**PRAKTIKUM 5 – DATA ANALITIK**

**REGRESI BERGANDA DENGAN SEBAGIAN PREDIKTOR**

**BERSIFAT KATEGORIK DAN REGRESI NON LINIER**

Jika variabel dependen-nya dihubungkan dengan lebih dari satu variabel independen, maka persamaan yang dihasilkan adalah persamaan regresi linier berganda (*multiple linier regression*). Dalam hal ini kita membatasi pada kasus dua peubah bebas X1 dan X2 saja. Dengan hanya dua peubah bebas, persamaan regresi contohnya menjadi :



Variabel terikat (respon) dalam regresi linier berganda harus bersifat kontinu sedangkan variabel bebas dapat seluruhnya bersifat kontinu atau sebagian bersifat kategorik. Bila peubah bebas (predictor) bersifat kategori maka dapat diubah menjadi sekumpulan variabel dummy dengan catatan bila dalam variabel kategorik tersebut terdapat n kategori maka harus terdapat sejumlah (n-1) variabel dummy.

Salah satu ukuran kebaikan model adalah dengan melihat koefisien determinasi R2 yang menyatakan proporsi keragaman variabel Y yang dapat dijelaskan oleh variabel X. Namun penggunaan yang lebih baik adalah dengan menggunakan nilai **R-Sq(adj)**, yang merupakan nilai estimasi yang tidak bias (*unbiased estimate*) dari populasi.

**Kasus 1 : Regresi dengan satu var independent bersifat kualitatif (2 kelas) 🡪1 var dummy**

Y =vol penjualan;

X1=pengeluaran iklan;

X2=media (1=tv, 0=koran)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Perusahaan | Volume Penjualan (Y) | Pengeluaran Iklan (X1) |  | Jenis Media (x2) |  | X2 |
| 1 | 620 | 250 |  | tv |  | 1 |
| 2 | 140 | 120 |  | koran |  | 0 |
| 3 | 500 | 175 |  | tv |  | 1 |
| 4 | 210 | 135 |  | koran |  | 0 |
| 5 | 460 | 180 |  | tv |  | 1 |
| 6 | 250 | 140 |  | koran |  | 0 |
| 7 | 200 | 130 |  | koran |  | 0 |
| 8 | 380 | 170 |  | tv |  | 1 |
| 9 | 400 | 200 |  | tv |  | 1 |
| 10 | 215 | 150 |  | koran |  | 0 |
| 11 | 395 | 175 |  | tv |  | 1 |
| 12 | 425 | 215 |  | tv |  | 1 |
| 13 | 235 | 145 |  | koran |  | 0 |
| 14 | 475 | 190 |  | tv |  | 1 |
| 15 | 195 | 130 |  | koran |  | 0 |

**General Regression (MINITAB)**

Regression Equation

X2 (1=tv,

0=koran)

0 (koran) Y = -75,412 + 2,07672 X1

1 (tv) Y = 53,2126 + 2,07672 X1

Coefficients

Term Coef SE Coef T P

Constant -11,0997 100,926 -0,10998 0,914

X1 2,0767 0,607 3,41996 0,005

X2 (1=tv, 0=koran)

0 -64,3123 21,425 -3,00176 0,011

Summary of Model

S = 46,0191 R-Sq = 91,06% R-Sq(adj) = 89,57%

PRESS = 54182,6 R-Sq(pred) = 80,95%

Hasil regresi mengindikasikan bahwa variabel kualitatif jenis media iklan yang digunakan berpengaruh terhadap volume penjualan. Perusahaan yang menggunakan media tv mempunyai volume penjualan lebih tinggi dibandingkan yang menggunakan surat kabar yang ditunjukkan oleh perbedaan koefisien titik potong dengan sumbu Y pada kedua persamaan regresi di atas yaitu -75,412 dan 53,2126 dengan perbedaan sebesar 128,625.

Koefisien regresi pengeluaran iklan dan jenis media iklan juga signifikan yang terlihat pada nilai p yang kurang dari 5 %.

