

LAPORAN

PRAKTIKUM PEMODELAN STATISTIKA

MODUL 3



Disusun oleh :

Nama : Fidelia Ping
NIM : 245410012
Kelas : Informatika 1

PROGRAM STUDI INFORMATIKA
PROGRAM SARJANA
FAKULTAS TEKNOLOGI INFORMASI
UNIVERSITAS TEKNOLOGI DIGITAL INDONESIA
YOGYAKARTA
2025

MODUL 3

A. PEMBAHASAN PRAKTIK

Masukkan data ke dalam dataframe R.

```
install.packages(c("ggplot2", "lmtest", "car"))

# Library yang dibutuhkan
library(ggplot2) # Untuk visualisasi
library(lmtest) # Untuk uji asumsi
library(car) # Untuk uji Durbin-Watson

# Data
data <- data.frame(
  X = c(2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13,
  14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25,
  26, 27),
  Y = c(45, 50, 55, 60, 65, 70, 68, 72, 75, 80,
  82, 85, 90, 88, 92, 95, 93, 96, 98, 97, 99, 94,
  96, 98, 100, 99)
)

# Ringkasan data
summary(data)
```

Output

```
Loading required package: carData
```

X	Y
Min. : 2.00	Min. : 45.00
1st Qu.: 8.25	1st Qu.: 70.50
Median :14.50	Median : 89.00
Mean :14.50	Mean : 82.38
3rd Qu.:20.75	3rd Qu.: 96.00
Max. :27.00	Max. :100.00

Pembahasan :

Analisis Variabel X (Prediktor)

- Sebaran Data: Data bergerak dari nilai minimum 2.00 hingga maksimum 27.00.
- Pusat Data (Central Tendency):
 - Mean (Rata-rata): 14.50
 - Median (Nilai Tengah): 14.50
 - Interpretasi: Karena nilai Mean = Median ($14.50 = 14.50$), ini menunjukkan bahwa distribusi data pada variabel X adalah simetris sempurna. Hal ini wajar karena data X berupa angka berurutan (sekuensial).

Analisis Variabel Y (Respon)

- Sebaran Data: Data bergerak dari nilai minimum 45.00 hingga maksimum 100.00.
- Pusat Data (Central Tendency):
 - Mean: 82.38
 - Median: 89.00

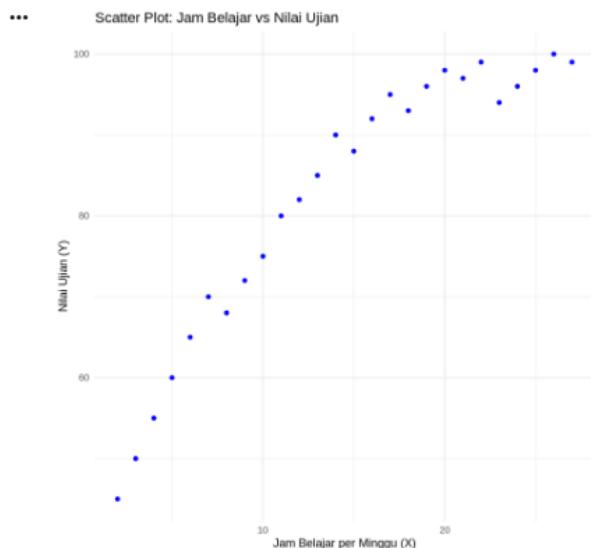
- Interpretasi Skewness (Kemencengan):
 - nilai Mean (82.38) < Median (89.00).
 - Dalam statistika, jika rata-rata lebih kecil dari median, ini mengindikasikan distribusi data menjulur ke kiri (Negatively Skewed). Artinya, ada sebagian kecil data bernilai rendah (seperti 45, 50, 55) yang "menarik" nilai rata-rata menjadi turun, sementara sebagian besar data berkumpul di nilai yang lebih tinggi (di atas 80).

