ANALISIS MODEL TERBAIK DATA AUTO-MPG

01112200028

Zefanya Ekanugraha

Zefanya Ekanugraha

2023-02-20

Deskripsi Masalah

Pada penelitian ini akan dicari model terbaik untuk memodelkan **mpg** pada data **autompg**. Namun data yang dipakai memiliki data hilang dan memiliki total sembilan variabel yang harus diseleksi mana yang harus dimasukkan sebagai variabel independen dan mana yang tidak.

Pertama didefinisikan dahulu library-library yang akan dipakai.

```
library(readxl)
data = read_xlsx("/Users/CATUR/Documents/Bahan Belajar/Kuliah 2023/Data Mining dan
Analisis Prediktif/UTS/autompg/auto-mpg-excel.xlsx") dim(data)
```

```
## [1] 398   9
```

Data meliputi 398 baris dan 9 kolom. Karena data yang dipakai hanya sedikit, yaitu 398 maka digunakan multiple linear regression dan bukan decision tree.

Deskripsi Data

Data yang digunakan diambil dari UCI machine learning repository (Quinlan, 1993). Variabel mpg merepresentasikan galon yang dihabiskan per mil, variabel cylinder merepresentasikan seberapa banyak silinder yang dimiliki suatu mobil, **displacement** merepresentasikan perpindahan mesin, **horsepower** merepresentasikan tenaga mobil, **weight** merepresentasikan berat mobil, **acceleration** menandakan waktu untuk akselerasi dari 0 sampai 60 mil per jam, **model year** menandakan tahun model dibuat, **origin** menandakan tempat produksi suatu mobil, antara lain Amerika (disimbolkan 1), Eropa (disimbolkan 2), dan Jepang (disimbolkan 3), **car name** menandakan nama kendaraan, lima merupakan variabel numerik dan empat merupakan variabel kategoris (ASA, 1983)

Pada data ini terdapat 6 variabel kosong. Menurut Miller & Forte (2017) data kosong, ada tiga jenisnya yaitu **Missing Completely at Random (MCAR)**, **Missing at Random (MAR)**, dan **Missing Not at Random (MNAR)**, untuk memeriksanya digunakan tes statistik Little.

Merubah tipe data

Menurut metadata, maka ada dua variabel yang merupakan data kategorik, yaitu **origin** dan **car names**. Namun, **car names** dan **model year** tidak dimasukkan ke regresi karena tidak dapat dikodekan, seperti **origin**, tanpa membuat variabel menjadi banyak.

Sehingga data memiliki 8 variabel independen

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8$$

$$x_1 = cylinders, x_2 = displacement, x_3 = horsepower, x_4 = weight, x_5 = acceleration$$

$$, x_6 = variabelsatu, x_7 = variabeldua, x_8 = variabeltiga$$

Imputasi Data

Karena p-value dari tes MCAR (0.264>0.01), maka menurut dokumentasi, H_0nya yaitu data bersifat MCAR tidak dapat ditolak, maka bisa menggunakan imputasi rata-rata (Little & Rubin, 2020)

```
mcar_test(data_mentah)
```

```
## # A tibble: 1 × 4
##   statistic    df p.value missing.patterns
##       <dbl> <dbl>   <dbl>            <int> ## 1      7.65     6   0.265
```

```
2 data_mentah$horsepower = impute_mean(data_mentah$horsepower)
```

Pengkodean variabel kategoris

```
variabel_dummy <- model.matrix(~ factor(data$origin) - 1, data = data_mentah)
colnames(variabel_dummy) <- c("variabelsatu", "variabeldua", "variabeltiga")
data_mentah = data_mentah[-c(7)]
data_mentah = cbind(data_mentah, variabel_dummy)
```

