Analisis Regresi Linier Ganda (Tugas 2 Bagian C)

-

27 Maret 2024

## Memanggil data

Data yang dipilih: Delivery Time Data, from Montgomery and Peck (1982). The aim is to explain the time required to service a vending machine (Y) by means of the number of products stocked (X1) and the distance walked by the route driver (X2).

delivery <- read.csv(file="C:/Users/LENOVO/OneDrive/Documents/Learning R/Semester 2/delivery.csv", head=T)  
str(delivery)

## 'data.frame': 25 obs. of 4 variables:  
## $ X : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ n.prod : int 7 3 3 4 6 7 2 7 30 5 ...  
## $ distance: int 560 220 340 80 150 330 110 210 1460 605 ...  
## $ delTime : num 16.7 11.5 12 14.9 13.8 ...

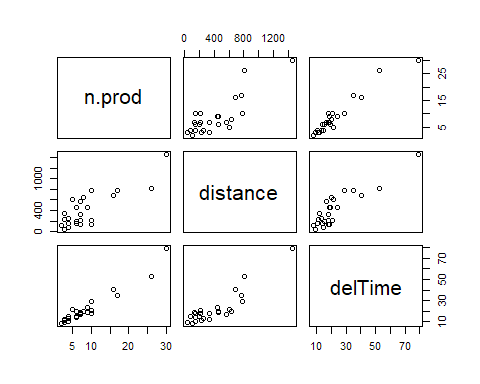
delivery.complete <- delivery[complete.cases(delivery),]  
delivery.complete

## X n.prod distance delTime  
## 1 1 7 560 16.68  
## 2 2 3 220 11.50  
## 3 3 3 340 12.03  
## 4 4 4 80 14.88  
## 5 5 6 150 13.75  
## 6 6 7 330 18.11  
## 7 7 2 110 8.00  
## 8 8 7 210 17.83  
## 9 9 30 1460 79.24  
## 10 10 5 605 21.50  
## 11 11 16 688 40.33  
## 12 12 10 215 21.00  
## 13 13 4 255 13.50  
## 14 14 6 462 19.75  
## 15 15 9 448 24.00  
## 16 16 10 776 29.00  
## 17 17 6 200 15.35  
## 18 18 7 132 19.00  
## 19 19 3 36 9.50  
## 20 20 17 770 35.10  
## 21 21 10 140 17.90  
## 22 22 26 810 52.32  
## 23 23 9 450 18.75  
## 24 24 8 635 19.83  
## 25 25 4 150 10.75

# 1. Menganalisis regresi berganda

### Membuat plot pencar

attach(delivery.complete)  
pairs(cbind(n.prod, distance, delTime))



# ERROR

Interpretasi:

* Terlihat ada hubungan linear positif antara jarak (distance) dengan waktu pengiriman (delTime) dan jumlah produk (n.prod).
* Korelasi antar dua variabel prediktor (jumlah produk dan jarak) sebesar 0.82 yang tergolong moderat. Hal ini mengindikasikan tidak terjadi multikolinearitas.
* Korelasi antar masing-masing variabel prediktor (jumlah produk dan jarak) dengan variabel respon (waktu pengiriman) berturut-turut sebesar 0,96 dan 0,89. Hal ini mengindikasikan adanya hubungan linear yang kuat antara masing-masing variabel prediktor dengan variabel respon. Sehingga asumsi linearitas dapat dianggap terpenuhi.

### Persamaan regresi dugaan

# Perhitungan menggunakan fungsi lm di program R  
mod.reg <- lm(delTime~n.prod+distance)  
summary(mod.reg)

##   
## Call:  
## lm(formula = delTime ~ n.prod + distance)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.7880 -0.6629 0.4364 1.1566 7.4197   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.341231 1.096730 2.135 0.044170 \*   
## n.prod 1.615907 0.170735 9.464 3.25e-09 \*\*\*  
## distance 0.014385 0.003613 3.981 0.000631 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.259 on 22 degrees of freedom  
## Multiple R-squared: 0.9596, Adjusted R-squared: 0.9559   
## F-statistic: 261.2 on 2 and 22 DF, p-value: 4.687e-16

