Projek Supervised Learning Uji Asumsi

Rafly Priyantama Ramadhan Bagaskara

2025-10-06

Soal 1

Membangkitkan Data yg sisaan yg ber autokorelasi (dibuktikan dengan uji Durbin Watson).

1.Persiapan Data

Kita akan mensimulasikan model regresi sederhana dengan error yang mengikuti proses **AR(1)**:

$$\epsilon_t = \rho \epsilon_{t-1} + u_t$$

dengan $u_t \sim N(0, \sigma^2)$.

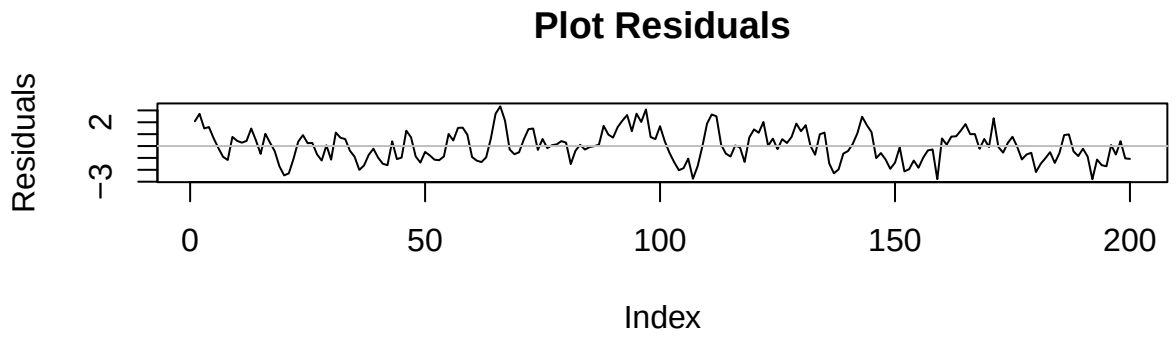
```
set.seed(123)

# Parameter simulasi
n     <- 200   # Jumlah Obs
beta0 <- 1.0   # Intercept
beta1 <- 2.5   # Koef Slope
sigma <- 1.0   # Std Dev
rho   <- 0.7   # koefisien autokorelasi

# Variabel independen
x <- rnorm(n, mean = 0, sd = 1)

# Bangkitkan error AR(1)
u <- rnorm(n, 0, sigma)
eps <- numeric(n)
eps[1] <- u[1]
for(t in 2:n) eps[t] <- rho * eps[t-1] + u[t]

# Variabel dependen
y <- beta0 + beta1 * x + eps

# Model regresi biasa
fit <- lm(y ~ x)
summary(fit)
```

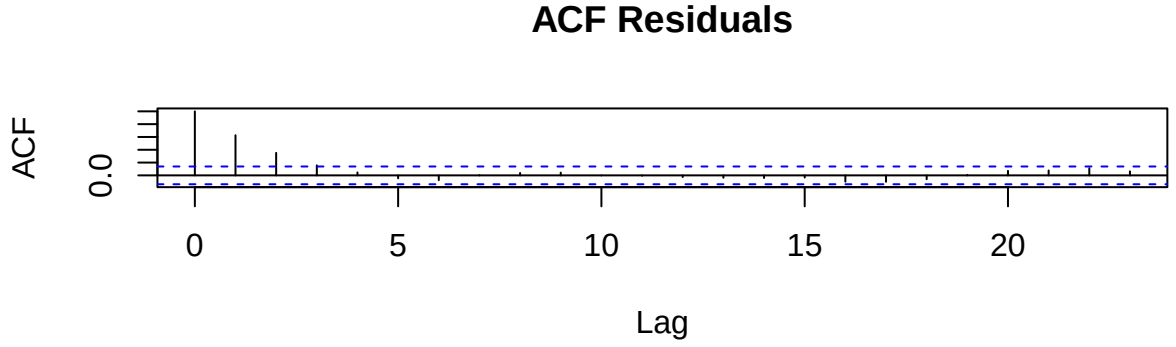
```
##
## Call:
## lm(formula = y ~ x)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7937 -0.9739 -0.0469  0.8434  3.3399
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.15262    0.09073   12.70   <2e-16 ***
## x            2.56205    0.09644   26.57   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.283 on 198 degrees of freedom
## Multiple R-squared:  0.7809, Adjusted R-squared:  0.7798
## F-statistic: 705.8 on 1 and 198 DF,  p-value: < 2.2e-16
```

2.Visualisasi dan Pemeriksaan Sisaan

```
res <- resid(fit)

par(mfrow = c(2,1))
plot(res, type = "l", main = "Plot Residuals", ylab = "Residuals")
abline(h = 0, col = "gray")
acf(res, main = "ACF Residuals")
```





```r
par(mfrow = c(1,1))
```

Dari plot di atas, terlihat bahwa residual memiliki pola berurutan (tidak acak sempurna), indikasi autokorelasi positif.

