Penyiapan Data

Kuliah 1 – STA1382 Teknik Pembelajaran Mesin

Septian Rahardiantoro

Materi Perkuliahan

No	Materi
1	Ruang Lingkup Pembelajaran Mesin
2	Penyiapan Data
3	Regresi Linier dan Regresi Logistik serta Evaluasi
4	Metode Berbasis Pohon (CART) dan Evaluasi
5	Neural Network
6	Support Vector Machine
7	Studi Kasus (Presentasi)
UTS	

No	Materi
8	Penggerombolan Berhierarki dan Evaluasi
9	Penggerombolan Non-hierarki dan Evaluasi
10	Reduksi Dimensi dengan PCA
11	Association Rule
12	Metode Ensemble
13	Kasus: Dosen Tamu
14	Studi Kasus (Makalah)
UAS	

Rencana Perkuliahan UTS

Kuliah : setiap Hari **Senin pukul 13:00 WIB**
RK. CCR 2.02 & 2.03
Praktikum : setiap Hari **Jumat pukul 16:00 WIB**

Tanggal	Materi	Pelaksanaan	Keterangan
22 Januari 2023	Penyiapan Data	Offline	
29 Januari 2023	Ruang Lingkup Pembelajaran Mesin	Offline	
5 Februari 2023	Regresi Linier dan Regresi Logistik beserta evaluasinya	Online	Tugas Kelompok*
12 Februari 2023	Metode berbasis pohon (CART) dan evaluasinya	Offline	
19 Februari 2023	Neural Network	Offline	Tugas Individu*
26 Februari 2023	Support Vector Machine	Online	
6 Maret 2023	Studi Kasus	Offline	Presentasi Tugas Kelompok*

*teknis dan waktunya akan dijelaskan lebih lanjut

Komponen Penilaian

Komponen penilaian:

Project	: **50%**
- Kelompok (Presentasi UTS)	: 15%
- Kelompok (Paper)	: 20%
- Individu	: 15%
UTS	: **20%**
UAS	: **20%**
Praktikum	: **10%**

Outline

- Data Hilang dan Metode mengatasinya
- Pencilan dan Metode penanganannya
- Metode transformasi data

Data Hilang

- Data hilang (missing value) → merupakan suatu kejadian ketika tidak ada nilai data yang terekam/ tercatat pada pengamatan untuk suatu peubah

- Penyebab data hilang:
  - Non-respon di survei: pertanyaan sensitif, tidak tahu jawaban, pengukuran berulang
  - Kondisi pengamatan rusak
  - Kesalahan peneliti atau entri data

- Analisis statistika yang tetap digunakan pada data hilang dapat menyebabkan kesimpulan yang diperoleh tidak merepresentasikan kondisi populasi yang sebenarnya.

- Selain itu, alasan mengapa terdapat data hilang penting diketahui untuk dapat menangani data yang tersisa dengan benar.

- Jika nilai hilang secara sistematis, analisis yang dilakukan mungkin akan bias.

Jenis Mekanisme Data Hilang

1. Missing Completely at Random (MCAR).
   - Jika peristiwa yang menyebabkan hilangnya nilai data tertentu tidak bergantung pada peubah yang dapat diamati dan yang tidak dapat diamati, dan terjadi seluruhnya secara acak. Serta hilangnya nilai peubah ini tidak ada hubungannya dengan objek yang sedang dipelajari.
   - Oleh karena itu, data hilang dan data lengkap tidak memberikan hasil berbeda saat analisis. Sehingga, kasus ini dapat dikatakan hilang secara acak penuh dari data.

Jenis Mekanisme Data Hilang (2)

2. Missing at Random (MAR).
   - Data hilang tergantung dari nilai yang diketahui, dengan demikian dapat dideskripsikan oleh peubah penjelas dalam data set. Menghitung nilai "penyebab" data hilang akan menghasilkan hasil tak bias dalam analisis.

3. Missing Not at Random (MNAR).
   - Data hilang disebabkan oleh kejadian atau materi yang tidak diukur. Nilai peubah yang hilang terkait dengan alasan hilangnya peubah tersebut.

- Jika data hilang termasuk dalam kategori MCAR, analisis yang dilakukan pada data tidak bias meskipun data hilang tersebut tidak ditangani. Namun, kondisi data hilang MCAR sangat jarang untuk ditemui. Dalam kasus MCAR, data pengamatan yang hilang tidak terkait dengan peubah apa pun: dengan demikian, pengamatan dengan data yang benar-benar diamati pada dasarnya adalah sampel acak dari populasinya.

- Jika data hilang termasuk dalam kategori MAR, data yang terdapat data hilang tersebut akan berbias jika langsung dilakukan analisis. Alangkah lebih baik jika pada nilai data hilang tersebut diduga terlebih dahulu menggunakan peubah-peubah lainnya yang lengkap sebelum melakukan analisis.

- Jika data hilang termasuk dalam kategiri MNAR, kondisi ini sangat sulit untuk ditangani karena hilangnya data berdasarkan alasan diluar kendali penelitian. Sehingga perlu ditelusuri penyebab hilangnya nilai data tersebut, serta penanganan yang tepat ketika hal ini terjadi.

Ilustrasi Kasus Data Hilang (1)

- Tidak terisinya beberapa pengamatan peubah jenis kelamin pada responden suatu survei. Sedangkan peubah jenis kelamin ini tidak bergantung pada peubah lainnya dalam survei tersebut.

- **Dalam hal ini data hilang sepenuhnya secara acak (MCAR)**.

Illustrasi Kasus Data Hilang (2)

- Misalnya tidak terisinya peubah tingkat pendapatan yang notabene nya dapat diduga dari peubah lainnya, seperti jenis pekerjaan, tingkat pendidikan, usia, dan lainnya.

- **Kondisi ini merupakan ilustrasi kasus MAR**, karena data hilang pada peubah tertentu dapat diduga berdasarkan data lengkap pada peubah lainnya.

Illustrasi Kasus Data Hilang (3)

- Pada survei mengenai depresi, mekanisme MNAR dapat terjadi jika seseorang gagal mengisi surveinya karena tingkat depresi yang dideritanya.