# Perhitungan secara manual di program R  
X <- cbind(1, n.prod, distance)  
Y <- delTime  
b <- solve(t(X)%\*%X)%\*%t(X)%\*%Y  
b

## [,1]  
## 2.34123115  
## n.prod 1.61590721  
## distance 0.01438483

Persamaan regresi dugaan yang dihasilkan adalah:

* **Intersep**: Nilai intersep () adalah 2.3412. Ini menunjukkan bahwa saat jumlah produk () dan jarak () sama dengan 0, waktu pengiriman () diperkirakan sebesar 2.3412 unit. Namun, dalam konteks praktis, nilai intersep ini mungkin tidak memiliki makna yang signifikan karena nilai nol untuk kedua variabel prediktor mungkin tidak relevan atau mungkin tidak mungkin dalam konteks yang sesungguhnya.
* **Koefisien untuk** : Koefisien untuk () adalah 1.6159. Ini menunjukkan bahwa jika jumlah produk () meningkat sebesar 1 unit dan jarak () tetap, waktu pengiriman () diperkirakan akan meningkat sekitar 1.6159 unit.
* **Koefisien untuk** : Koefisien untuk () adalah 0.0144. Ini menunjukkan bahwa jika jarak () meningkat sebesar 1 unit dan jumlah produk () tetap, waktu pengiriman () diperkirakan akan meningkat sekitar 0.0144 unit.
* **Kualitas model**: Multiple adalah 0.9596, yang menunjukkan bahwa model dapat menjelaskan sekitar 95.96% variabilitas dalam waktu pengiriman () dengan menggunakan jumlah produk () dan jarak () sebagai prediktor. Adjusted mengoreksi untuk jumlah variabel independen dalam model, dan nilainya adalah 0.9559, yang masih tinggi. F-statistic menunjukkan bahwa model secara keseluruhan signifikan, dengan p-value yang sangat rendah.

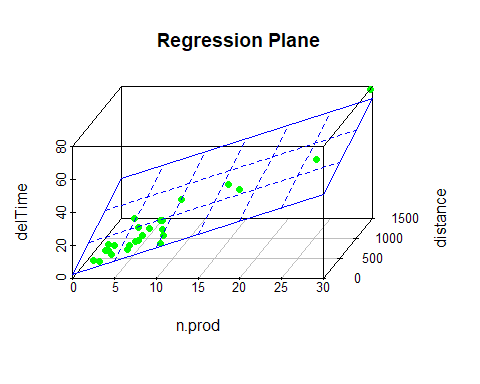
# Prediksi  
newdata = data.frame(n.prod=30, distance=80)  
predict(mod.reg, newdata=newdata)

## 1   
## 51.96923

Dari hasil prediksi menggunakan model regresi yang telah dibangun, diperoleh prediksi waktu pengiriman () untuk data baru dengan jumlah produk () sebesar 30 dan jarak () sebesar 80 adalah sekitar 51.97 unit waktu. Ini berarti berdasarkan model, jika terdapat 30 produk yang harus dikirimkan dengan jarak 80, diperkirakan waktu pengirimannya akan sekitar 51.97 unit waktu.

### Membuat plot persamaan regresi dugaan dan titik-titik pengamatan

library(scatterplot3d)  
plot3d <- scatterplot3d(n.prod, distance, delTime, angle=55, scale.y=, pch=16, color="green", main="Regression Plane")  
plot3d$plane3d(mod.reg, lty.box="solid", col="blue")



Persamaan regresi dugaan berikut berupa bidang:

# 2. Cek asumsi-asumsi dalam analisis regresi berganda

### 1) Asumsi Linearitas

library(GGally)

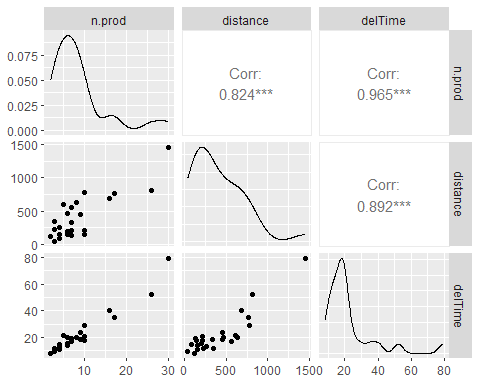
## Warning: package 'GGally' was built under R version 4.3.3