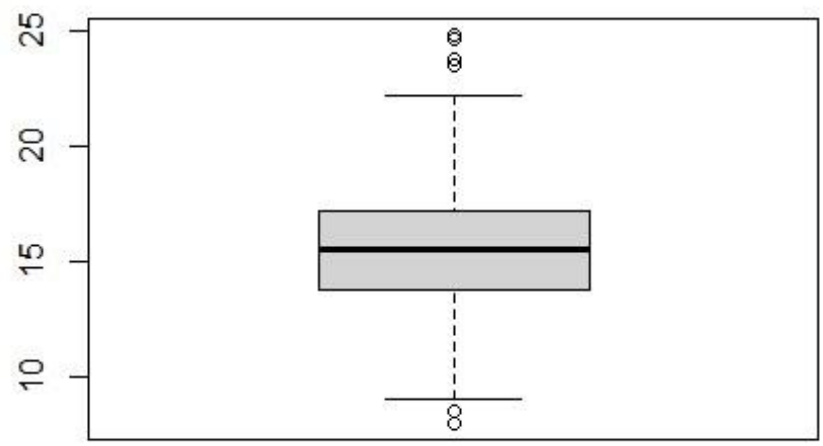
$$x_1 = \begin{cases} 1, \textit{jika origin berasal dari Amerika} \\ 0, \textit{jika tidak} \end{cases},$$

$$x_2 = \begin{cases} 1, & \textit{jika origin berasal dari Eropa} \\ 0, & \textit{jika tidak} \end{cases}$$

$$x_3 = f(x) = \begin{cases} 1, \textit{jika origin berasal dari Jepang} \\ 0, \quad \textit{jika tidak} \end{cases}$$

Exploratory data analysis

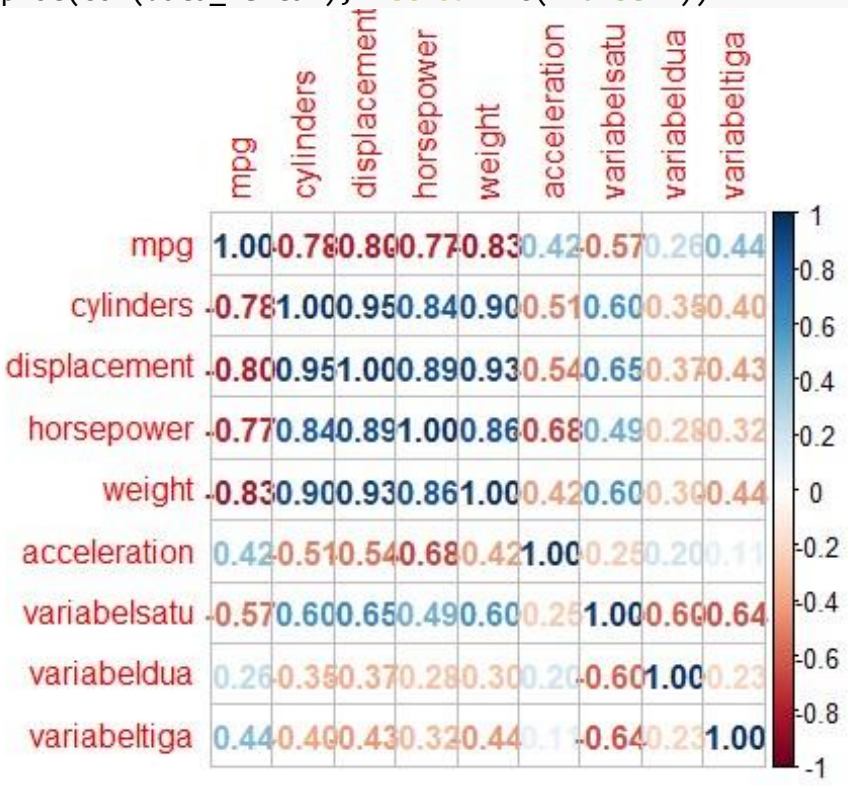
```
boxplot(data_mentah$acceleration)
```



Terlihat ada enam outlier pada variabel acceleration, namun tidak diapa-apakan karena data berupa mobil-mobil yang spesifik. Grafik lainnya ada di lampiran

