Analisis Faktor yang Mempengaruhi Harga Rumah di Boston

Zafyra Nur Rizqi

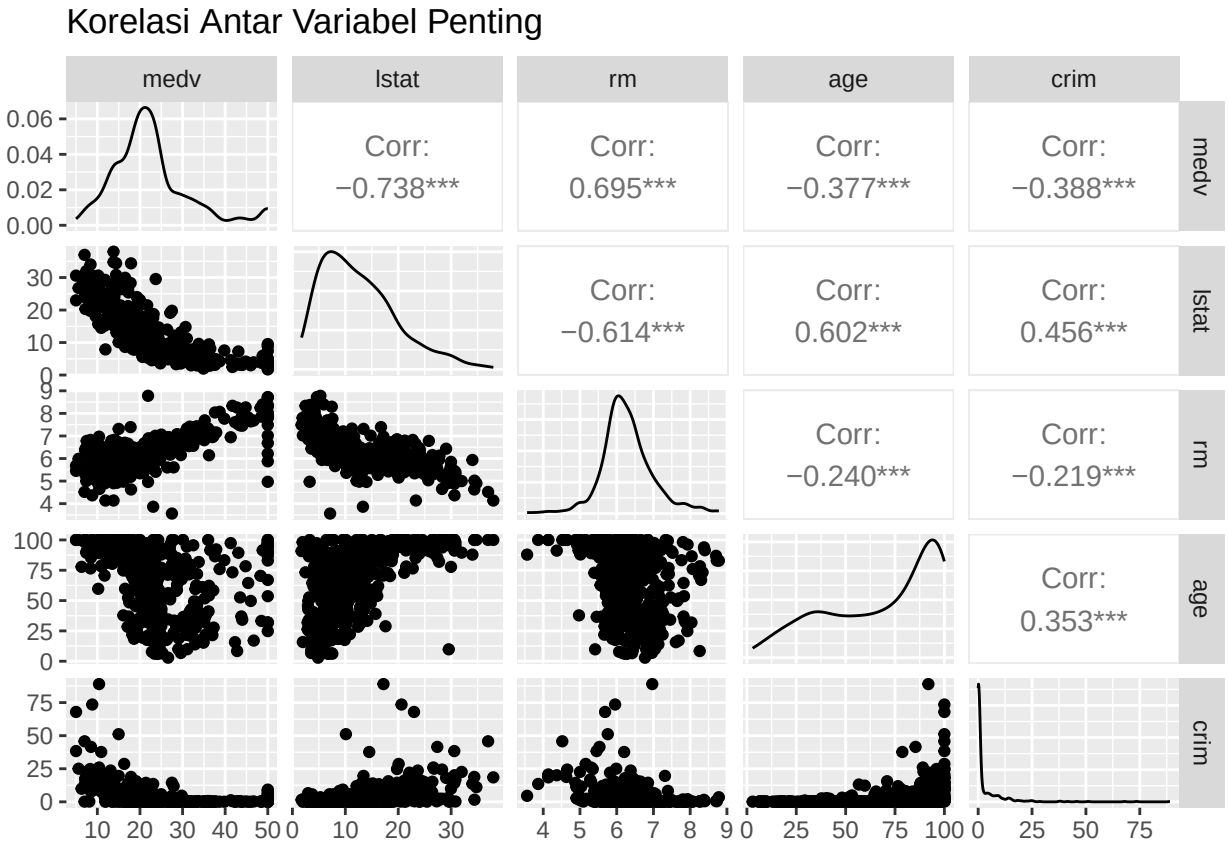
31 May 2025

Contents

1 Data

```
## # A tibble: 6 x 14
##      crim    zn indus  chas   nox    rm   age   dis   rad   tax ptratio black
##     <dbl> <dbl> <dbl> <int> <dbl> <dbl> <dbl> <dbl> <int> <dbl>   <dbl> <dbl>
## 1 0.00632    18  2.31     0 0.538  6.58  65.2  4.09     1   296    15.3  397.
## 2 0.0273      0  7.07     0 0.469  6.42  78.9  4.97     2   242    17.8  397.
## 3 0.0273      0  7.07     0 0.469  7.18  61.1  4.97     2   242    17.8  393.
## 4 0.0324      0  2.18     0 0.458  7.00  45.8  6.06     3   222    18.7  395.
## 5 0.0690      0  2.18     0 0.458  7.15  54.2  6.06     3   222    18.7  397.
## 6 0.0298      0  2.18     0 0.458  6.43  58.7  6.06     3   222    18.7  394.
## # i 2 more variables: lstat <dbl>, medv <dbl>
```

```
##       crim                zn              indus            chas        
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000  
##  1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000  
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000  
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917  
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000  
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000  
##       nox               rm              age              dis        
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130  
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100  
##  Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207  
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795  
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188  
##  Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127  
##       rad              tax           ptratio          black       
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32  
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38  
##  Median : 5.000   Median :330.0   Median :19.05   Median :391.44  
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67  
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23  
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90  
##      lstat            medv      
##  Min.   : 1.73   Min.   : 5.00  
##  1st Qu.: 6.95   1st Qu.:17.02  
##  Median :11.36   Median :21.20  
##  Mean   :12.65   Mean   :22.53  
##  3rd Qu.:16.95   3rd Qu.:25.00  
##  Max.   :37.97   Max.   :50.00
```




```
