit model?

- Gunakan fungsi `drop1` atau `dredge`
- Gunakan intuisi atau background information tentang subyek yang kita coba analisis

Seleksi model

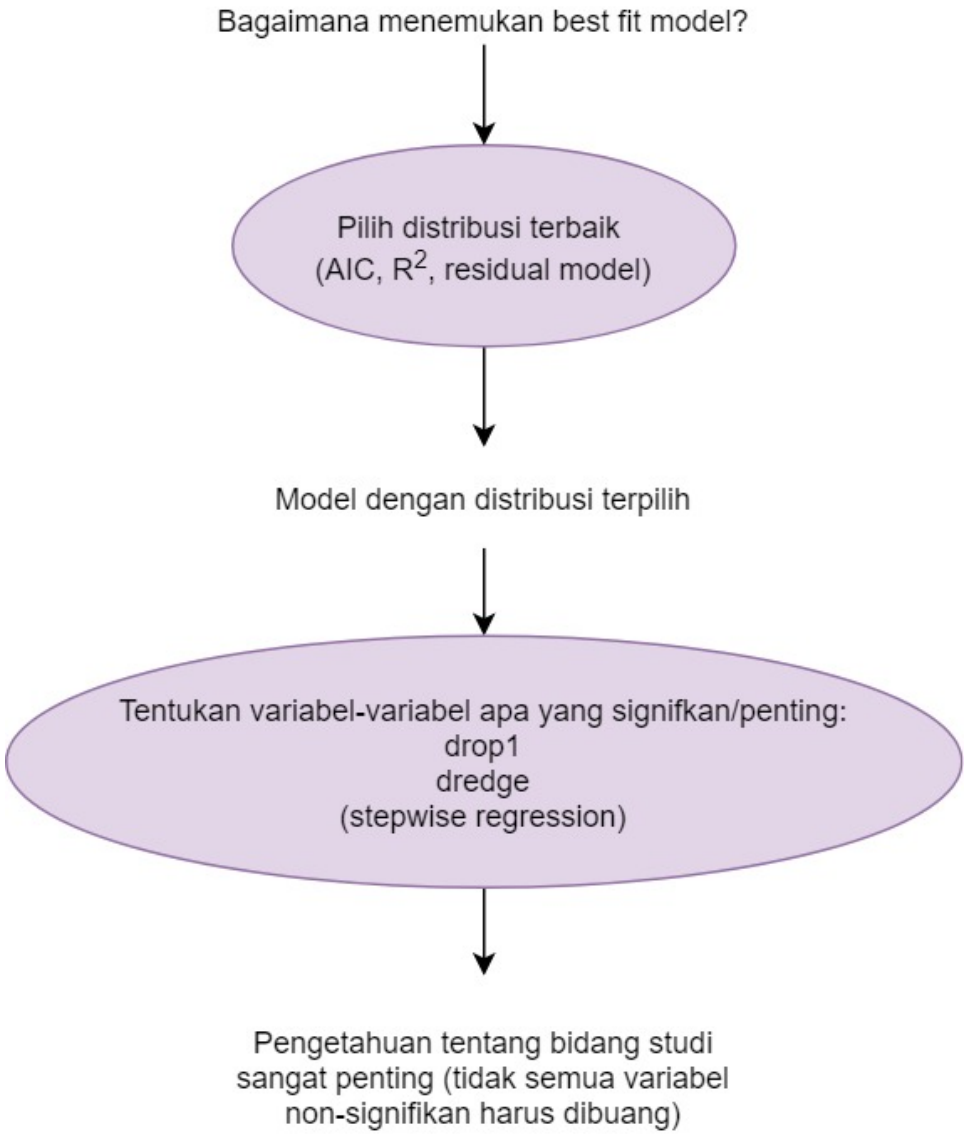
- Lihat kembali data hasil tangkapan ikan *Cephalopholis boenak* di Maluku Utara pada tahun 2017-2018.
- Apakah yang mempengaruhi jumlah tangkapan ikan *C. boenak*?
- Persamaan yang diuji:

$$IND = Bulan + Quarter + Boat + Crew + Hook + offset(DaS)$$

Dimana:

- IND = jumlah individu tertangkap per trip
- Tahun = tahun penangkapan
- Quarter = 1/4 bulan pertama, kedua, ketiga dan keempat
- Boat = nama pemilik kapal
- Crew = jumlah awak kapal
- Offset (DaS) = Days at Sea

Seleksi model



Seleksi model

Kita coba merunut permodelan yang akan kita pakai, sesuai tahap-tahap yang sudah kita pelajari tadi. Tahapannya adalah sebagai berikut:

```
library(MASS)
library(nlme)
library(MuMIn)
library(readxl)
library(lme4)
library(emmeans)
library(gam)
library(AER)
library(AICcmodavg)
library(rsq)
library(pscl)
library(lmtest)
```

