Tugas Besar ADS

Kaleb Filbert Istel

2025-11-21

Analisis Regresi dan Korelasi Antara Jam Belajar Mahasiswa terhadap IPK

Data preprocessing & Cleaning

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(readr)
library(stringr)
library(ggplot2)

raw_df <- read_csv("D:/Kuliah/Sem 3/Analisis Data Statistika (ADS)/TUBES/ads-dataset.csv", show_
col_types = FALSE)

head(raw_df)
```

```
## # A tibble: 6 × 20
##   NIM   `Program Studi :` `IPK Terakhir :` `Jenis Kelamin` `Tinggi Badan :`
##   <chr> <chr>             <chr>            <chr>           <chr>
## 1 01    Sains Data        4. 0             Laki-laki       160
## 2 02    Matematika        3.8              Laki-laki       169
## 3 03    Sains Data        3.4              Perempuan       150
## 4 04    Sains Data        3.86             Laki-laki       170
## 5 05    Sains Data        3.97             Laki-laki       170
## 6 06    Sains Data        3.35             Perempuan       154
## # ℹ 15 more variables: `Berat Badan :` <chr>, `Pendidikan terakhir` <chr>,
## #   `Rata-rata belajar perminggu (dalam jam) :` <chr>,
## #   `Apakah penerima beasiswa :` <chr>,
## #   `Asal Daerah (Kota beserta Provinsi) : \nEx : Padang, Sumatera Barat` <chr>,
## #   `Pekerjaan saat ini (selain kuliah)` <chr>,
## #   `Akses Internet (yang paling sering digunakan)` <chr>,
## #   `Keterlibatan Organisai` <chr>, …
```

```r
#Fungsi pembersih rerata jam belajar/minggu
clean_hours <- function(text){
  text_lower <- tolower(text)
  hours <- as.numeric(str_extract(text_lower, "\\d+"))

  # NA jika gaada jam/angka
  if(is.na(hours)){
    return(NA)
  }

  is_per_day <- str_detect(text_lower, "hari|perhari|per hari|/hari")
  is_per_week <- str_detect(text_lower, "minggu|perminggu|per minggu|/minggu")

  if (is_per_day){
    hours_per_week <- hours * 7 #jadi harian
  } else {
    hours_per_week <- hours
  }

  if(is.na(hours_per_week) || hours_per_week == 0 || hours_per_week>84){
    return(NA)
  }

  return(hours_per_week)
}

text_time<- raw_df$`Rata-rata belajar perminggu (dalam jam) :`

#fungsi pembersih IPK
clean_ipk <- function(x){
  if(is.na(x)) return(NA)

  # Ambil bagian angka
  value <- str_extract(x, "\\d+[\\.,]?\\d*")

  if (is.na(value)) return(NA)

  # Jika bentuknya "3." atau "3," → jadikan "3"
  value <- gsub("[\\.,]$", "", value)

  # Ganti koma ke titik untuk jadi desimal
  value <- gsub(",", ".", value)

  # Konversi ke numeric
  num <- as.numeric(value)

  # Validasi rentang IPK wajar
  if(is.na(num) || num < 0 || num > 4) return(NA)

  return(num)
}
```

```
ipkraw <- raw_df$`IPK Terakhir :`


df <- data.frame(
  ipk = sapply(ipkraw, clean_ipk),
  hours_per_week = sapply(text_time, clean_hours)
)

df_clean<- df %>%
  filter(!is.na(hours_per_week))

df_clean <- df_clean %>%
  filter(!is.na(ipk))

head(df_clean)
```

```
##    ipk hours_per_week
## 1 4.00              3
## 2 3.80             48
## 3 3.40             17
## 4 3.86              4
## 5 3.97             30
## 6 3.35              3
```