##              medv      lstat         rm        age       crim
## medv    1.0000000 -0.7376627  0.6953599 -0.3769546 -0.3883046
## lstat  -0.7376627  1.0000000 -0.6138083  0.6023385  0.4556215
## rm      0.6953599 -0.6138083  1.0000000 -0.2402649 -0.2192467
## age    -0.3769546  0.6023385 -0.2402649  1.0000000  0.3527343
## crim   -0.3883046  0.4556215 -0.2192467  0.3527343  1.0000000
```

Pengertian:

Saya memakai dataset Boston yang berasal dari package R yaitu MASS. Dataset ini berisi data properti di wilayah Boston. Beberapa variabel penting dalam dataset ini berupa medv (median harga rumah), lstat (persentase pennduduk dengan status sosial rendah), rm (rata-rata jumlah kamar/rumah), age (proporsi rumah tua), crim (tingkat kejahatan per kapita), dan indus (proporsi area bisnis non retail).

Adapula package library yang saya gunakan guna menunjang analisis ini, yaitu berupa MASS (dataset Boston), ggplot2 brom dan GGally (visualisasi), dplyr (manipulasi data), psych (statistik deskriptif), car (uji asumsi klasik), lmtest (uji homoskedastisitas).

Hasil ringkasan statistik deskriptif menunjukkan harga rumah (medv) memiliki nilai rata-rata 22.53, dengan nilai minimum 5 dan maksimum 50, kriminalitas (crim) sangat bervariasi dari 0.00632 hingga 88.98, dan terdapat korelasi negatif kuat antara medv dan lstat (-0.738), serta korelasi positif antara medv dan rm (0.695), yang menandakan keduanya adalah prediktor penting untuk harga rumah.

1.1 Model Regresi

```
model <- lm(medv ~ lstat + rm + age, data = Boston)
summary(model)
```

```
##
## Call:
## lm(formula = medv ~ lstat + rm + age, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.210  -3.467  -1.053   1.957  27.500
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.175311   3.181924  -0.369    0.712
## lstat       -0.668513   0.054357 -12.298   <2e-16 ***
## rm           5.019133   0.454306  11.048   <2e-16 ***
## age          0.009091   0.011215   0.811    0.418
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.542 on 502 degrees of freedom
## Multiple R-squared:  0.639,  Adjusted R-squared:  0.6369
## F-statistic: 296.2 on 3 and 502 DF,  p-value: < 2.2e-16
```

```
tidy(model, conf.int = TRUE)
```

```
## # A tibble: 4 x 7
##   term        estimate std.error statistic  p.value conf.low conf.high
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept) -1.18       3.18      -0.369 7.12e- 1  -7.43      5.08
## 2 lstat       -0.669      0.0544   -12.3   1.44e-30  -0.775    -0.562
## 3 rm           5.02       0.454     11.0   1.51e-25   4.13      5.91
## 4 age          0.00909    0.0112     0.811 4.18e- 1  -0.0129    0.0311
```