Pengembangan Model

```
corrplot(cor(data_mentah), method = c("number"))
```



Terlihat ada masalah multikolinieritas (variabel yang satu berkorelasi dengan yang lain), sehingga akan dilakukan seleksi variabel. Akan digunakan **forward stepwise**. Dengan metode stepwise

diketahui manakah variabel independen paling baik untuk dimasukkan dalam model

```
Residuals:
     Min       1Q   Median       3Q      Max
-12.4423  -2.8431  -0.3791   2.2430  14.8601

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  42.8760739  1.0934458  39.212  < 2e-16 ***
weight       -0.0050451  0.0005377  -9.383  < 2e-16 ***
horsepower   -0.0490782  0.0108839  -4.509 8.60e-06 ***
variabeltiga  2.6990839  0.6508396   4.147 4.13e-05 ***
variabeldua   1.2283868  0.6369403   1.929   0.0545 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.185 on 393 degrees of freedom
Multiple R-squared:  0.7162,    Adjusted R-squared:  0.7133
F-statistic: 247.9 on 4 and 393 DF,  p-value: < 2.2e-16

--------------- Variance Inflating Factor (VIF) ---------------
Multicollinearity Problem: Variance Inflating Factor (VIF) is bigger than 10 (Continuous Variable) or is bigger t
han 2.5 (Categorical Variable)
      weight   horsepower variabeltiga  variabeldua
    4.699729     3.918082     1.531390     1.336255
> |
```

Terlihat terpilih variabel **weight**, **horsepower**, **variabeldua**, dan **variabeltiga**. Juga dihitungkan VIF (ukuran multikolinieritas) tidak lebih dari 10 untuk variabel kontinu dan tidak lebih dari 2.5 untuk variabel kategoris. Sehingga, masalah multikolinieritas sudah selesai

```
model = lm(data_mentah$mpg ~ data_mentah$horsepower + data_mentah$weight +
data_mentah$variabeldua + data_mentah$variabeltiga) summary1 =
summary(model) summary1
```

```
##
## Call:
## lm(formula = data_mentah$mpg ~ data_mentah$horsepower + data_mentah$weight
+
##     data_mentah$variabeldua + data_mentah$variabeltiga) ##
## Residuals:
##      Min       1Q   Median       3Q      Max  ##
-12.4423  -2.8431  -0.3791   2.2430  14.8601
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)              42.8760739  1.0934458  39.212  < 2e-16 ***
## data_mentah$horsepower   -0.0490782  0.0108839  -4.509 8.60e-06 ***
## data_mentah$weight       -0.0050451  0.0005377  -9.383  < 2e-16 *** ##
data_mentah$variabeldua   1.2283868  0.6369403   1.929   0.0545 .   ##
data_mentah$variabeltiga  2.6990839  0.6508396   4.147 4.13e-05 *** ## --
-
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 ##
## Residual standard error: 4.185 on 393 degrees of freedom
## Multiple R-squared:  0.7162, Adjusted R-squared:  0.7133
## F-statistic: 247.9 on 4 and 393 DF,  p-value: < 2.2e-16
```

Saat keempat variabel tersebut dimasukkan ke dalam model regresi, maka terlihat bahwa variabel-variabel independen dapat menjelaskan 71.33 % dari data. Dari statistik F, juga dapat dilihat bahwa salah satu dari

$$\beta_1, \beta_2, \ldots \beta_4 \neq 0$$

.Namun ingin dilihat apakah masih ada model yang lebih baik.

Transformasi Akar

Transformasi dapat membuat model menjadi lebih baik (Mendenhall & Sincich, 2020). Dipilih transformasi akar pada variabel independen karena tidak mengubah nilai variabel kategoris yang bernilai nol maupun satu.

```
data_mentah$mpg = as.double(data_mentah$mpg)
data_terpakai = cbind(data_mentah$mpg,data_mentah$horsepower,
data_mentah$weight, data_mentah$variabeldua, data_mentah$variabeltiga)
data_terpakai = as.data.frame(data_terpakai) horsepower =
sqrt(data_terpakai$V2);weight = as.double(sqrt(data_terpakai$V3));
variabeldua = sqrt(data_terpakai$V4);variabeltiga = sqrt(data_terpakai$V5);
mpg = data_terpakai$V1 data_tertransformasi =
data.frame(mpg,horsepower,weight,variabeldua,variabeltiga)
data_tertransformasi = as.data.frame(data_tertransformasi) m2 =
lm(data_tertransformasi$mpg ~ data_tertransformasi$horsepower +
```

```
data_tertransformasi$weight
        + data_tertransformasi$variabeldua +
data_tertransformasi$variabeltiga)
ANOVA2 = anova(m2)
summary(m2)

##
## Call:
## lm(formula = data_tertransformasi$mpg ~ data_tertransformasi$horsepower +
##     data_tertransformasi$weight + data_tertransformasi$variabeldua +  ##
data_tertransformasi$variabeltiga)
##
## Residuals:
##      Min       1Q   Median       3Q      Max  ##
-12.1681  -2.6339  -0.2863   2.1551  14.5356
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      64.44226    1.98739  32.426  < 2e-16 ***
## data_tertransformasi$horsepower   -1.27952    0.23283  -5.495 7.02e-08 ***
## data_tertransformasi$weight       -0.53145    0.05918  -8.980  < 2e-16 ***
## data_tertransformasi$variabeldua   0.93031    0.62198   1.496  0.13553
## data_tertransformasi$variabeltiga  2.35721    0.63962   3.685  0.00026 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 ##
## Residual standard error: 4.066 on 393 degrees of freedom
## Multiple R-squared:  0.7322, Adjusted R-squared:  0.7294
## F-statistic: 268.6 on 4 and 393 DF,  p-value: < 2.2e-16
```