3.Uji Durbin–Watson

Uji ini digunakan untuk mendeteksi autokorelasi orde pertama antar sisaan berturut-turut.

```r
library(lmtest)

dw_result <- dwtest(fit)
dw_result
```

```
##
##  Durbin-Watson test
##
## data:  fit
## DW = 0.72828, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0
```

Soal 2

Membangkitkan data X yang saling multikolinearitas (dibuktikan dengan uji VIF).

1.Membangkitkan Data

Multikolinearitas terjadi ketika variabel independen dalam model regresi saling berkorelasi tinggi. Kondisi ini dapat menyebabkan estimasi koefisien menjadi tidak stabil dan interpretasi menjadi sulit.

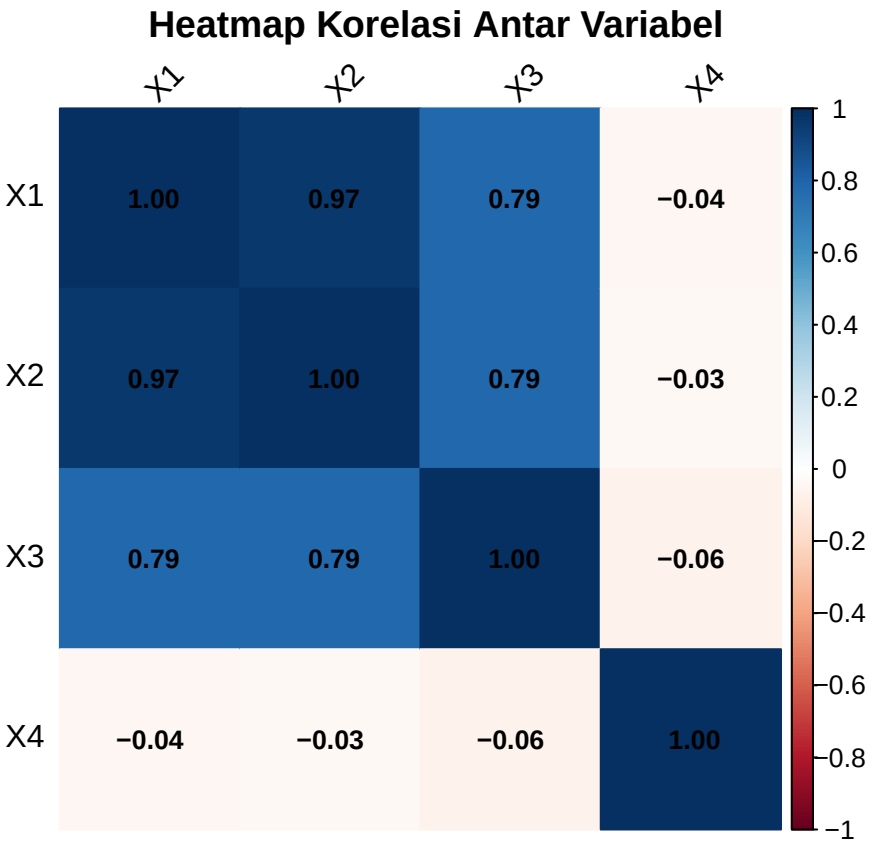
```r
library(car)      # untuk fungsi vif()
library(ggplot2)
library(GGally)   # untuk pairplot
library(corrplot) # untuk heatmap korelasi
```

```r
set.seed(123)
# Jumlah observasi
n <- 100

# Variabel independen
X1 <- rnorm(n, mean = 50, sd = 10)
X2 <- 0.8*X1 + rnorm(n, mean = 0, sd = 2)   # Sangat berkorelasi dengan X1
X3 <- 0.5*X1 + 0.3*X2 + rnorm(n, 0, 5)      # Kombinasi X1 & X2
X4 <- rnorm(n, mean = 100, sd = 20)         # Tidak berkorelasi dengan X1-X3

# Variabel dependen
Y <- 5 + 0.6*X1 + 0.4*X2 + 0.2*X3 + rnorm(n, 0, 5)

# Gabungkan ke dalam data frame
data_ml <- data.frame(Y, X1, X2, X3, X4)
head(data_ml)
```

```
##          Y       X1       X2       X3        X4
## 1 53.59158 44.39524 34.09538 43.42029  85.69516
## 2 51.64719 47.69823 38.67235 42.01288  84.94622
## 3 71.38116 65.58708 51.97628 47.06070  81.22923
## 4 59.23228 50.70508 39.86898 40.02921  78.94973
## 5 61.84444 51.29288 39.13106 35.31406  91.25681
## 6 67.94649 67.15065 53.63046 47.28323 106.62358
```