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 4.3.2

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

ggpairs(data=data.frame(n.prod, distance, delTime))



### 2) Asumsi Nonmultikolinearitas

library(car)

## Warning: package 'car' was built under R version 4.3.3

## Loading required package: carData

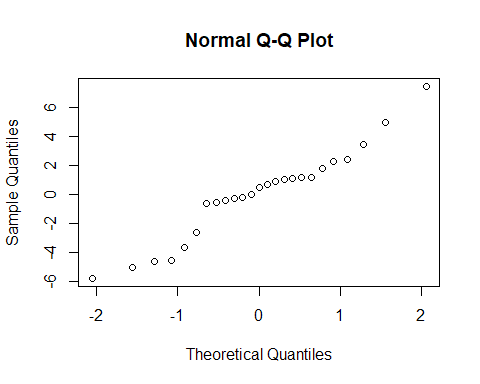
## Warning: package 'carData' was built under R version 4.3.3

vif(mod.reg)

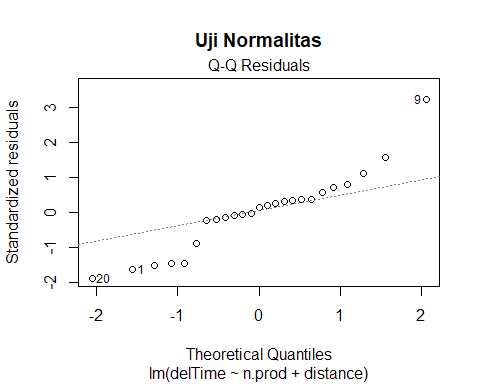
## n.prod distance   
## 3.118474 3.118474

### 3) Asumsi Normalitas

qqnorm(resid(mod.reg))



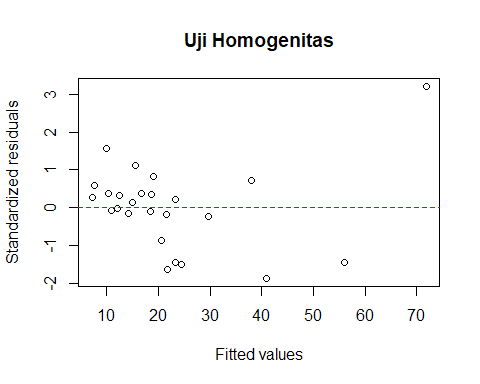
plot(mod.reg,2, main="Uji Normalitas")



Plot Normal Q-Q memperlihatkan bahwa titik-titik menyebar di sekitar garis diagonal dan mengikuti arah garis diagonal sehingga asumsi normalitas terpenuhi.

### 4) Asumsi Homogenitas

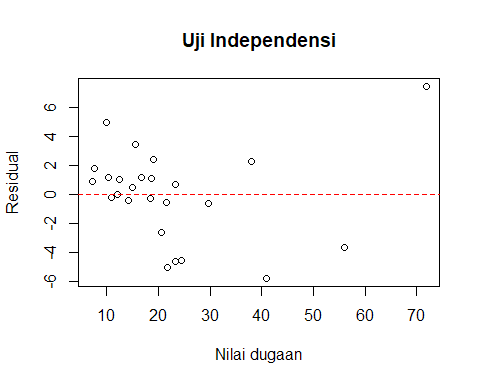
library(MASS)  
plot(fitted(mod.reg),stdres(mod.reg),xlab="Fitted values",ylab="Standardized residuals",main="Uji Homogenitas")  
abline(h=0,lty=2,col="red")



Plot fitted values vs standardized residuals di atas menunjukkan bahwa tidak ada pola yang kuat pada residual dan titik-titik menyebar secara acak di atas dan di bawah sumbu 𝑌=0 (dalam pita horizontal) sehingga asumsi homogenitas variansi terpenuhi.

### 5) Asumsi Independen

plot(fitted(mod.reg),resid(mod.reg),xlab="Nilai dugaan",ylab="Residual",main="Uji Independensi")  
abline(h=0,lty=2,col="red")



Plot uji independensi di atas menunjukkan bahwa titik-titik menyebar di sekitar garis horizontal 0, maka galat saling independen. Dapat disimpulkan bahwa asumsi independen terpenuhi.