```
model <- lm(ipk ~ hours_per_week, data = df_clean)
summary(model)
```

```
##
## Call:
## lm(formula = ipk ~ hours_per_week, data = df_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3114 -0.2375  0.0952  0.3886  0.7380
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.256552   0.047407  68.693   <2e-16 ***
## hours_per_week 0.002744   0.002264   1.212    0.226
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6487 on 323 degrees of freedom
## Multiple R-squared:  0.004529,   Adjusted R-squared:  0.001447
## F-statistic: 1.469 on 1 and 323 DF,  p-value: 0.2263
```

MODEL REGRESI: **IPK = 3.256552 +0.002744 * Jam belajar per minggu**

Koefisien Regresi:

Komponen	Nilai	Interpretasi
Intercept	**3.256552**	Saat jam belajar = 0, IPK yang diprediksi ≈ **3.26**.
hours_per_week (β1)	**0.002744**	**Setiap tambahan 1 jam belajar per minggu hanya menaikkan IPK sebesar 0.0027 poin**, dan secara statistik *tidak signifikan*.
p-value (β1)	**0.226**	Tidak signifikan pada taraf 5% → **jam belajar tidak terbukti memengaruhi IPK secara linear**.

Goodness of Fit (Kualitas Model)

Statistik	Nilai	Interpretasi
Multiple R-squared	**0.004529**	Model hanya menjelaskan **0.45%** variasi IPK.
Adjusted R-squared	**0.001447**	Setelah penyesuaian, kemampuan prediksi model makin kecil.
F-statistic	**1.469**	Model tidak signifikan.
p-value (model)	**0.2263**	Secara keseluruhan, model regresi **tidak signifikan**.

Karena model regresi sederhana ini hasil dari OLS (Ordinary least square)/ metode Kuadrat terkecil.

Kita sebaiknya mengecek apakah datanya memenuhi semua **asumsi klasik OLS nya.**

Karena jika memenuhi semua asumsi, model regresi ini akan bersifat **BLUE (Best Linear Unbiased Estimator)** model ini dapat digunakan untuk melakukan estimasi dan inferensi statistik secara valid dalam konteks hubungan IPK dan Jam belajar

Uji korelasi

```
# Korelasi Pearson
cor_xy <- cor(df_clean$hours_per_week, df_clean$ipk, method = "pearson")
cor_xy
```

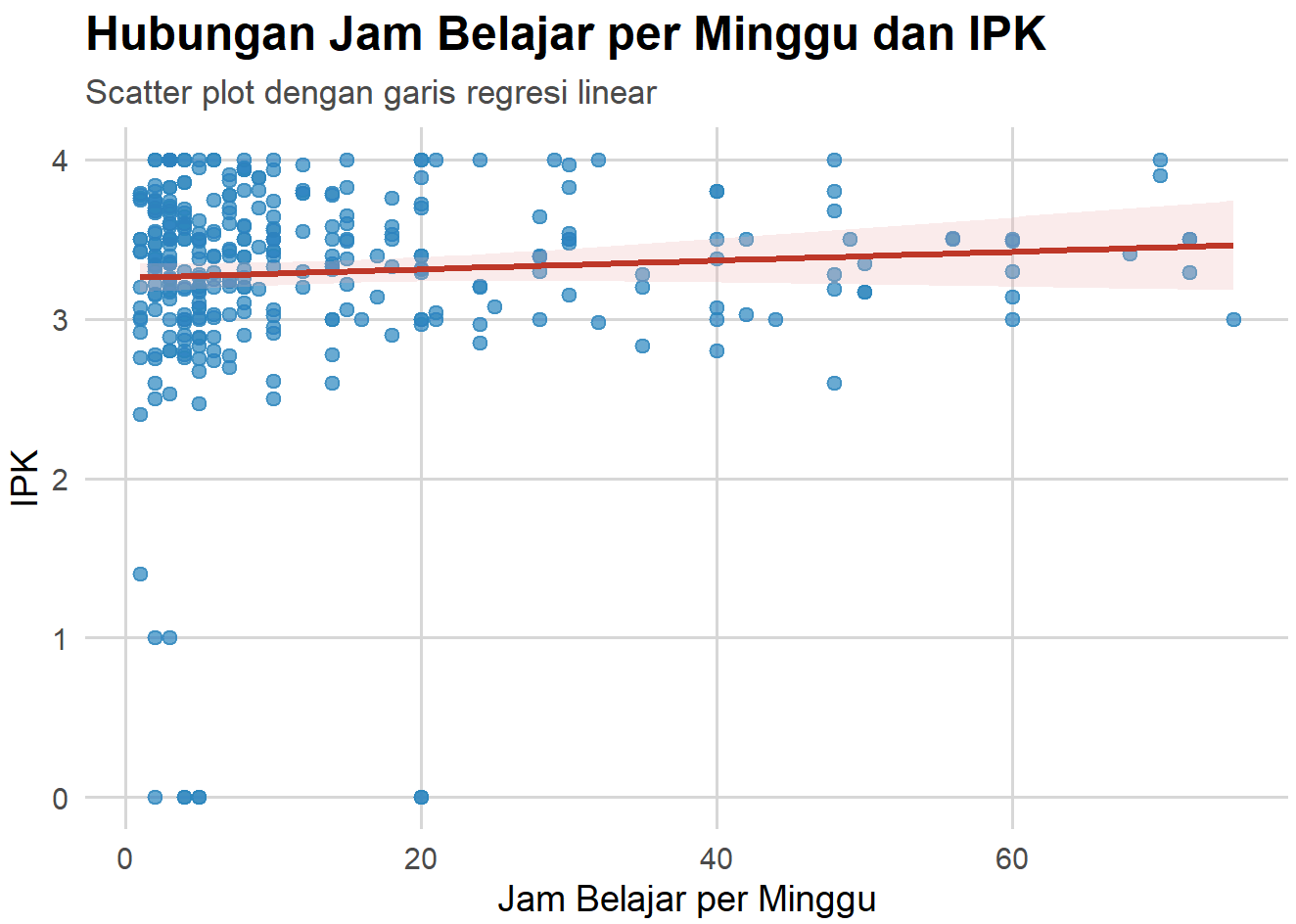
```
## [1] 0.06729458
```

Uji **Asumsi Klasik Regresi (OLS)**

1. Uji linearitas