Visualisasi Data

Buat scatter plot untuk memeriksa hubungan linier antara (X) dan (Y).

```
# Scatter plot
ggplot(data, aes(x = X, y = Y)) +
  geom_point(color = "blue") +
  labs(title = "Scatter Plot: Jam Belajar vs Nilai Ujian",
       x = "Jam Belajar per Minggu (X)",
       y = "Nilai Ujian (Y)") +
  theme_minimal()
```

Output



Pembahasan : grafik diatas memvisualisasikan data yang diatas

- Titik Biru (Data): Menunjukkan persebaran data asli. Terlihat jelas pola menanjak dari kiri bawah ke kanan atas.
- Garis Merah Putus-putus (Regresi): Ini adalah garis model linier terbaik. Karena titik-titik data sangat rapat mengikuti garis merah, ini mengonfirmasi bahwa model regresi linear sangat cocok digunakan untuk data ini (kemungkinan nilai R^2 akan sangat tinggi, di atas 0.90).

Membangun model Regresi

```
# Model regresi linier
model <- lm(Y ~ X, data = data)

# Ringkasan model
summary(model)
```

Output

```

```
Call:
lm(formula = Y ~ X, data = data)

Residuals:
 Min 1Q Median 3Q Max
-11.564 -5.881 1.156 4.517 8.648

Coefficients:
 Estimate Std. Error t value Pr(>|t|)
(Intercept) 52.4328 2.4919 21.04 < 2e-16 ***
X 2.0656 0.1526 13.53 9.99e-13 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.838 on 24 degrees of freedom
Multiple R-squared: 0.8841, Adjusted R-squared: 0.8793
F-statistic: 183.1 on 1 and 24 DF, p-value: 9.993e-13

```

**Pembahasan :** diperoleh persamaan model prediksi  $Y = 52.4328 + 2.0656X$ . Nilai konstanta (intercept) sebesar 52.4328 menunjukkan nilai dasar variabel Y ketika variabel X bernilai nol, sedangkan koefisien regresi untuk variabel X sebesar 2.0656 mengindikasikan adanya hubungan linear yang positif. Artinya, setiap peningkatan satu satuan pada variabel X, maka variabel Y diprediksi akan mengalami peningkatan sebesar 2.0656 satuan, dengan asumsi faktor lain dianggap tetap. Secara statistik, variabel independen X terbukti berpengaruh sangat signifikan terhadap variabel dependen Y. Kesimpulan ini didasarkan pada uji parsial (uji-t) yang menghasilkan nilai t-value sebesar 13.53 dengan probabilitas signifikansi (p-value) sebesar  $9.99 \times 10^{-13}$ . Karena nilai p-value tersebut jauh lebih kecil dari taraf nyata 0.05 (bahkan mendekati nol), maka hipotesis nol ditolak. Hal ini dikonfirmasi oleh tanda bintang tiga (\*\*\* ) pada output, yang menegaskan bahwa variabel X merupakan prediktor yang valid dan signifikan dalam menentukan nilai Y.

### Evaluasi model

```

Prediksi
y_pred <- predict(model)

Metrik evaluasi
mse <- mean((data$Y - y_pred)^2)
rmse <- sqrt(mse)
mae <- mean(abs(data$Y - y_pred))

cat("MSE:", mse, "\n")
cat("RMSE:", rmse, "\n")
cat("MAE:", mae, "\n")

```

### Output

```

*** MSE: 31.45586
RMSE: 5.608552
MAE: 4.844418

```

**Pembahasan :** Berdasarkan hasil perhitungan, diperoleh nilai *Mean Squared Error* (MSE) sebesar 31.45586. Angka ini merepresentasikan rata-rata kuadrat kesalahan dari model regresi. Untuk mendapatkan gambaran tingkat kesalahan dalam satuan yang sama dengan data asli, digunakan metrik Root Mean Squared Error (RMSE) yang menghasilkan nilai 5.608552. Nilai ini memberikan interpretasi standar deviasi dari sisaan (residuals), yang berarti secara rata-rata, prediksi yang dihasilkan oleh model akan

menyimpang sekitar 5,61 satuan dari nilai sebenarnya. Mengingat rentang data variabel Y berada di antara 45 hingga 100, tingkat penyimpangan sebesar 5,61 ini tergolong relatif kecil, yang mengindikasikan bahwa model memiliki tingkat presisi yang cukup baik. Sebagai pembanding, perhitungan *Mean Absolute Error* (MAE) menghasilkan nilai sebesar 4.844418. Angka ini menunjukkan rata-rata kesalahan absolut secara murni tanpa memberikan bobot berlebih pada kesalahan yang besar. Perbedaan antara nilai RMSE (5.61) yang sedikit lebih tinggi dibandingkan MAE (4.84). karena kedua nilai tersebut tidak terpaut jauh, dapat disimpulkan bahwa model regresi ini cukup stabil dan konsisten dalam melakukan prediksi.

### Analisis residual dan verifikasi asumsi