Setelah itu dilakukan pengujian hipotesis bahwa: H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$ H_a: Paling sedikit salah satu koefisien tidak bernilai nol (Mendenhall & Sincich, 2012). Terlihat bahwa model final p-valuenya < 0.01

```
summary2 = summary(m2)
```

Model Final yang terbentuk adalah

$$Y = \beta_0 + \beta_1\sqrt{x_1} + \beta_2\sqrt{x_2} + \beta\sqrt{x_3} + \beta\sqrt{x_4}$$

dengan x_1: horsepower, x_2: weight, x_3: variabeldua,x_4:variabeltiga

Analisis Residu

Model Saat diuji asumsi klasiknya, antara lain normalitas dari residu, heteroskedastisitas, autokolerasi, dan multikolinieritas (Mendenhall & Sincich, 2012).

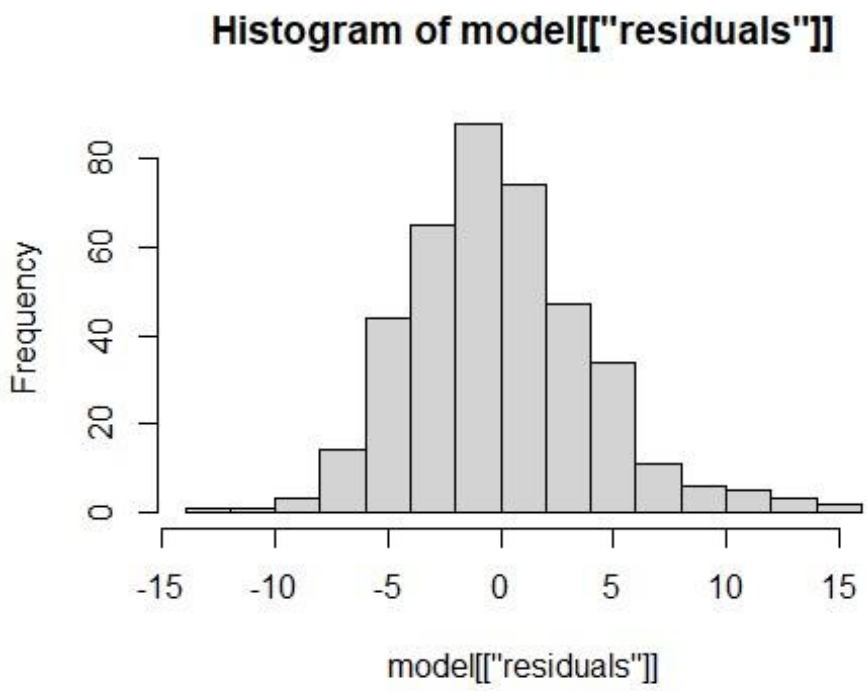
Tes kenormalan

Dilakukan uji kenormalan Kolmogorov-Smirnov terhadap residu

```
ANOVA = anova(m2)
ANOVA$`Sum Sq`[4]

## [1] 224.4895

hist(model[["residuals"]])
```




```
ks.test(m2[["residuals"]],"pnorm",mean(m2[["residuals"]]),
sd(m2[["residuals"]]))
```

```
## Warning in ks.test.default(m2[["residuals"]], "pnorm",
mean(m2[["residuals"]]),
## : ties should not be present for the Kolmogorov-Smirnov test

##
##  Asymptotic one-sample Kolmogorov-Smirnov test ##
## data:  m2[["residuals"]]
## D = 0.068475, p-value = 0.04787
## alternative hypothesis: two-sided
```