Analysis of Variance

Source DF Seq SS Adj SS Adj MS F P

Regression 2 258937 258937 129468 61,1347 0,000001

X1 1 239855 24770 24770 11,6961 0,005079

X2 (1=tv, 0=koran) 1 19082 19082 19082 9,0105 0,011031

Error 12 25413 25413 2118

Lack-of-Fit 10 19888 19888 1989 0,7199 0,706453

Pure Error 2 5525 5525 2763

Total 14 284350

**R script :**

|  |
| --- |
| > library(RMySQL)  > con = dbConnect(MySQL(), user = 'root', password = '', dbname =  + 'db\_reg', host = 'localhost')  > myQuery <- "select \* from reg;"  > reg\_stev <- dbGetQuery(con, myQuery)  > View(reg\_stev) |
| > head(reg\_stev)  Perusahaan Y X1 X2  1 1 620 250 1  2 2 140 120 0  3 3 500 175 1  4 4 210 135 0  5 5 460 180 1  6 6 250 140 0  > model=lm(Y ~ X1 + X2, data = reg\_stev)  > model  Call:  lm(formula = Y ~ X1 + X2, data = reg\_stev)  Coefficients:  (Intercept) X1 X2  -75.412 2.077 128.625  > summary(model)  Call:  lm(formula = Y ~ X1 + X2, data = reg\_stev)  Residuals:  Min 1Q Median 3Q Max  -74.707 -23.947 5.055 30.094 83.361  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) -75.4120 84.2261 -0.895 0.38821  X1 2.0767 0.6072 3.420 0.00508 \*\*  X2 128.6246 42.8498 3.002 0.01103 \*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 46.02 on 12 degrees of freedom  Multiple R-squared: 0.9106, Adjusted R-squared: 0.8957 |

**Kasus 2 : Regresi dengan satu var independent bersifat kualitatif (3 kelas) 🡪2 var dummy**

Y=gaji(juta) X1=masa kerja (tahun) X2=pendidikan (sma, diploma, sarjana)

00= sma, 10=diploma, 01=sarjana

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Gaji Y | Masa Kerja(X1) | Pendidikan |  | X2 | X3 |
| 3 | 1 | Sarjana |  | 0 | 1 |
| 2.4 | 5 | Diploma |  | 1 | 0 |
| 2.15 | 12 | SMA |  | 0 | 0 |
| 3.8 | 5 | Sarjana |  | 0 | 1 |
| 5 | 10 | Sarjana |  | 0 | 1 |
| 1.75 | 5 | SMA |  | 0 | 0 |
| 2 | 1 | Diploma |  | 1 | 0 |
| 2.2 | 3 | Diploma |  | 1 | 0 |
| 2.05 | 10 | SMA |  | 0 | 0 |
| 2.8 | 8 | Diploma |  | 1 | 0 |
| 3.2 | 2 | Sarjana |  | 0 | 1 |
| 1.5 | 1 | SMA |  | 0 | 0 |
| 1.55 | 2 | SMA |  | 0 | 0 |
| 3.6 | 4 | Sarjana |  | 0 | 1 |
| 1.9 | 8 | SMA |  | 0 | 0 |
| 3.5 | 15 | Diploma |  | 1 | 0 |
| 2.35 | 15 | SMA |  | 0 | 0 |
| 2.7 | 20 | SMA |  | 0 | 0 |
| 6 | 15 | Sarjana |  | 0 | 1 |
| 2.45 | 17 | SMA |  | 0 | 0 |