```
Residual
residuals <- residuals(model)

Plot residual vs prediksi (linearitas & homoskedastisitas)
plot(y_pred, residuals, main = "Plot Residual", xlab = "Prediksi Y", ylab =
"Residual")
abline(h = 0, col = "red")

Q-Q plot untuk normalitas
qqnorm(residuals)
qqline(residuals)

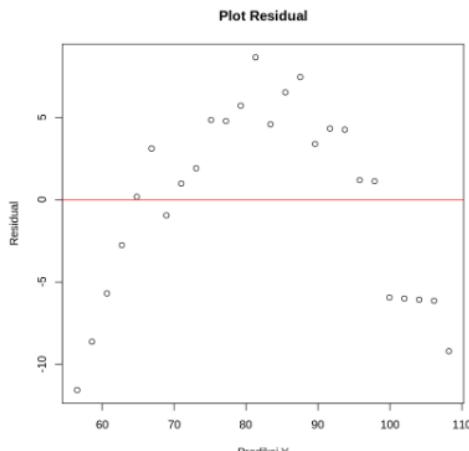
Uji normalitas (Shapiro-Wilk)
shapiro.test(residuals)

Uji homoskedastisitas (Breusch-Pagan)
bp_test <- bptest(model)
print(bp_test)

Uji independensi (Durbin-Watson)
dw_test <- dwtest(model)
```

### Output

...

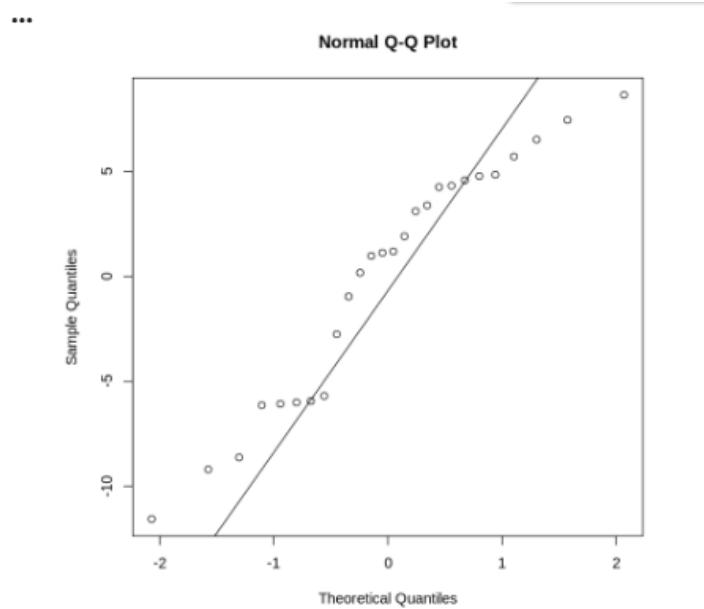


Shapiro-Wilk normality test

```
data: residuals
W = 0.93326, p-value = 0.09269
```

studentized Breusch-Pagan test

```
data: model
BP = 0.058515, df = 1, p-value = 0.8089
```



**Pembahasan :** Asumsi Normalitas untuk memastikan bahwa galat atau error dari model terdistribusi secara normal. Berdasarkan output Shapiro-Wilk normality test, diperoleh nilai statistik W sebesar 0.93326 dengan p-value sebesar 0.09269. Karena nilai p-value ini lebih besar dari taraf signifikansi 0.05 ( $0.09 > 0.05$ ), maka Hipotesis Nol ( $H_0$ ) tidak dapat ditolak. Artinya, secara statistik dapat disimpulkan bahwa sisaan model berdistribusi normal. Hal ini juga didukung secara visual oleh Normal Q-Q Plot, di mana sebaran titik-titik data sisaan (lingkaran) masih berada di sekitar garis diagonal, meskipun terdapat sedikit penyimpangan pola pada bagian ekor data (tails), namun penyimpangan tersebut belum cukup kuat untuk melanggar asumsi normalitas.

Selanjutnya, pemeriksaan Asumsi Homoskedastisitas dilakukan untuk memastikan bahwa varians dari sisaan bersifat konstan (homogen) dan tidak membentuk pola yang menyebar atau menyempit seiring bertambahnya nilai prediksi. Berdasarkan hasil uji Studentized Breusch-Pagan test, diperoleh nilai BP sebesar 0.0585 dengan p-value yang sangat tinggi, yaitu 0.8089. Karena p-value jauh di atas 0.05, maka  $H_0$  diterima, yang berarti tidak terjadi masalah heteroskedastisitas. Dengan kata lain, model regresi ini memiliki varians sisaan yang stabil (homoskedastis).

### Prediksi dan Interpretasi