Seleksi model

```
db.data<-read.csv("boenak.csv")
#menetapkan bulan sebagai data faktorial
#dengan urutan 1-12 (Jan-Des)
db.data$Bulan = factor(db.data$Bulan,
                       levels = c("1", "2", "3",
                      "4", "5", "6", "7", "8", "9",
                       "10", "11", "12"), order = T)
#menetapkan perempat bulan sebagai
#data faktorial dengan urutan 1, 2, 3, 4
db.data$Quarter = factor(db.data$Quarter,
                         levels = c("1", "2", "3",
                         "4"), order = T)
```

Seleksi model

```
#lain-lain sebagai data faktorial
db.data$Tahun = as.factor(db.data$Tahun)
db.data$Boat = as.factor(db.data$Boat)
db.data$Location = as.factor(db.data$Location)
#menetapkan beberapa data sebagai data numerik
db.data$Crew = as.numeric(db.data$Crew)
db.data$L = as.numeric(db.data$L)
db.data$Hook = as.numeric(db.data$Hook)
db.data$DaS = as.numeric(db.data$DaS)
db.data$IND = as.numeric(db.data$IND)
```

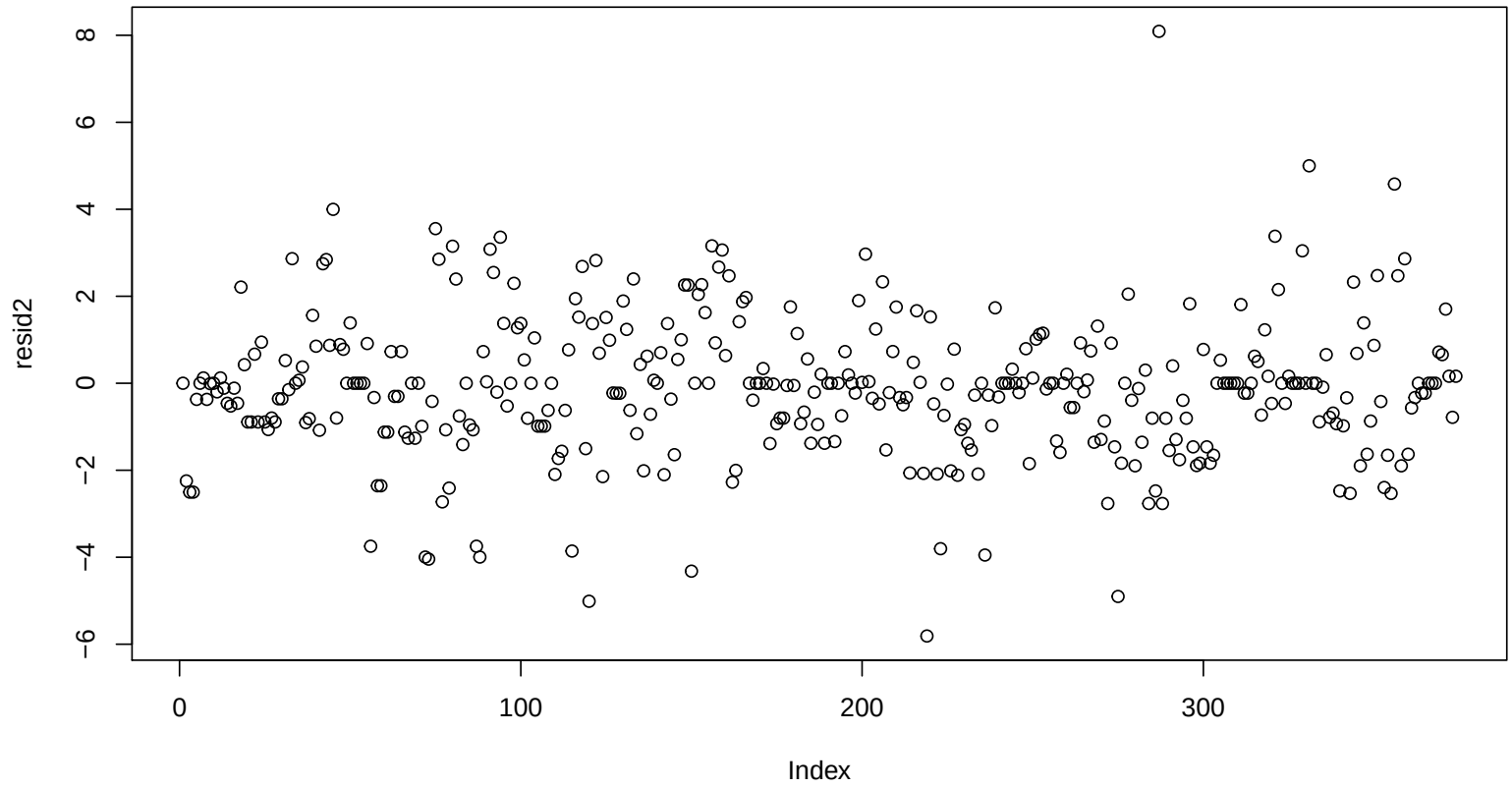
Seleksi model

```
for (i in 1:2618){
  if(db.data$IND[i] > 0){db.data$PE[i] = 1}
  else{db.data$PE[i] = 0}
}

db.data.Count = db.data[db.data$PE>0,]
#Linear regression (Gaussian)
glm.1 = glm(IND ~ Tahun + Quarter + Boat + Crew + L + Hook + offset(DaS),
            data = db.data.Count, family = gaussian)
resid1<-resid(glm.1)
```