**General regression (MINITAB)**

x2

0 y = 0.973004 + 0.107144 x1

1 y = 1.89428 + 0.107144 x1

2 y = 3.43928 + 0.107144 x1

Coefficients

Term Coef SE Coef T P

Constant 0,97300 0,206329 4,7158 0,000

Masa Kerja X1 0,10714 0,015908 6,7353 0,000

Pendidikan X2 0,92127 0,227207 4,0548 0,001

Pendidikan X3 2,46627 0,216523 11,3904 0,000

Summary of Model

S = 0,394194 R-Sq = 90,08% R-Sq(adj) = 88,22%

PRESS = 4,34619 R-Sq(pred) = 82,65%

Analysis of Variance

Source DF Seq SS Adj SS Adj MS F P

Regression 3 22,5683 22,5683 7,5228 48,413 0,0000000

Masa Kerja X1 1 2,3014 7,0490 7,0490 45,364 0,0000048

Pendidikan X2 1 0,1067 2,5548 2,5548 16,441 0,0009198

Pendidikan X3 1 20,1602 20,1602 20,1602 129,741 0,0000000

Error 16 2,4862 2,4862 0,1554

Total 19 25,0545

Fits and Diagnostics for Unusual Observations

Obs Gaji Y Fit SE Fit Residual St Resid

19 6 5,04644 0,213644 0,953561 2,87844 R

R denotes an observation with a large standardized residual.

|  |
| --- |
| > library(RMySQL)  > con = dbConnect(MySQL(), user = 'root', password = '', dbname =  + 'db\_reg', host = 'localhost')  > myQuery <- "select \* from reg2;"  > reg2\_stev <- dbGetQuery(con, myQuery)  Warning message:  In .local(conn, statement, ...) :  Decimal MySQL column 0 imported as numeric  > View(reg2\_stev) |
| > head(reg2\_stev)  Y X1 X2 X3  1 3.00 1 0 1  2 2.40 5 1 0  3 2.15 12 0 0  4 3.80 5 0 1  5 5.00 10 0 1  6 1.75 5 0 0  > model\_reg2=lm(Y ~ X1 + X2 + X3, data = reg2\_stev)  > model\_reg2  Call:  lm(formula = Y ~ X1 + X2 + X3, data = reg2\_stev)  Coefficients:  (Intercept) X1 X2 X3  0.9730 0.1071 0.9213 2.4663  > summary(model\_reg2)  Call:  lm(formula = Y ~ X1 + X2 + X3, data = reg2\_stev)  Residuals:  Min 1Q Median 3Q Max  -0.54642 -0.23959 -0.00857 0.11270 0.95356  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 0.97300 0.20633 4.716 0.000233 \*\*\*  X1 0.10714 0.01591 6.735 4.79e-06 \*\*\*  X2 0.92127 0.22721 4.055 0.000920 \*\*\*  X3 2.46627 0.21652 11.390 4.36e-09 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 0.3942 on 16 degrees of freedom  Multiple R-squared: 0.9008, Adjusted R-squared: 0.8822  F-statistic: 48.41 on 3 and 16 DF, p-value: 2.997e-08 |

**Kasus 3 : Regresi non linier dengan satu variabel independent**

An engineer at a computer manufacturing company wants to understand the relationship between strength of plastic and the predictors temperature. He suspects the relationship between temperature and strength is quadratic. He collects 14 samples. The engineer subjects the samples to various temperatures, then measures the strength of the plastic (File : Plastic.mtw)

|  |  |
| --- | --- |
| Temp X | Strength Y |
| 185 | 5150 |
| 183 | 5125 |
| 187 | 5123 |
| 188 | 5140 |
| 189 | 5195 |
| 189 | 5190 |
| 192 | 5150 |
| 195 | 5155 |
| 196 | 5156 |
| 198 | 5162 |
| 193 | 5172 |
| 196 | 5196 |
| 200 | 5063 |
| 202 | 5025 |
| 196 | 5105 |
| 197 | 5176 |
| 195 | 5160 |
| 200 | 5100 |
| 205 | 4956 |
| 206 | 4960 |
| 203 | 4975 |
| 199 | 5063 |
| 200 | 5090 |
| 206 | 4946 |
| 196 | 5175 |
| 207 | 4905 |
| 208 | 4760 |