```
library(ggplot2)

scatter<-ggplot(df_clean, aes(x = hours_per_week, y = ipk)) +
  geom_point(color = "#2E86C1", size = 2.3, alpha = 0.7) +          # titik data
  geom_smooth(method = "lm", se = TRUE, color = "#C0392B",
              linewidth = 1.2, fill = "#F5B7B1", alpha = 0.25) +    # garis regresi + band CI
  labs(
    title = "Hubungan Jam Belajar per Minggu dan IPK",
    subtitle = "Scatter plot dengan garis regresi linear",
    x = "Jam Belajar per Minggu",
    y = "IPK"
  ) +
  theme_minimal(base_size = 14) +
  theme(
    plot.title = element_text(face = "bold", size = 18),
    plot.subtitle = element_text(size = 13, color = "gray30"),
    panel.grid.minor = element_blank(),
    panel.grid.major = element_line(color = "gray85")
  )
scatter
```

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



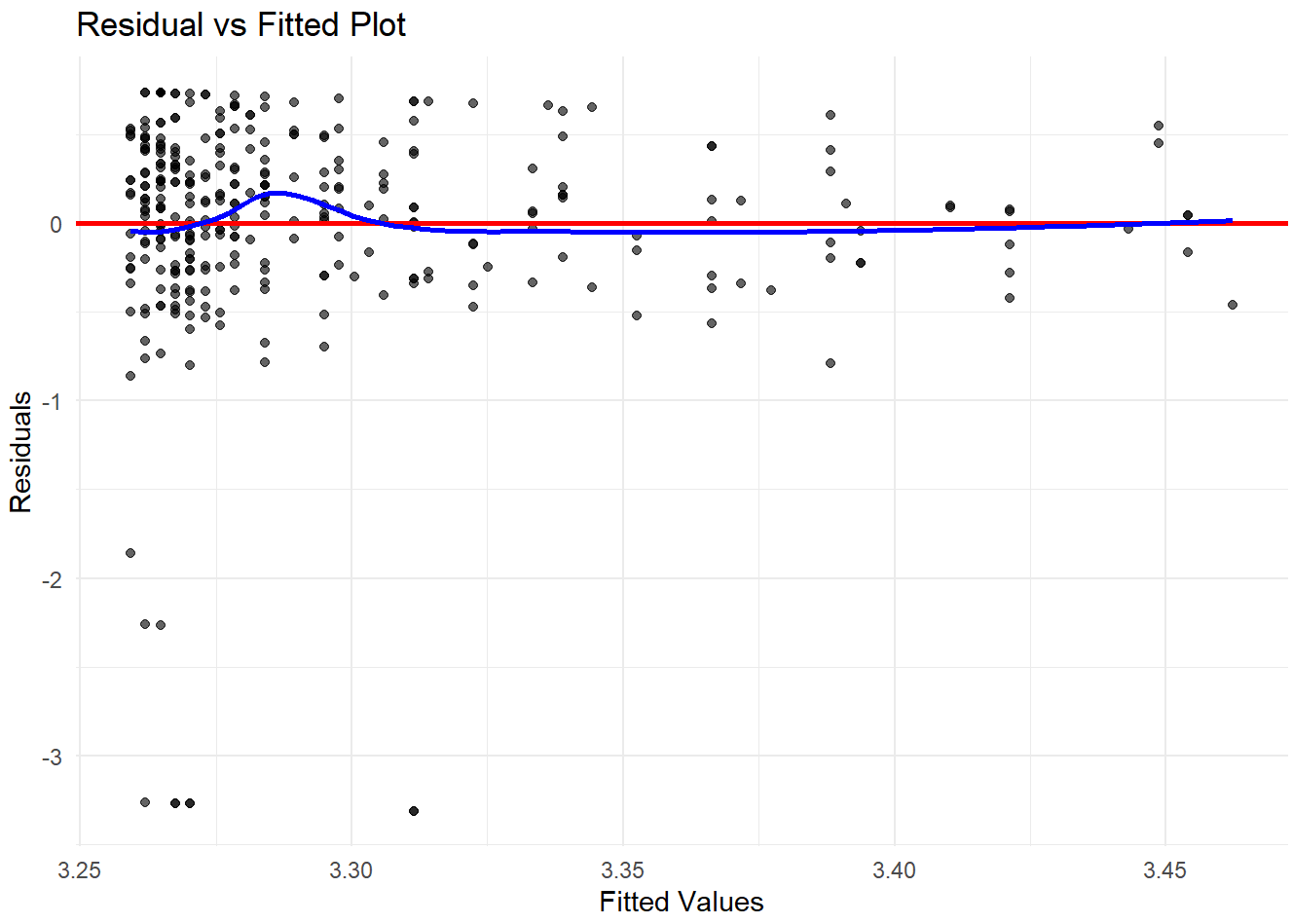
2. Ekspektasi nilai error = 0

```
# Residual dari model
res <- resid(model)
# Fit model
fit<-fitted(model)

#hitung rerata residual
mean_res<-mean(res)
mean_res
```

```
## [1] -9.275516e-18
```

```
# Membuat plot residual terhadap nilai prediksi
resfit<-ggplot(data.frame(fitted = fit, residuals = res),
       aes(x = fitted, y = residuals)) +
  geom_point(alpha = 0.6) +
  geom_hline(yintercept = 0, color = "red", linewidth = 1) +
  geom_smooth(method = "loess", se = FALSE, color = "blue") +
  labs(
    title = "Residual vs Fitted Plot",
    x = "Fitted Values",
    y = "Residuals"
  ) +
  theme_minimal()
resfit
```

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