```
Residual
residuals <- residuals(model)

Plot residual vs prediksi (linearitas & homoskedastisitas)
plot(y_pred, residuals, main = "Plot Residual", xlab = "Prediksi Y", ylab =
"Residual")
abline(h = 0, col = "red")

Q-Q plot untuk normalitas
qqnorm(residuals)
qqline(residuals)

Uji normalitas (Shapiro-Wilk)
shapiro.test(residuals)

Uji homoskedastisitas (Breusch-Pagan)
bp_test <- bptest(model)
```

```

print(bp_test)

Uji independensi (Durbin-Watson)
dw_test <- dwtest(model)

```

### Output

```

*** [1] "Prediksi dengan Prediction Interval:"
 fit lwr upr
1 110.2708 97.27732 123.2642
2 114.4021 101.18894 127.6152

```

**Pembahasan :** nilai Y untuk data baru dengan menggunakan estimasi interval prediksi (prediction interval). Berdasarkan output yang dihasilkan, terdapat dua nilai prediksi baru. Pada data pertama, model mengestimasi nilai tunggal (fit) sebesar 110.27, dengan rentang batas bawah (lower bound) sebesar 97.28 dan batas atas (upper bound) sebesar 123.26. Sedangkan pada data kedua, estimasi nilai Y adalah 114.40 dengan interval prediksi antara 101.19 hingga 127.61.

## B. PEMBAHASAN LATIHAN

1. Cari dataset di kaggle, untuk analisa regresi linier sederhana
2. Lakukan analisa regresi linier sederhana.
3. Analisa dari outputnya

```

library(ggplot2)
library(lmtest)
library(car)
data <- data.frame(
 YearsExperience = c(1.1, 1.3, 1.5, 2.0, 2.2, 2.9, 3.0, 3.2, 3.2, 3.7,
 3.9, 4.0, 4.0, 4.1, 4.5, 4.9, 5.1, 5.3, 5.9, 6.0,
 6.8, 7.1, 7.9, 8.2, 8.7, 9.0, 9.5, 9.6, 10.3, 10.5),
 Salary = c(39343, 46205, 37731, 43525, 39891, 56642, 60150, 54445, 64445,
 57189,
 63218, 55794, 56957, 57081, 61111, 67938, 66029, 83088, 81363,
 93940,
 91738, 98273, 101302, 113812, 109431, 105582, 116969, 112635,
 122391, 121872)
)