**General Regression Analysis: Strength Y versus Temp X (MINITAB)**

Regression Equation

Strength Y = -38711,3 + 459,45 Temp X - 1,20233 Temp X\*Temp X

Coefficients

Term Coef SE Coef T P

Constant -38711,3 4881,77 -7,92978 0,000

Temp X 459,5 49,83 9,21991 0,000

Temp X\*Temp X -1,2 0,13 -9,46251 0,000

|  |
| --- |
| Pembahasan : Kasus di atas diolah dengan menggunakan menu General Regression dalam Minitab. Hal ini karena model yang ingin digunakan adalah model kuadratik () yang lebih sesuai dengan pola datanya yang melengkung bila dibuat scatter plot. Berdasarkan uji statistik terhadap koefisien regresi di atas dapat dilihat bahwa seluruh koefisien regresi memiliki p value yang kurang dari taraf nyata (α) sebesar 1% yang berarti sangat signifikan. Hal ini menandakan bahwa ketiga koefisien secara sangat signifikan tidak sama dengan nol. |

Analysis of Variance

Source DF Seq SS Adj SS Adj MS F P

Regression 2 276023 276023 138012 126,950 0,000000

Temp X 1 178683 92414 92414 85,007 0,000000

Temp X\*Temp X 1 97341 97341 97341 89,539 0,000000

Error 24 26091 26091 1087

Lack-of-Fit 16 20690 20690 1293 1,915 0,177045

Pure Error 8 5402 5402 675

Total 26 302115

|  |
| --- |
| Berdasarkan Analysis of Variance (ANOVA) terlihat Regression memiliki p value < α (1%) berarti minimal terdapat satu koefisien regresi yang tidak sama dengan nol. Uji F pada ANOVA digunakan untuk mengevaluasi pengaruh semua variabel independen terhadap dependen). Hasil Lack of Fit juga menunjukkan tidak signifikan (p value > α) yang berarti model dapat diterima. |

Fits and Diagnostics for Unusual Observations

Obs Strength Y Fit SE Fit Residual St Resid

2 5125 5103,09 21,3071 21,9136 0,87089 X

27 4760 4836,52 17,3724 -76,5191 -2,73050 R

R denotes an observation with a large standardized residual.

X denotes an observation whose X value gives it large leverage.



**Bagaimana jika kasus 1 dimodelkan secara linier ? Nilai koefisien determinasinya (R-square) akan turun**

**General Regression Analysis: Strength Y versus Temp X**

Regression Equation

Strength Y = 7449,5 - 12,0059 Temp X

Coefficients

Term Coef SE Coef T P

Constant 7449,50 392,797 18,9653 0,000

Temp X -12,01 1,996 -6,0159 0,000

Summary of Model

S = 70,2658 **R-Sq = 59,14%** R-Sq(adj) = 57,51%

PRESS = 155356 R-Sq(pred) = 48,58%

Analysis of Variance

Source DF Seq SS Adj SS Adj MS F P

Regression 1 178683 178683 178683 36,1905 0,0000028

Temp X 1 178683 178683 178683 36,1905 0,0000028

Error 25 123432 123432 4937

Lack-of-Fit 17 118030 118030 6943 10,2827 0,0011422

Pure Error 8 5402 5402 675

Total 26 302115

Fits and Diagnostics for Unusual Observations

Obs Strength Y Fit SE Fit Residual St Resid

2 5125 5252,41 30,5092 -127,414 -2,01297 R

27 4760 4952,27 26,2888 -192,267 -2,95056 R

R denotes an observation with a large standardized residual.

Script R

|  |
| --- |
| > myQuery <- "select \* from reg3;"  > reg3\_stev <- dbGetQuery(con, myQuery)  > View(reg3\_stev)  > model\_reg3=lm(Y ~ X + X.2, data = reg3\_stev)  > model\_reg3  Call:  lm(formula = Y ~ X + X.2, data = reg3\_stev)  Coefficients:  (Intercept) X X.2  -38711.311 459.450 -1.202  > summary(model\_reg3)  Call:  lm(formula = Y ~ X + X.2, data = reg3\_stev)  Residuals:  Min 1Q Median 3Q Max  -76.519 -26.172 4.662 22.439 46.848  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) -3.871e+04 4.882e+03 -7.930 3.69e-08 \*\*\*  X 4.595e+02 4.983e+01 9.220 2.34e-09 \*\*\*  X.2 -1.202e+00 1.271e-01 -9.463 1.43e-09 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 32.97 on 24 degrees of freedom  Multiple R-squared: 0.9136, Adjusted R-squared: 0.9064  F-statistic: 127 on 2 and 24 DF, p-value: 1.721e-13 |

**Tugas :**

Gunakan data directmarketing.csv. Lakukan analisis regresi untuk melihat pengaruh dari salary dan gender terhadap amountspent serta pengaruh dari salary dan age terhadap amountspent. Lakukan pula uji asumsi terhadap kedua model regresi yang didapatkan.