Interpretasi:

1. **Residual mengambang di sekitar 0 → E(ε) = 0 terpenuhi**

Garis merah = garis residual nol.
Mayoritas titik tersebar di sekitar garis tersebut → *mean residual ≈ 0* ✓

Ini sejalan dengan hasil mean residual = **–9e–18**, yang praktis **nol sempurna** karena OLS.

2. Tidak ada pola melengkung besar → linearitas cukup terpenuhi

Garis biru (LOESS) sempat naik sedikit, lalu turun, tapi:

amplitudonya kecil, tidak membentuk pola bowl atau wave besar, dan tidak ada tren sistematis

Kesimpulan: **tidak ada indikasi kuat pelanggaran linearitas**.

3. **Homoskedastisitas: sebagian besar tampak wajar**

   (Jika titik menyebar acak di sekitar garis 0 dan lebar sebarannya relatif sama, asumsi homoskedastisitas cenderung terpenuhi. tidak terlihat pola “kipas” (fan shape) atau corong (semakin melebar / menyempit), yang berupa indikasi heteroskedastisitas.)

4. Ada outlier beberapa titik di bawah -2 dan -3 tapi tidak mengganggu tren

3. Homoskedastisitas (Varians Error Konstan)

Sudah bisa dilihat visualisasinya di graf Residuals vs Fitted plot sebelumnya, disini lebih ke uji numerik BP (Breusch-Pagan) nya

Jika varians error berubah-ubah (heteroskedastisitas), maka estimasi tetap tidak bias tetapi tidak efisien, dan uji statistik seperti uji t menjadi tidak valid.

Jika homoskedastisitas/varians error konstan, uji t menjadi **VALID.**

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
bptest(model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 1.7881, df = 1, p-value = 0.1812
```