Seleksi model

```
#Poisson
glm.2 = glm(IND ~ Tahun + Quarter + Boat + Crew + L + Hook + offset(DaS),
            data = db.data.Count, family = poisson)
resid2<-resid(glm.2)
plot(resid2)
```



Seleksi model

```
dispersiontest(glm.2)
```

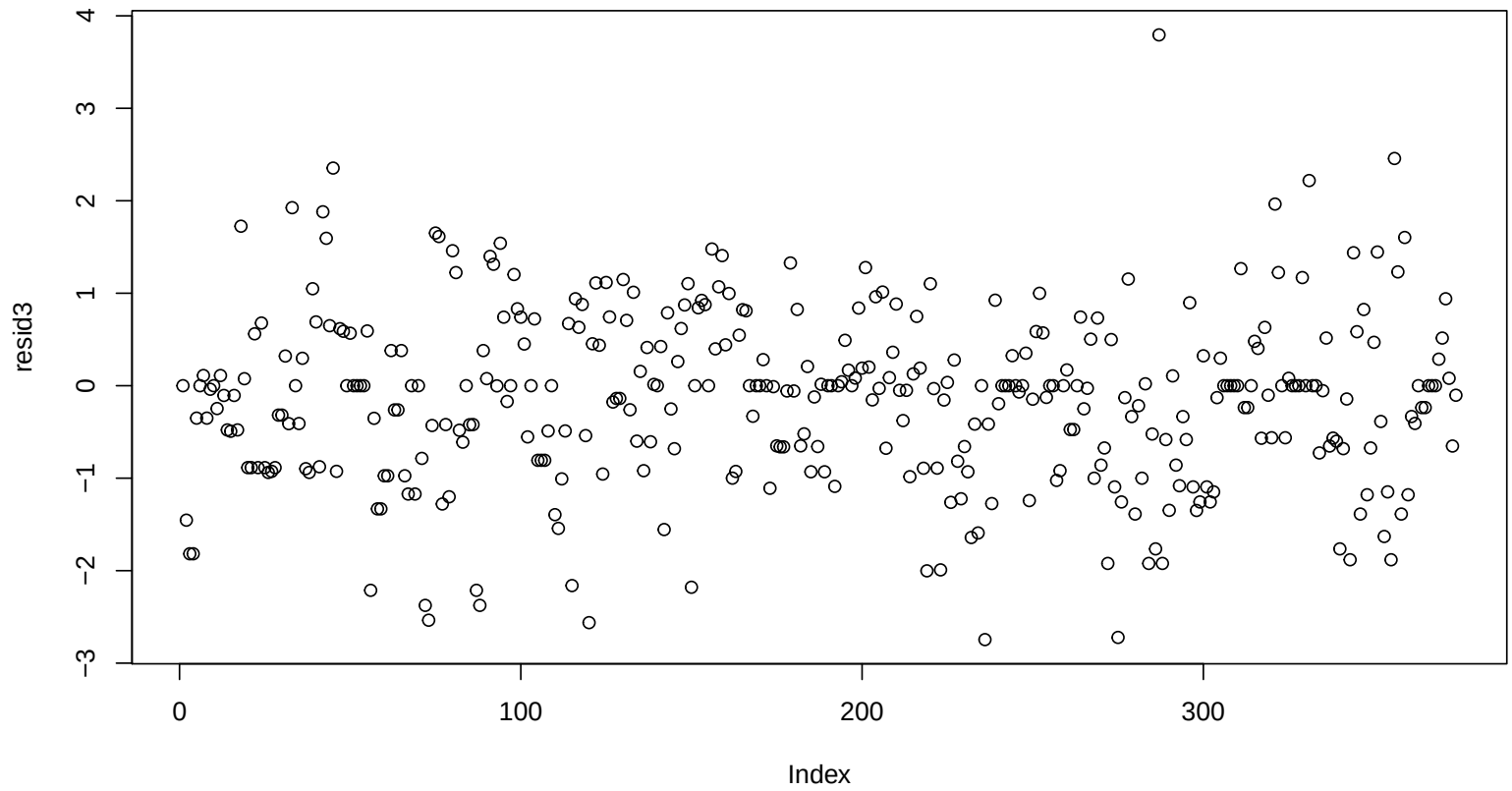
```
##
##  Overdispersion test
##
## data:  glm.2
## z = 4.1659, p-value = 1.551e-05
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
## dispersion
##   2.789862
```

```
glm.2b = glm(IND ~ Tahun + Quarter + Boat + Crew + L + Hook + offset(DaS),
             data = db.data.Count, family = quasipoisson)
summary(glm.2b)
```

```
##
## Call:
## glm(formula = IND ~ Tahun + Quarter + Boat + Crew + L + Hook +
##     offset(DaS), family = quasipoisson, data = db.data.Count)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.8112  -0.9431   0.0000   0.7139   8.0912
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              1.195510   2.661089   0.449  0.65361
## Tahun2018                0.936651   0.281687   3.325  0.00101 **
## Quarter.L               -0.317435   0.127156  -2.496  0.01315 *
## Quarter.Q               -0.717416   0.121148  -5.922 9.73e-09 ***
## Quarter.C               -0.157938   0.110561  -1.429  0.15431
## BoatABDUL HARIS          0.279415   2.488800   0.112  0.91069
```