2.Visualisasi Korelasi antar Variabel

```
# Hitung matriks korelasi
cor_matrix <- cor(data_ml[, -1]) # tanpa Y

# Plot heatmap korelasi
corrplot(cor_matrix, method = "color", addCoef.col = "black",
         tl.col = "black", tl.srt = 45, number.cex = 0.8,
         title = "Heatmap Korelasi Antar Variabel", mar = c(0,0,2,0))
```



Jika dua variabel memiliki nilai korelasi mendekati ± 1 (misalnya X1 dan X2 = 0.9), maka ada indikasi multikolinearitas yang kuat.

3.Analisis Regresi Linear

```r
# Membuat model regresi
model <- lm(Y ~ X1 + X2 + X3 + X4, data = data_ml)

# Ringkasan hasil regresi
summary(model)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4, data = data_ml)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.367  -2.901   0.373   3.439  12.328
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.769961   3.768314   3.123 0.002370 **
## X1           0.796133   0.212310   3.750 0.000304 ***
## X2           0.061353   0.257862   0.238 0.812449
## X3           0.149543   0.104676   1.429 0.156389
## X4          -0.005831   0.023765  -0.245 0.806702
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.897 on 95 degrees of freedom
## Multiple R-squared:  0.7653, Adjusted R-squared:  0.7554
## F-statistic: 77.43 on 4 and 95 DF,  p-value: < 2.2e-16
```

4.Uji Multikolinearitas (VIF)

```r
# Uji VIF
vif_values <- vif(model)
vif_values
```

```
##        X1        X2        X3        X4
## 15.505573 15.281330  2.714678  1.006429
```

```r
# Interpretasi sederhana
if (any(vif_values > 10)) {
  cat("\nTerdapat multikolinearitas tinggi (VIF > 10)\n")
} else if (any(vif_values > 5)) {
  cat("\nAda indikasi multikolinearitas sedang (VIF > 5)\n")
} else {
  cat("\nTidak ada masalah multikolinearitas\n")
}
```

```
##
## Terdapat multikolinearitas tinggi (VIF > 10)
```

5.Kesimpulan

- Dari hasil heatmap, terlihat bahwa beberapa variabel seperti X1 dan X2 saling berkorelasi tinggi.
- Hasil VIF = nilai VIF untuk X1 dan X2 > 10. Artinya terdapat multikolinearitas kuat di antara variabel independen.

Soal 3

Membangkitkan data dengan sisaan Normal baku (0,1) , dibuktikan dengan uji kenormalan.

1.Membangkitkan Data

kita mulai dengan membangkitkan data x dan error $\epsilon \sim N(0,1)$, lalu membentuk $y = \beta_0 + \beta_1 x + \epsilon$.

```
# Parameter dan jumlah sampel
n <- 200
beta0 <- 1.5
beta1 <- 2.0

# Variabel prediktor (X)
x <- runif(n, -3, 3)

# Error (sisaan sesungguhnya) ~ Normal(0,1)
eps <- rnorm(n, mean = 0, sd = 1)

# Variabel respon (Y)
y <- beta0 + beta1 * x + eps

# Data frame
df <- data.frame(x = x, y = y)
head(df)
```

```
##            x          y
## 1 -1.3582636 -0.1425149
## 2  0.5632016  2.5990562
## 3 -2.0388911 -2.6111126
## 4  2.1205814  4.2250952
## 5  2.0864349  6.4632552
## 6 -0.1326791  1.0239076
```

2.Membentuk Model Regresi dan Mengambil Residu

```
# Membentuk model regresi linear
fit <- lm(y ~ x, data = df)

# Mengambil residu dari model
resids <- residuals(fit)

# Ringkasan model
summary(fit)
```

```
##
## Call:
## lm(formula = y ~ x, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.49478 -0.70978  0.08427  0.73223  2.75478
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.47416    0.07250   20.33   <2e-16 ***
## x            1.95062    0.04456   43.78   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.024 on 198 degrees of freedom
## Multiple R-squared:  0.9064, Adjusted R-squared:  0.9059
## F-statistic:  1917 on 1 and 198 DF,  p-value: < 2.2e-16
```