Pengertian:

Analisis ini saya buat untuk mengetahui faktor-faktor apa saja yang memengaruhi harga rumah (medv) di Boston maka dari itu digunakanlah regresi berganda untuk memodelkan hubungan antara harga rumah dan beberapa variabel prediktor, yaitu lstat, rm dan age.

Hasil model regresi ini menunjukkan Intersep sebesar -1.175, tidak signifikan ($p = 0.712$), lstat mempunyai koefisien -0.669, sangat signifikan ($p < 2e\text{-}16$) yang berarti setiap kenaikan 1% populasi berstatus sosial rendah menurunkan harga rumah sekitar \$669, rm mempunyai koefisien 5.02, sangat signifikan ($p < 2e\text{-}16$) yang berarti setiap tambahan satu kamar meningkatkan harga rumah sekitar \$5,020, dan age mempunyai Koefisien 0.009 atau tidak signifikan ($p = 0.418$).

Hasil akhirnya adalah model ini menjelaskan 63.9% variasi harga rumah (R-squared = 0.639). Artinya, model ini cukup baik dalam menjelaskan hubungan antara variabel bebas dengan medv.

2 Uji Asumsi Klasik

2.1 Multikolineritas

```
vif(model)
```

```
##    lstat       rm      age
## 2.477304 1.675215 1.638542
```

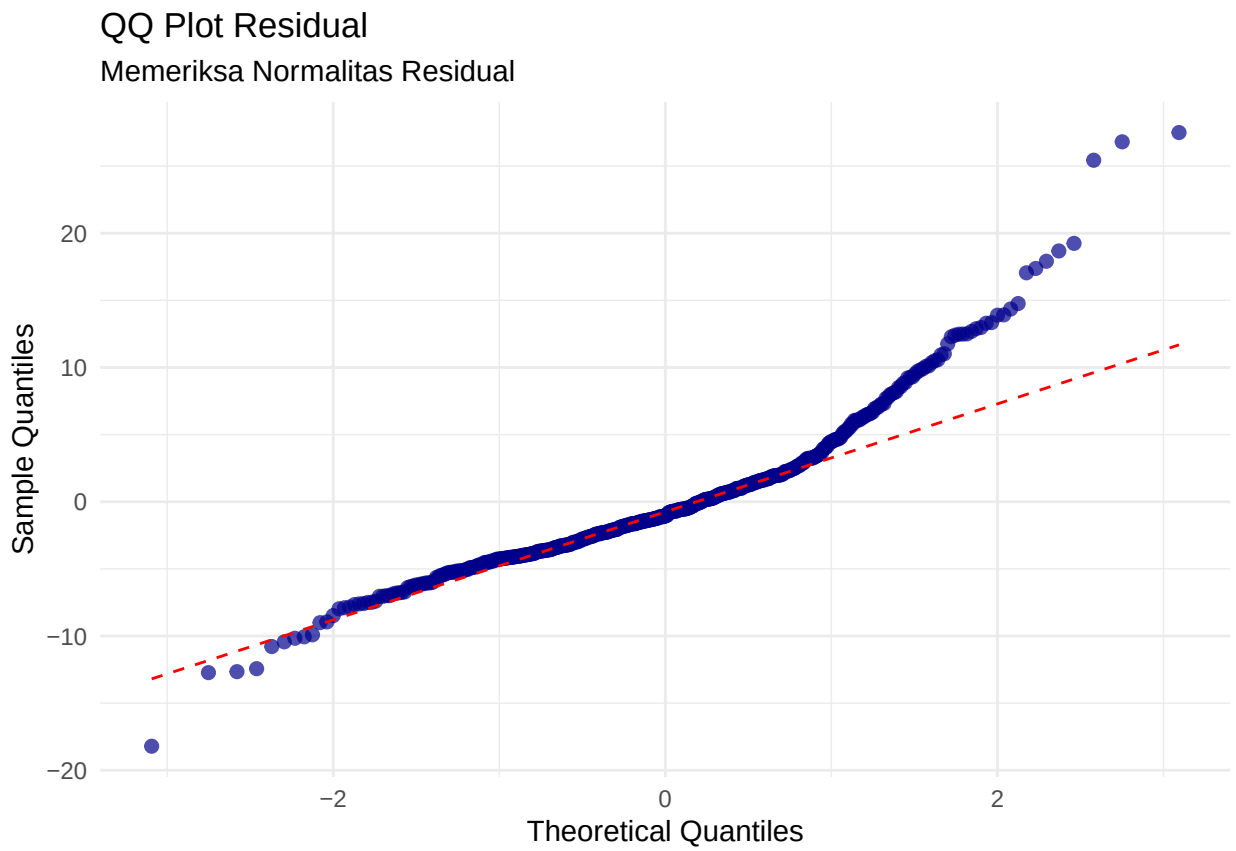
Pengertian:

Menghitung Variance Inflation Factor. VIF > 10 artinya terjadi multikolinearitas. Jika semua $<$ 10, aman.

Dari hasil yang didapatkan, semua nilai VIF < 10 maka tidak ada masalah multikolineritas.

2.2 Normalitas Residual

```
res <- resid(model)
ggplot(data = data.frame(residuals = res), aes(sample = residuals)) +
  stat_qq(color = "darkblue", size = 2, alpha = 0.7) +
  stat_qq_line(color = "red", linetype = "dashed") +
  labs(
    title = "QQ Plot Residual",
    subtitle = "Memeriksa Normalitas Residual",
    x = "Theoretical Quantiles",
    y = "Sample Quantiles"
  ) +
  theme_minimal()
```




```
shapiro.test(resid(model))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(model)
## W = 0.91406, p-value = 2.385e-16
```

Pengertian:

Fungsi ini untuk mengetaui apakah residual berdistribusi normal, dibantu dengan visualisasi datanya. Dilakukan juga uji statistik Shapiro-Wilk, yang dimana jika hasil nilai $p > 0.05$ = data normal.

Dari hasil yang didapatkan, W = 0.914 dan p-value = 2.385e-16 sehingga artinya, residual tidak berdistribusi normal, karena $p < 0.05$. Hal ini juga terlihat dari QQ Plot yang menyimpang dari garis lurus.

2.3 Homoskedastisitas

```r
bptest(model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 19.771, df = 3, p-value = 0.0001894
```

Pengertian:

Dilakukan uji Breusch-Pagan yang dimana jika hasil p-value > 0.05 menunjukkan residual memiliki varians konstan (tidak heteroskedastis).