Seleksi model

```
#Negative binomial
glm.3 = glm.nb(IND ~ Tahun + Quarter + Boat + Crew + L + Hook + offset(DaS),
               data = db.data.Count)
resid3<-resid(glm.3)
plot(resid3)
```




```
#residual 4.3, taken to be 1
```

Seleksi model

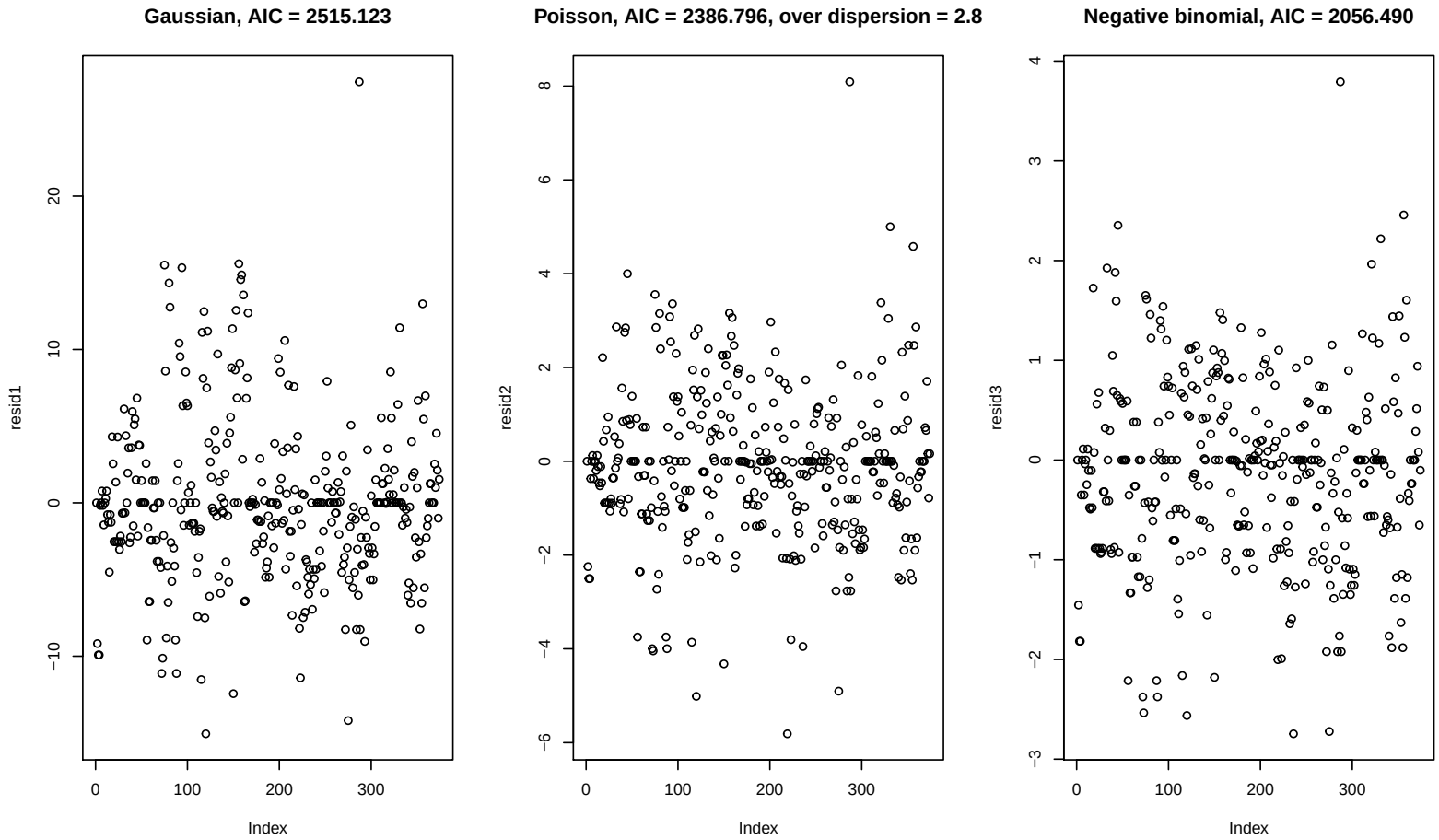
```
AIC(glm.1,glm.2,glm.3)
```

```
##         df      AIC
## glm.1 107 2515.123
## glm.2 106 2386.796
## glm.3 107 2056.490
```

```
#Bandingkan nilai AIC dan residual dari ketiga uji model di atas
```