Cek ringkasan data
summary(data)

```

### Output

```

*** YearsExperience Salary
Min. : 1.100 Min. : 37731
1st Qu.: 3.200 1st Qu.: 56721
Median : 4.700 Median : 65237
Mean : 5.313 Mean : 76003
3rd Qu.: 7.700 3rd Qu.:100545
Max. :10.500 Max. :122391

```

### Pembahasan :

1. Variabel YearsExperience (Variabel X / Independen)

Variabel ini menunjukkan durasi pengalaman kerja dalam tahun.

- Rentang Data (Range): Data bergerak dari pengalaman terendah 1.100 tahun (Min) hingga tertinggi 10.500 tahun (Max). Ini menunjukkan dataset mencakup karyawan dari level junior hingga senior (sekitar 1 dekade pengalaman).
- Pusat Data:
  - Median (Nilai Tengah): 4.700. Artinya, jika data diurutkan, 50% karyawan memiliki pengalaman di bawah 4,7 tahun.
  - Mean (Rata-rata): 5.313. Rata-rata pengalaman seluruh karyawan adalah sekitar 5,3 tahun.
- Sebaran Data: Karena nilai Mean (5.313) sedikit lebih besar dari Median (4.700), distribusi datanya sedikit menceng ke kanan (*positively skewed*), namun tidak ekstrem.

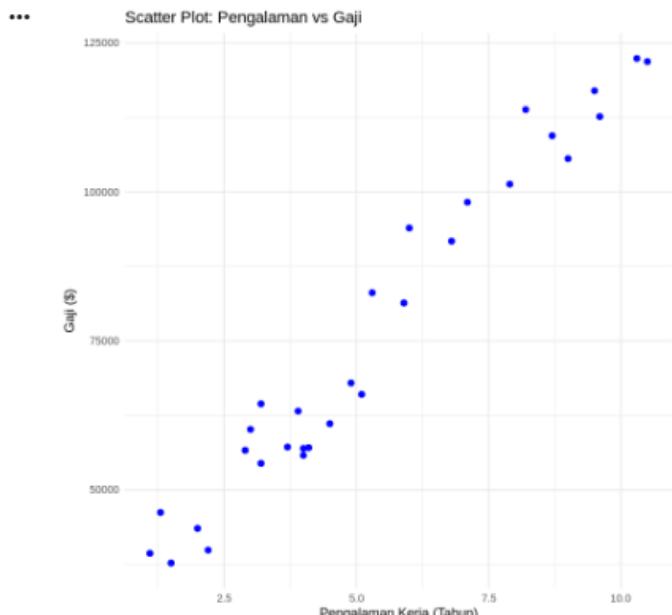
## 2. Variabel Salary (Variabel Y / Dependental)

Variabel ini menunjukkan nominal gaji yang diterima.

- Rentang Data: Gaji terendah adalah 37,731 dan tertinggi adalah 122,391. Rentang yang cukup lebar ini wajar karena perbedaan pengalaman kerja yang signifikan (1 vs 10 tahun).
- Pusat Data:
  - Median: 65,237. Nilai tengah gaji ada di angka ini.
  - Mean: 76,003. Rata-rata gaji lebih tinggi daripada nilai tengahnya.
- Indikasi Skewness: Perhatikan bahwa Mean (76,003) > Median (65,237) dengan selisih sekitar 10,000. Ini menunjukkan distribusi data gaji menjulur ke kanan (*positively skewed*). Artinya, ada sebagian karyawan dengan gaji yang cukup tinggi (di atas rata-rata) yang menarik nilai rata-rata menjadi naik.

```
Visualisasi Awal (Scatter Plot)
ggplot(data, aes(x = YearsExperience, y = Salary)) +
 geom_point(color = "blue", size = 2) +
 labs(title = "Scatter Plot: Pengalaman vs Gaji",
 x = "Pengalaman Kerja (Tahun)", y = "Gaji ($)") +
 theme_minimal()
```

## Output



**Pembahasan :** Visualisasi *Scatter Plot* di atas menampilkan pemetaan hubungan antara variabel independen "Pengalaman Kerja" (sumbu X) dan variabel dependen "Gaji" (sumbu Y). Berdasarkan grafik tersebut, terlihat pola penyebaran titik-titik data yang bergerak secara konsisten dari sudut kiri bawah menuju kanan atas. Pola ini mengindikasikan adanya korelasi linear positif yang sangat kuat; artinya, semakin lama pengalaman kerja seseorang (bertambahnya nilai X), maka besaran gaji yang diterima juga cenderung meningkat secara proporsional (bertambahnya nilai Y).

```
model <- lm(Salary ~ YearsExperience, data = data)

Menampilkan hasil analisis
summary(model)
```

## Output

```
...
Call:
lm(formula = Salary ~ YearsExperience, data = data)

Residuals:
 Min 1Q Median 3Q Max
-7958.0 -4088.5 -459.9 3372.6 11448.0

Coefficients:
 Estimate Std. Error t value Pr(>|t|)
(Intercept) 25792.2 2273.1 11.35 5.51e-12 ***
YearsExperience 9450.0 378.8 24.95 < 2e-16 ***