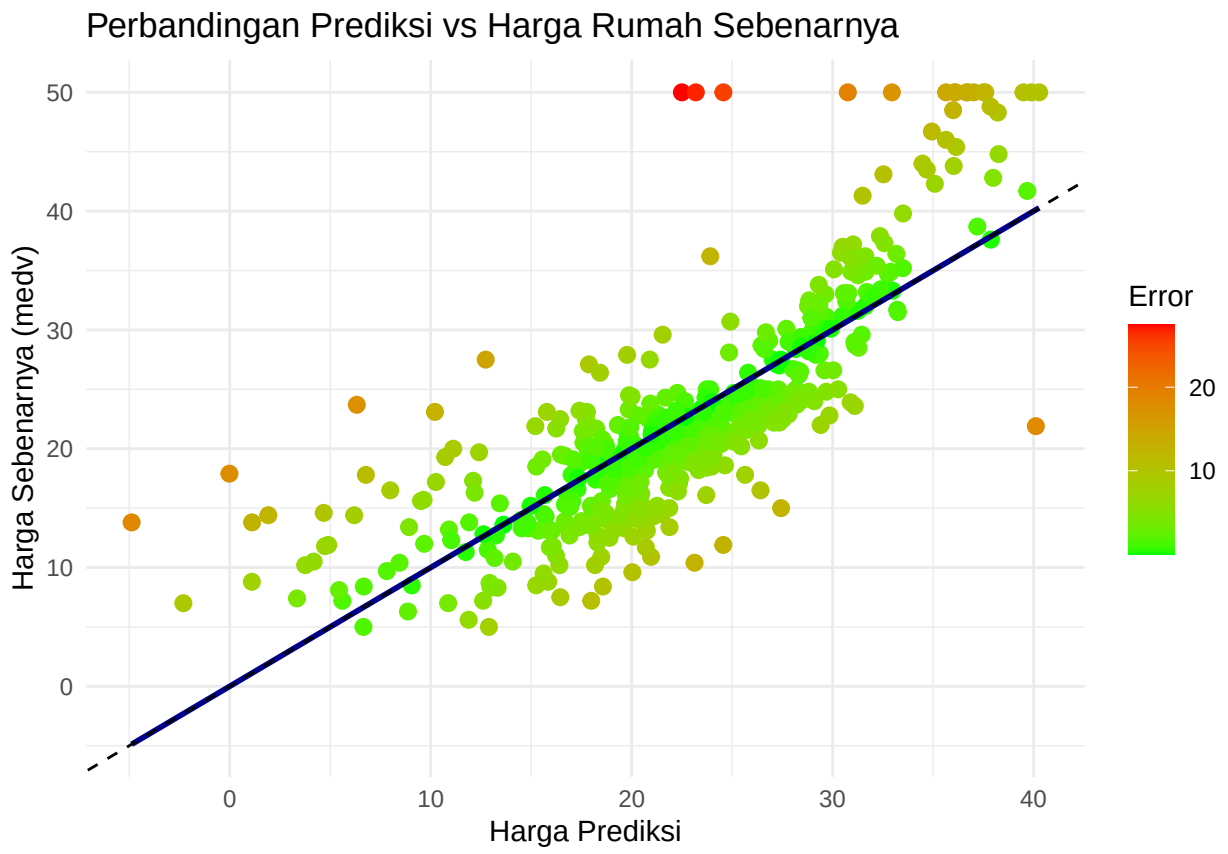
Dari hasil yang didapatkan, nilai BP $= 19.77$ dan p-value $= 0.0001894$ sehingga artinya terdapat indikasi heteroskedastisitas.

3 Visualisasi

3.1 Plot Prediksi vs Realisasi

```r
prediksi <- predict(model)
ggplot(data = Boston, aes(x = prediksi, y = medv)) +
  geom_point(aes(color = abs(prediksi - medv)), size = 2.5) +
  scale_color_gradient(low = "green", high = "red") +
  geom_smooth(method = "lm", se = FALSE, color = "darkblue") +
  geom_abline(slope = 1, intercept = 0, color = "black", linetype = "dashed") +
  labs(
    title = "Perbandingan Prediksi vs Harga Rumah Sebenarnya",
    x = "Harga Prediksi",
    y = "Harga Sebenarnya (medv)",
    color = "Error"
  ) +
  theme_minimal()
```

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



Pengertian:

Plot ini menunjukkan perbandingan antara harga rumah hasil prediksi dan harga sebenarnya. Garis diagonal (hitam) menunjukkan jika prediksi sempurna. Titik-titik yang jauh dari garis menunjukkan error prediksi. Semakin merah warnanya, semakin besar errornya.