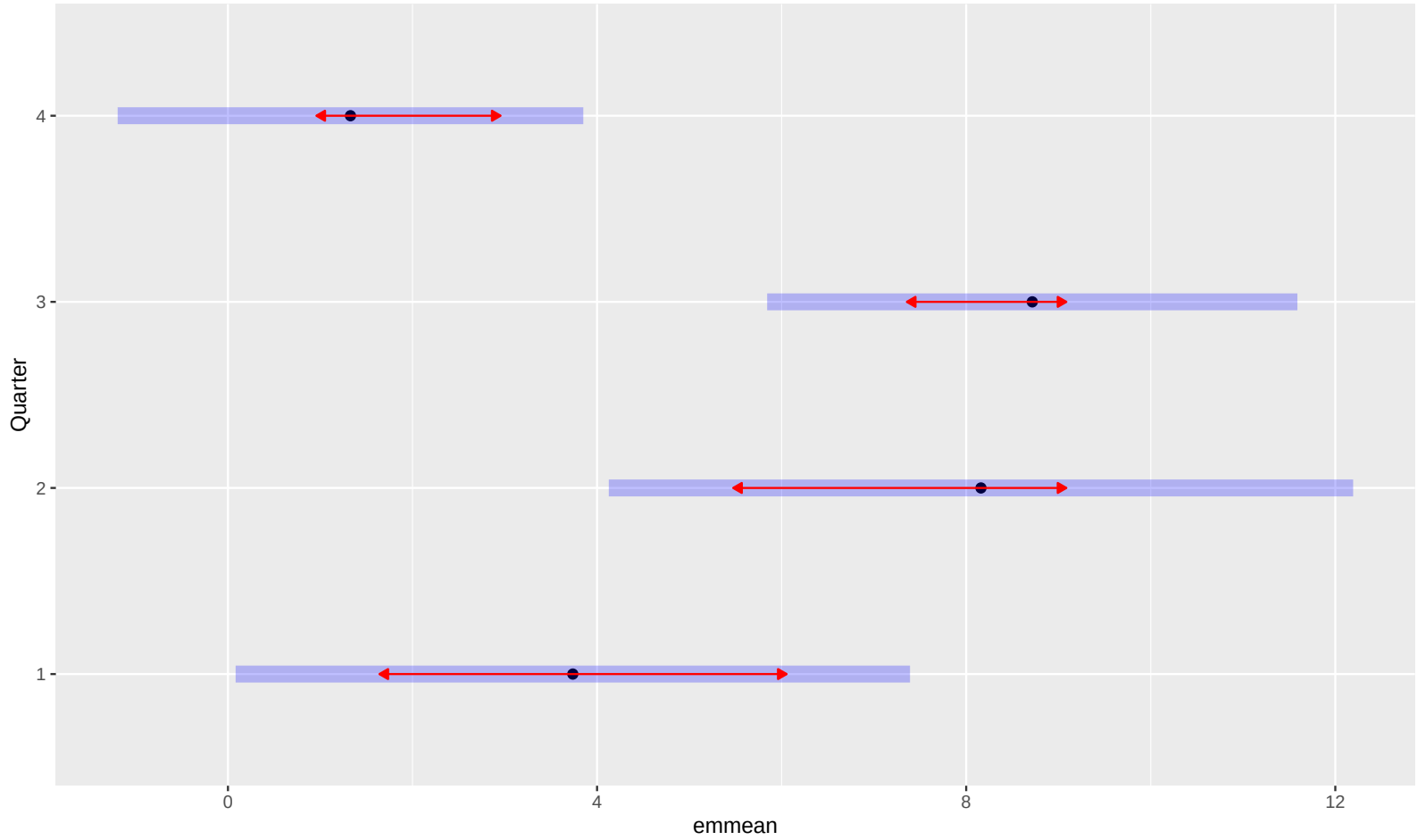
Seleksi model

```
par(mfrow=c(1,3))
plot(resid1, main="Gaussian, AIC = 2515.123")
plot(resid2, main="Poisson, AIC = 2386.796, over dispersion = 2.8")
plot(resid3, main="Negative binomial, AIC = 2056.490")
```



Seleksi model

```
#calculate mean value for "Quarter"
em.nb<-emmeans(glm.1, "Quarter") #A nesting structure was detected
#in the fitted model:
    #Bulan %in% Quarter
#pairwise comparisons
plot(em.nb, comparisons=T)
```



Mengkalkulasi nilai R2

- %R2 bisa membantu dalam memilih distribusi model terbaik.
- %R2 menjelaskan seberapa banyak variansi yang terjelaskan melalui model.
- R2 merupakan hasil output dari summary pada function `lm` untuk distribusi Gaussian.
- Bisa juga dihitung secara manual dari output summary model GLM.

```
#linear regression (Gaussian)
glm.0 = lm(IND ~ Tahun + Quarter + Boat + Crew + L + Hook + offset(DaS),
           data = db.data.Count)
summary(glm.0)
```

```
##
## Call:
## lm(formula = IND ~ Tahun + Quarter + Boat + Crew + L + Hook +
##     offset(DaS), data = db.data.Count)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.060  -2.664   0.000   1.575  27.459
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -6.8926    12.9214  -0.533 0.594179
```