Hipotesis tes homoskedastisitas:

H0 = homoskedastisitas/ tidak ada bukti heteroskedastisitas

H1 = terbukti heteroskedastisitas

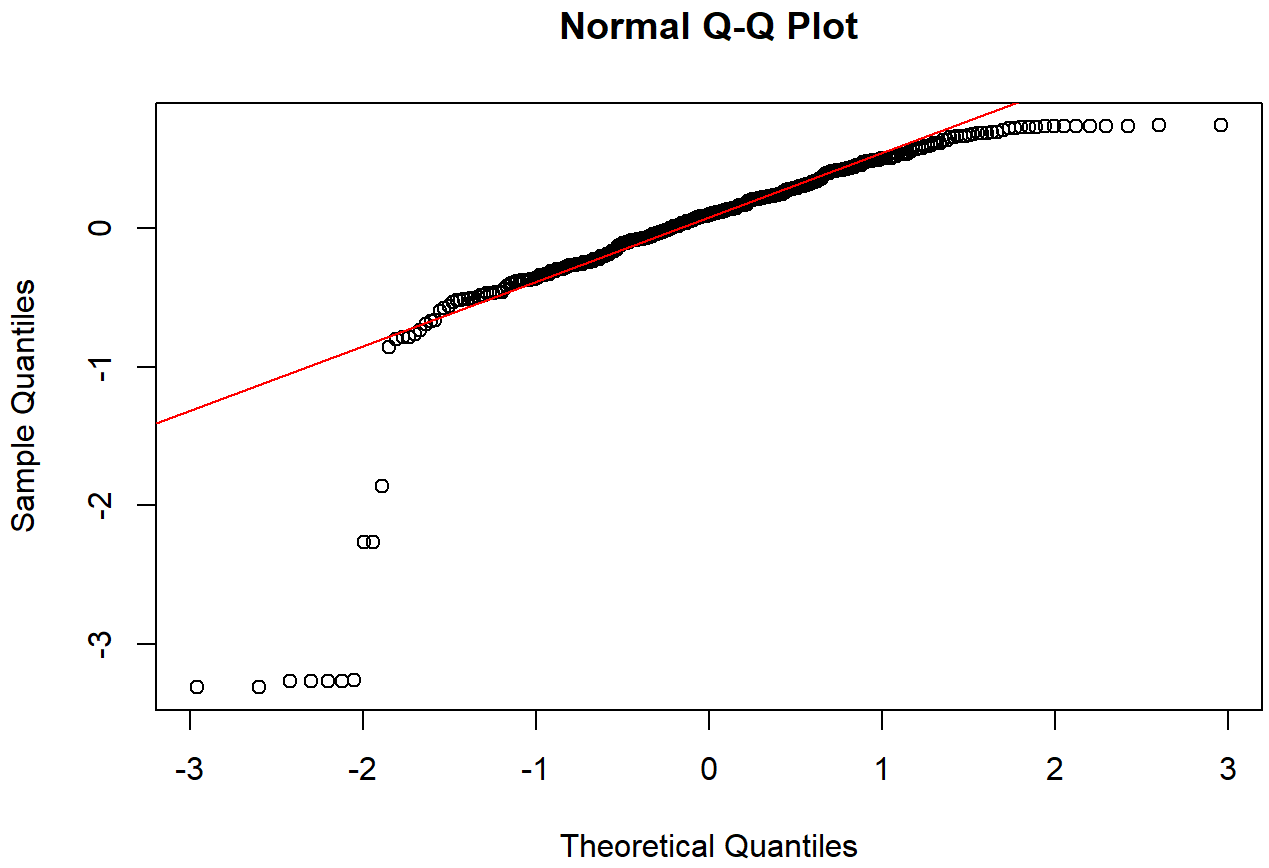
p-value (**0.1812**) > 0.05 → gagal tolak H0 -> tidak ada bukti heteroskedastisitas -> asumsi homoskedaskitas terpenuhi.