- **Data tersebut tidak hilang secara sembarangan (MNAR)**

Implikasi Data Hilang dalam Analisis

1. ketiadaan data pengamatan akan mengurangi kekuatan statistik, uji yang dihasilkan akan cenderung untuk tidak tolak hipotesis nol
2. adanya data hilang dapat meningkatkan galat baku (standard error)
3. data yang hilang dapat menyebabkan bias dalam pendugaan parameter
4. dapat mengurangi keterwakilan sampel sehingga melemahkan generalisasi dari hasil
5. adanya data hilang dapat mempersulit analisis penelitian, karena sebagian besar metode-metode analisis statistika dirancang untuk data lengkap

Metode Mengatasi Data Hilang (Teknik Sederhana)

1. Penghapusan Kasus (Listwise or Case Deletion)
   - dengan menghilangkan/ menghapus kasus pengamatan sepenuhnya jika terdapat data hilang pada salah satu peubah dalam analisis.
   - metode penghapusan kasus dapat digunakan sebagai alternatif penanganan data hilang yang cukup optimal apabila asumsi MCAR terpenuhi dengan cukup banyak sampel yang tersedia
2. Penghapusan Berpasangan (Pairwise Deletion)
   - penghapusan pengamatan data hilang hanya dilakukan pada peubah yang akan dianalisis, yang kemudian analisis data dapat diselesaikan pada subset data tergantung nilai pengamatan mana yang hilang.
   - pendekatan ini disarankan untuk diterapkan pada data hilang dengan mekanisme MCAR dan MAR.

Metode Mengatasi Data Hilang (Teknik Sederhana) (2)

3. Penggantian dengan Rataan (Mean Subtitution)
   - pendekatan ini mengganti nilai yang hilang pada peubah tertentu dengan nilai rataan sampelnya, sehingga menjadi seolah-olah data lengkap
   - pendekatan ini cukup baik diterapkan pada data hilang dengan mekanisme MCAR dan MAR
4. Pengisian dengan Regresi (Regression Imputation)
   - Pendekatan ini mempertahankan semua kasus pengamatan dengan mengganti nilai data yang hilang dengan nilai kemungkinan yang diduga oleh informasi peubah lain yang tersedia melalui pemodelan regresi
   - Pendekatan ini baik digunakan pada data hilang dengan mekanisme MAR.

Ilustrasi

- Diketahui data karakteristik ikan, yang terdiri dari 3 peubah: tinggi, lebar, dan panjang ikan.

```
dataikan <- read.csv("data ikan.csv", header=T)
str(dataikan)
summary(dataikan)
> str(dataikan)
'data.frame':   55 obs. of  3 variables:
$ Panjang: num  30 31.2 31.1 33.5 34 34.7 34.5 35 35.1 36.2 ...
$ Tinggi : num  11.5 12.5 12.4 12.7 12.4 ...
$ Lebar  : num  4.02 4.31 4.7 NA 5.13 ...
> summary(dataikan)
Panjang          Tinggi           Lebar
Min.   :16.20   Min.   : 4.147   Min.   :2.268
1st Qu.:26.75   1st Qu.: 7.060   1st Qu.:3.820
Median :35.00   Median :13.759   Median :4.927
Mean   :33.49   Mean   :12.097   Mean   :4.781
3rd Qu.:39.35   3rd Qu.:15.496   3rd Qu.:5.580
Max.   :46.50   Max.   :18.957   Max.   :6.750
                                 NA's   :2
```

Akan dilakukan analisis regresi:

Model 1: Tinggi ~ Panjang + Lebar

Model 2: Tinggi ~ Panjang

- Penanganan dengan Teknik Sederhana

```
##penghapusan kasus (listwise or case deletion)
datahilang <- dataikan[is.na(dataikan$Lebar),]
datahilang
datalengkap <- dataikan[!is.na(dataikan$Lebar),]
head(datalengkap)
model.1 <- lm(Tinggi~.,data=datalengkap)
model.1
model.2 <- lm(Tinggi~Panjang,data=datalengkap)
model.2
```

Model 1 dan model 2 menggunakan data lengkap tanpa pengamatan ke-4 dan ke-23

```
##penghapusan berpasangan (pairwise deletion)
model.1 <- lm(Tinggi~.,data=datalengkap)
model.1
model.2 <- lm(Tinggi~Panjang,data=dataikan)
model.2
```

Model 1 menggunakan data lengkap tanpa pengamatan ke-4 dan ke-23
Model 2 menggunakan data awal

```
##penggantian dengan rataan (mean substitution)
rataan.lebar <- mean(datalengkap$Lebar)
rataan.lebar
Lebar.2 <- ifelse(is.na(dataikan$Lebar),rataan.lebar,dataikan$Lebar)
datalengkap.2 <- data.frame(dataikan[,1:2],Lebar.2)
head(datalengkap.2)
model.1 <- lm(Tinggi~.,data=datalengkap.2)
model.1
model.2 <- lm(Tinggi~Panjang,data=datalengkap.2)
model.2
```

Model 1 dan model 2 menggunakan data lengkap dengan pengamatan ke-4 dan ke-23 diganti dengan nilai rataan peubahnya

```
##pengisian dengan regresi (regression imputation)
reg <- lm(Lebar~Tinggi,data=datalengkap)
prediksi.lebar <- predict(reg,newdata=datahilang)
prediksi.lebar
Lebar.3 <- dataikan$Lebar
Lebar.3[c(4,23)] <- prediksi.lebar
datalengkap.3 <- data.frame(dataikan[,1:2],Lebar.3)
head(datalengkap.3)
model.1 <- lm(Tinggi~.,data=datalengkap.3)
model.1
model.2 <- lm(Tinggi~Panjang,data=datalengkap.3)
model.2
```