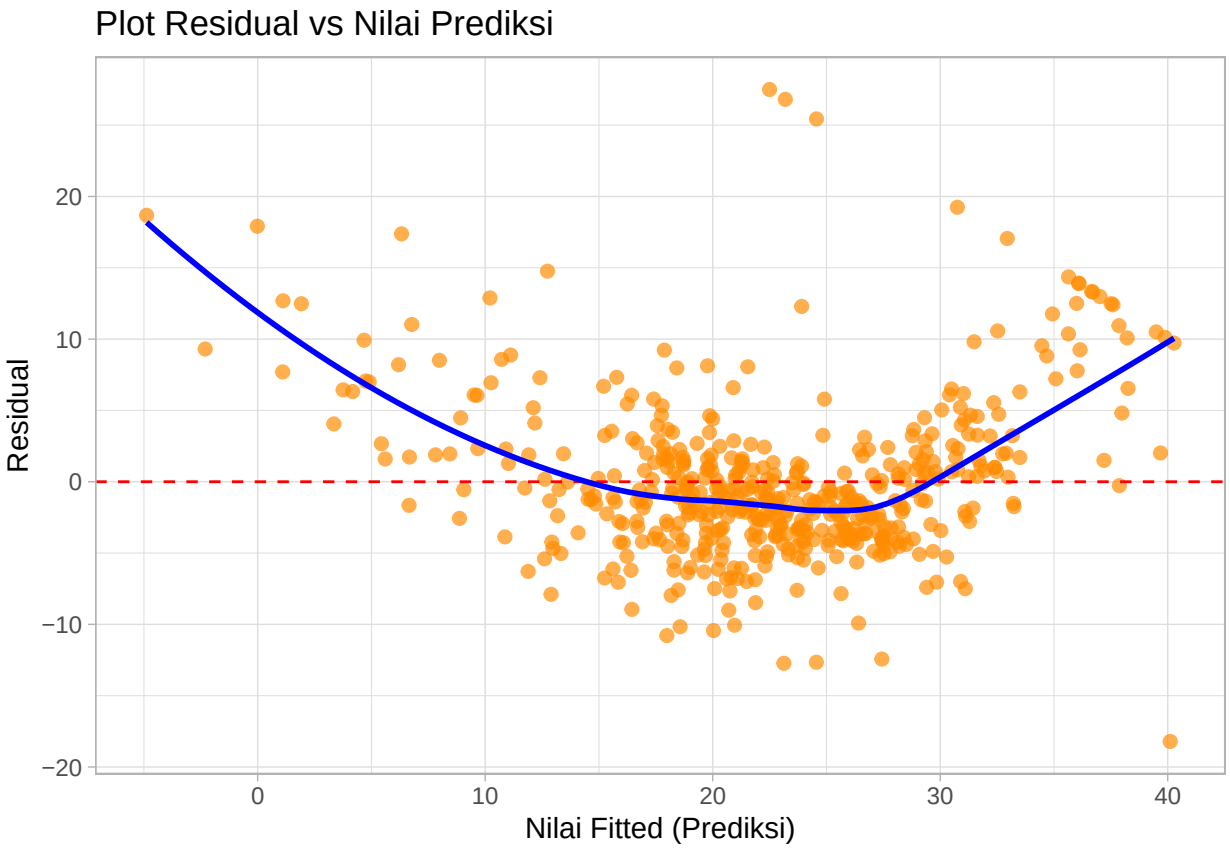
Model terlihat mampu menangkap pola umum, namun terdapat beberapa deviasi (error) besar.

3.2 Plot Residual vs Fitted

```
residuals <- resid(model)
fitted <- fitted(model)

ggplot(data = NULL, aes(x = fitted, y = residuals)) +
  geom_point(color = "darkorange", alpha = 0.7, size = 2) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  geom_smooth(method = "loess", se = FALSE, color = "blue") +
  labs(
    title = "Plot Residual vs Nilai Prediksi",
    x = "Nilai Fitted (Prediksi)",
    y = "Residual"
  ) +
  theme_light()
```