Saat digunakan $\alpha = 0.01$ maka H_0: residu berdistribusi normal tidak dapat ditolak

tes autokorelasi

```
durbinWatsonTest(m2)
##  lag Autocorrelation D-W Statistic p-value
##    1       0.5485777     0.8975319       0
##  Alternative hypothesis: rho != 0
```

Durbin-Watson jauh dari 2. Sehingga ada indikasi autokorelasi. Namun data tidak mendekati nol sehingga tidak ada korelasi yang kuat

Tes Multikolinieritas

```
vif(m2)
```

```
##   data_tertransformasi$horsepower       data_tertransformasi$weight
##                          4.097338                          4.915873
##  data_tertransformasi$variabeldua data_tertransformasi$variabeltiga
##                          1.350197                          1.567239
```

Seperti dalam tahap pemilihan variabel, VIF masih dapat ditoleransi (tidak melebihi 10 untuk variabel kuantitatif dan 2.5 untuk variabel kategoris).

Tes Heteroskedastisitas

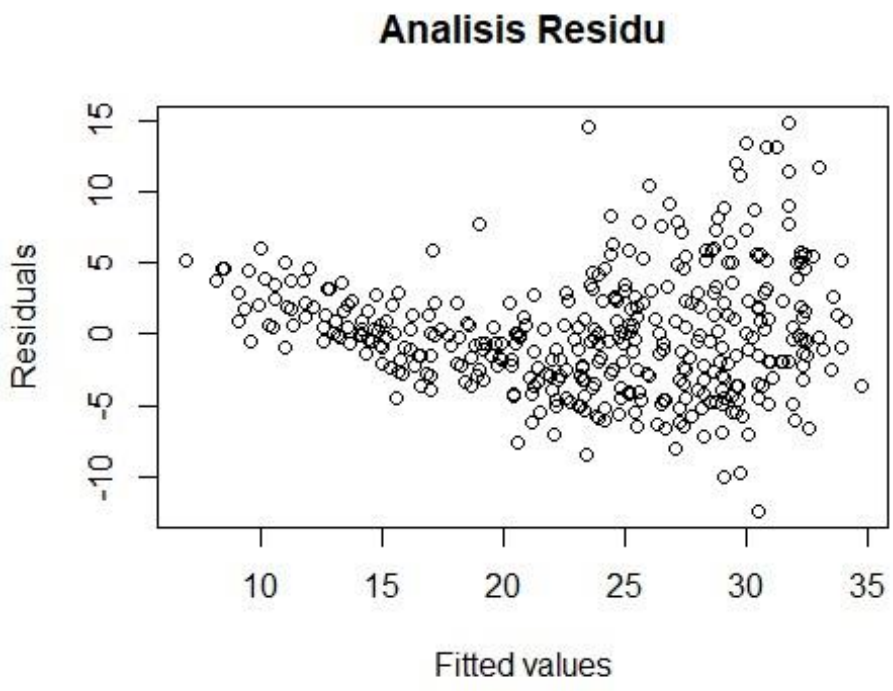
```
bptest(m2, studentize = FALSE)

##
##  Breusch-Pagan test
##
## data:  m2
## BP = 59.219, df = 4, p-value = 4.232e-12
```

Melalui uji Breusch-Pagan terlihat p-valuenya < 0.01, sehingga H_0: residu adalah homokesdastis ditolak, berarti ada indikasi heteroskedastisitas.

Sekarang dilakukan pengekstrakan data-data analisis residu diantaranya SSE yang ada dalam ANOVA.