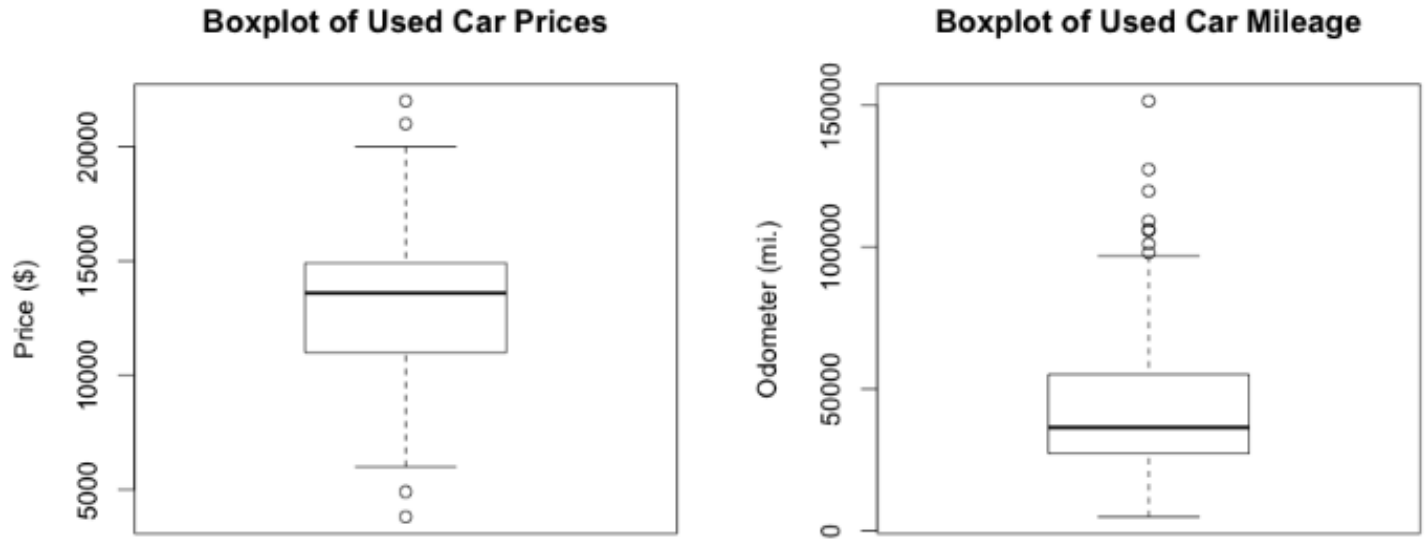
Model 1 dan model 2 menggunakan data lengkap dengan pengamatan ke-4 dan ke-23 diisi dengan nilai dugaan regresi

Pencilan (Outlier)

- Pencilan adalah titik dimana nilainya jauh dari nilai secara umum (yang diprediksi oleh model).

- Pencilan dapat muncul karena berbagai alasan, seperti pencatatan observasi yang salah selama pengumpulan data.

Ilustrasi pencilan dalam kasus univariat



- Nilai minimum dan maksimum diilustrasikan menggunakan garis yang memanjang di bawah dan di atas kotak; namun, sudah menjadi konvensi untuk hanya mengizinkan garis diperpanjang hingga minimum atau maksimum 1.5 kali IQR di bawah Q1 atau di atas Q3. Setiap nilai yang berada di luar ambang ini dianggap sebagai pencilan dan dilambangkan sebagai lingkaran atau titik.
- Misalnya, ingatlah bahwa IQR untuk peubah harga (Prices) adalah 3909 dengan Q1=10995 dan Q3=14904. Oleh karena itu, pencilan adalah nilai yang kurang dari 10995 – 1.5 * 3905 = 5137.5 atau lebih besar dari 14904 + 1.5 * 3905 = 20761.5.

Ilustrasi pencilan dalam kasus regresi linier

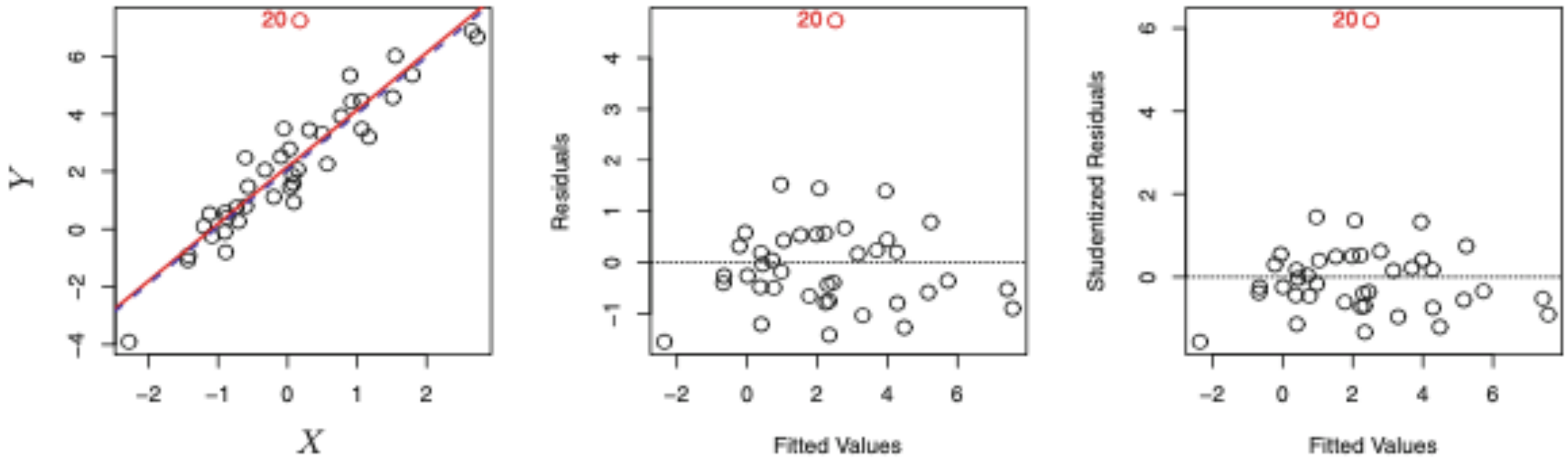


FIGURE 3.12. Left: *The least squares regression line is shown in red, and the regression line after removing the outlier is shown in blue.* Center: *The residual plot clearly identifies the outlier.* Right: *The outlier has a studentized residual of 6; typically we expect values between* -3 *and* 3.

- Dalam hal ini, menghapus pencilan memiliki efek yang kecil pada garis MKT: hal itu menyebabkan hampir tidak ada perubahan pada kemiringan, dan pengurangan intersep yang sangat kecil.

- Hal tersebut adalah tipikal untuk pencilan yang tidak memiliki nilai prediktor yang tidak biasa memiliki sedikit efek pada kuadrat terkecil.

- Namun, bahkan jika pencilan tidak memiliki banyak pengaruh pada dugaan MKT, hal itu dapat menyebabkan masalah lain.
  - Nilai JKG berubah → perubahan selang kepercayaan, p-value, R^2

- Pengamatan yang mutlak studentized residualnya lebih besar dari 3 dapat diidentifikasikan sebagai pencilan.

- Studentized Residual

Misalkan diketahui model regresi linier sebagai berikut:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

dengan

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix}$$

maka hat-matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$, dengan diagonal ke-$i$ matriks $\boldsymbol{H}$ adalah h_{ii}.

Akibatnya, studentized residual dapat dicari dengan:

$$t_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

dengan $\hat{\sigma}^2 = \frac{1}{n-m}\sum_{j=1}^{n} \hat{\varepsilon}_j^2$; m banyaknya parameter

- Jika kita yakin bahwa pencilan telah terjadi karena kesalahan dalam pengumpulan atau pencatatan data, maka salah satu solusinya adalah dengan menghilangkan observasi tersebut.