# 3. Inferensi bagi parameter (pengujian hipotesis atau interval kepercayaan)

# Uji hipotesis bagi signifikansi regresi linear ganda  
anova\_delivery <- anova(mod.reg)  
anova\_delivery

## Analysis of Variance Table  
##   
## Response: delTime  
## Df Sum Sq Mean Sq F value Pr(>F)   
## n.prod 1 5382.4 5382.4 506.619 < 2.2e-16 \*\*\*  
## distance 1 168.4 168.4 15.851 0.0006312 \*\*\*  
## Residuals 22 233.7 10.6   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

qf(0.95,df1=2,df2=22)

## [1] 3.443357

summary(mod.reg)

##   
## Call:  
## lm(formula = delTime ~ n.prod + distance)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.7880 -0.6629 0.4364 1.1566 7.4197   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.341231 1.096730 2.135 0.044170 \*   
## n.prod 1.615907 0.170735 9.464 3.25e-09 \*\*\*  
## distance 0.014385 0.003613 3.981 0.000631 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.259 on 22 degrees of freedom  
## Multiple R-squared: 0.9596, Adjusted R-squared: 0.9559   
## F-statistic: 261.2 on 2 and 22 DF, p-value: 4.687e-16

# 4. Pilih model regresi terbaik dengan salah satu metode seleksi

fit.stat <- function(model, data, MSE.all) {  
mod.fit <- lm(formula=model, data=data)  
sum.fit <- summary(mod.fit)  
n <- length(mod.fit$residuals)  
aic.mod <- AIC(object = mod.fit, k = 2)  
sbc.mod <- AIC(object = mod.fit, k = log(n))  
sse <- anova(mod.fit)$"Sum Sq"[length(anova(mod.fit)$"Sum Sq")]  
p <- length(mod.fit$coefficients)  
Cp <- sse/MSE.all - (n-2\*p)  
press <- sum(mod.fit$residuals^2/(1-lm.influence(mod.fit)$hat)^2)  
data.frame(p=p, SSE=sse, Rsq=sum.fit$r.squared, AdjRsq=sum.fit$adj.r.squared,  
 AIC.stat=aic.mod, SBC.stat=sbc.mod, PRESS=press, Cp=Cp)  
}

dat <- read.csv(file="C:/Users/LENOVO/OneDrive/Documents/Learning R/Semester 2/delivery.csv", head=T)  
str(dat)

## 'data.frame': 25 obs. of 4 variables:  
## $ X : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ n.prod : int 7 3 3 4 6 7 2 7 30 5 ...  
## $ distance: int 560 220 340 80 150 330 110 210 1460 605 ...  
## $ delTime : num 16.7 11.5 12 14.9 13.8 ...

# Menghitung MSE.all  
MSE.all <- mean((resid(mod.reg))^2)  
  
# Model dengan 2 variabel prediktor  
mod2.12 <- lm(log(delTime)~n.prod+distance, data=dat)  
fit.mod2.12 <- fit.stat(mod2.12, data=dat, MSE.all = MSE.all)  
fit.mod2.12

## p SSE Rsq AdjRsq AIC.stat SBC.stat PRESS Cp  
## 1 3 0.6267517 0.9055505 0.8969642 -13.20509 -8.329586 0.9040645 -18.93296

# Model dengan 1 variabel prediktor  
mod1.1 <- lm(log(delTime)~n.prod, data=dat)  
fit.mod1.1 <- fit.stat(mod1.1, data=dat, MSE.all = MSE.all)  
fit.mod1.1

## p SSE Rsq AdjRsq AIC.stat SBC.stat PRESS Cp  
## 1 2 0.8493149 0.872011 0.8664462 -7.608101 -3.951474 1.02719 -20.90916

mod1.2 <- lm(log(delTime)~n.prod, data=dat)  
fit.mod1.2 <- fit.stat(mod1.2, data=dat, MSE.all = MSE.all)  
fit.mod1.2

## p SSE Rsq AdjRsq AIC.stat SBC.stat PRESS Cp  
## 1 2 0.8493149 0.872011 0.8664462 -7.608101 -3.951474 1.02719 -20.90916