3.Uji Kenormalan Sisaan

Kita uji menggunakan **Shapiro–Wilk**.

```
# Shapiro-Wilk Test
shapiro_result <- shapiro.test(resids)
shapiro_result
```

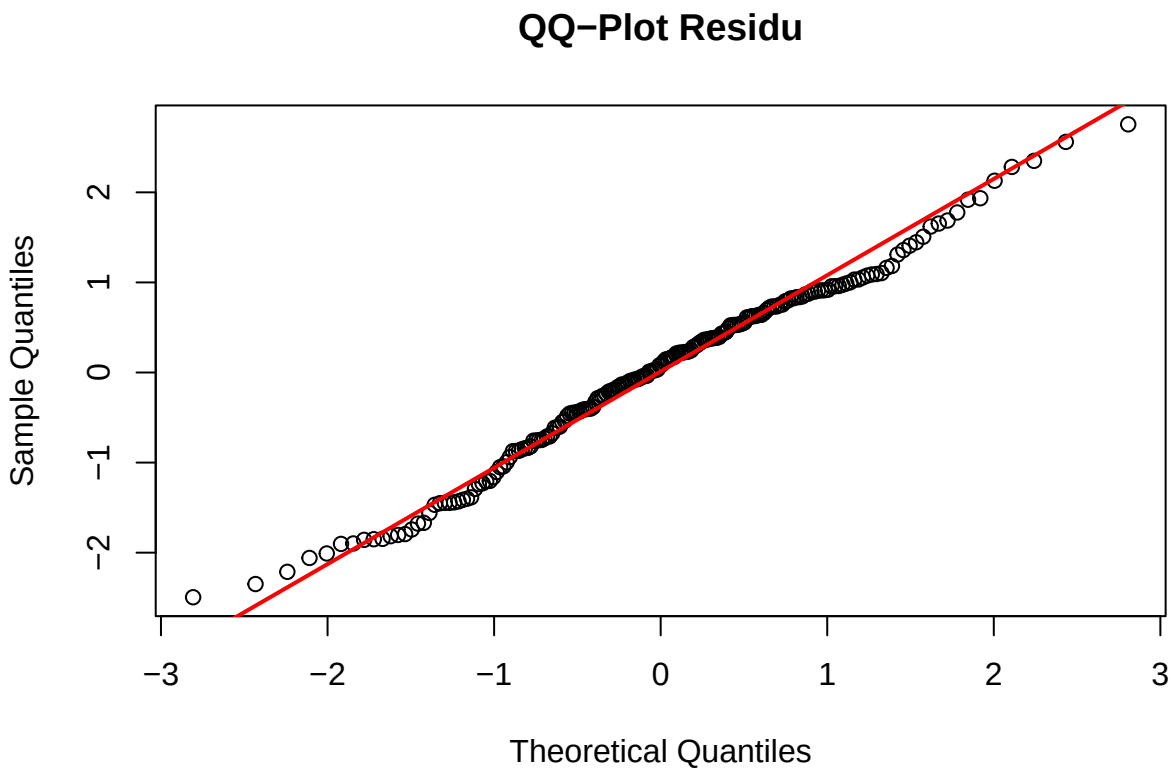
```
##
##  Shapiro-Wilk normality test
##
## data:  resids
## W = 0.98892, p-value = 0.1237
```

Interpretasi:

- Jika $p\text{-}value > 0.05 \rightarrow$ sisaan berdistribusi normal.
- Jika $p\text{-}value \leq 0.05 \rightarrow$ sisaan tidak berdistribusi normal.

4.Visualisasi Distribusi Sisaan

```
qqnorm(resids, main = "QQ-Plot Residu")
qqline(resids, col = "red", lwd = 2)
```



QQ-plot

5.Kesimpulan

```
## Berdasarkan:
## • Uji: 0.1237 > 0.05
## • QQ-Plot: Pola titik mengikuti garis normal
## Statistik Deskriptif Residual:
## • Mean = 0
## • Standar Deviasi = 1.0215
## • Distribusi: N( 0 , 1.02 ²)
```

Soal 4

Membangkitkan data dengan sisaan tidak homogen (dibuktikan dengan uji Breusch-Pagan).