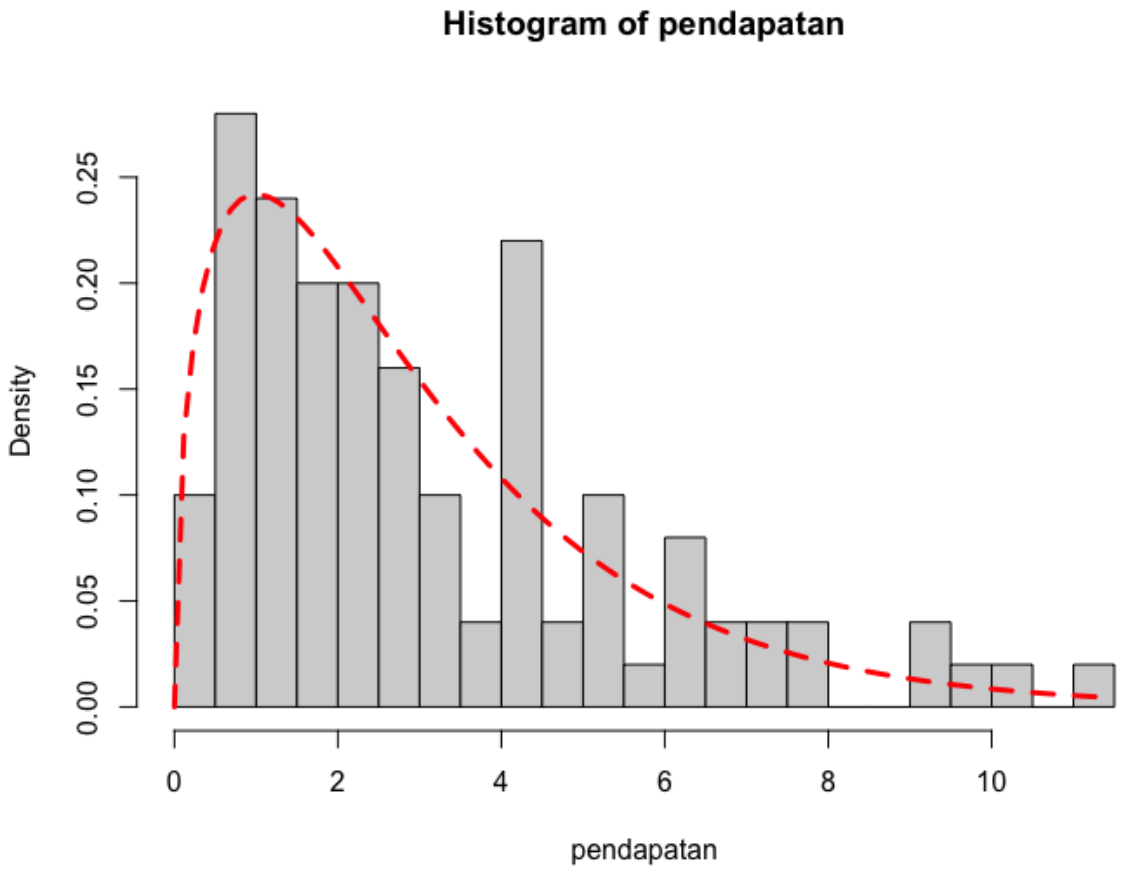
- Namun, kehati-hatian harus dilakukan, karena pencilan malah dapat menunjukkan kekurangan model, seperti prediktor yang hilang.

Transformasi Data

- Pada beberapa kasus, seringkali ditemui sebaran peubah yang diteliti tidak sesuai dengan asumsi.

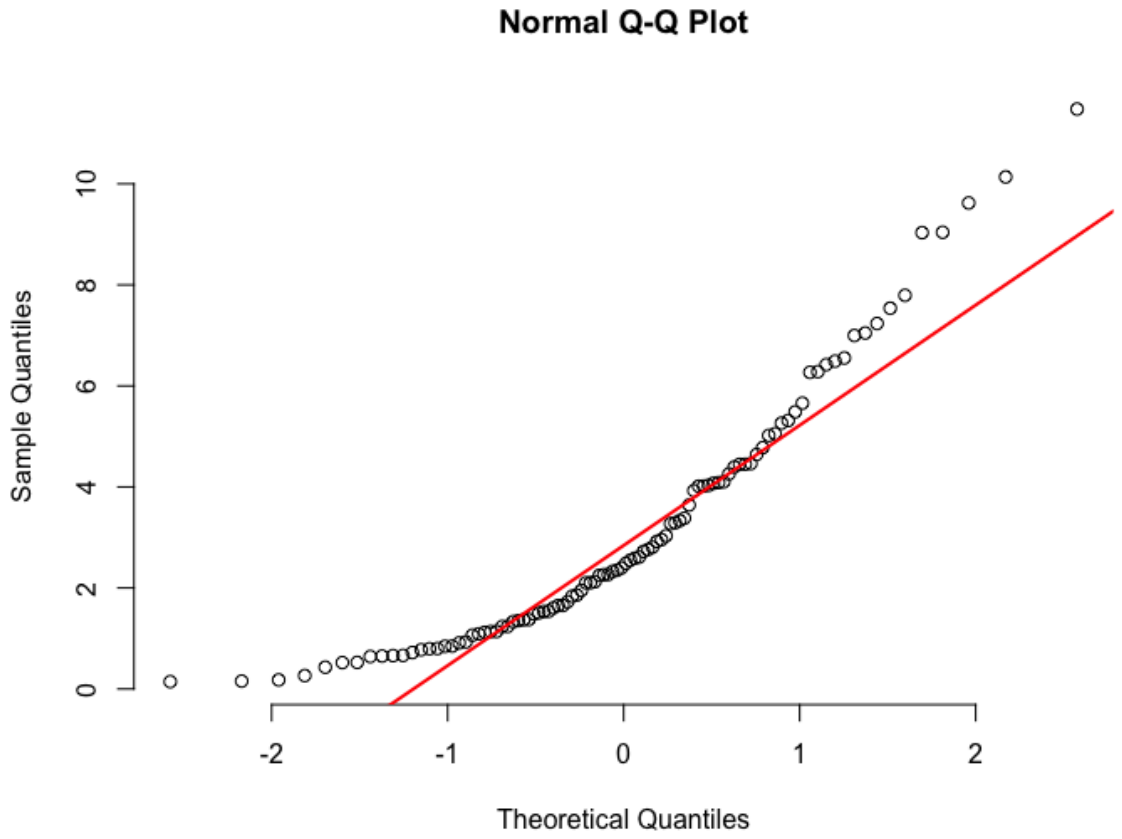
- Misalkan pada kasus peubah univariat yang mengharuskan memiliki sebaran normal yang simetris, seringkali ditemui bahwa nilai penyebarannya tidak simetris.
  - Contoh: Peubah pendapatan, ingin dilakukan uji hipotesis yang mengharuskan memiliki sebaran normal