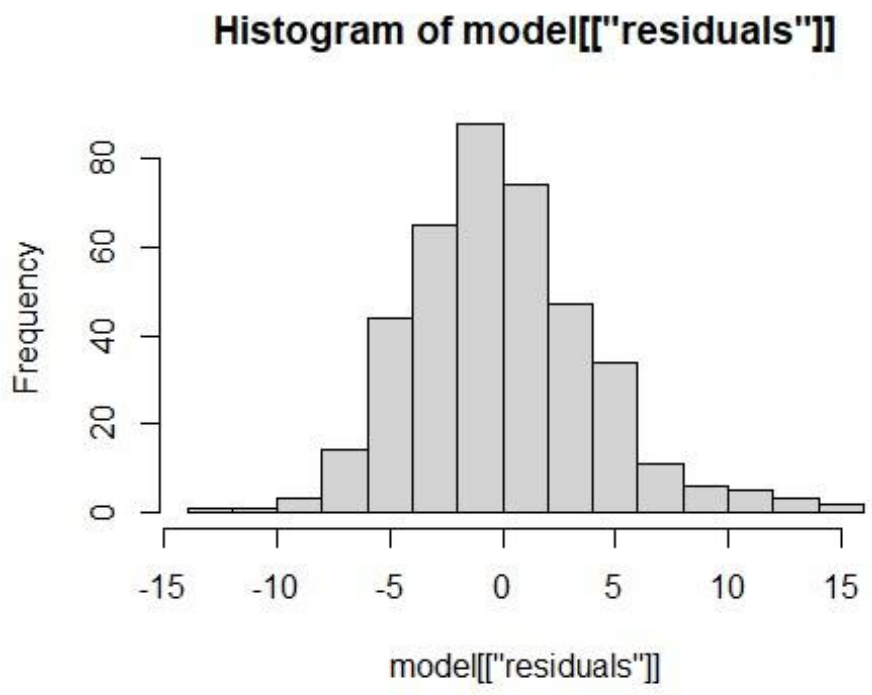
```
plot(model$fitted.values, model$residuals,main = "Analisis Residu",
xlab = "Fitted values", ylab = "Residuals")
```




```
ANOVA$`Sum Sq`[4] ## [1]

224.4895

hist(model[["residuals"]])
```



Prediksi

Akan coba diprediksi dua baris data baru

```
regresi = function(horsepower, weight, variabeldua, variabeltiga){
for(i in 1:length(horsepower))
  m2 <- lm(data_tertransformasi$mpg ~ data_tertransformasi$horsepower +
data_tertransformasi$weight + data_tertransformasi$variabeldua +
data_tertransformasi$variabeltiga)
  prediksi = coef(m2)[1] + coef(m2)[2]*horsepower + coef(m2)[3]*weight +
coef(m2)[4]*variabeldua + coef(m2)[5]*variabeltiga   return(prediksi)
}

regresi(12,41,1,0)

## (Intercept)  ##
28.22898
regresi(12,40,0,0)

## (Intercept)
##    27.83012
```

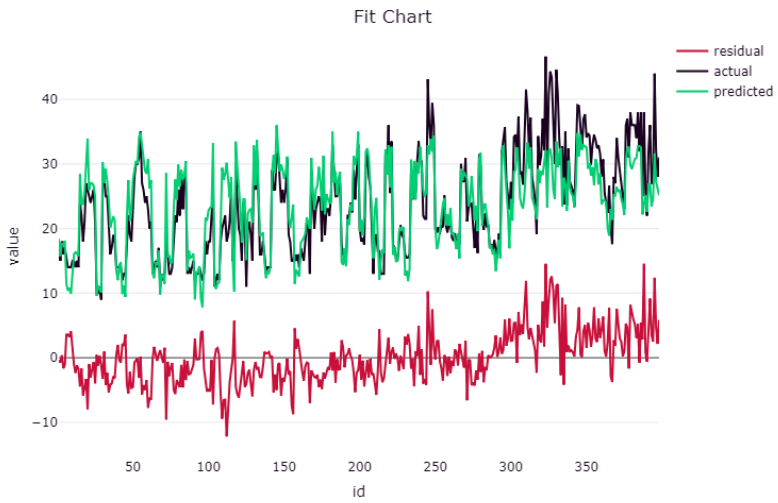
Evaluasi Model

```
ANOVA1 = anova(model) predicted1 =
model[["fitted.values"]] predicted2 =
m2[["fitted.values"]]
baris_pertama = paste("Didapatkan R kuadrat dari model pertama
adalah:",summary1[["adj.r.squared"]])
baris_kedua = paste("Didapatkan R kuadrat dari model final adalah:",
summary2[["adj.r.squared"]])
baris_ketiga = paste("didapatkan MSE dari model pertama adalah:",
ANOVA1$`Mean Sq`[5])
baris_keempat = paste("didapatkan MSE dari model final adalah:", ANOVA2$`Mean
Sq`[5])
baris_kelima = paste("didapatkan MAPE dari model pertama adalah:",
mape(data$mpg, predicted1))
baris_keenam = paste("didapatkan MAPE dari model final adalah",
mape(data$mpg, predicted2))
baris_ketujuh = paste("didapatkan RMSE dari model pertama adalah",
rmse(data$mpg,predicted1))
baris_kedelapan = paste("didapatkan RMSE dari model final adalah:",
rmse(data$mpg, predicted2)) tabel =
rbind(baris_pertama,baris_kedua,baris_ketiga,baris_keempat,baris_kelima,
baris_keenam, baris_ketujuh,baris_kedelapan) tabel
```

```
##                 [,1]
## baris_pertama   "Didapatkan R kuadrat dari model pertama adalah:
0.71329874130811"
## baris_kedua     "Didapatkan R kuadrat dari model final adalah:
0.729433087468349"
## baris_ketiga    "didapatkan MSE dari model pertama adalah:
17.5144683019821"
## baris_keempat   "didapatkan MSE dari model final adalah: 16.5288273749557"
## baris_kelima    "didapatkan MAPE dari model pertama adalah:
0.138169091540843"
## baris_keenam    "didapatkan MAPE dari model final adalah
0.132164235734104"
## baris_ketujuh   "didapatkan RMSE dari model pertama adalah
4.15865811207934"
## baris_kedelapan "didapatkan RMSE dari model final adalah:
4.03994786969377"
```