# Model dengan 1 variabel prediktor  
mod0 <- lm(log(delTime)~1,data= dat)  
fit.mod0 <- fit.stat(mod0, data = dat, MSE.all = MSE.all)  
fit.mod0

## p SSE Rsq AdjRsq AIC.stat SBC.stat PRESS Cp  
## 1 1 6.635841 0 0 41.78717 44.22492 7.200348 -22.29023

# Hasil  
result<-rbind(fit.mod0,fit.mod1.1,fit.mod1.2,fit.mod2.12)  
x.var <- c("Tanpa","X1","X2","X1X2")  
cbind(x.var,result)

## x.var p SSE Rsq AdjRsq AIC.stat SBC.stat PRESS  
## 1 Tanpa 1 6.6358411 0.0000000 0.0000000 41.787167 44.224918 7.2003484  
## 2 X1 2 0.8493149 0.8720110 0.8664462 -7.608101 -3.951474 1.0271897  
## 3 X2 2 0.8493149 0.8720110 0.8664462 -7.608101 -3.951474 1.0271897  
## 4 X1X2 3 0.6267517 0.9055505 0.8969642 -13.205089 -8.329586 0.9040645  
## Cp  
## 1 -22.29023  
## 2 -20.90916  
## 3 -20.90916  
## 4 -18.93296

#### Model regresi terbaik

mod2.12 <- lm(log(delTime)~n.prod+distance, data= dat)  
summary(mod2.12)

##   
## Call:  
## lm(formula = log(delTime) ~ n.prod + distance, data = dat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.37362 -0.11922 -0.00478 0.14208 0.22798   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.2935633 0.0567922 40.385 < 2e-16 \*\*\*  
## n.prod 0.0509854 0.0088412 5.767 8.42e-06 \*\*\*  
## distance 0.0005229 0.0001871 2.795 0.0106 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1688 on 22 degrees of freedom  
## Multiple R-squared: 0.9056, Adjusted R-squared: 0.897   
## F-statistic: 105.5 on 2 and 22 DF, p-value: 5.336e-12

Persamaan regresi dugaan:

Interpretasi dari hasil analisis regresi untuk model dengan dua variabel prediktor (n.prod dan distance) yang memprediksi logaritma waktu pengiriman () adalah sebagai berikut:

1. Nilai :

* Nilai intercept (konstanta) memiliki nilai sebesar 2.294 dengan standar error sekitar 0.057. Tanda positif pada koefisien menunjukkan bahwa ketika semua variabel independen (n.prod dan distance) bernilai 0 atau tidak mengalami perubahan, maka nilai logaritma waktu pengiriman () diperkirakan sekitar 2.294.

1. Koefisien regresi untuk variabel n.prod:

* Nilai koefisien untuk n.prod adalah sebesar 0.051 dengan standar error sekitar 0.009. Nilai t-statistic sebesar 5.767 menunjukkan bahwa koefisien ini signifikan secara statistik pada tingkat signifikansi 5%.
* Koefisien ini menunjukkan bahwa dengan setiap peningkatan satu unit dalam jumlah produk (n.prod), logaritma waktu pengiriman () diperkirakan akan meningkat sebesar 0.051, dengan menjaga variabel distance konstan.

1. Koefisien regresi untuk variabel distance:

* Nilai koefisien untuk distance adalah sebesar 0.001 dengan standar error sekitar 0.0002. Nilai t-statistic sebesar 2.795 menunjukkan bahwa koefisien ini juga signifikan secara statistik pada tingkat signifikansi 5%.
* Koefisien ini menunjukkan bahwa dengan setiap peningkatan satu unit dalam jarak (distance), logaritma waktu pengiriman () diperkirakan akan meningkat sebesar 0.001, dengan menjaga variabel n.prod konstan.

Dari hasil analisis tersebut, dapat disimpulkan bahwa kedua variabel prediktor, yaitu jumlah produk (n.prod) dan jarak (distance), berkontribusi secara signifikan terhadap prediksi logaritma waktu pengiriman (). Semakin tinggi jumlah produk dan jarak, semakin lama waktu pengiriman yang diperkirakan, dengan memperhitungkan pengaruh kedua variabel tersebut secara bersama-sama.