Mengkalkulasi nilai R2

- Multiple R2 = 0.55, atau 55%.
- Adjusted R-squarenya = 0.38 atau 38%.
- Bandingkan dengan fungsi `glm`.
- Adjusted R-square biasanya nilainya < multiple R2 karena memberikan penalti terhadap jumlah kovariat.

```
glm.1 = glm(IND ~ Tahun + Quarter + Boat + Crew + L + Hook + offset(DaS),
            data = db.data.Count, family = gaussian)
summary(glm.1)
```

```
##
## Call:
## glm(formula = IND ~ Tahun + Quarter + Boat + Crew + L + Hook +
##     offset(DaS), family = gaussian, data = db.data.Count)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -15.060   -2.664    0.000    1.575   27.459
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -6.8926    12.9214  -0.533 0.594179
## Tahun2018                  3.4953     1.4779   2.365 0.018738 *
## Quarter.L                 -1.4920     1.0850  -1.375 0.170237
## Quarter.Q                 -5.9059     1.1257  -5.247 3.15e-07 ***
## Quarter.C                 -0.9118     1.0982  -0.830 0.407162
## BoatABDUL HARIS            1.0766     7.8523   0.137 0.891047
## BoatABDULLAH              16.0238     7.3086   2.192 0.029205 *
```

Mengkalkulasi nilai R2

Hasilnya, kita memperoleh nilai `Null deviance` = 23119.2 dan `Residual deviance` = 8795.4. Maka R2 nya adalah=

$$R^2 = \frac{Null.deviance - Residual.deviance}{Null.deviance}$$

```
r1=(23119.2-10292)/23119.2
r1
```

```
## [1] 0.5548289
```

Hasil ini mirip dengan nilai multiple R-squared dengan function `lm`. Metode penghitungan ini bisa diaplikasikan pada distribusi Poisson dan Negative Binomial.

Seleksi model

Selanjutnya, untuk menggunakan fungsi drop1 dan dredge, maka berikut tahapan teknis penggunaannya:

```
#membandingkan model yang salah satu variabelnya
#di-drop terhadap model drop dan full model lainnya
drp1<-drop1(glm.3,test="Chi")
drp1
```

```
## Single term deletions
##
## Model:
## IND ~ Tahun + Quarter + Boat + Crew + L + Hook + offset(DaS)
##         Df Deviance    AIC    LRT  Pr(>Chi)
## <none>       330.42 2054.5
## Tahun    1   348.30 2070.4  17.88 2.357e-05 ***
## Quarter  3   393.06 2111.1  62.64 1.607e-13 ***
## Boat    98   658.17 2186.2 327.75 < 2.2e-16 ***
## Crew     1   335.63 2057.7   5.21  0.022483 *
## L        1   337.04 2059.1   6.62  0.010094 *
## Hook     1   338.07 2060.1   7.64  0.005698 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#variabel yang bila didrop AIC-nya jadi terendah
#dan tidak signifikan terhadap model bisa di-drop
#drop dilanjutkan sampai tidak ada lagi variabel
#yang tidak signifikan
```

Seleksi model

```
glm.3 = glm.nb(IND ~ Tahun + Quarter + Boat + Crew + L + offset(DaS),
               data = db.data.Count, link=log) #tidak ada yang di drop
drp1<-drop1(glm.3,test="Chi")
drp1
```

```
## Single term deletions
##
## Model:
## IND ~ Tahun + Quarter + Boat + Crew + L + offset(DaS)
##         Df Deviance    AIC     LRT  Pr(>Chi)
## <none>       330.97 2060.0
## Tahun    1   345.63 2072.7  14.657 0.0001289 ***
## Quarter  3   386.50 2109.6  55.526 5.304e-12 ***
## Boat    98   644.08 2177.2 313.105 < 2.2e-16 ***
## Crew     1   336.38 2063.4   5.408 0.0200455 *
## L        1   340.42 2067.5   9.445 0.0021167 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Seleksi model

Kita coba run model yang berbeda untuk bisa mempelajari metode drop variabel/kovariat.