Terlihat bahwa R^2 naik dan MSE menurun. Juga dari pengukuran MAPE dan RMSE terlihat turun galatnya.

Visualisasi model



Terlihat bahwa model cenderung tidak mampu untuk memprediksi data outlier.

Kesimpulan

Dengan dilakukannya transformasi akar dapat ditemukan model yang lebih baik daripada model sebelumnya (yang didapat dari seleksi variabel *stepwise*). Juga terlihat bahwa data outlier mempengaruhi penelitian (sehingga membuat galat bertambah). Untuk penelitian selanjutnya dapat ditentukan dan digunakan pemberat pada model maupun dapat menggunakan seleksi variabel yang lebih baik daripada seleksi *stepwise* sederhana, sehingga dapat juga ditemukan interaksi antara variabel independen.

Model year yang tidak masuk dalam model regresi ini juga dapat dimasukkan (variabel year terdiri dari 12 variabel berbeda, harus diatur cara pengkodeannya. Outlier juga dapat dikelola dengan metode-metode tertentu agar lebih baik. Juga dapat dilakukan metode penyembuhan heteroskedastisitas agar dapat dibuat residunya bersifat homokedastis.

Referensi

American Statistical Association. http://lib.stat.cmu.edu/datasets/cars.desc

Little, R., & Rubin, D. (2020). *Statistical Analysis with Missing Data, 3rd ed.*. Wiley.

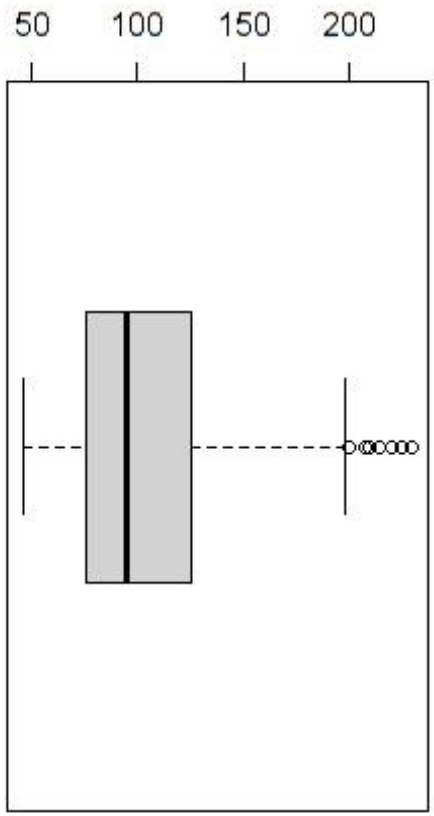
Mendenhall, William & Sincich, Terry. (2012). *A Second Course in Statistics : Regression Analysis*. Seventh Edition. Pearson Education,. Inc.