Ilustrasi




```
set.seed(123456)
pendapatan <- rchisq(100,3)
hist(pendapatan,breaks=30,freq=F)
x <- pendapatan
curve(dchisq(x,3),lty=2,col="red",lwd=3,add=T)
```

Tidak simetri (miring ke kanan)

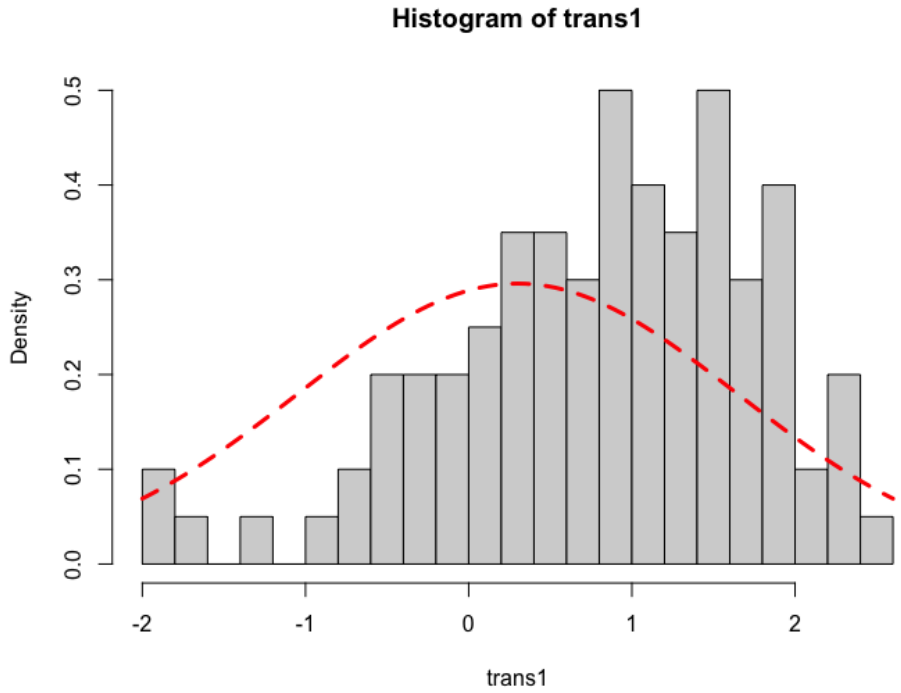


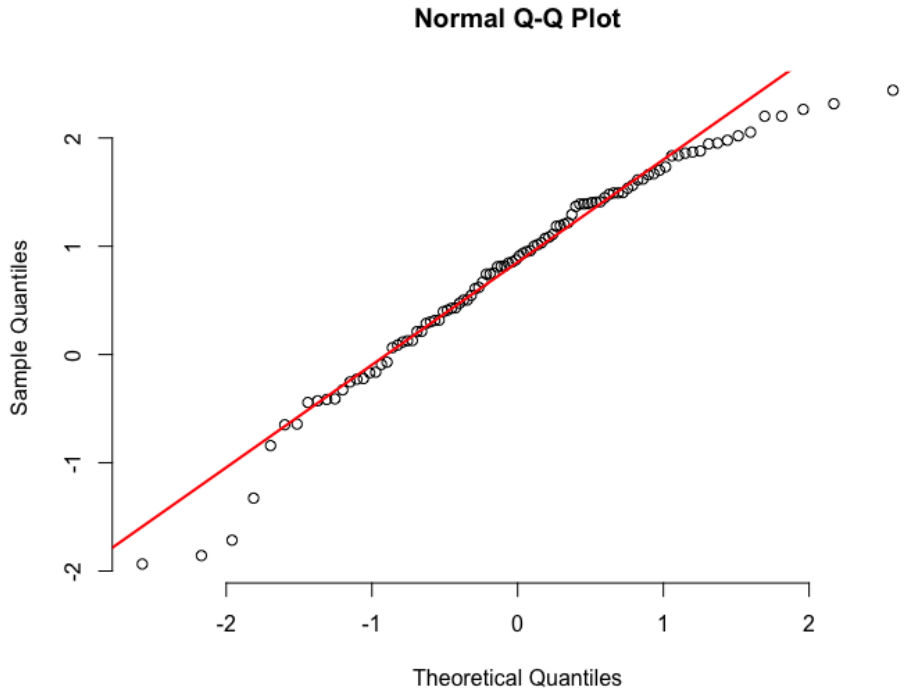

```
qqnorm(x, pch = 1, frame = FALSE)
qqline(x,col="red",lwd=2)
```