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

Residual standard error: 5788 on 28 degrees of freedom
Multiple R-squared: 0.957, Adjusted R-squared: 0.9554
F-statistic: 622.5 on 1 and 28 DF, p-value: < 2.2e-16
```

**Pembahasan :** Berdasarkan hasil estimasi parameter regresi menggunakan metode kuadrat terkecil (Ordinary Least Squares), diperoleh model persamaan prediksi  $\text{Salary} = 25792.2 + 9450.0\text{YearsExperience}$ . Nilai konstanta (intercept) sebesar 25.792,2 merepresentasikan nilai dasar gaji, yang dapat diinterpretasikan sebagai estimasi gaji awal bagi seseorang yang belum memiliki pengalaman kerja sama sekali (0 tahun). Sementara itu, koefisien regresi (slope) untuk variabel pengalaman kerja bernilai positif sebesar 9.450,0. Angka ini menunjukkan bahwa terdapat hubungan linier yang positif: setiap penambahan pengalaman kerja selama satu tahun, maka gaji karyawan diprediksi akan meningkat secara konstan sebesar 9.450 satuan mata uang. Evaluasi kelayakan model menunjukkan hasil yang sangat memuaskan. Secara parsial, variabel pengalaman kerja terbukti berpengaruh sangat signifikan terhadap gaji, ditunjukkan oleh nilai t-value sebesar 24.95 dengan p-value yang sangat kecil ( $< 2 \cdot 10^{-16}$ ), jauh di bawah taraf signifikansi 0.05. Kekuatan model ini juga dikonfirmasi oleh nilai Koefisien Determinasi (Multiple R-squared) sebesar 0.957. Artinya, model ini mampu menjelaskan 95,7% keragaman atau variasi yang terjadi pada data gaji, sedangkan sisanya sebesar 4,3% dijelaskan oleh faktor lain di luar model. Dengan nilai F-statistic yang tinggi (622.5), model ini dinyatakan valid dan sangat layak digunakan untuk memprediksi gaji berdasarkan pengalaman kerja.

```
a. Visualisasi Garis Regresi
ggplot(data, aes(x = YearsExperience, y = Salary)) +
```

```

geom_point(color = "blue") +
 geom_smooth(method = "lm", col = "red") +
 labs(title = "Model Regresi Linear: Salary ~ YearsExperience",
 subtitle = paste("R-Squared =", round(summary(model)$r.squared, 3)))
+
 theme_minimal()

b. Uji Normalitas Residual (Shapiro-Wilk)
residuals <- residuals(model)
shapiro_test <- shapiro.test(residuals)
print(shapiro_test)

c. Uji Homoskedastisitas (Breusch-Pagan)
bp_test <- bptest(model)
print(bp_test)

```

## Output

```

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

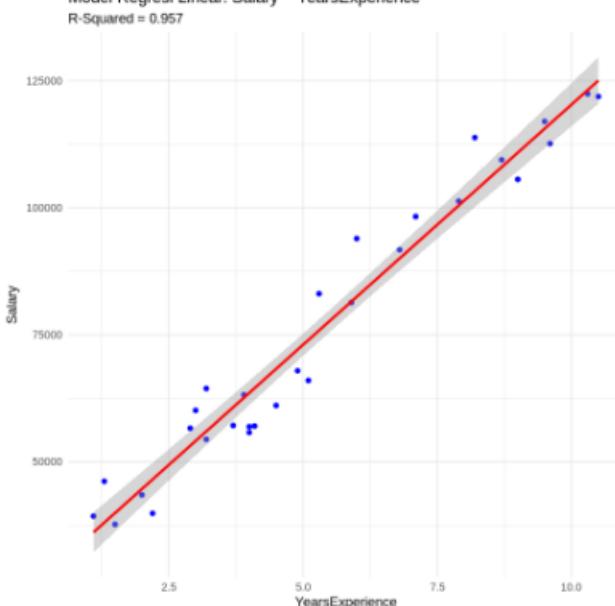
Shapiro-Wilk normality test

data: residuals
W = 0.95234, p-value = 0.1952

studentized Breusch-Pagan test

data: model
BP = 0.39905, df = 1, p-value = 0.5276

```



**Pembahasan :** Tahap evaluasi validitas model dilakukan melalui pengujian asumsi klasik untuk memastikan model regresi yang terbentuk bersifat Best Linear Unbiased Estimator (BLUE). Pertama, pengujian Normalitas Residual menggunakan metode Shapiro-Wilk menghasilkan nilai statistik W sebesar 0.95234 dengan p-value sebesar 0.1952. Karena nilai p-value tersebut lebih besar dari taraf signifikansi 0.05 ( $0.195 > 0.05$ ), maka hipotesis nol diterima, yang berarti sisaan (galat) dari model terdistribusi secara normal. Kedua, pengujian Homoskedastisitas menggunakan metode Studentized Breusch-Pagan menghasilkan p-value sebesar 0.5276. Nilai yang jauh di atas 0.05 ini mengindikasikan bahwa varians dari sisaan bersifat konstan (homogen) dan tidak terjadi gejala

heteroskedastisitas. Dengan terpenuhinya kedua asumsi ini, model regresi dinyatakan valid secara statistik