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

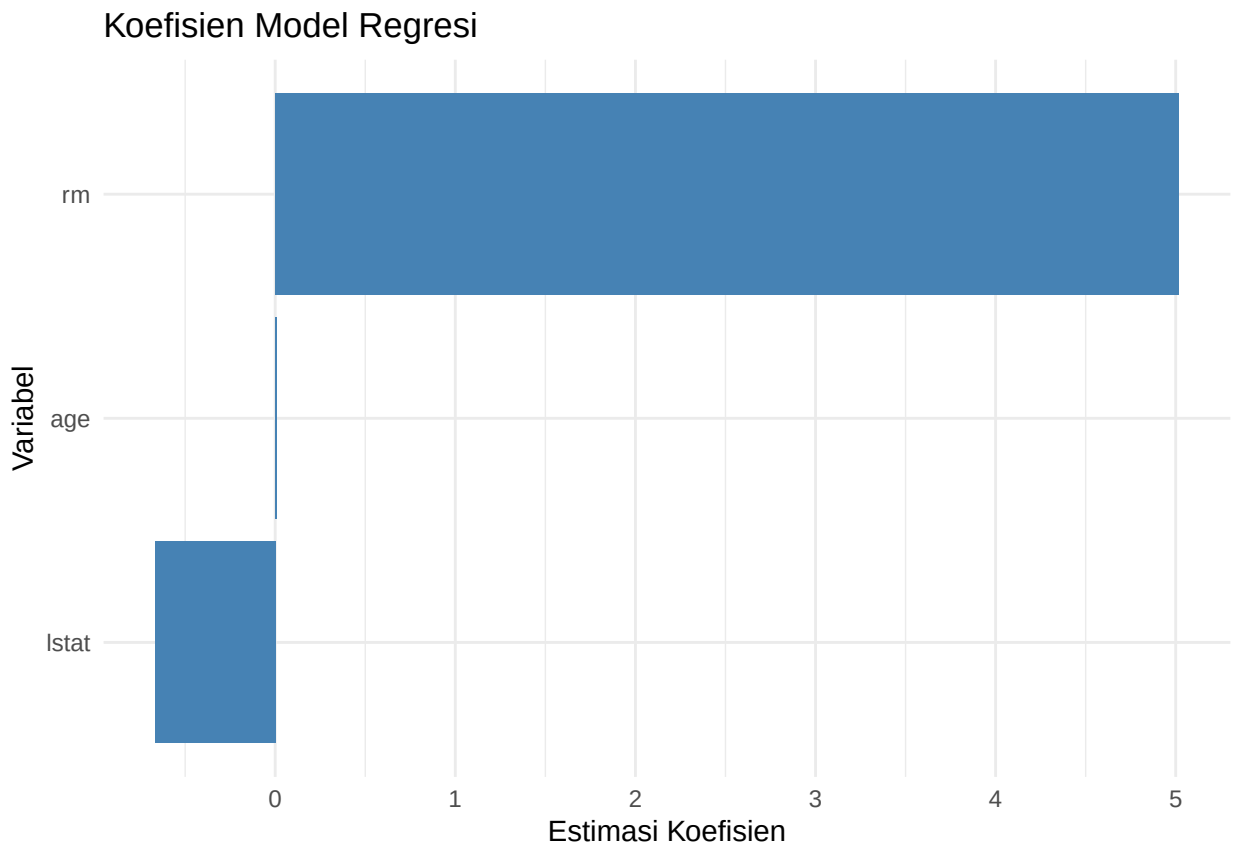


Pengertian:

Plot ini digunakan untuk melihat pola pada residual. Hasil menunjukkan pola menyebar tidak merata dan menyimpang dari garis horizontal (0), yang memperkuat adanya heteroskedastisitas. Idealnya, titik-titik menyebar secara acak di sekitar nol.

3.3 Plot Koefisien Model

```r
tidy(model) %>%
  filter(term != "(Intercept)") %>%
  ggplot(aes(x = reorder(term, estimate), y = estimate)) +
  geom_col(fill = "steelblue") +
  coord_flip() +
  labs(title = "Koefisien Model Regresi", x = "Variabel", y = "Estimasi Koefisien") +
  theme_minimal()
```



Pengertian:

Visualisasi ini menunjukkan nilai koefisien dari masing-masing prediktor. Dapat dilihat bahwa rm memiliki pengaruh positif dan paling besar, lstat berpengaruh negatif dan age memiliki pengaruh sangat kecil, mendekati nol.