Tidak menyebar normal

Transformasi → dapat dicobakan transformasi $\ln(Y)$ atau $\sqrt{Y}$

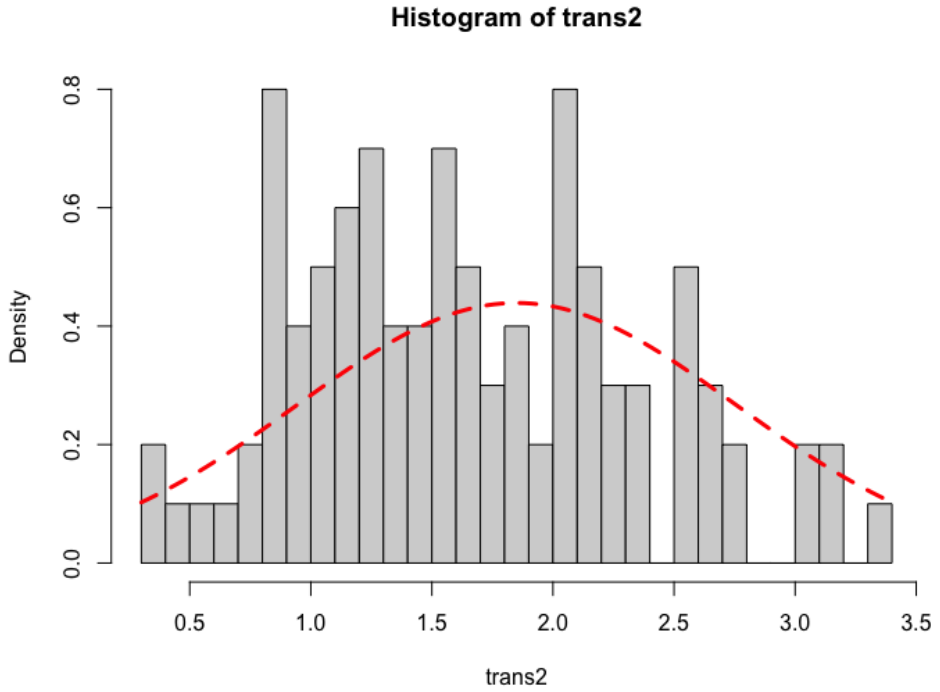
transformasi $\ln(Y)$

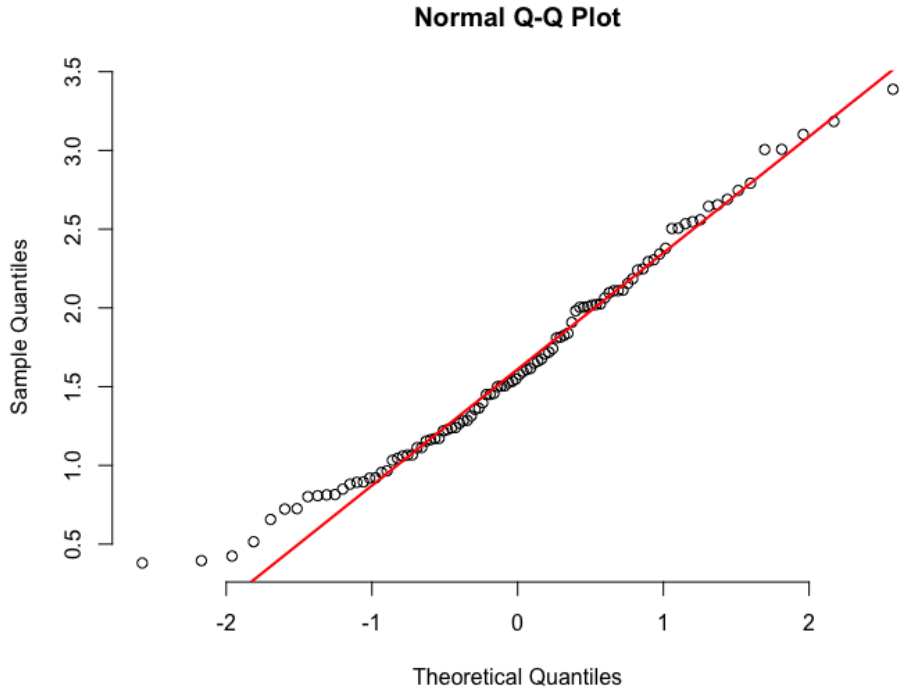





```
trans1 <- log(x)
hist(trans1,breaks=30,freq=F)
x <- trans1
curve(dnorm(x,mean(x),sd(x)),lty=2,col="red",lwd=3,add=T)
qqnorm(x, pch = 1, frame = FALSE)
qqline(x,col="red",lwd=2)
```

transformasi $\sqrt{Y}$






```
trans2 <- sqrt(pendapatan)
hist(trans2,breaks=30,freq=F)
x <- trans2
curve(dnorm(x,mean(x),sd(x)),lty=2,col="red",lwd=3,add=T)
qqnorm(x, pch = 1, frame = FALSE)
qqline(x,col="red",lwd=2)
```

Metode Box-Cox

- Transformasi Box Cox adalah transformasi peubah (peubah Y) yang tidak normal menjadi bentuk normal.

- Normalitas adalah asumsi penting untuk banyak metode statistik; jika data tidak normal, dengan menerapkan Box-Cox dapat menjalankan lebih banyak pengujian.

- Transformasi Box Cox dinamai ahli statistik George Box dan Sir David Roxbee Cox yang berkolaborasi pada makalah tahun 1964 dan mengembangkan teknik tersebut.

- Inti dari transformasi Box Cox adalah pangkat lambda (λ), yang bervariasi dari -5 hingga 5.

- Semua nilai λ dipertimbangkan dan nilai optimal untuk data yang akan dipilih; "Nilai optimal" adalah yang menghasilkan dugaan terbaik dari kurva sebaran normal.

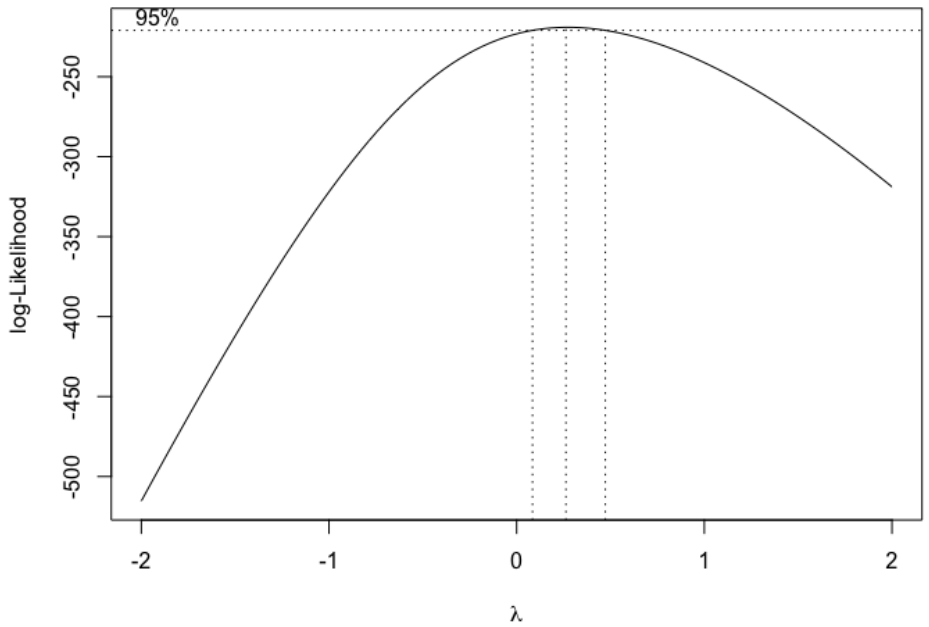
- Transformasi Y memiliki bentuk:

$$y(\lambda) = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda} & ; \lambda \neq 0 \\ \ln(y) & ; \lambda = 0 \end{cases}$$

- Namun, transformasi yang paling umum dijelaskan dalam tabel berikut:

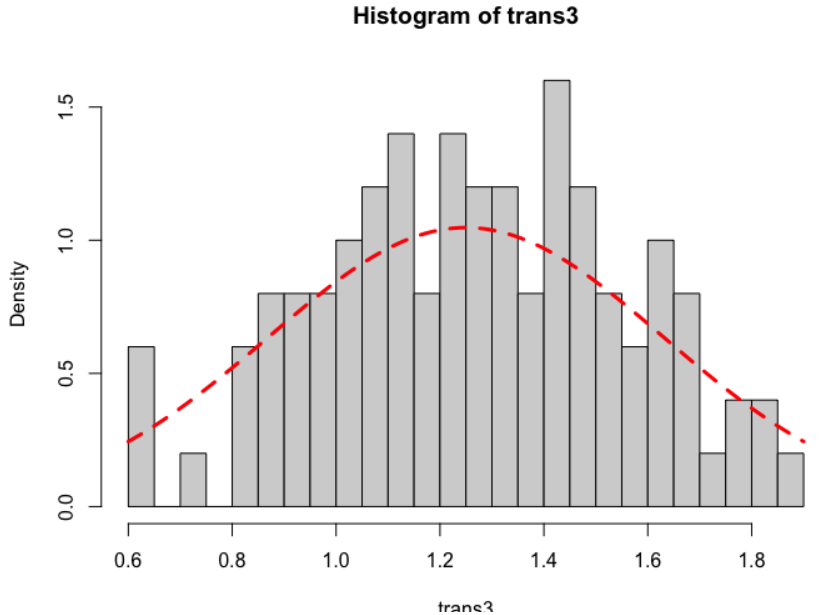
Nilai λ	Transformasi Y
-3	$Y^{-3} = {}^{1}/_{Y^3}$
-2	$Y^{-2} = {}^{1}/_{Y^2}$
-1	$Y^{-1} = {}^{1}/_{Y}$
-0.5	$Y^{-0.5} = {}^{1}/_{\sqrt{Y}}$
0	$\ln(Y)$
0.5	$Y^{0.5} = \sqrt{Y}$
1	$Y^1 = Y$
2	Y^2
3	Y^3

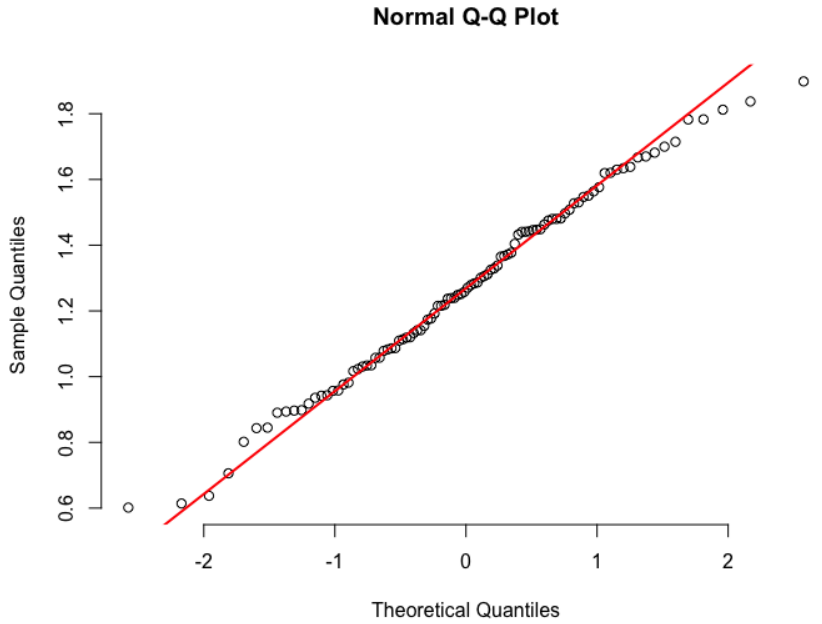
Transformasi boxcox



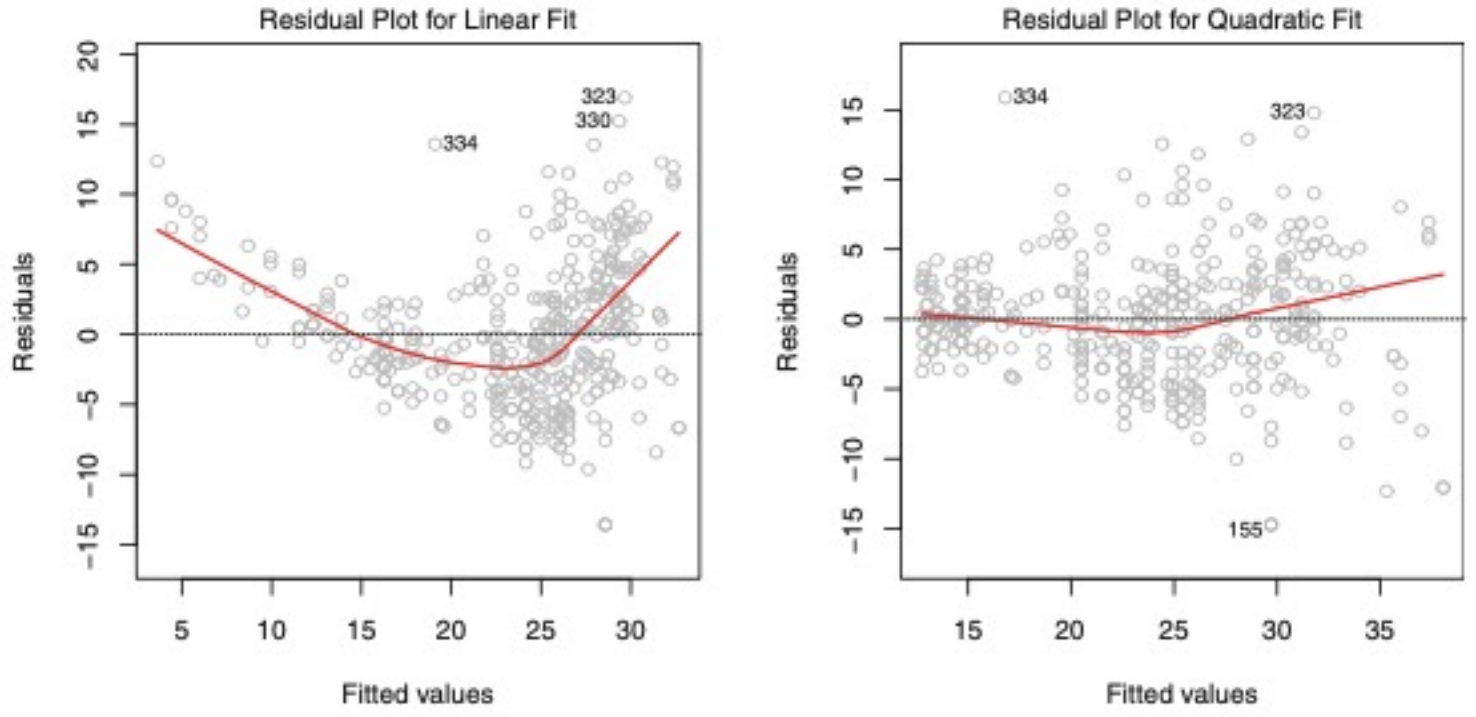

```
library(MASS)
bb <- boxcox(lm(pendapatan~1))
lambda <- bb$x[which.max(bb$y)]
lambda

trans3 <- pendapatan^(lambda)
hist(trans3,breaks=30,freq=F)
x <- trans3
curve(dnorm(x,mean(x),sd(x)),lty=2,col="red",lwd=3,add=T)
qqnorm(x, pch = 1, frame = FALSE)
qqline(x,col="red",lwd=2)
```