4. Uji error terdistribusi normal / normalitas residual

```
# Uji Shapiro-Wilk
shapiro.test(res)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res
## W = 0.71376, p-value < 2.2e-16
```

```
# Q-Q plot
qqnorm(res)
qqline(res, col = "red")
```



Residual **tidak berdistribusi normal**, terutama karena outlier kuat di bagian kiri dan sedikit penyimpangan di bagian kanan. Meski bagian tengah residual relatif normal, ekor distribusi yang melengkung menunjukkan pelanggaran asumsi normalitas signifikan.

Titik-titik yang mengikuti garis merah menunjukkan bahwa residual berdistribusi mendekati normal. Tapi karena banyak titik menyimpang jauh dari garis, terutama di ekor distribusi, maka indikasi ketidaknormalan muncul.

Interpretasi:

Uji Shapiro–Wilk memiliki hipotesis:

H0: residual berdistribusi normal

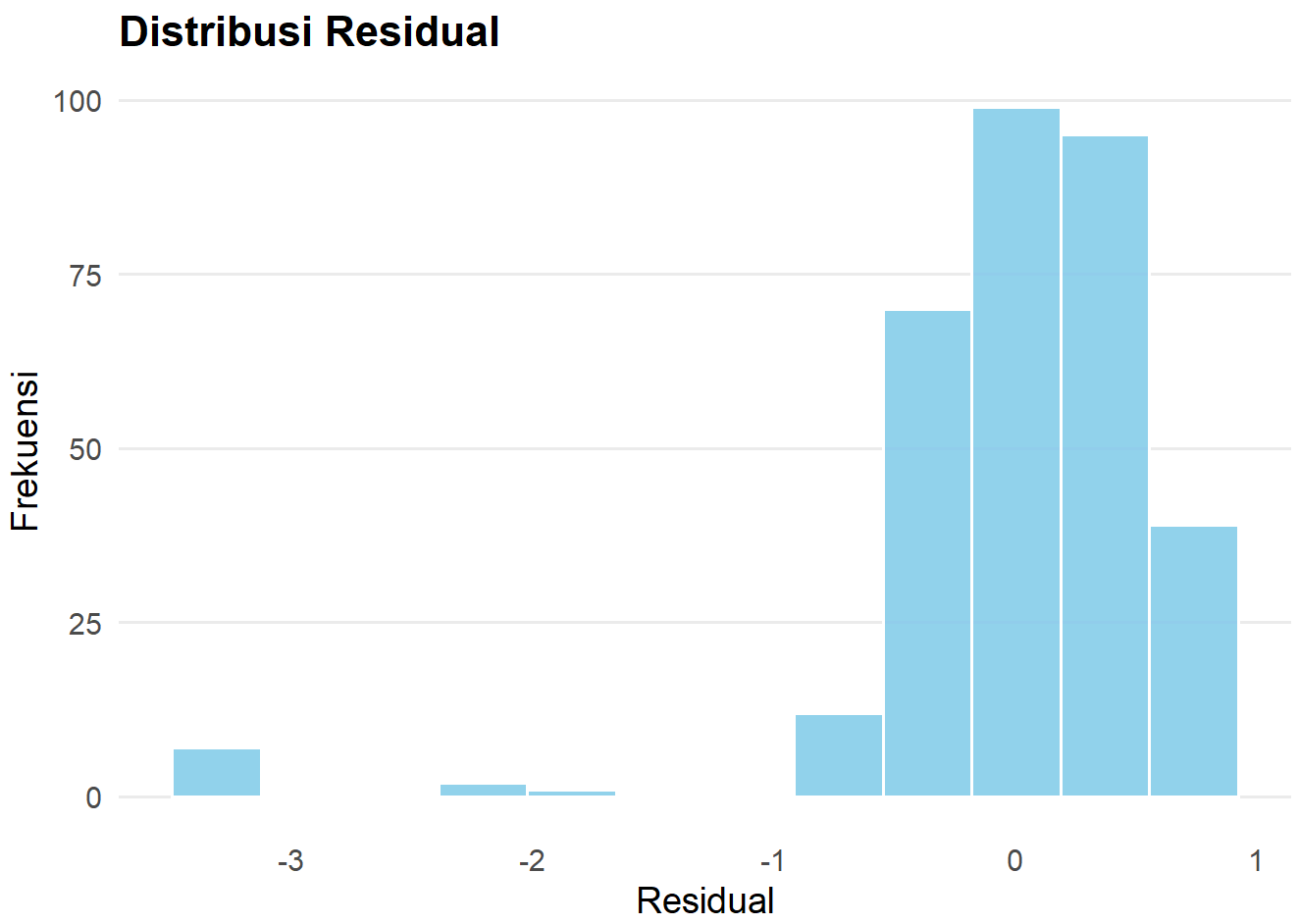
H1: residual *tidak* berdistribusi normal