**Model Regresi 1**

**Melihat pengaruh dari salary dan age terhadap amountspent**

> library(RMySQL)

> con = dbConnect(MySQL(), user = 'root', password = '', dbname =

+ 'db\_reg', host = 'localhost')

> myQuery <- "select \* from reg\_direct\_marketing;"

> reg\_direct\_marketing <- dbGetQuery(con, myQuery)

> myQuery <- "select \* from reg\_direct\_marketing;"

> reg\_direct\_marketing <- dbGetQuery(con, myQuery)

> View(reg\_direct\_marketing)

> head(reg\_direct\_marketing)

Y Salary Age X1 X2

1 755 47500 Old 0 1

2 1318 63600 Middle 1 0

3 296 13500 Young 0 0

4 2436 85600 Middle 1 0

5 1304 68400 Middle 1 0

6 495 30400 Young 0 0

> model\_dm=lm(Y ~ Salary+X1 + X2, data = reg\_direct\_marketing)

> model\_dm

Call:

lm(formula = Y ~ Salary + X1 + X2, data = reg\_direct\_marketing)

Coefficients:

(Intercept) Salary X1 X2

-61.18284 0.02236 -48.07869 232.79846

> summary(model\_dm)

Call:

lm(formula = Y ~ Salary + X1 + X2, data = reg\_direct\_marketing)

Residuals:

Min 1Q Median 3Q Max

-2434.5 -347.4 -50.8 278.2 3520.6

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -6.118e+01 4.715e+01 -1.298 0.194673

Salary 2.236e-02 8.949e-04 24.991 < 2e-16 \*\*\*

X1 -4.808e+01 6.394e+01 -0.752 0.452290

X2 2.328e+02 6.720e+01 3.464 0.000554 \*\*\*

---

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

Residual standard error: 679.2 on 996 degrees of freedom

Multiple R-squared: 0.502, Adjusted R-squared: 0.5005

F-statistic: 334.7 on 3 and 996 DF, p-value: < 2.2e-16

**Model Regresi 2**

**Melihat pengaruh dari salary dan gender terhadap amountspent**

> myQuery <- "select \* from reg\_direct\_marketing2;"

> reg\_direct\_marketing2 <- dbGetQuery(con, myQuery)

> View(reg\_direct\_marketing2)

> head(reg\_direct\_marketing2)

Y Salary Gender X1

1 755 47500 Female 0

2 1318 63600 Male 1

3 296 13500 Female 0

4 2436 85600 Male 1

5 1304 68400 Female 0

6 495 30400 Male 1

> model\_dm2=lm(Y ~ Salary+X1, data = reg\_direct\_marketing2)

> model\_dm2

Call:

lm(formula = Y ~ Salary + X1, data = reg\_direct\_marketing2)

Coefficients:

(Intercept) Salary X1

-25.1523 0.0218 38.6716

> summary(model\_dm2)

Call:

lm(formula = Y ~ Salary + X1, data = reg\_direct\_marketing2)

Residuals:

Min 1Q Median 3Q Max

-2180.6 -323.1 -53.7 282.8 3742.8

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -2.515e+01 4.680e+01 -0.537 0.591