```
new_data <- data.frame(YearsExperience = 12)
prediksi <- predict(model, new_data, interval = "confidence")
print(paste("Prediksi Gaji untuk 12 Tahun Pengalaman:", round(prediksi[1], 2)))
```

#### Output

```
... [1] "Prediksi Gaji untuk 12 Tahun Pengalaman: 139191.75"
```

**Pembahasan :** Berdasarkan output yang dihasilkan, model mengestimasi bahwa karyawan dengan pengalaman 12 tahun diprediksi akan memiliki gaji sebesar 139.191,75.

## C. PEMBAHASAN TUGAS

1. Cari dataset di open dataset yang lain selain di kaggle, untuk analisa regresi linier sederhana
2. Lakukan analisa regresi linier sederhana.
3. Analisa dari outputnya

```
data(cars)

Melihat struktur data
head(cars) # Menampilkan 6 baris pertama
str(cars) # Mengecek tipe data
```

#### Output

```
... A data.frame: 6 × 2
 speed dist
 <dbl> <dbl>
1 4 2
2 4 10
3 7 4
4 7 22
5 8 16
6 9 10
`data.frame': 50 obs. of 2 variables:
 $ speed: num 4 4 7 7 8 9 10 10 10 11 ...
 $ dist : num 2 10 4 22 16 10 18 26 34 17 ...
```

**Pembahasan :** Analisis dilakukan menggunakan dataset standar R bernama cars yang memuat 50 observasi. Data ini memiliki dua variabel numerik, yaitu:

- speed (Variabel Independen/X): Kecepatan mobil dalam satuan miles per hour (mph).
- dist (Variabel Dependen/Y): Jarak pengemahan yang dibutuhkan mobil untuk berhenti total dalam satuan feet (ft).

```
Membuat model regresi
model_cars <- lm(dist ~ speed, data = cars)

Menampilkan hasil statistik lengkap
summary(model_cars)
```

## Output

```
...
Call:
lm(formula = dist ~ speed, data = cars)

Residuals:
 Min 1Q Median 3Q Max
-29.069 -9.525 -2.272 9.215 43.201

Coefficients:
 Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791 6.7584 -2.601 0.0123 *
speed 3.9324 0.4155 9.464 1.49e-12 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared: 0.6511, Adjusted R-squared: 0.6438
F-statistic: 89.57 on 1 and 48 DF, p-value: 1.49e-12
```

## Pembahasan :

### Interpretasi Koefisien:

- Intercept ( $b_0 = -17.579$ ): Secara matematis, jika kecepatan mobil adalah 0 mph, jarak penggereman adalah -17.57 ft. Dalam konteks fisika nyata, jarak negatif tidak mungkin terjadi. Nilai ini berfungsi sebagai penyesuaian (*anchor*) garis regresi pada rentang data yang diobservasi dan tidak untuk diinterpretasikan secara harfiah pada titik nol.
- Slope ( $b_1 = 3.932$ ): Bernilai positif dan signifikan. Artinya, setiap kenaikan kecepatan mobil sebesar 1 mph, maka jarak penggereman diprediksi akan bertambah sejauh 3.93 feet (kaki).

### Kebaikan Model (Goodness of Fit):

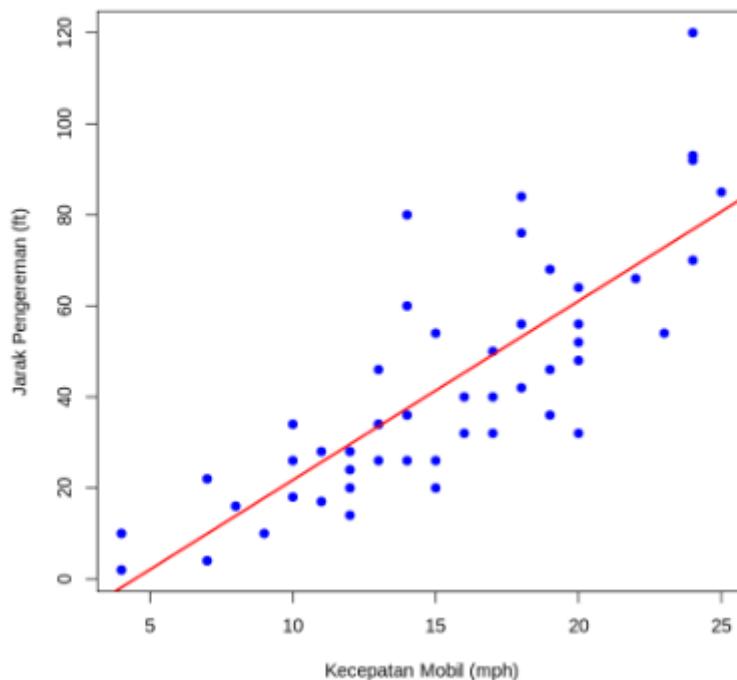
- Uji Signifikansi (Uji-t): Variabel speed memiliki *p-value* sebesar  $1.49 \times 10^{-12}$  (tanda bintang \*\*\*). Karena nilainya jauh di bawah 0.05, dapat disimpulkan bahwa kecepatan berpengaruh sangat signifikan terhadap jarak penggereman.
- R-Squared (\$R^2\$): Nilai *Multiple R-squared* sebesar 0.6511. Ini berarti model regresi mampu menjelaskan 65,11% variasi atau keragaman dari jarak penggereman. Sisanya sebesar 34,89% dijelaskan oleh faktor lain yang tidak tercatat dalam data ini (misalnya kondisi ban, jenis aspal, atau respon pengemudi).