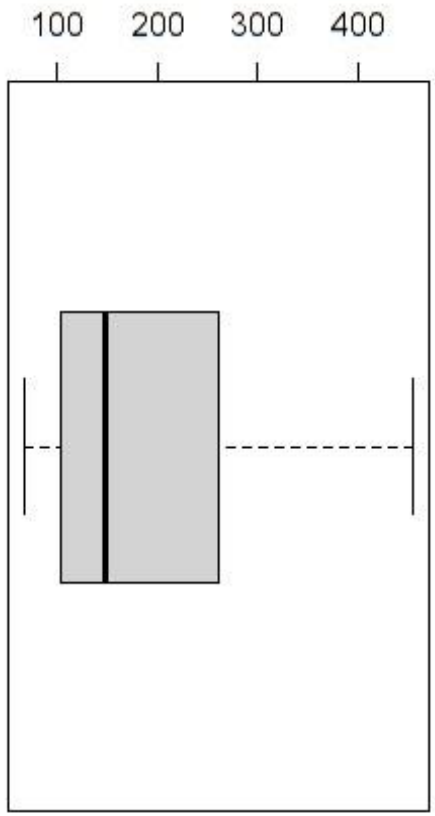
Sheather, S. (2009). *A Modern Approach to Regression with R*. Springer.

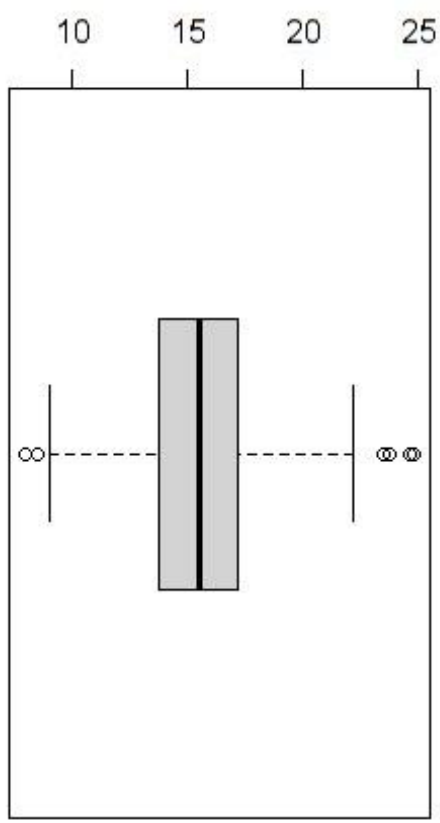
Quinlan, R. (1993). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

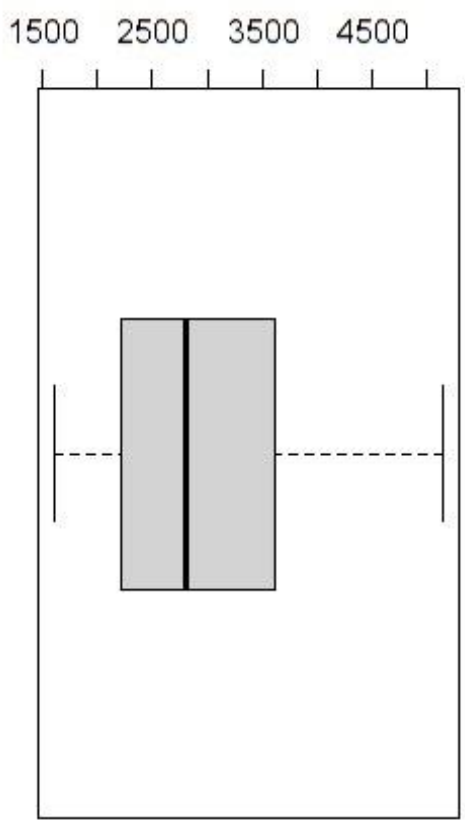
Lampiran

```
boxplot(data_mentah$displacement);boxplot(data_mentah$horsepower);boxplot(dat
a_mentah$weight);boxplot(data_mentah$acceleration)
```


50
100
150
200


100
200
300
400


10
15
20
25


1500
2500
3500
4500