1.Membangkitkan Data

Kita buat data regresi di mana varians error meningkat seiring nilai X.

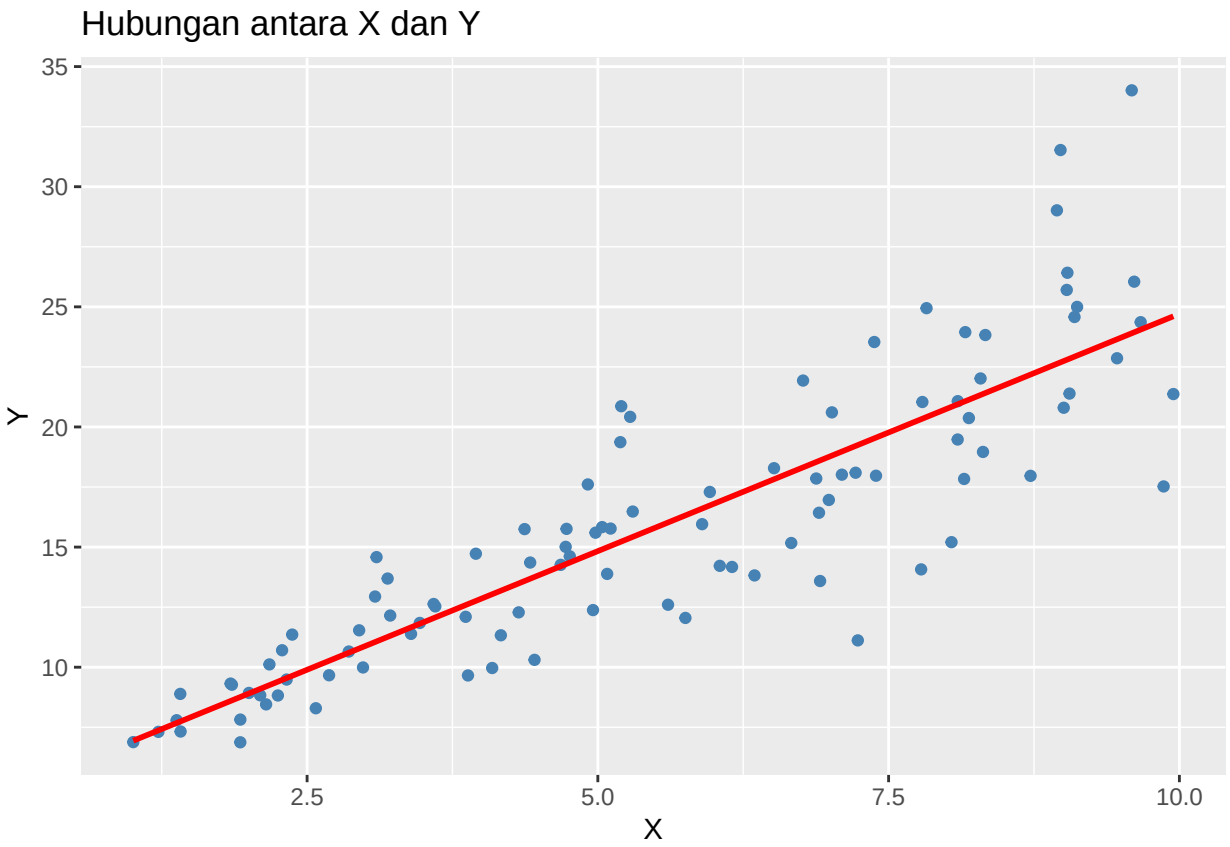
```
set.seed(123)

# Jumlah data
n <- 100

# Variabel independen
x <- runif(n, 1, 10)

# Sisaan tidak homogen: varians tergantung pada X
error <- rnorm(n, mean = 0, sd = x * 0.5)

# Variabel dependen (model linier)
y <- 5 + 2 * x + error

# Gabungkan menjadi data frame
data_hetero <- data.frame(x, y)
head(data_hetero)
```

```
##          x         y
## 1 3.588198 12.630874
## 2 8.094746 21.073953
## 3 4.680792 14.261251
## 4 8.947157 29.016863
## 5 9.464206 22.860040
## 6 1.410008  8.889135
```