```
Scatter plot dengan garis regresi
plot(cars$speed, cars$dist,
 main = "Analisis Regresi: Kecepatan vs Jarak Penggereman",
 xlab = "Kecepatan Mobil (mph)",
 ylab = "Jarak Penggereman (ft)",
 pch = 19, col = "blue") # Titik data biru
 # Menambahkan garis regresi
abline(model_cars, col = "red", lwd = 2) # Garis merah
```

## Output

...

Analisis Regresi: Kecepatan vs Jarak Pengereman



**Pembahasan :** Grafik *scatter plot* memvisualisasikan hubungan antara Kecepatan (sumbu X) dan Jarak Pengereman (sumbu Y).

- Titik Biru: Menunjukkan data observasi asli. Terlihat pola penyebaran yang bergerak dari kiri bawah ke kanan atas.
- Garis Merah: Adalah garis regresi linear terbaik (Y). Garis ini membelah data untuk meminimalkan *error*.
- Analisis Visual: Meskipun garis merah menangkap tren utama dengan baik, terlihat bahwa pada kecepatan tinggi (kanan atas), titik-titik data cenderung menyebar lebih luas (varians membesar) dibandingkan pada kecepatan rendah. Ini visualisasi yang konsisten dengan nilai  $R^2$  sebesar 65% (cukup baik, tapi tidak sempurna).

```
Menghitung korelasi
korelasi <- cor(cars$speed, cars$dist)
print(paste("Nilai Korelasi Pearson:", round(korelasi, 3)))
```

## Output

```
... [1] "Nilai Korelasi Pearson: 0.807"
```

**Pembahasan :** uji korelasi Pearson untuk melihat kekuatan hubungan linear. Hasil menunjukkan nilai korelasi ( $r$ ) sebesar 0.807.

Interpretasi: Nilai yang mendekati +1 ini mengindikasikan adanya hubungan linear positif yang kuat. Artinya, secara umum, semakin tinggi kecepatan mobil, semakin panjang pula jarak yang dibutuhkan untuk mengerem. Hubungan yang kuat ini menjadi landasan yang valid untuk melanjutkan ke analisis regresi linear.

## **D. KESIMPULAN**

Metode Ordinary Least Squares (OLS) yang diterapkan menggunakan R terbukti efektif untuk memodelkan hubungan sebab-akibat antara satu variabel independen (X) dan satu variabel dependen (Y). Hal ini dibuktikan dengan tingginya nilai koefisien determinasi ( $R^2$ ) pada ketiga kasus uji: 88,41% untuk kasus nilai ujian 10, 95,7% untuk kasus gaji 11, dan 65,11% untuk kasus jarak pengembaraan<sup>12</sup>