Karena:

p-value < 0.05, bahkan **sangat kecil (< 2.2e−16)**

Maka, Tolak H0
Residual tidak berdistribusi normal

```
library(ggplot2)

hist_gg <- ggplot(data.frame(residuals = res), aes(x = residuals)) +
  geom_histogram(
    bins = 12,
    fill = "skyblue",
    color = "white",
    alpha = 0.9
  ) +
  labs(
    title = "Distribusi Residual",
    x = "Residual",
    y = "Frekuensi"
  ) +
  theme_minimal(base_size = 14) +
  theme(
    plot.title = element_text(face = "bold", size = 16),
    panel.grid.minor = element_blank(),
    panel.grid.major.x = element_blank()
  )

hist_gg
```



export:

```
ggsave("histogram_residuals_ggplot.png",
       hist_gg,
       width = 8,
       height = 6,
       dpi = 300)
```

Interpretasi:

1. Histogram left skewed
2. terdapat lonceng, tapi terdapat ekor kecil selain itu

   Sebagian besar prediksi model cukup akurat (residual kecil), tetapi terdapat sejumlah kecil mahasiswa yang menyimpang jauh dari prediksi sehingga membuat distribusi residual menjadi **left-skewed dan tidak normal**. Model masih bisa digunakan, tetapi asumsi normalitas tidak sepenuhnya terpenuhi.

Hasil visualisasi menunjukkan bahwa asumsi **normalitas error** pada regresi linear tidak terpenuhi secara statistik.

Namun walaupun tidak terpenuhi asumsi ini, untungnya karena data besar (n>300):

Regresi OLS *tetap konsisten,* Uji t dan uji F *tetap robust,* Pelanggaran normalitas sering terjadi dan *tidak fatal.*

Kesimpulan

Model Regresi memenuhi asumsi OLS sehingga estimasi koefisien bersifat BLUE, namun hubungan antara jam belajar dan IPK sangat lemah dan tidak signifikan (p>0.05, r=0,067). Karena adanya non-normalitas residual akibat beberapa outlier, interpretasi inferensial harus dilakukan dengan kehati-hatian dan/atau metode robust.