2.Visualisasi Data

```
ggplot(data_hetero, aes(x = x, y = y)) +
  geom_point(color = "steelblue") +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Hubungan antara X dan Y",
       x = "X", y = "Y")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



3.Model Regresi Linear

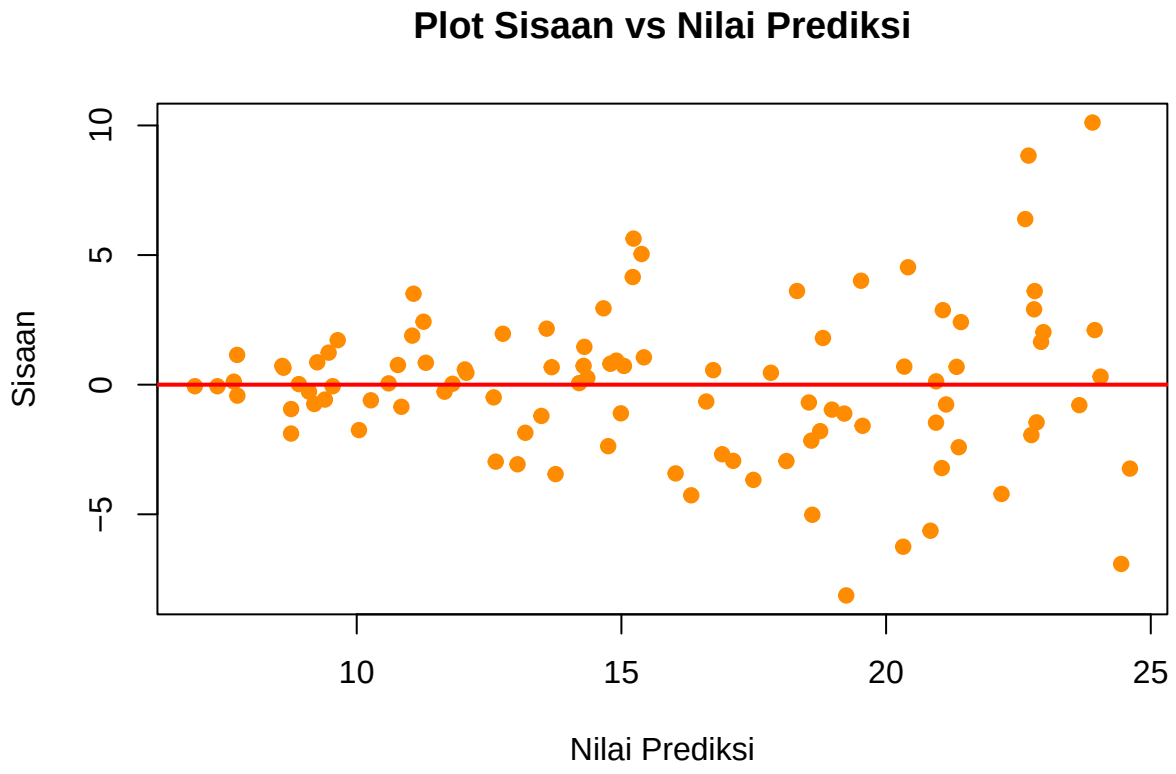
```
model <- lm(y ~ x, data = data_hetero)
summary(model)
```

```
##
## Call:
## lm(formula = y ~ x, data = data_hetero)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.1305 -1.6273  0.0422  1.2916 10.1121
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.9557     0.6977   7.103 1.97e-10 ***
## x             1.9753     0.1153  17.132  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.943 on 98 degrees of freedom
## Multiple R-squared:  0.7497, Adjusted R-squared:  0.7471
## F-statistic: 293.5 on 1 and 98 DF,  p-value: < 2.2e-16
```

4.Plot Sisaan vs Nilai Prediksi

Jika terdapat pola “kipas” atau “corong” pada plot sisaan, itu indikasi heteroskedastisitas.

```
plot(model$fitted.values, resid(model),
     main = "Plot Sisaan vs Nilai Prediksi",
     xlab = "Nilai Prediksi", ylab = "Sisaan",
     pch = 19, col = "darkorange")
abline(h = 0, col = "red", lwd = 2)
```



5.Uji Breusch-Pagan (BP Test)

```
bp_result <- bptest(model)
bp_result
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 16.153, df = 1, p-value = 5.842e-05
```

Jika p-value < 0.05, maka sisaan tidak homogen (terjadi heteroskedastisitas).

6.Kesimpulan

```
## Berdasarkan bukti
```

```
## • Uji  Breusch-Pagan: p-value = 5.84e-05
```

```
## • Plot Residual: Pola kipas yang terlihat jelas
```