Salary 2.180e-02 7.357e-04 29.626 <2e-16 \*\*\*

X1 3.867e+01 4.503e+01 0.859 0.391

---

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

Residual standard error: 687.2 on 997 degrees of freedom

Multiple R-squared: 0.4898, Adjusted R-squared: 0.4888

F-statistic: 478.6 on 2 and 997 DF, p-value: < 2.2e-16

**Uji Asumsi**

**Model Regresi 1, pengaruh dari salary dan age terhadap amountspent**

1.Normalitas

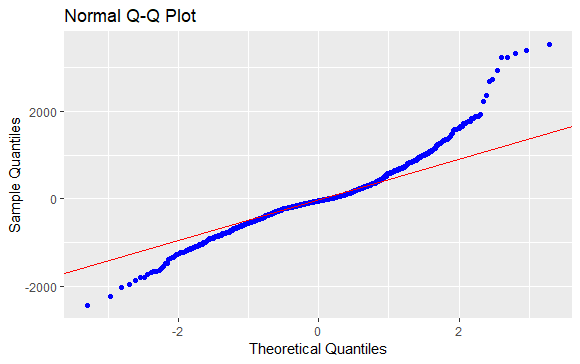
> library(olsrr)

> library(car)

> library(lmtest)

> library(ggpubr)

> ols\_plot\_resid\_qq(model\_dm)



>ols\_test\_normality(model\_dm)

-----------------------------------------------

Test Statistic pvalue

-----------------------------------------------

Shapiro-Wilk 0.9387 0.0000

Kolmogorov-Smirnov 0.0985 0.0000

Cramer-von Mises 87.3023 0.0044

Anderson-Darling 15.0475 0.0000

-----------------------------------------------

2. **Heteroskedastisitas**

> lmtest::bptest(model\_dm)

studentized Breusch-Pagan test

data: model\_dm

BP = 162.82, df = 3, p-value < 2.2e-16

**3. Autokorelasi**

> dwtest(model\_dm)

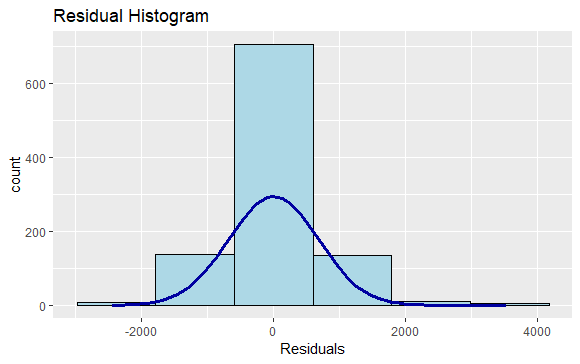
Durbin-Watson test

data: model\_dm

DW = 1.9794, p-value = 0.3717

alternative hypothesis: true autocorrelation is greater than 0

> ols\_plot\_resid\_hist(model\_dm)



**Uji Asumsi**

**Model Regresi 2, pengaruh dari salary dan gender terhadap amountspent**

1.Normalitas

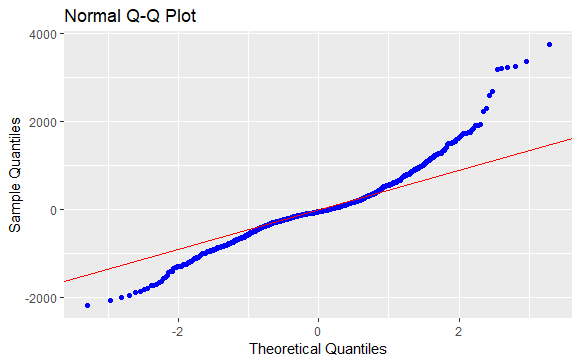
> library(olsrr)

> library(car)

> library(lmtest)

> library(ggpubr)

> ols\_plot\_resid\_qq(model\_dm2)



> ols\_test\_normality(model\_dm2)

-----------------------------------------------

Test Statistic pvalue

-----------------------------------------------

Shapiro-Wilk 0.9381 0.0000

Kolmogorov-Smirnov 0.1053 0.0000

Cramer-von Mises 86.7787 0.0043

Anderson-Darling 16.0621 0.0000

-----------------------------------------------

2. **Heteroskedastisitas**

> lmtest::bptest(model\_dm2)

studentized Breusch-Pagan test

data: model\_dm2

BP = 165.5, df = 2, p-value < 2.2e-16

**3. Autokorelasi**

> dwtest(model\_dm2)

Durbin-Watson test

data: model\_dm2

DW = 2.002, p-value = 0.5127

alternative hypothesis: true autocorrelation is greater than 0

> ols\_plot\_resid\_hist(model\_dm2)