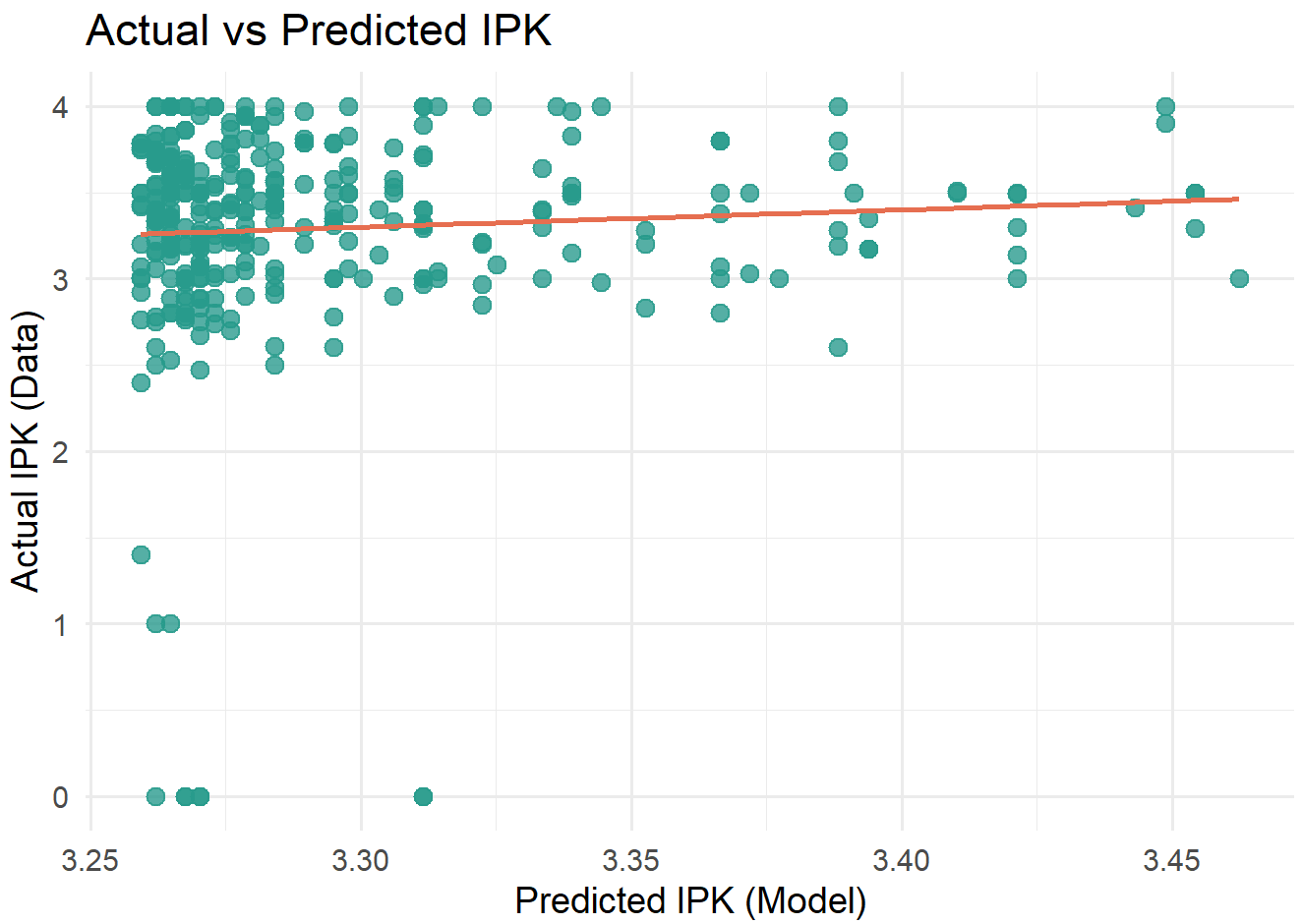
Forecasting Model Actual vs Predicted IPK

```
library(ggplot2)
library(dplyr)

# model regresi
model <- lm(ipk ~ hours_per_week, data = df_clean)

# buat data frame prediksi
df_plot <- df_clean %>%
  mutate(
    predicted_ipk = predict(model),
    residuals = residuals(model)
  )

# Plot Actual vs Predicted
p_actual_pred <- ggplot(df_plot, aes(x = predicted_ipk, y = ipk)) +
  geom_point(color = "#2A9D8F", size = 3, alpha = 0.8) +
  geom_smooth(method = "lm", color = "#E76F51", se = FALSE, size = 1) +
  labs(
    title = "Actual vs Predicted IPK",
    x = "Predicted IPK (Model)",
    y = "Actual IPK (Data)"
  ) +
  theme_minimal(base_size = 14)
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
# tampilkan
p_actual_pred
```

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


Actual vs Predicted IPK
Actual IPK (Data)
Predicted IPK (Model)
0
1
2
3
4
3.25
3.30
3.35
3.40
3.45