```
glm.3 = glm.nb(IND ~ Bulan + Quarter + Boat + Crew + L +
                 Hook + offset(DaS),
               data = db.data.Count, link=log) #drop Hook
drp1<-drop1(glm.3,test="Chi")
drp1
```

```
## Single term deletions
##
## Model:
## IND ~ Bulan + Quarter + Boat + Crew + L + Hook + offset(DaS)
##         Df Deviance    AIC     LRT  Pr(>Chi)
## <none>       329.44 2034.0
## Bulan    8   386.20 2074.8  56.762 2.003e-09 ***
## Quarter  0   329.44 2034.0   0.000
## Boat    98   619.98 2128.6 290.536 < 2.2e-16 ***
## Crew     1   332.61 2035.2   3.170  0.074982 .
## L        1   336.44 2039.0   6.996  0.008168 **
## Hook     1   330.17 2032.8   0.727  0.393974
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Seleksi model

```
glm.3 = glm.nb(IND ~ Bulan + Quarter + Boat + Crew + L +
                 offset(DaS),
               data = db.data.Count, link=log) #drop Quarter
drp1<-drop1(glm.3,test="Chi")
drp1 #model final (best fit)
```

```
## Single term deletions
##
## Model:
## IND ~ Bulan + Quarter + Boat + Crew + L + offset(DaS)
##         Df Deviance    AIC     LRT  Pr(>Chi)
## <none>       329.59 2032.8
## Bulan    8   390.99 2078.2  61.396  2.48e-10 ***
## Quarter  0   329.59 2032.8   0.000
## Boat    98   618.92 2126.1 289.330 < 2.2e-16 ***
## Crew     1   332.77 2034.0   3.175  0.074768 .
## L        1   337.80 2039.0   8.207  0.004172 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Seleksi model

Atau gunakan dredge

```
glm.3 = glm.nb(IND ~ Tahun + Quarter + Boat + Crew + L +
                 Hook + offset(DaS),
               data = db.data.Count, link=log, na.action = na.fail)
dd<-dredge(glm.3)
dd1<-subset(dd, delta < 4) #prinsipnya menjelaskan sekitar >80%
dd1
```

```
## Global model call: glm.nb(formula = IND ~ Tahun + Quarter + Boat + Crew + L + Hook +
##     offset(DaS), data = db.data.Count, na.action = na.fail, link = log,
##     init.theta = 3.656269462)
## ---
## Model selection table
##    (Int)     Crw     Hok       L Qrt Thn df    logLik   AICc delta weight
## 59 3.036 -0.9765         -0.2238   +   +  8 -1018.493 2053.4  0.00  0.723
## 63 2.902 -0.9733 0.00824 -0.2177   +   +  9 -1018.403 2055.3  1.92  0.277
## Models ranked by AICc(x)
```

```
dd2<-model.avg(dd, subset = cumsum(weight) <= 0.99,fit=TRUE)
emmeans(dd2,~c(Quarter), type = "links",
                         weights = "proportional",
                         at = list(L = 7, DaS = 1, Crew = 1, Hook = 9))
```

```
##  Quarter emmean    SE  df lower.CL upper.CL
##  1        -2.54 0.703 366    -3.93   -1.162
##  2        -3.34 0.707 366    -4.73   -1.955
##  3        -2.51 0.701 366    -3.89   -1.134
##  4        -1.56 0.720 366    -2.98   -0.145
##
## Results are averaged over the levels of: Tahun
## Results are given on the log (not the response) scale.
```

Seleksi model

Terdapat banyak nilai 0 pada hasil tangkapan ikan. Maka, salah satu pendekatan yang bisa kita lakukan antara lain dengan menggunakan distribusi Zero Inflated sebagai berikut:

```
#Kita siapkan dulu persamaannya. Sebagai catatan, dalam distribusi zero inflated,
#karena variabel 'Bulan' adalah nested di dalam 'Quarter', maka ia tidak bisa
#digabung permodelannya. Kita memilih untuk hanya memakai 'Quarter'
#sebagai variabel waktu.

f1<-formula(IND ~ Quarter+Crew+L+Hook+DaS|
              Quarter+Crew+L+Hook+offset(DaS),
            data = db.data)
Zip1<-zeroinfl(f1,data=db.data)
#Uji menggunakan distribusi Gaussian
Nb1A<-zeroinfl(f1,dist="negbin",link="logit", data=db.data)
#Uji menggunakan distribusi Negative Binomial
```

Seleksi model