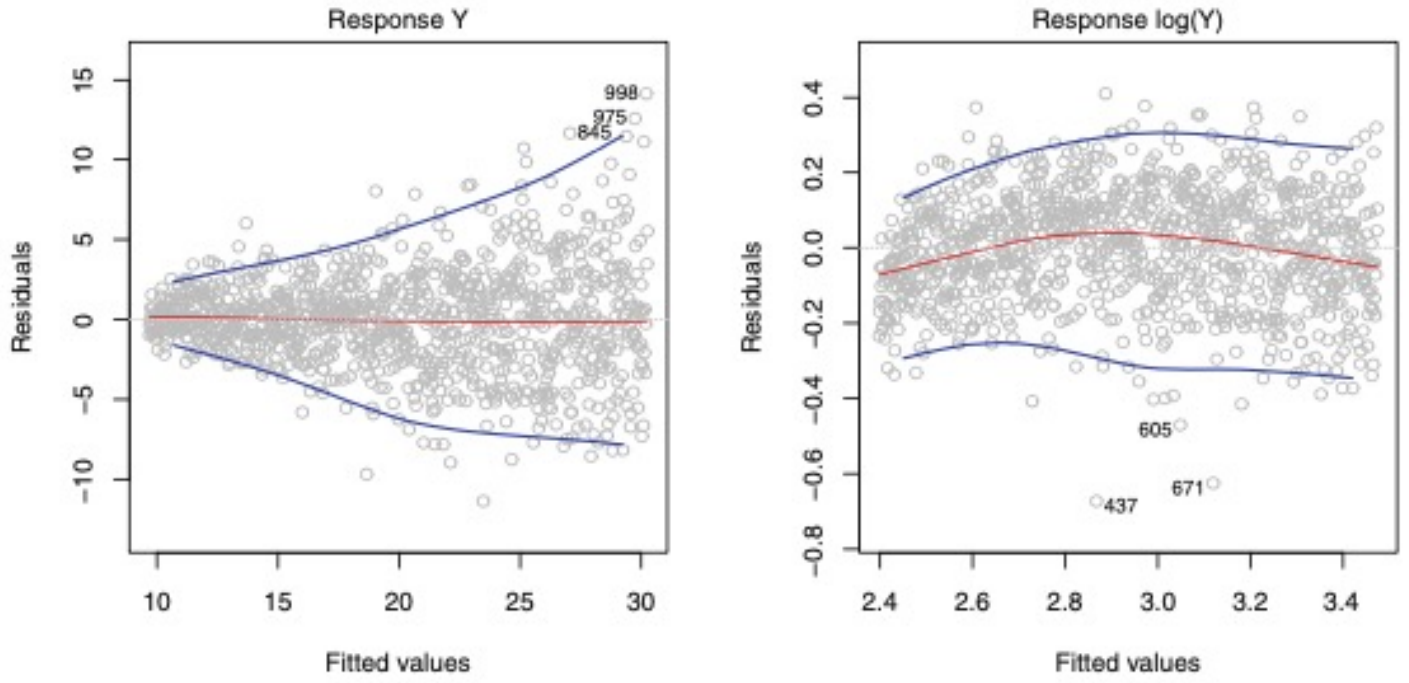




- Dalam kasus regresi, jika plot residual menunjukkan bahwa ada hubungan non-linear dalam data, maka pendekatan sederhana adalah dengan menggunakan transformasi prediktor non-linear, seperti $\ln(X)$, $\sqrt{X}$, dan X^2, dalam model regresi. Pendekatan non-linier ini dapat berupa pemodelan regresi non-linier.



- Ketika dihadapkan dengan masalah ketidakhomogenan ragam residual (heteroskedastisitas dalam kasus regresi), salah satu solusi yang mungkin adalah dengan transformasi Y menggunakan fungsi cekung seperti $\ln(Y)$ atau $\sqrt{Y}$.
- Transformasi seperti itu menghasilkan jumlah penyusutan yang lebih besar dari respons yang lebih besar, yang mengarah ke pengurangan heteroskedastisitas.



Latihan (isian singkat):

1. Seseorang tidak menghadiri tes narkoba karena orang tersebut menggunakan narkoba pada malam sebelumnya, merupakan contoh mekanisme data hilang …
2. Hilangnya pengamatan pada peubah jarak tempuh mobil pada pengumpulan data mengenai mobil bekas, yang notabenenya nilai peubah ini dapat diprediksi oleh peubah tahun mobil diproduksi, merupakan contoh mekanisme data hilang …
3. Tidak terisinya pengamatan pada peubah IPK karena terdapat kerusakan pada sistem entri data, merupakan contoh mekanisme data hilang …
4. Metode penanganan data hilang yang cenderung bersifat underestimate adalah …
5. Asumsi mekanisme data hilang yang harus dipenuhi agar metode pengisian dengan regresi valid adalah data hilang mengikuti mekanisme …

6. Metode penanganan data hilang yang valid diterapkan apabila mekanismenya MCAR dan ukuran contoh yang cukup banyak adalah …
7. Kriteria pengamatan pencilan pada boxplot adalah …
8. Berdasarkan studentized residual, pengamatan dalam regresi yang dianggap sebagai pencilan adalah …
9. Untuk nilai λ=0 pada transformasi box cox, maka transformasi yang dilakukan pada peubah Y adalah …
10. Transformasi pada peubah prediktor (X) pada model regresi dimaksudkan untuk …

Terima kasih ☺