```
#Cek residual
res.zip1<-resid(Zip1)
res.zip2<-resid(Nb1A)
#Uji likelihood test
lrtest(Zip1,Nb1A)
```

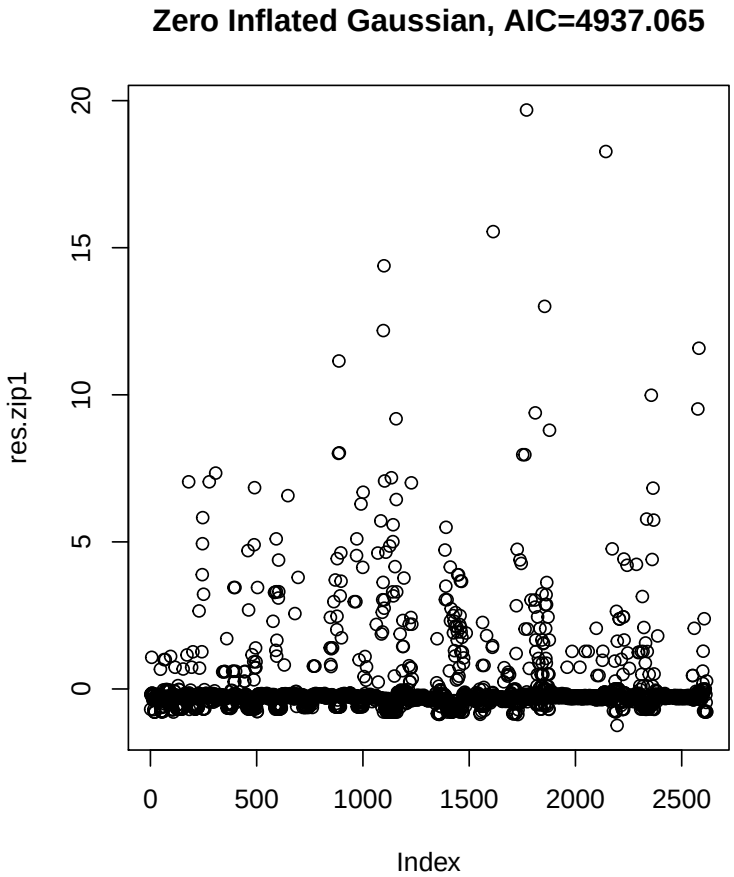
```
## Likelihood ratio test
##
## Model 1: IND ~ Quarter + Crew + L + Hook + DaS | Quarter + Crew + L +
##     Hook + offset(DaS)
## Model 2: IND ~ Quarter + Crew + L + Hook + DaS | Quarter + Crew + L +
##     Hook + offset(DaS)
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  15 -2453.5
## 2  16 -1922.5  1 1062.1  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

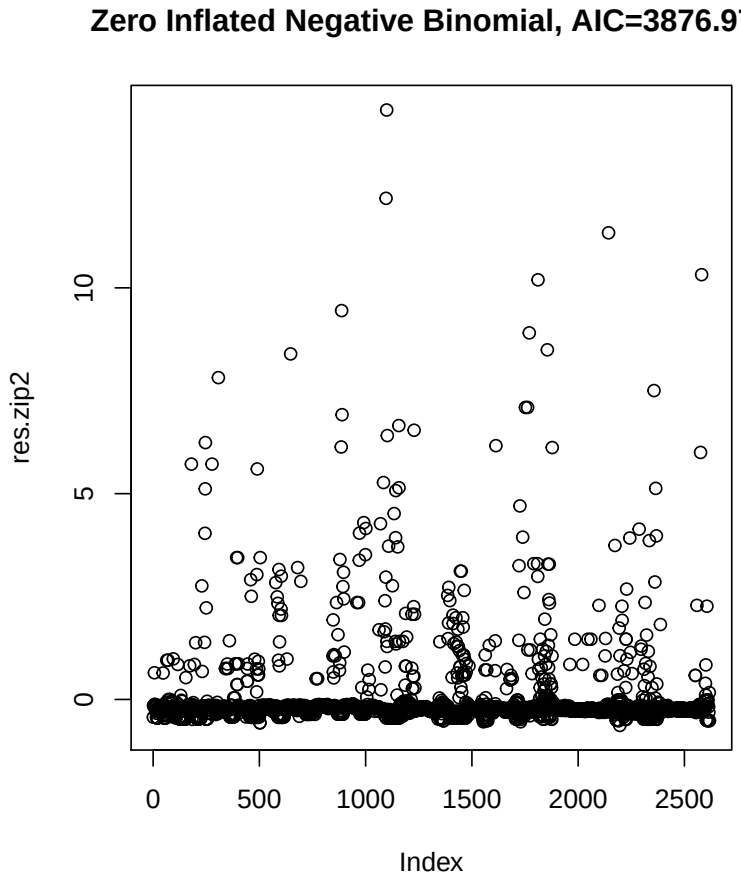
Seleksi model

```
#Hitung nilai AIC
AIC(Zip1,Nb1A)
```

```
##      df      AIC
## Zip1 15 4937.065
## Nb1A 16 3876.979
```

Seleksi model





QUIZ!

Tentukan distribusi apa yang bisa kita gunakan ketika menemukan kasus berikut ini:

1. Menguji apakah yang menyebabkan lampu jalanan menyala atau tidak setelah matahari tenggelam.
2. Menguji apakah yang menyebabkan suhu di negara tropis berubah dari nilai -4 ke +4 derajat Celcius.
3. Menguji apakah yang menyebabkan berat badan anak bayi meningkat di 3 tahun pertama umurnya.
4. Menguji apakah jumlah halaman buku fiksi dipengaruhi oleh *genre* buku tersebut.

References:

1. Gambar 2. By KendallVarent - Own work, Public Domain, https://commons.wikimedia.org/w/index.php?curid=10349551
2. Gambar 3. https://en.wikipedia.org/wiki/Q%E2%80%93Q_plot
3. Gambar 4. https://stats.idre.ucla.edu/spss/seminars/introduction-to-regression-with-spss/introreg-lesson2/
4. Gambar 5. https://uc-r.github.io/assumptions_homogeneity
5. Penjelasan mengenai variabel offset https://www.cscu.cornell.edu/news/statnews/stnews94.pdf
6. Zuur, et al. 2009. Sumber utama konten modul ini.